Instituto Tecnológico y de Estudios Superiores de Monterrey

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

Módulo 2 - Estadística
Reporte Final

Jorge Eduardo De León Reyna - A00829759

Septiembre 12, 2023

1. Resumen

La problemática central se enfoca en una empresa automovilística china que busca ingresar al mercado estadounidense y competir con fabricantes locales y europeos. Se plantea la pregunta sobre qué variables son más influyentes en la predicción del precio de los automóviles en Estados Unidos y que tan significativas son estas variables en la explicación de las variaciones de precios.

El enfoque inicial del análisis incluyó una exploración de datos para comprender su forma y distribución, seguida de una etapa de transformación de datos para prepararlos para modelos estadísticos. Los resultados principales, obtenidos a través de la aplicación de un modelo de regresión lineal y una prueba ANOVA, resaltaron la importancia de dos variables clave: los caballos de fuerza (horsepower) y el peso del automóvil (curb weight) en la predicción de precios en el mercado estadounidense.

2. Introducción

La empresa automovilística extranjera mencionada anteriormente se enfrenta al desafío de ingresar al mercado estadounidense. Es por eso que para lograrlo debe comprender los factores que influyen en el precio de los automóviles en este mercado, que difiere significativamente del chino. Este análisis se centra en identificar las variables determinantes del precio de los automóviles en Estados Unidos. y su influencia, brindando información crucial para tomar decisiones estratégicas.

La importancia de este estudio radica en que genera datos e información para la empresa automovilística para formular una estrategia exitosa en su incorporación a este nuevo país. A través de este análisis estadístico se busca desbloquear las puntos clave que permitirán a la empresa competir eficazmente en este nuevo mercado.

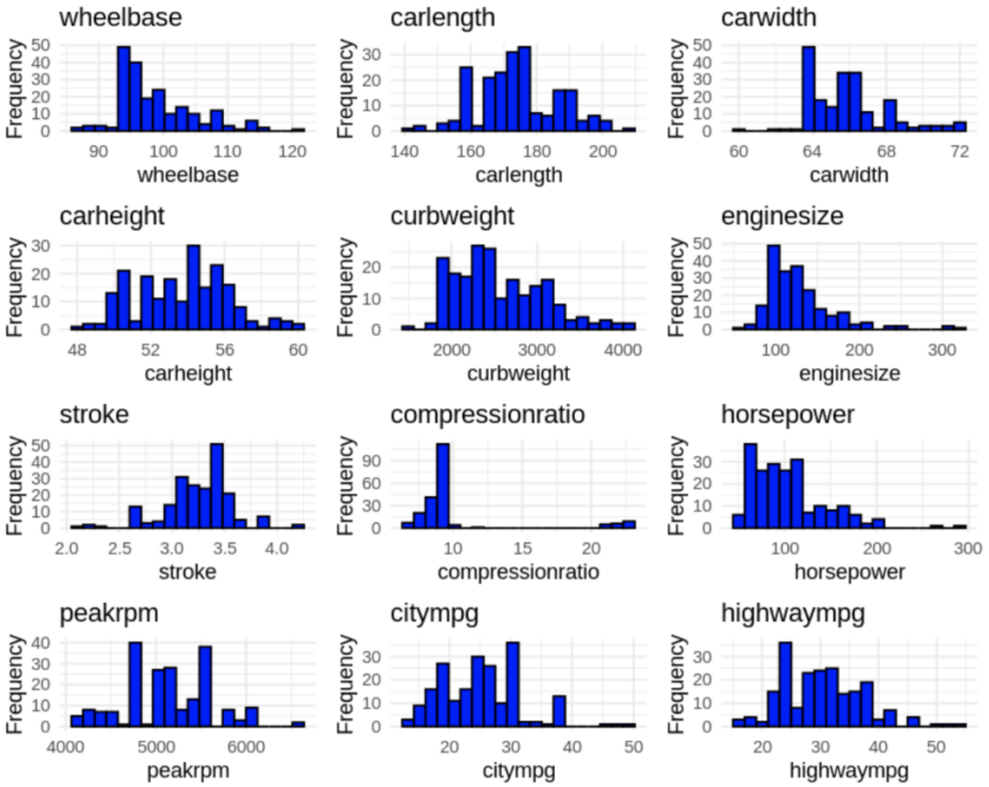
3. Análisis de los resultados

Análisis exploratorio

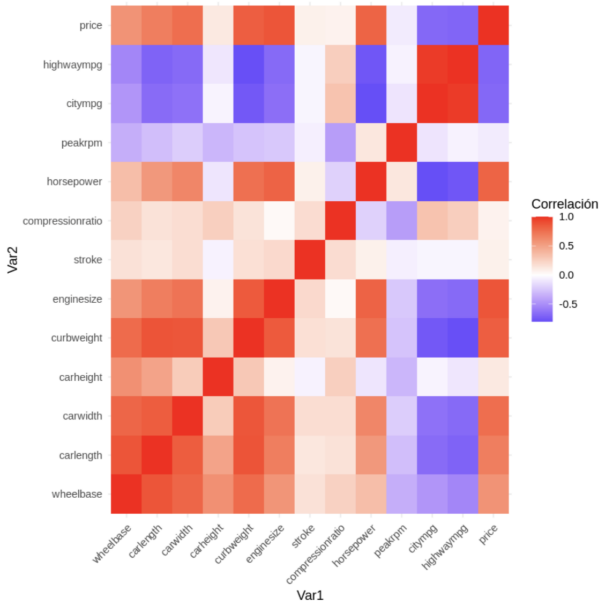
a) Medidas estadísticas: Como primer paso para el análisis se extrajeron las medidas estadísticas tanto para las variables numéricas como para las categóricas que componen el dataset.

b) Exploración visual de los datos: Para conocer sobre la distribución de los datos, datos atípicos y comportamiento general de los datos se hizo uso de boxplots e histogramas. A continuación se muestran algunos resultados:

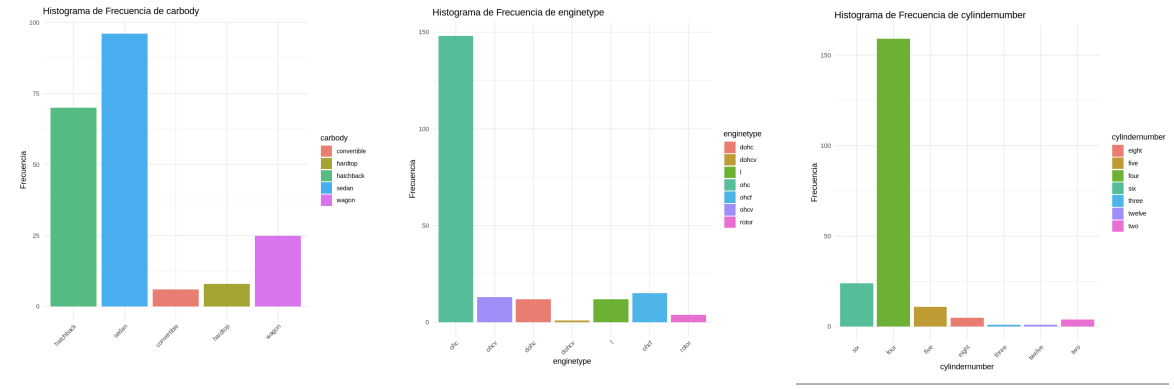
i) Histogramas de frecuencias de variables numéricas: En estos gráficos se puede generar una idea inicial del comportamiento y la distribución de los datos donde una de las principales observaciones que se puede hacer es el hecho de que ninguna variable presenta una distribución normal a simple vista.

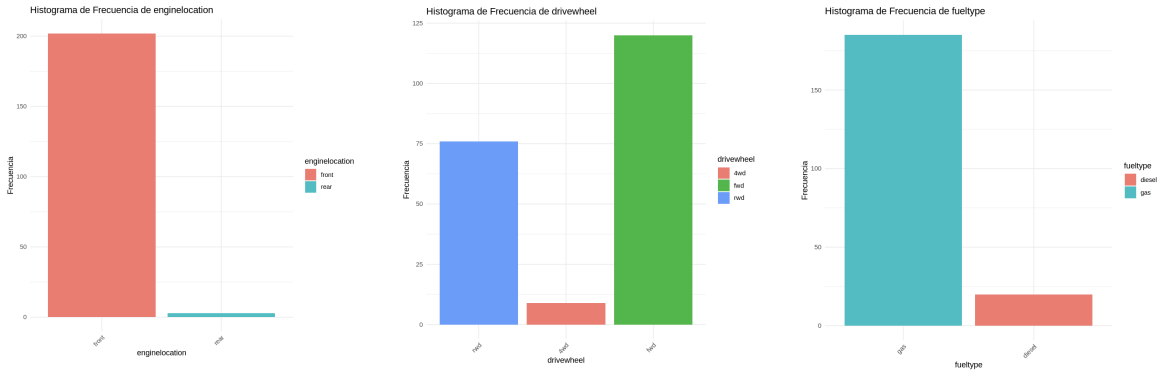


ii) Diagrama de correlación para variables numéricas: En este diagrama podemos observar cuales son las variables que más influyen con la variable objetivo (precio) lo cual nos da una guía para seguir sobre cuáles podrían ser las variables más importantes para la predicción deseada más adelante.

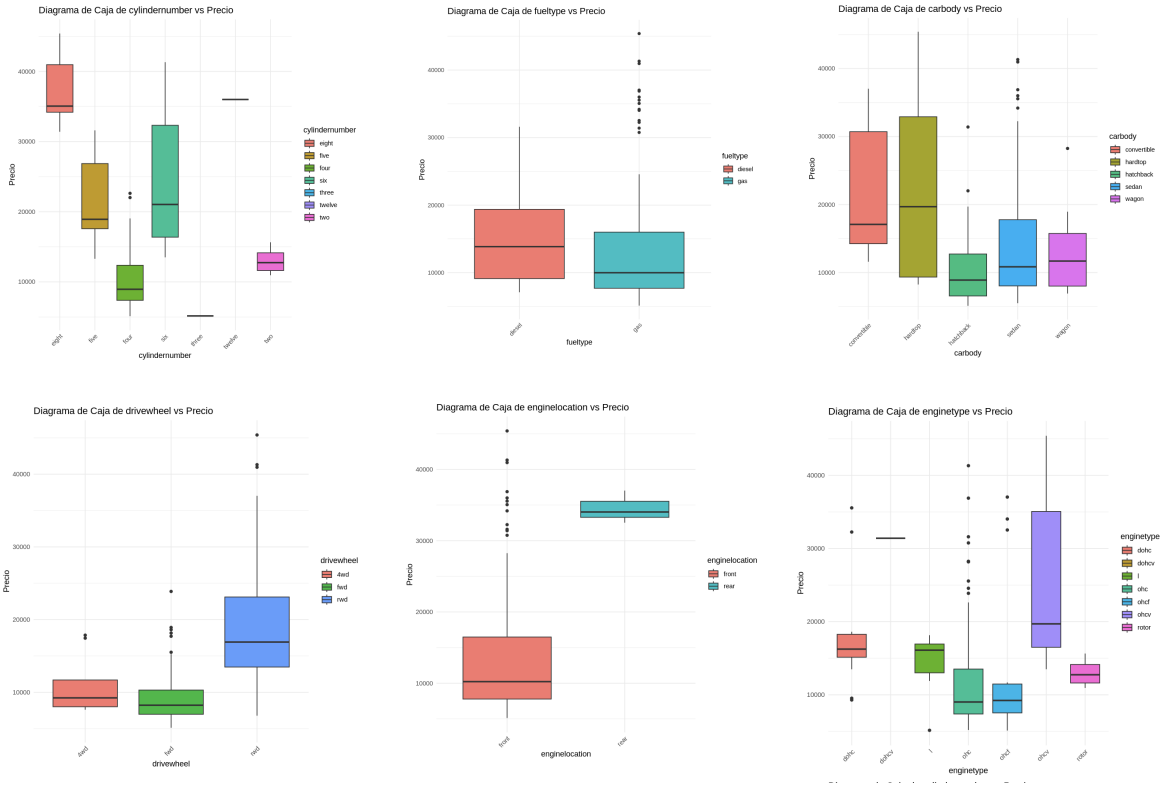


iii) Histograma de frecuencia de variables categorías: En estos gráficos, podemos tener un acercamiento inicial a la frecuencia de las variables categorías presentes en el set de datos. Como observación inicial se identifica que se presenta un desbalance en la frecuencia de las variables por lo que podría ser necesario un proceso de transformación o escalamiento.



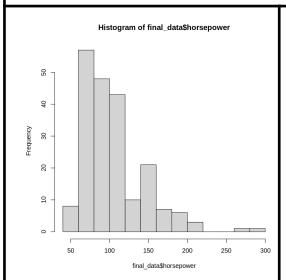
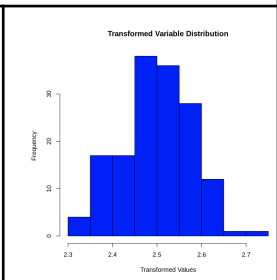
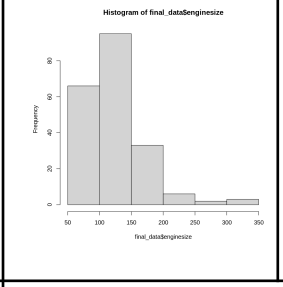
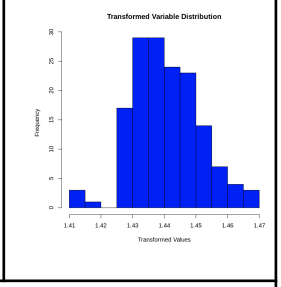
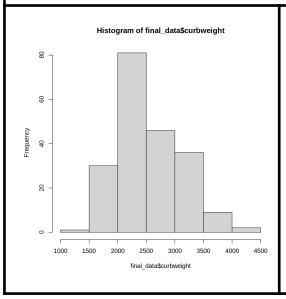
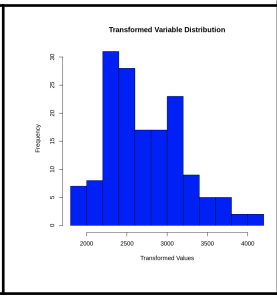
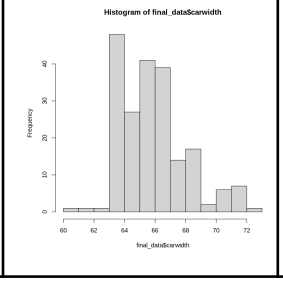
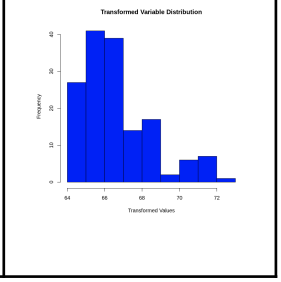


iv) Boxplot de distribución de variables categóricas contra el precio: En estos gráficos podemos entender de manera más clara la relación entre la frecuencia de las variables categóricas en relación a los rangos de precios, así, podemos identificar de la misma manera que con las variables categóricas los datos atípicos.



Transformación de datos

Después de haber seleccionado las variables más significativas acorde a su impacto en la variable objetivo (precio) se llevó a cabo un proceso de limpieza de datos atípicos y de transformación para buscar que su distribución pase a comportarse como una distribución normal. A continuación se muestra el resultado de aplicar el proceso de limpieza de datos atípicos y transformación de datos. Es importante mencionar que la eliminación de datos atípicos se llevó a cabo bajo un proceso iterativo donde la versión final que se presenta es la que mejores resultados obtiene en los modelos posteriores.

Transformación de “horsepower”		Transformación de “enginesize”	
			
Transformación de “curbweight”		**Transformación de “carwidht”**	
			

Como podemos ver, en algunas variables, al aplicar un proceso de transformación (boxcox en este caso) su distribución no mejora por lo que solo se aplicará solamente a las variables “horsepower” y “enginesize”, las cuales si tienen una mejoría en cuanto a su distribución acercándose más a una distribución normal. Por otro lado, las variables “carwidht” y “curbweight” se mantendrán igual al no presentar mejoría al aplicar la transformación.

Modelación y verificación del modelo

Para esta sección del reporte, se hizo uso de dos herramientas estadísticas: Regresión Lineal y Prueba ANOVA donde después de hacer distintas iteraciones con el modelo revisando su desempeño acorde a la significancia de las variables dentro de los modelos, el ajuste de los datos a la recta del modelo de regresión y las pruebas de normalidad de residuos como el Q-Q plot, el histograma de los mismos o el test de Kolmogorov-Smirnov se obtuvo que las variables más significativas para el modelo fueron “horsepower” y “curbweight”. A continuación se muestra el proceso de validación del modelo.

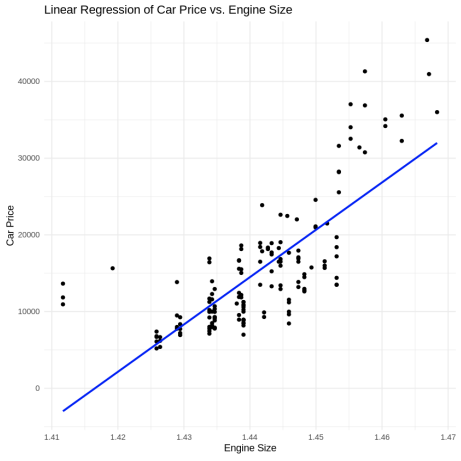
a) Desempeño general del modelo de regresión lineal

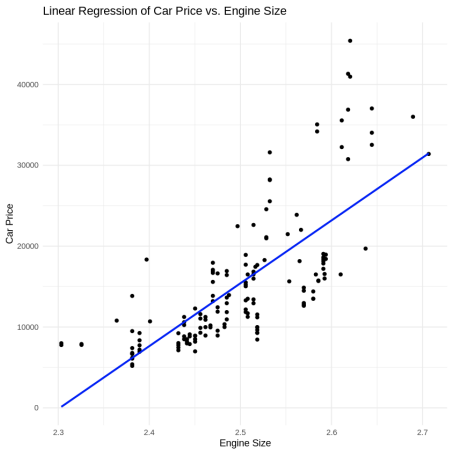
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)            -95616.940  14765.744  -6.476 1.25e-09 ***
final_data2$horsepower  33470.759   6715.184   4.984 1.68e-06 ***
final_data2$curbweight     10.014      1.098   9.119 4.44e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4514 on 151 degrees of freedom
Multiple R-squared:  0.7063,    Adjusted R-squared:  0.7024
F-statistic: 181.6 on 2 and 151 DF,  p-value: < 2.2e-16
```

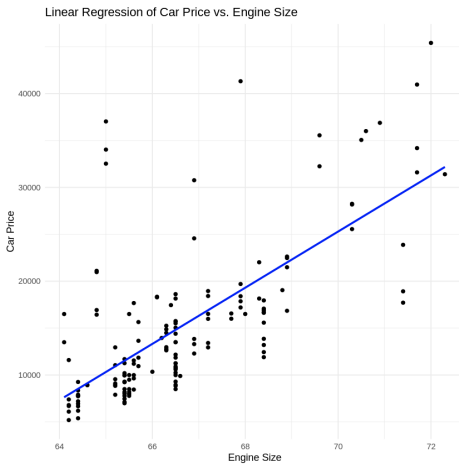
Como se puede observar, el modelo tiene un buen desempeño dado que:

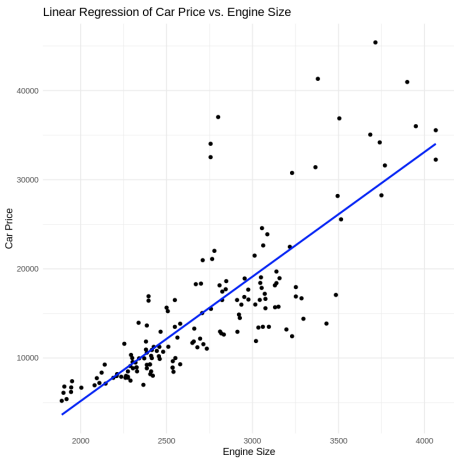
1. Coeficientes: Los coeficientes seleccionados para el modelo tienen un nivel de significancia alto.
2. Valor R Cuadrada: El valor de R Cuadrada Ajustada es cercano a 1 (0.7024) lo que nos habla de que el modelo describe en un 70.24% los datos reales.
3. Prueba de Hipótesis: El valor P es menor a un valor Alfa estándar de 0.05 por lo que es un elemento más para confirmar la validez y buen desempeño del modelo.

b) Ajuste de datos reales a la recta del modelo









Como se observa, la recta parece describir correctamente (con cierto grado de error) al comportamiento de los datos lo cual es un factor que apoya la validez del modelo. Sin embargo, al ser una prueba gráfica no se puede afirmar directamente la validez del modelo solo con esta prueba.

c) Pruebas de normalidad de residuos

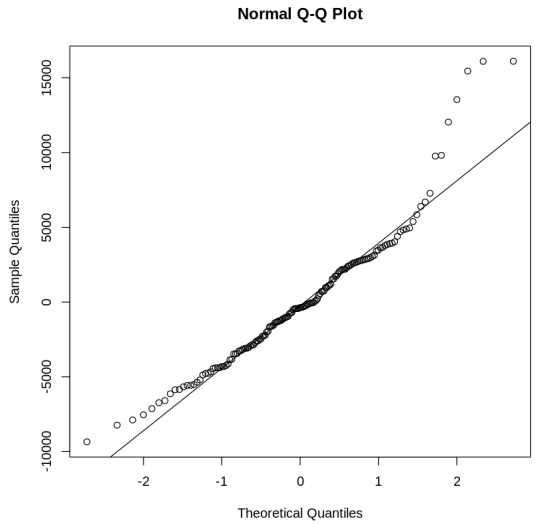
Para confirmar la validez del modelo, se recurrió a hacer un análisis de normalidad de residuos tanto de forma gráfica como como un test numérico. A continuación se enlistan las pruebas realizadas:

i) Prueba Kolmogorov-Smirnov:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuos
D = 0.073852, p-value = 0.3704
alternative hypothesis: two-sided
```

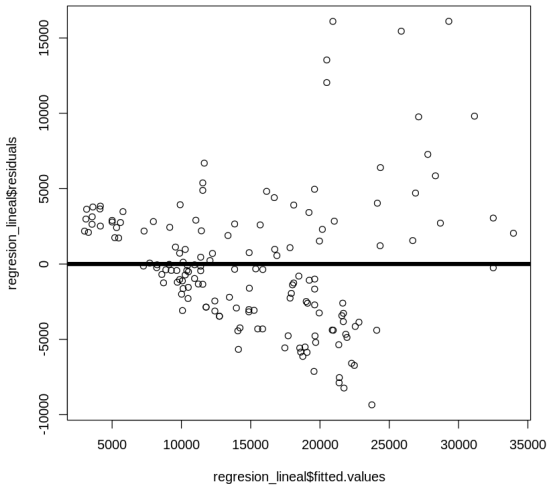
Como se puede observar, debido a que el valor P > 0.05 no se rechaza la H0 y se afirma que los datos se distribuyen normalmente.

ii) Q-Q Plot con distribución de los cuartiles de residuos.



En este Q-Q plot se puede ver una correcta distribución alrededor de la diagonal, lo cual es otro indicio que confirma la validez del modelo de regresión implementado.

iii) Homocedasticidad



En el análisis de homocedasticidad se puede observar que los residuos del modelo de regresión implementado se distribuyen relativamente de forma uniforme cerca del 0 lo cual es otro factor que apoya a la validez y buen desempeño del modelo.

d) Prueba ANOVA

Con el fin de tener otra herramienta estadística que nos ayudará a definir cuáles eran las variables más significativas para la predicción de la variable objetivo se hizo uso de la prueba de ANOVA.

```
Response: final_data2$price
                        Df     Sum Sq    Mean Sq F value    Pr(>F)
final_data2$horsepower   1 5705436757 5705436757 279.981 < 2.2e-16 ***
final_data2$curbweight   1 1694451598 1694451598  83.151 4.439e-16 ***
Residuals              151 3077071475   20377957
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Finalmente, como se mencionó anteriormente, después de distintas iteraciones en el modelo desarrollado podemos hacer un análisis de la significancia de cada una de las variables involucradas en el mismo por medio de la prueba ANOVA la cual podemos verificar por medio del valor p de cada variable individual. Con esto en mente, la prueba ANOVA nos indica que las variables con mayor significancia dentro del modelo dado un valor p menor a 0.05 (valor alfa seleccionado) son: “horsepower”, “curbweight”

4. Conclusión

Después de haber llevado a cabo un proceso de análisis descriptivo así como de generación de modelos estadísticos para encontrar las variables más significativas para predecir la variable objetivo los cuales incluyeron pruebas de hipótesis, la prueba ANOVA, gráficos del modelo y ajuste a los datos, así como pruebas de normalidad de residuos y homocedasticidad, podemos afirmar que el modelo desarrollado es significativo.

En primer lugar, la prueba ANOVA reveló la importancia global del modelo y sus variables predictoras, respaldando nuestras conclusiones. Además, las pruebas de hipótesis arrojaron valores P significativamente bajos, confirmado la relevancia del modelo y sus variables predictoras en la predicción del precio de los automóviles estadounidenses.

Además, la representación gráfica del modelo en relación con las variables más influyentes demostró un ajuste preciso, indicando que el modelo capta eficazmente las relaciones entre estas variables y los precios de los automóviles.

Por otro lado, las pruebas de normalidad de residuos y homocedasticidad proporcionaron evidencia sólida de que el modelo es adecuado para las expectativas estadísticas, lo que refuerza su validez y precisión.

Finalmente, este análisis estadístico nos da una base sólida y es válida para la toma de decisiones estratégicas de la empresa automovilística china en su búsqueda de ingresar y competir en el mercado estadounidense. Los resultados destacan la importancia de las variables identificadas y su influencia en los precios de los automóviles, lo que proporciona orientación esencial para el éxito en un mercado altamente competitivo.

5. Referencias

No se hizo uso de referencias bibliográficas para este reporte.

6. Anexos

Carpeta con Notebook de Análisis en R y Base De Datos:

https://drive.google.com/drive/folders/1RQF6tv9bMHvf5_Yw41bcqSFRRyq0uORZ?usp=sharing