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

1. Is earning more than 50K correlated with the education level, marital status, and hours worked per week?

Plots showing the relationship between income and each variable separately. For example, we will perform a logistic regression to show the the difference between individuals earning more than 50,000 a year and those who don't using the educational level as the independent variable.

2. Is hours worked per week correlated with age, relationship, education level, or sex?

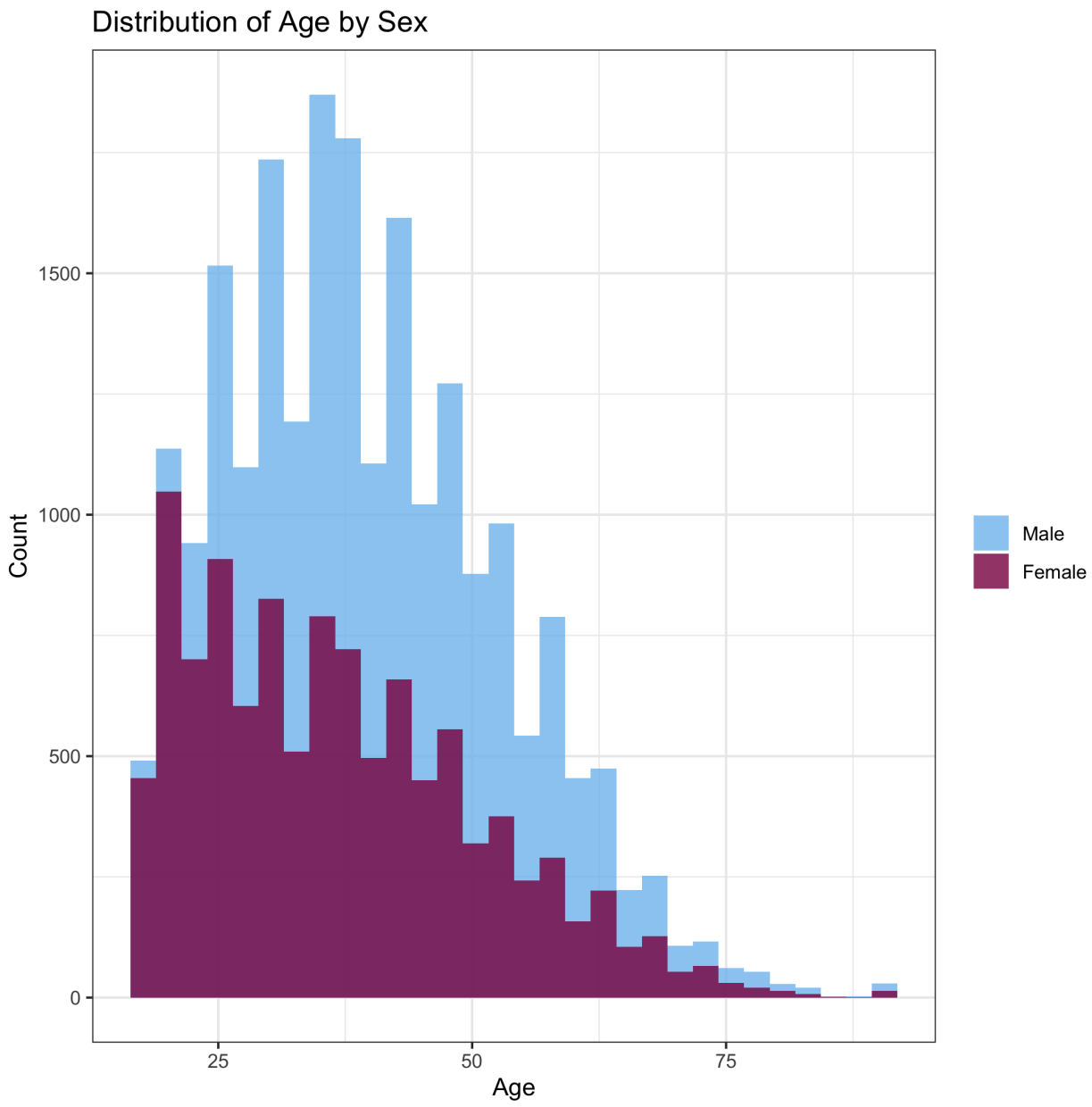
Plots showing the relationship between hours worked and each variable separately. For example, we will use the linear regression model to explore how hours at work is related to variables such as age, relationship, education level, and sex.

Exploratory Data Analysis

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

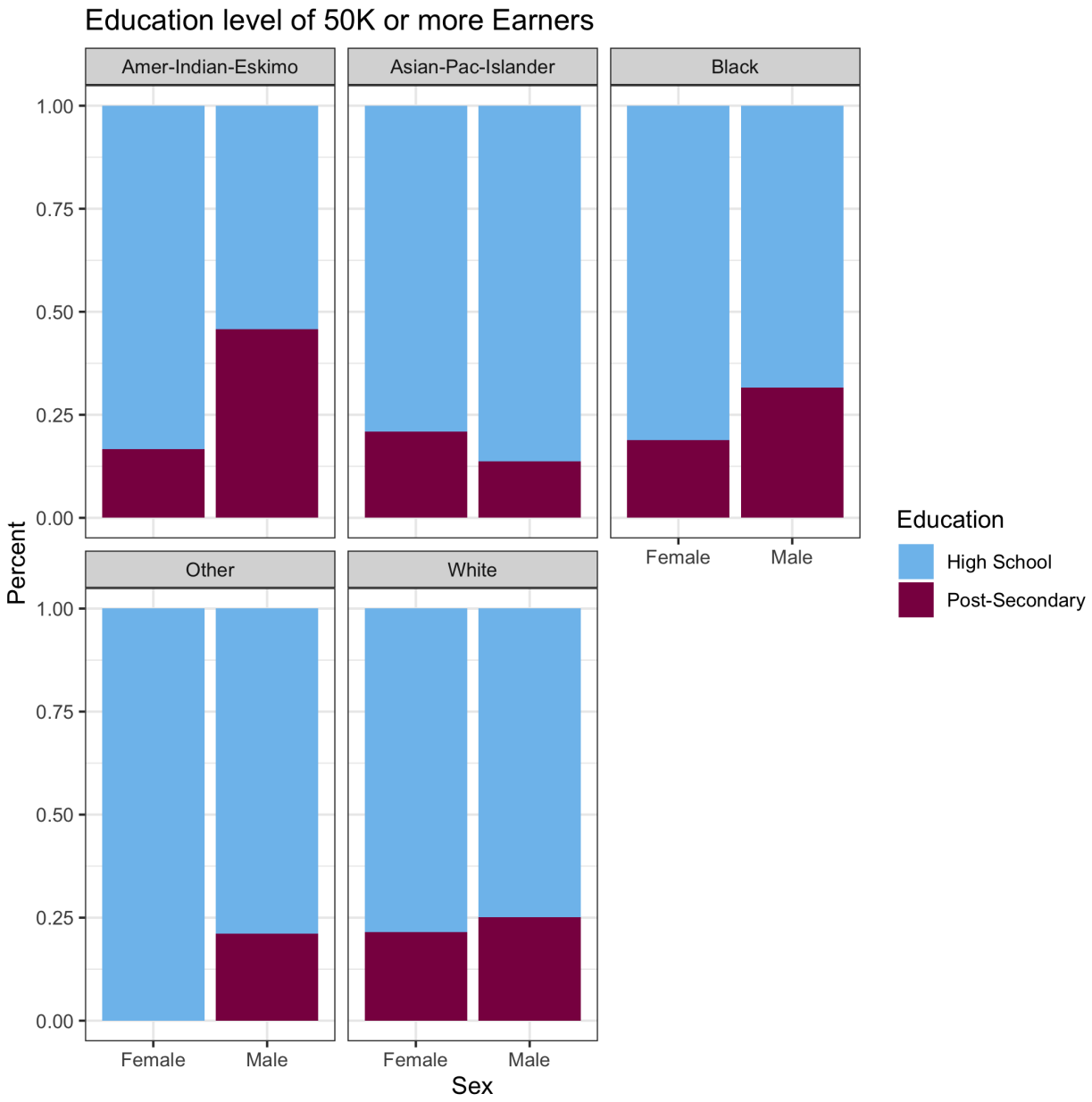
Age and Sex

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).



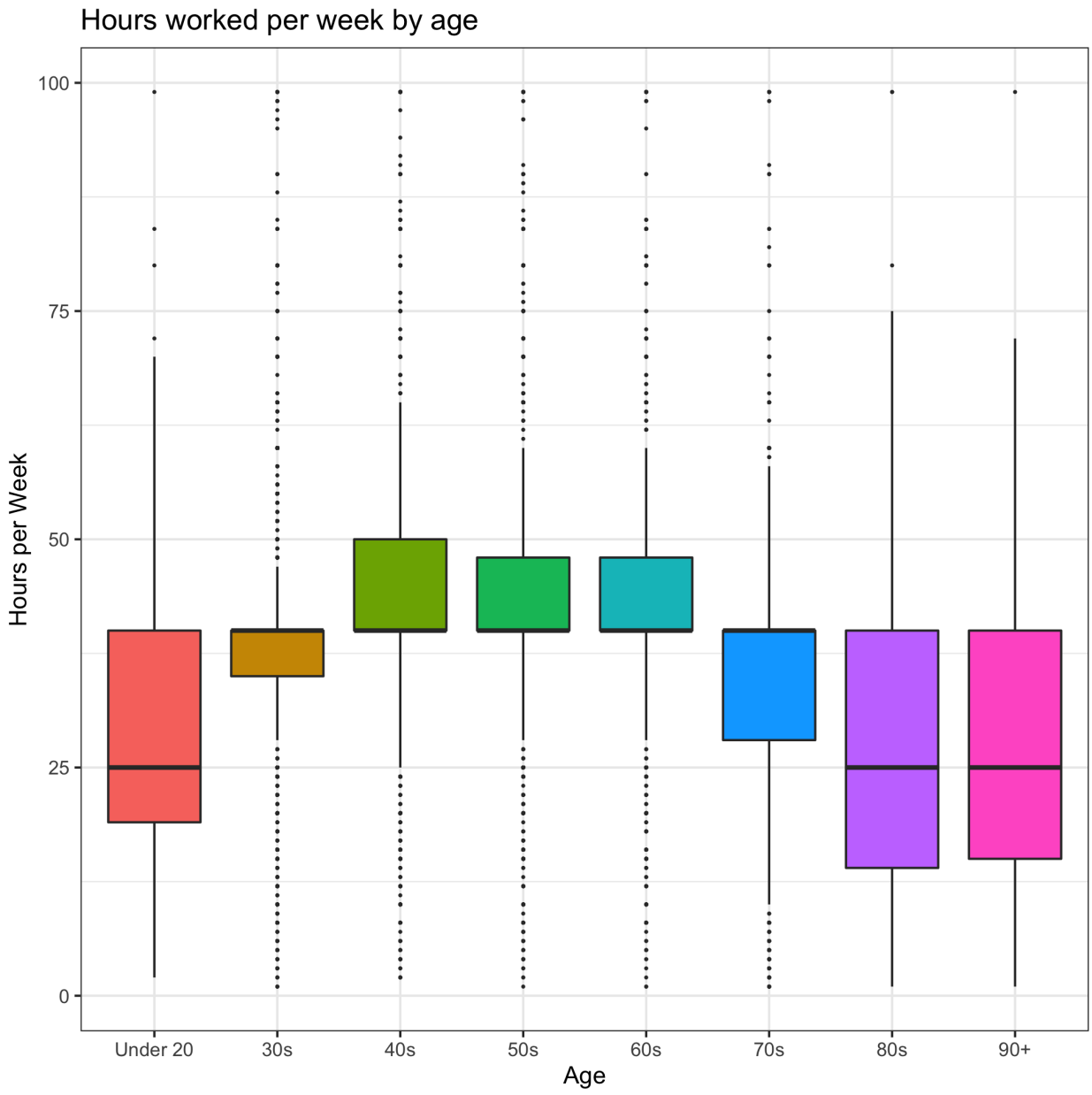
Educational Leve and Income

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.



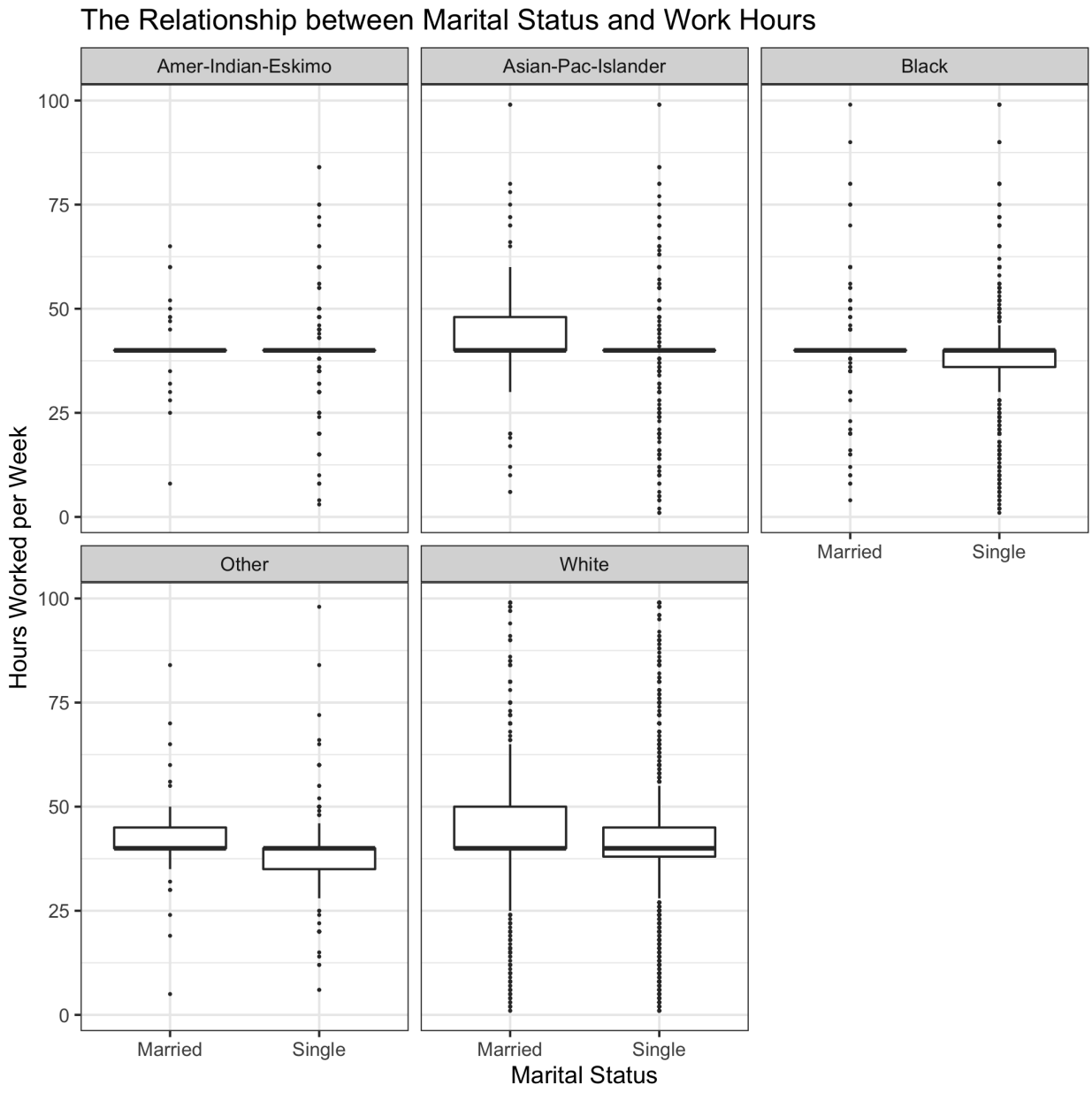
Number of Work Hours and Age

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

1. Is earning more than 50K correlated with the education level, marital status, and hours worked per week?

Plots showing the relationship between income and each variable separately. For example, we will perform a logistic regression to show the the difference between individuals earning more than 50,000 a year and those who don't using the educational level as the independent variable.

```
tidy(readRDS("../data/income_education.rds"))
```

```
## # A tibble: 16 x 5
```

```
##    term                  estimate std.error statistic   p.value
##    <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)              2.64      0.131    20.1   6.82e- 90
##  2 education11th            0.280     0.187     1.50  1.34e-  1
##  3 education12th           -0.148     0.224    -0.659 5.10e-  1
##  4 education1st-4th         0.653     0.436     1.50  1.34e-  1
##  5 education5th-6th         0.344     0.288     1.19  2.33e-  1
##  6 education7th-8th         0.0755    0.210     0.360 7.19e-  1
##  7 education9th             0.250     0.237     1.05  2.92e-  1
##  8 educationAssoc-acdm     -1.54      0.149   -10.3   8.62e- 25
##  9 educationAssoc-voc      -1.60      0.145   -11.1   2.10e- 28
## 10 educationBachelors      -2.30      0.134   -17.1   1.31e- 65
## 11 educationDoctorate      -3.69      0.173   -21.4   3.02e-101
## 12 educationHS-grad        -0.981     0.134    -7.31  2.64e- 13
## 13 educationMasters        -2.87      0.140   -20.5   3.01e- 93
## 14 educationPreschool      10.9      75.0       0.146 8.84e-  1
## 15 educationProf-school    -3.66      0.162   -22.6   2.88e-113
## 16 educationSome-college   -1.19      0.135    -8.86  8.10e- 19
```

```
augment(readRDS("../data/income_education.rds"))
```

```
## # A tibble: 32,561 x 9
##    income   education .fitted .se.fit .resid     .hat .sigma  .cooksd .std.resid
##    <fct>    <fct>       <dbl>   <dbl>  <dbl>    <dbl>  <dbl>    <dbl>      <dbl>
##  1 under_5~ Bachelors   0.344  0.0277  1.04   1.87e-4  0.987  8.27e-6      1.04
##  2 under_5~ Bachelors   0.344  0.0277  1.04   1.87e-4  0.987  8.27e-6      1.04
##  3 under_5~ HS-grad     1.66   0.0267  0.590  9.52e-5  0.987  1.13e-6      0.590
##  4 under_5~ 11th        2.92   0.133   0.324  8.51e-4  0.987  2.87e-6      0.324
##  5 under_5~ Bachelors   0.344  0.0277  1.04   1.87e-4  0.987  8.27e-6      1.04
##  6 under_5~ Masters    -0.227  0.0485  1.28   5.80e-4  0.987  4.56e-5      1.28
##  7 under_5~ 9th         2.89   0.198   0.329  1.95e-3  0.987  6.77e-6      0.329
##  8 over_50K HS-grad     1.66   0.0267 -1.92   9.52e-5  0.987  3.14e-5     -1.92
##  9 over_50K Masters    -0.227  0.0485 -1.08   5.80e-4  0.987  2.89e-5     -1.08
## 10 over_50K Bachelors   0.344  0.0277 -1.33   1.87e-4  0.987  1.65e-5     -1.33
## # ... with 32,551 more rows
```

```
glance(readRDS("../data/income_education.rds"))
```

```
## # A tibble: 1 x 7
##   null.deviance df.null  logLik    AIC    BIC deviance df.residual
##           <dbl>   <int>   <dbl>  <dbl>  <dbl>    <dbl>       <int>
## 1        35948.   32560 -15862. 31755. 31890.   31723.       32545
```

```
tidy(readRDS("../data/income_education.rds"))
```

```
## # A tibble: 16 x 5
##    term                  estimate std.error statistic   p.value
##    <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)              2.64      0.131    20.1   6.82e- 90
##  2 education11th            0.280     0.187     1.50  1.34e-  1
##  3 education12th           -0.148     0.224    -0.659 5.10e-  1
##  4 education1st-4th         0.653     0.436     1.50  1.34e-  1
```

```
##  5 education5th-6th        0.344      0.288     1.19  2.33e-  1
##  6 education7th-8th        0.0755     0.210     0.360 7.19e-  1
##  7 education9th            0.250      0.237     1.05  2.92e-  1
##  8 educationAssoc-acdm    -1.54       0.149   -10.3   8.62e- 25
##  9 educationAssoc-voc     -1.60       0.145   -11.1   2.10e- 28
## 10 educationBachelors     -2.30       0.134   -17.1   1.31e- 65
## 11 educationDoctorate     -3.69       0.173   -21.4   3.02e-101
## 12 educationHS-grad       -0.981      0.134    -7.31  2.64e- 13
## 13 educationMasters       -2.87       0.140   -20.5   3.01e- 93
## 14 educationPreschool     10.9       75.0       0.146 8.84e-  1
## 15 educationProf-school   -3.66       0.162   -22.6   2.88e-113
## 16 educationSome-college  -1.19       0.135    -8.86  8.10e- 19
```

```
readRDS("../data/income_marital_status.rds")
```

```
##
## Call:  glm(formula = income ~ marital_status, family = "binomial", data = data)
##
## Coefficients:
##                          (Intercept)      marital_statusMarried-AF-spouse
##                               2.1513                              -1.8889
##    marital_statusMarried-civ-spouse  marital_statusMarried-spouse-absent
##                              -1.9379                               0.2730
##         marital_statusNever-married              marital_statusSeparated
##                               0.8816                               0.5249
##               marital_statusWidowed
##                               0.2173
##
## Degrees of Freedom: 32560 Total (i.e. Null);  32554 Residual
## Null Deviance:       35950
## Residual Deviance: 28880     AIC: 28900
```

```
readRDS("../data/income_hours.rds")
```

```
##
## Call:  glm(formula = income ~ hours_per_week, family = "binomial", data = data)
##
## Coefficients:
##    (Intercept)  hours_per_week
##        3.10007        -0.04645
##
## Degrees of Freedom: 32560 Total (i.e. Null);  32559 Residual
## Null Deviance:       35950
## Residual Deviance: 34190     AIC: 34190
```

```
readRDS("../data/income_all.rds")
```

```
##
## Call:  glm(formula = income ~ education + marital_status + hours_per_week,
##     family = "binomial", data = data)
##
## Coefficients:
```

```
##                           (Intercept)                          education11th
##                               5.02265                               0.04631
##                          education12th                      education1st-4th
##                              -0.39628                               0.90683
##                       education5th-6th                      education7th-8th
##                               0.62407                               0.49644
##                          education9th                    educationAssoc-acdm
##                               0.40758                              -1.62378
##                     educationAssoc-voc                     educationBachelors
##                              -1.51804                              -2.42383
##                     educationDoctorate                       educationHS-grad
##                              -3.72508                              -0.85965
##                       educationMasters                     educationPreschool
##                              -2.97906                              11.73383
##                  educationProf-school                  educationSome-college
##                              -3.58791                              -1.33828
##      marital_statusMarried-AF-spouse      marital_statusMarried-civ-spouse
##                              -2.21161                              -2.04671
## marital_statusMarried-spouse-absent           marital_statusNever-married
##                               0.24019                               0.87213
##               marital_statusSeparated                 marital_statusWidowed
##                               0.25945                              -0.24583
##                         hours_per_week
##                              -0.02987
##
## Degrees of Freedom: 32560 Total (i.e. Null);  32538 Residual
## Null Deviance:       35950
## Residual Deviance: 24510     AIC: 24560
```

```
readRDS("../data/hours_age.rds")
```

```
##
## Call:
## lm(formula = hours_per_week ~ age, data = data)
##
## Coefficients:
## (Intercept)          age
##    38.03620      0.06224
```

```
readRDS("../data/hours_relationship.rds")
```

```
##
## Call:
## lm(formula = hours_per_week ~ relationship, data = data)
##
## Coefficients:
##                 (Intercept)   relationshipNot-in-family
##                      44.120                      -3.524
## relationshipOther-relative       relationshipOwn-child
##                      -7.114                     -10.851
##      relationshipUnmarried            relationshipWife
##                      -5.017                      -7.259
```

```r
readRDS("../data/hours_education.rds")
```

```
##
## Call:
## lm(formula = hours_per_week ~ education, data = data)
##
## Coefficients:
##           (Intercept)          education11th          education12th
##               37.0525                -3.1266                -1.2719
##      education1st-4th       education5th-6th       education7th-8th
##                1.2034                 1.8454                 2.3144
##          education9th    educationAssoc-acdm     educationAssoc-voc
##                0.9922                 3.4517                 4.5582
##    educationBachelors     educationDoctorate       educationHS-grad
##                5.5615                 9.9208                 3.5229
##      educationMasters     educationPreschool   educationProf-school
##                6.7838                -0.4055                10.3728
## educationSome-college
##                1.7998
```

```r
readRDS("../data/hours_sex.rds")
```

```
##
## Call:
## lm(formula = hours_per_week ~ sex, data = data)
##
## Coefficients:
## (Intercept)      sexMale
##      36.410        6.018
```

```r
  #Relationship of all variables together:
readRDS("../data/hours_sex.rds")
```

```
##
## Call:
## lm(formula = hours_per_week ~ sex, data = data)
##
## Coefficients:
## (Intercept)      sexMale
##      36.410        6.018
```

Results

Discussion

Conclusion