Using Housing Market Appreciation and Foursquare API to Determine the Best Locations for Future Business Development in Omaha, Nebraska

Charles Gambrel

A. Introduction

A.1 Background

The popular Chicago based pizza restaurant *Giordano's* hosted a Facebook competition in late 2017 to determine the newest location of their expanding franchise. Omaha, Nebraska won out of the several cities that participated, and the locals are expecting to try the new restaurant soon. However, simply determining the city to build a business based entirely off of the attention garnered from a Facebook vote is not enough to maximize the profitability of a new location. Every moderately sized city on the planet is comprised of several sections, whether these sections are organized by quadrant, zip code, borough, or neighborhood. And these different sections of a city can vary widely economically. Some sections may be thriving with a healthy housing market and an economically diverse business profile, whereas others may have stagnated with business owners and realtors alike struggling for sales. Therefore, it would be advantageous for property investors to be able to organize the city of question into zones and grade them by a metric to determine the best location for success.

A.2 Problem

For the purposes of this project, I will examine different sections of Omaha, Nebraska, organized by zip code. Real estate data regarding the sales prices of single-family homes will be utilized to construct the metric that will be used for grading. A good indicator of economic growth in a city is the sales price of a home. Since housing prices have a large variance, the price of a home per square foot will be used over the total sale price and this figure will be measured over a period of time to determine the appreciation rate. The most popular venues for the most desirable zip codes will then be found and this data will be used to cluster the zip codes into groups based on venue type.

A.3 Interest

This project will be of interest to three major groups. The central goal of this exercise is to provide business owners with data regarding the economic health of each region in a city to help them in determining the ideal area of a new location. However, data on the appreciation rates of

single family homes and nearby venue information by zip code can be of great use to realtors and homes buyers alike when making a big sale or a big purchase respectively.

B. Data acquisition and cleaning

B.1 Data acquisition

The majority of popular real estate sales websites are working toward the collection of market data and will often provide a great deal of this data for free. This project required a dataset of single family homes with details regarding the final sales price, the square footage, zip code, and the date of purchase. Because of either an overwhelming amount of data or a just recently implemented collection effort, the most popular real estate data sources (Zillow, Redfin, ect.) will only provide data ranging as far back as 2017. This will be acceptable as I am working to calculate recent market trends and since the Omaha housing market has remained largely steady since the recovery from the national housing crisis of 2008, the recent housing data will paint a reasonably accurate picture.

Datasets from Redfin were ultimately chosen as they already included the calculation for price per square foot and the site provided filters to narrow the data attained based on house size. Another dataset provided by Opendatasoft was used to attain city of Omaha zip codes and their corresponding geospatial coordinates. Finally, Foursquare API will be harnessed to fetch the most popular nearby venues for each of the highly appreciating zip codes.

B.2 Data cleaning

The free downloadable datasets provided by Redfin are limited in size to 350 houses. Redfin does this to prevent third parties bulk downloading the data for commercial use. As a result of the 350 data point limit and the need for an entire city examination, multiple datasets were downloaded pertaining to different areas of Omaha. All of the data was filtered to only include single family homes that sold from 2017 to present day. Perhaps the most efficient approach to load this data into Jupyter Notebooks would have been to first combine these datasets in Microsoft Excel to form a single, large collection of data. However, for the sake of the project, I decided to show all of my work and use the python package pandas for all processing.

Each dataset was uploaded to the IBM Watson Studio cloud storage and imported into a Jupyter Notebook. A total of 18 datasets were concatenated into a single dataset with 6,300 rows and 19 columns. Within this dataset were a large amount of duplicated values due to how the data needed to be downloaded. Also, the majority of the information provided with each data point was unnecessary for this project such as the year the house was built, the number of days it was on the market and so on. After dropping the unnecessary columns and duplicated rows the dataset was reduced to a size of 2,782 by 3 (Figure 1).

Out[220]:

	SOLD DATE	ZIP OR POSTAL CODE	$/SQUARE FEET
0	August-9-2019	68137	116.0
1	July-26-2019	68136	96.0
2	September-1-2017	68046	127.0
3	April-1-2020	68135	148.0
4	August-5-2019	68135	140.0

Figure 1. The cleaned dataset.

There was a lack of internal consistency in how the zip code of a house was recorded. Some zip codes included the unneeded four digit postal route code and so code was written to drop the unnecessary digits along with other zip codes that were obviously recorded incorrectly (Figure 2).

```
In [221]: df_omaha['ZIP OR POSTAL CODE'].unique()

Out[221]: array(['68137', '68136', '68046', '68135', '68154', '68164', '68116',
                 '68022', '68132', '68104', '68106', '68138', '68124', '68117',
                 '68114', '68118', '68127', '68130', '68105', '68134', '68122',
                 '68152', '68107', '68122-2269', '68111', '68144', '68116-2238',
                 '68108', '68111-0000', '68131', '68105-0000', '68106-3105',
                 '68110', '68142', '68007', '68112', '68166', '68135-3572',
                 '68107-1605', '68137-3730', '68137-3964', '68104-0000', '68146',
                 '68164-2237', '68022-0000', '68147', '68104-3554', '68016',
                 '51510', '68132-1705', '68114-0000', '68134-0000', '68144-0000',
                 '68112-3317', '68112-0000', '68112-3025', '68711-2', '68104-2307',
                 '68101', '68111-2703', '68110-1041', '68108-3536', '68105-1938',
                 '68107-0000', '68107-3312', '68109', '68107-1827', '68105-2204',
                 '68108-0000', '68105-2002', '68107-4327', '68107-3870', '68102',
                 '60107', '68107-3610', '68107-3906', '68147-1131', '38107',
                 '68107-3640', '68107-3728', '51503-0000', '68105-3635',
                 '68105-3774', '68107-1302', '68104-4502', '68104-3020',
                 '68104-3023', '68131-0000', '68104-3715', '68111-3950',
                 '68106-0000', '68104-2901', '68104-3811', '68106-1517',
                 '68104-4161', '68152-2252', '09197-2089', '68154-0000',
                 '68130-1970', '68154-2742', '68164-2510', '68164-6847',
                 '68142-0000', '68135-1227', '68135-1108', '68137-2486',
                 '68154-1146', '68135-1357'], dtype=object)
```

Figure 2. The lack of zip code internal consistency.

An additional dataset was downloaded from Opendatasoft that listed each Nebraskan zip code and the geospatial coordinates for each as well as extra information regarding time zone and daylight saving time flags. This extra information was dropped along with any zip codes pertaining to Nebraskan cities that were not Omaha. What remained was a new dataset only containing the zip codes of Omaha and their associated longitude and latitude (Figure 3).

Out[78]:

	Zip	City	Latitude	Longitude
4	68118	Omaha	41.263194	-96.171080
7	68180	Omaha	41.291736	-96.171104
11	68278	Omaha	41.264333	-95.946368
44	68181	Omaha	41.291736	-96.171104
53	68145	Omaha	41.291736	-96.171104

Figure 3. The data from Opendatasoft that lists each zip code with geospatial coordinates.

Later, Foursquare API will be used to attain the most popular venues of the best performing zip codes. This data containing datapoints of venue name, type, and location will be clustered using the K-Means algorithm in order to group the zip codes by venue type to provide more detail to real estate developers on the types of businesses each zip code favors.

B.3 Data organization

The data is now clean, but unusable in its current single dataframe. In order to contrast the different sections of Omaha, multiple dataframes were made from the existing mother dataset for each zip code. These individual zip code datasets were sorted by the date of sale after converting these dates to python readable datetime values (Figure 4). It will be from these individual subsets that the appreciation rate will be calculated.

Out[240]:

	SOLD DATE	$/SQUARE FEET	ZIP CODE
161	2017-04-14	87.0	68137
220	2017-04-21	98.0	68137
288	2017-04-21	137.0	68137
114	2017-05-02	88.0	68137
184	2017-05-02	101.0	68137

Figure 4. A dataframe containing values only pertaining to a unique zip code.

C. Methodology

C.1 Calculation and Visualization of the Appreciation Rates

Calculating the appreciation rate of a house is simple enough. Doing so retrospectively requires taking the change in price of a home and dividing by the amount of time that has caused that change. Therefore, with a real estate dataset of final sales price of several homes dating back a few years, calculating the average appreciation rate is possible. However, because I am grouping and comparing different homes to each other and not the same home to itself I decided to base my calculation on the

price per square foot metric as opposed to the overall sales price. Omaha is a growing city and so the land can be the most expensive part of the home, especially in the downtown areas, and it is for this reason that the price by square foot is a better way to compare houses belonging to the same area.

Each of the zip code dataframes were organized by date-time in ascending order before calculating the percent change of the price per square foot values. The average of these values was then found before dividing by the amount of time passed between the first and last entry. These results were then displayed in a bar plot (Figure 5) to see which of the zip codes performed at an above average rate in terms of appreciation.

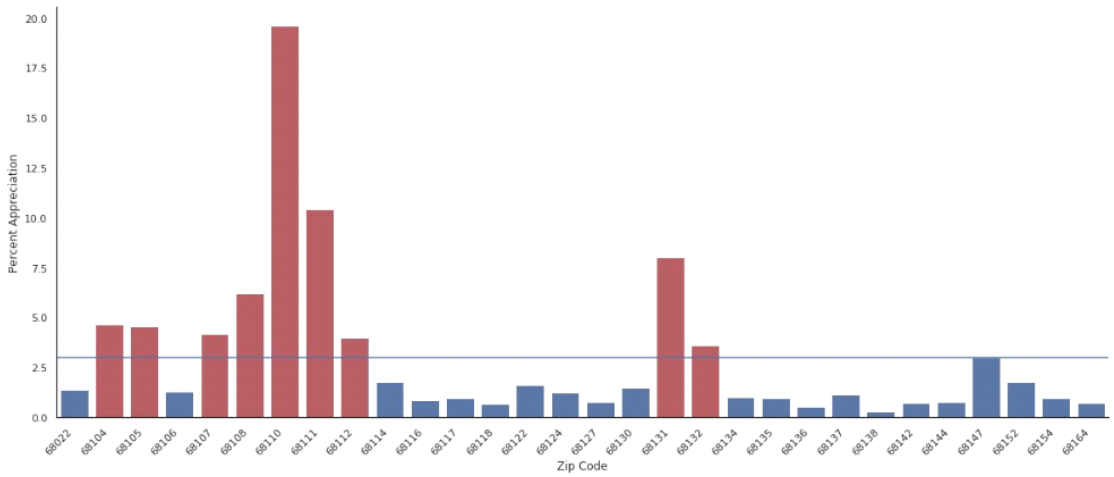

Figure 5. A bar plot showing the yearly appreciation rates of each Omaha zip code in comparison the city average.

The top performing zip codes were then separated from the rest and placed into a new dataframe along with additional data of zip code geospatial coordinates. By marking these areas on a map of Omaha (Figure 6) we are left with a visualization of the fastest growing areas of the city.

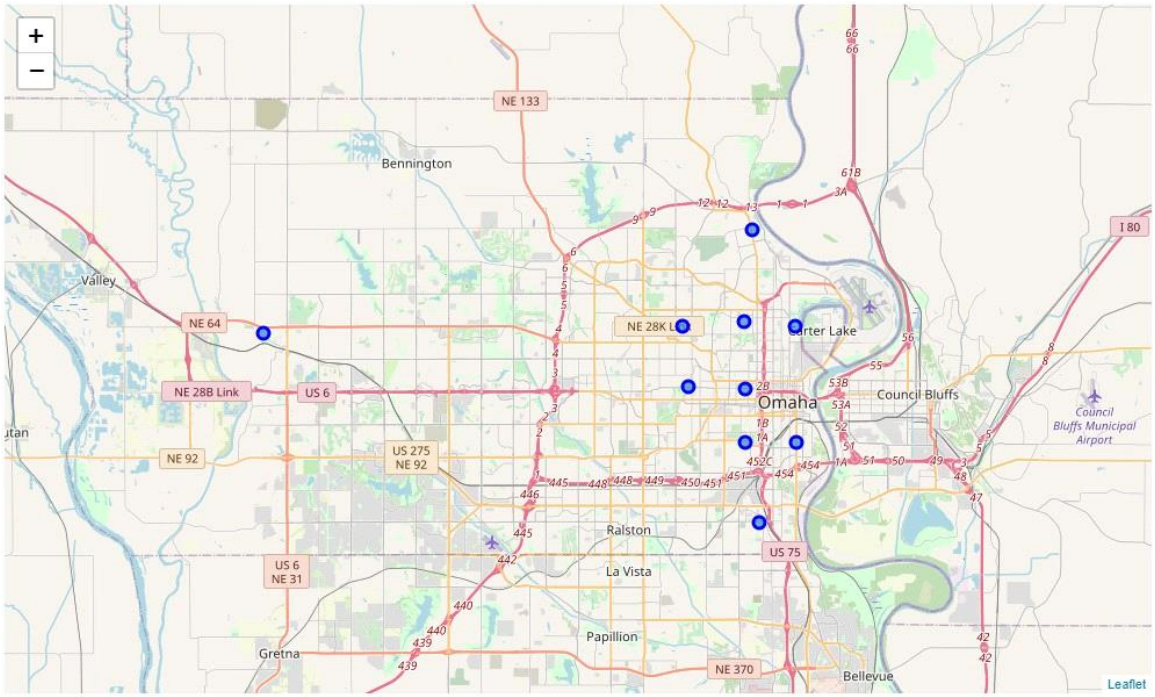

Figure 6. A map of Omaha, Nebraska featuring blue circles marking the top nine performing zip codes by yearly appreciation rate.

C.2 Examination of Most Popular Venues by Zip Code

Now that the scope has been narrowed from an entire city-wide view to a handful of the best areas for economic activity, it would be useful to real estate developers to group these areas based off of venue preference in order to understand what businesses are most supported. Foursquare API was harnessed to do so. From Foursquare a list of each top performing zip code's most popular venue name and category within a 1,000 meter radius was collected. The results were grouped into a dataframe before applying onehot encoding to prepare the data for machine learning.

C.3 K-Means Clustering

The machine learning algorithm k-means clustering was used to group the top performing zip codes based off of venue preference. This was an easy choice as k-means is perfectly suited to quickly group the zip codes based on preference without a training set of data. A quick trial and error determined five clusters to be the optimum amount for helpful grouping of the areas. These groups were then visualized by differing colors on the map of Omaha (Figure 7). For each group a dataframe was produced to display the top ten most common venue by type (Figure 8). A developer can now examine this information to determine which type of venue these zip codes are the most likely to support.

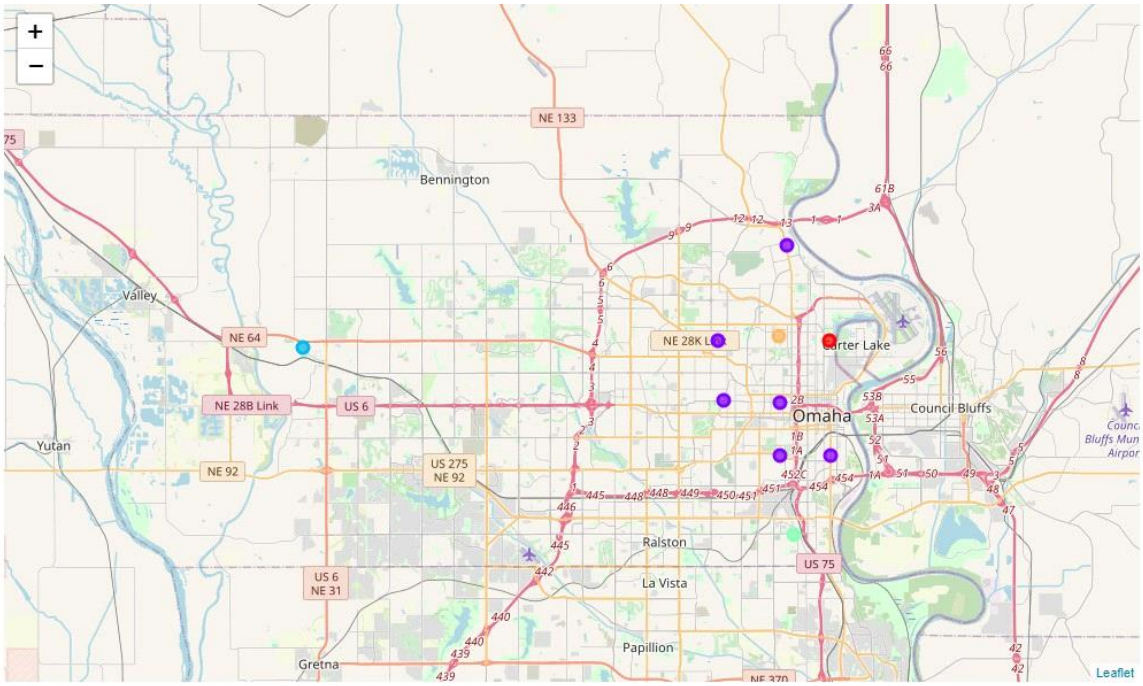

Figure 7. The same map of Omaha after grouping the zip codes with K-Means Clustering.

Out[309]:

	Zip	City	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
60	68107	Omaha	41.205198	-95.95539	3	Mexican Restaurant	Convenience Store	Discount Store	Gym / Fitness Center	Pizza Place	Supermarket	Dive Bar	Restaurant
104	68105	Omaha	41.240854	-95.96383	1	Mexican Restaurant	Park	Historic Site	Intersection	Dentist's Office	Greek Restaurant	Pool	Burger Joint
118	68110	Omaha	41.292321	-95.93427	0	Golf Course	Grocery Store	Thrift / Vintage Store	Discount Store	Shipping Store	Convenience Store	Baseball Field	Italian Restaurant
121	68104	Omaha	41.292445	-96.00060	1	Bar	Gas Station	Brewery	Restaurant	Pub	Pool	Burger Joint	Pizza Place
220	68132	Omaha	41.265650	-95.99741	1	Bar	Coffee Shop	American Restaurant	Gas Station	Pizza Place	Massage Studio	Playground	Park
311	68112	Omaha	41.334947	-95.95924	1	Pharmacy	Pizza Place	Historic Site	Sandwich Place	American Restaurant	Clothing Store	Convenience Store	Discount Store
324	68131	Omaha	41.264418	-95.96383	1	Bar	Park	Sandwich Place	Coffee Shop	Mexican Restaurant	Fast Food Restaurant	Gastropub	American Restaurant
418	68111	Omaha	41.294547	-95.96434	4	Gym / Fitness Center	Fast Food Restaurant	Discount Store	Bus Station	Grocery Store	IT Services	Gas Station	Fried Chicken Joint
605	68108	Omaha	41.240562	-95.93353	1	Dive Bar	Art Gallery	Sandwich Place	Mexican Restaurant	Discount Store	Performing Arts Venue	Pharmacy	Lawyer
0	68022	Omaha	41.289100	-96.24820	2	Park	Baseball Field	Construction & Landscaping	Home Service	Food Service	Dentist's Office	Diner	Discount Store

Figure 8. A dataframe featuring the most popular venues for each zip code.

D. Results

It was found that nine zip codes have an above city average appreciation rate after making the calculation based off of realtor listings of the price per square foot for single family homes. Visualizing this data shows the eastern section of the city to be greatly outperforming the rest with the exception of the recently annexed town of Elkhorn to the far west. These high appreciation rates illustrate the desirability of these areas and therefor indicate rapid future economic growth.

Clustering these areas using the k-means machine learning algorithm based off of venue preferences provided by Foursquare API allows for further area scouting. From the algorithm it is clear that the residents of cluster three seek out more active activities whereas the individuals in cluster two enjoy an active nightlife of bars and restaurants (Figure 9).

Cluster Two

```
In [315]: omaha_grouped.loc[omaha_grouped['Cluster Labels'] == 1, omaha_grouped.columns[[1] + list(range(5, omaha_grouped.shape[1]))]]
```

Out[315]:

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
104	Omaha	Mexican Restaurant	Park	Historic Site	Intersection	Dentist's Office	Greek Restaurant	Pool	Burger Joint	Dog Run	Brewery
121	Omaha	Bar	Gas Station	Brewery	Restaurant	Pub	Pool	Burger Joint	Pizza Place	Business Service	Music Venue
220	Omaha	Bar	Coffee Shop	American Restaurant	Gas Station	Pizza Place	Massage Studio	Playground	Park	New American Restaurant	Dive Bar
311	Omaha	Pharmacy	Pizza Place	Historic Site	Sandwich Place	American Restaurant	Clothing Store	Convenience Store	Discount Store	Farmers Market	Fast Food Restaurant
324	Omaha	Bar	Park	Sandwich Place	Coffee Shop	Mexican Restaurant	Fast Food Restaurant	Gastropub	American Restaurant	Lounge	Cocktail Bar
605	Omaha	Dive Bar	Art Gallery	Sandwich Place	Mexican Restaurant	Discount Store	Performing Arts Venue	Pharmacy	Lawyer	Smoke Shop	Italian Restaurant

Cluster Three

```
In [316]: omaha_grouped.loc[omaha_grouped['Cluster Labels'] == 2, omaha_grouped.columns[[1] + list(range(5, omaha_grouped.shape[1]))]]
```

Out[316]:

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Omaha	Park	Baseball Field	Construction & Landscaping	Home Service	Food Service	Dentist's Office	Diner	Discount Store	Dive Bar	Dog Run

Figure 9. Clusters two and three showing the preferences of each. The zip codes making up cluster two prefer bars and restaurants, whereas the individuals of cluster three seem to enjoy more outdoor activities.

E. Discussion

There are a large amount of variables to consider when scouting for real estate no matter what type real estate is being scouted for. For home buyers common needs are typical a close proximity to work, a good school district, and recreational options. For business owners tax rates, rent, and visibility must be taken into account. This project does not address these concerns, but it provides an excellent jumping off point into location selection. The volatility of the modern housing market is proof that the savvy home buyer must take the appreciation rate of their prospective houses into consideration as the home is the largest investment the majority of people make. For business owners, it is crucial to identify the target market and build nearby.

For this project the city of Omaha was divided by zip code, however there can still be significant differences between neighborhoods of the same zip code. For example, the zip code 68164 has a below average appreciation rate, but homes located in the 68164 neighborhood of Roanoke Estates have increased in value by about 4.6% in the past year alone. Closer examination of the area will reveal that Roanoke Estates, although a moderately desirable neighborhood, is situated near two section 8 (low income) housing programs, thereby resulting in an artificially low appreciation rate. Examining Omaha on a neighborhood basis instead of by zip code will remove this issue.

F. Conclusion

In this project, the appreciation rates for single family homes were calculated by zip code and those areas that were above the city average were grouped using the k-means machine learning algorithm

based off of most popular venue type. This information was overlayed onto a map of Omaha for visualization purposes. Appreciation rates and popular venue preferences are useful to real estate developers when scouting locations for new business construction. For instance, a bar franchisee would want to build in an area with a customer base that is both growing and likely to support a bar nightlife. Based off of the data processing in this project, the franchisee would be most interested in developing in the zip codes belonging to cluster two.

Data Sources:

Redfin

https://www.redfin.com/city/9417/NE/Omaha

Opendatasoft

https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/?refine.state=NE

Foursquare API

https://developer.foursquare.com/places