

Car Price Prediction



May 2, 2024

# Introduction

## Problem Statement

XYZ Auto, a XYZ automobile manufacturer, is planning to expand into the U.S. market by establishing a local manufacturing facility and producing cars to compete with U.S. and European brands. To facilitate this entry, XYZ Auto has engaged an automobile consulting firm to analyze the factors that influence car pricing in the U.S., recognizing that these factors might differ significantly from those in the XYZ market. The company seeks to understand:

1. Which variables are significant in predicting the price of a car in the American market?
2. How effectively do these variables explain the variations in car prices?

The consulting firm has compiled a comprehensive dataset from various market surveys covering different car types across the U.S. market. This analysis aims to decode the dynamics of car pricing to aid XYZ Auto's strategic entry into this new market.

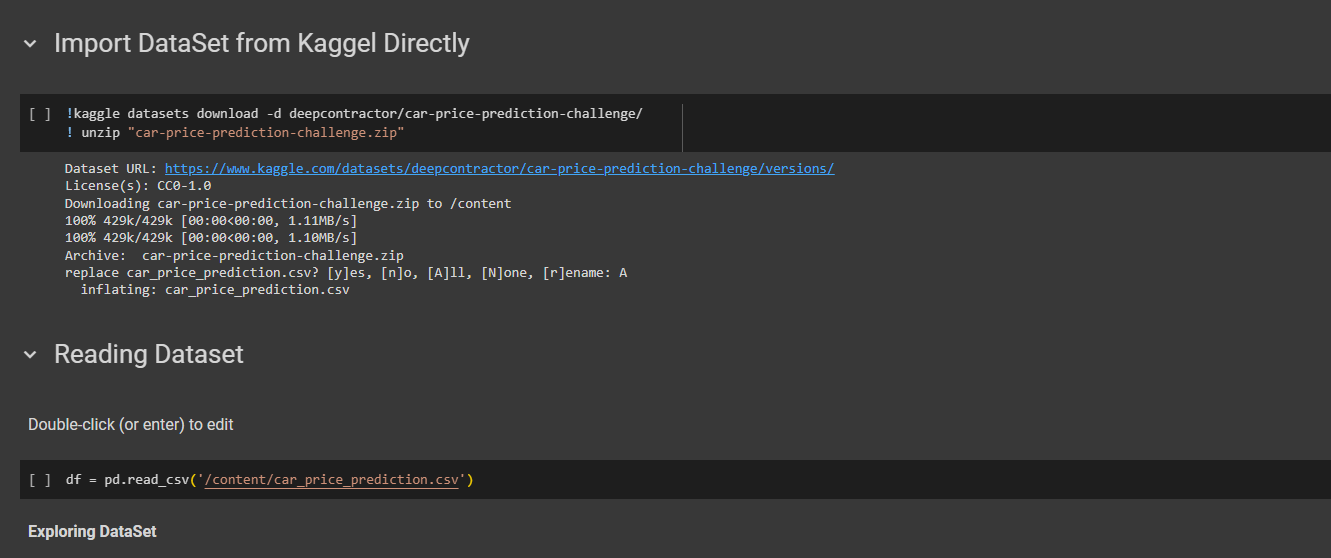
# Stage 1 : Exploratory Data Analysis Stage

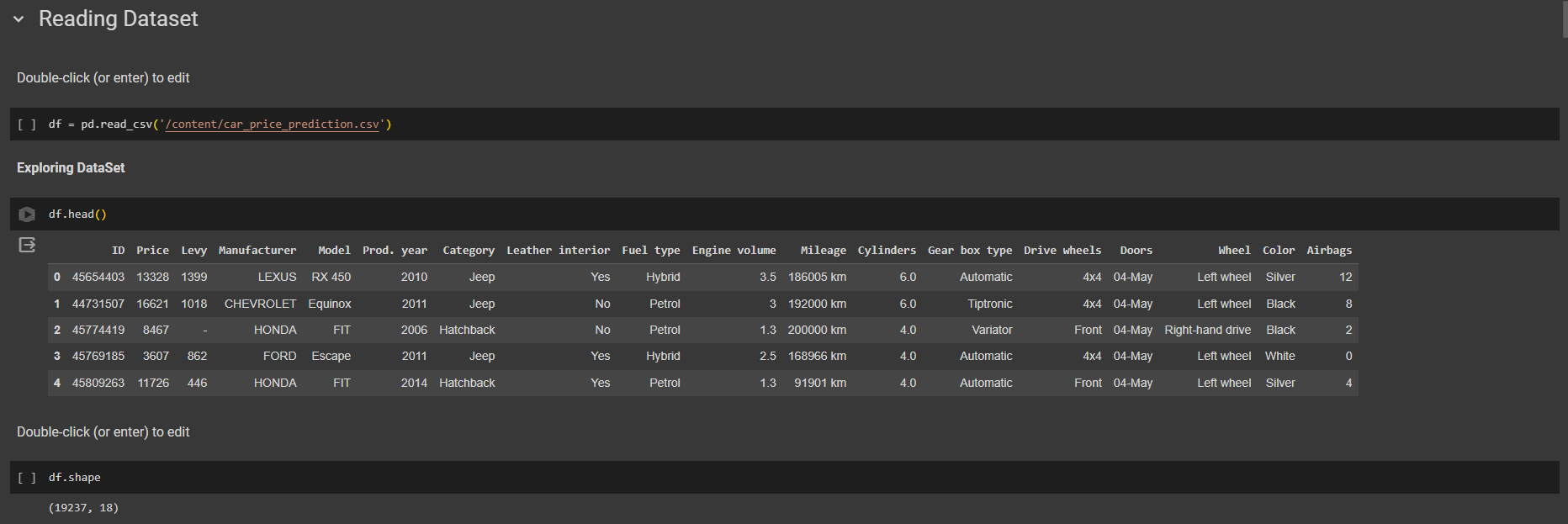
Stage 1 is most important preliminary stage and the purpose of exploratory data analysis is to obtain a thorough understanding of data, and inform about the choice of predictive analytics algorithms to be used, and expected performance of the software tool in real world settings.

# Exploratory Data Analysis

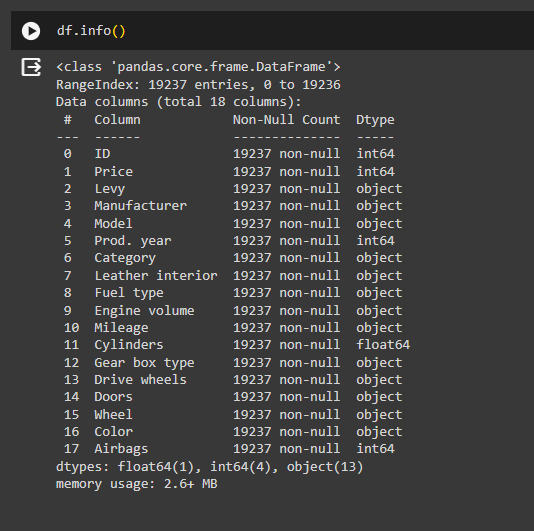
The first phase of the software development activity involved understanding the data, basic exploratory data analysis and visualization. Google Colab was chosen as the experimental environment. Before the exploratory data analysis can begin, some of steps required are:

## Download and Read the Dataset

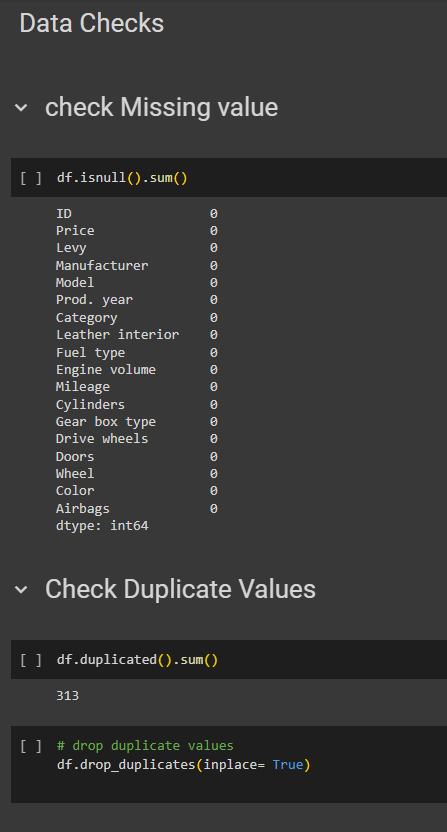




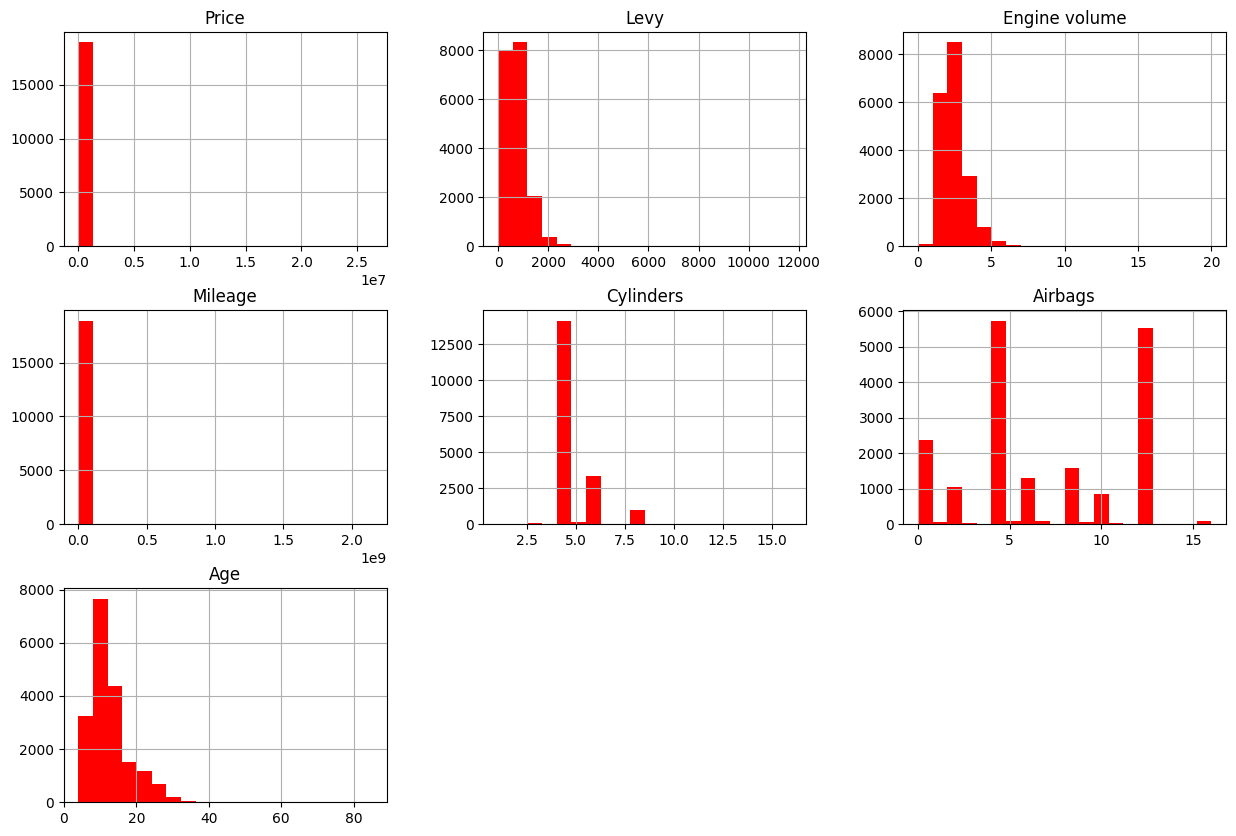
## Data variables Identification

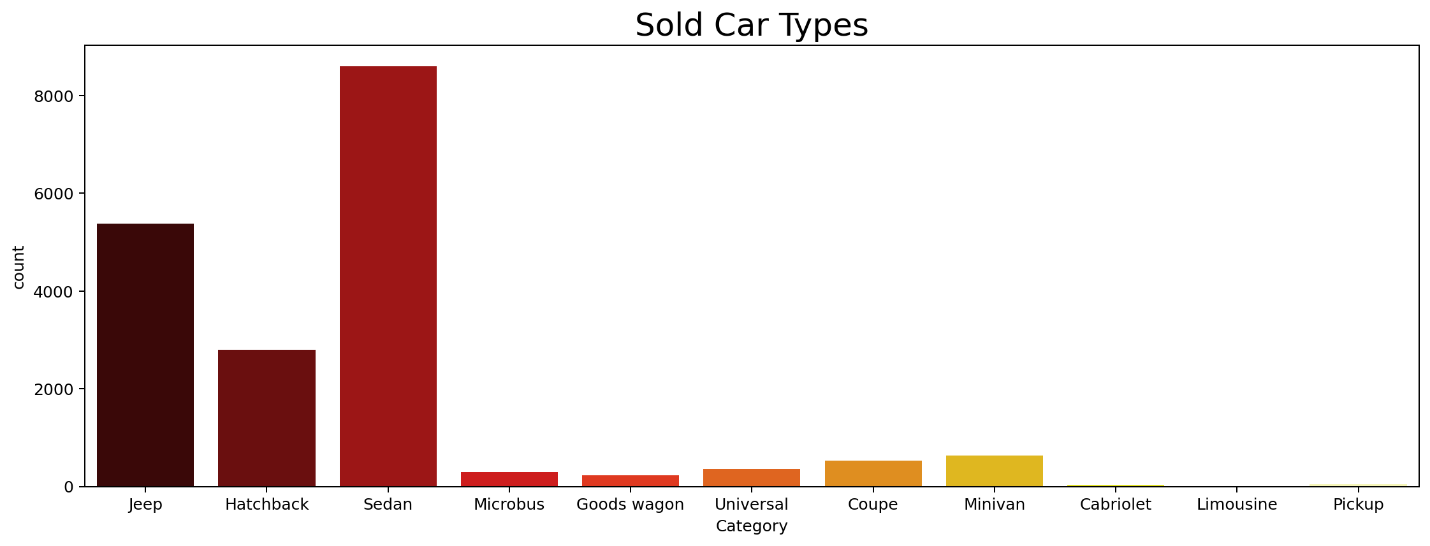


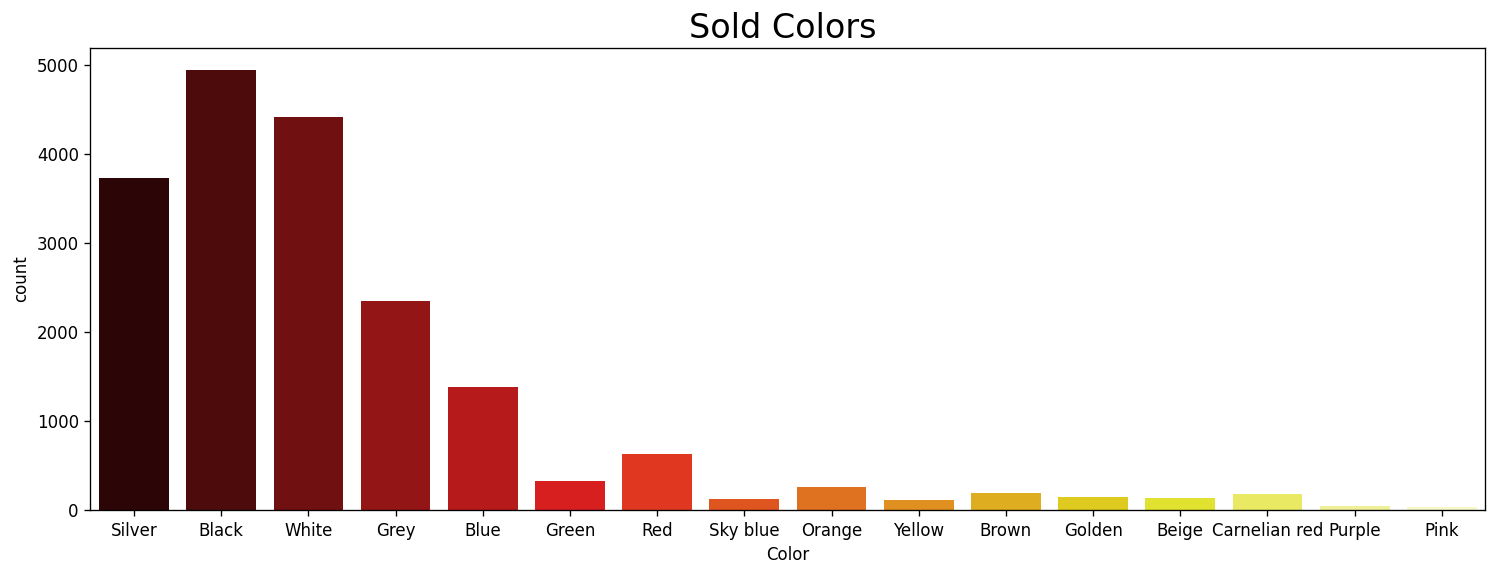
## Data set Check

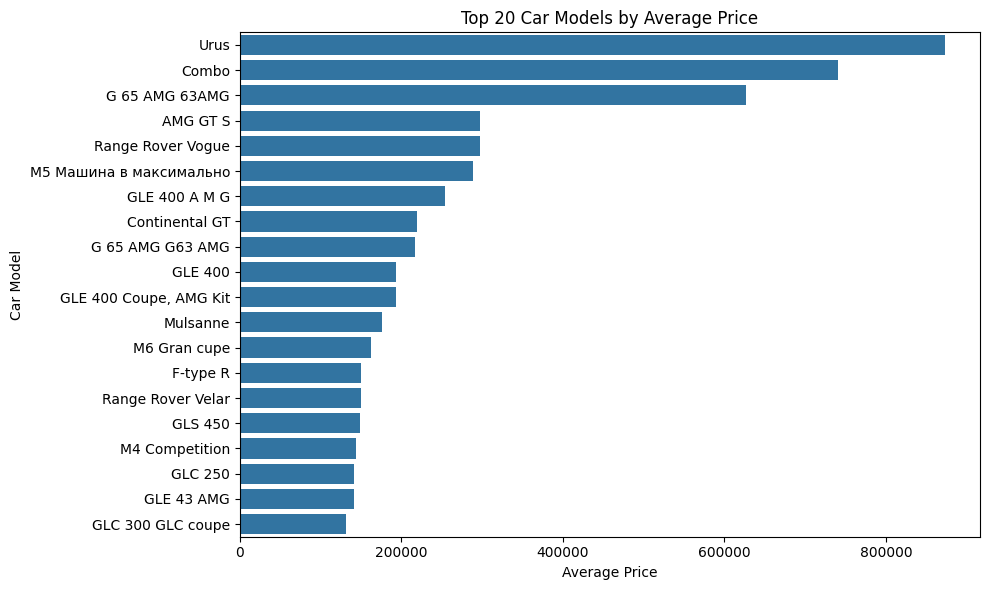


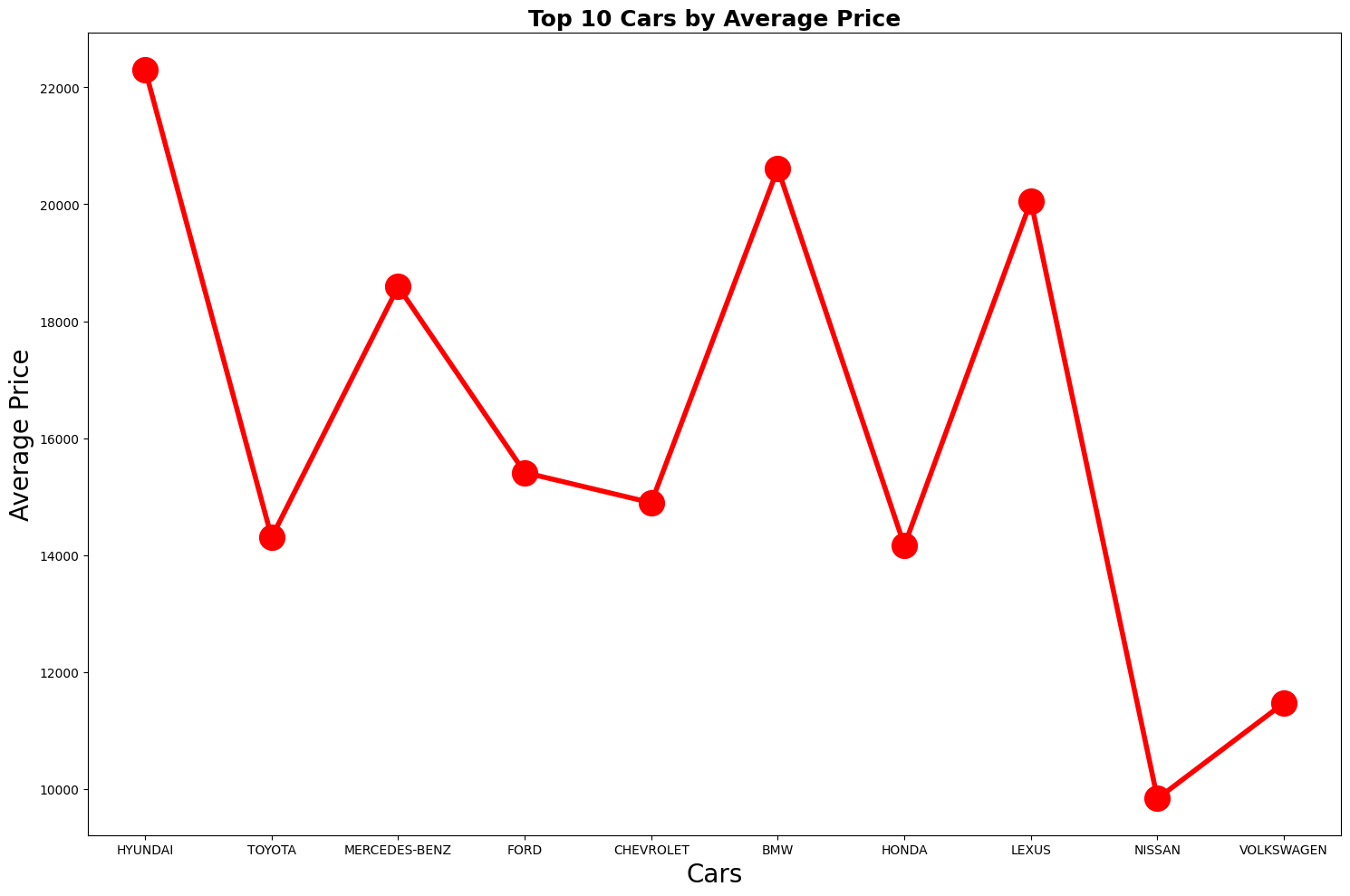
## Charts

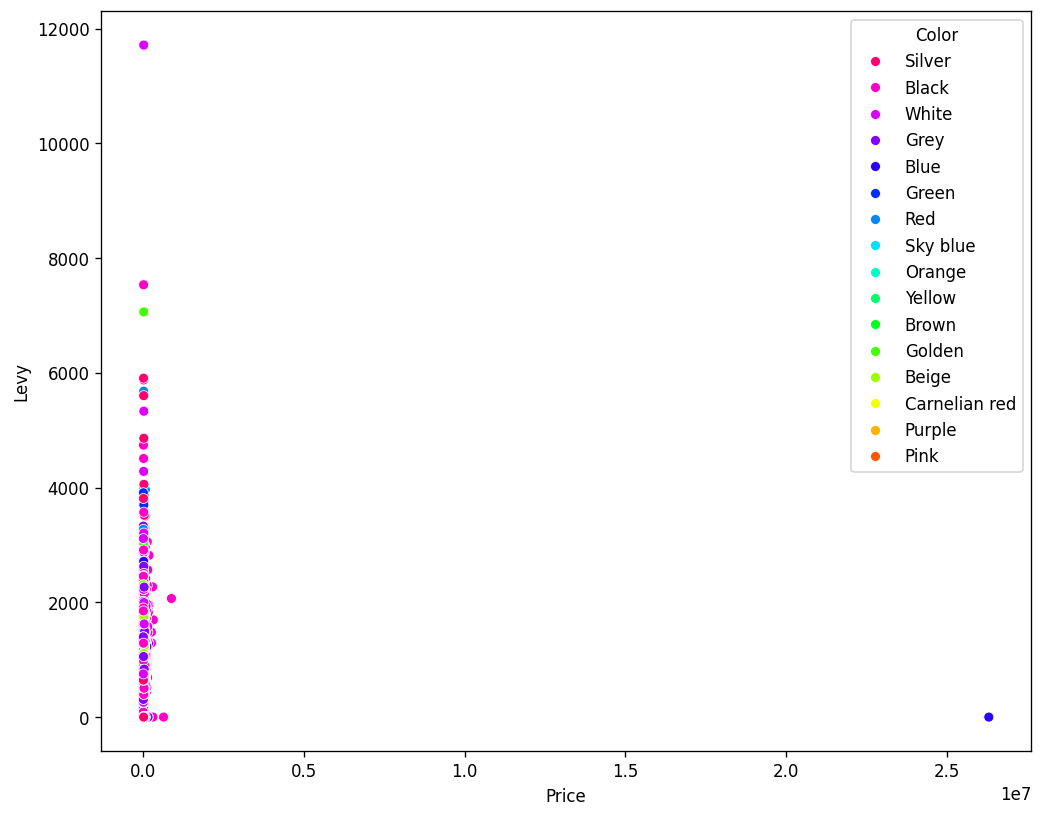




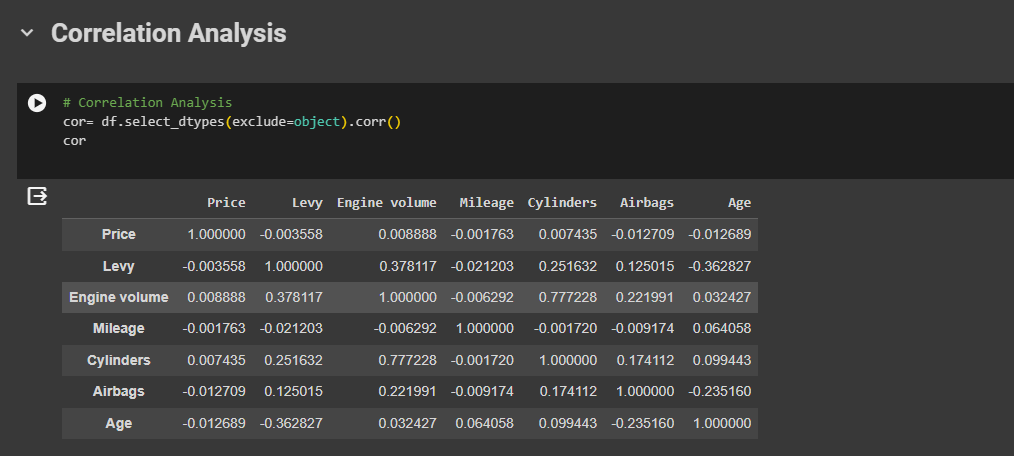


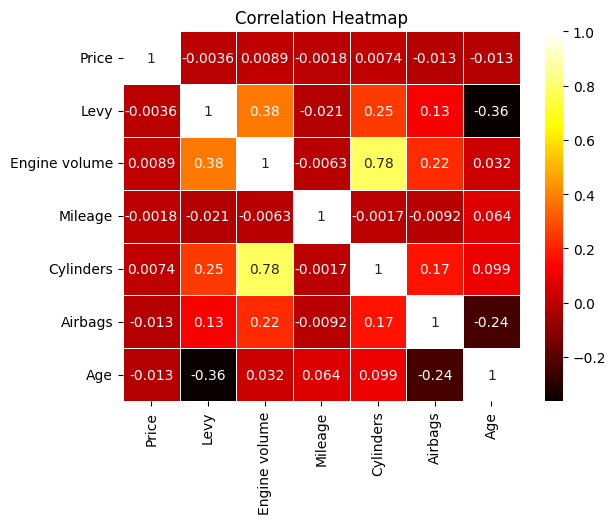


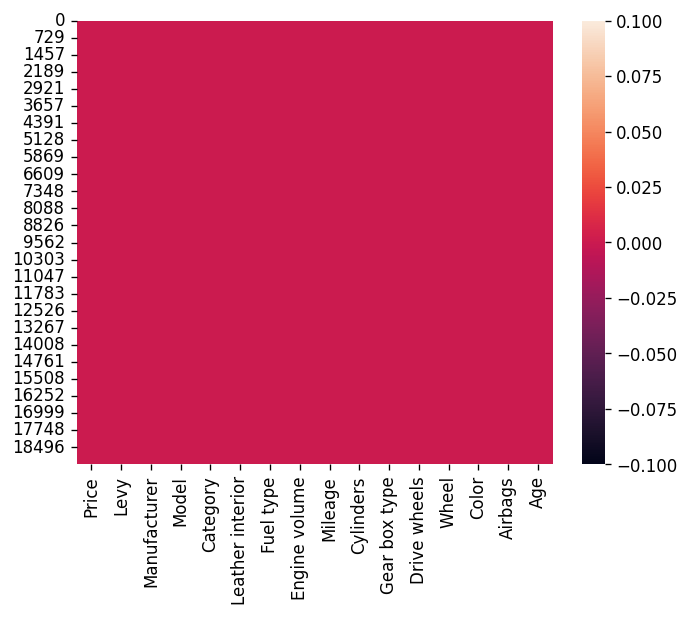




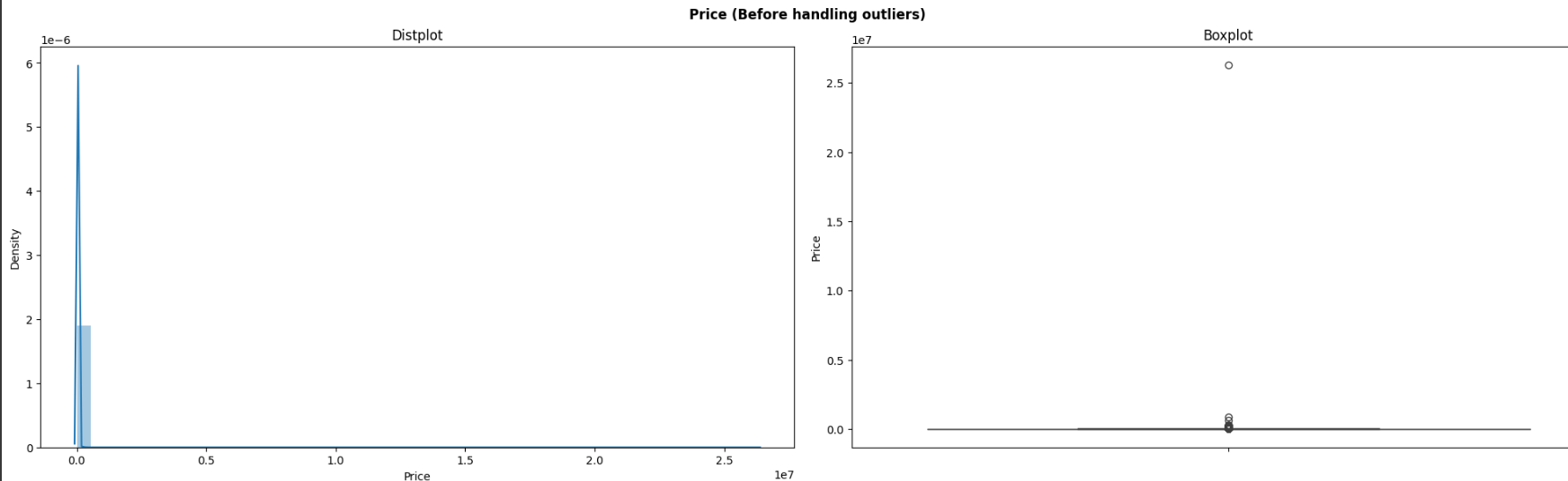
## Correlation Analysis

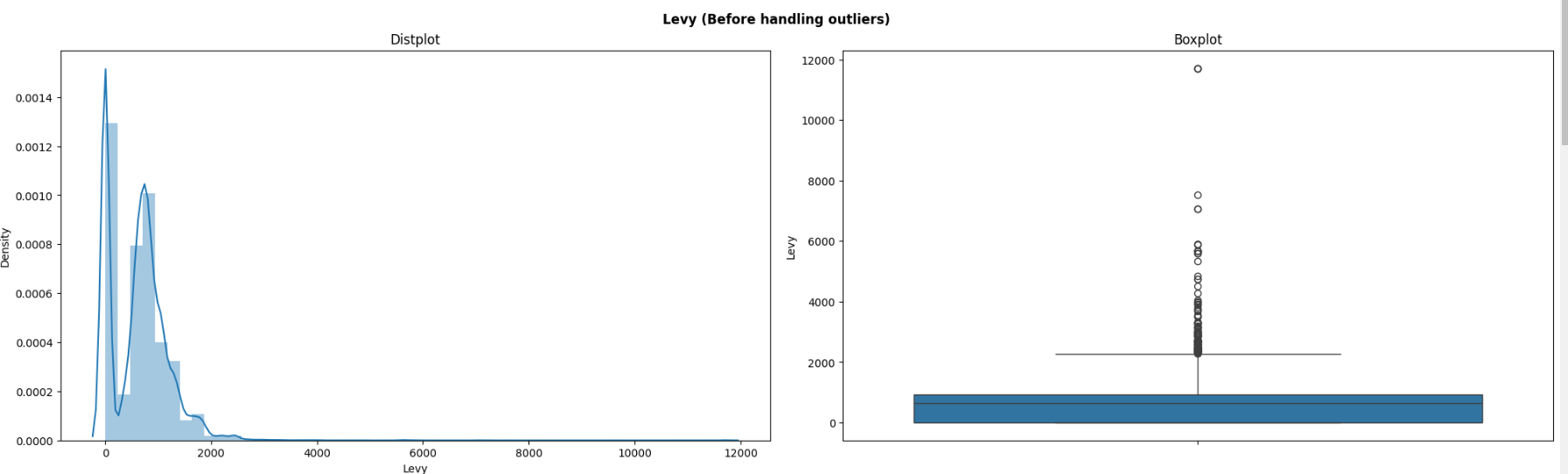


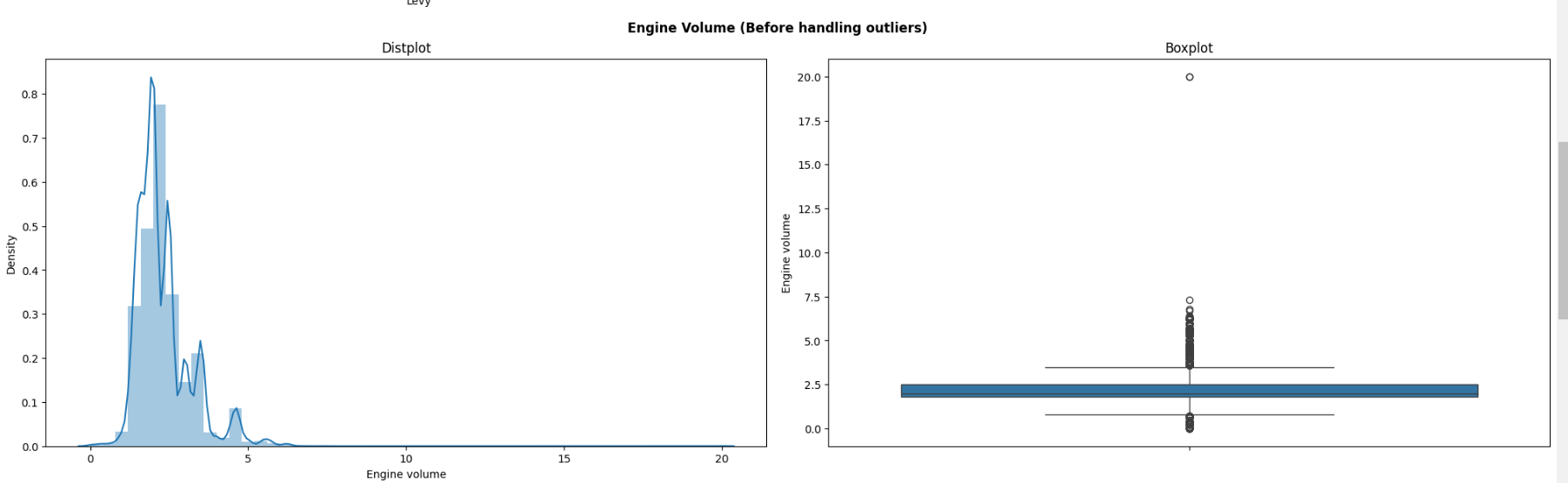


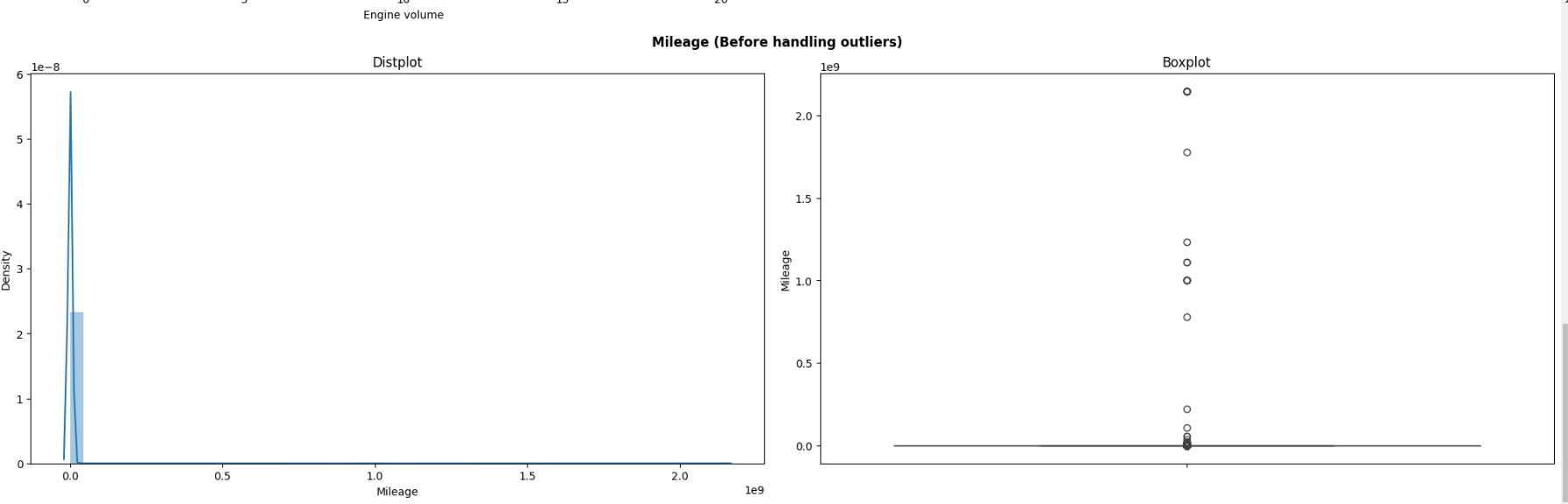


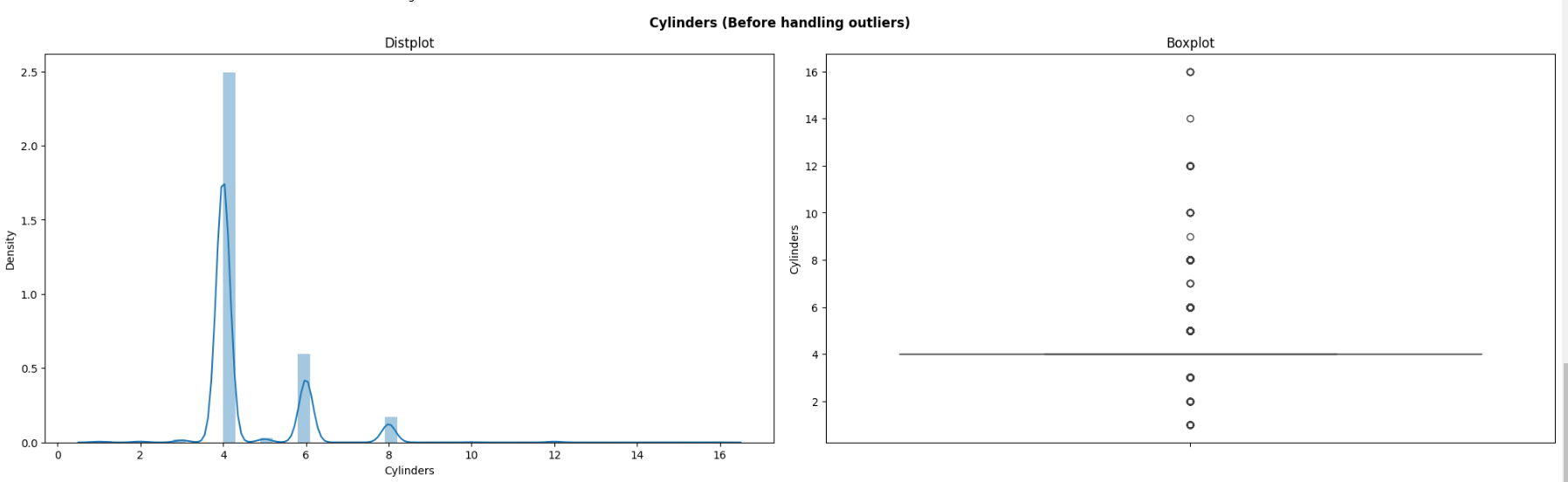
### Outlier Detection

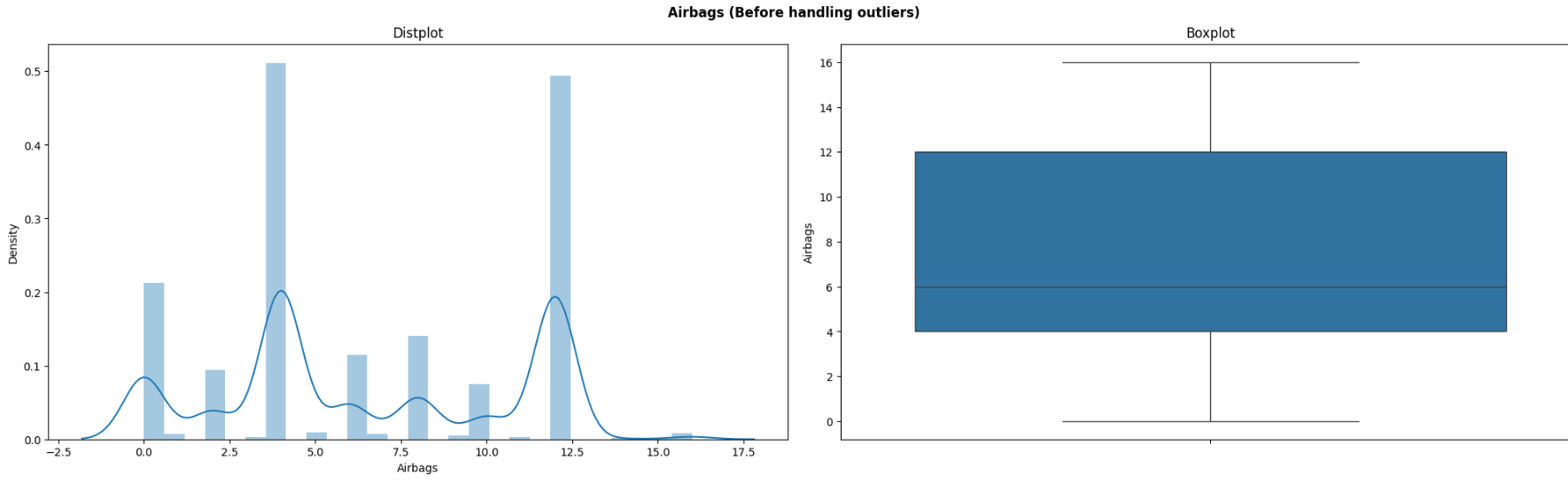


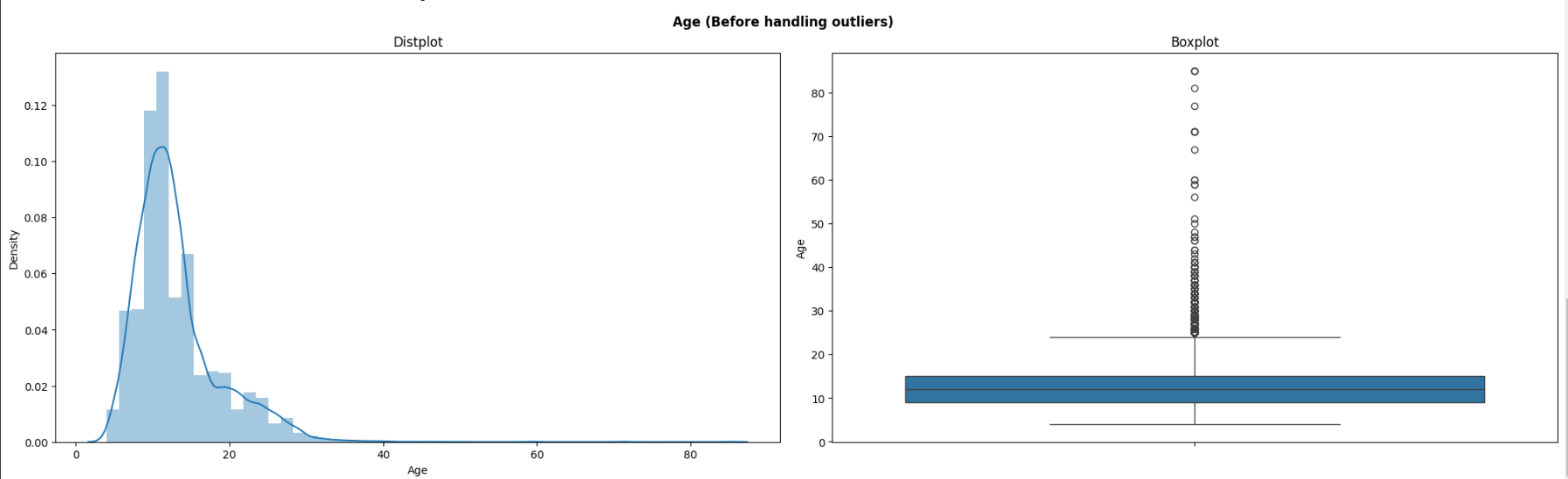


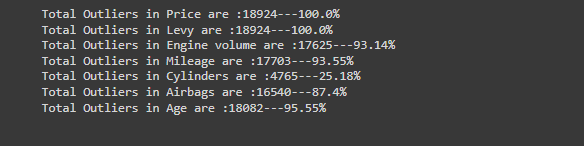




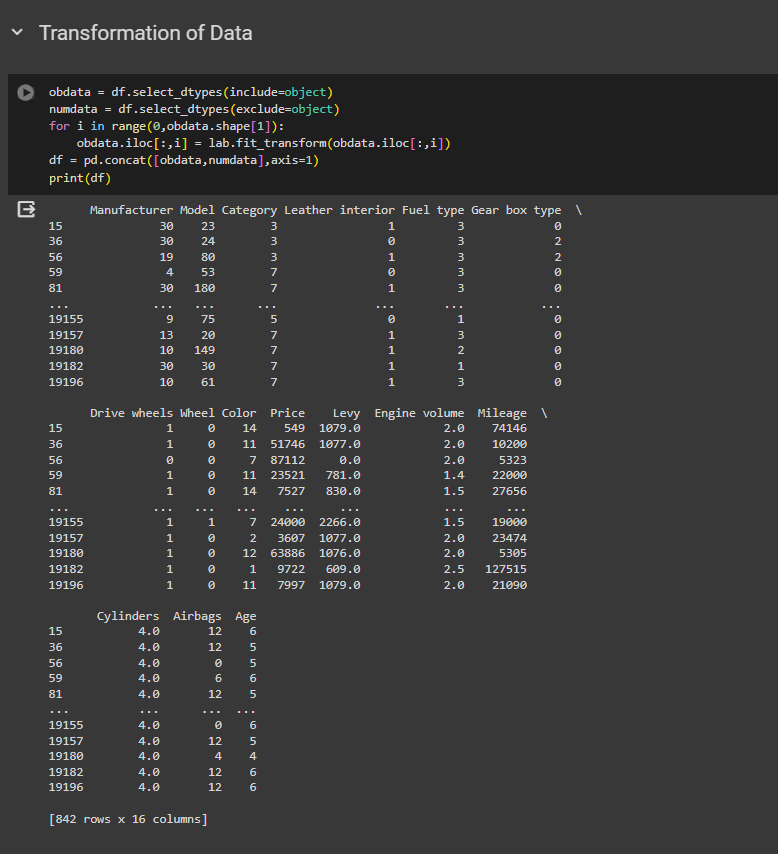






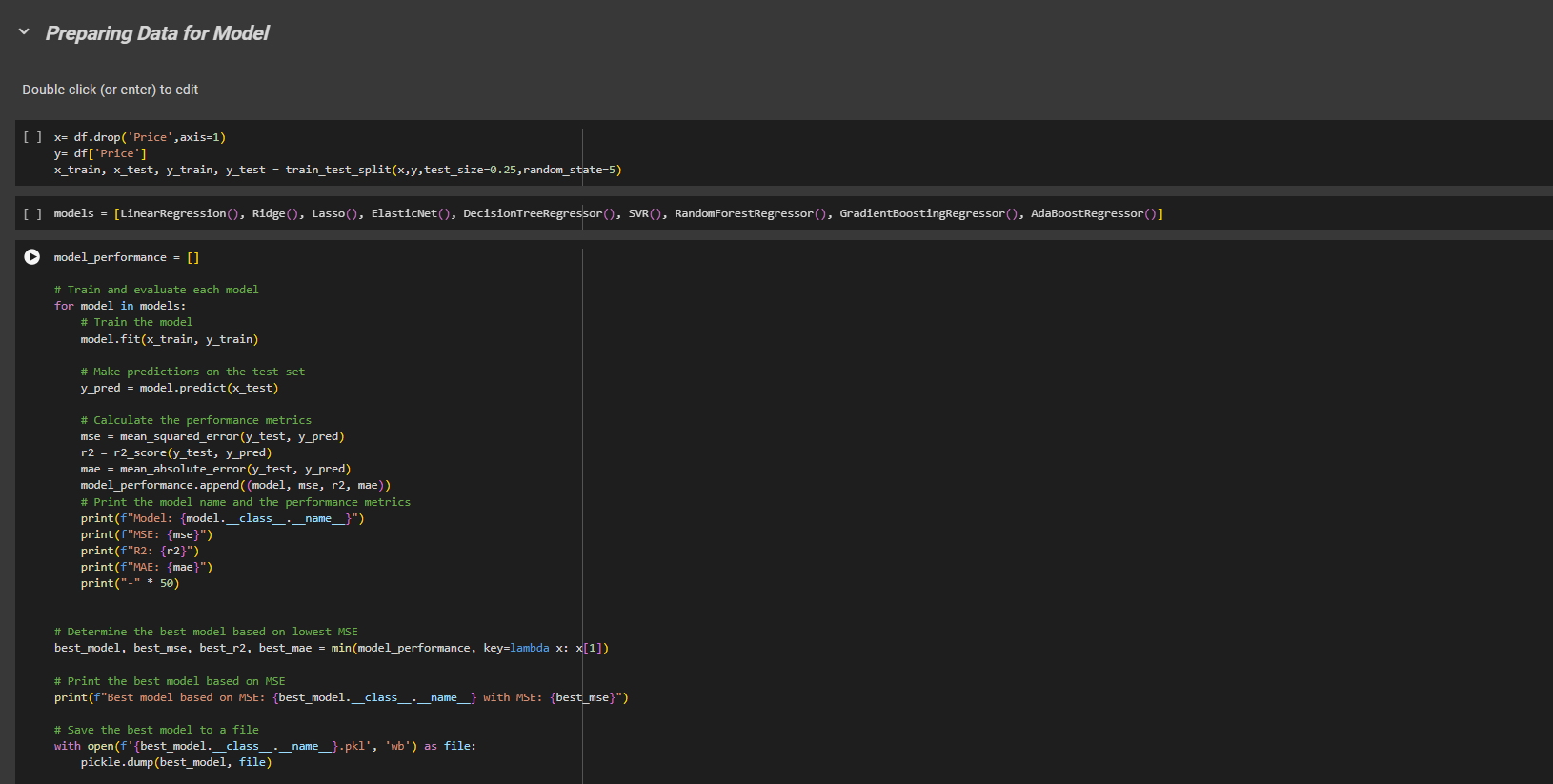


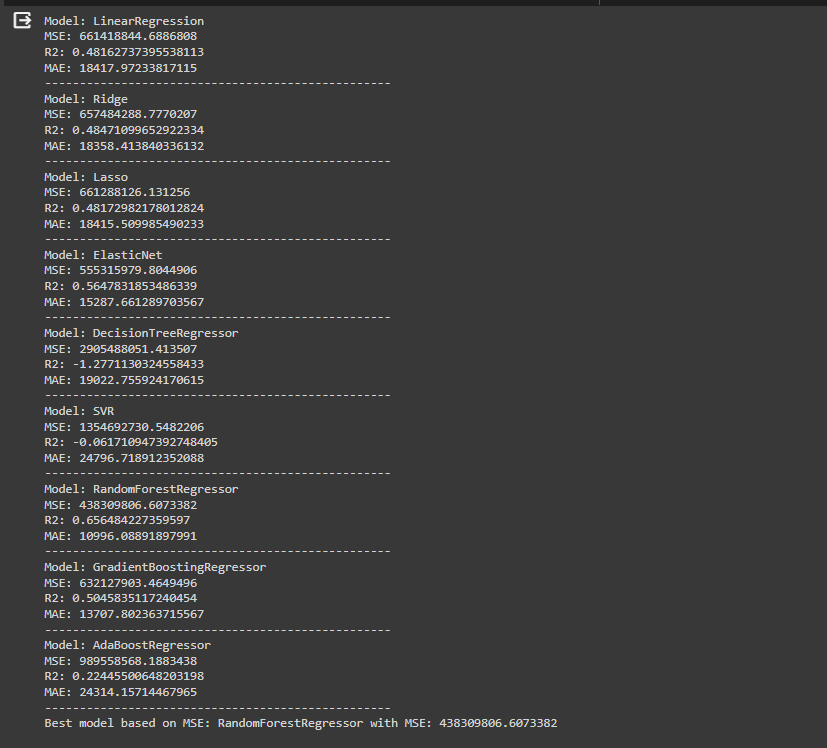
## Transformation data into number



## Stage 2 Training and Investigating multiple Regression algorithms

We have tested various regression models, including Linear Regression, Ridge, Lasso, Elastic Net, Decision Tree Regressor, Support Vector Regressor (SVR), Random Forest Regressor, Gradient Boosting Regressor, and AdaBoost Regressor. Each model is evaluated based on three metrics: Mean Squared Error (MSE), R-squared (R2), and Mean Absolute Error (MAE). The best model according to the MSE is highlighted, which in this case is the Random Forest Regressor.

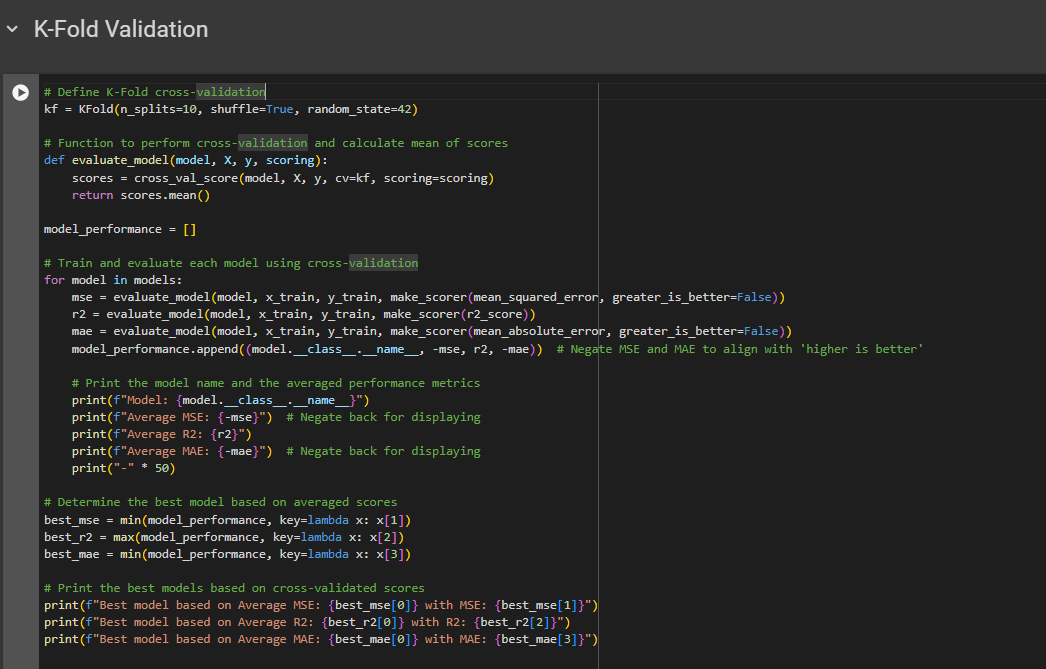


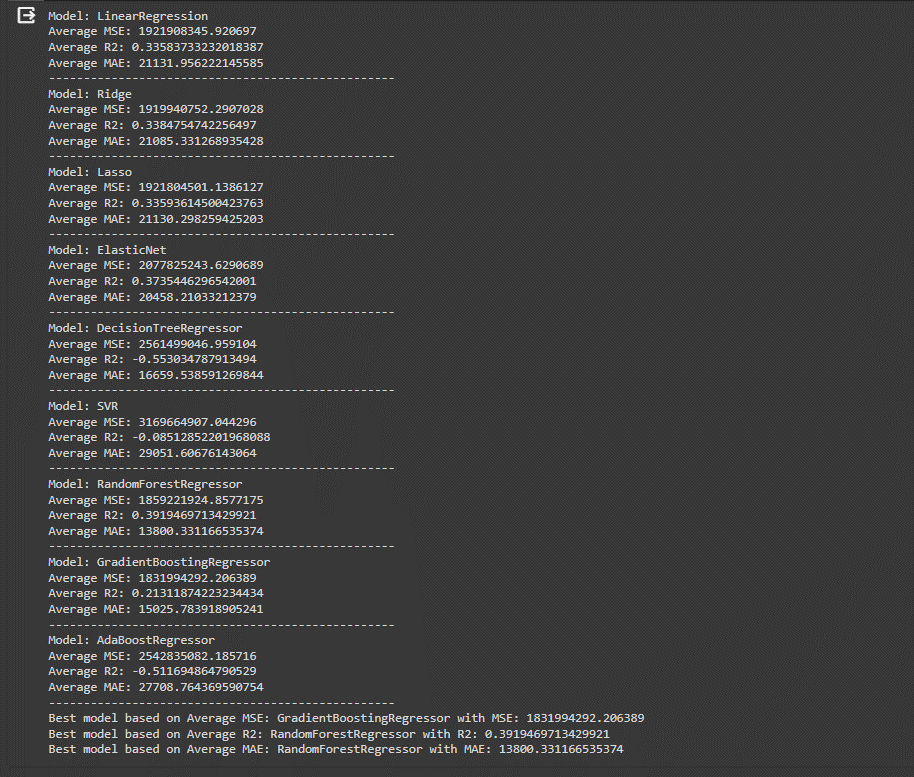


## K-Fold Validation

We have implemented K-Fold cross-validation to evaluate several machine learning models on a dataset. The code involves defining a cross-validation scheme with the KFold class, setting the number of splits to 10, and shuffling the data with a fixed random state for reproducibility.

A function evaluate model is defined to facilitate the computation of cross-validated scores for three different metrics: Mean Squared Error (MSE), R-squared (R2), and Mean Absolute Error (MAE). Notably, the code negates MSE and MAE scores to align with the convention that higher scores indicate better performance (since cross-validation in scikit-learn treats higher values as better).





# Stage 3A Deployment with Streamlit

