IST718 Big Data Analytics - Lab 2

Martin Alonso - 2019-02-23

Introduction

The objective of this project is provide a recommendation as to the best zip code in which to invest in based on the change in average household value. In order to achieve this, the average household value of 15,508 zip codes will be analyzed using time-series analysis, and a model will be built to predict how each zip code will change over the next 12-month period. The end result will provide investment recommendations on four zip codes based on risk-reward, where risk will be measured on the robustness of the model using Mean Absolute Error and reward will be based on the percentage change in value from the last point in the time-series to the last point in the forecasted data.

Along with the end result, the data will also be used to answer the following questions:

- Which Arkansas metropolitan area grew the most and would provide a suitable investment opportunity.
- What technique was used to down sample the data.
- What three zip codes provide the best investment opportunities.

Analysis

For this exercise, the data set that will be worked upon consists of 15,508 observations with 280 variables. Each observation consists of one zip code in the United States; for each zip code, there is a Region ID tag, the Region Name (or zip code), City, State, Metro area, County, and size of the zipcode. After this initial identification variables, the data has 273 additional variables, each consisting of the average household value for the zip code for each month between April, 1996 and December, 2018. Table 1 presents the first 10 variables and 5 observations of the data set.

Table 1

Region ID	Region Name	City	State	Metro	County Name	SizeRank	1996-04	1996-05	1996-06
84654	60657	Chicago	IL	Chicago-Naperville-Elgin	Cook County	1	334200	335400	336500
91982	77494	Katy	TX	Houston-The Woodlands-Sugar Land	Harris County	2	210400	212200	212200
84616	60614	Chicago	IL	Chicago-Naperville-Elgin	Cook County	3	498100	500900	503100
93144	79936	El Paso	TX	El Paso	El Paso County	4	77300	77300	77300
91940	77449	Katy	TX	Houston-The Woodlands-Sugar Land	Harris County	5	95400	95600	95800

Along with this data set, we've also added a second data set from the U.S. Department of Labor, which provides zip codes for the entire U.S., along with their city, state, metropolitan area. This data set was added because the Metro column in the original data set was incomplete. Similarly, there were many household value variables missing between 1996 and 2001. To account for these missing variables, we took the average household value for the metropolitan areas each zip code belonged to. Once these missing data were replaced, we proceeded to restructure the data.

Because of the way the data set is structured, it is very difficult to analyze. However, using Hadley Wickham's *Tidy Data* paper (2012), the data was restructured in such a way that each household value variable is now an observation, spread out for each zip code. Table 2 shows the first five rows and 9 columns of the restructured dataframe.

Table 2

date	RegionID	Region Name	City	State	Metro	County Name	SizeRank	Value
1996-04-01	58196	1001	Agawam	MA	Springfield	Hampden County	5957	113100
1996-05-01	58196	1001	Agawam	MA	Springfield	Hampden County	5957	112800
1996-06-01	58196	1001	Agawam	MA	Springfield	Hampden County	5957	112600
1996-07-01	58196	1001	Agawam	MA	Springfield	Hampden County	5957	112300
1996-08-01	58196	1001	Agaam	MA	Springfield	Hampden County	5957	112100

Transforming the data set into a tidy data set allows for much easier exploration and analysis. For starters, we now have a clearer idea that these 15,508 zip codes are spread out across 7,957 cities in 50 states and the District of Columbia which, for this data set, has been considered a separate state, giving the data a total of 51 states.

Given that the data set is a time-series data set, we want to know how household valuation changes month-to-month. The data was first grouped by nationally, meaning we had one single observation for each month. The data was then restricted to show change over time between 2009 and 2018. As seen in figure 1, average household value was falling prior to 2009, reaching a trough in 2012, and then climbing steadily up until 2018; and it appears the trend will continue.

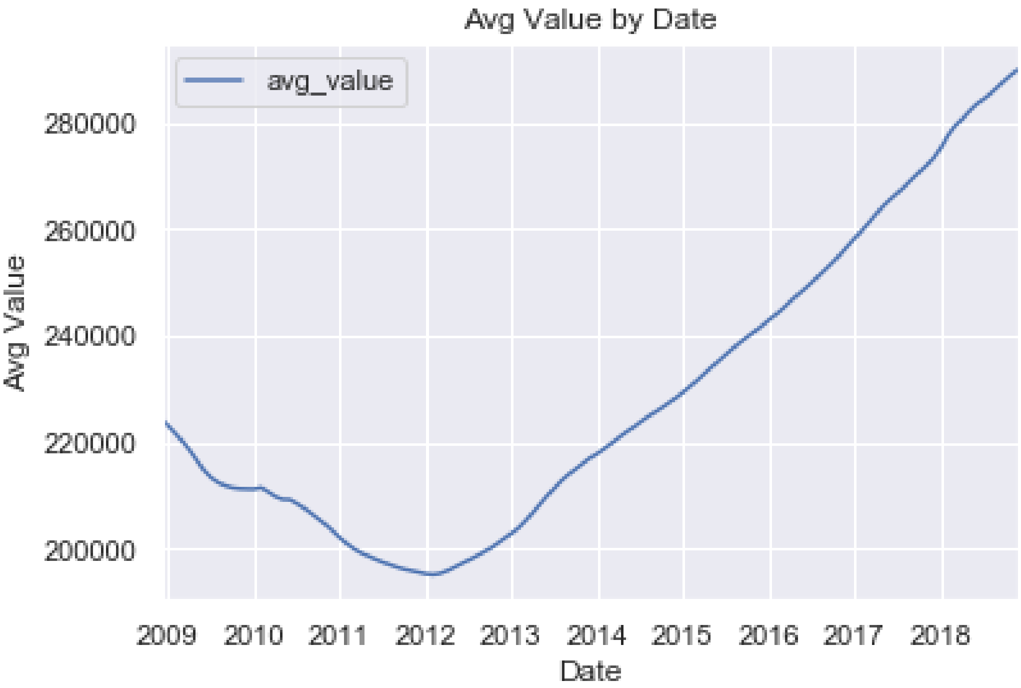

Figure 1: Average household value in the U.S.

It should come as no surprise that household value was deteriorating prior to 2009 because the U.S. household market was hit by a recession in 2007 that took five years to recover from. However, not all markets behaved the same. The data was plotted once again, figure 2, this time at the state level.

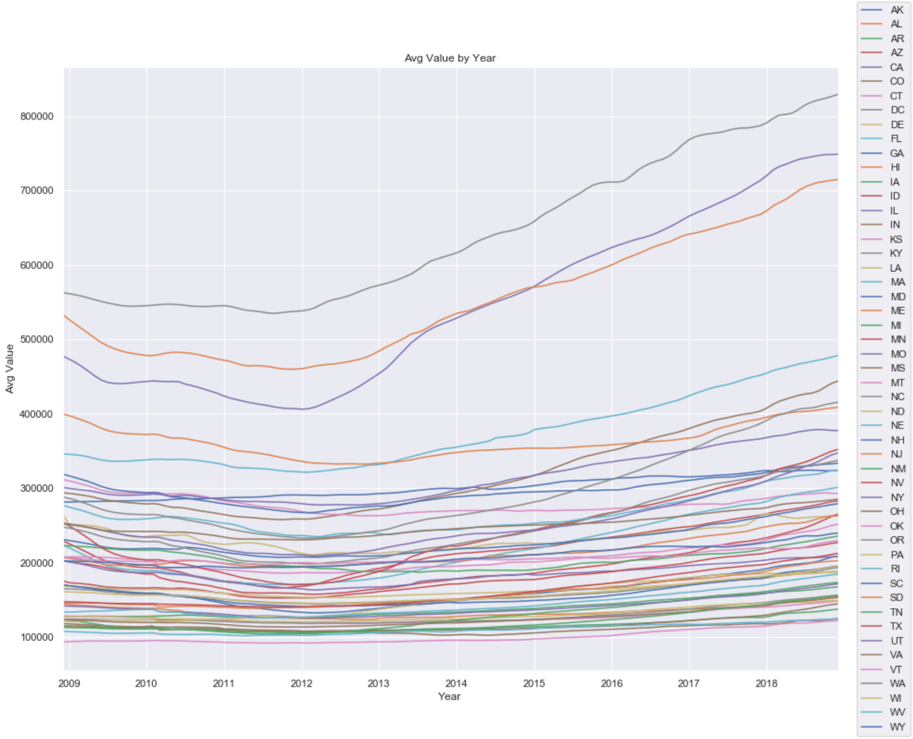

Figure 2: Month-to-month average value change by state

It is not clear by this second image which line belongs to which state. But what's interesting from this graph is that it is obvious that the recession did not hit all the states equally. In fact,

most states don’t see a dip in their household depreciation rate. Rather, they remain quite flat and see a growth after 2012. There are, nevertheless, three states that seem to see their household values depreciate, but recover strongly and can be thought responsible for driving up the average household value at the national level. These states are, in order of average household value by the end of 2018, the District of Columbia, California, and Hawaii.
But this graph does not tell the whole story. Even though these three states have grown significantly value-wise, it does not necessarily mean that they have grown significantly over the length of the sample. Figure 3 shows how each state evolves percentage-wise over the same time period, showing that five states have appreciated significantly. The average household values of Idaho, Utah, Indiana, Nevada, and Montana saw a rise of over 10 percent. If people living in states where household values are appreciating rapidly, or on the other hand can’t afford to live in a more expensive market, then they may have moved to one of these five states, helping increase the value of their markets.

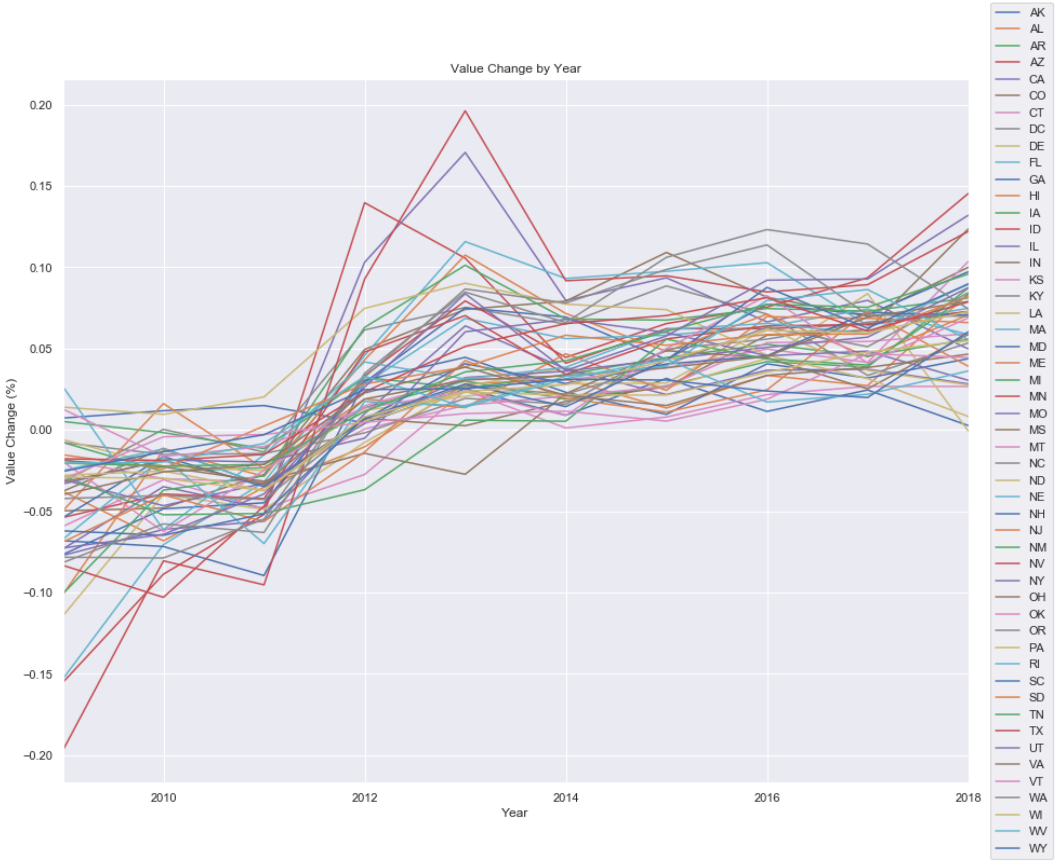

Figure 3: Month-to-month household value percentage change per state

If a person or company were to invest in real estate, it’s likely that they would have a bigger return on investment if they were to invest in one of the states mentioned. It’s probably likelier that the biggest return on investment comes from the states that have grown percentage-wise rather than value-wise.

In the spirit of further analysis, the state of Arkansas was analyzed to see whether it's household valuation pattern followed that of the U.S (figure 4). Furthermore, if a person were to invest in the state of Arkansas, it would be of interest to know how each metropolitan area within the state has changed over time (table 3). All but four of the Arkansas Metropolitan Areas more than doubled their average household value over the 22-year period that we have data for.

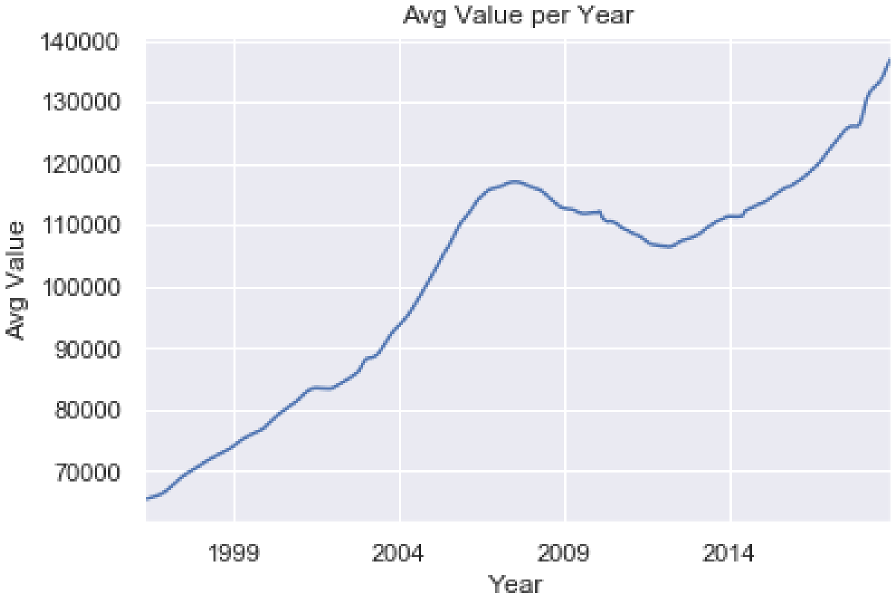

Figure 4: Average household value for Arkansas

Table 3: Household value change between Apr-96 and Dec-18.

Metropolitan Area	Valuation change (%)
Jonesboro	137.64
Fayetteville-Springfield--Rogers	129.91
Fort Smith	127.40
Paragould	125.90
Harrison	123.20
Arkansas Non-Metropolitan	116.18
Batesville	116.12
Mountain Home	112.79
Russellville	108.04
Hot Springs	105.35
Little Rock-North Little Rock-Conway	99.87
Texarkana	93.12
Searcy	76.78

Memphis	68.71
Magnolia	NA
Pine Bluff	NA

Discarding the Magnolia and Pine Bluff metropolitan areas, which had no data for 1996, the metropolitan areas of Jonesboro, Fayetteville-Springfield-Rogers (FSR), Searcy, and Memphis were plotted (figure 5), trying to determine which of these would prove the best investment opportunity.

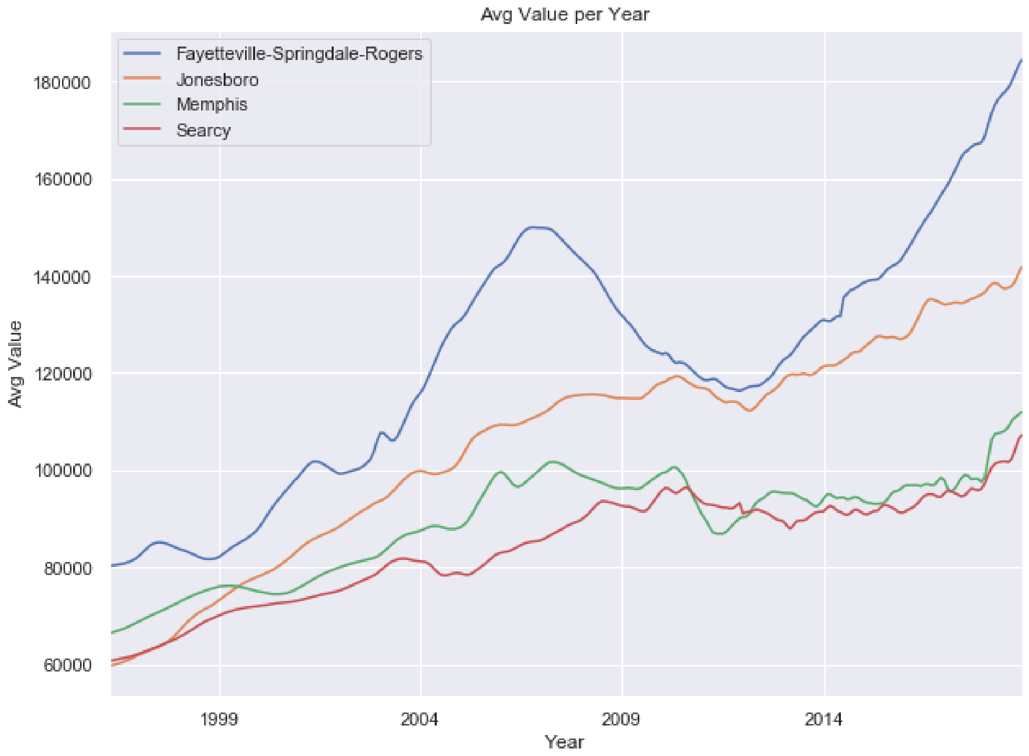

Figure 5

Judging on the data, a person wanting to get maximum return on investment would choose to invest in Jonesboro. By graphing the data, we can see that Jonesboro has been the more stable of the four markets; however, any person would be tempted to invest in FSR as a second choice. The problem is that FSR was hit hard by the recession and, if another recession were to happen, the pattern would probably repeat itself. That is not the case with Jonesboro, as its a smaller market and was not affected much by the recession.

Investing in Memphis or Searcy is not a bad option either. Both metro areas have seen growth, though not as big as the rest of Arkansas. They also weren’t hit hard by the recession. Instead, they both depreciated their household values during 2009 and 2012, recovering and growing at a very slow and steady rate.

With these insights into how markets fluctuate and grow, we can start building models to forecast the different zip codes.

Modeling

Because of the size of the data, we have decided to restrict our modeling efforts to those states that have seen either significant growth either value or percentage-wise. In this case, we will be focusing on California, Hawaii, Idaho, Indiana, Utah, Montana, Nevada, and the District of Columbia.

By restricting ourselves to the zip codes in these states, the data set has been reduced from 15,508 observations to 2,134, a significant loss of data, but necessary due to our limited computer processing power.

After subsetting the data, the first step was to build a partial autocorrelation plot to determine how many lagged values our model parameter might need. We chose three random zip codes from the data set, plotted them (figure 6), and found that all zip codes had a lagged value of one.

We ran an ARIMA model using this lag-value, along with 4 degrees of freedom, obtained by doing a grid search over 12-month periods for the entire time-series. We evaluated the model results (figure 7) and found the initial model to be acceptable within the training parameters.

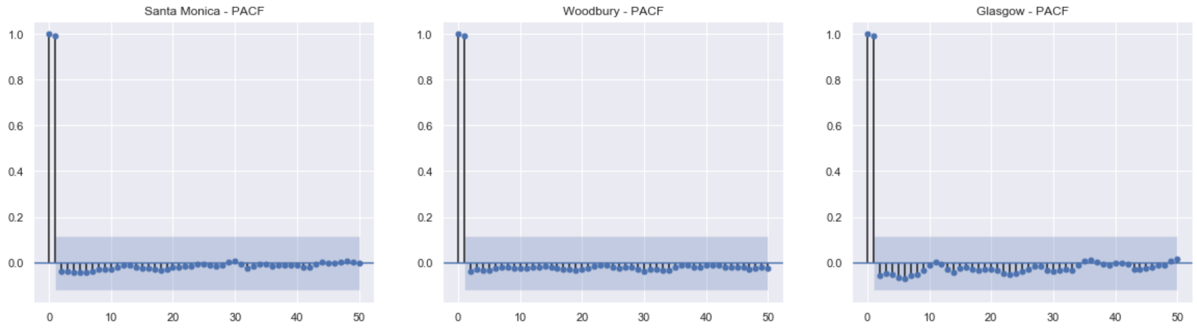

Figure 6

However, when running the model through the test set, we found that, when compared to the mean, the model was grossly overfitting, rising from an mean absolute error of 5,849 to 94,452. Though this was discouraging, the lower confidence intervals were very close to the actual values for 2018 (figure 8).

Nevertheless, we decided to build a model using the fbprophet package in Python, obtaining much better results when looking at the overall metrics at both the mean and lower CI interval, 64,263 and 17,909.

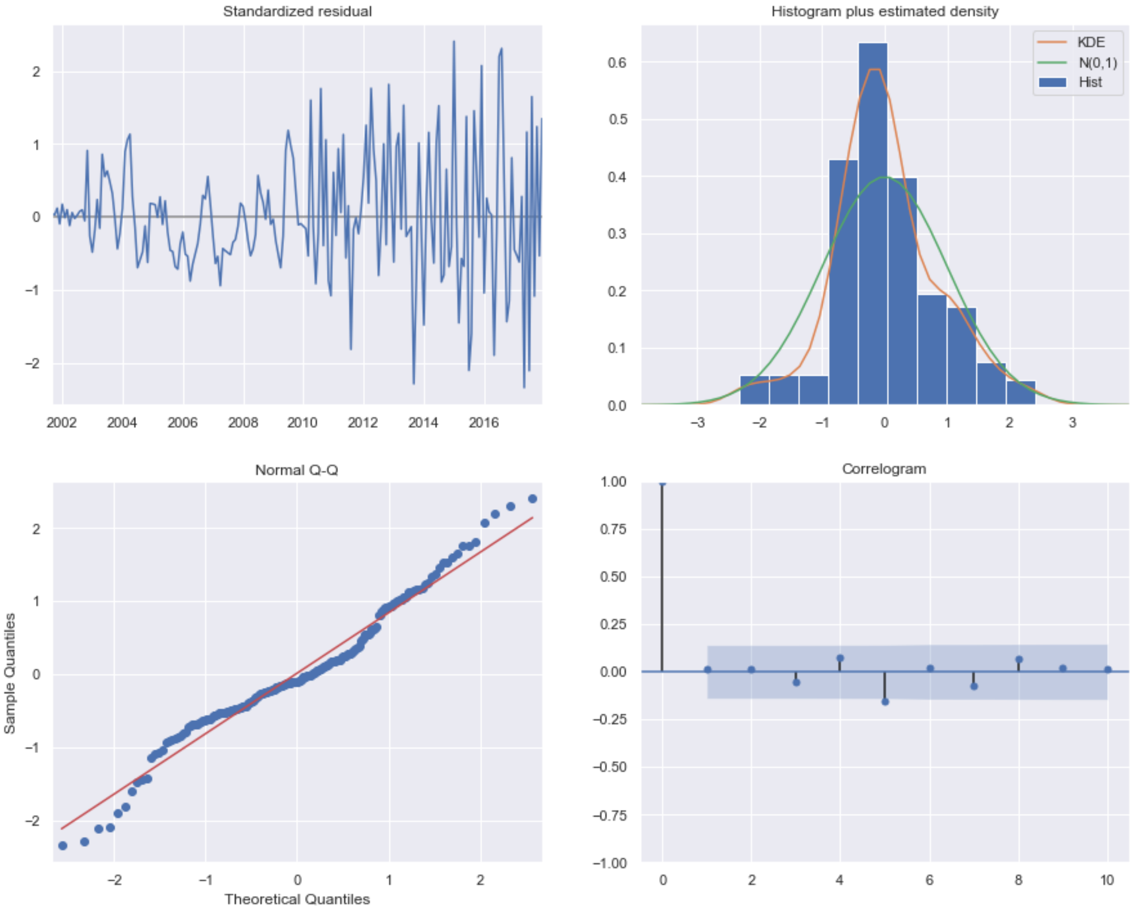

Figure 7

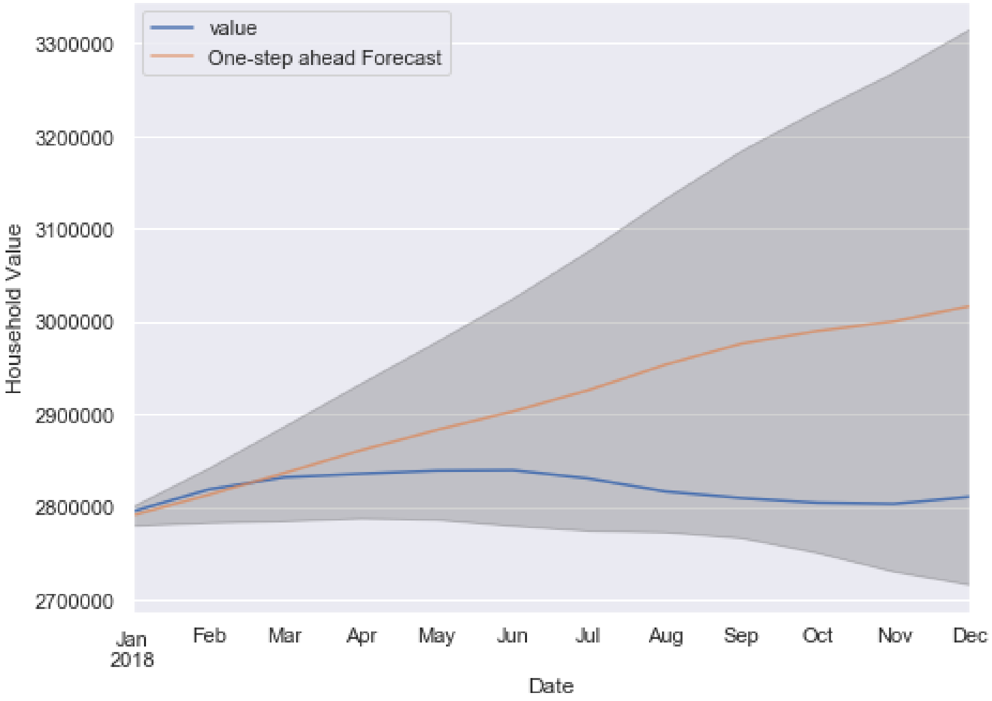

Figure 8

Given the results of the Prophet model, we decided to move forward with said choice, and proceeded to fit each zip code with the model, using their time-series data from 1996 to 2017 as the training set, and validating with their 2018 data.

Results

Having iterated over the 2,134 zip codes, we proceeded to compare the model results for the next 12-month period to the actual 12-month period in the test set. We measured the mean absolute error and R-squared among the predicted and actual results, and compared the forecasted change in value percentage-wise.

When we look at the distribution of the mean absolute error (figure 9), we find that it is very skewed towards the left, which is good, as it shows that the majority of the forecasted values are close to zero, and therefore there is not much deviation between what the model is predicting and the actual events that took place.

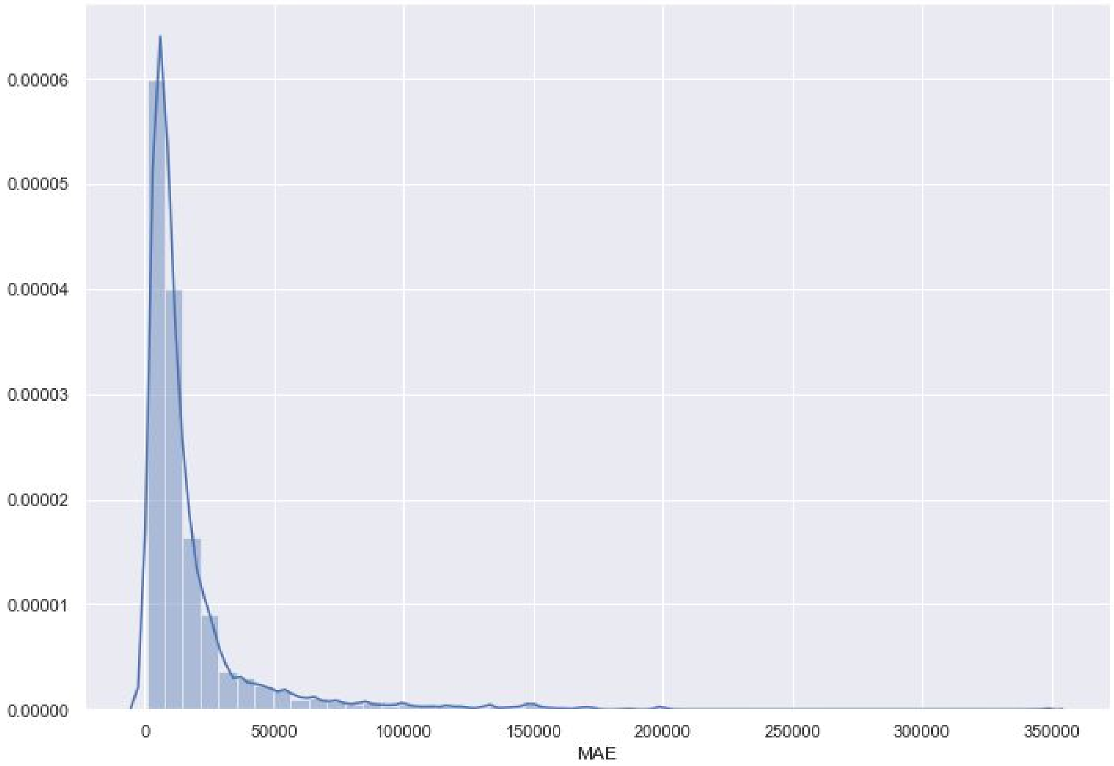

Figure 9

The same can not be said of the R-squared value, as the model returned over two-thirds of the forecasted values with R-squared scores below 0 (figure 10). Though initially thought to be an error made by the calculation, it is possible for these scores to have negative value as they show that the model really missed on the actual values given unforeseen patterns of conduct, similar to what was encountered when the model was tested on the Santa Monica 90403 zip code (figure 11).

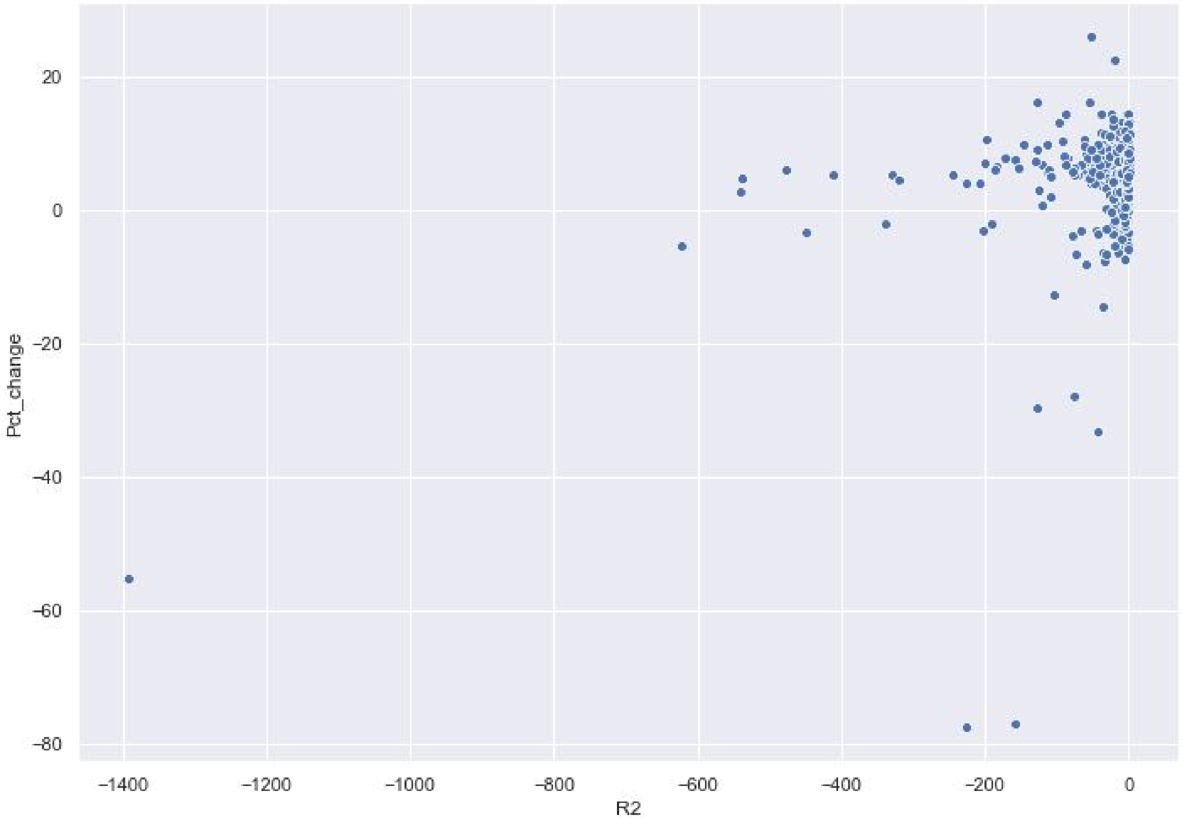

Figure 10

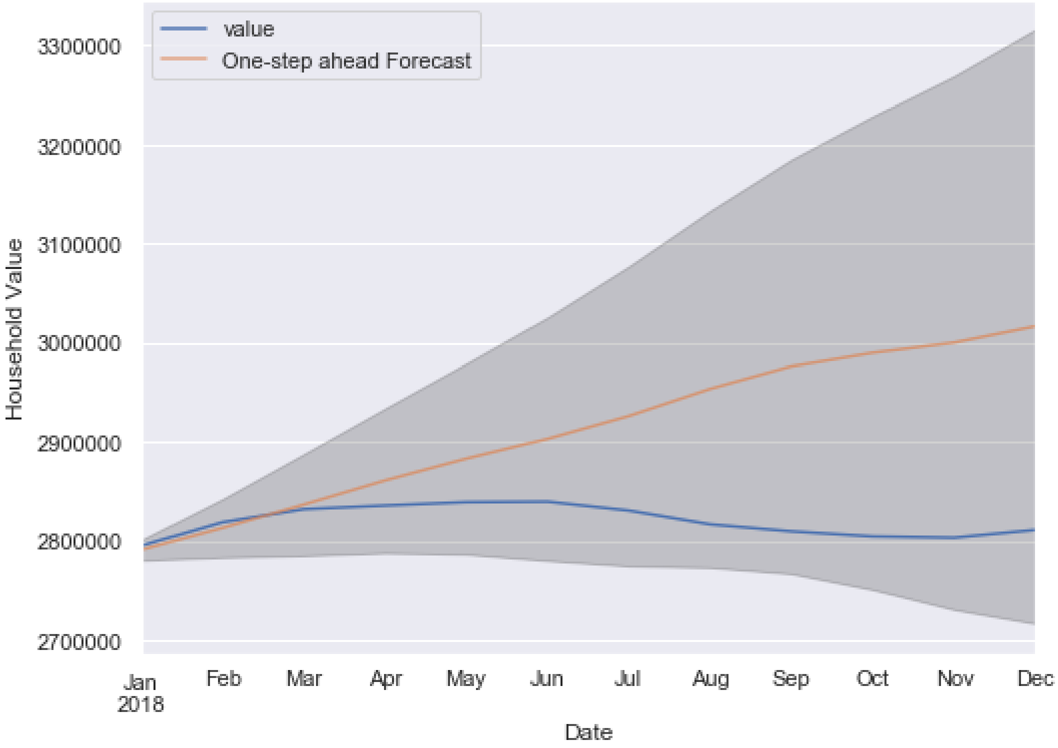

Figure 11

Because of these very large misses, we retained the results of the models that had R-square scores greater than zero. 616 zip codes were kept in this final data set (figure 12), that were then used to determine which areas would be best suited for high-low risk/high-low reward investments; along with the three best investments, based on R-squared value and high return on investment.

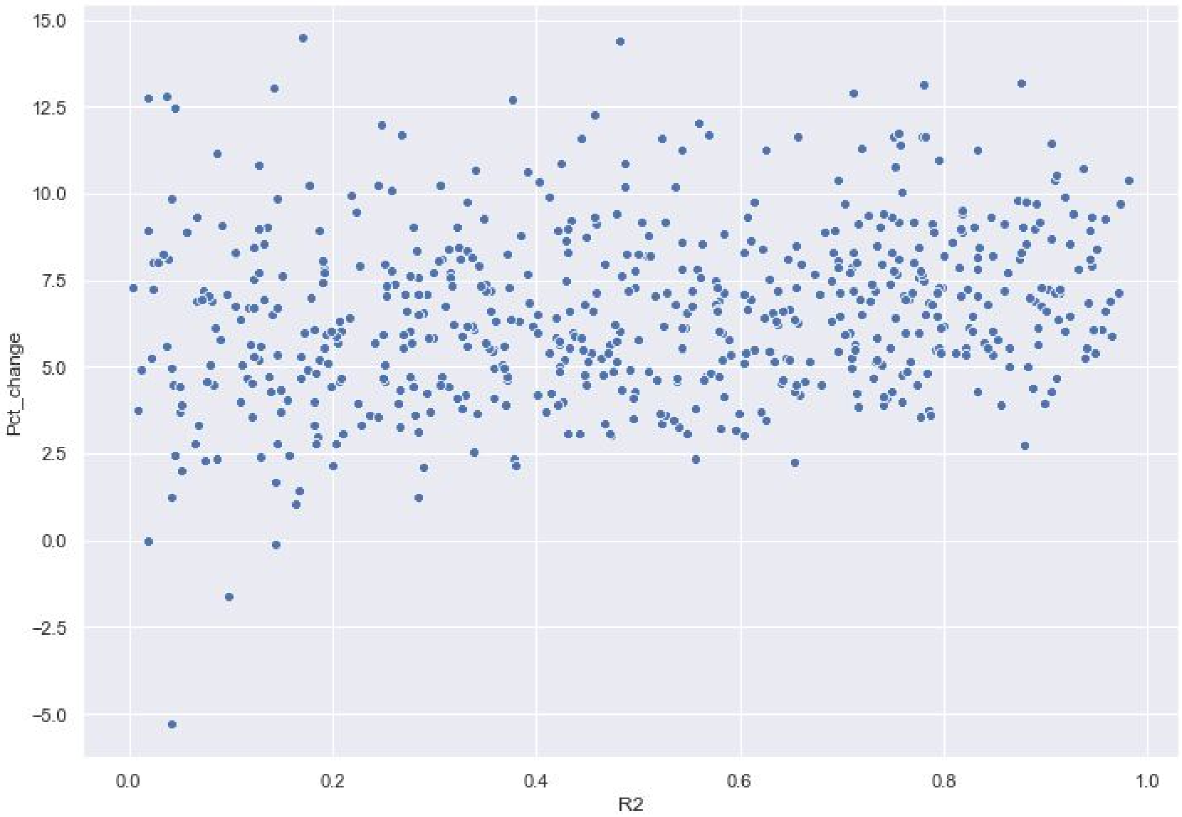

Figure 12

Conclusion and Answers

Before determining which three cities (and zip codes) are the best suited for investment by the SREIT team, we want to present four potential cities that mix high-low risk/high-low reward based on their R-squared and percentage-wise change in value over a 12-month period. We must note beforehand that models with low R-squared values are more high-risk, since the model is not able to determine a pattern between the forecasted and actual values. Conversely, low-risk models will be those with a high correlation between the forecasted and actual values (those to the right of Figure 12.)

Another note that we want to add: Though figure 12 shows that there are potential investment sites that have lost value over a 12-month period, we will not focus on those values since we believe that no sensible person would want to invest in a market that has depreciated over the past 12-months.

With those caveats established, we found the following results:

- Evansville, IN, 47720: Provides a high-risk, low-reward investment given that over the past 12-months, properties in Evansville have appreciated by 1.25 percent. This is not a very high ROI but we are not confident that this growth will be sustainable as the model throws an R-squared value of 0.042, showing that there were outside effects that might have affected the market and weren't captured by the model.
- Eastvale, CA, 92880: In the case of Eastvale, the model is also not confident on the growth of this market, given that the R-squared is almost negligible (0.005). However, Eastvale has shown considerable growth over the past calendar year, increasing its average property value by 7.31 percent.
- Oak Park, CA, 91377: The forecast model is very bullish on Oak Park, having an R-squared of 0.879. However, Oak Park has not grown much over the past year, appreciating its property value by 2.72 percent. Oak Park represents a very safe bet in any investment portfolio.
- San Francisco, CA, 94112: A popular city given its proximity to Silicon Valley, San Francisco is the best investment choice for low-risk/high-reward results; having a very robust R-squared of 0.982 and seeing a 10.4 percent growth over a 12-month period.

Cities like San Francisco, especially zip code 94112, are the type of investment opportunities that clients seek. They're very safe bets as they have remained consistent over 12-month periods and have also increased their property value. If we want two other option aside from San Francisco, it would be very wise to check property listings in markets like San Diego, CA, 92102, and Stockton, CA, 95206 (table 3).

Table 3

RegionName	City	State	CountyName	R-squared	Household appreciation (%)
94112	San Francisco	CA	San Francisco-Oakland-Hayward	0.982	10.4
95206	Stockton	CA	Stockton-Lodi	0.974	9.7
92102	San Diego	CA	San Diego-Carlsbad	0.971	7.1

It is curious that out of the six investment opportunities presented, five of them are in California. This speaks volumes that the model is focusing mostly on month-to-month value change more than percentage change in household appreciation. This could be amended if the model were to focus on both types of change rather than just property value.

Furthermore, the model didn't consider any outside factors such as employment rate, inflation, or average wage for each state or zip code. Adding this information could help us make a more informed decision on any potential investment opportunities but, with the current data, we are confident that the best three markets to invest in can all be found in California.

Resources

- https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
- https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3
  https://medium.com/@urban_institute/using-multiprocessing-to-make-python-code-faster-23ea5ef996ba
- https://dius.com.au/2018/09/04/time-series-forecasting-with-fbprophet/
- https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/
- Special thanks to Justin Clark who shared his multiprocess code along with providing advice on how to run it, despite not actually being able to run it.