Finding the best location for opening a new restaurant in Cleveland, Ohio

1 Introduction

1.1 Background

Whenever there is a plan of opening a new restaurant, there are many important aspects the potential owner has to be aware of. One of these aspects is the place where the restaurant might be located. The location of the restaurant has a huge impact on how successfull it will be. Two very important questions to be answered regarding the restaurant's location are the following:

1. How many people live nearby the restaurant?
2. How many other restaurants are already located nearby?

The answer to the first question is important as the majority of people might prefer to visit a restaurant in their neighborhood, rather than spending more time in the car to reach the place. The answer to the second question is important as it determines how stiff the competition in the region is. In this project we will compare different ZIP Codes of Cleveland, Ohio, based on their answers to the question above.

1.2 Problem

In order to compare different ZIP Codes of Cleveland based on the aspects declared above, we will need to combine different data (geographical, demographical, …) from different sources and merge them together. This project aims to give a suggestion for locations in Cleveland that might be suitable for opening a new restaurant.

1.3 Interest

This analysis is of interest for every person/company that plans on opening a new restaurant in the city of Ohio and needs to figure out a suitable location.

2 Data Acquisition and Cleaning

2.1 Data sources

The website https://zipcode.org/city/OH/CLEVELAND was scraped to receive a list of different ZIP Codes that belong to the city of Cleveland. I also scraped https://gist.github.com/erichurst/7882666 to receive the geo coordinates to each ZIP Code and https://www.zipdatamaps.com/zipcodes-cleveland-oh to gather the population of each ZIP Code. Last but not least, I used the Foursquare API https://api.foursquare.com for getting the venues nearby each ZIP Code.

2.2 Data Cleaning

The raw xml-data from https://zipcode.org/city/OH/CLEVELAND was filtered (using the BeautifulSoup library) so that it was able to be put in a Pandas DataFrame which contained the ZIP Codes in Cleveland:

	Zip Code
0	44101
1	44102
2	44103
3	44104

The raw data from https://gist.github.com/erichurst/7882666 was filtered and saved in a Pandas DataFrame that contained all ZIP Codes in the US along with their respective latitude and longitude. This dataframe was then merged into the first dataframe, so that a DataFrame was built which contained the ZIP Codes (only) of Cleveland along with their respective latitude and longitude. Some of the ZIP Codes were not assigned location data as they did not appear on https://gist.github.com/erichurst/7882666. I excluded these ZIP Codes from the DataFrame:

	Zip Code	Latitude	Longitude
0	44101	41.489355	-81.667393
1	44102	41.479174	-81.740603
2	44103	41.519415	-81.642123
3	44104	41.482230	-81.626784

The data scraped from https://www.zipdatamaps.com/zipcodes-cleveland-oh was filtered and stored in a DataFrame that contained the ZIP Codes of Cleveland along with their respective population. This information was again merged into the first DataFrame, so that in the end we got a DataFrame containing the ZIP Codes of Cleveland, their geocoordinates and their population. Some oft he ZIP Codes were not assigned a population number as they did not appear on https://www.zipdatamaps.com/zipcodes-cleveland-oh. I excluded these ZIP Codes from the DataFrame:

	Zip Code	Latitude	Longitude	Population
0	44102	41.479174	-81.740603	45014
1	44103	41.519415	-81.642123	18123
2	44104	41.482230	-81.626784	22640
3	44105	41.449476	-81.630289	40089

Based on the geo coordinates in this DataFrame, I used the explore-endpoint of the Foursquare API (https://api.foursquare.com) to get venues (incl. latitude, longitude, and category) nearby each ZIP Code. This resulted in a new DataFrame that had the following structure:

	Zip Code	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	44102	41.479174	-81.740603	78th Street Studios	41.484175	-81.739577	Art Gallery
1	44102	41.479174	-81.740603	Local West	41.482718	-81.735676	Sandwich Place
2	44102	41.479174	-81.740603	Banter Beer and Wine	41.482838	-81.735492	Food & Drink Shop
3	44102	41.479174	-81.740603	Don's Lighthouse	41.484667	-81.746104	Seafood Restaurant
4	44102	41.479174	-81.740603	Sweet Moses	41.483694	-81.731868	Dessert Shop

Using the columns „Venue Category“, I added the column „Restaurant“ to the DataFrame above. This column consists of boolean values which are True, if and only if the word „Restaurant“ could be found in the „Venue Category“:

	Zip Code	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Restaurant
0	44102	41.479174	-81.740603	78th Street Studios	41.484175	-81.739577	Art Gallery	False
1	44102	41.479174	-81.740603	Local West	41.482718	-81.735676	Sandwich Place	False
2	44102	41.479174	-81.740603	Banter Beer and Wine	41.482838	-81.735492	Food & Drink Shop	False
3	44102	41.479174	-81.740603	Don's Lighthouse	41.484667	-81.746104	Seafood Restaurant	True
4	44102	41.479174	-81.740603	Sweet Moses	41.483694	-81.731868	Dessert Shop	False

In the next step, I filtered out the venues that did not belong to any type of restaurant from the Data-Frame above:

	Zip Code	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Restaurant
0	44102	41.479174	-81.740603	Don's Lighthouse	41.484667	-81.746104	Seafood Restaurant	True
1	44102	41.479174	-81.740603	Luxe Kitchen & Lounge	41.483808	-81.731052	American Restaurant	True
2	44102	41.479174	-81.740603	Frank's Falafel House	41.481796	-81.729973	Falafel Restaurant	True
3	44102	41.479174	-81.740603	Villa Y Zapata	41.477276	-81.743501	Mexican Restaurant	True
4	44102	41.479174	-81.740603	Blue Habanero	41.484178	-81.729876	Mexican Restaurant	True

In the last step of Data Cleaning, I used the DataFrame above to count the Restaurants per ZIP Code and merge the result into the DataFrame that showed the ZIP Codes of Cleveland along with their latitudes, longitudes and population. For a better comparability, I also added to columns which normalize the „Population“ and „Restaurant“ columns using the min/max-method:

	Zip Code	Latitude	Longitude	Population	Population_normalized	Restaurant	Restaurant_norm
0	44102	41.479174	-81.740603	45014	1.000000	12.0	0.444444
1	44103	41.519415	-81.642123	18123	0.324160	9.0	0.333333
2	44104	41.482230	-81.626784	22640	0.437684	3.0	0.111111
3	44105	41.449476	-81.630289	40089	0.876222	6.0	0.222222
4	44106	41.505341	-81.605432	26896	0.544648	23.0	0.851852

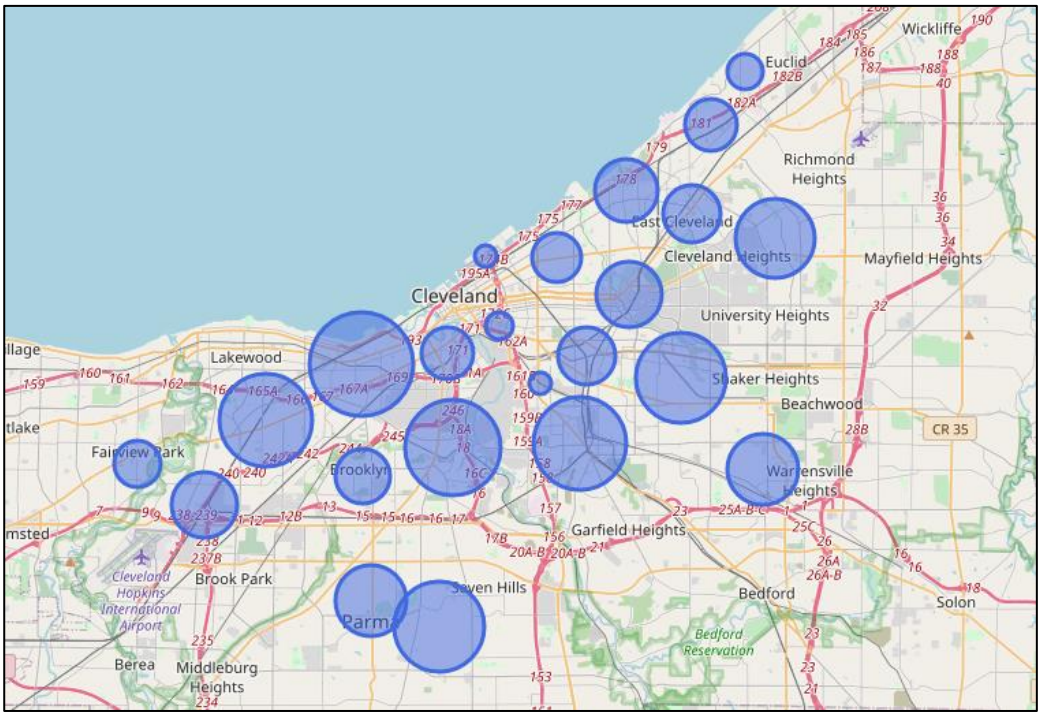
3 Methodology

3.1 Data Analysis

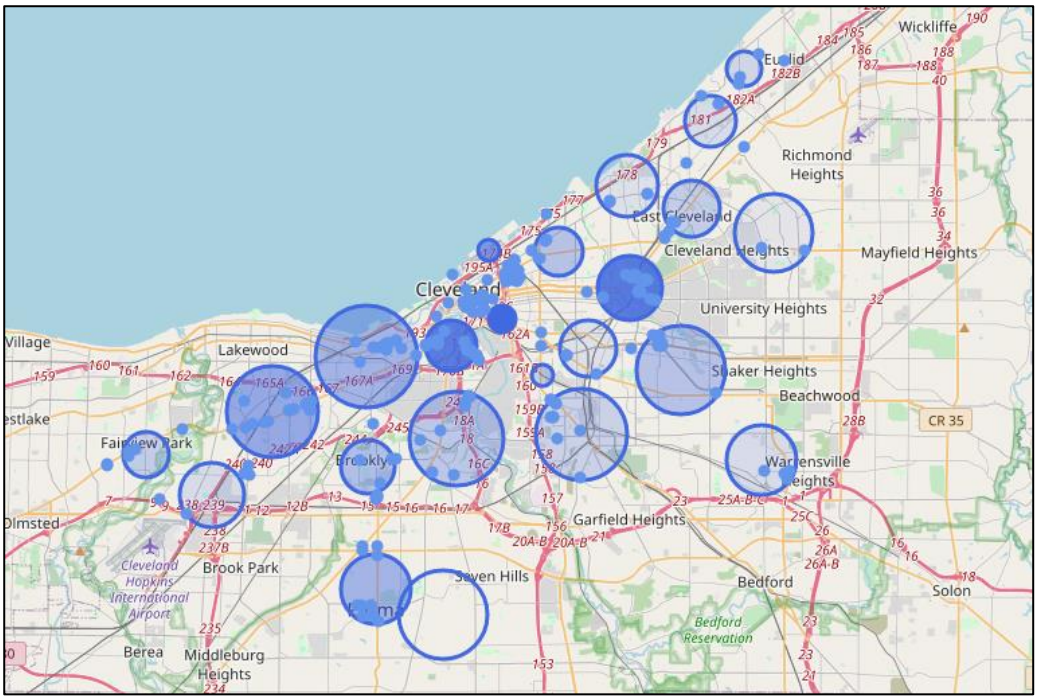
The first thing I analyzed was simply the geographical distribution of the ZIP Codes that I collected in the Data Section. One goal was to see whether the gathered information are correct. Another goal was to see how the different ZIP Codes distribute, i.e. whether there are some ZIP Codes that are very close to each other and/or ZIP Codes that are isolated from others
For this analysis, I created a Folium Map in which each data point represents one ZIP Code:

As we see in the map above, the ZIP Codes are satisfactorily-well distributed. Therefore, we did not have to adjust the data, for example by joining two ZIP Codes that are too close together or ignoring ZIP Codes that are too far away from the city centre.

In the second step, I visualized the population of each ZIP Code in the map. In the following map a higher radius of a circle marker means that the population in the according ZIP Code is higher than in ZIP Codes represented by smaller circles:



In the next step of analysis, I added the restaurants to the map above. Firstly, I added the location of each restaurant by adding a small blue dot for every restaurant. For better comparability, I additionally adjusted the fill opacity of every circle marker to represent the amount of restaurants nearby the ZIP Code it stands for. The higher the fill opacity of a circle, the higher the amount of restaurants nearby the according ZIP Code:

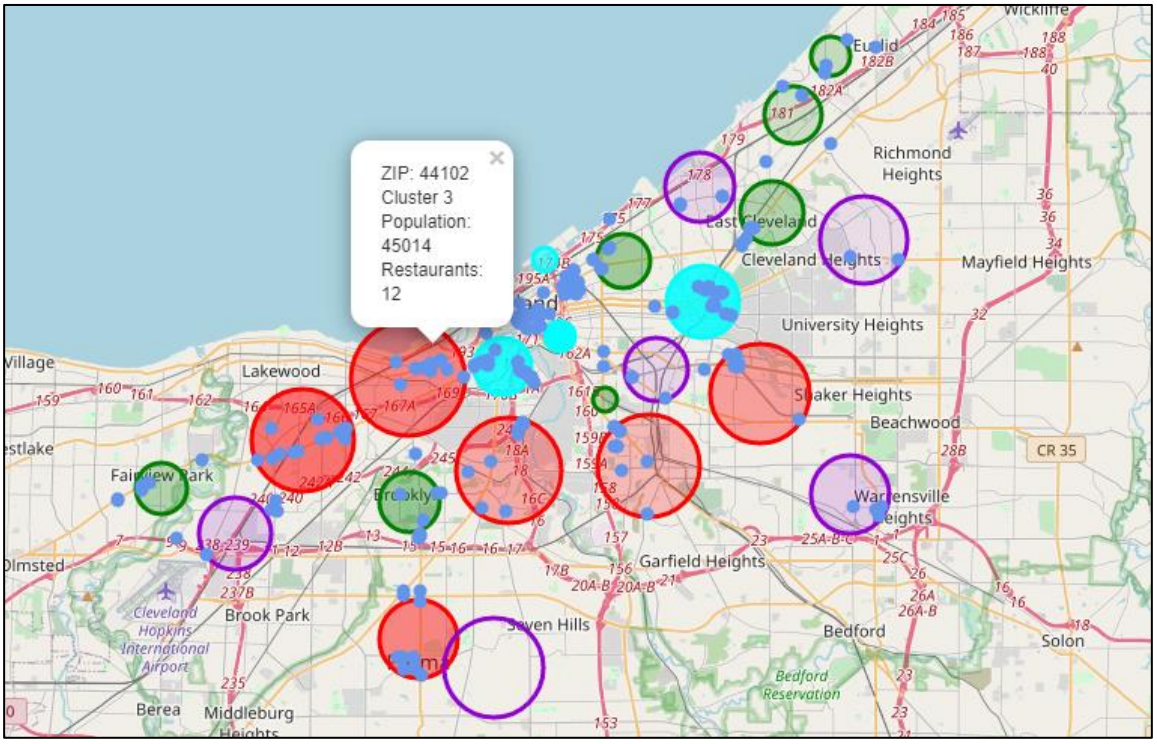


In the map above we see that there is no proportional relation between the population of a ZIP Code and the amount of restaurants nearby. Downtown there are smaller circles with high fill opacity, i.e. ZIP Codes with low population and a high density of restaurants. On the other hand there also bigger circles with high fill opacity (high population, high density of restaurants), bigger circles with low fill opacity.(high population, low density of restaurants) and smaller circles with low fill opacity (low population, low density of restaurants).

3.2 Clustering Algorithm

In the Data Analysis part we have already seen that there are different types of ZIP Codes regarding their population and restaurant density. In order to organize the ZIP Codes into different clusters, I used the KMeans-algorithm (from Scikit-Learn).

The goal of the KMeans-algorithm was to put the different ZIP Codes into clusters based on the min/max-normalized population and the min/max-normalized amount of restaurants. The following map shows the result of the KMeans-algorithm (with four centroids):



4 Results

Based on the data analysis and the KMeans-clustering, we can see that the ZIP Codes of Cleveland can be divided into four different groups with the following characteristics:

Cluster 0 (violet): medium high population / low restauraunt density
Cluster 1 (blue): low population / high restaurant density
Cluster 2 (green): medium low population / medium low restaurant density
Cluster 3 (red): high population / medium low restaurant density

The earned results lead us to answering the main question where to open a new restaurant in the city of Cleveland, Ohio.
In the ZIP Codes that belong to **Cluster 0 (violet)** there are not too many restaurants, but the population is relatively high. That means that the people living here may need to cover longer distances to reach a restaurant they want to visit. Therefore, it might be worth an idea to look at this ZIP Codes closer as the competition is not that big.
The ZIP Codes in **Cluster 1 (blue)** have a high restaurant density, but not many people are living here. These are the ZIP Codes located Downtown. People go shopping and/or working downtown,

so there are always many people around, although not many people live here. The amount of restaurants downtown show that this might also be a suitable location for opening a new restaurant, especially if the target is to attract people that occasionally walk by when shopping or visit a restaurant in their lunch break from work.
Cluster 2 (green) includes ZIP Codes that have a relatively low population and relatively low restaurant density. Not many people live nearby, neither there seems to be a lot of traffic. Therefore, these locations do not seem suitable for opening a new restaurant.
ZIP Codes in **Cluster 3 (red)** share the following characteristics. They have high population and medium low restaurant density. They might be suitable locations for a new restaurants because of the following reason. On one hand, there are many people that live nearby and therefore only had to cover short distances to visit the restaurant. On the other hand, there are only few restaurants which means that the competition is not too big.

6 Discussion

In the Results section I already hinted at some recommendations regarding the „best“ location to open a new restaurant in Cleveland. I pointed out that – for different reasons – the Clusters 0, 2 and 3 are suitable locations based on their population/restaurant-relation.
However, it is important to say that these recommendations are only indicators. For a fully reliable recommendation, there are many other aspects that need to be looked at. For example, it might not be sufficient to only look at the amount of people living nearby a specific location. The demographical structure – age, gender, employment etc. – is also a big factor that plays in a role in the decision making. Nevertheless, the recommendations made in the Results section give us a first hint at which ZIP Codes are worth having a more detailed look at them as the possible location for opening a new restaurant.

7 Conclusion

In this study, I analyzed the ZIP Codes in Cleveland, Ohio, in order to make recommendations on where to best open a new restaurant. Therefore, for every ZIP Code I gathered its location, its population and the restaurants nearby.
In the step of data analysis, the ZIP Codes were visualized along with their population and the amount of restaurants nearby. After that, we had a first sense of how the ZIP Codes differ from each other and how some are similar to others.
In order to group different ZIP Codes together, the KMeans-algorithm was used. This helped us to analyze the different locations better and make appropriate recommendation on where to open a new restaurant. However, in the Discussion section we also stated that this analysis only gives a first hint and helps on choosing locations that are worth analyzing them in more detail.