Chun-Wei Hu_CSE 578_Course Project Progress Report

1. Include your problem statement

There are are four problems in this project. First, this project is aiming to help UVW college to analyze how to marketing it's degree programs base on a key of 50K in individual income.

Second, clarify the facts that affects each individual to their income making. Which may compare two factors or above.

Third, make multiple charts and plot to help customers easily undetdstand the analysis and the difference between datasets.

Last but not least, predict each person's income to help analyze marketing efforts.

2. Describe the progress you have made so far, including the background work you completed

For the first problem, I finished cleaning data and separate data into two category: above 50K and below 50K.

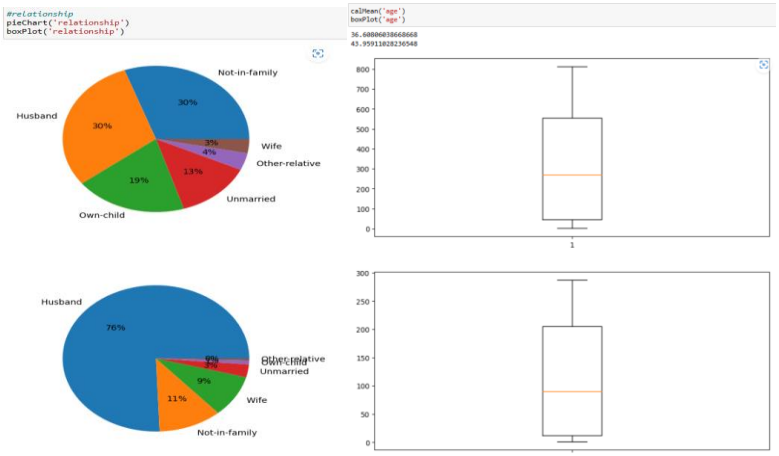
According to the second problem background, I finish creating charts for each attributes to see which affects the income most.

For the third question, I complete some of the multivariate analysis charts.

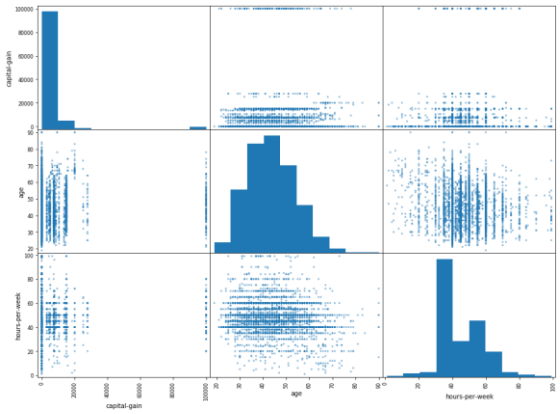
3. Summarize the specific tasks you have completed to date

First, I list all the data of each attributes and making sure what kind of charts fits them. If the unique data of the attributes are more than 6, I will choose not to use pie chart since the name will stick together and make it unclear.

Second, I make three charts for each columns: Pie chart, hist chart and Box plot, this allows me to identify which chart is more viewable for customers which helps them to get the informations they need.



Third, I tried to clearify which columns can be combined in order to give more information to the analysis, also try to pick a few specific charts of the combined data that makes customer more readable, this also helps me check whether the datas are clear spearate enough for me to proof my thoughts about my analysis.



Last but not least, I find out a proper way to predict the income. In order to make this happen, I find a proper range through the data charts and plots for each attributes that I think 70% or above are included. This will make the prediction's accuracy better. I will also create a function that gerenate the prediction and then compare to the original data to see if the prediction is accurate or not.

4. Discuss issues you have encountered thus far and your plan for solving them

At first, I can't read the data properly when I try to clean all the data with no ' ? 'in it. After I listed all the data in the column, I found out that all the datas have a space in front of it, so in order to clean the value or make use of the string, I have to also put a space before it and it worked.

The second issue I had met is that when I show charts of the data by calling "fig, ax=subplot..." at every block, it seems to also shows the result that the plot is using. But after I change the building chart part into a new function, the result will not show along with the charts.

Third, when I try to compare the data which the income is above 50,000 with the data under 50000, I found that both data did not have the same sample numbers. So I choose the first part of the data that it's icome is above 50,000 with the same number of the data under 50,000. By using the part of the data, I can at least starting to analyze the difference between the dataset of income above 50,000 and under 50,000.

5. Summarize the tasks you have yet to complete and how you intend to approach them. Be specific

The first problem I'm going to search for more information about making income across more than two attributes. This will also helps me to get a more accurate analysis for the marketing. I'll generate more charts for each unique attributes so that I can get a more detail reasearch on what will affects income

The second problem I'm going to solve it that making sure the 7000 more datas can represent the whole data that their incomes are above 50,000. I will try to draw charts of both 7000 dataset and the whole dataset and compare them together. If the stats shows that the percentage of both dataset are really similar, I will put it in the final report to show the customer that this part of data is reliable. On the other hand, if the data is not as aimiliar as the original dataset, I think I will try to separate the original dataset to find out which part will fit the most.