Allison R. Deming

IST 718: Big Data Analytics

Lab Exercise 2 Write Up

Dr. Lando

Syracuse University

The housing market is constantly changing, with it be rather expensive in most areas to rent, own, or anything in between. As an agent or investor, knowing when and where to buy. How much to invest, and what to do with each property is essential to actually making a profit in today’s market. How is the best way to go about acquiring this information to make good decision in the ever-changing market? Data and research are key. Here we will be taking a deep dive into the Syracuse Real Estate Investment Trust to know more about investments and how to avoid overpaying for a property and what to sell it for when you are ready.

About the Data

To develop a model to predict the best zip codes. We need more than just the information that was provided by our Zillow Data. Here we will be using two different data sets to further explore the real estate situation, both unemployment data as well as crime data.

1. Zillow is a popular real estate site designed for renters, buys, sellers, and investors. You can find the data at the following URL: <http://files.zillowstatic.com/research/public_csvs/zhvi/Zip_zhvi_uc_sfr_month.csv>. Here you will find that the dataset has 29,768 rows, and 324 columns. This includes information from 1996 until the present day.
2. The dataset for unemployment comes to us from the US Bureau of Labor and statistics and can be found at the following link: <https://www.bis.gov.lau/laucnty17/xlsx> . This has 3,224 rows and 10 columns. Here all or the rows are from 2007 or are about five years old.
3. The crime data comes from the FBI Crime Data Explorer. This includes many different locations in the United States and was collected from the URL https;//crime-data-explorer.app.cloud.gov/pages/explorer/crime/crime-trend. This has 4,942 rows and six different columns of information.

Exploring the Data

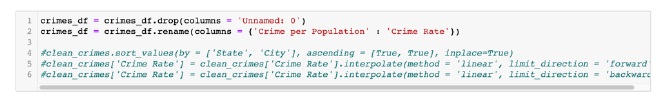
All three datasets were downloaded from the internet and imported into Anaconda using Python’s PANDAS (being made into data frames in the process). All three data frames needed an extensive amount of clean up and altering before they could work together to give us the answers that we were seeking. The Zillow data frame for example needed a lot of transformation to make the dates not appear as rows for the months and years, being converted into singular values. Rows were removed from the unemployment data set to make it useable as well. The FBI dataset had a few other tweaks that were needed, but PANDAS gives full control of these data frames and for the rest of the analysis. Below are what the first few rows and columns looked like for each.

A screenshot of a computer

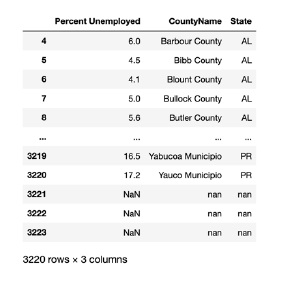
Description automatically generated

Building and Cleaning the Data Frame

The first dataset that I decided to clean was the FBI Crime data frame. First, I dropped the unnamed columns and renamed the Crime Per Population to a more manageable Crime Rate.



Next, I decided to work with the unemployment dataset. Here we were dealing with several unnamed columns that needed to be cleaned up. Unnamed 3 and Unnamed 9 were renamed Count/State and Percent Unemployed respectively. The indexes of 0-2 were dropped to focus on the data itself. The data was coped for scrubbing and cleaning , commas were split from the county name and state, the Percent Unemployed was converted to a float type. Then there were three columns left as seen below.



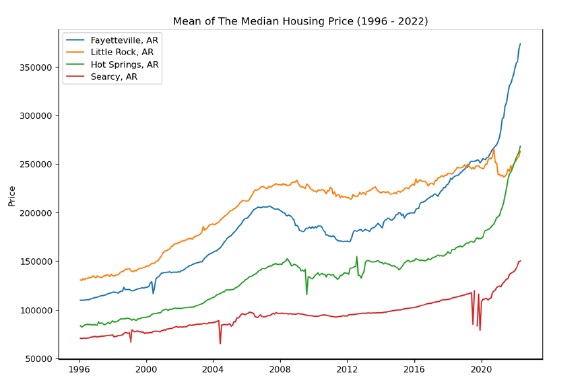
In the Zillow dataset, the columns that were cleaned up in this analysis were City, State, Metro, and County/Name which were all converted to a string type and the columns RegionID, RegionName, and SizeRank to an integer type. Now that all the datasets were ready, they were merged with the bigger dataset from Zillow into one giant dataset including the housing values, Crimes rates per state, and unemployment rate data. Here we want to look at each of the areas and see if there is crime in some of the initial areas and if those relate to other factors such as housing prices for the Real Estate Investment Trust. Overall, the data was paired down to 6,299 rows and 329 columns.

Scrubbing the Data Frame

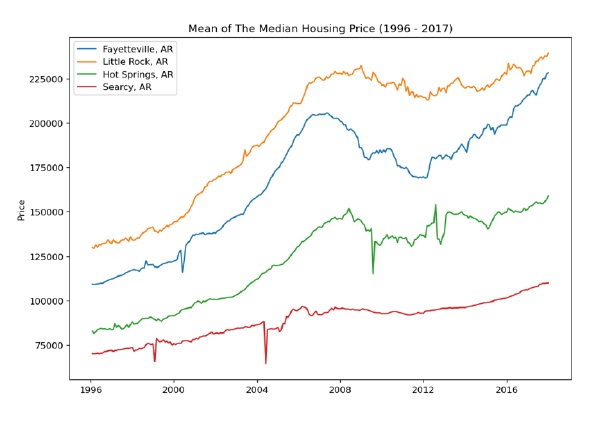
The most important thing that needs to be fixed in the dataset were the columns. In the merged data frame that I created the months and the years in the set needed to be transposed to take the median house price, the average crime rate per state, and the unemployment rate. This was done by creating a function that takes in the data frame, the state, and the city. This copies the data frame and gets the mean of all values and creates the data frame inside the function itself.

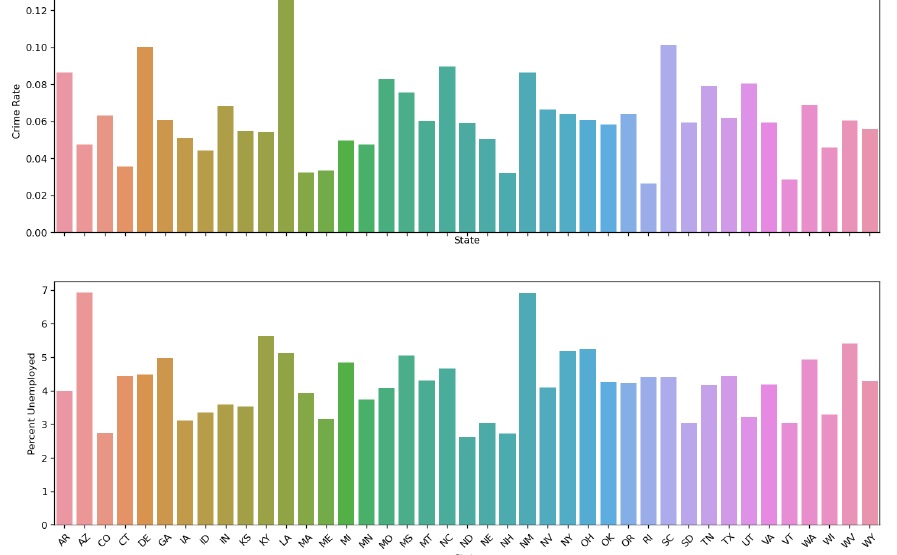
In-Depth Analysis and Recommendation

The first analysis looks at the average median housing price from the years of 1996 to Present for the following cities within Arkansas: Little Rock, Fayetteville, Hot Springs (Where I was born) and Searcy. As we can see in the time series analysis below the prices dipped in 2008 when the housing bubble burst (I bought I house in 2007 so this was real fun) and led to a recession in the United States. You can also see that since 2012, the average median housing price has been steadily increasing. Another change can be seen in 2020 when COVID hit and housing prices skyrocketed. Due to the shelter in place orders and a high demand for housing mixed with people moving out of more urbanized areas, price gouging, and people taking advantage of the market of those trying to buy a home.

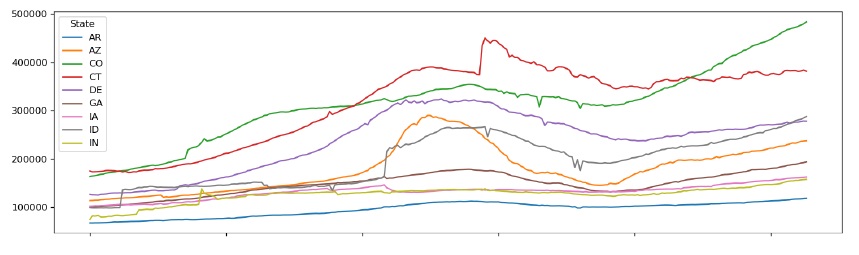


As the purpose of this study is to provide a recommendation for the top three best metro areas for the SERIT to invest in, we will construct an ARIMA Model using our data split into two sets: a test and a train set. As this is a time series and we are trying to make a prediction, we will use the training data as data from 1996 to 2017, and the testing data from 2018. The chart shows that the series is basically the same, and the average median housing price is steadily increasing over time.

 The next step in the process it basically to locate the states and cities within those states that would most likely be some of the best places to live in. Logically, we want to find states that have low crime as well as low unemployment. This would provide both a sense of safety that is valid for the consumer, as well as a good job market and hopefully financial security. Here we examine the states to see the average amount of crimes for the population, and the percentage of unemployment that is in our dataset. Below you can see that there are states with crime rates below 3.5% per capita, and we can see that New Hampshire meets this mark at 3.2%. On the unemployment side one of the lowest we see is North Dakota with a rate of 2.6%.



Job availability and crime rates are not the only things that are important when choosing a place to live. The line chart below we can compare the states to each other based upon the average median housing price per state, and we can see that Colorado has the highest median salary. Therefore, it would make sense that we could sell homes for more in these states leaving a bigger profit opportunity. We will explore Colorado as our third recommendation because it has the highest average housing value in the dataset.



* New Hampshire

A table of crime rate

Description automatically generated

We explored the cities in New Hampshire with the crime rates less that 1%. From that list I selected East Kingston as the first recommendation.

* North Dakota

A screenshot of a white screen

Description automatically generated

We explored the cities in North Dakota with Unemployment Rates of less than 3%. From that list I chose Devil Lake as the second recommendation.

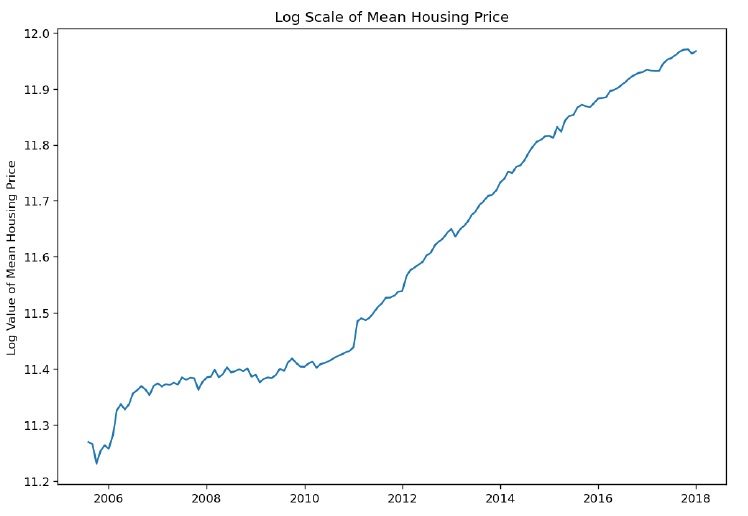
* Colorado

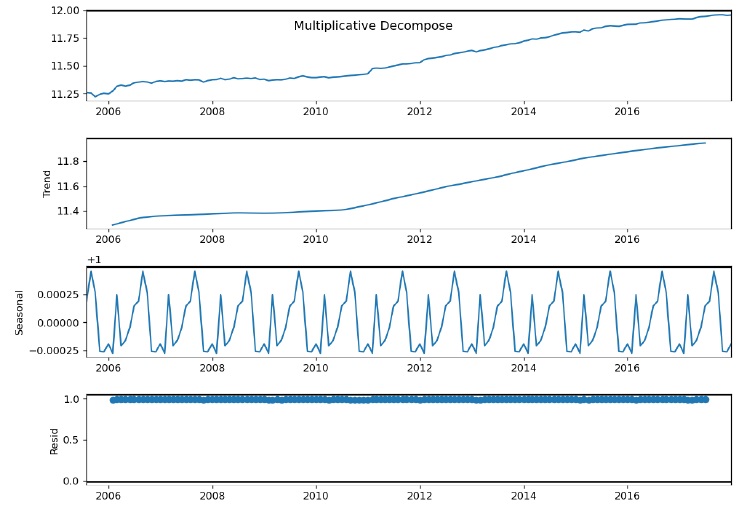
We explored Colorado as it had the highest mean housing price. From that we selected Snowmass Village.

A black text on a white background

Description automatically generatedTo avoid repetition, we will discuss our ARIMA model process using Devil Lake, ND as our recommendation. The other two cities we selected were evaluates with the same process though that you see below.

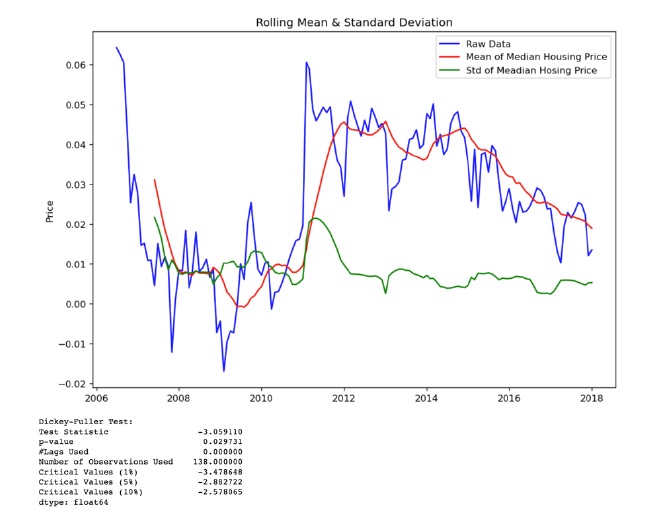
The first step was transforming the data to make it as stationary as possible. This was done by using a logistic scale on the housing values.

 Once the transformation was complete, the data was visually explored the seasonal decomposition of the data below:

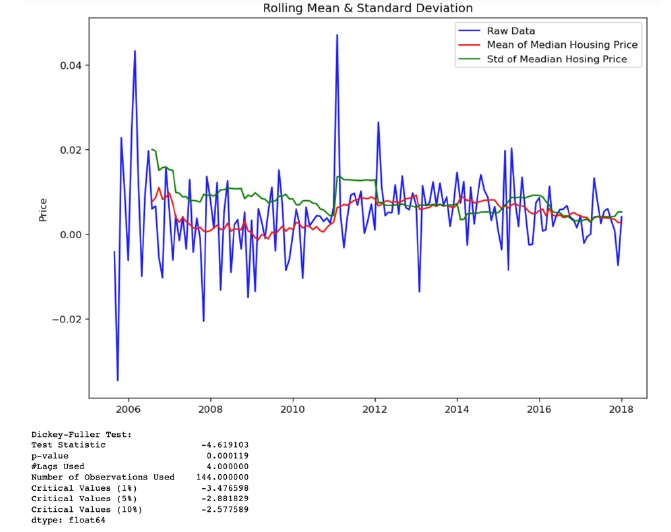


Two methods of differencing were explored and the metho that generated the lowest p-value from the Dickey-Fuller Stationary Test was used in the ARIMA model.

1. Log minus the moving average of the log

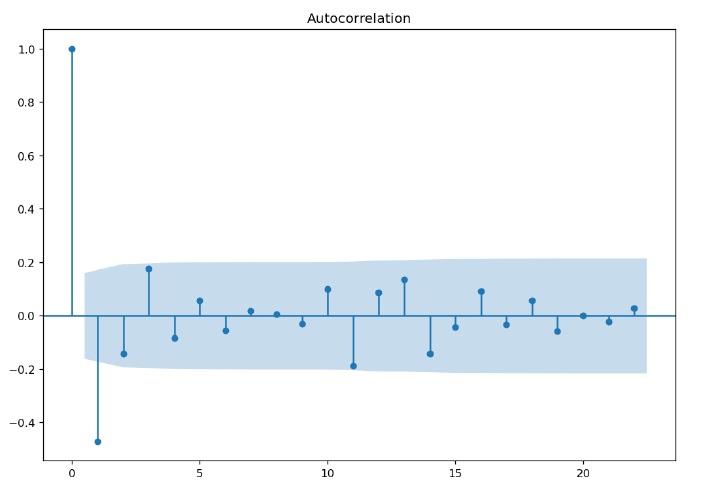


1. Log minus the previous log value

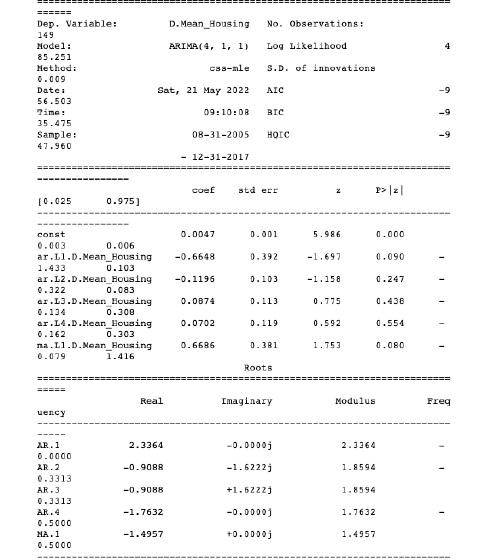


The second method has the lowest p-value of the two and is used to determine the Partial Autocorrelation p-value and the Autocorrelation q-value. The plot of the partial autocorrelation and autocorrelation differenced once d=1.

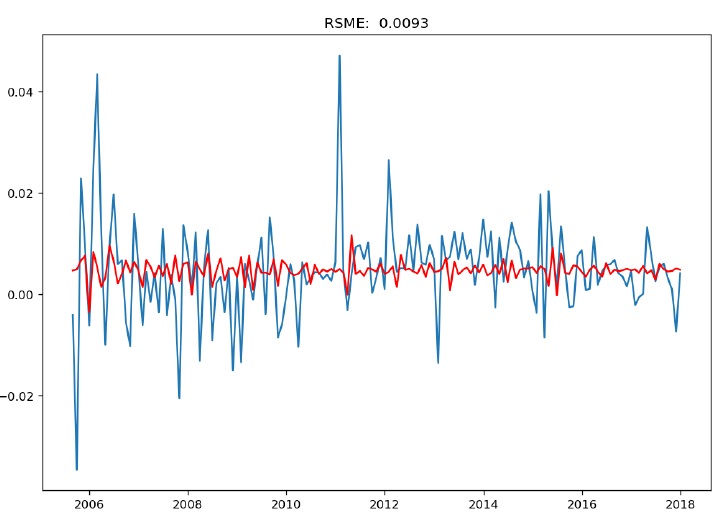
A graph with blue lines and dots

Description automatically generated

The number of lags in the partial autocorrelation is 4 and in our autocorrelation is 1. For our ARIMA model our order will be 4, 1, 1. Below is the results of our ARIMA Model.

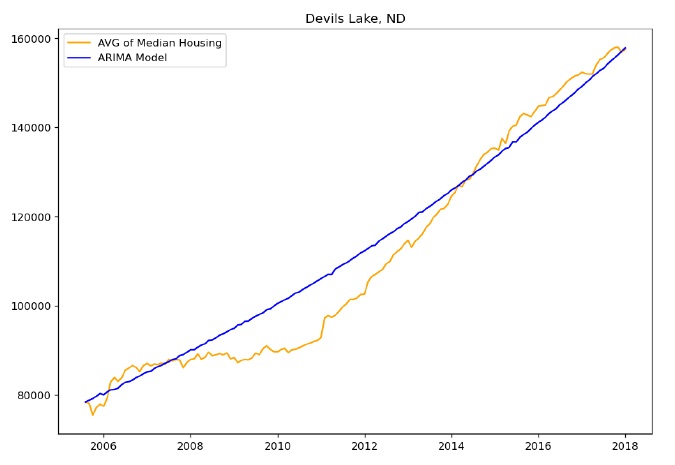


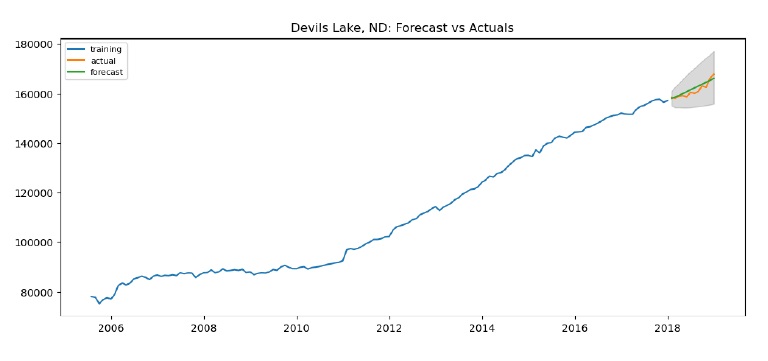
Using the Root Mean Squared Error (RMSE) we can determine the optimal ordering of our ARIMA Model.



When working with data over time, decomposition, is a great way to visualize the general patter of what is happening. But when I looked at the ARIMA model above to male specific predictions using this sophisticated approach. We were able to capture both the general upward trend, as well as the dramatic changes on seasonality, and do it in a way that can make very accurate predictions. In the three areas that we chose, we are able to plot out the ARIMA model, and the average median housing data in the same chart to understand visually how well the model is predicting.

The time series chart below displays how well the ARIMA model for Devil Lake, North Dakota did in the analysis. According to our chart, it shows that the prediction is very close to the real data that we fed into to training set when we compare it to the real data. The line that we see in our ARIMA model and the trend of the real data shows that it is more of a straight line. This shows that the rate of return can be low, but the risk is also low so we can get good returns in this area without carrying the high load of risk.

 In the chart below we wanted to see how confident we can be in our model when trying to predict the 2017 year. In our model below, we can see that our predictions are very close to the actual data fits in our confidence interval.



The next location we are going to look at is Snowmass Village, Colorado. According to the chart below we can see that this ARIMA model over-estimated the data.

A graph showing the growth of a house

Description automatically generated

When we look at the data for Snowmass Village’s confidence interval, the ARIMA Model is in our range and our actual data is also in the range. Showing a higher rate of return, but also a higher ate of risk.

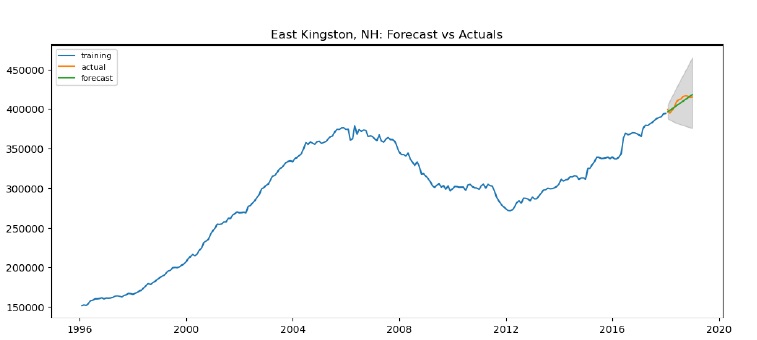
A graph showing the growth of snow glass

Description automatically generated with medium confidence

Finally, for our last area, we want to look at the ARIMA Model for East Kingston, New Hampshire. The lie shows that the chart can be a little unstable and when we looked at this it seems stationary. At the end of the chart, based upon the model, we can see that it Is a medium risk with a moderate return.

A graph showing the growth of housing prices

Description automatically generated And we can see again in this chart below, the ARIMA model and the actual data for East Kingston, New Hampshire is in the confidence interval.



By comparing the actual average median housing prices of 2018 with the forecast values of 2018 from the ARIMA Moel we can calculate the RMSE value. From the RMSE value we can assign a level of risk in using this model to forecast the average median housing price in the recommended cities (High = Lowest Accuracy, Medium = Moderate Accuracy, Low = Highest Accuracy).

|  |  |  |
| --- | --- | --- |
| Risk/Reward | Recommendation | RMSE |
| High Risk/High Return Value | Snowmass Village, CO | 70,903 |
| Medium Risk/Moderate Return Value | East Kingston, NH | 3,772 |
| Low Risk/Low Return Value | Devil Lake, ND | 1,426 |

Conclusions and Final Thoughts

With the employment rate data set we were able to create a brand new data frame, test and train the data for prediction on where to invest, and assign a risk level to the models based on its accuracy and can help the SEIT with future investment ventures. We were able to use the ARIMA Model to forecast where the data seems to trend, see how well the data was doing, and place a confidence interval for our ARIMA model to see if what was predicted was the best place to invest. We can recommend choosing these three areas based on this data to make an investment.