|  |  |
| --- | --- |
|  |  |

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

**Maya Mileva, Stefano Selenu, Meng Synn**

# FINAL PROJECT March 21, 2019

# IST 687 Applied Data Science Prof. J. Saltz and M. Syed

[1) Introduction 2](#_Toc4397975)

[2) Problem Definition, Business Questions, and Methodology 2](#_Toc4397976)

[3) Data Acquisition, Munging, and Classification 2](#_Toc4397977)

[3.1 Data Acquisition 2](#_Toc4397978)

[3.2 Data cleaning and munging 2](#_Toc4397979)

[3.3 Data Classification: a Short Summary 2](#_Toc4397980)

[4) Descriptive Statistics 2](#_Toc4397981)

[4.1 Data Structure 2](#_Toc4397982)

[4.2 EDA with Maps 2](#_Toc4397983)

[4.3 Reshaping the Data 2](#_Toc4397984)

[4.4 Box Plots and Histograms 2](#_Toc4397985)

[5) Modeling Techniques 2](#_Toc4397986)

[5.1 Linear Modeling 2](#_Toc4397987)

[5.2 Support Vector Machines (SVM) 2](#_Toc4397988)

[6) Summary 2](#_Toc4397989)

[A Appendices 2](#_Toc4397990)

[A.1 Appendix 1 2](#_Toc4397991)

[A.2 R code 2](#_Toc4397992)

# 1) Introduction

This project, performed in RStudio, deals 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 convenient CSV file format. Different versions of the dataset are available online at <https://www.kaggle.com/noriuk/us-education-datasets-unification-project>, where one can find different databases. Given that we had multiple options on the ways we could look at US K-12 Education through different data sets, we decided to reshape and model our own data sets as reported in the fourth section here below.

Our goal with this project is to understand if and how student enrollment and achievement in time correlates to revenue/expenditure ratio and the student demographic information about race and gender. In order to reach some actionable insight we decided to focus our attention in particular on three states: New York State, Massachusetts, and Florida. Why these three states? Our actionable insights will be directed toward New York States. Instead, we chose to compare New York State to Massachusetts and Florida for a simple reasons: they are the states with the highest and the lowest average scores in reading and math across the whole K-12 educational grade spectrum respectively.

# 2) Problem Definition, Business Questions, and Methodology

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]](#footnote-1) In fact, the collecting of data has often worked as an instrument to penalize educators. As the Center for Digital Education director Kecia Ray pointed out, data have unfortunately been used “as a zinger, a way to shut down schools and fire superintendents.”[[2]](#footnote-2)

This project grew 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. In other words, our team believes that an appropriate use of data and data analysis in education could change its course for years to come.

With this in mind, the purpose of this project is to identify relationships among revenue, expenditure, enrollment, and 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, which has worked in this context as our target state. 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:

* What are the relationships among revenue, expenditure, enrollment, student demographics, retention, and average scores (math and reading) in different states and years?
* What affects student retention (i.e. low attrition) in public high schools?
* To what extent do the revenue-expenditure ratios and the math/reading scores in 8th grade affect student retention/attrition?
* Are there any specific trends observable within 5 years or more?

In terms of methodology, our team used R, the programming language and software environment for statistical analysis, graphics representation and reporting. We used RStudio 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 .R file which can be opened directly in RStudio. Please see the attached .R file (the sections in this report are also indicated throughout the code).

# 3) Data Acquisition, Munging, and Classification

## 3.1 Data Acquisition

The first step 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 (see below).

## 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 decide to replace the NAs with 0. We decided not to fulfil blanks with an average value because we considered that that could have affected our final results. For this reason, our team considered that, in order to keep important data included in the rows with a few NAs, replacing the NAs with 0s was the best way to clean the data.

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. Lastly, we decided to exclude Alaska, District of Columbia, 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 (See below for maps).

## 3.3 Data Classification: a Short Summary

After cleaning the data set, we looked at the data available to us. Some of the most important variables in the dataset and their definitions can be the following – this is only a selection and does not want to be an all-encompassing list. See the Appendix 1 for a full list.

|  |  |  |
| --- | --- | --- |
| **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 data set, which includes 193 variables, we decided to group the variables into 8 main categories as follows:

|  |  |  |
| --- | --- | --- |
| **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. This data is of critical importance for our assumption of a possible correlation between expenditure and score and retention.
* **Category 6 to 8** 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, retention, expenditure, and assessment

# 4) Descriptive Statistics

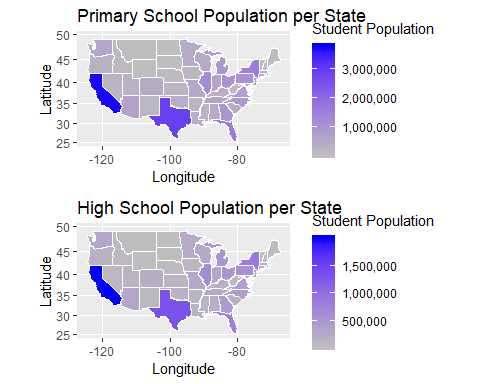
## 4.1 Data Structure

We started our exploratory data analysis (EDA) with a medium-sized version of around 10000 data points, with a medium-sized database (state\_all.csv) with 25 variables (columns) and 413 observations (rows). After cleaning the data set and mapping the data we realized that 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 dataset 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.

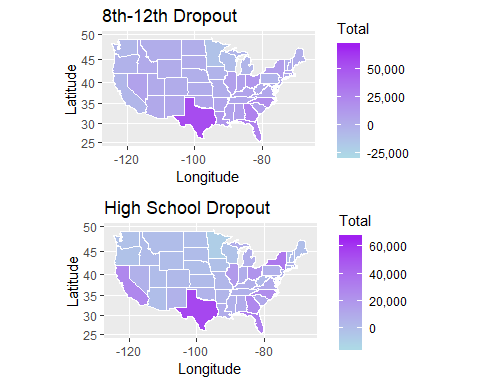
## 4.2 EDA with Maps

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. We first started visualizing the data through maps with ggplot (ggplot2 and ggmap packages) to have different color-coded maps for the following variables:

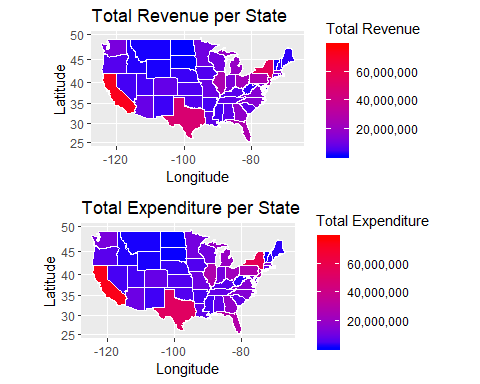
1. Primary school/High school populations



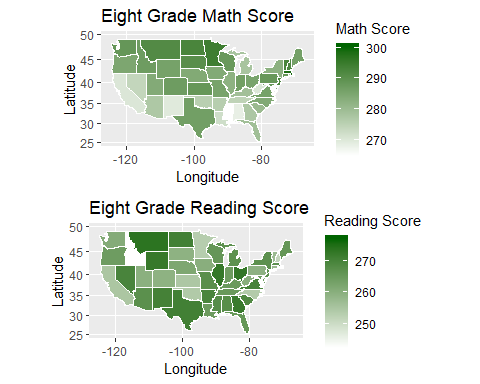
1. Enrollment changes between 8th and 12th grades and between 11th to 12th grade



1. Total revenue/total expenditure



1. Instruction expenditure
2. Math/Reading scores 8th grade

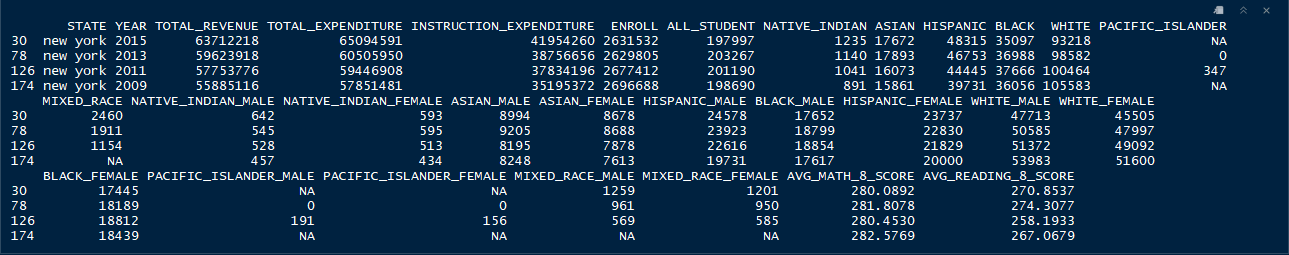


Creating these visualizations was beneficial to make things clearer and easier to understand, especially with such a large and multi-dimensional dataset as ours. This helped for instance map the changes in enrollment between 8th and 12th grade in comparison with 11th and 12th grades. It also allowed us to visualize first general assumptions about correlations between revenue and expenditure. From another perspective, we 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.

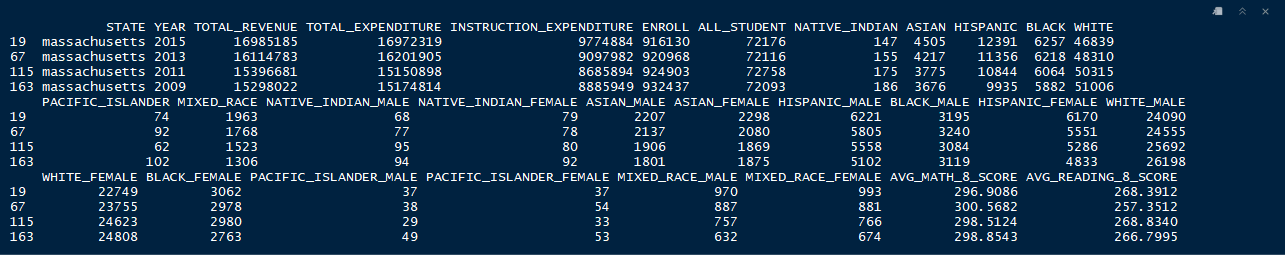
## 4.3 Reshaping the Data

Accurate analyses of large datasets require re-shaping the whole into smaller data subsets. After our general EAD of the extended 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. 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 data sets in chronological terms by recentering mostly on the 2009, 2011, 2013, and 2015 years. We refocused on them because the database appears to provide most complete data points for all variables in those years. Here are screenshots of the resulting datasets for:

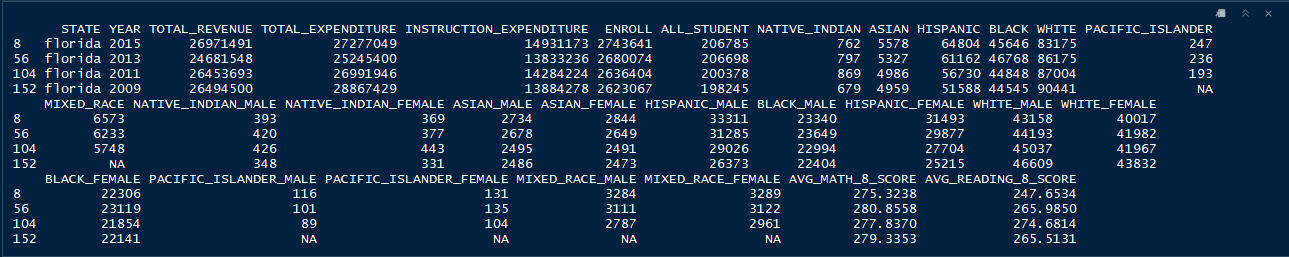
**New York State (our target data)**



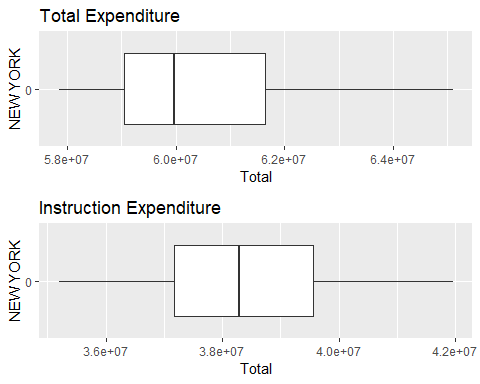
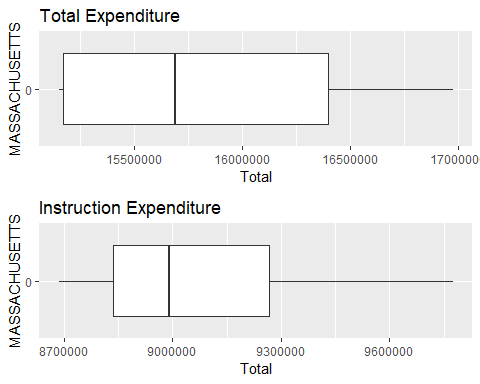
**Massachusetts (max. score comparable)**

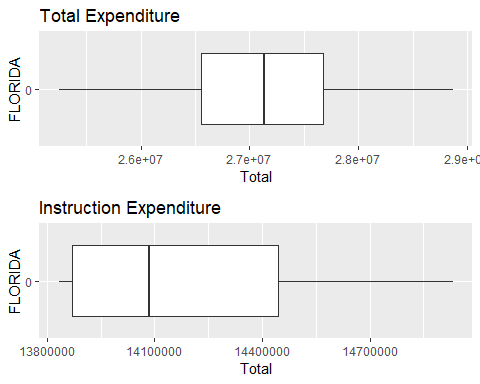


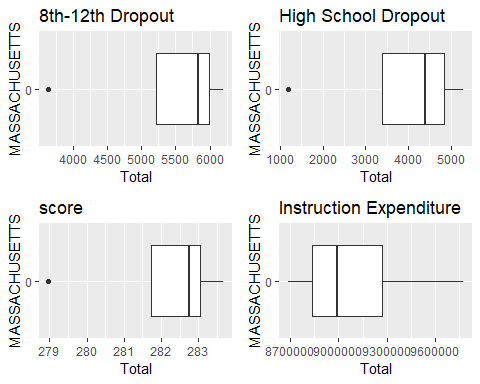
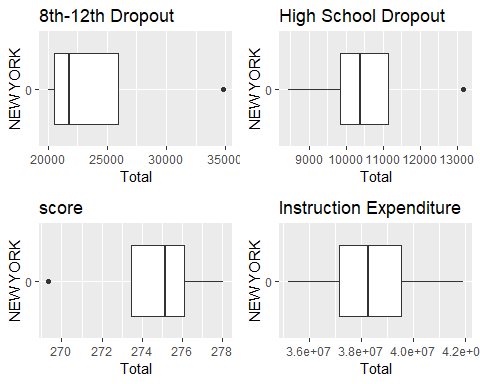
**Florida (min. score comparable)**

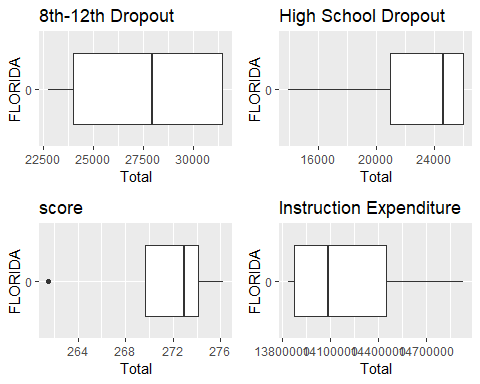


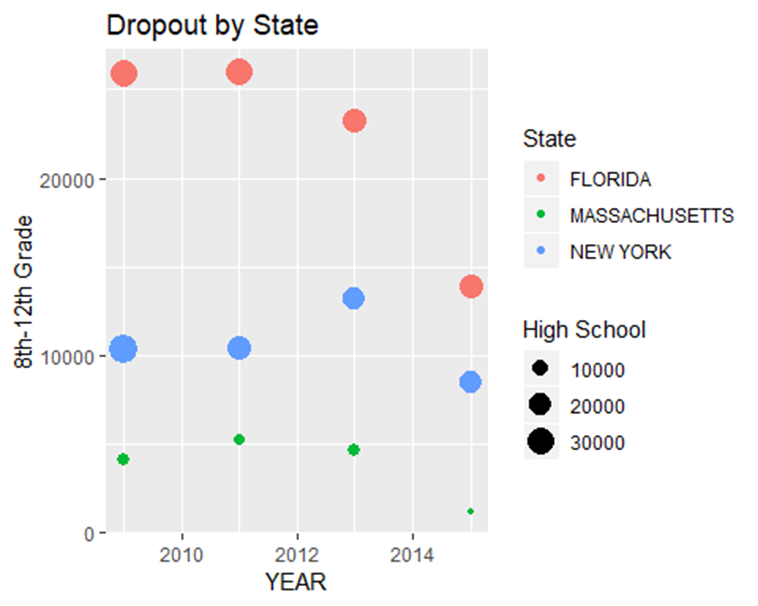
## 4.4 Box Plots and Histograms









# 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

We observed different intensity in the examined correlations in different states as follows.

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).

NY.1 High School retention ~ Instruction expenditure

Call:

lm(formula = HS\_DROP\_OUT ~ INSTRUCTION\_EXPENDITURE, data = dfRetention\_NY)

Residuals:

1 2 3 4

8418 -5367 -10878 7827

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.606e+05 9.418e+04 9.138 0.0118 \*

INSTRUCTION\_EXPENDITURE -5.664e-03 2.446e-03 -2.316 0.1466

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11820 on 2 degrees of freedom

Multiple R-squared: 0.7284, Adjusted R-squared: 0.5925

F-statistic: 5.363 on 1 and 2 DF, p-value: 0.1466

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 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).

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

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

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 also observed if scores correlates to instruction expenditure and we noticed that once again there is not one combination of factors valid for all states. Indeed, this correlation is not strong for New York state and Massachusetts, but has a higher correlation for Florida as the results here 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 observer 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 for the years 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.

# 6) Summary

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).

These are our team’s 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. Comparing New York state with other two states, we observed that for improving student retention, student grades in quantitative skills as well as humanities-oriented areas of study are not as crucial as instruction expenditure.
3. The state of New York might want to consider redistributing expenditure in order to have higher instruction expenditure to have higher retention rate.

# A Appendices

## A.1 Appendix 1

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\_PK\_AM  GRADES\_PK\_AS  GRADES\_PK\_HI  GRADES\_PK\_BL  GRADES\_PK\_WH  GRADES\_PK\_HP  GRADES\_PK\_TR  GRADES\_PK\_AMM  GRADES\_PK\_AMF  GRADES\_PK\_ASM  GRADES\_PK\_ASF  GRADES\_PK\_HIM  GRADES\_PK\_HIF  GRADES\_PK\_BLM  GRADES\_PK\_BLF  GRADES\_PK\_WHM  GRADES\_PK\_WHF  GRADES\_PK\_HPM  GRADES\_PK\_HPF  GRADES\_PK\_TRM  GRADES\_PK\_TRF | GRADES\_KG\_G  GRADES\_KG\_AM  GRADES\_KG\_AS  GRADES\_KG\_HI  GRADES\_KG\_BL  GRADES\_KG\_WH  GRADES\_KG\_TR  GRADES\_KG\_AMM  GRADES\_KG\_HP  GRADES\_KG\_ASM  GRADES\_KG\_AMF  GRADES\_KG\_HIM  GRADES\_KG\_ASF  GRADES\_KG\_HIF  GRADES\_KG\_BLM  GRADES\_KG\_BLF  GRADES\_KG\_WHM  GRADES\_KG\_WHF  GRADES\_KG\_HPM  GRADES\_KG\_HPF  GRADES\_KG\_TRM  GRADES\_KG\_TRF | GRADES\_4\_G  GRADES\_4\_AM  GRADES\_4\_AS  GRADES\_4\_HI  GRADES\_4\_BL  GRADES\_4\_WH  GRADES\_4\_HP  GRADES\_4\_TR  GRADES\_4\_AMM  GRADES\_4\_AMF  GRADES\_4\_ASM  GRADES\_4\_ASF  GRADES\_4\_HIM  GRADES\_4\_BLM  GRADES\_4\_HIF  GRADES\_4\_WHM  GRADES\_4\_WHF  GRADES\_4\_BLF  GRADES\_4\_HPM  GRADES\_4\_HPF  GRADES\_4\_TRM  GRADES\_4\_TRF | GRADES\_8\_G  GRADES\_8\_AM  GRADES\_8\_AS  GRADES\_8\_HI  GRADES\_8\_BL  GRADES\_8\_WH  GRADES\_8\_HP  GRADES\_8\_TR  GRADES\_8\_AMM  GRADES\_8\_AMF  GRADES\_8\_ASM  GRADES\_8\_ASF  GRADES\_8\_HIM  GRADES\_8\_HIF  GRADES\_8\_BLM  GRADES\_8\_BLF  GRADES\_8\_WHM  GRADES\_8\_WHF  GRADES\_8\_HPM  GRADES\_8\_HPF  GRADES\_8\_TRM  GRADES\_8\_TRF |
| **GRADE 1ST– 8TH** | **GRADE 9TH** | **GRADE 9TH -12TH** | **ALL GRADES** |
| GRADES\_1\_8\_G  GRADES\_1\_8\_AM  GRADES\_1\_8\_AS  GRADES\_1\_8\_HI  GRADES\_1\_8\_BL  GRADES\_1\_8\_WH  GRADES\_1\_8\_HP  GRADES\_1\_8\_TR  GRADES\_1\_8\_AMM  GRADES\_1\_8\_AMF  GRADES\_1\_8\_ASM  GRADES\_1\_8\_ASF  GRADES\_1\_8\_HIM  GRADES\_1\_8\_HIF  GRADES\_1\_8\_BLM  GRADES\_1\_8\_BLF  GRADES\_1\_8\_WHM  GRADES\_1\_8\_WHF  GRADES\_1\_8\_HPM  GRADES\_1\_8\_HPF  GRADES\_1\_8\_TRM  GRADES\_1\_8\_TRF | GRADES\_9\_G  GRADES\_9\_AM  GRADES\_9\_AS  GRADES\_9\_HI  GRADES\_9\_BL  GRADES\_9\_WH  GRADES\_9\_HP  GRADES\_9\_TR  GRADES\_9\_AMM  GRADES\_9\_AMF  GRADES\_9\_ASM  GRADES\_9\_ASF  GRADES\_9\_HIM  GRADES\_9\_HIF  GRADES\_9\_BLM  GRADES\_9\_BLF  GRADES\_9\_WHM  GRADES\_9\_WHF  GRADES\_9\_HPM  GRADES\_9\_HPF  GRADES\_9\_TRM  GRADES\_9\_TRF | GRADES\_9\_12\_G  GRADES\_9\_12\_AM  GRADES\_9\_12\_AS  GRADES\_9\_12\_HI  GRADES\_9\_12\_BL  GRADES\_9\_12\_WH  GRADES\_9\_12\_HP  GRADES\_9\_12\_TR  GRADES\_9\_12\_AMM  GRADES\_9\_12\_AMF  GRADES\_9\_12\_ASM  GRADES\_9\_12\_ASF  GRADES\_9\_12\_HIM  GRADES\_9\_12\_HIF  GRADES\_9\_12\_BLM  GRADES\_9\_12\_BLF  GRADES\_9\_12\_WHM  GRADES\_9\_12\_WHF  GRADES\_9\_12\_HPM  GRADES\_9\_12\_HPF  GRADES\_9\_12\_TRM  GRADES\_9\_12\_TRF | GRADES\_ALL\_G  GRADES\_ALL\_AM  GRADES\_ALL\_AS  GRADES\_ALL\_HI  GRADES\_ALL\_BL  GRADES\_ALL\_WH  GRADES\_ALL\_HP  GRADES\_ALL\_TR  GRADES\_ALL\_AMM  GRADES\_ALL\_AMF  GRADES\_ALL\_ASM  GRADES\_ALL\_ASF  GRADES\_ALL\_HIM  GRADES\_ALL\_HIF  GRADES\_ALL\_BLM  GRADES\_ALL\_BLF  GRADES\_ALL\_WHM  GRADES\_ALL\_WHF  GRADES\_ALL\_HPM  GRADES\_ALL\_HPF  GRADES\_ALL\_TRM  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 R code

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> [↑](#footnote-ref-1)
2. Ibid. [↑](#footnote-ref-2)