**Data Science question Bank**

**UNIT - 2**

1. What is data wrangling? Describe the processes involved in cleaning, transforming, merging, and reshaping datasets during data wrangling.

2. Explain the methods for combining and merging datasets. How is merging on index different from concatenating datasets? Provide Python suitable examples.

3. What are missing data? Describe the different techniques to handle missing data during data cleaning and preparation.

4. Discuss the methods to detect and handle outliers, noise, and anomalies in large datasets. Why is handling anomalies important in data preparation?

5. import pandas as pd

import numpy as np

data = {

'Customer\_ID': ['C100', 'C101', 'C102', 'C103', 'C104'],

'Name': ['john doe', 'Alice\_Green', 'BOB SMITH', 'diya patel', 'Eva\_Lee'],

'Age': [25, 33, 28, 45, 22],

'Purchase\_Amount': [120, 250, 300, 50, 5000],

'City': ['New York', 'los angeles', 'Chicago', 'chicago', 'NEW YORK'],

'Rating': ['4.5', '3.2', '4.0', '2.5', '4.8']

}

df = pd.DataFrame(data)

Write python code for the following

* Convert all names to Title Case
* Find mean and median 'Purchase\_Amount'
* Number of unique cities

Convert the 'Age' column into 3 bins: "Young (18-25)", "Adult (26-40)", "Senior (41+)"

6. import pandas as pd

data = {

'Emp\_ID': ['E001', 'E002', 'E003', 'E004', 'E005'],

'Full\_Name': ['raj kumar', 'SITA SHARMA', 'vijay\_verma', 'ANITA JAIN', 'rahul\_singh'],

'Salary': [50000, 72000, 65000, 48000, 150000],

'Department': ['HR', 'Finance', 'IT', 'it', 'HR'],

'Experience\_Years': [2, 8, 5, 1, 12],

'Performance\_Score': ['3.8', '4.5', '3.2', '4.0', '4.9']

}

df = pd.DataFrame(data)

* Convert all names in the Full\_Name column to Title Case (e.g., "raj kumar" → "Raj Kumar").
* Calculate the mean and standard deviation of the Salary column.
* Find the number of unique departments.
* Categorize the Experience\_Years into the following bins:

"Junior (0–3)",

"Mid-level (4–8)",

"Senior (9+)".

7, You are given two datasets about students’ academic records. Write Python code using **pandas** to perform the following tasks:

import pandas as pd

# DataFrame 1

data1 = {

'Student\_ID': ['S001', 'S002', 'S003', 'S004'],

'Name': ['john\_doe', 'ALICE SMITH', 'bob\_jones', 'Clara Brown'],

'Math\_Score': [85, 78, 92, 67]

}

df1 = pd.DataFrame(data1)

df1.set\_index('Student\_ID', inplace=True)

# DataFrame 2

data2 = {

'Student\_ID': ['S001', 'S002', 'S003', 'S004'],

'Science\_Score': [88, 74, 90, 72]

}

df2 = pd.DataFrame(data2)

df2.set\_index('Student\_ID', inplace=True)

**Tasks:**

1. Merge the two dataframes **on the Student\_ID index**.
2. Convert all names in the **Name** column to **Title Case** and replace underscores \_ with spaces.
3. Suppose there is noise in the **Math\_Score** column (e.g., scores less than 0 or greater than 100 are invalid). Write code to **detect and correct** such anomalies by capping values within the 0–100 range.
4. Create a **pivot table** showing the average **Math\_Score** and **Science\_Score**.

8. import pandas as pd

df1 = pd.DataFrame({

'ID': [1, 2, 3],

'Name': ['Alice', 'Bob', 'Charlie'],

'Dept': ['HR', 'Tech', 'Finance']

})

df2 = pd.DataFrame({

'ID': [2, 3, 4],

'Salary': [50000, 60000, 45000],

'Join\_Date': ['2020-01-15', '2019-05-20', '2021-11-10']

})

* Concatenate df1 and df2 (axis = 0 and axis = 1)
* Merge df1 and df2 (inner, left, right and outer)
* Join df1 and df2

9. You are given two DataFrames representing employee and project details. Write Python code to perform the following operations:

import pandas as pd

df\_a = pd.DataFrame({

'Emp\_ID': [101, 102, 103],

'Name': ['Raj', 'Anita', 'Vikram'],

'Project': ['Alpha', 'Beta', 'Gamma']

})

df\_b = pd.DataFrame({

'Emp\_ID': [102, 103, 104],

'Hours\_Worked': [160, 150, 170],

'Month': ['Jan', 'Jan', 'Jan']

})

**Tasks:**

1. Concatenate df\_a and df\_b using axis=0 and axis=1.
2. Merge df\_a and df\_b on Emp\_ID using:
   * inner join
   * left join
   * right join
   * outer join
3. Use the join() function to combine the two DataFrames. Before joining, set Emp\_ID as the index for both.

UINT III

1.Explain the various customization options available in Matplotlib to control axes, ticks, labels, legends, and annotations. Give suitable examples.

2.Discuss the differences between various plot types in Matplotlib and Seaborn like line plots, bar charts, histograms, and box plots with use-cases.

3. You are given a dataset representing student performance. Using **Matplotlib** and **Seaborn**, draw different types of plots to analyze the data:

import pandas as pd

data = {

'Student': ['Alice', 'Bob', 'Charlie', 'David', 'Eva'],

'Maths': [78, 85, 60, 90, 95],

'Science': [88, 75, 70, 85, 92],

'English': [82, 80, 65, 78, 88]

}

df = pd.DataFrame(data)

**Tasks:**

1. Plot a **line chart** showing marks in Maths, Science, and English for each student.
2. Create a **bar chart** comparing Maths marks for all students.
3. Draw a **scatter plot** of Science vs. Maths scores.
4. Plot a **box plot** for marks in all three subjects using Seaborn.
5. Create a **pair plot** of the dataset using Seaborn.
6. Customize one of the plots by changing line styles, colors, and adding legends, titles, and grid.
7. Save one of the plots as a PNG image.

4. *import seaborn as sns*

*import pandas as pd*

*import numpy as np*

*np.random.seed(42)*

*data = pd.DataFrame({*

*'Age': np.random.randint(20, 70, 100),*

*'Income': np.random.normal(50000, 15000, 100).astype(int),*

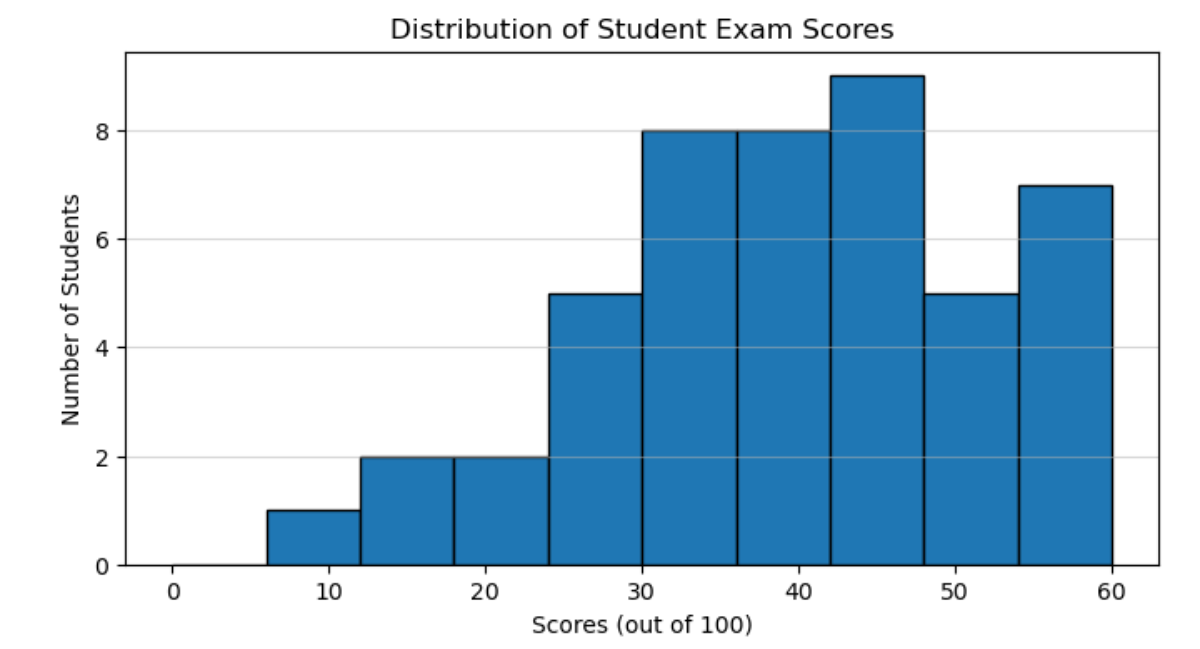
*'Spending': np.random.normal(0.8, 0.2, 100) \* np.random.normal(50000, 15000, 100).astype(int),*

*'Savings': np.random.uniform(1000, 20000, 100).astype(int)*

*})*

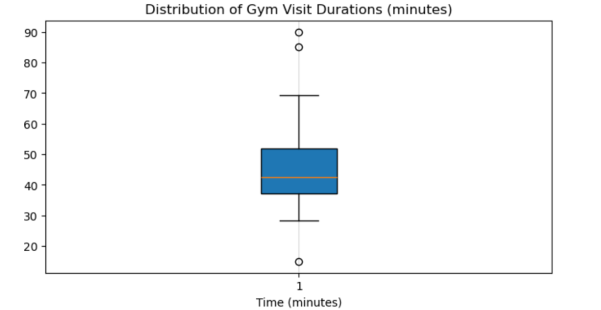
* Write python code to create the Pair plot
* If Savings vs. Age scatter plot had a negative slope, suggest solution.

If Age vs. Spending scatter plot shows no clear pattern. Interpret the relationship.

5. 

* If the passing score is 30, how many students failed?
* Estimate the total number of students represented in the histogram.

What type of distribution does this histogram represent?

6. 

* What is the median gym visit duration?
* How many outliers are shown in the plot?
* 50% of members stayed between Q1 and Q3. Ste True or False.
* If a visit lasts 100 minutes, would it be an outlier? Justify.

7. You are given the sales data for a store over 6 months:

months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun']

sales = [15000, 18000, 22000, 21000, 25000, 27000]

"Plot a line graph to show the store's sales trend over the first six months of the year. Use months as the X-axis and sales (in dollars) as the Y-axis."

"Create a bar chart representing the sales for each month from January to June. Which month had the highest and lowest sales?"

"Using the given sales data, draw a graph and analyze the change in sales between consecutive months. Highlight any months where sales dropped."

8.Create two subplots: the first showing a line plot of sales, the second showing the same data as a bar plot. Ensure each subplot has its own title and labeled axes.

9. Use Seaborn to plot a boxplot showing the distribution of sales data. Label axes and add a title.

10. Create a pair plot using a custom dataset with multiple sales-related features (e.g., monthly sales, customer count, profit). Analyze trends