**DEPARTMENT OF ROBOTICS & ARTIFICIAL INTELLIGENCE**

# Total Marks: 04

**Obtained Marks:**

Programming for Artificial Intelligence

**Assignment # 04**

**Last date of Submission: 20May 2024**

# Submitted To: Mr. Saad Irfan Khan

**Student Name: Syed Muhammad Mustafa**

# Reg. Number: 22108130

**DEPARTMENT OF ROBOTICS & ARTIFICIAL INTELLIGENCE**

***Instructions****: Copied or shown assignments will be marked zero. Late submissions are not entertained in any case.*

**CLO 4 – PLO B, D – C4**

**Scenario: Data Visualization with Matplotlib** **(4 Marks)**

Apply Matplotlib to visualize a publicly available dataset related to AI or ML. You are required to do the following:

1. Choose a dataset of your choice.
2. Explore data, handle missing values, and preprocess.
3. Create a line plot, bar plot, scatter plot, histogram, and pie chart using Matplotlib.
4. Use advanced features like subplotting and customization.
5. Analyze and interpret each plot's insights and discuss Matplotlib's utility for AI projects.

**Deliverables:**

1. Jupyter/Colab Notebook with code and visualizations (softcopy on GCR).
2. Dataset info, Code, annotated plots and a summary report (hardcopy submission) discussing insights and Matplotlib's effectiveness.

## Dataset Information:

|  |
| --- |
| The dataset contains various features related to cars, such as 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, and MSRP. Here is a brief overview of the dataset: |

## Dataset:

|  |
| --- |
| Make: 48 unique values  Model: 723 unique values  Year: 28 unique values (int64)  Engine Fuel Type: 9 unique values  Engine HP: 330 unique values (float64)  Engine Cylinders: 10 unique values (float64)  Transmission Type: 5 unique values  Driven Wheels: 4 unique values  Number of Doors: 4 unique values (float64)  Market Category: 71 unique values  Vehicle Size: 3 unique values  Vehicle Style: 16 unique values  highway MPG: 58 unique values (int64)  city mpg: 68 unique values (int64)  Popularity: 48 unique values (int64)  MSRP: 4680 unique values (int64) |
|  |

## Dataset Screenshot:

|  |
| --- |
|  |

­

## Code:

|  |
| --- |
| import pandas as pd  df = pd.read\_csv('Car Features and MSRP/data.csv')  df.head(5)  df.tail(5) |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| df.**shape**  df.**columns**  df.**dtypes** |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| df.info()  df.select\_dtypes(object).info()  df.select\_dtypes('float64').info()  df.select\_dtypes('int64').info() |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| \_object\_type\_columns = df.select\_dtypes(object).**shape**[1]  \_float\_type\_columns = df.select\_dtypes('float64').**shape**[1]  \_int\_type\_columns = df.select\_dtypes('int64').**shape**[1]  print(f'Columns with object data type  : {\_object\_type\_columns}')  print(f'Columns with float64 data type : {\_float\_type\_columns}')  print(f'Columns with int64 data type   : {\_int\_type\_columns}')  print(f'Shape of the entire dataset    : {df.**shape**[1]} ')  df.describe() |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| df.isnull().sum()  total\_nullvalues = df.isnull().sum().sum()  print(f'Total Null Values in the dataset : {total\_nullvalues}') |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| columns\_with\_nullvalues = df.**columns**[df.isnull().any()].tolist()  columns\_with\_nullvalues  print('Datatypes of columns with null values :')  for column in columns\_with\_nullvalues:      print(f'{column}: {df[column].**dtype**}') |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| print('Number of unique values in the features containing null values : ')  for column in columns\_with\_nullvalues:      print(f'{column} : {df[column].nunique()}')      print() |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| df\_NullFreeMarketCategory = df.copy()  df\_NullFreeMarketCategory = df\_NullFreeMarketCategory.dropna(*subset*=['Market Category'])  from sklearn.impute import SimpleImputer  impute\_categorical = SimpleImputer(*strategy*='most\_frequent')  df\_NullFreeMarketCategory[['Engine Fuel Type']] = impute\_categorical.fit\_transform(df\_NullFreeMarketCategory[['Engine Fuel Type']])  impute\_numerical = SimpleImputer(*strategy*='mean')  df\_NullFreeMarketCategory['Engine HP'] = impute\_numerical.fit\_transform(df\_NullFreeMarketCategory[['Engine HP']])  df\_NullFreeMarketCategory['Engine Cylinders'] = impute\_numerical.fit\_transform(df\_NullFreeMarketCategory[['Engine Cylinders']])  df\_NullFreeMarketCategory['Number of Doors'] = impute\_numerical.fit\_transform(df\_NullFreeMarketCategory[['Number of Doors']])  df.isnull().sum()  df\_NullFreeMarketCategory.isnull().sum() |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| from tabulate import tabulate  table = []  for column  in df\_NullFreeMarketCategory:      table.append([column, df\_NullFreeMarketCategory[column].nunique(), df\_NullFreeMarketCategory[column].**dtype**])  def table\_info():      print(tabulate(table, *headers* = ['Feature Name', 'Unique Values', 'Data Type'], *tablefmt* = 'grid'))  table\_info() |

## Output:

|  |
| --- |
|  |

## Code:

|  |
| --- |
| import matplotlib.pyplot as plt  data = df\_NullFreeMarketCategory  counts = data.groupby(['Year', 'Vehicle Size']).size().unstack(*fill\_value*=0)  counts.**plot**(*kind*='line', *marker*='o', *figsize*=(10, 6))  plt.title('Vehicle Size Trends Over Years')  plt.xlabel('Year')  plt.ylabel('Count')  plt.legend(*title*='Vehicle Size')  plt.grid(True)  plt.tight\_layout()  plt.show() |

## Output:

|  |
| --- |
|  |

## Insights:

|  |
| --- |
| 1. We can see the highest vehicle size trends was mostly in the favour of 'Midsize'. The 'Large' did peak at mid 2005 - 2010, however it quickly nosedived  2. Dominance of Midsize Vehicles: As noted, the line plot indicates that the highest vehicle size trends were mostly in favor of 'Midsize' vehicles. This suggests that midsize vehicles have been consistently popular over the years, maintaining a significant portion of the market share.  3. Temporal Peaks and Declines: The observation regarding the 'Large' vehicle size peaking around mid-2005 to 2010, followed by a decline, indicates a temporal fluctuation in consumer preferences or market dynamics. This could be influenced by various factors such as economic conditions, fuel prices, consumer preferences, and automotive industry trends. |

## Code:

|  |
| --- |
| plt.figure(*figsize*=(15, 8))  plt.bar(data['Make'], data['MSRP'])  plt.title('MSRP by Make')  plt.xlabel('Make')  plt.ylabel('MSRP')  plt.xticks(*rotation*=45, *ha*='right', *fontsize*=10)  plt.tight\_layout()  plt.grid(*axis*='y')  plt.show() |

## Output:

|  |
| --- |
|  |

## Insights:

|  |
| --- |
| 1. We can cleary see the highest MSRP by make is of the 'Bugatti', folloewd by 'Lamborghini' and 'MayBach'  2. Top MSRP Manufacturers: As highlighted, the visualization clearly shows that 'Bugatti' has the highest average MSRP among all makes, followed by 'Lamborghini' and 'Maybach'. This suggests that vehicles from these manufacturers are generally priced at a premium compared to others. |

## Code:

|  |
| --- |
| data\_Dodge = data[data['Make']=='Dodge']  plt.figure(*figsize*=(8, 8))  plt.scatter(data\_Dodge['Engine HP'], data\_Dodge['highway MPG'], *c*=data\_Dodge['Year'], *label*="Dodge")  plt.xlabel("Engine HP")  plt.ylabel("highway MPG")  plt.legend()  plt.title("Scatter Plot of Engine HP vs highway MPG for Dodge Cars")  plt.colorbar(*label*='Year')  plt.tight\_layout()  plt.show() |

## Output:

|  |
| --- |
|  |

## Insights:

|  |
| --- |
| 1. Trend Over Time: There seems to be a general trend of improvement in highway MPG over the years, particularly evident in the observations provided:     - In 1990, Dodge cars typically achieved highway MPG between 22.5 and 32.5, with engine horsepower ranging from 100 to 200.     - By 1995, there's a slight decrease in MPG, with a narrower range of around 22.5 to 25.0 MPG, but with a wider range of engine horsepower, typically between 200 and 400.     - From 2000 to 2005, there's a wider spread of highway MPG, ranging from 17.5 to 27.5, and a wider range of engine horsepower from 150 to 400, suggesting some variability in Dodge's offerings during this period.     - Between 2010 and 2015, there's a noticeable increase in both highway MPG and engine horsepower. MPG ranges from 20.0 to 30.0, with engine horsepower ranging from 180 to 700. This suggests a significant improvement in both fuel efficiency and engine performance over time.    2. Trade-off between HP and MPG: Generally, there seems to be a trade-off between engine horsepower and highway MPG, as evident from the scatter plot. Vehicles with higher horsepower tend to have lower MPG, and vice versa  3. Model Diversification: The spread of data points across different years suggests that Dodge has offered a diverse range of vehicle models with varying engine characteristics over time. This indicates that Dodge has adapted its product lineup to meet changing consumer preferences and regulatory requirements, resulting in a varied mix of vehicle options |

## Code:

|  |
| --- |
| data['Engine HP'].unique()  plt.figure(*figsize*=(10, 6))  plt.hist(data['Engine HP'], *bins*=30, *color*='crimson', *edgecolor*='black')  plt.title('Distribution of Engine Horsepower (HP)')  plt.xlabel('Engine HP')  plt.ylabel('Frequency')  plt.grid(True)  plt.tight\_layout()  plt.show() |

## Output:

|  |
| --- |
|  |

## Insights:

|  |
| --- |
| 1. Peak Frequency Range: The histogram peaks between 200 and 400 horsepower, indicating that a significant number of vehicles in the dataset have engine horsepower within this range. This suggests that engines with moderate to high horsepower are prevalent among the vehicles in the dataset.  2. Distribution Shape: The distribution appears to be right-skewed, with a gradual decrease in frequency as engine horsepower increases beyond the peak range. This suggests that while there are many vehicles with moderate horsepower, fewer vehicles have extremely high horsepower engines |

## Code:

|  |
| --- |
| df.**columns**  data['Engine Fuel Type'].unique()  plt.figure(*figsize*=(8, 8))  sizes = data['Engine Fuel Type'].value\_counts().**values**  labels = data['Engine Fuel Type'].value\_counts().**index**  total\_count = sum(sizes)  threshold = 5  def autopct\_fn(pct):      return f'{pct:.2f}%' if pct > threshold else ''  plt.pie(sizes, *labels*=None, *autopct*=autopct\_fn, *startangle*=140)  plt.title('Engine Fuel Type Distribution')  plt.axis('equal')  plt.legend(labels, *loc*="best", *fontsize*=10, *bbox\_to\_anchor*=(1, 0.5), *labels*=None)  plt.tight\_layout()  plt.show() |

## Output:

|  |
| --- |
|  |
|  |

## Insights:

|  |
| --- |
| 1.  Vehicles powered by regular unleaded fuel are widely manufactured and used (Prevalence of Traditional Fuel Types)  2.  Premium unleaded fuels, both recommended and required, collectively account for 42.48% of the distribution. This indicates a considerable market share for vehicles that require or recommend higher-octane gasoline.  3.  Flex-fuel vehicles epresent 10.78% of the distribution.This suggests a growing interest in vehicles that offer flexibility in fuel choice.  4.  Diesel, electric, and other flex-fuel variants represent relatively small percentages of the distribution, each below 5%. This indicates that while there is some adoption of alternative fuel technologies, they have not yet reached widespread usage compared to traditional gasoline-powered vehicles |

## Summary Report:

|  |
| --- |
| Vehicle Size Trends: The dominance of 'Midsize' vehicles over the years indicates consumer preference for midsize cars, likely due to a balance of size, fuel efficiency, and cost. The peak and subsequent decline of 'Large' vehicles around 2005-2010 suggest changing market conditions or consumer preferences.  MSRP by Make: The high MSRP of brands like Bugatti, Lamborghini, and Maybach highlights their position as luxury and performance car manufacturers. This contrasts with more affordable brands, emphasizing market segmentation.  Market Categories: The concentration in top market categories suggests the presence of popular segments that car manufacturers target, reflecting trends and consumer demand in the automotive market.  Dodge Cars: The scatter plot of Dodge cars reveals the relationship between engine horsepower and fuel efficiency over time, showing technological advancements and shifts in consumer preferences.  Engine Horsepower Distribution: The histogram shows that most cars have engine horsepower between 100 and 400 HP, indicating a focus on moderate performance vehicles.  Fuel Type Distribution: The pie chart indicates a preference for regular unleaded fuel, followed by premium options. The distribution reflects consumer choices and market availability of different fuel types |

## Effectiveness of Matplotlib:

|  |
| --- |
| Visualization: Matplotlib excels at displaying a wide range of data types and relationships, simplifying the identification of trends, category comparisons, and understanding of distributions.  Customization: The library's extensive customization options for plots (such as colors, labels, and grid lines) improve both clarity and visual appeal.  Interpretation: Annotated plots offer a clear, visual summary of complex datasets, making it easier to interpret and communicate insights effectively. |

|  |
| --- |
|  |