**Level 6 – Tasks**

**Task 1: Types of Statistics :**

**1. Descriptive Statistics:**

Descriptive statistics are used to summarize or describe the main features of a dataset. They help us understand the basic aspects of the data through numbers, such as averages, percentages, or standard deviations, without making predictions or inferences about a larger population.

The methods are,

* **Measures of Central Tendency:** Mean (average), Median (middle value), and Mode (most frequent value).
* **Measures of Dispersion:** Range, Variance, and Standard Deviation, which describe the spread or variability of the data.

**2. Inferential Statistics:**

Inferential statistics involve using data from a sample to make predictions or inferences about a larger population. These types of statistics are often used when it's impractical or impossible to collect data from the entire population, so we rely on sample data to draw conclusions.

The methods are,

* **Hypothesis Testing:** Used to determine if there is a significant difference or relationship between groups or variables.
* **Regression Analysis:** Used to predict the value of one variable based on the value of another.

Eg. The company surveys 200 customers and finds an average satisfaction rating of 4.2 out of 5, which is an example of descriptive statistics. Using this data, they predict that 85% of all customers will be satisfied, demonstrating inferential statistics by making predictions about a larger population based on the sample data.

**Task 2: Population vs. Sample :**

**Population**: All families in the city, as the researcher is ultimately interested in the average income of every family in the city.

**Sample**: The 100 households surveyed, which represent a subset of the entire population used to gather data and make inferences about the broader group.

**Task 3: Sampling Techniques :**

For this scenario, the best choice would be stratified sampling.

Because, it involves dividing the employees into distinct groups (or strata) based on department size. Since the company has departments of varying sizes, this technique ensures that each department is represented proportionally in the survey.

**Task 4: Types of Variables :**

* **Quantitative Variables** :
  + Number of pets is a quantitative variable since it represents the no. of pets and its numerical. It is a discrete variable because it gives whole value.
  + Weights in kg is also a quantitative because it represents the weights quantity in numerical. It is a continuous variable since it gives values within the range.
* **Qualitative variables :**
  + Blood Type is a qualitative variable because it represents the categories of data rather than its value.
  + Satisfaction level is also a qualitative variable because it also represents the categories.

**Task 5: Measures of Central Tendency :**

[3, 5, 5, 6, 7, 100]

Mean : 3 + 5 + 5 + 6 + 7 + 100 / 6 = 126 / 6 = 21

Median : The middle values are 5,6

(5 + 6) / 2 = 11/2 = 5.5

Mode : Frequent value = 5

Interpretation – The mean is larger because it contains an outlier.

**Task 6: Measures of Dispersion :**

[2, 4, 6, 8, 10]

Mean : 2 + 4 + 6 + 8 + 10 / 5 = 30 / 5 = 6

Variance : (2 - 6)2 = 16

(4 - 6)² = 4

(6 - 6)² = 0

(8 - 6)² = 4

(10 - 6)² = 16

16 + 4 + 0 + 4 + 16 = 40/5 = 8

Standard deviation : The standard deviation is the square root of the variance

= 2.828

Variance measures spread, and standard deviation shows typical deviation from the mean.

**Task 7: Z-Scores and Interpretation :**

For x = 60,

60 – 50 / 10 = 1

For x = 40,

40 – 50 / 10 = -1

Z-scores tell us how many standard deviations a data point is away from the mean. Positive z-scores indicate values above the mean, while negative z-scores indicate values below the mean.

**Task 8: Gaussian Distribution :**

Empirical rule -> μ ± 1σ

mean (μ) = 100, standard deviation (σ) = 15,

estimate the percentage of data between 85 and 115.

Lower range : 85 = 100 – 1(15) = μ - 1σ

It lies 1 sd below the mean

Higher range : 115 = 100 + 1(15) = μ + 1σ

It lies 1 sd above the mean

Approximately 68% of the data lies between 85 and 115

**Bonus Task: Box Plot and Outlier Detection :**

[1, 2, 3, 4, 5, 6, 7, 8, 9, 20]

calculate Q1, Q3, IQR, and identify outliers.

Q1 : 25/100\*10 = 2.5, Q1 = 3

Q3 : 75/100\*10 = 7.5, Q3 = 8

Inter Quartile Range : Q3 – Q1 = 7.5 – 2.5 = 5, IQR = 5

Lower fence : Q1 – 1.5\*IQR = 2.5 – 1.5\*5

2.5 – 7.5 = -5

Higher fnce : Q3 + 1.5\*IQR = 7.5 + 1.5\*5

= 7.5 + 1.5 = 15

Outliers : In the given dataset only 20 is