# Exploring the BRFSS data

## Setup

### Load packages

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

### Load data

Saved in the same direction, Renamed r/markdownfile to: Duke_uni_project.Rmd Renamed brfss2013.RData ; brfss_data.RData

```
load("brfss_data.RData")
```

---

# Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project between all of the states in the United States (US) and participating US territories and the Centers for Disease Control and Prevention (CDC).BRFSS is an ongoing surveillance system designed to measure behavioral risk factors for the non-institutionalized adult population (18 years of age and older) residing in the US, collecting surveillance data on risk behaviors through monthly telephone interviews.

We will use the Data to find out, more about high Cholesterol and Diabetes. Where is the Distribution of patients with high Cholesterol, does more activities prevents from the disease. The Data was gathered with monthly telephone Surveys.

We can not assume Causalities because of Lack of Experiments, the Date gathered is Observational, but we can Generalize. The Generalization is possible because of the Sample Size. In this Case the Average Sample size is Roundabout 400.000 People.

Bias that could be included; Linguistic, Cosmetic Bias especially for the ,“weekly-activities-min”, Data used in this case.

We can generalize, that with doing sports it is possible to reduce the possibilitie, of having high Cholesterol, by 17%. Also we found out that 9.9% of the questioned people had both illnesses.The Chance being diagnosed with Diabetes growths to 22%, if the patient has high Cholesterol.

---

# Part 2: Research questions

**Research question 1:** We want to know if people with Cholesterol also suffer with Diabetes? And if, what percentage has both diseases, what is the probabilitie being diagnosed with Diabetes give the knowing of having high Cholesterol ?

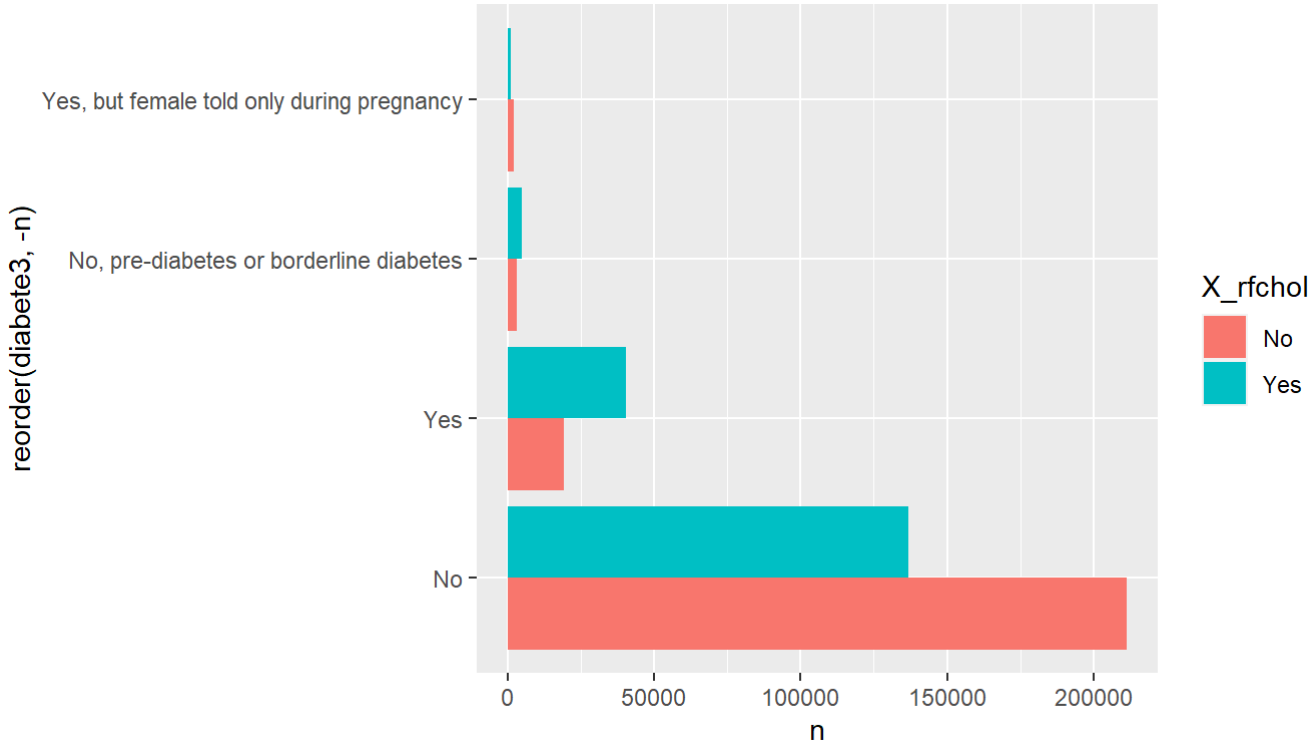
**Research question 2:** Does people who are Active are less common to be sick of Cholesterol?

**Research question 3:** How are the Distributions of Cholesterol, and weekly activities on 6 Groups of diffrent Ages?

# Part 3: Exploratory data analysis

**Research quesion 1:** We want to know if people with Cholesterol also suffer with Diabetes? And if, what percentage has both illnesses?

```
research_1 <- brfss2013 %>% count(diabete3 ,X_rfchol) %>% na.omit() %>% arrange(desc(group_by
        =n))

#Bar plot Horizontal for better reading of the Variables;
ggplot(research_1, aes(fill=X_rfchol, x=reorder(diabete3,-n),y=n)) +
  geom_bar(position ="dodge" ,stat="identity") + coord_flip()
```




```
#Calculations
tibble(research_1)
```

```
## # A tibble: 8 × 3
##   diabete3                                       X_rfchol      n
##   <fct>                                          <fct>     <int>
## 1 No                                             No       211471
## 2 No                                             Yes      136731
## 3 Yes                                            Yes       40323
## 4 Yes                                            No        19404
## 5 No, pre-diabetes or borderline diabetes        Yes        4844
## 6 No, pre-diabetes or borderline diabetes        No         3121
## 7 Yes, but female told only during pregnancy     No         2371
## 8 Yes, but female told only during pregnancy     Yes        1336
```

```
#Summarizing, total Sample Size
research_1 %>% summarise(sum(n, na.rm = TRUE))
```

```
##   sum(n, na.rm = TRUE)
## 1               419601
```

```
#Filtering, Diabetes=Yes, Cholesterol=Yes
rs_math1 <-research_1 %>% filter(diabete3=="Yes",X_rfchol=="Yes")

#Filtering the Pregnants with both Answers Yes
rs_math2 <-research_1 %>% filter(diabete3=="Yes, but female told only during pregnancy", X_rf
         chol=="Yes")

#Results, Dataframes from the Filtering;
tibble(rs_math1)
```

```
## # A tibble: 1 × 3
##   diabete3 X_rfchol     n
##   <fct>    <fct>    <int>
## 1 Yes      Yes      40323
```

```
tibble(rs_math2)
```

```
## # A tibble: 1 × 3
##   diabete3                                  X_rfchol     n
##   <fct>                                     <fct>    <int>
## 1 Yes, but female told only during pregnancy Yes       1336
```

Surveyed = 419.601 pos.Chol+Diab =41.659

= 0.0992

The Possibilitie to have Diabetes and a high Cholesterol is by 0.0992.

But,

it would be pretty surprising getting diagnostic with both disease without a previous problem. We will assume that the Person is diagnosed with high Cholesterol. What are now the Possibilities being Diabetes positive.

Calculations Fast-Forward :

Surveyed = 419601 positive Chol = 183234

Possibilities Cholesterol = 0.436

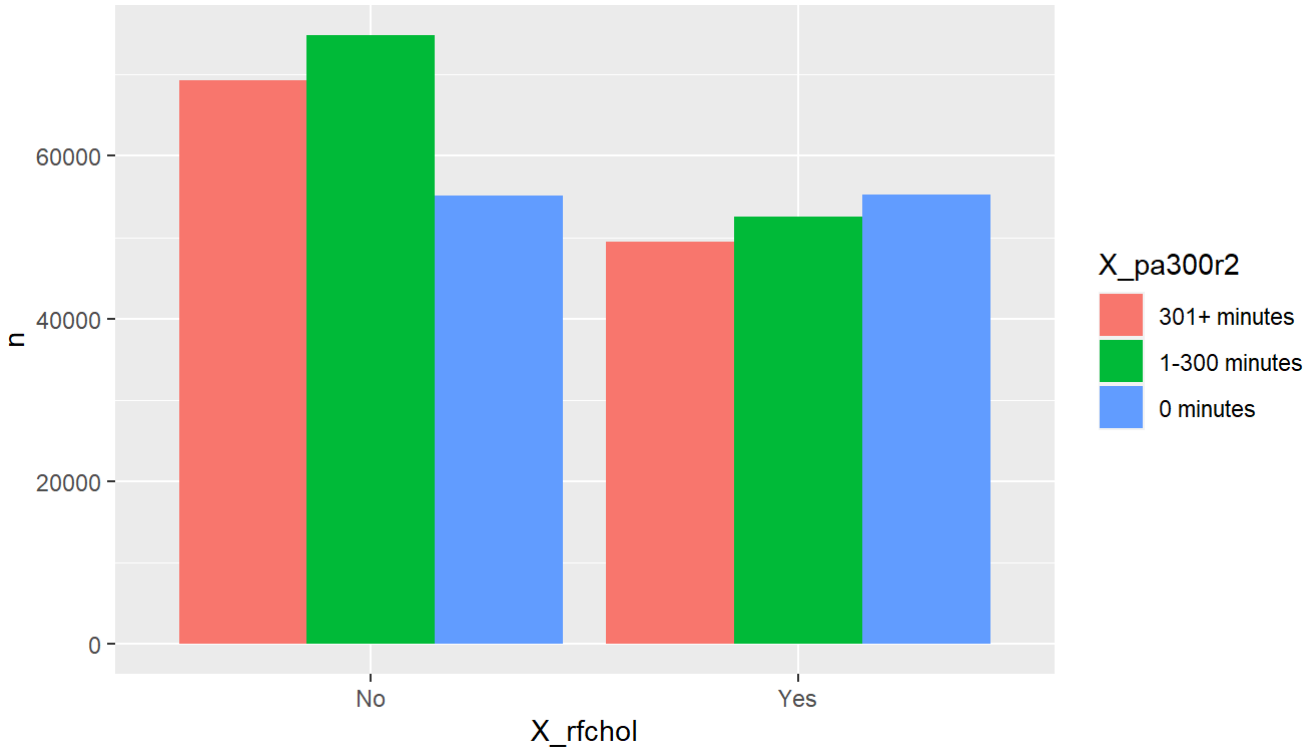
positive Chol = 183234 pos.Chol+Diab = 41659

**Possibilities pos.Chol+Diab Given pos.Chol = 0.2273**
the Chance of having Diabetes diagnosed doubled now, with the Information having high Cholesterol.

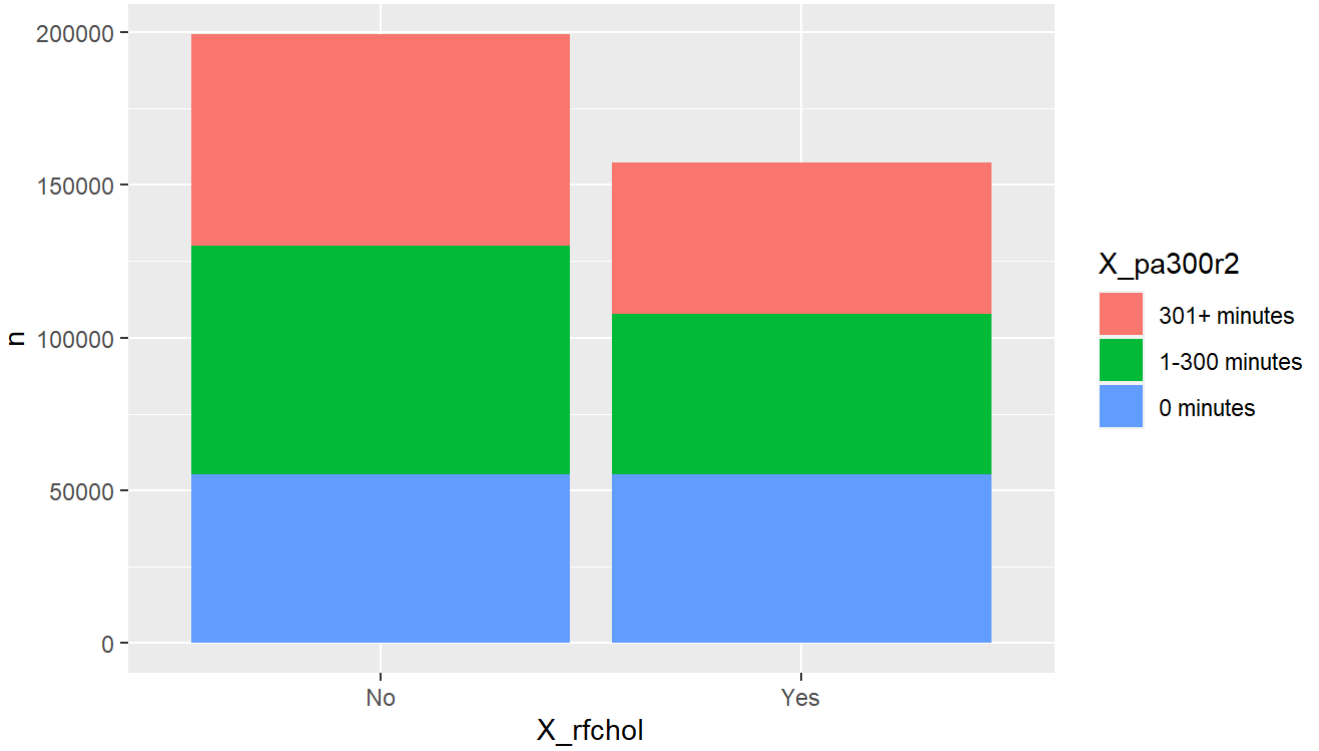
**Research quesion 2:**

How does the distribution look like, with the Variable of being Active? Does people who are Active are less sick of Cholesterol?

```
#Barcharts for better Visualisation

research_2 <- brfss2013 %>% count(X_rfchol ,X_pa300r2) %>% na.omit() %>% arrange(desc(by_grou
        p = n))

ggplot(research_2, aes(fill=X_pa300r2, y=n, x=X_rfchol)) +
    geom_bar(position="dodge", stat="identity")
```




```
ggplot(research_2, aes(fill=X_pa300r2, y=n, x=X_rfchol)) +
    geom_bar(position="stack", stat="identity")
```




```
#Summarizing for Calculation

tibble(research_2)
```

```
## # A tibble: 6 × 3
##   X_rfchol X_pa300r2         n
##   <fct>    <fct>         <int>
## 1 No       1-300 minutes 74791
## 2 No       301+ minutes  69275
## 3 Yes      0 minutes     55237
## 4 No       0 minutes     55169
## 5 Yes      1-300 minutes 52484
## 6 Yes      301+ minutes  49460
```

When we do simple math, summarizing and dividing, we find out that Sport can be a Indicator. Just by looking at the stacked Bar, we can guess the difference (20%) with Sports.

Here are the Differences in the Levels of activities :

p1=Cholesterol positive p2=Cholesterol negative

0 Activities = p1 = 0.5003 p2 = 0.4996

1-300min Activities = p1 = 0.4123 p2 = 0.5876 difference of ~0.17

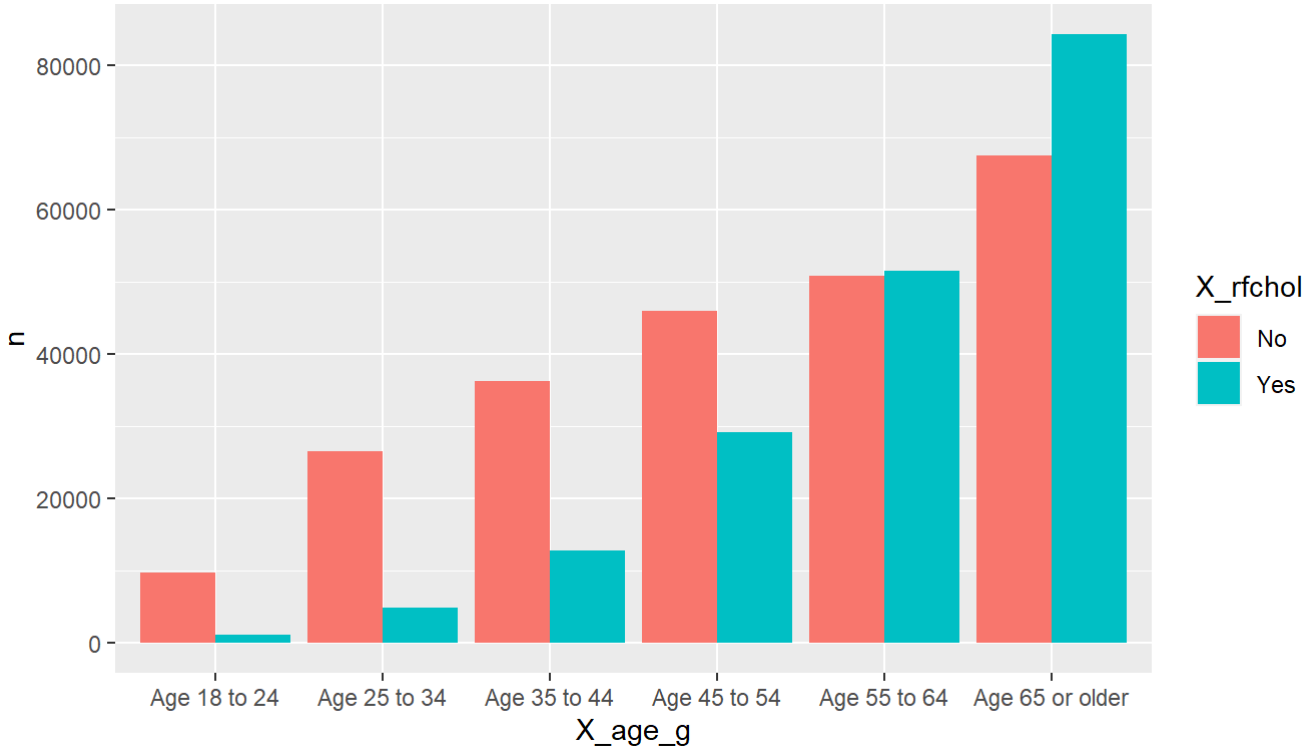
+301min Activities = p1 = 0.4165 p2 = 0.5834 difference of ~ 0.17

We now know, that doing sports reduces the Possibilities of having high Cholesterol by 17%! Not to forget, this is one Variable, to reduce high Cholesterol, it might be helpful to do researches in Food, Sleep, Stress, Smoking, Alcohol as well.

**Research quesion 3:**

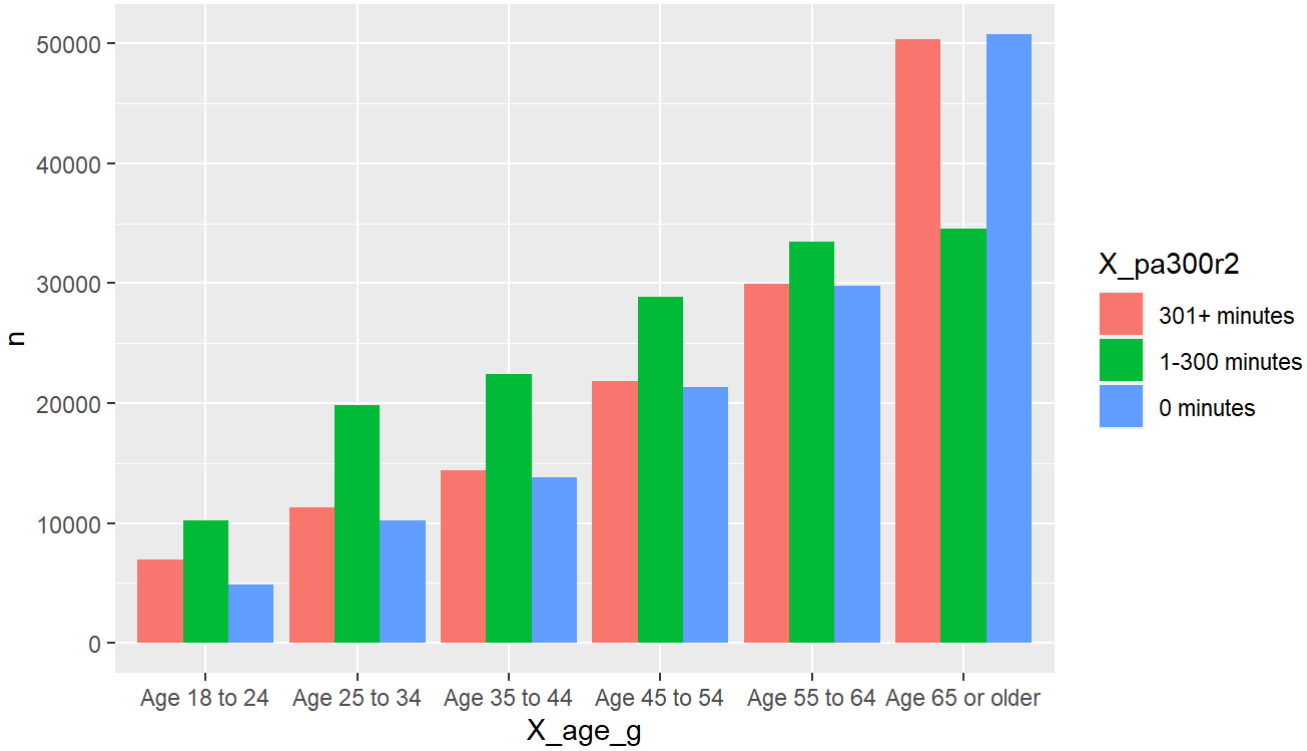
How is the Distributions of Cholesterol, shown on 6 different Age Groups?

```
#Creating a new Data.Frame
research_3 <- brfss2013 %>% count(X_age_g ,X_rfchol) %>% na.omit() %>% arrange(desc(group_by=
        n))

#Visualization of Age and positiv Cholesterol
ggplot(research_3, aes(fill=X_rfchol ,y=n ,x=X_age_g)) +
  geom_bar(position = "dodge" ,stat = "identity")
```




The Distrobution is Left skewed, we can see that with a higher Age there comes a higher possibilitie having a high Cholesterol level. With the Age of 55, the Possibilities of having high Cholesterol is >0.50. That could strengthen our previous researches, but for that we would need another few Variables to compare, like Activities/Age.

```
research_4 <- brfss2013 %>% count(X_age_g ,X_pa300r2) %>% na.omit()

ggplot(research_4, aes(fill=X_pa300r2, y=n, x=X_age_g)) + geom_bar(position="dodge", stat =
        "identity")
```




The Data does match the Assumption, that with an higher Age, the Activities are falling back. Clearly visible that in the Age >55 the '1-300min' is falling back.

This was Created by Sasa Aleksic

During 6-9.February.2023