Car Price EDA

Submitted in partial fulfillment of the requirements of

BDA Mini Project

for

Final Year of Computer Engineering

By

Deep Shahane	19102B0052
Riya Ingale	19102B0030
Samiksha Pansare	19102B0021
Rutvik Narkar	19102B0060

Under the Guidance of

Prof. Prakash Parmar

Department of Computer Engineering

Vidyalankar Institute of Technology
Wadala(E), Mumbai-400437

University of Mumbai
2022-23

CERTIFICATE OF APPROVAL

This is to certify that the project entitled

“Car Price EDA”

is a bonafide work of

Deep Shahane	**19102B0052**
Riya Ingale	**19102B0030**
Samiksha Pansare	**19102B0021**
Rutvik Narkar	**19102B0060**

submitted to the University of Mumbai in partial fulfillment of

BDA Mini Project

for

Final Year of Computer Engineering

Guide	Head of Department	Principal
Prof. Prakash Parmar	Dr. Sachin Bojewar	Dr. Sunil Patekar

Mini Project Report Approval

This project report entitled **Car Price EDA** by

1.	**Deep Shahane**	**19102B0052**
2.	**Riya Ingale**	**19102B0030**
3.	**Samiksha Pansare**	**19102B0021**
4.	**Rutvik Narkar**	**19102B0060**

is approved for BDA Mini Project for the Final Year of Engineering.

Internal Examiner External Examiner

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name of student	Roll No.	Signature
1) Deep Shahane	19102B0052	
2)Riya Ingale	19102B0030	
3)Samiksha Pansare	19102B0021	
4)Rutvik Narkar	19102B0060	

Date:

Place:

Acknowledgments

This project wouldn’t have been possible without the support, assistance, and guidance of a number of people to whom we would like to express our gratitude. First, we would like to convey our gratitude and regards to our mentor **Prof. Prakash Parmar** for guiding us with his constructive and valuable feedback and for his time and efforts. It was a great privilege to work and study under his guidance.

We would like to extend our heartfelt thanks to our Head of Department, **Dr. Sachin Bojewar** for overseeing this initiative which will, in turn, provide every Vidyalankar student a distinctive competitive edge over others.

We appreciate everyone who spared time from their busy schedules and participated in the survey. Lastly, we are extremely grateful to all those who have contributed and shared their useful insights throughout the entire process and helped us acquire the right direction during this research project.

Abstract

The modern automobile is a complex technical system employing subsystems with specific design functions. Some of these consist of thousands of component parts that have evolved from breakthroughs in existing technology or from new technologies such as electronic computers, high-strength plastics, and new alloys of steel and nonferrous metals.

Exploratory Data Analysis is essential for any business. It allows data scientists to analyze the data before coming to any assumption. It ensures that the results produced are valid and applicable to business outcomes and goals. Explorative Data Analysis is a process where one learns about the data, forms insights and identifies important columns (features) that can be user to tell a story or later formulate a ML problem. The aim of performing Explorative Data Analysis is to find the the features that effect the prices of an Auto Mobile.

Table of Contents

Introduction

The modern automobile is a complex technical system employing subsystems with specific design functions. Some of these consist of thousands of component parts that have evolved from breakthroughs in existing technology or from new technologies such as electronic computers, high-strength plastics, and new alloys of steel and nonferrous metals.

Explorative Data Analysis is a process where one learns about the data, forms insights and identifies important columns (features) that can be user to tell a story or later formulate a ML problem. The aim of performing Explorative Data Analysis is to find the the features that effect the prices of an Auto Mobile.

The project is based on the exploratory data analysis in automobile manufacturing using machine learning with python. We use different modules and library functions. Based on the data analysis using these properties and find the accurate and detailed information about automobile vehicles.

Our project goes through 2 phases: -

1) Data Cleaning: The first step is we'd like to clean and format the info. (That is because computers don't seem to be smart when it involves working out the difference between an image or text once we send it in), that the very first thing we do is typically clean the info so all our data are in one file and text is being processed separately.

2) Exploratory Data Analysis: Exploratory Data Analysis refers to the critical process of performing initial investigations on data so on discover patterns,to spot anomalies, to test hypothesis and to test assumptions with the assistance of summary statistics and graphical representations.

Problem Definition

At present, industry has spent Transfer amount has increased subsequently over the years. So, transfers play an important role in vehicle development. in step with the manufacture analysis produce their products basically for supplies to the vehicle manufacturers. These products are manufactured as per the drawings and specifications of the vehicle producer. Since the Automotive manufacturer's aim is to manage the value, the supplier is left with a meager margin when he supplies his produce to the vehicle manufacturers. These supplies are referred to as supplies to Original Equipment Manufacturers.

Exploratory Data analysis on Automobile analytics allows companies to form decisions supporting performance of their manufacturing products. Also, after analysis car companies and peoples to induce relevant information about the vehicles. The accuracy is extremely questionable during this case. So, with the assistance of machine learning algorithms, we are able to determine the accurate information of the vehicles.

Literature Survey

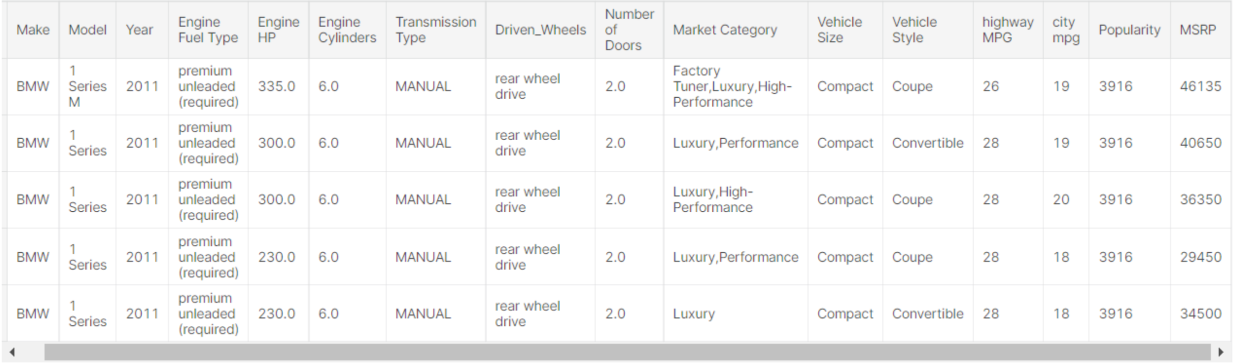
A car (or automobile) is a wheeled motor vehicle that is used for transportation. Most definitions of cars say that they run primarily on roads, seat one to eight people, have four wheels, and mainly transport people instead of goods. The year 1886 is regarded as the birth year of the car, when German inventor Carl Benz patented his Benz Patent-Motorwagen. Cars became widely available during the 20th century. One of the first cars affordable by the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company. Cars were rapidly adopted in the US, where they replaced animal-drawn carriages and carts. In Europe and other parts of the world, demand for automobiles did not increase until after World War II. The car is considered an essential part of the developed economy.

This Repository is used in identify the important drivers for the sales and predict the new cars sales for the given model by using linear regression algorithm through the Statistical Approach. Code for this problem is built on python 3.7 and has been written on Jupyter Notebook. (Attached as LR - Prediction of Car Sales_updated.ipynb file). Business Context for this repository is given in attached CAR SALES PREDICTION - Regression Case Study PDF file and Data available for this same is in Car_sales.csv.

In [3] the author illustrates about Sales prediction is the current numerous trend in which all the business companies thrive and it also aids the organization or concern in determining the future goals for it and its plan and procedure to achieve it. The data about car sales are derived from various sources. Sales of cars does not contain any independent variable since various factors such as horsepower; model, width, fuel type, height, price, city-mileage, highway-mileage and manufacturer are the various features that influence the sales. In car sales prediction we first implement the methodology of analytic hierarchy process in order to get varied idea about how well the various criteria in our data set works and after this we apply the machine learning algorithms such as Linear regression, Random tree to get the best clusters and we process them in to random forest to get best accurate feature out of it. which is ultimately followed by Technique for Order of preference by similarity to ideal solution (TOPSIS) an tool which helps the researcher to arrive at a verdict when he/she faces the one or more pattern selection problem and the final resultant derived from all these methods gives the fittest feature which influences the customer in purchasing the car which indirectly gives the company or the research market a result in predicting the future sales for cars. Hence this paper not only provides its users with some stats, it also serves as a guiding guardian by providing accurate results for purchasing a car.

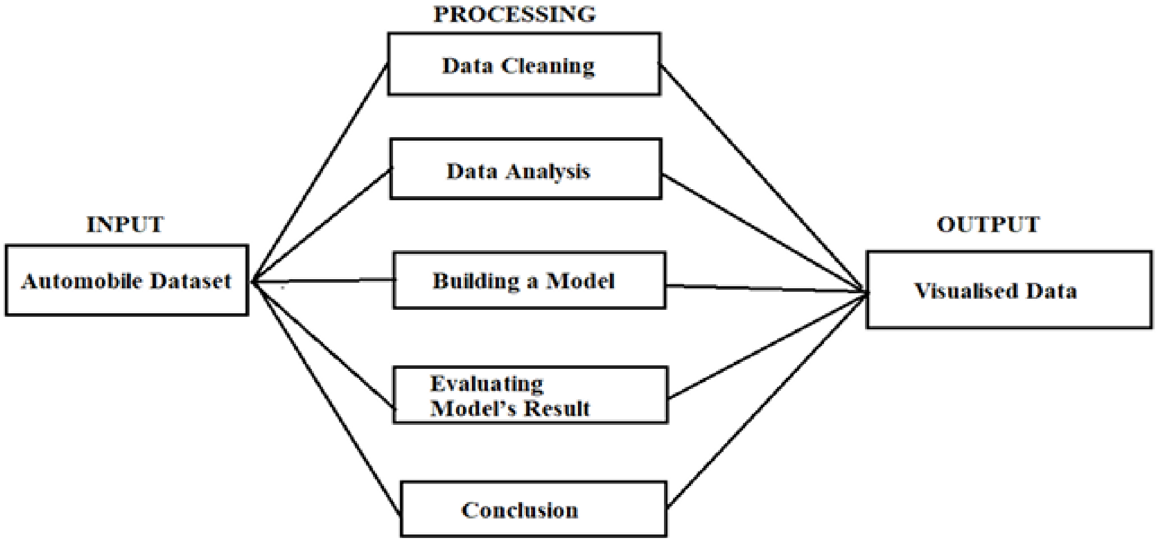
Dataset

Car’s dataset with features including make, model, year, engine, and other properties of the car used to predict its price. It is scraped from Edmunds and Twitter.



Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg	Popularity	MSRP
BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916	46135
BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	28	19	3916	40650
BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916	36350
BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916	29450
BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertible	28	18	3916	34500

```
Shape of the Data:  (11914, 16)
```



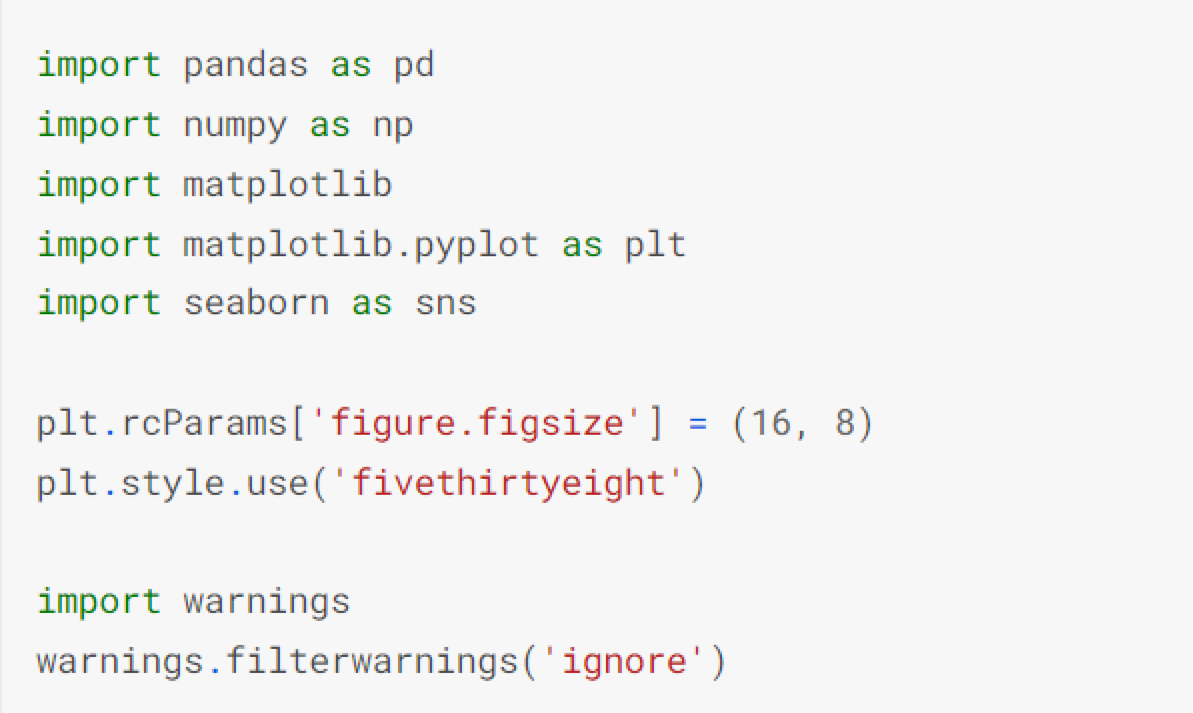
Proposed System

Our System goes through the following phases: -

1. Introduction

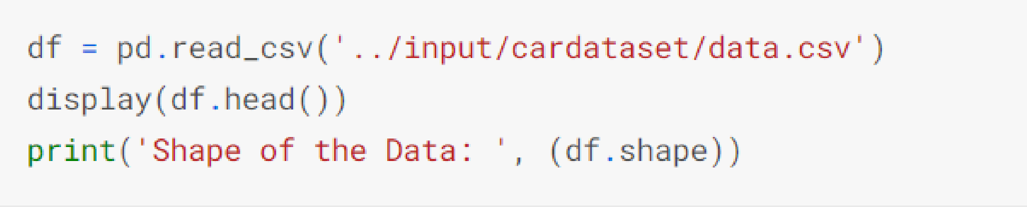
Car’s dataset with features including make, model, year, engine, and other properties of the car used to predict its price. Scraped from Edmunds and Twitter. Effects of features on the price. Predict the price of Cars using different Variables

2. Importing Libraries

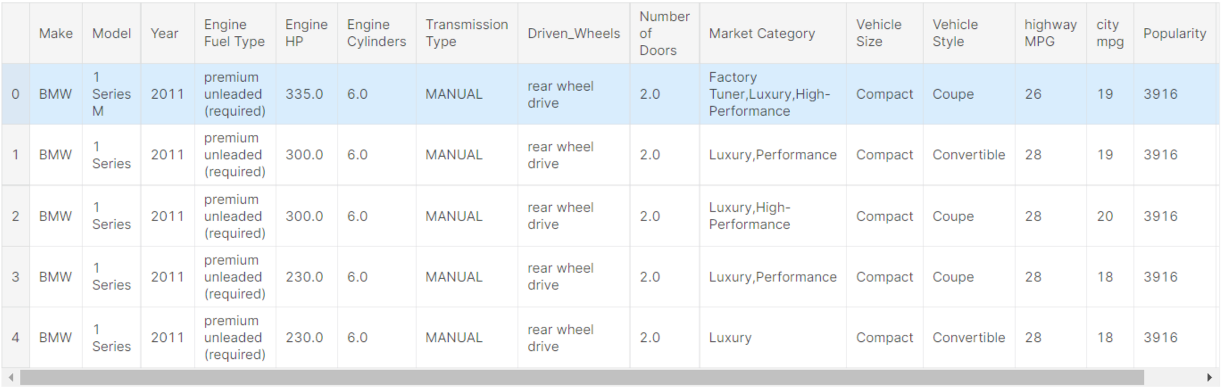


```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns

plt.rcParams['figure.figsize'] = (16, 8)
plt.style.use('fivethirtyeight')

import warnings
warnings.filterwarnings('ignore')
```

3. Loading and Reading Data

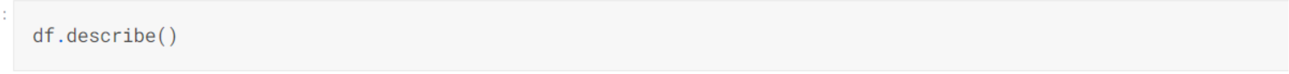


```
df = pd.read_csv('../input/cardataset/data.csv')
display(df.head())
print('Shape of the Data: ', (df.shape))
```

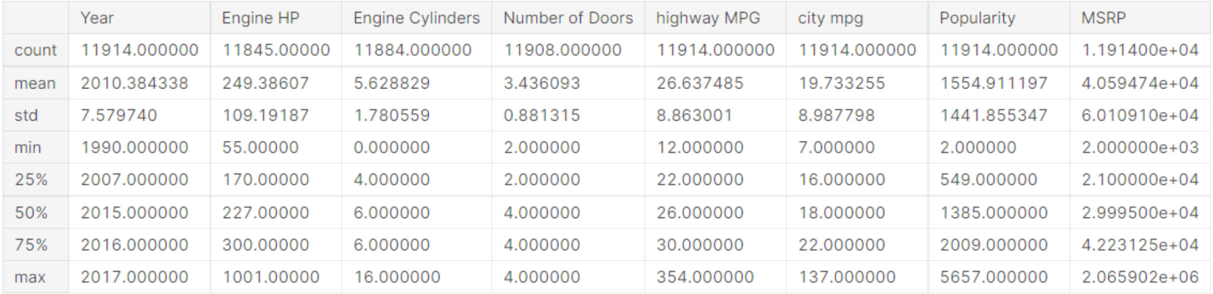


	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg	Popularity
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	28	19	3916
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertible	28	18	3916

4. Data Description and initial cleaning

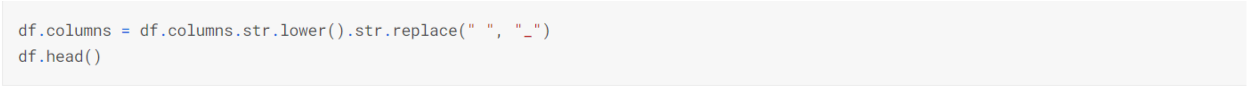


```
df.describe()
```

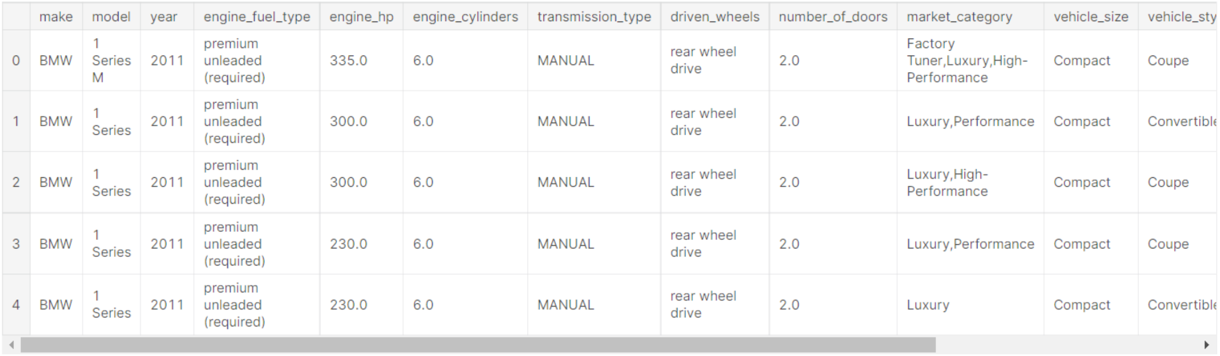


	Year	Engine HP	Engine Cylinders	Number of Doors	highway MPG	city mpg	Popularity	MSRP
count	11914.000000	11845.00000	11884.000000	11908.000000	11914.000000	11914.000000	11914.000000	1.191400e+04
mean	2010.384338	249.38607	5.628829	3.436093	26.637485	19.733255	1554.911197	4.059474e+04
std	7.579740	109.19187	1.780559	0.881315	8.863001	8.987798	1441.855347	6.010910e+04
min	1990.000000	55.00000	0.000000	2.000000	12.000000	7.000000	2.000000	2.000000e+03
25%	2007.000000	170.00000	4.000000	2.000000	22.000000	16.000000	549.000000	2.100000e+04
50%	2015.000000	227.00000	6.000000	4.000000	26.000000	18.000000	1385.000000	2.999500e+04
75%	2016.000000	300.00000	6.000000	4.000000	30.000000	22.000000	2009.000000	4.223125e+04
max	2017.000000	1001.00000	16.000000	4.000000	354.000000	137.000000	5657.000000	2.065902e+06

4.1 Cleaning Strings in Column and Values



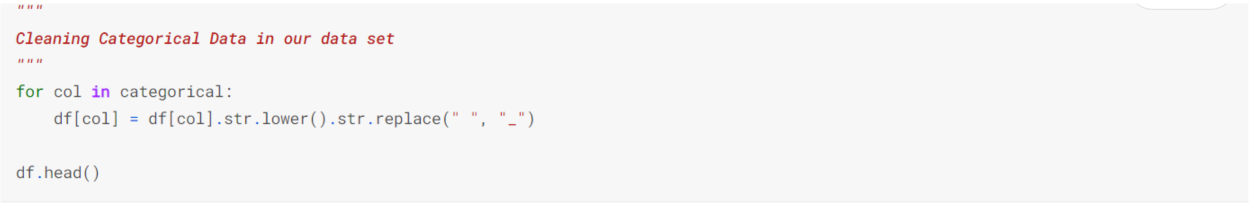
```
df.columns = df.columns.str.lower().str.replace(" ", "_")
df.head()
```



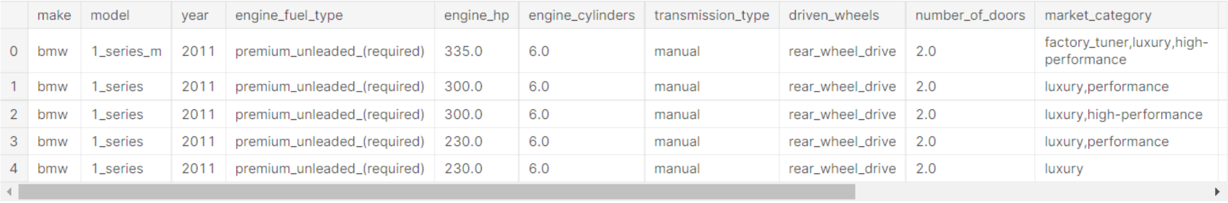
	make	model	year	engine_fuel_type	engine_hp	engine_cylinders	transmission_type	driven_wheels	number_of_doors	market_category	vehicle_size	vehicle_sty
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertibl
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertibl

4.2 Making List of Categorical Columns

4.3 Cleaning Categorical Data in our data set

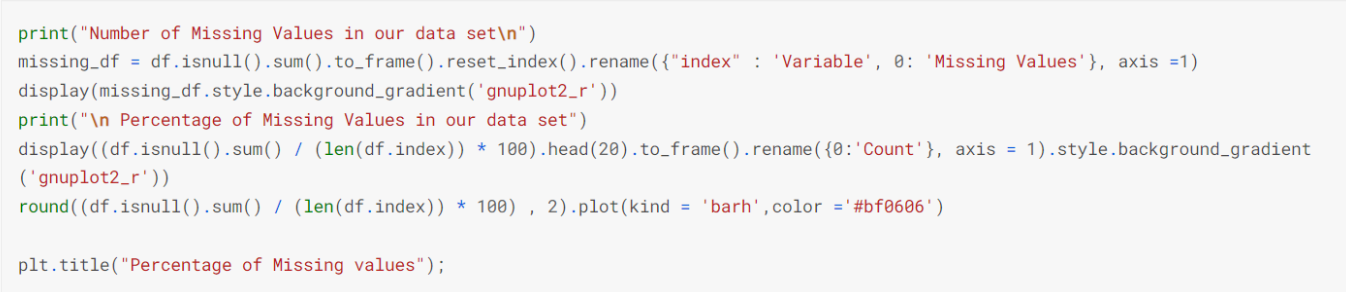


```
"""
Cleaning Categorical Data in our data set
"""
for col in categorical:
    df[col] = df[col].str.lower().str.replace(" ", "_")

df.head()
```



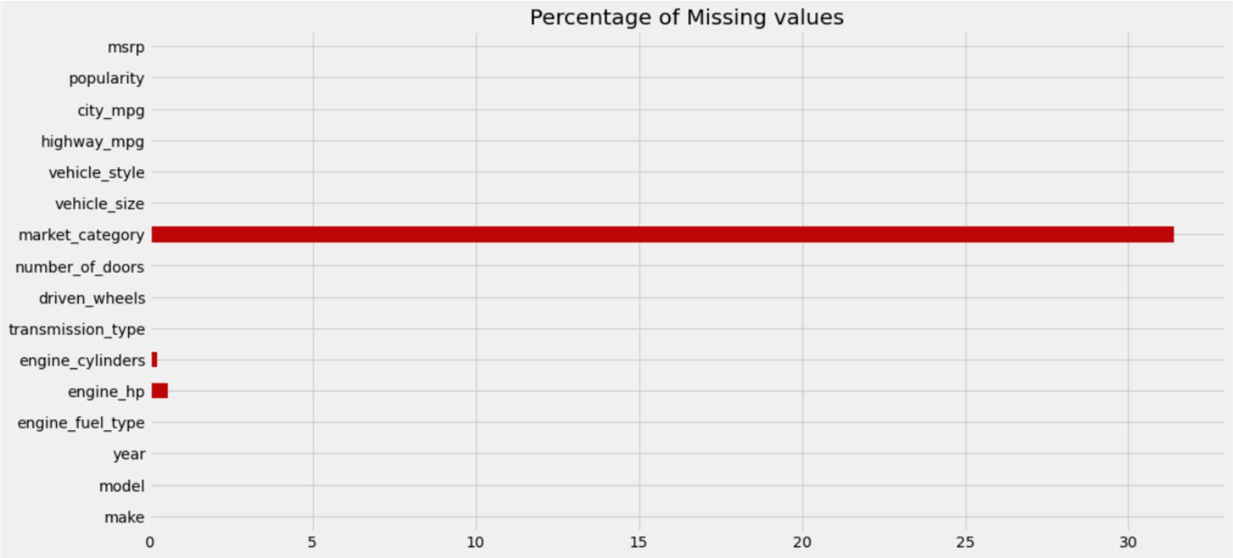
	make	model	year	engine_fuel_type	engine_hp	engine_cylinders	transmission_type	driven_wheels	number_of_doors	market_category
0	bmw	1_series_m	2011	premium_unleaded_(required)	335.0	6.0	manual	rear_wheel_drive	2.0	factory_tuner,luxury,high-performance
1	bmw	1_series	2011	premium_unleaded_(required)	300.0	6.0	manual	rear_wheel_drive	2.0	luxury,performance
2	bmw	1_series	2011	premium_unleaded_(required)	300.0	6.0	manual	rear_wheel_drive	2.0	luxury,high-performance
3	bmw	1_series	2011	premium_unleaded_(required)	230.0	6.0	manual	rear_wheel_drive	2.0	luxury,performance
4	bmw	1_series	2011	premium_unleaded_(required)	230.0	6.0	manual	rear_wheel_drive	2.0	luxury

5. Missing Values



```
print("Number of Missing Values in our data set\n")
missing_df = df.isnull().sum().to_frame().reset_index().rename({"index" : 'Variable', 0: 'Missing Values'}, axis =1)
display(missing_df.style.background_gradient('gnuplot2_r'))
print("\n Percentage of Missing Values in our data set")
display((df.isnull().sum() / (len(df.index)) * 100).head(20).to_frame().rename({0:'Count'}, axis = 1).style.background_gradient
('gnuplot2_r'))
round((df.isnull().sum() / (len(df.index)) * 100) , 2).plot(kind = 'barh',color ='#bf0606')

plt.title("Percentage of Missing values");
```

5.1 Treating Missing Values

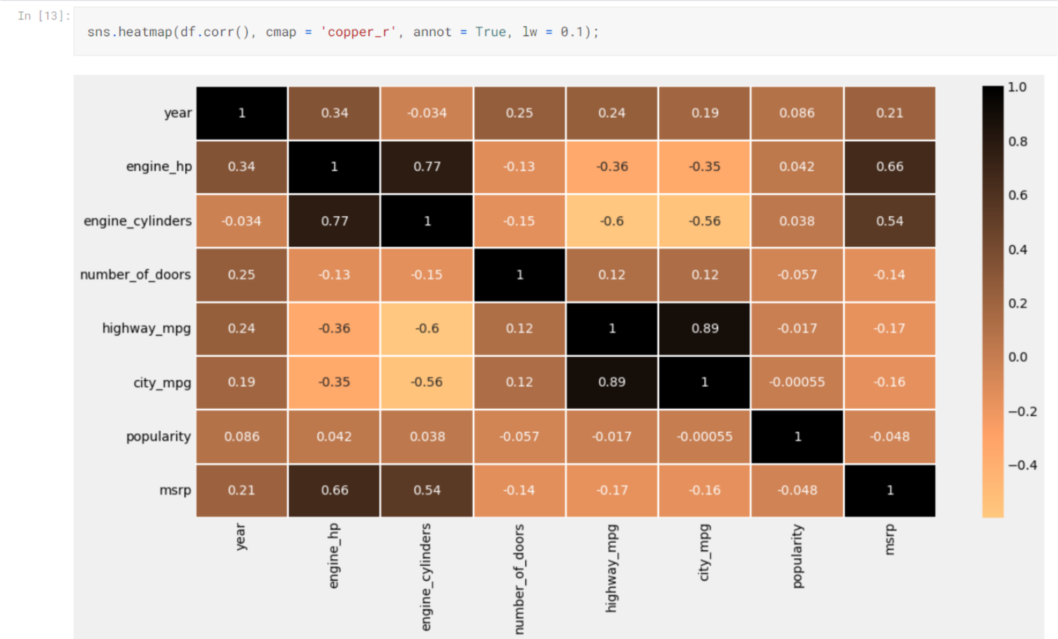


6. Checking duplicates in our data set.

6.1 Dropping the duplicates

7. Checking Correlation

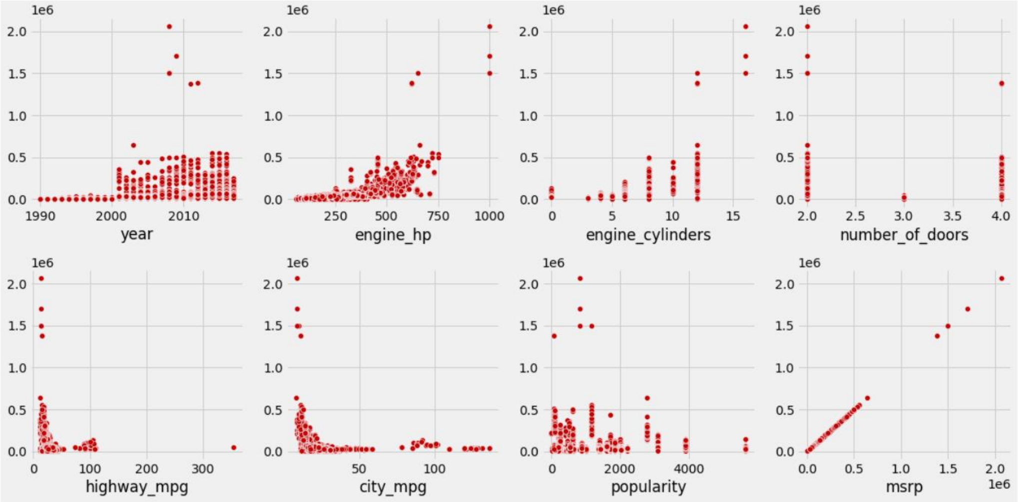
8. Checking Relation between all variables



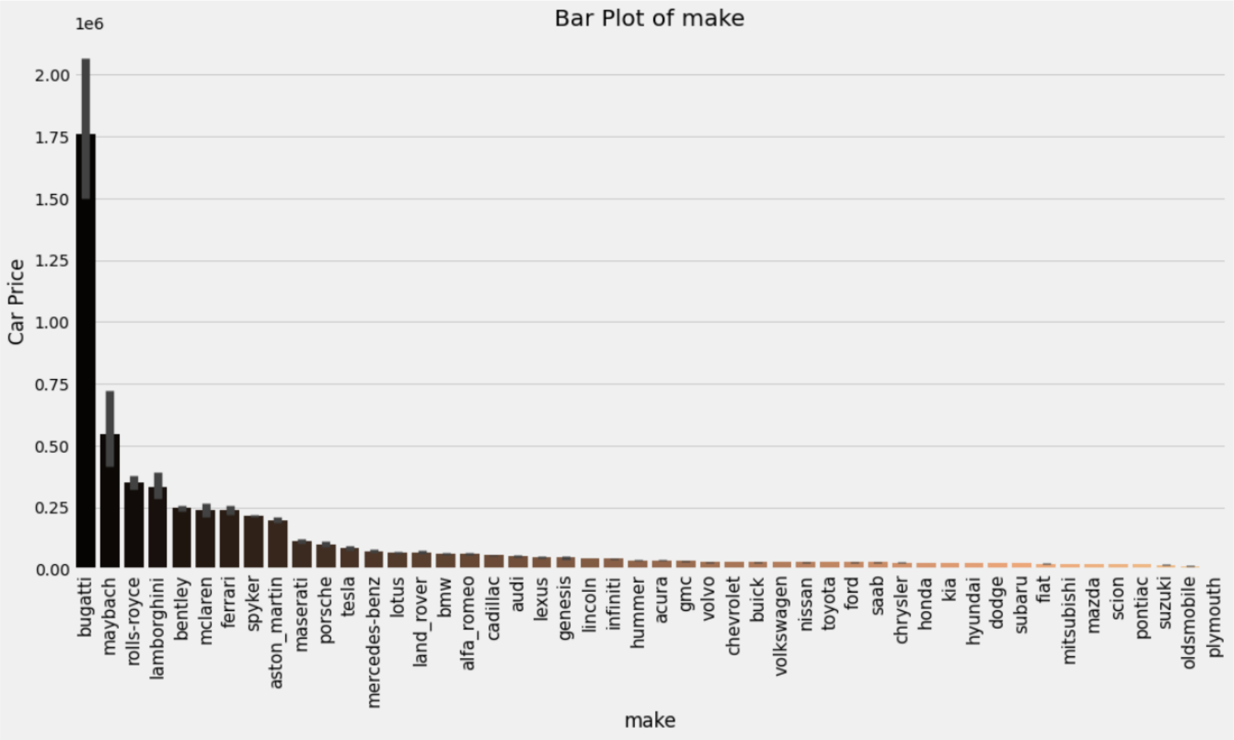
```
In [13]: sns.heatmap(df.corr(), cmap = 'copper_r', annot = True, lw = 0.1);
```

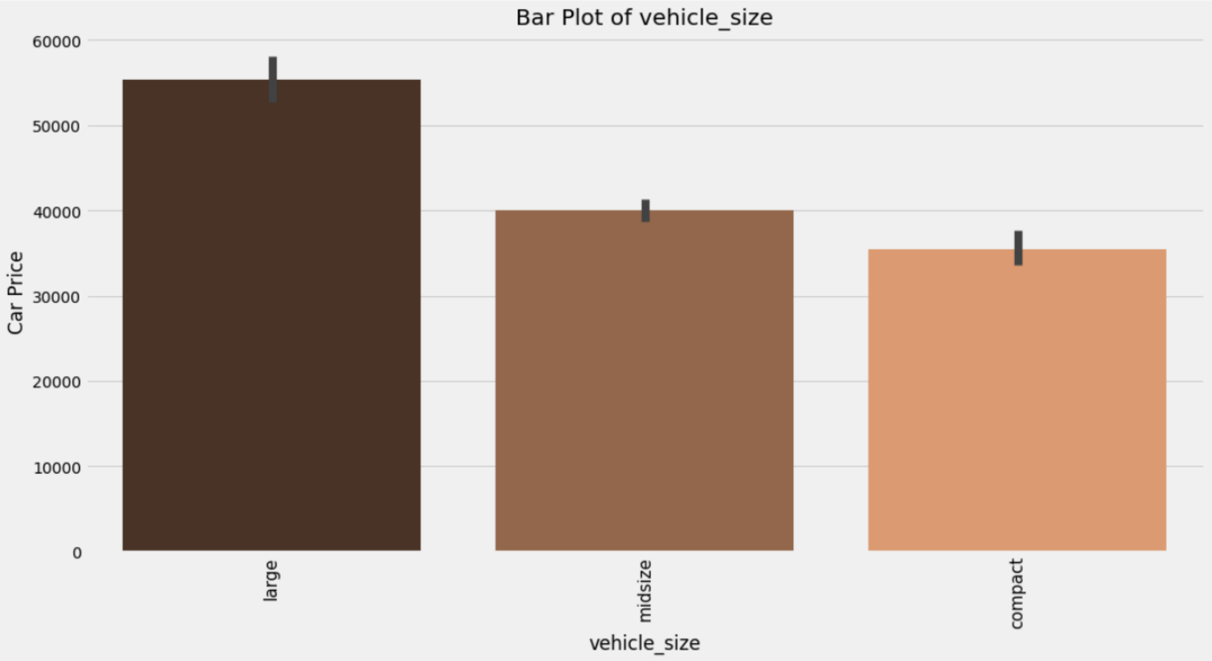
	year	engine_hp	engine_cylinders	number_of_doors	highway_mpg	city_mpg	popularity	msrp
year	1	0.34	-0.034	0.25	0.24	0.19	0.086	0.21
engine_hp	0.34	1	0.77	-0.13	-0.36	-0.35	0.042	0.66
engine_cylinders	-0.034	0.77	1	-0.15	-0.6	-0.56	0.038	0.54
number_of_doors	0.25	-0.13	-0.15	1	0.12	0.12	-0.057	-0.14
highway_mpg	0.24	-0.36	-0.6	0.12	1	0.89	-0.017	-0.17
city_mpg	0.19	-0.35	-0.56	0.12	0.89	1	-0.00055	-0.16
popularity	0.086	0.042	0.038	-0.057	-0.017	-0.00055	1	-0.048
msrp	0.21	0.66	0.54	-0.14	-0.17	-0.16	-0.048	1

8.1 Distribution and relationship of Numerical Variables with dependent variable



8.1.1 Distribution and relationship of Numerical Variables with dependent variable


Bar Plot of make
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Car Price
bugatti
maybach
rolls-royce
lamborghini
bentley
mclaren
ferrari
spyker
aston_martin
maserati
porsche
tesla
mercedes-benz
lotus
land_rover
bmw
alfa_romeo
cadillac
audi
lexus
genesis
lincoln
infiniti
hummer
acura
gmc
volvo
chevrolet
buick
volkswagen
nissan
toyota
ford
saab
chrysler
honda
kia
hyundai
dodge
subaru
fiat
mitsubishi
mazda
scion
pontiac
suzuki
oldsmobile
plymouth
make


Bar Plot of vehicle_size
60000
50000
40000
30000
20000
10000
0
Car Price
large
midsize
compact
vehicle_size

Conclusion

These bar graphs represent individual categorical variable relations with dependent variables. Many groups in every variable have led to high prices in cars.

Some of them are:

convertible and coupe in vehicle_style, large in vehicle_size, exotic, luxury, performance in market_category, all wheel drive and rear wheel drive in driven_wheels, automated-manual in transmission type, bugatti, maybach in vehicle_size

References

1]https://www.kaggle.com/code/melikedilekci/eda-cars-india-dataset/notebook#1-||-INTRODUCTION

2] https://github.com/rishika1444/Car-Sales-Prediction

3] Madhuvanthi, K. & Kailasanathan, Nallakaruppan & N C, Senthilkumar & Somayaji, Siva. (2019). Car Sales Prediction Using Machine Learning Algorithmns. International Journal of Innovative Technology and Exploring Engineering. 8.

4] EDA (Exploratory Data Analysis) on Used Car Sales