Section-1

K. Haripriya Sudheera
DA364536.
EN12025082~~77~~2

1A) For categorical column in univariate analysis, two graphs are commonly used:

1. Bar graph:- This is a simple and effective way to visualize the distribution of categorical data. The xaxis represents the categories, and ~~they~~ a y-axis represents the frequency or count of each category. Each bar represents a category and its ~~highest~~ height corresponds to the number of observation in that category. Bar graphs are easy to interpret and can be used to compare the frequencies of different categories.

2. Piechart ~~At~~ It is another popular choice. It represents the distribution of categories a circle, divided into Slices. Each slice represents a category, and its size corresponds to the proportion of observations in that category. Pie charts are useful for showing the relative frequencies of categories can be visually appealing. However, they can be difficult to interpret for large numbers of categories

2A) for numerical columns in Univariate analysis, two graphs are commonly used:

1. Histogram: This is a graphical representation of the distribution of a numerical variable. The x-axis represents the range of variable, and the y-axis represents the frequency or count of observations within each range. The histogram is constructed by dividing the range into intervals and counting the numbes of observations that

fall within each interval. This graph is useful for understanding the shape, center, and spread of the distribution.

Boxplot:- This is a graphical summary of the distribution of a numerical variable. It shows the median, quartiles and outliers of the data. The box represents the interquartile range (IQR), which contains 50% of the data. The median is shown as a line within the box. The whiskers extend from the box to the minimum and maximum values, excluding outliers. Outliers are points that lie outside of 1.5 times the IQR. Box plots are useful for comparing the distributions of multiple numerical variables and identifying outliers.

3A) **Mean, Median, Mode :**

There are three common measures of central tendency used in statistics to describe the middle or typical value of dataset.

Mean: This is the sum of all values divided by the total number of values. It's often referred to as the averag. It's suitable for normally distributed data without extoreme outliers. For example, calculating the average test score for a class of students.

Median:- This is the middle value in a dataset when the values are arranged in order. It's not affected by extreme outliers. It's suitable for skewed data or when there are extreme values. For example, finding the median income in a neighbourhood where there are a few very high earners.

Mode: This is the most frequent value in dataset. There can be one mode, multiple modes or no mode. Its suitable for categorical data or when you want to know the most common value. For example, determining the most popular color of car sold in a dealership.

Scenario:

Imagine you're analyzing the salaries of employees at a company.

- Mean: If the salaries are normally distributed without any significant outliers, the mean would be a good measure of central tendancy. It would give a representative average salary.
- Median: If there are a few very high earning executives the mean might be skewed upwards. In this case, the median would be a better measure as it wouldnt be affected by the extereme values.
- Mode: If there are a large number of employees earning the same salary (eg. minimum wage), the mode would be useful to identify the most common salary level.

4A) Box plots:-

A box plot is a graphical representation of the distribution of a numerical dataset. It provides a visual summary of the key characteristics, including the median, quartiles and potential outliers.

Key components:-

- Median: The vertical line within the box represents the median, which is the middle value of dataset. It seperates the lower half and upper half of the data.

- Quartiles! The box itself is divided into four equal parts by the median and two other lines. These lines represent the quartiles.

  -> first quartile (Q1): The bottom edge of the box represents Q1, which separates the lowest 25% of the data from the rest.

  -> Third Quartile (Q3): The top edge of the box represents Q3, which separates the highest 25% of the data from the rest.

  -> InterQuartile range (IQR): The distance between Q1 & Q3 is the IQR, which represents the middle 50% of the data.

- whiskers:- The lines extending from the box are called whiskers. They typically represent the minimum & maximum values excluding outliers

- Outliers:- Points that lie outside of the whiskers are considered potential outliers. They are often plotted individually as points. Outliers can indicate unusual or extreme values in the data.

Interpretation:-

- Shape: The shape of the box plot can reveal the distribution of the data. A symmetric box plot indicates a roughly normal distribution, while ~~a~~ skewed box plot suggests a skewed distribution.

- Central Tendency: The median represents the central value of the data.
- Spread: The IQR measures the spread the middle 50%. of the data. Longer the whiskers indicate a wider spread.
- Outliers: Outliers can be identified by their distance from the whiskers. They may require further investigation to understand their cause and impact on the analysis.

Ex:- If box plot for a dataset has a median near the center of the box, a relatively small IQR & no outliers. it suggests a fairly symmetrical distribution with a concentrated middle & limited extreme values.

6A) Constructing a Histogram: Step-by-step.

1. Determine the Range:- Find the minimum and maximum values of the numerical column. This will give you the overall range of your data.
2. Choose the Number of Bins: The number of bins determines the level of detail in your histogram. A smaller number of bins will create a coarser representation, while a larger number will create a finer representation. Generally, b/n 5 & 15 is good starting point.
3. Calculate bin width: Divide the range of your data by the desired number of bins to get bin width. This will be the size of each interval in your histogram.

4. Create Bins:- Starting from the minimum value, create bins of equal width using the calculated bin width.

5. Count observations: Count the number of observations that fall into each bin.

6. Plot the Histogram: Draw a bar for each bin, with the height of the bar representing the frequency of observations in that bin.

Impact of bin choice:-

The choice of bins significantly impacts the interpretation of a histogram:

- Too few bins: Using too few bins can hide important details in the data. It may create a very smooth histogram, making it difficult to identify patterns or shapes.

- Too many bins: Using too many bins can create a very jagged histogram, makes difficult.

- Appropriate no. of bins: The ideal no. of bins depend on the specific data set & desired level of detail.

7A) Correlational is a statistical measure that quantifies the strength and direction of the linear relationship between two numerical variables. It ranges from -1 to 1:

→ -1: Perfect negative correlation, indicating a strong inverse relationship. As one variable increases, the other decreases proportionally.

→ 0: No correlation, indicating no linear relationship b/n the variables.

→ 1: Perfect positive correlation, indicating a strong direct relationship. As one variable increases, the other increases proportionally.

The correlation coefficient is often calculated using the Pearson's correlation coefficient, which measures the covariance b/n the two variables divided by the product of their standard deviations.

Example:

consider a relationship between height and weight in a group of people. If we find a positive correlation between height & weight, it means that taller people tend to weigh more, and shorter people tend to weigh less. The strength of the correlation would indicate how closely this relationship holds. A correlation coefficient of 0.8, for example, would suggest a strong positive relationship, but not a perfect one.