Problem Statement

Public schools are state and federally funded by pupil enrollment counts. Each school year the enrollment counts differs from the previous years enrollment count. However, staffing and logistics for the school need to be determined at least 6 months before the start of the school year. It would be significantly beneficial for school district leadership to be able to predict the enrollment counts for a particular school/school district in California for the next school year. Historical public school enrollment data is publicly available from 1981 - 2018 for California School districts by school. The data available is disaggregated by school, ethnicity, gender and grade level. It is available at the California Department of education at https://www.cde.ca.gov/ds/

Hypothesis

It has been a general observation in California bay area school districts that the enrollment count has been dropping. In San Jose Unified School District the enrollment count had dropped by 10% from 2016 to 2017. It was the general understanding that due to increase in rental and home prices, some of the population had migrated to other neighborhoods. I would like to pursue the hypothesis that population migration into and out of a neighborhood is a factor that influences the enrollment count.

There is another data set also available at the California Department of Education - which contains data about the Free and Reduced Meal Program(FRMP). Students eligible for the FRMP come from households below a certain income level.

By analysing this data one can get a sense of the income distribution of households in the school district neighborhood at 3 income levels - i) qualifies for the free lunch ii) qualifies for the reduced lunch and iii) does not qualify for the FRMP

It would be interesting to analyse the rental and housing prices in the neighborhoods. I have to find a public data source for this yet.

Assessment performance data is available also on the California Department of Education. Perhaps assessment data can be analysed to see if elementary middle school performance determines the high school enrollment count. This data is available from the year 1998-2013.

Approach to problem solving

1. Observe the enrollment data - performing Exploratory Data Analysis. The data for all years is not in the same format - so I expect some data munging cleanup tasks to prepare the data for analysis.
2. Superimpose the Free and Reduced Meal Plan data to further disaggregate the data by income level.
3. Record observations. Prepare a report of findings.
4. Perform appropriate inferential statistics and machine learning methods to project enrollment count for next year. May need to come up with other ideas for what influences enrollment count.

Deliverables

1. All code
2. A summary report of all findings with meaning visualizations
3. A clickable map of all school districts in California with a report of projected enrollment counts (It's a maybe - would be nice to learn how to code that)