Kwasi Mensah

IST 718

Big Data Analytics

Lab Exercise 2

**Objective:**

1) Obtain data and understand data structures and data elements.

2) Scrub data using scripting methods, to include debugging, for data manipulation in R and other tools.

3) Explore data using essential qualitative analysis techniques including descriptive statistics.

4) Model relationships between data using the appropriate analytical methodologies matched to

the information and the needs of clients and users.

5) Interpret the data, model, analysis, and findings. Communicate the results in a meaningful way.

**Introduction**:

This exercise provides an opportunity to demonstrate our ability to combine data sets and produce meaningful analysis. We will use the datasets that were given and any that were manually acquired to try to forecast the best zip code for SREIT to invest in the Real Estate Market in the United States. We will try to accomplish this by using the ARIMA model as making use of RMSE (Root Mean Squared Error) well as. In addition, we will also take a closer look at the state of Arkansas specifically and look at the housing value in the state when looking at four specific metro areas.

**Question:**

“How can we recommend what would be the top three zip codes that provide the best investment opportunity for the Syracuse Real Estate Investment Trust (SREIT) by way of prediction through forecasting?” by use of time series and model forecasting for predictions.

**Data Sets Used for Analysis (Data Obtained):**

The data frames below are a part of two separate python notebooks due to system memory constraints, when trying to create the times series forecasting and finding the RMSE for prediction purposes.

*‘df’ Data Frame (29532 Rows x 336 Columns)*

* This is our main data frame for our lab exercise. This data set was given to us in our course but could be accessed at the Zillow website under the Single-Family Residence data set categorized by metro in the US. This set has the median housing value of each zip code in the us dating back to 1996 to April 2023.

*‘df\_labor’ Data Frame (3142 Rows x 10 Columns)*

* This data frame is from the U.S Bureau of Labor Statistics, specifically the Labor Force for the year 2022. This dataset provides us information on the unemployment rates in the US categorized by county.

*‘zwar’ Data Frame (27176 Rows x 7 Columns)*

* This transformed data frame is used for analysis of the Arkansas metro area’s median housing value specifically from the year 1997 to present*.*

*‘zw’ Data Frame (11930 Rows x 326 Columns)*

* This is the combined data frame of df\_labor and df after data scrubbing was performed. This data set has Zillow data combined with the unemployment rate that were able to be matched by like county names.

*‘df\_invest ‘Data Frame (6851 Rows x 265 Columns)*

* This is the data frame created that has the median housing value for the year 2018 and on, starting with 1997. This data frame has the zip code that remained after filtering out the zip codes with higher than the national average unemployment rate, with each zip code as the row index and each year and month as the columns.

*‘df\_bl’ Tuple (264 Rows x 1 Columns)*

* This is the transformed data frame for our baseline time series analysis model. This data set has the years of 1997-2017 and the average housing value for that respective month for the specified year.

*‘df\_ts’ Data Frame (316 Rows x 6851 Columns)*

* This is the data frame created for our model that forecasts the average median housing value for the year 2018 and on, starting with 1997. This data frame has the zip code that remained after filtering out the zip codes with higher than the national average unemployment rate, with each zip code as a column and each year and month as the rows.

**Data Cleaning (Scrub):**

First, I looked for the total number of null values in our main data set from Zillow. When this was performed, I found that the number of null values: 2068145. Because we noticed that more than 50% of the data frame contained null values, we decided not to fill the missing values by using the mean of the housing value or use the ‘bfill’ function provided in python through pandas as a method of backward filling in the missing values. When the nulls were dropped, we remained with a data set with 11589 rows. We next renamed the ‘RegionName’ column to ‘ZipCode’ and dropped the additional state column that appeared in the data set when pulling it from Zillow through use of a python IDE. The ‘RegionType’ column was also removed because it only had 1 identical observation throughout representing that the region are ‘zip’ or zip codes. We then removed any columns that have the year 1996 in them to reduce noise when creating models because we are focused on analyzing the data starting from 1997 and on. For the data frames created that focused on the zip codes specifically a function was created in python to convert them to a string and add the number ‘0’ if the zip code has less than 5 numbers. Lastly a function was created to remove the ‘,’ and state abbreviation from the county names in both the labor and the Zillow datasets and removed the month from the years in certain data frames when focusing on zip codes.

Another function was created to convert the county names in the “CountyName” column to lowercase so that merging data will be easier perform by like column. In addition, the function removed any empty spaces in the observations and removed the terms like ‘city’ and ‘municipality’ to match the format of the county names in the dataset collected from the BLS website. Next to further scrub our data set I used the unemployment rate data set collected from BLS to find the national unemployment ratee. Because we are looking for key locations for SREIT to invest in, I felt it best to remove any counties with an unemployment rate that is lower than the national average which was about 3.67%. This down sampling method helped to improve the processing power when trying to create models. After this down sampling was performed, we have a total of 6851 zip codes remaining for our model prediction.

**Exploration:**

When we look at the times series created for the Arkansas Metro Area, we see that the average housing value overall for the metros is below.

**A picture containing text, font, screenshot, line

Description automatically generated**

We also created a visualization depicting the Housing Values and how they spread for each respective metro area. This graph helps us to see why average housing value overall for these metro areas are not too far off from each other in price.

A picture containing diagram, plot, screenshot, text

Description automatically generated

We built an initial baseline model from our general time series dataset created in our python notebooks. We took the average of the median housing value for each zip code. Then we created a for loop to perform a simple grid search to find the best coordinates for our ARIMA model. We then created a training and testing set that has 252 rows and 1 column, and 12 rows, and 1 column in the testing set. This training set will be used to create our simple ARIMA model for analysis. When this was performed, we found that the best combinations with the lowest RMSE score was (5,0,5). This optimal coordinate is clear to see when comparing the other coordinates in graphical form. These specific coordinates gave us a prediction closest to the actual value.

**A picture containing text, font, screenshot

Description automatically generated**

With the baseline ARIMA model created we developed two visualizations. The first is the One-Step Forecast for using the training set (1997-2017) and predicting how will the housing prices increase in 2018. We see that it is predicted that the market will increase in value for homes in the US. Next, we see a similar visualization, but it is the 500-step forecast predicting the housing values in the future. We see that the graph does correctly predict that with the current housing market will continue to increase then it shows a drop most likely due to the number of observations used to perform this prediction.

A picture containing text, plot, diagram, screenshot

Description automatically generated A picture containing text, plot, diagram, line

Description automatically generated

When trying to make predictions to find the best zip code to invest in we created a training and testing set 264 rows and 6851 columns in the training set and 12 rows and 6851 columns in this testing set. When we ran the for loop to find the RMSE for each zip code using the optimal coordinates we got from our base line model, we were able to get 6825 out of 6851 observation’s RMSE which was then put into a new data frame and csv file because this section took over 2 hours to run due to memory constraints on my system. We will use this to perform our AIRMA model for prediction due to the library our system was able to successfully use only taking 1 argument.

**Interpretation:**

For our initial Exploratory analysis, we decided to focus on the state of Arkansas. When looking at Arkansas we will look at the metro areas of the Hot Springs, Little Rock, Fayetteville, Searcy Arkansas. Frist an initial Time Series plot was created to get a view of Arkansas Housing value change over time. We see below that the median housing value has drastically increased over time and the time series increases in a positive direction. We also see that between 1997 and 2018 there are many time periods where the median housing value remained stagnant. Lastly, we see noticeable jumps in housing prices after 2005 and 2016, and a drop in 2008 when the Great Recession occurred, and the housing market crashed.

A picture containing text, diagram, screenshot, plot

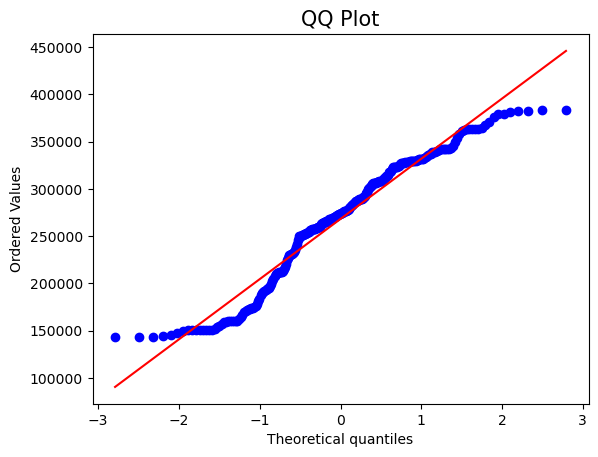
Description automatically generated

Next, we will look at our time series created for each of the four metros being observed below.

A picture containing text, plot, diagram, line

Description automatically generated

We see in this time series that Fayetteville Arkansas had the biggest increase in median housing value followed by Little Rock, then Hot Spring and in last is Searcy which also has the least number of observations among the four counties. We can see for all four of the times series that there is an upward trend. We would want to further research to see what factors went into the housing value of the homes in the capital of Arkansas (Little Rock) maintaining the highest housing value throughout different time periods, and why Fayetteville now has a significant difference in housing value when compared to the other three metros.

Next, we will begin modeling for our prediction for the SREIT to find the best locations the investment firm should make their next real estate investments. 

Fist we look at the QQ Plot created to check if our baseline time series created is normally distributed. We see that the points for out time series are mostly going in an upward direction which tells us that our baseline time series is not normally distributed in one direction. Next, we will look at our ACF & PACF charts created. We see in our ACF chart that effects deteriorate slowly meaning that there is a high correlation between our observations in the time series and the past values in the time series. When we look at the PACF chart we see that there are lags at 3,4,5,6,7,8, 11, & 15, these lags let us know we could use this baseline model for our ARIMA model.

A picture containing text, screenshot, line, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generated

When we look at the ADF Statistics below we see that we get a p-value of .484794, which is above our threshold of .05, telling us that we cannot reject the null hypothesis. As a result, can now determine that our time series is non-stationary, and we do not need to perform differencing for our time series to become so.

A picture containing text, font, screenshot, algebra

Description automatically generated

When we use our previously found best coordinates, we see that 5,0,5 gives us the best Actual vs. Predicted chart closest to the actual housing value in the US for 2018. We will see below the top 3 coordinates A vs P charts created to show why the best coordinates were chosen.

(9,0,3)

A picture containing text, plot, line, screenshot

Description automatically generated

(7,0,2)

A picture containing text, plot, screenshot, line

Description automatically generated

(5,0,5)

A picture containing text, plot, line, diagram

Description automatically generated

We can see that the Actual vs Predicted for (5,0,5) line graph follows the most similar pattern. Below are the descriptive statistics of our ARIMA model created for our base line model.

A screenshot of a computer

Description automatically generated with low confidence

Lastly, we used our coordinates of 5,0,5 to create multiple ARIMA models for our transformed data frame that has the zip codes as the columns and the years and months as the index. We build an ARIMA model for each respective zip code then find the RMSE for each zip code using a for loop which can be seen below. We did this to reduce the time needed to run the model on each zip code individually.

A screenshot of a computer program

Description automatically generated with medium confidence

We then saved the output in a new csv file and data frame and combined it with the merged data frame that had the housing values combined with the unemployment rates. When doing research, we learned that the ‘Size Rank’ column has to do with how urban an area is, which may be key information for our stakeholder when deciding on where to invest with our recommendations. When we grouped our new data frame by the RMSE, Size Rank and unemployment rate we got the following as our top 5 recommendations for real estate investment, which can be seen below.

A picture containing text, screenshot, font

Description automatically generated

**Conclusion**:

We can recommend that the Syracuse Real Estate Investment Trust (SREIT) invest in the states of Texas with an unemployment rate of 2.8%, Georgia, with an unemployment rate of 2.7% and Tennessee with an unemployment rate of 2.8%, giving confidence to our investors that these states have lower unemployment rate meaning there is a higher chance of getting a return on their investment. If specifically looking at counties Travis County, TX 78660, Gwinnett County, GA 30044, and Davidson County, TN 37013. Respective they all have a size rank (the lowest being a larger urban population and area) of 4, 20, 25 making them fall in the top 25 of the size rank in the US. In addition, each of the top 3 have very low RMSE scores giving us confidence there is better accuracy when looking at the performance of our prediction.

Although it is very time consuming to create the models and clean the data so that it is usable for time series, it is important to use other modeling techniques to compare the performance of the models when making accurate predictions. Due to system constraints on my machine, I was unable to perform different modeling techniques. I kept getting errors when trying to gradient boost my decision trees that were created, and when trying to use Facebook ‘fbprophet’ I kept getting errors trying to install the package, and when manually downloading the tar file it was unable to load into my environment. In the future I would like to do more research to find out how I could successfully test these alternate methods and compare the results if these packages are not already retired.

**References**:

[**Real Estate Investing in Texas. Can you predict housing prices in Texas… | by Hazel Donaldson | Towards Data Science**](https://towardsdatascience.com/real-estate-investing-in-texas-bbe794df1613#:~:text=It%20also%20contained%20a%20variable,the%20higher%20the%20size%20rank.)

[**Time Series Analysis on Zillow’s Housing Data | by Fernando Aguilar | Medium**](https://medium.com/@feraguilari/time-series-analysis-modfinalproyect-b9fb23c28309#:~:text=The%20data%20contains%20descriptive%20variable%20indicating%20the%20zip,More%20urbanized%20zip%20codes%20have%20a%20higher%20rank.)

[**Time series / date functionality — pandas 2.0.1 documentation (pydata.org)**](https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html)

[**A Guide to Time Series Forecasting in Python | Built In**](https://builtin.com/data-science/time-series-forecasting-python)

[**The Complete Guide to Time Series Analysis and Forecasting | by Marco Peixeiro | Towards Data Science**](https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775)

[**A Guide to Time Series Forecasting with ARIMA in Python 3 | DigitalOcean**](https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3)

[**A Guide to Time Series Visualization with Python 3 | DigitalOcean**](https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-visualization-with-python-3)