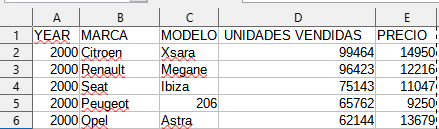
**Análisis de datos en R**

Alfredo Talavera Ramajo

### Especialización Inteligencia Artificial y Big Data · 03/2022

Esta práctica consistirá en el análisis de los datos de los 5 coches por año más vendidos desde el 2000 hasta el 2014

En primer lugar crearemos un archivo *.csv* con el cual trabajaremos mas tarde



\* Un ejemplo del *.csv* creado en excel, el cual tiene como columnas: El año de fabricación, la marca del coche, el modelo, las unidades que vendió en ese año y el precio al que se vendió en ese año

Una vez creado el csv podemos pasar a R estudio

Para comenzar a trabajar necesitamos importar las librerías que vayamos a necesitar

# · Importamos las librerias necesarias

library(dplyr)

library(ggplot2)

library(reshape2)

library(plotrix)

install.packages("car")

library(car)

crPlots(model\_1)

install.packages("gapminder")

library(gapminder)

install.packages("devtools")

devtools::install\_github("gertstulp/ggplotgui")

library("ggplotgui")

# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- #

Vamos a leer el *.csv* y a guardarlo en una variable llamada datos

# ··· ··· ··· ··· ··· ··· ··· ··· ··· ··· Abrimos el archivo.csv y lo mostramos ··· ··· ··· ··· ··· ··· ··· ··· ··· ··· #

datos <- read.csv("coches.csv",sep = ",", header = TRUE)

datos

class(datos)

# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

Por preferencia personal se crearán 5 variables las cuales son las 5 columnas que tiene el *.csv* y también por comodidad a la hora de utilizarlas

# ··· ··· ··· ··· ··· ··· Nos creamos una variable para cada dato extraido de la columna ··· ··· ··· ··· ··· ··· #

X <- datos$YEAR

Y <- datos$PRECIO

Z <- datos$MARCA

V <- datos$UNIDADES.VENDIDAS

K <- datos$MODELO

# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

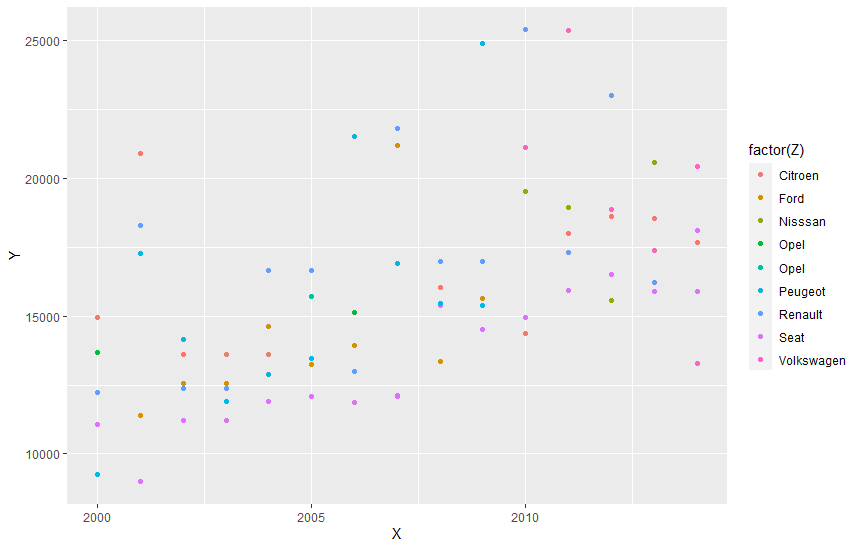
Vamos a poner el supuesto que somos una empresa que quiere analizar el mercado de coches, vamos a representar distintas gráficas con distintos datos a lo largo de la práctica para tener la mayor cantidad de información posible

En esta primera gráfica vemos la representación del precio por año de cada empresa, en la cual vemos el incremento de precio gradual desde el 2000 hasta el 2014

# ··· ··· ··· ··· ··· ··· ··· ··· ··· ··· Representamos mediante graficas ··· ··· ··· ··· ··· ··· ··· ··· ··· ··· #

# · Representamos el año frente al precio y lo coloreamos dependiedo de la marca

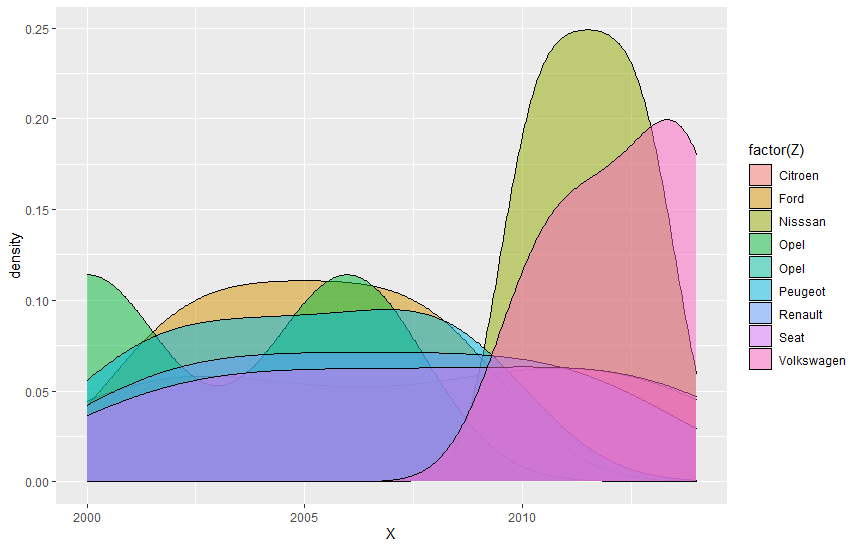
ggplot(datos, aes(X, Y)) + geom\_point(aes(colour = factor(Z)))



Ahora vamos a representar la densidad de ventas que tienen las empresas por año

# · Representamos la densidad de las marcas por año

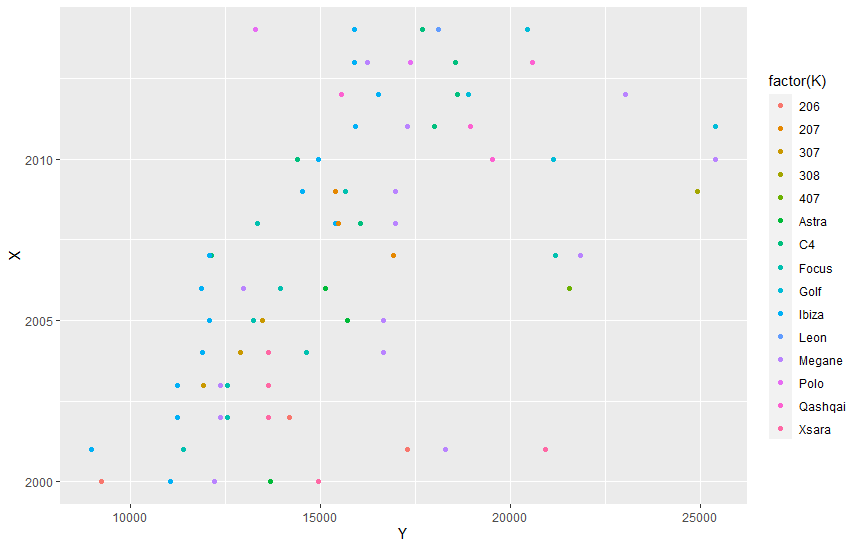
ggplot(data = datos) + geom\_density(aes(x=X,fill=factor(Z)),bins=10, position = "identity",alpha = 0.5)



También nos interesa saber el precio por año de cada modelo que se ha vendido, en la grafica vemos como hay algunos modelos que se siguen manteniendo a lo largo de los años y otros que en camio desaparecen u otros que aparecen nuevos

# · Representamos por precio y año pero pintamos por el modelo de coche

ggplot(datos, aes(Y, X)) + geom\_point(aes(colour = factor(K)))

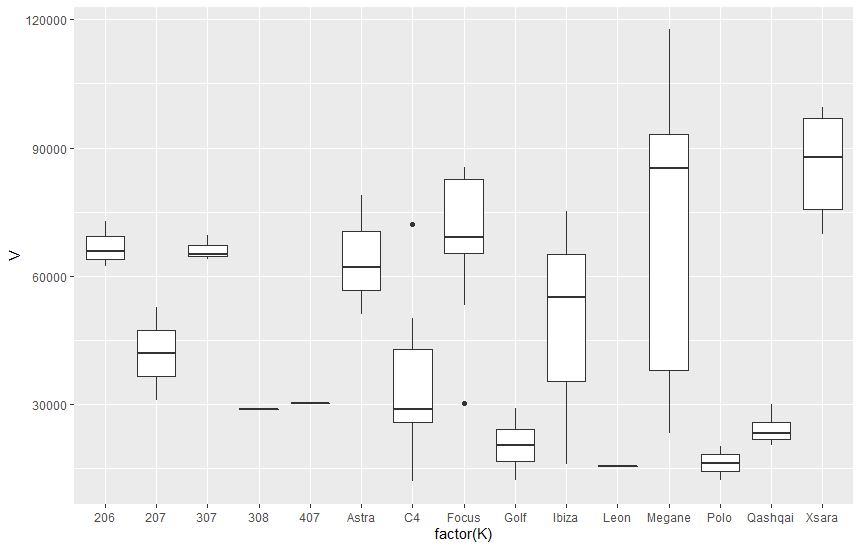


Ahora vamos a utilizar los BoxPlot que es un diagrama de caja es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles

Y vemos cada modelo por el numero de ventas totales que ha tenido

# · Dibujamos con un BoxPlot los modelos de coches por unidades vendidas

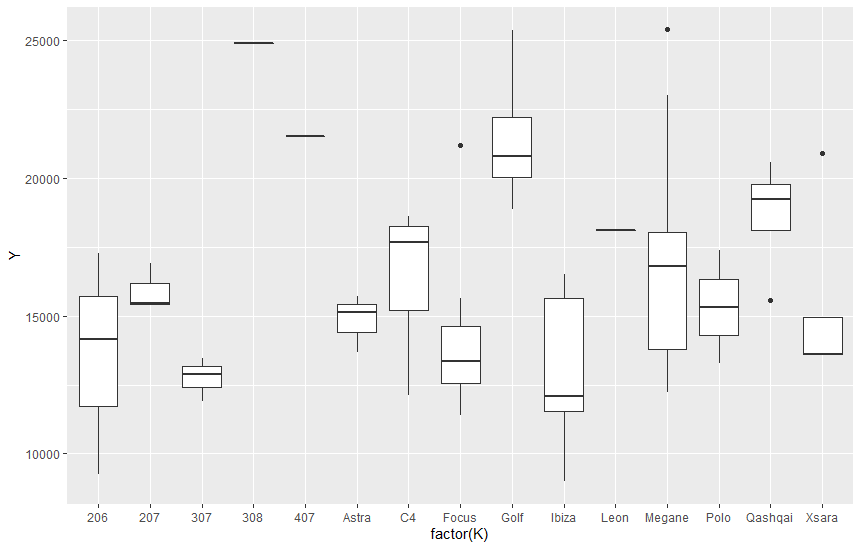
ggplot(data = datos) + geom\_boxplot(aes(x=factor(K), y=V))



Tambien con los BoxPlot podemos visualizar el precio que ronda cada modelo de coche

# · Dibujamos con un BoxPlot los modelos de coches por precio

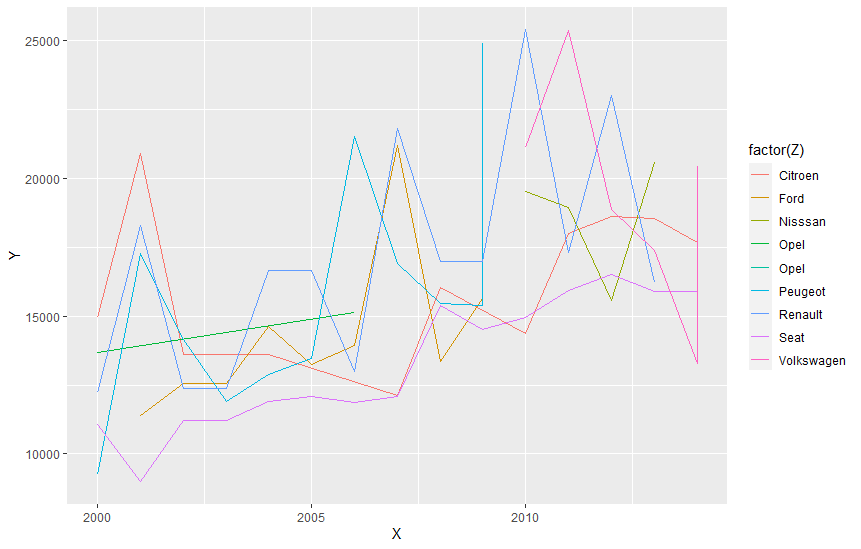
ggplot(data = datos) + geom\_boxplot(aes(x=factor(K), y=Y))



Como empresa que busca analizar el mercado, nos interesará saber la evolución de los precios a lo largo de los años de cada empresa, los cuales podemos apreciar como suben

# · Vemos como a lo largo de los años las marcas cambian de precio

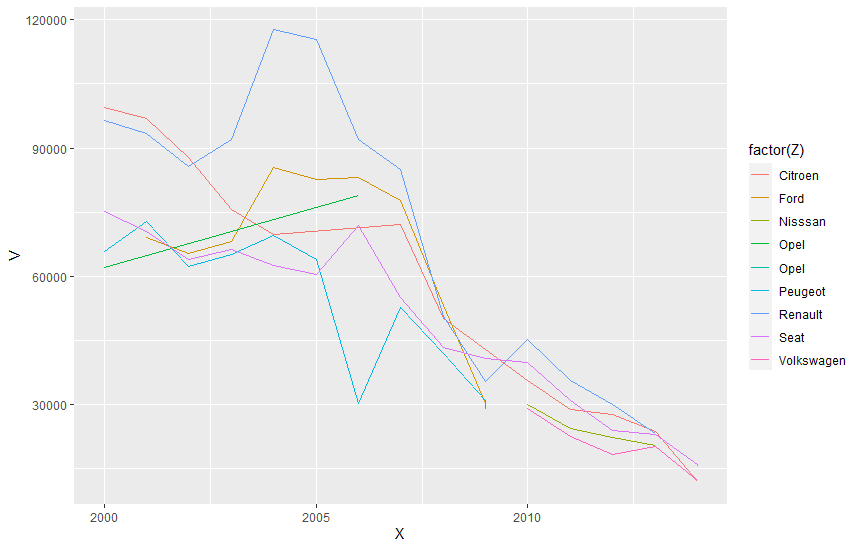
ggplot(datos) + geom\_line(aes(x=X, y=Y,colour = factor(Z)))



Si analizamos detenidamente la siguiente grafica nos daremos cuenta del descenso del numero de ventas de coches nuevos decae claramente, recordar que los coches son los 5 coches mas vendidos por año desde el 2000 hasta el 2014 y vemos un descenso en picado

# · Vemos como a lo largo de los años las marcas pierden unidades de venta

ggplot(datos) + geom\_line(aes(x=X, y=V,colour = factor(Z)))



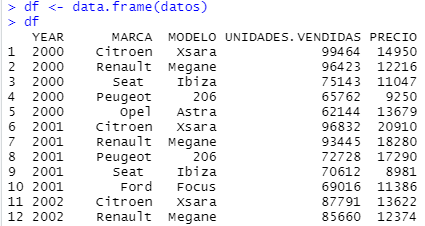
RStudio nos permite trabajar con dataFrames, el cual lo obtenemos a raiz de un csv

# ··· ··· ··· ··· ··· ··· ··· ··· ··· ··· Convertimos a DataFrame ··· ··· ··· ··· ··· ··· ··· ··· ··· ··· #

df <- data.frame(datos)

df

# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---



Procedemos a hallar la media del precio en general a lo largo de los años y la suma total de las unidades vendidas

Lo cual nos vendrá bien para mas adelante

# ··· ··· ··· ··· ··· Sacamos la media de precio y de unidades vendidas en general ··· ··· ··· ··· ··· #

mediaUnidadesPrecio <- mean(Y)

sumaUnidadesVendidas <- sum(V)

# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

En las siguientes graficas vamos a representar el precio por los años y las unidades vendidas a lo largo de los años, con la diferencia que va a haber una linea de color roja y verde que nos indica la media de precio total y la media de unidades vendidas totales

Y podemos apreciar como el precio va subiendo con los años pero las unidades vendidas caen

# ··· ··· ··· ··· ··· Representamos años por precio con una linea roja que es la media del precio ··· ··· ··· ··· ··· #

ggplot(mapping = aes(x = X, y = Y)) +

labs(x = "AÑOS", y = "PRECIO") +

geom\_point() +

geom\_hline(mapping = aes(yintercept = mean(Y)), color = "red")

# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- #

# ··· ··· ··· ··· ··· Representamos años por unidades vendidas con una linea verde que es la media de las unidades vendidas ··· ··· ··· ··· ··· #

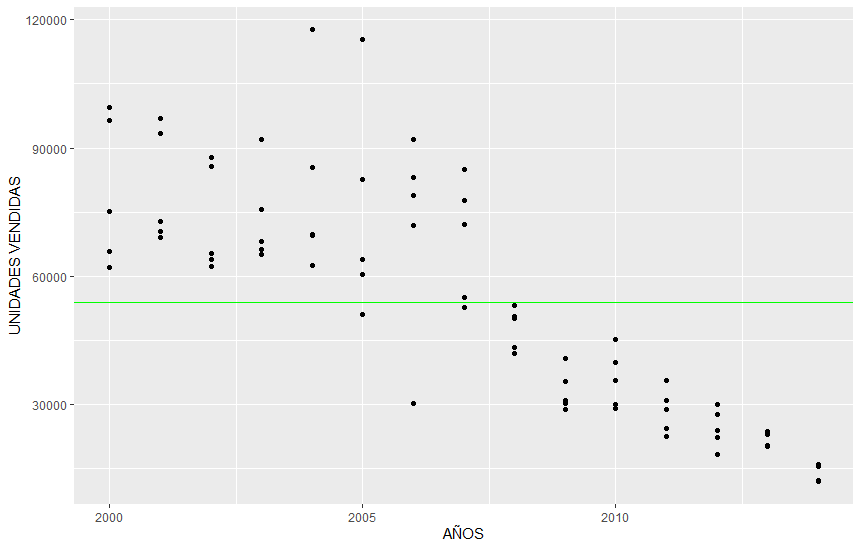
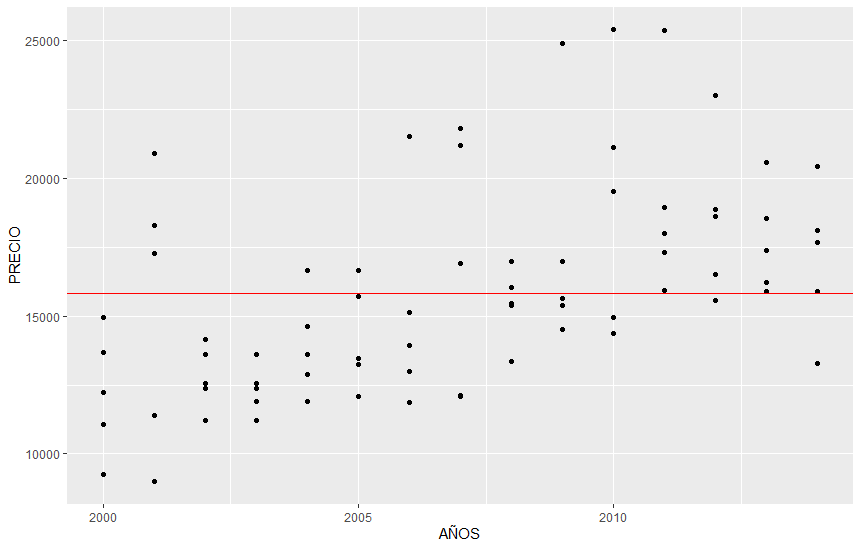
ggplot(mapping = aes(x = X, y = V)) +

labs(x = "AÑOS", y = "UNIDADES VENDIDAS") +

geom\_point() +

geom\_hline(mapping = aes(yintercept = mean(V)), color = "green")

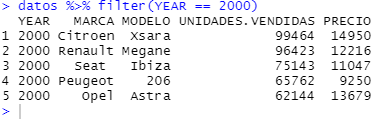
# --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- #



En RStudio nos permite mediante %>% filtrar lo que queramos que esté en la tabla

Si escribimos:

datos %>% filter(YEAR == 2000)



Podremos comprobar los datos que hay en el año 2000

Y si escribimos:

cochesYear2000 <- datos %>% filter(YEAR == 2000)

Lo podremos guardar en una variable

Vamos a proceder a ver el precio de los coches que tenían en el año 2000, utilizando la variable anteriormente mencionada, la cual actúa de filtro, ya que lo hemos guardado ahí

# · Sacamos el precio de los coches en el año 2000 y trazamos una linea roja que es la media del precio en el año 2000

ggplot(mapping = aes(x = cochesYear2000$YEAR, y = cochesYear2000$PRECIO)) +

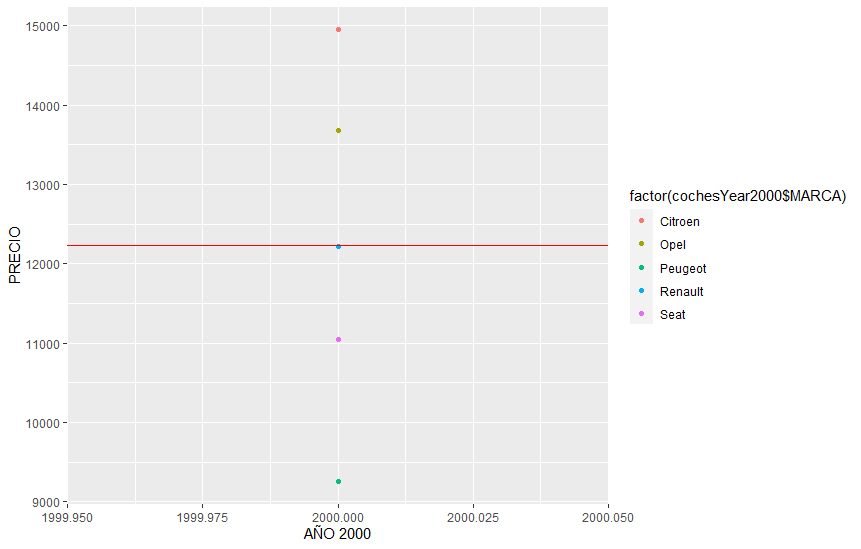
labs(x = "AÑO 2000", y = "PRECIO") +

geom\_point(aes(colour = factor(cochesYear2000$MARCA))) +

geom\_hline(mapping = aes(yintercept = mean(cochesYear2000$PRECIO)), color = "red")

La parte de: mean(cochesYear2000$PRECIO Podemos sacar cualquier dato que se asemeje con ese filtro, es decir, si los coches son del año 2000 la media de precio que va a hacer no es del precio general, si no del precio de ese año, mas o menos actua como una tabla

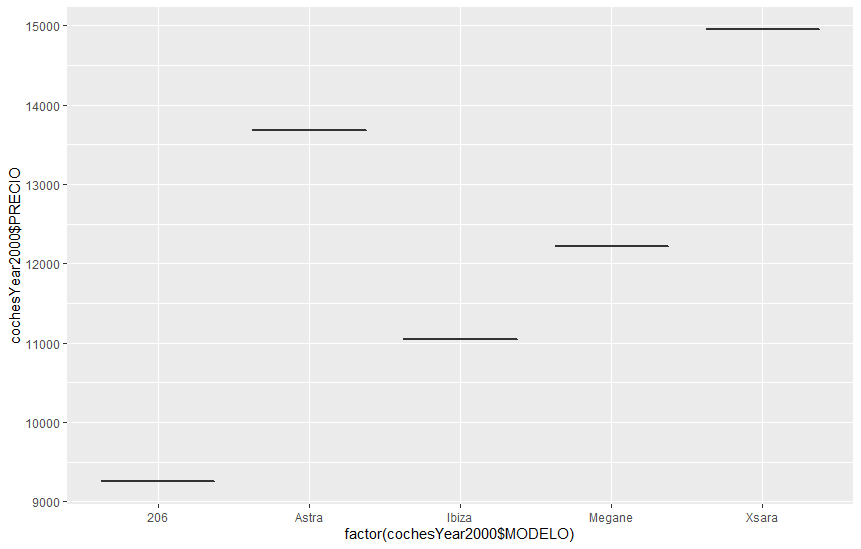
Vemos el precio de los coches en el año 2000 con la media de precio de ese año



Volvemos a utilizar los boxplot para hallar con la variable de los coches del 2000 el precio que tenían los distintos modelos

# · Dibujamos con un boxplot los modelos de los coches por el precio

ggplot(data = cochesYear2000) + geom\_boxplot(aes(x=factor(cochesYear2000$MODELO), y=cochesYear2000$PRECIO))



Con la libreria PIE nos permite hacer diagramas de sectores con colores (diagramas de queso) con los cuales vamos a ver el volumen de ventas (unidades vendidas) de las marcas en el año 2000

# · Hacemos un diagrama de sectores con las unidades vendidas de cada marca en el año 2000

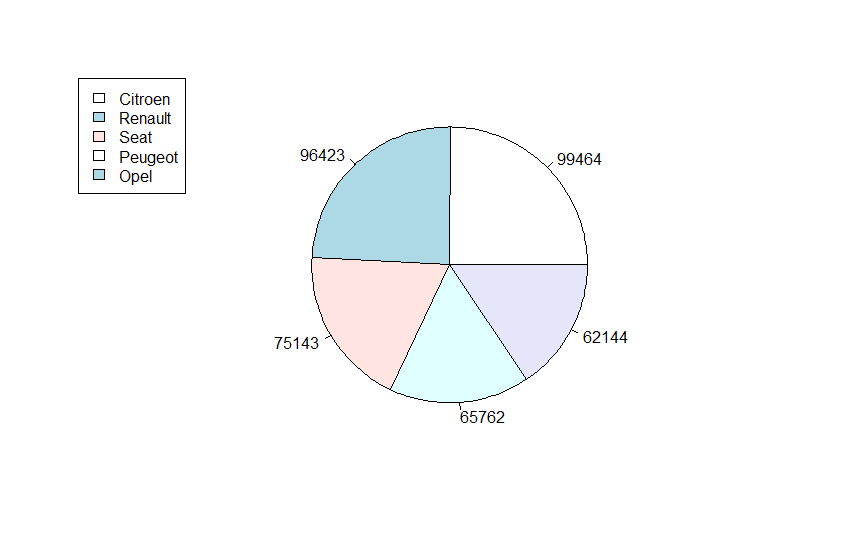
count\_3 <- c(cochesYear2000$UNIDADES.VENDIDAS)

sum(count\_3)

pie(count\_3, labels = paste0(count\_3))

legend("topleft", legend = c(cochesYear2000$MARCA),

fill = c("white", "lightblue", "mistyrose"))



Todo lo anteriormente mencionado con el filtro del año 2000 se puede replicar con cualquier dato, en mi caso lo he hecho todo igual hasta el 2014 pero solo voy a poner el del año 2000 ya que es igual siempre

Con la función group\_by podemos agrupar los datos que pertenezcan a ese mismo y la función summarise() funciona de forma análoga a la función mutate, excepto que en lugar de añadir nuevas columnas crea un nuevo data frame y todo ello lo podemos guardar en una variable la cual nos sirve para agrupar los datos por el año y crear la variable promPrecio que es para ver el precio medio de los coches, o la media del precio

# · Hacemos una grafica del precio medio por años para ver como evoluciona

mediaPrecioYears <- datos %>% group\_by(YEAR) %>% summarise(promPrecio=mean(PRECIO))

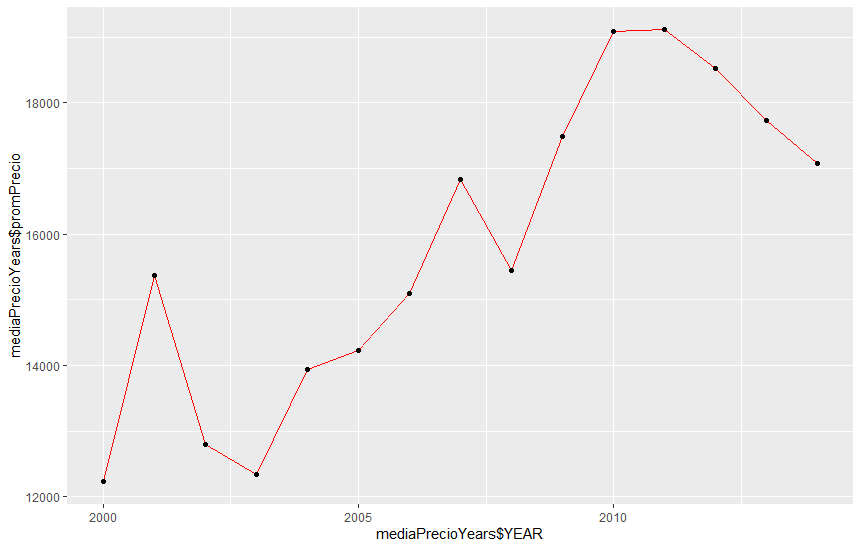
mediaPrecioYears

ggplot(data=mediaPrecioYears, aes(x=mediaPrecioYears$YEAR, y=mediaPrecioYears$promPrecio, group=1)) +

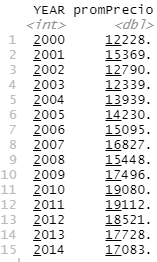
geom\_line(color="red")+

geom\_point()

Vemos como el precio medio sube considerablemente, ademas que podemos suponer que fue por la crisis de 2008



Podemos ver que la variable mediaPrecioYears contiene el precio medio por cada año



En contraste nos vamos a fijar en la media de las unidades vendidas por años para fijarnos en la diferencia con respecto al anterior y asi vemos como evoluciona

# · Hacemos una grafica de la media de las unidades vendidas por años para ver como evoluciona

mediaUnidadesVendidasYears <- datos %>% group\_by(YEAR) %>% summarise(promUnidadesVendidas=mean(UNIDADES.VENDIDAS))

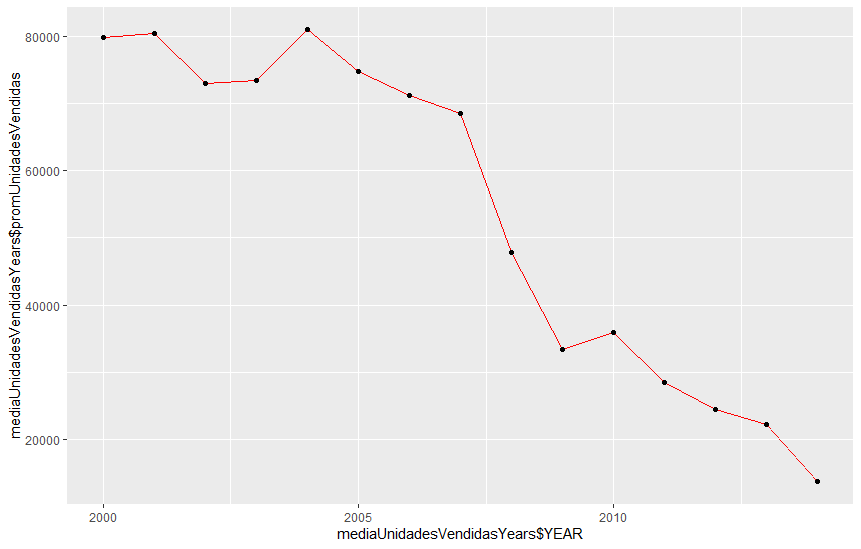
mediaUnidadesVendidasYears

ggplot(data=mediaUnidadesVendidasYears, aes(x=mediaUnidadesVendidasYears$YEAR, y=mediaUnidadesVendidasYears$promUnidadesVendidas, group=1)) +

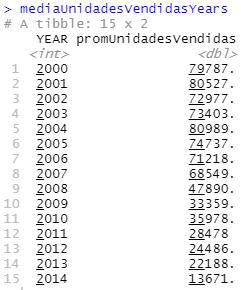
geom\_line(color="red")+

geom\_point()

Como vemos, las ventas de 80000 que habia en el año 2000 a menos de 20000 que hay en 2014



La variable que guarda nuestra media de unidades vendidas:

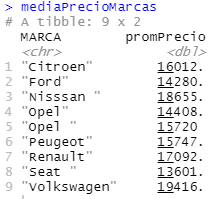


Tambien nos interesa saber en que segmento de precio se encuentra cada marca, ya sea segmento premium, cohe medio… Entonces para saberlo simplemente podemos hacer una media del precio de cada marca y que nos lo saque en una tabla y así podemos analizar a la competencia

# · Sacamos la tabla de la media del precio de las marcas

mediaPrecioMarcas <- datos %>% group\_by(MARCA) %>% summarise(promPrecio=mean(PRECIO))

mediaPrecioMarcas



Como vemos en primera posicion esta Volkswagen con 19416 € de media

Ya que el precio de cada marca fluctúa y va cambiando con el paso de los años por distintos motivos, es interesante ver su evolución a lo largo de los años, lo interesante es ver el cambio en cada marca y como algunas irrumpen en el mercado y otras que practicamente desaparecen

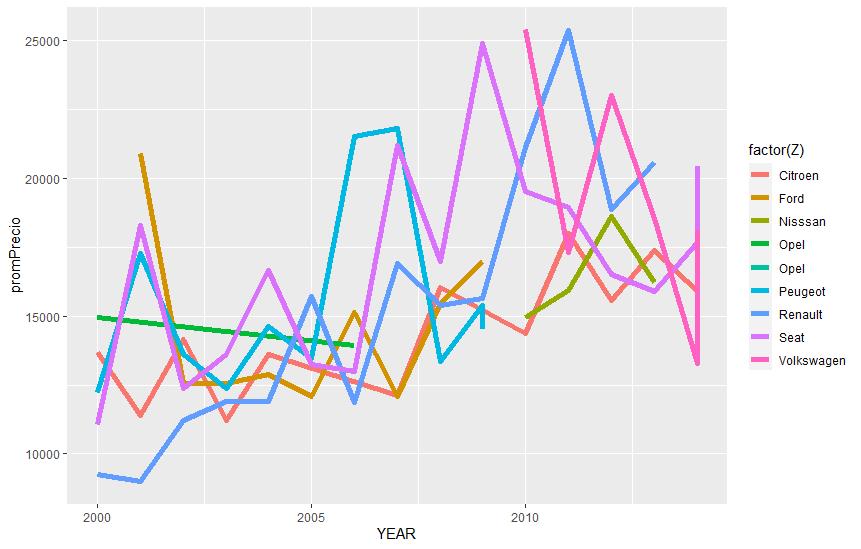
# · Realizamos la grafica del precio medio de cada marca a lo largo de los años

datos %>%

group\_by(YEAR, UNIDADES.VENDIDAS) %>%

summarize(promPrecio=mean(PRECIO)) %>%

ggplot(aes(x=YEAR, y=promPrecio, color= factor(Z))) + geom\_line(size=2)



Para tener mas cantidad de información vamos a separar el mercado en 3 tipos, el mercado de gama baja, que supongamos que está en un precio inferior a los 10000, recordando que estamos trabajando sobre los 5 coches más vendidos hasta 2014, tambien lo dividiremos en gama media y gama alta, que trabajaremos con ello más adelante

# · Filtramos el precio a menos de 10000 para sacar los coches de gama baja y dibujamos los puntos donde se encuentran y una linea roja que delimita el precio medio

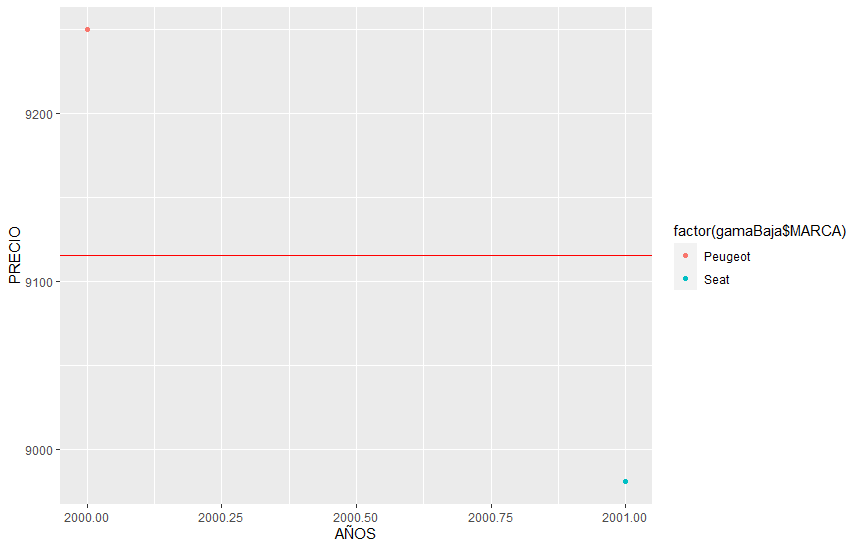
gamaBaja <- datos %>% filter(PRECIO < 10000)

Para ello lo filtraremos por precio que sea inferior a 10000 y lo guardaremos en una variable

Y en la tabla nos mostrará que unicamente solo 2 coches pertenecen a esta categoría



Y realizamos la gráfica, en la cual vemos como en el año 2000 había un Peugeot que sobrepasaba los 9200 € y en 2001 había un Seat que no llegaba a los 9000 €



Y hacemos lo mismo con la gama media y gama alta:

# · Filtramos el precio entre 10000 y 20000 para sacar los coches de gama media y dibujamos los puntos donde se encuentran y una linea roja que delimita el precio medio

gamaMedia <- datos %>% filter(PRECIO >= 10000, PRECIO <= 20000)

# Mostramos la tabla

datos %>% filter(PRECIO >= 10000, PRECIO <= 20000)

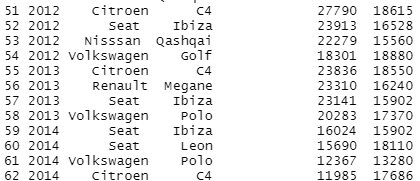
ggplot(mapping = aes(x = gamaMedia$YEAR, y = gamaMedia$PRECIO)) +

labs(x = "Valores predichos", y = "Residuos") +

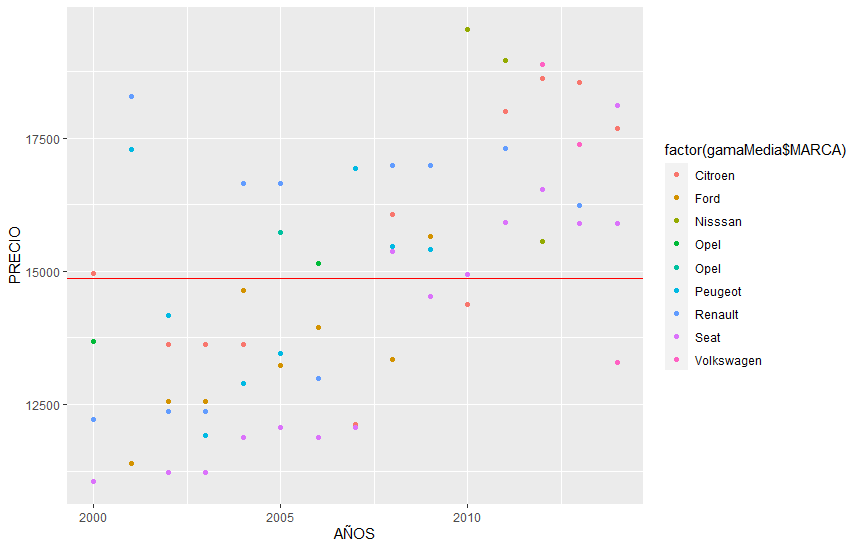
geom\_point(aes(colour = factor(gamaMedia$MARCA))) +

geom\_hline(mapping = aes(yintercept = mean(gamaMedia$PRECIO)), color = "red")

Vemos que a diferencia de la gama baja, la gama media tiene muchisimos mas vehiculos, en este caso como vemos hay 62 coches



Y la gráfica:



Vemos como el precio medio en la gama media esta en unos 15000 € y que es un segmento lleno de coches, marcas y es muy competitivo, además que podemos extraer como conclusión es que el grueso de las ventas está en el segmento de gama media y por tanto podemos decir que el mercado español desde el 2000 hasta el 2014 es un mercado o una sociedad en su reflejo de clase media

Pasamos a la gama alta:

# · Filtramos el precio a mas de 20000 para sacar los coches de gama alta

gamaAlta <- datos %>% filter(PRECIO > 20000)

# Mostramos la tabla

datos %>% filter(PRECIO > 20000)

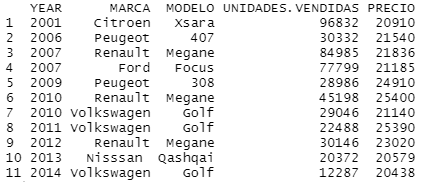
ggplot(mapping = aes(x = gamaAlta$YEAR, y = gamaAlta$PRECIO)) +

labs(x = "AÑOS", y = "PRECIO") +

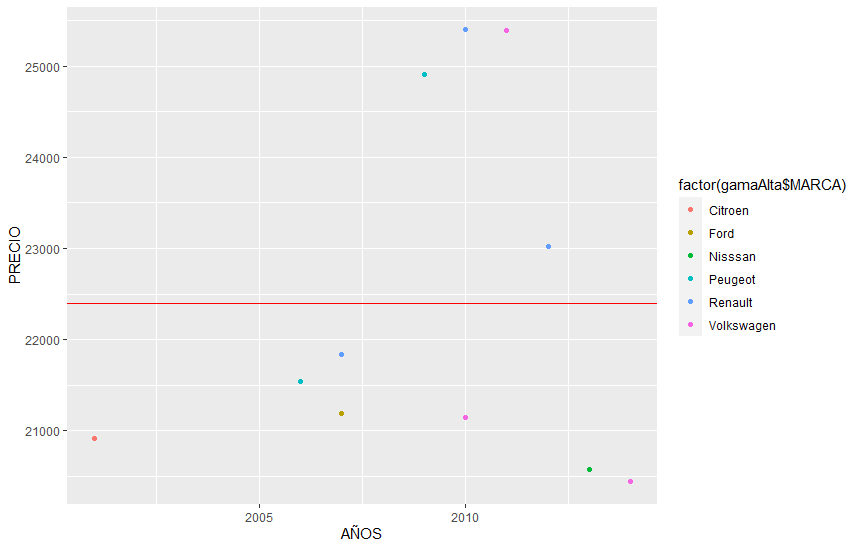
geom\_point(aes(colour = factor(gamaAlta$MARCA))) +

geom\_hline(mapping = aes(yintercept = mean(gamaAlta$PRECIO)), color = "red")

Vemos como solo hay 11:



Y vemos que el precio medio está sobre los 22500€



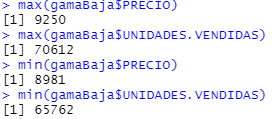
Ahora nos interesará ver el precio mínimo y máximo de cada segmento, además, del máximo y el mínimo de las unidades vendidas

maxPrecioBajo <- max(gamaBaja$PRECIO)

maxUnidadesVendidasPrecioBajo <- max(gamaBaja$UNIDADES.VENDIDAS)

minPrecioBajo <- min(gamaBaja$PRECIO)

minUnidadesVendidasPrecioBajo <- min(gamaBaja$UNIDADES.VENDIDAS)



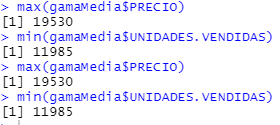
Lo mismo con la gama media:

maxPrecioMedio <- max(gamaMedia$PRECIO)

maxUnidadesVendidasPrecioMedio <- min(gamaMedia$UNIDADES.VENDIDAS)

minPrecioMedio <- max(gamaMedia$PRECIO)

minUnidadesVendidasPrecioMedio <- min(gamaMedia$UNIDADES.VENDIDAS)



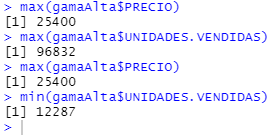
Y con la gama alta:

maxPrecioAlto <- max(gamaAlta$PRECIO)

maxUnidadesVendidasPrecioAlto <- max(gamaAlta$UNIDADES.VENDIDAS)

minPrecioAlto <- max(gamaAlta$PRECIO)

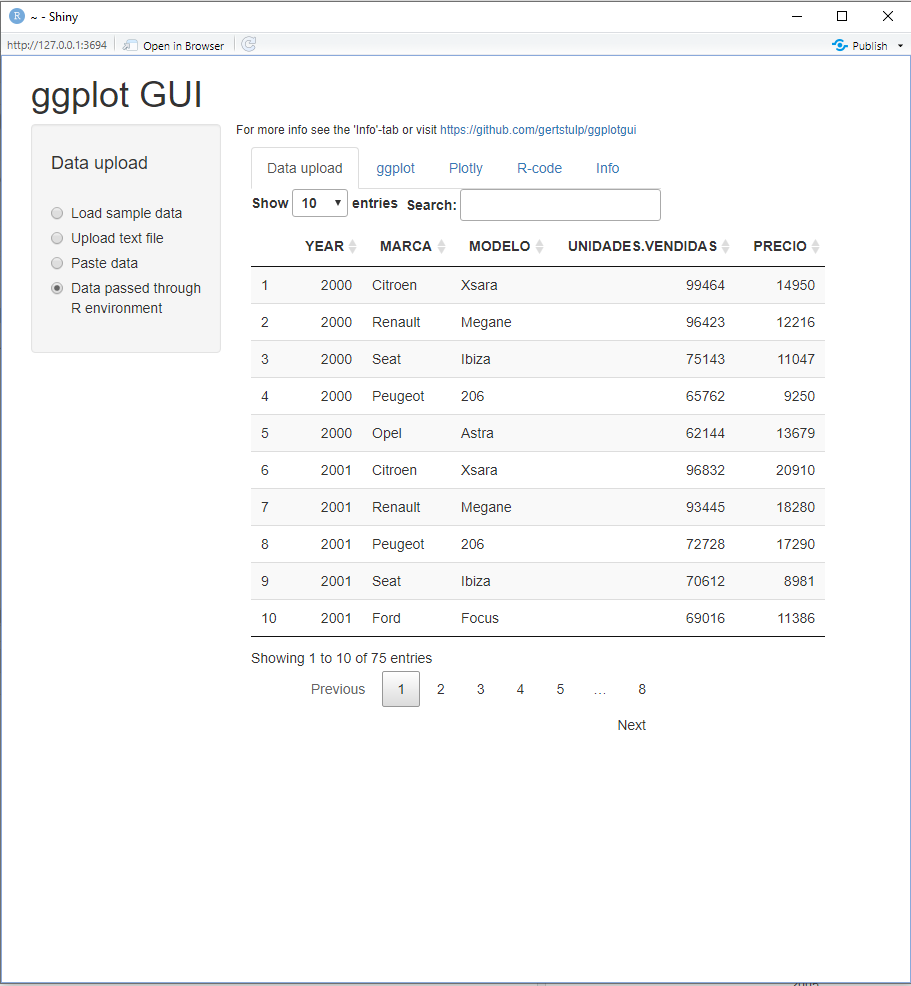
minUnidadesVendidasPrecioAlto <- min(gamaAlta$UNIDADES.VENDIDAS)



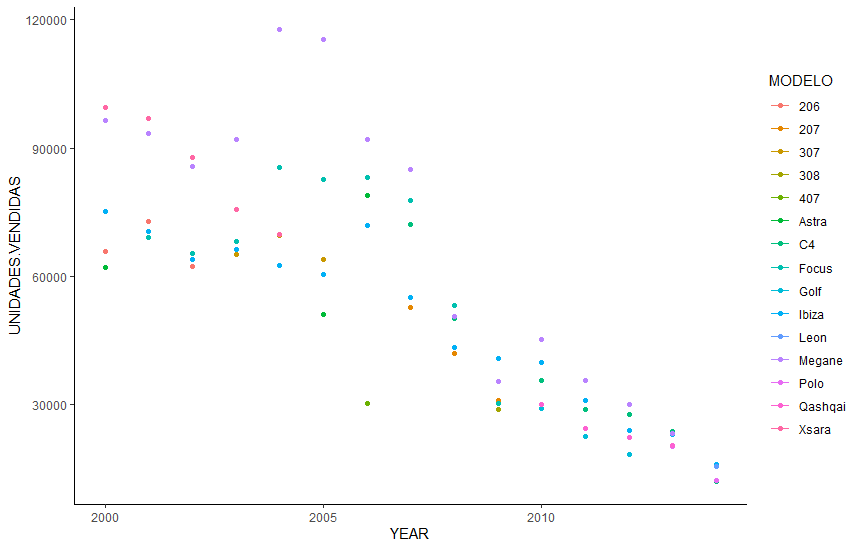
Otra funcionalidad de RStudio es el *ggplotgui* el cual nos puede beneficiar ya que tiene una interfaz de usuario que nos facilita mucho las cosas y además es muy intuitiva

Para utilizarla solo nos bastará con: (datos es nuestra variable donde hemos leido el csv)

ggplot\_shiny(datos)



Y podremos interactuar con ella y hacer un montón de funcionalidades que tiene, como por ejemplo el crear gráficas como la siguiente:



Que son las unidades vendidas de cada modelo por año

R también nos permite seleccionar los datos que queramos y mostrarlos como es el caso:

Renault <- datos[datos$MARCA == "Renault",]

ggplot(data=Renault, aes(x=YEAR, y=UNIDADES.VENDIDAS, group=1)) +

geom\_line(color="red")+

geom\_point()

Citroen <- datos[datos$MARCA == "Citroen",]

ggplot(data=Citroen, aes(x=YEAR, y=UNIDADES.VENDIDAS, group=1)) +

geom\_line(color="red")+

geom\_point()

Seat <- datos[datos$MARCA == "Seat",]

ggplot(data=Seat, aes(x=YEAR, y=UNIDADES.VENDIDAS, group=1)) +

geom\_line(color="red")+

geom\_point()

Ford <- datos[datos$MARCA == "Ford",]

ggplot(data=Ford, aes(x=YEAR, y=UNIDADES.VENDIDAS, group=1)) +

geom\_line(color="red")+

geom\_point()

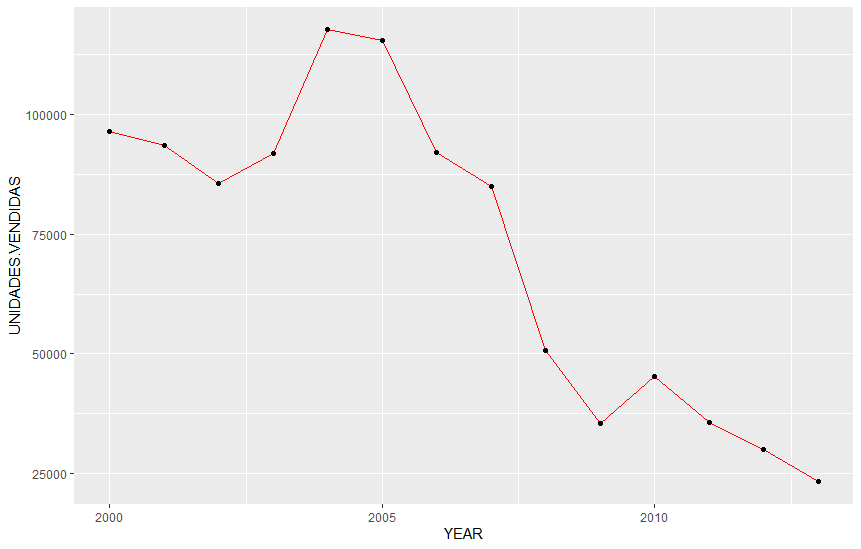
Opel <- datos[datos$MARCA == "Opel",]

ggplot(data=Opel, aes(x=YEAR, y=UNIDADES.VENDIDAS, group=1)) +

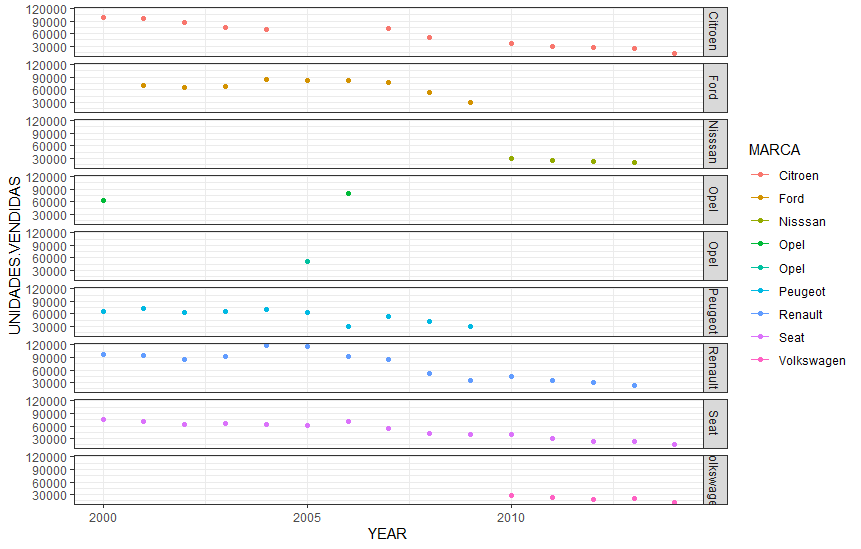
geom\_line(color="red")+

geom\_point()

Seleccionamos la marca concreta y la guardamos dentro de una variable para luego poder hacerle una gráfica para cada una de las unidades vendidas por año

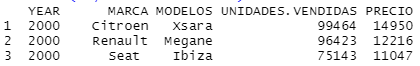


En este caso la grafica es de las unidades vendidas por año de la marca Renault, pero como no quiero sobrecargar de graficas, con la ayuda de ggplotgui vamos a agrupar todas las marcas en una misma imagen:



rename(df, MODELOS = MODELO)

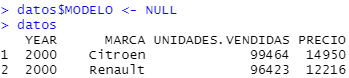
Para acabar, en r podemos cambiar el nombre de las columnas con la anterior sentencia (Pero del datframe)



Y este sería el resultado, hemos cambiado MODELO por MODELOS

Tambien podemos borrar columnas con:

datos$MODELO <- NULL



Y así quedaría sin la columna MODELO

Bibliografía

Datos de los coches más vendidos en España desde el 2000 hasta el 2014:

https://www.cursocerrajeria.com/inicio/cursos-de-apertura-de-coches/coches-m%C3%A1s-vendidos-en-espa%C3%B1a-por-a%C3%B1os/

Precio de los coches en ese año:

Se han ido buscando individualmente en:

https://www.km77.com/