Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

Load data

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.

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

Part 1: Data. Learn something about the data set.

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). The BRFSS is administered and supported by CDC's Population Health Surveillance Branch, under the Division of Population Health at the National Center for Chronic Disease Prevention and Health Promotion. 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. The BRFSS was initiated in 1984, with 15 states collecting surveillance data on risk behaviors through monthly telephone interviews. Over time, the number of states participating in the survey increased; by 2001, 50 states, the District of Columbia, Puerto Rico, Guam, and the US Virgin Islands were participating in the BRFSS. Today, all 50 states, the District of Columbia, Puerto Rico, and Guam collect data annually and American Samoa, Federated States of Micronesia, and Palau collect survey data over a limited point- in-time (usually one to three months). In this document, the term "state" is used to refer to all areas participating in BRFSS, including the District of Columbia, Guam, and the Commonwealth of Puerto Rico.

The BRFSS objective is to collect 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. Factors assessed by the BRFSS in 2013 include tobacco use, HIV/AIDS knowledge and prevention, exercise, immunization, health status, healthy days — health-related quality of life, health care access, inadequate sleep, hypertension awareness, cholesterol awareness, chronic health conditions, alcohol consumption, fruits and vegetables consumption, arthritis burden, and seatbelt use.

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.

Health characteristics estimated from the BRFSS pertain to the non-institutionalized adult population, aged 18 years or older, who reside in the US. In 2013, additional question sets were included as optional modules to provide a measure for several childhood health and wellness indicators, including asthma prevalence for people aged 17 years or younger.

Part 2: Research questions

Research question 1: As a first question, we might be interested in exploring the relationship between Cholesterol awareness.Focusing on the question, we want to know how many people are aware about cholesterol. To achieve this, we familiarize ourselves with the variables bloodcho(Ever had Blood cholesterol checked), cholchk(How long Since cholesterol checked), and toldhi2(Ever Told Blood Cholesterol High).

Research question 2: As a second question, we might be interested in exploring the chronic health conditions data of atleast 4 variables. **Research question 3:** As a third question, we might be interested in exploring the Exercise(physical activity) survey. We will explore the different types of physical activity, exercise in past 30 days and we will also analyze the exeroft1 variable.

Part 3: Exploratory data analysis

Research question 1:

```
brfss2013 %>%
  select(bloodcho,cholchk,toldhi2) %>%
  str()
```

```
## 'data.frame':    491775 obs. of  3 variables:
##  $ bloodcho: Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...
##  $ cholchk : Factor w/ 4 levels "Within past year",..: 1 1 4 1 2 1 1 1 1 1 ...
##  $ toldhi2 : Factor w/ 2 levels "Yes","No": 1 2 2 1 2 1 2 1 1 2 ...
```

These are all categorical data, however they are recorded as characters (text strings) as opposed to factors.

An easy way of tabulating these data to see how many times each level of is to use the group_by() function along with the summarise() command:

```
brfss2013 %>%
  group_by(bloodcho) %>%
  summarise(count=n())
```

```
## # A tibble: 3 x 2
##   bloodcho  count
##   <fct>     <int>
## 1 Yes      423868
## 2 No        58897
## 3 <NA>       9010
```

This set has NA entries. NA entries need special targeting because they do not actually exist (they are different to the text “NA” or a variable saved with the name NA).

We want the entries that are not NAs we can use the Not operator, !, to indicate “we want all the ones that are not NA”:!is.na(). Hence we can filter out all non NAs in our dplyr chain:

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

```
## # A tibble: 2 x 2
##   bloodcho  count
##   <fct>     <int>
## 1 Yes      423868
## 2 No        58897
```

From this we can say that 58897 are not aware about cholesterol, regretting the NA values.Which means that out of 482765 responses approximately around 12% people are not aware of cholesterol.

We can also similarly view the levels and number of occurrences of these levels in the cholchk variable:

```
brfss2013 %>%
  group_by(cholchk) %>%
  summarise(count=n())
```

```
## # A tibble: 5 x 2
##   cholchk              count
##   <fct>                <int>
## 1 Within past year    323331
## 2 Within past 2 years  49549
## 3 Within past 5 years  29985
## 4 5 or more years ago  15732
## 5 <NA>                 73178
```

This set also has NA entries,so once again we have to use the filter function.

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

```
## # A tibble: 4 x 2
##   cholchk              count
##   <fct>                <int>
## 1 Within past year    323331
## 2 Within past 2 years  49549
## 3 Within past 5 years  29985
## 4 5 or more years ago  15732
```

This shows that around 77% of have checked their cholesterol within past year. Similarly doing the same for toldhi2 variable,

```
brfss2013 %>%
  filter(!is.na(toldhi2)) %>%
  group_by(toldhi2) %>%
  summarise(count=n())
```

```
## # A tibble: 2 x 2
##   toldhi2  count
##   <fct>    <int>
## 1 Yes     183501
## 2 No      236612
```

From this we can find that 183501 people have been told that their blood cholesterol is high.

Research question 2:

```
brfss2013 %>%
  select(cvdinfr4,cvdcrhd4,cvdstrk3,asthma3) %>%
  str()
```

```
## 'data.frame':    491775 obs. of  4 variables:
##  $ cvdinfr4: Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ cvdcrhd4: Factor w/ 2 levels "Yes","No": NA 2 2 2 2 2 2 1 2 2 ...
##  $ cvdstrk3: Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ asthma3 : Factor w/ 2 levels "Yes","No": 1 2 2 2 1 2 2 2 2 2 ...
```

These are all categorical data, however they are recorded as characters (text strings) as opposed to factors.

An easy way of tabulating these data to see how many times each level of is to use the group_by() function along with the summarise() command:

```
brfss2013 %>%
  filter(!is.na(cvdinfr4)) %>%
  group_by(cvdinfr4) %>%
  summarise(count=n())
```

```
## # A tibble: 2 x 2
##   cvdinfr4  count
##   <fct>     <int>
## 1 Yes       29284
## 2 No       459904
```

Therefore out of a total of 489188, 29284 has been diagnosed with heart attack. Which is around 6% approximately.

```
brfss2013 %>%
  filter(!is.na(cvdcrhd4)) %>%
  group_by(cvdcrhd4) %>%
  summarise(count=n())
```

```
## # A tibble: 2 x 2
##   cvdcrhd4  count
##   <fct>     <int>
## 1 Yes       29064
## 2 No       458288
```

This shows that out of 487352, 29064 has been Diagnosed With Angina Or Coronary Heart Disease

```
brfss2013 %>%
  filter(!is.na(cvdstrk3)) %>%
  group_by(cvdstrk3) %>%
  summarise(count=n())
```

```
## # A tibble: 2 x 2
##   cvdstrk3  count
##   <fct>     <int>
## 1 Yes       20391
## 2 No       469917
```

Here out of 490308, 20391 has been Diagnosed With A Stroke, which is around 4.1% approximately.

```
brfss2013 %>%
  filter(!is.na(asthma3)) %>%
  group_by(asthma3) %>%
  summarise(count=n())
```

```
## # A tibble: 2 x 2
##   asthma3  count
##   <fct>    <int>
## 1 Yes      67204
## 2 No      423012
```

Out of 490216 people, 67204 has been told they had asthma, which is around 13.7% respectively.

Research question 3:

```
brfss2013 %>%
  select(exerany2,exract11,exeroft1) %>%
  str()
```

```
## 'data.frame':    491775 obs. of  3 variables:
##  $ exerany2: Factor w/ 2 levels "Yes","No": 2 1 2 1 2 1 1 1 1 1 ...
##  $ exract11: Factor w/ 75 levels "Active Gaming Devices (Wii Fit, Dance, Dance revolutio
n)",..: NA 64 NA 64 NA 6 64 64 7 64 ...
##  $ exeroft1: int  NA 105 NA 205 NA 102 220 102 102 220 ...
```

First we will analyze the extract11 variable, which is a factor.

```
brfss2013 %>%
  filter(!is.na(exract11)) %>%
  group_by(exract11) %>%
  summarise(count=n())
```

```
## # A tibble: 75 x 2
##    exract11                                                                count
##    <fct>                                                                   <int>
##  1 "Active Gaming Devices (Wii Fit, Dance, Dance revolution)"                308
##  2 "Aerobics video or class"                                                8154
##  3 "Backpacking"                                                              43
##  4 "Badminton"                                                                39
##  5 "Basketball"                                                             2251
##  6 "Bicycling machine exercise"                                             7223
##  7 "Bicycling"                                                              8565
##  8 "Boating (Canoeing, rowing, kayaking, sailing for pleasure or camping)"   192
##  9 "Bowling"                                                                682
## 10 "Boxing "                                                                307
## # ... with 65 more rows
```

From the above data we can say that there are 75 different types of physical activity. And out of those 75 Walking has the more number of people.

```
brfss2013 %>%
  filter(!is.na(exerany2)) %>%
  group_by(exerany2) %>%
  summarise(count=n())
```

```
## # A tibble: 2 x 2
##   exerany2  count
##   <fct>     <int>
## 1 Yes      332464
## 2 No       125282
```

From the above data out of 457746, 332464 has exercised in the past 30 days.

```
brfss2013 %>%
  filter(!is.na(exeroft1)) %>%
  group_by(exeroft1) %>%
  summarise(count=n())
```

```
## # A tibble: 126 x 2
##    exeroft1 count
##       <int> <int>
##  1        0     1
##  2        1     1
##  3        2     1
##  4       13     1
##  5      101 20061
##  6      102 38777
##  7      103 59275
##  8      104 29172
##  9      105 31221
## 10      106 10840
## # ... with 116 more rows
```

Now here considering exeroft1 like a categorical data, we get 126 rows with 230 as the most occurrences.