Course Project Final Report for CSE 578: Data Visualization

Maclay Teefey

I. Introduction

THIS project aims to use census data to understand the demographic factors that most impact income in order to better market UVW College and improve their enrollment. Using univariate and multivariate analysis of the data, this project showcases five user scenarios encompassing eight census categories to understood the applications of the data analysis. This project focuses on the eight census categories with the highest correlation to income: Marital Status, Education, Age, Hours Worked Per Week, Sex, Occupation, Relationship, and Capital Gain. For each user story, I have created visualizations of the relationship between income and one or more census categories.

A. Preparing the Data

As part of my exploration process, I visualized the untransformed census data as both table and distribution charts of the features. When I visualized the fnlwgt feature, there was no clear pattern for how the fnlwgt feature relates to the other data that were relatively self-explanatory. After reviewing the documentation, I understood that the feature is calculated to weight various socio-economic characteristics of the census takers to group similar individuals by their socio-economic characteristics. Then I calculated the correlation of the final weight feature with the income feature, and the fnlwgt feature had a correlation value of 0.009463. This correlation value is less than the five other numeric features in the data set, with the second-lowest correlation value being 0.150526 for the capital-loss data. Due to the low correlation with the income, and the relatively unintuitiveness of the feature compared to the other categories, I have decided to remove the feature from the cleaned data set.

The raw census data includes many data entries with the value set as a question mark. To decide how to handle the entries without data, I have analyzed how many census takers have incomplete data, and 7.37% of the census population have at least one entry with no data (2399 census takers out of 32561 total). The only features with no data are the occupation, native-country, and workclass columns with 1,836 rows have both workclass and occupation set to a question mark. Because of the large percentage of individuals with at least one entry with no value, I decided to not remove the rows with no value entries. I assume that there are multiple factors that could cause someone to not display a value for each census categories, so I replaced the '?' entries with "No Value".

To better understand which categories factor the most into the income of the census takers, I calculated the correlation matrix of every factor versus income. To accomplish this I transformed both the numeric and categorical features. Many numeric features have non-normal distribution of values. For example, over 45% of all the census takers stated that they work 40 hours per week, with the second most popular choice being 50 hours per week at 8.5% of the population. Uneven data distributions can result in the prediction and correlation techniques not being applied accurately across the data set. Therefore, for all numeric features, I applied a scaling technique. For categorical features, the data need to be transformed into numeric values in order for both prediction and correlation techniques to understand the features. I have decided to represent each category using a one-hot distribution, because many of the categorical features do not have an intuitive ordering. For example, the relationship feature has values Wife, Own-child, Husband, Not-in-family, Other-relative, and Unmarried. If I transformed the feature to correspond to an indexing of the list of values, the 'Wife' value would be interpreted by correlation and prediction techniques to be more similar to the 'Own-child' value than the 'Unmarried'.

After transforming the data, I have calculated the correlation between all the features and the income rate, with the highest absolute correlation being the marital status of 'Married-civ-spouse' (civilian spouse) at 0.439802. After analyzing the correlation matrix, I decided upon using the following factors to for my user stories: Marital Status, Education, Age, Hours Worked Per Week, Sex, Occupation, Relationship, and Capital Gain.

II. User Stories

A. User Story 1

The marketing director is curious whether sex and marital status have any bearing on income. Due to the few categories for sex and marital status, I decided to use a mosaic plot to understand the distribution of census data and how each large category differs in percentage of individuals making over $50,000. When creating the mosaic plot, I decided to condense many of the values in marital status into three categories: 'Married', 'Separated', and 'Never Married'. This is because many of the categories are more understandable and fit better into one category. For example, couples who are married that have one or more of them in the armed forces of the United States were counted in a separate category as civilian marriages. Using

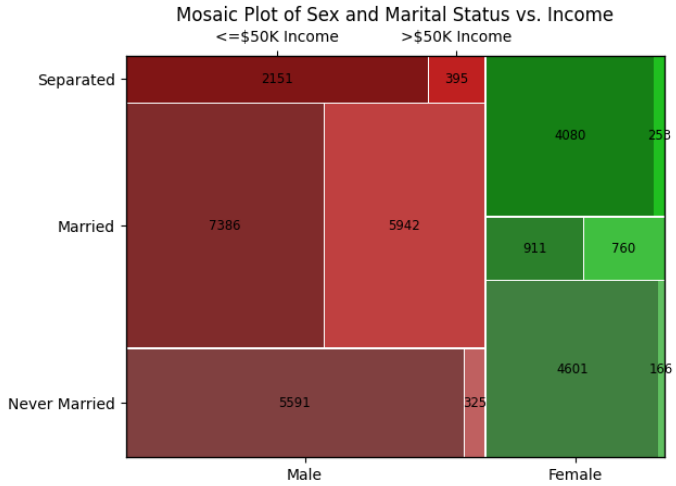

Fig. 1. Mosaic Plot of Sex and Marital Status vs. Income for User Story 1

the statmodels mosaic plot, the figure included text in each box describing what categories the box corresponded to. This created issues analyzing the box, because the text couldn't fit into most boxes with the 'Separated' and 'Never Married' boxes overlapping the most. In order to fix this, I originally removed all the text inside the boxes, but I found that understanding the exact percentages for each category was very difficult. I decided that displaying the counts for each category and only the counts for each category was the most useful text to include inside of the boxes. Displaying the counts for each box eliminates the need to understand the box in the context of the line splits and limited the amount of text that would display outside of the boxes. After calculating and displaying the count for each box, marketers can more easily understand which box is bigger than the other by seeing which box had the bigger number. When analyzing the chart, it is easy to see that the individuals currently married are more likely to have an income of over $50,000 than separated and never married couples. The chart displays that men on average are more likely to have above $50,000 in income for each marital status. These two data points can allow the marketing team at UVW college to more effectively market financial aid to groups like unmarried women who are more likely to make $50,000 or less.

B. User Story 2

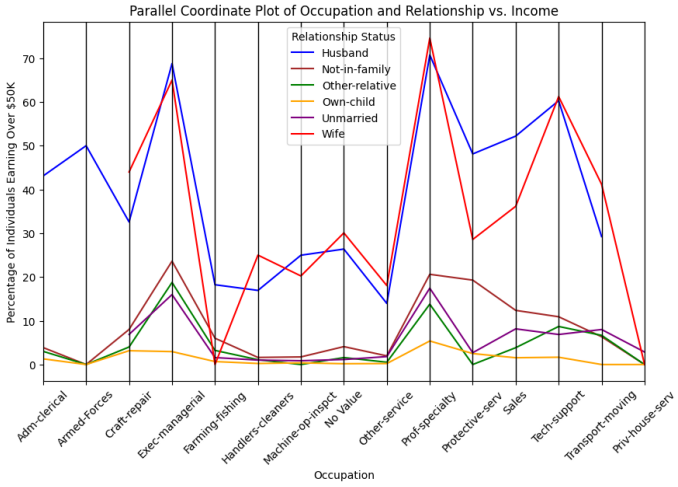

Fig. 2. Coordinate Plot of Occupation and Relationship vs. Income for User Story 2

As a member of the UVW marketing team, I want to understand how relationship status and occupation impact the income rate of individuals. I have decided to investigate these two variables due to the relationship data including both sex and marriage status in the categorical values 'Husband' and 'Wife'. The contrast between the income levels of the husband and wife values can show the differences in money offered in various occupation based on sex and can show which fields have the largest differences between being married and unmarried. To illustrate these differences best, I decided to use a parallel coordinate

plot, which can quickly show contrasts between two separate categorical variables with different values. When creating the chart, I decided to use a qualitative color scheme to illustrate that there is no inherent ordering between the categorical values. However, I decided to set the 'Husband' and 'Wife' lines in contrasting colors, red and blue, in order to clearly see the difference between those two lines. This chart shows the impact of marriage on income levels, because the lines for wife and husband remain above the other lines in the chart except for when someone who is the wife of someone works in the farming and fishing industry. This puts them below every other line for that occupation. The 'Husband' and 'Wife' lines illustrate occupations such as farming, protective services, and sales, which have large percentage gaps of individuals below $50,000 income.

C. User Story 3

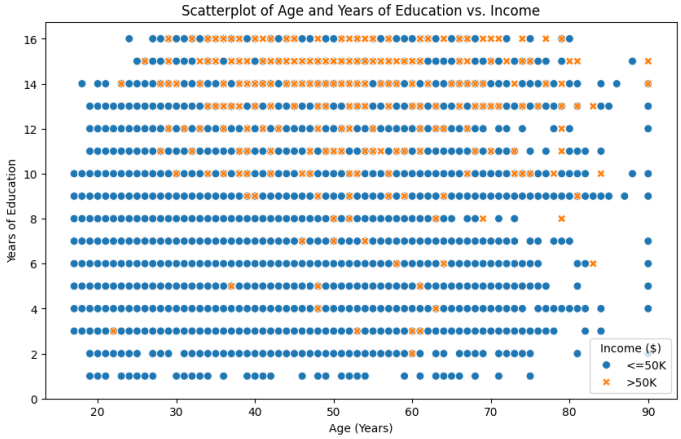

Fig. 3. Scatter Plot of Age and Years of Education vs. Income for User Story 3

Due to their correlation between eachother, I want to understand whether age or years of education impact income more. For example, individuals that are 15 mathematically cannot have over 15 years of education, so the best way of disentangling these factors is to plot them together. I decided to use a scatterplot for my visualization, because the large number of individuals for each value of age and years of education show trends well in a scatterplot. Because the data is displayed in a scatterplot, I decided to display each point using income for both color and point type. Because less than three quarters of individuals make less than $50,000 in this data set, I decided to display each instance of having above the $50,000 threshold using both a different color and shape than the other data points. Due to the large amount of data made up from each point, I originally tried to use percentages to show one value for each point on the graph. However, the data seemed to be more confusing and failed to eliminate the issues with the large number of data points. I also removed the grid lines for this visualization, because marketing agents would be able to see the correlation easier without having the grid lines take up more space on the chart. The scatterplot shows that the age of an individual has less of a factor in determing the individual's income than how many years of education a person has taken. Individuals with less than eight years of education are less likely to make over $50,000, with the highest concentration of individuals with over a $50,000 income at over 14 years of education.

D. User Story 4

As a member of the UVW marketing team, I want to know how the amount of capital gain an individual reports correlates with their income and whether they make over $50,000 a year. In order to understand if the capital gain amount impacts the users, I had to understand the distribution of the data. The vast majority of individuals in the dataset (88.2%) had $0 in capital gain, with the lowest non-zero value set to $114. The highest non-zero capital gain value is $99,999, which I assume to be the highest possible value input, because multiple users set their capital gain to be exactly $99,999. I decided to use a univariate distribution for this data due to the large population of individuals having a single value for the category. With most categories having multiple categorical values, it would not be useful to illustrate how most categories would have the majority of individuals with a categorical value having a capital gain value of $0. Both bar charts and line charts could show the varying percentages at each level of capital gain grouping. However, the high percentage of individuals having 0 as a value, and users having large gaps between capital gain values make line graphs not as useful in showing off the difference between users with $0 in capital gains and various groups of capital gain amounts. When creating the bar chart, I split the groups of capital gain values into round numbers to make it easier for marketing agents to understand the general trend of the data, and to illustrate that there is not enough data at non-round value to be confident where one group stops and another group begins. To describe how income differs across each group, I decided to use the percentage of the group's individuals that have an

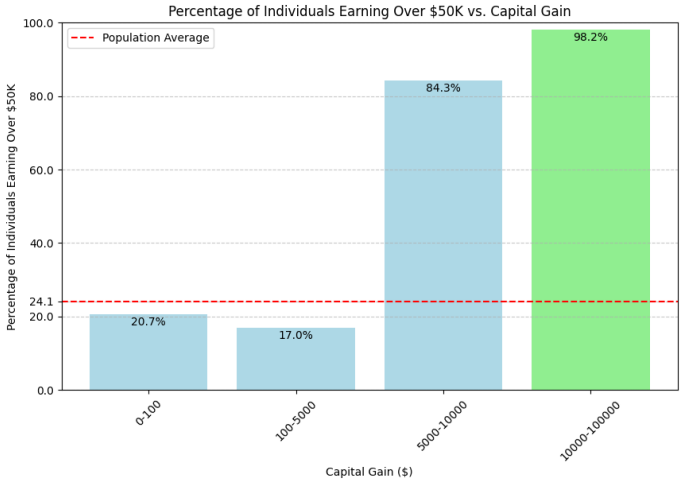

Fig. 4. Bar Chart of Income vs. Capital Gain for User Story 4

income of over $50,000 and plotted the percentage of the whole data set that has an income of over $50,000. The line of average percentage allows the UVW marketing team to quickly assess which groups make less money than other groups. The visualization illustrates that the users who do not earn capital gains are below the average income level, and that individuals that earn over $5,000 in capital gains are well above the average population in income level.

E. User Story 5

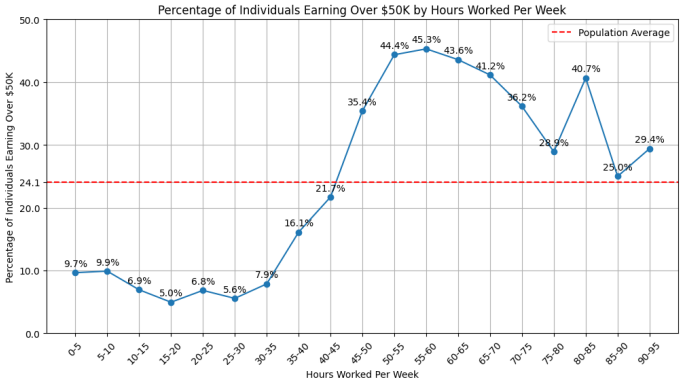

Fig. 5. Line Chart of Income vs. Hours Worked Per Week for User Story 5

When marketing to individuals who work, I want to know how the amount of hours the person works per week affects their income level. In order to decide how I should visualize the relationship between income and hours per week worked, I had to figure out the distribution of hours worked and if it would be useful to analyze it with other variables. The hours per week distribution roughly resembles a normal distribution with a large peak of individuals working 40 hours a week (45.8%) and the next three highest reported hours worked including 45 and 35 hours. Although the hours per week worked is described as being continuous, most individuals reported their hours worked in five hour intervals with the minimum hours worked at 0 hours worked, and the maximum hours worked at 95 hours worked. I assume that the 0 hours worked means that the individual is unemployed, and the 95 hours worked is the maximum hours individuals could report because it is the highest hours individuals could input while being below 100 hours worked. Due to users having a small range of hours worked over a roughly normal distribution, I decided to use a line chart to represent the changing percent of users with over $50,000 income across the hours individuals could work. To represent the groups of five hours individuals would report for their hours worked, I created bins of five hours counting up from 0 hours to 95 hours. To quickly determine which bins were below the population average, I displayed a line at 24.1%, the percentage of individuals across the data set that reported an income of over $50,000. The line graph shows that working 45 hours or more correlated well with an above average income level, with the line graph crossing the population average at the 45-hour mark. The line graph also illustrates that individuals that either are underworked (less than 35 hours worked per week) and overworked (over 70 hours worked per week) are more likely to not make $50,000 in income than someone who works between 50 and 60 hours a week. The marketing team at UVW College can use this

visualization and how many hours a person works per week to determine the people that would need and be receptive of scholarships and grants to alleviate college tuition.

III. Future Work

As part of the project, I decided to not create a machine learning model to best predict the income level of census takers with census data. In the future, I would like to assess which models are the most accurate in their predictions and use models that use decision trees to help assess the most impact features in relation to the census taker's income. When I created the correlation matrix, I saw that one of the highest correlation coefficients was 'relationship_Husband'. After discovering how high of a factor being male and married was, I attempted to understand which pairs of factors had the most impact. Due to time restraints, I decided against persuing down this path, but in the future, I would want to analyze more pairs of factors and understand which pairs have the most impact on income. After seeing the large difference between having over $5,000 in capital gains, I believe that it would be useful to visualize more census categories to how they correlate with income.