El objetivo de la práctica 5 es intentar explicar la mayor parte información sobre las comunidades autónomas ( lo que se especifica en el archivo comaut.sav (<https://www.uv.es/mlejarza/datamine/comaut.sav>) ) con un número muy inferior de factores explicativos de **fácil interpretación** y capaces de dar cuenta de la **mayor parte** de la variabilidad de todas variables.

También se pretende utilizar la información de estos **factores** explicativos para **agrupar** las comunidades autónomas en grupos homogéneos según su comportamiento en ellos.

Para llevar el cabo objetivo el proceso de análisis incluirá:

1. Ensayar la obtención de las componentes principales.
	* **Decidir** cuántos factores pueden o deben considerarse en función de su explicabilidad / interpretabilidad y de su capacidad para dar cuenta de la variabilidad de las variables originales en su conjunto y particularizadamente.
* **Discutir** la conveniencia o no de eliminar alguna variable del análisis, o de generar algún indicador nuevo.
* **Ensayar** una solución rotada ( varimax) y **decidir** si es preferible o no.
* **Guardar** la solución elegida

2. Probar una solución alternativa por Máxima Verosimilitud (ML). ( Hay que tener en cuenta que la matriz de correlación de la totalidad de las variables presenta una muy elevada multicolinealidad lo que puede producir resultados numéricamente demasiado cercanos a cero en el cálculo de los determinantes que den un mensaje de error . Esto ha ocurrido al usar SPSS , quizá ocurra también con R. Una posible forma de solución es no considerar la totalidad de las variables. Por ejemplo : la opción de obtener 3 factores ML del conjunto de variables : densidad, Ventasaotras, Comprasdeotras, autoabas, ratioVentasCompras, Gastoporpersona ,Gastoporunidadconsumo, pibpercap, tempIndus , tempServ, tbuscandoempleo ; funcinó bien, sin ningún problema ) .

* + **Decidir** cuántos factores pueden o deben considerarse en función de su explicabilidad / interpretabilidad y de su capacidad para dar cuenta de la variabilidad de las variables originales en su conjunto y particularizadamente.
* **Discutir** la conveniencia o no de eliminar alguna variable del análisis, o de generar algún indicador nuevo.
* **Ensayar** una solución rotada ( varimax) y **decidir** si es preferible o no.
* **Comparar** la solución elegida con la elegida en 1 ( CP o CP-rotada) y **optar** por una de ellas
* **Guardar** las soluciones

3. Aplicar el método de ejes principales ( éste sí se ha podido comprobar que funciona bien con la totalidad de las variables originales usando R )

* + **Decidir** cuántos factores pueden o deben considerarse en función de su explicabilidad / interpretabilidad y de su capacidad para dar cuenta de la variabilidad de las variables originales en su conjunto y particularizadamente.
* **Discutir** la conveniencia o no de eliminar alguna variable del análisis, o de generar algún indicador nuevo.
* **Ensaya**r una solución rotada ( varimax) y **decidir** si es preferible o no.
* **Comparar** la solución con las obtenidas en 1 y en 2 ( CP y ML ) y **optar** por una de ellas
* **Guardar** las soluciones

4. **Comparar CRÍTICAMENTE** las soluciones obtenidas ( originales y rotadas ) y **optar por una** de ellas **justificadamente.**

5. Emplear esta solución elegida para llevar a cabo una **clusterización jerárquica** por el método de Ward y obtener JUSTIFICADAMENTE una **solución de agrupación** de comunidades autónomas que se considere suficientemente adecuada en términos de tamaño y homegeneidad interna. Listar la agrupación y comentar los resultados de agrupación de las distintas comunidades autónomas.

6. Aplicar los oportunos análisis de la varianza sobre **factores** ( que se entiende han sido ya interpretados y adecuadamente NOMBRADOS ) y sobre aquellas **variables originales** que se piense que pueden resultar más fecundas ( Es de utilidad aquí repasar los resultados de anova ya llevados a cabo en la práctica 3 )

7. Con todos los análisis obtenidos ( especialmente los puntos 4,5 y 6**) discutir** las diferencias entre la agrupación obtenida aquí y los resultados finales de la práctica 3.

8. Es interesante considerar que la utilización en el clustering de la distancia de Mahalanobis equivale , como se ha comentado muchas veces , a aplicar el mismo método de clustering a la **totalidad de las componentes principales con la distancia Euclidea**. **No es lo que se ha hecho aquí.** ¿Es algo parecido? ¿cabe esperar resultados muy distintos? ¿por qué? ( Puede hacerse el clustering con todas las CP para comprobarlo **pero no es necesario si se argumenta** adecuadamente)

9. Si a lo largo de los distintos puntos anteriores no han sido ya contestadas, aprovechar las conclusiones finales para contestar las siguientes cuestiones.

¿cuáles son los principales factores del diferente comportamiento socio-economico-comercial de las distintas comunidades autónomas?¿cuál es el más importante de estos factores? ¿Cómo se cifra / cómo se valoraría su importancia? ¿Cómo quedan agrupadas las comunidades autónomas según su comportamiento es estos factores?¿ ¿Qué comunidades autónomas se comportan más parecidamente? ¿Existen comunidades autónomas de comportamiento SINGULAR?¿Cuáles son y en que consiste su singularidad? ¿qué grupos de comunidades autónomas son más similares? ¿Cuáles más diferentes ¿ Qué variables/ factores marcan mayores diferencias en las agrupaciones de Com. Autónomas? ¿Cuáles son los parecidos y diferencias entre la(s) agrupación(es) obtenida(s) en la práctica 3 y las obtenidas aquí?