1. Which of the following statements about pdf and cdf?

1) The cumulative distribution function is the probability that a random variable X, will take a value exactly equal to x.
2) The probability density function is the probability that a random variable X, will take a value equal to or less than the specified value.
3) The probability density function is the probability that a random variable X, will take a value exactly equal to x.
4) The cumulative distribution function is the probability that a random variable X, will take a value equal to or less than the specified value.

Ans: option (3) and option (4) are correct

2. What are bins in a histogram? How does changing the number of bins will modify the plot?

Ans:-A histogram displays numerical data by grouping data into "bins" of equal width.Bins are also sometimes called "intervals", "classes", or "buckets".

The number of bins affects the appearance of a graph. If there are few bins, the graph will be unrefined and will not represent the data well. If there are too many bins, many of the bins will be unoccupied and the graph may have too much detail.

3. How do you interpret Boxplot results? What do different boundaries of boxplot signify? What is the significance of whiskers in boxplot ?

Ans:Half the scores are greater than or equal to this value and half are less. The Middle box represents the middle 50% of scores for the group. The range of Scores from lower to upper quartile is referred to as the inter-quartile range. The The middle 50% of scores fall within the inter-quartile range.

Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

median (Q2/50th Percentile): the middle value of the dataset.

first quartile (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the dataset.

third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the "maximum") of the dataset.

interquartile range (IQR): 25th to the 75th percentile.

whiskers (shown in blue)

outliers (shown as green circles)

"maximum": Q3 + 1.5*IQR

minimum: Q1 -1.5*IQR

What defines an outlier, "minimum", or"maximum" may not be clear yet. The next section will try to clear that up for you.

A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the Data distribution through their quartiles.The lines extending parallel from the boxes are known as the "whiskers", which are used to indicate variability outside The upper and lower quartiles.

4. What are the different ways in which we can handle NaN values for
i) Categorical features
ii) Numerical features
Ans: **For categorical features:-**

i)Method 1: Filling with most occurring class
ii)Method 2: Filling with unknown class
iii)Method 3: Using Categorical Imputer of sklearn-pandas library

Numerical Features:-
i) A Simple Option: Drop Columns with Missing Values
ii) A Better Option: Imputation
iii) An Extension To Imputation