Milestone Report

Marvin Edmond

September 1, 2017

```
# Set working directory
setwd("~/School/Springboard/Capstone")
# Load in project data and dplyr and tidyr
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.3.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.3.3
```

Predicting Income from U.S. Census Bureau Survey

The aim of my capstone was to predict whether income exceeds $50k using machine learning algorithms. Data from the U.S. Census Bureau survey was collected and analyzed. Train and test sets were used in the process.

About U.S. Census Data

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, I gives 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.

Cleaning The Data

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 occured 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.

```
adult <- read_csv("~/School/Springboard/Capstone/Data/adult.data",
col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   X2 = col_character(),
##   X3 = col_integer(),
##   X4 = col_character(),
##   X5 = col_integer(),
##   X6 = col_character(),
##   X7 = col_character(),
##   X8 = col_character(),
##   X9 = col_character(),
##   X10 = col_character(),
##   X11 = col_integer(),
```

```
##   X12 = col_integer(),
##   X13 = col_integer(),
##   X14 = col_character(),
##   X15 = col_character()
## )
```

```r
View(adult)


# Add column names to data
?colnames
```

```
## starting httpd help server ...
```

```
##  done
```

```r
# Also convert imported dataset to table dataframe
census_data <- tbl_df(adult)


colnames(census_data) <- c("Age", "Work_Class", "FNLWGT", "Education","Education_Number", "Marital_Statu

# Checking for missing values in my dataset
summary(census_data)
```

```
##       Age         Work_Class            FNLWGT         Education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
##  Education_Number Marital_Status      Occupation         Relationship
##  Min.   : 1.00    Length:32561       Length:32561       Length:32561
##  1st Qu.: 9.00    Class :character   Class :character   Class :character
##  Median :10.00    Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      Race               Sex             Capital_Gain    Capital_Loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1078   Mean   :  87.3
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  Hours_Per_Week  Native_Country          NA
##  Min.   : 1.00   Length:32561       Length:32561
##  1st Qu.:40.00   Class :character   Class :character
##  Median :40.00   Mode  :character   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

```r
sum(is.na(census_data$age))
```

```
## Warning: Unknown or uninitialised column: 'age'.
```

```
## Warning in is.na(census_data$age): is.na() applied to non-(list or vector)
## of type 'NULL'

## [1] 0
```

```r
sum(is.na(census_data$Age))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Work_Class))
```

```
## [1] 0
```

```r
# I notice that all missing values contain a question mark (?). I will have to convert these values int
census_data[ census_data == "?"] <- NA


# Now that all missing values within the data frame have been converted to an NA value, I can now perfo
sum(is.na(census_data$Age))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Work_Class))
```

```
## [1] 1836
```

```r
sum(is.na(census_data$FNLWGT))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Education))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Education_Number))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Marital_Status))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Occupation))
```

```
## [1] 1843
```

```r
sum(is.na(census_data$Relationship))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Race))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Sex))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Capital_Gain))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Capital_Loss))
```

```
## [1] 0
```

```r
sum(is.na(census_data$Native_Country))
```

```
## [1] 583
```

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:

- Hours per week vs Education (separated by sex)
- Age vs Education
- Education vs Gender

Initially, NA values were scattered throughout several columns of the data set. Several inline commands were used to determine most repeated values and fill in those missing values.

```r
# Using a table to provide a list of all possible values in a chosen category and the number of times i

sort(table(census_data$Work_Class, useNA="ifany"))
```

```
##
##     Never-worked      Without-pay      Federal-gov     Self-emp-inc
##                7               14              960             1116
##        State-gov             <NA>        Local-gov Self-emp-not-inc
##             1298             1836             2093             2541
##          Private
##            22696
```

```r
sort(table(census_data$Occupation, useNA="ifany"))
```

```
##
##      Armed-Forces   Priv-house-serv   Protective-serv      Tech-support
##                 9               149               649               928
##   Farming-fishing Handlers-cleaners  Transport-moving              <NA>
##               994              1370              1597              1843
## Machine-op-inspct     Other-service             Sales      Adm-clerical
##              2002              3295              3650              3770
##   Exec-managerial      Craft-repair    Prof-specialty
##              4066              4099              4140
```

```r
sort(table(census_data$Native_Country, useNA="ifany"))
```

```
##
##         Holand-Netherlands                   Scotland
##                          1                         12
##                   Honduras                    Hungary
##                         13                         13
## Outlying-US(Guam-USVI-etc)                 Yugoslavia
##                         14                         16
##                       Laos                   Thailand
##                         18                         18
##                   Cambodia            Trinadad&Tobago
##                         19                         19
##                       Hong                    Ireland
##                         20                         24
##                    Ecuador                     France
##                         28                         29
##                     Greece                       Peru
##                         29                         31
##                  Nicaragua                   Portugal
##                         34                         37
##                       Iran                      Haiti
##                         43                         44
##                     Taiwan                   Columbia
##                         51                         59
##                     Poland                      Japan
##                         60                         62
##                  Guatemala                    Vietnam
##                         64                         67
##         Dominican-Republic                      Italy
##                         70                         73
##                      China                      South
##                         75                         80
##                    Jamaica                    England
##                         81                         90
##                       Cuba                      India
##                         95                        100
##                El-Salvador                Puerto-Rico
##                        106                        114
##                     Canada                    Germany
##                        121                        137
##                Philippines                       <NA>
##                        198                        583
##                     Mexico              United-States
##                        643                      29170
```

```r
# NA values will now be filled with its corresponding most repeated value within its column.

census_data$Work_Class[is.na(census_data$Work_Class)] <- "Private"
census_data$Occupation[is.na(census_data$Occupation)] <- "Prof-specialty"
census_data$Native_Country[is.na(census_data$Native_Country)] <- "United-States"

# Checking the sum of NA values within the entire data set will reveal any remaining missing values
```

```r
sum(is.na(census_data))
```

```
## [1] 0
```

```r
# Add column name to predictor values

colnames(census_data)[15] <- "Income"
```

The replacement of NA values permitted exploratory data analysis to begin.

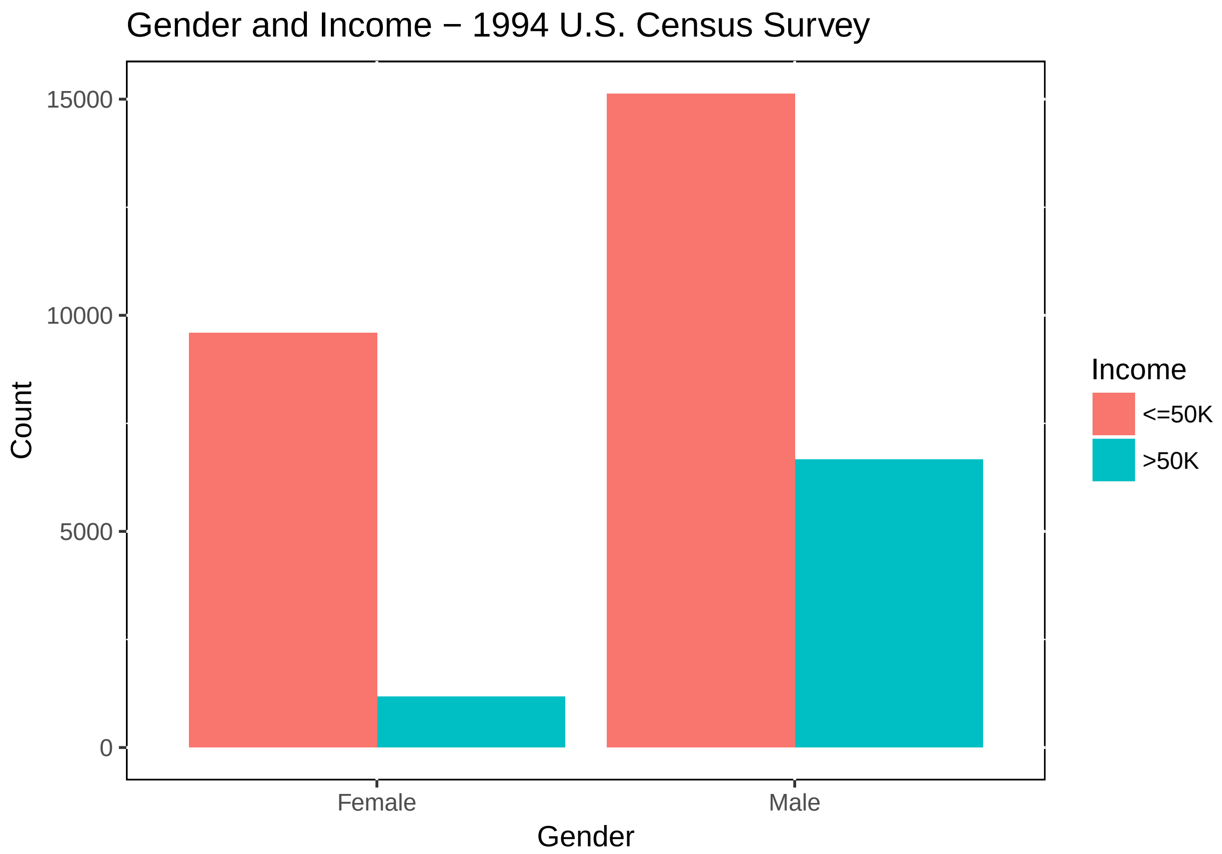
```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

B.

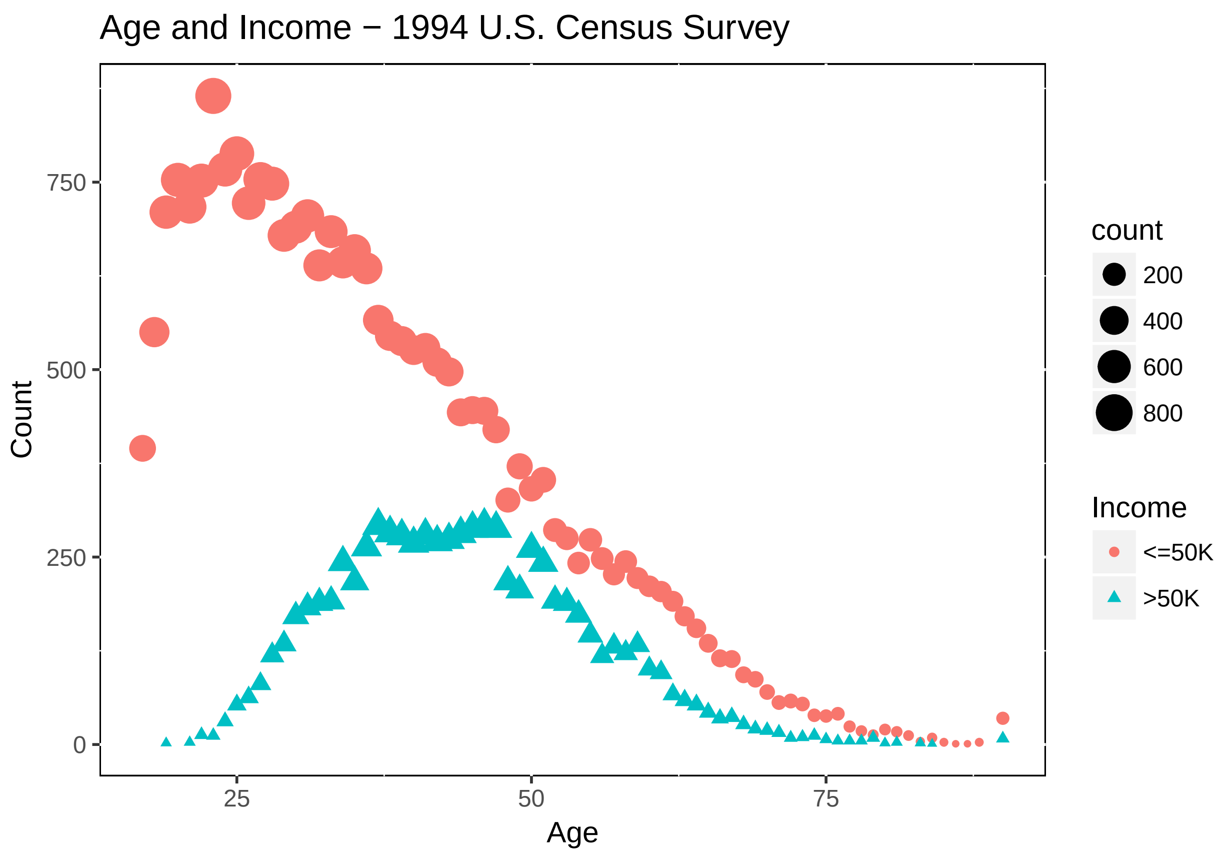
One of the preliminary investigations for correlations to be performed was the effect of gender on annual income. A bar graph separating the data by gender and color-coded by income was used. Applying a color designation of income provided a more complete picture of any differences.

```r
gender_income_plot <- ggplot(data=census_data, aes(x=Sex, fill=Income)) + geom_bar(position="dodge", al

gender_income_plot + theme(panel.background = element_rect(fill='white', colour='black'))
```



Correlating Age and Income was meant to provide further insight into any longer term trends. What the data reveals will be compared with other visualizations to uncover implicit correlations and for a more holistic perspective.

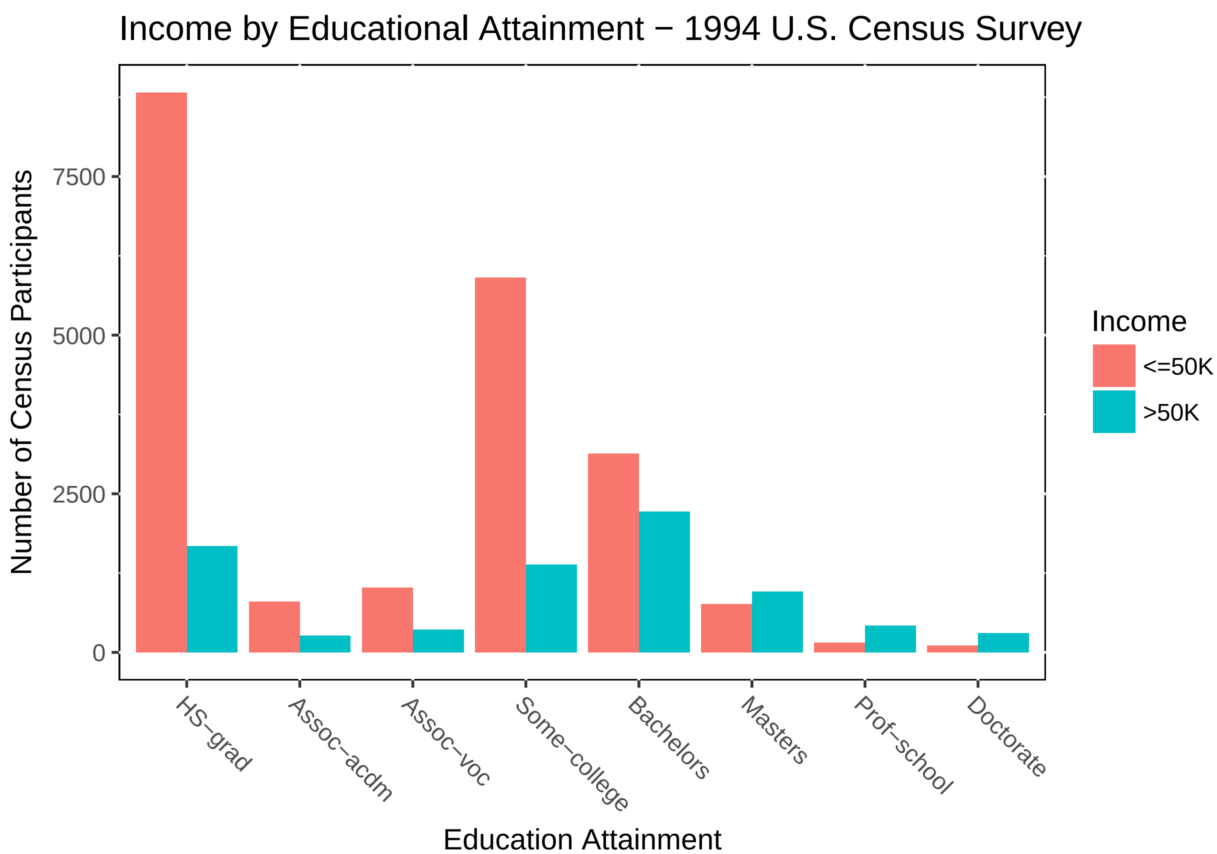
```
age_income_plot <- ggplot(data=census_data, aes(x=Age, y=..count..)) + geom_point (aes(colour=Income, sh

age_income_plot + theme(panel.background = element_rect(fill='white', colour='black'))
```



Last but not least, a plot comparing income and educational attainment is essential. To be able to find any income disparities related to education will help in determining a correlation betwen level of attainment and overall earnings.

```
income_education_plot <- ggplot(data=census_data, aes(x=Education, fill=Income)) + geom_bar(position="d

income_education_plot + theme(panel.background = element_rect(fill='white', colour='black'), axis.text.
```

```
## Warning: Removed 4253 rows containing non-finite values (stat_count).
```



III. Results

A. Gender and Income

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.

Overall, males brought in more income as a whole compared to females. This could be due to a preference for employing males in the job market. The underlying issue of gender discrimination producing labor and pay gaps between males and females is exposed. Gender would be a key attribute in a regression model due to it's influence on distribution of income.

B. Age and Income

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. However, as time progresses entry-level

employees are promoted and enter mid-level or senior-level positions. Also, attaining higher levels of education put individuals in a better position to receive better job offers.

Furthermore, dividends from investments such as stocks, bonds, IRAs, and pensions can explain the continued shrinking of the income gap for older individuals. Those who participate early in retirement plans reap many benefits at an older age. The overall decrease in census participants as age increases could be due to one's own mortality or health issues. Age is an ideal attribute for a regression model because it proves an excellent predictor of income level. Age and income follows a trend.

C. Income and Education

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/yr 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/yr decreases and the probability of making over $50k/year increases.

***Education levels below high school were removed because the small amount of data pertaining to grade school levels were insignificant. Education is a great attribute for a regression model because the various levels of attainment have highly correlated income levels, i.e. Doctorate holders are compensated more as a whole than solely high school graduates.**

D. Future Analysis

Future analysis could be done to investigate the effects of race on annual income. Gender discrimination is an important issue which affects employment opportunities but the pairing of race should be closely studied to reveal any insightful results. Also, including the participant's native country would be an interesting factor to examine and how it affects the amount of income earned.