1 Background, problem, and questions to be answered:

The public, private, and home schools provide education in the United States. State governments set overall educational standards, often mandate standardized tests for K–12 public school systems, and supervise, usually through a board of regents, state colleges and universities. Funding comes from the state, local, and the federal government. Private schools are generally free to determine their own curriculum and staffing policies, with voluntary accreditation available through independent regional accreditation authorities. About 87% of school-age children attend public schools, about 10% attend private schools, and roughly 3% are home-schooled.

The United States spends more per student on education than any other country. In 2014, the Pearson/Economist Intelligence Unit rated US education as 14th best in the world, just behind Russia. In 2015, the Programme for International Student Assessment rated U.S. high school students #40 Globally in Math and #24 in Science and Reading.

It has been alleged, since the 1950s and especially in recent years that American schooling is undergoing a crisis in which academic performance is behind other countries, such as Russia, Japan, or China, in core subjects. Congress passed the National Defense Education Act in 1958 in an attempt to rectify these problems and a series of other legislative acts in later decades such as No Child Left Behind. According to the Organization for Economic Cooperation and Development, however, American students of 2012 ranked 25th in math, 17th in science, and 14th in reading compared with students in 27 other countries. In 2013, Amanda Ripley published The Smartest Kids in the World (And How They Got That Way), a comparative study of how the American education system differs from top-performing countries such as Finland and South Korea.

Despite the demonstrated link between economic growth and education standards, high schools and colleges sharply disagree about the college readiness of high school graduates, in that 44% of college faculty believe that incoming students are not ready for writing at the college level, while 90% of high school teachers believe exiting students are well prepared.

The first school compulsory attendance law in the United States was enacted in Massachusetts in 1852. Other states soon followed, but it was not until the early 20th century that all states in the union had such a law. In the 19th and early 20th century the enforcement of compulsory

school attendance laws by the school or government officials was usually quite sketchy, but in those years, the student who dropped out of school had a reasonably good chance of finding a job, since there was high demand for semi-skilled and even unskilled laborers. The nature of the American workplace as it now exists puts the young person without a high school diploma who is seeking a job with an adequate wage in a precarious situation. Even a high school diploma may not be sufficient to put the young person on a good path with the opportunity to move up the employment ladder. Additional education may be required, whether it is a community college, a technical certification program, or a four-year college.

Every year, over 1.2 million students drop out of high school in the United States alone. That's a student every 26 seconds or 7,000 a day. About 25% of high school freshmen fail to graduate from high school on time. The U.S., which had some of the highest graduation rates of any developed country, now ranks 22nd out of 27 developed countries. In 2010, 38 states had higher graduation rates. Vermont had the highest rate, with 91.4% graduating. And Nevada had the lowest with 57.8% of students graduating.

For this project, my focuses is on major educational issues in the United States and analyze the root cause of high school dropout issue and provided a detailed report on the following:

1. A summarized visualization of high school dropout factories by location
2. A comparison of 3-5 most high-school graduation rate district against 3-5 least high-school graduation rate district
3. Identifying the most prominent race in each district having least high-school graduation rate
4. Identifying the district with the most change in high-school graduation rate and potential reasons for the change
5. Building a predictive model of most high-school graduation rate in each district using machine learning
6. Finally, after addressing #5, I wanted to identify the most salient features/variables used by the model for predicting most high-school graduation rate, within the limitations of my dataset

2 Potential Clients

There are two different types of clients that could be interested in the findings from this project. The first type of clients would be the US online and print media that cover socioeconomic and urban issues. These clients are magazines that take an active interest in stories driven by socially relevant issues and are backed by data analytics, for creating awareness within the public while simultaneously enhancing the quality of their readership. For example, US online media such as US News and Gates Foundation would fall under this category. I also anticipate interest from Government funded bodies and non-profits offering job placement services, and subsidized education services for youth and adults.

3 Datasets used, data wrangling, and data exploration

EDFacts[1] is a U.S. Department of Education (ED) initiative to collect, analyze, and promote the use of high-quality, pre-kindergarten through grade 12 data. These datasets are organized into 4 different categories. I used two datasets from EDFacts portal consisting of District & school level statistics on graduation rates and performance on math & reading/language art assessments by race/ethnicity, gender, disability, English proficiency, socioeconomic status, and homeless & migrant status[2] for two years 2011 and 2014.

3.1 Reading in data

Each dataset was provided as a raw dataset in CSV format for 2011 and 2014, which are imported as Pandas data frame. Each raw dataset resulted in two pandas data frames. Initially, I did not foresee any use for the graduation data as I felt that the adjusted cohort graduation rates (ACGR) datasets would be sufficient to address my problem.

The following formula provides an example of how the four-year adjusted cohort graduation rate would be calculated for the cohort entering 9th grade for the first time in the 2011-12 school year and graduating by the end of the 2014-15 school year:

Formula for Calculating the Four-Year Adjusted-Cohort Graduation Rate

$$\frac{\text{Number of cohort members who earned a regular high school diploma by the end of the 2014-15 school year}}{\text{Number of first-time 9th graders in fall 2011 (starting cohort) plus students who transferred in, minus students who transferred out, emigrated, or died during school years 2011-12, 2012-13, 2013-14, and 2014-15}}$$

3.2 Initial data exploration

After initial data exploration, I have found that adjusted cohort graduation rate is present in one dataset but there are 3 more datasets which has the remaining variables I am considering for analysis. Hence, I had to work on merging 4 datasets and coming up with a single dataset which can be used in the analysis.

I checked the first five rows in the adjusted cohort graduation rate data frames for both 2011 and 2014, and noticed that some columns are having suppressing data for very small groups of students to protect individual student's identity. I realized this would pose a challenge for making comparisons between 2011 and 2014.

[1] https://www.ed.gov

[2] https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html

The first five rows of the 2011 adjusted cohort graduation rates dataframe are shown below:

	STNAM	FIPST	LEAID	LEANM	ALL_COHORT_1112	ALL_RATE_1112	MAM_COHORT_1112	MAM_RATE_1112	MAS_COHORT_1112	MAS_RATE_1112	...
0	ALABAMA	01	0100005	Albertville City	268	83	NaN	NaN	NaN	NaN	...
1	ALABAMA	01	0100006	Marshall County	424	79	2	PS	1	PS	...
2	ALABAMA	01	0100007	Hoover City	1042	91	1	PS	71	85-89	...
3	ALABAMA	01	0100008	Madison City	836	91	4	PS	44	GE90	...
4	ALABAMA	01	0100011	Leeds City	117	70-74	NaN	NaN	NaN	NaN	...

3.3 Checking for missing values

Because it is often easy to identify specific individuals when data are presented for a very small number of students, the graduation rate has been suppressed for all subgroups for which there are 1-5 students in the cohort. These suppressions are identified by 'PS'. To further protect the privacy of students, and to prevent any data suppressed in Step One from being recalculated by subtracting other reported groups data from the "All Students" group, the Education Department has reported the graduation rates for all medium-sized groups as a range (e.g., <20% or 70-74%)

The magnitude of the reported ranges is determined by the size of the group whose data are being reported. For example, subgroups with the fewest students (6-15) are reported with the widest ranges (e.g., <50% or ≥50%). As the number of students in the group increases, the magnitude of the range decreases, until there are more than 300 students in a subgroup, at which point the graduation rate is reported as a whole number percentage. The ranges used for varying sized groups are presented in the following Table.

Number of Students in the Subgroup	Ranges Used for Reporting the Graduation Rate for that Subgroup
6-15	<50%, ≥50%
16-30	≤20%, 21-39%, 40-59%, 60-79% ≥80%
31-60	≤10%, 11-19%, 20-29%, 30-39%, 40-49%, 50-59%, 60-69%, 70-79%, 80-89%, ≥90%
61-300	≤5%, 6-9%, 10-14%, 15-19%, 20-24%, 24-29%, 30-34%, 35-39%, 40-44%, 45-49%, 50-54%, 55-59%, 60-64%, 65-69%, 70-74%, 75-79%, 80-84%, 85-89%, 90-94%, ≥95%
More than 300	≤1%, [whole number percentages] 2%, 3%, . . ., 98%, ≥99%

During data cleanup, I have used a random generator to replace the suppressed and missing values in the adjusted cohort graduation rates dataframe. This was done by generating a random number between the range given in the original dataframe and replace the corresponding value. In the final step of data wrangling I check the following to verify the data integrity:

1. If there are records with more than 100% graduation rate
2. If there are records with less than 0% graduation rate
3. If there are records with a non-integer graduation rate

I also checked to see if there were any record with adjusted cohort graduation rate as 0 and decided to drop these rows from my analysis. To do this, I dropped all the columns in the dataframes and checked to count after dropping the rows with values. After a comparison, I found that 9 & 27 records are dropped from the adjusted cohort graduation rates dataframe of 2011 and 2014 respectively.

3.4 Renaming column titles

The column titles of the adjusted cohort graduation rates dataframes for 2011 and 2014 are shown below.

2011 Columns	2014 Columns	Column Description	New Column Name
STNAM	STNAM	State Name	STNAM
FIPST	FIPST	The two-digit State code	FIPST
LEAID	LEAID	District NCES ID	LEAID
LEANM	LEANM	District Name	LEANM
ALL_COHORT_1112	ALL_COHORT_1415	Total number of students within the four-year adjusted-cohort	ALL_COHORT
ALL_RATE_1112	ALL_RATE_1415	Rate of students who graduated within the four-year adjusted-cohort	ALL_RATE
MAM_COHORT_1112	MAM_COHORT_1415	Total number of American Indian/Alaska Native students within the four-year adjusted-cohort	MAM_COHORT
MAM_RATE_1112	MAM_RATE_1415	Rate of American Indian/Alaska Native students who graduated within the four-year adjusted-cohort	MAM_RATE
MAS_COHORT_1112	MAS_COHORT_1415	Total number of Asian/Pacific Islander students within the four-year adjusted-cohort	MAS_COHORT
MAS_RATE_1112	MAS_RATE_1415	Rate of Asian/Pacific Islander students who graduated within the four-year adjusted-cohort	MAS_RATE
MBL_COHORT_1112	MBL_COHORT_1415	Total number of Black students within the four-year adjusted-cohort	MBL_COHORT
MBL_RATE_1112	MBL_RATE_1415	Rate of Black students who graduated within the four-year adjusted-cohort	MBL_RATE
MHI_COHORT_1112	MHI_COHORT_1415	Total number of Hispanic students within the four-year adjusted-cohort	MHI_COHORT
MHI_RATE_1112	MHI_RATE_1415	Rate of Hispanic students who graduated within the four-year adjusted-cohort	MHI_RATE
MTR_COHORT_1112	MTR_COHORT_1415	Total number of Multiracial students within the four-year adjusted-cohort	MTR_COHORT
MTR_RATE_1112	MTR_RATE_1415	Rate of Multiracial students who graduated within the four-year adjusted-cohort	MTR_RATE
MWH_COHORT_1112	MWH_COHORT_1415	Total number of White students within the four-year adjusted-cohort	MWH_COHORT
MWH_RATE_1112	MWH_RATE_1415	Rate of White students who graduated within the four-year adjusted-cohort	MWH_RATE
CWD_COHORT_1112	CWD_COHORT_1415	Total number of students with disabilities within the four-year adjusted-cohort	CWD_COHORT
CWD_RATE_1112	CWD_RATE_1415	Rate of students with disabilities who graduated within the four-year adjusted-cohort	CWD_RATE
ECD_COHORT_1112	ECD_COHORT_1415	Total number of economically disadvantaged students within the four-year adjusted-cohort	ECD_COHORT
ECD_RATE_1112	ECD_RATE_1415	Rate of economically disadvantaged students who graduated within the four-year adjusted-cohort	ECD_RATE
LEP_COHORT_1112	LEP_COHORT_1415	Total number of students with limited English proficiency within the four-year adjusted-cohort	LEP_COHORT
LEP_RATE_1112	LEP_RATE_1415	Rate of students with limited English proficiency who	LEP_RATE

graduated within the four-year adjusted-cohort

The column titles were very long with years as suffix. To make data access and data operations more manageable, I shortened all the column titles into smaller names by using a dictionary.

An example of shortened column titles for the 2011 adjusted cohort graduation rates data frame is shown below:

YEAR	STATE	LEAID	ALL_COHORT	ALL_RATE	MAM_RATE	MAS_RATE	MBL_RATE	MHI_RATE	MTR_RATE	MWH_RATE	CWD_RATE	ECD_RATE
2011	ALABAMA	0100005	288	83	0	0	74	60	0	88	74	72
2011	ALABAMA	0100006	424	79	18	0	4	64	0	79	71	71

An example of shortened column titles for the 2014 adjusted cohort graduation rate data frame is shown below:

YEAR	STATE	LEAID	ALL_COHORT	ALL_RATE	MAM_RATE	MAS_RATE	MBL_RATE	MHI_RATE	MTR_RATE	MWH_RATE	CWD_RATE	ECD_RATE
2014	ALABAMA	0100005	252	91	0	35	83	86	55	91	33	78
2014	ALABAMA	0100006	369	91	20	0	95	99	0	89	86	85

3.5 Normalizing Cohort data

I realized while doing some initial analysis that the adjusted cohort graduation rates dataframes does not have the % of students from different race which is an important variable in determining why students from some states are doing better. Hence, I calculated the following columns and added to the adjusted cohort graduation rates dataframes.

Column Name	Column Description	Column Rule
MAM_RATIO	% of American Indian/Alaska Native students appeared	MAM_RATIO = (MAM_COHORT/ALL_COHORT) * 100
MAS_RATIO	% of Asian/Pacific Islander students appeared	MAS_RATIO = (MAS_COHORT/ALL_COHORT) * 100
MBL_RATIO	% of Black students appeared	MBL_RATIO = (MBL_COHORT/ALL_COHORT) * 100
MHI_RATIO	% of Hispanic students appeared	MHI_RATIO = (MHI_COHORT/ALL_COHORT) * 100
MTR_RATIO	% of Multiracial students with appeared	MTR_RATIO = (MTR_COHORT/ALL_COHORT) * 100
MWH_RATIO	% of White students appeared	MWH_RATIO = (MWH_COHORT/ALL_COHORT) * 100
CWD_RATIO	% of students appeared with disabilities	CWD_RATIO = (CWD_COHORT/ALL_COHORT) * 100
ECD_RATIO	% of economically disadvantaged students appeared	ECD_RATIO = (ECD_COHORT/ALL_COHORT) * 100
LEP_RATIO	% of students appeared with limited English proficiency	LEP_RATIO = (LEP_COHORT/ALL_COHORT) * 100

3.6 Further exploration

Next, I compared the means of all the critical factors which can impact the adjusted cohort graduation rates in 2011 with 2014. I found that the % of American Indian, Hispanic and Asian students are increased in 2014 and the % of Black and White students is decreased in 2014. The % of students, who are proficient in Math and ELA were the two most significant factors that can contribute towards the adjusted cohort graduation rate. This finding perhaps justifies why the nationwide adjusted cohort graduation rate is moved from 82.25% to 84.26%.

The mean of each normalized critical factor impacting the adjusted cohort graduation rates data for 2011 and 2014:

Critical Factor	2011	2014
% Appeared - American Indian	2.02	2.60
% Appeared - Asian	1.95	2.06
% Appeared - Black	10.00	9.32
% Appeared - Hispanic	11.53	12.63
% Appeared - White	72.99	71.41
% Appeared - Two or More Races	1.21	1.81
% Appeared - Children with disabilities	13.71	13.46
% Appeared - Economically disadvantaged	41.76	45.19
% Appeared - Limited English proficient	2.74	2.86
% Proficient - Math	63.60	52.48
% Proficient - ELA	72.40	61.95
% of Charter School	8.24	8.25
% Public School	101.05	100.91
% of Students on free/reduced lunch	43.99	44.62
% of secondary teachers	45.50	44.86
% of secondary counselors	0.90	0.88

3.7 Comparison of states having most and least Cohort rate

The state wise mean adjusted cohort graduation rate in 2011 and 2014 (i.e., ALL_RATE) were sorted to find the top five state having lowest graduation rate for both these years.

Top five states having lowest graduation rate in 2011 and 2014:

2011	
STATE	% Graduated
ALASKA	60.06
ARIZONA	62.73
LOUISIANA	69.55
OREGON	69.96
DISTRICT OF COLUMBIA	70.00

2014	
STATE	% Graduated
ARIZONA	65.43
DISTRICT OF COLUMBIA	69.43
ALASKA	70.70
OREGON	73.74
NEW MEXICO	75.00

Four of the top five states having lowest graduation rate matched in 2011 and 2014.

Next, I compared the top five states having highest graduation rate for these two years.

Top five states having highest graduation rate in 2011 and 2014:

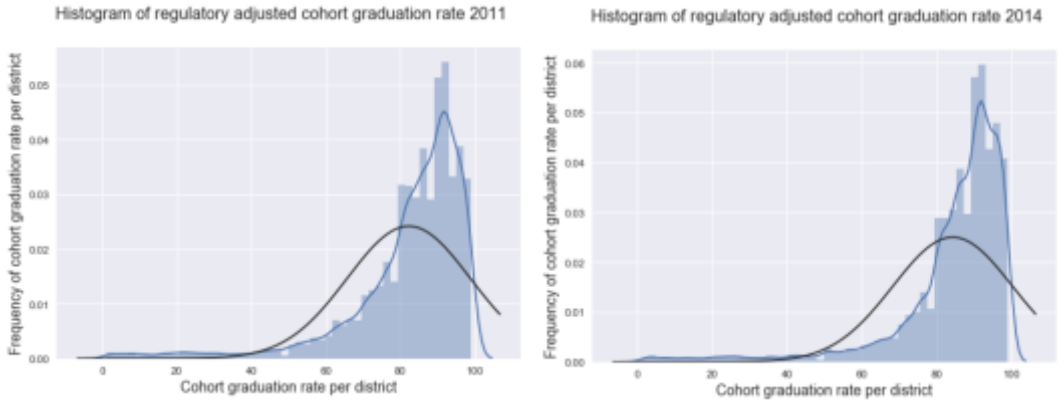
2011	
STATE	% Graduated
IOWA	88.20
NEW JERSEY	88.66
TENNESSEE	89.29
PENNSYLVANIA	89.35
WISCONSIN	90.86

2014	
STATE	% Graduated
ALABAMA	89.74
IOWA	89.85
NEW JERSEY	90.82
WISCONSIN	91.06
KENTUCKY	91.28

Again, I noticed that three of the top five states having highest graduation match in 2011 and 2014.

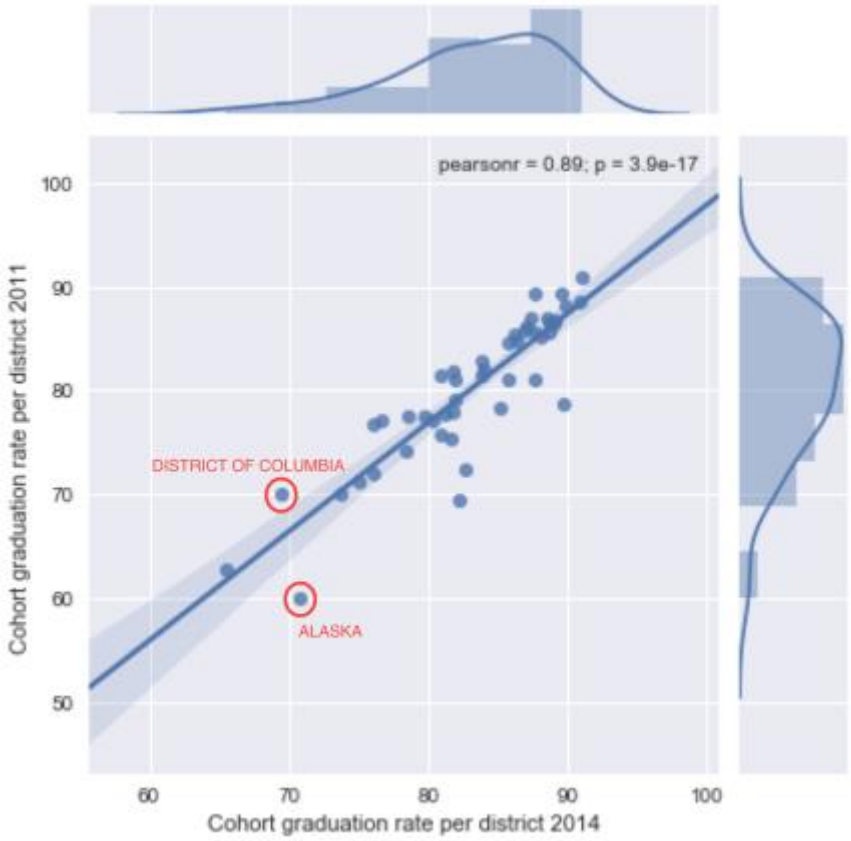
Next, I plotted a few visualizations of the adjusted cohort graduation rates data. I compared graduation rate per district in 2011 and 2014 as univariate distributions using Seaborn by

plotting a histogram and a kernel density estimate. A kernel density estimation allows us to estimate the probability density function of a random variable from a finite set of data. So, it allows us to look at the adjusted cohort graduation data as a continuous probability distribution rather than a histogram.



I realized that an easier way to compare both years would be to show the adjusted cohort graduation rate in a scatterplot. This also showed how correlated both were. In the plot below, 2014 adjusted cohort graduation rates data is on the x-axis and 2011 adjusted cohort graduation rates data is on the y-axis.

As expected, there was a strong correlation of 0.89 in adjusted cohort graduation rates for both years. In addition to outliers in the plot, I also noticed two points, which did not fall along the general trend. These two points refer to the adjusted cohort graduation rate in two states. These are states where the adjusted cohort graduation rate in 2011 is 60.06 and 70.00, and the adjusted cohort graduation rate in 2014 is between 70.70 and 69.43. I found that these two states were Alaska and District of Columbia. We discussed how Alaska, despite having a low adjusted cohort graduation rate, had an increase in adjusted cohort graduation rate in 2014. Likewise, District of Columbia seemed to have a drop in adjusted cohort graduation rate as well.


pearsonr = 0.89; p = 3.9e-17
DISTRICT OF COLUMBIA
ALASKA
Cohort graduation rate per district 2011
Cohort graduation rate per district 2014
100
90
80
70
60
50
60
70
80
90
100

3.8 States with maximum change in Cohort rate

I calculated the percentage of increase or decrease in adjusted cohort graduation rate from 2011 to 2014 by (a) computing the difference in adjusted cohort graduation rate between the two years for each state, (b) dividing this difference with the adjusted cohort graduation rate in 2011 for that state, and (c) converting this value into a percentage.

The five states with the maximum decrease in adjusted cohort graduation rate from 2011 to 2014 were the following.

State	% Change in Graduation Rate from 2011 to 2014
PENNSYLVANIA	-1.86%
COLORADO	-0.89%
NORTH DAKOTA	-0.82%
DISTRICT OF COLUMBIA	-0.82%
MICHIGAN	-0.61%

The five states with the maximum increase in adjusted cohort graduation rate from 2011 to 2014 were the following.

State	% Change in Graduation Rate from 2011 to 2014
NEW HAMPSHIRE	8.89%
ALABAMA	13.87%
GEORGIA	14.19%
ALASKA	17.73%
LOUISIANA	18.20%

Positive values of percentage change in graduation rate from 2011 to 2014 indicate an increase in graduation rate and negative values indicate a decrease in graduation rate, from 2011 to 2014. I wanted to know how these states with the most increase and decrease in adjusted cohort graduation rate compared against some of the top and bottom states based on graduation states for 2014.

I already discussed how Alaska, having a low graduation rate, had a decrease in graduation rate in 2014. This becomes obvious when we look at the percentage change in graduation rate value which shows a 17.73% increase from 2011 to 2014. We also noticed the decrease in graduation rate for District of Columbia in the scatterplot. In support of this finding, the percentage change in graduation rate reveals a 0.82% decrease in graduation rate for the state of District of Columbia from 2011 to 2014.

3.9 Identifying major racial, ethnic groups and socio-economic factors

I looked at the adjusted cohort graduation rates data for 2011 and 2014 and decided to focus on four different racial and ethnic groups – (1) Asian, (2) Hispanic, (3) Black, (4) White. Most of

these categories were already available as columns in the 2011 adjusted cohort graduation rates data frame but the data was not normalized. So, I calculated % of these racial and ethnic groups by dividing the number of students from one racial and ethnic group by the total number of students and multiples that by 100.

I also wanted to consider the assessment data for 2011 and 2014 and decided to use the percentage of students in the district that scored at or above proficient in Math and Reading/Language Arts. Following a same procedure, I did with the adjusted cohort graduation rates data, I read the assessment data for 2011 and 2014 into separate Pandas data frames, selected only the most relevant columns: Total number of students that completed an assessment in Math & Reading/Language Arts, Percentage of students that scored at or above proficient in Math & Reading/Language Arts.

STATE	LEAID	ALL_MTH_PCT	ALL_RLA_PCT
ALABAMA	0100005	84	84
ALABAMA	0100006	90	82
ALABAMA	0100007	97	94
ALABAMA	0100008	95	93
ALABAMA	0100011	74	84