Arda Kocaman
1231699432

Ira A. Fulton Schools of Engineering
Arizona State University

CSE 578: Data Visualization

Project Final Report

Abstract

This report outlines the process of developing marketing profiles based on income data for the UVW College. The data shows various factors influencing an individual's income such as age, education level, how many hours they work, marital status, capital gain, etc. After the data was cleaned for analysis, some visualizations supported the user insights.

I. PROBLEM STATEMENT

Knowing the prospective customer's profile can help companies make targeted offers. The US Census Bureau supplied data for adults that could help us gain some insights on what factors determine the salary threshold of $50k. Like all real-world data, this data also has wrong and missing values that could be treated to utilize safe analysis. The dataset is then transformed or aggregated to be prepared for the analysis. The analysis includes trends, correlations, statistical descriptions, etc. The features analyzed on their distribution as well as their relation to each other and how it affects salary. Visualizing these analyses is crucial to comprehend the insights that they offer and makes them more digestible.

II. DATA PREPROCESSING

The analysis process started with getting to know the data with an initial exploratory analysis. It has 32561 data points with 15 features. The last feature is salary which we are interested in finding out. There are both numerical and categorical variables with the salary being a binary class (<50k and >50k). Salary is not perfectly balanced but still acceptable, with %76 of the data points belonging to <50k. For the analysis part, only *fnlwght* parameter seems unimportant, as it has a very high percentage of unique values and of course the given meaning of it.

III. ASSUMPTIONS

First of all the data used for the analysis is assumed to align with the business context. The target customers of the UVW College represent similar patterns with the US Census data. Another assumption is the statistical significance of the data, meaning there are enough data points for inference. There are more than 32k rows and pretty low missing values for most of the columns, so it's assumed that the conclusions from the analysis are statistically significant. Understanding the bias in the dataset is crucial. Some features like sex are unevenly distributed, so during analysis the ratio should be considered instead of the number of data points.

IV. DATA TRANSFORMATION

Before the analysis, some features need to be cleaned or transformed. The dataset does not have missing values but instead, some parameters have *?* values. However, this is not detrimental as there are only three columns that have it, and the rate of the unknown is around 5% for *workclass* and *occupation* and only 1.8% for *native-country.* The unknown values are deleted because they are low in percentage. The *native-country* feature also has other unknown values that do not belong to any known country but since their number is so low they are also deleted.

Some categorical values are being aggregated for the analysis. This made the analysis simpler, easier to interpret, and left out the data sparsity. For example, the *native-country* columns have many country values that were hard to analyze and visualize. The countries are grouped into their continents to have some more general insight. Also with the *hours-per-week* parameter the case was similar, there are hours between 1-99 but the data seem to show a Gaussian distribution with the mean of 40 hours per week. So this feature was split into two from the mean point.

V. ANALYSIS

Both numerical and categorical features were examined among the 14 available. A Pearson correlation was used to assess the correlation between features, visualized through a heatmap. To analyze the correlation between features and salary, the data type of the salary column was converted from object to binary. Numerical features such as *capital-gain*, *education*, and *hours-per-week* appear to have the most significant impact on the salary threshold. Conversely, categorical features required encoding to determine their correlation with salary. *Marital status* and *gender* were found to have the most positive effects, with being married and male having greater influence.

For univariate analysis, we can look at the distribution of the parameters. The most influential ones, age and education, have different semantics. *Age* is continuous between 17-90, so grouping into bins and plotting the histogram for the dataset shows the distribution correctly. Dividing into 12 bins yields 6-year bins. It appears that there are more young adults than older ones. Young adults are pretty steady until the age of 40, around 750 individuals per age.

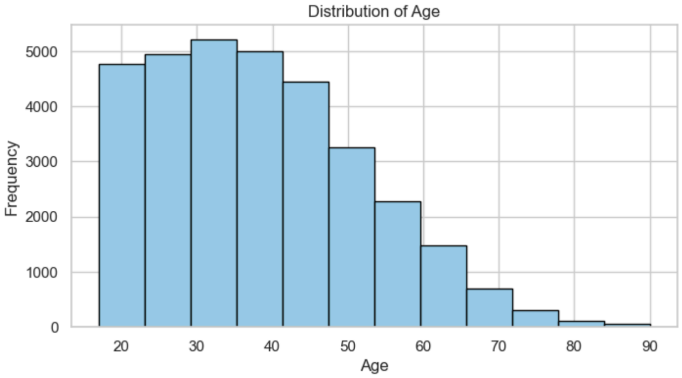

Figure 1: *Age* distribution histogram

For education level, the most frequent one is high school grad, following college and bachelor's. The further and lower education levels are less common and becoming less as they increase or decline. Since education is categorical, a line graph explaining its frequency in the dataset is sufficient.

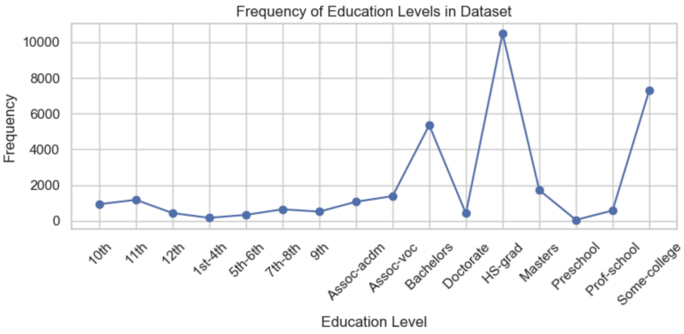

Figure 2: *Education* distribution line chart

Multivariate analysis was done in order to find out one or more feature's effects on the salary. The first analysis is the individual's *sex* vs. *salary*, and visualizing this on a mosaic plot reveals that men proportionally earn more than >50k.

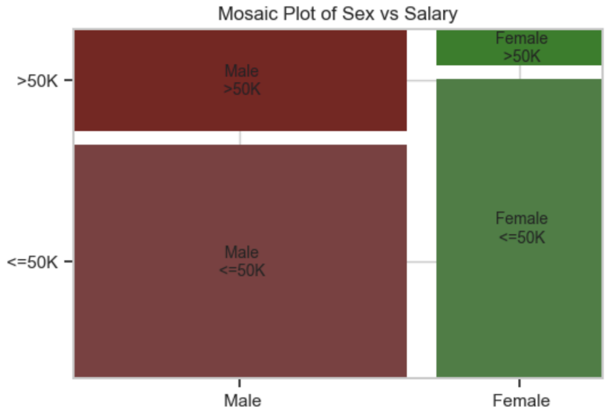

Figure 3: *Sex* vs *salary* mosaic plot

Another factor that was highly correlated with *salary* was *marital-status*. A stacked bar chart can show us what proportion of each martial class is earning. The chart shows that being married to a civil spouse proportionally is more probable than any of the statuses by a high margin.

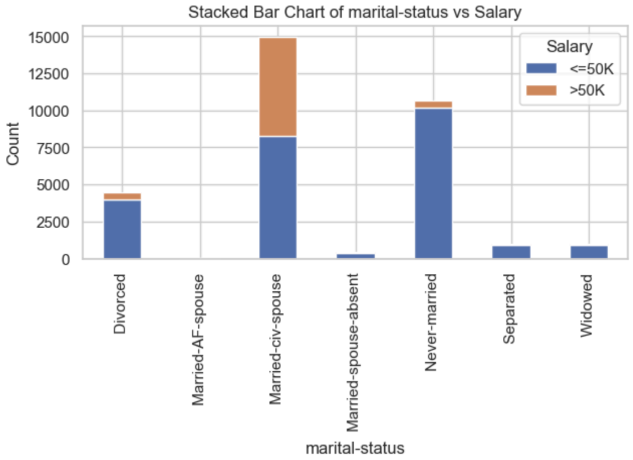

Figure 4: *Marital status* vs *salary* stacked bar chart

Age and educational level vs salary was an interesting multivariate plot. A scatter plot was chosen to visualize this relationship. On the y-axis, there is education level and on the x-axis there is age. There is a clear trend on the education level and the number of green (>50k salary) data points, and a much weaker trend on the age. On the top right of the plot, the greens are even greater than the red ones.

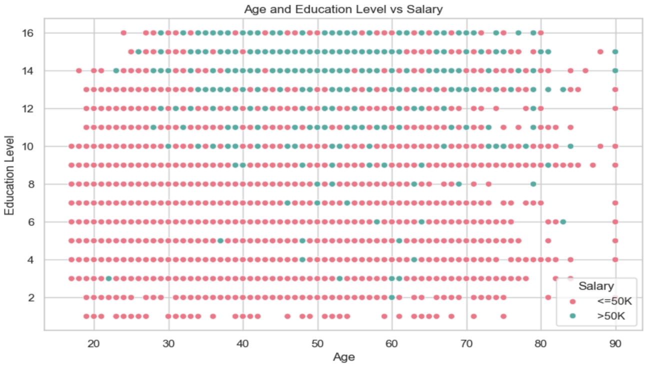

Figure 5: *Education-number* and *Age* vs *Salary* scatter plot

The occupation and relationship features verify these insights. The occupations that require higher education levels like being a professor or older age requiring occupations like executives are the ones that make the most more than $50k.

The next analysis was based on the *native-country*. The aim is to find out how backgrounds from different countries affect the income. Salary is transformed into a ratio, the number of rows above the data points over the whole, grouped according to the country. A choropleth graph was plotted. As the color of the country goes yellow, means more representative of that country earned above the threshold, as it goes to blue it's below. Iran backgrounded immigrants earned above the threshold the most by ratio, whereas central Americans fell below the threshold.

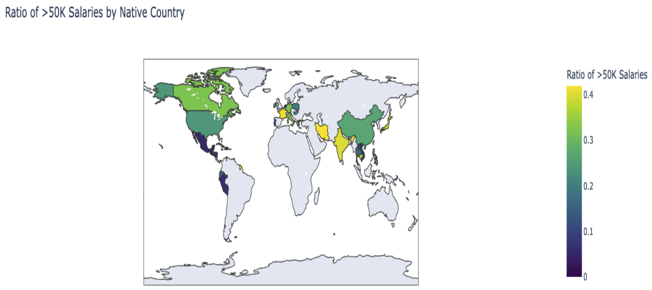

Figure 6: *Native Country* cloropleth plot

The last analysis answers a broader question, how do *education-number*, *hours-per-week*, and *capital-gain* was affect salary? These features are particularly chosen because their effect on income is bigger, an analysis would answer both the interconnection with each other and provide how strong their effect on income is. Since this accounts for three features it's not easy to visualize with the traditional ways. To tackle high-dimension and also a high number of data points for a proper representation, a 3D and interactive plot is used. python's plotly library enables interactive visualizations a user can rotate, zoom, and hover over to see the particular data point's exact value. Especially around 10k capital gain and 40 hours per week, many data points piled up, so zooming in to see in detail helps. According to the plot, it seems there is almost no one who makes less than $50k on the level of $100k capital gain. Education number is highly distinctive as expected, as the level increases the rate of high incomes is increasing. Hours of work do not have this distinct difference, as both red and blue data points seem to be scattered around without a clear trend.

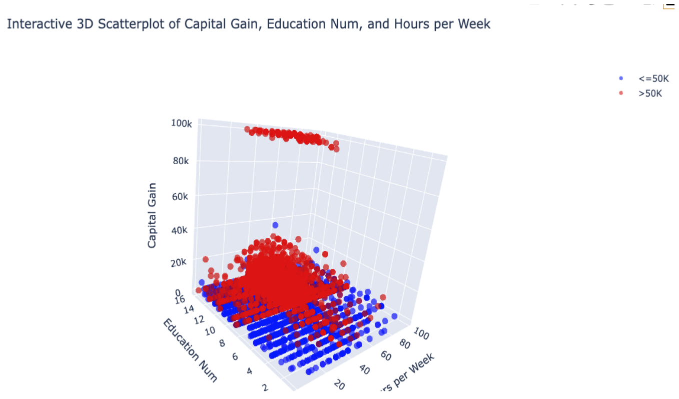

Figure 7: *Hours per Week* and *Education Number* and *Capital Gain vs Salary* scatter plot

VI. CHALLENGES

The first challenge was the data cleaning process. There was indeed missing data as "?" and there should be a decision made for them to fill it with 0 or delete it. Because they're low in size, I decided to delete them.

The second challenge was the data transformation process. While checking the correlation, I realized that more than half of the features including the salary were categorical, meaning an encoding was required to measure the correlation. I did one-hot-encoding for salary because there is no hierarchy among values that can be plain and simple one and zero.

Another challenge was the data sparsity for visualizations. I tried to solve this issue by aggregating the values into a higher hierarchy. For example, in the *native-country* feature, there are more than 10 countries and the visualization does not fit the screen. So mapping them to their continents, I was able to generalize them on the plot. Unfortunately this fell short as there are not enough number of country that could generalize the continents. This violates the assumption that visualizations and insights are representative of general, so a choropleth plot according to the countries is drawn.

VII. FURTHER WORK

This work can be further continued with prediction models. This task would be binary classification. This analysis gives a good base for starting it. As there are 14 features, some may be eliminated to overcome overfitting. There seem to be collinear features like education and education level. Some other transformations may be required, for example, data normalization for scaling the values between 0 and 1. After evaluating the model's accuracy with metrics like AUC score, the marketing department uses this model's output before communicating with prospective customers.

Also, there could be monitoring and data augmentation processes carried out. Analytics and prediction are ongoing business processes that need to be monitored regularly. If the analysis seems off after it's deployed to the business, action could be taken to change the data source or reconsider the assumptions. This analysis can be further improved by adding current student's protection. Of course privacy and data protection issues should be addressed, and some techniques with data hashing for student privacy may be a required step in the preprocessing.

VIII. CONCLUSION

This analysis started with the question 'How a data that shows various features of adults help us to gain insights about an individual's income level?'. This question was answered by extracting insights from feature's effect on each other and of course their effect on salary. The columns used in the analysis with the salary were: *age, education, sex, maritial status, education number, native country, capital gain,* and *hours per week.* Given all the insights and visualizations, some user stories can be concluded for this work:

1- There is a stable number of young adult individuals (<40) that peaked in early 30s. After that, there's a constant drop. It has an intermediate to strong effect on salary.

2- The level of education is skewed into high school, college, and bachelor's. Lower than high school and higher than bachelor's are getting less and less as it moves to the end.

3- Men are more likely to earn more than $50k than women.

4- Couples earn more than singles. Especially married to a civil spouse has the highest rate among all. Combined with the previous story, as expected, married men (also the husband in relationship status) have the highest rate.

5- As an individual gets older into their adulthood, and their education level gets advanced, the probability of earning more than $50k drastically improves. Other related feature values like occupation being a professor or executive support this.

6- Individuals who have more capital gain are more likely to earn $50k. The trend is clear, especially over the $40k capital gain.