CSE 578: Course Project Final Report

Shachi Shah, *Student, CSE 578*

Abstract

This report analyzes the demographic factors that influence income levels using the UCI Adult dataset. The primary objective is to identify patterns and correlations between socio-economic attributes and income to assist UVW College in optimizing its marketing strategies. Through various data visualizations, this study examines attributes such as education, age, occupation, and working hours to determine their impact on income categories ($\leq$50K or $>$50K). The findings provide actionable insights for designing targeted educational programs and strategic decision-making. The results show that education level and occupation play a significant role in income disparities, whereas age and work hours alone do not strongly predict higher income.

Index Terms

Data visualization, income analysis, demographic factors, education, occupation, socio-economic attributes, UCI Adult dataset, marketing strategy.

I. Goals and Business Objective

ANALYZING demographic factors influencing income levels is a crucial task for various industries, including education and marketing. This project leverages the UCI Adult dataset to explore the correlation between different socio-economic attributes and income levels. The insights obtained will help UVW College refine its marketing strategies by identifying potential students who are most likely to benefit from its educational programs.

The study focuses on determining the key characteristics of individuals earning above or below $50,000 annually and visualizing patterns that can provide actionable insights.

Key goals include:

1) Identify Significantly Factors Affecting Income: (e.g., age, education, work, hours, occupation, etc.)

2) Develop Meaningful Visualizations: to understand income distribution and its relation to demographic attributes.

3) Provide Data-Driven Recommendations: to UVW College for marketing and strategic decision-making.

4) Ensure Data Cleanliness and Integrity: for accurate analysis and future predictive modeling.

II. Assumptions

1) Income Distribution Bias: The dataset is naturally imbalanced, with a larger proportion of individuals earning less than 50K. This imbalance may affect the generalization of insights for individuals earning above 50K.

2) Key Influences: Income is assumed to be primarily influenced by education level, occupation, age, hours worked, and capital gain/loss. Other factors such as geographic location and social networks are not considered due to dataset limitations.

3) Data Integrity: The dataset is assumed to be a fair representation of the general working population. To maintain consistency, missing values were removed rather than imputed, ensuring that only complete and reliable data points are included.

4) Marketing Implication: Implications: Higher education and stable employment are assumed to be desirable characteristics for upskilling programs. It is assumed that UVW College will prioritize these groups for their targeted marketing efforts.

5) No Predictive Modeling: The project focuses purely on data analysis and visualization, without implementing machine learning models. It is assumed that findings from visual analysis will be sufficient to guide decision-making.

6) Work Hours vs. Income Assumption: While higher work hours might contribute to increased income, it is assumed that additional factors like job role, experience, and industry type play a significant role in determining wages.

7) Capital Gain Influence: It is assumed that individuals with higher capital gains are more likely to be financially stable, which correlates with higher income levels. However, this may not fully account for investment-based incomes.

8) Educational Attainment Assumption: Higher education levels are assumed to be a strong determinant of income, but external factors such as skill specialization and job market demand are not included in the analysis.

III. User Stories

1) User Story #1 - Marketing Strategist: As a marketing strategist in UVM College, they would want to analyze whether age influences income levels, so that they can determine the financial stability of potential students. *Attributes Used: Age, Income.* Reasoning is that older individuals may have accumulated experience leading to higher salaries. This insight can help UVW College identify financially stable candidates for higher education programs.

2) User Story #2 - Marketing Directors: Marketing directors are interesting in the relationship between education levels and income to determine which degree holdres are likely to entoll in advanced courses. *Attributes Used: Education, Income.* Reason is that higher education often correlates with better salaries. Understanding this pattern helps in designing courses tailored to professionals looking to upskill.

3) User Story #3 - Research Teams: Research teams would like to analusze Occupation vs. Income to assess which professional groups have higher earning potential. *Attributes Used: Occupation, Income.* Reasoning is that different occupations yield varying salary ranges. This analysis helps UVW College target professionals in high-income brackets who may seek specialized programs.

4) User Story #4 - Data Analysts: Data analysts would want to explore how age, hours worked per week, and capital gain impact income to derive multivariate patterns in income distribution. *Attributes Used: Age, Hours Per Week, Capital Gain, Income.* Reasoning is that income is influenced by multiple factors. Understanding how these interact helps in identifying trends among high-income earners.

5) User Story #5 - College Career Counselors: College career counselors would want to understand how working hours correlate with income levels to design and recommend programs that cater to future working professionals. *Attributes Used: Hours Per Week, Income.* Reasoning is that working often balance job responsibilities with education. This insight helps tailor flexible courses for professionals who aim to increase their earning potential.

IV. Visualizations

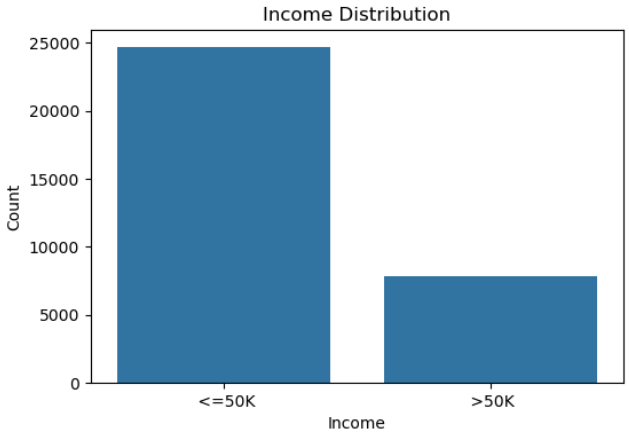

Fig. 1. Income Distribution

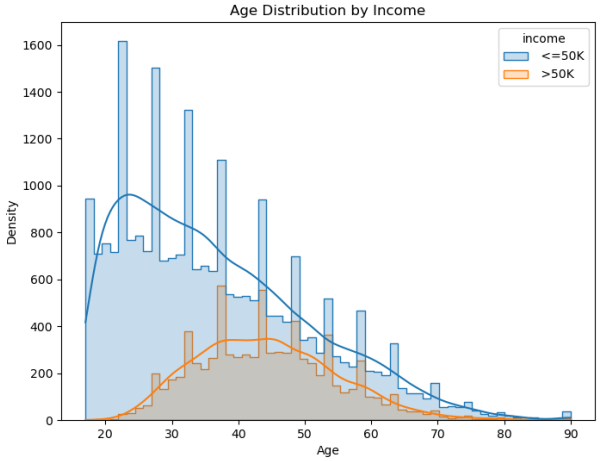

Fig. 2. Age Distribution by Income

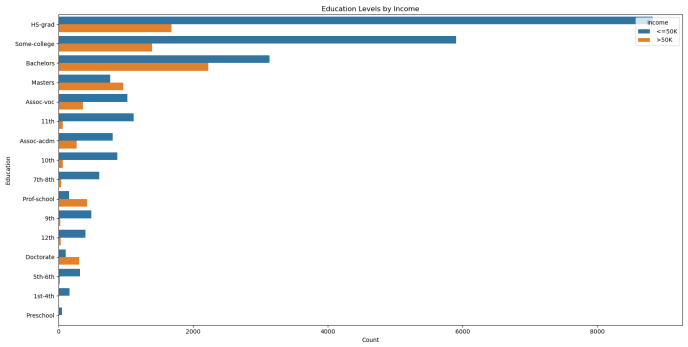

Fig. 3. Education Levels by Income

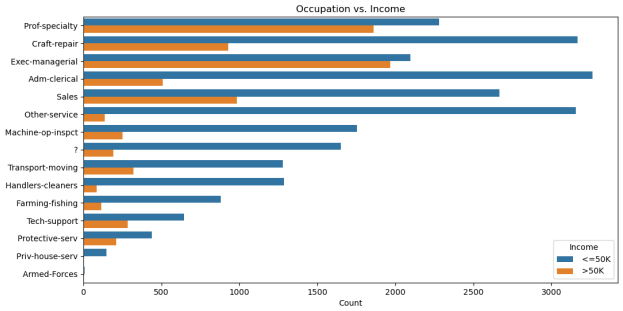

Fig. 4. Occupation vs. Income

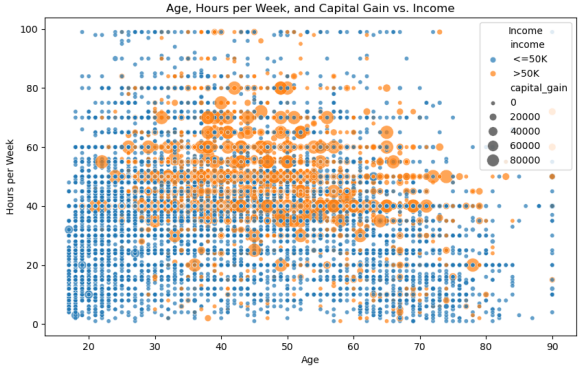

Fig. 5. Age, Hours per Week, and Capital Gain vs. Income

A. Figure 1: Income Distribution (Univariate Analysis)

This figure helps to understand the overall distribution of income levels that is crucial for determining income disparities. The dataset reveals that the majority of individuals earn $\leq$50K, highlighting an income imbalance. This countplot was used to display categorical income distribution. Since most individuals fall in the 50K category, marketing strategies should focus on this group for career-advancement programs.

B. Figure 2: Age Distribution (Univariate Analysis)

This figure helped determine whether older individuals are more likely to earn >50K. While there was a slight increase in income with age, a significant number of older individuals still fall in the $\leq$50K category. A histogram with KDE (Kernel Density Estimate) was used to visualize the age distribution by income. Age alone is not a strong predictor of income; other factors such as education and occupation have a greater impact.

C. Figure 3: Education Levels by Inncome (Bivariate Analysis)

Higher education is often linked to better job opportunities and higher income levels. Individuals with higher education levels (Masters, Doctorate, and Bachelors) have a significantly higher probability of earning >50K. A countplot was used to compare education levels against income categories. Marketing efforts should be directed towards bachelor's degree holders looking to upskill, as they represent a key segment likely to enroll in educational programs.

D. Figure 4: Occupation vs. Income (Bivariate Analysis)

To analyze which occupations have the highest proportion of high earners. Executive, managerial, and professional roles show a higher proportion of individuals earning >50K, whereas service-oriented jobs have lower income levels. A countplot was used to analyze occupation-based income distribution. High-income occupations could be targeted for specialized executive education programs.

E. Figure 5: Age, Hours per Week, and Capital Gain vs. Income (Multivariate Analysis)

To analyze how multiple factors interact to impact income. Higher-income individuals tend to work more hours and have higher capital gains, but age alone is not a strong predictor. A scatterplot with hue and size variations was used to visualize trends. Capital gain has a significant impact on income, but work hours alone do not guarantee higher earnings. This suggests that financial investments may play a role in distinguishing high earners.

V. Questions

During the course of the analysis, several challenges were encountered:

1) Missing Values: Approximately 1,800 rows contained missing values in attributes such as workclass, occupation, and native-country. Handling missing values was crucial for ensuring data quality. Since these missing values were distributed across multiple variables, careful consideration was required to decide whether to impute or remove them. Removing these rows ensured that the analysis remained unbiased but also resulted in the loss of some data.

2) Income Imbalance: The dataset had a significant skew, with about 76% of individuals earning ≤50K. This imbalance made it difficult to draw equally weighted insights across income categories. Visualizations required careful interpretation, and statistical normalization techniques were considered but not implemented in this phase of the project.

3) Correlations Between Attributes: Some attributes, such as education, showed a strong relationship with income, while others, such as hours worked per week, presented a more complex pattern. The challenge was to determine which attributes had the highest predictive power for income level and how these attributes interacted in multivariate relationships. This required exploratory data analysis techniques such as scatter plots and correlation heatmaps.

VI. Solutions

To address the challenges that were faced, the following solutions were implemented:

1) Data Cleaning: Rows with missing values were removed to prevent bias and ensure accurate analysis. While imputation techniques such as mean substitution were considered, it was decided that removing missing values would maintain the dataset's integrity without introducing artificial patterns.

2) Normalized of Data: Rows with missing values were removed to prevent bias and ensure accurate analysis. While imputation techniques such as mean substitution were considered, it was decided that removing missing values would maintain the dataset's integrity without introducing artificial patterns.

3) Multivariate Analysis: To handle complex relationships between attributes, scatter plots and pair plots were used to visually inspect correlations. These analyses helped confirm that education and occupation were the strongest determinants of income. Additionally, multivariate plots were used to explore the interaction of age, hours worked, and capital gains to ensure a thorough understanding of their influence on income.

4) Feature Engineering Considerations: Although not implemented in this phase, potential strategies such as creating new categorical bins for age groups or aggregating occupations into broader categories were explored for future analysis.

VII. Not Doing for Future Work

1) Predtictive Modeling: A supervised machine learning model (e.g., logistic regression, decision trees, or neural networks) could be developed to predict income levels based on demographic attributes. This would allow for a more precise assessment of income determinants and could be used to develop targeted marketing strategies for UVW College.

2) Geographic Segmentation: The dataset includes a variable for native country, which was not extensively analyzed in this report. Future research could explore how income levels vary across different geographic locations and whether certain countries have a stronger correlation with high-income individuals.

3) Expanded Feature Engineering: New features could be derived, such as aggregating education levels into fewer categories (e.g., High School, Bachelor's, Advanced Degree) or creating new interaction terms between key variables like work hours and education. These features could improve the robustness of future models.

4) Incorporating External Socioeconomic Data: Future work could involve merging the dataset with additional economic indicators, such as unemployment rates, inflation, or industry growth statistics. This would provide a more comprehensive understanding of how macroeconomic factors influence income levels.

By implementing these enhancements, future studies can provide even deeper insights into income distribution and help UVW College refine its approach to targeting students based on financial stability and career progression.

VIII. Conclusion

The analysis of the UCI Adult dataset has provided significant insights into the factors influencing income levels. Education and occupation emerged as the strongest determinants of income, whereas factors like age and work hours showed limited predictive power. The findings suggest that higher education and professional roles correlate with greater financial success, reinforcing the importance of academic and career advancement programs.

For UVW College, these insights can drive targeted marketing efforts toward individuals seeking career growth and skill enhancement. Given the strong influence of education on income, programs designed for upskilling and executive education should be prioritized. Additionally, capital gains were found to be a critical differentiator among high-income earners, indicating that financial literacy and investment-oriented education could also be valuable offerings.

Future studies could incorporate predictive modeling techniques and external socio-economic data to enhance the analysis. Expanding the dataset to include geographic and industry-specific variables would further refine the understanding of income distribution trends. By continuously leveraging data-driven insights, UVW College can develop strategies to effectively engage its target audience and support their professional aspirations.

Acknowledgment

The author would like to thank Professor Ghayekhloo and the Teaching Assistants for instructing this course and providing the resources to complete this course's final project.