Diego Valdes

IST 718

Lab 2

Aug 9, 2019

**Data Exploration**

The analysis presented in this paper used data provided by the course work from Zillow of real estate prices by zip code for the fifty states of the United States and Washington D.C.. Each row represents one zip code and provides the state, city, metro area and real estate prices from April 2019 to June 2019, but the starting point for all analysis was January 1997.

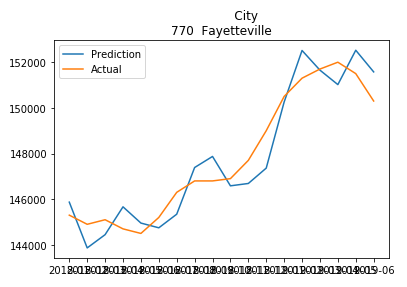
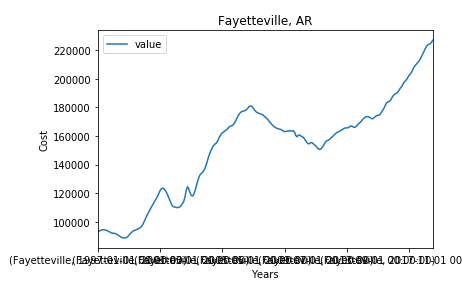
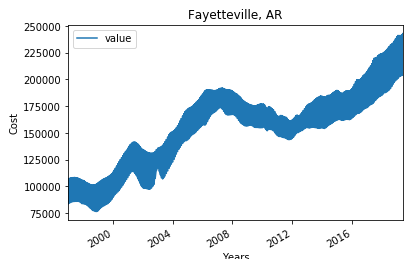
Cleaning the data comprised determining the number/circumstances of NaN values. It was discovered that many zip codes had years of pricing data. Options to replace the data were evaluated (taking the mean of the entire time span), but this method was not used because of the many fluctuations that can occur in the real estate market. These rows were removed. Zip codes with less than 4 characters or higher than 5 characters were also removed. The only consideration given to preserve some of this data was given to zip codes with 4 characters, as there is the possibility that these examples are missing the first zero, ie: 1758 -> 01758. This assumption, while may be valid, once an attempt was made to apply the fix, it was quickly discovered that it was possible to have multiple matches for zip codes with 4 characters. The effort to pick the correct zip code, while possible, required a level of effort that was deemed to severe for the intended analysis, mainly due to dropping all of these values still resulted in 84% of the original data remaining to be analyzed.

**Analysis and Results**

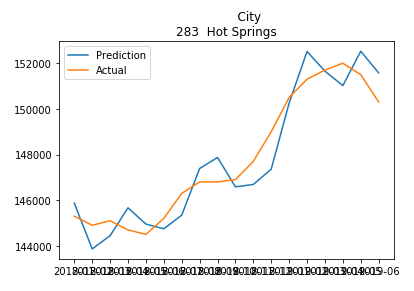
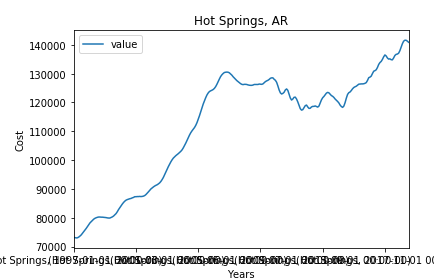
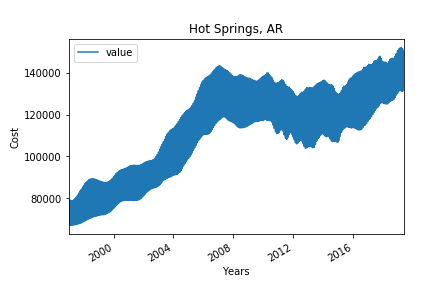
**Arkansas**

Initial analysis began with the great state of Arkansas, with the cities: Hot Springs, Fayetteville, Searcy, and Little Rock. For each area, plots were generated that examined all of the real estate values, the mean real estate value, and a model was built to predict the price of the real estate. The analysis was completed as the first step to refine the process of creating a model by zip code so it could then be applied to the entire nation, and to examine the results on a smaller scale before scalled to the entire dataset. ARIMA was used to generate the prediction.

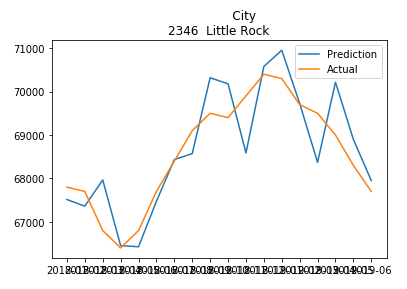
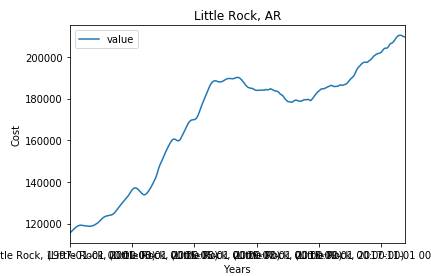
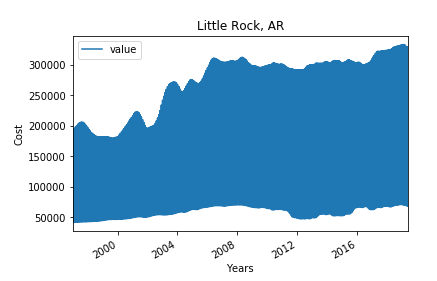
**Fayetville, AR**



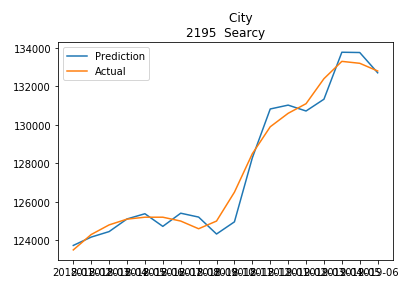
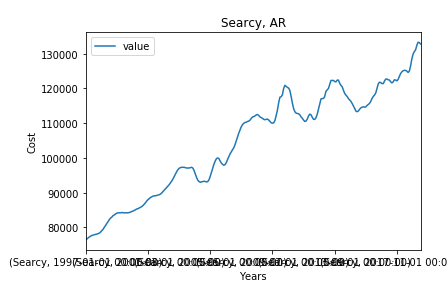
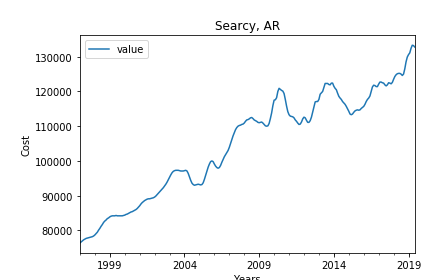
**Hot Springs, AR**



**Little Rock, AR**



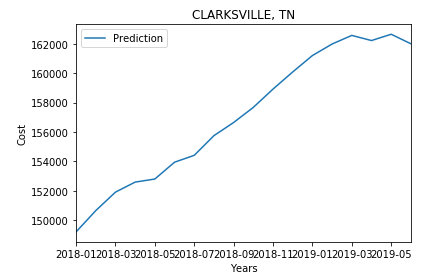
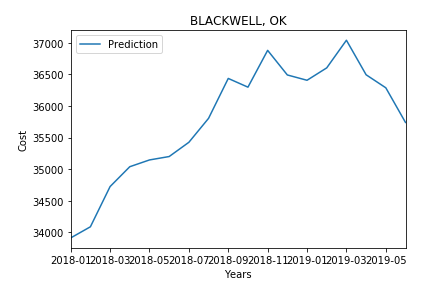
**Searcy, AR**



**United States**

Models were developed using ARIMA. The dataset was divided into training (1997 – 2017) and test (2018 – 2019), and pivoted by region name for processing through the model. Each zipcode was then processed and the results recorded. Choices for the SREIT were based on the strength of the model (low RMSE), and Return on Investment (ROI). The model produced ROI values of greater than 40%, but these models had many of the worse RMSE values.

Recommendations:

* Clarksville, TN
  + RMSE: 236.7
  + Expected ROI: 9%
  + 
* BlackWell, OK
  + RMSE: 244.2
  + Expected ROI: 9.2%
  + 
* Toccoa, GA
  + RMSE: 245.4
  + Expected ROI: 11.2%
  + 