IBM - Coursera
Data Science Professional Certificate

Final Capstone Project Report

"The best home in Minneapolis for me"

Tenzin Kunsang
August 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 the city of Minneapolis in the United States, I have considered living in the city. For this project, I was inspired to delve into home values and rental costs 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 be using methods such as K-means and linear regression to create predictive analytics of home values in Minneapolis. Real estate values are determined by many factors and different buyers have different priorities. Some factors 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
2. Home size and usable space
3. Location

I will cluster neighborhoods based on location and find the best model by regressing home size, usable space, and neighborhood comps to get a sense of what home value the buyer might be landing on to get a home of their choice or vice versa.

---

[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 values index in Minneapolis, the scraped dataset from Zillow didn't contain enough information. Therefore I manually downloaded it from this 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 apify. Unfortunately this API didn't readily return neighborhood names so I looked for homes in each neighborhood by searching for homes at every 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.

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

From K-means elbow method, there were three clusters.
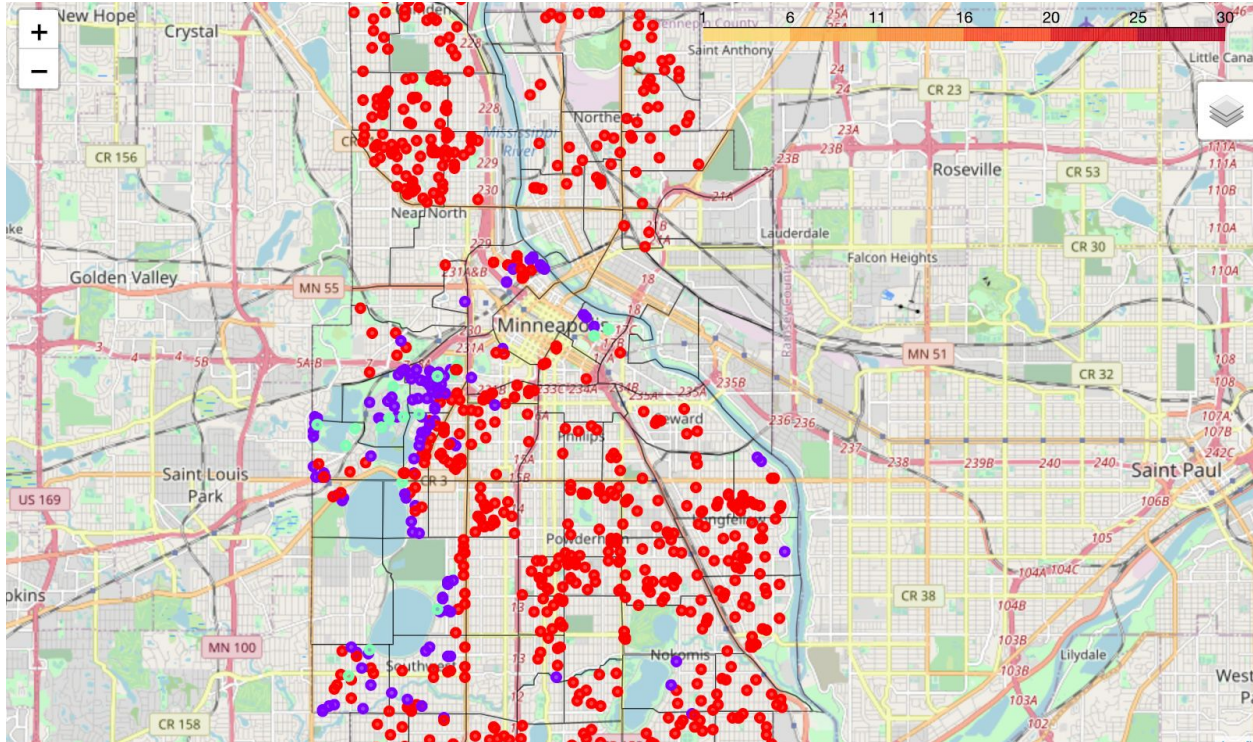
Summary of the three clusters show that cluster 0 is comparatively cheaper (average price: 316900) and has much less living area (mean 1611 sqft). It also has on average less bathrooms and bedrooms than the later two clusters.

Labels	bathrooms	bedrooms	sqft	price	yearBuilt
0	1.867982	3.021244	1611.693475	3.169008e+05	1935.657056
1	3.875969	4.069767	3631.589147	1.070467e+06	1950.007752
2	4.950000	4.450000	5647.450000	3.033235e+06	1951.250000

Visual depiction of how home sale price varies in different clusters.
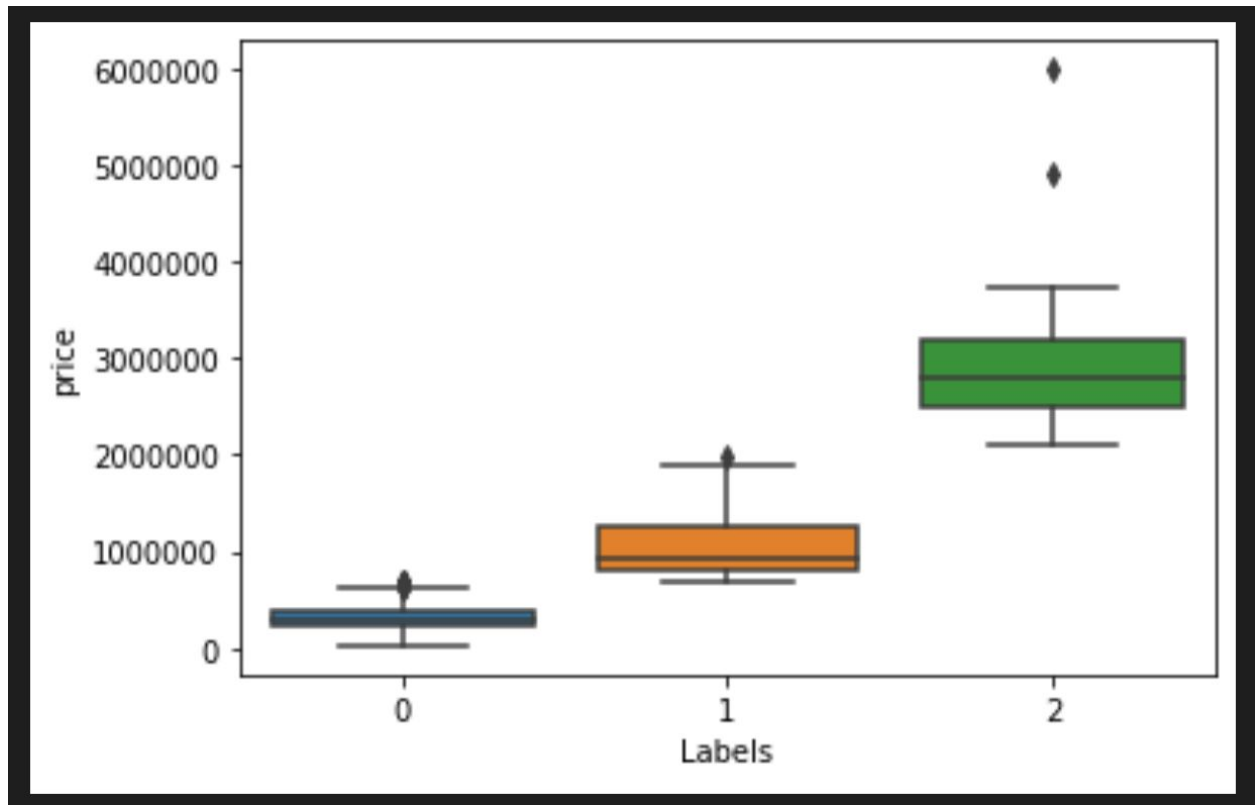
Living area positively and strongly related to price, with a correlation of 0.86.
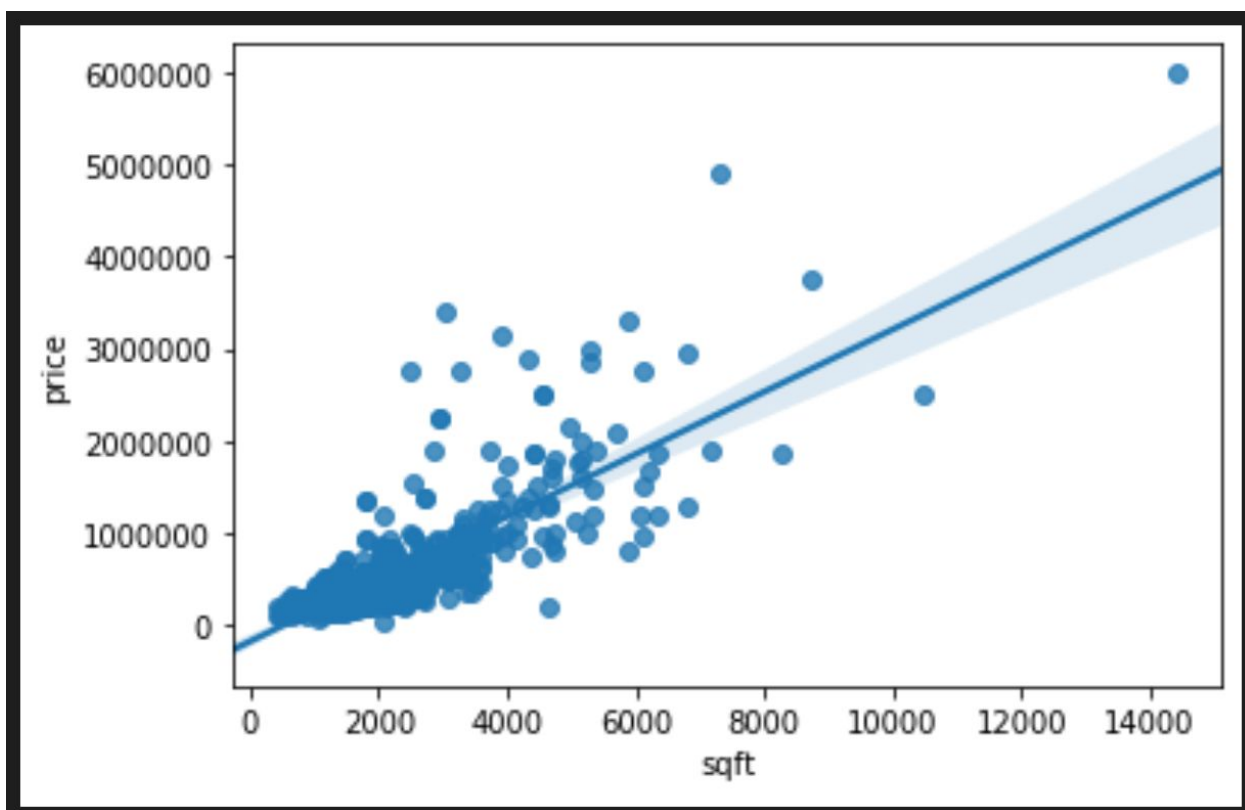
http://sls.gatech.edu/sites/default/files/documents/Toolkit-Docs/hong_fieldguide_zillow.pdf