Exploring the BRFSS data

Setup

Load packages

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

Load data

```
load("brfss2013.RData")
```

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

The observations in the sample are collected via surveys conducted via calls to both landlines and cell phones numbers in all 50 states of the United States of America, as well as the District of Columbia and three U.S. territories.

The landline and cell phone numbers are generated at random using Random Digit Dialing (RDD) techniques. Because there is random sampling done, the result of the analysis can be generalized to the population of interest.

Because there is no random assignment done, the result of the analysis cannot be used for causal inference.

Part 2: Research questions

Research quesion 1: What is the most common chronic health condition that people have been told that they have?

Knowing about this can help us be careful to avoid getting the most common chronic health condition.

Research quesion 2: For arthritis management where doctor suggest to lose weight, what is the typical weight?

Knowing that weight will motivate us to be careful about our weight to avoid getting arthritis due to body weight.

Research quesion 3: What are the top 10 types of exercise with the least percentage of people being told that they have chronic health condition?

Knowing this information may motivate us to participate in one of these types of exercise.

Part 3: Exploratory data analysis

Research quesion 1: What is the most common chronic health condition that people have been told that they have?

```
chronic_cond <- brfss2013 %>%
  summarise(
    heart_attack = sum(cvdinfr4 == 'Yes', na.rm = TRUE)
    , angina_or_chd = sum(cvdcrhd4 == 'Yes', na.rm = TRUE)
    , stroke = sum(cvdstrk3 == 'Yes', na.rm = TRUE)
    , asthma = sum(asthma3 == 'Yes', na.rm = TRUE)
    , skin_cancer = sum(chcscncr == 'Yes', na.rm = TRUE)
    , other_cancer = sum(chcocncr == 'Yes', na.rm = TRUE)
    , copd = sum(chccopd1 == 'Yes', na.rm = TRUE)
    , arthritis = sum(havarth3 == 'Yes', na.rm = TRUE)
    , depressive_disorder = sum(addepev2 == 'Yes', na.rm = TRUE)
    , kidney_disease = sum(chckidny == 'Yes', na.rm = TRUE)
    , diabetes = sum(diabete3 == 'Yes', na.rm = TRUE)
    )

chronic_cond
```

```
##   heart_attack angina_or_chd stroke asthma skin_cancer other_cancer  copd
## 1        29284         29064  20391  67204       45446        47074 40660
##   arthritis depressive_disorder kidney_disease diabetes
## 1    165152               95779          15901    62363
```

To make the summary statistics easier to interpret, I will rotate the table counterclockwise, and sort them by their count in descending order.

```
chronic_cond <- gather(chronic_cond, 'condition_type', 'count') %>%
  arrange(desc(count))

chronic_cond
```

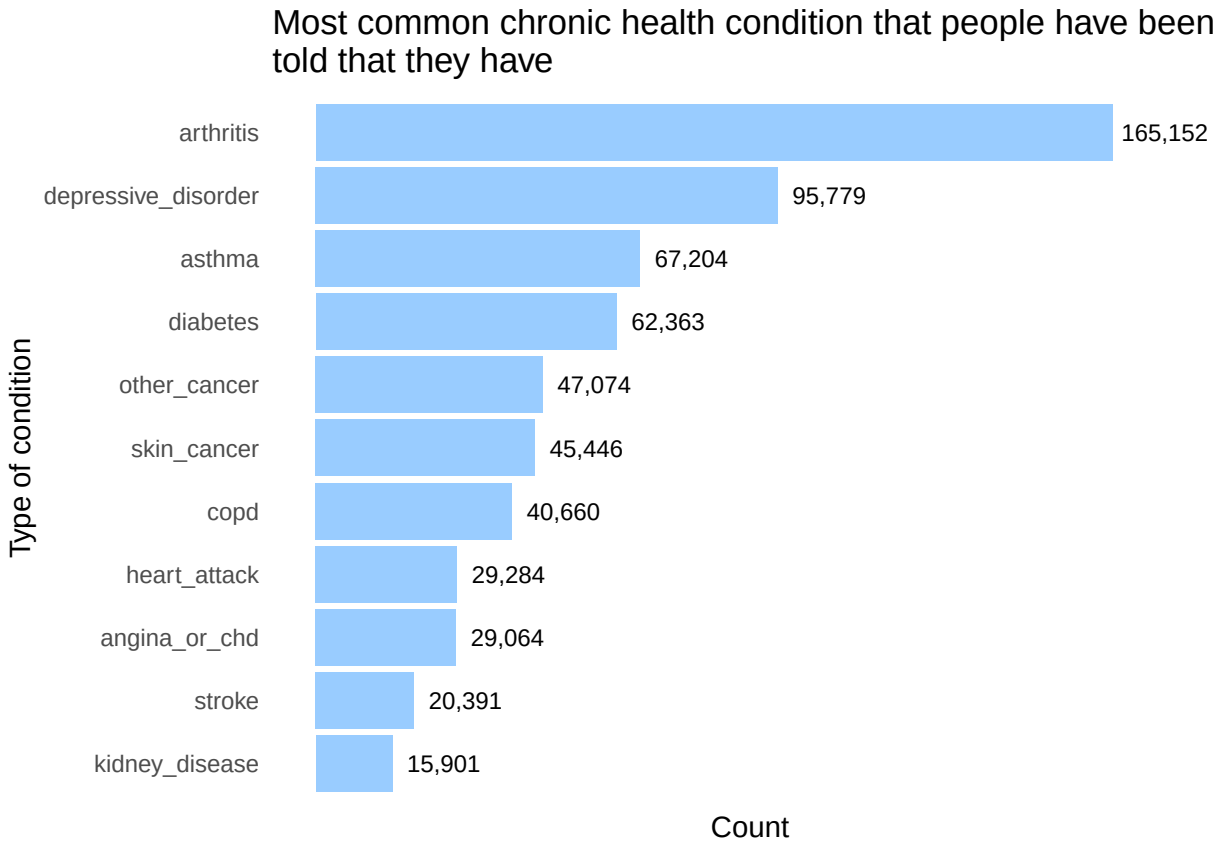
```
##         condition_type  count
## 1            arthritis 165152
## 2  depressive_disorder  95779
## 3               asthma  67204
## 4             diabetes  62363
## 5         other_cancer  47074
## 6          skin_cancer  45446
## 7                 copd  40660
## 8         heart_attack  29284
## 9        angina_or_chd  29064
## 10              stroke  20391
## 11      kidney_disease  15901
```

From the summary statistics, it shows that arthritis is the most common chronic health condition that people have been told that they have.

I will generate a bar plot for better understanding.

```
chronic_cond$condition_type <-
  factor(
    chronic_cond$condition_type
    , levels = chronic_cond$condition_type[order(chronic_cond$count)]
    )
```

```
ggplot(data = chronic_cond, aes(x = count, y = condition_type)) +
  geom_col(fill = '#99ccff') +
  ggtitle('Most common chronic health condition that people have been
told that they have') +
  xlab('Count') +
  ylab('Type of condition') +
  scale_x_continuous(labels = NULL) +
  theme_minimal() +
  theme(axis.ticks.x = element_blank()
        , axis.ticks.y = element_blank()
        , panel.grid = element_blank()) +
  geom_text(aes(label = format(count, big.mark = ","))
            , hjust = -0.1, size = 3) +
  coord_cartesian(xlim = c(0, 180000))
```



Research quesion 2: For arthritis management where doctor suggest to lose weight, what is the typical weight?

```
having_arthritis <- brfss2013 %>% filter(arthwgt == 'Yes')

having_arthritis <- having_arthritis %>%
  mutate(weight2_num = as.numeric(as.character(weight2)))

having_arthritis %>%
  summarise(
    average_weight = mean(weight2_num, na.rm = TRUE)
    , standard_deviation = sd(weight2_num, na.rm = TRUE)
```

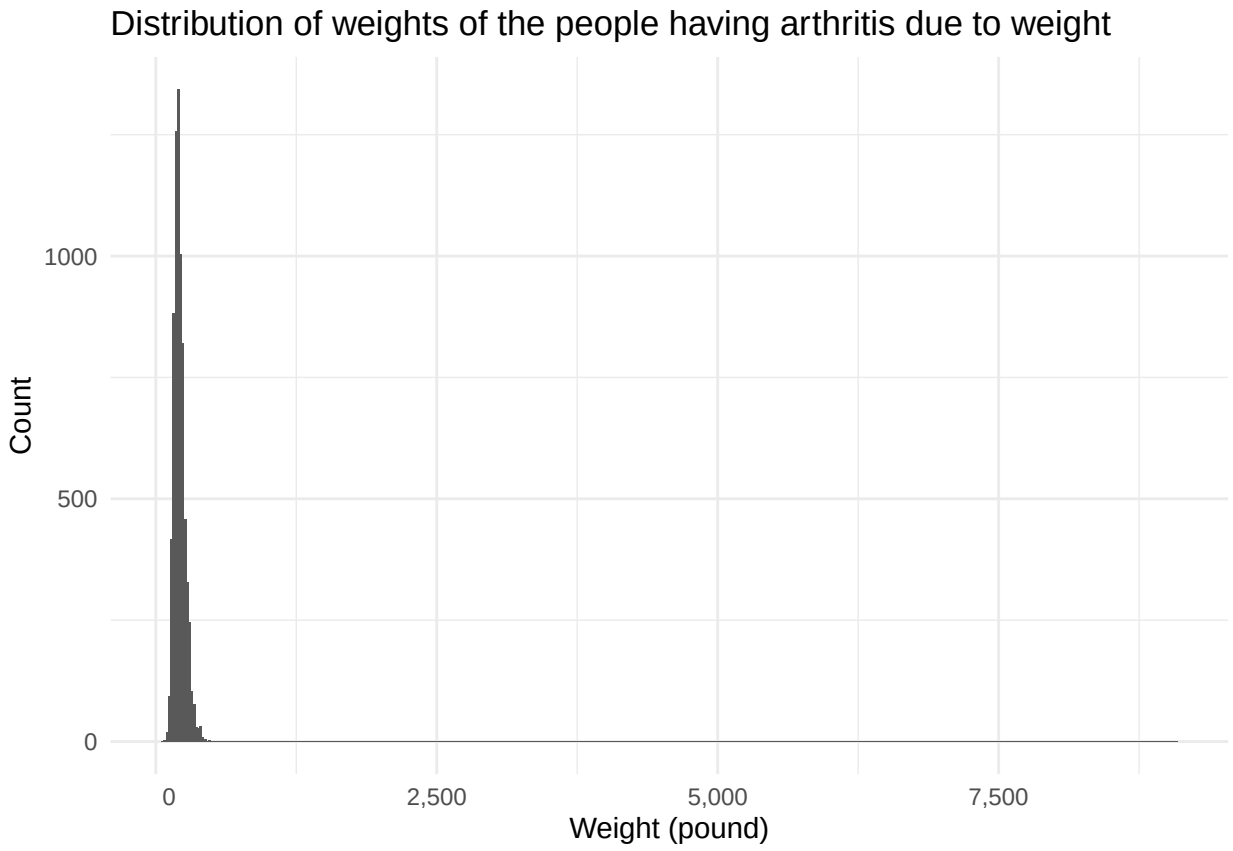
```
    , max_weight = max(weight2_num, na.rm = TRUE)
    )
```

```
##   average_weight standard_deviation max_weight
## 1       216.0014           156.4615       9077
```

From the summary statistics, it shows that the average weight for people having arthritis due to weight is 216 pounds. The standard deviation of 156 pounds is very large, so there should be an outlier, which is confirmed by the max weight of 9077 pounds.

I will generate a histogram so that it the outlier is more obvious shown by the limit on the X axis.

```
ggplot(data=having_arthritis, aes(x=weight2_num)) +
  geom_histogram(binwidth = 20) +
  ggtitle('Distribution of weights of the people having arthritis due to weight') +
  xlab('Weight (pound)') +
  ylab('Count') +
  scale_x_continuous(labels = function(x) format(x, big.mark = ",",
                                                 scientific = FALSE)) +
  theme_minimal()
```



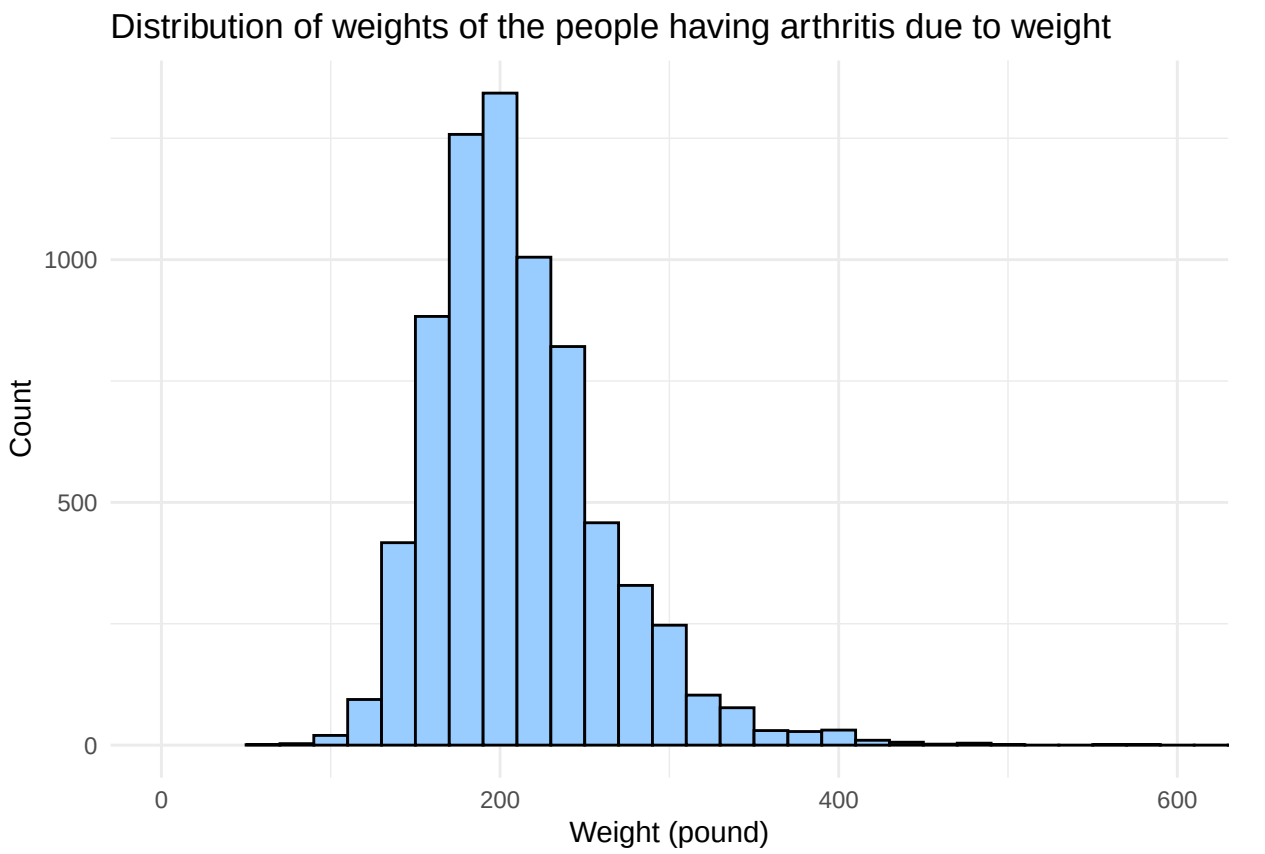

```
quantile(having_arthritis$weight2_num, c(0.25, 0.5, 0.75), na.rm = TRUE)
```

```
## 25% 50% 75%
## 180 203 240
```

So to be more accurate on the typical weight of the people who have arthritis due to weight, it is the middle 50% of the weights, which is from 180 pounds (i.e. 25th percentile) to 240 pounds (i.e. 75th percentile).

I will generate a histogram with limit on the X axis to exclude the outlier.

```r
ggplot(data=having_arthritis, aes(x=weight2_num)) +
  geom_histogram(binwidth = 20, fill = '#99ccff', color = 'black') +
  ggtitle('Distribution of weights of the people having arthritis due to weight') +
  xlab('Weight (pound)') +
  ylab('Count') +
  theme_minimal() +
  coord_cartesian(xlim = c(0, 600))
```



Research quesion 3: What are the top 10 types of exercise with the least percentage of people being told that they have chronic health condition?

```r
brfss2013 <- brfss2013 %>%
  mutate(told_have_chronic_cond=ifelse(
                                cvdinfr4 == 'Yes'
                                | cvdcrhd4 == 'Yes'
                                | cvdstrk3 == 'Yes'
                                | asthma3 == 'Yes'
                                | chcscncr == 'Yes'
                                | chcocncr == 'Yes'
                                | chccopd1 == 'Yes'
                                | havarth3 == 'Yes'
                                | addepev2 == 'Yes'
                                | chckidny == 'Yes'
                                | diabete3 == 'Yes'
                                , 'Yes', 'No'
                                )
         )
```

```
exercise_and_chr_cond <- brfss2013 %>%
  filter(
    told_have_chronic_cond %in% c('Yes', 'No')
    , !is.na(exract11)) %>%
  group_by(exract11) %>%
  summarise(
    percent_told_have_chronic_cond = round(sum(told_have_chronic_cond == 'Yes')
                                           / n() * 100
                                           , 1)
    ) %>%
  rename(exercise_type = exract11) %>%
  arrange(percent_told_have_chronic_cond)

top_10_exercise_types <- head(exercise_and_chr_cond, 10)
top_10_exercise_types
```

```
## # A tibble: 10 x 2
##    exercise_type  percent_told_have_chronic_cond
##    <fct>                                  <dbl>
##  1 Lacrosse                                23.1
##  2 Soccer                                  25.5
##  3 Rugby                                   26.3
##  4 Touch football                          28
##  5 Basketball                              31.4
##  6 Running                                 31.5
##  7 Hockey                                  32.5
##  8 Volleyball                              32.6
##  9 Rock Climbing                           34.2
## 10 Jogging                                 36.1
```

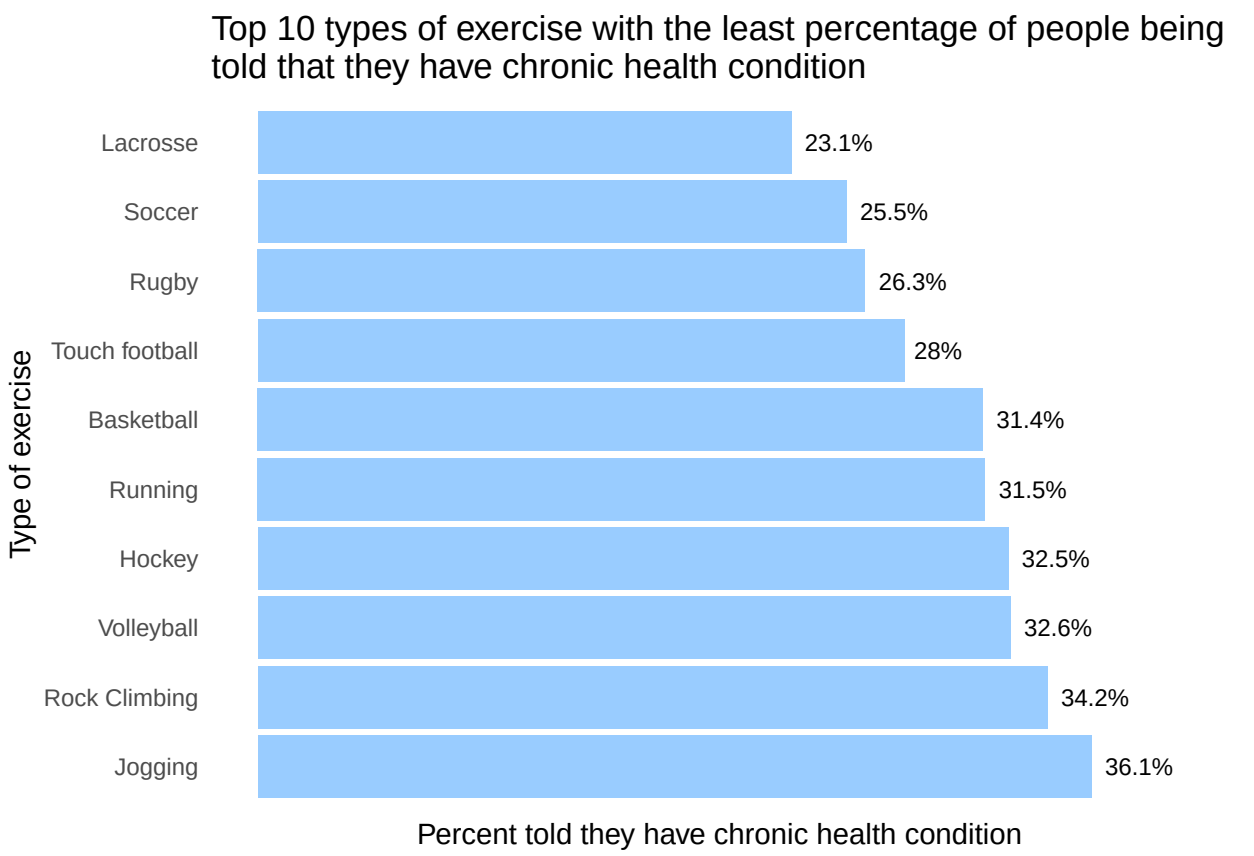
The above summary statistics shows the top 10 types of exercise with the least percentage of people being told that they have chronic health condition. It shows that most of them are high intensity exercise.

I will generate a bar plot for better understanding.

```
top_10_exercise_types$exercise_type <-
  factor(
    top_10_exercise_types$exercise_type
    , levels = top_10_exercise_types$exercise_type[
        order(
        top_10_exercise_types$percent_told_have_chronic_cond
        , decreasing=TRUE
        )
      ]
    )

ggplot(data=top_10_exercise_types
       , aes(x=percent_told_have_chronic_cond
             , y=exercise_type)) +
  geom_col(fill = '#99ccff') +
  ggtitle('Top 10 types of exercise with the least percentage of people being
told that they have chronic health condition') +
  xlab('Percent told they have chronic health condition') +
  ylab('Type of exercise') +
  geom_text(aes(label = paste(percent_told_have_chronic_cond, '%', sep = '')))
```

```
          , hjust = -0.2, size = 3) +
  scale_x_continuous(labels = NULL) +
  theme_minimal() +
  theme(axis.ticks = element_blank()
        , panel.grid = element_blank()) +
  coord_cartesian(xlim = c(0, 40))
```



I will also do a summary statistics of the top 10 types of exercise with the most percentage of people being told that they have chronic health condition, so that you can compare the percentages from both statistics.

```
bottom_10_exercise_types <- exercise_and_chr_cond %>%
  arrange(desc(percent_told_have_chronic_cond)) %>%
  head(n=10)

bottom_10_exercise_types
```

```
## # A tibble: 10 x 2
##    exercise_type                                    percent_told_have_chronic~
##    <fct>                                                                <dbl>
##  1 Tai Chi                                                               77.2
##  2 Household Activities (vacuuming, dusting, home re~                    73.2
##  3 Bicycling machine exercise                                            71.1
##  4 Bowling                                                               70.5
##  5 Gardening (spading, weeding, digging, filling)                        69.8
##  6 Snorkeling                                                            68.8
##  7 Fishing from river bank or boat                                       68.3
##  8 Raking lawn                                                           67.5
```

```
##  9 Upper Body Cycle (wheelchair sports, ergometer, e~                  66.7
## 10 Mowing lawn                                                         66
```