CSE 578 Project Progress Report

Problem Statement

The XYZ corporation is developing marketing profiles for one of their clients, UVW College. The objective of the marketing profiles is to enhance their enrollment rates by identifying individuals that earn more or less than $50,000 annually, so the college can tailor their marketing strategies accordingly. Our task as an analyst employed at the XYZ corporation is to utilize the data provided by the US Census Bureau to identify what key factors among the 13 variables that are in the dataset affect one's income. The final goal of this project is to utilize the data to create 5 story-telling visualizations about the features in the dataset. Finally, this project will be handed off to be used in a model to accurately predict the income of individuals so they can tailor their marketing strategies accordingly.

Progress Made & Background Work

To begin this project, I observed the different types of features within the dataset and identified what data types are present within the dataset. For example, there are nominal variables such as the 'native-country' and 'race' feature, ordinal variables such as the 'education' feature, and most of the other variables are of the ratio variable type such as 'age,' 'capital-gains,' and 'hours-per-week.' Next, I looked over the data and noticed that there were missing values within the dataset using the '?' value. Moving on, I loaded the data into a Dataframe using the pandas library. The next step I took was to replace all the '?' values in the Dataframe with the NaN datatype from the NumPy library, so I can easily drop the entries using the 'dropna()' method in the pandas library. Furthermore, as instructed by Professor Ghayekhloo, I dropped the 'fnlwgt' feature in the dataset. The final step in the data preprocessing process was to split the data into two Dataframes based on the income feature. To summarize my current progress in the project so far, I've been able to create my first visualization, grouped bar chart between Capital Gains & Income, that might give the client some insight into which client earns more or less than $50,000 annually. Furthermore, I was able to discover what visualizations might not be insightful to the user by exploring the native-country feature using a heat map. I will go into more details in the next section.

Specific Tasks Completed

The first user story that I chose was to create a grouped bar chart between the 'Income' feature & 'Capital Gains' feature. Initially, I split my dataset into two categories based on the 'Income' feature, one group earning $50,000/year and the other earning more than $50,000/year. To identify what I would be plotting from the 'Capital Gains' feature, I observed the different values within the feature. What I noticed that the values were either of 2 types, they were either zero or non-zero values. So, I used a dictionary to split it into the categories mentioned above. This decision was made based on researching the docs provided by the matplotlib library. Next, I used the capabilities of the Dataframes to fill in the dictionary accordingly. Furthermore, I utilized the notes mentioned in the second module indicating the shortcomings of induvial bars about precision of values and added the exact value of each bar on top each bar respectively. Moving on, I'll talk about other plots I've worked on and the challenges that arose from them.

Issues Encountered and Plan Moving Forward

Before generating the plot mentioned above, I first explored the possibility of creating a heat map between the 'native-country' and 'income' features. To begin this process, I imported the GeoDataFrame of the world using the geopandas library. Next, I observed all the unique values mentioned in our dataset and the difference in spelling between our dataset's strings and the GeoDataFrame's strings, which I adjusted accordingly. Before merging the datasets, I noticed that there are a couple of countries that would not have valid entries in the GeoDataFrame's entries, so I dropped the entries that contained the following countries 'Outlying-US(Guam-USVI-etc)', 'England', and 'South'. It is important to mention that each of these countries had less than 10 entries each in either the 'less_than_50k' or 'more_than_50k' dataset. After plotting the data, I noticed that the data has an outlier, which is the country of the USA. This was caused by the USA having more entries than any other country by a magnitude of over 1000. This issue leads the heat map visualization have little to no value to the user, as all other countries had similar values. Moving forward, I should pay attention to features where most of its values are in a single value and avoid those, as these visualizations would not provide us with any insights.

Tasks yet to be Completed and Approach

As of now, I have 4 user stories remaining. I plan on creating a mosaic plot either between a binary feature such as 'sex' and the 'income' feature or choosing a feature of nominal data type to compare with the 'income' feature in hopes of finding a correlation. Next, I plan on using the 'workclass' feature using a bar or pie chart to visualize the correlation if there exists any. When it comes to the two final user stories, I have not decided on what features to explore, but I plan on visualizing more continuous features using a box-and-whisker plot and/or line charts. Moving forward, I plan on extensively researching the dataset to ensure that the features that I am visualizing do not end up with a similar issue to the heat map visualization. Furthermore, it is safe to assume that if I'm able to achieve a generally accurate/acceptable visualization, I intend on enhancing the plots to ensure that they adhere to all the principles/rules mentioned in the earlier modules.