**PROBLEM STATEMENT**

To understand whether the density of selected types of businesses in and surrounding neighborhood contributes to home price appreciation for that neighborhood. When looking for a home to either buy or rent, prospective residents want to be able to project either the growth of the equity they’ve put into their homes or rents they can expect to pay in the future. A tour of the neighborhood and assessment of surrounding shops and businesses might be factors in price appreciation and would be something prospective buyers and renters could easily observe on their own.

My hypothesis is that businesses such as cafes, bakeries, and fitness centers would be positively correlated with home price appreciation, while big box department stores could have a dampening effect, and certain businesses with relatively low foot traffic such as dentists would have little to no effect.

The primary target variable will be home prices. Rental prices could also be explored as a second target variable.

**OUTLINE OF POTENTIAL METHODS AND MODELS**

To begin, exploratory analysis will be conducted on the data to understand the relative concentrations of these businesses, and any correlation between individual business types and home price trends. Correlation between business types should also be considered. Linear regression may be a suitable model for this, and logistic regression is also a possibility if density of certain business types (such as whether there is a department store within the zip code) follows a more binary pattern.

**EXPLANATION OF AVAILABLE DATA**

Home values data exists from multiple sources. Zillow maintains and makes publicly available a database of historical home sales aggregated by geography.

Information on business storefronts can also be found on a number of platforms which makes their data available to the public via APIs. For this project Yelp’s database will be used to find the concentration of specific business types in or in the vicinity of each zip code used for the research. Due to the very large data set, and restrictions set on the amount of data these APIs allow users to extract, this project will likely extract data for and focus on specific geographic region, such as the Bay Area.

Update: Due to individual zip codes having varying boundary areas and populations, there may be further need to normalize each zip code (i.e. data record) for each data set according to some sizing factor (possibly area or population count).

**OUTSTANDING QUESTIONS, ASSUMPTIONS, RISKS, CAVEATS**

* Data for some zip codes could be incomplete, either due to availability from the source, or even historical existence of a given zip code. Therefore, some zip codes may be excluded from the research.
* Some smaller zip codes, may have relatively few home sales for a given month, which could skew month-to-month pricing data (and to a degree, also business density).
* Some zip codes may be primarily commercial instead of residential. This could play a role in how the residential portions of the community are priced, but generally speaking, this is not factored into the analysis.
* Of course, only specific business types are considered for this analysis. Due to the very large number of business categories which are available and the limitations in extracting this information, the business types considered for this analysis has been pre-chosen and are not necessarily the ones which will be most insightful.
* Due to the difficulty in obtaining information of individual store openings and closings, this analysis will likely assume that business counts are not time-dependent – that is, there will likely be an assumption that businesses were active for the duration of the research period.

**DEMONSTRATE DOMAIN KNOWLEDGE**

This research was partially inspired by a [blog/article published by Zillow](http://www.zillow.com/blog/starbucks-home-values-170734/) which looked at real estate markets affected by the presence of a Starbucks café location, which confirmed a personal intuition.

**DEFINE GOALS AND CRITERIA, TO EXPLAIN WHAT SUCCESS LOOKS LIKE**

The initial goal will be to set out to construct a well-fitting model to explain home value trends over time. However if in fact there is no correlation (or an unexpected correlation) between the chosen factors, that would also be considered a success for the purpose of this analysis.