CSE 578: Systems Documentation Report

Team Members

Rosty Hnatyshyn

Kyle Gonzalez

Nicholas Seah

Sang-Hun Sim

Arvind Hasti

Chandana Tarur Veerabhadraiah

Roles and responsibilities

The product owner is the UVW marketing company, which is working closely with XYZ University to increase its enrollment through targeted marketing. The stakeholders are the University's staff, who hope to have more students enrolled to fund grants and other projects, and the analysts at UVW who hope to fulfill their contractual obligations with the University and bolster the enrollment at XYZ through their efforts. Each member of the team is responsible for one of the attribute visualizations. Each member of the team is also responsible for communicating their ideas with the rest of the group and providing feedback on other team members' ideas. Each team member is responsible for showing up to meetings at 2 PM every Thursday. Finally, each member of the team is expected to help with editing and polishing the final report and presentation.

Team Members

- Rosty Hnatyshyn - Clean data, analyze data, attribute 1 (marital status) visualization, write report
- Kyle Gonzalez - Build machine learning model, attribute 2 (relationship) visualization
- Nicholas Seah - Data analysis, attribute 3 (capital gain) visualizations, feature importance chart
- Sang-Hun Sim - Analyze machine learning results, attribute 4 (occupation) visualization, parallel set plot
- Arvind Hasti - Data analysis, attribute 5 (education) visualization
- Chandana Tarur Veerabhadraiah - Data analysis, attribute 6 (age) visualization

Team goals and business objective

Our goal is to help the analysts at UVW to figure out who to market to based upon their income. We need to build visualizations that clearly show the relationship between a person's attributes and income. We also need to build a machine learning model that can predict a person's income, based upon these attributes - we can use the results from the model to determine what attributes affect a person's income.

Assumptions

- The dataset needs to be cleaned before use
- The analysts at UVW need to have clear visualizations and explanations to go with them
- None of the data that we are using will be distributed to third parties without the consent of the users

User Stories

Story #1: As an analyst at UVW, I would like to know what attributes affect a person's income the most so that I could know what to look for in a potential student at XYZ University.

Story #2: As an analyst at UVW, I would like to know if a person's marital status could impact whether or not they make more than $50K a year. This would allow me to target my marketing towards people who make less than $50K a year with this attribute.

Story #3: As an analyst at UVW, I would like to know if a person's relationship status could impact whether or not they make more than $50K a year. This would allow me to target my marketing towards people who make less than $50K a year with this attribute.

Story #4: As an analyst at UVW, I would like to know if a person's capital gains could impact whether or not they make more than $50K a year. This would allow me to target my marketing towards people who make less than $50K a year with this attribute.

Story #5: As an analyst at UVW, I would like to know if a person's occupation could impact whether or not they make more than $50K a year. This would allow me to target my marketing towards people who make less than $50K a year with this attribute.

Story #6: As an analyst at UVW, I would like to know if a person's education could impact whether or not they make more than $50K a year. This would allow me to target my marketing towards people who make less than $50K a year with this attribute.

Story #7: As an analyst at UVW, I would like to know if a person's age could impact whether or not they make more than $50K a year. This would allow me to target my marketing towards people who make less than $50K a year with this attribute.

Story #8: As an analyst at UVW, I would like to be able to predict a person's income, based upon attributes found in the dataset. This would help me decide whether or not I should contact that person and try to market XYZ University to them.

Visualizations

Feature Importance: Shown in Figure 1 of the appendix, this visualization shows which attributes are most influential in determining whether or not a person makes more than $50K a year. It is not surprising that marital status is so influential, as people typically file taxes jointly - this tends to increase their income on paper. A relationship where you need to take care of a family also creates the need for more income - this fact makes it easier to separate people who make more than $50K from those who make less. Capital gains also directly correlate with

income - as, the more you make from investing, the more money you have. A person's education has been shown to highly influence their salary; people with doctorate degrees tend to make more than those with less education. A person's occupation also has a direct impact on their salary - an executive tends to make a lot more than someone who works a trade. Finally, as a person ages, they tend to get promoted and make more money than someone who just started their career at 20 years old.

Parallel Set: We used the parallel set chart to further confirm what we found in the feature importance chart - this graph clearly shows the groups of people who tend to make over 50K a year. They are usually married, educated males that work white-collar jobs. This immediately cuts out a good number of marketing targets for the analysts - they don't need to market XYZ University to people who have attended college or are working highly skilled jobs.

Marital Status: We used a mosaic plot to show the splits among categorical data shown in Figure 3 of the appendix. There are some very pure splits in this attribute group: people who are not currently married (never married, divorced, separated, or widowed) tend to make less than $50k a year. The people in these groups could be good marketing targets. The excitement of a fresh start could sound appealing to people who have lost their families.

Relationship: We used a mosaic plot to show the splits among categorical data shown in Figure 4 of the appendix. Once again, the splits in income for the relationship attribute demonstrate that unmarried people or people without families should be targeted because they overwhelmingly tend to make under $50k a year.

Capital Gains: We built two stacked bar charts to explore capital gains - one to explore the relationship between income and making no capital gains vs making some capital gains, and another to "drill down" and further explore the subset of people who make at least some capital

gains. From the graph, shown in Figure 5 of the appendix, it's clear that people who tend to have at least some kind of investment (i.e capital gains) tend to make more than 50K a year. This is especially true once the gains go over 8K a year. This makes sense as people who have money to invest probably make more than enough to be comfortable.

Occupation: We used a mosaic plot to show the splits among categorical data which is shown in Figure 6 of the appendix. White-collar workers tend to be more educated, as shown in the parallel sets plot. Therefore, it makes sense for the marketing department at UVW to avoid sending mail to people who are in this category. On the other hand, we can see that people who work in blue-collar fields overwhelmingly tend to make less than 50K - farmers, cleaners, etc. These people may be looking for a way to make more money, and therefore they may be good targets to market XYZ University to. Moreover, people with white-collar jobs are unlikely to change specializations - after having gone through college/training once, why would anyone try to do it again?

Education: We used a mosaic plot to show the splits among categorical data shown in Figure 7 of the appendix. Since we are marketing a university, it makes sense to avoid soliciting people who have already attended college, unless it's for a graduate program. Marketing a university to high school students makes a lot of sense since they tend to be the dominant demographic on college campuses, as well as targeting people who only have a high school diploma - as shown above, they tend to make less than 50K a year and the fact that college tends to increase your salary could be a major selling point for people in that group. It also made sense to consolidate people who haven’t finished high school together, as they cluttered the mosaic plot. They could be considered as one category and definitely should be marketed to as they are likely to be college students in the future.

Age: We used a split bar chart to show the distribution of people's income based on their age group, shown in figure 8 of the appendix. The group with the highest percentage of salaries over 50K tend to be around thirty to forty years old. The marketing department at UVW should avoid marketing to this group and instead focus on people in their twenties because this group has the lowest number of people making 50K a year or more. They could also consider reaching out to people in their senior years, as they also tend not to make as much money, yet have a lot of free time.

Tools

- Juypter notebook to organize the code
- Github for easier collaboration and code organization
- Python with the following libraries:
    - Matplotlib for visualizations
    - Statsmodels for mosaic plots
    - Pandas for handling data and visualizations
    - Scikit-learn for machine learning
    - Seaborn for machine learning analysis
    - Plotly for the parallel set plot

Questions

- Is there a visualization that gives an overview of the splits in the dataset?
    - A Parallel Sets plot gave us good results when trying to demonstrate the relationship between all of the attributes and the income.
- What is the best machine learning model for this dataset?

    - After trying a few different models to fit the dataset like GaussianNB, DecisionTreeClassifier, and RandomForestClassifier, it was determined that the Random Forest algorithm yielded the best accuracy.
- What visualizations work best for the categorical attributes?
    - Mosaic plots worked best to show the splits in the categorical attributes.
- What visualizations work best for the continuous attributes?
    - Binning the continuous data and placing them in stacked bar charts effectively showed the splits in the binned groups.

Further work and future goals

- Increase accuracy of machine learning model
- Build interactive visualizations
- Investigate less important features to see if anything was missed
- Build visualizations exploring relationships between different attributes instead of only focusing on attributes vs income

Appendix

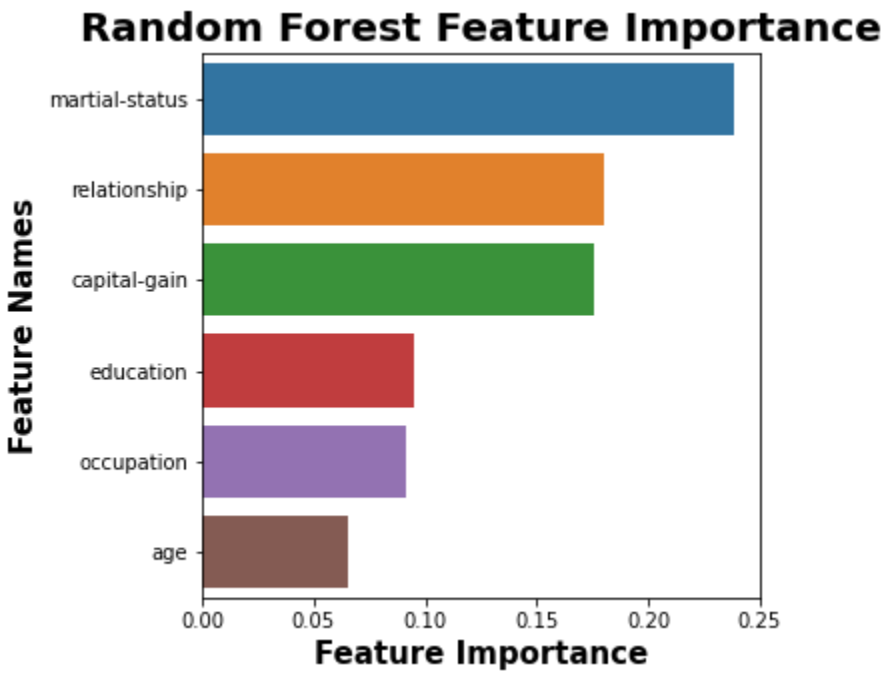

Figure 1: Feature Importance Chart

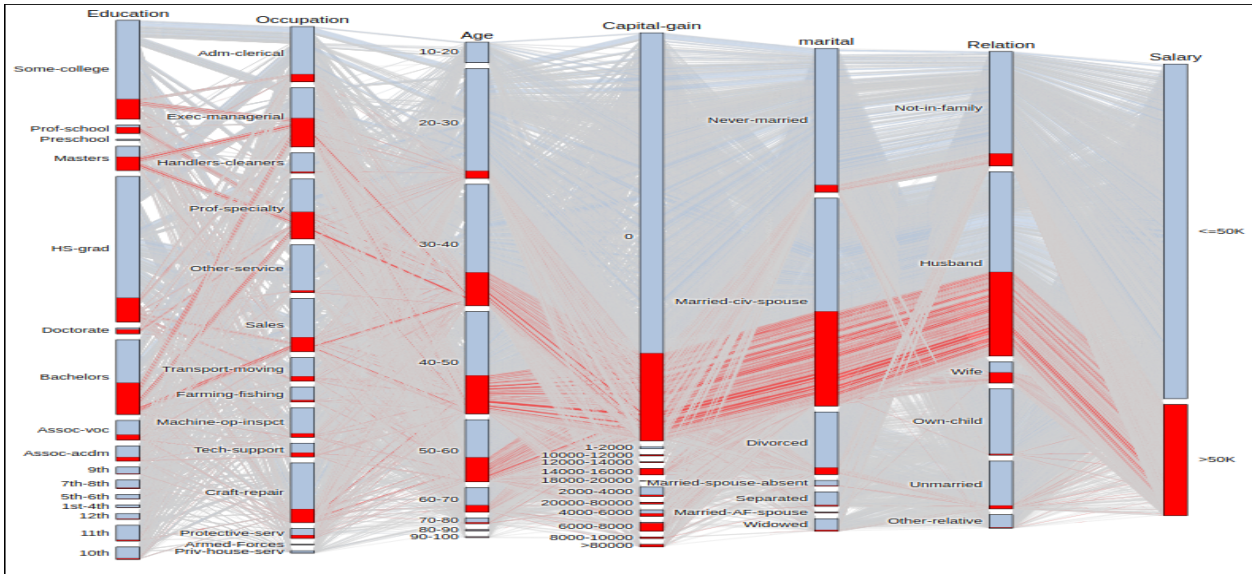

Figure 2: Parallel Set Plot

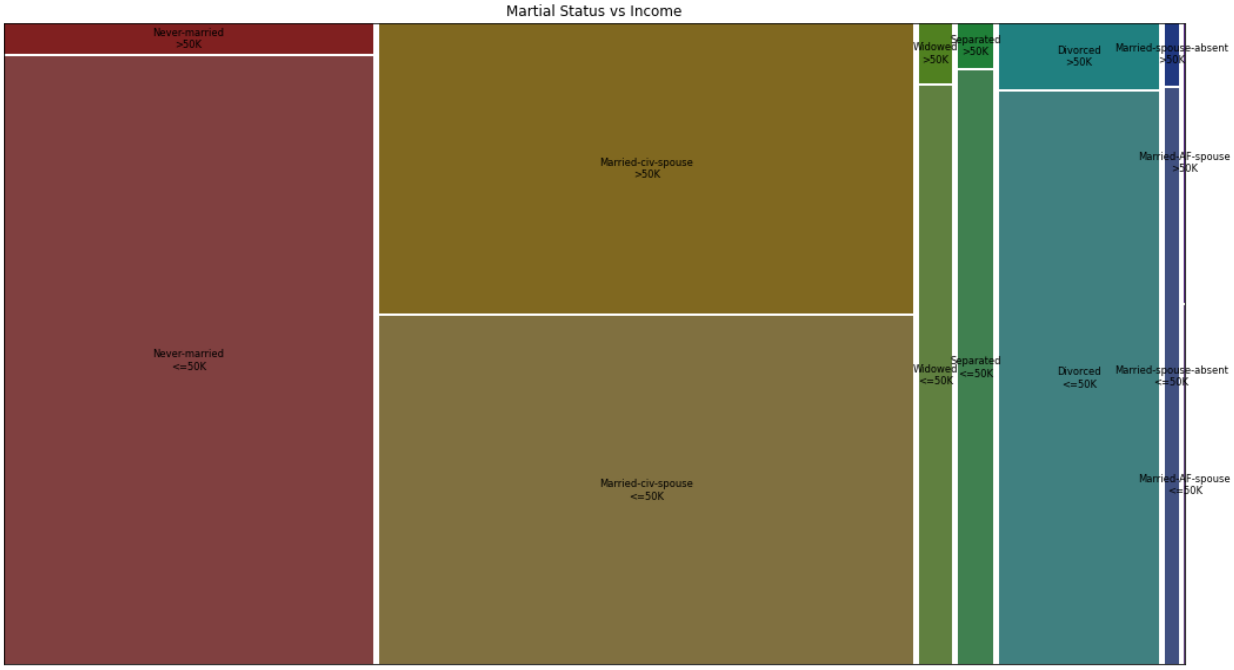

Figure 3: Marital Status Mosaic Plot

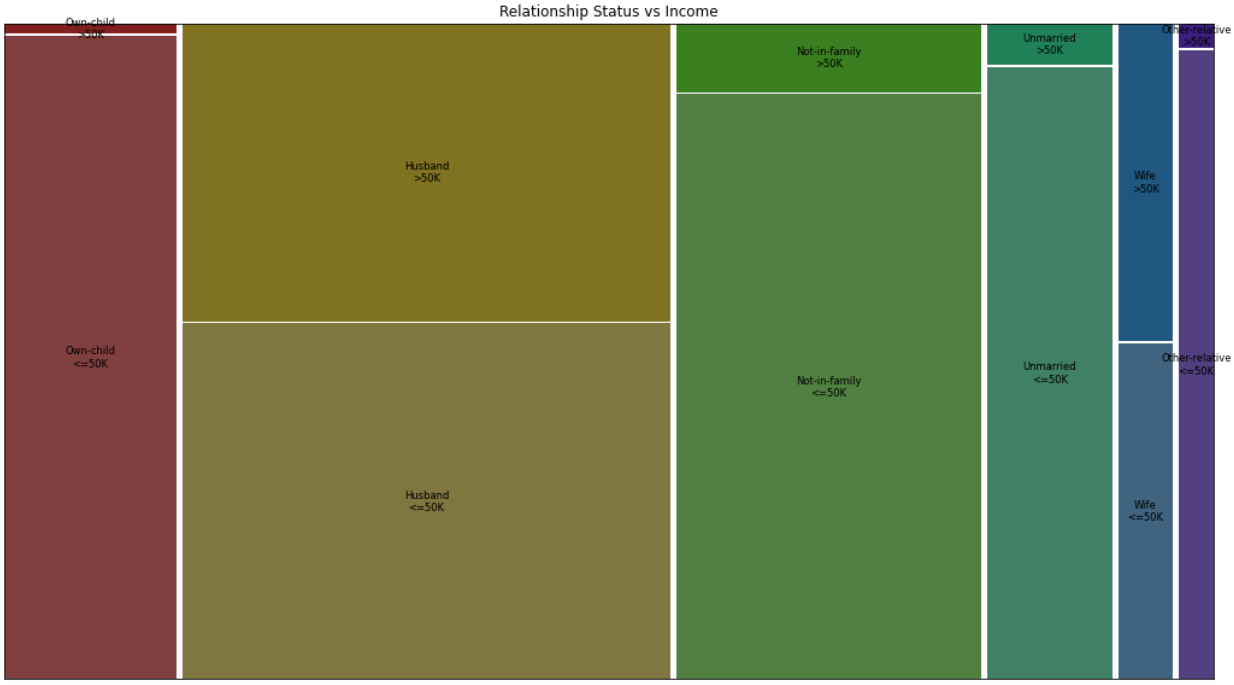

Figure 4: Relationship Mosaic Plot

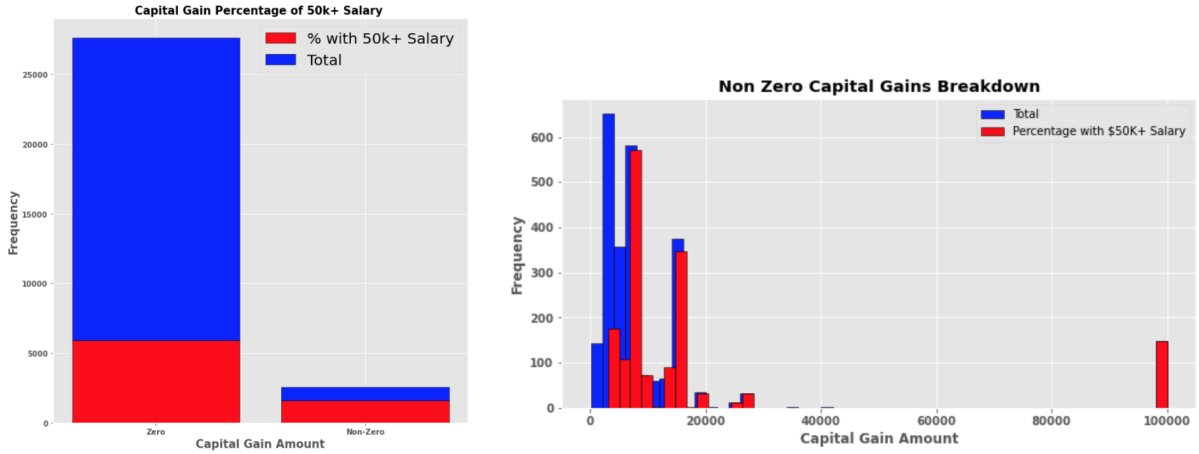

Figure 5: Capital Gains Stacked Bar Charts

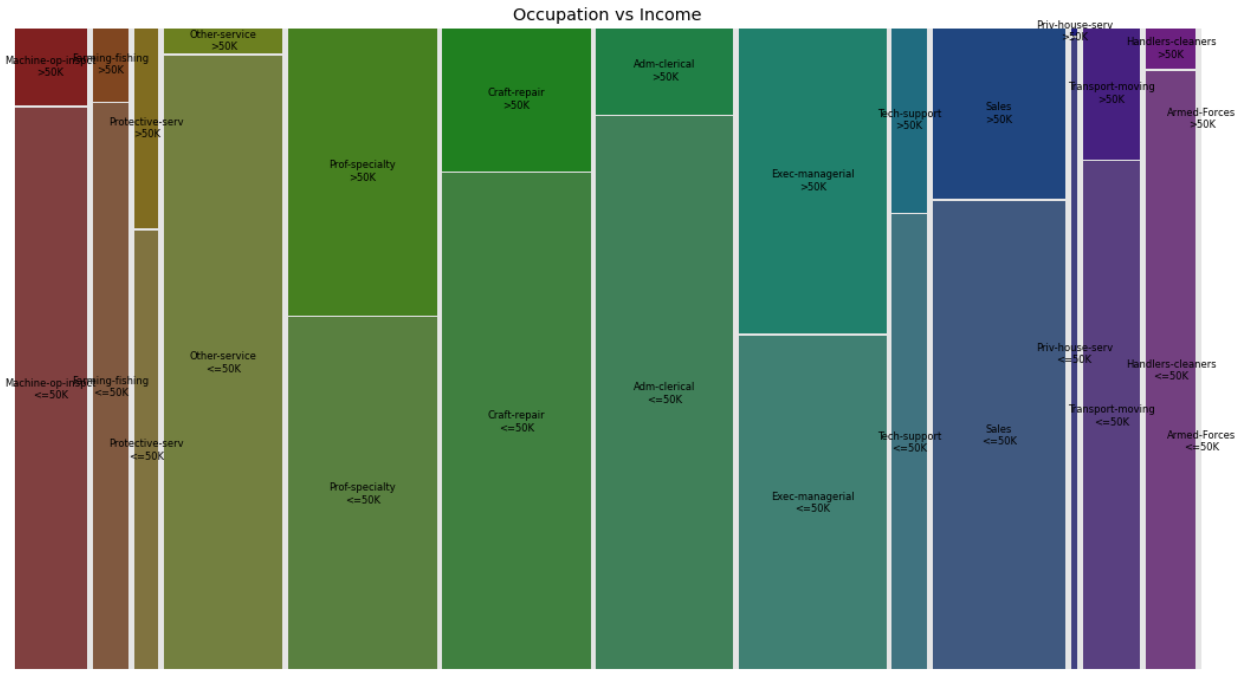

Figure 6: Occupation Mosaic Plot

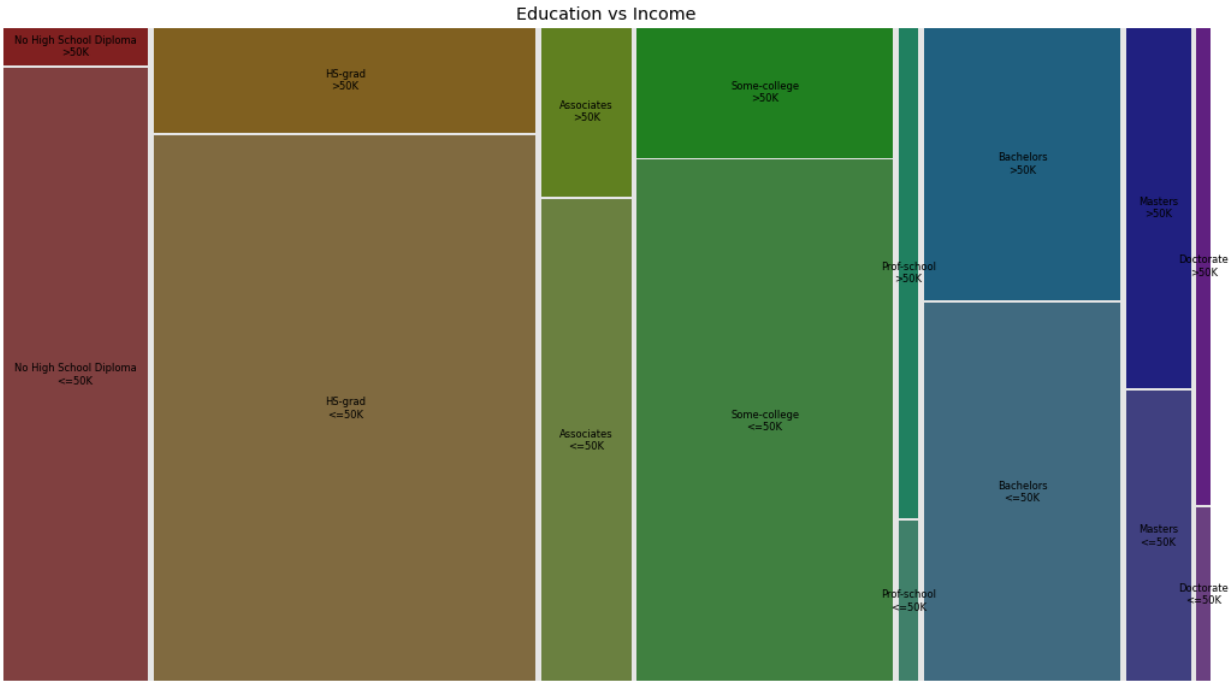

Figure 7: Education Mosaic Plot

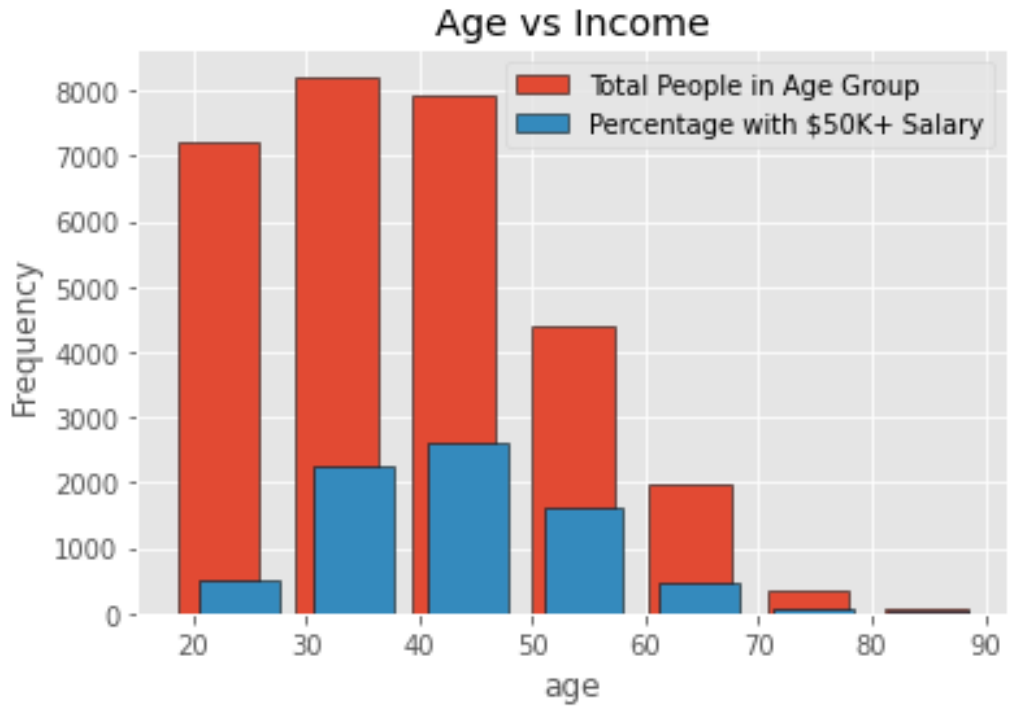

Figure 8: Age Stacked Bar Chart