**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

**Pre-processing Dataset:**

1. Choose variables most related to topic, defined 515 variables

2. Clean data, eliminate most of the missing values in excel

3. Fill in average value for no value columns in Python

4. Use programming, visualization, classification technics to analyze correlation between factors and results

**Problem Statement and Hypothesis:**

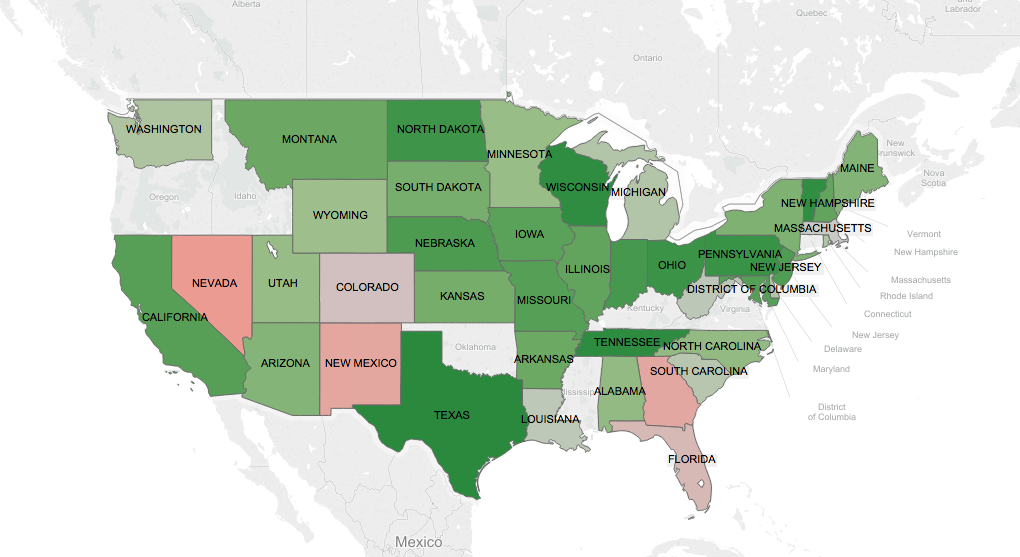
There are lots of factors can affect graduation rate, finding the root causes can help government set up new roles to help students and school increase graduation rate by 2020.

Some of the hypotheses are:

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

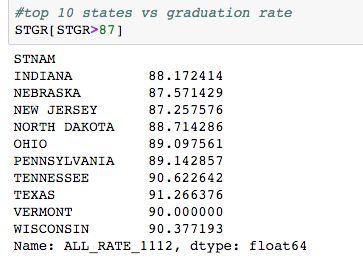
**Analysis:**

1. Graduation rate vs State level

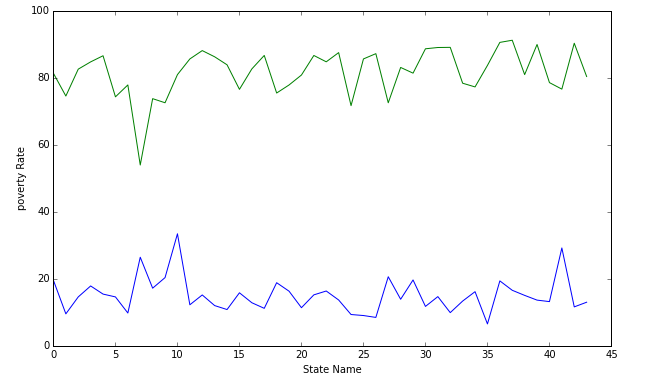


This map shows the graduation rate for each state. All the data are above 50%, so I set up red to be close to 50% and green as to be close to 100%. On the map, we can see, northern part of the country has relatively higher graduation rate than southern part. And even mid-land states has relatively higher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

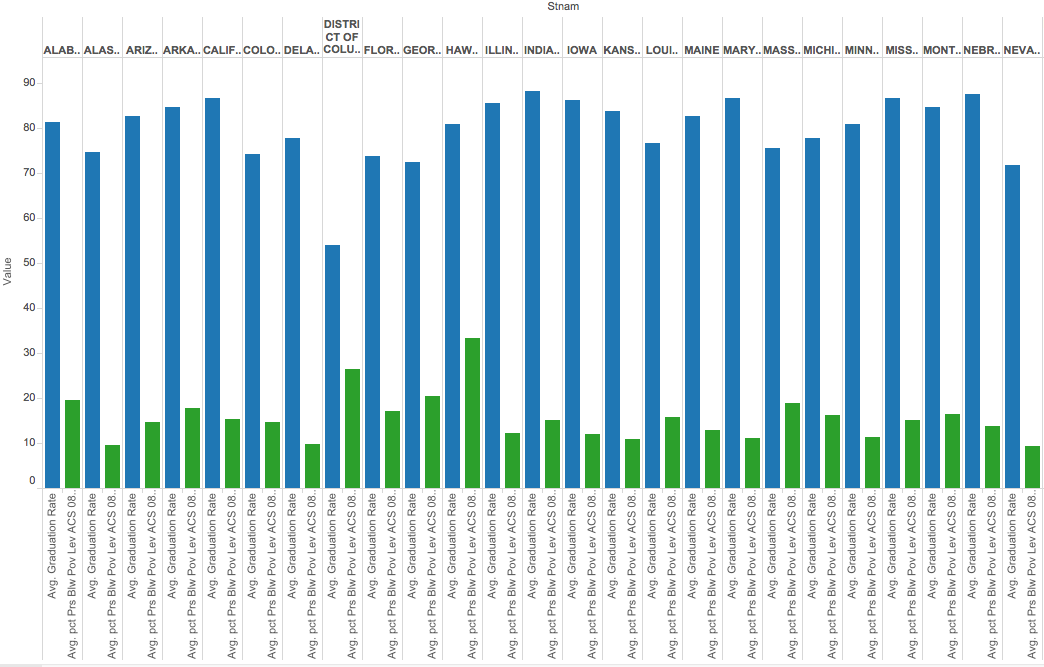
I also grouped up graduation rate by states in python to see top 10 states of highest graduation rate. As showing below:



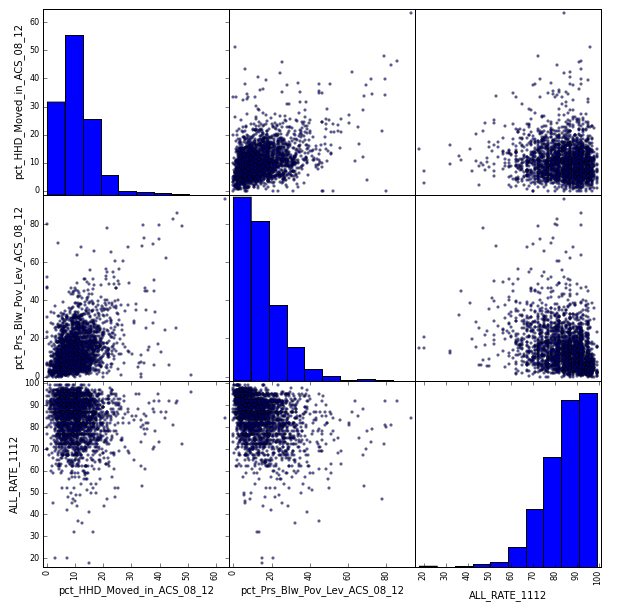
2. Poverty vs. Graduation rate



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.

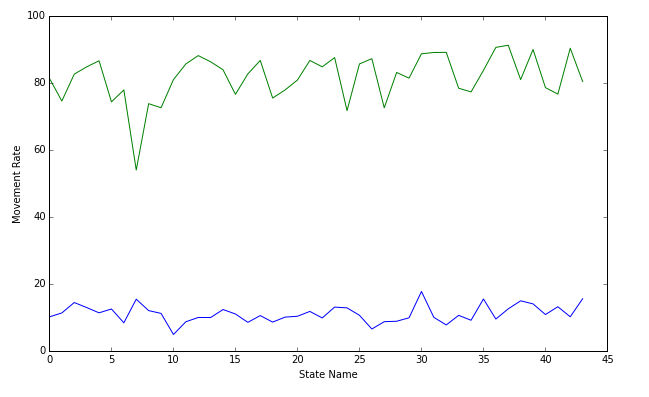


This graph shows similar results with last one. However, we can see that not every state with higher poverty rate lead to lower graduation rate. Which means, poverty level is one of the factors affect graduation rate, but not the root factor.

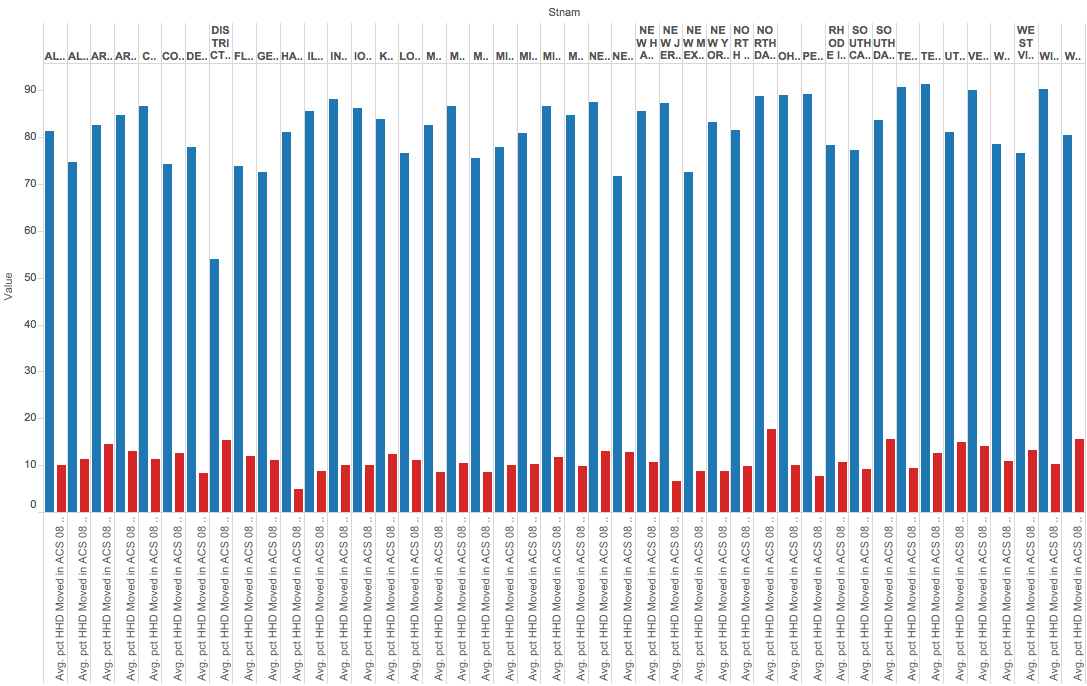


Based on these graphs, we can see that, poverty has negative correlation with graduation rate. The higher the poverty level, the lower the graduation rate. So the government is better to provide extra subsidy for families with high poverty level. It can help them to overcome some difficulties in living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

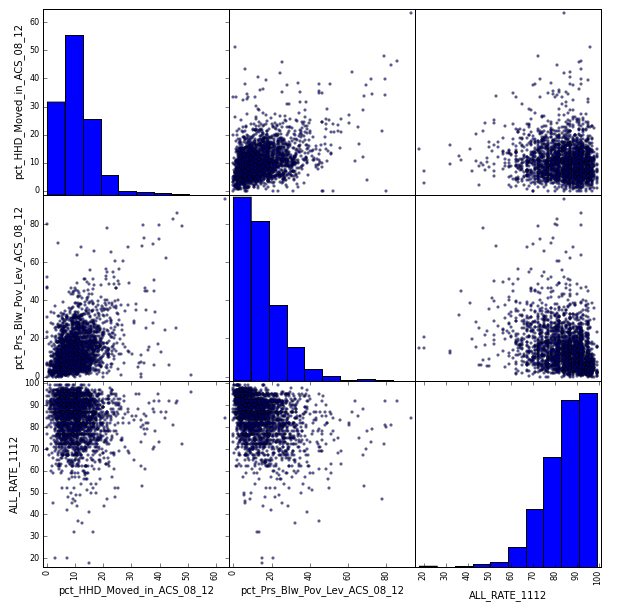
3. Movement rate vs. graduation rate



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.



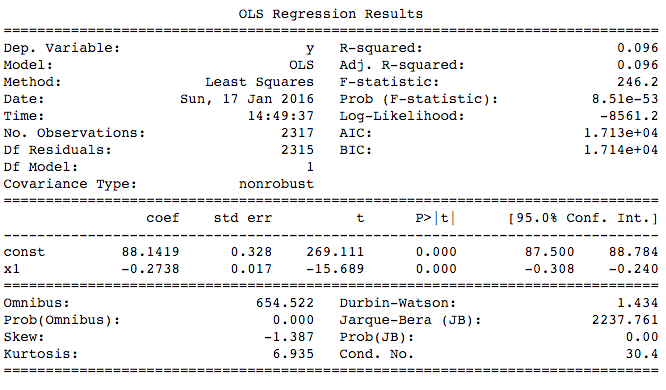
Based on these graphs, we can see that, household movement rate has negative correlation with graduation rate. The more frequency a household moves, the lower the graduation rate.

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.

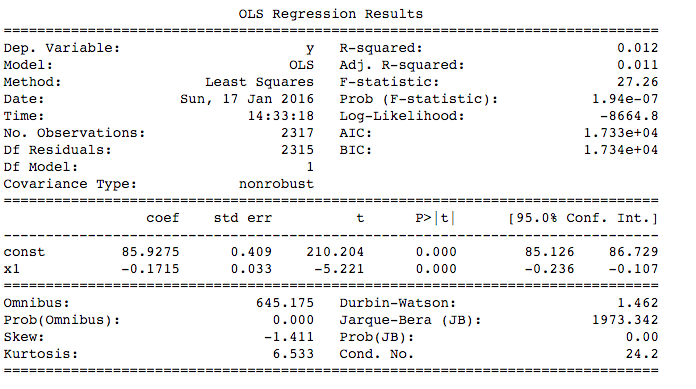
The government should regulate the process of transfer and movement, make it easy and simple to complete. And allow new moving family lower time to finish all the process, and require student showing up as little as possible.

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.

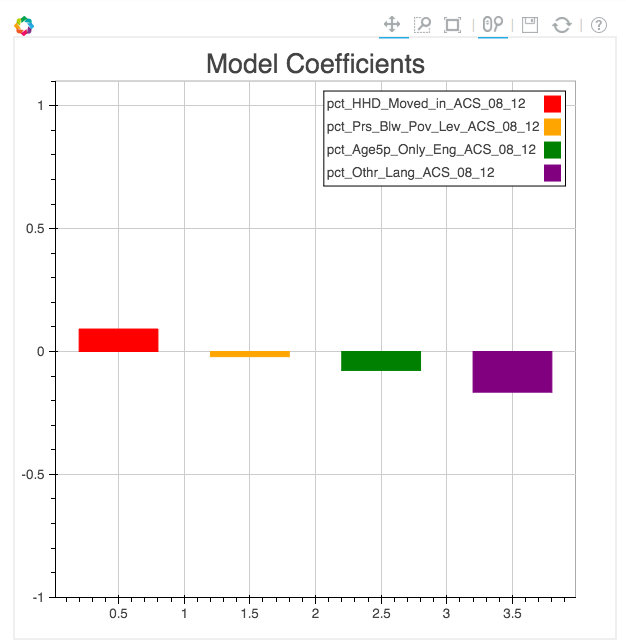
Poverty vs GR:



Movement vs GR



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

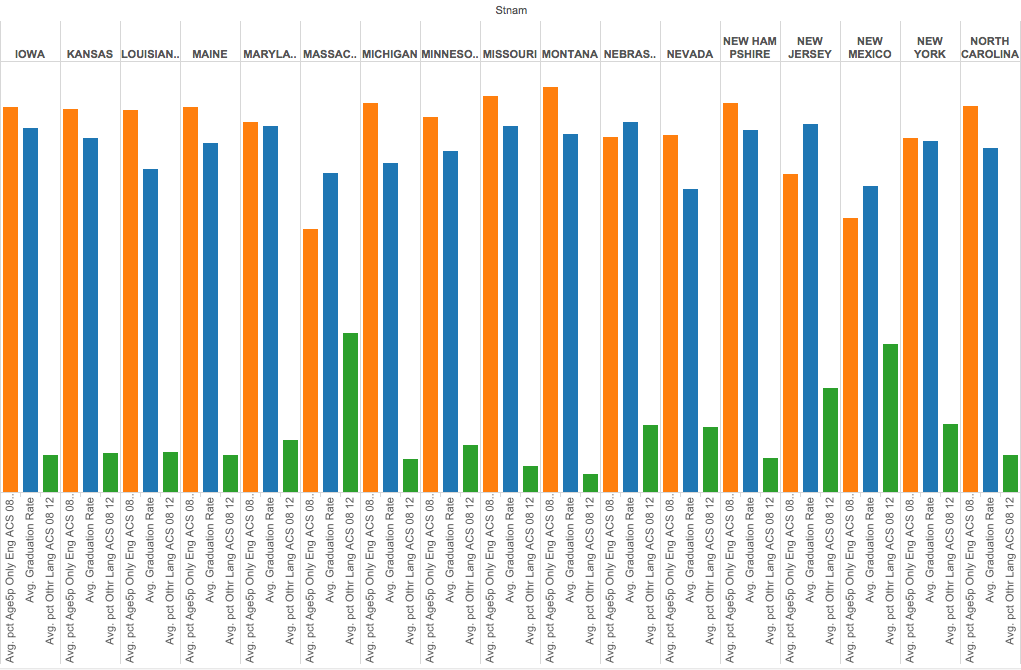


From the graph above, speaks other language rather than English and household movement has relatively high coefficient with graduation rate, which means these 2 variables are more affective on graduation rate.

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

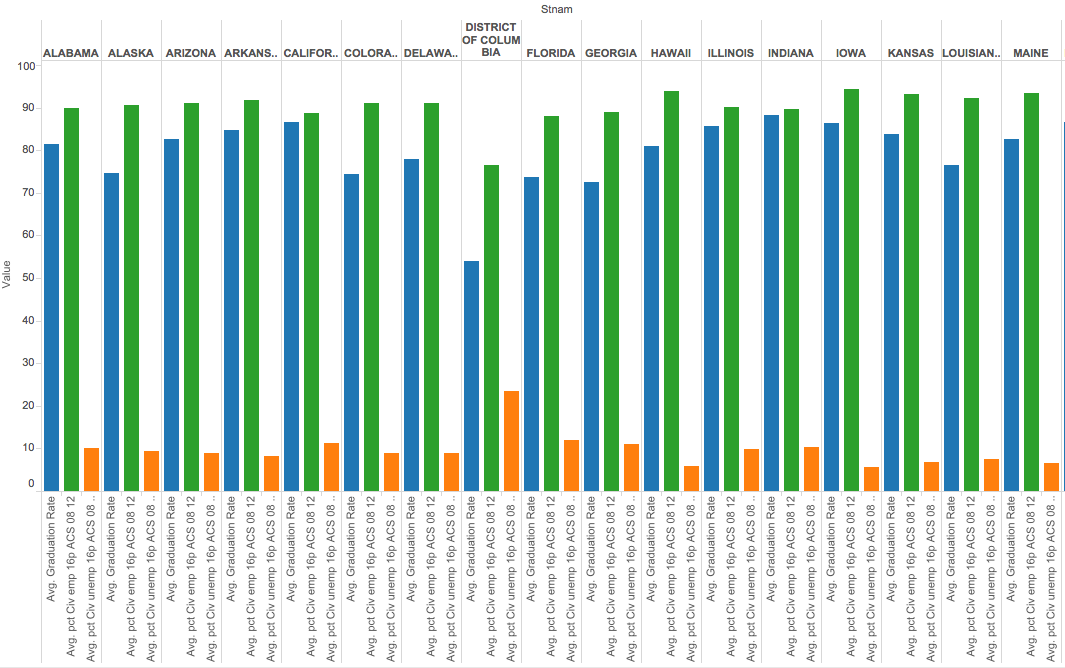
To help students who speaks English not well enough, schools could set up English tutoring to help these students to improve their English and better understanding about their learnings.

5. English Level vs GR



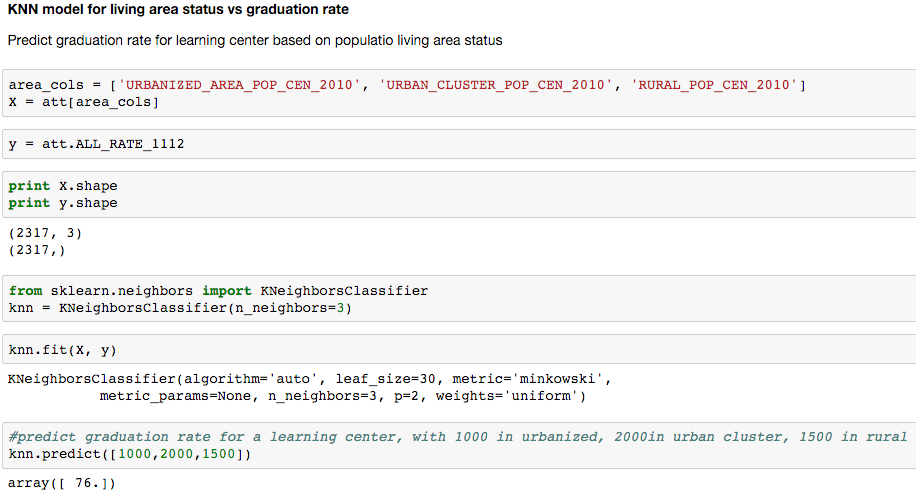
The higher percentage of students who speak English only, the lower percentage of students who speak other language as well. We can analyze the correlation between English speaking only students with graduation rate. We can see, higher percentage of English-speaking student lead to higher graduation 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.

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



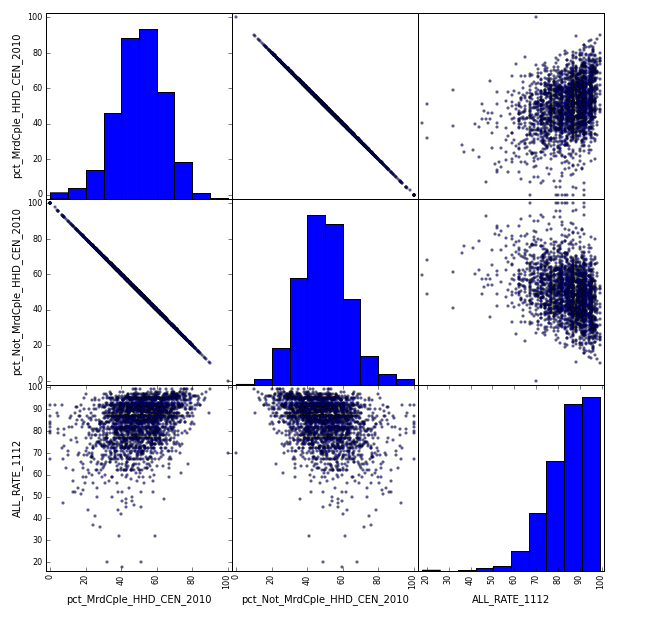
The reason I made this analysis is that I thought having a job while going to school may lead to lack of time on homework and study. However, from the graph above, we can see that, high rate of employment does not lead to high graduation rate, which means employment status of students do not contribute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

7. Built a KNN model based population living area status to predict further graduation rate for new school center



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.

8. 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.