Análisis - Compra-venta de coches usados en UK

Marta Gómez / Juan Fco Nieto

12/17/2021

Contents


```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')

if (!require('tidyr')) install.packages('tidyr'); library('tidyr')

# Comprobación de media winsor
if (!require('skimr')) install.packages('skimr'); library('skimr')
```

0.0.0.0.1 Cargar librerías

```
used_cars <- read.csv('aprox100KUsedCars.csv', stringsAsFactors = TRUE)

sample_n(used_cars,10)  %>% knitr::kable()
```

0.0.0.0.2 Lectura del fichero y preparación de los datos

manufacturer	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
merc	A Class	2014	12000	Manual	43937	Diesel	20	70.6	1.5
toyota	Auris	2017	18000	Automatic	17558	Hybrid	135	70.6	1.8
ford	Mustang	2020	42489	Automatic	3500	Petrol	145	22.1	5.0
bmw	X3	2016	18995	Semi-Auto	34274	Diesel	145	54.3	2.0
bmw	X1	2016	13990	Manual	43529	Diesel	125	60.1	2.0
merc	C Class	2019	34499	Semi-Auto	3822	Petrol	145	44.1	2.0
audi	Q3	2016	17584	Manual	21002	Diesel	160	49.6	2.0
toyota	Yaris	2011	6495	Automatic	26335	Other	30	54.3	1.3
vw	Polo	2016	9990	Manual	20665	Petrol	20	60.1	1.2
vauxhall	Mokka X	2017	11280	Manual	14800	Petrol	145	47.1	1.4

```
sk <- skim(used_cars)
```

1 Descripción del dataset

1.1 Propósito

Nos encontramos con un dataset con con 99187 filas y 10 columnas que representan 99187 ofertas en portales de compraventa de coches usados en UK. Este dataset ha sido UK creado en Julio 2020 por el usuario 'Aditya' (https://www.kaggle.com/adityadesai13) a través de web scraping.

El dataset está orientado a crear un modelo de coches usados para hacer predicciones sobre la variable target "price", interpretándose como un análisis del precio de mercado.

Citando al usuario:

"I collected the data to make a tool to predict how much my friend should sell his old car for compared to other stuff on the market, and then just extended the data set. Then made a more general car value regression model."

El usuario hizo el dataset para averiguar el precio de mercado de un coche en concreto mediante regresión lineal.

2 Integración y selección de los datos de interés a analizar.

En cuanto a la integración, este dataset proviene de una única fuente, pero de múltiples ficheros que han sido mergeados tal como se expresa en el README del proyecto con un script de Ruby.

2.1 Descripción de las variables

Entre las 10 variables del dataset podemos encontrar:

2.1.1 Variables categóricas

- manufacturer: Fabricante del automóvil. Variable categórica nominal con 9 categorías diferentes.
- model: Modelo del automóvil. Variable categórica nominal con 195 categorías diferentes.
- transmission: Tipo de transmisión. Variable categórica nominal con 4 categorías. Transmisión manual, automática, semiautomática y otras.

- fuelType: Tipo de combustible. Variable categórica nominal con 5 categorias. Diesel, eléctrico, híbrido, gasolina y otros.

```
sk %>% yank('factor') %>% select(c(skim_variable, n_unique)) %>% rename(Variable=skim_variable, Niveles=
```

Variable	Niveles
manufacturer	9
model	195
transmission	4
fuelType	5

2.1.2 Variables numéricas

- year: Año de matriculación del coche. Variable numérica discreta.
- price: Precio en Libras que se colocó en el portal de compraventa a fecha Julio 2020. Variable numérica continua.
- mileage: Millaje. Millas que el coche ha recorrido desde su puesta en funcionamiento. (En España utilizamos Kilometraje, porque medimos esta distancia en kilómetros). Variable numérica continua
- tax: Impuesto de circulación (en Libras). Dependiendo de los años del vehículo, emisión de gases (sobre todo) y otros factores este impuesto varía. Variable numérica continua
- mpg: Consumo de combutible del vehículo en millas por galon. Variable numérica continua
- engineSize: Tamaño del motor en litros. Variable numérica continua.

```
sk %>% yank('numeric') %>% select(c(skim_variable, mean, sd, p25, p50, p75, p0, p100)) %>% rename(Variab
```

Variable	Media	Desviación Típica	Q1	Q2/Mediana	Q3	Mínimo	Máximo
year	2017.088	2.124	2016.0	2017.0	2019.0	1970.0	2060.0
price	16805.348	9866.773	9999.0	14495.0	20870.0	450.0	159999.0
mileage	23058.914	21148.524	7425.0	17460.0	32339.0	1.0	323000.0
tax	120.300	63.151	125.0	145.0	145.0	0.0	580.0
mpg	55.167	16.139	47.1	54.3	62.8	0.3	470.8
engineSize	1.663	0.558	1.2	1.6	2.0	0.0	6.6

Como podemos observar tanto en las variables categóricas como numéricas no hay valores vacios, no quiere decir que la consistencia de los valores sea total. Esto está sujeto a más pruebas, ejemplo: una medida distinta a 0 litros de capacidad de motor tendría sentido para un vehículo diesel, gasolina o híbrido pero no eléctrico.

Por otra parte para el análisis, para hacerlo más completo cabe hacer comparaciones binomiales planteando hipótesis con nuevas variables como "Alta cilindrada/Baja cilindrada", "Alto consumo/Bajo consumo", "Eléctrico/Combustibles fósiles".

3 Limpieza de los datos

3.1 Análisis de valores nulos

A continuación se muestra la no existencia de valores nulos en el data set:

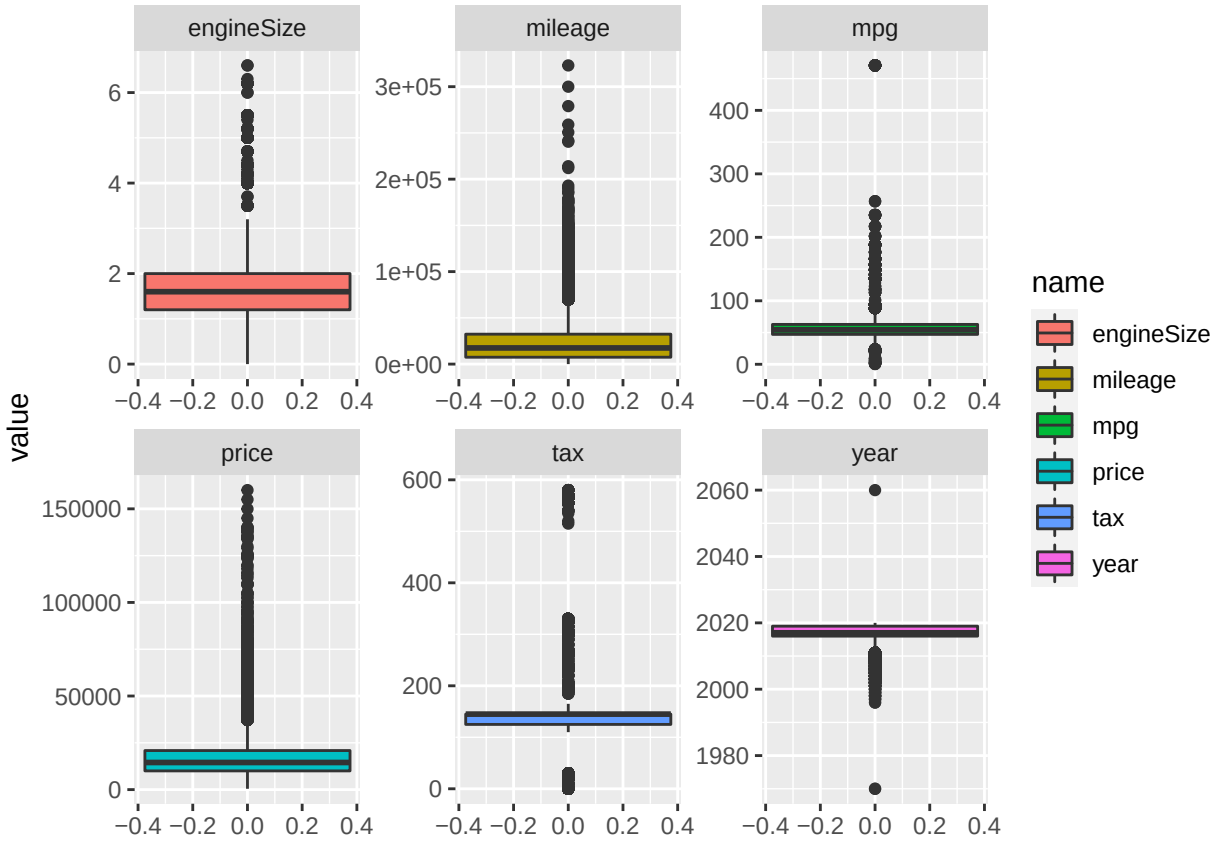
```r
is_blank <-  function(x){
  return (any(x=="") || anyNA(x))
}
used_cars %>% summarise_all(.funs=c('is_blank'), )
```

```
##   manufacturer model  year price transmission mileage fuelType   tax   mpg
## 1        FALSE FALSE FALSE FALSE        FALSE   FALSE    FALSE FALSE FALSE
##   engineSize
## 1      FALSE
```

3.2 Análisis de valores atípicos

```r
vars <- c("year", "price", "mileage", "tax", "mpg", "engineSize")
used_cars %>% select(vars) %>% pivot_longer(cols=vars, values_drop_na = TRUE) %>% ggplot(.) + facet_wra
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(vars)‘ instead of ‘vars‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```



Como podemos ver todas las variables tienen valores atípicos superiores y solo algunas inferiores. Y es por eso que nos deberíamos preguntar que hacer con esto valores, analicemos cada caso.

3.2.1 Identificación y tratamiento de outliers

```
get_outliers <- function(x){
  q1 = quantile(x, c(0.25))
  q3 = quantile(x, c(0.75))
  iqr = q3-q1
  result = sapply(x, function(y){
    if(y < q1 - 1.5*iqr){
      y
    }else if(y > q3+1.5*iqr){
      y
    }else{
      NA
    }
  }
  )
  return(list(result, q1 - 1.5*iqr, q3+1.5*iqr))
}
```

Con la presente función veremos todos los outliers de las distintas variables:

```
final <- list()
outliers_analysis <- used_cars %>% select(vars) %>% sapply(., get_outliers)
for(i in vars){
  ej <- na.omit(outliers_analysis[,i][[1]])

  final <- cbind(final, c(paste(sample(ej, 3), sep=" ", collapse=" / "),  outliers_analysis[,i][[2]], ou
}

colnames(final) = vars
rownames(final) = c("Ejemplos", "Mínimo", "Máximo", "Conteo")
data.frame(final) %>% knitr::kable()
```

	year	price	mileage	tax	mpg	engineSize
Ejemplos	2006 / 2010 / 2009	58994 / 45000 / 41391	94000 / 101800 / 73575	20 / 30 / 30	2.8 / 156.9 / 166.2	4 / 4 / 5.5
Mínimo	2011.5	-6307.5	-29946	95	23.55	-2.22044604925031e-16
Máximo	2023.5	37176.5	69710	175	86.35	3.2
Conteo	1737	3669	3902	28815	939	650

Como podemos ver a parte de outliers podemos ver inconsistencias en las distintas variables de nuestro dataset:

- Year: El año de matriculación no puede ser superior al año de la recogida de los datos (Julio 2010), por seguridad solo se aceptarán valores inferiores o iguales a 2020. Como los coches inferiores o iguales a 2011 son minoritarios y pueden sesgar nuestros análisis también los vamos a retirar.

```
original_used_cars <- used_cars
used_cars <- used_cars %>% filter(!(year < 2011.5 | year > 2019))
```

Hemos reducido 6202 filas. El total de filas ahora es 92985.

- Price: No podemos aceptar que haya precios negativos en nuestro dataset. Tampoco vamos a trabajar con valores atípicos en cuanto a precios puesto que pueden sesgar nuestro estudio alterando medidas de tendencia central y dispersión. Eliminamos outliers y precios inferiores a 0

```
used_cars <- used_cars %>% filter(!(price <= 0 | price > 37176.5))
```

El total de filas ahora es 90076, se han elimina 2909.

Podríamos imputar este valor, con kNN por ejemplo, pero dado que es nuestra variable principal de estudio vamos a llevarnos los mejores valores a la etapa de análisis.

- Mileage: de igual manera no podemos aceptar mileage negativo por ello eliminaremos de las negativas y los valores atípicos.

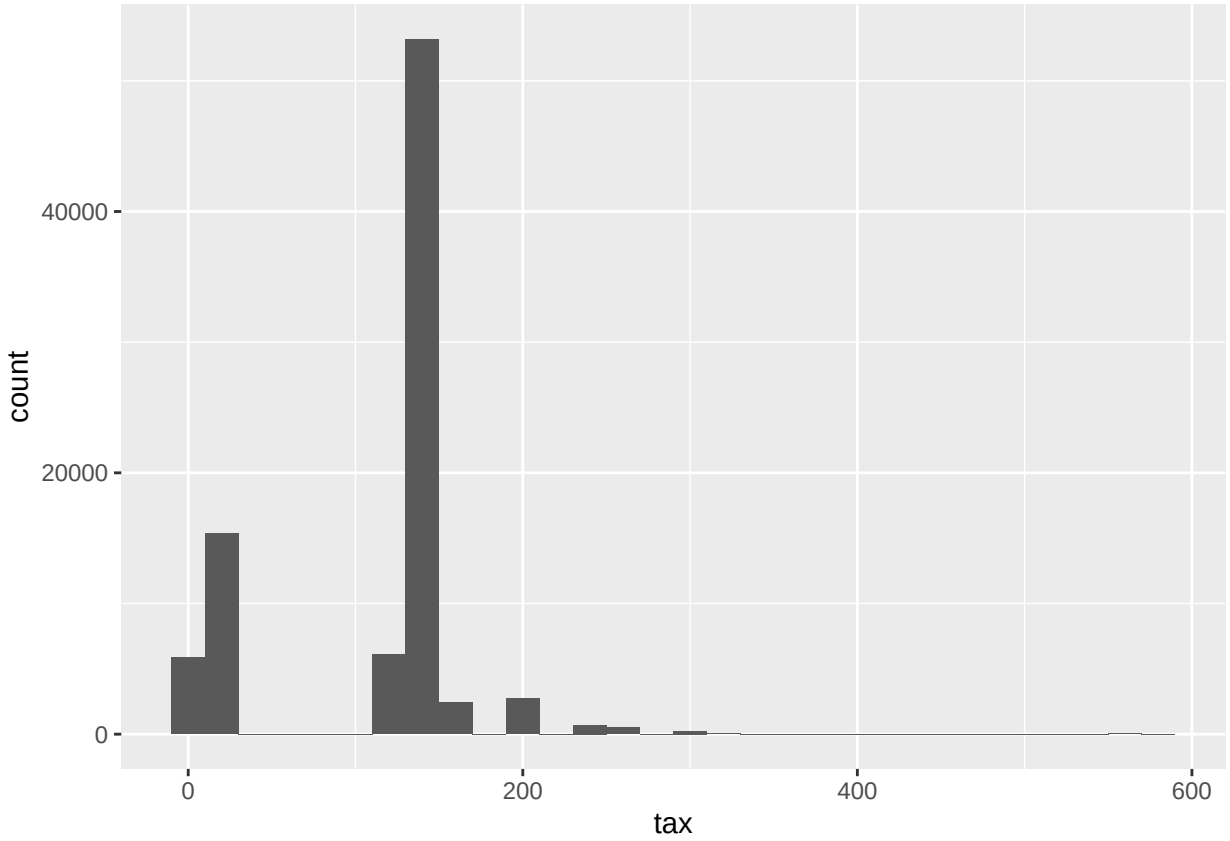
```
used_cars <- used_cars %>% filter(!(mileage <= 0 | mileage > 69710))
```

El total de filas ahora es 87251, se han eliminado 2825.

- Tax: podemos entender que la relacion de impuestos no es lineal y si está estratificada no corresponde el presente análisis de outliers. Veamos un histograma:

```
ggplot(used_cars) + geom_histogram(aes(tax))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Vemos que hay bastante volumen de gente que paga menos de 50 por lo que podemos pensar que no tiene por qué ser un error, en cambio por la parte superior, si es muy raro que se pague por encima de 200. Eliminaremos esta banda, para evitar sesgos.

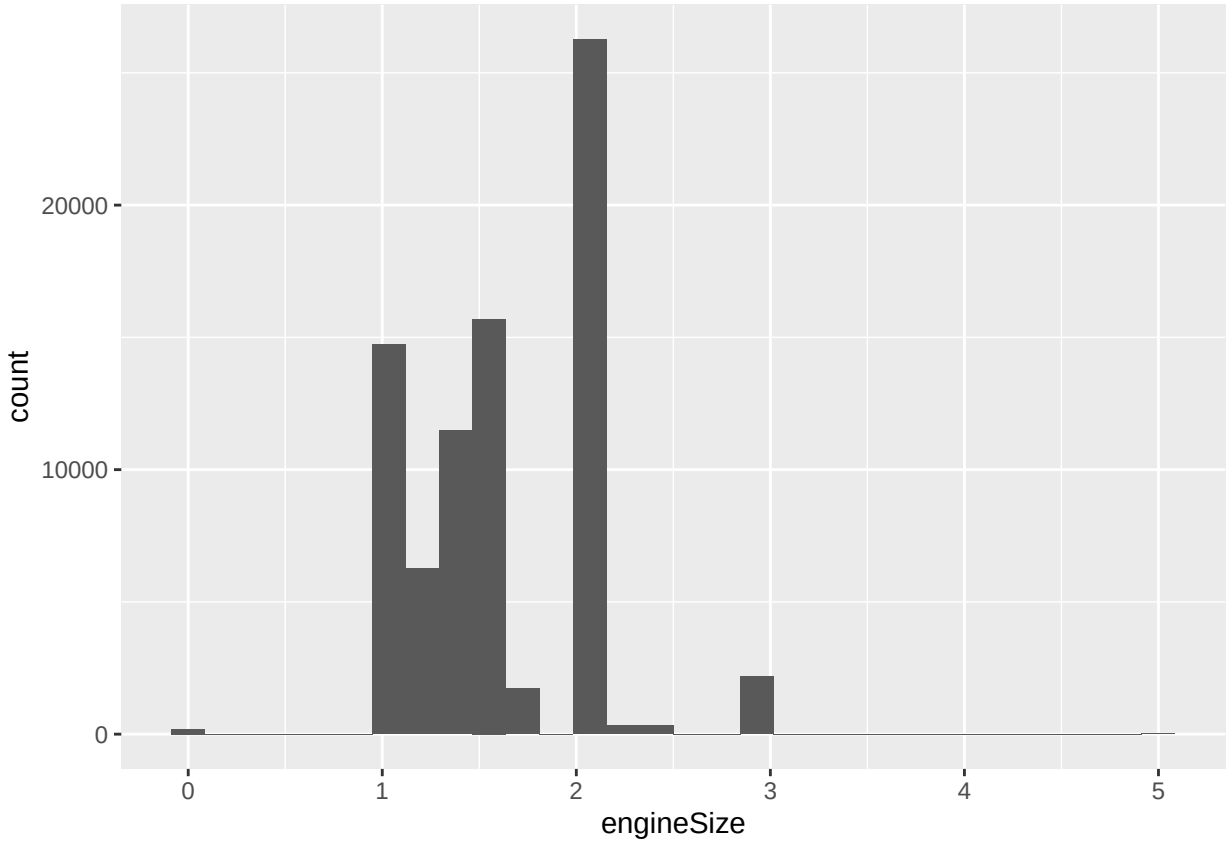
```
used_cars <- used_cars %>% filter(!(tax <= 0 | tax > 200))
```

El total de filas ahora es 79216, se han eliminado 8035.

- Mpg: debemos tener en cuenta que nuestro datasets contiene tanto coches híbridos como eléctricos por lo que habrá que mantener los datos cuyo mpg sea 0 o bajo puesto que hay que tener en cuenta estas categorías. Se podría hacer una manipulación distinta de los outliers para los diferentes segmentos de fuelType pero para no aumentar complejidad la mantendremos tal como está.
- engineSize: de igual manera un motor eléctrico deberá tener volumen 0, otro sería considerado una inconsistencia.

```
ggplot(used_cars) + geom_histogram(aes(engineSize))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



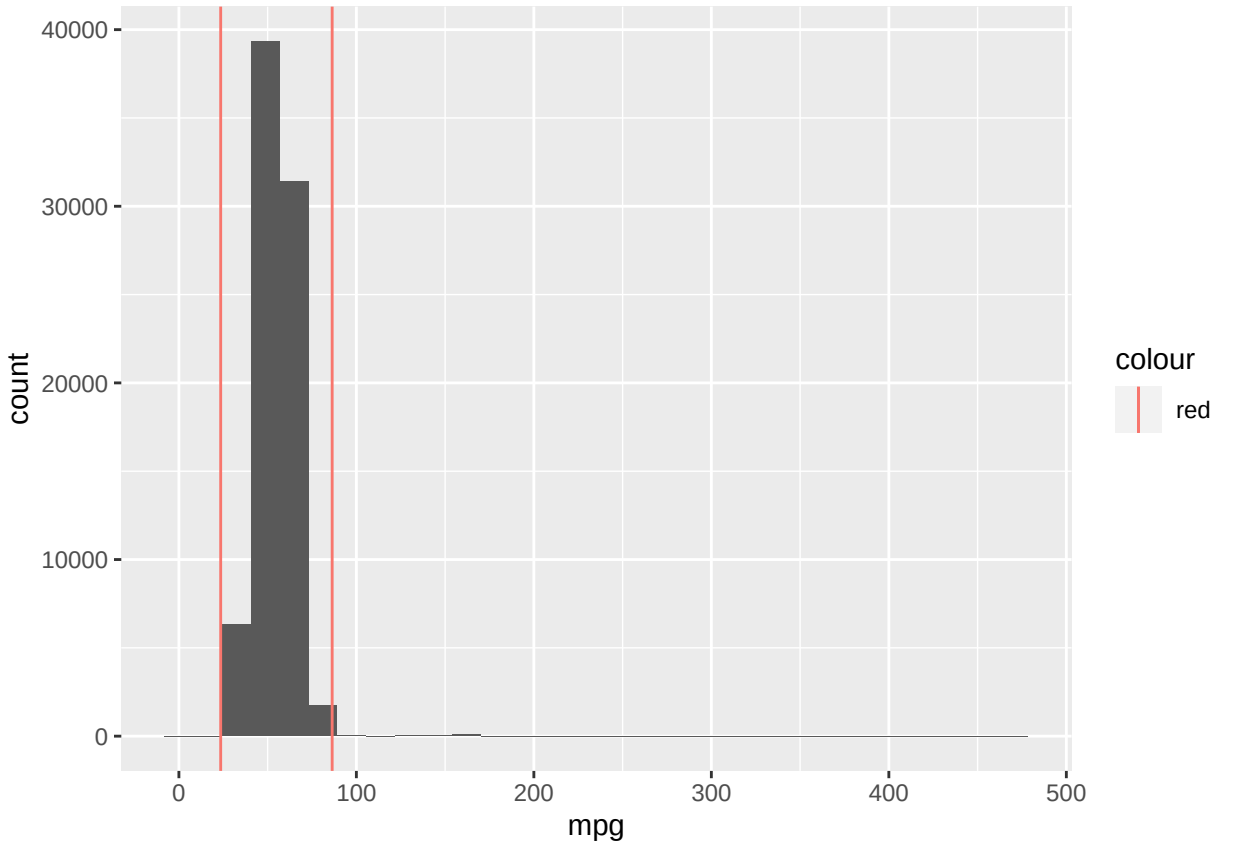
Como podemos ver el valor de 0 es muy muy bajo pues tan solo hay 6 vehículos eléctricos que finalmente eliminaremos dado el bajo volumen, no podemos hacer predicciones ni análisis con un volumen tan bajo. Por lo que finalmente eliminaremos outliers superiores volumen de motor superior a 3.2 y 0.

```
used_cars <- used_cars %>% filter(!(engineSize <= 0 | engineSize > 3.2))
```

Como hemos decidido que no vamos a tomar los vehículos eléctricos por su baja frecuencia eliminaremos también los outliers inferiores de mpg, ya que si ploteamos su histograma tiene sentido:

```
ggplot(used_cars) + geom_histogram(aes(mpg)) + geom_vline(aes(xintercept=23.55, color="red")) + geom_vl
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
used_cars <- used_cars %>% filter(!(mpg <= 23.55 | mpg > 86.35))
```

Finalmente tras el análisis de outliers el conjunto de datos numérico que tenemos tiene unas estadísticas:

```
sk2 <- skim(used_cars)
sk2 %>% yank('numeric') %>% select(c(skim_variable, mean, sd, p25, p50, p75, p0, p100)) %>% rename(Varia
```

Variable	Media	Desviación Típica	Q1	Q2/Mediana	Q3	Mínimo	Máximo
year	2017.284	1.581	2016.0	2017.0	2019.0	2012	2019
price	15859.292	6877.007	10495.0	14499.0	19990.0	2400	37116
mileage	20366.434	15467.256	8011.0	16635.0	29797.0	1	69700
tax	122.514	49.314	125.0	145.0	145.0	10	200
mpg	54.634	9.619	47.9	55.0	61.4	28	86
engineSize	1.601	0.452	1.2	1.5	2.0	1	3

Con una dimension de 78728 filas por 10 columnas, donde se han eliminado 20459 filas.

3.3 Inconsistencias del dominio del problema

Las principales incosistencias entre variables podían ocurrir entre los fuelType="Electric" y los atributos referentes a un motor que utiliza combustible en lugar de ser alimentado por una batería. (mpg/engineSize).

También, como hemos visto, el año de matriculación no puede ser superior a la fecha de scrapping, incluso hay fechas superiores a la fecha actual.

En cuanto aL tratamiento de variables cualitativas dado que hay variables que han sido eliminadas debemos reformular el dataset para eliminar los valores que no se usan de los niveles de los factores.

```
used_cars[] <- lapply(used_cars, function(x) if(is.factor(x)) factor(x) else x)

sk3 <- skim(used_cars)
sk3 %>% yank('factor') %>% select(c(skim_variable, n_unique, top_counts)) %>% rename(Variable=skim_varia
```

Variable	Niveles	Top
manufacturer	9	for: 14573, vw: 12333, vau: 11927, mer: 10152
model	153	Fie: 4742, Gol: 3891, Foc: 3771, C C: 3131
transmission	4	Man: 46771, Sem: 17733, Aut: 14218, Oth: 6
fuelType	4	Pet: 46096, Die: 31186, Hyb: 1306, Oth: 140

Cabe a interpretación si nos interesa mantener valores como "Others" para transmission (6 ocurrencias) y para fuelType (140) ya que más que aportar valor podrían introducir incertidumbre en según que análisis. Por otra parte podríamos considerar acotar el análisis a coches con gasolina y diésel dado que el volumen de coches híbridos es bajo (tan sólo el 3.1% de los coches de la muestra).

Haremos únicamente la eliminación de "Other"s:

```
used_cars <- used_cars %>% filter(!(transmission=="Other" | fuelType=="Other"))

used_cars[] <- lapply(used_cars, function(x) if(is.factor(x)) factor(x) else x)

sk3 <- skim(used_cars)
sk3 %>% yank('factor') %>% select(c(skim_variable, n_unique, top_counts)) %>% rename(Variable=skim_varia
```

Variable	Niveles	Top
manufacturer	9	for: 14573, vw: 12268, vau: 11924, mer: 10151
model	153	Fie: 4742, Gol: 3873, Foc: 3771, C C: 3131
transmission	3	Man: 46748, Sem: 17733, Aut: 14101
fuelType	3	Pet: 46092, Die: 31184, Hyb: 1306

Las dimensiones finales de nuestro dataset son 78582 filas x 10 columnas.

Como tónica general se ha procedido siempre a la eliminación de valores extremos, tras un análisis concienzudo de las variables en el dominio del problema, debido a que contabamos con un número amplio de registros en nuestro dataset. Si hubiera sido más reducido quizás tendríamos que pensar en buscar la manera de no eliminar dimensionalidad si no imputar estos valores por media, mediana, clasificadociones, regresiones, etc.