IBM - Coursera
Data Science Professional Certificate

Final Capstone Project Report

"The best home in Minneapolis for me"

Tenzin Kunsang
September 2020

Introduction

This report is a part of Coursera's Data Science Professional Certificate provided by IBM. The certificate includes 9 courses such as Data Science Methodology, Databases and SQL, Data Analysis, Data Visualization, and Machine Learning. The requirement for the final report is to use FourSquare API to explore or compare neighborhoods or cities of our choice. The learner can decide what problem they will focus on and what methodologies will be used to solve that problem.

Having just graduated from Carleton College, which is about 40 minutes away from Minneapolis, I have considered living in the city. For this project, I was inspired to delve deeper into home values to look for a place in Minneapolis. The city, along with St. Paul, makes up the 'Twin Cities.' Located in the Midwest of the United States, it is the second most densely populated city in the region behind Chicago.[1] The median rental cost is $985 and the median house value is $235,900.[2]

For this project, I will mainly be using the K-means clustering method to evaluate the home values in Minneapolis based on neighborhood and homes. Real estate values are determined by many factors and different buyers have different priorities. Some parameters that many people consider when buying a home are location, home size, usable space, upgrades, local market, and neighborhood comps.[3]

This report will present three main factors:

1. Neighborhood Comps - homes located in the neighborhood with similar size, condition, and features
2. Home size - living area (sqft)
3. Location - venues in the vicinity of a home

We will analyze the neighborhood and home clusters based on these parameters.

---

[1] https://en.wikipedia.org/wiki/Minneapolis
[2] https://worldpopulationreview.com/us-cities/minneapolis-mn-population
[3] https://www.opendoor.com/w/blog/factors-that-influence-home-value

The target audiences for this report are

- Buyers looking for a new home in Minneapolis
- Property Investors
- Realtors and agents
- The curious ones

Data Description

Before I extract any data, I wanted to get the best and most updated list of neighborhoods in Minneapolis. Therefore, I scraped the list from wikipedia page: https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Minneapolis

Current average home values in each neighborhood is important to consider when looking at homes. For the home value index (HVI) in Minneapolis, the scraped dataset from Zillow didn't contain enough information. Therefore I manually downloaded it from the following page: https://www.zillow.com/minneapolis-mn/home-values/

Sales prices of similar homes in neighborhoods that have been sold recently gives a good estimation of what price range the home of your choice might land in. Some other important features when buying a home are the home size, age and condition, and neighborhood comps. I scraped these data from Zillow using a platform called Apify. Unfortunately this API didn't readily return neighborhood names so I manually searched for homes in each neighborhood in Minneapolis separately on the platform.
https://apify.com/

Many families consider the quality of local schools, employment opportunities, proximity to shopping etc. before buying a home. I used Foursquare API to get a sense of the neighborhood locations by looking at venues closeby. Moreover, I used this API to get the nearby venues for

each home to get a better idea of how homes in the same neighborhood differ based on the distance from nearby venues.

Methodology

Scraped through the wikipedia's page containing names of neighborhoods in Minneapolis. Used FourSquare API to get the latitude and longitude of all the neighborhoods. Cleaned up the resulting dataset. Attached the zestimate values of these neighborhoods by getting the home value index by neighborhood dataset from Zillow and appending the new data onto the original dataset by the name of the neighborhood. (resulting dataset picture) Used FourSquare API to get the venues within a radius of 500m from the center of each neighborhood. This, in turn, returned venues along with the category that these venues belonged in. (resulting dataset picture) One-hot encoded the unique venue categories in columns.. (mention examples of venue categories). In addition to the columns of venues categories, I found the top 10 categories for each neighborhood and these categories are thus the entries for the 1st Most Common column, 2nd Most Common column, and so on. The final dataset (picture). Based on features including the top common columns, venue categories, home value index, MoM, QoQ, YoY, five and ten years annualized, the best k-value found using the elbow method is 3. (picture of elbow method). Finally implemented unsupervised machine learning and clustered the neighborhoods into 3 groups.

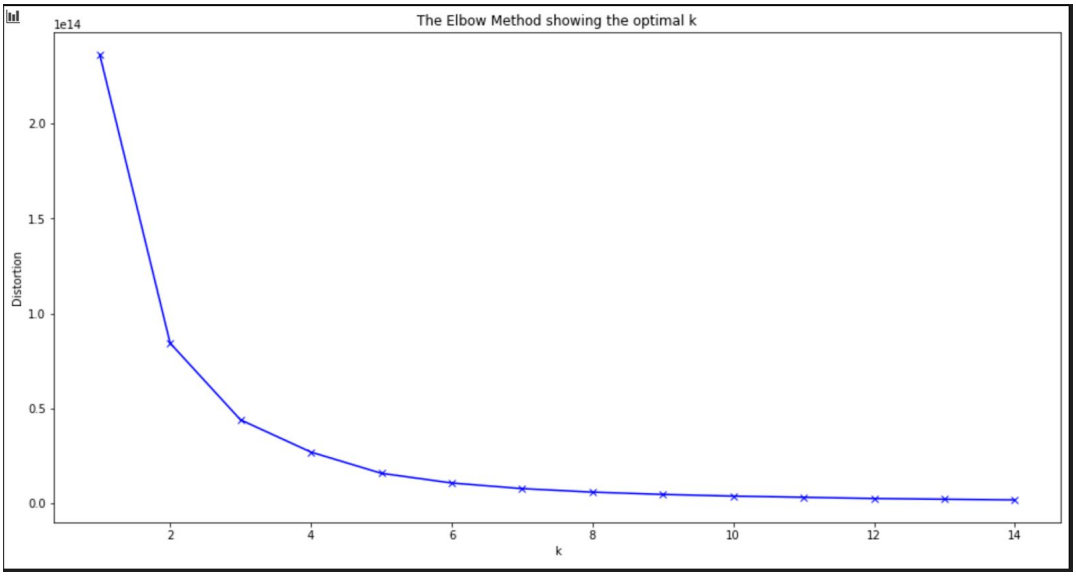
Used Apify to search for homes in each neighborhood. Resulting features: zip code, bathrooms, bedrooms, year built, description, sqft, and sale price. Regression analysis to see how number of bathrooms and bedrooms, year built, and sqft correlate to the sale price. Cluster of these homes. Merge this dataset with the dataset from the previous part and based on location, neighborhood comps, and home size and usable space, see how the final cluster changes.

Results

After scraping the homes of each neighborhood in Minneapolis and cleaning up the dataset, there were a total of 811 homes from 63 unique neighborhoods. The main parameters in the resulting table include

- Number of bedrooms
- Number of bathrooms
- Living area in square feet
- Home sale price
- Year built
- Description of the home
- Neighborhood name

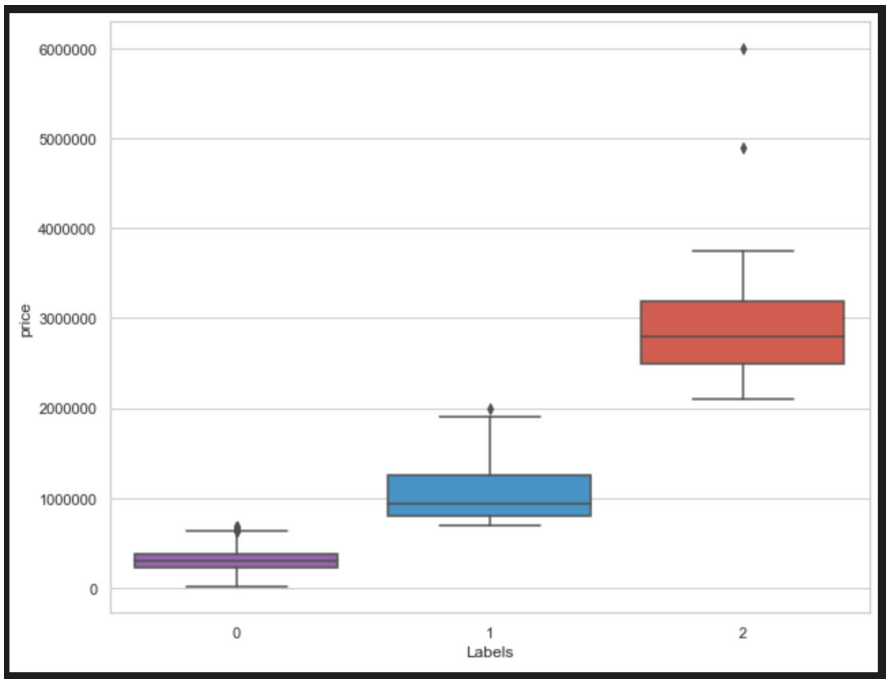
Using the FourSquare API to search for homes near each home and cleaning up the resulting dataset, we received a total of more than 800 venues with 320 unique venue categories. These venue categories include school, museum, bar, restaurant, shopping mall, park, and many more. Thus, each home has columns of home information such as bedrooms and sale price and number of venues of all the venue categories. Based on these parameters, the elbow method displayed that the homes in Minneapolis can be divided into 3 clusters.



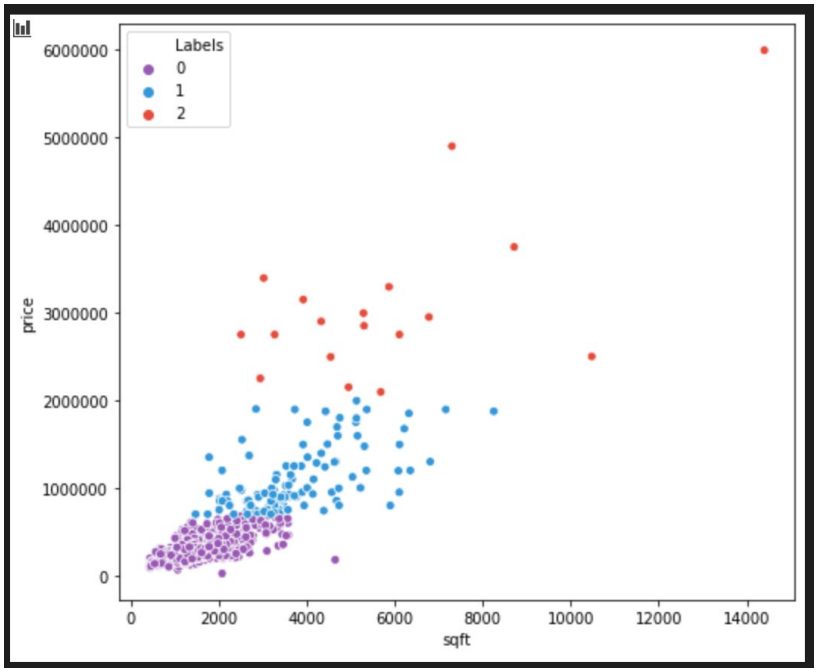
As shown in the table below, the first cluster of homes is comparatively cheaper (average price: $316,900) and has much less living area (average: 1611 sqft). On the other hand, the second cluster of homes were built much more recently. They are much higher priced than the first group of homes, with an average living area of 3631 sq ft and average home sale price of $1,070,467. Finally, the third cluster has much more living area (average: 3631 sqft) and has an average home selling price of $3,033,235. Moreover, the average home price of all the clusters are higher than the median home price of Minneapolis ($235,900), thus resulting in a right skewed graph for the limited set of homes in our dataset.

Cluster	Bathrooms	Bedrooms	Living Area (sqft)	Sale Price ($)	Year Built
0	1.87	3.02	1611.69	316,900	1935.66
1	3.86	4.07	3631.59	1,070,467	1950.01
2	4.95	4.45	5647.45	3,033,235	1951.25

Following figure displays how the home sale price differs for different cluster groups. It is important to note that the results are not standardized. There are 659, 129, and 20 homes in each cluster, respectively.



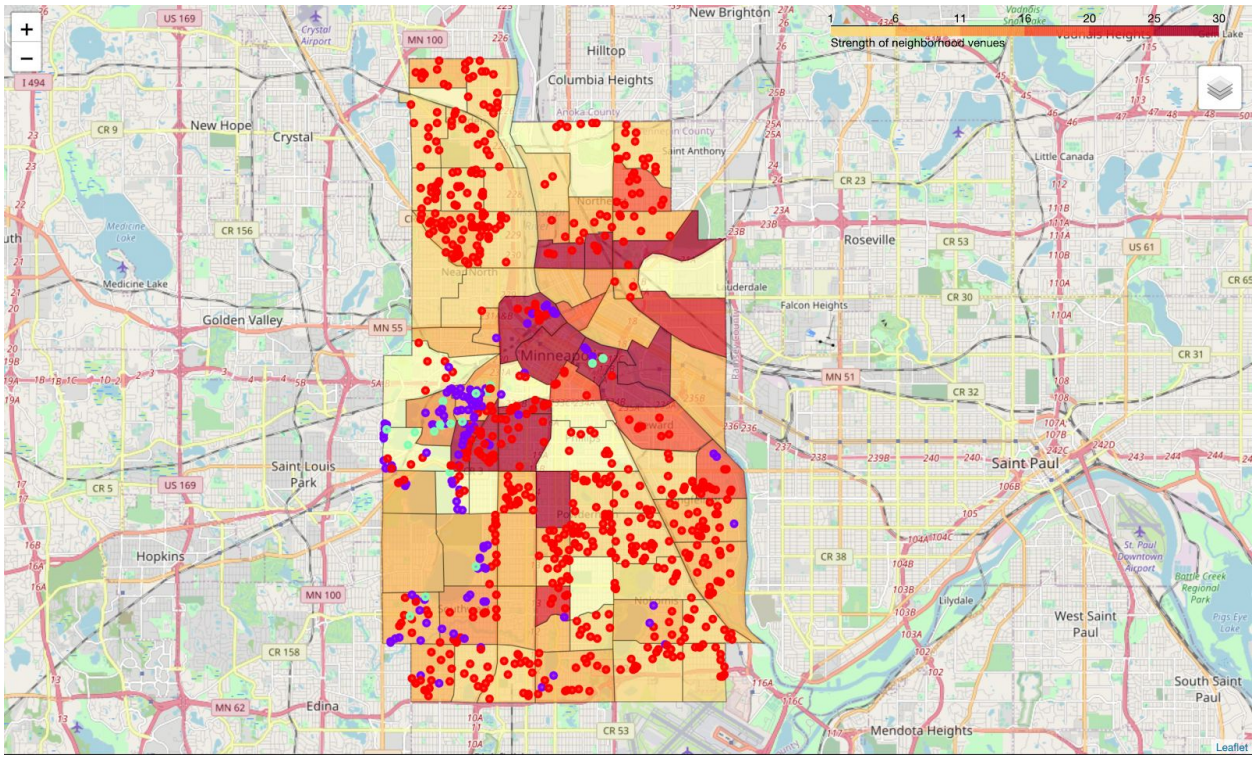
Additionally, we found that the living area has a strong, positive relationship with the price of homes (correlation: 0.86). Picture below shows a scatterplot of how the living area's size compare with the home price, colored by cluster groups.



Our result also shows that while number of bedrooms of a home in Minneapolis has a weak, positive correlation with price (0.37), the number of bathrooms has a strong and positive correlation with price (0.72).

The map below displays all the homes in our dataset, where the popups are colored by cluster groups. Each neighborhood is further outlined on the map and visualized using choropleth in Python's Folium library -- the higher the number of venues found in a neighborhood, the

darker the shade.



Discussion

Need more data. No homes in some of the neighborhoods. Standardizing might give a better idea. Would help to have more than just the basic info about homes; commute and walk scores in Zillow would be great data to have. Word cloud of home descriptions would be pretty great too.

Conclusion

We found that in the city of Minneapolis, there are three main clusters of homes based on location venues, home sale price, living area, basic home details such as number of bedrooms and bathrooms. Generally, homes in the western region of Minneapolis are more expensive because almost all the homes belonging in the second and third cluster groups lie there. Without any surprise, there are far more venues in downtown Minneapolis than anywhere else in the city.

Works Cited

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