PITDSC Project

2023-01-16

The purpose of this project was to determine what variables provide strong evidence for school performance, and was presented to the City of New York. The project looked into a variety of variables, such as neighborhood, income, test scores, survey responses, internet speeds, etc. This project was split into 5 parts: Data Cleaning, Linear Regression, Correlation, Principal Component Analysis (PCA), Outliers, and Map. Only portions of each section have been included as examples. From the study, poverty was found to be the strongest indicator of assessment performance across schools. Outliers were found and proposed to the hearing committee as examples of positively performing schools despite restricted finances. The key characteristic discussed was emphasis on English Language education. The specific outliers had this similarity in place due to a variety of reasons: high proportion of immigrant students, disabilities program, etc. The key takeaway from this was that high performing but low funded schools were able to sustain performance on the level of their well funded counterparts in part because of a school-wide emphasis on learning equity for students with a non-traditional learning experience, specifically those not from New York or those with disabilities impeding the learning experience. The project proposed an emphasis on a more balanced approach to test score evaluation: underperformance can be more readily addressed by focusing on students who may have disadvantages rather than solely focusing on high scores. Schools with policies supporting these students appear to have trickle down effects to native, English speakers, however further research is needed to confirm this.

1. Packages

```
library(factoextra)
```

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

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✔ tibble  3.1.8     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
## ✔ purrr   0.3.4
## ── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(Hmisc)
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
##
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(sjPlot)
```

```
## Install package "strengejacke" from GitHub (`devtools::install_github("strengejacke/strengejacke")`) to load a
ll sj-packages at once!
```

```
library(dplyr)
library(Rcpp)
```

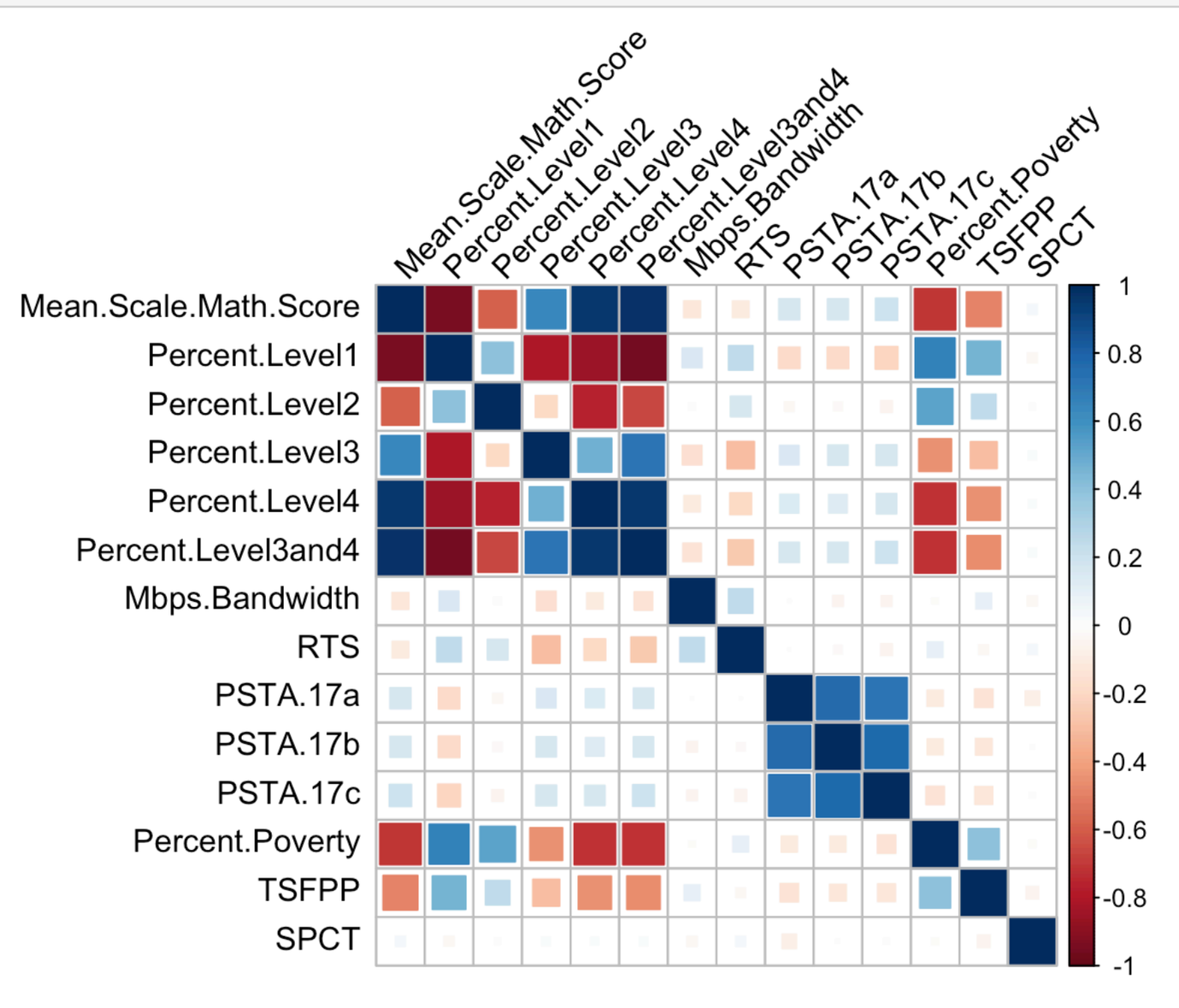
2. Data Cleaning For presentation, this section has been omitted. A master dataset was used by merging the three levels of education
3. Linear Regression (Master)

```
#load in our data
df <- read.csv("/Users/nathaniellowe/Library/CloudStorage/GoogleDrive-thenathaniellowe@gmail.com/My Drive/Archive
/PITDSC/PIT-DSC-STEM-Inequality/data/2018_Master_Data.csv") %>% select(c(9:15,18,20:23,29:30))

#rename columns
df <- df %>% rename(RTS = Ratio.of.Full.Time.Licensed.STEM.Teachers.to.Students,
                    PSTA.17a = Percent.of.STEMTeachers.That.Agree.With.17a,
                    PSTA.17b = Percent.of.STEMTeachers.That.Agree.With.17b,
                    PSTA.17c = Percent.of.STEMTeachers.That.Agree.With.17c,
                    TSFPP = Total.School.Funding.per.Pupil,
                    SPCT = Salary.Per.Classroom.Teacher,
                    Percent.Level1 = X.Level1,
                    Percent.Level2 = X.Level2,
                    Percent.Level3 = X.Level3,
                    Percent.Level4 = X.Level4,
                    Percent.Level3and4 = X.Level3.4)

#scaled the dataframe, then turned it back to a dataframe for linear regression analysis
df <- data.frame(scale(df))

#linear regression on percent math scores and mean math scores
mathscore <- lm( Mean.Scale.Math.Score ~ Mbps.Bandwidth + RTS + PSTA.17a + PSTA.17b + PSTA.17c + Percent.Poverty
                + TSFPP + SPCT, df)

percentOne <- lm(Percent.Level1 ~  Mbps.Bandwidth + RTS + PSTA.17a + PSTA.17b + PSTA.17c + Percent.Poverty
                + TSFPP + SPCT, df)


percentTwo <- lm(Percent.Level2 ~  Mbps.Bandwidth + RTS + PSTA.17a + PSTA.17b + PSTA.17c + Percent.Poverty
                + TSFPP + SPCT, df)

percentThree <- lm(Percent.Level3 ~  Mbps.Bandwidth + RTS + PSTA.17a + PSTA.17b + PSTA.17c + Percent.Poverty
                + TSFPP + SPCT, df)

percentFour <- lm(Percent.Level4 ~  Mbps.Bandwidth + RTS + PSTA.17a + PSTA.17b + PSTA.17c + Percent.Poverty
                + TSFPP + SPCT, df)
percentThreeFour <- lm(Percent.Level3and4 ~  Mbps.Bandwidth + RTS + PSTA.17a + PSTA.17b + PSTA.17c + Percent.Pove
rty
                + TSFPP + SPCT, df)

#showcase the results with tab_model
tab_model(mathscore, percentOne, percentTwo, percentThree, percentFour, percentThreeFour)
```

	Mean Scale Math Score			Percent Level 1			Percent Level 2			Percent Level 3			Percent Level 4			Percent Level 3 and 4		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	0.01	-0.04 – 0.07	0.594	-0.01	-0.06 – 0.05	0.849	0.00	-0.06 – 0.07	0.905	-0.01	-0.07 – 0.05	0.761	0.01	-0.04 – 0.06	0.758	0.00	-0.05 – 0.05	0.905
Mbps Bandwidth	-0.07	-0.12 – -0.02	**0.006**	0.08	0.02 – 0.13	**0.004**	-0.01	-0.08 – 0.05	0.702	-0.07	-0.14 – -0.01	**0.026**	-0.04	-0.10 – 0.01	0.094	-0.06	-0.11 – -0.01	**0.017**
RTS	-0.03	-0.08 – 0.02	0.207	0.16	0.11 – 0.21	**<0.001**	0.13	0.07 – 0.19	**<0.001**	-0.23	-0.29 – -0.17	**<0.001**	-0.12	-0.17 – -0.07	**<0.001**	-0.17	-0.22 – -0.13	**<0.001**
PSTA 17a	0.03	-0.05 – 0.11	0.458	-0.04	-0.12 – 0.05	0.371	0.00	-0.10 – 0.11	0.935	0.02	-0.08 – 0.12	0.683	0.03	-0.05 – 0.11	0.482	0.03	-0.05 – 0.11	0.446
PSTA 17b	0.04	-0.06 – 0.13	0.451	-0.04	-0.14 – 0.05	0.385	0.01	-0.11 – 0.13	0.884	0.08	-0.04 – 0.19	0.193	0.01	-0.09 – 0.10	0.875	0.03	-0.06 – 0.12	0.475
PSTA 17c	0.04	-0.05 – 0.12	0.418	-0.04	-0.13 – 0.05	0.422	0.01	-0.10 – 0.12	0.897	0.02	-0.09 – 0.12	0.754	0.03	-0.06 – 0.11	0.533	0.03	-0.05 – 0.11	0.509
Percent Poverty	-0.59	-0.65 – -0.53	**<0.001**	0.56	0.49 – 0.62	**<0.001**	0.46	0.38 – 0.54	**<0.001**	-0.39	-0.46 – -0.31	**<0.001**	-0.60	-0.66 – -0.54	**<0.001**	-0.61	-0.67 – -0.55	**<0.001**
TSFPP	-0.30	-0.35 – -0.24	**<0.001**	0.28	0.22 – 0.34	**<0.001**	0.10	0.02 – 0.17	**0.009**	-0.18	-0.25 – -0.11	**<0.001**	-0.25	-0.31 – -0.19	**<0.001**	-0.26	-0.32 – -0.21	**<0.001**
SPCT	0.02	-0.04 – 0.08	0.507	-0.01	-0.07 – 0.06	0.775	-0.00	-0.08 – 0.08	0.943	0.00	-0.07 – 0.08	0.927	0.01	-0.05 – 0.07	0.765	0.01	-0.05 – 0.07	0.772
Observations	683			683			683			683			683			683		
R^2 / R^2 adjusted	0.553 / 0.548			0.528 / 0.522			0.250 / 0.241			0.328 / 0.321			0.532 / 0.527			0.583 / 0.578		

2. Correlation (Master)

```
#load dataset
df <- read.csv("/Users/nathaniellowe/Library/CloudStorage/GoogleDrive-thenathaniellowe@gmail.com/My Drive/Archive
/PITDSC/PIT-DSC-STEM-Inequality/data/2018_Master_Data.csv") %>%
    select(c(9:15,18,20:23,29:30))

#rename columns
df <- df %>% rename(RTS = Ratio.of.Full.Time.Licensed.STEM.Teachers.to.Students,
                    PSTA.17a = Percent.of.STEMTeachers.That.Agree.With.17a,
                    PSTA.17b = Percent.of.STEMTeachers.That.Agree.With.17b,
                    PSTA.17c = Percent.of.STEMTeachers.That.Agree.With.17c,
                    TSFPP = Total.School.Funding.per.Pupil,
                    SPCT = Salary.Per.Classroom.Teacher,
                    Percent.Level1 = X.Level1,
                    Percent.Level2 = X.Level2,
                    Percent.Level3 = X.Level3,
                    Percent.Level4 = X.Level4,
                    Percent.Level3and4 = X.Level3.4)

#want to turn data frame to matrix so that rcorr() works
M <- rcorr(as.matrix(df), type = "pearson")

#we want to plot the corr matrix, so we do M$r
M$r %>% cor_plot(method = "square")
```

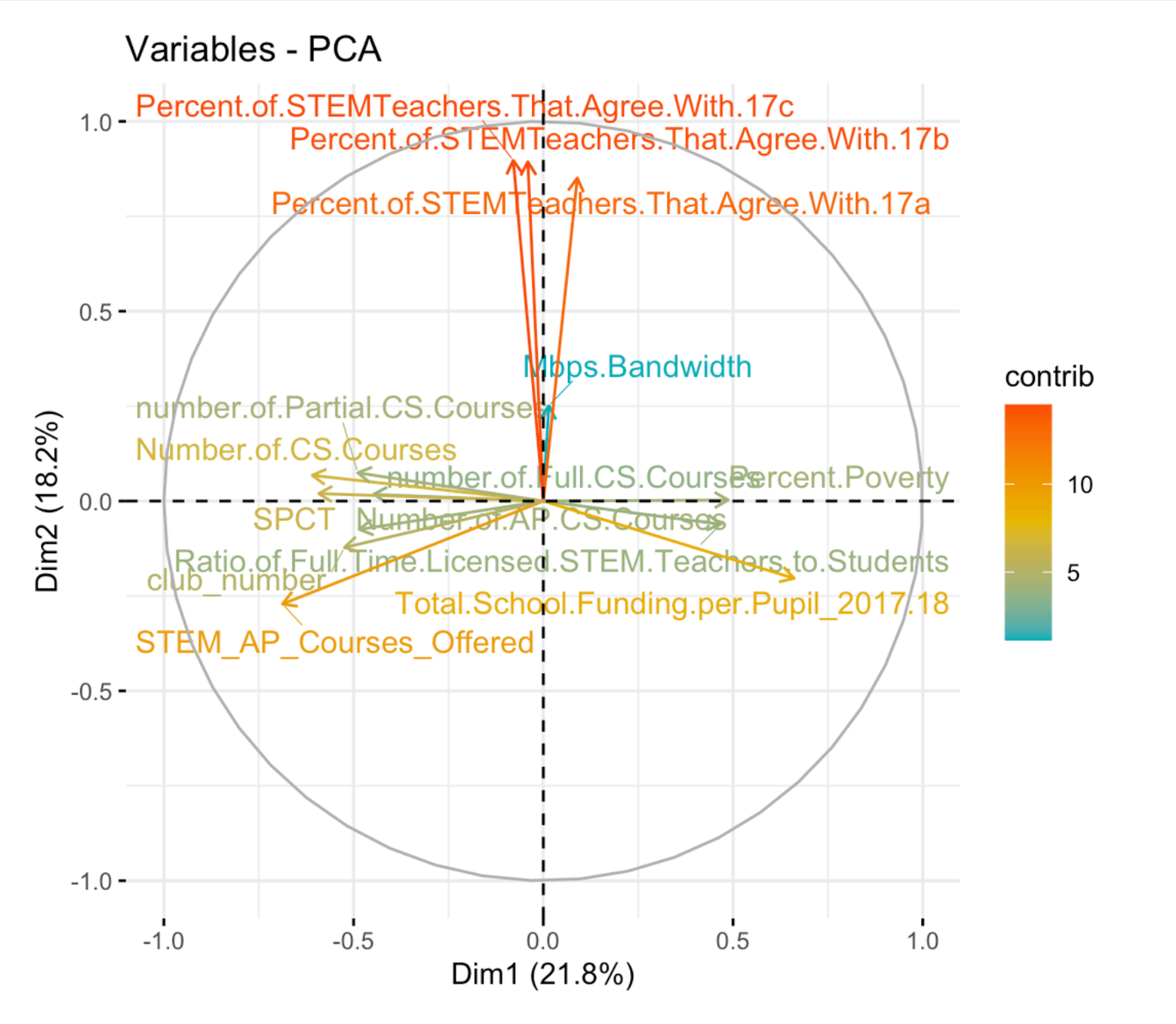


3. PCA (High School only)

```
#loads in dataset
hs <- read.csv("/Users/nathaniellowe/Library/CloudStorage/GoogleDrive-thenathaniellowe@gmail.com/My Drive/Archive
/PITDSC/PIT-DSC-STEM-Inequality/data/2018_HS_Data.csv")
#add colum SPCT (salary per classroom teacher)
hs <- hs %>% mutate(SPCT = Classroom.Teachers_2017.18 / (Classroom.Teachers.w..0.3.Years.Experience_2017.18 +
                                                          Classroom.Teachers.w..More.than.3.Years.Experience_2017.18
) )


hs <- hs %>% select(-c(1:27,29:30,36,40:44,46))
#take away the rows with NA (can't do PCA otherwise)
hs <- hs %>% na.omit()

#compute PCA
hs.pca <- prcomp(hs, center = TRUE, scale = TRUE)

#showcase the result
summary(hs.pca)
```

```
## Importance of components:
##                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.747 1.5941 1.3816 1.13396 1.03162 0.97009 0.88507
## Proportion of Variance 0.218 0.1815 0.1363 0.09185 0.07602 0.06722 0.05595
## Cumulative Proportion  0.218 0.3995 0.5359 0.62773 0.70375 0.77097 0.82692
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.76107 0.72089 0.69045 0.58765 0.54664 0.45091
## Proportion of Variance 0.04137 0.03712 0.03405 0.02467 0.02134 0.01452
## Cumulative Proportion  0.86829 0.90542 0.93947 0.96413 0.98548 1.00000
##                             PC14
## Standard deviation     4.684e-16
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

```
#graph/visualize the variables
fviz_pca_var(hs.pca,
             col.var = "contrib",
             gradient.cols = c("#00AFBB","#E7B800", "#FC4E07"),
             repel = TRUE )
```



4. Outliers (HS only)

```
#outliers of our dataset
#msle and percent poverty

#load in our data
hs_data <- read.csv("/Users/nathaniellowe/Library/CloudStorage/GoogleDrive-thenathaniellowe@gmail.com/My Drive/Ar
chive/PITDSC/PIT-DSC-STEM-Inequality/data/2018_HS_Data.csv")


#mean/sd calculations for STEM tests

hs_mean_test_Algebra <- mean(hs_data$Mean.Score_Common.Core.Algebra, na.rm = TRUE) %>% round(digits = 2)
hs_sd_test_Algebra <-  sd(hs_data$Mean.Score_Common.Core.Algebra, na.rm = TRUE) %>% round(digits = 2)

hs_mean_test_Algebra2 <- mean(hs_data$Mean.Score_Common.Core.Algebra2, na.rm = TRUE) %>% round(digits = 2)
hs_sd_test_Algebra2 <-  sd(hs_data$Mean.Score_Common.Core.Algebra2, na.rm = TRUE) %>% round(digits = 2)

hs_mean_test_Geometry <- mean(hs_data$Mean.Score_Common.Core.Geometry, na.rm = TRUE) %>% round(digits = 2)
hs_sd_test_Geometry <-  sd(hs_data$Mean.Score_Common.Core.Geometry, na.rm = TRUE) %>% round(digits = 2)


hs_mean_test_LE <- mean(hs_data$Mean.Score_Living.Environment, na.rm = TRUE) %>% round(digits = 2)
hs_sd_test_LE <-  sd(hs_data$Mean.Score_Living.Environment, na.rm = TRUE) %>% round(digits = 2)

hs_mean_test_ES <- mean(hs_data$Mean.Score_Physical.Settings.Earth.Science, na.rm = TRUE) %>% round(digits = 2)
hs_sd_test_ES <-  sd(hs_data$Mean.Score_Physical.Settings.Earth.Science, na.rm = TRUE) %>% round(digits = 2)

hs_mean_test_Chemistry <- mean(hs_data$Mean.Score_Physical.Settings.Chemistry, na.rm = TRUE) %>% round(digits = 2
)
hs_sd_test_Chemistry <-  sd(hs_data$Mean.Score_Physical.Settings.Chemistry, na.rm = TRUE) %>% round(digits = 2)

hs_mean_test_Physics <- mean(hs_data$Mean.Score_Physical.Settings.Physics, na.rm = TRUE) %>% round(digits = 2)
hs_sd_test_Physics <-  sd(hs_data$Mean.Score_Physical.Settings.Physics, na.rm = TRUE) %>% round(digits = 2)



#mean/sd for poverty
hs_mean_pov <- mean(hs_data$Percent.Poverty, na.rm = TRUE) %>% round(digits = 2)
hs_sd_pov <- sd(hs_data$Percent.Poverty, na.rm = TRUE) %>% round(digits = 2)


#1 sd above the mean for STEM tests
one_sd_test_Algebra <- hs_mean_test_Algebra + 1*hs_sd_test_Algebra
one_sd_test_Algebra2 <- hs_mean_test_Algebra2 + 1*hs_sd_test_Algebra2
one_sd_test_Geometry <- hs_mean_test_Geometry + 1*hs_sd_test_Geometry
one_sd_test_LE <- hs_mean_test_LE + 1*hs_sd_test_LE
one_sd_test_ES <- hs_mean_test_ES + 1*hs_sd_test_ES
one_sd_test_Chemistry <- hs_mean_test_Chemistry + 1*hs_sd_test_Chemistry
one_sd_test_Physics <- hs_mean_test_Physics + 1*hs_sd_test_Physics


#1 sd above the mean for poverty
one_sd_pov_hs<- hs_mean_pov + 1*hs_sd_pov


#filter to get the outliers that we want
outlier_hs_Algebra <- hs_data %>% filter(Percent.Poverty >= one_sd_pov_hs & Mean.Score_Common.Core.Algebra >= one
_sd_test_Algebra)
outlier_hs_Algebra2 <- hs_data %>% filter(Percent.Poverty >= one_sd_pov_hs & Mean.Score_Common.Core.Algebra2 >= o
ne_sd_test_Algebra2)
outlier_hs_Geometry <- hs_data %>% filter(Percent.Poverty >= one_sd_pov_hs & Mean.Score_Common.Core.Geometry >= o
ne_sd_test_Geometry)
outlier_hs_LE <- hs_data %>% filter(Percent.Poverty >= one_sd_pov_hs & Mean.Score_Living.Environment >= one_sd_te
st_LE)
outlier_ES <- hs_data %>% filter(Percent.Poverty >= one_sd_pov_hs & Mean.Score_Physical.Settings.Earth.Science >=
one_sd_test_ES)
outlier_Chemistry <- hs_data %>% filter(Percent.Poverty >= one_sd_pov_hs & Mean.Score_Physical.Settings.Chemistry
>= one_sd_test_Chemistry)
outlier_hs_Physics <- hs_data %>% filter(Percent.Poverty >= one_sd_pov_hs & Mean.Score_Physical.Settings.Physics
>= one_sd_test_Physics)

#create a dataset with all the outlier data with schools
outlier_data1 <- outlier_Chemistry %>% full_join(outlier_hs_Algebra)
```

```
## Joining, by = c("dbn", "school_name", "Borough", "grade_span", "Latitude",
## "Longitude", "Total.Tested_Common.Core.Algebra",
## "Mean.Score_Common.Core.Algebra",
## "Percent.Students.Took.The.Test.For.CCAlgebra",
## "Total.Tested_Common.Core.Algebra2", "Mean.Score_Common.Core.Algebra2",
## "Percent.Students.Took.The.Test.For.CCAlgebra2",
## "Total.Tested_Common.Core.Geometry", "Mean.Score_Common.Core.Geometry",
## "Percent.Students.Took.The.Test.For.CCGeometry",
## "Total.Tested_Living.Environment", "Mean.Score_Living.Environment",
## "Percent.Students.Took.The.Test.For.LE",
## "Total.Tested_Physical.Settings.Earth.Science",
## "Mean.Score_Physical.Settings.Earth.Science",
## "Percent.Students.Took.The.Test.For.ES",
## "Total.Tested_Physical.Settings.Chemistry",
## "Mean.Score_Physical.Settings.Chemistry",
## "Percent.Students.Took.The.Test.For.Chemistry",
## "Total.Tested_Physical.Settings.Physics",
## "Mean.Score_Physical.Settings.Physics",
## "Percent.Students.Took.The.Test.For.Physics", "Mbps.Bandwidth",
## "Full.Time.Certified.STEM.Teachers", "Part.Time.Certified.STEM.Teachers",
## "Ratio.of.Full.Time.Licensed.STEM.Teachers.to.Students",
## "Number.of.CS.Courses", "Number.of.AP.CS.Courses", "number.of.Full.CS.Courses",
## "number.of.Partial.CS.Courses", "Total.Teacher.Response.Rate",
## "Percent.of.STEMTeachers.That.Agree.With.17a",
## "Percent.of.STEMTeachers.That.Agree.With.17b",
## "Percent.of.STEMTeachers.That.Agree.With.17c", "K.12.Enrollment_2017.18",
## "K.12.FRPL.Count_2017.18",
## "Classroom.Teachers.w..0.3.Years.Experience_2017.18",
## "Classroom.Teachers.w..More.than.3.Years.Experience_2017.18",
## "Classroom.Teachers_2017.18", "Total.School.Funding.per.Pupil_2017.18",
## "School_Type", "Percent.Poverty", "STEM_AP_Courses_Offered", "club_number")
```

```
outlier_data <- outlier_data1 %>%  full_join(outlier_hs_Algebra2)
```


```
## Joining, by = c("dbn", "school_name", "Borough", "grade_span", "Latitude",
## "Longitude", "Total.Tested_Common.Core.Algebra",
## "Mean.Score_Common.Core.Algebra",
## "Percent.Students.Took.The.Test.For.CCAlgebra",
## "Total.Tested_Common.Core.Algebra2", "Mean.Score_Common.Core.Algebra2",
## "Percent.Students.Took.The.Test.For.CCAlgebra2",
## "Total.Tested_Common.Core.Geometry", "Mean.Score_Common.Core.Geometry",
## "Percent.Students.Took.The.Test.For.CCGeometry",
## "Total.Tested_Living.Environment", "Mean.Score_Living.Environment",
## "Percent.Students.Took.The.Test.For.LE",
## "Total.Tested_Physical.Settings.Earth.Science",
## "Mean.Score_Physical.Settings.Earth.Science",
## "Percent.Students.Took.The.Test.For.ES",
## "Total.Tested_Physical.Settings.Chemistry",
## "Mean.Score_Physical.Settings.Chemistry",
## "Percent.Students.Took.The.Test.For.Chemistry",
## "Total.Tested_Physical.Settings.Physics",
## "Mean.Score_Physical.Settings.Physics",
## "Percent.Students.Took.The.Test.For.Physics", "Mbps.Bandwidth",
## "Full.Time.Certified.STEM.Teachers", "Part.Time.Certified.STEM.Teachers",
## "Ratio.of.Full.Time.Licensed.STEM.Teachers.to.Students",
## "Number.of.CS.Courses", "Number.of.AP.CS.Courses", "number.of.Full.CS.Courses",
## "number.of.Partial.CS.Courses", "Total.Teacher.Response.Rate",
## "Percent.of.STEMTeachers.That.Agree.With.17a",
## "Percent.of.STEMTeachers.That.Agree.With.17b",
## "Percent.of.STEMTeachers.That.Agree.With.17c", "K.12.Enrollment_2017.18",
## "K.12.FRPL.Count_2017.18",
## "Classroom.Teachers.w..0.3.Years.Experience_2017.18",
## "Classroom.Teachers.w..More.than.3.Years.Experience_2017.18",
## "Classroom.Teachers_2017.18", "Total.School.Funding.per.Pupil_2017.18",
## "School_Type", "Percent.Poverty", "STEM_AP_Courses_Offered", "club_number")
```


```
#now want to include/exclude some columns
outlier_data <- outlier_data %>% dplyr::select(c(1:6))
outlier_data$Address <- NA
outlier_data$Website <- NA

#add the addresses and website, since only 3 outlier schools I will enter it manually
outlier_data$Website[1] <- "https://www.languageandinnovation.org/"
outlier_data$Website[2] <- "https://www.internationalcommunityhs.org/"
outlier_data$Website[3] <- "https://www.thewae.org/"

outlier_data$Address[1] <- "925 Astor Ave, Bronx, NY 10469"
outlier_data$Address[2] <- "345 Brook Avenue, Bronx, NY 10454"
outlier_data$Address[3] <- "456 White Plains Road, Third Floor, Bronx, NY 10473"
```

5. Map See https://thenml.shinyapps.io/webApp/