|  |  |
| --- | --- |
| **TITLE** | Data Visualization II |
| **PROBLEM STATEMENT/ DEFINITION** | 1. Use the inbuilt dataset 'titanic' as used in the assignment#7. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names: 'sex' and 'age')  2. Write observations on the inference from the above statistics. |
| **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. |

**Assignment No. 9**

* **Aim:**

**Summary statistics, data visualization, boxplot for the features on the ‘titanic’ dataset or any other dataset.**

* **Problem Statement / Definition:**

Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names: 'sex' and 'age')

* + Write observations on the inference from the above statistics.
* **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.

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.

Boxplot is a very interesting plot that basically plots a 5 number summary. to get 5 number summary 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).

**Titanic dataset:**

It is one of the most popular datasets used for understanding machine learning basics. It contains information of all the passengers aboard the RMS Titanic, which unfortunately was shipwrecked. This dataset can be used to predict whether a given passenger survived or not. The csv file can be downloaded from Kaggle.

Description:

Text, letter

Description automatically generated

Following is a box plot for distribution of age with respect to each gender along with the information about whether they survived or not.

import numpy as np

import pandas pd

import matplotlib.pyplot as plt

import seaborn as sns

from seaborn import load\_dataset

#titanic dataset

data = pd.read\_csv("titanic\_train.csv")

sns.boxplot(data['Sex'], data["Age"], data["Survived"])

plt.show()

