Jay Mahapatra

Lab 6- Can we predict which three zip codes provide the best investment opportunity for the Syracuse Real Estate Investment Trust (SREIT)?

IST 718- Big Data Analysis

2/23/2019

**Goal:**

* First understand the characteristics of the housing data across zip codes and then understand what elements are important in determining those prices. Then use that knowledge to narrow down which zip codes to focus on and then run a forecast on those zip codes using the Facebook Prophet forecasting model to predict housing prices.

**Hypothesis:**

* A positive change in sales from year to year will represent the best investment opportunity for SREIT. In addition, population growth (specifically the change in population from year to year) is the biggest factor in change in sales. Thus, zip codes that have the biggest increase in population from year to year will have the biggest increase in sales from year to year which in turn represents the best opportunity for SREIT.

**Data Sources:**

Zillow Data - files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_SingleFamilyResidence.csv

* + - Median Sales data 1997-2018
    - Zip Codes
    - City
    - Rank Size
    - Metro Area
    - County Name
* Loaded data in two parts and dropped columns prior to 1/1997 (see #Obtain Lab 6 .ipynb).

Population Data - <https://simplemaps.com/data/us-cities>

* Updated regularly with data from the census up to 2016.
* Loaded data in and combined with Zillow data to then run a regression analysis (see #Loaded in US Cities population data Lab 6 .ipynb).
  + - Population 12/2000
    - Population 12/2013
    - Population growth from 12/2000-12/2013

**Summary Statistics:**

* Not surprisingly, zip codes in New York City and San Francisco represent 60% of the top 10 average housing prices in 2018.
* Both cities are peninsulas that have limited space for development and thus the price for land is lucrative and at a premium.
* (Graphs re-created in excel for visual quality but originally created in Colab)
* Over the duration of the time range the top zip codes have shown a 7x-14x increase in sales.
* All of the zip codes are centered around up and coming neighborhoods in established cities where young people are now settling.
* In addition, many of these places were formerly considered undesirable places/unsafe places to live within their respective cities.
* These zip codes include Kensington in Philadelphia, Williamsburg in Brooklyn within New York City, Jersey City in New Jersey, and Long Beach in Los Angeles all of which are noted place where young people are currently settling within these respective cities as well as are neighborhoods in or are adjacent cities to established cities.
* (Graphs re-created in excel for visual quality but originally created in Colab)
* (I also looked at short term (6/18-12/18 and mid-term 12/17-12/18 percent change but in running forecasts found much less certainty in those numbers)

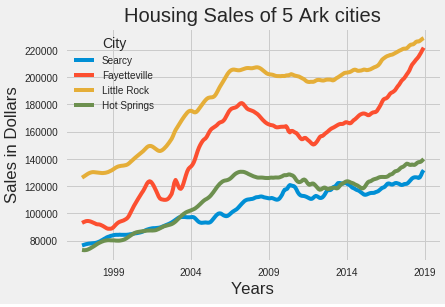
**Regression Analysis**

* Two regression analyses were run to determine if change in sales, population, and population growth have an affect on current housing prices per the hypothesis stated above.
* In looking at the change in sales, change in population, and current populations effect on 2013 housing prices the following was determined:
  + The factors only accounted for 37% of why a housing price was the way it was in 2013.
  + Population in the year 2000 and 2013 as well as population growth were both not significant factors.
  + Change in sales is a significant value. An increase in sales corresponds to a $492K increase in a house price with a standard error of +/- $26K
* Another regression was run that looked at the effect of change in sales, size rank, and previous sales price on current housing price for 12/2017 compared to 12/2018.
* The following was determined:
  + The factors accounted for 99% of the 2018 sales price.
  + Change in sales was the most significant factor. An increase in sales percentage corresponded to a $204K increase in a housing price with a standard error of approx. $4K
* Both regressions suggest that a larger positive change in sales is an OK indicator of a zip codes increased value.

**Questions:**

***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 metro area level***



* Values are averaged together each year by the city/metro area level.

***Develop models for forecasting average median housing value by zip for 2018***

***Use historical data from 1997 through 2017 as your training data***

***Integrate data from other sources***

* See regression analysis section.
* A linear regression that combined population data from 2000 and 2013 as well as the sales data from Zillow were used to determine which factors were important in determining the price of a house. Another regression was done to look at simply change of sales and size rank on current housing price.
* The regression showed that change in sales is an OK indicator of housing price.

***What technique/algorithm/decision process did you use to down sample?***

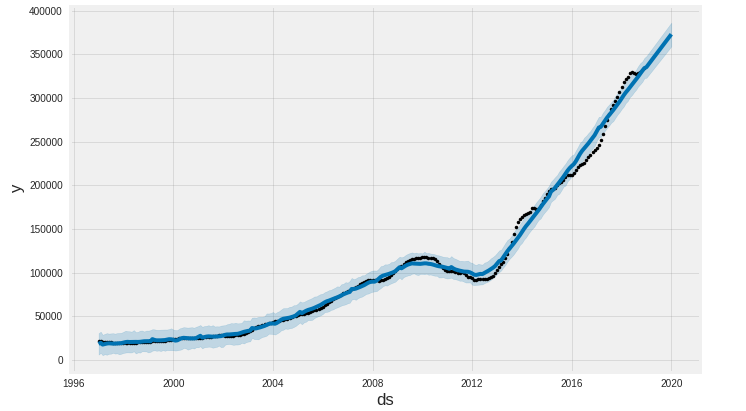
* Summary statistics, multiple linear regressions, and even the Facebook Prophet forecasting model were used to down sample to zip codes that would provide the best investment opportunity.
  + Multiple linear regression was used to determine that change in sales does contribute to a positive increase in a house’s current price.
  + Summary statistics combined with outside research on demographic information was used to determine which zip codes had the highest change in sales at different time periods (long-1997-2018,mid-12/2017-12/2018, short-6/2018-12/2018).
  + The Facebook profit model was used to determine that a long-term change in sales time interval was the best investment opportunity as a shorter time frame proved less predictable and consistent.

***What three zip codes provide the best investment opportunity for the SREIT?***

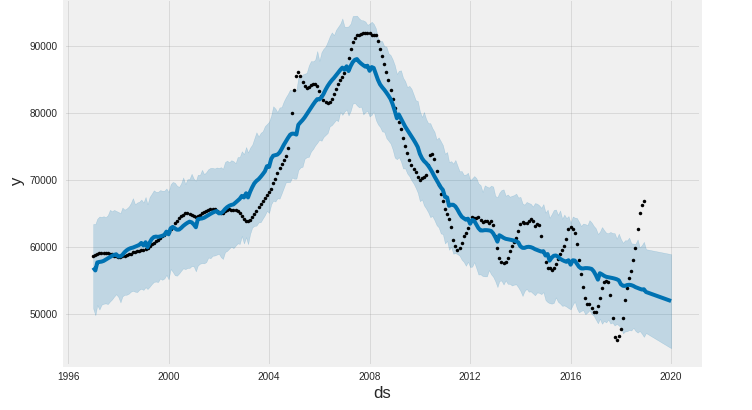
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Zip Code | Area | 12/2019  Lower Bound | 12/2019  Forecast | 12/2019  Upper Bound | Percent Increase from 12/2018 |
| 19146 | Northwest side of Philadelphia | $359,037 | $372,892 | $386,135 | 7%-15% |
| 11222 | Little Poland/Brooklyn-New York City | $2,031,696 | $2,140,302 | $2,242,755 | 4%-15% |
| 11211 | Williamsburg/Brooklyn-New York City | $1,731,885 | $1,795,684 | $1,862,666 | 5%-13% |

***Why?***

* First based upon the percent change in sales from 1997-2018 these three zip codes represent the highest overall positive change in sales during the time-period.
* Second these three zip codes show stable and consistent growth compared to other zip codes that have had higher shorter-term growth. As a result, the long-term growth zip codes are more predictable when forecasted and represent less risk in losing money when investing.
  + For example, here are the housing prices from 1997-2018 for zip code 19146 in Philadelphia as well as zip code 31906 in Columbus which had one of the highest percent change in sales from 2017 to 2018.
  + As a result, the forecast is much more predictable for Philadelphia compared to Columbus and even suggests the growth in Columbus is an outlier:
    - 19146-Philadelphia



* + - 31906-Columbus



* Finally, these long-term sales growth areas possess characteristics that make it primed to increase in value which include:
  + Underdeveloped neighborhoods surrounded by established cities and/or growing cities suggesting the money to support these neighborhoods will not dry up as they get developed.
  + These neighborhoods are also likely to get developed as more and more people look to move to these cities.
  + These zip codes are inhabited by lots of young people suggesting they are looking for places to rent or soon will be looking to buy a home. For example, here is the description for Brooklyn zip code 11211 per unitestateszipcodes.org.
    - “The people living in ZIP code 11211 are primarily white. The number of people in their late 20s to early 40s is extremely large while the number of middle-aged adults is small. There are also an extremely large number of single adults and an extremely small number of families. The percentage of children under 18 living in the 11211 ZIP code is large compared to other areas of the country”

**Conclusion/Next Steps:**

* The results and recommendation both concur with and contradict the original hypothesis. It was right to assume that the change in sales accounts for a decent amount of a house’s current price. However, it was wrong to think that population or growth in population accounts for a housing price or even helps to account for the change in sales.
* If I was to continue this analysis further, I would like to include age-demographic data by zip code to see if that has a large affect on current housing prices as that seems to be a consistent characteristic with the zip codes that had the largest percent increase from 1997-2018.
* In addition, I would like to create a function that would take the percent increase/decrease in sales year by year and then average them for each of the zip codes to show which zip codes on average consistently have an increase in price as I think those zip codes would be even better investments then the zip codes I recommended.

**References:**

* Smit/Beth were a huge help!
* files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_SingleFamilyResidence.csv
* <https://simplemaps.com/data/us-cities>
* unitestateszipcodes.org