

# Momento de Retroalimentación: Módulo 1 Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos (Portafolio Análisis)

Jorge Eduardo de Leon Reyna - A00829759

Consideraciones: Para el desarrollo de esta actividad se hizo uso de la herramienta ChatGPT para tres tareas:

- 1. Comentar el código con el fin de hacerlo más legible y entendible para cualquier persona.
- 2. Detección y corrección de errores de sintaxis del código.
- 3. Consulta de funcionamiento de funciones y guía de sintaxis del código.

## Descripción del problema

Una empresa automovilística china aspira a entrar en el mercado estadounidense. Desea establecer allí una unidad de fabricación y producir automóviles localmente para competir con sus contrapartes estadounidenses y europeas. Contrataron una empresa de consultoría de automóviles para identificar los principales factores de los que depende el precio de los automóviles, específicamente, en el mercado estadounidense, ya que pueden ser muy diferentes del mercado chino. Esencialmente, la empresa quiere saber:

- Qué variables son significativas para predecir el precio de un automóvil
- Qué tan bien describen esas variables el precio de un automóvil

## Analisis

### 1. Exploración y preparación de la base de datos (Portafolio de Análisis)

#### a) Exploración de la base de datos

```
1 #Cargando datos
2 data = read.csv("/content/precios_autos.csv")
3 head(data)
```

A data.frame: 6 × 21

	symboling	CarName	fueltype	carbody	drivewheel	enginelocation	wheelbase	carlength	carwidth	carheight	...	enginetype
	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	...	<chr>
1	3	alfa-romero giulia	gas	convertible	rwd	front	88.6	168.8	64.1	48.8	...	di
2	3	alfa-romero stelvio	gas	convertible	rwd	front	88.6	168.8	64.1	48.8	...	di
3	1	alfa-romero Quadrifoglio	gas	hatchback	rwd	front	94.5	171.2	65.5	52.4	...	o
4	2	audi 100 ls	gas	sedan	fwd	front	99.8	176.6	66.2	54.3	...	i
5	2	audi 100ls	gas	sedan	4wd	front	99.4	176.6	66.4	54.3	...	i
6	2	audi fox	gas	sedan	fwd	front	99.8	177.3	66.3	53.1	...	i

```
1 # Instalacion de paquetes necesarios
2 install.packages("ggplot2")
3 install.packages("gridExtra")
4 install.packages("reshape2")

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

```
1 install.packages("caret")

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

Calcula medidas estadísticas apropiadas para las variables cuantitativas (media, desviación estándar, cuantiles, etc) y para las cualitativas (cuantiles, frecuencias)

```
1 # separacion de variables numericas y categoricas
2 categorical_vars <- c("symboling", "CarName", "fueltype", "carbody", "drivewheel", "enginelocation", "enginetype", "cylindernum")
3 numeric_vars <- c("wheelbase", "carlength", "carwidth", "carheight", "curbweight", "enginesize", "stroke", "compressionratio",
4                  "horsepower", "peakrpm", "citympg", "highwaympg", "price")

1 # calculo de medidas estadisticas de valores numericos
2
3 # funcion para calcular medidas estadisticas
4 calculate_statistics <- function(x) {
5   c(mean = mean(x, na.rm = TRUE),      # Calcula la media de 'x' ignorando los valores nulos
6     median = median(x, na.rm = TRUE),  # Calcula la mediana de 'x' ignorando los valores nulos
7     sd = sd(x, na.rm = TRUE),          # Calcula la desviación estándar de 'x' ignorando los valores nulos
8     q25 = quantile(x, probs = 0.25),   # Calcula el primer cuartil (percentil 25) de 'x'
9     q50 = quantile(x, probs = 0.50),   # Calcula el segundo cuartil (mediana) de 'x'
10    q75 = quantile(x, probs = 0.75))    # Calcula el tercer cuartil (percentil 75) de 'x'
11 }
12
13 # Para cada variable numérica en el conjunto de datos 'data', se aplica la función 'calculate_statistics',
14 # lo que da como resultado una lista de vectores con las medidas estadísticas calculadas.
15 statistics_list <- lapply(data[numeric_vars], calculate_statistics)
16
17 # Se combina la lista de vectores en una matriz, donde cada fila representa una variable y cada columna una medida estadística
18 statistics_matrix <- do.call(rbind, statistics_list)
19
20 # La matriz se convierte en un DataFrame, lo que facilita su manipulación y visualización.
21 statistics_df <- as.data.frame(statistics_matrix)
22
23 # Los nombres de las variables numéricas se asignan como nombres de fila en el DataFrame 'statistics_df'.
24 rownames(statistics_df) <- numeric_vars
25
26 # Finalmente, se imprime en la consola el DataFrame 'statistics_df' que contiene las medidas estadísticas calculadas.
27 print(statistics_df)
```

	mean	median	sd	q25.25%	q50.50%	q75.75%
wheelbase	98.756585	97.00	6.021776	94.50	97.00	102.40
carlength	174.049268	173.20	12.337289	166.30	173.20	183.10
carwidth	65.907805	65.50	2.145204	64.10	65.50	66.90
carheight	53.724878	54.10	2.443522	52.00	54.10	55.50
curbweight	2555.565854	2414.00	520.680204	2145.00	2414.00	2935.00
enginesize	126.907317	120.00	41.642693	97.00	120.00	141.00
stroke	3.255415	3.29	0.313597	3.11	3.29	3.41
compressionratio	10.142537	9.00	3.972040	8.60	9.00	9.40
horsepower	104.117073	95.00	39.544167	70.00	95.00	116.00
peakrpm	5125.121951	5200.00	476.985643	4800.00	5200.00	5500.00
citympg	25.219512	24.00	6.542142	19.00	24.00	30.00
highwaympg	30.751220	30.00	6.886443	25.00	30.00	34.00
price	13276.710571	10295.00	7988.852332	7788.00	10295.00	16503.00

```
1 # calculo de medidas estadisticas de valores categoricos
2
3 # Cargar la biblioteca requerida
4 library(knitr)
5
6 # Función personalizada para calcular estadísticas de variables cualitativas
7 calculate_qualitative_statistics <- function(x) {
8   frequency_table <- table(x) # Calcula una tabla de frecuencias para la variable 'x'
9   most_common_value <- names(frequency_table)[which.max(frequency_table)] # Encuentra el valor más común en la tabla de frecue
10
11 # Devuelve una lista con la tabla de frecuencias y el valor más común
12 list(frequency_table = frequency_table,
13      most_common_value = most_common_value)
14 }
15
16 # Aplica la función 'calculate_qualitative_statistics' a cada variable cualitativa en el conjunto de datos 'data',
17 # lo que resulta en una lista de listas con las tablas de frecuencias y los valores más comunes.
```

```

18 qualitative_statistics_list <- lapply(data[categorical_vars], calculate_qualitative_statistics)
19
20 # Imprime los resultados como tablas
21 for (i in seq_along(categorical_vars)) {
22   cat("Variable:", categorical_vars[i], "\n")
23
24   # Crea un DataFrame con las columnas "Frequency" y "Most_Common_Value" usando los datos de la lista correspondiente.
25   table_data <- data.frame(Frequency = qualitative_statistics_list[[i]]$frequency_table,
26                             Most_Common_Value = qualitative_statistics_list[[i]]$most_common_value)
27
28   # Imprime la tabla utilizando el formato "markdown" proporcionado por la función 'kable'.
29   print(kable(table_data, format = "markdown"))
30
31   cat("\n")
32 }
33

```

Variable: symboling

Frequency.x	Frequency.Freq	Most_Common_Value
-2	3	0
-1	22	0
0	67	0
1	54	0
2	32	0
3	27	0

Variable: CarName

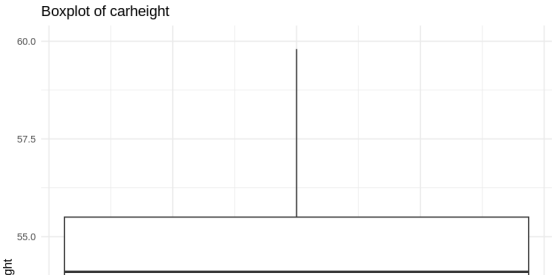
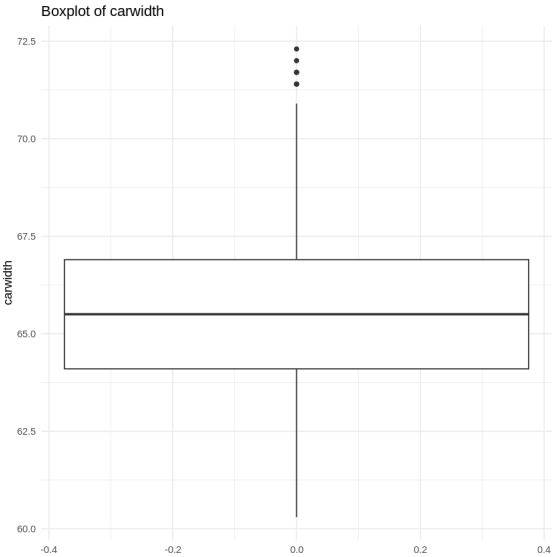
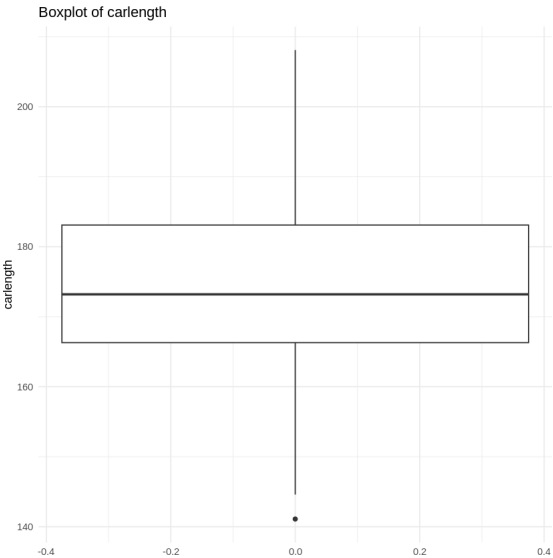
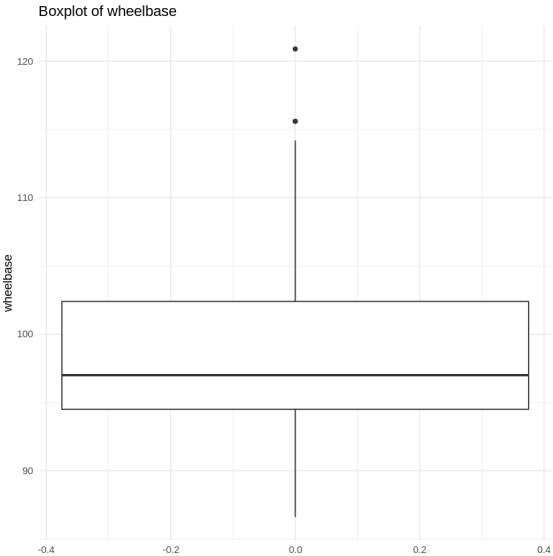
Frequency.x	Frequency.Freq	Most_Common_Value
alfa-romero giulia	1	peugeot 504
alfa-romero Quadrifoglio	1	peugeot 504
alfa-romero stelvio	1	peugeot 504
audi 100 ls	1	peugeot 504
audi 100ls	2	peugeot 504
audi 4000	1	peugeot 504
audi 5000	1	peugeot 504
audi 5000s (diesel)	1	peugeot 504
audi fox	1	peugeot 504
bmw 320i	2	peugeot 504
bmw x1	1	peugeot 504
bmw x3	2	peugeot 504
bmw x4	1	peugeot 504
bmw x5	1	peugeot 504
bmw z4	1	peugeot 504
buick century	1	peugeot 504
buick century luxus (sw)	1	peugeot 504
buick century special	1	peugeot 504
buick electra 225 custom	1	peugeot 504
buick opel isuzu deluxe	1	peugeot 504
buick regal sport coupe (turbo)	1	peugeot 504
buick skyhawk	1	peugeot 504
buick skylark	1	peugeot 504
chevrolet impala	1	peugeot 504
chevrolet monte carlo	1	peugeot 504
chevrolet vega 2300	1	peugeot 504
dodge challenger se	1	peugeot 504
dodge colt (sw)	1	peugeot 504
dodge colt hardtop	1	peugeot 504
dodge coronet custom	1	peugeot 504
dodge coronet custom (sw)	1	peugeot 504
dodge d200	1	peugeot 504
dodge dart custom	1	peugeot 504
dodge monaco (sw)	1	peugeot 504
dodge rampage	1	peugeot 504
honda accord	2	peugeot 504
honda accord cvcc	1	peugeot 504
honda accord lx	1	peugeot 504
honda civic	3	peugeot 504
honda civic (auto)	1	peugeot 504
honda civic 1300	1	peugeot 504

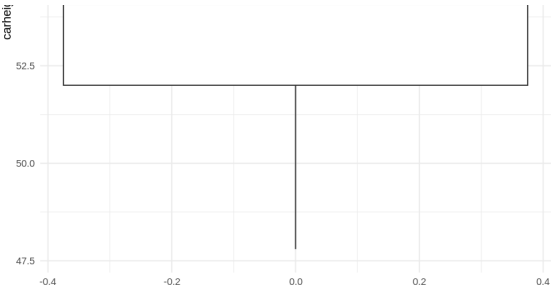
Explora los datos usando herramientas de visualización:

#### 1. Variables cuantitativas:

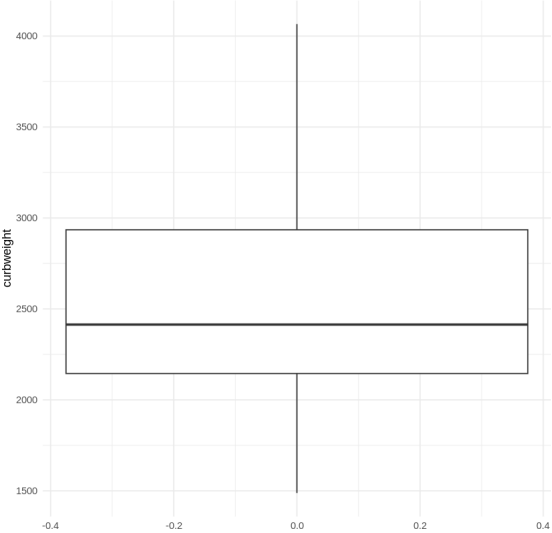
- Medidas de posición: cuartiles, outlier (valores atípicos), boxplots
- Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica
- Analiza colinealidad (coeficiente de correlación y diagramas de dispersión)

```
1 # Boxplots para variables numericas
2
3 # Cargar la biblioteca necesaria para hacer gráficos
4 library(ggplot2)
5
6 # Crear una función para generar y mostrar boxplots para cada variable
7 generate_boxplots <- function(data) {
8   # Iterar sobre cada columna numérica
9   for (col in numeric_vars) {
10    # Crear un objeto de gráfico de ggplot con un boxplot
11    p <- ggplot(data, aes(y = .data[[col]])) +
12      geom_boxplot() + # Agregar el geom para el boxplot
13      labs(title = paste("Boxplot of", col), # Asignar título al gráfico
14           y = col) + # Etiqueta del eje Y
15      theme_minimal() # Aplicar un tema minimalista al gráfico
16
17    print(p) # Mostrar el gráfico
18  }
19 }
20
21 # Llamar a la función para generar y mostrar los boxplots
22 generate_boxplots(data)
23
```

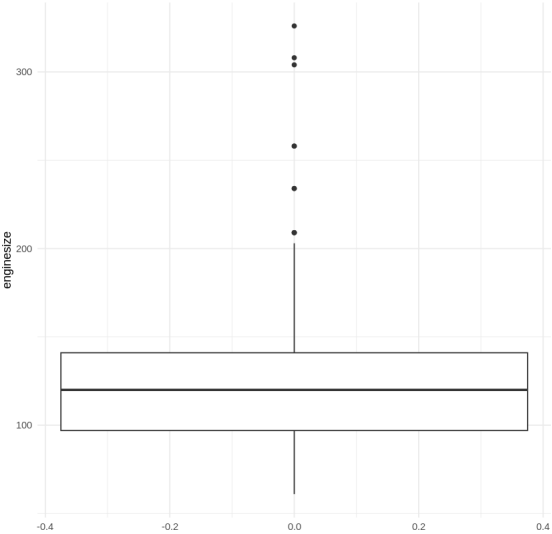




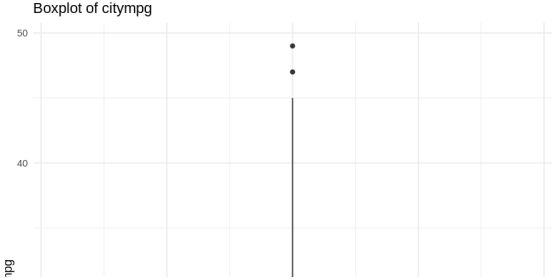
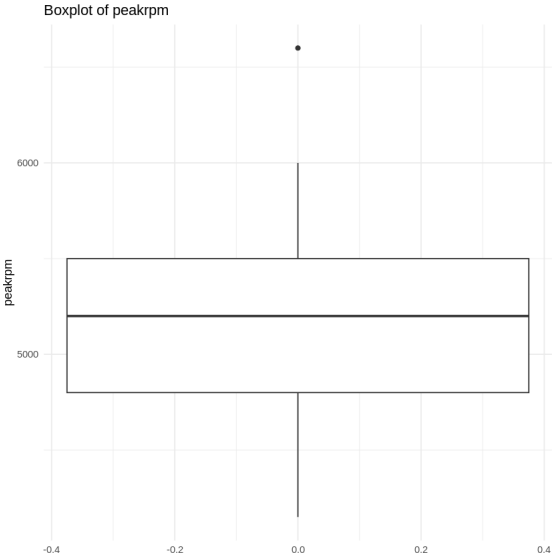
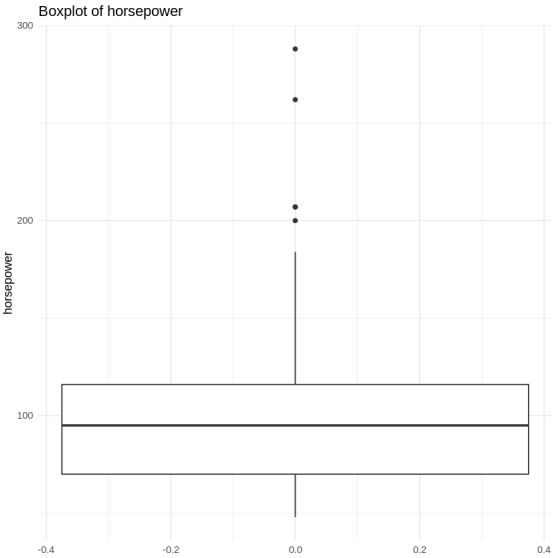
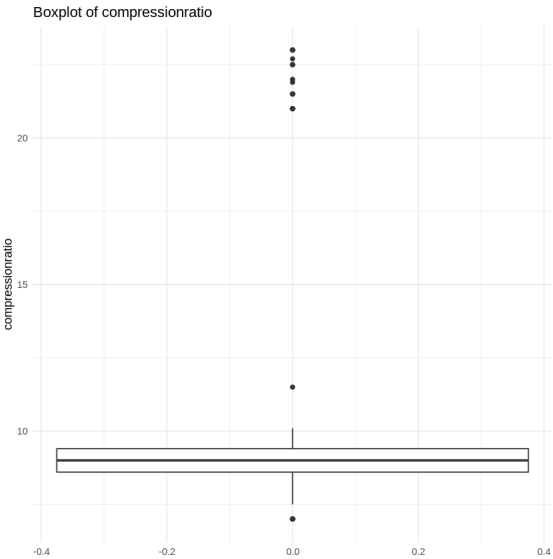
Boxplot of curbside weight

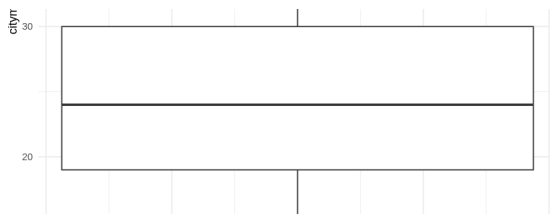


Boxplot of engine size



Boxplot of stroke

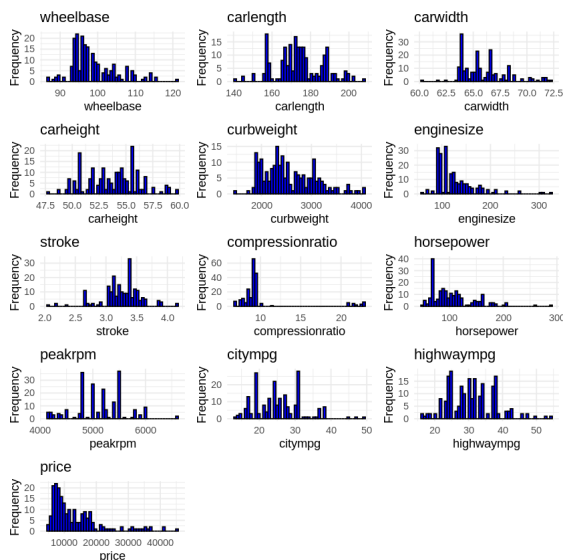




```

1 #Histogramas de frecuencia por rangos de variables numericas
2
3 # Cargar las bibliotecas necesarias para hacer gráficos y organizar
4 library(ggplot2)
5 library(gridExtra)
6
7 # Crear una función para generar histogramas por rangos para cada variable
8 generate_histograms_by_ranges <- function(data) {
9   # Crear una lista para almacenar los gráficos generados
10  plots <- list()
11
12  # Iterar sobre cada columna numérica
13  for (col in numeric_vars) {
14    # Crear un objeto de gráfico de ggplot con un histograma
15    p <- ggplot(data, aes(x = .data[[col]])) +
16      geom_histogram(binwidth = NULL, bins = 50, fill = "blue", color = "black") +
17      labs(title = paste(col), # Asignar título al gráfico
18           x = col,           # Etiqueta del eje X
19           y = "Frequency") + # Etiqueta del eje Y
20      theme_minimal()         # Aplicar un tema minimalista al gráfico
21    plots[[col]] <- p # Agregar el gráfico a la lista 'plots' bajo el nombre de la columna actual
22  }
23
24  grid.arrange(grobs = plots, ncol = 3) # Organizar y mostrar los gráficos en una matriz (3 columnas en este caso)
25 }
26
27 # Llamar a la función para generar y organizar los histogramas por rangos en una matriz
28 generate_histograms_by_ranges(data)

```



```

1 # Matriz de correlacion de variables numericas
2
3 # Calcular los coeficientes de correlación
4 correlation_matrix <- cor(data[, numeric_vars])
5
6 # Cargar las bibliotecas necesarias para hacer gráficos
7 library(ggplot2)
8 library(reshape2)
9
10 # Convertir la matriz de correlación al formato largo
11 correlation_long <- melt(correlation_matrix)
12
13 # Crear un mapa de calor para la matriz de correlación
14 heatmap_plot <- ggplot(correlation_long, aes(x = Var1, y = Var2, fill = value)) +
15   geom_tile() +
16   scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0,

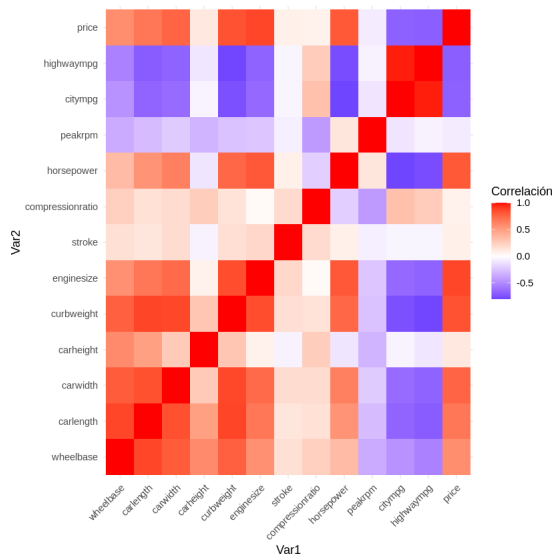
```



```

17         name = "Correlación") +
18     theme_minimal() +
19     theme(axis.text.x = element_text(angle = 45, hjust = 1))
20
21 # Mostrar el mapa de calor
22 print(heatmap_plot)
23
24

```



## 2. Variables categóricas:

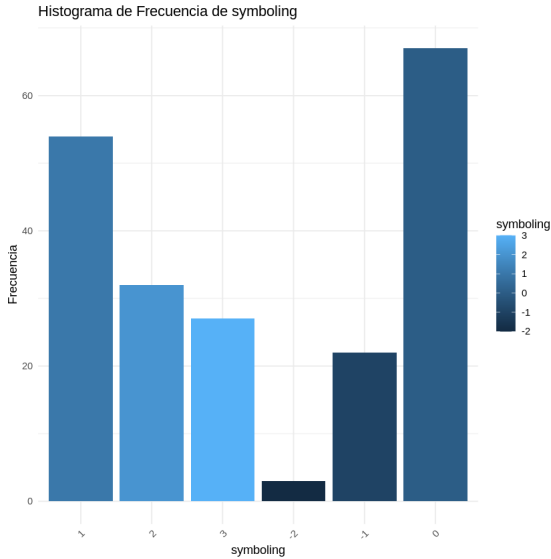
- Distribución de los datos (diagramas de barras, diagramas de pastel)
- Analiza asociación o colinealidad (diagramas de caja y bigote de precio por categoría y barras por categoría)

```

1 # Histograma para distribución de frecuencias para variables categoricas
2
3 # Iterar a través de cada variable y crear un histograma de frecuencia
4 for (var in categorical_vars) {
5     # Crear el gráfico de histograma usando ggplot2
6     plot <- ggplot(data, aes(x = reorder(.data[[var]], -table(.data[[var]])), fill = .data[[var]])) +
7         geom_bar() +
8         labs(title = paste("Histograma de Frecuencia de", var),
9              x = var,
10             y = "Frecuencia") +
11         theme_minimal() +
12         theme(axis.text.x = element_text(angle = 45, hjust = 1))
13
14 # Mostrar el gráfico
15 print(plot)
16 }
17

```

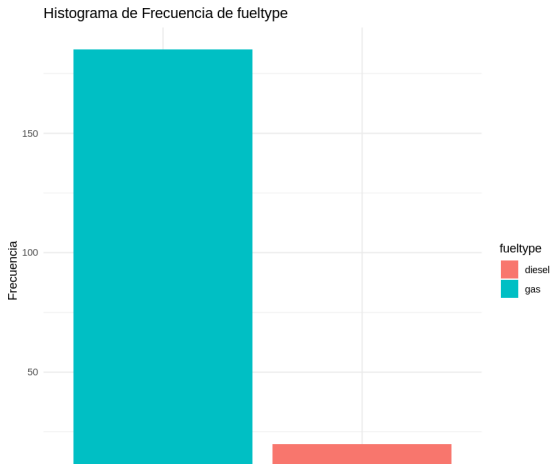
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"  
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"  
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"  
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"

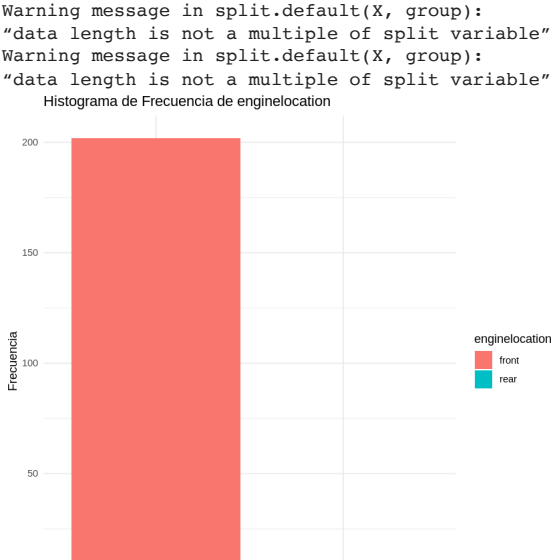
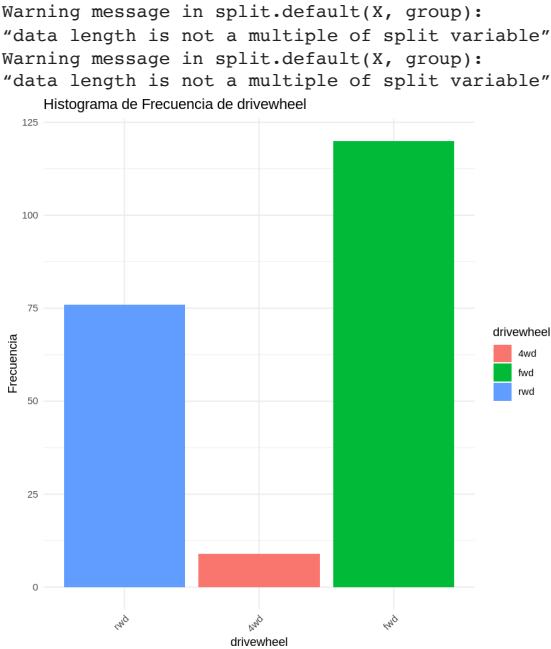
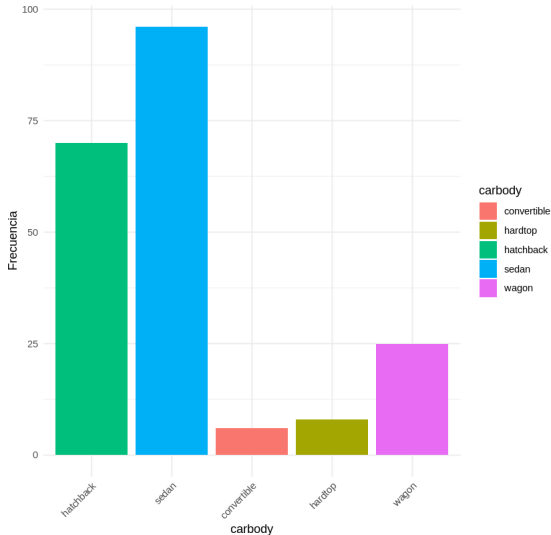
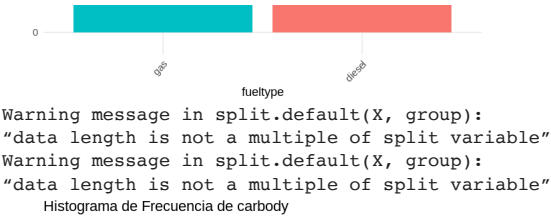


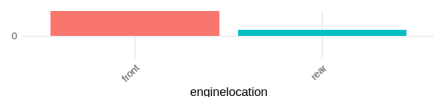
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"  
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"

?	honda civic	mazda glc deluxe	nissan leaf	porcshee panamera
?	honda civic (auto)	mazda rx-4	nissan note	porcshe boxter
?	honda civic 1300	mazda rx-7 gs	nissan nv200	porcshe cayenne
?	honda civic 1500 gl	mazda rx2 coupe	nissan otti	porcshe macan
?	honda civic cvcc	mercury cougar	nissan rogue	renault 12tl
?	honda prelude	mitsubishi g4	nissan teana	renault 5 gtl
?	isuzu D-Max	mitsubishi lancer	nissan titan	saab 99e
?	isuzu D-Max V-Cross	mitsubishi mirage	Nissan versa	saab 99gle
?	isuzu MU-X	mitsubishi mirage g4	peugeot 304	saab 99le
?	jaguar xf	mitsubishi montero	peugeot 504	subaru
?	jaguar xj	mitsubishi outlander	peugeot 504 (sw)	subaru baja
(sw)	jaguar xk	mitsubishi pajero	peugeot 505s turbo diesel	subaru brz
	mazda glc deluxe	nissan clipper	peugeot 604sl	subaru dl
	mazda rx3	nissan dayz	plymouth cricket	subaru r1
	mazda 626	nissan fuga	plymouth duster	subaru r2
	mazda glc	nissan gt-r	plymouth fury gran sedan	subaru trezia
	mazda glc 4	nissan juke	plymouth fury iii	subaru tribeca
	mazda glc custom	nissan kicks	plymouth satellite custom (sw)	toyota carina
	mazda glc custom l	nissan latio	plymouth valiant	toyota celica gt

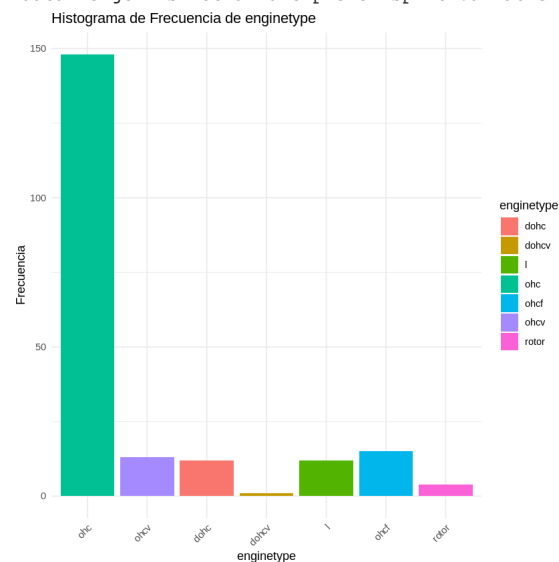
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"  
Warning message in split.default(X, group):  
"data length is not a multiple of split variable"



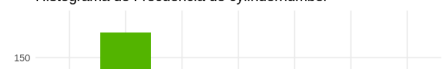




Warning message in split.default(X, group):  
 "data length is not a multiple of split variable"  
 Warning message in split.default(X, group):  
 "data length is not a multiple of split variable"



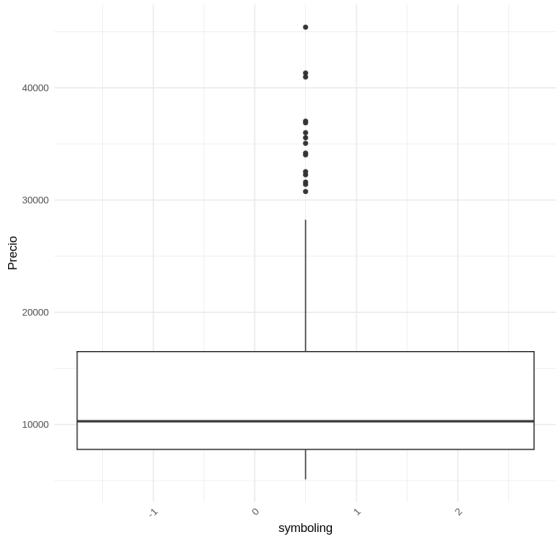
Histograma de Frecuencia de cylindernumber



```
1 #Boxplots para comparar cada variable categorica con el precio (variable objetivo)
2
3 # Iterar a través de cada variable y crear un diagrama de caja
4 for (var in categorical_vars) {
5   # Crear el diagrama de caja usando ggplot2
6   plot <- ggplot(data, aes(x = .data[[var]], y = price, fill = .data[[var]])) +
7     geom_boxplot() +
8     labs(title = paste("Diagrama de Caja de", var, "vs Precio"),
9          x = var,
10         y = "Precio") +
11     theme_minimal() +
12     theme(axis.text.x = element_text(angle = 45, hjust = 1))
13
14 # Mostrar el gráfico
15 print(plot)
16 }
17
```

```
Warning message:
"Continuous x aesthetic
i did you forget `aes(group = ...)`?"
Warning message:
"The following aesthetics were dropped during statistical transformation: fill
i This can happen when ggplot fails to infer the correct grouping structure in
the data.
i Did you forget to specify a `group` aesthetic or to convert a numerical
variable into a factor?"
```

Diagrama de Caja de symboling vs Precio



axe	honda civic	mazda glc deluxe	nissan leaf	porsche panamera
ipe (turbo)	honda civic (auto)	mazda rx-4	nissan note	porsche boxster
	honda civic 1300	mazda rx-7 gs	nissan nv200	porsche cayenne
	honda civic 1500 gl	mazda rx2 coupe	nissan otti	porsche macan
	honda civic cvcc	mercury cougar	nissan rogue	renault 12i
lo	honda prelude	mitsubishi g4	nissan teana	renault 5 gl
)	isuzu d-max	mitsubishi lancer	nissan titan	saab 99e
:	isuzu d-max v-cross	mitsubishi mirage	nissan versa	saab 99gle
	isuzu mu-x	mitsubishi mirage g4	peugeot 304	saab 99le
	jaguar xf	mitsubishi montero	peugeot 504	subaru
jm	jaguar xj	mitsubishi outlander	peugeot 504 (sw)	subaru baja
jm (sw)	jaguar xk	mitsubishi pajero	peugeot 505s turbo diesel	subaru brz
	mazda glc deluxe	nissan clipper	peugeot 604sl	subaru di
	mazda rx3	nissan dayz	plymouth cricket	subaru r1
	mazda 626	nissan fuga	plymouth duster	subaru r2
	mazda glc	nissan gt-r	plymouth fury gran sedan	subaru trezia
	mazda glc 4	nissan juke	plymouth fury ii	subaru tribeca
	mazda glc custom	nissan kicks	plymouth satellite custom (sw)	toyota carina
	mazda glc custom i	nissan latio	plymouth valiant	toyota celica gt

Diagrama de Caja de fueuetype vs Precio

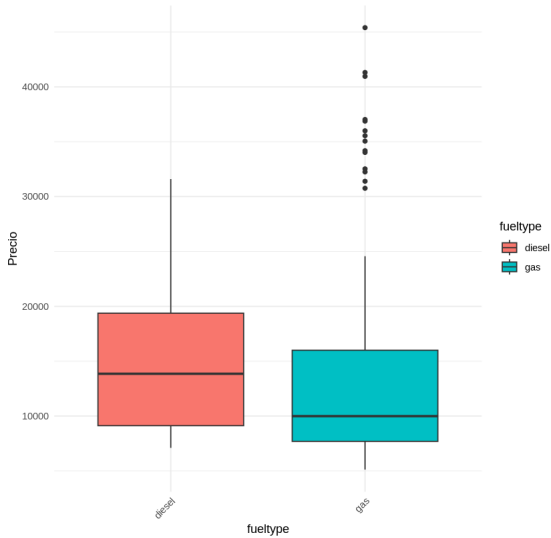


Diagrama de Caja de carbody vs Precio

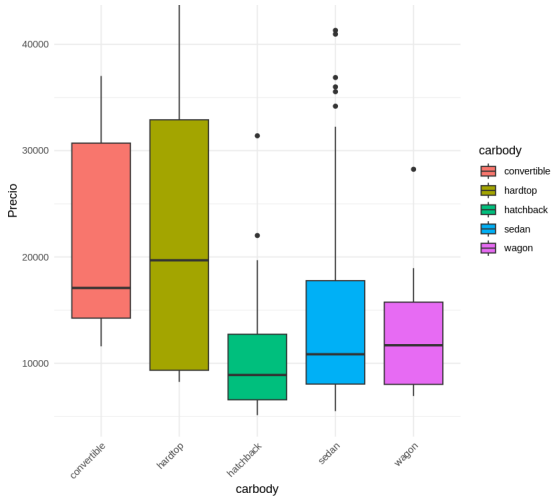


Diagrama de Caja de drivewheel vs Precio

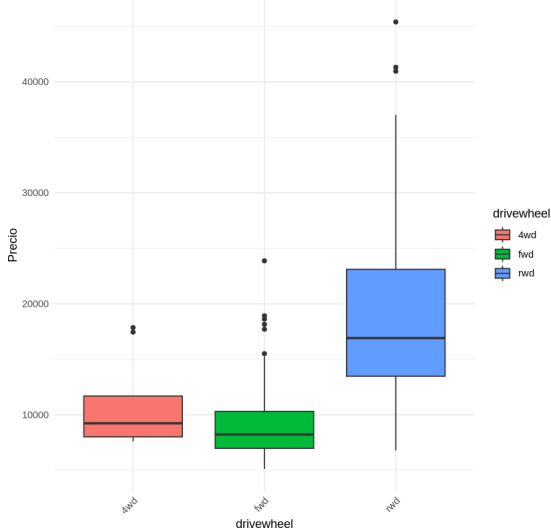


Diagrama de Caja de enginelocation vs Precio

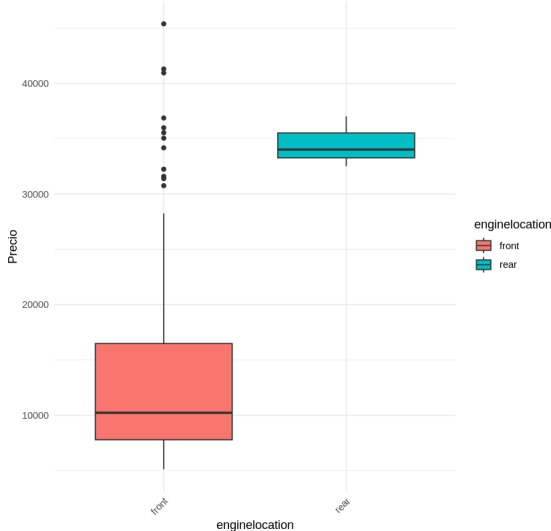
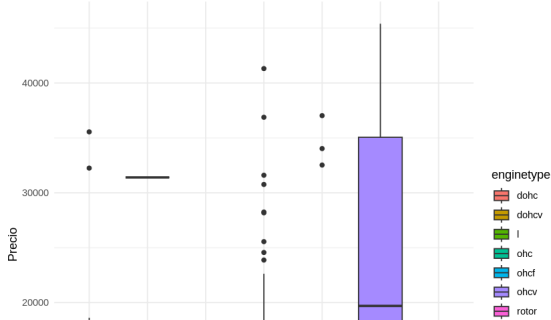
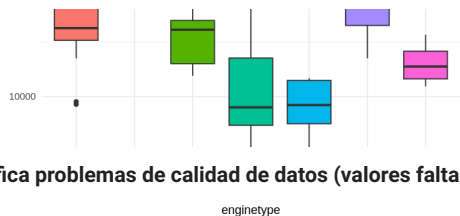


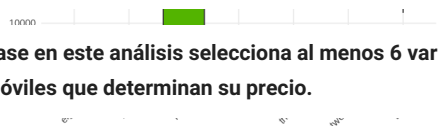
Diagrama de Caja de enginetype vs Precio





### Identifica problemas de calidad de datos (valores faltantes, outliers, etc).

- Identificación de Valores Atípicos: Mediante el uso de diagramas de caja (boxplots) se ha detecto que las variables **carwidth**, **enginesize**, **stroke**, **horsepower** y **price** presentan una mayor cantidad de valores atípicos en comparación con el resto de variables lo cual sugiere que estos atributos podrían contener observaciones que se desvían significativamente del comportamiento general de los datos.
- Evaluación de Sesgo Aparente: La representación grafica de histogramas ha permitido visualizar el sesgo en ciertas variables. En particular, se observa un sesgo aparente en las variables **wheelbase**, **enginesize**, **compressionratio**, **horsepower** y **price**. Estos resultados indican que la distribución de datos en estas variables no es simétrica y tiende a inclinarse hacia un extremo, lo que podría tener implicaciones en los análisis posteriores.
- Identificación de Correlaciones Relevantes: Mediante el cálculo de la matriz de correlación, se ha destacado un conjunto de variables que muestran una fuerte correlacion con la variable objetivo. Estas variables son: **horsepower**, **enginesize**, **curbweight**, **carwidth**, **highwaympg** y **citympg**. La existencia de correlaciones significativas entre estas características y la variable objetivo sugiere que podrían desempeñar un papel crucial en la predicción o explicación de la variable objetivo.



### Con base en este análisis selecciona al menos 6 variables que consideres importantes para el análisis de las características de los automóviles que determinan su precio.

Para la seleccion de las variables a utilizar se tomaron en cuenta la correlacion con la variable objetivo (el precio) para despues pasar a un procesamiento posterior donde se escalen y transformen estas variables para poder ser utilizadas con efectividad.

Para el primer factor se consideran por tanto las siguientes variables: **horsepower**, **enginesize**, **curbweight**, **carwidth**, **cylindernumber**, **drivewheel** (estas ultimas dos categoricas en basea su frecuencia y a lo observado en su relacion con la variable price en los boxplot correspondientes).

Posterior a esto, se identifica que en ciertas variables seleccionadas se presentan datos atípicos importantes y distribuciones no normales lo que podría complicar su posterior utilizacion en un modelo por lo que tendran que ser procesadas mediante metodos de transformacion y escalamiento.

#### ▼ b) Preparación de la base de datos

Selecciona el conjunto de datos a utilizar.

- Maneja datos categóricos: transforma a variables dummy si es necesario.
- Maneja apropiadamente datos atípicos.

```
1 # Seleccion de columnas
2 filtered_data = data[, c("horsepower", "enginesize", "curbweight", "carwidth", "cylindernumber", "drivewheel", "price")]
3 head(filtered_data)
```

A data.frame: 6 × 7

	horsepower	enginesize	curbweight	carwidth	cylindernumber	drivewheel	price
	<int>	<int>	<int>	<dbl>	<chr>	<chr>	<dbl>
1	111	130	2548	64.1	four	rwd	13495
2	111	130	2548	64.1	four	rwd	16500
3	154	152	2823	65.5	six	rwd	16500
4	102	109	2337	66.2	four	fwd	13950
5	115	136	2824	66.4	five	4wd	17450
6	110	136	2507	66.3	five	fwd	15250

```
1 # Se aplica one-hot encoding para las variables categoricas
2
3 # Se aplica el one-hot encoding a los datos
4 encoded_data_matrix <- model.matrix(~ . - 1, data = filtered_data)
5
```

```

6 final_data <- as.data.frame(encoded_data_matrix)
7
8 # Imprimos el resultado
9 head(final_data)

```

	horsepower	enginesize	curbweight	carwidth	cylindernumbereight	cylindernum
	<dbl>	<dbl>	<dbl>	<dbl>		<dbl>
1	111	130	2548	64.1		0
2	111	130	2548	64.1		0
3	154	152	2823	65.5		0
4	102	109	2337	66.2		0
5	115	136	2824	66.4		0
6	110	136	2507	66.3		0

Transforma los datos en caso necesario.

- Revisa si es necesario discretizar los datos
- Revisa si es necesario escalar y normalizar los datos

Dada la naturaleza de los datos se plantea implementar un metodo de transformacion para que las variables horsepower y enginesize tomen un comportamiento mas cercano a una distribucion normal.

## Conclusion

En este primer analisis se pudo aprender sobre el comportamiento de los datos, tanto cualitativos y los cuantitativos. Haciendo uso de funciones de R pudimos conocer sobre las distribuciones, comportamientos, datos atípicos, etc. asi como tambien sobre sus representaciones graficas como los histogramas y los diagramas de caja los cuales nos permitieron seleccionar que variables podrían funcionar mejor para un futuro modelo.

Finalmente, se observo que algunas variables presentan comportamiento no ideal para su implementacion en un modelo por lo que se propone la implementación de una tecnica de transformación para un mejor desempeño.

1