Seeing is Recruiting: Visualizing Income Factors for Targeted Enrollment

Austin Hudgins

alhudgins@ASU.edu

1. 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 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.

1. Assumptions
2. The data is correct.
3. 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.

1. User Stories
2. *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 makes up what portion of the job market for incomes equal to or 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 then $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 label of the graph and the values as the data that makes up the percentage.

1. *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 someone’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 each gender and if they were born in the US make above and below the target income price. 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 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 either sex. The total number of each occurrence of the key is the value for said key and then used statsmodel library to graph it.

1. *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 gains changes every year 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 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 was the maximum of 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.

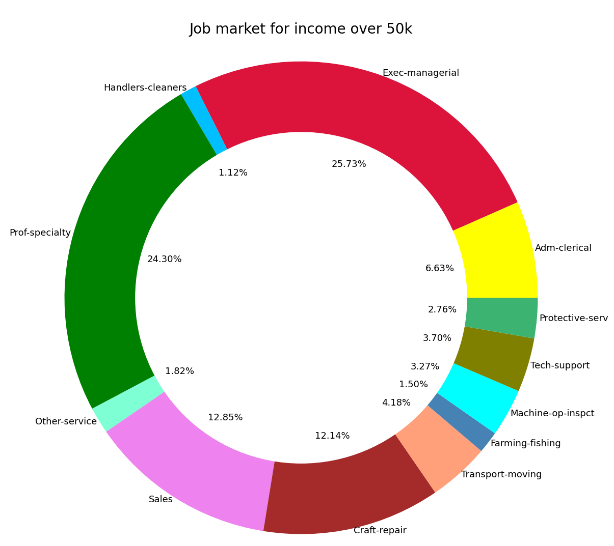
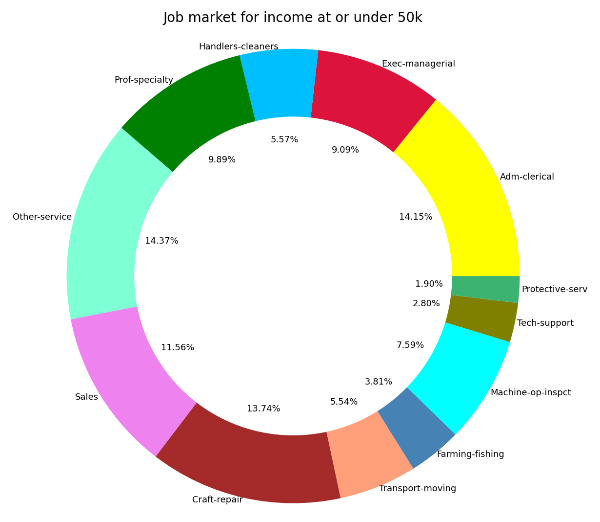
1. *As a member of the UVW Marketing Team, I want to know if working for a education level can affect your capital loss based on your income level so that we know what type of Jobs and Education we should target our advertisements too.*

.

1. *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.*

.

1. Visualizations



*Figure 1. donut graphs for User Story 1.*

*A screenshot of a graph

Description automatically generated*

*Figure 3. Mosaic Plot for User Story 2.*

A graph of a graph

Description automatically generated with medium confidence

*Figure 2. Line graph for User Story 3.*

*A graph of a graph with numbers and points

Description automatically generated with medium confidence*

*Figure 4. Scatter Plot for User Story 4.*

*A white paper with black lines

Description automatically generated*

*Figure 4. Box Plots for User Story 5.*

1. Questions
2. How do we know a factor affects someone’s income?
3. 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 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.

1. 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 dsata a machine learning model would allow a more accurate way of finding predicting factors that could determine income.

1. Factors

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/.