CSE 578: Project Progress Report

Shachi Shah
School of Computing and Augmented Intelligence
Arizona State University Online
Azusa, California, United States of America
spshah22@asu.edu

I. Problem Statement

It is difficult to analyze demographic factors that influence the income levels, therefore the objective of this project is to analyze demographic factors to their income levels using the UCI Adult dataset. The end goal is to assist UVW College in tailoring its marketing strategies to reach individuals most likely to benefit from its degree programs and other various factors such as age and even marital status. This analysis will inform tailored marketing strategies and guide the development of applications that predict income based on input parameters. Specifically, I aim to identify the characteristics of individuals earning above or below $50,000 annually and develop visualizations to support actionable insights.

II. Background Work and Progress Made

A. *Background Work*

From other mini-assignments such as the *Dino Fun World Assignment* and the *Graphing Dino World Assignment* I was able to complete the graphing techniques and data points in order to analyze the date afterward. For example, I use 'import matplotlib.pyplot' that I used in practice from the other course assignments for all the different barcharts. Barcharts were my choice of visualization and therefore was able to properly compare different distributions.

B. *Progress Made*

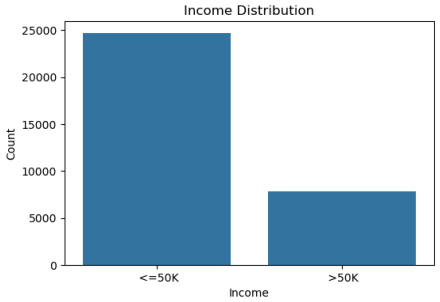

Figure 1: Income Distribution

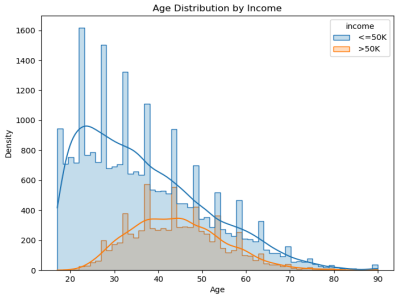

Figure 2: Age Distribution by Income

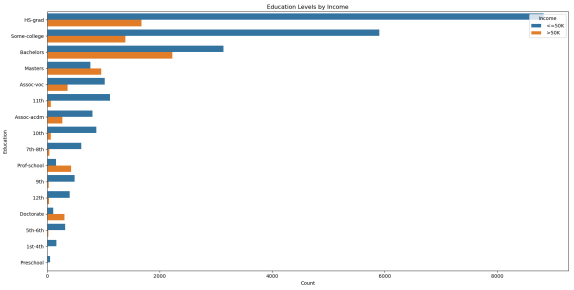

Figure 3: Education Levels by Income

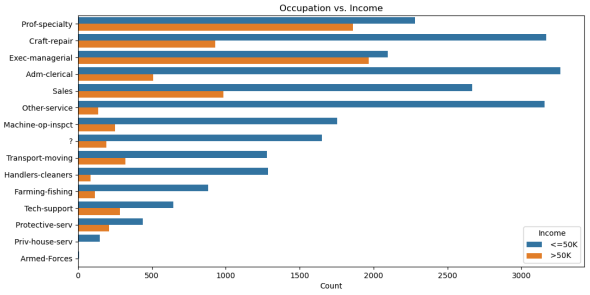

Figure 4: Occupation vs. Income

Significant progress has been made in data preprocessing and exploratory data analysis (EDA). Above are correlations graphed from the script I have developed to demonstrate the different types of barcharts I generated from the python data script I made. The following tasks have been completed from that script:

1. Data Loading: The dataset was successfully loaded into a Pandas DataFrame, and appropriate column names were assigned based on the `adult.names.txt` metadata.
2. Data Cleaning: Missing values, represented by "?", were identified and removed to ensure a

clean dataset for analysis. This step reduced the dataset size from 32,561 rows to 30,162 rows.

3. Exploratory Data Analysis (EDA): Initial univariate and bivariate visualizations were created:
    a. Income Distribution: A count plot of income categories highlighted the imbalance between individuals earning "<=50K" and ">50K".
    b. Education vs. Income: A visualization was generated to analyze the relationship between education levels and income (see Figure 3).
    c. Occupation vs. Income: An additional bivariate analysis examined income distribution across different occupations (see Figure 4).
4. Dataset Export: I designed my python script so that the cleaned dataset was saved as `cleaned_adult_data.csv` for further analysis in the future.

These steps overall provided a foundational understanding of the dataset's structure and key trends and will help me prepare for my Final Report.

III. Challenges and Solutions

A. *Challenges*

1. Approximately 1,800 rows containing missing values in the `workclass`, `occupation`, and `native-country` columns. This especially took the most time to figure out how to properly clean in order to read this as empty data and not pre-filled broken data that would have disrupted the data from being accurate when I generate my graphs.
2. Income categories were imbalanced, with a significant majority earning "<=50K". As per requirements, I needed the majority earnings to be balanced for more accurate data. Again, the data had to be cleaned in order for this to work.

B. *Solutions*

1. Rows with missing values were dropped to maintain data integrity. Alternative methods, such as imputation, were considered but not implemented to avoid potential bias. This way I do not mistakenly do something else to miscalculate the data and have a more accurate analysis through graph generation.
2. The category imbalances will be addressed in future multivariate analyses by normalizing counts or focusing on stratified subsets of the data.

IV. Next Steps and Tasks

A. *Multivariate Analyses*

1. Future visualizations will explore interactions between multiple variables (e.g., age, education, and hours-per-week vs. income).
2. Correlation analysis will be performed to quantify relationships between numerical attributes like age, capital gain, and hours worked.

B. *Advanced Visualizations*

1. Generate multivariate plots using Seaborn to uncover deeper patterns.
2. Enhance current visualizations with annotations to improve interpretability.

C. *Reporting and Documentation*

1. Begin drafting the final report, ensuring all visualizations are appropriately documented and contextualized.
2. Address remaining questions about influential factors and actionable insights.

D. *Plan to Complete Remaining Tasks*

1. Expanding EDA to include multivariate visualizations and correlation analysis.
2. Summarizing findings in a professional and structured final report.
3. Incorporating feedback from this progress report to refine the analyses.

With the current progress and outlined next steps, the project is well-positioned to meet the final deliverable's objectives.

References

[1] Ghayekhloo, Samira. "Course Project Progress Report" *CSE 578 - Data Visualization*, Ira A. Fulton Schools of Engineering, 13 January 2025.