Zip Code Recommendations

LAB 2

Michelle Mak
IST 718 | BIG DATA ANALYTICS

Overview

This case study investigates how we can predict three zip codes that provide the best investment opportunity for the Syracuse Real Estate Investment Trust (SREIT). Using historical data of housing prices, it will calculate the compound average growth rate of every zip code from 1997 - 2018 in order to identify the zip codes with the best CAGR. After the data has been down sampled, linear regression will be performed using FBProphet to predict future house pricing trends.

Research Objectives

This investigation is carried out in 3 parts:

1. Exploratory Analysis

The data will be explored with a quick exercise by developing a time series plot for the Arkansas metropolitan areas of Hot Springs, Little Rock, Fayyetteville, and Searcy. The time series plot will be limited to the period of 1997 to present.

2. Down Sampling

Due to the largeness of the dataset, down sampling will be performed through the measurement of Compound Annual Growth Rate (CAGR). The 10 zip codes with the highest CAGR will be further investigated in part 3.

3. Forecasting

Regression models will be created for the 10 identified zip codes, using historical data from 1997 - 2017 as training data. Housing prices from 2018 will then be used to test the model and the Root Mean Squared Error will be calculated for each zip codes and used as a second metric to determine the best zip codes for investment.

Data Cleaning and Preprocessing

Base data is provided by the IST 718 course via Zillow:(files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv) The data initially has 15,533 zip codes and 281 columns of data.

The data must first be scrubbed and transformed from wide format to long format so that the housing prices can be grouped by year.

Initial scrubbing changes formatting for consistency (i.e. column names to lower case, renaming of columns, etc.) The zip code column also needs to be formatted so that leading 0's will still appear. Then, null values are identified and replaced or dropped. Finally, data types are transformed and the melt function is applied to flip the year columns to rows, which can then be grouped together by mean.

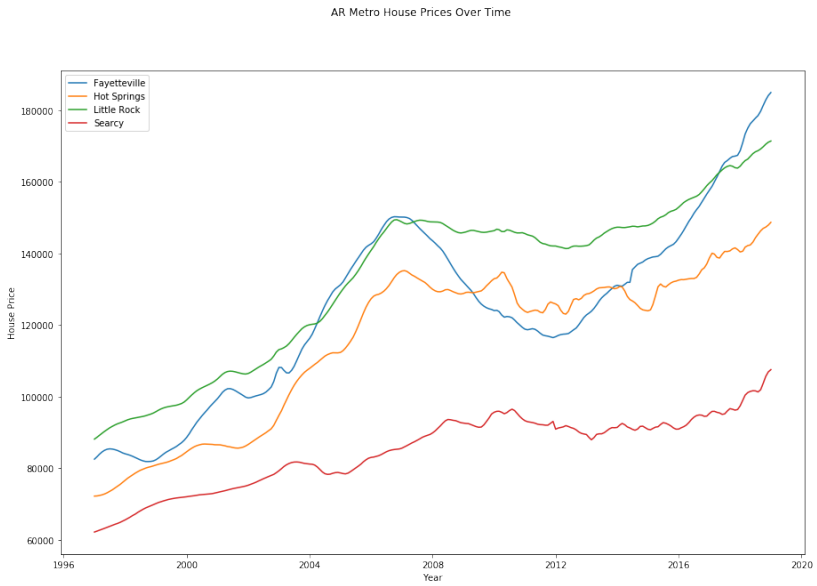
```
zillow_group = zillow_melt.groupby(['regionid', 'zipcode', 'city', 'state', 'metro', 'countyname', 'sizerank', 'year',
'date'])['houseprice'].mean().reset_index()
zillow_group.head()
```

	regionid	zipcode	city	state	metro	countyname	sizerank	year	date	houseprice
0	58196	01001	Agawam	MA	Springfield	Hampden County	5955	1997	1997-01-01	111900.0
1	58196	01001	Agawam	MA	Springfield	Hampden County	5955	1997	1997-02-01	112100.0
2	58196	01001	Agawam	MA	Springfield	Hampden County	5955	1997	1997-03-01	112300.0
3	58196	01001	Agawam	MA	Springfield	Hampden County	5955	1997	1997-04-01	112500.0
4	58196	01001	Agawam	MA	Springfield	Hampden County	5955	1997	1997-05-01	112800.0

PART 1: Develop a Time Series Plot for Arkansas Metro Areas

The data for the metro areas of Hot Springs, Little Rock, Fayetteville, and Searcy are called and put into their own datasets. The identifiers used to call each metro area includes the state as well, in case there are multiple metropolitan areas with the same name outside of Arkansas.

The four datasets are then concatenated into a new dataset, which is then used to find the mean of each metropolitan area by year. This is the dataset that is ultimately used to create the time series plot.

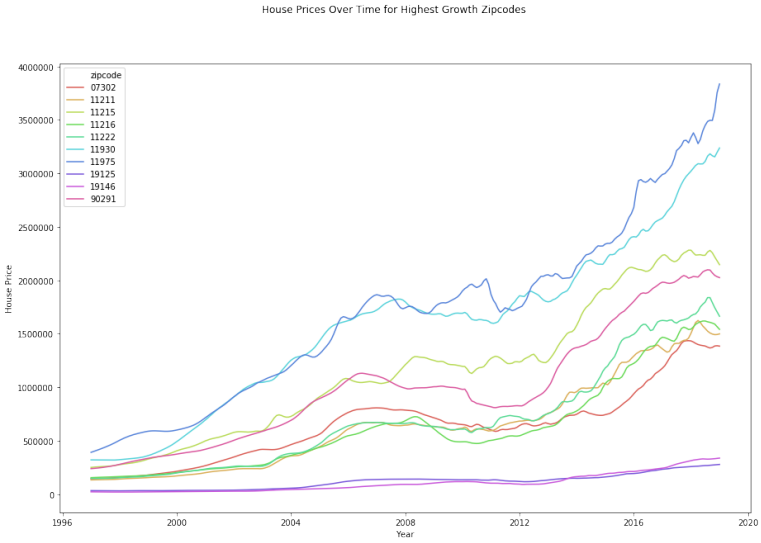


PART 2: Down Sampling with CAGR

The ten best investment opportunities will be identified by looking at Compound Annual Growth Rate (CAGR) for each zip code. This will be completed by creating a column that calculates the CAGR based on the housing prices at the beginning of 1997 and the end of 2018. The dataset will be sorted in descending order of CAGR to identify the zip codes with the best annual growth.

	zipcode	city	state	1997-01	2018-12	CAGR
690	19146	Philadelphia	PA	21400.0	334300	12.704988
1159	11222	New York	NY	146600.0	1703900	9.341441
104	11211	New York	NY	133100.0	1492900	8.996770
476	11216	New York	NY	154100.0	1566000	8.099913
12207	11930	Amagansett	NY	320300.0	3196400	7.943918
167	07302	Jersey City	NJ	143500.0	1387900	7.681153
14538	11975	Wainscott	NY	390900.0	3755400	7.625822
105	11215	New York	NY	249800.0	2181600	6.877066
1837	90291	Los Angeles	CA	238500.0	2033600	6.699356
3560	19125	Philadelphia	PA	32800.0	276000	6.603017

Looking at initial results, it seems like we have a very solid list of zipcodes that reflect substantial growth since 1997. But looking at the housing prices from December 2018, there is a large difference between the highest and lowest housing price. Let's graph the results to get better understanding of the data.

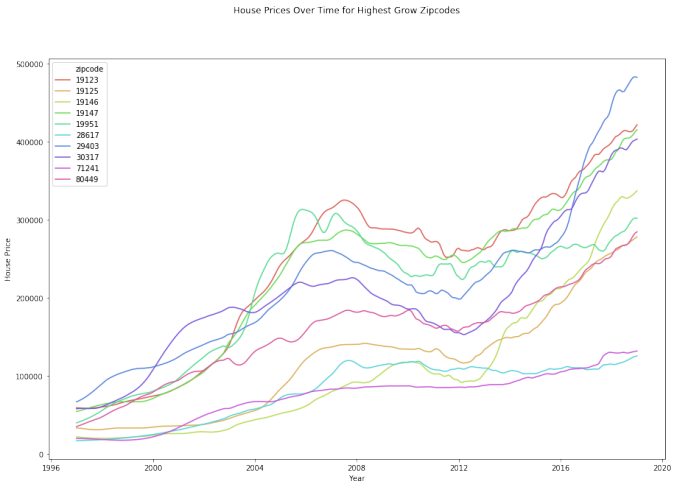


Results initially show that 19146 (Philadelphia), 11222 (New York), and 11211 (New York) have the highest rate of compound annual growth from the years 1997 - 2018, which would theoretically make them the top contenders for the SREIT fund.

However, upon further analysis of the visualized data, it is clear that the two zip codes from New York, which demonstrate enormous growth, are actually very expensive.

Thus, we will attempt the CAGR analysis again -- only this time, we will filter out all recent house prices above $500,000 to allow for maximal growth potential.

	zipcode	city	state	1997-01	2018-12	CAGR
690	19146	Philadelphia	PA	21400.0	334300	12.704988
3560	19125	Philadelphia	PA	32800.0	276000	6.603017
14931	80449	Hartsel	CO	34600.0	282400	6.385333
14377	19951	Harbeson	DE	39300.0	302000	5.973363
802	19147	Philadelphia	PA	53800.0	412300	5.955291
13969	28617	Crumpler	NC	16500.0	124500	5.853160
4694	29403	Charleston	SC	66300.0	483100	5.629045
5057	19123	Philadelphia	PA	58700.0	417600	5.479554
6789	30317	Atlanta	GA	57300.0	401900	5.392628
7972	71241	Farmerville	LA	19300.0	130900	5.191455



PART 3: FORECASTING WITH PROPHET

So far, the zip codes of 19146, 19125, and 80449 look the most promising as they exhibit affordable median housing prices and have the highest growth rates in this tier. To further solidify the options, linear regression models will be computed for all 10 zip codes for the CAGR_df2.

The melted Zillow data (zillow_group) must be further processed in order to create the model. First, the dataset needs to be broken out into training and testing datasets for each zip code. Then, training set for each zip code will be run through fbprophet. The Root Mean Square Error will be computed based on the prediction of the model and saved into a new column of CAGR_df2 to provide another decision metric.

RMSE is calculated by finding the square root of the mean squared error, which is computed using mean_squared_error from the sklearn.metrics.

zipcode	city	state	1997-01	2018-12	CAGR	RMSE
19146	Philadelphia	PA	21400.0	334300	12.704988	20009.363188
19125	Philadelphia	PA	32800.0	276000	6.603017	2610.831148
80449	Hartsel	CO	34600.0	282400	6.385333	10661.546786
19951	Harbeson	DE	39300.0	302000	5.973363	15397.082609
19147	Philadelphia	PA	53800.0	412300	5.955291	16930.756735
28617	Crumpler	NC	16500.0	124500	5.853160	7825.614127
29403	Charleston	SC	66300.0	483100	5.629045	51881.058094
19123	Philadelphia	PA	58700.0	417600	5.479554	5037.458003
30317	Atlanta	GA	57300.0	401900	5.392628	14113.845364
71241	Farmerville	LA	19300.0	130900	5.191455	5523.488619

Conclusion

Now that we have the preliminary information needed to make an educated investment decision, we will sort the table be descending CAGR, ascending RMSE, and descending housing prices in December of 2018. These three views will show us the zip codes with the top growth rate, most stability, and most affordable housing costs respectively. This analysis will help us the recommendation.

Recommendation

Based on the results, the recommendation for the SREIT would be 19146, 19125, and 28917.

19146 (Philadelphia, PA):

This is the top recommended zipcode because it exhibits enormous annual growth at 12.7%. It is also midrange in median housing price. The only downside of this zip code is that the RMSE is quite high. But, as the saying goes: high risk, high reward.

19125 (Philadelphia, PA):

The next recommended zipcode derives from the fact that it has the lowest RMSE at 2610, in addition to having the 2nd highest CAGR within this batch. The low RMSE will help to balance out the portfolio, since 19146 is such high risk.

28617 (Crumpler, NC):

The final recommended zipcode is 28617 because it has the lowest median housing cost, has an above average RMSE score. It is ranked #6 in terms of CAGR, but the low cost and above average stability make it a more worthwhile investment.

References

CAGR: https://www.investopedia.com/terms/c/cagr.asp
CAGR: https://stackoverflow.com/questions/37355924/pandas-calculate-cagr-with-slicing
FBProphet: https://facebook.github.io/prophet/docs/quick_start.html#python-api
RMSE: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde