Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos (Gpo 101)

Reporte final "El precio de los autos"

Francisco Mestizo Hernández
A01731549

Monterrey, Nuevo León

Septiembre 11, 2023

Resumen

Con este reporte se busca dar un análisis de las características de los coches fabricados por empresas americanas y europeas. El fin es entender el mercado para que una empresa China pueda establecer precios competitivos.

Para hacer el modelo, primero se hará la selección de 6 variables. Para decidir las variables se usarán técnicas como analizar la distribución de los datos, su correlación, la desviación estándar y los datos atípicos.

Cuando los datos estén seleccionados, se normalizarán y se hará una regresión lineal para las variables cuantitativas y un ANOVA para las variables cualitativas. Estos modelos serán después analizados con diferentes técnicas, como la verificación de normalidad para la regresión lineal.

Los principales resultados obtenidos fueron que se puede aproximar el precio de un coche basándose en variables que tomen en cuenta características del motor o de la forma del vehículo.

Introducción

El problema que se busca solucionar es la llegada de una empresa China de autos que quiere entrar al mercado de Estados Unidos. La empresa está buscando analizar los coches americanos y europeos para poder establecer precios competitivos para sus coches. Para hacer este análisis se cuenta con una base de datos que cuenta con el precio de un coche y lista las siguientes características de los automóviles: Symboling, car name, fuel type, carbody, drive wheel, engine location, wheel base, car length, car width, car height, curb weight, engine type, cylinder number, engine size, stroke, compression ratio, horse power, peak rpm, city mpg, highway mpg.

Análisis

Esta sección del reporte busca explicar paso a paso lo que se hizo para llegar al modelo final. De todas formas, para entender completamente algunas decisiones o valores obtenidos, se recomienda visitar el reporte listado en los anexos.

Para comenzar con la limpieza de los datos, podemos comenzar analizando como se comporta cada columna. Para verificar que los datos que se utilizarán sean efectivos para el

análisis comenzamos verificando que las columnas no tengan valores faltantes. Este no es el caso para este set de datos, ya que todos los registros están completos.

Debido a que el análisis es diferente para las variables cualitativas y cuantitativas, se hará por separado.

Variables cuantitativas

Viendo el archivo de excel, podemos ver qué columnas incluyen valores cuantitativos (números). Estas variables son: Wheelbase, carlength, carwidth, car height, curbweight, enginesize, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg y price (variable dependiente)

Primero, para el análisis comenzamos imprimiendo la desviación estándar de cada variable, el valor mínimo, el máximo, los cuartiles, la media y la mediana. Por el momento estos datos no nos dan muchas pistas sobre las variables. Solamente la desviación estándar nos puede dar una idea de que tan separados están los datos de la media.

Podemos generar un histograma para ver la distribución de los datos, así como un diagrama de caja y bigotes para cada variable. Lo que trataremos de observar aquí es si los datos se comportan de forma normal o no.

El diagrama de caja y bigote nos ayuda a ver valores outliers, después de la linea roja (más lejos de 1.5 desviaciones estándar) o valores extremos, después de la línea azul (más lejos de 3 desviaciones estándar).

Aunque con los diagramas podemos ver si hay una concentración alta de outliers, no podemos saber con precisión cuántos datos son, por eso imprimimos cuántos posibles valores outliers tiene cada columna. En general casi todas las columnas tienen pocos outliers, aunque hay algunas como el stroke o compression ratio que pasan de los 20 outliers.

Para considerar una variable para el análisis final, es importante saber si se relaciona con la variable de salida (precio). El coeficiente de correlación nos muestra números que van del -1 al 1. Si tomamos el valor absoluto de estos, entre más se acerquen al 1, quiere decir que tienen una correlación más alta con el precio. El signo nos dice si la correlación es descendiente (negativa) o ascendente (positiva).

Y para ver de forma gráfica cómo interactúan estas variables con el precio final, se pueden hacer gráficos de dispersión para ver el comportamiento en una gráfica.

Variables cualitativas

Por otro lado tenemos las variables categóricas o cualitativas, estas son: Symboling, carname, fueltype, carbody, drivewheel, enginelocation, enginetype y cylindernumber.

Primero, generamos tablas donde podemos ver todas las categorías de cada variable cualitativa. Dentro de estas tablas se muestran las categorías, la frecuencia de cada una y el porcentaje de la frecuencia que tiene. Y para hacer más visual esta información podemos generar gráficas de barras.

Para estas variables es más complicado verificar que haya una correlación con la variable de salida. Aún asi, podemos hacer gráficas de caja y bigote para cada categoría de cada variable. Aquí podremos ver si se muestra algún comportamiento que nos pueda relacionar las variables con el precio del auto.

Selección de variables

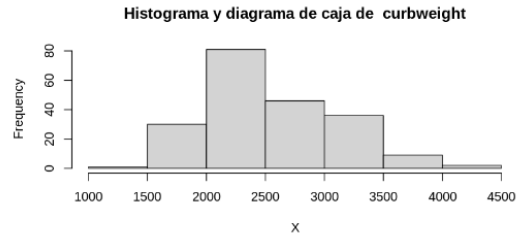
Es importante mencionar que en esta sección se hará un análisis de las seis variables que se elegirán para potenciar las ventas de los coches. Para no hacer muy largo el reporte, solamente se adjuntan gráficos y datos para las variables seleccionadas, pero para ver los resultados para todas las variables se puede consultar el reporte del anexo.

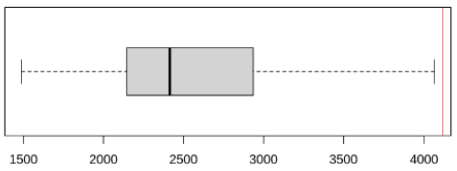
Primero, las variables cuantitativas. Una de las principales razones para seleccionar una variable o no, es el coeficiente de correlación. Por esto, se listan las variables que tienen un coeficiente de correlación alto con el precio:

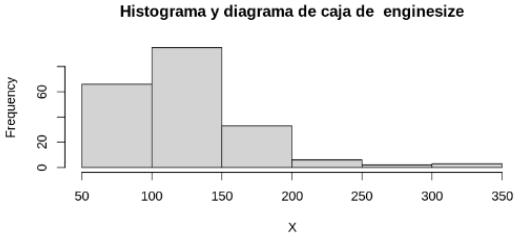
Variable	**Coeficiente de correlación**
Engine size	0.87
Horse power	0.81
City MPG	-0.69
Highway MPG	-0.70
Curb weight	0.84

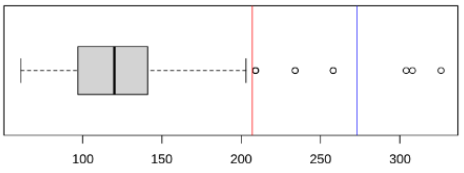
Además, estas variables tienen entre 0 y 10 outliers. Por lo tanto, se puede analizar cada uno de los datos para ver si se pueden eliminar o si se deben mantener para el análisis.

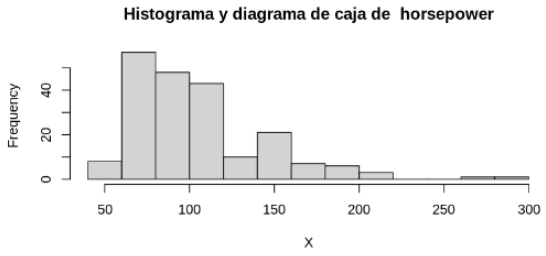
Finalmente, los datos no son normales, casi todas las variables elegidas presentan un ligero sesgo a la derecha. Pero con transformaciones podemos tratar de normalizar estas variables. La falta de normalidad se puede ver en las siguientes gráficas.

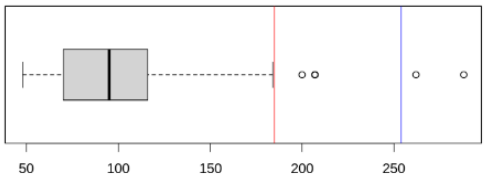


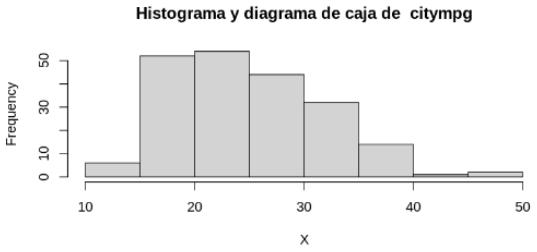


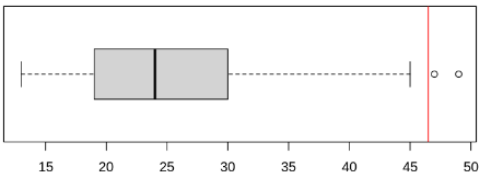


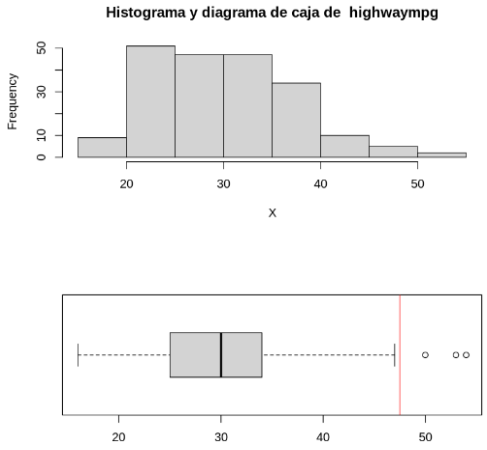




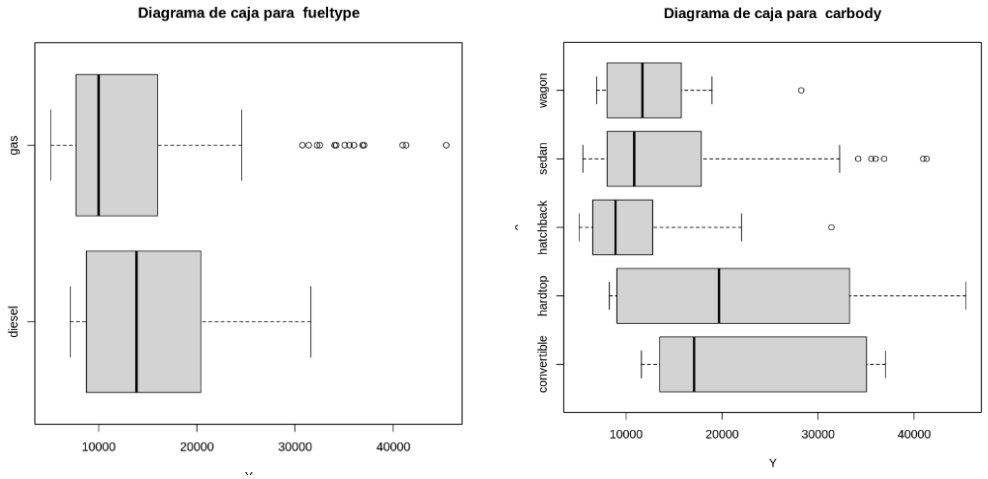


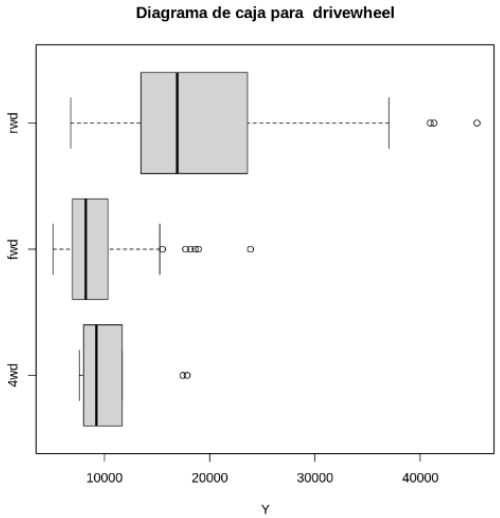






Sobre las variables cualitativas, se eligieron: Carbody, fuel type y drive wheel. Esta elección es porque en los diagramas de caja y bigote se puede observar que dependiendo la categoría en la que está un vehículo, puede aumentar o disminuir su precio.





Hay variables que también presentan este comportamiento, como el enginelocation, pero no se seleccionó, porque de todos los datos, el 90% tienen el motor al frente. Y no hay un cambio tan significativo en el precio por la localización del motor.

Finalmente, es importante considerar estas variables ya que el tamaño del motor o sus caballos de fuerza pueden aumentar su precio ya que son autos más veloces. Por lo mismo, la eficiencia de gasolina (highwaympg y citympg) pueden influir, ya que un motor más fuerte, puede tener peor eficiencia de gasolina.

Además, el curbweight nos dirá el peso total del coche, y entre más accesorios tenga o el motor que tiene, el peso puede aumentar.

Por otro lado, la forma del coche (carbody) puede hacer referencia a si un auto es de lujo o no. Esto mismo puede pasar con el fueltype, ya que hay gasolinas que son más caras que otras.

Creación del modelo lineal

Como ya se había mencionado anteriormente, se utilizará un modelo lineal para las variables cuantitativas. Lo más recomendable es que las variables que se usen para este modelo estén normalizadas. Entonces podemos normalizar los datos utilizando la transformación de boxcox y se usará la lambda exacta obtenida por el modelo. Únicamente normalizamos las variables que son cuantitativas, ya que las cualitativas no se pueden normalizar. Y para eliminar las variables influyentes se utilizará la distancia de cook.

Para tratar de encontrar el mejor modelo se harán tres. El primero será el modelo con los datos normalizados, el segundo será el modelo con los datos sin normalizar y el tercero será eliminando los datos influyentes. A continuación se muestran los resultados de los tres modelos.

Modelo lineal con variables normalizadas

```
Call:
lm(formula = Y ~ Nn$enginesize + Nn$horsepower + Nn$curbweight +
    Nn$carbody)

Residuals:
     Min       1Q   Median       3Q      Max
-11539.2  -2618.0   -486.4   1860.9  16999.0

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3751512.2   450135.1  -8.334 1.32e-14 ***
Nn$enginesize         561475.8   286579.8   1.959 0.051497 .
Nn$horsepower          98702.8    30296.8   3.258 0.001322 **
Nn$curbweight        1810387.6   355706.9   5.090 8.35e-07 ***
Nn$carbodyhardtop       -140.9     2329.5  -0.060 0.951846
Nn$carbodyhatchback    -5043.0     1867.1  -2.701 0.007515 **
Nn$carbodysedan        -3789.5     1834.8  -2.065 0.040199 *
Nn$carbodywagon        -6802.2     2030.3  -3.350 0.000967 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4301 on 197 degrees of freedom
Multiple R-squared:   0.72,     Adjusted R-squared:  0.7101
F-statistic: 72.38 on 7 and 197 DF,  p-value: < 2.2e-16
```

Modelo lineal con variables sin normalizar

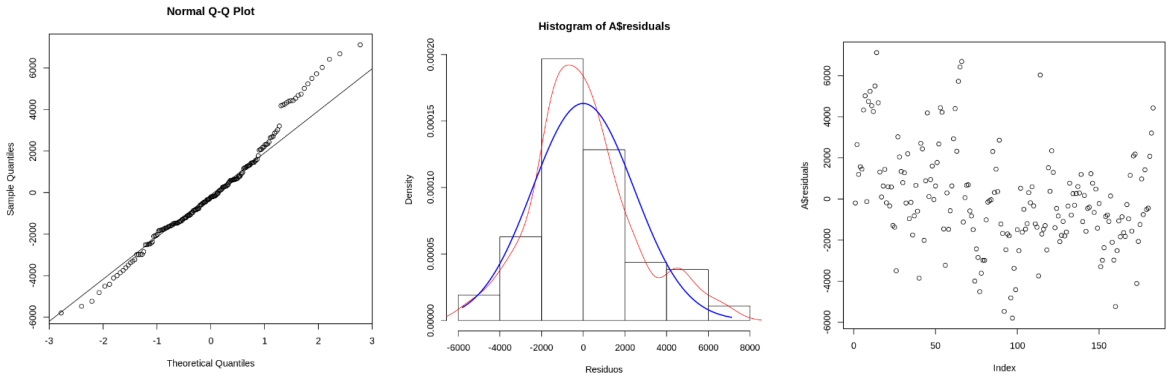
```
Call:
lm(formula = Y ~ N$enginesize + N$horsepower + N$curbweight +
    N$carbody)

Residuals:
    Min      1Q  Median      3Q     Max
-9144.4 -1632.9   -24.1  1515.8 13830.7

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -9427.1368  1951.9260  -4.830 2.74e-06 ***
N$enginesize         68.2372    12.9575   5.266 3.63e-07 ***
N$horsepower         50.3111    10.7358   4.686 5.18e-06 ***
N$curbweight          4.9859     0.9795   5.090 8.33e-07 ***
N$carbodyhardtop  -1561.3180  1807.8562  -0.864 0.388842
N$carbodyhatchback -4800.3262  1436.4233  -3.342 0.000996 ***
N$carbodysedan    -3358.9306  1414.1982  -2.375 0.018502 *
N$carbodywagon    -5435.5524  1568.0380  -3.466 0.000647 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3323 on 197 degrees of freedom
Multiple R-squared:  0.8329,    Adjusted R-squared:  0.827
F-statistic: 140.3 on 7 and 197 DF,  p-value: < 2.2e-16
```

Modelo lineal con variables sin normalizar y eliminando datos influyentes

```
Call:
lm(formula = YI ~ NI$enginesize + NI$horsepower + NI$curbweight +
    NI$carbody)

Residuals:
    Min      1Q  Median      3Q     Max
-5794.9 -1482.2  -220.1  1252.3  7118.3

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.424e+04  1.034e+03 -13.767  < 2e-16 ***
NI$enginesize    3.094e+01  1.265e+01   2.446  0.01541 *
NI$horsepower    3.024e+01  1.002e+01   3.017  0.00293 **
NI$curbweight    7.642e+00  8.996e-01   8.495  7.9e-15 ***
NI$carbodysedan  1.225e+03  4.247e+02   2.884  0.00442 **
NI$carbodywagon -1.278e+03  6.634e+02  -1.927  0.05558 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2478 on 177 degrees of freedom
Multiple R-squared:  0.8338,    Adjusted R-squared:  0.8291
F-statistic: 177.6 on 5 and 177 DF,  p-value: < 2.2e-16
```

En estos tres modelos las variables utilizadas son influyentes y podemos ver que el tercer modelo es el mejor ya que tiene el valor F más alto (177.6). Entonces, podemos hacer la validación de este modelo. En el reporte se muestra la validación para otros modelos, pero aquí solamente se mostrarán los resultados para el mejor modelo. Para verificar que el modelo es apropiado para el conjunto de datos, podemos comprobar los siguientes puntos: Normalidad de los residuos, verificación de media cero, y homocedasticidad e independencia.

Prueba de normalidad de Shapiro Wilk

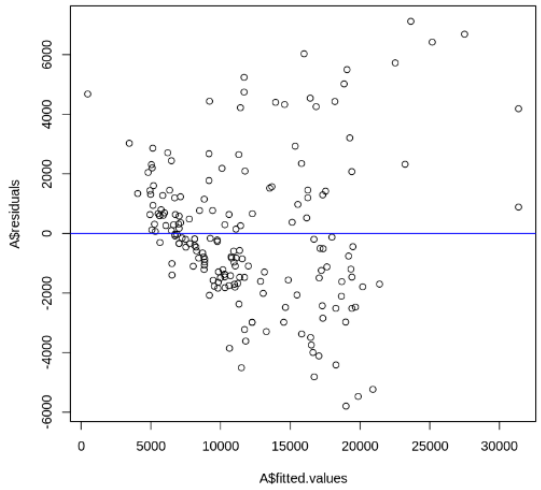


La prueba nos arroja estos resultados, podemos ver que en el QQ Plot tenemos la cola de arriba un poco pronunciada, pero la de abajo está muy cerca de la línea. En el segundo gráfico se ilustra que los datos son casi normales, ya que la línea roja es muy cercana a la azul (distribución ideal). Además los residuos de la tercera gráfica se encuentran muy esparcidos en toda la gráfica y no parece que tengan un comportamiento que nos indique que estamos obviando alguna variable.

Prueba de media cero

El resultado del test para la t de student nos dice que la media no es 0, por lo tanto el resultado del modelo puede no se completamente confiable. Esto es lo mismo que veíamos arriba en las gráficas, donde hay una cola que está un poco más pronunciada.

Homocedasticidad



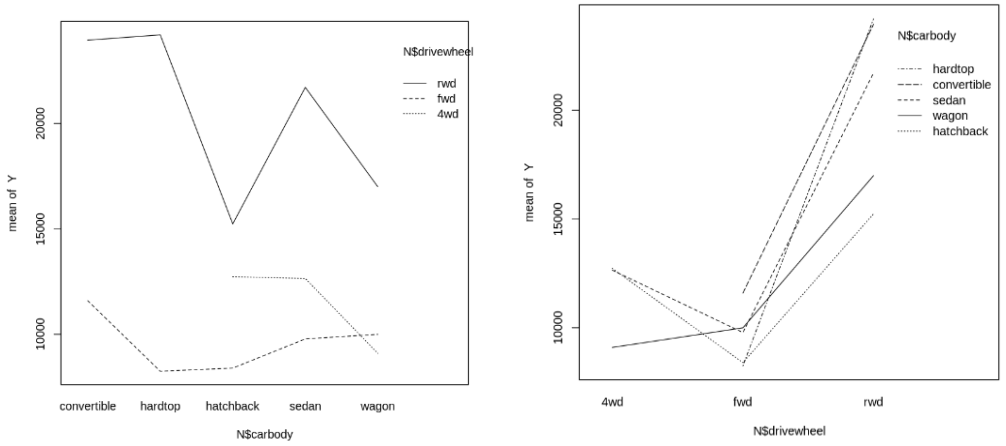
Viendo la gráfica podemos decir que los datos si presentan un comportamiento de homocedasticidad e independencia.

Creación del modelo ANOVA

Podemos hacer un segundo modelo utilizando el método de ANOVA. Es importante mencionar que este modelo utiliza variables categóricas y las compara con la variable dependiente del precio.

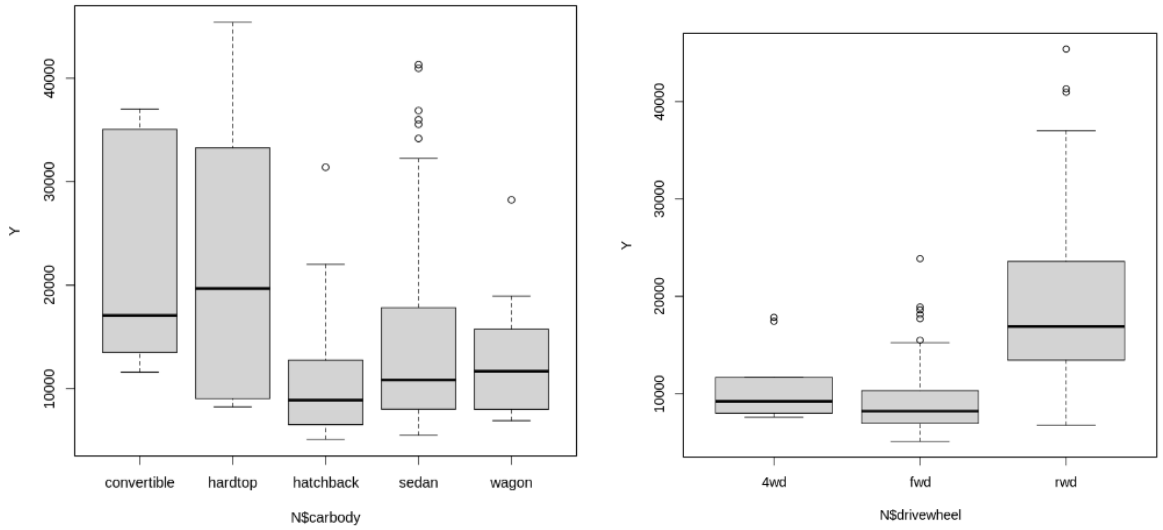
Utilizaremos este modelo porque la mayoría de las variables categóricas quedaron fuera del modelo lineal, y aquí podremos comprobar si con el ANOVA podemos encontrar más variables que sean significativas para el modelo.

Los resultados para las interacciones que conseguimos nos dicen que las variables sí son importantes para el modelo porque las líneas chocan entre ellas. Esto lo podemos ver en las siguientes gráficas.



Las dos variables utilizadas para este modelo son significativas, ya que tienen un valor p muy pequeño, por lo que no rechazamos la hipótesis.

En los diagramas de caja y bigotes podemos ver que la distribución de los datos es bastante amplia, es decir, que sus intervalos de confianza se encuentran separados. Por esto se sigue sustentando que son variables significativas para el precio final que tendrá el coche.



Conclusión

Después de realizar las pruebas para los modelos, podemos concluir que las variables seleccionadas en su mayoría pueden dar una buena explicación para el precio que tienen esos coches.

La empresa puede tomar en cuenta que para variar el precio de un coche es muy importante tomar en cuenta el tamaño del motor, los caballos de fuerza que tiene el mismo o

la tecnología que se usa para la tracción de las ruedas. Además, la forma del auto puede ser que determine si pertenece a una línea de lujo o no, y por eso afectará en su precio.

Aun así, es importante mencionar que esto no implica que las variables que no fueron utilizadas para la generación de modelos no sean significativas para determinar el precio, pero se recomienda altamente que la empresa tome en cuenta las variables mencionadas anteriormente para entender los precios de los coches que se encuentran en el mercado americano. Y así, poder establecer precios de sus carros tomando en cuenta esas mismas variables para competir en el mercado.

Anexo

1. Reporte generado en R para el análisis completo de las variables y los modelos. https://colab.research.google.com/drive/1aLgrR_p-AwMN-XaD597TqSlmdtGv2cYX?usp=sharing