Capstone 1:

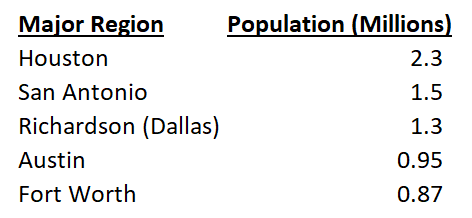
Texas Education

*(Predicting Percentage of Students Who Will Graduate College Within 4-Years, Based on the Features of the School District They Attended High School in)*

***“We Would Like Our Child to Attend College. Do You Know Which School Districts Improve the Likelihood of this Happening?”***

Many parents across the country find themselves asking each other this exact question. Their motive is simple, they are looking to provide their child with a quality education. You hear it all the time, education is key! It has been proven time and time again that a quality education can unlock doors for the average individual. It’s common belief that having your child attend college will lead to a more comfortable life, which is what all concerned parents want for their child transitioning into adulthood.

In this study I aimed to help parents considering the move to or within the “major regions” (listed below) of Texas. These particular regions were chosen as they represent the areas with the most economic opportunity, making them prime locations for families and talented educators. The information provided is strictly educational in providing parents with “food for thought” ahead of a potential move. Parents may even use the data to rule out a move entirely should their current school district be favorable among other options.



In order to achieve this goal, we first must address the common question from parents mentioned at the top of the page. While this question has good intentions, I don’t necessarily believe this is the right question to ask for your child. Parents need to ask “which school districts are proven to contain a higher percentage its graduates earn a college degree within four years?” instead of simply “which school districts increase the likelihood of my child getting into college?”

The honest truth is that there are many colleges willing to accept your child even if they had poor college admission test results. Why is this the case? Colleges are operated as businesses. They will gladly collect expensive tuition checks until the student fails out (not prepared) or earns a degree (prepared student). Depending on the college, there’s often a considerable dropout rate for freshman who have come in unprepared.

In a 2011 Harvard study “Pathways to Prosperity”, the U.S. contained the highest college dropout rate among industrialized nations. Among four-year colleges, just 56% of students graduated within six years (not four). One must remember that more time spent at a college means more money spent or more debt accumulated. Financial pressure and academic disqualification remain the top two reasons why a student drops out of college. This is why I chose to focus on four-year graduation, as that should be the goal.

To answer the college graduation question, historical data was collected on the percentage of students who were able to graduate college in four years’ time after graduating high school for each respective school district. School district features were also collected to measure their effect on the resulting college graduation percentage. Utilizing all the historical data for the major regions, I aimed to build a predictive model that would help estimate the percentage of students going to college (from a particular district) that will graduate within four years’ time. Later in the report I will go into more detail about the chosen school district features, but for now I’ll provide a brief summary as to what they are in case the reader has not seen them before.

The SAT and ACT are both college admissions tests that indicate a student’s readiness for college. Thus, you can see their importance for this study. The SAT includes a Math and a Reading/Writing section that are both scored out of 800 points for a maximum possible score of a 1600. The ACT includes a Math, Reading, English, and Science section which are all scored out of 36. A student’s composite score on the ACT is the average score of all these sections.

AP exams are taken at the end of the school year for those enrolled in AP classes in high school. These AP classes are comparable to introductory college classes and cover several subjects like calculus, physics, literature, biology, and psychology, for example. At the end of the class, students have the option to take AP exams for college credit. The tests are scored out of five with most colleges nationwide accepting a passing score of a three. In this study we will use a three as a passing score translating into college credit.

Though it doesn’t represent our main concern in assessing the quality of a school district for a student, the percentage of students who enroll into college after graduation is an important feature to explore. Sure, part of this feature is influenced by the wealth of parents residing in the district, but it also can be influenced by impressive test scores that increase the amount of scholarship money offered by colleges to students.

A family’s financial state does not only affect college enrollment, but also its ability to reside in a particular school district. Housing prices/taxes in the respective school districts are not all created equal, as demand can be higher for certain areas. This demand is affected by several features, but a major one is certainly the quality of education. I acquired each district’s historical wealth per average daily attendance (“Wealth/ADA”) data for later analysis.

Before we move further, I find it important to clarify how college enrollment percentage and college graduation percentage is assessed for a particular year and to note that this study strictly focused on Texas colleges (statistics do not include students who enrolled into colleges outside the state). If we say the year is 2013, then the percentage of students who enrolled into colleges in the fall of 2013 is denoted as “Enrolled 4-Year (%)” and the number of students enrolling is denoted as “Enrolled 4-Year”. The percentage of those students who were able to graduate four years later (2017) is denoted as “Graduated 4-Year (%)”, representing the target variable I attempted to predict.

Now that we understand the desired data for each school district, let’s go over how I collected and wrangled it for later analysis/modeling.

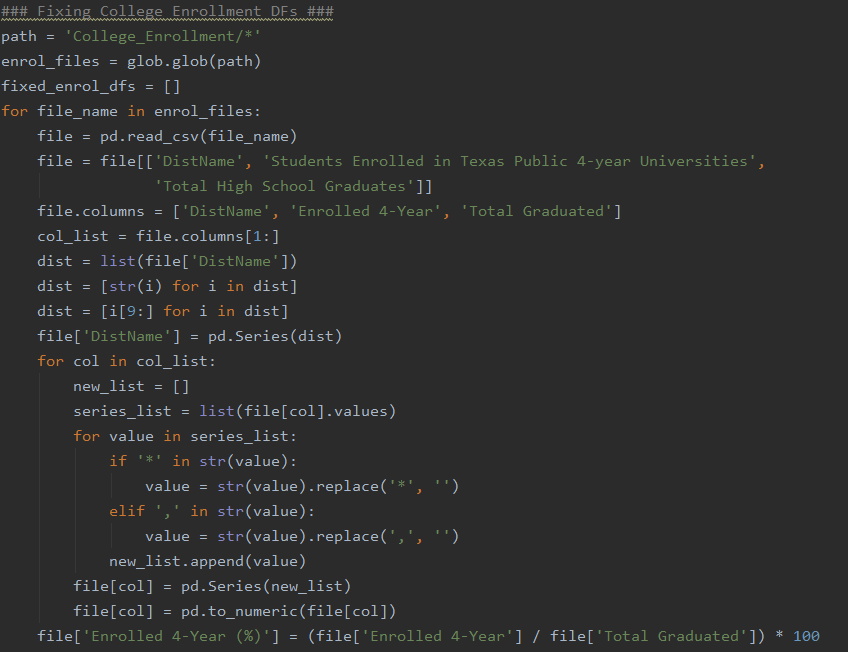
**Data Wrangling**

SAT, ACT, AP exam, and Wealth/ADA datasets were downloaded from the Public Education Information Management System (“PEIMS”) on the Texas Education Agency’s website. At the time of this project, the latest data out was from the class of 2017 and the earliest was from the class of 2011, resulting in seven classes of full historical data for each respective school district. Having the data from before the class of 2011 would be nice, but I also do believe there’s a sense of “what have you done for me lately” that goes into assessing a school districts value to a student. The datasets on college enrollment and college graduation were downloaded from the Texas Public Education Information Resource (“TPEIR”) website. Again, It’s important to note that this data strictly focuses on Texas colleges.

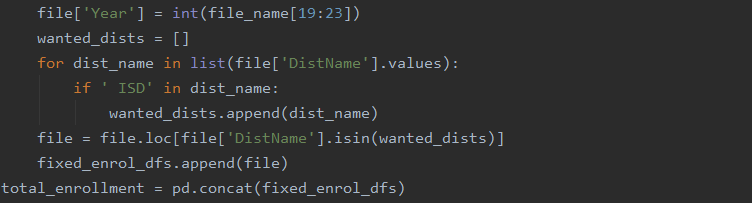
From there I needed to recognize what variables were important in each dataset and remove the unnecessary ones using my knowledge/experience of once being a high schooler in Texas. The ACT, SAT, and AP datasets all contained breakdowns by ethnicity for scores, but I filtered these to simply be “all students” for each respective district. The SAT and ACT datasets contained average scores (“SAT-Total”, ”ACT-Composite”) for each school district and participation data was also included in both. The AP dataset contained data on participation, amount of AP exams taken in each respective district, and the amount of passing exams (scored 3 or above).

For each dataset type mentioned above, I created a folder containing the separate classes with the goal of later creating one “Total” dataset (Ex: Total\_AP) containing respective district data for 2011 – 2017. With the different “Total” datasets in place, I could merge them into one final dataset containing all the district features. Before being able to merge all the data and perform analysis, the data needed to be cleaned and wrangled.

First up were the datasets on college enrollment. The main problem here was that district names contained an ID number and name in all caps (Ex: 4825170 KATY ISD). The ID number is not necessary, so I got rid of it to leave the District name in all caps. I also needed fix numerical data that contained “\*” (data not available) or was represented as a string with a comma (Ex: 1,244). From there I was able to calculate the percentage high school students who were able to enroll into a four-year college that fall.

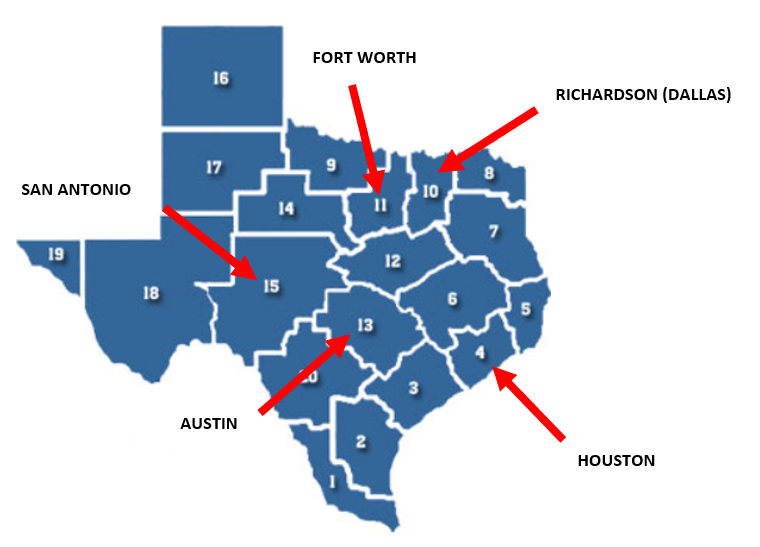


The respective classes were added to each dataset and public districts were extracted using district names that contained “ISD” (Independent School Districts). The datasets included academies/prep schools, but I wanted to strictly focus on public school districts.

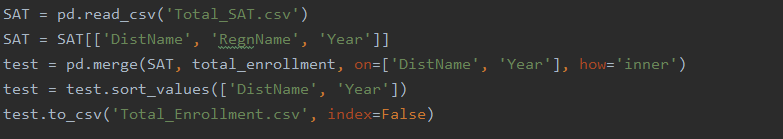


After each enrollment dataset was cleaned, it was appended to a list to later be concatenated into one total (2011 – 2017) DataFrame. To spare the reader from having to look at too much code, this relative process was repeated for each dataset type. The dataset taking the most code to clean was the SAT due to 2011 – 2016 scores needing to be converted from the old scoring system to the new (which is utilized in the 2017 scores). This was done using CollegeBoard’s SAT concordance tables.

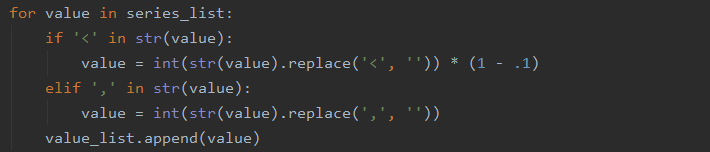
I noticed that the enrollment data also did not contain each district’s respective region name, which was important in meeting my desire to focus on the major regions. To give the reader a visual perspective of the size of the major regions that districts belong to, I have provided the image below. This is how the Texas Education Agency splits up the educational regions for its data.



Luckily the SAT datasets contained which region each district belonged to. Using the total SAT dataset, I performed an inner merge to provide the enrollment dataset with respective region names for the school districts.



For the AP datasets, they all contained numerical approximations in the form of strings (Ex: <60) and some instances of the string with a comma problem. I decided to be consistent in decreasing the number by 10% for each of the “less than” cases.



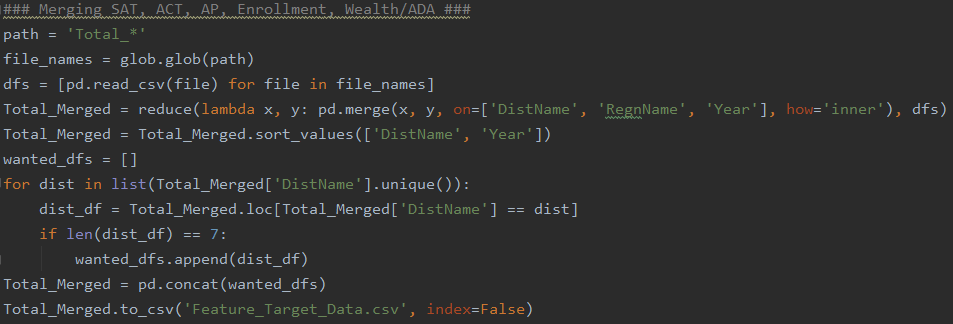
Upon fixing the numerical data, I was then able to add the number of AP exams taken per student in each district for each year.



Besides filtering for public schools in the major regions, I also decided to remove districts that did not take more than 50 AP exams. The removed districts represent small ones with very limited data. Rather than play a guessing game for the features of these small districts (which has a large effect on statistics involving percentages), we can just exclude them from our analysis as families very rarely decide to move to these districts for educational purposes.



With all the DataFrames containing 2011 – 2017 data for public school districts in the major regions of Texas, I was then able to merge them all into one DataFrame



Seven years of historical data gives a parent a good idea of what they can likely expect from a particular school district’s features going forward.

For the college graduation data (the target), I manually inputted the number of students who were able to earn their college degree within four years of 2011 – 2014 (what was available). To clarify for the reader, this can be read as the number of students who earned their college degree in 2015 – 2018 that belonged to the high school classes of 2011 – 2014.

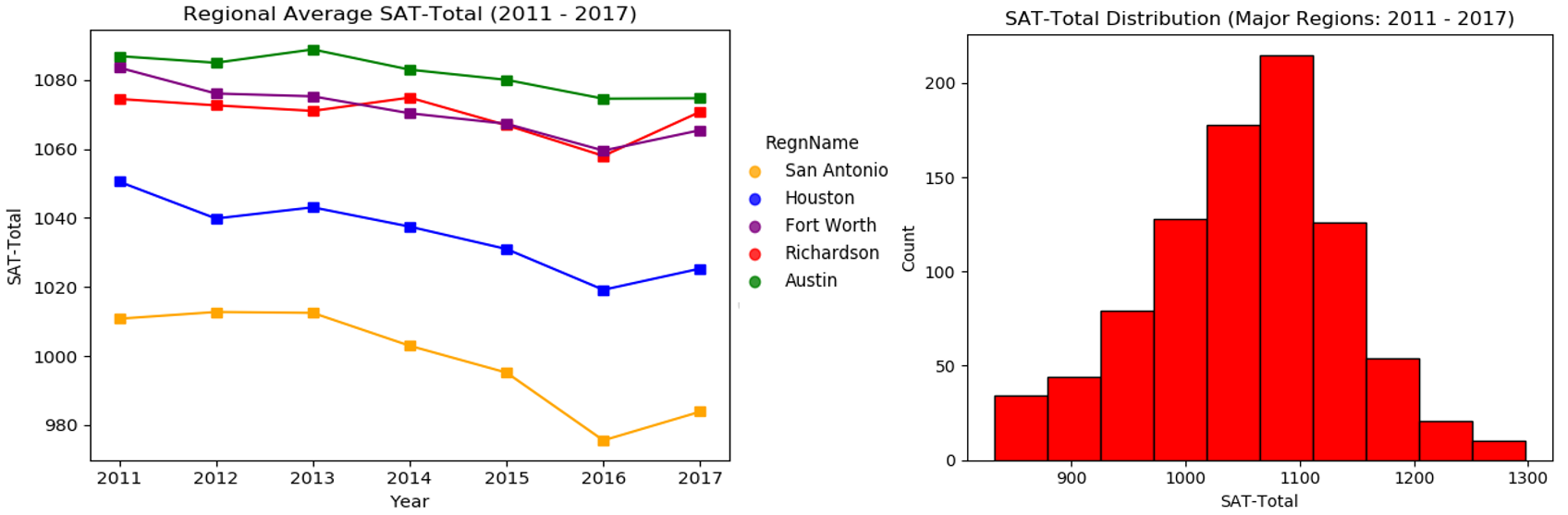
With the number of students graduating college within four years and the number of students who enrolled into college, I was then able to calculate the percentage of students who were able to earn their degree.

Following all this data wrangling/cleaning, I was left with a dataset that included all feature and target data for the classes of 2011 – 2014. For 2015 – 2017, only the feature data was available and the target data (college graduation percentage in 2019 – 2021) is unknown. We can try to later predict the target after establishing a satisfactory machine learning model.

**Data Analysis: School District Features**

*(SAT)*

Let’s start of by viewing the regional averages for each class year and the distribution of district-level scores.

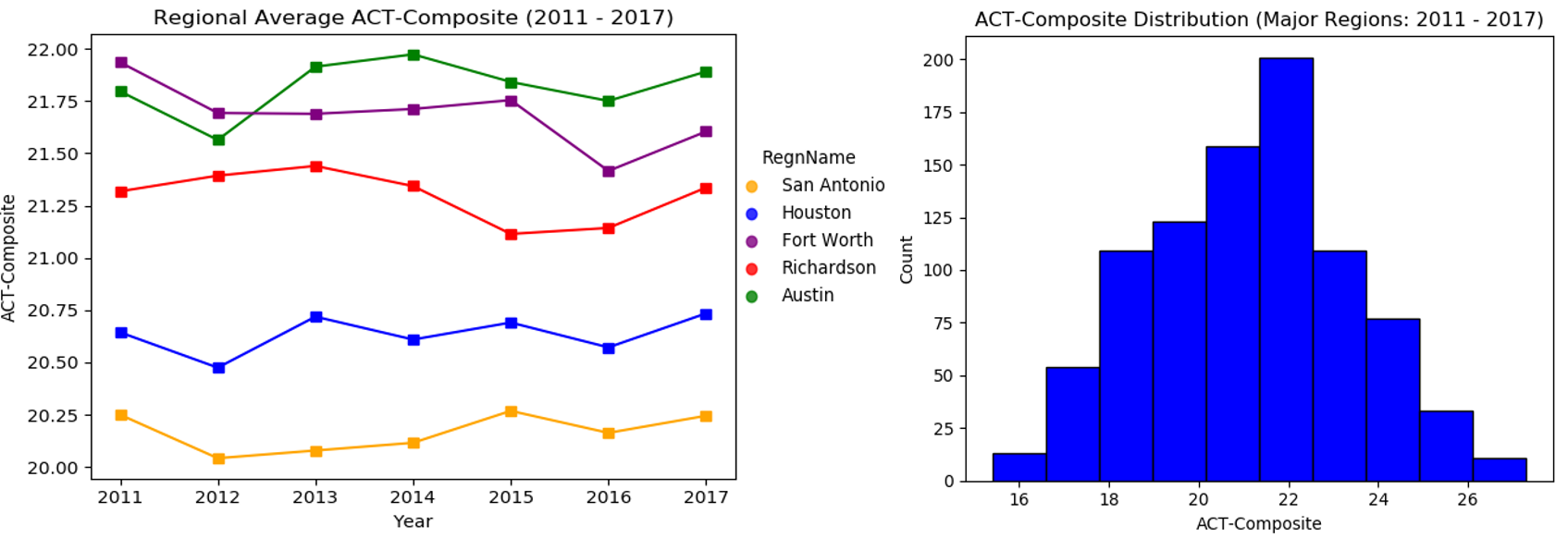


Mean District-Level SAT-Total:

1052.8

We can see that Austin performed the best while Fort Worth and Richardson were neck and neck. The gap between the top region and San Antonio is quite large. This will, unfortunately for the region of San Antonio, remain a common theme throughout this analysis.

*(ACT)*

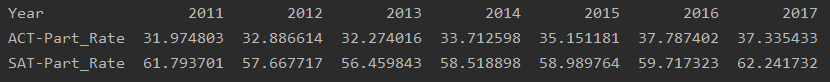


Mean District-Level ACT-Composite:

21.1

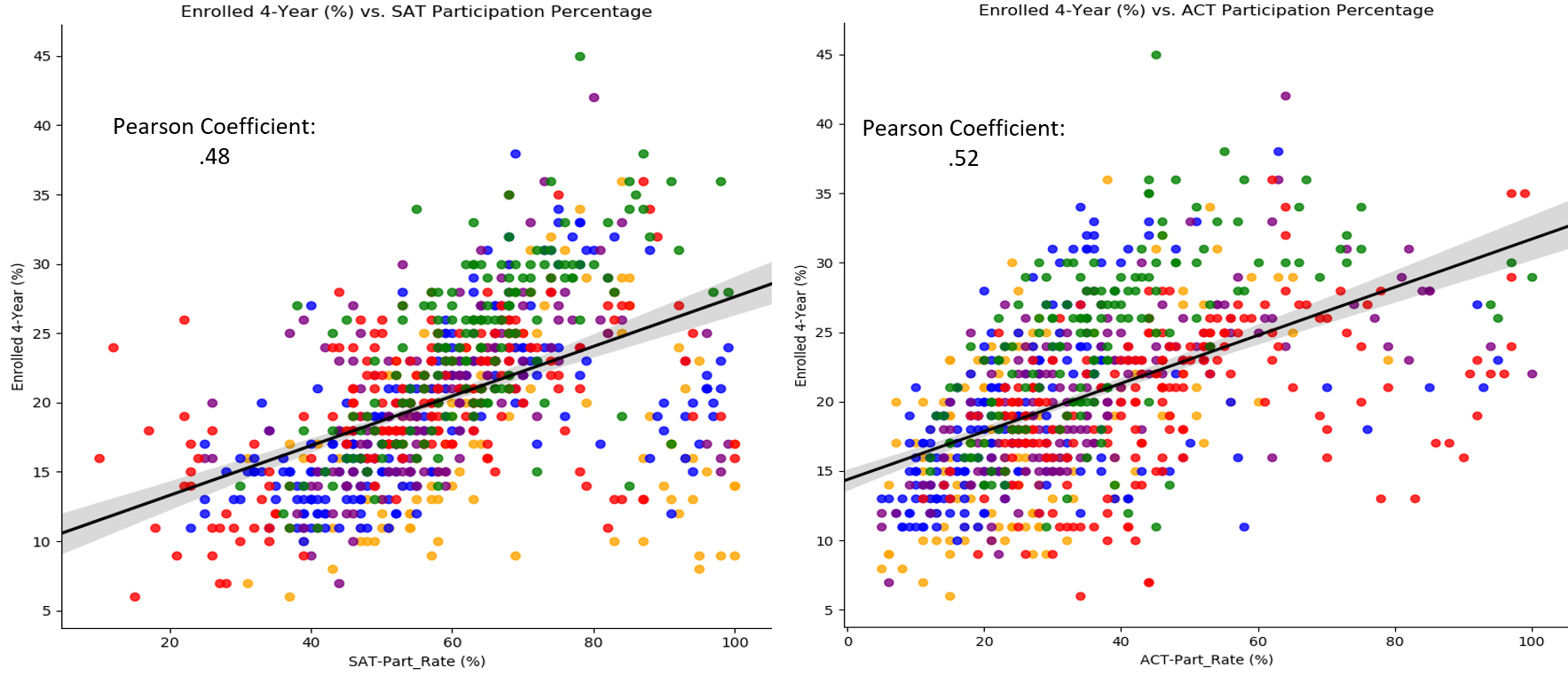
It appears that the districts located in Austin and Fort Worth achieved the best average scores on the ACT with Richardson not too far behind. As with SAT scores, there appears to be a slight uptick in the average scores for the class of 2017 (improved educators/smarter kids?).

*(SAT/ACT Participation %)*



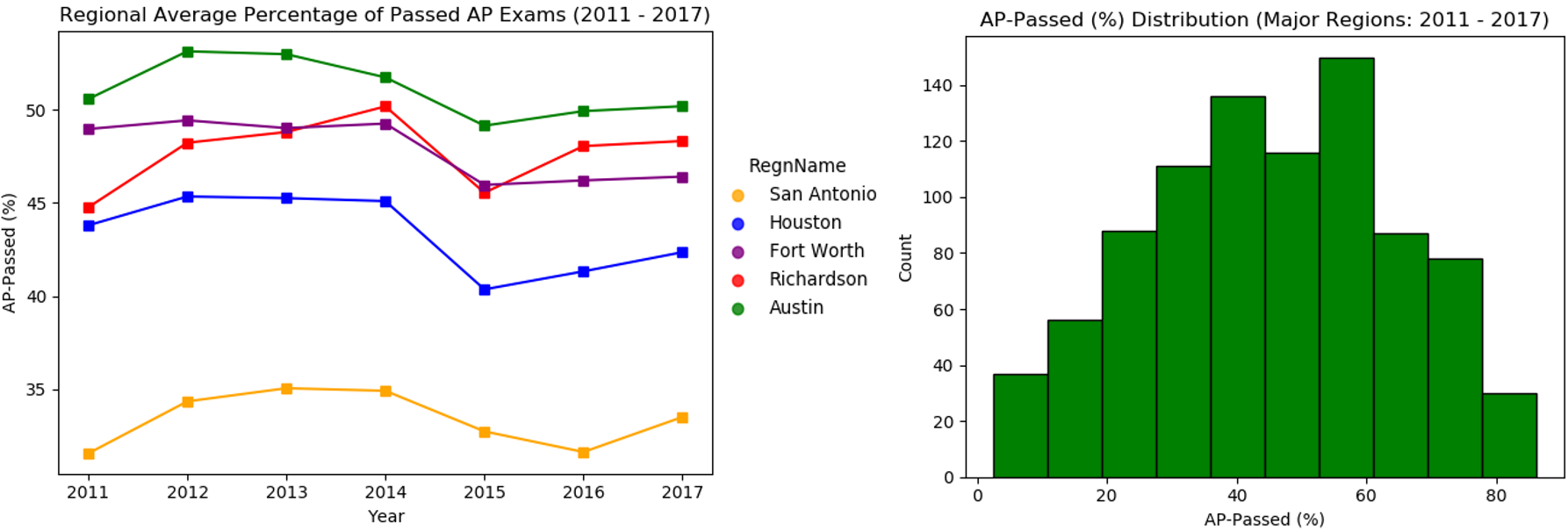
Something interesting I discovered is that more students consistently choose to take the SAT than the ACT. Why is this? Well… it’s very tough to know for certain without surveying high school students across Texas, but I’ll provide a possible theory. The SAT is the test students hear the most about while growing up from friends, parents, and teachers. You could say it’s the “standard test” students feel they have to take in high school.

It’s good to see in the later years that ACT participation grew, but it is also quite concerning that students are not more encouraged to take the ACT as well as the SAT. Taking two tests instead of just one immediately increases the chance of a student doing well on at least one. Doing well on one of the college admission tests is all it can take to get into college and earn scholarship money. With scholarship money, enrolling into college becomes more likely/possible for a student. As you can see below, both SAT and ACT participation rates contain positive correlation with Texas college enrollment percentage. Why not take both??



*(AP Exams)*

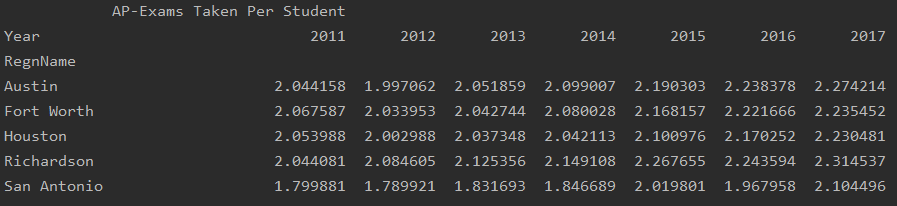
Note: In this study, a score of three or above on an AP exam is considered passing. Most colleges accept this score.



Mean District-Level AP-Passed (%):

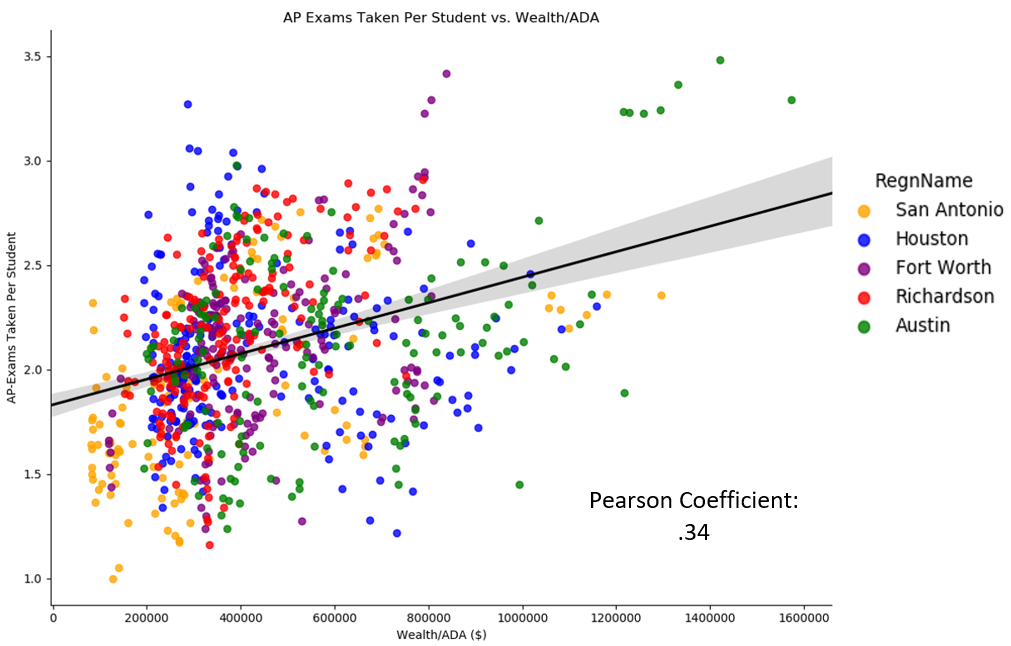
45.1

The regions of Austin, Richardson (Dallas), and Fort Worth appear to contain the best passing percentages. I found it interesting to also take a look at the availability of AP classes to students. One could argue that more availability to AP classes would result in a student being able to take more exams and earn more college credit/gain more college-level exposure.



From the standpoint of AP class/exam availability, it appears that Richardson (Dallas) gives the most opportunity to students to earn college credit. Austin tends to offer the second most opportunity, which is quite impressive when also taking into account that Austin contained the best passing percentage.

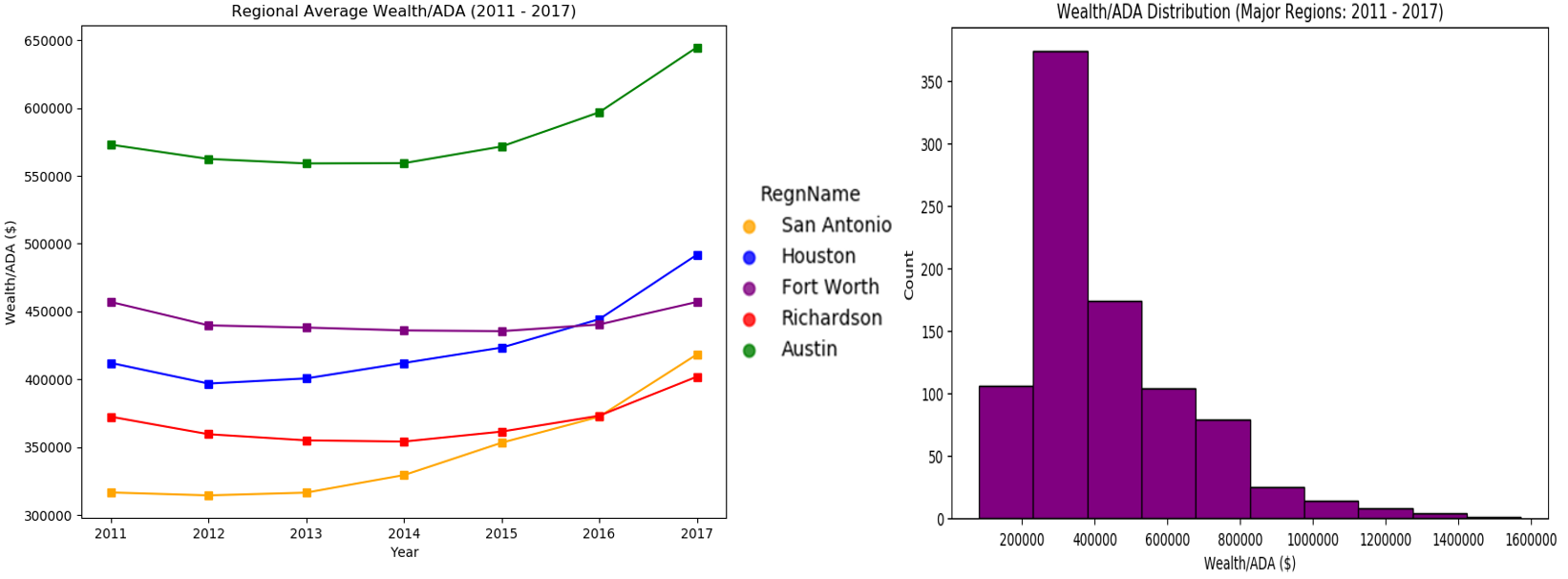
The next figure will offer a transition to our next analysis of the wealth per average daily attendance (“Wealth/ADA”) feature of school districts. Let’s take a look at how Wealth/ADA affected the amount of AP exams taken per student throughout the classes of 2011 - 2017. One might hypothesize that we could see a positive correlation as wealth/funding would bring in more qualified teachers and increase the number of AP classes offered to students.



As anticipated, there was indeed a positive correlation. The cluster on the top rights represents Eanes ISD in Austin, which averaged the largest Wealth/ADA. The school that averaged the second highest Wealth/ADA was Alamo Heights ISD in San Antonio. It’s interesting to view the difference in Wealth/ADA between the Alamo Heights ISD cluster and the rest of the districts from San Antonio.

*(Wealth/ADA)*

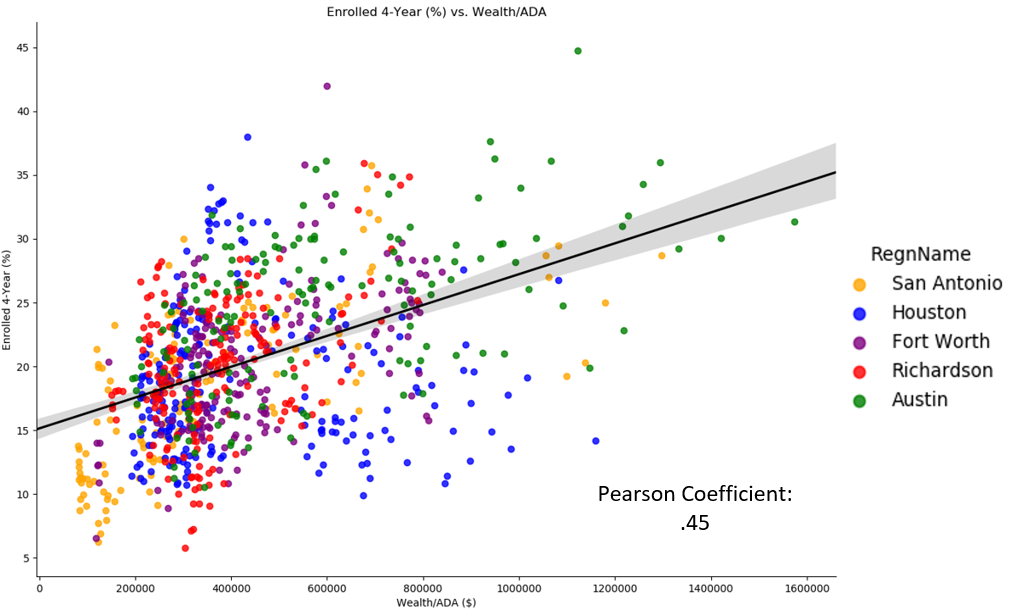
Austin held a healthy lead over its nearest competitors in Fort Worth and Houston. Overall, it appears that Wealth/ADA has been increasing in recent years. Something we could later choose to explore is the average property tax for homes in each region.



Mean District-Level Wealth/ADA ($):

431,718

It should be interesting view how Wealth/ADA correlated with college enrollment percentage for the different class years as well. It would not be unreasonable to hypothesize that a student’s ability to enroll into college increases with wealth.



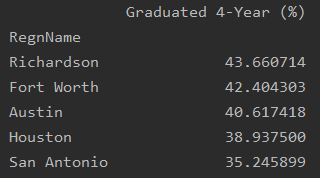
Though it certainly isn’t the only factor influencing college enrollment percentage, Wealth/ADA does indeed contain a positive correlation with college enrollment percentage. As mentioned before, this can partially be attributed to “wealthier” students having more access to college.

Let’s say a student has poor test scores, but comes from a family with money. There are several colleges that are still willing to accept this student. Even if the student ends up failing out, the college will still be entitled to tuition payments.

With that being said we still have to acknowledge that there are plenty of students in the districts containing higher Wealth/ADA that are taking advantage of the quality education made available to them in testing well and earning scholarships/being more attractive to colleges. Though I can’t necessarily prove it with data right now, it’s not egregious to assume that many of the parents in these “wealthier” areas are well educated and therefore, encourage/press their children to do well in school as they know firsthand what education can bring to an individual’s life.

**Data Analysis: Percent of Students Earning College Degree Within Four Years**

For the classes of 2011 – 2014, let’s look at the average percentage of students who were able to earn their college degree within four years by region.

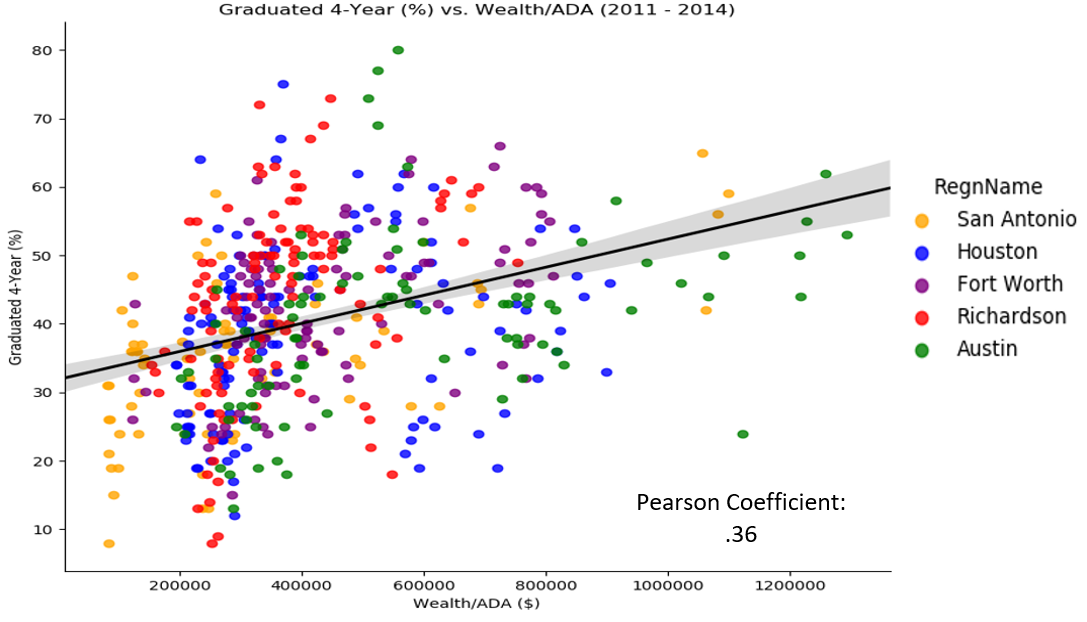


Looks like Richardson (Dallas) contained the highest percentage of students who were able to earn their college degree within four years. There’s roughly an 8% difference between Richardson (Dallas) and San Antonio, which isn’t too much of a surprise as San Antonio, on average, contained poor school district features.

(Wealth/ADA Effect on College Graduation %)

As an exercise, let’s assume that most of the students attending high school in the wealthier districts are simply going to college because they can afford it and there’s a college out there willing to accept them for their money. We could then expect to see a very poor correlation between Wealth/ADA and those who graduate from college within four years.

For the less fortunate areas, the students who attended college most likely earned scholarship. We could assume these students contained a greater chance of earning their degree as they were more prepared, hurting the positive correlation as well. To test our assumptions, let’s take a look at the actual data from the classes of 2011 – 2014 (college graduation year: 2015 – 2018).



The correlation may be lower than the one for college enrollment percentage (pearson coefficient: 0.45), but the difference is not as extreme as one would expect from the assumptions made in the exercise above. As with college enrollment percentage, Wealth/ADA is not the only feature determining the percentage of students earning their degree within four years. We’ve already been presented with several other features that contain their own respective influence on college graduation percentage. These features work together, with some having more influence than others, in predicting our target variable.

It’s also important to consider that our school district features contain some influence on each other (as presented previously in several figures). This could present some multicollinearity issues in building a regression model to predict our target. In the data analysis jupyter notebook included with this project, I extracted several features that I would later explore dropping to improve model performance. These features are a bit redundant or were used to engineer other features. Despite some of these features being dropped in building my final model, they are all included in the complete feature-target dataset as information for those who are curious.

**Machine Learning:**

Note: I recommend the reader follows along with the Machine\_Learning notebook open for code/results and more info on how I ended up with my final model

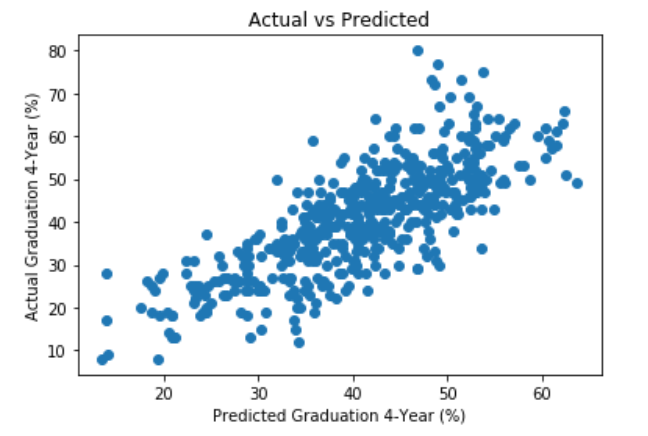
(OLS)

To commence building a predictive model, I trained an ordinary-least-squares (“OLS”) model utilizing training data with all of the collected school district features. In printing out a summary of the model, r-squared (“r2”) was found to be .96 and I also received a warning about multicollinearity as the condition number was large.

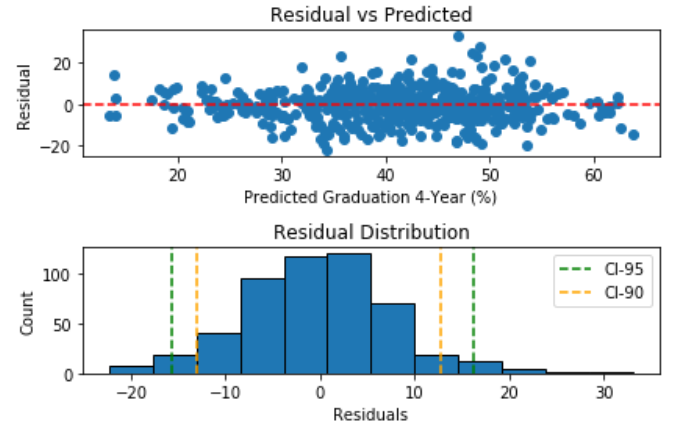
Keeping all other parameters constant, I decided to drop some of the features that were found to cause multicollinearity within my model from the correlation matrix I built in the Data\_Analysis notebook and the base OLS model summary. In training the new model and printing out the updated summary, the r-squared decreased slightly to .959 and the multicollinearity warning disappeared.

I then utilized linear regression in scikit learn to the view resulting root-mean-squared-error (“RMSE”), r2, and mean-absolute-percentage-error (“MAPE”) metrics for evaluating model performance with all features versus dropping the features causing multicollinearity. I used several different testing sizes and 1000 different random states for each testing size. I then took the average metric scores for each testing size for all features and after dropping certain features. I did this to simply eliminate bias for how the data was split up into training and testing sets to make a reasonable conclusion on how dropping certain features affects model performance. In total, I had 508 records available for training my model, so the splits and random states have more influence on the model performance than a model with thousands of records. For all the different testing sizes and random states, I could not reasonably state that dropping certain features resulted in a better performance metrics.

I then proceeded to view performance metrics after training both Lasso and Ridge regression models. Both failed to beat the Linear Regression model’s performance on the testing set. To further view the Linear Regression’s performance, I plotted an actual versus predicted scatter plot for the target variable.



We can see that there’s some variation in the predicted percentages compared to the actual percentages. The correlation between the two is not as strong as I’d like it to be. As another visual, I plotted the resulting residuals with the predicted graduation percentages. The distribution of residuals is also provided below.



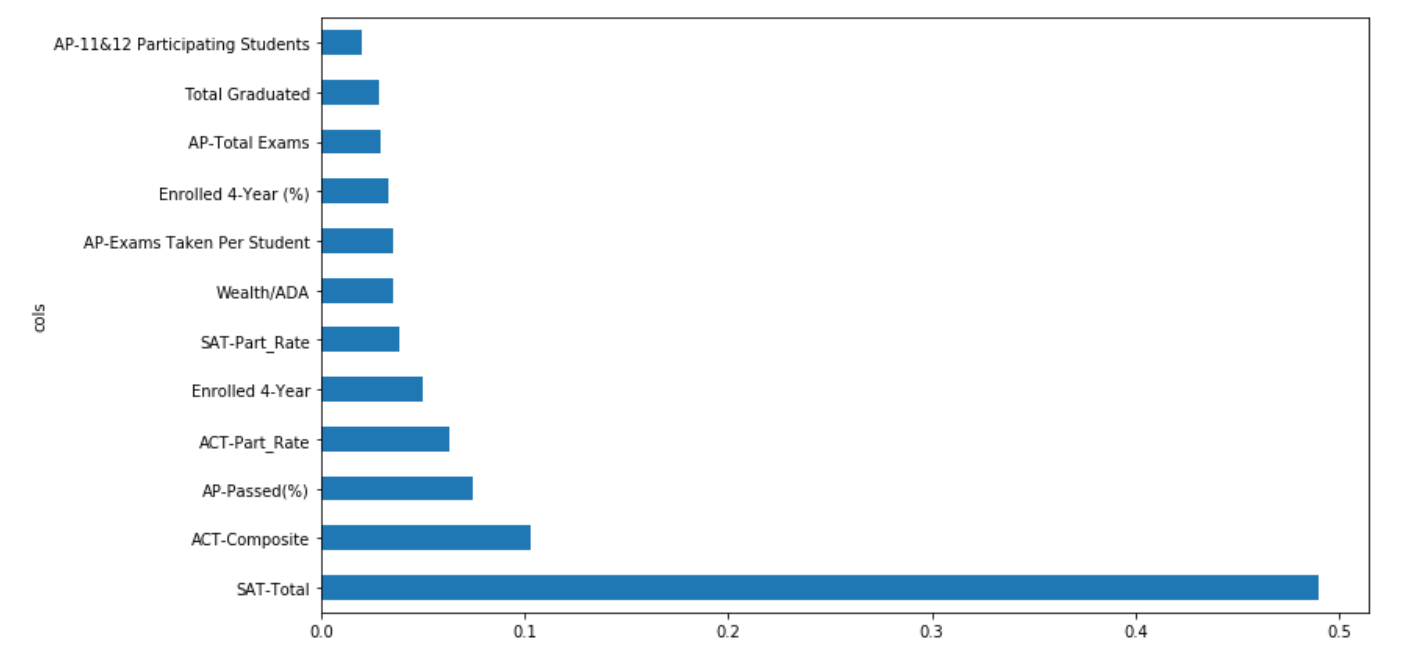
The above served as a base model that I later aimed to improve on. To do so, I chose to utilize a Random Forest Regressor (“RFR”) and evaluate the resulting performance in comparison to the base model. I believed the RFR would help solve my concerns with the size of the dataset when generating splits, adding a further element of randomness to prevent overfitting.

The number of features that can be split on at each node is limited to some percentage of the total, which also helps to ensure that the ensemble model doesn’t rely too heavily on any individual feature. This makes fair use of all potentially predictive features and helps issues of multicollinearity.

In training my first RFR, I did not tune any parameters and used all features. The resulting RMSE, r2, and MAPE on the test set were all comparable to the base Linear Regression model. In evaluating the model, I also printed out the mean r2 and RMSE after performing a 4-fold cross validation.

An out of bag (“OOB”) score was also provided. In the implementation of the RFR algorithm, each tree is trained on roughly 2/3 of the total training set. As the forest is being constructed, each tree can then be tested on the data not used in building that tree. This results in an OOB score that we can use as another comparison metric for model performance.

In previously stating concerns about redundant features that don’t add much value to our model, I thought it would be a good idea to view the resulting feature importance plot after implementing the base RFR model.



After dropping the bottom four features in the above figure, I trained a new RFR to evaluate if there was an improvement in model performance. For RMSE, r2, and MAPE, the model improved. The resulting average r2 and RMSE for the four-fold cross validation also improved slightly.

At this point, I had not yet tuned any hyperparameters, which could help improve the model performance. As a good exercise to gain some insight, I explored the resulting OOB errors from RFRs with different max features (number of features to consider when looking for best split) and number of estimators (trees in the forest).

OOB Error Plot