Individual Contribution Report

Subhash Chandra Bose Vuppala
1219970986
svuppal6@asu.edu

REFLECTION

This project aims to develop targeted marketing profiles based on income as a key demographic in the decision-making. We used the United States Census Bureau data to look at various features that can affect a person's income, including age, sex, education, occupation, capital gain, hours per week and marital status, etc. The key target is determining if an individual has an income above or below $50,000 with relevant attributes. It is crucial to make sense out of the data to accomplish the goal. I was able to use my learnings in the coursework into application while working on this project dataset.

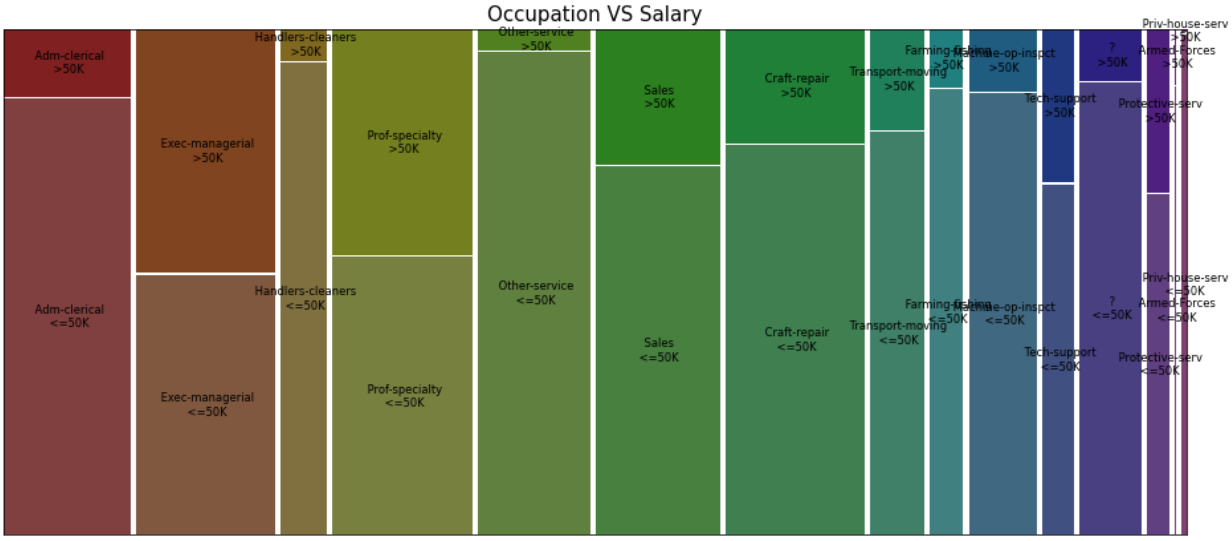
With this project, I learned the necessity of Data exploration and Visualization, which resulted in several inferences that helped solve the main task. Apart from the concepts I learned during the course and assignments, I went beyond doing my research to learn visualization tools and find patterns in the data. I started cleaning data and removing undefined values in the work class and native country column. While doing the visualizations, I got a chance to explore seaborn and plotly, which made my work easy to plot mosaic plots and count plots. Then, I worked on finding the correlations between the features to gain a better understanding.

As one of my teammates made visualizations using Tableau, I got to know about Tableau dashboards. Several inferences that I was able to develop by doing this project and some of those are discussed along with my methodology in the next section.

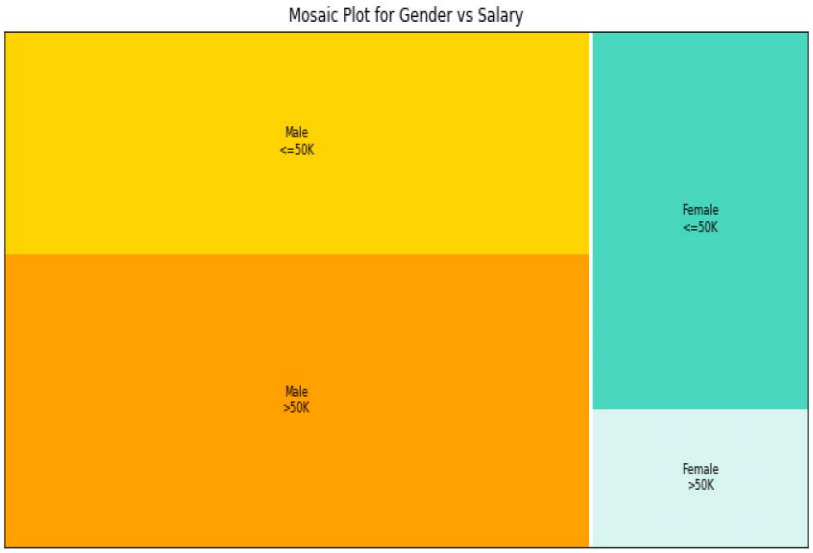
ANALYSIS

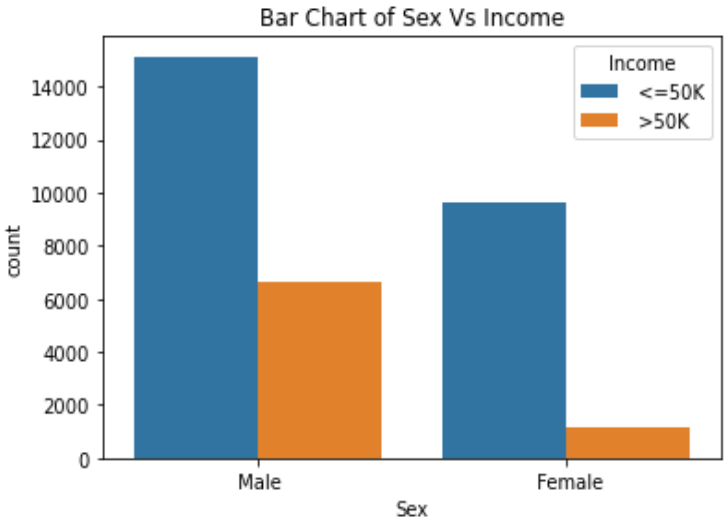
Before doing analysis, I used data preprocessing techniques to analyze the given data and clean the data, like fixing missing data. I, along with my team, used python's pandas libraries to store and manipulate the data as needed for the models. To find the features that are best predictors of income and the accuracy of the predictions, we fit the data to a neural network and found the best features as age, fnlwgt, capital gain, capital loss, and hours per week. We found the feature importance by using the Extra Trees Classifier and Select K best. The top features of the Extra Trees classifier include fnlwgt, age, capital gain, whereas Select K best includes capital-gain, educational-num, relationship, age, etc. Then, I used Spearman's feature correlation and Pearson's feature correlation to find the correlation between the features.

The use of visualizations is key to achieve the business objective. Visualizations are used to obtain knowledge and patterns from the dataset. I visualized occupation, Marital status, and sex by using a bar chart and mosaic plot. The following are the visualizations made by me

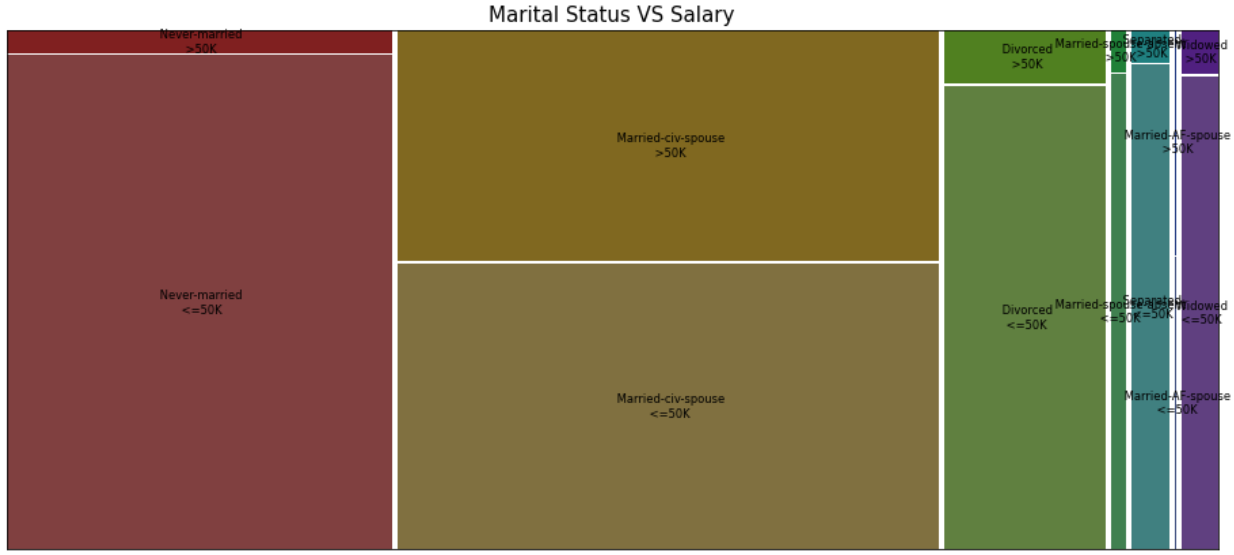


The above mosaic plot shows the relationship between occupation and salary. From the mosaic plot, I analyzed that for most categorical data, the dispersion of the two classes is exceptionally skewed, indicating that this component can be utilized to recognize among the two classes. We can also find that individuals with occupations such as “Exe-managerial” and “Prof-speciality” are more likely to earn 50K income.





The second feature which I have visualized is Sex. The above bar chart and mosaic plot shows the relationship between Sex and Income. From this visualization, I analyzed that males with income higher than 50K are high compared to females with income higher than 50K, and males with income less than 50K are high in number compared to females with income less than 50K.



The third feature which I have visualized is Marital Status. The above mosaic plot shows the relationship between marital status and Income. From the Mosaic Plot, we can conclude that Married-civ-spouse is more likely to earn more than 50K income, and never married are more likely to earn less than 50K income.

With the visualizations done by my teammates, I was able to understand the following relationships between the particular feature and Income.

1. Individuals at an older age are likely to earn a higher salary than individuals at a younger age.
2. Individuals that work a large number of hours are more likely to earn a salary greater than 50K.
3. Individuals that are Wife/Husband are far more likely to have an income of >50k.

4. People with an Income>50k are more statistically likely to have a marital status of married and present.
5. People whose age between 35 and 50 are likely to have a salary above 50K.
6. Some outliers between age 72 and 90 have Income above 50K.
7. Most of the people who work 40 to 50 hours per week have Income above 50K.
8. It was found that age, sex, education and hours-per-week have the strongest positive correlation.

These types of inferences would not be possible without appropriate visualizations and visual elements.
Some of the things that I learned from the work were:

1. A correlation matrix is used to determine the relationship between two variables and a variable with high correlation should be focused.
2. Tableau is a powerful visualization tool that converts the raw data into a very easily understandable format.

Finally, this project helped me to work on multivariate analysis, specifically on categorical data. After completing the project, I have had the opportunity to reflect on the importance of Data Visualization towards drawing meaningful conclusions from the dataset.