EXPERIMENT 5

AIM:

To perform data exploration and visualization for given dataset

THEORY:

Data mining comprises the discovery of interesting patterns and knowledge from large sets of data. This may include finding new data, confirming theory and models, etc. Data mining is not simply mining for data but mining knowledge from data, and thus it is also known as knowledge discovery

The steps in data mining include:

- Acquiring a dataset (including creating one)
- Cleaning and preprocessing
- Reduction to necessary components
- Basic exploration and visualization
- Choosing the task and algorithm
- Using this algorithm to get necessary output

Cleaning and preprocessing includes:

- Removing noisy data
- Removing outliers depending on their undue effext
- Handling missing data
    - Represent it as NULL (not considered for arithmetic, logical comparisons
    - Represent it as a default value (0 for numeric, “” for string)
    - Represent it as a central tendency
    - Represent it as a central tendency based on the other attributes class
- Maintain time sequence, changes etc.

Data exploration is a process that involves performing basic analysis on a dataset in order to understand it better. It could be used to uncover some patterns, and also understand

characteristics of the dataset, such as type of distribution, tendencies, variations etc. It helps with the cleaning process as werll as the reduction process

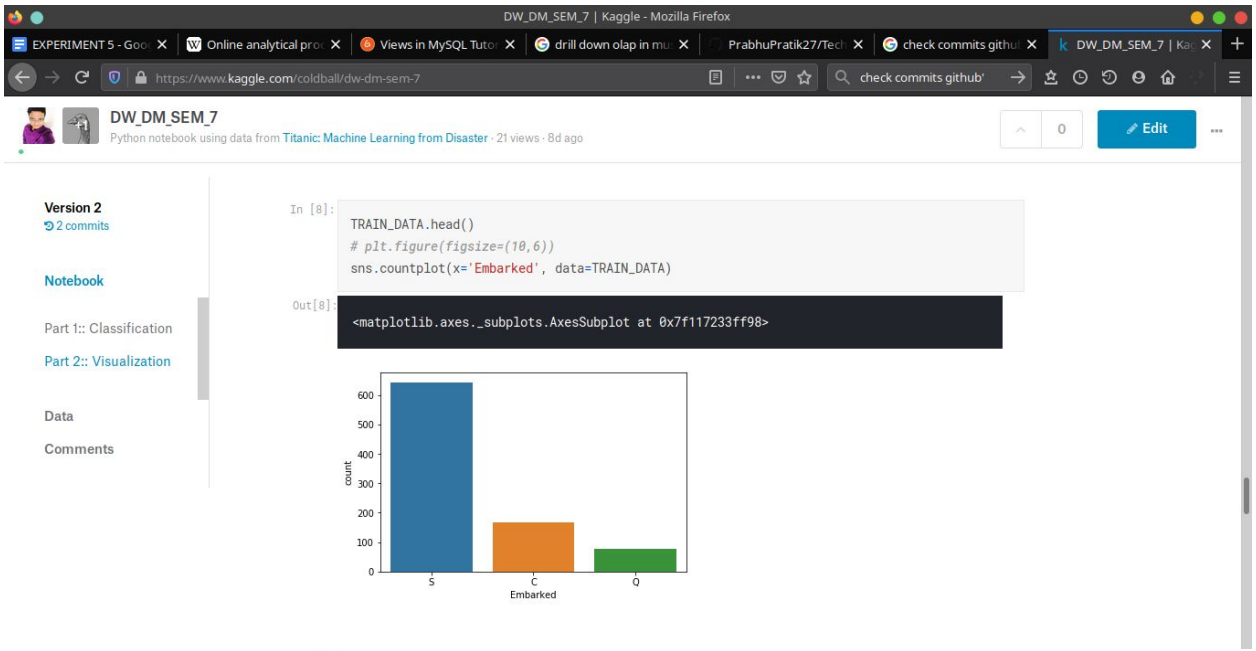
Data visualization can use mathematical methods (sucha s finding statisticla values) or graphical methods (data visualization)

Data visualization is the process where the data is represented in a grapihcal way to realizae some patterns, outliers, correlations, etc. in it. Some useful visualizations are:

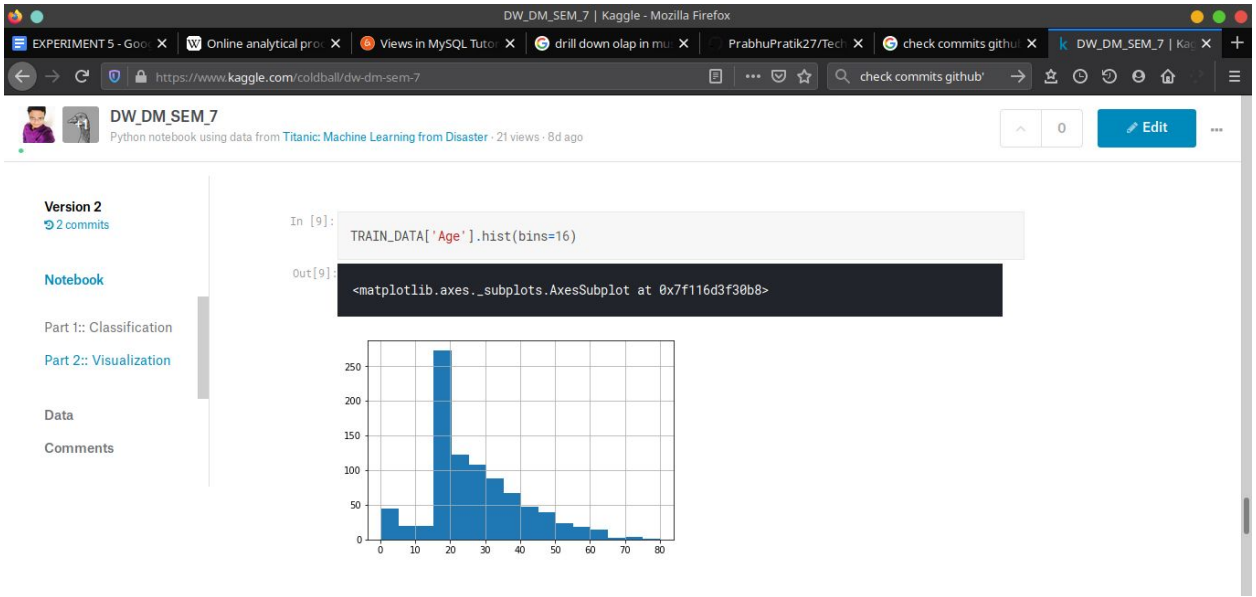
- Scateer plot:
    - Two variables are plotted, one on x axis and one on y, to find corellations between them
    - Additional variables can be plotted with 3d models, multiple plots, hues, size of point etc.
    - Useufl to understand trends between data, as well as corellation
    - Limited to only few variables at a time
- Bar charts:
    - A bar chart is a simple comparison chart for class data
    - Data of various classes may be aggregated on a given measure to get a bar chart
    - This can provide easy visual comparison of the classes
    - Multiple classes can be used with different colors, different dimensions et
    - Multiple measures can also be used similarly
    - Bar charts are good for discrete class data. Histograms are functionally equivalent for continuous data. Here, discrete classes are made by choosing ranges.
- Heatmaps:
    - These are comparisons of corellations between pairs of variables
    - A square is drawn with rows and columns for attributes
    - The intersection of a row and column is a cell
    - The cells gradient represents its corellation

IMAGES:

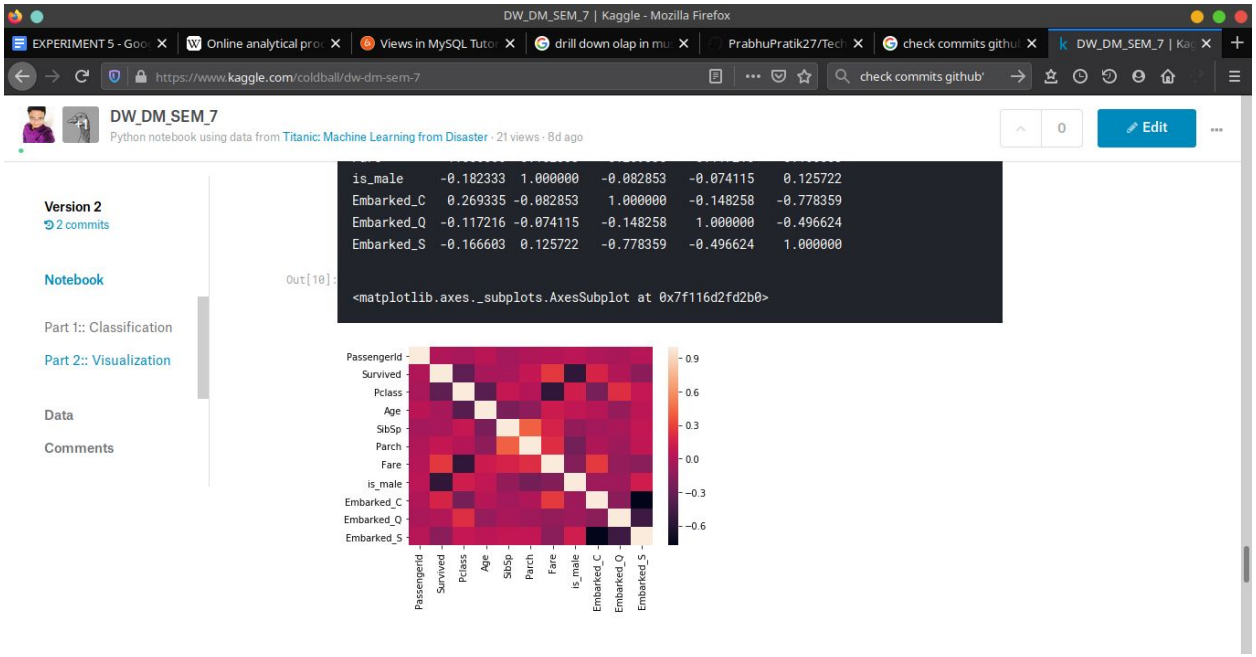
BAR CHART:



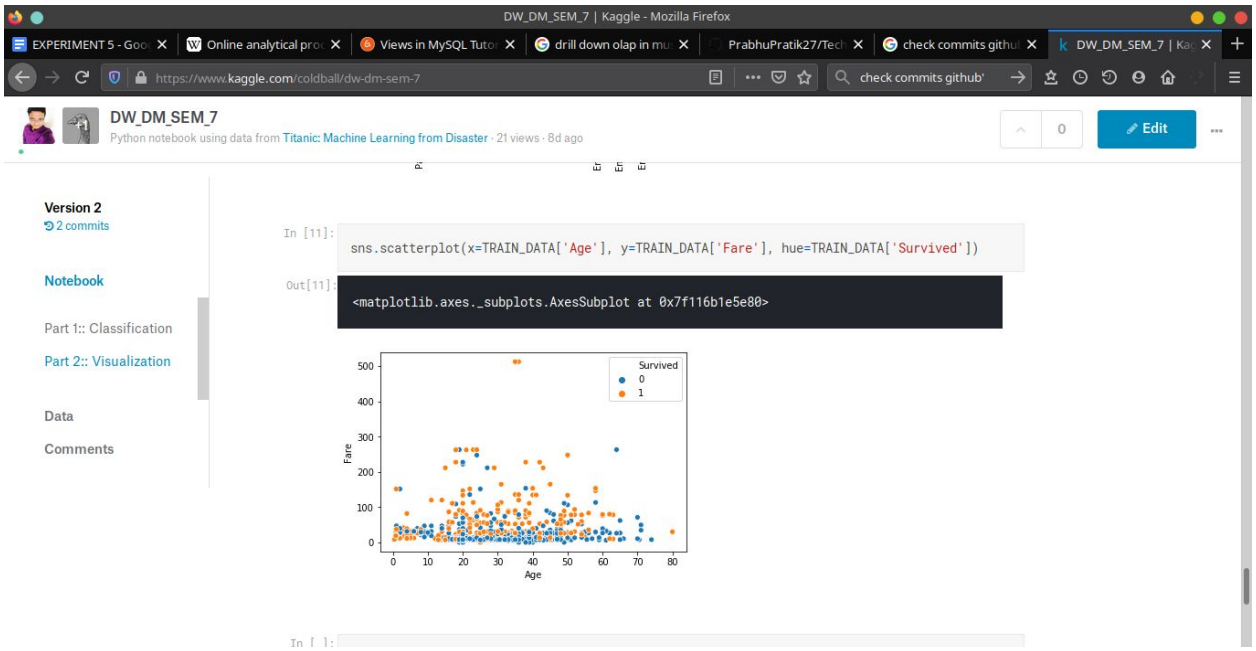
HISTOGRAM:



HEATMAP:



SCATTER:



CONCLUSION:

Thus, various visualizations have been created