Kyle Welch

IST 718

Lab 2

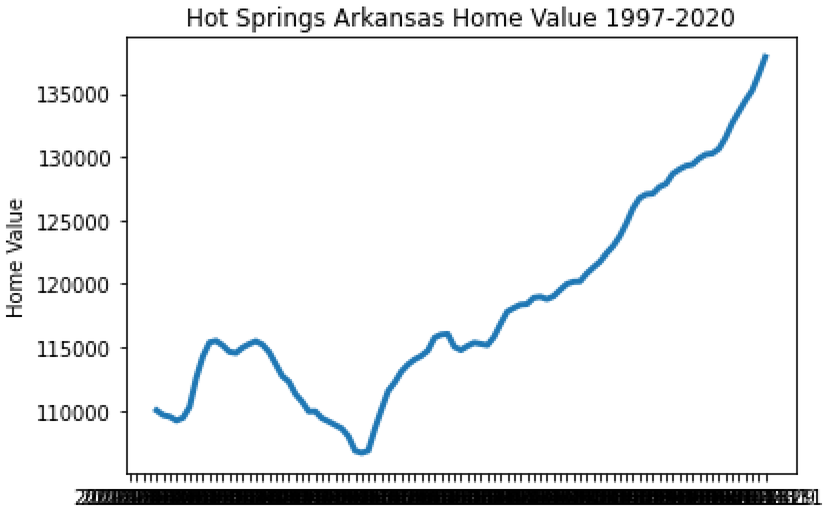
## Cleaning

The data was first read into a pandas data frame from the csv hosted on Zillow's website. The first step taken was to create a new data frame for just the Arkansas data so that it would not be lost when a majority of the rows were dropped. Initially, the entire dataset was intended to be used; however, due to technical difficulties the data set was reduced to 1100 zip codes instead of 30000+. Next, a new data frame was created with only the zip codes and home values from 1997 onwards. Using the melt function in pandas the data frame was transformed from wide to long, maintaining the zip code column, but introducing a date and value column that are titled 'ds' and 'y' for Prophet. The dates in the new ds column were originally column headers, so a conversion to datetime was necessary. Finally, the zip code column was renamed from 'RegionName' to 'Zip' for ease of use.

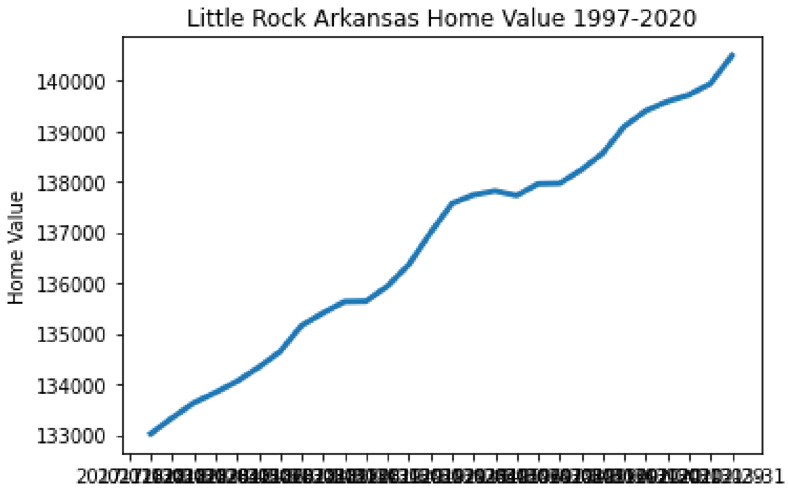
## Arkansas

The Arkansas data went through almost the exact same cleaning process with the only major difference being that the 'Metro' column was also kept and used when the data frame was melted from wide to long. From there a new data frames were created for each metro area, Hot Springs, Little Rock, Fayetteville, and Searcy. Lastly, each new metro area data frame grouped the housing values by dates and then calculated the mean housing value per date to make the time series plots for each metro area.

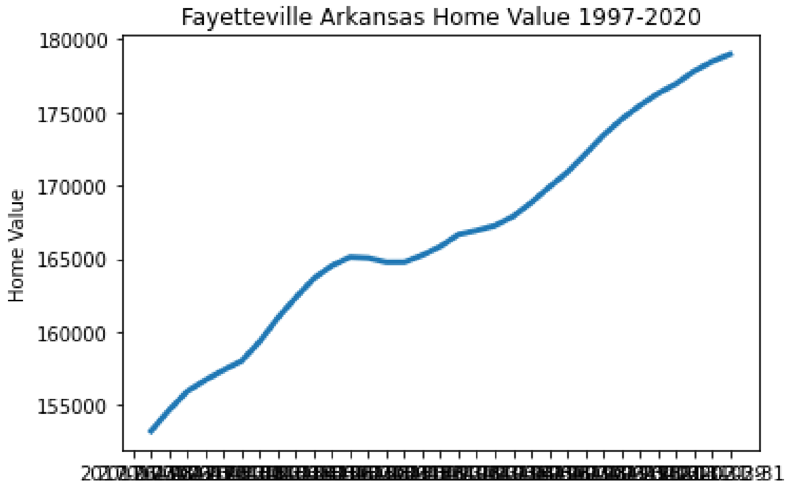
Hot Springs:



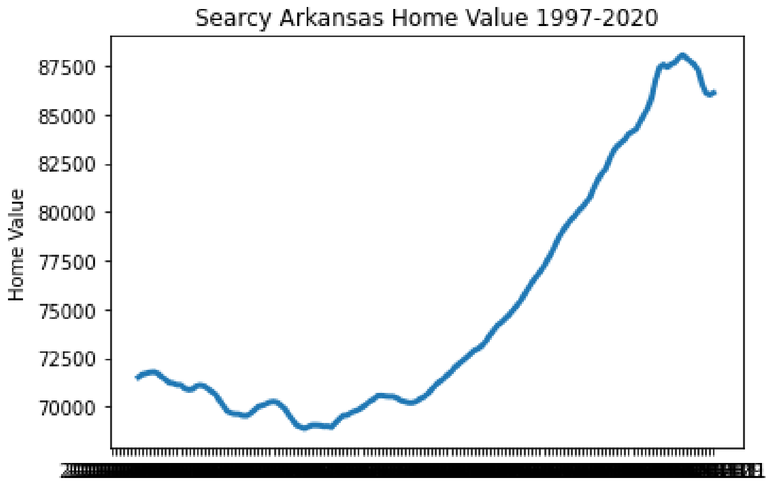
Little Rock:



Fayeteville:



Searcy:



Hot Springs and Fayetteville both have the largest gains over the timespan at roughly 25000 dollars; however, Hot Springs had a much lower starting average cost so their percentage of growth is greater. Searcy had the lowest starting value and increased roughly 12500 dollars. Little Rock had the least amount of growth in the time series of 7500 dollars.

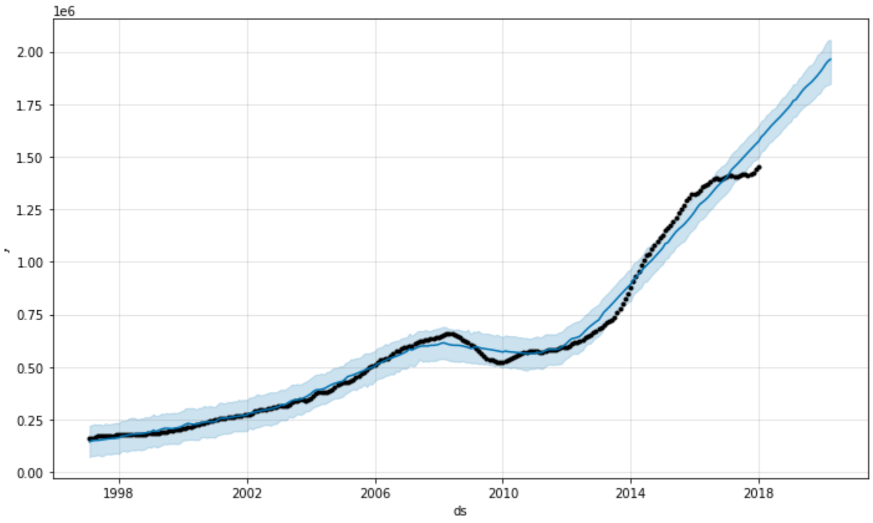
## Prophet

A ’train’ data frame was created that only included the house values by zip code to the end of 2017, then the Prophet model was created and tasked with forecasting the home values for 27 months, (the end of March 2020, this is where the Zillow data ends). The Prophet model was ran for each unique zip code to create 1100 forecasts. The forecasts were stored in a dictionary which was then converted to a data frame. From here some cleaning of the data was necessary, a new data frame was created to include, the zip code, the predicted March 2020 housing price, the current end of 2017 housing price, the projected growth (predicted value less the current value), and the projected percent (predicted percent growth).

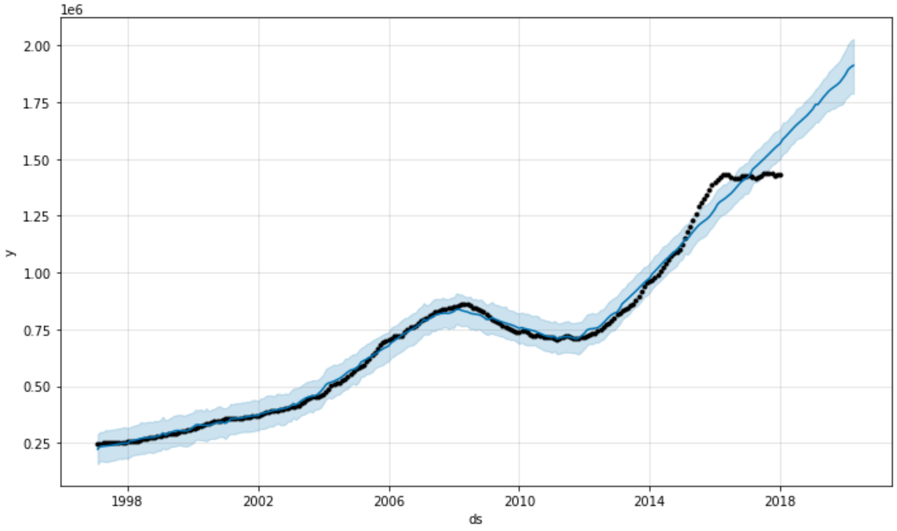
## Questions

- **What technique/algorithm/decision process did you use to down sample? (BONUS FOR NOT DOWN SAMPLING)**
  My original intent was to not down sample; however, due to technical difficulties I had to drop the majority of the rows using the following function, df = df[:-29364], leaving me with 1100 zip codes.

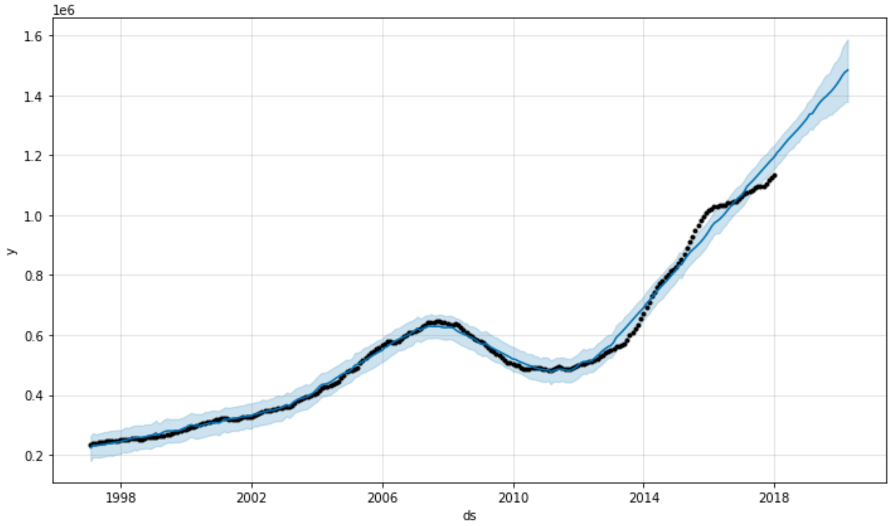
- **What three zip codes provide the best investment opportunity for the SREIT?**
  The three zip codes that provide the best investment opportunity for the SREIT at the end of 2017 according to the model are:
  **11216 (Brooklyn, NY) – 35.27% Projected Growth**



### 11225 (Brooklyn, NY) – 33.53% Projected Growth



### 11221 (Brooklyn, NY) – 30.98% Projected Growth



- **Why?**
  The best investment opportunity was determined by the zip codes that had the largest projected growth in terms of percentage return on investment. Another way that the investment opportunity could have been determined would have been to use return on investment in terms of raw dollars, which then the best zip codes would be 10021, 11201, 10014 (all of which are still New York City zip codes). However, these zip codes have considerably less growth in terms of percent increase. Additionally, they are all significantly more expensive and would tie up a lot of capital that may be better invested elsewhere.