CPSC 4800

ASSIGNMENT 3

DATA EXPLORATION

Sukhmani Kaur Sandhu

Student ID: 100376033

I used the info() and head() function to get an overview of the dataset. It shows that the dataset contains 891 entries and 12 columns. It also showed that there are some missing values for cabin, age and embarked.

I performed exploratory data analysis on the titanic data set. The three hypotheses I considered for it were:

- Association of survival rate and class:
  I plotted a bar graph which showed that first class passengers had more survival rate than second class passengers. Third class passengers survived the least. This may be due to preference given to first class passengers when evacuating.
  Also, first class passengers had more chances of survival than dying as compared to second and third class passengers who had more chances of dying

- Association of survival rate and gender:
  I plotted a bar graph which showed that women have more chances of survival than men. Out of 342 people who survived, 233 were female while 109 were male. Out of the people who did not survive women were less than hundred while the count of men was more than 450.

- Association of survival rate and age:
  To find the relation between survival and age, I plotted a histogram. It could be derived that out of survivors, people between age 20-40 has the highest survival rate, while those above 40 has low chance of survival. Surprisingly, the number of people who died is also highest among age group 20-40 while others has low rate of death. This may be because the number of people in this age group was much larger than other.

Link to repo: https://github.com/sandhusukhmani/Assignment3.git