US Education by State - 1992

Improvements: This analysis only provides information for 1992, so there are only 50 observations. This can limit the analysis, so an improvement that can be made is to perform this analysis on state throughout many years.

Another modification is to improve the aesthetics of the visualizations to make it easier to read without using the interpretations. This can be done by exploring the features of ggplot2 further.

Introduction

This report explores US education data by state in 1992. One data set, acquired from the carData package in R, provides information like the population of the state in 1992, average SAT scores for both components (math and verbal), and the percentage of graduating high-school students in the state who took the SAT exam. Another data set, acquired from Kaggle, provides education data per state from 1992 to 2019. It includes information like state expenditure related to education and a breakdown of students enrolled in schools by school year. Finally, the third data set simply provides the state associated with the state abbreviation. This was acquired from the World Population Review, just to make joining the two previous data sets easier.

Education has always been a topic of interest to me, especially as I delve deeper into the U.S. education as a student myself. I have experienced the education system in depth in two different states, spending the first 18 years of my life in Colorado before coming to Texas for college. As such, the discrepancies between education in both states has always intrigued me, and I have been interested in understanding this in a broader context as well. Although this information is outdated, it is always valuable to understand historical data to get a grasp of if and how the education system has developed for the better over the years. One of the associations that I expect to find is that the states who tend to spend more on education will see higher average scores on standardized tests. Another association is that states who have less participation in tests will see higher average scores as well. I will also perform dimension reductionality to see if there is a difference in education quality and expenditure amongst regions and states in 1992.

Tidying & Cleaning

```
library(carData)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.0.5     ✓ dplyr   1.0.3
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.0
```

```
## ── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
### IMPORTING DATA
#load States data from cardata package into environment, name it States.edu1, and convert rownames to column
data(States)
States.edu1 <- rownames_to_column(States, "State")
#set working directory and import second set of education and states data
setwd("~/Downloads")
States.edu2 <- read.csv("states_all.csv")
#import state names & abbreviation data
State.Abbr<- read.csv("csvData.csv")

### CLEANING STATES.EDU1 FOR JOINING PURPOSES
#rename column on States.edu1 so that there is a match when joining two datasets
States.edu1 <- States.edu1%>%
  rename(Code = State)
#recode data in States.edu1 so that they match with States.Abbr instances
States.edu1 <- States.edu1 %>%
  mutate(Code = recode(Code, 'CN' = 'CT'))
```

Before joining the three datasets, it is important that there is a matching key variable between the three. The States Abbreviations dataset and the States dataset from R had a discrepancy in the way Connecticut was abbreviated, so there needed to be some cleaning of the dataset to fix this.

Joining

```
### JOINS
#Join States.edu1 and state abbreviation datasets
States.edu1 <- States.edu1%>%
  full_join(State.Abbr, by="Code") %>%
  select(-Abbrev)
#Reorder columns for better readability
States.edu1 <- States.edu1[c(1, 9, 2, 3, 4, 5, 6, 7, 8)]
#Recode all states in States.edu1 to match States.edu2 instances
States.edu1$State = toupper(States.edu1$State)
States.edu1$State <- gsub(" ", "_", States.edu1$State)
#Create a primary_key column in States.edu1
States.edu1 <- States.edu1 %>%
  mutate(PRIMARY_KEY = paste("1992", State, sep = "_"))
#Join States.edu1 with States.edu2 on the primary key column
States.edu3 <- States.edu1%>%
  left_join(States.edu2, by="PRIMARY_KEY")%>%
  select(-10, -11, -13, -12, -21, -28, -29, -34)
```

To join all three datasets into one, there were two join statements required. I chose a full join for the State data from R and the abbreviations data. Conveniently enough, both sets took data for 50 states and for Washington D.C., so there were 51 observations for both. An inner join, left join, full join, and right join would have done the same thing.

For the second join (between the newly created dataset and the kaggle data set), I wanted only 1992 data for each state. I created a matching Primary key on both sets and used an inner join to drop all the instances in the Kaggle data set that did not show data for 1992.

Summary Statistics and Usage of Core dplyr Functions

```
### SUMMARY STATISTICS AND USAGE OF CORE DPLYR FUNCTIONS

#Create categorical variable SAT_Participation, measuring how high their percentage of students taking the S
AT is to other states
summary(States.edu3$percent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.00   11.50   25.00   33.75   57.50   74.00
```

```
States.edu3 <- States.edu3 %>%
  mutate(SAT_Participation = case_when(percent <= 11.50 ~ "Low",
                                       percent > 11.50 & percent <= 57.5 ~ "Average", percent > 57.5 ~ "Hig
h") )

#1 - Number of distinct regions
States.edu3 %>%
  summarize(n_distinct(region))
```

```
##   n_distinct(region)
## 1                  9
```

```
#2 - Mean of SAT Verbal for all states in the mountain region
States.edu3 %>%
  filter(region == "MTN") %>%
  summarize(mean(SATV))
```

```
##   mean(SATV)
## 1    461.875
```

```
#3 - Mean of SAT Math per region
States.math.means <- States.edu3 %>%
  select(region, dollars, SATM) %>%
  group_by(region) %>%
  summarize(mean_SATM = mean(SATM), mean_dollars = mean(dollars)) %>%
  arrange(desc(mean_SATM))

States.math.means
```

```
## # A tibble: 9 x 3
##   region mean_SATM mean_dollars
##   <fct>      <dbl>        <dbl>
## 1 WNC         551.         4.47
## 2 ESC         520.         3.77
## 3 MTN         513.         4.34
## 4 ENC         509.         5.39
## 5 WSC         503          3.83
## 6 PAC         482.         5.61
## 7 NE          470          6.40
## 8 MA          469.         8.06
## 9 SA          459.         5.55
```

```
#4 - Mean and standard deviation of average teacher's salary in the state, grouped by SAT participation rate, arranged from highest to lowest
States.edu3 %>%
  group_by(SAT_Participation) %>%
  summarize (mean_pay = mean(pay), sd_pay = sd(pay)) %>%
  arrange(desc(mean_pay))
```

```
## # A tibble: 3 x 3
##   SAT_Participation mean_pay sd_pay
##   <chr>                <dbl>  <dbl>
## 1 High                  35.8   4.41
## 2 Average               30.9   4.28
## 3 Low                   26.1   3.09
```

```
#5 - Median of Instruction Expenditure as a percentage of Total Expenditure, grouped by SAT participation rate
States.edu3 <- States.edu3 %>%
  mutate(Ins.As.Per.Tot = round((INSTRUCTION_EXPENDITURE/TOTAL_EXPENDITURE) * 100))

States.edu3 %>%
  group_by(SAT_Participation) %>%
  summarize(medperc = median(Ins.As.Per.Tot)) %>%
  arrange(desc(medperc))
```

```
## # A tibble: 3 x 2
##   SAT_Participation medperc
##   <chr>               <dbl>
## 1 High                   56
## 2 Low                    55
## 3 Average                52
```

```
#6 - Variation in Local Revenue
States.edu3 %>%
  summarize(sd = sd(LOCAL_REVENUE))
```

```
##        sd
## 1 2547132
```

```
#7 - Five number summary of capital outlay expenditure
States.edu3 %>%
  summarize(quantile(CAPITAL_OUTLAY_EXPENDITURE))
```

```
##   quantile(CAPITAL_OUTLAY_EXPENDITURE)
## 1                              14685.0
## 2                             118782.5
## 3                             194081.0
## 4                             525026.5
## 5                            2044688.0
```

```
#8 - Find the number of states that fall in each category for SAT Participation rates
States.edu3 %>%
  select(SAT_Participation) %>%
  group_by(SAT_Participation) %>%
  summarize(n())
```

```
## # A tibble: 3 x 2
##   SAT_Participation `n()`
## * <chr>             <int>
## 1 Average              25
## 2 High                 13
## 3 Low                  13
```

```
#9 - Find the number of states that fall in each region
States.edu3 %>%
  group_by(region) %>%
  summarize(n())
```

```
## # A tibble: 9 x 2
##   region `n()`
## * <fct>  <int>
## 1 ENC        5
## 2 ESC        4
## 3 MA         3
## 4 MTN        8
## 5 NE         6
## 6 PAC        5
## 7 SA         9
## 8 WNC        7
## 9 WSC        4
```

```
#10 - Find the average number of students enrolled in 12th grade
States.edu3 %>%
  summarize(mean(GRADES_12_G))
```

```
##   mean(GRADES_12_G)
## 1          47657.57
```

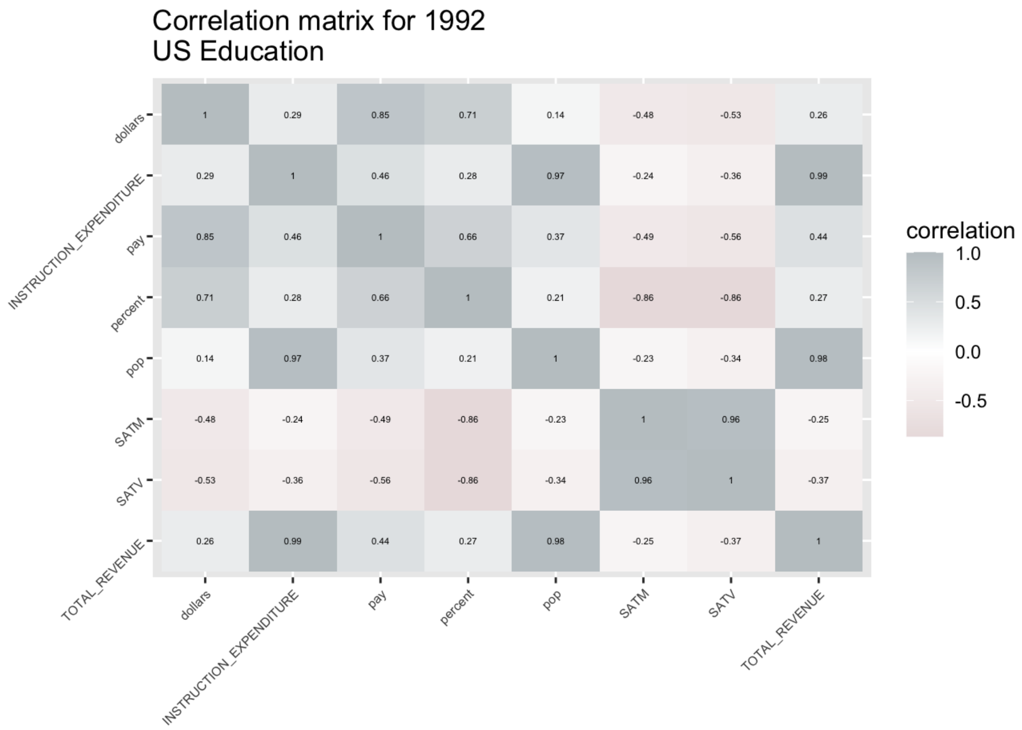
```
#11 - Correlation matrix for numeric variables

States.edu3.num <- States.edu3 %>%
  column_to_rownames("State") %>%
  select_if(is.numeric)

States.edu3.num <- States.edu3.num %>%
  select(1, 2, 3, 4, 5, 6, 7, 12)

cor(States.edu3.num)
```

```
##                                pop       SATV       SATM    percent    dollars
## pop                      1.0000000 -0.3381028 -0.2300418  0.2100687  0.1436745
## SATV                    -0.3381028  1.0000000  0.9620359 -0.8627954 -0.5268313
## SATM                    -0.2300418  0.9620359  1.0000000 -0.8581495 -0.4844477
## percent                  0.2100687 -0.8627954 -0.8581495  1.0000000  0.7111474
## dollars                  0.1436745 -0.5268313 -0.4844477  0.7111474  1.0000000
## pay                      0.3677244 -0.5559238 -0.4853306  0.6630098  0.8476737
## TOTAL_REVENUE            0.9815998 -0.3667500 -0.2480021  0.2702711  0.2628193
## INSTRUCTION_EXPENDITURE  0.9661415 -0.3581571 -0.2400128  0.2802296  0.2909489
##                                pay TOTAL_REVENUE INSTRUCTION_EXPENDITURE
## pop                      0.3677244     0.9815998               0.9661415
## SATV                    -0.5559238    -0.3667500              -0.3581571
## SATM                    -0.4853306    -0.2480021              -0.2400128
## percent                  0.6630098     0.2702711               0.2802296
## dollars                  0.8476737     0.2628193               0.2909489
## pay                      1.0000000     0.4430113               0.4558008
## TOTAL_REVENUE            0.4430113     1.0000000               0.9938057
## INSTRUCTION_EXPENDITURE  0.4558008     0.9938057               1.0000000
```

1. This data explores education data among 9 different regions across the United States.
2. The mean of the average score of graduating high-school students on the SAT Verbal in the Mountain region is about 461.88.
3. Region WNC had the highest mean SAT math score in 1992 with region ESC closely following, but these regions place nowhere near the top when it comes to mean state spending on public education per student.
4. The states that had high SAT participation rates tended to have the higher rates of participation in the SAT, but also came with greater variance.
5. The states that allocated the highest proportion of their total expenditure towards instruction and education had the highest SAT participation rates.
6. The variation in local revenue was very high among states in 1992, leading me to believe that there is an unequal distribution in the amount that different states earned.
7. Other expenditures, such as capital outlay, also take priority for states. Some states had significantly higher capital outlay expenditures than others, with this data showing a range of $2,030,003. This can explain the difference in instruction expenditure, but of course it is important to keep in mind how this relates in proportion to total revenue per state.
8. 25 states followed an average SAT participation rate, 13 states had a low SAT participation rate, and 13 states had a high SAT participation rate.
9. The SA region has the most number of states (9) while the MA region has the least number of states (3). This is important to keep in mind when trying to analyze this data based on region.
10. The average number of students enrolled in the 12th grade per state in 1992 was 47657.57. This is lower than I had expected.
11. Correlation matrix showing correlation between all numeric variables. Findings discussed in heatmap section.

Visualizatons

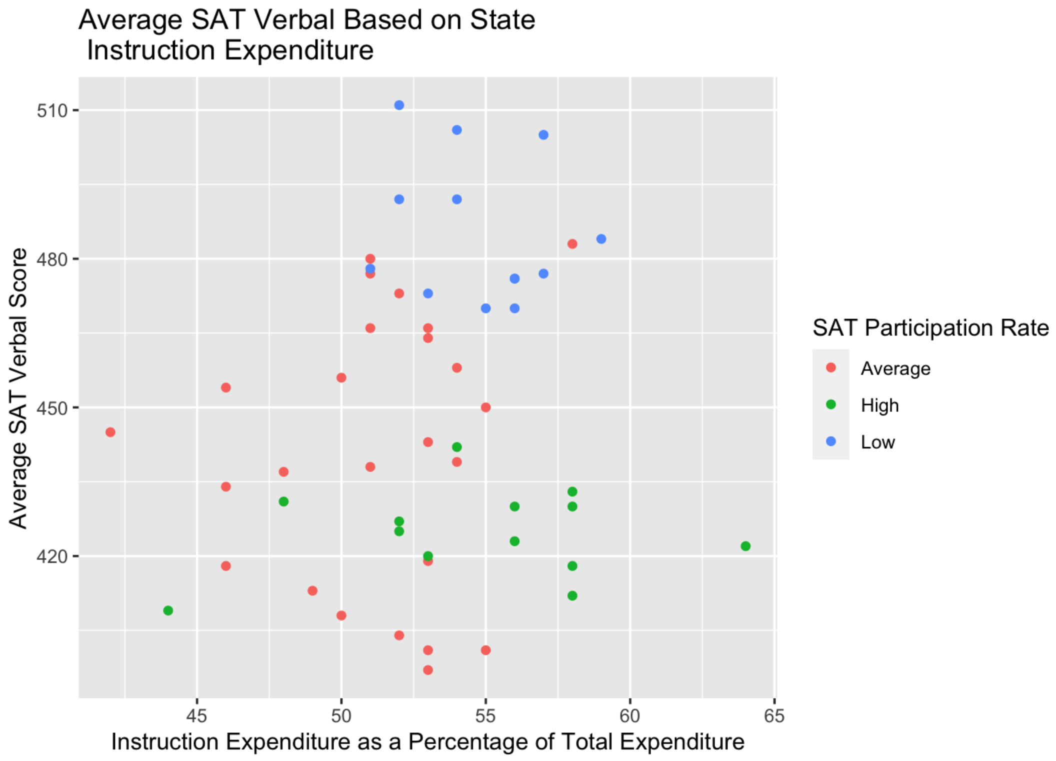
```
cor(States.edu3.num) %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  # Specify variables are displayed alphabetically from top to bottom
  ggplot(aes(rowname, factor(other_var, levels = rev(levels(factor(other_var)))), fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutral
  #scale_fill_gradient2(low="red",mid="white",high="blue") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 1.5) +
  # Give title and labels
  labs(title = "Correlation matrix for 1992 \nUS Education", x = "", y = "") + scale_fill_gradient2(low = "
#e7dbdb", high = "#c1c8cb", mid="white") + theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 6
)) + theme(axis.text.y = element_text(angle = 45, hjust = 1, size = 6))
```



This correlation matrix gives the correlation between all numeric variables. Some that stood out to me specifically are the strong negative correlation between percent & SATV and percent & SATM. This confirms an association I thought would occur initially: the lower the SAT participation rate is, the higher the average scores would be. This is because at that point, the people taking the exam would be the ones who want to go to college so they would likely take steps to increase their score.

Another thing that stood out to me was that there was a moderately high negative correlation of the dollars variable (state spending on public education) against SAT Verbal and against SAT Math. This means that the more money put into public education, the lower the SAT scores tended to be that year. This altered from what I had expected.

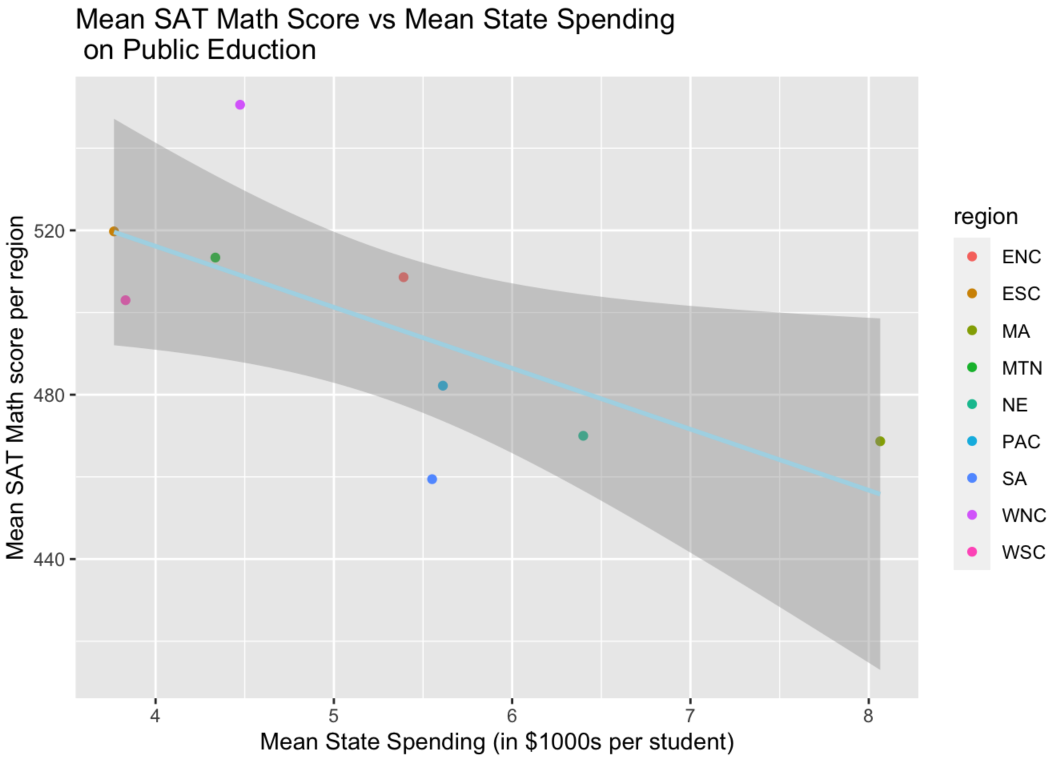
```
#Plot to show the relation between SAT Verbal scores and Instruction Expenditure as a percentage of Total Ex
penditure, mapped against the SAT Participation Level aesthetic
ggplot(States.edu3, aes(Ins.As.Per.Tot, SATV))+ geom_point(aes(color = SAT_Participation)) +
xlab("Instruction Expenditure as a Percentage of Total Expenditure") + ylab("Average SAT Verbal Score") + gg
title("Average SAT Verbal Based on State \n Instruction Expenditure") + labs(color = "SAT Participation Rate
") + scale_fill_gradient2(low = "#e7dbdb",
  high = "#c1c8cb", mid="white")
```



This plot shows the average SAT Verbal Score based on the proportion of total expenditure that was put into instruction expenditure. From the graph, the highest SAT Verbal scores were states that put in the most amount of money (and also had the lowest SAT participation rates).

```
#Plot to show the relation between SAT Verbal scores and Instruction Expenditure as a percentage of Total Ex
penditure, mapped against the SAT Participation Level aesthetic
ggplot(States.math.means, aes(mean_dollars, mean_SATM)) +
geom_point(aes(color = region)) + geom_smooth(method = lm, col= "light blue") +
ylab("Mean SAT Math score per region") + xlab("Mean State Spending (in $1000s per student)") + ggtitle("Mean
SAT Math Score vs Mean State Spending \n on Public Eduction")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



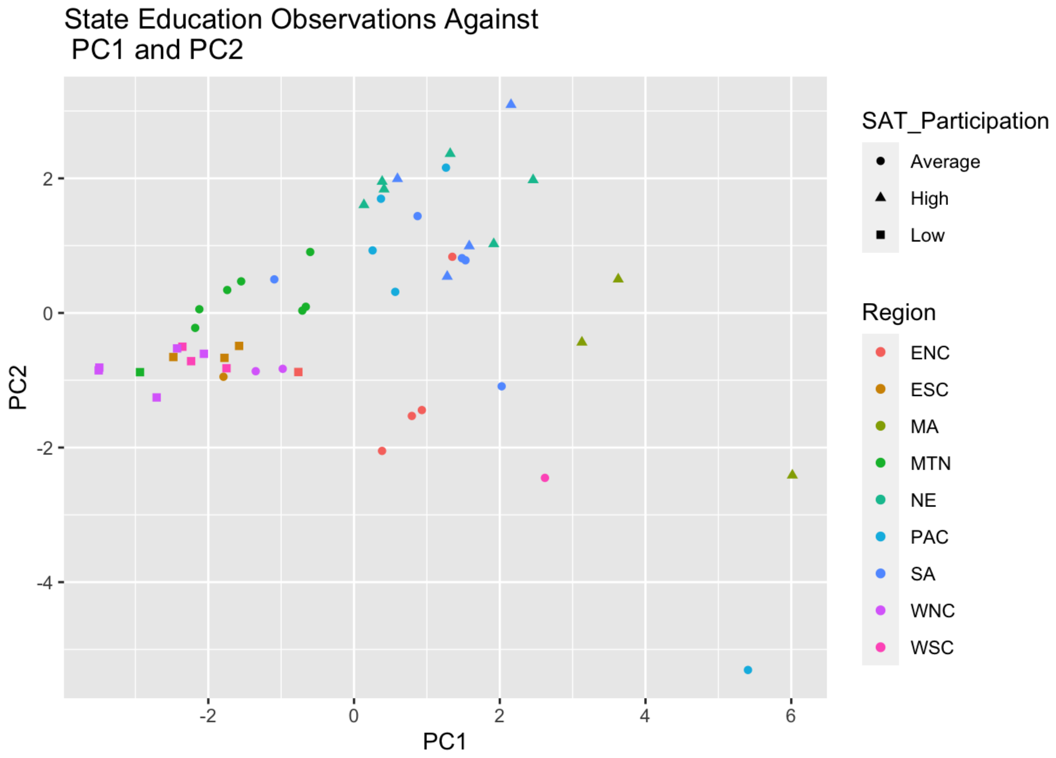
This graph depicts a weak negative correlation between mean state spending in $1000s per student for each region and the mean SAT math score per region. The more money put into public education per student, the lower the SAT Math scores tended to be that year. For example, region MA showed the highest public education spending but the lowest mean scores on the SAT Math portion.

Principal Component Analysis/Clustering

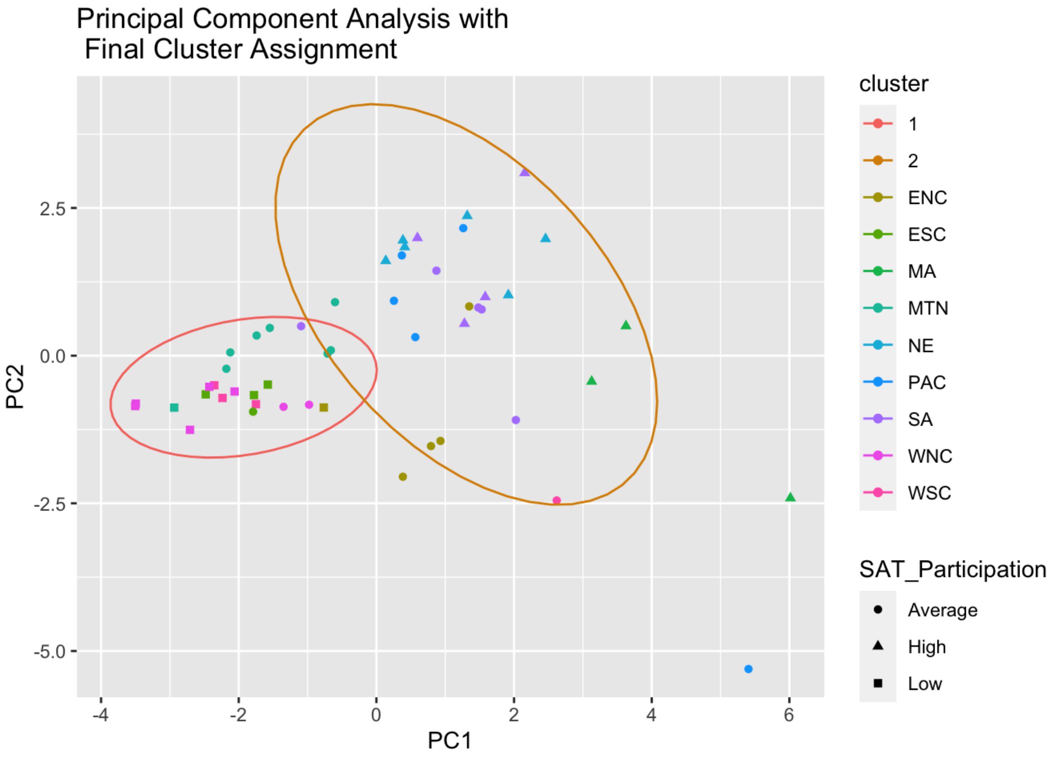
```
### PCA
#Only keep numeric variables and scale
States.edu3.scaled <- States.edu3.num %>%
  mutate_if(is.numeric, scale)

#Perform PCA
States.edu3.pca <- States.edu3.scaled %>%
  # Scale to 0 mean and unit variance (standardize)
  #scale() %>%
  prcomp()

#Find the percentage of variance explained by PCs
summary(States.edu3.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1488 1.4794 0.9450 0.39062 0.31289 0.17368 0.13210
## Proportion of Variance 0.5771 0.2736 0.1116 0.01907 0.01224 0.00377 0.00218
## Cumulative Proportion  0.5771 0.8507 0.9624 0.98143 0.99367 0.99744 0.99962
##                            PC8
## Standard deviation     0.05502
## Proportion of Variance 0.00038
## Cumulative Proportion  1.00000
```

```
#Save matrix x from PCA as data frame with region column and SAT participation
pca_data <- data.frame(States.edu3.pca$x, Region = States.edu3$region, SAT_Participation = States.edu3$SAT_Participation)

#plot along PC1 and PC2
pca_data %>%
  ggplot(aes(PC1, PC2)) + geom_point(aes(color = Region, shape=SAT_Participation)) + ggtitle("State Education Observations Against \n PC1 and PC2")
```



```
### PAM CLUSTERING
#Use pam clustering to identify clustering of State education data
library(cluster)

pam1 <- pca_data %>%
  select(PC1, PC2) %>%
  pam(k=2)

# Mutate cluster appropriately
pamclust <- pca_data %>%
  mutate(cluster = as.factor(pam1$clustering))

# Make a plot of data colored by final cluster assignment
pamclust %>%
  ggplot(aes(PC1, PC2, color = cluster)) +
  geom_point(aes(color=Region, shape = SAT_Participation)) + stat_ellipse(aes(group = cluster)) + ggtitle("Principal Component Analysis with \n Final Cluster Assignment")
```



Through this PCA, I was able to determine the ideal number of PCs to use (two). This is from the rule of thumb that we should only keep PCs until the cumulative proportion of variance is greater than 80%. PC1 accounts for about 57.7% of the total variance in the dataset, while PC2 accounts for about 27.4%.

Through PAM clustering, it can also be seen that the data falls in two clusters, with regions like PAC and SA tending to fall in the same cluster and states with high SAT Participation Levels tending to fall in the same cluster as well. This would make sense, as states with similar SAT participation levels tend to perform in a similar fashion among all other variables (like SAT and other test scores). The same goes when looking at regions as well. This goes to show that factors like quality of education and education expenditure are not evenly distributed across the United States, as one would hope is the case in order to ensure equitable education no matter where a student is in the United States.

References: Garrard, R. (2020, April 13). U.S. education Datasets: Unification project. Retrieved March 22, 2021, from https://www.kaggle.com/noriuk/us-education-datasets-unification-project

List of State Abbreviations. (n.d.). Retrieved March 22, 2021, from https://worldpopulationreview.com/states/state-abbreviations

```
```