CSE578: Data Visualization

Portfolio

Nourhan ElNaggar - nelnagga@asu.edu
Arizona State University

Abstract-This portfolio report is for Data Visualization project CSE579. Most of the content was taking from system documentation report[1], Executive Report Presentation[2] and Individual Contribution Report [3] that was made earlier for the-course's project.

I. Introduction

This project was made for UVW College, a local college looking to assist it in marketing its degree programs and increase the number of prospective students depending on a salary as a key demographic and other important features related to citizens' demographics. We are given a dataset of the United States Census Bureau with different features such as age, gender, education, occupation and others that will be used to create marketing profiles for the UVW college.

Our main focus of the data exploration and visualization is to connect between the given demographic features of the individuals and their income feature that focuses on $50K as the key number of salary.

II. Description of solution

As a team we worked on analyzing the dataset as well as creating visualizations that demonstrate the relation between different attributes column and how they have an impact on individuals income. Below are some of the analyzed attributes with their visualizations.

a. Age

As shown on figure 1, the distributions of the individuals age ranges from people less than 20 to people who are above the age of 90, and the main age group that is having an income greater than 50k are people between 30 to 50, while other age groups are mainly receiving an income less than 50k. Moreover, as shown in figure 2 the skewness of the data income is different between the two graphs. when income is less than 50k is positively skewed while when income is greater than 50k is less skewed. The age group targeted by UVW college programs are people under 30 years old.

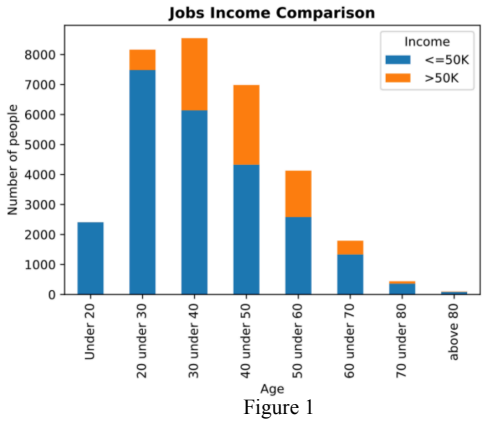

Figure 1

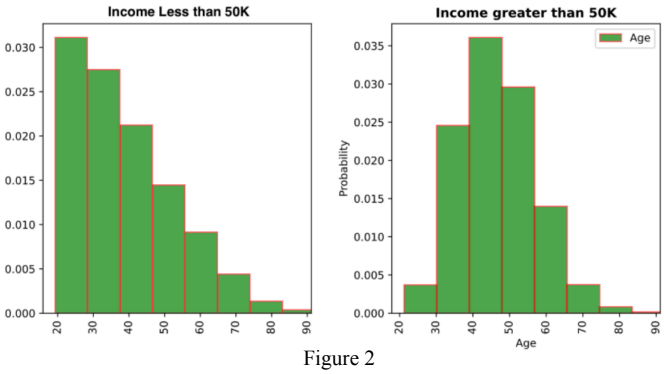

Figure 2

b. Sex

A pie chart comparison was created to compare between the male and female genders in terms of income. As shown in figure 3, 30.6% of men have an income greater than 50k. While figure 4 shows that only 10.9% of women have an income greater than 50k.

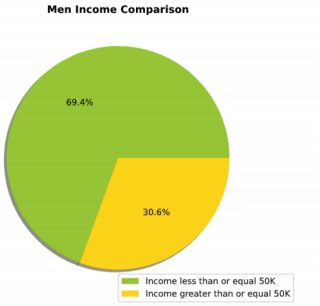

Figure 3

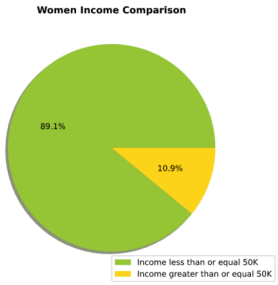

Figure 4

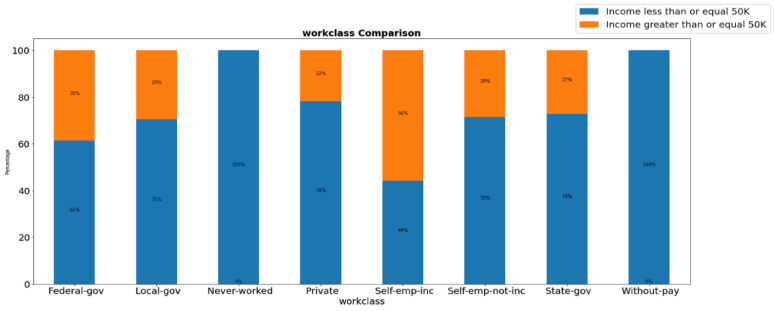

Figure 6

c. Occupation

For the occupations comparison, a stacked bar with percentages was created to compare between different individual's occupations with respect to their income. Figure 5 shows how an occupation is having a great impact on an individual's income, as we can see the top two occupations with the highest income are Exec-managerial and Prof-specialty with income percentage higher than 50k of 48% and 45% respectively.

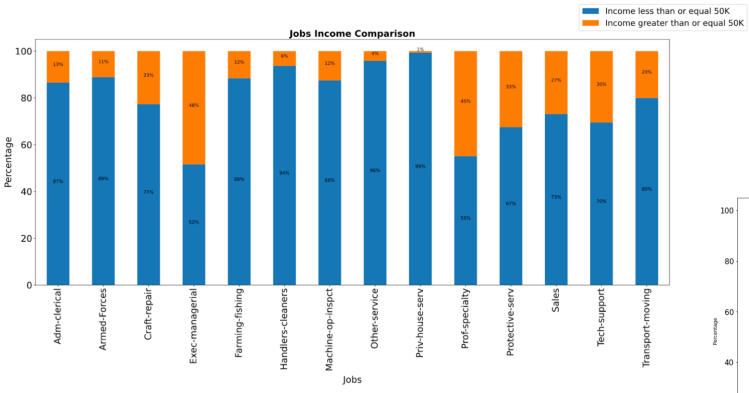

Figure 5

d. Workclass

"In the comparison between the work-class and income. Using again the same bar chart with percentage. We could see that, Never worked and working without-pay are 100% less than 50K which is very intuitive.Also, it is very intuitive that the self-emp-inc, which refers to people who work for themselves in corporate entities, has the highest percentage for having income greater than 50K. Followed by working in the federal government to be the second highest work-class having a high percentage to have an income greater than 50K. So, having Figure 6 in our visualizations is very helpful and intuitive."[1]

e. Marital status

"Comparing the marital status against the income found out to have an obvious relationship with one another. Again, following the same stacked bar with percentage as in occupation comparison. In Figure 7, the two categories Married-civ-spouse, which corresponds to a civilian spouse while married, and Married-AF-Spouse- which corresponds to Married spouse in the armed forces, are the two main categories that most probably have an income which is greater than 50K with percentage 45% and 43% respectively. This is very close to its half. On the other hand, all the other categories smaller than 10% of it is greater than 50K." [1]

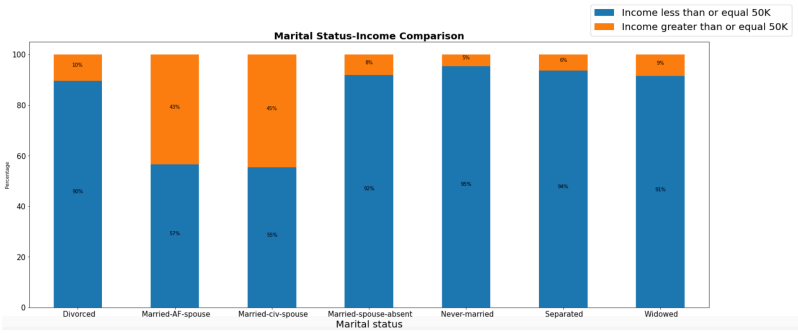

Figure 7

III. Results

As the pervious section shows the visualizations, we can reach results that shows how each attribute affects the income and how the UVW collage and others who are interested on the data can use it to generate marketing profiles. The top attributes that affects individual's income are education, occupation, age and marital status respectively. Education is the top factor that affects individual's income. UVW college can use it to reach people with lower certificates with their marketing plans.

second factor is individual's occupation, the top occupations with salary more than $50k are Exec-Managerial and Prof-specialityUVW colleges can use these job titles as investors for their programs. While they can target people with occupations like Priv-houseservice and other job services -as they have the lowest income percentages- to benefit from their programs.

Third factor is the age, the main age group that is having an income greater than 50k are people between the ages of 30 to 50, while other age groups are mainly receiving an income less than 50k. We can clearly see that people under 30 years old need more support as they have less income than others.

And there are other factors like the sex, although that gender affects the income -as the visualizations showed- and that male's income are higher than female's income, however females these days have worked on different sectors that empowered them to gain an income more than 50k and that shows that women can reach higher percentages of income over time just like men. This difference between males and females can help UVW college to target more females and empower them with their programs.

IV. Contribution

Starting form the team distributions that happened on week one, I created a slack room for the team to ease the communication between us and I asked for a meeting to get to know each other as well as to explore the needed data to start working on the project. After taking some time to explore the data, I was leading the team telling them what we need to deliver from attending the live events and taking notes about what needs to be done. Starting from week 4, we jumped into coding and creating visualizations. I was responsible for creating the visualizations of age, sex and occupations that were mentioned on the description of the solution above.

I worked on their visualizations according to their types.

After creating the visualizations, my team liked how the stacked percentages bar diagram for comparing between different categories was intuitive. So, I shared my code for them to use it for visualizing their assigned attributes as well. For the executive report presentation, I contributed on adding the occupation and age slides as well as writing needed descriptions for each slide and adding the last slide for summary. And for the System documentation report, I worked on writing the introduction, roles and responsibilities, team goal and business objectives as well as creating all the user stories based on the business objectives and visualizations. Moreover, I added my visualizations and their descriptions.

V. Lessons learned

1- The importance of taking the time needed to explore the dataset very well because it will affect the whole visualization process.

2- How to decide which chart to use to represent certain variables in the dataset, for example bar charts will be best showing comparison between different categories, while pie chart is really powerful if we want to show the percentages between a small set of categories and the histogram is best to show the skewness and distribution frequency of the data.

3- One of the best design practices is communication between the team members, communication can be a very powerful asset that helps sharing insights that might be really useful as well as taking and giving feedbacks to other people.

4- We need to take care of the visual variables when creating a visualization, such as size, shapes and mainly colors as they have a huge impact on how the data is anticipated and viewed.

5- The best way to show the relation between two variables are by drawing charts to represent it. "Without data visualization, it is challenging to identify the correlations between the relationship of independent variables. By making sense of those independent variables, we can make better business decisions. "

References

[1] Nourhan ElNaggar, Myan Sherif, Qossay Jarad, Abdullah Mohammed April 19, 2021. System documentation report

[2] Nourhan ElNaggar, Myan Sherif, Qossay Jarad, Abdullah Mohammed Mar 19, 2021. Executive Report Presentation

[3] Nourhan ElNaggar April 26, 2021. Executive Report Presentation

[4] June 10th, 2020
https://analytiks.co/importance-of-data-visualization/#:~:text=Data%20visualization%20gives%20us%20a,outliers%20within%20large%20data%20sets