**EDA assignment**

Ans1.

Option3- The probability density function is the probability that a random variable X, will take a value exactly equal to x

Option4- The cumulative distribution function is the probability that a random variable X, will take a value equal to or less than the specified value.

Ans2.

For plotting a histogram we show the frequency of data by grouping it into "bins" of equal width. Each bin is plotted as a bar whose height depicts how many data lies in that bin

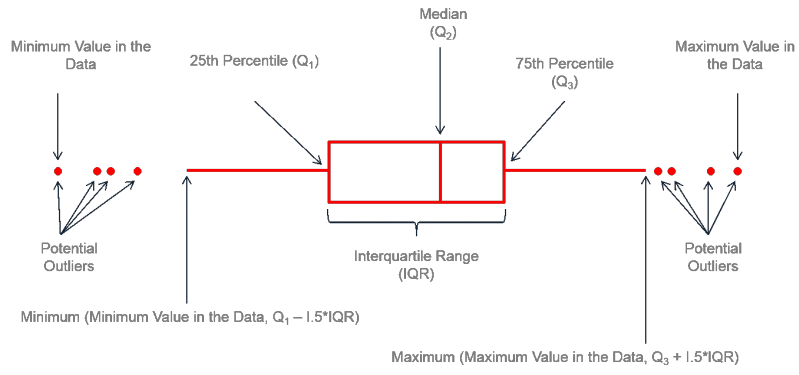
Eg: finding number of students as per age groups

Here we can create groups of interval 0 to 5, 6 to 10, 11 to 15 ….they are called bins

If we change the number of bins, the plot will get modified based on the changes we make.

Taking same eg above, if we make groups of interval 0 to 20, 21 to 40 and so on…there we might not get proper insight from the data as it would contain too much data. Similarly if we create bins of small groups , there is a possibility that the graph can get cluttered and even some groups might not have any data.

Ans3.



1. The rectangular box of the boxplot is: the interquartile range.

2. The left boundary of the box, is first quartile ie 25th Percentile

3. The right boundary is third quartile i.e 75th Percentile.

4. The median is represented by a vertical bar in the center of the box.

5. The leftmost point of the chart is the minimum which is the smallest number and the r rightmost is the maximum which is the largest number in the set.

The whiskers – they can be any value and basically represent the minimum or maximum value

Ans4.

Handling NaN values:

1. Categorical feature:
2. Deleting rows with missing values
3. Replace with most frequent values from the category or with NA
4. Create classifier algorithm to predict the value eg KNN
5. Apply unsupervised ML
6. Numerical feature:
7. Fill missing value with Mean/Median
8. Deleting rows with missing values
9. Using imputer of sklearn-pandas library

References – Geeks for Geeks, youtube