Andrew Floyd

CS3001 - Intro to Data Science

October 7th, 2018

Hw #2

Clink on this link for my github repo: https://github.com/afzm4/cs3001hw2

Task #1

a. Discrete, Ordinal
b. Continuous, Ratio
c. Discrete, Nominal
d. Discrete, Nominal
e. Continuous, Ratio
f. Discrete, Ordinal
g. Continuous, Interval (or Ratio if in K)

Task #2

1. You would use correlation for this because we are looking at a linear relationship
2. I would once again use correlation for this one due to the linear relationship of the x and y along with fact that a ratio would work well for this comparison.

Task #3

Subtask 1:

- Q1: Features include: Survival, Pclass, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked (Name?)
- Q2: Categorical features include: Survival, Sex, Pclass (ordinal) and Embarked
- Q3: Numerical features include: Age, Fare, SibSp and Parch
- Q4: Mixed data types include: Ticket and Cabin
- Q5: It appear that Cabin, Age and Embarked features have some null values while Cabin and Age also have some incomplete values

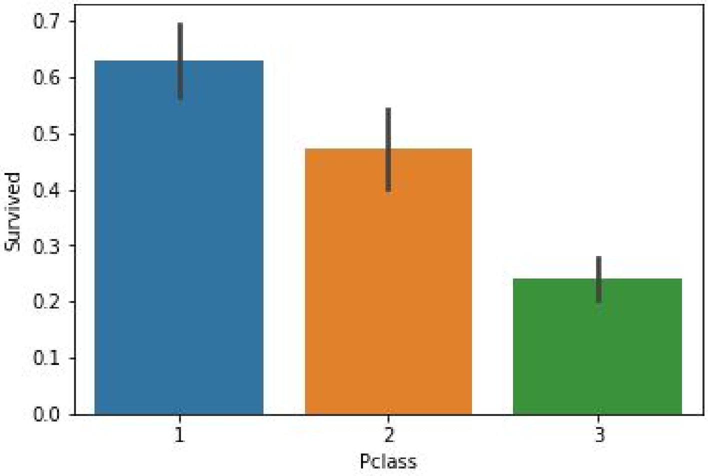
- Q6: It seems that we have some numerical values (ints and floats), while the rest of the features appear to be strings/objects
- Q7:
    - Age:
        - Count: 1046
        - Mean: 29.881138
        - STD: 14.413493
        - MIN: 0.17
        - 25% Percentile: 21.00
        - 75% Percentile: 39.00
        - MAX: 80.00
    - Fare:
        - Count: 1038
        - Mean: 33.295479
        - STD: 51.758668
        - MIN: 0.00
        - 25% Percentile: 7.895800
        - 75% Percentile: 31.275000
        - MAX: 512.329200
    - SibSp:
        - Count: 1309
        - Mean: 0.498854
        - STD: 1.041658
        - MIN: 0.00
        - 25% Percentile: 0.00
        - 75% Percentile: 1.00
        - MAX: 8.00
    - Parch:
        - Count: 1309

- Mean: 0.385027
- STD: 0.865560
- MIN: 0.00
- 25% Percentile: 0.00
- 75% Percentile: 0.00
- MAX: 9.00

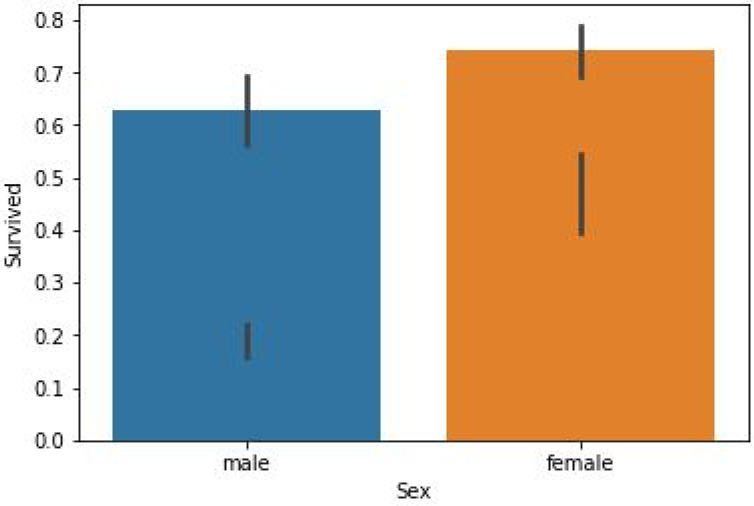
- Q8:
  - Sex:
    - Count: 1309
    - Unique: 2
    - Top: male (0.0)
    - FREQ: 843
  - Survived:
    - Count: 891
    - Unique: 2
    - Top: dead (0.0)
    - FREQ: 549
  - Pclass:
    - Count: 1309
    - Unique: 3
    - Top: 3
    - FREQ: 709
  - Embarked:
    - Count: 1307
    - Unique: 3
    - Top: S
    - FREQ: 914

Subtask #2:

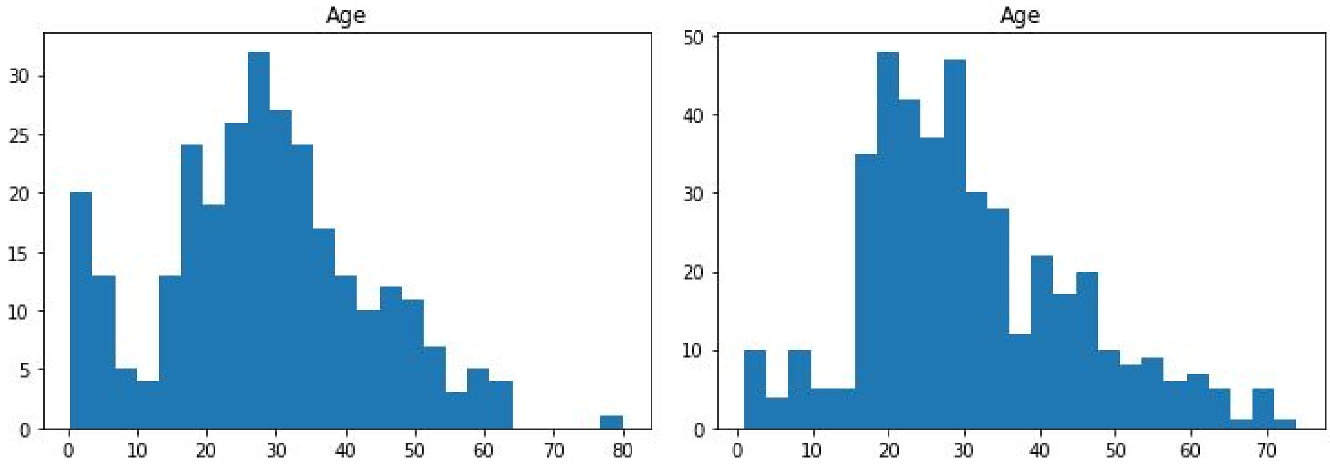
- Q9: As seen below (graph from code), Pclass=1 does have a high correlation with Survived, so we should include it.



- Q10: As seen below (another graph from code), female's do have a higher chance of surviving than males



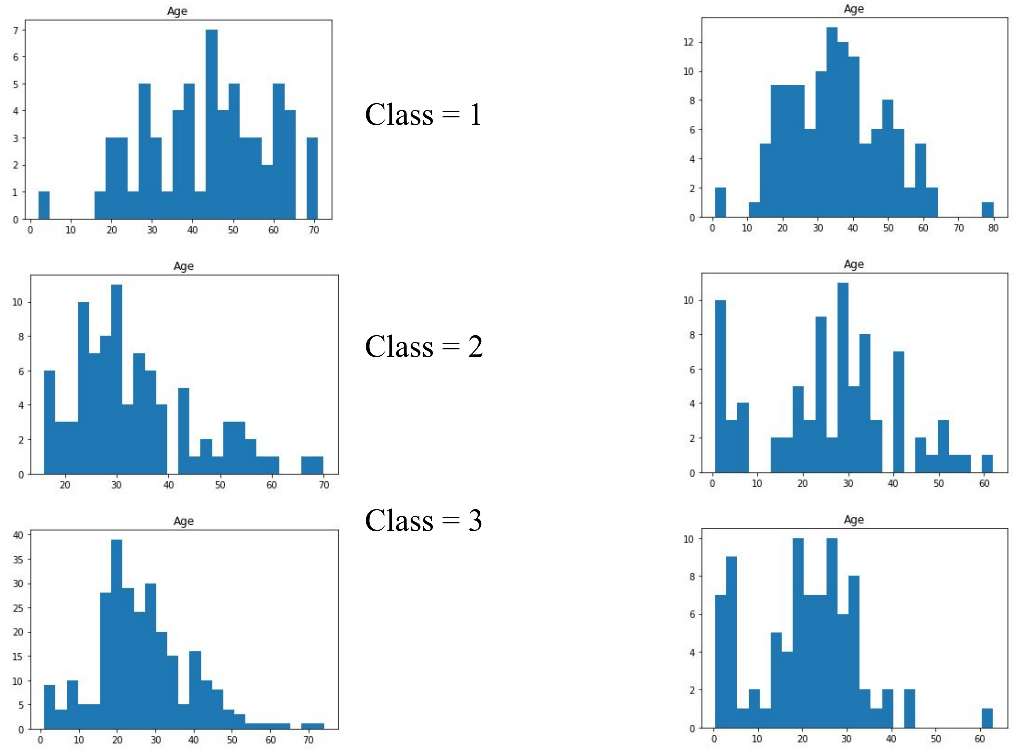
- Q11: (left one is those who survived, right is those who died)



  - It appears that infants had close to double the amount live than die, so the survival rate would be fairly high
  - All passengers age = 80 survived
  - 144 People aged 15-25 died while 79 lived, so much more died than survived
  - Yes, we should include Age. We complete the age value later on in Q17
  - Yes, we should consider banding Age
- Q12:

Survived = 0

Survived = 1

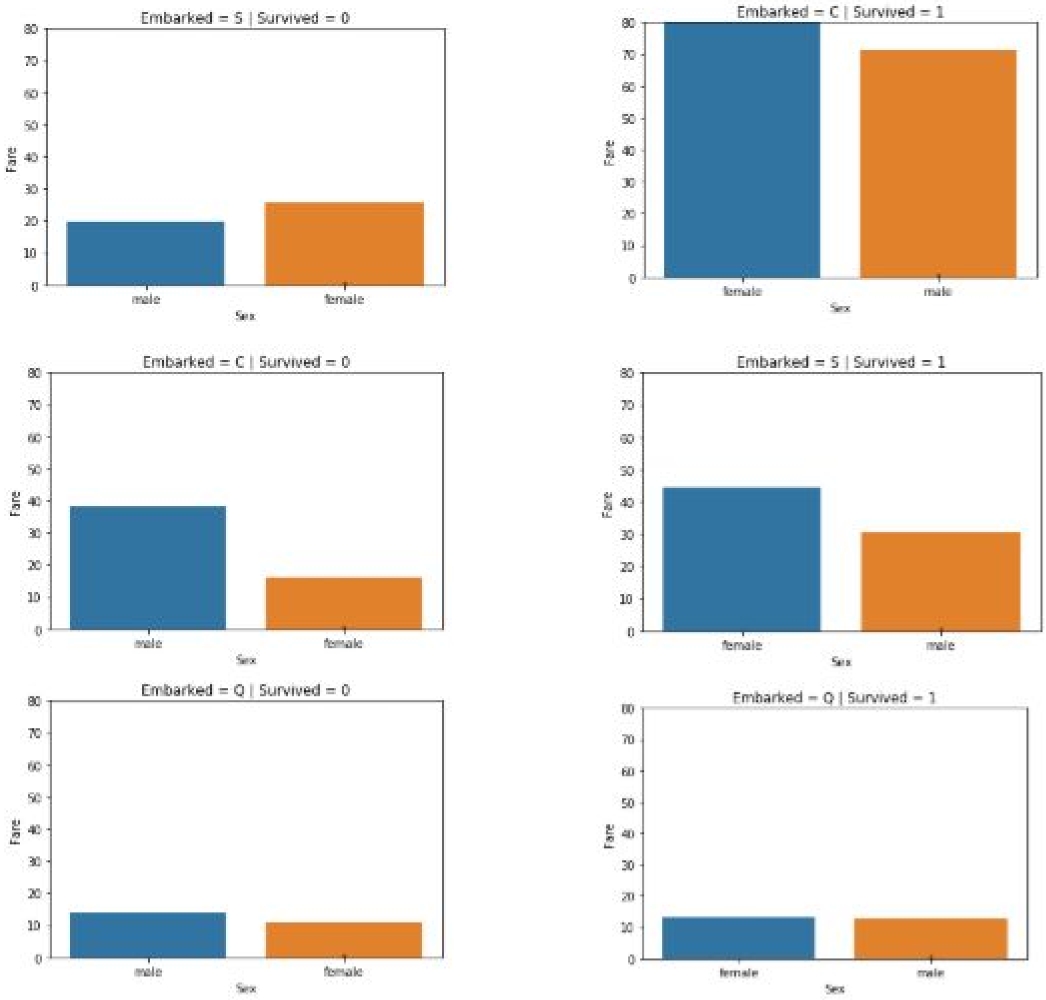


Class = 1

Class = 2

Class = 3

  - It appears that Pclass=3 has the most passengers, but the majority of them did not survive
  - All of infants in class 2 survive, while about half survive in class 3
  - The majority of passengers in Pclass=1 do survive
  - Yes, it appears that Pclass=1 is mostly older passengers, while Pclass=2 is more balanced while Pclass=3 has much more younger passengers.
  - Yes, we should include Pclass in our model
- Q13:



    - It seems that in general, the higher paying the passenger, the better the survival rate (but at port Q, this doesn’t seem to be the case)
    - It does appear that there is a significant relationship between embarking locations and survival rate
    - Yes, we should consider banding fare
- Q14:
    - Ticket duplicate rate: 0.2903
    - There doesn’t seem to be a correlation between Ticket and Survived
    - It looks like we should drop the Ticket feature
- Q15:
    - The Cabin feature is not complete
    - Cabin has 1014 NULL values in the combined dataset
    - Since an overwhelming number of values are NULL, we should drop the Cabin feature
- For Q16-Q20, all answers are done in the code