Laboratorio 3

Juan Sebastian Cortes Chia

Jhonatan Rios Tapiero

Seminario De Big Data

ELIAS BUITRAGO BOLIVAR

Universidad ECCI

Facultad de ingeniera

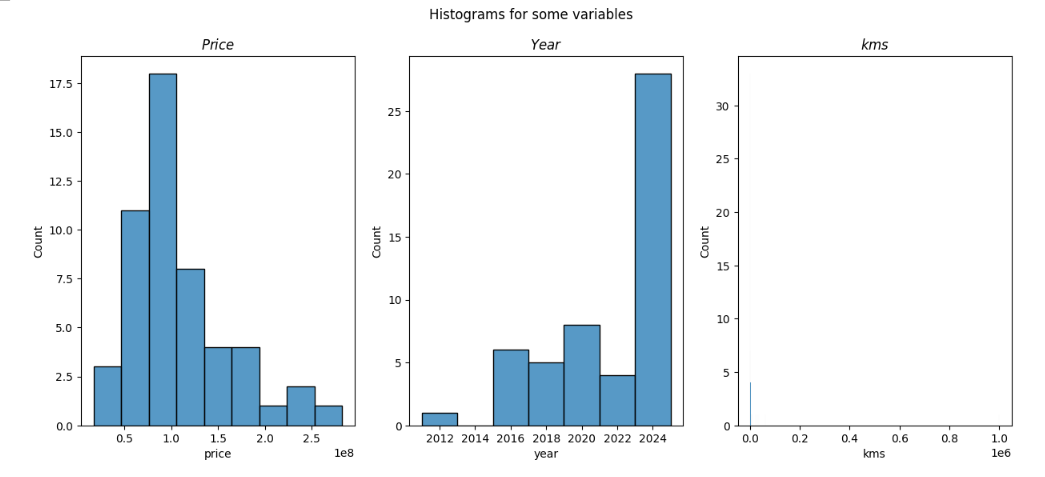
Desarrollo

En el desarrollo de este laboratorio se realiza análisis acerca del webscraping con la marca Mazda con el modelo 3, en este análisis podemos ver factores determinantes como precios, modelos, etc.

En la primera grafica podemos ver los siguientes datos que se encuentran almacenados en una tabla, esto previamente se transformó y se limpiaron los datos para obtener esta data:



Luego de realizar el proceso de transformación y limpieza de los datos, se empezó a realizar su data exploration, creando inicialmente histogramas con las variables Price,Year, kms, como se visualiza en la siguiente imagen:

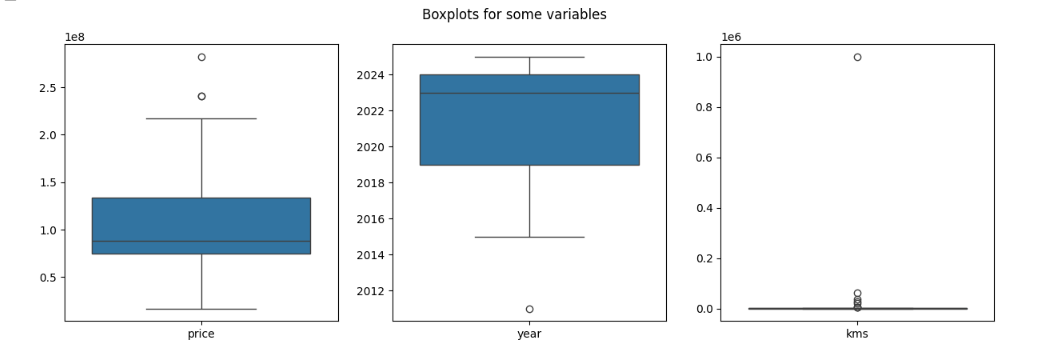


De lo cual podemos inferir lo siguiente de acuerdo a las variables del histograma:

Precio: El histograma muestra una tendencia en la que los precios más bajos son más frecuentes. Esto indica que hay más elementos (probablemente vehículos) con precios bajos en el conjunto de datos. La escala del eje x sugiere que los precios están en alguna unidad multiplicada por 108

Año: Este histograma muestra la distribución de los años. Hay picos notables en ciertos años, lo que sugiere que esos años tienen más ocurrencias dentro del conjunto de datos. Esto podría indicar los años en los que se vendieron o fabricaron más vehículos.

Kilómetros (Kms): Este histograma muestra que los vehículos con menor kilometraje son más frecuentes en el conjunto de datos. La escala del eje x sugiere que la distancia recorrida está en kilómetros multiplicada por 106



En el análisis de los boxplots para las variables Price, year y kms. Se puede inferir lo siguiente:

Price

* Rango intercuartílico (IQR): La mayor parte de los datos de price se encuentra entre aproximadamente 5x y 1.5 x
* Mediana: La línea dentro de la caja indica la mediana, que está cerca de 1x. Esto sugiere que la mitad de los datos de price están por debajo de esta cantidad y la otra mitad por encima.
* Valores Atípicos: Se observan varios puntos por encima de 2.5 x , por lo que podemos inferir que estos valores son significativamente mayores que el resto de los datos y pueden influir en el análisis realizado.

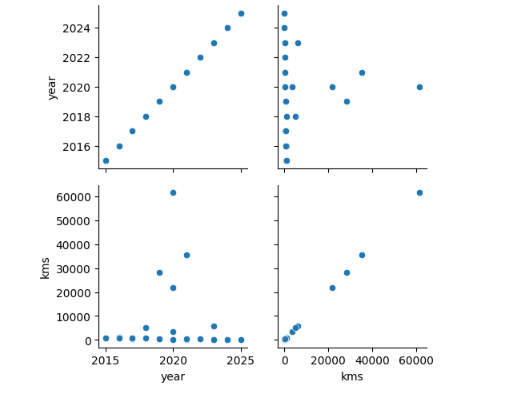
Year

* Rango intercuartílico (IQR): Los datos de year se concentran principalmente entre los años 2018 y 2023.
* Mediana: La mediana se sitúa cerca del año 2022, indicando que la mitad de los datos son de años anteriores a 2022 y la otra mitad son de años posteriores.
* Valores atípicos: Hay un punto atípico por debajo del año 2014. Este valor es significativamente menor que el resto de los datos.

Kms

* Rango intercuartílico (IQR): La mayoría de los datos de kms están cerca de cero, indicando que una gran parte de los registros tienen valores bajos en kilómetros.
* Mediana: La mediana está muy cerca de cero, lo que sugiere que al menos el 50% de los datos tienen valores extremadamente bajos de kilómetros.
* Valores atípicos: Se observan varios puntos atípicos, con un valor extremadamente alto alrededor de 1x. Este valor es significativamente mayor que el resto y debemos tener en cuenta que debe ser analizado cuidadosamente para obtener mejores resultados en nuestro análisis.

Luego de realizar el análisis en los boxplots, procedemos a realizar el análisis exploratorio sobre la data que estamos revisando:



De esta grafica podemos inferir lo siguiente:

Relación entre year y year:

La gráfica diagonal de arriba a la izquierda muestra una línea perfecta, lo cual es esperado ya que compara la misma variable consigo misma.

Relación entre year y kms:

Las gráficas superior derecha e inferior izquierda muestran la relación entre year y kms.

No parece haber una relación lineal clara entre year y kms. Los datos de kilómetros (kms) están dispersos a lo largo de los años sin un patrón evidente.

La mayoría de los datos de kms están concentrados en valores bajos, independientemente del año. Sin embargo, hay algunos puntos con valores de kms más altos, que se dispersan a lo largo de los años.

Hay un valor atípico significativo alrededor de los años 2020-2021 con un kms muy alto, lo que podría requerir una investigación más profunda.

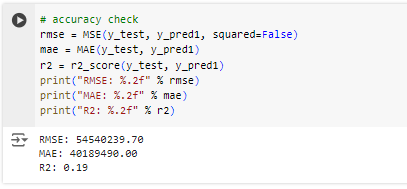
Relación entre kms y kms:

La gráfica diagonal de abajo a la derecha muestra una línea perfecta, lo cual es esperado ya que compara la misma variable consigo misma.

Evaluación de los modelos:

Multivariate lineal regression

De este modelo podemos ver que nos da los siguientes valores acerca del rendimiento que se obtuvo:



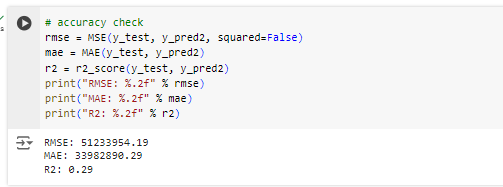
Al inferir estos valores podemos ver lo siguiente:

RMSE (54,540,239.70): Al tener un RMSE alto, podemos determinar que el modelo evaluado tiene errores de predicción significativos. En este caso, el valor es bastante elevado, lo que indica que las predicciones del modelo están bastante alejadas de los valores reales en promedio.

MAE (40,189,490.00): Al tener un MAE alto, significa que las predicciones del modelo no son precisas. En este caso, el valor de MAE es también muy alto, lo que confirma que el modelo tiene un rendimiento deficiente en términos de precisión de predicción.

R² (0.19): En nuestro caso el R² es de 0.19 indica que solo el 19% de la variabilidad en los datos observados es explicada por el modelo. Esto sugiere que el modelo tiene un bajo poder explicativo y no es muy eficaz para predecir los valores.

Light GBM



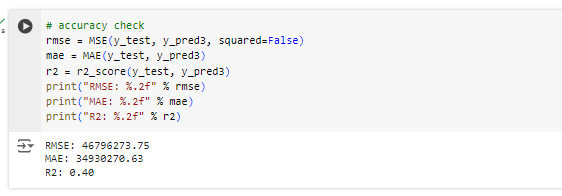
De este modelo podemos ver que nos da los siguientes valores acerca del rendimiento que se obtuvo:

RMSE (51233954.19): Un RMSE de 51233954.19 indica que el modelo tiene un error promedio de 51233954.19 en sus predicciones. Cuanto más pequeño sea el valor de RMSE, mejor será el rendimiento del modelo.

MAE: 33982890.29: Con este MAE que nos brinda el modelo, podemos determinar que las predicciones del modelo se desvían en 33982890.29 unidades, lo cual nos ayuda a saber que el modelo es eficiente ya que tiene un valor bajo.

R2(0.8929): El R2 que nos brinda el modelo, nos ayuda a validar que el modelo puede explicar el 89.29% de la variabilidad en los datos de prueba.

RandomForestRegressor



RMSE(46796273.75): Este valor proporcionado por el modelo lo que nos ayuda a validar es que nuestro modelo tiene un error promedio de 46796273.75 en sus predicciones. Y de acuerdo a la definición del RMSE cuanto más pequeño sea el valor de RMSE, mejor será el rendimiento del modelo.

MAE(34930270.63): Este valor significa que significa que, en promedio, las predicciones del modelo se desvían en 34930270.63 unidades, lo cual no es optimo para el modelo, debido a que el valor tiene que ser bajo para que sea optimo.

R2(-0.40): Este valor tan bajo lo que indica es que el modelo no se ajusta bien a los datos y este puede tener un mal rendimiento en este contexto donde está siendo evaluado para ver si es optimo o no.