CSE 578: Data Visualization Project

Author: Jerry Barboza

***Abstract*– The purpose of this report is to find factors on the data that determine an individual's income by creating data visualizations. The data visualizations I will be creating will help us understand any correlations among individuals earning under $50,000. This will then assist UVW in marketing their degree programs. This report will contain at least 8 features across 5 user stories that I will be prioritizing based on the selected attributes.**

I. Goals and Business Objective

The main goal for this project is to find any correlations between attributes that make a good prediction on having a salary over $50K and what causes people have a salary of $50K or less. Before creating these data visualizations that will help UVW's marketing, I will first need to clean the data. I noticed that there are some rows that contain the value '?'; therefore, I must remove these rows so my dataset can be more accurate and clean for my data visualizations. I also cleaned the data by adding names for the columns and updating the cleaned dataset to a new csv file that automatically saved to my desktop. Once the data was cleaned, the business objective was to create five stories with data visualizations that would help me find any correlations with these attributes and salary.

II. Assumptions

Before doing any data analysis and creating data visualizations, I have few assumptions about this data set. Since we have the data containing the education, I would have an educated guess about college graduates are more likely to have a salary above $50K verses someone who is not a college graduate. Also based on the education level someone has, it will impact on the salary. Hence someone with a master's or Doctorate degree will have a higher probability of earning a salary above $50K vs someone with only a bachelor's degree. Therefore, as a member of the UVW marketing team, I can market the degrees to students just starting their college studies or even high-school students who are thinking about attending college but are still not sure if college is for them.

III. User Stories

a. Story #1: Correlation Matrix

I created a Correlation Matrix using the columns: 'age', 'fnlwft', 'education_num', 'captital_gain', 'hours_per_hour', and 'salary', where I then was able to understand if there were any correlations between these columns. For the salary to be part of the correlation matrix, I made the salary '>50K' have a value of 1 and '<=50K' have a value of 0. The strongest correlation I was able to find was between salary and number of educations with a positive correlation of 0.335. Therefore, the higher the education number, the more likely the person will have a salary greater than $50K. The second strongest correlation was between salary and hours worked per week with a positive correlation of 0.229. Age and capital gain also had some correlation with salary since the older the person is, the more likely they will have a salary above $50K and having a good capital gain are more likely to have similar outcomes. We can also see that most of these attributes have a weak correlation which means that it might not be statistically significant to count them towards our UVW marketing program. As a member of the UVW marketing, and after

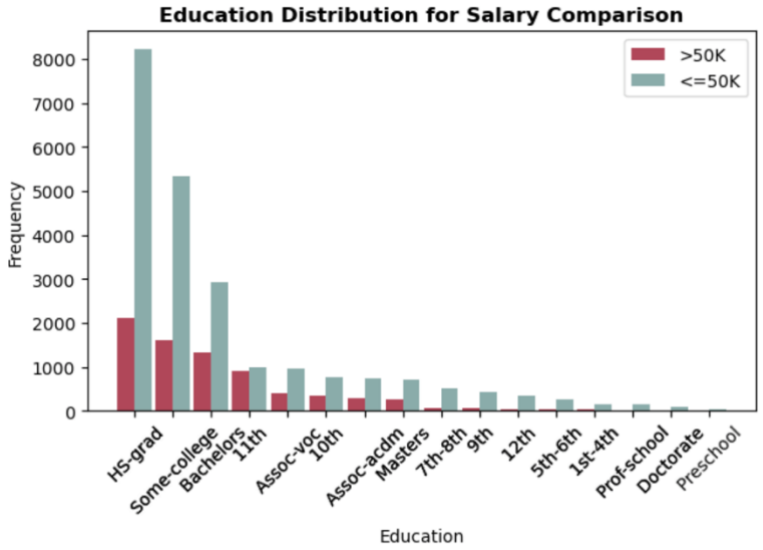
finding a moderate correlation between education_num and salary, it will be important to use these findings for the marketing of the college programs.

Correlation Matrix

	age	fnlwft	education_num	capital_gain	captital_loss	hours_per_week	salary
age	1.000	-0.077	0.044	0.080	0.060	0.102	0.242
fnlwft	-0.077	1.000	-0.045	0.000	-0.010	-0.023	-0.009
education_num	0.044	-0.045	1.000	0.124	0.080	0.153	0.335
capital_gain	0.080	0.000	0.124	1.000	-0.032	0.080	0.221
captital_loss	0.060	-0.010	0.080	-0.032	1.000	0.052	0.150
hours_per_week	0.102	-0.023	0.153	0.080	0.052	1.000	0.229
salary	0.242	-0.009	0.335	0.221	0.150	0.229	1.000

b. Story #2: Bar Plots for Education

For this story I wanted to see if there is a correlation within education and salary, and as I assumed, there was a correlation between these two attributes. I created a two bar charts side by side where we can see a comparison between salaries and education for both salaries ">$50K" and "<=$50K". On the right of the bar plot, I included the ratios, where a higher ratio represents the higher the probability of having a salary over 50K. As expected, professional school, doctorate and masters have the highest rations which means that higher education contributes to having a salary over 50K. People who didn't graduated high school are more likely to earn under 50K.

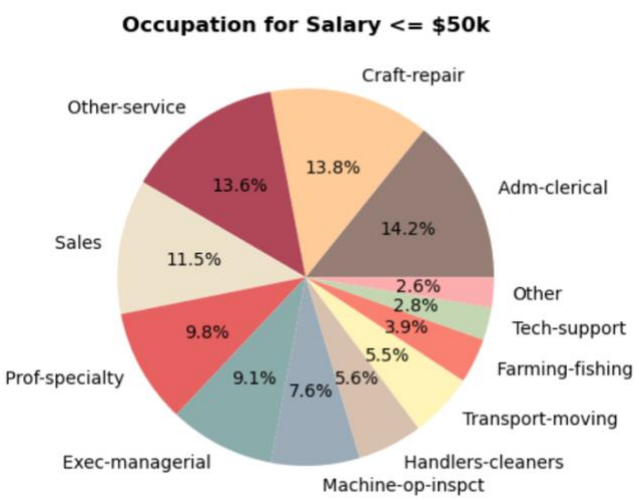


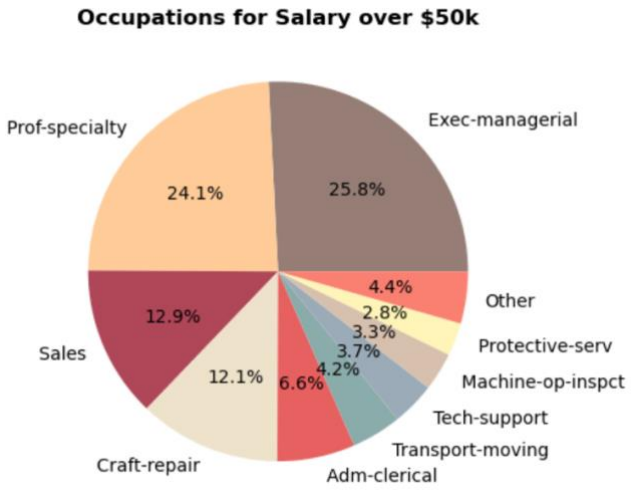

```
The higher the ratio, the more likely the salary will be >50K:

Prof-school: 2.99
Doctorate: 2.95
Masters: 1.29
Bachelors: 0.73
Assoc-voc: 0.36
Assoc-acdm: 0.34
Some-college: 0.25
HS-grad: 0.20
12th: 0.08
10th: 0.08
7th-8th: 0.07
11th: 0.06
9th: 0.06
5th-6th: 0.04
1st-4th: 0.04
Preschool: 0.00
```

c. Story #3: Pie Charts for Occupations

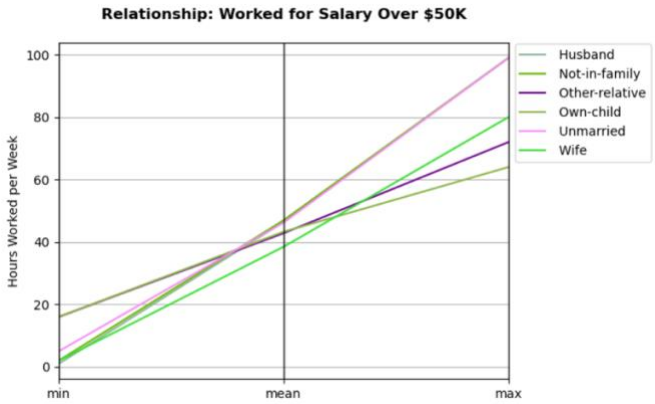
For this story I created two pie charts: one for occupation with a salary of $50K or less, and the 2nd pie chart for occupations with a salary over $50K. From the pie chart "Occupations for Salary over $50K" we can see that 49.9% of the occupations are from professional specialty and execution managerial. There was a smaller percentage on people doing these two occupations and earning $50K or less a year; therefore, we can assume that people following these two occupations are more likely to have a salary of over $50K versus the other occupations.





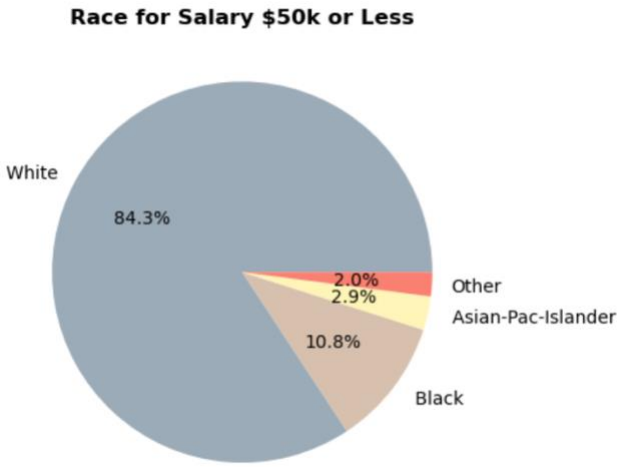
d. Story #4: Parallel Coordinate Plot: Relationship and hours worked per week.

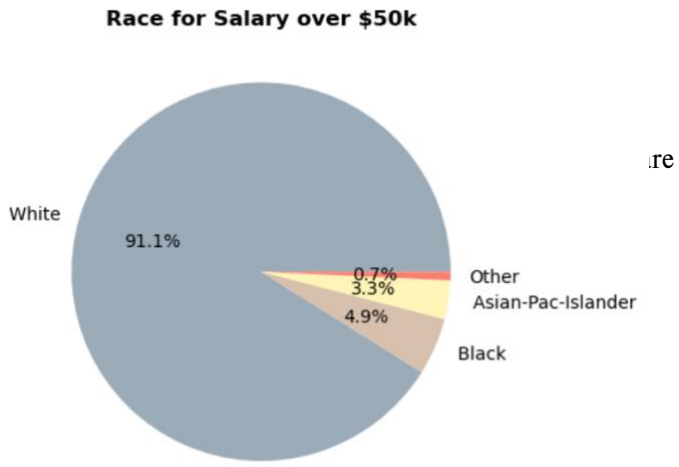
For this story, I decided to create a parallel coordinate plot to see if there was a correlation on relationship versus hours worked per week and see how they compare with each other for both people making $50K or less and people making over $50K. As shown on the left parallel coordinate plot, you can see that the plots look very similar on the categories of the relationship. All six relationships had a very similar minimum, average, and maxed hours worked per week for salaries $50K or less. We can see that people who have relationship of owning child have the lowest average of hours worked per week comparing the other relationships but tends to have similar minimum and maximum as the rest of the relationship categories. On the other hand, on the right parallel coordinate plot, we can see that the relationship categories are more distinct from each other when the person earns over $50K. Someone who is a wife and earn over $50K, have the lowest mean of hours work from the other 5 relationship categories, making the mean just under 40 hours per week. Someone who is a husband, not in a family or is unmarried are more likely to work more hours and also some of them even working almost 100 hours per week. I was able to find these findings too on the parallel coordinates plot for people earning $50K or less since for all 6 categories from relationship, we were able to find some from every category working maximum of close to 100 hours per week, while people who make over $50K, I found out that a wife, other-relative, and own-child have a lower max hours worked per week.



e. Story #5: Pie Chart for Race

To explore income distribution by race, I created two pie charts comparing the racial breakdown of individuals earning ≤$50K versus >$50K. While the dataset is predominantly composed of individuals identifying as white, the proportion of white individuals increases from 84.3% in the lower income group to 91.1% in the higher income group. Meanwhile, the representation of Black individuals drops from 10.8% to 4.9%, and those identifying as Other or Asian Pacific Islander also decrease.





IV. Problems Encountered and Solutions

One of the main issues I encountered was getting outliers since some of these outliers were too extreme that I wasn't sure if they were typo mistakes from the data entry. Therefore, I was thinking to remove some of these extreme outliers from the data. Another problem I encountered at first was being able to use sqlite3 library for the csv file, however at the end I ended up just using python as the main and only programming language for this project. I also had some trouble finding a moderate correlation between some of these attributes using a correlation matrix as shown on the image below:

	age	fnlwft	education-num	capital-gain	captital-loss	hours-per-week
age	1.000	-0.077	0.037	0.078	0.058	0.069
fnlwft	-0.077	1.000	-0.043	0.000	-0.010	-0.019
education-num	0.037	-0.043	1.000	0.123	0.080	0.148
capital-gain	0.078	0.000	0.123	1.000	-0.032	0.078
captital-loss	0.058	-0.010	0.080	-0.032	1.000	0.054
hours-per-week	0.069	-0.019	0.148	0.078	0.054	1.000

The salary wasn't part of the correlation matrix until I ended up making the salary of '<=50K' have a value of 0, and '50K' with a value of 1, and was able to find a moderate correlation and few stronger correlations from previous correlation matrix but were still weak to be a significant finding for this project. The updated correlation matrix was mentioned in story #1.

V. Not Doing

Some things I am not doing now but plan to do in the future include applying more statistical analysis to the dataset. This will help increase the marketing at UVW. While my five stories do help with UVW's marketing of their degree programs, there are still more stories to include that will provide additional help in improving UVW's marketing.

REFERRENCES

Zach (2021) *What is considered to be a 'weak' correlation?*, *Statology*. Available at: https://www.statology.org/what-is-a-weak-correlation/ (Accessed: 02 July 2023).