**Analysis of Root Factors of High School Graduation Rate**

**Introduction/Motivation:**

In 2004, there are only 71.7% students were graduated from high school successfully. In 2013, it increased to 81.4% with 1.8 million more people graduated successfully.

The government set up a goal that 90% of students will graduate from high school successfully in 2020.

To achieve this goal, it is necessary to find out root causes lead to high school graduation rate and predict future graduation rate for new school centers based on each factors.

**Purpose:**

- To research for root causes affect high-school graduation rate and predict consequences dropout rate base on each factors.

- Use result to give reasonable recommendations to help government achieve higher graduation rate.

- Data: http://datafordiplomas.devpost.com/details/resources 2010 census data merged 2013 graduation data

- 2317 records with missing value - 565 variables including census/graduation factors

- Detailed graduation data with graduation related variables Mapping logic of censes tract overlap school districts

- Average temperature data on state level

- Average winter temperature data on state level

- School system quality/safety data

**Pre-processing Dataset:**

1. Deleted missing value and error data as much as possible
2. Only select calculated percentage variables, which is about 200 variables
3. Clean data, eliminate missing values in python
4. Went along data dictionary and choose most typical features and analyze feature importance
5. Use Tableau and Python to analyze correlation between important factors and target factor

**Problem Statement and Hypothesis:**

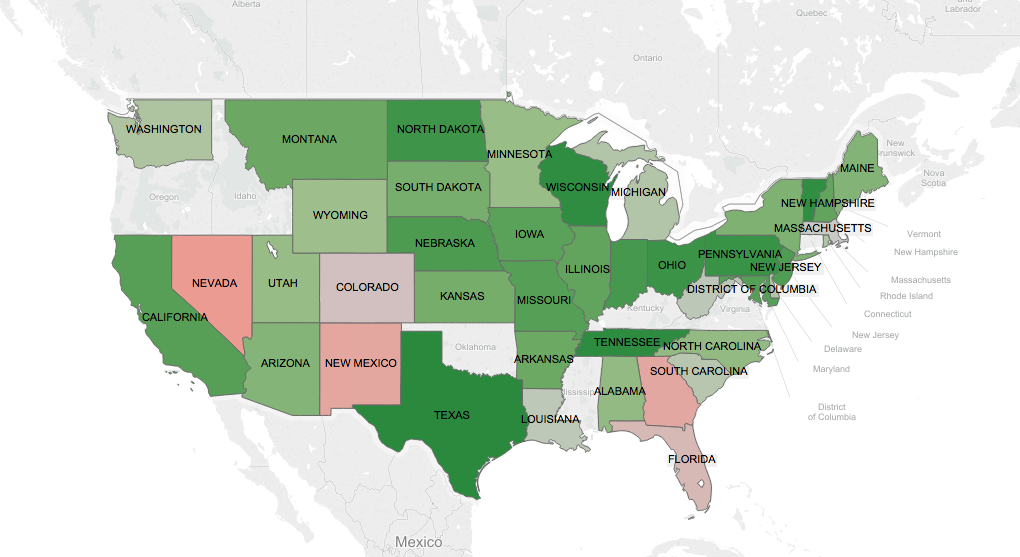
There are lots of factors can affect graduation rate, finding the root causes can help government set up new rules to help schools increase graduation rate by 2020.

Some of the hypotheses are:

* Weather and location of states can affect graduation rate.
* Families with high movement rate and poverty level have lower graduation rate
* Students with employed are having less time study, lead to lower graduation rate
* Temperature and weather could affect students’ study and graduation rate

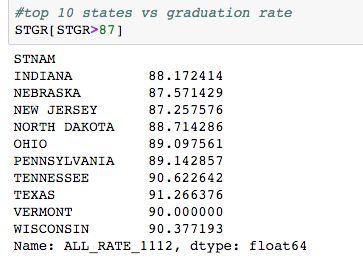
**Analysis:**

1. Graduation rate on State level



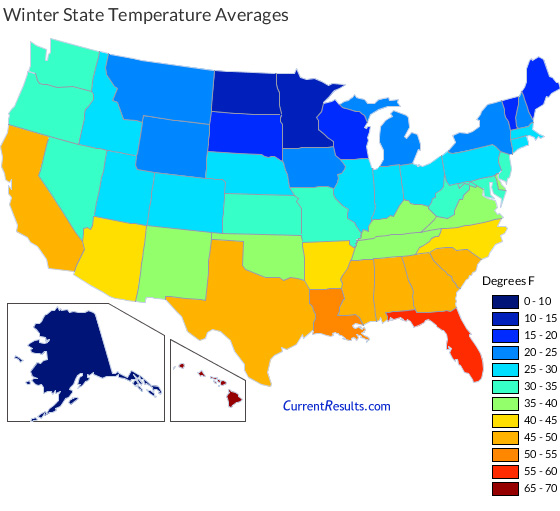
This map shows the graduation rate for each state. All the data are above 50%, so I set up red to be close to 50% and green as to be close to 100%. On the map, we can see, northern part of the country has relatively higher graduation rate than southern part. And even mid-land states has relatively higher graduation rate than others.

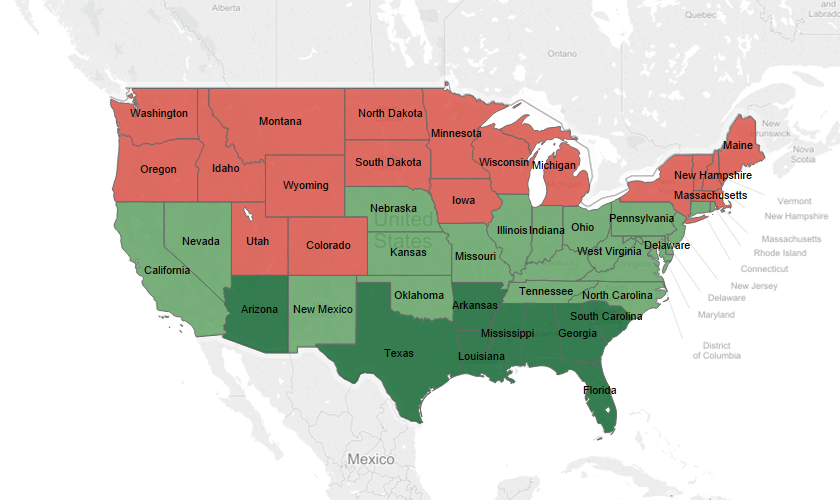
I also grouped up graduation rate by states in python to see top 10 states of highest graduation rate. As showing below:



2. Average temperature vs. Graduation rate

I looked up average annual temperature and average winter temperature by states, trying to find out the impact of temperature on graduation rate. Temperature maps shows below:

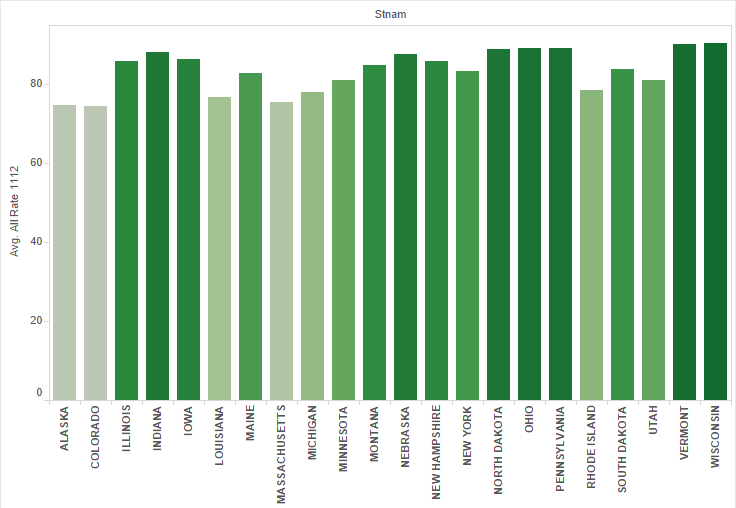




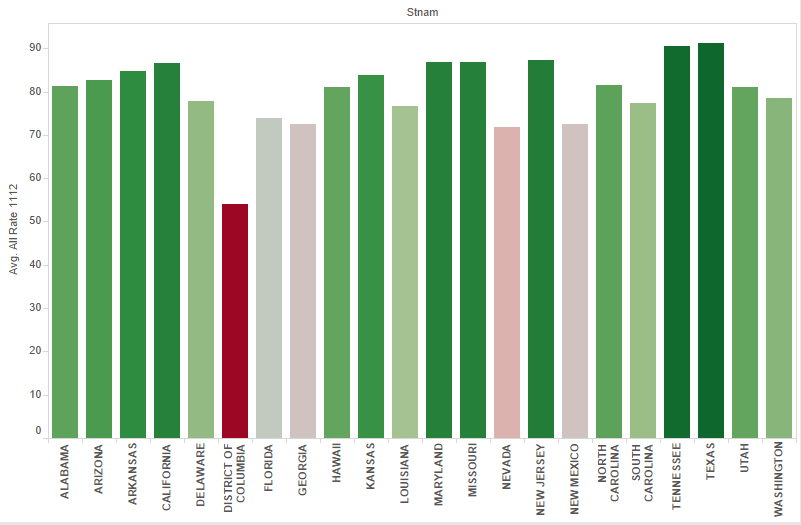
I imported average temperature data into Tableau to see the weather in different state. Red means cold and dark green means warm.

I separate states by winter temperature. I defined states with temperature above 35 degree as south state, others as north. Graduation rate for different parts shows below:

North



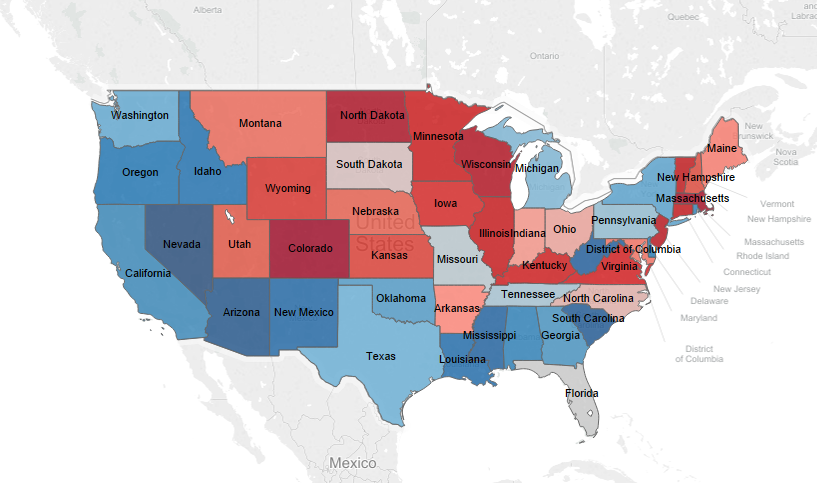
South:



From these two charts, we can see that southern states have relatively lower graduation rate than northern states. It might because, southern state consist of so many tourism cities, students may put so much time on making money and pay little effort on their study. To solving this problem, government should make strict rules about under-age laboring. And set up higher salary rate for people with high-school diploma to encourage high school finishing rate.

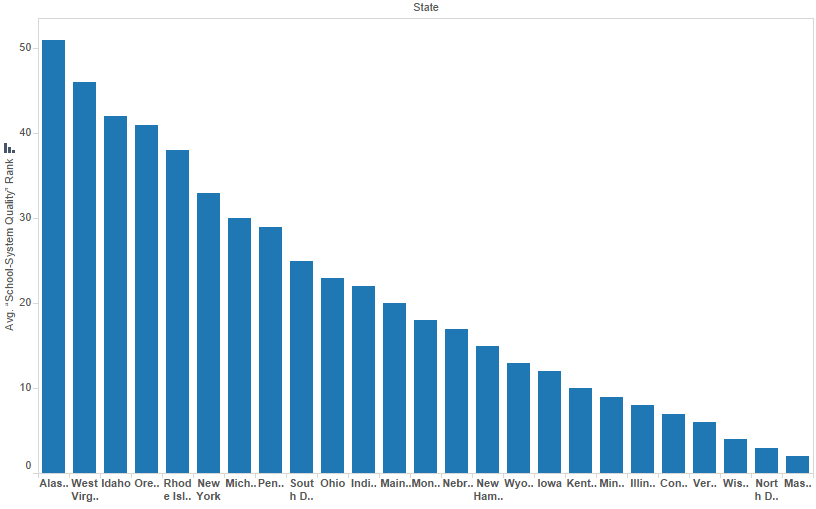
3. School-System Quality ranking on State Level

School-system quality ranking: red-low(good), blue-high(bad)

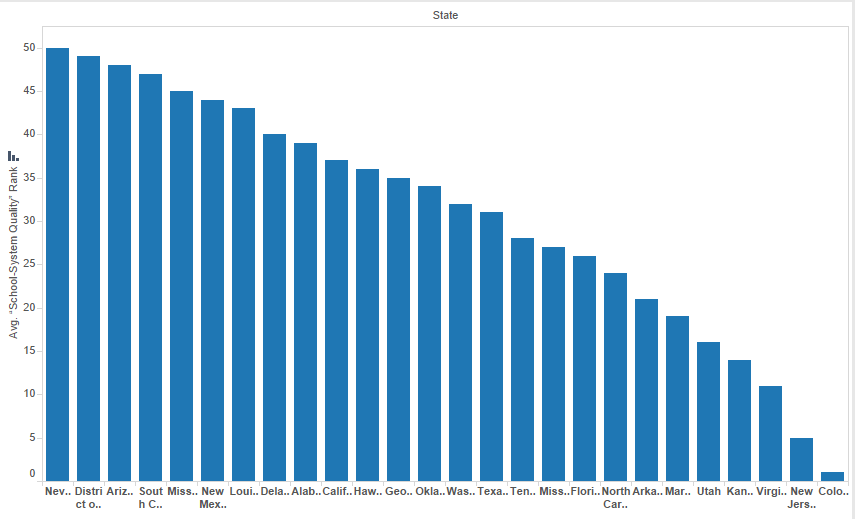


These two graphs show that middle and northern states have better school system and eastern and southern states have relatively higher safety scores. To figure out if these two variables are contributed to graduation rates, I split these states to north and south using the same split methods as temperatures.

North School system Quality ranking:



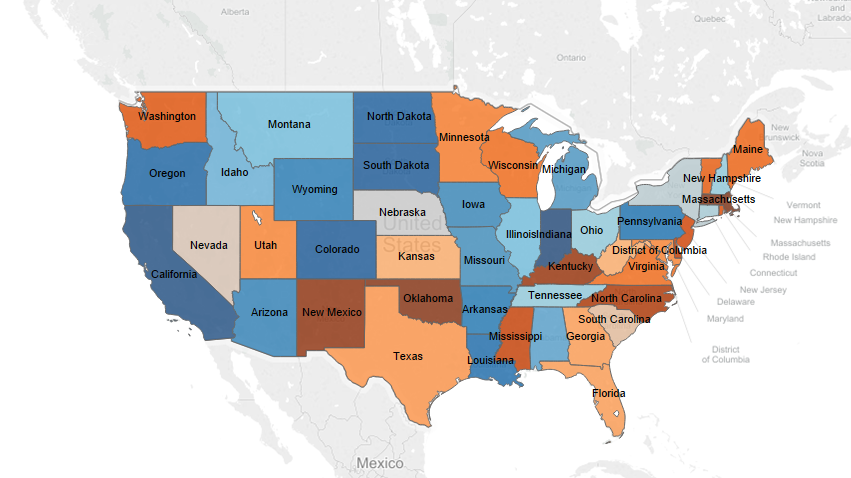
South school system quality ranking



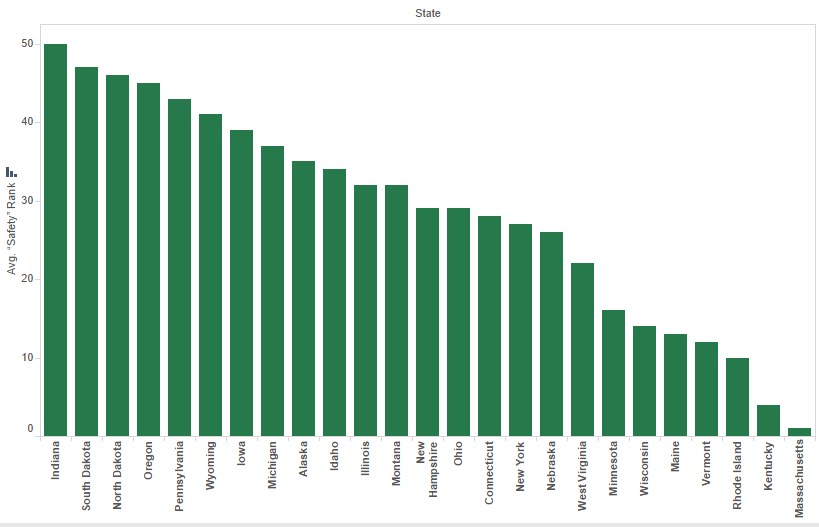
Northern states rank overall higher than southern states. Due to my conclusion above, northern states have higher graduation rate than southern states. School-system quality plays a role on effect the graduation rate. Government should improve school-system quality in southern states to increase graduation rate.

4. School Safety ranking vs. graduation rate

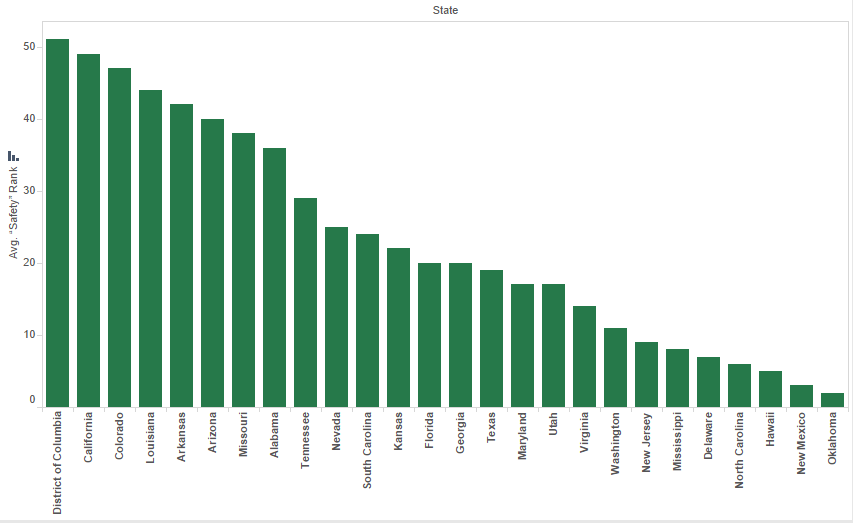
School safety ranking: orange-low (good), blue-high (bad)



North School Safety:



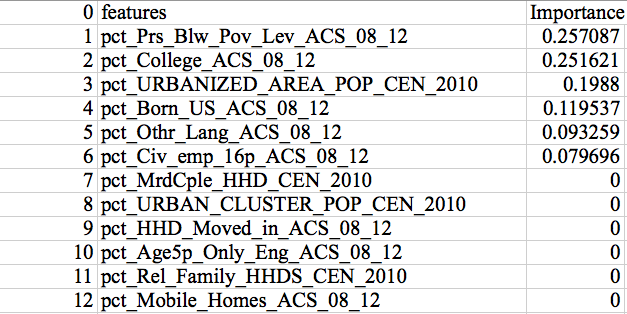
South school safety:

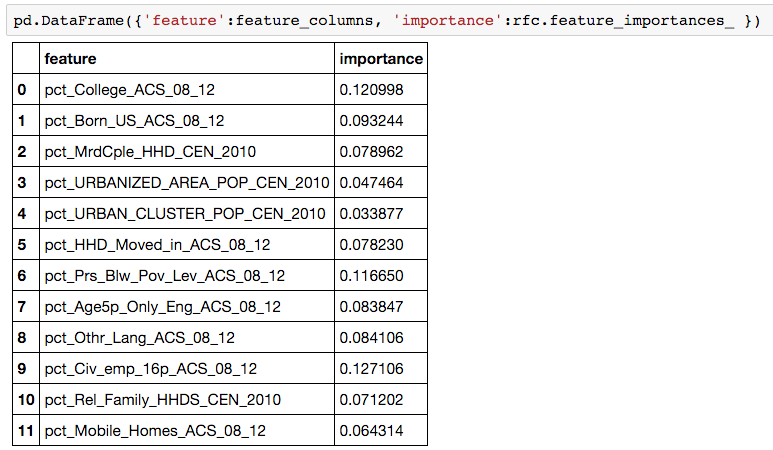


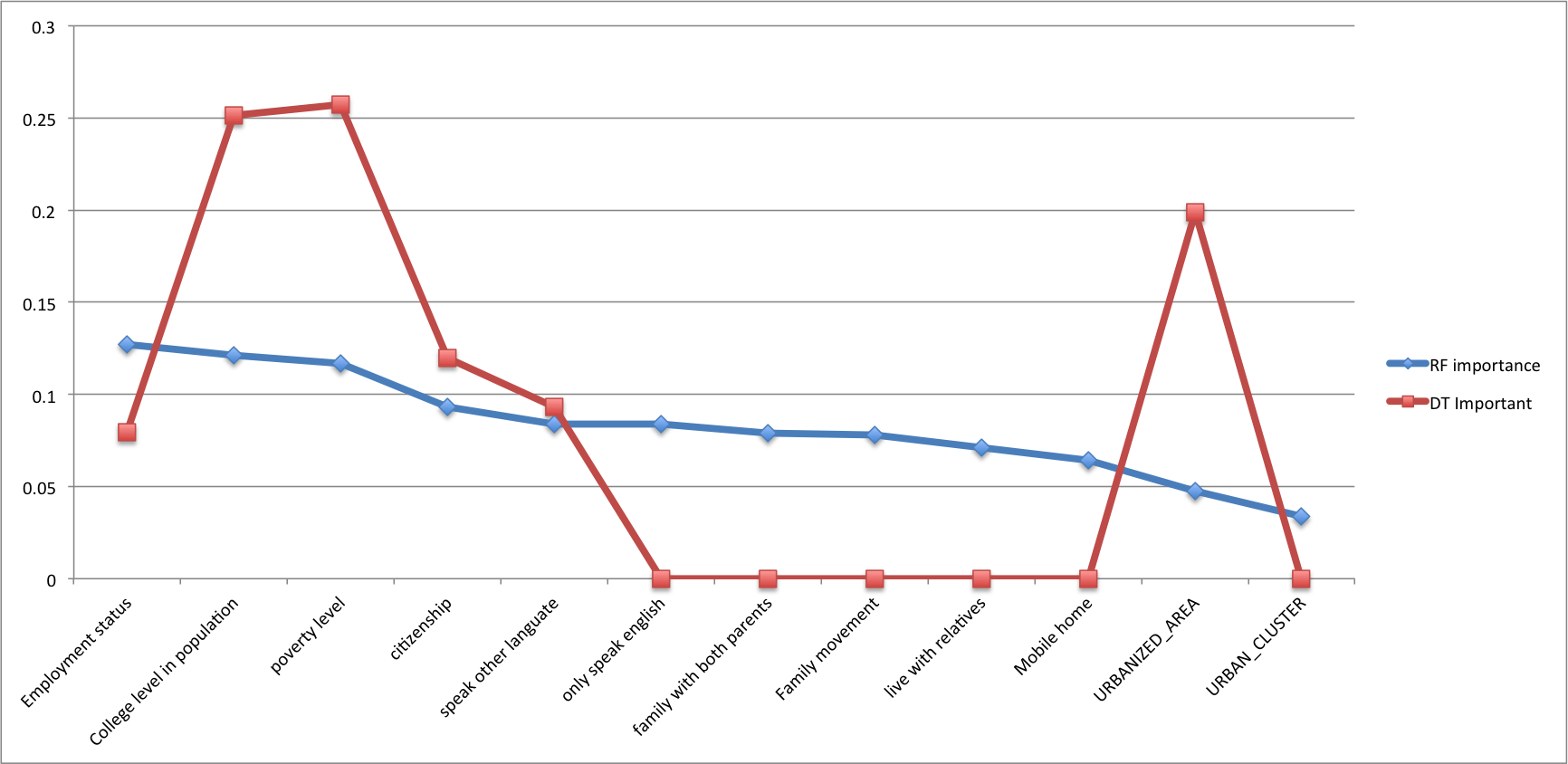
Southern states have higher school safety than northern schools. Because northern states have higher graduation rate than southern states, which means school safety does not contribute much on graduation rate.

5. Feature Importance:

I picked up 12 features that might contribute to graduation rate, and ran the feature importance analysis on both decision tree and random forest models. Results shows below:

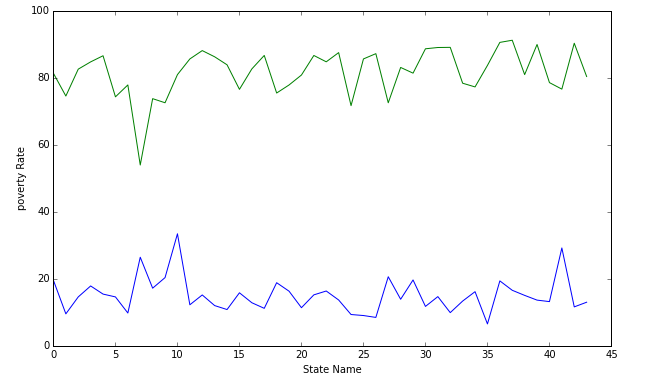




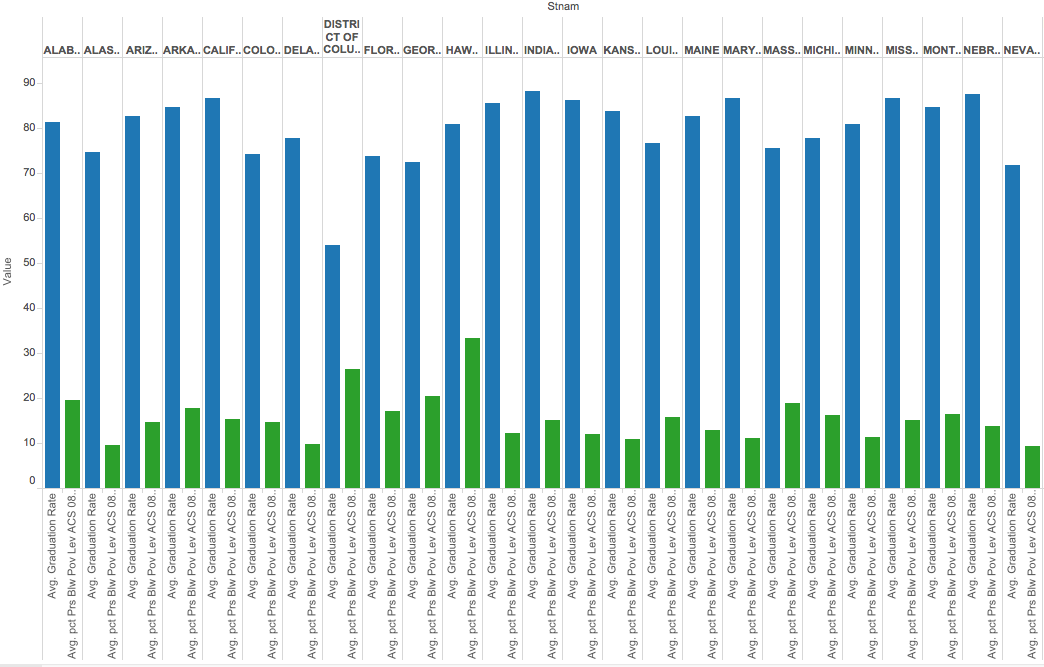


I used both decision tree and random forest models to analyze feature importance, and transfer the result into a graph above. It shows that Employment status, college level among population, poverty level, citizenship and language speaking are most important features related to gradation rate. Which means we should focus analysis on these features and figure out solutions.

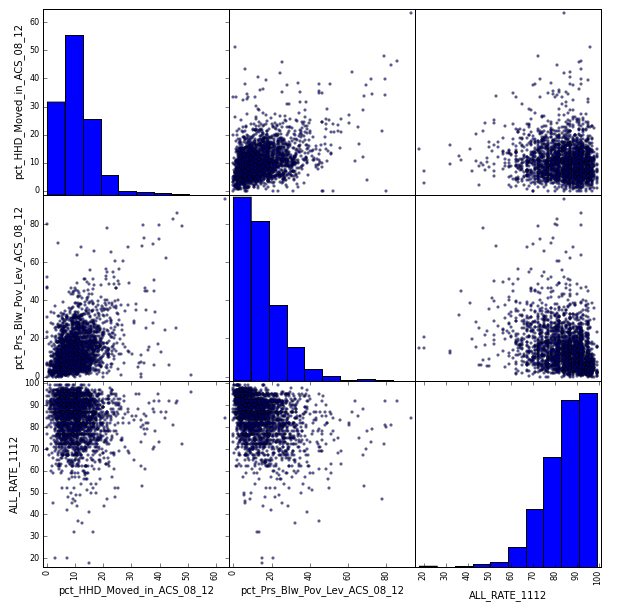
6. Poverty vs. Graduation rate



The top line shows the graduation rate and bottom line shows poverty rate. We can see for some obviously high poverty state, the graduation rate is relatively lower as well. Just as #7 of the state, which is District of Columbia.

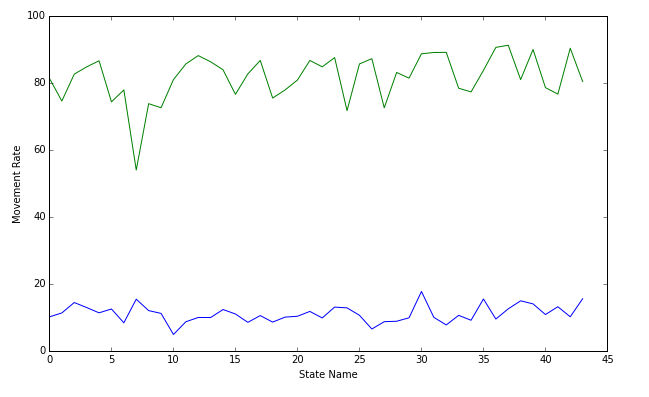


This graph shows similar results with last one. However, we can see that not every state with higher poverty rate lead to lower graduation rate. Which means, poverty level is one of the factors affect graduation rate, but not the root factor.

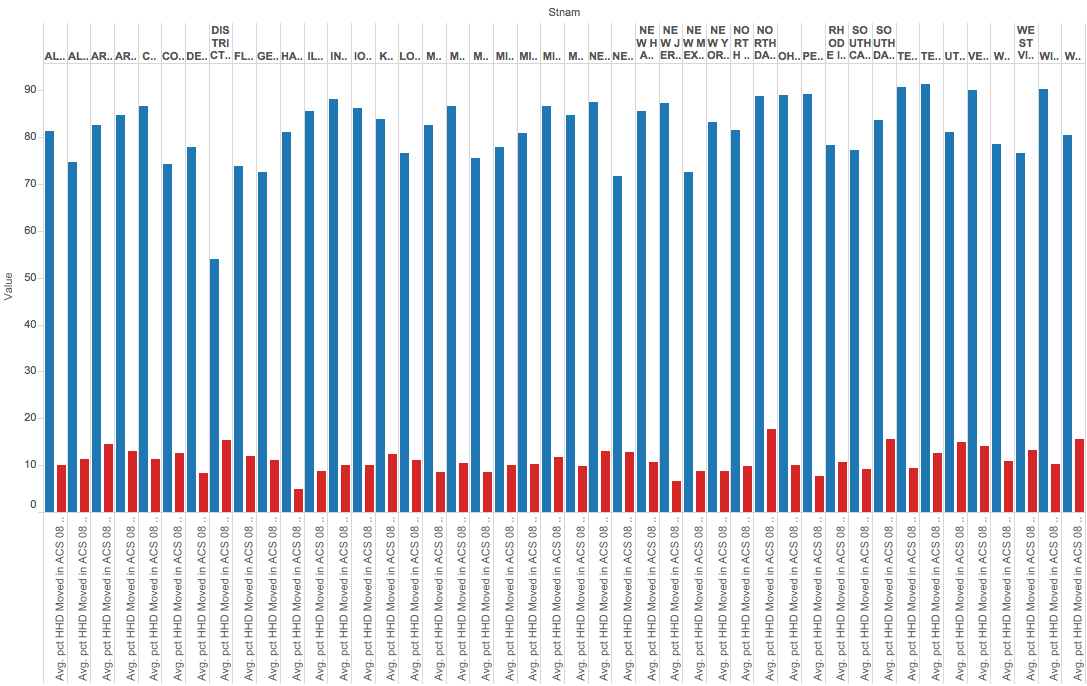


Based on these graphs, we can see that, poverty has negative correlation with graduation rate. The higher the poverty level, the lower the graduation rate. So the government is better to provide extra subsidy for families with high poverty level. It can help them to overcome some difficulties in living and help students more focus on study.

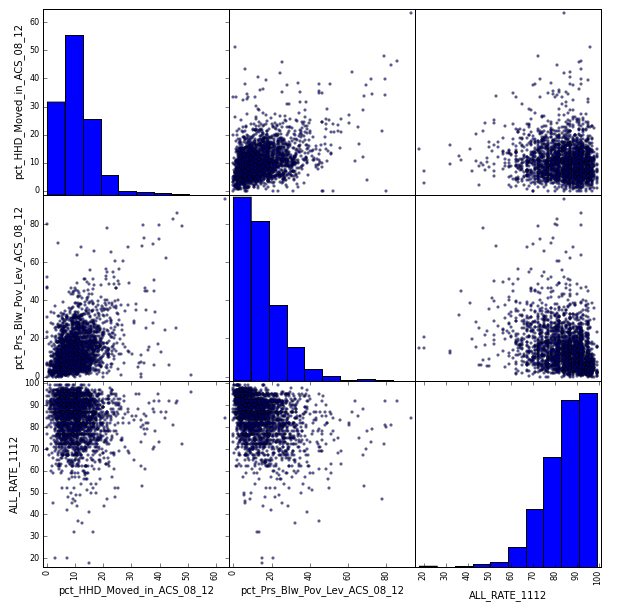
7. Movement rate vs. graduation rate



The top line shows the graduation rate and bottom line shows household movement rate. We can see that graduation rate fluctuates with movement rate. Just like #30 state, as the movement rate goes up, the graduation rate suddenly goes down, when the movement rate decreases for #31 state, the graduation rate goes up again. Which means household movement rate has negative correlation with graduation rate.



This graph shows similar results with last one. However, we can see that not every state with higher movement rate lead to lower graduation rate. Which means, movement level is one of the factors affect graduation rate, but not the root factor.



Based on these graphs, we can see that, household movement rate has negative correlation with graduation rate. The more frequency a household moves, the lower the graduation rate.

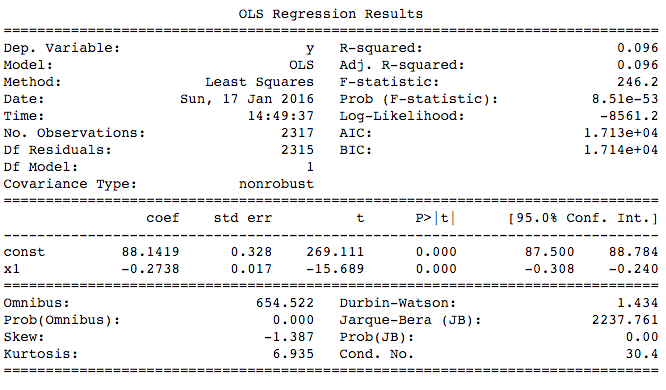
Each time a household moves and the student make a transfer, they has to face the complex and slow process of transfer and movement. It may take as long as a few weeks to complete the process and require the students to take off a few days to go through the process. So it will increase the absent rate of new students, which lead to lower graduation rate.

The government should regulate the process of transfer and movement, make it easy and simple to complete. And allow new moving family more time to finish all the process, and require student showing up as little as possible.

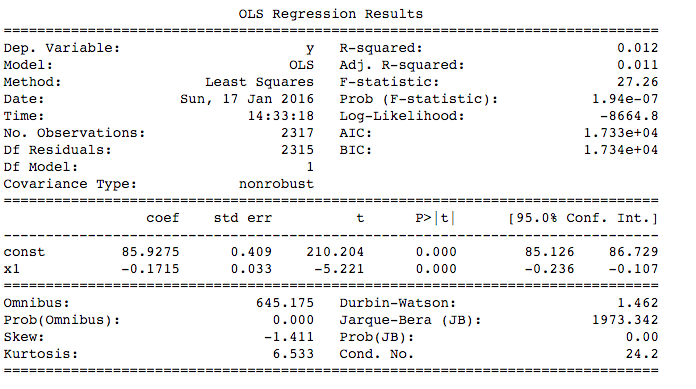
8. Linear regression model on poverty and movement

Similar results also showed on linear regression model. The negative x1 shows the negative correlation between two variables, but because the number is pretty small, it means the correlation is not very strong.

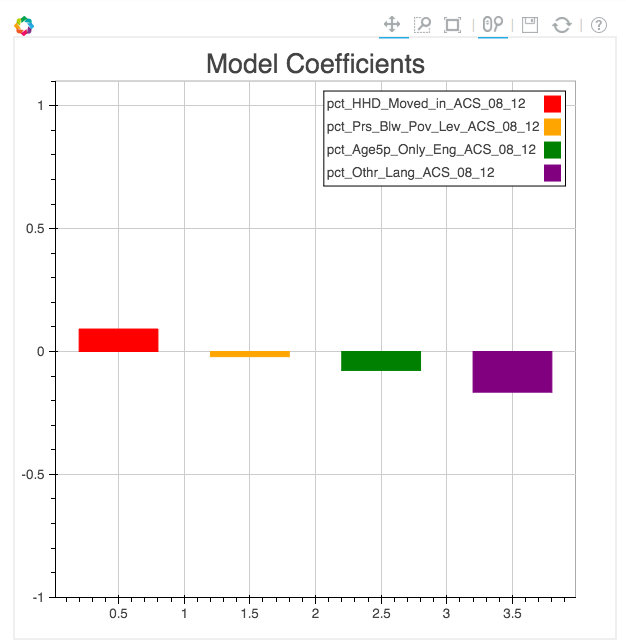
Poverty vs GR:



Movement vs GR



#### 9. Logistic Regression and Cross Validation on poverty level, movement rate, English level, etc.

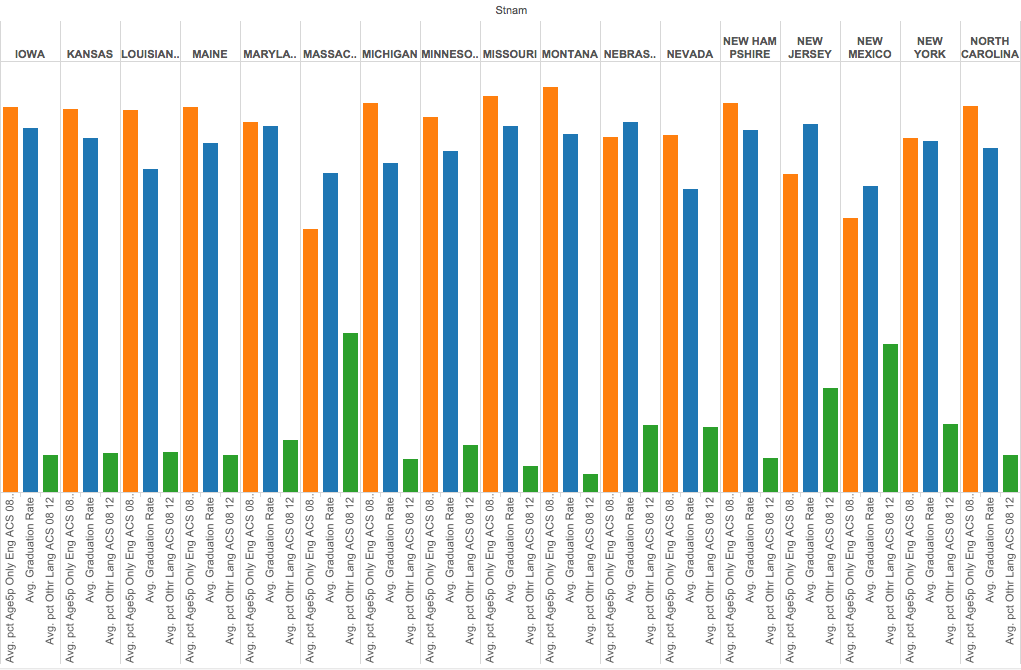


From the graph above, speaks other language rather than English and household movement has relatively high coefficient with graduation rate, which means these 2 variables are more affective on graduation rate.

Because speaks language rather than English variable has highest, but negative coefficient, which means it has negative correlation with graduation rate.

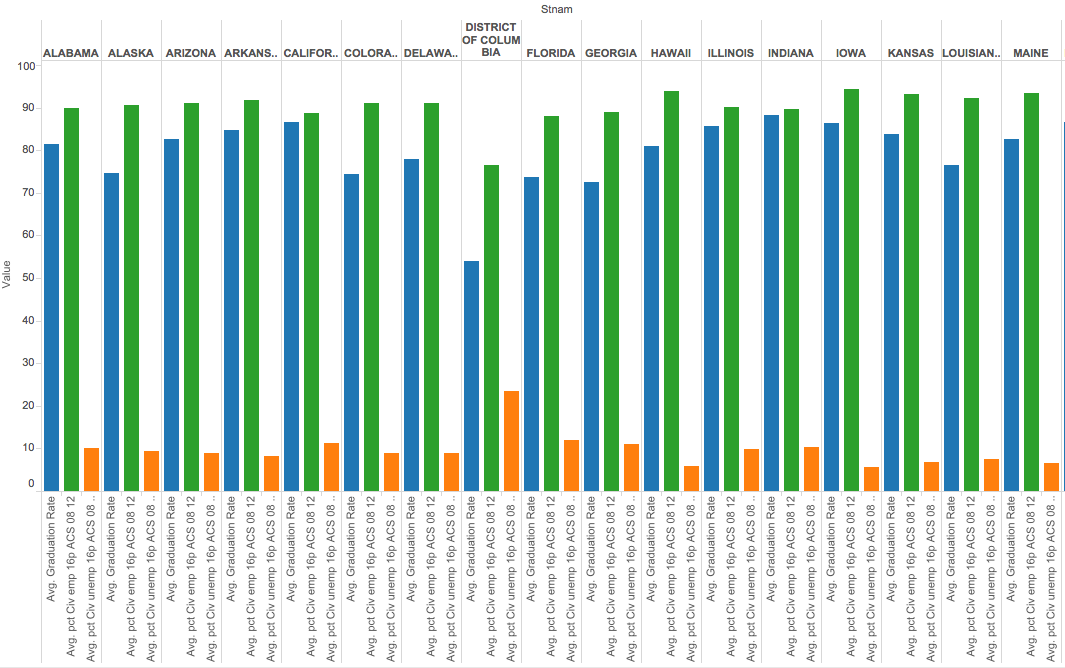
To help students who speaks English not well enough, schools could set up English tutoring to help these students to improve their English and better understanding about their learnings.

10. English Level vs GR



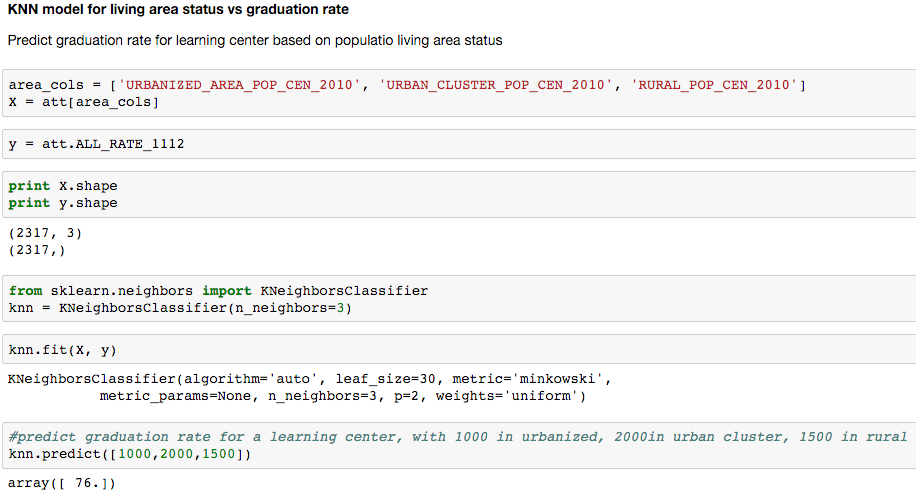
The higher percentage of students who speak English only, the lower percentage of students who speak other language at the same time. We can analyze the correlation between English speaking only students with graduation rate. We can see, higher percentage of English-speaking student lead to higher graduation rate. This may show the difficulties for other language-speaking students to understand their learning and get good scores. As my suggestion above, schools can provide tutoring for these group of students on English and study.

11. Age 16+, employed or not vs. GR



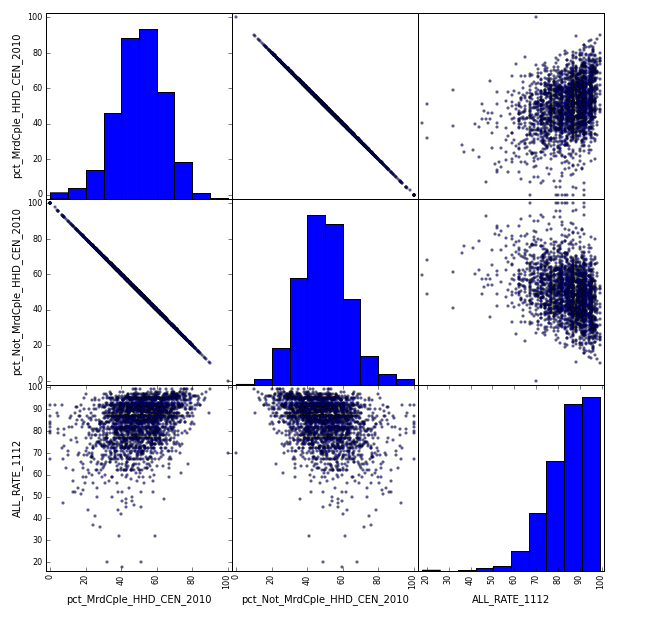
The reason I made this analysis is that I thought having a job while going to school may lead to lack of time on homework and study. However, from the graph above, we can see that, high rate of employment does not lead to high graduation rate, which means employment status of students do not contribute to graduation rate much.

12. Built a KNN model based population living area status to predict further graduation rate for new school center or following years



I built a model using KNN method to predict future graduation rate based on existing data. On the example, I used living area status, such as living in large city, urban or rural. Every year, school center can use the model and demographic information of students to predict graduation rate of students. Similar model can also be used on other variables, such as family status, age, etc.

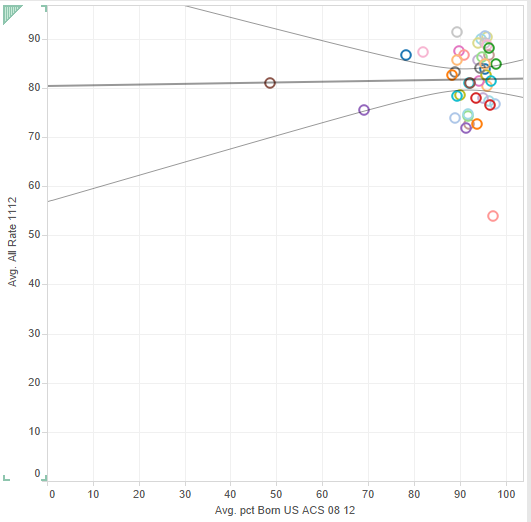
13. Linear regression of GR vs. family marriage status (single family or not)



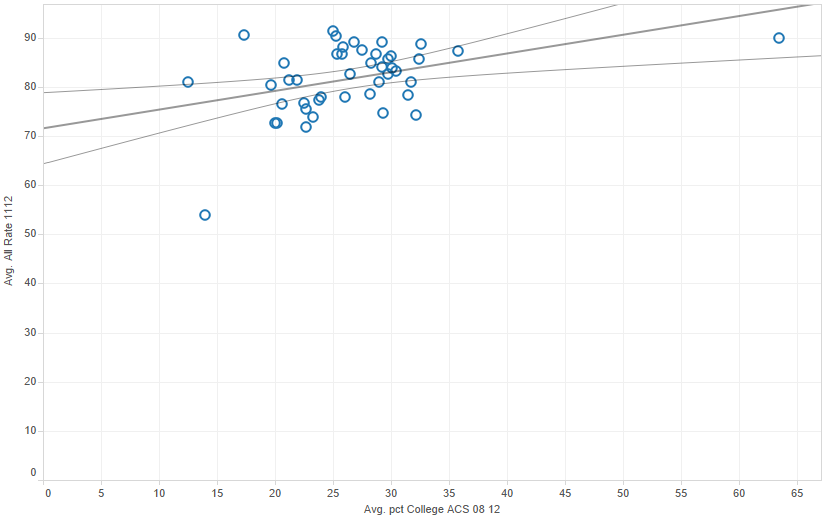
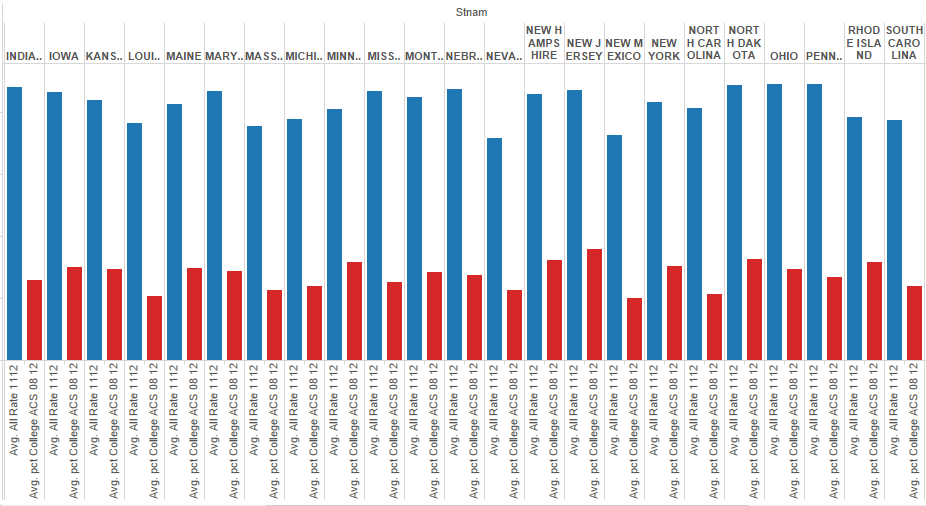
These graphs show the correlation of student family status and graduation rate. We can see that high percentage of family with both parents lead to high graduation rate, while high percentage of single-family lead to decreasing of graduation rate. This might because single families have more difficulties in living and students may not as confident as other students. Schools can build connection with single families and give concern about these students to see if they have problem in living and learning and may give them even some discount on tuition.

14. Citizenship vs Graduation Rate

Although the feature importance shows citizenship has impact on graduation rate, the graph below does not show positive or negative correlation.



15. Population education level vs. graduation rate



From the graph above, we can see that population education level has positive correlation with graduation rate, which means the higher education level the population is, the higher the graduation rate. This might because high education level of population provides students better teaching, tutoring and a better atmosphere to study. So for lower graduation rate schools, they should hire higher education level students and provide more tutoring to help students finish homework and understand knowledge.

**Conclusion/Implementation:**

1. Northern states have higher average graduation rate than southern states, due to better school-system quality in northern states. Government should improve school system in southern states, such as better teaching, tutoring program, better facilities, etc.

2. Poverty level has negative contribution to graduation rate, which means poor families are facing more difficulties to study and get good scores. Government can lower the tuition of poverty families and providing more benefits for families with students going to high school.

3. The more frequency a household moves, the lower the graduation rate. The government should regulate the process of transfer and movement, make it easy and simple to complete. And allow new moving family more time to finish all the process, and require student showing up as little as possible.

4. Higher percentage of English-speaking student leads to higher graduation rate. Schools should provide more tutoring for non-native students on English learning and study.

5. Students with both parents have higher graduation rate than single family. Schools can build connection with single families and give concern about these students to see if they have problem in living and learning. Government can provide more subsidies for single families as well.

6. Population education level has impact on high school graduation rate. Schools should have stricter recruiting requirement and hire higher education level teachers.

**Challenges and Successes:**

1. So many variables and missing values, find out the most valuable variables are challenging
2. Mass up types of variables makes model building challenging and difficult.
3. Successfully build KNN model to predict graduation rate base on new data.
4. Successfully find out some of the root causes related to graduation rate and give suggestion

**Potential Business Extension:**

Based on my suggestion, government and schools can make improvement on their policy and school system. They can also know the focus of their tutoring and teaching. It would be more efficiency to achieve the goal of 90% graduation rate in 2020.

**Reference:**

https://www.sciencenews.org/article/continental-divide-2014-temperature

http://weather.org/weatherorg\_records\_and\_averages.html

https://www.currentresults.com/Weather/US/average-annual-state-temperatures.php

https://wallethub.com/edu/states-with-the-best-schools/5335/