**TFM** [**MÁSTER EN BUSINESS ANALYTICS Y BIG DATA**](https://campusformacionnebrija.imf.com/general/prt/prg/master.php?codcurso=MBIGDATA3&varp=335) **- IMF**

Análisis criminológico en función de la situación económica y el nivel educativo de España

*Memoria del proyecto de Data Science.*

*Laura Esteban de Pedro*

**ÍNDICE**

[**RESUMEN** 2](#_Toc56441819)

[**INTRODUCCIÓN** 2](#_Toc56441820)

[**Objetivo** 2](#_Toc56441821)

[**Orígenes de datos** 2](#_Toc56441822)

[**METODOLOGÍA** 4](#_Toc56441823)

[**Preprocesamiento de datos** 4](#_Toc56441824)

[**Resumen del proceso** 4](#_Toc56441825)

[**Carga de datos** 4](#_Toc56441826)

[**Análisis de los delitos** 13](#_Toc56441827)

[**Inspección por variables** 13](#_Toc56441828)

[**Análisis sobre los datos de los delitos** 18](#_Toc56441829)

[**Análisis sobre los datos del paro** 22](#_Toc56441830)

[**Análisis en función del tipo del delito** 26](#_Toc56441831)

[**Modelo Regresión Lineal** 28](#_Toc56441832)

[**Técnicas de selección de variables** 29](#_Toc56441833)

[**Modelos regularizados** 35](#_Toc56441834)

[**Comparación de Modelos** 41](#_Toc56441835)

[**Análisis de condenados** 43](#_Toc56441836)

[**Clustering** 45](#_Toc56441837)

[**Obtención del número de clúster** 46](#_Toc56441838)

[**Creación de los modelos** 49](#_Toc56441839)

[**Resultados** 51](#_Toc56441840)

[**Validación de los clústers** 52](#_Toc56441841)

[**CONCLUSIÓN** 54](#_Toc56441842)

# **RESUMEN**

Este proyecto se centra en el análisis del índice de criminalidad basándose en los datos del PIB, del gasto medio y del nivel educativo por habitante de España desde el año 2002 hasta el 2010.

A priori, antes de comenzar con el estudio, parto de la teoría de que el aumento del paro, un nivel económico bajo y un nivel educativo pobre puede favorecer el crecimiento de la tasa de criminalidad debido a la necesidad de supervivencia del ser humano. En situaciones de pobreza y hambruna puede ser que aumenten los hurtos, robos o ataques a otras personas. En una población con un nivel educativo bajo, puede aumentar el desorden público, contra el patrimonio o contra la autoridad.

Pero esto es sólo una hipótesis. Veremos si el análisis realizado afirma esta teoría o se verán reflejados otros resultados totalmente opuestos.

# **INTRODUCCIÓN**

## **Objetivo**

El objetivo del estudio es analizar los distintos factores y ver hasta qué punto influyen en el índice de criminalidad a lo largo de los años. En caso de que se vea una cierta dependencia se creará un modelo capaz de predecir la tasa de criminalidad para los años siguientes.

Se llevará a cabo otro tipo de estudio sobre los perfiles de las personas condenadas desde el año 2013 al 2019, a través de la información de la comunidad autónoma donde se produjo el crimen y la edad y el sexo del condenado. Se intentará identificar ciertos patrones que a simple vista son más difíciles de apreciar.

## **Orígenes de datos**

Se parte de la información obtenida del INE (Instituto Nacional de Estadística) y de las estadísticas proporcionadas por el Ministerio de Educación y Formación Profesional, dónde se realizaron distintas búsquedas para obtener los informes.

**Datos sobre los delitos y condenados:**

<https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176793&menu=resultados&idp=1254735573206>

**Datos demográficos (informe 3.1**[**Población por comunidades y ciudades autónomas y sexo.**](https://www.ine.es/jaxiT3/Tabla.htm?t=2853&L=0)**):**

<https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=523>

**Datos sobre el nivel educativo de España:**

<http://estadisticas.mecd.gob.es/EducaDynPx/educabase/index.htm?type=pcaxis&path=/Formacionyml/EPA_2020_05/NivFor&file=pcaxis&l=s0>

**Datos del PIB y del paro:**

<https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176825&menu=resultados&idp=1254735576508>

Los informes están descargados dentro de la carpeta “Informes Originales” dividido en deferentes categorías en función de los datos que contengan.

* Delitos: Contiene la información del número de delitos cometidos en función del tipo del delito y la comunidad autónoma donde se ha producido desde el año 1998 al 2019.
* Demografía: datos demográficos por comunidad autónoma, necesarios para calcular la tasa de delitos y la tasa de paro desde el año 1996 al 2019.
* Económico: Esta categoría se divide en dos subcategorías que contienen los datos del PIB y el gasto medio por habitante desde el año 1995 al 2010, y la segunda subcategoría que contiene el número de persona paradas desde el año 1994 al 2010.
* Educación: datos del nivel de educación (Inferior a 2º etapa de Educación Secundaria, 2º etapa Educación Secundaria y Educación Superior) por edad, sexo y comunidad autónoma desde el año 2002 al 2019.
* Personas condenadas: número de personas condenadas por edad, sexo y comunidad autónoma desde el año 2013 al 2019.

# **METODOLOGÍA**

## **Preprocesamiento de datos**

### **Resumen del proceso**

El preprocesado y la unificación de los datos se ha realizado en un notebook de Python 3 llamado Proceso de obtención de datos.ipynb

Cada informe tiene un formato específico y distinto al resto, incluso el mismo informe puede cambiar el formato en los distintos años. Por ello es necesario hacer un pre procesado de los datos y limpiarlos para poder realizar los distintos análisis. Para cada estudio se preparará un dataset distinto que contendrá información realizada para ese análisis. Estos dos dataset se guardan en ficheros dentro de la carpeta “Datos Procesados”.

El primero de ellos, llamado “datos\_delitos.csv”, contiene la tasa de delitos en función del tipo del delito, el año, la comunidad autónoma, el PIB, el gasto medio, la tasa de paro y el nivel educativo. Con estos datos se creará un modelo de regresión lineal que sea capaz de analizar la dependencia de estas variables y ver su nivel de influencia sobre la tasa de delitos, además de realizar una predicción de ésta sobre los siguientes años.

El segundo dataset se llama “datos\_condenados.csv” y contiene el número de condenados en función de la comunidad autónoma, el año, el sexo y la edad de la persona. A partir de esta información se intentará perfilar los distintos grupos distintos de delincuentes que pueda haber en función a estas características y ver cómo va evolucionando a lo largos de los años.

### **Carga de datos**

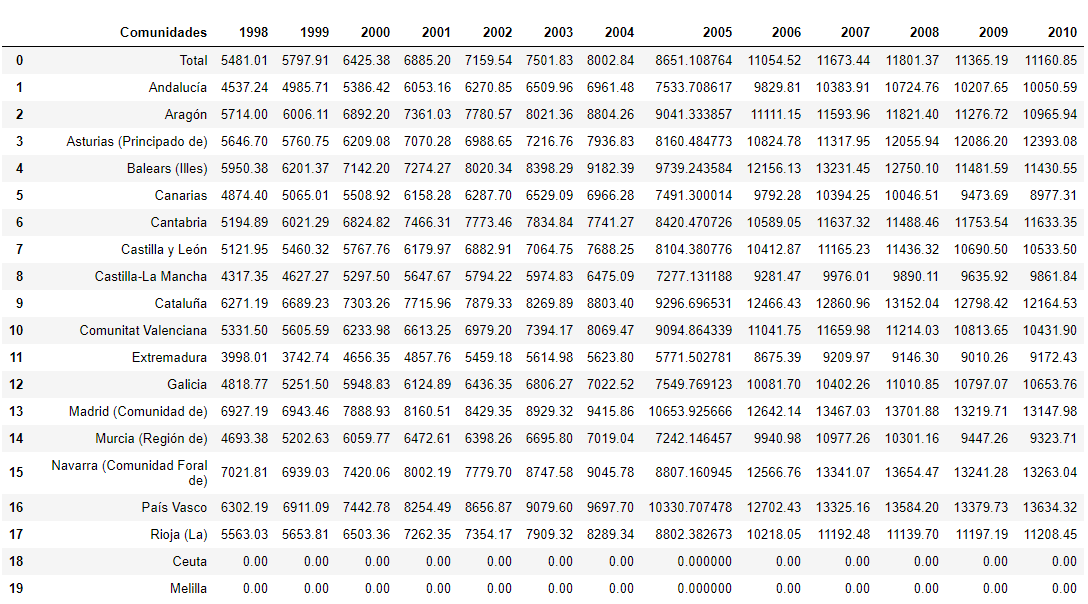
Antes de iniciar la carga de cada uno de los ficheros se importan las dos librerías pandas y numpy que se van a utilizar en este proceso.

También se crea un diccionario con el nombre de las comunidades autónomas y una función que se encarga de unificarlos nombres de las comunidades autónomas de los distintos ficheros. Tanto el diccionario como la función “unificar\_nombre\_comunidad\_autonoma(df)” se utilizan en la carga de varios ficheros.

**Fichero PIB por habitante**

Se obtienen los datos del PIB por habitante. Para ello se realizan las siguientes acciones que también se detallan en el código:

* Se renombra la columna de las comunidades y se limpia el nombre de las columnas de los años.
* Se eliminan las columnas vacías que no nos interesan.
* Se renombran los nombres de las comunidades para unificarlos con el resto de los ficheros.
* Se eliminan los registros que no corresponden a las comunidades autónomas, como son los registros correspondientes a las provincias.
* Se formatea la columna Comunidades como una variable categórica.
* Se redimensiona el dataframe para pasar los años a una única columna llamada *Anio*.
* Se formatea la nueva columna del año y la columna que contiene los datos del PIB para redondear los datos a 2 decimales.



**Fichero Gasto por habitante**

Se obtienen los datos del gasto medio por habitante. Para ello se realizan las siguientes acciones que también se detallan en el código:

* Se renombra la columna de las comunidades y se limpia el nombre de las columnas de los años.
* Se eliminan las columnas vacías que no nos interesan.
* Se renombran los nombres de las comunidades para unificarlos con el resto de los ficheros.
* Se eliminan los registros que no corresponden a las comunidades autónomas, como son los registros correspondientes a las provincias.
* Se formatea la columna Comunidades como una variable categórica.
* Se redimensiona el dataframe para pasar los años a una única columna llamada *Anio*.
* Se formatea la nueva columna del año.



**Fichero de los datos del paro**

Se obtienen los datos del paro. Para ello se realizan las siguientes acciones que también se detallan en el código:

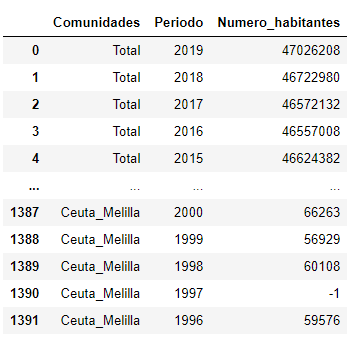
* Se obtienen los datos del paro.
* Se renombra la columna de las comunidades y se limpia el nombre de las columnas de los años.
* Se renombran los nombres de las comunidades para unificarlos con el resto de los ficheros.
* Se formatea la columna Comunidades como una variable categórica.
* Se eliminan los registros que no corresponden a las comunidades autónomas, como los registros correspondientes a las provincias.
* Se eliminan las columnas vacías que no interesan para el estudio.
* Se redimensiona el dataframe para pasar los años a una única columna llamada *Anio*.
* Se formatea la nueva columna del año.
* La columna Número de parados del fichero refleja el dato en miles de habitantes, por lo tanto, se multiplica el valor por 1000.



**Fichero de los datos demográficos**

Se obtienen los datos demográficos. Para ello se realizan las siguientes acciones que también se detallan en el código:

* Se obtienen los datos demográficos.
* Se formatea la columna Total y se reemplazan los valores nulos por -1.
* Se renombra la columna de las Comunidades y la columna Total por *Numero\_habitantes.*
* Se eliminan todas las columnas que no nos interesan (las referentes al desglose por sexo).
* Se renombran los nombres de las comunidades para unificarlos con el resto de los ficheros.
* Se formatea la columna Comunidades como una variable categórica.



**Fichero de los datos del nivel educativo**

Se obtienen los datos del nivel educativo de España. Para ello se realizan las siguientes acciones que también se detallan en el código:

* Se obtienen los datos del nivel educativo de España
* Se renombra la primera columna de las comunidades
* Se eliminan todas las columnas que no nos interesan (las referentes al desglose por sexo)
* Se elimina la primera fila que no aporta nada de información
* Se eliminan las filas del desglose por rango de edad
  + Se renombran los nombres de las comunidades para unificarlos con el resto de los ficheros
* Se formatea la columna Comunidades como una variable categórica

Se separan los distintos niveles de educación en distintos dataframes: Educación inferior a la 2º etapa de secundaria, Educación 2º etapa de secundaria y Educación Superior.

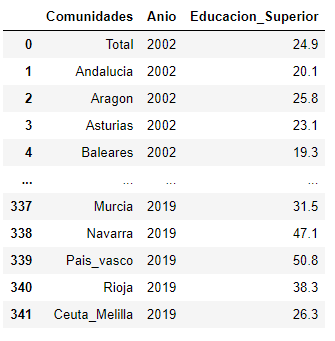
Educación inferior a la 2º etapa de secundaria



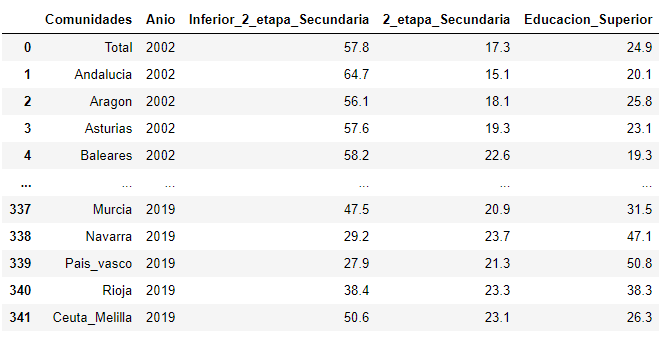
Educación de la 2º etapa de secundaria



Educación superior



Se unifican los 3 dataframes de educación



**Fichero de datos de los delitos**

Se obtienen los datos de los delitos. Para estos informes hay que tener en cuenta el formato de cada uno de ellos, ya que cambia según el año, por eso se hace distintas cargas. Para ello se realizan las siguientes acciones que también se detallan en el código:

Carga de los delitos a partir del año 2012

* Se obtienen los datos de los delitos a partir del 2012
* Se renombran y ordenan las distintas columnas
* Se formatea la columna Numero\_Delitos y se actualiza los nulos por -1

Carga de los delitos del 2007 al 2012

* Se obtienen los datos de los delitos desde el 2007 al 2012
* Se renombran y ordenan las distintas columnas
* Se formatea la columna Numero\_Delitos y se actualiza los nulos por -1
* Sólo obtenemos los datos totales sin el desglose por sexo

Carga de los delitos del 1998 al 2006

* Se obtienen los datos de los delitos desde el 1998 al 2006
* Se renombran y ordenan las distintas columnas
* Se formatea la columna Numero\_Delitos y se actualiza los nulos por -1

Se renombran los nombres de las comunidades para unificarlos con el resto de los ficheros

Se formatea la columna Comunidades como una variable categórica



**Se unifica toda la información en un único dataframe para su posterior análisis**

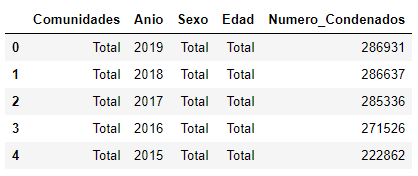
* Se crea un dataframe llamado df\_datos\_economicos\_educacion\_delitos que contiene toda la información que se añadirá a continuación uno a uno.
* Se unifican los dataframe del producto interior bruto y del gasto por habitante.
* Se incluyen los datos de parados.
* Se incluyen los datos de la población.
* Se elimina la columna del Periodo, no es necesaria porque ya tenemos el año.
* La información de parados, pib y gastos para las comunidades de Ceuta y Melilla vienen por separado, en cambio los datos sobre el nivel de educación y de delitos nos llegan unificados para las 2 comunidades, con lo cual se analizarán los datos conjuntos. Para ello, se unifican las 2 comunidades calculando la media y el sumatorio.
* Se eliminan las etiquetas "mean" de las columnas calculadas y se redondean a 2 decimales las columnas de PIB y Gasto.
* Se calcula la tasa del paro.
* Se añade la información del nivel educativo.
* Se añade la información de los delitos.
* La información de número de delitos para las comunidades de Ceuta y Melilla vienen por separado, con lo que se unifican las 2 comunidades calculando el sumatorio.
* Se eliminan las etiquetas "mean" de las columnas calculadas.
* Se calcula la tasa del delito.
* Se ordenan las columnas del dataframe.

**Se simplifica el listado de tipos de delitos para facilitar el análisis**

* Se mira cuántos tipos distintos de delitos existen.
* Hacer un análisis por tipo de delito es inviable por lo que se buscan los delitos más frecuentes entre los registros con mayor tasa de delitos.
* Se crea un diccionario que contiene el nombre de los tipos de delitos.
* Se crea una función llamada agrupar\_tipos\_delitos que renombra el nombre de los tipos de delitos.

**Fichero de los datos de las personas condenadas**

* Se obtienen los datos de las personas condenadas según sexo y edad.
* Se renombran y ordenan las distintas columnas.
* Se formatea la columna Numero\_Delitos y se actualiza los nulos por -1.
* Se renombran los nombres de las comunidades para unificarlos con el resto de los ficheros.
* Se formatea la columna Comunidades como una variable categórica.



**Se guarda la información en un fichero para tratar la información más tarde**

Se guarda el primer dataframe generado con la tasa de delitos en un fichero llamado datos\_delitos.csv dentro de la carpeta Datos Procesados.

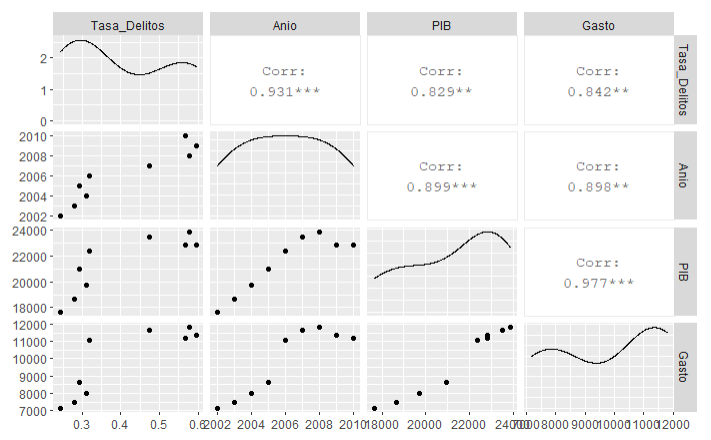
El segundo dataframe con los datos de las personas condenadas se guarda en la misma ubicación con el nombre datos\_condenados.csv

## **Análisis de los delitos**

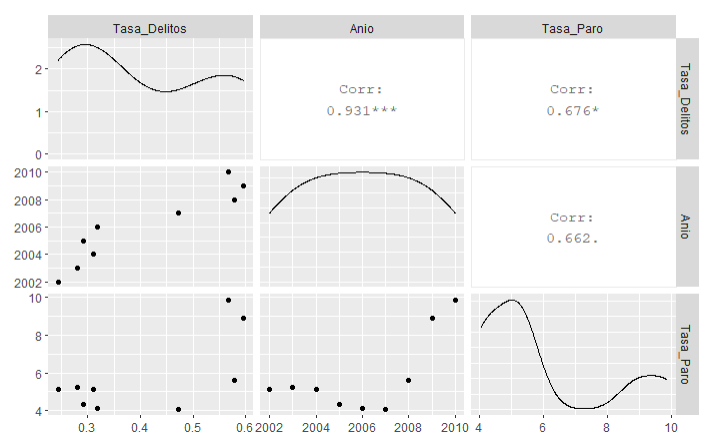
Antes de empezar con el modelo para analizar la dependencia de las variables se realiza un análisis exploratorio de los datos recogidos.

### **Inspección por variables**

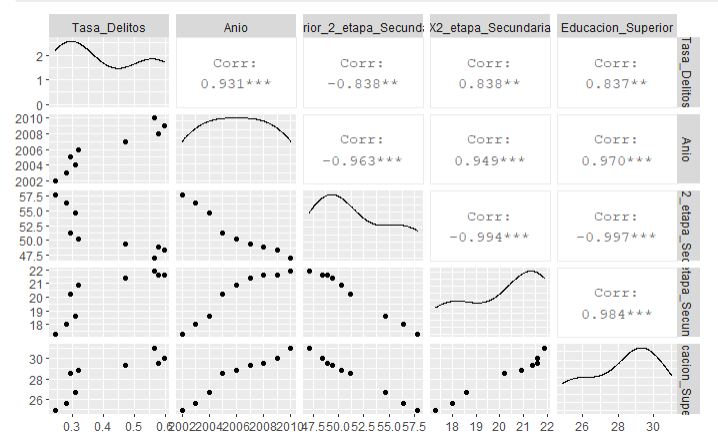
Inspección por variables de la distribución de la tasa de delitos en función de cada atributo visualmente.



Los gráficos sobre las variables Anio, PIB y Gasto se puede deducir decir que son relevantes ya que tienen una cierta tendencia de que cuando crece una de estas variables influye positivamente en la tasa de delitos.

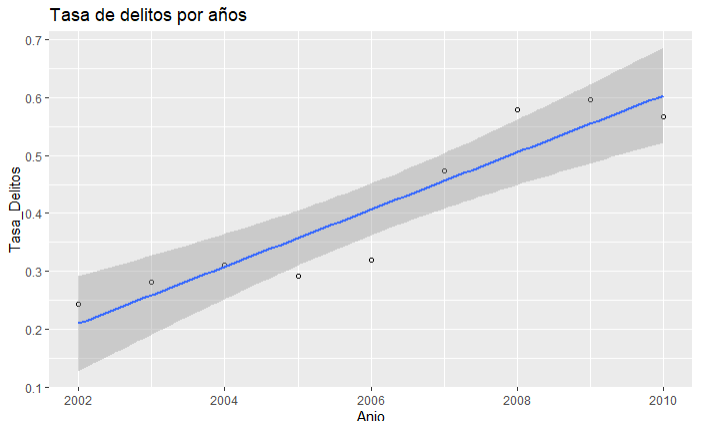


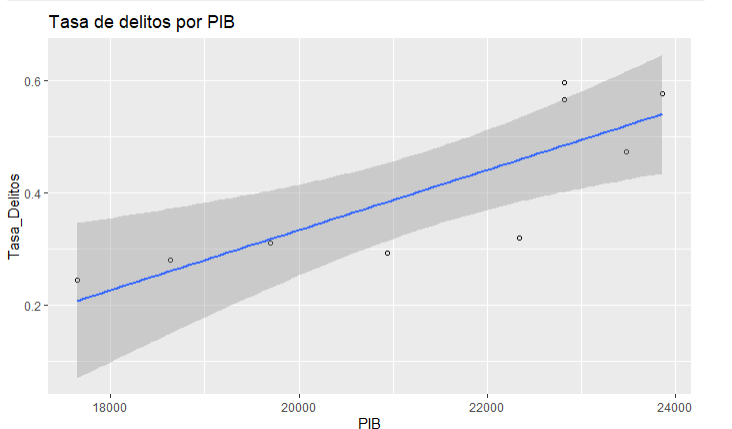
Con el gráfico sobre la tasa del paro no se puede decir lo mismo. No se ve que haya una relación lineal sobre la variable de de la tasa de delitos.

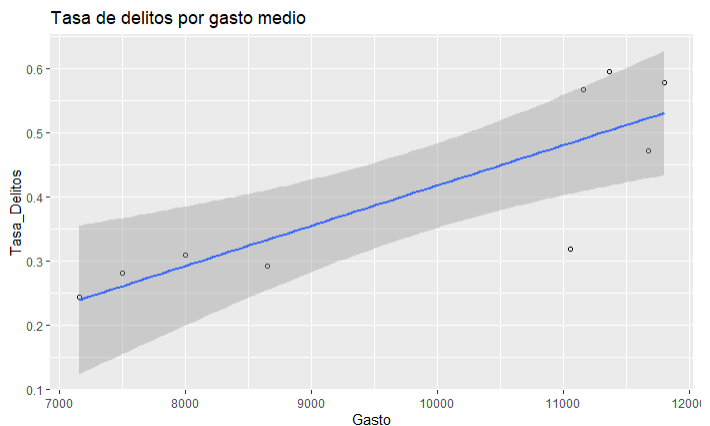


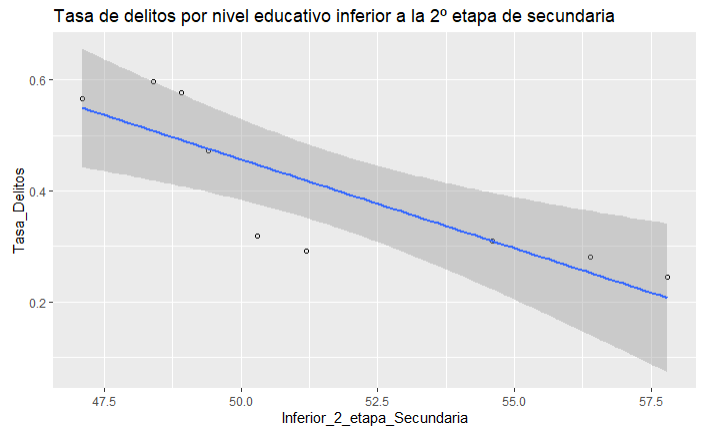
Sobre los gráficos de educación, también se ve que tiene una cierta dependencia negativa en el caso de la variable Inferior\_2\_etapa\_Secundaria y una dependencia lineal positiva con respecto a las otras dos variables: X2\_etapa\_Secundaria, Educacion\_Superior.

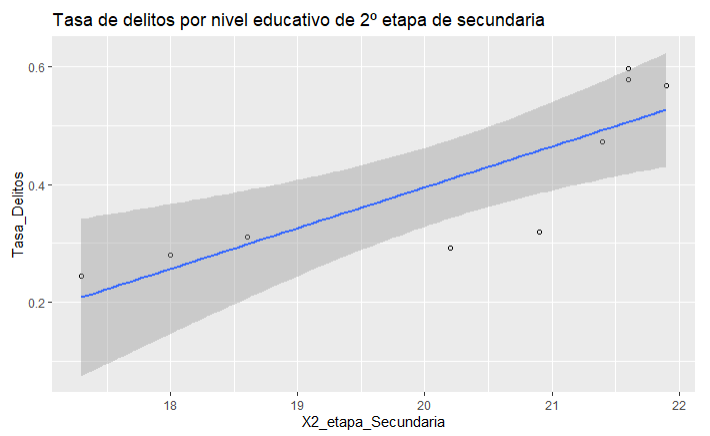
A continuación, se va a analizar la varianza y el sesgo de cada una de las variables con respecto a la variable de tasa de delitos:

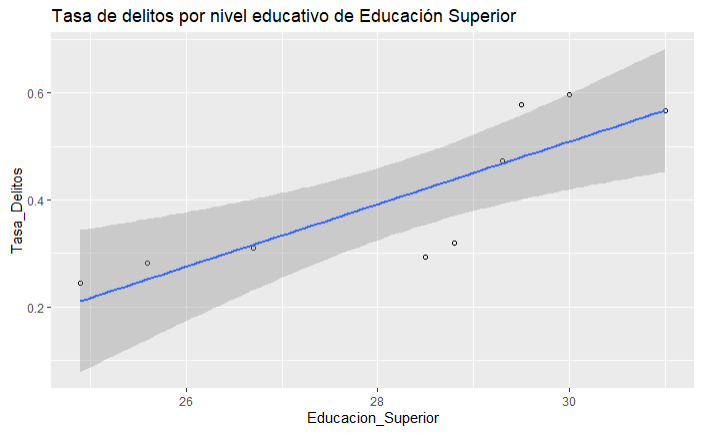










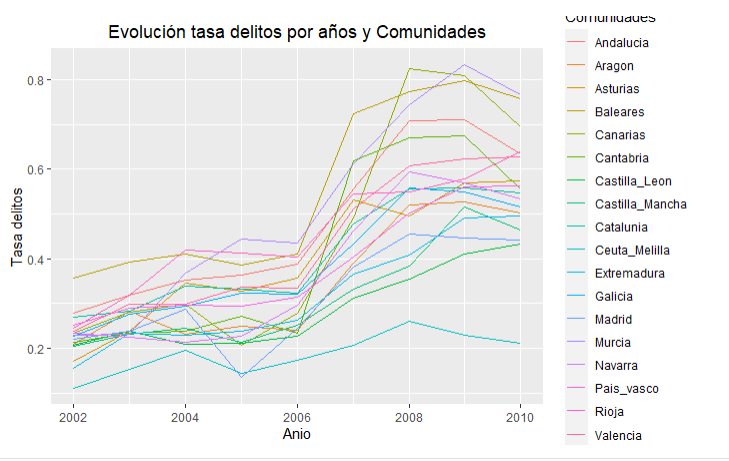


Todos ellos reflejan una alta varianza, y en alguno de ellos también se puede apreciar un alto sesgo como en los gráficos por PIB, por gasto medio y por nivel educativo de 2º etapa de secundaria.

Una vez realizada la inspección por variables se comienza con el análisis del conjunto de datos.

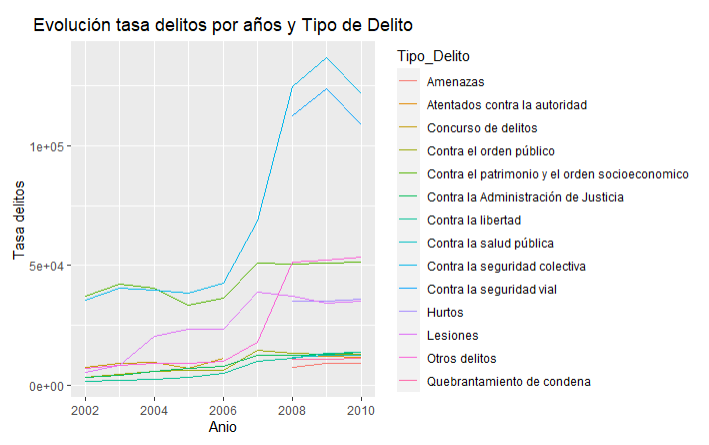
### **Análisis sobre los datos de los delitos**

En primer lugar, se hace un análisis de la evolución de la tasa de delitos por años y comunidades autónomas para todos los tipos de delitos:



Se puede apreciar que la mayoría de las comunidades tienen una tendencia a aumentar la tasa de delitos, aunque se ve un cierto aplanamiento de la curva a partir del año 2008. Ceuta y Melilla se alejan un poco de la media del resto de comunidades siendo las que menos delitos registran.

En el siguiente gráfico se visualiza la evolución de los delitos por años y tipo de delitos para todas las comunidades autónomas:



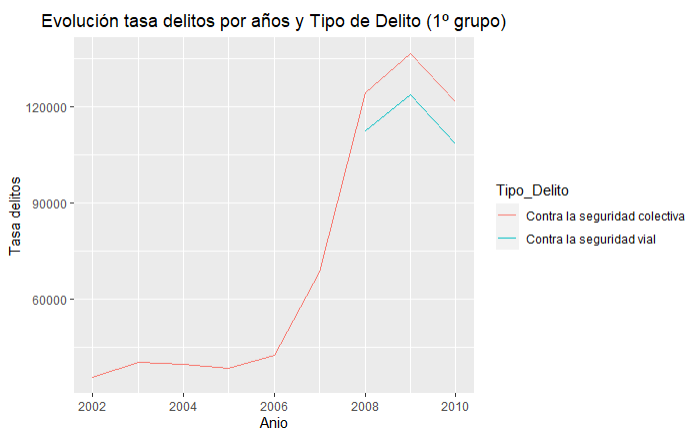
En el gráfico se puede apreciar 3 grupos que parecen seguir el mismo comportamiento.

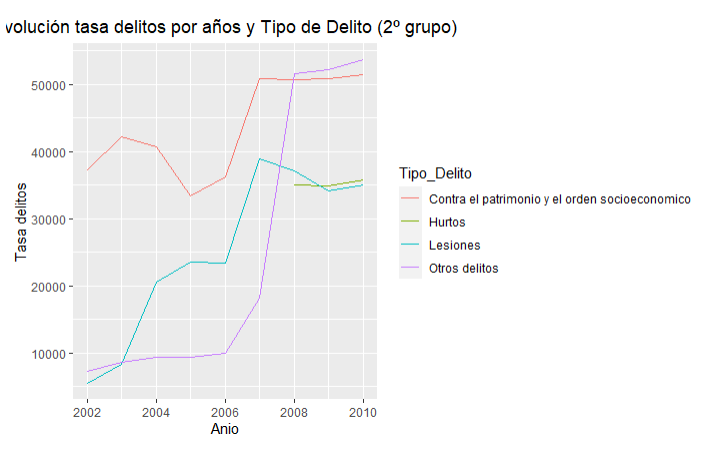
El primero de ellos que lo componen dos tipos de delitos que destacan claramente respecto al resto y que han crecido considerablemente desde 2006. Éstos son contra la seguridad colectiva y la seguridad vial, éste último añadido recientemente en 2008 como nuevo tipo de delito.

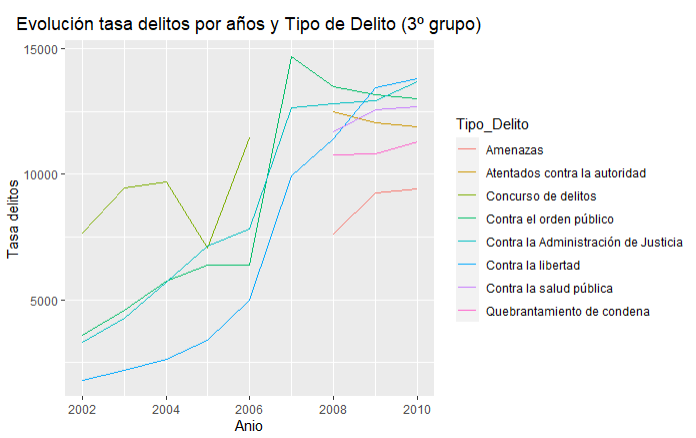
El segundo grupo se encuentra en la franja media compuesto por 4 tipos de delitos.

Por último, el tercer grupo se mantiene casi sin pendiente, manteniendo la misma tasa a lo largo de los años.

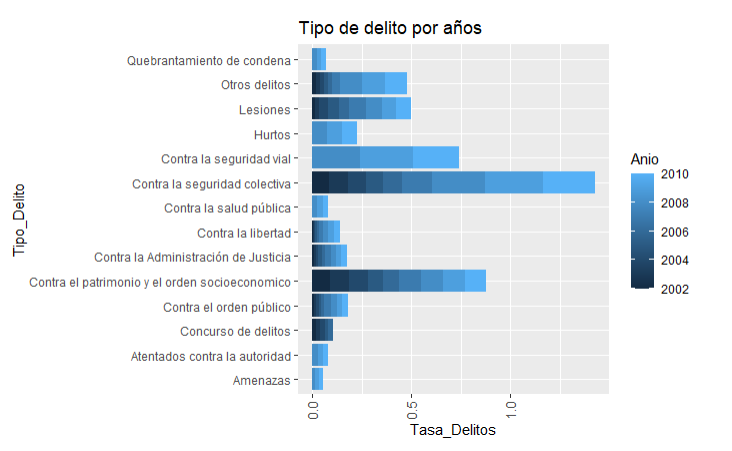
A continuación, se amplían estos tres grupos para poder analizarlos más en detalle.





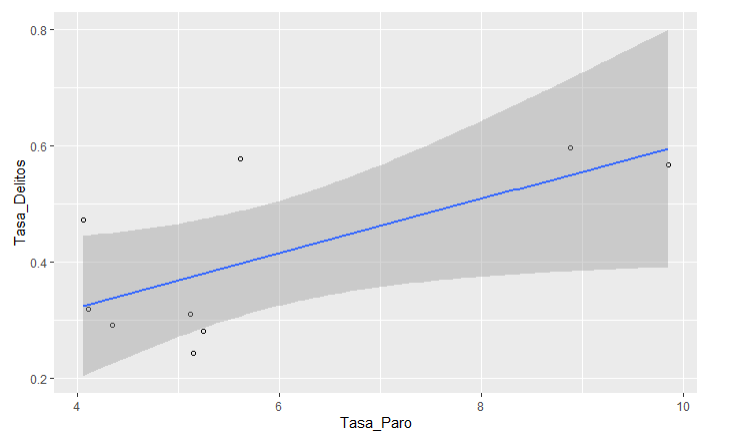


Se muestra otro tipo de visualización para analizar la evolución de cada uno de los tipos de delitos a lo largo de los años:



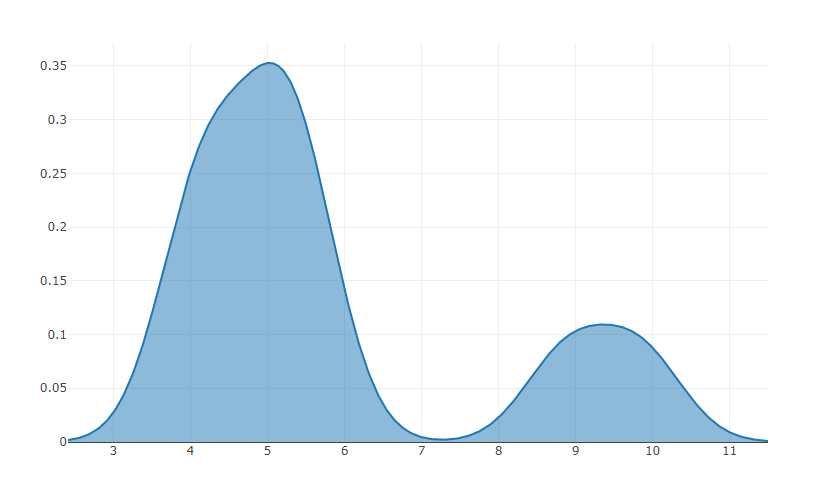
### **Análisis sobre los datos del paro**

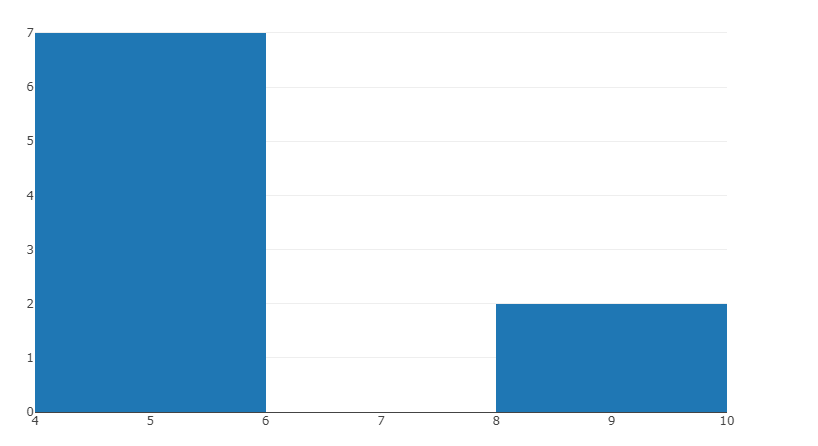
En la inspección de la variable tasa de delitos se ha visto que la tasa de paro no sigue una distribución lineal. A continuación, se analiza tanto la tasa de paro como la tasa de delitos siguen una distribución normal.



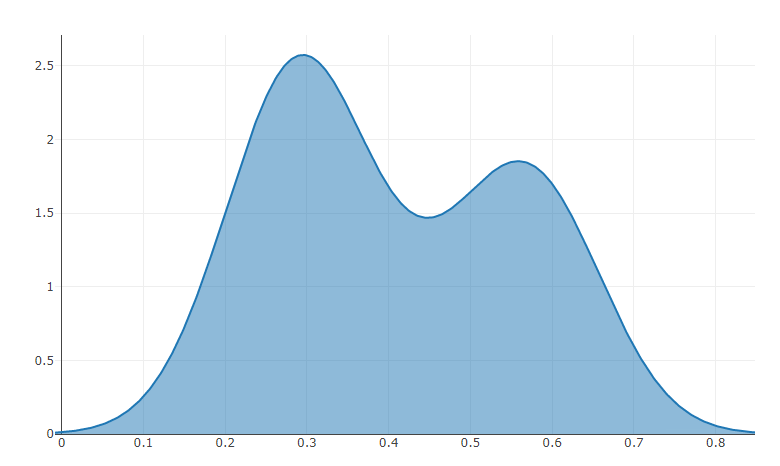
Como se ve en el gráfico, las dos variables no tienen una dependencia lineal.

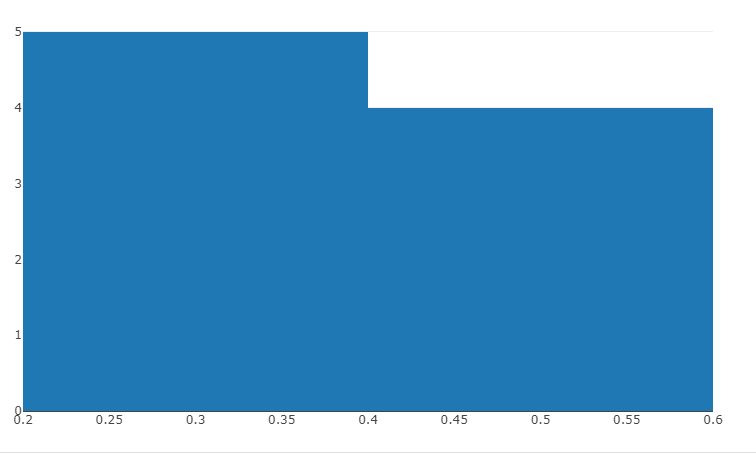
A continuación, se muestra la distribución y el histograma de los datos del paro:





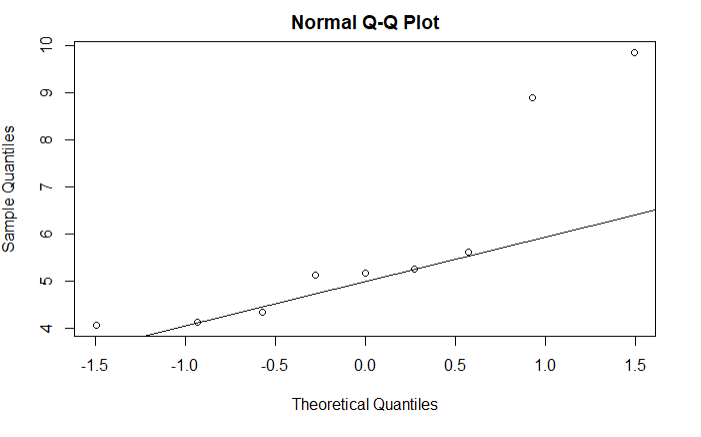
Y la distribución y el histograma de los datos del delito:

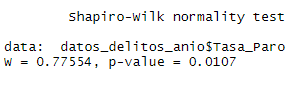




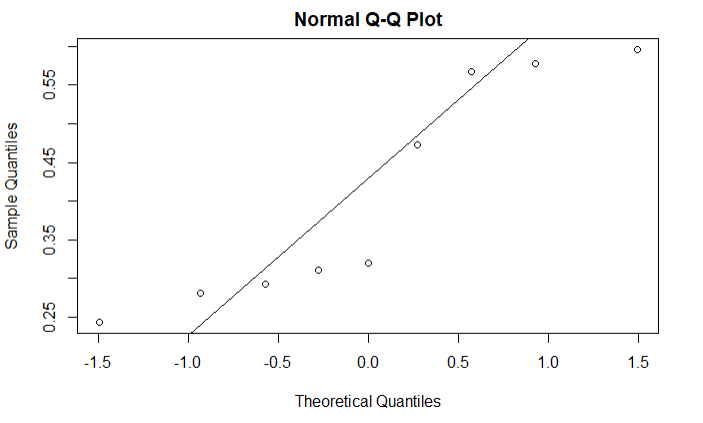
A primera vista no parece que ninguna tenga una distribución normal y aplicando el test de shapiro se comprueba que efectivamente ninguno de las dos variables se acerca a dicha distribución.

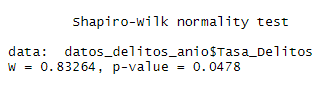
Test Shapiro sobre los datos del paro:



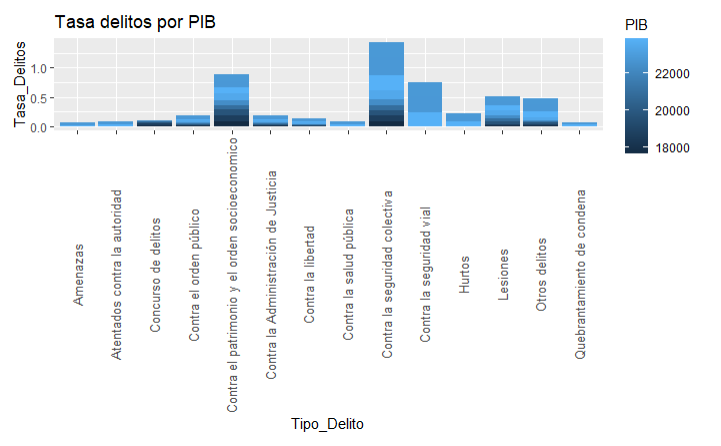


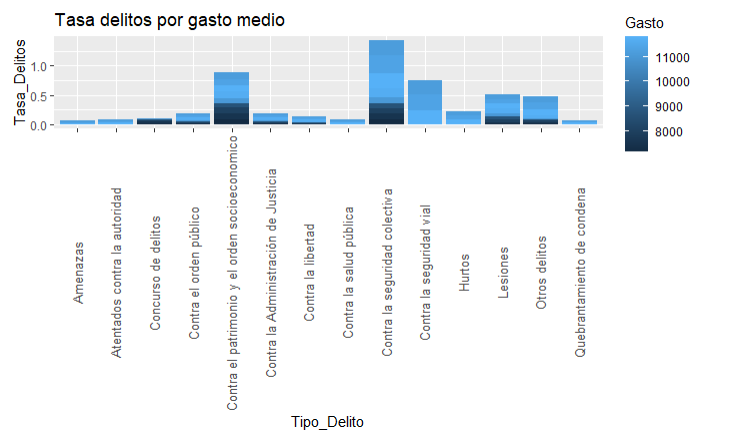
Test Shapiro sobre los datos de los delitos:

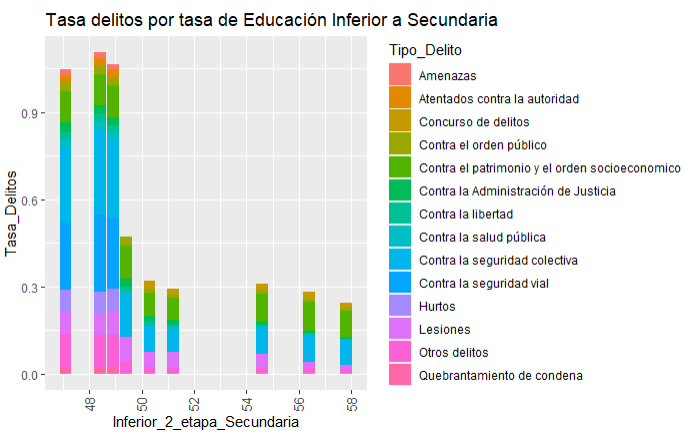


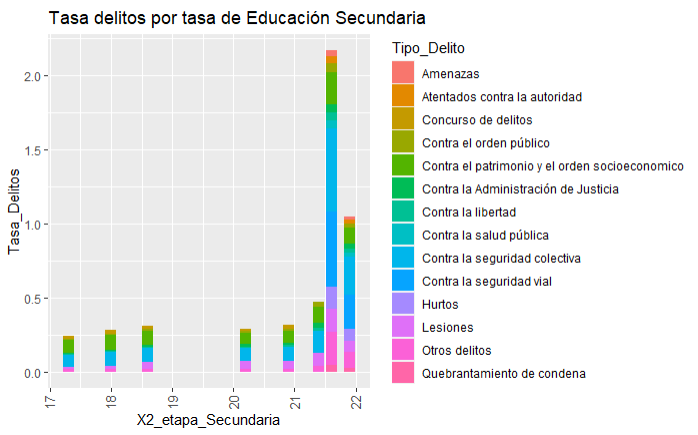


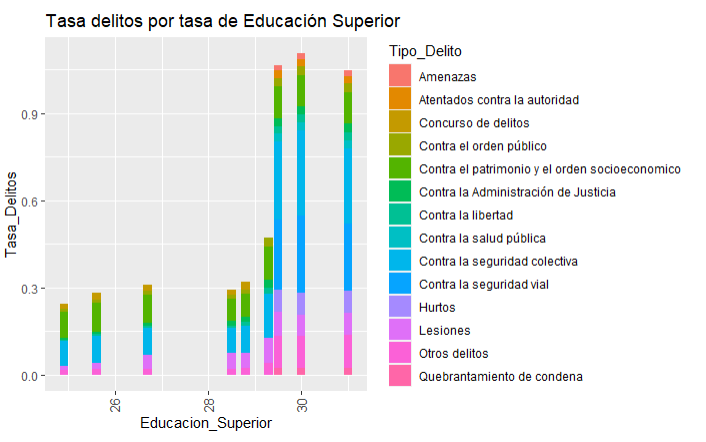
### **Análisis en función del tipo del delito**











## **Modelo Regresión Lineal**

El objetivo es encontrar el modelo que permita predecir con mayor precisión la tasa de delitos de España haciendo uso de técnicas de selección de variables y modelos regularizados.

Se debe de buscar un modelo cuya estructura sea lo más simple posible para que tenga una baja varianza, pero llegando a un equilibrio para que no suba demasiado el sesgo. Se intenta conseguir este equilibro estableciendo unos criterios estadísticos como los utilizados para este estudio:

* R2 ajustado, que es capaz de explicar la variabilidad de la variable respecto a los predictores penalizando la complejidad del modelo.
* Residual Sum of Squares RSS: la suma de cuadrados residuales es la suma de los cuadrados de los residuos. Es una medida de la discrepancia entre los datos y un modelo de estimación.
* AIC: se basa en la verosimilitud penalizando la complejidad del modelo.

Sin embargo, lo que realmente permite cuantificar cómo de útil es un modelo no es el test error, por lo que es esta medida en la que hay que basarse para elegir entre modelos con diferente número de predictores. Generalmente, al hablar de test error se hace referencia al test mean square error (test-MSE), que equivale al test RSS dividido por el número de observaciones MSE=RSS / n.

Durante el estudio se van a comparar los siguientes métodos de regresión: ordinary least squares (OLS), Subset Selection, Ridge, Lasso y Elastic Net.

Antes de comenzar se obtienen los datos que nos interesan del dataframe. No se estudia el desglose por tipos de delitos, ya que los datos del PIB, el gasto medio y el paro sólo se encuentran a nivel de comunidades. Eliminamos esta variable porque no es influyente para el estudio.

Se divide aleatoriamente el set de datos en dos grupos, uno se empleará para entrenar los modelos (datos\_delitos\_espana\_train) y el otro para validarlos (datos\_delitos\_espana\_test).

### **Técnicas de selección de variables**

Se utilizan dos técnicas que posteriormente se van a comparar para ver cuál de ellas obtiene el mejor modelo.

* StepAIC del paquete MASS. Automatiza la selección de modelos con la función step. Se elige la dirección both que empieza en el modelo base hasta el más complejo y luego realiza el proceso al revés, seleccionando por pasos el mejor modelo en función de los atributos elegidos usando el criterio AIC.
* Best Subset Selection del paquete leaps. Se usa cuando interesa encontrar subconjuntos de variables dependientes para optimizar las características de un modelo.

#### **Modelo OLS**

Se comienza con el método de regresión más simple de todos, ordinary least squares (OLS).

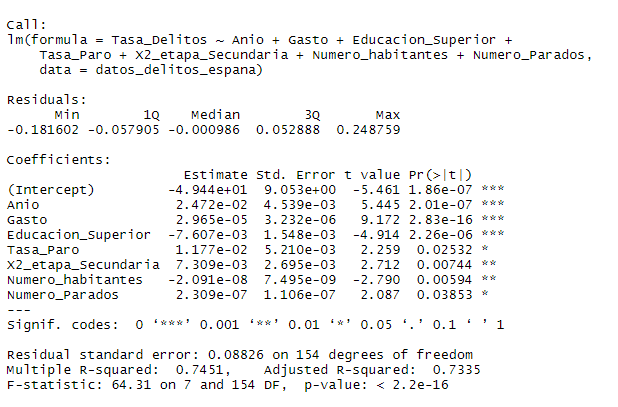
Para seleccionar las características para obtener el mejor modelo se usa el método StepAIC de la librería MASS que automatiza la selección de modelos con la función step.

Se elige la dirección both que empieza en el modelo base hasta el más complejo y luego realiza el proceso al revés, seleccionando por pasos el mejor modelo en función de los atributos elegidos usando el criterio AIC.

Se obtiene el siguiente modelo:

*lm(formula = Tasa\_Delitos ~ Anio + Gasto + Educacion\_Superior + Tasa\_Paro + X2\_etapa\_Secundaria + Numero\_habitantes + Numero\_Parados, data = datos\_delitos\_espana)*

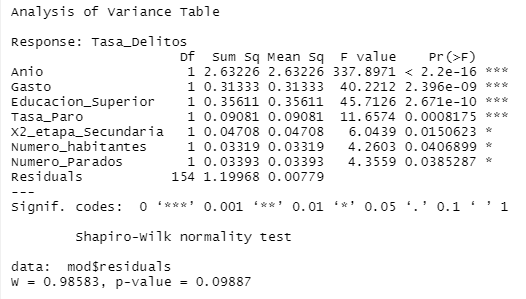
Se obtienen los resultados del modelo obtenido:

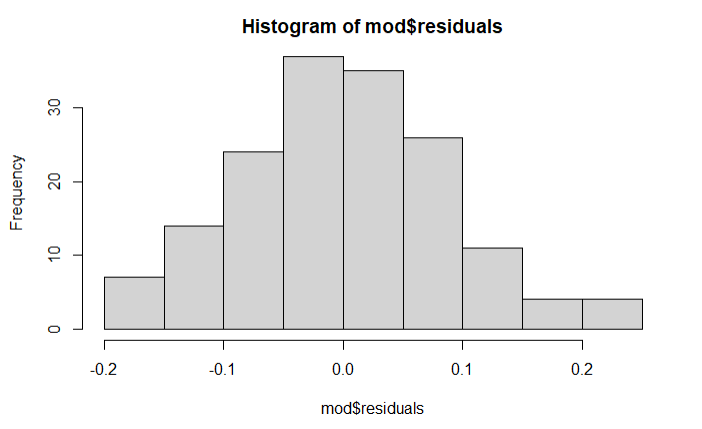


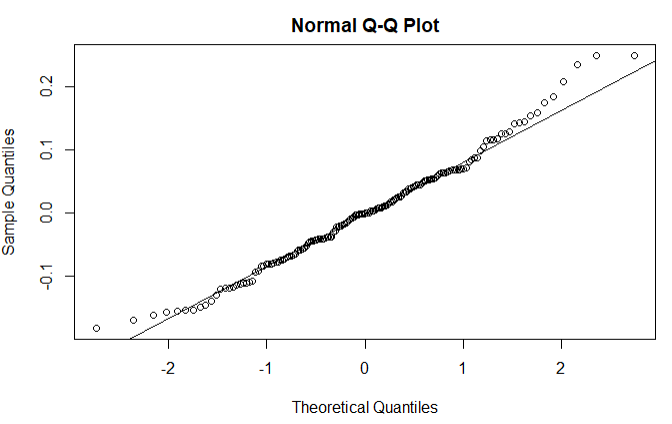
Con el valor de p-value < 2.2e-16 -se puede decir que existe relación lineal.

El R2 nos indica que aproximadamente el 75% de la variabilidad en la variable Tasa\_Delitos es explicada por las variables Anio, Gasto, Educacion\_Superior, Tasa\_Paro, X2\_etapa\_Secundaria, Numero\_habitantes, Numero\_Parados.

A continuación, se verifica el ajuste de este modelo, pintando sus residuos y verificando que sean normales, viendo su AIC. En este caso los residuos tienen un fuerte sesgo hacia la derecha que indica que el modelo no está siendo muy adecuado.







Se obtienen las predicciones del modelo y se guardan en el dataframe datos\_delitos\_espana con la predicción de la tasa de delitos en la columna Tasa\_Delitos\_Pred\_OLS.

Se calcula el valor de test\_MSE y se guarda en test\_MSE\_OLS para luego compararlo con el resto de los modelos.

#### **Modelo Best Subset Selection mediante k-cross-validation**

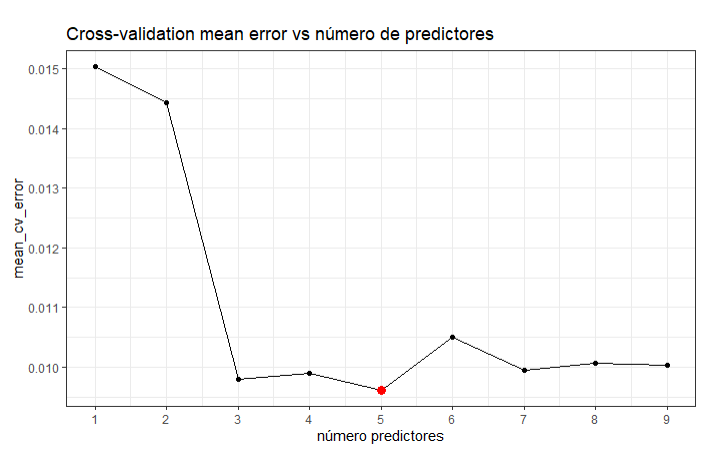
Los métodos conocidos como subset selection de la librería leaps tienen la finalidad de identificar y seleccionar, de entre todos los predictores disponibles, aquellos que están más relacionados con la variable respuesta y así crear el mejor modelo. Dentro de este grupo se diferencian: best subset selection y stepwise selection (forward, backward e hybrid). Al igual en el método de Step, para un mismo conjunto de datos, no todos tienen por qué converger en un mismo modelo final.

El esquema general de los métodos de subset selection consiste en:

* Crear un conjunto de modelos, todos los posibles (best subset) o bien un conjunto de ellos (stepwise), mediante diferentes combinaciones de los predictores disponibles.
* Para cada posible estructura de modelo se selecciona el mejor basándose en el estimador del error.
* Los modelos finalistas de cada tamaño se comparan entre ellos para identificar el mejor basándose en la estimación del test error.

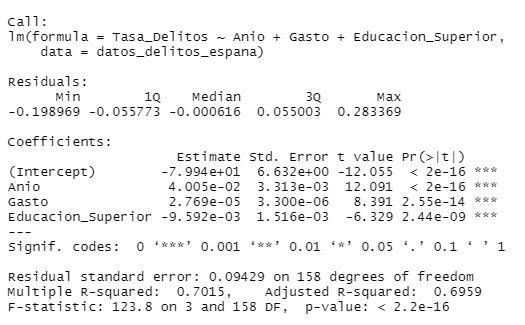
Para este estudio se va a utilizar el Best Subset Selection consiste en evaluar todos los posibles modelos que se pueden crear por combinación de los predictores disponibles.

Se validará el modelo con el método K-Cross-Validation para evaluar los resultados del análisis y garantizar que son independientes de la partición entre datos de entrenamiento y prueba, calculando la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones.



El modelo identificado mediante k-Cross-Validation es el formado por 5 predictores que es el que menor cv test error estimado tiene, el gráfico muestra que a partir de 3 predictores la mejora es mínima. Acorde al principio de parsimonia, según el cual se recomienda emplear de entre los modelos buenos el más simple, el modelo más adecuado es el de 3 predictores.

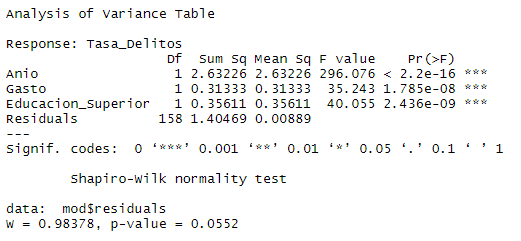
Se obtienen los resultados del modelo seleccionado con 3 variables:

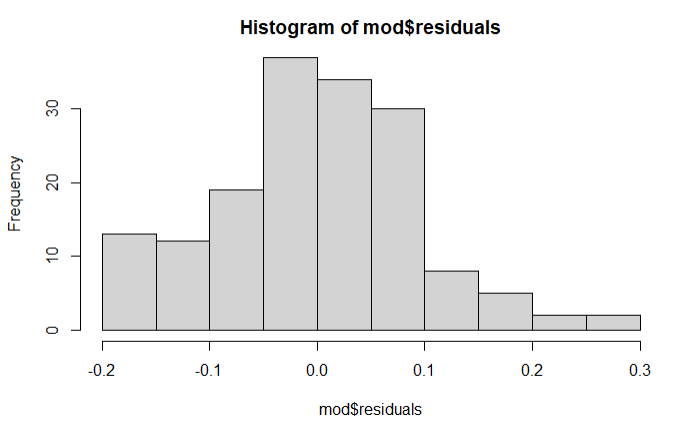


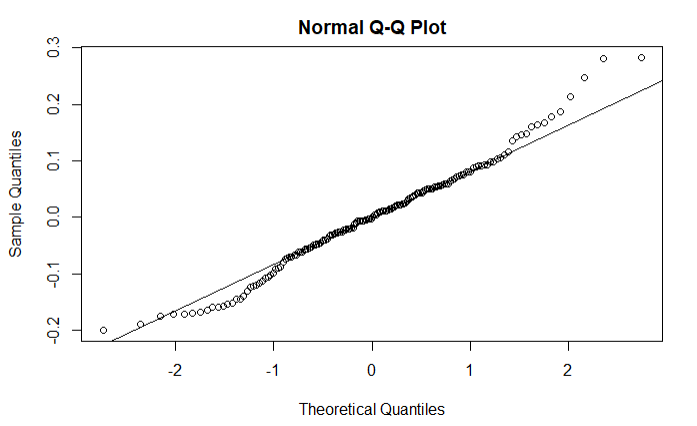
Con el valor de p-value < 2.2e-16 2.2e-16 se puede decir que existe relación lineal.

El R2 nos indica que aproximadamente el 70% de la variabilidad en la variable Tasa\_Delitos es explicada por las variables Anio, Gasto, Educacion\_Superior

A continuación, se verifica el ajuste de este modelo, pintando sus residuos y verificando que sean normales, viendo su AIC. En este caso los residuos tienen un fuerte sesgo hacia la derecha que indica que el modelo no está siendo muy adecuado.







Una vez empleado todas las observaciones de datos\_delitos\_espana\_train, y se calcula el test-MSE empleando el set de datos datos\_delitos\_espana\_test y se guarda en test\_MSE\_subset para luego compararlo con el resto de los modelos.

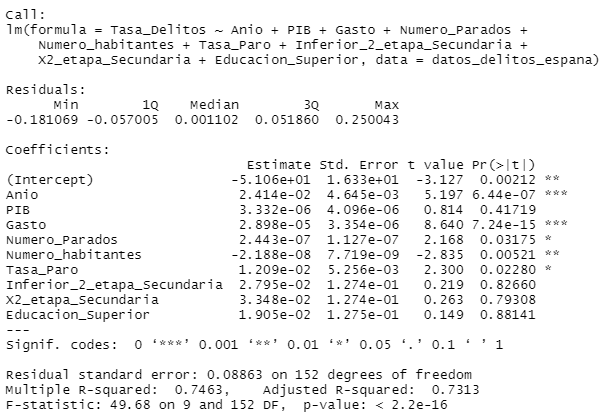
Se obtienen las predicciones del modelo y se guardan en el dataframe datos\_delitos\_espana con la predicción de la tasa de delitos en la columna Tasa\_Delitos\_Pred\_Subset.

### **Modelos regularizados**

#### **Modelo Ridge**

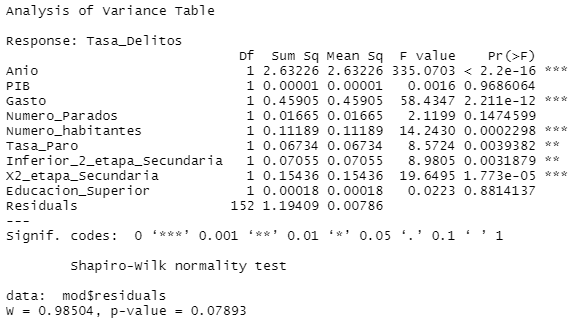
Este modelo produce una restricción continua sobre los valores de los coeficientes de cada variable que hace que estos sean más interesantes y potentes como modelos de predicción. El valor natural de los coeficientes es 0, penalizando la adjudicación de valor. En este modelo no se obtienen coeficientes finales nulos, aunque sean muy pequeños.

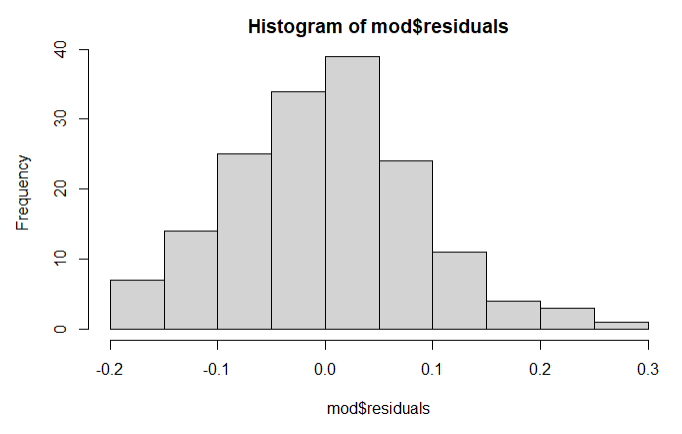
Se analizan los datos del modelo obtenido:

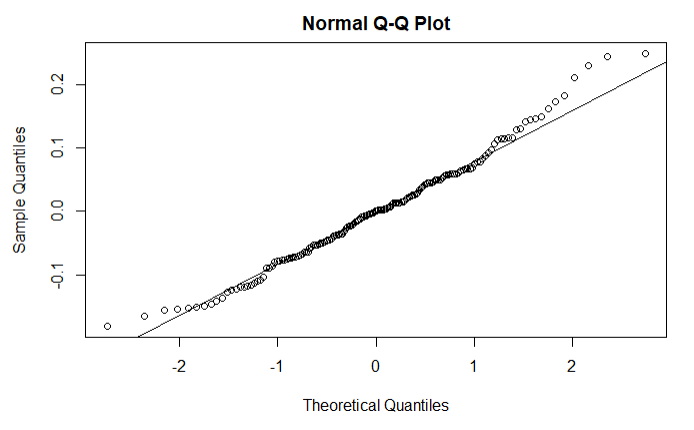


El R2 nos indica que aproximadamente el 75% de la variabilidad en la variable Tasa\_Delitos es explicada por las variables Anio, PIB, Gasto, Numero\_Parados, Numero\_habitantes, Tasa\_Paro, Inferior\_2\_etapa\_Secundaria, X2\_etapa\_Secundaria, Educacion\_Superior.

SA continuación, se verifica el ajuste de este modelo, pintando sus residuos y verificando que sean normales, viendo su AIC. En este caso los residuos tienen un fuerte sesgo hacia la derecha que indica que el modelo no está siendo muy adecuado.







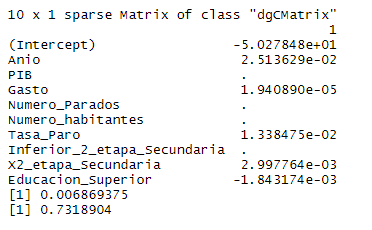
Se calcula el test-MSE empleando el set de datos datos\_delitos\_espana\_test y se guarda en test\_MSE\_ridge para luego compararlo con el resto de los modelos.

Se obtienen las predicciones del modelo y se guardan en el dataframe datos\_delitos\_espana con la predicción de la tasa de delitos en la columna Tasa\_Delitos\_Pred\_Ridge.

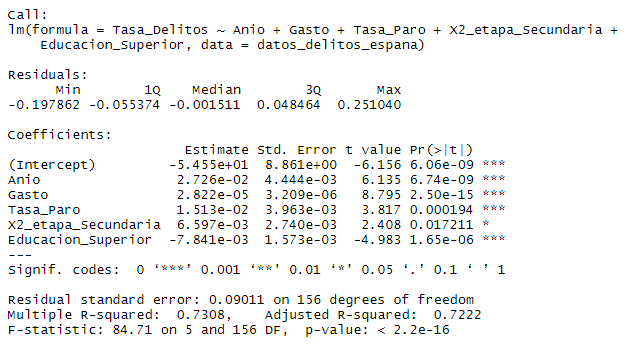
#### **Modelo Lasso**

Este modelo produce una restricción continua sobre los valores de los coeficientes de cada variable que hace que estos sean más interesantes y potentes como modelos de predicción. El valor natural de los coeficientes es 0, penalizando la adjudicación de valor. A diferencia del modelo de Ridge este modelo establece coeficientes finales con valor nulo (los que no son importantes).

Resultado del modelo:

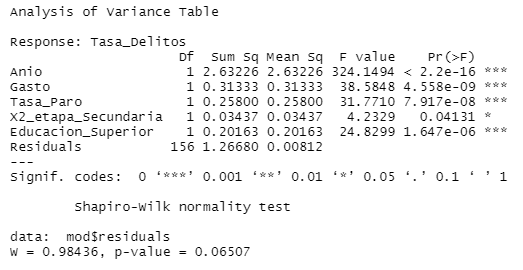


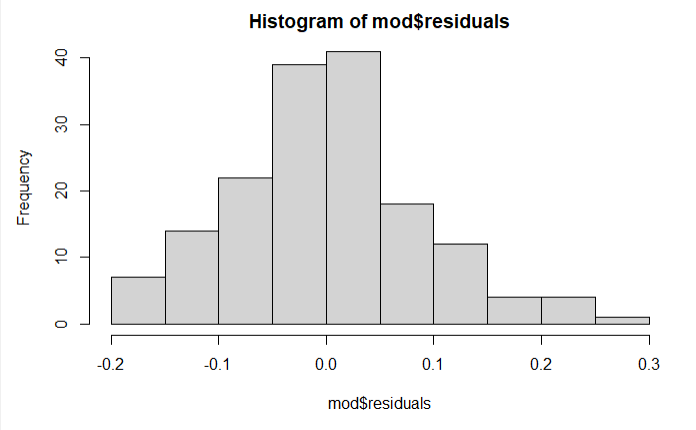
Se analizan los datos del modelo obtenido con las 5 variables seleccionadas.

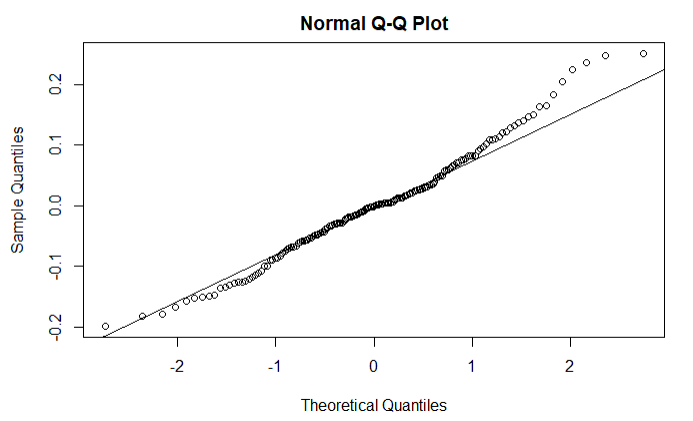


El R2 nos indica que aproximadamente el 73% de la variabilidad en la variable Tasa\_Delitos es explicada por las variables Anio, Gasto, Tasa\_Paro, X2\_etapa\_Secundaria, Educacion\_Superior.

Como en el resto de modelos, se verifica el ajuste de este modelo, pintando sus residuos y verificando que sean normales, viendo su AIC. En este caso los residuos tienen un fuerte sesgo hacia la derecha que indica que el modelo no está siendo muy adecuado.







Se calcula el test-MSE empleando el set de datos datos\_delitos\_espana\_test y se guarda en test\_MSE\_lasso para luego compararlo con el resto de los modelos.

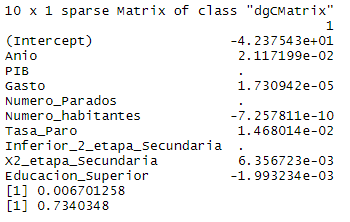
Se obtienen las predicciones del modelo y se guardan en el dataframe datos\_delitos\_espana con la predicción de la tasa de delitos en la columna Tasa\_Delitos\_Pred\_Lasso.

#### **Modelo Ridge**

Este modelo es un término medio entre el modelo de Ridge y el de Lasso. Esto permite generar un modelo en el que solo algunos de los coeficientes sean no nulos, manteniendo las propiedades de regularización de Ridge. El parámetro alpha regula el peso dado a la regularización impuesta por Ridge y por Lasso. Desde este punto de vista Elastic Net es un superconjunto de ambos modelos.

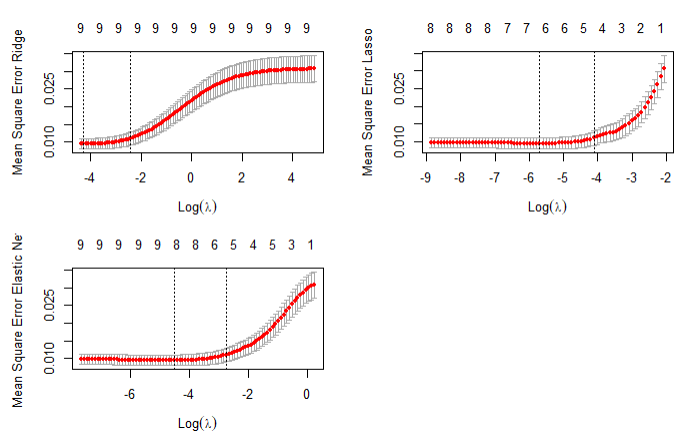
Para determinar el valor de alpha se ha probado con varios valores, obteniendo cuyo valor R2 sea mayor, en este caso para alpha = 0.1 R2 es 0.7340348.

El resultado obtenido es el siguiente:



En este caso el R2 nos indica que aproximadamente el 73% y no se hace análisis de la distribución del modelo porque tiene los mismos parámetros que el modelo de Lasso.

Para los métodos de Ridge, Lasso y Elastic Net consiguen, empleando sus respectivos valores óptimos de λ, reducir el MSE (test error) a unos niveles muy parecidos. La ventaja del modelo final obtenido por Lasso Y Elastic Net es que es mucho más simple ya que contiene únicamente 5 predictores.



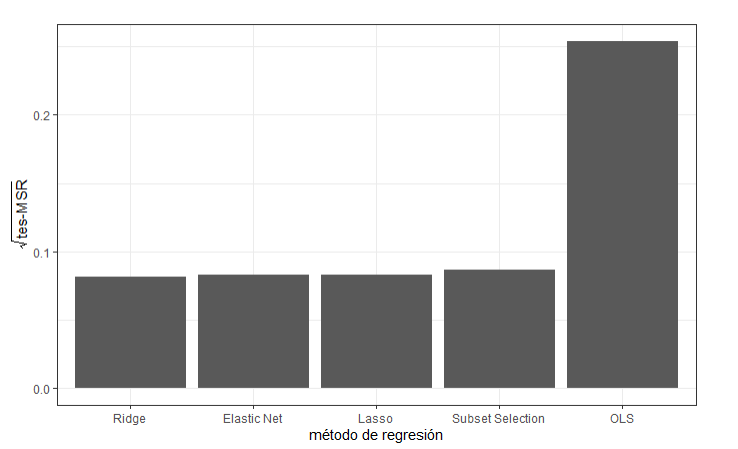
Se calcula el test-MSE empleando el set de datos datos\_delitos\_espana\_test y se guarda en test\_MSE\_elastic para luego compararlo con el resto de los modelos.

Se obtienen las predicciones del modelo y se guardan en el dataframe datos\_delitos\_espana con la predicción de la tasa de delitos en la columna Tasa\_Delitos\_Pred\_Elastic.

### **Comparación de Modelos**

Se comparan los resultados de los valores de test\_MSE obtenido de los distintos modelos.





#### **Conclusión**

Viendo la distribución de los residuos de cada modelo, se puede decir que ninguno de los modelos obtenidos son demasiado buenos como para predecir la tasa de delincuencia.

Aun así, entre los resultados obtenidos, el que consigue mayor precisión (menor test-MSE) es Ridge regression. Cabe destacar el modelo OLS, donde se obtiene un elevado test-MSE, siendo el peor modelo de todos con diferencia.

Para poder comparar las predicciones realizadas de cada uno de ellos se ha añadido una columna al dataframe datos\_delitos\_espana con las predicciones de cada uno de los modelos, ya mencionado anteriormente en el documento, aunque como ya se ha comentado, ninguno de ellos es lo bastante fiable.

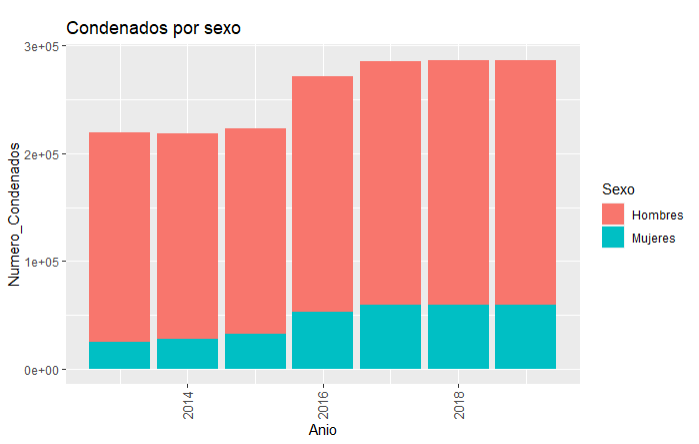
## **Análisis de condenados**

Se realiza un estudio sobre los perfiles de las personas condenadas desde el año 2013 al 2019, a través de la información de la comunidad autónoma donde se produjo el crimen y la edad y el sexo del condenado. Se intentará identificar ciertos patrones que a simple vista son más difíciles de apreciar.

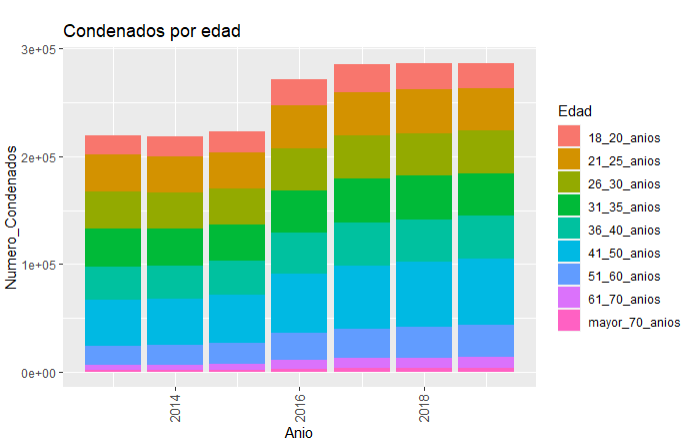
Los datos se obtienen del fichero datos\_condenados.csv de la carpeta Datos Procesados.

Antes de nada, se hace un análisis exploratorio de los datos:

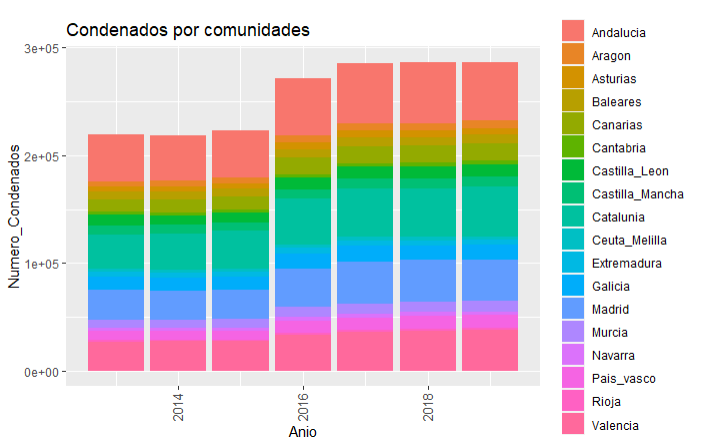
Se analizan los condenados por sexo:



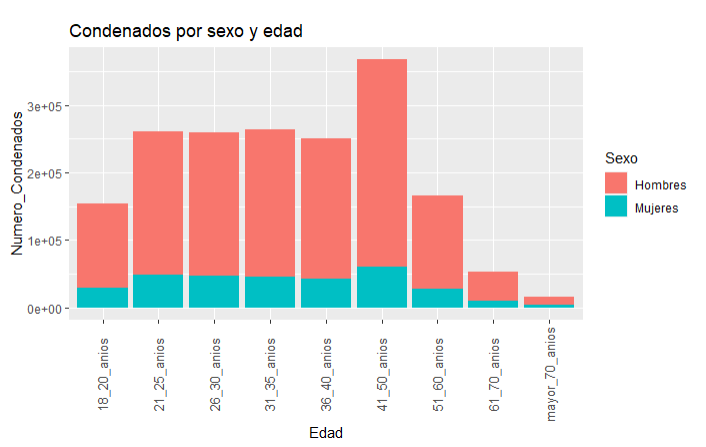
Se analizan los condenados por edad:



Se analizan los condenados por comunidad autónoma:



Se analizan los condenados por sexo y edad:



## **Clustering**

Se va a utilizar métodos de aprendizaje automático no supervisado para encontrar los distintos perfiles, a priori desconocidos.

Para ello se van a utilizar 3 tipos de clúster para poder compararlos más tarde:

* **Kmeans**: agrupa las observaciones en K clusters distintos, donde el número K se determina previamente a través del método del codo o silhouette.
* **Kmedoids**: es un método de clustering muy similar a K-means en cuanto a que ambos agrupan las observaciones en K clusters. La diferencia es que, en K-medoids, cada cluster está representado por una observación presente en el cluster (medoid), mientras que en K-means cada cluster está representado por su centroide, que se corresponde con el promedio de todas las observaciones del cluster pero con ninguna en particular. K-medoids es un método de clustering más robusto que K-means, por lo es más adecuado cuando el set de datos contiene outliers o ruido.
* **HClust**: es un método basado en el cluster jerárquico, el cual construye un árbol que representa las relaciones de similitud entre los distintos elementos.

Todos los métodos de clustering tienen que determinar qué distancia se emplea dentro del contexto del clustering como cuantificación de la similitud o diferencia entre observaciones.

En este caso se utilizará la distancia euclídea que define la longitud del segmento que une 2 observaciones.

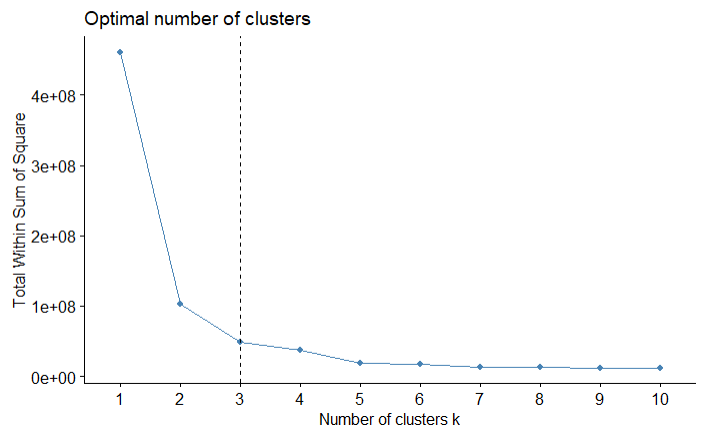
Antes de comenzar es necesario convertir los datos categóricos a dummies para que todas las variables estén representadas por un valor numérico.

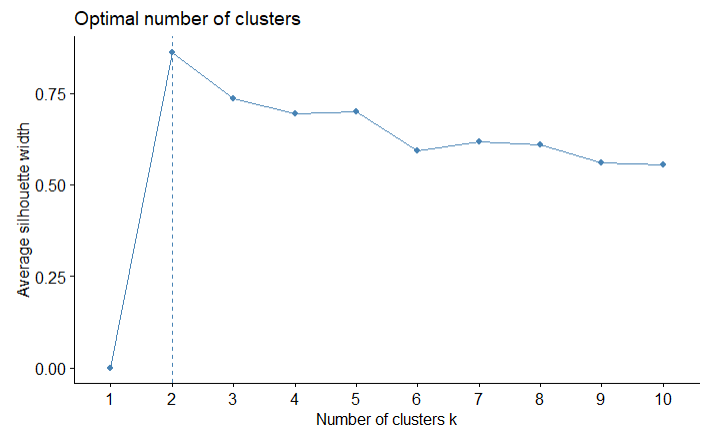
### **Obtención del número de clúster**

Se analiza el número de clústers necesarios para identificar los distintos perfiles para cada uno de los métodos. Para ello se usa la función fviz\_nbclust del paquete factoextra que automatiza el proceso del método del codo y de la silhouette.

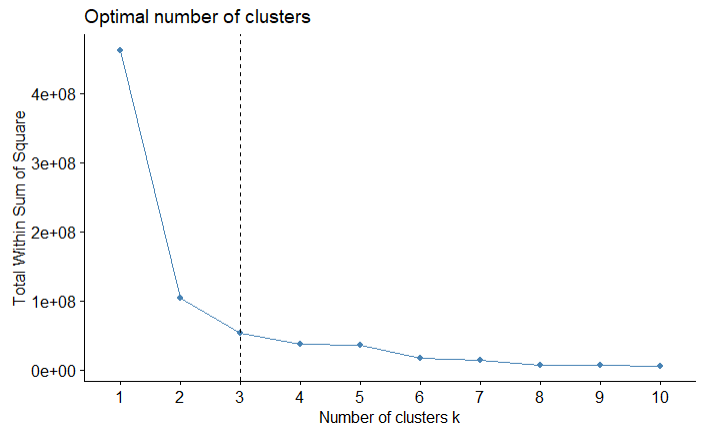
Dado que el método del codo no se aprecia bien el número de clusters seleccionados, se escoge el método de la silhouette para determinar los clusters. En este caso, para los 3 modelos, el número de clusters son 2.

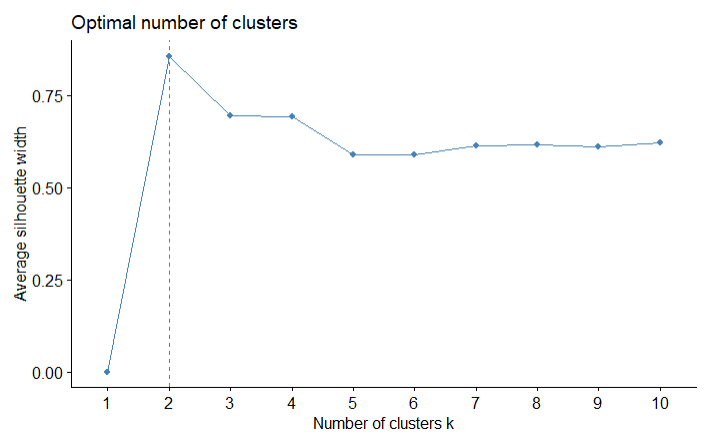
KMEANS



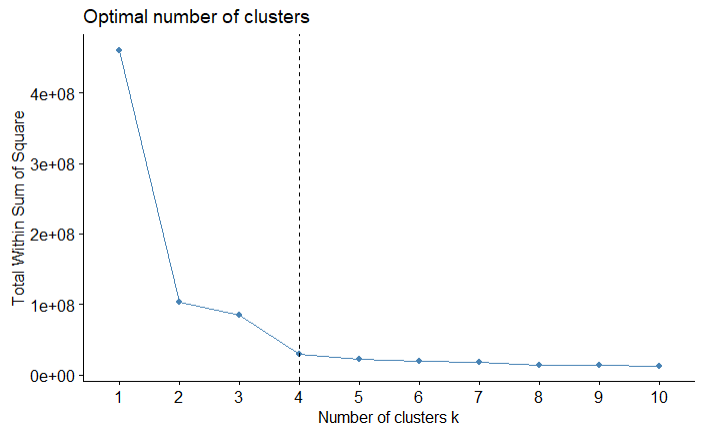


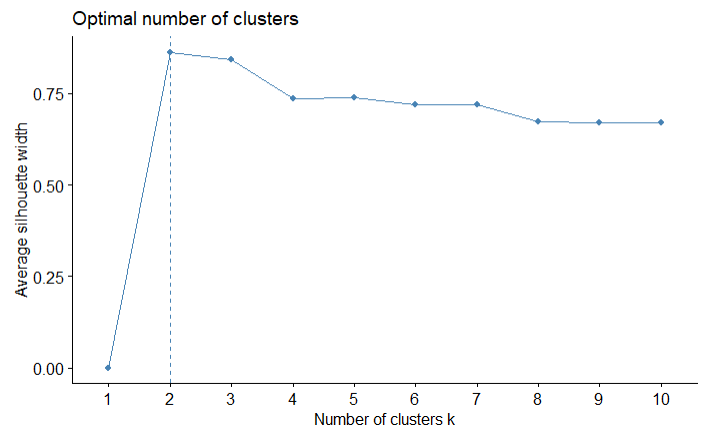
KMEDOIDS





HCLUST





### **Creación de los modelos**

Antes de comenzar con la creación de los modelos, se normalizan los datos para que todas las variables tengan el mismo peso en el análisis.

Se crean los 3 modelos de clúster y se guarda el grupo de la segmentación al que pertenece cada registro en las distintas columnas del dataset datos\_condenados\_dummies\_2013 (Segmentos\_kmeans, Segmentos\_kmdeoids, Segmentos\_hclust)

**KMEANS**

Modelo\_kmeans=kmeans(datos\_condenados\_dummies\_2013\_nor,NUM\_CLUSTERS)

datos\_condenados\_dummies\_2013$Segmentos\_kmeans=Modelo\_kmeans$cluster

**KMEDOIDS**

Modelo\_kmedoids=pam(datos\_condenados\_dummies\_2013\_nor,NUM\_CLUSTERS, metric = "euclidean")

datos\_condenados\_dummies\_2013$Segmentos\_kmdeoids=Modelo\_kmedoids$cluster

**HCLUST**

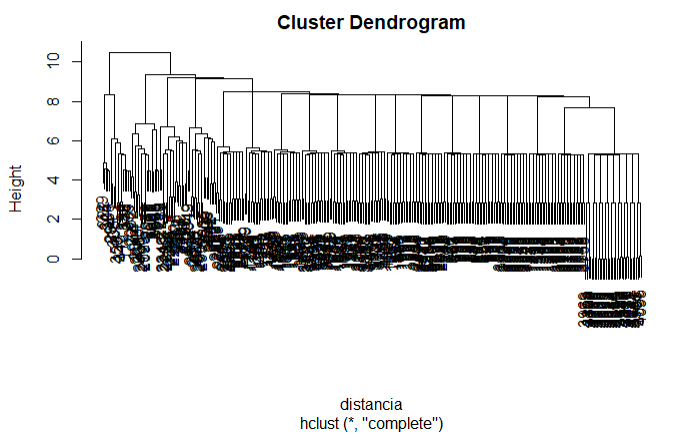
distancia <- dist(datos\_condenados\_dummies\_2013\_nor, method = "euclidean")

Modelo\_hclust<- hclust(distancia, method = "complete")

datos\_condenados\_dummies\_2013$Segmentos\_hclust = cutree(Modelo\_hclust, k = NUM\_CLUSTERS)

plot(Modelo\_hclust)

Se muestra el deograma generado por el modelo de HClust.



Los tamaños de cada clúster de cada uno de los 3 modelos resultantes son los siguientes:

Kmeans:

* Grupo 1: 264
* Grupo 2: 78

Kmedoids:

* Grupo 1: 171
* Grupo 2: 171

HClust:

* Grupo 1: 18
* Grupo 2: 324

### **Resultados**

A continuación, se obtienen los perfiles de cada grupo de cada uno de los modelos analizados.

Para el año 2013 se observan los siguientes perfiles:

KMEANS:

* Grupo 1: Perfil repartido por todas las comunidades menos en Valencia, con un porcentaje menor en Ceuta y Melilla. El mismo porcentaje de hombres y mujeres, de todas las edades menos en las franjas de los 61 a 70 años.
* Grupo 2: Perfil repartido por todas las comunidades por igual, con un porcentaje menor en Ceuta y Melilla. El mismo porcentaje de hombres y mujeres, de todas las edades con un elevado porcentaje del 70% en las edades comprendidas entre los 61 y 70.

KMEDOIDS:

* Grupo 1: Perfil repartido por todas las comunidades por igual, con un mayor porcentaje en Ceuta y Melilla. Predomina un mayor porcentaje de hombres de todas las edades, excepto mayores de 70 años, aunque predominan las edades de entre 26 y 30 años.
* Grupo 2: Perfil repartido por todas las comunidades por igual, con un mayor porcentaje en Ceuta y Melilla. Predomina un mayor porcentaje de mujeres de todas las edades, excepto las edades comprendidas entre 26 y 30 años, aunque predomina la edad mayor de 70 años.

HCLUST:

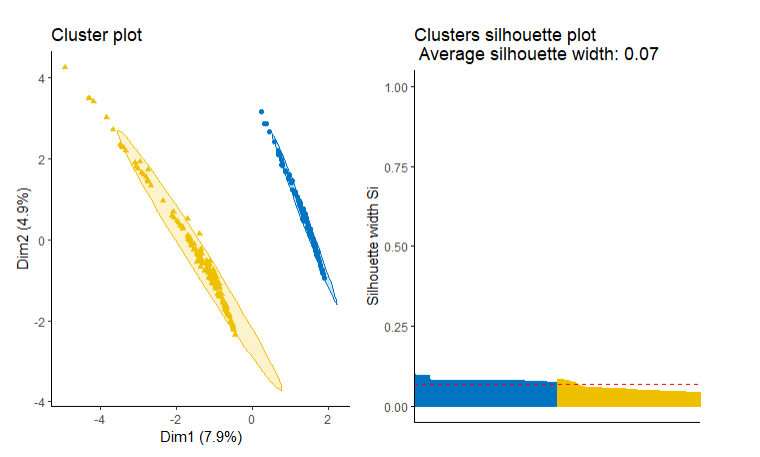
* Grupo 1: Perfil centrado en Andalucía con el mismo porcentaje de hombres y mujeres por igual, repartido entre todas las edades.
* Grupo 2: Perfil repartido por igual en todas las comunidades excepto en Andalucía con el mismo porcentaje de hombres y mujeres y repartido entre todas las edades.

### **Validación de los clústers**

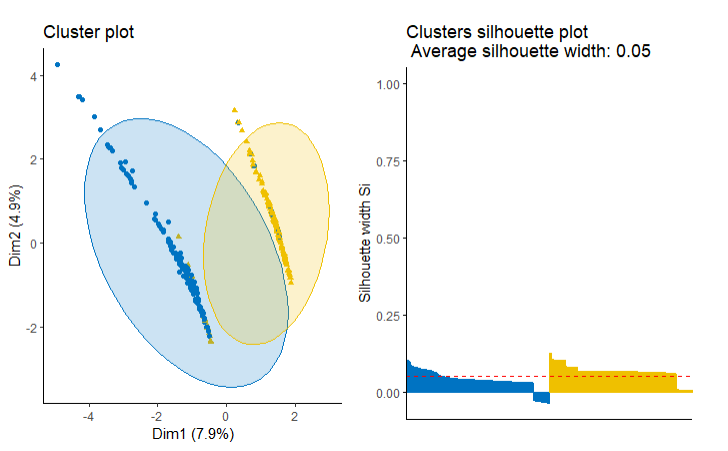
Se puede validar los clústers resultantes a través de los coeficientes silhouette. Esto se consigue con la combinación de las funciones eclust y fviz\_silhouette del paquete factoextra.

Tal y como se puede observar en las gráficas, el modelo que realiza la agrupación más correcta es la de Kmeans.

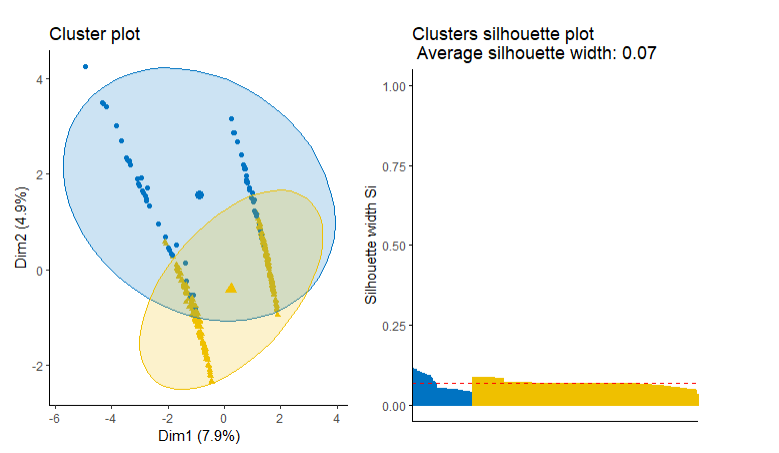
Validación KMEANS



Validación KMEDOIDS



Validación HCLUST



# **CONCLUSIÓN**

Una vez realizado el estudio se ha visto que los datos proporcionados no son suficientes o lo bastante buenos como para poder obtener un modelo válido para la predicción de la tasa de delincuencia. En cambio, sí se ha podido observar cómo el modelo de Kmeans ha conseguido obtener dos perfiles definidos de los criminales.

Aunque este estudio no ha finalizado con los resultados esperados, al menos sirve para rechazar la hipótesis inicial sobre las variables influyentes en el índice de delincuencia en España y concluir con un par de perfiles de delincuentes válidos.

Por otro lado, me ha servido para aprender nuevas técnicas, ver más en profundidad funciones, validaciones y métodos, y descubrir nuevas librerías de R.

En el ámbito de los modelos no supervisados he aprendido nuevos métodos como HCluster y Kmedoids y nuevas validaciones para validar esos modelos como el uso de los coeficientes silhouette a través de la librería factoextra.

En cuanto al análisis de regresión lineal, he visto nuevas técnicas y métodos como el Best Subset Selection y Elastic Net bastantes útiles para obtener un modelo de regresión lineal y cómo obtener las predicciones de una manera automática para calcular el valor de test\_MSE necesario para la comparación de modelos.

En definitiva, me ha gustado bastante el desarrollo realizado. Me servido para enfrentarme a un proyecto de Data Science de principio a fin, planteando un problema, recogiendo los datos iniciales, analizando cada una de las variables por separado y en su conjunto y finalmente realizar dos estudios de machine learning, uno sobre un modelo supervisado y otro sobre un modelo no supervisado.