1. Exploratory Analysis

New York City’s Public Schools were struggling with overcrowded for some schools while seats available in other schools for years. Therefore, NYC is looking for an improve data models to better predict the future enrollments. Based on this target, we explore the data set includes students’ number in each grade from 2001-2012 school year to 2010-2011 school year first.

To know better about the correlation between numbers of students in each grade each year, we plotted the following graphs.

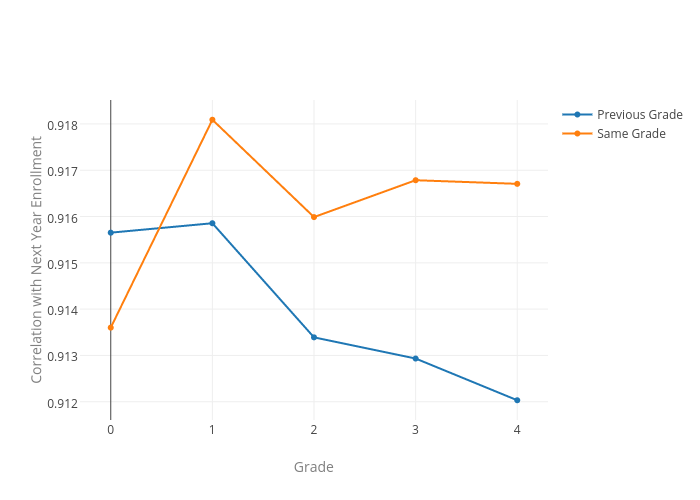
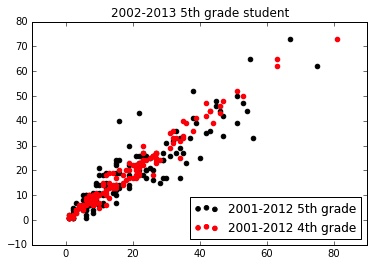


Figure  Figure

The black dots show the correlation between 2002 school year 5th grade student numbers and 5th grade student numbers in the previous year, while the red dots show the correlation between 2002 school year 5th grade student numbers and 4th grade in the previous year. In this case, as we can see, the 5th grade student in 2002 may more correlated to the lower grade student in the year before. This is also make a lot of sense, since usually when students finish lower grade study they will head to a higher grade study. However, this is not always the case.

As we can see from the Figure 2, if we sum up the data from all census tracts, we may carefully draw the conclusion that for most grades, same grade in this and the next year seems more correlated. This also sets up a frame of data we are going to use to predict different grade student number.

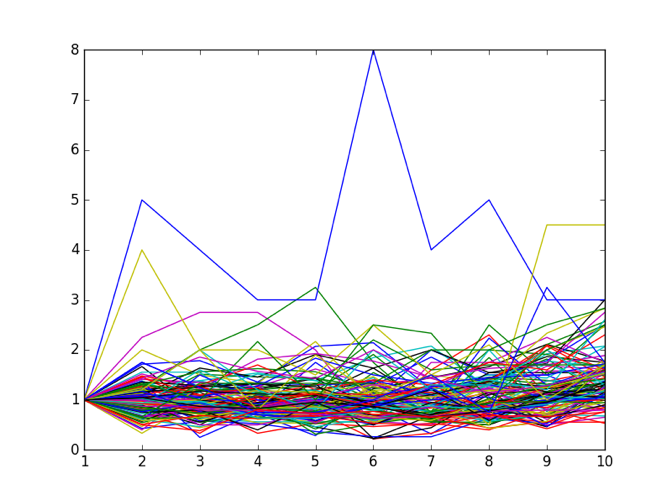
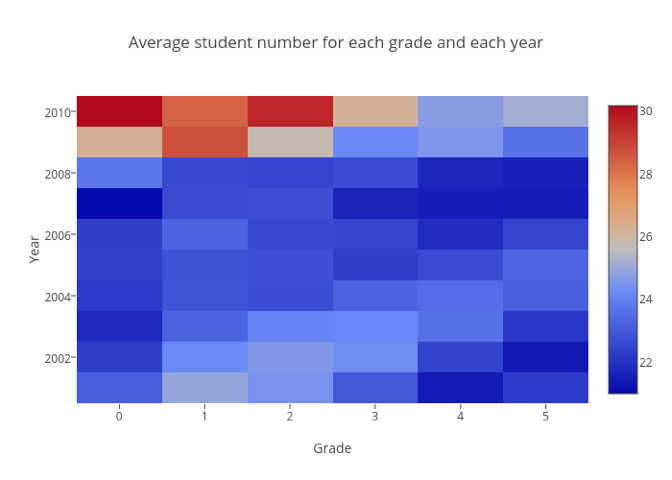


Figure Figure

The heat map in Figure 3 shows the change of student numbers in a clearer way. As we can see, the average number of new students for kindergarten increases from 2001-2012 to 2010-2011. Figure 4 shows the ratio of student number in this year to the previous year in different tracts. For most census tracts, the ratios are quite stable around 1 which means the number of students didn’t change to much. For those very unstable tracts, we will consider to deal with separately.