Data Preprocessing and Analyze by pivoting features of Titanic dataset

You can download the Kaggle Titanic dataset from https://www.kaggle.com/c/titanic/data. (The Titanic.zip is uploaded to modules->data for your fast reference.) The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.
The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.
We also include gender_submission.csv, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

Data Dictionary

VariableDefinitionKey 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

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

Let us start with acquiring data: The Python Pandas packages helps us work with our datasets. We start by acquiring the training and testing datasets into Pandas DataFrames. We also combine these datasets to run certain operations on both datasets together.

```
train_df = pd.read_csv('../input/train.csv')
test_df = pd.read_csv('../input/test.csv')
combine = [train_df, test_df]
```

There are three clarifications for homework 1: (1). You only need to answer all questions in homework 1 by analyzing the train set; (2) Ticket is mixed data types. Because ticket mixes numeric and alphanumeric data types. The value of the cabin is alphanumeric.; (3) The sample graph in Q13 is wrong. The correct graph is just the reversed version. The observation is caused by a bug in Pandas plot.

Q1: In training set, which features are available?

Q2: In training set, which features are categorical?

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

Q4: In training set, which features are mixed data types (Carbin is not mixed data type: [C19 C18 C17], Ticket is a mixed data type: 342354, SA/12434324)?

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

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

Q7: In training set, 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?

Q8: In training set, 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?

Q9: In training set, 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?

Q10: In training set, are Women (Sex=female) were more likely to have survived?

Q11: In training set, 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 histograms 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 figures,

- 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?

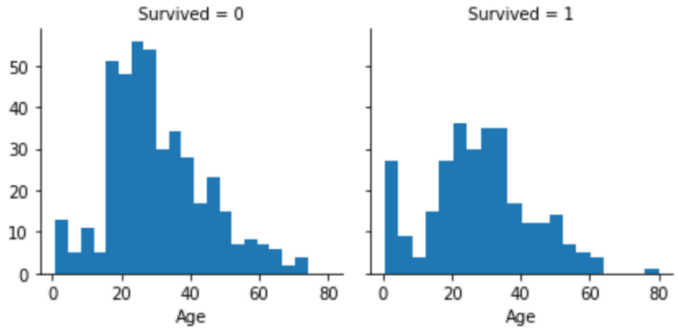

Figure 1: a sample histogram plot of age

Q12: In training set, 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:

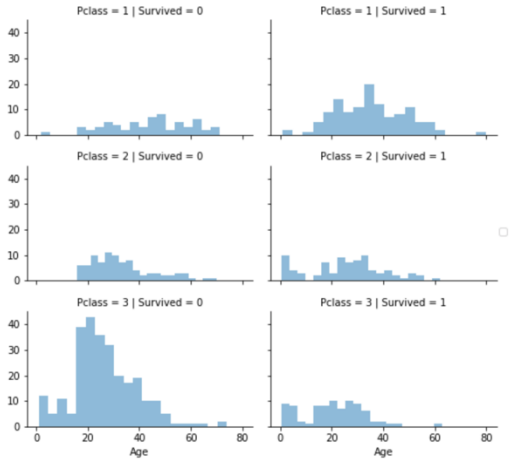

Figure 2: a sample histograms plot of age, Pclass, and survivied.

Please plot the 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?

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 Survivied. Here is a sample plot (Figure 3):

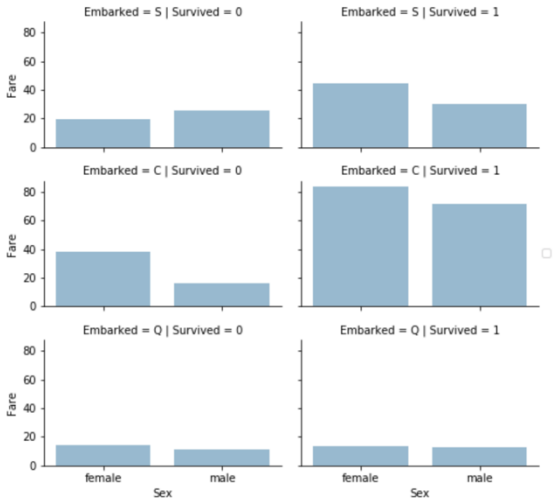

Figure 3: a sample figure of the correlations of Embarked, Sex, Fare, and Survivied (notice that your plot might not be the same with the sample plot)

And answer the following questions:

- Do higher fare paying passengers have better survival?
- Should we consider banding fare feature?

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?

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?

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.

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.

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.

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.

Q20: In the training set, 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

Additional Questions:

- Approximately how many hours did you spend on this assignment?
- Which aspects of this assignment did you find most challenging? Were there any significant stumbling blocks?
- Which aspects of this assignment did you like? Is there anything you would have changed?

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.