William Frazee
CAP5610
Code repository: https://github.com/Spyrix/CAP5610_HW1 – In titanic.py
I have commented out the lines in my code that plot my charts, but those should be obvious.

Q1: In training set, which features are available?

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Q2: In training set, which features are categorical?
Survived, Sex, Embarked, cabin, pclass
Q3: In training set, which features are numerical (e.g., discrete, continuous, or time series based)?
age, sibsp, Parch, fare, pclass, survived
Q4: In training set, which features are mixed data types?
Ticket because ticket is alphanumerical and numerical at the same time.
Q5: In training set, which features contain blank, null or empty values? In test set, which features contain blank, null or empty values?
Training:
PassengerId: 0
Survived: 0
Pclass: 0
Name: 0
Sex: 0
Age: 177
SibSp: 0
Parch: 0
Ticket: 0
Fare: 0
Cabin: 687
Embarked: 2

Testing:
PassengerId: 0
Pclass: 0
Name: 0

Sex: 0
Age: 86
SibSp: 0
Parch: 0
Ticket: 0
Fare: 1
Cabin: 327
Embarked: 0

Q6: In training set, what are the data types (e.g., integer, floats or strings) for various features?

PassengerId: int
Pclass: int
Name: string
Sex: string
Age: float (Some of the ages have .5 attached to indicate an estimation)
SibSp: int
Parch: int
Ticket: string
Fare: float
Cabin: string
Embarked: character

Q7: To understand the distribution of numerical feature values across the samples, please list the properties, including count, mean, std, min, 25% percentile, 50% percentile, 75% percentile, max, of numerical features?

```
count    714.000000
mean      29.699118
std       14.526497
min        0.420000
25%       20.125000
50%       28.000000
75%       38.000000
max       80.000000
Name: Age, dtype: float64

count    891.000000
mean       0.523008
std        1.102743
min        0.000000
25%        0.000000
50%        0.000000
75%        1.000000
max        8.000000
Name: SibSp, dtype: float64

count    891.000000
mean       0.381594
std        0.806057
min        0.000000
25%        0.000000
50%        0.000000
```

```
75%        0.000000
max        6.000000
Name: Parch, dtype: float64

count    891.000000
mean      32.204208
std      49.693429
min       0.000000
25%        7.910400
50%       14.454200
75%       31.000000
max      512.329200
Name: Fare, dtype: float64

count    891.000000
mean       2.308642
std        0.836071
min        1.000000
25%        2.000000
50%        3.000000
75%        3.000000
max        3.000000
Name: Pclass, dtype: float64

count    891.000000
mean     0.383838
std      0.486592
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      1.000000
Name: Survived, dtype: float64
```

Q8: To understand the distribution of categorical features, we define: count is the total number of categorical values per column; unique is the total number of unique categorical values per column; top is the most frequent categorical value; freq is the total number of the most frequent categorical value. Please list the properties, including count, unique, top, freq, of categorical features?

```
count      891
unique       2
top       male
freq       577
Name: Sex

count     889
unique      3
top         S
freq      644
Name: Embarked
```

```
count     204
unique     147
top        G6
freq        4
Name: Cabin
```

Q9: Can you observe significant correlation (average survivied ratio>0.5) among the group of Pclass=1 and Survived? If Pclass has significant correlation with Survivied, we should include this feature in the predictive model. Based on your computation, will you include this feature in the predictive model?

```
Pclass  Survived      Passengers
1        0              80
         1              136
2        0              97
         1              87
3        0              372
         1              119
```

Pclass 1 passngers had a survival rate of 63%.
Pclass 2 passengers had a survival rate of 47.3%

Pclass 3 passengers had a survival rate of 24.2%

Therefore, there does seem to be a significant correlation between Pclass and Survived, with survival trending downward with pclass value. This makes sense because upper class passengers on the titanic were likely to be closer to the deck and thus able to more easily reach limited lifeboats than lower class passengers. Therefore, we should definitely include this feature in the model.

Q10: Are Women (Sex=female) were more likely to have survived?

```
Sex       Survived   PassengerId
female  0             81
          1             233
male
          0             468
          1             109
```

Women had a survival rate of 74.2%
Men had a survival rate of 18.9%.
So, women were more likely to have survived. This makes sense because the captain had ordered to give lifeboat preference to women and children.

Q11: Let us start by understanding correlations between a numeric feature (Age) and our predictive goal (Survived). A histogram chart is useful for

analyzing continuous numerical variables like Age where banding or ranges will help identify useful patterns. The histogram can indicate distribution of samples using automatically defined bins or equally ranged bands. This helps us answer questions relating to specific bands (e.g., infants, old). Please plot the histogram plots between ages and Survived (Figure 1 is an example), and answer the following questions:

• Do infants (Age <=4) have high survival rate?

More infants survived than did not survive. This makes sense given the aforementioned preference to women and children.

• Do oldest passengers (Age = 80) survive?

These passengers appear to have an equal rate of surviving and not surviving.

• Do large number of 15-25 year olds not survive?

This group has a comparatively equal rate of surviving and not surviving.

• Based on your analysis of the histograms:Should we consider Age in our model training? (If yes, then we should complete the Age feature for null values.)

I believe so. More infants survived than did not survive, and as age increases (40-50+) the number of non-survivors also increases.

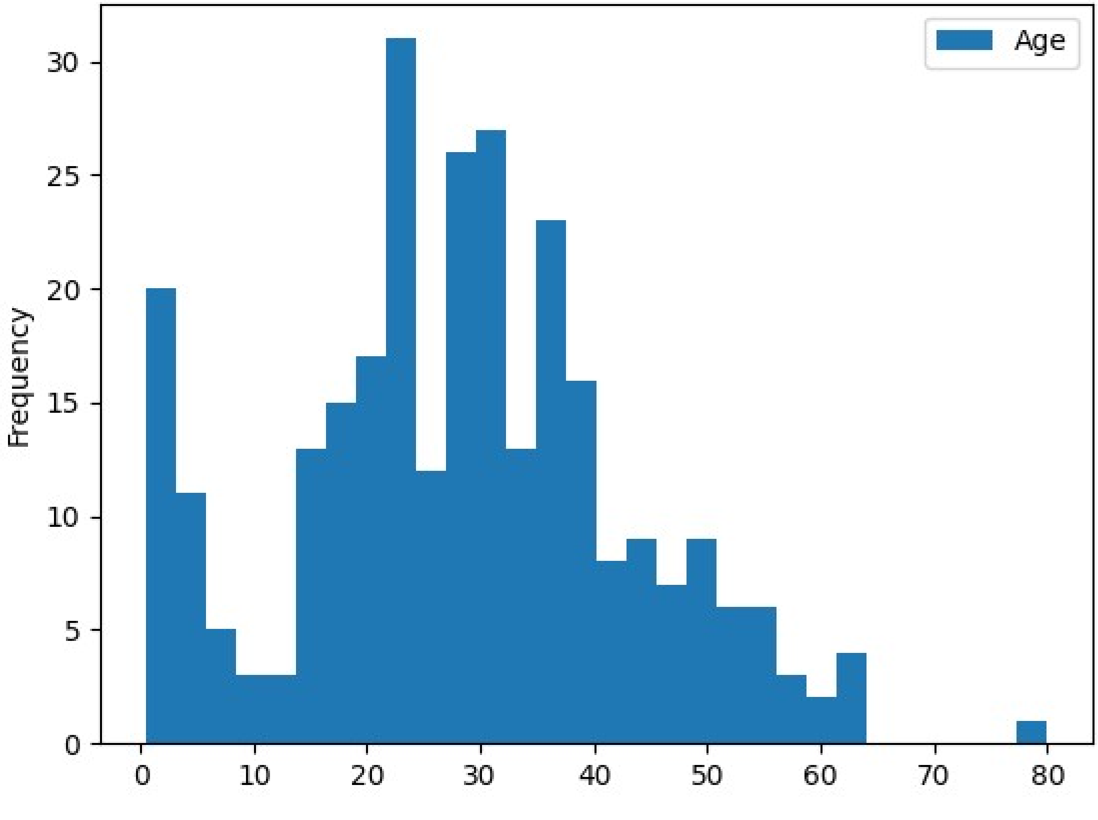
• Should we should band age groups?

Yes. There are pretty clear divides in both histograms. I would say that the bands ought to be:

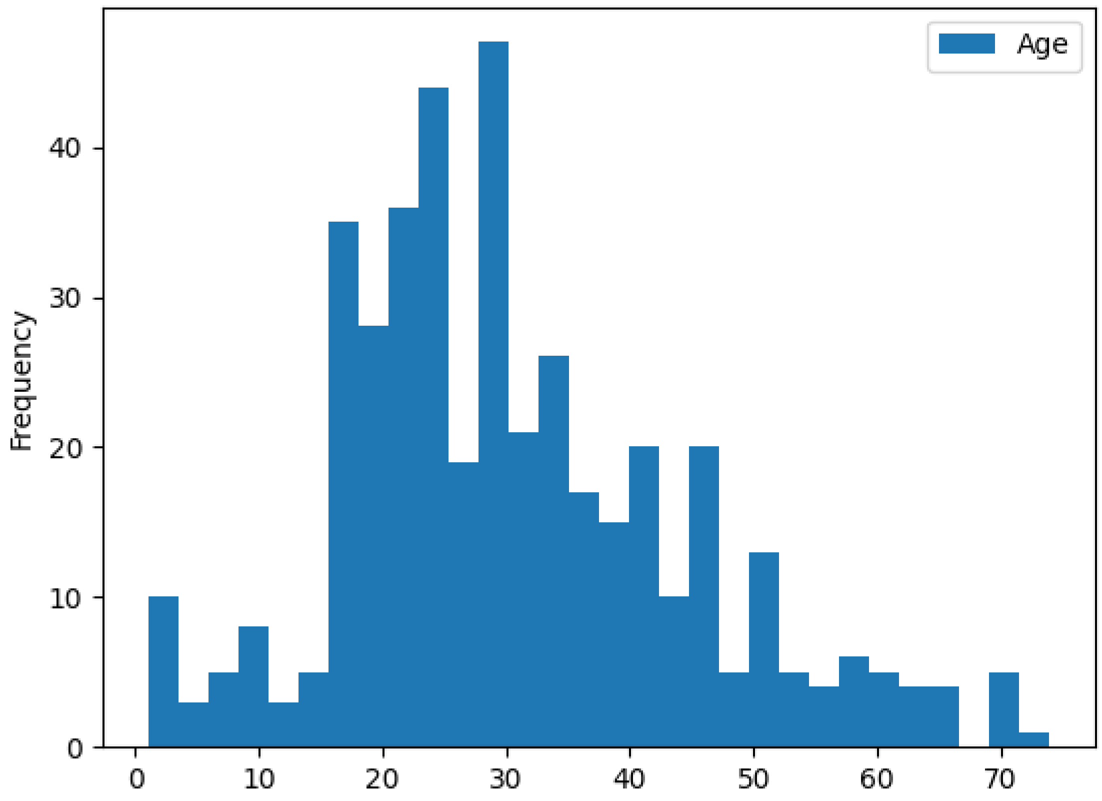
0-4
4-15
15-30
30-40
40-60
60-80+

Based on the visual distinction of areas in the histogram.

Histogram of age vs Survived = 1



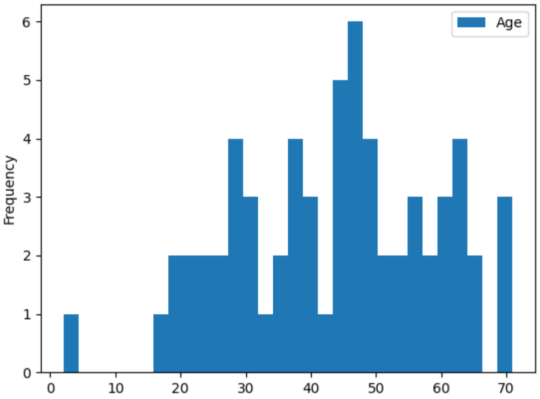
Histogram of Age vs Survived = 0

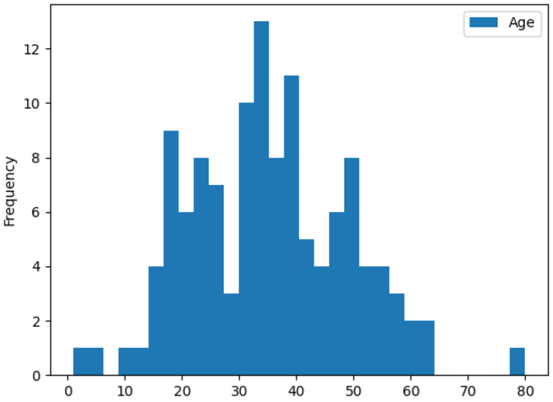


Q12: We can combine three features (age, Pclass, and survivied) for identifying correlations using a single plot. This can be done with numerical and categorical features which have numeric values. Please plot the histogram plot using python, and answer the following questions:

- **Does Pclass=3 have most passengers, however most did not survive?**
  - Pclass 3 had the most passengers, and most in fact did not survive.
- **Do infant passengers in Pclass=2 and Pclass=3 mostly survive?**
  - Yes
- **Do most passengers in Pclass=1 survive?**
  - Yes. The scale on the histogram of those who survived in pclass1 is considerably higher than the scale on the histogram of the non-survivors in pclass1.
- **Does Pclass vary in terms of Age distribution of passengers?**
  - Not considerably. The majority of people are between 20-50 years old.
- **Should we consider Pclass for model training?**
  - Yes. It appears that Pclass does correlate with survival, in such that the rate of survival goes: 1>2>3.

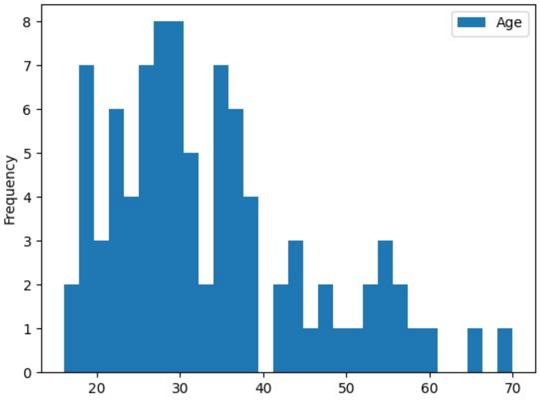
Pclass = 1
Survived = 0 Survived = 1

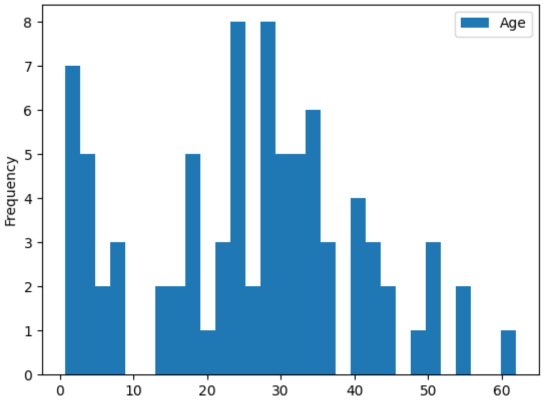




Pclass = 2
Survived = 0

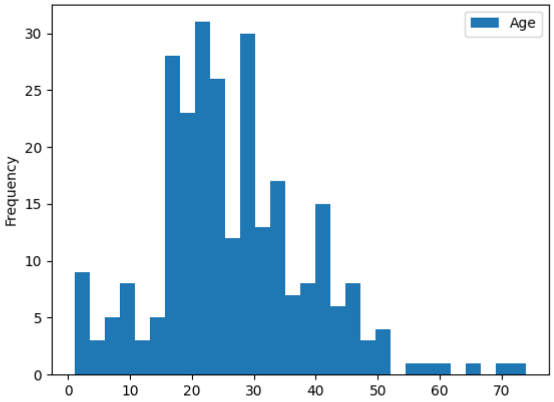
Survived = 1

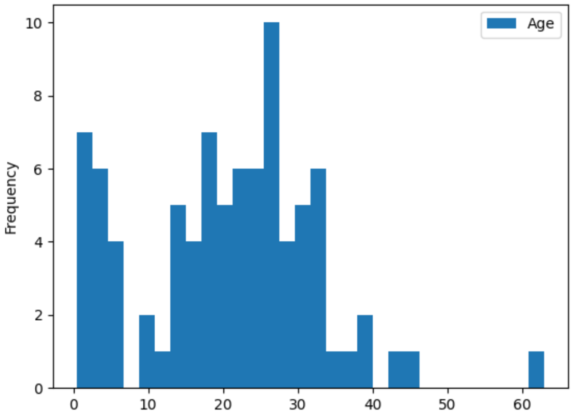




Pclass = 3
Survived = 0

Survived = 1

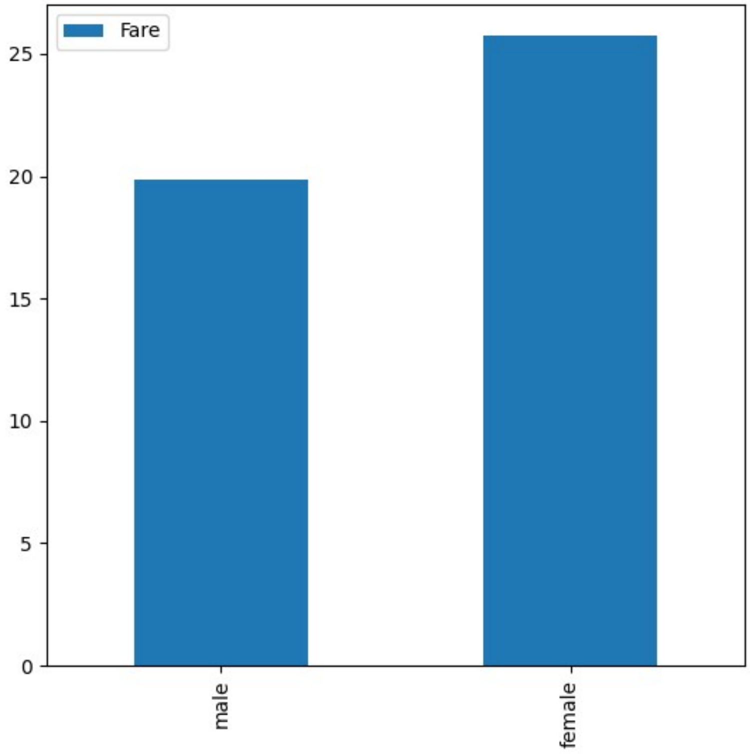




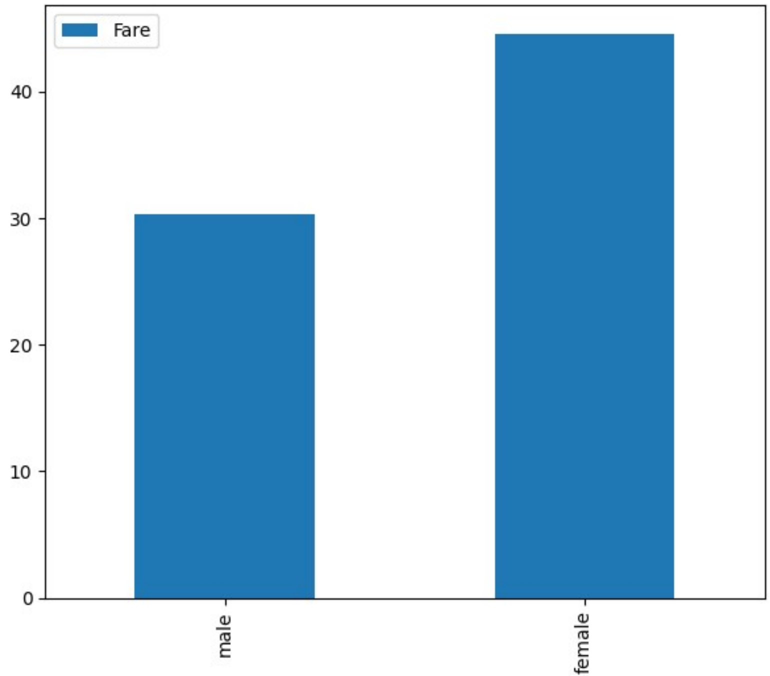
Q13: In training set, we want to correlate categorical features (with non-numeric values) and numeric features. We can consider correlating Embarked

(Categorical non-numeric), Sex (Categorical non-numeric), Fare (Numeric continuous), with Survived (Categorical numeric). Please plot a figure to illustrate the correlations of Embarked, Sex, Fare, and Survived.

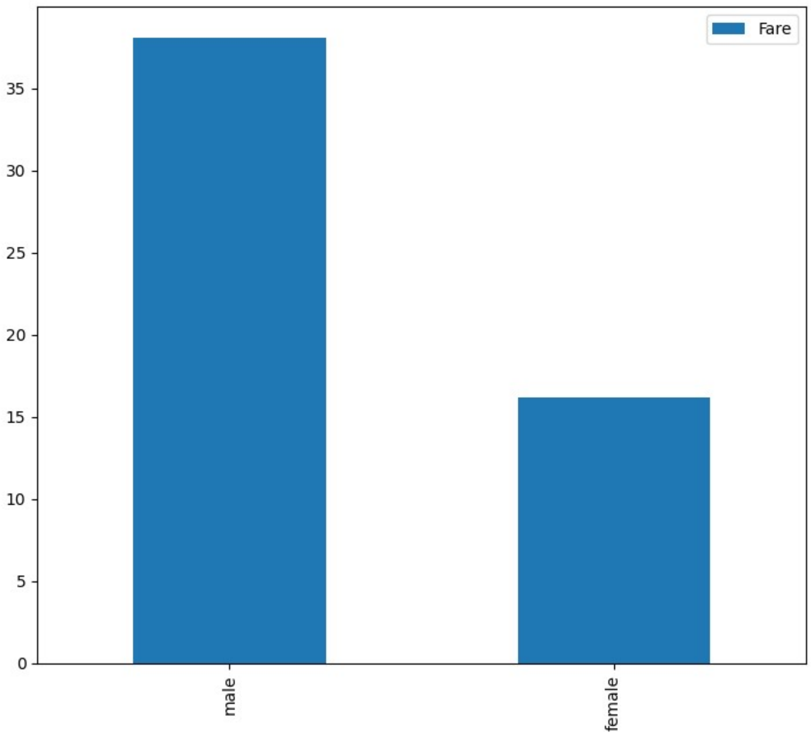
Embarked = S and Survived = 0



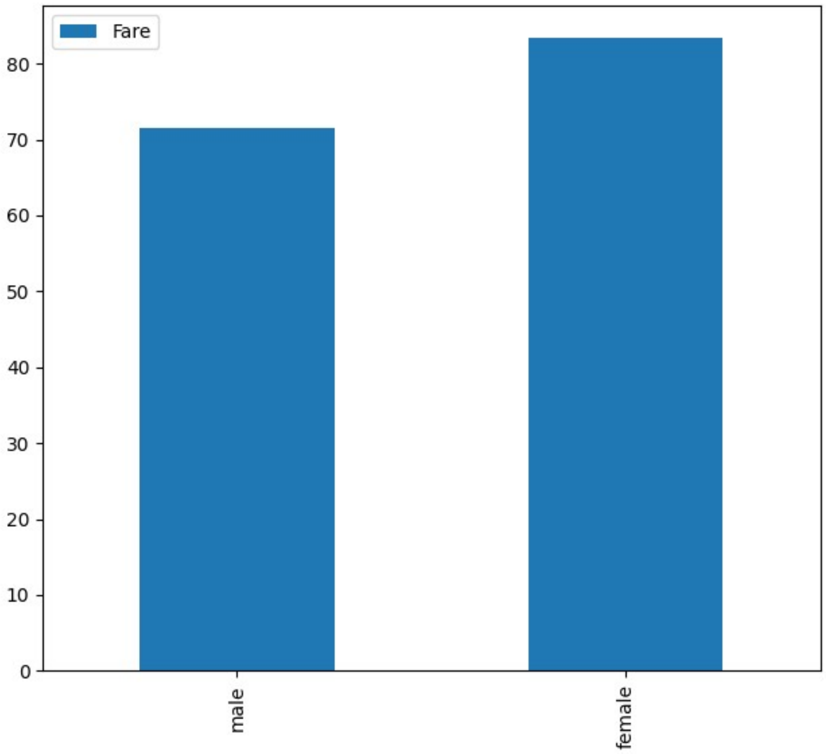
Embarked =S and Survived = 1



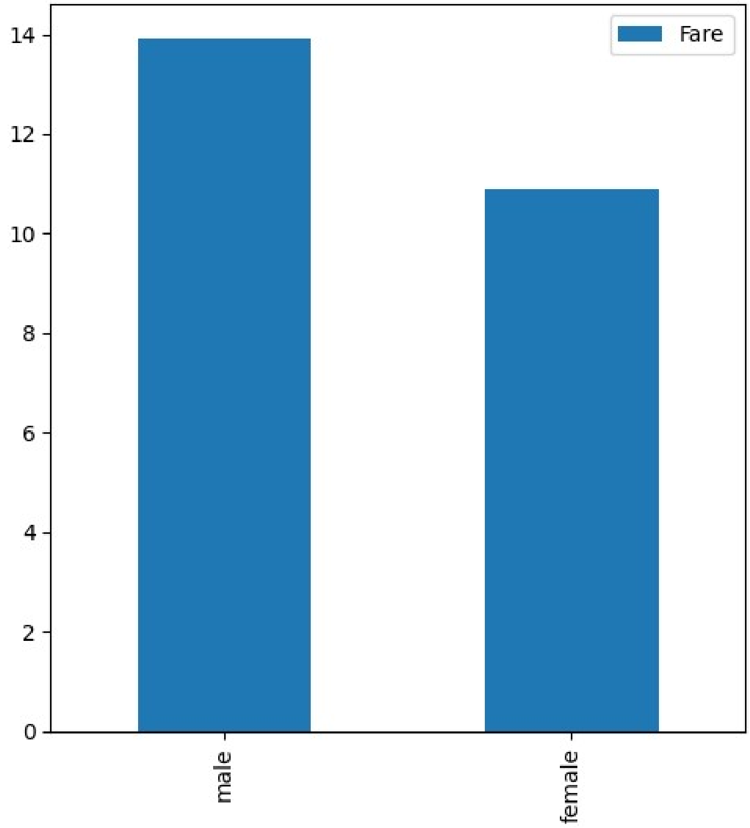
Embarked = C and Survived = 0



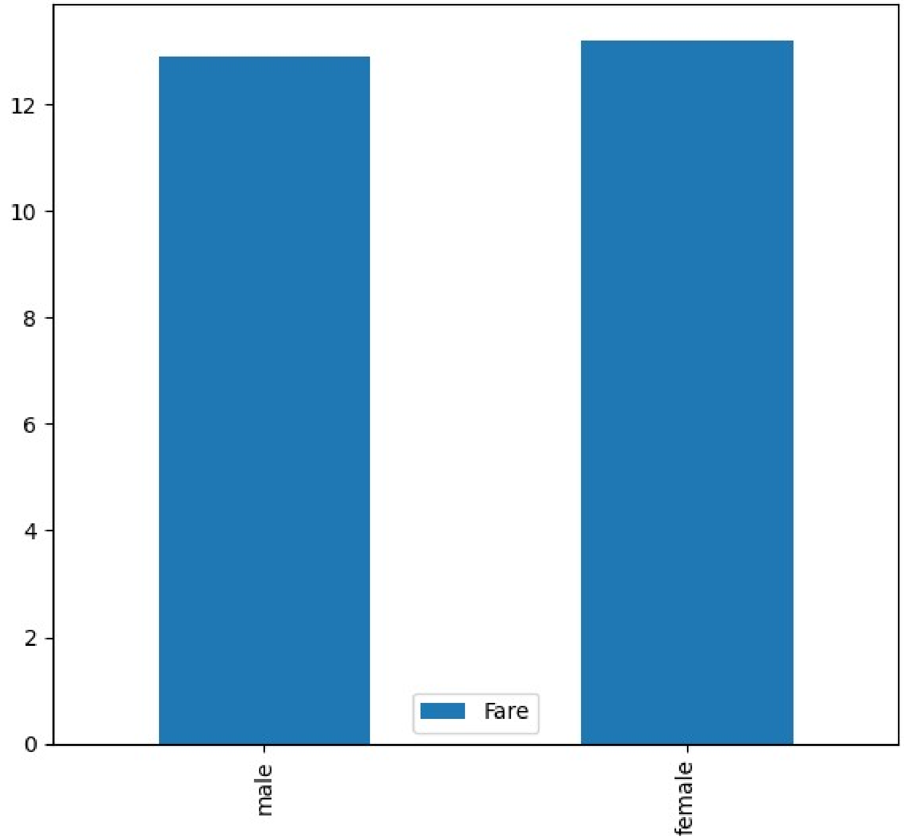
Embarked = C and Survived = 1



Embarked = Q and Survived = 0



Embarked = Q and Survived = 0



• Do higher fare paying passengers have better survival?

On average, yes. For both male and female survivors who embarked from C and S, their average fare was higher than those who did not survive and embarked from the same place. For those who embarked from Q, the average fare did not change very much when considering survivors. But the the difference between survivors and non-survivors from ports C and S is great enough for it to be considered significant.

It is possible that this is related to pclass in some way (passengers with higher pclasses paying a larger fare).

• Should we consider banding fare feature?

Yes. The fare feature is continuous and has such a wide range of possible values that it would be more meaningful to band.

Q14: In training set, what is the rate of duplicates for the Ticket feature? Is there a correlation between Ticket and survival? Should we drop the Ticket feature?

There are 210 duplicates in the ticket feature. It does not make sense for the ticket number to correlate to survivability in any way, it’s essentially a random number.

Therefore, we should drop it.

Q15: In the training set, Is the Cabin feature complete? How many null values there are in the Cabin features of the combined dataset of training and test dataset? Should we drop the Cabin feature?

The cabin feature is not complete.
687/891 = 77.1% null values in the training data
327/418 = 78.2% null values in testing data.
I think we should drop this feature because over 75% of the data is missing from this column in both sets. At that rate, no replacement method that we use would be very accurate.

Q16: In the training set, we can convert features which contain strings to numerical values. This is required by most model algorithms. Doing so will also help us in achieving the feature completing goal. In this question ,please convert Sex feature to a new feature called Gender where female=1 and male=0.

See code under comment # Q16. I used np.where to do the replacement, rename to change the name to Gender, and astype to covert it to an int.

Q17: In the training set, we start estimating and completing features with missing or null values. We will first do this for the Age feature. We can consider three methods to complete a numerical continuous feature. A simple way is to generate random numbers between mean and standard deviation. More accurate way of guessing missing values is to use the K-Nearest Neighbor algorithm to select the top-K most similar data points, and then use the top-K most similar data points to impute the missing values of ages.

See code under comment # Q17. I used the KNN method. I created a different dataframe in this case, because I believe that the KNNinputer requires that all columns be numeric. I couldn’t get it working otherwise. Obviously, if it were necessary, I could replace the old Age feature with the new Age feature from the new dataframe.

Q18: In the training set, complete a categorical feature: Embarked feature takes S, Q, C values based on port of embarkation. Our training dataset has some missing values. Please simply fill these with the most common occurrences.

See code under comment # Q18. According to the describe function, S is the most common value in Embarked. So I made the replacement using fillna.

Q19: In the training set, complete and convert a numeric feature. Please complete the Fare feature for single missing value in test dataset using mode to get the value that occurs most frequently for this feature.

See code under comment # Q19. I calculated the mode of Fare and did fillna to replace the missing test data.

Q20: In the training set, convert the Fare feature to ordinal values based on the FareBand defined follows:

See code under comment # Q20. I created bins that followed the FareBand values and used pd.cut to slice up the Fare feature to fit the labels detailed under ‘Ordinal Fare Indicator’.

Approximately how many hours did you spend on this assignment?

About 20.

Which aspects of this assignment did you find most challenging? Were there any significant stumbling blocks?

The most challenging aspect was the code. I knew what I wanted to do, I just needed to wrestle with pandas to get it all working. I am not a pandas expert, having only limited experience, so I had to get up to speed on a couple of things.

The other stumbling block was with the way some of the questions were worded. For example, Q17 details three ways that we can deal with missing values, but never specifies what method to use. I just chose to use KNN to understand how it worked.

The sample plots were a little confusing too. For example, in Q13, I did not understand how the sample plots given related to the question at hand. The question asks you to correlate fare, sex, embarked, and survived. However, due to the placement of the labels, I initially thought that embarked and survived only had something to do with the first four plots, while sex only had something to do with the last plot. I think this is because the first four plots had the sex x axis label covered up.

Which aspects of this assignment did you like? Is there anything you would have changed?

I enjoyed the coding aspect. It was a good experience to put to use some of the things we’ve been learning in class. If I had to change anything, it would be to clean up the wording of the questions a little and to make things a little clearer. I was able to resolve most of my questions by speaking with the TA, but it would just have been an overall smoother process to have had clearer questions in the first place.