Census Data

Charmayne Patterson

7/23/2019

Introduction

In order to explore the potential relationships between several variables related to the status of African Americans in Philadeplhia in 1847, I must start off by reading the census data.

```
library(tidyverse) #The swiss knife!
library(RCurl) #To read the csv file from GitHub directly
```

#Data Wrangling The original data frame consisted of census data and consequently included numerous variables. I was able to utilize several data wrangling techniques to transform it into a document that can be easily analyzed. These techniques include removing unnecessary variables, changing "NA" to "0", removing "NA" for numerical data, and changing the names of some columns.

Removing Unnecessary Variables: The original data frame included 46 variables. Many of these variables were not necessary for the classification study that I endeavor to undertake. Removing the variables reduced the data frame to 10 variables. Presently, all of the variables included should be relevant to the capstone project.

Changing "NA" to "0", Removing "NA" for Numerical Data: In order to have the data more actively reflect the characteristics of the specific variables, I opted to change the "NA" assigned to several variables to "0". This made the data more easily understandable as variables such as birthplace could more easily be identified, established. The designation of "NA" was also removed for all numerical data but was maintained for categorical data.

Changing Columns Names: Of the ten variables that are included in the data frame, several of them underwent a name change. When the data was imported into R Studio, some variables names were too long, and others did not accurately describe the particular data. Therefore, the names of the variables were changed to shorter, concise names that more accurately describe the data in the column.

Now that the packages are loaded, I can get the data into R.

```
x <- getURL('https://raw.githubusercontent.com/cmayne5/Cleaned-Data-Set--Census3/master/census3.csv')
census <- read_csv(x)
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

Exploratory Data Analysis

Now that we have ingested the data, let us start exploring it. The first thing to do is to have a quick look at our data. For that, we can use the `str()` function or the more human readable (and intuitive), `glimpse()` function.

```
glimpse(census)
```

```
## Observations: 2,851
## Variables: 11
## $ X1                 <dbl> 1, 2, 4, 5, 7, 9, 15, 16, 17, 19, 20, 23, ...
## $ `Family Size`      <dbl> 3, 3, 6, 7, 4, 3, 3, 4, 2, 6, 2, 3, 2, 4, ...
## $ Males              <dbl> 1, 2, 1, 5, 3, 1, 1, 2, 1, 4, 1, 2, 1, 2, ...
## $ Females            <dbl> 2, 1, 5, 2, 1, 2, 2, 2, 1, 2, 1, 1, 1, 2, ...
```

```
## $ Natives             <dbl> 2, 1, 2, 6, 2, 1, 1, 1, 1, 5, 0, 3, 0, 2, ...
## $ `Not Natives`       <dbl> 1, 2, 4, 1, 2, 2, 2, 3, 1, 1, 2, 0, 2, 2, ...
## $ `Male Occupation`   <chr> "Watchman", "1 lab 1 Waterman", "drives dr...
## $ `Female Occupation` <chr> "Washer", "Washer", "eating house", "domes...
## $ Born.slaves         <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, ...
## $ Can.read            <dbl> 1, 1, 0, 5, 2, 1, 1, 2, 1, 5, 0, 1, 1, 3, ...
## $ Can.write           <dbl> 1, 1, 0, 5, 2, 1, 1, 2, 1, 5, 0, 2, 0, 3, ...
```

Now I can begin exploring the data. The first thing to do is to have a quick look at it, noticing that there are 2851 rows in our dataset with 11 variables. It added the `X1` variable on its own while importing the data. Not only can I see my variables, but I also now know the type of variables, for example, `Male Occupation` is a characteer variable. Should it be a factor?based on the numerous observations and 11 variables, it is obvious that the data must be cleaned in order to make it more manageable and readable.

While carrying out any analysis, it is always best to start off with a question. It is also important to know the metadata, or the data that describes the data. For example, I see the variable `Can.read` has values ranging from 0 to 5 (we will of course check this), but I have no idea what these values mean. I would have assumed that this might be a binary value with 0 indicating the person cannot read and 1 indicating that the person can read. Or this could also mean the total number of people in a household who can read. That sounds more plausible to me. Either way, I will look at the metadata or a data dictionary, if one exists. I can also observe the different values available in the `Can.read` and `Can.write` variables.

```
table(census$Can.read)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  23  31
## 784 585 690 333 206 126  52  44  14   8   3   2   1   1   2
```

```
table(census$Can.write)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  11
## 976 832 511 250 116  71  51  27   9   5   3
```

Looking at this, I am now more inclined to think that these two variables show me the number of people who can read or write in a family. So with these assumptions, what are the questions that come to my mind? Personally I would really like to understand the impact of proffessions, gender and other variables available, on the level of education. For the 'level of education' I have two variables available, viz. `Can.read` and `Can.write`. I also know the size of a family. So perhaps it would be best to divide these two variables with the size of the family and look at this as a percentage to say what percent of a family is well versed with reading or writing. Prior to that I also want to check for discrepancies in the dataset. Given the family size, the sum of the Males and Females should be equal to the family size, considering this era only provided binary options for gender.

```
#Create a new variable called GenderSum and then look at the difference
census <- census %>% mutate(
  GenderSum = Males + Females,
  FamDiff = `Family Size` - GenderSum
)

which(census$FamDiff > 0)
```

```
##  [1]   71   78   95  105  145  178  295  368  421  428  432  651  787  868
## [15] 1008 1054 1487 1523 1557 1964 2472
```

There are about 21 values that don't quite match up.Row 71 provides an example. While it says that the family size is 5, it has only 1 male and 2 females. This doesn't quite add up. So maybe removing these values and the two additional columns we created, as well as the `X1` column.

```
census <- census %>% filter(FamDiff == 0) %>% select( -GenderSum, -FamDiff)
```

Now I can mutate a new variable baased on the percentages discussed above.

```
census <- census %>% mutate(
  Read = round((Can.read/`Family Size`)* 100, 2),
  Write = round((Can.write/`Family Size`)* 100, 2)
)
```

I could also perhaps combine these two newly created variables and have one cummulative variable which is just the mean of these two, to signify the overall education level. After creating this new variable I won't need the original variables.

```
census <- census %>% mutate(
  EduLevel = round((Read + Write)/2, 2)
)
```

I must check and see that are `EduLevel` is not greater that 100%.

```
which(census$EduLevel > 100)
```

```
## [1]  144  165  189  238  392  475 1220
```

Looking at row 144, I see that the family size is 3, but the number of people who can read or write is 6 and 4 respectively. This alerts me to the fact that something is wrong here. I will need to filter for only those values that are less than or equal to 100.

```
census <- census %>% filter(EduLevel <= 100) %>% select(-(Can.read:Write))
```

Ok, so far so good. Now I must address the task of making some sense of the male and femal occupations. There are too many occupations and so they need to be reduced. Looking at the length of unique values of male and female occupation.

```
length(unique(census$`Male Occupation`))
```

```
## [1] 1415
```

```
length(unique(census$`Female Occupation`))
```

```
## [1] 788
```

My suuggestion here would be to combine the unique values for both male and female occupation, then save it as a csv file. Moving forward it can be opened in Excel and then put a common name for multiple categories. For example, for male occupation, I see porter, porter $25.00 month sometimes, porter irregular, porter sometimes 1.25 d, etc. These can all be combined as "porter". This will make the analysis more manageable.

```
Occupations <- as.data.frame(unique(c(unique(census$`Male Occupation`), unique(census$`Female Occupatio
colnames(Occupations) <- 'Occupation'
write_csv(Occupations, 'Occupations.csv')
```

I have attached the csv with the male and filled in a few values as well.

#Now we can replace our Male and Female occupations with the shorter names census‘$MaleOccupation$‘ <
$-occ$ShortName[match(census‘$MaleOccupation$‘, occOccupation)] census‘$FemaleOccupation$‘ < $-occ$ShortName[match(censu
“‘

The above code should provide mangeable names to work with. I can then look at various plots based on the occupation of the males and females. I will answer some of the questions below with plots.

1. What was the most common profession for males?
2. Waht was the most common profession for females?
3. How do these professions measure up when viewed from the education level?

4. Which is the least educated profession?
5. Which is the most educated profession?
6. Is there a relationship between family size and Education Level?
7. Is there any difference between genders as far as education level is concerned?
8. Does being born a slave have any bearing on the education level?
9. What about natives? How are they doing academically?

All these questions can help me to think about the data.

Machine Learning

From a predictive perspective, I am inclined towards predicting the education level of families given all the other variables available. For this, I would treat this aas a regression problem. For regression, the education level is already a numeric variable. I anticipate that education will be the dependent variable and will select perhaps 4 independent variables. It should be straight forward.