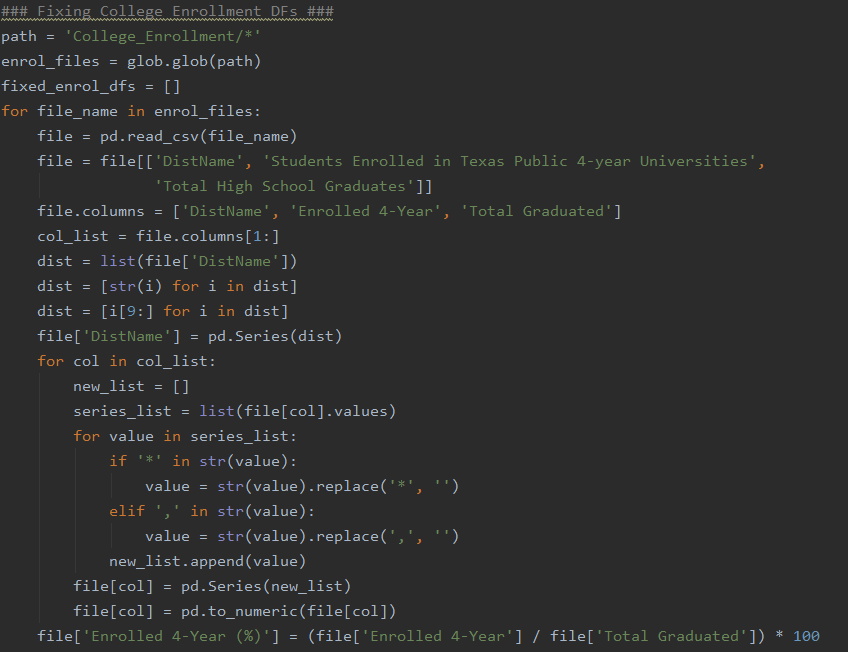
**Data Wrangling**

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. It’s important to note that this data strictly focuses on Texas colleges.

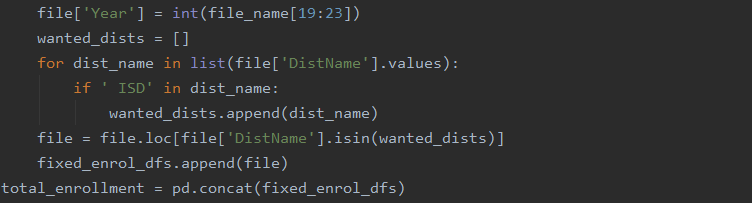
From there I needed to recognize what variables were important in each dataset and remove the unnecessary ones using my knowledge/experience of once being a high schooler in Texas. The ACT, SAT, and AP datasets all contained breakdowns by ethnicity for scores, but I filtered these to simply be “all students” 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 problems here were that district names contained an ID number and name in all caps (Ex: 4825170 KATY ISD). The ID number is not really 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.

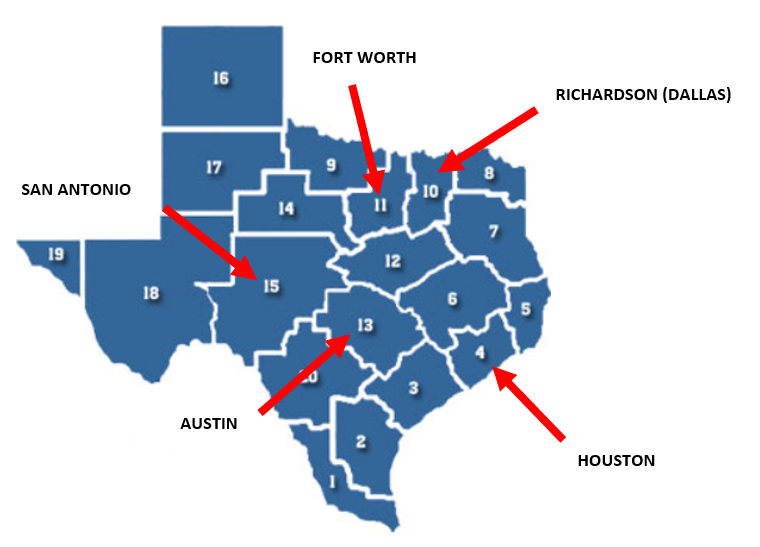


The respective classes were added to each dataset and public districts were extracted using district names that contained “ISD” (Independent School Districts). The datasets included academies/prep schools, but I wanted to strictly focus on public school districts.

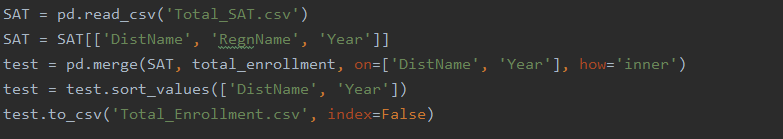


After each enrollment dataset was cleaned, it was appended to a list to later be concatenated into one total (2011 – 2017) DataFrame. To spare the reader from having to look at too much code, this relative process was repeated for each dataset type. The dataset taking the most code to clean was the SAT 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.

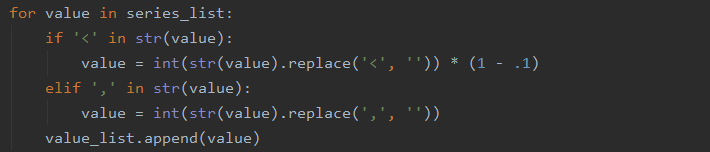
I noticed that the enrollment data also did not contain each district’s respective region name, which was important in meeting my desire to focus on the major regions. To give the reader a visual perspective of the size of the major regions that the districts belong to, I have provided the image below.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



Luckily the SAT datasets contained this data and were utilized to perform an inner merge to provide the enrollment dataset with respective region names for the school districts.



The yearly AP datasets 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.



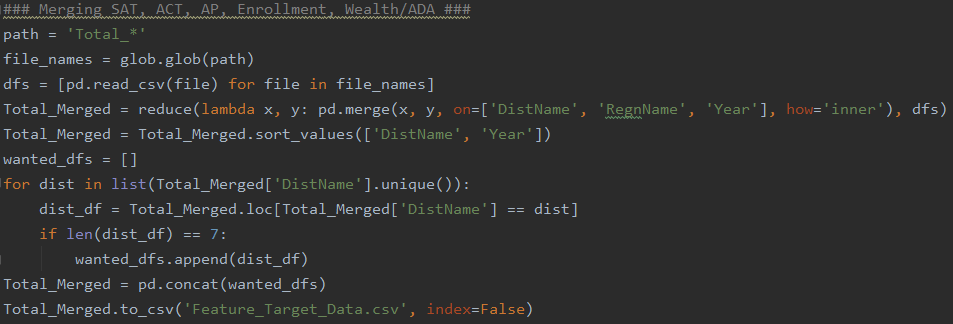
Upon fixing the numerical data, I was then able to add the number of AP exams taken per student in each district for each year.



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.

I was essentially 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 is unknown (what we can try to predict later after establishing a satisfactory model).