Final Report

Carleena Ortega and Saelin Bjornson

15/03/2020

Contents

Adult Income

Introduction

The Dataset

Who: The data set was extracted by Barry Becker from the 1994 Census database and is donated by Silicon Graphics
What: This is a multivariate dataset with categorical and integer variables. It contains the predicted income of individuals from the census with attributes including age, marital status, work class, education, sex, and race.

When: The data is from a 1994 census.
Why: The data set is found in the University of California Irvine Machine Learning Repository, and was used for ML prediction of whether a person makes over or under 50K a year based on their attributes.
How: The census data was collected by survey.

The Research Questions

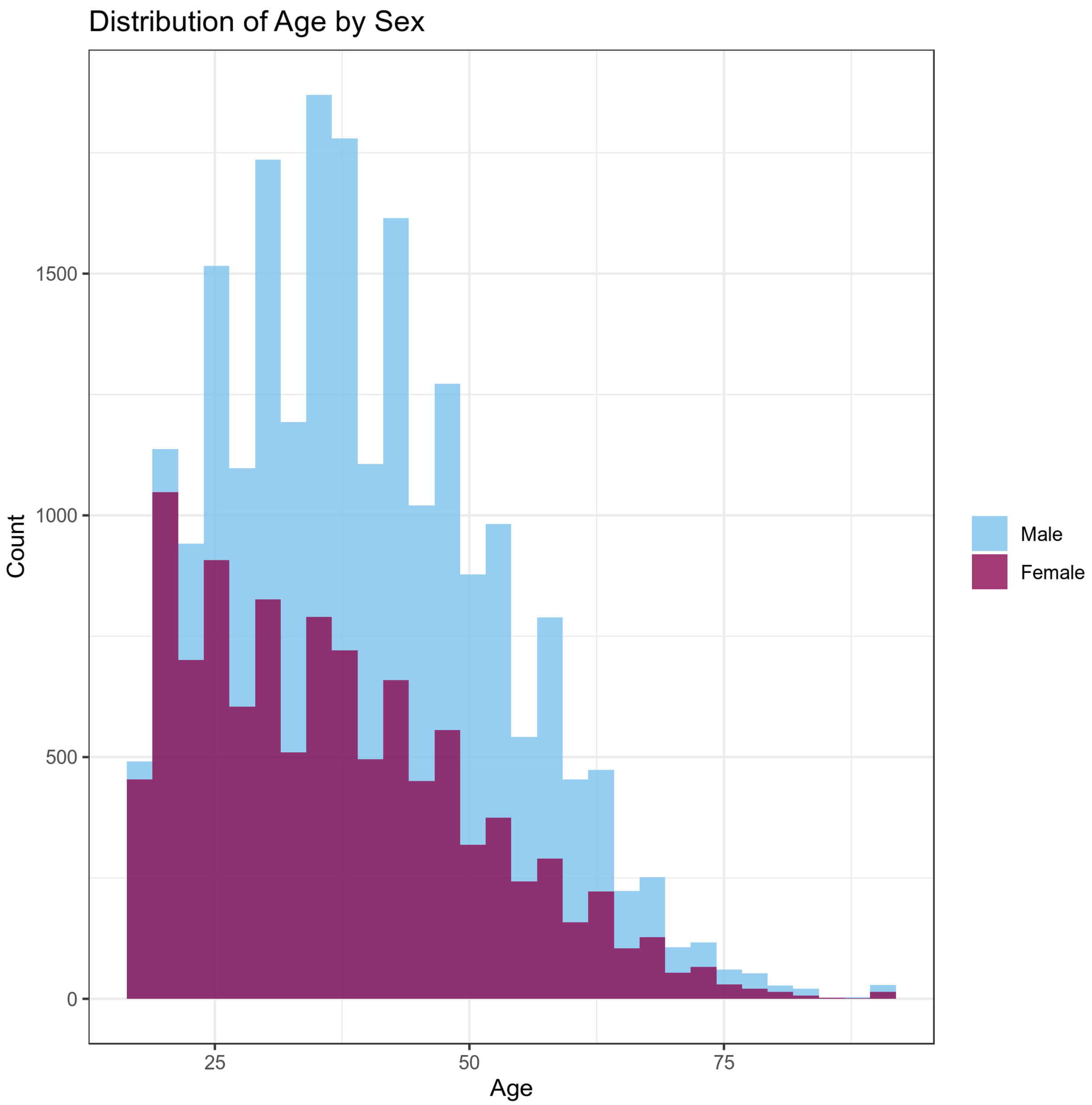
Are the number of hours someone works per week correlated with their age, relationship, education level or sex?

Exploratory Data Analysis

In this section, we will get to know our dataset better by exploring the relationship between certain factors.

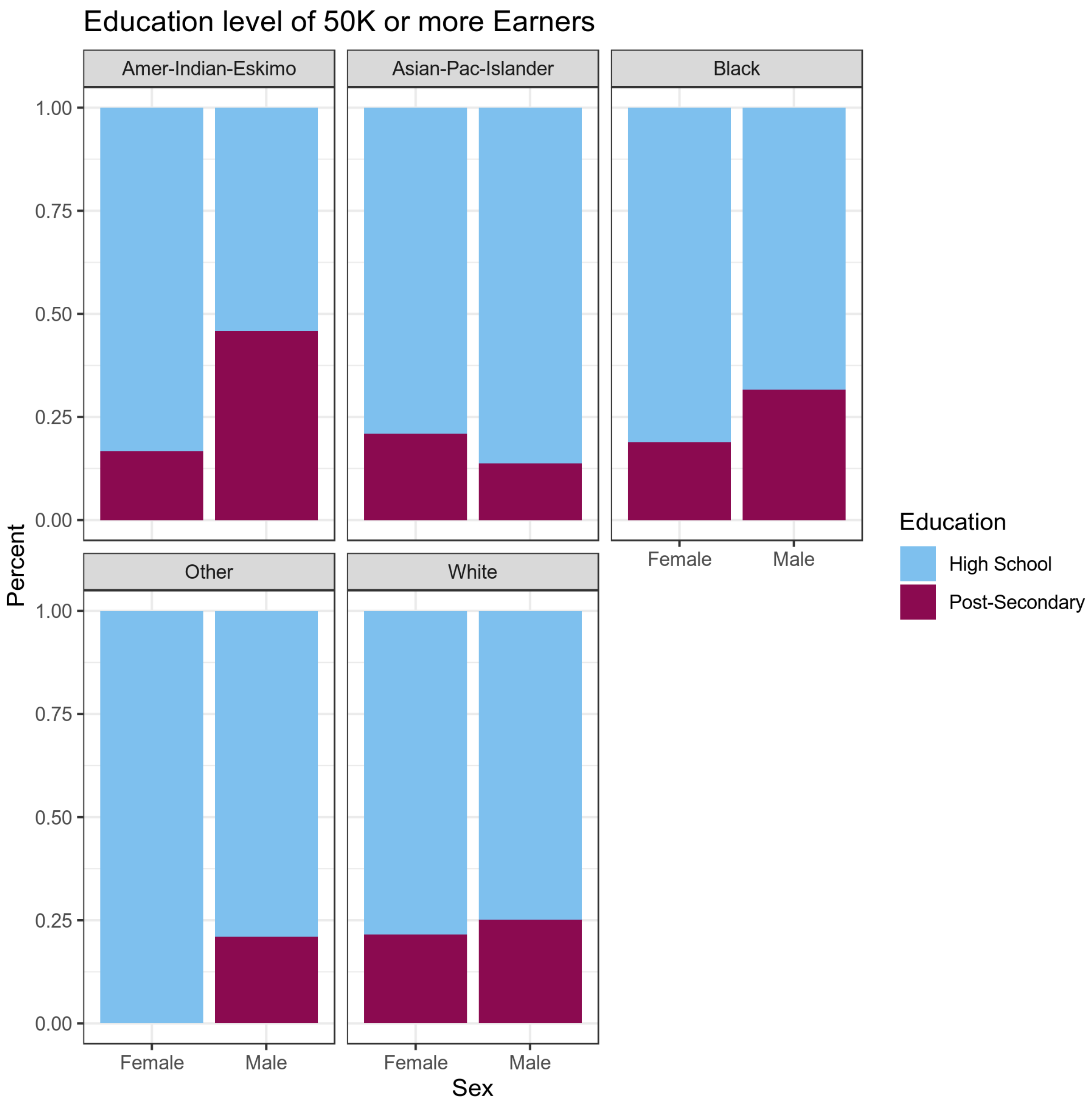
Age and Sex

The plot below shows that there are more male employees than female employees and that the majority of working males are older than working females since the male (blue curve) have a peak shifted to the right with respect to female (red peak).



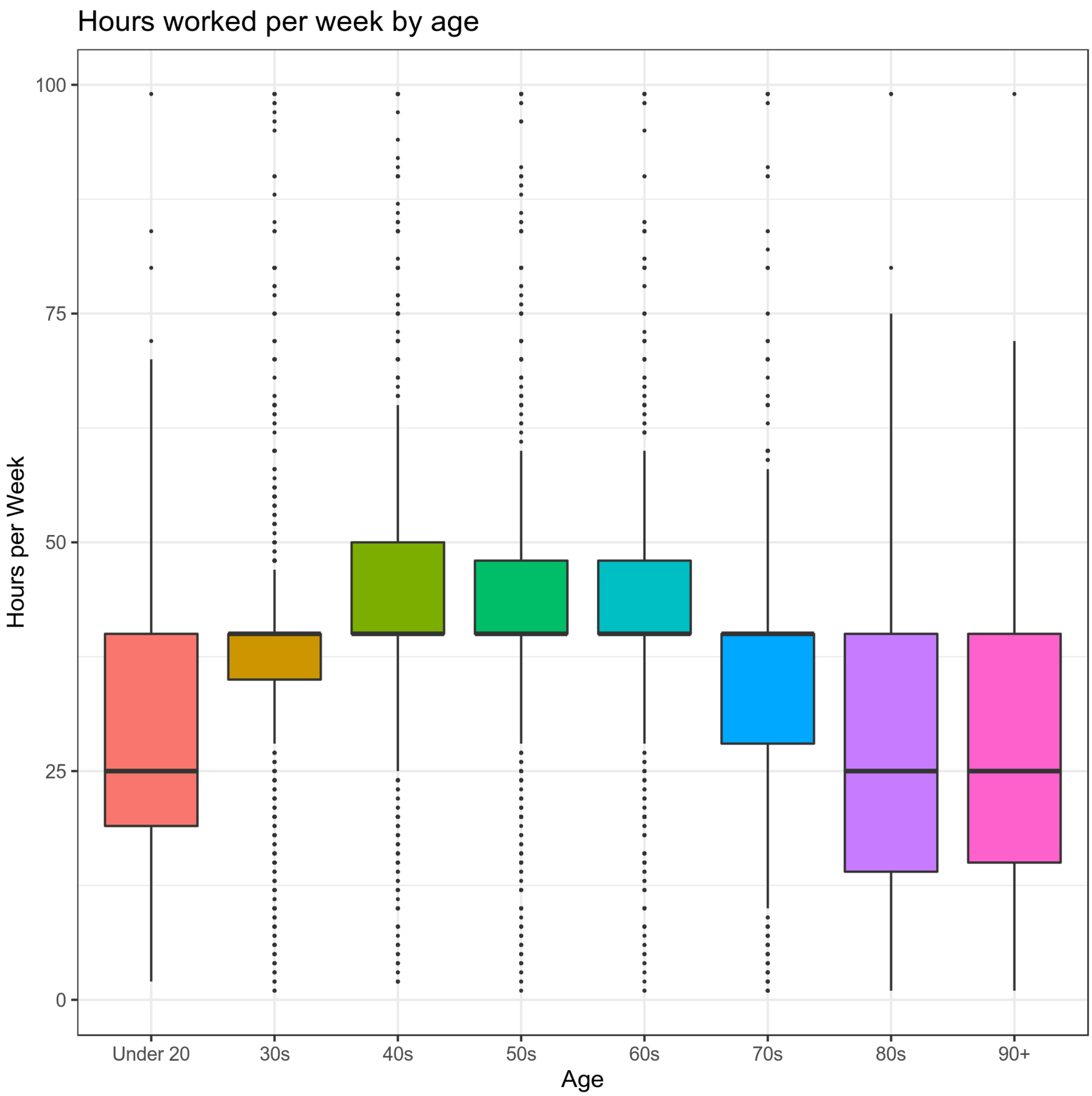
Educational Leve and Income

We observe from the following graphs that a majority of individuals earning greater than $50,000 a year only accomplished high school irrespective of sex or ethnical background.



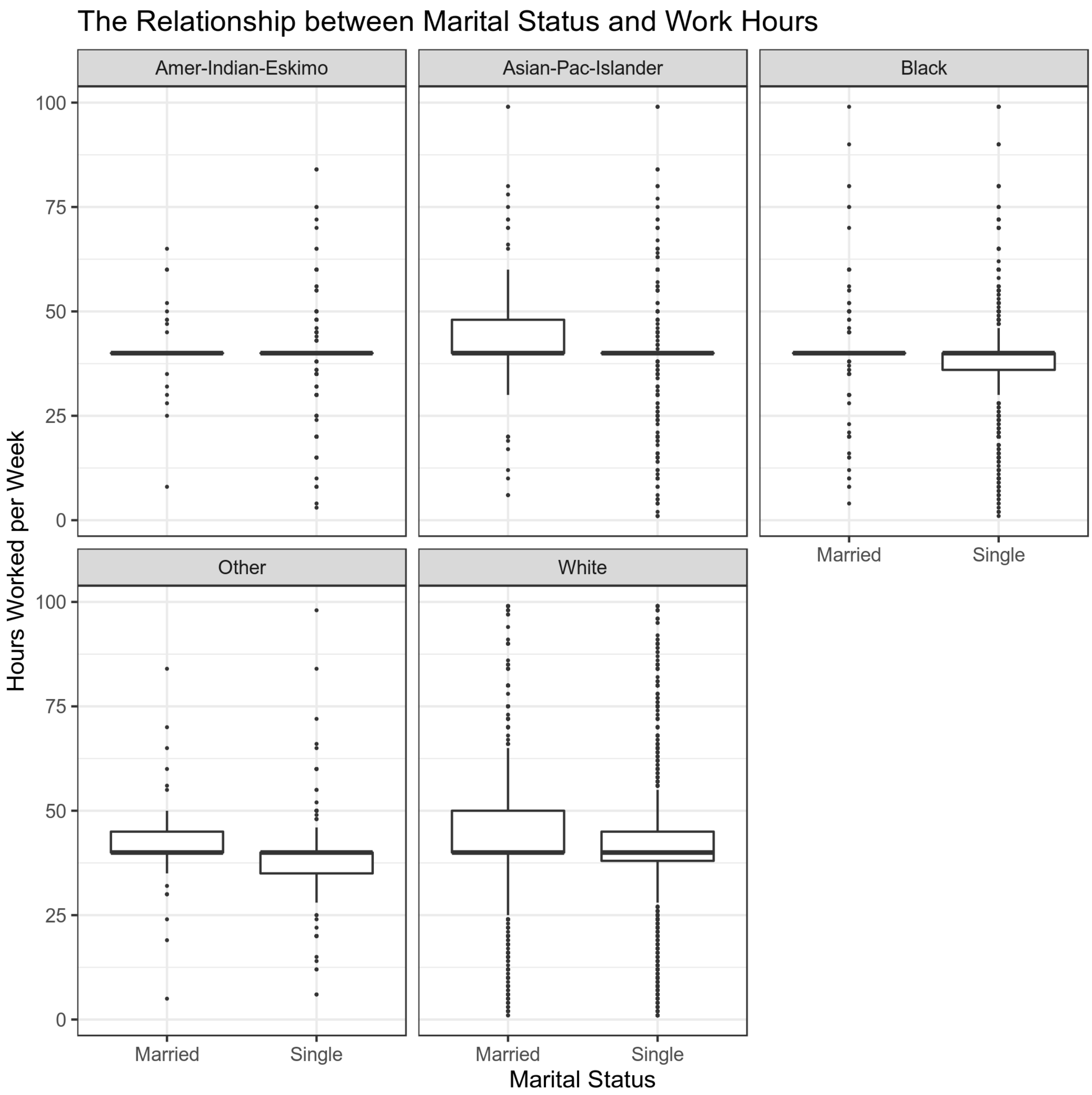
Number of Work Hours and Age

We can deduce from the graph below that individuals work the most hours between their 40's and 60's (probably full time at 40 hours or more a week) and that employees under 20 and over 80 years of age work the same number of hours (probably part time at 25 hours)



Marital Status and Number of Hours Worked

The plot below shows that the working hours between married individuals and single employees are similar.



Analysis

We have performed linear regression of hours per week vs. each variable separately, as well as linear regression using all these variables together.

This was done using the lm function of the purr package. For example: lm(hours_per_week~education,data)

For categorical variables sex, education and relationship, the intercept is the defaul reference group, where the "estimate" is the mean of that group, and the "estimates" of all other variables are the differences in means between that group and the reference. The statistic is the t-statistic comparing these means, with a given p-value reporting the significance of this difference.

Results

Hours vs. Relationship

```
## # A tibble: 6 x 5
##   term                       estimate std.error statistic   p.value
##   <chr>                         <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                   44.1      0.102     432.  0.       
## 2 relationshipNot-in-family     -3.52     0.164     -21.4 3.32e-101
## 3 relationshipOther-relative    -7.11     0.389     -18.3 1.60e- 74
## 4 relationshipOwn-child        -10.9      0.194     -55.9 0.       
## 5 relationshipUnmarried         -5.02     0.225     -22.3 1.04e-109
## 6 relationshipWife              -7.26     0.314     -23.1 1.42e-117
```

In this case, we are comparing the mean hours worked per week of husbands (intercept) to each other relationship category. It appears that husbands work more than any other age group, (as seen by the negative estimates), with significant p-values in each case.

Hours vs. Sex

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    36.4      0.116     314.        0
## 2 sexMale         6.02     0.142      42.5       0
```

From this output, we can see that the average hours worked per week for women is 36.1, and men work 6 more hours per week on average.

Hours vs. Age

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  38.0      0.205       186.  0.      
## 2 age           0.0622   0.00500      12.4 2.01e-35
```

Age

Hours vs. Education

```
## # A tibble: 16 x 5
##    term                  estimate std.error statistic  p.value
##    <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)             37.1       0.397    93.4   0.      
##  2 education11th           -3.13      0.531    -5.89  4.01e- 9
##  3 education12th           -1.27      0.704    -1.81  7.10e- 2
##  4 education1st-4th         1.20      1.02      1.19  2.36e- 1
##  5 education5th-6th         1.85      0.773     2.39  1.70e- 2
##  6 education7th-8th         2.31      0.620     3.73  1.90e- 4
##  7 education9th             0.992     0.665     1.49  1.36e- 1
##  8 educationAssoc-acdm      3.45      0.543     6.36  2.09e-10
##  9 educationAssoc-voc       4.56      0.513     8.88  7.05e-19
## 10 educationBachelors       5.56      0.430    12.9   3.33e-38
## 11 educationDoctorate       9.92      0.716    13.9   1.57e-43
## 12 educationHS-grad         3.52      0.414     8.51  1.78e-17
## 13 educationMasters         6.78      0.492    13.8   4.70e-43
## 14 educationPreschool      -0.405     1.74     -0.233 8.16e- 1
## 15 educationProf-school    10.4       0.642    16.2   1.69e-58
## 16 educationSome-college    1.80      0.421     4.27  1.94e- 5
```

In this analysis, the default reference group (intercept) is a 10th grade education. It appears that those with an 11th grade education work 3 hours less (significant p-value), whereas all other with no more than a high school education work the same amount (no significant p-values).

Every other education level higher than highschool worked significantly more hours, as seen by positive estimates of each group and low p-values.

All Variables

```
## # A tibble: 6 x 5
##   term             estimate std.error statistic   p.value
##   <chr>               <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)        40.6       0.496    81.9   0.       
## 2 sexMale             4.04      0.179    22.6   1.23e-112
## 3 education11th      -2.26      0.504    -4.49  7.31e-  6
## 4 education12th      -0.536     0.668    -0.801 4.23e-  1
## 5 education1st-4th    0.305     0.964     0.317 7.51e-  1
## 6 education5th-6th    0.575     0.734     0.783 4.34e-  1
```

Discussion

Overall we believe a there is a more informative way to analyze our research question than the default lm() parameters.

Our data is quite biased in that there are much more men (21790) than women (10771) in data set. This is especially apparent at older ages, for instance there are 742 women over 60 and 1590 men over 60.

Conclusion

More appropriate linear regression analysis need to be done