**Ryan Timbrook**

NetID: RTIMBROO

Course: IST 718 Big Data Analytics

Term: Summer, 2019

Assignment: Investment Opportunity Zip Codes Prediction

Table of Contents

[1 Introduction 3](#_Toc16377283)

[1.1 Research question: 3](#_Toc16377284)

[2 Analysis and Models 4](#_Toc16377285)

[2.1 About the Data 4](#_Toc16377286)

[2.1.1 OBTAIN the Zillow Residential data 4](#_Toc16377287)

[2.1.2 Data Exploration & Cleaning 4](#_Toc16377288)

[2.1.3 AR Metro Region Data Transformations 4](#_Toc16377289)

[2.1.4 Arkansas Metro Region Exploratory Data & Visualizations 6](#_Toc16377290)

[2.1.5 Arkansas Metro Region Timeseries Plots 7](#_Toc16377291)

[2.2 Models 10](#_Toc16377292)

[2.2.1 Model Timeseries Forecast, ZipCodes Details 11](#_Toc16377294)

[2.2.2 Top ZipCode Choose Matrix 11](#_Toc16377295)

[2.2.3 Model Visualizations 11](#_Toc16377297)

# Introduction

## 

## Research question:

Can we predict which three zip codes provide the best investment opportunity for the Syracuse Real Estate Investment Trust (RSEIT)?

* Using the base data available from [Zillow](http://localhost:8888/notebooks/syr_ds/ist718/block_2/lab2/files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv)
  + Review the data - clean as appropriate
  + Provide an initial data analysis to include:
    - 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**
  + Using data from Zillow:
    - Develop model(s) for forecasting **average median housing value by zip code** for **2018**
    - Use the historical data from **1997 to 2017 as the training data**
    - Integrate data from other sources (like; Bureau of Labor Statistics and Census data) to improve upon the base model(s)
  + Answer the following questions:
    - What technique/algorithm/decisions process did you use to down sample?
    - What three zip codes provide the best investment opportunity for the SREIT?
    - Why?

# Analysis and Models

## About the Data

### OBTAIN the Zillow Residential data

Using the base data available from [Zillow](http://localhost:8888/notebooks/syr_ds/ist718/block_2/lab2/files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv)

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](https://wp.zillowstatic.com/3/ZHVI-InfoSheet-04ed2b.pdf).

### Data Exploration & Cleaning

#### SCRUB / CLEAN

Clean and perform initial transformations steps of the data

Subset Zillow on the AR 'Metro' regions for initial time series investigation and plotting.  
Zillow Dataset on these primary metro areas, 1997 to present:  
Note: 61 observations, some observations don't have complete date records.

* Metro Names:
  + 'Fayetteville-Springdale-Rogers'
  + RegionID:89749 | Records start on: 2003-07
  + RegionID:89717 | Records start on: 2014-07
* 'Hot Springs'
* 'Little Rock-North Little Rock-Conway'
* 'Searcy'
  + RegionID:89370 | Records start on: 2012-01

**Zipcode = RegionName**

Steps taken to clean and scrub the data:

* Remove columns dates prior to 1997
  + ['1996-04','1996-05','1996-06','1996-07','1996-08','1996-09','1996-10','1996-11','1996-12']
* Remove any possible white space in column values
  + ['City','State','Metro','CountyName']
* Rename RegionName to ZipCode for clarity
* Create integer id mapping to zipcode for future lookups
  + id\_to\_zipcode = {i:z for i,z in enumerate(zillow.ZipCode)}
  + zipcode\_to\_id = {z:i for i,z in enumerate(zillow.ZipCode)}

### AR Metro Region Data Transformations

The following filtering techniques were taken for the AR Metro region analysis:

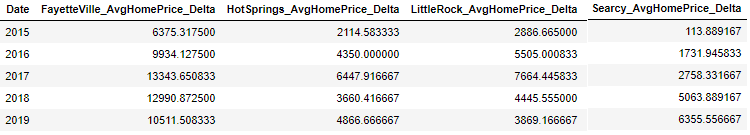
* Subset the zillow dataset to the AR metro areas
  + fayetteVille = 'Fayetteville-Springdale-Rogers'
  + hotSprings = 'Hot Springs'
  + lrNorth\_lrConway = 'Little Rock-North Little Rock-Conway'
  + searcy = 'Searcy'

Subset dataset shapes:

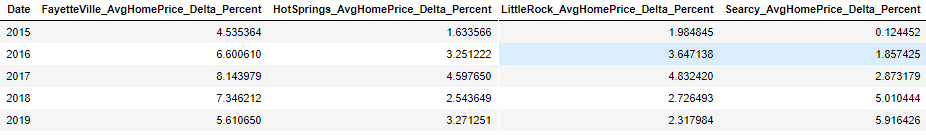
* Fayetteville-Springdale-Rogers: shape (21, 277)
* Hot Springs: shape (4, 277)
* Little Rock-North Little Rock-Conway: shape (30, 277)
* Searcy: shape (6, 277)

Data transformation to support Facebook Prophet Timeseries modeling package were made as followes:

* The input to Prophet is always a dataframe with two columns: ds and y. The ds (datestamp) column should be of a format expected by Pandas, ideally YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric, and represents the measurement we wish to forecast.
  + Average Median Home Prive Value by Metro Region and Year were generated to create a net value annual change region matrix:
    - Average Median Home Price
    - Annual Median Home Price Net Change Year over Year



* + - Annual Median Home Price as a percent Net Change Year over Year



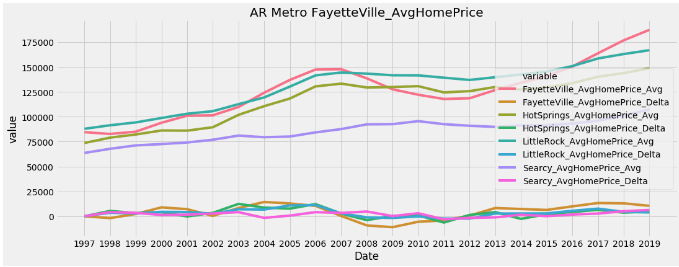
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

### Arkansas Metro Region Exploratory Data & Visualizations

#### Annual Net Price Changes by Metro Region

|  |
| --- |
| Figure: Annual Net Median Home Price Change |
| Figure: Annual Percent Net Median Home Price Change |

Figure: Annual Net Price Change, Percent and Value

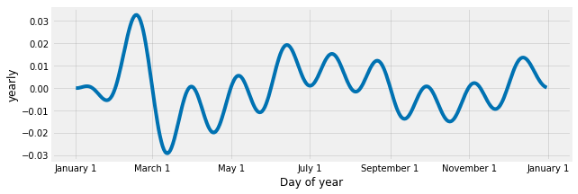
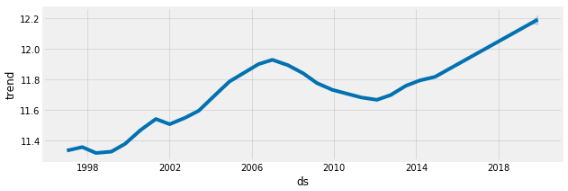
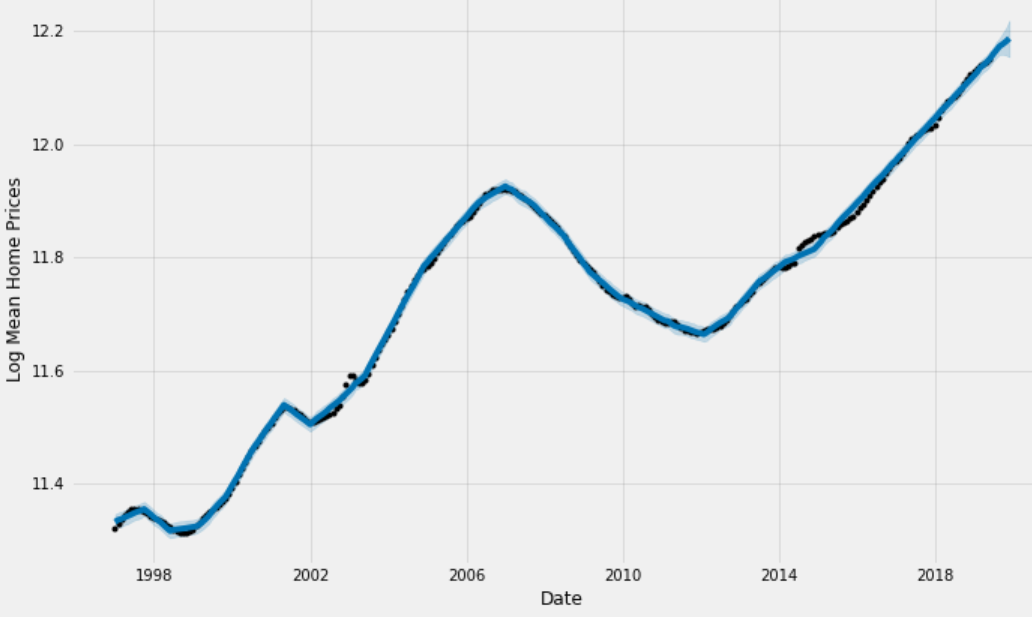


### Arkansas Metro Region Timeseries Plots

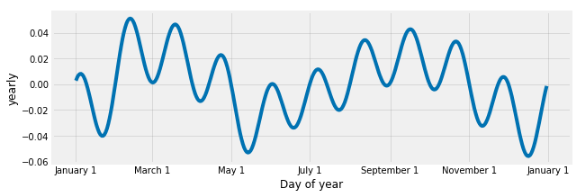
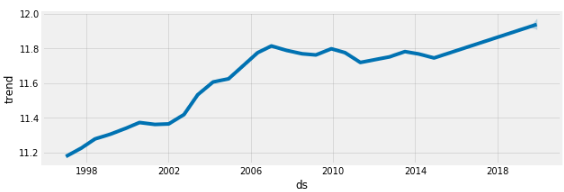
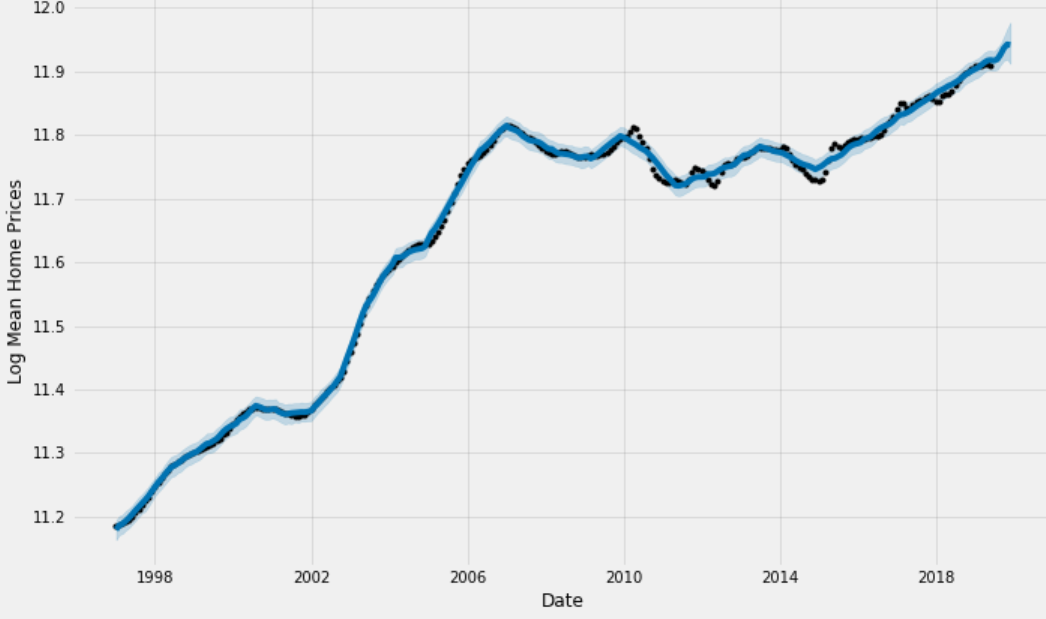
Provide an intial data analysis to include:

* 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

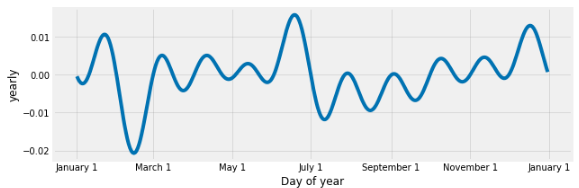
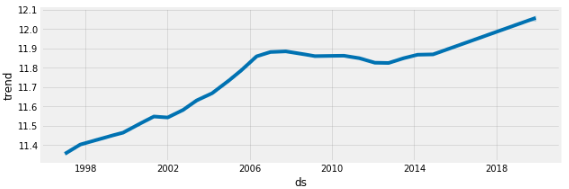
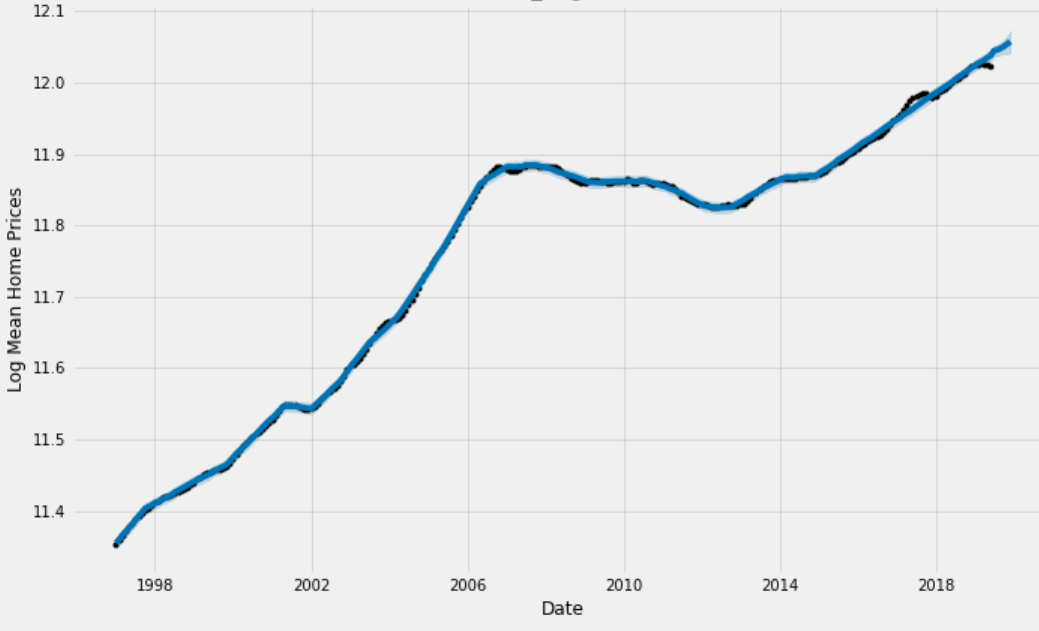
#### Figure: FayetteVille, Log Median Home Timeseries Forecast



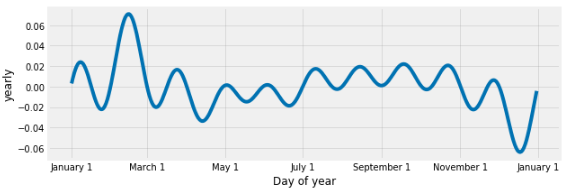
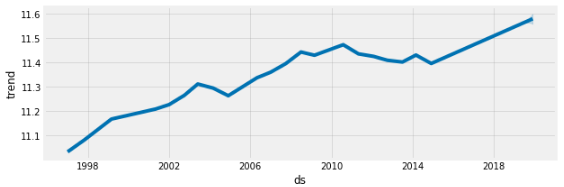
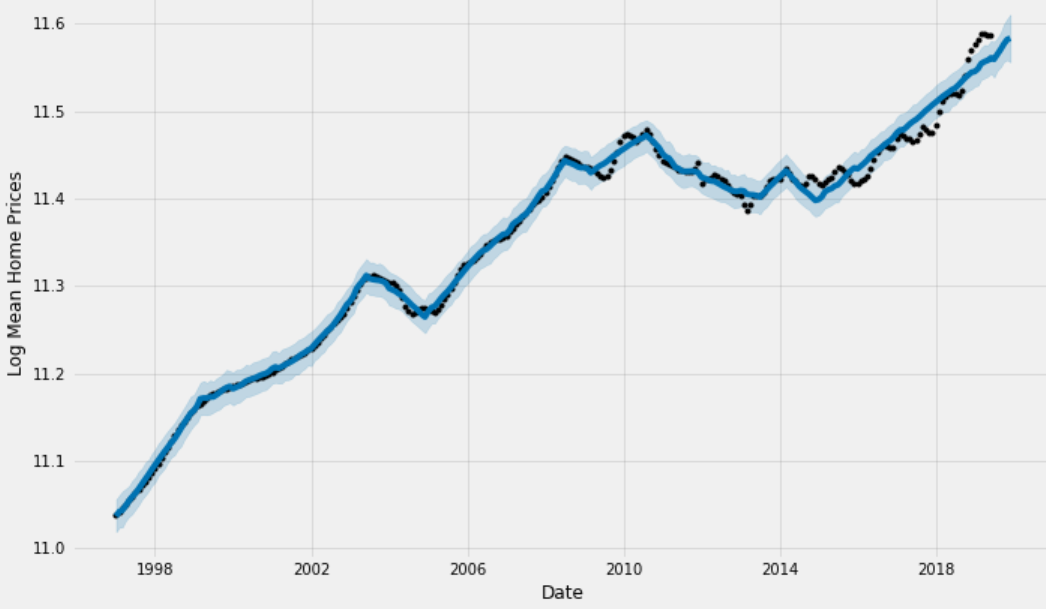
#### Figure: HotSprings, Log Median Home Timeseries Forecast



#### Figure: Little Rock, Log Median Home Timeseries Forecast



#### Figure: Searcy, Log Median Home Timeseries Forecast



## Models

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

* Use the historical data from 1997 through 2017 as the training data
* Integrate data from other sources, like the Bureau of Labor Statistics and Census data to improve upon the base model(s)
  + Capital markets and economics: **using time series analysis**
    - Seasonal unemployment, Price/return series, Risk analysis

**The model should answer these questions:**

* What technique/algorithm/decision process was used to down sample?
  + Methodology taken was to select the top performing zipcodes based on thier prior five year average net percent home median value change year-over-year. (PCHV)
* What three zip codes provide the best investment opportunity for the SREIT? And Why?

**Top three best performing zipcodes over the prior 5 years:**

* 30315: PCHV 22.15%
* 30032: PCHV 19.73%
* 29405: PCHV 18.55%

**Modeling Algorithms and Python packages used:**

* Algorithm:
  + Time series: prophet
  + Picking:
    - selecting top performing zip codes:
      * Methodology taken was to select the top performing zipcodes based on thier prior five year average net percent home median value change year-over-year. (PCHV)
      * Future enhancements would be to provide performing picking based on best NPV scores.

**Percent change of housing value --- PCHV**

### Model Timeseries Forecast, ZipCodes Details

### Top ZipCode Choose Matrix

### 

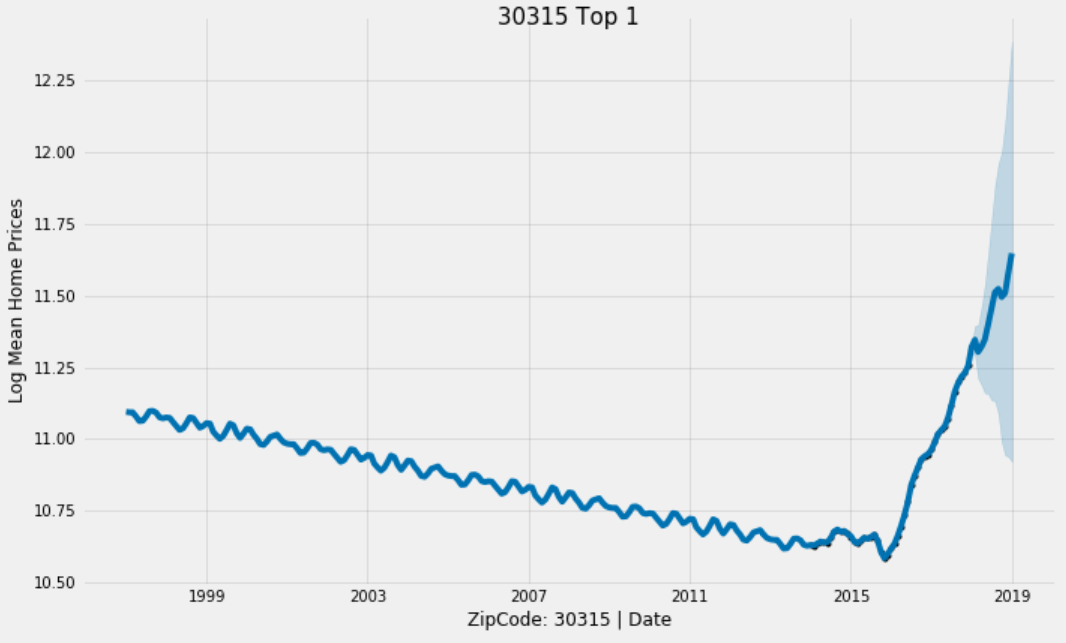
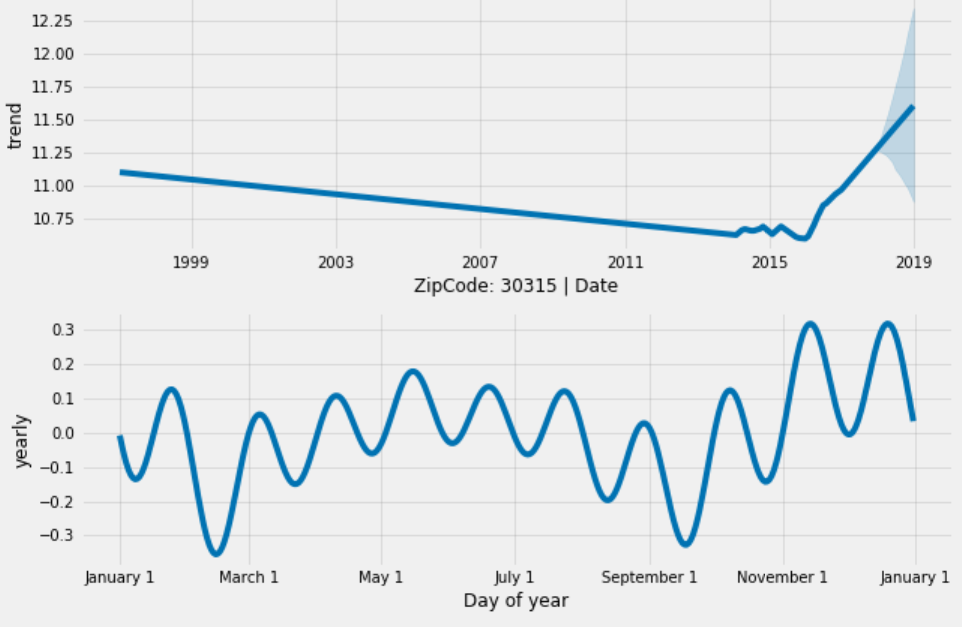


**Top three best performing zipcodes over the prior 5 years:**

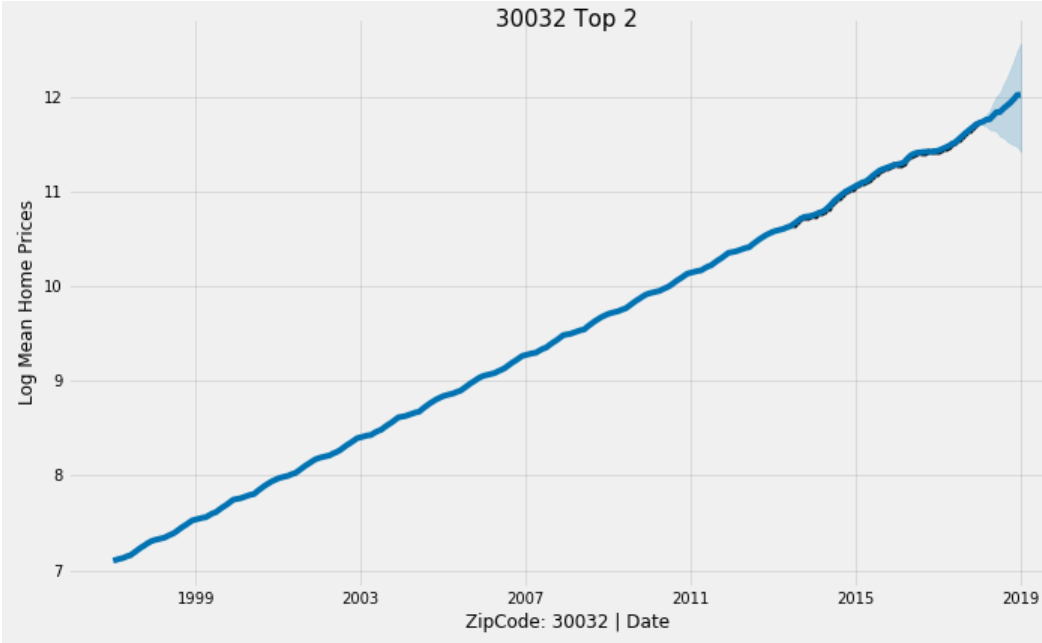
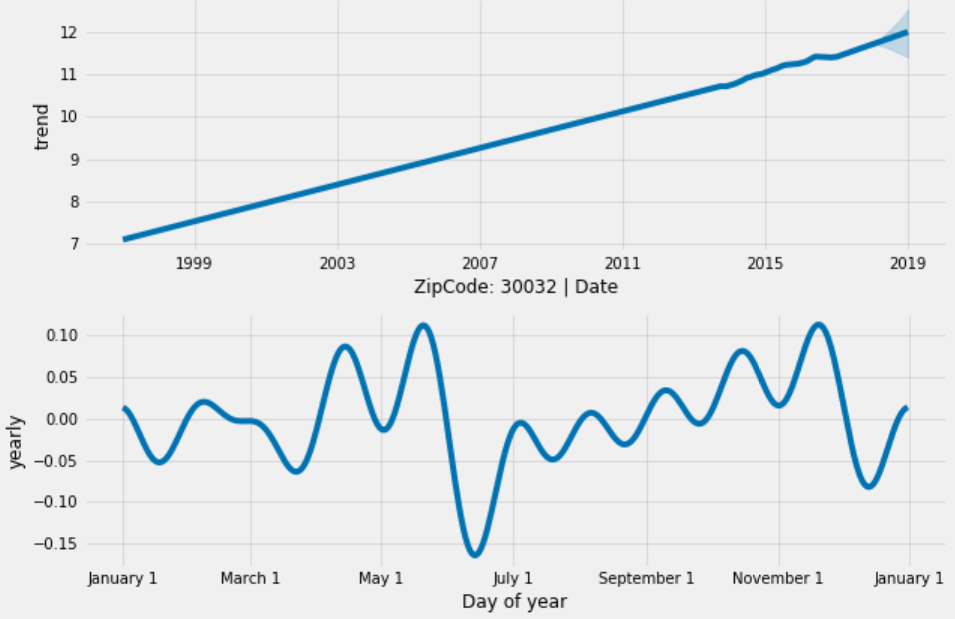
* 30315: PCHV 22.15%
* 30032: PCHV 19.73%
* 29405: PCHV 18.55%

### Model Visualizations

#### Figure: 30315 - Top 1 Performing ZipCode - Last Five Years



#### Figure: 30032 - Top 2 Performing ZipCode - Last Five Years



#### Figure: 29405 - Top 3 Performing ZipCode - Last Five Years

