Minneapolis-St. Paul Residential Real Estate Prices and Venues

Mark Hanson

May, 2019

1. Introduction

Minneapolis - St. Paul's (MSP) and the surrounding metropolitan area is home to numerous public and private companies including several on the Fortune 500 list. Some notable companies with headquarters or major operations in the area are United Health Group, Target, Best Buy, US Bank, Well Fargo, Ameriprise Financial, Thrivent Financial, Securian Financial, Supervalu, General Mills, CHS, Cargill, Hormel, Land O Lakes, C.H. Robinson, Xcel Energy, 3M, Medtronic, Honeywell, Ecolab, and Patterson. With such a diverse set of large companies in the MSP area there are bound to people that relocate to the area for work but that are unfamiliar with the different cities and surrounding suburbs that comprise the MSP metro area.

With over 140 zip codes in the seven county metro, there are a lot of real estate markets to explore. Realtors can help. However, while they may be familiar with entire metro area in general, they usually specialize in specific cities and suburbs (i.e. south west metro).

Many factors will come into play when deciding where to start looking for a home to buy. Location and price range are usually the chief considerations when deciding where to start looking. But other factors such what types of venues are in the area are also important to many prospective buyers.

1.1. Business Problem

In an effort to bring some rigor and data science to the problem of characterizing the various suburbs within the MSP metro area, an analysis is presented that clusters the various zip codes that comprise the seven county metro by median home price range. To add some additional detail, the most prevalent venues in each zip code are added to the median home value cluster data set.

1.2. Interested Audiences

Realtors, prospective buyers, current home owners, residential or commercial developers or anyone seeking to better understand the residential real estate market in the Minneapolis – St. Paul seven county metro will be interested in this data analysis and the results.

2. Data

2.1. Data Sources

To perform this analysis, current and historical residential real estate value data was acquired, analyzed, and joined with both venue data and location data. Current and historical real estate value data were acquired from Zillow[1] - an online real estate database company that has data of median home values by zip code available to download for free. Listings of which zip codes are in the 7 counties - Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington - that make up the MSP metro area were

[1] Home value data attribution: "Data acquired from Zillow.com/data on April 10, 2019. Aggregated data on this page is made freely available by Zillow for non-commercial use." - https://www.zillow.com/research/data/

obtained from Capital Impact[2] - a site that has a lot of basic civic information about places in the United States. The source for the latitude and longitude of each of these zip codes is CivicSpace Labs[3]. GeoJSON data that specifies that polygon shapes for these zip codes originates as U.S. Census[4] data but was obtained from OpenDataDE[5] in a GitHub posting. Finally, venue data was provided by Foursquare[6].

Zillow's median home value data for zip codes in Minnesota and Wisconsin were downloaded and imported into a Juypter notebook as a pandas dataframe from a comma separated values (CSV) file. This was done for expediency in this singular analysis but an application programming interface (API) is also an option for accessing the data as part of an application. Ultimately this data was filtered down to just data from the seven counties that make up the MSP metro area.

2.2. Cleaning Data

Since the Zillow data is parsed down to the zip code level as opposed to just the city level – cities can contain multiple zip codes – the Python geolocator library didn't provide fine enough latitude and longitude resolution for use in characterizing locations with Foursquare data. To enable searching each zip code in Foursquare, another pandas dataframe was created with data obtained from CivicSpace Labs containing the zip code and latitude and longitude of a centralized point in each zip code in the MSP seven county metro area. The two dataframes were then merged by matching zip codes resulting in a single dataframe with geo-coordinates for each zip code, as well as city, state, county, and current and historical median home value data.

3. Methodology

3.1. Exploratory Data Analysis

At this point in the analysis the latitude and longitude for each zip code were used to obtain venue data for each zip code with the intent of merging it with the median home value data and using a k-means algorithm to cluster the zip codes based on the combined dataset. K-means clustering machine learning algorithm was chosen because the intent with this data is to find patterns and summarize the unlabeled yet inherently similar data.

Two issues arose while pursuing this approach. First, some zip codes were identified with no venues at all within the search radius. Expanding the search radius to 1 mile corrected this problem which is likely due to both rural zip codes having lower venue density and Foursquare venue data being less complete than say Google's dataset owing to a difference in the number of active users of the respective services. Second, using a k-means algorithm with a normalized feature set to cluster zip codes on the combined dataset resulted in homogeneity between clusters despite efforts to weight the clustering towards median home value as a differentiating factor.

3.2. Modeling

While still attempting to cluster the combined home value / venues data set, various numbers of clusters were evaluated using the k_inertia metric. Graphing k_inertia versus the number of clusters should allow

[2] Zip code by county source attribution: http://ciclt.net/sn/clt/capitolimpact/gw_default.aspx

[3] Zip code centroid data attribution: Copyright 2004 CivicSpace Labs, License: Creative Commons Attribution-ShareAlike, Dataset Identifier: us-zip-code-latitude-and-longitude, Modified: February 9, 2018 10:31 AM, Reference: https://boutell.com/zipcodes/

[4] Zip code GeoJSON attribution primary data attribution: https://www.census.gov/cgi-bin/geo/shapefiles2010/layers.cgi

[5] Zip code GeoJSON attribution final data attribution: https://github.com/OpenDataDE/State-zip-code-GeoJSON/blob/master/README.md

[6] Foursquare developer API attribution: https://foursquare.com/developers/apps

selection of an appropriate value for the number of clusters as the region corresponding to the "knee" or inflection point in the graph where further reductions in k_inertia reaches the point of diminishing returns for further increases in the number of clusters. Using the combined median home value / venues dataset, no distinct inflection point was observed.

Since clustering on the combined home value / venues dataset did not produce useful results and recalling that price range is the primary factor of interest in the residential real estate market, a shift in approach was made to cluster the data on median home value itself and then merge other data with the clustering result set. With this approach, k_inertia versus number of clusters had more of an inflection point. Based on the inflection point the number of clusters to be formed from the k-means algorithm was set to 10.

Choosing 10 clusters produces distinct price ranges, Figure 1, as the boxplots of price versus cluster show. Six of the 10 clusters have between 14 and 24 members (zip codes) the remaining four clusters have 8, 6, 3, and 1 member with the single member cluster clearly being an outlier that belongs in a class by itself.

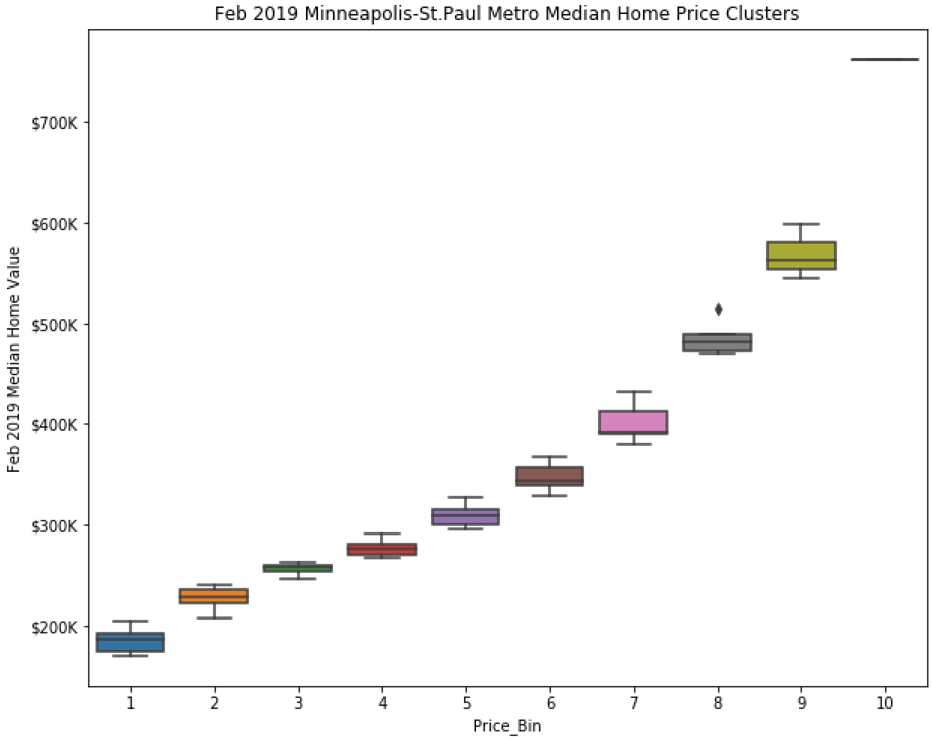
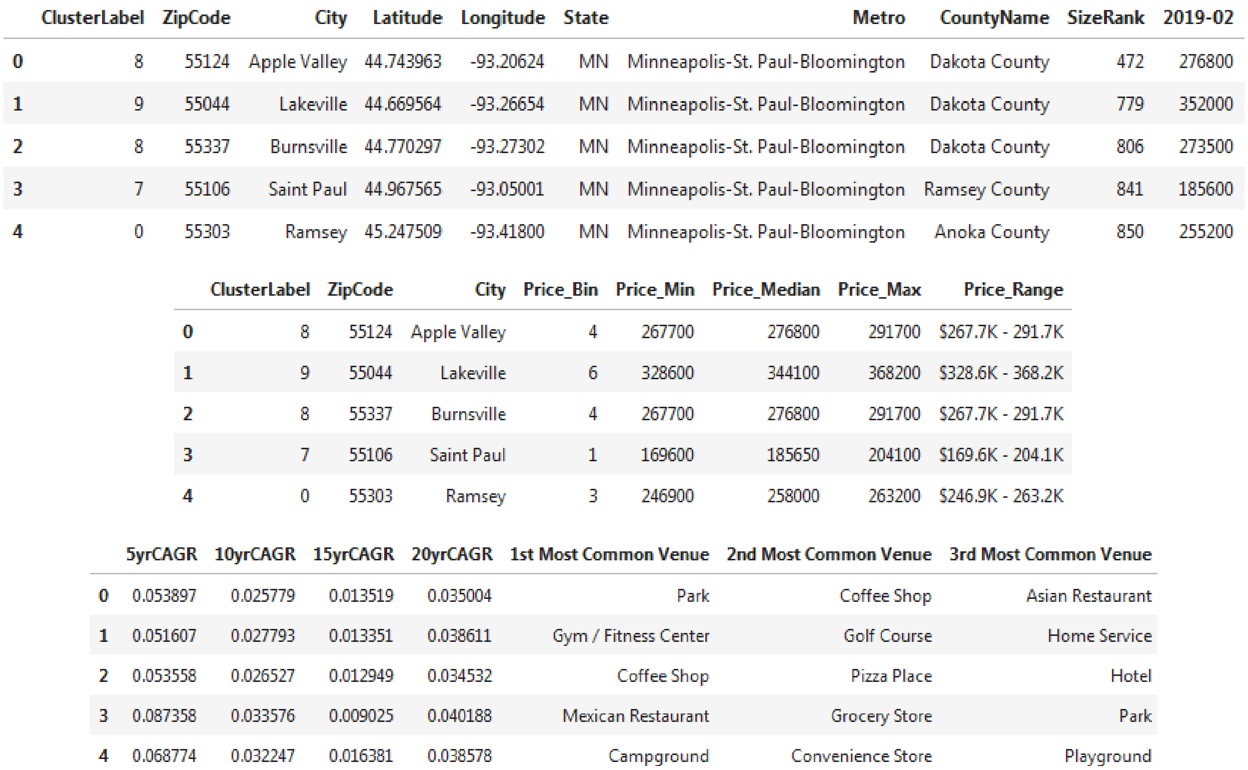


Figure 1: Median home price cluster box plots arranged in ascending order by assigning price bins corresponding to cluster labels

4. Results

With clustering by median home value complete, additional analysis was done and a summary dataframe created to store the results. Additional analysis included calculating for each cluster a price bin, minimum, median, maximum, and range. Compound annual growth rate (CAGR) over 5, 10, 15, and 20 year periods were also calculated and added to the summary. Finally, the top 10 venues within one mile of each zip code approximate geographic center were obtained from Foursquare and added to the summary. Snippets of the summary dataframe are shown, Figure 2. The entire Jupyter notebook is posted in GitHub. More details instructions on how to access it are provided later in the report.



	ClusterLabel	ZipCode	City	Latitude	Longitude	State	Metro	CountyName	SizeRank	2019-02
0	8	55124	Apple Valley	44.743963	-93.20624	MN	Minneapolis-St. Paul-Bloomington	Dakota County	472	276800
1	9	55044	Lakeville	44.669564	-93.26654	MN	Minneapolis-St. Paul-Bloomington	Dakota County	779	352000
2	8	55337	Burnsville	44.770297	-93.27302	MN	Minneapolis-St. Paul-Bloomington	Dakota County	806	273500
3	7	55106	Saint Paul	44.967565	-93.05001	MN	Minneapolis-St. Paul-Bloomington	Ramsey County	841	185600
4	0	55303	Ramsey	45.247509	-93.41800	MN	Minneapolis-St. Paul-Bloomington	Anoka County	850	255200

	ClusterLabel	ZipCode	City	Price_Bin	Price_Min	Price_Median	Price_Max	Price_Range
0	8	55124	Apple Valley	4	267700	276800	291700	$267.7K - 291.7K
1	9	55044	Lakeville	6	328600	344100	368200	$328.6K - 368.2K
2	8	55337	Burnsville	4	267700	276800	291700	$267.7K - 291.7K
3	7	55106	Saint Paul	1	169600	185650	204100	$169.6K - 204.1K
4	0	55303	Ramsey	3	246900	258000	263200	$246.9K - 263.2K

	5yrCAGR	10yrCAGR	15yrCAGR	20yrCAGR	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	0.053897	0.025779	0.013519	0.035004	Park	Coffee Shop	Asian Restaurant
1	0.051607	0.027793	0.013351	0.038611	Gym / Fitness Center	Golf Course	Home Service
2	0.053558	0.026527	0.012949	0.034532	Coffee Shop	Pizza Place	Hotel
3	0.087358	0.033576	0.009025	0.040188	Mexican Restaurant	Grocery Store	Park
4	0.068774	0.032247	0.016381	0.038578	Campground	Convenience Store	Playground

Figure 2: Snippets of pandas dataframe summarizing median home value by zip code dataset

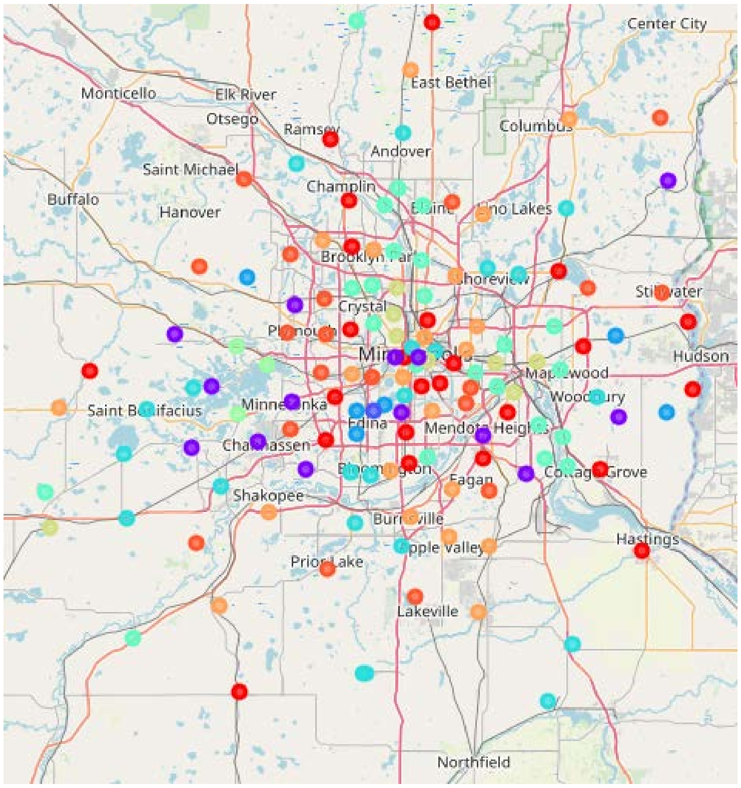
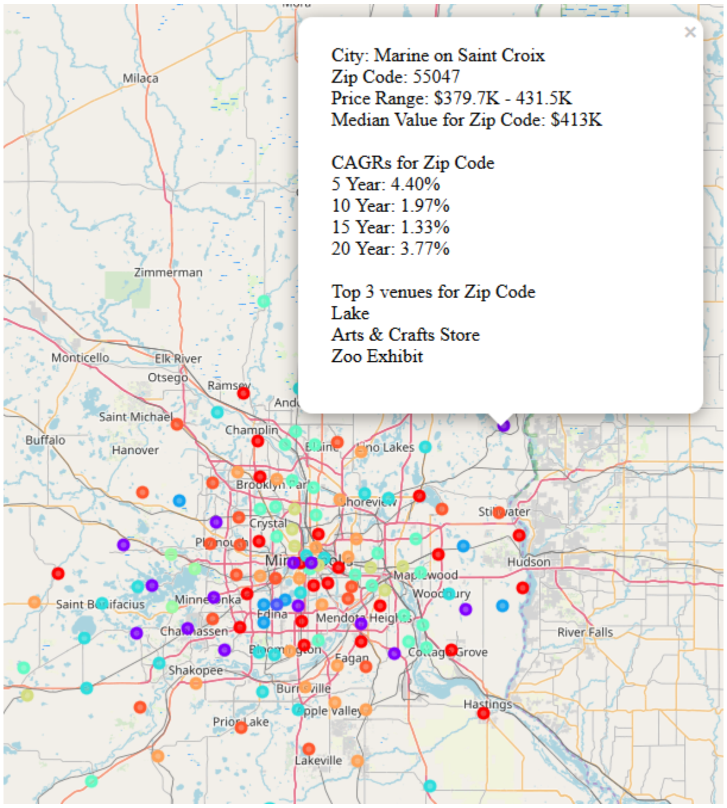


Figure 3: Cluster map with color coded circle markers

Price bins were defined to put the clusters in ascending order of price range. Each price bin was further summarized to again display the price range and median home price of the zip codes that are members of the cluster. In addition, the top 10 venues for each cluster as a whole were calculated and displayed.

"Location, location, location" is the saying in real estate – meaning of course that in addition to price, location is an extremely important factor guiding transactions and interest. With all the analysis summarized in pandas dataframes, visualizations were created to help users better understand the data. As a starting point, a cluster map, Figure 3, with a circle marker showing which price bin it falls into was created using the folium library.



City: Marine on Saint Croix
Zip Code: 55047
Price Range: $379.7K - 431.5K
Median Value for Zip Code: $413K

CAGRs for Zip Code
5 Year: 4.40%
10 Year: 1.97%
15 Year: 1.33%
20 Year: 3.77%

Top 3 venues for Zip Code
Lake
Arts & Crafts Store
Zoo Exhibit

Figure 4: Cluster map with pop-up dialog boxes

While a cluster map helps to show where in the MSP seven county metro area zip codes with similar median home values are located, it isn't very easy to get that information from quickly. Adding pop-up boxes, Figure 4, for Juypter notebook users to click that list the city, zip code, price range, median home value, compound annual growth rate history, and top 3 venues makes understanding the map easier.

But it's still not as clear as if the data was shown in a choropleth map. As a next step, a choropleth map was created that keyed on the price bin for the zip code Figure 5. Of course choropleth maps require geoJSON data which as noted above was obtained from a GitHub posting by OpenDataDE; the raw underlying data for it coming from the U.S. Census Bureau. In the GitHub posting geoJSON data was posted by state. Since only geoJSON data for the zip codes in the MSP seven county metro were required, python code was used to loop through the larger geoJSON file for all of Minnesota and write a smaller geoJSON file for just the seven counties of interest.

One issue of note that arose while creating the choropleth map was that the zip code data in the geoJSON was stored as string datatype while that in the dataframe was stored as int datatype. Converting one to match the other datatype eliminated the error code from the datatype mismatch.

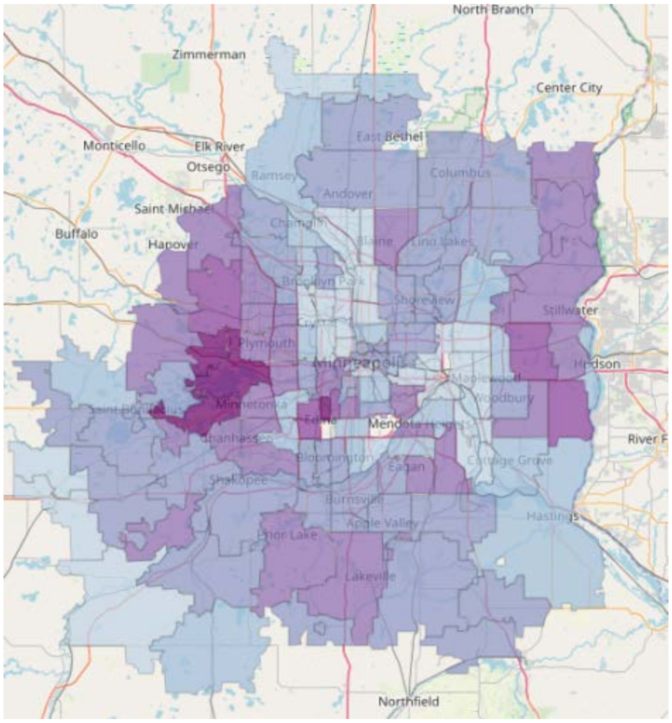
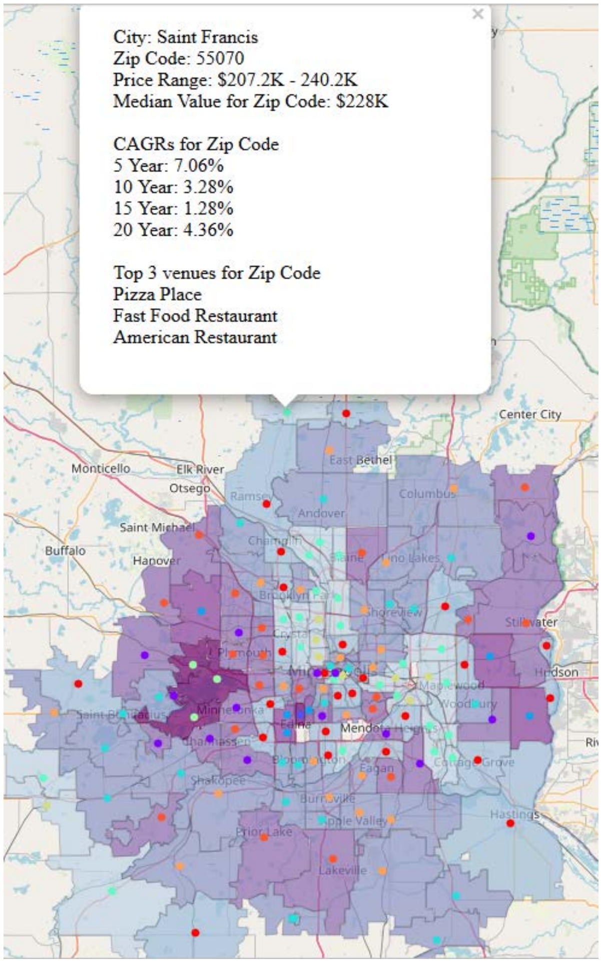


Figure 5: Choropleth map of median home value by zip code for MSP seven county metro area



City: Saint Francis
Zip Code: 55070
Price Range: $207.2K - 240.2K
Median Value for Zip Code: $228K

CAGRs for Zip Code
5 Year: 7.06%
10 Year: 3.28%
15 Year: 1.28%
20 Year: 4.36%

Top 3 venues for Zip Code
Pizza Place
Fast Food Restaurant
American Restaurant

Figure 6: Combined choropleth cluster map with pop-ups

Choropleth maps make it much easier to see which zip codes fall into which price bins and where they fall geographically. But this choropleth map lacks the circle marker pop-ups that allow Jupyter notebook users to click on areas of interest to drill down and get more information. To provide the best of both map types, they were combined into a single map with all the visualizations, Figure 6.

5. Discussion

Zip codes with higher median home values tend to be on the western and southern side of the Minneapolis – St. Paul metro area. Some possible explanations for this are that these areas are closer to Lake Minnetonka which is valued as a large and beautiful area lake, that there is a well-developed freeway connecting these areas to downtown Minneapolis, and that many of the largest employers are located either in the southwest metro or in downtown Minneapolis. Minneapolis is west of St. Paul and has a larger business district. St. Paul is to the east, is the state capital, and as such has more government jobs. There are, however, a few large employers with operations in the east metro including 3M, Ecolab, and Patterson.

Although more of the zip codes in the southwest metro have higher median value homes, there is also a fairly broad range of median home prices within a commutable distance to virtually anywhere in the seven county MSP area. In addition, as noted earlier there are similar venues in many of the zip codes in the MSP metro area. So finding a home in the right price range close to work with the right attractions in the area isn't usually a problem.

It is noteworthy that there is a substantial variation in the size of the various zip codes with the more urban zip codes being smaller and the more rural zip codes larger. This size of the search radius used in Foursquare was a mile, which was chosen to be small enough so as not to have significant overlap between zip codes yet large enough to characterize the venues in the zip code. For some of the very large zip codes the one mile search radius may not adequately characterize the venues encompassed within it. The location of the zip code centroid could have a big impact on the venues that are found as well.

6. Conclusion

The combined choropleth circle marker map with pop-up dialog boxes provides an efficient way to explore the residential real estate market in the Minneapolis – St. Paul metro area and get a high level assessment of current median home prices, historical growth rates, as well as venues that can be found in various zip codes.

Background on this analysis

This analysis was completed as a capstone project for Coursera's IBM Data Science Professional certification and using a Jupyter notebook. More information on Jupyter notebooks can be found at the following URL: https://jupyter.org/

Accessing GitHub posting

GitHub is a version control system used primarily by application developers. The Jupyter notebook that was used for this analysis can be downloaded from GitHub, along with the supporting JSON and CSV files at the following URL: https://github.com/chiefs18/Coursera_Capstone

Viewing and interacting with the Jupyter Notebook

Clicking on the circle markers to get pop-up boxes with more information is only possible when interacting with the Jupyter notebook in an online viewer after having executed the relevant code blocks up to the point where the map is created.

The Jupyter notebook can viewed with an online notebook viewer at the following URL: https://nbviewer.jupyter.org/github/chiefs18/Coursera_Capstone/blob/master/MH_Data_Sci_Capstone_JupyterNB.ipynb