Shuhan Zhang

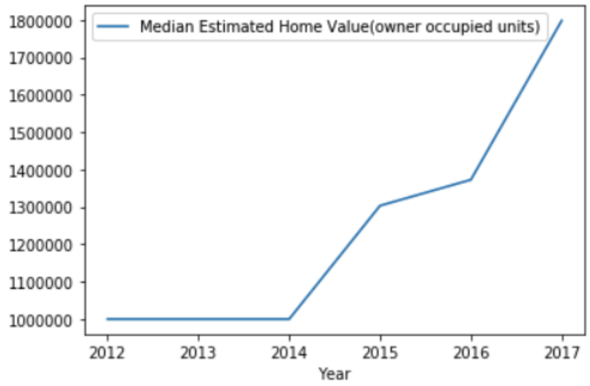
Aspen Capital

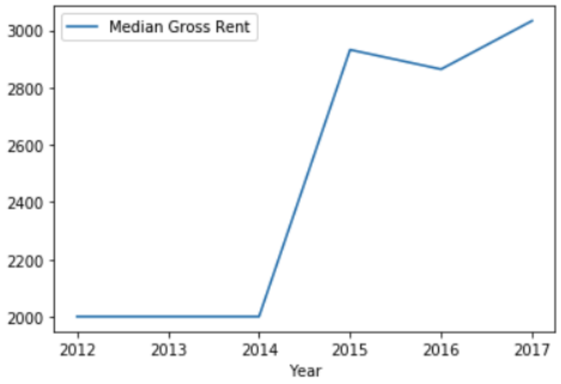
10/25/2022

Take-home Data Science Case

CONCLUSION

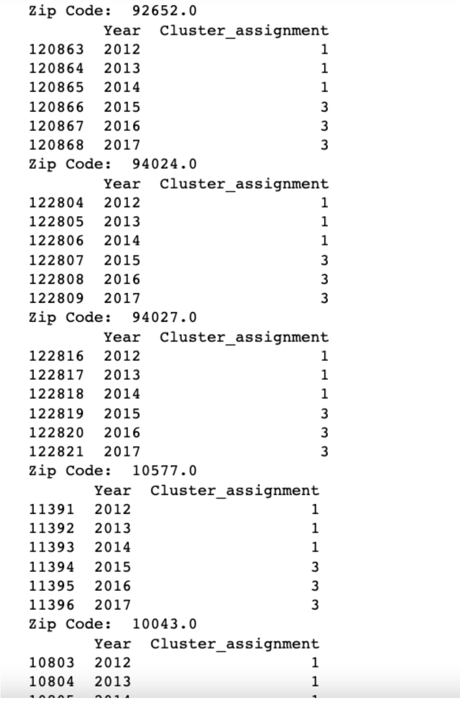
1. The top 10 zip codes to focus on property investment are: **92652, 94024, 94027, 10577, 10043, 94022, 94020, 94301, 90272, 94925**. These are the top 10 zip codes which have a high forecasted index value based on the consideration of both the median home value and median rent for year 2023. By plotting both the median home value and median rent, we can tell there is a huge lift from 2012 to 2017. Let's look at an example of zip code 92652.



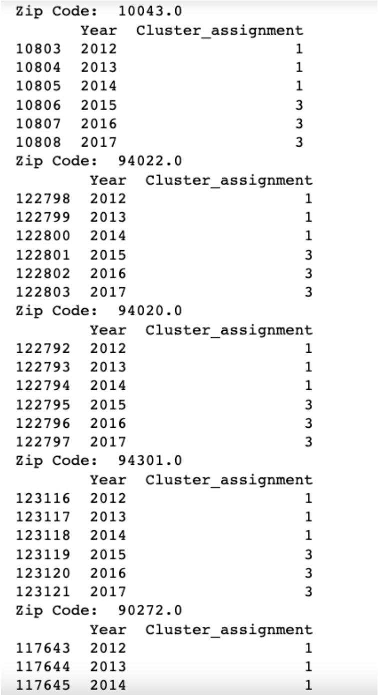


2. We can observe from the results that almost all the 10 zip codes are from the same state which is California, with 2 from New York. Also, we can notice that the zip codes are pretty close to each other. Also, based on the clustering result from the KMeans model, we can see that physically close zip codes got assigned to the same clusters for almost

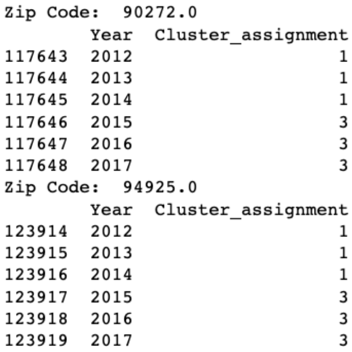
each individual year from 2012 to 2017 (see pics below). So, we can observe a trend that zip codes which are close to each other will have a relatively similar mutual effect on each other, positive or negative.



```
Zip Code:  92652.0
        Year  Cluster_assignment
120863  2012                   1
120864  2013                   1
120865  2014                   1
120866  2015                   3
120867  2016                   3
120868  2017                   3
Zip Code:  94024.0
        Year  Cluster_assignment
122804  2012                   1
122805  2013                   1
122806  2014                   1
122807  2015                   3
122808  2016                   3
122809  2017                   3
Zip Code:  94027.0
        Year  Cluster_assignment
122816  2012                   1
122817  2013                   1
122818  2014                   1
122819  2015                   3
122820  2016                   3
122821  2017                   3
Zip Code:  10577.0
       Year  Cluster_assignment
11391  2012                   1
11392  2013                   1
11393  2014                   1
11394  2015                   3
11395  2016                   3
11396  2017                   3
Zip Code:  10043.0
       Year  Cluster_assignment
10803  2012                   1
10804  2013                   1
```



```
Zip Code:  10043.0
       Year  Cluster_assignment
10803  2012                   1
10804  2013                   1
10805  2014                   1
10806  2015                   3
10807  2016                   3
10808  2017                   3
Zip Code:  94022.0
        Year  Cluster_assignment
122798  2012                   1
122799  2013                   1
122800  2014                   1
122801  2015                   3
122802  2016                   3
122803  2017                   3
Zip Code:  94020.0
        Year  Cluster_assignment
122792  2012                   1
122793  2013                   1
122794  2014                   1
122795  2015                   3
122796  2016                   3
122797  2017                   3
Zip Code:  94301.0
        Year  Cluster_assignment
123116  2012                   1
123117  2013                   1
123118  2014                   1
123119  2015                   3
123120  2016                   3
123121  2017                   3
Zip Code:  90272.0
        Year  Cluster_assignment
117643  2012                   1
117644  2013                   1
117645  2014                   1
```



```
Zip Code:  90272.0
        Year  Cluster_assignment
117643  2012                   1
117644  2013                   1
117645  2014                   1
117646  2015                   3
117647  2016                   3
117648  2017                   3
Zip Code:  94925.0
        Year  Cluster_assignment
123914  2012                   1
123915  2013                   1
123916  2014                   1
123917  2015                   3
123918  2016                   3
123919  2017                   3
```

MAIN APPROACH

To answer the questions of what are the zip codes which should be focused on investing, the first thing is to analyze the dataset. The census dataset consists of different attributes of one zip code at different tract number from year 2012 to 2017 like the population, the home value, the income level, the average rent etc. For the sake of this project, I decided to mainly focus on the **home value** and **rent** as the two key metrics.

As I want to consider both home value and rent, I decide to generate a new index which combines these two metrics together. **The new investment index is** $i = home\ value \times rent$**.**

The main approach is to build a time series first-order **ARIMA** model to forecast the future investment index value of each zip code by using the census data from 2012 to 2017.

ASSUMPTIONS

1. The housing market doesn't fluctuate a lot because of the pandemic. I am only able to model the data which is from 2012 to 2017 to forecast the future value after 2017.
2. The cost of the property investment remains constant or irrelevant. In a more realistic situation, it would be better to take into account the cost factors.
3. There is no huge inflation or deflation.

FUTURE WORK TO BE DONE

1. Data Preprocessing. In the data preprocessing step, I decided to drop the rows as long as it has one empty data entries. However, if I have more time, I would only drop the rows where the zip code is missing, and check each column to see what are the feature is and decide a way to impute missing values.
2. Metric Selection. Here in this project, due to time constraints, I only selected home value and rent as the two main factors to look at and keep track of. However, I do believe that there is a better way to come up with a new investment index by figuring out what are the key components of property investment with investment experts and business experts. If there is more room for this project, I also would suggest to collect the direct investment data which includes the information of property investment returns. In this way, we can build supervised learning model to explore the most important factors which could lead to the highest investment returns.

3. Time Series Model. Currently, I'm using the first order auto-regressive time series model because it is relatively more simple. If I have more time, I would try several additional time series settings, and check whether the forecasting result is consistent across different models, so to ensure the stability.
4. Closeness of zip codes. To answer the second question of what is observed for zip codes which are close together, if there is more time, I would try several more clustering models and see whether physically close zip codes will fall into the same cluster or relatively far away clusters.
5. Forecasting for more years. If I can have more time and more data, I would forecast the index values for 5 more years instead of only 2023.

APPENDIX

You can find my code here:

https://github.com/nahuhs/Property-Investment-Modeling/blob/main/ShuhanZhang_workbook.ipynb

(The code probably takes too long to render, but you can download the file and change the suffix to .ipynb, and it should work fine)