Assignment DA-1

Title: Data visualisation and analysis

Problem statement
Download the iris flower dataset or any other dataset into dataframe. Use python to perform the following
1) How many features are there and their types?
2) Compute and display summary statistic for each feature in dataset
3) Data visualisation - Create a histogram for each feature in dataset to illustrate feature distribution
   Plot histogram. Create boxplot for feature in dataset. All boxplot should be combined in single plot. Compare distributions and identify outliers.

Objectives
To understand basics of data visualisation and description techniques
To gain hands on experience of using libraries such as numpy, pandas, matplotlib and seaborn

Outcomes
To be able to implement and execute script which allows data visualisation thereby enabling us to gain better understanding of dataset

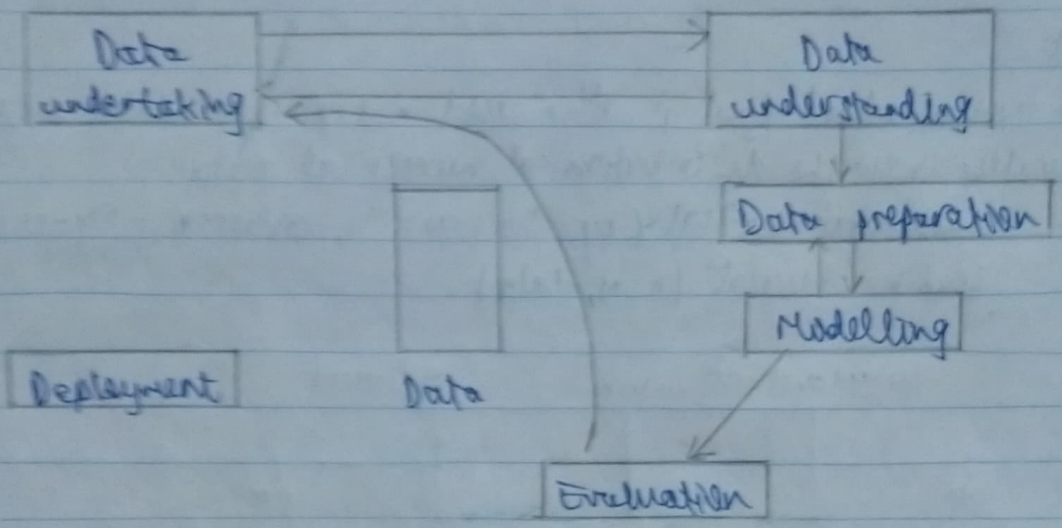
Software requirements
Jupyter notebook, Python3, associated libraries and packages

Theory related concepts

Data Analysis
It is a process of inspecting, cleansing, transforming and modeling data with aims to discover useful information and patterns in data
The purpose is to extract useful information

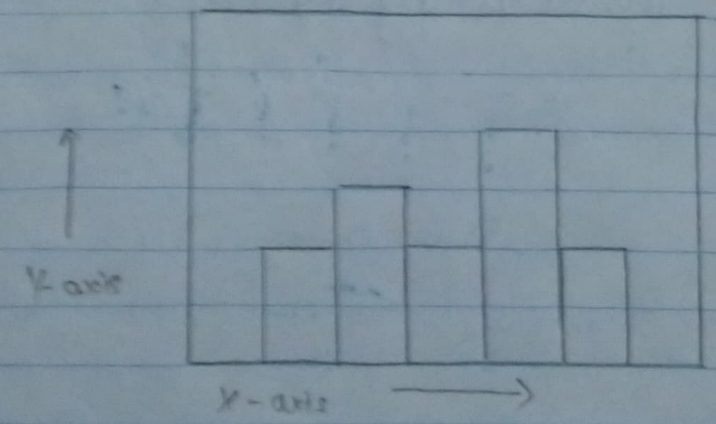
Types of data analysis

1. Text analysis / data mining
   It is a method to discover a pattern in large data using databases or data mining tools.
2. Statistical analysis
   Includes collection, analysis of data
   i) Inferential analysis - analyzes samples from complete data
   ii) Descriptive analysis - analyzes complete data or sample of summarised data
3. Predictive analysis
4. Descriptive analysis
5. Diagnostic analysis



Data visualization
Graphical representation of information and data. By virtual elements like charts, graphs, and maps provides way

Histogram
Histogram is an approximate representation of distribution of normal data
- Height of bars in histogram show observed value
- Uses vertical column to display data
- Pandas library histogram draw function groups values a full range a bin and draw all bins in axes

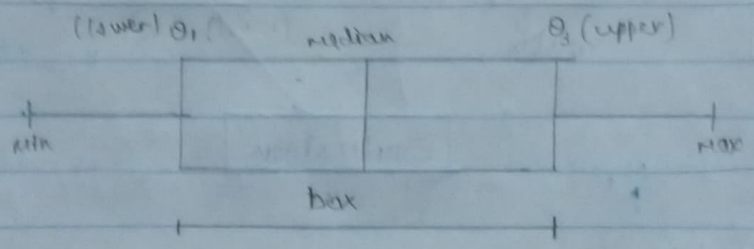


Boxplot

It is a type of chart that is often used in exploring data analysis to visually show the distribution of numerical data

Syntax: dataframe.boxplot(by = "x-axis", column = "y-axis")

seaborn.boxplot(x, y, data)



Algorithm

1. Import libraries pandas, matplotlib, seaborn
2. Load dataset into dataframe
3. Use info function to get number of feature and datatypes
4. Use describe function to get statistical info of dataset
5. Use dataframe hist() to plot histogram for each feature

Conclusion

Thus the task of feature description and data visualisation on flower dataset was successfully implemented