Data Visualization Course Project

Vishal Anand Muruganandam, 1225435679, vmuruga7@asu.edu

I. Goals and Business Objectives

As a data analyst at XYZ corporation under UVW College, I aim to examine demographic and financial data to pinpoint potential customer segments for UVW's marketing courses, primarily focusing on income levels. Utilizing the provided dataset, I will pick pertinent factors and craft informative visuals to grasp the traits of these target groups. This analysis will aid in tailoring marketing tactics to bolster enrollment in UVW's programs. Key factors include age, race, education level, capital gain, country of origin, weekly work hours, sex, and employment status.

II. Assumptions

1) The provided dataset is presumed to be both accurate and precise, with the exception of some missing data points. This indicates its reliability and capacity to offer meaningful insights.
2) The supplied data is presumed to be current and pertinent, ensuring its relevance to the present moment. Outdated or irrelevant data can lead to misinterpretations and faulty decisions.
3) The provided data is assumed to be exhaustive and inclusive, offering a holistic perspective without any gaps. Incomplete data is just as risky as inaccurate data, potentially leading to uninformed decisions in the absence of a full understanding.

III. User Stories

1) As a data analyst at XYZ under UVW, I'm tasked with outlining the age breakdown of individuals earning below $50K, aiming to pinpoint potential age cohorts for targeted college marketing efforts.
2) In my role as a data analyst at XYZ for UVW, I'm requested to illustrate the count of individuals earning above and below $50K within various racial communities, aiding in understanding which communities could be receptive to course marketing initiatives.
3) Working as a data analyst at XYZ for UVW, I'm required to display the correlation between capital gain, education level, and income bracket ($50K and below, and above $50K), informing the design of courses tailored to different educational backgrounds within the demographic.
4) In my capacity as a data analyst at XYZ employed by UVW, I'm tasked with exploring the interplay between weekly work hours, marital status, and education years among those earning less than $50,000 annually, informing course marketing strategies tailored to specific demographic needs.
5) As a data analyst at XYZ within UVW, my task involves presenting data on weekly work hours, age, and employment class for individuals earning $50K or less, enhancing understanding of target group demographics and creating comprehensive student profiles.
6) As a data analyst at XYZ for UVW, I'm tasked with showcasing the gender distribution within both salary brackets greater than $50K and less than or equal to $50K categorized by occupation, shedding light on the demographic composition of prospective students.
7) In my role as a data analyst at XYZ for UVW, I'm requested to exhibit the correlation among capital gain, weekly work hours, and education years to offer insights valuable for UVW's marketing team in their decision-making process.

IV. Visualizations

A. Histogram

To address user story 1 for UVW, I was tasked with identifying the optimal age group for their marketing strategy. I opted for a histogram because it accurately represents the age distribution, enabling us to pinpoint the target age group for course marketing. The implicationsare highlighted below:

1) This histogram analysis for user story 1 will assist UVW's marketing team in identifying the age group they should target with their courses.
2) The histogram indicates that most individuals in the dataset were over 15 years old, and those between 15 and 35 constituted a significant portion of those earning less than or equal to 50K.

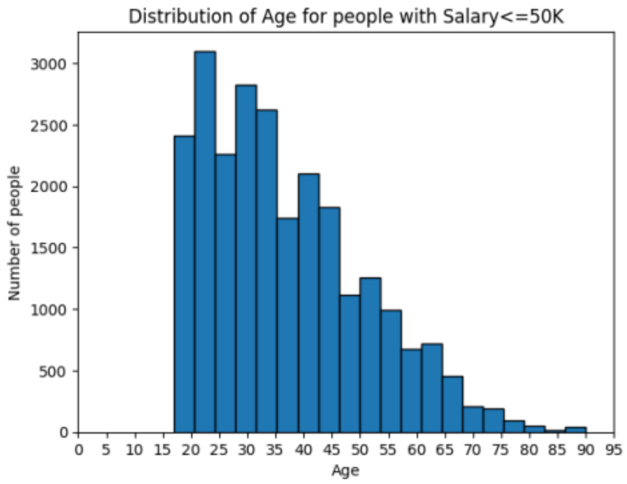

Fig. 1: Histogram for user story 1

B. Stacked Bar Chart

To fulfill user story 2 aimed at demonstrating the demographic breakdown of individuals earning above and below $50,000 to UVW's marketing team, a stacked bar chart is the optimal visualization choice. This type of chart effectively illustrates the distribution of individuals earning above and below $50,000 within each racial group. While the initial graph displays a skewed dataset, the subsequent one presents a more balanced dataset. Both charts highlight that despite the dataset's imbalance in the first instance, the representation of minority groups concerning individuals earning above and below $50,000 remains largely consistent. The implications are highlighted below:

1) The stacked bar chart associated with user story 2 serves to assist UVW's marketing team in targeting various communities more effectively.
2) Upon analysis, it becomes evident that although a significant portion of individuals earning below $50,000 belong to a particular community, the stacked bar chart accurately portrays the distribution across communities for both income brackets. Notably, the Black and Asian-Pacific Islander communities exhibit a higher proportion of individuals earning below $50,000 compared to those earning above this threshold.

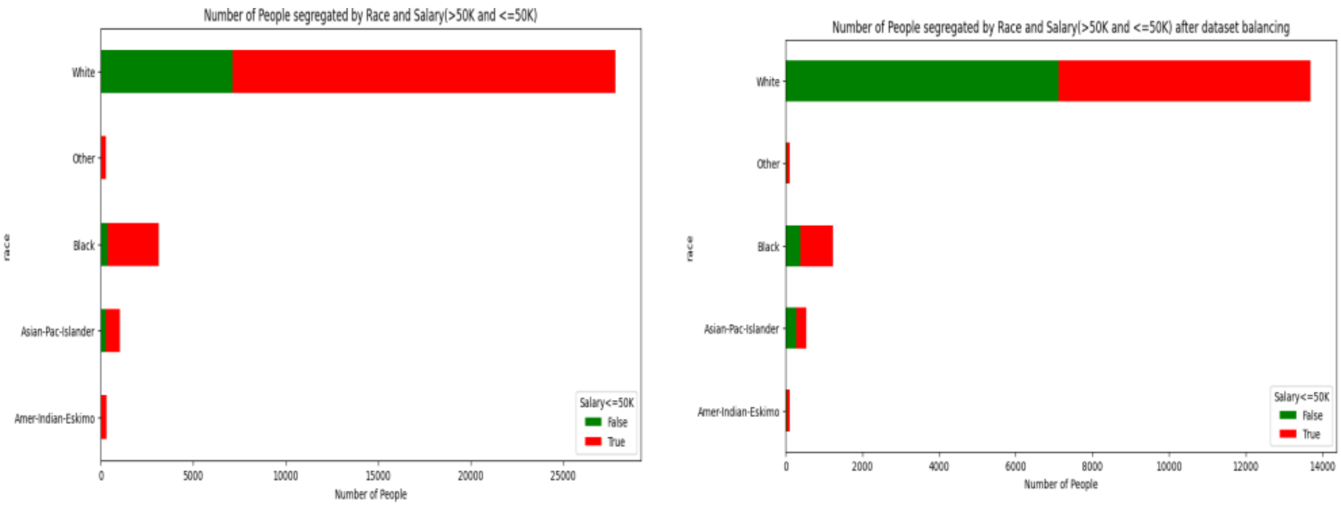

Fig. 2: Stacked Bar charts for user story 2

C. Scatter Plot

To fulfill the requirements of user story 3, which involves displaying the distribution of capital gains and years of education for individuals earning above and below $50,000, we propose creating a scatter plot. This plot will use color coding to distinguish

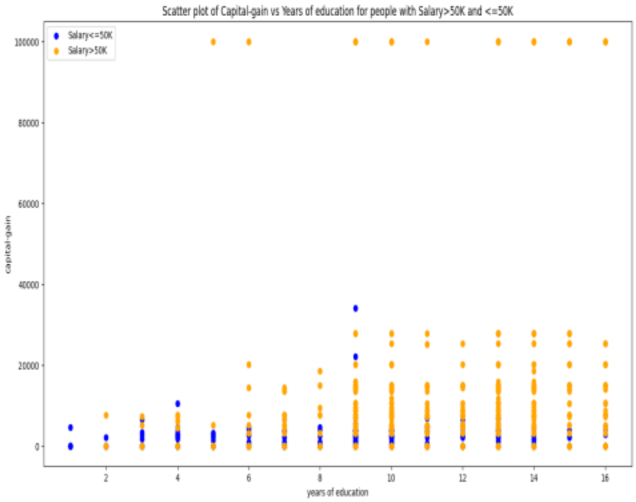

Fig. 3: Scatter Plot for user story 3

between those earning less than or equal to $50,000 and those earning more. By doing so, we aim to grasp the correlation between these variables and their impact on salary levels. To ensure a balanced dataset, we have employed undersampling for those earning less than or equal to $50,000, allowing us to examine the influence of education and capital gain on salary. Our observations are as follows:

1) The graph indicates a positive correlation between the number of years of education and the proportion of individuals earning over $50,000.
2) Additionally, the graph reveals an upward trend in capital gains among those earning over $50,000 as the number of years of education increases.
3) Notably, there is a significant majority of individuals earning over $50,000 who exhibit higher levels of education and capital gains.

D. Bubble Chart

To fulfill user story 4, which requires analyzing the connection between marital status, hours worked per week, and years of education, I employed a bubble chart. This chart illustrates the average hours worked per week and years of education, with bubble size representing the number of individuals in different marital status categories. Each labeled bubble represents a category, displaying the average years of education and hours worked per week for individuals within that category. Utilizing statistics, we gain insight into the demographic profile, aiding the marketing team in targeting potential enrollees. The key findings are highlighted below:

1) Within the same marital status group, individuals with Salary greater than $50K and Salary less than or equal to $50K exhibit varying average hours worked and education levels.
2) Individuals within the same marital status but different salary brackets provide further evidence supporting the correlation between higher education and salary, as noted in a previous analysis.
3) This chart assists the UVW marketing team in refining social profiles for targeted marketing campaigns. It highlights the available time certain social groups may dedicate to courses based on their education level, aiding in course selection.

E. Multi-Scatter Plot

To fulfill user story-5, which requires expanding demographic data to locate social profiles of potential course recipients for UVW College, aiming to boost enrollment, I employed a color-coded scatter plot. This plot illustrates the average age and weekly work hours across various workclass categories. Key insights are drawn below:

1) The plot indicates a segment of individuals classified as ’without-pay,’ who work more than 30 hours weekly. UVW College could target this group with courses focused on financial empowerment and offer financial aid.
2) Additionally, the ’never-worked’ category presents an opportunity for targeted marketing and financial aid, given the average age falls in the early 20s, aligning with the majority earning Salary less than or equal to $50K.

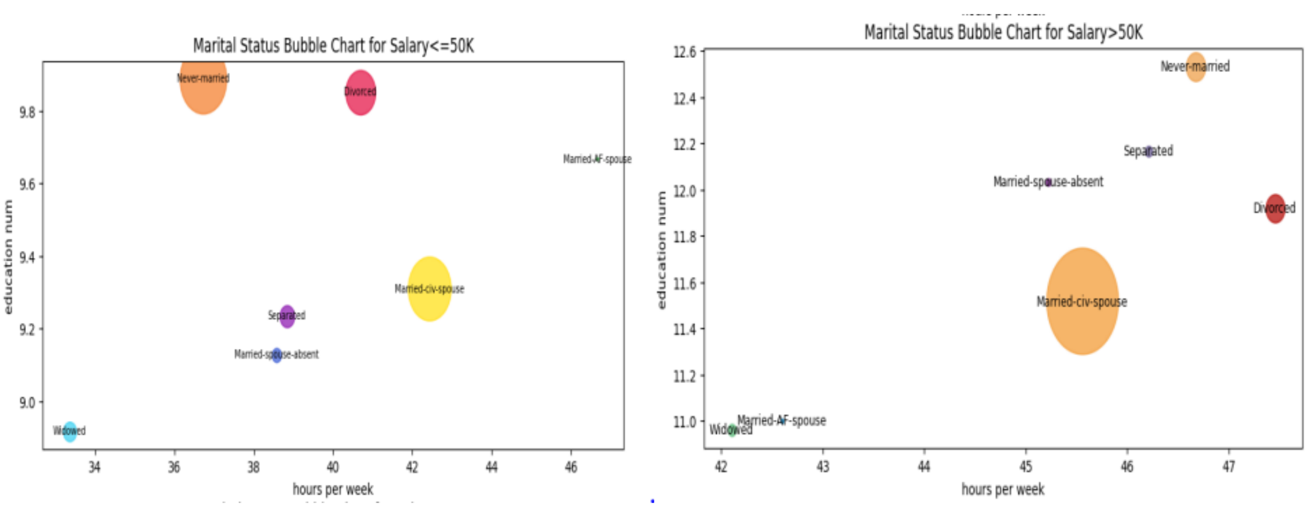

Fig. 4: Bubble charts for user story 4

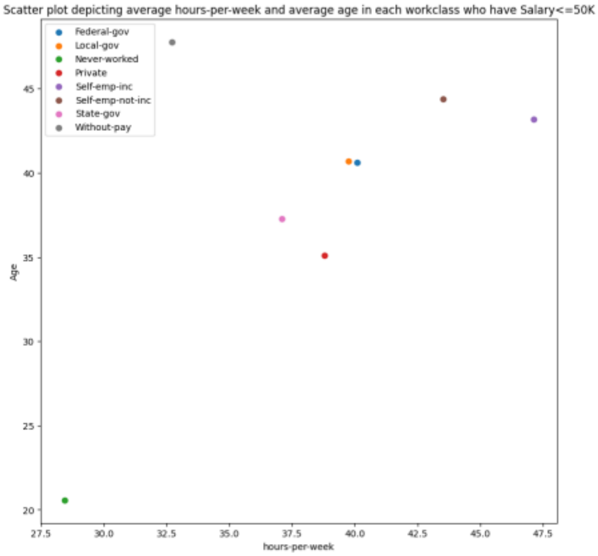

Fig. 5: Multi-Scatter plot for user story 5

F. *Mosaic Plot*

To fulfill user story 6, we're tasked with displaying the breakdown of male and female populations across various occupations in our dataset. I opted for a mosaic plot to accurately illustrate this scenario. I've generated plots for both instances where Salary is less than or equal to 50K and where it exceeds 50K. Key observations are:

1) The plots reveal a disparity: when Salary is less than or equal to 50K, women outnumber men in certain occupations, contrasting sharply with instances where Salary exceeds 50K, where no occupation is predominantly female.
2) Notably, despite the dataset being predominantly male-centric, we observe certain occupations where women dominate. This suggests a significant concentration of women in specific job roles, sparking further inquiry.

G. *Heatmap*

To fulfill user story 7, which involves analyzing the relationship among capital gain, years of education (education-num), and hours worked per week (hours-per-week), I opted to address it by employing a heatmap. This heatmap will offer insights into the correlations between capital gain, hours worked per week, and years of education. The observations from the heatmaps indicate:

1) The analysis reveals a notably stronger correlation between capital gain and education-num (0.11 for Salary greater than $50K and 0.011 for Salary less than or equal to $50K), supporting previous findings suggesting that individuals with a Salary greater than $50K tend to have higher levels of education (education-num).

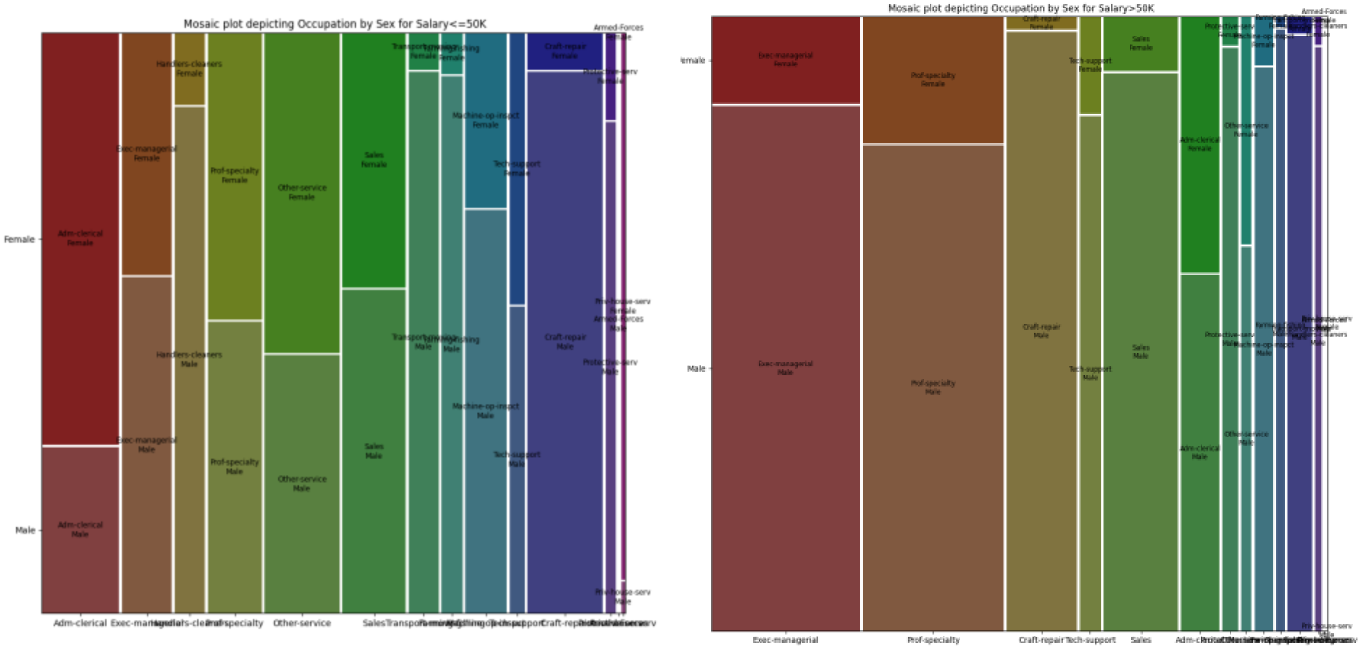

Fig. 6: Mosaic plots for user story 6

2) Additionally, there is a notable difference in the correlation between capital gain and hours worked per week (hours-per-week).

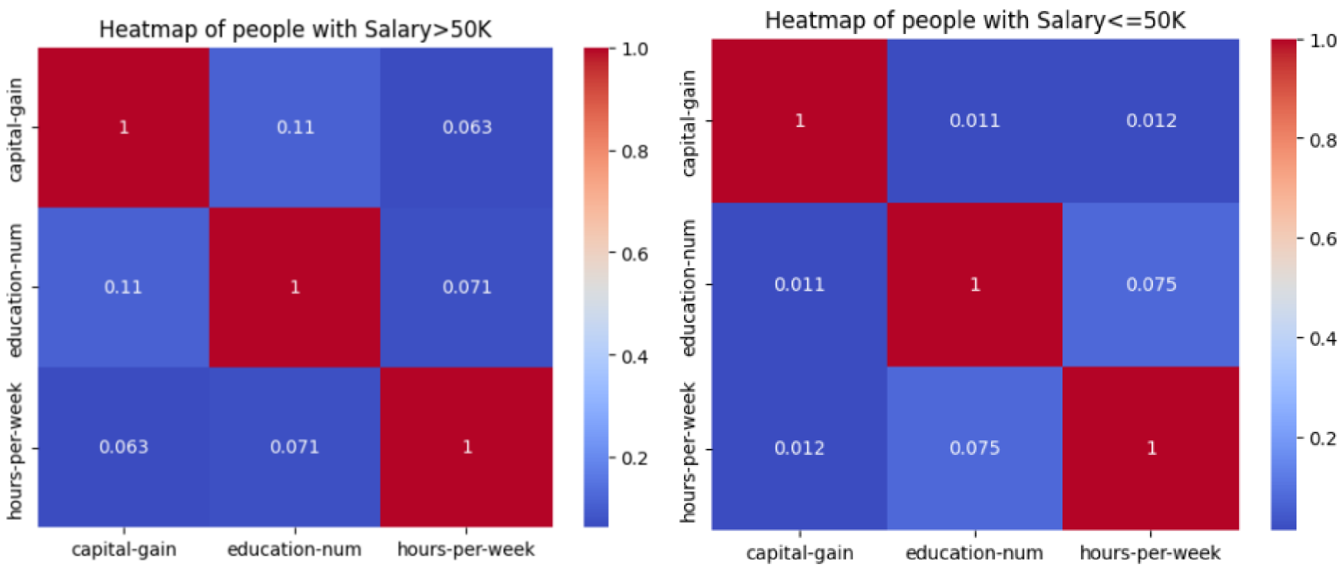

Fig. 7: Heatmaps for user story 7

V. Questions

1) Initially, the dataset posed a unique challenge as it was in a .Data format, which was unfamiliar to operate with at that time. To tackle this, I embarked on learning about the .Data format and how to interpret it. Employing Python's functions like lists, split(), and replace(), alongside Pandas' DataFrame() function, I constructed the necessary dataframe for project analysis.

2) Dealing with missing values in the dataframe presented another hurdle. To address this, I meticulously examined the dataset to identify rows containing ’?’. Once identified, I developed a method to pinpoint these rows and subsequently replaced the ’?’ with a None value using the replace() function.
3) The project introduced a significant challenge with a heavily skewed dataset, where individuals earning Salary¡=50K outnumbered those with Salary greater than $50K by nearly threefold. This disparity led to complications when attempting to utilize scatter plots for certain user stories, as one class dominated the other, hindering any meaningful inferences. To mitigate this, I opted for undersampling of the dominant dataset and created a balanced dataset using the resample function from the sklearn library. This approach allowed for a more equitable representation of both classes in the scatter plot, enabling better analysis of the data points involved.

VI. Future Work

In the future, my plan involves creating a tool that automates the detection of skewed features upon dataset loading. This tool will sport an interactive interface, allowing users to choose features and plots, while also providing error messages when plots cannot be generated. Throughout this project, I’ve pinpointed crucial attributes that influence a person’s salary class, and I aim to utilize machine learning methods for predictive analysis.