Name: Omkar Gaikwad
Batch: L1
Roll no: 31126

PICT, PUNE

DSBDAL
Assignment-9

Title: Data Visualization - II

Problem Statement:

1. Use the inbuilt dataset 'titanic' as used in above problem. Plot a boxplot for distribution of age with respect to each gender along with the information about whether they survived or not (column name: 'sex' and 'age')
2. Write observations on the inference from the above statistics.

Learning outcomes:

- To understand various visualization techniques using seaborn python library.
- To apply appropriate plotting technique to visualize how the survivability of passengers depended on their 'sex' and 'age'
- Describe the observations made by using each plot/ graph.

Learning Outcomes:

Student will be able to:

- perform basic data visualization to appropriate graphs
- make observations on the 'survived' of passenger and how it varied wrt 'sex' and 'age'

PICT, PUNE

Software and hardware Requirements:

Linux OS (ubuntu), Intel i5 - 8th gen (8GB RAM)
python 3.8, Jupyter Notebook.

Theory:

Data visualization -

It is representation of data through use of common graphics, such as charts, plots, info graphics and even animations.

Seaborn library:

It is data visualization library built on top of matplotlib and closely integrated with pandas data structure in python. Visualization the central part of seaborn which helps in exploration and understanding of the data.

Advantage of using seaborn:

- Better aethetics
- Nicer Built in plots
- easy customizability
- statistically minded plots.

Categorical data type:

It represent the types of data which may be divided into groups. Examples are sex, educationlevel, class etc.

Plots used in the Analysis:

1] Barplot: It is basically used to aggregate the categorical data according to some methods (like mean). We choose a categorical column on x-axis and numeric column on y-axis

2] Boxplot: It is used to detect outliers in a in a group of numerical data through quartiles. Boxplot summarizes a sample data using 25th, 50th, 75th percentiles

3] swarmplot: A swarmplot is a type of scatterplot that is used for representing categorical variables. It is similar to stripplot, but it avoids the overlapping of points

4] violin plot: It is used to visualize the distribution of numerical data of different variables The advantage of a violinplot is that it can show nuances in the distribution That aren't perceptible in a boxplot.

5) Stripplot: It is used to draw a scatter plot based on category. We choose a categorical column on x-axis and numeric on y-axis.

PICT, PUNE

Observations:

- The dataset has 891 rows and 15 columns, 2 of which are numerical (fare and age)
- The mean age of passengers is 29.4 [using barplot]
- The mean age of passengers who survived was 25-30 (male) and (28-30) female
- The boxplot shows that about 50% of the passengers were between 20-40 age.
- The dominance of blue hue in ~~term~~ males in the swarmplot indicates that more than 50% males did not survive the disaster.
- On the other hand dominance of orange hue in females in swarmplot indicates that more than 50% females survived in disaster.

Conclusion:

Thus, we have successfully applied various visualization techniques and inferred the survival probability of passengers based on 'sex' and 'age' in the titanic dataset.