Lucy Zhang
ID:1229576766
02/01/2023

Course Project Progress Report

As a data analyst at XYZ company, I have been given a task to bolster enrollment at UVW College by using the United States Census Bureau data and focusing on a $50,000 salary as a key demographic to create data visualizations. In this project, I will review all the features in the data to determine what are the key factors for higher and lower income, and I will calculate the correlation between them and visualize with graphs accordingly.

The background work I have completed is exploring the given data set and cleaning the data set. After opening the data sets and matching values with corresponding name columns, I found there were some value duplicates, and I managed to drop the duplicated data. Some data marked as "?" as a missing value in the data set, and I have replaced them with "None" and removed the "fnlwgt" attribute as instructed.

Each person has 13 features of data in the cleaned data set, and I would like to compare 8 of the features with income to determine the key factors for further analysis. From the given features, I assumed age, work class, education, marital status, occupation, race, sex, and native country are more relevant to income compared to education grade, relationship, hours per week, and capital gain or loss.

The progress so far included analyzing values in each feature and comparing features with income by using a corresponding visualizing graph. More specifically, I tried to find the best visualizing tool for each feature. For example, comparing Box and Whisker Plots with Pie Chart for the "age" feature since it's numerical data, and whether Pie Chart has a better visual appearance than Mosaic Plot for "sex" feature data. Also, I've written down my obsecration about each visualization.

Observation on each feature with visualizing graph:

- Age: By using Box and Whisker Plots, I observed around age 25 to 47 fall into the interquartile lower income range, and the median is around age 35. On the other hand, around age 35 to 51 falls into the interquartile range of higher pay, and the median is around age 45.
- Work Class: By using a Pie Chart, I observed the percentage of people who works in a private firm that makes less or equal to $50,000 is more significant than people who make more, and the percentage of people who are self-employed in a higher-income is higher than people in lower income group.
- Education: Using a Line Graph, I observed that a large percentage of people in the higher income group have bachelor's degrees, and people in the lower income group have a large portion with only a high school degree. Only a few or none with higher income have an education level lower than a high school degree.
- Marital Status: By using Mosaic Plot, I observed that only a tiny portion of people who never married make more, and about half of married individuals in this data set have lower incomes.
- Occupation: By using the Bar Chart, I observed that the group of people who makes more have a career of executive managerial, professional specialty, and sales have a higher percentage compared to people who make less. Conversely, the rest of the occupations in this data set have a higher rate with people with lower income.

- Race: By using Pie Chart, I observed that white and Asian Pacific Islanders races have nearly 30% of people with higher incomes, but black, Indian Eskimo, and other races, only 10% of people of their race make a higher income.
- Sex: By using Mosaic Plot, I observed that males in this data set are almost twice the size of females, and the percentage of males who make more in the male group is around 1/3. However, only less than 1/6 of females fall into the higher income group.
- Country: By using Bar Chart, I observed that outlying areas of the U.S. and Holland Netherlands have very few or no people making more than $50,000. Also, Philippines and Honduras have less than 10% of people with higher income. Other countries have a similar ratio of around 15% of people in the higher income group.
- Based on the observations of the data, we can conclude that age, education, marital status, gender, and occupation significantly affect income. For example, a 45-year-old self-employed married man has a higher probability falls into a higher income range than a 35-year-old single woman who works in a private firm.

User stories that I prioritized for this project based on selected attributes and observations:

1. Race diversity in school has always been an important topic. From the visualization of Race vs. Income observation, the school can market to black, Indian Eskimos, and other races other than white and Asian with different approaches. Such as offering financial aid and scholarships specifically for students of targeted races to make higher education more accessible.
2. According to the Country vs. Income graph observation, we can create a program that targets international students who have high school diplomas and are interested in obtaining a higher level of education in the United States and improving their chances of finding a higher-paying job. Also, the school can provide grand or some financial relief plan for students from Outlying US, Holland Netherlands, Philippines, and Honduras.
3. Since data showed that a large portion of never-married individuals has lower income, I assumed the majority are under age 45, have a full-time job during the daytime, and live alone. Thus, we can create a night school program on skill development, such as classes focusing on Executive Management, Professional Specialty, and Sales.

One of the issues and challenges encountered during this project was coding Mosaic Plot using matplotlib, which we have yet to learn, but this issue has already been resolved by using statsmodels module. Another problem is that the education level in the data set has no rank, and it's also resolved by manually typing in the rank accordingly. While graphing Race vs. Income, I first used two Pie Charts for higher and lower income and found the visualization presentation unclear in terms of comparison. I resolved this by switching it to a Bar Chart with a percentage-based representation.

There still are tasks that need to be completed, such as finishing writing user stories that I prioritized for the project based on selected attributes and finding the best visualization I learned so far for each scenario. Also, improving the visualization on each graph accordingly. Such as if colors are too similar and labels overlap.

The approach I'm planning to complete my tasks is optimizing my graph, computing correlations between features and income, and providing different plans for bolstering enrollment at UVW college by using the information provided by visualizing graphs accordingly. Present the charts with more precise visualization by implementing the color and labels on the graphs.