Data Science – Set Question – Scenario Based Problem Solving

1. **Scenario:** An e-commerce company tracks delivery times (in minutes) for 15 orders:
   `[25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95]`

   The company wants to analyze the delivery performance using percentiles and detect if there are any unusual delivery times.

   Question 1:
   Calculate Q1 and Q3.

   Answer:

   Q1 can be obtained by finding the percentile value of 25th position.

   Q3 can be obtained by finding the percentile value of 75th position.

```
Q1 = np.percentile(Dataset, 25)
print("Q1 = ", Q1)
```

```
Q1 =  42.5
```

```
Q3 = np.percentile(Dataset, 75)
print("Q3 = ", Q3)
```

```
Q3 =  77.5
```

   Question 2:
   Find the Interquartile Range (IQR).

   Answer:

   IQR can be calculated by subtracting Q3 from Q1

   IQR = Q3 – Q1

```
IQR = Q3 - Q1

print("IQR = ", IQR)
```

```
IQR =  35.0
```

   Question 3:
   Detect Outliers using the IQR method.

   We can detect Lower Bound and Upper Bound Outliers using below formulas:
   LBO = Q1 – 1.5 * IQR
   UBO = Q3 + 1.5 * IQR

```
LBO = Q1 - (1.5 * IQR)

print("Lower Bound Outlier value is: ", LBO)
```

```
Lower Bound Outlier value is:  -10.0
```

```
UBO = Q3 + (1.5 * IQR)

print("Upper Bound Outlier value is: ", UBO)
```

```
Upper Bound Outlier value is:  130.0
```

Conclusion:
No Outliers detected.

---

2. **Scenario: Student Score Analysis**

A teacher is analyzing the mathematics scores of students in her class. The scores are:
[45, 50, 55, 60, 60, 62, 63, 65, 90, 95]

Question 1:
Calculate the mean, median, and mode of the scores.

$$\text{Mean} = \frac{45, 50, 55, 60, 60, 62, 63, 65, 90, 95}{10}$$

Median is the mid value of the dataset

Mode gives the value which occurred repeatedly

```
from scipy import stats
Dataset =  [45, 50, 55, 60, 60, 62, 63, 65, 90, 95]
Dataset = np.array(Dataset)
mean = Dataset.mean()
median = np.median(Dataset)
mode = stats.mode(Dataset, keepdims=True)
print("Mean = ", mean)
print("Median = ", median)
print("Mode = ", mode.mode[0])
```

```
Mean =  64.5
Median =  61.0
Mode =  60
```

Question 2:
Explain why the median might be a better representation than the mean in this case.

Answer:
The **median** is often a better representation than the **mean** when the dataset contains **outliers** or extreme values. The mean gets pulled in the direction of

unusually high or low scores, which can distort the overall picture of student performance. In contrast, the median is not affected by outliers, as it only considers the middle value of the dataset. This makes the median a more reliable indicator of the typical student's performance in such cases.

3. **Scenario: Grocery Store Customer Analysis**
   A grocery store manager tracks how many customers visit the store daily for a month:
   [5, 10, 8, 15, 20, 5, 12, 14, 10, 18]
   Question 1:
   Create a frequency distribution table for this data

```python
import pandas as pd

Dataset = [5, 10, 8, 15, 20, 5, 12, 14, 10, 18]

Data = pd.DataFrame({
    "Customers": pd.Series(Dataset).unique(),
    "Frequently visited (daily)": pd.Series(Dataset).value_counts().values
})

print(Data)
```

```
   Customers  Frequently visited (daily)
0          5                           2
1         10                           2
2          8                           1
3         15                           1
4         20                           1
5         12                           1
6         14                           1
7         18                           1
```

4. **Scenario**: **Real Estate Model Analysis**
   A real estate model has three variables:
   - House Size
   - Number of Rooms
   - Number of Bathrooms

   Question 1:
   How can you detect multicollinearity in this model?

   Answer:
   - A **Variation Inflation Factor (VIF)** helps to detect multicollinearity in the Regression Analysis.
   - Multicollinearity is when there is a correlation between Predictors in a model.

   Formula:

$$VIF = \frac{}{1 - R2}$$

- If the **VIF** value is **> 10,** then there is a Multicollinearity in the model.

---

5. **Scenario : Medicine Effectiveness Study**
   A company made a new medicine to lower blood pressure. They gave it to one group and gave a fake pill (placebo) to another group.
   Question 1:
   How can the company check if the new medicine works?

   Answer:
   - As we have been provided with two different groups and we want to check with one condition then we can go head with **T test.**
   - First we need to frame Null and Alternative Hypothesis
     - **Null Hypothesis:** The new medicine will not Lower the Blood Pressure
     - **Alternative Hypothesis:** The new medicine will Lower Blood Pressure

   - When T test is performed we will get P value. If P value is **< 0.05** then we can conclude the new medicine **will Lower** the **Blood Pressure** that means the medicine is **effective.**

---

6. **Scenario: Identifying Outliers in Sales data**

   A company wants to find any unusual spikes in sales.

   Question 1:
   How can the company detect outliers in their sales data?

   Answer:

   1) Calculate **Q1** and **Q3** (the 25th and 75th percentiles) of the dataset using the percentile formula.
   2) Compute the **Interquartile Range (IQR)** as:
      IQR=Q3–Q1
   3) Determine the **Lower Bound (LBO)** and **Upper Bound (UBO)** for outliers using the formulas:
      LBO = Q1 – (1.5 × IQR)
      UBO = Q3 + (1.5 × IQR)
   4) Identify outliers below the lower bound:
      - If any value < LBO → Outlier detected.
      - Otherwise → No outliers below LBO.
   5) Identify outliers above the upper bound:

- If any value > UBO → Outlier detected.
- Otherwise → No outliers above UBO.

7. **Scenario: Understanding Customer Satisfaction**
   A restaurant conducted a survey to rate customer satisfaction on a scale of 1 to 5:
   `[5, 4, 4, 5, 3, 4, 5, 2, 4, 3]`
   Question 1: How can the restaurant summarize the overall satisfaction?

   Answer:
   - As we have the ratings as a data, we can check the repeated rating using **Mode.**
   - If the Mode value is **4** or **5** then it means Customer are very well satisfied with this Restaurant.
   - If we want to study more about the Customer satisfaction level then we can also calculate Mean and Median.

```
Mode = stats.mode(Data, keepdims=True)
print("The Mode value is : ", Mode.mode[0])
```

```
The Mode value is :  4
```

```
Data =  np.array([5, 4, 4, 5, 3, 4, 5, 2, 4, 3])

Mean = Data.mean()
print("The Mean value is : ",Mean)
```

```
The Mean value is :  3.9
```

```
Median = np.median(Data)
print("The Median value is : ", Median)
```

```
The Median value is :  4.0
```