Exploring the BRFSS data

Load packages

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

Load data

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

Part 1: The Data

The data for this analysis was collected via phone interviews in 2013 by the CDC through its annual Behavior Risk Factor Surveillance System (BRFSS). The goal of BRFSS is to collect data on health practices and behaviors linked to diseases and injuries. It is an observational study because it does not impose subjects to certain treatments and does not interfere with how the data arises. The observations are collected using stratified sampling. The population is divided into stratas - the states, and then the population is randomly sampled within each strata. Because of this, and the fact that random assignment is not used, generalizability can be established, however, causality cannot be determined.

Part 2: The Research Questions

Below are research questions that are investigated in this analysis.

Research quesion 1: Are people in some states more likely to complete the interview than others?

Research quesion 2: Are fruit consumption and exercise standard across all states?

Research quesion 3: Are vetarans more disposed to alcohol and tobacco use?

Part 3: Exploratory Data Analysis

Research quesion 1: Are people in some states more likely to complete the interview than others?

Though it is often said that New Yorkers are notably less friendly and Southerners are more friendly than other parts of the country, this will be evaluated depending on the rate of interview completion for each geographical region interviewed. This question could give insight as to whether some regions of the country are more inclined to answer phone surveys and potentially more amiable than other parts of the country.

Below is the head of the data frame that will be used for the first question:

```
question1<-data.frame(brfss2013$X_state,brfss2013$dispcode, stringsAsFactors = FALSE) #use stringAsFactors=FAL
SE so I can change the order of the states on the plot<--more on this later
question1<-question1[complete.cases(question1), ] #gets rid of NAs
colnames(question1) <- c("State","Final Disposition")
head(question1, 5)
```

```
##     State   Final Disposition
## 1 Alabama Completed interview
## 2 Alabama Completed interview
## 3 Alabama Completed interview
## 4 Alabama Completed interview
## 5 Alabama Completed interview
```

The completion rate for each state is then calculated by using: (completed interviews/total number of interviews) *100. The head of the resulting data frame is shown below:

```
disp_totals_df<- question1%>%
  group_by(State)%>%
  summarise(`Completion Rate` = round(sum(`Final Disposition`=='Completed interview')*100/n(), digits = 1))
head(disp_totals_df,5)
```

```
## # A tibble: 5 x 2
##        State `Completion Rate`
##       <fctr>             <dbl>
## 1    Alabama              90.3
## 2     Alaska              88.7
## 3    Arizona              80.5
## 4   Arkansas              86.2
## 5 California              75.2
```

The mean, median, and IQR for the completion rates is calculated below.

```
completion_rate_mean <- round(mean(disp_totals_df$`Completion Rate`), digits = 1)
completion_rate_median <- round(median(disp_totals_df$`Completion Rate`), digits = 1)
completion_rate_IQR <- round(IQR(disp_totals_df$`Completion Rate`), digits = 1)
completion_rate_SD <- round(sd(disp_totals_df$`Completion Rate`), digits = 1)

cat("Mean:", completion_rate_mean, "% \n")
```

```
## Mean: 88.3 %
```

```
cat("Median:", completion_rate_median, "% \n")
```

```
## Median: 88.6 %
```

```
cat("IQR:", completion_rate_IQR, "%\n")
```

```
## IQR: 4.6 %
```

```
cat("Std. Dev:", completion_rate_SD, "%")
```

```
## Std. Dev: 4.4 %
```

The median (88.6%) is slighly higher than the mean (88.3%) which suggests that the data is slightly skewed to the left.

Plotly plots strings factors, aka word/categorical variables, alphabetically. In order to plot the completion rates in descending order numerically, these string factors (the states' names) needed to have their level orders reset. Basically the states needed to be re-ordered based on their completion rates, not their spelling, which is the plotly default.

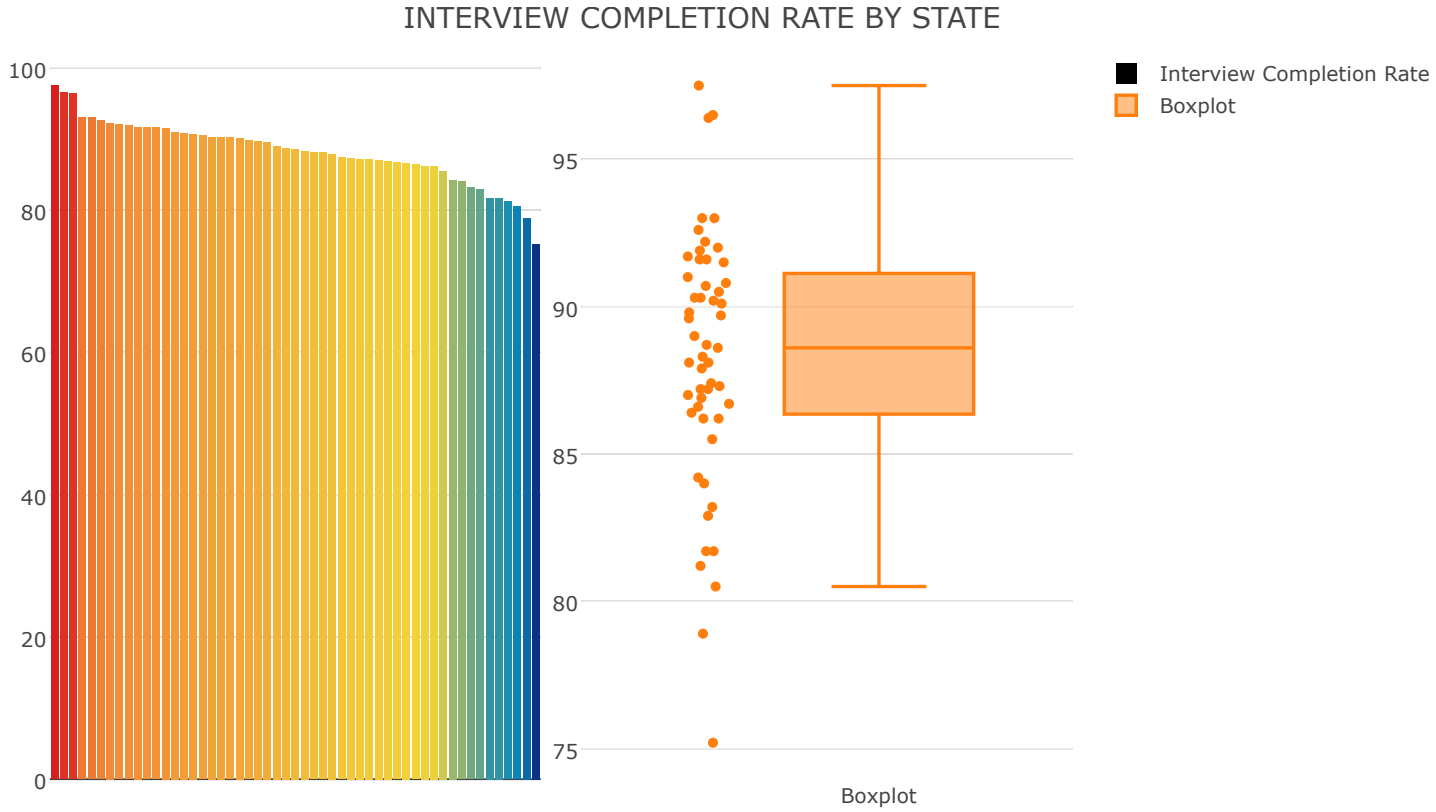
```
disp_totals_df$State <- factor(disp_totals_df$State, levels = unique(disp_totals_df$State)[order(disp_totals_d
f$`Completion Rate`, decreasing = TRUE)])
```

```
f <- plot_ly(disp_totals_df,
        x = ~State,
        y = ~`Completion Rate`,
        type = 'bar',
        name = "Interview Completion Rate",
        marker = list(
          color = ~`Completion Rate`,
          colorscale = 'Portland'
        )) %>%
  layout(
    title = 'INTERVIEW COMPLETION RATE BY STATE',
    xaxis = list(
      title = "States",
      showticklabels = FALSE
      ),
    yaxis = list(
      title = "Completion Rate"
      )
  )

ff <- plot_ly(disp_totals_df,
              y = ~`Completion Rate`,
              type = "box",
              boxpoints = "all",
              jitter = 0.3,
              pointpos = -1.8,
              width = 900,
              height = 500,
              hoverinfo = "y",
              name = 'Boxplot')

LASDOS<-subplot(f,ff)

LASDOS
```



Given that this is not a histogram, the skew that is seen in the summary statistics cannot be determined on this bar chart. However, it should be noted that California is an outlier, sitting almost 3 standard deviations away from the mean, at a survey completion rate of 75.2%. On the other hand, Puerto Rico sits at a survey completion rate of 97.1%, which is a full percentage point higher than the next, Tennessee. New York, notably, had the 6th lowest completion rate, 81.7%

Further analysis: When grouped by Time Zone or by region (Northeast, South, Midwest, etc.) are differences observed?

Research quesion 2: Are fruit consumption standard across all states?

Maintaining a healthy lifestyle is sought after by many. Science has backed that this is best achieved in the kitchen - starting with the food one eats. The organic food industry boasts $40+ billion in sales, of which more than 90% were in the food industry. Fruit is often incorporated into healthy lifestyles and can be used as a measure of “healthiness”, but much like all things - too much fruit is not good.

The factors (variables) names listed on the CDC’s website, did not align with the data. As a result, the console was throwing an error, so the grep function was used to search for all of the variables with similar names in the hopes of finding the correct variable name.

```
grep("frut", names(brfss2013), value = TRUE)
```

```
## [1] "ssbfrut2"  "frutda1_"  "X_frutsum"
```

```
grep("pain", names(brfss2013), value = TRUE)
```

```
## [1] "joinpain"  "painact2"  "rlivpain"  "X_paindx1"
```

The observations that were listed as ‘NA’ for missing values were removed from the data frame.

```
NROW(is.na(brfss2013$X_state))
```

```
## [1] 491775
```

```
NROW(is.na(brfss2013$frutda1_))
```

```
## [1] 491775
```

The fruit factor, frutda1_, corresponds to the fruit intake in 1 day. The head of the data frame for that will be used for this question is shown below.

```
question2 <- data.frame(brfss2013$X_state,brfss2013$frutda1_)
question2 <- question2[complete.cases(question2), ] #gets rid of NAs
colnames(question2) <- c("State","Computed Fruit Consumed")
head(question2, 5)
```

```
##     State Computed Fruit Consumed
## 1 Alabama                     400
## 2 Alabama                       3
## 3 Alabama                      43
## 4 Alabama                      20
## 5 Alabama                       7
```

The total fruit consumption in 1 day by each state is computed and the head of this data frame is shown below.

```
fruitdf<-question2%>%
  group_by(State)%>%
  summarise(`Computed Fruit Consumed Mean` = round(mean(`Computed Fruit Consumed`),digits = 1))

head(fruitdf,5)
```

```
## # A tibble: 5 x 2
##        State `Computed Fruit Consumed Mean`
##       <fctr>                          <dbl>
## 1    Alabama                           79.3
## 2     Alaska                          108.6
## 3    Arizona                          103.9
## 4   Arkansas                           88.5
## 5 California                          129.2
```

The mean, median, IQR, and standard deviation is computed for the total fruit consumption across all states.

```
fruit_mean <- round(mean(fruitdf$`Computed Fruit Consumed Mean`), digits = 1)
fruit_median <- round(median(fruitdf$`Computed Fruit Consumed Mean`), digits = 1)
fruit_IQR <- round(IQR(fruitdf$`Computed Fruit Consumed Mean`), digits = 1)
fruit_SD <- round(sd(fruitdf$`Computed Fruit Consumed Mean`), digits = 1)
fruit_range <- max(fruitdf$`Computed Fruit Consumed Mean`) - min(fruitdf$`Computed Fruit Consumed Mean`)

cat("Mean:", prettyNum(fruit_mean, big.mark=",",scientific=FALSE), "\n") #prettyNum() adds the comma separator
to the numbers
```

```
## Mean: 101.3
```

```
cat("Median:", prettyNum(fruit_median, big.mark=",",scientific=FALSE), "\n")
```

```
## Median: 104.2
```

```
cat("IQR:", prettyNum(fruit_IQR, big.mark=",",scientific=FALSE), "\n")
```

```
## IQR: 16.6
```

```
cat("Std. Dev:", prettyNum(fruit_SD, big.mark=",",scientific=FALSE), "\n")
```

```
## Std. Dev: 13.2
```

```
cat("Range:", prettyNum(fruit_range, big.mark=",",scientific=FALSE))
```

```
## Range: 69.2
```

Given that the median is higher than the mean, this suggests that the data is skewed to the left.

Similar to the first question, in order to plot in decreasing order, the state names had to be reordered.

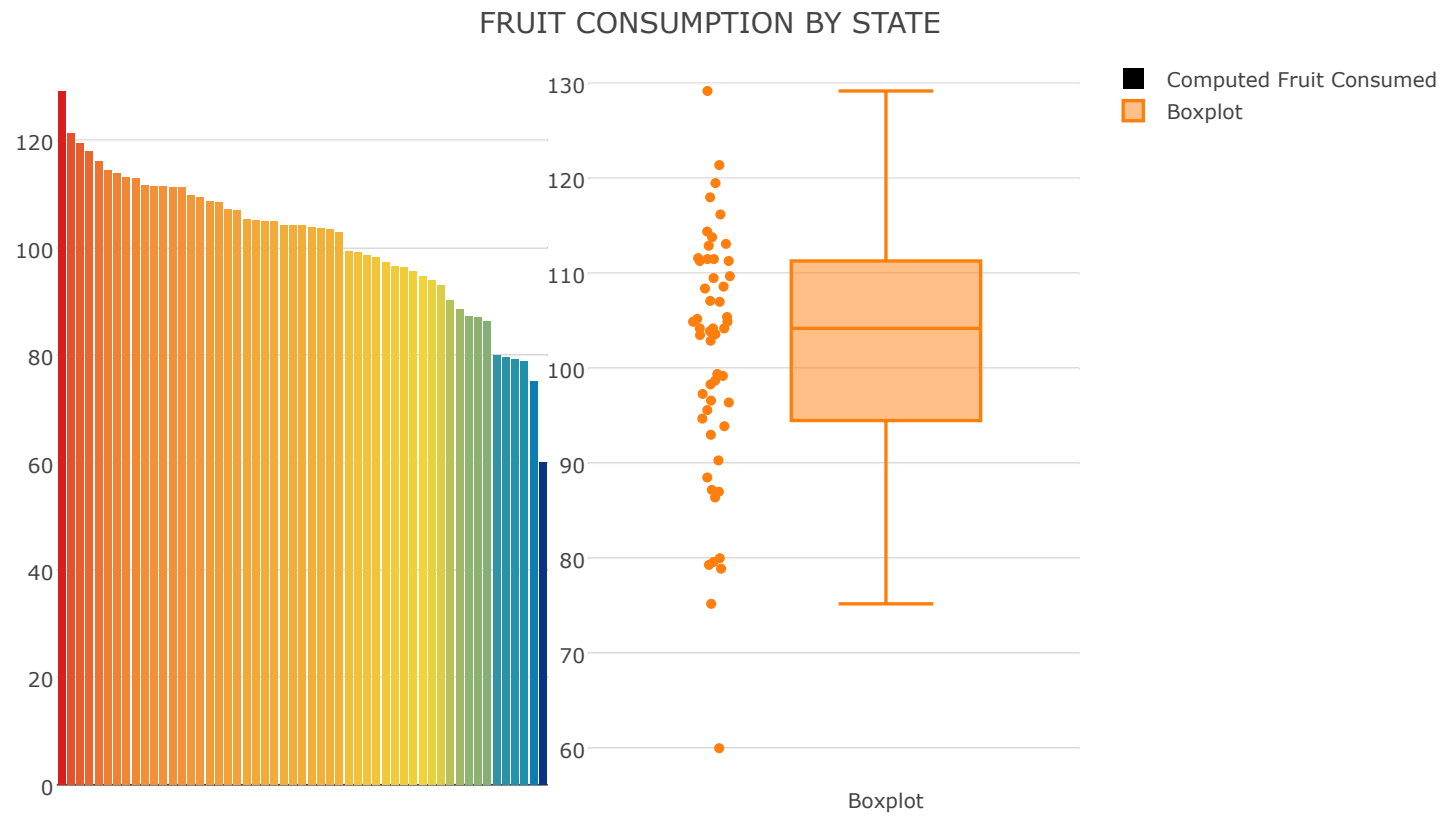
```
fruitdf$State <- factor(fruitdf$State, levels = unique(fruitdf$State)[order(fruitdf$`Computed Fruit Consumed M
ean`, decreasing = TRUE)])
```

```
g <-plot_ly(fruitdf,
        x = ~State,
        y = ~`Computed Fruit Consumed Mean`,
        type = 'bar',
        name = "Computed Fruit Consumed",
        marker = list(
          color = ~`Computed Fruit Consumed Mean`,
          colorscale = 'Portland'
        ),
         width = 1000,
         height = 500) %>%
  layout(
    title = 'FRUIT CONSUMPTION BY STATE',
    xaxis = list(
      title = "States",
      showticklabels = FALSE
      ),
    yaxis = list(
      title = "Total Fruit Consumption"
      )
  )

gg <- plot_ly(fruitdf,
              y = ~`Computed Fruit Consumed Mean`,
              type = "box",
              boxpoints = "all",
              jitter = 0.3,
              pointpos = -1.8,
              width = 900,
              height = 500,
              hoverinfo = "y",
              name = 'Boxplot')

BOTHPLOTS<-subplot(g,gg)

BOTHPLOTS
```



In the first run of the analysis, the total fruit consumed by state was calculated. However in the plot produced, Florida was an obvious outlier and sat at over 3,000,000 fruits consumed in 1 day, which upon further thought - total fruit consumed by state seemed an inaccurate measure. This is especially so because the number of people interviewed in each state varied greatly, and thus the amount of fruit consumed varied greatly. In the second iteration, the mean of the computed fruit consumed was calculated. This has been deemed a better indicator and has a lot less bias than the previous iteration. The second iteration is plotted above. California (129.2) and Puerto Rico (60.0) are notable outliers on differing ends of the spectrum. California is ~2 standard deviations above the mean, while Puerto Rico is ~3 standard deviations below the mean. Puerto Rico's large variance is a contributing factor to the skew of the data. One observation that is surprising is that Hawaii is in the top 50% with a mean of 105.2 computed fruits consumed per day. Fruit, and food in general, is more expensive in Hawaii given that almost everythng has to be flown in to the state. It is surprising that even with high prices of fruit products, Hawaii sits surprisingly high on the spectrum.

Future Analysis: Compute total fruit consumed per capita and per person interviewed.

It should be noted that it has not been determined how the CDC computed the "computed fruit consumed" factor.

Research quesion 3: Are vetarans more disposed to alcohol and tobacco use?

It has long been noted that a signficant number of veterans suffer from PTSD after service, but how does the prevalence of drinking alcohol compare with non-veterans?

The head of the data frame used to answer this question is below.

```
question3<-data.frame(brfss2013$veteran3,brfss2013$smoke100,brfss2013$drnkany5) #makes df for Research questio
n #3 pt2
question3<-question3[complete.cases(question3), ] #gets rid of NAs
colnames(question3) <- c("Veteran Status","Smoke Status", "Drink Status")
head(question3, 5)
```

```
##   Veteran Status Smoke Status Drink Status
## 1             No          Yes           No
## 2             No           No          Yes
## 3             No          Yes           No
## 4             No           No           No
## 5             No          Yes           No
```

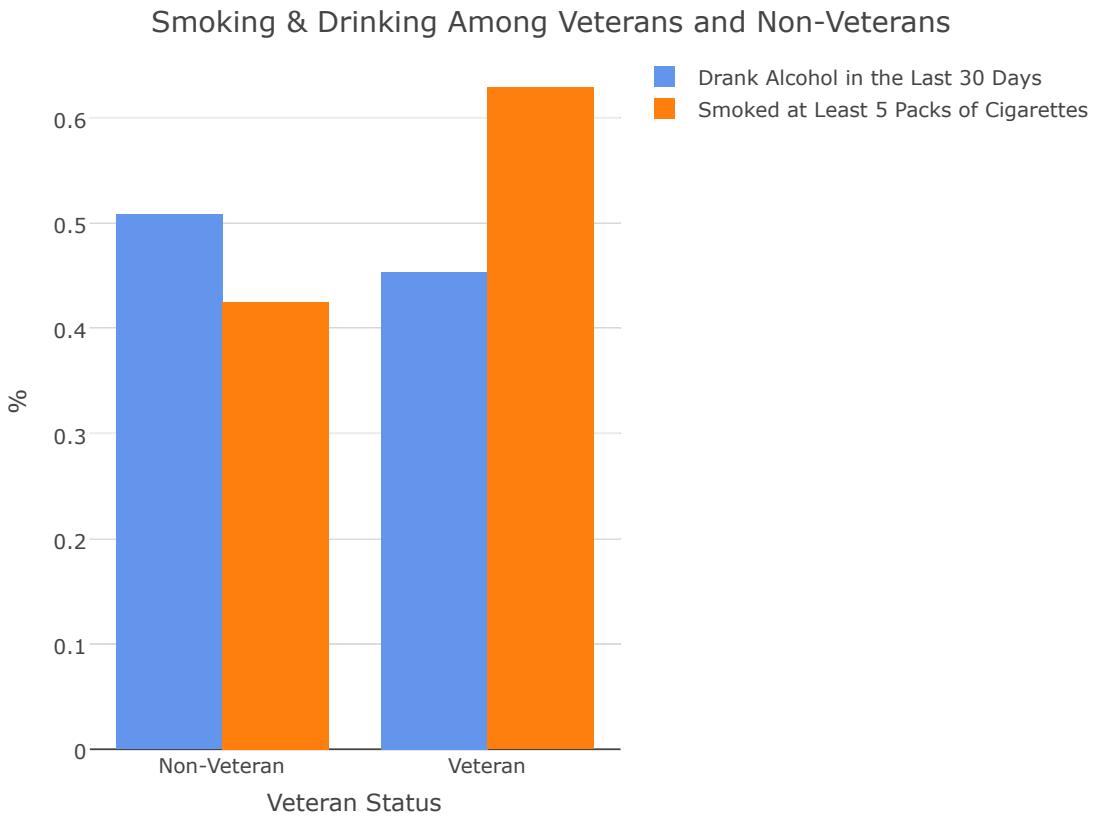
The above code creates the data frame that will be used to make the tibble, a variation of the standard data frame, smokedrinkTIB. The tibble will be summarized and the percentage of smokers and people who have had alcohol in the last 30 days will be calculated based on veteran status. This tibble is below.

```
smokedrinkTIB<-question3%>%
group_by(`Veteran Status`)%>%
summarise(alcperct = sum(`Drink Status` == 'Yes')/n(), smokeperct = sum(`Smoke Status` == 'Yes')/n())

smokedrinkTIB
```

```
## # A tibble: 2 x 3
##   `Veteran Status`  alcperct smokeperct
##             <fctr>     <dbl>      <dbl>
## 1              Yes 0.4535990  0.6285709
## 2               No 0.5080064  0.4246721
```

```
plot_ly(smokedrinkTIB,
        x = c("Veteran", "Non-Veteran"),
        y = ~alcperct,
        type = 'bar',
        name = "Drank Alcohol in the Last 30 Days",
        marker = list(
          color = 'cornflowerblue')
        ) %>%
  add_trace(y = ~smokeperct,
            name = 'Smoked at Least 5 Packs of Cigarettes',
            marker = list(
          color = 'deeppink2')) %>%
  layout(
    title = 'Smoking & Drinking Among Veterans and Non-Veterans',
    xaxis = list(
      title = "Veteran Status",
      labels=c("Non-Veteran", "Veteran")
      ),
    yaxis = list(
      title = "%"
      ),
    barmode = 'group'
  )
```



The graphs show that there is no difference of significance between veterans and non-veterans based on drinking habits in the last 30 days. However, there is a more stark contrast in smoking habits, as is expected. Given the generalizability of the study, these observations can be generalized, but no causal relationship can be determined.

*Source: https://www.ota.com/news/press-releases/19031 (https://www.ota.com/news/press-releases/19031)