*Assignment 1: - Data Visualization on Cars Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| *Darshan Ruparel*  *AML-2103 Visualization for AI and ML,*  *Artificial Intelligence and Machine Learning,*  *Lambton College,*  *Toronto, Canada* | *Nisargkumar K. Patel*  *AML-2103 Visualization for AI and ML,*  *Artificial Intelligence and Machine Learning,*  *Lambton College,*  *Toronto, Canada* | *Anusha Gundeti*  *AML-2103 Visualization for AI and ML,*  *Artificial Intelligence and Machine Learning,*  *Lambton College,*  *Toronto, Canada* | *Gursimran Kaur*  *AML-2103 Visualization for AI and ML,*  *Artificial Intelligence and Machine Learning,*  *Lambton College,*  *Toronto, Canada* |

**Abstract:**

**Purpose –** This report details our experience with EDA on the Cars dataset, where the focus is on getting as many insights as possible by visualizing data. The data goes through various processes of data cleaning before visualization, and each step is discussed thoroughly later in the report. The final goal is to prepare the dataset for machine learning models by gathering intel from data visualization. This task also involves handling interactive graphs and storing/displaying images in various formats.

**Design/approach:**

The assignment asks for several methods that need to be completed, including:

1. Removing Duplicate Date.
2. Missing/Null value handling
3. Outlier detection
4. Creation of interactive graph

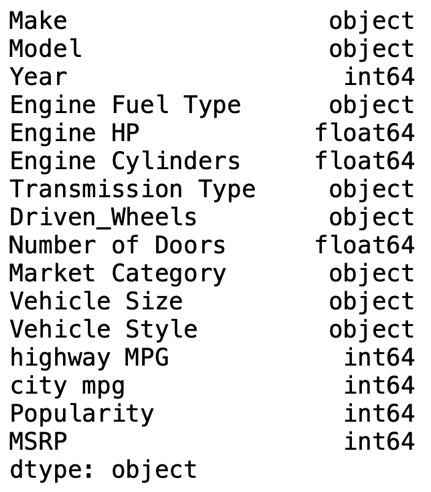
**Keywords:** Exploratory Data Analysis (EDA), Matplotlib, Seaborn, Plotly, Preprocessing, Visualization, Duplicated data, Null value handling, Outlier detection, Interactive graphs.

1. **INTRODUCTION**

Data Visualization has become essential for businesses in every industry to gain insight into trends and patterns that can improve decision-making. The dataset contains various features which affect the car price, and we can find which factor is majorly correlated with the car price. Python libraries offer multiple features that make the exploratory data analysis (EDA) process more efficient and effective. Our analysis includes the process of removing duplicate data, and missing values, identifying outliers, and creating interactive charting using the "Go" method from Plotly. The outcome of the analysis provides valuable insight into the dataset, including correlations between diverse factors which might have an impact on the car's price.

1. **LITERATURE REVIRW**

The dataset provided includes information on various attributes of vehicles, as shown below figure with their datatypes.



The "Make" and "Model" features provide information about the car manufacturer and the model of the vehicle. The "Year" indicates the year in which the car was released. The dataset provided includes information on various attributes of vehicles, as shown below figure with their datatypes.

The "Engine fuel type" provides information on the types of fuel used by the car's engine. The "Engine HP" and "engine cylinders" includes information about the engine's power and the number of cylinders that the model has.

The "Transmission type" say which type of transmission is used by that car. The "driven wheels" factor says which formation of the drive that car uses.

The "Number of doors" says the number of doors that the vehicle has, and information about the type of market segment the car is intended for is provided by the "Market category" feature. The "vehicle size" element provides details about the car's dimensions, whereas the "vehicle style" attribute offers details on the car's body design.

The "highway MPG" and "city MPG" describes the fuel efficiency of the vehicle in miles per gallon for highway and city driving conditions, respectively.

The number of times the vehicle has been viewed on Edmunds.com, which is how the "popularity" attribute measures the popularity of the car, provides information about the popularity of the car.

Finally, the "MSRP" provides intel on the suggested retail price of the vehicle.

1. **HANDELING MISSING VALUE**

The Dataset has many missing values and have handle each missing value care fully

Table

Description automatically generated

Here we can see that total 5 features have missing values in.

We have handled missing value from each column separately.

**1 Handling missing value in Engine Fuel type**

As we can see in above picture that this column has only 3 missing value so we checked which car has missing value in this feature.

**Table

Description automatically generated**

we can see that we have nan values in Suzuki’s model so checked for other Suzuki model and everyone has similar fuel type which is “regular unleaded”, so we filled nan value with that



**2 Removing nulls from Engine\_hp**

Here in this column, we had many nulls, so we have created separate data frames for each model from make columns such as ['FIAT', 'Lincoln', 'Ford', 'Honda', 'Mitsubishi,' 'Chevrolet', 'Nissan', 'Mercedes-Benz', 'Tesla,' 'Toyota,' 'Kia'] we have found average for Engine\_hp feature for each of newly created data frame and replaced the nan with an appropriate average in the primary data frame.

For ex: if we have null three nulls in engine\_hp in fiat, then we find the average for the rest of the fiat model in engine\_hp and replace that null value with the standard.

**Creation of data frame as per Make**

Text

Description automatically generated

**Finding mean for each engine\_hp column from each data frame**

Text

Description automatically generated

**Replacing mean with appropriate mean in main data frame**

**Table

Description automatically generated with medium confidence**

Here we had a problem in tesla car because not a single tesla car an engine horse power in original data frame as we can see in below picture.

Table

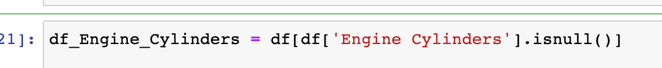
Description automatically generated

So, we replaced the nan value with total average of the column



3 **Handling Missing values in Engine\_cylinders**

First we created separate data frame which has nan in engine cylinders column

****

Then we checked which car has nan in Engine Cylinder and what was their engine type

Table

Description automatically generated

And we got to know that either electric cars of ['Chevrolet', 'Volkswagen', 'Mitsubishi', 'Toyota'] or Mazda.

Electric car uses electricity so there would be no use of cylinders in it And Mazda rx-7 and rx-8 are sports cars that were designed with unique rotary engines, which are different from traditional piston engines, so it is not using piston and does not have any cylinder

So, we replace nan with 0 in this entire column



**4 Removing Null Values from Numbers of doors Column.**

Here again we created a new data frame where number of doors has nan value and check which are the cars that have nan in number of doors column.

# A picture containing text Description automatically generated

We got to know that only Ferrari and tesla had missing values in a number of door columns, and we searched for a particular model from both cars and got to know that Ferrari mostly has two doors in every model, and tesla has 4 in most of the models, but still, we find a mode in a number of doors column where the model is Ferrari and Tesla we found out that in this data set too the most frequent number is 2 and 4 for Ferrari and tesla, so we replaced null values with respective mode.

Graphical user interface, text

Description automatically generated with medium confidence

**5 Removing Nulls from Market Category column**

Here we thought of dropping the Market Category column because it has multiple unique values and it will take so much time to do the analysis for each category, so we thought it would be best if we drop the entire column.

**Checking nulls after handling missing values from original data frame**

Table

Description automatically generated

Now we can see that we have a clean data set with no missing values in it.

1. **DATA DUPLICATION HANDLING**

Data duplication has become very common for each data, which can lead to biased or incorrect results if not handled properly.

Data duplication can occur in datasets when multiple observations have the same values for all attributes.

First, we checked how many duplicates we had in our cleaned dataset.Graphical user interface, text, application

Description automatically generated

We can see that 806 rows have duplicate values. so we removed it from our data to avoid any false prediction.

Graphical user interface, text, application

Description automatically generated

It is crucial to remember that duplicate handling must be done cautiously because removing duplicates can result in data loss and isn't always necessary for the research subject being examined. Documenting the duplicate handling procedure and any potential biases or restrictions the procedure may have introduced is also crucial.

1. **MEASURE OF CENTRAL TENDENCY AND DISPERSION**

The Center or typical value of a dataset is described using central tendency measures like the mean, median, and mode. Adding together all the values in a dataset and dividing by the total number of values yields the mean. The mode is the value that appears the most frequently in the dataset, and the median is the midpoint value when the values are ordered in either ascending or descending order.

To describe the spread or variability of a dataset, dispersion measures like the range and standard deviation are used. The range represents the difference between the dataset's highest and lowest values. How far apart the values are from the mean is indicated by the standard deviation. When the standard deviation is smaller, the values are more closely packed around the mean; when it is bigger, the values are more widely dispersed.

Here we have created separate Data frame for numeric columns and measure the Central Tendency and Dispersion.

**Mean**

Text

Description automatically generated

**Median**

Text

Description automatically generated

**Mode**

Text

Description automatically generated

**Standard Deviation**

Text

Description automatically generated

**Range**

**Text

Description automatically generated**

1. **OUTLIER DETECTION AND GRAPHICAL PRESENTATION (GRAPH FROM DATA)**

Data visualization and outlier detection techniques applied to the 'Engine HP', 'highway MPG', 'city mpg', 'MSRP' columns in a dataset using Box plot method there are ways other ways too to detect outliers but we have used box plot as it gives proper representation of outliers.

Graphical user interface, text, application

Description automatically generated

In the below picture we can the outliers in Engine HP and Highway Mpg Column

Chart, box and whisker chart

Description automatically generated

Below figure represent the Outlier in City MPG and MSRP column

Chart, box and whisker chart

Description automatically generated

We have also find out that which are the datapoints lien in between the IQR range

Graphical user interface, text, application, email

Description automatically generated

From above numbers we can say that if we remove the outlier there is a high chance of data loss so instead, we thought of doing log transform and analyze the data with the help of histogram

Graphical user interface, text, application

Description automatically generated

**Chart, bar chart

Description automatically generated**

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

We thought log transformation would be better option for Engine HP as it rectifies the skewness

Since in this assignment we are not focusing on outlier removal and just focused on visualization we will not handle it.

1. **COMPARE DIFFERENT FEATURES WITH RESPECT TO FREQUENCY COUNTS.**

We plot histogram for each column to visualize the frequency count.

The below code will draw histogram for each column living model column alone because it has too models so we leave that column.

**Graphical user interface, text, application

Description automatically generated**

**Text

Description automatically generated**

Here in above code, we do a value count in Make column and plot bar graph for each manufacturer the outcome of the bar graph is shown in below figure

Chart, histogram

Description automatically generated

We can say that the maximum cars are from Chevrolet and minimum are from Spyker in given Data.

**Comparing various feature against each other**

**Text

Description automatically generated**

Here we plot the scatter plot for Engine HP and MSRP column to find the relation between them

Chart, scatter chart

Description automatically generated

We can say that maximum datapoints are between 100 to 500 HP and as HP increases the price of the car also increase.

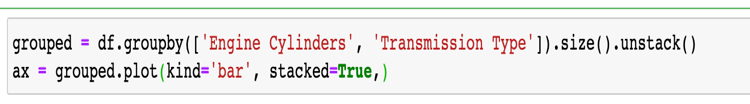
Text

Description automatically generated

The above Code will visualize the scatter plot for Highway MPG and City MPG

Chart, scatter chart

Description automatically generated



The above code, will plot Stacked bar graph in Engine Cylinders and Transmission Type column

Chart, bar chart

Description automatically generated

1. **DISPLAY YOUR FINDINGS WITH VARIOUS GRAPHS LIKE**

**Heat Map**

**A picture containing text

Description automatically generated**

**Graphical user interface, application, Teams

Description automatically generated**We can see from above graph that price of car has correlation with Engine HP and number of cylinders and High anticorrelation between Engine Cylinders and highway MPG

**Distribution in Cars from year column**

**Graphical user interface, text, application

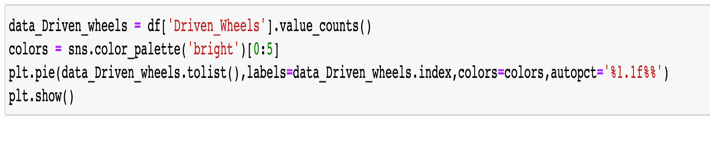
Description automatically generated**

**Outcome:**

Chart, histogram

Description automatically generated

From above Graph we can say that that there is significant demand in cars from 2016 onwards.



The above Code will plot a pie chart in Driven Wheels column

Chart, pie chart

Description automatically generated

From above graph we can say that maximum cars have front wheel drive

Graphical user interface, text, application

Description automatically generated

The above Code will plot a pie chart in vehicle\_size column.

Chart, pie chart

Description automatically generated

From the above graph we can say that the maximum cars from the data are of either compact or midsize format.

1. **CREATE STATIC GRAPHS AND INTERACTIVE GRAPHS (DYNAMIC) LEVERAGING MATPLOTLIB AND SEABORN**

The code demonstrates how to create interactive scatter plots using Matplotlib and Plotly, respectively.

Graphical user interface, text, application, email

Description automatically generated

The Above code will plot a interactive plot using plotly between MSRP and Year column in form scatter plot

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

The above code will plot interactive pie chart on vehicle style column, and it will represent each style with its proportion in data.

Chart, sunburst chart

Description automatically generated

1. **SAVE GRAPHS IN VARIOUS FORMATS (.JPG, .PNG, HTML, JSON, BYTE STRING ETC.)**

The code demonstrates how to save a interactive plot in various formats using **import plotly.io as pio**

Text

Description automatically generated

Below code will use Savefig function from matplotlib and save the image in different format.

Text

Description automatically generated

1. **ENCAPSULATE AND SHOW IMAGE INTO HTML PAGE**

The bar graph created by this code is saved as a.png file. After that, it generates a piece of HTML code that shows the plot and saves it as a.html file. The HTML code snippet is then shown by the code in the Jupyter notebook.Text

Description automatically generated

1. **Implementing Model**

Here we have implemented RandomForestRegressor model on numeric from below code.

Graphical user interface, text, application, email

Description automatically generated

Here We have multiple independent variable and one dependant variable which MSRP to predict the price.

And our model has provided the accuracy of .92

1. **CONCLUSION**

In conclusion, data visualization is an important component of data analysis that aids in offering insightful information about the data. Powerful Python tools like Matplotlib and Seaborn let us build a variety of static and dynamic visualizations. Bar charts, pie charts, histograms, scatter plots, heat maps, and box plots are just a few examples of the different graphs and charts we can make. These visualizations are available for saving in a number of file types, including jpg, png, pdf, SVG, and tif. For better presentation, we may also incorporate these photographs into an HTML page. We can effectively convey our findings to stakeholders and make data-driven decisions by utilizing these tools and strategies.

1. References

*Text properties and layout#*. Text properties and layout - Matplotlib 3.7.1 documentation. (n.d.). Retrieved March 17, 2023, from <https://matplotlib.org/stable/tutorials/text/text_props.html>

*Text properties and layout#*. Text properties and layout - Matplotlib 3.7.1 documentation. (n.d.). Retrieved March 17, 2023, from  
[https://matplotlib.org/stable/tutorials/text/text\_props.html](https://seaborn.pydata.org/)

*Low-code data app development*. Plotly. (n.d.). Retrieved March 17, 2023, from <https://plotly.com/>