**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.
* We saw the value\_counts() for each column to get a rough idea about the data.

**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.
* We plotted boxplot() to use the IQR method , but here also we saw that the all datapoints were between the upper limit and the lower limit so no outliers were found.
* 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.
* Next, we created a pie chart based on Region, here also we saw an equal representation of each region.
* We also found the standard deviation of Mileage and skewness of Fuel\_Efficiency. 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 and 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 and Categorical Analysis = Barplot between Fuel\_Type and Price , Boxplot between Fuel\_Type & Horsepower with hue being the Condition , Violinplot between Condition & Price
* Categorical and Categorical Analysis = Heatmap between Fuel\_Type & Condition, pairplot of the full dataset based on Fuel\_Type
* We also made a pivot table between Top\_Speed , Year , Fuel\_Type which showed almost same max of Top\_Speed each year. We also printed the heatmap for it.