Instituto Tecnológico y de Estudios Superiores de Monterrey

Reporte final, "El precio de los autos"

Francisco Castorena Salazar, A00827756

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

TC3006C.101

Dra. Blanca R. Ruiz Hernández

10 de Septiembre del 2023

Resumen

Este informe se enfoca en identificar los principales factores que afectan los precios de los automóviles en Estados Unidos. Utilizamos métodos estadísticos y técnicas como Box Cox, análisis exploratorio, correlaciones y modelos de regresión para seleccionar las variables clave. Posteriormente, aplicamos pruebas y análisis para validar nuestras hipótesis. Los resultados ayudan a comprender mejor la influencia de las variables en los precios de los automóviles en el mercado estadounidense para así llegar a las conclusiones que se mostrarán al final del documento.

1. Introducción

1.1. Sobre la problemática

La problemática a resolver se define como sigue:

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 contra partes 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 se presenta como la base de datos que se compartió a los miembros de esta clase, así como un diccionario de términos que describe brevemente cada variable, que es y el tipo de datos que contiene (categóricos, numéricos continuos). Para consultar la base de datos y el diccionario de términos se puede acceder dado clic en el siguiente enlace: ENLACE.

1.2. Sobre la importancia de la problemática

El hacer análisis estadístico sobre estos datos es crítico para abordar con éxito esta problemática, al proporcionar una base cuantitativa sólida para la toma de decisiones estratégicas, se le permite a la empresa comprender y responder a las dinámicas del mercado de forma informada y efectiva. La aplicación de este tipo de análisis permite tener un punto de vista más amplio amplio, el cual se considera que es necesario en un contexto global y de constante cambio.

2. Análisis de los resultados

El análisis hecho sobre la base de datos se realizó utilizando R en un archivo de tipo R Markdown, si se desea consultar el notebook en formato pdf ó Rmd con el código se puede dar clic en el siguiente enlace: ENLACE.

2.1. Carga y limpieza de la base de datos

La base de datos sobre la cual se realizó el análisis tenía 205 instancias (sin contar headers) y 21 columnas, de las cuales 7 fueron del tipo categóricas y 13 de tipo numéricas tanto con únicamente números enteros como numéricas continuas, a continuación se muestran los nombres de las variables y su clasificación.

- Categóricas: symboling, fueltype, carbody, drivewheel. enginelocation, enginetype, cylindernumber.
- Numéricas: wheelbase, carlength, carwidth, carheight, curbweight, enginesize, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg, price.

Se buscó por valores nulos en la base de datos por medio de la función $colSums(is.na(df))$, no se encontró ningun valor nulo en la base de datos.

2.2. Análisis de distribución de datos para variables numéricas

Para analizar si las variables numéricas se comportaban acorde a la distribución normal, primeramente se realizaron histogramas sobre todas estas variables. Como ejemplo de los histogramas arrojados en el notebook véase la figura 1.

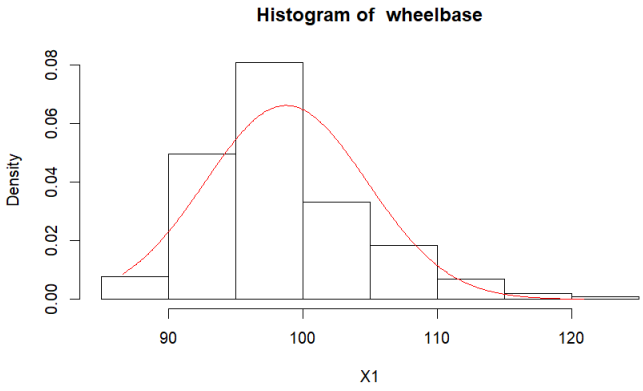

Figura 1: Histograma de la variable wheelbase con su función de densidad de probabilidad normal correspondiente a su media y desviación estándar.

Si se desea consultar un histograma en específico, véase el notebook en formato pdf compartido en en enlace anterior.

Después de ver los histogramas de las variables numéricas se concluyo que todas "encajaban" de forma correcta sobre una distribución normal para ser consideradas como normales, sin embargo había presencia de distintos grados de sesgo para ciertas variables, como la variable price (véase figura 2) que es de gran importancia debido a que buscamos determinar como se ve afectada por otro tipo de variables que podrían tener otro tipo de distribución.

Por otra parte se analizaron medidas estadísticas sobre las variables cuantitativas como lo son la media, cuantiles, mediana y moda. A excepción de las variables price, horsepower y compressionratio que son las

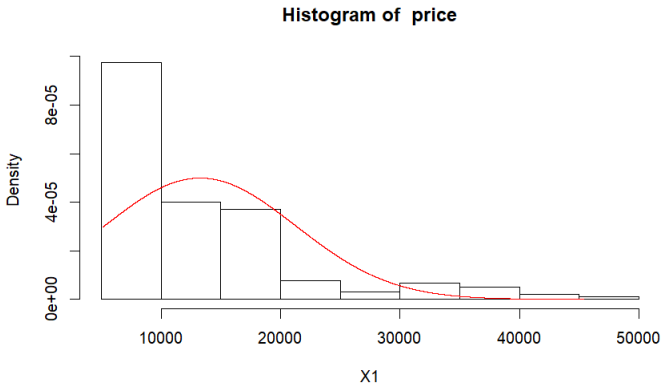

Figura 2: Histograma de la variable price con su función de densidad de probabilidad normal correspondiente a su media y desviación estándar.

variables que presentan mayor sesgo en sus datos originales, se pudo observar que las distancias entre mínimos y máximos de la mediana eran hasta cierto punto simétricas, lo cual pueden indicar un sesgo bajo, por otra parte la media y la mediana en estas variables eran muy parecidas, todos estos factores mencionados pueden ser indicios de un comportamiento acorde a la distribución normal.

2.2.1. Observaciones sobre diagramas de caja de variables cuantitativas

Para cada una de las variables cuantitativas se realizó su respectivo diagrama de caja, para observar la distribución de datos en cuartiles e identificar valores atípicos. Véase figura 3 como ejemplo de los diagramas de caja obtenidos.

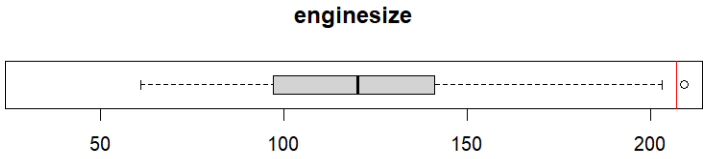

Figura 3: Diagrama de caja de la variable enginesize.

En los boxplots se puede observar como los valores están mayormente agrupados entre el cuartil 1 y el cuartil 3, lo cual es un indicio de normalidad ya que esto nos dice que los datos suelen estar agrupados alrededor de la mediana.

Para identificar valores extremos se definía un valor el cual era 1.5 veces la distancia entre el cuartil 1 y el cuartil 3 y se sumaba al cuartil 3, si el valor de algún dato era mayor a esta, se consideraba como valor extremo. Cabe mencionar que se tuvieron muy pocos datos atípicos, se tenía 1 o ningún dato atípico en todas las variables cuantitativas, lo cual es bueno para poder implementar modelos lineales, ya que la recta se ajusta mejor sobre las zonas donde hay mayor densidad de datos.

En primer momento no se eliminan los datos atípicos, se analizará el comportamiento de las variables y las pruebas de significancia con todos los datos, sin embargo en partes posteriores del documento se explica que si serán tratados para que encajen mejor a una distribución normal.

2.3. Análisis de medidas estadísticas sobre variables cualitativas (cuantiles y frecuencias)

Se analizó la frecuencia y distribución de datos de las variables cuantitativas, se pudo observar que las frecuencias de aparición de los datos respecto al total no era balanceado entre clases, como se puede observar en la figura 5, todas las variables tenían alguna clase que predominaba sobre las demás, por otra parte las variables cuantitativas tenían entre 2 y 7 valores por columna, lo cual considero que es un buen número para hacer análisis sobre grupos, no se tienen demasiados grupos por variable, lo cual hace que no se pierda valor por grupo y estos puedan ser diferenciados más fácilmente por sus características.

2.3.1. Diagramas de barras y de pastel

Se realizaron diagramas de barras y pie charts para analizar de forma visual la distribución de los datos por las frecuencias de sus distintos valores. Véase figura 4 y 5 como ejemplos de los tipos de gráficas que se obtuvieron en esta sección del análisis.

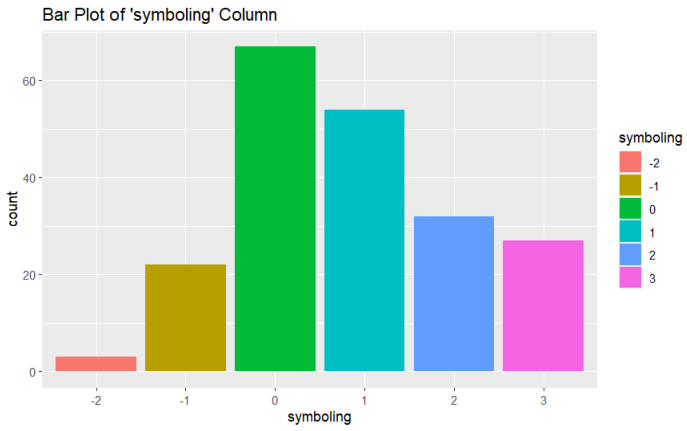

Figura 4: Distribución por tipo de datos, pordría considerarse en este caso como un histograma.

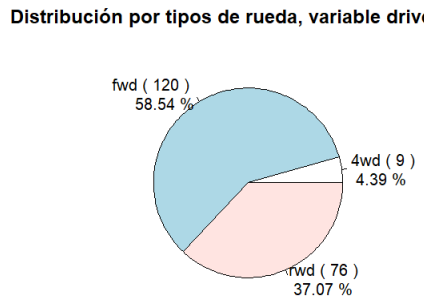

Figura 5: Diagrama de pastel de la variable drivewheel, se pueden observar los porcentajes de cada tipo de valor que tomaba la variable.

2.4. Análisis de correlación

Para identificar aquellas variables que estaban mayormente relacionadas con la variable price, así como identificar aquellas que estaban más relacionadas entre sí, se realizó una matriz de correlación de coeficientes de Pearson con las variables numéricas de la base de datos. Debido a que se obtuvo una matriz de correlación de dimensiones 13x13, era algo difícil identificar aquellas variables que solían representar correlaciones altas con una o varias variables, por lo que primeramente se realizó un heatmap, que sirviera como apoyo visual para esto. Véase figura 6.

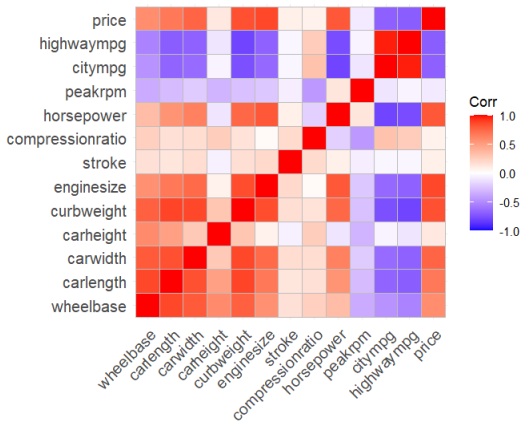

Figura 6: Heatmap de correlación entre variables numéricas del conjunto de datos, podemos ver que aquellos recuadros con colores rojos y azules intensos son aquellos con mayor correlación.

Para la selección de variables se tiene que tener en cuenta aquellas variables que tengan mayor correlación con la variable price, sin embargo tenemos que procurar que la correlación entre variables predictoras sea bajo, esto para evitar colinealidad, aún y bajo el supuesto de que dos variables tienen una correlación alta con la variable price, y también alta entre ellas, puede que la varianza explicada de la variable price por

ambas variables en un modelo multilineal sea casi el mismo al que explican por separado, debido a que suelen comportarse de la misma manera ante el cambio de la variable price.

Por lo tanto, para identificar y seleccionar aquellas variables que pueden ser significativas sobre la variable price, se tomará en cuenta que tengan una alta correlación con la variable price procurando evitar colinealidad, la siguiente subsección del documento nos ayudará a identificar aquellas variables con alta colinealidad.

2.5. Análisis por Componentes Principales (PCA)

Se dio uso del análisis de componentes principales para identificar aquellos grupos de variables que tenían más correlación entre si y aquellas que tenían menos, se dió uso de gráficas como Loading Plots (véase figura 7) para identificar aquellas variables que tienen alta correlación con la variable price así como aquellos que tienen una correlación inversa.

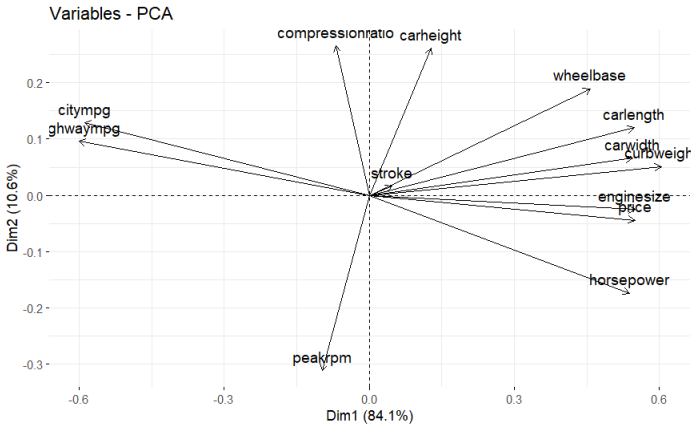

Figura 7: Load plot que muestra las variables con mayor correlación.

Por ejemplo podemos observar que la variable price es en buena medida inversamente proporcional a la highwaympg, debido a que queremos variables que describan la mayor cantidad de variabilidad de la variable price para futuros análisis, no solo es importante escoger variables muy correlacionadas como curbweight, sino también otras que tengan una correlación menor pero inversamente proporcional o cercana a 0 para ayudar a definir esta varianza de forma más amplia.

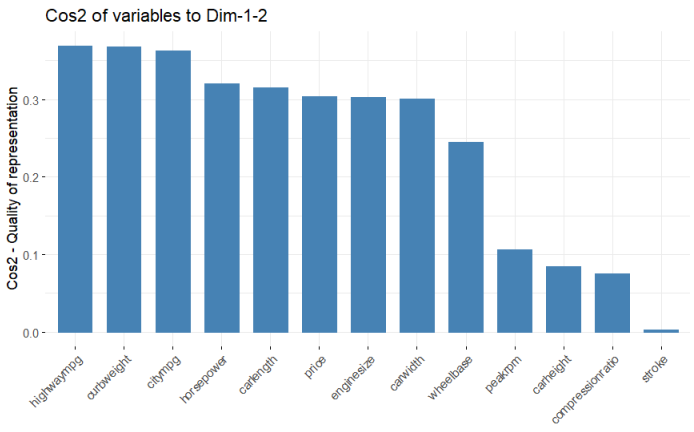

Figura 8: Bar plot que muestra las variables con mayor peso sobre los componentes creados para la reducción de dimensionalidad de los datos originales

En la figura 8 podemos observar aquellas variables que por si solas, pueden explicar una mayor varianza de la base de datos al ser combinadas de forma lineal entre ellas mismas, por lo cual, no es una mala idea, para los análisis posteriores de hipótesis, seleccionar a variables que puedan describir altamente a la base de datos por si sola.

2.6. Selección de variables para futuros análisis de significancia sobre la variable price

Ante todos los análisis estadísticos exploratorios realizados previamente, ahora podemos tener una idea más clara de que variables tienen buenos indicios ára describir de mejor forma el cambio de la variable price, teniendo en cuenta aspectos como correlación, normalidad, escala de los datos y tipos de datos, se decidió que sobre las siguientes 7 variables se harán análisis de significancia sobre la variable price, estás son **highwaympg, curbweight, horsepower, carheight, wheelbase, enginesize, symboling (variable categórica)**.

2.7. Prueba de Anderson-Darling sobre variable seleccionadas

Se realizó la prueba de normalidad de Anderson-Darling para analizar la normalidad de cada una de las variables continuas seleccionadas (incluyendo variable price) para hacer pruebas de significancia.

En esta prueba se tiene la hipótesis nula que nos dice que los datos no siguen una distribución normal.

Los resultados específicos del test para cada variable pueden ser encontrados en el notebook donde se hizo el análisis. Después de realizar el test para cada variable (menos symboling), se pudo observar que el p-value es menor a 0.05, por lo que podemos rechazar la hipótesis nula que dice que los datos no siguen una distribución normal, por lo que concluimos con un 95 % de confianza que todas las variables seleccionadas para análisis de significancia siguen una distribución normal (incluyendo price).

2.8. Análisis de sesgo y curtosis de las variables

Para hacer modelos de regresión lineal con variables que siguen la distribución normal, se considera mejor que estas tengan poco sesgo y una curtosis cercana a 3, esto para que las predicciones se ajusten mejor a la recta que minimiza el MSE de los datos.

Teniendo esto en cuenta, se decidió hacer un análisis de sesgo y curtosis para todas las variables continuas (highwaympg, curbweight, horsepower, carheight, wheelbase, enginesize), con y sin transformación de box-cox. La transformación de Box-Cox se realiza con el fin de disminuir el sesgo y curtosis de los datos, de forma que los datos se acoplen mejor a la distribución normal.

2.8.1. Sobre la transformación Box-Cox

La transformación Box-Cox es una técnica utilizada en estadísticas y análisis de datos para mejorar la normalidad y estabilizar la varianza en distribuciones de datos asimétricas o heterocedásticas. Fue desarrollada por los estadísticos George Box y Sir David Cox. El propósito principal de esta transformación es hacer que los datos se ajusten mejor a las suposiciones de muchas técnicas estadísticas, como regresión lineal y análisis de varianza, que asumen normalidad y homocedasticidad.

2.8.2. Resultados de transformación Box-Cox sobre las variables seleccionadas

La transformación de Box-Cox sobre las variables seleccionadas redujo en todas las variables, el sesgo y el nivel de curtosis, para la variable price (véase figura 9), que era la que tenía un mayor sesgo, disminuyo la calificación de sesgo original obtenida de 1.764 a 0.095 con la transformación hecha, por otra parte la variable de mayor valor de curtosis (enginesize) con un valor de 8.14, disminuyo a 3.27, lo cual es un valor muy bueno, recordando que se quiere tener una curtosis cercana a 3.

Figura 9: Histograma de los valores de la variable price, aplicando la transformación Box-Cox.

Recordando que los modelos lineales asumen normalidad y homocedasticidad de las variables, se considera que es mejor realizar los análisis y modelos a seguir sobre los datos transformados y no los originales, ya que como vimos anteriormente, se mejoró la noormalidad de los datos en gran medida.

2.9. Tukey test para estimar el cambio de la variable price, sobre la variable symboling(categórica)

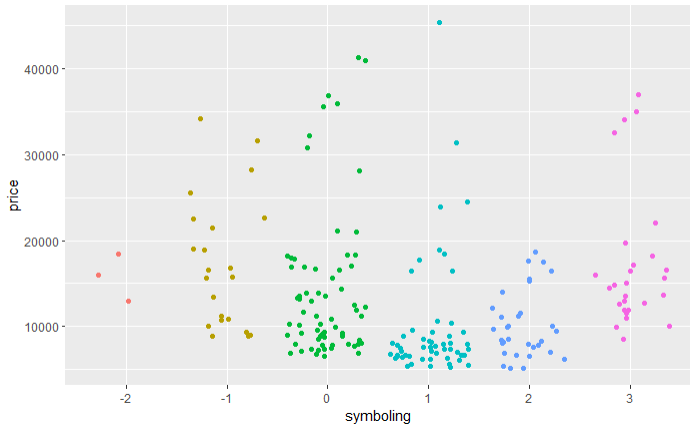

Figura 10: Scatter Plot por tipo de valor categórico en la variable symboling y valores respectivos para la variable price

En la figura 10 podemos observar como no hay una clara diferencia de varianza en los precios dependiendo del tipo de valor categórico en la variable symboling, se ve como para todo tipo de valor categórico, los precios tienden a estar concentrados en un mismo precio (alrededor de 10000).

Sin embargo para comprobar o rechazar las hipótesis en la cual asumimos que el cambio de valor en la variable symboling no afecta el valor de la variable price, haremos uso del test de Tukey, el cual no solo nos permite definir si hay significancia de una variable categórica sobre una continua, sino que también nos dice cuales grupos dentro de la variable categórica son aquellos que tienen una diferencia significativa para la variable continua.

En nuestro caso tenemos 6 grupos para la variable categórica, el test de Tukey en R nos mostrará aquellos grupos sobre los cuales se tiene una diferencia significativa en la variable price.

Los grupos de la variable symboling sobre los cuales se hayo una diferencia significativa en la variable price fueron los siguientes: (-1,1),(-1,2),(0,1),(1,3),(2,3).

Por lo que concluimos que se tienen diferencias de precios en automóviles dependiendo del grupo de symboling al que pertenecen.

2.10. Análisis de regresión sobre las variables numéricas transformadas

Se pudo observar en los modelos lineales creados para cada variable numérica respecto a la variable price, que la media de los residuos es cercana a 0, y estos se distribuyen de forma simétrica, el valor F estadístico es alto y el p-value de todos los modelos es bajo, menor a 0.05, la varianza explicada de los modelos va desde el 2 % con la variable carheight, al 81 % con la variable curbweight.

También se realizarán regresiones multilineales para encontrar aquel modelo multilineal, que explica mayor variablidad de la variable price.

2.10.1. Análisis de modelo multi-lineal con variables numéricas transformadas

Para encontrar el mejor modelo multi-lineal que explique mejor la varianza de la variable price, se hará uso de la función step() en R a la cual se le alimenta un modelo de regresión multi-lineal con todas las variables $linearmodel < -lm(price\ enginesize + curbweight + highwaympg + horsepower + wheelbase + carheight, data = transfdf)$ de esta forma la función step() en R se utiliza en combinación con el modelo lineal lm() para realizar una selección de características automática y encontrar el mejor modelo lineal que explique la varianza de una variable respuesta (en este caso price). Su objetivo es identificar un subconjunto de predictores (variables independientes) que proporcionen la mejor capacidad predictiva en función de algún criterio de selección, como el AIC (Criterio de Información de Akaike) o el BIC (Criterio de Información Bayesiano).

Después de correr la función step(lm()) en R, se obtuvo que el mejor modelo predictor para la variable price era únicamente tomando en cuenta a la variable curbweight y la variable horsepower como variables independientes.

Al analizar este modelo se obtiene una media de residuos sumamente cercana a 0, los residuos son simétricos en valores mínimos y máximos respecto a la media, se obtiene un F-estadístico alto, que es lo que buscamos, por otra parte el p-value es menor a 0.05, por lo que el modelo es significante para explicar la varianza de la variable price. Finalmente encontramos que la varianza explicada por este modelo es del 85.5 %, un valor alto para nuestro caso.

3. Conclusiones

En conclusión, basándonos en el análisis estadístico realizado, hemos identificado las siguientes conclusiones clave:

- Variables Significativas: Después de una cuidadosa selección, hemos determinado que las variables más significativas para predecir el precio de los automóviles en el mercado estadounidense son curbweight y horsepower. Estas dos variables tienen un peso considerable en la explicación de la varianza en el precio.
- Importancia de Symboling: La variable categórica Symboling también ha demostrado ser relevante, ya que existe una diferencia significativa entre sus clases. Esto sugiere que el nivel de seguridad percibido de un automóvil, representado por Symboling, influye en su precio en el mercado estadounidense.
- Modelo Optimo: El mejor modelo identificado combina curbweight y horsepower, logrando explicar el 85 % de la varianza en el precio. Esto indica que esta combinación de variables es altamente predictiva y podría ser la base para estrategias de fijación de precios y toma de decisiones en la producción de automóviles.

En resumen, este análisis estadístico proporciona información valiosa a la empresa automovilística china interesada en ingresar al mercado estadounidense. Conocer las variables más influyentes en el precio de los

automóviles les permitirá tomar decisiones estratégicas informadas y aumentar sus posibilidades de éxito en un mercado altamente competitivo.

4. Anexos

Para acceder a la carpeta donde se encuentra el análisis estadístico hecho en R markdown y también en formato pdf y a la base de datos, acceder a la siguiente liga: `https://drive.google.com/drive/folders/1lNanOw_mqOtVQmxLpENUFXXWWu5bYjPa?usp=sharing`