SAT Scores in NYC: Project Part 1

Description of Data

Background Information

The data used in this project describes the average SAT scores at all acreddited high schools in New York City in the 2014-15 schools year. It was published by New York City Department of Education with data from the College Board, and made available on kaggle.com. (Ref. 1)

Explanation of Data

Each of the 435 rows of the data represents an observation of one high school. The columns of the data used in this project are School Name, Borough (the NYC borough in which the school is located), Percent White, Percent Black, Percent Hispanic, Percent Asian (these 4 columns give the proportion of students at the school who identify as the respective race), and Percent Tested (the proportion of students from the school that took the SAT). An additional column, Average Score (SAT Total) was added to the data for this project. Each of its values is the sum of the 3 average partial scores for each section of the SAT. This project uses only the 374 schools for which all columns of interest were filled in.

Data Selection

This data represents the population of all NYC high school students in the 2014-2015 school year. It was compiled by the New York City Department of Education in conjunction with the College Board, the organization that administers the SAT, with data reported from each schools to the aforementioned organizations.

Potential Issues

Some issues in the data may exist. The removal of 61 schools from the data due to missing score values may create bias for schools that were better prepared to supply SAT data, which

were the most common missing statistics. The data only describes proportions of Whites, Blacks, Hispanics, and Asians, so racial minorities are overgeneralized. Additionally, all percentages are rounded to the nearest hundredth and all score averages are rounded to the nearest integer, which may cause some rounding bias.

Importing and Cleaning Data

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
scores_raw <- read.csv("scores.csv")
scores <- scores_raw[is.na(scores_raw$Average.Score..SAT.Writing.)==FALSE &
                       scores_raw$School.Name!="Forest Hills High School",
                       c("School.Name", "Borough",
                         "Percent.White", "Percent.Black","Percent.Hispanic",
                         "Percent.Asian", "Percent.Tested",
                         "Average.Score..SAT.Total.")] #Ref.2
names(scores) <- c("school", "borough", "white_prop", "black_prop",
                     "hispanic_prop", "asian_prop", "tested_prop", "score")
scores[,"white_prop"] <- as.numeric(sub("%", "", scores[,"white_prop"]))
scores[,"black_prop"] <- as.numeric(sub("%", "", scores[,"black_prop"]))
scores[,"hispanic_prop"] <- as.numeric(sub("%", "", scores[,"hispanic_prop"]))
scores[,"asian_prop"] <- as.numeric(sub("%", "", scores[,"asian_prop"]))
scores[,"tested_prop"] <- as.numeric(sub("%", "", scores[,"tested_prop"]))#Ref.3
```

The above code loads ggplot2, imports the data from scores.csv as scores_raw, then cleans scores_raw into data frame scores by removing rows with missing values, selecting and renaming columns of interest, and converting percentages into useful numeric values.

Numerical Summary

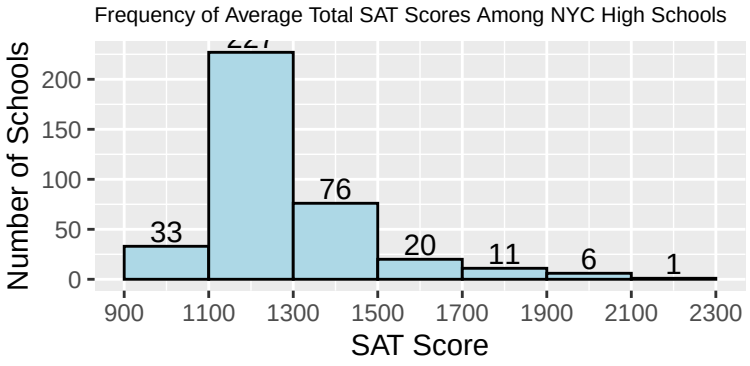
Summary for All Schools:

```
round(summary(scores[,"score"]), 0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     924    1157    1226    1275    1327    2144
```

This summary shows us that, across all schools, there is a mean SAT score of 1275 and a median of 1226. This indicates a slight right skew, as the mean is greater than median, and a moderate spread of the data.

```
# Ref. 5
ggplot(scores, aes(x=score)) +
  geom_histogram(binwidth=200, color="black", fill="lightblue") +
  labs(title="Frequency of Average Total SAT Scores Among NYC High Schools",
       x="SAT Score", y="Number of Schools") +
  stat_bin(aes(label=..count..), geom="text", binwidth=200, vjust=-.2) +
  scale_x_continuous(breaks=seq(900, 2300, 200)) +
  theme(plot.title=element_text(size=8))
```



This histogram confirms the median between 1100 and 1300 and the right skew of the distribution for all schools' average SAT scores. Stuyvesant High School boasts the highest average SAT score at 2144, being the sole resident of its bin.

Summary for Schools By Racial Majority:

```
race_summaries <- round(cbind(c(Count=nrow(scores[scores$white_prop>50,]),
                                Tested=mean(scores[scores$white_prop>50,
                                                   "tested_prop"]),
                                summary(scores[scores$white_prop>50,"score"])),
                              c(Count=nrow(scores[scores$black_prop>50,]),
                                Tested=mean(scores[scores$black_prop>50,
```

```
                                                    "tested_prop"]),
                                      summary(scores[scores$black_prop>50, "score"])),
                                    c(Count=nrow(scores[scores$hispanic_prop>50,]),
                                      Tested=mean(scores[scores$hispanic_prop>50,
                                                    "tested_prop"]),
                                      summary(scores[scores$hispanic_prop>50,
                                                    "score"])),
                                    c(Count=nrow(scores[scores$asian_prop>50,]),
                                      Tested=mean(scores[scores$asian_prop>50,
                                                    "tested_prop"]),
                                      summary(scores[scores$asian_prop>50,
                                                    "score"]))), 0)
colnames(race_summaries) <- c("White", "Black", "Hispanic", "Asian")#Ref. 4
race_summaries
```

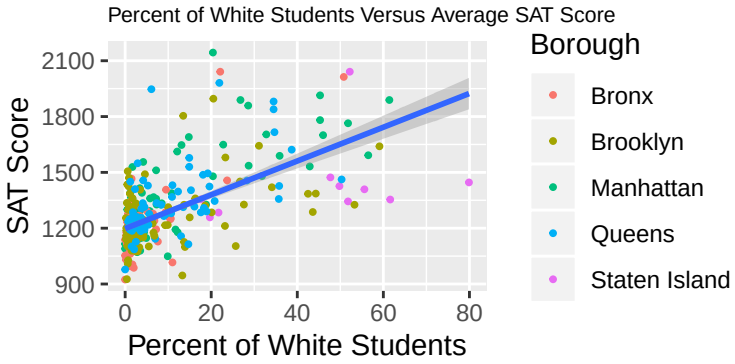
```
##         White Black Hispanic Asian
## Count      12    95      169    10
## Tested     81    61       60    91
## Min.     1326  1009      924  1127
## 1st Qu.  1395  1159     1141  1529
## Median   1527  1192     1203  1772
## Mean     1607  1211     1205  1727
## 3rd Qu.  1795  1247     1276  1972
## Max.     2041  1506     1556  2144
```

This matrix gives the count, tested percentage, and 5-number summaries for schools with students primarily of each race. This metric leaves out the 88 schools (23.5% of total) that have no racial majority, but still allows for significant differences to be seen. The interquartile range (IQR) for average SAT scores for White-majority schools is (1395, 1795), which lies entirely above the IQRs for Black- and Hispanic-majority schools of (1159, 1247) and (1141, 1276), respectively. The 10 Asian-majority schools have yet higher 1st and 3rd quartile values of (1529, 1927).

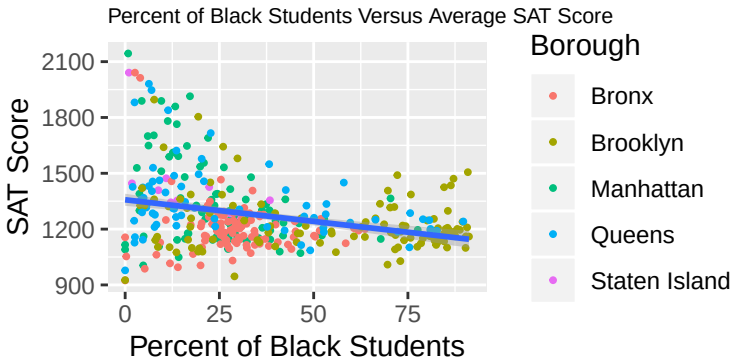
The percentages in the row "Tested" reinforce this pattern of racial inequality by showing that groups of schools with higher average SAT scores also have higher proportions of students taking the SAT, with a notable 20% gap between only 61% of students at Black-majority schools taking the SAT and 81% of students at White-majority schools taking the SAT.

Graphical Summary

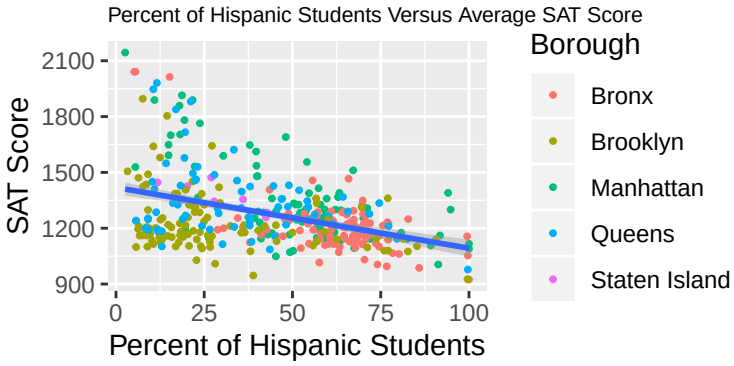
```
ggplot(scores, aes(x=white_prop, y=score)) +
  geom_point(aes(color=borough), size=.7) + geom_smooth(method=lm) +
  labs(title="Percent of White Students Versus Average SAT Score",
       x="Percent of White Students", y="SAT Score", color="Borough") +
  theme(plot.title=element_text(size=8))
```



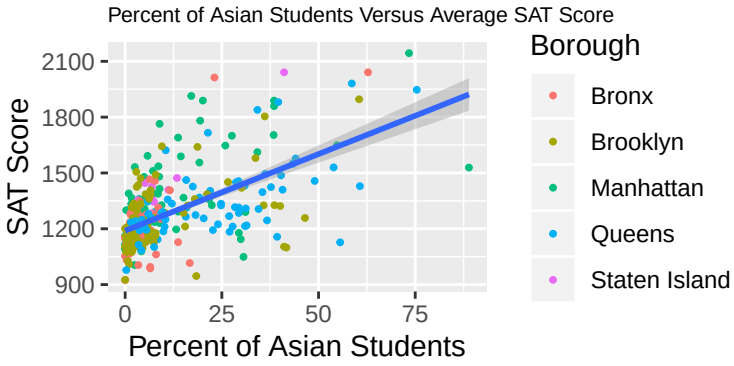

```
ggplot(scores, aes(x=black_prop, y=score)) +
  geom_point(aes(color=borough), size=.7) + geom_smooth(method=lm) +
  labs(title="Percent of Black Students Versus Average SAT Score",
       x="Percent of Black Students", y="SAT Score", color="Borough") +
  theme(plot.title=element_text(size=8))
```




```
ggplot(scores, aes(x=hispanic_prop, y=score)) +
  geom_point(aes(color=borough), size=.7) + geom_smooth(method=lm) +
  labs(title="Percent of Hispanic Students Versus Average SAT Score",
       x="Percent of Hispanic Students", y="SAT Score", color="Borough") +
  theme(plot.title=element_text(size=8))
```

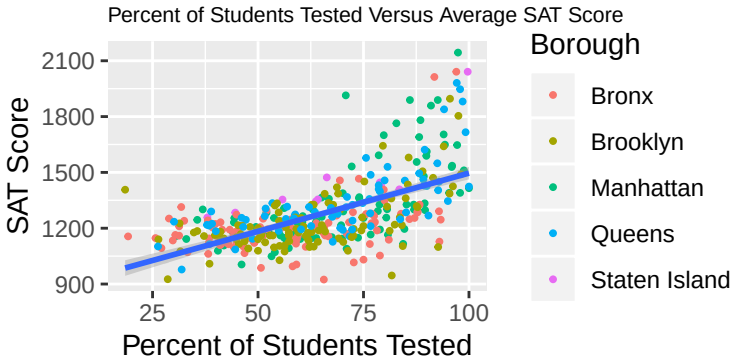



```
ggplot(scores, aes(x=asian_prop, y=score)) +
  geom_point(aes(color=borough), size=.7) + geom_smooth(method=lm) +
  labs(title="Percent of Asian Students Versus Average SAT Score",
       x="Percent of Asian Students", y="SAT Score", color="Borough") +
  theme(plot.title=element_text(size=8))
```



This set of scatterplots visually shows the racial trends discovered in the Numerical Summary section. While percentages of Black and Hispanic students have negative correlations with SAT score, percentages of White and Asian students have positive correlations with SAT score. Clustering of colors in each of these four scatterplots indicates a relationship between Borough and racial composition of schools, giving evidence to the additional geographic factors that influence SAT score in addition to racial ones.

```
ggplot(scores, aes(x=tested_prop, y=score)) +
  geom_point(aes(color=borough), size=.7) +
  labs(title="Percent of Students Tested Versus Average SAT Score",
       x="Percent of Students Tested", y="SAT Score", color="Borough") +
  geom_smooth(method=lm) +
  theme(plot.title=element_text(size=8))
```



This scatterplot shows the weak curvilinear relationship between the proportion of students who take the SAT at a school and that school's average SAT score. By coloring the points by Borough, it is evident that the schools with the highest average SAT scores are primarily in Queens and Manhattan, shedding light on another potential factor influencing students' SAT scores: location. As can be seen by clusters of colors on the 4 scatterplots describing racial composition versus SAT score, a school's average SAT score, racial composition, and location within NYC are all related.

Conclusion

The analysis of this data shows relationships between racial composition of a school and its percentage of students who take the SAT, racial composition of a school and its average SAT score, and percentage of students who take the SAT and SAT score. Of course, correlation cannot be equated to causation. It is possible that these factors influence each other in any order, or indirectly through other factors not analyzed in this project.

Because the x-axis of a graph is often assumed to be causative of the resulting y-axis values, it is necessary to be extremely careful when interperting the graphics in this project. It is highly unlikely that students' race has a direct impact their SAT score, as could be erroneously concluded from the scatterplot showing that schools with greater percentages of Black students have lower average SAT scores. Instead, one must be aware that a myriad of additional factors associated with race and SAT score, such as income, educational opportunity, and discriminatory zoning, likely also play a role in creating this correlation. The most confident justifiable conclusion that can be made from the analysis performed in this project is as follows: There is clear evidence of racial and geographic inequality in SAT scores.

References

(1) https://www.kaggle.com/nycopendata/high-schools

(2) http://www.programmingr.com/examples/remove-na-rows-in-r/

(3) https://stackoverflow.com/questions/8329059/how-to-convert-character-of-percentage-into-numeric-in-r

(4) https://stat.ethz.ch/R-manual/R-devel/library/base/html/colnames.html

(5) https://stackoverflow.com/questions/13540955/ggplot2-figure-size-with-rmarkdown