Experiment No. 10

Aim: Data Visualization III

Problem Statement:
Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as:
1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a boxplot for each feature in the dataset.
4. Compare distributions and identify outliers.

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) 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 studied & given inference for list of dataset features, feature distribution and identified outliers.