Lab Exercise 2

Bing-Je Wu

IST 718 – big data analytics

Table of Contents

[Introduction 2](#_Toc32326047)

[Dataset 2](#_Toc32326048)

[Methods for Analysis 2](#_Toc32326049)

[Data and Data Processing 2](#_Toc32326050)

[Exploratory Analysis 3](#_Toc32326051)

[Data Modeling – AR Model and ARIMA Model 4](#_Toc32326052)

[The Naïve Model – Baseline 4](#_Toc32326053)

[ARIMA Model – Part 1 (General Time Series) 6](#_Toc32326054)

[ARIMA Model – Part 2 (Time Series for each Zip Code) 7](#_Toc32326055)

[Final Model 8](#_Toc32326056)

[Questions 9](#_Toc32326057)

[Conclusion 9](#_Toc32326058)

[Reference 9](#_Toc32326059)

# Introduction

Real estate is an asset form with limited liquidity relative to other investments. Real estate investing involves the purchase, ownership, management, rental and/or sale of real estate for profit. If these factors are not well understood and managed by the investor, real estate becomes a risky investment. The lab exercise is to build ARIMA models to forecast the future median price of single-family house by zip codes. There are few questions we are going to answer during this lab exercise:

1. Provide an initial data analysis of the following Arkansas metro areas: Hot Springs, Little Rock, Fayetteville, Searcy
2. What technique/algorithm/decision process did you use to down sample?
3. What three zip codes provide the best investment opportunity for the SREIT? Why?

# Dataset

This lab exercise contains 2 datasets from different sources: (1) The SingleFamilyResidence dataset is provided by the instructor as the main dataset for the entire lab exercise (2) The Labor Force Data by County (2018 Annual Averages) dataset is from the U.S Bureau of Labor Statistics as supplement for providing the information of unemployment information by county. (1. <http://files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv>) (2. <https://www.bls.gov/lau/laucnty18.xlsx>).

# Methods for Analysis

A baseline model, simple AR model, was conducted by using the general time series data, taking average of the house price for the candidate zip codes, as a benchmark standard in high level to compare the performance of ARIMA models for general time series data. The grid search technique was implemented on building optimal ARIMA model for general time series data in high level. With the optimal combination of the order (p, d, q), 7,425 ARIMA models were trying to be built for each of the zip code time series data and 5,748 of ARIMA models for zip codes were built successfully. A list of best opportunities for real estate table was sorted by the size rank factor, unemployment rate factor and root mean square errors (RMSE) which is generated from each zip code ARIMA model with the combination of the order (p,d,q) of the general time series ARIMA model. The top three zip codes were selected as the potential opportunities for real estate investment and new ARIMA models were built for finding the best combination of the order (p,d,q) for three of the zip code time series.

# Data and Data Processing

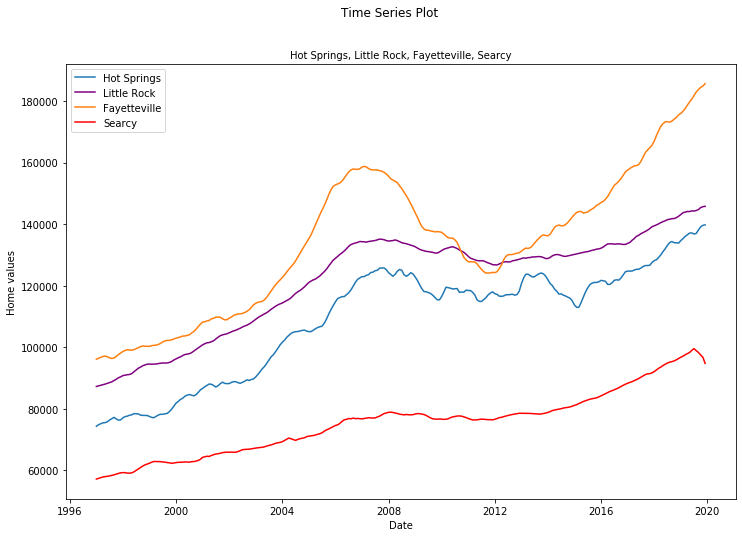
After importing the five datasets, some cleaning steps need to be performed. For the SingleFamilyResidence dataset, the data should be normalized, such as transforming the words were used in ‘CountyName’ column, converting uppercase to lowercase, removing leading or trailing spaces and unwanted string, and changing the ‘RegionName’ column from numeric datatype to string datatype and padding 0s if the length of strings is not equal to 5 as a complete zip code format. To answer the research questions using the data from 1997 to 2018, zip codes with missing values in the time frame were excluded. That left 13,916 zip codes in the main dataframe, called as df\_1997.

The employment information as another indicator can be used on real estate investment in order to reduce the risk of losing money. A good real estate investing strategy is to find their targeting market. Lower unemployment rate means there are more demand of buying houses, renting an apartment. Thus, the Labor Force Data by County (2018 Annual Averages) dataset, called as df\_unemployed\_rate was used to join with df\_1997. The rule for join is as following:

1. df\_1997 dataset left join with df\_unemployed\_rate dataset on county name and state

By having the complete dataset and combined features, another filter can be applied to get the candidate zip codes for real estate investment. An average unemployment rate of all county in the united states, 3.98, was used as the filter standard to exclude the zip codes that have unemployment rate higher than the average in national level. After filtering out the high unemployment rate, there were 7,425 zip codes left as candidate zip codes for real estate investment.

# Exploratory Analysis

Hot Springs, Little Rock, Fayetteville, Searcy Metro Area in Arkansas metropolitan areas were selected for analysis. The four metro areas have 5 zip codes, 49 zip codes, 31 zip codes and 12 zip codes respectively. Each of metro area took the average of the median price for single-family house of zip codes. A time series plot for the four metropolitans in average price was shown below. 

The time series plot shows that there are clear upward trends for each metropolitan. The overall price of single-family house is higher at Fayetteville metro area. Fayetteville metro area is the home to the University of Arkansas. The college town brings up the economic in that area. The population started to growth in 2010. It also provides some of the opportunities for business and career in 2016. Therefore, we can see, with the growing population and business effort, there is a clear upward trend after 2012. The great recession at the year of 2008 affected this area a lot which brings the dramatic decrease in housing market after 2008. Companies were cutting their capital and young graduate had hard time to find a job. Therefore, the housing price market declined drastically during that time. And it took a while to get back the market and keep growing until 2019. This place could be an option for investor to consider investing on the real estate in this area but with higher risk.

Little Rock metro area is centered by Little Rock city which is the capital and most populous city of the Arkansas state. With the capital having diversity cultural, prosper in economic, government, and transportation center, the metro area has steady grown during the past 20 years. The great recession seems to have not too much impact on that area. Thus, it can be a perfect place for real estate investment but with higher cost.

Hot Springs is a resort city in the state of Arkansas. Also, it has hot springs Hot Springs National Park in that area. With many sightseeing spots in the area, the market of single-family house is growing in overall. The variation of the housing price can be explained by the business nature of the area. The great recession in the year of 2008 might have affected the housing price market.

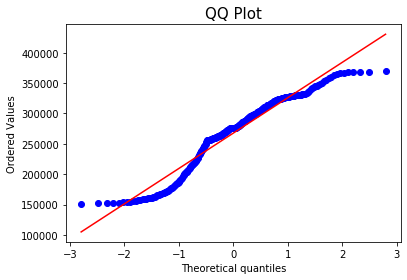
The lowest one is Searcy metro area. It does not have too much special factor that can be compared with other three area. Therefore, the overall hosing price is low but with a steady growth trend. The great recession seems to have no impact at all for the Searcy metro area. It can be a good option for conservative investors wih lower captial.

# Data Modeling – AR Model and ARIMA Model

Models were built by implementing the train-test split technique. 252 observations at the year range from 1997 to 2017 were used as training set. 12 observations at the year of 2018 were used as test set.

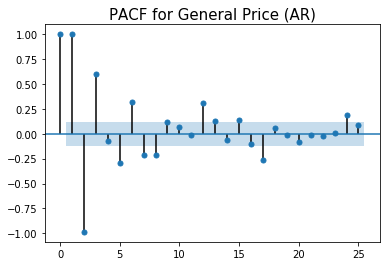
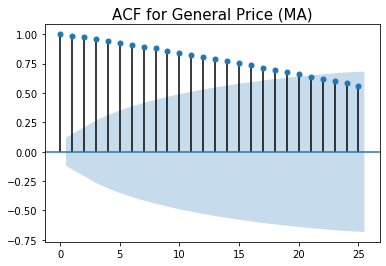
## The Naïve Model – Baseline

An AR model for general time series data was built as a benchmark standard for the order (p, d, q) selection in grid search. The general time series data is the average of combination of 7,425 zip codes. The general time series has been checked if it follows the normal distribution by examining the QQ Plot as shown below.



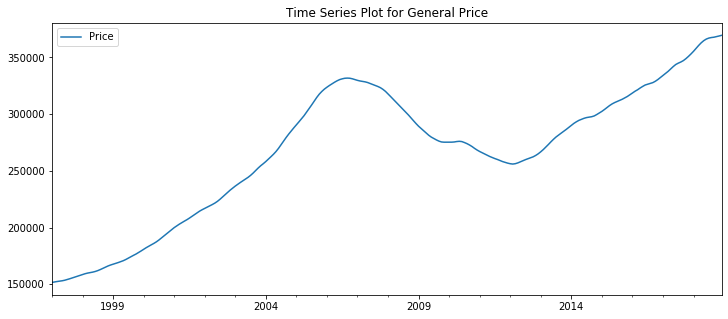
The plot shows that the 7,425 data points does not align with the diagonal line indicating that the general time series is not following normal distribution. It also implies that we cannot use the elegant statistics of normal distributions to make successful forecasts however this is what we usually expect from time series data and we can move forward.

The ACF plot and PACF plot for the general time series data are shown below.



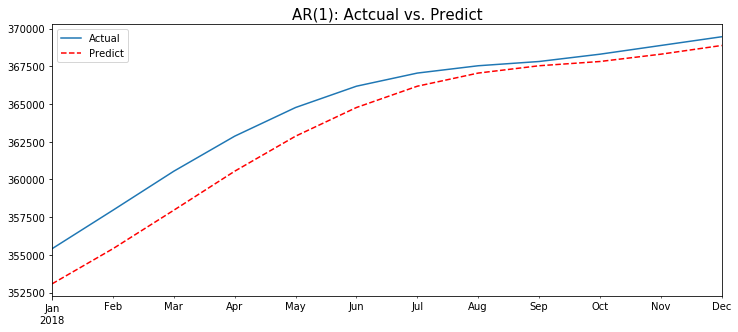
The ACF plot shows the effects deteriorate slowly. The PACF shows that there many significant lags, such as Lag 1, 2, 3, 5, 6, 7, 8. And this also indicates that the general time series can be used for AR model. The Augmented Dickey-Fuller unit root test proves that the time series is non-stationary. Since the simple AR model is used as a benchmark standard, we do not need to perform differencing to convert the general time series to stationary as the requirement of AR model. The issue of stationary can be resolved by implementing the integration in ARIMA model.

|  |
| --- |
| ADF Statistic: -2.086293  p-value: 0.250062  Critical Values:  1%: -3.457  5%: -2.873  10%: -2.573  \*\* Fail to reject the null hypothesis. Data has unit root and is non-stationary \*\* |



|  |
| --- |
| The baseline RMSE: 1587.984 |

The AR(1) model , using general time series data with pandas shift method, yielded the RMSE at 1587.984 as the benchmark standard to select the optimal combination of order (p, q, d). The time series actual vs predict plot in 2018 for AR(1) model is shown below. We can see that AR(1) perform pretty well on the prediction of price in 2018



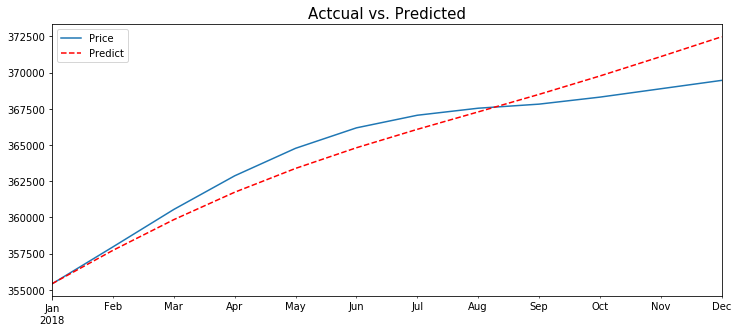
## ARIMA Model – Part 1 (General Time Series)

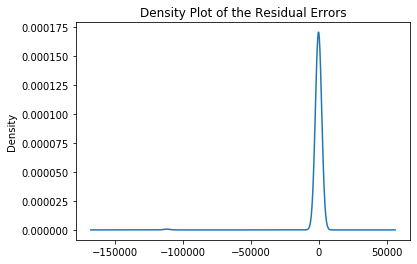
An optimal ARIMA model was built for the general time series in high level by using grid search techinque. The snapshot of the output is shown as below. The best combination of order (p, d, q) = (5, 0, 4)

|  |
| --- |
| .  .  ARIMA(4, 2, 3), RMSE=5376.0509  ARIMA(5, 0, 0), RMSE=1593.8  ARIMA(5, 0, 1), RMSE=1576.2328  ARIMA(5, 0, 2), RMSE=1944.8953  ARIMA(5, 0, 4), RMSE=**1391.8805** . . |

|  |
| --- |
| ARMA Model Results  ==============================================================================  Dep. Variable: Price No. Observations: 252  Model: ARMA(5, 4) Log Likelihood -1588.638  Method: css-mle S.D. of innovations 125.090  Date: Mon, 10 Feb 2020 AIC 3199.277  Time: 14:46:27 BIC 3238.101  Sample: 01-01-1997 HQIC 3214.899  - 12-01-2017  ===============================================================================  coef std err z P>|z| [0.025 0.975]  -------------------------------------------------------------------------------  const 2.631e+05 2.86e+04 9.212 0.000 2.07e+05 3.19e+05  ar.L1.Price 3.4740 0.000 2.88e+04 0.000 3.474 3.474  ar.L2.Price -4.5629 0.000 -3.34e+04 0.000 -4.563 -4.563  ar.L3.Price 2.4844 0.000 1.1e+04 0.000 2.484 2.485  ar.L4.Price -0.1824 0.000 -1503.327 0.000 -0.183 -0.182  ar.L5.Price -0.2134 0.000 -635.441 0.000 -0.214 -0.213  ma.L1.Price -0.6965 0.035 -19.897 0.000 -0.765 -0.628  ma.L2.Price 0.2938 0.037 7.906 0.000 0.221 0.367  ma.L3.Price -0.7238 0.042 -17.407 0.000 -0.805 -0.642  ma.L4.Price 0.9650 0.034 28.619 0.000 0.899 1.031 |

The optimal model is actual the ARMA (5, 4). Surprisingly, integration was not used in the model for the better performance.

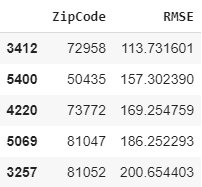




The line plot of the residual errors has a clear pattern around 0, and the density plot of the residual errors shows that the residual is skewed to the left. Those indicators imply that the residual is not white noise and the prediction outcome is not perfect.

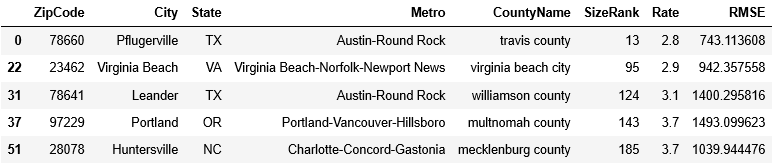
## ARIMA Model – Part 2 (Time Series for each Zip Code)

7,425 zip codes were used to build ARIMA models with the order (p, d, q) = (5, 0, 4). There were 5,748 models were successfully built. And the top five zip codes with lower RMSE are 72958, 50435, 73722, 81047 and 81052. However, we cannot consider only the RMSE alone as our standard to pick a place for real estate investment. Other factors should be included as well, such as size rank, unemployment rate, etc.…



## Final Model

To find the top three zip codes for real estate investment, we need to combine the factor of the size rank, unemployment rate and RMSE together. A table which is sorted by ‘SizeRank’, ‘Rate’ and ‘RMSE’ is shown below. From the table, we can see the top three zip codes are 78660, 23462 and 78641. Models for those three zip codes were adjusted to find if there were better combination of order (p, d, q) for each zip code.



Three models were able to find the better combination of order (p, d, q). Zip code, 78660, 23462 and 78641 have the order of (2, 0, 4), (2, 0, 1) and (4, 0, 4) with the RMSE at 726.5509, 425.9606 and 1513.9206 respectively.

|  |  |
| --- | --- |
| zipcode = '78660'; ARIMA(2, 0, 4) | zipcode = '23462'; ARIMA(2, 0, 1) |
| zipcode = '78641'; ARIMA(4, 0, 4) |  |

# Questions

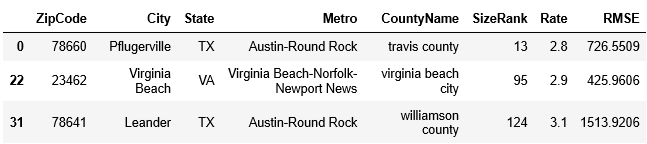
1. **Provide an initial data analysis of the following Arkansas metro areas: Hot Springs, Little Rock, Fayetteville, Searcy?**

The time series plot shows that there are clear upward trends for each metropolitan. Fayetteville metro area has higher variation but with strong growing trend. It is a place for investor who is willing to take at higher risk with higher return of investment. Little Rock metro area and Searcy metro area have lower variation and steady growth trends. A lower risk taker of investor can invest real estate on both areas depending on the level of capital. Hot Springs metro area might not be good choice of real estate investment for a conservative investor because of the nature of business in that area.

1. **What technique/algorithm/decision process did you use to down sample?**

First, I excluded the zip codes with missing data from 1997 to 2018. It had 13,916 zip codes remained. Then, I merged the unemployment information based on the county. Using the unemployment feature to exclude zip codes that have unemployment rate higher than the average unemployment rate, 3.98. After filtering out the unwanted zip codes, there were still 7,425 zip codes left as our candidates. Then, by running the ARIMA models for each zip code with the order (p, d, q) gathered from the general time series ARIMA model, there were 5,748 as our candidates for real estate investment.

1. **What three zip codes provide the best investment opportunity for the Syracuse Real Estate Investment Trust(SREIT)? Why?**

****

Zip codes, 78660, 23462, 78641 are the best opportunity for the SREIT. Its' unemployment rates are way below average, 3.9, and have low RMSE (after adjusting the order of p, d, and q), indicating that the model is predicting better. Two zip codes, 78660 and 78641, are near the Austin metro area which is the place that has a diverse mix of government employees, college students, musicians, high-tech workers, blue-collar workers, and a vibrant LGBT community, also known as Silicon Hill. For the zip code, 23462, it is located between Virginia Beach and Norfolk city. Many of people may choose the place that is between the workplace and relaxing place. On top of that, having the models with low root mean square error gives the investor more room on profit. Thus, those three zip codes are the best opportunity for the SREIT.

# Conclusion

Prediction time series analysis could be time consuming. There are many need to be considered while building a model. Other Model can be implemented in the future such as SARIMAX, ARIMAX to cover the factor of seasonality, and exogenous factors. Prophet can be applied when the holiday factor is considered.

# Reference

Real estate investing from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Real_estate_investing) ; 2 Hot Springs, Arkansas from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Hot_Springs,_Arkansas)  
3 Little Rock, Arkansas from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Little_Rock,_Arkansas) ; 4 Fayetteville, Arkansas from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Fayetteville,_Arkansas)  
5 Searcy, Arkansas from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Searcy,_Arkansas) ; 6 Austin, Texas from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Austin,_Texas)