3)Write a program to perform Exploratory Data Analysis (EDA) for Classification using Pandas and Matplotlib

from google.colab import files
uploaded = files.upload()

Dataset link:-https://gist.github.com/noamross/e5d3e859aa0c794be10b
Or search for (Cars.csv ml dataset)

```python
# Importing the required libraries
import pandas as pd
import numpy as np
import seaborn as sns #visualisation
import matplotlib.pyplot as plt #visualisation
%matplotlib inline
sns.set(color_codes=True)
```

```python
import io
df = pd.read_csv(io.BytesIO(uploaded['CARS.csv']))

df.head(5)
```

```python
# Removing irrelevant features
df = df.drop(['Model','DriveTrain','Invoice', 'Origin', 'Type'],
axis=1)
df.head(5)
```

```python
df.info()
```

```python
df.describe()
```

```python
# Getting the number of instances and features
df.shape
```

```python
# Getting the dimensions of the data frame
df.ndim
```

```python
#Finding duplicate data
df = df.drop_duplicates(subset='MSRP', keep='first')
df.count()
```

```python
# Finding the null values
```

```python
print(df.isnull().sum())
```

```python
# Printing the null value rows
df[240:242]
```

```python
# Filling the rows with the mean of the column
val = df['Cylinders'].mean()
df['Cylinders'][247] = round(val)
val = df['Cylinders'].mean()
df['Cylinders'][248]= round(val)
```

```python
# Removing the formatting
df['MSRP'] = [x.replace('$', '') for x in df['MSRP']]
df['MSRP'] = [x.replace(',', '') for x in df['MSRP']]
df['MSRP']=pd.to_numeric(df['MSRP'],errors='coerce')
```

```python
#detecting outliers
sns.boxplot(x=df['MSRP'])
```
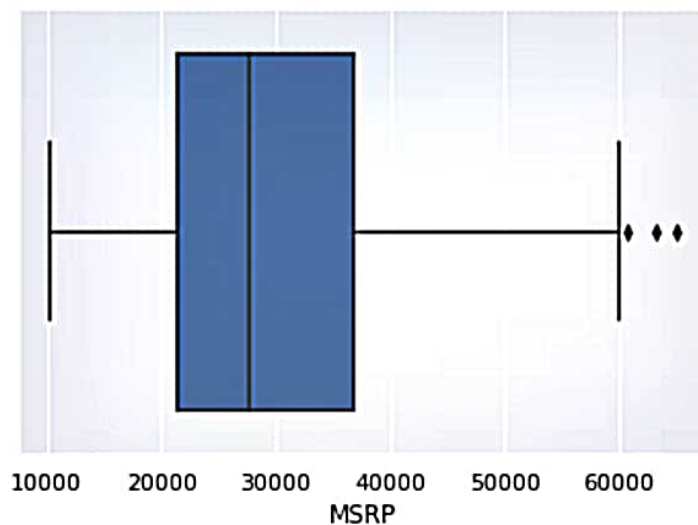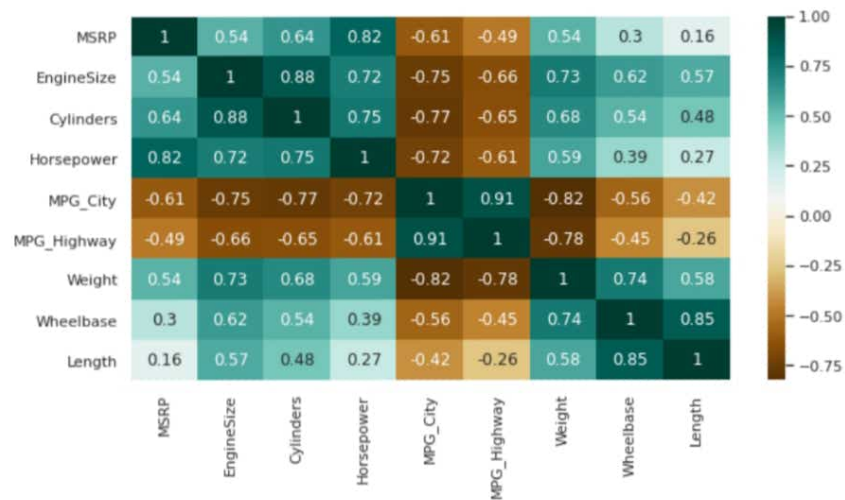


```python
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3-Q1
print(IQR)
```

```python
df = df[~((df < (Q1-1.5 * IQR)) |(df > (Q3 + 1.5 * IQR))).any(axis=1)]
```

```python
sns.boxplot(x=df['MSRP'])
```

```python
# Plotting a heat map
plt.figure(figsize=(10,5))
c= df.corr()
```

```
sns.heatmap(c,cmap='BrBG',annot=True)
```

| | MSRP | EngineSize | Cylinders | Horsepower | MPG_City | MPG_Highway | Weight | Wheelbase | Length |
|---|---|---|---|---|---|---|---|---|---|
| MSRP | 1 | 0.54 | 0.64 | 0.82 | -0.61 | -0.49 | 0.54 | 0.3 | 0.16 |
| EngineSize | 0.54 | 1 | 0.88 | 0.72 | -0.75 | -0.66 | 0.73 | 0.62 | 0.57 |
| Cylinders | 0.64 | 0.88 | 1 | 0.75 | -0.77 | -0.65 | 0.68 | 0.54 | 0.48 |
| Horsepower | 0.82 | 0.72 | 0.75 | 1 | -0.72 | -0.61 | 0.59 | 0.39 | 0.27 |
| MPG_City | -0.61 | -0.75 | -0.77 | -0.72 | 1 | 0.91 | -0.82 | -0.56 | -0.42 |
| MPG_Highway | -0.49 | -0.66 | -0.65 | -0.61 | 0.91 | 1 | -0.78 | -0.45 | -0.26 |
| Weight | 0.54 | 0.73 | 0.68 | 0.59 | -0.82 | -0.78 | 1 | 0.74 | 0.58 |
| Wheelbase | 0.3 | 0.62 | 0.54 | 0.39 | -0.56 | -0.45 | 0.74 | 1 | 0.85 |
| Length | 0.16 | 0.57 | 0.48 | 0.27 | -0.42 | -0.26 | 0.58 | 0.85 | 1 |

```
# Plotting a scatter plot
fig, ax = plt.subplots(figsize=(5,5))
ax.scatter(df['Horsepower'], df['MSRP'])
plt.title('Scatter plot between MSRP and Horsepower')
ax.set_xlabel('Horsepower')
ax.set_ylabel('MSRP')
plt.show()
```

Output:-



Scatter plot between MSRP and Horsepower