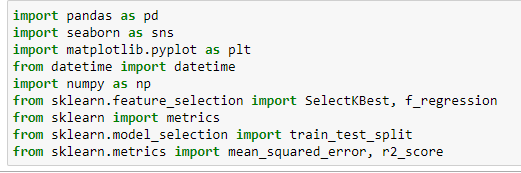
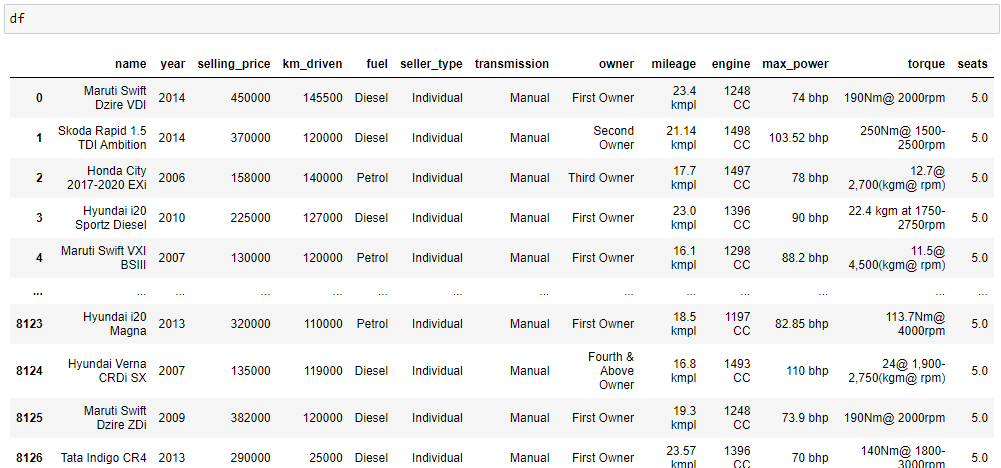
#### Introduction

#### This report presents a comprehensive analysis of a dataset containing details about used cars, including attributes such as manufacturing year, selling price, kilometers driven, fuel type, seller type, transmission type, ownership history, mileage, engine capacity, maximum power, and seating capacity. The objective is to explore and understand the factors influencing the selling price of used cars. Through data cleaning, transformation, and exploratory data analysis, key trends and correlations are identified, providing insights into the used car market. This analysis serves as a foundation for predictive modeling and further statistical analysis to better understand price determinants in the used car market.

**Importing Libraries**



#### Dataset Used

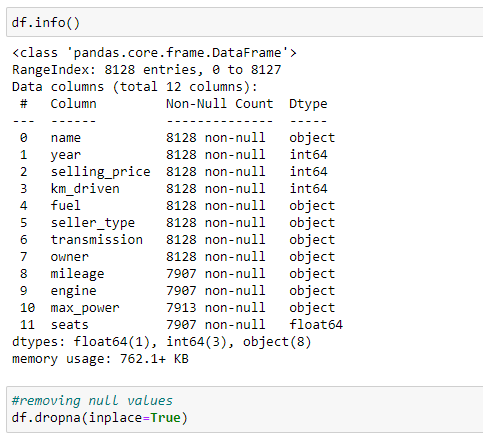
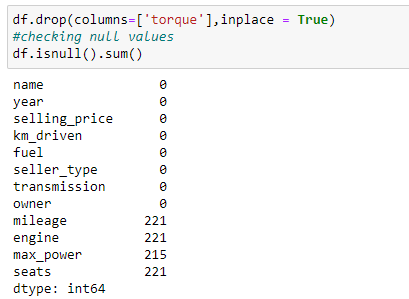


#### ****Data Overview****

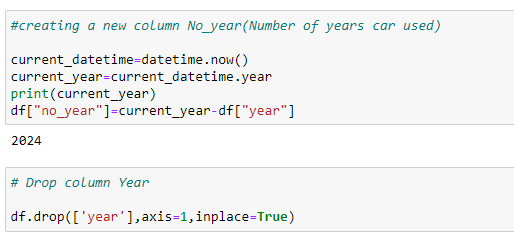
The dataset contains information about used cars, with attributes such as the car's name, manufacturing year, selling price, kilometers driven, fuel type, seller type, transmission type, ownership history, mileage, engine capacity, maximum power, and number of seats.

#### ****Data Cleaning and Preparation****

* **Removing Null Values**: We identified and removed null values present in the mileage, engine, max power, and seats columns.
* **Handling Duplicates**: We detected and eliminated duplicate records to ensure the dataset's integrity.



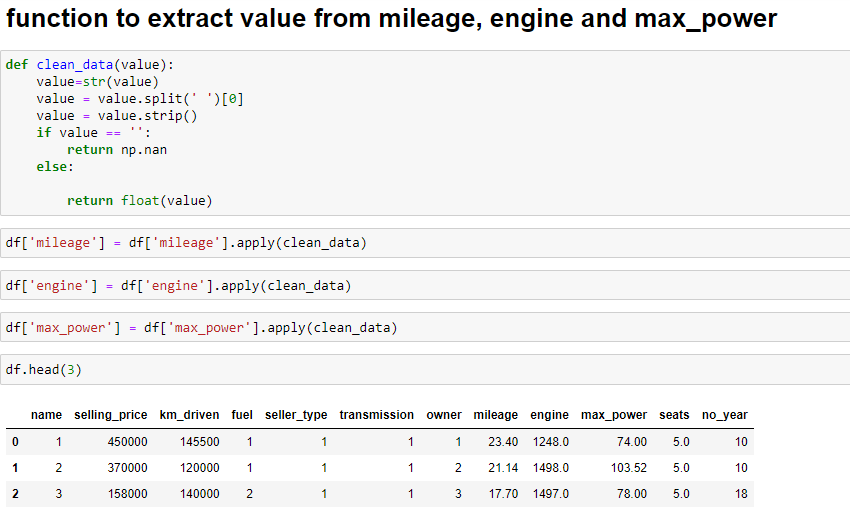
* **Extracting Useful Features**:
  + **No. of Years Used**: A new column, **“no\_year”,** was created to represent the number of years a car has been used, calculated by subtracting the manufacturing year from the current year (2024).
* **Encoding Categorical Variables**:
  + Fuel type, seller type, transmission type, owner type, and car brand names were encoded as numerical values for better analysis and model training.





#### ****Data Transformation****

* **Modifying Mileage, Engine, and Max Power**: These columns originally contained text values with units. We extracted the numerical part of these values for analysis.
* **Brand Name Extraction**: The car brand was extracted from the name column and encoded into numerical values.

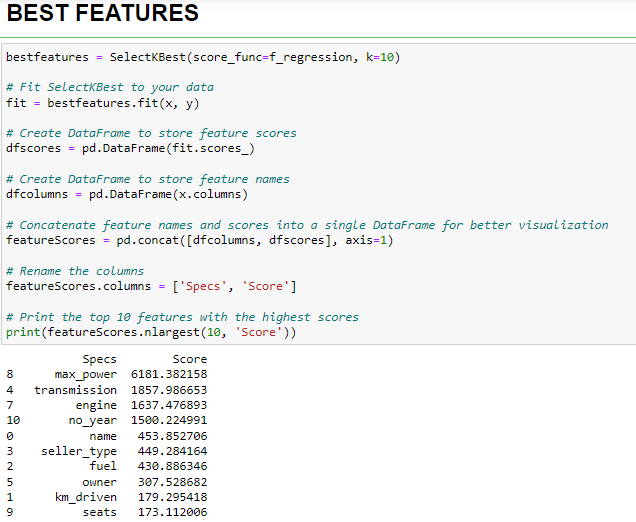


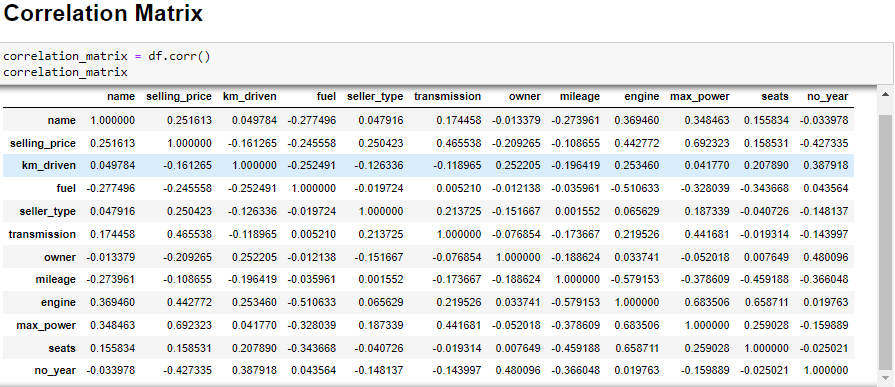
#### ****Exploratory Data Analysis****

* **Distributions**: We analyzed the distribution of key numerical features (selling price, kilometers driven, mileage, engine capacity, max power, and number of years used) using distribution plots.
  + **Selling Price**: The distribution is right-skewed, indicating that most cars are sold at a lower price.
  + **Kilometers Driven**: This distribution is also right-skewed, with most cars having fewer kilometers driven.
  + **Mileage**: There is a wide range in mileage, but a significant number of cars offer moderate mileage.
  + **Engine Capacity**: The distribution shows a concentration of cars with smaller engine sizes.
  + **Max Power**: Most cars have lower maximum power values.
  + **Number of Years Used**: The distribution shows a mix of newer and older cars.
  + **Box plots**: Box plots were used to identify outliers in the dataset.Significant outliers were present in the selling price, kilometers driven, mileage, engine capacity, and max power columns.

**Correlation Analysis -** A correlation matrix was generated to understand the relationships between numerical variables.

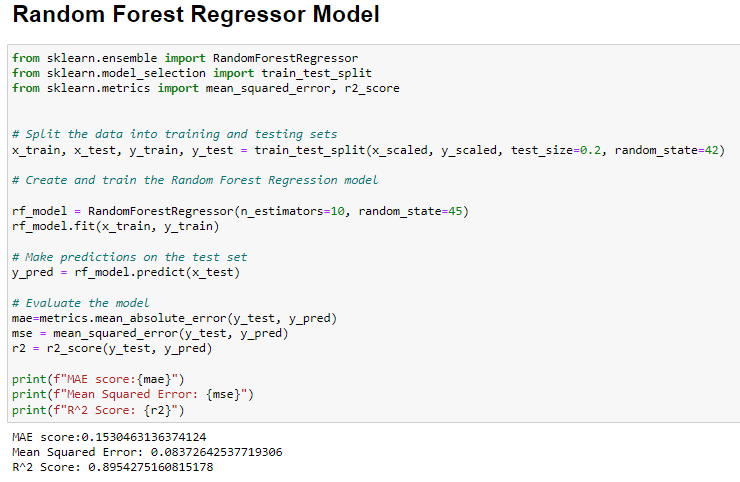
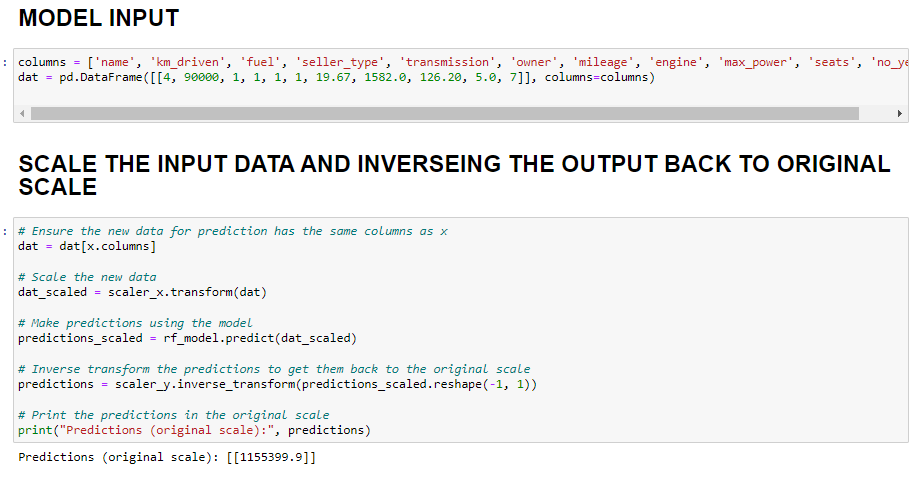
* **Positive Correlations**:
  + Selling price is positively correlated with engine capacity and max power.
  + Engine capacity and max power also show a strong positive correlation.
* **Negative Correlations**:
  + Selling price is negatively correlated with the number of years used, indicating that older cars tend to sell for less. Mileage has a negative correlation with engine capacity and max power.



**

#### ****Insights****

* **Factors Influencing Car Prices**: Engine capacity, max power, and the age of the car are significant factors affecting the selling price. Cars with higher engine capacity and max power tend to sell for more, while older cars generally fetch lower prices.
* **Model Training**: Implemented Random Forest regression model to predict the selling price and evaluate their performance using metrics like Mean Squared Error (MSE), Mean Absolute Error(MAE) and R-squared (R²) values.

****