CS518 - Final Project Progress Report

Nora Quick

Executive Summary

This progress report will have three main sections to it. These sections are a description of the dataset and the questions that will be answered in the final report, explanation of response variables used in the chose models and methods used to answer the questions listed, and an exploratory plot to introduce some of the sata/variables.

Body

Dataset Desciption & Questions

The dataset I chose to work with is the cencus data that provides adult income information. This dataset includes many things such as income (greater than or less than $50K), age, working class, education, marital status, and more. In additional to this there are three data files provided for this set including adult.data, adult.names, and adult.test. For this report the most important data is the adult.data which holds all of the variables listed above and for simplicitiy will the the only file worked with.

1. Do either age or hours work increase the likelyhood of earning more than $50K a year? If so, which one or both and what is the most significant age/hours worked.
2. What are the causes, if any, of divorce?

Reponse Variables and Methods Used to Address the Questions

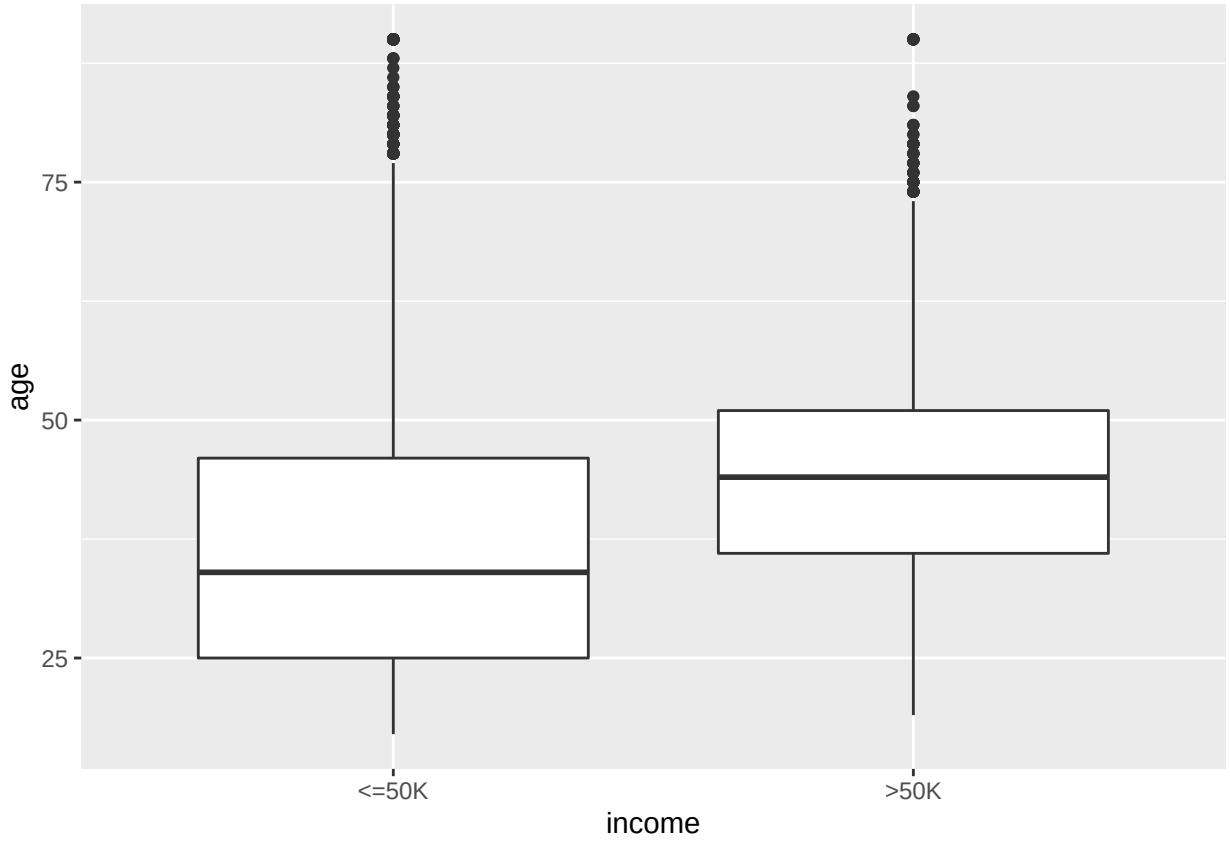
The two main reponse variables I will be using are income and marital-status.

I will be using generalized linear model wiht poisson distribution for the first question to find a age and hours worked for an income above and below $50K. I will be using a mixed effect model to look at the different variables and their influence on marital-status

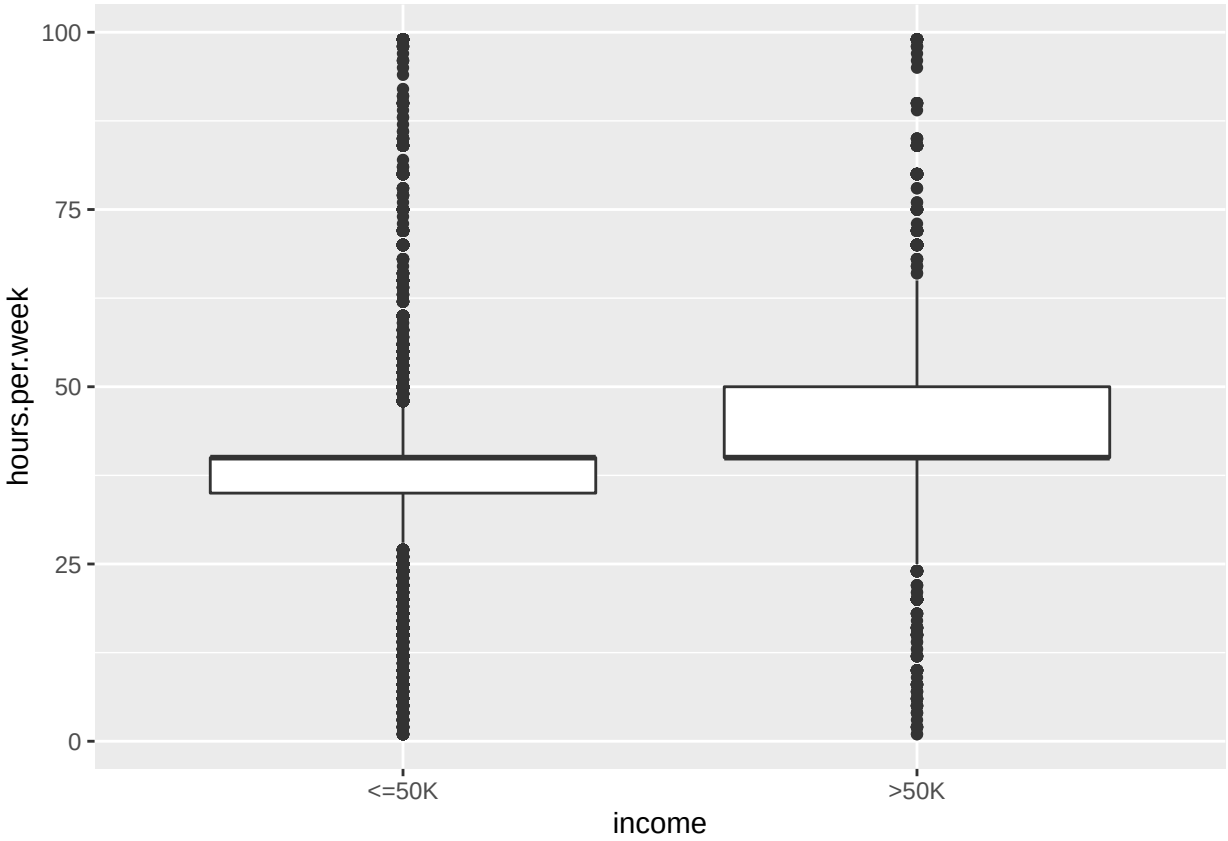
An Exploraroty Plot

```
cencus <- read.csv("adult.csv")

ggplot(cencus, aes(x = income, y = age)) +
  geom_boxplot()
```




```
ggplot(cencus, aes(x = income, y = hours.per.week)) +
  geom_boxplot()
```



Conclusions/Discussion

Looking at the data table and exploratory graphs above for question one my initial thoughts are that age will be a significant explanatory variable in income but number of hours worked a week will not be significant or if it is it will be smaller than age.

Looking at the data for questions two I believe that carrer and education level are the two most singificant explanatory variables for marital status, specifically causes for divorce.

Appendices

Currently there is not enough excess data to provide additional information or technical analysis.