

## **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.

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 state against 3-5 least high-school graduation rate state
3. Identifying the state with the most change in high-school graduation rate and potential reasons for the change
4. Building a predictive model of most high-school graduation rate in each district using machine learning
5. Finally, after addressing #4, 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<sup>1</sup> 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<sup>2</sup> 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

---

<sup>1</sup> <https://www.ed.gov>

<sup>2</sup> <https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html>

students to protect individual student's identity. I realized this would pose a challenge for making comparisons between 2011 and 2014.

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 142 & 128 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
FIPST	FIPST	The two-digit State code	FIPCD
LEAID	LEAID	District NCES ID	LEAID
ALL_COHORT_1112	ALL_COHORT_1415	Total number of students within the four-year adjusted-cohort	ALLA
ALL_RATE_1112	ALL_RATE_1415	Rate of students who graduated within the four-year adjusted-cohort	ALLP
MAM_COHORT_1112	MAM_COHORT_1415	Total number of American Indian/Alaska Native students within the four-year adjusted-cohort	MAMA
MAM_RATE_1112	MAM_RATE_1415	Rate of American Indian/Alaska Native students who graduated within the four-year adjusted-cohort	MAMP
MAS_COHORT_1112	MAS_COHORT_1415	Total number of Asian/Pacific Islander students within the four-year adjusted-cohort	MASA
MAS_RATE_1112	MAS_RATE_1415	Rate of Asian/Pacific Islander students who graduated within the four-year adjusted-cohort	MASP
MBL_COHORT_1112	MBL_COHORT_1415	Total number of Black students within the four-year adjusted-cohort	MBLA
MBL_RATE_1112	MBL_RATE_1415	Rate of Black students who graduated within the four-year adjusted-cohort	MBLP
MHI_COHORT_1112	MHI_COHORT_1415	Total number of Hispanic students within the four-year adjusted-cohort	MHIA
MHI_RATE_1112	MHI_RATE_1415	Rate of Hispanic students who graduated within the four-year adjusted-cohort	MHIP
MTR_COHORT_1112	MTR_COHORT_1415	Total number of Multiracial students within the four-year adjusted-cohort	MTRA
MTR_RATE_1112	MTR_RATE_1415	Rate of Multiracial students who graduated within the four-year adjusted-cohort	MTRP
MWH_COHORT_1112	MWH_COHORT_1415	Total number of White students within the four-year adjusted-cohort	MWHA
MWH_RATE_1112	MWH_RATE_1415	Rate of White students who graduated within the four-year adjusted-cohort	MWHP
CWD_COHORT_1112	CWD_COHORT_1415	Total number of students with disabilities within the four-year adjusted-cohort	CWDA
CWD_RATE_1112	CWD_RATE_1415	Rate of students with disabilities who graduated within the four-year adjusted-cohort	CWDP
ECD_COHORT_1112	ECD_COHORT_1415	Total number of economically disadvantaged students within the four-year adjusted-cohort	ECDA
ECD_RATE_1112	ECD_RATE_1415	Rate of economically disadvantaged students who graduated within the four-year adjusted-cohort	ECDP
LEP_COHORT_1112	LEP_COHORT_1415	Total number of students with limited English proficiency within the four-year adjusted-cohort	LEPA
LEP_RATE_1112	LEP_RATE_1415	Rate of students with limited English proficiency who graduated within the four-year adjusted-cohort	LEPP

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	FIPCD	LEAID	ALLA	ALLP	ALLG	MAMA	MAMP	MAMG	MASA	MASP	MASG	MBLA	MBLP	MBLG
0	2011	01	0100005	268	83	222	0	0	0	0	0	0	6	97	6
1	2011	01	0100006	424	79	335	2	0	0	1	0	0	4	0	0

### 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
MAMR	% of American Indian/Alaska Native students appeared	$MAMR = (MAMA/ALLA) * 100$
MASR	% of Asian/Pacific Islander students appeared	$MASR = (MASA/ALLA) * 100$
MBLR	% of Black students appeared	$MBLR = (MBLA/ALLA) * 100$
MHIR	% of Hispanic students appeared	$MHIR = (MHIA/ALLA) * 100$
MTRR	% of Multiracial students with appeared	$MTRR = (MTRA/ALLA) * 100$
MWHR	% of White students appeared	$MWHR = (MWHA/ALLA) * 100$
CWDR	% of students appeared with disabilities	$CWDR = (CWDA/ALLA) * 100$
ECDR	% of economically disadvantaged students appeared	$ECDR = (ECDA/ALLA) * 100$
LEPR	% of students appeared with limited English proficiency	$LEPR = (LEPA/ALLA) * 100$

### 3.6 Assessment data exploration

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

### 3.7 Poverty data exploration

Now I want to parse the small area income and poverty estimates for all district for 2011 and 2014. Following a same procedure, I read the poverty data for 2011 and 2014 into separate Pandas data frames, selected only the most relevant columns: Estimated Total Population(TOTP), Estimated Population between 5-17(CHLD) and Estimated number of relevant children 5 to 17 years old in poverty(CHIP).

I realized while doing some initial analysis that the poverty data has various number of relevant children 5 to 17 years old in poverty. Hence, I calculated the following columns and added to the poverty dataframes.

Column Name	Column Description	Column Rule
CHIR	% of Children per district	$CHIR = (CHLD / TOTP) * 100$
POVR	% of Children under poverty per district	$MASR = (CHIP / CHLD) * 100$



### 3.8 Elementary/Secondary data exploration

Now I want to parse the elementary/secondary data for all district for 2011 and 2014. Following a same procedure, I read the poverty data for 2011 and 2014 into separate Pandas data frames, selected only the most relevant columns. After renaming the column titles of the elementary/secondary dataframes for 2011 and 2014 are shown below.

2011 Columns	2014 Columns	Column Description	New Column Name
FIPST	FIPST	The two-digit State code	STATE
LEAID	LEAID	ANSI/FIPS State Code	LEAID
Total Schools	Total Schools	Total number of school per district	TOTSC
Total Charter Schools	Total Charter Schools	Total number of charter school per district	TOTCH
Total Public Schools	Total Public Schools	Total number of public charter school per district	TOTPS
Total Students	Total Students	Total number of Total Students per district	TOTS
Free and Reduced Lunch Students	Free and Reduced Lunch Students	Total of Free Lunch Eligible students and Reduced-price Lunch Eligible students	FRLS
Grade 12 Students - male	Grade 12 Students - male	Total number of male students in 12th grade	GR12M
Grade 12 Students - female	Grade 12 Students - female	Total number of female students in 12th grade	GR12F
Pupil/Teacher Ratio	Pupil/Teacher Ratio	Calculated Pupil Teacher Ratio	PTR
Full-Time Equivalent	Full-Time Equivalent	Total number of Full Time Equivalent teachers per district	FTE
Secondary Teachers	Secondary Teachers	Total number of secondary teachers per district	SET
Total Staff	Total Staff	Total number of staff per district	TOTSF
Secondary Guidance Counselors	Secondary Guidance Counselors	Total number of secondary guidance counselors per district	SEGC

I realized while doing some initial analysis that the elementary/secondary data is not normalized. Hence, I added to the elementary/secondary dataframes the following columns after normalization:

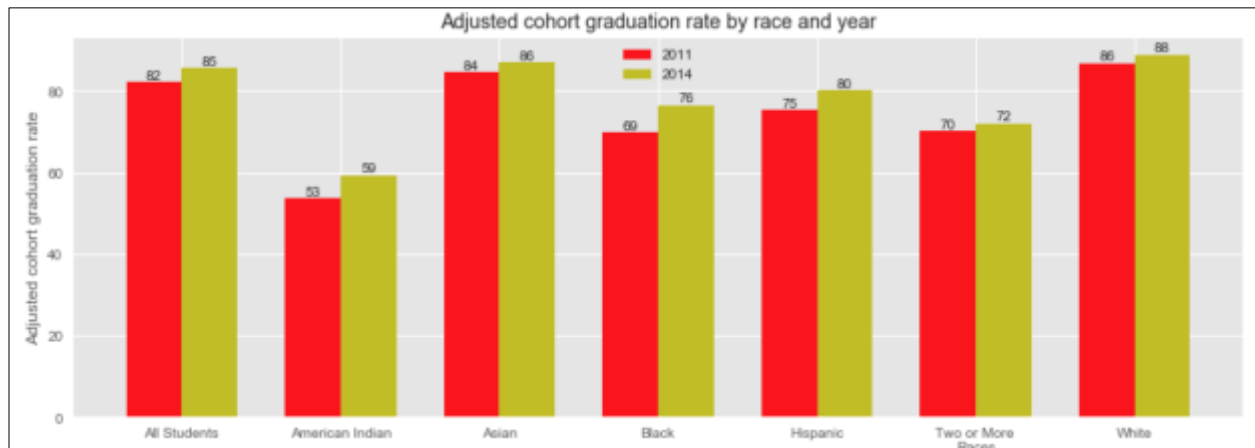
Column Name	Column Description	Column Rule
CHSR	% of Charter School	$MAMR = (MAMA/TOTSC) * 100$
MLSR	% of male students	$MLSR = (GR12M/TOTS) * 100$
FLSR	% of female students	$FLSR = (GR12F/TOTS) * 100$
FRLR	% of Students on free/reduced lunch	$FRLR = (FRLS/TOTS) * 100$
SETR	% of secondary teachers	$SETR = (SET/FTE) * 100$
SECR	% of secondary counselors	$SECR = (SEGC/TOTSF) * 100$

### 3.9 Exploration of Combined Dataset

After normalizing the cohort dataset, assessment dataset, poverty dataset and elementary secondary dataset, these datasets are merged and generated a combined dataframe. Now we will try to analyze the statement made by Department of Education that "In school year 2014-15, the adjusted cohort graduation rate (ACGR) for public high school students rose to 83%."

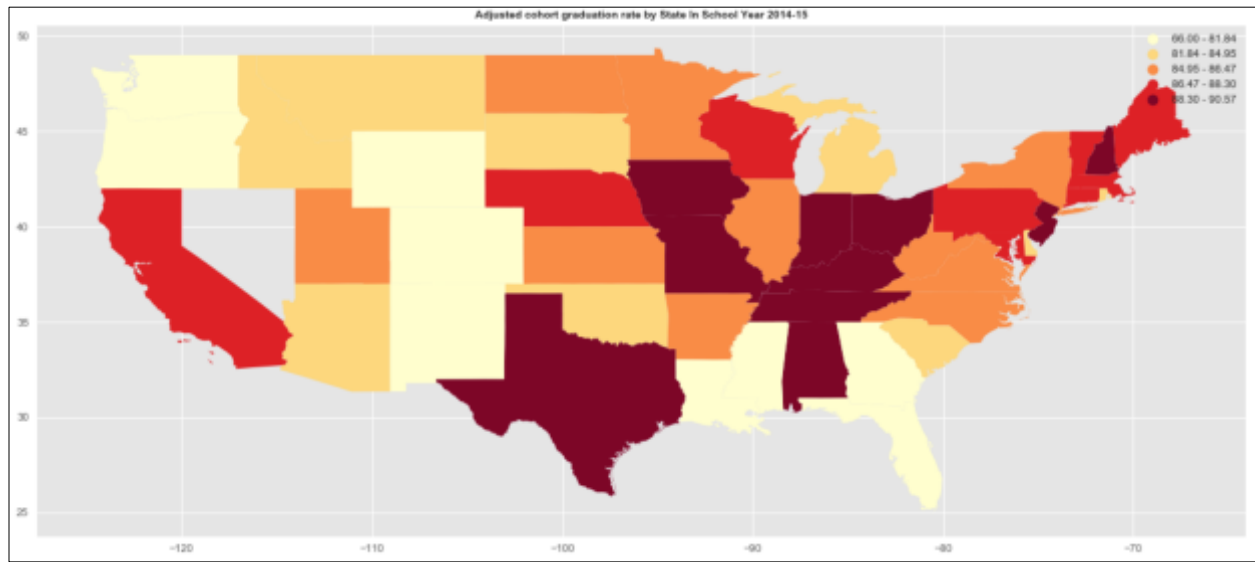
After verifying the clean regulatory adjusted cohort graduation rate for all district, we can see that in school year 2014-15 the graduation rate at the national level is 85%, an increase from 82% in 2011-12.

After calculating the cohort graduation rate for major racial and ethnic groups, we can see that in school year 2014-15 white students have the highest ACGR of 88.0%, followed by Asian/Pacific Islander 86.0%, Hispanic 80.0%, Black 76.0%, Two or more races 72% and American Indian/Alaska Native 59.0%.



From the Adjusted cohort graduation rate by race and year we can conclude that the increase in the overall cohort graduation rate is contributed by each race but they may not have contributed at the same rate. Later we will analyze which race has more impact on the overall graduation rate.

Now check the states which are doing better than other state. In order to, have a better visual representation we have plotted the choropleth map with the data from school year 2014-15 after removing Hawaii and Alaska.



From the above heat map, it is very clear that the school district in Northeast and South region is doing better than the school district in Midwest and west region. Now check how these states are doing over time.

### 3.10 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	% of change
% Appeared - American Indian	1.90	2.60	36.84
% Appeared - Asian	1.89	1.98	4.76
% Appeared - Black	8.41	7.81	-7.13
% Appeared - Hispanic	10.66	11.72	9.94
% Appeared - White	75.72	73.95	-2.34
% Appeared - Two or More Races	1.20	1.82	51.67
% Appeared - Children with disabilities	12.98	12.59	-3.00
% Appeared - Economically disadvantaged	40.44	43.75	8.18
% Appeared - Limited English proficient	2.68	2.76	2.99
% Proficient - Math	64.19	51.52	-19.74
% Proficient - ELA	73.76	61.71	-16.34

% of Children Population	16.91	16.68	-1.36
% of Children under poverty	18.97	18.76	-1.11
% of Charter School	1.08	1.02	-5.56
Pupil/Teacher Ratio	14.48	14.55	0.48
% of male students in G12	4.12	4.04	-1.94
% of female students in G12	3.93	3.85	-2.04
% of Students on free/reduced lunch	43.12	44.16	2.41
% of secondary teachers	44.74	44.36	-0.85
% of secondary counselors	0.89	0.88	-1.12

### 3.11 Comparison of states having most and least Cohort rate

The state wise mean adjusted cohort graduation rate in 2011 and 2014 (i.e., ALLP) 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
<b>DISTRICT OF COLUMBIA</b>	54.01
NEVADA	64.38
<b>OREGON</b>	69.01
<b>ALASKA</b>	69.08
GEORGIA	70.60

2014	
STATE	% Graduated
<b>DISTRICT OF COLUMBIA</b>	66.00
NEW MEXICO	69.81
<b>OREGON</b>	74.62
<b>ALASKA</b>	75.33
FLORIDA	77.88

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.

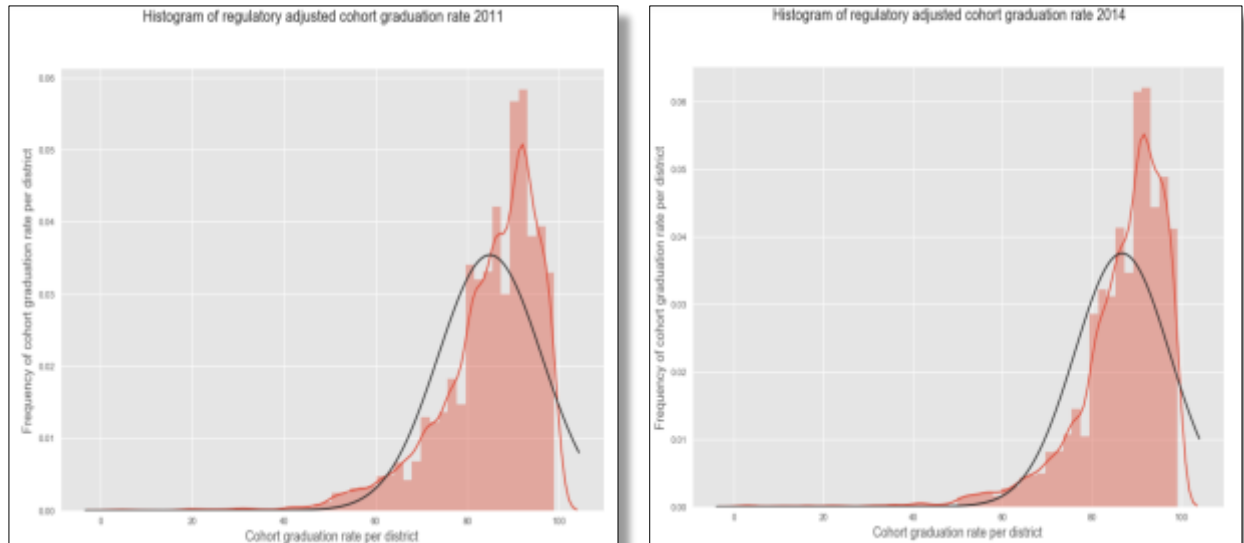
Top five states having highest graduation rate in 2011 and 2014:

2011	
STATE	% Graduated
WISCONSIN	87.29
VERMONT	87.78
<b>OHIO</b>	88.13
IOWA	88.67
<b>TEXAS</b>	88.98

2014	
STATE	% Graduated
MISSOURI	89.76
INDIANA	89.87
<b>IOWA</b>	89.91
<b>TEXAS</b>	90.25
TENNESSEE	90.57

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

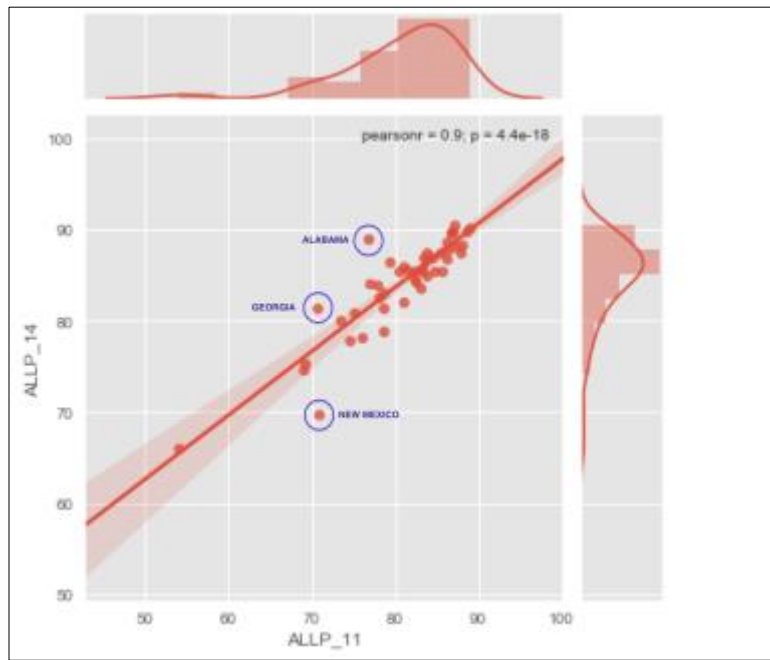
Next, I plotted a few visualizations of the adjusted cohort graduation rates data. I compared graduation rate per district in 2011 and 2014 as univariate distributions using Seaborn by plotting a histogram and a kernel density estimate. A kernel density estimation allows us to estimate the probability 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.

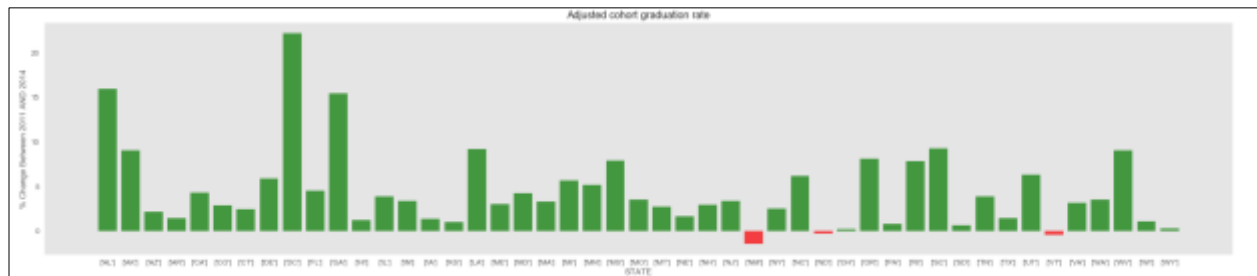
I realized that an easier way to compare both years would be to show the adjusted cohort graduation rate in a scatterplot. This also showed how correlated both were. In the plot below, 2011 adjusted cohort graduation rates data is on the x-axis and 2014 adjusted cohort graduation rates data is on the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

As expected, there was a strong correlation of 0.9 in adjusted cohort graduation rates for both years. In addition to outliers in the plot, I also noticed three points, which did not fall along the general trend. These two points refer to the adjusted cohort graduation rate in three states. I found that these three states were New Mexico, Georgia and Alabama. Among these three states New Mexico is the only state which is showing negative growth in adjusted cohort graduation rate.



### 3.12 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.



From the above histogram, it is clear that most states shown positive growth in cohort graduation rate.

But NEW MEXICO, NORTH DAKOTA, VERMONT are the states showing negative growth in cohort graduation rate.

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.

## 4 Prediction Using Machine Learning

### 4.1 4.1 Feature selection

Prior to performing model fitting, I wanted to take a look at the limited set of features I had and filter out any unnecessary features, keeping the regression problem in mind.

I merged the cohort dataset, assessment dataset, poverty dataset and elementary secondary dataset, for each year separately. This resulted in obtaining two data frames with 21 columns – 20 features/independent variables, and 1 dependent variable. Then I computed cross-correlations of all 20. This allowed me to look for any correlations among variables and remove variables that were confounding each other. Two strongly correlated variables can decrease the robustness of a machine learning model as one variable can potentially inhibit the effect of another.

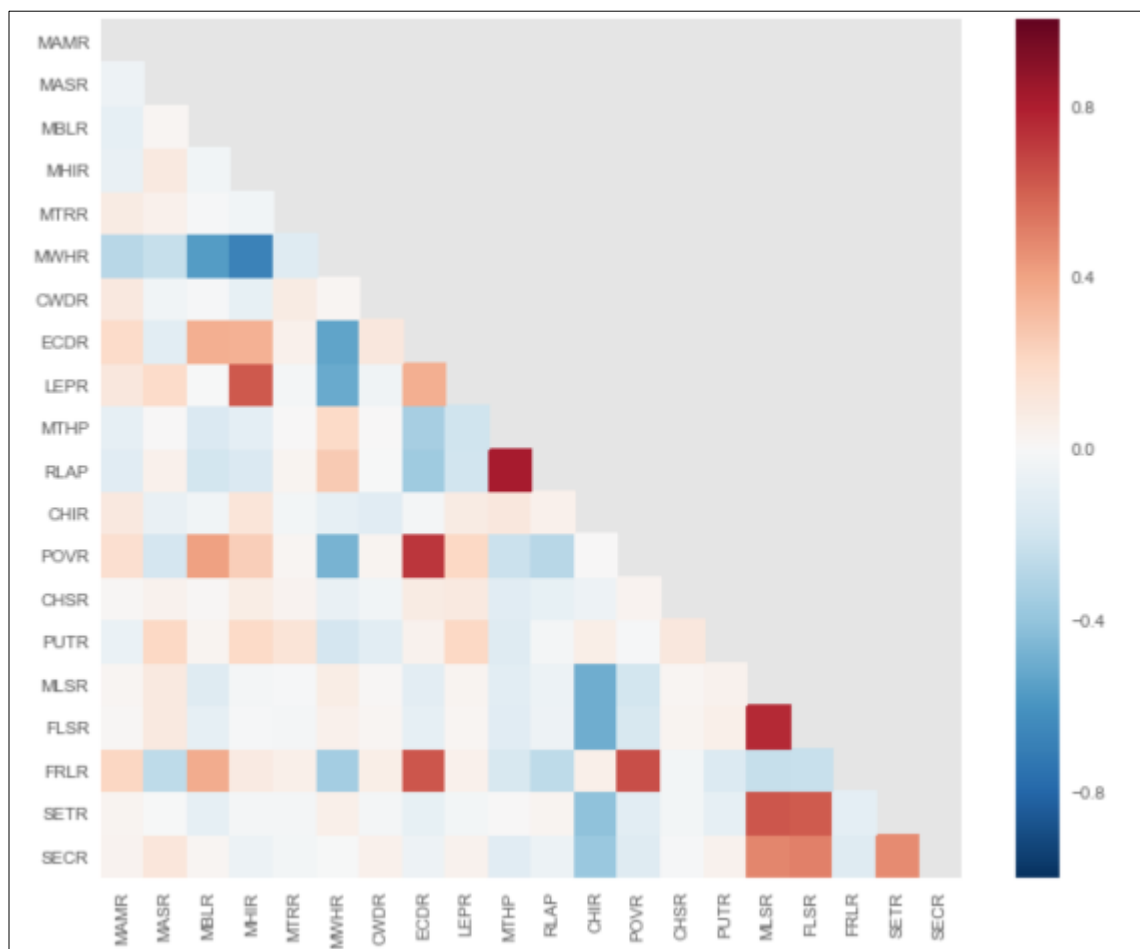
Full Name	2011 Column	2014 Column
% Appeared - American Indian	MAMR	MAMR
% Appeared - Asian	MASR	MASR
% Appeared - Black	MBLR	MBLR
% Appeared - Hispanic	MHIR	MHIR
% Appeared - White	MWHR	MWHR
% Appeared - Two or More Races	MTRR	MTRR
% Appeared - Children with disabilities	CWDR	CWDR
% Appeared - Economically disadvantaged	ECDR	ECDR
% Appeared - Limited English proficient	LEPR	LEPR
% Proficient - Math	MTHP	MTHP
% Proficient - ELA	RLAP	RLAP
% of Children Population	CHIR	CHIR
% of Children under poverty	POVR	POVR
% of Charter School	CHSR	CHSR
Pupil/Teacher Ratio	PUTR	PUTR
% of male students in G12	MLSR	MLSR
% of female students in G12	FLSR	FLSR
% of Students on free/reduced lunch	FRLR	FRLR
% of secondary teachers	SETR	SETR
% of secondary counselors	SECR	SECR

Cross-correlation table for 2014 is provided below.

	MAMR	MASR	MBLR	MHIR	MTRR	MWHR	CWDR	ECDR	LEPR	MTHP	RLAP	CHIR	POVR	CHSR	PUTR	MLSR	FLSR	FRLR	SETR	SECR
MAMR	1	-0.051	-0.086	-0.07	0.08	-0.29	0.11	0.19	0.11	-0.091	-0.12	0.1	0.160	0.078	-0.068	0.018	0.01	0.22	0.03	0.032
MASR	-0.051	1	0.021	0.097	0.05	-0.24	-0.039	-0.11	0.20	0.0093	0.047	-0.075	-0.18	0.045	0.21	0.099	0.1	-0.26	-0.0014	0.12
MBLR	-0.086	0.021	1	-0.038	-0.01	-0.57	-0.01	0.36	-0.0036	-0.15	-0.19	-0.038	0.410	0.092	0.028	-0.13	-0.093	0.37	-0.09	0.022
MHIR	-0.07	0.097	-0.038	1	-0.031	-0.67	-0.08	0.35	0.62	-0.1	-0.16	0.13	0.25	0.076	0.2	-0.016	-0.014	0.088	-0.021	-0.061
MTRR	0.08	0.05	-0.01	-0.031	1	-0.13	0.083	0.052	-0.023	0.002	0.025	-0.031	0.017	0.037	0.14	-0.011	-0.021	0.061	-0.021	-0.027
MWHR	-0.29	-0.24	-0.57	-0.67	-0.13	1	0.02	-0.54	-0.51	0.2	0.26	-0.089	-0.47	-0.074	-0.18	0.072	0.05	-0.35	0.062	-0.00087
CWDR	0.11	-0.039	-0.01	-0.08	0.083	0.02	1	0.12	-0.04	0.0036	-0.0056	-0.12	0.028	-0.033	-0.12	0.012	0.019	0.066	-0.018	0.048
ECDR	0.19	-0.11	0.36	0.35	0.052	-0.54	0.12	1	0.36	-0.34	-0.36	-0.016	0.72	0.086	0.044	-0.1	-0.092	0.63	-0.081	-0.047
LEPR	0.11	0.2	-0.0036	0.62	-0.023	-0.51	-0.04	0.36	1	-0.21	-0.2	0.081	0.21	0.098	0.21	0.024	0.019	0.05	-0.028	0.04
MTHP	-0.091	0.0093	-0.15	-0.1	0.002	0.2	0.0036	-0.34	-0.21	1	0.81	0.11	-0.22	-0.12	-0.13	-0.12	-0.12	-0.17	0.0072	-0.12
RLAP	-0.12	0.047	-0.19	-0.16	0.025	0.26	-0.0056	-0.36	-0.2	0.81	1	0.054	-0.28	-0.081	-0.02	-0.058	-0.053	-0.26	0.026	-0.055
CHIR	0.1	-0.075	-0.038	0.13	-0.031	-0.089	-0.12	-0.016	0.061	0.11	0.054	1	0.00033	-0.048	0.069	-0.5	-0.5	0.061	-0.41	-0.38
POVR	0.16	-0.18	0.41	0.25	0.017	-0.47	0.028	0.72	0.21	-0.22	-0.28	0.00033	1	0.039	-0.0084	-0.19	-0.17	0.65	-0.11	-0.13
CHSR	0.0078	0.045	0.0092	0.076	0.037	-0.074	-0.033	0.086	0.098	-0.12	-0.061	-0.048	0.039	1	0.11	0.02	0.03	-0.028	-0.024	-0.015
PUTR	-0.068	0.21	0.028	0.2	0.14	-0.18	-0.12	0.044	0.21	-0.13	-0.02	0.069	-0.0084	0.11	1	0.041	0.062	-0.15	-0.087	0.044
MLSR	0.018	0.099	-0.13	-0.016	-0.011	0.072	0.012	-0.1	0.024	-0.12	-0.058	-0.5	-0.19	0.02	0.041	1	0.76	-0.24	0.63	0.49
FLSR	0.01	0.1	-0.093	-0.014	-0.021	0.05	0.019	-0.092	0.019	-0.12	-0.053	-0.5	-0.17	0.03	0.062	0.76	1	-0.23	0.61	0.51
FRLR	0.22	-0.26	0.37	0.088	0.061	-0.35	0.066	0.63	0.05	-0.17	-0.26	0.061	0.65	-0.028	-0.15	-0.24	-0.23	1	-0.095	-0.13
SETR	-0.03	-0.0014	-0.09	-0.021	-0.021	0.082	-0.018	-0.081	-0.028	0.0072	0.026	-0.41	-0.11	-0.024	-0.087	0.63	0.61	-0.095	1	0.47
SECR	0.032	0.12	0.022	-0.061	-0.027	-0.00087	0.048	-0.047	0.04	-0.12	-0.055	-0.38	-0.13	-0.015	0.044	0.49	0.51	-0.13	0.47	1

I plotted these correlations to make them more understandable.

Correlation plot for 2014:





## 4.2 Model selection and evaluation