|  |  |
| --- | --- |
| **TITLE** | Data Visualization I |
| **PROBLEM STATEMENT/ DEFINITION** | 1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see if we can find any patterns in the data.  2. Write a code to check how the price of the ticket (column  name: 'fare') for each passenger is distributed by plotting a histogram. |
| **OBJECTIVE** | to make it easier to identify patterns, trends and outliers in large data sets. |
| **S/W PACKAGES AND HARDWARE APPARATUS USED** | **Operating System recommended** :- 64-bit Open source Linux or its derivative  **Programming tools recommended**: - JAVA/Python/R/Scala |
| **REFERENCES** | 1)Chirag Shah, “A Hands-On Introduction To Data Science”, Cambridge University Press,(2020), ISBN : ISBN 978-1-108-47244-9. 2.  2)Wes McKinney, “Python for Data Analysis”, O' Reilly media, ISBN : 978-1-449-31979-3. |
| **STEPS** | **Refer to student activity flow chart if found necessary by subject teacher and relevant to the subjectmanual.**  **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. |

**Theory:**

**Data Visualization**

Data Visualization is the presentation of data in pictorial format. It is extremely important for Data Analysis, primarily because of the fantastic ecosystem of data-centric Python packages. And it helps to understand the data, however, complex it is, the significance of data by summarizing and presenting a huge amount of data in a simple and easy-to-understand format and helps communicate information clearly and effectively

**Pandas and Seaborn** is one of those packages and makes importing and analyzing data much easier. In this article, we will use Pandas and Seaborn to analyze data.

## ****Pandas****

[**Pandas**](https://www.geeksforgeeks.org/python-pandas-dataframe/) offer tools for cleaning and process your data. It is the most popular Python library that is used for data analysis. In pandas, a data table is called a dataframe.

## ****Seaborn****

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It is built on the top of [matplotlib](https://www.geeksforgeeks.org/python-introduction-matplotlib/) library and also closely integrated into the data structures from [pandas](https://www.geeksforgeeks.org/introduction-to-pandas-in-python/).

**Installation**

For python environment :

pip install seaborn

## Seaborn: statistical data visualization

Seaborn helps to visualize the statistical relationships, To understand how variables in a dataset are related to one another and how that relationship is dependent on other variables, we perform statistical analysis. This Statistical analysis helps to visualize the trends and identify various patterns in the dataset.

These are the plot will help to visualize:

* Line Plot
* Scatter Plot
* Box plot
* Point plot
* Count plot
* Violin plot
* Swarm plot
* Bar plot
* KDE Plot