Stats 32 Final Project: SAT Scores by California County

Kaitlyn Tang (SUNet ID: k8lyn10g)

May 16, 2020

Introduction

This project is an analysis of average SAT scores in California by county. Specifically, this project explores a possible correlation between the average SAT score of a county in California and the county's percentage of high school students eligible for Free or Reduced Price Meals.

The data for SAT scores comes from data.world (https://data.world/education/california-sat-report-2015-2016) and PrepScholar (https://blog.prepscholar.com/average-sat-scores-over-time) (https://data.world/education/california-sat-report-2015-2016 (https://data.world/education/california-sat-report-2015-2016) and https://blog.prepscholar.com/average-sat-scores-over-time (https://blog.prepscholar.com/average-sat-scores-over-time)). The data on percentage of students on Free or Reduced Price Meals (FRPM) was provided by the California Department of Education (https://www.cde.ca.gov/ds/sd/sd/filessp.asp) (https://www.cde.ca.gov/ds/sd/sd/filessp.asp (https://www.cde.ca.gov/ds/sd/sd/filessp.asp)). Lastly, the California county map data comes from R's map package.

As the SAT is an important component of access to higher education, this study aims to analyze the impact which socioeconomic status may factor into SAT scores. In order to explore this topic, we will focus on four main questions:

1. How has the national average for SAT scores changed throughout the past few decades?
2. What is the distribution of average SAT scores across California counties?
3. What is the distribution of percentages of high school students on Free or Reduced Price Meals across California counties?
4. Is there a correlation between a California county's average SAT score and its percentage of high school students on Free or Reduced Price Meals?

Data Analysis

Loading Packages

```
library(tidyverse)
library(maps)
library(plotly)
```

Data Visualization 1

This is an analysis of the overall trends of SAT scores from 1972 to 2019 using this data (https://blog.prepscholar.com/average-sat-scores-over-time). Between the years 2006 and 2016, “Writing” served as a testing category alongside “Critical Reading”. To weigh these two components evenly, the average of these two sections was taken to represent the current singular section “Evidence-Based Reading and Writing”.

```
timeTrends <- read_csv("avg_sat_over_time.csv")
```

```
## Warning: Missing column names filled in: 'X5' [5], 'X6' [6], 'X7' [7], 'X8' [8]
```

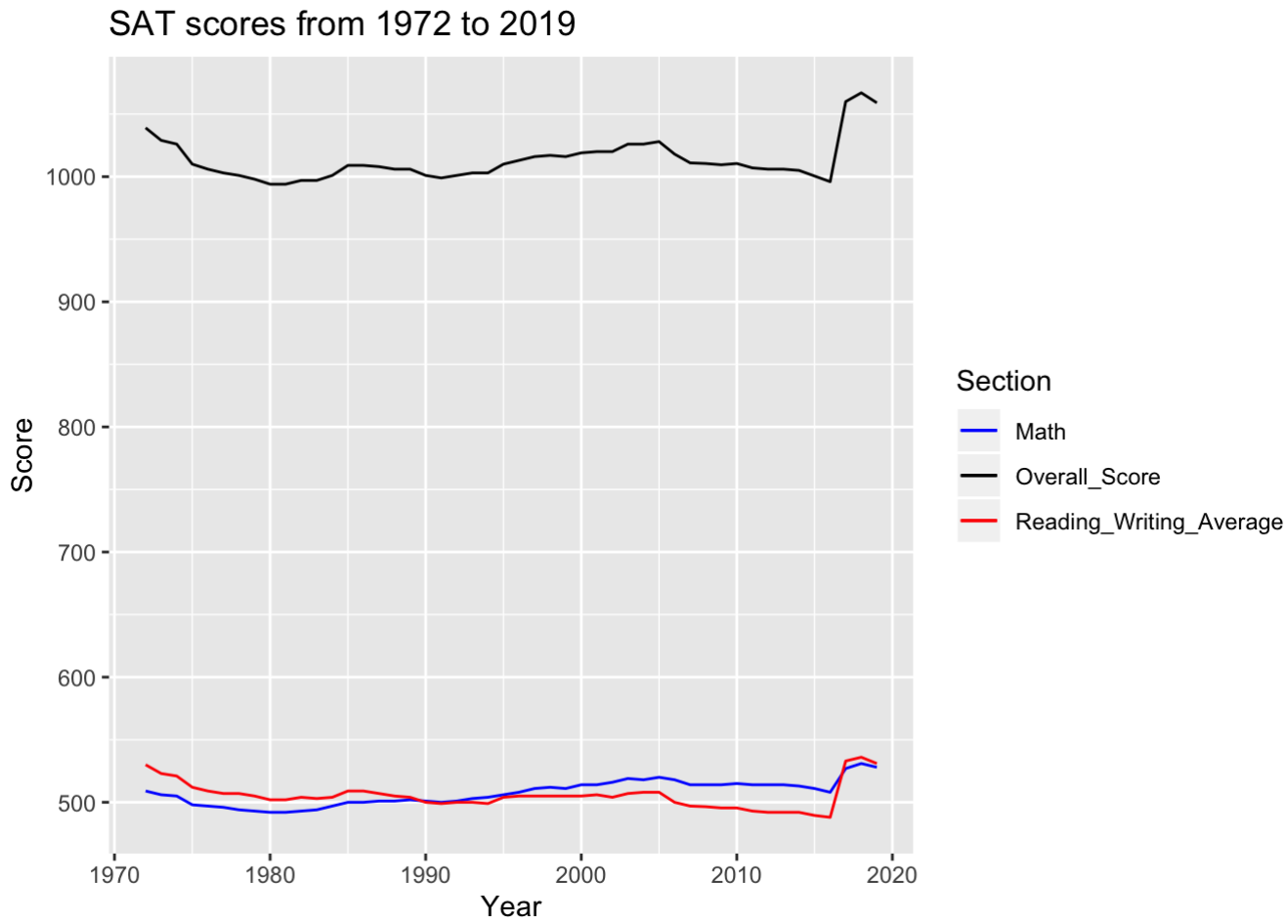
```
## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Math = col_double(),
##   `Critical Reading` = col_double(),
##   Writing = col_double(),
##   X5 = col_logical(),
##   X6 = col_logical(),
##   X7 = col_logical(),
##   X8 = col_logical()
## )
```

```
timeTrends <- timeTrends %>% select(c(1,2,3,4))
timeTrends <- mutate(timeTrends, Reading_Writing_Average = rowMeans(timeTrends[c(3,4)],
 na.rm=TRUE))
timeTrends<- mutate(timeTrends, Overall_Score = rowSums(timeTrends[c(2,5)]))
head(timeTrends)
```

```
## # A tibble: 6 x 6
##    Year  Math `Critical Reading` Writing Reading_Writing_Average Overall_Score
##   <dbl> <dbl>              <dbl>   <dbl>                   <dbl>         <dbl>
## 1  1972   509                530      NA                     530          1039
## 2  1973   506                523      NA                     523          1029
## 3  1974   505                521      NA                     521          1026
## 4  1975   498                512      NA                     512          1010
## 5  1976   497                509      NA                     509          1006
## 6  1977   496                507      NA                     507          1003
```

This is a lineplot showing the average Math, Evidence-Based Reading and Writing, and overall SAT scores from 1972 to 2019.

```
SATTime <- ggplot(data = timeTrends, aes(x = Year)) + geom_line(data = timeTrends, aes(x
= Year, y = Math, color = "Math")) + geom_line (data = timeTrends, aes(x = Year, y = Rea
ding_Writing_Average, color = "Reading_Writing_Average")) + geom_line(data = timeTrends,
aes(x = Year, y = Overall_Score, color = "Overall_Score")) + ggtitle("SAT scores from 19
72 to 2019") +labs(x = "Year", y = "Score", color="Section") + scale_color_manual(values
=c("blue", "black", "red"))
plot(SATTime)
```



The line graph indicates that the average math, evidence-based reading and writing, and overall SAT scores have followed similar trends between 1972 and 2019. While these averages were fairly constant between 1972 and 2015, in the past five years the averages for math, evidence-based reading and writing, and overall SAT scores has increased dramatically.

Data Visualization 2

This is an analysis of average SAT scores for California counties in the testing period between 2015 and 2016. (NOTE: For the testing period from 2015 to 2016 the maximum SAT score was 2400).

Using this data (https://data.world/education/california-sat-report-2015-2016), the average SAT score for each county was calculated. Any school that had no reported data was dropped from the dataset, and a weighted average SAT score (each school's score was weighted by the percentage of the county's test takers that were from the school) was calculated for each county.

```
SAT_2015_2016 <- read.csv("SAT_2015_2016.csv",stringsAsFactors=FALSE)
SAT_2015_2016 <- select(SAT_2015_2016, c(4,5,6,7,8,9,10))
SAT_2015_2016 <- SAT_2015_2016 %>% rename("District_Name" = dname, "County_Name" = cnam
e, "12th Graders" = enroll12, "Number_Test_Takers" = NumTstTakr, "Avg_Read" = AvgScrRea
d, "Avg_Math" = AvgScrMath, "Avg_Writ" = AvgScrWrit)
SAT_2015_2016 <- mutate(SAT_2015_2016, Avg_Read = na_if(Avg_Read, "*"), Avg_Math = na_if
(Avg_Math, "*"), Avg_Writ = na_if(Avg_Writ, "*"))
SAT_2015_2016$Avg_Read <- as.numeric(as.character(SAT_2015_2016$Avg_Read))
SAT_2015_2016$Avg_Math <- as.numeric(as.character(SAT_2015_2016$Avg_Math))
SAT_2015_2016$Avg_Writ <- as.numeric(as.character(SAT_2015_2016$Avg_Writ))
SAT_2015_2016 <- drop_na(SAT_2015_2016)
SAT_2015_2016 <- mutate(SAT_2015_2016, Overall_Score = rowSums(SAT_2015_2016[c(5,6,7)],
 na.rm=FALSE))
SAT_2015_2016 <- SAT_2015_2016 %>% group_by(County_Name) %>% mutate(County_test_takers =
sum(Number_Test_Takers)) %>% transform(Percent_testers = Number_Test_Takers / County_tes
t_takers) %>% transform(Weighted_Score = Percent_testers * Overall_Score)
SAT_2015_2016_Counties <- SAT_2015_2016 %>% group_by(County_Name) %>% summarize(County_S
AT = sum(Weighted_Score))
head(SAT_2015_2016)
```

```
##                       District_Name County_Name X12th.Graders
## 1                                                      492835
## 2                                    Alameda          16662
## 3 Alameda County Office of Education     Alameda            263
## 4 Alameda County Office of Education     Alameda             88
## 5                    Alameda Unified     Alameda            858
## 6                    Alameda Unified     Alameda             37
##   Number_Test_Takers Avg_Read Avg_Math Avg_Writ Overall_Score
## 1             214262      484      494      477          1455
## 2               8611      517      534      515          1566
## 3                 95      395      378      388          1161
## 4                 92      391      376      386          1153
## 5                472      527      543      514          1584
## 6                 35      572      612      530          1714
##   County_test_takers Percent_testers Weighted_Score
## 1             214262     1.000000000    1455.000000
## 2              25679     0.335332373     525.130496
## 3              25679     0.003699521       4.295144
## 4              25679     0.003582694       4.130846
## 5              25679     0.018380778      29.115152
## 6              25679     0.001362981       2.336150
```

```
head(SAT_2015_2016_Counties)
```

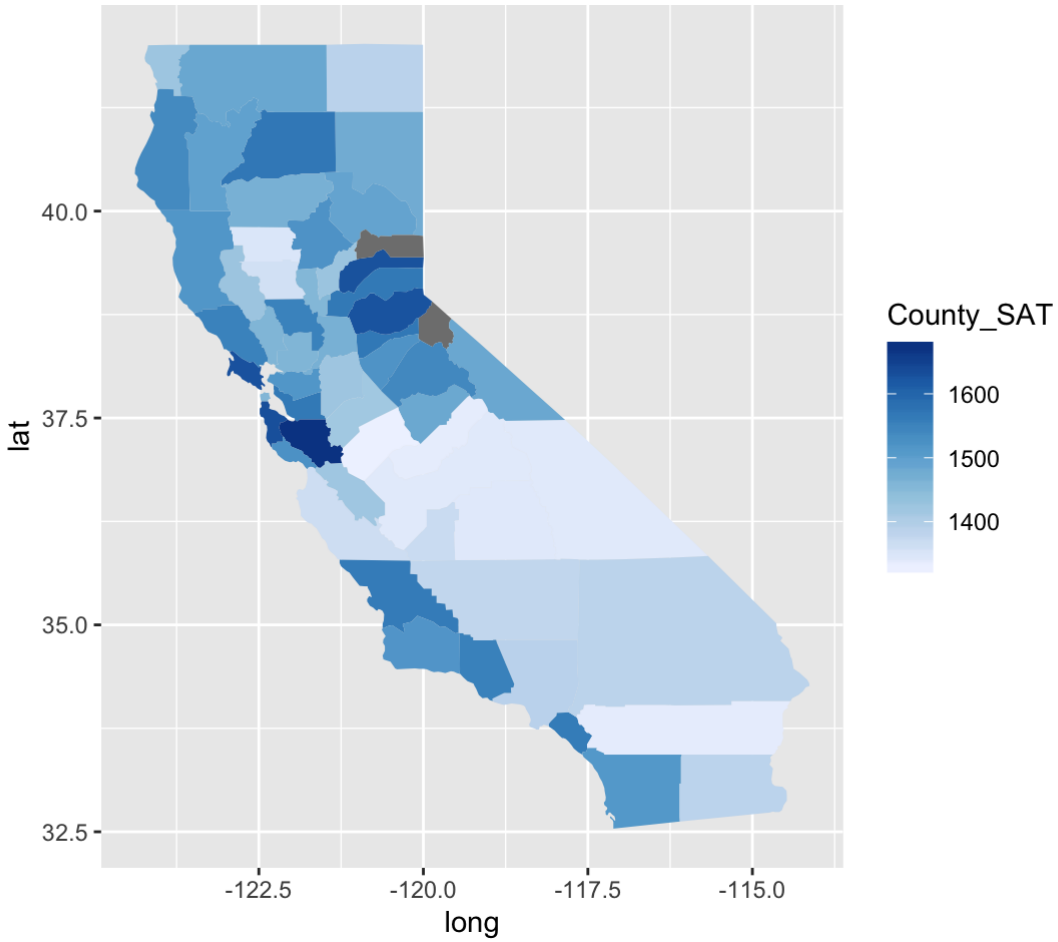
```
## # A tibble: 6 x 2
##   County_Name County_SAT
##   <chr>            <dbl>
## 1 ""               1455
## 2 "Alameda"        1567.
## 3 "Amador"         1569.
## 4 "Butte"          1520.
## 5 "Calaveras"      1520.
## 6 "Colusa"         1360.
```

This is a map of the average SAT score for each county in California for the academic year 2015-2016, using R's maps package and the dataset of average SAT scores for each county in California.

```
county_data <- map_data("county")
CA_data <- county_data %>% filter(region == "california")
CA_map <- ggplot(data = CA_data) + geom_polygon(mapping = aes(x = long, y = lat, group =
group)) + coord_quickmap()
SAT_2015_2016_Counties$County_Name <- tolower(SAT_2015_2016_Counties$County_Name)
CAMap_SAT_data <- left_join(CA_data, SAT_2015_2016_Counties, by = c("subregion" = "Count
y_Name"))
head(CAMap_SAT_data)
```

```
##        long      lat group order     region subregion County_SAT
## 1 -121.4785 37.48290   157  6965 california   alameda    1567.21
## 2 -121.5129 37.48290   157  6966 california   alameda    1567.21
## 3 -121.8853 37.48290   157  6967 california   alameda    1567.21
## 4 -121.8968 37.46571   157  6968 california   alameda    1567.21
## 5 -121.9254 37.45998   157  6969 california   alameda    1567.21
## 6 -121.9483 37.47717   157  6970 california   alameda    1567.21
```

```
CA_SAT_MAP <- ggplot(data = CAMap_SAT_data) + geom_polygon(mapping = aes(x = long, y = l
at, group = group, fill = County_SAT)) + coord_quickmap() + scale_fill_distiller(directi
on = 1)
plot(CA_SAT_MAP)
```



On the map, a higher average SAT score (on a scale out of 2400) is corresponded with darker shades of blue and lower average SAT scores with lighter shades of blue. Two counties are missing data, as indicated by the grey shaded regions. Looking at the map it appears that regions along the coast and within Northern California have higher average SAT scores than regions in southern and eastern California.

Data Visualization 3

This is an analysis of the distribution of percentage of high school students on Free or Reduced Price Meals across counties of California for the academic school year of 2015-2016.

For each county in California, the percentage of high school students on Free or Reduced Price Meals during the academic year 2015-2016 was found by finding the proportion of a county's high school students on FRPM out of the total enrolled high school students in the county using data provided by the California Department of Education (https://www.cde.ca.gov/ds/sd/sd/filessp.asp).

```
FreeRedMeal_1516 <- read.csv("2015-2016_FreeReduced.csv", header = TRUE)
FreeRedMeal_1516_HighSchools <- subset(FreeRedMeal_1516, School.Type == "High Schools (P
ublic)" )
FreeRedMeal_1516_HighSchools <- FreeRedMeal_1516_HighSchools %>% select(c(5,9,18,21,22))
FreeRedMeal_1516_County <- FreeRedMeal_1516_HighSchools %>% group_by(County.Name) %>% mu
tate(Total.County.Enrollment = sum(Enrollment...K.12.)) %>% mutate(Total.FRPM.Enrollment
= sum(FRPM.Count...K.12.)) %>% mutate(Percent.FRPM.Enrollment = Total.FRPM.Enrollment /
 Total.County.Enrollment * 100)
FreeRedMeal_1516_County_CA <- FreeRedMeal_1516_County %>% group_by(County.Name) %>% summ
arize(Percent.FRPM.Enrollment = mean(Percent.FRPM.Enrollment))
head(FreeRedMeal_1516_HighSchools)
```

```
##    County.Name           School.Type Enrollment...K.12. FRPM.Count...K.12.
## 1      Alameda High Schools (Public)                407                274
## 18     Alameda High Schools (Public)               1718                312
## 19     Alameda High Schools (Public)                379                 42
## 20     Alameda High Schools (Public)                165                154
## 36     Alameda High Schools (Public)               1210                206
## 43     Alameda High Schools (Public)                366                268
##    Percent......Eligible.FRPM...K.12.
## 1                               67.3%
## 18                              18.2%
## 19                              11.1%
## 20                              93.3%
## 36                              17.0%
## 43                              73.2%
```

```
head(FreeRedMeal_1516_County_CA)
```

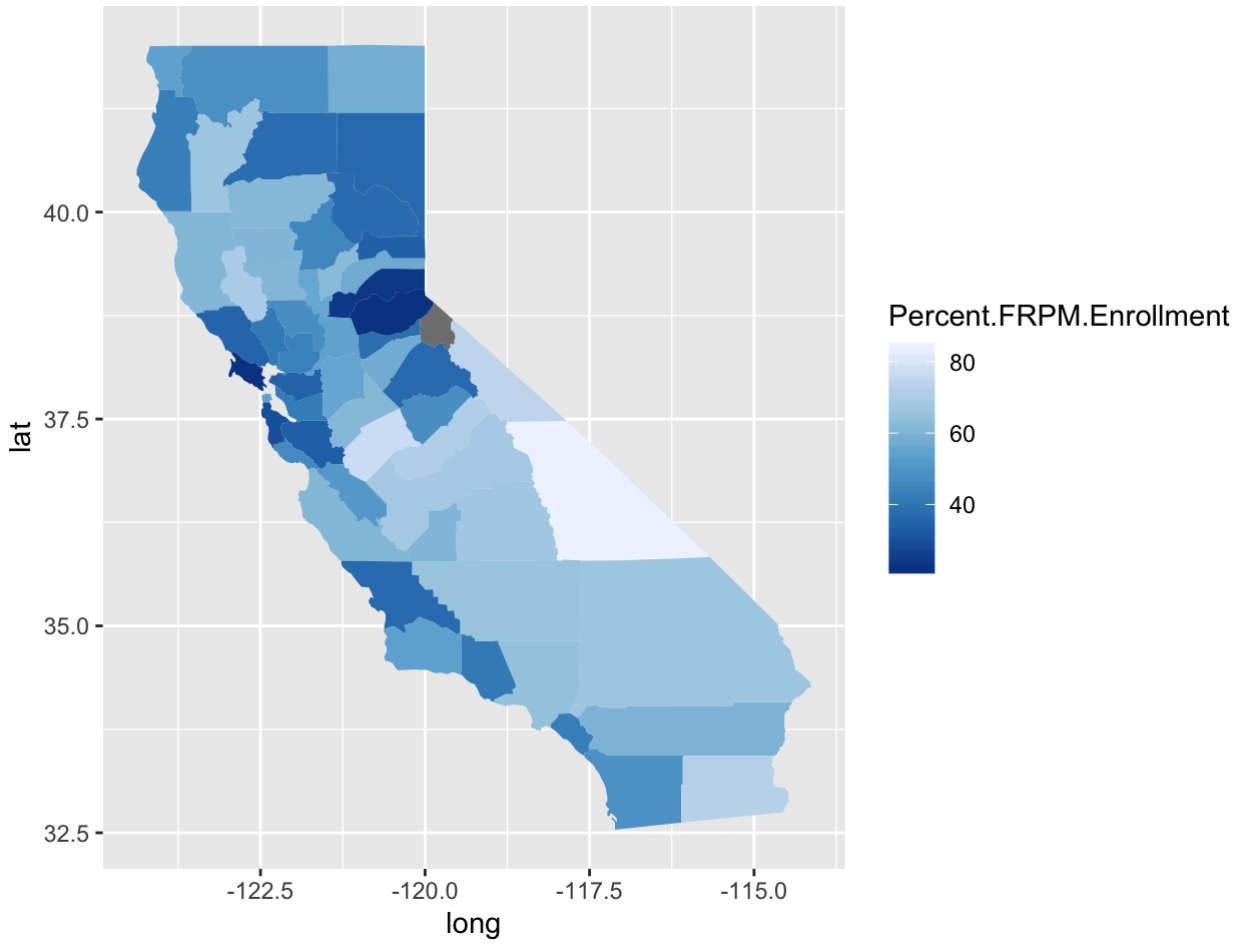
```
## # A tibble: 6 x 2
##   County.Name  Percent.FRPM.Enrollment
##   <fct>                         <dbl>
## 1 Alameda                        42.1
## 2 Amador                         38.6
## 3 Butte                          45.6
## 4 Calaveras                      57.3
## 5 Colusa                         60.8
## 6 Contra Costa                   34.5
```

This is a map of the distribution of percentages of public high school students across California counties eligible for Free or Reduced Price Meals for the academic year 2015-2016 using R's maps package and the California data on percentages of high school students on FRPM.

```
county_data <- map_data("county")
CA_data <- county_data %>% filter(region == "california")
CA_map <- ggplot(data = CA_data) + geom_polygon(mapping = aes(x = long, y = lat, group =
group)) + coord_quickmap()
FreeRedMeal_1516_County_CA$County.Name <- tolower(FreeRedMeal_1516_County_CA$County.Nam
e)
CAMap_FRPM_data <- left_join(CA_data, FreeRedMeal_1516_County_CA, by = c("subregion" =
"County.Name"))
head(CAMap_FRPM_data)
```

```
##        long      lat group order     region subregion Percent.FRPM.Enrollment
## 1 -121.4785 37.48290   157  6965 california   alameda                 42.07334
## 2 -121.5129 37.48290   157  6966 california   alameda                 42.07334
## 3 -121.8853 37.48290   157  6967 california   alameda                 42.07334
## 4 -121.8968 37.46571   157  6968 california   alameda                 42.07334
## 5 -121.9254 37.45998   157  6969 california   alameda                 42.07334
## 6 -121.9483 37.47717   157  6970 california   alameda                 42.07334
```

```
CA_FRPM_MAP <- ggplot(data = CAMap_FRPM_data) + geom_polygon(mapping = aes(x = long, y =
lat, group = group, fill = Percent.FRPM.Enrollment)) + coord_quickmap() + scale_fill_dis
tiller(direction = -1)
plot(CA_FRPM_MAP)
```



On the map, a lower percentage of high school students in a county on Free or Reduced Price Meals is represented by a darker shade of blue while a higher percentage of high school students in a county on Free or Reduced Price Meals is represented by a lighter shade of blue. One county is missing data as indicated by the grey shaded region. The map indicates that counties along the coast of California and in the northern region tend to have a lower percentage of students on Free or Reduced Price Meals than counties in southern and eastern regions of the state. Overall, the regions on the map with a lower percentage of high school students on Free or Reduced Price Meals tend to be the regions that had higher average SAT scores on the previous data visualization.

Data Visualization 4

This is an analysis of the correlation between a California county's average SAT score and its percentage of students on Free or Reduced Price Meals for the academic year 2015-2016.

This is data of a California county's average SAT score and its percentage of students on Free or Reduced Price Meals during the 2015-2016 school year.

```
SAT_FRPM_county <- full_join(SAT_2015_2016_Counties, FreeRedMeal_1516_County_CA, by = c(
"County_Name" = "County.Name") )
SAT_FRPM_county <- drop_na(SAT_FRPM_county)
head(SAT_FRPM_county)
```

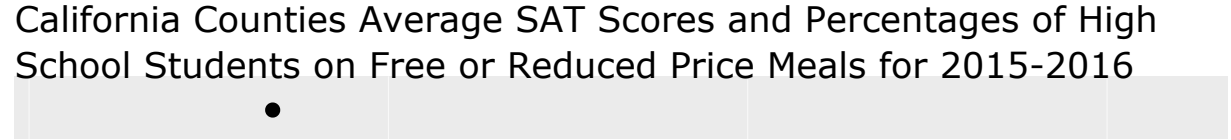
```
## # A tibble: 6 x 3
##   County_Name  County_SAT Percent.FRPM.Enrollment
##   <chr>             <dbl>                   <dbl>
## 1 alameda           1567.                    42.1
## 2 amador            1569.                    38.6
## 3 butte             1520.                    45.6
## 4 calaveras         1520.                    57.3
## 5 colusa            1360.                    60.8
## 6 contra costa      1515.                    34.5
```

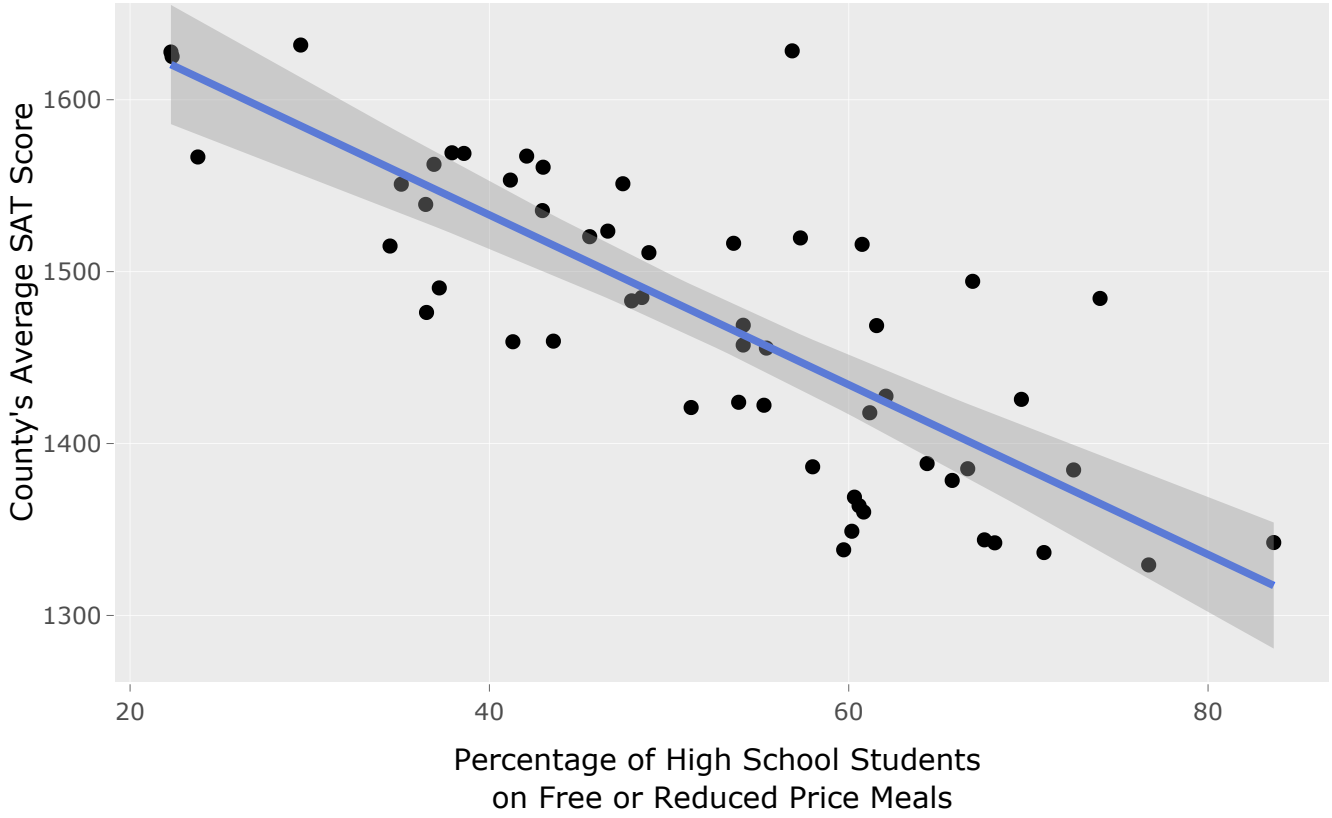
This is a scatterplot of the correlation between a California county's average SAT score and its percentage of high school students on Free or Reduced Price Meals for the 2015-2016 academic school year.

```
plotFRPM_SAT <- ggplot(SAT_FRPM_county, aes(x = Percent.FRPM.Enrollment, y = County_SA
T)) + geom_point(aes(z = County_Name, y1 = County_SAT, x1 = Percent.FRPM.Enrollment)) +
ggtitle("California Counties Average SAT Scores and Percentages of High \nSchool Student
s on Free or Reduced Price Meals for 2015-2016") + labs(x = "Percentage of High School S
tudents \non Free or Reduced Price Meals", y = "County's Average SAT Score") + stat_smoo
th(method = "lm")
```

```
## Warning: Ignoring unknown aesthetics: z, y1, x1
```

```
ggplotly(plotFRPM_SAT, tooltip = c("z", "x1", "y1"))
```





The scatterplot indicates a negative correlation between the percentage of high school students in a California county that is on Free or Reduced Price Meals and the county's average SAT score for the academic year 2015-2016. The correlation is seen by the downward trend of each county's data as well as the line of best fit which has a negative slope. To see each individual county's percentage of Free and Reduced Price Meals and average SAT score, hover over each data point.

Conclusion

Through this analysis of a California county's average SAT score and its percentage of high school students on Free or Reduced Price Meals in 2015-2016, there appears to be a slight negative correlation between the two. More specifically, to address the initial questions:

1. It appears that the average math, evidence-based reading and writing, and overall SAT scores have stayed fairly constant in the past few decades. However, in the past five years, the scores have increased slightly.
2. The distribution of the average SAT score for California counties from 2015-2016 indicates that counties on the coast and in northern regions of the state tend to have higher average SAT scores while regions in eastern and southern California tend to have lower average SAT scores.
3. The distribution of the percentage of high school students on Free or Reduced Price Meals for California counties from 2015-2016 indicates that counties on the coast and in northern regions of the state tend to have lower percentages of high school students on Free or Reduced Price Meals while regions in eastern and southern California tend to have higher percentages of high school students on Free or Reduced Price Meals.
4. There appears to be a slight negative correlation between a California county's percentage of high school students on Free or Reduced Price Meals and the county's average SAT score for the academic year 2015-2016.

While this project appears to show a slight negtive correlation between a California county's average SAT score and percentage of high school students on Free or Reduced Price Meals, the study has many limitations. This project only analyzed the academic year 2015-2016 and did not look at any other year's data. Additionally, the study looked at California counties as a whole, which often are quite large and have diverse demographics within themselves. While my project proposal suggested analyzing California by various regions such as school districts, for this project I chose to focus specifically on California counties. Future interesting areas for study would be investigating other academic years or analyzing individual public school districts.