Correlation Analysis between Property/Rental Values and the Gold Line Service Expansion

Eric Pan, Phuong Ngo, Shayne Hu

Problem Definition

For this project, a total number of two datasets and other sources of information are used to carry out real estate analysis and visualization. These datasets, which comprise of house listing price and rental price will be explained later, are two time series datasets collected by realtor.com and zillow.com respectively. They span over the years of 2012 to 2019. Hence, from a data science standpoint, we will be carrying out time series analysis in order to extract meaningful statistics and other characteristics of our rental and housing price datasets.

From a business perspective, information extracted from the datasets will be used to help us determine whether there is a correlation between the expansion of the Metro Gold Line and the price of housing and rental in the nearby areas. This means that we will turn to data mining and analysis to extract some kind of pattern from the datasets over the years and compare them to certain time stamps (announcement dates and completion dates of the Gold Line) [1] to prove our stated hypothesis.

Description of Background

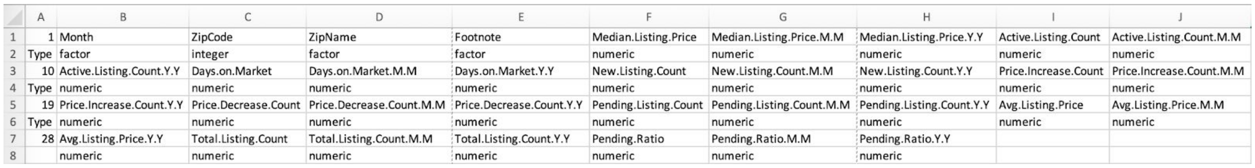
In machine learning and statistics, time series analysis helps deal with data so that we can identify the characteristics or nature of a phenomenon represented by sequences of observations and in turn help forecast or predict future values of the time table [2]. Once a pattern is identified, we can interpret and integrate with other data to use it in supporting our hypothesis or investigation of a phenomenon, which in this case is the correlation between house and rental prices and the expansion of the Metro Gold Line.

With this project, we would like to see if our hypothesis about the Gold Line and real estate prices in nearby area is true. There are a lot more factors that can affect housing and rental prices but we want to only focus on the expansion of transportation and how big of an effect it has on the price increase.

As Los Angeles remains one of the most popular cities in the United States with people moving to the city for work and school, looking for a reasonable place to live remains one of the main concerns for people who just moved to Los Angeles or want to relocate to the city to be closer to work. A lot of people are willing to commute to work in exchange for cheaper rent. Freeway and metro expansions allow that to happen, but they can also be a factor that makes a neighborhood more desirable, hence our investigation into how big of a role the expansion is towards housing and rental prices.

Description of Dataset

There are two major datasets we are going to use. First is house listing price history dataset which is accessible at realtor.com [3]. It includes all history housing data which are based on zip code across the US from 2012 May to 2019 April. This dataset has 1259408 rows and each row is a monthly record of 34 variables in a zip code area (Record date and zip code together form the row key).



	A	B	C	D	E	F	G	H	I	J
1	1	Month	ZipCode	ZipName	Footnote	Median.Listing.Price	Median.Listing.Price.M.M	Median.Listing.Price.Y.Y	Active.Listing.Count	Active.Listing.Count.M.M
2	Type	factor	integer	factor	factor	numeric	numeric	numeric	numeric	numeric
3	10	Active.Listing.Count.Y.Y	Days.on.Market	Days.on.Market.M.M	Days.on.Market.Y.Y	New.Listing.Count	New.Listing.Count.M.M	New.Listing.Count.Y.Y	Price.Increase.Count	Price.Increase.Count.M.M
4	Type	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric
5	19	Price.Increase.Count.Y.Y	Price.Decrease.Count	Price.Decrease.Count.M.M	Price.Decrease.Count.Y.Y	Pending.Listing.Count	Pending.Listing.Count.M.M	Pending.Listing.Count.Y.Y	Avg.Listing.Price	Avg.Listing.Price.M.M
6	Type	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric
7	28	Avg.Listing.Price.Y.Y	Total.Listing.Count	Total.Listing.Count.M.M	Total.Listing.Count.Y.Y	Pending.Ratio	Pending.Ratio.M.M	Pending.Ratio.Y.Y		
8		numeric	numeric	numeric	numeric	numeric	numeric	numeric		

Table 1. Summary of 34 variables in realtor.com housing history data

The second dataset provides only rental price data from 2010 September to 2019 September which is accessible at zillow.com [4]. It has 32218 rows and 116 columns. Each row uses RegionName/zip code as row key and the last 109 columns in each row contain the monthly rental price.

Because both datasets does not cover the data back to 2009 October, time when the expansion was put on the agenda, that means we have to analyze the influences brought by the completion of expansion. After preliminary visualizations on different variables in the house listing dataset, we find the median price best represents the trend, so we only extract median price in 21 zip code areas (would be discussed later) from 2014 January to 2018 April (2 years before the completion and 2 years after the completion) from the raw dataset. Similarly, we extract data in the same time frame from the rental price dataset. Both extracted datasets have very high quality (no need to treat missing values), now we are good to start our analysis on these two processed datasets.

Description of Methods Used

Before we start our analysis on the data, we first assume a simplified model that the housing cost in one place, for example at a certain zip code, will be increasing gradually over time. In other words, the average increasing rate over a relatively long period of time is constant. So, if we find a difference between the average increase rates in two periods of time with the same length at the same location, we may come to a conclusion that something happens between those two periods and results in the difference. Besides this, we also need a set of control data so that we can exclude the influence brought by factors we are not interested in, like the cut of property taxes. In our case, we pick 11 zip codes which are in the expansion area of gold line, 91107, 91775, 91006, 91024, 91016, 91706, 91010, 91702,91722, 91741 as our training group. And we randomly pick 10 zip codes, 90064, 90249, 90808, 90630, 90621, 90631, 92805, 90068, 91506 and 91762 around LA area (but ensuring that these areas are far away from the gold line extension) as the control group. The location of zip code in both training and control group is shown below.

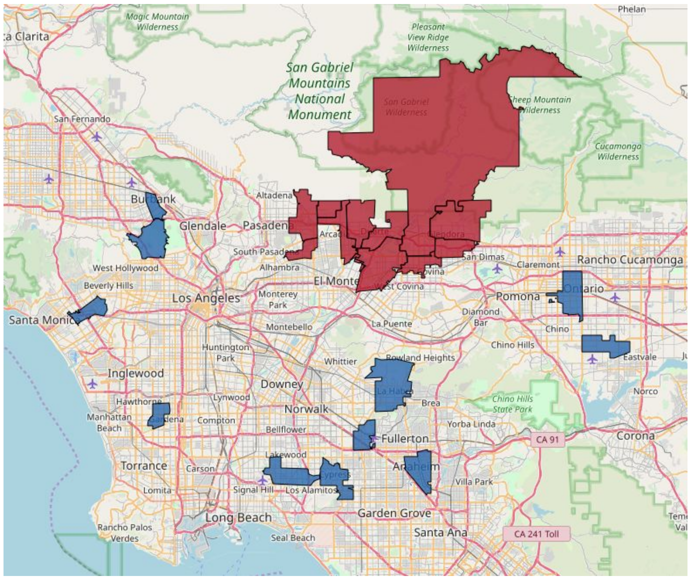


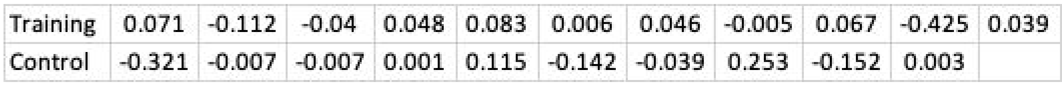
Figure 1. Red for training group and blue for control group

After setting 2016 March as the expansion milestone (date when Phase 2A extension entering service) as the period delimiter, in housing price dataset, we calculated the increasing rate from 2014 Jan to 2016 Jan (before milestone) and 2016 April to 2018 April (after milestone) in both training group and control group.

For pattern reading and analysis, we incorporate Python and Matplotlib.Pyplot package to plot our time series by features for training house listing price data and training rental price data. These features are already extracted for us by realtor.com, and they are median and average of the time series. By doing this, we hope to identify a hike in housing and rental prices and any patterns that come with it.

Experiment: Analysis Results

In training group of 11 zip code areas, the difference between the increasing rate before and after milestone are shown below.



Training	0.071	-0.112	-0.04	0.048	0.083	0.006	0.046	-0.005	0.067	-0.425	0.039
Control	-0.321	-0.007	-0.007	0.001	0.115	-0.142	-0.039	0.253	-0.152	0.003	

Table 2. Difference of increasing rate in house listing price dataset

We notice that there are outliers in the table, so we apply the [1st Quartile - 1.5 IQR, 3rd Quartile + 1.5 Quartile] model to remove outliers which are -0.425 in the first group and -0.321 and 0.253 in the second group. After this, we get the box plot of the two sets of data.

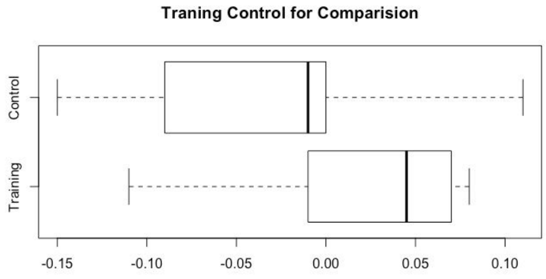


Figure 2. Box plot for training group and control group in housing cost dataset

Obviously, in the training data, the difference in rate is larger than that in the control group, which means the increasing rate becomes larger in training group after the gold line expansion.

In order to make the increasing rate difference more intuitive as well as to make it possible to feed it in Choropleth maps in folium library (this solution does not take negative values) [5], we employ an index conversion on the raw difference of increasing rate. In details, we add 0.16 to all the difference of rate to remove negative numbers, multiply the outcome by 100 then take the ceiling. So, according to the conversion principle, the larger the index, the more rapid growth of the housing cost in a certain area.

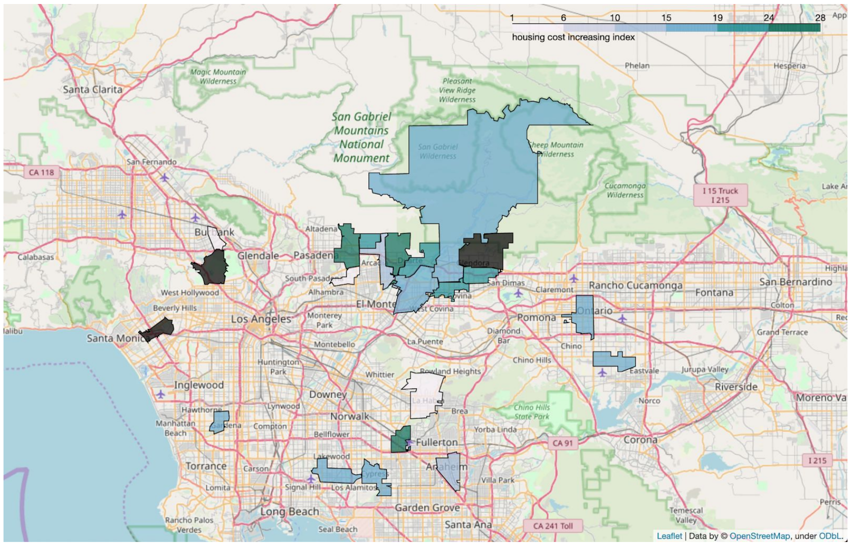
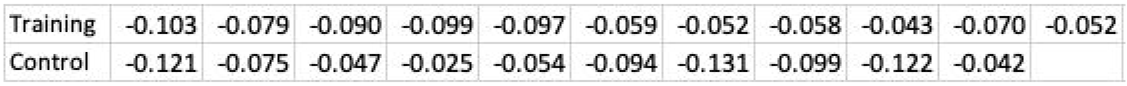


Figure 3. Housing cost increasing index visualization (Larger index means larger increasing rate difference and areas with black are outliers excluded in the previous step)

In the figure above, we can see that areas around the gold line expansion project have a larger rise in increasing rate after 2016 Mar.

Similarly, we then employ the full analysis procedure discussed above on the other dataset, the rental cost data retrieved from zillow.com. The raw difference in both groups is shown in the table below.



Training	-0.103	-0.079	-0.090	-0.099	-0.097	-0.059	-0.052	-0.058	-0.043	-0.070	-0.052
Control	-0.121	-0.075	-0.047	-0.025	-0.054	-0.094	-0.131	-0.099	-0.122	-0.042	

Table 3. Difference of increasing rate in rental dataset

The box plot for these two sets of data is shown below. From it, we can see that there is no outstanding discrepancy between the two groups in rental cost rise.

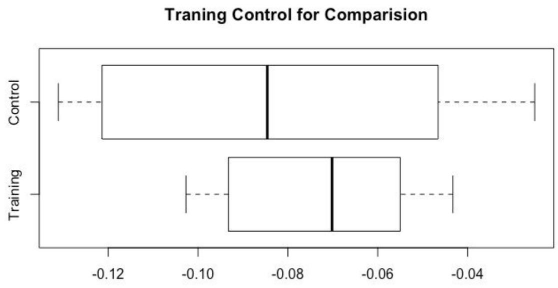


Figure 4. Box plot for training group and control group in rental cost dataset

The visualization of increasing rate index is shown below. Consistent with the result obtained from box plot, the geographical visualization of increasing rate index can not separate training group and control group clearly.

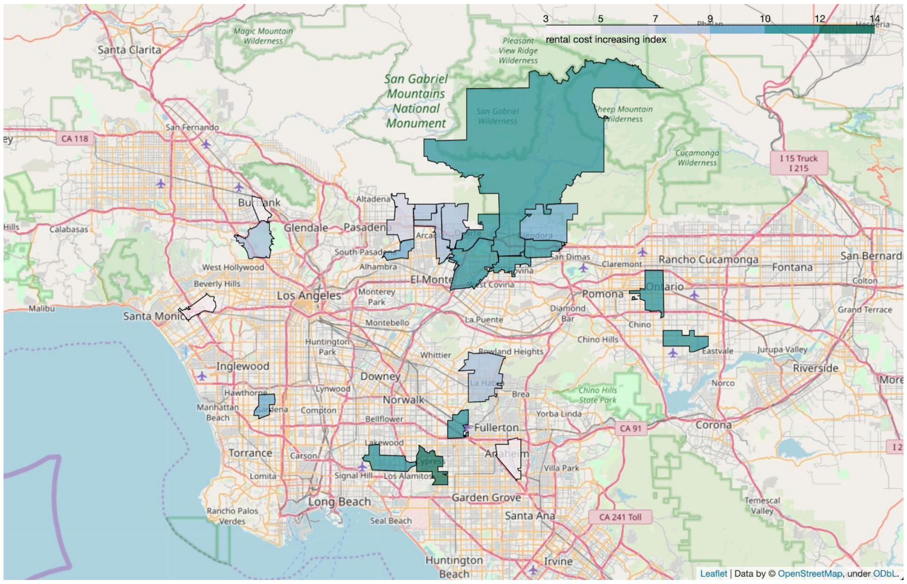

3
5
7
9
10
12
14
rental cost increasing index
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 5. Rental cost increasing index visualization (Larger index means larger increasing rate difference)

Observation and Conclusion

Based on our analysis, we can get to the conclusion that the expansion of metro gold line does stimulate the increase of house prices in its serving area to some extent. On the other hand, the expansion project does not influence the house rental market in the same area. Oppositely, the rental market within our investigating area experienced a slight decline in the increasing rate after the completion of the expansion.

Then we make an induction that, from an investment or self-use perspective, it may be a good idea to purchase a property before a new metro line being operational, as the price of property would increase faster later. But we do not recommend that purchasing a property just for renting it out. The rental market does not respond positively to the metro line expansion, so the return on investment will be lower than expected.

Other than the programmatic observations and conclusions above, another interesting finding which could be the topic of future analysis is that although the rental market in the gold line newly serving areas experienced a retreat after 2016 April according to our analysis, areas near the end of the expansion route present a relatively low decreasing trend. So our question is: Is there an attribute or a set of attributes is accountable for this phenomenon and how can we proceed to analyze it?

References:

1. The Gold Line Foothill Extension, https://en.wikipedia.org/wiki/Gold_Line_Foothill_Extension
2. http://www.statsoft.com/textbook/time-series-analysis#main
3. All residential historical data based on zip code, https://www.realtor.com/research/data/
4. ZRI Time Series: Multifamily, SFR, Condo/Co-op based on zip code, https://www.zillow.com/research/data/
5. Visualizing Data at the ZIP Code Level with Folium, Finn Qiao, https://towardsdatascience.com/visualizing-data-at-the-zip-code-level-with-folium-d07ac983db20