|  |  |  |  |
| --- | --- | --- | --- |
| **TITLE** | | Data Visualization III | |
| **PROBLEM STATEMENT/ DEFINITION** | | 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.   1. Create a histogram for each feature in the dataset to illustrate the feature distributions. 2. Create a box plot for each feature in the dataset. 3. Compare distributions and identify outliers. | |
| **OBJECTIVE** | | To implement the data visualization techniques | |
| **S/W PACKAGES AND HARDWARE APPARATUS USED** | | 1. Operating System : 64-bit Open source Linux or its derivative  2. Programming Languages: PYTHON/R | |
| **REFERENCES** | | * Mark Gardner, “Beginning R: The Statistical Programming Language”, Wrox Publication, ISBN: 978-1-118-16430-3 * David Dietrich, Barry Hiller, “Data Science and Big Data Analytics”, EMC education services, Wiley publications, 2012, ISBN0-07-120413-X * Luis Torgo, “Data Mining with R, Learning with Case Studies”, CRC Press, Talay and Francis Group, ISBN9781482234893 | |
| **STEPS** | | **Refer to student activity flow chart if found necessary by subject teacher and relevant to the subject manual.**  **Describe steps only.** | |
| **INSTRUCTIONS FOR WRITING JOURNAL** | | 1. Title 2. Problem statement 3. Learning objective 4. Learning outcome 5. Theory (includes methods, libraries and functions, 6. Analysis (as per assignment), 7. conclusion. | |

P:F:-LTL-UG / 03 / R1

**Assignment No. 10**

* **Aim:**

**Summary statistics, data visualization, histogram and boxplot for the features on the Iris dataset or any other dataset.**

* **Problem Statement / Definition:**
  + 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:
    - List down the features and their types (e.g., numeric, nominal) available in the dataset.
    - Create a histogram for each feature in the dataset to illustrate the feature distributions.
    - Create a box plot for each feature in the dataset.
* **Prerequisites**
  + Database management system, Python/R programming
* **Learning Objectives**
  + Learn to use dataset, dataframes, features of dataset in an application
  + Learn to compute summary statistics for the features.
  + Learn to use visualization techniques.
* **Learning Outcome:**
  + Students will be able to compute statistics on the features of the dataset, use histograms and boxplot on the features of the dataset.
* **Theory:**

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

**Iris flower dataset:**

The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms.

Attribute Information:

-> sepal length in cm

-> sepal width in cm

-> petal length in cm

-> petal width in cm

-> class:

Iris Setosa

Iris Versicolour

Iris Virginica

Number of Instances: 150

**Summary statistic:**

Mean, standard deviation, regression, sample size determination and hypothesis testing are the fundamental data analytics methods.

Mean: The sum of all the data entries divided by the number of entries.

* 1. ****



* 1. Range: The difference between the maximum and minimum data entries in the set.

Range = (Max. data entry) – (Min. data entry)

Standard deviation:

The standard deviation measure variability and consistency of the sample or population. In most real-world applications, consistency is a great advantage. In statistical data analysis, less variation is often better.

Logo

Description automatically generated with medium confidence



Variance: The average squared deviation from the mean is also known as the variance.

Percentile: Let p be any integer between 0 and 100. The pth percentile of data set is the data value at which p percent of the value in the data set are less than or equal to this value.

• How to calculate percentiles: Use the following steps for calculating percentiles for small data sets.

• Step 1: Sort the data in ascending order (from smallest to largest)



• Step Step 3: 2: Calculate ith = the 100 where p is the percentile and n is the sample size.

Step 3: If i is an integer the pth percentile is the mean of the data values in position i and i+1.If i is not an integer then round up to the next integer and use the value in this position.

Summary statistic on Iris dataset:

Summary Statistics:

Min Max Mean SD Class Correlation

sepal length: 4.3 7.9 5.84 0.83 0.7826

sepal width: 2.0 4.4 3.05 0.43 -0.4194

petal length: 1.0 6.9 3.76 1.76 0.9490 (high!)

petal width: 0.1 2.5 1.20 0.76 0.9565 (high!)

Class Distribution: 33.3% for each of 3 classes.

**Box Plot:**

A boxplot shows the distribution of the data with more detailed information. It shows the outliers more clearly, maximum, minimum, quartile(Q1), third quartile(Q3), interquartile range(IQR), and median. You can calculate the middle 50% from the IQR.

Chart, box and whisker chart

Description automatically generated

**Histogram:**

Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying [probability distribution](https://citoolkit.com/articles/probability-distributions/) of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

A histogram is a value distribution plot of numerical columns. It basically creates bins in various ranges in values and plots it where we can visualize how values are distributed. We can have a look where more values lie like in positive, negative, or at the center(mean)

Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although [histograms](https://citoolkit.com/articles/histogram/) are better in determining the underlying distribution of the data, [box plots](https://citoolkit.com/articles/box-plot/) allow you to compare multiple data sets better than histograms as they are less detailed and take up less space. It is recommended that you plot your data [graphically](https://citoolkit.com/articles/graphical-analysis/) before proceeding with further statistical analysis.

Chart, histogram

Description automatically generated

Histogram for Sepal Length

Chart, histogram

Description automatically generated

Histogram for Petal Length