CSE578 Project Progress Report

Milind Parab

(ASU ID 1233342788)
Arizona State University
Tempe, USA

mgparab@asu.edu

***Abstract*— XYZ Corporation is helping UVW College increase enrollment by developing marketing profiles based on demographic data. Focusing on individuals earning $50,000 or more, using data from the U.S. Census Bureau, this report analyzes factors influencing income, such as age, gender, education, marital status, and occupation. The goal is to create targeted marketing profiles that will inform UVW College's outreach.**

I. Introduction

In today's data-driven world, targeted marketing strategies are essential for reaching the right audience. XYZ Corporation has partnered with UVW College to create detailed marketing profiles aimed at boosting enrollment. By focusing on income, specifically the $50,000 salary threshold, UVW College aims to identify prospective students for its programs.

Using a rich dataset from the United States Census Bureau, which includes demographic factors like age, gender, education, marital status, and occupation, the goal is to segment individuals into those earning less than and more than $50,000. These profiles will help UVW College tailor programs, tuition options, and delivery methods to meet the needs of each income group.

This report outlines the analysis process, development of marketing profiles, and creation of a predictive application to assist the marketing team in refining their strategies and targeting the right audience.

II. Problem Statement

UVW College aims to use income as a key demographic indicator for marketing its degree programs. The challenge is to identify and analyze the factors that distinguish individuals earning above or below $50,000. The goal is to develop an application that can:

1. Identify the factors influencing a person's income.
2. Segment individuals into two categories: those earning less than $50,000 and those earning more.
3. Predict an individual's income based on demographic variables to enable targeted marketing.

III. Dataset Summary

The dataset consists of **32,561 records** (individuals) and **15 columns** (features). We will focus on analyzing the following features after excluding the fnlwgt column and addressing missing values represented as "?" (which will be replaced with NaN):

Features:

Data Preprocessing Steps:

1. **Exclusion of fnlwgt**: The fnlwgt column will be excluded as it is used for statistical sampling and is not directly relevant for the analysis.
2. **Handling Missing Data**: Any "?" values in the dataset will be replaced with NaN, which will allow us to handle missing data more effectively using techniques such as imputation or removal.

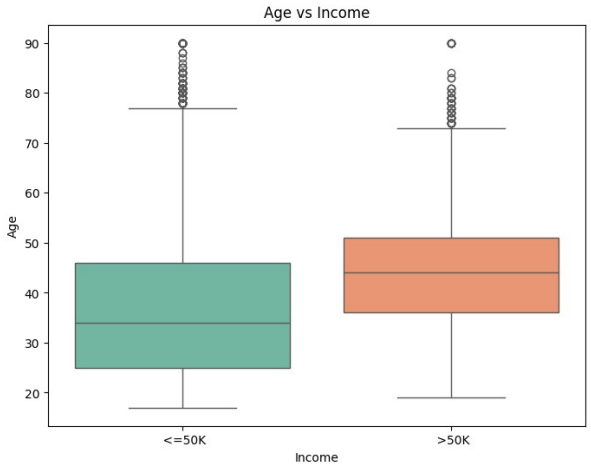
IV. Exploratory Data Analysis (EDA)

After data cleaning and preprocessing, we will perform **Exploratory Data Analysis (EDA)** to gain insights into the dataset, identify trends, relationships, and patterns. EDA is essential to understand the data before building any predictive models.

Case 1: "Examining the Relationship Between Age and Income"

Analysis of the Box Plot reveals a discernible relationship between age and income, indicating a tendency for older individuals to have a higher likelihood of earning over $50K.

1. Visualization: Box Plot - Age vs. Income (<=50K vs. >50K)



This box plot demonstrates the distribution of age across two income categories: <=50K and >50K. It visually compares the median, quartiles, and range of ages within each income group, highlighting potential differences in age distribution between those earning less than or equal to $50,000 and those earning more than $50,000.

2. Design Process:

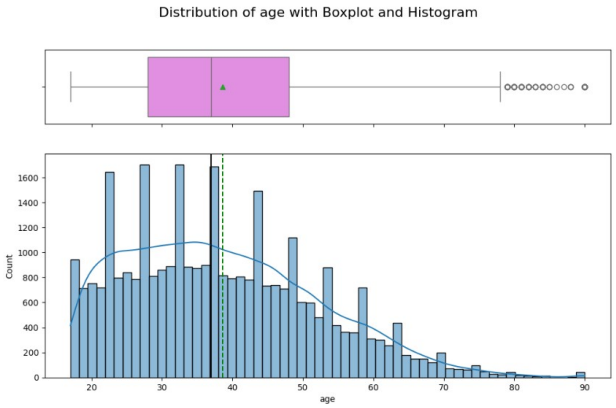
- Data Preparation:

The dataset was cleaned and organized to include age and income information for each individual.

The income variable was categorized into two groups: <=50K and >50K.

- Chart Selection:

A box plot was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the distribution of age across two income categories and to highlight potential differences in age distribution.



3. Conclusion:

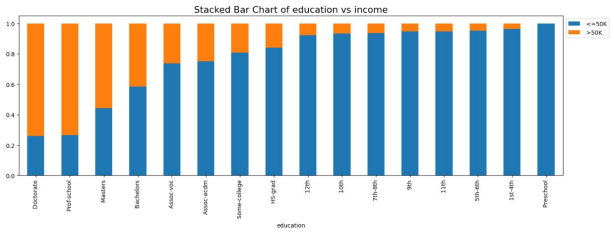
- The box plot effectively demonstrates a relationship between age and income. The median age for individuals earning >50K is noticeably higher than the median age for those earning <=50K, suggesting that older individuals are more likely to earn more.

Case 2: "The Impact of Education on Earning Potential"

Visualization: Stacked Bar Chart - Education vs. Income (<=50K vs. >50K)

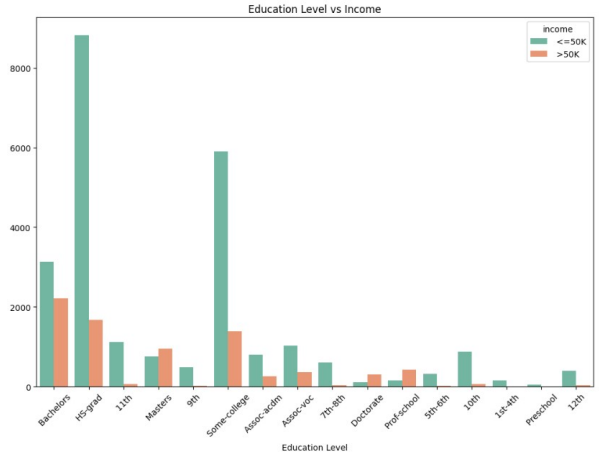
1. Explanation of the Visualization and Rationale:

- Demonstration: This stacked bar chart demonstrates the relationship between education level and income, specifically the proportion of individuals earning <=50K and >50K within each education category. It visually compares the distribution of income across different education levels.
- Rationale: The stacked bar chart was chosen to effectively show the composition of income groups within each education category. It allows for a direct comparison of the relative proportions of <=50K and >50K earners for each level of educational attainment. This visualization is ideal for showcasing how the distribution of income changes as education level increases. It's clear and intuitive, making it easy to understand the correlation between education and income.



2. Design Process:

- **Data Preparation:** The initial step involved organizing and cleaning the data to group individuals by their education level and income bracket (<=50K and >50K). The data was then aggregated to calculate the proportion of each income group within each education category.
- **Chart Selection:** A stacked bar chart was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the composition of income groups within each education category and to facilitate easy comparisons across different education levels.



3. Conclusion:

The stacked bar chart effectively demonstrates the strong positive correlation between education level and earning potential. The visualization clearly shows that individuals with higher education levels are significantly more likely to earn over $50K.

- The chart highlights the importance of education as a pathway to economic opportunity and higher income levels.
- The "inversion point" around the "Some-college" and "Assoc-acdm" categories provides a clear visual indication of the threshold where education begins to significantly impact income.
- The high proportion of >50K earners in the Doctorate and Prof-school categories underscores the economic benefits of pursuing advanced degrees.

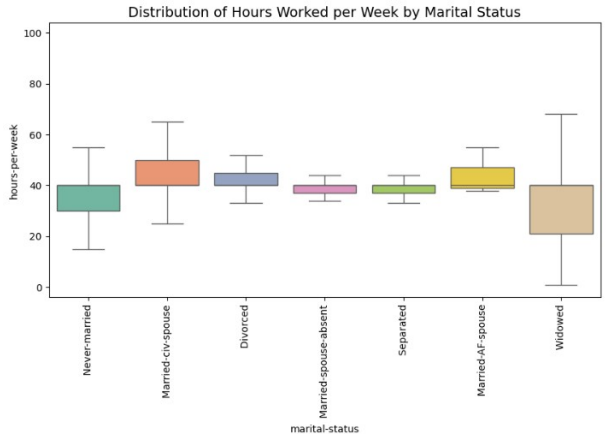
In conclusion, this stacked bar chart provides a clear and compelling visual representation of the relationship between education and income, making it a valuable tool for understanding the economic impact of educational attainment.

Case 3: "Exploring the Impact of Marital Status and Work Hours on Income"

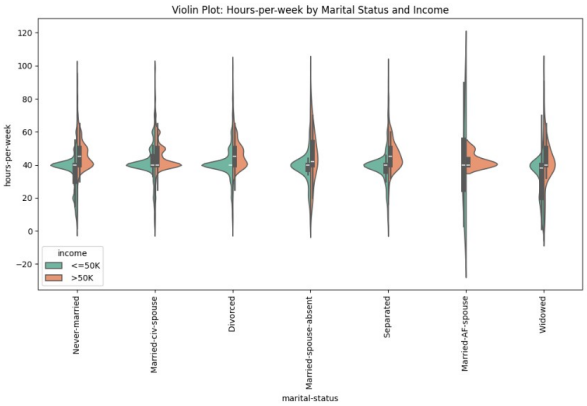
Visualization: Violin Plot - Hours-per-week by Marital Status and Income.

1. Explanation of the Visualization and Rationale:

- **Demonstration:** This violin plot demonstrates the distribution of hours worked per week across different marital status categories, further segmented by income levels (<=50K and >50K). It visually compares the shape and spread of the data, highlighting differences in typical hours worked and the range of hours worked for each group.



- **Rationale:** A violin plot was chosen to effectively visualize the distribution of a continuous variable (hours-per-week) across multiple categorical variables (marital status and income). It provides a more nuanced view compared to a box plot by showing the probability density of the data at different values. This allows us to observe not just the median and quartiles, but also the overall shape and skewness of the distribution for each group. The separation by income within each marital status category allows for a direct comparison of how income influences working hours within each marital status.



2. Design Process:

- **Data Preparation:** The initial step involved organizing and cleaning the data to group individuals by their marital status and income bracket (<=50K and >50K). The data was then aggregated to calculate the distribution of hours worked per week for each group.
- **Chart Selection:** A violin plot was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the distribution of a continuous variable (hours-per-week) across multiple categorical variables (marital status and income).

- **Axis Definition:** The x-axis was defined to represent the marital status categories, arranged for clear comparison. The y-axis was defined to represent the hours worked per week, ranging from -20 to 120 to encompass the entire range of values.
- **Violin Creation:** Violin plots were created for each marital status category, split by income level (<=50K and >50K). The width of each violin represents the density of data points at different values of hours worked per week.

3. Conclusion:

The violin plot effectively demonstrates the complex relationship between marital status, income, and hours worked per week. The visualization clearly shows that:

- **Married-civ-spouse Stands Out:** Individuals who are "Married-civ-spouse" and earn >50K tend to work significantly longer hours compared to other groups. This suggests a strong correlation between working longer hours and higher income for married individuals with civilian spouses.
- **Marital Status Influences Work Patterns:** The distribution of hours worked varies across different marital status categories, indicating that marital status plays a role in influencing work patterns.

In conclusion, this violin plot provides a comprehensive and insightful view into the interplay between marital status, income, and hours worked per week. It highlights the importance of considering these factors together to understand work patterns and income disparities.

Case 4 : Interesting Case: The "Workaholic Senior" Phenomenon

We're interested in exploring the relationship between age, hours worked per week, and income, specifically focusing on the potential for older individuals to work longer hours and earn more. We're particularly curious to see if there's evidence of a "workaholic senior" phenomenon, where older individuals continue to work long hours, potentially to maintain income or pursue personal fulfillment.

Visualization: Scatter Plot - Age vs. Hours-per-week by Income

1. Explanation of the Visualization and Rationale:

- **Demonstration:** This scatter plot demonstrates the relationship between age, hours worked per week, and income (<=50K and >50K). Each point represents an individual, with the x-coordinate representing age, the y-coordinate representing hours worked per week, and the color representing income. It allows for a visual exploration of potential trends and patterns between these three variables.
- **Rationale:** A scatter plot was chosen to effectively visualize the relationship between two continuous variables (age and hours-per-week) while also incorporating a categorical variable (income) through color coding. This type of plot is ideal for identifying potential correlations, clusters, and outliers. It allows us to observe whether there's a trend of older individuals working longer hours, and whether this trend is associated with higher income.

2. Design Process:

- **Data Preparation:** The initial step involved cleaning and organizing the data to include age, hours-per-week, and income information for each individual. The income variable was categorized into <=50K and >50K.
- **Chart Selection:** A scatter plot was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the relationship between two continuous variables and to incorporate a categorical variable.

3. Conclusion:

The scatter plot provides valuable insights into the relationship between age, hours worked per week, and income.

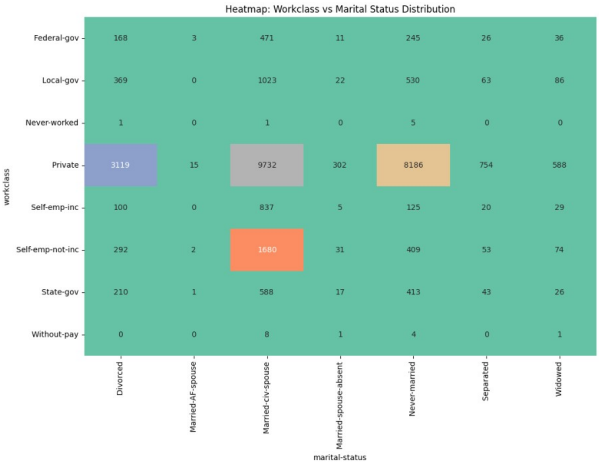
- **Higher Concentration of >50K at Older Ages:** There's a higher concentration of individuals earning >50K in the older age ranges, particularly above 60. This aligns with the understanding that older individuals may

have accumulated more experience and skills, leading to higher-paying positions.

- **Potential "Workaholic Seniors":** While not a dominant trend, there are individuals in the older age ranges (70+) who work significantly longer hours (60+ per week) and earn >50K. This provides some evidence for the "workaholic senior" phenomenon.
- **Spread of Hours Worked:** The hours worked per week are spread across a wide range for all age groups, indicating that various factors influence work patterns, regardless of age.
- **Income Distribution:** The distribution of income is not uniform across age groups. There's a tendency for the >50K group to be more prevalent in the older age ranges, but there are also individuals earning <=50K in these ranges.

Case 5: "Analyzing the Relationship Between Workclass and Marital Status"

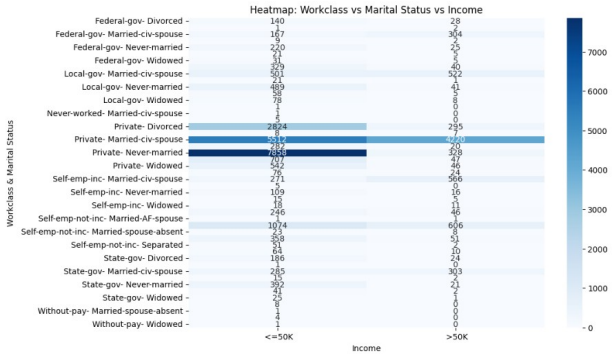
This heatmap reveals the distribution of individuals across different workclasses and marital statuses. The "Private" workclass dominates, with the highest counts, particularly for "Married-civ-spouse" and "Never-married" individuals. "Local-gov" and "Federal-gov" show significant numbers, with "Married-civ-spouse" being the most common. Self-employed categories have lower counts, while "Never-worked" and "Without-pay" show minimal representation. The heatmap highlights the prevalence of the "Private" workclass and the strong association between marital status and workclass distribution.



Visualization: Heatmap - Workclass vs. Marital Status vs. Income

1. Explanation of the Visualization and Rationale:

- **Demonstration:** This heatmap visualizes the distribution of individuals across different combinations of workclass, marital status, and income levels (<=50K and >50K). It displays the count of individuals in each specific category, with color intensity representing the magnitude of these counts.
- **Rationale:** A heatmap was chosen to effectively represent the relationship between three categorical variables. It allows for a clear and concise visual representation of the data, highlighting patterns and concentrations of individuals across different categories. The color gradient enables quick identification of categories with higher or lower counts, making it easy to compare and contrast different combinations.



2. Design Process:

- Data Preparation:

  The dataset was cleaned and organized to create a table with counts of individuals for each unique combination of workclass, marital status, and income.The counts were calculated for each category, which would be the numerical data represented by the heatmap.

- Chart Selection:

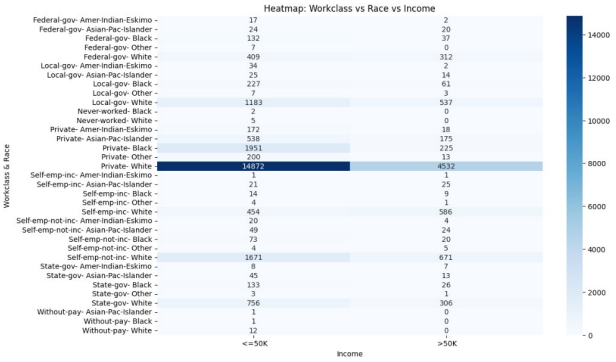
  A heatmap was selected as the most appropriate visualization due to its ability to effectively display the relationships between multiple categorical variables and the magnitude of counts.

3. Conclusion:

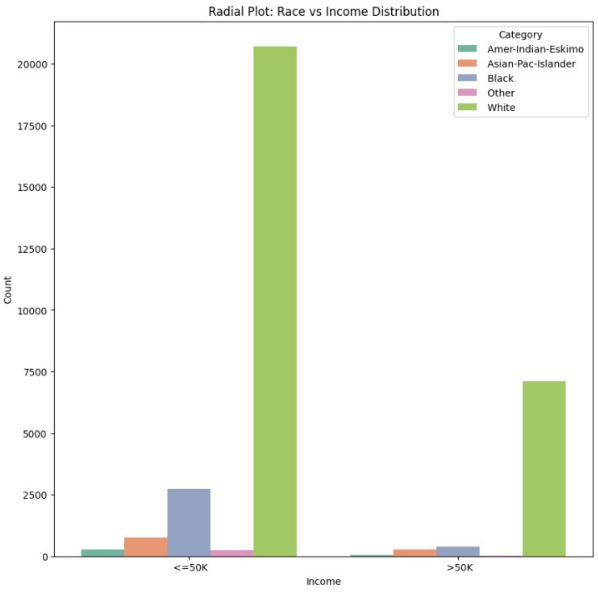
- The heatmap effectively highlights the dominant role of the "Private" workclass, particularly for "Married-civ-spouse" and "Never-married" individuals. This indicates a significant concentration of individuals in these categories.
- The visualization also reveals a clear pattern of income distribution across different workclasses and marital statuses. For example, within the "Private" workclass, there's a significant number of individuals in both

income brackets, but the distribution varies depending on marital status.

Case 6: "Disparities in WorkClass Distribution Across Racial Groups"



This heatmap illustrates the distribution of individuals across different workclasses, races, and income brackets (<=50K and >50K). The "Private" workclass, particularly for White individuals, dominates with the highest counts, especially in the <=50K category. Significant numbers of White individuals are also seen in the >50K bracket within the "Private" sector. Government sectors ("Federal-gov" and "Local-gov") show a higher concentration of White individuals, with a notable portion earning >50K. Black individuals are predominantly in the "Private" and "Local-gov" sectors, largely in the <=50K bracket. Other racial groups show smaller counts across all categories. The heatmap highlights the disparity in income and representation across different workclasses and racial groups, with a clear concentration of White individuals in the higher-income bracket, particularly within the "Private" sector.



This radial plot, more accurately described as a grouped bar chart, illustrates the income distribution (<=50K and >50K) across different racial categories. White individuals dominate both income brackets, particularly in the <=50K range. While still significant, the number of White individuals earning >50K is notably lower. Black individuals show a substantial presence in the <=50K category but are significantly underrepresented in the >50K category. Other racial groups (Amer-Indian-Eskimo, Asian-Pac-Islander, and Other) show relatively low counts across both income brackets. The chart highlights a clear disparity in income distribution, with White individuals having a significantly larger presence in both income brackets, and Black individuals being disproportionately represented in the lower income bracket.

V. Marketing Strategies:

1. Target Yonger Individuals for Higher Enrollment

2. Highlight the Economic Benefits of Higher Education

3. Appeal to Married Individuals with Higher Work Hours- Offer Online Option

4. Cater to Older Adults Who Continue Working

5. Focus on the Private Sector for Targeted Outreach

6. Focus on Racial Equity and Inclusion in Marketing Campaigns

7. Tailor Programs to Work-Life Balance Needs - highlights the need for flexibility in education for individuals working in various sectors and balancing family responsibilities.

8. Use Data to Refine Targeted Advertising - The data reveals significant patterns based on age, education, marital status, and work class that influence income levels.

9. Address Income Inequities in Marketing Strategies - The income disparities across racial groups, especially the underrepresentation of Black individuals in the >$50K category, suggest that marketing strategies need to address socio-economic challenges and barriers.

VI. Conclusion:

By using these insights and recommendations, XYZ Corporation can help UVW College create targeted marketing strategies that not only increase enrollment but also promote inclusivity and diversity. Focusing on key demographics—such as older professionals, individuals with higher education, married individuals, and those in the private sector—while addressing income and racial disparities will help UVW College appeal to a broad range of prospective students and meet their educational goals.