US Education Descriptive and Predictive Analysis

Ramadan Gannud
Maaz Bin Aftab
DSC 424 Spring 2019

Executive Summary

The standard of education in the United States of America is characterized by its quality and diversity. This includes aspects like socioeconomic backgrounds, number of students enrolled, size of classes, and different types of financial resources provided to schools by government. These aspects were generated using different educational bodies like board of education, which is a collection of members elected by local elections. These bodies' primary responsibility is to check the academic quality of different schools as well as provide funding reports and resources.

The main purpose of this report is to document multivariate statistical techniques that present findings and extract variances important to our study objectives. This research uses these techniques in order to explore meaningful patterns in the data. This educational data set was utilized to study different features of students' achievements in relevance to other aspects. This data was aggregated from all public schools across the US in the years from 1992 to 2017.

The US Educational Dataset is a multivariate data frame sourced from Kaggle. The data was originally sourced from the National Center for Education Statistics (NCES) and the U.S. Census Bureau. The data set is designed to bring different aspects of the US education and see how they interact together. It had 25 variables and a total of 1492 observations from different states. The data is broken down into three different variable types which are Enrollments, Financials and Student Achievements in each state every year.

Different statistical techniques like Multiple Regression Analysis, and Principle Components Analysis were applied in order to study the academic achievements for fourth and eighth grade students. These techniques produce different models and check for similar patterns between different aspects in order to provide insights for different educational organizations and institutions. Principal Component Analysis is a statistical technique used for feature extraction. It's used to emphasize variation and bring out strong patterns in a dataset.

This study finds that factors like State revenue, Federal revenue and Total number of students enrolled in grade 1-8 played an important role in predicting the average math score for 8th grade students, whereas Federal revenue, Local revenue, Total number of students enrolled in Pre-Kindergarten, and Total number of students enrolled in 4th grade were able to predict the average score for the 4th grade students in each state.

In future, different variables could be added to the database using more information about students and their families. Different social and demographic variables could help disclose more patterns and interesting communalities between students and these variables. These aspects could help schools and educational institutions improve their systems and quality of education.

In conclusion, education is a very important element in our society. Knowledge Discovery Process can play a crucial role which allows many Data Mining techniques to extract knowledge from row data and obtain wisdom. In this paper, we found that different scores for different grades vary together. In addition, all financial and school enrollments aspects connect in accordance to certain patterns and they shaped as one component.

Abstract

Higher education standards for the 21st century continues to promote discoveries in the field through learning analytics (LA). The purpose of this study was to examine American students' Math and Reading scores in relevance to number of students and financial expenditures to identify groups exhibiting common patterns of responses. This study investigated the scores of American students through the adoption of students' enrollments in different grades. The scores of students were examined through Linear Regression, Principle Component Analysis, and Logistic Regression. PCA have identified two components and the results indicated that variables in each component are highly correlated. All students' enrollments that include number of students in different grades in addition to financial features belonged to component 1. Students' performance and academic achievements belonged to component 2. The first component has explained 75.5 percent component variability.

Linear regression on the other hand have identified the important variables that could import a good fit in a model. Math scores were used to generate the linear model using other values as predictors. After discarding insignificant variables using different statistical tests that weighs the strength of the evidences, two models were generated. First model consists of 4 predictors (FEDERAL_REVENUE, GRADES_PK_G, LOCAL_REVENUE, GRADES_4_G) and the second model consists of 3 predictors (FEDERAL_REVENUE, STATE_REVENUE, GRADES_1_8_G).

This research may provide practical implications by proposing effective ways to better understand and target students' math and reading scores. Research results may also provide direction for developing successful strategies in improving students' scores.

Introduction

The US Educational Institutions have seen different variations in students' performance due to a number of factors including students' Enrollment, States' revenues, and Expenditures. The study has administered American students at public schools in all American states for 25 years (1993-2017). Data is sourced from the U.S. Census Bureau and the National Center for Education Statistics (NCES). The 'ENROLL' feature represents the U.S. Census Bureau data value (financial data), while the column 'GRADES_ALL' represents the NCES data value (demographic data). Though the two organizations correspond on this matter, these values (which are ostensibly the same) do vary. This dataset consists of 25 variables and 1492 observations. Since identification columns (PRIMARY_KEY, STATE, YEAR) are categorical, they would be removed to perform statistical analysis. NAEP test score data is only available for certain years. Thus, missing values had to be treated for analytical purposes as well. Also, a repeated row was deleted in addition to some rows with zero values. Dimensions of this dataset were dropped to 403 observations and 22 numeric variables.

Principal Component Analysis was applied to educational data to explore the factors thought to be very important in improving the students' performance. PCA was performed in R-mode using correlation and covariance for educational data. PCA chooses a set of variables that are highly correlated with each other to create a subset out of them. PCA chooses these variables to also be quite uncorrelated with other subset of variables which are combined together to form a factor.

Methodology:

Principal Component Analysis (PCA):

principal component analysis technique was implemented in R using correlation and covariance for educational data. Principal component analysis is based on the linear equation as follows:

$Yij = \beta_{i1}X_{1j} + \beta_{i2}X_{2j} + \beta_{i3}X_{3j} + \ldots + \beta_{ip}X_{pj}$ where i, j = 1, 2, … p.

First, correlation matrix was used to identify correlations between different variables. In the correlation matrix, the sizes of some correlation place constraints on the sizes of others. To illustrate, high bi-variate correlation matrix provides components. Second, it's important to look at frameworks of individual correlations where match acute relationships are balanced due to correlation between two variables. Third, under certain conditions, matrices can be diagonalized. The diagonalization of a matrix gives a matrix with positive numbers on the main diagonal and all non-diagonal elements become zeros. These positive elements represent the variances from the correlation matrix. This can be written as follows: L =V' RV.

The columns of matrix V are called eigenvectors, while the elements on the main diagonal of L matrix are eigenvalues. Each eigenvector corresponds to each eigenvalue. Factor Analysis keep only components with correlated variables by keeping the ones that have large corresponding eigenvalues which are normally the ones with eigenvalue greater than 1. Finally, when these correlation matrices are diagonalized, the information is repackaged. The resulting correlation matrix is determined using the product of the matrices of eigenvectors and the square root of the eigenvalues as follows: R = (V L) (LV'). This equation in brief represent the product of factor loading matrix and its transpose.

The communality for a variable is the sum of squared loading (SSL) across components which is the proportion of variance accounted for by the components. It can be calculated by dividing the sum of square loadings (SSL) with the number of variables. Furthermore, the covariance proportion for a component is obtained by dividing the sum of square loadings (SSL) for this component with the sum of communalities.

Multiple Regression Analysis:

Multiple Regression analysis technique on the educational dataset was conducted using R as well. The main purpose of this technique was to learn more about the relationships between different independent (Predictor) variables and the dependent (Response) variable. The main objective goal of multiple regression is to fit a straight line to a number of points so that the squared standard deviations of the observed values from that line are minimized. Hence, it's also sometimes known as the least square's estimation method. The line is basically a two-dimensional space that is define by the equation **Y= a + b*X.** Here, Y is the dependent variable which can be expressed in terms of the constant (a) and a slope (b) times the X as the dependent variable. In case of multivariate analysis, there exists to be more than one predicting or X variables therefore the line could not be visualized in 2d space but can be computed in the exact manner. Generally, the multiple regression will estimate a linear equation of the form:

$\mathbf{Y = a + b_1*X_1 + b_2*X_2 + ... + b_p*X_p}$.

The multiple regression analysis represents the independent contributions of each predicting variable to the prediction of response variables. Another way to represent it is if a dependent variable Y is correlated to independent variable X, the amount of change in Y could be interpreted by each individual independent variable X by having other variables constant. R squared which is also known as the coefficient of determination is generally used to evaluate the fitness of a model. R squared is equal to 1 minus the ratio of residual variability. The Coefficient of determination measures the proportion of variance in the independent variables explained by the dependent variable.

In statistical analysis the correlation coefficient R measures the degree to which two or more independent (X) variables are related to the dependent (Y). While performing multiple regression analysis, different assumptions must be held, multiple collinearity is one of them. The issue of multiple collinearity arises when there is a significant relationship between independent variables. These independent variables are considered correlated and cause problems when fitting the model and interpreting the results. Multicollinearity can be diagnosed by using scatter plots or correlation coefficient matrix between X-variables. Correlation coefficients that are 0.7 or higher considered as strong multicollinearity. Also, multicollinearity can be detected using the variance inflation factor VIF. If VIF is bigger than 10, that means the multicollinearity is problematic.

Discussion and Results

Descriptive Analysis:

In this study we discuss the results of principal component analysis, regression analysis, and logistic regression for 8 enrollment features of students in different grades, 10 state financials, and 4 students' math and reading scores. The following table (table) shows the descriptive statistics of the mentioned features.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
ENROLL	1	403	953349.397	1097437.229	671715.000	727023.350	562181.164	84146.000	6307022.000	6222876.000	2.894
TOTAL_REVENUE	2	403	10575588.650	12748786.152	6158756.000	7820613.183	5759293.134	812982.000	73958896.000	73145914.000	2.634
FEDERAL_REVENUE	3	403	950082.886	1297672.440	559732.000	668271.139	495834.814	65052.000	9990221.000	9925169.000	3.735
STATE_REVENUE	4	403	4891952.536	5905222.361	3014993.000	3729114.588	2702951.782	303132.000	42333637.000	42030505.000	3.351
LOCAL_REVENUE	5	403	4733553.208	6028822.097	2446818.000	3385247.223	2658549.394	28531.000	34941513.000	34912982.000	2.255
TOTAL_EXPENDITURE	6	403	10685658.084	13003620.251	6318241.000	7845005.495	5920174.508	802768.000	74766086.000	73963318.000	2.650
INSTRUCTION_EXPENDITURE	7	403	5530190.278	6846025.028	3197895.000	4033086.031	2939748.206	402602.000	41954260.000	41551658.000	2.786
SUPPORT_SERVICES_EXPENDITURE	8	403	3117919.243	3608489.804	1907907.000	2361999.118	1736504.146	204117.000	21693675.000	21489558.000	2.470
OTHER_EXPENDITURE	9	403	483958.169	567314.956	336579.000	369287.149	307055.356	27608.000	3759373.000	3731765.000	2.946
CAPITAL_OUTLAY_EXPENDITURE	10	403	1046100.980	1485925.129	616478.000	714303.607	568932.924	20070.000	10223657.000	10203587.000	3.370
GRADES_PK_G	11	403	22571.486	34607.711	12795.000	15735.690	15677.012	399.000	249524.000	249125.000	4.171
GRADES_KG_G	12	403	72578.861	84140.635	52937.000	55584.752	45355.699	5863.000	506831.000	500968.000	2.968
GRADES_4_G	13	403	73017.223	84305.706	52715.000	55917.449	44716.699	5975.000	493415.000	487440.000	2.931
GRADES_8_G	14	403	73722.548	84149.040	52289.000	56654.523	44455.761	6173.000	500143.000	493970.000	2.863
GRADES_12_G	15	403	64737.318	74252.914	41836.000	50124.276	39919.005	6079.000	498403.000	492324.000	3.104
GRADES_1_8_G	16	403	588898.896	678710.221	423510.000	450592.211	362910.828	48722.000	3929869.000	3881147.000	2.897
GRADES_9_12_G	17	403	288712.007	331928.808	191816.000	221406.594	165842.153	26002.000	2011865.000	1985863.000	2.919
GRADES_ALL_G	18	403	900182.390	1036925.051	636640.000	687973.944	540925.127	78028.000	5926631.000	5848603.000	2.884
AVG_MATH_4_SCORE	19	403	237.223	8.023	238.678	237.957	6.912	209.021	253.421	44.400	-0.897
AVG_MATH_8_SCORE	20	403	280.284	8.347	281.607	280.726	7.528	252.380	300.568	48.188	-0.498
AVG_READING_4_SCORE	21	403	219.879	6.679	221.169	220.253	6.139	199.702	236.774	37.072	-0.470
AVG_READING_8_SCORE	22	403	263.659	6.909	265.139	264.227	6.081	236.379	277.191	40.812	-0.944

It is evident from the table above that the eighth grader math and reading results for students who took the National Assessment of Educational Progress (NAEP) for states are better in the eight grade students than they are in the fourth grader. The average math score for the 4th graders is 237 while it's 280 for the 8th grader students. Furthermore, the average reading for results in the 4th graders is 219 while it's 264 for the 8th graders.

Correlation coefficients between 8 enrollment features of students in different grades are observed from the table below which shows high positive correlation between these variables. Also, these variables are still high positively correlated with different state financial variables. In addition, the 10 different financial variables are observed to be positively high correlated as well. Students' results seem to be significantly correlated with each other except for the average reading scores for 8 graders which have no relationship with any of the other math and reading scores. Moreover, the 4 students' math and reading scores variables have no relationship with the rest of the financial and enrollment features in the dataset.

	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE	STATE_REVENUE	LOCAL_REVENUE	TOTAL_EXPENDITURE
ENROLL	1.000	0.923	0.923	0.925	0.848	0.926
TOTAL_REVENUE	0.923	1.000	0.923	0.972	0.964	0.999
FEDERAL_REVENUE	0.923	0.923	1.000	0.920	0.836	0.923
STATE_REVENUE	0.925	0.972	0.920	1.000	0.877	0.970
LOCAL_REVENUE	0.848	0.964	0.836	0.877	1.000	0.964
TOTAL_EXPENDITURE	0.926	0.999	0.923	0.970	0.964	1.000
INSTRUCTION_EXPENDITURE	0.883	0.989	0.883	0.953	0.969	0.990
SUPPORT_SERVICES_EXPENDITURE	0.929	0.995	0.927	0.966	0.958	0.993
OTHER_EXPENDITURE	0.968	0.940	0.948	0.943	0.860	0.940
CAPITAL_OUTLAY_EXPENDITURE	0.934	0.920	0.902	0.917	0.852	0.928
GRADES_PK_G	0.789	0.711	0.732	0.637	0.723	0.720
GRADES_KG_G	0.997	0.908	0.920	0.912	0.830	0.910
GRADES_4_G	0.998	0.913	0.916	0.915	0.837	0.915
GRADES_8_G	0.999	0.921	0.921	0.924	0.845	0.923
GRADES_12_G	0.986	0.943	0.947	0.951	0.859	0.943
GRADES_1_8_G	0.999	0.914	0.916	0.915	0.839	0.916
GRADES_9_12_G	0.997	0.935	0.934	0.940	0.856	0.936
GRADES_ALL_G	0.999	0.921	0.923	0.921	0.847	0.923
AVG_MATH_4_SCORE	-0.077	0.083	0.018	0.030	0.142	0.079
AVG_MATH_8_SCORE	-0.121	0.024	-0.065	-0.036	0.100	0.021
AVG_READING_4_SCORE	-0.143	0.004	-0.131	-0.070	0.104	0.000
AVG_READING_8_SCORE	-0.014	0.012	0.002	0.009	0.016	0.015

	INSTRUCTION_EXPENDITURE	SUPPORT_SERVICES_EXPENDITURE	OTHER_EXPENDITURE	CAPITAL_OUTLAY_EXPENDITURE
ENROLL	0.883	0.929	0.968	0.934
TOTAL_REVENUE	0.989	0.995	0.940	0.920
FEDERAL_REVENUE	0.883	0.927	0.948	0.902
STATE_REVENUE	0.953	0.966	0.943	0.917
LOCAL_REVENUE	0.969	0.958	0.860	0.852
TOTAL_EXPENDITURE	0.990	0.993	0.940	0.928
INSTRUCTION_EXPENDITURE	1.000	0.975	0.897	0.884
SUPPORT_SERVICES_EXPENDITURE	0.975	1.000	0.947	0.915
OTHER_EXPENDITURE	0.897	0.947	1.000	0.921
CAPITAL_OUTLAY_EXPENDITURE	0.884	0.915	0.921	1.000
GRADES_PK_G	0.671	0.720	0.749	0.767
GRADES_KG_G	0.866	0.913	0.964	0.922
GRADES_4_G	0.871	0.918	0.963	0.928
GRADES_8_G	0.880	0.927	0.966	0.932
GRADES_12_G	0.899	0.950	0.978	0.930
GRADES_1_8_G	0.872	0.919	0.963	0.929
GRADES_9_12_G	0.893	0.941	0.974	0.937
GRADES_ALL_G	0.879	0.927	0.967	0.933
AVG_MATH_4_SCORE	0.086	0.089	0.011	0.006
AVG_MATH_8_SCORE	0.029	0.028	-0.066	-0.053
AVG_READING_4_SCORE	0.028	0.003	-0.096	-0.104
AVG_READING_8_SCORE	0.022	0.011	0.000	-0.005

	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
ENROLL	0.789	0.997	0.998	0.999	0.986	0.999	0.997	0.999
TOTAL_REVENUE	0.711	0.908	0.913	0.921	0.943	0.914	0.935	0.921
FEDERAL_REVENUE	0.732	0.920	0.916	0.921	0.947	0.916	0.934	0.923
STATE_REVENUE	0.637	0.912	0.915	0.924	0.951	0.915	0.940	0.921
LOCAL_REVENUE	0.723	0.830	0.837	0.845	0.859	0.839	0.856	0.847
TOTAL_EXPENDITURE	0.720	0.910	0.915	0.923	0.943	0.916	0.936	0.923
INSTRUCTION_EXPENDITURE	0.671	0.866	0.871	0.880	0.899	0.872	0.893	0.879
SUPPORT_SERVICES_EXPENDITURE	0.720	0.913	0.918	0.927	0.950	0.919	0.941	0.927
OTHER_EXPENDITURE	0.749	0.964	0.963	0.966	0.978	0.963	0.974	0.967
CAPITAL_OUTLAY_EXPENDITURE	0.767	0.922	0.928	0.932	0.930	0.929	0.937	0.933
GRADES_PK_G	1.000	0.790	0.791	0.781	0.740	0.791	0.765	0.796
GRADES_KG_G	0.790	1.000	0.998	0.996	0.982	0.998	0.993	0.997
GRADES_4_G	0.791	0.998	1.000	0.998	0.982	1.000	0.995	0.999
GRADES_8_G	0.781	0.996	0.998	1.000	0.986	0.999	0.998	0.999
GRADES_12_G	0.740	0.982	0.982	0.986	1.000	0.983	0.995	0.986
GRADES_1_8_G	0.791	0.998	1.000	0.999	0.983	1.000	0.995	1.000
GRADES_9_12_G	0.765	0.993	0.995	0.998	0.995	0.995	1.000	0.997
GRADES_ALL_G	0.796	0.997	0.999	0.999	0.986	1.000	0.997	1.000
AVG_MATH_4_SCORE	0.049	-0.088	-0.090	-0.077	-0.021	-0.087	-0.054	-0.073
AVG_MATH_8_SCORE	0.015	-0.132	-0.130	-0.123	-0.068	-0.128	-0.102	-0.116
AVG_READING_4_SCORE	-0.055	-0.161	-0.155	-0.146	-0.106	-0.153	-0.129	-0.143
AVG_READING_8_SCORE	0.014	-0.009	-0.012	-0.017	-0.010	-0.013	-0.014	-0.013

	AVG_MATH_4_SCORE	AVG_MATH_8_SCORE	AVG_READING_4_SCORE	AVG_READING_8_SCORE
ENROLL	-0.077	-0.121	-0.143	-0.014
TOTAL_REVENUE	0.083	0.024	0.004	0.012
FEDERAL_REVENUE	0.018	-0.065	-0.131	0.002
STATE_REVENUE	0.030	-0.036	-0.070	0.009
LOCAL_REVENUE	0.142	0.100	0.104	0.016
TOTAL_EXPENDITURE	0.079	0.021	0.000	0.015
INSTRUCTION_EXPENDITURE	0.086	0.029	0.028	0.022
SUPPORT_SERVICES_EXPENDITURE	0.089	0.028	0.003	0.011
OTHER_EXPENDITURE	0.011	-0.066	-0.096	0.000
CAPITAL_OUTLAY_EXPENDITURE	0.006	-0.053	-0.104	-0.005
GRADES_PK_G	0.049	0.015	-0.055	0.014
GRADES_KG_G	-0.088	-0.132	-0.161	-0.009
GRADES_4_G	-0.090	-0.130	-0.155	-0.012
GRADES_8_G	-0.077	-0.123	-0.146	-0.017
GRADES_12_G	-0.021	-0.068	-0.106	-0.010
GRADES_1_8_G	-0.087	-0.128	-0.153	-0.013
GRADES_9_12_G	-0.054	-0.102	-0.129	-0.014
GRADES_ALL_G	-0.073	-0.116	-0.143	-0.013
AVG_MATH_4_SCORE	1.000	0.884	0.753	0.079
AVG_MATH_8_SCORE	0.884	1.000	0.806	0.067
AVG_READING_4_SCORE	0.753	0.806	1.000	0.009
AVG_READING_8_SCORE	0.079	0.067	0.009	1.000

Principle Component Analysis:

These correlation coefficients that appear to be strong between some variables and week between some others yield the principle component analysis technique. These highly correlated variables then can be combined together to form different components. In PCA, variances accounted for different factors initialize and extract communalities for different variables. The cumulative proportion of variances accounted for the factors is always equal to 1.0. These communalities are the proportion of variance for each variable by the rest of the variables. Using the *prcomp* in R, principle components and their importance could be generated as follows:

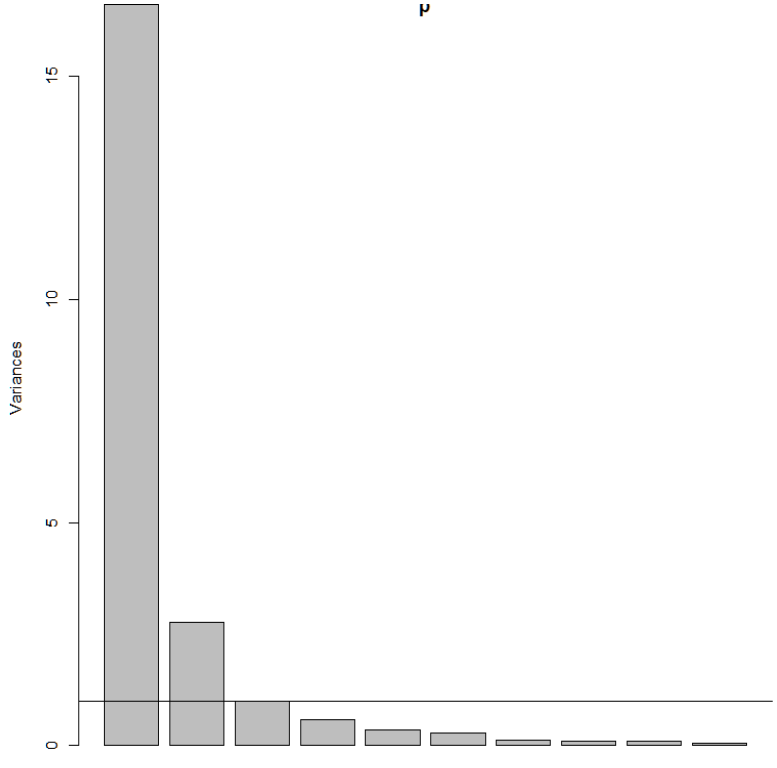
```
Importance of components:
                            PC1       PC2       PC3       PC4       PC5       PC6       PC7       PC8       PC9  PC10  PC11  PC12
Standard deviation     2.152e+07 2.167e+06 1.069e+06 6.270e+05 4.626e+05 4.305e+05 2.841e+05 1.631e+05 9.559e+04 23566 20905 15097
Proportion of Variance 9.856e-01 9.990e-03 2.430e-03 8.400e-04 4.600e-04 3.900e-04 1.700e-04 6.000e-05 2.000e-05     0     0     0
Cumulative Proportion  9.856e-01 9.956e-01 9.981e-01 9.989e-01 9.993e-01 9.998e-01 9.999e-01 1.000e+00 1.000e+00     1     1     1
                          PC13 PC14 PC15 PC16  PC17  PC18  PC19  PC20   PC21      PC22
Standard deviation     5225 3708 2170 1320 10.27 6.766 3.114 2.419 0.1963 3.463e-10
Proportion of Variance    0    0    0    0  0.00 0.000 0.000 0.000 0.0000 0.000e+00
Cumulative Proportion     1    1    1    1  1.00 1.000 1.000 1.000 1.0000 1.000e+00
```

According to the table above, 22 components were needed using the covariance to explain 100% of total variation for this data. The first component has accounted for 98.5% of the variance which is considered very high. The second component accounts for less than 1 percent of the variance.

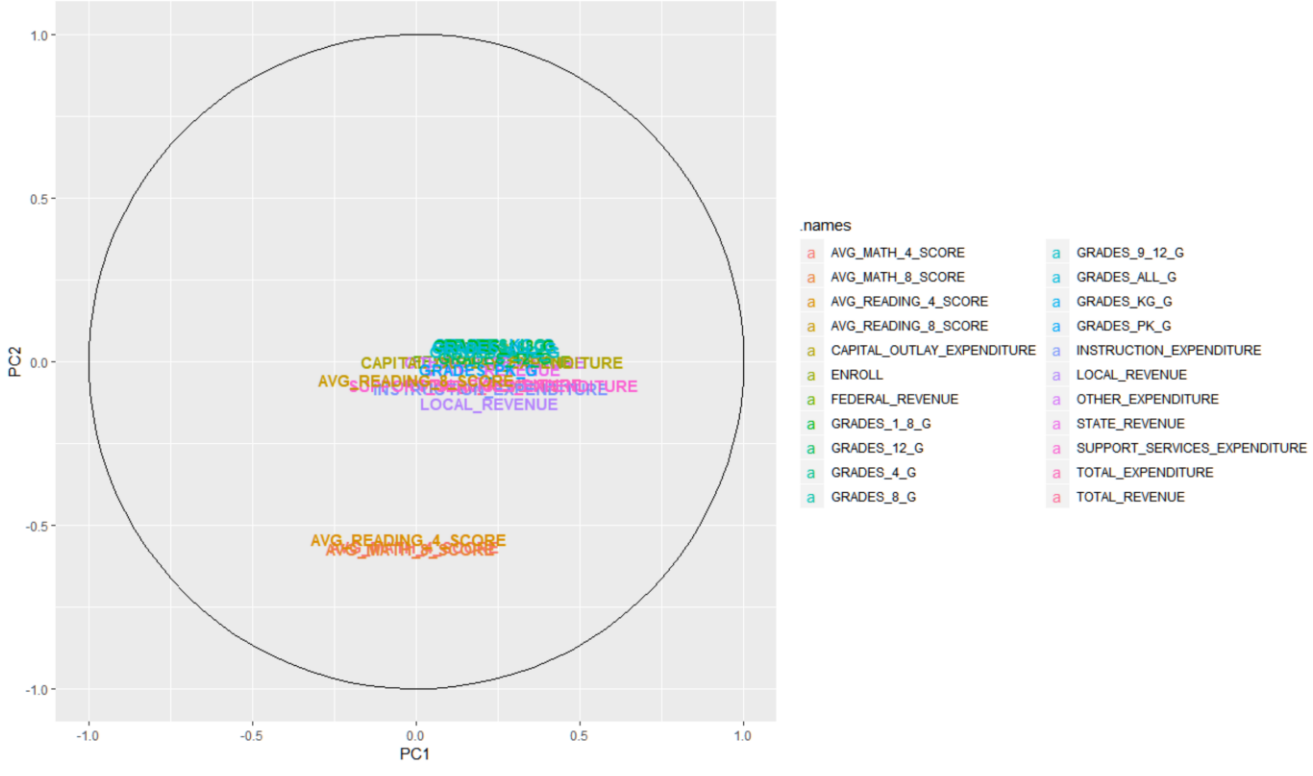
After conducting PCA with the correlation matrix using TRUE scale, the importance of components has changed. The proportion of variance in the first component has decreased from 0.985 to 0.755. The proportion of variance for the other components have changed as well. The cumulative proportion of variance that is accounted for the first two components is 0.88 and 0.92 for the first 3 components as it's shown in the following table. The standard deviations have also dropped using the PCA with the correlation matrix. PCA on correlation is much more informative and reveals some structure in the data and relationships between variables.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     4.0772 1.6590 0.99983 0.75748 0.58838 0.53339 0.34762 0.31737 0.30087 0.21089 0.17896 0.13128 0.07880 0.05696
Proportion of Variance 0.7556 0.1251 0.04544 0.02608 0.01574 0.01293 0.00549 0.00458 0.00411 0.00202 0.00146 0.00078 0.00028 0.00015
Cumulative Proportion  0.7556 0.8807 0.92615 0.95224 0.96797 0.98090 0.98640 0.99097 0.99509 0.99711 0.99857 0.99935 0.99963 0.99978
                          PC15    PC16    PC17    PC18    PC19     PC20     PC21     PC22
Standard deviation     0.04752 0.03459 0.02558 0.01929 0.01672 0.009405 2.558e-08 7.356e-17
Proportion of Variance 0.00010 0.00005 0.00003 0.00002 0.00001 0.000000 0.000e+00 0.000e+00
Cumulative Proportion  0.99988 0.99994 0.99997 0.99998 1.00000 1.000000 1.000e+00 1.000e+00
```

Scree plot has determined 10 components. The number of components with a variance higher than 1 in the scree plot are 2 with a big difference in variance between them as it's shown in the figure below.



The next plot shows the principle components and how different variables are spread around these components. It's clear how correlated variables cluster together to form different components.



Factors that have small variances do not fit in the principle components solution and should be discarded from the analysis. According to the scree plot and the importance of components table, only the first two components should be retained and discard the rest because they explain a very small proportion of the variance. These first two factors account for 88 percent of the variance. In addition, every eigenvalue corresponds to a different component and usually only factor with large eigenvalues are kept for the analysis.

After rotating the component matrix using varimax function in R, the component loading matrix then shows the correlation between components and variables. The first column shows the correlation between the first component and different variables. The second column is the correlation between the second component and different variables as well. Using a cutoff of 0.4 would reveal strong correlation between components and their respective variables. These variables were originally correlated and combined to form these components as it was stated previously. Components are interpreted using the variables that are highly correlated with it and have loadings bigger than 0.4.

```
Loadings:
                              RC1    RC2
ENROLL                         0.983
TOTAL_REVENUE                  0.975
FEDERAL_REVENUE                0.949
STATE_REVENUE                  0.960
LOCAL_REVENUE                  0.916
TOTAL_EXPENDITURE              0.976
INSTRUCTION_EXPENDITURE        0.944
SUPPORT_SERVICES_EXPENDITURE   0.976
OTHER_EXPENDITURE              0.978
CAPITAL_OUTLAY_EXPENDITURE     0.952
GRADES_PK_G                    0.788
GRADES_KG_G                    0.976
GRADES_4_G                     0.978
GRADES_8_G                     0.982
GRADES_12_G                    0.986
GRADES_1_8_G                   0.979
GRADES_9_12_G                  0.987
GRADES_ALL_G                   0.983
AVG_MATH_4_SCORE                      0.938
AVG_MATH_8_SCORE                      0.951
AVG_READING_4_SCORE                   0.904
AVG_READING_8_SCORE

                 RC1   RC2
SS loadings    16.608 2.768
Proportion Var  0.755 0.126
Cumulative Var  0.755 0.881
```

To show that these components are giving information with less correlation between each other, correlation matrix could be performed on the scores of these two different components after calculating scores. The score correlation matrix shows an insignificant negative relationship between both components with a correlation coefficient equals to -0.19.

```
            RC1        RC2
RC1  1.0000000 -0.1923796
RC2 -0.1923796  1.0000000
```

Regression Analysis:

To apply multiple regression analysis on the data, the research question was broken down into two parts to see if different independent variables have any impact on the average math score for 4^{th} and 8^{th} grade students. Therefore, two different models were initially run; the first one is to predict the math score of 4^{th} graders as response variables and the other model uses the math scores of 8th graders as a response variable. According to the correlation table above in the descriptive analysis above, many independent variables were highly correlated with each other. Hence, in order to account for multicollinearity, some variables were removed to reduce the confidence intervals of the coefficients. Both manual and feature selection methods were used to come up with the best fit for the model. The values of Variation Inflation Factor (VIF) were also observed using the *DescTools* package in R to avoid any multicollinearity issues. Both Backward and Stepwise selection methods would not be appropriate to use in this case since they don not account for multicollinearity issues. Models were then compared on the basis of their adjusted R-squares to determine which model can predict these dependent variables the most. Due to the sensitivity of the data, the multiple regression approach was used more as an exploratory analysis technique rather than predicting values. Hence, the data was not split into training and testing data. With the help of multiple regression analysis, two different models were finalized and equations for both models are given below. Also, the multiple regression results are shown below.

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.798e+02  6.213e-01 450.342  < 2e-16 ***
education$FEDERAL_REVENUE -7.251e-08  1.047e-06  -0.069 0.944818
education$STATE_REVENUE    7.891e-07  2.292e-07   3.443 0.000635 ***
education$GRADES_1_8_G    -6.863e-06  1.955e-06  -3.510 0.000498 ***
```

AVERAGE_MATH_4_SCORE = 2.367e+02 + 1.846e-06*FEDERAL_REVENUE +4.278e-05* GRADES_PK_G + 8.161e-07*LOCAL_REVENUE -9.150e-05* GRADES_4_G

AVERAGE_MATH_8_SCORE = 2.798e+02 -7.251e-08*FEDERAL_REVENUE + 7.891e-07* STATE_REVENUE -6.863e-06*GRADES_1_8_G

Different methods of multivariate analysis were applied to see if some interesting insights can be uncovered around students' results in math and reading. There might be more latent variables can be added for analysis to improve the model. This research may contribute to existing decision-making literature by providing more evidence and suggestions for adaptation in applying this framework to understand college students' performance across different school grades in different states. However, this research could be expanded by providing assessment data for different backgrounds and cultures. Moreover, expand all data sources to work at the district level, in addition to the state level. Future work could also use a version of financial data that accounts for inflation as well.

Conclusion:

The principal Component Analysis was used to explore the effects of different enrollment and financial variables which are insignificantly affecting the students' math and reading scores for 4th and 8th grades. In the case of finding financial variables are highly correlated with each other and with enrollment variables as well. This explains that the more student enrollments there are, the more the population is and the higher the financial income and expenditures are. To know the variability between the components, it was found that first two components account for exactly 88.6% of the total variability. These two components adequately describe the whole correlation matrix.

The multiple regression analysis was able to identify different predicting variables that could cause noticeable impacts yet small on the average math scores for students in different states. The financial status and the numbers of students' enrollments in each state played a vital role in determining the students' academic performance.

Finally, it was obvious to see that states' financials, expenditures, and students' enrollments had high loadings in component 1. In component 2, it was found that math and reading results had the high scores. The important finding of this study is that the scores vary together for different grades in different states and different years. On the other hand, all financials, number of enrollments as well vary together in different states and different years.