Segmenting the City of Omaha, Nebraska by Zip Code to Determine the Best Placement of Businesses

Charles Gambrel

1. Introduction

1.1 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.

1.2 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.

1.3 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.

2. Data acquisition and cleaning

2.1 Data acquisition

The majority of popular real estate sales websites work towards 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 highly appreciating zip code.

2.2 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 have already 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 8.

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 they were dropped along with other zip codes that were obviously recorded incorrectly (Figure 1).

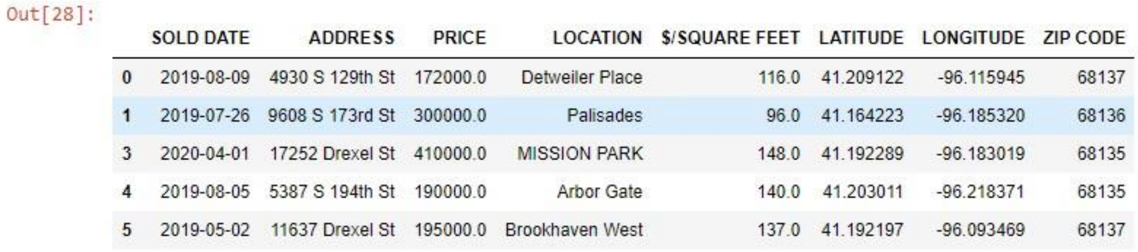
An additional dataset was downloaded from Opendatasoft that listed each Nebraskan zip code and the geospatial coordinates for each as well as extra information regarding time zone and daylight saving time flags. This extra information was dropped along with any zip codes pertaining to Nebraskan cities that were not Omaha. What remained was a new dataset only containing the zip codes of Omaha and their associated longitude and latitude (Figure 2).

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.

2.3 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 3). It will be from these individual subsets that the appreciation rate will be calculated.

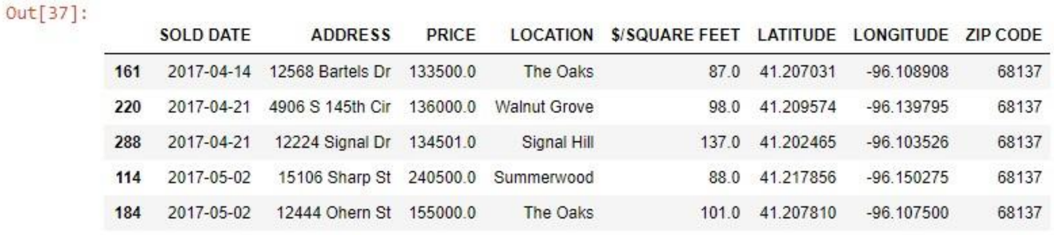
Figure 1. The cleaned dataset.



Out[28]:

	SOLD DATE	ADDRESS	PRICE	LOCATION	$/SQUARE FEET	LATITUDE	LONGITUDE	ZIP CODE
0	2019-08-09	4930 S 129th St	172000.0	Detweiler Place	116.0	41.209122	-96.115945	68137
1	2019-07-26	9608 S 173rd St	300000.0	Palisades	96.0	41.164223	-96.185320	68136
3	2020-04-01	17252 Drexel St	410000.0	MISSION PARK	148.0	41.192289	-96.183019	68135
4	2019-08-05	5387 S 194th St	190000.0	Arbor Gate	140.0	41.203011	-96.218371	68135
5	2019-05-02	11637 Drexel St	195000.0	Brookhaven West	137.0	41.192197	-96.093469	68137

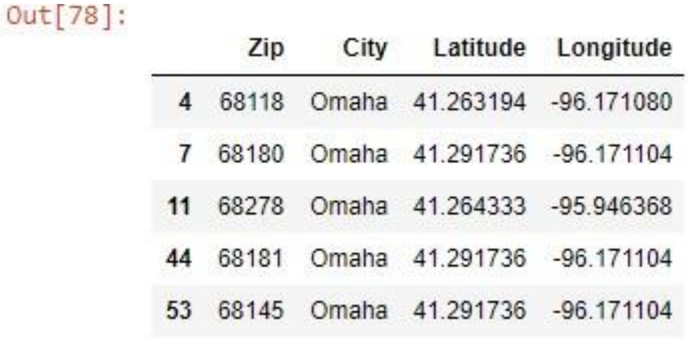
Figure 2. A dataframe containing values only pertaining to a unique zip code.



Out[37]:

	SOLD DATE	ADDRESS	PRICE	LOCATION	$/SQUARE FEET	LATITUDE	LONGITUDE	ZIP CODE
161	2017-04-14	12568 Bartels Dr	133500.0	The Oaks	87.0	41.207031	-96.108908	68137
220	2017-04-21	4906 S 145th Cir	136000.0	Walnut Grove	98.0	41.209574	-96.139795	68137
288	2017-04-21	12224 Signal Dr	134501.0	Signal Hill	137.0	41.202465	-96.103526	68137
114	2017-05-02	15106 Sharp St	240500.0	Summerwood	88.0	41.217856	-96.150275	68137
184	2017-05-02	12444 Ohern St	155000.0	The Oaks	101.0	41.207810	-96.107500	68137

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



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