Part 3: Complete this tutorial (Steps 14&15) and then use what you learn to create your own case study.

The data set is available in Weekly Resources/Data

**Case Study: Analyze data to predict who will Survive the Titanic**

1. Load the data from the “train.csv” file into a DataFrame.
2. Display the dimensions of the file (so you’ll have a good idea the amount of data you are working with.
3. Display the first 5 rows of data so you can see the column headings and the type of data for each column.
   1. Notice that Survived is represented as a 1 or 0
   2. Notice that missing data is represented as “NaN”
   3. The Survived variable will be the “target” and the other variables will be the “features”
4. Think about some questions that might help you predict who will survive:
   1. What do the variables look like? For example, are they numerical or categorical data. If they are numerical, what are their distribution; if they are categorical, how many are they in different categories?
   2. Are the numerical variables correlated?
   3. Are the distributions of numerical variables the same or different among survived and not survived? Is the survival rate different for different values? For example, were people more likely to survive if they were younger?
   4. Are there different survival rates in different categories? For example, did more women survived than man?
5. Look at summary information about your data (total, mean, min, max, freq, unique, etc.) Does this present any more questions for you? Does it lead you to a conclusion yet?
6. Make some histograms of your data (“A picture is worth a thousand words!”)
   1. Most of the passengers are around 20 to 30 years old and don't have siblings or relatives with them. A large amount of the tickets sold were less than $50. There are very few tickets sold where the fare was over $500.
7. Make some bar charts for variables with only a few options.
   1. Ticket and Cabin have more than 100 variables so don’t do those!
8. To see if the data is correlated, make some Pearson Ranking charts
   1. Notice that in my sample code, I have saved this png file.
   2. The correlation between the variables is low (1 or -1 is high positive or high negative, 0 is low or no correlation) These results show there is “some” positive correlation but it’s not a high correlation.
9. Use Parallel Coordinates visualization to compare the distributions of numerical variables between passengers that survived and those that did not survive.
   1. That’s a cool chart, isn’t it?! Passengers traveling with siblings on the boat have a higher death rate and passengers who paid a higher fare had a higher survival rate.
10. Use Stack Bar Charts to compare passengers who survived to passengers who didn’t survive based on the other variables.
    1. More females survived than men. 3rd Class Tickets had a lower survival rate. Also, Embarkation from Southampton port had a lower survival rate.
11. Some of my questions have been answered by seeing the charts but in some ways, looking at this much data has created even more questions.
    1. Now it’s time to reduce some of the features so we can concentrate on the things that matter! There features we will get rid of are: "PassengerId", "Name", "Ticket" and "Cabin". (ID doesn’t really give us any useful data, Ticket and Cabin have too many variables. Name might reflect that they are related but we’re keeping the category about siblings (for now).
    2. We can also fill in missing values. (Cabin has some missing values but we are dropping that feature.) Age has some missing values so I’ll fill in with the average age. Embarked also has some missing so I’ll the most common.
12. If you go back and look at the histograms of Fare, you’ll see that it is very skewed…many low cost fares, not very many high cost fares. Log Transformation is a good method to use on highly skewed data.
13. Convert your categorical data into numbers (Sex, PClass, Embark)
14. Training - Split your data into two sets: Training and Testing.
15. Evaluation – Remember, we are trying to predict if a passenger has survived or not so this is a classification problem. There are many algorithms that could be used but we’re going to use logistic regression.
    1. Metrics for the evaluation:
       1. Confusion Matrix (you should get 84% - pretty good)
       2. Precision, Recall & F1 score (all 3 were very good)
       3. ROC curve (the dotted line is the randomly guessed so anything above that is good metric)