Course Project Final Report CSE578

Jeremy Escobar
School of Computing and Augmented Intelligence
Arizona State University
Tempe, USA
jgescoba@asu.edu

***Abstract*— In a contemporary society, data plays a crucial role in enhancing our lives, guiding decisions, discovering solutions, among other things. In CSE 578 Data Visualization taught me how to employ various forms of visual representations, such as charts and graphs, to simplify the analysis of patterns and trends. I utilized these tools for a range of analyses, including market research, customer insights, and income trends. This report will showcase my efforts in generating diverse visualizations to aid XYZ Corporation in formulating marketing tactics derived from the United States Census Bureau. Therefore, enabling UVW College to create marketing strategies to increase enrollment based on the visualized data.**

Keywords*— *visualization, analyses, marketing.

I. GOALS AND BUSINESS OBJECTIVES

This initiative aims to identify key factors impacting individuals' earning potential using visual aids, drawing on data from the United States Census Bureau. My role as Data Analyst involves collaborating with XYZ Corporation to enhance UVW College's enrollment. I am specifically tasked with focusing on the pivotal indicator of whether an individual's income falls above or below the $50,000 threshold. My objective is to produce visual representations that the marketing department can use to target these specific demographic segments effectively.

XYZ Corporation specializes in crafting detailed marketing visuals based on data analysis, subsequently offering these profiles to various companies for targeted marketing efforts. As part of my responsibilities at XYZ, I have a new project with UVW College, which seeks to bolster its enrollment numbers. UVW College has pinpointed income, particularly the $50,000 benchmark, as a key demographic marker to guide the marketing of its academic programs. My mission is to develop insightful marketing visualizations using information from the US Census Bureau, with a special focus on identifying the income bracket as a determinant factor.

II. ASSUMPTIONS

Constructing a comprehensive data visualization from a dataset with numerous variables necessitates several critical assumptions to accurately convey the data.

a. Initially, I must trust in the authenticity of the dataset provided by the United States Census Bureau. The foundation of any visualization lies in genuine data; otherwise, discrepancies can distort the intended narrative, presenting a misleading picture.

b. Another key presumption is the completeness of the data. Any holes or missing information can compromise the integrity of the visualization, preventing it from fully representing the chronicle.

c. Accuracy of the data is paramount. Erroneous data can result in inaccurate representations, leading to incorrect conclusions and potentially unexpected visual outcomes.

d. The timeliness of the data is also essential. For the visualization to be relevant and valuable to XYZ Corporation, it must reflect the most current data available, enabling informed decision-making.

e. Lastly, it's vital that the data encompasses a diverse range of individuals. Diversity ensures that the analysis is not skewed towards a particular group, offering a balanced view that can cater to a broader audience and provide a more accurate evaluation.

III. User Stories

User Story 1. **Age and Income Analysis for UVW College Marketing Initiatives:** In partnership with XYZ Corporation, my role is to assess the influence of age on income levels, focusing on the pivotal $50,000 mark. Utilizing SQL to extract age and income data from the Census Bureau's database, I created a histogram that visualizes the age distribution within different income brackets. This visual tool will assist UVW College in developing marketing strategies tailored to age groups most likely to benefit from further education to achieve or exceed an annual income of $50,000.

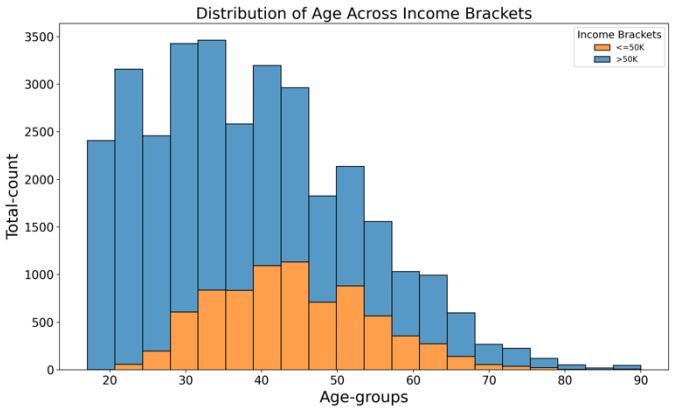
User Story 2. **Marital Status & Education, Influence on Earnings for Targeted Marketing:** My analysis is centered on the impact of marital_status on earnings, with the goal of identifying trends that can inform UVW College's marketing campaigns. By using a mosaic plot, I investigate the intersection of education and marital status, providing UVW College with insights to tailor marketing messages to demographics that are more likely to earn below or above $50,000 based on marital status and education level.

User Story 3. **Age & Capital Gain Trend Predictors for UVW College's Enrollment Strategy:** Exploring capital_gains, and age as potential predictors of income, I aim to inform UVW College's enrollment strategy by identifying the factors that determine earning potential. A scatter plot analysis allows us to understand how these variables interact, enabling UVW College to target its marketing to individuals who could benefit from further education.

User Story 4. **Native Country, Income Analysis for UVW College's Marketing Profiles:** Focusing on the link between native_country and income levels, this user story aims to categorize professions by income brackets to aid UVW College's market segmentation. Pie charts based on native country origins excluding the United States highlight the income distribution among different demographic groups, providing a basis for UVW College's targeted marketing strategies.

User Story 5. **Sex, Race and Education num's Impact on Income for UVW College's Target Demographic:** This analysis assesses how varying levels of education_num, sex and race correlate with income. By generating a parallel coordinates plot, I provide UVW College with a visual tool to discern which educational qualifications impact earning potential, enabling them to focus their marketing efforts on demographics that would benefit most from higher education to surpass the $50,000 income threshold.

IV. Visualizations

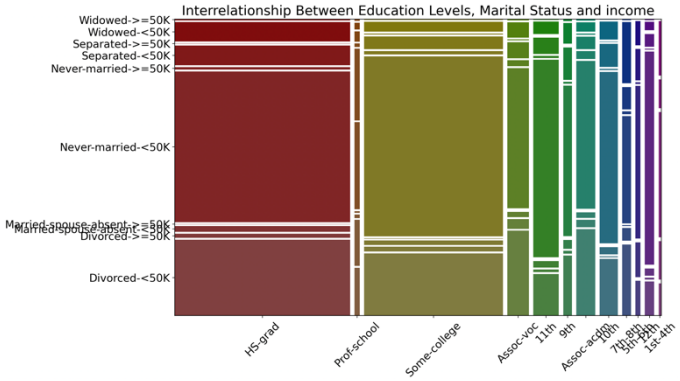


User Story 1 - Bar (univariate):

Description: In this analysis, I explore how age, and income play a role in determining whether individuals earn above or below $50,000 annually.

Procedure: Using pandas, I extract and cleaned age and income data from the adult_data database, then applied matplotlib and seaborn to create a color-coded histogram with sns.histplot, visualizing age distribution across income brackets. This approach concisely illustrates the age-income relationship through visual analysis.

Conclusion: The histogram provides a comprehensive visual representation of the age distribution across various income levels, emphasizing the crucial role age plays in earning potential. It vividly highlights the concentration of certain age groups within specific income brackets, suggesting a potential correlation between age and income. For instance, it may show trends such as a higher concentration of younger individuals in lower income brackets and a gradual shift towards higher income brackets as age increases. It also opens avenues for deeper exploration into how age-related factors such as experience, education, and professional development opportunities contribute to an individual's financial success. Aiding UVW College in developing marketing strategies tailored for certain age groups.

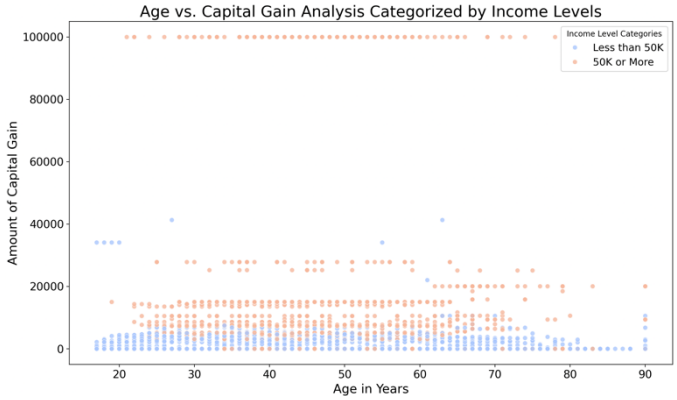


User Story 2 - Mosaic Graph (multivariate):

Description: This analysis explores the connection between education levels and marital status, and how these factors interrelate.

Procedure: I start by extracting data on education and marital status from the *adult_data* database, excluding higher education levels to focus on more in-complete forms of education. Using *pandas* for data preparation and the *statsmodels* library for visualization, I created a mosaic plot that illustrates the relationship between the two variables. This plot method allows for a clear visual comparison of different education levels against marital statuses.

Conclusion: The mosaic plot reveals a pattern in the association between education levels and marital statuses, suggesting how these factors may interact to influence societal roles and expectations. The visualization highlights the complexity of these relationships, offering insights into the diverse educational backgrounds within marital status categories. Allowing UVW College great insight into potential students to market too.

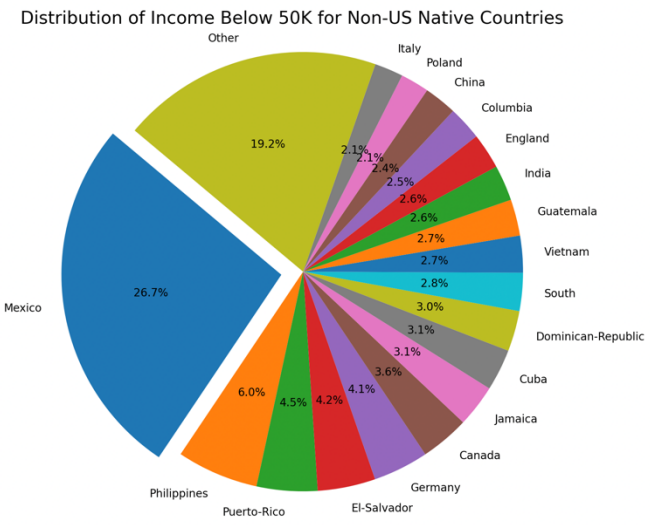


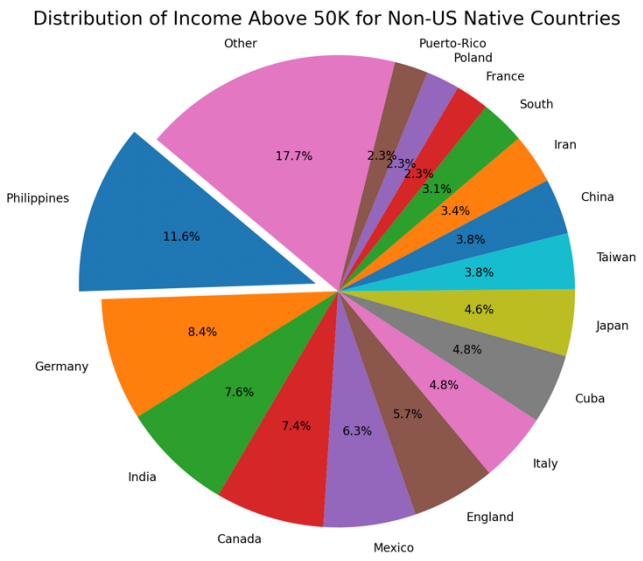
User Story 3 - Scatter Plot Graph (multivariate):

Description: This analysis investigates how working age, capital gain, and income interact to influence an individual's income.

Procedure: I begin by extracting data on hours per week, capital gain, and education number from the *adult_data* database, ensuring all records are complete. Utilizing *pandas* for data handling and *matplotlib* along with *seaborn* for visualization, I created a scatter plot. This plot maps capital gains against age, with a color gradient representing income, to visually explore these variables' interplay and their collective impact on income potential.

Conclusion: The scatter plot vividly demonstrates the complex relationship between an individual's age, capital gains, and income, revealing insights into how these factors may correlate with earning potential. The color-coded representation of income adds depth to the analysis, suggesting that age may correlate with increased capital gains and potentially higher income, especially when combined with longer capital gains.



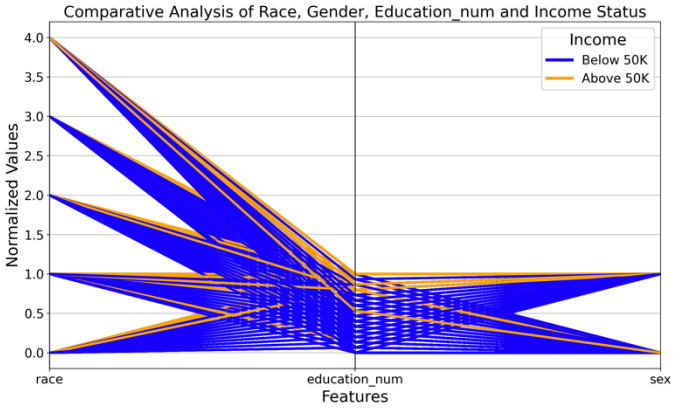


User Story 4 - Pie Chart Graph (2x univariate):

Description: This analysis examines the income levels of American citizens based on their ancestral origins, focusing on the distinction between those earning above and below $50,000, excluding data from individuals of United States origin to highlight diversity.

Procedure: I extract information on the ancestral origins and income levels from the adult_data database, specifically excluding individuals identified with the United States as their native country to emphasize the variety of backgrounds. Using *pandas*, income data is categorized into two groups: below and above $50,000. For visual clarity and to manage the diversity of origins, I consolidate fewer common origins into an 'Other' category if they constitute less than 2% of the dataset. Pie charts are then generated for each income category using *matplotlib*, illustrating the proportion of individuals in each income bracket by their ancestral origins.

Conclusion: The pie charts provide a visual breakdown of income levels among American citizens with diverse ancestral origins, excluding those from the United States. This approach sheds light on the economic integration and success of different ethnic groups within the U.S., revealing patterns that might assist UVW College in their marketing efforts. For example, Mexican decent appears to be a good demographic to target earning income below $50k.



User Story 5 - Parallel Coordinates (multivariate):

Description: This investigation examines the interplay between race, education level, gender, and income, aiming to uncover how these factors collectively influence earning potential in the United States.

Procedure: I retrieve data encompassing race, education number, sex, and income from the *adult_data* database, ensuring no entries are missing. To facilitate analysis, categorical data are encoded numerically, and education numbers are normalized to scale these values between 0 and 4.0. Utilizing *matplotlib* and *pandas*, a parallel coordinates plot is constructed, mapping each factor against income levels to visually represent their relationships. Colors differentiate income categories, offering immediate insight into how these variables intersect to affect earnings.

Conclusion: The parallel coordinates plot vividly contrasts the dynamics between race, education, gender, and income status, revealing nuanced patterns of socio-economic stratification. This visualization method effectively highlights disparities and trends, suggesting that a complex interplay of demographic factors contributes to income variation among American citizens, giving UVW College the necessary visual data it can use to make marketing easier.

V. QUESTIONS

What visualization techniques are appropriate for categorical and continuous data?

- Based on the attributes I selected for analysis, certain visualization techniques are more appropriate due to the nature of the data either continuous or categorical. For categorical data, bar charts, pie charts, and mosaic plots are effective. Conversely, line charts, scatter

plots, and histograms better suit continuous data. This distinction was emphasized during our coursework.

- I really wanted to try and find more graphs and Ideas to use for displaying the data, but my challenges were easily overcome by using what I learned in the course versus trying stuff I knew nothing about. So sticking to what I learned proved to be highly beneficial for the project in the end.

What tool did you use for data visualization in Python, and why?

- To construct the necessary visualizations, I employed Matplotlib, Seaborn, Pandas, Sqlite3 and Statsmodel in Python, all great tools for data visualization and graphical plotting, which enhanced the presentation of my findings.
- I had issues with legends not correctly showing the colors for income correctly, I had to correct the colors manually. I was having issues with parallel coordinates not displaying the data correctly, so I added print statements to fix my issues through debugging. Then I was finally able to compute the data accurately for display.

Why did you choose PyCharm for this project?

- For this project, despite the availability of various environments, I opted for PyCharm due to its simplicity and cross-platform compatibility, which facilitated the development process. It also has a ability to view the database, quickly and efficiently without having to write code to view tables and columns almost instantly.
- I tried using Jupyter notebook but it failed because I was on a Mac and I looked into VSCode as well, but I found that PyCharm's built in SQL database reader was the best option for me.

How did you manage data extraction and filtering in Python?

- Data extraction and filtering from the *adult_data* file was accomplished using *Pandas*, a powerful 2D data structure ideal for managing tables divided into rows and columns. The comma-delimited data was effectively organized into separate arrays within Pandas, allowing for efficient manipulation and analysis.
- Many instances when I viewed the data on the graphs it was too much data. I had to for instance begin filtering out some the data to prevent there being too much information. examples are the mosaic plot, I excluded people with higher education and in the pie charts I bunched together data that was less than 2%. Without applying these filters and more, the visual data would have not been a good representation for UVW College to see.

VI. Not Doing

In the future, I plan to explore areas of data science that I have not yet delved into. One such area is data mining, which involves the extraction of patterns and knowledge from large datasets. By integrating data mining techniques, I aim to uncover deeper insights and trends that are not immediately apparent. This will involve learning and applying complex algorithms to sift through data effectively, a skill set that will significantly enhance my analytical capabilities in CSE572.

Additionally, I intend to further harness machine learning capabilities within Python to refine predictions and enhance the accuracy of visual data engineering. This ambition includes adopting a more meticulous approach to data selection and visualization beyond the methods I currently use, such as mosaic plots, bar histograms, and parallel coordinates. I plan to incorporate advanced visualization techniques, such as network graphs, which I have not used previously. These techniques will allow for more granular filtering based on variables like educational attainment, ensuring the relevance and precision of the insights generated. This progression towards more sophisticated data analysis and visualization methods will enable me to produce more accurate and insightful visual representations of complex datasets. Therefore, XYZ Corporation using me as an avenue to produce marketing pieces that will aide in a company's marketing success long-term.