Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
```

Load data

```
setwd("C:/Users/51937/Documents/CURSOS/Git_R/Projects")
load("brfss2013")
```

Part 1: Data

BRFSS Survey Data and Documentation

The data is obtained through surveys that are carried out by telephone in each participating state, this design is stratified with a random methodology, as the same website mentions: BRFSS divides the telephone numbers into two groups, or strata, that are sampled separately.

Now, knowing how the sample is obtained, based on that, experimental studies can be carried out, and due to the randomness we can determine the generalizability, for the target population, of an existing relationship between variables, for example, with the relationship that we determine between the level of income and the consumption of alcoholic beverages in the sample, we can accept that the relationship continues to be fulfilled in the target population, on the other hand, if we randomly assign individuals to a study, after having carried out the random sampling, we can determine that more that an observed relationship between variables exists causality from one to another.

Part 2: Research questions

Research quesion 1:

What is the number of respondents in the sample who have more than 5 children and have an income of less than $ 10,000? This question is of interest to me because the information obtained makes us think closely about the conditions of life offered by families with lower incomes and who take care of more than two children who require expenses in health, food and education.

Research quesion 2: From the sample, what educational level of the veterans has the most observations? What is the occupation of the veterans at the time of being surveyed?

This question is of interest to me because in my country and in most of Latin America the service military is taken on a voluntary or compulsory basis by young people who do not have economic resources, They neither study nor work and sometimes commit crimes, on the other hand, in Latin America, studies have

shown that after finishing military service, reinsertion into work is difficult for those who carried out the service. With the question posed, we can see which educational level has more observations by veterans as well as the employment situation in which they are.

Research quesion 3: From the sample drawn: Who has greater variability in the hours worked per week, men or women? Are more men who work than women who work?

The importance of this question lies in the gender gaps between men and women in the labor market where they are expressed as the shortest hours allowed to work for women as well as in the salary differences between both genders.

Part 3: Exploratory data analysis

Research quesion 1: To answer this question we first create a variable that contains respondents with incomes less than $ 10,000 and equal to or greater than $ 10,000.

```
brfss2013<-brfss2013 %>% mutate(incless = ifelse(income2=="Less than $10,000","Less than $10,
         000","More than $10,000"))
```

In the same way, with the variable “children” we create another variable that contains the families that have more than 5 children or less and equal to 5 children.

```
brfss2013 <- brfss2013 %>% mutate(child5 = ifelse(children>5, "children>5", "children<=5"))
```

To detail what we have advanced, and answer the research question, we represent the observations in a table in absolute or relative terms, taking into account that the relative terms are from the total sample. So as follows:

```
table(brfss2013$incless,brfss2013$child5)
```

```
##
##                     children<=5 children>5
##   Less than $10,000       25318         84
##   More than $10,000      393261       1037
```

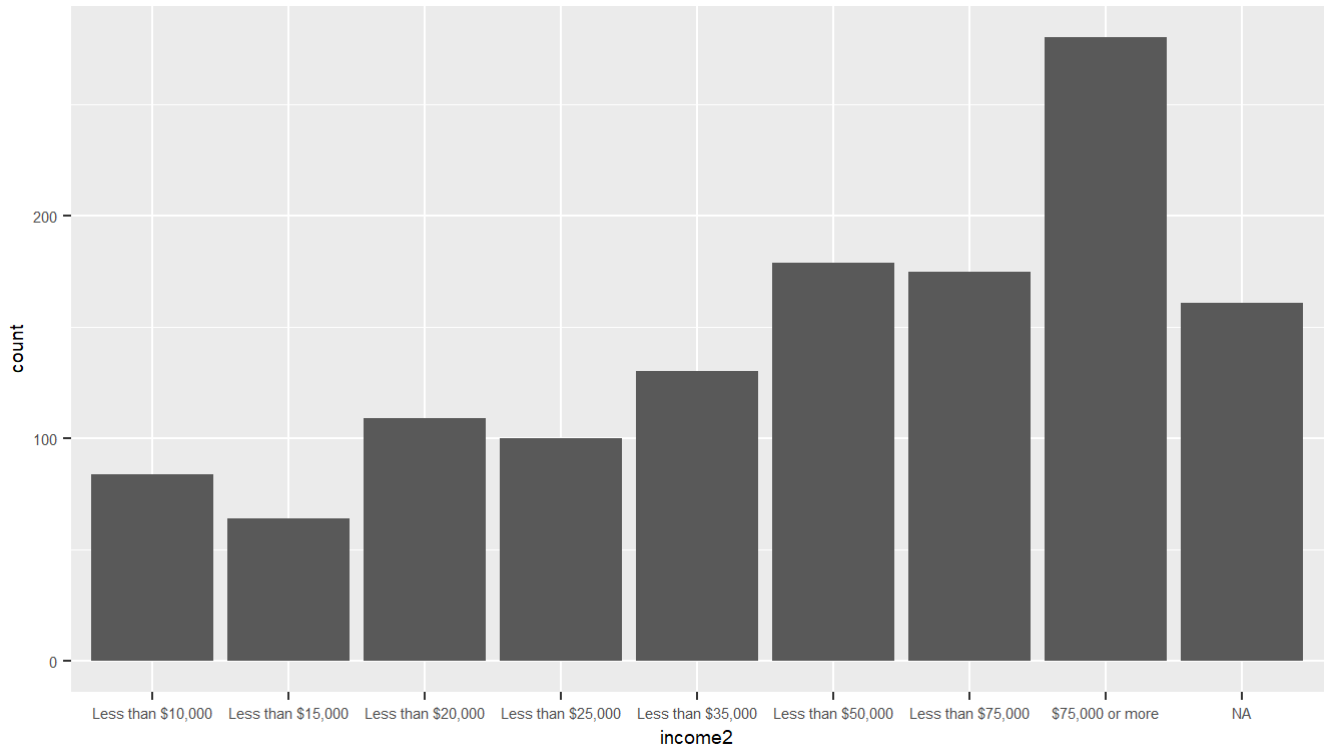
```
prop.table(table(brfss2013$incless,brfss2013$child5))
```

```
##
##                     children<=5  children>5
##   Less than $10,000 0.060324041 0.000200143
##   More than $10,000 0.937005004 0.002470812
```

Executing the line of code we determine that the families that have incomes less than $ 10,000 and have more than 5 children are 84, representing 0.02% of the total sample.

Finally, to graphically represent the response, we create a new database that contains the variables of interest, in this case “child5” and “income2”, but only for the observations “children> 5” of “child5”

```
filtchi<-brfss2013 %>% filter(child5=="children>5") %>% select(child5,income2)
ggplot(data=filtchi,aes(x=income2))+
  geom_bar()+
  theme(text=element_text(size=7))
```



The statement is also verified where a small part of the sample has more than 5 children with income less than $ 10,000 and it can be observed that most of the sample that has more than 5 children has income greater than $ 75,000.

Research quesion 2: To answer the first part of the question, we group the sample according to the variables “veteran3” and “educate”

```
brfss2013 %>% group_by(veteran3,educa) %>% summarise(count=n())
```

```
## `summarise()` has grouped output by 'veteran3'. You can override using the `.groups` argum
ent.
```

```
## # A tibble: 21 x 3
## # Groups:   veteran3 [3]
##    veteran3 educa                                                  count
##    <fct>    <fct>                                                  <int>
##  1 Yes      Never attended school or only kindergarten                31
##  2 Yes      Grades 1 through 8 (Elementary)                          929
##  3 Yes      Grades 9 though 11 (Some high school)                   2090
##  4 Yes      Grade 12 or GED (High school graduate)                 17403
##  5 Yes      College 1 year to 3 years (Some college or technical school) 18906
##  6 Yes      College 4 years or more (College graduate)             21908
##  7 Yes      <NA>                                                     179
##  8 No       Never attended school or only kindergarten               643
##  9 No       Grades 1 through 8 (Elementary)                        12448
## 10 No       Grades 9 though 11 (Some high school)                  26030
## # ... with 11 more rows
```

If we want to be more specific and only focus on veterans we use the “filter” command as follows:

```
brfss2013 %>% group_by(veteran3, educa) %>% filter(veteran3=="Yes") %>% summarise(count=n())
```

```
## `summarise()` has grouped output by 'veteran3'. You can override using the `.groups` argum
ent.
```

```
## # A tibble: 7 x 3
## # Groups:   veteran3 [1]
##   veteran3 educa                                                  count
##   <fct>    <fct>                                                  <int>
## 1 Yes      Never attended school or only kindergarten                31
## 2 Yes      Grades 1 through 8 (Elementary)                          929
## 3 Yes      Grades 9 though 11 (Some high school)                   2090
## 4 Yes      Grade 12 or GED (High school graduate)                 17403
## 5 Yes      College 1 year to 3 years (Some college or technical school) 18906
## 6 Yes      College 4 years or more (College graduate)             21908
## 7 Yes      <NA>                                                     179
```

We observed that 21,908 veterans have been in college 4 years or more or are college graduates.

In the same way, to observe the employment situation of veterans, we group the sample according to the variables “veteran3” and “employ1”

```
brfss2013 %>% group_by(veteran3,employ1) %>% summarise(count=n())
```

```
## `summarise()` has grouped output by 'veteran3'. You can override using the `.groups` argum
ent.
```

```
## # A tibble: 27 x 3
## # Groups:   veteran3 [3]
##    veteran3 employ1                           count
##    <fct>    <fct>                             <int>
##  1 Yes      Employed for wages                18559
##  2 Yes      Self-employed                      4417
##  3 Yes      Out of work for 1 year or more     1209
##  4 Yes      Out of work for less than 1 year   1025
##  5 Yes      A homemaker                         430
##  6 Yes      A student                           611
##  7 Yes      Retired                           31025
##  8 Yes      Unable to work                     3850
##  9 Yes      <NA>                                320
## 10 No       Employed for wages                183495
## # ... with 17 more rows
```

And filtering only for those who are veterans:

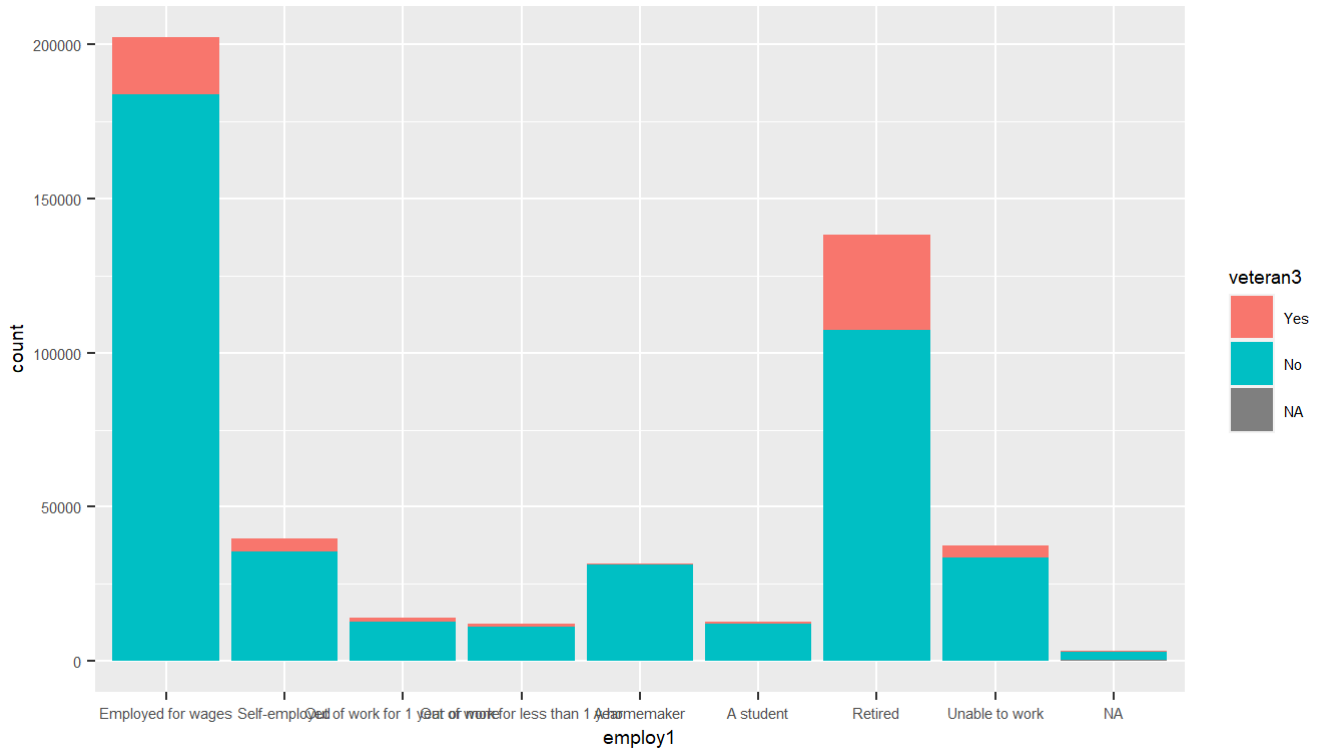
```
brfss2013 %>% group_by(veteran3, employ1) %>% filter(veteran3=="Yes") %>% summarise(count=n
         ())
```

```
## `summarise()` has grouped output by 'veteran3'. You can override using the `.groups` argum
ent.
```

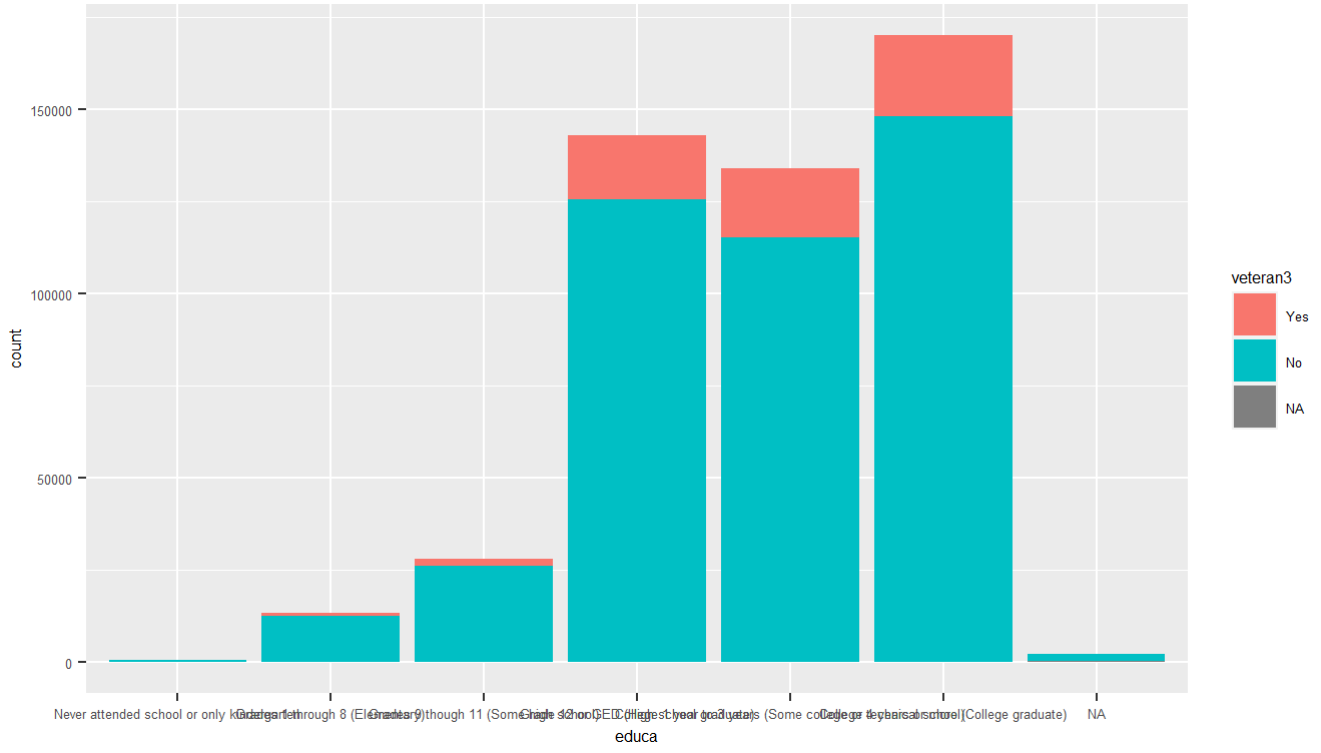
```
## # A tibble: 9 x 3
## # Groups:   veteran3 [1]
##   veteran3 employ1                           count
##   <fct>    <fct>                             <int>
## 1 Yes      Employed for wages                18559
## 2 Yes      Self-employed                      4417
## 3 Yes      Out of work for 1 year or more     1209
## 4 Yes      Out of work for less than 1 year   1025
## 5 Yes      A homemaker                         430
## 6 Yes      A student                           611
## 7 Yes      Retired                           31025
## 8 Yes      Unable to work                     3850
## 9 Yes      <NA>                                320
```

Graphic representation:

```
ggplot(data=brfss2013, aes(x=employ1, fill=veteran3))+
  geom_bar()+
  theme(text=element_text(size=7))
```




```
ggplot(data=brfss2013, aes(x=educa, fill=veteran3))+
  geom_bar()+
  theme(text=element_text(size=6))
```



We can see that a large part of the veterans in the sample are “Retired”, this category occupies the first place with 31,025 observations, in second place it is found that the labor situation “Employed for wages” with 18,559 observations. On the other hand, it is also observed that the category that is repeated the most in the educational level of veterans is “College 4 years or more (College graduate)”.

From what has been observed, with this first approximation, we could extract some differences with respect to Latin America, however, we must take into account the generalization of what is applied.

Research quesion 3:

To answer this, the first part of the question, we first group the database with the group_by command, then we filter the observations that are not “na” with the filter command, finally we make a summary of statistics which are the IQR and the standard deviation of the variable of interest, in this case “scntwrk1”, which tells us how many hours a week the respondent works.

```
brfss2013 %>% group_by(sex) %>%  filter(!is.na(sex),!is.na(scntwrk1)) %>%
  summarise(iqr=IQR(scntwrk1), sdesv = sd(scntwrk1))
```

```
## # A tibble: 2 x 3
##   sex      iqr sdesv
##   <fct>  <dbl> <dbl>
## 1 Male      10  15.5
## 2 Female    11  15.6
```

For the second part of the question, where we want to determine if there are more men who work compared to women who work, the variable “employ1” will be used, first we group the sample by the variables “sex” and “employ1”, then we filter the observations “na” and finally we count the observations with the command “summarize”

```
brfss2013 %>% group_by(sex, employ1) %>% filter(!is.na(sex), !is.na(employ1)) %>%
  summarise(count=n())
```

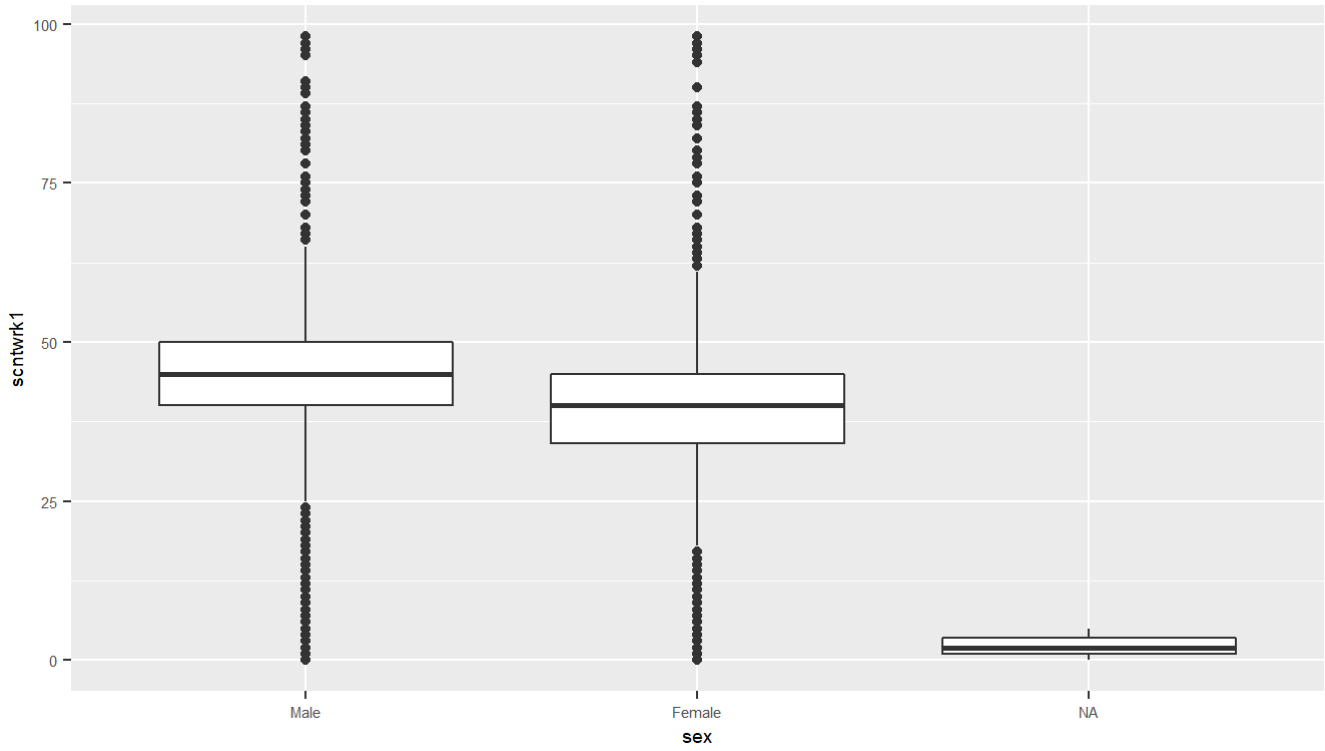
```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups` argument.
```

```
## # A tibble: 16 x 3
## # Groups:   sex [2]
##    sex    employ1                          count
##    <fct>  <fct>                            <int>
##  1 Male   Employed for wages                91055
##  2 Male   Self-employed                    23081
##  3 Male   Out of work for 1 year or more     5830
##  4 Male   Out of work for less than 1 year   5709
##  5 Male   A homemaker                         610
##  6 Male   A student                          5382
##  7 Male   Retired                           54893
##  8 Male   Unable to work                    13367
##  9 Female Employed for wages               111145
## 10 Female Self-employed                    16751
## 11 Female Out of work for 1 year or more     8243
## 12 Female Out of work for less than 1 year   6532
## 13 Female A homemaker                      31036
## 14 Female A student                         7300
## 15 Female Retired                          83366
## 16 Female Unable to work                   24086
```

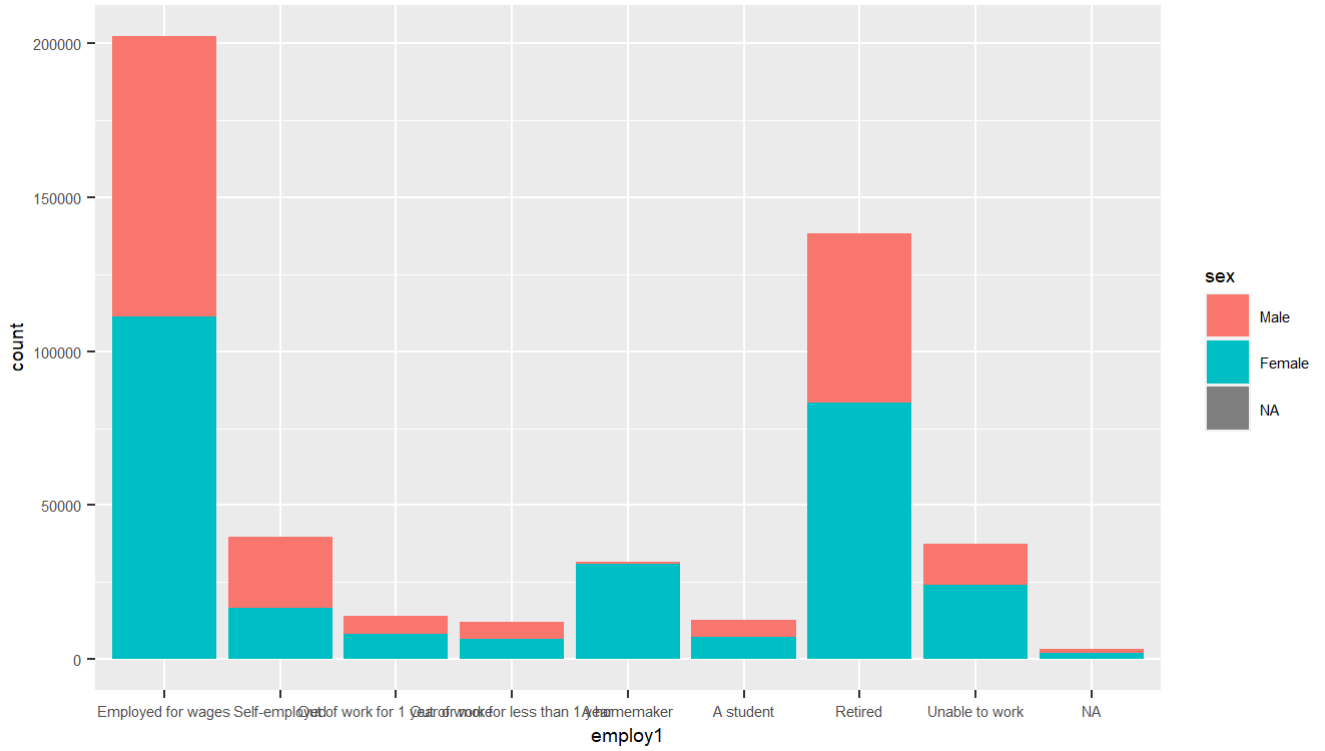
Graphic representation:

```
ggplot(data= brfss2013, aes(x=sex, y = scntwrk1))+
  geom_boxplot()+
  theme(text=element_text(size=7))
```

```
## Warning: Removed 459413 rows containing non-finite values (stat_boxplot).
```




```
ggplot(data=brfss2013, aes(x=employ1, fill=sex))+
  geom_bar()+
  theme(text=element_text(size=7))
```



We can see that from the sample extracted the IQR is 10 for men and 11 for women, this means that 50% of the data is in a greater range and has a greater dispersion as seen in the boxplot, we can also observe that the standard deviation of both groups differs by 0.1, which means that the data have some similarity

in their variability. On the other hand, with what was observed for the second question, we verify that for the sample women have more employment for wages than men and are also retired to a greater degree, these observations are against our initial belief and encourage more research.