Exploring the BRFSS data

Tiago de Almeida Silva

Setup

Load packages

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

Load data

```
load("brfss2013.Rdata")
```

Part 1: Data

The information in the sample were collected though a system of health-related telephone surveys that gathered state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.The survey was carried out in all 50 states in US as well as the District of Columbia and three U.S. territories. Since 2011, BRFSS conducts both landline telephone- and cellular telephone-based surveys. In conducting the BRFSS landline telephone survey, interviewers collect data from a randomly selected adult in a household. In conducting the cellular telephone version of the BRFSS questionnaire, interviewers collect data from an adult who participates by using a cellular telephone and resides in a private residence or college housing.

The data were collected through a random sample of adults in US and the implication of this data collection is to gather uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population.

Part 2: Research questions

Research quesion 1:

It is well known that people with Arthritis feel lots of pain and in many cases it can turn their lives into a nightmare because they cannot do basic things such as walking, bathing, going to a market and so on. For this reason I will check the impact of this disease on people's lives.

Research quesion 2:

Diabetes is one of the main responsible for blindness caused by a disease among the adult population around the world and thinking about that I decided to compare and analyse if there is any relation between the illness and blindness in this sample from USA. Moreover, I will check using a variable from the dataset if people who have diabetes and are blind were told about this relation.

Research quesion 3:

In the last question I will analyse if there is any relation among sleep disorders, depression and hypertension because according to (Batal et al., 2011) these disorders are a risk factor for both high blood pressure and depression.

Part 3: Exploratory data analysis

Research quesion 1:

Firstly I organized the data selecting the variables I wanted to work with (havarth3 ,diffwalk, diffdres and diffalon) in the object named "artrite" and after that I turned the factor data (yes/no) into numeric ones in order to get some useful insights for my plots.

Just to clarify the variables mentioned above:

- havarth3: People Told Have Arthritis;
- diffwalk: Difficulty Walking Or Climbing Stairs;
- diffdres: Difficulty Dressing Or Bathing;
- diffalon: Difficulty Doing Errands Alone.

```
artrite <- brfss2013 %>%
   select(havarth3 ,diffwalk, diffdres, diffalon) %>%
  mutate(diffwalk_bin = recode(diffwalk,
                               "Yes" = 1,
                               "No" = 0),
         diffdress_bin = recode(diffdres,
                               "Yes" = 1,
                               "No" = 0),
         diffalon_bin = recode(diffalon,
                               "Yes" = 1,
                               "No" = 0)) %>%
  select(everything(), -(2:4))
```

Secondly I separated the object, that contains 491.775 parameters, into two where the first one is related to people without arthritis and the another one concerns people with. I did that because I wanted to show the huge difference that exists between these two groups when it comes to problems caused by this disease. Once I got the data I was looking for I created two different plots regarding people with and without arthritis.

First of all, I created a new object named "noarth" to separate people without the illness.

```
noarth <- artrite %>%
  filter(havarth3 == "No")

noarth <- noarth %>% mutate(total = rowSums(noarth[ ,2:4], na.rm = TRUE))
```

I found there are 323.653 people out of 491.775 in this condition (no arthritis) which represents 65.81% of total.

```r
paste(nrow(artrite), " Total Sample", sep = "")
```

```
## [1] "491775 Total Sample"
```

```r
paste(nrow(noarth), " People without Arthritis ", sep = "")
```

```
## [1] "323653 People without Arthritis "
```

```r
paste("Percentage ", round(nrow(noarth)*100/nrow(artrite), digits = 2), "%", sep= "")
```

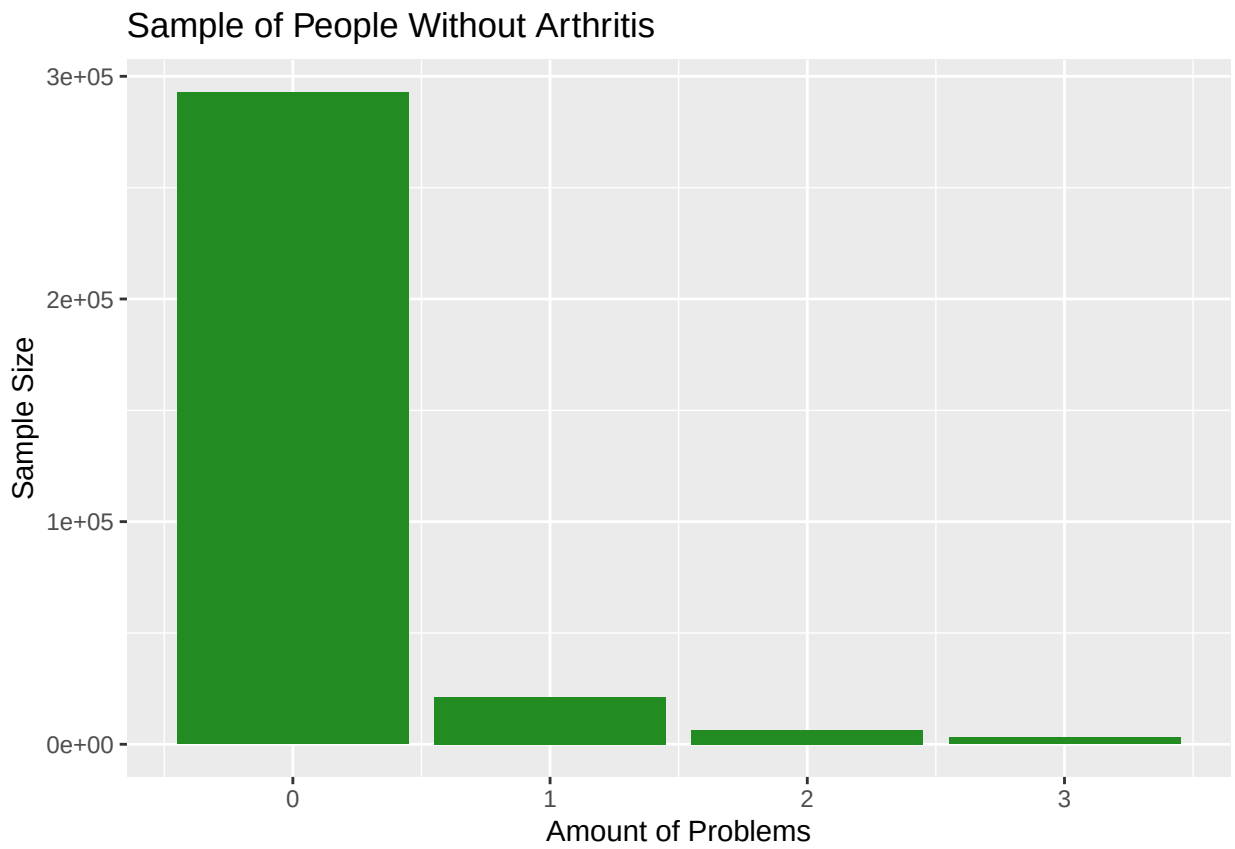
```
## [1] "Percentage 65.81%"
```

And I made a plot using ggplot showing the amount of problems people without arthritis said they have. Note that these 3 problems are the ones I mentioned on the description of this first case study:

- Difficulty Walking Or Climbing Stairs;
- Difficulty Dressing Or Bathing;
- Difficulty Doing Errands Alone.

If a person said he or she had some issue to walk or wear a t-shirt for example, they got 1 in the respective variable and 0 in case of a negative feedback. In other words, I changed the "yes/no" answers to 1/0 where 1 stands for "yes" and 0 for "no" due to it is easier to work and manipulate data like that.

```r
ggplot(noarth) +
  geom_bar(aes(x = total), fill = "forest green") +
  labs(x = "Amount of Problems",
       y = "Sample Size",
       title = "Sample of People Without Arthritis")
```

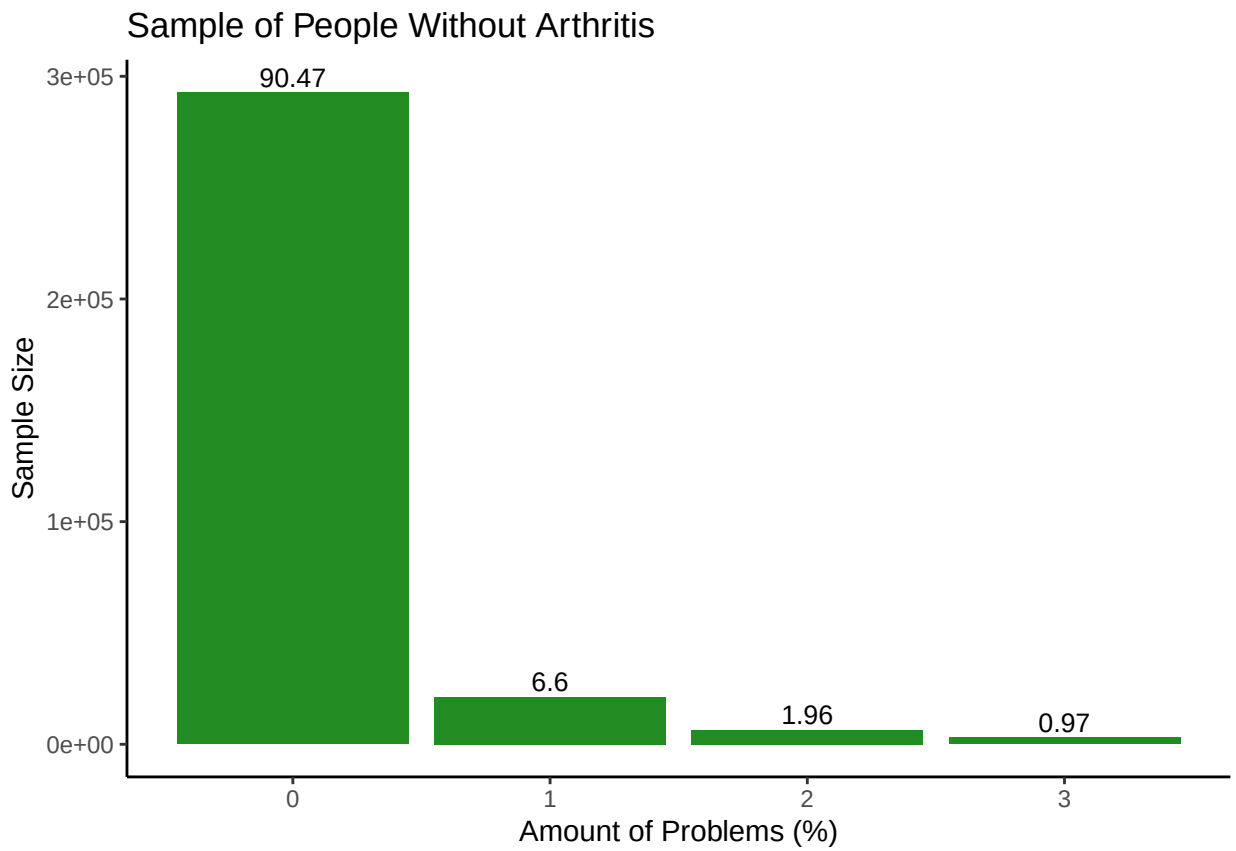


As we can see, the majority of people without the disease have no problems (0) that are commonly related to it. The result was very satisfactory because we were expecting for numbers like that. Below we have a summary of the finds:

```
noarth_summary <- noarth %>% count(total)

noarth_summary <- noarth_summary %>%
  mutate(percent = round(noarth_summary$n*100/nrow(noarth), digits = 2)) %>%
         rename("n_of_problems" = total,
                "total" = n)

print(noarth_summary)
```

```
##   n_of_problems  total percent
## 1             0 292805   90.47
## 2             1  21354    6.60
## 3             2   6352    1.96
## 4             3   3142    0.97
```

```
ggplot(noarth_summary) +
  geom_col(aes(x = n_of_problems, y = total), fill = " forest green") +
  geom_text(aes(x = n_of_problems, y = total, label = percent), vjust = -0.3, size = 3.5, ) +
  labs(x = "Amount of Problems (%)",
       y = "Sample Size",
       title = "Sample of People Without Arthritis") +
  theme_classic()
```



- 90.47% of the people without arthritis have no pain or any difficulty to accomplish daily basic functions;
- Less than 1% said they were facing all the 3 problems, which is a very low number.

Now I got the data for people with arthritis. Let's see the outcomes...

```
arth <- artrite %>%
  filter(havarth3 == "Yes")

arth <- arth %>% mutate(total = rowSums(arth[ ,2:4], na.rm = TRUE))
```

As expected due to the previous outcomes, there are 165.152 people out of 491.775 with arthritis which represent 33.58% of total.

```
paste(nrow(artrite), " Total Sample", sep = "")
```

```
## [1] "491775 Total Sample"
```

```
paste(nrow(arth), " People with Arthritis", sep = "")
```

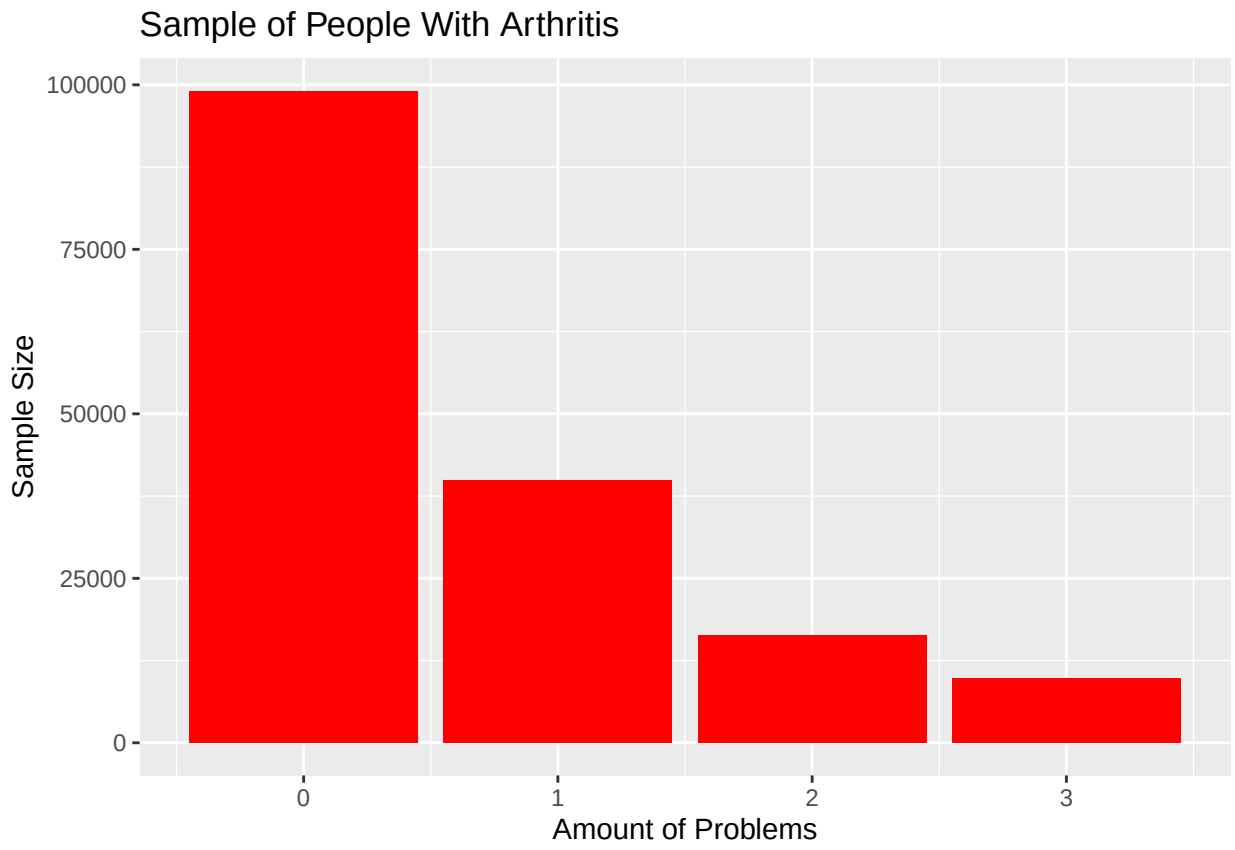
```
## [1] "165152 People with Arthritis"
```

```
paste("Percentage ", round(nrow(arth)*100/nrow(artrite), digits = 2), "%",sep = "")
```

```
## [1] "Percentage 33.58%"
```

The chart shows a completely different pattern in comparison to the first one. In this one we can see there is a huge amount of people with daily problems that are commonly related to arthritis.

```
ggplot (arth) +
  geom_bar(aes(x = total), fill = "red") +
  labs( x = "Amount of Problems",
        y = "Sample Size",
        title = "Sample of People With Arthritis")
```

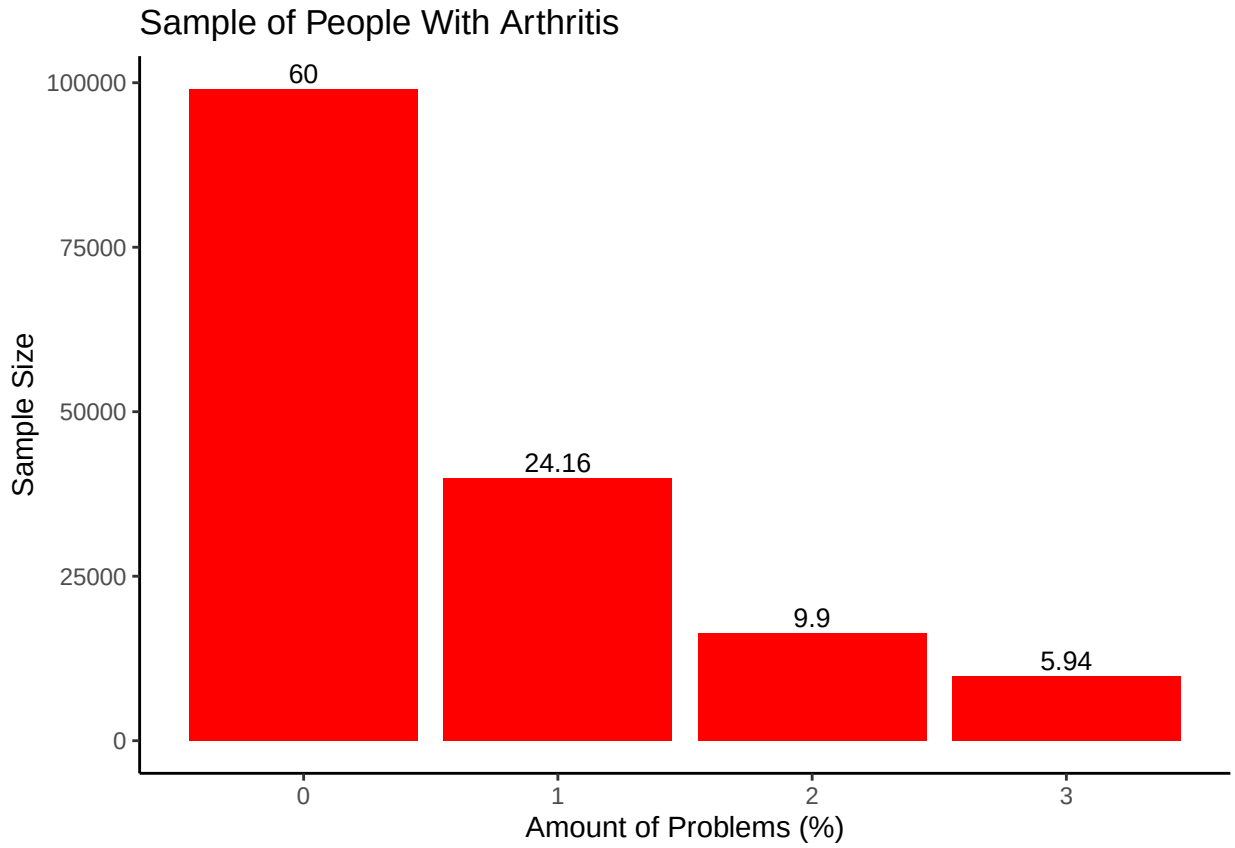


This outcome was not surprising because people with the illness usually have some difficulty to do some daily activities, even the basic ones like walking or wearing a shirt or a pant. You can see below a table showing all the results, including the respective percentage:

```
arth_summary <- arth %>% count(total)

arth_summary <- arth_summary %>%
  mutate(percent = round(arth_summary$n*100/nrow(arth), digits = 2)) %>%
   rename("num_of_problems" = total,
                "total" = n)

print(arth_summary)
```

```
##   num_of_problems total percent
## 1               0 99085   60.00
## 2               1 39908   24.16
## 3               2 16355    9.90
## 4               3  9804    5.94
```

```
ggplot(arth_summary) +
  geom_col(aes(x = num_of_problems, y = total), fill = "red") +
  geom_text(aes(x= num_of_problems, y = total, label = percent), vjust = -0.3, size = 3.5) +
  labs( x = "Amount of Problems (%)",
        y = "Sample Size",
        title = "Sample of People With Arthritis") +
  theme_classic()
```



The amount of people who find some activities very difficult due to joint pain increased significally as expected because people on this sample have arthritis. We saw in the first graph that more than 90% of people without the disease had no issues with pain, while in this one the percentage drops to 60%. Although we can see there is a strong relation between arthritis and the problems mentioned in this case study, we cannot say there is a causation due to it is necessary to analyse every single case to make sure these problems are not caused by other factors.

Research quesion 2:

Firstly I organized the data again selecting the variables I wanted to work with (diabete3, blind, diabeye) in the object named "diabetes" and after that I turned the factor data (yes/no) into numeric ones in order to get some useful insights for my plots.

Just to clarify the variables mentioned above:

- diabete3: (Ever Told) You Have Diabetes;
- blind: Blind Or Difficulty Seeing;
- diabeye: Ever Told Diabetes Has Affected Eyes.

```
diabetes <- brfss2013 %>%
  select(diabete3, blind, diabeye) %>%
  mutate(diabete3 = recode(diabete3,
                           "No, pre-diabetes or borderline diabetes" = "No",
                           "Yes, but female told only during pregnancy" = "Yes"),
         blind_bin = recode(blind,
                            "Yes" = 1,
                            "No" = 0),
         dbeye_bin = recode(diabeye,
                            "Yes" = 1,
                            "No" = 0)) %>%
  select(everything(), -(2:3))
```

At this time I separated and selected only the data about people with diabetes because the rest of the it did not present relevant information for what I am interested in.

```
diab <- diabetes %>%
  filter(diabete3 == "Yes")

diab <- diab %>%
  mutate(total = rowSums(diab[ ,2:3], na.rm = TRUE))
```

We can see below only 66.965 people out of 491.775 said they have diabetes and it represents 13.62% of the total sample.

```
paste(nrow(diabetes), " Total Sample", sep = "")
```

```
## [1] "491775 Total Sample"
```

```
paste(nrow(diab), " Total of people with diabetes", sep = "")
```

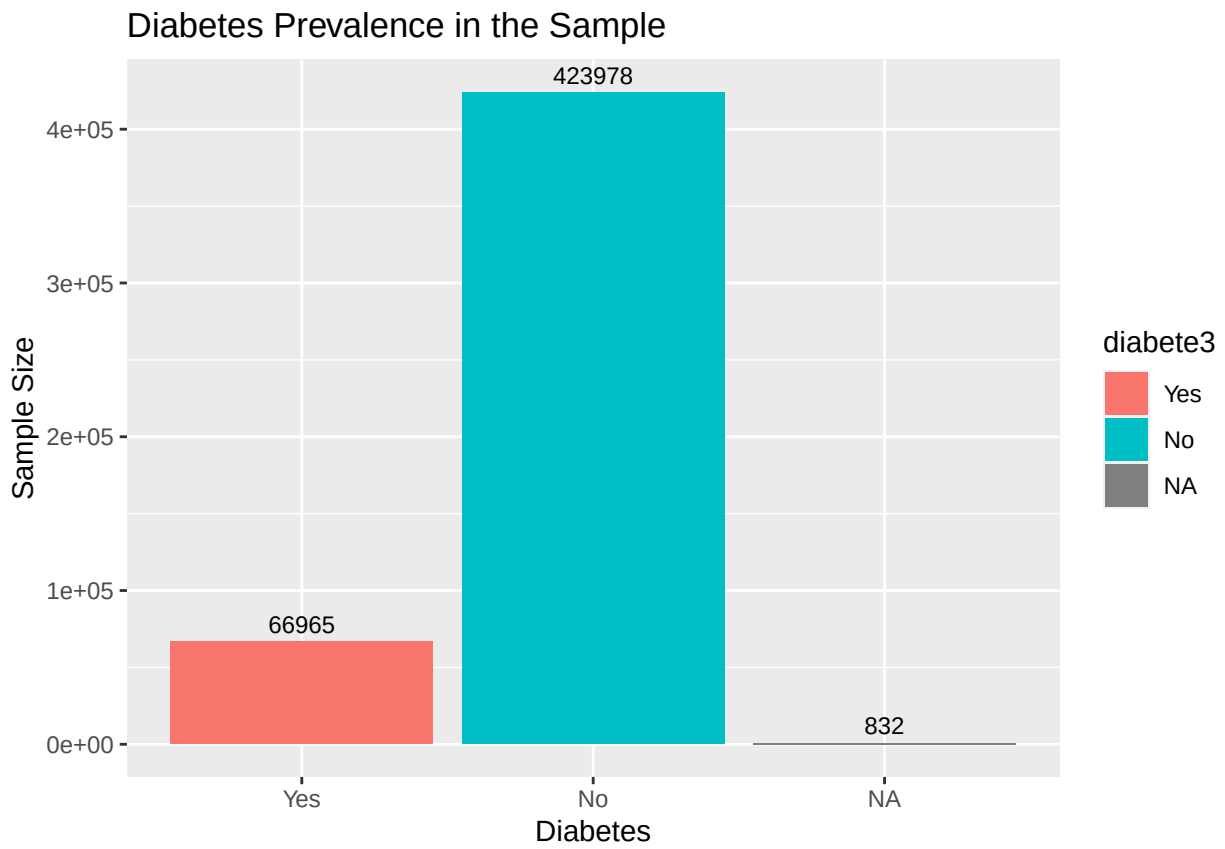
```
## [1] "66965 Total of people with diabetes"
```

```
paste("Percentage ", round(nrow(diab)*100/nrow(diabetes), digits = 2), "%", sep = "")
```

```
## [1] "Percentage 13.62%"
```

Visualization of the previous finds:

```
diabetes_teste <- diabetes %>%
  count(diabete3)

ggplot(diabetes_teste) +
  geom_col(aes(x = diabete3, y = n, fill = diabete3)) +
  geom_text(aes(x = diabete3, y = n, label = n), vjust = -0.5, size = 3.3) +
  labs(title = "Diabetes Prevalence in the Sample",
       x = "Diabetes",
       y = "Sample Size")
```



I used a different approach with this data because I want to know how many people with diabetes are blind and how many of them were told this illness could cause it. I created a new object named "diabetes_summary" to get a better overview of my dataset.

```
diabetes_summary <- diab %>% count(blind_bin, dbeye_bin) %>%
  filter(blind_bin == 1 | dbeye_bin == 1) %>%
  rename("blindness" = blind_bin,
         "awareness" = dbeye_bin,
         "total" = n)


print(diabetes_summary)
```

```
##   blindness awareness total
## 1         0         1  4990
## 2         1         0  2358
## 3         1         1  1907
## 4         1        NA  3099
## 5        NA         1   150
```

Looking at the table above it is possible to get some useful data which suit some parameters and the results are interesting! Fo example, only 10.52% of people who have diabetes were aware this disease is one of the main causes of blindness among the adult population over the world.

```r
per_awar <- diab %>% filter(diab$dbeye_bin == 1) %>%
  count(dbeye_bin)

per_awar <- per_awar %>%
  mutate(per_awr = round(per_awar$n*100/nrow(diab), digits = 2)) %>%
         rename("awareness" = dbeye_bin,
                "total" = n,
                "percentage" = per_awr)


print(per_awar)
```

```
##   awareness total percentage
## 1         1  7047      10.52
```

A quick summary to explain better the other results I got:

- There are 7.364 blind people out of 66.965 with diabetes (11%);
- 5.457 blind individuals in 7.364 were not aware of the relation between diabetes and blindness (74.10%);
- 2.85% of the total of people with diabetes are blind and were told about the relation.

```r
blind_awar <- diabetes_summary %>%
  filter(blindness == 1 & awareness ==1) %>%
  count(total) %>%
  mutate(blind_awar_per = round(1907*100/7364, digits = 2))


print(blind_awar)
```

```
##   total n blind_awar_per
## 1  1907 1           25.9
```

Now I finally have the result I was looking for this second question:

- Only 1.907 blind people with diabetes (25.90%) knew about the relation and that's a very low number and it suggests it should be created an awareness campaign to let people know about the risks of getting blind because of diabetes.

Perhaps if people had more knowledge about that, many of them could have avoided losing their vision. But again, we cannot confirm any causation here as well.

Research quesion 3:

Firstly I selected and organized the variables I wanted to work with. In this question I picked the following ones:

- sleptim1: How Much Time Do You Sleep;
- addepev2: Ever Told You Had A Depressive Disorder;
- bphigh4: Ever Told Blood Pressure High

```r
insomnia1 <- brfss2013 %>%
  select(sleptim1, bphigh4, addepev2)
```

As we can see in the dataset above the "sleptim1" variable only gives us how may hours an individual says he or she sleeps and for this reason I need to clear this data because I want to work with information regarding people with sleep disorders. People who sleep less than 6 hours a day are considered to have some sleep disorder (Pereira, 2021) and I will use this fact to filter my data.

```r
insomnia <- brfss2013 %>%
  filter(sleptim1 <= 6) %>%
  select(sleptim1, bphigh4, addepev2) %>%
  rename("sleep_hours" = sleptim1,
         "hypertension" = bphigh4,
         "depression" = addepev2)
```

Amount of people according to their sleep hours:

```r
insomnia_summary2 <- insomnia %>%
  mutate(hypertension = recode(hypertension,
                               "Yes, but female told only during pregnancy" = "Yes",
                               "Told borderline or pre-hypertensive" = "Yes")) %>%
  count(sleep_hours)

print(insomnia_summary2)
```

```
##   sleep_hours      n
## 1           0      1
## 2           1    228
## 3           2   1076
## 4           3   3496
## 5           4  14261
## 6           5  33436
## 7           6 106197
```

A quick overview of my dataset:

```r
paste(nrow(insomnia1), " Total Sample", sep = "")
```

```
## [1] "491775 Total Sample"
```

```r
paste(nrow(insomnia), " People with a Potential Sleep Disorder", sleep = "")
```

```
## [1] "158695  People with a Potential Sleep Disorder "
```

```r
paste("Percentage ", round(nrow(insomnia)*100/nrow(insomnia1), digits = 2), "%", sep ="")
```

```
## [1] "Percentage 32.27%"
```

We can see 32.27% of the people in the sample (158.695) can potentially have some sleep disorders like insomnia, narcolepsy or sleep apnea.

Now I summarized my dataset to facilitate my data analysis. I will only select people who suffers at least of one of both diseases mentioned in this case study and will sum the amount of people according to their sleep hours.

```
insomnia_summary1 <- insomnia %>%
  mutate(hypertension = recode(hypertension,
                               "Yes, but female told only during pregnancy" = "Yes",
                               "Told borderline or pre-hypertensive" = "Yes")) %>%
  filter(hypertension == "Yes" | depression == "Yes") %>%
  count(sleep_hours)

print(insomnia_summary1)
```

```
##   sleep_hours     n
## 1           0     1
## 2           1   155
## 3           2   819
## 4           3  2623
## 5           4  9842
## 6           5 20059
## 7           6 55532
```

Looking at the dataset above I could find how many people who sleep less than 6 hours suffer of either hypertension or depression, or even both.

```
paste(sum(insomnia_summary1$n), " People with depression or/and hypertension", sep = "")
```

```
## [1] "89031 People with depression or/and hypertension"
```

```
paste("Percentage ", round(sum(insomnia_summary1$n)*100/nrow(insomnia), digits = 2), "%" ,sep = "")
```

```
## [1] "Percentage 56.1%"
```

The finds of this analysis are quite chocking because more than half of the sample has a disease that is highly connected to a poor sleep habit.

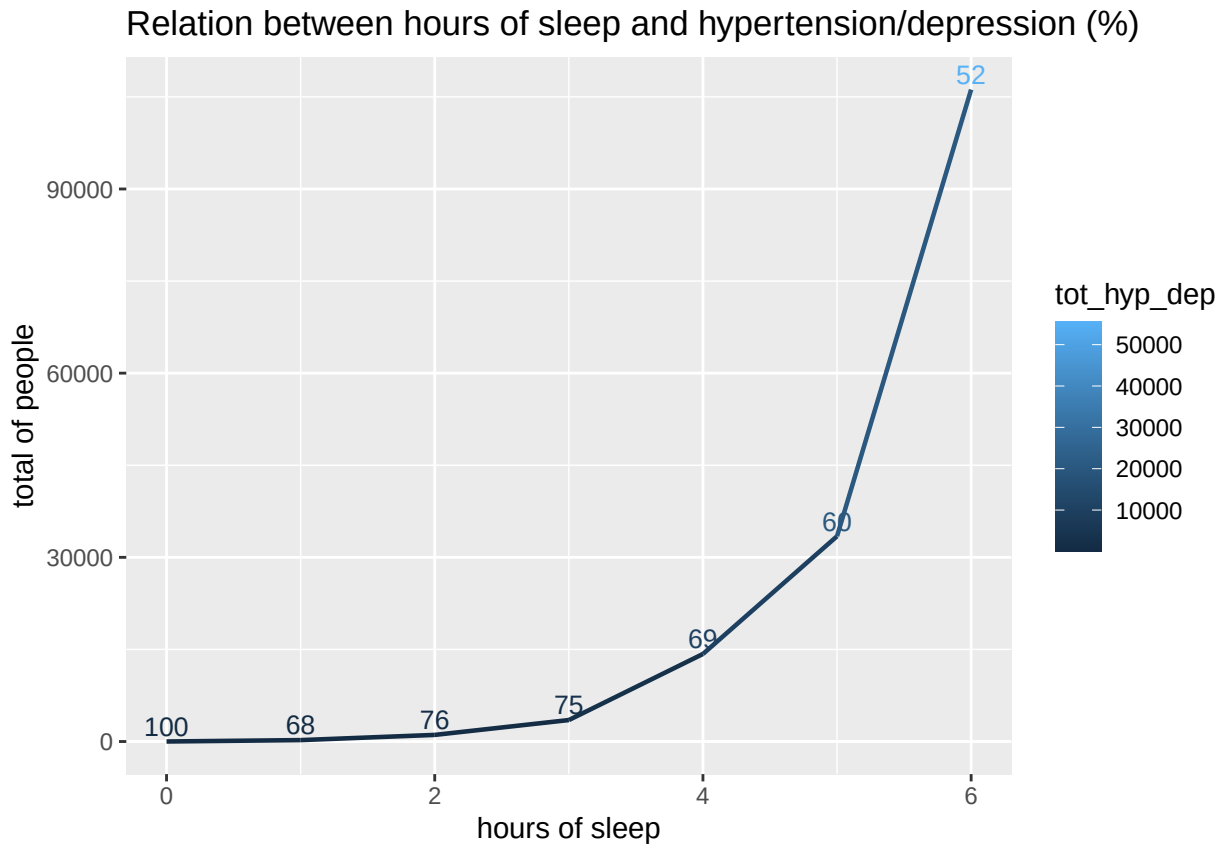
Now I just need to merge the "insomnia_summary1" dataframe into the "insomnia_summary2" one to be able to create a proper plot about the results I want to.

```
completo <- left_join(insomnia_summary1, insomnia_summary2, by = "sleep_hours") %>%
  rename("tot_hyp_dep" = n.x,
         "total" = n.y)

completo <- completo %>%
  mutate(percent = round(completo$tot_hyp_dep*100/completo$total))


print(completo)
```

```
##   sleep_hours tot_hyp_dep  total percent
## 1           0           1      1     100
## 2           1         155    228      68
## 3           2         819   1076      76
## 4           3        2623   3496      75
## 5           4        9842  14261      69
## 6           5       20059  33436      60
## 7           6       55532 106197      52
```

Finally I can generate my chart about the possible relation among sleep disorders and hypertension/depression

```
ggplot(completo, aes(x = sleep_hours, y = total, color = tot_hyp_dep)) +
  geom_line(size = 0.8) +
  geom_text(aes(label = percent), vjust = -0.3, hjust = 0.5, size = 3.5) +
  labs (title = "Relation between hours of sleep and hypertension/depression (%) ",
        x = "hours of sleep",
        y = "total of people")
```



As we can see, there is a considerable relation between poor sleep habits to hypertension and depression. The less hours someone sleeps, the more chance he or she has to get a disease associated to that taking into consideration the people in our sample.

Once more, we cannot affirm there is a causation in our sample because we need to get more detailed information about every single person who took part in the survey.

According to the graph:

- 68% of people who sleep only 1 hour a day have at least one of both illnesses;
- 76% of people who sleep only 2 hours a day have at least one of both illnesses;
- 75% of people who sleep only 3 hours a day have at least one of both illnesses;
- 69% of people who sleep only 4 hours a day have at least one of both illnesses;
- 60% of people who sleep only 5 hours a day have at least one of both illnesses;
- 52% of people who sleep only 6 hours a day have at least one of both illnesses.

PS. I am not taking into consideration the 100% concerning 0 (no sleep) because there was only one individual in the sample and it is not statistically relevant. It is probably an outlier.