2019-0703 IST 707 Data Analytics

Lab2 (week 6)

Ryan Timbrook
NetID: RTIMBROO
Course: IST 718 Big Data Analytics
Term: Summer, 2019

Assignment: Investment Opportunity Zip Codes Prediction

Lab2 (week 6)

Table of Contents

Lab2 (week 6)

1 Introduction

1.1 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
  - 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?

2 Analysis and Models

2.1 About the Data

2.1.1 *OBTAIN the Zillow Residential data*

Using the base data available from Zillow

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.

2.1.2 *Data Exploration & Cleaning*

2.1.2.1 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)}

2.1.3 *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

Date	FayetteVille_AvgHomePrice_Delta	HotSprings_AvgHomePrice_Delta	LittleRock_AvgHomePrice_Delta	Searcy_AvgHomePrice_Delta
2015	6375.317500	2114.583333	2886.665000	113.889167
2016	9934.127500	4350.000000	5505.000833	1731.945833
2017	13343.650833	6447.916667	7664.445833	2758.331667
2018	12990.872500	3660.416667	4445.555000	5063.889167
2019	10511.508333	4866.666667	3869.166667	6355.556667

Date	FayetteVille_AvgHomePrice_Delta_Percent	HotSprings_AvgHomePrice_Delta_Percent	LittleRock_AvgHomePrice_Delta_Percent	Searcy_AvgHomePrice_Delta_Percent
2015	4.535364	1.633566	1.984845	0.124452
2016	6.600610	3.251222	3.647138	1.857425
2017	8.143979	4.597650	4.832420	2.873179
2018	7.346212	2.543649	2.726493	5.010444
2019	5.610650	3.271251	2.317984	5.916426

FayetteVille_AvgHomePrice	
count	270.000000
mean	125629.047333
std	26780.873237
min	81878.950000
25%	102230.260000
50%	125547.500000
75%	142832.500000
max	188995.240000

HotSprings_AvgHomePrice	
count	270.000000
mean	116036.481481
std	22265.684348
min	72150.000000
25%	90106.250000
50%	126250.000000
75%	131612.500000
max	149075.000000

LittleRock_AvgHomePrice	
count	270.000000
mean	129622.271704
std	22889.048391
min	85206.670000
25%	106229.165000
50%	139551.670000
75%	144176.670000
max	167096.670000

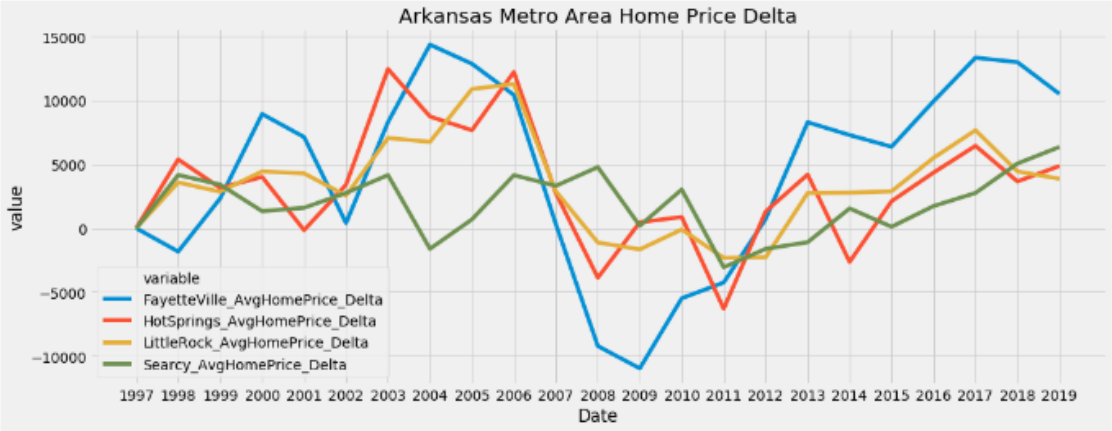
Searcy_AvgHomePrice Stats:	
count	270.000000
mean	85328.345704
std	10560.427209
min	62180.000000
25%	77475.000000
50%	89486.665000
75%	92545.000000
max	107816.670000

Lab2 (week 6)

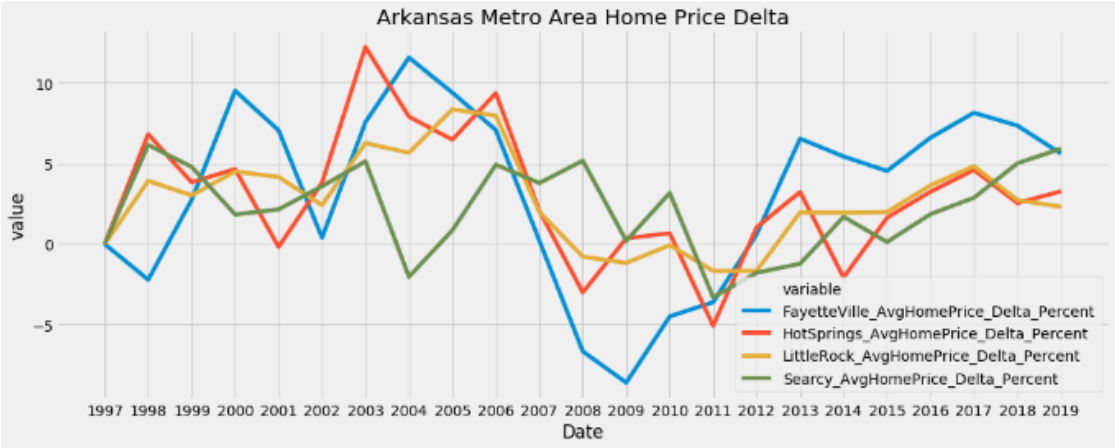
2.1.4 *Arkansas Metro Region Exploratory Data & Visualizations*

2.1.4.1 Annual Net Price Changes by Metro Region

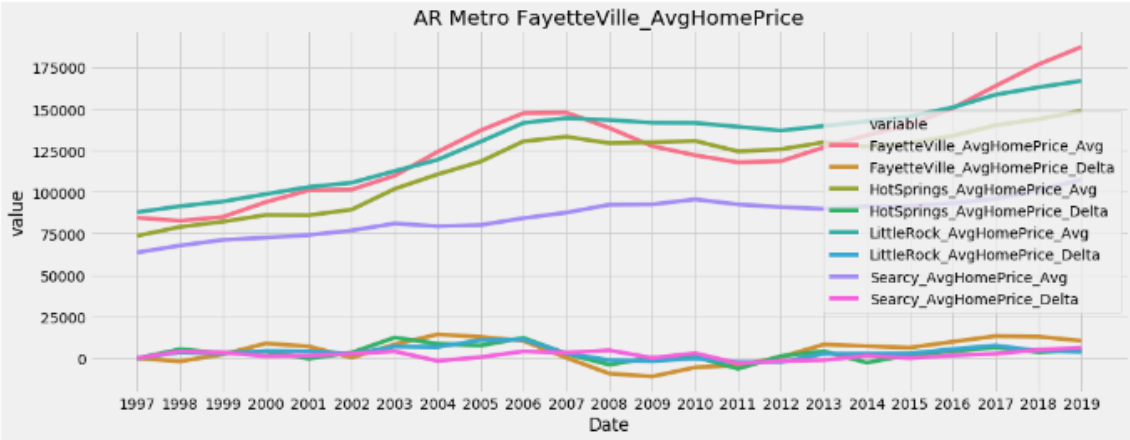
2.1.4.1.1..1 Figure: Annual Net Median Home Price Change



2.1.4.1.1..2 Figure: Annual Percent Net Median Home Price Change



2.1.4.1.1..3 Figure: Annual Net Price Change, Percent and Value



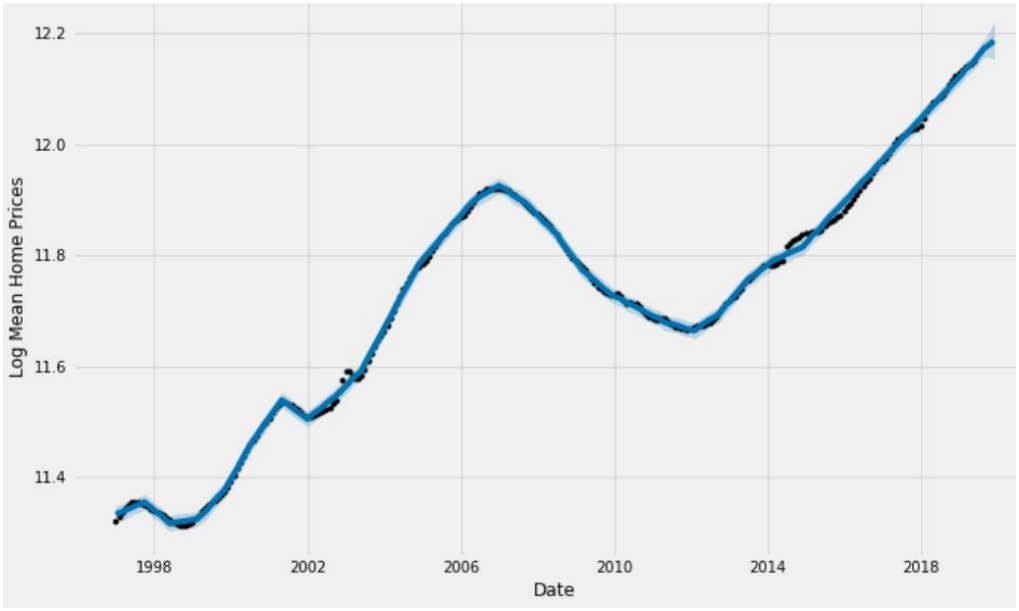
Lab2 (week 6)

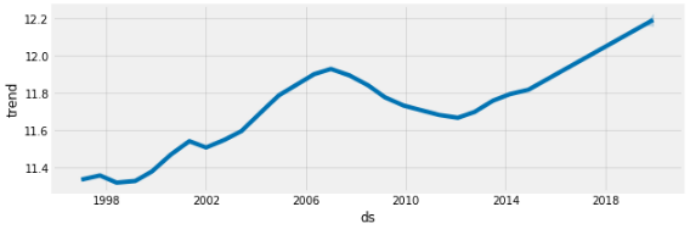
2.1.5 *Arkansas Metro Region Timeseries Plots*

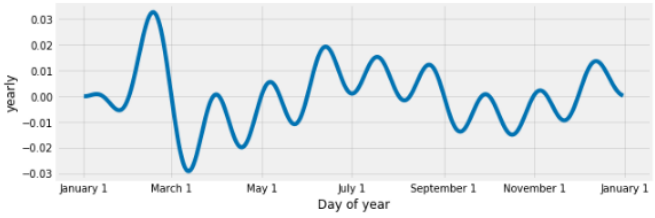
Provide an intial data analysis to include:

- 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

2.1.5.1 Figure: FayetteVille, Log Median Home Timeseries Forecast

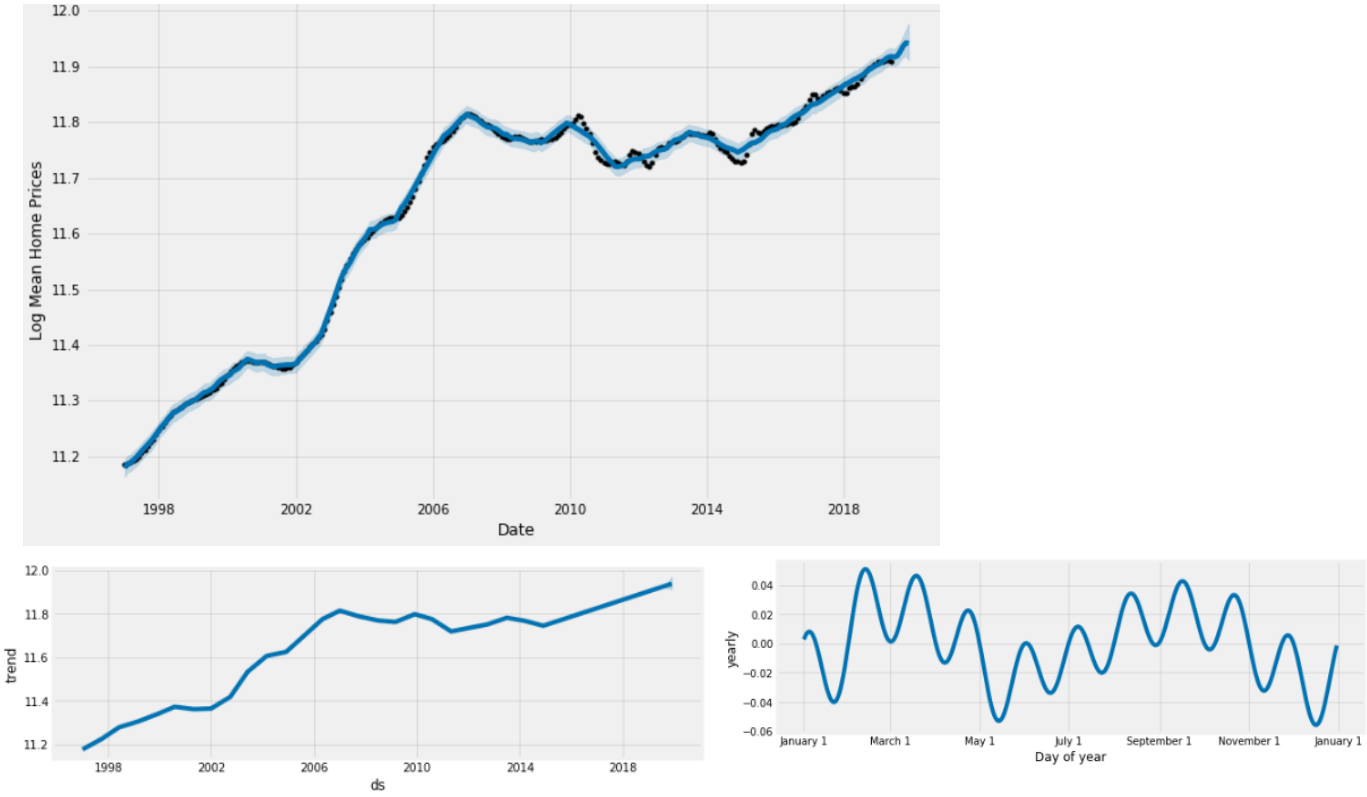




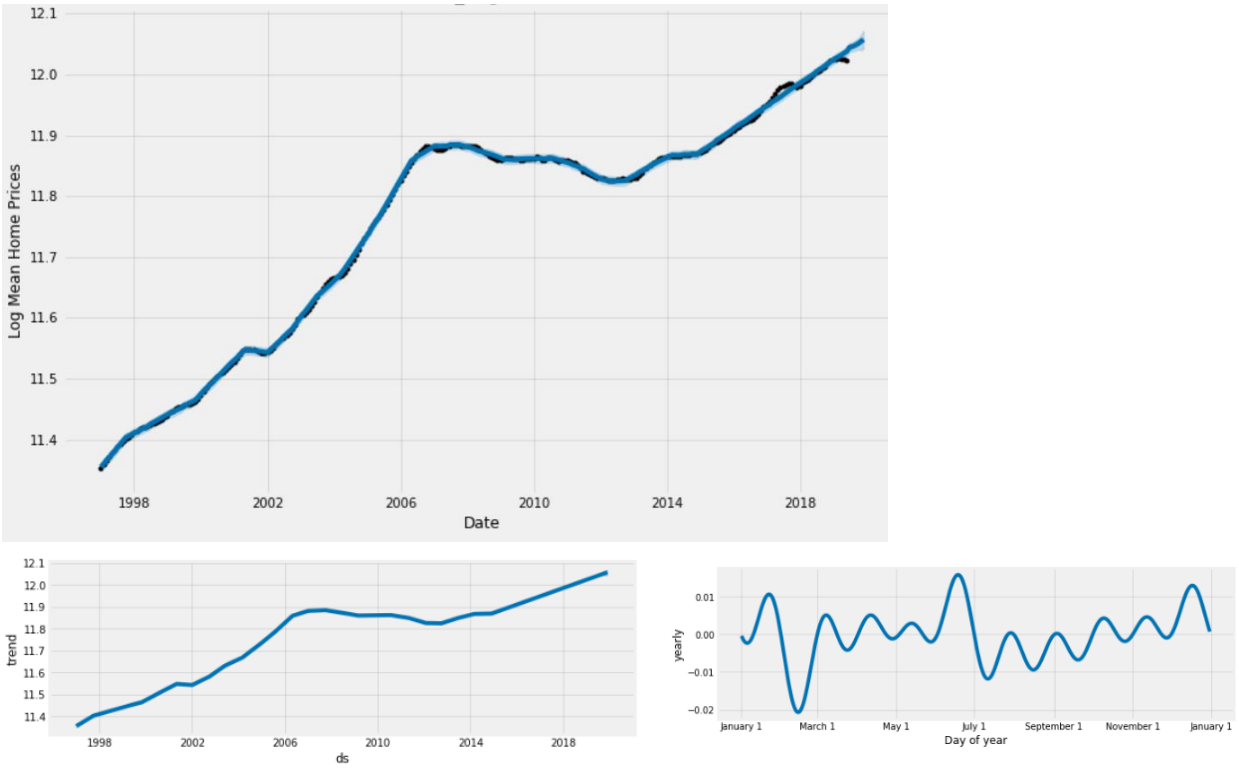


Lab2 (week 6)

2.1.5.2 Figure: HotSprings, Log Median Home Timeseries Forecast

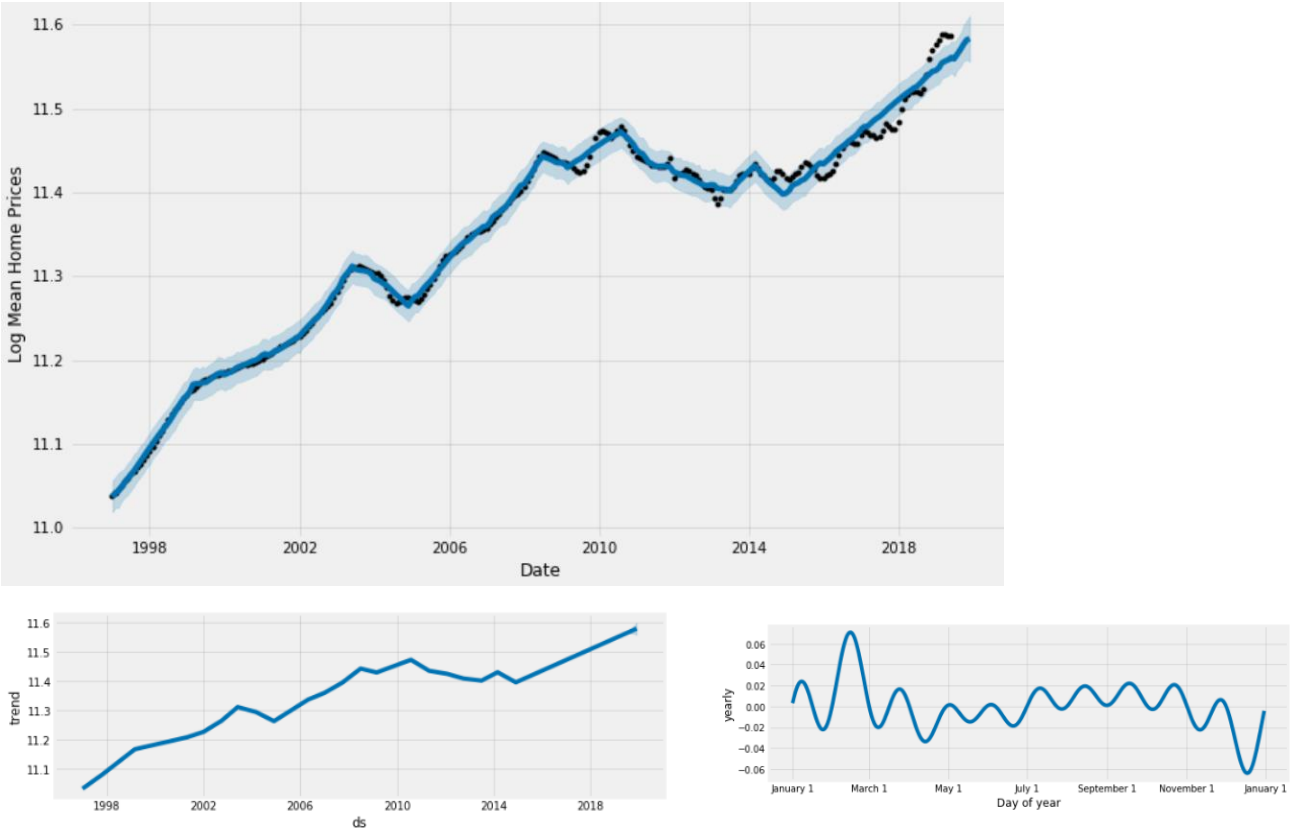


2.1.5.3 Figure: Little Rock, Log Median Home Timeseries Forecast



Lab2 (week 6)

2.1.5.4 Figure: Searcy, Log Median Home Timeseries Forecast



Lab2 (week 6)

2.2 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

Lab2 (week 6)

2.2.1 *Model Timeseries Forecast, ZipCodes Details*

2.2.2 *Top ZipCode Choose Matrix*

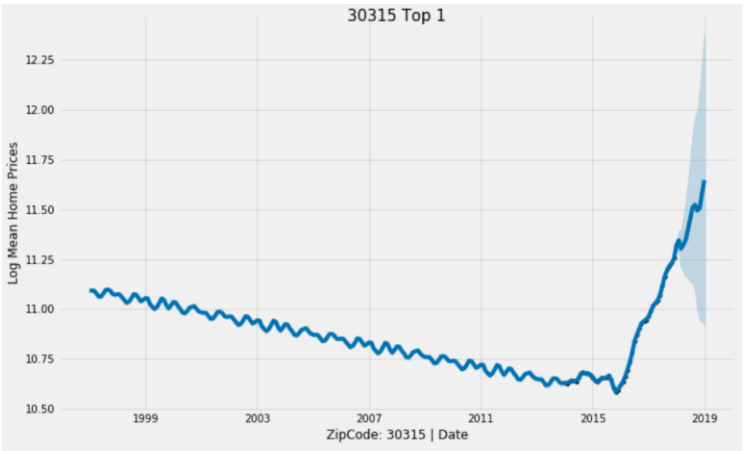
ZipCode	OneYearAvg	TwoYearAvg	ThreeYearAvg	FourYearAvg	FiveYearAvg
30315	33.053245	29.848357	26.043715	21.801677	22.149399
30032	28.471072	22.467730	23.055558	24.644119	19.727696
29405	27.446025	26.511627	19.336513	19.442008	18.547235
33805	22.389435	23.205006	19.139119	20.336521	17.014016
34785	27.272727	22.323290	18.771192	16.418539	16.957150
7050	28.118209	21.207066	16.696146	17.540963	16.712477
33714	20.922128	21.233407	19.921016	19.579158	16.331142
18102	24.529305	22.053236	17.361840	16.023187	15.993514
30288	24.226681	20.019458	18.553583	16.721854	15.904315
29661	24.605137	22.048447	17.972283	14.320429	15.789259

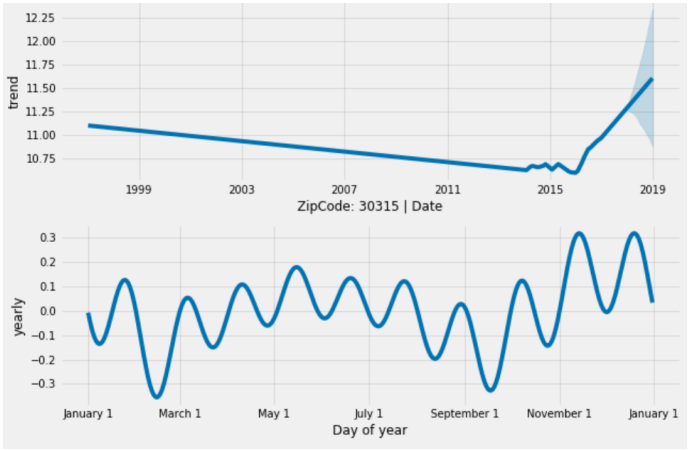
Top three best performing zipcodes over the prior 5 years:

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

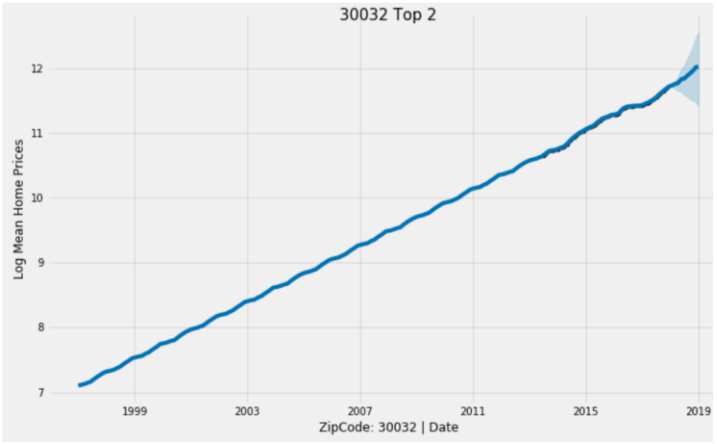
2.2.3 *Model Visualizations*

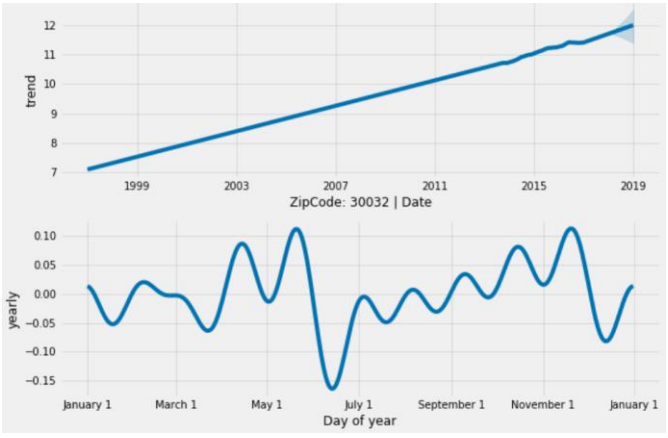
2.2.3.1 Figure: 30315 - Top 1 Performing ZipCode - Last Five Years





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





Lab2 (week 6)

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