Data Visualization Course Project

Rohith Krishna Papani

Arizona State University
rpapani@asu.edu

Goals and Business objectives

As a data analyst at XYZ corporation working for UVW's college, my goal is to analyze a dataset of demographic and financial information and identify potential target groups for UVW College's marketing courses based on their income levels (salary). The dataset provided will be used to select relevant attributes and develop effective visualizations to understand the characteristics of the target group. The analysis will assist in the creation of targeted marketing strategies to promote UVW's courses and programs.Key Attributes : 1. Age 2.Race 3. Education Num 4. Capital Gain 5. Native Country 6. Hours-per-week 7. Sex 8.work class

Assumptions

1. The given dataset is assumed to be accurate and precise apart from the missing data points for some entries. This will signify that the data is free from errors and provides us with meaningful insights.

2. The given data is assumed to be timely and relevant This means that the data is collected at the right moment in time and remains relevant to the current situation. If the data is outdated or irrelevant, it can misrepresent the situation and drive inaccurate decisions.

3. The given data is assumed to be complete and comprehensive.This means that the data is without gaps and provides a comprehensive view of the overall picture. Incomplete data can be as dangerous as inaccurate data, and without a complete picture, uninformed actions can occur.

User Stories

1. As a data analyst at XYZ working for UVW, I am asked to provide the age distribution of people earning less than 50K to identify potential age groups for the college to target in their marketing efforts.

2. As a data analyst at XYZ working for UVW, I am asked to show the number of people with salaries less than and greater than 50K from each community (Race) to understand the communities to which the courses can be marketed.

3.As a data analyst at XYZ working for UVW, I am asked to show the capital gain, education years for people earning more than than 50K and less than equal to 50K to market different courses according to the education level of the demographic.

4.As a data analyst at XYZ working for UVW, I am asked to understand the relationship between hours worked per week, marital status, and number of years of education for individuals earning less than $50,000 per year, so the college can market courses as per the requirements which could be presented by the target groups .

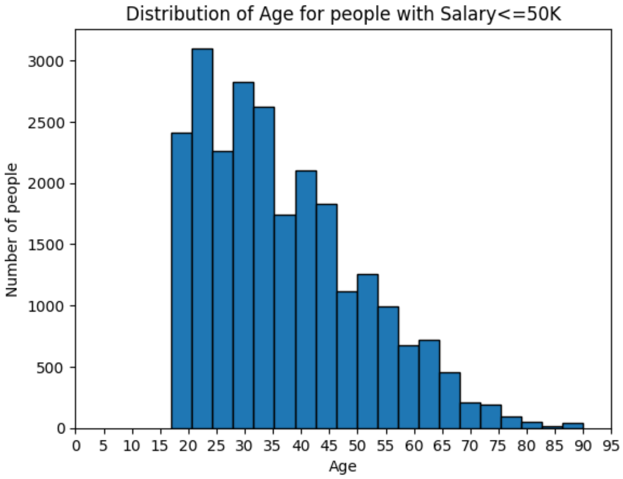
5.As a data analyst at XYZ working for UVW, I am asked to show hours-per-week , age and work-class for people. making a salary less than or equal to 50K to better understand the target groups' demographics and make a social profile of target students.

6. As a data analyst at XYZ working for UVW, I am asked to show the distribution of Male and Females belonging to both classes of Salary >50K and Salary<=50K , segregated according to the occupation, to understand the demographic makeup of the target student population.

7. As a data analyst at XYZ working for UVW, I am asked to show the marketing team at UVW the relationship between Capital-gain , hours-per-week and years of education to provide valuable insights in the decision making process.

Visualizations

1. Histogram



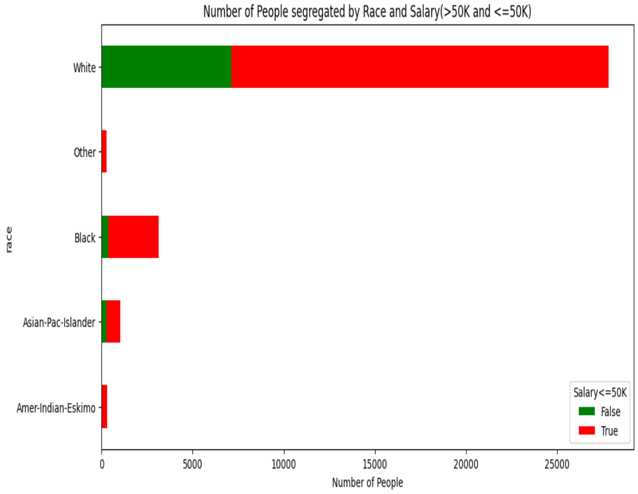
To deal with user story 1 , where UVW needed me to find the target age group according to which they can draw their marketing strategy. I chose a histogram as it accurately depicts the age distribution which could be utilized to finding the target age group to market courses.

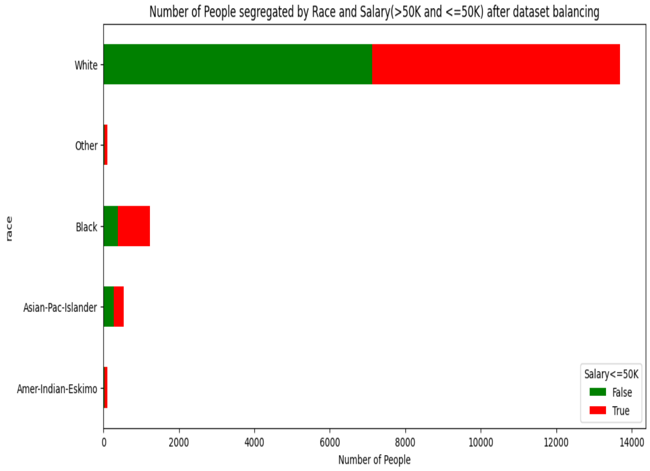
Inferences:

1. The above histogram plot for user story 1 will help the marketing department at UVW in determining the target age group to which they will be marketing their courses to.

2. The above histogram suggests that the majority of people in the dataset belonged to age >15 and age<=35 harbored a major part of the people with Salary less than equal to 50K.

2. Stacked Bar chart





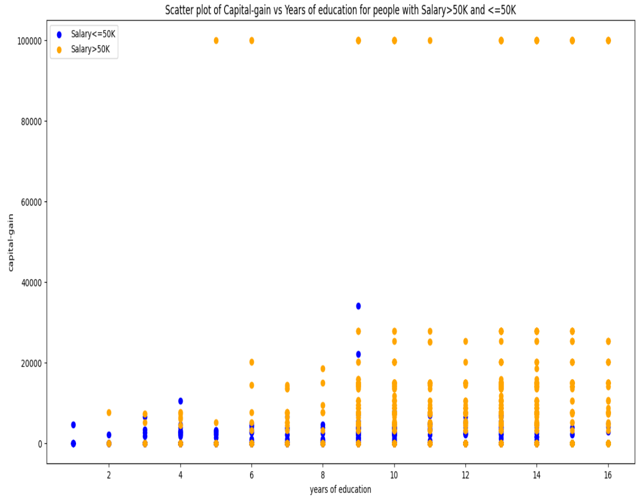
To complete the user story 2 which was intended to show the marketing team at UVW the race-based make up of people with Salary >50K and Salary<=50K. The best choice of plot would be a stacked bar chart as it depicts the proportion of people making salary<=50K and number of people making Salary>50K. The above graphs as titled depict the stacked chart of each race making Salary>50K and <=50K but the first plot has a skewed dataset but the second plot has a balanced dataset. I plotted both to showcase that even though the dataset presents an imbalance in the first-case , the compositions of minority groups in accordance with people with Salary>50K and Salary<50K remains almost identical.

Inferences:

1. The above plot for user story 2 will help the marketing team at UVW college in their marketing efforts to reach out to different communities.

2. On analysis It is found that although a majority of people with Salary<=50K belong to one community , the stacked bar chart shows the true picture by showing the proportion from each community who belong to class with Salary<=50K and Salary>50K. The Black and Asian-Pac-Islander communities were found to have a major proportion of people with less than 50K than people with more than 50K.

3. Scatter Plot



To complete user story 3 which demands us to show the capital-gain and years of education people making Salary>50K and Salary<=50K to identify the capital-gain and education-years distribution. I believe the scatter plot with color coding to differentiate people with Salary<=50K and Salary>50K aids us to understand the relationship between each of the variables and their bearing on the Salary being less than equal to 50K or greater than 50K. The dataset was made balanced by undersampling the Salary<=50K, as we want to learn about the bearing of years of education and capital gain on people's Salary.

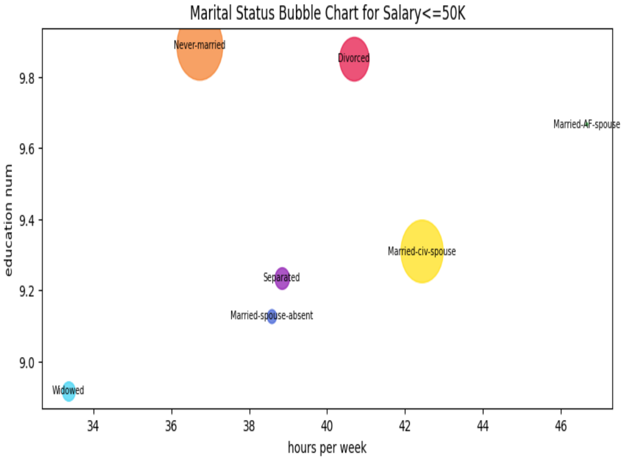
Inferences:

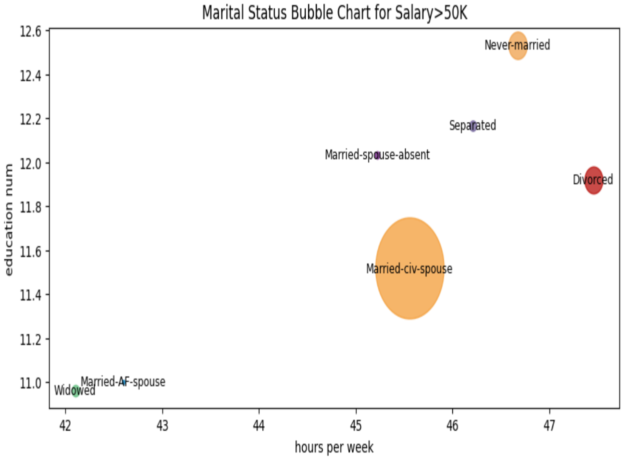
1. From the graph we can infer that as the number of years of education increases there is an increase in people with Salary>50K.

2. We also observe from the graph that even for people with Salary>50K there is a rise in capital-gain with increasing number years of education.

3. We observe that there is an overwhelming majority of people with Salary>50K with increase in years of education and capital-gain.

4. Bubble chart





To complete user story 4 which demands us to find the relationship between marital-status , hours worked per week(hours-per-week) and years of education

(education-num). I adopted the bubble chart between average number of hours worked per week and average years of education and size signifying the number of people classified under different categories of marital-status. Each labeled bubble signifies the category the people in the dataset belong to and their average years of education and average number of hours they work per week. Statistics helps us in getting an idea about the demographic of the people to whom then the marketing team can profile to design a marketing campaign to increase enrollment.

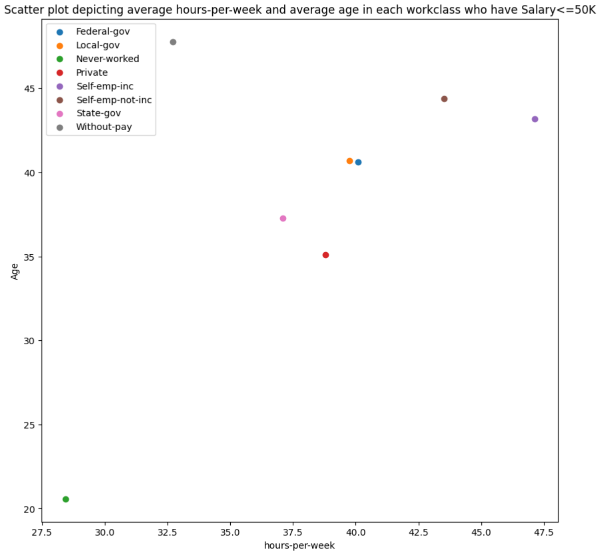
Inferences:

1. From the plots we observe that the same class of people(in this instance , people belonging to the same marital-status class) have differing levels of average hours worked and average years of education in the case of them earning Salary>50K and Salary<=50K.

2. The people belonging to the same marital-status class but belonging to different Salary class(Salary>50K and Salary<=50K) provide us with more conclusive evidence of observation we made in an earlier user story where more years of education acted as a strong indicator towards their Salary.

3. This plot helps the UVW marketing team to find one more parameter for crafting the social profile of people to whom they intend to market courses to. This plot indicates the number of hours the particular social group can devote towards classes , based on years of education to find the kind of courses they might need.

5. Multi - Scatter Plot



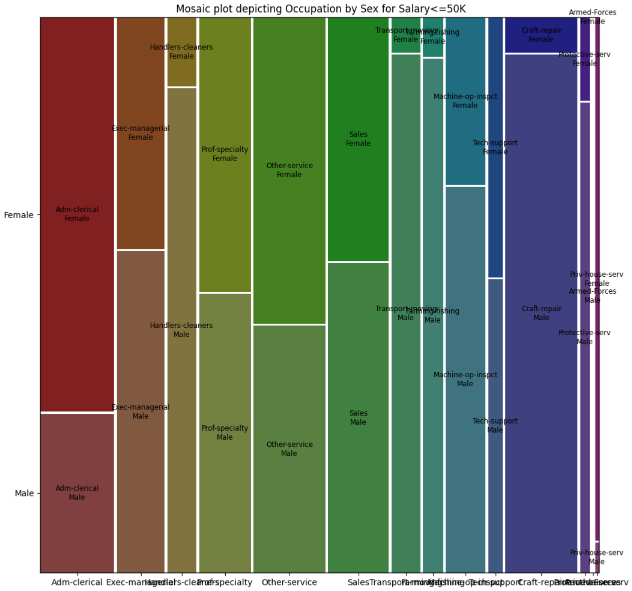
To complete user story-5 which demands us to further their information about demographics to find social profiles of people to whom they can market courses to increase enrollment in UVW college using workclass , age and hours of work per week. I utilized the color-coded scatter plot to represent my aggregate data which shows the average age and hours of work per week belonging to each of the categories of workclass.

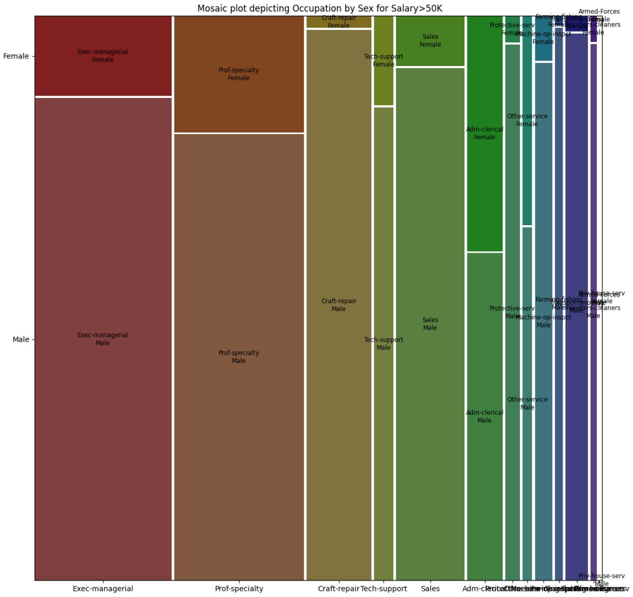
Inferences:

1. From the plot above we can observe that there is a class of people who have fallen under class called without-pay and work for over 30 hours a week. The UVW college can market financially empowering courses and financial aid to people falling under this category.

2. Also the workclass category called never-worked provide a good target group to market courses and offer financial aid as the average age of the group seems to be in early 20's falling in the band that represents the majority of people earning Salary<=50K

6. Mosaic Plot





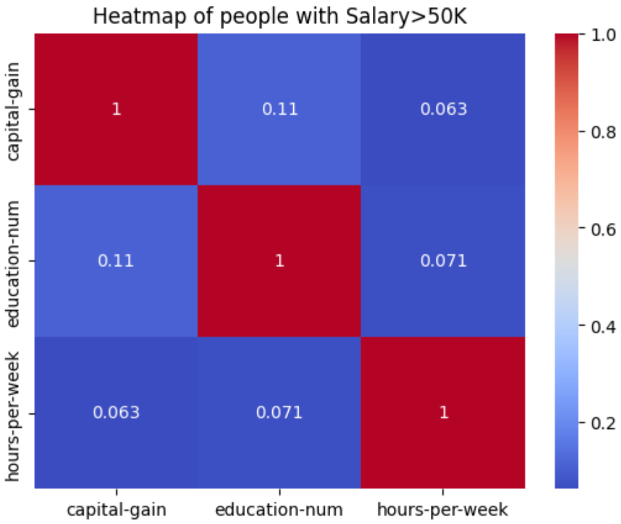
To complete user story 6 we are asked to show the distribution of male and female population for each occupation found in the dataset. I utilized the the mosaic to depict the scenario accurately. I have depicted the plots for the cases where Salary<=50K and Salary>50K.

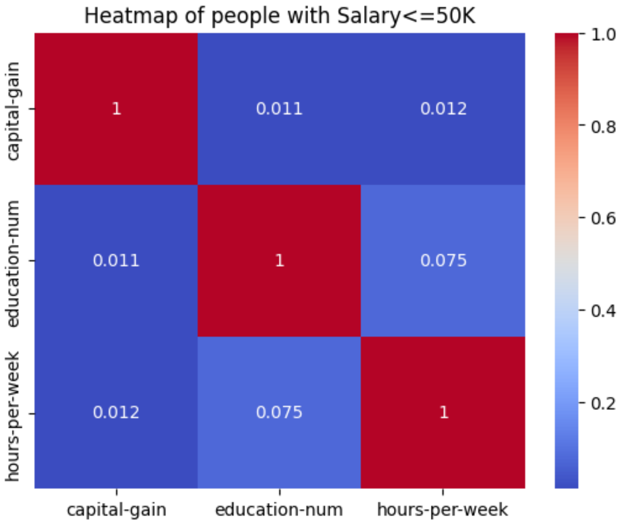
Inferences:

1. From the plots given we observe that in the cases where Salary<=50K , women population dominates the male population for some of the same occupations in stark contrast to the case where Salary>50K where no occupation is dominated by women.

2. It is noteworthy to observe that we see some jobs are being dominated by women despite majority of entries collected in the dataset being skewed with male entries which leads us to an interesting avenue which indicates high concentration of women in few of the occupations.

7.Heat Map





To complete user story 7 which requires us to find a relationship between Capital-gain , years of education(education-num) ,and hours worked per week(hours-per-week). I chose to take care of this user story by utilizing heatmap which will provide me with inferences about correlations between capital-gain, hours-per-week and years of education(education-num).
Inference:

1. On observing both the heatmaps It is observed that there is a much stronger correlation between capital-gain and education-num(0.11 for Salary>50K and 0.011 for Salary<=50K) which might stand as a corroborative evidence to the previous evidence made about people with Salary>50K having more years of education(education-num).

2. We also observe a significant difference in correlation between capital-gain and hours worked per week(hours-per-week).

Future Work

In the future , I intend to build an automated tool with an interactive UI that can detect the features which are skewed when I just load the dataset. This tool should be able to give the user options to select the features and option to select plots along with error messages and when the plot is not possible.

During the course of this project I have identified key attributes that determine the Salary class of the person. I intend to apply machine learning techniques to do prediction.

Questions

1. The Dataset initially was in a .Data format which presented a unique challenge due to operational unfamiliarity with .Data format at that point in course.

Solution implemented : I started by learning about .Data format and learned to read the format. I then utilized lists , split() , replace() functions in python along with DataFrame() function from Pandas library to develop the dataframe upon which I did the analysis for the project

2. The Dataframe presented a new challenge in terms of missing values.

Solution implemented: I explored the dataset to find if the rows did indeed have '?'. I implemented a method to find out if its the case and display these rows and then I replaced the '?' with None value using replace() function.

3. This project presented a highly skewed dataset with people earning Salary<=50K being almost 3 times the number of people with Salary>50K. This meant that when I tried to utilize scatter plots for some user stories one class over-dominated another , obscuring any inferences possible.

Solution implemented : I performed undersampling of the dominant dataset and made a balanced dataset using the sklearn library and resample function. This helped me solve the issue with a scatter plot where data points from both the classes(Salary<=50K and Salary) are involved.