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 Variables

Variable	Type	Description
age	int	Age of individual

Variable	Type	Description
workclass	chr	e.g. private, self-emplowed, federal government, never worked, etc.
fnlwgt	int	Final weights: weighted sums of the socio-economic characteristics of the individual. People with similar demographics have similar weights.
education	chr	Highest education recieved
educationnum	factor	Numerical code for highest education recieved
marital_status	int	e.g. married, never married, divorced, etc.
occupation	chr	Occupation of individual
relationship	chr	Relation of individual in family. e.g. wife, child, husband, unmarried
race	chr	Asian-Pacific Islander, Native American, White, Black, other
sex	chr	Male or Female
capital_gain	int	Profit from capital assets such as investments, real estate, etc.
capital_loss	int	Loss from capital assets
hours_per_week		The number of hours that the individual works per week
country	chr	Country of origin
income	chr	Whether individual is predicted to make over or under 50K

mentioned the renamed/ grouped factors here

The Research Questions

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

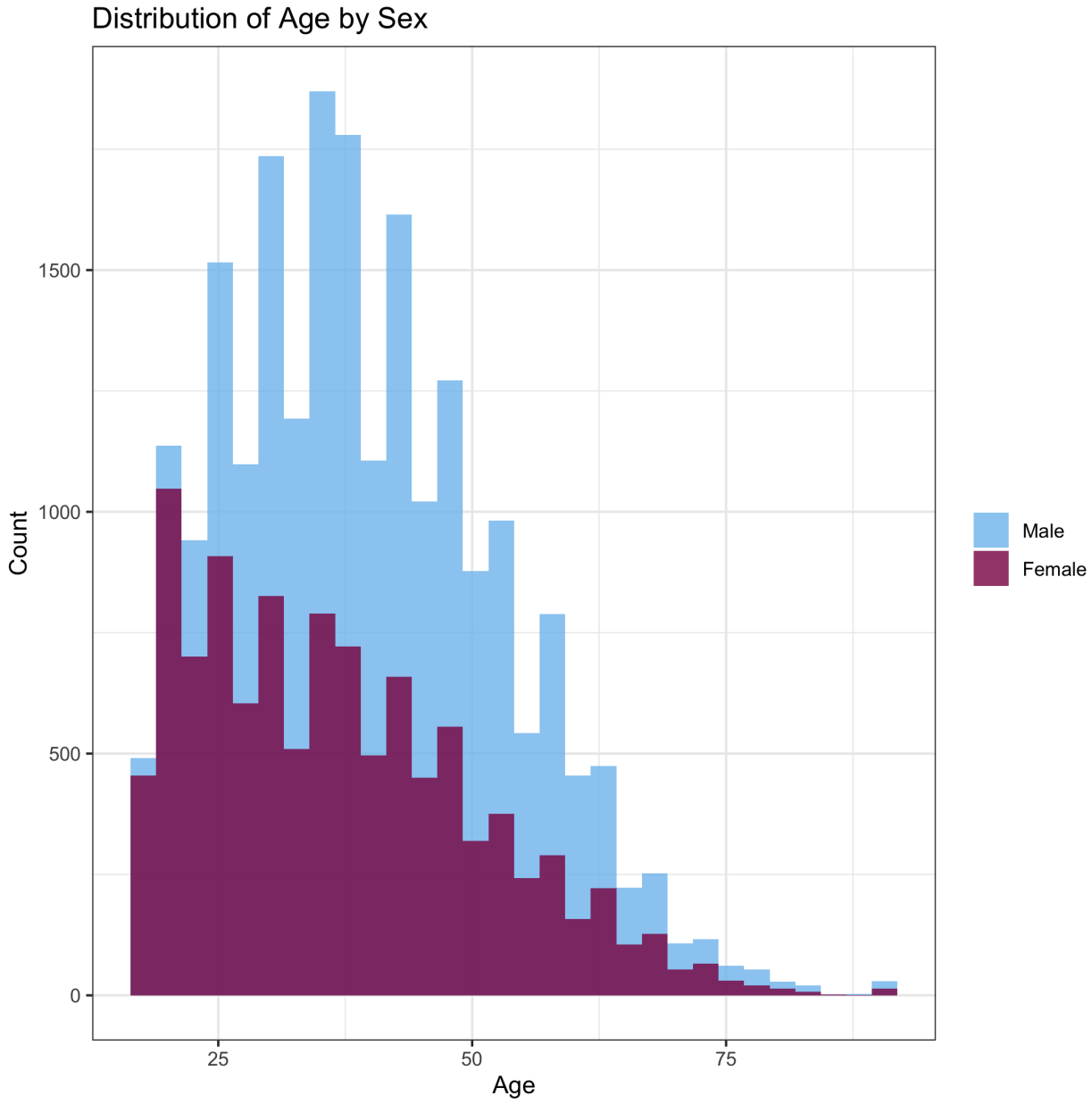
Plots showing the relationship between hours worked and each variable separately. For example, we will use the linear regression model to explore how hours at work is related to variables such as age, relationship, education level, and sex.

Exploratory Data Analysis

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

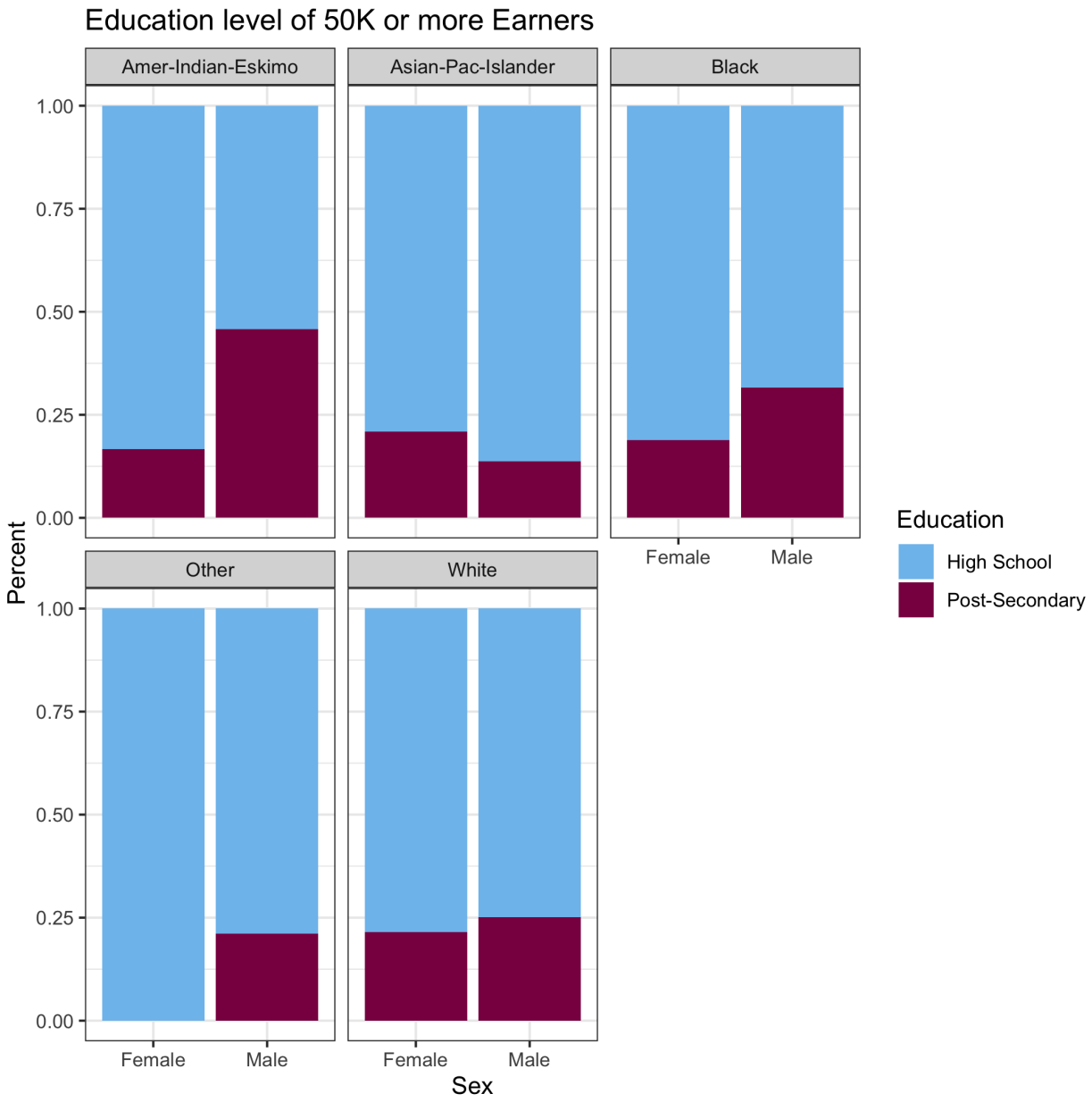
Age and Sex

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).



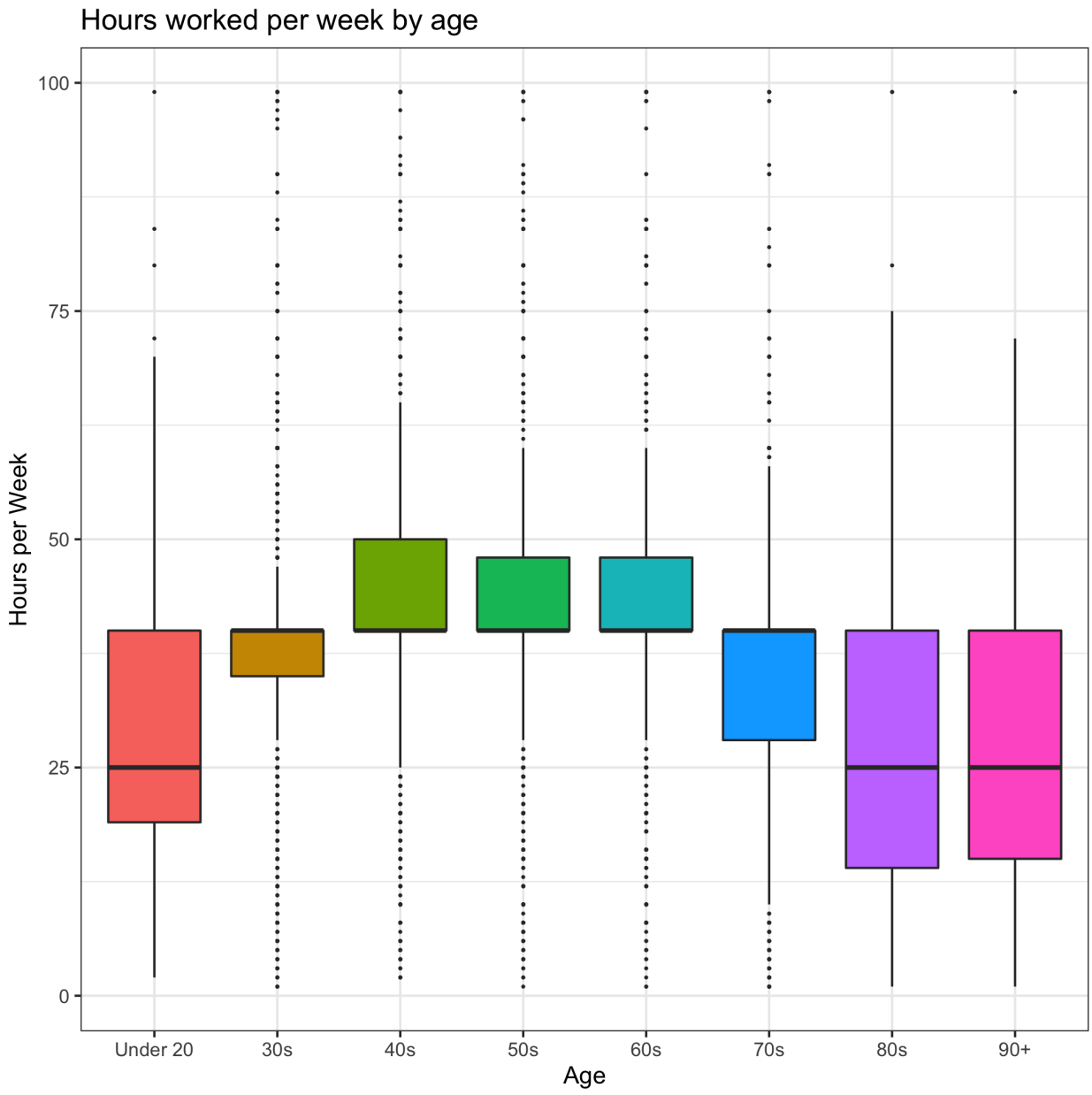
Educational Leve and Income

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.



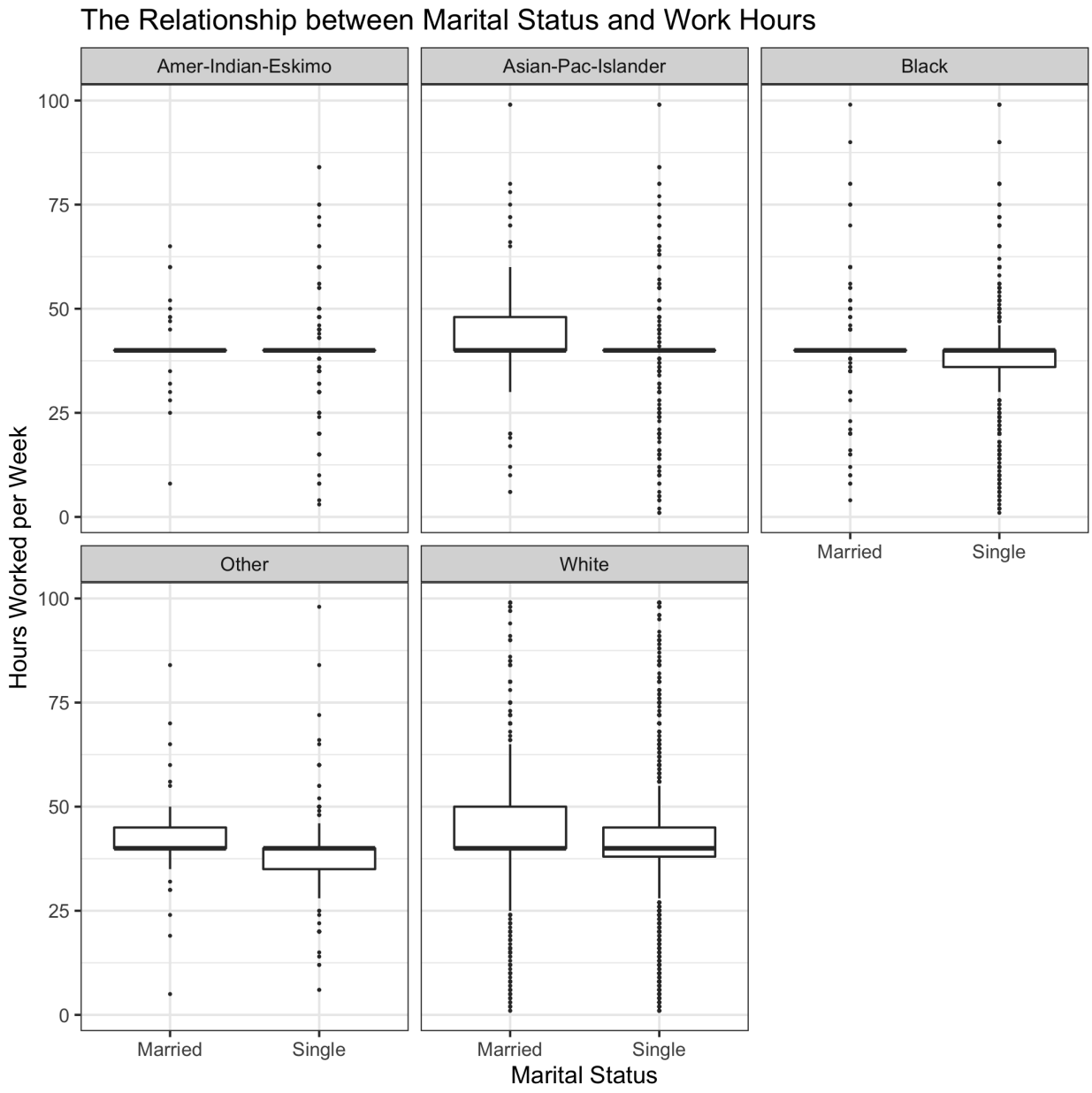
Number of Work Hours and Age

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.

«««< HEAD

#{r} `tidy(readRDS("../data/income_education.rds")) augment(readRDS("../data/income_education.rds")) glance(readRDS("../data/income_education.rds"))`

======= »»»> upstream/master #"'{r, readRDS for analysis#2} readRDS("../data/hours_age.rds")

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

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

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

#Relationship of all variables together: readRDS("../data/hours_all.rds")

```

**Results**

**Discussion**

**Conclusion**
```