Data Analysis and Statistical Inference - Project Proposal

Mariame M

September 20, 2014

Impact of Number of Siblings on Education Level

Background

Only children have been the subjects of numerous studies, sometimes stigmatized as spoiled brats and other times as high achiever. In this project, we will analyze the relationship between the number of siblings a person has or had, and her/his level of education. The question we will try to asnwer in this analysis is thus the following:

Are people with no or less siblings better educated?

To answer that question, we will use data from the General Social Survey (GSS) Cumulative File 1972-2012, which provides a sample of selected indicators on contemporary American society. Detailed information on this file can be found at https://d396qusza40orc.cloudfront.net/statistics%2Fproject%2Fgss1.html.

Variables

From the GSS file, we will use the following variables:

- **sibs**: respondent's number of brothers and sisters - numeric variable;
- **degree**: respondent's highest degree - ordinal variable.

Data Processing

We first load the required libraries:

```r
setwd("~/Repositories/Coursera_DataAnalysis_Duke/Project/")

library(reshape2)
library(ggplot2)
```

We then load the data (and cache it) and get some summary statistics:

```r
load(url("http://bit.ly/dasi_gss_data"))

# only keep count of siblings, degree and constant family income
data <- gss[,c("sibs","degree")]

# get statistics
summary(data)
```

```
##       sibs                  degree     
##  Min.   : 0.0   Lt High School:11822  
##  1st Qu.: 2.0   High School   :29287  
##  Median : 3.0   Junior College: 3070  
##  Mean   : 3.9   Bachelor      : 8002  
##  3rd Qu.: 5.0   Graduate      : 3870  
##  Max.   :68.0   NA's          : 1010  
##  NA's   :1679
```

We see that all three variables contain missing values (NA). In order to prevent this data to create bias in our analysis, we will in those missing values using the following strategy:

- sibs: the median number of siblings being 3, we use that number to fill in the NAs in that variable;

- degree: most people have responded "High School" as the highest degree they obtained, we use that value to fill in the NAs in that variable.

```
# use median number of siblings
sibs.default <- median(data$sibs, na.rm = TRUE)
data[is.na(data$sibs),"sibs"] <- sibs.default

# use "High School"
degree.default <- "High School"
data[is.na(data$degree),"degree"] <- degree.default
```

Exploratory Analysis

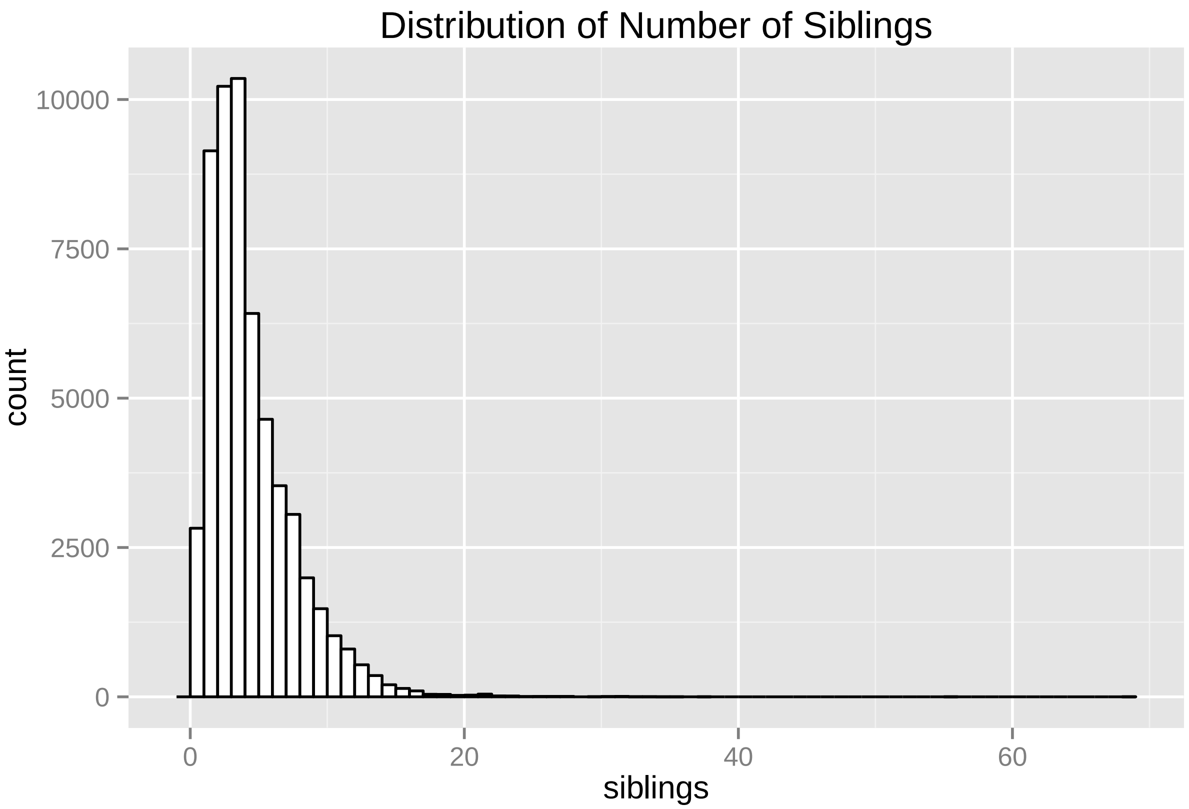
In this section we will draw a few exploratory plots to get a first impression of each variable and the relationships between variables.

1. Single Variable Analysis

1.1. Number of Siblings

We observe that the distribution for the number of siblings is right-skewed and limited to zero on the left. Both of these observations are expected as one cannot have a negative number of siblings. Similarly, we do expect the count of respondents to decrease as the number of siblings increases:

```
ggplot(data, aes(x=sibs)) +
    geom_histogram(binwidth=1, colour="black", fill="white") +
    xlab("siblings") +
    ggtitle("Distribution of Number of Siblings")
```



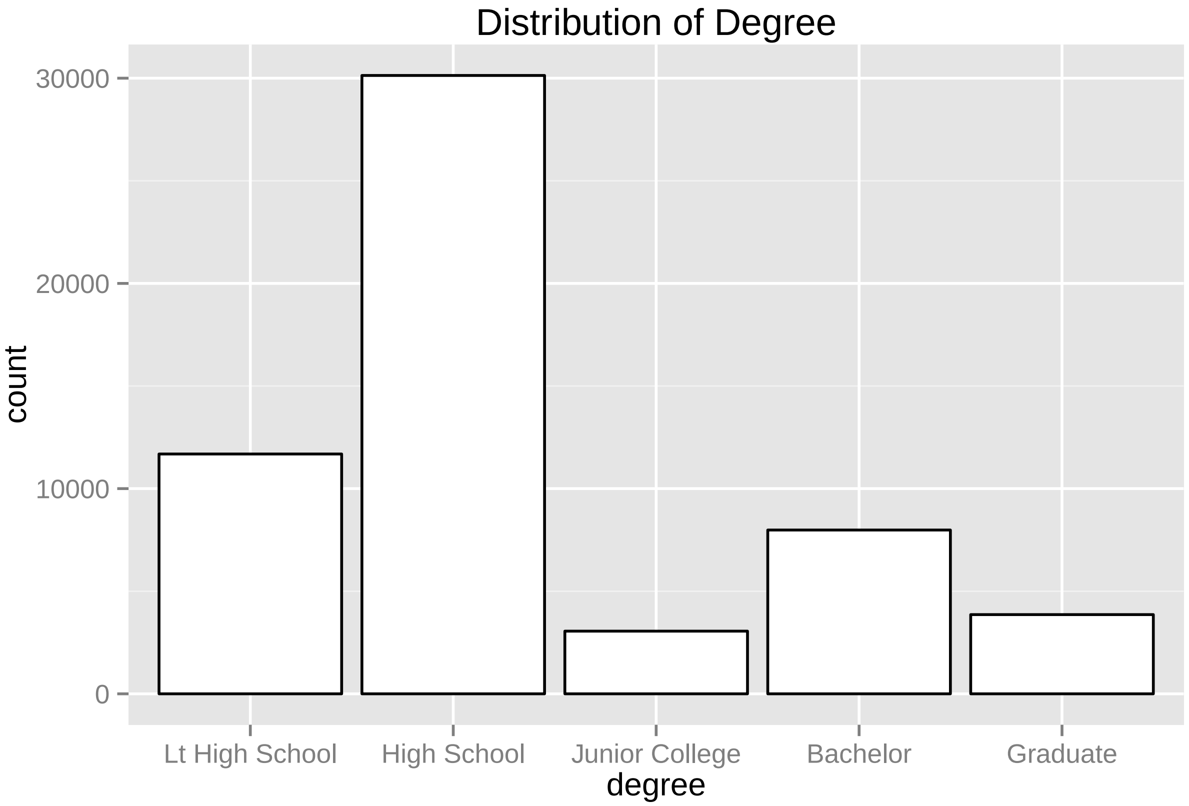
We see that there is a very small number of respondents who have extreme numbers of siblings (well over 60). These data points represents outliers which are likely to skew the resuts of our analysis. Therefore, we remove these entries from our data and will focus on respondents who have 15 or less siblings:

```
data <- data[data$sibs <= 15,]
```

1.2. Education Level

Most respondents hold a high school diloma as their highest degree:

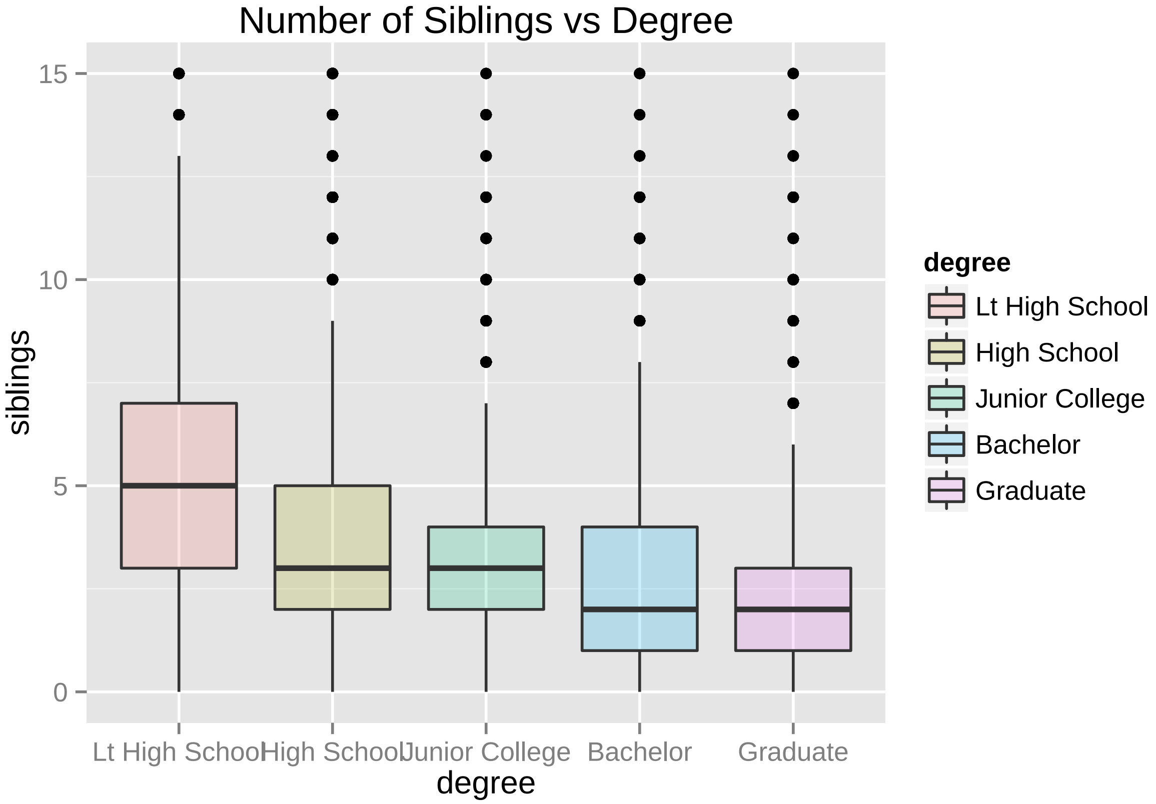
```
ggplot(data, aes(x=degree)) +
    geom_histogram(color="black", fill="white") +
    ggtitle("Distribution of Degree")
```



1.3. Number of Siblings vs Degree

On the boxplots below, we see that as the level of education increases, the number of siblings decreases. For the two highest degrees (bachelor and graduate), we see that while their median number of siblings is similar, their IQR clearly differs, showing a lower IQR as the degree level is higher:

```
ggplot(data, aes(x=degree, y=sibs, fill=degree)) +
    geom_boxplot(alpha=0.2) +
    xlab("degree") +
    ylab("siblings") +
    ggtitle("Number of Siblings vs Degree")
```



Conclusion

From the data gathered and the exploratory analysis provided above, there seems to be a negative relationship between number of siblings and level of education achieved (i.e. the less siblings, the higher the degree). This however does not seem to be entirely true for family income, as we have seen that only children do not necessariy have the highest incomes.

In the second part of this project, we will run hypothesis testings to further analyze these first impressions.