# R coding for public policy

## Assignment 4

[Name removed]

## Assignment Instruction:

Once you download and open Assignment4.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 4

## In assignment 4, 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

#import and clean the SAT 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

```
SAT<-read.csv("https://data.cityofnewyork.us/api/views/f9bf-2cp4/rows.csv?accessType=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.Score)
```

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

#Problem 1 [SAT] use quantile() to calculate the 88th percentile overall SAT score?

```
quantile(TotalSAT.Score, prob=0.88, na.rm=TRUE)
```

```
##  88%
## 1374
```

#Problem 2 [SAT] use mean() to calculate the percentage of schools with a reading score below 453?

```
mean(SAT$Reading.Score<453)
```

```
## [1] 0.8693587
```

#Problem 3 [SAT] use ifelse() to create a binary variable top12: 1 - top 12% of the overall SAT score and 0 - other;

```
top12<-ifelse(TotalSAT.Score>=1374,1,0)
top12<-na.omit(top12)
```

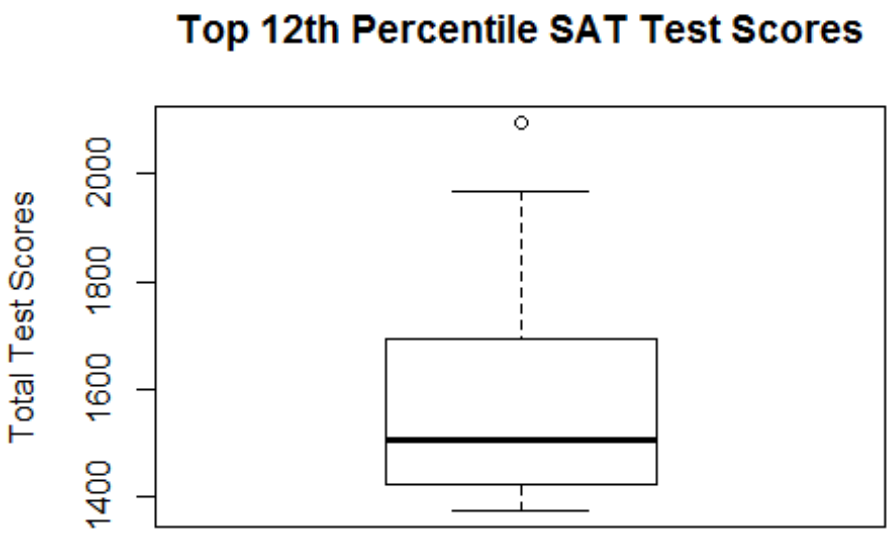
#Problem 4 [SAT] use summary() to check the distributions of overall SAT score by top12

```
top12df<-subset(SAT, subset = (TotalSAT.Score>=1374))
bottom88df<-subset(SAT, subset = (TotalSAT.Score<1374))
summary(top12df)
```

```
##      DBN            School.Name        N.test.takers    Reading.Score
##  Length:52          Length:52          Min.   :  30.0   Min.   :412.0
##  Class :character   Class :character   1st Qu.:  92.0   1st Qu.:468.0
##  Mode  :character   Mode  :character   Median : 139.5   Median :497.0
##                                        Mean   : 286.6   Mean   :515.6
##                                        3rd Qu.: 302.2   3rd Qu.:550.2
##                                        Max.   :1277.0   Max.   :679.0
##    Math.Score    Writing.Score   TotalSAT.Score
##  Min.   :440.0   Min.   :431.0   Min.   :1374
##  1st Qu.:489.8   1st Qu.:466.8   1st Qu.:1424
##  Median :523.0   Median :492.5   Median :1504
##  Mean   :543.3   Mean   :516.2   Mean   :1575
##  3rd Qu.:576.5   3rd Qu.:555.0   3rd Qu.:1684
##  Max.   :735.0   Max.   :682.0   Max.   :2096
```

#Problem 5 [SAT] use boxplot() to visualize the distributions of overall SAT scores by top12

```
boxplot(x=top12df$TotalSAT.Score, main="Top 12th Percentile SAT Test Scores",
ylab="Total Test Scores")
```




#Problem 6 [SAT] Use the binary variable created in Problem 3 as the outcome, run a simple logistic regression to model the relationship between top12 (Y) and the average math score (X). name the model mlr, get the summary for mlr

```
top12df<-subset(SAT, subset = (TotalSAT.Score>=1374))


mlr<-glm(top12 ~ SAT$Math.Score, data=SAT, family=binomial)
summary(mlr)

##
## Call:
## glm(formula = top12 ~ SAT$Math.Score, family = binomial, data = SAT)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.15940  -0.11233  -0.03655  -0.01688   2.55913
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -34.14914    5.30778  -6.434 1.24e-10 ***
## SAT$Math.Score   0.07026    0.01117   6.290 3.16e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

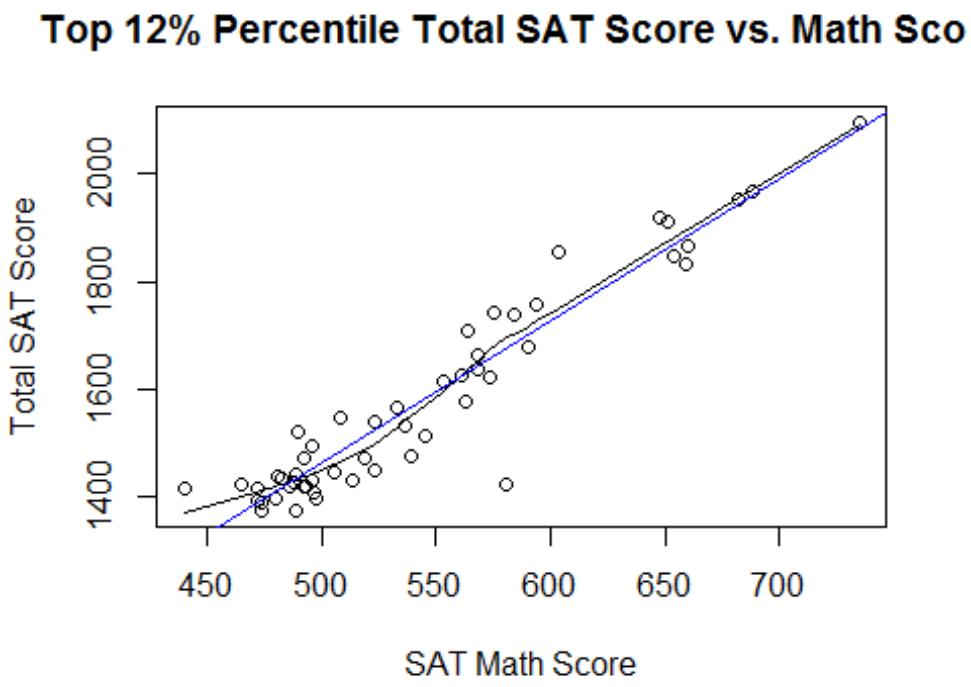
```
##     Null deviance: 314.800  on 420  degrees of freedom
## Residual deviance:  80.682  on 419  degrees of freedom
## AIC: 84.682
##
## Number of Fisher Scoring iterations: 8
```

#Problem 7 [SAT] *bonus* use lm() to run a simple linear regression testing the relationship between the overall SAT score (Y) and the SAT Math Score (X) among the top 12% school. Name the model lm1, run a summary of lm1. ## Then use plot(), lines() and abline() to visualize the relationship between the overall SAT score and the SAT math score for the top 12%, give the linear trend line (abline) a different color

```
lm1<-lm(top12df$TotalSAT.Score ~ top12df$Math.Score)
summary(lm1)
```

```
##
## Call:
## lm(formula = top12df$TotalSAT.Score ~ top12df$Math.Score)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -250.776  -27.427    0.811   29.744  120.402
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          138.362     67.343   2.055   0.0452 *
## top12df$Math.Score     2.644      0.123  21.499   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.18 on 50 degrees of freedom
## Multiple R-squared:  0.9024, Adjusted R-squared:  0.9004
## F-statistic: 462.2 on 1 and 50 DF,  p-value: < 2.2e-16
```

```
plot(top12df$Math.Score,top12df$TotalSAT.Score, main="Top 12% Percentile Tota
l SAT Score vs. Math Score", ylab="Total SAT Score", xlab="SAT Math Score")
lines(lowess(top12df$Math.Score,top12df$TotalSAT.Score))
abline(lm(top12df$TotalSAT.Score ~ top12df$Math.Score), col="blue")
```


Top 12% Percentile Total SAT Score vs. Math Sco
Total SAT Score
1400
1600
1800
2000
450
500
550
600
650
700
SAT Math Score