Data 607: Assignment One

Zachary Safir

Introduction

For the first homework assignment, we were given a simple task. Find an interesting article and dataset from the following website: https://data.fivethirtyeight.com. I chose this article https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/, which contained a story about which Halloween candy people liked the most.

Loading the Data

```
library(tidyverse)
library(knitr)
github <- "https://raw.githubusercontent.com/"
repo <- "fivethirtyeight/data/master/candy-power-ranking/"
data <- "candy-data.csv"

candy <- paste0(github, repo, data) %>% read.csv()
```

Checking for Missing Data

The first thing to check is for any missing values. As shown in the table below, this data is NA value free.

```
kable(colSums(is.na(candy)),format = "pandoc",col.names  =c("NA Values"),
caption = "Number of NA Values")
```

Table 1: Number of NA Values

	NA Values
competitorname	0
chocolate	0
fruity	0
caramel	0
peanutyalmondy	0
nougat	0
crispedricewafer	0
hard	0
bar	0
pluribus	0
sugarpercent	0
pricepercent	0
winpercent	0

Adjusting Column Names

When looking through the data, I noticed something off about the column names. Every name was put in lowercase, even if it contained two separate words. For example, 'competitorname'. While this does not get in the way of analyzing the data, it makes it unnecessarily harder to read. It makes much more sense to use either underscores or capitalization to separate words. So, the first thing I did was change the two-name columns in the latter mentioned way.

```
candy <- candy %>% rename(CompetitorName = competitorname, PeanutAlmond = peanutyalmondy,
    CrispedRiceWafer = crispedricewafer, OnePiece = pluribus, SugarPercent = sugarpercent,
    PricePercent = pricepercent, WinPercent = winpercent)
```

Subset One of the data

Seen below is code used to create a subset of the data that focuses soley on how the price affects the overall win percent.

```
priceData <- candy %>% select(CompetitorName, PricePercent, WinPercent)
```

Subset Two of the data

Then I created a second subset of the data which contained information about the type of candy each are.

```
typeOfCandy <- candy %>% select(CompetitorName, chocolate, fruity, caramel, PeanutAlmond,
    nougat, WinPercent, CrispedRiceWafer, hard, bar, OnePiece)
```

Conclusion

In summary, I chose an interesting article, downloaded the data they used, and played around with the variables. The findings in the original article covered the most interesting parts of this data set.