Alex Hyman
Lab 6

Introduction

The Syracuse Real Estate Investment Trust (SREIT) has entrusted me to determine which zip codes would offer the best investment properties for the trust. For this analysis, I will utilize historical housing prices and adjusted gross income data[1] to create an ARIMA model that will forecast the housing prices in December 2018. The final models will be built with training data from 2011 through 2017. After deciding on which zip codes to recommend, I will utilize the data that is available for 2018 to see how the recommendations are performing.

The Data

Before any models were created, I performed a brief inspection of the comma-separated file. The file contained 15,388 rows and 277 columns and had column names that included the regionID, Region Name (zip code), City, State, Metro Name, County Name, Size Rank, and a year-month value starting in April of 1996 and ending in September of 2018. Each of the rows were a different zip code, and were organized by the size rank of the zip code. The values in the year columns were numeric and indicated the average price of a single family home in the given zip code.

To see how the housing prices were trending over time, I looked at the data to see that the median average price of a single family home in the first period (April 1996) was $96,500, and the month had an interquartile range of $73,800. The median average price for a house at the last period in the data set was $197,800 with an IQR of $187,000. This quick overview of the data suggests that the prices of single family homes increased overtime, and that the data has a heavy positive skew (IQR is nearly as large as the median!).

The csv file was also checked for any instances of missing data. In the first month of data collection (April 1996), there were 1122 different zip codes that were missing the average price of a single family home. The first month in which there were not any zip codes missing data was January 2015. A plot of the number of zip codes missing the average single family housing price over time is provided was created, and is provided in Figure 1.

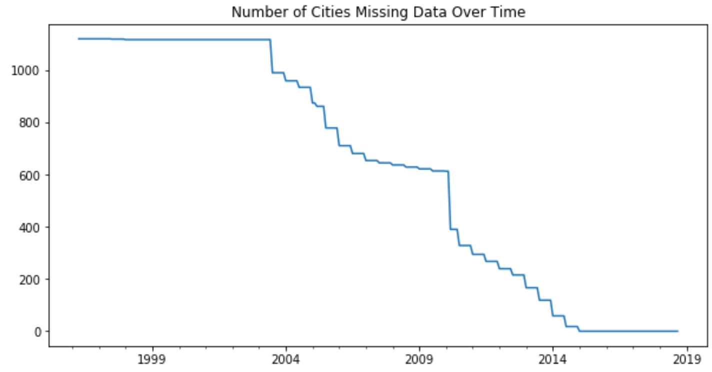


Figure 1 Number of Zip Codes Missing Values

[1] Acquired https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi

The adjusted gross income data came in six different files, one file for each year between 2011 and 2016. Each of the files contained about 166,000 rows and between 70 to 150 different column. Each row contained a zip code, but zip codes were duplicated multiple times. This was due to the different tiers in the AGI reporting. To determine the average AGI for a zip code in a given year, I needed to sum the total AGI reported within the zip code, and divide it by the total number of returns collected from that zip code. There were also some zip codes that were not available within a state, and they were removed from the data before modeling. The AGI data was also reported in 1000's of dollars, therefore I multiplied each of the calculated AGIs by 1000.

Cleaning the Data

To make this data usable for time series analysis, I brought the data into a pandas data frame, and created a series that combined the zip code and the city name of that zip code. I then removed the first seven columns of the data frame and transposed the data so each of the rows were a different month, and each of the columns were a different zip code. Next, a new series was created that contained the zip code its city name, this series was inserted to replace the old column names, so I could tell which city is being evaluated, as well as the zip code. The index of the data frame was also converted into datetime objects so time series methods and analysis could be conducted. This final formatting had the name of the zip code and city as the column and the month-year combination as the index. Each of the data objects in the data frame were the average single family home price in a given zip code at a given point in time.

The AGI data that I was planning on using as an exogenous variable in the modeling process is reported on a yearly basis. The ARIMA model, that I planned to use to forecast future housing prices, requires both the endogenous data (housing prices) and exogenous data to have the same frequency. Because I felt that more information would be lost in the model if I were to convert the housing prices into yearly data, I converted the AGI data into data reported monthly. This was accomplished by using a linear interpolation of the AGI data between the ends of the years.

The interpolation was useful in adding data between reported AGIs, but there was data missing from January 2011 to November 2011 and from January 2017 to December 2017. Additionally, I needed to have anticipated AGI data if I wanted to forecast housing data with this model; this meant I also needed to have AGI data from January 2018 to December 2018. To account for this missing data, I extended the interpolated rate I used from December 2015 to December 2016 until December 2018. The data missing from 2011 was accounted for by extending the interpolated rate of the AGI from December 2011 to December 2012 in the reverse direction. Figure 2 shows the AGI data that had been interpolated and when the known data has been reported for zip code 30024.

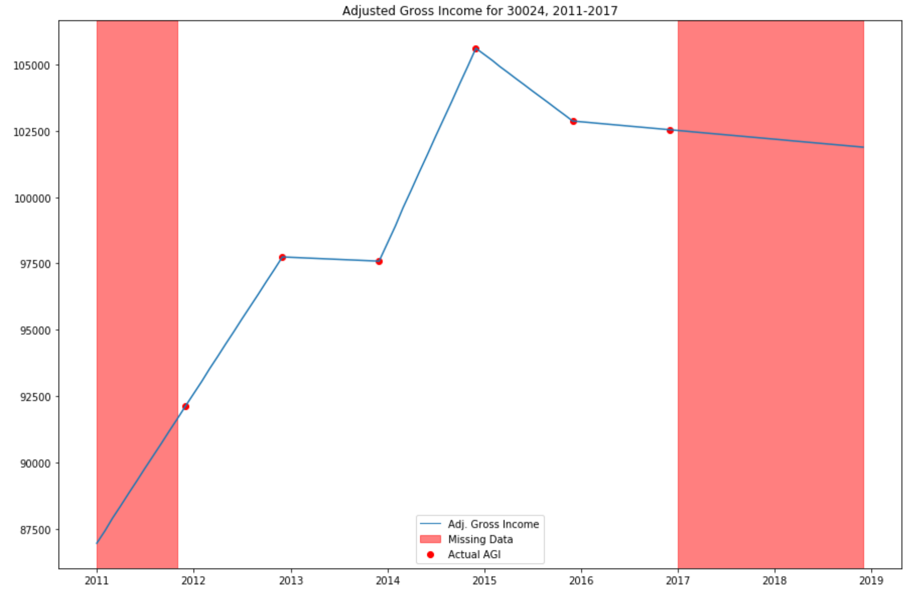


Figure 2 Interpolated Adjusted Gross Income

Time Series Exploration

Now that the data was formatted in a usable manner for time series analysis, I wanted to conduct some preliminary time series analysis/data exploration. Data for the Arkansas cities of Hot Springs, Little Rock, Fayetteville, and Searcy were chosen as the subjects of the data exploration. Indexing these cities showed that there were two zip codes in Hot Springs, eight zip codes in Little Rock, three zip codes in Fayetteville, and one zip code in Searcy. The zip codes were then converted into four different time series grouping the different zip codes by the city, and then taking the mean of the housing prices. This resulted in a time series for each city that had the average price of the average home in each zip code in the city. The average housing prices over time for each city was then plotted to see the movements in average price. This plot of the average housing prices for the Arkansas cities is provided in Figure 3.

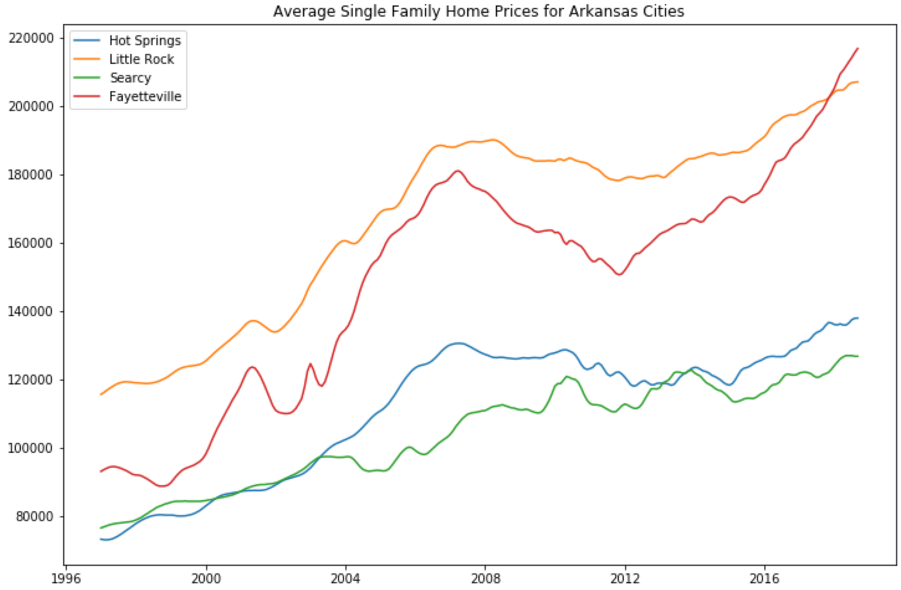


Figure 3 Average Home Prices for Arkansas Cities

The plot in Figure 3 shows that over time, all three of the cities had their home values increase in price. The most recent data shows that Fayetteville has the highest single family home value, followed by Little Rock, Hot Springs, and then Searcy. In about 2002, the average single family home prices in Little Rock, Fayetteville, and Hot Springs began to increase at a faster rate while Searcy remained fairly steady with its trend upward trend in average single family home price. Beginning in about late 2007, the trend in home values in the cities of Little Rock, Fayetteville, and Hot Springs began to decline to the point where the housing prices began decreasing in value; however, the trend in the Searcy housing prices remained steady and continued to move in an upward direction. Finally, around 2012, the housing prices once again began to increase in Little Rock, Fayetteville, and Hot Springs at an increased rate and the housing prices in Searcy once again remained steady.

To see how related the time series for the four Arkansas cities were, a heat map of the correlations was created and is provided in Figure 4. This heat map shows that in general, all four of the time series move together; however, the relationship between Searcy and the other three cities is not as strong as the relationships the other three cities have with each other (R values of 0.97, 0.98, and 0.99).

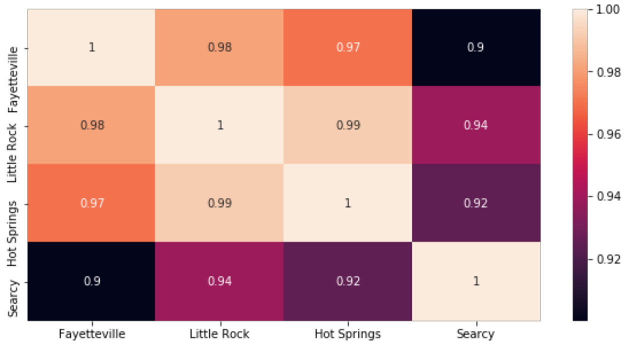


Figure 4 Correlations for the Four Arkansas Cities

While looking at how the time series data behaves over time is helpful, a time series is generally composed of three different parts: a seasonal component, a trend component, and a white noise component. These components of a time series can provide insight into how the data behaves that may not have been apparent in an overview of all the data available. For example, a plot of the seasonality of the four cities in Arkansas is provide in Figure 5.

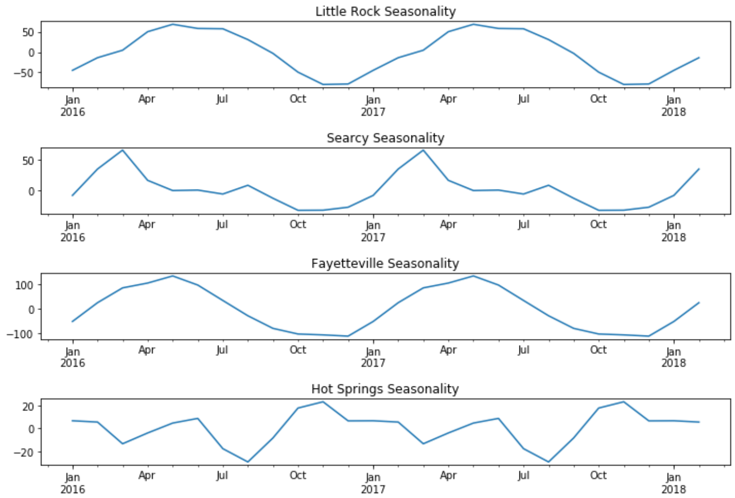


Figure 5 Seasonality Extracted from the Four Arkansas Cities

Figure 5 shows that each of the Arkansas cities have a 12-month seasonality period and the seasonality does not really have a huge effect on the prices. The largest seasonal variation was about +/- $100 in Fayetteville, depending on the month. Fayetteville and Little Rock's seasonality generally follows the same trajectory of having the peak prices in May and the lowest prices around November/December. Searcy's seasonality has similar characteristics as Little Rock and Fayetteville, but the peaks and valleys for the seasonal housing prices happen about two months earlier, with the peak prices occurring in March, and lowest prices occurring in October. The seasonality of Hot Springs does not appear to be very cyclical, and the effects of seasonality only appear to contribute to a difference of about +/- $20 depending on the month.

Additionally, the trend component of a time series can was analyzed to provide insight into how the data has been moving over time. The trend component essentially removes the effects of the seasonality and white noise to provide the overall trajectory of the housing prices in each of the cities. The trends for each of the cities have plotted and provided in Figure 6.

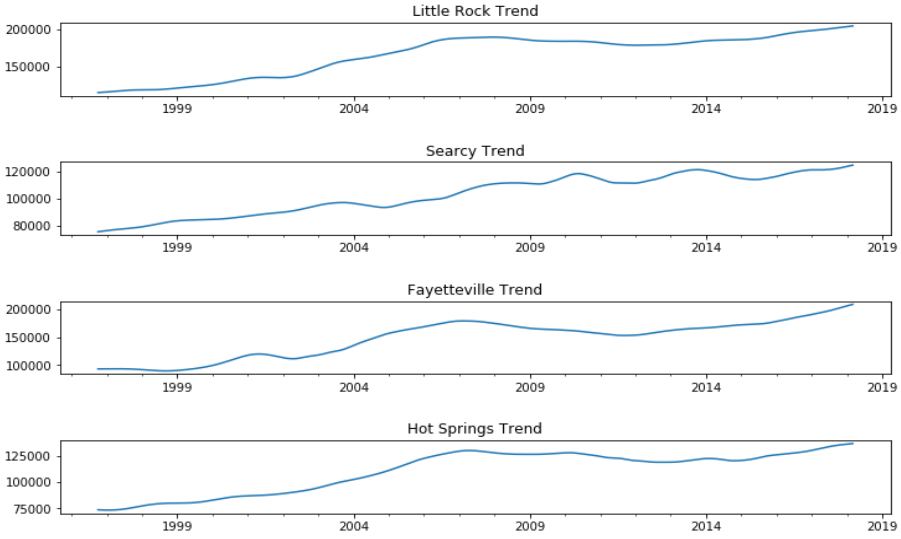


Figure 6 Housing Prices Trend for Arkansas Cities

The trends for the four cities show a pattern similar to the one extracted from the aggregated time series. All four of the Arkansas cities have housing prices increased in value over time. Little Rock, Fayetteville, and Hot Springs all saw an increase in the rate of prices increasing in value (slope increases more than it had been) beginning around 2002, and then the trend in housing prices decreased to a point in which single family home houses are decreasing in value. Searcy's trend line is generally pretty steady with a positive slope, which indicates that the housing crisis in 2008 did not have too much of an effect on the value of a house in Searcy, Arkansas.

The autocorrelations and partial autocorrelations of the cities in Arkansas were also evaluated, and each showed a similar trend. To generalize, the autocorrelations for each time series had a linear decay in the autocorrelation value, and autocorrelation became insignificant at around the lag-24 position. The partial autocorrelations shows strong significance at the lag-1 position (> 0.95), but became insignificant at the lag-2 position. This information informs us that the autocorrelation values at lags greater than one are heavily influenced by the partial autocorrelation at the lag-1 position. This indicated that really only the lag-1 position is affecting the next data point, and it is likely that a model forecasting home values will have an AR(1) component. The ACF and PACF plots also did not show significant signs of seasonality in the housing prices data. The autocorrelation and partial autocorrelation plots for Little Rock are provided in Figure 7. The other three cities have very similar ACF and PACF plots, and therefore the plots are shown only in the jupyter notebook

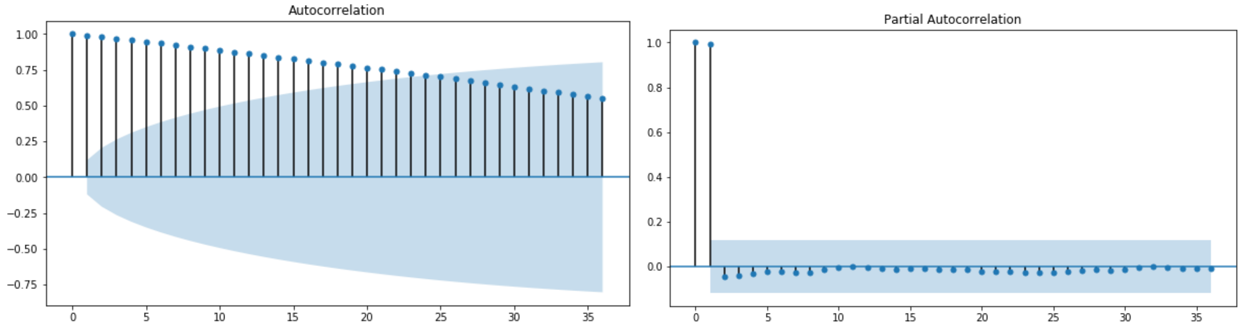


Figure 7 ACF and PACF for Little Rock, Arkansas

Modeling Arkansas

After analyzing some time series components of the cities in Arkansas, I wanted to try and forecast the average home price within these cities. While this modeling is likely not going to be representative of the rest of the United States, I wanted to compare the results of a lot of parameter variables on a much smaller sample size. Additionally, this preliminary modeling will provide me with a general idea of what the best parameters could look like when a larger sample across the United States is taken.

For this modeling, I chose to use the ARIMA model because the differencing in the algorithm would keep the time series stationary, and I am already pretty confident that there is an autoregressive component to time series due to the PACF being significant at only lag-1 [likely an AR(1)]. Each of the Arkansas cities that had been a part of the EDA process were evaluated with every combination of AR and MA order between 1-4, and with both first order and second order differencing. The data the model was trained was monthly housing prices from January 2007 through December 2017. The model was then forecasted for the next nine months, and then evaluated with the mean squared error of the nine known data housing prices in 2018.

After looking at the results of this testing method, there was not a consensus of a singular model that would be best for forecasting the home prices for all of the cities. The best ARIMA models for Little Rock and Hot Springs had an AR(1) component however, Fayetteville's best model had an AR(3) component, and Searcy's best model had an AR(2) component. Second-order differencing appeared to work better for Hot Springs and Fayetteville, but first-order differencing worked better for both Little Rock and Searcy. The MA order also appeared to be all over the place for each of the cities, with Little Rock and Fayetteville's best model having an MA(1) component, Hot Springs' best model had an MA(2) component, and Searcy's best model had an MA(4) component. The best models for each city based on RMSE were plotted and the forecasted values and confidence intervals provided by the model were compared to the actual housing prices. The plots of each of the city's best performing model and the actual values are provided in Figure 8.

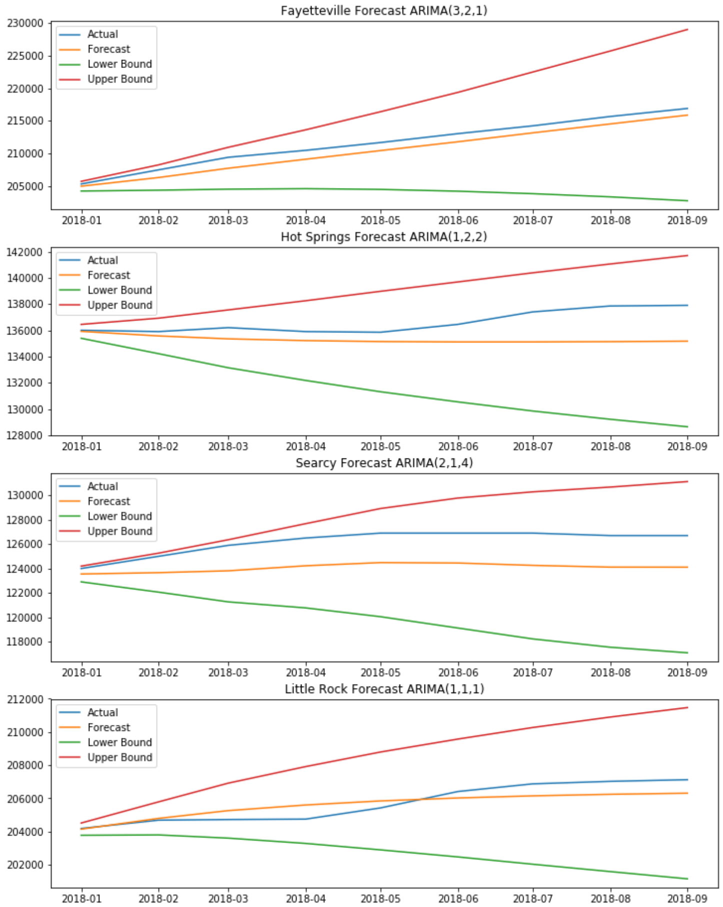


Figure 8 Best Performing ARIMA Model for Arkansas Cities

Choosing the Models

The two different time series that were available had date ranges of 1996-2017 and 2011-2017. Because we wanted to use both datasets in creating a model, only the housing data from 2011 to 2017 will be used to accommodate the missing AGI data. Additionally, Figure 1 had previously shown that there are zip codes missing monthly housing data until 2015; therefore I removed all zip codes from the housing dataset and the AGI dataset that did not have complete data from 2011-2017.

After removing the missing zip codes, we were still left with 14,600 zip codes to model. While it would be possible to create an ARIMA model for every zip code remaining, it would take a significant amount of time to find the parameters that best fit the zip code. To reduce the number of zip codes that would be modeled, I decided to limit the modeling to zip codes that had a December 2017 average housing price below $400,000. This limitation reduced the number of zip codes available to model, and it prevented us from having an investment that had an excessive initial cost. This filtering left us with 12,256 different zip codes that could be modeled.

With too many zip codes still left to model, I created a map to show which states have had the greatest percent change in housing price since 2011. A generalization of the performance of the zip codes within the state was included by plotting the median percent change of the zip codes with the state. States that had the greatest percent change in housing prices since 2011 were plotted in darker green, and states with a smaller percent change were plotted in light green. This figure is presented in Figure 9[2].

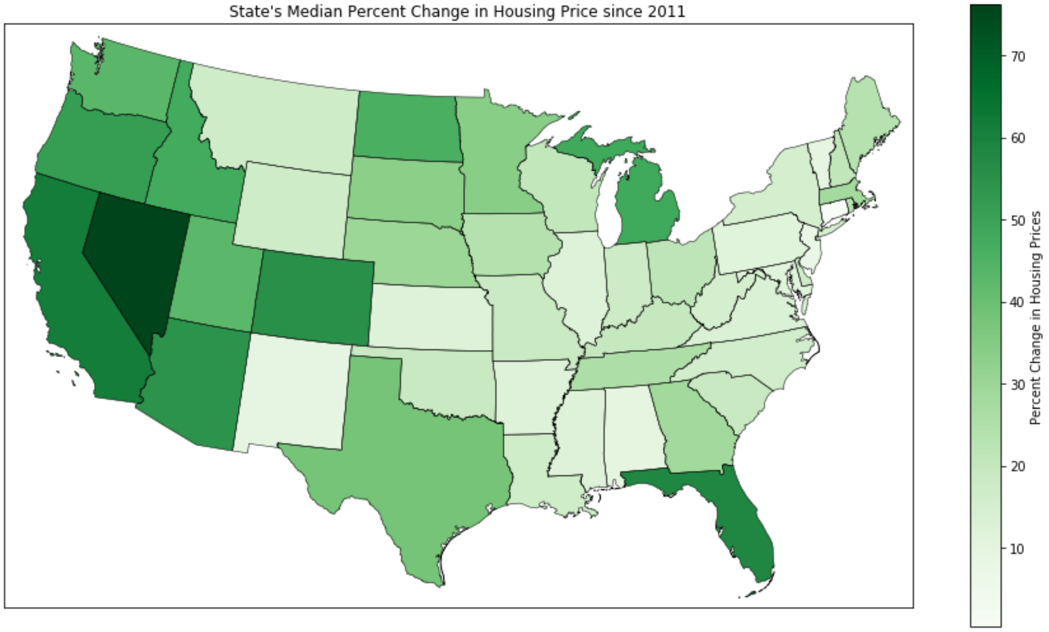

State's Median Percent Change in Housing Price since 2011
70
60
50
40
30
20
10
Percent Change in Housing Prices

Figure 9 State's Median Percent Change (2011-2017)

[2] Shape file from USDA https://infosys.ars.usda.gov/svn/code/weps1/trunk/weps.install/db/gis/st99_d00/

To narrow down the zip codes even further, I decided to only model zip codes that are in states that have been in the top 10 in highest percent increase in housing prices from 2011 to 2017 (NV, CA, FL, CO, AZ, DC, OR, MI, ID, and ND). After filtering out states that have not had high increases in housing prices, I was left with 2,294 zip codes, a manageable amount of zip codes.

Sampling a few of the remaining zip codes from this filtered list showed that the zip codes remaining followed the same trend of the autocorrelation decaying in a linear manner, and the partial autocorrelation being significant at only the lag-1 position, without much seasonal variance. This informed me that most of the zip codes will have an AR(1) component in the time series analysis. These ACF and PACF plots also showed that if there is a seasonal component to these time series, it is not significant enough to include in the model. However, the optimal differencing order and the moving average order for the zip codes was still unknown.

To determine the best mode for each of the zip codes, I created a loop that fit an ARIMA model with the exogenous AGI data to every zip code. Because I was confident that an AR(1) would be the most appropriate for this modeling, an AR(!) component was fixed for each test. Inside the zip code loop, I created a first-order differencing model and a second-order differencing model for each zip code, and I tested an MA order of 1-4 for each zip code.

Each loop utilized the housing and AGI data from 2011-2016 as a training set, and the housing data from 2017 as the test set. Every model created for each zip code calculated the root mean square error of the model on the test set, and the Bayesian information criterion (BIC) of the model. For each zip code, the order that resulted in the lowest BIC and the order that had the lowest RMSE was saved into a list. This list was used to see if the best RMSE models were generally the same as the best BIC models (they were not). Due to the difference in performance of the best BIC and best RMSE, I decided the most representative model of the zip code would be the model that resulted in the best BIC as RMSE in a time series can be greatly affected by randomness/white noise.

Recommending Zip Codes

In my opinion, the best zip codes to recommend were the zip codes that had the highest returns (reward), but with smaller confidence intervals (risk). I used the forecasted price for each zip code to calculate the return on each investment property as of December 2017. To account for the risk involved in the zip code, I used two metrics: lower bound percent change and the normalized confidence interval. The lower bound percent change was the percent difference in price between the December 2017 price and the forecasted lower bound confidence interval on December 2018. I calculated the standardized confidence interval by dividing the confidence interval width by the mean predicted price in December 2018. This allowed for the confidence interval to be a comparable metric to measure forecasted risk between properties at difference prices.

To limit my decision to a few of the zip codes with the best returns, I sorted the zip codes by the forecasted percent change and looked at the ten highest projected zip codes. The top ten zip codes are provided in Table 1.

Table 1 Highest 10 Forecasted Percent Change

Zip Code	City	State	Forecasted Percent Change	Lower Bound Percent Change	Normalized Conf. Interval
89060	Pahrump	NV	49.00%	21.88%	28.17%
83804	Blanchard	ID	42.41%	12.11%	40.20%
34465	Beverly Hills	FL	34.70%	18.09%	18.74%
34785	Wildwood	FL	33.28%	4.46%	42.93%
95245	Mokelumne Hill	CA	30.71%	-18.72%	71.12%
83334	Hansen	ID	29.44%	6.65%	34.47%
33805	Lakeland	FL	29.00%	10.24%	27.26%
85544	Pine	AZ	28.97%	2.38%	41.13%
92339	Forest Falls	CA	27.69%	2.88%	38.72%
89108	Las Vegas	NV	27.44%	11.76%	22.13%

An initial look at the projections show some huge potential returns on investment, but some of the properties looked to be quite risky as well. Due to the large risk in terms of the confidence interval size, I decided to eliminate zip codes 83804, 34785, 95245, 83334, 85544, and 92339 from being potential investment properties. To differentiate the remaining four zip codes, I plotted the historical housing prices from 2006-2017 (Figure 10).

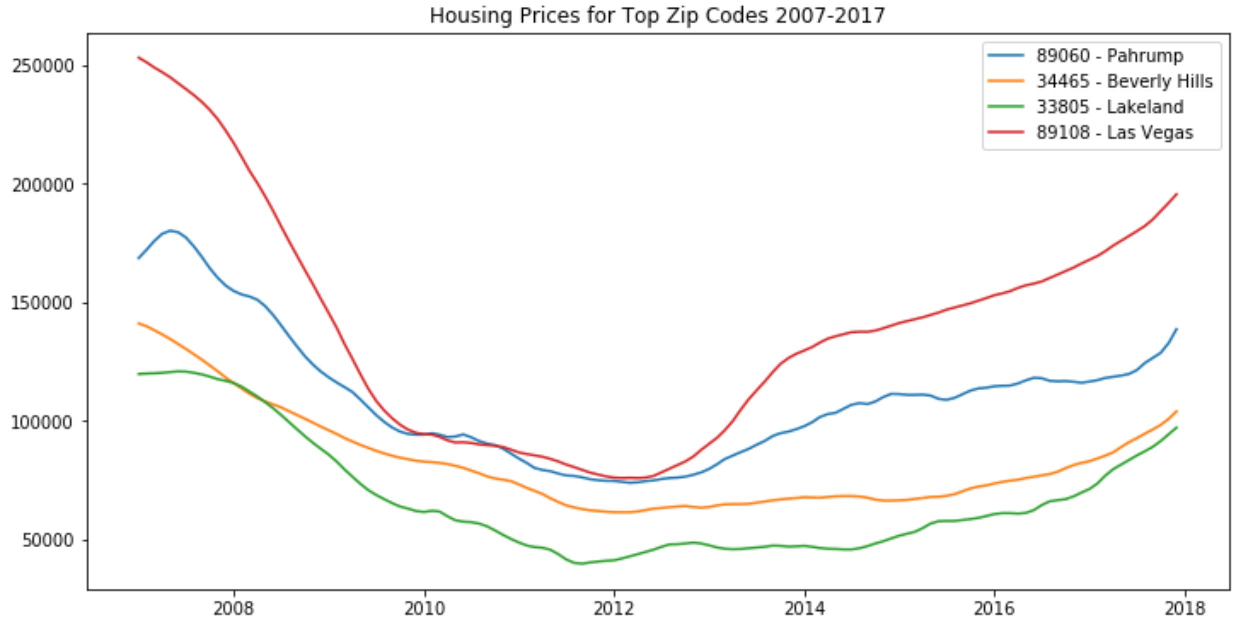
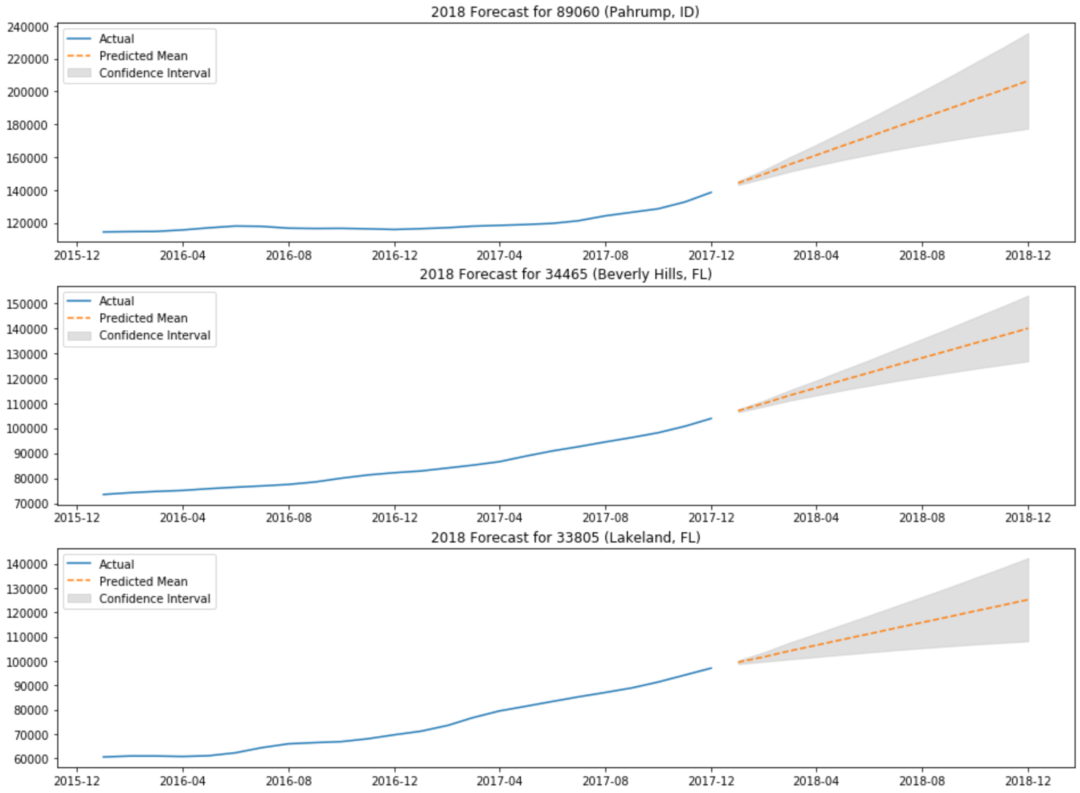


Figure 10 Historical Prices for Top Zip Codes

The figure shows that each of the zip codes were affected by the housing crisis in 2008, but the housing prices in Las Vegas were more heavily affected than the other three zip codes. While Las Vegas housing prices appear to have recovered since the housing crisis, the swings in trend in the Las Vegas real estate pricing appear to vary often and wildly, and therefore would be too risky as a long term investment. Therefore the three zip codes I have decided to recommend the SREIT due to the high returns and the limited risk are **89060** (Pahrump, NV), **34465** (Beverly Hills, FL), and **33805** (Lakeland, FL).

The first zip code I have recommended is 89060 in Pahrump, NV. The average single family home in Pahrump is currently priced at $138,600 and is forecasted to be priced between $177,429 and $235,611, with a mean forecast of $206,520 in December 2018. Pahrump, NV is also projected to provide the largest return on investment with a mean forecasted return of 49%. The next zip code I have recommended is 34465 in Beverly Hills, FL. A single family home is currently priced at $103,900 and is forecasted to be priced between $126,842 and $153,070 with a mean forecast of $139,956.30 in December 2018. The last zip code recommended is 33805 in Lakeland, FL. An average single family house in 33805 is currently priced at $97,100 and is forecasted to be priced between $108,185 and $142,333, with a mean forecast of $125,259.50 in December 2018. The mean predictions and the confidence interval for each of the zip codes recommended are provided in Figure 11. As forecasted, a purchase of an average single family home in each of these zip codes in December 2017, would likely result in a profit between $72,000 and $191,415, or returns between 21.5% and 56.4% with a mean return of 38.9% in December 2018.


2018 Forecast for 89060 (Pahrump, ID)

2018 Forecast for 34465 (Beverly Hills, FL)

2018 Forecast for 33805 (Lakeland, FL)

Figure 11 Forecasts for Top Zip Codes in 2018

Actual Performance of Zip Codes

Since our recommendation to the SREIT in December 2017 to purchase homes in zip codes 89060, 34465, and 33805, nine time periods have passed and we would like to check on the status of the investment properties. Figure 12 shows the actual values of the housing prices from January 2018 to September 2018, as well as our forecast for the year.

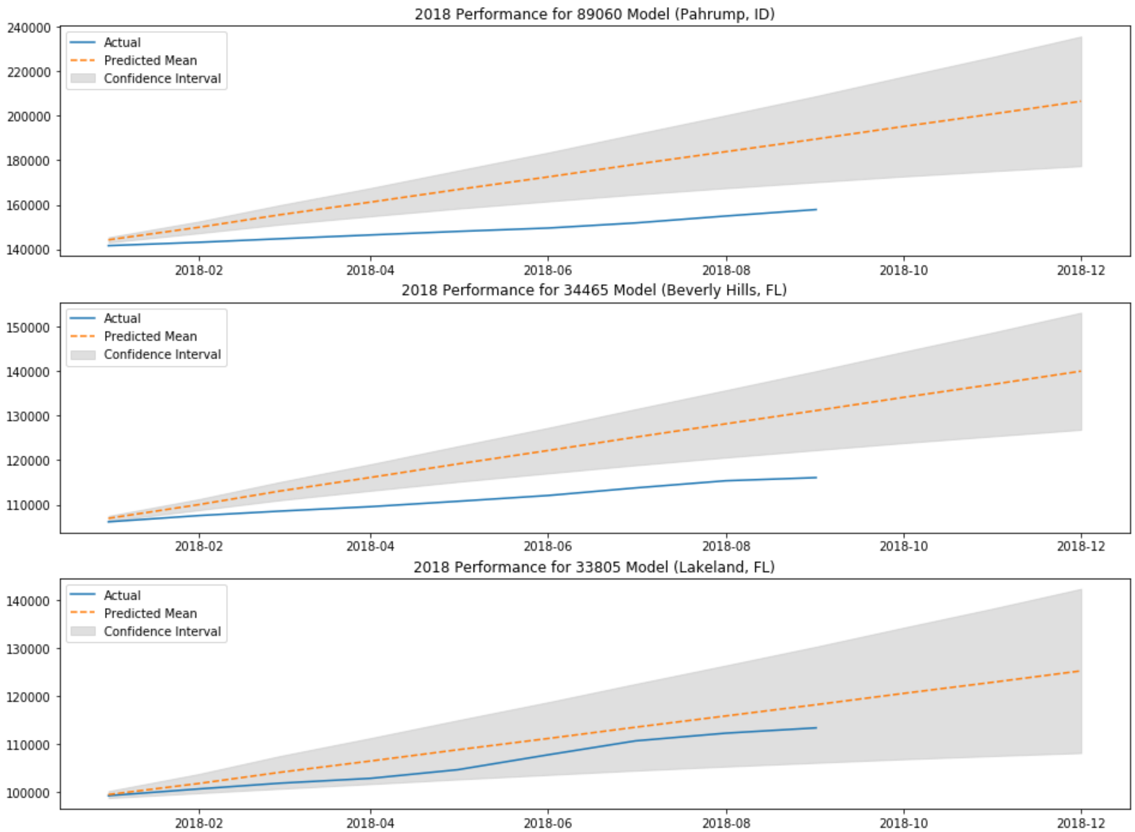


Figure 12 Model Performance for 2018

Of the three zip codes recommended, only the model for 33805 (Lakeland, FL) has performed within the forecasted confidence interval. The root mean square error for each of the zip code were calculated to be $20,531.71, $9,184.04, and $3,203.52 for 89060, 34465, and 33805 respectively. It would make sense that 89060 (Pahrump, NV) had a RMSE much larger than the other zip codes, as it was already the highest priced, and it was forecasted to perform the best.

While the housing prices have generally not performed as forecasted, the zip codes recommended have had returns of 13.92%, 11.74%, and 16.79% since December of 2017, which sounded quite promising. When comparing the percent change of the recommended zip codes to the percent change of zip codes across the United States, they were in the 94th, 90th, and 97th percentile of returns; so while our model was too optimistic in how well the housing prices in the zip codes would behave, each of the zip codes are in the 90th percentile or greater of returns. In fact, If we had purchased an average single family home in each of the zip codes, we would have made a profit of $47,800.00 since December 2017; a return of 14.08% on the initial investment.