# Capstone 1 – Milestone Report

**What is the problem you want to solve?**

I would like to predict the percentage of students who will earn a college degree within four years after graduating high school from a certain school district. This problem strictly applies to students graduating high school in the major regions of Texas (Houston, Austin, San Antonio, Richardson (Dallas), Austin) that will attend a Texas college.

**Who is your client and why do they care about this problem?**

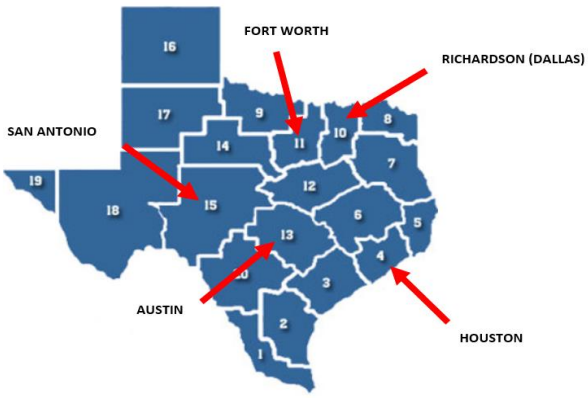
The clients are parents considering a move to or within the major regions of Texas. They'd like for their child to be prepared to earn a college degree and live a more comfortable life. In specifying they want their child to graduate within four years, the clients are looking to avoid a situation in which their child takes more time to earn their degree (more tuition money spent) or even fails out (worst case scenario with minimal return on investment). In strictly focusing on Texas colleges, the clients are also looking to avoid expensive out-of-state tuition.

## Data Wrangling

Before we move further, I find it important to clarify how college enrollment and college graduation percentages are assessed for a particular high school class year. If we say the year is 2013, then the percentage of students who enrolled into college in the fall is denoted as "Enrolled 4-Year (%)" and the number of students enrolling is denoted as "Enrolled 4-Year". The percentage of enrolled students who were able to graduate four years later (2017) is denoted as "Graduated 4-Year (%)", the target variable I attempted to predict.

SAT, ACT, AP exam, and wealth per average daily attendance ("Wealth/ADA") datasets were all downloaded from the Public Education Information Management System ("PEIMS") on the Texas Education Agency's Website [1]. The datasets on college enrollment and college graduation were downloaded from the Texas Public Education Information Resource's ("TPEIR") Website [2]. To give the reader a visual representation of the major regions, I have provided the figure below.

[1] https://tea.texas.gov/
[2] https://www.texaseducationinfo.org/




To clarify for anyone unfamiliar with the American education system, The SAT and ACT are college admission tests, while AP exams offer a student the chance to earn college credit. Wealth/ADA is simply the property value of each school district divided by its average daily attendance. The property value comes from the Texas state comptroller and is the basis for each school district's local property tax collections. 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.

From the SAT and ACT datasets, I extracted the average scores ("SAT-Total", "ACT-Composite") and participation percentages (percentage of available students who took each respective test) for each school district. From the AP exam datasets, I extracted the total number of AP exams taken, and the amount of passing exams (Note: A score of a 3 or above was considered passing), and participation percentage for each respective district. The number of students who graduated high school ("Total Graduated") and the number of students who enrolled into a four-year college that fall ("Enrolled 4-Year") were taken from each of the respective enrollment datasets. Wealth/ADA was straightforward in just providing the figure for all districts.

For each dataset type (SAT, ACT, AP, Enrollment, Wealth/ADA), I sliced the datasets to only include public school districts (those containing "ISD" in the district name) within the major regions. Some of the dataset types required further cleansing or provided room for feature engineering, which I have outlined below. Once the cleansing of every class year was finalized, it was appended to a list to later be concatenated into one total DataFrame (Ex: "Total_SAT") containing all of the data for the classes of 2011 - 2017.

Further Cleansing:

- For the classes of 2011 – 2016, I adjusted "SAT-Total" scores to be equivalent to CollegeBoard's new scoring system out of 1600 (previously out of 2400, new scoring introduced in 2016) using CollegeBoard's concordance tables [3].
- The enrollment datasets contained district names that included an ID number (Ex: 4825170 KATY ISD). The ID number was not necessary, so I got rid of it to leave the district name in all caps. I also needed fix numerical data that was represented as "*" (data not available) or contained a string with a comma (Ex: 1,244). From there I was able to calculate the percentage of graduating high school students who were able to enroll into a four-year college that fall ("Enrolled 4-Year (%)").
- For the AP datasets, numerical approximations in the form of strings (Ex: <60) and some instances of the string with a comma problem were present. I decided to be consistent in decreasing the number by 10% for each of the "less than" cases. I was then able to calculate the average number of AP exams taken per student ("AP-Exams Per Student") and the passing percentage ("AP- Passed (%)") for each district.
- For the enrollment and Wealth/ADA datasets, each district's region was not provided. Using the "Total_SAT" dataset, I performed an inner merge to obtain the respective region names for each school district.

With all the respective DataFrames containing class of 2011 – 2017 data for public school districts in the major regions of Texas, I then merged them all into one DataFrame ("Feature_Target_Data").

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. 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 ("Graduated 4-Year (%)").

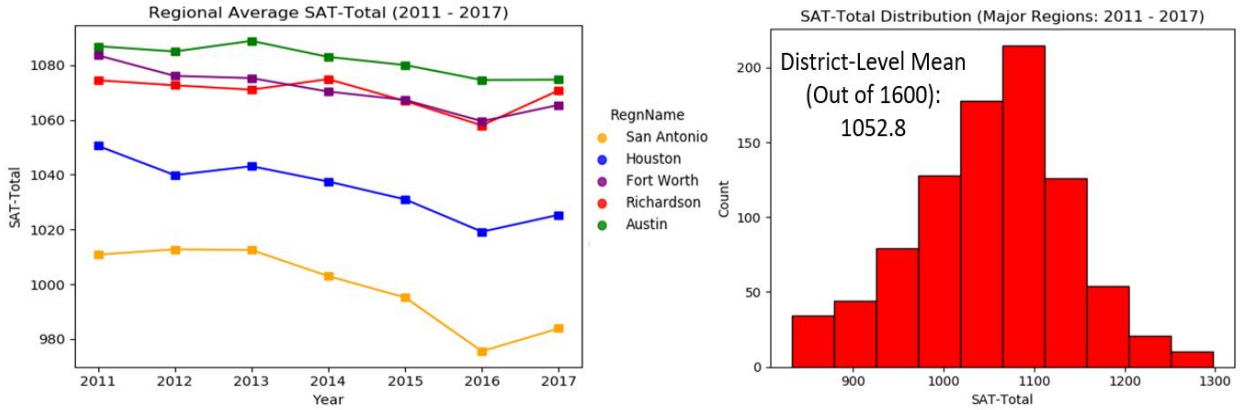
I was then left with a dataset that included all the feature and target data for the classes of 2011 – 2014. For 2015 – 2017, only the feature data was available and the target data (college graduation percentage in 2019 – 2021) was unknown. Later in the report, I will attempt to predict the target for these class years after establishing a satisfactory machine learning model.

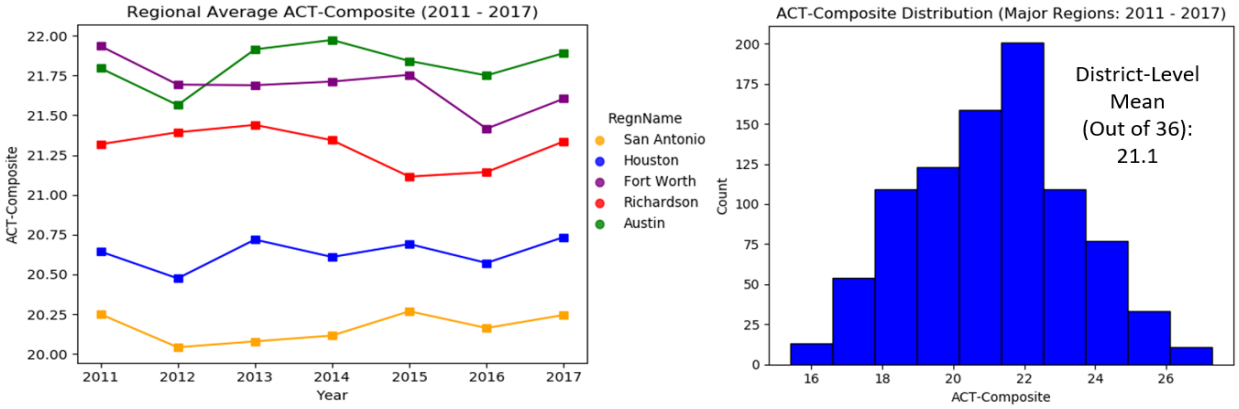
## **Data Storytelling / Inferential Statistics**

*(SAT & ACT)*

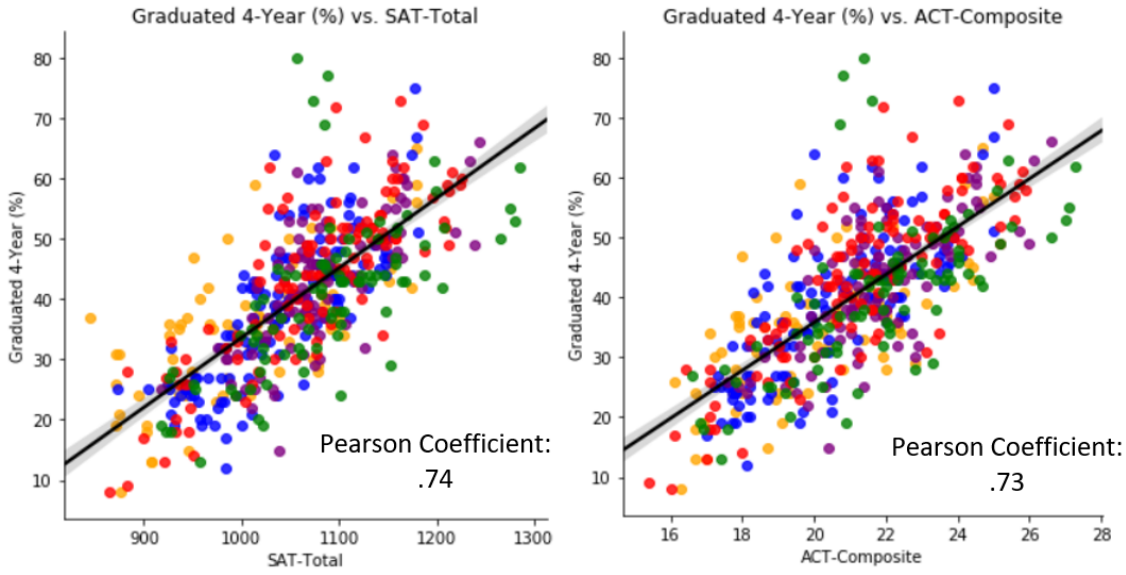
Let's start of by viewing the regional averages for each class year and the distribution of district-level scores for both the SAT and ACT.

[3] https://collegereadiness.collegeboard.org/pdf/higher-ed-brief-sat-concordance.pdf




We can see that Austin performed the best while Fort Worth and Richardson were close (especially for the SAT). The gap between the top region and San Antonio is quite large for both tests, this will remain a common theme for the other features as well. Now that we have seen how the different regions performed, let's see how increasing district-level scores affected the college graduation percentage for the classes of 2011 – 2014.

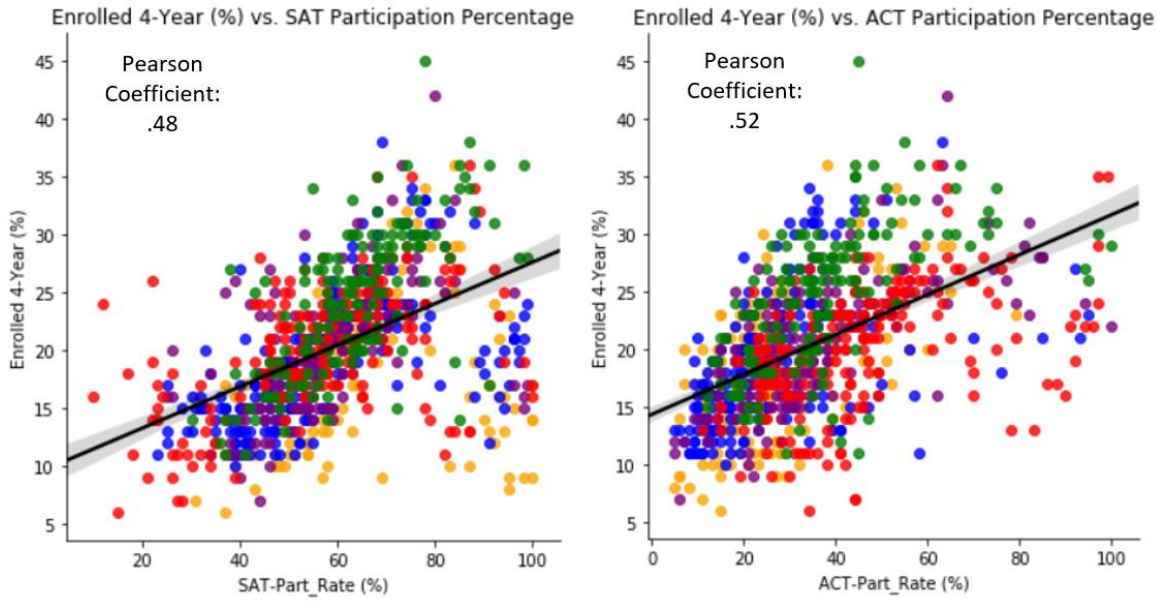


The SAT and ACT both contained a strong positive correlation with college graduation percentage. The correlations make complete sense as colleges use these tests to identify students that they believe are more likely to do well at their institution.

*(SAT/ACT Participation %)*

| Year | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|
| ACT-Part_Rate | 32.007874 | 32.937008 | 32.346457 | 33.771654 | 35.204724 | 37.826772 | 37.409449 |
| SAT-Part_Rate | 61.866142 | 57.771654 | 56.511811 | 58.559055 | 59.039370 | 59.755906 | 62.267717 |

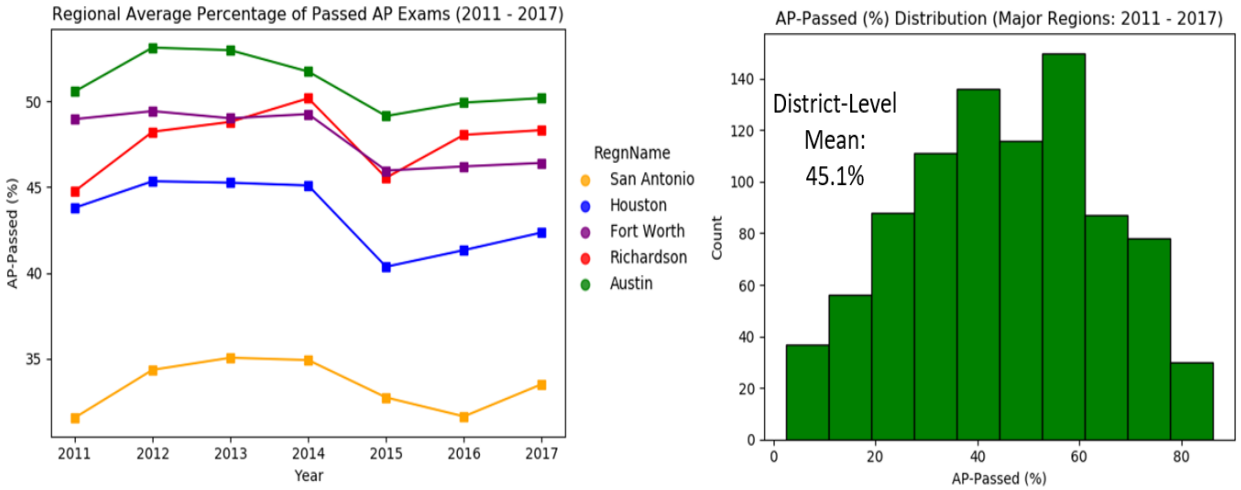
When looking at the classes of 2011 – 2017, I found that a greater percentage of students consistently chose to take the SAT compared to the ACT. Why is this? Well, we could see if choosing to take the SAT historically contained a stronger correlation with college enrollment percentage. If it turns out that this was indeed the case, this trend wouldn't be very surprising. Let's take a look.




Interestingly enough, we see that SAT participation did not contain a stronger correlation with college enrollment than ACT participation. The correlations were actually quite similar with ACT participation even containing a slightly stronger correlation.

As it relates to the client's question/goal, the first step towards earning a college degree is being admitted into college. The above figures indicate that taking both of the college admission tests would be a good idea in helping a student achieve college admission.

*(AP Exams)*

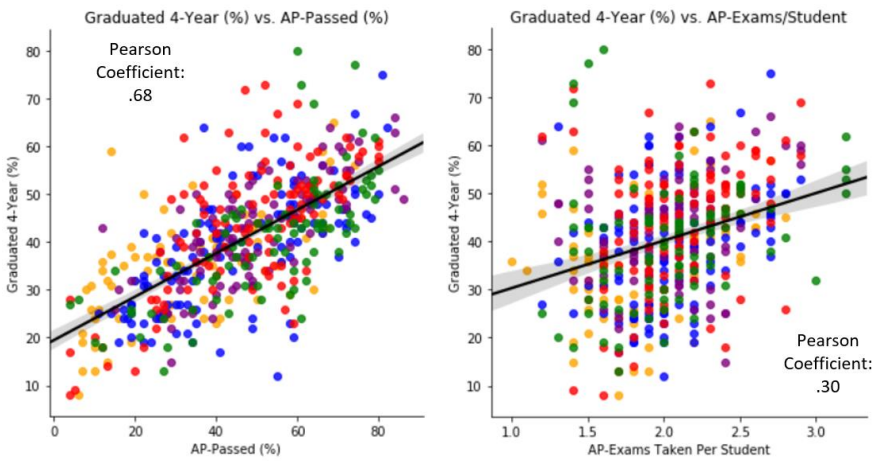



The regions of Austin, Richardson (Dallas), and Fort Worth appear to have contained the best passing percentages. I found it interesting to also take a look at the availability of AP classes to students. One could argue that more availability to AP classes would result in a student being able to take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

**AP-Exams Taken Per Student**

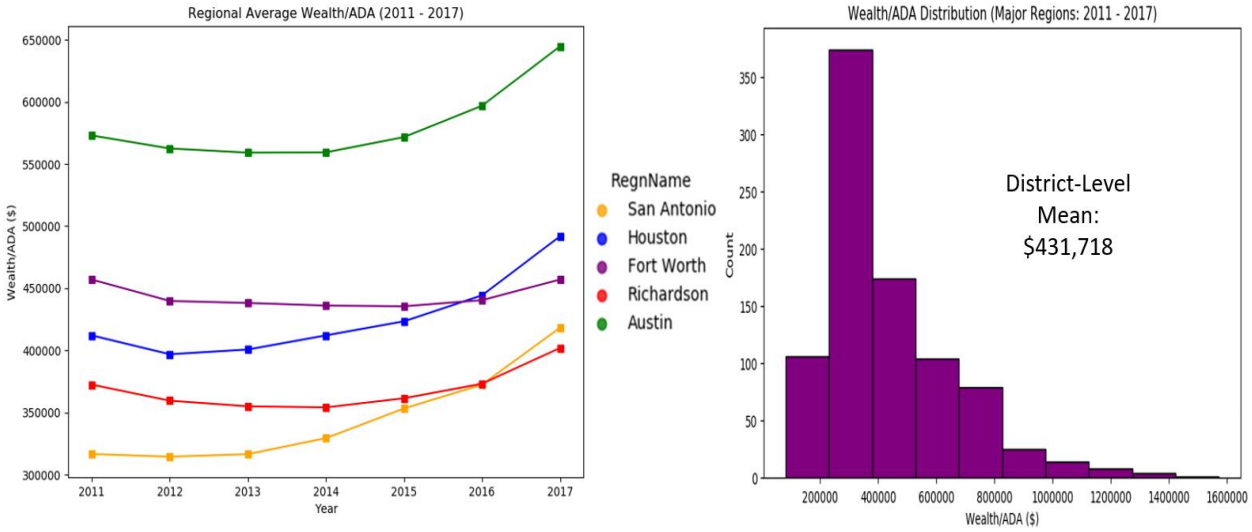
| Year | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|
| **RegnName** | | | | | | | |
| Austin | 2.043478 | 1.990909 | 2.050000 | 2.081818 | 2.200000 | 2.240909 | 2.271429 |
| Fort Worth | 2.065385 | 2.034615 | 2.038462 | 2.084615 | 2.161538 | 2.226923 | 2.238462 |
| Houston | 2.065625 | 2.009375 | 2.043750 | 2.050000 | 2.103125 | 2.178125 | 2.228125 |
| Richardson | 2.035714 | 2.082143 | 2.128571 | 2.142857 | 2.264286 | 2.242857 | 2.307143 |
| San Antonio | 1.811111 | 1.794737 | 1.831579 | 1.842105 | 2.005263 | 1.968421 | 2.105000 |

From the standpoint of AP class/exam availability, students attending high school in Richardson (Dallas) had the most opportunity to gain exposure to college-level courses and earn college credit. For the two AP exam – related features mentioned above, let's view their respective correlations with college graduation percentage.




When logically thinking about the two features, it's no surprise that the passing percentage contained the stronger correlation with college graduation percentage. What good is having access to more AP classes/exams if the student isn't prepared to prove they contain a college-level understanding of the material? Though it contains a weaker correlation with the target variable, the amount of AP exams taken per student still has some importance to the clients as it allows them the opportunity to further save money on college while giving their child more college-level exposure.

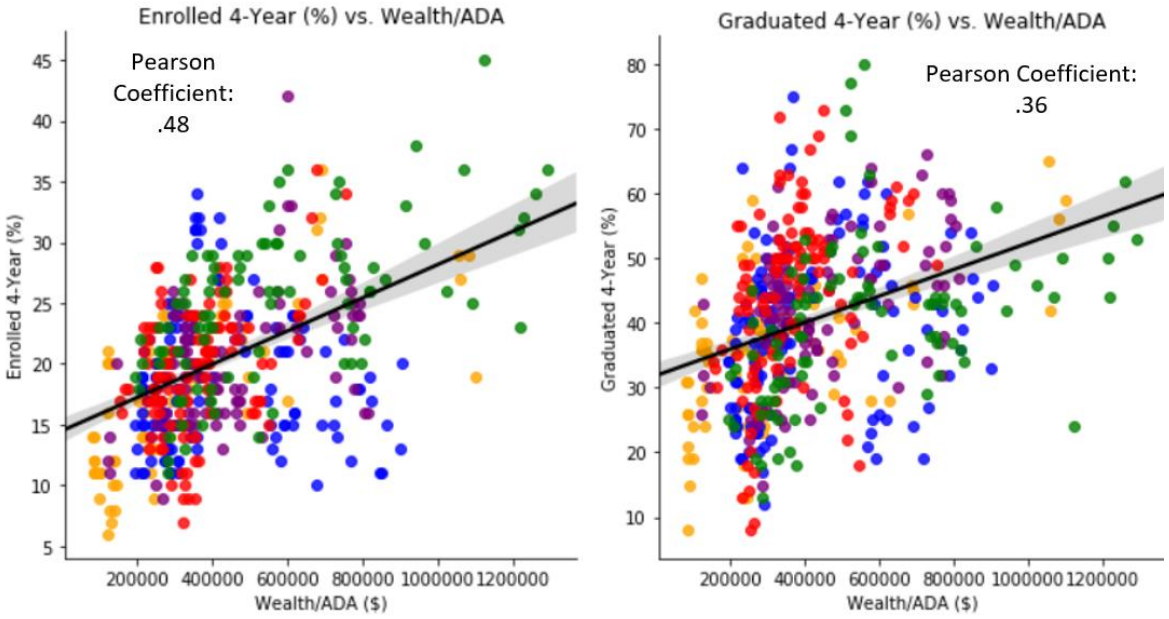
*(Wealth/ADA)*




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.

*(Wealth/ADA's Influence on College Enrollment (%) Versus College Graduation (%))*

Did Wealth/ADA contain a greater positive correlation with college enrollment percentage or college graduation percentage? Let's see what we can extract from the classes of 2011 – 2014.



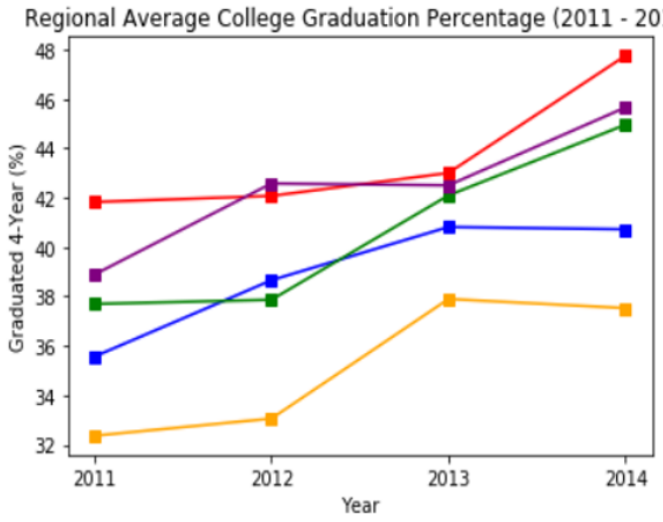

Above, we can see that the pearson correlation between Wealth/ADA and college enrollment percentage was roughly 0.48, indicating a moderate positive relationship between the two variables. Though it certainly isn't the only variable influencing college enrollment percentage, wealth does increase an individual's ability to afford college tuition.

The pearson correlation between Wealth/ADA and college graduation percentage was roughly 0.36, indicating that the wealth of the district a student attends high school in has less influence on college graduation percentage than college enrollment percentage.

When it comes to the more common question among parents about their child getting into college, these statistics tell us that wealth will contribute more to their desired outcome. Maintaining focus on the client's specific question, Wealth/ADA contains less influence in predicting college graduation percentage.

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

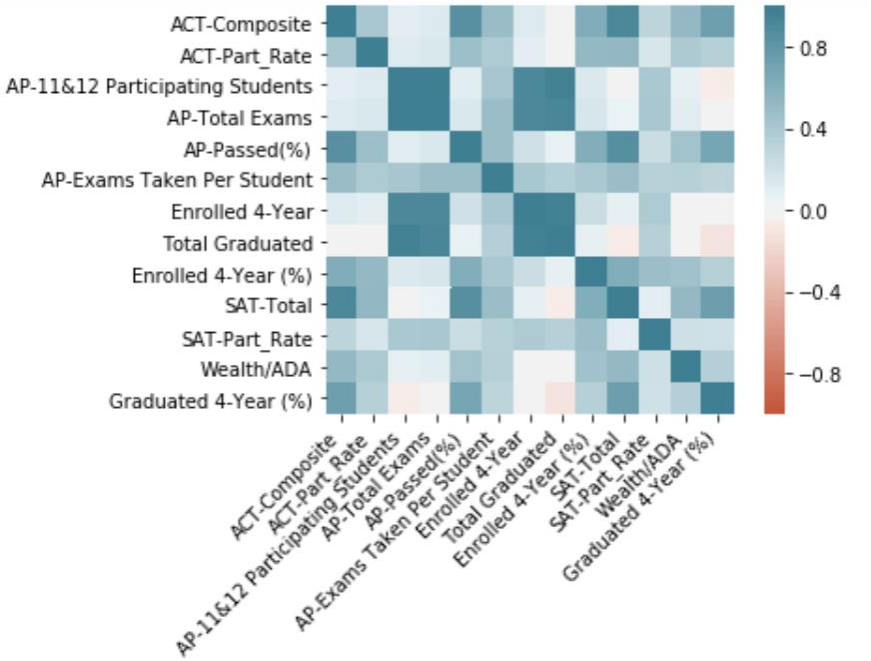
For the classes of 2011 – 2014, let's look at the average percentage of students who were able to earn their college degree within four years by region.




| RegnName | Graduated 4-Year (%) |
|---|---|
| Richardson | 43.660714 |
| Fort Worth | 42.404303 |
| Austin | 40.617418 |
| Houston | 38.937500 |
| San Antonio | 35.245899 |

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.

It's also important to consider that the school district features contain some influence on each other. Some of the features above are a bit redundant as they were used to engineer other features (mentioned in the data wrangling section). This could present some multicollinearity issues in building a regression model to predict our target. Let's explore the respective correlations among all the variables in the next figure.



The features that are prime candidates to be dropped are “Total Graduated”, “AP-11&12 Participating Students”, and “AP-Total Exams”. These features contained slightly negative correlations / no correlation with college graduation percentage and strong correlations with features they helped engineer.