**Capstone Project**

Arthur: Wilson Lau Date: 2015 November

**Capstone Milestone Report**

**Introduction**[**¶**](http://localhost:8888/notebooks/DataStoryOne/DataStoryOneHousePriceAnalysis.ipynb#Introduction)

The problem I want to tackle in my capstone project is to detect investment opportunities for cash flow real estate property.

**The Problem**

Real estate investment is a passive income opportunity for investors to accumulate return without active management. Huge long term appreciation is very much possible for real estate. However, appreciation often happens in short period of time and the appreciation is heavily affected local market condition. It is difficult for out-of-the-market investors to detect the trend and acquire properties before price goes up. My capstone project is trying to help investors to detect real estate investment opportunities and be able to predict which zip code will experience long term appreciation in the near future.

My project is divided into two stages:

**Stage One**

* the first stage is to use anomaly detection technique to discover which zipcodes have success sprike of upward price appreciation.
* The Anomaly detection is compared new monthly updated home price with historical price as well as comparing with the home price of the surrending areas based on neighborhoods, city, menpolian area, and state.

**Stage Two**

* Based on the result of stage one, more data will be collected for the zipcodes that have sudden increase of home price.
* Those new dataset will be used to figure out what feature contributes to the price appreciation and a predictive model will be created using those features.

**Dateset**

Dataset used is downloaded from Quandl's Zillow data. This dataset is a monthly time series recording the change of home price.

**What important fields and information does the data set have?**

The important columns for this dataset is time(date end of the month), home price and zipcode.

**What are its limitations i.e. what are some questions that you cannot answer with this data set?**

This dataset actually have 22 columns. For example, it has columns for different home size e.g. 1 bedroom, 2 bedroom, 3 bedroom, price per square foot, average listing price, etc... Since it seems that most of these columns are highly corelate to home price, they may be very help for predicting home price.

**What kind of cleaning and wrangling did you need to do?**

Quandl's API is very powerful and very flexible. At the same time, it is a bit complicated to pull the home price. One way to pull home price is to pass zipcode as input parameter, and Quandl will return the historical price data with all the columns related to that zipcode. So, to get the data, I will need to use a zipcode dataset that contains all the zipcode of the US and then use the Quandl API to pull data one zipcode at a time.

**Are there other datasets you can find, use and combine with, to answer the questions that matter?**

To order to group home price data by city, state, etc.., I will need to find other dataset that relates individual zipcodes with their city name and state name. I will need to merge the dataset together in order to find out which zipcode belongs to which city or state.