Actividad Integradora 2

Juan Bernal

2024-09-06

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. Las variables recopiladas vienen descritas en el 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:

- Primer grupo. Distancia entre los ejes (wheelbase), tipo de gasolina que usa y caballos de fuerza

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?

Con el grupo de variables seleccionadas realiza el siguiente procesamiento de los datos:

1. Exploración de la base de datos

```
data = read.csv('precios_autos.csv')
head(data)
```

```
##   symboling                 CarName fueltype     carbody drivewheel
## 1         3       alfa-romero giulia      gas convertible        rwd
## 2         3      alfa-romero stelvio      gas convertible        rwd
## 3         1 alfa-romero Quadrifoglio      gas   hatchback        rwd
## 4         2              audi 100 ls      gas       sedan        fwd
## 5         2               audi 100ls      gas       sedan        4wd
## 6         2                 audi fox      gas       sedan        fwd
##   enginelocation wheelbase carlength carwidth carheight curbweight enginetype
## 1          front      88.6     168.8     64.1      48.8       2548       dohc
## 2          front      88.6     168.8     64.1      48.8       2548       dohc
## 3          front      94.5     171.2     65.5      52.4       2823       ohcv
## 4          front      99.8     176.6     66.2      54.3       2337        ohc
## 5          front      99.4     176.6     66.4      54.3       2824        ohc
## 6          front      99.8     177.3     66.3      53.1       2507        ohc
##   cylindernumber enginesize stroke compressionratio horsepower peakrpm citympg
## 1           four        130   2.68              9.0        111    5000      21
## 2           four        130   2.68              9.0        111    5000      21
## 3            six        152   3.47              9.0        154    5000      19
## 4           four        109   3.40             10.0        102    5500      24
## 5           five        136   3.40              8.0        115    5500      18
## 6           five        136   3.40              8.5        110    5500      19
##   highwaympg price
## 1         27 13495
## 2         27 16500
## 3         26 16500
## 4         30 13950
## 5         22 17450
## 6         25 15250
```

1. Calcula medidas estadísticas apropiadas para las variables:

```
x = data.frame(data$fueltype, data$horsepower, data$wheelbase, data$price)
x$data.fueltype = ifelse(x$data.fueltype == "gas", 1, 0) # Transforma los "gas" a 1 y "diesel" a 0

head(x)
```

```
##   data.fueltype data.horsepower data.wheelbase data.price
## 1             1             111           88.6      13495
## 2             1             111           88.6      16500
## 3             1             154           94.5      16500
## 4             1             102           99.8      13950
## 5             1             115           99.4      17450
## 6             1             110           99.8      15250
```

```
x1 = subset(x, x$data.fueltype==1)
x0 = subset(x, x$data.fueltype==0)
```

1. cuantitativas (media, desviación estándar, cuantiles, etc)

```
summary(data$wheelbase)
```

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

```
print('Desviación estándar')
```

```
## [1] "Desviación estándar"
```

```
sd(data$wheelbase)
```

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

```
summary(data$horsepower)
```

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

```
print('Desviación estándar')
```

```
## [1] "Desviación estándar"
```

```
sd(data$horsepower)
```

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

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

```
prop.table(table(data$fueltype))
```

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

```
table(data$fueltype)
```

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

2. Analiza la correlación entre las variables (analiza posible colinealidad entre las variables)

```
cor(x)
```

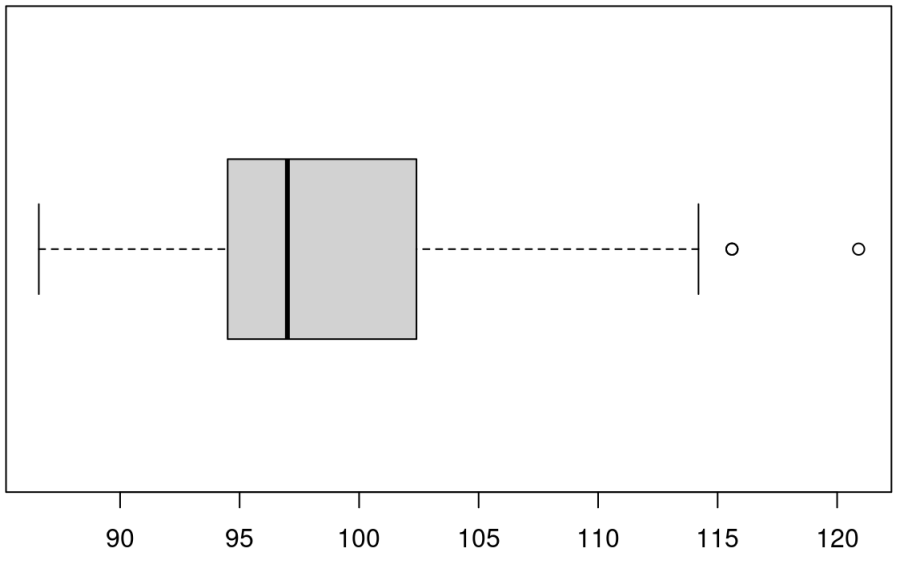
```
##                 data.fueltype data.horsepower data.wheelbase data.price
## data.fueltype       1.0000000       0.1639262     -0.3083459 -0.1056795
## data.horsepower     0.1639262       1.0000000      0.3532945  0.8081388
## data.wheelbase     -0.3083459       0.3532945      1.0000000  0.5778156
## data.price         -0.1056795       0.8081388      0.5778156  1.0000000
```

3. Explora los datos usando herramientas de visualización (si lo consideras necesario):

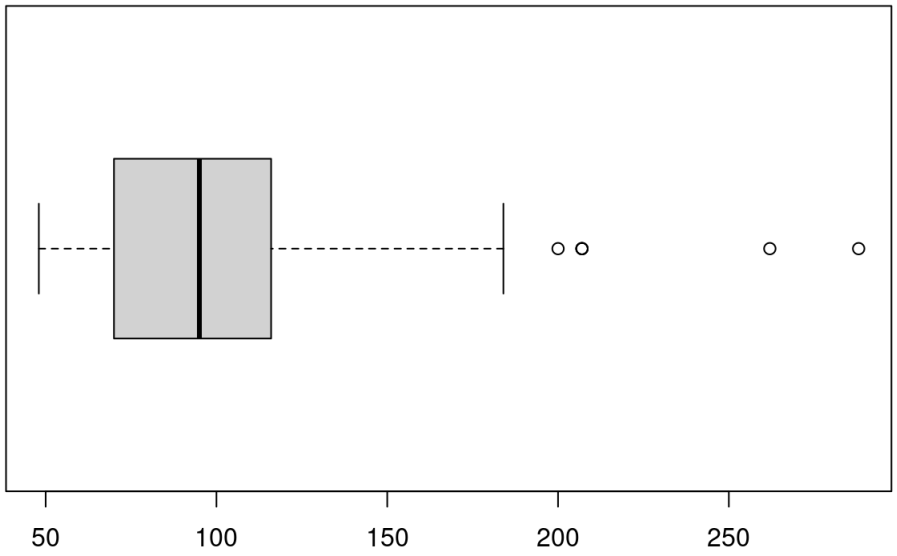
1. Variables cuantitativas:

* Boxplot (visualización de datos atípicos)

```
boxplot(x$data.wheelbase, horizontal = TRUE)
```

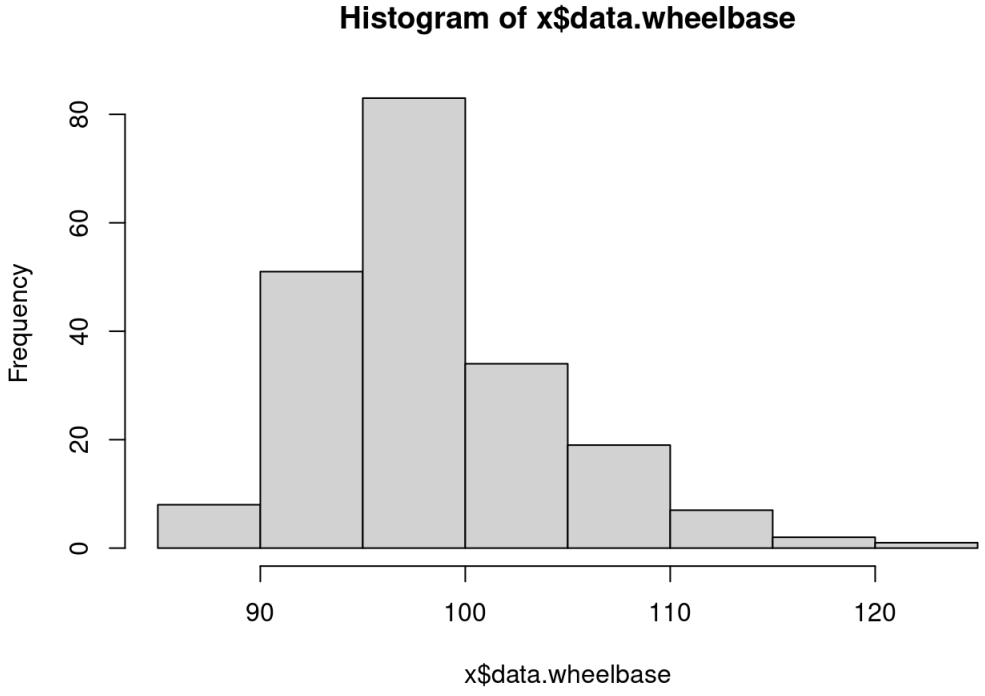



```
boxplot(x$data.horsepower, horizontal = TRUE)
```

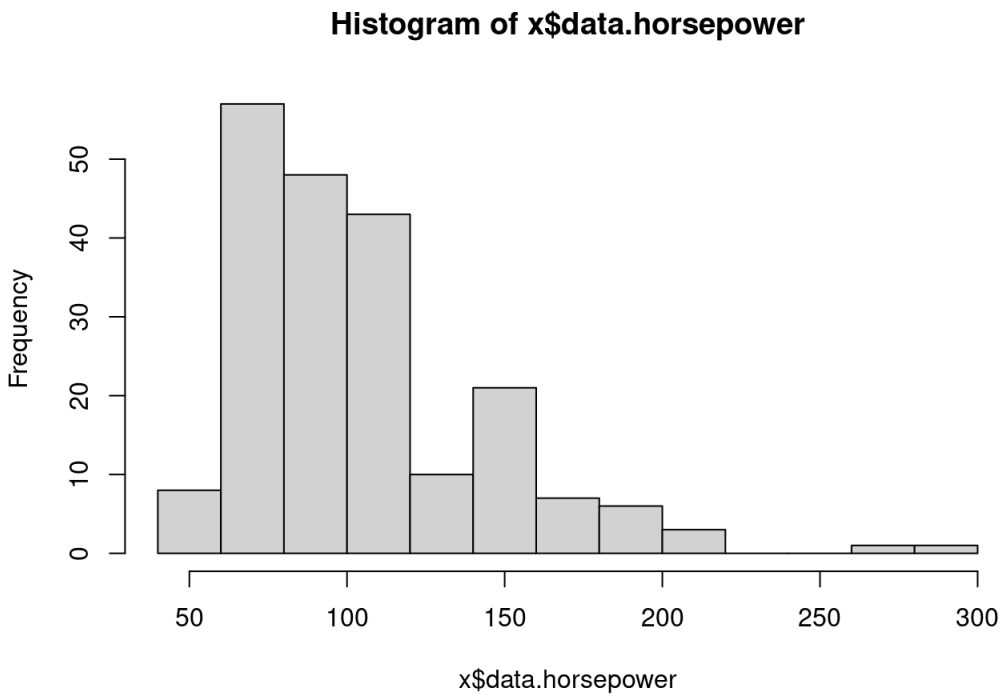


* Histogramas

```
hist(x$data.wheelbase)
```

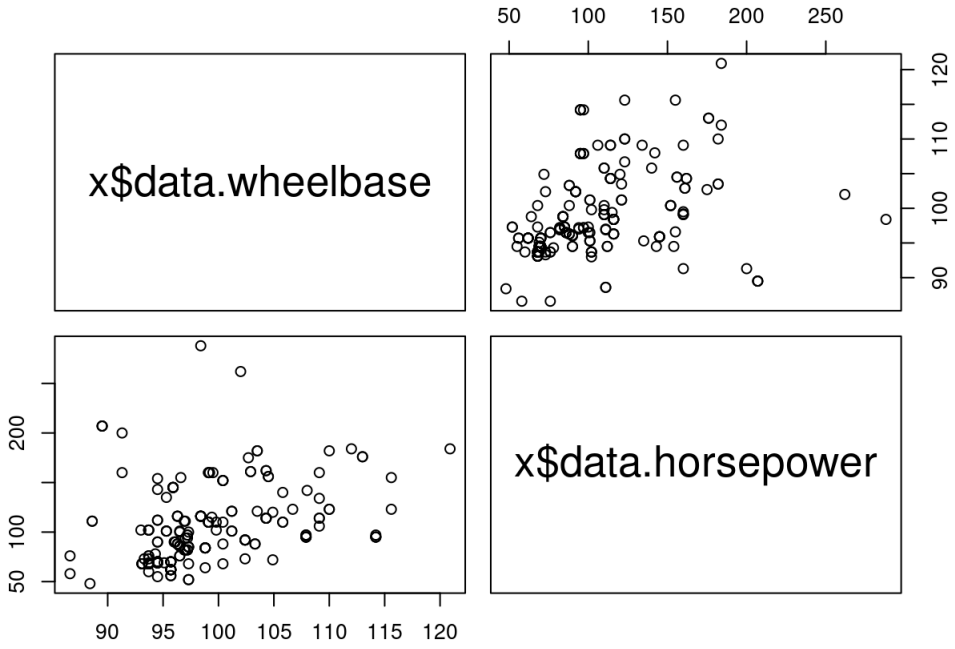



```
hist(x$data.horsepower)
```



* Diagramas de dispersión y correlación por pares

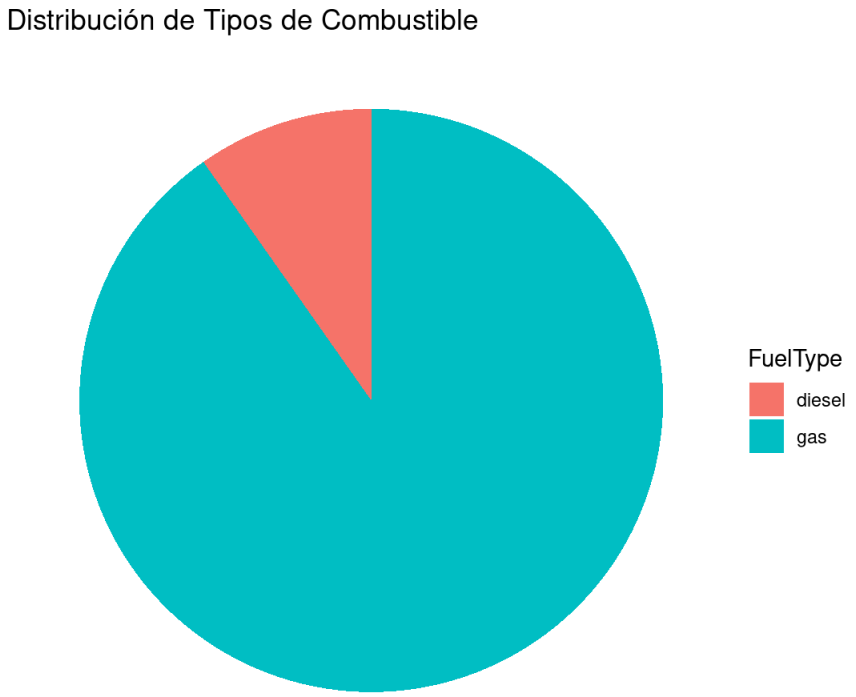
```
pairs(x$data.wheelbase ~ x$data.horsepower)
```



2. Variables categóricas

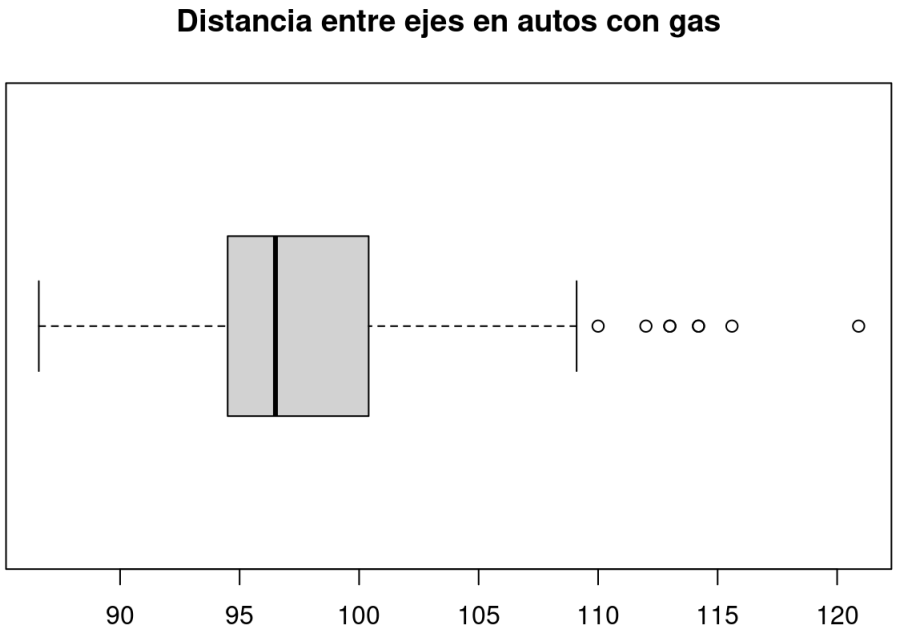
* Distribución de los datos (diagramas de barras, diagramas de pastel)

```
library(ggplot2)
conteos <- as.data.frame(table(data$fueltype))
colnames(conteos) <- c("FuelType", "Count")
ggplot(conteos, aes(x = "", y = Count, fill = FuelType)) +
geom_bar(width = 1, stat = "identity") +
coord_polar(theta = "y") +
labs(title = "Distribución de Tipos de Combustible") +
theme_void()
```

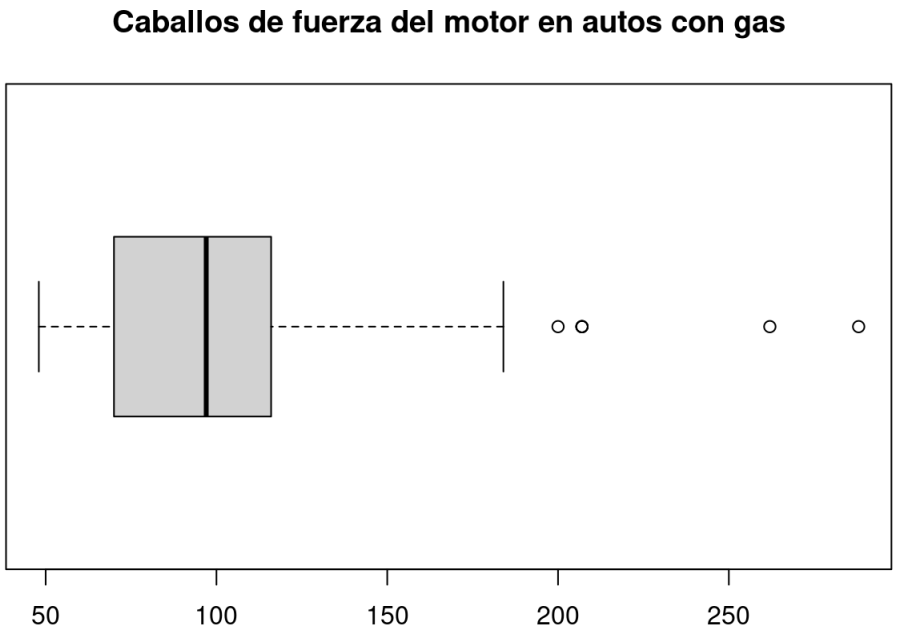


* Boxplot por categoría de las variables cuantitativas

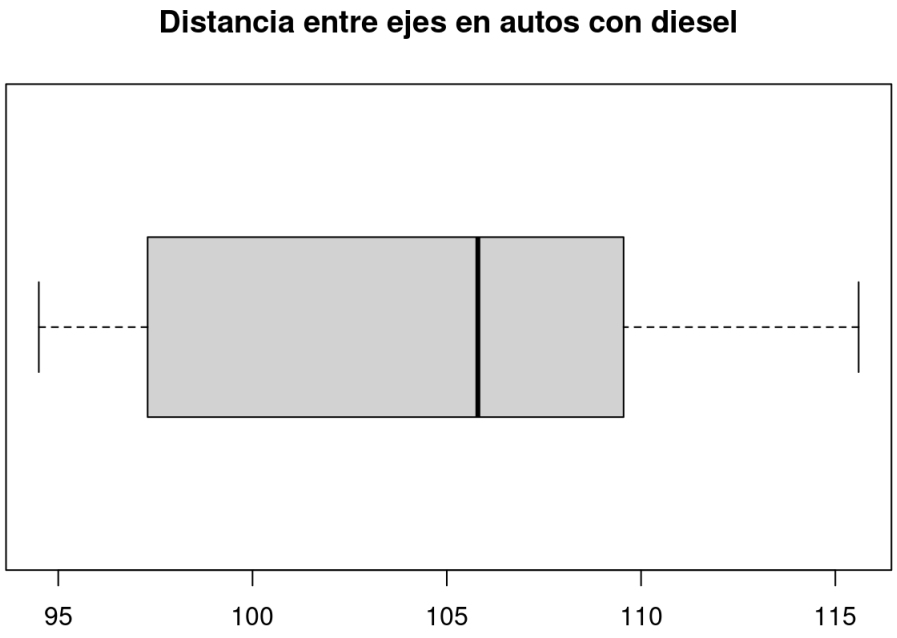
```
boxplot(x1$data.wheelbase, horizontal = TRUE, main = 'Distancia entre ejes en autos con gas')
```



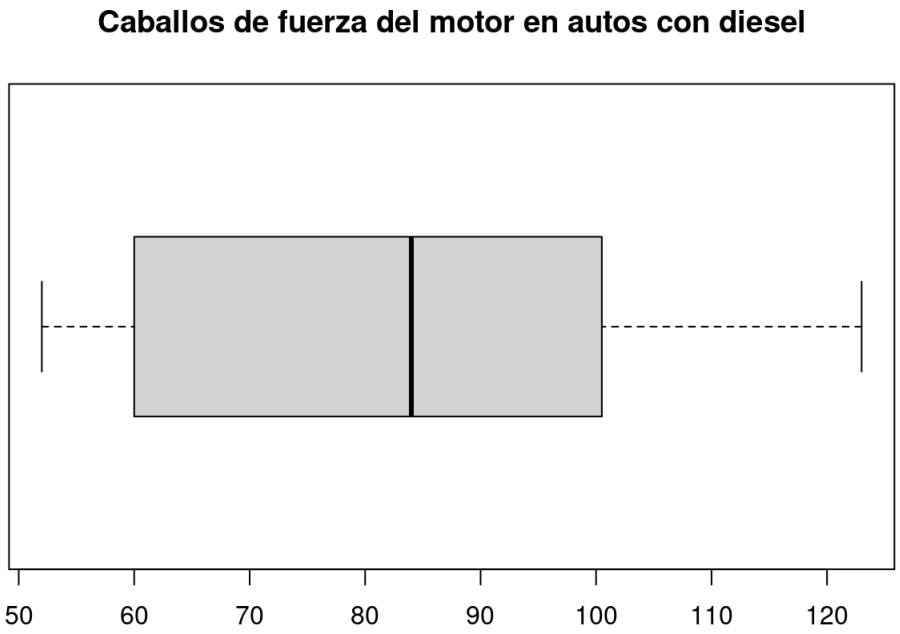
```
boxplot(x1$data.horsepower, horizontal = TRUE, main = 'Caballos de fuerza del motor en autos con gas')
```




```
boxplot(x0$data.wheelbase, horizontal = TRUE, main = 'Distancia entre ejes en autos con diesel')
```




```
boxplot(x0$data.horsepower, horizontal = TRUE, main = 'Caballos de fuerza del motor en autos con diesel')
```



2. 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.

Se propone un modelo en donde solo se consideren los caballos de fuerza, otro donde se considere únicamente la distancia entre los ejes, y dos modelos más que consideren los caballos de fuerza en función del tipo de combustible que requiera el auto (gas o diesel).

Nótese que no se propuso ningún modelo con más de dos variables, debido a que no se desea manejar modelos multivariables.

2. Para cada uno de los modelos propuestos:

1. Realiza la regresión entre las variables involucradas

Modelo de predicción del precio de un auto en función de los caballos de fuerza del motor:

- E_hp = -3721.761 + 163.263*Horsepower

```
rl2 = lm(x$data.price~x$data.horsepower)
summary(rl2)
```

```
##
## Call:
## lm(formula = x$data.price ~ x$data.horsepower)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.5  -2350.4   -711.1   1644.6  19081.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3721.761    929.849  -4.003 8.78e-05 ***
## x$data.horsepower   163.263      8.351  19.549  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4717 on 203 degrees of freedom
## Multiple R-squared:  0.6531, Adjusted R-squared:  0.6514
## F-statistic: 382.2 on 1 and 203 DF,  p-value: < 2.2e-16
```

Modelo de predicción del precio de un auto en función de la distancia entre los ejes:

- E_wb = -62426.7 + 766.6*Wheelbase

```
rl = lm(data$price~x$data.wheelbase)
summary(rl)
```

```
##
## Call:
## lm(formula = data$price ~ x$data.wheelbase)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12675  -3364  -1956   1264  30847
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -62426.7     7519.0  -8.303 1.42e-14 ***
## x$data.wheelbase    766.6       76.0  10.087  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6536 on 203 degrees of freedom
## Multiple R-squared:  0.3339, Adjusted R-squared:  0.3306
## F-statistic: 101.7 on 1 and 203 DF,  p-value: < 2.2e-16
```

Modelo de predicción del precio de un auto en función de los caballos de fuerza del motor a gas:

- E_hp_g = -4714.538 + 166.734*Horsepower

```
rl3 = lm(x1$data.price~x1$data.horsepower)
summary(rl3)
```

```
##
## Call:
## lm(formula = x1$data.price ~ x1$data.horsepower)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11904.3  -1831.6   -394.4   1458.9  19435.5
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -4714.538    904.611  -5.212 5.02e-07 ***
## x1$data.horsepower   166.734      7.966  20.931  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4346 on 183 degrees of freedom
## Multiple R-squared:  0.7054, Adjusted R-squared:  0.7038
## F-statistic: 438.1 on 1 and 183 DF,  p-value: < 2.2e-16
```

Modelo de predicción del precio de un auto en función de los caballos de fuerza del motor a diesel:

- E_hp_d = -7731.37 + 279.09*Horsepower

```
rl4 = lm(x0$data.price~x0$data.horsepower)
summary(rl4)
```

```
##
## Call:
## lm(formula = x0$data.price ~ x0$data.horsepower)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5582.6 -1718.7   -54.9  1304.8  5980.6
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -7731.37    2225.19  -3.474  0.00271 **
## x0$data.horsepower   279.09      25.24  11.057 1.86e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2856 on 18 degrees of freedom
## Multiple R-squared:  0.8717, Adjusted R-squared:  0.8645
## F-statistic: 122.3 on 1 and 18 DF,  p-value: 1.862e-09
```

2. Analiza la significancia del modelo:

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

Hipótesis:

- $H_0 : \beta = 0$ El modelo no es significante
- $H_1 : \beta \neq 0$ El modelo es significante

Dado el valor de significancia de 0.04 y que todos los p-value de los modelos son menores, los cuatro modelos son significativos.

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

- $H_0 : \beta_0 = \beta_1 = 0$
- $H_1 : \exists \beta_i \neq 0$

Dado el valor de significancia de 0.04 y los p-value de cada beta, las betas de todos los modelos son significantes, es decir, la distancia entre los ejes y los caballos de fuerza del motor son todos significantes para determinar el precio de un auto.

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

El modelo que considera únicamente los caballos de fuerza del motor explica un 65.31% de la variación de los datos.

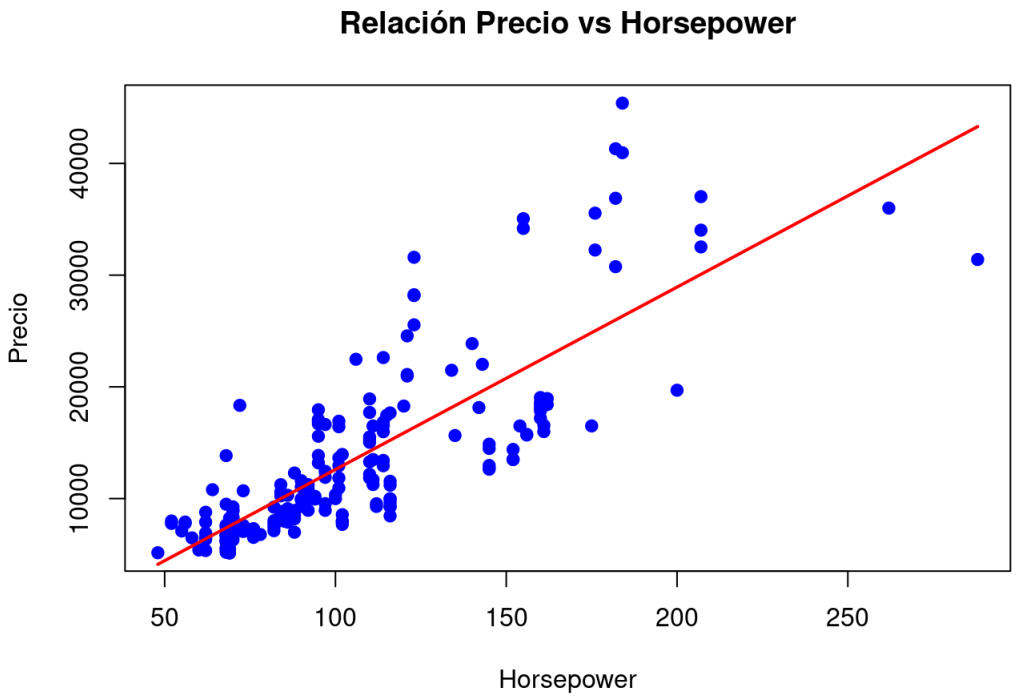
El modelo que considera la distancia entre los ejes explica el 33.39% de la variación de los datos.

El modelo que considera los caballos de fuerza un motor a gas explica un 70.54% de la variación del 90% datos.

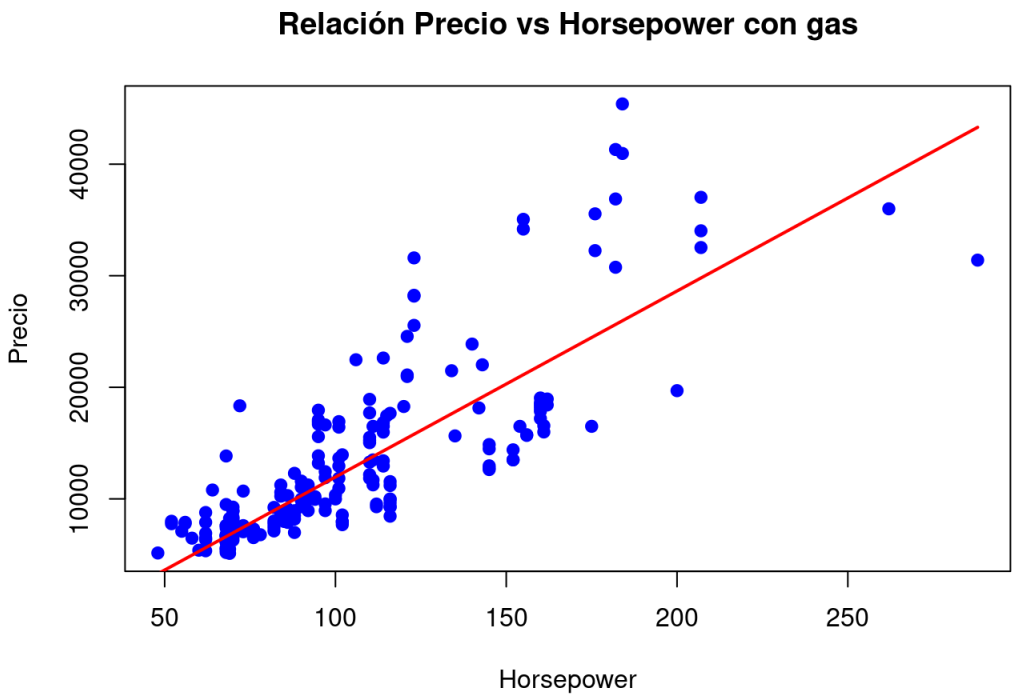
El modelo que considera los caballos de fuerza un motor a diesel explica un 87.17% de la variación del 10% datos.

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

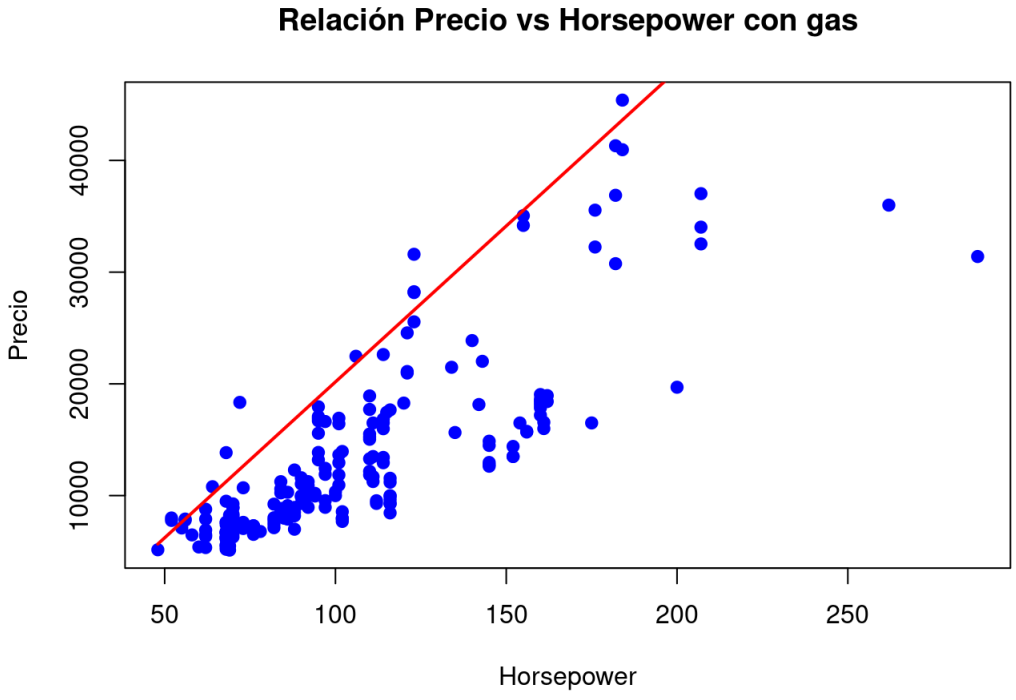
```
b0 = rl2$coefficients[1] # Beta 0
b1 = rl2$coefficients[2] # Beta 1

p = function(x){b0 + b1*x}

plot(data$horsepower, data$price, col = 'blue', pch = 19, ylab = "Precio", xlab = "Horsepower", main = "Relación Precio vs H
orsepower")
xx = seq(min(x$data.horsepower), max(x$data.horsepower), 0.01)
lines(xx, p(xx), col = 'red', lwd=2)
```



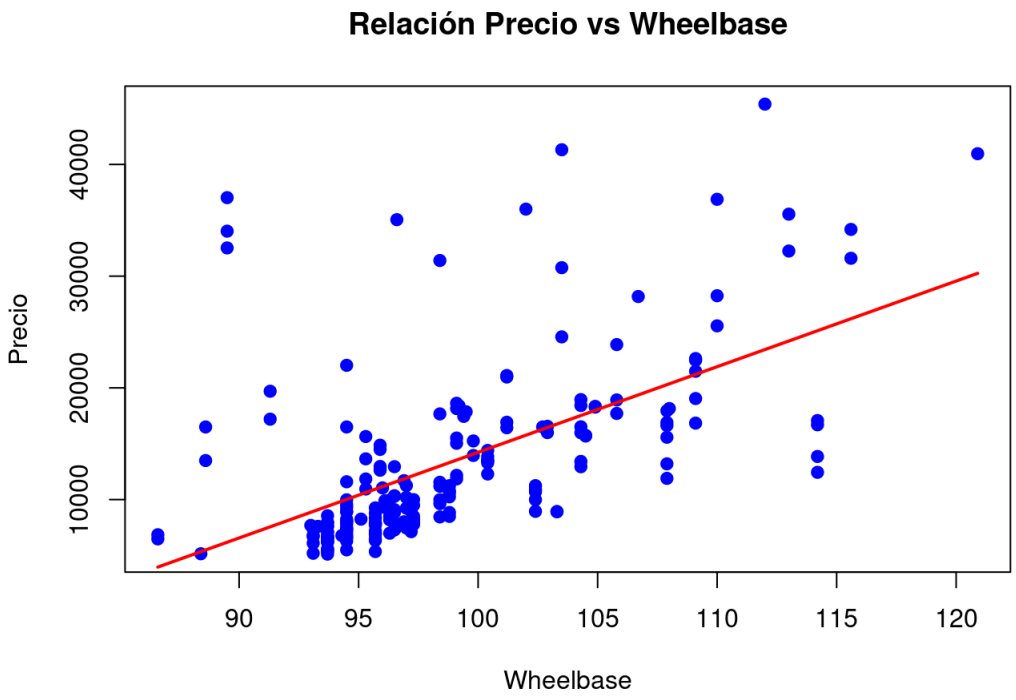

```
b0 = rl3$coefficients[1] # Beta 0
b1 = rl3$coefficients[2] # Beta 1

p = function(x){b0 + b1*x}

plot(x$data.horsepower, x$data.price, col = 'blue', pch = 19, ylab = "Precio", xlab = "Horsepower", main = "Relación Precio
vs Horsepower con gas")
xx = seq(min(x$data.horsepower), max(x$data.horsepower), 0.01)
lines(xx, p(xx), col = 'red', lwd=2)
```




```
b0 = rl4$coefficients[1] # Beta 0
b1 = rl4$coefficients[2] # Beta 1

p = function(x){b0 + b1*x}

plot(x$data.horsepower, x$data.price, col = 'blue', pch = 19, ylab = "Precio", xlab = "Horsepower", main = "Relación Precio
vs Horsepower con gas")
xx = seq(min(x$data.horsepower), max(x$data.horsepower), 0.01)
lines(xx, p(xx), col = 'red', lwd=2)
```



```
b0 = rl$coefficients[1] # Beta 0
b1 = rl$coefficients[2] # Beta 1

p = function(x){b0 + b1*x}

plot(x$data.wheelbase, x$data.price, col = 'blue', pch = 19, ylab = "Precio", xlab = "Wheelbase", main = "Relación Precio vs Wheelbase")
xx = seq(min(x$data.wheelbase), max(x$data.wheelbase), 0.01)
lines(xx, p(xx), col = 'red', lwd=2)
```



5. Interpreta en el contexto del problema cada uno de los análisis que hiciste.

Se analizaron cuatro modelos significativos para la determinación del precio de un auto; uno que considera los caballos de fuerza del motor, otro con la distancia entre los ejes del auto, y otros dos que consideran los caballos de fuerza del motor dependiendo de si es de gas o de diesel. Mediante el análisis de significancia de las variables en ambos modelos y la explicación de la variación de datos que ofrecen, podemos notar los mejores modelos son los que consideran los caballos de fuerza del motor, pues explican al menos el 65% de la variación del precio. Además, por las gráficas podemos observar que el modelo que considera el horsepower con motor a gas y el horsepower con ambos combustibles tienen rectas que se acomodan muy bien a los datos.

3. Analiza la validez de los modelos propuestos:

1. Normalidad de los residuos

Prueba de hipótesis:

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

Regla de decisión: $p - value < \alpha$ se rechaza H_0

```
library(nortest)
ad.test(rl2$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  rl2$residuals
## A = 4.8029, p-value = 6.267e-12
```

```
ad.test(rl3$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  rl3$residuals
## A = 4.426, p-value = 5.041e-11
```

```
ad.test(rl4$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  rl4$residuals
## A = 0.36324, p-value = 0.4058
```

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

```
##
##  Anderson-Darling normality test
##
## data:  rl$residuals
## A = 15.605, p-value < 2.2e-16
```

Dados los valores p de todos los modelos, sabemos con un 97% de confianza que el único modelo que no tiene evidencia suficiente para rechazar la hipótesis inicial es el modelo de horsepower con diesel. Es decir, el modelo que predice el precio considerando el horsepower en carros con diesel proviene de una población normal, y los otros modelos no.

2. Verificación de media cero

Prueba de hipótesis:

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

Regla de decisión: $p - value < \alpha$ se rechaza H_0

```
t.test(rl2$residuals)
```

```
##
##  One Sample t-test
##
## data:  rl2$residuals
## t = 8.0373e-17, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -647.9614  647.9614
## sample estimates:
##    mean of x
## 2.641356e-14
```

```
t.test(rl3$residuals)
```

```
##
##  One Sample t-test
##
## data:  rl3$residuals
## t = 2.4626e-16, df = 184, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -628.6496  628.6496
## sample estimates:
##    mean of x
## 7.846696e-14
```

```
t.test(rl4$residuals)
```

```
##
##  One Sample t-test
##
## data:  rl4$residuals
## t = 1.6097e-17, df = 19, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1300.993  1300.993
## sample estimates:
##    mean of x
## 1.000589e-14
```

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

```
##
##  One Sample t-test
##
## data:  rl$residuals
## t = -2.4209e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -897.8812  897.8812
## sample estimates:
##     mean of x
## -1.102446e-13
```

Dados los valores p de todos los modelos, observemos que ninguno tiene la suficiente evidencia para rechazar la hipótesis inicial, por lo que todos los modelos tienen errores con media 0.

3. Homocedasticidad, linealidad e independencia

Prueba de hipótesis para homocedasticidad:

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

Regla de decisión: $p - value < \alpha$ se rechaza H_0

Prueba de hipótesis para independencia:

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

Prueba de hipótesis para linealidad:

- H_0 : No hay términos omitidos que indican linealidad
- H_1 : Hay una especificación errónea en el modelo que indica no linealidad

Regla de decisión: $p - value < \alpha$ se rechaza H_0

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
dwtest(rl2) # Test de Durbin-Watson para Independencia
```

```
##
##  Durbin-Watson test
##
## data:  rl2
## DW = 0.79229, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(rl2) # Test de Breusch-Pagan para Homocedasticidad
```

```
##
##  studentized Breusch-Pagan test
##
## data:  rl2
## BP = 54.573, df = 1, p-value = 1.497e-13
```

```
resettest(rl2)
```

```
##
##  RESET test
##
## data:  rl2
## RESET = 5.1766, df1 = 2, df2 = 201, p-value = 0.006424
```

```
dwtest(rl3) # Test de Durbin-Watson para Independencia
```

```
##
##  Durbin-Watson test
##
## data:  rl3
## DW = 0.95227, p-value = 2.8e-13
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(rl3) # Test de Breusch-Pagan para Homocedasticidad
```

```
##
##  studentized Breusch-Pagan test
##
## data:  rl3
## BP = 56.272, df = 1, p-value = 6.309e-14
```

```
resettest(rl3)
```

```
##
##  RESET test
##
## data:  rl3
## RESET = 7.3283, df1 = 2, df2 = 181, p-value = 0.0008702
```

```
dwtest(rl4) # Test de Durbin-Watson para Independencia
```

```
##
##  Durbin-Watson test
##
## data:  rl4
## DW = 1.4141, p-value = 0.06718
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(rl4) # Test de Breusch-Pagan para Homocedasticidad
```

```
##
##  studentized Breusch-Pagan test
##
## data:  rl4
## BP = 0.79779, df = 1, p-value = 0.3718
```

```
resettest(rl4)
```

```
##
##  RESET test
##
## data:  rl4
## RESET = 4.019, df1 = 2, df2 = 16, p-value = 0.03853
```

```
dwtest(rl) # Test de Durbin-Watson para Independencia
```

```
##
##  Durbin-Watson test
##
## data:  rl
## DW = 0.56645, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(rl) # Test de Breusch-Pagan para Homocedasticidad
```

```
##
##  studentized Breusch-Pagan test
##
## data:  rl
## BP = 0.01538, df = 1, p-value = 0.9013
```

```
resettest(rl)
```

```
##
##  RESET test
##
## data:  rl
## RESET = 10.65, df1 = 2, df2 = 201, p-value = 4.016e-05
```

Dados los valores p de todos los modelos, sabemos con un 97% de confianza que:

- El modelo que considera el horsepower no contiene homocedasticidad, ni independencia, ni linealidad en sus errores.
- El modelo que considera el horsepower en carros a gas no contiene homocedasticidad, ni independencia, ni linealidad en sus errores.
- El modelo que considera el horsepower en carros a diesel contiene homocedasticidad, independencia y linealidad en sus errores.
- El modelo que considera la distancia entre los ejes del auto contiene homocedasticidad en sus errores, pero no independencia ni linealidad.

4. Interpreta cada uno de los analisis que realizaste

Aún cuando el modelo que considera el horsepower a diesel pasa la validez de los errores del modelo, los datos que considera constituyen únicamente un 10% de todos los datos, por lo que no es representativo y no generaliza adecuadamente. Y descartamos el modelo que considera wheelbase aún cuando este presenta homocedasticidad porque su explicación de la varianza del precio es muy baja.

Esto nos deja con dos modelos, que serían el que considera horsepower y el que considera horsepower en carros a gas, este último, a diferencia del modelo que considera el diesel, si es representativo, pues se constituye de un 90% de los datos del modelo, pero no termina de generalizar los datos de los carros a diesel. Ambos modelos tienen una muy buena explicación de la varianza del precio, y la recta de ajuste se "ajusta" a los datos originales. La desventaja de estos modelos es que dado que no presentan homocedasticidad, linealidad o independencia las inferencias estadísticas se vuelven menos confiables y las predicciones del modelo pierden precisión.

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

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

El mejor modelo encontrado es el que predice el precio de un auto de acuerdo a la capacidad de caballos de fuerza de un motor a gas, pues explica un 70% de la variación del 90% de los datos. Argumentando que es mejor dado que usa menos datos que el modelo que considera el horsepower con ambos tipos de combustible y sigue dando una buena explicación de los datos.

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

Tanto la capacidad de caballos de fuerza del motor, como el tipo de combustible y la distancia entre los ejes del auto, todos influyen en el precio final del auto. La variable de caballos de fuerza influye mucho más que el tipo de combustible y la distancia entre ejes, pues tiene sentido que una mayor capacidad conduzca a más trabajo y gasto de materiales para el proveedor, resultando en un precio más alto. La distancia entre ejes influye a menor medida en el precio, pues solo es una manera de decidir el largo del carro, resultando en un gran incremento del precio, pues los demás componentes siguen siendo los mismo. Y por último, el tipo de combustible influye debido a su influencia en el motor, pues los diferentes tipos requieren de diferentes procesos para la combustión interna.

Ahora, dado el modelo que se realizó para predecir el precio, diríamos que las variables que influyen de mayor manera en el precio son la capacidad de caballos de fuerza del motor y el hecho de que el motor funcione con gas.

3. 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:

1. Calcula los intervalos para la variable Y

```
Ip=predict(object=rl3,interval="prediction",level=0.97)
```

```
## Warning in predict.lm(object = rl3, interval = "prediction", level = 0.97): predictions on current data refer to _future_
responses
```

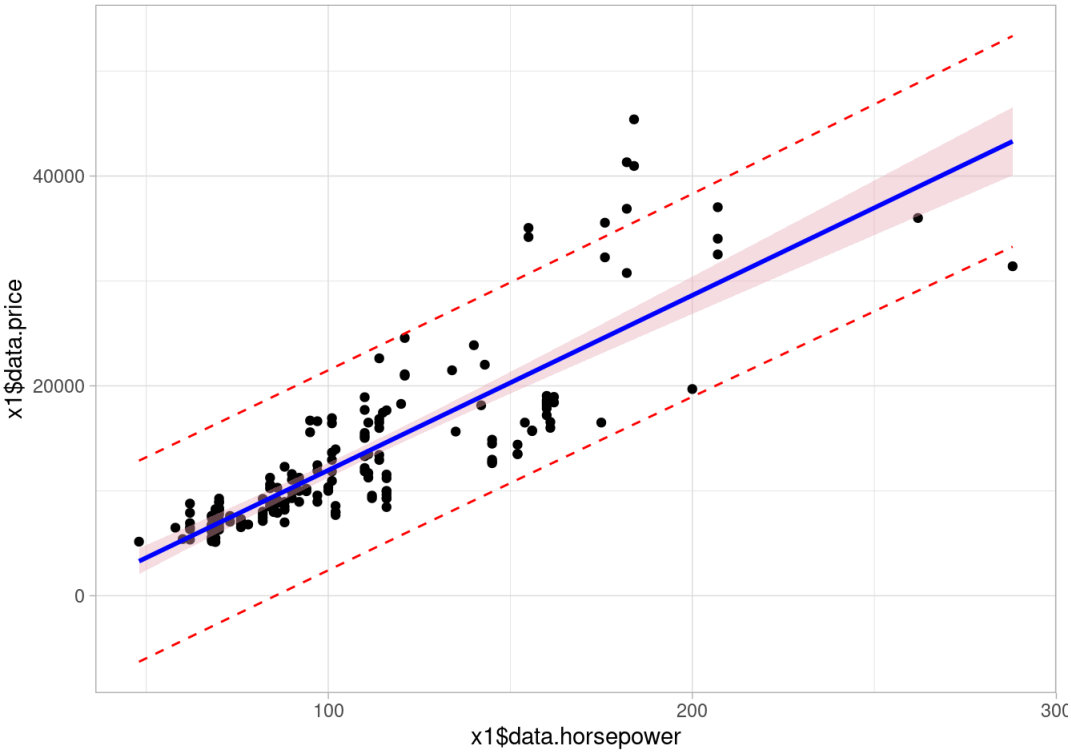
```
datos=cbind(x1,Ip)
```

2. 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.

La categoría más importante de la variable "fueltype" o "tipo de combustible" es "gas". Es decir, es importante que el auto sea de combustible por gas, pues es la predominancia de los datos.

3. Grafica por pares de variables numéricas

```
ggplot(datos,aes(x=x1$data.horsepower, y=x1$data.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()
```



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.

No se realizó con la categoría de combustible “diesel” porque este representa únicamente un 10% de los datos, y no sería representativo.

3. Interpreta en el contexto del problema

En la gráfica se observa la línea azul central que es la línea de regresión lineal que mejor ajusta los datos, la banda rojo claro alrededor de la línea de tendencia representa el intervalo de confianza y las líneas punteadas rojas representan los límites de predicción, que muestran el rango dentro del cual se espera que caigan los valores individuales del peso para una estatura dada.

La gráfica muestra que existe una tendencia de que a mayor caballos de fuerza de capacidad tenga el motor, mayor será el precio del auto, con cierta variabilidad alrededor de esta tendencia. Además de ciertos datos fuera del intervalo de predicción, dando a entender que el modelo presenta áreas de mejora, ya sea incluyendo más variables, o que estos solo sean casos extremos.

4. 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?

Propondría una nueva agrupación de variables que sigan incluyendo horsepower, para poder explicar el 35% de variación restante según el modelo de regresión con horsepower, sin tomar en cuenta el hecho de que no se cumplen los supuestos de normalidad de errores del modelo.

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

```
summary.data.frame(data)
```

```
##    symboling         CarName            fueltype           carbody
##  Min.   :-2.0000   Length:205         Length:205         Length:205
##  1st Qu.: 0.0000   Class :character   Class :character   Class :character
##  Median : 1.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.8341
##  3rd Qu.: 2.0000
##  Max.   : 3.0000
##   drivewheel        enginelocation       wheelbase        carlength
##  Length:205         Length:205         Min.   : 86.60   Min.   :141.1
##  Class :character   Class :character   1st Qu.: 94.50   1st Qu.:166.3
##  Mode  :character   Mode  :character   Median : 97.00   Median :173.2
##                                        Mean   : 98.76   Mean   :174.0
##                                        3rd Qu.:102.40   3rd Qu.:183.1
##                                        Max.   :120.90   Max.   :208.1
##     carwidth       carheight       curbweight    enginetype
##  Min.   :60.30   Min.   :47.80   Min.   :1488   Length:205
##  1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145   Class :character
##  Median :65.50   Median :54.10   Median :2414   Mode  :character
##  Mean   :65.91   Mean   :53.72   Mean   :2556
##  3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935
##  Max.   :72.30   Max.   :59.80   Max.   :4066
##  cylindernumber       enginesize        stroke      compressionratio
##  Length:205         Min.   : 61.0   Min.   :2.070   Min.   : 7.00
##  Class :character   1st Qu.: 97.0   1st Qu.:3.110   1st Qu.: 8.60
##  Mode  :character   Median :120.0   Median :3.290   Median : 9.00
##                     Mean   :126.9   Mean   :3.255   Mean   :10.14
##                     3rd Qu.:141.0   3rd Qu.:3.410   3rd Qu.: 9.40
##                     Max.   :326.0   Max.   :4.170   Max.   :23.00
##    horsepower       peakrpm        citympg        highwaympg        price
##  Min.   : 48.0   Min.   :4150   Min.   :13.00   Min.   :16.00   Min.   : 5118
##  1st Qu.: 70.0   1st Qu.:4800   1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7788
##  Median : 95.0   Median :5200   Median :24.00   Median :30.00   Median :10295
##  Mean   :104.1   Mean   :5125   Mean   :25.22   Mean   :30.75   Mean   :13277
##  3rd Qu.:116.0   3rd Qu.:5500   3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16503
##  Max.   :288.0   Max.   :6600   Max.   :49.00   Max.   :54.00   Max.   :45400
```

Nos enfocaremos en la correlación de las variables numéricas, dado que estas son más fáciles de manejar, y también dan suficiente información según nuestra experiencia con la variable “Horsepower”.

```
new = data[, sapply(data, function(x) !is.factor(x) & !is.character(x))]
cor(new)
```

```
##                    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
```

Según el análisis de correlación de las variables numéricas de la base de datos, las variables que más importancia tienen para decidir el precio de un carro son wheelbase (distancia entre los ejes), carwidth (ancho del carro), carlength (largo del carro), curbweight (peso del carro), enginesize (tamaño del carro), citympg (kilometraje en ciudad), highwaympg (kilometraje en carretera) y horsepower (caballos de fuerza del motor).