# UVW Enrollment Demographic Analysis by Income

Ari Argoud

**Abstract**

This report presents the data analysis conducted for UVW College's marketing application. This project is aimed at bolstering enrollment by targeting specific income demographics. This is achieved by utilizing data provided by the United States Census Bureau to identify key predictive variables for demographics with income above and below $50,000. By understanding the characteristics and demographics of individuals with income above and below this threshold, UVW College can tailor its marketing strategies to effectively target potential students. The report outlines the project's objectives, data sources, analysis methods, and proposed application functionalities to predict income and guide marketing efforts.

## I. Introduction

XYZ Corporation specializes in leveraging data to create comprehensive marketing profiles that are sold to various companies for targeted marketing purposes. This project is one such profile. UVW College, a local educational institution, is seeking to enhance enrollment numbers and has identified the $50,000 income threshold as a crucial metric in determining marketing criteria for its degree programs. This project aims to develop marketing profiles of potential students utilizing data provided by the United States Census Bureau. Key variables including gender, education, marital status, occupation, age, and more, are assessed for their relevance in the determination of whether income exceeds the $50,000 threshold.

This report details the objectives, methodologies, and outcomes of the data analysis conducted for UVW College's marketing application. By identifying and understanding the factors influencing income levels, the report aims to guide application development, and provide a firm basis of knowledge to facilitate the creation of targeted campaigns and strategies to attract individuals fitting specific demographics. The insights derived from the analysis and the proposed application functionalities are essential tools in achieving UVW College's enrollment goals.

## II. Goals and Business Objectives

The marketing team at UVW College aims to develop an application that can identify the factors determining an individual's income. To this end, the project utilizes United States Census Bureau data to group and analyze the factors that may contribute significantly to an individual's income, with the ultimate goal of facilitating the development of their proposed model or application. The application will have the capability of predicting an individual's income from different input parameters, enabling tailored marketing efforts to effectively engage with prospective students.

Though the analysis primarily attempts to guide application development, it is also intended to provide an understanding of key marketing demographics to the UVW college marketing team. Regardless of further application development, the analysis should enable the team to design targeted marketing campaigns by tailoring tuition amounts, program concentrations, and even the choice between on-campus and online programs with the intention of effectively engaging with their desired demographics.

## III. Assumptions

The data provided by the United States Census Bureau is assumed to be accurate and reliable for estimating the features of the targeted demographics. The Project assumes that the $50,000 income threshold is a meaningful differentiator for prospective students. The project assumes that the UVW marketing team will have access to similar data for potential students as the data used for the analysis conducted, described below.

The data contains 32,561 unique entries, each with non-null values for all variables. The variables tracked are as follows:

- age: A continuous variable, measured in years.
- workclass: A nominal variable representing the employment type.
- fnlwgt: A continuous internal metric describing intra-state similarity between entrants.
- education: An ordinal variable describing the maximum level of education.
- education-num: A continuous variable describing the maximum level of an individual's education.
- marital-status: A nominal variable representing marital status.
- occupation: A nominal variable indicating job type.
- relationship: A nominal variable defining familial relationship.
- race: A nominal variable indicating the individual's race.
- sex: A binary variable, signifying the individual's gender.

A. Argoud was a student with the School of Computing and Augmented Intelligence, Arizona State University, Phoenix, AZ 85004 USA e-mail (aargoud@asu.edu)

Report received June 1, 2023.

- capital-gain: A continuous variable indicating an increase in capital asset values.
- capital-loss: A continuous variable indicating a decrease in capital asset values.
- hours-per-week: A continuous variable, representing hours worked per week.
- native-country: A nominal variable denoting the individual's country of origin.

Various metrics describing the seven continuous variables were calculated as follows in Figure 1:

| | Age | Fnlwgt | Edu-num | Capital-gain | Capital-loss | Hours/wk |
|---|---|---|---|---|---|---|
| Count | 32561 | 32561 | 32561 | 32561 | 32561 | 32561 |
| Mean | 38.581647 | 1.897784e+05 | 10.080679 | 1077.648844 | 87.303830 | 40.437456 |
| Std | 13.640433 | 1.055500e+05 | 2.572720 | 7385.292085 | 402.960219 | 12.347429 |
| Min | 17.000000 | 1.228500e+04 | 1.000000 | 0.000000 | 0.000000 | 1.000000 |
| 25th percentile | 28.000000 | 1.178270e+05 | 9.000000 | 0.000000 | 0.000000 | 40.000000 |
| 50th percentile | 37.000000 | 1.783560e+05 | 10.000000 | 0.000000 | 0.000000 | 40.000000 |
| 75th percentile | 48.000000 | 2.370510e+05 | 12.000000 | 0.000000 | 0.000000 | 45.000000 |
| Max | 90.000000 | 1.484705e+06 | 16.000000 | 99999.00000 | 4356.00000 | 99.000000 |

Fig. 1: Metrics

## IV. User Stories

In the context of this paper, 'user stories' are hypothetical scenarios that illustrate the objectives and requirements of various stakeholders involved in the UVW College marketing application project. Each user story is presented from the perspective of a user, typically a member of the project team or an end-user of the application. They outline a specific need or question that the user might have with respect to the data analysis and application functionalities.

- User Story #1: A UVW marketing team member would like to find out if education number and hours worked per week are determinant factors for an individual's income label.
- User Story #2: The director of marketing is interested to learn what kind of effects marital status and relationship have on income.
- User Story #3: An advertising coordinator would like to know if sex is a significant factor in influencing income label to better market towards different gender demographics.
- User Story #4: A UVW application design team member would like to know if occupation and sex have more combined predictive value than sex on its own.
- User Story #5: A marketing operations manager is curious about the predictive efficacy of capital gain and loss on income.

August 26, 2015

## V. Visualization

Throughout this section, individuals or distributions of individuals earning less than $50,000 in income will be represented with blue while individuals or distributions of individuals earning greater than $50,000 in income will be represented with orange. Proportional plots are used when possible to provide a machine learning oriented understanding of the relationships between income and the variables of interest. The design methodology for all the plots created is consistent, so it will not be repeated unless there is some notable deviation.

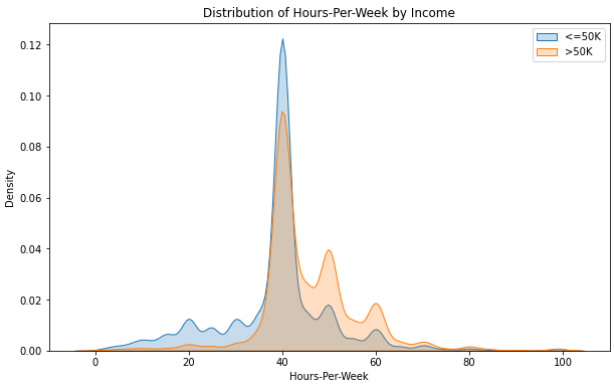


Fig. 2: Hours Worked Per Week KDE

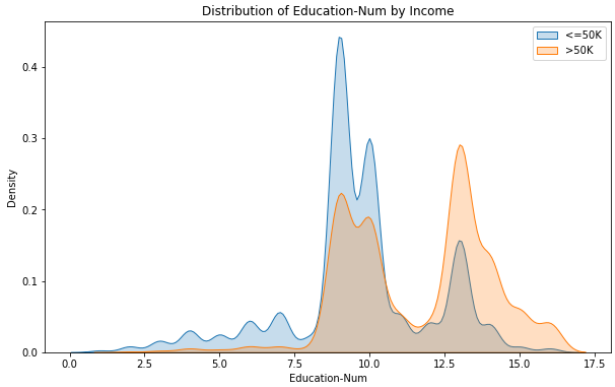


Fig. 3: Education Number KDE

### A. *User Story #1 Visualizations: Hours Worked per Week and Education Number*

Figures 2 and 3 are kernel density estimation (KDE) plots of hours worked per week and education number by income respectively. This visualization type was chosen over histograms due to the relatively uninterrupted nature of the data. The plots are intended to give an idea of the differences in the uni-variate distributions of these variables before multivariate analysis was performed. It is notable that both hours worked per week and education number seem to have a positive correlation with income over $50,000.

Figure 4 is a scatter-plot of education number against hours worked per week. It is apparent that these combined metrics carry some predictive value, as individuals with both a higher education number and higher hours per week are observed to predominantly earn greater than $50,000 in income, while individuals with a lower education number working fewer hours per week are observed to predominantly earn less than $50,000 in income. Though the KDE plots imply this kind of distribution exists, when these two variables are analyzed together, the relationship becomes even more evident.

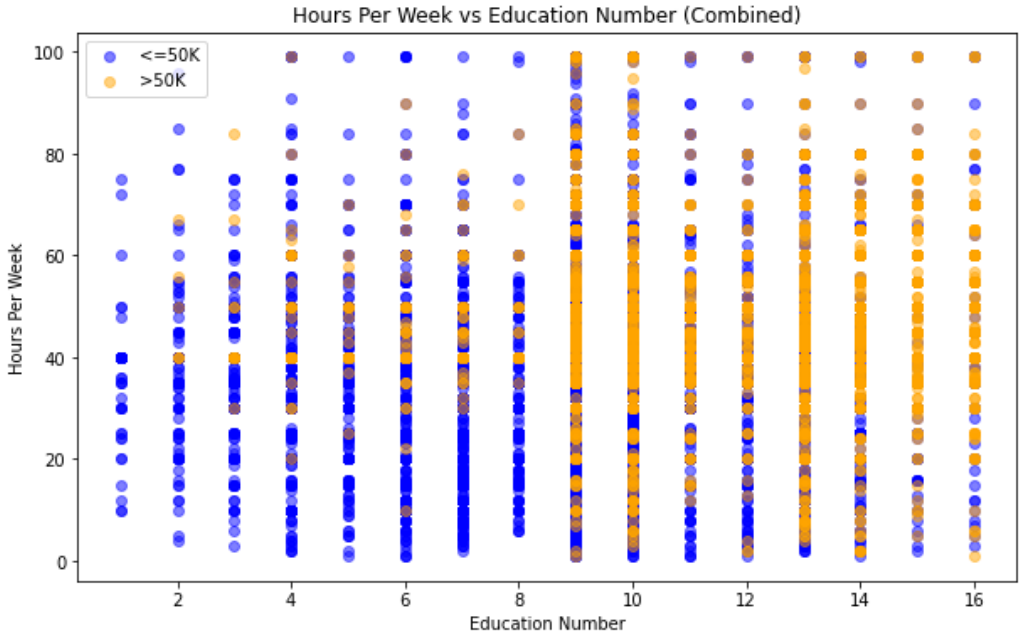


Fig. 4: Scatter Plot of Education Number vs Hours Worked Per Week

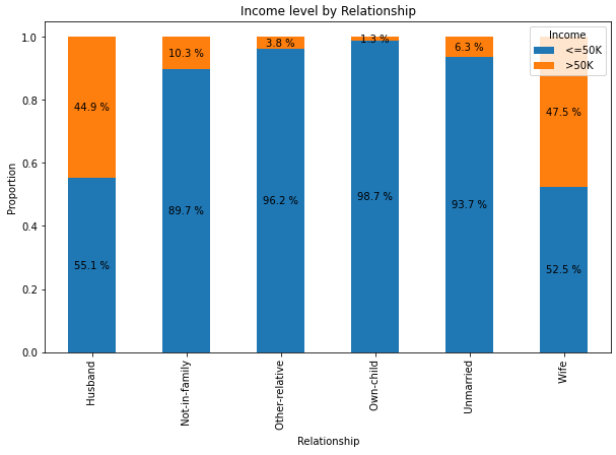


Fig. 5: Proportional Stacked Bar Plot of Income by Relationship

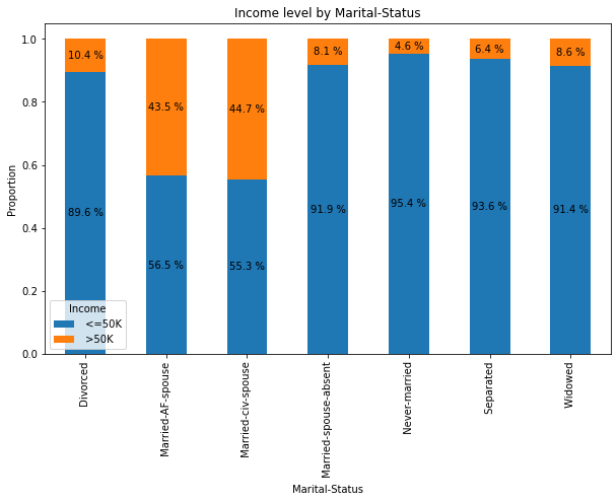


Fig. 6: Proportional Stacked Bar Plot of Income by Marital Status

### B. User Story #2 Visualizations: Relationship and Marital Status

Figures 5 and 6 are stacked bar plots of income by relationship and marital status respectively. These were chosen to provide estimations of the uni-variate predictive value of each of the categories therein. The Own-child and Other-relative categories of the relationship bar plot, as well as the Never-married category of the marital status bar plot have particularly high predictive significance for individuals generating less than $50,000 in income. It is also worth noting that married couples seem to have a significantly higher likelihood of making more than $50,000 in income relative to the other categories examined.

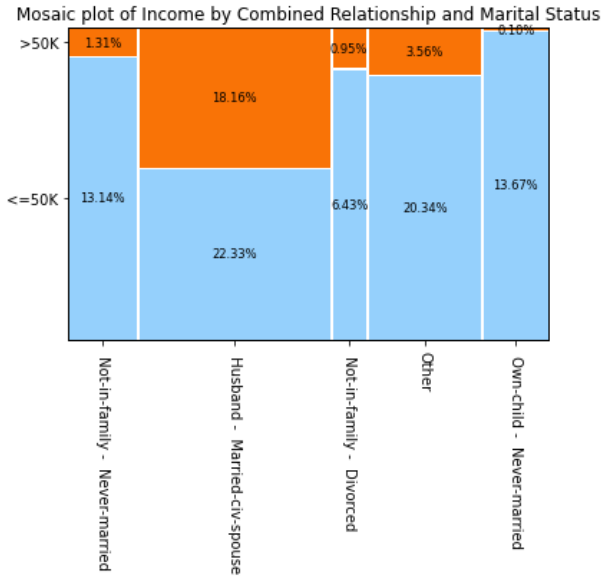


Fig. 7: Mosaic Plot of Income by Relationship and Marital Status

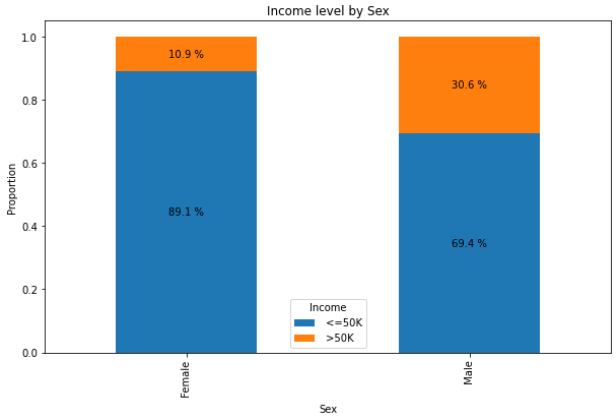


Fig. 8: Proportional Stacked Bar Plot of Income by Sex

Figure 7 is a mosaic plot of income by relationship and marital status. Categories consisting of less than 5% of the total data have been combined into the 'other' category for ease of viewing. The mosaic successfully reinforces the predictive qualities implied by the stacked bar plots and gives further predictive value to individuals with who fall under the Own-child and Never-married categories. Such individuals almost certainly make less than $50,000 in income.

### C. User Story #3 Visualization: Sex

Figure 8 is a stacked bar plot of sex by income. This plot was chosen to provide a general understanding of the income differential between men and women before further analysis was performed. From this plot it is apparent that men are significantly more likely to make more than $50,000 in income.

### D. User Story #4 Visualizations: Sex and Occupation

Figure 9 is a paired, stacked bar plot. For each occupation (with the exception of Armed-Forces due to a lack of female data) there is a bar for male and female, allowing for the intra-occupation comparison of income between sexes. This is intended to highlight occupations where sex may have a higher or lower than average predictive value. Of particular note are the Exec-managerial, Prof-specialty, Sales, Tech-support, and Protective-services categories, which exceed the already existing wage differential in favor of men by more than 10%.

The other notable insight from these plots is the occupations wherein sex is significantly less of a predictive factor than average. These include Priv-house-service, Handlers-Cleaners, and Other-service.

### E. User Story #5 Visualizations: Capital Gain and Loss

Figures 10 and 11 are histograms of capital gain and loss by income respectively. This visualization type was chosen over KDE due to the relatively sparse and non-continuous nature of the distributions. It should be noted that the majority of the data points for these variables were zero values, which were excluded for ease of viewing and analysis.

When present, capital gain has significant predictive capability. Any individual with capital gain greater than $8,000 is almost guaranteed to have income greater than $50,000. Capital loss is less effective towards this end, but may still be able to provide some efficacy if used as a stochastic predictor.

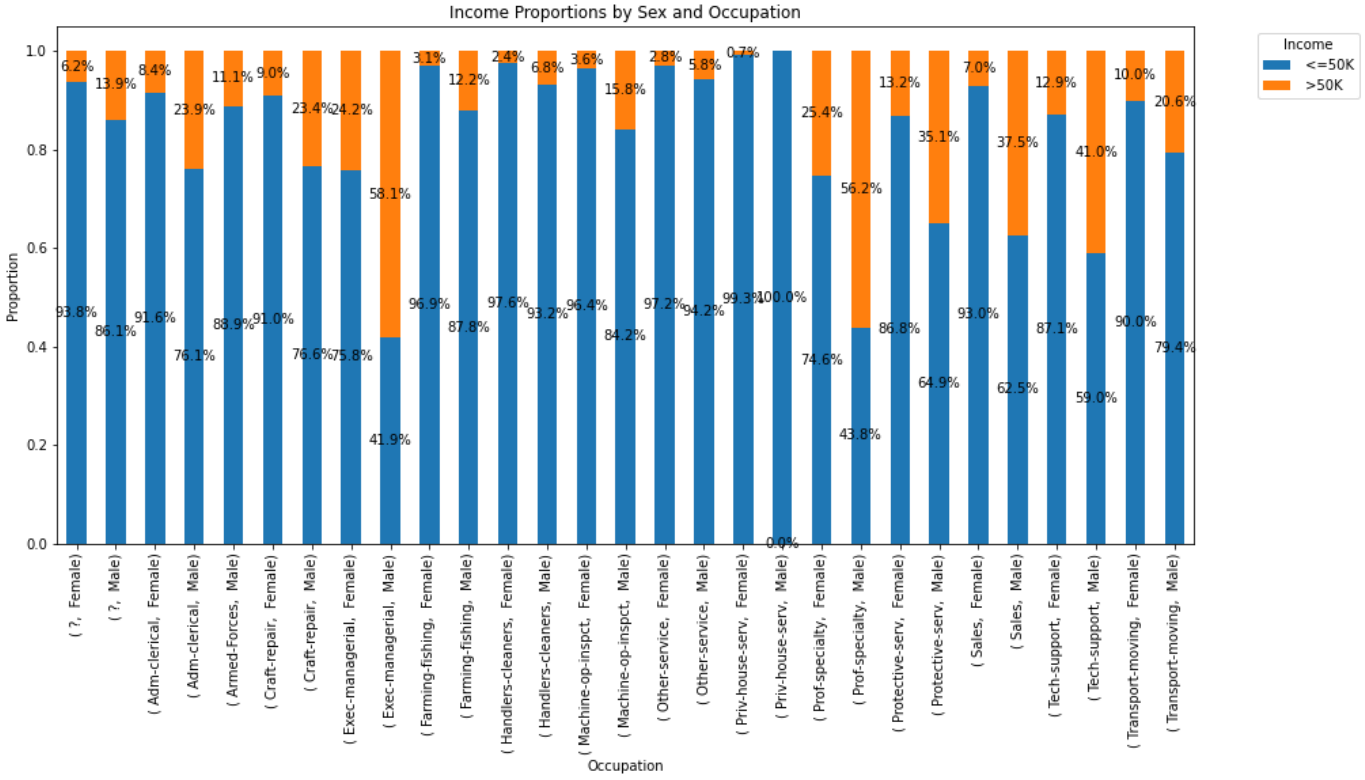


Fig. 9: Proportional Stacked Bar Plot

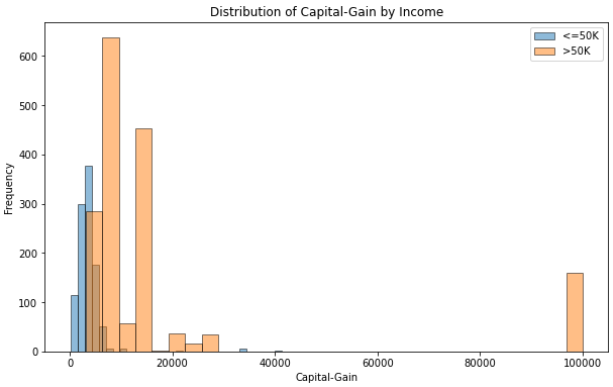


Fig. 10: Capital Gain Histogram

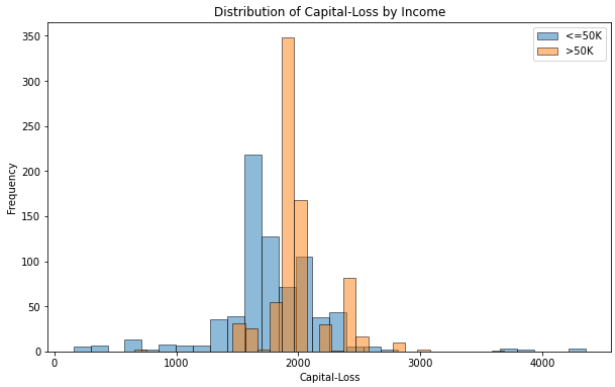


Fig. 11: Capital Loss Histogram

## VI. Questions

Do hours worked per week or level of education achieved impact the likelihood of making over $50,000 in income?

- Answer: Yes. It can be observed from the KDE plots of these variables that there is a roughly positive correlation between both hours worked per week and income, as well as level of education (education num) and income. That said, this correlation was much more concrete when these variables are visualized together via scatter-plot

Is sex a determining factor in whether or not an individual makes over $50,000 in income? Is this impacted by occupation?

- Answer: Yes. Sex was analyzed by itself via stacked bar graph, and in the context of occupation via paired stacked bar graph. It does indeed play a substantial role in determining whether an individual makes over $50,000. This impact can either be embellished or muted by depending on occupation.

Are married couples more likely to make over $50,000 in income?

- Answer: Yes. As visualized by both bar plots of relationship and marital status, as well as a mosaic plot of both these variables, individuals who are currently married have a significantly higher likelihood of making over $50,000 with respects to the other categories examined.

## VII. Not doing

- Perform multivariate analysis of age against other variables with relation to income.
- Further explore relationships between sex and variables other than those discussed in the paper with respects to income.
- Engineer combination features and perform further analysis in the hopes of achieving favorable distributions.
- Discard features that do not provide significant predictive value.
- Determine various machine learning models that might appropriately model the data.
- Test these models against each other with various sets of input parameters to determine which is best suited to the use-case.