Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS

TRABAJO FIN DE MASTER

ESTUDIO RETROSPECTIVO DEL SÍNDROME CORONARIO AGUDO

AUTOR: GUSTAVO ALBERTO ZARAGOZA GARCÍA

TUTORES: VALERO LAPARRA PÉREZ-MUELAS

ANTONIO FÉLIX DE CASTRO

SEPTIEMBRE 2024

Índice

I Introducción **3**

II Obtención de los datos **5**
II.1 Petición inicial de los datos 5
II.2 Fuentes de información y datos solicitados 6

III Limpieza de los datos y construcción del corpus de información **9**
III.1 Conjuntos de datos brutos y limpieza 9
III.1.1 Variables socio-demográficas 9
III.1.2 Nivel de cronicidad 10
III.1.3 Datos Atención Continuada de Atención Primaria 11
III.1.4 Datos Hospitalizaciones 15
III.1.5 Datos Urgencias Hospitalarias 19
III.1.6 Datos Emergencias 23
III.2 Construcción del conjunto de episodios 25
III.2.1 Agrupar los episodios por SIPs 26
III.2.2 Agrupar los episodios por fecha 26
III.2.3 Observaciones del objeto final 28

IV Perfil socio-demográfico del paciente de SCA **31**
IV.1 Extracción de los datos 31
IV.1.1 Recorrido asistencial 31
IV.1.2 Recorrido diagnóstico 32
IV.1.3 Lugar y Edad 32
IV.1.4 Sexo 33
IV.1.5 Alta 33
IV.1.6 Unión de toda la información extraída en un dataframe de *pandas* 34
IV.2 Análisis de los datos 35
IV.2.1 Datos cruzados con los departamentos de salud 46
IV.2.2 Datos cruzados con grupos de edad 51
IV.2.3 Datos acumulados por sexo 56

V Conclusiones y temas de investigación futuros **59**

Resumen

El Síndrome Coronario Agudo es una de las expresiones clínicas más frecuentes de la enfermedad cardiovascular. Al ser ésta una de las patologías más frecuentes en el mundo desarrollado, existe una gran cantidad de información al respecto recopilada en los entes públicos. Si bien existen diversos estudios sobre ésta, en la Conselleria de Sanitat de la Generalitat Valenciana en particular no hay muchas líneas de investigación abiertas. Esto se debe a la complicación para obtener estos datos y la diversidad de fuentes de información que deben unirse para poder trabajar con ellos. Es por eso que surge la idea de este trabajo, donde pasamos el tedioso proceso de solicitud de los datos, para poder unirlos. Una vez con el conjunto de datos accesible y ordenado, nos centramos sobretodo en los perfiles de los pacientes y los protocolos llevados a cabo tanto por pacientes como por profesionales, para poder extraer información que pueda utilizarse desde los organismos públicos para poder mejorar el servicio.

Resum

El Síndrome Coronari Agut és una de les expressions clíniques més freqüents de la malaltia cardiovascular. En ser aquesta una de les patologies més comunes al món desenvolupat, existeix una gran quantitat d'informació al respecte recopilada als organismes públics. Tot i que hi ha diversos estudis sobre aquesta, a la Conselleria de Sanitat de la Generalitat Valenciana en particular no hi ha moltes línies d'investigació obertes. Això es deu a la complicació per obtenir aquestes dades i a la diversitat de fonts d'informació que s'han d'unir per poder treballar amb elles. És per això que sorgeix la idea d'aquest treball, on passem pel tediós procés de solůlicitud de les dades per poder unir-les. Una vegada tenim el conjunt de dades accessible i ordenat, ens centrem sobretot en els perfils dels pacients i els protocols seguits tant per pacients com per professionals, per tal d'extraure informació que puga ser utilitzada pels organismes públics per millorar el servei.

Abstract

Acute Coronary Syndrome is one of the most frequent clinical expressions of cardiovascular disease. As it is one of the most common pathologies in the developed world, there is a large amount of information on this topic collected by public entities. Although there are various studies on it, in the Department of Health of the Generalitat Valenciana, in particular, there are not many open lines of research. This is due to the difficulty of obtaining this data and the diversity of information sources that need to be combined in order to work with them. This is why the idea for this work arose, where we go through the tedious process of requesting the data in order to unify them. Once the data set is accessible and organized, we focus mainly on the profiles of the patients and the protocols followed by both patients and professionals, in order to extract information that can be used by public organizations to improve the service.

I Introducción

La principal causa de muerte en el mundo desarrollado a día de hoy, es la enfermedad cardiovascular, que actualmente, es responsable de 1.8 millones de fallecimientos anuales, es decir, alrededor del 20 % de las muertes en Europa. En particular, una de sus expresiones clínicas más frecuentes es el Síndrome Coronario Agudo (SCA), tanto con elevación del segmento ST como sin ella. La mortalidad de esta enfermedad depende de factores muy diversos, como pueden ser, entre otras, la edad del paciente, sus patologías previas, o el tiempo transcurrido hasta su tratamiento y los protocolos aplicados. Es sabido que aproximadamente la mitad de los pacientes que sufren un SCA fallecen dentro de la primera hora desde su comienzo, antes incluso de poder ser trasladados a un centro hospitalario. Es por esto que es relevante estudiar los motivos por los que algunos pacientes no solicitan ayuda a su debido tiempo, cuales son los factores de prevención que pueden tenerse en cuenta para poder actuar a tiempo, y, en general, cuál es el recorrido trazado por pacientes y médicos, tanto antes de sentir los síntomas fuertes como tras el aviso de urgencia.

Diversos estudios han sido realizados a lo largo de los últimos 100 años acerca de la enfermedad cardiovascular. Probablemente el más conocido de ellos es el Framingham Heart Study (FHS) [2], fruto de la colaboración entre National Heart, Lung and Blood Institute y la Universidad de Boston. Este estudio, el cual sigue en activo tras más de 70 años de investigación y se encuentra actualmente estudiando a la tercera generación de pacientes, se centra en encontrar factores de riesgo para la enfermedad cardiovascular. Algunos de los que ha encontrado han sido factores tan comunes en la población como el consumo de tabaco, las altas presiones arteriales o los altos niveles de colesterol.

Con el objetivo de contextualizar nuestra investigación, es interesante también tomar en cuenta estudios epidemiológicos centrados en España. La Revista Española de Cardiología publicó en 2013 un estudio epidemiológico del SCA [4], que estimaba el número de casos y la tendencia de esta enfermedad de 2005 a 2049. Este estudio preveía un aumento de los casos de SCA en España en las décadas vinientes, causado en gran parte por envejecimiento general de la población. Conclusiones muy parecidas se podían leer en este otro artículo publicado en 2014 en la misma revista [5], el cual añadía como factores importantes para el aumento de casos de enfermedades cardiovasculares la inmigración y la globalización de la dieta occidental, aunque haciendo hincapié en el descenso progresivo de la mortalidad.

Si bien existen muchas líneas de investigación abiertas a este respecto, sigue habiendo diversos campos en los que la investigación todavía no ha progresado tanto, como lo es el comportamiento del paciente que presenta síntomas compatibles con el SCA. Es importante poder conocer el perfil del paciente que sufre de esta afección, para poder tener en cuenta el mayor número de factores posibles de cara a la lucha contra esta enfermedad. Es por ello que surge la idea de realizar esta investigación, focalizada en este campo todavía por conocer.

En el caso concreto de esta investigación, nos dedicaremos a responder las siguientes preguntas: ¿Cuáles son los perfiles demográficos más comunes de los pacientes afectados por el SCA? ¿Cuál es el recorrido asistencial por el que son atendidos? ¿Es éste correcto en una gran mayoría de los casos? ¿Cuál es la vía de entrada del paciente al sistema sanitario? ¿Es la mortalidad muy elevada dentro de los pacientes atendidos? ¿Cómo varían las respuestas a estas preguntas dependiendo del lugar de procedencia, la edad, la vía de entrada o el sexo del paciente?

Además de estas investigaciones, tal y como se verá a lo largo del trabajo, los datos relativos a un único episodio de infarto en la Conselleria de Sanitat se encuentran dispersos entre diversas fuentes de información. Esto se debe a que, a lo largo de un único suceso, el paciente puede pasar por distintos servicios sanitarios, los cuales toman constancia de los pacientes que los utilizan de manera independiente a los demás. Uno de los objetivos principales de este trabajo será la elaboración de un correcto corpus de información, capaz de almacenar todos los episodios relacionados con infartos en la Comunitat Valenciana entre 2016 y 2019. De esta manera, podremos transferir el código y los resultados a los trabajadores de la Conselleria para que estos puedan realizar sus propias investigaciones, además de poder construir nuevos conjuntos de información de manera análoga.

Es importante reseñar que este trabajo se enmarca en una investigación de mayor dimensión llevada a cabo junto con la propia Conselleria de Sanitat. Es por ello, que en este trabajo cubriremos una parte de ella, junto con la construcción del corpus de la información. Existen diversos enfoques de investigación que se están llevando a cabo con este conjunto de datos, que no serán mencionadas en este documento.

Dividiremos este trabajo en cuatro secciones principales. La primera de ellas se centrará en la obtención de los datos. Estos fueron solicitados a la Generalitat Valenciana, y hubo que seguir un tedioso proceso para obtenerlos. En segundo lugar, colocaremos el foco en la limpieza de los datos y creación del corpus de la información para permitir su posterior manipulación. En tercer lugar, trataremos de responder a las problemáticas propuestas sobre el tipo de pacientes que sufren de esta afección, y extraeremos los datos relevantes para resolverlas. Por último, comentaremos los resultados del análisis exploratorio de estos y extraeremos las métricas más adecuadas para poder responder a las preguntas planteadas.

II Obtención de los datos

II.1 Petición inicial de los datos

El primer paso en la realización de este trabajo consistió en la obtención de los datos. Al ser estos datos de carácter público, tuvimos que realizar una petición de datos a la Generalitat Valenciana. En particular, puesto que se trata de datos del ámbito sanitario, tuvimos que realizar la solicitud a la Conselleria de Sanitat, al sistema PROSIGA (SIA-GAIA).

PROSIGA (SIA-GAIA) es el sistema de información de la asistencia ambulatoria de la Conselleria de Sanidad. A través de este canal el personal investigador puede solicitar los datos que requiera de la historia clínica de pacientes, tanto de atención primaria como de especializada.

Para realizar esta solicitud, tuvimos que reunir la siguiente documentación:

1. Visto bueno de la Comisión Ética del Departamento, en este caso la CEIC de la Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (FISABIO).

2. Memoria del proyecto.

3. Visto Bueno del Gerente o Comisionado (en caso de que los datos se refieran a un solo Departamento).

4. Visto Bueno de la Dirección General de Asistencia Sanitaria.

Por motivos de confidencialidad, únicamente podremos adjuntar a este trabajo la Memoria del proyecto (*páginas 1 a 13 de los documentos anexos al trabajo*), que se corresponde con el PIC (Protocolo de Investigación Clínica) que detallaremos más adelante.

El primer paso, y el más largo, fue lograr el visto bueno del CEIC (Comité de Ética de Investigación Clínica) de FISABIO. Para ello, tuvimos que presentar una serie de requisitos, detallados en los Procedimientos Normalizados de Trabajo (de ahora en adelante PNT) facilitados por la propia fundación en su página web [3].

Para obtener este visto bueno, tuvimos que presentar un total de 10 documentos, algunos redactados por nosotros mismos, y otros rellenando las plantillas facilitadas en los anexos de los PNT. Los documentos presentados son los siguientes:

1. Solicitud de evaluación (*anexo 0 de los PNT*): formulario a cumplimentar para solicitar la evaluación al CEI.

2. Compromiso del investigador principal (Gustavo Alberto Zaragoza García) (*anexo 2 de los PNT*): documento de compromiso con la investigación que deben firmar investigador/a principal e investigadores/as colaboradores/as.

3. Compromiso del colaborador (Antonio Félix De Castro) (*anexo 3 de los PNT*): documento de compromiso con la investigación que deben firmar investigador/a principal e investigadores/as colaboradores/as. Equivalente al anterior punto con los colaboradores.

4. Compromiso del colaborador (Valero Laparra Pérez-Muelas) (*anexo 3 de los PNT*): documento de compromiso con la investigación que deben firmar investigador/a principal e investigadores/as colaboradores/as. Equivalente al anterior punto con los colaboradores.

5. Compromiso del investigador principal y los investigadores colaboradores (disponible en el sitio web de FISABIO [3]): documento de compromiso con la investigación que deben firmar investigador/a principal e investigadores/as colaboradores/as. Muy similar al segundo documento, pero mucho más detallado. Presentamos ambas versiones, puesto que no teníamos claro si era necesario presentarlas por separado o era suficiente con una.

6. Formulario de datos anonimizados (disponible en el sitio web de FISABIO [3]): Informe de seguridad y riesgo de reidentificación en solicitudes de cesion de datos anonimizados o pseudoanonimizados firmado por el/la investigardor/a principal.

7. Currículum Vitae firmado y fechado del investigador principal Gustavo Alberto Zaragoza García.

8. Currículum Vitae firmado y fechado del investigador colaborador Antonio Félix de Castro.

9. Currículum Vitae firmado y fechado del investigador colaborador Valero Laparra Pérez-Muelas.

10. Protocolo de Investigación Clínica (PIC) (*páginas 1 a 13 de los documentos anexos al trabajo*): protocolo completo (no resumen) con el proyecto del estudio definitivo, siguiendo la estructura del punto 9 de los PNT.

Una vez obtenido el visto bueno del CEIC de Fisabio, el resto de pasos fueron más sencillos. Presentamos este visto bueno junto con el PIC, que explicaba en detalle el objetivo de la investigación, y las aprobaciones tanto del comisionado como de la dirección general fueron cuestión de resolver los trámites administrativos. Una vez reunidos todos los documentos, realizamos la solicitud y rápidamente nos fueron remitidos los datos por correo electrónico directamente por la Generalitat Valenciana.

II.2 Fuentes de información y datos solicitados

Para la realización de este trabajo, tuvimos que solicitar datos de tres fuentes distintas del sistema sanitario. Las fuentes son las siguientes:

1. El sistema de información del Servicio de Emergencias Sanitarias de la Comunitat Valenciana (SESCV), que proporciona los episodios clasificados como dolor torácico, los episodios clasificados como dificultad respiratoria asociada a dolor torácico, y los episodios con diagnóstico registrado compatible con SCA.

2. Los hospitales facilitan los datos provenientes de dos fuentes diferentes:

   (a) Episodios de urgencias: se requieren datos de los episodios con diagnóstico registrado compatible con SCA.

(b) Episodios de hospitalación: se requieren datos de los episodios de con diagnóstico principal registrado compatible con SCA o con procedimiento principal de angioplastia realizado.

3. Atención Primaria suministra los datos de los episodios de la Actividad Continuada con diagnóstico registrado compatible con SCA.

Toda la extracción y anonimización de los datos fue llevada a cabo por el Servicio de Análisis de Sistemas de la Información Sanitaria a instancias de la comisión PROSIGA. En cuanto a los datos solicitados para el estudio, se propuso la extracción de los datos con el siguiente protocolo de obtención:

1. Extracción de todos los episodios de hospitalización, urgencias hospitalarias, de Atención Continuada de Atención Primaria y del Servicio de Emergencias Sanitarias de la Comunitat Valenciana con diagnóstico principal del episodio compatible con SCA (pero en este caso se incorporará los casos clasificados como posible SCA en el triaje previo de la llamada al 112).

2. El conjunto de SIP obtenidos de esa extracción se obtendrá todos los episodios de hospitalización, urgencias, emergencias y atención continuada de cada paciente durante los años solicitados del estudio.

El rango temporal en el que se enmarcan los datos empleados en este estudio es del 1 de enero de 2016 al 31 de diciembre de 2019. La lista de atributos solicitados es la siguiente:

1. Del paciente anonimizado individual:

    - Edad.
    - Sexo.
    - Grupo de Cronicidad.
    - Nacionalidad (si existe).

2. De cada episodio de atención sanitaria:

    (a) Hospitalización.
        - Departamento atención.
        - Centro de atención.
        - Servicio de ingreso.
        - Servicio de alta.
        - Fecha de ingreso.
        - Fecha de alta.
        - Diagnóstico (principal y secundarios).
        - Procedimientos realizados (primario y secundarios).
        - Circunstancia al alta.

    (b) Urgencias hospitalarias.
        - Departamento atención.

- Centro de atención.
- Fecha de ingreso en servicio de urgencias.
- Fecha de alta.
- Tiempos: momento inicio triaje, momento final triaje, momento atención, momento ing en observación, momento alta administrativa, momento alta.
- Triaje Manchester.
- Origen entrada (homologado y propio).
- Motivo urgencia (homologado y propio).
- Diagnóstico (principal y secundarios).
- Procedimientos (principal y secundarios).
- Circunstancia al alta.

(c) Servicio de Emergencias Sanitarias de la Comunitat Valenciana.

- Fecha.
- Municipio de atención.
- Departamento de atención.
- Clasificación triaje 112.
- Tiempos de atención: entrada llamada, tiempo asistencia, tiempo de llegada a hospital, tiempo de transferencia.
- Diagnóstico.
- Maniobras.
- Finalización.

(d) Atención Continuada de Atención Primaria.

- Fecha.
- Departamento de salud.
- Centro.
- Diagnóstico.
- Procedimiento.
- Circunstancia al alta.

En el PIC se pueden encontrar más detalles en cuanto a la lista de diagnósticos compatibles con SCA solicitados en la petición, codificados tanto en CIE-9 como en CIE-10 (sistema de codificación de diagnósicos que se cambió en el año 2016).

III Limpieza de los datos y construcción del corpus de información

Una vez nos fueron entregados los datos, el primer paso fue realizar una correcta manipulación de estos para luego poderlos manejar adecuadamente.

III.1 Conjuntos de datos brutos y limpieza

Los datos fueron entregados en seis archivos diferentes, de tipos *.txt* y *.xlsx*. Los describiremos de manera individual para detallar el contenido de cada uno de ellos.

III.1.1 Variables socio-demográficas

Las variables socio-demográficas de los pacientes nos fueron entregadas en un archivo de texto llamado ***SD2628_SCA_VARIABLES_SOCIODEMOGRAFICAS.sql.log.enc.txt***, cuyo tamaño es 2596 KB. Para abrir este archivo, hemos usado el método *read_table* de la librería *pandas*. En este dataset, encontramos 55876 registros, con cinco variables. Existe un valor ausente en cuatro de ellas, todos en la misma fila. Veamos las variables en detalle:

- *NUM_SIP*: Esta variable se corresponde con el número SIP de los 55876 pacientes totales incluidos en cualquiera de los conjuntos de datos entregados. Se identifica con la variable de mismo nombre del resto de fuentes de datos. Es el identificador alfanumérico obtenido después de someter el número SIP de cada paciente al proceso de anonimización de los datos. Puesto que corresponde con un identificador alfanumérico, para trabajar con ella la transformamos en tipo *string* o cadena de texto.

- *COD_SEXO*: Esta variable es una codificación numérica del sexo del paciente en cuestión. Esta variable contiene un *1* si el paciente es un hombre o un *2* si es una mujer. Contiene un valor nulo en la fila 55876, la misma que las demás variables de este conjunto de datos. Puesto que corresponde con una etiqueta, para trabajar con ella la transformamos en tipo *string*.

- *DESC_SEXO*: Esta variable es la descripción del sexo del paciente en cuestión. Contiene el valor *Hombre* si el paciente es un hombre o el valor *Mujer* si es una mujer. Contiene un valor nulo en la fila 55876, la misma que las demás variables. Puesto que corresponde con una etiqueta, para trabajar con ella la transformamos en tipo *string*.

- *FECHA_NACIMIENTO*: Esta variable corresponde a la fecha de nacimiento del paciente en cuestión. Podemos encontrar esta fecha en formato *%d-%b-%y* (es decir, *01-ENE-24*). Contiene un valor nulo en la fila 55876, la misma que las demás variables. Puesto que corresponde con una fecha, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y-%m-%d*, el cual será nuestro formato de referencia para fechas.

- *COD_PAIS*: Esta variable es la codificación numérica de la nacionalidad del paciente en cuestión. Contiene 114 valores numéricos distintos, correspondientes al etiquetado de cada nacionalidad por el INE. Contiene un valor nulo en la

fila 55876, la misma que las demás variables. Puesto que corresponde con una etiqueta, para trabajar con ella la transformamos en tipo *string*.

Únicamente dos transformaciones, más allá de la intervención en los tipos de los datos, ya descrita en el apartado correspondiente a cada variable, han sido realizadas en este dataset en la primera lectura.

En primer lugar, existe un SIP para el cual todos los valores son nulos, en particular, el SIP *K0bIBXlP9pY948d8C3HF8w==*. En este caso, hemos tomado la decisión de no eliminar estos datos, para evitar conflictos a la hora de unir este dataset con los datasets que correspondan a algún episodio sufrido por este paciente. Simplemente sustituiremos los valores ausentes por la cadena de texto *Desconocido*.

En segundo lugar, existen fechas de nacimiento que pueden llevarnos a confusión por el formato de las fechas en el documento de texto de entrada. En particular, esto se debe a que el año está indicado únicamente con dos cifras. En los valores más bajos, no existe manera de saber si se trata de pacientes nacidos a principio del siglo XX o del siglo XXI (por ejemplo, un registro cuya fecha de nacimiento registrada es *01-ENE-00* podría ser nacida en el año 1900 o en el año 2000). Para minimizar este riesgo de error, puesto que el rango de nuestros registros es desde el año 2016 hasta el año 2019, daremos el valor *Desconocido* para la variable *FECHA_NACIMIENTO* en todos aquellos registros cuyo año de nacimiento en los datos sea inferior a 09. De esta manera, los pacientes no serán nunca menores de 16 años (los más jóvenes nacidos en el 1999), ni mayores de 110 años (los más longevos nacidos en 1909).

Se puede encontrar una primera observación generalista de las variables junto con el resultado de las transformaciones realizadas en el notebook *0_e_demograficos_y_cronicidad_csv.ipynb*. Por otra parte, el código correspondiente a las transformaciones realizadas se encuentran en la definición de la función *transformar_tipos_cronicidad_2*, en el archivo *help_functions.py*.

III.1.2 Nivel de cronicidad

En medicina, el concepto crónico se refiere a algo que continúa durante un período de tiempo prolongado. Una enfermedad crónica generalmente dura mucho tiempo y no desaparece en forma rápida o fácil. La Conselleria de Sanidad utiliza un sistema de clasificación de pacientes según el grado de cronicidad de su enfermedad. Esta clasificación nos fue facilitada en un archivo de texto llamado ***SD2628_NIVEL_CRONICIDAD.enc***, cuyo tamaño es 284594 KB, y que contiene el nivel de cronicidad asignado a cada paciente en cada visita a algún centro médico en la cual éste ha sido anotado. Este dataset contiene 3313119 registros y cuatro variables. Además, no encontramos ningún valor nulo. Veamos las variables en detalle:

- *NUM_SIP*: Esta variable se corresponde con el número SIP de los 54326 (número de valores distintos en esta variable) pacientes incluidos en cualquiera de los conjuntos de datos entregados, cuya cronicidad ha sido registrada en el sistema sanitario al menos en una ocasión entre 2016 y 2019. Se identifica con la variable de mismo nombre del resto de fuentes de datos. *NUM_SIP* es el identificador alfanumérico obtenido después de someter el número SIP de cada paciente al proceso de anonimización de los datos. Puesto que corresponde con un identificador alfanumérico, para trabajar con ella la transformamos en tipo *string* o cadena de texto.

- *FECHA_CONSULTA*: Esta variable se corresponde con la fecha de la consulta en la cual se le asignó al respectivo paciente el respectivo nivel de cronicidad. Podemos encontrar esta fecha en formato *%d- %b- %y*. Puesto que corresponde con una fecha, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d*.

- *COD_NVL_CRONI*: Esta variable es la codificación numérica del nivel de cronicidad asignado al paciente en cuestión. Existen cuatro valores distintos asignados a esta variable, correspondientes a cada uno de los niveles de cronicidad, a los que hay que sumar el valor *-1*, asignado cuando está registrado un valor vacío. Esta variable contiene números enteros, pero, al igual que en casos anteriores, al tratarse de una etiqueta, la transformaremos en un string.

- *DESC_NVL_CRONI*: Esta variable es la descripción del nivel de cronicidad asignado al paciente en cuestión. Existen cuatro valores distintos asignados a esta variable, correspondientes a cada uno de los niveles de cronicidad, a los que hay que sumar el valor *[Vacío]*. Puesto que corresponde con una etiqueta, para trabajar con ella la transformamos en tipo *string*.

La intervención en los tipos de los datos, ya descrita en el apartado correspondiente a cada variable, junto con algún ligero cambio de nombre con respecto a las variables originales (*COD_NVL_CRONI* y *DESC_NVL_CRONI* se denominan *COD_NIVEL-_CRONICIDAD* y *DESC_NIVEL_CRONICIDAD* en el documento *txt* original, y fueron modificadas por motivos de practicidad, para que fuesen más cortos) han sido realizadas en este dataset en la primera lectura.

Se puede encontrar una primera observación generalista de las variables junto con el resultado de las transformaciones realizadas en el notebook *0_e_demograficos_y_cronicidad_csv.ipynb*. Por otra parte, el código correspondiente a las transformaciones realizadas se encuentran en la definición de la función *transformar_tipos_cronicidad_2*, en el archivo *help_functions.py*.

III.1.3 Datos Atención Continuada de Atención Primaria

Los datos provenientes de Atención Continuada de Atención Primaria nos fueron entregados en un documento de tipo hoja de cálculo, llamado ***SD2628_URGENCIAS-_ATENCION_PRIMARIA.xlsx***, de tamaño 38057KB. Contiene los datos con las variables solicitadas de aquellos pacientes que han recibido algún diagnóstico compatible con SCA (ya sea en Atención Primaria o en alguna de las otras tres vías de entrada al sistema sanitario tratadas en este trabajo) entre 2016 y 2019; y han hecho uso del sistema de Atención Primaria, ya sea por un motivo relacionado con el SCA o no. Para abrir este archivo, hemos usado el método *read_excel* de la librería *pandas*. Podemos encontrar 556495 registros en total, y 14 columnas en este conjunto de datos. Cinco de éstas contienen valores ausentes, tal y como explicaremos en el punto correspondiente a cada una de ellas (en caso de no mencionar valores ausentes, significa que esa variable no contiene ninguno, aunque puede tener valores indicados como nulos o vacíos). Veamos las variables en detalle:

- *NUM_SIP*: Esta variable se corresponde con el número SIP de los 40939 (número de valores distintos en esta variable) pacientes compatibles con la descripción hecha en el párrafo anterior. Se identifica con la variable de mismo nombre del

resto de fuentes de datos. *NUM_SIP* es el identificador alfanumérico obtenido después de someter el número SIP de cada paciente al proceso de anonimización de los datos. Puesto que corresponde con un identificador alfanumérico, para trabajar con ella la transformamos en tipo *string*.

- *Fecha*: Esta variable se corresponde con la fecha de la consulta en la cual el paciente hizo uso del servicio de Atención Primaria. Podemos encontrar esta fecha en formato *%Y- %m- %d*, que coincide con el formato de referencia antes mencionado. Puesto que corresponde con una fecha, transformaremos esta variable en tipo *datetime64[ns]*, guardando el mismo formato con el que la encontramos.

- *Hora Inicio Consulta*: Esta variable se corresponde con la hora a la cual inició la correspondiente consulta en el servicio de Atención Primaria. La hora está indicada en formato *%H: %M* (es decir, *00:00*). Puesto que esta variable se corresponde con una hora, que por sí sola no nos dará ninguna información, no le realizaremos ninguna transformación y mantendremos esta variable en tipo *object*, tal y como es leída en un inicio por python. No obstante, como explicaremos en futuras secciones, sí la emplearemos como parte de otras operaciones.

- *Hora Fin Consulta*: Esta variable se corresponde con la hora a la cual finalizó la correspondiente consulta en el servicio de Atención Primaria. La hora está indicada en formato *%H: %M* (es decir, *00:00*). Esta variable contiene un total de 10105 valores ausentes, correspondientes a visitas en las que no se tomó la hora de salida. Puesto que esta variable se corresponde con una hora, que por sí sola no nos dará ninguna información, no le realizaremos ninguna transformación y mantendremos esta variable en tipo *object*, tal y como es leída en un inicio por python.

- *Código Departamento*: Esta variable es la codificación numérica del departamento de salud al cual pertenece el centro en el cual se ha realizado la consulta en cuestión. Existen 24 valores distintos asignados a esta variable, uno por cada departamento de salud. Esta variable contiene números enteros, pero, al igual que en casos anteriores, al tratarse de una etiqueta, la transformaremos en un *string*.

- *Descripción Departamento*: Esta variable es la descripción del departamento de salud al cual pertenece el centro en el cual se ha realizado la consulta en cuestión. Existen 24 valores distintos asignados a esta variable, uno por cada departamento de salud. Esta variable contiene cadenas de texto, compuestas por el nombre de cada uno. Puesto que corresponde con una etiqueta, para trabajar con ella la transformamos en tipo *string*.

- *Código Centro*: Esta variable es la codificación numérica del centro de salud en el cual se ha realizado la consulta en cuestión. Existen 213 valores distintos asignados a esta variable, uno por cada centro de salud de la comunidad valenciana. Esta variable contiene números enteros, pero, al igual que en casos anteriores, al tratarse de una etiqueta, la transformaremos en un *string*.

- *Descripción Centro*: Esta variable es la descripción del centro de salud en el cual se ha realizado la consulta en cuestión. Existen 213 valores distintos asignados a esta variable, uno por cada centro de salud. Esta variable contiene cadenas de

texto, compuestas por el nombre de cada centro. Puesto que corresponde con una etiqueta, para trabajar con ella la transformamos en tipo *string*.

- *Código Diagnóstico*: Esta variable es la codificación del diagnóstico asignado al paciente tras la consulta en Atención Primaria. Esta codificación se rige mediante los códigos CIE-9 o CIE-10, ya mencionados anteriormente en este trabajo. Existen 5723 valores distintos asignados a esta variable. Existen un total de 127618 valores ausentes en esta variable, correspondientes a consultas sin diagnóstico conocido. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Descripción Diagnóstico*: Esta variable es la descripción del diagnóstico asignado al paciente tras la consulta en Atención Primaria. Esta descripción se corresponde al diagnóstico codificado mediante los sistemas CIE-9 o CIE-10 presentes en la columna anterior. Existen 5598 valores distintos asignados a esta variable. Existen un total de 127618 valores ausentes en esta variable, correspondientes a consultas sin diagnóstico conocido. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Código Procedimiento*: Esta variable es la codificación del procedimiento realizado al paciente durante la consulta en Atención Primaria. Esta codificación se puede regir mediante los códigos CIE-9 o CIE-10, ya mencionados anteriormente en este trabajo, así como mediante la codificación NIC (Nurse International Codification), que clasifica las intervenciones de enfermería. Existen 1256 valores distintos asignados a esta variable. Existen un total de 291227 valores ausentes en esta variable, correspondientes a consultas en las cuales no se ha sometido al paciente a ningún procedimiento o no se ha dejado constancia de ello. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Descripción Procedimiento*: Esta variable es la descripción del procedimiento realizado al paciente durante la consulta en Atención Primaria. Esta descripción se corresponde al procedimiento codificado mediante los sistemas CIE-9, CIE-10 o NIC presentes en la columna anterior. Existen 1229 valores distintos asignados a esta variable. Existen un total de 291227 valores ausentes en esta variable, correspondientes a consultas en las cuales no se ha sometido al paciente a ningún procedimiento o no se ha dejado constancia de ello. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Código Circunstancia de Alta*: Esta variable es la codificación numérica de la circunstancia en la cual se le ha dado el alta de Atención Primaria al paciente en una consulta. Esta codificación es arbitraria, y es la misma en los datos correspondientes a las demás etapas del sistema sanitario. En este dataset, tenemos 4 valores distintos para esta columna. En otras fuentes de datos veremos que existen hasta 21 codificaciones distintas para esta variable. Podemos encontrar un total de 0 valores ausentes. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Descripción Circunstancia de Alta*: Esta variable es la descripción de la circunstancia en la cual se le ha dado el alta de Atención Primaria al paciente en una consulta. Al igual que en casos anteriores, esta columna va emparejada con su código equivalente. En este dataset, tenemos 4 valores distintos para esta columna.

Al igual que con su código correspondiente, en otras fuentes de datos veremos que existen hasta 21 descripciones distintas para esta variable. Podemos encontrar un total de 0 valores ausentes. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

Más allá de la intervención en los tipos de los datos, ya descrita en el apartado correspondiente a cada variable, en esta primera lectura únicamente hemos realizado una transformación en las variables pertenecientes a este dataset. Hemos creado la columna denominada *FECHA_REFERENCIA*, compuesta de la concatenación de las columnas *Fecha* y *Hora Inicio Consulta*. Esta columna será creada en las cuatro fuentes de datos que nos han sido facilitadas, con el objetivo de tomar esta fecha como referencia temporal a la hora de construir un episodio de infarto (proceso que explicaremos más adelante). Esta nueva variable es de tipo *datetime64[ns]* y su formato es *%Y-%m-%d %H:%M:%S*.

También es importante reseñar, aunque no fuese realizada en este caso ninguna modificación, que existe una diferencia entre el número de valores diferentes en las variables *Código Diagnóstico* (5723) y *Descripción Diagnóstico* (5598), a pesar de referirse ambas al diagnóstico recibido por el paciente. Esta diferencia se debe a que una misma descripción de diagnóstico puede haber sido asignada a distintos códigos, ya sea porque haya sido codificado en algunos episodios en CIE-9 y en otros en CIE-10, por referirse a diagnósticos ligeramente diferentes pero en esencia similares. Un ejemplo de este último supuesto es el caso del diagnóstico *AMIGDALITIS CRONICA*, que podemos encontrar con los códigos *474.00*, el diagnóstico específico de amigdalitis crónica, y *474.0*, el diagnóstico general para amigdalitis y adenoiditis crónicas. Otro podría ser el caso del diagnóstico *CEGUERA Y BAJA VISION-NIVEL DE DETERIORO SIN MAS ESPECIFICACION*, que podemos encontrar con los códigos *369.00*, que corresponde a la ceguera sin especificar, *369.20*, que corresponde a la ceguera de ambos ojos, o *369.70*, que corresponde a la ceguera de un solo ojo. Hay un total de 124 diagnósticos afectados. El caso inverso (un mismo código y dos descripciones distintas) no existe en este dataset. No realizamos ninguna modificación, puesto que la información almacenada en los códigos puede ser distinta de un registro a otro pese a estar acompañada de la misma descripción, y por lo tanto hacer cambios supondría perder información.

Al igual que en el caso de los diagnósticos, existe una diferencia muy similar entre el número de valores únicos en las variables *Código Procedimiento* (1256) y *Descripción Procedimiento* (1229). Ésta se puede deber a que una misma descripción de procedimiento puede haber sido asignada a distintos códigos. No obstante, una misma descripción también puede referirse a procedimientos ligeramente diferenciados por un pequeño detalle, o a mismos procedimientos en pacientes con diferentes condiciones. Un ejemplo de este último supuesto es el caso del procedimiento *COMPROBAR INDICACION Y POSIBLES CONTRAINDICACIONES*, que podemos encontrar con los códigos *0910.99* en el caso de estar en un supuesto de realizar una estabilización, inmovilización y/o protección de una parte corporal del paciente; o *4039.99*, en el caso de estarle realizando un electrocardiograma. Hay un total de 21 procedimientos afectados. El caso inverso (un mismo código y dos descripciones distintas) no existe en este dataset. Por el mismo motivo que en el caso de los diagnósticos, no realizamos ninguna modificación.

Tras la primera lectura, hemos unido este dataframe con el correspondiente a los datos demográficos, de manera a tener unificados en uno las variables de atención primaria con las demográficas. Para esto, una operación de tipo *left join*, tomando

como referencia izquierda el dataframe de atención primaria y derecha el de datos demográficos; y como clave la variable *NUM_SIP*, ha sido suficiente y no ha supuesto complicación alguna.

Posteriormente, hemos hecho la misma operación para unir estos datos resultantes con los datos de cronicidad. Nuevamente una operación análoga a la anterior, sustituyendo los datos demográficos por los de cronicidad, ha sido suficiente para realizar esta unión de manera exitosa.

El resultado de estas transformaciones y uniones ha sido almacenado en un documento *csv*, mucho más fácilmente manejable, denominado *at_primaria_para_grafs_con_cron.csv*.

Se puede encontrar una primera observación generalista de las variables junto con el resultado de las transformaciones realizadas y las uniones de los distintos conjuntos de datos en el notebook *0_a_at_prim_csv.ipynb*. Por otra parte, el código correspondiente a las transformaciones realizadas se encuentran en la definición de la función *transformar_tipos_UAP_2*, en el archivo *help_functions.py*.

III.1.4 Datos Hospitalizaciones

Los datos provenientes de hospitalizaciones nos fueron entregados en un documento de texto, llamado ***SD2628_Hospit_CMBD.enc.txt***, de tamaño 118489KB. Contiene los datos con las variables solicitadas de aquellos pacientes que han recibido algún diagnóstico compatible con SCA (ya sea siendo hospitalizados o en alguna de las otras tres vías de entrada al sistema sanitario tratadas en este trabajo) entre 2016 y 2019; y han sido hospitalizados o han recibido algún tipo de intervención en la unidad de cardiología de algún hospital (sin necesidad de ser ingresados), ya sea por un motivo relacionado con el SCA o no. Para abrir este archivo, hemos usado el método *read_excel* de la librería *pandas*. Podemos encontrar 135229 registros y 75 columnas en este conjunto de datos, en las cuales no encontraremos valores ausentes, aunque sí tal vez valores marcados como nulos o vacíos. Veamos su composición en detalle:

- *NUM_SIP*: Esta variable se corresponde con el número SIP de los 51202 (número de valores distintos en esta variable) pacientes compatibles con la descripción hecha en el párrafo anterior. Se identifica con la variable de mismo nombre del resto de fuentes de datos. *NUM_SIP* es el identificador alfanumérico obtenido después de someter el número SIP de cada paciente al proceso de anonimización de los datos. Puesto que corresponde con un identificador alfanumérico, para trabajar con ella la transformamos en tipo *string*.

- *COD_HOSPITAL*: Esta variable es la codificación numérica del hospital en el cual se ha realizado el ingreso en cuestión. Existen 33 valores distintos asignados a esta variable, uno por cada hospital de la Comunitat Valenciana, que son números enteros del 1 al 33. Esta variable contiene números enteros, pero, al igual que en casos anteriores, al tratarse de una etiqueta, la transformaremos en un *string*.

- *DESC_HOSPITAL*: Esta variable es la descripción del hospital en el cual se ha realizado el ingreso en cuestión. Existen 33 valores distintos asignados a esta variable, uno por cada hospital de la Comunitat Valenciana, correspondientes a los nombres de los mismos. Al tratarse de una etiqueta, la transformaremos en un *string*.

- *COD_DEPARTAMENTO*: Esta variable es la codificación numérica del departamento de salud al cual pertenece el hospital en el cual se ha realizado el ingreso en cuestión. Existen 31 valores distintos asignados a esta variable: los números 1 a 24 por cada uno de los departamentos de salud, más los valores: *80* para referirse al hospital provincial de Castellón, *HACLE01* para referirse al hospital de la Magdalena, *HACLE02* para referirse al hospital de Doctor Moliner, *HACLE03* para referirse al hospital Padre Jofré, *HACLE04* para referirse al hospital de Sant Vicent del Raspeig, *HACLE05* para referirse al hospital de la Pedrera y *HACLE06* para referirse al hospital de la Crónicos de Mislata. Esta variable contiene números enteros y cadenas de texto, y además, al tratarse de una etiqueta, la transformaremos en un *string*.

- *DESC_DEPARTAMENTO*: Esta variable es la descripción del departamento de salud al cual pertenece el centro en el cual se ha realizado la consulta en cuestión. Existen 31 valores distintos asignados a esta variable: los nombres de cada uno de los departamentos de salud, más los valores: *HOSPITAL PROVINCIAL DE CASTELLÓN*, *Hospital La Magdalena*, *Hospital Doctor Moliner*, *Hospital Pare Jofré*, *Hospital Sant Vicent del Raspeig*, *Hospital La Pedrera* y *Hospital de Crónicos de Mislata (Antiguo Hospital Militar de Valencia)*. Puesto que corresponde con una etiqueta, para trabajar con ella transformamos esta variable en tipo *string*.

- *COD_SERV_INGRESO*: Esta variable es la codificación del servicio del hospital a través del cual se ha realizado el ingreso. Contiene 79 valores distintos asignados a esta variable, compuestos por un código de tres letras que representa cada servicio, además del valor *'-'*, que interpretamos como la ausencia de servicio. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *COD_SERV_ALTA*: Esta variable es la codificación del servicio del hospital a través del cual se ha realizado el alta. Contiene 78 valores distintos asignados a esta variable, compuestos por un código de tres letras que representa cada servicio, además del valor *'-'*, que interpretamos como la ausencia de servicio. Estos valores son los mismos que en la variable anterior, a excepción de los valores *'DIE'*, *'UDA'* y *'-'*, presentes entre los servicios de ingresos y no los de altas, y los valores *'DII'* y *'GER'*, presentes entre los servicios de altas y no los de ingresos. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *FECHA_INGRESO*: Esta variable se corresponde con la fecha del ingreso en el hospital. Podemos encontrar esta fecha en formato *%d- %b- %y %H: %M: %S* (es decir, *1-ENE-2024 00:00:00*), pero la parte correspondiente a la hora es *00:00:00* en todos los registros. Puesto que corresponde con una fecha, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d'*.

- *FECHA_INGRESO_HORA*: Esta variable se corresponde con la hora del ingreso en el hospital. Es un número entero entre el 0 y el 23. Para poder trabajar con esta variable, la cual tendremos que combinar en el futuro con la fecha de ingreso, la transformaremos en tipo *string*.

- *FECHA_INGRESO_MINUTO*: Esta variable se corresponde con el minuto del ingreso en el hospital. Es un número entero entre el 0 y el 59. Para poder trabajar

con esta variable, la cual tendremos que combinar en el futuro con la fecha de ingreso, la transformaremos en tipo *string*.

- *FECHA_ALTA*: Esta variable se corresponde con la fecha en la que el paciente recibe el alta. Podemos encontrar esta fecha en formato *%d- %b- %y %H: %M: %S*, pero, al igual que en la variable *FECHA_INGRESO*, la parte correspondiente a la hora es *00:00:00* en todos los registros. Puesto que corresponde con una fecha, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y-%m- %d'*.

- *FECHA_ALTA_HORA*: Esta variable se corresponde con la hora en la que el paciente recibe el alta. Es un número entero entre el 0 y el 23. Para poder trabajar con esta variable, la cual tendremos que combinar en el futuro con la fecha de ingreso, la transformaremos en tipo *string*.

- *FECHA_ALTA_MINUTO*: Esta variable se corresponde con el minuto en el que el paciente recibe el alta. Es un número entero entre el 0 y el 59. Para poder trabajar con esta variable, la cual tendremos que combinar en el futuro con la fecha de ingreso, la transformaremos en tipo *string*.

- *DIAG_PRINCIPAL*: Esta variable contiene la codificación del diagnóstico principal asignado al paciente durante el ingreso mediante los códigos CIE-9 y CIE-10, junto con la descripción del mismo. Existen 5645 valores distintos asignados a esta variable. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- Variables *DIAG_SEC_02, DIAG_SEC_03, ..., DIAG_SEC_30*: Estas variables contienen la codificación de los hasta 30 diagnósticos secundarios asignados al paciente durante su ingreso, junto con las descripciones de los mismos. Gran parte de los registros de estas variables contiene el valor *'-'* que indica la ausencia de diagnósticos secundarios, más todavía cuando más aumenta el número de diagnóstico secundario, puesto que no siempre existen. Estas variables son etiquetas, y por lo tanto las transformaremos en un *string*.

- *PROC_PRINCIPAL*: Esta variable contiene la codificación del procedimiento principal asignado al paciente durante el ingreso mediante los códigos CIE-9, CIE-10 y NIC, junto con la descripción del mismo. Existen 6192 valores distintos asignados a esta variable. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- Variables *PROC_SEC_02, PROC_SEC_03, ..., PROC_SEC_30*: Estas variables contienen la codificación de los hasta 30 procedimientos secundarios asignados al paciente durante su ingreso, junto con las descripciones de los mismos. Gran parte de los registros de estas variables contiene el valor *'-'* que indica la ausencia de procedimientos secundarios, más todavía cuando más aumenta el número de diagnóstico secundario, puesto que no siempre existen. Estas variables son etiquetas, y por lo tanto las transformaremos en un *string*.

- *COD_CIRC_ALTA*: Esta variable es la codificación numérica de la circunstancia en la cual se le ha dado el alta hospitalaria al paciente en una consulta. Esta codificación es arbitraria, y es la misma en los datos correspondientes a las demás

etapas del sistema sanitario. En este dataset, tenemos 15 posibles valores para esta columna. En otras fuentes de datos veremos que existen hasta 21 codificaciones distintas para esta variable. Podemos encontrar un total de 0 valores ausentes. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *DESC_CIRC_ALTA*: Esta variable es la descripción de la circunstancia en la cual se le ha dado el alta hospitalaria al paciente en una consulta. Al igual que en casos anteriores, esta columna va emparejada con su código equivalente. En este dataset, tenemos 15 posibles valores para esta columna. Al igual que con su código correspondiente, en otras fuentes de datos veremos que existen hasta 21 descripciones distintas para esta variable. Podemos encontrar un total de 0 valores ausentes. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

Más allá de la intervención en los tipos de los datos, ya descrita en el apartado correspondiente a cada variable, en esta primera lectura tuvimos que realizar una serie de transformaciones en las variables pertenecientes a este dataset para poder manipular los datos.

En primer lugar, realizamos dos modificaciones en las variables *FECHA_INGRESO* y *FECHA_ALTA* de manera análoga. La primera de ellas consistió en conservar en estas variables tan solo la parte referente a la fecha, eliminando así el *00:00:00* que acompañaba a todos y cada uno de los registros. La segunda consistió en sustituir los strings correspondientes a los meses de enero (*ENE*), abril (*ABR*), agosto (*AGO*) y diciembre (*DIC*) por sus contrapartes equivalentes en inglés (respectivamente, *JAN*, *APR*, *AUG* y *DEC*), para facilitar la lectura de estas variables como fecha por las funciones predefinidas por python.

También creamos las variables *FEC_HOR_ING* y *FEC_HOR_ALT*, compuestas por la fecha de ingreso unida a la hora y el minuto almacenados en las variables correspondientes. Además, renombramos la variable *FEC_HOR_ING* para que sea la variable *FECHA_REFERENCIA* de este dataset.

Tras estas modificaciones, unimos este dataframe con el correspondiente a los datos demográficos, de manera a tener unificados en uno las variables de hospitalizaciones con las demográficas. Para esto, análogamente al caso de los datos de atención primaria, aplicamos una operación de tipo *left join*, tomando como referencia izquierda el dataframe de hospitalizaciones y derecha el de datos demográficos; y como clave la variable *NUM_SIP*. No supuso complicación alguna.

Posteriormente, hicimos la misma operación para unir estos datos resultantes con los datos de cronicidad. Aplicamos una operación análoga a la anterior, sustituyendo los datos demográficos por los de cronicidad, no obstante, en este caso no fue suficiente para unir la totalidad del dataframe. Esto se debe a que no en todas las fechas en las que se realizó el ingreso hospitalario se tomó nota de la cronicidad del paciente. Por lo tanto, para aquellos registros del dataframe de hospitalizaciones cuya fecha exacta no esté almacenada en el registro correspondiente a su SIP del dataframe de cronicidad, tomamos el registro de cronicidad con fecha anterior más cercano, en caso de existir; y en caso contrario, marcamos ese valor como desconocido. El código correspondiente a esta unión de los valores no exactos se puede encontrar en la función *unir_cronicidad_CMBD_2*, que podemos encontrar en el archivo *help_functions.py*.

El resultado de estas transformaciones y uniones ha sido almacenado en un documento *csv*, mucho más fácilmente manejable, al cual hemos llamado *hospitalizaciones_para_grafs_con_cron_alternat.csv*.

Se puede encontrar una primera observación generalista de las variables junto con el resultado de las transformaciones realizadas y las uniones de los distintos conjuntos de datos en el notebook *0_b_hospitalizaciones_csv.ipynb*. Por otra parte, el código correspondiente a las transformaciones realizadas se encuentran en la definición de la función *transformar_tipos_CMBD_2*, en el archivo *help_functions.py*.

III.1.5 Datos Urgencias Hospitalarias

Los datos provenientes de urgencias hospitalarias nos fueron entregados en un documento de texto, llamado ***SD2628_URGENCIAS_HOSPITALARIAS.enc.txt***, de tamaño 81268KB. Contiene los datos con las variables solicitadas de aquellos pacientes que han recibido algún diagnóstico compatible con SCA (ya sea en el servicio de urgencias hospitalarias o en alguna de las otras tres vías de entrada al sistema sanitario tratadas en este trabajo) entre 2016 y 2019; y han utilizado este servicio, ya sea por un motivo relacionado con el SCA o no. Para abrir este archivo, hemos usado el método *read_excel* de la librería *pandas*. Podemos encontrar 263761 registros y 25 variables en este conjunto de datos. Encontraremos valores ausentes en hasta cuatro de ellas, los cuales explicaremos en el punto correspondiente a cada una de ellas (en caso de no mencionar valores ausentes, significa que esa variable no contiene ninguno, aunque puede contener valores indicados como nulos o vacíos). Veamos su composición en detalle:

- *NUM_SIP*: Esta variable se corresponde con el número SIP de los 52907 (número de valores distintos en esta variable) pacientes compatibles con la descripción hecha en el párrafo anterior. Se identifica con la variable de mismo nombre del resto de fuentes de datos. *NUM_SIP* es el identificador alfanumérico obtenido después de someter el número SIP de cada paciente al proceso de anonimización de los datos. Puesto que corresponde con un identificador alfanumérico, para trabajar con ella la transformamos en tipo *string*.

- *Depart_Atencion*: Esta variable corresponde al nombre del departamento de salud al cual pertenece el hospital en el cual se ha realizado la consulta en cuestión. Existen 26 valores distintos asignados a esta variable: los nombres de cada uno de los departamentos de salud, más los valores: *C.H. PROVINCIAL CASTELLON* y *H.SANT VICENT DEL RASPEIG*. Puesto que corresponde con una etiqueta, para trabajar con ella transformamos esta variable en tipo *string*.

- *Centro_Atencion*: Esta variable corresponde al nombre del hospital en el cual se ha realizado la consulta en cuestión. Existen 27 valores distintos asignados a esta variable, correspondientes a los nombres de cada uno de los hospitales que tienen servicio de urgencias. Puesto que corresponde con una etiqueta, para trabajar con ella transformamos esta variable en tipo *string*.

- *Momen_Registro*: Esta variable se corresponde con la fecha y la hora del registro del paciente en el sistema de urgencias. Podemos encontrar esta fecha en formato *%d\ %b\ %y %H:%M:%S* (es decir, *1\ENE\2024 00:00:00*). Puesto que

corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *Momen_Alta*: Esta variable se corresponde con la fecha y la hora a la que el paciente recibe el alta del sistema de urgencias. Podemos encontrar esta fecha en formato *%d\ %b\ %y %H: %M: %S*. Existen 11009 valores ausentes para esta variable, correspondientes a consultas en las cuales el momento de alta no ha sido registrado. Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *Momen_Ini_Triaje*: Esta variable se corresponde con la fecha y la hora del inicio del proceso de triaje Manchester tras la llegada del paciente al sistema de urgencias. Podemos encontrar esta fecha en formato *%d\ %b\ %y %H: %M: %S*. Existen 20704 valores ausentes para esta variable, correspondientes a consultas en las cuales el momento de inicio del proceso de triaje no ha sido registrado. Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *Momen_Fin_Triaje*: Esta variable se corresponde con la fecha y la hora de finalización del proceso de triaje Manchester tras la llegada del paciente al sistema de urgencias. Podemos encontrar esta fecha en formato *%d\ %b\ %y %H: %M: %S*. Existen 20462 valores ausentes para esta variable, correspondientes a consultas en las cuales el momento de fin del proceso de triaje no ha sido registrado. Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *Momen_Atencion*: Esta variable se corresponde con la fecha y la hora en la que se atiende al paciente en el sistema de urgencias. Podemos encontrar esta fecha en formato *%d\ %b\ %y %H: %M: %S*. Existen 4123 valores ausentes para esta variable, correspondientes a consultas en las cuales el momento de atención al paciente no ha sido registrado. Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *Momen_Alta_Administrativa*: Esta variable se corresponde con la fecha y la hora a la que el paciente recibe el alta del sistema de urgencias. Podemos encontrar esta fecha en formato *%d\ %b\ %y %H: %M: %S* (es decir, *1\ENE\2024 00:00:00*). Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *Cod_Prio_Manchester*: Esta variable es la codificación numérica de la clasificación otorgada al paciente en el triaje Manchester. Contiene 7 valores distintos, que son los números enteros del 1 al 5, cada uno de los cuales indica el nivel de prioridad correspondiente, además de los valores *-1* cuando el valor está vacío, y *0* cuando el triaje no aplica a este paciente. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Prio_Manchester*: Esta variable es la descripción de la clasificación otorgada al paciente en el triaje Manchester. Contiene 7 valores distintos, que corresponden a los niveles de severidad 1 a 5, más los valores *[Vacío]* y *[No aplica]*. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Cod_Origen_Ingreso*: Esta variable es la codificación numérica del origen del ingreso del paciente. Contiene 18 valores distintos asignados a esta variable, que son 16 números enteros entre *1* y *18* correspondientes a distintos orígenes posibles, más los números *-2* para los registros sin referencia y *-1* para los registros vacíos. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *Origen_Ingreso*: Esta variable es la descrpición del origen del ingreso del paciente. Contiene 18 valores distintos asignados a esta variable, que son 16 descripciones de distintos orígenes de los pacientes, más los valores *[Sin referencia]* y *[Vacío]*. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *Cod_Punto_Entrada*: Esta variable es la codificación numérica del punto de entrada del paciente a las urgencias hospitalarias. Contiene 6 valores distintos asignados a esta variable, que son 4 números enteros entre *1* y *4* correspondientes a distintos puntos de entrada, más los números *-2* para los registros sin referencia y *-1* para los registros vacíos. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *Punto_Entrada*: Esta variable es la descripción del punto de entrada del paciente a las urgencias hospitalarias. Contiene 6 valores distintos asignados a esta variable, que son 4 descripciones de los distintos puntos de entrada, más los valores *Sin referencia* y *Vacío*. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *Cod_Motivo_Urg*: Esta variable es la codificación numérica del motivo de la urgencia por la cual ha acudido el paciente a urgencias hospitalarias. Contiene 9 valores distintos asignados a esta variable, que son 8 números enteros entre *1* y *8* correspondientes a distintos motivos de urgencia, más el número *-2* para los registros sin referencia. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *Motivo_Urg*: Esta variable es la descripción del motivo de la urgencia por la cual ha acudido el paciente a urgencias hospitalarias. Contiene 9 valores distintos asignados a esta variable, que son 8 descripciones de distintos motivos de urgencias, más el valor *Sin referencia*. Al tratarse de una etiqueta, transformaremos esta variable en un *string*.

- *CIE_Cod_Diag*: Esta variable contiene la codificación del diagnóstico principal asignado al paciente durante el ingreso mediante los códigos CIE-9 y CIE-10. Existen 6292 valores distintos asignados a esta variable, a los que hay que sumar dos valores más, el *-1* para indicar registros vacíos y *-2* para indicar registros sin referencia. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *CIE_Cod_Diag_2*: Esta variable contiene la codificación del diagnóstico secundario asignado al paciente durante el ingreso mediante los códigos CIE-9 y CIE-10. Existen 3252 valores distintos asignados a esta variable, a los que hay que sumar dos valores más, el *-1* para indicar registros vacíos y *-2* para indicar registros sin referencia. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *Cod_Proce*: Esta variable contiene la codificación del procedimiento principal aplicado al paciente durante el ingreso mediante los códigos CIE-9, CIE-10 y NIC. Existen 323 valores distintos asignados a esta variable, a los que hay que sumar dos valores más, el *-1* para indicar registros vacíos y *-2* para indicar registros sin referencia. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *Descrip_Proce*: Esta variable contiene la descripción del procedimiento principal aplicado al paciente durante el ingreso mediante los códigos CIE-9, CIE-10 y NIC. Existen 335 valores distintos asignados a esta variable, a los que hay que sumar dos valores más, el *[Vacío]* y *[Sin referencia]*. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *Cod_Proce_2*: Esta variable contiene la codificación del procedimiento secundario aplicado al paciente durante el ingreso mediante los códigos CIE-9, CIE-10 y NIC. Existen 183 valores distintos asignados a esta variable, a los que hay que sumar dos valores más, el *-1* para indicar registros vacíos y *-2* para indicar registros sin referencia. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *Descrip_Proce_2*: Esta variable contiene la descripción del procedimiento secundario aplicado al paciente durante el ingreso mediante los códigos CIE-9, CIE-10 y NIC. Existen 189 valores distintos asignados a esta variable, a los que hay que sumar dos valores más, el *[Vacío]* y *[Sin referencia]*. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *Cod_Circunst_Alta*: Esta variable es la codificación numérica de la circunstancia en la cual se le ha dado el alta del servicio de urgencias hospitalarias al paciente. Esta codificación es arbitraria, y es la misma en los datos correspondientes a las demás etapas del sistema sanitario. En este dataset, tenemos 18 valores distintos en esta columna. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

- *Circunst_Alta*: Esta variable es la descripción de la circunstancia en la cual se le ha dado el alta del servicio de urgencias hospitalarias al paciente. Al igual que en casos anteriores, esta columna va emparejada con su código equivalente. En este dataset, tenemos 18 valores distintos para esta columna. Esta variable se trata de una etiqueta, y por lo tanto la transformaremos en un *string*.

Más allá de la intervención en los tipos de los datos, ya descrita en el apartado correspondiente a cada variable, en esta primera lectura únicamente modificamos el nombre de la variable *MOMENTO_REGISTRO* a *FECHA_REFERENCIA*, tal y como hicimos en los conjuntos de datos previos.

También es reseñable destacar que, al igual que ocurría con los datos de atención primaria, existe una diferencia muy similar entre el número de valores únicos en las variables *Cod_Proce* (325) y *Descrip_Proce* (337). El mismo caso ocurre entre las variables *Cod_Proce_2* (183) y *Descrip_Proce_2* (189). Esto puede deberse a que un procedimiento igual ha sido escrito con caracteres o palabras distintos (como el caso del código *39.98*, que tiene como descripciones *CONTROL DE HEMORRAGIA, N. E. O. M.* y *CONTROL DE HEMORRAGIA, NEOM*. En este caso, no realizaremos tampoco

ninguna transformación, puesto que, mientras exista la codificación, es irrelevante que las descripciones sean ligeramente diferentes. Existen diccionarios de los códigos CIE-9, CIE-10 y NIC donde se pueden consultar las descripciones oficiales correspondientes a cada caso.

Tras estas modificaciones, unimos este dataframe con el correspondiente a los datos demográficos, de manera a tener unificados en uno las variables de urgencias hositalarias con las demográficas. Para esto, análogamente al caso de los conjuntos de datos anteriores, aplicamos una operación de tipo *left join*, tomando como referencia izquierda el dataframe de urgencias hospitalarias y derecha el de datos demográficos; y como clave la variable *NUM_SIP*. No supuso complicación alguna.

Posteriormente, como en los casos anteriores, intentamos unir el dataframe resultante de la operación anterior con el de cronicidad. Sin embargo, al igual que en el caso de los datos de hospitalizaciones, realizar una operación del tipo *left join* fue insuficiente para unir la totalidad del dataframe, debido de nuevo a que no en todas las fechas en las que se realizó la consulta en urgencias hospitalarias se tomó nota de la cronicidad del paciente. Por lo tanto, para aquellos registros cuya fecha exacta no esté almacenada en el registro correspondiente a su SIP del dataframe de cronicidad, aplicaremos una operación análoga a la utilizada en el caso del dataframe de hospitalizaciones: tomamos el registro de cronicidad con fecha anterior más cercano, en caso de existir; y en caso contrario, marcamos ese valor como desconocido. El código correspondiente a esta unión de los valores no exactos se puede encontrar en la función *unir_cronicidad_URG_HOSP_2*, que podemos encontrar en el archivo *help_functions.py*.

El resultado de estas transformaciones y uniones ha sido almacenado en un documento *csv*, cuyo nombre es *urg_hospitalarias_para_grafs_con_cron_alternat.csv*.

Se puede encontrar una primera observación generalista de las variables junto con el resultado de las transformaciones realizadas y las uniones de los distintos conjuntos de datos en el notebook *0_c_urg_hospitalarias_csv.ipynb*. Por otra parte, el código correspondiente a las transformaciones realizadas se encuentran en la definición de la función *transformar_tipos_UHOSP_2*, en el archivo *help_functions.py*.

III.1.6 Datos Emergencias

Los datos provenientes del servicio de emergencias nos fueron entregados en un documento de texto, llamado ***Datos_emergencias.enc.txt***. Esta variable debería contener los datos con las variables solicitadas de aquellos pacientes que han recibido algún diagnóstico compatible con SCA (ya sea en el servicio de emergencias o en alguna de las otras tres vías de entrada al sistema sanitario tratadas en este trabajo) entre 2016 y 2019; y han utilizado este servicio, ya sea por un motivo relacionado con el SCA o no. Para abrir este archivo, hemos usado el método *read_excel* de la librería *pandas*. No obstante, por un error durante la extracción por parte del servicio encargado de proveerlos, este conjunto tan solo contiene aquellos datos de los pacientes que han recibido un diagnóstico compatible con SCA en el propio servicio de emergencias. Podemos encontrar 19921 registros y 12 variables en este conjunto de datos. Encontraremos valores ausentes en hasta cuatro de ellas, los cuales explicaremos en el punto correspondiente a cada una de ellas (en caso de no mencionar valores ausentes, significa que esa variable no contiene ninguno, aunque puede contener valores indicados como nulos o vacíos). Veamos su composición en detalle:

- *SIP*: Esta variable se corresponde con el número SIP de los 16849 (número de valores distintos en esta variable) pacientes compatibles con la descripción hecha en el párrafo anterior. Se identifica con la variable de mismo nombre del resto de fuentes de datos. *SIP* es el identificador alfanumérico obtenido después de someter el número SIP de cada paciente al proceso de anonimización de los datos. Puesto que corresponde con un identificador alfanumérico, para trabajar con ella la transformamos en tipo *string*.

- *FECHA*: Esta variable se corresponde con la fecha y la hora a la que se inicia el proceso de solicitud de atención del servicio de urgencias. Podemos encontrar esta fecha en formato *%d\ %m\ %y %H:%M* (es decir, 01\01\2024 0:00). Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y-%m-%d %H:%M:%S*.

- *MUNICIPIO*: Esta variable corresponde al nombre del municipio en el cual se solicitó la actuación del servicio de emergencias. Existen 397 valores distintos asignados a esta variable: los nombres de cada uno de los municipios, más el valor *Sin Determinar*. Puesto que corresponde con una etiqueta, para trabajar con ella transformamos esta variable en tipo *string*.

- *DEPARTAMENTO*: Esta variable corresponde al nombre del departamento de salud en el cual se ha provisto el servicio de emergencias. Existen 25 valores distintos asignados a esta variable: los nombres de cada uno de los departamentos de salud, más el valor *Sin Determinar*. Puesto que corresponde con una etiqueta, para trabajar con ella transformamos esta variable en tipo *string*.

- *TIPIFICACIÓN (4 nivel)*: Esta variable corresponde a la tipificación otorgada al episodio por el servicio de emergencias, realizada en cuatro niveles. Existen 197 valores distintos asignados a esta variable. Puesto que corresponde con una etiqueta, para trabajar con ella transformamos esta variable en tipo *string*.

- *HORA ENTRADA LLAMADA A CICU*: Esta variable se corresponde con la fecha y la hora a la que se la llamada ha sido atendida por la CICU. Podemos encontrar esta fecha en formato *%d\ %m\ %y %H:%M*. Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y-%m-%d %H:%M:%S*.

- *HORA ASISTENCIA*: Esta variable se corresponde con la fecha y la hora a la que se ha realizado la asistencia al paciente, en caso de haber sido asistido. Podemos encontrar esta fecha en formato *%d\ %m\ %y %H:%M*. Existen 2427 valores ausentes en esta variable, correspondiente a casos en los cuales no se ha realizado ninguna asistencia. Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y-%m-%d %H:%M:%S*.

- *HORA TRANSFERENCIA*: Esta variable se corresponde con la fecha y la hora a la que se ha realizado la transferencia del paciente a otro servicio, en caso de haber sido transferido. Podemos encontrar esta fecha en formato *%d\ %m\ %y %H:%M*. Existen 2861 valores ausentes en esta variable, correspondiente a casos en los cuales no se ha realizado ninguna asistencia. Puesto que corresponde con una

fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *HORA FINALIZACIÓN*: Esta variable se corresponde con la fecha y la hora a la que ha finalizado la actuación del servicio de emergencias. Podemos encontrar esta fecha en formato *%d\ %m\ %y %H: %M*. Existen 4056 valores ausentes en esta variable, correspondiente a casos en los cuales no se ha anotado esta hora de finalización. Puesto que corresponde con una fecha y una hora, transformaremos esta variable en tipo *datetime64[ns]*, con fecha en formato *%Y- %m- %d %H: %M: %S*.

- *DIAGNOSTICO*: Esta variable contiene la codificación del diagnóstico asignado al paciente durante la atención por el servicio de emergencias mediante los códigos CIE-9 junto con la descripción del mismo. Al tener únicamente los registros de los pacientes con diagnósticos compatibles con SCA, existen únicamente 38 valores distintos en esta variable. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *MANIOBRAS*: Esta variable contiene la descripción de las maniobras realizadas al paciente en caso de haberle sido realizada alguna durante la actuación del servicio de emergencias. Existen un total de 5156 valores distintos y 3795 registros con valores ausentes en esta variable. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

- *FINALIZACION*: Esta variable contiene la descripción de la circunstancia en las cual se da por finalizada la atención del servicio de emergencias. Existen un total de 20 valores distintos en esta variable. Puesto que se trata de una etiqueta, la transformaremos en un *string*.

Más allá de la intervención en los tipos de los datos, ya descrita en el apartado correspondiente a cada variable, en esta primera lectura modificamos el nombre de la variable *FECHA* para transformarla en *FECHA_REFERENCIA*, tal y como hicimos en los conjuntos de datos previos, y el nombre de la variable *SIP* para transformarla en *NUM_SIP*, por coherencia con el resto de fuentes de datos.

Tras estas modificaciones, unimos este dataframe con el correspondiente a los datos demográficos, mediante el mismo procedimiento que en los tres dataframes anteriores, y con el dataframe de cronicidad, siguiendo un proceso análogo al empleado con los casos de hospitalizaciones y urgencias hospitalarias. El código de la función que une los la cronicidad a los registros con fechas no exactas se puede encontrar en la función *unir_cronicidad_Emg_2*, que podemos encontrar en el archivo *help_functions.py*. El resultado de estas transformaciones y uniones ha sido almacenado en un documento *csv*, llamado *emergencias_para_grafs_con_cron.csv*.

Se puede encontrar una primera observación generalista de las variables junto con el resultado de las transformaciones realizadas y las uniones de los distintos conjuntos de datos en el notebook *0_d_emergencias_csv.ipynb*. Por otra parte, el código correspondiente a las transformaciones realizadas se encuentran en la definición de la función *transformar_tipos_emergencias*, en el archivo *help_functions.py*.

III.2 Construcción del conjunto de episodios

Una vez realizada la primera observación y limpieza de cada uno de los conjuntos de datos, pasamos a la etapa de construcción del corpus de información. El principal

objetivo de este apartado es la creación de un único conjunto de datos que reúna la información recopilada a través de las cuatro fuentes de datos proporcionadas, agrupada por episodios. Con este objetivo, desarrollamos los notebooks *1_Episodios_por_sip.ipynb* y *2_Episodios_por_fecha_en_cada_sip.ipynb* Veamos la composición de cada uno de los notebooks:

III.2.1 Agrupar los episodios por SIPs

En esta primera etapa, el objetivo es crear un diccionario de python cuyas claves sean cada uno de los SIPs de los pacientes. Dentro de cada clave, podremos encontrar otro diccionario que contenga como clave un identificador de cada uno de los servicios: *UAP* para el servicio de Atención Primaria, *UHOSP* para el servicio de urgencias hospitalarias, *CMBD* para el servicio de hospitalizaciones y *Emergencias* para el servicio de emergencias. Utilizaremos estos mismos identificadores para el resto del trabajo. Dentro de cada una de estas, encontraremos un filtrado del dataframe de cada uno de los servicios (unidos todos con la información demográfica y de cronicidad), para contener únicamente la información relativa al respectivo SIP.

Para construir este diccionario, hemos seguido un proceso iterativo. Describiremos este proceso para el caso de los datos de atención primaria, y será análogo para los otros tres servicios.

En primer lugar, leemos los datos relativos a atención primaria almacenados en el punto anterior en el archivo *at_primaria_para_grafs_con_cron.csv*.

Tras esto, transformaremos cada variable en su respectivo tipo, de acuerdo con lo descrito en el apartado anterior, gracias a la función *transformar_tipos_UAP_3*, que se puede encontrar en el archivo *help_functions.py*. La única particularidad que tuvimos que añadir a esta función, fue en la variiable *FECHA_NACIMIENTO*. Para poder transformar esta variable en tipo *datetime64[ns]*, tuvimos que sustituir los valores *Desconocido* por valores ausentes (en python, *None*). Almacenamos este dataframe en la variable *data_1*.

Después, creamos la lista *unique_sips_UAP*, que contiene la lista de SIPs únicos en el dataframe recién leído, y el diccionario vacío llamado *unique_sips_with_dataframes*.

Por último, iteramos sobre la lista *unique_sips_UAP*. En cada iteración *i*, realizamos una copia del dataframe *data_1* filtrado por aquellos registros que contengan el *i* en la variable *NUM_SIP*. Posteriormente, comprobamos si el diccionario *unique_sips_with_dataframes* contiene la clave *i*. En caso de que no la contenga, asignamos un diccionario vacío como valor para esta clave. Por último, dentro de este diccionario, creamos la clave *UAP*, y almacenamos dentro la copia del dataframe filtrado recién creada.

Realizamos este proceso de manera análoga para las otras tres tres fuentes de datos, almacenando en el diccionario contenido dentro de cada clave SIP los datos de hospitalizaciones con la clave *CMBD*, los datos de urgencias hospitalarias con la clave *UHOSP* y los datos de emergencias con la clave *Emergencias*.

Por último, almacenamos el diccionario resultante en el archivo de tipo *pkl* llamado *unique_sips_with_dataframes_4.pkl*, el cual reutilizaremos en la siguiente etapa.

III.2.2 Agrupar los episodios por fecha

En esta segunda etapa trabajaremos con el mismo diccionario que acabamos de crear y de almacenar en el documento *pkl*. El objetivo en esta ocasión es, dentro de

cada clave relativa a cada uno de los diferentes SIPs, agrupar los datos por cada episodio de infarto. Consideraremos que dos registros pertenecen al mismo episodio de infarto si la diferencia en las fechas almacenadas en la variable *FECHA_REFERENCIA* de cada registro es de menos de un día.

Antes de cualquier acción, creamos el diccionario *unique_sips_separated_by_episodes*, que contendrá los resultados finales de todo el procedimiento que describiremos a continuación. Este diccionario tendrá una clave por cada identificador SIP, de manera similar al objeto creado en el anterior notebook, el cual también leemos a partir del documento *unique_sips_with_dataframes_4.pkl* y almacenamos nuevamente en la variable *unique_sips_with_dataframes*).

Seguimos un proceso en el cual iteraremos a través de los identificadores *SIP* que encontramos en las claves del objeto *unique_sips_with_dataframes_4.pkl*. Sean *sip* una clave cualquiera, y *dict_sip* el diccionario almacenado dentro de ella. A continuación, describiremos el proceso seguido para agrupar sus datos asociados por episodios.

La primera acción a realizar, es modificar el nombre de las columnas de todos los dataframes, para identificarlas con cada uno de los servicio. Esto se debe a que existen nombres idénticos en variables pertenecientes a diferentes fuentes de datos (como por ejemplo, *FECHA_REFERENCIA*). Realizaremos este cambio en todas las columnas salvo *NUM_SIP*, que, puesto que es el identificador común en todas las fuentes de datos, es interesante mantener el mismo nombre para realizar operaciones entre las tablas.

Con este cometido, concatenaremos los nombres de la columna y de cada uno de los servicios, separándolos con el caracter _. Por ejemplo, imaginemos que tratamos los datos relativos a atención primaria almacenados en *dict_sip*. En este caso, añadiremos a todos los identificadores de las columnas diferentes de *NUM_SIP* el string *_UAP* (siguiendo con el caso anteriormente mencionado, la variable *FECHA_REFERENCIA* pasará a llamarse *FECHA_REFERENCIA_UAP*).

En segundo lugar, creamos la lista *unique_dates*, la cual almacena los valores únicos de fechas almacenados en las variables *FECHA_REFERENCIA* pertenecientes a cada una de las fuentes de datos presentes en *dict_sip*.

En tercer lugar, creamos una lista de listas, llamada *unique_ranges_dates_def*, que contiene las fechas que habíamos almacenado en la lista *unique_dates*, agrupadas por un día de diferencia. Para ello seguimos el siguiente proceso: Para empezar, creamos las listas vacías *counted_dates* y *unique_ranges_dates*. Iteramos sobre la lista *unique_dates*, y llamamos a cada iteración *date_1*. Si esta fecha no se encuentra dentro de la lista *counted_dates*, creamos la lista vacía *fechas* y le añadimos la fecha *date_1*. Después, volvemos a iterar sobre la lista *unique_dates*, llamando *date_2* a cada iteración. Comprobamos si la diferencia entre *date_1* y *date_2* es inferior a un día, y en caso afirmativo, adherimos *date_2* a la lista *fechas*. Posteriormente, añadimos el objeto lista *fechas* a la lista *unique_ranges_dates*, que será una lista de listas. Por último, creamos la lista *unique_ranges_dates_def*, que contendrá las mismas listas que *unique_ranges_dates*, pero ordenadas cronológicamente en su interior.

Por último, creamos la clave *sip* en el diccionario *unique_sips_separated_by_episodes*, que contiene un diccionario vacío. Ahora, iteramos dentro de las claves del diccionario almacenado en *unique_sips_with_dataframes*, llamando *key* a cada iteración. Cada una de éstas corresponderá a uno de los servicios de salud que contenga datos asociados a *sip*.

Iteramos después dentro de la variable *unique_ranges_dates_def*, para obtener ca-

da una de las listas de fechas separadas por un día. Llamemos a cada una de estas listas *dates* y al número de la iteración (numeración python, es decir, *0* para la primera iteración, 1 para la segunda...). Iterando dentro de las fechas almacenadas en la lista *dates* (llamemos a la fecha iterada *rg*), comprobamos si el servicio de salud *key* contiene algún registro cuya fecha referencia esté separada por menos de un día de *rg*. En caso afirmativo, creamos la clave *k* (en caso de no existir) en el diccionario contenido en la clave *sip* de *unique_sips_separated_by_episodes*. La clave *k* será el identificador del episodio en cuestión. Como última instancia, introducimos en este nuevo diccionario una clave con el identificador del servicio, unido con el número de veces que ha aparecido este servicio en este episodio, ya que un paciente puede haber pasado dos veces por el mismo servicio en un mismo episodio de infarto (como podría ser en caso de realizar dos traslados en ambulancia, primero a un hospital y luego a otro más especializado, los cuales se etiquetarían como *Emergencias_1* y *Emergencias_2*). En esta clave, almacenaremos el dataframe relativo al servicio *key*, filtrado por aquellos registros cuya fecha de referencia sea *rg*.

En algunos casos, puede que este dataframe filtrado que acabamos de mencionar contenga más de un registro, a pesar de referirse a una única visita. Esto se puede deber a que un paciente reciba más de un diagnóstico o se vea sometido a más de un procedimiento, en una fuente de datos que no recoge los diagnósticos o procedimientos secundarios.

En la última celda de este notebook, almacenamos el diccionario resultante en el archivo *unique_sips_separated_by_episodes_4.pkl*.

III.2.3 Observaciones del objeto final

En este archivo *.pkl*, hemos logrado aunar de manera ordenada todos los episodios proporcionados en nuestros datos en un único objeto. Este archivo es fácilmente extraíbe en un diccionario de python (como era el objeto del que lo hemos guardado). No obstante, existen una serie de observaciones que tenemos que realizar al respecto del diccionario resultante, a tener en cuenta en caso de querer utilizar este conjunto de datos para poder realizar una investigación futura.

Como se indicó en el apartado dedicado a la solicitud de los datos, no solamente nos fueron proporcionados los episodios de infarto pertenecientes a cada uno de los servicios de salud. En los datos que tenemos a nuestra disposición, podemos encontrar todos los registros existentes de consultas médicas realizadas entre 2016 y 2019, a pacientes que, en algún momento durante este mismo periodo, hayan recibido un diagnóstico compatible con SCA. Por ilustrarlo con un ejemplo, si un paciente recurrió a una consulta en cualquiera de los cuatro servicios consultados y recibió un diagnóstico compatible con SCA el 27 de julio de 2017, pero también acudió al centro de salud por un resfriado el 15 de noviembre de 2019, no solamente tendremos los registros relacionados con el primer incidente, sino también los registros relacionados con el segundo. Esta solicitud de información extra se realizó para tener registro de la totalidad del recorrido asistencial del paciente, puesto que, si únicamente solicitábamos los datos relacionados con diagnósticos compatibles con SCA, podríamos estar perdiendo parte de ellos. Pongamos el caso de un paciente que acude a su centro de salud con un dolor en el pecho, y es derivado al hospital con un diagnóstico compatible con SCA. Cuando llega al hospital, su dolencia es diagnosticada como acidez estomacal. Si únicamente solicitamos los datos relativos a diagnósticos compatibles con SCA, perderíamos esta parte del recorrido, que puede ser muy interesante para diversas investigaciones. No

obstante, esto supone una gran cantidad de episodios que no contienen ningún diagnóstico de infarto. Esto deberá ser tenido en cuenta a la hora de tratar con estos datos, para eliminar este exceso de información en caso de ser irrelevante.

A pesar de lo dicho en el punto anterior, podemos observar algunos pocos *SIPs* para los cuales no existe ningún episodio que contenga ningún diagnóstico compatible con SCA. Aunque esto pueda dar que pensar al principio que puede deberse a un error en el proceso de construcción del corpus de información, si miramos un poco más en profundidad, existe un motivo por el cual estos casos existen. En el protocolo de investigación clínica podemos encontrar la siguiente frase, relativa a los diagnósticos solicitados:

> *La codificación de entidades clínicas compatibles con SCA utilizando CIE-9 permite la diferenciación de los anteriores grupos asignando los siguientes códigos a cada uno de ellos:*
>
> - *SCA: 410, 411, 413 y todos los que incluyen a nivel de cuarto y quinto dígito.*
> - *IAM: 410, 411 y todos los que incluyen a nivel de cuarto y quinto dígito.*
>
> *La codificación de entidades clínicas compatibles con SCA utilizando CIE-10 permite una diferenciación de las entidades clínicas utilizando los siguientes códigos:*
>
> - *SCA: I20.0, I20.1, I20.8, I20.9, I21.01, I21.02, I21.09, I21.11, I21.19, I21.21, I21.29, I21.3, I22.0, I22.1, I22.8, I22.9, I21.4, I22.2, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.7, I23.8, I24.0, I25.110, I25.111, I25.119, I25.700, I25.701, I25.710, I25.711, I25.720, I25.721, I25.730, I25.731, I25.750, I25.751, I25.760, I25.761, I25.790, I25.791, I24.8, I24.9, I25.118, I25.5, I25.6, I25.708, I25.709, I25.718, I25.719, I25.728, I25.729, I25.738, I25.739, I25.758, I25.759, I25.768, I25.769, I25.798, I25.799.*
> - *IAM: I21.01, I21.02, I21.09, I21.11, I21.19, I21.21, I21.29, I21.3, I21.4, I22.0, I22.1, I22.2, I22.8, I22.9, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.7, I23.8.*
>
> *El cambio de codificación también ha afectado al procedimiento de la Angioplastia Coronaria Transluminal Percutánea, pasando de un único código en CIE-9, el 00.66 a 372 posibles códigos en CIE-10, los comenzados por 027 seguidos de hasta 4 dígitos más, con 4 posibles valores en el primero, 3 en el segundo, 10 en el tercero y dos en el cuarto dígito. No se listan los 372 códigos en el presente documento por razones de espacio.*

Si prestamos atención a los códigos de los episodios que no contienen ningún diagnóstico compatible con SCA, podremos observar que, en su totalidad, siempre contienen en alguno de los diagnósticos, una parte de uno de los de la lista. Es el caso del SIP **TvZ32M33HYD6dZ0bJakOyA==**, que contiene hasta 7 episodios distintos. Ninguno de ellos contiene un diagnóstico de los que podemos encontrar en la lista solicitada. No obstante, en el episodio de clave *4*, podemos encontrar el diagnóstico: *C50.411* -

Neoplasia maligna de cuadrante superior, que contiene un 411 entre su codificación, motivo por el cual podemos encontrar a este paciente entre los solicitados.

Por último, también debemos resaltar el amplio tiempo de ejecución gastado en trabajar sobre este objeto. Este archivo *unique_sips_separated_by_episodes_4.pkl* pesa alrededor de dos Gb. Es por ello que los recursos de computación requeridos para unir estos datos son muy potentes, y el tiempo de computación ha sido muy largo, lo cual debe ser tenido en cuenta a la hora de ejecutar el código.

IV Perfil socio-demográfico del paciente de SCA

En este apartado, expondremos el resultado de la investigación llevada a cabo con el conjunto de datos que hemos extraído en la fase anterior. Recordemos las preguntas que planteábamos al inicio del trabajo: ¿Cuáles son los perfiles demográficos más comunes de los pacientes afectados por el SCA? ¿Cuál es el recorrido asistencial por el que son atendidos? ¿Es éste correcto en una gran mayoría de los casos? ¿Cuál es la vía de entrada del paciente al sistema sanitario? ¿Es la mortalidad muy elevada dentro de los pacientes atendidos? ¿Cómo varían las respuestas a estas preguntas dependiendo del lugar de procedencia, la edad, la vía de entrada o el sexo del paciente?

Para responder a estas preguntas, llevaremos a cabo en primer lugar una serie de transformaciones para poder obtener un conjunto de datos mucho más manipulable, después hablaremos del análisis realizado sobre estos, y finalmente expondremos las conclusiones extraídas.

IV.1 Extracción de los datos

Como se ha visto en el apartado anterior correspondiente a la construcción del corpus de la información, los datos relativos al recorrido asistencial completo de un paciente durante un episodio de atención están distribuidos en distintas fuentes y es difícilmente manejable. Si bien ahora disponemos de un archivo que contiene todos los episodios almacenados y ordenados, ahora necesitamos transformarlo en otro más sencillo, libre de los episodios no relativos a SCA anteriormente mencionados, y que contenga únicamente la información que se desea utilizar para este estudio. Con este objetivo, creamos el notebook *3_Extraer_Datos_Para_Investigacion.ipynb*, donde trabajaremos nuevamente con el diccionario que acabamos de crear y de almacenar en el archivo *unique_sips_separated_by_episodes_4.pkl*, que leeremos en la variable *unique_sips_separated_by_episodes*.

IV.1.1 Recorrido asistencial

La primera operación la realizaremos dentro de cada clave relativa a cada uno de los diferentes SIPs, es decir, dentro de cada episodio. Extraeremos el recorrido asistencial del paciente, es decir, extraeremos una lista donde figuren de manera ordenada cuáles son los servicios en los que ha sido tratado el paciente durante el episodio. Con este objetivo, iteramos sobre el diccionario *unique_sips_separated_by_episodes* para acceder a cada uno de los sips almacenados en las claves de este diccionario (sea *sip* cada iteración), y lo hacemos nuevamente para acceder a cada uno de los episodios de cada sip (llamamos *episodio* a cada iteración).

Para cada *episodio*, creamos el diccionario vacío *fuentes*. Después, iteramos sobre el diccionario almacenado en la clave *episodio*, para extraer cada una de las claves, que pertenecen a los servicios de salud que ha ido visitando el paciente durante el episodio (sea j una de estas iteraciones, y sea *df* el dataframe almacenado en ella). Puesto que las claves j son la concatenación del identificador de cada servicio y el número de veces que el paciente ha visitado este servicio en este episodio, extraemos solamente el identificador del servicio y lo llamamos k. En el diccionario *fuentes*, almacenamos en la clave k el valor de la primera fila de *df* (es la misma fecha en todas las filas). Por último, aplicamos la función *ordered_keys_fechas* al diccionario *fuentes*, que podemos

encontrar en el archivo *help_functions.py*. Esta función devuelve una lista de los servicios en los que ha sido tratado el paciente durante el episodio, la cual guardaremos en el propio diccionario almacenado en *episodio*, con la clave *Recorrido*.

IV.1.2 Recorrido diagnóstico

En segundo lugar, ejecutaremos una operación parecida a la anterior. Igual que antes, iteraremos sobre *unique_sips_separated_by_episodes* para acceder a cada *episodio* dentro de cada *sip*. Extraeremos el recorrido diagnóstico del paciente, es decir, una lista donde figuren de manera ordenada cuáles son los diagnósticos que ha recibido el paciente durante el episodio.

En cada *episodio*, creamos el diccionario vacío *diagnosticos*. Después, accedemos análogamente al punto anterior a cada clave del diccionario almacenado en *episodio*, llamando *j* a cada una de ellas y *df* al dataframe que contiene. De nuevo, extraemos únicamente el identificador del servicio contenido en *j*, llamado *k*. En el diccionario *diagnosticos*, creamos la clave *k*, que contiene a su vez un diccionario con las siguientes claves: *Fecha*, que contiene el valor de *FECHA_REFERENCIA* encontrada en la primera fila de *j*, y *Diagnostico*, que contiene el valor de la primera fila de la columna correspondiente al diagnóstico principal del respectivo servicio. Por último, con la función *ordered_keys_diagnostico*, que se puede encontrar en el archivo *help_functions.py*, construimos una lista que contenga los diagnósticos almacenados en las claves *Diagnostico*, ordenados por la clave *Fecha*, que corresponderán al recorrido diagnóstico. Almacenaremos el resultado en el propio diccionario *episodio*, con la clave *Recorrido diagnostico*.

IV.1.3 Lugar y Edad

La siguiente operación a realizar es extraer las siguientes variables:

- El lugar donde se ha iniciado cada episodio, en particular, el departamento de salud en el que se ha producido la entrada en el sistema sanitario.
- La edad del paciente al inicio el episodio.

Para ello, realizaremos una iteración similar nuevamente sobre el diccionario *unique_sips_separated_by_episodes* tal y como en el apartado anterior. Sea *sip* cada iteración sobre el las claves del diccionario, y *episodio* cada iteración sobre las claves contenidas en *sip* correspondientes a cada episodio. Comprobamos previamente que la clave *episodio* es un número, puesto que, como veremos en el futuro, podría tomar valores como *Sexo*.

Aquí, comprobamos cuál es el servicio por el cual arranca el recorrido asistencial del *episodio*. Para esto, comprobamos cuál es el primer valor de la lista guardada en la clave *Recorrido* correspondiente al respectivo *episodio*. Llamemos a este servicio *servicio*.

Buscamos entonces el dataframe asociado a la primera consulta en el servicio *servicio* (por ejemplo, *UHOSP_1*), y extraemos:

- La descripción del departamento de salud almacenado en la primera fila de la columna que se refiera a este valor en el respectivo *servicio*. Este resultado lo almacenamos en la clave *Lugar*, dentro del diccionario almacenado en *episodio*. Este proceso está descrito en la función *determinar_lugar*, incluida en el fichero *help_functions.py*.

- El valor de la *EDAD* almacenada en la primera fila. Este resultado lo almacenamos en la clave *Edad*, dentro del diccionario almacenado en *episodio*. Este proceso está descrito en la función *determinar_edad*, incluida en el fichero *help_functions.py*. Como apunte, los registros cuyo valor en la variable *EDAD* sea *Desconocido* los sustituimos por *-1*, para mantener el tipo entero de la variable.

Debemos detenernos un momento en la variable *Lugar*, puesto que, para este caso, hemos tomado la descripción del departamento de salud correspondiente (es decir, el nombre). No obstante, en las cuatro fuentes de datos correspondientes existen diferentes maneras de referirse a los departamentos, además de existir algunas denominaciones como *Hospital Doctor Moliner*, que corresponden erróneamente al centro donde se atendió al paciente en lugar de al respectivo departamento de salud. Es por eso, que, tras extraer la variable *Lugar*, la modificaremos para darle un nombre unificado a todos los casos. Para esto, buscaremos una cadena de texto que se encuentre en todos los valores respectivos a un mismo lugar, y le otorgamos a todos un mismo valor (por ejemplo, todos los registros que contengan la cadena *ARNAU* tomarán el valor *DEPARTAMENT DE SALUT DE VALENCIA ARNAU DE VILANOVA LLIRIA*). Además, también identificamos aquellos casos en los que, como dijimos anteriormente, el nombre almacenado es el del hospital, y no el del departamento, y le damos el nombre del departamento correspondiente (tomando de nuevo el caso que hemos descrito antes, el valor *Hospital Doctor Moliner* será sustituido por su departamento de salud correspondiente, es decir, *DEPARTAMENT DE SALUT DE VALENCIA ARNAU DE VILANOVA LLIRIA*).

IV.1.4 Sexo

Extraeremos ahora la variable relativa al sexo del paciente. Puesto que esta variable no cambia entre distintos episodios, no nos hará falta en este caso iterar por cada uno de ellos, tan sólo accederemos al primero para tomar esta información. No obstante, sí que lo haremos sobre el diccionario *unique_sips_separated_by_episodes* para acceder a cada paciente. Sea *sip* cada iteración sobre el las claves del diccionario.

Aquí, comprobamos igual que en el anterior proceso cuál es el servicio por el cual arranca el recorrido asistencial del primer episodio, tomando el primer valor de la lista guardada en la clave *Recorrido*. Llamemos a este servicio *servicio*.

Buscamos entonces el dataframe asociado a la primera consulta en el servicio *servicio*, y extraemos la descripción del sexo almacenado en la primera fila de la columna que se refiera a este valor en el respectivo *servicio*. Este resultado lo almacenamos en la clave *Sexo*, dentro del diccionario almacenado en *sip*. Este proceso está descrito en la función *determinar_sexo*, incluida en el fichero *help_functions.py*.

IV.1.5 Alta

En la siguiente etapa extraemos la variable relativa al alta recibida por el paciente al final de cada episodio. Realizaremos una iteración parecida a la hecha en los casos de las variables *Lugar*, *Cronicidad* y *Edad*, es decir, iteraremos sobre el diccionario *unique_sips_separated_by_episodes*, y después sobre cada una de las iteraciones para acceder a los episodios. Sea *sip* cada iteración sobre el las claves del diccionario, y *episodio* cada una de las iteraciones en su interior, previa comprobación de que es un número

entero, que marque efectivamente es relativo a un episodio, y no está almacenando una de las variables (como *Lugar*, *Cronicidad* o *Edad*).

Contrariamente a los casos anteriores, puesto que para extraer esta variable buscamos el final del recorrido asistencia, comprobamos cuál es el servicio en el cual finaliza éste para el respectivo *episodio*. Para esto, comprobamos cuál es el último valor de la lista guardada en la clave *Recorrido* correspondiente al respectivo *episodio*. Llamemos a este servicio *servicio*.

Buscamos entonces el dataframe asociado a la última consulta en el servicio *servicio*. En caso de que *servicio* sea *CMBD*, *UHOSP* o *UAP*, tenemos que existen variables respectivas a la codificación y la descripción de la circunstancia de alta. Puesto que esta codificación es común a los tres servicios, extraeremos el valor almacenado en la columna relativa a este código, y no aquella relativa a la descripción, para evitar ligeras diferencias que puedan existir de escritura. También creamos el diccionario *altas_dict*, que contiene como claves cada una de las codificaciones contenidas en los dataframes mencionados, y como valores las respectivas descripciones. Utilizamos este diccionario para darle a la codificación extraída su respectiva descripción, la cual almacenamos en la clave *Alta*, en el diccionario almacenado en *sip*. No obstante, en algunas ocasiones este valor puede estar ausente en la primera fila. Esto se puede deber, como dijimos anteriormente, a que un paciente reciba dos diagnósticos o procedimientos en una visita y estos sean tomados en dos registros distintos a pesar de corresponder a la misma visita. En este caso, tomamos el código de la circunstancia de alta de la segunda fila.

Para el caso de que este *servicio* final del *Recorrido* sea *Emergencias*, el valor que almacenaremos en la clave *Alta* será el valor de la columna *FINALIZACIÓN*. Aunque este valor no esté en el mismo abanico de valores posibles que los que hemos visto en los tres otros servicios, es indicativo de lo que se hace con el paciente después de atenderle, y por lo tanto, es lo más cercano que tenemos.

IV.1.6 Unión de toda la información extraída en un dataframe de *pandas*

El objetivo final de todo este proceso es unir toda la información relativa a cada episodio que queramos utilizar en el estudio en un objeto manejable para su manipulación y visualización. El formato elegido ha sido el muy común *dataframe* de la librería *pandas*.

Para crear este dataframe, creamos el diccionario vacío *todo_comun_df*, que contendrá una clave por cada columna que queramos crear, y dentro de cada una de estas claves, encontraremos una lista vacía. Iremos rellenando esta lista con los respectivos valores que hayamos extraído de los episodios en las fases previas. Las claves serán: *Episodio*, *Sip*, *Edad*, *Fecha_Inicio*, *Recorrido*, *Entrada*, *Salida*, *Alta*, *Cronicidad*, *Sexo* y *Lugar*.

Para proveer a este diccionario de la información deseada, iteraremos sobre las claves de *unique_sips_separated_by_episodes* (sea *sip* la clave iterada), y dentro de ésta iteraremos sobre las claves del diccionario almacenado en ella (sea *episodio* la clave iterada, asegurándonos de que es un número y no una cadena de texto como *Sexo*). En cada iteración, añadimos a cada lista el valor correspondiente, veamos en detalle lo añadido a cada lista:

- *Episodio*: El número de iteración, que actuará como identificador de cada episodio.
- *Sip*: El valor de *sip*, que actuará como identificador del paciente.

- *Recorrido*: La lista almacenada en la clave *Recorrido*, dentro de cada *episodo*. Representa el recorrido asistencial seguido por el paciente durante el episodio de infarto. Aunque que es una lista, ésta se almacenará en el dataframe como una cadena de texto, y no como una lista, por los tipos de valores que se pueden almacenar. Deberemos tenerlo en cuenta a la hora de manipularlo en el análisis.

- *Recorrido_Diagnostico*: La lista almacenada en la clave *Recorrido diagnostico*, dentro de cada *episodo*. Representa los distintos diagnósticos asignados al paciente durante el recorrido asistencial. Igual que en el caso de la variable *Recorrido*, se almacenará en el dataframe como una cadena de texto.

- *Entrada*: El primer valor de la lista almacenada en la clave *Recorrido* dentro de cada *episodio*, que indica la vía de entrada del paciente al sistema sanitario.

- *Salida*: El último valor de la lista almacenada en la clave *Recorrido* dentro de cada *episodio*, que indica el último servicio utilizado por el paciente en el sistema sanitario.

- *Alta*: La cadena de texto almacenada en la clave *Alta*, dentro de cada *episodio*. Representa el tipo de alta dada al paciente al finalizar el recorrido asistencial.

- *Sexo*: La cadena de texto almacenada en la clave *Sexo*, dentro de cada *sip*. Representa el sexo del paciente, en caso de existir, en caso contrario, tomará el valor *Desconocido*.

- *Fecha_Inicio*: La fecha almacenada en el primer registro encontrado en la variable *FECHA_REFERENCIA* perteneciente al dataframe correspondiente al servicio de entrada al sistema sanitario del paciente en cada *episodio*. Representa la fecha en la cual da inicio la atención médica al paciente.

- *Lugar*: La cadena de texto almacenada en la clave *Lugar*, dentro de cada *episodio*. Representa el departamento de salud en el cual se ha producido la entrada al sistema sanitario del paciente.

Finalmente, guardaremos el dataframe resultante en el archivo *datos_para_representar_graficas_lat_lon_5.csv*.

IV.2 Análisis de los datos

Para realizar el análisis de los datos, hemos dividido el trabajo en diversos notebooks, enfocados cada uno en un tipo de análisis. Veamos primero la operación común a realizar en todas las etapas, y después el trabajo parte por parte.

IV.2.0.1 Operaciones comunes

La operación común a realizar en todos los notebooks es la de filtrar el dataset por aquellos episodios que hayan recibido en algún momento un diagnóstico compatible con SCA. Veamos el proceso seguido para realizar este filtrado.

En primer lugar, por supuesto, extraemos en un dataframe de pandas el archivo *datos_para_representar_graficas_lat_lon_5.csv*. Después, creamos una expresión regular que contenga los códigos CIE-9 y CIE-10 de estos diagnósticos. Finalmente, con

la función *pd.Series.str.contains*, comprobamos cuáles son las filas que contienen alguno de estos diagnósticos, y desechamos las demás.

Previamente a la realización de esta operación, podíamos observar que el dataframe contenía 482030 filas, es decir, 482030 episodios. Sin embargo, una vez hemos hecho estos cambios, vemos que el número se ha reducido considerablemente, hasta los 60638.

Los diagnósticos que no se pueden encontrar en ninguno de los episodios con los que trataremos son: *I23.4, I23.5, I25.701, I25.711M I25.721, I25.731, I25.738, I25.739, I25.751, I25.760, I25.761, I25.768, I25.769, I25.791* y *00.66*.

IV.2.0.2 Análisis general

En primer lugar, realizaremos un análisis exploratorio general de los datos, para contestar a las preguntas relativas al prefil general del paciente. Todas las operaciones relativas a este análisis general se pueden encontrar en el notebook *4_a_Estudio_general_datos_referencia.ipynb*. Podremos observar que el dataframe filtrado como describimos en el apartado anterior se almacena en la variable *cruzar_datos_df_filtered*. Lo primero que podemos observar es que este dataframe contiene 46946 identificadores diferentes en la columna *Sip*. Esto significa que hasta 8930 de los 55876 identificadores que observamos al principio fueron eliminados por no contener ningún diagnóstico compatible con SCA en ninguno de los episodios, lo cual se debe, como explicamos con anterioridad, a un error de extracción de los datos.

IV.2.0.3 Datos acumulados por sexo

En primer lugar, echemos un vistazo a los datos acumulados por sexo. Podemos observar que existe la siguiente distribución por sexos:

Sexo	**Episodios**	**% episodios**	**Individuos únicos**	**% Individuos únicos**
Hombre	41144	67.85	31586	67.28
Mujer	19191	31.65	15358	32.71
Desconocido	303	0.50	2	$\leq$0.01

Cuadro 1: Episodios totales y número de individuos por sexo (variable *Sexo)*.

Como podemos observar, el número de individuos únicos que han sufrido de algún episodio de SCA entre 2016 y 2019 es claramente superior entre los hombres (67.28 %) con respecto a las mujeres (32.71 %). Además, si miramos con respecto a los episodios, en lugar de los individuos únicos, podemos ver que la diferencia es aún mayor, siendo un 67.91 % de hombres con respecto a un 32.09 % de mujeres.

IV.2.0.4 Episodios acumulados por servicio de entrada

Echemos ahora un vistazo a los datos acumulados por servicio de entrada al sistema sanitario. Aplicando la función *porcentajes_entradas* que se puede encontrar en el archivo *help_functions.py*, podemos observar la siguiente distribución:

Entrada	Episodios	% episodios
UHOSP	31115	51.31
UAP	12269	20.23
Emergencias	11150	18.39
CMBD	6104	10.07

Cuadro 2: Episodios totales por servicio de entrada (variable *Entrada*) al sistema sanitario.

Claramente, la vía de entrada más repetida al sistema sanitario son las urgencias hospitalarias. Este dato es esperable, puesto que los infartos suelen presentar síntomas dolorosos y que requieren de atención inmediata, por lo cual la gente tiende a acudir a servicios de urgencias. Llama la atención también el alto porcentaje de episodios que empiezan en centros de atención primaria, relacionados con pacientes que, o bien subestiman los síntomas que sufren, o bien viven lejos de un centro hospitalario donde se les pueda atender de manera inmediata. Es igualmente interesante hablar sobre el relativamente bajo porcentaje de casos que entran a través del servicio de emergencias, siendo una afección que, como ya hemos mencionado, requiere de la mayor rapidez posible a la hora de atenderla. Esto se puede deber a que los pacientes pueden preferir desplazarse por sus propios medios a los servicios de urgencias si se sienten capaces, a pesar de ser esto contraproducente, puesto que no reciben la atención durante el desplazamiento. Relacionado con este punto, puesto que hay muchas personas que únicamente acuden al servicio de emergencias en caso de tener una afección que les impida moverse por sí solos, en muchos casos el paciente fallece antes de que el servicio de emergencias le pueda atender, lo cual no se recoge en este conjunto de datos. Esta mala actuación por parte de los pacientes será un punto a mencionar en las conclusiones, en cuanto a la necesidad de solicitar ayuda de manera inmediata en caso de sentir síntomas compatibles con esta afección. Además, debemos recordar que estos resultados se ven influenciados por la falta de episodios no relacionados con SCA que hay en esta fuente, tal y como mencionamos al inicio del trabajo. Por esto, podemos estar perdiendo casos que no hayan sido diagnosticados como un episodio de SCA durante la atención de emergencias, pero que luego en el resto de servicios efectivamente se haya visto que sí lo era.

IV.2.0.5 Episodios acumulados por servicio de salida

Echemos ahora un vistazo a los datos acumulados por servicio de salida del sistema sanitario, es decir, por el servicio en el que el paciente recibe el alta. Aplicando la función *porcentajes_salidas* que se puede encontrar en el archivo *help_functions.py*, podemos observar la siguiente distribución:

Entrada	**Episodios**	**% episodios**
CMBD	44569	73.50
UHOSP	7970	13.14
UAP	4442	7.33
Emergencias	3657	6.03

Cuadro 3: Episodios por servicio de salida (variable *Salida*) del sistema sanitario.

Como podemos observar, el 73.50 % de los pacientes finalizan el recorrido asistencial en una hospitalización. Efectivamente, éste es el protocolo a seguir durante un episodio de infarto (como indica el plan código infarto de la Generalitat Valenciana [1]), puesto que es una patología muy grave que puede requerir de intervenciones u observación sobre el paciente. No obstante, podemos observar que el 26.50 % de los pacientes no han sido derivados al hospital, lo cual se puede deber a diversos motivos:

- El paciente ha fallecido.
- El paciente ha sido derivado a otro servicio de manera correcta, pero no ha podido (o no ha querido) acudir.
- El episodio había sido diagnosticado al principio como un infarto, pero el diagnóstico ha sido después corregido por otro que no requiere de ingreso hospitalario.
- Posteriormente a su hospitalización o intervención hospitalaria, el paciente ha acudido a otro servicio del sistema sanitario.
- El paciente ha recibido una mala asistencia sanitaria y no ha seguido el protocolo de infartos.

En cualquier caso, ese 23.50 % de los casos es muy llamativo y se sale del protocolo establecido. Si tomamos una mirada más profunda, podemos observar más bien el porcentaje de episodios en los que el paciente no ha pasado en ningún momento por el servicio de hospitalización, es decir, no ha sido ni ingresados ni intervenidos en el hospital (no podemos encontrar *CMBD* en su recorrido asistencial). Podemos observar que 48503 de los episodios que tenemos han pasado por el servicio de hospitalización al menos en una ocasión, lo cual significa que en 12135 episodios esto no ha ocurrido. Por lo tanto, ya podemos decir que en estos 12135 casos (alrededor de un 20.01 %) el protocolo no se ha seguido bien, puesto que estos pacientes no han sido ni ingresados ni intervenidos en un hospital a pesar de mostrar síntomas compatibles con un diagnóstico de SCA.

IV.2.0.6 Episodios acumulados por recorrido asistencial

Buscamos ahora las conclusiones que podemos extraer de la variable *Recorrido*. Si acumulamos los datos con respecto a los valores contenidos, podemos observar que existen 343 recorridos asistenciales distintos entre todos los episodios. No obstante, la gran mayoría de ellos presentan menos de 5 casos, y se podrían considerar como outliers. Por ello, nosotros nos centraremos en aquellos recorridos que podemos encontrar en, por lo menos, 1000 episodios diferentes, que son los siguientes:

Recorrido	**Episodios**	**% episodios**
[UHOSP, CMBD]	23501	38.76
[CMBD]	5868	9.68
[UAP, UHOSP, CMBD]	4609	7.60
[Emergencias, UHOSP, CMBD]	3499	5.77
[Emergencias, UHOSP]	2725	4.49
[UHOSP]	2707	4.46
[UAP]	2139	3.53
[Emergencias]	1876	3.09
[Emergencias, CMBD]	1504	2.48
[UHOSP, Emergencias, CMBD]	1225	2.02

Cuadro 4: Episodios por recorridos (variable *Recorrido*) con más de 1000 instancias.

Describamos cada uno de estos recorridos y veamos si efectivamente son correctos:

- [**UHOSP, CMBD**]: paciente que acude a urgencias y es hospitalizado. Recorrido más común, protocolo bien llevado a cabo, puesto que un paciente con un infarto de gravedad debe ser ingresado.

- [**CMBD**]: paciente que es directamente hospitalizado. Protocolo bien llevado a cabo, puesto que un paciente con un infarto que acude al hospital y es inmediatamente ingresado debido a la gravedad de su patología, o es directamente intervenido en el hospital con una cirugía de urgencias en el departamento de cardiología.

- [**UAP, UHOSP, CMBD**]: paciente que acude a una consulta de atención primaria, es derivada a urgencias y ahí es ingresado en el hospital. Protocolo bien llevado a cabo, puesto que el paciente es derivado del centro de atención primaria a las urgencias para poder tratar su grave patología, donde es inmediatamente ingresado.

- [**Emergencias, UHOSP, CMBD**]: paciente que solicita el servicio de emergencias, es transportado a urgencias y ahí es ingresado en el hospital. Protocolo bien llevado a cabo, puesto que el paciente llama al servicio de emergencias al sentirse grave, los cuales lo transportan a las urgencias para poder tratar su grave patología, donde es inmediatamente ingresado.

- [**Emergencias, UHOSP**]: Paciente que solicita el servicio de emergencias, y es transportado a urgencias donde es atendido. El protocolo es bien llevado a cabo por el servicio de emergencias, mientras que el servicio de urgencias hospitalarias puede haber cometido un fallo tal y como vimos en el apartado de las salidas.

- [**UHOSP**]: paciente que acude al servicio de urgencias debido a su patología. El servicio de urgencias hospitalarias puede haber cometido un fallo tal y como vimos en el apartado de las salidas.

- [**UAP**]: paciente que acude al centro de atención primaria y ahí finaliza su recorrido. El servicio de atención primaria puede haber cometido un fallo tal y como vimos en el apartado de las salidas.

- [**Emergencias**]: Paciente que solicita el servicio de emergencias. El servicio de emergencias puede haber cometido un fallo tal y como vimos en el apartado de las salidas.

- [**Emergencias, CMBD**]: Paciente que solicita el servicio de emergencias, y es trasladado directamente al hospital. En este caso, el paciente es trasladado directamente al hospital, donde, debido a la gravedad de su caso, es hospitalizado conforme entra y/o intervenido de urgencia en el servicio de cardiología.

- [**UHOSP, Emergencias, CMBD**]: Paciente que acude a urgencias, y necesita ser trasladado en ambulancia por el servicio de emergencias a otro hospital que disponga de unidad de cardiología, la cual no está disponible en todos. En este caso, el paciente es trasladado por necesitar una intervención que no se le puede dar en el hospital que aloja las urgencias de origen.

Estos son los diez recorridos con más de 1000 casos. Como podemos ver, no existen errores graves en las fases intermedias de estos casos, sino que todos los posibles cuestionamientos se encuentran en la fase final (donde los pacientes no son hospitalizados).

Entre estos diez recorridos suman el 81.89 % de los casos que tenemos disponibles, lo cual supone una gran mayoría de los casos el recorrido asistencial.

IV.2.0.7 Episodios acumulados por departamento de atención

Buscamos ahora información en la variable *Lugar*, que contiene el departamento de entrada de cada uno de los episodios del dataframe. Además, solicitamos también, en una consulta fuera de la realizada al inicio del trabajo, que nos fuesen facilitada la población de cada departamento de salud. Aquí podemos ver los episodios acumulados por departamento de salud:

Departamento	Ep.Totales	% Totales	Poblacion	Ratio episodios / población
VALÈNCIA - LA FE	4379	7.22	290447	0.0151
LA RIBERA	3712	6.12	264023	0.0141
VALÈNCIA ARNAU DE VILANOVA	3664	6.04	324034	0.0113
SANT JOAN D'ALACANT	3490	5.76	225153	0.0155
TORREVIEJA	3206	5.29	188602	0.0170
CASTELLÓ	2967	4.89	284183	0.0104
ALACANT - HOSPITAL GENERAL	2883	4.75	280535	0.0103
VALÈNCIA - DR. PESET	2870	4.73	280052	0.0102
VALÈNCIA CLÍNIC-LA MALVA-ROSA	2777	4.58	348137	0.0080
VALÈNCIA - HOSPITAL GENERAL	2770	4.57	367149	0.0075
MARINA BAIXA	2587	4.27	184956	0.0140
MANISES	2525	4.16	208384	0.0121
DÉNIA	2394	3.95	174194	0.0137
GANDIA	2386	3.93	179272	0.0133
ELDA	2321	3.83	189629	0.0122
ONTINYENT-XÀTIVA	2227	3.67	194800	0.0114
LA PLANA	2099	3.46	187616	0.0112
ELX - HOSPITAL GENERAL	2096	3.46	169599	0.0124
SAGUNT	1970	3.25	154712	0.0127
ORIHUELA	1925	3.17	169580	0.0114
ELX-CREVILLENT	1650	2.72	157314	0.0105
ALCOI	1408	2.32	137336	0.0102
VINARÒS	1250	2.06	90365	0.0138
REQUENA	1056	1.74	51622	0.0205
Sin Determinar	26	0.04	0	-

Cuadro 5: Episodios por departamento de salud (variable *Lugar*).

Lo primero que debemos reseñar es que el valor guardado en la columna *% Población* no indica que esa sea la proporción de habitantes de ese departamento que haya sufrido de episodios de SCA entre 2016 y 2019. Esto es, primero, porque estamos contando el número de episodios, y no de individuos únicos, y además, porque no disponemos de los datos de empadronamiento de los pacientes, simplemente sabemos dónde empiezan los episodios.

Como podemos observar, aunque como es lógico los departamentos de salud con más habitantes tienden a tener más casos, porcentualmente no es el caso. El departamento de salud de Requena es el más llamativo, puesto que es por un amplio margen el que mayor ratio de episodios por habitante muestra. En su contraparte, dos departamentos centrados en la ciudad de València como son el departamento de salud de Valencia - Hospital General y el departamento de salud Clínic - Malva-Rosa muestran los mínimos ratios de todo el conjunto de datos.

IV.2.0.8 Episodios acumulados por tipo de alta

Como hemos visto, muchos episodios no tienen el final de recorrido que deberían, puesto que no siempre acaban en una hospitalización. Pasemos ahora a ver cuáles son los tipos de alta. Lo primero que podemos observar es que hay hasta 36 tipos de alta distintos, por ello, veremos nuevamente aquellos que tienen más de 1000 episodios:

Alta	**Episodios**	**% Totales**
Domicilio	22657	37.36
Equipo atención primaria	12369	20.40
Consultas externas	11230	18.52
Traslado Hospital de agudos	4400	7.26
TRANSPORTE SANITARIO SECUNDARIO	2829	4.67
Éxitus	2304	3.80
Urgencias	1816	2.99

Cuadro 6: Episodios por tipo de alta (variable *Alta*) con más de 1000 instancias.

En total, el 95 % de los episodios están incluidos en estos 7 tipos de altas. Comentemos cada una de ellas por separado:

- *Domicilio*: El paciente es enviado a su domicilio. Es un alta correcta si el paciente proviene de una hospitalización, ya ha sido estabilizado y ha estado en observación, y no requiere de más vigilancia exhaustiva.

- *Equipo atención primaria*: El paciente es derivado al equipo de atención primaria. Esto significa que el paciente es mandado a casa con visitas recurrentes a su centro de atención primaria, o de los profesionales de atención primaria a su domicilio, para tenerlo en continua observación. Esta puede ser un alta correcta si el paciente proviene de una hospitalización, donde ya ha sido estabilizado y ha estado en observación, y no requiere de más vigilancia exhaustiva.

- *Consultas externas*: El paciente es dado de alta pero debe acudir a consultas externas del hospital para ser sometido a exámenes y/o pruebas diagnósticas. Puede ser un alta correcta si proviene de una hospitalización, donde ya ha sido estabilizado y ha estado en observación, y no requiere de más vigilancia exhaustiva.

- *Traslado Hospital de agudos*: El paciente es trasladado a un hospital de agudos, es decir, uno donde se traten las enfermedades no crónicas. Este alta es incorrecta, o este episodio puede estar incompleto, puesto que, si un paciente es trasladado al hospital de agudos, deberíamos de tener los datos del respectivo ingreso.

- *TRANSPORTE SANITARIO SECUNDARIO*: Este alta es una de las pertenecientes a la fuente de emergencias, y se refiere a un transporte realizado entre dos centros hospitalarios. Esto puede suceder, por ejemplo, si el paciente ha acudido a urgencias de un hospital sin servicio de cardiología, y que necesita ser atendido en otro hospital donde sí tengan los medios que requiere su patología. Al igual que en el caso anterior, este tipo de alta se trata de un error, puesto que deberíamos tener el regisro posterior al traslado del paciente.

- *Éxitus*: Este alta corresponde al fallecimiento del paciente. Es un alta correcta.

- *Urgencias*: Este alta corresponde al traslado del paciente a urgencias hospitalarias. Igual que en casos anteriores, es un tipo de alta errónea, puesto que deberíamos de tener el registro de urgencias hospitalarias correspondiente.

Como podemos ver, existen diversas altas incorrectas en los registros, especialmente referentes a los . Esto se puede deber a diversas razones:

- Un paciente que haya decidido volver a su domicilio en lugar de dirigirse al centro al que ha sido derivado. Esto es una negligencia por parte del paciente.
- Al llegar al centro de destino, los datos del paciente no han sido tomados, o han sido tomados con más de un día de diferencia. Esto es una negligencia por parte de la atención en el centro de destino.
- El paciente ha fallecido durante su traslado entre centros, cuando éste no se ha producido por parte del servicio de emergencias.
- El código de alta ha sido tomado de manera errónea por parte del último centro de éste recorrido, provocando así éste error. Esto es una negligencia por parte de la atención en este centro.

Puesto que hemos visto que diversos tipos de alta no son fiables, tan solo estudiaremos uno de ellos: los casos de Éxitus. Estos se refieren a cuando un paciente fallece durante su atención en cualquiera de los cuatro servicios del sistema sanitario estudiados. Si bien los pacientes pueden fallecer por el mismo episodio de infarto cuando están fuera del sistema sanitario (como acabamos de explicar, por ejemplo, pueden fallecer en el trayecto entre dos servicios), podemos estar seguros de que los casos marcados como éxitus no son un error en una gran mayoría de los casos (no se da por fallecido a un paciente por error). Puede llamar la atención que el porcentaje sea relativamente bajo (alrededor de 3.80 %), pero esto tiene una explicación: el paciente ha sido atendido por el sistema sanitario, lo cual reduce en grandísima medida la mortalidad. Esto no significa que la mortalidad de la enfermedad sea 3.80 % en general, puesto que la mayoría de infartos mortales se producen cuando el paciente no ha solicitado la ayuda del sistema sanitario.

IV.2.0.9 Episodios acumulados por edad

Acumulamos ahora los episodios por edad, siendo cuidadosos para eliminar aquellas cuyo valor es *-1* antes de realizar los cálculos. Las métricas principales de esta variable son las siguientes:

count	60274
mean	68.67
std	13.37
min	17
25 %	59
50 %	70
75 %	79
max	111

Cuadro 7: Estadísticos descriptivos generales de la variable *Edad*.

Como podemos observar, por lo menos la mitad de los registros se encuentran entre los 59 y los 79 años, que es la edad que concentra la mayoría de casos. Esto se ve igualmente reflejado en el histograma de la misma variable:

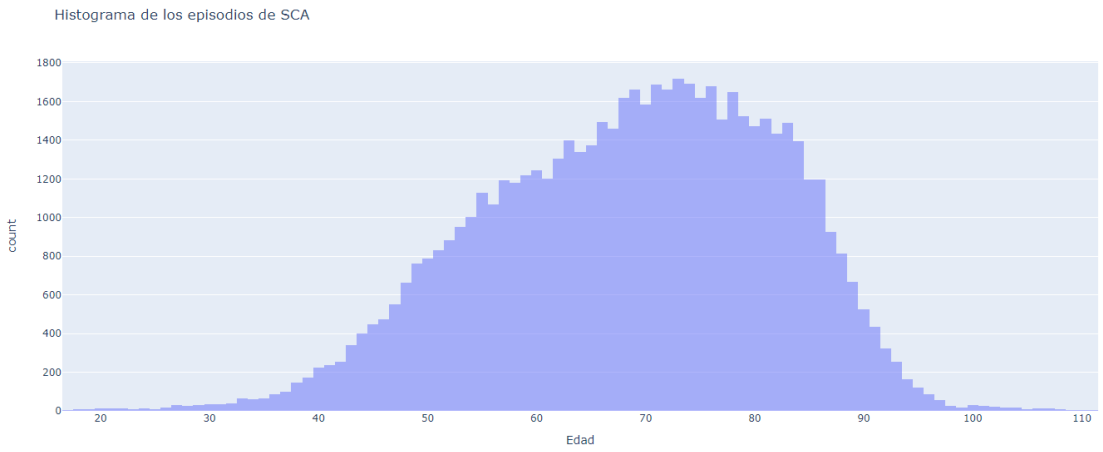

Figura 1: Histograma de la variable *Edad*.

Sabemos que la edad es un factor que no afecta de igual manera a los hombres que a las mujeres, puesto que los hombres tienden a tener episodios de SCA siendo más jóvenes que las mujeres (como podemos ver en este estudio publicado en la revista española de cardiología [6]). Vamos a ver si esto se cumple en este caso:

Para realizar los cálculos que veremos a continuación, hemos eliminado el valor *Desconocido* de la variable *Sexo*. Las métricas de edad acumuladas por sexo son las siguientes:

	Hombre	Mujer
count	41116	19158
mean	66.86	72.57
std	13.10	13.13
min	17	17
25 %	57	64
50 %	68	74
75 %	77	82
max	111	109

Cuadro 8: Estadísticos descriptivos de la variable *Edad* por *Sexo*.

Y el histograma diferenciado por sexos se puede ver aquí:

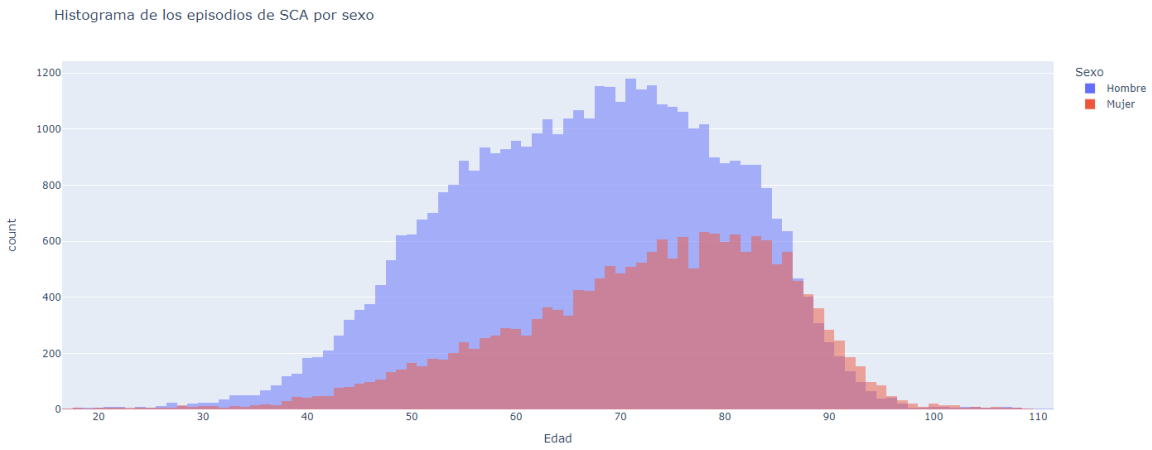

Figura 2: Historgrama de la variable *Edad* por *Sexo*.

Efectivamente, todos los indicadores apuntan a que no estamos en un caso especial, y que la media de edad de los hombres en este grupo es inferior a la de las mujeres. Los números son claros, teniendo las mujeres una edad mayor en todos los estadísticos comprobados. En cuanto al histograma, podemos ver que los hombres el número de casos crece mucho a partir de los 45 años, alcanza el pico de máximo número de casos alrededor de los 70 años, y a partir de los 80 van decreciendo en picado; mientras que para las mujeres, el crecimiento es más progresivo, alcanza los máximos valores entre los 75 y los 85 años, y después sufre una bajada. Para comprobar estas diferencias, hemos realizado un t-test, planteando el siguiente contraste de hipótesis:

Si μ_{Hombre} y μ_{Mujer} son las medias de edad de hombres y mujeres en este grupo:

H_0: no hay diferencia entre las medias de edad entre hombres y mujeres que sufren infartos en la Comunitat Valenciana: $\mu_{Hombre} = \mu_{Mujer}$.

H_a: si hay diferencia entre las medias de edad entre hombres y mujeres que sufren infartos en la Comunitat Valenciana: $\mu_{Hombre} \leq \mu_{Mujer}$.

Los detalles de las condiciones necesarias para la aplicación del test se pueden encontrar en el mismo notebook. El resultado del estadístico salido del test es -49.81, que supone un p-valor directamente redondeado directamente en 0. Si tomamos un nivel de significación $\alpha = 0{,}05$, que es el que tomaremos en todos los tests de aquí en adelante, podemos rechazar la hipótesis nula.

Como conclusiones de este notebook, podemos extraer lo siguiente:

- Tenemos muchos más casos de hombres (31586) que de mujeres (15358) que han sufrido un episodio con diagnóstico compatible con SCA (más del doble de casos).
- La edad media de los hombres con diagnósticos compatibles con SCA es en promedio menor a la media de edad de las mujeres que lo padecen.
- Un porcentaje muy alto de los pacientes (por lo menos un 20.01 %) no ha sido atendido de manera eficiente a pesar de acudir al sistema sanitario.

- Un alto porcentaje de los pacientes acude al centro de atención primaria (12269 casos, un 20.23 %), a pesar de sentirse con síntomas compatibles con una afección tan grave como es el SCA. Es llamativo que el porcentaje de episodios cuya entrada al sistema sanitario es la vía de Emergencias es menor (11150 casos, el 18.39 %).

- Un muy alto porcentaje de las altas más comunes es erróneo (9045 casos entre *Traslado Hospital Agudos*, *TRANSPORTE SANITARO SECUNDARIO* y *Urgencias*, por lo menos un 14.91 % de casos erróneos).

Veamos ahora diversos cruces de las variables para extraer conclusiones más profundas.

IV.2.1 Datos cruzados con los departamentos de salud

La primera variable que nos vamos a encargar de cruzar con las demás es la variable *Lugar*, que contiene el departamento de salud de entrada de cada uno de los pacientes. En el punto anterior no hemos conseguido mucha información relevante a este respecto más allá de algunas observaciones superficiales, por lo que queremos ver cómo influye la localización del paciente en las demás variables.

Para ello, hemos construido el notebook *4_b_Estudio_por_departamento.ipynb*, donde se podrán ver en más detalle los cálculos que realizaremos. Hemos realizado el mismo estudio para todos los departamentos de salud, por lo cual, detallaremos el estudio realizado para uno de ellos, y después hablaremos de las conclusiones que hemos podido extraer del conjunto. Por supuesto, los 26 episodios cuya localización es indeterminada que hemos visto en el punto anterior serán ignorados en este notebook.

A lo largo de este, y de los siguientes notebooks, realizaremos una serie de contrastes de hipótesis. En su mayor parte, realizaremos comparaciones con respecto a la población valenciana en general, que es la población que estamos estudiando. En cualquiera de estos casos, tomaremos como referencia los números extraídos del apartado anterior, que serán aquellos con los que realizaremos las comparaciones. Aunque no sean estadísticos poblacionales sino muestrales, los tomaremos como referencia para ver si existen diferencias entre la muestra general y las muestras filtradas por cada uno de los datos.

IV.2.1.1 Un ejemplo: el departamento de València Arnau de Vilanova-Llíria

En este notebook, hemos realizado los mismos contrastes para obtener la misma información de todos los departamentos de salud. Puesto que exponer las operaciones realizadas en los 24 departamentos sería muy largo y repetitivo, expondremos el caso del departamento de València Arnau de Vilanova-Llíria. Para obtener los datos relativos a este departamento, aplicamos la función *filtrar_grupo*, que podemos encontrar en el archivo *help_functions.py*.

Después de una primera exploración de la variable *Edad* realizada gracias a la función *describir_grupo* que encontramos en el archivo *help_functions.py*, la primera operación que hemos realizado es el mismo contraste de hipótesis llevado a cabo para el caso general, con el objetivo de verificar si a partir de esta muestra podemos afirmar que la edad media de los hombres que sufren de episodios de SCA en este departamento

es inferior a la de las mujeres. El contraste de hipótesis es:

Si μ_{Hombre} y μ_{Mujer} son las medias de edad de hombres y mujeres en este grupo:

H_0: no hay diferencia entre las medias de edad entre hombres y mujeres que sufren infartos en el departamento de salud de València Arnau de Vilanova - Llíria: $\mu_{Hombre} = \mu_{Mujer}$.

H_a: si hay diferencia entre las medias de edad entre hombres y mujeres que sufren infartos en el departamento de salud de València Arnau de Vilanova - Llíria: $\mu_{Hombre} \leq \mu_{Mujer}$.

El código para verificar las condiciones necesarias para poder realizar el t-test que necesitamos para resolver este contraste se realizan en las funciones *testear_normalidad* y *test_levene*. Por otra parte, el código que contiene el test realizado está en la función *ttest*. Las tres funciones están en el archivo *help_functions.py*. En este caso, obtenemos un p-valor del orden de 10^{-34}, muy inferior al nivel de significación $\alpha = 0{,}05$ establecido para todos los casos, por lo tanto podemos afirmar que efectivamente la edad media es menor para hombres que para mujeres en casos de SCA del departamento de València Arnau de Vilanova - Llíria.

En segundo lugar, queremos verificar si existe una diferencia significativa entre la proporción de individuos de sexo masculino y femenino que sufren de SCA en este departamento y en la Comunitat Valenciana. Como hemos visto antes, el porcentaje de hombres que han sufrido de SCA en la Comunitat en este periodo es 67.28 %, por un 32.71 % de mujeres. En el caso del departamento de València Arnau de Vilanova - Llíria, el porcentaje de hombres es de 66.45 % (1929 individuos) por 33.55 % de mujeres (974 individuos). Siendo estos los datos de los que disponemos, trataremos de resolver el contraste de hipótesis siguiente:

Si P_{Hombre} y P_{Mujer} representan respectivamente las proporciones de hombres y mujeres en esta muestra:

H_0: $P_{Hombre} = 67{,}28$ y $P_{Mujer} = 32{,}71$.

H_a: $P_{Hombre} \neq 67{,}28$ y $P_{Mujer} \neq 32{,}71$.

Para resolver este contraste, realizamos un test χ^2, cuyas condiciones necesarias para su aplicación están verificadas en la propia función creada para la resolución del test, la función *contrastar_proporciones* alojada en el archivo *help_functions.py*. En este caso, el p-valor obtenido es 0.34, que es muy superior al nivel de significación $\alpha = 0{,}05$, por lo tanto no podemos descartar la hipótesis nula, y podemos considerar que no existen diferencias significativas entre el porcentaje de hombres y mujeres que sufren de SCA en el departamento de València Arnau de Vilanova - Llíria con respecto a la Comunitat Valenciana.

En el siguiente paso, veremos la influencia del departamento de salud en la variable edad. En este caso, la media de edad de los individuos que padecen de un episodio de SCA en el departamento de València Arnau de Vilanova - Llíria es de 68.22 años,

mientras que en la Comunitat Valenciana es de 68.67 años. Es por ello que planteamos el siguiente contraste de hipótesis:

Si μ representa la media de edad de este grupo y $\mu_0 = 68{,}67$ es la media de edad en la Comunitat Valenciana,

H_0: $\mu = \mu_0$

H_a: $\mu \leq \mu_0$

Planteamos este contraste porque, en este caso, la media de edad es 68.20 años, que es inferior a los 68.67 de la Comunitat Valenciana. En otros departamentos, con una media de edad superior a la de la Comunitat, plantearemos el contraste contrario.

Planteamos un t-test para resolver este contraste, cuyos detalles de las condiciones necesarias están verificadas en la propia función creada para la resolución del test, la función *contrastar_media_edad* alojada en el archivo *help_functions.py*. En este caso, obtenemos un p-valor de 0.025, inferior al nivel de significación $\alpha = 0{,}05$. Por lo tanto, podemos afirmar que existe una diferencia significativa entre las medias de edad de este grupo y de la Comunitat Valenciana.

Realizamos los contrastes análogos para las medias de edad de hombres y mujeres. En este caso, la media de edad de los hombres que sufren de SCA en el departamento de València Arnau de Vilanova-Llíria es 66.24 (en la Comunitat 66.86), mientras que la media en el caso de las mujeres la media de edad es 72.18 (en la Comunitat 72.57). Gracias a la función *contrastar_media_edad_sexo*, que encontraremos en el archivo *helper_functions.py* y que contiene el código tanto para verificar las condiciones necesarias como para realizar el t-test, obtenemos que existe evidencia significativa de que, en el caso de los hombres que han sufrido un episodio de SCA en el departamento de València Arnau de Vilanova-Llíria, la media de edad es inferior a 66.86 (p-valor 0.01). No obstante, en el caso de las mujeres, obtenemos un p-valor igual a 0.16, superior al nivel de significación, que por lo tanto nos dice que la media de edad poblacional de este grupo no es inferior a 72.57 años.

Posteriormente, pasamos a contrastar el porcentaje de episodios que no han pasado por el servicio de hospitalizaciones (CMBD). Queremos comprobar si el lugar de procedencia de los pacientes puede tener alguna influencia positiva o negativa en el correcto funcionamiento del recorrido asistencial.

Como hemos visto en el caso general, alrededor del 79.99 % de los casos habían sido derivados en algún momento al servicio CMBD. Si miramos los datos del departamento de València Arnau de Vilanova-Llíria, podemos ver que en este caso alrededor de 80.93 % de los episodios han sido derivados a CMBD en algún momento, mientras que 19.06 % no lo han sido nunca (en contraste con el 20.01 % del caso de toda la Comunitat). Planteamos el siguiente contraste de hipótesis:

Si P_{CMBD} y P_{No_CMBD} representan, respectivamente, las proporciones de individuos que han pasado y no han pasado por CMBD en la muestra, realizaremos el siguiente contraste de hipótesis:

H_0: $P_{CMBD} = 79{,}99$ y $P_{No\_CMBD} = 20{,}01$.

H_a: $P_{CMBD} \neq 79{,}99$ y $P_{No\_CMBD} \neq 20{,}01$.

Para resolver este contraste, realizamos un test χ^2, cuyas condiciones necesarias para su aplicación están verificadas en la propia función creada para la resolución del test, la función *contrastar_proporciones_CMBD* que podemos ver en el archivo *help_functions.py*. En este caso concreto, obtenemos que el p-valor es 0.15, que es superior al nivel de significación $\alpha = 0{,}05$, por lo tanto, no podemos decir que las proporciones observadas en el departamento de València Arnau de Vilanova - Llíria sean significativamente diferentes a las observadas para la Comunitat.

Quedándonos en el tema del recorrido asistencial, vamos a comprobar ahora si la proporción de pacientes que entran al sistema sanitario mediante el servicio de atención primaria (UAP) es la misma en el departamento de València Arnau de Vilanova - Llíria que en toda la Comunitat. En el caso general hemos visto que aproximadamente el 20.23 % de los episodios entró al sistema sanitario mediante el servicio de atención primaria. No obstante, en este caso vemos que el porcentaje es mucho mayor, siendo un 31.13 % de los episodios. Planteamos el siguiente contraste de hipótesis:

Si P_{UAP} y P_{No_UAP} representan, respectivamente, las proporciones de individuos que han entrado y no han entrado por UAP al sistema sanitario en la muestra, realizaremos el siguiente contraste de hipótesis:

H_0: $P_{UAP} = 20{,}23$ y $P_{No\_UAP} = 79{,}77$.

H_a: $P_{UAP} \neq 20{,}23$ y $P_{No\_UAP} \neq 79{,}77$.

Nuevamente realizamos un test χ^2, cuyas condiciones necesarias para su aplicación están verificadas en la propia función creada para la resolución del test, la función *contrastar_proporciones_UAP* que podemos ver en el archivo *help_functions.py*. El p-valor que obtenemos es del orden de 10^{-111}, que nos indica que podemos rechazar la hipótesis nula y afirmar que hay efectivamente una diferencia significativa con respecto a la proporción obtenida en el caso general. Podemos encontrar una explicación en este caso en la composición del departamento, que abarca muchos municipios entre València y Ademuz, con tan solo dos hospitales, uno en València (Arnau de Vilanova), y el otro en Llíria. Estos quedan muy lejos de algunos pueblos como Aras de los Olmos o Ademuz, lo cual puede obligar al paciente a acudir al centro de atención primaria más cercano para ser tratado con la mayor brevedad posible.

Por último, comprobaremos cómo es la proporción de casos de Éxitus en este departamento en comparación al caso de toda la Comunitat. En el departamento de València Arnau de Vilanova-Llíria, el porcentaje de casos de Éxitus es del 3.25 %, mientras que en la Comunitat Valenciana en general es de 3.8 %. Planteamos el siguiente contraste de hipótesis:

Si P_{Ex} y P_{No_Ex} representan, respectivamente, las proporciones de individuos que han finalizado y no han finalizado en éxitus la muestra, realizaremos el siguiente contraste de hipótesis:

H_0: $P_{Ex} = 3{,}80$ y $P_{No\_Ex} = 96{,}20$.

H_a: $P_{Ex} \neq 3{,}80$ y $P_{No\_Ex} \neq 96{,}20$.

Tal y como en los casos anteriores, realizamos un test χ^2, cuyas condiciones necesarias para su aplicación están verificadas en la propia función creada para la resolución del test, la función *contrastar_proporciones_Exitus* que podemos ver en el archivo *help_functions.py*, y que nos devuelve un p-valor de 0.08. Puesto que es superior al nivel de significación $\alpha = 0{,}05$, no podemos rechazar la hipótesis nula. En consecuencia, no podemos afirmar que la diferencia entre el porcentaje de casos de éxitus en este grupo y el porcentaje observado en la Comunitat sea significativa.

IV.2.1.2 Resultados en todos los departamentos

Una vez hemos visto el proceso que hemos llevado a cabo en un ejemplo, vamos a proceder a presentar los resultados obtenidos en todos los casos. Para ello, hemos aunado todos en una única tabla resumen, que se encuentra al final del notebook, que no podemos situar aquí por motivos de espacio. También podemos encontrar la tabla en los *documentos anexos a este trabajo*, a partir de la página 14. Las conclusiones que podemos extraer de este notebook son las siguientes:

En primer lugar, el porcentaje de episodios que entran al sistema sanitario por atención primaria es la variable estudiada en este notebook que más se diferencia entre departamentos. Únicamente el departamento de salud de Elda presenta un porcentaje parecido al general de la Comunitat Valenciana. Puesto que uno de los objetivos de este trabajo es buscar los fallos cometidos tanto por el paciente como por el sistema sanitario a la hora de tratar con un episodio de SCA, es relevante mencionar que encontramos hasta 14 departamentos que presentan un porcentaje más alto de episodios cuyo recorrido asistencial se inicia en el servicio de atención primaria. Entre ellos, destacan los casos del departamento de la Plana (38.25 % de los casos), Dénia (38.01 % de los casos) y Manises (34.42 % de los casos). En muchos de estos departamentos podemos encontrar un factor común muy claro: corresponden a departamentos grandes, con una gran cantidad de municipios que disponen de centros de atención primaria, pero no de hospitales a menos de 20 minutos en coche.

En segundo lugar, también hay una importante variación del porcentaje de episodios que pasan por el servicio de hospitalizaciones en función del departamento de salud que estemos tratando. En este caso, en 7 departamentos no hemos visto diferencias significativas con el porcentaje encontrado en la Comunitat, mientras que con 10 de ellos hemos visto un porcentaje significativamente menor. Llama la atención que entre ellos encontremos los casos de los departamentos de València - La Fe (74.47 %) y València - Hospital General (75.52 %), los cuales, al cubrir principalmente el área urbana de la ciudad de València, no nos dan a pensar ningún motivo de lejanía o de falta de medios para no hospitalizar a ningún paciente. También destaca en este lado negativo el departamento de salud de Requena, donde únicamente el 55.94 % de los episodios

fueron derivados al hospital.

En tercer lugar, los porcentajes de éxitus no varían demasiado, encontrándonos con hasta 17 departamentos cuyas diferencias porcentuales no son significativas con respecto a los datos de la Comunitat Valenciana. No obstante, este dato nos sirve para ilustrar, en compañía de los anteriores, el particular caso del departamento de salud de Dénia. En este caso, podemos ver que los tres datos mencionados hasta ahora en este apartado fallan: el porcentaje de pacientes que entra al sistema por atención primaria es significativamente mayor (38.01 %), el porcentaje de episodios derivados al servicio CMBD es significativamente menor (77.12 %), y el porcentaje de casos de éxitus es significativamente mayor (5.41 %) que en los datos de toda la Comunitat Valenciana.

Además, encontramos diferencias significativas en bastantes departamentos con respecto de la edad media de los pacientes que sufren de SCA. Es destacable nuevamente el caso de Requena, con una edad media bastante mayor tanto en general (71.76 años) como en ambos sexos (69.84 años los hombres y 75.46 años las mujeres), que se puede deber, como en la mayoría de los departamentos que encontramos con una media de edad mayor, a la mayor media de edad de los habitantes de los municipios pequeños, ya que los más jóvenes tienden a marcharse de los pueblos. Por último, encontramos diferencias significativas en cuanto a porcentajes de hombres y mujeres con respecto a los porcentajes de referencia en apenas 8 departamentos.

IV.2.2 Datos cruzados con grupos de edad

En el siguiente apartado, dividiremos los datos por grupos de edad y observaremos cómo varían las diferentes variables que tenemos en función del grupo en el que nos encontremos. Hemos dividido los grupos de edad en: menores de 30 años, mayores de 100 años, y grupos divididos cada 10 años entre los 30 y los 100. Todas las operaciones relativas a este apartado se encuentran en el notebook *4_c_Estudio_por_grupos_de_edad.ipynb*. El número de episodios que encontramos por cada grupo es el siguiente:

Grupo de edad	Episodios	Episodios acumulados	Porcentaje	Porcentaje acumulado
70-80	16321	16321	27.05	27.05
60-70	14094	30415	23.36	50.41
80-90	12106	42521	20.06	70.47
50-60	10242	52763	16.98	87.45
40-50	4353	57116	7.21	94.66
90-100	2010	59126	3.34	98.00
30-40	804	59930	1.33	99.33
Menor de 30	248	60178	0.41	99.74
Mayor de 100	157	60335	0.26	100.00

Cuadro 9: Episodios acumulados por grupo de edad.

IV.2.2.1 Un ejemplo: los pacientes menores de 30 años

Al igual que cuando agrupamos los datos por departamento de salud, en este caso presentaremos los resultados obtenidos para un único grupo, porque las operaciones realizadas para los demás serán análogas. Posteriormente, presentaremos todos los resultados en una tabla común. En este caso, los presentaremos para el grupo de los menores de 30, al ser el primero con el que hemos tratado en el notebook. Realizamos nuevamente el filtrado del grupo con la función *filtrar_grupo*.

La primera operación realizada es análoga a la que hemos visto en el caso anterior sobre las proporciones de hombres y mujeres en el grupo. En el histograma 2, en el apartado en el que mirábamos las variables de manera general, hemos visto que, en toda la comunidad valenciana, hubo más casos de hombres que de mujeres afectados por el SCA entre 2016 y 2019, pero los episodios no parecen acumularse en los mismos grupos de edad. Es por ello que observar la evolución de las proporciones conforme pasan los años puede ser muy interesante. En este caso, en lugar de comparar las edades a las proporciones de referencia obtenidas en la Comunitat Valenciana, queremos ver si la proporción de hombres y mujeres no sigue una distribución uniforme para algún grupo, es decir, si podemos afirmar significativamente que en cierto grupo de edad hay más individuos de un sexo que del otro. Por ello, planteamos el siguiente contraste de hipótesis:

Si P_{Hombre} y P_{Mujer} representan, respectivamente, las proporciones de hombres y mujeres en la muestra:

H_0: $P_{Hombre} = 50$ y $P_{Mujer} = 50$.

H_a: $P_{Hombre} \neq 50$ y $P_{Mujer} \neq 50$.

Para resolver este contraste, igual que en los casos anteriores, realizamos un test χ^2, cuyo código está incluido en la función *comparar_proporciones_entre_si*, que está en el archivo *helper_functions.py*. En este caso, tenemos que el 58.19 % (135 casos) de los episodios de SCA (135 casos) (135 casos) en menores de 30 años han sido sufridos por hombres, mientras que el otro 41.81 % (97 casos) han sido sufridos por mujeres. El p-valor obtenido con el test χ^2 es 0.01, que es menor que el nivel de significación $\alpha = 0{,}05$. Por lo tanto, podemos afirmar que, en este grupo, el porcentaje de individuos según el sexo no sigue una distribución uniforme. En particular, para este grupo existe un mayor porcentaje de hombres que de mujeres que han sufrido de un SCA.

A partir de aquí, seguiremos un guión análogo al seguido en el notebook anterior, cuando agrupamos los datos por departamentos. El siguiente paso consistirá en ver si la proporción de pacientes que entran al sistema sanitario mediante el servicio UAP es la misma en los menores de 30 años que en toda la Comunitat. En el caso general hemos visto que aproximadamente el 20.23 % de los episodios entró al sistema sanitario mediante el servicio de atención primaria, mientras que entre los menores de 30, este porcentaje se eleva hasta el 55.24 %, con 137 episodios iniciados en Atención Primaria. Realizamos el siguiente contraste de hipótesis:

Si P_{UAP} y P_{No_UAP} representan, respectivamente, las proporciones de individuos que han entrado y no han entrado por UAP al sistema sanitario en la muestra, realizaremos el siguiente contraste de hipótesis:

H_0: $P_{UAP} = 20{,}23$ y $P_{No\_UAP} = 79{,}77$.

H_a: $P_{UAP} \neq 20{,}23$ y $P_{No\_UAP} \neq 79{,}77$.

Igual que en el caso anterior, lo resolvemos utilizando el test χ^2 con la función

contrastar_proporciones_UAP, el cual nos da como resultado un p-valor del orden de 10^{-69}, muy inferior al nivel de significación $\alpha = 0{,}05$. Por lo tanto, podemos afirmar que en el grupo de los menores de 30, la distribución de los episodios que inician en UAP es diferente a la que habíamos obtenido en los datos generales de la Comunitat Valenciana.

Posteriormente, quedándonos en el tema de los recorridos, pasamos a contrastar el porcentaje de episodios que no han pasado por el servicio de hospitalizaciones (CMBD). Queremos comprobar si la edad puede tener algún efecto en esto, tratando de una manera u otra a los pacientes en función de ella.

Alrededor del 79.99 % de los casos habían sido derivados en algún momento al servicio CMBD en el total de la Comunitat Valenciana. Si miramos los datos de los menores de 30 años, podemos ver que solamente alrededor del 35.99 % de los episodios han sido derivados a CMBD en algún momento, mientras que el 64.92 % no lo han sido nunca (en contraste con el 20.01 % del caso de toda la Comunitat). Planteamos el siguiente contraste de hipótesis:

Si P_{CMBD} y P_{No_CMBD} representan, respectivamente, las proporciones de individuos que han pasado y no han pasado por CMBD en la muestra, realizaremos el siguiente contraste de hipótesis:

H_0: $P_{CMBD} = 79{,}99$ y $P_{No\_CMBD} = 20{,}01$.

H_a: $P_{CMBD} \neq 79{,}99$ y $P_{No\_CMBD} \neq 20{,}01$.

Para resolver este contraste, realizamos un test χ^2 una vez más con la función *contrastar_proporciones_CMBD*. Obtenemos un p-valor del orden de 10^{-69}, que evidencia que la gran diferencia que observábamos a simple vista es significativa. En este caso, se puede deber a una infravaloración del riesgo que puede tener un episodio de SCA en una persona tan joven. De hecho, con la función *porcentajes_entradas*, podemos comprobar que únicamente un episodio de los 248 que encontramos en este grupo de edad pasó por el servicio de hospitalizaciones, lo cual supone un paupérrimo 0.40 %.

Por último, comprobaremos nuevamente cómo es la proporción de casos de Éxitus en este grupo de edad en comparación al caso de toda la Comunitat. En este grupo encontramos un único caso de éxitus, que supone un 0.40 % del total, mientras que en la Comunitat Valenciana en general este porcentaje es de 3.8 %. Planteamos el siguiente contraste de hipótesis:

Si P_{Ex} y P_{No_Ex} representan, respectivamente, las proporciones de individuos que han finalizado y no han finalizado en éxitus la muestra, realizaremos el siguiente contraste de hipótesis:

H_0: $P_{Ex} = 3{,}80$ y $P_{No\_Ex} = 96{,}20$.

H_a: $P_{Ex} \neq 3{,}80$ y $P_{No\_Ex} \neq 96{,}20$.

Tal y como en los casos anteriores, realizamos un test χ^2 otra vez con la función *contrastar_proporciones_Exitus*, que nos devuelve un p-valor de 0.006. Este p-valor

es inferior al nivel de significación $\alpha = 0{,}05$, y por lo tanto podemos afirmar que la diferencia con las proporciones que estamos tomando como referencias es significativa. En este caso, era esperable obtener un porcentaje menor de casos de éxitus, puesto que la salud de las personas más jóvenes tiende a estar mejor conservada que la de los más longevos.

IV.2.2.2 Resultados en todos los grupos de edad

Vamos a presentar ahora los resultados obtenidos en todos los casos. Al igual que en la división por departamentos, hemos juntado todos en una única tabla, que es la siguiente:

Edad	**Sexo dominante**	$P_{UAP} = 20{,}23$ **y** $P_{No\_UAP} = 79{,}77$	$P_{CMBD} = 79{,}99$ **y** $P_{No\_CMBD} = 20{,}01$	$P_{Ex} = 3{,}80$ **y** $P_{No\_Ex} = 96{,}20$
Menor de 30	Hombre	Más UAP	Menos CMBD	Menos éxitus
30-40	Hombre	Más UAP	Menos CMBD	Menos éxitus
40-50	Hombre	Más UAP	Igual	Menos éxitus
50-60	Hombre	Más UAP	Más CMBD	Menos éxitus
60-70	Hombre	Menos UAP	Más CMBD	Menos éxitus
70-80	Hombre	Menos UAP	Más CMBD	Igual
80-90	Hombre	Más UAP	Menos CMBD	Más éxitus
90-100	Mujer	Más UAP	Menos CMBD	Más éxitus
Mayor de 100	Igual	Más UAP	Menos CMBD	Más éxitus

De los cálculos realizados en este notebook, lo primero que llama la atención es algo más bien esperable: conforme más avanza la edad de los pacientes, parece existir un mayor riesgo de éxitus. Esto lo puede confirmar también la siguiente gráfica, relativa a la evolución del porcentaje de los casos de éxitus con respecto a la edad:

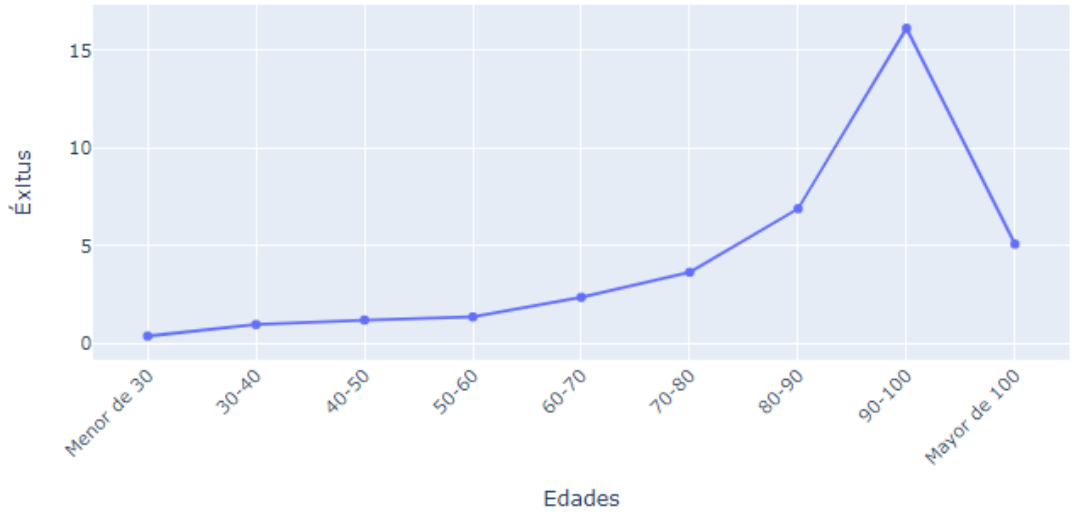

Figura 3: Evolución del porcentaje de casos de éxitus por grupo de edad.

Como podemos ver, el porcentaje aumenta a cada grupo de edad, a excepción del grupo de los mayores de 100, donde existe una gran caída. Esta tendencia se debe al deterioro de la salud que acompaña al paso de los años. Es por ello que es importante hacer tomar conciencia a la población más longeva de lo que significa tener un episodio infarto y lo que puede hacer para prevenirlo o para tratarlo de la manera más exitosa posible. Es sabido, como mencionamos en la introducción, que la detección precoz de

estos episodios es clave para su intervención exitosa, por lo tanto es importante que este grupo de edad, cuya salud es más vulnerable, aprenda a identificarlo y a actuar en consecuencia en caso de que verse víctimas de uno.

Por otra parte, es importante reseñar la relación que tenemos entre la edad y el sexo. Como podemos ver en la tabla, encontramos una proporción significativamente mayor de hombres que de mujeres hasta los 90 años, y a partir de los 100 la diferencia no es estadísticamente significativa. No obstante, veamos en detalle la diferencia en los porcentajes de episodios de hombres y mujeres que nos encontramos con respecto al paso del tiempo:

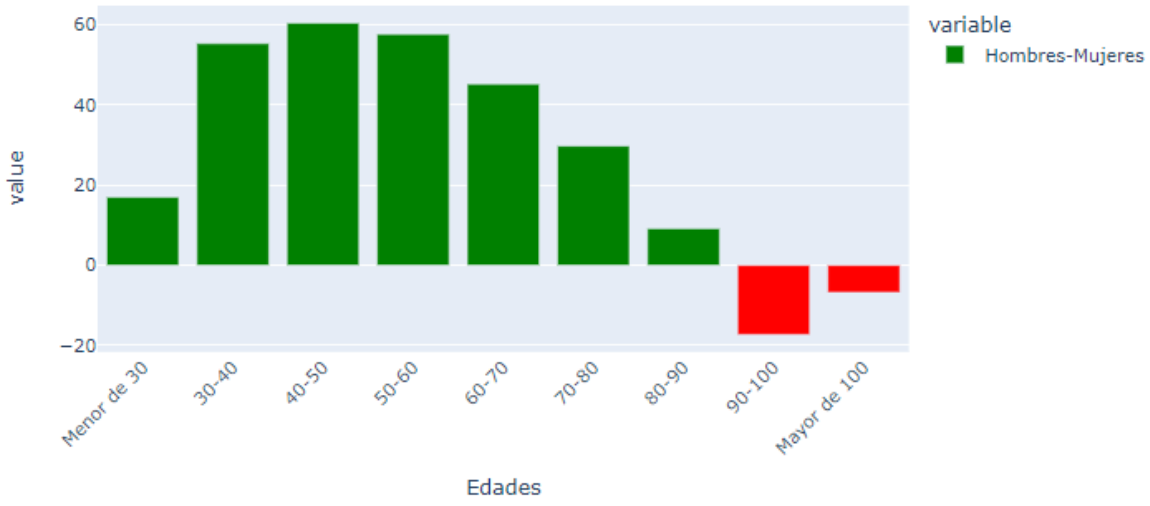

Figura 4: Evolución de la diferencia de porcentaje de hombres y mujeres por grupo de edad.

En los primeros grupos, podemos ver que la diferencia entre hombres y mujeres se tiende a acrecentar. Esto se debe a que el sexo masculino alcanza el punto de máxima concentración de casos de infarto a una edad más temprana, como vimos en la gráfica general. Conforme pasa el tiempo, las mujeres son las que empiezan a presentar un mayor riesgo de infarto, llegando incluso a superar en número de casos a los hombres a partir de los 90 años.

En otro orden de cosas, podemos echar un ojo también a una gráfica similar, pero respectiva a los porcentajes de episodios que comienzan en atención primaria, y los episodios que pasan por el servicio de hospitalización:

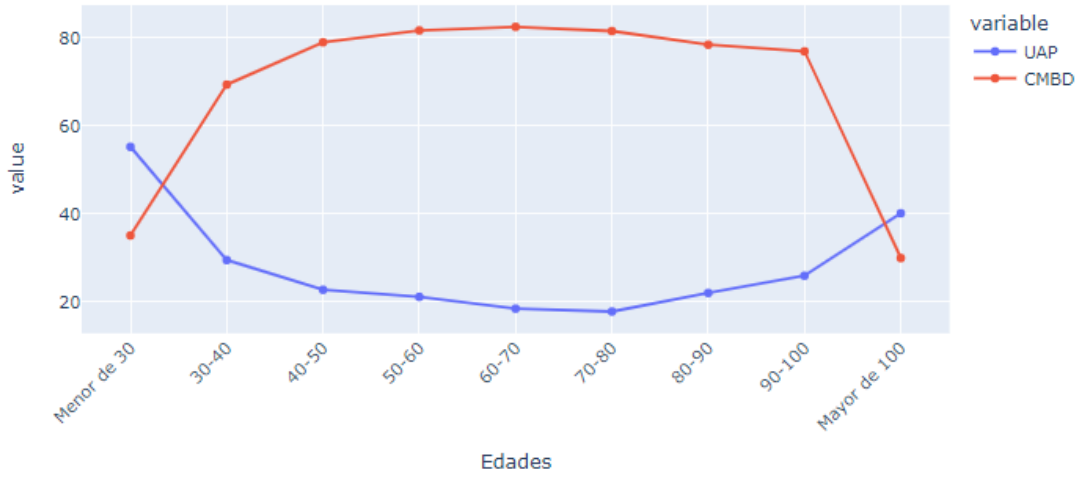

Figura 5: Evolución del porcentaje de casos de entrada UAP y de CMBD en función del grupo de edad.

En esta gráfica podemos ver como las tendencias de episodios que se inician en atención primaria y que son hospitalizados son inversamente similares, y tiene motivos similares, que son errores frecuentes tanto por parte del sistema sanitario como por parte de los pacientes.

En primera instancia, en los casos más jóvenes (hasta los 40 años) podemos observar claramente que los pacientes entran mucho más al sistema sanitario por UAP, y son menos veces enviados al servicio de hospitalización. Por otra parte, en este mismo segmento podemos ver que los pacientes también son enviados al servicio de hospitalización en muchas menos ocasiones. En los casos entre los 40 y los 60 años, los individuos siguen sin acudir a atención primaria tanto como en los números promedio de toda la Comunitat, pero sin embargo sí que son enviados igual o más veces al servicio de hospitalización. En los siguientes grupos, de los 60 a los 80 años, el porcentaje de personas que entran por UAP es significativamente menor al resto del conjunto de datos, mientras que el porcentaje de personas que son enviadas a CMBD es mayor. Por último, a partir de los 80 años, los pacientes vuelven a entrar al sistema más por el servicio de atención primaria, mientras que son menos hospitalizados de nuevo. Deberá ser la Conselleria de Sanitat quien elabore su investigación interna para observar qué está ocurriendo con los protocolos en cada tramo de edades, tanto para encontrar motivos sociológicos para explicar por qué los pacientes entran más o menos veces por atención primaria, como motivos de organización para mirar por qué los pacientes son enviados más o menos veces al servicio hospitalario.

IV.2.3 Datos acumulados por sexo

En tercer lugar, acumularemos los datos por sexo para cruzar esta variable con aquellas que no hemos cruzado todavía (es decir, entradas, hospitalizaciones y éxitus). En este caso, al ser únicamente dos posibilidades, hablaremos de todas las operaciones a la vez. Veamos por partes los cálculos que hemos realizado. Todas ellas se pueden encontrar en el notebook *4_d_Estudio_Sexos.ipynb*.

En primer lugar, veremos una gráfica con las proporciones observadas por sexos en estas tres variables:

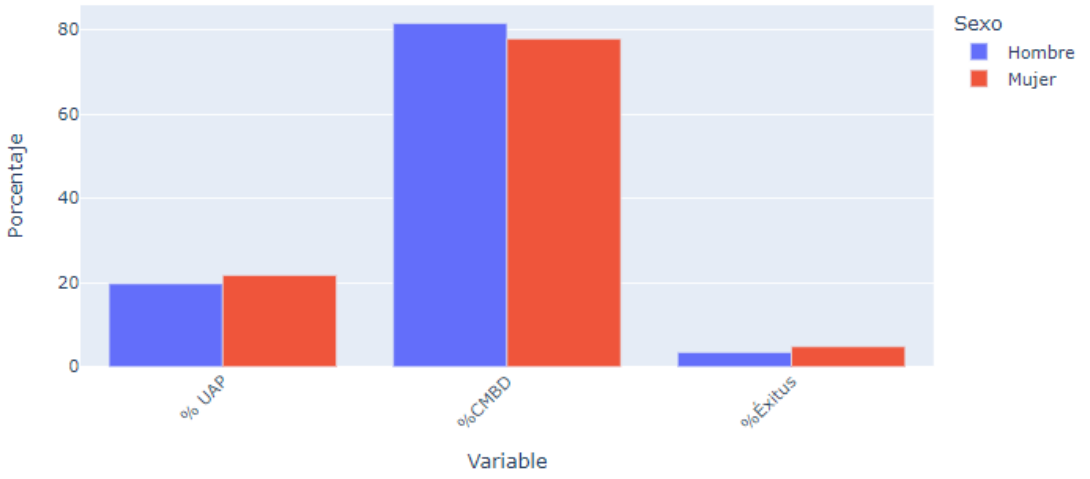

Figura 6: Diferencias porcentuales en diversas variables en función del sexo.

Como podemos observar, se ven unas leves diferencias entre sexos dependiendo de las variables: el porcentaje de individuos que entran por UAP y el porcentaje de casos de éxitus es mayor en las mujeres, mientras que el porcentaje de casos hospitalizados es mayor en los hombres.

Queremos saber si podemos asumir que esta diferencia entre proporciones es estadísticamente significativa. Para ello, plantearemos el siguiente contraste de hipótesis, igual para los tres casos:

Si P_H y P_M representan, respectivamente, las proporciones de hombres y mujeres estudiadas, realizaremos el siguiente contraste de hipótesis:

H_0: $P_M = P_H$.

H_a: $P_M \geq P_H$ (o H_a: $P_M \leq P_H$ en el caso del porcentaje de CMBD).

Hemos resuelto todos estos contrastes de hipótesis realizando un z-test que se puede encontrar en la función *contrastar_prop_mayor_menor*, en el archivo *helper_functions.py*. Podemos realizar este test porque tenemos que el total de casos es superior a 30 para ambos sexos, y los valores dentro de la variable sexo son independientes.

Como resultado, en todos los casos hemos obtenido un p-valor inferior a 0. Podemos entonces afirmar que las diferencias entre las proporciones que hemos visto son estadísticamente significativas. Por lo tanto, podemos extraer una serie de conclusiones. En primer lugar, vemos que las mujeres (21.70 %) tienden a entrar más por atención primaria que los hombres (19.70 %). Además, también las mujeres son enviadas en menor proporción (77.87 %) al servicio de CMBD que los hombres (81.56 %). Estos números encajan con el alto número de mujeres de más avanzada edad que sufren infartos, puesto que como hemos visto en el punto anterior, las personas con edades superiores a 80 años encajan también con estos dos puntos.

Por último, las mujeres presentar una mayor proporción de altas por éxitus (4.75 %) que los hombres (3.39 %). Diversos estudios afirman que, proporcionalmente, las mujeres suelen morir más por infartos que los hombres en general (véanse los estudios

[6, 4, 5]). Estos datos nos reflejan que esto no solo sucede en los casos generales, sino que en los episodios en los que se acude al sistema sanitario para solicitar ayuda esto también ocurre.

V Conclusiones y temas de investigación futuros

Este trabajo ha sido un largo proceso en el que hemos tenido que lidiar con varias cosas. En primer lugar, tuvimos que realizar una solicitud de datos públicos, con todas las restricciones que hemos presentado en la primera sección. Hemos visto que no es un proceso sencillo, pues requiere de una información muy detallada de la utilidad que se les va a dar, de las personas implicadas, y requiere de pasar por diversos filtros que alargan mucho el proceso hasta obtenerlos.

Una vez obtenidos los datos, hemos podido ver que los datos manejados por la Conselleria de Sanitat no son ni mucho menos sencillos de tratar. Existen diversos puntos de mejora posibles para ellos, que veremos a continuación:

- Los datos provienen de cuatro fuentes diferentes y hay que unificarlos antes de poder realizar cualquier estudio, lo cual supone un tedioso proceso.
- Cada uno de los cuatro servicios sanitarios con los que hemos contado para este trabajo tiene su propia fuente de datos. No únicamente esto, sino que cada una de las fuentes de datos los toma de una manera diferente: los nombres de las variables pueden no ser iguales para dos variables que representan lo mismo. (ejemplo: *SIP* y *NUM_SIP*).
- Los valores no son uniformes para las mismas variables en diferentes fuentes, sobretodo en los strings (ejemplo: los nombres de departamentos están en algunas fuentes en castellano y en otras en valenciano).
- El proceso de extracción de datos no ha sido del todo correcto, faltando algunos datos provenientes de la fuente de emergencias.

Realizar cambios en base a estas observaciones facilitaría no solo los trabajos de investigación, sino también el uso aplicado de estos para la propia Conselleria de Sanitat, que podría emplearlos de manera mucho más sencilla en su propio beneficio.

Cuando hemos conseguido un objeto de datos mucho más manejable, hemos conseguido realizar un estudio de las variables más relevantes relacionadas con los episodios de infarto sucedidos en la Comunitat Valenciana entre los años 2016 y 2019. De aquí, hemos podido extraer una serie de conclusiones relevantes:

- Existen muchos episodios finalizados con un tipo de alta que no da pie a la finalización, como pueden ser *Urgencias* o *Traslado al hospital de agudos*.
- Se ha observado una mayor mortalidad dentro del sistema sanitario en las personas de los grupos de edad más longevos.
- Existe un porcentaje mayor de casos de SCA en hombres, pero la mortalidad dentro del sistema sanitario es mayor en las mujeres.
- Los hombres que sufren de SCA son más jóvenes que las mujeres.

- Existe un alto porcentaje (20.01 %) de pacientes que entra al sistema sanitario por la vía menos eficiente para el tratamiento de esta enfermedad, que es la vía de atención primaria. Además, existe una gran variabilidad en este porcentaje en función de:
  - El departamento de atención. En los departamentos con mayor población rural, la vía de acceso de atención primaria suele ser más utilizada que en aquellos centrados en las grandes urbes de la Comunitat.
  - La edad. Tanto los grupos de edad más jóvenes como los más longevos presentan un mayor porcentaje de entradas por la vía de atención primaria.
  - El sexo. Las mujeres con SCA acuden al sistema sanitario en mayor porcentaje que los hombres.
- Hay un porcentaje similar de episodios de SCA que nunca pasan por el servicio de hospitalización (20.23 %), y por lo tanto incumplen el protocolo de infartos de la Generalitat Valenciana. Además, existe una gran variabilidad en este porcentaje en función de:
  - El departamento de atención. Nuevamente en los departamentos con mayor población rural, los episodios de infarto son derivados en menor porcentaje al sistema hospitalario.
  - La edad. Una vez más, tanto los grupos de edad más jóvenes como los más longevos presentan un menor porcentaje de paso por el sistema de hospitalizaciones.
  - El sexo. Las mujeres con SCA son hospitalizadas en un menor porcentaje que los hombres.

Por lo tanto, podemos afirmar que existen diversos puntos en los que se puede centrar la atención para mejorar el sistema sanitario, en particular, en la mejora de los protocolos llevados a cabo tanto por profesionales como por pacientes.

Como punto final, es importante recalcar que este trabajo está inmerso en una investigación junto con la Conselleria de Sanitat, en la que todavía hay bastantes vías de exploración de donde se pueden extraer muchas más conclusiones. Tenemos muchas variables que todavía no hemos explorado, como pueden ser la cronicidad de los pacientes, los protocolos aplicados sobre ellos, o los diagnósticos secundarios. Además, también se pueden aplicar algoritmos de predicción para ayudar a predecir patrones de comportamiento, tanto de pacientes como de sanitarios, para ayudar a encontrar los puntos flacos del sistema y poderlos mejorar de manera eficiente.

Referencias

[1] Plan código infarto de la Generalitat Valenciana. URL https://www.san.gva.es/documents/d/assistencia-sanitaria/plan_codigo_infarto_cv.

[2] Framingham Heart Study. Three generations of research on heart disease. URL https://www.framinghamheartstudy.org/.

[3] Comité Ético de Investigación de Salud Pública FISABIO. URL https://fisabio.san.gva.es/es/investigacion/comite-etico-de-investigacion-de-salud-publica/.

[4] I. R. Dégano, R. Elosua, and J. Marrugat. Epidemiología del síndrome coronario agudo en españa: estimación del número de casos y la tendencia de 2005 a 2049. *Revista Española de Cardiología*, Junio 2013. URL https://www.revespcardiol.org/es-epidemiologia-del-sindrome-coronario-agudo-articulo-S0300893213001267.

[5] I. Ferreira-González. Epidemiología de la enfermedad coronaria. *Revista Española de Cardiología*, Febrero 2014. URL https://www.revespcardiol.org/es-epidemiologia-enfermedad-coronaria-articulo-S0300893213004855.

[6] A. Sambola, F. J. Elola, J. L. Ferreiro, N. Murga, L. Rodríguez-Padial, C. Fernández, H. Bueno, J. L. Bernal, Ángel Cequier, F. Marín, and M. Anguita. Impacto de las diferencias de sexo y los sistemas de red en la mortalidad hospitalaria de pacientes con infarto agudo de miocardio con elevación del segmento st. *Revista Española de Cardiología*, Noviembre 2021. URL https://www.revespcardiol.org/es-impacto-de-las-diferencias-de-sexo-y-los-articulo-S0300893220304723.