Assignment-2

1. Explain how data visualization assists in identifying trends and anomalies during Exploratory Data Analysis (EDA). Compare the usefulness of summary statistics and visual analysis in understanding data patterns.
2. Describe the concept of *feature scaling* and explain the difference between Standardization and Normalization. How does improper scaling affect distance-based algorithms?
3. Write Python code to generate a **box plot** and a **violin plot** for the same dataset. Explain how interpretations differ between both visualizations.
4. You are given a dataset containing *employee working hours, efficiency score, and job role*. Explain how you would use a **pairplot** to understand variable relationships. What patterns or insights can you derive?
5. Create an interactive **bar chart** using *Plotly* to compare average monthly sales for different regions. Describe how tooltips, hover labels, and interactive legends enhance data exploration
6. List the differences between *right-skewed* and *left-skewed* distributions. What would a long left tail in a dataset indicate?
7. In Matplotlib, which arguments allow you to:

   - Rotate x-axis labels
   - Add a grid
   - Add a title and axis labels

   Explain each with an example.

8. What is a **density plot** and how does it differ from a histogram? Mention one use case where density plots provide better insights.
9. Explain the difference between *subplot()*, *subplots()*, and *add_subplot()* in Matplotlib. When would you prefer using a **facet grid** instead of subplots?
10. You are designing a dashboard to analyze **customer churn** for a telecom company. Describe which interactive components (filters, dropdowns, sliders) you would add and what visualizations (heatmap, funnel chart, stacked bar) would best support managerial decisions.
11. Perform **Mean Imputation** and **Median Imputation** , **forward and backward fill**, **KNN Imputer** (K=2) one by one on the following dataset and show all calculation steps:

Attr1	Attr2	Attr3
10	NaN	2
NaN	20	4
15	NaN	1
25	30	NaN

Compute the imputed values for each missing entry.

12. **A. Category Grouping for Pie Chart**

You have the following market share distribution:

Brand	Market %
A	25
B	18
C	7
D	2
E	1
F	27
G	20

Write Python code to group all categories with **<5%** share into **"Others"**, and then plot a pie chart.

12. B. Bar Chart Customization

Write Python code to draw a bar chart for student marks in 5 subjects. Roughly draw the diagram in your answer sheet using pencils.
Include the following customizations:

- Display value labels above each bar
- Change bar colors
- Rotate x-axis labels
- Add title and axis labels
- Add edge color and transparency

13. What is the IQR method used for in statistics? Why is the IQR method preferred over standard deviation for detecting outliers in non-normal distributions? Which quartiles are used when applying the IQR method for outlier detection?

14. Define lower fence and upper fence in the context of outlier detection using IQR. How does the IQR method determine whether a data value is an outlier? Given Q1 = 12 and Q3 = 26, calculate: IQR-Lower fence, Upper fence

15. A dataset has Q1 = 18 and Q3 = 44. Using the IQR rule, determine if the value **90** is an outlier. For the dataset: **10, 12, 14, 15, 18, 22, 25, 27, 40**

- Find Q1 and Q3
- Compute IQR
- Identify outliers using the IQR rule

16. You are given: Q1 = 52, Q3 = 72. Identify whether the following values are outliers: **30, 55, 68, 75, 102**

17. Calculate the IQR and detect outliers for the set: **5, 7, 9, 12, 15, 18, 22, 55**

18. Question 5 (Challenge Question). Given the sales distribution of bikes across cities:

City	Sales

Delhi	500
Mumbai	620
Bangalore	410
Kolkata	280
Chennai	350

Write Python code to:

- Create a pie chart
- Explode the highest and lowest values
- Show percentages rounded to nearest integer
- Add a shadow and legend
- Rotate so Mumbai slice begins at 60 degrees

19. Create a pie chart for a survey of favourite social media platforms:

- Instagram: 40
- YouTube: 60
- WhatsApp: 50
- Snapchat: 20
- Facebook: 30

Include the following:

a) Explode the social media platform with the **lowest count**.
b) Show values as **percentages (two decimal places)**.
c) Apply a **start angle of 45 degrees**.
d) Add a **legend positioned outside** the chart.
e) Add a **shadow** and make the chart perfectly circular.

20. 📌 Scenario Question: Line Chart

A fitness tracking company recorded the daily step count of a user over one week. The data is shown below:

Day	Steps
Monday	5,200
Tuesday	6,800
Wednesday	7,500
Thursday	6,200
Friday	8,900
Saturday	10,500
Sunday	9,800

Using Python and Matplotlib, create a line chart to visualize this data.

Your task is to:

- Plot the days on the x-axis and step counts on the y-axis.
- Change the line style to dashed (--).
- Set the line color to blue.
- Increase the line width to 3.
- Add circular markers ('o') on each data point.
- Set the marker size to 10 and marker face color to red.
- Add a title to the chart: "Weekly Step Count Trend".
- Label the axes:
    - X-axis: "Days"
    - Y-axis: "Number of Steps"
- Add a grid to make the chart easier to read.

21. What is the purpose of using plt.figure() before plotting in Matplotlib? Explain with an example where it becomes necessary. How does plt.subplot() help in creating multiple visualizations within the same figure? Describe its format and give an example. Compare plt.subplot() and plt.subplots(). When would you prefer one over the other?

22. A financial analyst wants to compare: A line plot showing monthly revenue trends. A scatter plot showing customer growth. A bar chart showing profit per region. Explain how plt.figure() and plt.subplot() would help organize these three visualizations in one layout.

23. What is a waffle chart, and how does it differ from a pie chart when visualizing proportions? Give one real-world scenario where a waffle chart would be more effective than a bar chart, and justify your choice.

24. A health analytics company is studying how **exercise frequency** relates to **BMI and age**. The dataset includes columns like Age, BMI, Exercise_Hours_Per_Week, and Gender.

You want to:

- Visualize the relationship between **BMI and exercise hours**
- Use color to represent **gender**
- Make the scatter plot interactive so hovering reveals details about each person

Task:
Describe how you would create an **interactive scatter plot in Plotly** and mention at least **two insights** you might gain from the visualization.

25. A car manufacturer is analyzing **vehicle performance data**. The dataset includes: `Engine_Size`, `Fuel_Efficiency`, `Horsepower`, and `Car_Type`.

You want to:

- Compare **horsepower vs. fuel efficiency**
- Use different colors for **car types (Sedan, SUV, Hatchback, etc.)**
- Add hover tooltips showing engine size

Task:
Write how you would create a scatter plot in Plotly to visualize these relationships and explain what patterns you would expect to find.

26. An e-commerce company wants to analyze the **age distribution of its customers** using the column `Customer_Age`.

You want to:

- Create a histogram using Plotly to visualize the distribution
- Adjust the number of bins
- Use color to enhance readability
- Make the chart interactive (hover to see counts)

Task:
Explain how you would build this histogram in Plotly and describe what insights (e.g., dominant age groups, skewness) could be gained from it.

27. A human resources department is studying salary structure. The dataset contains `Annual_Salary` values for all employees.

You need to:

- Plot a histogram of employee salary distribution
- Add transparency (`opacity`) to highlight overlapping bin densities
- Enable interactive tooltips to display salary ranges
- Add titles and axis labels

Task:
Describe how to create this histogram in Plotly and discuss what trends it might reveal (e.g., salary clustering, possible inequality).

28. A retail company wants to compare the total sales of five product categories: **Electronics, Clothing, Groceries, Furniture, and Toys.**
The dataset includes `Category` and `Total_Sales`.

You want to:

- Create a **bar chart** showing total sales per category
- Apply different colors to each bar
- Add hover labels displaying exact sales values
- Add axis labels and a title

Task:
Explain how you would create this bar chart in Plotly and describe what insights you might gain (e.g., best-selling vs least-selling categories).

29. A digital marketing team tracks website visitor monthly. The dataset includes `Month` and `Visitors`.

You want to:

- Create a **bar chart** showing visitor count per month
- Use a single color shade with slight opacity
- Add a tooltip showing exact visitor numbers
- Sort months chronologically

Task:
Explain how to create this bar chart and state what trends you might notice (e.g., seasonal peaks).

30. What is multicollinearity in the context of regression analysis, and why is it considered a problem in predictive modeling? Explain how multicollinearity affects the interpretation of regression coefficients. Differentiate between partial correlation and multicollinearity in multiple regression models.

31. You are building a linear regression model to predict house prices using variables such as Square_Footage, Number_of_Rooms, Living_Area, and Property_Size.

32. You observe that Square_Footage, Living_Area, and Property_Size are highly correlated with each other. What does this indicate? How would you detect and quantify the degree of multicollinearity? Suggest two possible solutions to handle this issue.

33. How Histogram is different from Bar chart? Explain with the help of Python code of each.

34. You build a multiple regression model and find the following VIF values for predictors:

Variable	VIF
Education Level	2.4
Income	9.8
Years_of_Experience	11.5
Age	3.2

- Which variables indicate multicollinearity?
- Should any variable be removed? Why or why not?
- Propose two techniques (e.g., feature engineering, dimensionality reduction) to address the issue.

35. Suppose You are working as a data analyst for a retail company that operates in multiple cities. Management wants a dashboard that tracks key business metrics in real time. The dataset includes fields like City, Sales, Profit, Customer_Segment, and Order_Date.

How would you design an interactive dashboard (using tool Dash) that allows managers to :

- View total sales and profit trends over time
- Filter data by city and customer segment

- Quickly identify top-performing cities or products?

What visualizations and layout choices would you use to make the dashboard both informative and easy to use? Use dash python library for development of dashboards.

36. If you accidentally use Merge instead of Append while combining monthly sales files with identical structures, what problem might occur in your final Power BI dataset? Provide one real-world reporting impact.

37. What could go wrong if you use a Left Outer Join instead of a Full Outer Join while merging two customer and transaction tables? Give a possible impact on the final business dashboard in PowerBI.

38. You are building a Power BI report using a dataset with the fields:
Employee ID, Hours Worked, Hourly Rate, and Bonus Percent.

Your task is to calculate the employee's **Total Compensation**, which includes both wages and bonus.

You need to:

- Display the **Total Compensation value for each employee** in a table visual
- Show the **Total Company Compensation** in a card visual

Questions:

1. **Write the DAX formula to calculate Total Compensation both as:**
   - A **Calculated Column**
   - A **Measure**
2. **Explain the difference between using a calculated column and a measure in Power BI.**
   Specifically focus on:
   - When they are calculated
   - Whether they depend on user-selected filters
   - Storage impact
3. **Which one (calculated column or measure) would be more appropriate for the dashboard analysis, and why?**
   Support your answer with a real reporting reason.

39. A business analyst is studying relationships among **12 employee performance metrics** such as Productivity Score, Attendance Rate, Overtime Hours, Customer Feedback Rating, Project Completion Time, and Training Hours across 500 employees.

- Which visualization method(s) would best help analyze relationships among all performance metrics simultaneously?
- How would you interpret visual cues to identify highly correlated or weakly related variables?
- Why would using individual bar charts or line charts be inadequate for analyzing multi-variable relationships in this case?

40. A teacher wants to study the pattern of students' exam scores to evaluate overall performance. The marks (out of 100) are:
[55, 60, 72, 80, 65, 70, 90, 45, 85, 78, 62, 50, 95, 88, 74]

- Write Python code using **Seaborn** to create a histogram of the marks and include a **KDE (kernel density estimate) curve**.
- Provide a short interpretation of what the histogram suggests about how the students performed.

41. A retail analytics team wants to identify premium customers by calculating the total number of purchases where the **order value exceeds ₹75,000**.

- Write a DAX measure using **CALCULATE()** and **FILTER()** to count only those high-value transactions.
- Explain why a basic **COUNT()** or **COUNTROWS()** would not be sufficient in this case.
- Discuss how applying such conditional calculations can support customer loyalty program targeting and pricing strategies.

42. An IoT-enabled greenhouse records temperature data every hour from three separate zones: **A, B, and C**.
The system engineer wants to visualize and compare temperature patterns to spot any anomalies.

Write a Python program using **Plotly** to generate an **interactive line chart** with markers that displays hourly temperature readings for each zone. Make sure the chart includes:

- Hour values labeled along the x-axis
- Unique colors and marker styles for each zone
- A clear title and properly labeled axes