R coding for public policy

Assignment 5

[Name removed]

Assignment Instruction:

Once you download and open Assignment5.rmd in R studio,

Please complete all the problems in the empty line between ”‘{r} and’” You can add more empty lines by press Enter in the empty line You can also click the green arrow next to each code chunk to check your code Please include only the relevant codes in the chunks Once you complete all the problems, click “Knit”

Submit the knitted word document and your markdown as attachments to NYU Classes - Assignment - Assignment 5

In assignment 5, we will use the following datasets:

1. The SAT dataset from assignment 2 and lecture 3 link: https://data.cityofnewyork.us/Education/2012-SAT-Results/f9bf-2cp4 *make sure you find the correct url for the csv file

The Baby names dataset from lecture 3 link: https://health.data.ny.gov/Health/Baby-Names-Beginning-2007/jxy9-yhdk *make sure you find the correct url for the csv file

#import and clean the SAT dataset and the baby dataset, remove the missing rows, correct the column classes and rename the columns for your convenience. Create a new variable for the overall SAT score

```r
#SAT
SAT<-read.csv("https://data.cityofnewyork.us/api/views/f9bf-2cp4/rows.csv?acc
essType=DOWNLOAD", stringsAsFactors = FALSE)

SAT[, 3:6]<-apply(SAT[, 3:6], 2, as.numeric)

## Warning in apply(SAT[, 3:6], 2, as.numeric): NAs introduced by coercion

## Warning in apply(SAT[, 3:6], 2, as.numeric): NAs introduced by coercion

## Warning in apply(SAT[, 3:6], 2, as.numeric): NAs introduced by coercion

## Warning in apply(SAT[, 3:6], 2, as.numeric): NAs introduced by coercion

names(SAT) <- c("DBN", "School.Name", "N.test.takers", "Reading.Score", "Math
.Score", "Writing.Score")
```

```
TotalSAT.Score <-as.numeric(SAT$Reading.Score+SAT$Math.Score+SAT$Writing.Scor
e)


SAT<-cbind(SAT,TotalSAT.Score)
SAT<-na.omit(SAT)

#Baby - double check w/ class notes
baby<-read.csv("https://health.data.ny.gov/api/views/jxy9-yhdk/rows.csv?acces
sType=DOWNLOAD")


a<-sapply(baby, is.numeric)
a

##       Year First.Name     County        Sex      Count
##       TRUE      FALSE      FALSE      FALSE       TRUE

baby[, !a]<-lapply(baby[, !a], toupper)
baby<-na.omit(baby)
```

#Problem 1 [Baby] use filter() a new baby sample, babya, for “QUEENS” and “BRONX” in 2014.

```
babya<- baby %>%
  filter(Year == 2014, County == "QUEENS" | County == "BRONX")
```

#Problem 2 [Baby] use select() to create a baby sample, babyb, with the following columns: Sex, Year, Count.

```
babyb<-select(baby, Sex, Year, Count)
```

#Problem 3 [Baby] use filter() and select() to create a baby sample, babyc, for “QUEENS” and “BRONX” counties in 2014, with the following columns: Sex, Year, Count. You are encouraged to use %>%.

```
babyc<-baby %>%
  filter(Year == 2014, County == "QUEENS" | County == "BRONX") %>%
  select(Sex, Year, Count)
```

#Problem 4 [Baby] use table() and prop.table() to create a frequency table and the column percent table for variable Year (rows) and Sex(columns) in babyc

```
table(babyc$Year,babyc$Sex)

##
##          F   M
##   2014 931 816

prop.table(table(babyc$Year,babyc$Sex),1)
```

```
##
##                 F         M
##   2014 0.5329136 0.4670864
```

#Problem 5 [SAT] use quantile() to find the top 22% of SAT overall score, then use ifelse() to create a binary variable top22: 1 - top 22% of the overall SAT score and 0 - other;

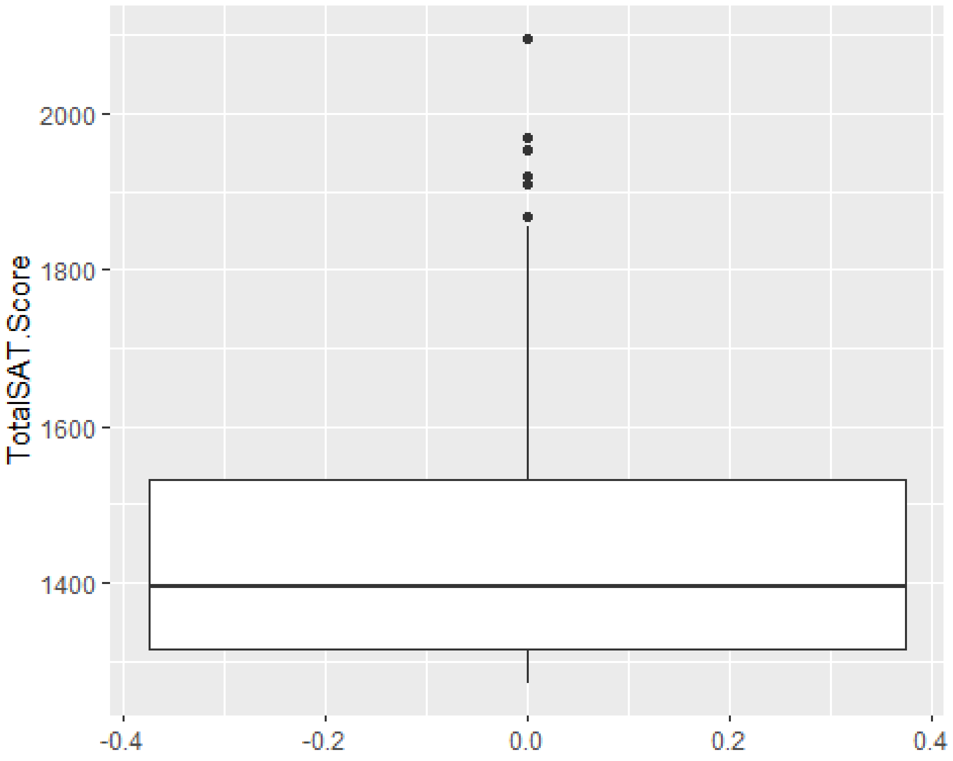
```
quantile(TotalSAT.Score, prob=0.78, na.rm=TRUE)

##    78%
## 1271.6

top22<-ifelse(TotalSAT.Score>=1271.6,1,0)
top22<-na.omit(top22)
```

#Problem 6 [SAT] use ggplot() and geom_boxplot() to create a boxplot of overall SAT scores by top22

```
top22df<-subset(SAT, subset = (TotalSAT.Score>=1271.6))
ggplot(top22df, aes(y=TotalSAT.Score)) + geom_boxplot()
```



#Problem 7 [SAT] *bonus* research group_by(), summarise() and mutate () from dplyr, you can google on your own or refer to chapter 5.6 of R for Data science - link: http://r4ds.had.co.nz/

use dplyr to

### 1. summarize the mean, median and variance for SAT overall score by top 30 ### 2. create a boxplot of overall SAT scores by top30.

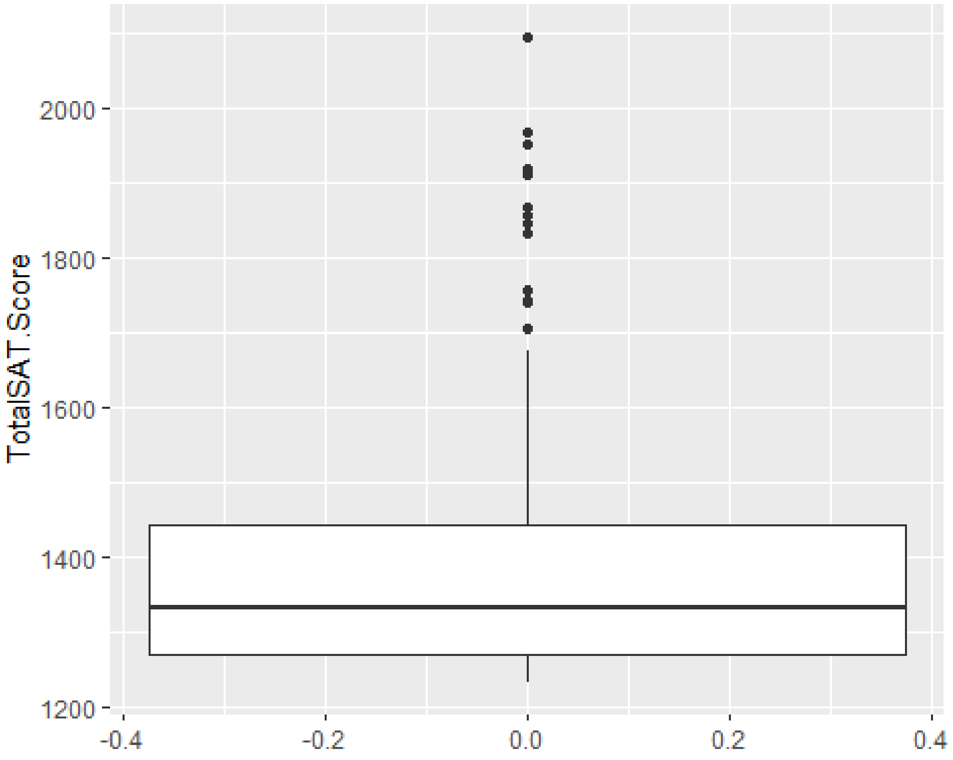
###You are encouraged to use %>%

```
#Quantile Result 1233 for top 30
SAT %>%
  filter(TotalSAT.Score>=1233)%>%
  summarize(TOT_mean_top33=mean(TotalSAT.Score),
            TOT_median_top30=median(TotalSAT.Score),
            TOT_var_top30=var(TotalSAT.Score)
            )
```

```
##   TOT_mean_top33 TOT_median_top30 TOT_var_top30
## 1       1405.118             1333      35789.25
```

```
top30<-SAT %>%
  filter(TotalSAT.Score>=1233)


ggplot(top30, aes(y=TotalSAT.Score)) + geom_boxplot()
```


TotalSAT.Score