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()
## )
```

```
View(adult)


# Add column names to data
?colnames
```

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

```
##  done
```

```
# 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
```

```
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
```

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

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

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

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

```
# 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
```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
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"))
```

```
##
##         Holland-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
```

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

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

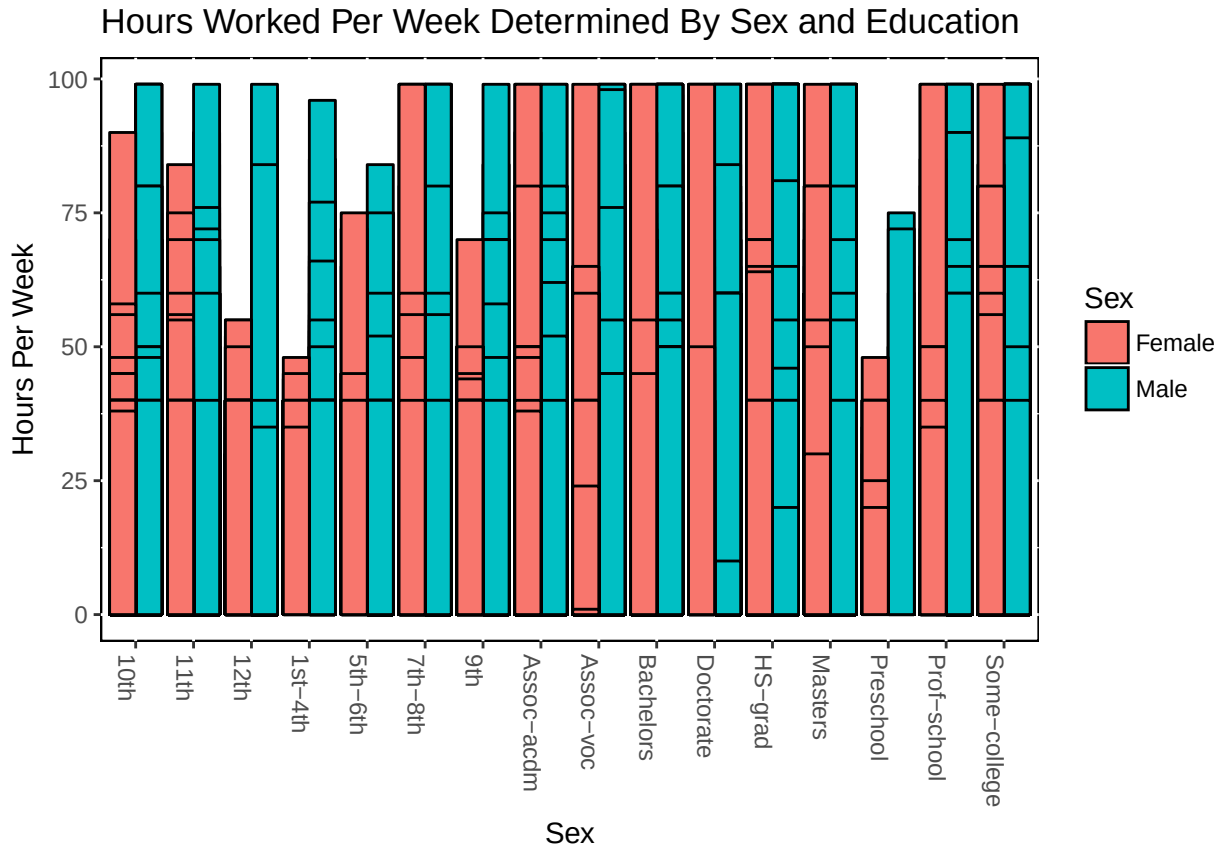
The replacement of NA values permitted exploratory data analysis to begin.

```
library(ggplot2)
```

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

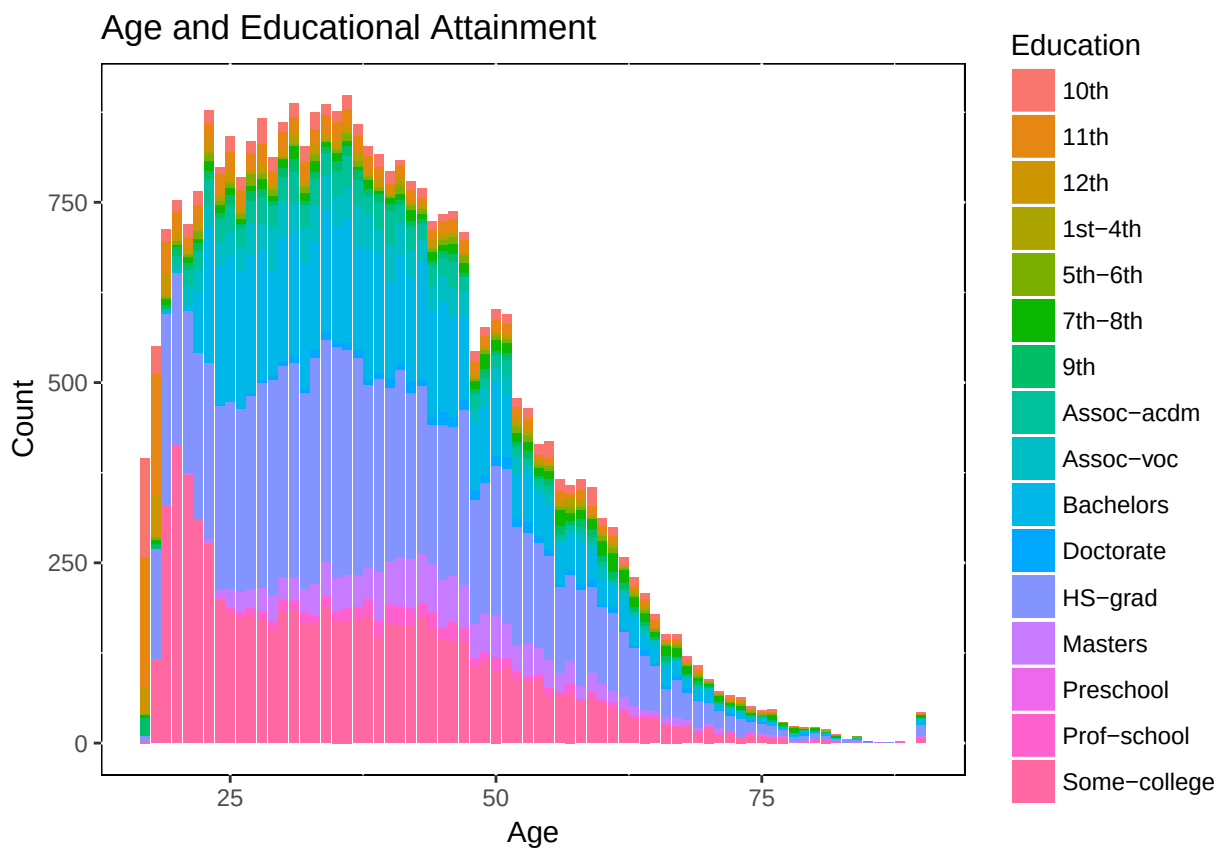
One of the preliminary investigations for correlations to be performed was the effect of sex and education on the number of hours worked. A bar graph separating the data by sex and color-coded by education was used. The color designation of education provided a better insight into the direct effects of educational attainment.

```
sex_education_hours_plot <- ggplot(data=census_data, aes(x=Education, y=Hours_Per_Week, fill=Sex)) + ge

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



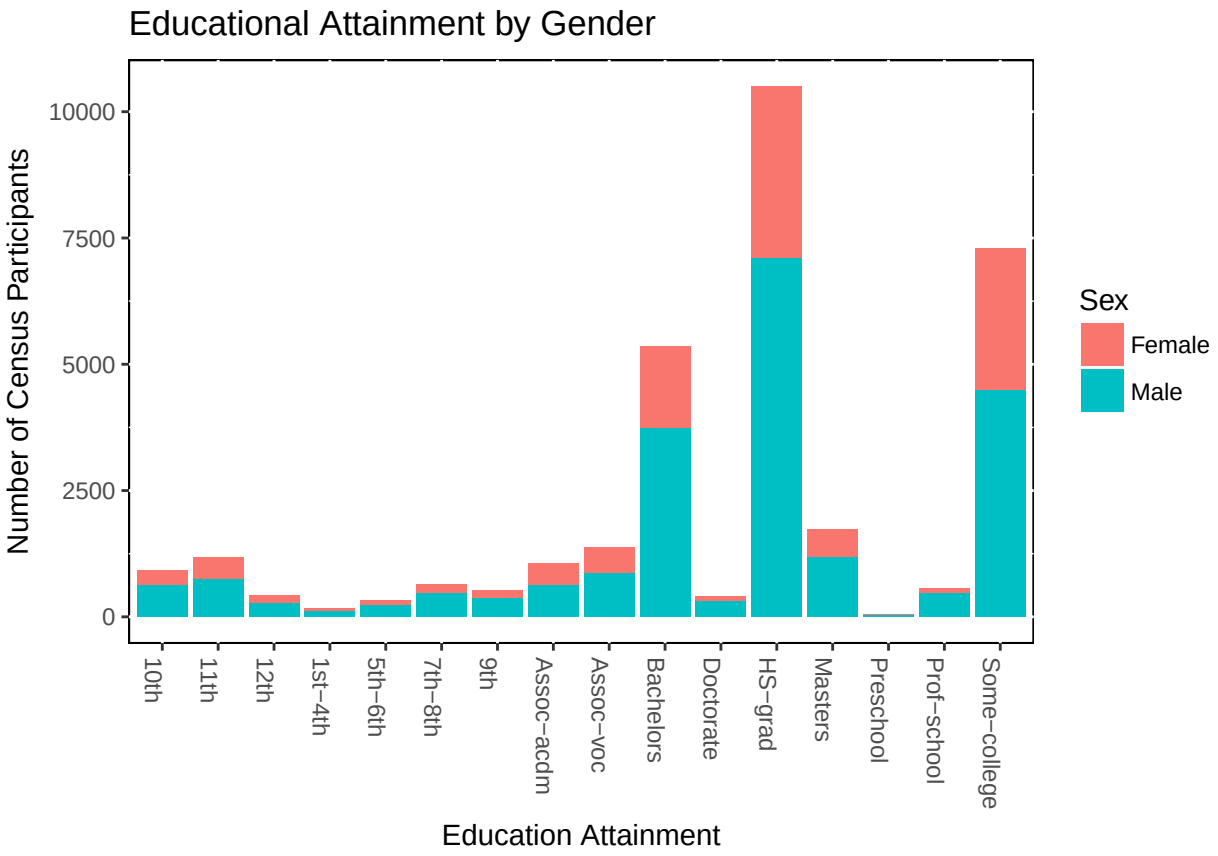
The effects of age on educational attainment was meant to provide further insight into its longer term effects. What the data reveals will be compared with other visualizations to uncover implicit correlations and for a more holistic perspective.

```
age_education_plot <- ggplot(data=census_data, aes(x=Age, fill=Education)) + geom_bar(alpha=1, width=0.

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



Last but not least, a plot comparing gender and educational attainment is essential. To be able to find any educational gaps in gender will help in determining more precise factors for possible career outcome, and overall earnings.

```
gender_education_plot <- ggplot(data=census_data, aes(x=Education, fill=Sex)) + geom_bar() + xlab("Educa

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



A. Hours Per Week and Education

The effect of educational attainment on number of hours worked for female laborers are evident. Females whose highest level of education was high school, worked less hours per week compared to their male counterparts. On the other hand, males worked a similar amount of hours per week regardless of education level. This can possibly represent the bigger problem of gender discrimination, with the lack of education exacerbating the issue.

Regardless of a male participants' native country or education level, they continued to dominate the majority of working hours in the United States.

B. Age and Education

The decreasing amount of continued education among adults over 35 years of age is not surprising. According to the Department of Education, approximately 70% percent of Americans only hold a high school diploma. The largest proportions of education level are "HS Grad" and "Some College". These two account for most of the data represented in the graph. The data concurs with the research analysis of the DOE.

The importance of higher education was not a stressed a few decades ago. A high school diploma was good enough to find a decent job in many workplaces. In today's world, a college education is now the least amount of schooling needed for entry-level jobs. This puts

tremendous pressure on adolescents to undertake advanced learning and outperform their peers. The increased amount of college students among young adults directly reflects this.

C. Gender and Education

Due to the Census data being from a 1994 survey, 23 years ago, the gender gap in education can be shocking at first. In modern times, females now account for close to 50% of higher education recipients from recent survey responses. The lack of advanced education among females directly reflected their opportunities for employment in the workplace. High ranking positions began to fill with women as their educational attainment began to rise.

Men accounted for the majority of higher education and also held higher level positions.