GRADO EN ECONOMÍA

CURSO ACADÉMICO 2019/2020

TRABAJO FIN DE GRADO

Análisis estadístico multivariante por Comunidades Autónomas.

AUTOR/A

ANA GÁMIR DE LAS HERAS

DIRECTOR/A

CARMEN MARÍA SORDO GARCÍA

FECHA

23 / 09 / 2020

ÍNDICE

Análisis estadístico multivariante por Comunidades Autónomas.

Ana Gámir de las Heras.

Resumen

Este trabajo tiene como objetivo analizar las 19 Comunidades Autónomas que formas nuestro país, a través de unas variables escogidas, con el fin de presentar similitudes y diferencias entre las zonas. Aplicaremos diferentes técnicas de análisis de datos multivariante como el análisis factorial, análisis cluster y análisis de la varianza. Con ello extraeremos una serie de conclusiones donde detectaremos debilidades y fortalezas de cada comunidad, y propondremos posibles políticas económicas a aplicar en cada caso para revertir los problemas o por el contrario potenciar las virtudes.

Palabras clave: Análisis descriptivo, análisis factorial, análisis cluster, análisis anova, política económica.

1. INTRODUCCIÓN.

Nuestra investigación tiene el objetivo de estudiar las Comunidades Autónomas españolas, concretamente de conocer la situación económica y social de cada una de ellas. El planteamiento que seguiremos en el trabajo será el siguiente. En primer lugar, conocer en qué se parecen y en qué se diferencian las regiones de nuestro país, es decir, agruparemos las Comunidades en función de sus similitudes y diferencias. En segundo lugar, trataremos de detectar los problemas que existen en las zonas a estudiar, así como los puntos fuertes de las mismas, de tal forma que se propondrán medidas para mejorar las debilidades o reforzar las fortalezas.

Dados unos objetivos claros, desarrollaremos el trabajo de una forma uniforme y ordenada, es decir, aplicando la metodología de forma clara con el fin de analizar, explicar y aclarar todos los resultados.

Para realizar nuestro análisis manejaremos datos de corte transversal, concretamente para el año 2019. La principal fuente que hemos utilizado para seleccionar dichos datos de la investigación ha sido el INE (Instituto Nacional de Estadística) además de la página del Ministerio de Fomento.
De toda la información recopilada hemos escogido las siguientes variables:

1. Índice de precios al consumo.
2. Producto interior bruto per cápita.
3. Tasa de paro.
4. Tasa de actividad.
5. Ocupados.
6. Número de hipotecas.
7. Precio de la vivienda.
8. Condenados por violencia de género.

Tras haber hecho una exploración exhaustiva de datos, finalmente estos son los 8 seleccionados para llevar a cavo nuestra investigación. A partir de estos 8, se realizarán 19 observaciones en cada variable, una por cada comunidad autónoma española, de tal manera que podremos hacer una comparativa fiable para el año a estudiar, en este caso, el 2019.

Vamos a dividir el trabajo en 4 análisis claramente diferenciados entre sí de los cuales extraeremos una serie de conclusiones y deducciones acerca de nuestras 19 comunidades autónomas.

- Análisis descriptivo.
- Análisis factorial.
- Análisis cluster.
- Análisis anova.

En primer lugar, a través del análisis descriptivo, expondremos de forma gráfica las variables con el fin de ordenar la información y poder hacer una comparación que nos pueda adelantar los resultados finales.

En segundo lugar, con el análisis factorial, trataremos de reducir la dimensionalidad, de tal manera que, agruparemos las variables para buscar el mínimo número de dimensiones que sean capaces de explicar la máxima información contenida en los datos.

En tercer lugar, el análisis cluster agrupará las observaciones, en nuestro caso, las comunidades autónomas, utilizando como variables las que se obtuvieron en el análisis factorial. Este análisis se llevará a cabo con la intención de obtener una división por grupos de las comunidades autónomas que sean diferentes entre ellos y parecidos dentro de ellos con el fin de sacar unas conclusiones más fácilmente.

En cuarto y último lugar, comprobaremos si el análisis cluster realizado en el apartado anterior ha agrupado las comunidades de una manera eficiente o no. Para llevar a cabo este contraste, conocido como Anova, se realizará una comparativa entre las medias de los distintos grupos.

2. ANÁLISIS DE DATOS Y RESULTADOS.

2.1. Exposición de las variables.

La variable objeto de estudio sobre la que centraremos nuestro análisis va a ser la Comunidad Autónoma, en concreto analizaremos 19 casos dado que son el número en las que se divide nuestro país.
Esta variable será estudiada a partir de otras que explicaremos detalladamente a lo largo de este punto. El estudio se llevará a cabo mediante 4 tipos de análisis con el fin de concluir qué problemas existen en cada comunidad (o grupo de comunidades) y qué políticas o medidas económicas podrían llevarse a cabo para paliar dichos problemas.

Para poder comenzar el análisis, en primer lugar, se escogieron una gran cantidad de variables con el fin de estudiar al detalle nuestras observaciones (las 19 comunidades autónomas). De todas esas variables, se descartaron varias dado que no aportaban la información suficiente para aportar información y conseguir sacar unas conclusiones fiables. Por tanto, debido a que algunas de ellas podrían sesgar en exceso nuestros resultados se eliminaron siendo finalmente 8 las seleccionadas para aportar información sobre nuestras 19 observaciones. A continuación, detallamos cada una de las variables que se van a incluir en nuestro análisis:

1. Índice de precios al consumo (IPC): Medida estadística que recoge cómo han evolucionado los precios de los bienes y servicios que consume una población residente en viviendas familiares en España. Indica la tasa de inflación de una zona, con él analizaremos la competitividad y el poder adquisitivo.

2. Producto interior bruto per cápita (PIBpc): “PIB por habitante” Indicador económico que mide la relación existente entre el nivel de renta de un país y su población.

3. Tasa de paro (tparo): Mide el nivel de desocupados en relación con la población activa, es decir, la parte de la población en edad, condiciones y disposición de trabajar que no dispone de puesto de trabajo.

4. Tasa de actividad (tactiv): Índice que mide el nivel de empleo de un país.

5. Ocupados (ocup): Indica las personas de 16 o más años que tienen un trabajo por cuenta ajena o que ejercen una actividad por cuenta propia.

6. Número de hipotecas (hipot): Ofrece información sobre el número de hipotecas que se constituyen sobre los bienes inmuebles.

7. Precio de la vivienda (pviv): Mide la evolución de los precios de compraventa de las viviendas de precio libre, tanto nuevas como de segunda mano, a lo largo del tiempo.

8. Condenados por violencia de género (viogen): Indica el número de personas condenadas por violencia de género, obtenido a partir de la explotación estadística del Registro central.

Agregando la información obtenida de las variables podremos estudiar cada Comunidad Autónoma y además haremos comparativas entre ellas para poder obtener mejores resultados.

2.2. Análisis descriptivo.

Como hemos mencionado anteriormente, para llegar a elegir las 8 variables con las que finalmente vamos a hacer el estudio, estas han pasado por un proceso de selección con el fin de obtener los mejores resultados y que ninguna de ellas nos produjese sesgos a la hora de realizar los análisis.

En este epígrafe vamos a expresar mediante estadística descriptiva y gráficos el estado de las variables para cada zona.

La tabla 1 está compuesta por 4 estadísticos descriptivos aplicados a cada una de nuestras variables seleccionadas. Esta tabla está compuesta en primer lugar, por la columna en la que se encuentran las variables, en segundo por el número de observaciones que se van a estudiar, en este caso 19 (CCAA) para cada variable. En tercer lugar, el mínimo valor y en cuarto, el máximo. La quinta columna es la media de nuestra serie de datos y, por último, la desviación típica, que ofrece información sobre la dispersión media de las variables.

Tabla 1. Estadísticos descriptivos.

	N	Mínimo	Máximo	Media	Desv. Típica
Índice de precios al consumo	19	1,7	2,7	2,16842105	0,24287225
PIB per cápita	19	19073	35876	25361,2632	5235,09432
Tasa de paro	19	8,2	25,9	15,2421053	5,05704303
Tasa de actividad	19	50,9	63,3	57,8578947	3,35050838
Número de ocupados	19	29,3	3391	1024,79474	1066,6417
Número de hipotecas	19	416	92967	25926	29796,8122
Precio vivienda	19	105,76475	141,142	118,387184	9,68325638
Condenados violencia género	19	53	6110	1605	1628,76269

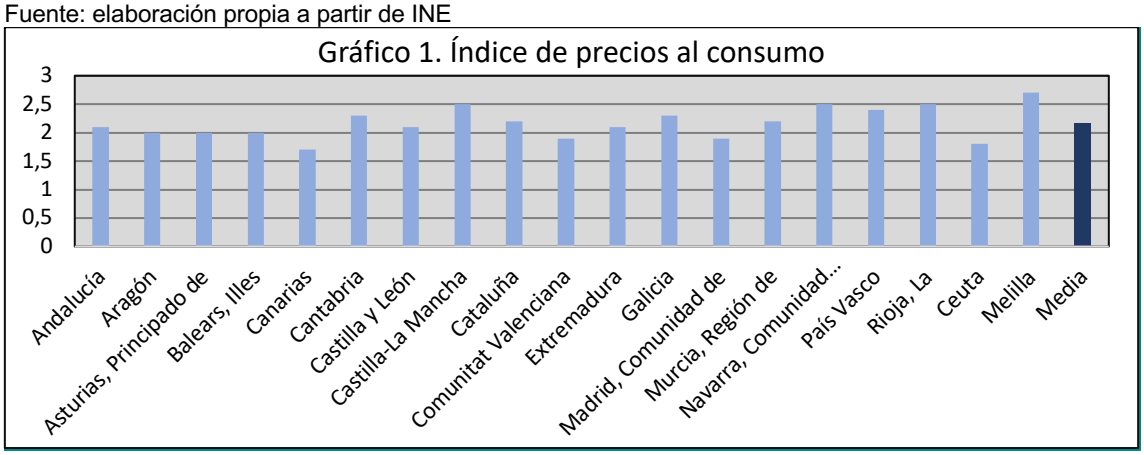
Fuente: elaboración propia a partir de INE

Analizando los resultados obtenidos en la tabla 1, podemos observar a primera vista las enormes diferencias que nos vamos a encontrar entre unas comunidades y otras dada la gran diferencia entre los mínimos y los máximos de todas las variables (aunque no en todas se da la misma desigualdad).

Como introdujimos, nuestro objetivo es realizar una comparativa de las Comunidades Autónomas entre ellas, por lo que, a continuación, vamos a desagregar la información por zonas.

En el gráfico 1, vemos reflejada la tasa de inflación de cada Comunidad Autónoma. No existe una diferencia muy notable entre ellas, pero podemos apreciar que Melilla, La Rioja, Navarra, País Vasco y Castilla la Mancha están por encima del resto, lo cual significa que han sufrido una pérdida de competitividad debido a la pérdida de poder adquisitivo.

Fuente: elaboración propia a partir de INE

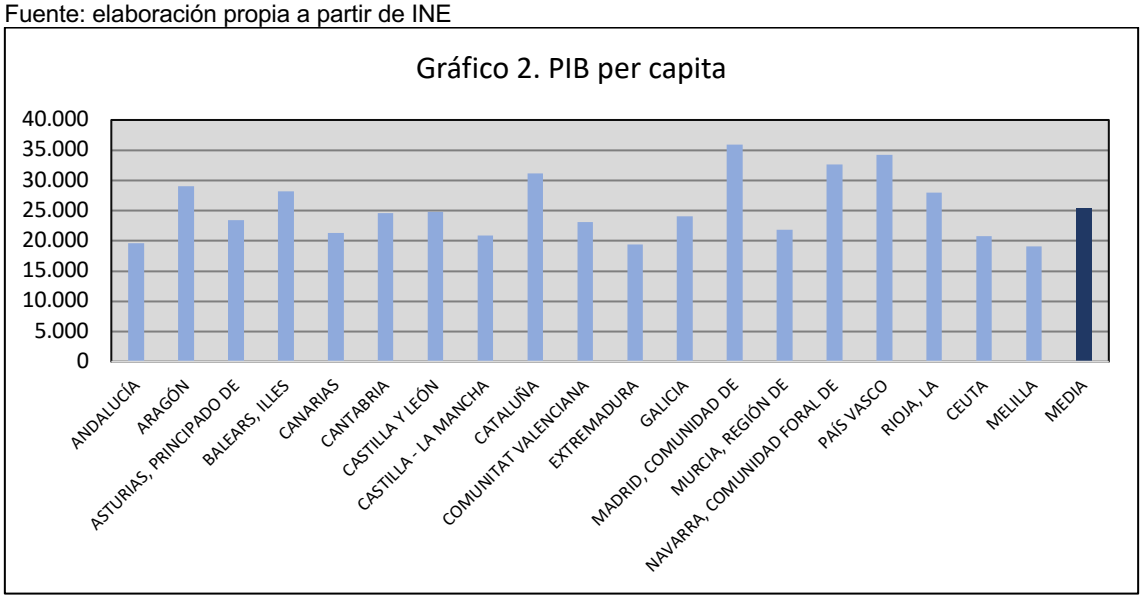


En el gráfico 2, vemos diferencias notables en términos de PIB per cápita, lo cual significa que existe una gran desigualdad económica dentro del territorio nacional.

Análisis estadístico multivariante por CCAA.

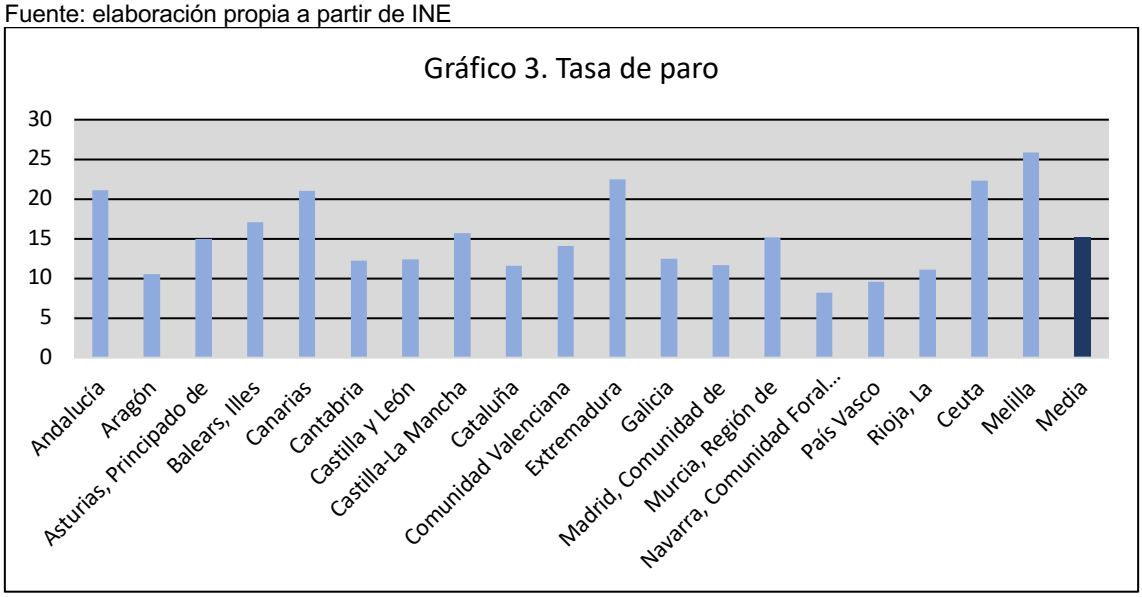
Madrid, País Vasco, Navarra y Cataluña tienen un índice elevado a diferencia de regiones como Melilla, Extremadura, Andalucía y Castilla la Mancha con valores inferiores a la media.

Fuente: elaboración propia a partir de INE



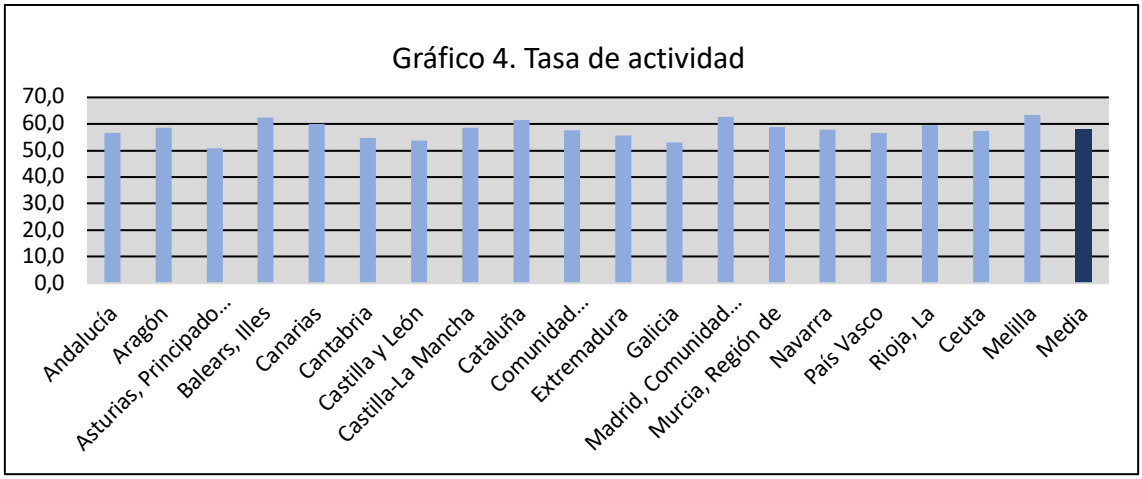
Con la información del gráfico 3, el cual nos indica la tasa de paro, podemos percibir las desigualdades existentes entre las comunidades autónomas en cuanto al mercado de trabajo. En una peor situación, es decir, con una tasa de paro superior a la media se encuentran Melilla, Extremadura y Ceuta. Por el contrario, en situación prácticamente de pleno empleo se sitúan Navarra, País Vasco y Aragón. Destacamos la elevada media española de la tasa de paro en el año a estudiar, 2019 (15,2%).

Fuente: elaboración propia a partir de INE



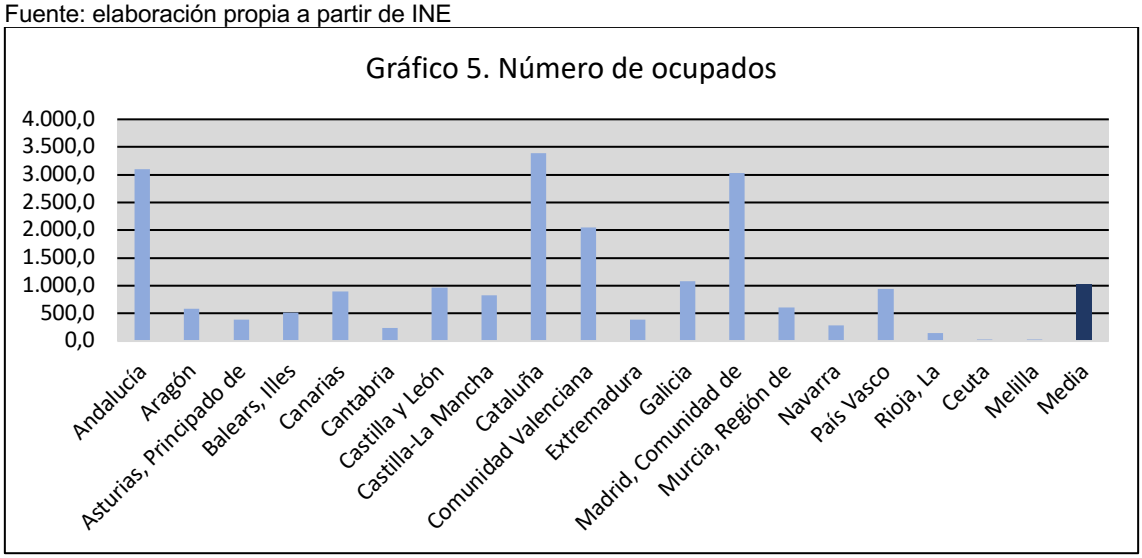
Al contrario de otras variables, la tasa de actividad expuesta en el gráfico 4 no presenta grandes diferencias entre unas observaciones y otras. Si que apreciamos que zonas con más actividad económica suelen tener tasa de actividad en el empleo superiores como es el caso de Madrid y Cataluña.

Fuente: elaboración propia a partir de INE

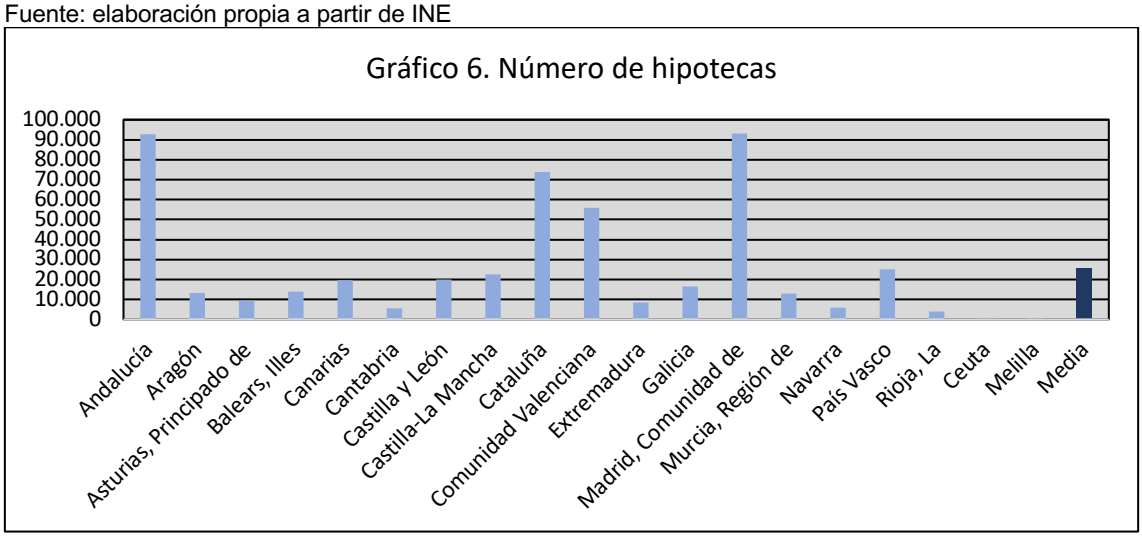


Si observamos simultáneamente el gráfico 5 (número de ocupados) y el gráfico 6 (número de hipotecas) apreciamos que en ambos existe una enorme desigualdad entre las 19 comunidades. Destacan muy por encima del resto y de la media española Madrid, Cataluña y Andalucía dado que ambas variables son directamente proporcionales al número de habitantes.

Fuente: elaboración propia a partir de INE



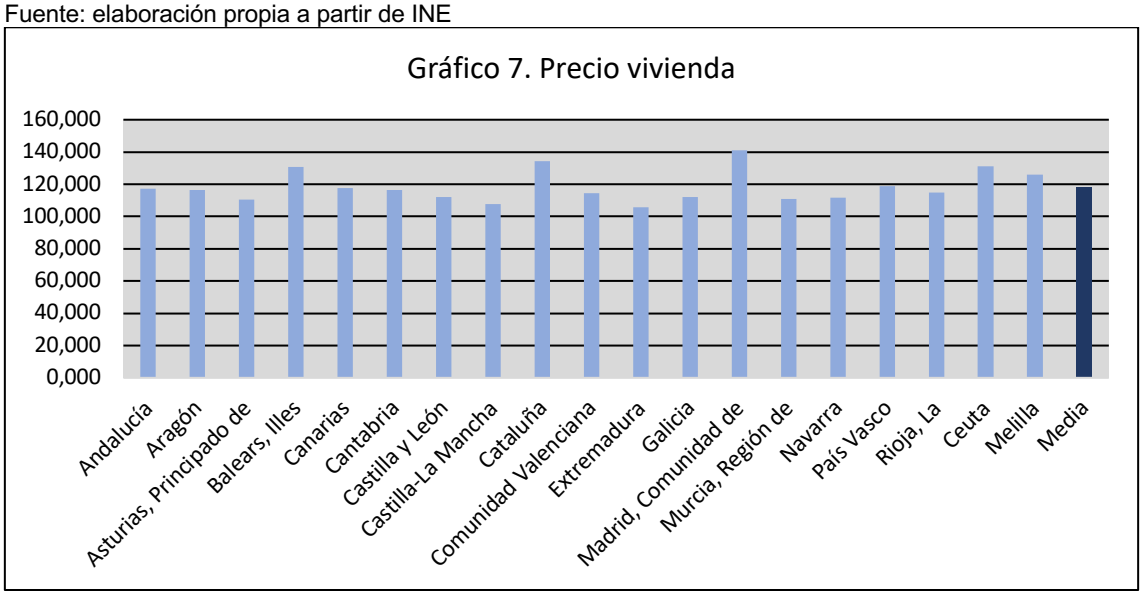
Fuente: elaboración propia a partir de INE



Análisis estadístico multivariante por CCAA.

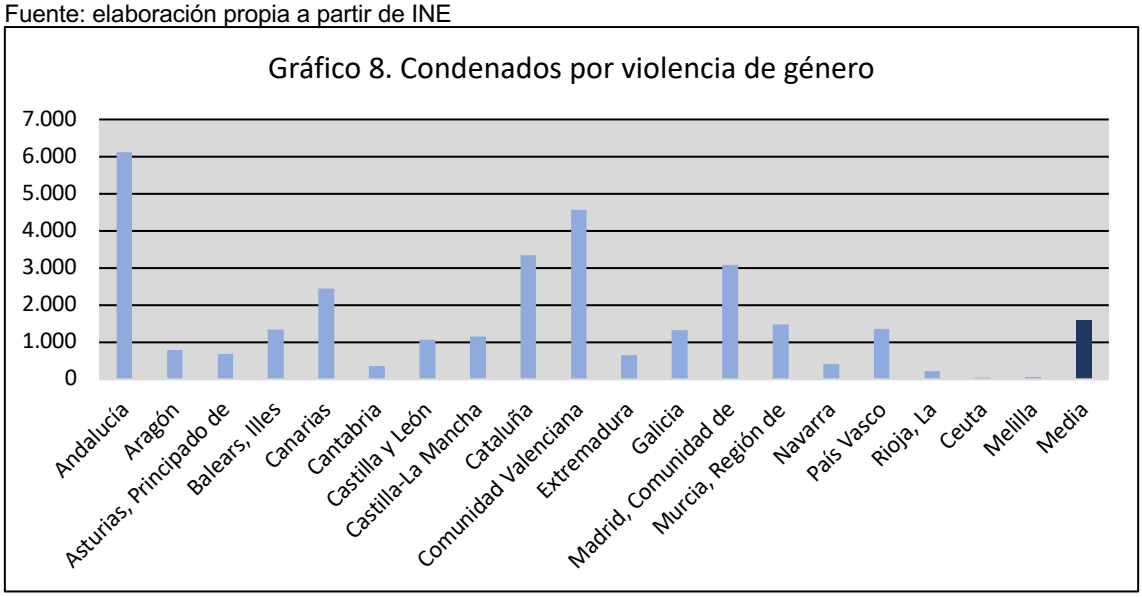
En el siguiente, el gráfico 7, observamos que no existen grandes diferencias en el precio de la vivienda, las 19 comunidades están en torno a la media, quedando por encima, Madrid, con el precio más elevado de todo el país, seguido de Cataluña, que es menor que el de la capital, pero sigue superando la media.

Fuente: elaboración propia a partir de INE



Para concluir la primera impresión sobre las diferencias y similitudes entre las comunidades autónomas, el gráfico 8 nos muestra los condenados en el año a estudiar, 2019, por violencia de género. Vemos claramente que superan la media española Andalucía, Comunidad Valenciana y Madrid y que apenas aportan datos Ceuta y Melilla. Debemos aclarar que el motivo por el que Ceuta y Melilla presentan valores prácticamente inapreciables en el gráfico es porque los datos están presentados en bruto y no como porcentaje de los habitantes, es decir, que Ceuta y Melilla presenten valores mínimos es debido a su escasa población, no porque no existan casos. En cambio, si que resulta alarmante el resultado de la Comunidad Valenciana, ya que el dato es muy elevado con respecto a sus habitantes y comparado con el resto de Comunidades.

Fuente: elaboración propia a partir de INE



Este análisis descriptivo nos ha servido para apreciar a primera vista las diferencias y similitudes más evidentes de las 19 comunidades autónomas. En cambio, para poder llevar a cabo nuestro análisis de manera más detallada aplicaremos técnicas de análisis estadístico multivariante, concretamente análisis factorial y análisis cluster.

2.3. Análisis factorial.

Con el análisis factorial, la intención que tenemos es agrupar las variables que nos repitan información, es decir, las variables en las que exista una fuerte incorrelación. De tal manera, que las 8 variables seleccionadas se satures en pocos factores, con los cuales podamos explicar el contenido, pero de una manera más clara y sintética.

A continuación, utilizaremos el conjunto de datos tfg_R, que muestra datos de 8 variables escogidas para las diferentes comunidades autónomas de España en el año 2019. (Los datos importados son de elaboración propia a partir de datos extraídos del INE).

En análisis factorial se encuentra implementado a través de la función de R factanal. Para comenzar aplicaremos el test formal para determinar el número de factores a considerar utilizando el algoritmo de máxima verosimilitud (method = "mle"). Probamos a utilizar de 1 a 3 factores:

```
> sapply(1:3, function(x)
+   factanal(tfg_R[,2:9], factors = x, method = "mle")$PVAL)
   objective    objective    objective
0.0002517117 0.0139024841 0.5039713720
```

Los resultados obtenidos sugieren que utilizar 3 factores es el número adecuado para recoger las varianzas en los datos. Tomando 3 factores no podemos rechazar la hipótesis nula de que 3 factores son suficientes dado que el p-valor > 0,05.

```
> factanal(tfg_R[,2:9],factors=3)# factors = x, method = "mle")

Call:
factanal(x = tfg_R[, 2:9], factors = 3)

Uniquenesses:
                IPC       PIB per capita         TASA DE PARO    TASA DE ACTIVIDAD
              0.870                0.085                0.005                0.484
          OCUPADOS NÚMERO DE HIPOTECAS      PRECIO VIVIENDA VIOLENCIA DE GÉNERO
              0.034                0.008                0.161                0.108

Loadings:
                   Factor1 Factor2 Factor3
IPC                -0.299   0.155  -0.130
PIB per capita              0.881   0.367
TASA DE PARO               -0.975   0.211
TASA DE ACTIVIDAD                   0.711
OCUPADOS            0.946   0.188   0.192
NÚMERO DE HIPOTECAS 0.958   0.149   0.229
PRECIO VIVIENDA     0.236   0.112   0.878
VIOLENCIA DE GÉNERO 0.943

               Factor1 Factor2 Factor3
SS loadings      2.858   1.826   1.562
Proportion Var   0.357   0.228   0.195
Cumulative Var   0.357   0.586   0.781

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 6.31 on 7 degrees of freedom.
The p-value is 0.504
```

Tras examinar los resultados obtenidos en el análisis factorial comprobamos por qué variables está formado cada factor:

- El Factor 1 (F1) está construido por el número de ocupados, el número de hipotecas y los condenados por violencia de género.

  Expresado como un índice → F1 (factor 1) = 0,946 ocup + 0,958 n_hipo + 0,943 Violen

  De esta ecuación podemos decir que todas las variables explicativas tienen una gran elasticidad sobre la variable explicada, siendo el nº de hipotecas la que mayor impacto tiene sobre el factor1.

- El Factor 2 (F2) lo forman el índice de precios al consumo, el PIB per cápita y la tasa de paro.

  Expresado como un índice → F2 (factor 2) = 0,155 IPC + 0,881 PIBpc -0,975 Tparo

  De esta ecuación podemos decir que el PIBpc es la variable que más peso tiene sobre el factor. Además, podemos apreciar que la tasa de paro tiene una relación inversa con el factor, es decir, que cuanto mayor es el paro, tanto más se verá perjudicado el factor2.

- El Factor 3 (F3) formado por la tasa de actividad y el precio de la vivienda.

  Expresado como un índice → F3 (factor 3) = 0,711 T_act + 0,878 P_vivienda

En esta ecuación existe una relación directa entre las variables y el factor, es decir, que cuanto mayor sean los valores de las variables más impacto tendrá sobre el factor3.

A continuación, expondremos un breve desarrollo teórico sobre el análisis factorial:

Dado un conjunto de q variables observadas $x = (x_1, x_2, \ldots, x_q)$, se asume que estas están relacionadas con otro conjunto de k variables latentes o factores comunes $f_1, f_2, \ldots, f_k$, donde $k < q$ mediante un modelo de regresión de la siguiente forma:

$$x_1 = \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1k}f_k + u_1,$$

$$x_2 = \lambda_{21}f_1 + \lambda_{22}f_2 + \cdots + \lambda_{2k}f_k + u_2,$$

$$\ldots$$

$$x_q = \lambda_{q1}f_1 + \lambda_{q2}f_2 + \cdots + \lambda_{qk}f_k + u_q.$$

Los términos λ_{ij} son los coeficientes de regresión de las variables x_i sobre los factores comunes f_j. En análisis factorial, estos coeficientes de regresión se conocen como pesos o cargas y muestran como cada una de las variables observadas x_i dependen de los factores comunes. Estos pesos se utilizan por ello para interpretar los factores, es decir, podemos resumir la expresión anterior como:

$$\mathbf{x = \Lambda f + u,}$$

1. **Λ** es una matriz $q \times k$ de constantes desconocidas $(k < q)$. Se conoce con el nombre de matriz de carga.

2. **f** es un vector de $q \times 1$ variables latentes. Se asume que estos factores no observados siguen una distribución normal tipificada $(f \sim N_q(0, I))$.

3. **u** es un vector $(q \times 1)$ de perturbaciones no observadas. Recoge el efecto de todas las variables distintas de los factores que influyen sobre x. Supondremos que u tiene una distribución $N_p(0, \psi)$ donde ψ es diagonal, y que las perturbaciones están incorrelacionadas con los factores f.

Vamos a detallar los principales elementos devueltos por la función “factanal”. En primer lugar, guardaremos el objeto creado por la función con el nombre “modelo”.

```
> modelo <- factanal(tfg_R[,2:9], factors = 3, scores = "regression")
```

A partir del objeto modelo, tenemos las siguientes componentes:

- Matriz de cargas (Λ):

```
> unclass(modelo$loadings)
                          Factor1     Factor2      Factor3
IPC                   -0.298938697  0.15476128 -0.130238942
PIB per capita         0.058846268  0.88138775  0.367252837
TASA DE PARO          -0.007922309 -0.97498552  0.210582592
TASA DE ACTIVIDAD      0.093885969 -0.03834801  0.711318205
OCUPADOS               0.945584812  0.18786761  0.192153988
NÚMERO DE HIPOTECAS    0.958059406  0.14853869  0.228583376
PRECIO VIVIENDA        0.235669215  0.11247525  0.877817408
VIOLENCIA DE GÉNERO    0.942654012 -0.05835832 -0.008794124
```

- Términos de variabilidad específica (uniquenesses) (ψ_{ii}):

```
> unclass(modelo$uniquenesses)
               IPC      PIB per capita        TASA DE PARO   TASA DE ACTIVIDAD
       0.869726193         0.084817815         0.005000000         0.483750110
         OCUPADOS NÚMERO DE HIPOTECAS     PRECIO VIVIENDA VIOLENCIA DE GÉNERO
       0.033651800         0.007808071         0.161244740         0.107921664
```

- Número de factores (k):

```
> unclass(modelo$factors)
[1] 3
```

- Matriz de covarianzas (correlación) (Σ):

```
> unclass(modelo$correlation)
                          IPC PIB per capita TASA DE PARO TASA DE ACTIVIDAD   OCUPADOS
IPC                1.00000000   0.0702984025  -0.17549156        0.07430154 -0.2714188
PIB per capita     0.07029840   1.0000000000  -0.78253914        0.24385947  0.2812680
TASA DE PARO      -0.17549156  -0.7825391386   1.00000000        0.18680756 -0.1506911
TASA DE ACTIVIDAD  0.07430154   0.2438594714   0.18680756        1.00000000  0.2151314
OCUPADOS          -0.27141883   0.2812680455  -0.15069112        0.21513137  1.0000000
NÚMERO DE HIPOTECAS -0.28831641   0.2738182251  -0.10414279        0.24469065  0.9778152
PRECIO VIVIENDA   -0.25956328   0.4301479482   0.07311970        0.65177005  0.4255543
VIOLENCIA DE GÉNERO -0.37013017  -0.0008567701   0.04733139        0.13034706  0.8768032
                  NÚMERO DE HIPOTECAS PRECIO VIVIENDA VIOLENCIA DE GÉNERO
IPC                        -0.2883164      -0.2595633       -0.3701301659
PIB per capita              0.2738182       0.4301479       -0.0008567701
TASA DE PARO               -0.1041428       0.0731197        0.0473313909
TASA DE ACTIVIDAD           0.2446906       0.6517701        0.1303470624
OCUPADOS                    0.9778152       0.4255543        0.8768032205
NÚMERO DE HIPOTECAS         1.0000000       0.4407169        0.8927131069
PRECIO VIVIENDA             0.4407169       1.0000000        0.1970319787
VIOLENCIA DE GÉNERO         0.8927131       0.1970320        1.0000000000
```

Calculamos los valores estimados de cada factor para cada una de las observaciones:

```
> scores <- factanal(tfg_R[,2:9], factors = 3, scores = "regression")$scores
> View(scores)
> head(scores)
        Factor1     Factor2    Factor3
[1,]  2.5945718 -1.29695522 -0.4509193
[2,] -0.5489276  0.92147647 -0.1651383
[3,] -0.3821607 -0.12305186 -0.8398339
[4,] -0.7013088 -0.07617415  1.3585178
[5,] -0.0205817 -1.12505541  0.1624098
[6,] -0.6727006  0.48060856 -0.5690911
```

Utilizaremos los valores obtenidos con la función “scores” para realizar una representación gráfica de los datos en la figura 1:

```
> par(mfrow = c(1,3))
> plot(scores[ ,1:2], asp = 1, )
> text(scores[ ,1], scores[ ,2], abb.var, cex = .7, pos = 3)
> plot(scores[ ,c(1,3)], asp = 1)
> text(scores[ ,1], scores[ ,3], abb.var, cex = .7, pos = 3)
> plot(scores[ ,c(2,3)], asp = 1)
> text(scores[ ,2], scores[ ,3], abb.var, cex = .7, pos = 3)
```

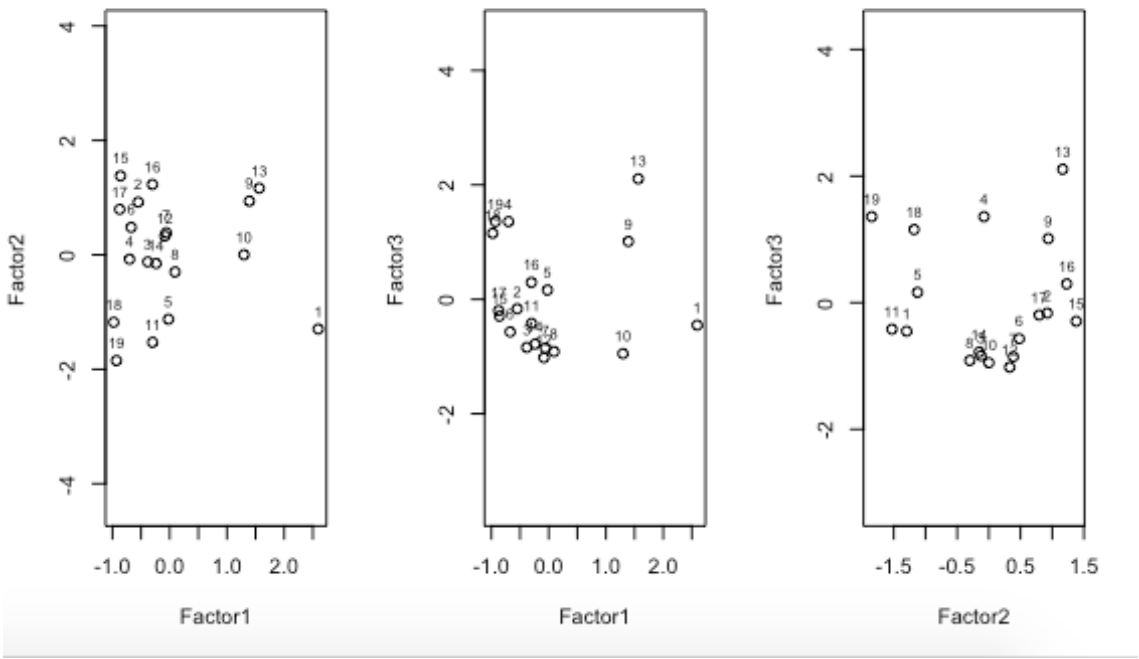


Figura 1: Diagramas de dispersión de los valores de los tres factores para el conjunto de datos de tfg_R.

2.4 Análisis Cluster.

A continuación, llevaremos a cabo el análisis cluster, con el cual podremos clasificar las observaciones, en este caso comunidades autónomas, en grupos, con el fin de distinguir en qué se parecen y en qué se diferencian.

Cada grupo realizado por el análisis será homogéneo respecto a las variables utilizadas para caracterizarlo y, además, los grupos serán lo más distintos posibles unos de otros respecto de las variables consideradas.

Para llevarlo a cabo ejecutaremos el siguiente agrupamiento en R:

```
hc.single<-hclust(dist(tfg_R[,2:9]),method="single")
plot(hc.single)
```

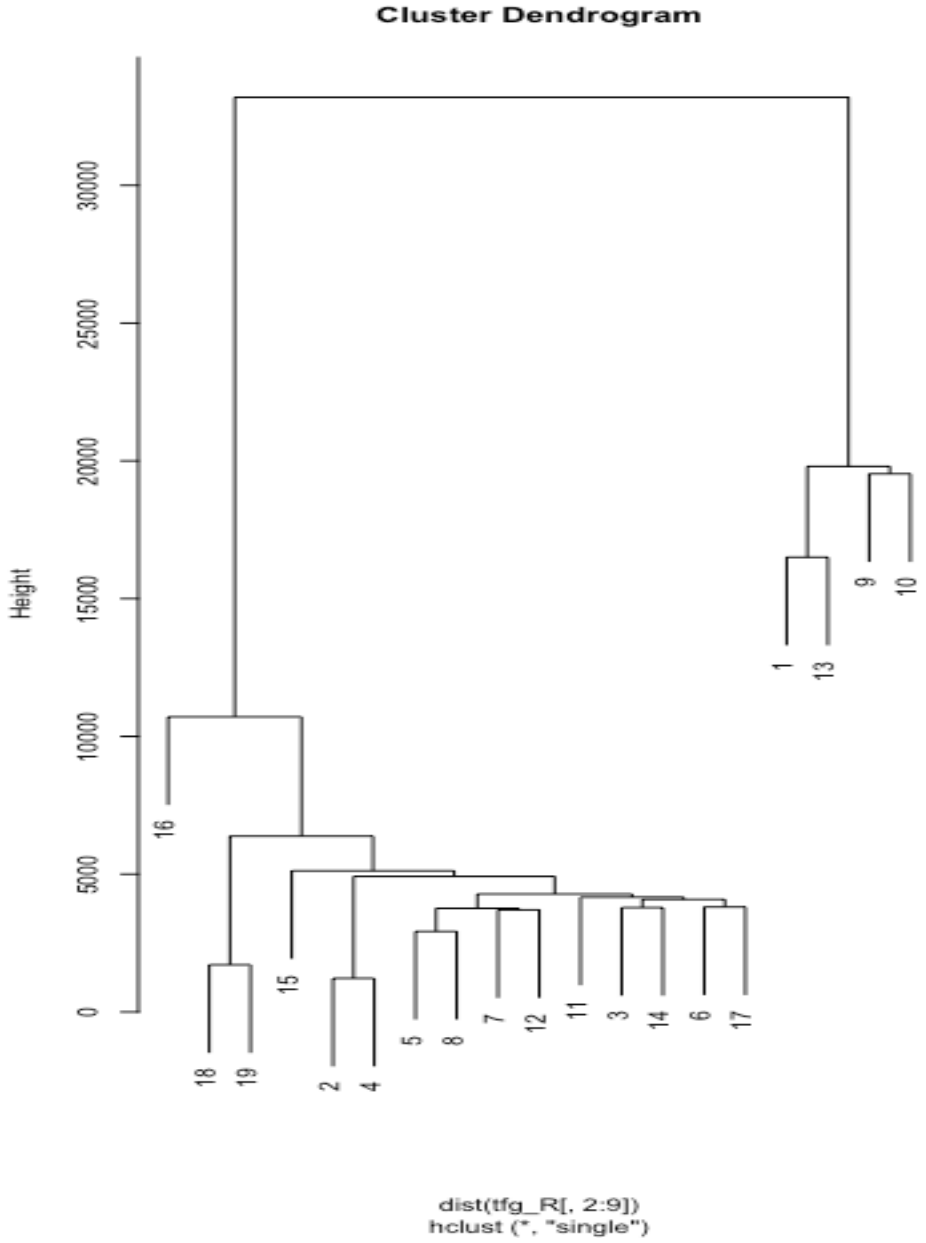
Utilizaremos la función “hclust” para ejecutar el análisis cluster jerárquico sobre un conjunto de distancias y métodos diferentes, “dist” es la matriz de distancia y, a través del argumento “method” especificaremos el método de unión que vamos a aplicar entre los grupos, es decir, la estrategia del método cluster a usar.
Los agrupamientos que obtengamos están contenidos en objetos de la clase “hclust” y representaremos el dendrograma con la función “plot”.

Existen distintos métodos de unión aplicables en el análisis Cluster. A continuación, haremos una breve explicación sobre ellos para poder analizar los resultados y sacar conclusiones sobre las agrupaciones resultantes.

1. Single: La distancia o similitud entre dos clusters viene provocada por la mínima distancia (o máxima similitud) entre sus componentes. Calcula las disimilitudes entre pares de observaciones y guarda la menor de ellas. Puede dar lugar a dendrogramas muy ramificados en que las observaciones se unen de una en una en cada interacción.
2. Complete: La distancia o similitud entre dos cluster viene dada por la máxima distancia (o mínima similitud) entre sus componentes. Calcula todas las disimilitudes entre pares de observaciones y guarda la mayor de ellas.
3. Centroid: Calcula la distancia entre dos clusters, por la diferencia entre sus centroides, es decir, entre los vectores de medias de las variables medidas sobre los individuos cluster.
4. Average: Disimilitud media entre clusters. Calcula todas las disimilitudes entre pares de observaciones dentro del cluster A y dentro del cluster B, y guarda la media de todas ellas.
5. Ward.D: Minimiza el total dentro de la varianza del cluster. En cada paso el par de cluster con distancia mínima entre ellos son mezclados.
6. Ward.D2: Es el método de Ward.D añadiendo que las diferencias se elevan al cuadrado antes de actualizar el cluster.

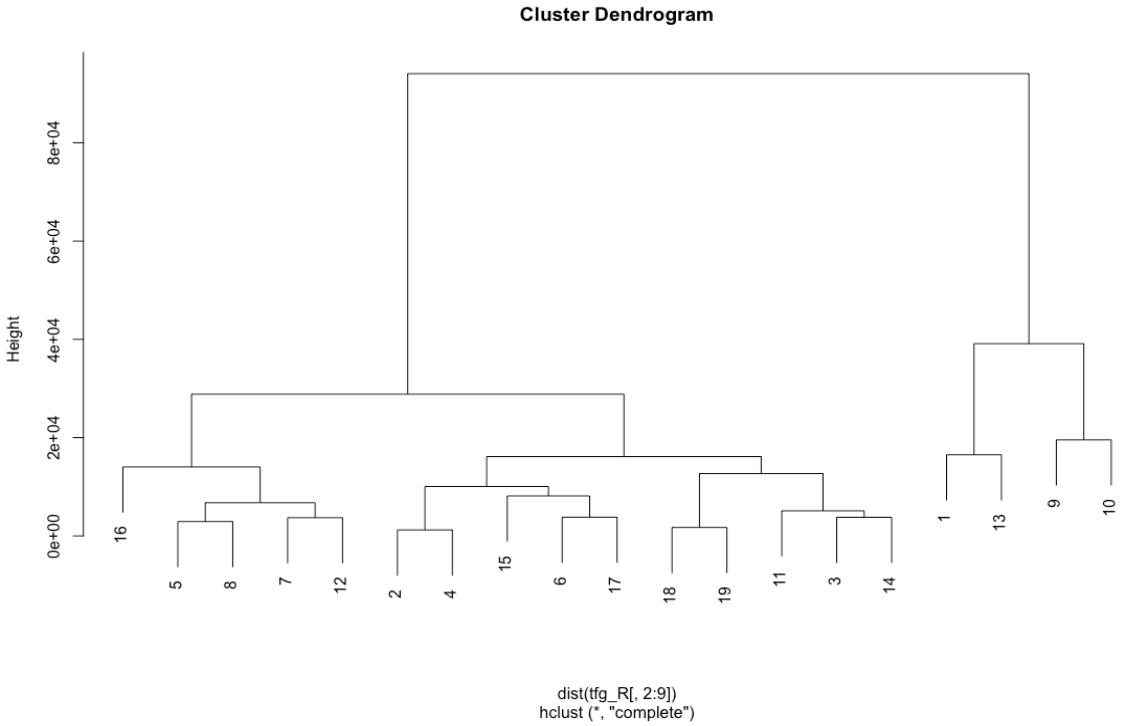
2.4.1. Método single.

```
hc.single<-hclust(dist(tfg_R[,2:9]),method="single")
plot(hc.single)
```



2.4.2. Método complete.

```
hc.complete<-hclust(dist(tfg_R[,2:9]),method="complete")
plot(hc.complete)
```

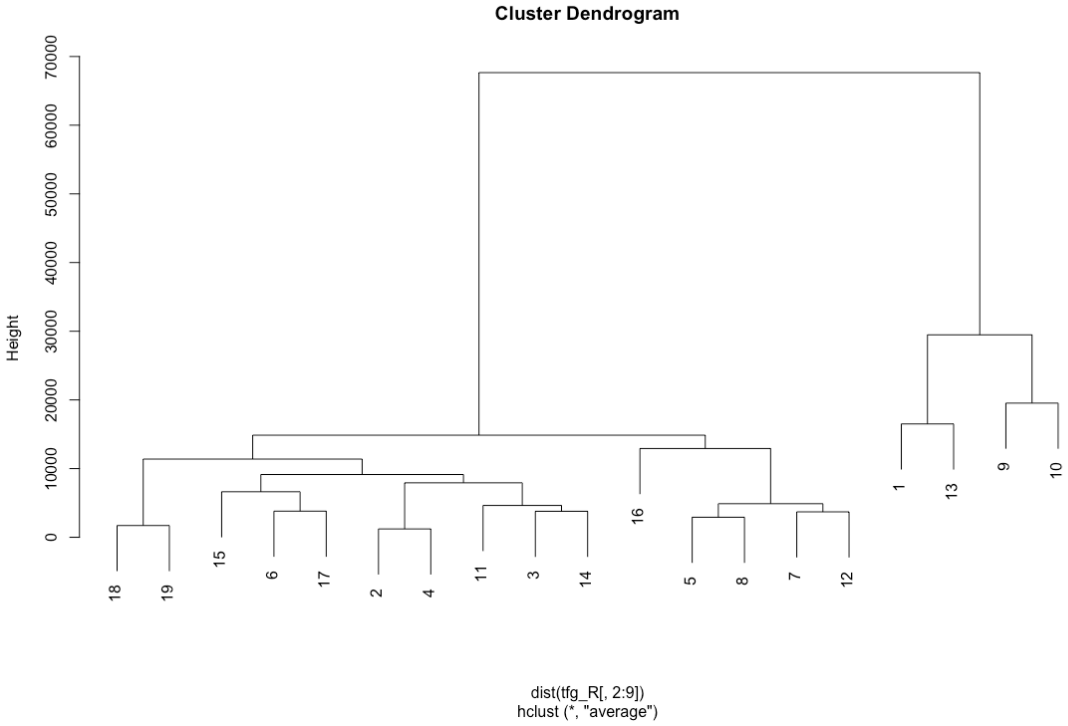


2.4.3. Método centroid.

```
hc.centroid<-hclust(dist(tfg_R[,2:9]),method="centroid")
plot(hc.centroid)
```

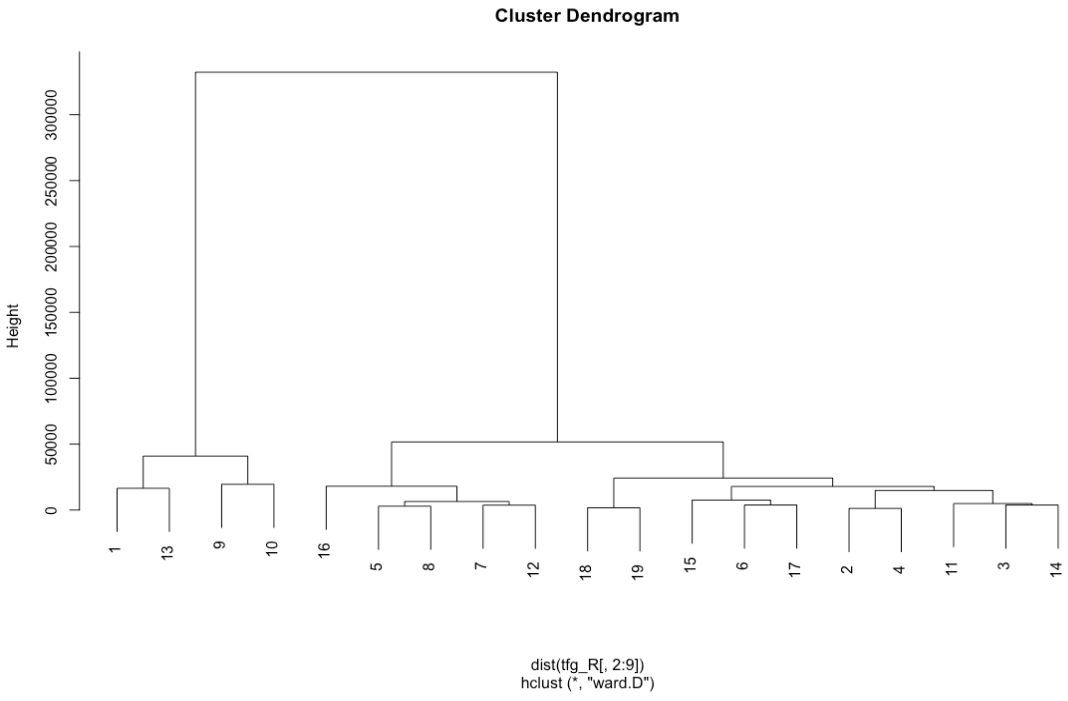
2.4.4. Método average.

```
hc.average<-hclust(dist(tfg_R[,2:9]),method="average")
plot(hc.average)
```



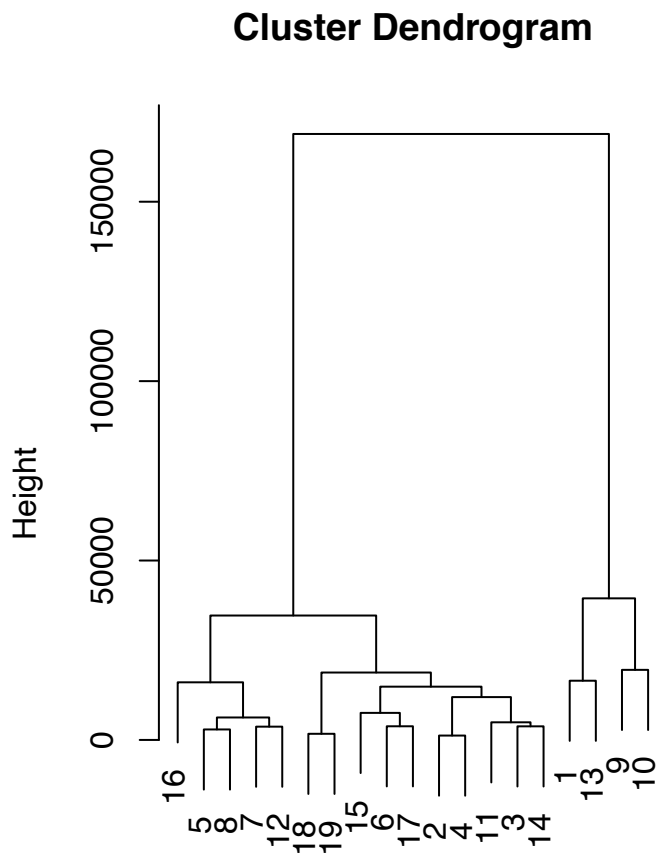
2.4.5. Método ward.D.

```
hc.wardD<-hclust(dist(tfg_R[,2:9]),method="ward.D")
plot(hc.wardD)
```



2.4.6. Método ward.D2.

```
hc.wardD2<-hclust(dist(tfg_R[,2:9]),method="ward.D2")
plot(hc.wardD2)
```



La parte baja de los dendrogramas resultantes identifican las observaciones de nuestro análisis, es decir, las 19 comunidades autónomas de nuestro país.

A simple vista podemos ver claramente que con todos los métodos cluster aplicados se generan dos grandes grupos, que posteriormente se dividen en más ramificaciones. Estos dos grupos están compuestos, en primer lugar, por Andalucía, Madrid, Cataluña y Comunidad Valenciana y en segundo lugar por las 15 comunidades autónomas restantes: País Vasco, Canarias, Castilla la Mancha, Castilla y León, Galicia, Aragón, Islas Baleares, Navarra, Cantabria, La Rioja, Ceuta, Melilla, Extremadura, Asturias e Islas Baleares.

Esta división tan evidente viene provocada, como adelantamos en el análisis descriptivo, por el número de habitantes de las comunidades, ya que en casi todas las variables que hemos analizado en este trabajo los resultados se han visto claramente

correlacionados con el número de habitantes de las comunidades.

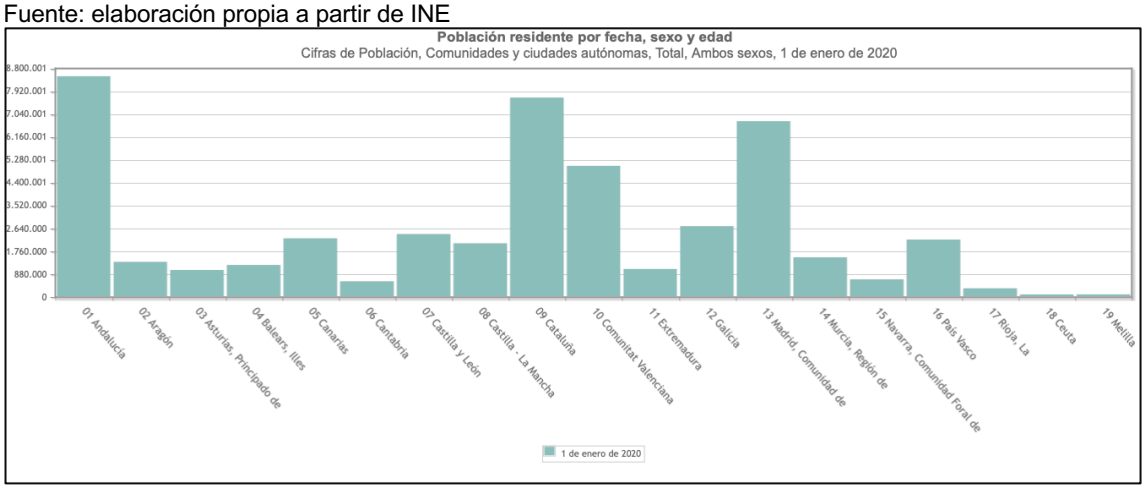
Las 4 comunidades que componen el primer grupo tienen un número de habitantes muy superior a las otras 15, por tanto, como ya hemos mencionado, al realizar los análisis, (sobre todo con los datos expresados en valores absolutos), estas 4 variables quedaban por encima del resto.

Si el análisis de los dendrogramas lo hacemos más detalladamente, es decir, el corte de la altura no es tan alto, la división con todos los métodos aplicados se divide en 3 grupos:

C1. Andalucía, Madrid, Cataluña y Comunidad Valenciana.
C2. País Vasco, Canarias, Castilla la Mancha, Castilla y León y Galicia.
C3. Aragón, Islas Baleares, Navarra, Cantabria, La Rioja, Ceuta, Melilla, Extremadura, Asturias e Islas Baleares.

Para poder explicar con mayor facilidad los resultados obtenidos con los métodos cluster hemos extraído del INE las cifras de población por comunidades autónomas.

Fuente: elaboración propia a partir de INE



En el gráfico el primer grupo con un mayor número de habitantes se ve claramente diferenciado del resto.
El segundo grupo, de la misma manera que sucede en los dendrogramas, es superior al tercero, pero las diferencias no son tan elevadas, por eso el corte de altura es más próximo.

Finalmente, tras haber estudiado detalladamente los resultados de los 6 métodos, vamos a seleccionar el método Ward.D debido a que nos aporta unos resultados más ordenados y podemos apreciar con mayor claridad los cluster formados.
Hemos decidido hacer el corte de altura en d=5000, de tal forma que obtendremos los 3 grupos ya mencionados afectados por las cifras poblacionales.

- Analizando C1 (Andalucía, Madrid, Cataluña y Comunidad Valenciana) apreciamos que presentan valores bastante elevados tanto de F1, F2 y F3. Como ya hemos adelantado esto es consecuencia en parte del número de habitantes de este conglomerado ya que es más elevado que el resto de comunidades. Podemos confirmar con los resultados que es el conglomerado más avanzado en términos de desarrollo económico y social y de calidad de vida, con un PIB per cápita superior a la media, una alta tasa de actividad, así

como unos datos de ocupación e hipotecas muy superiores al resto de comunidades.

- Estudiando el C2 (País Vasco, Canarias, Castilla la Mancha, Castilla y León y Galicia) destacamos los datos recogidos en el análisis de condenados por violencia de género ya que los resultados son bastante elevados en comparación con el resto de Comunidades, teniendo en cuenta la población de las mismas. Estos resultados tendrán como consecuencia que sean unas zonas más inseguras que el resto, lo cual afectará directamente al modo de vida en dichas regiones.
- Por último, el conglomerado C3 (Aragón, Islas Baleares, Navarra, Cantabria, La Rioja, Ceuta, Melilla, Extremadura, Asturias e Islas Baleares) presenta un escaso desarrollo económico y social. En algunas regiones de C3 más que en otras ya que este conglomerado está formado por un gran número de Comunidades. Presentan bajos niveles de PIB per cápita, así como una elevada tasa de paro en muchas de las comunidades que lo componen.

Los grupos de conglomerados han sido creados a través del análisis cluster en función de las variables que hemos incluido en nuestro estudio, pero, además, debemos mencionar que nuestros datos a analizar están en valores brutos, no en porcentajes sobre la población, por lo tanto, estos grupos se han visto afectados, además de por las variables, por las cifras poblacionales de nuestras 19 Comunidades Autónomas.

2.5. Análisis Anova.

Para finalizar nuestro análisis, vamos a llevar a cabo el ANOVA o análisis de la varianza. Se trata de un método estadístico que determinará si los conjuntos de muestras aleatorias de la variable proceden de una misma población o de poblaciones distintas.

Para llevarlo a cabo añadiremos una nueva variable, que llamaremos “cluster” ya que está formada por los 3 grupos que hemos seleccionado en el análisis anterior. Esta variable actuará como factor a la hora de realizar el ANOVA. En este epígrafe compararemos los valores de las variables frente a este nuevo factor denominado “cluster”.

Tendremos en cuenta que este último análisis es un test por lo que debemos aceptar o rechazar las hipótesis que planteemos en función los resultados obtenidos.

La hipótesis nula a contrastar a través del Análisis de la Varianza puede ser establecida como igualdad de efectos:

$$Ho: \mu_1 = \mu_2 = .. = \mu_G$$

$$H_1 : \text{alguna distinta}$$

La cuestión que queremos abordar es si los grupos obtenidos con el análisis cluster están correctamente creados o por el contrario la división se ha hecho de manera artificial.

Si las medias son iguales es que la divisón es artificial así que vamos a proceder a realizar los cálculos.

Para calcular las medias de cada grupo de conglomerados directamente calculamos las medias de las columnas.

```
> colMeans(cluster[,2:4])
       C1        C2        C3
10061.364  6046.295  4145.816
```

Vemos claramente que el C1 tiene una media muy elevada en comparación con los otros dos grupos, por el contrario, C3 se corresponde con la media más baja, quedando entre ambos el C2. Apreciamos unas diferencias muy grandes, pero lo que nos interesa es saber si dichas diferencias son o no significativas.

Comprobaremos para ello si las varianzas son constantes para cada conglomerado.

```
> apply(cluster[,2:4],MARGIN = 2, FUN = var)
       C1        C2        C3
343883526 109665101  75254154
```

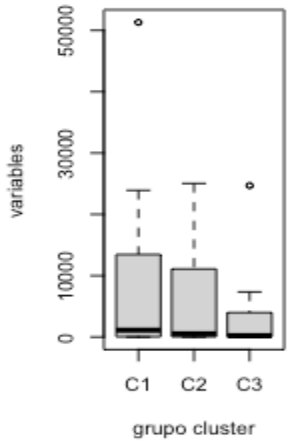
La varianza del grupo C1 es bastante más del cuádruple que la de C3. Aplicaremos a continuación el test de homogeneidad de las varianzas de Fligner-Killeen con el fin de comprobar si la diferencia entre las varianzas es signiticativa. Utilizamos la función fligner.test:

```
> fligner.test(cluster[,2:4])

        Fligner-Killeen test of homogeneity of variances

data:  cluster[, 2:4]
Fligner-Killeen:med chi-squared = 5.1895, df = 2, p-value =
0.07466
```

Nuestra hipótesis nula es que las varianzas son homogéneas frente a la hipótesis alternativa de que no lo son. Al realizar el test obtenemos como resultado un p-valor de 0,07466, valor superior a 0.05, por tanto, con un nivel de confianza del 95%, no rechazaremos la hipótesis nula. No encontramos evidencia de que las varianzas sean inhomogéneas y por lo tanto podemos continuar con el ANOVA.

Utilizaremos la función boxplot para mostrar los datos mediante un diagrama de cajas ya que estamos analizando variables explicativas de tipo categórico.

```
> boxplot(cluster[,2:4], ylab = "variables", xlab = "grupo cluster")
```



A partir de este gráfico parece que C1 es significativamente mayor que C3. No obstante, para analizar las varianzas de manera más formal realizaremos cálculos con ayuda de R:

En primer lugar, definimos el vector Y concatenando las columnas de la tabla cluster.

```
> y <- with(cluster[,2:4], c(C1, C2, C3))
```

De acuerdo con la definición de SSY, lo calculamos del siguiente modo:

```
> (SSY <- sum((y - mean(y))^2))
[1] 3847556259
```

La varianza no explicada (SSE), se calcula para cada uno de los grupos cluster por separado:

```
> (SSE <- sum(apply(cluster[,2:4], MARGIN = 2, FUN = function(x) (x - mean(x))^2)))
[1] 3701619462
```

La medida en que SSE es menor que SSY es una indicación de la magnitud de la diferencia entre las medias. Cuanto mayor sea la diferencia entre los valores de los datos de nuestros conglomerados, mayor será la diferencia entre SSE y SSY; dicho de otro modo, mayor será la varianza explicada (SSA). Este es el principio en que se basa el análisis de las varianzas: podemos hacer inferencia sobre la diferencia entre medias mirando las diferencias entre las varianzas (o entre sumas de cuadrados). Por último, la varianza explicada por el factor suelo se calculará simplemente como la diferencia entre la varianza total (SSY) y la no explicada (SSE):

```
> (SSA <- SSY - SSE)
[1] 145936797
```

Los resultados del análisis de las varianzas se representan en una tabla característica, que suele contener típicamente (al menos) 6 columnas que indican, de izquierda a derecha:

- la fuente de variación
- la suma de cuadrados atribuible a dicha fuente
- los grados de libertad de dicha fuente
- la varianza de dicha fuente (tradicionalmente designada como medias cuadradas en lugar de varianza)
- el estadístico F (prueba la hipótesis nula de que dicha fuente de variación no es significativamente distinta de cero)
- p-valor asociado al valor de F (si p < 0:05, rechazamos la hipótesis nula con un nivel de confianza del 95%)

La media cuadrática se obtiene como el cociente entre cada suma de cuadrados y sus respectivos grados de libertad. La varianza del error, s2, es la media cuadrática residual (la media de los cuadrados de la varianza no

explicada, SSE). En este ejemplo se corresponde con la media de las varianzas de cada uno de los grupos por separado.

```
> apply(cluster[,2:4],MARGIN = 2, FUN = var)
       C1        C2        C3
343883526 109665101  75254154

> (s2 <- mean(apply(cluster[,2:4], MARGIN = 2, FUN = var)))
[1] 176267593
```

El estadístico F es la media cuadrática de los grupos dividida por la varianza del error, y contrasta la hipótesis nula de que no hay diferencias significativas entre las medias correspondientes a los diferentes grupos. Si se rechaza la hipótesis nula, se acepta la hipótesis alternativa de al menos una de las medias de alguno de los grupos es significativamente diferente del resto.
Por lo tanto, tenemos que la media cuadrática viene dada por el cociente entre SSA y los grados de libertad del tipo de suelo:

```
> (MC <- SSA/2)
[1] 72968399
```

y la F por lo tanto:

```
> (F = MC/s2)
[1] 0.4139638
```

La cuestión ahora es decidir si F es suficientemente grande como para poder rechazar la hipótesis nula.

Para tomar decisiones aplicaremos la función aov para llevar a cabo anova en R.

Antes de aplicarla, realizaremos una pequeña readaptación del formato de los datos de entrada. Utilizaremos un vector de datos y otro vector con el factor a considerar (grupos cluster), y crearemos una tabla de datos, para aplicar el formato de entrada de datos de tipo fórmula característico en la creación de modelos en R:

```
> (y <- as.vector(as.matrix(cluster[,2:4])))
 [1]     2.2750 23919.5000    15.2250    57.4000  2101.7250 51299.0000   117.7857
 [8]  2978.0000     2.2000 25041.2000    14.2400    56.3800   940.5400 20736.6000
[15]   113.5998  1465.6000     2.2100 24692.0000    15.9900    57.9100   318.5800
[22]  7355.2000   117.4408   607.2000

> (grupos <- rep(names(cluster[,2:4]), each = 8))
 [1] "C1" "C1" "C1" "C1" "C1" "C1" "C1" "C1" "C2" "C2" "C2" "C2" "C2" "C2" "C2" "C2" "C3"
[18] "C3" "C3" "C3" "C3" "C3" "C3" "C3"

> tabla.datos <- cbind.data.frame(y, grupos)
> head(tabla.datos)
          y grupos
1     2.275     C1
2 23919.500     C1
3    15.225     C1
4    57.400     C1
5  2101.725     C1
6 51299.000     C1
```

Una vez creada la tabla de datos, podemos proceder con el ANOVA mediante la función aov:

```
> summary(aov(y ~ grupos, data = tabla.datos))
            Df    Sum Sq   Mean Sq F value Pr(>F)
grupos       2 1.459e+08  72968399   0.414  0.666
Residuals   21 3.702e+09 176267593
```

Apreciamos en la tabla el p-valor como Pr(>F). Dado que el p-valor obtenido es 0.666 es mayor que 0.05 no rechazamos la hipótesis nula, lo cual quiere decir que todos los efectos del factor grupos son iguales.

3. CONCLUSIONES: Principales resultados y recomendaciones.

En este último apartado expondremos los resultados finales de todos los análisis que hemos llevado a cabo.

- En cuanto al análisis descriptivo de las variables, podemos concluir que Ceuta, Melilla y Castilla-La Mancha son zonas que ofrecen resultados bastante desfavorables, elevada tasa de inflación, un PIBpc bajo, alta tasa de paro, etc. Además, observamos zonas como Extremadura o Andalucía con valores poco positivos, en términos generales. En cambio, en sentido positivo resaltaremos Comunidades como Madrid y Cataluña que presentan niveles mejores que el resto del país en casi todos los ámbitos estudiados.

- A través del análisis factorial agrupamos las distintas variables seleccionadas en factores con el fin de que tuvieran carácter de indicadores y de esa manera poder interpretar los resultados.

  Nuestras variables se saturaron en tres factores:

  - F1. Número de ocupados, número de hipotecas y detenidos por violencia de género.
  - F2. IPC, PIBpc y tasa de paro.
  - F3. Tasa de actividad y precio de la vivienda.

- Mediante el análisis cluster agrupamos por Comunidades Autónomas, en función de las similitudes y diferencias de las mismas, obteniendo así tres nuevos grupos o conglomerados:

  - C1. Andalucía, Madrid, Cataluña y Comunidad Valenciana.
  - C2. País Vasco, Canarias, Castilla la Mancha, Castilla y León y Galicia.
  - C3. Aragón, Islas Baleares, Navarra, Cantabria, La Rioja, Ceuta, Melilla, Extremadura, Asturias e Islas Baleares.

Ahora, una vez expuestos los resultados obtenidos, vamos a analizar los grupos (o conglomerados) de tal manera que expondremos los problemas que hemos detectado en los mismos y posteriormente propondremos una serie de medidas de política económica con el fin de paliar los problemas previamente expuestos:

- Analizando C1 (Andalucía, Madrid, Cataluña y Comunidad Valenciana) apreciamos que presentan valores bastante elevados tanto F1, F2 y F3.
  Los resultados obtenidos en C1 nos llevan a confirmar que es el conglomerado más avanzado en términos de desarrollo económico y social y de calidad de vida. Como problema podemos detectar los elevados precios de las viviendas

en las comunidades que lo componen. Este problema no es una tarea fácil de resolver, es más, en nuestro país llevamos muchos años tratando de solucionarlo y las políticas aplicadas suelen fracasar debido a la mentalidad tan clara que tenemos de que “si no compramos hoy, mañana será más caro”. Esta mentalidad es consecuencia de que los precios de las viviendas los últimos años sólo han tenido una tendencia creciente, por tanto, nadie espera que, si no compras “hoy”, en un futuro exista la posibilidad de que el precio sea inferior.
Podríamos proponer aplicar políticas llevadas a cabo por el gobierno como regular el precio del suelo o incrementar las ayudas directas y las subvenciones.

- Estudiando los problemas de C2 (País Vasco, Canarias, Castilla la Mancha, Castilla y León y Galicia) destacamos los datos recogidos en el análisis de condenados por violencia de género ya que los resultados son bastante elevados en comparación con el resto de Comunidades, teniendo en cuenta la población de las mismas. Estos resultados tendrán como consecuencia que sean unas zonas más inseguras que el resto, lo cual afectará directamente al modo de vida en dichas regiones.
Para tratar de revertir esta situación, nuestra propuesta es incrementar las medidas de seguridad y de control de tal modo que hagan reducir los maltratos en las Comunidades que forman C2.

- Por último, el conglomerado C3 (Aragón, Islas Baleares, Navarra, Cantabria, La Rioja, Ceuta, Melilla, Extremadura, Asturias e Islas Baleares) presenta como problema fundamental el escaso desarrollo económico y social, con un PIB per cápita inferior a la media y unos datos de parados bastante elevados.
Para poder hacer frente al problema de la elevada tasa de paro proponemos medidas políticas como el apoyo a los autónomos y emprendedores, ya que, a diferencia con otros países, en el nuestro reciben pocas ayudas a nivel económico. Con ello podríamos recortar un gran número de parados en nuestro país y sobre todo en las comunidades más afectadas como lo son las que forman este conglomerado.

Bibliografía

- Boletín estadístico online – Información estadística – Ministerio de Fomento disponible en <https://apps.fomento.gob.es/BoletinOnline2/?nivel=2&orden=36000000>
- España: población por género y región 2019 – Statista disponible en <https://es.statista.com/estadisticas/473637/poblacion-de-espana-por-genero-y-comunidad-autonoma/>
- INE. Instituto Nacional de Estadística disponible en https://www.ine.es
- Matematicas.unex.es disponible en <http://matematicas.unex.es/~trinidad/mui/tutorial.R.pdf>
- Medidas para bajar los precios de la vivienda disponible en <https://www.eleconomista.es/opinion-blogs/noticias/10059788/08/19/Medidas-para-bajar-los-precios-de-la-vivienda.html>
- Métodos jerárquicos de análisis cluster, capítulo 3 disponible en https://www.ugr.es/~gallardo/pdf/cluster-3.pdf
- Personales.unican.es disponible en <https://personales.unican.es/gonzaleof/r/comandos.pdf>
- Poza Lara, Carlos (2005): “Análisis estadístico multivariante por Comunidades Autónomas: diferencias y similitudes” Universidad Antonio de Nebrija.
- RStudio disponible en <https://rstudio.com/products/rstudio/>
- Sordo García, Carmen María (2019): “Análisis Anova: “Comparaciones entre grupos de observaciones” Universidad de Cantabria.
- Sordo García, Carmen María (2019): “Análisis Factorial” Universidad de Cantabria.
- Sordo García, Carmen María (2019): “Técnicas de agrupación: Análisis cluster” Universidad de Cantabria.