AP Exams & Demographics at NYC Schools

DATA 602 Project Part 2: Specification

Dan Smilowitz

1 Project Statement

I will investigate relationships and correlations between demographics and enrollment in and scores on Advanced Placement (AP) exams for high schools in New York City. I will attempt to see if these relationships are statistically significant and visualize the results.

2 Project Data

Data will come from report from the New York City Department of Education, as hosted on NYC OpenData, specifically:

- Reports detailing annual number of AP exams taken and passed
- Profile of school demographic information

2.1 Data Manipulation

As the two primary data components (test scores and demographics) are located in different data files, they will have to be merged to match demographic attributes to scores for each school. Test score data will have to be converted to determine a pass rate. Additionally, the number of test takers (from one dataset) will have to be compared to total school enrollment (from the other dataset) to understand the portion of students taking AP exams. This manipulation will be done using a combination of base Python and `pandas`.

3 Data Analysis

In order to determine the strength of the relationship between AP test metrics and demographics, analysis of the relationship between AP scores and AP enrollment will be compared to a demographic measures (primarily race and gender). This will include:

- Correlation analysis
- Regression analysis
- Tests for statistical significance

These analyses will be performed using one or both of `NumPy` and `SciPy`.

3.1 Visualization

The analyses outlined in the section above will be presented graphically to show the relationship between variables of interest. Plotting the independent variables (demographic characteristics) vs. the dependent variables (AP test scores and AP class enrollment) will allow the correlation and linear regression calculations to be represented visually. Visualizations will be prepared using `matplotlib`, as well as possibly incorporating `bokeh` or `plotly`.

4 Findings

The key findings from this project will be presented in the following forms:

- The correlation between *dependent variable* and *independent variable* is X
- There *(does/does not)* exist a statistically significant relationship between *dependent variable* and *independent variable*.
  - The equation describing this relationship is given by *[equation]*