ZILLOW HOUSE LISTINGS

Alejandro Mora

SYRACUSE UNIVERSITY iSCHOOL 2/17/2021

Problem

Good investment possibility predictions of wants and needs, can be a task worth doing to make sure the best result is received. If there are specific areas of which an investor can put their money into or specifications of what they can invest into, the right investment can be found. Multiple resources are available for finding a good investment, whether it be a housing market website, multiple databases, or realtors if they educated on a specific market area for growing potential. A thing that would have to be done manually would be forecasting future prices of potential real estate. Zillow offers databases on market data to allow access to areas of estate prices over time. With given data from the past, future data can be predicted. Many investors, whether they are individuals or companies can use this technique to see what is available to them and what can be a good investment for the future. A company looking to get into investing is the Syracuse Real Estate Investment Trust (SREIT). They are trying to find the best possible investment opportunity within the United States and are looking for the best Zip code areas that have those opportunities.

Data

As mentioned before Zillow offers datasets on housing and other real estate properties within their databases, valuing them over time. Multiple datasets can be obtained based on what real estate is trying to be acquired. For this forecast modeling, “SingleFamilyResidence” data will be obtained and looked at for potential investment. Within the data set it has plenty of observations and variables to work with to see what models can be built.

Variable Name	**Type**	**Description**	**Example**
RegionID	Int	Region ID Number	‘61639’
SizeRank	Int	Ranking of Zip Code area based on population and size, Urbanized ranking (Hong, 2017)	‘1’ ‘2’ ‘3’
RegionName	Int	Zip Code number	‘10025’
RegionType	String	Region type	‘Zip’
StateName	String	Abbreviation of state	‘NY’
State	String	Abbreviation of state	‘NY’
City	String	Location of City	‘New York’

Metro	String	Location of Metro area around city	‘New York-Newark-Jersey City’
CountyName	String	Name of County	‘New York County’
1996-1-31	Float	Price of listing at listed year (First Year, column 9)	‘364892.0’
….	…	…	…
2020-3-31	Float	Ending year, column 300	…

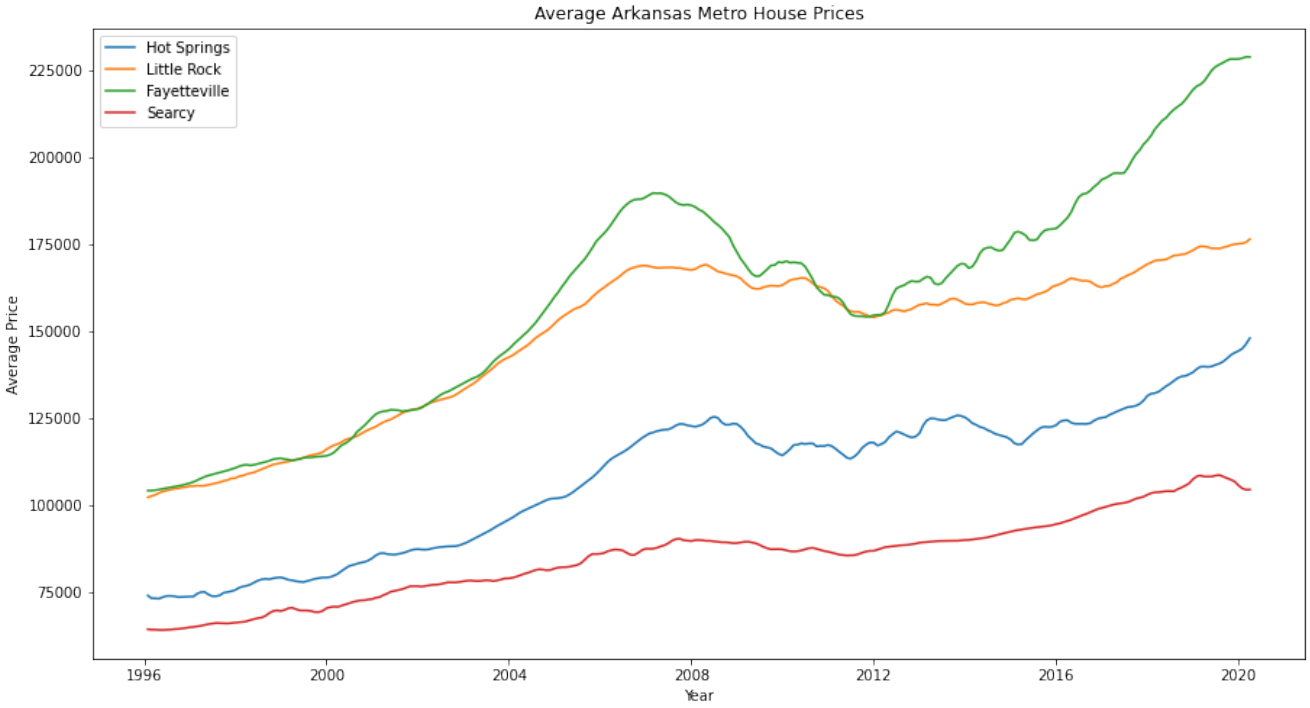
Many variables are not needed for time series analysis or they are non-essential to the dataset. Some can be taken out for being repetitive. Another transformation that must be done is to rearrange the time variables and prices for easier understanding when dealing with time series. The data set will have to be rearranged as follows and another variable will be created to help determine zip code ranking:

New Dataset:

Variable Name	**Type**	**Description**	**Example**
RegionName	Category	Zip Code number	‘10025’
SizeRank	Int	Ranking of Zip Code area based on population and size, Urbanized ranking (Hong, 2017)	‘1’ ‘2’ ‘3’
State	String	Abbreviation of state	‘NY’
City	String	Location of City	‘New York City’
Metro	String	Location of Metro area around city	‘New York-Newark-Jersey City’
CountyName	String	Name of County	‘New York County’
Percent Change	float	Percentage change of price from 1996-1-31 to 2019-12-31	‘2.66’
Year	datetime	Year-month-day	‘1996-01-31’
Price	float	Price listing at given year/date	‘364892.0’

This data set is helpful in time series analysis as dates are now variable values/observations rather than variables themselves. One thing that must be done when determining which investment is the best is deciding what factors are to be considered when looking at what is best for the company. No factors are listed as preferences, so the company is looking for the best area in which the United States has to offer for investments in general. The variable, SizeRank, has a good definition in finding good investments, the higher the ranking, the more urbanized the area is, which can have good investment opportunities.

The data can be sub-set to see if there are any trends along the data that can be seen better. Regarding the Arkansas metro area, the average pricing per zip code can be taken and plotted to see any trends.



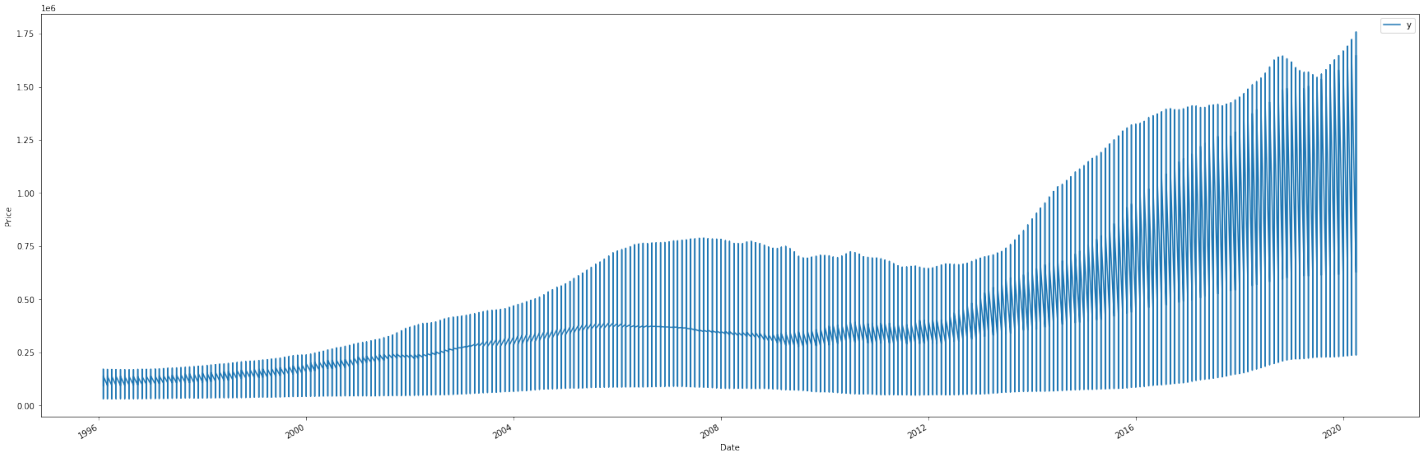
The data seems to dip after 2008 but continuously rises afterwards, this can be a sign from the recession taking place affecting the housing market. Since it is something that is not a seasonal occurrence, it can be left in the model amongst all prices as it affected the whole market around that time.

To narrow down the options, the top twenty five percent of the rankings can be taken into consideration. After down sampling down to the top twenty five percent, the top ten percent of percentage changes over time can be taken for the higher return on the real estate over the beginning of the data. Ten listings will be used for analysis and in order to remove area and city bias, the top listings by percentage change will all include the top zip codes from different urbanized areas. If that is not done, the top 10 listings are listed all around the Los Angeles area and even multiple listings in the same zip code, the top one can be taken for consideration.

<u>Top 10 areas/zip codes:</u>

- Los Angeles (90027)
- Oakland (94610)
- New York (11216)
- San Diego (92104)
- Palm Springs (92264)
- Jersey City (07302)
- Philadelphia (19125)
- Atlanta (30310)
- Alameda (94501)
- Boston (02125)

With these listings, the average prices can be taken within each zip code and used for model analysis to see which one offers better investment opportunities.

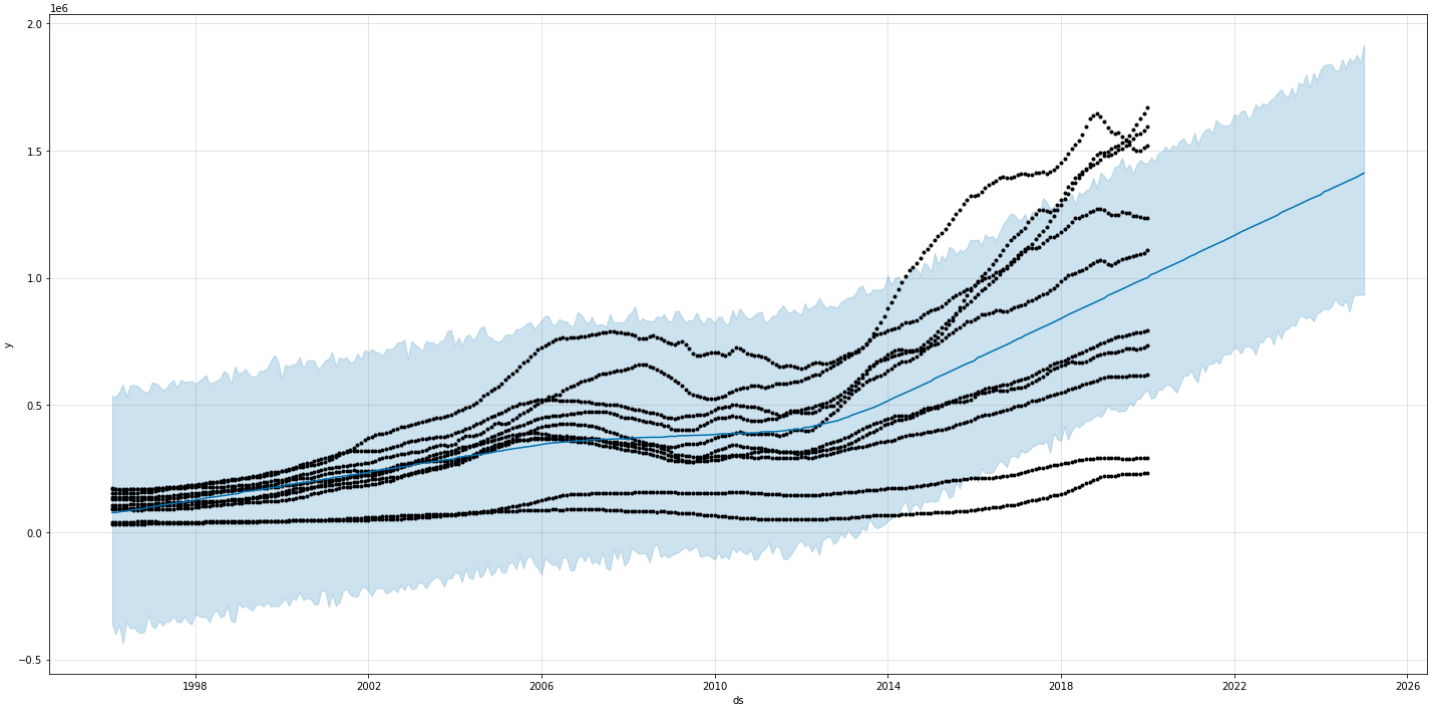


After 20 years of data, the averages tend to spread out more within all the zip codes. This can be due to physical area attraction to consumers or opportunities within the areas themselves which can cause higher or lower real estate prices. Some tend to stay the same, while others are increasing exponentially.

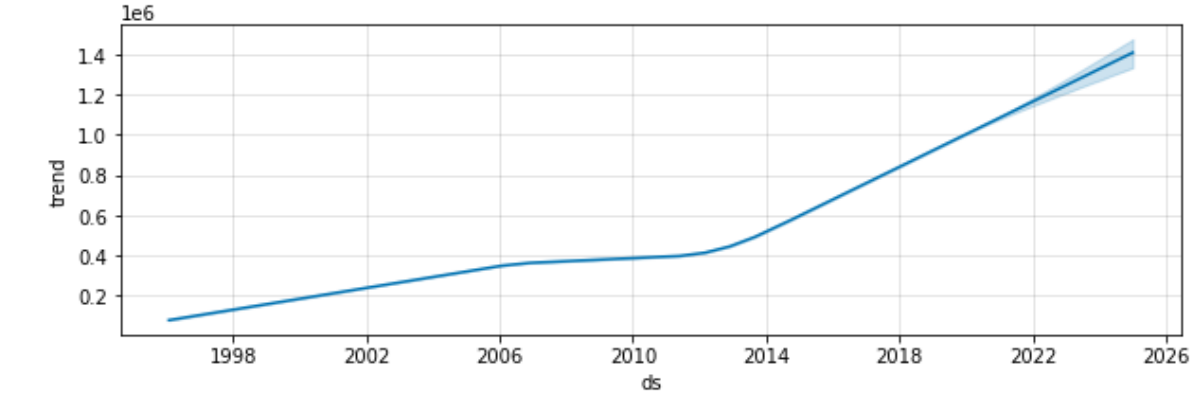
Model

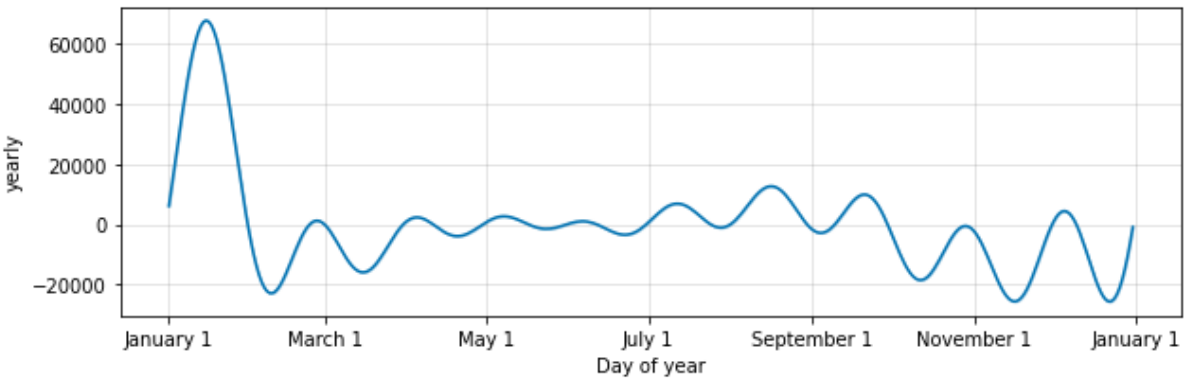
The model used for analysis is an additive time series model within python called “Prophet” created by Facebook, all components are added together to determine future prices. This is used because there is a recurring trend within the data as seen on the plot above of house prices increasing upward. The small dip in prices was due to an unforeseen event within the country that is not recurring throughout the years. With this being said, an additive model can be used taking everything from the past for forecasting. The data for the model is down-sampled from the whole data set, this is because since there are no specific requirements for having the best investment, they are hypothetically created (top 25% Size Rank, top 10% Percent Change, top 10 in those Percent Change). Confidence interval will be set to .95 for good accuracy levels. And data will be forecasted 60 months past the cutoff date which is, 2020-1-1. Last five dates predicted:

index	ds	yhat	yhat_lower	yhat_upper
343	2024-08-31	1.382526e+06	866108.249612	1.863124e+06
344	2024-09-30	1.388581e+06	927003.332834	1.847662e+06
345	2024-10-31	1.395795e+06	934565.518135	1.876599e+06
346	2024-11-30	1.403470e+06	933230.420415	1.846740e+06
347	2024-12-31	1.411387e+06	933691.856513	1.916945e+06



Some prices are seen to outward of the upper and lower predicted evaluations, the ones not desirable would be the areas that are steadily decreasing away from the projected trendline.





Recommendations and Outcomes

With implementing the model created on the subset data, the model can be fitted into the data and help create a trend difference in actual prices and where they are headed. The actual average price given on the last day recorded can be taken and compared to the trend line. Some prices showed decreasing values away from the predicted trend which showed decreasing or no investment promise for the future. After using the additive forecasting model. The highest difference from actual price and predicted trend gave us these three zip codes with high difference and percent change within the data set.

	RegionName	TrendDiff	PercentChange(In Whole Numbers)
2875	**90027**	664809.532027	13.177882
2879	**94610**	592592.532027	11.615760
2872	**11216**	517118.532027	10.002523

Recommended Zip Code areas for investments:

90027 – Los Angeles, California
94610 – Alameda, California
11216 – Brooklyn, New York

Data and References:

Zillow Housing Data:

https://files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv

Variable Dictionary Help:

https://sls.gatech.edu/sites/default/files/documents/Toolkit-Docs/hong_fieldguide_zillow.pdf