**Q1: In training set, which features are available?**

In the training set, we have the following features: socio-economic status, name, sex, age, sibling/spouse, parent/child, ticket, fare, cabin, and embarked. Here I am not including whether the passenger survived since that is what we are predicting.

**Q2: In training set, which features are categorical?**

Socio-economic class, sex, and embarked were the only truly categorical data types.

**Q3: In training set, which features are numerical (e.g., discrete, continuous, or time series based)?**

Age, passengerId, survived, Pclass, Sibsp, Parch, and the cabin **number**, could be considered discrete data types. The fare is considered a continuous data type.

**Q4: In training set, which features are mixed data types?**

Cabin number is a mix of each discrete and categorical data types, since people could be grouped together by cabin (A, B, C, D, or E), but they also have a unique number assigned to each letter. The ticket is also mixed, as it has a port identifier as well as a unique ticker number. I would say parch and sibsp could be mixed data types as well, since two people with the same number don’t necessarily have the same number of siblings/parents/children (if I interpreted the variable correctly).

**Q5: In training set, which features contain blank, null or empty values? In test set, which features**

**contain blank, null or empty values?**

In the training set, the age, cabin, and embarked features all contain missing values. In the testing dataset, the age, fare, and cabin features contain missing values. There are different missing features in each dataset.

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

In the training dataset, the data types of the features are as follows:

* Pclass -> int
* Name -> str
* Sex -> str (though could be represented as an int)
* Age -> int
* Sibsp -> int
* Parch -> int
* Ticket -> str and/or int
* Fare -> float
* Cabin -> str and/or int
* Embarked -> str (though could be represented as an int)

**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?**

**From Python Script (looks like it matches, they were computed with numpy on train set)**

**PassengerId Survived Pclass Age SibSp**

**count 891.000000 891.000000 891.000000 714.000000 891.000000**

**mean 446.000000 0.383838 2.308642 29.699118 0.523008**

**std 257.209383 0.486319 0.835602 14.516321 1.102124**

**min 1.000000 0.000000 1.000000 0.420000 0.000000**

**25% 223.500000 0.000000 2.000000 20.125000 0.000000**

**50% 446.000000 0.000000 3.000000 28.000000 0.000000**

**75% 668.500000 1.000000 3.000000 38.000000 1.000000**

**max 891.000000 1.000000 3.000000 80.000000 8.000000**

**Parch Fare**

**count 891.000000 891.000000**

**mean 0.381594 32.204208**

**std 0.805605 49.665534**

**min 0.000000 0.000000**

**25% 0.000000 7.910400**

**50% 0.000000 14.454200**

**75% 0.000000 31.000000**

**max 6.000000 512.329200**

**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?**

**Pclass Sex Embarked**

**Count 891 891 891**

**Top 3 male S**

**Freq 491 577 644**

**Unique 3 2 3**

**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 Survived, we should include this**

**feature in the predictive model. Based on your computation, will you include this feature in the**

**predictive model?**

**pclass and survived have a weak correlation of -0.338481,**

**Therefore pclass won't be included in the features for this model**

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

**35.241302 percent of women survived.**

**64.758698 percent of men survived.**

**Women are less likely to survive.**

**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?**

**• Do oldest passengers (Age = 80) survive?**

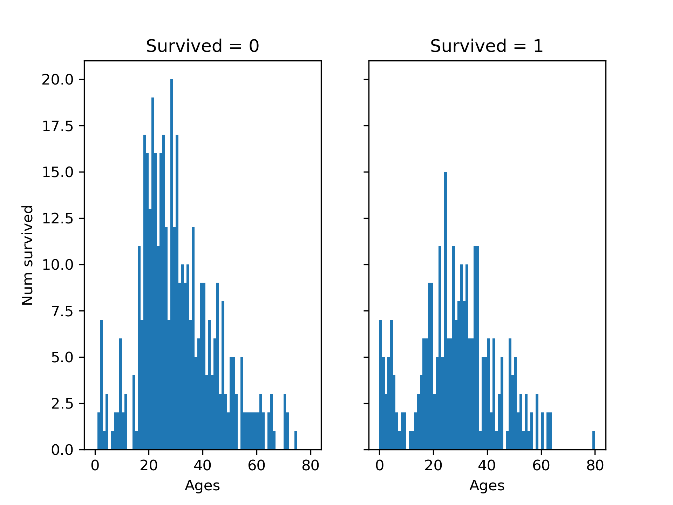
**• Do large number of 15-25 year olds not survive?**

**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.)**

**• Should we should band age groups?**

****

* **Infants (Age <= 4) have a decently high survival rate**
* **Looks like all the oldest passengers survived.**
* **A large number of 15-25yr olds don’t survive.**
* **Based on this, we should include age in the model training, since it seems to have a large impact on whether or not someone survived.**
* **We should band the age groups, since specific age groups have much higher chances of survival than others.**

**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.**

**Here is an example plot:**

**Figure 2: a sample histograms plot of age, Pclass, and survivied.**

**Please plot the histogram plot using python, and answer the following questions:**

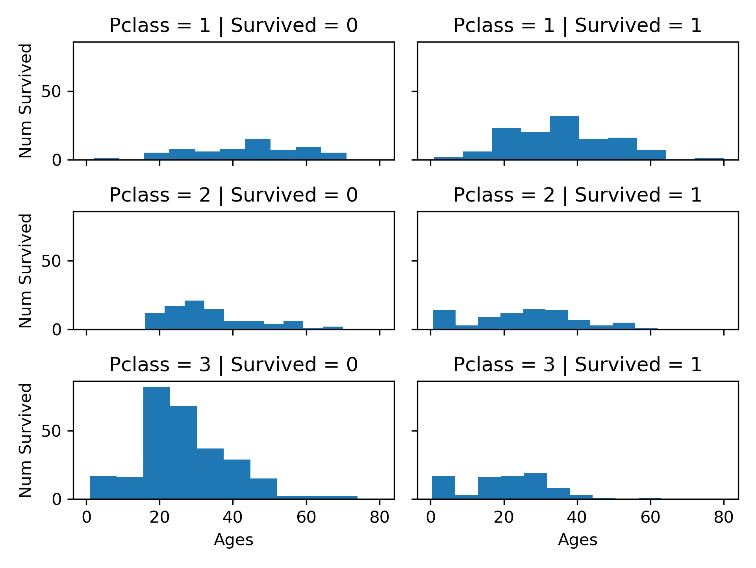
**• Does Pclass=3 have most passengers, however most did not survive?**

**• Do infant passengers in Pclass=2 and Pclass=3 mostly survive?**

**• Do most passengers in Pclass=1 survive?**

**• Does Pclass vary in terms of Age distribution of passengers?**

**• Should we consider Pclass for model training?**

****

**Based on my above graphs:**

* **Pclass=3 has the most passengers, but most did NOT survive.**
* **Infant passengers in Pclass=2 and Pclass=3 mostly survived**
* **The majority of Pclass=1 passengers survived**
* **It looks like Pclass=3 has a lower average age than Pclass=1**
* **Pclass should be considered, since it seems to have a direct correlation with num survived**

**Q13: 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 histogram figure to**

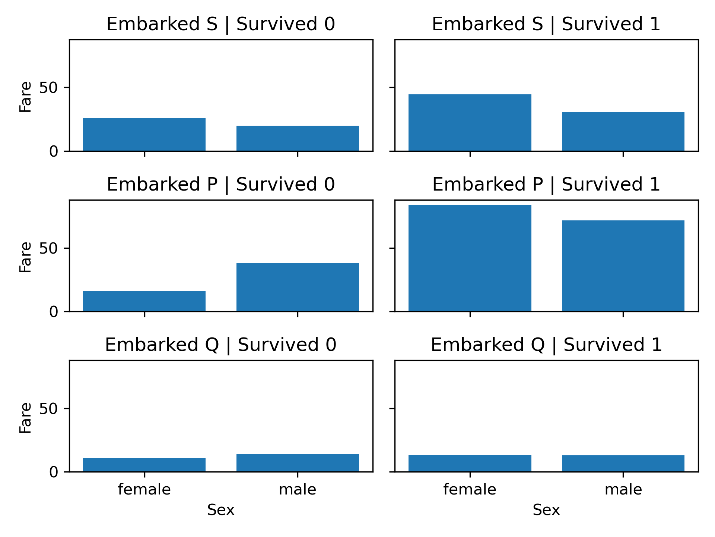
**illustrate the correlations of Embarked, Sex, Fare, and Survivied. Here is a sample plot:**

**Figure 3: a sample figure of the correlations of Embarked, Sex, Fare, and Survivied**

**And answer the following questions:**

**• Do higher fare paying passengers have better survival?**

**• Should we consider banding fare feature?**

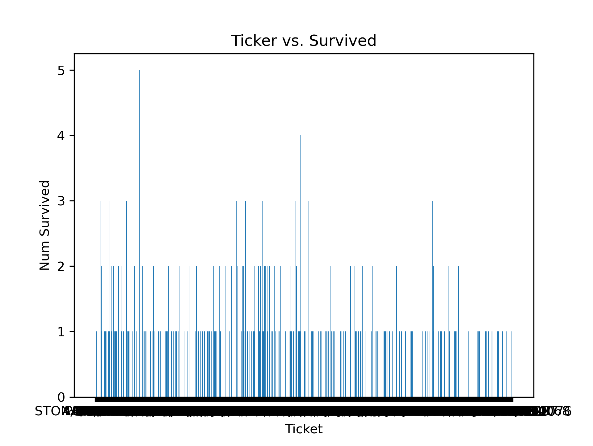
****

Based on my above graph:

* It seems like higher fare payed passengers have a high chance of survival, except for those embarking from port Q
* It is hard to say whether to band the fare feature from these charts, since we can’t see the actual distribution of fares. Though if fares of different pclass or ports are in well-defined groups, then it would be worth it to band the fare feature.

**Q14: 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 duplicate values for ticket, so about 23.57 percent.

Based on the above graph, there is not much correlation between ticket and survived, therefore it should be dropped as a feature.

**Q15: 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?**

There are 891 passengers, but only 204 entries for cabin.

This means there are 687 null values.

Therefore, the cabin feature won't be used.

**Q16: 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.**

Done in Python code.

**Q17: 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.**

Done in Python Code

**Q18: Completing 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.**

**S 644**

**C 168**

**Q 77**

Done in Python code

**Q19: Completing and converting 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.**

Fare Mode == 8.050000

Done in Python code.

**Q20: Convert the Fare feature to ordinal values based on the FareBand defined follows:**

**Ordinal**

**Fare**

**Indicator**

**FareBand Survivied**

**0 (-0.001,**

**7.91] 0.197309**

**1 (7.91,**

**14.454] 0.303571**

**2 (14.454,**

**31.0] 0.454955**

**3 (31.0,**

**512.329] 0.581081**

**Please submit a PDF report. In your report, please answer each question with your**

**explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At**

**the end of your PDF, please include a website address (e.g., Github, Dropbox, OneDrive,**

**GoogleDrive) that can allow the TA to read your code.**