Homework 2

Steven Barnett

9/10/2020

Problem 3

a) Sensory Data

First, I am going to pull down the dataset and save it to local storage

Remote data: http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat

```r
## sensory_data_url <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
## sensory_data <- fread(sensory_data_url, skip = 1, fill = TRUE, data.table = FALSE)
## saveRDS(sensory_data, "dwnldd_data/sensory_data_raw.RDS")
sensory_data <- readRDS("dwnldd_data/sensory_data_raw.RDS")
```

The dataset has some incorrect row lengths due to indices being included in the data. We will fix these using Base R functions.

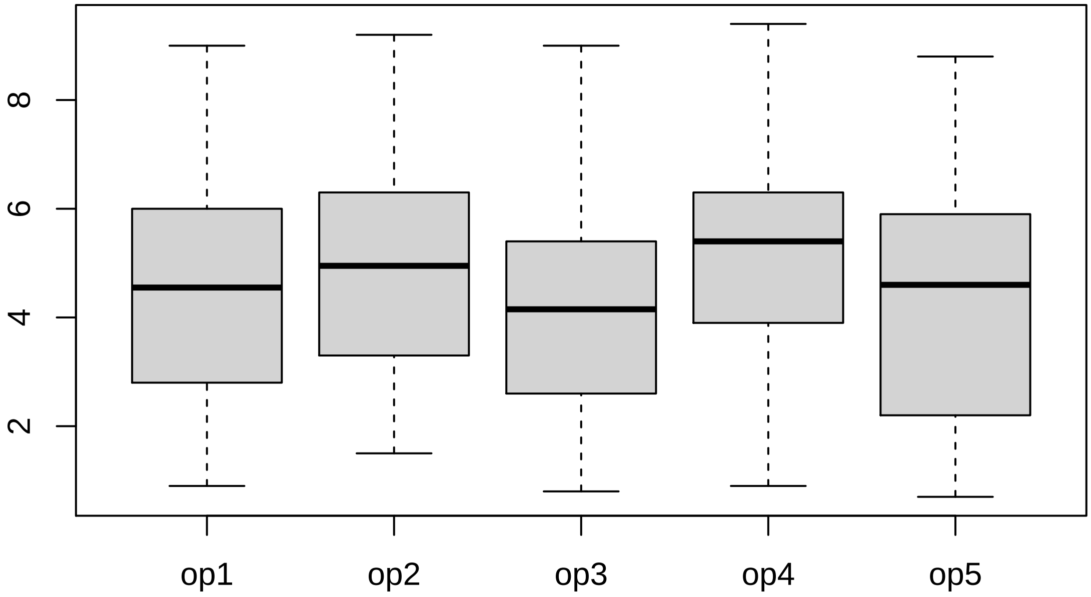
```r
sensory_data_base_r <- sensory_data
for (i in seq(from = 1, to = 30, by = 3)) {
  sensory_data_base_r[i,] = sensory_data_base_r[i,][2:6]
}
sensory_data_base_r <- subset(sensory_data_base_r, select = -c(6))
names(sensory_data_base_r) <- c("op1", "op2", "op3", "op4", "op5")
```

We will attempt to clean the dataset using tidyverse functions

```r
#sensory_data_tidyverse <- sensory_data
#sensory_data_tidyverse <- select(sensory_data_tidyverse, -6)
#rename(sensory_data_tidyverse, op1 = Item, op2 = 1, op3 = 2, op4 = 3, op5 = 4)
```

After cleaning the dataset using Base R and tidyverse functions, I am going to display the cleaned data

op1	op2	op3	op4	op5
Min. :0.900	Min. :1.500	Min. :0.800	Min. :0.900	Min. :0.700
1st Qu.:2.850	1st Qu.:3.450	1st Qu.:2.650	1st Qu.:3.925	1st Qu.:2.250
Median :4.550	Median :4.950	Median :4.150	Median :5.400	Median :4.600
Mean :4.593	Mean :5.063	Mean :4.167	Mean :5.193	Mean :4.267
3rd Qu.:5.950	3rd Qu.:6.225	3rd Qu.:5.400	3rd Qu.:6.275	3rd Qu.:5.800
Max. :9.000	Max. :9.200	Max. :9.000	Max. :9.400	Max. :8.800



b) Gold Medal Data

First, I will pull the data down and store it locally

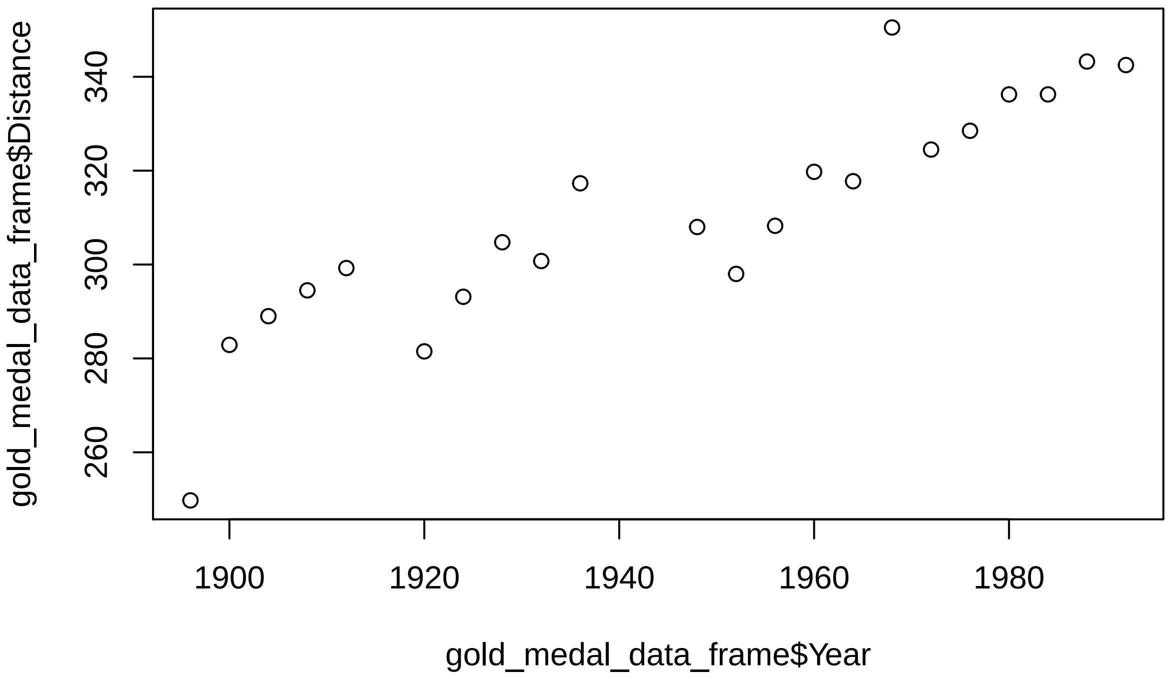
```r
## gold_medal_data_url <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
## gold_medal_data <- fread(gold_medal_data_url, fill=TRUE)
## saveRDS(gold_medal_data, "dwnldd_data/gold_medal_data_raw.RDS")
gold_medal_data <- readRDS("dwnldd_data/gold_medal_data_raw.RDS")
```

The dataset has some incorrect row lengths due to indices being included in the data. We will fix these using Base R functions.

```r
gold_medal_data_base_r <- gold_medal_data
gold_medal_data_frame <- data.frame(Year=integer(), Distance=numeric())
for (i in seq(from = 1, to = 11, by = 2)) {
  for (j in seq(from = 1, to = 6, by = 1)) {
    data_row <- gold_medal_data_base_r[j,]
    year <- data_row[[i]]
    distance <- data_row[[i+1]]
    gold_medal_data_frame <- rbind(gold_medal_data_frame, c(year, distance))
  }
}
names(gold_medal_data_frame) <- c("Year", "Distance")
gold_medal_data_frame <- na.omit(gold_medal_data_frame)
gold_medal_data_frame$Year <- gold_medal_data_frame$Year + 1900
```

After cleaning the dataset using Base R and tidyverse functions, I am going to display the cleaned data

Year	Distance
Min. :1896	Min. :249.8
1st Qu.:1921	1st Qu.:295.4
Median :1950	Median :308.1
Mean :1945	Mean :310.3
3rd Qu.:1971	3rd Qu.:327.5
Max. :1992	Max. :350.5


gold_medal_data_frame$Distance
gold_medal_data_frame$Year