</talentlabs>

Data Analysis
Assignment Chapter 2

Instructions

1. You can take help from the lecture notes to revise the concepts that we have covered
2. Choose the best suitable answer and submit the word document
3. Please submit the assignment through TalentLabs Learning System.

Question 1. (5 points):
There are different methods of summarizing data. In this exercise you have to identify the methods which best fit the following scenarios:

#	Scenario	Best Method
E.g.	Average Salary of software engineers in the market	Average
1	Highest and lowest marks of students in the subject of Math	Max, Min
2	Outlier product sold by a company	Minimum (Q1 – 1.5*IQR), Maximum (Q3 + 1.5*IQR)
3	Most trending topic on Internet	Mode
4	Average exam score with different weight for different subjects	Weighted Mean
5	Number of products sold by product category	Sum, Group by

Question 2:
Suppose that a marketing firm conducts a survey of 100 households to determine the average number of Air Conditioners each household owns. Every household in the sample has at least one Air Conditioner and no household has more than four. Find the expected number of Air Conditioners per household. (Hint: Weighted Mean)

Number of refrigerators per Household	Number of Households
1	47
2	18
3	23
4	12

Weighted sum = (1 x 47) + (2 x 18) + (3 x 23) + (4 x 12) = 200

Weighted mean = 200 / (47+18+23+12)
= 2

Expected number of Air conditioners per household = 2

Question 3:
Consider a data set having the number of toys owned by different kids in a society. Find the median number of toys found in the society
Number of toys owned by each kid = [1, 2, 7, 6, 4, 3, 3, 8, 7, 6]

Median = (4 + 6)/2
= 5

Question 4 (1 point):
Measures of Spread is a type of data summary, and it helps you understand the spread of your data set. In the given multiple choices, you need to identify the methods which do not belong to Measures of Spread.

☐ A – Interquartile Range
☒ B – Weighted Mean
☐ C – Data Skewness
☐ D – Range
☒ E – Median

Question 5:
Consider a store that sells different categories of products

Skin care	Frozen Foods	Imported Cookies	Chocolates	Clothes	Electronics	Stationery

Now they have a dataset of all the sales (Number of sales of each product every day, with categories information) that they made **in one year**. If they want to get some insights from that data:

a) What are the metrics that can be generated from the dataset of products sold using different data aggregation methodologies? List 2 of them. (2 points)
b) And explain the significance of insights generated using the data aggregation methods in part(a)? (1 point)

a)

Metrics	Data Aggregation method	Description
Total sales	SUM	Aggregate of the quantity and price of all products sold.
Total sales by product category	SUM, GROUP BY	Sum of sales grouped by individual product categories.

b) By analyzing **Total Sales**, we can gain insights into our performance against sales targets.

We breakdown the **total sales into product category** that are currently or potentially driving the most revenue.

This enables us to make informed decision about allocating additional resources to capitalize those profitable product categories.

Question 6 (2 points):
There are several ways of Data Summary, identifying which of the following falls under the methods of Data Aggregation.

☒ A – Groupby
☐ B – Mode
☒ C – Unique Values
☒ D – Count
☐ E – Quartiles

Question 7 (2 points):
Data Skewness is the difference of some values from most other values in a dataset. It brings the asymmetry. As we know that there are two types of Data Skew; can you formulate an example of Positive and Negative Skew in Data?

Skewness	Example
Positive Skew	**Distribution of income**, with small proportion of individuals earning huge portion of the total income.
Negative Skew	**Retirement age**, where most people tend to retire at their 50s while significantly less people retire at 40s

</talentlabs>

Question 8 (2 points):

Let's take an example of a data set recorded at different branches of a bank where they recorded the time taken by an ATM for doing one transaction. The Maximum and minimum of the IQR of data set is as follows:

Maximum of IQR = 12 min

Minimum of IQR = 3 min

Now identify the Outlier ATM's involved in the experiment

Time taken by ATM (in minutes) = [7,5,3,9,6,22,10,11,4,2]

Mean = (2 + 3 + 4 + 5 + 6 + 9 + 10 + 11 + 22) / 9
= 8

Std deviation

$$= \sqrt{\frac{(2-8)^2+(3-8)^2+(4-8)^2+(5-8)^2+(6-8)^2+(9-8)^2+(10-8)^2+(11-8)^2+(22-8)^2}{9-1}}$$

x	z-score
2	-0.9798
3	-0.8165
4	-0.6532
5	-0.4899
6	-0.3266
9	0.163299
10	0.326599
11	0.489898
22	2.28619

Z-scores beyond certain critical values signal outliers:

- 90% Confidence Interval: z-score > 1.645
- 95% Confidence Interval: z-score > 1.96
- 99% Confidence Interval: z-score > 2.58

Ans: **22 is considered an outlier in both 90% and 95% confidence intervals.**