Actividad Integradora 2

Andrés Villarreal González

2024-09-06

Actividad Integradora 2

Leyendo los datos

```
data <- read.csv("precios_autos.csv")
```

Seleccionamos variables del primer grupo

```
gas <- data$fueltype
dist <- data$wheelbase
cab <- data$horsepower
price <- data$price
df <- data.frame(gas, dist, cab, price)
```

cuantitativas (media, desviación estándar, cuantiles, etc)

```
summary(dist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   86.60   94.50   97.00   98.76  102.40  120.90
```

```
summary(cab)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    48.0    70.0    95.0   104.1   116.0   288.0
```

```
summary(price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5118    7788   10295   13277   16503   45400
```

```
mean(dist)
```

```
## [1] 98.75659
```

```
mean(cab)
```

```
## [1] 104.1171
```

```
mean(price)
```

```
## [1] 13276.71
```

```
sd(dist)
```

```
## [1] 6.021776
```

```r
sd(cab)
```

```
## [1] 39.54417
```

```r
sd(price)
```

```
## [1] 7988.852
```

```r
quantile(dist)
```

```
##    0%   25%   50%   75%  100%
##  86.6  94.5  97.0 102.4 120.9
```

```r
quantile(cab)
```

```
##   0%  25%  50%  75% 100%
##   48   70   95  116  288
```

```r
quantile(price)
```

```
##    0%   25%   50%   75%  100%
##  5118  7788 10295 16503 45400
```

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

```r
table(gas)
```

```
## gas
## diesel    gas
##     20    185
```

```r
prop.table(table(gas))
```

```
## gas
##     diesel        gas
## 0.09756098 0.90243902
```

Matriz de correlación

```r
# Calcular la matriz de correlación
cor_matrix <- cor(data[, c("wheelbase", "horsepower", "price")])

# Mostrar la matriz de correlación
print(cor_matrix)
```

```
##            wheelbase horsepower     price
## wheelbase  1.0000000  0.3532945 0.5778156
## horsepower 0.3532945  1.0000000 0.8081388
## price      0.5778156  0.8081388 1.0000000
```

No parece haber una colinealidad fuerte entre las variables independientes (wheelbase y horsepower), pero ambas están moderadamente correlacionadas con el precio, especialmente horsepower.

Explora los datos usando herramientas de visualización

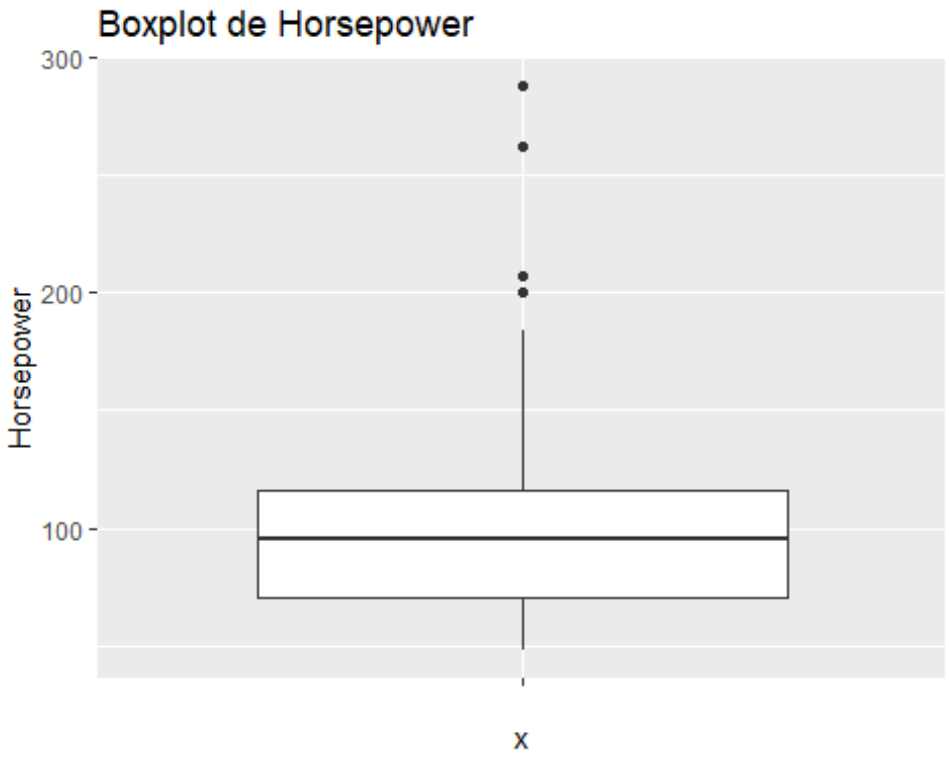
Variables Numericas

Boxplots

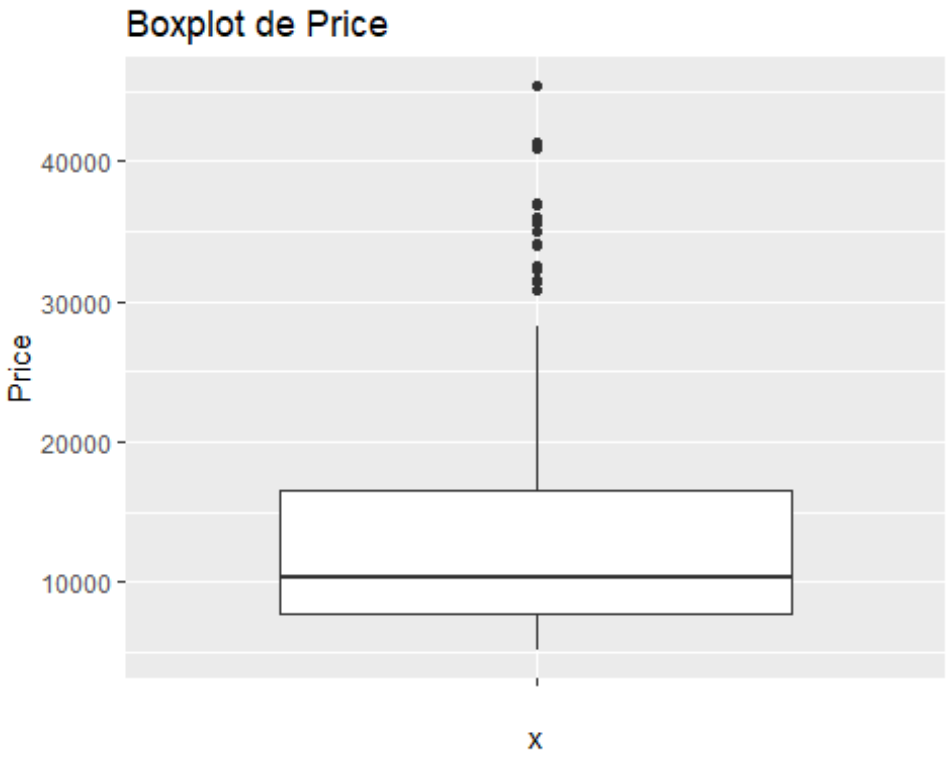
```r
# Cargar las librerías necesarias
library(ggplot2)

# Boxplot para wheelbase
ggplot(data, aes(x = "", y = wheelbase)) +
  geom_boxplot() +
  labs(title = "Boxplot de Wheelbase", y = "Wheelbase")
```


```r
# Boxplot para horsepower
ggplot(data, aes(x = "", y = horsepower)) +
  geom_boxplot() +
  labs(title = "Boxplot de Horsepower", y = "Horsepower")
```

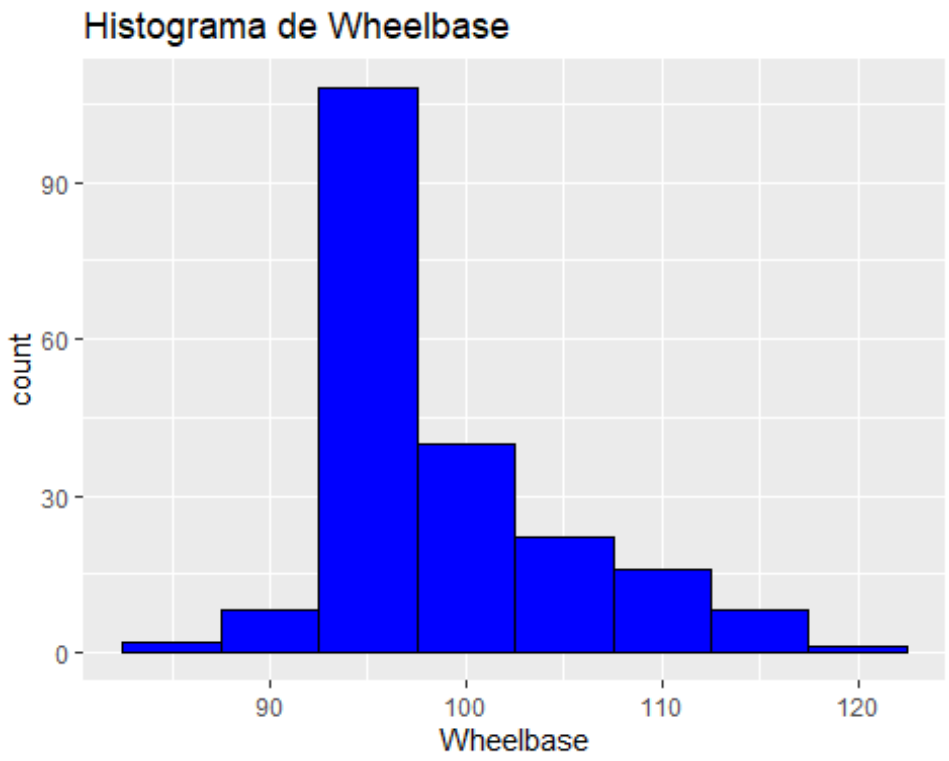



```r
# Boxplot para price
ggplot(data, aes(x = "", y = price)) +
  geom_boxplot() +
  labs(title = "Boxplot de Price", y = "Price")
```

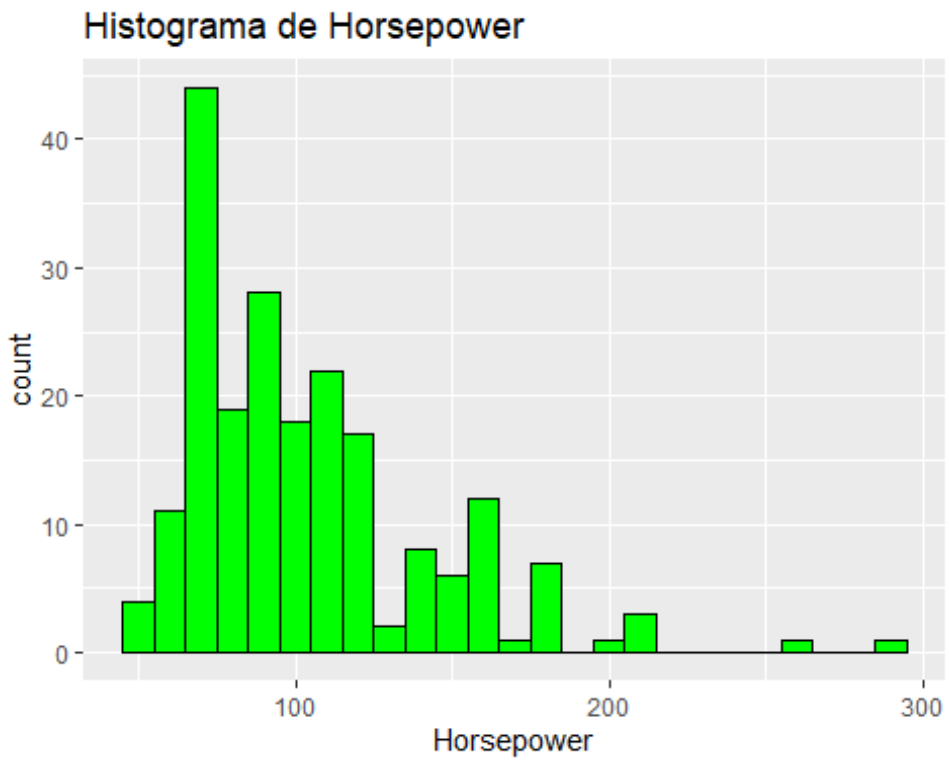


Histogramas

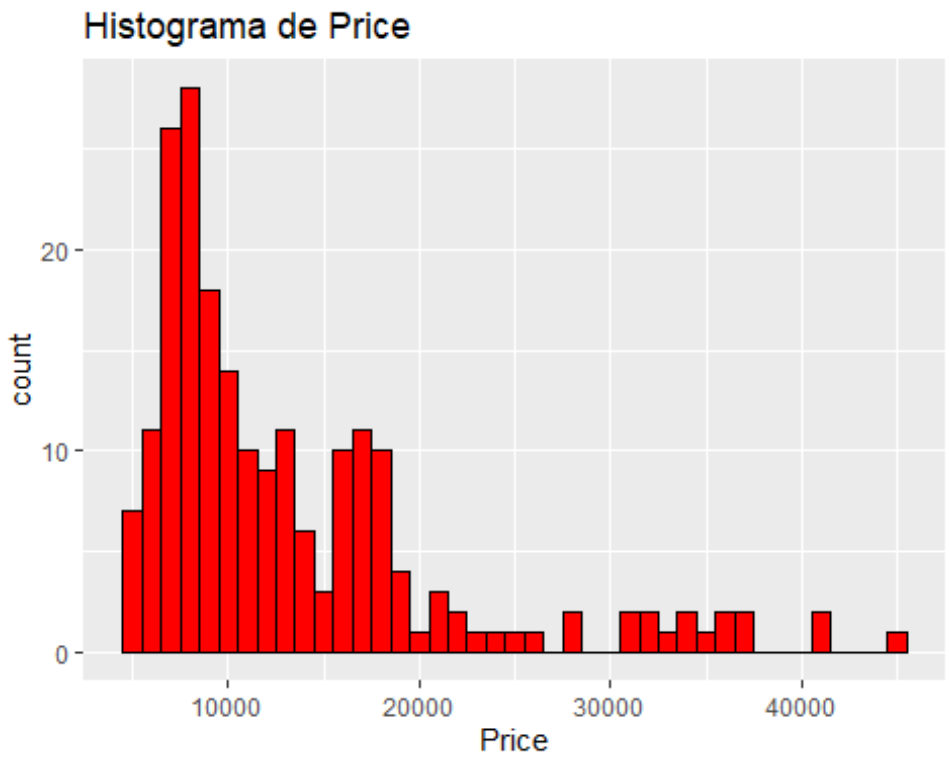
```
# Histograma para wheelbase
ggplot(data, aes(x = wheelbase)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Histograma de Wheelbase", x = "Wheelbase")
```




```r
# Histograma para horsepower
ggplot(data, aes(x = horsepower)) +
  geom_histogram(binwidth = 10, fill = "green", color = "black") +
  labs(title = "Histograma de Horsepower", x = "Horsepower")
```



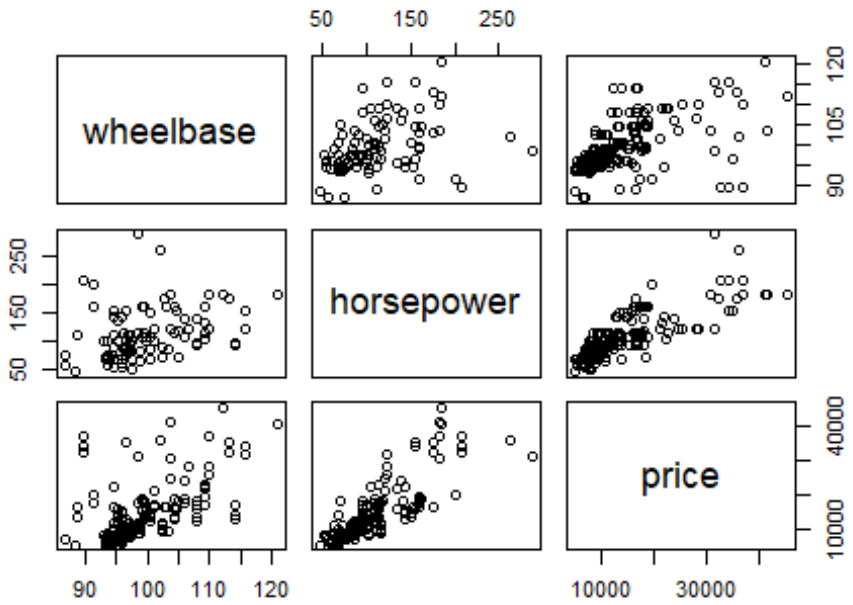

```r
# Histograma para price
ggplot(data, aes(x = price)) +
  geom_histogram(binwidth = 1000, fill = "red", color = "black") +
  labs(title = "Histograma de Price", x = "Price")
```



Diagramas de dispersion y correlación por pares

```
# Pares de variables con scatterplot
pairs(~wheelbase + horsepower + price, data = data,
      main = "Scatterplot de pares de Wheelbase, Horsepower y Price")
```

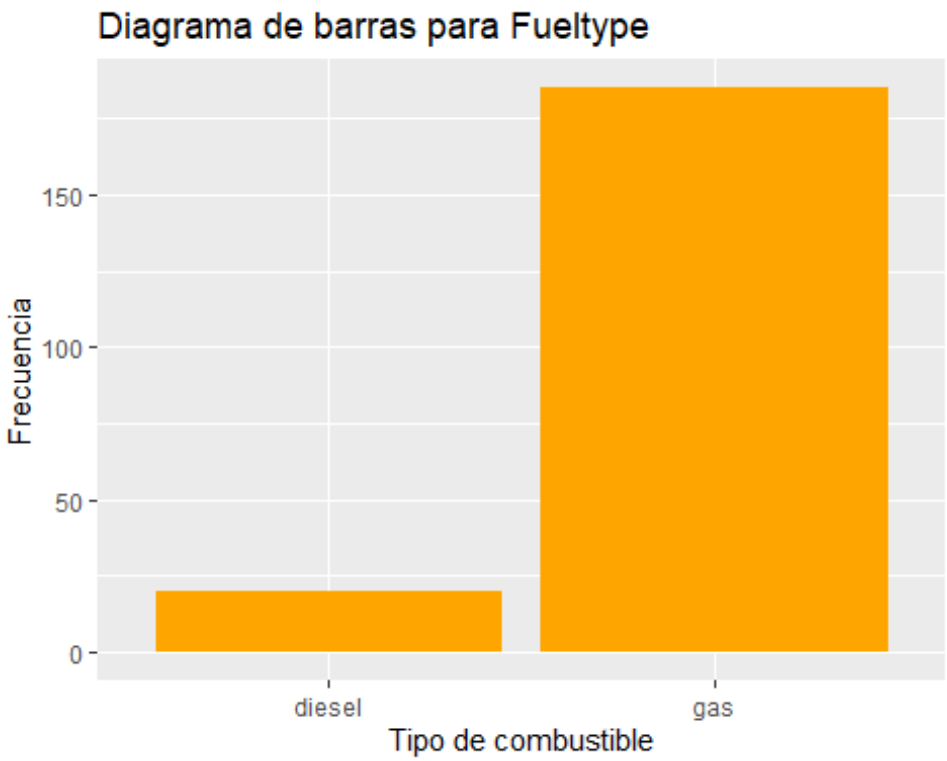
catterplot de pares de Wheelbase, Horsepower y Pric



Variables categoricas

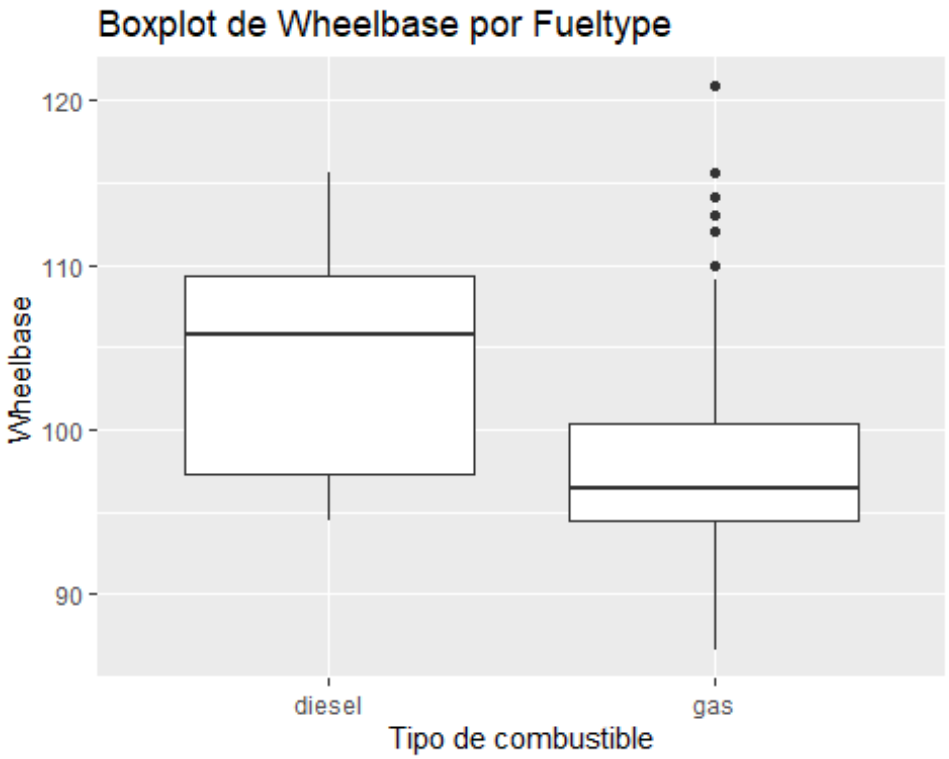
Diagrama de barras

```
# Diagrama de barras para fueltype
ggplot(data, aes(x = fueltype)) +
  geom_bar(fill = "orange") +
  labs(title = "Diagrama de barras para Fueltype", x = "Tipo de
combustible", y = "Frecuencia")
```

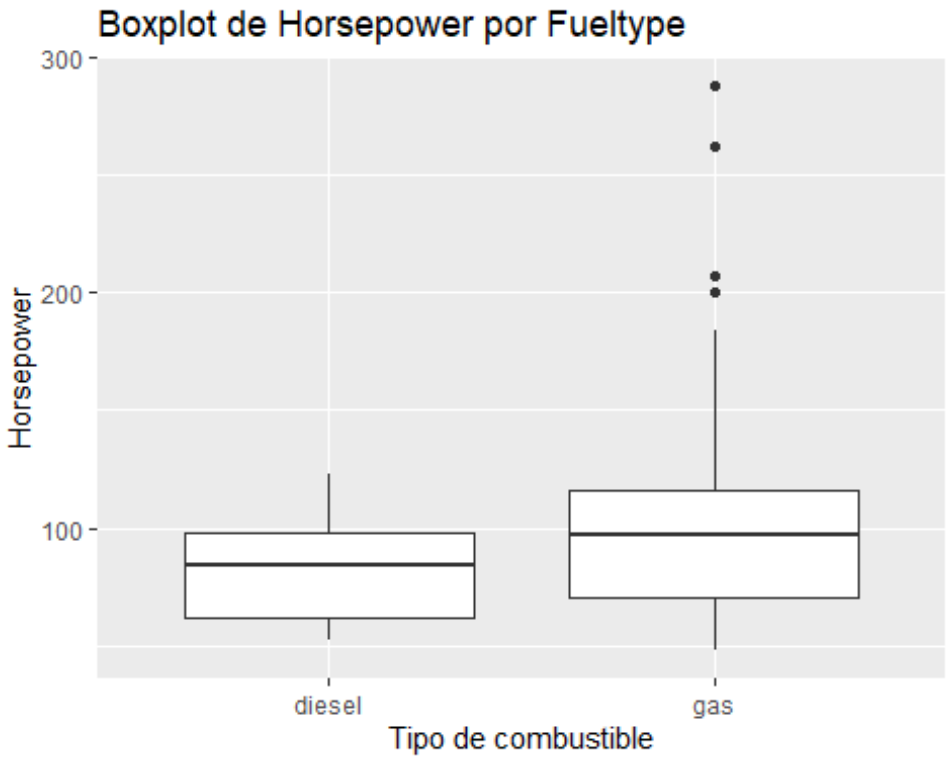


Boxplots

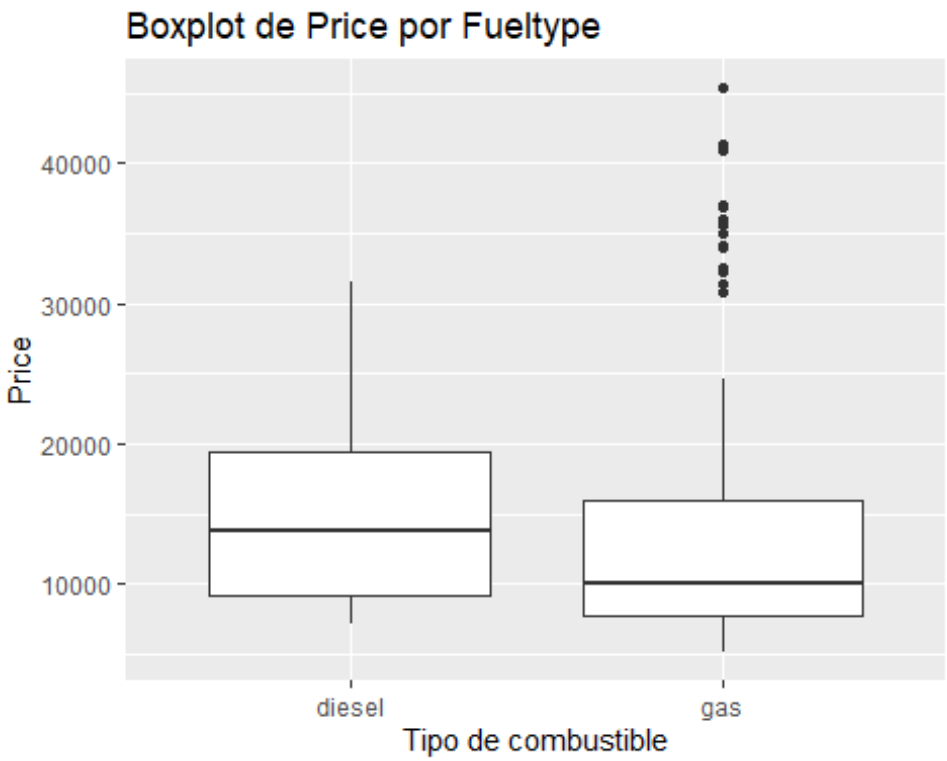
```
# Boxplot para wheelbase por fueltype
ggplot(data, aes(x = fueltype, y = wheelbase)) +
  geom_boxplot() +
  labs(title = "Boxplot de Wheelbase por Fueltype", x = "Tipo de
combustible", y = "Wheelbase")
```




```r
# Boxplot para horsepower por fueltype
ggplot(data, aes(x = fueltype, y = horsepower)) +
  geom_boxplot() +
  labs(title = "Boxplot de Horsepower por Fueltype", x = "Tipo de
combustible", y = "Horsepower")
```




```r
# Boxplot para price por fueltype
ggplot(data, aes(x = fueltype, y = price)) +
  geom_boxplot() +
  labs(title = "Boxplot de Price por Fueltype", x = "Tipo de
combustible", y = "Price")
```



Modelación y verificacion del modelo

Modelo 1

```
# Modelo 1: Regresión lineal múltiple con wheelbase, horsepower y
fueltype
model1 <- lm(price ~ wheelbase + horsepower + fueltype, data = data)

# Resumen del modelo
summary(model1)

##
## Call:
## lm(formula = price ~ wheelbase + horsepower + fueltype, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8650  -2191   -197   1606  15816
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34754.325   5314.194  -6.540 4.99e-10 ***
## wheelbase      364.657     52.594   6.933 5.48e-11 ***
## horsepower     148.323      7.723  19.205  < 2e-16 ***
## fueltypegas  -3794.450   1009.750  -3.758 0.000225 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3884 on 201 degrees of freedom
## Multiple R-squared:  0.7671, Adjusted R-squared:  0.7636
## F-statistic: 220.7 on 3 and 201 DF,  p-value: < 2.2e-16
```

Modelo 2

```
# Modelo 2: Regresión lineal múltiple con horsepower y wheelbase
model2 <- lm(price ~ horsepower + wheelbase, data = data)

# Resumen del modelo
summary(model2)

##
## Call:
## lm(formula = price ~ horsepower + wheelbase, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8403.9 -2303.7  -227.6  1608.4 15640.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44998.311   4707.546  -9.559  < 2e-16 ***
## horsepower     139.425      7.586  18.379  < 2e-16 ***
## wheelbase      443.095     49.818   8.894 3.33e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4008 on 202 degrees of freedom
## Multiple R-squared:  0.7507, Adjusted R-squared:  0.7482
## F-statistic: 304.2 on 2 and 202 DF,  p-value: < 2.2e-16
```

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

H_0: El modelo no es significativo. Todos los $\beta_i = 0$ H_1: El modelo es significativo. Al menos algún $\beta_i \neq 0$

Valida la significancia de βi con un alfa de 0.04

H_0: La variable no es significativa. $\beta_i = 0$ H_1: La variable es significativa. $\beta_i \neq 0$

```
cat("Modelo 1:","\n")

## Modelo 1:

anova(model1)

## Analysis of Variance Table
##
## Response: price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
```

```
## wheelbase    1 4346878264 4346878264 288.116 < 2.2e-16 ***
## horsepower   1 5427172318 5427172318 359.719 < 2.2e-16 ***
## fueltype     1  213049228  213049228  14.121 0.0002246 ***
## Residuals  201 3032539552   15087261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("Modelo 2:","\n")
```

```
## Modelo 2:
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## horsepower   1 8502974873 8502974873  529.21 < 2.2e-16 ***
## wheelbase    1 1271075709 1271075709   79.11 3.333e-16 ***
## Residuals  202 3245588780   16067271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para ambos modelos (Modelo 1 y Modelo 2), rechazamos H0 y concluimos que los modelos son significativos. Es decir, al menos una de las variables independientes tiene un efecto significativo sobre el precio.

Debido al valor p que es muy chico de todas las variables podemos concluir que todas son significativas para el modelo.

Porcentaje de variación explicada por el modelo

```
cat("R^2 modelo 1:",summary(model1)$adj.r.squared,"\n")
```

```
## R^2 modelo 1: 0.7636032
```

```
cat("R^2 modelo 2;",summary(model2)$adj.r.squared,"\n")
```

```
## R^2 modelo 2; 0.7482478
```

El Modelo 1 explica aproximadamente el 76.36% de la variación en el precio de los autos. Este es un buen nivel de ajuste.

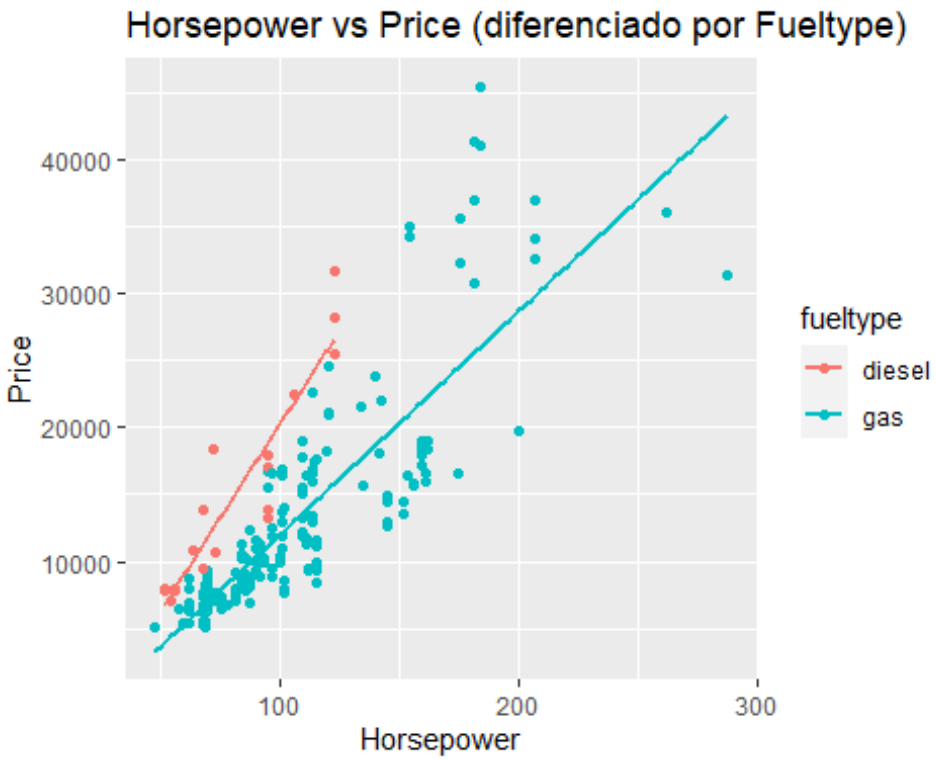
El Modelo 2 explica aproximadamente el 74.82% de la variación en el precio de los autos. Este es un buen nivel de ajuste.

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

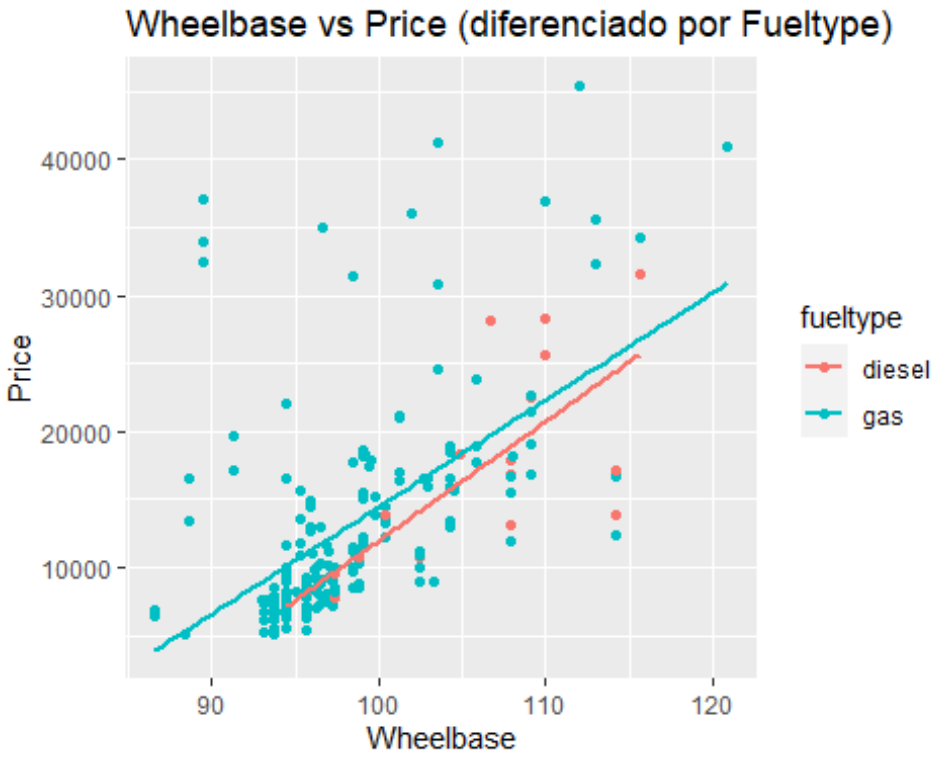
Primer modelo (separando por Fueltype)

```
# Gráfico de dispersión con recta de mejor ajuste, diferenciada por
fueltype
```

```r
ggplot(data, aes(x = horsepower, y = price, color = fueltype)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  labs(title = "Horsepower vs Price (diferenciado por Fueltype)", x =
"Horsepower", y = "Price")
```

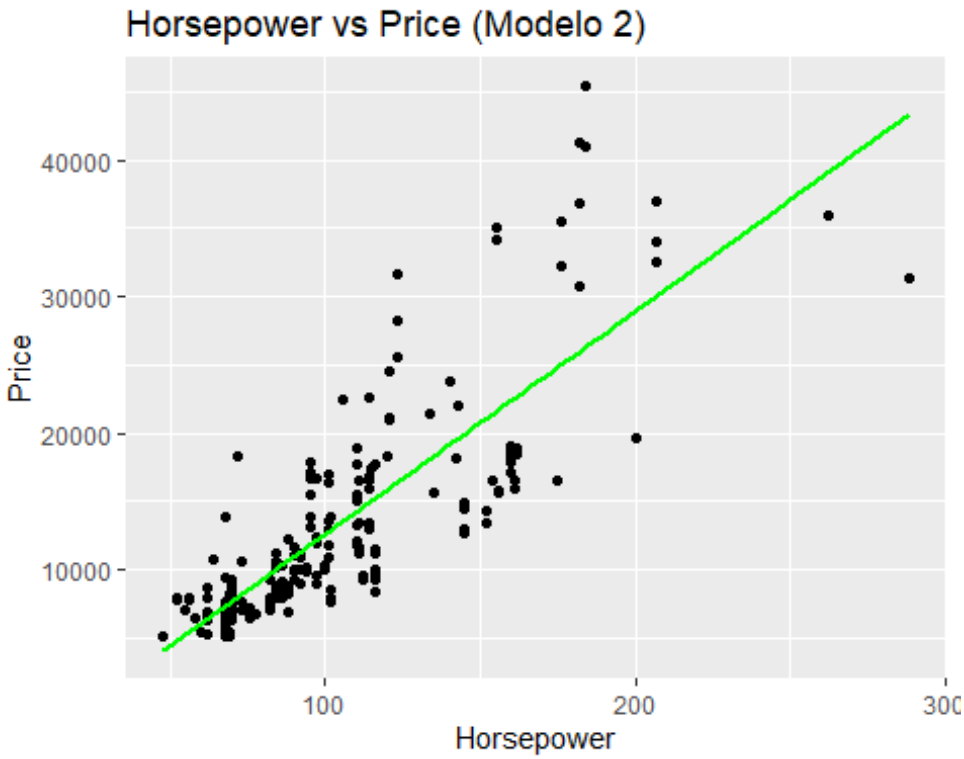



```r
# Gráfico de dispersión con recta de mejor ajuste, diferenciada por
fueltype
ggplot(data, aes(x = wheelbase, y = price, color = fueltype)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  labs(title = "Wheelbase vs Price (diferenciado por Fueltype)", x =
"Wheelbase", y = "Price")
```

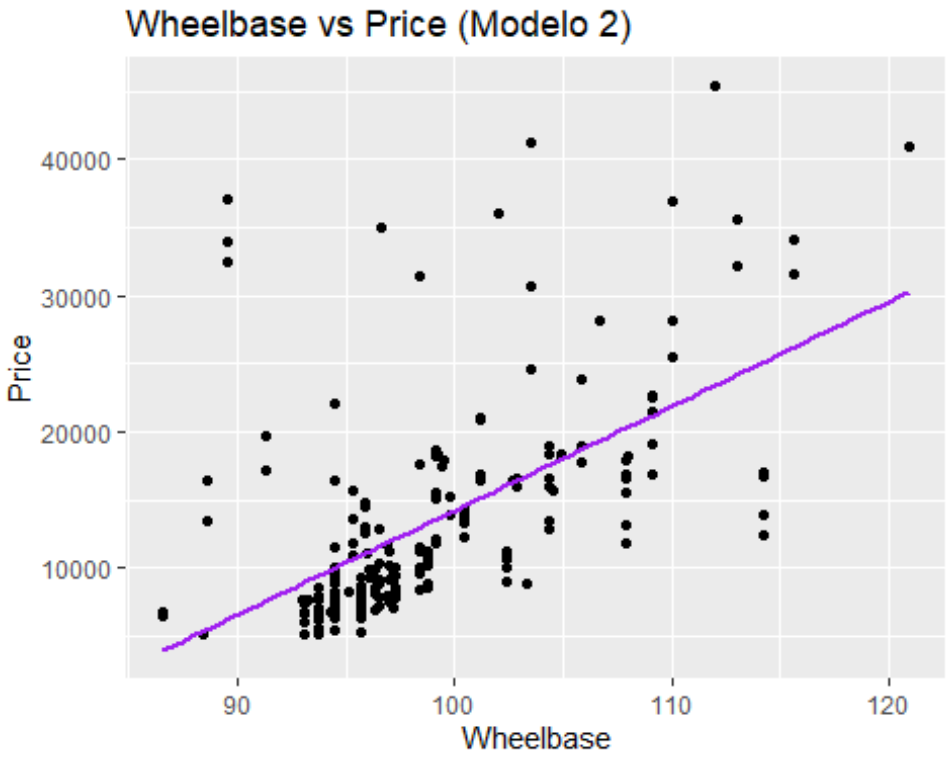


Segundo modelo (separando por fueltype)

```
# Gráfico de dispersión con recta de mejor ajuste para horsepower y price (Modelo 2)
ggplot(data, aes(x = horsepower, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "green") +
  labs(title = "Horsepower vs Price (Modelo 2)", x = "Horsepower", y = "Price")
```




```r
# Gráfico de dispersión con recta de mejor ajuste para wheelbase y price (Modelo 2)
ggplot(data, aes(x = wheelbase, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "purple") +
  labs(title = "Wheelbase vs Price (Modelo 2)", x = "Wheelbase", y = "Price")
```



Normalidad de residuos

H_0: Los datos provienen de una población normal H_1: Los datos no provienen de una población normal

```
library(nortest)
ad.test(model1$residuals)
```

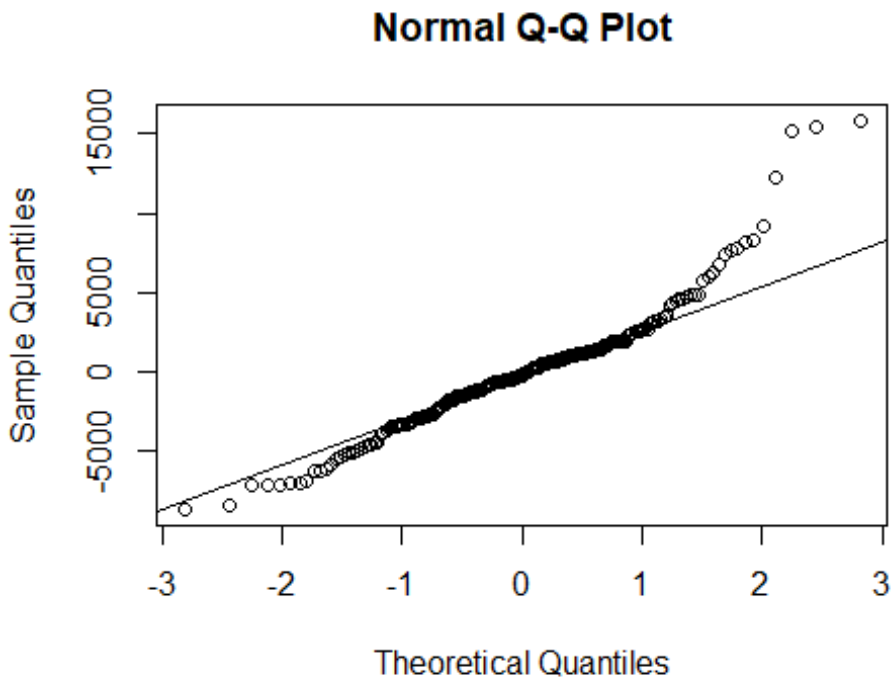
```
##
##  Anderson-Darling normality test
##
## data:  model1$residuals
## A = 2.7561, p-value = 5.82e-07
```

```
ad.test(model2$residuals)
```

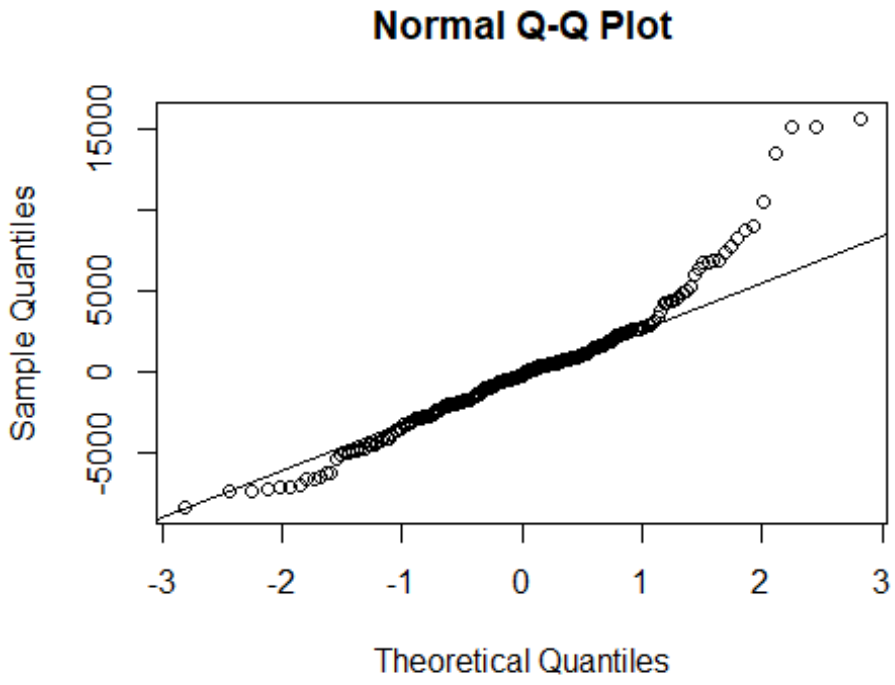
```
##
##  Anderson-Darling normality test
##
## data:  model2$residuals
## A = 2.8064, p-value = 4.385e-07
```

Debido al valor bajo del p-valor para ambos modelos rechazamos H0, por lo que se puede decir que los datos no siguen normalidad

```
qqnorm(model1$residuals)
qqline(model1$residuals)
```




```r
qqnorm(model2$residuals)
qqline(model2$residuals)
```



Verificación de media cero

$H_0: \mu = 0$ $H_1: \mu \neq 0$

```
t.test(model1$residuals)
```

```
##
##  One Sample t-test
##
## data:  model1$residuals
## t = 5.6678e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -530.9376  530.9376
## sample estimates:
##    mean of x
## 1.526259e-13
```

```
t.test(model2$residuals)
```

```
##
##  One Sample t-test
##
## data:  model2$residuals
## t = 2.4215e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -549.2714  549.2714
## sample estimates:
##    mean of x
## 6.745823e-14
```

Para ambos modelos el valor p es alto por lo que no tenemos evidencia suficiente para rechazar H0

Homocedasticidad e Independencia

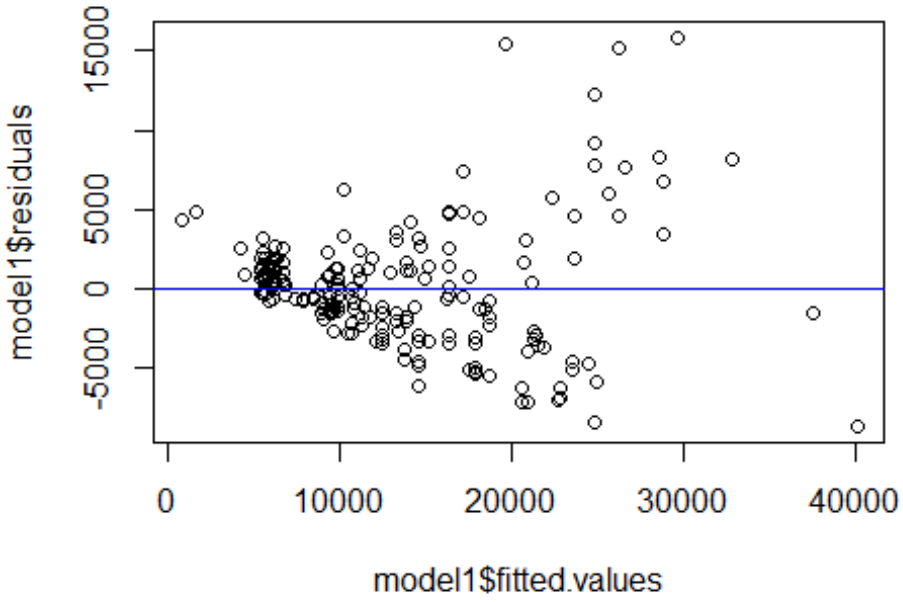
Homocedasticidad

H_0: La varianza de los errores es constante (homocedasticidad) H_1: La varianza de los errores no es constante (heterocedasticidad)

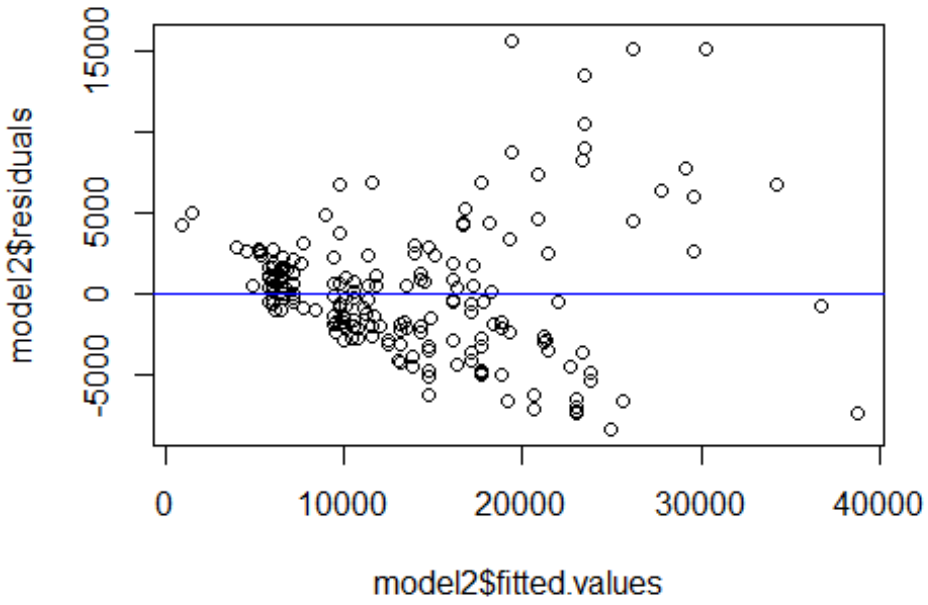
Independencia

H_0: Los errores no están correlacionados H_1: Los errores están correlacionados

```
plot(model1$fitted.values,model1$residuals)
abline(h=0, col="blue")
```




```
plot(model2$fitted.values,model2$residuals)
abline(h=0, col="blue")
```




```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

dwtest(model1)

##
##  Durbin-Watson test
##
## data:  model1
## DW = 0.97856, p-value = 4.496e-14
## alternative hypothesis: true autocorrelation is greater than 0

bgtest(model1)

##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  model1
## LM test = 57.79, df = 1, p-value = 2.916e-14

dwtest(model2)

##
##  Durbin-Watson test
##
## data:  model2
## DW = 0.98038, p-value = 5.339e-14
## alternative hypothesis: true autocorrelation is greater than 0

bgtest(model2)

##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  model2
## LM test = 57.47, df = 1, p-value = 3.432e-14
```

Para ambos modelos el valor p es muy bajo por lo que rechazamos ambas hipotesis nulas

```
# Filtrar los datos para la categoría 'gas'
data_gas <- subset(data, fueltype == "gas")

# Ajustar el modelo solo para los autos que usan gasolina (gas)
best_model_gas <- lm(price ~ horsepower + wheelbase, data = data_gas)
```

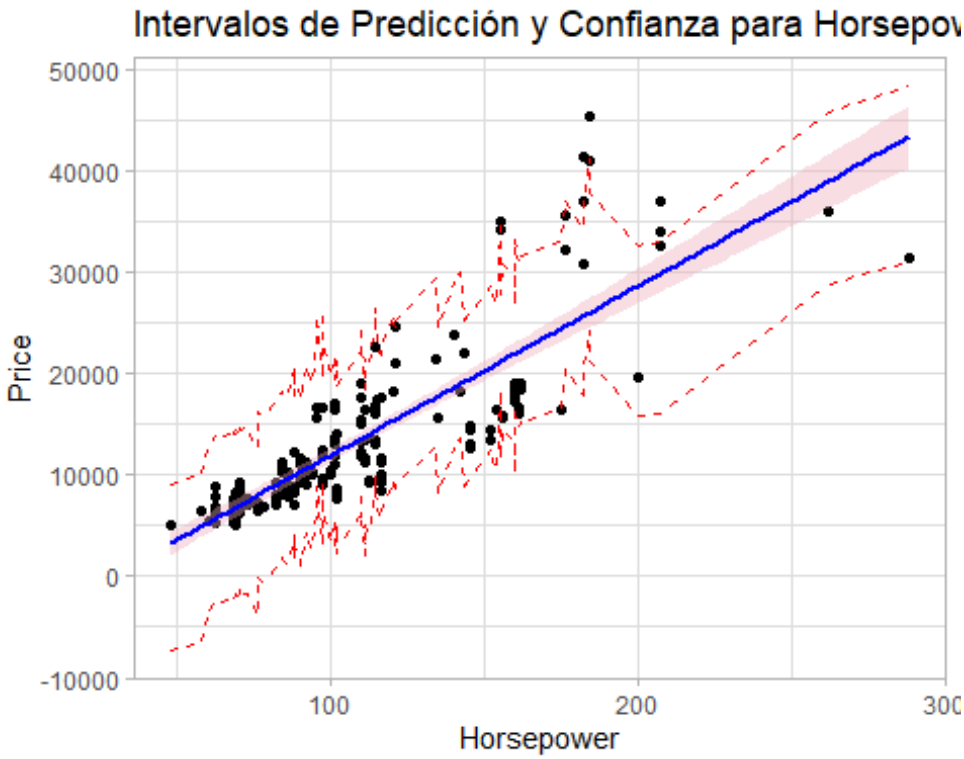
```r
# Generar los intervalos de predicción con un nivel de confianza del 97%
intervals_gas <- predict(object = best_model_gas, interval =
"prediction", level = 0.96)

## Warning in predict.lm(object = best_model_gas, interval =
"prediction", : predictions on current data refer to _future_ responses

# Añadir los intervalos de predicción a los datos originales filtrados
data_gas_with_intervals <- cbind(data_gas, intervals_gas)

# Cargar ggplot2 para graficar
library(ggplot2)

# Crear la gráfica con intervalos de predicción y confianza para la
categoría 'gas'
ggplot(data_gas_with_intervals, aes(x = horsepower, y = price)) +
  geom_point() +  # Puntos de los datos reales
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +  # Límite
inferior del intervalo de predicción
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +  # Límite
superior del intervalo de predicción
  geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.96, col
= "blue", fill = "pink2") +  # Línea de regresión con intervalo de
confianza
  labs(title = "Intervalos de Predicción y Confianza para Horsepower vs
Price (Gas)",
       x = "Horsepower", y = "Price") +
  theme_light()  # Tema de la gráfica
```



Conclusiones

Concluye sobre el mejor modelo que encontraste y argumenta por qué es el mejor?

El Modelo 1 tiene un R^2 de 0.7636, lo que significa que explica aproximadamente el 76.36% de la variación en el precio de los autos. En comparación, el Modelo 2 tiene un R^2 de 0.7482. Aunque la diferencia no es muy grande, el Modelo 1 tiene una capacidad explicativa ligeramente superior.

¿Cuáles de las variables asignadas influyen en el precio del auto? ¿de qué manera lo hacen? Horsepower es la variable que más influye en el precio de manera positiva: a mayor potencia del auto, mayor es el precio. Wheelbase también tiene un impacto positivo: autos con mayor distancia entre ejes tienden a ser más caros. Fueltype (gasolina vs diésel) tiene un impacto negativo: autos a gasolina tienden a ser más económicos en comparación con autos a diésel.

¿propondrías una nueva agrupación de las variables a la empresa automovilísitica? Sí, propondría una nueva agrupación de variables que divida los aspectos del automóvil en grupos relacionados con el rendimiento mecánico, características físicas y características adicionales o de lujo. Esta nueva agrupación mejoraría la capacidad de predicción de los modelos y permitiría una mejor interpretación de los factores que afectan el precio del automóvil en diferentes segmentos del mercado.