Experiment No. 9

Aim: Data Visualization II

Problem Statement:

1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age')
2. Write observations on the inference from the above statistics.

Theory:

Numerical Data

Analyzing Numerical data is important because understanding the distribution of variables helps to further process the data. Most of the time you will find much inconsistency with numerical data so do explore numerical variables.

1) Histogram

A histogram is a value distribution plot of numerical columns. It basically creates bins in various ranges of values and plots them where we can visualize how values are distributed.

2) Distplot

Distplot is also known as the second Histogram because it is a slightly improved version of the Histogram. Distplot gives us a KDE(Kernel Density Estimation) over histogram which explains PDF(Probability Density Function) which means what is the probability of each value occurring in this column.

3) Boxplot

Boxplot is a very interesting plot that basically plots a 5 number summary. to get a 5 -number summary of some terms we need to describe.

- Median – Middle value in series after sorting
- Percentile – Gives any number which is number of values present before this percentile like for example 50 under 25th percentile so it explains total of 50 values are there below 25th percentile
- Minimum and Maximum – These are not minimum and maximum values, rather they describe the lower and upper boundary of standard deviation which is calculated using Interquartile range(IQR).

IQR = Q3 - Q1
Lower_boundary = Q1 - 1.5 * IQR
Upper_bounday = Q3 + 1.5 * IQR

Here Q1 and Q3 are 1st quantile(25th percentile) and 3rd Quantile(75th percentile).

Conclusion: Hence we have observed that inference from given statistics.