PICT, PUNE

Aarkin Kanvik
31140

DSBDAL

Assignment - 8

- Problem statement:

i) Use inbuilt dataset 'titanic'. The dataset contains 891 rows & 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.

ii) Write a code to check how the price of the ticket for each passenger is distributed by plotting a histogram.

- Learning Objectives:

To learn & understand data visualization in python using seaborn library.

- Software Used:

i) Python 3.8.1
ii) Jupyter Notebook
iii) Operating System: Ubuntu

- Hardware Used:

i) Name: Dell Optiplex - 3020
ii) CPU: Intel i5 - 4590
iii) RAM: 8 GB DDR3 @ 1600 MHz

PICT, PUNE

31140

Theory:

Data Visualization is one of the steps of data science process, which states that after data has been collected, processed & modelled, it must be visualised for conclusions to be made. It provides a quick & effective way to communicate information.

Seaborn is a visualization python library for statistical graphics plotting. It provides beautiful default styles & color palettes to make statistical plot more attractive.

Seaborn aims to make visualisation the central part of exploring & understanding data. It provides dataset oriented APIs, so that we can switch between different visual styles for same variables for better understanding the dataset.

Six main types of plots in Seaborn library:

i) Relational plots
ii) Categorical plots
iii) Distribution plots
iv) Regression plots
v) Matrix plots
vi) Multi-plot grids

PICT, PUNE

31140

- Histogram:
They are visualization tools that represent the distribution of a set of a continuous data.

In a histogram, the data is divided into bins (usually on the x-axis) & the count of data points that fall into each bin corresponding to the height of the bar above that bin.

- Analysis:
The dataset has a shape of (891, 12). The rows with null values in 'Age', 'Cabin' & 'Embarked' columns were found. For column 'Age' the null values were substituted with the mean. For 'Embarked' the null values were filled with the mode of the column & 'cabin' column was dropped. Thus leaving us with new shape (712, 11). Also, 'survived' column was typecasted to bool.
Used the pandas inbuilt function to find correlations between the columns & then used Seaborn's Heatmap to show the correlations.

Seaborn's histplot plotted the histogram to find out the distribution of passengers over fare which indicated that as the fare increases the number of passengers paying that fare decreases.

PICT, PUNE

31140

Conclusion:

Thus, we successfully visualised the titanic dataset with the help of Seaborn library's Heatmap & Histograms.