CSE: 578 Data Visualization Project Progress Report

Sahil Yogesh Hadke | 1229679960 | shadke1@asu.edu

Problem Statement

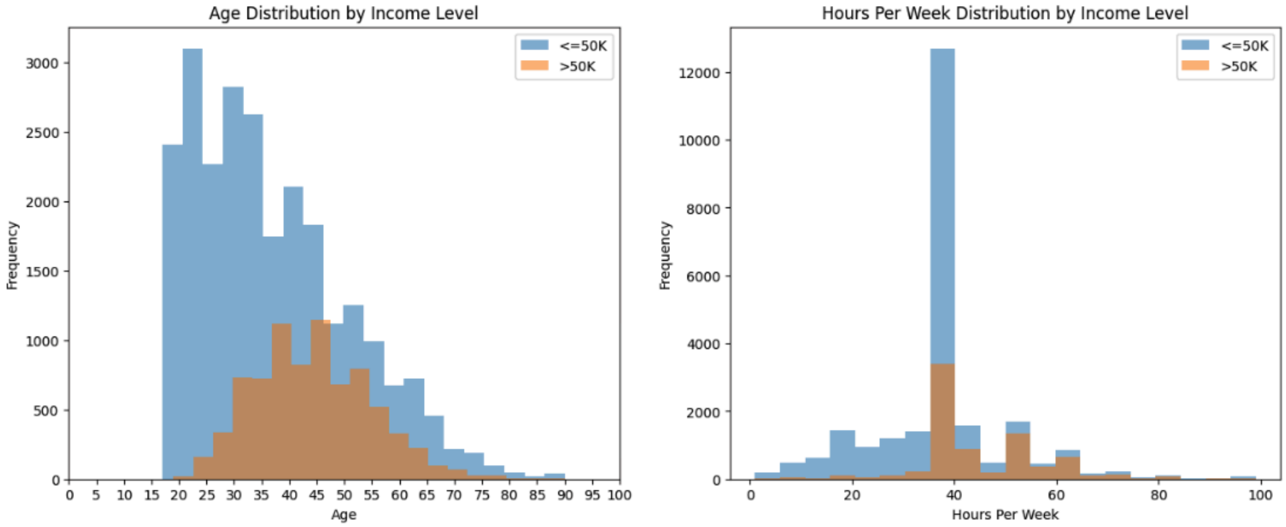
The problem statement revolves around creating visualizations for user stories to develop targeted marketing profiles for UVW College, ultimately increasing enrollment by tailoring marketing efforts to specific demographic groups based on income as the key demographic. The college seeks to leverage data from the United States Census Bureau to identify key demographic factors that influence an individual's income, including age, gender, education status, marital status, and occupation. By analyzing these factors, the goal is to understand characteristics and plot charts to analyze the earnings of individuals earning less than and more than $50,000 annually and tailor marketing efforts accordingly.

The challenge lies in efficiently analyzing the vast amount of Census Bureau data to extract meaningful insights that can inform marketing strategies. This requires a comprehensive understanding of statistical analysis methods, and data visualization techniques to identify trends and patterns in the data.

Progress Made

The project has made substantial progress, establishing a solid foundation for further development. Through the acquisition of advanced Python libraries like Pandas, NumPy, and Matplotlib, the groundwork for efficient data manipulation and visualization has been laid, signaling readiness for practical work with the dataset. A fundamental understanding of the dataset's categorical and non-categorical data types is essential for effective data preprocessing and analysis. Moreover, a clear comprehension of the problem statement and dataset columns is crucial for identifying relevant factors influencing income and devising targeted marketing strategies.

Two key charts, a histogram and a stacked bar chart, have been finalized along with corresponding user stories. These visualizations promise valuable insights into income distribution among different demographic groups. For instance, the age distribution histogram indicates a larger curve for ages 35-45, suggesting that individuals in this age bracket are more likely to earn over $50,000 annually. Conversely, the histogram for individuals earning less than or equal to $50,000 annually is left-skewed, indicating that more people in their early 20s and 30s earn this amount. Additionally, the histogram for hours per week distribution with income as a hue suggests that individuals working around 40-60 hours per week are more likely to earn over $50,000 annually. Notably, the income distribution by workclass chart reveals that more than 50% of self-employed individuals earn over $50,000 annually, contrasting with other workclasses where more than half of the people earn less than or equal to $50,000 annually.



Above are two of the histograms I have plotted to have insights on distribution of income.

Issues or Challenges Encountered

The project encountered several challenges that required innovative solutions and continuous learning. Data preprocessing posed a significant hurdle, particularly in handling missing values and determining which attributes to retain for analysis. This step is critical as it directly impacts the quality and reliability of the insights derived from the data.

Adjusting figure size, colors, and scales on axis to enhance the readability of charts was another challenge. While seemingly trivial, these adjustments are crucial for ensuring that the visualizations effectively convey the intended message and are easily understood by the audience.

The categorical nature of many variables in the dataset also presented a challenge. Categorical variables require specialized plotting techniques to visualize effectively, which required time to research and implement.

Plan to Resolve Issues

To address the challenges encountered, a structured plan has been devised. For further data preprocessing, the focus will be on exploring more techniques to handle missing values and selecting the most appropriate attributes for analysis as per different charts. This will involve further research and experimentation to ensure that the data is cleaned effectively and only relevant variables are retained.

Improving the readability of charts will require a deeper understanding of visualization principles and continued experimentation with different color combinations, figure sizes, and axis settings. This will involve studying the theory behind effective chart design and applying these principles to create visually appealing and informative visualizations.

In terms of choosing user stories and plotting charts, the plan is to continue with the iterative process of trial and error until the most suitable plots are finalized. This will involve refining the user stories to ensure they align with the project's objectives and selecting charts that effectively communicate the insights derived from the data.

Finally, studying the theory behind advanced plots such as pie charts, doughnut charts, mosaic plots and heatmaps will be a priority. This will involve researching these techniques in depth and exploring how they can be applied to the project to gain deeper insights into the data.

Tasks To Be Completed and Detailed Plan to Complete

The tasks to be completed for the project include refining existing charts, for example, by plotting subplots of histograms for all numerical values in the dataset to provide a comprehensive overview of their distribution and relationship with income. Additionally, exploring additional chart types such as mosaic plots, doughnut charts, parallel coordinate plots and various bar chart variations will help visualize categorical variables and their impact on income, enhancing the depth and breadth of insights gained from the data. Further study and research will be conducted to identify additional user stories that can provide valuable insights into income trends and demographics, enabling the development of more targeted marketing strategies for UVW College.

To ensure that the project is well-documented for future reference and replication, the entire process, including data preprocessing steps, chart development, user story creation, and insights gained, will be documented. Finally, the refined charts will be implemented, additional chart types will be explored, and the entire process will be documented according to the detailed plan to ensure that all aspects of the project are completed in a timely and efficient manner.