DAI – 101 Assignment – Report

~ Utkarsh Kumar 23114101

Initial Analysis

- We are going to use a dataset involving details about various cars brands and their models. The dataset is : cars.csv

- First, we read the csv with the help of pandas using read_csv() function. Then we inspected the dataset by various functions involving describe(), info(), shape, etc.

- We saw that there were a lot many columns, some seemed to be irrelevant also, so we identified and dropped those columns.

- Next, we checked for missing values, the column “Modification” had more than 50% missing values, so we decided to drop it. This way we handled the missing values from the dataset.

- We checked for duplicate rows in the data but luckily there were no duplicate records. We saw that the Country column didn’t match with its content, so we changed the column name to Region.

- A review using value_counts() helped in understanding the distribution of values across various columns.

Handling Outliers

- Now we are starting to search for outliers in our Price column. We plotted a displot() and we saw that the curve is not a normal curve, thereby z-score method to remove outliers is not applicable here.

- A subsequent boxplot using the IQR method showed that all data points were within the upper and lower limits, indicating no significant outliers in the Price column.

- Hence there are no outliers in the Price column.

EDA Univariate Analysis

- We started plotting a countplot() on Brand column. And we saw that all the companies were in almost equal amount. Ferrari has the maximum count while Lamborghini has the minimum count.

- A pie chart for the Region column indicated an even representation across different regions.

- Calculations for the standard deviation of Mileage and the skewness of Fuel_Efficiency were performed.

- Also making histogram of Mileage, displot of Fuel_Efficiency , boxplot of Horsepower. These plots also show that data is evenly distributed among all fields.

EDA Bivariate / Multivariate Analysis

- Numerical vs Numerical Analysis = We printed the correlation table among various columns like Price, Mileage, Fuel_Efficiency, Horsepower. Also plotting heatmap for the same.

  Positive Correlations: Price & Mileage, Price & Fuel_Efficiency, Price & Horsepower

  Negative Correlations: Mileage & Fuel_Efficiency, Mileage & Horsepower, Fuel_Efficiency & Horsepower

  Scatterplot is also a good method comparing Mileage, Fuel_Efficiency, incorporating hue in form of Fuel_Type.

- Numerical vs Categorical Analysis = Barplot between Fuel_Type and Price , Boxplot between Fuel_Type & Horsepower with hue being the Condition , Violinplot between Condition & Price.

- Categorical vs Categorical Analysis = A heatmap displaying the relationship between Fuel_Type and Condition was constructed. A PairPlot of the dataset based on Fuel_Type helped visualize the interrelations.

- A pivot table between Top_Speed, Year, and Fuel_Type indicated that the maximum top speed remained fairly consistent across the years. A heatmap of this pivot table further emphasized this observation.

Results / Discussion

- **Initial Data Insights:**
  The first look at the data presented a need for cleaning; among other things, columns with a high percentage of missing values were better dropped.
- **Distribution And Outlier Treatment:**
  Univariate analysis has noted that most of the variables appear to be evenly distributed. No outliers were found in the dataset.
- **Correlation and Relationships:**
  Multivariate analysis involving correlation matrices and heat maps has yielded positive and negative relationships amongst the key variables. These can lead to predictive models.
- **Visualisations:**
  A wide range of countplots, pie charts, scatterplots, barplots, boxplots, and heat maps were utilized to provide a comprehensive picture of the dataset.

Conclusion

In this report, the complete analysis of the cars.csv dataset, including thorough data cleaning, outlier treatment, and exploratory data analysis, has taken place.

Further confirmations include : Cleaning-the dataset included removal of columns that do not contribute to analysis and handling of missing values. Univariate analysis and visualization indicated that data were well-distributed. In multivariate analysis, significant relationships between variables were established in readiness for additional predictive endeavours.