DATA 606 Data Project Proposal

Jai Jeffryes

10/8/2019

Data Preparation

Set up

```
library(psych)
library(ggplot2)
```

Download data (set not to execute)

```
library(lodown) # Anthony Damico's lodown package, available from GitHub; ajdamico/lodown.

# Set download directory
dl_dir <- file.path(getwd(), "DATA606", "Project", "data")

# Get list of BRFSS files.
brfss_cat <-get_catalog("brfss", output_dir = dl_dir)
brfss_cat <- subset(brfss_cat, year == 2018)
# Download
brfss_cat <- lodown("brfss" , brfss_cat)
```

Prepare data (set not to execute)

After preparing the data, I saved it to a file so repetition of this chunk is unnecessary.

```
brfss_df <- readRDS(file.path("data", "2018 main.rds"))

# There are a lot of columns in these data frames. This will subset the data
# to include only the variables we are interested in. We will also rename
# the columns to be more descriptive.

variables_to_keep <- c(
    "xstate", "fmonth", "imonth", "iday", "iyear",
    "genhlth", "exerany2",
    "sex1", "height3",
    "diabete3", "diabage2", "prediab1",
    "xageg5yr", "xage65yr", "htin4", "xbmi5", "xbmi5cat", "xrfbmi5"
)

brfss_df <- brfss_df[ variables_to_keep ]; gc()

names(brfss_df) <- c(
    "state", "file_month", "interview_month", "interview_day", "interview_year",
    "general_health", "exercise",
    "sex", "height",
    "has_diabetes", "diabetes_age", "is_prediabetic",
```

```
    "age", "is_65", "height_inches", "bmi", "bmi_category", "is_overweight_or_obese"
)

# Factors. Use code book and Emacs macros to build.

# state
state_labels <- c(
    "Alabama", "Alaska", "Arizona", "Arkansas", "California",
    "Colorado", "Connecticut", "Delaware", "District", "Florida",
    "Georgia", "Hawaii", "Idaho", "Illinois", "Indiana",
    "Iowa", "Kansas", "Kentucky", "Louisiana", "Maine",
    "Maryland", "Massachusetts", "Michigan", "Minnesota", "Mississippi",
    "Missouri", "Montana", "Nebraska", "Nevada", "New Hampshire",
    "New Jersey", "New Mexico", "New York", "North Carolina", "North Dakota",
    "Ohio", "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island",
    "South Carolina", "South Dakota", "Tennessee", "Texas", "Utah",
    "Vermont", "Virginia", "Washington", "West Virginia", "Wisconsin",
    "Wyoming", "Guam", "Puerto Rico"
)

state_levels <- c(
    1, 2, 4, 5, 6, 8, 9, 10, 11, 12,
    13, 15, 16, 17, 18, 19, 20, 21, 22, 23,
    24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
    34, 35, 36, 37, 38, 39, 40, 41, 42, 44,
    45, 46, 47, 48, 49, 50, 51, 53, 54, 55,
    56, 66, 72
)

brfss_df$state <- factor(brfss_df$state,
                         labels = state_labels,
                         levels = state_levels
                         )


# general_health
general_health_labels <- c(
    "Excellent",
    "Very good",
    "Good",
    "Fair",
    "Poor",
    "Don't know/Not sure",
    "Refused"
)
brfss_df$general_health <- factor(brfss_df$general_health,
                                  labels = general_health_labels,
                                  levels = c(1:5, 7, 9))

# exercise
exercise_labels = c(
    "Yes",
    "No",
```

```
    "Don't know/Not sure",
    "Refused"
)
brfss_df$exercise <- factor(brfss_df$exercise,
                                labels = exercise_labels,
                                levels = c(1, 2, 7, 9))

# bmi_category
bmi_category_labels = c(
    "Underweight",
    "Normal weight",
    "Overweight",
    "Obese"
)
brfss_df$bmi_category <- factor(brfss_df$bmi_category,
                                labels = bmi_category_labels,
                                levels = c(1:4))

# has_diabetes
has_diabetes_labels <- c(
    "Yes",
    "Yes, female told only during pregnancy",
    "No",
    "No, pre-diabetes or borderline",
    "Don't know/Not sure",
    "Refused"
)
brfss_df$has_diabetes <- factor(brfss_df$has_diabetes,
                              labels = has_diabetes_labels,
                              levels = c(1:4, 7, 9))

saveRDS(brfss_df, file.path("data", "BRFSS_2018_subset.RDS"))
```

Load subset

```
brfss_df <- readRDS(file.path("data", "BRFSS_2018_subset.RDS"))
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Are exercise level and body mass index predictive of diabetes?

Cases

What are the cases, and how many are there?

The cases are noninstitutionalized adults residing in the United States. There are 437,436 observations in the dataset.

Data collection

Describe the method of data collection.

Data collection for the Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project between all of the states in the United States and participating US territories and the Centers for Disease Control and Prevention. 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. The data are responses to questions from phone based surveys.

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The Centers for Disease Control and Prevention and the states of the U.S. and participating territories collected the data. CDC publish BRFSS data, and the annual results for year 2018 are available here. For this project, data acquisition was facilitated by the R `lodown` package. Instructions for installing `lodown` are here.

Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2018.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response variable is the case's answer to the survey question, "Have you ever been told you have diabetes?" It is categorical, or qualitative.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

The explanatory variables are body mass index (BMI), which is provided in the dataset as a computed variable given by height and weight, and the case's answer to the survey question, "During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?"

BMI is a numerical, or quantitative, variable. The dataset also provides binned ranges of BMI as categories. Exercise is categorical, or qualitative.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```r
describe(brfss_df$bmi)
```

```
##    vars      n    mean     sd median trimmed   mad  min  max range skew
## X1    1 402174 2826.73 637.88   2728 2764.44 536.7 1205 9873  8668 1.43
##    kurtosis   se
## X1     4.88 1.01
```

```r
table(brfss_df$has_diabetes, useNA='ifany')
```

```
##
##                                    Yes
##                                  60703
## Yes, female told only during pregnancy
##                                   3857
##                                     No
##                                 363757
##         No, pre-diabetes or borderline
##                                   8263
##                    Don't know/Not sure
##                                    567
##                                Refused
##                                    265
##                                   <NA>
##                                     24
```

```r
prop.table(table(brfss_df$has_diabetes, useNA='ifany')) * 100
```

```
##
##                                    Yes
##                            13.877001436
## Yes, female told only during pregnancy
##                             0.881728984
##                                     No
##                            83.156621769
##         No, pre-diabetes or borderline
##                             1.888962042
##                    Don't know/Not sure
##                             0.129618961
##                                 Refused
##                             0.060580291
##                                    <NA>
##                             0.005486517
```

```r
table(brfss_df$bmi_category, useNA='ifany')
```

```
##
##   Underweight Normal weight    Overweight         Obese          <NA>
##          6776        123522        143878        127998         35262
```

```r
prop.table(table(brfss_df$bmi_category, useNA='ifany')) * 100
```

```
##
##   Underweight Normal weight    Overweight         Obese          <NA>
##      1.549027     28.237731     32.891212     29.260966      8.061065
```

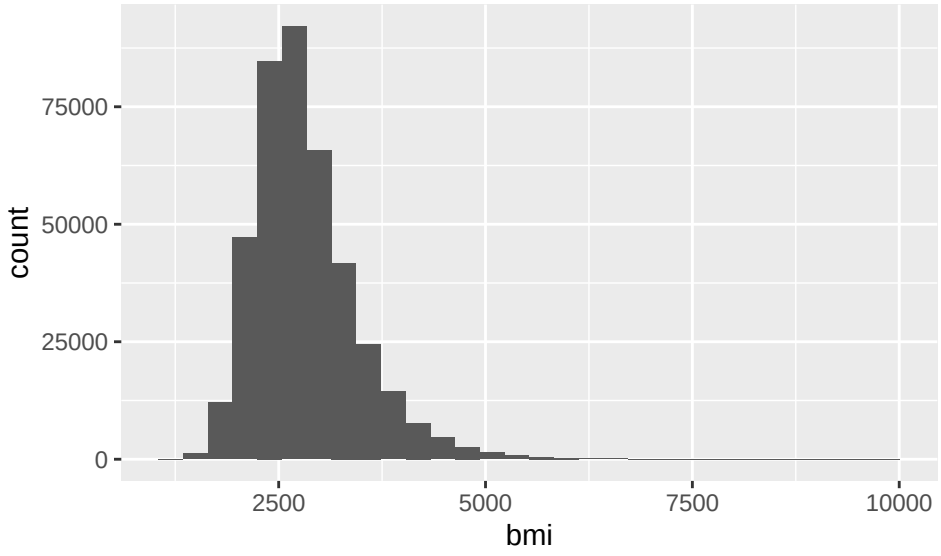
```r
table(brfss_df$exercise, useNA='ifany')
```

```
##
##                  Yes                   No Don't know/Not sure
##               326472               110269                  482
##              Refused                 <NA>
##                  188                   25
```

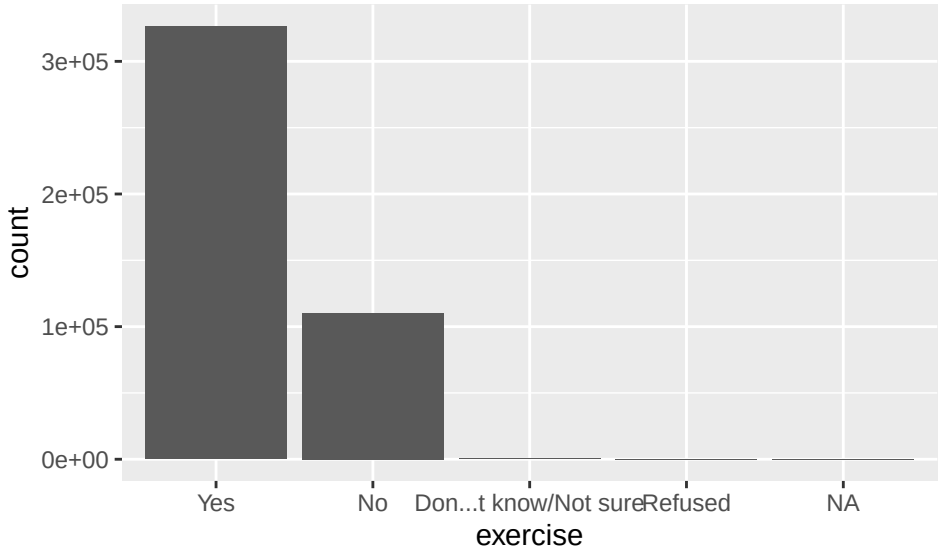
```r
prop.table(table(brfss_df$exercise, useNA='ifany')) * 100
```

```
##
##                  Yes                   No Don't know/Not sure
##        74.633089183        25.208030432         0.110187547
##              Refused                 <NA>
##         0.042977716         0.005715122
```

```r
ggplot(brfss_df, aes(x=bmi)) + geom_histogram()
```



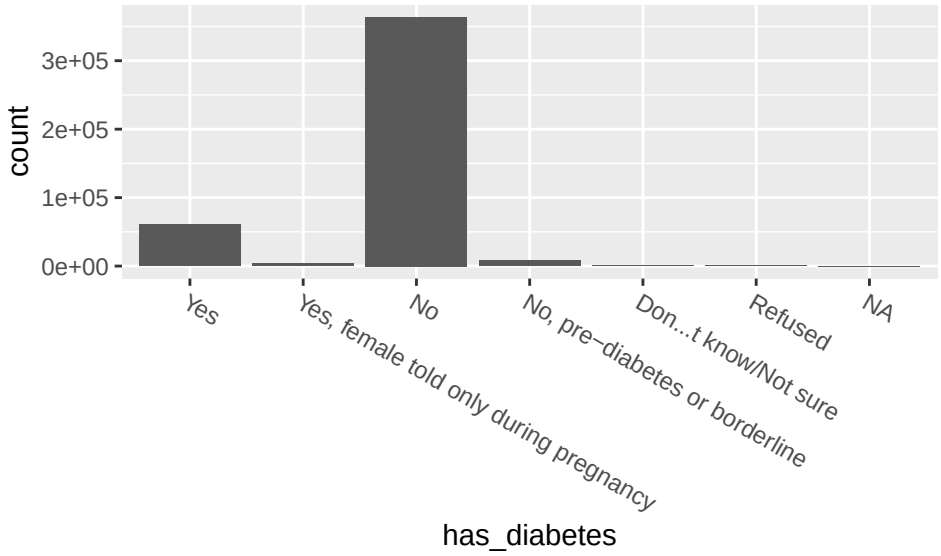

```r
ggplot(brfss_df, aes(x=exercise)) + geom_histogram(stat = "count")
```




```
    theme(axis.text.x = element_text(angle = -30, hjust = 0)) # Justification 0 = left, 1 = right
```

```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : num -30
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
ggplot(brfss_df, aes(x=has_diabetes)) + geom_histogram(stat = "count") +
    theme(axis.text.x = element_text(angle = -30, hjust = 0))
```



Notes for later

Candidate variables for my research question and in case I need record identification or demographcis, etc.

- DIABETE3. Ever told you have diabetes?
- DIABAGE2. How old were you when you were told you have diabetes?
- PREDIAB1. Have you ever been told you have pre-diabetes or borderline?

Section: Record identification.

- _STATE.
- FMONTH. File month.
- IMONTH. Interview month.
- IDAY. Interview day.
- IYEAR. Interview year.

Section: Health status.

- GENHLTH. Would you say your general health is:

Section: Exercise.

- EXERANY2. Did you exercise in the past month?

Section: Demograhics.

- SEX1. What is your sex?
- WEIGHT2.
- HEIGHT3.

Section: Calculated variables.

- _AGE5YR. Binned by 5 years.
- _AGE65YR. Two-level age category.
- HTIN4. Reported height in inches.
- _BMI5. Body mass index.
- _BMI5CAT. Four categoreis of BMI.
- _RFBMI5. Overweight or obese.