Lab 3: Data Visualization

Our goal in this lab will be to practice with exploratory data analysis, specifically focusing on different visualization techniques. We will continue with our use of the Titanic data with an eye towards thinking about what variables we should include in a machine learning model once we get to the modeling stage of the data science process. We will leverage the work done last week to fill missing values in the age column this week.

1 Reading in the data

1. First read the data into your python development environment using pandas. You'll need to import pandas and then apply the '.read_csv' method. We will be using an updated data set with the age variable missing values filled using the mean, the median, and the mode of the non-missing values, as well as by using kNN to learn and predict the missing values.

2. As in the last Lab, we will need to type the data appropriately. Convert the following variables to categorical variables: Pclass, Sex, and Embarked. You may find the .astype method to be useful.

3. Answer the following question in your writeup:

   (a) For each variable in the data set, identify the type of the data it contains and articulate the reason for your assessment. Refer to the typology from the data types lecture.

2 Classification

1. In the last lab you identified a candidate for a label/target/response/dependent variable for a supervised machine learning problem – we will proceed with the Survived variable as our response variable and begin to build a case for a particular selection of features. We will do this using visualizations.

2. Begin with comparing Pclass to Survived using a heatmap. You may find the .heatmap method from Seaborn to be useful. To use it, you will need to aggregate the data into a dataframe with the following form:

Survived/Pclass	1	2	3
0	?	?	?
1	?	?	?

   Each cell should contain the fraction of passengers of a given class for each survival outcome. For example, the upper right cell should contain the fraction of all third class passengers who did not survive. You may find the .pivot_table method from pandas to be useful. You can check your calculations–each column should sum to one.

3. Next, determine the best visualization to compare each of the following variables to the Survived variable: Fare, Sex, SibSp, Parch, Embarked. You will need to use the answers you came up with for question 1.3.a to do this effectively.

4. Create these visualizations.

5. Finally, determine the best visualization to compare Age with Survived. For each of Age_fill_mean, Age_fill_median, Age_fill_mode, and Age_fill_KNN create a visualization to compare this to Survived.

6. Answer the following questions in your writeup:

   (a) What are appropriate plot types for comparing each variable to Survived and why?

   (b) What would you need to observe in each visualization type to conclude that there might be a predictive relationship between the variable and Survived?

   (c) Identify any predictive relationships in your data by interpreting your plots–what do the x and y axes mean? What do they tell us about the relationships between each variable and Survived? Provide the plots for the potentially predictive variables in your writeup. Explain why it does or does not make sense that that variable would be related to Survived – not in terms of the evidence from your visualizations, but in terms of what the variable measures. What do you learn about Edwardian society?

   (d) Does that manner in which you fill missing values in the Age column affect your conclusions on the existence of a predictive relationship between Age and Survived? How do you know?

   (e) Consider the PassengerID and Cabin variables. We did not include them in our analysis as potential predictors for Survived. What would we need to do to these variables if we did want to use them. Suggest one transformation you could apply to each to make use of it and why it needs to be transformed.

   (f) Which variables would you want to include in a machine learning model? Is a plot that appears 'predictive' necessary or sufficient for that variable to have a predictive relationship with Survived?

3 Comparing features

1. When we do machine learning if two features are highly correlated with one another it is of limited value to include both as predictors (and indeed it may even hurt model performance).

2. Consider the Age and Fare variables. We will compare these visually to see if they are correlated with one another. Make a scatterplot and a jointplot of Age_fill_KNN vs Fare. You may find the .scatterplot and .jointplot methods from Seaborn to be useful. If you use the latter you will need to set kind = 'kde'.

3. Answer the following question in your writeup:

   (a) Do you observe any relationship between Age_fill_KNN and Fare? Given the answer to this question is there any reason not to include both in a machine learning model at this point?

   (b) Why would you not include two correlated variables?

4 Submission Instructions

- Write up the answers to the questions in a short 2 page word document. Add footnotes identifying which sentence addresses which questions. Use complete sentences – your goal is to explain your dataset to the audience, what you've done to it, and what you've learned about it. Accordingly, you may seek to emulate some of the sections of the whale paper describing their data. Include the appropriate plots you've generated as mentioned above. Convert this to pdf and submit it. Submit your .ipynb file as well.

- The grading rubric for this assignment will be available in Canvas.

- NO OTHER SUBMISSION TYPES WILL BE ACCEPTED.

- **Late policy**: 5% of total points deducted per day for three days – after that no submissions allowed.