INTERNSHIPSTUDIO

Project Title:

Exploratory Data Analysis of Car Features

Context

As a data scientist, the majority of your time will be spent on data pre-processing i.e. making sure you have the right data in the right format. Once this is done, you get a sense of your dataset through applying some descriptive statistics and then, you move on to the exploration stage wherein you plot various graphs and mine the hidden insights. In this project, you as a data scientist are expected to perform Exploratory data analysis on how the different features of a car and its price are related. The data comes from the Kaggle dataset "Car Features and MSRP". It describes almost 12,000 car models, sold in the USA between 1990 and 2017, with the market price (new or used) and some features.

Objective

The objective of the project is to do data pre-processing and exploratory data analysis of the dataset

Data Description

```
# Make                  Car Make
# Model                 Car Model
# Year                  Car Year (Marketing)
# Engine Fuel Type      Engine Fuel Type
# Engine HP             Engine HorsePower (HP)
# Engine Cylinders      Engine Cylinders
# Transmission Type     Transmission Type
# Driven_Wheels         Driven Wheels
# Number of Doors       Number of Doors
# Market Category       Market Category
# Vehicle Size          Size of Vehicle
# Vehicle Style         Type of Vehicle
# highway MPG           Highway MPG
```

```
# city mpg              City MPG
# Popularity            Popularity (Twitter)
# MSRP                  Manufacturer Suggested Retail Price
```

Steps

1. Import the dataset and the necessary libraries, check datatype, statistical summary, shape, null values etc.
2. Are there any columns in the dataset which you think are of less relevance. If so, give your reasoning and drop them.
3. Rename the columns "Engine HP": "HP", "Engine Cylinders": "Cylinders", "Transmission Type": "Transmission", "Driven_Wheels": "Drive Mode","highway MPG": "MPG-H", "city mpg": "MPG-C", "MSRP": "Price"
4. Check for any duplicates in the data, check for null values and missing data and remove them.
5. Plot graphs of various columns to check for outliers and remove those data points from the dataset.
6. What car brands are the most represented in the dataset and find the average price among the top car brands.
7. Plot the correlation matrix and document your insights.
8. Perform EDA and plot different graphs and document your findings (Try to see how other variables affect the price of the car)
9. (Extra Credits)Split the dataset into 80 and 20 ratio and build a machine learning model with Price as the target variable
10. (Extra Credits)Try different algorithms and check their performance over metrics like R square, RMSE, MAE etc and document your findings