Actividad Integradora 2

Jacobo Hirsch Rodriguez

2024-09-06

leemos el archivo csv

```
A=read.csv("./precios_autos.csv") #leer la base de datos
```

```
str(A)
```

```
## 'data.frame':    205 obs. of  21 variables:
##  $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ CarName         : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa-romero
Quadrifoglio" "audi 100 ls" ...
##  $ fueltype        : chr  "gas" "gas" "gas" "gas" ...
##  $ carbody         : chr  "convertible" "convertible" "hatchback" "sedan" ...
##  $ drivewheel      : chr  "rwd" "rwd" "rwd" "fwd" ...
##  $ enginelocation  : chr  "front" "front" "front" "front" ...
##  $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ carlength       : num  169 169 171 177 177 ...
##  $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curbweight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
##  $ enginetype      : chr  "dohc" "dohc" "ohcv" "ohc" ...
##  $ cylindernumber  : chr  "four" "four" "six" "four" ...
##  $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ peakrpm         : int  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
##  $ citympg         : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highwaympg      : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price           : num  13495 16500 16500 13950 17450 ...
```

A mi me toco analizar el problema con las siguientes variables : Distancia entre los ejes (wheelbase), tipo de gasolina que usa (fueltype) y caballos de fuerza (horsepower)

convetimos la unica variable que no es numerica a numerica para analizarla

```
A$fueltype_numeric <- ifelse(A$fueltype == "gas", 1, 0)
```

```
new_data = data.frame(A$wheelbase, A$fueltype_numeric, A$horsepower, A$price)

str(new_data)
```

```
## 'data.frame':    205 obs. of  4 variables:
##  $ A.wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ A.fueltype_numeric: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ A.horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ A.price           : num  13495 16500 16500 13950 17450 ...
```

#Exploración de la base de datos

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

```
medias <- colMeans(new_data)
desviaciones <- apply(new_data, 2, sd)
minimos <- apply(new_data, 2, min)
maximos <- apply(new_data, 2, max)

medidas <- data.frame(Medias = medias, Desviaciones = desviaciones, Minimos = minimo
s, Maximos = maximos)
print(medidas)
```

```
##                         Medias Desviaciones Minimos Maximos
## A.wheelbase          98.756585    6.0217757    86.6   120.9
## A.fueltype_numeric    0.902439    0.2974465     0.0     1.0
## A.horsepower        104.117073   39.5441668    48.0   288.0
## A.price           13276.710571 7988.8523317  5118.0 45400.0
```

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

```
correlacion <- cor(new_data)
print(new_data)
```

```
##     A.wheelbase A.fueltype_numeric A.horsepower  A.price
## 1          88.6                  1          111 13495.00
## 2          88.6                  1          111 16500.00
## 3          94.5                  1          154 16500.00
## 4          99.8                  1          102 13950.00
## 5          99.4                  1          115 17450.00
## 6          99.8                  1          110 15250.00
## 7         105.8                  1          110 17710.00
## 8         105.8                  1          110 18920.00
```

```
## 9         105.8                  1          140 23875.00
## 10         99.5                  1          160 17859.17
## 11        101.2                  1          101 16430.00
## 12        101.2                  1          101 16925.00
## 13        101.2                  1          121 20970.00
## 14        101.2                  1          121 21105.00
## 15        103.5                  1          121 24565.00
## 16        103.5                  1          182 30760.00
## 17        103.5                  1          182 41315.00
## 18        110.0                  1          182 36880.00
## 19         88.4                  1           48  5151.00
## 20         94.5                  1           70  6295.00
## 21         94.5                  1           70  6575.00
## 22         93.7                  1           68  5572.00
## 23         93.7                  1           68  6377.00
## 24         93.7                  1          102  7957.00
## 25         93.7                  1           68  6229.00
## 26         93.7                  1           68  6692.00
## 27         93.7                  1           68  7609.00
## 28         93.7                  1          102  8558.00
## 29        103.3                  1           88  8921.00
## 30         95.9                  1          145 12964.00
## 31         86.6                  1           58  6479.00
## 32         86.6                  1           76  6855.00
## 33         93.7                  1           60  5399.00
## 34         93.7                  1           76  6529.00
## 35         93.7                  1           76  7129.00
## 36         96.5                  1           76  7295.00
## 37         96.5                  1           76  7295.00
## 38         96.5                  1           86  7895.00
## 39         96.5                  1           86  9095.00
## 40         96.5                  1           86  8845.00
## 41         96.5                  1           86 10295.00
## 42         96.5                  1          101 12945.00
## 43         96.5                  1          100 10345.00
## 44         94.3                  1           78  6785.00
## 45         94.5                  1           70  8916.50
## 46         94.5                  1           70  8916.50
## 47         96.0                  1           90 11048.00
## 48        113.0                  1          176 32250.00
## 49        113.0                  1          176 35550.00
## 50        102.0                  1          262 36000.00
## 51         93.1                  1           68  5195.00
## 52         93.1                  1           68  6095.00
## 53         93.1                  1           68  6795.00
## 54         93.1                  1           68  6695.00
## 55         93.1                  1           68  7395.00
```

```
## 56         95.3                  1          101 10945.00
## 57         95.3                  1          101 11845.00
## 58         95.3                  1          101 13645.00
## 59         95.3                  1          135 15645.00
## 60         98.8                  1           84  8845.00
## 61         98.8                  1           84  8495.00
## 62         98.8                  1           84 10595.00
## 63         98.8                  1           84 10245.00
## 64         98.8                  0           64 10795.00
## 65         98.8                  1           84 11245.00
## 66        104.9                  1          120 18280.00
## 67        104.9                  0           72 18344.00
## 68        110.0                  0          123 25552.00
## 69        110.0                  0          123 28248.00
## 70        106.7                  0          123 28176.00
## 71        115.6                  0          123 31600.00
## 72        115.6                  1          155 34184.00
## 73         96.6                  1          155 35056.00
## 74        120.9                  1          184 40960.00
## 75        112.0                  1          184 45400.00
## 76        102.7                  1          175 16503.00
## 77         93.7                  1           68  5389.00
## 78         93.7                  1           68  6189.00
## 79         93.7                  1           68  6669.00
## 80         93.0                  1          102  7689.00
## 81         96.3                  1          116  9959.00
## 82         96.3                  1           88  8499.00
## 83         95.9                  1          145 12629.00
## 84         95.9                  1          145 14869.00
## 85         95.9                  1          145 14489.00
## 86         96.3                  1           88  6989.00
## 87         96.3                  1           88  8189.00
## 88         96.3                  1          116  9279.00
## 89         96.3                  1          116  9279.00
## 90         94.5                  1           69  5499.00
## 91         94.5                  0           55  7099.00
## 92         94.5                  1           69  6649.00
## 93         94.5                  1           69  6849.00
## 94         94.5                  1           69  7349.00
## 95         94.5                  1           69  7299.00
## 96         94.5                  1           69  7799.00
## 97         94.5                  1           69  7499.00
## 98         94.5                  1           69  7999.00
## 99         95.1                  1           69  8249.00
## 100        97.2                  1           97  8949.00
## 101        97.2                  1           97  9549.00
## 102       100.4                  1          152 13499.00
```

```
## 103       100.4                 1           152 14399.00
## 104       100.4                 1           152 13499.00
## 105        91.3                 1           160 17199.00
## 106        91.3                 1           200 19699.00
## 107        99.2                 1           160 18399.00
## 108       107.9                 1            97 11900.00
## 109       107.9                 0            95 13200.00
## 110       114.2                 1            97 12440.00
## 111       114.2                 0            95 13860.00
## 112       107.9                 1            95 15580.00
## 113       107.9                 0            95 16900.00
## 114       114.2                 1            95 16695.00
## 115       114.2                 0            95 17075.00
## 116       107.9                 1            97 16630.00
## 117       107.9                 0            95 17950.00
## 118       108.0                 1           142 18150.00
## 119        93.7                 1            68  5572.00
## 120        93.7                 1           102  7957.00
## 121        93.7                 1            68  6229.00
## 122        93.7                 1            68  6692.00
## 123        93.7                 1            68  7609.00
## 124       103.3                 1            88  8921.00
## 125        95.9                 1           145 12764.00
## 126        94.5                 1           143 22018.00
## 127        89.5                 1           207 32528.00
## 128        89.5                 1           207 34028.00
## 129        89.5                 1           207 37028.00
## 130        98.4                 1           288 31400.50
## 131        96.1                 1            90  9295.00
## 132        96.1                 1            90  9895.00
## 133        99.1                 1           110 11850.00
## 134        99.1                 1           110 12170.00
## 135        99.1                 1           110 15040.00
## 136        99.1                 1           110 15510.00
## 137        99.1                 1           160 18150.00
## 138        99.1                 1           160 18620.00
## 139        93.7                 1            69  5118.00
## 140        93.7                 1            73  7053.00
## 141        93.3                 1            73  7603.00
## 142        97.2                 1            82  7126.00
## 143        97.2                 1            82  7775.00
## 144        97.2                 1            94  9960.00
## 145        97.0                 1            82  9233.00
## 146        97.0                 1           111 11259.00
## 147        97.0                 1            82  7463.00
## 148        97.0                 1            94 10198.00
## 149        96.9                 1            82  8013.00
```

```
## 150        96.9                  1           111 11694.00
## 151        95.7                  1            62  5348.00
## 152        95.7                  1            62  6338.00
## 153        95.7                  1            62  6488.00
## 154        95.7                  1            62  6918.00
## 155        95.7                  1            62  7898.00
## 156        95.7                  1            62  8778.00
## 157        95.7                  1            70  6938.00
## 158        95.7                  1            70  7198.00
## 159        95.7                  0            56  7898.00
## 160        95.7                  0            56  7788.00
## 161        95.7                  1            70  7738.00
## 162        95.7                  1            70  8358.00
## 163        95.7                  1            70  9258.00
## 164        94.5                  1            70  8058.00
## 165        94.5                  1            70  8238.00
## 166        94.5                  1           112  9298.00
## 167        94.5                  1           112  9538.00
## 168        98.4                  1           116  8449.00
## 169        98.4                  1           116  9639.00
## 170        98.4                  1           116  9989.00
## 171        98.4                  1           116 11199.00
## 172        98.4                  1           116 11549.00
## 173        98.4                  1           116 17669.00
## 174       102.4                  1            92  8948.00
## 175       102.4                  0            73 10698.00
## 176       102.4                  1            92  9988.00
## 177       102.4                  1            92 10898.00
## 178       102.4                  1            92 11248.00
## 179       102.9                  1           161 16558.00
## 180       102.9                  1           161 15998.00
## 181       104.5                  1           156 15690.00
## 182       104.5                  1           156 15750.00
## 183        97.3                  0            52  7775.00
## 184        97.3                  1            85  7975.00
## 185        97.3                  0            52  7995.00
## 186        97.3                  1            85  8195.00
## 187        97.3                  1            85  8495.00
## 188        97.3                  0            68  9495.00
## 189        97.3                  1           100  9995.00
## 190        94.5                  1            90 11595.00
## 191        94.5                  1            90  9980.00
## 192       100.4                  1           110 13295.00
## 193       100.4                  0            68 13845.00
## 194       100.4                  1            88 12290.00
## 195       104.3                  1           114 12940.00
## 196       104.3                  1           114 13415.00
```

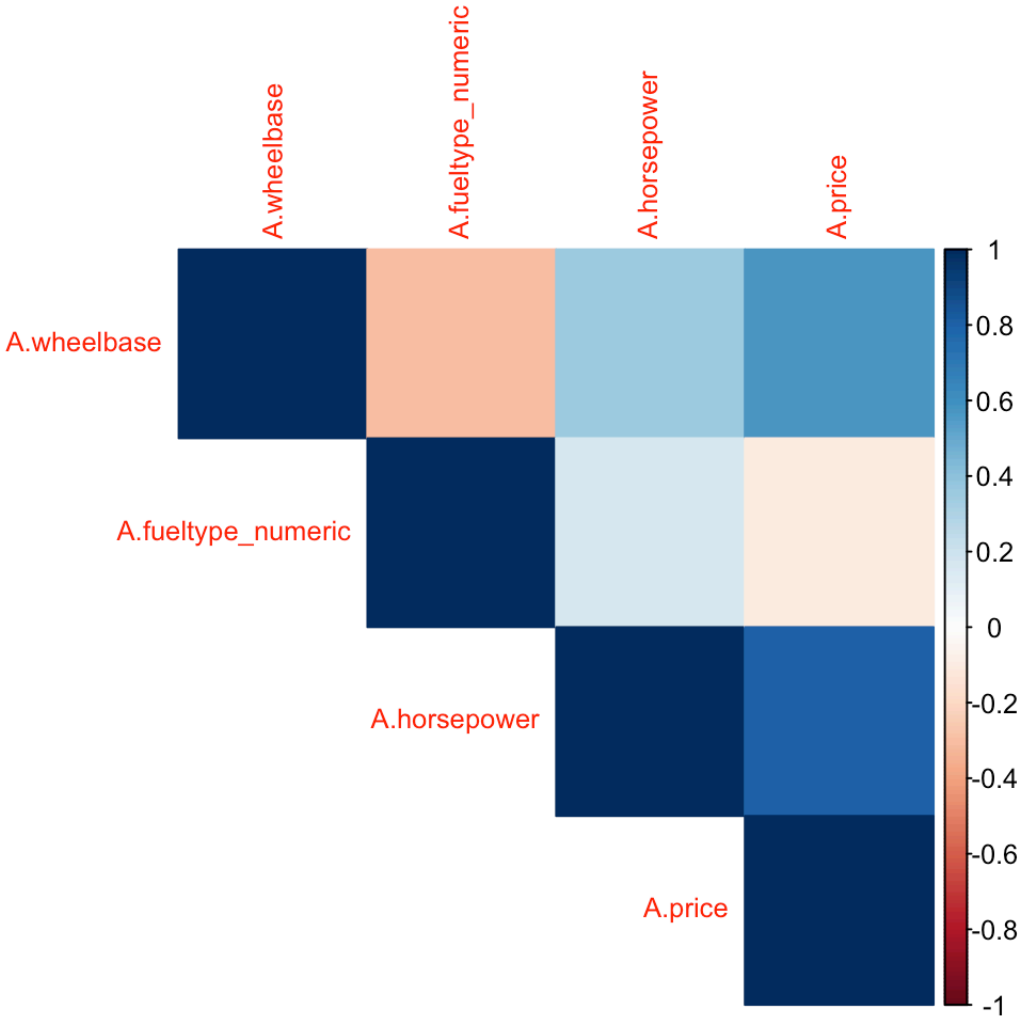
```
## 197       104.3                   1          114 15985.00
## 198       104.3                   1          114 16515.00
## 199       104.3                   1          162 18420.00
## 200       104.3                   1          162 18950.00
## 201       109.1                   1          114 16845.00
## 202       109.1                   1          160 19045.00
## 203       109.1                   1          134 21485.00
## 204       109.1                   0          106 22470.00
## 205       109.1                   1          114 22625.00
```

visualizamos el resultado de la matriz de correlacion con la liibreria corrplot

```
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
# Visualizar la matriz de correlación con un heatmap
corrplot(correlacion, method = "color", type = "upper", tl.cex = 0.8)
```



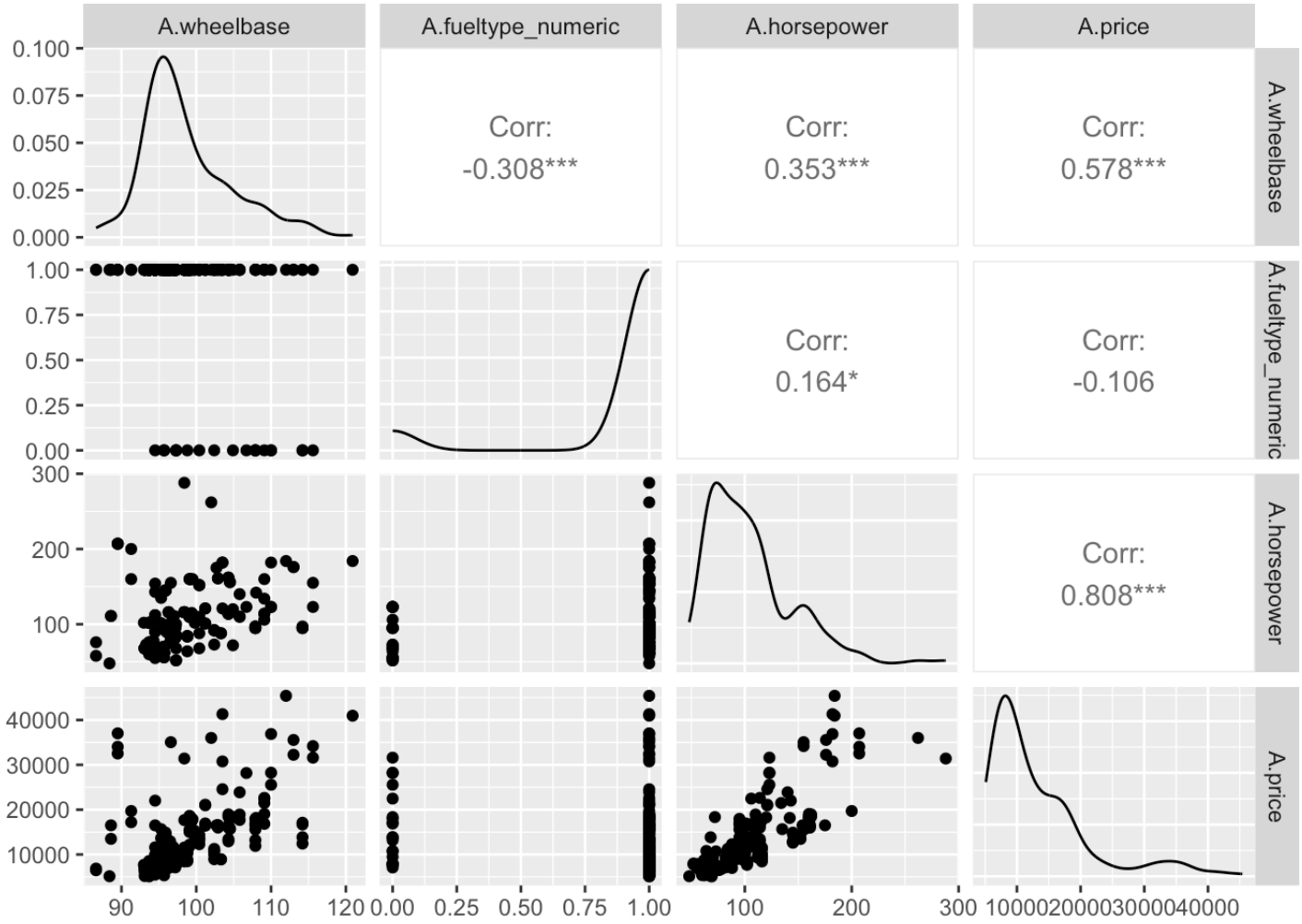
de esta grafica podemos observar que la variable que más relacion tiene (sin si misma) es la de precio con horsepower, seguido del precio con wheelbase, lo cual es un buen indicio para poder predecir el precio de los automoviles, tenemos dos variables que nos pueden ser de utilidad para una regresión.

```
# Instalar y cargar GGally si no lo tienes

library(GGally)
```

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

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

```
# Generar el gráfico de pares para ver la relación entre múltiples variables
ggpairs(new_data)
```



el siguiente gráfico confirma las afirmaciones anteriores. en donde la correlacion de pearson másy fuerte de precio con las demás variables es horsepower y wheelbase, en los graficos de dispersión se puede observar una tendencia lineal con estas dos variables.

#Modelación y verificación del modelo

para hacer nuestro modelo de regresión lineal, únicamente voy a utilizar las dos variables que pueden explicar de la mejor forma el precio

para ello voy a crear un nuevo dataframe que contenga únicamente estas variables (en realidad no es necesario pero me gusta y se me hace más limpio :)), antes haré unas variables aparte para meterlas al dataframe

```
wheelbase = A$wheelbase

horsepower = A$horsepower

price = A$price
```

```
ultimate_data = data.frame(wheelbase, horsepower, price)
```

modelo 1 (sin interacción):

```
modelo_sin_interaccion_act <- lm(price ~ wheelbase + horsepower, data = ultimate_dat
a)
```

ahora vamos a ver los resultados:

```
summary(modelo_sin_interaccion_act)
```

```
##
## Call:
## lm(formula = price ~ wheelbase + horsepower, data = ultimate_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 ***
## wheelbase      443.095     49.818   8.894 3.33e-16 ***
## horsepower     139.425      7.586  18.379  < 2e-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
```

modelo 2 (con interacción):

```
modelo_con_interaccion_act <- lm(price ~ wheelbase * horsepower, data = ultimate_dat
a)
```

ahora vamos a ver los resultados:

```
summary(modelo_con_interaccion_act)
```

```
##
## Call:
## lm(formula = price ~ wheelbase * horsepower, data = ultimate_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8847  -2050   -177   1350  15889
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -17059.574  14377.287  -1.187   0.2368
## wheelbase               155.900    148.256   1.052   0.2943
## horsepower              -89.721    111.777  -0.803   0.4231
## wheelbase:horsepower      2.342      1.140   2.055   0.0412 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3977 on 201 degrees of freedom
## Multiple R-squared:  0.7558, Adjusted R-squared:  0.7522
## F-statistic: 207.4 on 3 and 201 DF,  p-value: < 2.2e-16
```

modelo 3 horsepower:

```
modelo_con_horsepower <- lm(price ~ horsepower, data = ultimate_data)
```

ahora vamos a ver los resultados:

```
summary(modelo_con_horsepower)
```

```
##
## Call:
## lm(formula = price ~ horsepower, data = ultimate_data)
##
## 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 ***
## 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 4 wheelbase:

```
modelo_con_wheelbase <- lm(price ~ wheelbase, data = ultimate_data)
```

ahora vamos a ver los resultados:

```
summary(modelo_con_horsepower)
```

```
##
## Call:
## lm(formula = price ~ horsepower, data = ultimate_data)
##
## 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 ***
## 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
```

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)

hipotesis nula: el modelo no es signifivativo

hipotesis alternativa : el modelo es significativo

regla de decision : si el valor p es menor o igual a alfa p <= 0.04, el modelo es estadisticamente significativo y rechazas la hipotesis nula

modelo 1) sin interacción:

```
# Verificar la significancia del modelo completo
prueba_a_1 <- summary(modelo_sin_interaccion_act)$coefficients[2,4] < 0.04
if (prueba_a_1) {
  print("El modelo es significativo con un nivel de alfa = 0.04")
} else {
  print("El modelo no es significativo con un nivel de alfa = 0.04")
}
```

```
## [1] "El modelo es significativo con un nivel de alfa = 0.04"
```

modelo 2) con interaccion:

```
# Verificar la significancia del modelo completo
prueba_a_2 <- summary(modelo_con_interaccion_act)$coefficients[2,4] < 0.04
if (prueba_a_2) {
  print("El modelo es significativo con un nivel de alfa = 0.04")
} else {
  print("El modelo no es significativo con un nivel de alfa = 0.04")
}
```

```
## [1] "El modelo no es significativo con un nivel de alfa = 0.04"
```

modelo 3) con horsepower:

```
# Verificar la significancia del modelo completo
prueba_a_3 <- summary(modelo_con_horsepower)$coefficients[2,4] < 0.04
if (prueba_a_3) {
  print("El modelo es significativo con un nivel de alfa = 0.04")
} else {
  print("El modelo no es significativo con un nivel de alfa = 0.04")
}
```

```
## [1] "El modelo es significativo con un nivel de alfa = 0.04"
```

modelo 4) con wheelbase:

```
# Verificar la significancia del modelo completo
prueba_a_4 <- summary(modelo_con_wheelbase)$coefficients[2,4] < 0.04
if (prueba_a_4) {
  print("El modelo es significativo con un nivel de alfa = 0.04")
} else {
  print("El modelo no es significativo con un nivel de alfa = 0.04")
}
```

```
## [1] "El modelo es significativo con un nivel de alfa = 0.04"
```

##Valida la significancia de $\hat{\beta}i$ con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera de cada una de ellas)

hipotesis nula = los coeficientes del modelo no son significativos y por ende son iguales a 0

hipotesis alternativa = los coeficientes del modelo no son signfificativos y por ende son diferentes de 0

regla de decisión = si el valor p

modelo 1) sin interacción

```
# Verificar la significancia de los coeficientes
coef_significativos_1 <- summary(modelo_sin_interaccion_act)$coefficients[,4] < 0.04
print(coef_significativos_1)
```

```
## (Intercept)   wheelbase  horsepower
##        TRUE        TRUE        TRUE
```

para este modelo muestra que todos los coeficientes son estadisticamente significativos

modelo 2) con interacción

```
# Verificar la significancia de los coeficientes
coef_significativos_2 <- summary(modelo_con_interaccion_act)$coefficients[,4] < 0.04
print(coef_significativos_2)
```

```
##          (Intercept)            wheelbase           horsepower
##                FALSE                FALSE                FALSE
## wheelbase:horsepower
##                FALSE
```

para este modelo muestra que ninguno de los coeficientes son estadisticamente significativos

modelo 3) con horsepower

```
# Verificar la significancia de los coeficientes
coef_significativos_3 <- summary(modelo_con_horsepower)$coefficients[,4] < 0.04
print(coef_significativos_3)
```

```
## (Intercept)  horsepower
##        TRUE        TRUE
```

para este modelo ambos coeficientes son significativos

modelo 4) con wheelbase

```
# Verificar la significancia de los coeficientes
coef_significativos_4 <- summary(modelo_con_wheelbase)$coefficients[,4] < 0.04
print(coef_significativos_4)
```

```
## (Intercept)   wheelbase
##        TRUE        TRUE
```

para este modelo ambos coeficientes fueron significativos

##Indica cuál es el porcentaje de variación explicada por el modelo. modelo 1) sin interaccion:

```
# R-cuadrado del modelo
r_cuadrado_act_1 <- summary(modelo_sin_interaccion_act)$r.squared
print(paste("El porcentaje de variación explicada por el modelo es:", r_cuadrado_act_
1 * 100, "%"))
```

```
## [1] "El porcentaje de variación explicada por el modelo es: 75.0715923086211 %"
```

modelo 2) con interaccion:

```
# R-cuadrado del modelo
r_cuadrado_act_2 <- summary(modelo_con_interaccion_act)$r.squared
print(paste("El porcentaje de variación explicada por el modelo es:", r_cuadrado_act_
2 * 100, "%"))
```

```
## [1] "El porcentaje de variación explicada por el modelo es: 75.5844106194618 %"
```

modelo 3) con horsepower:

```
# R-cuadrado del modelo
r_cuadrado_act_3 <- summary(modelo_con_horsepower)$r.squared
print(paste("El porcentaje de variación explicada por el modelo es:", r_cuadrado_act_
3 * 100, "%"))
```

```
## [1] "El porcentaje de variación explicada por el modelo es: 65.3088356490231 %"
```

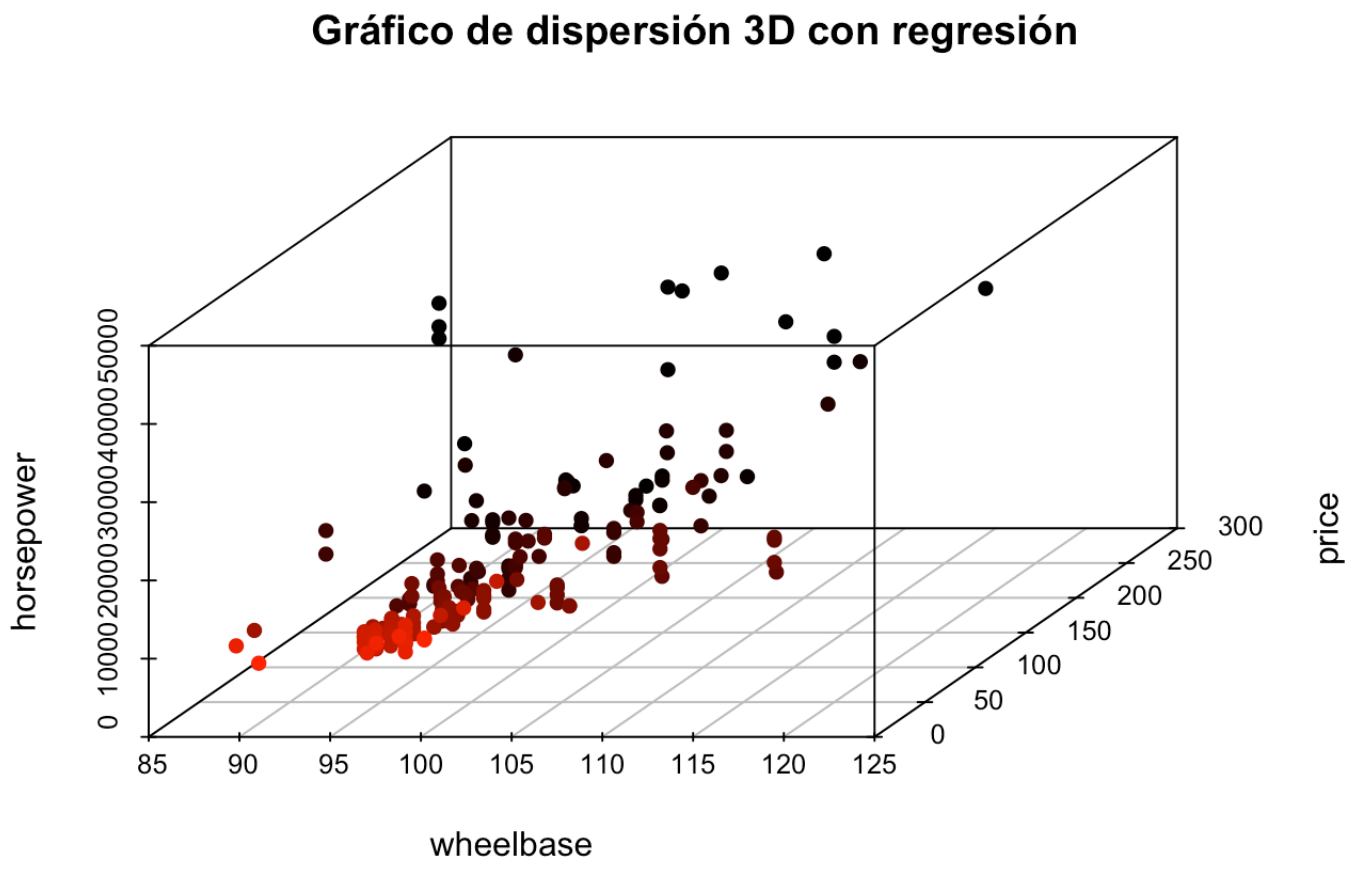
modelo 4) con wheelbase

```
# R-cuadrado del modelo
r_cuadrado_act_4 <- summary(modelo_con_wheelbase)$r.squared
print(paste("El porcentaje de variación explicada por el modelo es:", r_cuadrado_act_
4 * 100, "%"))
```

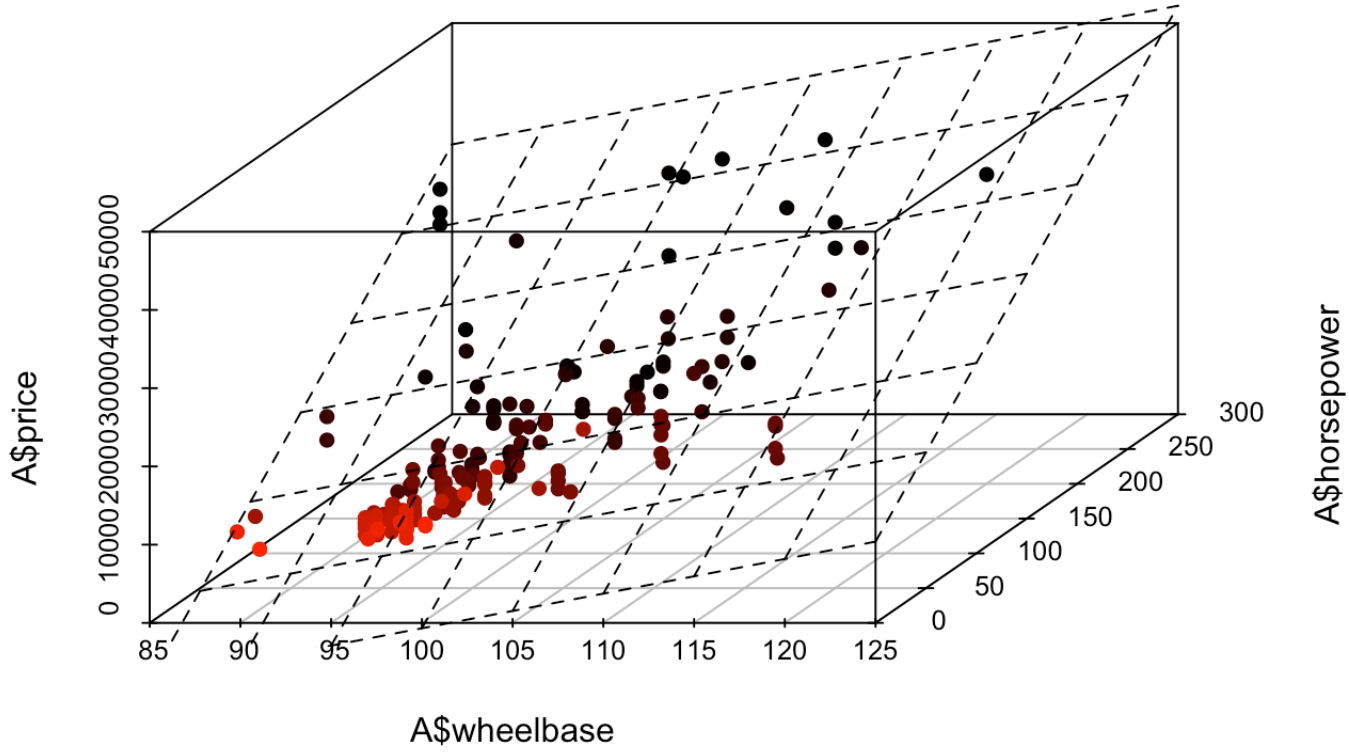
```
## [1] "El porcentaje de variación explicada por el modelo es: 33.3870865629716 %"
```

##Dibuja el diagrama de dispersión de los datos por pares y la recta de mejor ajuste. modelo 1) sin interaccion:

```
library(scatterplot3d)

# Crear el gráfico de dispersión 3D
scatterplot3d(A$wheelbase, A$horsepower, A$price, pch = 16, highlight.3d = TRUE,
              main = "Gráfico de dispersión 3D con regresión",
              xlab = "wheelbase", ylab = "price",
              zlab = "horsepower")
```

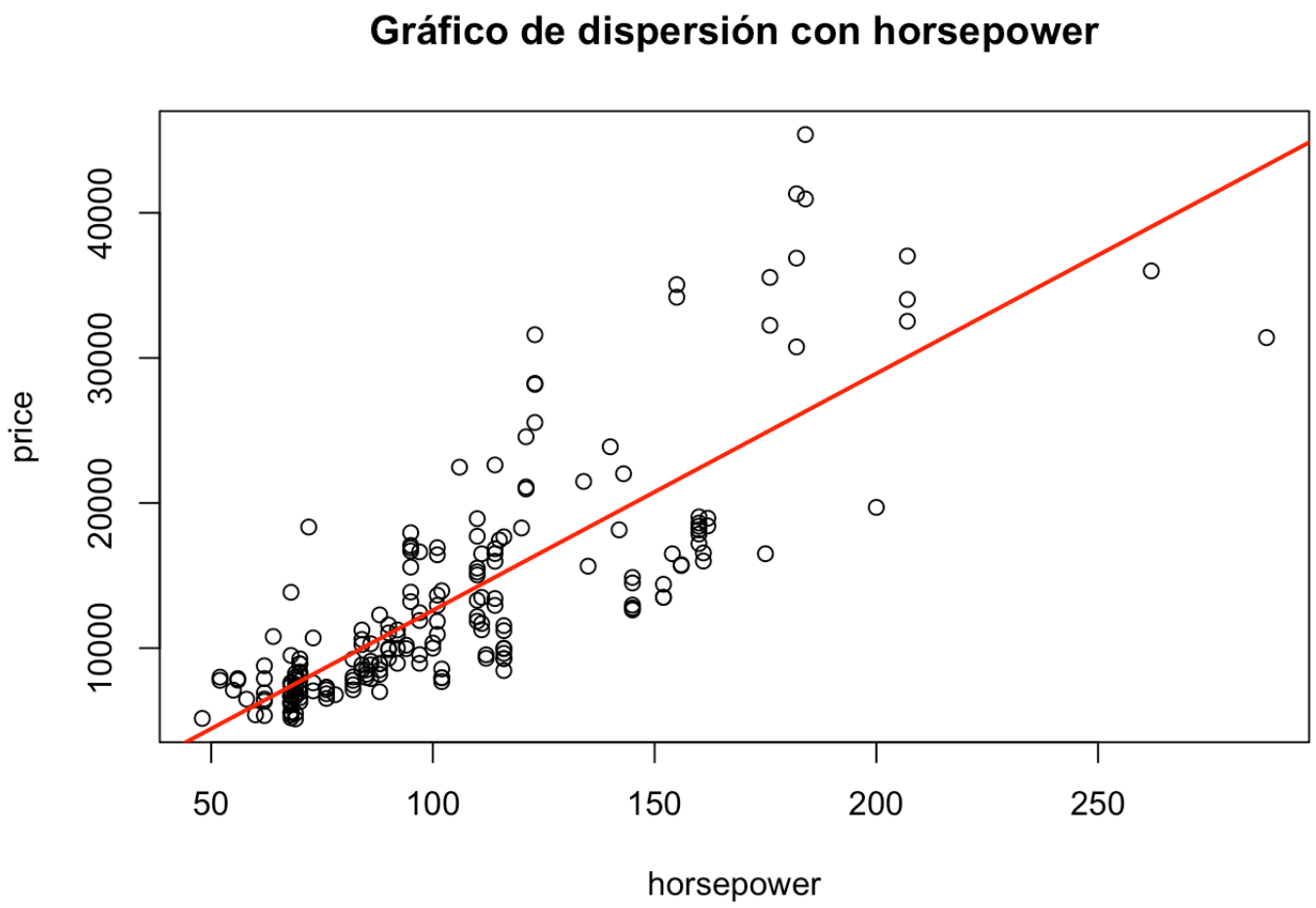



```
# Dibujar el plano de regresión
scatter_one <- scatterplot3d(A$wheelbase, A$horsepower, A$price, pch = 16, highlight.
3d = TRUE)
scatter_one$plane3d(modelo_sin_interaccion_act)
```



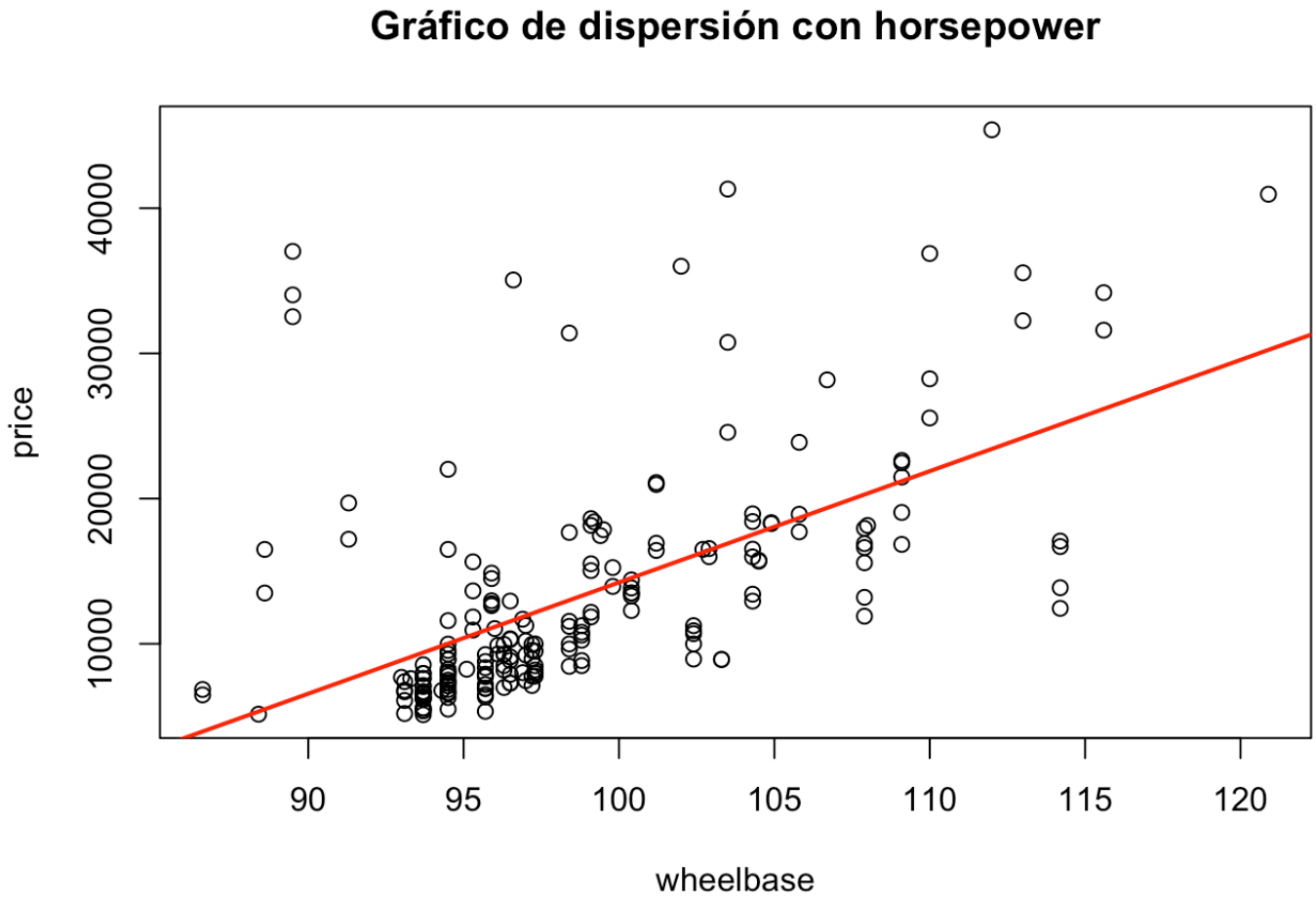
modelo 3) con horsepower:

```
# Crear un gráfico de dispersión y ajustar la recta de regresión
plot(ultimate_data$horsepower, A$price , main = "Gráfico de dispersión con horsepowe
r",
     xlab = "horsepower", ylab = "price")

# Agregar la línea de regresión al gráfico
abline(modelo_con_horsepower, col = "red", lwd = 2)
```



modelo 4) con wheelbase:

```
# Crear un gráfico de dispersión y ajustar la recta de regresión
plot(ultimate_data$wheelbase, A$price , main = "Gráfico de dispersión con horsepowe
r",
     xlab = "wheelbase", ylab = "price")

# Agregar la línea de regresión al gráfico
abline(modelo_con_wheelbase, col = "red", lwd = 2)
```



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

El único modelo que parece no ser significativo es el modelo que posee interacción, aunque curiosamente para este conjunto de datos parece ser junto con el modelo que no posee interacción el modelo que mayor porcentaje de variación explicada posee, aunque tampoco sus coeficientes sean significativos como pasa contrariamente con el modelo que no posee interacción. En general el modelo que mejor se comporto es el modelo que no posee interacción ya que el modelo es significativo con un alfa de 0.04%, los coeficientes parecen ser significativos y el porcentaje de variación explicada es de 75%. Los otros dos modelos de una variable dependiente poseen porcentajes muy bajos de viaración explicada, lo suficiente para ni siquiera tomarlos en cuenta para los analisis posteriores.

##Analiza la validez de los modelos propuestos:

primero escribimos los residuos

```
residuos_con_interaccion_act <- residuals(modelo_con_interaccion_act)
residuos_sin_interaccion_act <- residuals(modelo_sin_interaccion_act)
```

Normalidad de los residuos

hipotesis nula : los residuos siguen una distribución normal.

hipotesis alternativa : los residuos no siguen una distribución normal

regla de decisión : Si el valor p es menor a 0.05, se rechaza la hipótesis nula y se concluye que los residuos no siguen una distribución normal.

sacamos la prueba de shapiro-wilk

modelo 1) sin interaccion:

```
# Prueba de Shapiro-Wilk
shapiro.test(residuos_sin_interaccion_act)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos_sin_interaccion_act
## W = 0.93465, p-value = 5.933e-08
```

modelo 2) con interaccion:

```
# Prueba de Shapiro-Wilk
shapiro.test(residuos_con_interaccion_act)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos_con_interaccion_act
## W = 0.92229, p-value = 6.187e-09
```

Verificación de media cero

hipotesis nula : la media de los errores es igual a 0

hiporesis alrernativa : la media de los errores es diferente de 0

regla de decisión con prueba t : Si el valor p es menor que 0.05, se rechaza la hipótesis nula y se concluye que la media de los residuos es significativamente diferente de 0.

modelo 1) sin interaccion:

```
media_sin_interaccion_act = mean(residuos_sin_interaccion_act)
print(media_sin_interaccion_act)
```

```
## [1] 4.081635e-13
```

realizamos la prueba t para el modelo sin interacción

```
# Realizar la prueba t para verificar si la media es 0
prueba_t_act_2 <- t.test(residuos_sin_interaccion_act, mu = 0)

# Mostrar el resultado de la prueba
print(prueba_t_act_2)
```

```
##
##  One Sample t-test
##
## data:  residuos_sin_interaccion_act
## t = 1.4651e-15, 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
## 4.081635e-13
```

modelo 2) con interaccion:

```
media_con_interaccion_act = mean(residuos_con_interaccion_act)
print(media_con_interaccion_act)
```

```
## [1] 1.286602e-13
```

realizamos la prueba t para el modelo con interacción

```
# Realizar la prueba t para verificar si la media es 0
prueba_t_act_1 <- t.test(residuos_con_interaccion_act, mu = 0)

# Mostrar el resultado de la prueba
print(prueba_t_act_1)
```

```
##
##  One Sample t-test
##
## data:  residuos_con_interaccion_act
## t = 4.6666e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -543.5923  543.5923
## sample estimates:
##    mean of x
## 1.286602e-13
```

Homocedasticidad

hipotesis nula: la varianza de los errores es constante (hay homocedasticidad)

hipotesis alternativa : la varianza de los errores no es constante (hay heterocedasticidad)

regla de decisión: si el valor p es menor o igual que 0.05 entonces rechazamos la hipotesis nula

```
# Cargar el paquete lmtest
library(lmtest)
```

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

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

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

modelo 1) sin interaccion:

```
# Realizar la prueba de White
prueba_white_sin_interaccion_act <- bptest(modelo_sin_interaccion_act, ~ fitted(model
o_sin_interaccion_act) + I(fitted(modelo_sin_interaccion_act)^2))

# Mostrar los resultados de la prueba
print(prueba_white_sin_interaccion_act)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_sin_interaccion_act
## BP = 52.3, df = 2, p-value = 4.398e-12
```

vamos a verificar la hipotesis nula

```
# Extraer el valor p de la prueba
p_value_white_sin_interaccion_act <- prueba_white_sin_interaccion_act$p.value

# Comparar el valor p con el umbral de significancia de 0.05
if (p_value_white_sin_interaccion_act > 0.05) {
  print("No se rechaza la hipótesis nula: no hay evidencia de heterocedasticidad.")
} else {
  print("Se rechaza la hipótesis nula: hay evidencia de heterocedasticidad.")
}
```

```
## [1] "Se rechaza la hipótesis nula: hay evidencia de heterocedasticidad."
```

modelo 2) con interaccion:

```
# Realizar la prueba de White
prueba_white_con_interaccion_act <- bptest(modelo_con_interaccion_act, ~ fitted(model
o_con_interaccion_act) + I(fitted(modelo_con_interaccion_act)^2))

# Mostrar los resultados de la prueba
print(prueba_white_con_interaccion_act)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_con_interaccion_act
## BP = 44.627, df = 2, p-value = 2.038e-10
```

```
# Extraer el valor p de la prueba
p_value_white_con_interaccion_act <- prueba_white_con_interaccion_act$p.value

# Comparar el valor p con el umbral de significancia de 0.05
if (p_value_white_con_interaccion_act > 0.05) {
  print("No se rechaza la hipótesis nula: no hay evidencia de heterocedasticidad.")
} else {
  print("Se rechaza la hipótesis nula: hay evidencia de heterocedasticidad.")
}
```

```
## [1] "Se rechaza la hipótesis nula: hay evidencia de heterocedasticidad."
```

independencia

hipotesis nula: los errores no estan correlacionados

hipotesis alternativa : los errores estan correlacionados

regla de decisión: si el valor p de la prueba es menor a 0.05 se rechaza la hipotesis nula

modelo 1) sin interaccion:

```
# Realizar la prueba de Durbin-Watson
prueba_dw_sin_interaccion_act <- dwtest(modelo_sin_interaccion_act)

# Mostrar los resultados de la prueba
print(prueba_dw_sin_interaccion_act)
```

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

verificamos que el valor p no sea menor a 0.05

```
# Extraer el valor p
p_value_dw_sin_interaccion_act <- prueba_dw_sin_interaccion_act$p.value

# Interpretar los resultados
if (p_value_dw_sin_interaccion_act > 0.05) {
  print("No hay evidencia de correlación en los residuos (hipótesis nula no rechazad
a).")
} else {
  print("Hay evidencia de correlación en los residuos (hipótesis nula rechazada).")
}
```

```
## [1] "Hay evidencia de correlación en los residuos (hipótesis nula rechazada)."
```

modelo 2) con interaccion:

```
# Realizar la prueba de Durbin-Watson
prueba_dw_con_interaccion_act <- dwtest(modelo_con_interaccion_act)

# Mostrar los resultados de la prueba
print(prueba_dw_con_interaccion_act)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_con_interaccion_act
## DW = 1.0509, p-value = 1.575e-12
## alternative hypothesis: true autocorrelation is greater than 0
```

verificamos que el valor p no sea menor a 0.05

```
# Extraer el valor p
p_value_dw_con_interaccion_act <- prueba_dw_con_interaccion_act$p.value

# Interpretar los resultados
if (p_value_dw_con_interaccion_act > 0.05) {
  print("No hay evidencia de correlación en los residuos (hipótesis nula no rechazad
a).")
} else {
  print("Hay evidencia de correlación en los residuos (hipótesis nula rechazada).")
}
```

```
## [1] "Hay evidencia de correlación en los residuos (hipótesis nula rechazada)."
```

linealidad

para verificar la linealidad vamos a utilizar una prueba RESET

La prueba RESET de Ramsey (Regression Equation Specification Error Test) es utilizada para detectar posibles errores de especificación en un modelo de regresión lineal. La prueba examina si hay variables omitidas o si la forma funcional del modelo es incorrecta.

hipotesis nula: no hay términos omitidos que indican linealidad

hipotesis alternativa: hay una especificación errónea en el modelo que indica no linealidad

modelo 1) sin interaccion:

vamos a hacer la prueba RESET

```
# Realizar la prueba RESET de Ramsey
prueba_reset_sin_interaccion_act <- resettest(modelo_sin_interaccion_act)

# Mostrar los resultados de la prueba
print(prueba_reset_sin_interaccion_act)
```

```
##
##  RESET test
##
## data:  modelo_sin_interaccion_act
## RESET = 12.027, df1 = 2, df2 = 200, p-value = 1.169e-05
```

extraemos el valor p para evalualr la regla de decisión

```
# Extraer el valor p
p_value_reset_sin_interaccion_act <- prueba_reset_sin_interaccion_act$p.value

# Interpretar los resultados
if (p_value_reset_sin_interaccion_act > 0.05) {
  print("No se rechaza la hipótesis nula: el modelo no parece tener errores de especi
ficación.")
} else {
  print("Se rechaza la hipótesis nula: es probable que haya errores de especificación
en el modelo.")
}
```

```
## [1] "Se rechaza la hipótesis nula: es probable que haya errores de especificación
en el modelo."
```

modelo 2) con interaccion:

vamos a hacer la prueba RESET

```
# Realizar la prueba RESET de Ramsey
prueba_reset_con_interaccion_act <- resettest(modelo_con_interaccion_act)

# Mostrar los resultados de la prueba
print(prueba_reset_con_interaccion_act)
```

```
##
##  RESET test
##
## data:  modelo_con_interaccion_act
## RESET = 9.9821, df1 = 2, df2 = 199, p-value = 7.392e-05
```

extraemos el valor p para evalualr la regla de decisión

```
# Extraer el valor p
p_value_reset_con_interaccion_act <- prueba_reset_con_interaccion_act$p.value

# Interpretar los resultados
if (p_value_reset_con_interaccion_act > 0.05) {
  print("No se rechaza la hipótesis nula: el modelo no parece tener errores de especi
ficación.")
} else {
  print("Se rechaza la hipótesis nula: es probable que haya errores de especificación
en el modelo.")
}
```

```
## [1] "Se rechaza la hipótesis nula: es probable que haya errores de especificación
en el modelo."
```

##Interpreta cada uno de los analisis que realizaste

conclusión para la normalidad de los residuos: los resultados del valor p para ambas pruebas son menos a 0.05 por lo que se puede rechazar la hipotesis nula para ambos modelos.

conclusión para la verificación de la media 0: para ambos casos el valor p fue de 1, el valor más alto posible que indica que no hay evidencia para descartar la hipotesis nula.

conclusion para la homocedasticidad: ninguno de los dos modelos pudo pasar la prueba de heterocedasticidad, parece ser que en ambos modelos hay heterocedasticidad

conclusión para la independencia: en ambos modelos se rechaza la hiporesis nula y hay evidencia de correlación entre los residuos.

conclusión para la linealidad: en ambos modelos es muy probable que haya errores de especificación

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

Ambos modelos presentan problemas en varios de los supuestos clave de la regresión lineal. A partir de esto, podemos inferir que ninguno de los modelos es completamente adecuado en su forma actual. Sin embargo, si tuviera que elegir entre los dos eligiría el modelo que no posee interacción ya que los coeficientes parecen ser significativos y el porcentaje de varianza explicado es suficientemente alto con un 75% .

#Intervalos de predicción y confianza 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

```
# Intervalos de confianza y predicción
predicciones <- predict(modelo_sin_interaccion_act, newdata = ultimate_data,
                        interval = "confidence", level = 0.95)

predicciones_pred <- predict(modelo_sin_interaccion_act, newdata = ultimate_data,
                             interval = "prediction", level = 0.95)
```

```
# Carga la librería
library(plotly)
```

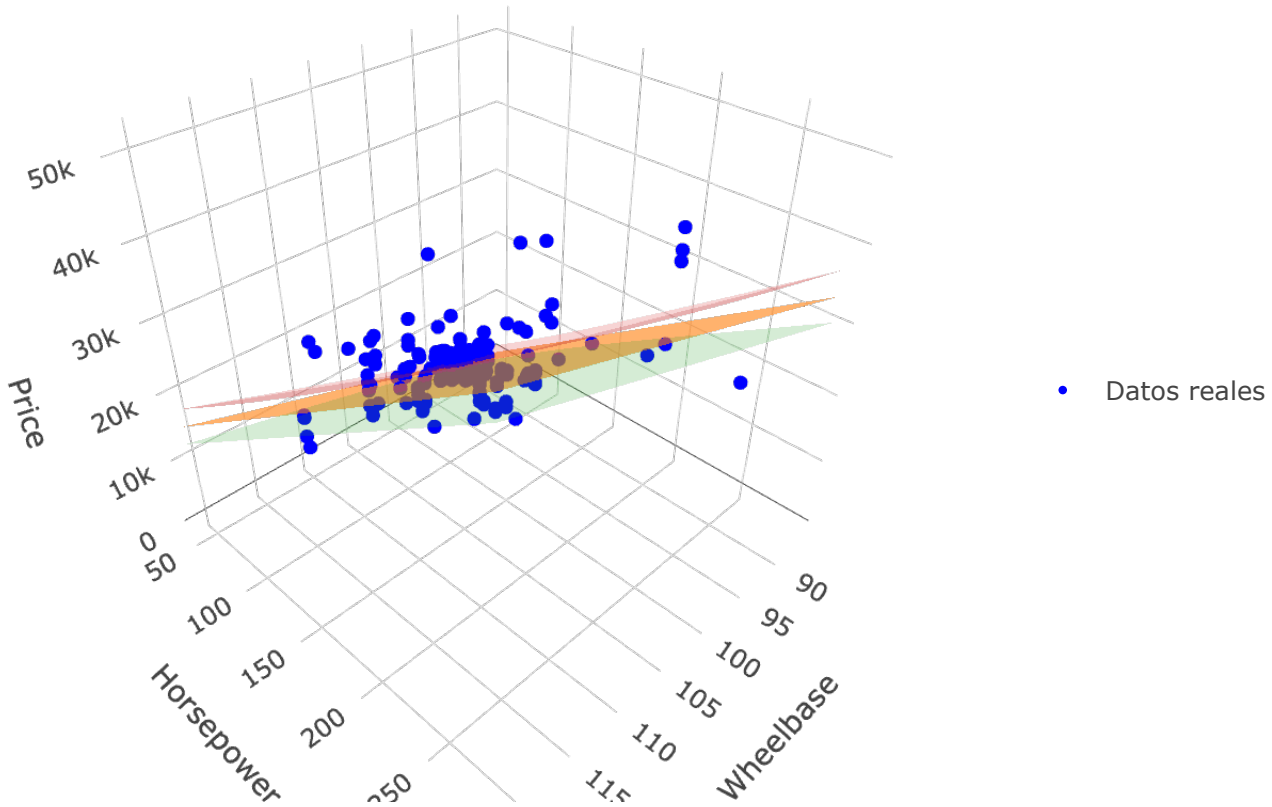
```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
# Paso 1: Crear una cuadrícula de valores de wheelbase y horsepower
wheelbase_vals <- seq(min(ultimate_data$wheelbase), max(ultimate_data$wheelbase), len
gth.out = 100)
horsepower_vals <- seq(min(ultimate_data$horsepower), max(ultimate_data$horsepower),
length.out = 100)

# Crear una cuadrícula con todas las combinaciones posibles de wheelbase y horsepower
grid <- expand.grid(wheelbase = wheelbase_vals, horsepower = horsepower_vals)

# Paso 2: Realizar las predicciones para cada combinación de wheelbase y horsepower
pred_grid <- predict(modelo_sin_interaccion_act, newdata = grid, interval = "confiden
ce")

# Añadir las predicciones y los intervalos al data frame de la cuadrícula
grid$pred_price <- pred_grid[, "fit"]
grid$lwr_conf <- pred_grid[, "lwr"]
grid$upr_conf <- pred_grid[, "upr"]

# Paso 3: Crear la gráfica 3D con datos reales y predicciones
plot_ly() %>%

  # Añadir los puntos de datos reales
  add_markers(data = ultimate_data, x = ~wheelbase, y = ~horsepower, z = ~price,
              marker = list(size = 4, color = 'blue'), name = "Datos reales") %>%

  # Añadir la superficie de predicciones
  add_trace(data = grid, x = ~wheelbase, y = ~horsepower, z = ~pred_price,
            type = 'mesh3d', opacity = 0.6, name = "Precio Predicho") %>%

  # Añadir la superficie del límite inferior del intervalo de confianza
  add_trace(data = grid, x = ~wheelbase, y = ~horsepower, z = ~lwr_conf,
            type = 'mesh3d', opacity = 0.2, name = "Límite Inferior IC", color = I('g
reen')) %>%

  # Añadir la superficie del límite superior del intervalo de confianza
  add_trace(data = grid, x = ~wheelbase, y = ~horsepower, z = ~upr_conf,
            type = 'mesh3d', opacity = 0.2, name = "Límite Superior IC", color = I('r
ed')) %>%

  # Personalizar los ejes
  layout(scene = list(xaxis = list(title = 'Wheelbase'),
                      yaxis = list(title = 'Horsepower'),
                      zaxis = list(title = 'Price')),
         title = "Predicciones y Datos Reales con Intervalos de Confianza")
```



como podemos observar de forma grafica, nuestro modelo es increiblemente malo prediciendo el precio real, tenemos un monton de datos fuera de los limites.

De acuerdo a mi análisis inicial mi variable categorica inicial (fueltype) no es significativa aunque existe una posibilidad de que únicamente podamos utilizar uno de los posibles valores de fueltype para poder explicar el precio de un automovil.

#Más allá: ¿propondrías una nueva agrupación de las variables a la empresa automovilísitica? Definitivamente propondría tomar diferentes variables para poder explicar otro modelo, aunque realmente no hicimos ninguna transformación de las variables para que pudiera corregir la no linealidad y la heterocedasticidad, las pruebas de linealidad parecían indicar que faltaban variables explicativas para nuestros modelos.

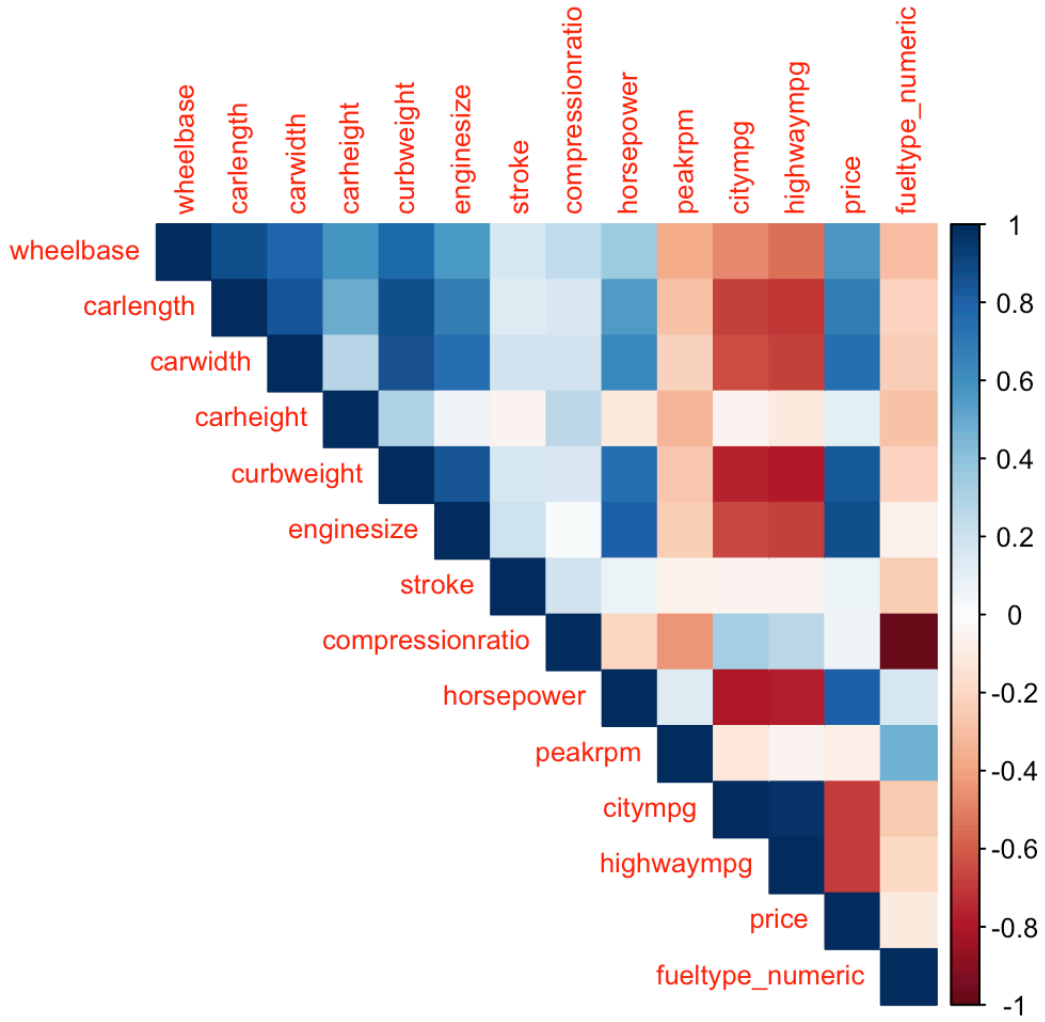
vamos a hacer un análisis muy leve sobre que variables hubieran podido ser útiles, pero antes haré un nuevo dataframe considerando las variables que a simple vista podrían ayudar (los mas obvios), removiendo primary keys, nombres etc.

```
last_analysis <- subset(A, select = -c(CarName, fueltype,symboling,drivewheel,enginet
ype,carbody,enginelocation,cylindernumber))
```

```
summary(last_analysis)
```

```
##    wheelbase        carlength        carwidth       carheight
##  Min.   : 86.60   Min.   :141.1   Min.   :60.30   Min.   :47.80
##  1st Qu.: 94.50   1st Qu.:166.3   1st Qu.:64.10   1st Qu.:52.00
##  Median : 97.00   Median :173.2   Median :65.50   Median :54.10
##  Mean   : 98.76   Mean   :174.0   Mean   :65.91   Mean   :53.72
##  3rd Qu.:102.40   3rd Qu.:183.1   3rd Qu.:66.90   3rd Qu.:55.50
##  Max.   :120.90   Max.   :208.1   Max.   :72.30   Max.   :59.80
##    curbweight     enginesize        stroke      compressionratio
##  Min.   :1488   Min.   : 61.0   Min.   :2.070   Min.   : 7.00
##  1st Qu.:2145   1st Qu.: 97.0   1st Qu.:3.110   1st Qu.: 8.60
##  Median :2414   Median :120.0   Median :3.290   Median : 9.00
##  Mean   :2556   Mean   :126.9   Mean   :3.255   Mean   :10.14
##  3rd Qu.:2935   3rd Qu.:141.0   3rd Qu.:3.410   3rd Qu.: 9.40
##  Max.   :4066   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
##  fueltype_numeric
##  Min.   :0.0000
##  1st Qu.:1.0000
##  Median :1.0000
##  Mean   :0.9024
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
library(corrplot)

last_cor = cor(last_analysis)

# Visualizar la matriz de correlación con un heatmap
corrplot(last_cor, method = "color", type = "upper", tl.cex = 0.8)
```



Esto nos muestra a todas las variables numericas que pueden estar relacionadas. Como dije anteriormente, muy leve.