Data Visualization Course Project

Nishal Chandra Reddy, 1226003456, nchand28@asu.edu

I. Goals and Business objective

In my capacity as a data analyst at XYZ Corporation, I work in conjunction with UVW College to conduct an analysis of a dataset that includes demographic and financial information from the US Census Bureau. The goal is to identify target consumers for UVW College's marketing courses depending on income level, allowing for the development of targeted marketing strategies. To do this, I will choose key criteria and build effective visualizations to better understand the characteristics of these target demographics.

This research will assist UVW College in customizing its marketing strategies for successfully advertising its programs and courses to a wide variety of demographic and income groups. This procedure will take into account key characteristics such as age, race, education level, capital gain, weekly working hours, sex, education, and work class.

II. Assumptions

1) The dataset is believed to be collected properly and reliably, giving a solid foundation for analysis and it is expected that the dataset is representative of the population being described, allowing for reasonable generalizations.
2) The variables in the dataset are considered to be correctly specified and labeled, which allows for a clear grasp of their meanings and ensures consistency in analysis.
3) The dataset is presumed to be largely error-free, meeting ethical and legal requirements for data collection and use, though some additional cleaning and preparation may be required.
4) It is expected that any modifications or preparation processes performed on the dataset are suitable and do not create bias or distort the original data.

III. User Stories

1) One of my jobs as a data analyst at UVW's XYZ is to look at the age distribution of those earning $50,000 or less and in excess of $50,000. The goal of this study is to identify prospective age groups that UVW College should target in their marketing activities.
2) As a data analyst at XYZ for UVW, it is my obligation to investigate how incomes in various racial communities—both under and beyond $50,000—are divided. This inquiry will help us better understand the demographic mix of future UVW course audiences.
3) My work at UVW as a data analyst at XYZ is to investigate the capital gain and educational backgrounds of persons earning over $50,000 a year and less than or equal to $50,000. This study would make it simpler to market various courses based on these economic groups' educational attainment.
4) Understanding the relationship between years of schooling, marital status, and hours worked per week for persons earning less than $50,000 annually is one of my primary responsibilities as a data analyst at XYZ for UVW. With the help of this survey, UVW will determine who the target audience for its financial aid initiatives is.
5) My job at UVW as a data analyst at XYZ is to create detailed student profiles and improve comprehension of target group demographics by presenting data on age, employment class, and weekly work hours for those making $50,000 or less.
6) In my capacity as a data analyst at XYZ for UVW, my job is to present the gender distribution by occupation in the pay ranges more than and less than $50,000 in order to provide insight into the demographics of potential students.
7) As a data analyst at XYZ for UVW, I analyze the association between capital gain, weekly labor hours, and school years to inform the marketing team's decision-making process.

IV. Visualizations

A. Histogram

The age distribution of UVW's target audience was revealed through the histogram analysis of User Story 1. It was shown that younger people, especially those in the 20–45 age range, are more likely to earn less than $50,000, while those who make more than $50,000 are mostly in the 35–55 age range. This implies that middle-aged people have a greater chance of earning more money. Furthermore, the study revealed a lower representation of people under 20 and over 60 in the dataset, suggesting a lesser prevalence of these age groups in UVW's intended student body for marketing courses. All things considered, this data will assist UVW's marketing staff in determining the ideal age range to target with their offerings.

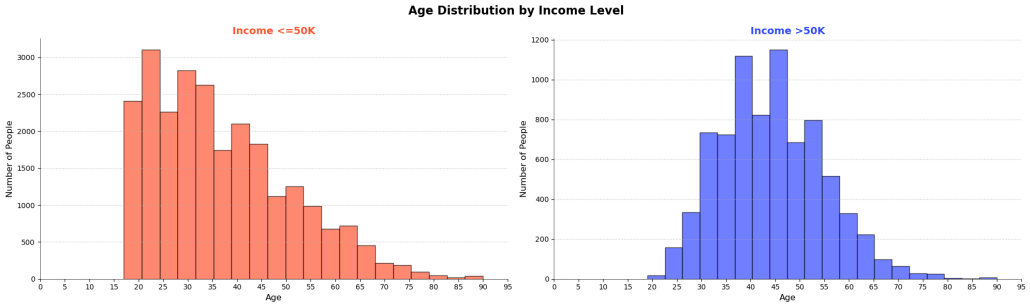

Fig. 1: Histogram of age for different Income groups

B. Bar Chart

Significant differences in wealth across ethnic groups were displayed in the bar chart for User Story 2. The distribution of both income levels among localities was depicted in the graphic with accuracy. Some groups had fewer earners in this category than white people, who made up the majority of the earners. All people but white people made at least $50,000. This emphasizes how important it is to have laws that deal with economic inequality. Implications:

1) The graphic makes it easier for UVW's marketing staff to target populations.
2) Compared to those making more than $50,000, a greater percentage of people in the Black and Asian-Pacific Islander populations are employed below that level.

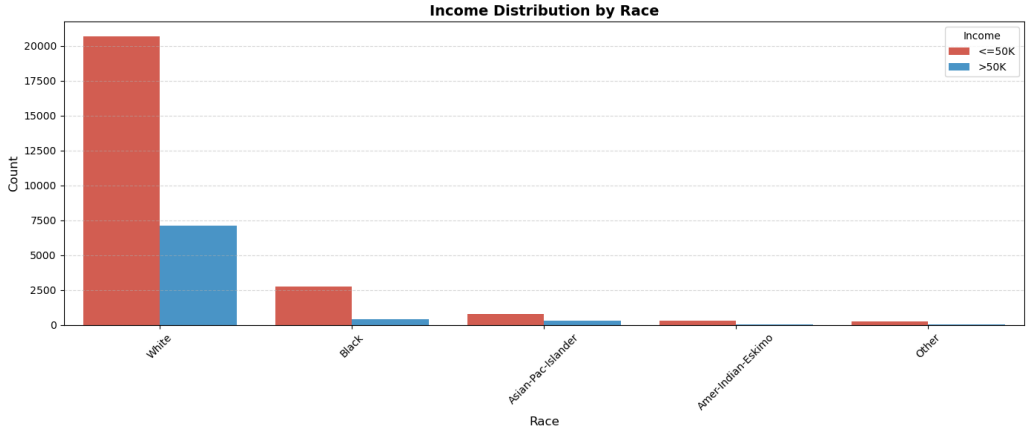

Fig. 2: Bar chart of Race for different Income groups

C. Scatter Plot

A thorough investigation was carried out in User Story 3 to comprehend the relationship among income levels, years of education, and capital gain. An helpful visual aid for illustrating these linkages was a scatter plot. The plot's conclusions were really eye-opening. It was shown that those with greater education levels were more likely to earn above $50,000 and to have larger financial gains. This implies that income levels and education levels are positively correlated.

Moreover, a clear concentration of those with lower incomes and less education was shown by the scatter plot. This emphasizes the existence of economic inequality between various educational groupings. Because marketing plans may be tailored using these information, they are extremely important. For example, educational items may be marketed to those with lower earnings and less education in an effort to improve their financial situation. Conversely, wealthier incomes may be presented with investment possibilities that take use of their financial resources. Organizations may successfully target their audience and cater to their unique wants and preferences by personalizing strategies in this way.

D. Bubble Chart

A bubble chart was used for User Story 4, which examines the association between years of education, hours worked per week, and marital status. The average number of hours worked per week and years of education are visually represented in this figure, where the size of each bubble denotes the proportion of people in each marital status category. Every labeled bubble is a distinct marital status group and shows the mean years of education and weekly hours worked for those in that category.

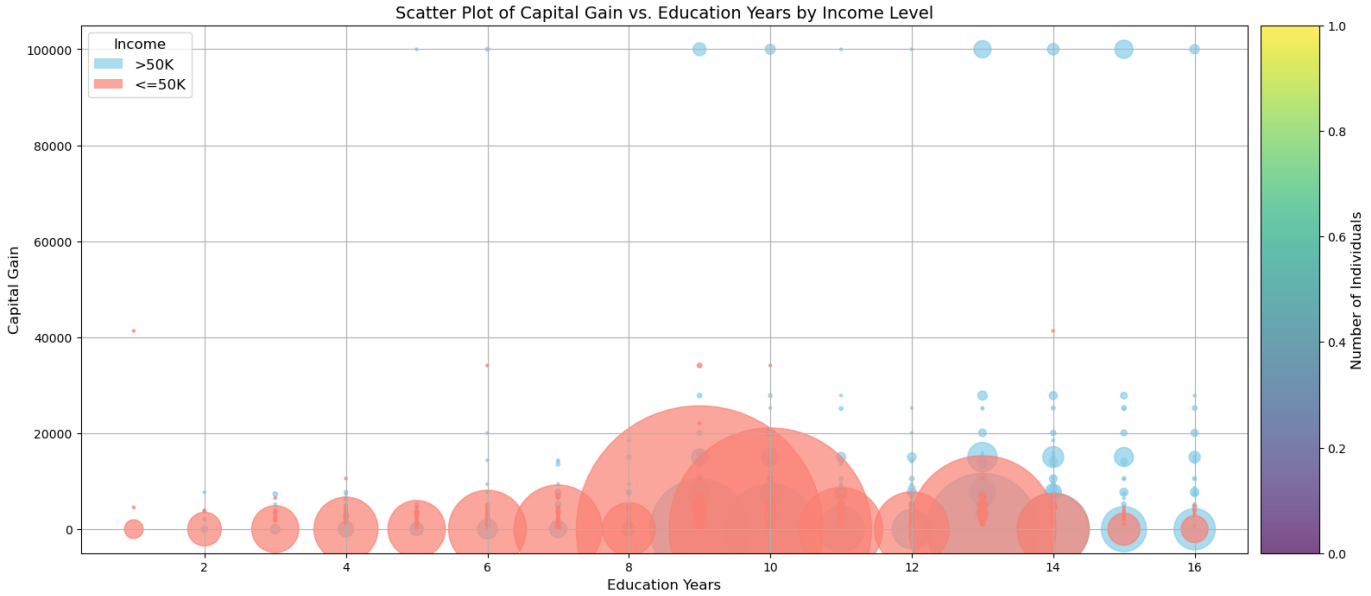

Fig. 3: Scatter Plot of Capital Gain, Education Years by Income Level

The marketing staff may better target potential registrants by using the demographic profile that was discovered via the use of statistical research. Key Findings:

1) The analysis revealed that within the same marital status group, individuals earning over $50K and those earning less than or equal to $50K exhibit varying average hours worked per week and levels of education. This suggests that income level is not the sole determinant of work hours and education level.
2) Furthermore, individuals within the same marital status but different salary brackets provided additional evidence supporting the correlation between higher education and income level, consistent with previous analyses. This reinforces the importance of education in determining earning potential.
3) The bubble chart serves as a valuable tool for the UVW marketing team, enabling them to refine social profiles for targeted marketing campaigns. It provides insights into the available time that certain social groups may dedicate to courses based on their education level, which can aid in course selection and campaign planning.

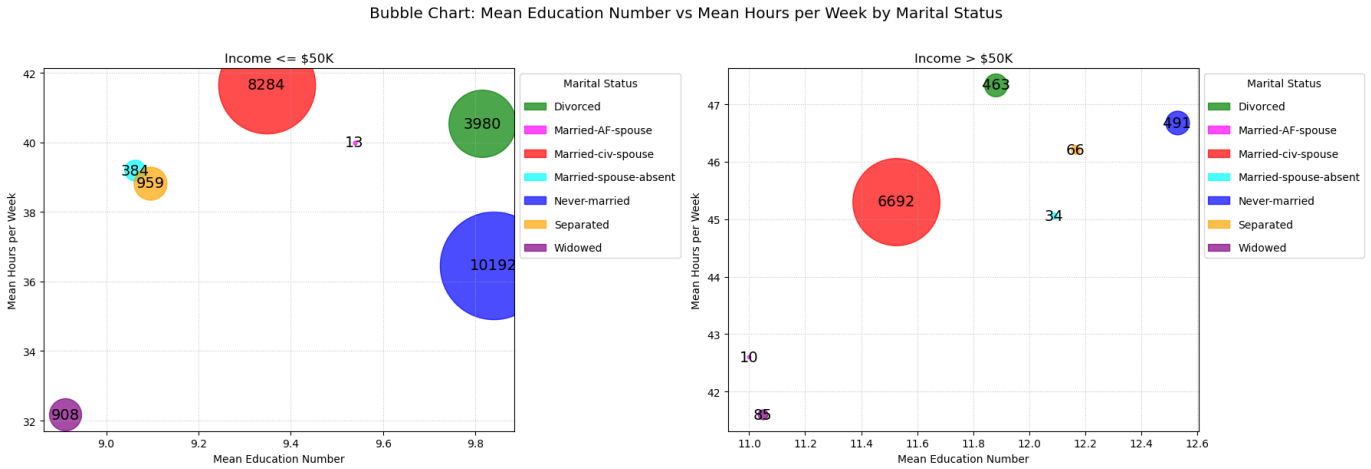

Fig. 4: Bubble chart of Education, Hours worked and Marital Status

E. Multi-Scatter Plot

I carried out an analysis to identify possible course beneficiaries and increase demographic data in order to satisfy User Story 5 and improve enrollment at UVW College. This involved visualizing the average age and weekly labor hours across different work-class groups using a color-coded scatter plot. The following is a summary of the conclusions drawn from this analysis:

1) The scatter plot identified a unique group of people who work more than thirty hours a week and are categorized as ”without-pay.” Courses on financial empowerment may be offered to this demographic, and UVW College might provide financial aid to increase the accessibility of these courses. UVW College might boost enrollment and draw in more students by customizing its programming to suit the requirements of this particular group.
2) Those who have never worked are included in the ’never-worked’ group, which is another opportunity found in the analysis. The typical age of this group is in their early 20s, suggesting that they may be interested in pursuing more education even though they do not have any job experience. In order to increase enrolment in UVW College’s courses, this demographic also corresponds with the bulk of people whose salaries are less than or equal to $50,000, indicating that they could profit from focused marketing and financial aid.

Through the utilization of these information, UVW College may focus on particular demographic groups and offer financial assistance and customized marketing to attract a diverse student body to its courses.

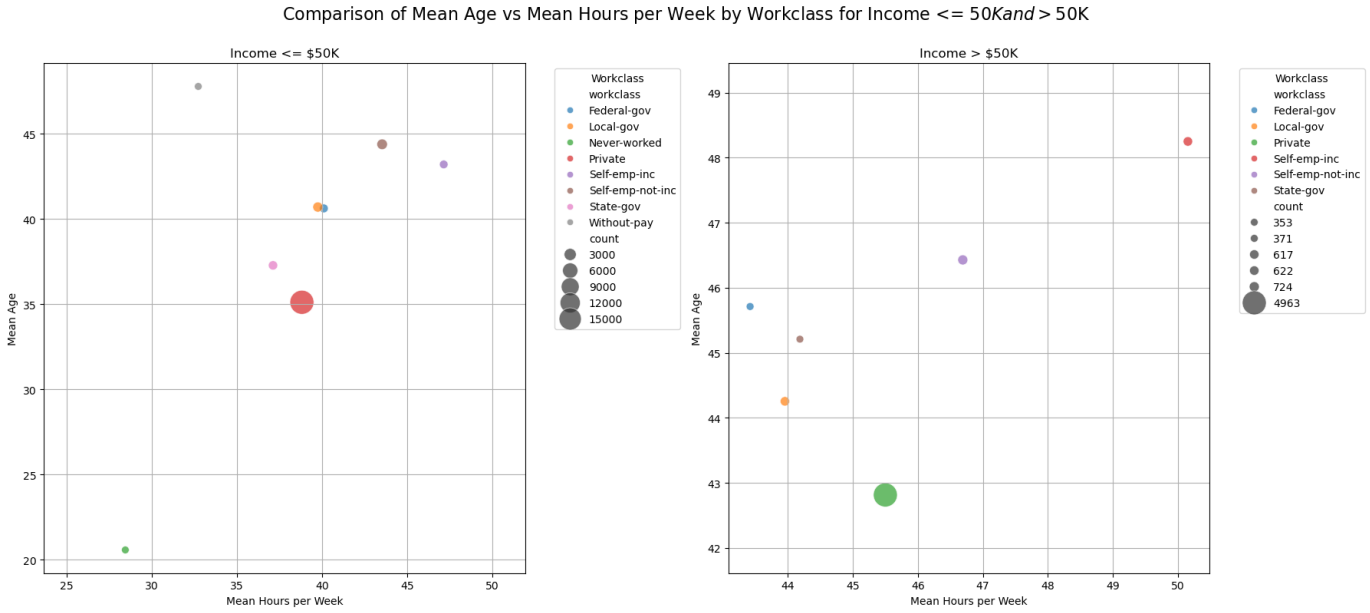

Fig. 5: Comparison of Mean Age vs Mean Hours per Week by Work-class for different Incomes

F. Mosaic Plot

A mosaic plot was chosen to accurately depict the situation for User Story 6, which entails portraying the distribution of male and female populations across various vocations in our dataset. Plots were created for both situations in which the income is greater than $50,000 and situations in which it is less than $50,000. The following are the main findings from these plots:

1) A discrepancy between male and female populations across occupations was shown by the mosaic plots. Women outnumbered males in several jobs where the compensation was less than or equal to $50,000, indicating a gender disparity in lower-income categories. Nonetheless, no occupation was predominately held by women in situations where the wage exceeded $50,000, suggesting a more even distribution of the sexes in higher income levels.
2) Although the dataset was mostly focused on males, women dominated a few professions. This indicates that women are disproportionately concentrated in particular employment roles, which calls for more research to determine the mechanisms influencing this trend.

In order to achieve gender equality in the workplace, further study and focused interventions may be required in areas where gender discrepancies persist, as shown by the mosaic plots, which overall offered insightful information on the gender distribution across income groups and occupations.

G. Heat Map

A heatmap was utilized to depict the relationships between capital gain, years of schooling, and hours worked per week in order to answer User Story 7. The heatmap presents the data in an understandable and straightforward manner while shedding light on the connections between different factors. The following are the findings from the heatmaps:

1) The study showed that capital gain and education-num were significantly more correlated (0.11 for salaries above $50,000 and 0.011 for salaries under $50,000). This confirms earlier research indicating those who make over $50,000 annually typically have higher levels of education.

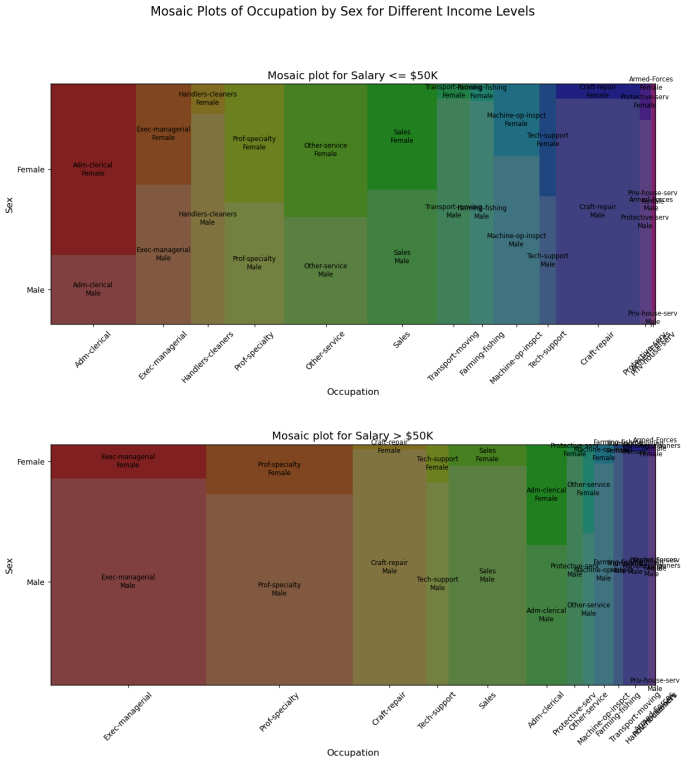

Fig. 6: Mosaic Plot of Occupation and Sex

2) Furthermore, there is a discernible variation in the relationship between weekly hours worked and capital gain. This discrepancy emphasizes how intricate the connections between these factors are and how much more research is necessary to properly comprehend how they interact.

All things considered, the heatmap is a useful tool for analyzing and displaying the connections between hours worked each week, education level, and capital gain. It provides a clear and succinct depiction of the data, making it possible for researchers to see trends and patterns that might not be immediately obvious from the raw data.

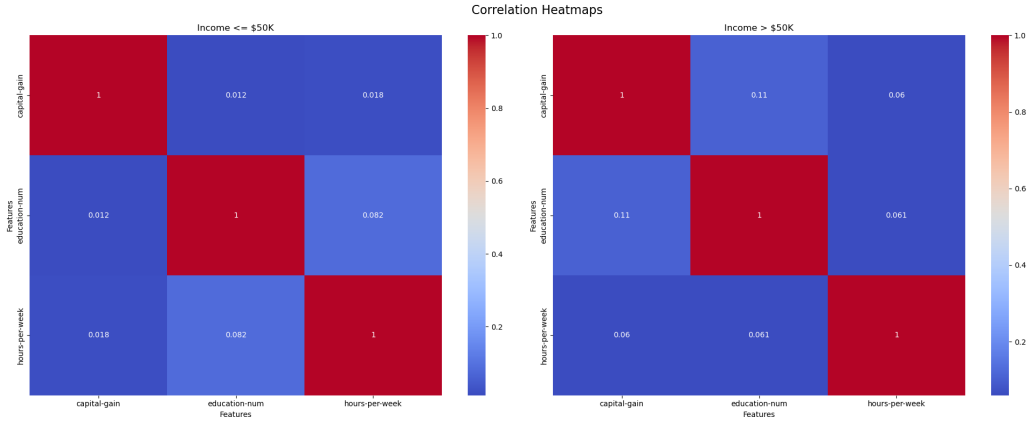

Fig. 7: Heat Map using Capital Gain, Education and Hours per week vs Income

V. Questions

The dataset presented a special hurdle to me when I initially saw it in the.Data format since I was not familiar with its structure. I started a study trip to comprehend the format and learn how to read it in order to deal with this. I was able to create the required dataframe for the project analysis by combining Pandas' DataFrame() method with Python's functionalities. Managing the dataframe's missing values presented another difficulty. Carefully going over the dataset, I found the rows that included '?'. After identifying them, I created a method to find these rows and used the replace() function to change the '?' to a None value.

The highly skewed nature of the dataset made the undertaking extremely difficult. People making less than or equal to $50,000 were almost three times as numerous as those making more than $50,000. Because one class predominated the other due to this skewness, it was challenging to draw conclusions that were useful when using scatter plots for specific user stories. In order to solve this, I used the sklearn library's resample function to produce a balanced dataset and undersampled the dominating dataset. Better analysis of the data points was made possible by this method, which guaranteed a more equal representation of both groups in the scatter plot.

Furthermore, learning how to appropriately organize and identify the data was necessary for me to operate with the.data file type. The analysis became much more complicated when handling unusual characters and missing data values. In addition, the multidimensional nature of the data made analysis and visualization difficult. It was difficult to choose characteristics for multivariate plots, such as scatterplots, that would offer useful insights and aid in creating a target profile for course marketing. Programmatically organizing data for multivariate analysis, such as scatter plots, was another challenge I encountered.

VI. Future Work

To enhance my data visualization techniques, I plan to review user stories 1 and 2 thoroughly. This review will help me select appropriate groups for analysis using oversampling and under sampling methods, especially for attributes where missing values have been replaced with None. I also aim to explore simpler, self-calibrated values to improve my understanding of grouping processes and identify the most effective charts for visualization.

Additionally, I intend to create more user stories that utilize control graphs and mosaic plots to gain further insights into the dataset. These stories will help me explore the data from different angles and uncover hidden patterns or trends.

Furthermore, I plan to develop a program that automatically identifies skewed features in the dataset upon import. This tool will have an interactive interface that allows users to select features, create plots, and receive error alerts if needed. By automating this process, I aim to streamline the analysis and visualization of the dataset.

Overall, despite the challenges faced, my data analysis has progressed significantly. My future work will focus on building tools to expedite the analysis process, exploring novel approaches to dataset analysis, and refining data visualization techniques to extract meaningful insights from the data.