PROJECT MILESTONE-1

Name: Nishal Chandra Reddy **ASU ID:**1226003456

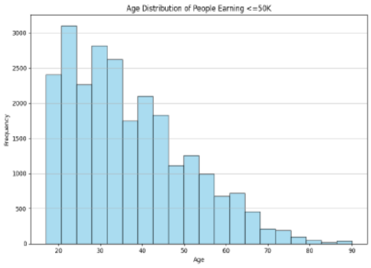
Problem Statement: As a data analyst at XYZ Corporation, one of my responsibilities is working with UVW College to examine a dataset that includes financial and demographic data. Based on income levels, target audiences for UVW College's marketing courses are to be identified. Dataset made available by the US Census Bureau. To comprehend the traits of the target group, we choose pertinent criteria and create powerful visualizations. With the use of this data, UVW College will be able to develop focused marketing plans that will successfully promote its programs and courses to a diverse range of demographic and income groups. Factors such as age, race, education num, capital gain, native country, hours-per-week, sex, education and work class are important considerations.

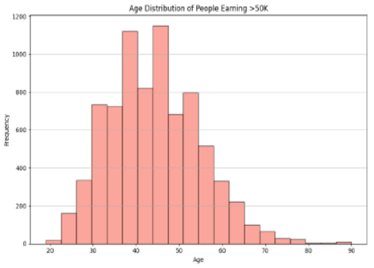
User Stories:

1. One of my responsibilities as a data analyst at UVW's XYZ is to examine the age distribution of people who make $50,000 or less. The purpose of this research is to determine prospective age groups that UVW College should focus on in their marketing campaigns.
2. As a data analyst at XYZ for UVW, it is my responsibility to examine how earnings in various racial communities—both less than and more than $50,000—are distributed. The demographic makeup of prospective target audiences for UVW's courses will be better understood thanks to this investigation.
3. My job at UVW as a data analyst at XYZ is to examine the capital gain and schooling histories of people making more than $50,000 and less than or equal to $50,000. Marketing various courses according to the educational attainment of these income groups would be made easier with the use of this study.
4. Understanding the correlation between years of education, marital status, and hours worked per week for those making less than $50,000 annually is a key responsibility of mine as a data analyst at XYZ for UVW. A target group for financial aid programs will be identified by UVW with the aid of this study.
5. I work as a data analyst at XYZ for UVW, where my job is to analyze weekly hours worked and work class of those who make over $50,000 against those who make $50,000 or less. The purpose of this report is to shed light on the target student groups' demographics for UVW marketing initiatives.

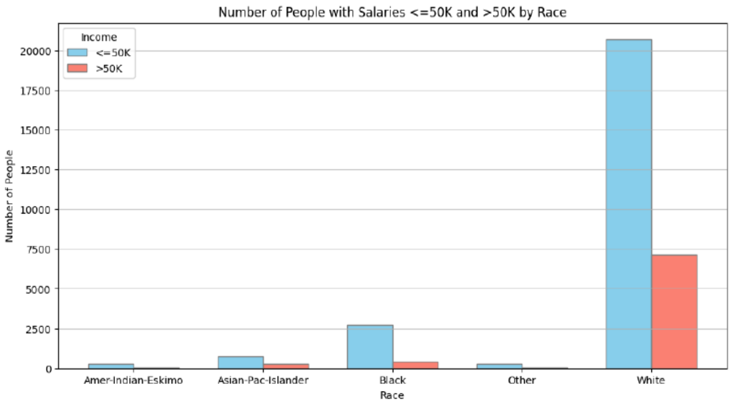
Progress Made: I carefully reviewed the problem statement and noted key attributes like age, race, education, capital gain, native country, hours worked per week, sex, and work class. The data was in a .data file format, which posed challenges. I used the pandas library to create a structured dataframe, eliminating extra spaces with the split function. Handling missing values in 'work class', 'native country', and 'occupation', which contained '?' values, was also necessary.

After cleaning and structuring the data, an analysis provided insights for the marketing team. A histogram for User Story 1 showed age distribution for income groups. Younger individuals peak between 20-45 years, more likely earning <=50K. Those earning >50K are mostly 35-55, suggesting a higher likelihood for middle-aged individuals.

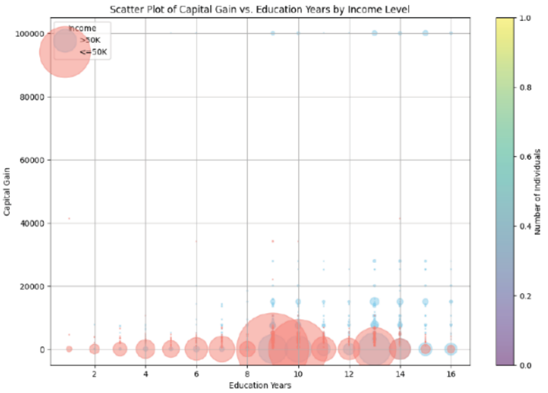




For User Story 2 after creating a stacked bar chart, I observed a stark income disparity among different racial groups. The chart indicated that while a significant number of white individuals earn more than 50K, other racial groups have fewer individuals in this income bracket. The majority of individuals from all races, except white, predominantly earn less than or equal to 50K. This visualization underscores the importance of addressing economic inequality through targeted policies and initiatives to ensure equal opportunities for all racial groups.



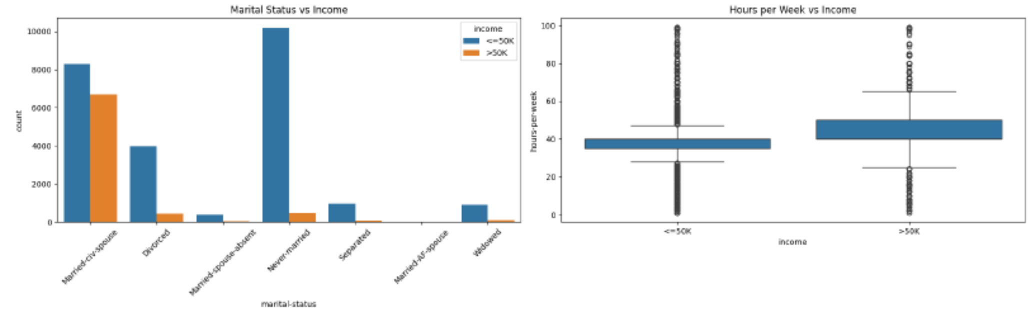
For User Story 3, I created a scatter plot showing capital gain and education years, color-coded by income level. It highlights trends like higher education correlating with increased capital gain and a concentration of lower-income individuals with less education. This underscores income disparity, with higher education and capital gain linked to earning over 50K. These insights could help tailor strategies, offering educational products to lower-income individuals and investment opportunities to higher earners, to better target the audience.



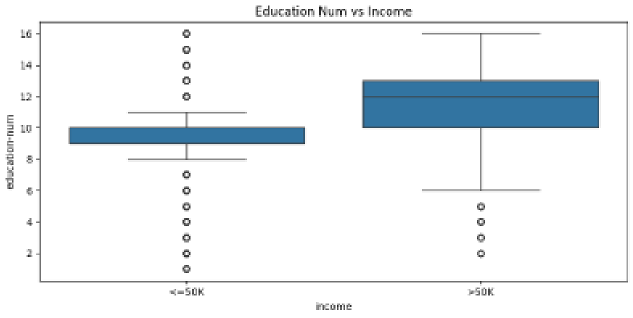
I completed User Story 4 by creating bar and box plots for hours worked per week, marital status, and years of education vs. income, revealing key insights:

• Marital Status vs. Income: Married individuals with civilian spouses tend to have a higher proportion of >50K income, suggesting a positive impact on income.

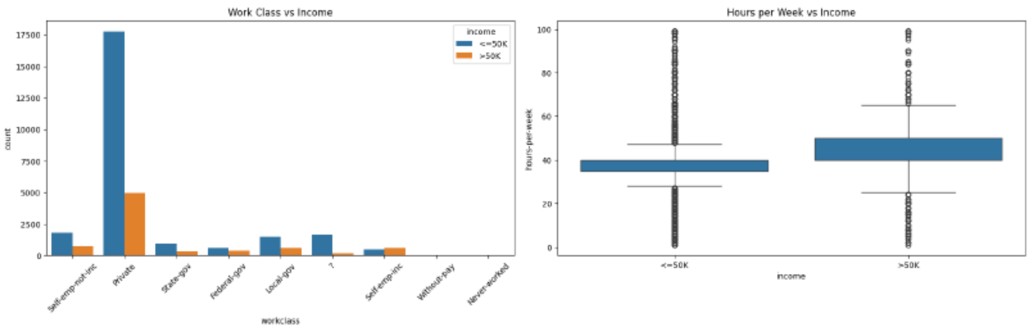
• Hours per Week vs. Income: Those earning >50K work more hours per week, indicating a link between higher income and longer work hours.



• Years of Education vs. Income: More education is associated with >50K income, highlighting education's role in higher income.



User Story 5 involved creating bar plots and box plots for hours-per-week, work-class, and income (salary) to understand their relationship. The bar plot for work class and income showed most people are in the private sector, with a smaller proportion earning >50K. This suggests income varies within work classes. The box plot for hours per week and income indicated those earning >50K work more hours, but some work long hours and earn <=50K. These insights can inform targeted marketing strategies.



Challenges Encountered: I had to learn data labelling and structure in order to work with a .data files, which was difficult. Complexity was increased by managing special characters and missing data. There was a significant salary bias in the statistics. It was difficult to organize the data programmatically and analyses the multivariate data for useful insights, particularly for scatterplots.

Solutions: I used pandas, NumPy, and Matplotlib to clean and arrange the data. Created useful data exploration tools such as histograms and stacked bar charts. Created a user-defined method to categories data based on salary, which will help with visualization in User Story 2. User stories one and two supplied UVW with meaningful data about age distribution and race-based marketing opportunities.

Future Plan: Even with the obstacles, I have made great strides toward my goal of identifying prospective target audiences for UVW College's marketing courses by analyzing the financial and demographic data. Next, I'll review user stories 1 and 2 to improve data visualization. To visualize properties with '?' replaced with None, I plan to employ oversampling and under sampling to identify appropriate groups for study. I hope to use self-calibrated numbers to better grasp grouping operations and examine which charts perform best based on course ideas. I plan to create additional user stories using mosaic plots and control graphs to identify qualities.