Exploring the BRFSS data

Will Breckwoldt

10/4/2021

Intro to Probability and Data with R Final Project: Analyzing BFSS Data Duke University William Breckwoldt Date: ‘r format(Sys.time(), "%d %B, %Y)’

Project Report

Import Libraries

```
library(devtools)
library(shiny)
install_github("StatsWithR/statsr")
library(statsr)
library(dplyr)
library(ggplot2)
library(dplyr)
```

Load Data

```
load("C:/Users/Scott/Downloads/brfss2013.Rdata")
```

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `brfss2013`. Delete this note when before you submit your work.

Part 1: Data

For this project, we will be using the 2013 Behavioral Risk Factor Surveillance System (BRFSS), an observational study conducted by the Center for Disease Control and Prevention (CDC). The BRFSS is administered and supported by the CDC's Population Health Surveillance Branch, under the Division of Population Health at the National Center for Chronic Disease Prevention and Health Promotion. The observations are collected by telephone surveys that collect health-related risk behaviors, chronic health conditions, and use of preventive services. The data in the BRFSS is only from non-institutionalized adults, aged 18 years or older in all 50 states as well as the District of Columbia and three U.S. territories.The BRFSS is the premier health survey system of the United States and is the largest continously conducted heath survey system in the world. The 2013 BRFSS has 491,775 observations with 330 variables.

However, the BRFSS does not take into account the town or county within a state that the observation is located and the population of states are not well depicted by the amount of survey results(i.e. California has 11,518 observations and Kansas has 23,282 observations). Additionally, the BRFSS relies on self-reported data, therefore information could be easily falsified or withheld. Despite these concerns, there have been many studies that have examined issues related to the reliability and validity of the BRFSS and the system's ability to provide both valid national estimates, within state estimates and comparisons across states that have proven that the BRFSS is accurate.

These studies have demonstrated that the BRFSS has the highest reliability when compared to other national surveys. In this project, we will be looking for correlation among variables, instead of causation. We may be

able to identify an explanatory variable as a variable suspected of affecting the other(s), however, this does not guarantee that the relationship between the two is actually causal, even if there is an association between the variables. Therefore, the project must generalize observations by state or country and the specificity of my research questions will therefore be limited to state boundaries and correlation among variables.

Part 2: Research Questions

Research Question 1

Is there any correlation between race and diabetes in the United States?

It is widely believed that diabetes tends to effect African Americans more than other races in the United States. This question will be interesting to answer because it can either confirm or go against these beliefs.

Research Question 2

Do people who report poor general health smoke cigarettes regularly or have smoked in the past?

This will be very interesting question to answer as we know that cigarettes have ill effects on the human body, but people also tend to quit smoking once they realize they are in poor heath.

Research Question 3

Does the amount of children a mother (or female guardian) effect the amount that she works?

We know that children are dependant on their mothers, but also expensive. It would be interesting to see whether or not the amount a women works increases or decreases with the amount of children they have. Historically, mothers do not work, however, in the past few decades that has changed drastically.

Part 3: Exploratory Data Analysis

Research Question 1 Is there any correlation between race and diabetes in the United States?

First, filter the observations using the rrclass2 variable to make a BRFSS database for African Americans in the United States.

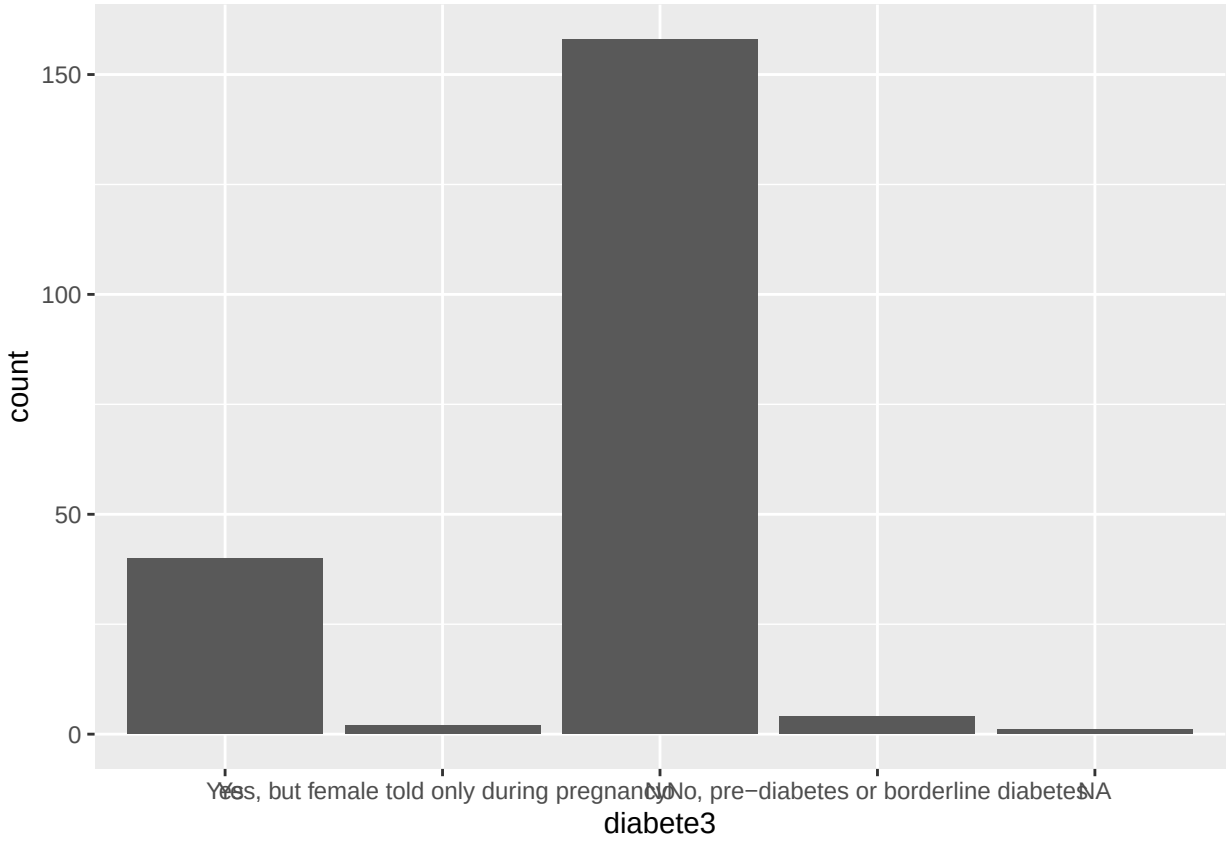
```
AAdata <- brfss2013 %>%
  filter(rrclass2 == "Black or African American")
```

Unfortunately, more than 99% of respondants gave no response for the rrclass2 variable, however, we are still able to gather 205 observations.

Next, let us find the percentages of African Americans who have been told they have diabetes.

```
AAdata %>%
  filter(diabete3=="Yes")
```

```
ggplot(data = AAdata, aes(x = diabete3))+
  geom_histogram(stat="count")
```



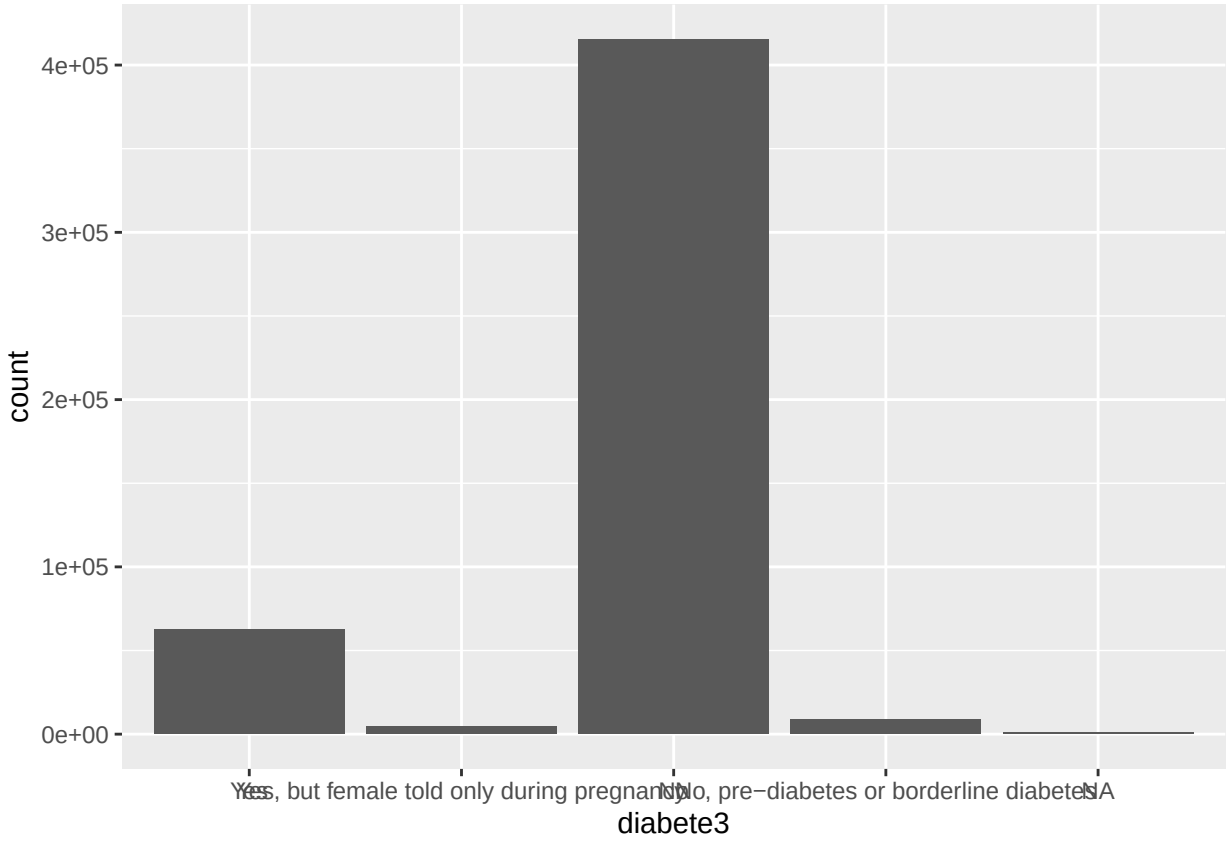
Now, we have the percentage of African Americans in the BRFSS who have or have had diabetes. (40/205= 19.5%)

Let us compare this with the percentage of all Americans who have been told they have diabetes.

```
brfss2013 %>%
  filter(diabete3=="Yes")
```

We see that there are 62,363 observations, therefore, 62,363 out of 491,775 (12.7%) people answered that they have been told they have diabetes.

```
ggplot(data = brfss2013, aes(x = diabete3))+
  geom_histogram(stat="count")
```



By examining these values, one can see the significant difference between the percentage of all Americans compared to African Americans who have been told they have diabetes.

Research Question 2 Do people who report poor general health smoke cigarettes regularly or have smoked in the past?

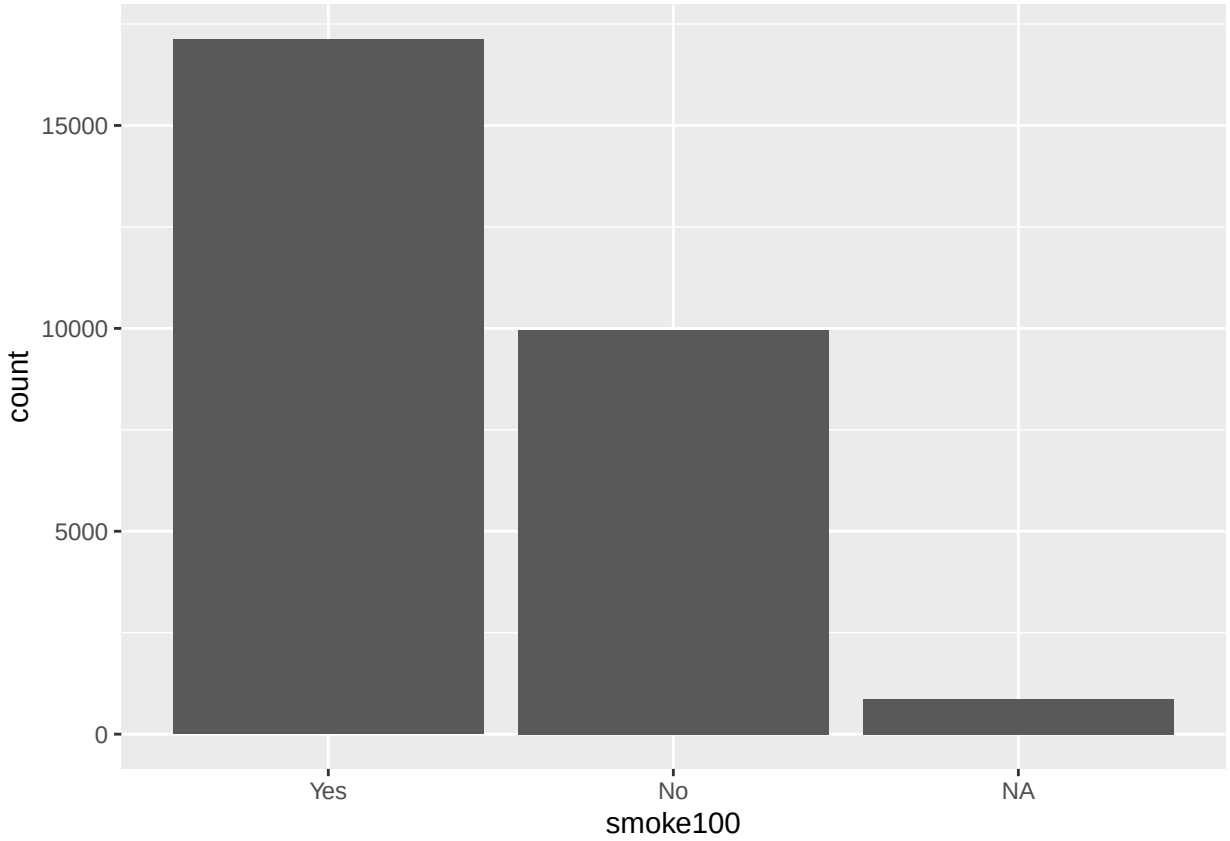
First, we will make a database for Americans who reported that they have poor health using the variable genhlth.

```
ph_data <- brfss2013 %>%
  filter(genhlth=="Poor")
```

This dataset has 27,951 observations. Therefore 27,951 out of 491,775 (5.68%) 2013 BRFSS respondents reported they had poor general health.

Now, with this data set, let us create a plot with number of people on the y-axis and amount of people who have or have not smoked at least 100 cigarettes on the x-axis.

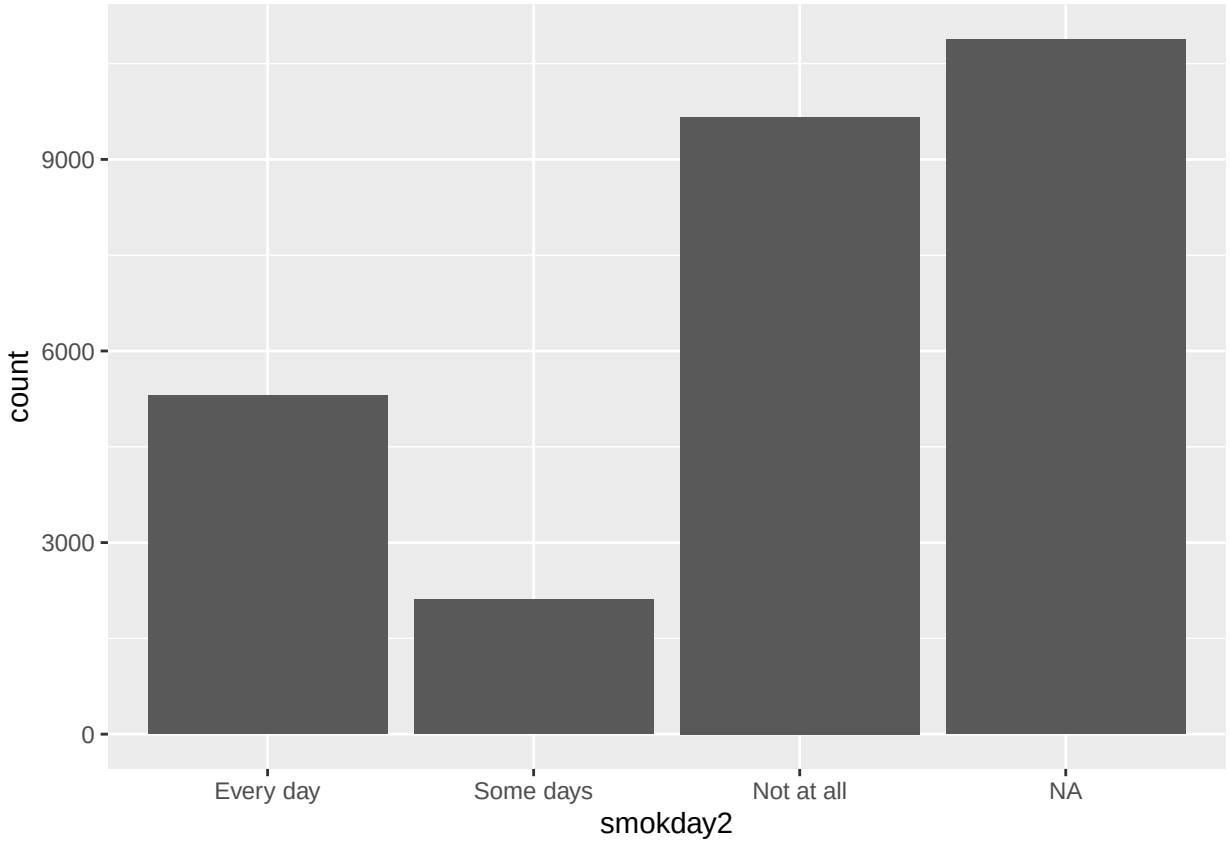
```
ggplot(data = ph_data, aes(x = smoke100))+
  geom_histogram(stat="count")
```



We can see from that plot that the majority of respondents who said they have poor general health have smoked over 100 cigarettes in their lifetime.

Next, we will make a plot to see if the frequency of days now smoking cigarettes has any correlation with poor health.

```
ggplot(data = ph_data, aes(x = smokday2))+
  geom_histogram(stat="count")
```



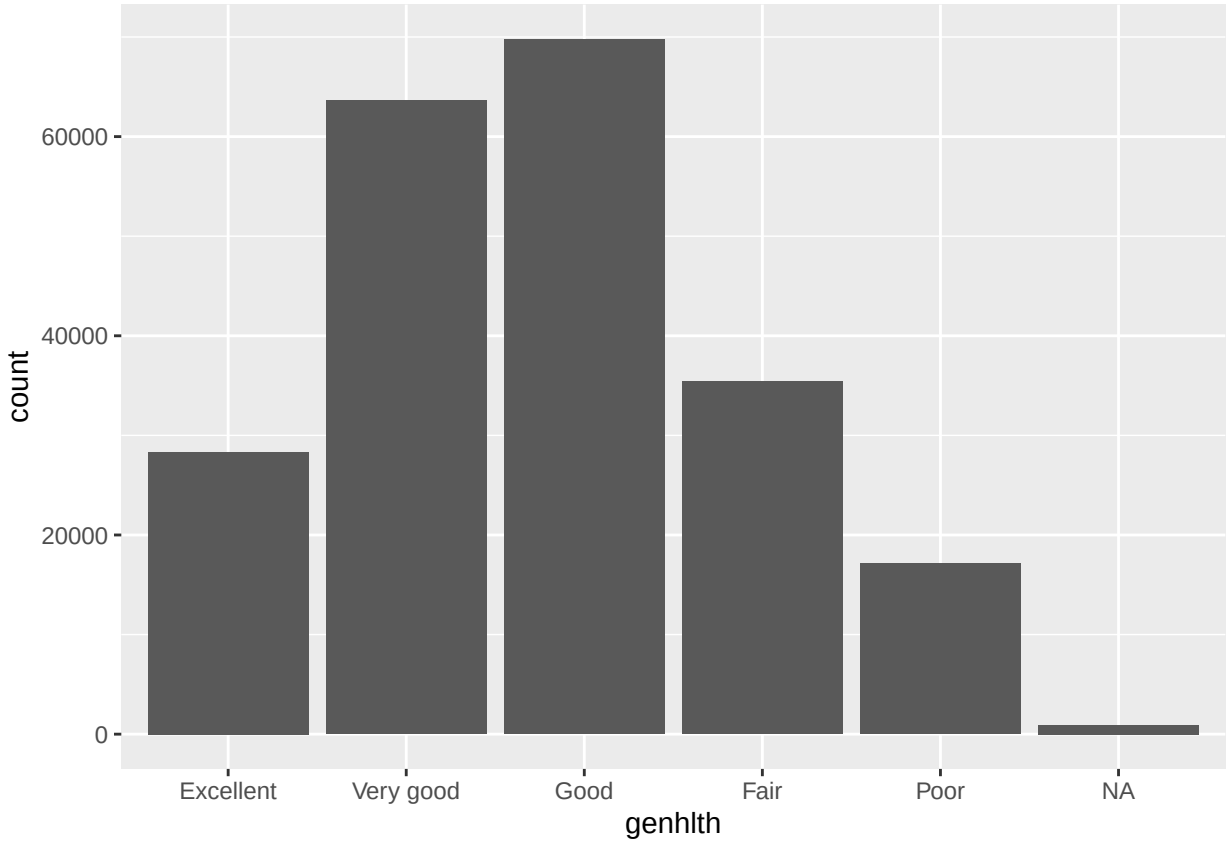
Oddly, it seems that the frequency of cigarettes now smoking has less of a correlation with poor general health. This may be because people stop smoking once they have poor general health or maybe a number of people who smoke regularly refused to respond to this survey question as we can see that the "NA" column has the greatest count. Additionally, it could be that people who no longer smoke, smoked 100 cigarettes in their lifetime and are either improving their health or have not worsened it, whereas those with poor general health that continue to smoke tend to expire quicker.

Nonetheless, if this is accurate, it appears that it would be in one's best interest, with their general health in mind, to not smoke 100 cigarettes in their lifetime as the majority of people who have poor general health have.

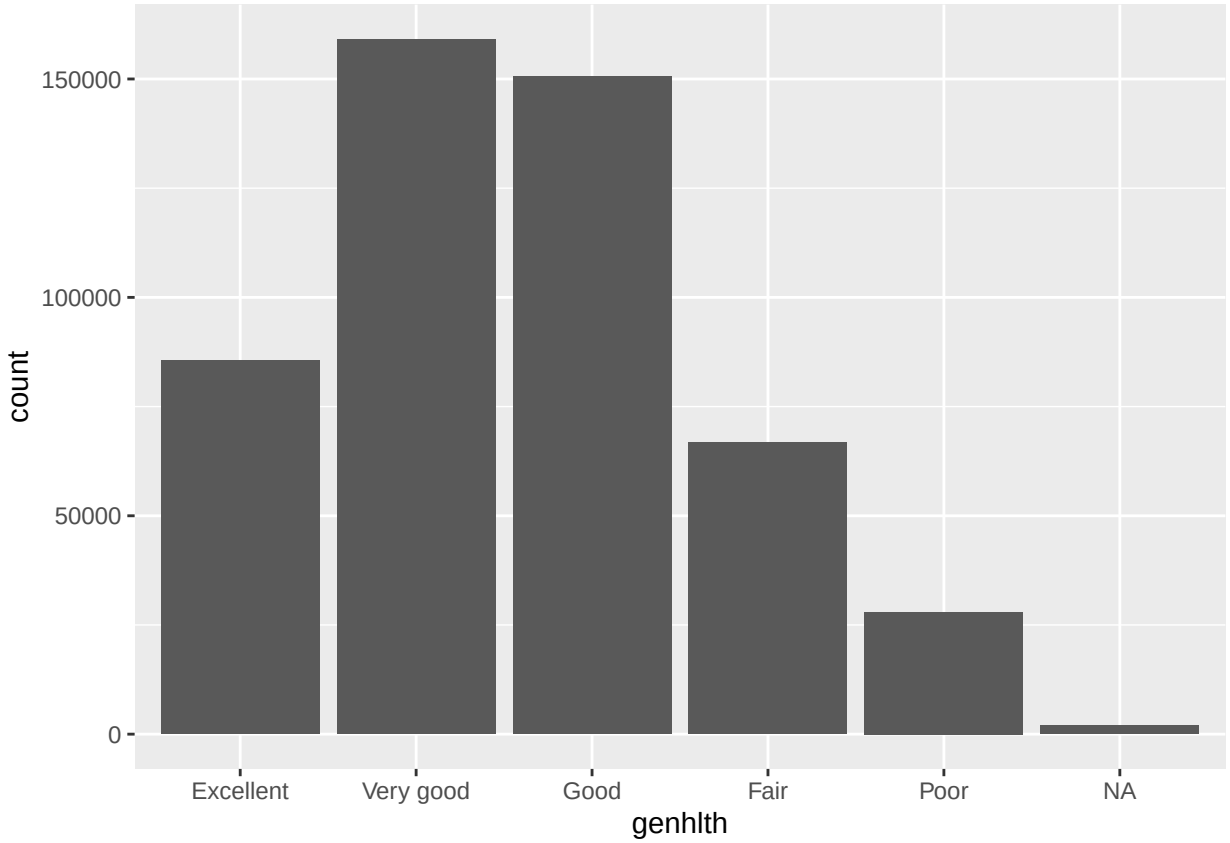
Before we move on to the next research question. Let us create one more histogram that examines the amount of people who have smoked at least 100 cigarettes and their reported general health.

```
cig100_data <- brfss2013 %>%
  filter(smoke100=="Yes")
```

```
ggplot(data = cig100_data, aes(x = genhlth))+
  geom_histogram(stat="count")
```




```
ggplot(data = brfss2013, aes(x = genhlth))+
  geom_histogram(stat="count")
```



By comparing the two plots, it appears there may be a correlation between smoking 100 cigarettes in one's lifetime and their reported general health.

Research Question 3

Does the amount of children a mother (or female guardian) effect the amount that she works?

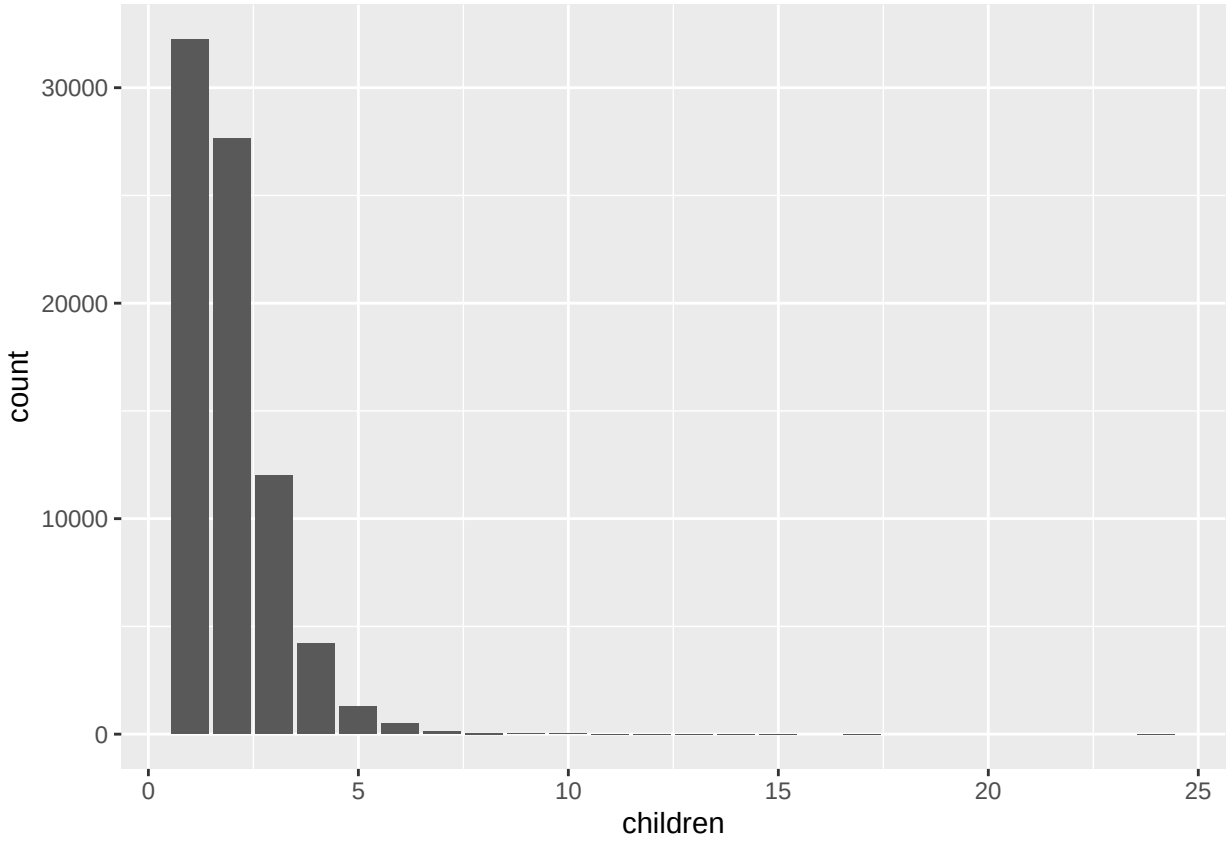
```
fem_data <- brfss2013 %>%
  filter(sex=="Female")
```

We have 290,455 female respondants.

```
mom_data <- fem_data %>%
  filter(children >= 1)
```

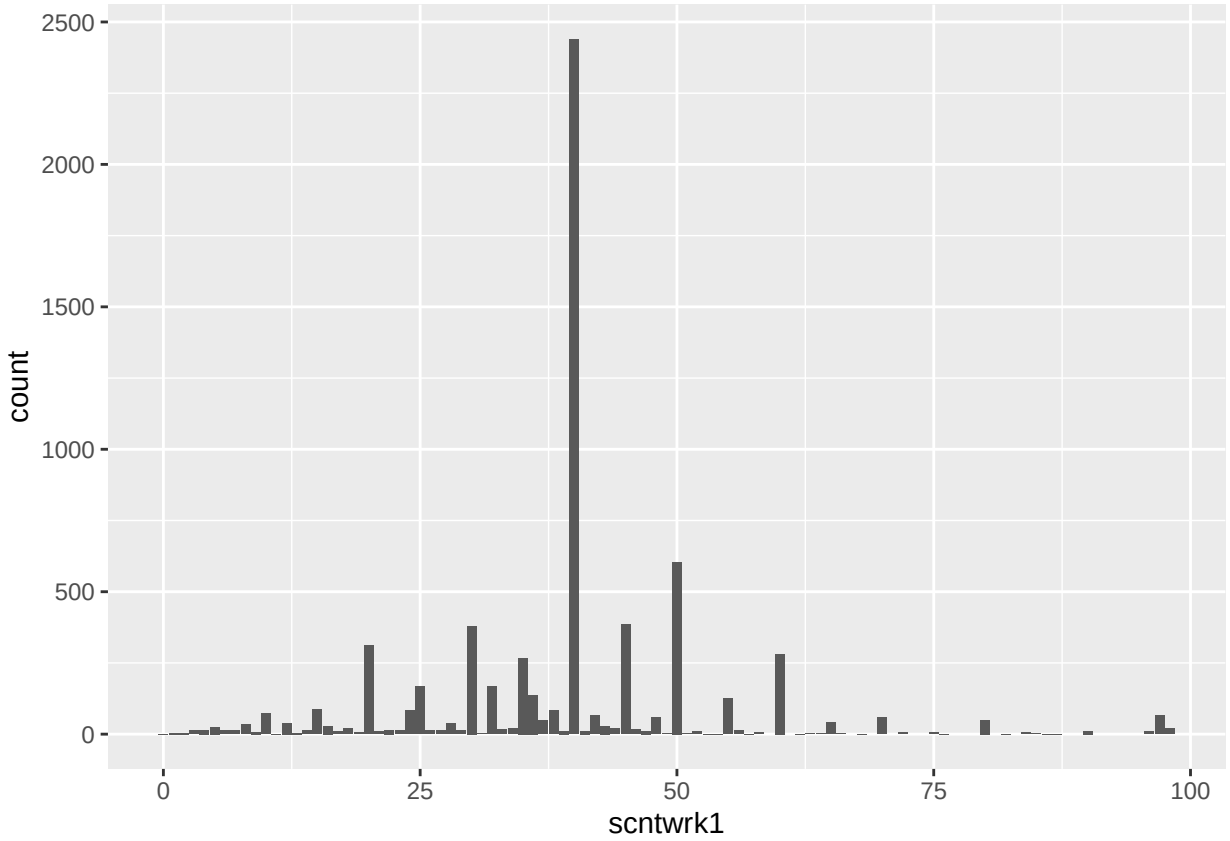
Now, we have 78,134 observations of females with children in their household. Let us make a histogram with this data.

```
ggplot(data = mom_data, aes(x = children))+
  geom_histogram(stat="count")
```



Here we have a histogram demonstrating the counts of the number of children females with children in their household have.

```
ggplot(data=mom_data, aes(x=scntwrk1))+
  geom_histogram(stat="count")
```



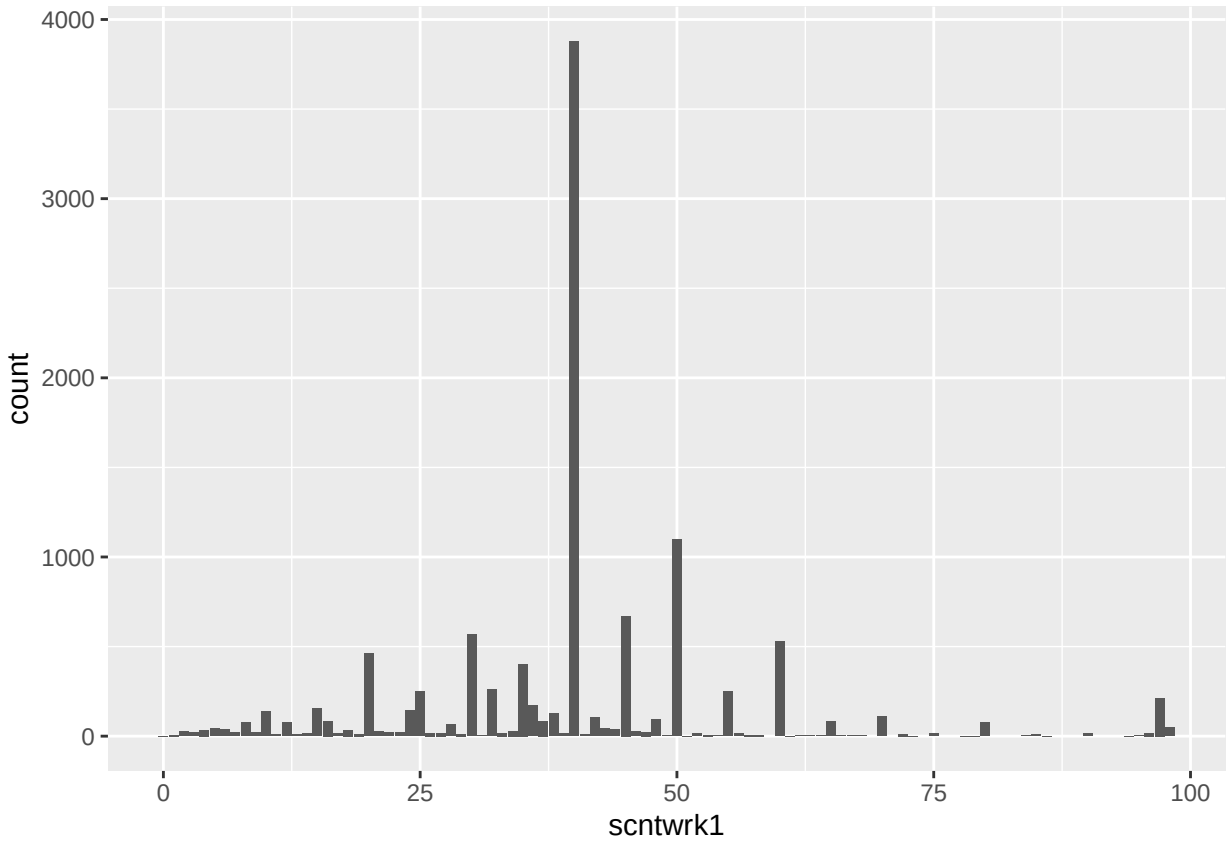
Now we have a histogram demonstrating the amount of hours females with children in their household have.

Now let us compare this to women without children.

```
nomom_data <- fem_data %>%
  filter(children==0)
```

We have 211,044 females without children in the household.

```
ggplot(data=nomom_data, aes(x=scntwrk1))+
  geom_histogram(stat="count")
```



##As we can see from the plots, it appears that women without children in their household work slightly more than those with children in the household, however the difference is not significant.