Clustering of Redfin Houses – your next perfect house

Shanshan Liang, Coursera Learner

2018-11-04

# Abstract

House hunting process is complex and time-consuming. Traditional approach of hiring an agent heavily depends on the agent’s empirical experiences, also the customer may not be presented with complete list of houses for consideration. In order to save time in providing a complete picture for hunting the next house, the author uses the combination of web scraping of the available houses on the Redfin website and leverage the FourSquare API to collect surrounding venues. An unsupervised machine learning model – clustering is applied to cluster the houses. This approach presents all the available houses with their characteristics in clusters, which then help the customer make further research and hopefully next steps in house hunting quicker and easier.

# Introduction

Buying a house is a critical decision people make in various stages of lives. When people move into the new areas for jobs or family reasons, they settle down by embarking the house hunting journey. When life situation changes, they need a bigger house or a smaller house in the existing neighborhood that they’ve already fallen love with, they start the house hunting again. House hunting is important without doubt, yet a very complicated and resource intensive process. It is complicated due to people take various sets of factors in the decision making process – for example, someone may value the school district a lot, while as other people may not care much. There are many aspects of the houses people need to evaluate, with typically hundreds of houses on the market for research, all makes it a little intimidating in terms of the amount of the houses people need to go through before making a decision.

During this course, we’ve learned how to use FourSquare API to collect nearly venues data to cluster the neighborhoods. In this capstone project, I am leveraging this tool, together with the Redfin website to collect selected/custom information about houses for sale in my own zip code, as an example, to illustrate that house hunting process can be streamlined, customized and flexible. I hope this idea/process can help people gain more information about the targeted housing market, save time and increase the comfort level after fully analyzing the house market.

In this project, the stakeholder is myself – that is, the information that would be typically obtained from the external stakeholders is instead determined by myself. During this report, I also describe briefly the application of using this approach to satisfy a broader audience’s need in real life.

# Data

Redfin data is collected via web scraping in Python. First, we identify the geography of interest, in my case, is my current 5 digit zip codes 21043. In real life, it could be a new area someone will relocate to, it could be the zip code of the job site if short commute is important, it could be a broader area if someone wants to fully analyze the nearby areas. Second, we collect the information about each house in the search results. Web addresses of the houses for sale are saved in a list and we loop through this list to get information of each house. To use the FourSquare API, geo coordinates of the latitude and longitude of each house are also collected.

The surrounding venue data is collected via FourSquare API using the explore query. Depending on the density of the interested venue, one could adjust the radius when query the place so it returns the desired number for the analysis. After the venue data is collected for each house, in the result set, I filter the venue type based on my preferences – the places that I go to most frequently already or interested to spend time on.

# Methodology

**Feature Selection**

In this study, I only collect selected aspects of the houses rather than every available information from the Redfin website - based on what I value the most in buying a house. In real life situations, I expect the stakeholder would discuss their needs with the data scientists. Similarly, only selected venue types are chosen to feed into the model based on the fact that most people are interested in a limited number of venues. In real life application, we could let the customer define or choose the venue types that they are most interested in. The features that I am particularly interested are:

* total price
* square feet
* price per square feet
* school rating (elementary, middle and high schools)
* school distance (elementary, middle and high schools)

**Feature Scaling**

Redfin data and venue data is then merged together and to be prepared for fitting the clustering algorithm. The venue data is already standardized by taking the average of the frequency of each Venue Category. But Redfin data varies in number scales such as prices are very large numbers, and school rating is between 6 and 10 in this data set. Due to the variability of the value, I decide to perform feature scaling on most of the Redfin data including the price, square feet, price per square feet, school distance. I’ve intentionally left school ratings out during the feature scaling process. The reason is that I value the school ratings a lot more than the other features of the houses, that I do want school ratings to stand out in determining the clustering results.

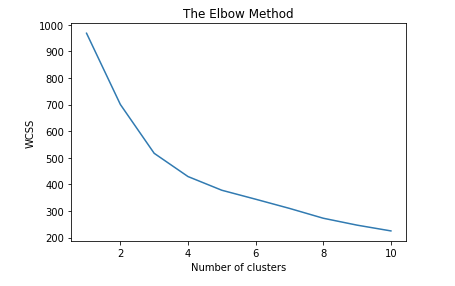
**K-Means Clustering**

Unsupervised machine learning model of clustering helps to group items together without the need of the dependent variable, or a label before running the algorithm. It has applications such as customer segmentation, identify groups to better understand the data. The feature set of the houses are important inputs in determining the cluster outputs. K-means clustering is used in this project. In order to determine the number of clusters, we will employ the “elbow method” and choose the number of cluster from the turning point.

**The Elbow Method**

Choosing the number of clustering is necessary before fitting the K-means clustering algorithm. Based on below chart that shows the WCSS value for number of clusters ranges from 1-10. The turning point is around 3 or 4. It is obvious that after 5 and slope is very smooth, indicating that the increased number of clusters does not add too much explanation power. But it is not evident 3 clusters or 4 clusters should be used.

**Figure 1. The Elbow Method from K-Means algorithm**

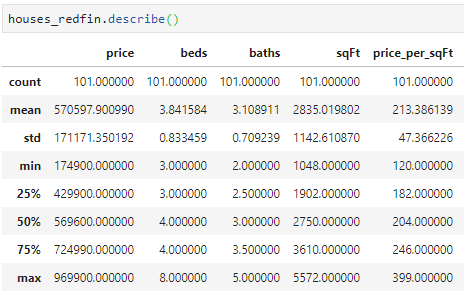


In this project, I’ve ran the algorithm separately using 3 clusters and 4 clusters. Based on the clustering results, I’ve found 4 clusters make a little more sense and thus presented the results below based on choosing 4 clusters.

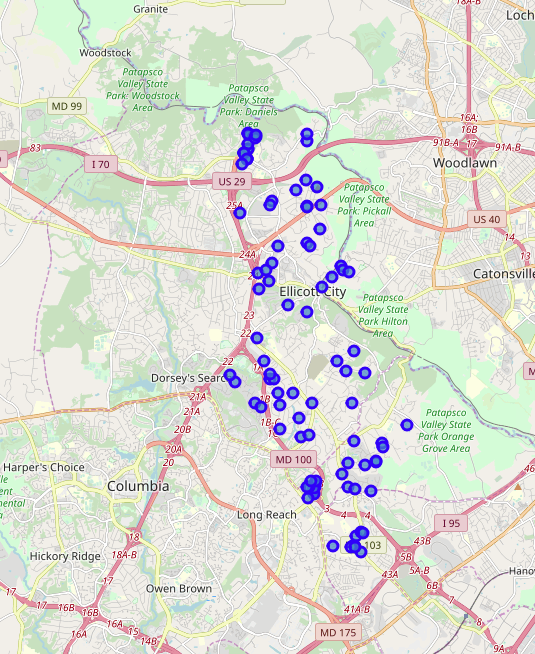
# Results

**Redfin Houses Data**

The Redfin houses data is collected via the web scraping as of October 24, 2018 by searching the Redfin website of the 5 digit zip code of 21043. A total of 129 houses are scraped and output to a JSON file. The stakeholder (me) has basic requirement as to the type of houses that she is interested, that is, houses with at least 3 bedrooms and 2 bathrooms. There is 101 houses that satisfy this basic requirement and thus form the pool of the clustering analysis. The data is pretty clean, with no missing data in any of the field in the Redfin data set.

**Table 1. Summary of Redfin houses**

**Figure 2. Houses imposed on Ellicott City Map**



**Venues Data**

For each of the houses, I collect the top 100 venue records within 3000 meters as of October 25, 2018. For the 101 houses, a total of 7400 venue data is collected.

**Table 2A. Summary of Venue Data**



In the course exercises, all the venue categories are fed into the clustering model. In this project, based on the venue categories from the 101 houses, I choose 16 of distinct categories based on my family’s preferences. The rationale behind it is that most people have their developed behavior pattern or life style, which then reflected in the data that some of the categories are highly utilized than others.

The venue data is then further prepared to calculate the average frequency of occurrence of each venue categories. Each house is one row with categories in columns.

**Table 2B. Sample houses of Average Occurrence of each Venue Categories**

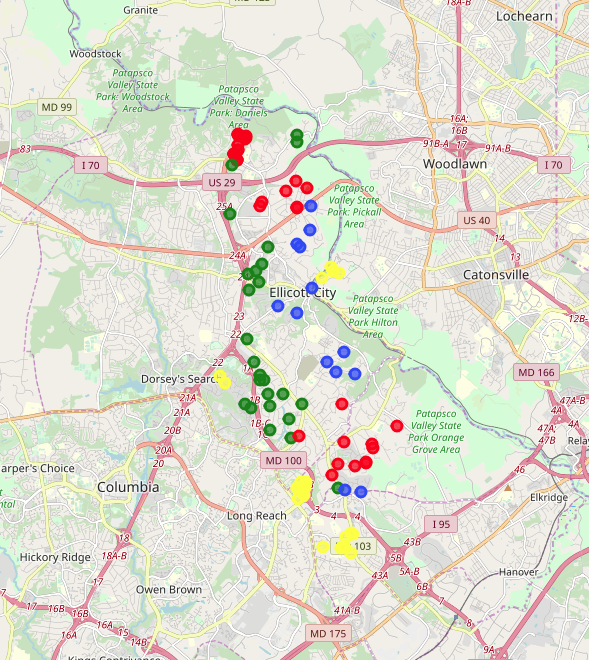
# 

**Clustering Results**

Four clustered are generated in the 101 houses based on 25 features – 9 of which are from the Redfin data and 16 of which are from the venue data. In Figure 3, the four clusters are colored Cluster 0 (yellow), Cluster 1 (blue), Cluster 2 (green) and Cluster 3 (red). In order to comprehend the results, Redfin data and venue data is merged again, but slightly different from how we prepare the data for the K-means clustering, in this case, we keep the original Redfin data and summarizes the venue data by top frequencies.

The complete list of houses with Clustering Labels is attached in the Appendix 1 in this report.

**Figure 3. Four Clusters by K-Means algorithm**



By analyzing the four clusters, I’ve summarized below table to show the distinct four groups of houses. One pattern that I notice is that school rating does play an important role in separating the houses, this is consistent with the fact that I have not performed feature scaling of the school ratings. For me personally, I have two small children who will be in school in a few years, so I am mostly interested in Cluster 1-3. Based on other factors such as budget, the relative location to highways (from the map), the actual school names and the surrounding venues, this clustering result makes my life much easier when hunting for my next perfect house. I will probably go through each house in the Cluster 1 to find out more details.

**Table 3. Characteristics and Potential Customers for each Cluster**

|  |  |  |
| --- | --- | --- |
| **Clusters** | **Characteristics** | **Customers** |
| Cluster 0 (yellow) | Relative low overall school ratings, most common venues are American Restaurant, Coffer Shop and Gym | Singles or Couple without kids |
| Cluster 1 (blue) | Highest overall school ratings, most common venues are Restaurants, Coffer Shop and Wine Shop | Family with kids |
| Cluster 2 (green) | Middle School Rating Score 7, most common venues are Restaurants, State Parks and Theaters | Family with kids with higher budget |
| Cluster 3 (red) | Some 7s in Elementary and Middle School Rating Score, relative smaller houses, most common venues are Restaurants, Coffee Shop and Playground | Family with kids with medium budget |

# Discussion

In this project, I’ve employed the Redfin website, FourSquare API and the K-Means algorithm to cluster the houses. This approach lays a good foundation to streamline the house hunting process if someone wants to DIY versus hiring an agent. However, it has limitations such as feature selection – is any of the Redfin house features may be excluded due to it is duplicative nature of other features, or cross-validation, if we use more sophisticated approaches to minimize the bias by cross-validation, would we possibly get different results. Part of it is due to the depth that this course covers, it is a very good starting point, and encourages students to continue their journey in the data sciences domain to get more involved in the methodology and mathematics theories supporting the machine learning algorithms. In addition, I am the sole stakeholder in this project, so the results are highly influenced by my preferences.

# Conclusion

In this Capstone project, I provided an innovative approach in applying K-means algorithms to cluster the Redfin houses. I’ve developed a framework to streamline the house hunting process by web scraping the houses from the Redfin website, querying the venues using FourSquare API, and K-Means algorithms for the houses clustering. This process is also flexible to incorporate stakeholders’ input/requirement at various stages such as the basic filtering of the houses of interests, selected venue categories for the clustering or all of them. In conclusion, the clustering results of the houses gives me tremendous insights in this neighborhood and equipped me well in finding my next house.

# Reference

Segmenting and Clustering Neighborhoods in New York City. <https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DP0701EN/DP0701EN-2-2-1-Foursquare-API-py-v1.0.ipynb>

**Appendix 1. Clustering Results** 