

Assignment No. 10

Page No.	
Date	

Name - Kuldharan Sumit Dattatraya

Class - TE

Division - Div 4

Subject - DSBDA

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 & give the inference as:

1. List down the features & 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 box plot for each feature in the dataset.
4. Compare distributions & identify outliers.

Theory -

- 1) How to deal with outliers of various dependent & independent variables in regression analysis?
- Sometimes outliers are bad data & should be excluded such as types.

Different methods to detect outliers

Univariate - boxplot outside 1.5 inter-quartile range

Bivariate - with the help of scatterplot

Multivariate - Mahalanobis D² distance

ways to finding outliers -

- sorting method
- data visualization method
- statistical tests (Z-score)
- Interquartile range method.

Pros of removing outlier -

- the number is clearly on unintentional error
- the number is an intentional error
- the outlier comes from a different population

Cons of removing outlier -

- the outlier is a legitimate observation from your desired population
- I chopped an outlier & looked at my data again & now a new outlier popped up.
- making a definitive rule can be hard to defend.

Why are outliers to be treated carefully-

outliers increases the variability in your data which decreases statistical power.

What cause outliers-

It arise due to changes in system behaviour Fraudulent behaviour, human error, instrument error or simply through natural deviation in populations.

How do you determine the distribution fits my data best-

probability plots might be the best way to determine whether your data follow a particular distribution. If your data follow the straight line on the graph.

can be distribution be normal if it has outliers?

Yes the normal distribution data can have outliers.

Types of Distribution-

- Direct Distribution
- Indirect Distribution
- Intensive Distribution
- Exclusive Distribution
- selective Distribution

Why Distribution is important:

Distribution are important for statistics because we need to collect the sample & estimate the parameters of the population distribution. Hence distribution is necessary to make inferences about the overall population.

Conclusion -

Hence, we learned about data visualization & Outlier.