COLLEGE SCORE CARD (USA UNIVERSITIES DATA FOR 2013 - 2014)

ISAAC OWUSU AKOWUAH

January 1, 2018

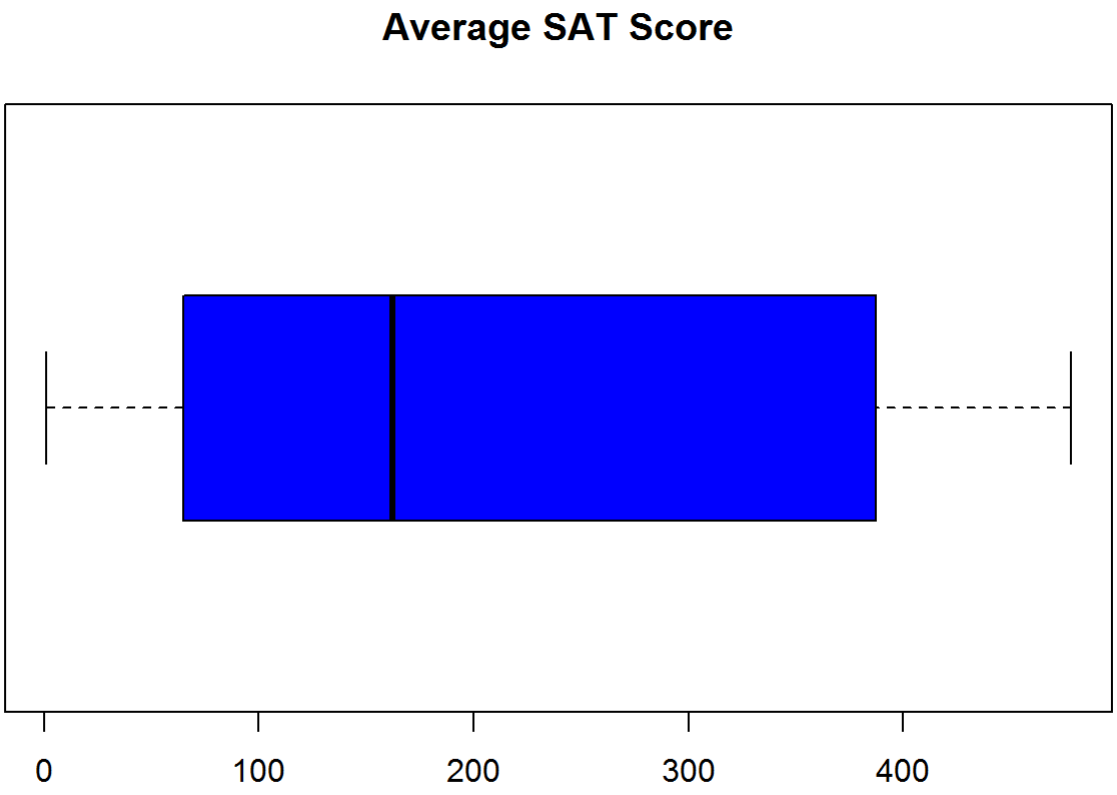
College Score Card holds data from both private and public universities in United States of America, which is updated every year. This data helps to provide information to students and guardians seeking higher education, public, government, policy makers and stakeholders concerning tertiary education. In this project, we ask questions and then answer with this available data. \

1. What is the average SAT score for universities in 2013 - 2014 academic year?

```
##Importing data into R
MERGED2013_14_PP <- read.delim("H:/github files/CollegeScorecard_Raw_Data/MERGED2013_14_PP.csv")
##Subsetting
new<-subset(MERGED2013_14_PP,MERGED2013_14_PP$SAT_AVG!='NULL')
##Median
median(as.numeric(new$SAT_AVG))
```

```
## [1] 162
```

```
##Box and Whisker Plot
boxplot(as.numeric(new$SAT_AVG),horizontal=TRUE, main="Average SAT Score",col = c("blue"))
```



2. What is the difference in completion rate between students from low and high income family for 4 year college?

```
##Subsetting
diff<-subset(MERGED2013_14_PP,LO_INC_COMP_ORIG_YR4_RT!='PrivacySuppressed' & LO_INC_COMP_ORIG_YR
4_RT!='NULL' & MD_INC_COMP_ORIG_YR4_RT!='PrivacySuppressed' & MD_INC_COMP_ORIG_YR4_RT!='NULL'& H
I_INC_COMP_ORIG_YR4_RT!='PrivacySuppressed'& HI_INC_COMP_ORIG_YR4_RT!='NULL')
##Average Completion for low income students
a<-median(as.numeric(diff$LO_INC_COMP_ORIG_YR4_RT))
a
```

```
## [1] 1487
```

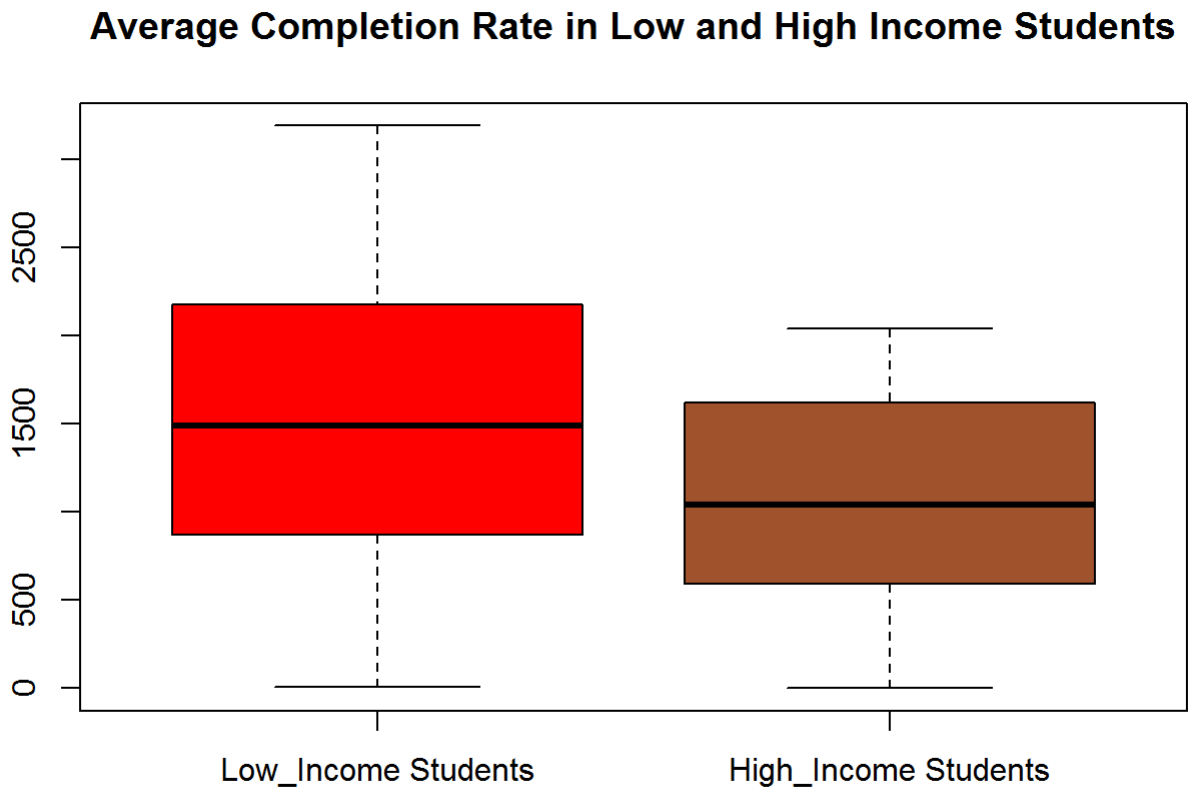
```
##Average Completion for high income students
b<-median(as.numeric(diff$HI_INC_COMP_ORIG_YR4_RT))
b
```

```
## [1] 1041
```

```
m=abs(a-b)
m
```

```
## [1] 446
```

```
boxplot(as.numeric(diff$LO_INC_COMP_ORIG_YR4_RT),as.numeric(diff$HI_INC_COMP_ORIG_YR4_RT), main="Average Completion Rate in Low and High Income Students",names=c("Low_Income Students","High_Income Students"),col = c("red","sienna"))
```



3. What is the completion rate in Females and Males?

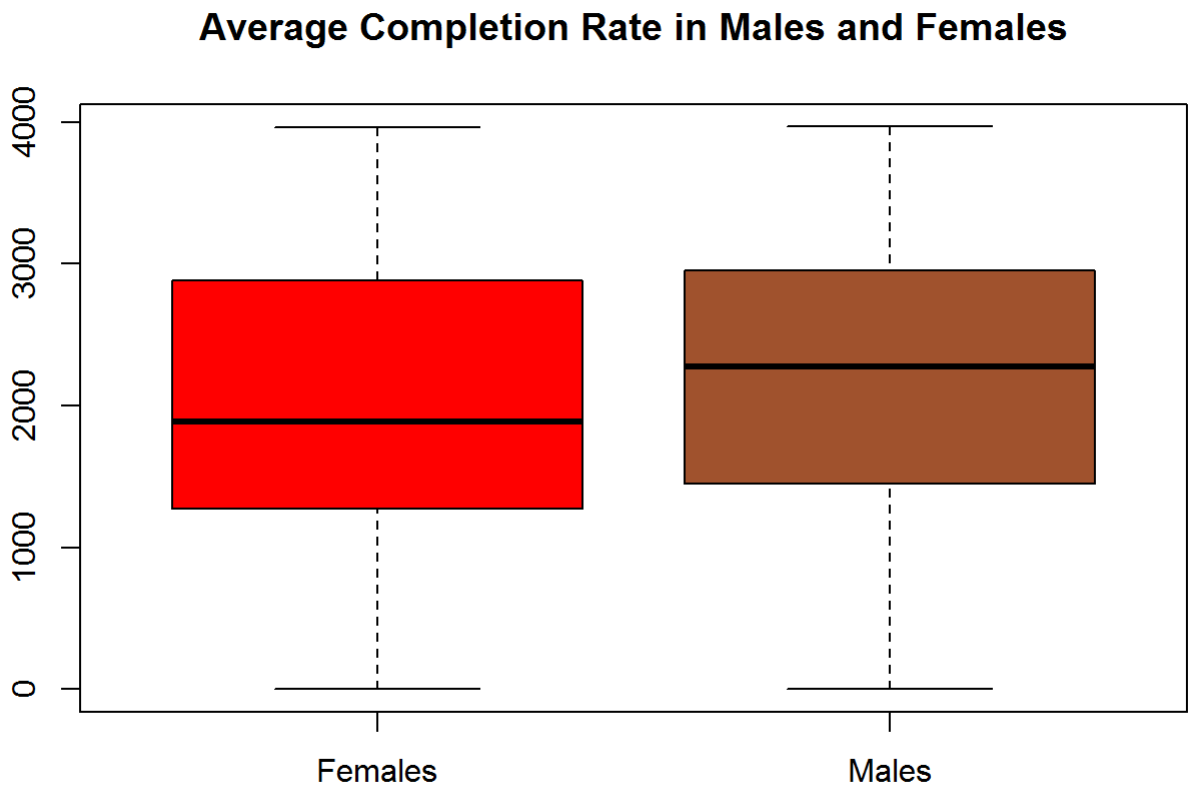
```
##Average in Females
median(as.numeric(MERGED2013_14_PP$UGDS_WOMEN))
```

```
## [1] 2313
```

```
##Average in Males
median(as.numeric(MERGED2013_14_PP$UGDS_MEN))
```

```
## [1] 2052
```

```
boxplot(as.numeric(diff$UGDS_WOMEN),as.numeric(diff$UGDS_MEN), main="Average Completion Rate in Males and Females",names=c("Females","Males"),col = c("red","sienna"))
```



4. How diverse is the race in the Universities?

```
##Subsetting
div<-subset(MERGED2013_14_PP,UGDS_WHITE!='NULL'&UGDS_WHITE!='0'& UGDS_BLACK!='NULL'&UGDS_BLACK!=
'0'& UGDS_HISP!='NULL'&UGDS_HISP!='0'& UGDS_ASIAN!='NULL'& UGDS_ASIAN!='0'& UGDS_AIAN!='NULL'&UG
DS_AIAN!='0'& UGDS_NHPI!='NULL'&UGDS_NHPI!='0'& UGDS_2MOR!='NULL'& UGDS_2MOR!='0'& UGDS_NRA!='NU
LL'&UGDS_NRA!='0'& UGDS_UNKN!='NULL'&UGDS_UNKN!='0')
##Maximum race in a university
div$max<-apply(div[,293:301],1,max)
##Minimum race in a university
div$min<-apply(div[,293:301],1,min)
##Difference in maximum and minimum
div$diff<-as.numeric(div$max)-as.numeric(div$min)
##Maximum difference in race
max(div$diff)
```

```
## [1] 0.9752
```

```
##Row with maximum difference
which(div$diff=='0.9752')
```

```
## [1] 1481
```

```
##The university and city with high dominance of a specific race
div[1481,4:5]
```

```
##                         INSTNM   CITY
## 3768 Laredo Community College Laredo
```

```
##The university high and low dominance race
div[1481,1744:1745]
```

```
##         max    min
## 3768 0.9754 0.0002
```

```
##Race and their value of race dorminance
div[1481,293:301]
```

```
##      UGDS_WHITE UGDS_BLACK UGDS_HISP UGDS_ASIAN UGDS_AIAN UGDS_NHPI
## 3768     0.0137     0.0016    0.9754     0.0016    0.0005    0.0002
##      UGDS_2MOR UGDS_NRA UGDS_UNKN
## 3768    0.0006   0.0009    0.0055
```

```
##Subsetting universities with difference in high and low race greater the 0.5
Divmore<-subset(div,div$diff>=0.50)
##Number of universities in the dataset
nrow(div)
```

```
## [1] 1907
```

```
##Number of universities with race difference greater than 0.5
nrow(Divmore)
```

```
## [1] 1410
```

```
##Ratio of universities out of whole universities with difference greater than 0.5
nrow(Divmore)/nrow(div)
```

```
## [1] 0.7393812
```

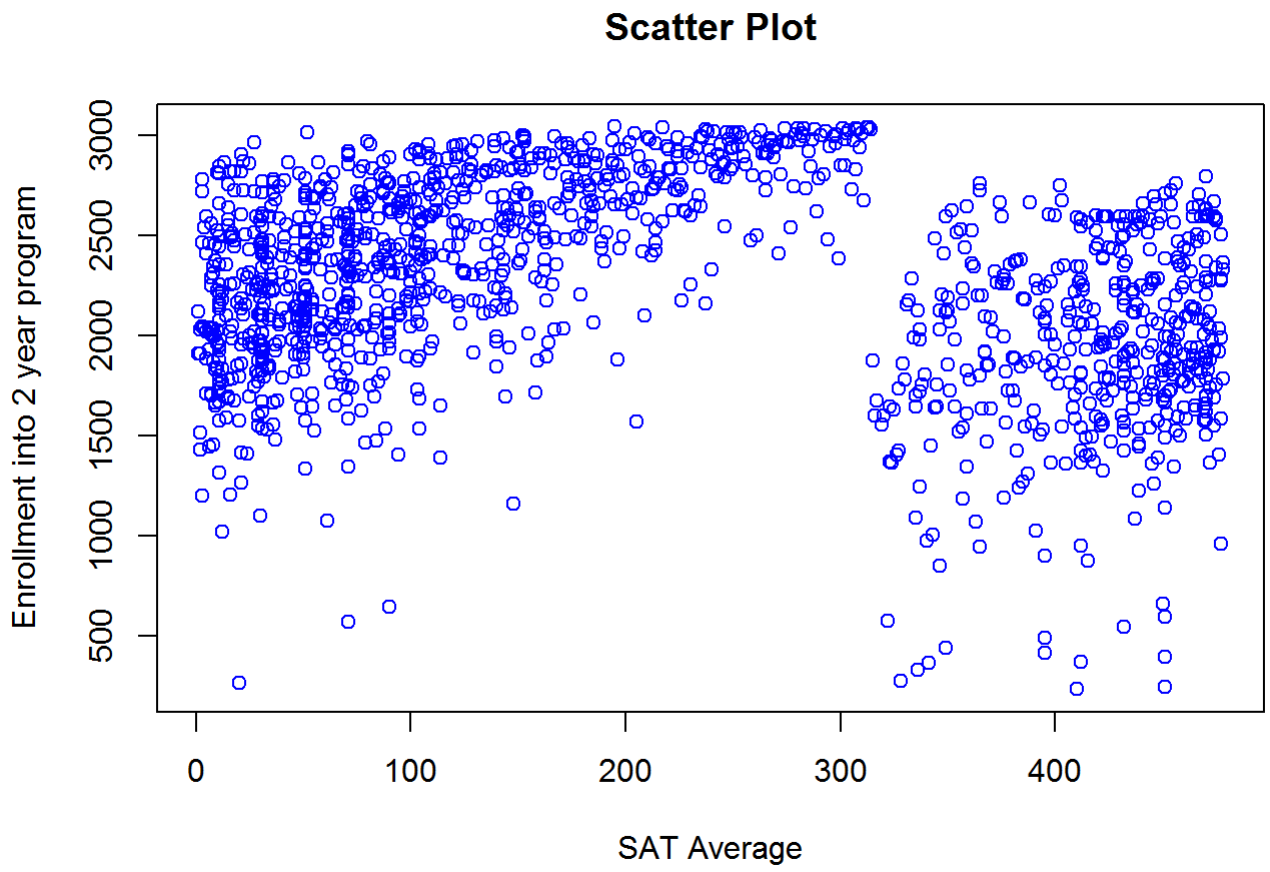
It is noted Laredo Community College in the city of Laredo has much diversity (More of one specific race). Laredo Community College has more Hispanic of a fraction of 0.9754 and very less of Native Hawaiian/Pacific Islander, a fraction of 0.0002.It is also revealed about 74 percent of the universities have over 50 percent difference in the minimum and maximum race of students. This means many universities have a more students of a specific race.

5. What is the Pearson Correlation between Average SAT Score and Enrollment of 2 year university program?

```
##Subsetting
Pcor<-subset(MERGED2013_14_PP,SAT_AVG!='NULL' & ENRL_ORIG_YR2_RT!='PrivacySuppressed'&ENRL_ORIG_
YR2_RT!='NULL')
##Pearson Correlation Test
cor(as.numeric(Pcor$SAT_AVG),as.numeric(Pcor$ENRL_ORIG_YR2_RT),method = "pearson")
```

```
## [1] -0.2185499
```

```
plot(as.numeric(Pcor$SAT_AVG),as.numeric(Pcor$ENRL_ORIG_YR2_RT),main = "Scatter Plot",ylab = "En
rollment into 2 year program",xlab = "SAT Average",col="blue")
```



6. What is the results perfoming a two sample t-test for completion in Low and High Income University?

```
t.test(as.numeric(diff$LO_INC_COMP_ORIG_YR4_RT),as.numeric(diff$HI_INC_COMP_ORIG_YR4_RT), var.eq
ual=TRUE, paired=FALSE)
```

```
##
##  Two Sample t-test
##
## data:  as.numeric(diff$LO_INC_COMP_ORIG_YR4_RT) and as.numeric(diff$HI_INC_COMP_ORIG_YR4_RT)
## t = 28.722, df = 7916, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  430.4045 493.4571
## sample estimates:
## mean of x mean of y
##  1527.573  1065.642
```

```
log10(2.2*exp(-16))
```

```
## [1] -6.606289
```