



# U.S. Census Bureau

PREDICTING INCOME LEVEL USING CENSUS DATA

MARVIN EDMOND

08-13-2018

SPRINGBOARD CAPSTONE



# Contents

1. U.S. CENSUS BUREAU OVERVIEW
2. PROJECT PURPOSE
3. DATA SOURCES
4. DATA TRANSFORMATION AND ANALYSIS
5. MACHINE LEARNING APPROACH
6. RESULTS
7. SUMMARY

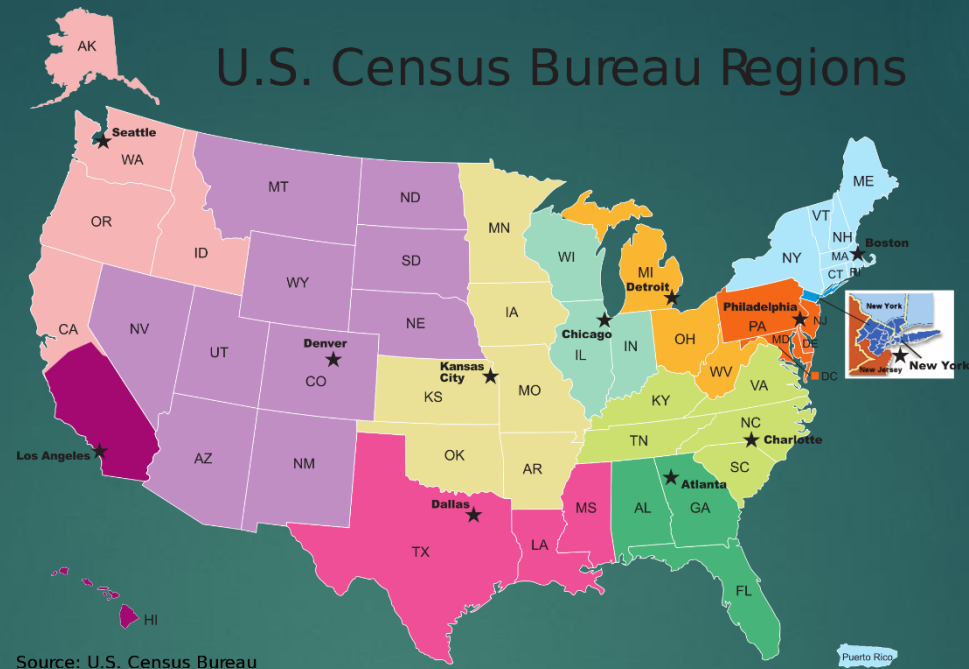


# U.S. Census Bureau Overview




- ▶ The U.S. Census Bureau has been headquartered in Suitland, Md. since 1942, and currently employs about 4,285 staff members.
- ▶ The Census Bureau is part of the U.S. Department of Commerce and is overseen by the Economics and Statistics Administration (ESA) within the Department of Commerce.
- ▶ The Economics and Statistics Administration provides high-quality economic analysis and fosters the missions of the U.S. Census Bureau and the Bureau of Economic Analysis.

# Project Purpose



- ▶ Wanting to practice my newly acquired machine learning skills, I searched for a project which would be interesting.
- ▶ Census data is always available with an abundance of information but, determining which specific variables to use to predict annual income was a fun challenge.
- ▶ Even though the project was simple, it gave me further insight into the power of machine learning.

# Data Sources



The Census Bureau did a wonderful job in bringing some order to the available datasets. A plentiful amount of variables were provided for thorough analysis. The data used in my project is from the 1994 Census survey.

### **Data Files:**

- ▶ Data Folder
- ▶ Data Set Description


### **The Data Folder includes:**

- ▶ Train and test data sets.
- ▶ Variable names for each column.





# Data Transformation and Analysis



The data provided by the Census Bureau is semi-unstructured but the data dictionary helped out tremendously in helping to clean the data. A few issues occurred while wrangling with the data which were:

- ▶ Reassigning easy to read variable names to the data.
- ▶ While checking for missing values, I noticed missing values contained a '?' instead of an NA value.
- ▶ All missing data were later converted to NA values.

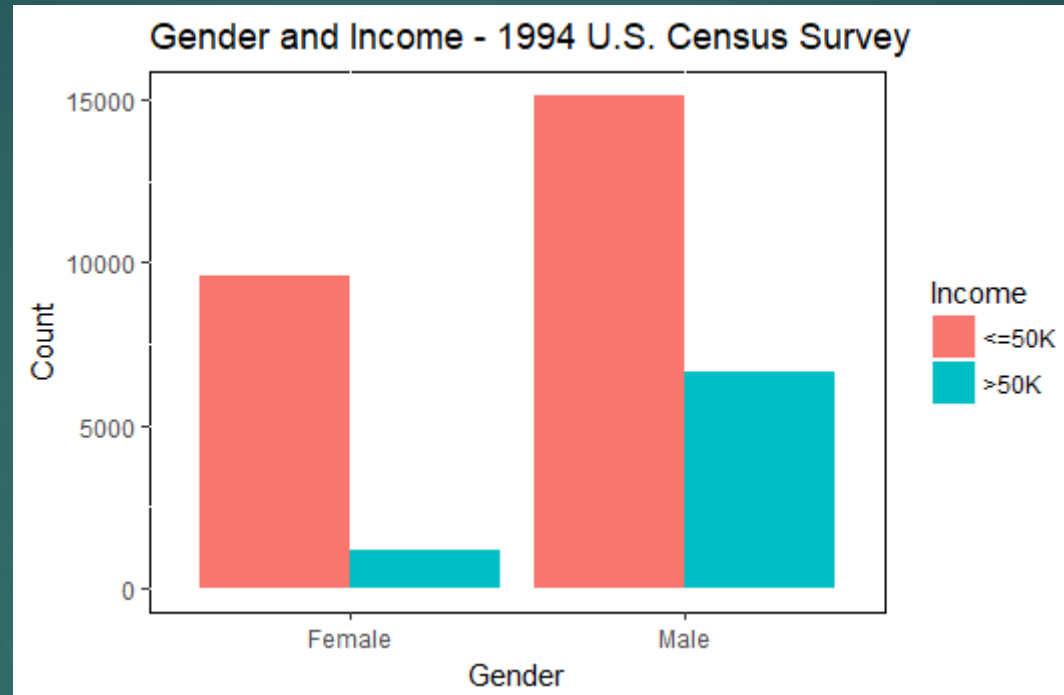
## Data Transformation and Analysis Continued..

- ▶ Important information that the data contains are age, gender, work class, occupation and education level. These factors help to create a profile which can be further analyzed to increase predictability for a predetermined income level. The use of character defining traits produces more efficient training of data sets further strengthening algorithms when it comes to testing.
- ▶ However, the data set does provide some limitations. The absence of specified states/cities in the survey makes it impossible to determine which regions have the highest income level. This piece of information could have further aided the algorithms in determining if an individual makes over a certain amount of income per year. Also, the knowledge of state tax levels would help us to determine which areas of the U.S. did individuals retain more of their earnings.

Data exploration is vital for understanding your data before performing further analysis. Familiarizing yourself with the data visually, quickly helps to determine correlation between variables. Investigating correlation amongst several variables could provide valuable insights pertaining to my capstone involving U.S. Census data. Variables for investigating correlation:

- ▶ Income vs Education
- ▶ Age vs Income
- ▶ Income vs Gender

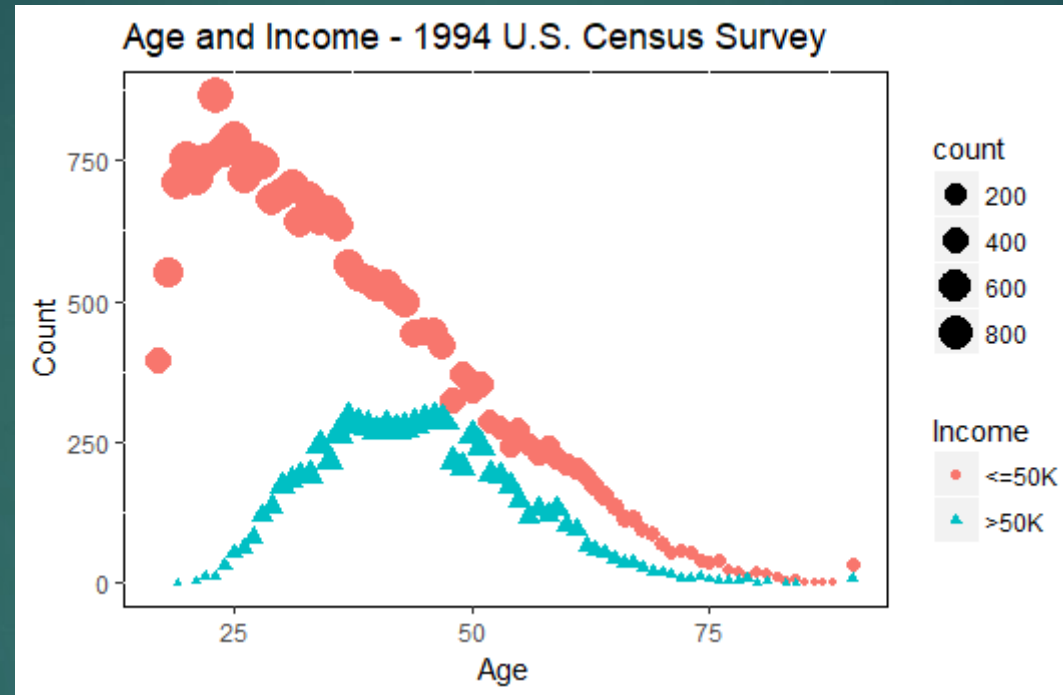
## Data Transformation and Analysis Continued..



The effects of gender on annual income for female laborers are evident.

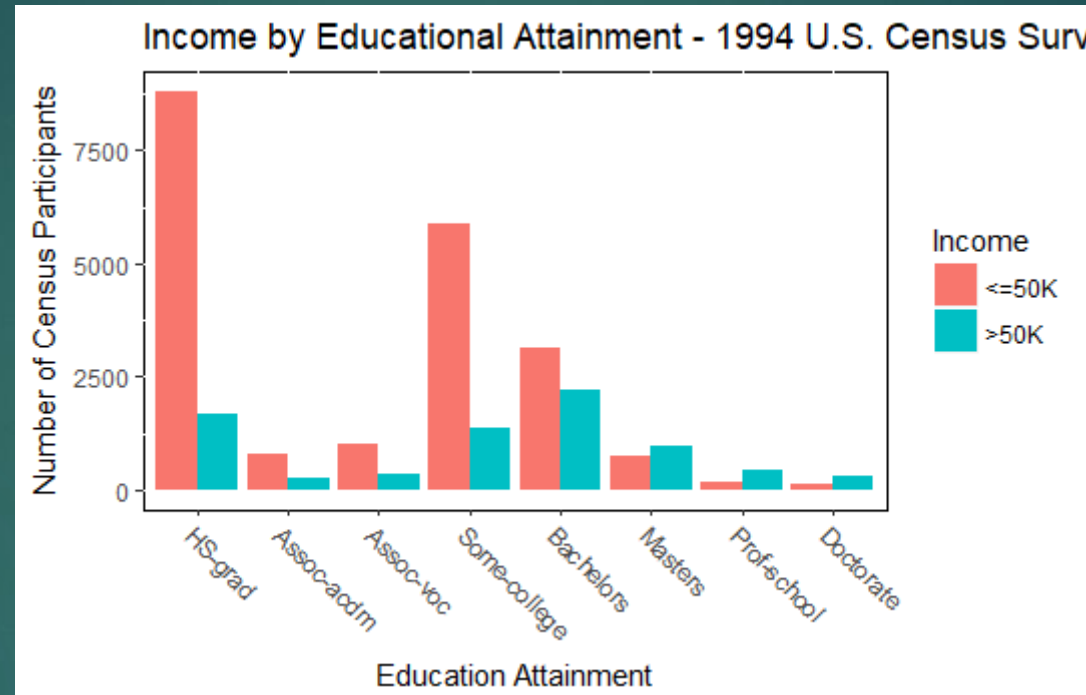
- ▶ The percentage of females that are compensated over \$50K/year compared to their overall aggregated income is a very small amount.
- ▶ Furthermore, the percentage of males that make over \$50k/year compared to their aggregated overall income is far greater than their female counterparts. This represented a huge pay gap, reminiscent of that time period.

## Data Transformation and Analysis Continued..



- ▶ Individuals between 18-35 years old have a wider disparity of income.
- ▶ The majority of this age group's annual income is less than or equal to \$50k/yr. As individuals grow older the gap in annual income begins to shrink.
- ▶ 18-25 year olds are usually in school either full-time or work part-time jobs. Students and recent graduates are navigating various career paths so generous employment offers are few and far between.

## Data Transformation and Analysis Continued..



- ▶ Individuals attaining only a high school diploma are more than likely to make less than or equal to \$50k/year in 1994, as well as associate and bachelor degree holders.
- ▶ As advanced levels of education are sought, the probability of making over \$50k/year rises in proportion.
- ▶ Progressing from a masters level to professional school, then finally a doctorate, the probability of making less than or equal to \$50k/year decreases and the probability of making over \$50k/year increases.

# Machine Learning Approach