Assignment - A1 41403

- Title: Summary statistics, data visualization & boxplot for the features on the Iris dataset or any other dataset

- Objective: Learn to use dataset, dataframes, features in an application
  - Learn to compute summary statistics for the features
  - Learn to use visualization technique

- problem statement: Download the Iris dataset or any into a dataframe. Use python & perform following
  - How many features are there & what are their types.
  - compute & display summary statistics for each feature available in the dataset.
  - Data visualization. Create a histogram for each feature in the dataset to illustrate the feature distribution.
  - Create a boxplot for each feature in the dataset. All of the boxplots should be combined.
- outcome: we will be able to compute statistics on the features of the dataset., use histograms & box plot.

41403

S/w H/w : Ubuntu, Python 3

Aim

Theory: Data analysis is a process of inspecting cleaning transforming & modelling data with the goal of discovering useful information, informing conclusions & supporting decision making. Data analysis has multiple facets & approaches encompassing diverse techniques under a variety of names while being used in different business, science & social science domains. A data set is a collection of data. Most commonly a data set corresponds to the contents of a single database table.

- Important terms:
mean, SD, regression, simple size determination & hypothesis testing are the fundamental data analytics methods

Mean: The sum of all the data entries divided by the no. of entries.
Range: The diff. b/w max & min in data.
SD: The standard deviation measures variability & consistency of sample or population In most real world problem is advantage

41403

- Mean:

$$\text{mean} = \frac{\text{sum}}{N}$$

- Variance:

$$VAR = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

The formula for SD is basically just square root of variance

$$SD = \sqrt{VAR}$$

- Algorithm

x : iris dataset

How many feature are there & their types

x.dtypes.

Compute & display summary

x.describe

create histogram

plt.hist(x['feature'], bins=15)

plt.show().

create combined boxplot

x.boxplot()

- Conclusion: we have successfully conducted the data visualization of iris dataset