Public School Data Analysis: Expenditure, Grades, and Retention in New York State

Maya Mileva | Stefano Selenu | Meng Synn

FINAL PROJECT

IST 687 Applied Data Science

March 28, 2019

Prof. J. Saltz and M. Syed

1) Introduction

Although the explosion of data accumulation in public schools and universities has increased the demand for people who understand data and its potential in the educational field, many obstacles still remain for educational data scientists. In a general sense, the majority of people might agree that data, used the right way, is knowledge, but, in the K-12 world, data scientists may have to fight to prove their worth. "While the business community has invested in data as a driver of success, many educators feel lukewarm about it."[1] Not rarely, in fact, the collecting of data has worked – as the Center for Digital Education director Kecia Ray pointed out – as an instrument to penalize educators and "a way to shut down schools and fire superintendents."[2]

This project grew up from the opposite spirit. That is, from the idea that data analysis and interpretation can help schools to meet high marks and can encourage educators and administrators to view old problems in new ways. With this in mind, in this project we deal with a dataset containing aggregated information on US K-12 Education since 1992. This dataset is designed to bring together multiple facets of U.S. education data into one CSV file format. Different versions of the database are available online at https://www.kaggle.com/noriuk/us-education-datasets-unification-project. Given that we had multiple options on the ways we could look at US K-12 Education through different databases, we decided to reshape and model our own data as reported in the section 3.3 Data Structure and Classification and Appendix 1 here below.

2) Problem Definition, Business Questions, and Methodology

The purpose of this project is to identify relationships among expenditure, enrollment, and student achievement (math and reading scores) in public schools in different states and years in order to advance some actionable insights in particular for the state of New York. In order to have a better point of observation for New York state, we examined two other states in particular, Massachusetts (the state with the highest total average score for math and reading in the US), and Florida (the state with the lowest total average score for math and reading in the US).

The following general questions have driven our investigation:

[1] Quoted from Adam Stone, "Will Data Scientists Have a Big Impact on Education?", https://www.govtech.com/education/k-12/Will-Data-Scientists-Have-a-Big-Impact-on-Education.html

[2] Ibid.

- What are the relationships among expenditure, retention, and average scores (math and reading) in different states and years?
- What factors influence student retention/attrition in public high schools?
- To what extent do the ratios between total expenditure and instruction expenses affect student enrollment after 8th grade?
- Do student enrollment after 8th grade correlate to the math/reading scores in that grade?
- Are there any specific trends observable within the most recent 5 years?

As for the methodology, our team used RMarkdown in RStudio, the programming language and software environment for statistical analysis, data visualization, and reporting. We used RMarkdown to write and run all of the code for the project. Due to the large amount of code accumulated, we have decided to submit it as a separate file that can be opened directly in RStudio. Please see the attached .R file, which is also accessible by clicking on the R icon in the Appendix 3.

3) Data Acquisition, Munging, and Classification

3.1 Data Acquisition

The first step in our project consisted in downloading the database file (in csv format) into our local hard drive and then in importing the data into RStudio. We used the read.csv command and stored the full dataset into the dfUSEducation variable. We first decided to analyze the full dataset after cleaning it, then we reorganized the full dataset into subsets.

We are using the following codes to create the data frame:

```
# Create dataframe from states_all_extended.csv file
dfUSEducation <- read.csv("../USEducationDataset/states_all_extended.csv",stringsAsFactors = FALSE)
```

3.2 Data cleaning and munging

The dataset was overall in good shape, but we had to make some decisions with regard to data munging. First, we decided to focus only on the entries where the data was as complete as possible and to ensure that every insight was actually data-driven. For that reason, we decided to replace the NAs with 0s. We chose not to fulfil blanks with an average value because we considered that such a choice could have affected our final results. For this reason, our team deemed that, in order to keep important data included in the rows with few NAs, replacing the NAs with 0s was the best way to clean the data.

```
# Replace NAs with 0s
dfUSEducation[is.na(dfUSEducation)] <- 0
```

We also noticed that states with compound names as New York, New Mexico, and others, were indicated with the use of underscores. We decided to remove the underscores to uniform all state names.

STR_REPLACE_ALL function is used to clean up state names with the underscore:

```
# Remove "_" from STATE
dfUSEducation$STATE <- str_replace_all(dfUSEducation$STATE, '_', ' ')
```

Lastly, we decided to exclude Alaska and Hawaii from the list of states composing the US in order to work on the lower 48 states. This has helped us visualizing the data by means of maps and charts (see section 4.1 EDA with Maps and Charts).

We are utilizing SQLDF function to create the data frame for the lower 48 states:

```
# Get 9 years of good data for the lower 48 states from 1996 to 2015
dfUSEducation_ALL <- sqldf("SELECT * FROM dfUSEducation sa JOIN dfUSStates s
            ON sa.STATE = s.StateName")

# Exclude 1996,2000,2003,2005,2007 as they do not contain full spectrum of races and genders
dfUSEducation <- sqldf("SELECT * FROM dfUSEducation sa JOIN dfUSStates s
            ON sa.STATE = s.StateName
            AND sa.YEAR IN (2009,2011,2013,2015)")
```

3.3 Data Structure and Classification: A Short Summary

We started our exploratory data analysis (EDA) with a medium-sized database (state_all.csv) of around 10000 data points divided into 25 variables (columns) and 413 observations (rows). After cleaning the dataset and mapping the data, we realized that the type of variables contained in that dataset were not sufficient to achieve useful actionable insights. For that reason, we looked at different databases available in the same Kaggle kernel and we decided to utilize the extended version included in the state_all_extended.csv file. This database is composed of a total of 1492 observations and 193 variables in its uncleaned version. After cleaning the data, the dataset narrowed down to 342 observations and 193 variables, for a total of around 66,000 data points (see the Appendix 1 for a full list of all variables included in the state_all_extended.csv database). We had thus to proceed with further selections in order to focus and reach useful actionable insights.

Among the most important variables in the dataset we chose are the following.

	Variable Name	Meaning
1.	**"STATE"**	The name of the State in the United States
2.	**"YEAR"**	The year the data refers to. Years included are from 1996 to

		2015
3.	**"ENROLL"**	Total student enrollment
4.	**"TOTAL_REVENUE"**	Total revenue available to the public schools per year in a specific school
5.	**"FEDERAL_REVENUE"**	The revenue provided by the Federal government per year for each State
6.	**"STATE_REVENUE"**	The revenue provided by the State per year
7.	**"LOCAL_REVENUE"**	The revenue provided by the city per year in each State
8.	**"TOTAL_EXPENDITURE"**	Total expenses encountered yearly by all public schools in each State
9.	**"INSTRUCTION_EXPENDITURE"**	Total expenses for instruction encountered yearly by all public schools in each State
10.	**"SUPPORT_SERVICES_EXPENDITURE"**	Total expenses for support services encountered yearly by all public schools in each State
11.	**"OTHER_EXPENDITURE"**	Various other expenses encountered yearly by all public schools in each State
12.	**"CAPITAL_OUTLAY_EXPENDITURE"**	Expenses encountered yearly by all public schools in each State for capital outlay (that is, money spent to acquire, maintain, repair, or upgrade capital assets, which may include technology, land, facilities, or other business necessities that are not expended during normal use).
13.	**"GRADES_PK_G"**	Number of enrolled students in pre-kindergarten schools per year in each State
14.	**"GRADES_KG_G"**	Number of enrolled students in kindergarten schools per year in each State
15	**"GRADES_4_G"**	Number of enrolled students in public schools in the 4th grade per year in each State
16.	**"GRADES_8_G"**	Number of enrolled students in public schools in the 8th grade per year in each State
17.	**"GRADES_12_G"**	Number of enrolled students in public schools in the 12th grade per year in each State
18.	**"GRADES_1_8_G"**	Number of enrolled students in public schools between 1st and 8th grades per year in each State
19.	**"GRADES_9_12_G"**	Number of enrolled students in public schools between 9th and 12th grades per year in each State
20.	**"GRADES_ALL_G"**	Total number of enrolled students in public schools in all grades per year in each State
21.	**"AVG_MATH_4_SCORE"**	The average grade math scores in the 4th grade of all public schools per year in each State
22.	**"AVG_MATH_8_SCORE"**	The average grade math scores in the 8th grade of all public schools per year in each State

23.	**"AVG_READING_4_SCORE"**	The average grade reading scores in the fourth grade of all public schools per year in each State
24.	**"AVG_READING_8_SCORE"**	The average grade reading scores in the 8th grade of all public schools per year in each State

To help our reflections on the extended dataset, we also decided to group the variables into the following 8 main categories:

	Category Name	**Variables included**
1.	**STATE**	**STATE**
2.	**YEAR**	**YEAR**
3.	**TOTAL ENROLLMENT**	**ENROLL**
4.	**REVENUE**	**"TOTAL_REVENUE", "FEDERAL_REVENUE", "STATE_REVENUE", "LOCAL_REVENUE"**
5.	**EXPENDITURE**	**"TOTAL_EXPENDITURE", "INSTRUCTION_EXPENDITURE", "SUPPORT_SERVICES_EXPENDITURE", "OTHER_EXPENDITURE", "CAPITAL_OUTLAY_EXPENDITURE"**
6.	**SPECIFIC ENROLLMENT PER GRADE**	**"GRADES_PK_G", "GRADES_KG_G", "GRADES_4_G", "GRADES_8_G", "GRADES_12_G", "GRADES_1_8_G", "GRADES_9_12_G", "GRADES_ALL_G" and more in Appendix 1**
7.	**STUDENT DEMOGRAPHIC INFORMATION (RACE AND GENDER)**	**See Appendix 1**
8.	**ASSESSMENT (MATH AND READING SCORES)**	**"AVG_MATH_4_SCORE", "AVG_MATH_8_SCORE", "AVG_READING_4_SCORE", "AVG_READING_8_SCORE"**

- **Category 1 to 3** contain basic information on state, year, and total enrollment of students.
- **Category 4 and 5** includes financial information about total revenue and its specific divisions at the federal, state, and local levels as well as data about expenditure and its divisions among instruction, support services, capital outlay, and other expenditure. This data is of critical importance for our assumption of possible correlations among expenditure, score, and retention.
- **Category 6 to 8** instead provides information revolving around students: the total number of student enrollment, the specific number of student enrollment in different K-12 grades, student demographic information such as race and gender, and, finally, student scores in math and reading in different grades.

- For our data visualization and predictions we then organized the data into dependent and independent variables as follows:
  - Independent v.: state, year, revenue, and student demographic
  - Dependent v.: enrollment, expenditure, retention (that we named "dropout" in our dataset and plots), and assessment (that we named "score")

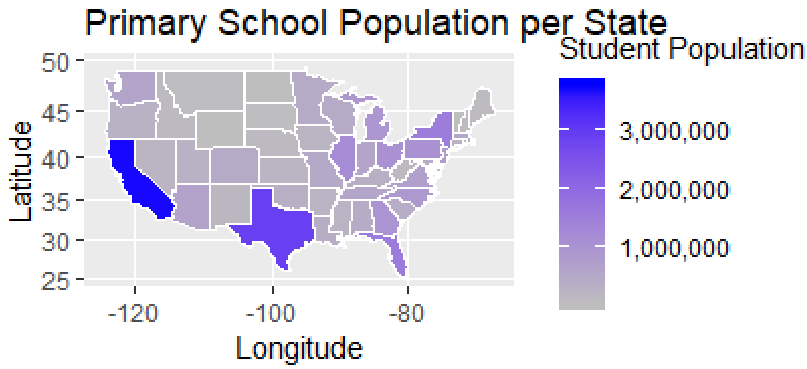
4) Descriptive Statistics

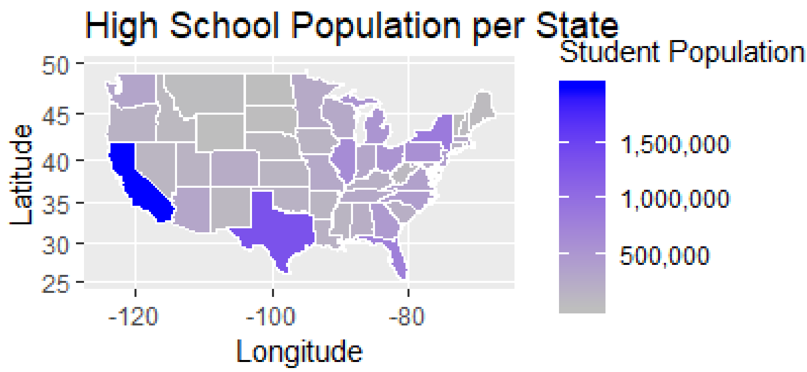
4.1 EDA with Maps and Charts

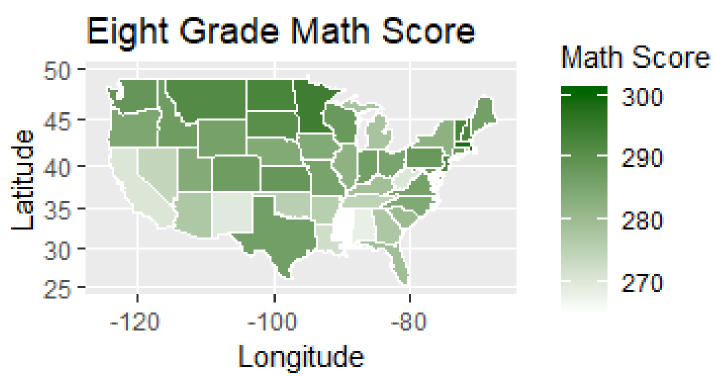
In the early stages of our project – more precisely after classifying the type of variables included in the dataset (see above in section 3) –, we furthered our exploratory analysis to gain some basic insights into our data. The following code snippets were used to create maps by means of GGPLOT, GGMAP, and GRID.ARRANGE packages:

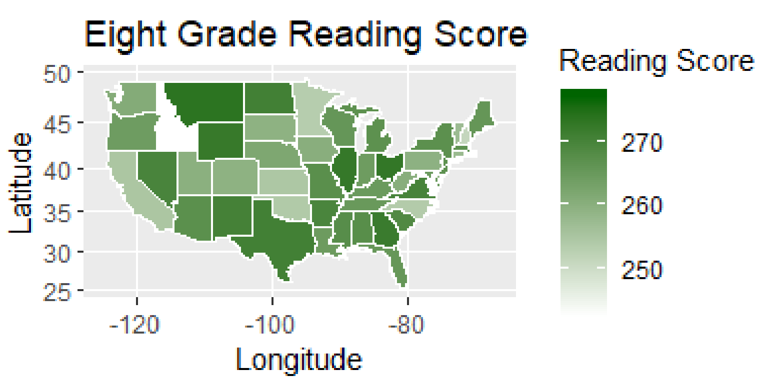
```
# Plot basic U.S. map using state dataframe
map.Primary <- ggplot(dfUSEducation, aes(map_id = STATE)) +
    geom_map(map = us, aes(fill=dfUSEducation$GRADES_1_8_G), color="white") +
    expand_limits(x = us$long, y = us$lat) +
    coord_map() +
    scale_fill_continuous(low = "gray", high = "blue", name = "Student Population", label = sca
les::comma) +
    theme(legend.position = "right") +
    ggtitle("Primary School Population per State") +
    labs(x= "Longitude") +
    labs(y= "Latitude")

# Plot basic U.S. map using state dataframe
map.HighSchool <- ggplot(dfUSEducation, aes(map_id = STATE)) +
    geom_map(map = us, aes(fill=dfUSEducation$GRADES_9_12_G), color="White") +
    expand_limits(x = us$long, y = us$lat) +
    coord_map() +
    scale_fill_continuous(low = "gray", high = "blue", name = "Student Population", label = sca
les::comma) +
    theme(legend.position = "right") +
    ggtitle("High School Population per State") +
    labs(x= "Longitude") +
    labs(y= "Latitude")

# Arrange the plots via grid.arrange
grid.arrange(map.Primary, map.HighSchool)
```

Through the codes above, we visualized different color-coded maps for the variables Primary/High school populations and Math/Reading scores for 8th grade.

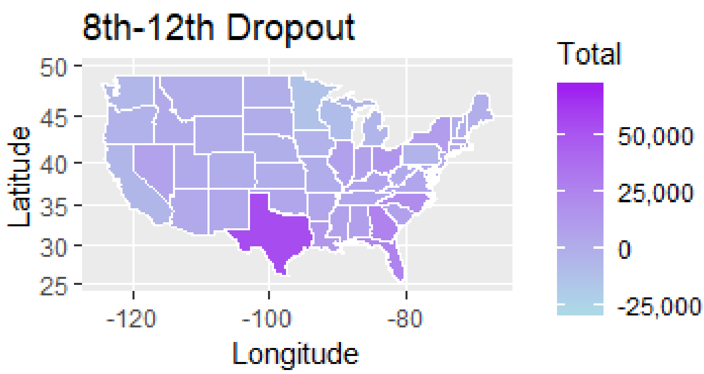

Primary School Population per State
Student Population
3,000,000
2,000,000
1,000,000
Latitude
50
45
40
35
30
25
-120
-100
-80
Longitude

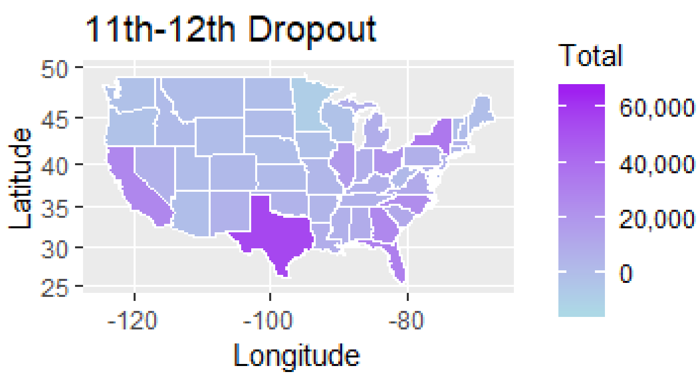

High School Population per State
Student Population
1,500,000
1,000,000
500,000
Latitude
50
45
40
35
30
25
-120
-100
-80
Longitude


Eight Grade Math Score
Math Score
300
290
280
270
Latitude
50
45
40
35
30
25
-120
-100
-80
Longitude

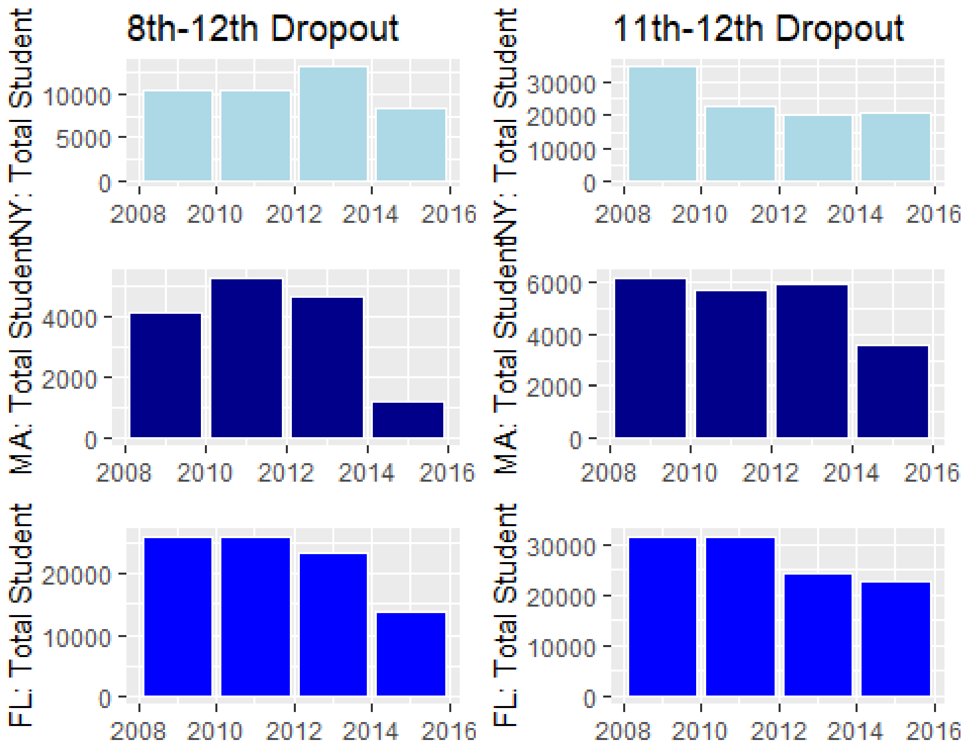

Eight Grade Reading Score
Reading Score
270
260
250
Latitude
50
45
40
35
30
25
-120
-100
-80
Longitude

Creating these and other visualizations (see below) was beneficial to make things clearer and easier to understand, especially with such a large and multi-dimensional dataset as ours. This helped for instance map the changes in enrollment between 8th and 12th grade in comparison with 11th and 12th grades. See, for instance, the following map about the enrollment changes between 8th and 12th grades and between 11th and 12th grade.



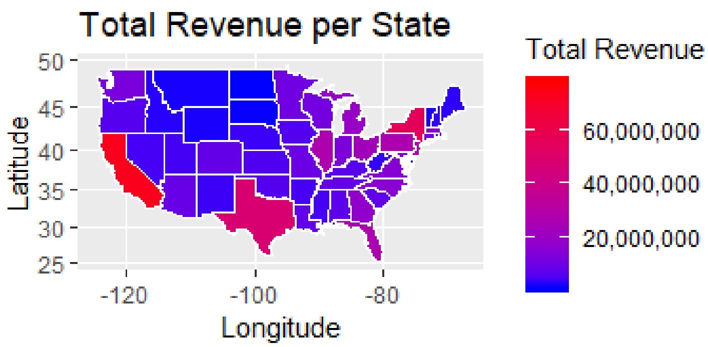


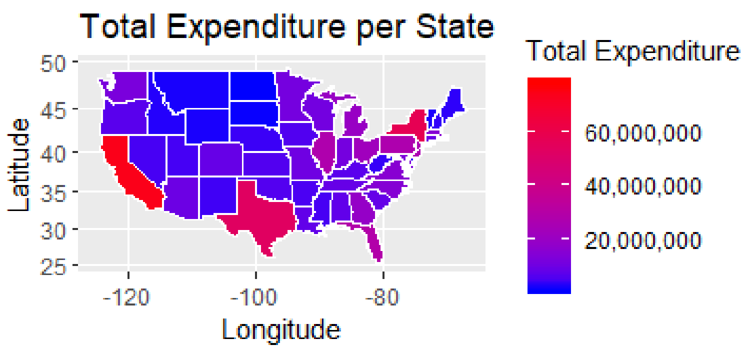
If we look at the map closely, we can see that the student attrition in New York state is evidently higher in the final year of high school enrollment (11th-12th grade). This observation is also confirmed by the comparative bar charts below with regard to NY state, Massachusetts, and Florida.



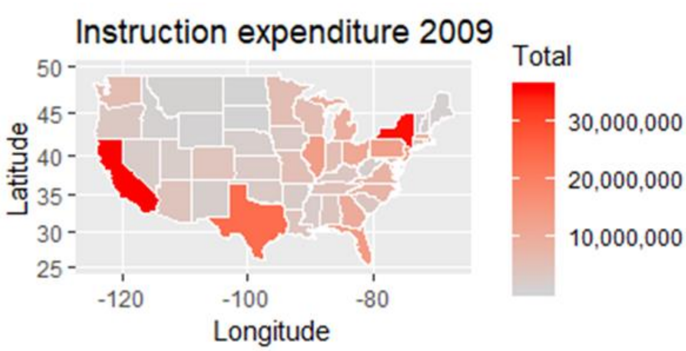
Moreover, these graphic representations allowed us to visualize first general assumptions about correlations between revenue and expenditure as the maps here below display.

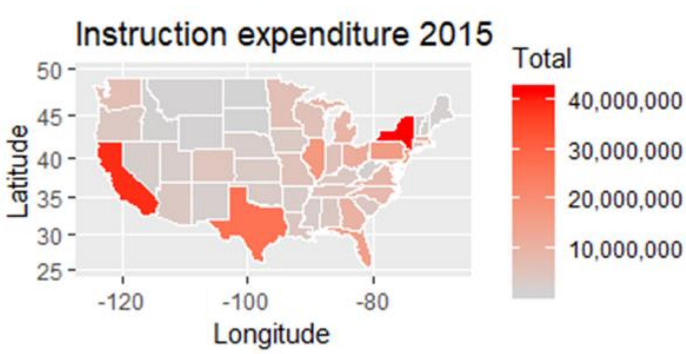
1) Total revenue/total expenditure





2) Instruction expenditure 2009 ~ 2015





Continuing our exploratory analysis, we then realized that the distribution of scores across states was not as easy to predict without a close look at the data. This in turn contributed to the ways we re-shaped the data into different subsets.

4.2 Reshaping the Data

Accurate analyses of large datasets require re-shaping the whole into smaller data subsets. After our general EAD of our database, we decided to center our analysis on New York as our target state in comparison with the states with the highest and lowest average math and reading scores, that is, Massachusetts and Florida respectively. We refocused on them because the database appears to provide most complete data points for all variables in those years. Since our business questions focus on assessment and retention in US high schools, we created subsets on the math and reading scores for 8th, 9th, and 9th-12th grades and on student attrition (or drop out) from 8th to 12th grade. We used the sqldf function included in the sqldf package. We also reshaped the datasets in chronological terms by recentering mostly on the 2009, 2011, 2013, and 2015 years.

In preparation of reshaping our data, we created a function called SelectDataByYear() that takes SQL string and year as parameters. This function then returns the data frame back to the caller:

```
# Create a function to return specific dataset
SelectDataByYear <- function(sql, year)
{ strSQL <- sql
 if (year > 0) { strSQL <- paste(sql, "WHERE YEAR = ", year) }
 df <- sqldf(strSQL)
 return (df)
}
```

Retention and Retenion_ALL are SQL strings designed to hold specific columns:

```
# Selecting only fields needed and for the years in (2009,2011,2013,2015)
Retention <- "SELECT STATE,YEAR,TOTAL_REVENUE,TOTAL_EXPENDITURE,
      INSTRUCTION_EXPENDITURE,GRADES_ALL_G ALL_STUDENT_ALL,
      GRADES_8_G ALL_STUDENT_8,AVG_MATH_8_SCORE,
      AVG_READING_8_SCORE,GRADES_9_12_G ALL_STUDENT_9_12,
      GRADES_12_G ALL_STUDENT_12,
      (AVG_MATH_8_SCORE + AVG_READING_8_SCORE)/2 SCORE,
      (GRADES_8_G - GRADES_12_G) DROP_OUT_8_12 ,
      (((GRADES_9_12_G - GRADES_12_G)/3) - GRADES_12_G) HS_DROP_OUT
      FROM dfUSEducation"
# Selecting fields needed from the entire dataset
Retention_ALL <- "SELECT STATE,YEAR,TOTAL_REVENUE,TOTAL_EXPENDITURE,
      INSTRUCTION_EXPENDITURE,GRADES_ALL_G ALL_STUDENT_ALL,
      GRADES_8_G ALL_STUDENT_8,AVG_MATH_8_SCORE,AVG_READING_8_SCORE,
      GRADES_9_12_G ALL_STUDENT_9_12,GRADES_12_G ALL_STUDENT_12,
      (GRADES_8_G - GRADES_12_G) DROP_OUT_8_12,
```

```
    (((GRADES_9_12_G - GRADES_12_G)/3) - GRADES_12_G) HS_DROP_OUT
   FROM dfUSEducation_ALL"
```

Create dataset for NY, MA, and FL:

```
# Prepare datasets for charting
dfRetention.NY <- sqldf('SELECT YEAR, DROP_OUT_8_12, HS_DROP_OUT
            FROM dfRetention WHERE STATE ="new york"')
dfRetention.MA <- sqldf('SELECT YEAR, DROP_OUT_8_12, HS_DROP_OUT
          FROM dfRetention WHERE STATE ="massachusetts"')
dfRetention.FL <- sqldf('SELECT YEAR, DROP_OUT_8_12, HS_DROP_OUT
          FROM dfRetention WHERE STATE ="florida"')
```

Here are screenshots of the resulting datasets for:

New York State (our target data)

	STATE	YEAR	TOTAL_REVENUE	TOTAL_EXPENDITURE	INSTRUCTION_EXPENDITURE	ENROLL	ALL_STUDENT	NATIVE_INDIAN	ASIAN	HISPANIC	BLACK	WHITE	PACIFIC_ISLANDER
30	new york	2015	63712218	65094591	41954260	2631532	197997	1235	17672	48315	35097	93218	NA
78	new york	2013	59623918	60505950	38756656	2629805	203267	1140	17893	46753	36988	98582	0
126	new york	2011	57753776	59446908	37834196	2677412	201190	1041	16073	44445	37666	100464	347
174	new york	2009	55885116	57851481	35195372	2696688	198690	891	15861	39731	36056	105583	NA

	MIXED_RACE	NATIVE_INDIAN_MALE	NATIVE_INDIAN_FEMALE	ASIAN_MALE	ASIAN_FEMALE	HISPANIC_MALE	BLACK_MALE	HISPANIC_FEMALE	WHITE_MALE	WHITE_FEMALE
30	2460	642	593	8994	8678	24578	17652	23737	47713	45505
78	1911	545	595	9205	8688	23923	18799	22830	50585	47997
126	1154	528	513	8195	7878	22616	18854	21829	51372	49092
174	NA	457	434	8248	7613	19731	17617	20000	53983	51600

	BLACK_FEMALE	PACIFIC_ISLANDER_MALE	PACIFIC_ISLANDER_FEMALE	MIXED_RACE_MALE	MIXED_RACE_FEMALE	AVG_MATH_8_SCORE	AVG_READING_8_SCORE
30	17445	NA	NA	1259	1201	280.0892	270.8537
78	18189	0	0	961	950	281.8078	274.3077
126	18812	191	156	569	585	280.4530	258.1933
174	18439	NA	NA	NA	NA	282.5769	267.0679

Massachusetts (max. score comparable)

	STATE	YEAR	TOTAL_REVENUE	TOTAL_EXPENDITURE	INSTRUCTION_EXPENDITURE	ENROLL	ALL_STUDENT	NATIVE_INDIAN	ASIAN	HISPANIC	BLACK	WHITE
19	massachusetts	2015	16985185	16972319	9774884	916130	72176	147	4505	12391	6257	46839
67	massachusetts	2013	16114783	16201905	9097982	920968	72116	155	4217	11356	6218	48310
115	massachusetts	2011	15396681	15150898	8685894	924903	72758	175	3775	10844	6064	50315
163	massachusetts	2009	15298022	15174814	8885949	932437	72093	186	3676	9935	5882	51006

	PACIFIC_ISLANDER	MIXED_RACE	NATIVE_INDIAN_MALE	NATIVE_INDIAN_FEMALE	ASIAN_MALE	ASIAN_FEMALE	HISPANIC_MALE	BLACK_MALE	HISPANIC_FEMALE	WHITE_MALE
19	74	1963	68	79	2207	2298	6221	3195	6170	24090
67	92	1768	77	78	2137	2080	5805	3240	5551	24555
115	62	1523	95	80	1906	1869	5558	3084	5286	25692
163	102	1306	94	92	1801	1875	5102	3119	4833	26198

	WHITE_FEMALE	BLACK_FEMALE	PACIFIC_ISLANDER_MALE	PACIFIC_ISLANDER_FEMALE	MIXED_RACE_MALE	MIXED_RACE_FEMALE	AVG_MATH_8_SCORE	AVG_READING_8_SCORE
19	22749	3062	37	37	970	993	296.9086	268.3912
67	23755	2978	38	54	887	881	300.5682	257.3512
115	24623	2980	29	33	757	766	298.5124	268.8340
163	24808	2763	49	53	632	674	298.8543	266.7995

Florida (min. score comparable)

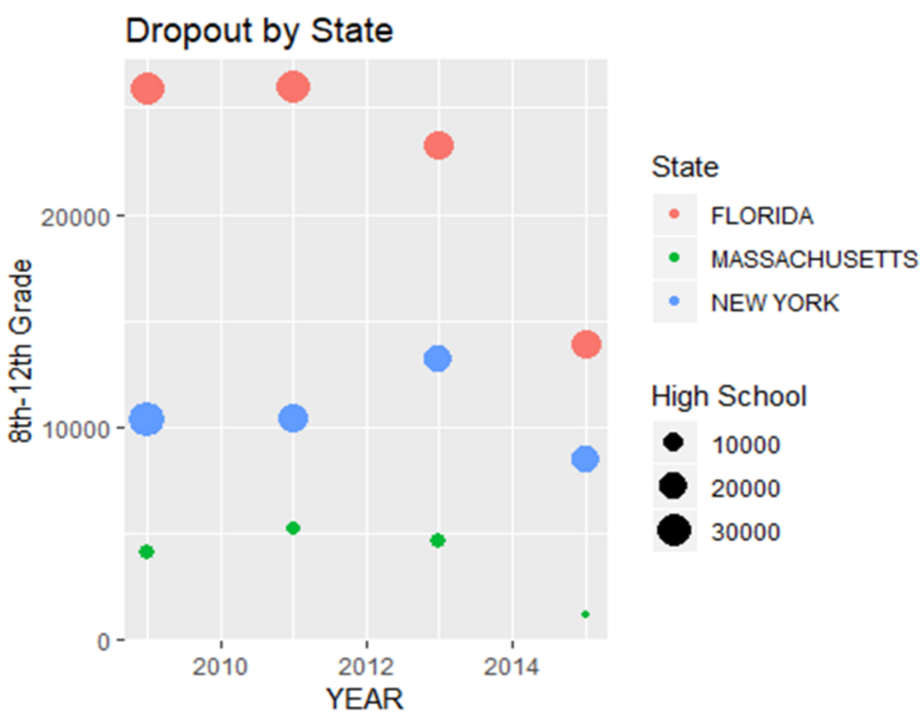
	STATE	YEAR	TOTAL_REVENUE	TOTAL_EXPENDITURE	INSTRUCTION_EXPENDITURE	ENROLL	ALL_STUDENT	NATIVE_INDIAN	ASIAN	HISPANIC	BLACK	WHITE	PACIFIC_ISLANDER
8	florida	2015	26971491	27277049	14931173	2743641	206785	762	5578	64804	45646	83175	247
56	florida	2013	24681548	25245400	13833236	2680074	206698	797	5327	61162	46768	86175	236
104	florida	2011	26453693	26991946	14284224	2636404	200378	869	4986	56730	44848	87004	193
152	florida	2009	26494500	28867429	13884278	2623067	198245	679	4959	51588	44545	90441	NA

	MIXED_RACE	NATIVE_INDIAN_MALE	NATIVE_INDIAN_FEMALE	ASIAN_MALE	ASIAN_FEMALE	HISPANIC_MALE	BLACK_MALE	HISPANIC_FEMALE	WHITE_MALE	WHITE_FEMALE
8	6573	393	369	2734	2844	33311	23340	31493	43158	40017
56	6233	420	377	2678	2649	31285	23649	29877	44193	41982
104	5748	426	443	2495	2491	29026	22994	27704	45037	41967
152	NA	348	331	2486	2473	26373	22404	25215	46609	43832

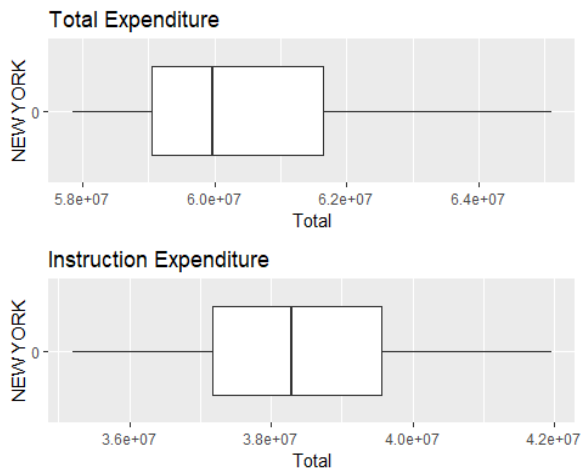
	BLACK_FEMALE	PACIFIC_ISLANDER_MALE	PACIFIC_ISLANDER_FEMALE	MIXED_RACE_MALE	MIXED_RACE_FEMALE	AVG_MATH_8_SCORE	AVG_READING_8_SCORE
8	22306	116	131	3284	3289	275.3238	247.6534
56	23119	101	135	3111	3122	280.8558	265.9850
104	21854	89	104	2787	2961	277.8370	274.6814
152	22141	NA	NA	NA	NA	279.3353	265.5131

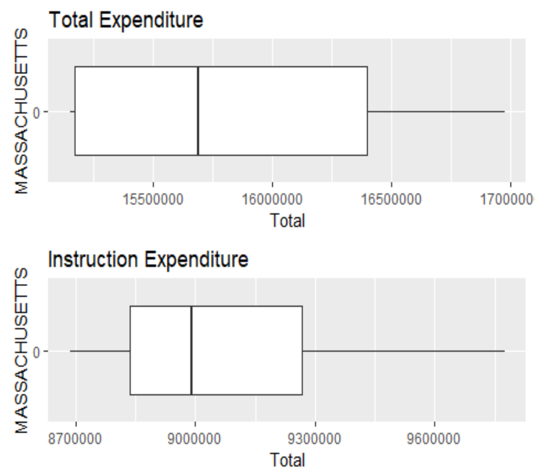
4.3 Box Plots, Scatter Plots, and Histograms

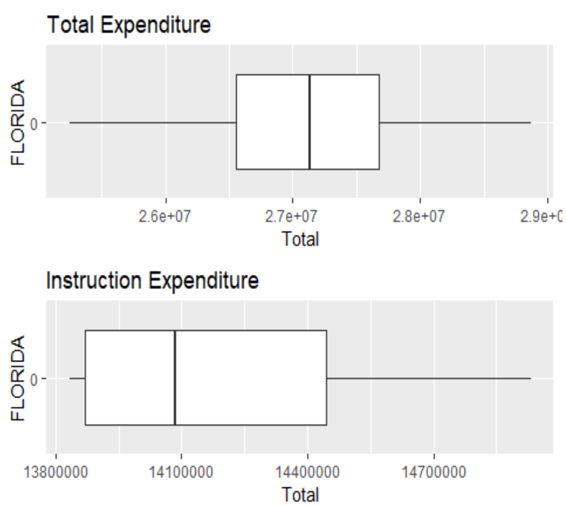
Box plots enable data observers to study the distributional characteristics of a group of scores as well as the level of the scores. The following plots (scatter plots, box plots, and histograms) allow us to visualize comparatively the characteristics of scores for total expenditure and instruction expenditure. First, it is clear that all states examined in recent years have improved in terms of student retention as the following scatter plot attests.



Nonetheless, there are evident differences among the three states in particular when we analyze the ratios between total and instruction expenditures. As the plots below also show, the instruction expenditure for New York state corresponds to the 63.3% of total expenditure with a median value of around $38 millions against a median value for total expenditure of around 60 millions. Both Massachusetts and Florida return a lower percentage of around 57.3%, with a median value for instruction expenditure of around $9millions vs. a total expenditure median value of around 15.7 millions, for Massachusetts, and 51.85%, with a median instruction expenditure of around $14millions vs. a median total expenditure of more than 27 millions, for Florida.



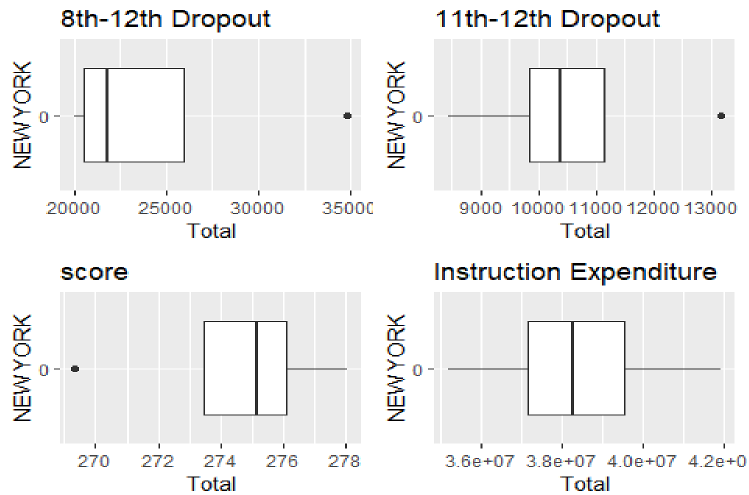


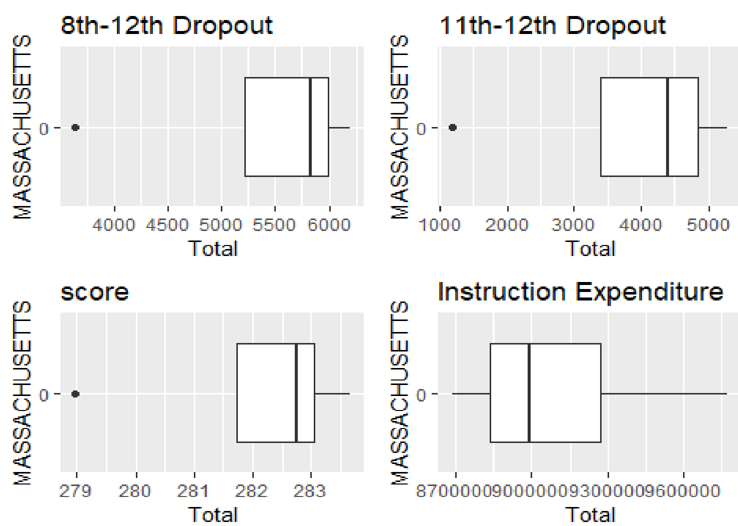


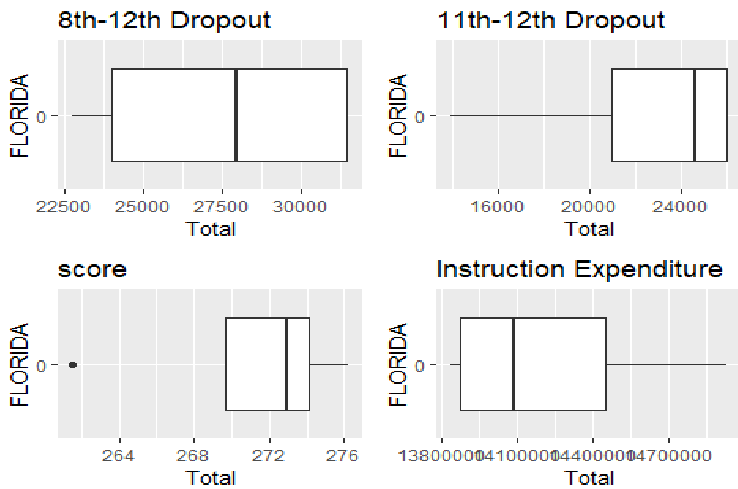
The above boxplots were produced by using GGPLOT(Boxplot) and GRID.ARRANGE packages as follows:

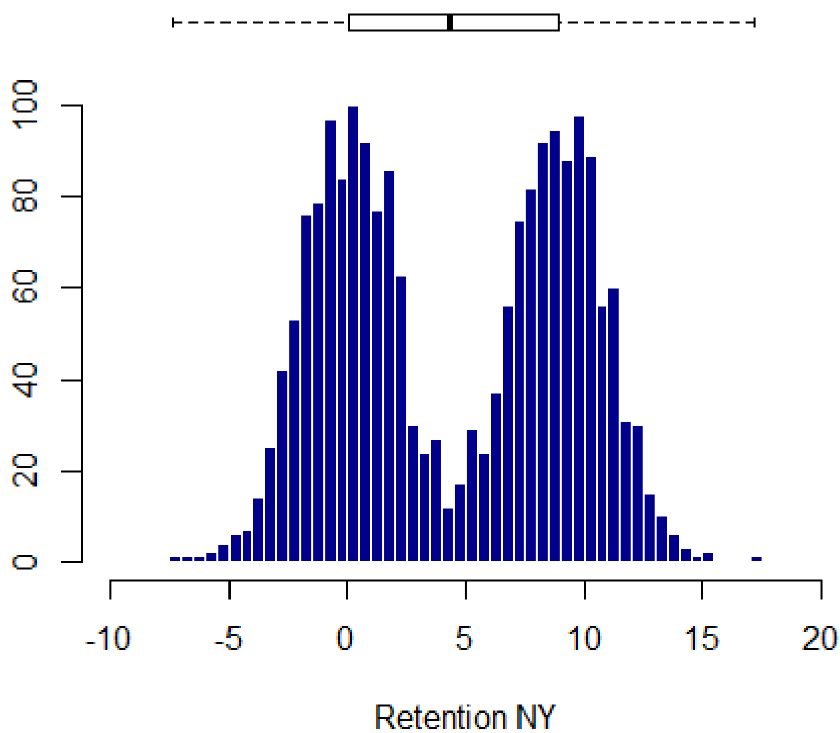
```
# Create chart to show total expenditure for NY
NY_TOT_EXPEND_BP <- ggplot(dfGrades_8NewYork, aes(x=factor(0),
dfGrades_8NewYork$TOTAL_EXPENDITURE)) +
            geom_boxplot()+coord_flip() +
            labs(x="NEW YORK") +
            labs(y="Total") +
            ggtitle("Total Expenditure")
# Create chart to show instruction expenditure for NY
NY_TOT_INST_BP <- ggplot(dfGrades_8NewYork, aes(x=factor(0),
dfGrades_8NewYork$INSTRUCTION_EXPENDITURE)) +
            geom_boxplot()+coord_flip() +
            labs(x="NEW YORK") +
            labs(y="Total") +
            ggtitle("Instruction Expenditure")
# Combining the charts via grid.arrange
grid.arrange(NY_TOT_EXPEND_BP,NY_TOT_INST_BP)
```

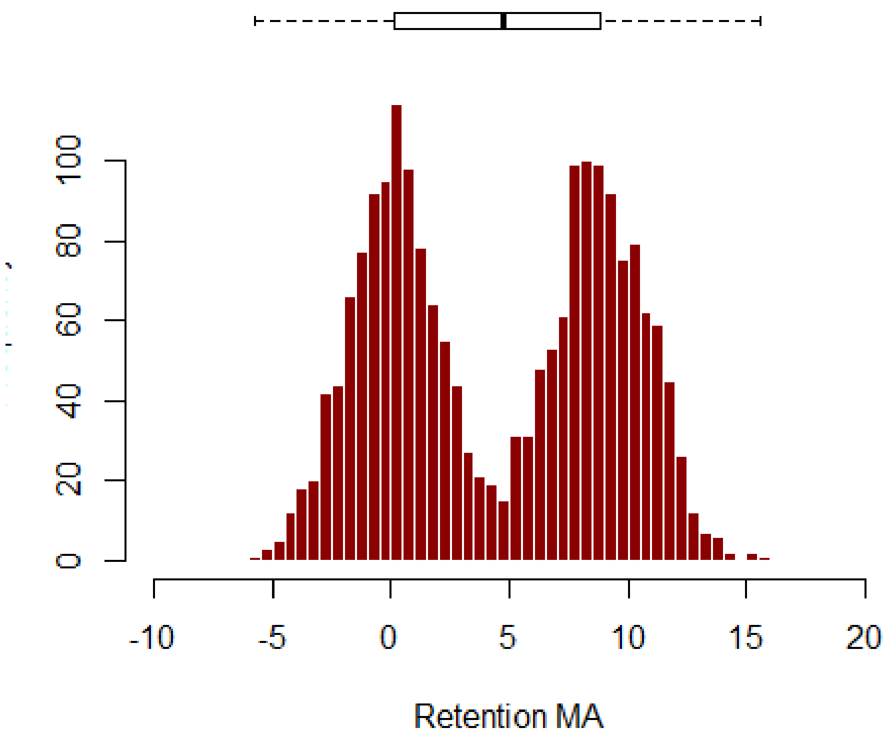
But, how do instruction expenditures, dropout, and score relate in those three states? The following plots visualize exactly these relationships.



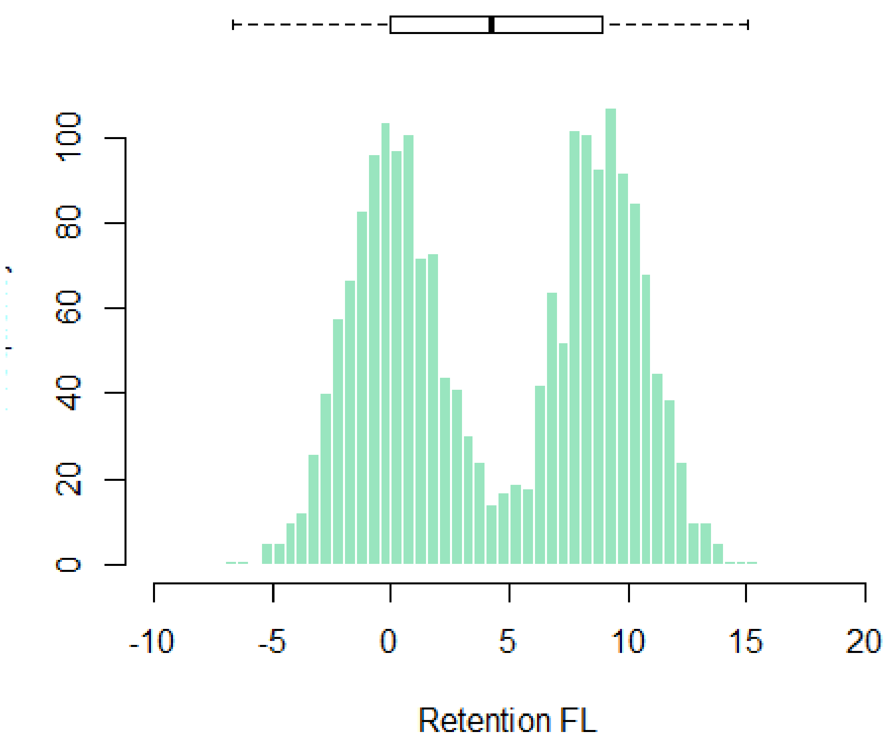



8th-12th Dropout
FLORIDA
0
22500
25000
27500
30000
Total
11th-12th Dropout
16000
20000
24000
score
264
268
272
276
Instruction Expenditure
13800000
14100000
14400000
14700000


0
20
40
60
80
100
-10
-5
5
10
15
20
Retention NY



0
20
40
60
80
100
-10
-5
0
5
10
15
20
Retention MA



0
20
40
60
80
100
-10
-5
0
5
10
15
20
Retention FL

5) Modeling Techniques

In our analysis, we focused on finding possible factors that could affect student retention/attrition in public high schools. What variables might affect retention? Is retention correlated to student scores in 8th grade or to schools' expenditure? Or both?

We tried to answer these questions by means of two inferential statistics models, linear modeling and support vector machine. We operated our calculations for the three states of New York, Massachusetts, and Florida for the following categories:

1) Retention (8-12 and 9-12) ~ Instruction expenditure
2) Retention (8-12 and 9-12) ~ Scores at the 8th grade
3) Scores at the 8th grade ~ Instruction expenditure

5.1 Linear Modeling

In this section, we report the findings and plots of our linear modeling for New York state, Massachusetts, and Florida for the dropout, score, and instruction expenditure variables. To see the whole set of values that our linear modeling returned see Appendix 2: Linear Modeling Results. Here is a sample code that we used to create LM models:

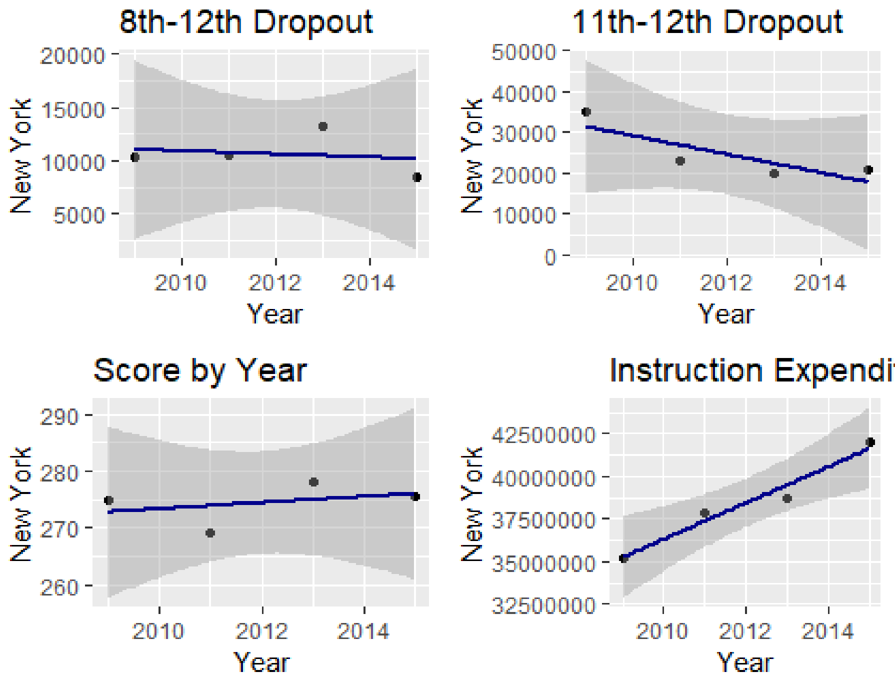
```
# Models for Retention NY
NY_mHS_DROP_OUT_EXPENDLM <- lm(HS_DROP_OUT ~ INSTRUCTION_EXPENDITURE, data=dfRet
ention_NY)

NY_mHS_DROP_OUT_ScoreLM <- lm(HS_DROP_OUT ~ AVG_MATH_8_SCORE + AVG_READING_8_SC
ORE,
                    data=dfRetention_NY)

NY_mDROP_OUT_8_12_ScoreLM <- lm(DROP_OUT_8_12 ~  AVG_MATH_8_SCORE + AVG_READING_8
_SCORE,
                     data=dfRetention_NY)
```

Sample code on how model is tested using PREDICT():

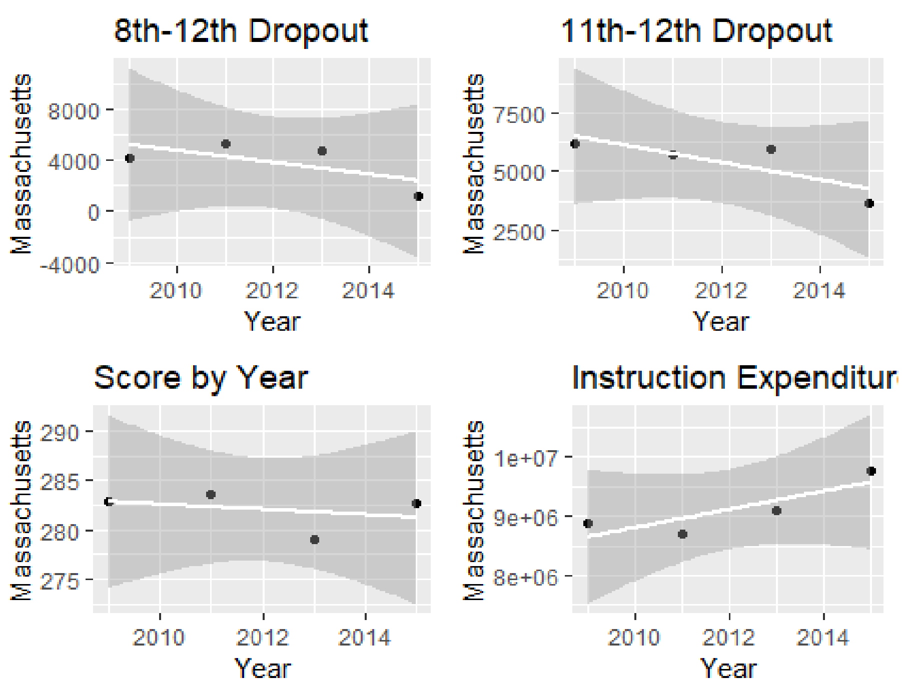
```
# Test model
NY_pmHS_DROP_OUT_EXPENDLM <- predict(NY_mHS_DROP_OUT_ScoreLM,
                        dfRetention_NY, type='response')
NY_pmHS_DROP_OUT_ScoreLM <- predict(NY_mHS_DROP_OUT_ScoreLM,
                        dfRetention_NY, type='response')
NY_pmDROP_OUT_8_12_ScoreLM <-  predict(NY_mDROP_OUT_8_12_ScoreLM,
                         dfRetention_NY, type='response')
```

As expected, different intensity is evident in the examined correlations in different states. Let's see closely:

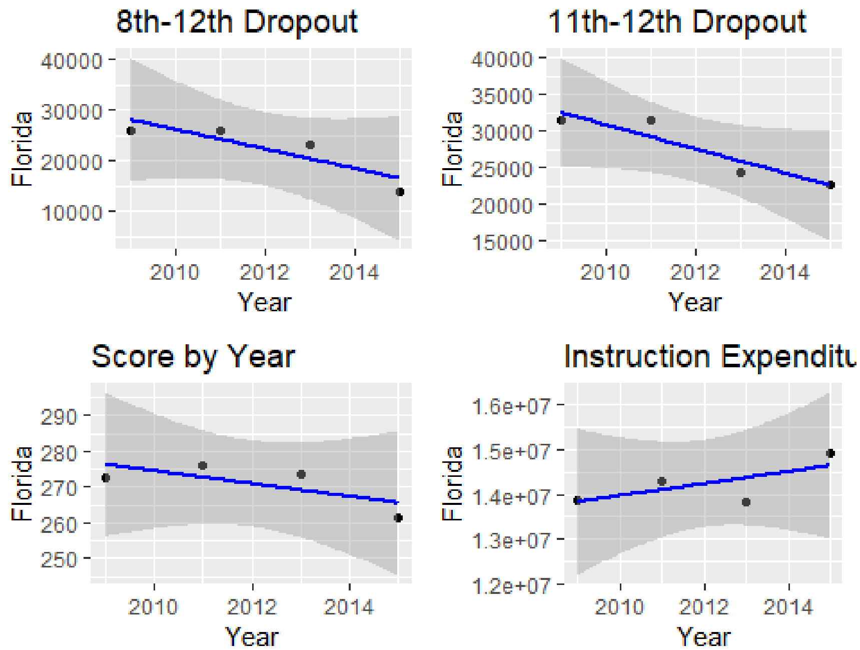
New York State presents stronger correlation between student retention and instruction expenditure (NY.1), but weak correlation between retention and 8th grade scores (NY.2 and NY.3).



Massachusetts reverses the correlation. Indeed, we have stronger correlation between retention and high scores in 8th grade (see MA.3), but weak correlation between retention and instruction expenditure (MA.1 and MA.2).



Florida appears to offer a third condition: retention correlates to both instruction expenditure (FL.1) and 8th grade scores (FL.2 and FL.3).



These findings lead to confirm a general principle that grounds our actionable suggestions at the end of the project (see the Summary below): *there is not one combination of factors that works similarly for all states. Each state can have different correlations among the following three possible ones.*

COMBINATION	STRONG CORRELATION	WEAK CORRELATION	STATE
1	Retention ~ Instruction exp.	Retention ~ Scores	NY
2	Retention ~ Scores	Retention ~ Instruction exp.	MA
3	Retention ~ Instruction exp. Retention ~ Scores		FL

Additionally, we observed that the correlation between scores and retention is not as strong as one might assume before analyzing the data. The strength of this correlation is different for the three states; for New York state and Massachusetts it appears to be quite weak, whereas it is much stronger for Florida as the results below attest.

New York:

```
Call:
lm(formula = AVG_MATH_8_SCORE + AVG_READING_8_SCORE ~
INSTRUCTION_EXPENDITURE,
    data = dfGrades_8NewYork)

Residuals:
     30      78     126     174 
 2.6473 -9.8498  7.0955  0.1069 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)  
(Intercept)             5.270e+02  7.003e+01   7.526   0.0172 *
INSTRUCTION_EXPENDITURE 5.679e-07  1.818e-06   0.312   0.7844  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.786 on 2 degrees of freedom
Multiple R-squared:  0.0465,	Adjusted R-squared:  -0.4302 
F-statistic: 0.09755 on 1 and 2 DF,  p-value: 0.7844
```

Massachusetts

```
Call:
lm(formula = AVG_MATH_8_SCORE + AVG_READING_8_SCORE ~
INSTRUCTION_EXPENDITURE,
    data = dfGrades_8Massachusetts)

Residuals:
    19     67    115    163 
 1.313  2.752 -6.152  2.087 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)   
(Intercept)              5.756e+02  5.645e+01  10.197  0.00948 **
INSTRUCTION_EXPENDITURE -1.268e-06  6.189e-06  -0.205  0.85658   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.075 on 2 degrees of freedom
Multiple R-squared:  0.02057,	Adjusted R-squared:  -0.4691 
F-statistic: 0.042 on 1 and 2 DF,  p-value: 0.8566
```

Florida

```
Call:
lm(formula = AVG_MATH_8_SCORE + AVG_READING_8_SCORE ~
INSTRUCTION_EXPENDITURE,
    data = dfGrades_8Florida)

Residuals:
      8      56     104     152
-4.036 11.758 -3.080 -4.642

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              8.309e+02  1.562e+02   5.321   0.0336 *
INSTRUCTION_EXPENDITURE -2.031e-05  1.097e-05  -1.852   0.2052
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.633 on 2 degrees of freedom
Multiple R-squared:  0.6317,   Adjusted R-squared:  0.4476
F-statistic:  3.43 on 1 and 2 DF,  p-value: 0.2052
```

5.2 Support Vector Machines (SVM)

The Support Vector Machine is another method used to observe correlations among variables that can help the analyst to produce actionable insights. Overall, the SVM, which we operated with the ksvm function in RStudio, confirmed the results returned through linear modeling.

First, we divided the whole data set into two subsets, one to train the machine, the second to test the datasets in order to obtain a prediction. The SVM returned a prediction very close to the actual values we examined, which confirms that the correlation between instruction expenditure and student retention is a strong one to validate good predictions for all states examined. For Florida, the SVM also confirmed that 8th grade math and reading scores are factors that affect retention.

Sample code on how KSMV model is created:

```
# Models for Retention NY
NY_mHS_DROP_OUT_EXPEND <- ksvm(HS_DROP_OUT ~ INSTRUCTION_EXPENDITURE,
                data=NY_tr.Retention, kernel = "rbfdot",
                kpar="automatic", C=1, cross=2, prob.model=TRUE)

NY_mHS_DROP_OUT_Score <- ksvm(HS_DROP_OUT ~ AVG_MATH_8_SCORE + AVG_READING_8_SCO
RE,
                data=NY_tr.Retention, kernel = "rbfdot",
                kpar="automatic", C=1, cross=2, prob.model=TRUE)

NY_mDROP_OUT_8_12_Score <- ksvm(DROP_OUT_8_12 ~ AVG_MATH_8_SCORE + AVG_READING_8_
SCORE,
```

```
                data=NY_tr.Retention, kernel = "rbfdot",
                kpar="automatic", C=1, cross=2, prob.model=TRUE)
```

Sample code on how model is tested using PREDICT():

```
# Model is created on line 946
pNY_mHS_DROP_OUT_EXPEND <- predict(NY_mHS_DROP_OUT_EXPEND, NY_ts.Retention)
pNY_mHS_DROP_OUT_EXPEND.Error <- (NY_ts.Retention$HS_DROP_OUT - pNY_mHS_DROP_OUT_E
XPEND )
pNY_mHS_DROP_OUT_EXPEND.rmse <-  rmse(pNY_mHS_DROP_OUT_EXPEND.Error)
print(pNY_mHS_DROP_OUT_EXPEND.rmse)
```

```
# NEW YORK - Print results
NY_mHS_DROP_OUT_EXPEND

## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  65.397595920409
##
## Number of Support Vectors : 15
##
## Objective Function Value : -4.228
## Training error : 0.093005
## Cross validation error : 200532831
## Laplace distr. width : 25824.85

NY_mHS_DROP_OUT_Score

## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.280707562478635
##
## Number of Support Vectors : 16
##
## Objective Function Value : -12.4477
## Training error : 1.057574
## Cross validation error : 808501896
## Laplace distr. width : 42568.96
```

```
# MASSACHUSETTS - Print results
MA_mHS_DROP_OUT_EXPEND

## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  3.9263737571042
##
## Number of Support Vectors : 13
##
## Objective Function Value : -3.7128
## Training error : 0.084244
## Cross validation error : 12876582
## Laplace distr. width : 9287.054

MA_mHS_DROP_OUT_Score
```

```
## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  164.859042036737
##
## Number of Support Vectors : 16
##
## Objective Function Value : -9.0734
## Training error : 0.707954
## Cross validation error : 15127350
## Laplace distr. width : 1526.169
```

```
# FLORIDA - Print results
FL_mHS_DROP_OUT_EXPEND

## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  8.05720275132499
##
## Number of Support Vectors : 15
##
## Objective Function Value : -3.9278
## Training error : 0.09226
## Cross validation error : 87229152
## Laplace distr. width : 33962.06

FL_mHS_DROP_OUT_Score

## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  155.835817328262
##
## Number of Support Vectors : 17
##
## Objective Function Value : -9.5268
## Training error : 0.622685
## Cross validation error : 222099984
## Laplace distr. width : 0
```

6) Summary

After this long journey, let's sum up the outcomes we obtained from our data analysis:

New York State: stronger correlation between student retention and instruction expenditure (NY.1), but weak correlation between retention and 8th grade scores (NY.2 and NY.3).

Massachusetts: stronger correlation between retention and high scores in 8th grade (see MA.3), but weak correlation between retention and instruction expenditure (MA.1 and MA.2).

Florida: retention correlates to both instruction expenditure (FL.1) and 8th grade scores (FL.2 and FL.3).

With these results in mind, our team would provide the following actionable suggestions to the future administrations of New York state:

1. Always keep in mind the general principle according to which: *there is not one combination of factors that works similarly for all states. Each state can have different correlations among different factors.*

2. By comparing New York state with other two states, we observed that for improving student retention, student grades in quantitative skills (math) as well as humanities-oriented areas of study (reading) are not as crucial as instruction expenditure.

3. To keep improving public school retention rates, the state of New York might want to consider keeping high instruction expenditure as it has been doing in recent years.

A Appendices

A.1 Appendix 1: Full List of Variables

Here you can find the complete list of 193 columns of the state_all_extended.csv database accompanied by a legend that explains each single acronym (see below).

Category 1: State

STATE

Category 2: Year
YEAR

Category 3: Total enrollment

ENROLL

Category 4: Revenue

TOTAL_REVENUE
FEDERAL_REVENUE
STATE_REVENUE
LOCAL_REVENUE

Category 5: Expenditure

TOTAL_EXPENDITURE
INSTRUCTION_EXPENDITURE
SUPPORT_SERVICES_EXPENDITURE
OTHER_EXPENDITURE
CAPITAL_OUTLAY_EXPENDITURE

Category 6: Enrollment And Retention (grades) + Category 7: Student Demographic Information (race and gender)

PRE-SCHOOL	KINDERGARDEN	4TH GRADE	8TH GRADE
GRADES_PK_G	GRADES_KG_G	GRADES_4_G	GRADES_8_G
GRADES_PK_AM	GRADES_KG_AM	GRADES_4_AM	GRADES_8_AM
GRADES_PK_AS	GRADES_KG_AS	GRADES_4_AS	GRADES_8_AS
GRADES_PK_HI	GRADES_KG_HI	GRADES_4_HI	GRADES_8_HI
GRADES_PK_BL	GRADES_KG_BL	GRADES_4_BL	GRADES_8_BL
GRADES_PK_WH	GRADES_KG_WH	GRADES_4_WH	GRADES_8_WH
GRADES_PK_HP	GRADES_KG_TR	GRADES_4_HP	GRADES_8_HP
GRADES_PK_TR	GRADES_KG_AMM	GRADES_4_TR	GRADES_8_TR
GRADES_PK_AMM	GRADES_KG_HP	GRADES_4_AMM	GRADES_8_AMM
GRADES_PK_AMF	GRADES_KG_ASM	GRADES_4_AMF	GRADES_8_AMF
GRADES_PK_ASM	GRADES_KG_AMF	GRADES_4_ASM	GRADES_8_ASM
GRADES_PK_ASF	GRADES_KG_HIM	GRADES_4_ASF	GRADES_8_ASF

GRADES_PK_HIM	GRADES_KG_ASF	GRADES_4_HIM	GRADES_8_HIM
GRADES_PK_HIF	GRADES_KG_HIF	GRADES_4_BLM	GRADES_8_HIF
GRADES_PK_BLM	GRADES_KG_BLM	GRADES_4_HIF	GRADES_8_BLM
GRADES_PK_BLF	GRADES_KG_BLF	GRADES_4_WHM	GRADES_8_BLF
GRADES_PK_WHM	GRADES_KG_WHM	GRADES_4_WHF	GRADES_8_WHM
GRADES_PK_WHF	GRADES_KG_WHF	GRADES_4_BLF	GRADES_8_WHF
GRADES_PK_HPM	GRADES_KG_HPM	GRADES_4_HPM	GRADES_8_HPM
GRADES_PK_HPF	GRADES_KG_HPF	GRADES_4_HPF	GRADES_8_HPF
GRADES_PK_TRM	GRADES_KG_TRM	GRADES_4_TRM	GRADES_8_TRM
GRADES_PK_TRF	GRADES_KG_TRF	GRADES_4_TRF	GRADES_8_TRF
GRADE 1ST– 8TH	**GRADE 9TH**	**GRADE 9TH -12TH**	**ALL GRADES**
GRADES_1_8_G	GRADES_9_G	GRADES_9_12_G	GRADES_ALL_G
GRADES_1_8_AM	GRADES_9_AM	GRADES_9_12_AM	GRADES_ALL_AM
GRADES_1_8_AS	GRADES_9_AS	GRADES_9_12_AS	GRADES_ALL_AS
GRADES_1_8_HI	GRADES_9_HI	GRADES_9_12_HI	GRADES_ALL_HI
GRADES_1_8_BL	GRADES_9_BL	GRADES_9_12_BL	GRADES_ALL_BL
GRADES_1_8_WH	GRADES_9_WH	GRADES_9_12_WH	GRADES_ALL_WH
GRADES_1_8_HP	GRADES_9_HP	GRADES_9_12_HP	GRADES_ALL_HP
GRADES_1_8_TR	GRADES_9_TR	GRADES_9_12_TR	GRADES_ALL_TR
GRADES_1_8_AMM	GRADES_9_AMM	GRADES_9_12_AMM	GRADES_ALL_AMM
GRADES_1_8_AMF	GRADES_9_AMF	GRADES_9_12_AMF	GRADES_ALL_AMF
GRADES_1_8_ASM	GRADES_9_ASM	GRADES_9_12_ASM	GRADES_ALL_ASM
GRADES_1_8_ASF	GRADES_9_ASF	GRADES_9_12_ASF	GRADES_ALL_ASF
GRADES_1_8_HIM	GRADES_9_HIM	GRADES_9_12_HIM	GRADES_ALL_HIM
GRADES_1_8_HIF	GRADES_9_HIF	GRADES_9_12_HIF	GRADES_ALL_HIF
GRADES_1_8_BLM	GRADES_9_BLM	GRADES_9_12_BLM	GRADES_ALL_BLM
GRADES_1_8_BLF	GRADES_9_BLF	GRADES_9_12_BLF	GRADES_ALL_BLF
GRADES_1_8_WHM	GRADES_9_WHM	GRADES_9_12_WHM	GRADES_ALL_WHM
GRADES_1_8_WHF	GRADES_9_WHF	GRADES_9_12_WHF	GRADES_ALL_WHF
GRADES_1_8_HPM	GRADES_9_HPM	GRADES_9_12_HPM	GRADES_ALL_HPM
GRADES_1_8_HPF	GRADES_9_HPF	GRADES_9_12_HPF	GRADES_ALL_HPF
GRADES_1_8_TRM	GRADES_9_TRM	GRADES_9_12_TRM	GRADES_ALL_TRM
GRADES_1_8_TRF	GRADES_9_TRF	GRADES_9_12_TRF	GRADES_ALL_TRF

Category 8: Assessment (math and reading scores)

AVG_MATH_4_SCORE
AVG_MATH_8_SCORE
AVG_READING_4_SCORE
AVG_READING_8_SCORE

LEGEND

Grades_ALL_AS	Number of students whose ethnicity was classified as "Asian"
Grades_ALL_ASM	Number of male students whose ethnicity was classified as "Asian"
Grades_ALL_ASF	Number of female students whose ethnicity was classified as "Asian"
The represented races include	
AM	American Indian or Alaska Native
AS	Asian
HI	Hispanic/Latino
BL	Black or African American
WH	White
HP	Hawaiian Native/Pacific Islander
TR	Two or More Races
The represented genders include	
M	Male
F	Female

A.2 Appendix 2: Linear Modeling Results

NEW YORK

NY.1 High School retention ~ Instruction expenditure

```
Call:
lm(formula = HS_DROP_OUT ~ STATE + AVG_MATH_8_SCORE + AVG_READING_8_SCORE,
    data = dfRetention_Copy)

Residuals:
   Min     1Q Median     3Q    Max
-19701  -8402   -871   4905  32801

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2904571.7   564545.9   5.145 0.000188 ***
STATEflorida         -826263.2    14961.6 -55.226  < 2e-16 ***
STATEmassachusetts  -1132738.4    53042.5 -21.355 1.66e-11 ***
STATEnew york        -792312.1    19471.9 -40.690 4.30e-15 ***
STATEohio           -1002313.2    30178.6 -33.213 5.88e-14 ***
AVG_MATH_8_SCORE       -4704.7     2084.3  -2.257 0.041849 *
AVG_READING_8_SCORE     -546.3      574.2  -0.951 0.358735
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16040 on 13 degrees of freedom
Multiple R-squared:  0.9991,  Adjusted R-squared:  0.9987
F-statistic:  2392 on 6 and 13 DF,  p-value: < 2.2e-16
```

NY.2 High School retention ~ Scores at the 8th grade

```
Call:
lm(formula = HS_DROP_OUT ~ AVG_MATH_8_SCORE + AVG_READING_8_SCORE,
    data = dfRetention_NY)

Residuals:
     1      2      3      4
  8100  -5314 -10600   7814

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2706302    2301459  -1.176    0.449
AVG_MATH_8_SCORE       13294       8457   1.572    0.361
AVG_READING_8_SCORE    -1456       1417  -1.027    0.491

Residual standard error: 16350 on 1 degrees of freedom
Multiple R-squared:   0.74,   Adjusted R-squared:  0.2201
F-statistic: 1.423 on 2 and 1 DF,  p-value: 0.5099
```

NY.3 Retention 8th-12th grades ~ Scores at the 8th grade

```
Call:
lm(formula = DROP_OUT_8_12 ~ AVG_MATH_8_SCORE + AVG_READING_8_SCORE,
    data = dfRetention_NY)

Residuals:
      1       2       3       4 
-1391.0   912.6  1820.4 -1342.0 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -236753.0   395243.4  -0.599    0.656
AVG_MATH_8_SCORE        839.2     1452.4   0.578    0.666
AVG_READING_8_SCORE      42.4      243.4   0.174    0.890

Residual standard error: 2808 on 1 degrees of freedom
Multiple R-squared:  0.3132,   Adjusted R-squared:  -1.06 
F-statistic: 0.228 on 2 and 1 DF,  p-value: 0.8287
```

MASSACHUSETTS

MA.1 High School retention ~ Instruction expenditure

```
Call:
lm(formula = HS_DROP_OUT ~ INSTRUCTION_EXPENDITURE, data = dfRetention_MA)

Residuals:
      1       2       3       4 
 1744.6  -537.6 -1426.3   219.2 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)   
(Intercept)             1.918e+05  1.830e+04  10.481  0.00898 **
INSTRUCTION_EXPENDITURE 3.258e-03  2.007e-03   1.624  0.24594   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1645 on 2 degrees of freedom
Multiple R-squared:  0.5686,   Adjusted R-squared:  0.3529 
F-statistic: 2.636 on 1 and 2 DF,  p-value: 0.2459
```

MA.2 High School retention ~ Scores at the 8th grade

```
Call:
lm(formula = HS_DROP_OUT ~ AVG_MATH_8_SCORE + AVG_READING_8_SCORE,
    data = dfRetention_MA)

Residuals:
       1        2        3        4 
 1657.20 -1342.46  -292.77   -21.98 

Coefficients:
```

```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          826389.7   583605.6   1.416    0.391
AVG_MATH_8_SCORE      -1786.8     1603.9  -1.114    0.466
AVG_READING_8_SCORE    -268.0      445.8  -0.601    0.655

Residual standard error: 2153 on 1 degrees of freedom
Multiple R-squared:  0.6308,	Adjusted R-squared:  -0.1077 
F-statistic: 0.8541 on 2 and 1 DF,  p-value: 0.6077
```

MA.3 Retention 8th-12th grades ~ Scores at the 8th grade

```
Call:
lm(formula = DROP_OUT_8_12 ~ AVG_MATH_8_SCORE + AVG_READING_8_SCORE, 
    data = dfRetention_MA)

Residuals:
       1        2        3        4 
-601.701  487.423  106.298    7.979 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -748320.6   211897.2  -3.532    0.176
AVG_MATH_8_SCORE        2157.1      582.3   3.704    0.168
AVG_READING_8_SCORE      406.2      161.9   2.510    0.241

Residual standard error: 781.7 on 1 degrees of freedom
Multiple R-squared:  0.9389,	Adjusted R-squared:  0.8167 
F-statistic: 7.684 on 2 and 1 DF,  p-value: 0.2472
```

FLORIDA

FL.1 High School retention ~ Instruction expenditure

```
Call:
lm(formula = HS_DROP_OUT ~ INSTRUCTION_EXPENDITURE, data = dfRetention_FL)

Residuals:
    1     2     3     4 
-4416 -8528  9236  3708 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.514e+05  1.585e+05   1.585    0.254
INSTRUCTION_EXPENDITURE 2.624e-02  1.113e-02   2.357    0.143

Residual standard error: 9779 on 2 degrees of freedom
Multiple R-squared:  0.7353,	Adjusted R-squared:  0.6029 
F-statistic: 5.555 on 1 and 2 DF,  p-value: 0.1425
```

FL.2 High School retention ~ Scores at the 8th grade

```
Call:
lm(formula = HS_DROP_OUT ~ AVG_MATH_8_SCORE + AVG_READING_8_SCORE,
    data = dfRetention_FL)

Residuals:
    1     2     3     4 
-9720  2374  5970  1376 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1409291.7   914721.6   1.541    0.367
AVG_MATH_8_SCORE      -1930.1     3693.7  -0.523    0.693
AVG_READING_8_SCORE    -938.3      767.3  -1.223    0.436

Residual standard error: 11730 on 1 degrees of freedom
Multiple R-squared:  0.8095,	Adjusted R-squared:  0.4285 
F-statistic: 2.125 on 2 and 1 DF,  p-value: 0.4365
```

FL.3 Retention 8th-12th grades ~ Scores at the 8th grade

```
Call:
lm(formula = DROP_OUT_8_12 ~ AVG_MATH_8_SCORE + AVG_READING_8_SCORE,
    data = dfRetention_FL)

Residuals:
      1       2       3       4 
 2270.2  -554.4 -1394.3  -321.5 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -235663.3   213647.8  -1.103    0.469
AVG_MATH_8_SCORE        542.6      862.7   0.629    0.643
AVG_READING_8_SCORE     405.8      179.2   2.265    0.265

Residual standard error: 2740 on 1 degrees of freedom
Multiple R-squared:  0.9236,	Adjusted R-squared:  0.7707 
F-statistic: 6.042 on 2 and 1 DF,  p-value: 0.2764
```

A.3 Appendix 3: R code

Mileva_Selenu_Synn_IST687_Final_Project.Rmd