**Data Science And Big Data Practical Assignment NO.8**

**Title: Data Visualization 1**

**Problem Statement:**

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**: Students should be able to perform the data visualization using Python Libraries on -any open-source dataset.

**Pre-requisite:**

1. Basic of Python programming

2. Concept of Data Analysis Fundamentals, Data visualization technique, Data visualization principles

3. Basics of Python libraries Matplotlib and Seaborn.

**Theory:**

Data visualization is the graphical representation of information and data. By using visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. It involves leveraging various python libraries and tools specifically designed for creating visual representations of data. Some of the most commonly used libraries for data visualization in Python include:

* Seaborn

Seaborn is built on top of Matplotlib and provides a higher-level interface for creating attractive and informative statistical graphics. It simplifies the process of creating complex visualizations such as categorical plots, relational plots, and distribution plots.

* Matplotlib

Matplotlib is a powerful and widely used plotting library in Python. It provides a wide variety of plots and customization options, allowing users to create basic plots like line plots, scatter plots, histograms, bar charts, and more.

* Pandas

Pandas, a popular data manipulation library in Python, also provides basic plotting functionality through its integration with Matplotib. Users can quickly create simple plots directly from Pandas Data-Frames.

**Know your Data**

Dataset used is Titanic. csv' dataset.

Import dataset in your jupyter notebook

The dataset contains 891 rows and 15 columns and contains information about the passengers

who boarded the unfortunate Titanic ship. The original task is to predict whether or not the

passenger survived depending upon different features such as their age, ticket, cabin they

We will use the Seaborn library to see if we can find any

boarded, the class of the ticket, etc.

patterns in the data.

**Data Preprocessing**

Data should not contain any null or incorrect data. It should be correct and accurate analysis

so that the graphs made can be correctly interpreted.

**Finding patterns of data**

Patterns of data can be find out with the help of different types of plots.**Histogram**

A histogram is a graphical representation of the distribution of numerical data. It

consists of a series of adjacent rectangles, or bins, where the width of each bin

represents a range of values, and the height represents the frequency of occurrences

within that range. Histograms are particularly useful for visualizing the frequency

distribution of continuous data.

**Key Concepts of Histogram:**

* Frequency Distribution: Histograms display the frequency distribution of data

by depicting how often values fall within certain ranges (bins). The height of

each bar represents the frequency or count of values falling within that bin.

* Bins: Bins are intervals or ranges into which the data is divided. The number

of bins and their width influence the granularity of the histogram and can

affect the interpretation of the data.

* Skewness and Symmetry: The shape of the histogram can reveal important

characteristics of the data, such as skewness (asymmetry) and symmetry. A

symmetrical distribution appears approximately mirror-like on either side of

the centre, while a skewed distribution is asymmetrical.

* Central Tendency and Dispersion: It provide insights into the central tendency

(mean, median, mode) and dispersion (variance, standard deviation) of the

cluster and how widely they vary.

data. The position and spread of the histogram indicate where most values

* Outliers: Outliers, or extreme values, can be identified visually as data points

that lie far away from the bulk of the data in the histogram. Analysing outliers

can help detect anomalies or errors in the dataset.

Interpretation of Histogram

* Shape of the Distribution: Assess the overall shape of the histogram to understand the

underlying distribution of the data. Common shapes include symmetric (bell-shaped),

skewed (left-skewed or right-skewed), uniform, or multimodal (having multiple

peaks).The shape provides insights into the central tendency, variability, and presence

of any anomalies in the data.

* Central Tendency: Identify measures of central tendency such as the mean, median.

and mode. Central tendency measures indicate where the bulk of the data is centre

within the distribution.

* Variability: Evaluate the spread or variability of the data around the central tendency.

Variability is assessed by observing the width of the distribution and the dispersion of

data points within the histogram.

* Skewness and Kurtosis: Determine whether the distribution is symmetric or skewed

Skewness indicates asymmetry in the distribution, with the tail of the distribution

extending more towards one side. Kurtosis measures the peaked ness or flatness of the distribution.

* Outliers and Anomalies: Identify any outliers or extreme values that deviate

significantly from the main distribution. Outliers may indicate errors in data

collection, unique observations, or interesting phenomena in the dataset.

within the data that provide insights into its underlying structure.

**When to Use Histograms:**

* Frequency Distribution Analysis: It are ideal for visualizing the distribution of

numerical data, including understanding the shape, central tendency, and variability of

the data

* Comparison of Distributions: Use to compare the distributions of different datasets or

subsets of data.

* Detecting Patterns and Anomalies: Analyse the shape of the histogram to identify

patterns, trends, or anomalies in the data.

* Data Preprocessing: It can aid in data preprocessing tasks such as identifying data

skewness, choosing appropriate data transformations, and detecting outliers.

* Detecting Outliers: Histograms can highlight outliers - observations that fall far

outside the typical range of values in the dataset. Outliers are often visible as bars that

extend beyond the typical range of values or as isolated bars with unusually high

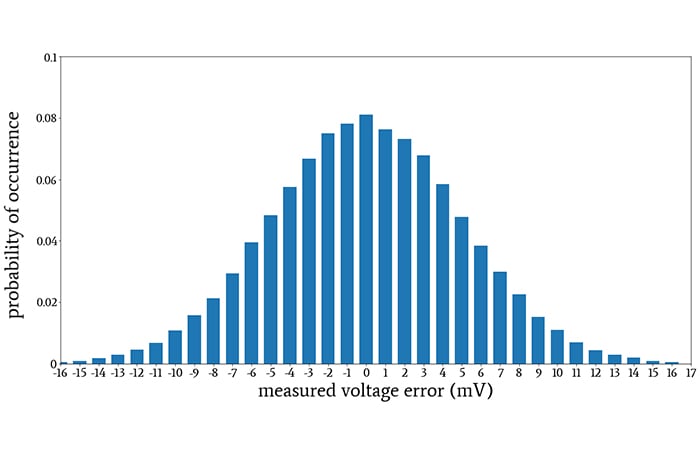
frequencies.

* Exploratory Data Analysis: Histograms are commonly used in exploratory data

analysis (EDA) to gain insights into the structure and characteristics of the data. By

visually inspecting the histogram, analysts can identify patterns, trends, and anomalies

that may warrant further investigation.



**Countplot:**

A countplot is a visualization tool used to represent the frequency distribution of categorical variables. It counts the number of occurrences of each category and displays them as bars.

* When to Use Countplot?

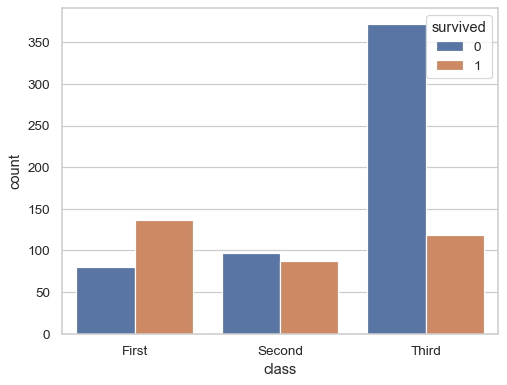
1. Analyzing the distribution of categorical variables – It helps understand how many times each category appears in the dataset.
2. Comparing the frequency of different categories – Useful for identifying dominant or underrepresented categories.
3. Detecting imbalances in data – Helps assess whether a dataset is skewed toward certain categories.
4. Understanding relationships – Can be used with hue (color-coded grouping) to compare distributions across multiple categories.

* **Interpretation of Countplot**

1. Height of the bars represents the count – Taller bars indicate higher frequency.
2. Evenly distributed bars – Suggests that all categories have similar occurrences.
3. Unevenly distributed bars – Indicates that some categories are more dominant than others.
4. Multiple colors (hue parameter) – Allows comparison within subgroups to see if a specific category has a higher occurrence in a particular group

* **Used for**

1. Best suited for categorical variables.
2. Helps in visual comparison of different categories.
3. Useful for identifying patterns and imbalances in the dataset.
4. Can be enhanced with hue, order, and different orientations for better insights.

**Conclusion:**

We have successfully this assignment on Data Visualization and have gained insights and knowledge about finding patterns using various graphs using the titanic dataset.