Iteration 3

Updates

The only comments we have received after Iteration 2 were concerned with our presentation style. Therefore, we were not able to reflect these comments in the repository.

We however did perform some minor updates to existing code from part 2:

- we finalized getting distance data for each high school
- we found CDS codes for the high schools in our main dataset
- we finalized merging test score data with our main dataset
- we fixed errors in our SAT data preprocessing:
    - Accounting for the SAT Writing test in the years of 2006-2007
    - Properly parsing the CDS code
    - Fixing the format of the data for SAT scores from 2015

Implementation Summary

- /preprocessing.ipynb contains all the same initial preprocessing stuff for the GPA and count data. (All three of us)
- /distances.ipynb uses Google Maps to find distances for each campus to each school, and appends this to previous data. (Nick)
- /TestScores.ipynb merges all the test score data into two data frames and resaves them. (Michal)
- /MergingTestScores.ipynb merges the two test score frames into one and resaves them (Michal)
- /FindingCdsNumbers.ipynb takes the test dataset from the previous notebook and matches school names to CDS numbers, so that we can combine the test score data with the GPA, count, and distance data. Do that, and save the result (Michal)
- /DataExploration.ipynb gives some visualizations for the initial GPA and count data. (Nelson and Nick)
- /Modeling.ipynb does all the splitting, normalizing, and then actual fitting with our linear and baseline models (All three of us contributed to this)

All the data can be found in the /data catalog.