Actividad Integradora 2

Héctor Hibran Tapia Fernández - A01661114

2024-09-06

Actividad Integradora 2

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

Con base en varias encuestas de mercado, la consultora ha recopilado un gran conjunto de datos de diferentes tipos de automóviles en el mercado estadounidense que presenta en el siguiente archivo “Download” archivo. Las variables recopiladas vienen descritas en el diccionario de términos “diccionario de términos”. Por un análisis de correlación, la empresa automovilistica tiene interés en analizar las variables agrupadas de la siguiente forma para hacer el análisis de variables significativas:

1. **Primer grupo.** Distancia entre los ejes (wheelbase), tipo de gasolina que usa y caballos de fuerza
2. **Segundo grupo.** Altura del auto, ancho del auto y si es convertible o no.
3. **Tercer grupo.** Tamaño del motor (ensinesize), carrera o lanzamiento del pistón (stroke) y localización del motor en el carro

Selecciona uno de los tres grupos analizados (te será asignado por tu profesora) y analiza la significancia de las variables para predecir o influir en la variable precio. ¿propondrías una nueva agrupación a la empresa automovilísitica?

I. Exploración de la base de datos

1. Exploración de la base de datos

-> Calcula medidas estadísticas apropiadas para las variables:

- cuantitativas (media, desviación estándar, cuantiles, etc)

```
df = read.csv("./precios_autos.csv")
df_tercer_grupo = df[, c("enginetype", "enginelocation", "stroke", "price")]

library(psych)
describe(df_tercer_grupo)
```

```
##                 vars   n     mean      sd   median  trimmed     mad     min
## enginetype*        1 205     4.01    1.05     4.00     4.04    0.00    1.00
## enginelocation*    2 205     1.01    0.12     1.00     1.00    0.00    1.00
## stroke             3 205     3.26    0.31     3.29     3.28    0.21    2.07
## price              4 205 13276.71 7988.85 10295.00 11747.30 4901.48 5118.00
##                    max   range  skew kurtosis     se
## enginetype*       7.00     6.0 -0.53     3.13   0.07
## enginelocation*   2.00     1.0  8.02    62.70   0.01
## stroke            4.17     2.1 -0.68     2.04   0.02
## price          45400.00 40282.0  1.75     2.89 557.97
```

```
summary(df_tercer_grupo)
```

```
##   enginetype        enginelocation        stroke          price
##  Length:205         Length:205         Min.   :2.070   Min.   : 5118
##  Class :character   Class :character   1st Qu.:3.110   1st Qu.: 7788
##  Mode  :character   Mode  :character   Median :3.290   Median :10295
##                                        Mean   :3.255   Mean   :13277
##                                        3rd Qu.:3.410   3rd Qu.:16503
##                                        Max.   :4.170   Max.   :45400
```

- cualitativas: cuantiles, frecuencias (puedes usar el comando table o prop.table)

```
df_cualitativas = df_tercer_grupo[sapply(df_tercer_grupo, is.factor) | sapply(df_tercer_
grupo, is.character)]

lapply(df_cualitativas, function(col) {cat("\nVariable:", colnames(as.data.frame(col)),
"\n")
  print(table(col)) # absolutas
  print(prop.table(table(col))) # relativas
})
```

```
##
## Variable: col
## col
##  dohc dohcv     l   ohc  ohcf  ohcv rotor
##    12     1    12   148    15    13     4
## col
##        dohc       dohcv           l         ohc        ohcf        ohcv
## 0.058536585 0.004878049 0.058536585 0.721951220 0.073170732 0.063414634
##       rotor
## 0.019512195
##
## Variable: col
## col
## front  rear
##   202     3
## col
##      front       rear
## 0.98536585 0.01463415
```

```
## $enginetype
## col
##        dohc       dohcv           l         ohc        ohcf        ohcv
## 0.058536585 0.004878049 0.058536585 0.721951220 0.073170732 0.063414634
##       rotor
## 0.019512195
##
## $enginelocation
## col
##      front       rear
## 0.98536585 0.01463415
```

-> Analiza la correlación entre las variables (analiza posible colinealidad entre las variables)

```
df_cuantitativas = df_tercer_grupo[sapply(df_tercer_grupo, is.numeric)]
matriz_correlacion = cor(df_cuantitativas)
matriz_correlacion
```

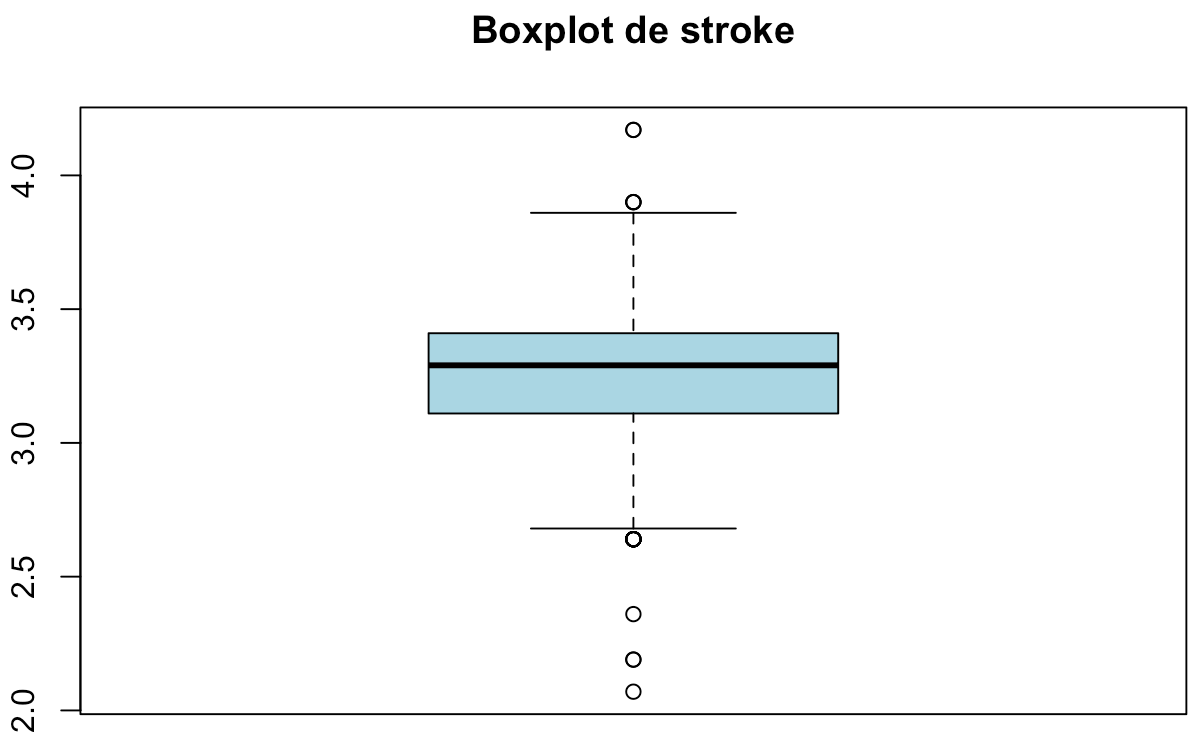
```
##            stroke      price
## stroke 1.00000000 0.07944308
## price  0.07944308 1.00000000
```

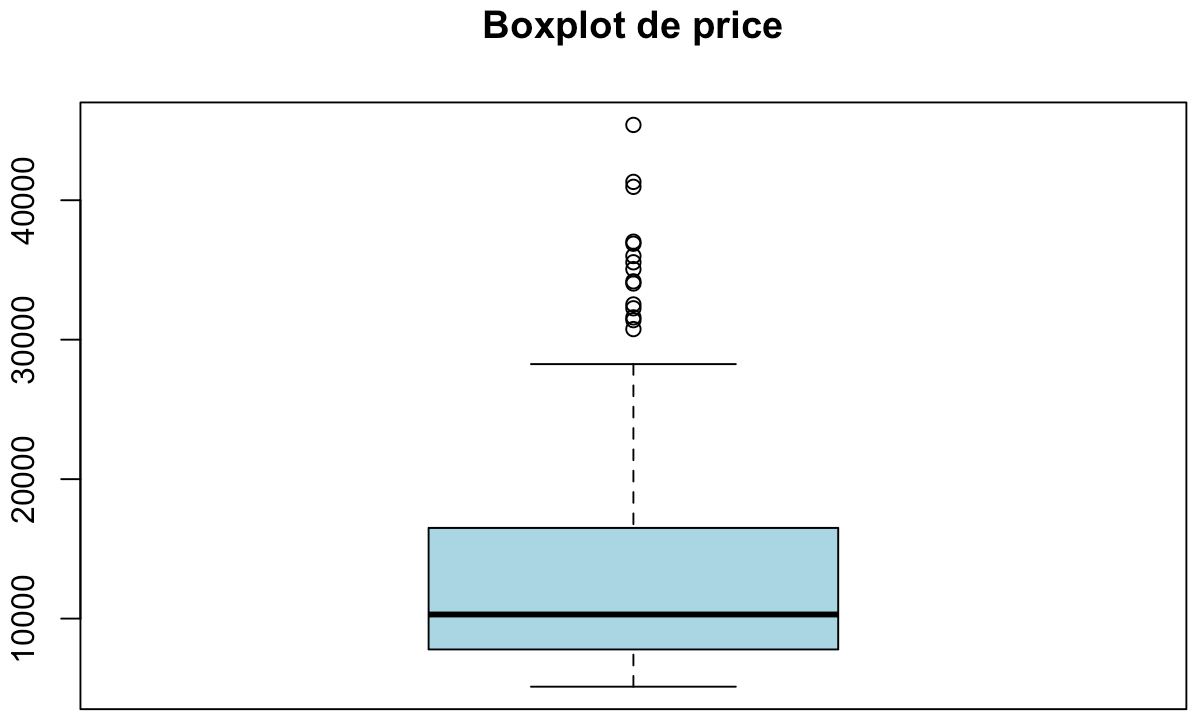
Hay colinealidad positiva muy baja entra las variables numéricas.

-> Explora los datos usando herramientas de visualización (si lo consideras necesario):

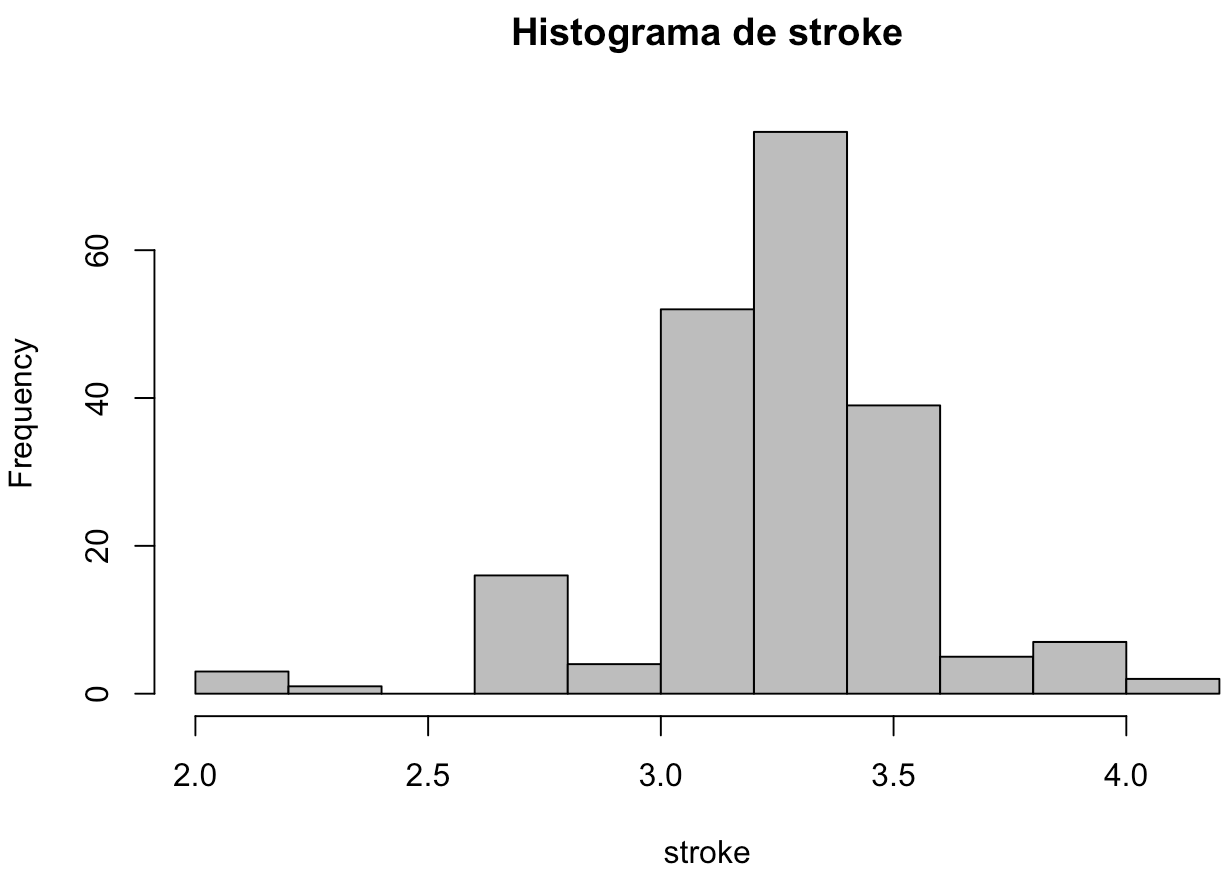
- Variables cuantitativas: Boxplot (visualización de datos atípicos), Histogramas, Diagramas de dispersión y correlación por pares.

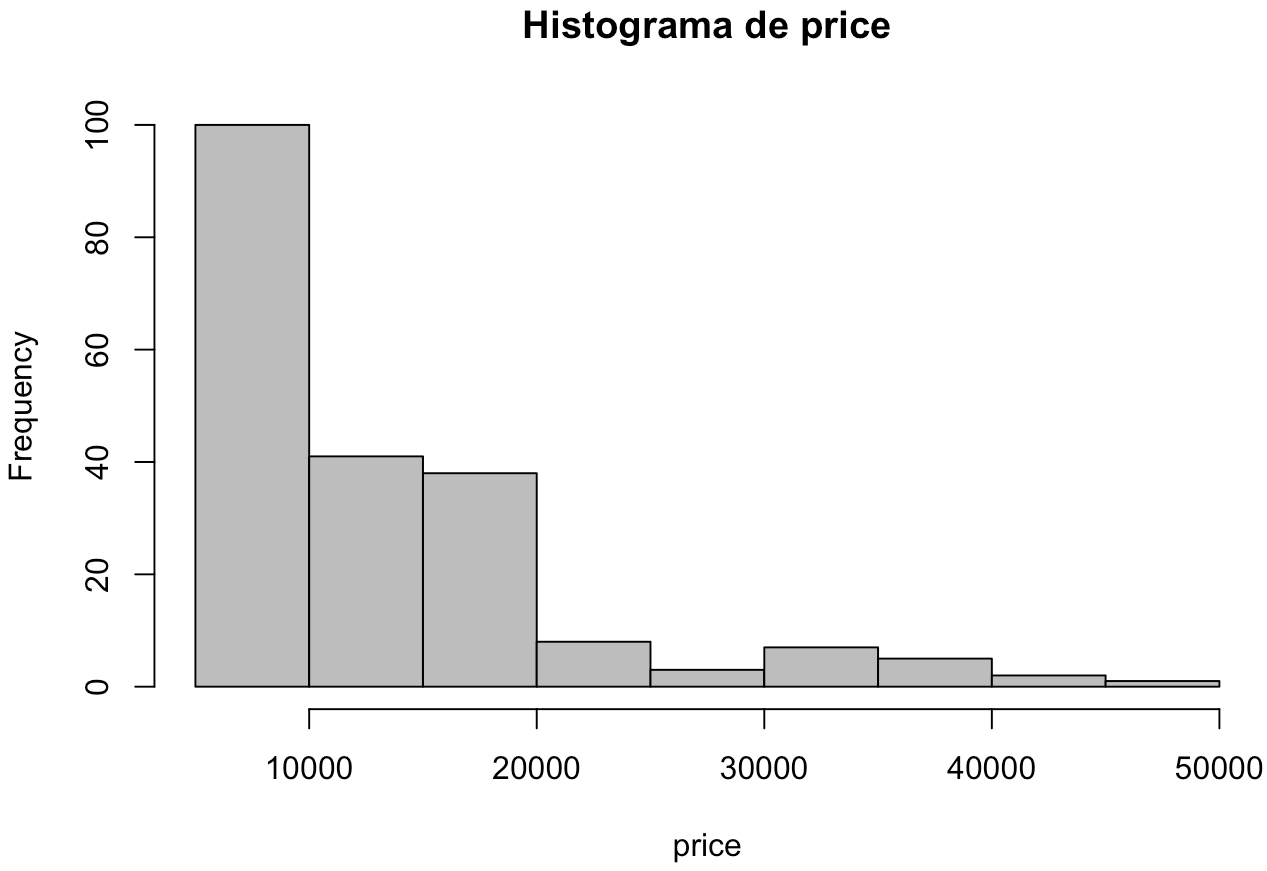
```
for (col_name in colnames(df_cuantitativas)) {boxplot(df_cuantitativas[[col_name]], main
= paste("Boxplot de", col_name), col = "lightblue", outline = TRUE)}
```


Boxplot de stroke
4.0
3.5
3.0
2.5
2.0

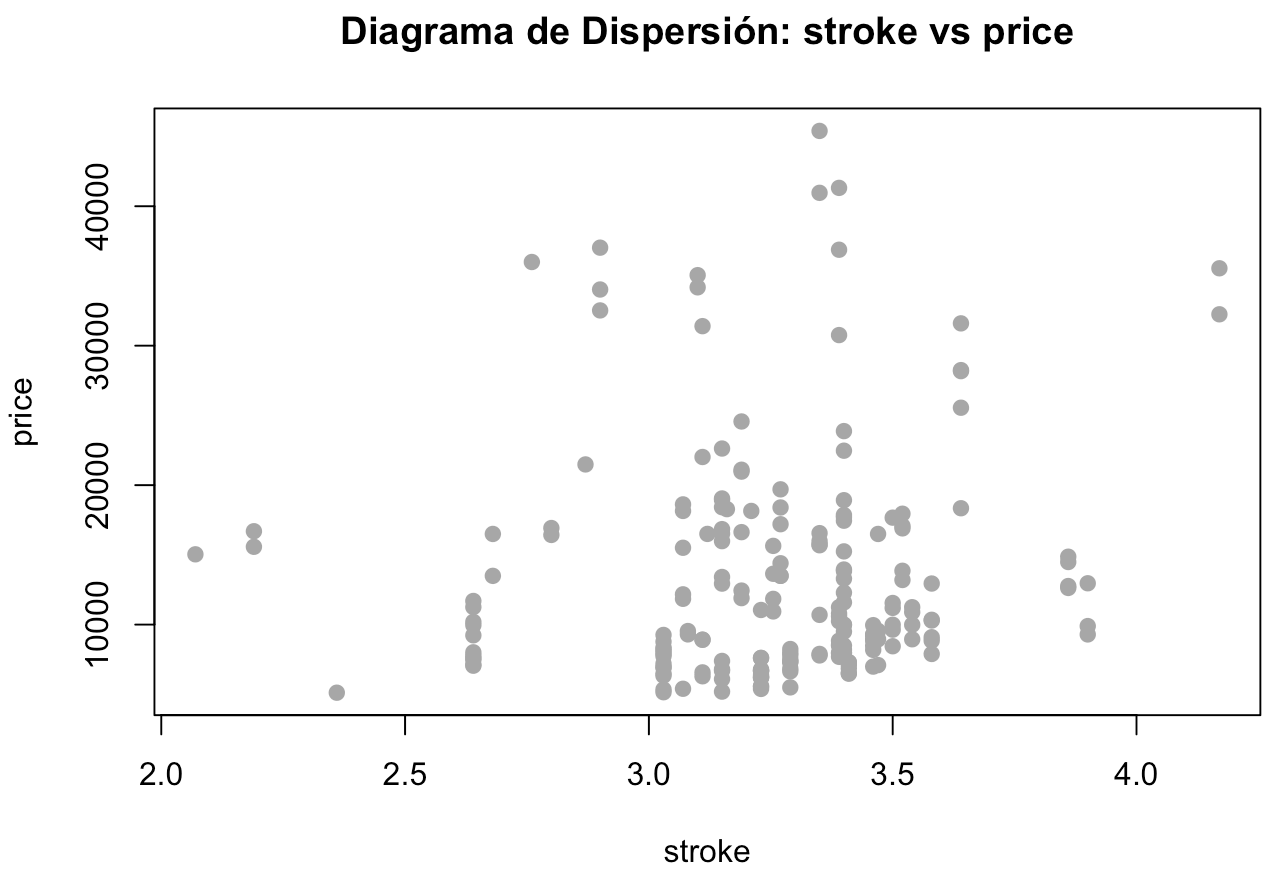

Boxplot de price
40000
30000
20000
10000

```
for (col_name in colnames(df_cuantitativas)) {hist(df_cuantitativas[[col_name]], main =
paste("Histograma de", col_name),xlab = col_name, col = "gray", border = "black")}
```


Histograma de stroke
Frequency
0
20
40
60
2.0
2.5
3.0
3.5
4.0
stroke


Histograma de price
Frequency
0
20
40
60
80
100
10000
20000
30000
40000
50000
price

```
plot(df_cuantitativas$stroke, df_cuantitativas$price, main = "Diagrama de Dispersión: st
roke vs price", xlab = "stroke", ylab = "price", pch = 19, col = "darkgray")
```




```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.2.3
```

```
## Loading required package: ggplot2
```

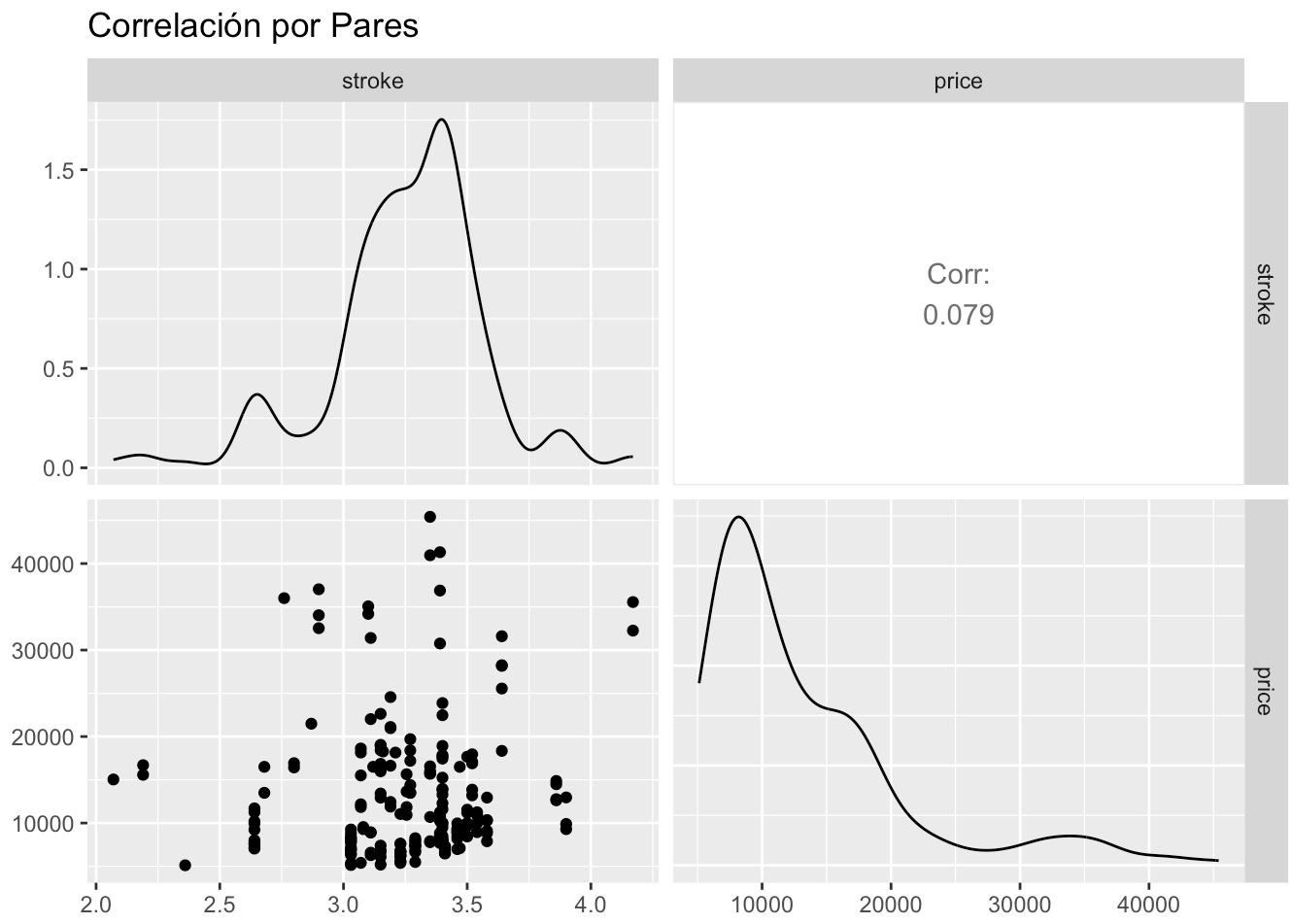
```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

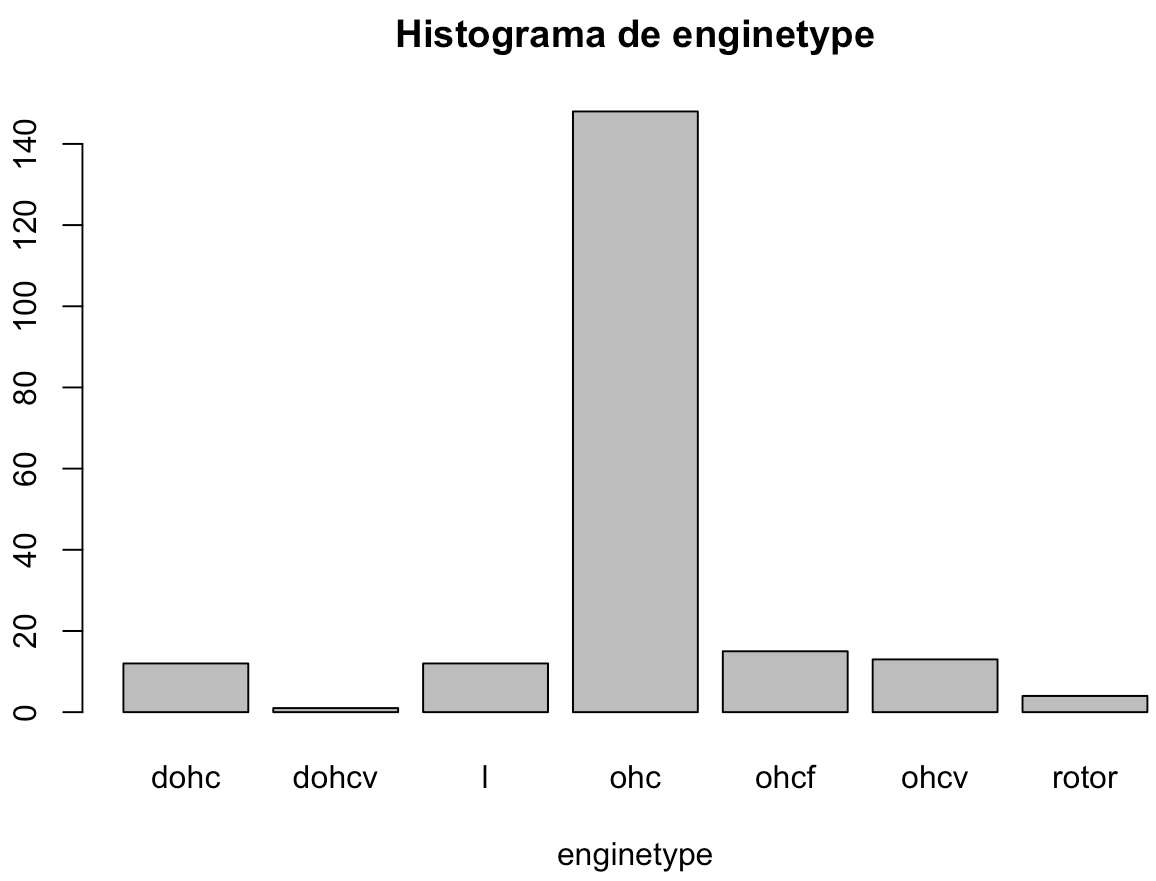
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(df_cuantitativas, title = "Correlación por Pares")
```



- Variables categóricas: Distribución de los datos (diagramas de barras, diagramas de pastel) y Boxplot por categoría de las variables cuantitativas

```
for (col_name in colnames(df_cualitativas)) {barplot(table(df_cualitativas[[col_name]]),
main = paste("Histograma de", col_name),xlab = col_name, col = "gray", border = "blac
k")}
```


Histograma de enginetype
0
20
40
60
80
100
120
140
dohc
dohcv
l
ohc
ohcf
ohcv
rotor
enginetype


Histograma de enginelocation
0
50
100
150
200
front
rear
enginelocation

```
for (col_name in colnames(df_cualitativas)) {pie(table(df_cualitativas[[col_name]]), mai
n = paste("Diagrama de Pastel de", col_name),col = rainbow(length(table(df_cualitativas
[[col_name]])))))}
```

Diagrama de Pastel de enginetype

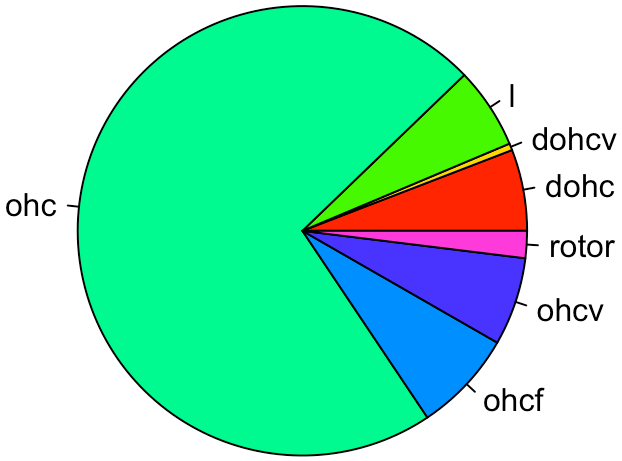
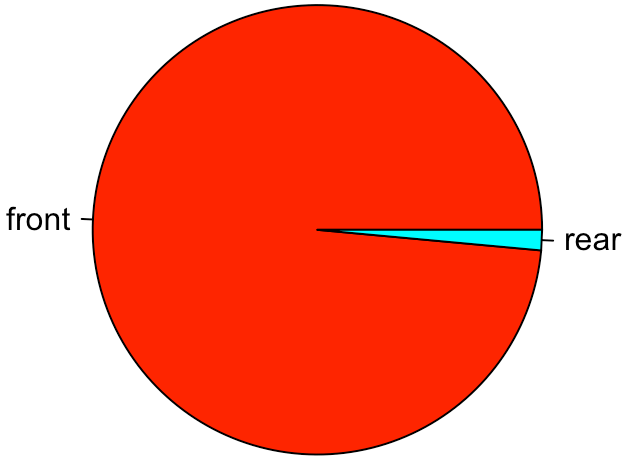
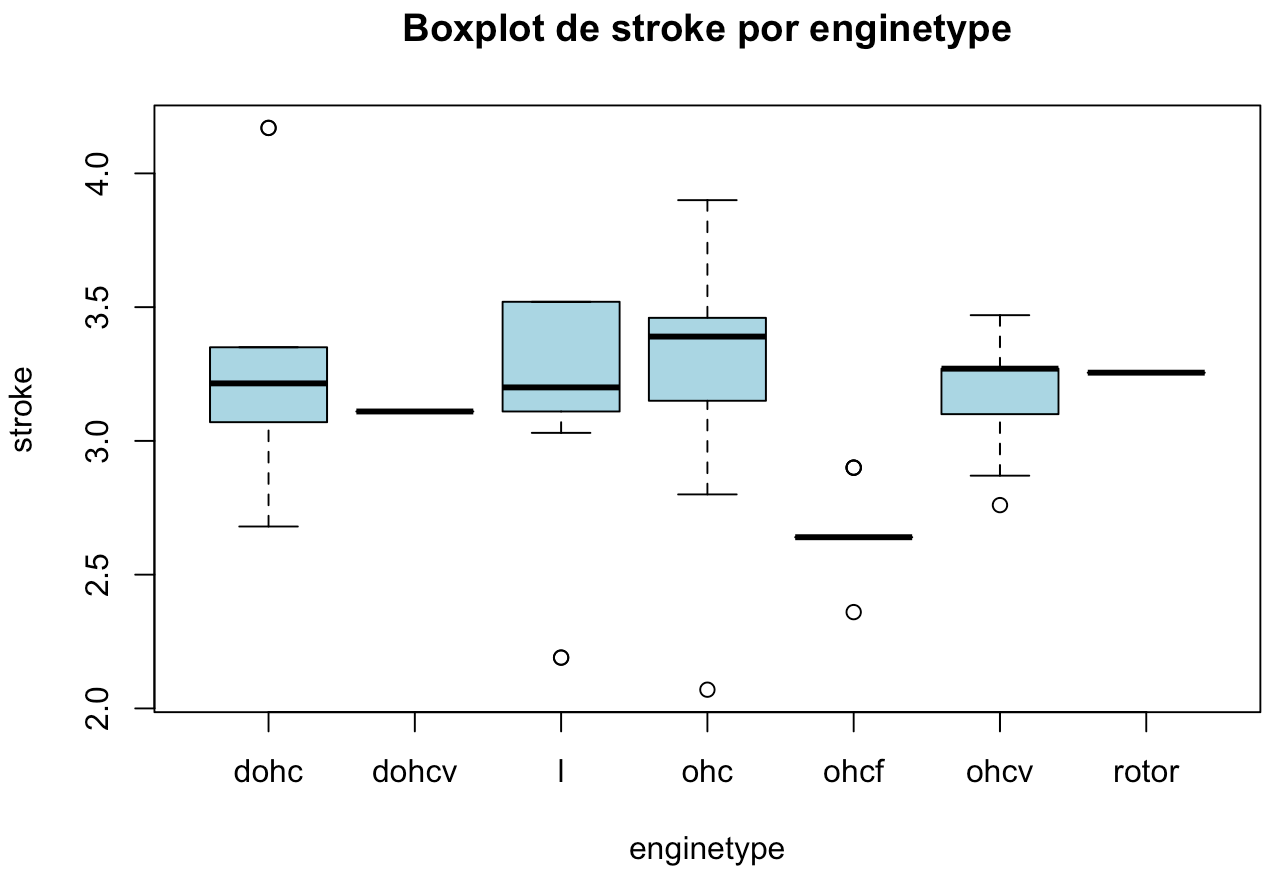
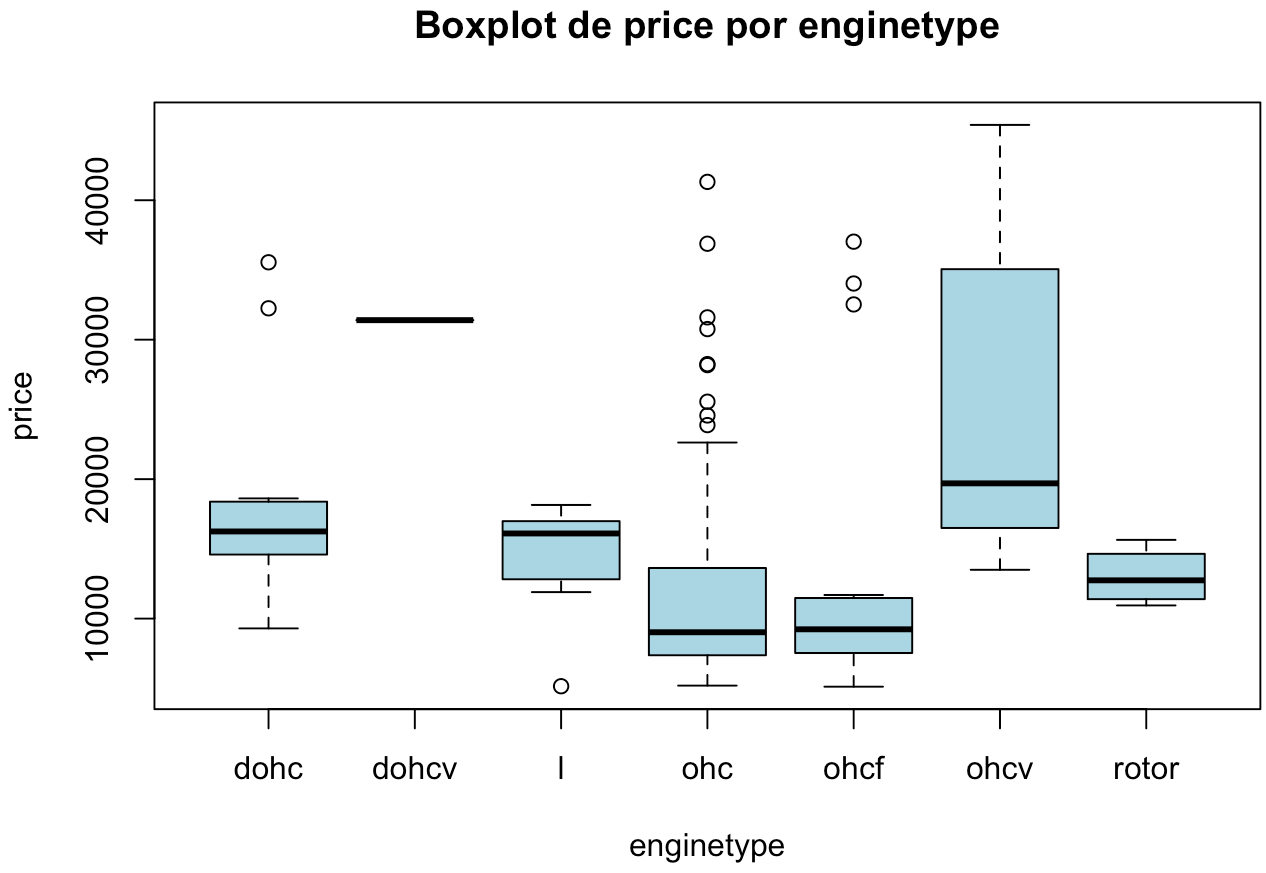


Diagrama de Pastel de enginelocation

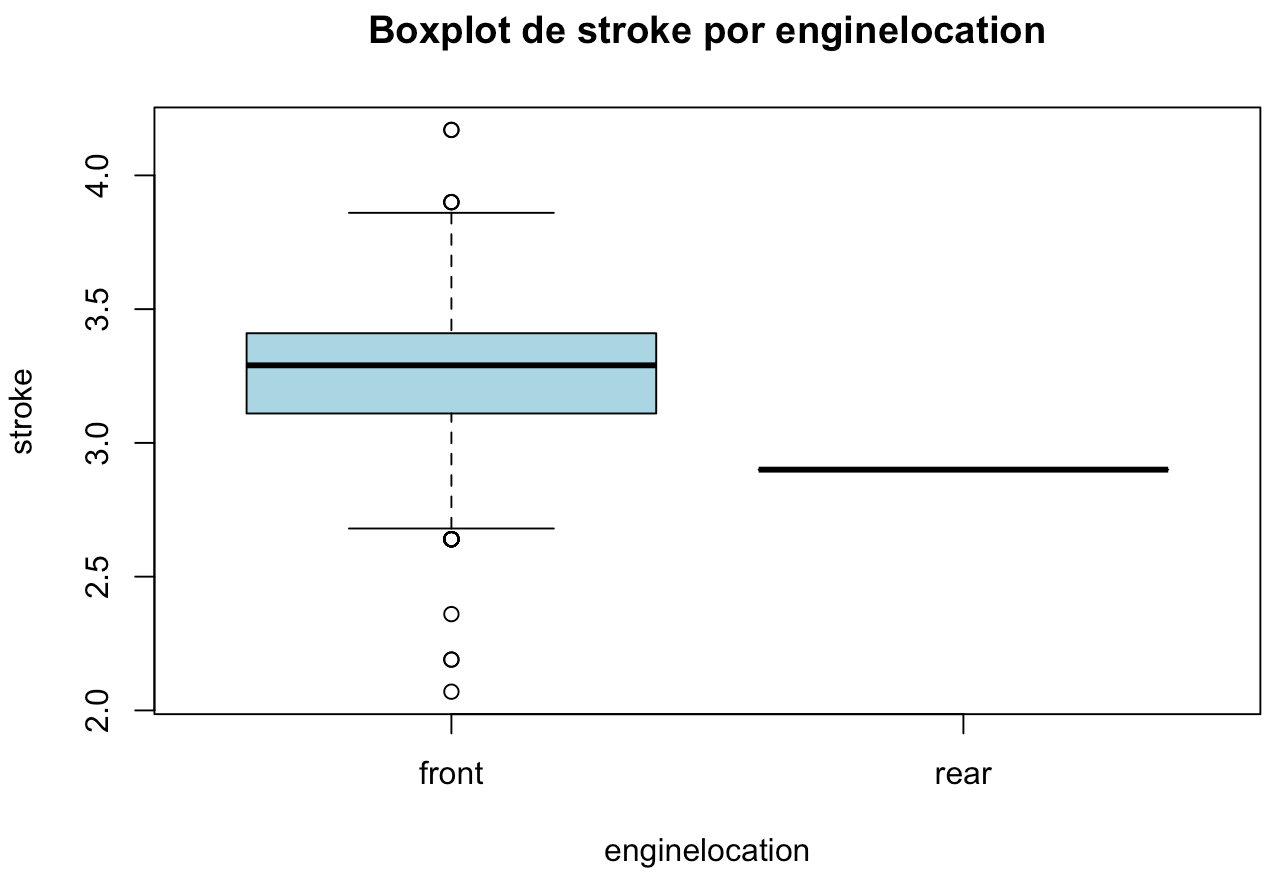


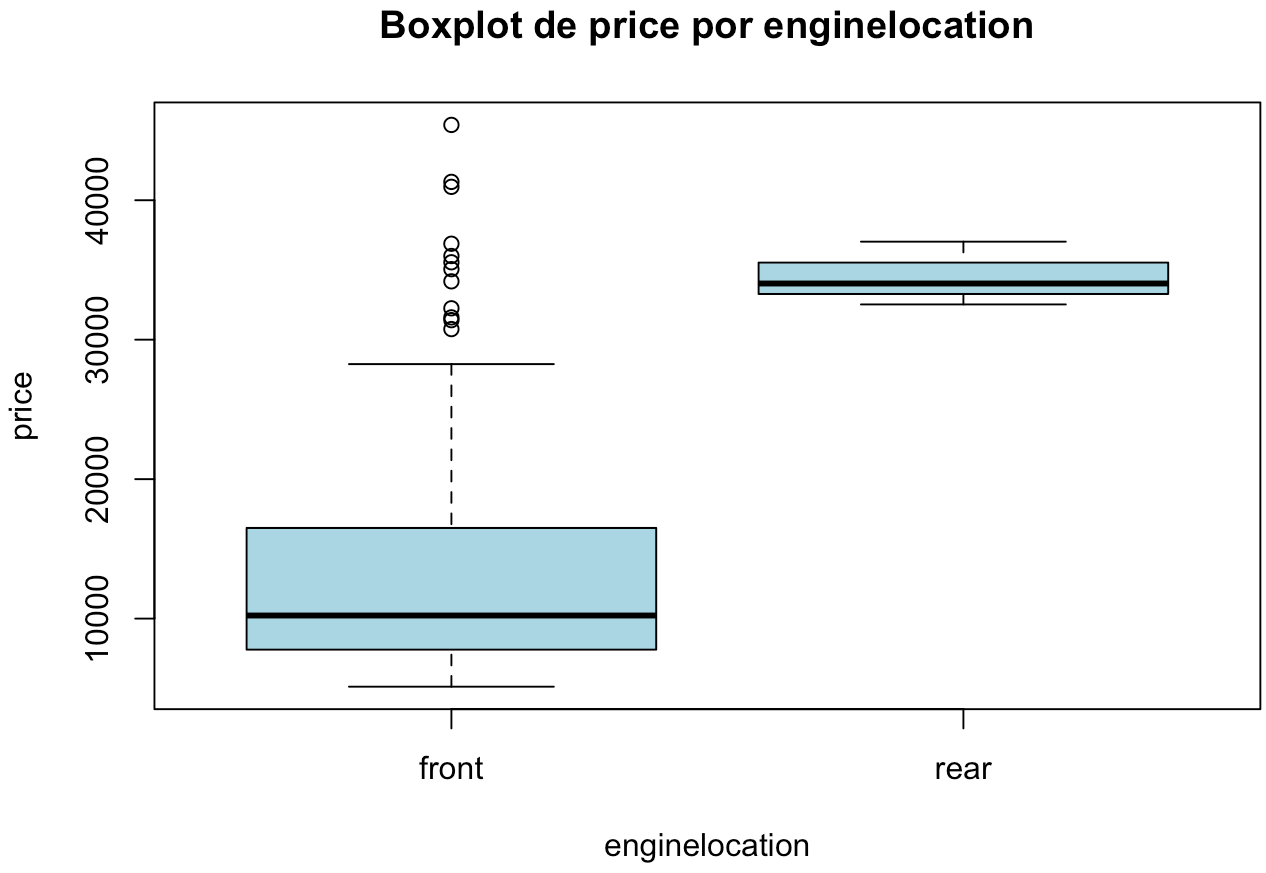
```
for (quant_col in colnames(df_cuantitativas)) {boxplot(df_cuantitativas[[quant_col]] ~ d
f_cualitativas[["enginetype"]], main = paste("Boxplot de", quant_col, "por enginetype"),
xlab = "enginetype", ylab = quant_col, col = "lightblue")}
```


Boxplot de stroke por enginetype
stroke
2.0
2.5
3.0
3.5
4.0
dohc
dohcv
l
ohc
ohcf
ohcv
rotor
enginetype


Boxplot de price por enginetype
price
10000
20000
30000
40000
dohc
dohcv
l
ohc
ohcf
ohcv
rotor
enginetype

```
for (quant_col in colnames(df_cuantitativas)) {boxplot(df_cuantitativas[[quant_col]] ~ d
f_cualitativas[["enginelocation"]],
main = paste("Boxplot de", quant_col, "por enginelocation"),xlab = "enginelocation", yla
b = quant_col, col = "lightblue")}
```


Boxplot de stroke por enginelocation
stroke
2.0
2.5
3.0
3.5
4.0
front
rear
enginelocation


Boxplot de price por enginelocation
price
10000
20000
30000
40000
front
rear
enginelocation

Modelación y verificación del modelo

1. Encuentra la ecuación de regresión de mejor ajuste. Propón al menos 2 modelos de ajuste para encontrar la mejor forma de ajustar la variable precio.

```
modelo1 = lm(price ~ enginetype + enginelocation + stroke, data = df_tercer_grupo)
M1 = summary(modelo1)
M1
```

```
##
## Call:
## lm(formula = price ~ enginetype + enginelocation + stroke, data = df_tercer_grupo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11876  -3888  -2138   1838  29484
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4854       6041   0.803 0.422674
## enginetypedohcv       13984       6777   2.064 0.040380 *
## enginetypel           -2947       2666  -1.105 0.270322
## enginetypeohc         -6717       1954  -3.438 0.000716 ***
## enginetypeohcf        -6882       2900  -2.373 0.018606 *
## enginetypeohcv         7312       2608   2.804 0.005554 **
## enginetyperotor       -4982       3756  -1.327 0.186213
## enginelocationrear    24842       4228   5.876 1.78e-08 ***
## stroke                 4039       1749   2.310 0.021930 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6504 on 196 degrees of freedom
## Multiple R-squared:  0.3631, Adjusted R-squared:  0.3371
## F-statistic: 13.97 on 8 and 196 DF,  p-value: 5.309e-16
```

```
modelo2 = lm(price ~ enginetype * enginelocation * stroke, data = df_tercer_grupo)
M2 = summary(modelo2)
M2
```

```
##
## Call:
## lm(formula = price ~ enginetype * enginelocation * stroke, data = df_tercer_grupo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10699  -3831  -2000   2254  29541
##
## Coefficients: (15 not defined because of singularities)
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              -27487.9    13436.7  -2.046 0.042146
## enginetypedohcv                           15691.6     6697.0   2.343 0.020149
## enginetypel                               41502.6    18524.7   2.240 0.026212
## enginetypeohc                             28580.3    15353.4   1.861 0.064203
## enginetypeohcf                             1129.8    63912.9   0.018 0.985914
## enginetypeohcv                            94693.2    32818.6   2.885 0.004357
## enginetyperotor                           -4702.9     3696.1  -1.272 0.204777
## enginelocationrear                        22207.8     7924.4   2.802 0.005592
## stroke                                    13889.7     4053.5   3.427 0.000748
## enginetypedohcv:enginelocationrear             NA         NA      NA       NA
## enginetypel:enginelocationrear                 NA         NA      NA       NA
## enginetypeohc:enginelocationrear               NA         NA      NA       NA
## enginetypeohcf:enginelocationrear              NA         NA      NA       NA
## enginetypeohcv:enginelocationrear              NA         NA      NA       NA
## enginetyperotor:enginelocationrear             NA         NA      NA       NA
## enginetypedohcv:stroke                         NA         NA      NA       NA
## enginetypel:stroke                        -13695.0     5699.5  -2.403 0.017221
## enginetypeohc:stroke                      -10738.7     4625.3  -2.322 0.021296
## enginetypeohcf:stroke                       -552.3    24210.7  -0.023 0.981823
## enginetypeohcv:stroke                     -27041.7    10177.9  -2.657 0.008550
## enginetyperotor:stroke                         NA         NA      NA       NA
## enginelocationrear:stroke                      NA         NA      NA       NA
## enginetypedohcv:enginelocationrear:stroke      NA         NA      NA       NA
## enginetypel:enginelocationrear:stroke          NA         NA      NA       NA
## enginetypeohc:enginelocationrear:stroke        NA         NA      NA       NA
## enginetypeohcf:enginelocationrear:stroke       NA         NA      NA       NA
## enginetypeohcv:enginelocationrear:stroke       NA         NA      NA       NA
## enginetyperotor:enginelocationrear:stroke      NA         NA      NA       NA
##
## (Intercept)                              *
## enginetypedohcv                          *
## enginetypel                              *
## enginetypeohc                            .
## enginetypeohcf
## enginetypeohcv                           **
## enginetyperotor
## enginelocationrear                       **
## stroke                                   ***
## enginetypedohcv:enginelocationrear
## enginetypel:enginelocationrear
## enginetypeohc:enginelocationrear
## enginetypeohcf:enginelocationrear
```

```
## enginetypeohcv:enginelocationrear
## enginetyperotor:enginelocationrear
## enginetypedohcv:stroke
## enginetypel:stroke                        *
## enginetypeohc:stroke                      *
## enginetypeohcf:stroke
## enginetypeohcv:stroke                     **
## enginetyperotor:stroke
## enginelocationrear:stroke
## enginetypedohcv:enginelocationrear:stroke
## enginetypel:enginelocationrear:stroke
## enginetypeohc:enginelocationrear:stroke
## enginetypeohcf:enginelocationrear:stroke
## enginetypeohcv:enginelocationrear:stroke
## enginetyperotor:enginelocationrear:stroke
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6399 on 192 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3585
## F-statistic:  10.5 on 12 and 192 DF,  p-value: 7.752e-16
```

2. Para cada uno de los modelos propuestos:

-> Realiza la regresión entre las variables involucradas. LISTO, SE HIZO ARRIBA.

-> Analiza la significancia del modelo:

- Valida la significancia del modelo con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera)

```
p_valor_modelo1 = M1$coefficients[1, 4]
p_valor_modelo2 = M2$coefficients[1, 4]

if (p_valor_modelo1 < 0.04) {
    cat("Se rechaza la hipótesis nula (H0) para el modelo 1. Hay evidencia suficiente pa
ra afirmar que existe una relación lineal significativa.\n")
} else {
    cat("No se rechaza la hipótesis nula (H0) para el modelo 1. No hay evidencia suficie
nte para afirmar que existe una relación lineal significativa.\n")
}
```

```
## No se rechaza la hipótesis nula (H0) para el modelo 1. No hay evidencia suficiente pa
ra afirmar que existe una relación lineal significativa.
```

```
if (p_valor_modelo2 < 0.04) {
    cat("Se rechaza la hipótesis nula (H0) para el modelo 2. Hay evidencia suficiente pa
ra afirmar que existe una relación lineal significativa.\n")
} else {
    cat("No se rechaza la hipótesis nula (H0) para el modelo 2. No hay evidencia suficie
nte para afirmar que existe una relación lineal significativa.\n")
}
```

```
## No se rechaza la hipótesis nula (H0) para el modelo 2. No hay evidencia suficiente pa
ra afirmar que existe una relación lineal significativa.
```

- Valida la significancia de β̂i con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera de cada una de ellas)

```
for (i in 1:length(p_valor_modelo1)) {
  if (p_valor_modelo1[i] < 0.04) {
    cat("Se rechaza la hipótesis nula (H0) para el coeficiente β̂", i, " en el modelo 1.
Hay evidencia suficiente para afirmar que el coeficiente es significativo.\n")
  } else {
    cat("No se rechaza la hipótesis nula (H0) para el coeficiente β̂", i, " en el modelo
1. No hay evidencia suficiente para afirmar que el coeficiente es significativo.\n")
  }
}
```

```
## No se rechaza la hipótesis nula (H0) para el coeficiente β̂ 1  en el modelo 1. No hay
evidencia suficiente para afirmar que el coeficiente es significativo.
```

```
for (i in 1:length(p_valor_modelo2)) {
  if (p_valor_modelo2[i] < 0.04) {
    cat("Se rechaza la hipótesis nula (H0) para el coeficiente β̂", i, " en el modelo 2.
Hay evidencia suficiente para afirmar que el coeficiente es significativo.\n")
  } else {
    cat("No se rechaza la hipótesis nula (H0) para el coeficiente β̂", i, " en el modelo
2. No hay evidencia suficiente para afirmar que el coeficiente es significativo.\n")
  }
}
```

```
## No se rechaza la hipótesis nula (H0) para el coeficiente β̂ 1  en el modelo 2. No hay
evidencia suficiente para afirmar que el coeficiente es significativo.
```

- Indica cuál es el porcentaje de variación explicada por el modelo.

```
R2_modelo1 <- M1$r.squared
R2_modelo2 <- M2$r.squared

cat("El porcentaje de variación explicada por el modelo 1 es: ", R2_modelo1 * 100, "%
\n")
```

```
## El porcentaje de variación explicada por el modelo 1 es:  36.30861 %
```

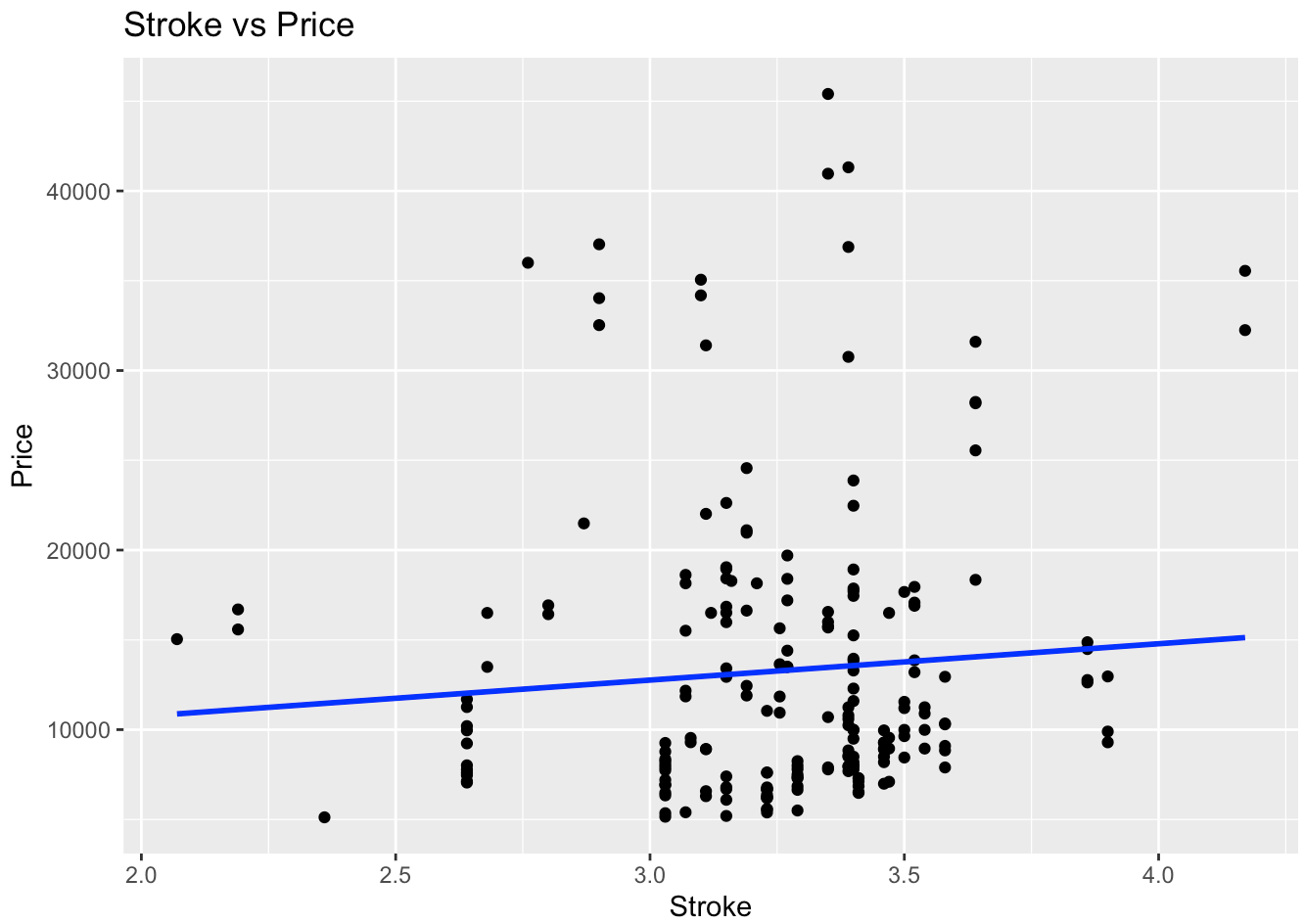
```
cat("El porcentaje de variación explicada por el modelo 2 es: ", R2_modelo2 * 100, "%
\n")
```

```
## El porcentaje de variación explicada por el modelo 2 es:  39.61935 %
```

- Dibuja el diagrama de dispersión de los datos por pares y la recta de mejor ajuste.

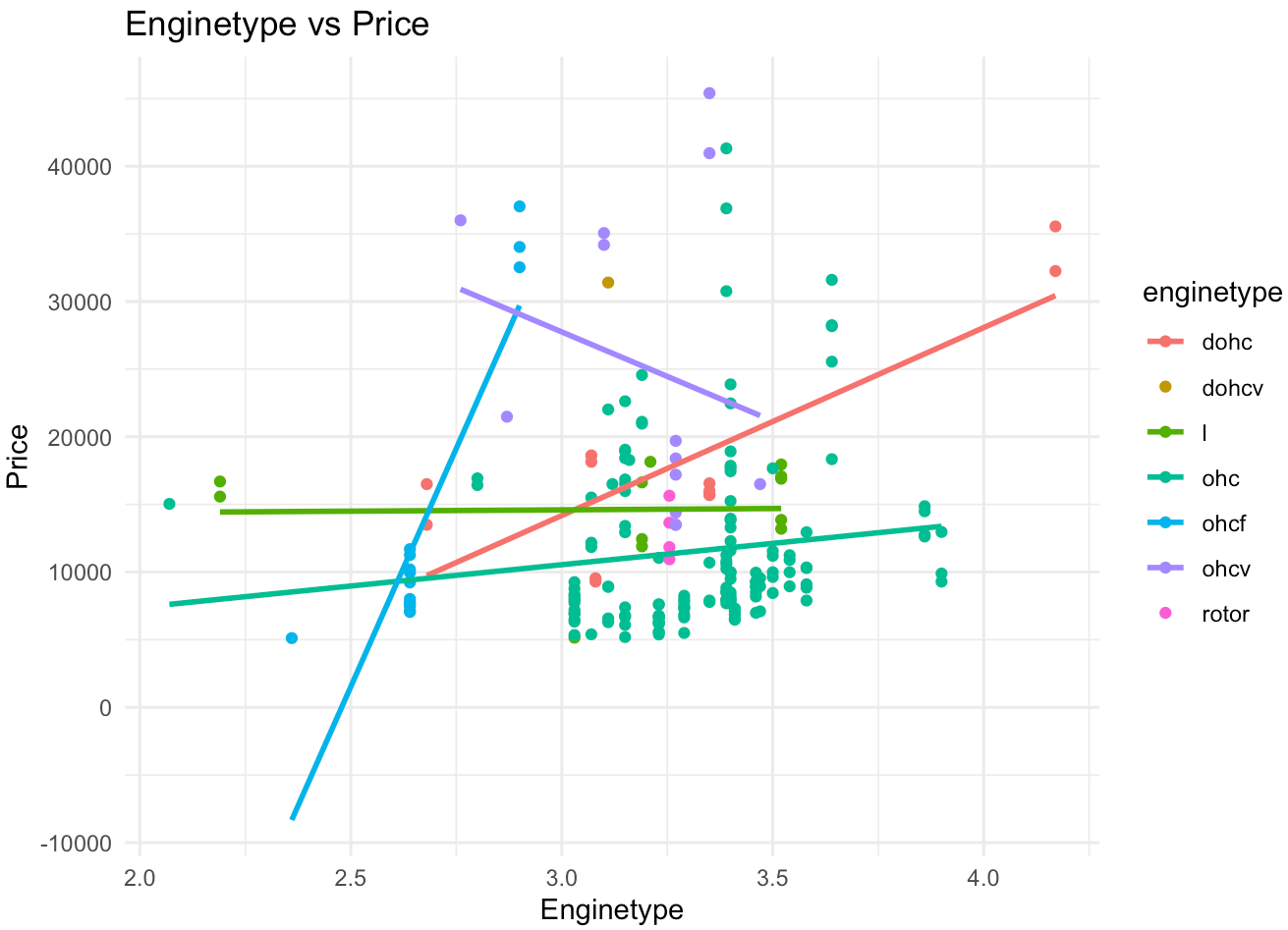
```
library(ggplot2)
library(ggpubr)

p1 = ggplot(df_tercer_grupo, aes(x = stroke, y = price)) + geom_point() + geom_smooth(me
thod = "lm", se = FALSE, color = "blue") + labs(title = "Stroke vs Price", x = "Stroke",
y = "Price")
ggarrange(p1, ncol = 1, nrow = 1)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



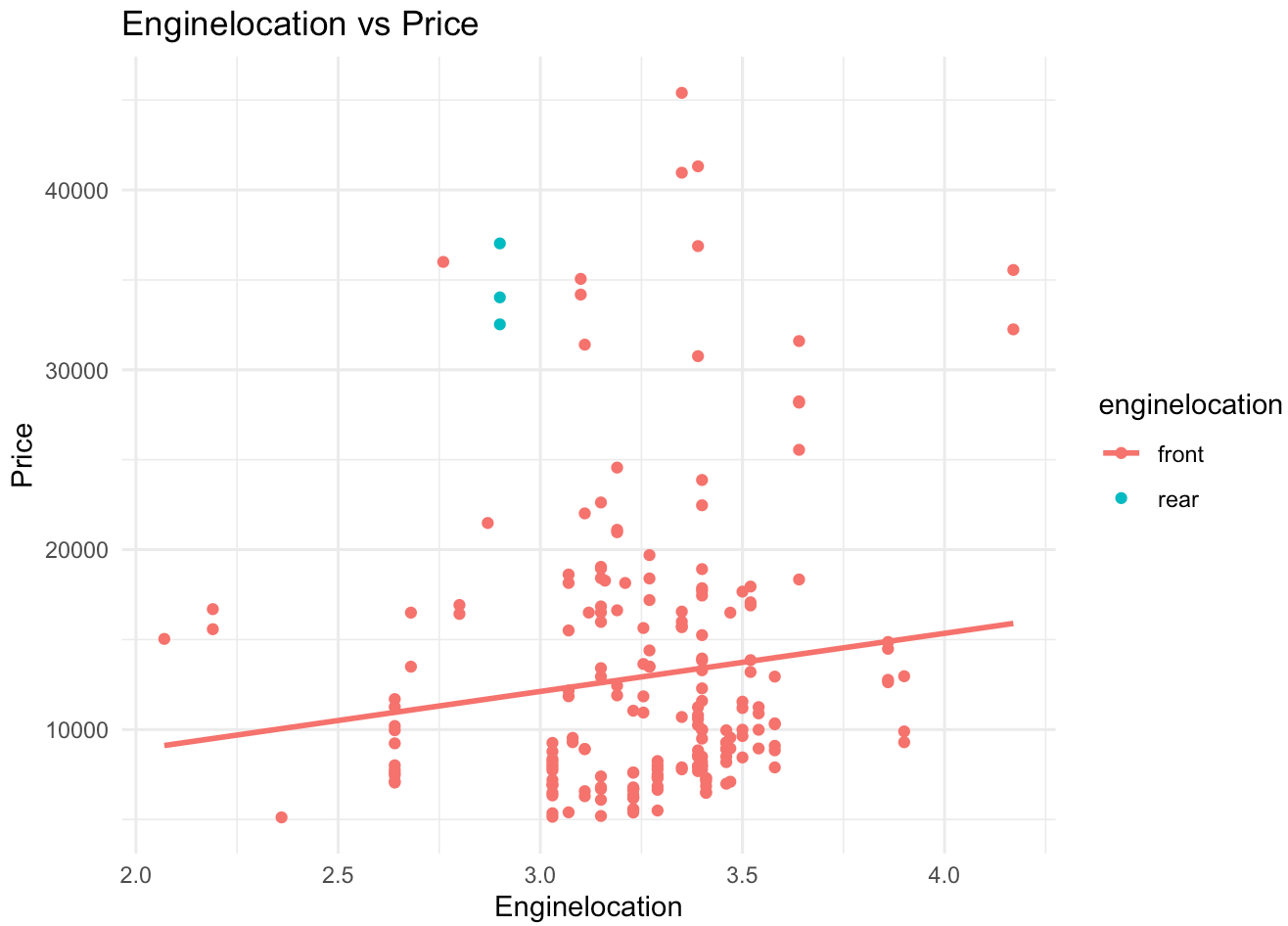
```
p2 = ggplot(df_tercer_grupo, aes(x = stroke, y = price, color = enginetype)) + geom_poin
t() + geom_smooth(method = "lm", se = FALSE) + labs(title = "Enginetype vs Price", x =
"Enginetype", y = "Price") + theme_minimal()
print(p2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
p3 = ggplot(df_tercer_grupo, aes(x = stroke, y = price, color = enginelocation)) + geom_
point() + geom_smooth(method = "lm", se = FALSE) + labs(title = "Enginelocation vs Pric
e", x = "Enginelocation", y = "Price") + theme_minimal()
print(p3)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



3. Analiza la validez de los modelos propuestos:

-> Normalidad de los residuos

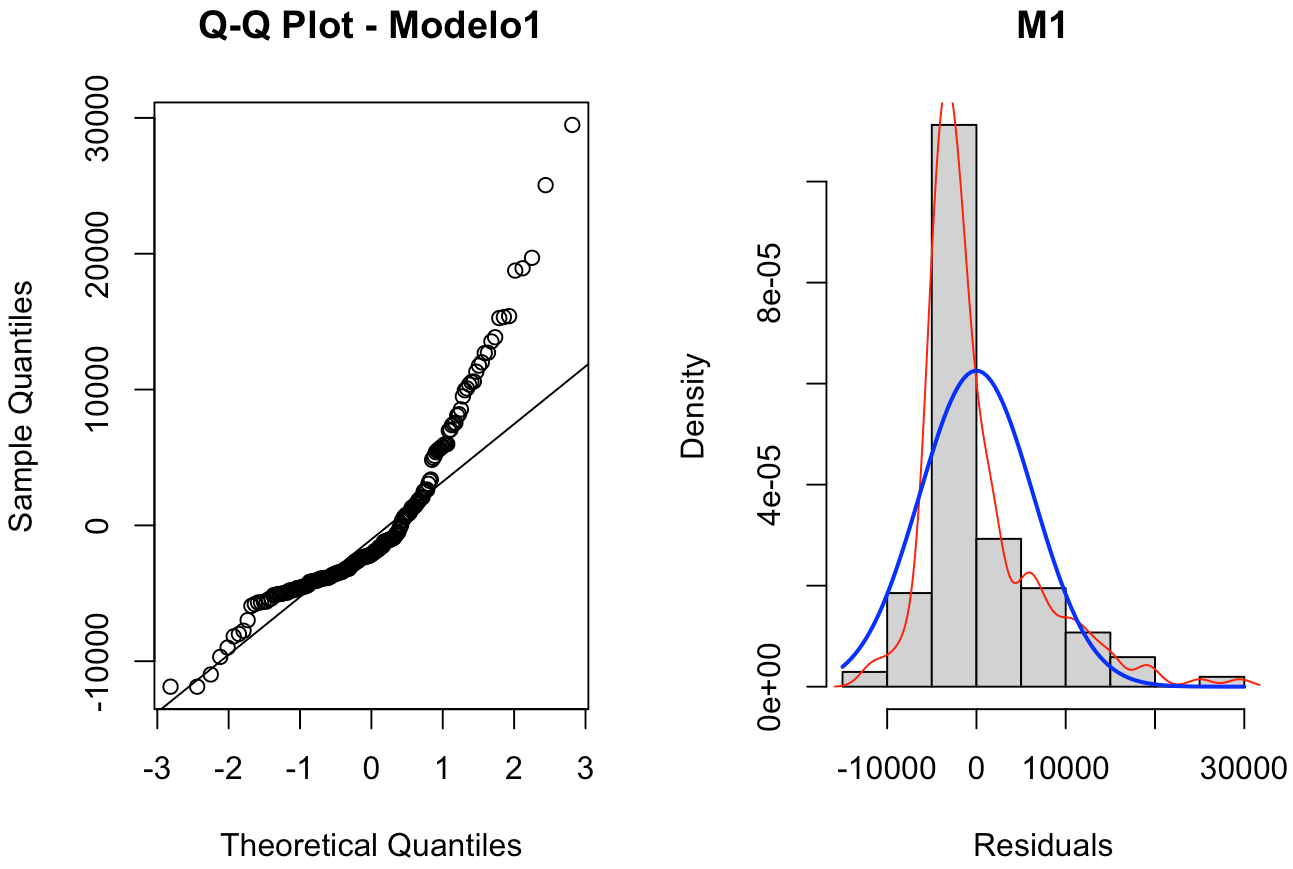
```
library(nortest)
ad.test(M1$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  M1$residuals
## A = 10.37, p-value < 2.2e-16
```

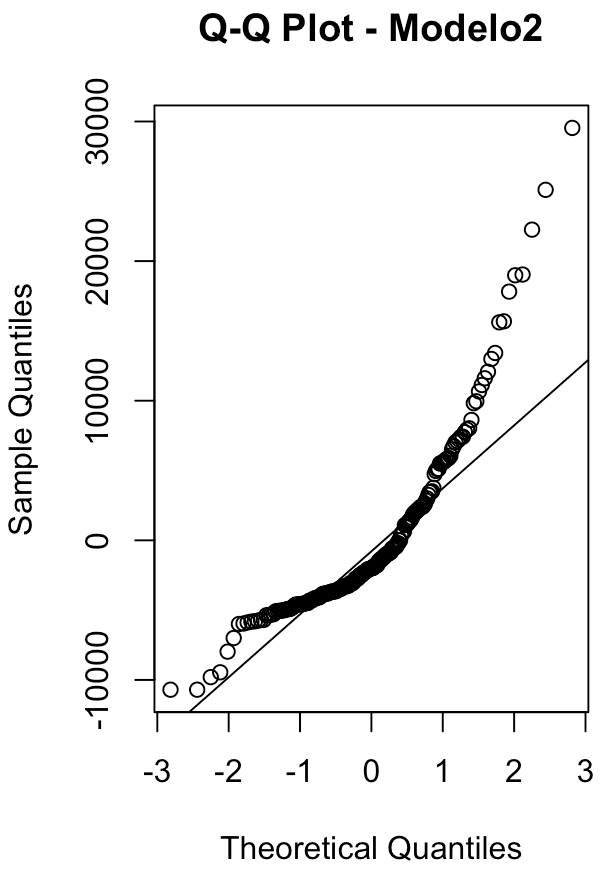
```
ad.test(M2$residuals)
```

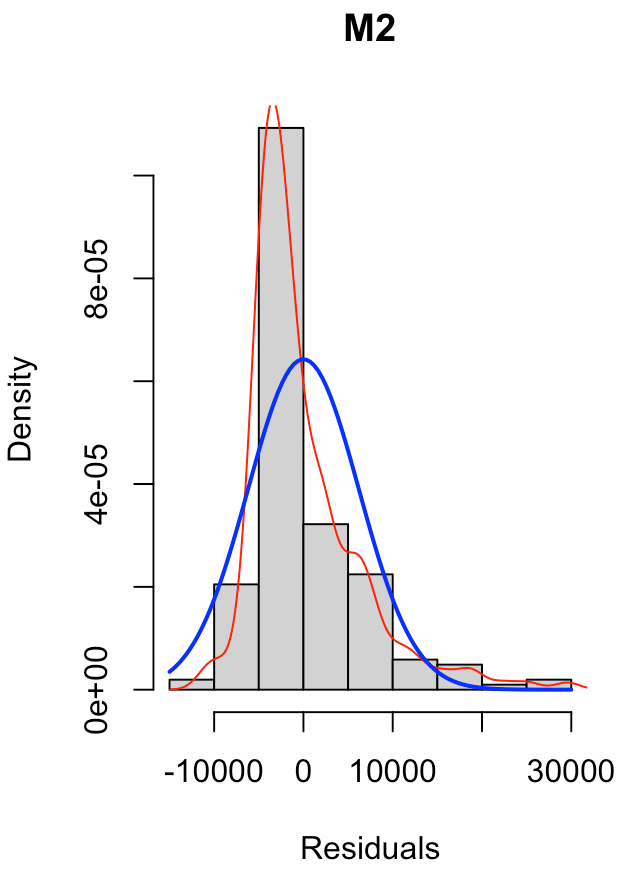
```
##
##  Anderson-Darling normality test
##
## data:  M2$residuals
## A = 9.6978, p-value < 2.2e-16
```

```
par(mfrow = c(1,2))
qqnorm(M1$residuals, main = "Q-Q Plot - Modelo1")
qqline(M1$residuals)
hist(M1$residuals, freq = FALSE, main = "M1", xlab = "Residuals", ylab = "Density")
lines(density(M1$residuals), col = "red")
curve(dnorm(x, mean = mean(M1$residuals), sd=sd(M1$residuals)), add = TRUE, col = "blu
e", lwd = 2)
```




```
par(mfrow = c(1,1))

par(mfrow = c(1,2))
qqnorm(M2$residuals, main = "Q-Q Plot - Modelo2")
qqline(M2$residuals)
hist(M2$residuals, freq = FALSE, main = "M2", xlab = "Residuals", ylab = "Density")
lines(density(M2$residuals), col = "red")
curve(dnorm(x, mean = mean(M2$residuals), sd=sd(M2$residuals)), add = TRUE, col = "blu
e", lwd = 2)
```






```
par(mfrow = c(1,2))
```

Se muestran los QQplots de cada modelo donde medimos y comparamos como se comportan los datos, comparandolos con los de una distribución normal, el desvío de datos muestra la diferencia que se tiene versus una normal, tamboén se puede hacer otro tipo de comparaciones con otrs distribuciones pero la normal es la más usada.

De esta forma podemos verificar la linealidad de los datos, y pordemos confirmar que no se comportan completamente como una normal.

-> Verificación de media cero

```
t.test(M1$residuals)
```

```
##
##  One Sample t-test
##
## data:  M1$residuals
## t = 2.7897e-15, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -877.9708  877.9708
## sample estimates:
##    mean of x
## 1.242237e-12
```

- El valor estadístico t: 2.7897e-15, es increíblemente pequeño, muy cercano a 0. Indica diferencia entre la media de los residuales y 0. (La diferencia en este caso en casi nula)
- Df: Se refiere a los grados de libertad: # Datos - # Parámetros
- P-Value: 1, es demasiado alto indica lo que nos dice que no hay evidencia suficiente para rechazar la hipótesis nula de que la media de los residuales es igual a 0. (Por cierto esta es la H0)
- 95 percent confidence interval: Estamos 95% seguros de que la media de los residuales está dentro de este (-877.9708 877.9708) rango.
- Media de los residuales: Es prácticamente 0.

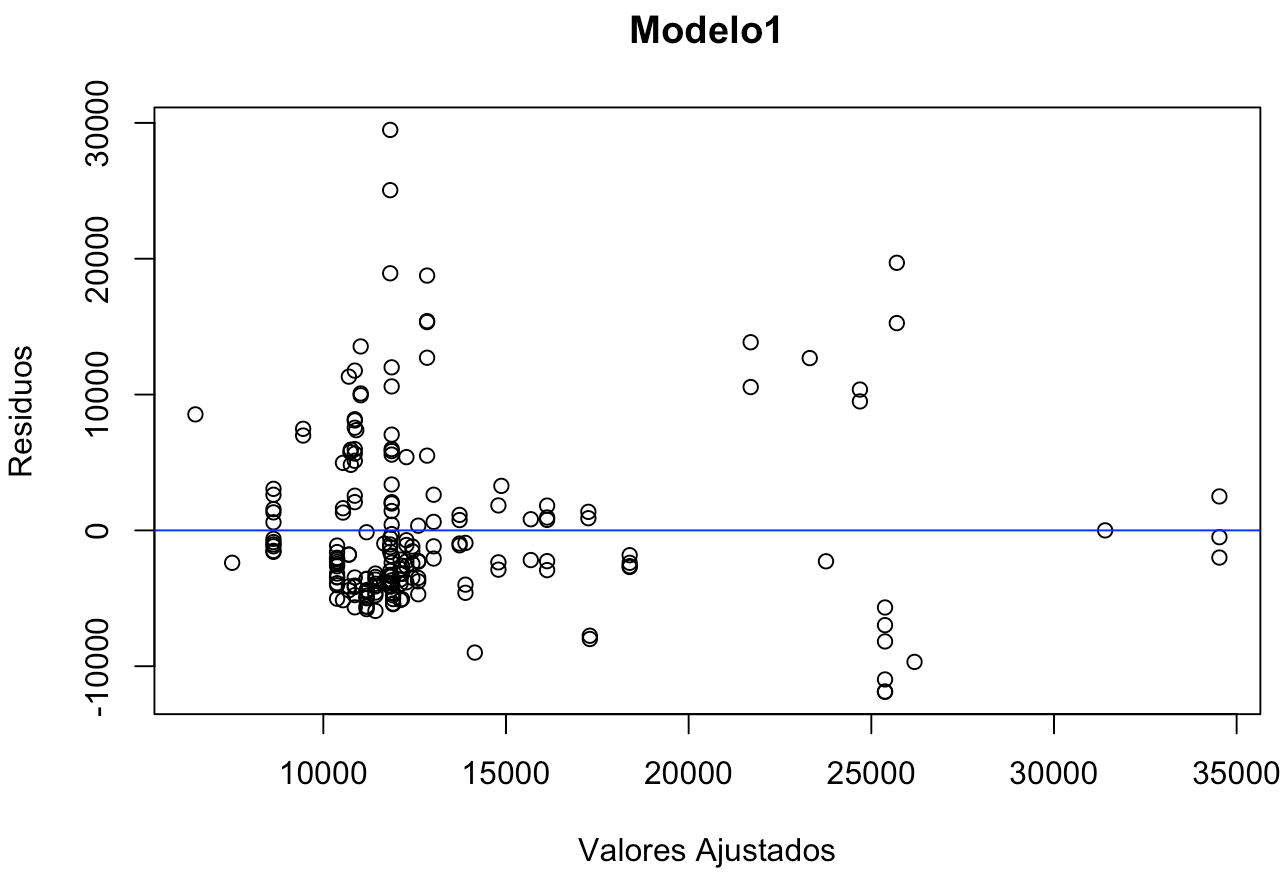
```
t.test(M2$residuals)
```

```
##
##  One Sample t-test
##
## data:  M2$residuals
## t = 9.8234e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -854.8474  854.8474
## sample estimates:
##    mean of x
## 4.259097e-13
```

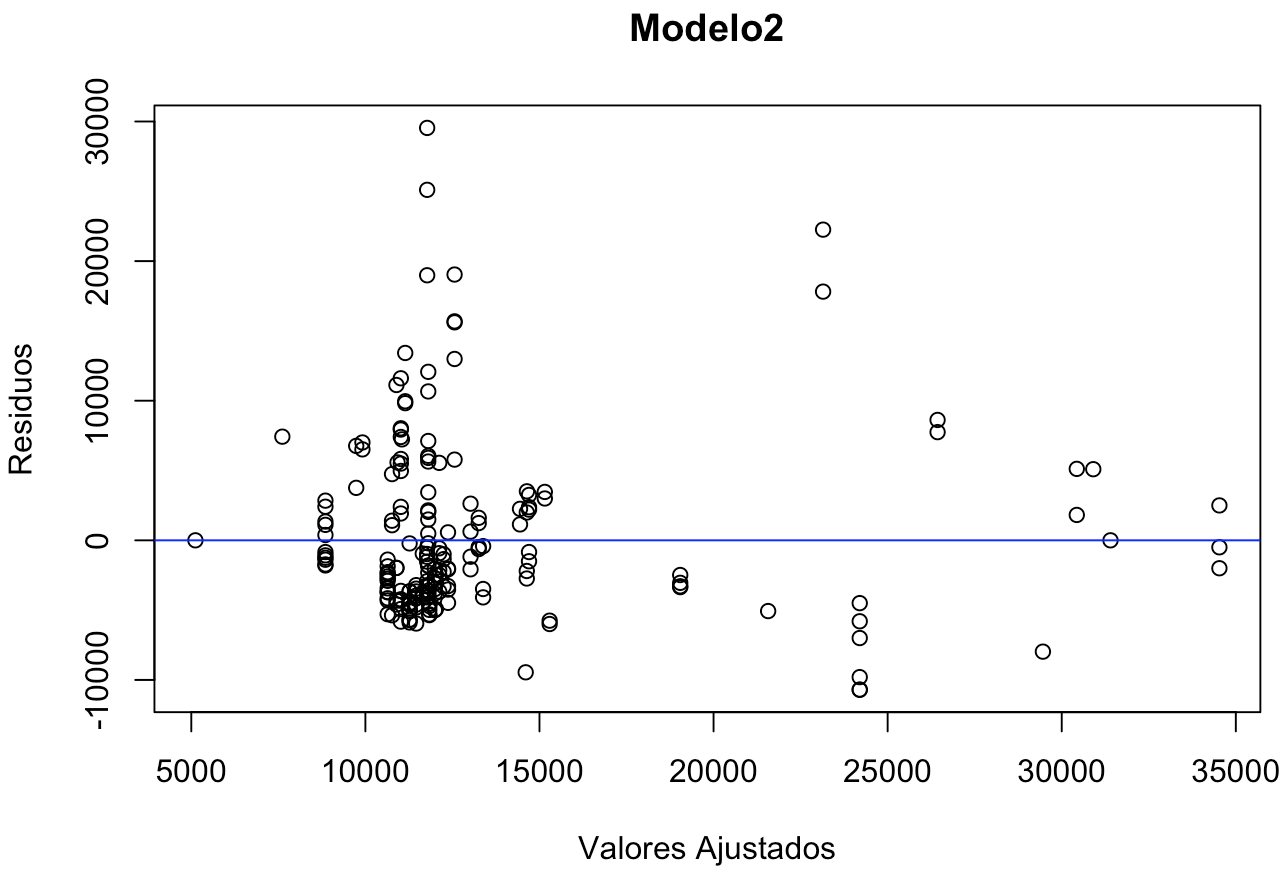
- El valor estadístico t: 9.8234e-16, es increíblemente pequeño, muy cercano a 0. Indica diferencia entre la media de los residuales y 0. (La diferencia en este caso en casi nula)
- Df: Se refiere a los grados de libertad: # Datos - # Parámetros
- P-Value: 1, es demasiado alto indica lo que nos dice que no hay evidencia suficiente para rechazar la hipótesis nula de que la media de los residuales es igual a 0. (Por cierto esta es la H0)
- 95 percent confidence interval: Estamos 95% seguros de que la media de los residuales está dentro de este (-854.8474 854.8474) rango.
- Media de los residuales: Es prácticamente 0.

-> Homocedasticidad, linealidad e independencia

```
plot(modelo1$fitted.values, modelo1$residuals, main = "Modelo1", xlab = "Valores Ajustad
os", ylab = "Residuos")
abline(h = 0, col = "blue")
```




```
plot(modelo2$fitted.values, modelo2$residuals, main = "Modelo2", xlab = "Valores Ajustad
os", ylab = "Residuos")
abline(h = 0, col = "blue")
```



-> Interpreta cada uno de los analisis que realizaste

Modelo 1: Este modelo es una regresión lineal cada variable que está en este modelo afecta de manera independiente y lineal, NO se tiene intereacción entre ellas, lo que significa que cada variable tiene su coeficiente que describe el efecto que se tiene sobre la variable predictora “price”.

Por otro lado el Modelo 2, cada variable tiene interacciones entre ellas, lo que significa que además de tener en cuenta el efecto directo de cada variable en el precio, también se tiene el efecto de cada como cada una variables depende de los valores de las otras.

Para ambos modelos podemos decir que los datos se ajustan bien en términos de los residuales.

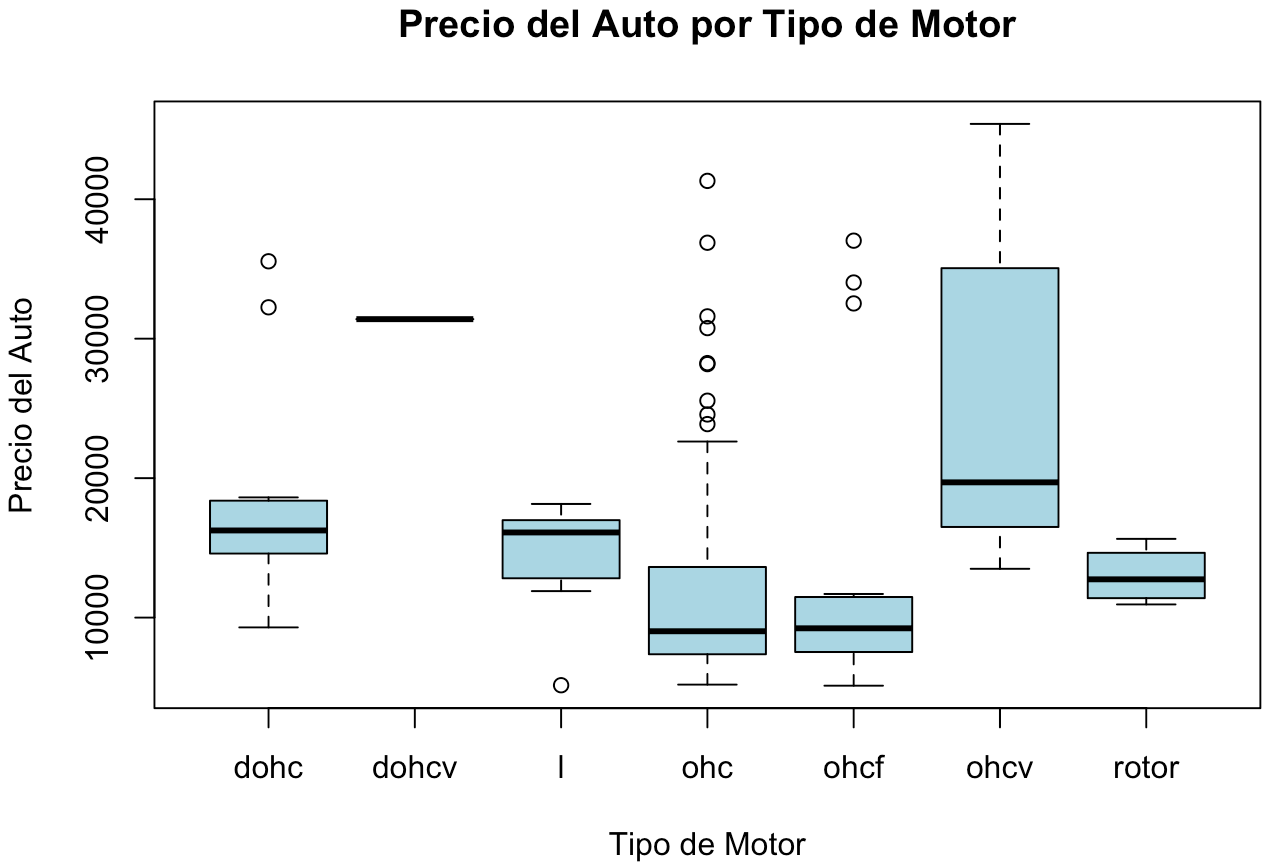
4. Emite una conclusión final sobre el mejor modelo de regresión lineal y contesta la pregunta central:

-> Concluye sobre el mejor modelo que encontraste y argumenta por qué es el mejor.

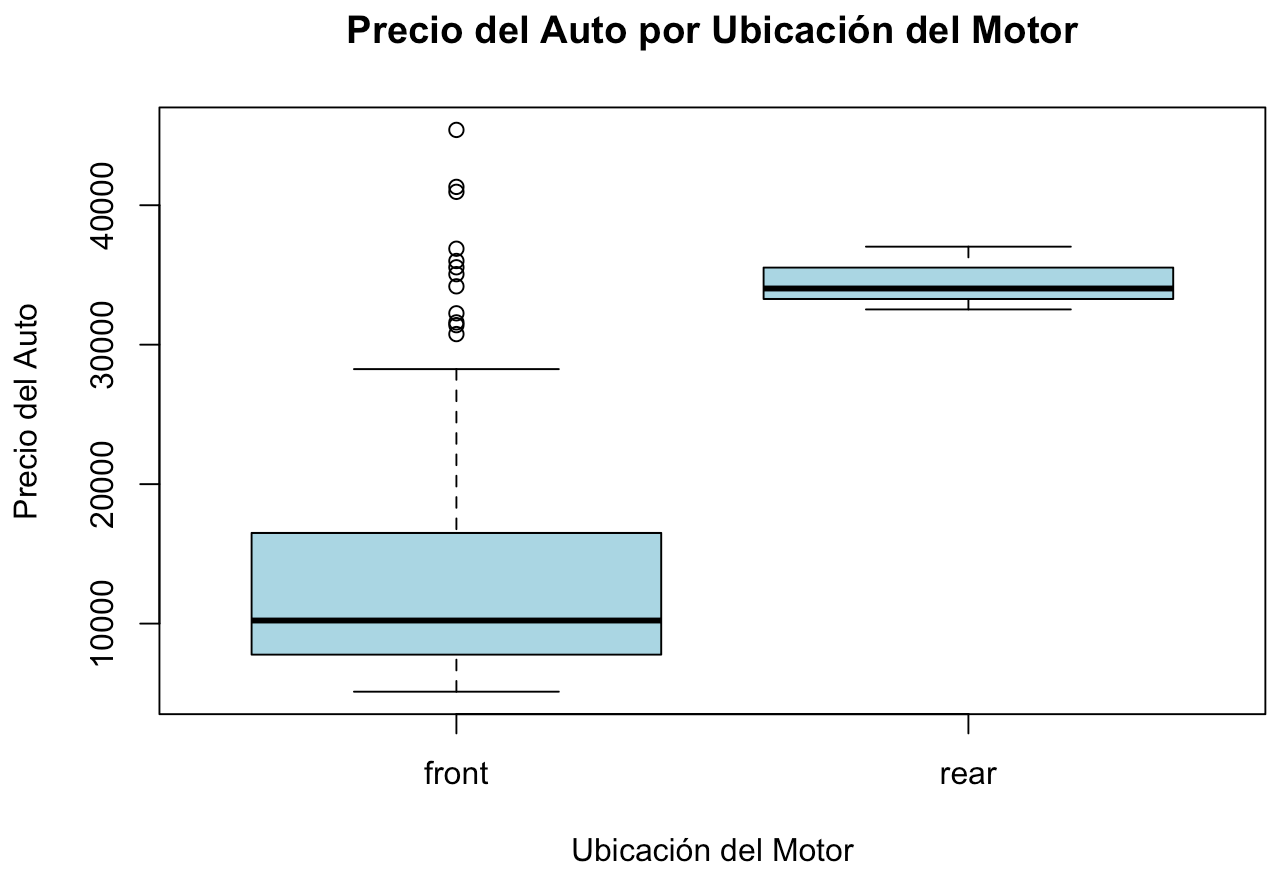
- Creo que el mejor modelo es el Modelo 1 debido a que es más simple además en la comparación de sus R2s no es tán grande la diferencia como se espera de un modelo más complejo y completo como el Modelo 2, para el Modelo 2 hay muchas interacciones que no son significativas, o que de plano en el summary se muestran como NA. Comparando los Errores Estándar se muestra lo mismo, el Modelo 2 tiene un ES ligeramente menor pero la complejidad del modelo 2 no se compara con la simplicidad del modelo 1, por estás razones considero que el modelo 1 es mejor.

-> ¿Cuáles de las variables asignadas influyen en el precio del auto? ¿de qué manera lo hacen?

```
boxplot(price ~ enginetype, data = df_tercer_grupo, main = "Precio del Auto por Tipo de
Motor", ylab = "Precio del Auto", xlab = "Tipo de Motor", col = "lightblue")
```




```
boxplot(price ~ enginelocation, data = df_tercer_grupo, main = "Precio del Auto por Ubic
ación del Motor", ylab = "Precio del Auto", xlab = "Ubicación del Motor", col = "lightbl
ue")
```




```
M1 = summary(modelo1)
M1
```

```
##
## Call:
## lm(formula = price ~ enginetype + enginelocation + stroke, data = df_tercer_grupo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11876  -3888  -2138   1838  29484
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4854       6041   0.803 0.422674
## enginetypedohcv       13984       6777   2.064 0.040380 *
## enginetypel           -2947       2666  -1.105 0.270322
## enginetypeohc         -6717       1954  -3.438 0.000716 ***
## enginetypeohcf        -6882       2900  -2.373 0.018606 *
## enginetypeohcv         7312       2608   2.804 0.005554 **
## enginetyperotor       -4982       3756  -1.327 0.186213
## enginelocationrear    24842       4228   5.876 1.78e-08 ***
## stroke                 4039       1749   2.310 0.021930 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6504 on 196 degrees of freedom
## Multiple R-squared:  0.3631, Adjusted R-squared:  0.3371
## F-statistic: 13.97 on 8 and 196 DF,  p-value: 5.309e-16
```

Tomando en cuenta lo anterior las variables que son estadísticamente significativas en el modelo, son las que tienen un valor p menor a 0.05, lo que influencia en el precio del auto.

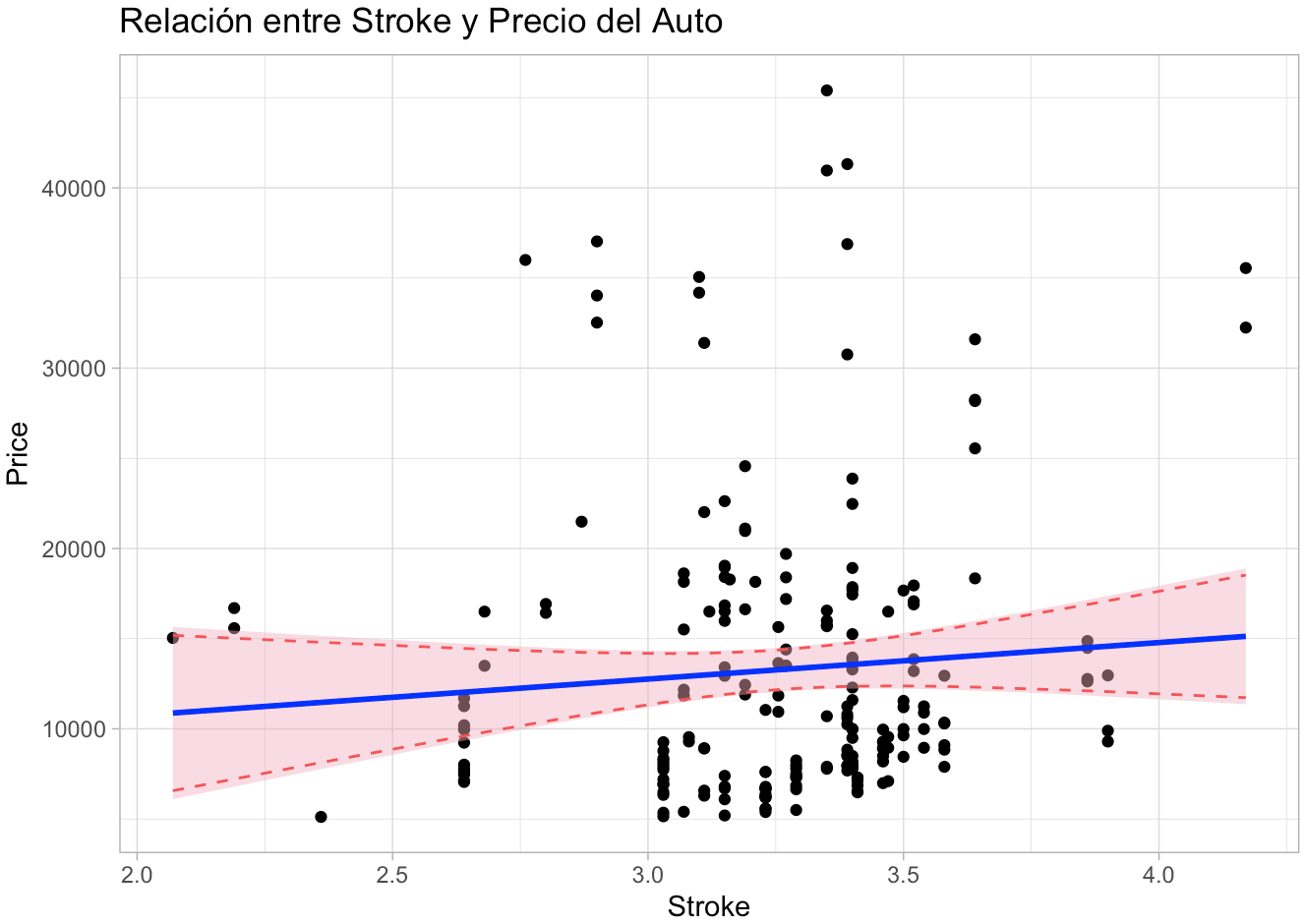
- enginetypedohcv
- enginetypeohc
- enginetypeohcf
- enginetypeohcv
- enginelocationrear
- stroke

III. Intervalos de predicción y confianza

1. Con los datos de las variables asignadas construye la gráfica de los intervalos de confianza y predicción para la estimación y predicción del precio para el mejor modelo seleccionado:

-> Calcula los intervalos para la variable Y

```
modelo_stroke <- lm(price ~ stroke, data = df_tercer_grupo)
data_numerica = data.frame(stroke = df_tercer_grupo$stroke)
predicciones = predict(modelo_stroke, newdata = data_numerica, interval = "confidence")

df_tercer_grupo$fit = predicciones[, "fit"]
df_tercer_grupo$lwr = predicciones[, "lwr"]
df_tercer_grupo$upr = predicciones[, "upr"]

p1 = ggplot(df_tercer_grupo, aes(x = stroke, y = price)) + geom_point() + geom_line(aes
(y = lwr), color = "red", linetype = "dashed") +  geom_line(aes(y = upr), color = "red",
linetype = "dashed") + geom_smooth(method = "lm", formula = y ~ x, se = TRUE, level = 0.
97, col = "blue", fill = "pink2") + theme_light() + labs(title = "Relación entre Stroke
y Precio del Auto", x = "Stroke", y = "Price")
ggarrange(p1, ncol = 1, nrow = 1)
```



-> Selecciona la categoría de la variable cualitativa que, de acuerdo a tu análisis resulte la más importante, y separa la base de datos por esa variable categórica.

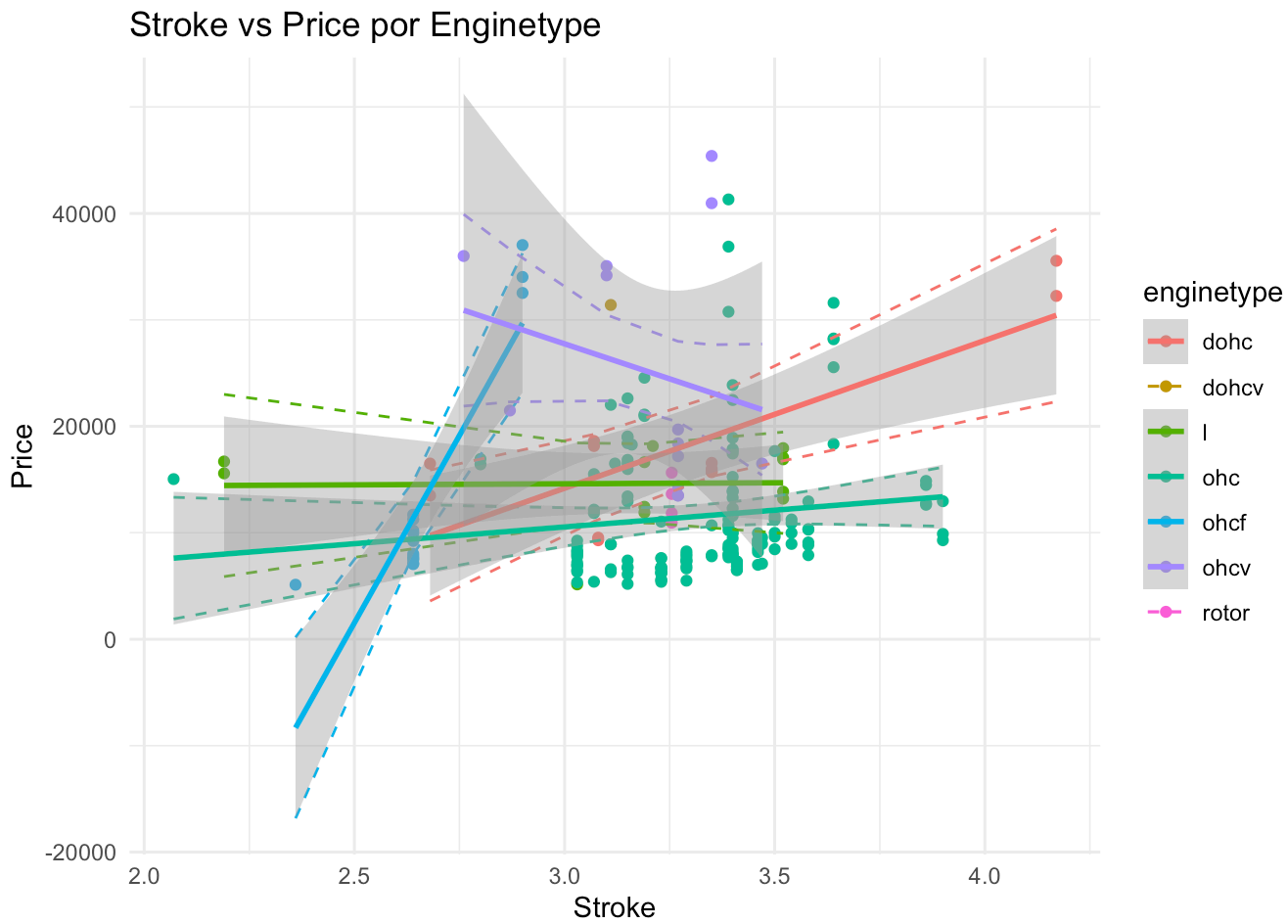
```
df_motor_trasero = subset(df_tercer_grupo, enginelocation == "rear")
df_motor_delantero = subset(df_tercer_grupo, enginelocation == "front")
```

-> Grafica por pares de variables numéricas

```
modelo_stroke_enginetype = lm(price ~ stroke * enginetype, data = df_tercer_grupo)
data_categorica = data.frame(stroke = df_tercer_grupo$stroke, enginetype = df_tercer_gru
po$enginetype)
predicciones = predict(modelo_stroke_enginetype, newdata = data_categorica, interval =
"confidence")
```

```
## Warning in predict.lm(modelo_stroke_enginetype, newdata = data_categorica, :
## prediction from a rank-deficient fit may be misleading
```

```
df_tercer_grupo$fit = predicciones[, "fit"]
df_tercer_grupo$lwr = predicciones[, "lwr"]
df_tercer_grupo$upr = predicciones[, "upr"]

p2 <- ggplot(df_tercer_grupo, aes(x = stroke, y = price, color = enginetype)) + geom_poi
nt() + geom_line(aes(y = lwr, color = enginetype), linetype = "dashed") + geom_line(aes
(y = upr, color = enginetype), linetype = "dashed") + geom_smooth(method = "lm", formula
= y ~ x, se = TRUE, level = 0.97) + theme_minimal() + labs(title = "Stroke vs Price por
Enginetype", x = "Stroke", y = "Price")

print(p2)
```



2. Puedes hacer el mismo análisis para otra categoría de la variable cualitativa, pero no es necesario, bastará con que justiques la categoría seleccionada anteriormente.

Se eligió esa categoría porque es la que es estadísticamente más significativa.

IV. Más allá:

-> Contesta la pregunta referida a la agrupación de variables que propuso la empresa para el análisis: ¿propondrías una nueva agrupación de las variables a la empresa automovilísitica?

```
datos = read.csv("./precios_autos.csv")
colnames(datos)
```

```
##  [1] "symboling"        "CarName"          "fueltype"         "carbody"
##  [5] "drivewheel"       "enginelocation"   "wheelbase"        "carlength"
##  [9] "carwidth"         "carheight"        "curbweight"       "enginetype"
## [13] "cylindernumber"   "enginesize"       "stroke"           "compressionratio"
## [17] "horsepower"       "peakrpm"          "citympg"          "highwaympg"
## [21] "price"
```

Teniendo en cuenta todas esas variables, definitivamente no me iría directamente a lo que tiene sentido lógico o empírco armar los grupos, me iría por un análisis de mercado o análisis estadístico que me ayude a agrupar las

-> Retoma todas las variables y haz un análisis estadístico muy leve (medias y correlación) de cómo crees que se deberían agrupar para analizarlas.

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
numericas <- datos %>%select(where(is.numeric))
correlacion <- cor(numericas, use = "complete.obs")
print(correlacion)
```

```
##                    symboling  wheelbase  carlength   carwidth   carheight
## symboling         1.000000000 -0.5319537 -0.3576115 -0.2329191 -0.54103820
## wheelbase        -0.531953682  1.0000000  0.8745875  0.7951436  0.58943476
## carlength        -0.357611523  0.8745875  1.0000000  0.8411183  0.49102946
## carwidth         -0.232919061  0.7951436  0.8411183  1.0000000  0.27921032
## carheight        -0.541038200  0.5894348  0.4910295  0.2792103  1.00000000
## curbweight       -0.227690588  0.7763863  0.8777285  0.8670325  0.29557173
## enginesize       -0.105789709  0.5693287  0.6833599  0.7354334  0.06714874
## stroke           -0.008735141  0.1609590  0.1295326  0.1829417 -0.05530667
## compressionratio -0.178515084  0.2497858  0.1584137  0.1811286  0.26121423
## horsepower        0.070872724  0.3532945  0.5526230  0.6407321 -0.10880206
## peakrpm           0.273606245 -0.3604687 -0.2872422 -0.2200123 -0.32041072
## citympg          -0.035822628 -0.4704136 -0.6709087 -0.6427043 -0.04863963
## highwaympg        0.034606001 -0.5440819 -0.7046616 -0.6772179 -0.10735763
## price            -0.079978225  0.5778156  0.6829200  0.7593253  0.11933623
##                   curbweight  enginesize       stroke compressionratio
## symboling        -0.2276906 -0.10578971 -0.008735141      -0.17851508
## wheelbase         0.7763863  0.56932868  0.160959047       0.24978585
## carlength         0.8777285  0.68335987  0.129532611       0.15841371
## carwidth          0.8670325  0.73543340  0.182941693       0.18112863
## carheight         0.2955717  0.06714874 -0.055306674       0.26121423
## curbweight        1.0000000  0.85059407  0.168790035       0.15136174
## enginesize        0.8505941  1.00000000  0.203128588       0.02897136
## stroke            0.1687900  0.20312859  1.000000000       0.18611011
## compressionratio  0.1513617  0.02897136  0.186110110       1.00000000
## horsepower        0.7507393  0.80976865  0.080939536      -0.20432623
## peakrpm          -0.2662432 -0.24465983 -0.067963753      -0.43574051
## citympg          -0.7574138 -0.65365792 -0.042144754       0.32470142
## highwaympg       -0.7974648 -0.67746991 -0.043930930       0.26520139
## price             0.8353049  0.87414480  0.079443084       0.06798351
##                   horsepower     peakrpm     citympg  highwaympg       price
## symboling         0.07087272  0.27360625 -0.03582263  0.03460600 -0.07997822
## wheelbase         0.35329448 -0.36046875 -0.47041361 -0.54408192  0.57781560
## carlength         0.55262297 -0.28724220 -0.67090866 -0.70466160  0.68292002
## carwidth          0.64073208 -0.22001230 -0.64270434 -0.67721792  0.75932530
## carheight        -0.10880206 -0.32041072 -0.04863963 -0.10735763  0.11933623
## curbweight        0.75073925 -0.26624318 -0.75741378 -0.79746479  0.83530488
## enginesize        0.80976865 -0.24465983 -0.65365792 -0.67746991  0.87414480
## stroke            0.08093954 -0.06796375 -0.04214475 -0.04393093  0.07944308
## compressionratio -0.20432623 -0.43574051  0.32470142  0.26520139  0.06798351
## horsepower        1.00000000  0.13107251 -0.80145618 -0.77054389  0.80813882
## peakrpm           0.13107251  1.00000000 -0.11354438 -0.05427481 -0.08526715
## citympg          -0.80145618 -0.11354438  1.00000000  0.97133704 -0.68575134
## highwaympg       -0.77054389 -0.05427481  0.97133704  1.00000000 -0.69759909
## price             0.80813882 -0.08526715 -0.68575134 -0.69759909  1.00000000
```

Con esto armaría:

Grupo con el tamaño y peso del vehículo: wheelbase, carlength, carwidth, curbweight

Grupo con el motor y rendimiento: enginesize, horsepower, curbweight, price

Grupo con el consumo de combustible: citympg, highwaympg, horsepower