SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE NAME: Machine Learning Laboratory **COURSE CODE:** DJS22L602

CLASS: Third Year B.Tech **SEM:** VI

NAME:Dipti Agarwal **DIV:** IT1-1 **ROLL:** I047

EXPERIMENT NO. 1

CO1 - Gain knowledge about basic concepts of Machine Learning.

TITLE: Data Preprocessing

AIM / OBJECTIVE:

To perform data preprocessing in terms of handling, missing data, removing outliers, eliminating duplicate rows and modifying the datatype, etc.

DESCRIPTION OF EXPERIMENT:

Python is an easy-to-learn programming language, which makes it the most preferred choice for beginners in Data Science, Data Analytics, and Machine Learning. It also has a great community of online learners and excellent data-centric libraries. With so much data being generated, it becomes important that the data we use for Data Science applications like Machine Learning and Predictive Modeling is clean. But what do we mean by clean data? And what makes data dirty in the first place? Dirty data simply means data that is erroneous. Duplicacy of records, incomplete or outdated data, and improper parsing can make data dirty. This data needs to be cleaned. Data cleaning (or data cleansing) refers to the process of "cleaning" this dirty data, by identifying errors in the data and then rectifying them. Data cleaning is an important step in and Machine Learning project, and we will cover some basic data cleaning techniques (in Python) in this article.

Cleaning Data in Python

We will now separate the numeric columns from the categorical columns.

Missing values

We will start by calculating the percentage of values missing in each column, and then storing this information in a DataFrame.

Drop observations

One way could be to drop those observations that contain any null value in them for any of the columns. This will work when the percentage of missing values in each column is very less.

Remove columns (features)

Another way to tackle missing values in a dataset would be to drop those columns or features that have a significant percentage of values missing.

Impute missing values

There is still missing data left in our dataset. We will now impute the missing values in each numerical column with the median value of that column.

Outliers

An outlier is an unusual observation that lies away from the majority of the data. Outliers can affect the performance of a Machine Learning model significantly.

Duplicate records

Data can sometimes contain duplicate values. It is important to remove duplicate records from your dataset before you proceed with any Machine Learning project. In our data, since the ID column is a unique identifier, we will drop duplicate records by considering all but the ID column.

Fixing data type

Often in the dataset, values are not stored in the correct data type. This can create a problem in later stages, and we may not get the desired output or may get errors while execution.

PROCEDURE:

The dataset provides information about vehicles, including their make, model, year, engine type, horsepower, transmission, drive type, number of doors, market category, size, style, fuel efficiency (highway and city MPG), popularity, and suggested retail price (MSRP). It helps compare vehicle features and performance.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
# Load the dataset
df = pd.read_csv('/content/data 1.csv')  # replace with your actual file path

# Display the first few rows of the dataset
print("First 5 rows of the dataset:")
print(df.head())

# 1. Examine the data for missing values, identify data types, and inconsistencies
# Checking for missing values
missing_data = df.isnull().sum()
print("\nMissing Values:")
print(missing_data)

# Checking data types
data_types = df.dtypes
print("\nData Types:")
print(data_types)

# Summary of the dataframe
df_info = df.info()
print("\nDataframe Info:")
print(df_info)

# Statistical summary
summary_statistics = df.describe()
print("\nSummary Statistics:")
print(summary_statistics)

# 2. Handle missing data
# Handling missing values: Filling missing numerical values with the median or mode
df['Engine HP'] = df['Engine HP'].fillna(df['Engine HP'].median())
df['Engine Cylinders'] = df['Engine Cylinders'].fillna(df['Engine
Cylinders'].mode()[0])  # Fixed line
df['Number of Doors'] = df['Number of Doors'].fillna(df['Number of Doors'].mode()[0])

# 3. Calculate summary statistics
# 3. Calculate summary statistics (only for numeric columns)
numeric_columns = df.select_dtypes(include=[np.number])  # Select only numeric
columns

# Now calculate summary statistics on numeric columns
mean_values = numeric_columns.mean()
median_values = numeric_columns.median()
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
mode_values = numeric_columns.mode().iloc[0]  # mode() returns a DataFrame, so get the first row
std_dev_values = numeric_columns.std()

print("\nMean Values:")
print(mean_values)

print("\nMedian Values:")
print(median_values)

print("\nMode Values:")
print(mode_values)

print("\nStandard Deviation Values:")
print(std_dev_values)


# 4. Data Transformation
# Scaling numerical features
scaler = StandardScaler()
df[['Engine HP', 'Engine Cylinders']] = scaler.fit_transform(df[['Engine HP', 'Engine Cylinders']])

# Encoding categorical columns (Example: Transmission Type)
df['Transmission Type'] = df['Transmission Type'].map({'Automatic': 0, 'Manual': 1})

# 5. Data Visualization
# Univariate Analysis (Histogram for numerical features)

plt.figure(figsize=(10, 6))
sns.histplot(df['Engine HP'], kde=True)
plt.title('Distribution of Engine HP')
plt.xlabel('Engine HP')
plt.ylabel('Frequency')
plt.show()

plt.figure(figsize=(10, 6))
sns.boxplot(x=df['MSRP'])
plt.title('Boxplot of MSRP')
plt.show()

# Bivariate Analysis (Scatter plot between Engine HP and MSRP)
plt.figure(figsize=(10, 6))
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
sns.scatterplot(x=df['Engine HP'], y=df['MSRP'])
plt.title('Scatter plot of Engine HP vs MSRP')
plt.xlabel('Engine HP')
plt.ylabel('MSRP')
plt.show()

# Correlation heatmap for numerical variables
df_encoded = df.copy()
for col in df.select_dtypes(include=['object']).columns:  # Convert categorical
columns
    df_encoded[col] = LabelEncoder().fit_transform(df[col])

plt.figure(figsize=(10,5))
c = df_encoded.corr()
sns.heatmap(c, cmap="BrBG", annot=True)
plt.show()

# 6. Multivariate Analysis (Pairplot to visualize relationships between features)
sns.pairplot(df[['Engine HP', 'Engine Cylinders', 'MSRP', 'highway MPG']])
plt.suptitle('Pairplot of Selected Features', y=1.02)
plt.show()
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

First 5 rows of the dataset:

	Make	Model	Year	Engine Fuel Type	Engine HP
0	BMW	1 Series M	2011	premium unleaded (required)	335.0
1	BMW	1 Series	2011	premium unleaded (required)	300.0
2	BMW	1 Series	2011	premium unleaded (required)	300.0
3	BMW	1 Series	2011	premium unleaded (required)	230.0
4	BMW	1 Series	2011	premium unleaded (required)	230.0

	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors
0	6.0	MANUAL	rear wheel drive	2.0
1	6.0	MANUAL	rear wheel drive	2.0
2	6.0	MANUAL	rear wheel drive	2.0
3	6.0	MANUAL	rear wheel drive	2.0
4	6.0	MANUAL	rear wheel drive	2.0

	Market Category	Vehicle Size	Vehicle Style
0	Factory Tuner,Luxury,High-Performance	Compact	Coupe
1	Luxury,Performance	Compact	Convertible
2	Luxury,High-Performance	Compact	Coupe
3	Luxury,Performance	Compact	Coupe
4	Luxury	Compact	Convertible

	highway MPG	city mpg	Popularity	MSRP
0	26	19	3916	46135
1	28	19	3916	40650
2	28	20	3916	36350
3	28	18	3916	29450
4	28	18	3916	34500

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
Missing Values:
Make                    0
Model                   0
Year                    0
Engine Fuel Type        3
Engine HP              69
Engine Cylinders       30
Transmission Type       0
Driven_Wheels           0
Number of Doors         6
Market Category      3742
Vehicle Size            0
Vehicle Style           0
highway MPG             0
city mpg                0
Popularity              0
MSRP                    0
dtype: int64

Data Types:
Make                  object
Model                 object
Year                   int64
Engine Fuel Type      object
Engine HP            float64
Engine Cylinders     float64
Transmission Type     object
Driven_Wheels         object
Number of Doors      float64
Market Category       object
Vehicle Size          object
Vehicle Style         object
highway MPG            int64
city mpg               int64
Popularity             int64
MSRP                   int64
dtype: object
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

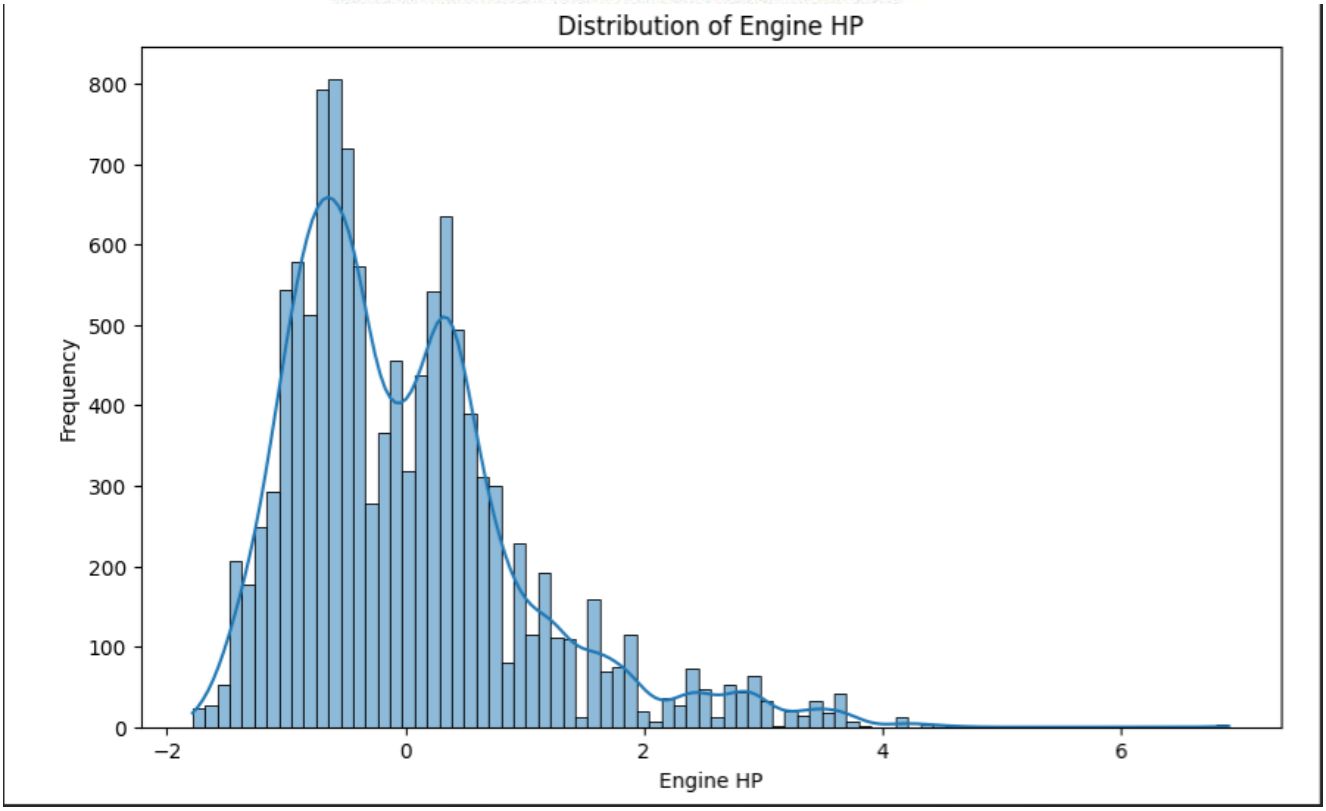
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11914 entries, 0 to 11913
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Make               11914 non-null  object
 1   Model              11914 non-null  object
 2   Year               11914 non-null  int64
 3   Engine Fuel Type   11911 non-null  object
 4   Engine HP          11845 non-null  float64
 5   Engine Cylinders   11884 non-null  float64
 6   Transmission Type  11914 non-null  object
 7   Driven_Wheels      11914 non-null  object
 8   Number of Doors    11908 non-null  float64
 9   Market Category    8172 non-null   object
 10  Vehicle Size       11914 non-null  object
 11  Vehicle Style      11914 non-null  object
 12  highway MPG        11914 non-null  int64
 13  city mpg           11914 non-null  int64
 14  Popularity         11914 non-null  int64
 15  MSRP               11914 non-null  int64
dtypes: float64(3), int64(5), object(8)
memory usage: 1.5+ MB

Dataframe Info:
None
```

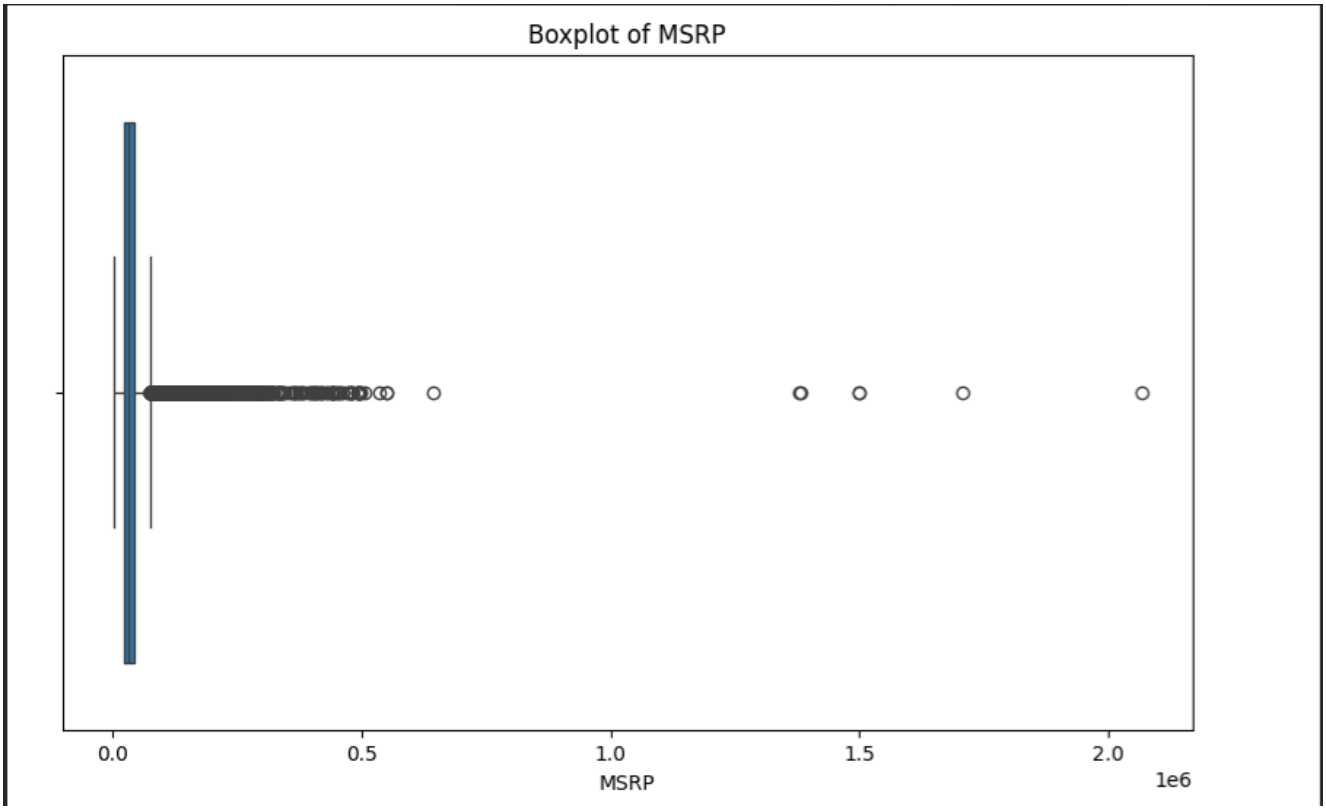
```
Summary Statistics:
                Year    Engine HP  Engine Cylinders  Number of Doors  \
count  11914.000000  11845.00000      11884.000000     11908.000000
mean    2010.384338    249.38607          5.628829         3.436093
std        7.579740    109.19187          1.780559         0.881315
min     1990.000000     55.00000          0.000000         2.000000
25%     2007.000000    170.00000          4.000000         2.000000
50%     2015.000000    227.00000          6.000000         4.000000
75%     2016.000000    300.00000          6.000000         4.000000
max     2017.000000   1001.00000         16.000000         4.000000

        highway MPG      city mpg    Popularity          MSRP
count  11914.000000  11914.000000  11914.000000  1.191400e+04
mean      26.637485     19.733255   1554.911197  4.059474e+04
std        8.863001      8.987798   1441.855347  6.010910e+04
min       12.000000      7.000000      2.000000  2.000000e+03
25%       22.000000     16.000000    549.000000  2.100000e+04
50%       26.000000     18.000000   1385.000000  2.999500e+04
75%       30.000000     22.000000   2009.000000  4.223125e+04
max      354.000000    137.000000   5657.000000  2.065902e+06
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

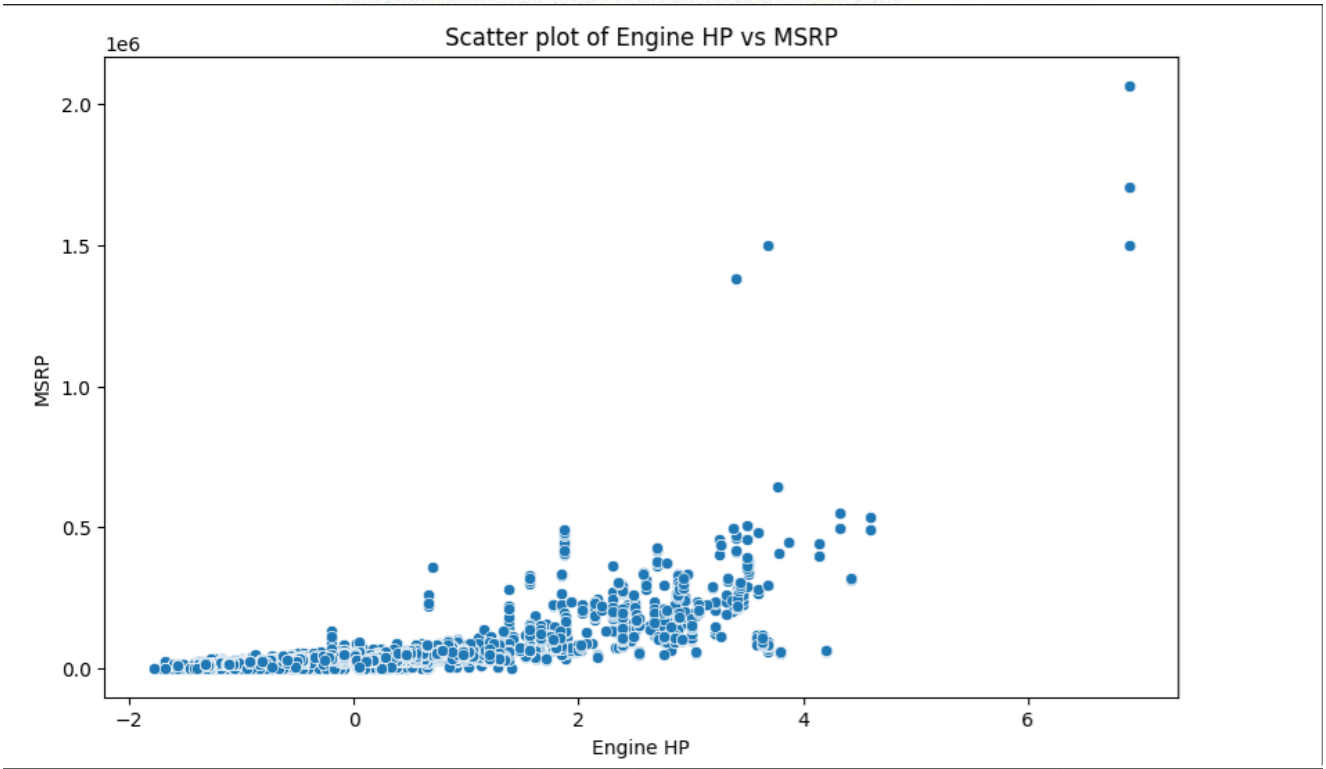
```
Mean Values:
Year                  2010.384338
Engine HP              249.256421
Engine Cylinders         5.624727
Number of Doors          3.436377
highway MPG             26.637485
city mpg                19.733255
Popularity            1554.911197
MSRP                 40594.737032
dtype: float64

Median Values:
Year                  2015.0
Engine HP              227.0
Engine Cylinders         6.0
Number of Doors          4.0
highway MPG             26.0
city mpg                18.0
Popularity            1385.0
MSRP                 29995.0
dtype: float64

Mode Values:
Year                  2015.0
Engine HP              200.0
Engine Cylinders         4.0
Number of Doors          4.0
highway MPG             24.0
city mpg                17.0
Popularity            1385.0
MSRP                  2000.0
Name: 0, dtype: float64

Standard Deviation Values:
Year                     7.579740
Engine HP              108.888444
Engine Cylinders         1.780189
Number of Doors          0.881184
highway MPG              8.863001
city mpg                 8.987798
Popularity            1441.855347
MSRP                 60109.103604
dtype: float64
```

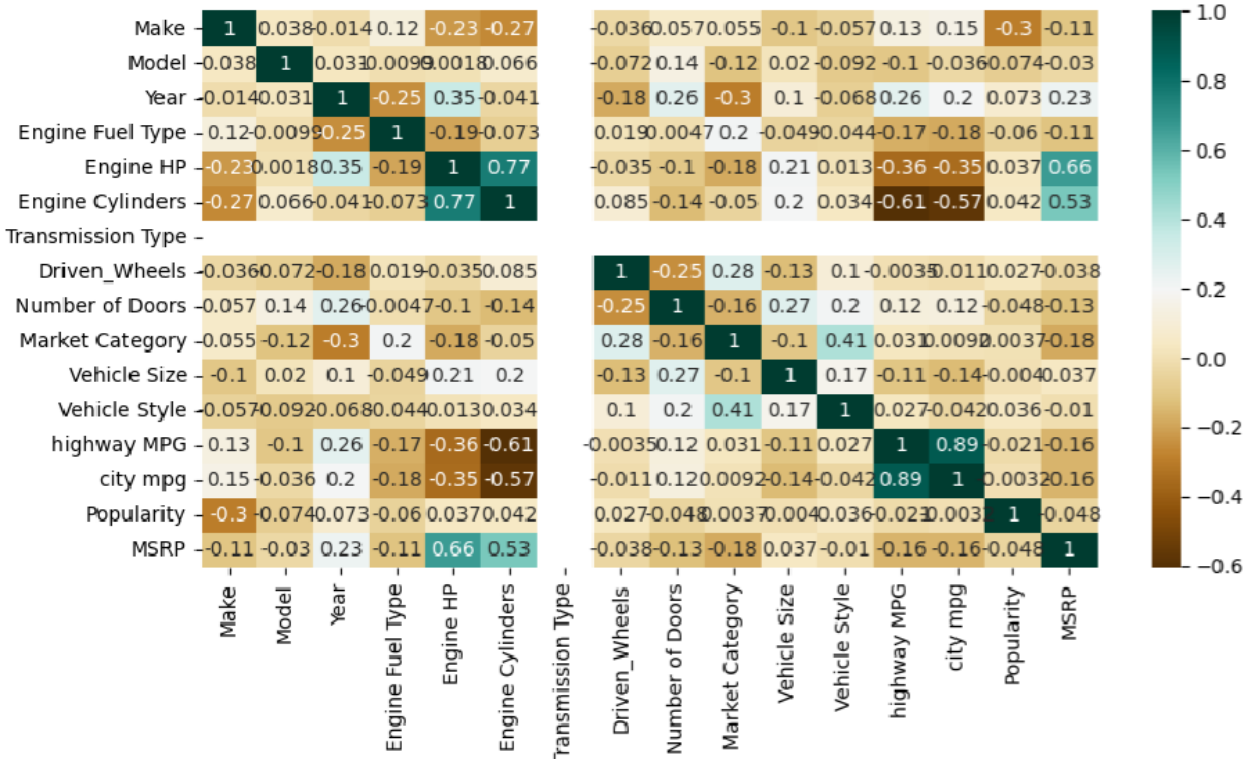
SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



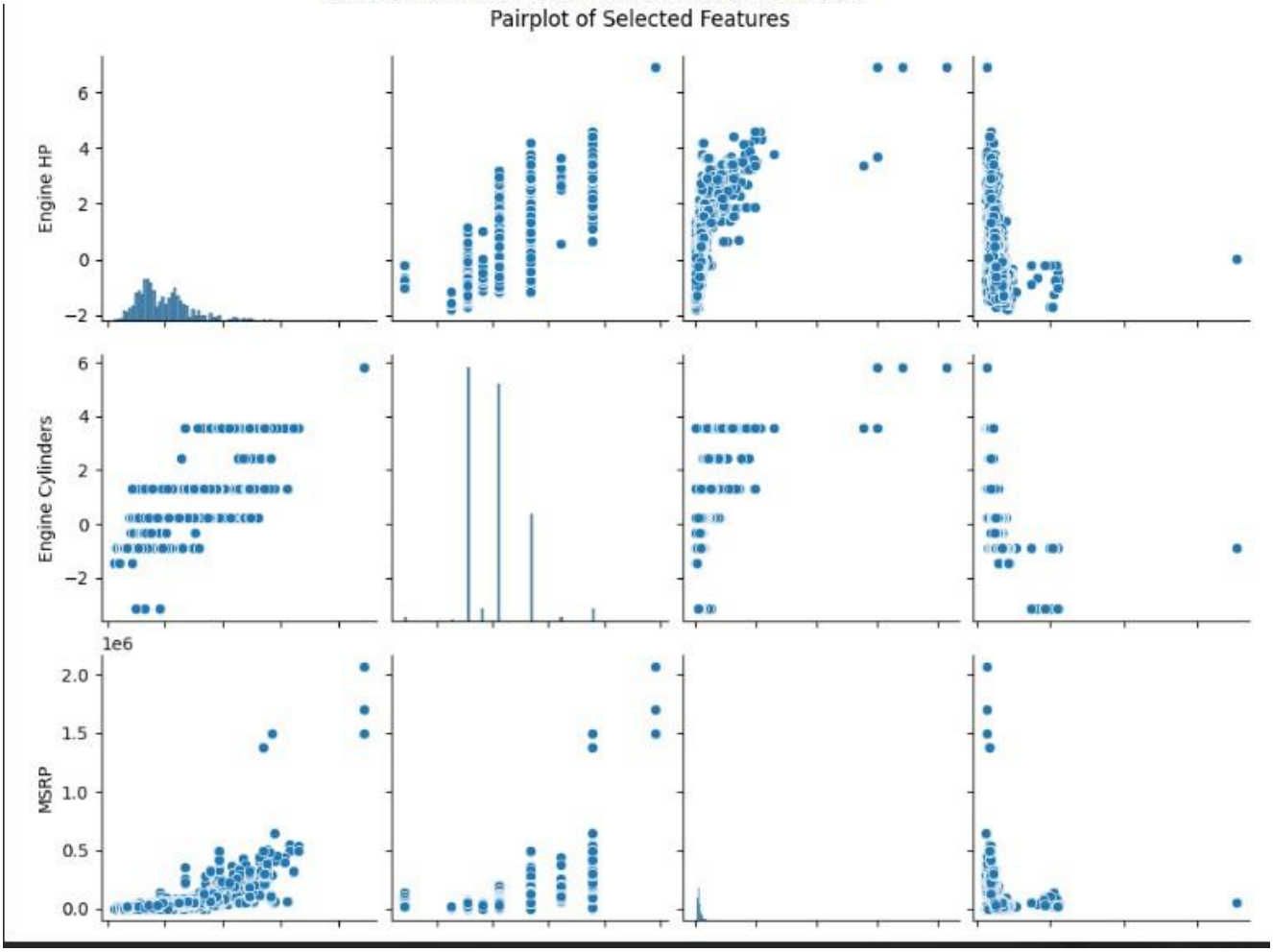


SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

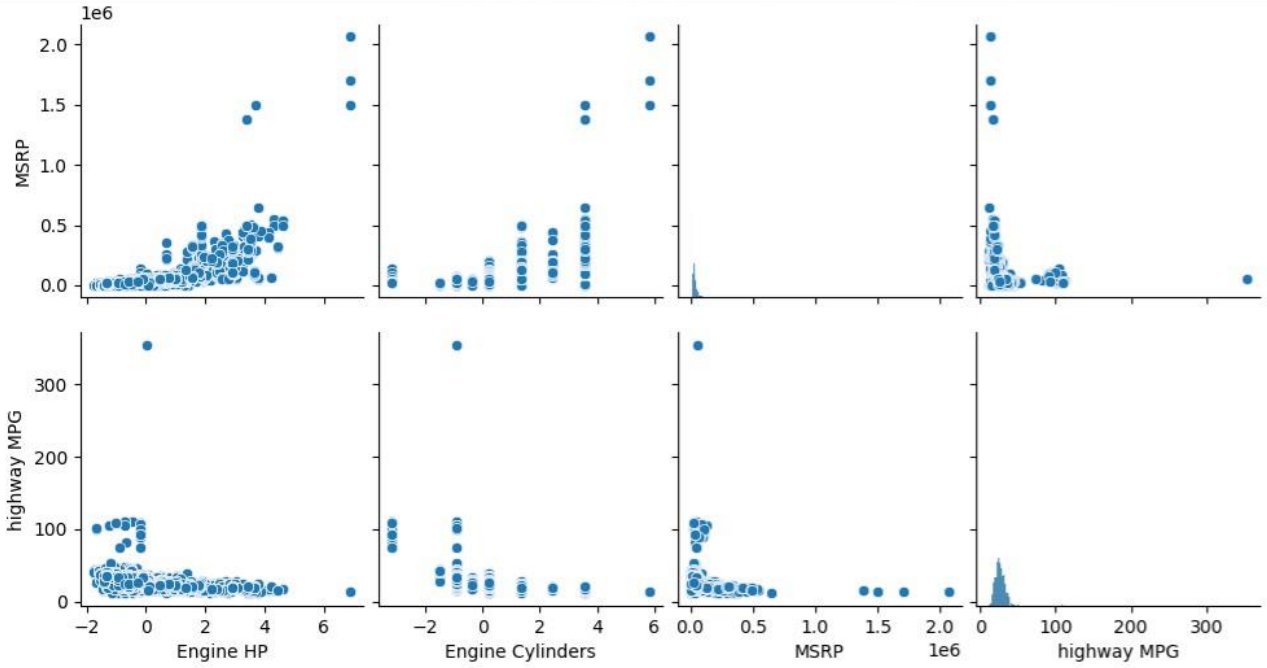




SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



- **QUESTIONS:**

1. What is the average highway MPG for vehicles with different engine fuel types?

2. How does engine horsepower vary across different vehicle sizes?

3. Is there a correlation between the number of engine cylinders and the MSRP of a vehicle?

4. Which transmission type is most common among the most popular vehicle models?

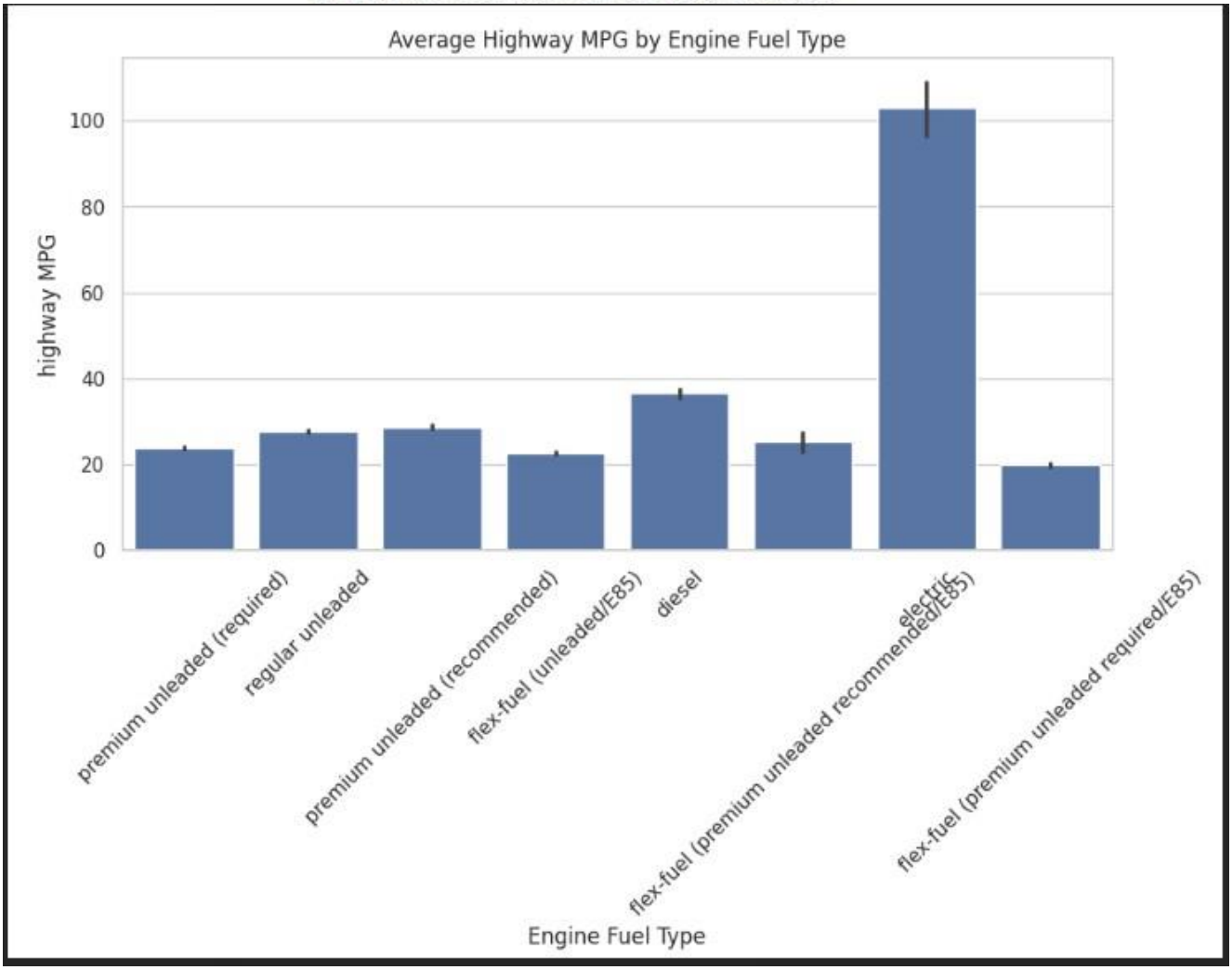
5. How does city MPG compare between all-wheel drive and front-wheel drive vehicles?

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

# Load the dataset (replace 'cars.csv' with the actual filename)
df = pd.read_csv("cars.csv")

# Drop rows with missing values to avoid errors in analysis
df.dropna(inplace=True)

# Set Seaborn style
sns.set(style="whitegrid")

# 1. Average Highway MPG by Engine Fuel Type
```

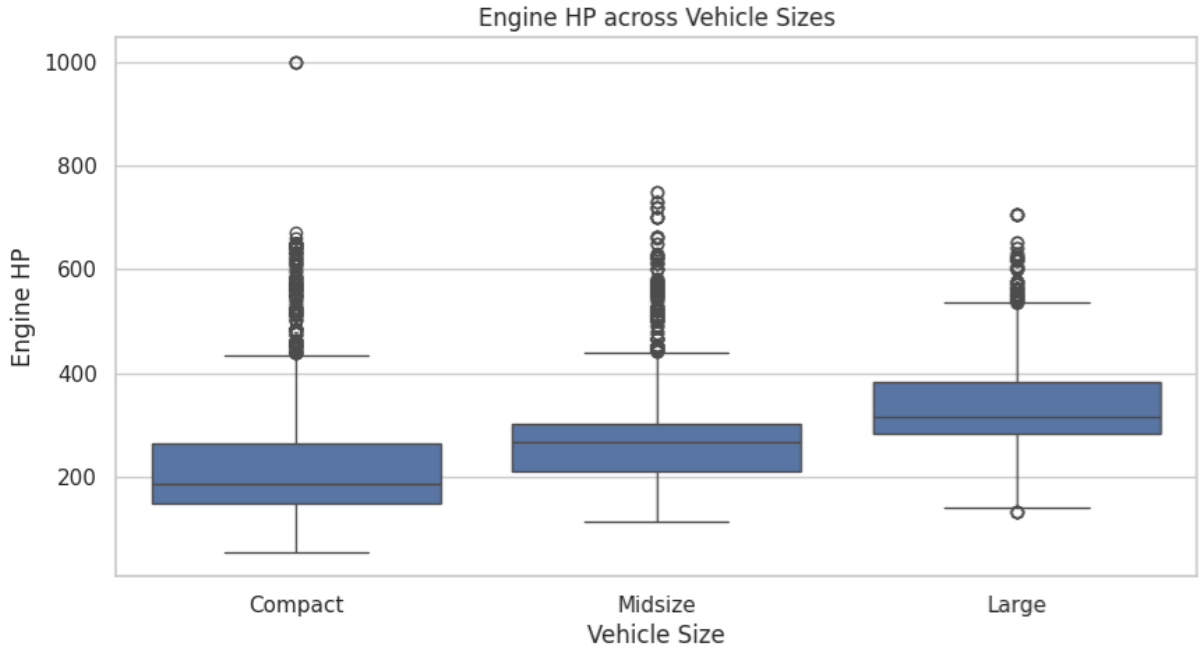
SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

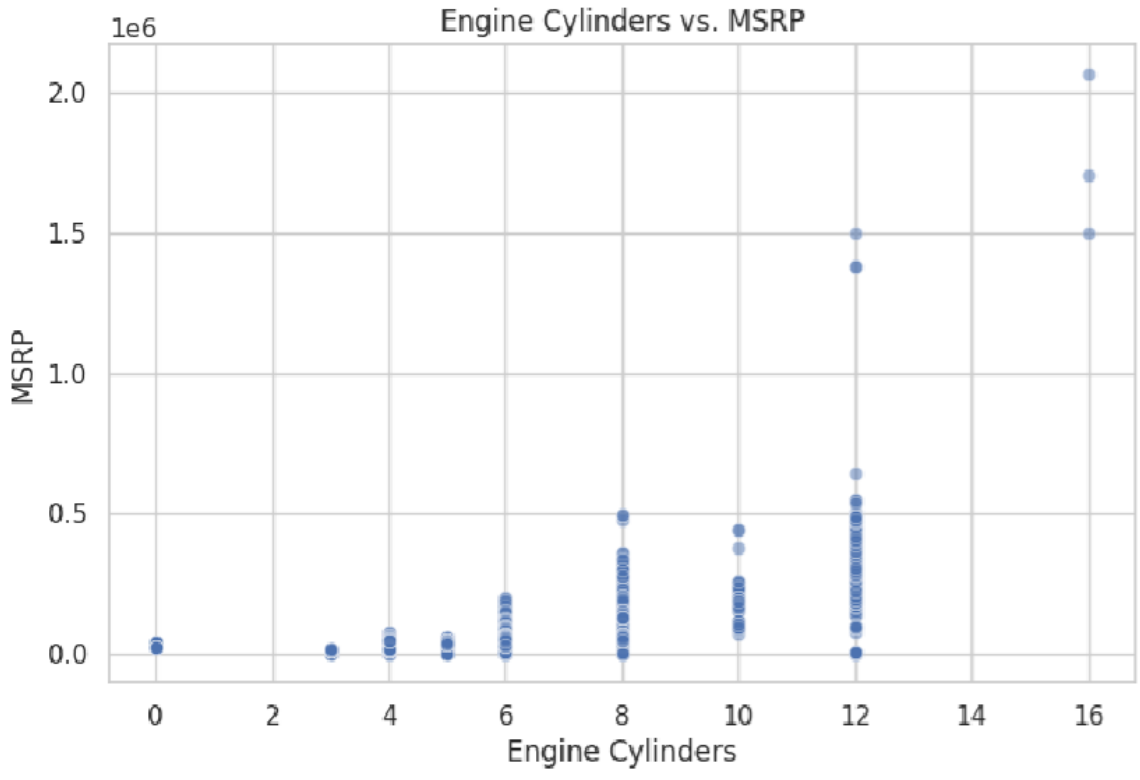
```
plt.figure(figsize=(10, 5))
sns.barplot(x="Engine Fuel Type", y="highway MPG", data=df, estimator=lambda x:
x.mean())
plt.xticks(rotation=45)
plt.title("Average Highway MPG by Engine Fuel Type")
plt.show()

# 2. Engine Horsepower vs. Vehicle Size
plt.figure(figsize=(10, 5))
sns.boxplot(x="Vehicle Size", y="Engine HP", data=df)
plt.title("Engine HP across Vehicle Sizes")
plt.show()

# 3. Correlation between Engine Cylinders and MSRP
plt.figure(figsize=(8, 5))
sns.scatterplot(x="Engine Cylinders", y="MSRP", data=df, alpha=0.5)
plt.title("Engine Cylinders vs. MSRP")
plt.show()

# 4. Count of Transmission Types in Most Popular Models
top_models = df.groupby("Make")["Popularity"].sum().nlargest(10).index
popular_cars = df[df["Make"].isin(top_models)]

plt.figure(figsize=(10, 5))
sns.countplot(x="Transmission Type", data=popular_cars,
order=popular_cars["Transmission Type"].value_counts().index)
plt.xticks(rotation=45)
plt.title("Most Common Transmission Types in Popular Car Models")
plt.show()

# 5. City MPG Comparison for Different Driven Wheels
plt.figure(figsize=(10, 5))
sns.boxplot(x="Driven_Wheels", y="city mpg", data=df)
plt.xticks(rotation=45)
plt.title("City MPG by Driven Wheels")
plt.show()
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

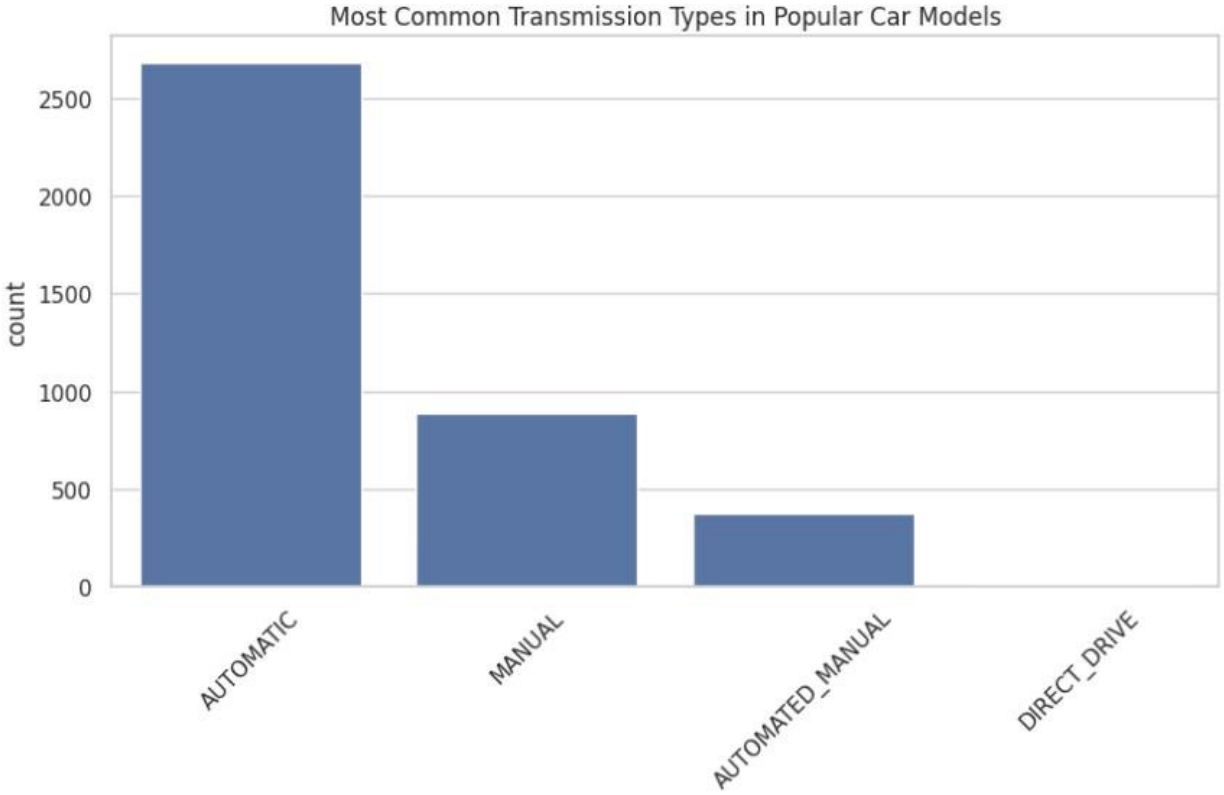


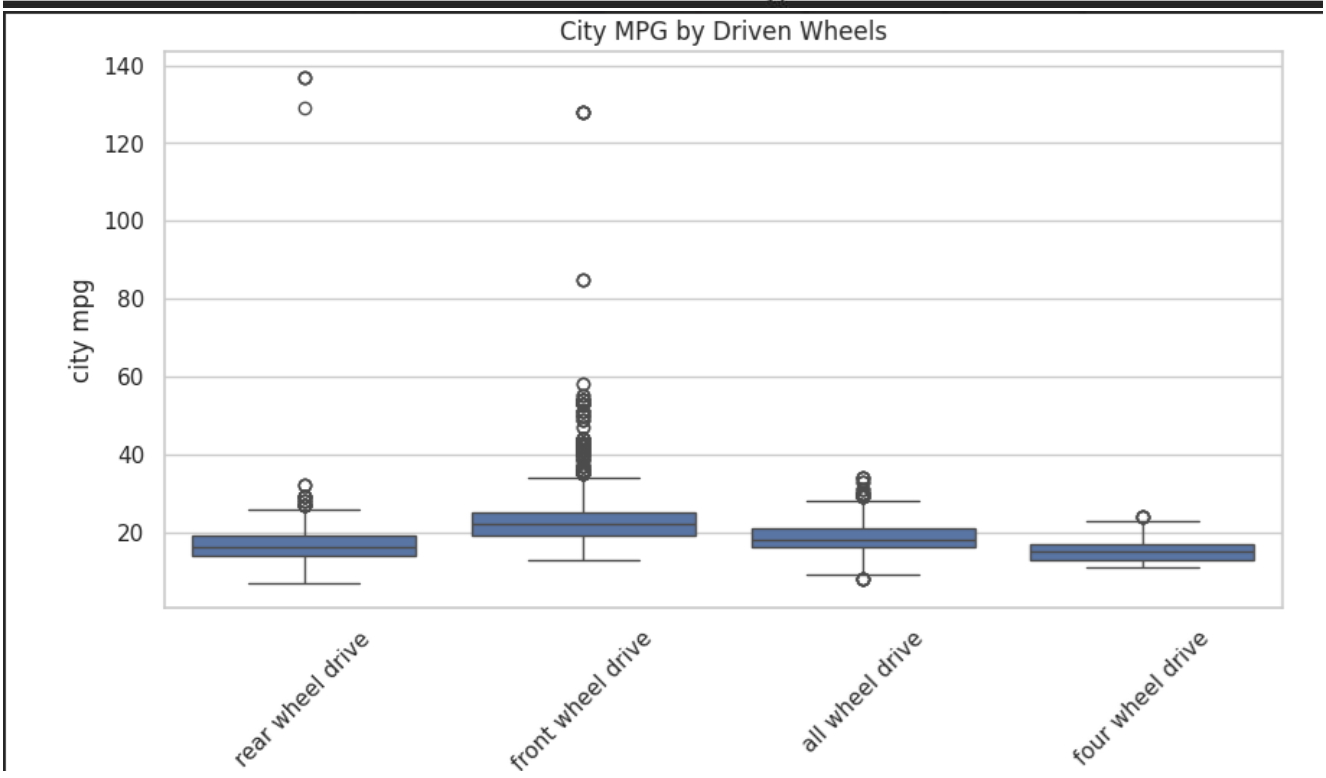
SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)





SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)





OBSERVATIONS / DISCUSSION OF RESULT:

1. List all the tasks that you have considered while doing preprocessing.

2. Compare the raw input data with cleaned data and list the benefits.

CONCLUSION:

Performed data preprocessing in terms of handling, missing data, removing outliers, eliminating duplicate rows and modifying the datatype, etc.

REFERENCES:

(List the references as per format given below and citations to be included the document)

1. Ethem Alpaydın, “Introduction to Machine Learning”, 4th Edition, The MIT Press, 2020.

2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012.

3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.

4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1st Edition, O'reilly, 2016.

5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012.

Website References:

[1] https://www.sciencedirect.com/topics/engineering/data-preprocessing

[2] https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/

[3]https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/