**Syracuse University**

**IST-652 Final Project Report**

ThulasiRam RuppaKrishnan

IST 652

Professor Jason Anastasopoulos

Contents

[Project Goal 3](#_Toc26740863)

[Introduction 3](#_Toc26740864)

[Analysis and Models 5](#_Toc26740865)

[**About the data** 5](#_Toc26740866)

[**Models** 13](#_Toc26740867)

[Results 20](#_Toc26740868)

[Conclusion 25](#_Toc26740869)

## 

## **Project Goal**

The goal of the project is to predict the Zillow home value index and find out the zip codes that are great for investment opportunity

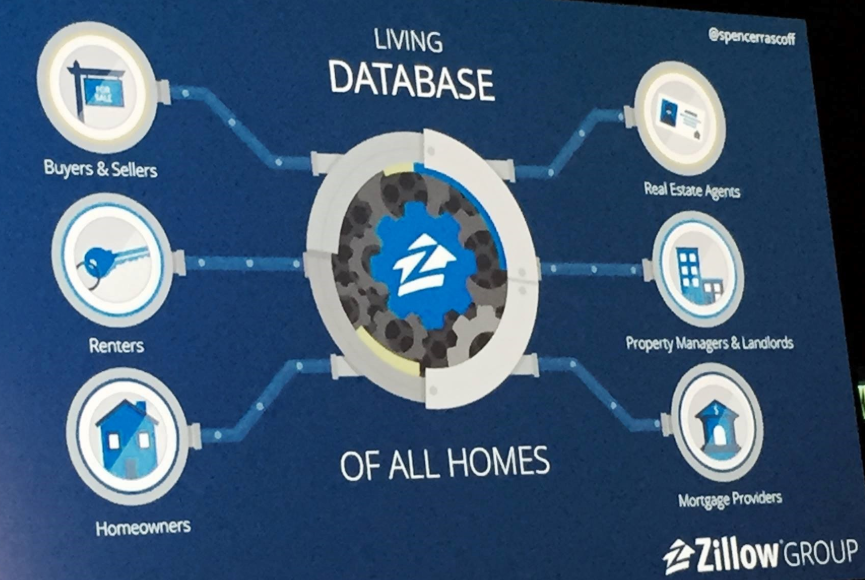
## **Introduction**

Buying and owning real estate is an exciting investment strategy, that can be both satisfying and lucrative. Unlike stock and bond investors, prospective real estate owners can use leverage to buy a property by paying a portion of the total cost up front, then paying off the balance, plus interest, over time. While a traditional mortgage generally requires a 20% to 25% down payment, in some cases, a 5% down payment is all it takes to purchase an entire property



Real estate appraisal, property valuation or land valuation is the process of developing an opinion of value, for real property (usually market value). Real estate transactions often require appraisals because they occur infrequently and every property is unique (especially their condition, a key factor in valuation), unlike corporate stocks, which are traded daily and are identical (thus a centralized Walrasian auction like a stock exchange is unrealistic). The location also plays a key role in valuation. However, since property cannot change location, it is often the upgrades or improvements to the home that can change its value. Appraisal reports form the basis for mortgage loans, settling estates and divorces, taxation, and so on. Sometimes an appraisal report is used to establish a sale price for a property.

Zillow Group, Inc., or simply Zillow, is an American online real estate database company that was founded in 2006



Zillow has data on 110 million homes across the United States, not just those homes currently for sale. In addition to giving value estimates of homes, it offers several features including value changes of each home in a given time frame (such as one, five, or 10 years), aerial views of homes, and prices of comparable homes in the area. It can access appropriate public data, it also provides basic information on a given home, such as square footage and the number of bedrooms and bathrooms. Users can also get current estimates of homes if there was a significant change made, such as a recently remodeled kitchen. Zillow provides an application programming interface (API) and developer support network.

Real estate is generally a great investment option. It can generate ongoing passive income and can be a good long-term investment if the value increases over time. You may even use it as a part of your overall strategy to begin building wealth.

## **Analysis and Models**

### **About the data**

(files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_SingleFamilyResidence.csv)

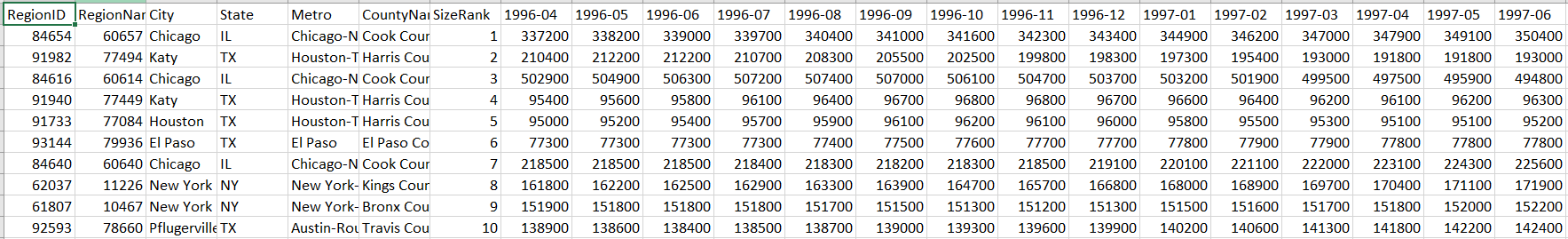
The dataset (Zip\_Zhvi\_SingleFamilyResidence.csv) is a comma-separated files with the following variables

|  |  |
| --- | --- |
| **RegionID** | Unique Identifier for a Region |
| **RegionName** | Region identified by zip code |
| **City** | City identified by zip code |
| **State** | State identified by zip code |
| **Metro** | Metro identified by zip code |
| **CountyName** | County identified by zip code |
| **SizeRank** | Size identified based on population |
| **1996-04** | Zhvi by zip code and year month |
| **1996-05** | Zhvi by zip code and year month |
| **1996-06** | Zhvi by zip code and year month |
| **. . .** | Zhvi by zip code and year month |
| **2019-08** | Zhvi by zip code and year month |
| **2019-09** | Zhvi by zip code and year month |

**Home Values (ZHVI)**

Zillow Home Value Index (ZHVI): A smoothed, seasonally adjusted measure of the median estimated home value across a given region and housing type. It is a dollar-denominated alternative to repeat-sales indices. Zillow also publishes home value and other housing data for local markets, as well as a more detailed methodology and a comparison of ZHVI to the S&P CoreLogic Case-Shiller Home Price Indices.

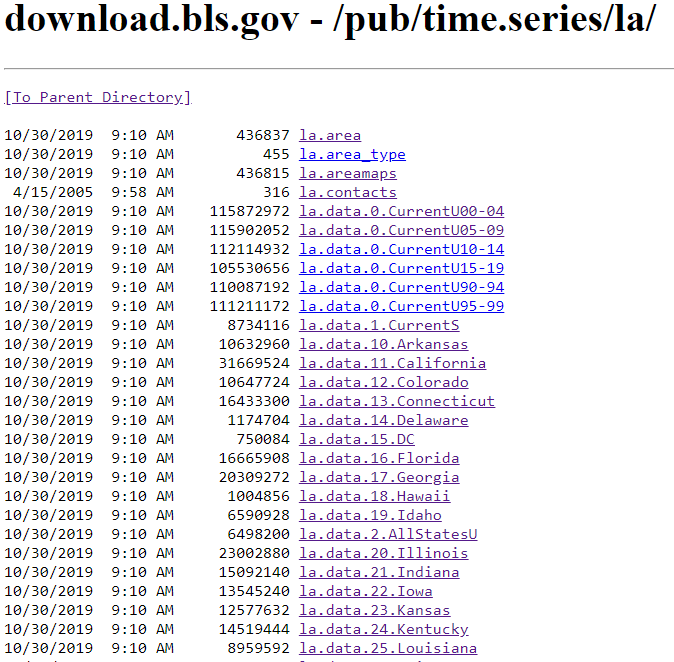
**Table 1.1** given below shows the sample dataset from Zip\_Zhvi\_SingleFamilyResidence.csv.



**Table 1.1 Sample Zip\_Zhvi\_SingleFamilyResidence dataset**

**Additional dataset**

Unemployment rate from Bureau of Labor Statistics and Census data



This dataset contains the following information

|  |  |
| --- | --- |
| **series\_id** | Unique identifier for a metric published by bureau LASST100000000000003 represents Unemployment Rate by state where the two digits (10) followed by "LASST" represents the identifier for state (10 🡪 Delaware) |
| **year** | year when the metrics value is recorded |
| **period** | identifier for month when the metric value is recorded |
| **value** | metric value where the metric in this case is unemployment rate |
| **footnote\_codes** | notes/comments if applicable |

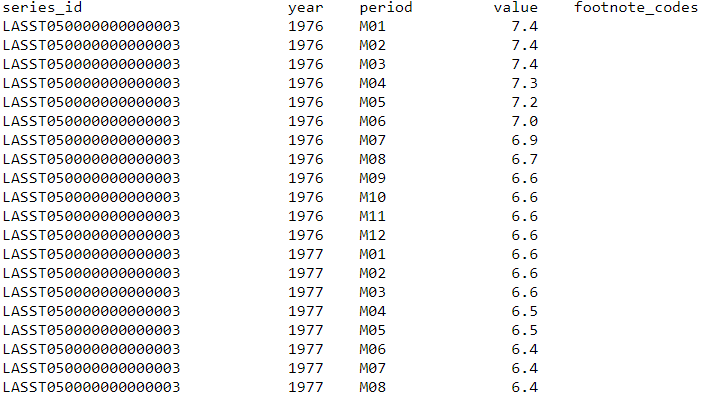
**Series data file description**



**State file layout**



**Table 1.2** given below shows the sample dataset from [la.data.10.Arkansas](https://download.bls.gov/pub/time.series/la/la.data.10.Arkansas).



**Table 1.2 Sample** [**la.data.10.Arkansas**](https://download.bls.gov/pub/time.series/la/la.data.10.Arkansas)**. dataset**

**Answer the following business question using data from Zillow:**

The research question is can we predict best investment zip codes?

• Using the base data available from Zillow

(files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_SingleFamilyResidence.csv)

o Review the data – clean as appropriate

▪ Develop time series plots for the following Arkansas metro areas:

• Hot Springs, Little Rock, Fayetteville, Searcy

• Present all values from 1997 to present

**• Average at the metro area level**

o Develop model(s) for forecasting average median housing value by zip code for 2018

o Use the historical data from 1997 through 2017 as training data

o Integrate data from other sources (Bureau of Labor Statistics and Census data) to improve upon base model(s)

After cleaning the dataset, exploratory data analysis is performed on top of this dataset to study each variable and its interaction with one another

**Figure 1.1**: This gives the distribution of median housing value for Hot Springs, Little Rock, Fayetteville, Searcy and its trend

**Figure 1.2**: Shows the boxplot and annual trend compared to each other to see the variations and any seasonal input for Fayetteville

**Figure 1.3**: Shows the boxplot and annual trend compared to each other to see the variations and any seasonal input for Hot Springs

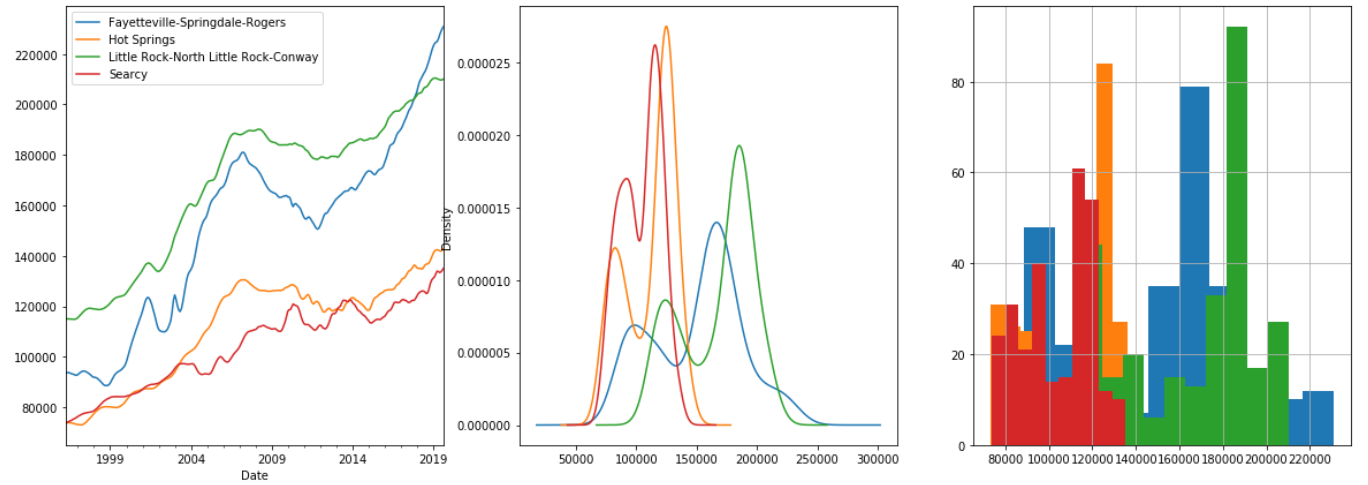
**Figure 1.4**: Shows the boxplot and annual trend compared to each other to see the variations and any seasonal input for Little Rock

**Figure 1.5**: Shows the boxplot and annual trend compared to each other to see the variations and any seasonal input for Searcy

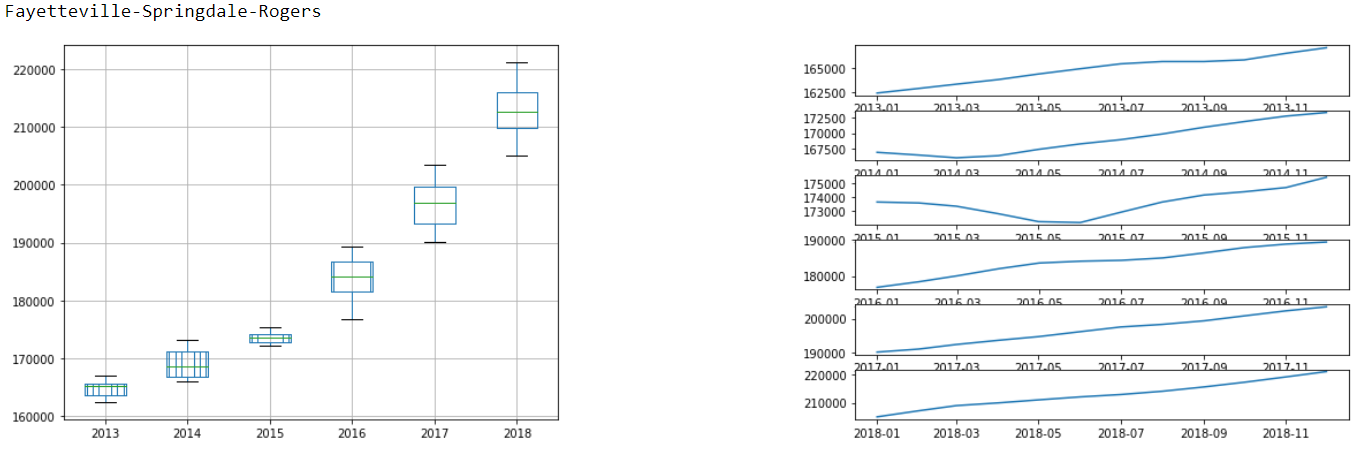
**Figure 1.6**: Shows the stationary observation of **Hot** Springs, Little Rock, Fayetteville, Searcy

**Figure 1.7**: Shows the autocorrelation and partial auto correlation details on **Hot** Springs, Little Rock, Fayetteville, Searcy

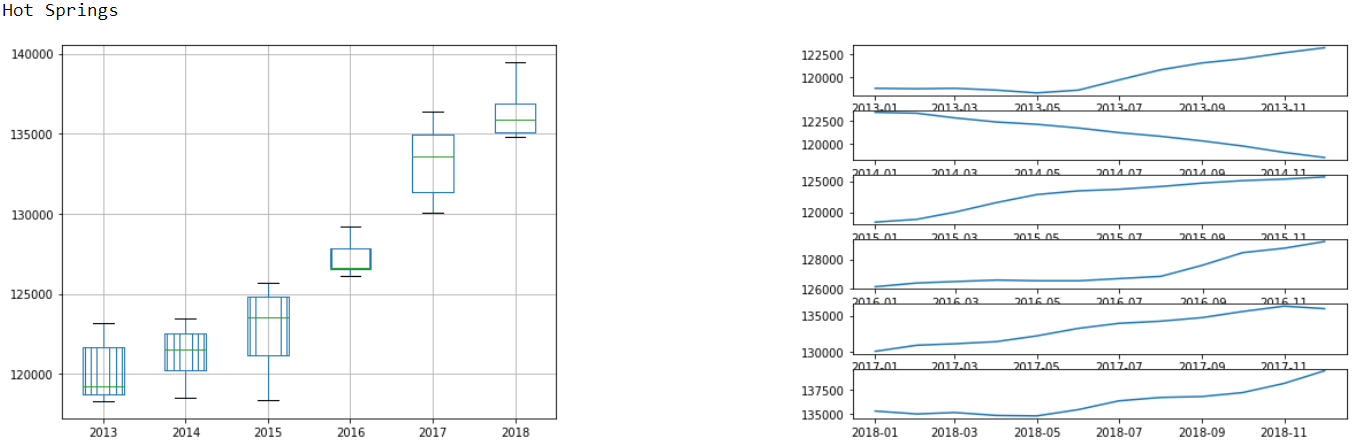
**Figure 1.8**: This gives the correlation matrix between all the dependent variable



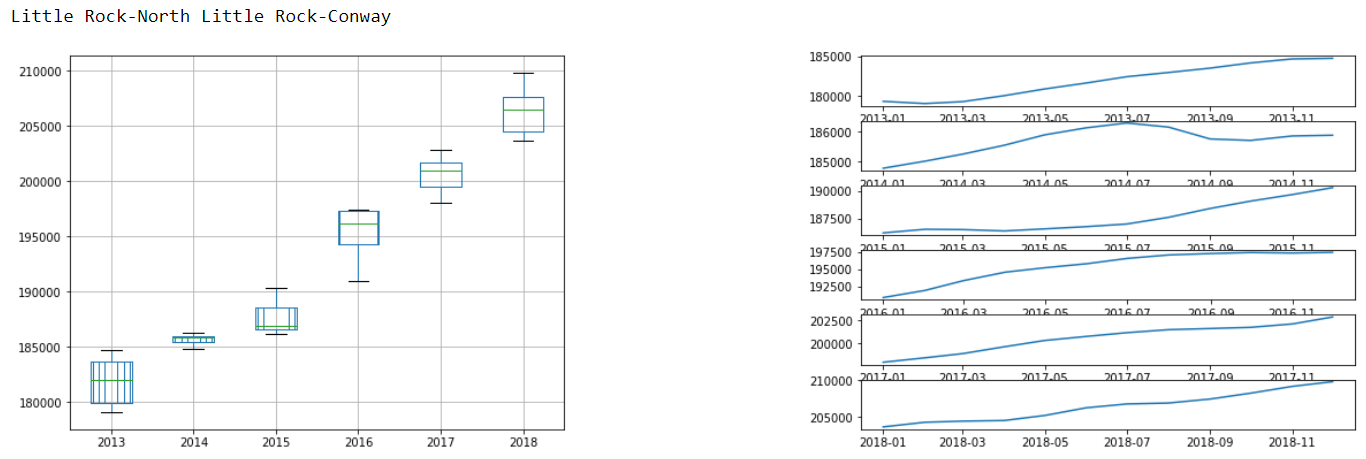
**Figure 1.1**



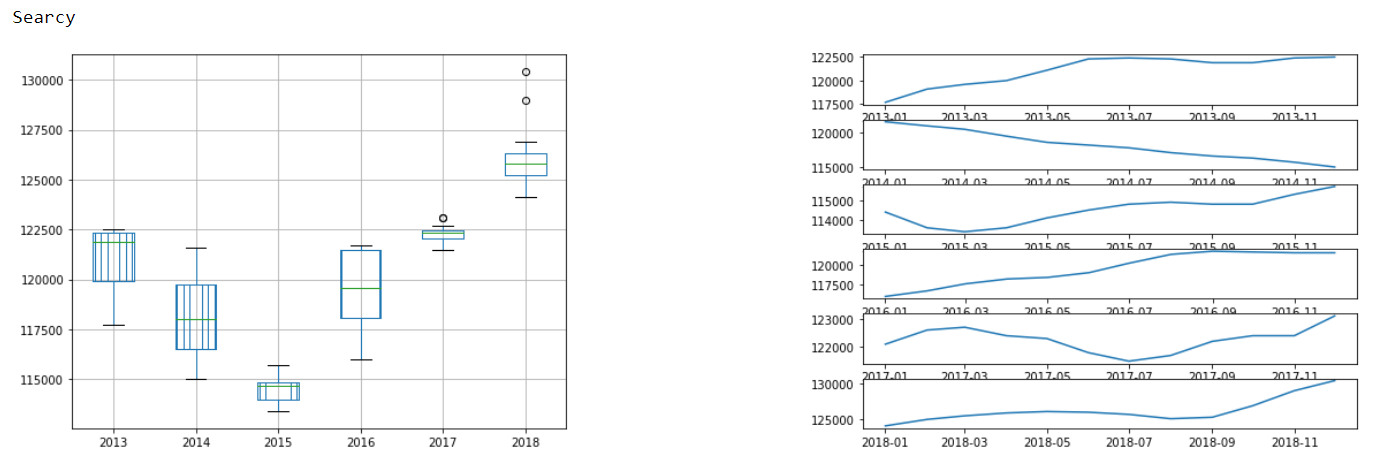
**Figure 1.2**



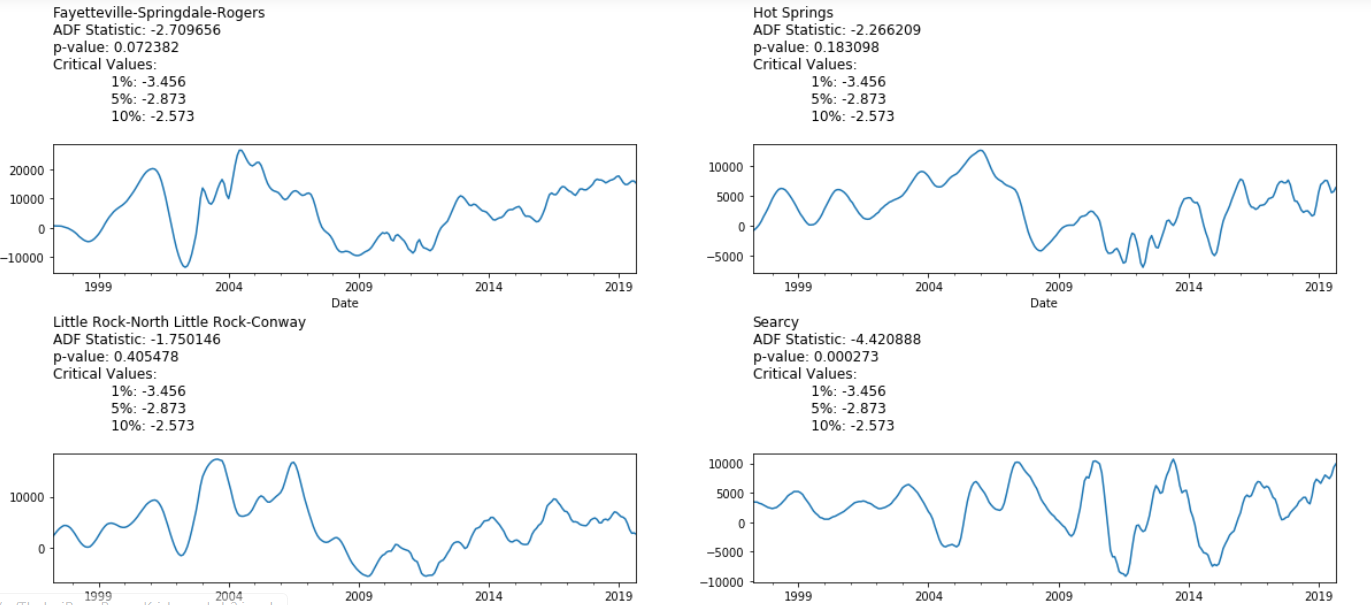
**Figure 1.3**



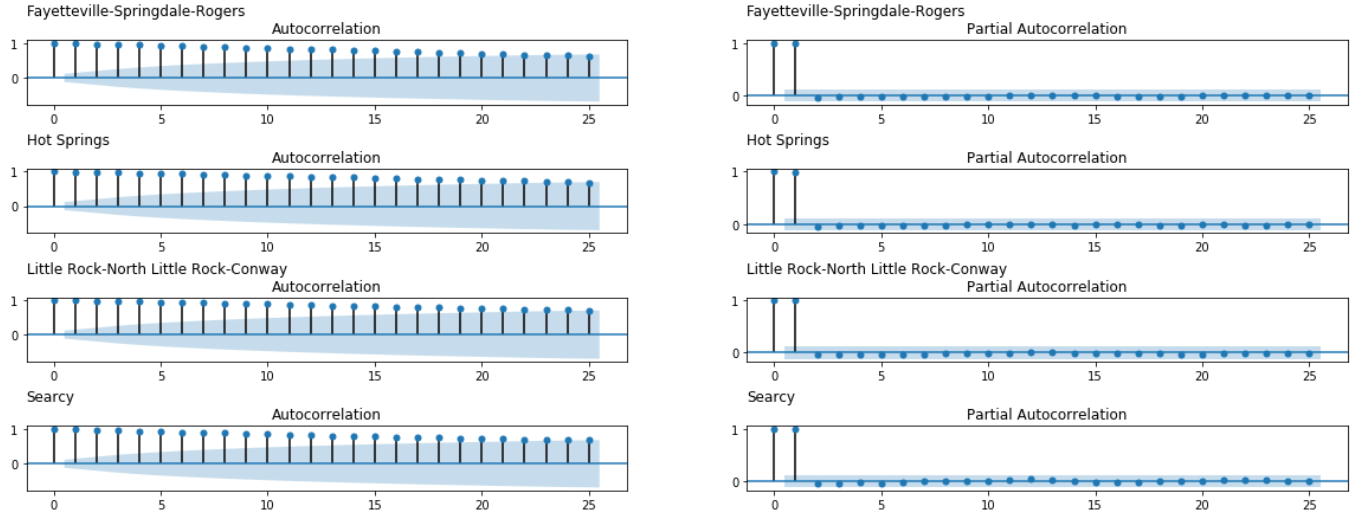
**Figure 1.4**



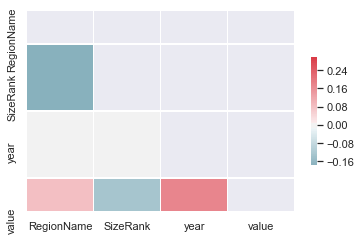
**Figure 1.5**



**Figure 1.6**

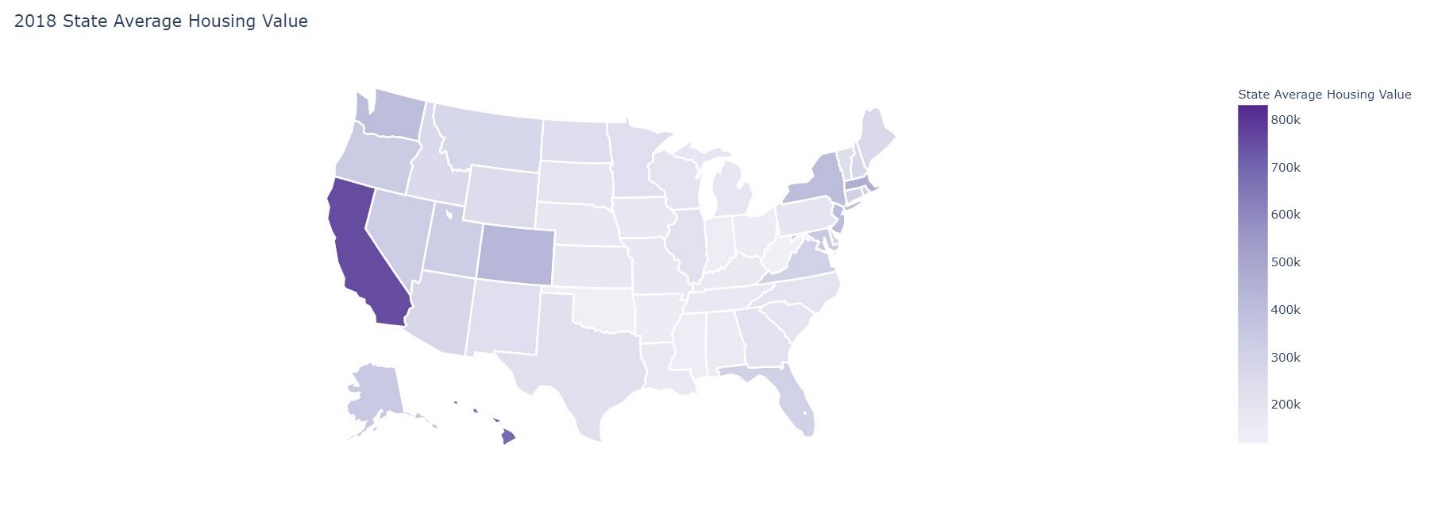


**Figure 1.7**

****

**Figure 1.8**

State Average Housing Value is plotted in the US MAP and the figure is shown below



### **Models**

In this exercise, models are developed using ARIMA and ordinary least squares.

**ARIMA** is an acronym that stands for AutoRegressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data.

**OLS** Ordinary least squares is a type of linear least squares method for estimating the unknown parameters in a linear **regression** model

#### **ARIMA**

An ARIMA model is a class of statistical models for analyzing and forecasting time series data.

It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.

This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.

I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.

MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model, also called the lag order.

d: The number of times that the raw observations are differenced, also called the degree of differencing.

q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model.

Adopting an ARIMA model for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious, but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.

#### **OLS Regression**

OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function.

Geometrically, this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression surface – the smaller the differences, the better the model fits the data. The resulting estimator can be expressed by a simple formula, especially in the case of a simple linear regression, in which there is a single regressor on the right side of the regression equation.

The OLS estimator is consistent when the regressors are exogenous, and optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated. Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances. Under the additional assumption that the errors are normally distributed, OLS is the maximum likelihood estimator.

OLS is used in fields as diverse as economics (econometrics), data science, political science, psychology and engineering (control theory and signal processing).

#### **Model 1: # ARIMA timeseries model for forecasting ZHVI**

ARIMA model is constructed for the few metro areas using p, d and q ranging from various input values as given below

p\_values = range(0, 7)

d\_values = range(0, 3)

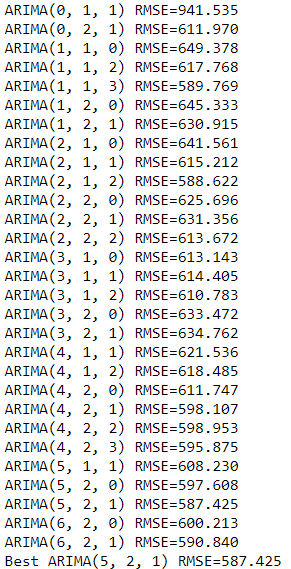
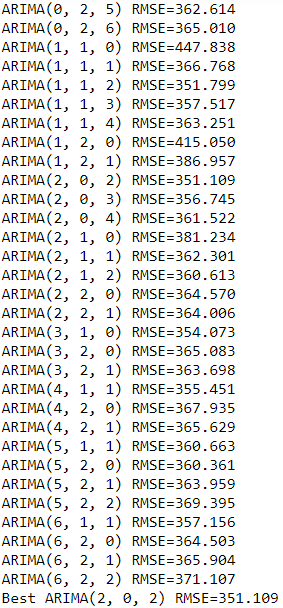
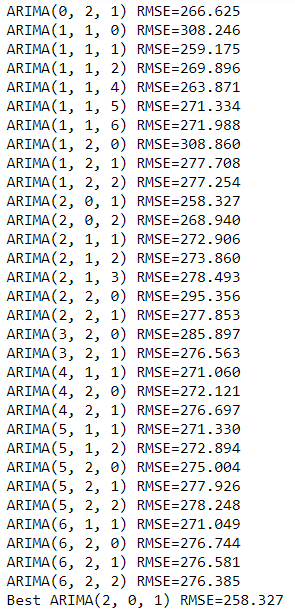
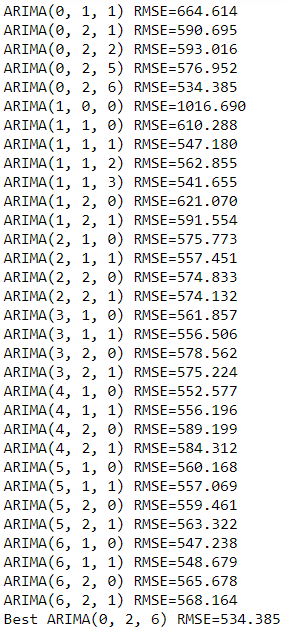
q\_values = range(0, 7)

**Metro areas for which the model is used are as follows**

* Fayetteville
* HotSprings
* LittleRock
* Searcy

**Model Outcome**

**Fayetteville HotSprings LittleRock Searcy**

#### **Model 2: # OLS Regression for predicting ZHVI using Zillow data**

OLS Regression Results

==============================================================================

Dep. Variable: value R-squared: 0.986

Model: OLS Adj. R-squared: 0.986

Method: Least Squares F-statistic: 2.449e+04

Date: Sun, 10 Nov 2019 Prob (F-statistic): 0.00

Time: 10:00:53 Log-Likelihood: -3.3450e+06

No. Observations: 286604 AIC: 6.692e+06

Df Residuals: 285766 BIC: 6.700e+06

Df Model: 837

Covariance Type: nonrobust

=======================================================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------------------------------

Intercept 9.335e+05 1.91e+04 48.917 0.000 8.96e+05 9.71e+05

State[T.AL] -1697.1159 3241.207 -0.524 0.601 -8049.791 4655.560

State[T.AR] 4.027e+04 3812.401 10.562 0.000 3.28e+04 4.77e+04

.

.

.

State[T.WA] 7.183e+04 6954.182 10.329 0.000 5.82e+04 8.55e+04

State[T.WI] 2.012e+04 1613.876 12.468 0.000 1.7e+04 2.33e+04

State[T.WV] -1.427e+04 4044.592 -3.527 0.000 -2.22e+04 -6338.180

State[T.WY] 3.48e+04 9614.084 3.620 0.000 1.6e+04 5.36e+04

Metro[T.Abilene] 5073.9595 3748.282 1.354 0.176 -2272.569 1.24e+04

Metro[T.Ada] 2671.0174 6141.204 0.435 0.664 -9365.572 1.47e+04

.

.

.

Metro[T.Yuma] 4806.1553 2699.941 1.780 0.075 -485.653 1.01e+04

Metro[T.Zanesville] 853.9452 3610.300 0.237 0.813 -6222.144 7930.034

RegionName -1.1717 0.143 -8.211 0.000 -1.451 -0.892

SizeRank -0.0563 0.014 -4.026 0.000 -0.084 -0.029

year -444.0871 8.957 -49.580 0.000 -461.643 -426.532

previous\_value 1.0400 0.000 3609.201 0.000 1.039 1.041

==============================================================================

Omnibus: 254070.345 Durbin-Watson: 0.879

Prob(Omnibus): 0.000 Jarque-Bera (JB): 3799396378.165

Skew: 2.565 Prob(JB): 0.00

Kurtosis: 567.032 Cond. No. 2.95e+20

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 3.16e-25. This might indicate that there are

strong multicollinearity problems or that the design matrix is singular.

#### **Model 3: OLS Regression for predicting ZHVI using dataset from Zillow and Centre for Bureau and Labor Statistics**

OLS Regression Results

==============================================================================

Dep. Variable: value R-squared: 0.987

Model: OLS Adj. R-squared: 0.987

Method: Least Squares F-statistic: 2.682e+04

Date: Sun, 10 Nov 2019 Prob (F-statistic): 0.00

Time: 22:27:06 Log-Likelihood: -3.3319e+06

No. Observations: 286604 AIC: 6.666e+06

Df Residuals: 285765 BIC: 6.674e+06

Df Model: 838

Covariance Type: nonrobust

=======================================================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------------------------------

Intercept -6.814e+04 1.92e+04 -3.545 0.000 -1.06e+05 -3.05e+04

State[T.AL] -1.925e+04 3098.951 -6.212 0.000 -2.53e+04 -1.32e+04

State[T.AR] 2.308e+04 3644.415 6.334 0.000 1.59e+04 3.02e+04

.

.

.

Metro[T.Yuba City] 463.9552 2089.655 0.222 0.824 -3631.710 4559.621

.

.

.

Metro[T.Yuma] 3173.4506 2579.937 1.230 0.219 -1883.155 8230.056

Metro[T.Zanesville] 1208.2214 3449.810 0.350 0.726 -5553.311 7969.753

RegionName -1.2081 0.136 -8.860 0.000 -1.475 -0.941

SizeRank -0.0588 0.013 -4.399 0.000 -0.085 -0.033

year 78.5126 9.126 8.603 0.000 60.625 96.400

u\_rate -5174.1171 31.368 -164.949 0.000 -5235.598 -5112.637

previous\_value 1.0389 0.000 3771.851 0.000 1.038 1.039

==============================================================================

Omnibus: 286911.663 Durbin-Watson: 0.758

Prob(Omnibus): 0.000 Jarque-Bera (JB): 5432354376.739

Skew: 3.291 Prob(JB): 0.00

Kurtosis: 677.432 Cond. No. 5.95e+20

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 7.76e-26. This might indicate that there are

strong multicollinearity problems or that the design matrix is singular.

## **Results**

Test data set results are captured for each model specification and are given below

#### **Model 1: # ARIMA timeseries model for forecasting ZHVI**

**Best ARIMA model for Fayetteville, HotSprings, LittleRock and Searcy are as follows**

Fayetteville: Best ARIMA(5, 2, 1) RMSE=587.425

HotSprings: Best ARIMA(2, 0, 2) RMSE=351.109

LittleRock: Best ARIMA(2, 0, 1) RMSE=258.327

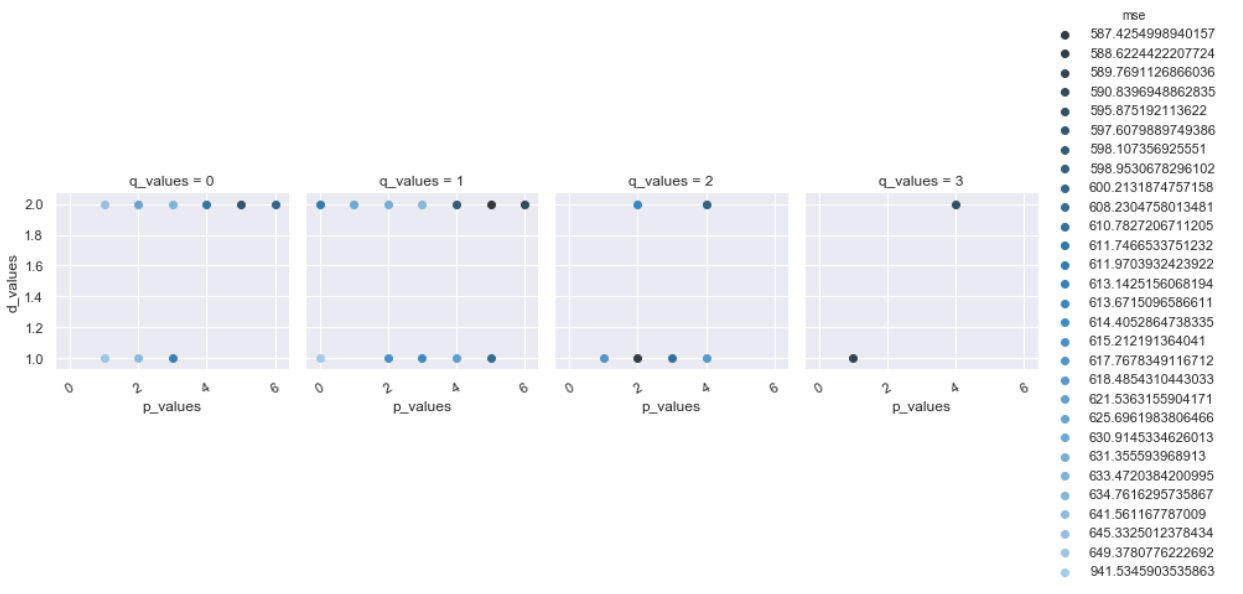
Searcy: Best ARIMA(0, 2, 6) RMSE=534.385

**Figure 2.1**: Shows the scatter plot on root mean squared value for Fayetteville for different p, d and q values

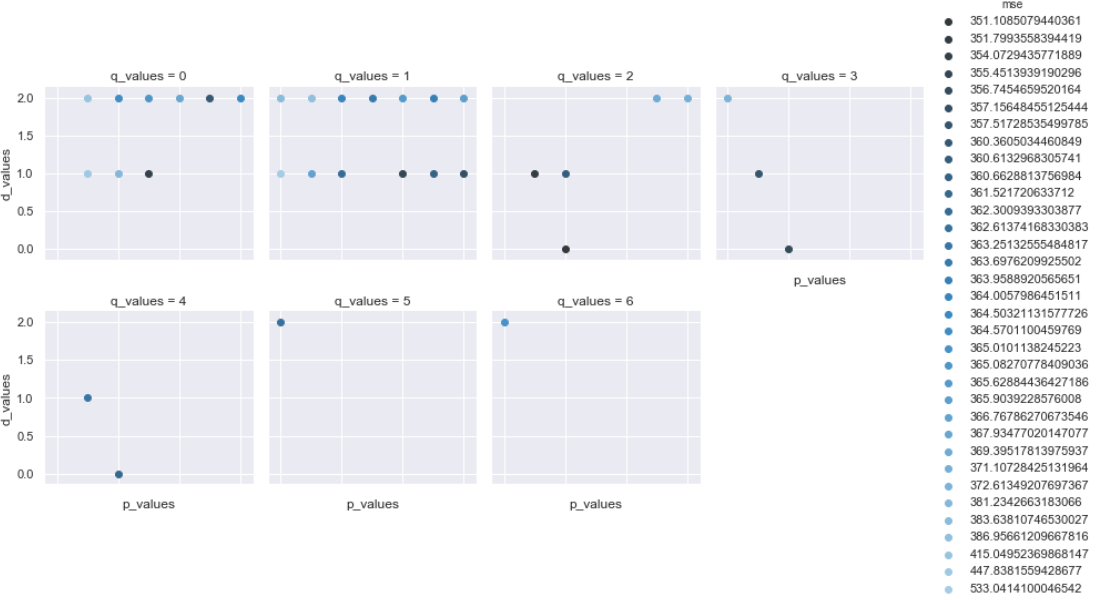
**Figure 2.2**: Shows the scatter plot on root mean squared value for HotSprings for different p, d and q values

**Figure 2.3**: Shows the scatter plot on root mean squared value for LittleRock for different p, d and q values

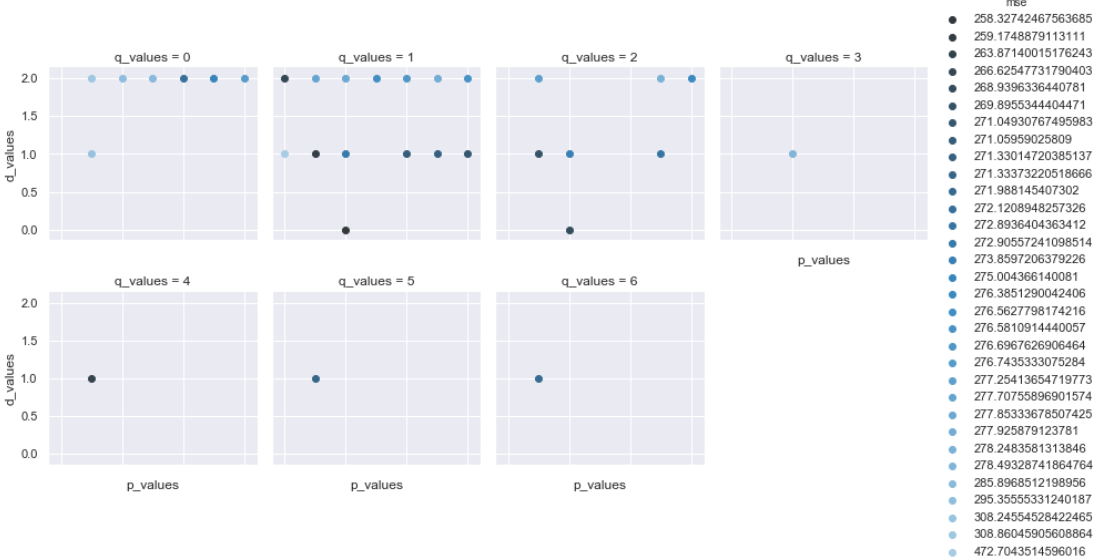
**Figure 2.4**: Shows the scatter plot on root mean squared value for Searcy for different p, d and q values



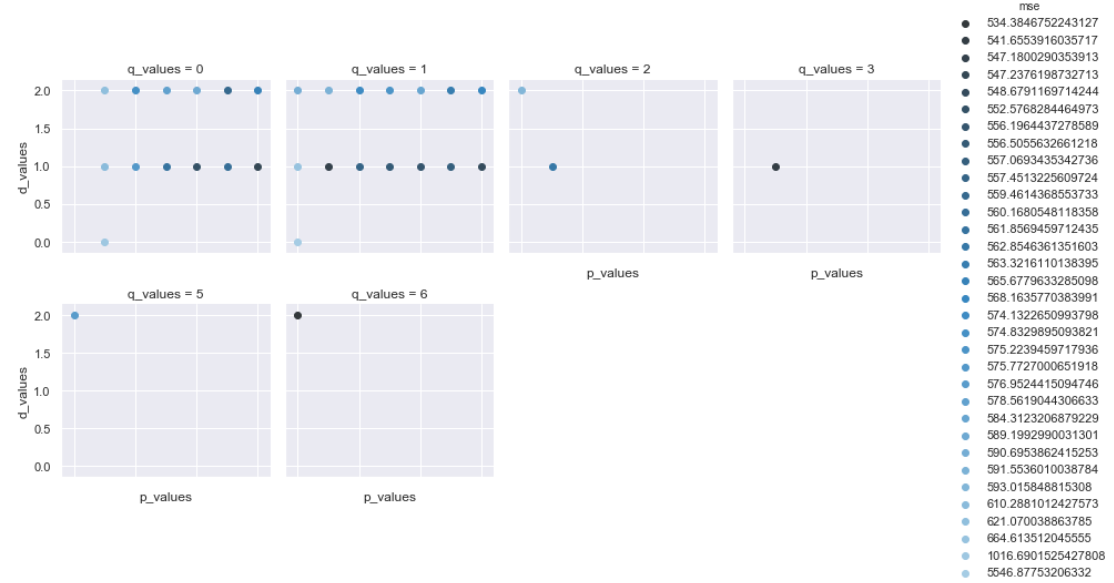
**Figure 2.1**



**Figure 2.2**



**Figure 2.3**



**Figure 2.4**

#### **Model 2: # OLS Regression for predicting ZHVI using Zillow data**

Proportion of Test Set Variance Accounted for: 0.995

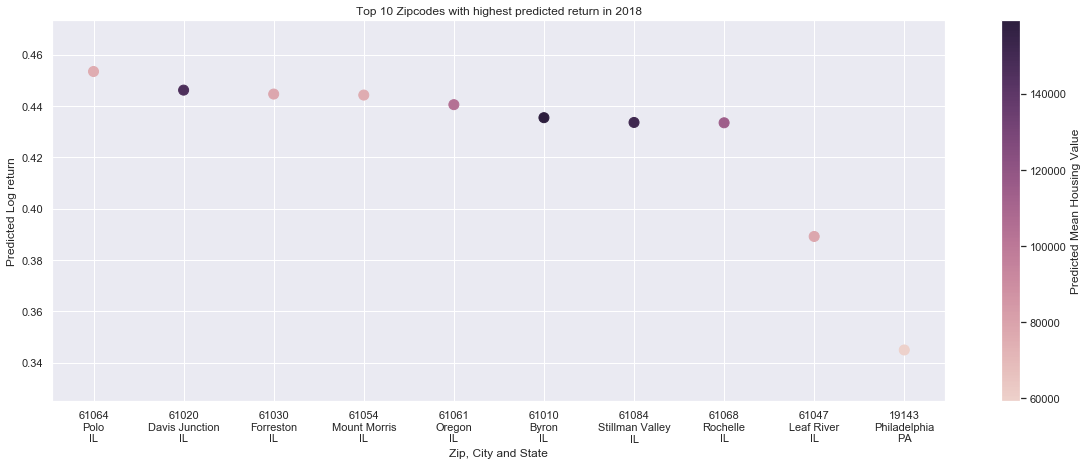
**Figure 2.5**: Shows the top 10 zip codes which yielded consistently higher returns in the past

**Figure 2.6**: Shows the top 10 zip codes with highest return in 2018

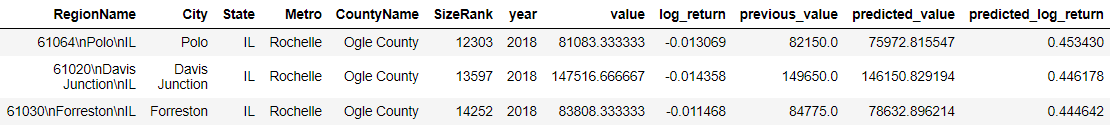
**Table 2.1**: It gives the top 3 zipcodes which are best for investment for the year 2018 based on the maximum log return

****

**Figure 2.5**

****

**Figure 2.6**



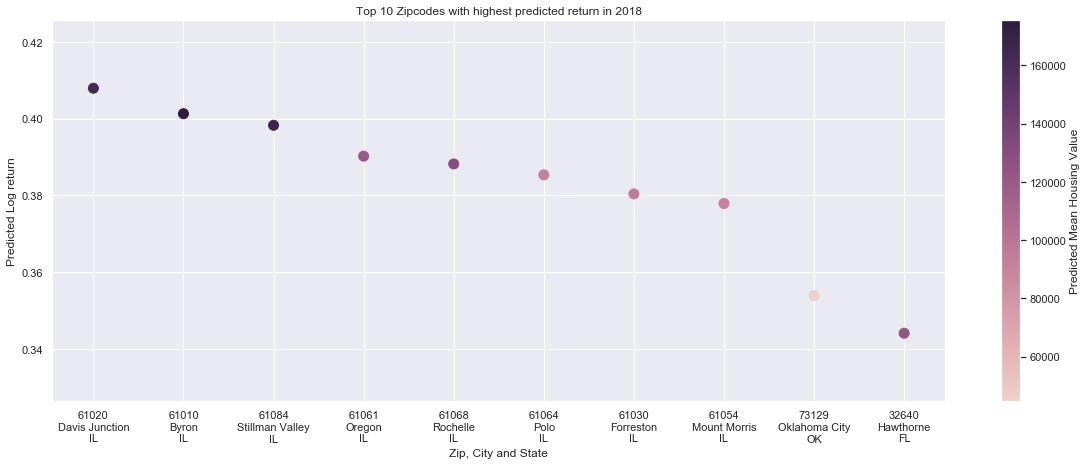
**Table 2.1**

#### **Model 3: OLS Regression for predicting ZHVI using dataset from Zillow and Centre for Bureau and Labor Statistics**

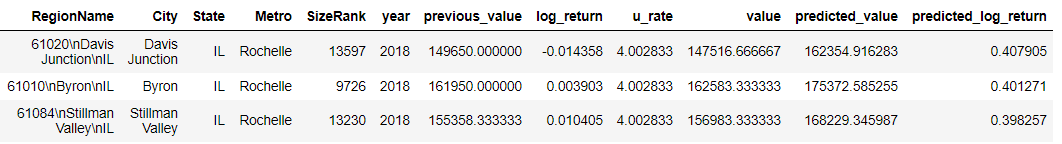
Proportion of Test Set Variance Accounted for: 0.995

**Figure 2.7**: Shows the top 10 zip codes with highest return in 2018

**Table 2.2**: It gives the top 3 zipcodes which are best for investment for the year 2018 based on the maximum log return

****

**Figure 2.7**



**Table 2.2**

## **Conclusion**

ARIMA time series plots are developed for the Arkansas metro areas Hot Springs, Little Rock, Fayetteville, Searcy using all the data from 1997 till present and the results are given as follows

**Best ARIMA model**

Fayetteville: Best ARIMA(5, 2, 1) RMSE=587.425

HotSprings: Best ARIMA(2, 0, 2) RMSE=351.109

LittleRock: Best ARIMA(2, 0, 1) RMSE=258.327

Searcy: Best ARIMA(0, 2, 6) RMSE=534.385

Model(s) for forecasting average median housing value by zip code for 2018 is constructed using OLS regression without any down sampling

The threes zip codes which are best for the investment opportunity is as follows and they are chosen because of the high predicted log return value.

