Seeing is Recruiting: Visualizing Income Factors for Targeted Enrollment

Austin Hudgins
alhudgins@ASU.edu

I. Goals and Business Objectives

This paper's main goal is to showcase different data visualizations techniques that are used to find factors to determine if someone's income is above or below $50,000. These factors are found through U.S. Census data and include information Age, Occupation and Gender just to name few. The meeting of this goal will help XYZ Corporation develop its marketing profiles which can be sold to companies or for this application to be used at schools like UVW College to bolster their enrollment. This application will be able to showcase some of these major factors so they can find future students. If done correctly this can be scaled and used across all campuses to help any college meant and or exceed their enrollment goals.

II. Assumptions

1. The data is correct.
2. The data was collected in an unbiased manner.

Since we are given the data from the United States Census Bureau, we are not sure if all or any of the data is correct or if it was collected in a matter that would not bias the results. If it comes out the data is incorrect or was collected incorrectly then the graphs could vastly be different with complete opposite observations. This could be tested with a smaller sample size to see if they get results that compare roughly the same to the results that are represented in this paper.

III. User Stories

1. *As a member of the UVW Marketing Team, I want to know what occupations make more or less than $50,000 so that I can effectively advertise to possible students that want to switch careers and make more money.*

For user story one, I created two different donut graphs, to show what occupation make up what portion of the job market for incomes less than $50,000 and over $50,000. I did this because a member of the UVW Marketing Team needs to know what occupations can make more or less than $50,000. Using a donut chart, you can see that "Exec-Managerial" jobs can make above and below the target income, but it makes up a much larger portion of jobs over $50,000.

With this information the marketing team can advertise to occupations like "Handlers-cleaners" who want to make more money, because very few of them make over $50,000. This could cause enrollment to increase if UVW college offers business like classes or certifications that would help transition them to a new career allowing them a higher chance to make more than $50,000.

I created this by first combing "Armed-Forces" and "Priv-house-serv" in to the "Other-Service". I did this because they took up less than 1%, which leads to the data being hard to read. For incomes above and below the target price I add the occupations to a dictionary as the key and the number of times that occupations shows up as its value in the dictionary. I then use the keys as the labels of the graph and the values as the data that makes up the percentage.

2. *As a manager of the UVM Financial Aid Team, I want to know what factors affect possible students outside the US so that I know what type of students might be offered scholarships or grants to come to UVM.*

For user story two, I created a mosaic plot, to capture how a person's country of birth and their gender could affect if they have an income above or below the target income of $50,000. I created it because a manager of the

UVM Financial Aid Team wanted to know what factors could affect income, so they know who to offer scholarships and grants to. The mosaic plot allows you to see what percentage of the data makes up each gender and if they were born in the US make with an income above and below the target. I chose this because it allows the manager at UVM to know what type of student would most likely benefit or qualify for a grant/scholarship.

This information allows the UVM financial aid team to see that on average women are more likely to be making less than the target income and might would benefit for a grant or scholarship. This is a huge factor for many households because prices of school have steadily been increasing so giving more grants to women could increase the number of students that enroll that wouldn't other wise be able to afford it. While being born outside the US has some effect, it is not nearly as big as a factor as gender. This could save UVM money by not offering more scholarships to potential students born outside the US because they distributed around the target income just like potential students born in the US.

I created this by first grabbing only the columns of data I needed, "income", "native-country", and "sex". Then looped through the data and skipped rows that included null data from one of the 3 columns, this allows me to only get data with complete information so it would not skew the results, like if one of the columns was missing a large chunk of data. Next, I added the data to a dictionary where the keys are every combination of income above and below the target income, native country being in or out of the US, and thier sex. The total number of each occurrence of the key is the value for said key and then used statsmodel library to graph it.

3. *As a head of the UVM Career Success Center, I want to know how a person's capital gain could affect their income over their lifetime so I can talk with potential students at different points in their life about coming to our college.*

For user Story three I created a line graph for above and below the target income that compares average capital gain to the person's age. I did this because the head of UVM's Career Success Center wanted to know how a person's capital gain over their life influences their income. I chose a line graph because age is an ordinal data, and it shows you how a person's capital gain changes every year throughout their life, on average. With this graph the UVM Career Success Center can see if a person is making above or below their target income solely by seeing their age and average capital gain and target advertisement more specifically to potential students that fit the criteria of what they might be looking for.

Except for a few ages, UVM will be able to see that almost all people that make below the target income have less than $1,000 of capital gain. This is the opposite for almost every age making more than the target income. The exception being ages 20, 81 and 85, this could be do outliers in the data or just coincidence.

The line graph was created by first separating the data and only grabbing what columns of data I would need, "age", "capital-gain", and "income". I then removed all the data that was more than 2 standard deviations away from the mean for all nonzero data. This was done because there was an unusual amount of data that was "9999". This could have been because a lot of people made that much capital gain, but it seemed more likely that it was the maximum for the field, so it defaulted the data to "9999". Either way that portion of the data was removed to not create an unnatural skew to the data. I added the data to a dictionary where the income and age were the key, and the value was an array with capital gains for that age. I then averaged each age and sorted the dictionary by age, using the keys for the x axis and values for the y axis.

4. *As a member of the UVW Marketing Team, I want to know if education level can affect your capital loss based on your income level so that we know what type of Jobs and Education we should target our advertisements too.*

For user story 4, I created a scatter plot that compares the education level to the average capital loss for income above and below $50,000. I did this because the UVM marketing team wanted to about how or if education and capital loss affected your potential income so they can more effectively target advertisements. Using this graph, you can see on average among most education levels, people with income over $50,000 have more capital loss than incomes below $50,000.

All education levels except for two of them, show people above the target income have more capital loss. This will allow the UVM to target advertisements towards their target audience more effectively. If they need enrollment of people above the target income, UVM should target advertisements towards higher capital loss as the education level increases, specifically over $100 of capital loss. This will allow them to have the best shot at

hitting their target audience. Without this they could be advertising for people that cannot afford to go to school, wasting their market budget and not increasing enrollment.

I created this graph by first adding the data I need to a data frame, “Capital-loss”, “education-num”, and “income”. Then created a dictionary for income above and below the targe, with the keys being the education level. Then looped through the data frame, ignoring rows with null data, and added the capital loss to the dictionary based on its income and education. After I averaged the capital loss per education and used the keys as the x-axis and the average capital loss as the y-axis.

5. *As a member of the UVW Marketing Team, I want to know how hours worked per week are affected by their income so I can accurately advertise to potential students.*

For user story 5, I created a box plot showing the distribution of the number of hours worked by people above and below the target income on average. I did this because the UVM marketing team wanted to see how hours worked per week will affect someone’s income. It will be able to show UVM’s marketing team how they can advertise to people above or below a certain number of hours to be able to target the most people at once, so they have more effective advertising.

With this graph UVM’s marketing team can see the relationship between hours worked per week and its income. They will be able to target people working less than 40 hours per week to see if they would be interested in enrolling to have the potential to increase their income. This also will show them what percent of people make up the majority of each income, if UVM advertise to people working over 40 hours there are some but they are outliers and would probably not have a successful advertisement causing a small if any enrollment increase.

This was created by first grabbing the data I needed, “hours -per-week” and “income”, then adding it to a data frame. After I created a dictionary with two keys, above and below the target income, and made the value an array. I then appended to the array if there was not a null value in the row and the income was matching the key. The x-axis became the 2 box plots, over and under target income, and the y-axis being hours worked per week.

IV. VISUALIZATIONS

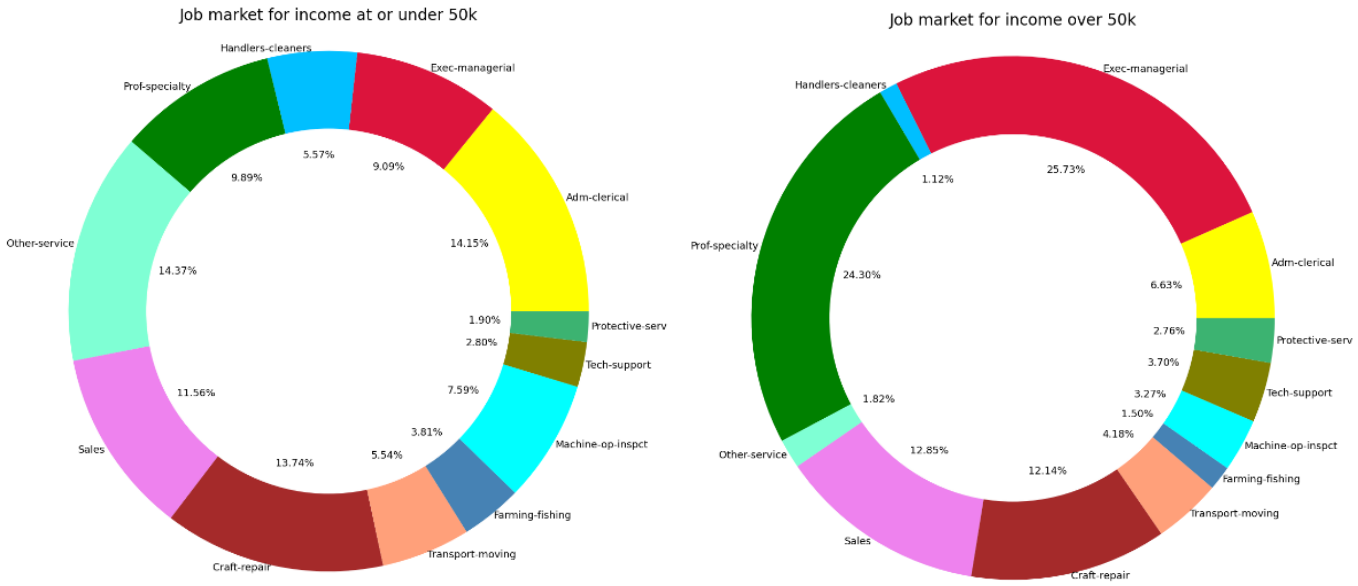

Figure 1. donut graphs for User Story 1.

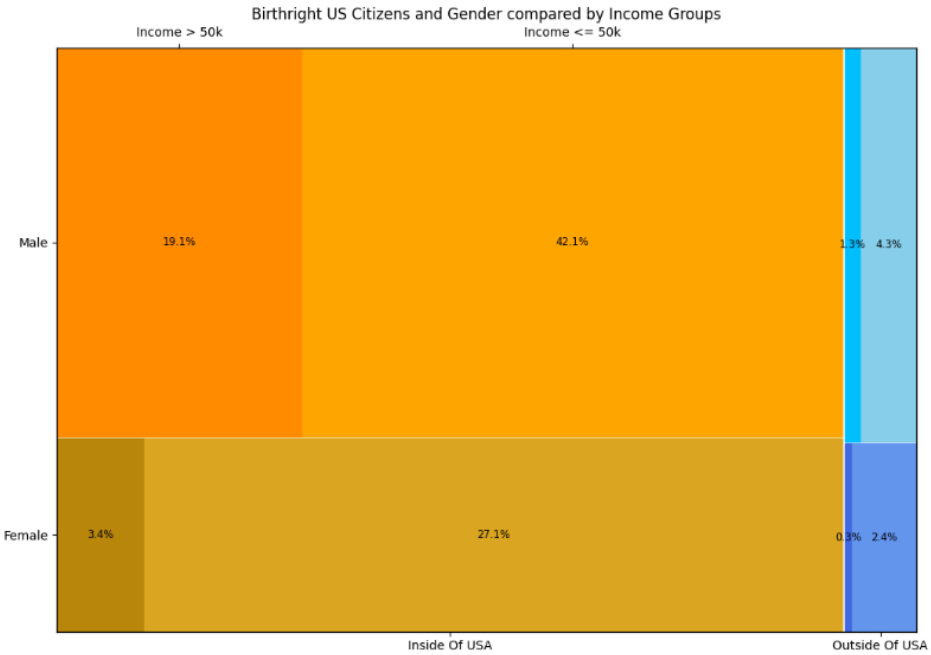

Figure 3. Mosaic Plot for User Story 2.

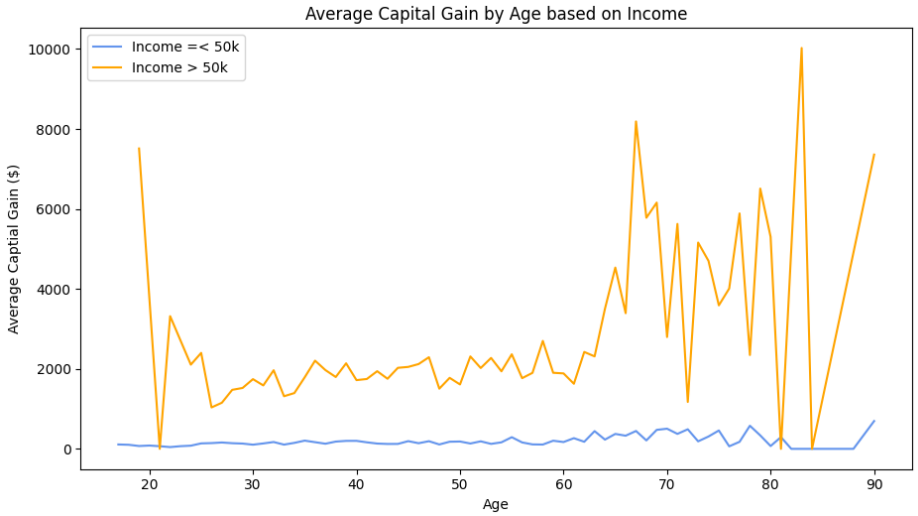

Figure 2. Line graph for User Story 3.

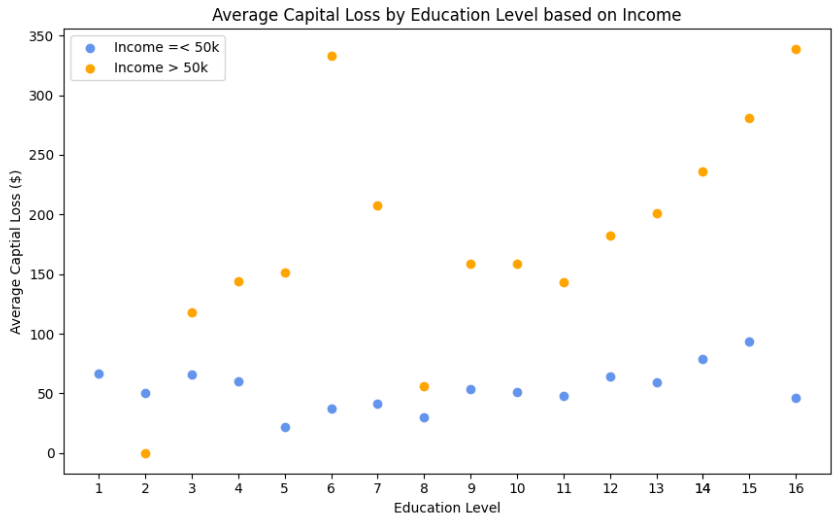

Figure 4. Scatter Plot for User Story 4.

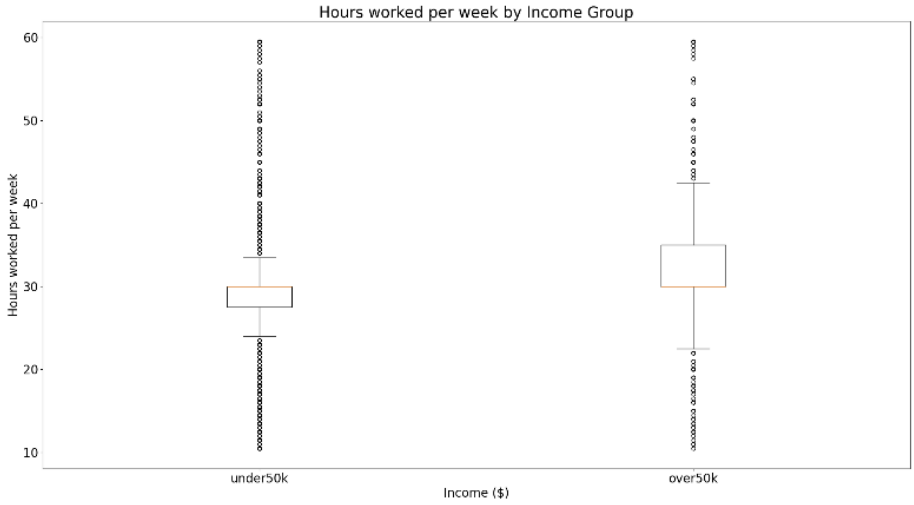

Figure 4. Box Plots for User Story 5.

V. QUESTIONS

1. How do we know a factor affects someone's income?
2. What data is clean and which data is dirty?

When looking at a graph it should be easy to understand the data but drawing conclusions from the data can be a little tricky. This is because it's hard to tell the difference between correlation and causation. When adding more variables like multivariable graphs, it becomes even harder, one variable could have no effect on income, but both together have a large impact on income. The way I handled this is by first looking at how the data affects income by itself, then in combination with another variable. By looking at induvial factors first you can see if there is a greater correlation together then when they are apart from each other.

Looking at data it's hard to tell what clean data is and what is not. To solve this, I start by removing the obvious problems, null data/empty data. The next step I took is removing outliers, any data that is two standard deviations away. Both these methods help clean the data to make sure the data doesn't unnaturally skew factors towards or way from affecting income.

VI. NOT DOING

Due to time constraints and the scope of this project I did not make a machine learning model. This, however, would make a great addition to the application. You could use supervised learning algorithms like Naïve Bayes to classify the data as above or below the target income. While graphing is a great way to see and understand the data a machine learning model would allow a more accurate way of finding predicting factors that could determine income.

VII. FACTORS

Throughout this paper we went over several factors including Age, Gender, Capital Gain and Loss, Occupation, Native Country, Education Level, Hour worked per Week and of course Income. Some of them had a clear correlation with income while others had some and others almost no correlation. The main factors that UVM can use to predict income would have to be Gender, Capital Gain and Loss, and Education Level. Capital gain and loss is an obvious one and could be explained by if you have more money, you can invest more and therefore gain more, or if you make more money you can afford to invest more and lose it. Education is most likely UVM best metric, in the data you can find that the higher education you are more likely to make over the target income and have higher capital gain, this would be a great selling point to increase enrollment. While Gender was a clear factor that showed being male increase your chance of making more than the target income whether born in the US or not, I won't touch on that in this paper because I think it would need more research and expertise.

Some factors that were not so clear would-be Occupation, Age, Native Country, and Hours worked per week. While some occupations did have more people making above the target income it, the majority had an even distribution of making above or below the target income, which make since because the median household income in the US was around $50,000 when this data was captured. Age, I thought would be a clear factor for predicting income because you have longer time to make more money but the distribution of people making more than target income stayed even though out all ages. Then there was almost no effect when looking at people who were born outside the US. When taking in to account that there is more data for people born in the US, people born outside the US are making the same as US born citizens.

REFERENCES

[1] "Python Sort Dictionary by Key – How to Sort a Dict with Keys," freeCodeCamp.org, May 25, 2023. https://www.freecodecamp.org/news/python-sort-dictionary-by-key/.

[2] "statsmodels.graphics.mosaicplot.mosaic - statsmodels 0.15.0 (+213)," www.statsmodels.org. https://www.statsmodels.org/devel/generated/statsmodels.graphics.mosaicplot.mosaic.html (accessed Feb. 23, 2024).

[3] "Choosing Another Color Palette for a Mosaic Plot," Stack Overflow. https://stackoverflow.com/questions/61704718/choosing-another-color-palette-for-a-mosaic-plot (accessed Feb. 23, 2024).

[4] "Python | Sort Python Dictionaries by Key or Value," GeeksforGeeks, Jul. 24, 2018. https://www.geeksforgeeks.org/python-sort-python-dictionaries-by-key-or-value/.