Data Visualization Project Report

Name - Shashank Navad, ASUID - 1229407428, snavad@asu.edu

A. Goals and business objectives

Discovering hidden patterns and connections between various dataset features through data visualization is becoming more and more helpful in predicting challenging events. For a huge dataset to be visually appealing, these relationships and patterns are essential. An adult income dataset—a popular unbalanced machine learning dataset—was employed for this project.
The purpose of this project is to use a dataset including demographic data to develop a profile for a college marketing team. The college will be able to meet its enrollment target with this profile. The correlations between an individual's various demographic factors will be found in this report. These results will be used to construct a compelling and educational profile that will assist in suggesting a main demographic group for college enrollment.

B. Assumptions

- Data Accuracy -

I assume the information I've been given to be precise and correct. It is essential that the data be error-free and presents accurate information free from false information. However, if I don't know why the data is being used, making sure it's accurate and precise could become difficult or more costly than necessary.

- Collection, Timeliness, and Validity of Data -

I make the assumption that the methods used to acquire the data are unbiased and that the data is processed consistently without affecting other occurrences. Furthermore, there aren't many gender alternatives in the provided dataset, and open-ended responses aren't permitted. Notwithstanding these restrictions, I believe the dataset to be legitimate and verified.

- Reliability -

I make sure that our dataset is not in conflict with any outside data sources. Furthermore, I firmly believe that none of the data we have will be disclosed to outside parties without the explicit approval of the relevant users.

- Feature Selection -

I assume that the characteristics that have the most effects on class prediction will be able to give me more interpretable patterns. I thus concentrate on unbiased features in the majority of my studies and visualizations.

C. User Stories, Visualizations and questions

1) User story 1 - As a member of the UVW marketing team, is gender a relevant factor in determining the income label?:
To analyze the categorical variable gender I created two functions.

Visualization and questions - One of the functions is designed to analyze a single categorical column and produce a pie chart to visualize the distribution of categories within the column based on income. The insight gained through this function is that those with income greater than 50K are predominantly male with 85% men and only 15% women. While for the people in less than 50K category there are 62% men and 38% women. We can gather from this data that the majority of the data contains data from men however there is a disparity in income greater than 50K with a higher proportion of men in that category.
Question - Is there a disparity in income among people belonging to different genders?
To investigate this we can then create a second function that analyzes per unique value. Initially, the function finds all unique values in the specified column. After obtaining the unique values, the function visualizes the distribution of income classes for each unique value. For each unique value, it creates a new pie chart to represent the distribution of income classes. Additionally, it calculates the count of individuals with income above and below $50K for each unique value and displays this information on the pie chart. Through this function we find that there is indeed a gender disparity between those with a higher income class than a lower income class. Amongst the males a total of 69% with a total count of 13984 turned out to have incomes

less than 50K and 31% with a total count of 6396 turned out to have incomes greater than 50K. Amongst the females a total of 89% with a total count of 8670 turned out to have incomes less than 50K and 11% with a total count of 1112 turned out to have incomes greater than 50K. Through this we can discern that the majority of people with an income greater than 50K are men confirming that there is a gender disparity.

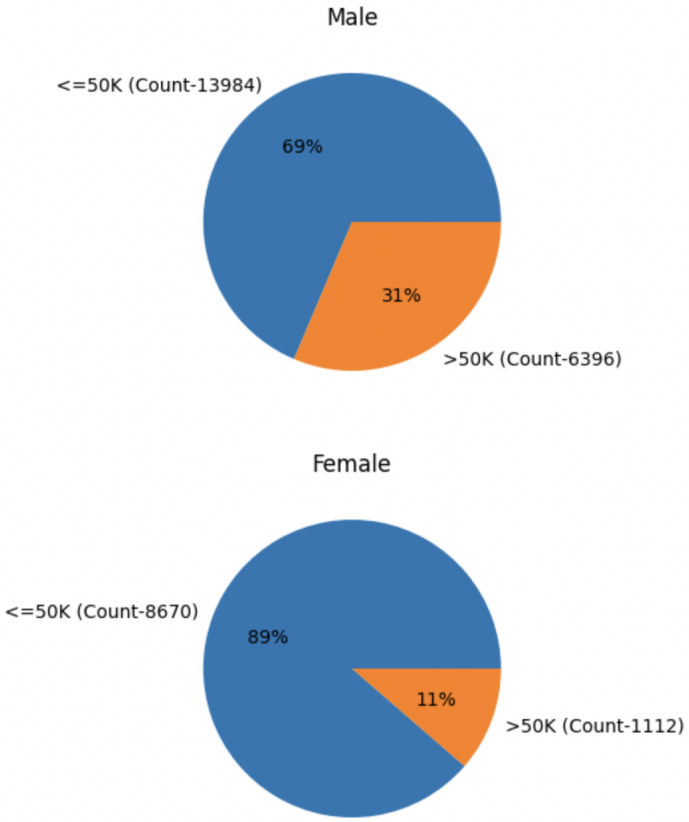

Fig. 1. Pie Chart analyzing gender per unique value

2) User story 2 - As a member of the UVW marketing team, is occupation a relevant factor in determining the income label?: Visualization and questions - A mosaic plot is a graphical representation used to visualize the relationship between two or more categorical variables. It divides a rectangle into smaller rectangles, with each area proportional to the frequency of the combination of categories. Mosaic plots are especially useful for exploring the association and interaction between categorical variables.

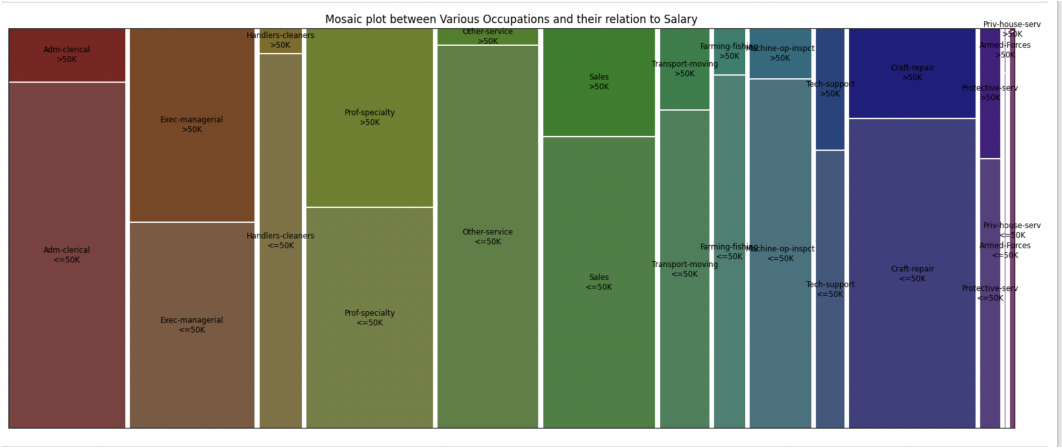

Fig. 2. Mosaic plot between Various Occupations and their relation to Salary

Question - Is there a disparity in income among people belonging to different occupations?
Some valuable insights were gained by using this visualization. Most people with clerical occupation turned out to have an income below the 50K threshold. Among managers and professors the distribution between the two classes was uniform. Most tech-support and sales, craft-repair professionals had an income lesser than the 50K threshold.

3) User story 3 - As a member of the UVW marketing team, how does age, marital-status and hours-per-week affect the income label and how are they interrelated?:
Visualization and questions - To examine and evaluate whether these variables are interrelated we use the scatter plot visualization. The function creates scatter plots for three pairs of columns of the DataFrame, distinguishing between two income classes. The function initializes the plotting area and then loops through each pair of columns to create a scatter plot. For each pair, it assigns blue color to instances where the 'class' column is 'less than or equal to 50K' and yellow color to instances where the 'class' column is 'greater than 50K'. This function is valuable for visualizing the relationship between different pairs of columns in a dataset, especially when comparing how each pair relates to the 'class' column.

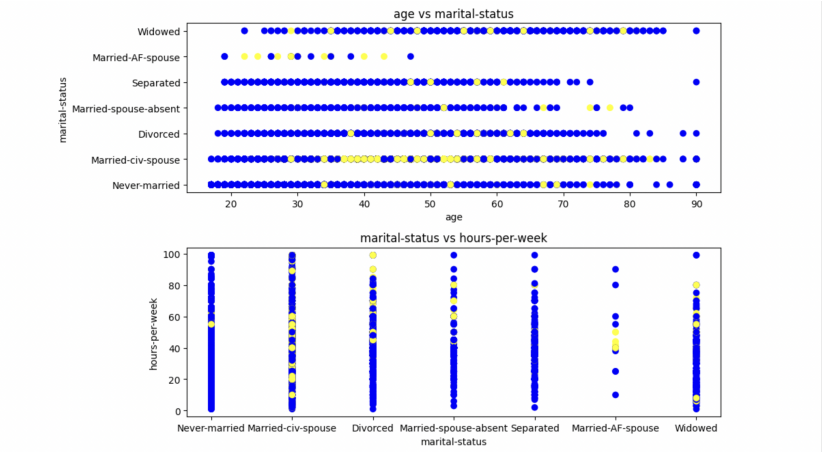

Fig. 3. Scatter plot between age, marital-status and hours-per-week Part 1

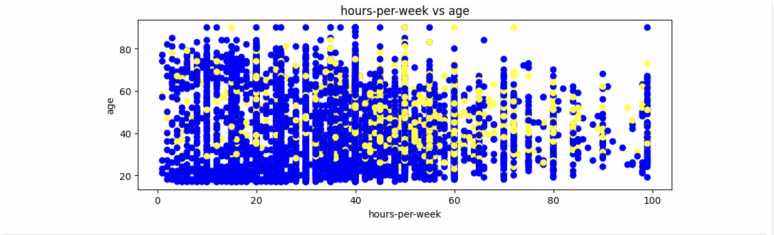

Fig. 4. Scatter plot between age, marital-status and hours-per-week Part 2

A total of three plots are constructed one of which is the age vs marital status and their corresponding income plot. Some of the insights that are gained through this are that people who are in the greater than 50K income category have an age between 35 and 60 and come under the married to a civilian spouse category.
Question - How are the variables age, marital-status and income interrelated?
The second plot maps marital status vs hours per week and their corresponding income. Through this plot we find that barring a few outliers, regardless of the number of hours of work per week the people in the never-married category find themselves in the income class below 50K. Most people with an income greater than 50K in the married with civilian spouse category tend to work between 40 and 60 hours. The people who are divorced and have an income greater than 50K tend to work longer hours ranging between 50 and a 100 hours. We can observe a similar distribution among those with higher income who are married with an absent spouse with their working hours ranging from 50 to 80 hours. Among those who are married to a person belonging to the armed forces we find the people with a higher income tend to work about 30 to 50 hours per day.
A third plot was constructed comparing hours-per-week with age. As expected people with a younger age tended to work less hours a week and have an income lower than 50K. Most people with a higher income class tended to be between the ages of 40 and 60 and worked more than 40 hours per week.

4) User story 4 - As a member of the UVW marketing team, how does education number, age and capital-gain affect the income label and how are they interrelated?: To do this we create two functions. One to analyze the contents of the capital-gain variable and another to create a parallel coordinate plot of the three variables.

Visualization and questions -
Question - How is the capital-gain variable distributed?
The first function computes and displays the statistical summary of the capital-gain column. It calculates and prints the mean, median, and standard deviation for both the ”less than or equal to 50K” and ”greater than 50K” classes. Following this, it creates a visualization including box plots and histograms for the same column, allowing for a visual comparison of the data distribution between the two income classes. It is observed that in the capital-gain column that those with a class ”greater than 50K” tended to have a much higher capital-gain with the mean for the higher income class coming up to approximately $3937.68 while those with the ”less than or equal to 50K” class had a mean capital-gain of approximately $148.89.

The second function normalizes the ’education-number’, ’age’, and ’capital-gain’ features using Min-Max scaling. After scaling, it selects a random sample of 50 instances for both the ”less than or equal to 50K” and ”greater than 50K” classes. These samples are combined into a single DataFrame. Finally, it generates a parallel coordinate plot to visualize the relationship between ’education-number’, ’age’, and ’capital-gain’, with the lines colored according to the ’class’.

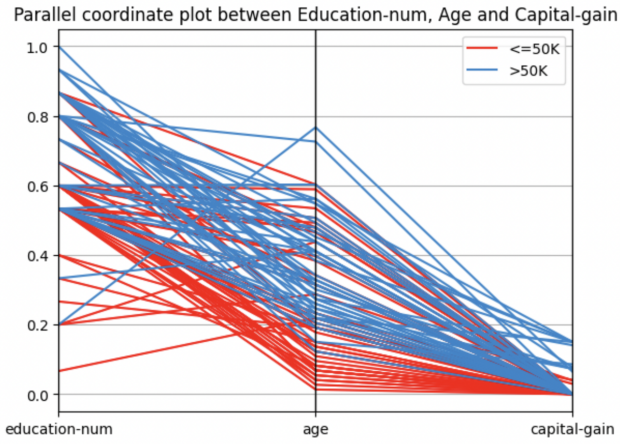

Fig. 5. Parallel coordinate plot between Education-number, Age and Capital-gain

As expected those with a higher education tend to have a greater proportion of people in the greater than 50K category. Those with this kind of higher education also tend to be older and have a higher amount of capital-gain. In this way the relationships between the three variables were established using a parallel coordinate plot.

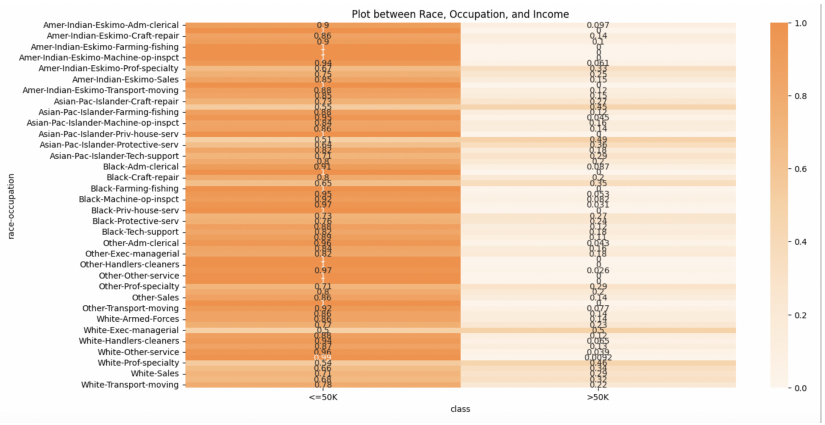

Fig. 6. Plot between Race, Occupation, and Income

5) User story 5 - As a member of the UVW marketing team, how are occupation and race correlated?:
Visualization and questions - To analyze the two variables occupation and race I created a heat-map shown above using the

'seaborn' library.
Question - What are some target customers for the marketing team?
The disparity between the races can be highlighted in the figure as well as some target customers for the marketing team. In the Asian-Pacific Islander category we find that those in the craft and repair profession as well as the private housing category have about 50% of people in the 'greater than 50K' category. Other such examples include White Executive-Managers and Professors who also have about 50% of people in the 'greater than 50K' category.

D. Future work

1. For my next steps, I plan to delve into machine learning modeling after visualizing the data and identifying patterns. I aim to explore building predictive models to forecast enrollment likelihood or target specific demographic groups more effectively. I'll be considering classification algorithms like logistic regression, decision trees, and random forests to classify individuals based on demographic factors.
2. I'm also interested in conducting segmentation analysis to identify distinct groups of individuals within the dataset based on demographic characteristics. This analysis will provide deeper insights into the diversity of our target audience and help tailor marketing strategies to different segments.
3. As part of my future work, I'll extend my analysis to predictive modeling by forecasting enrollment numbers or predicting the likelihood of an individual enrolling in college based on their demographic attributes. This may involve time-series analysis or building regression models to predict enrollment rates accurately.
4. Another aspect I'm keen to explore is feature engineering. I'll delve into creating new features from existing ones or incorporating external datasets to capture more information about individuals. This will help enhance the predictive power of our models and provide more accurate insights.
5. Additionally, I plan to develop interactive visualization tools or dashboards to allow stakeholders to explore the data and insights more dynamically. This will involve using tools like Tableau or Plotly to create interactive charts and graphs that enable users to drill down into specific aspects of the data.

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.
[2] CSE 578 Lecture Slide/Videos.