CS 418 Introduction to Data Science: Final Report

Restaurant Analysis using Food Inspection and Yelp Data

Group members:

Name & UIN (Captain):	Jason James Dsouza	677667371
Name & UIN:	Kunal Shah	667580441
Name & UIN:	Harsh Jethwani	670509818
Name & UIN:	Shivali Singh	665535339

1. Introduction

The project explores restaurants in Chicago and clusters them based on their location. This helps to classify the city into different zones (neighborhoods) and compare metrics such as average rating, average risk and the size of the clusters. The project is aimed to help customers or tourists in Chicago find the most dominant area for a specified cuisine to dine out. It can also act as an aid to new restaurateurs to explore the city to find localities where their restaurant is likely to be profitable. This is done by further analyzing the clusters to obtain information that can help identify areas with high competition or low interest for a particular cuisine.

As new international students in Chicago, we wanted to explore Chicago through its food lens, given the cultural diversity and breadth of cuisines available. Our motivation for this project came from us relocating to a new country and having problems finding places to dine out, due to the presence of an overwhelming variety of cuisines. We then realized how difficult it must be for tourists, picky eaters and people with dietary restrictions to find areas worth exploring. Thus, to solve this problem, we decided to make a project that can act as a guide to the food scene in Chicago. It can also help the native population explore different cuisines they haven't already tried, as well as aid new restaurateurs to explore the city to find localities where their restaurant is likely to be profitable.

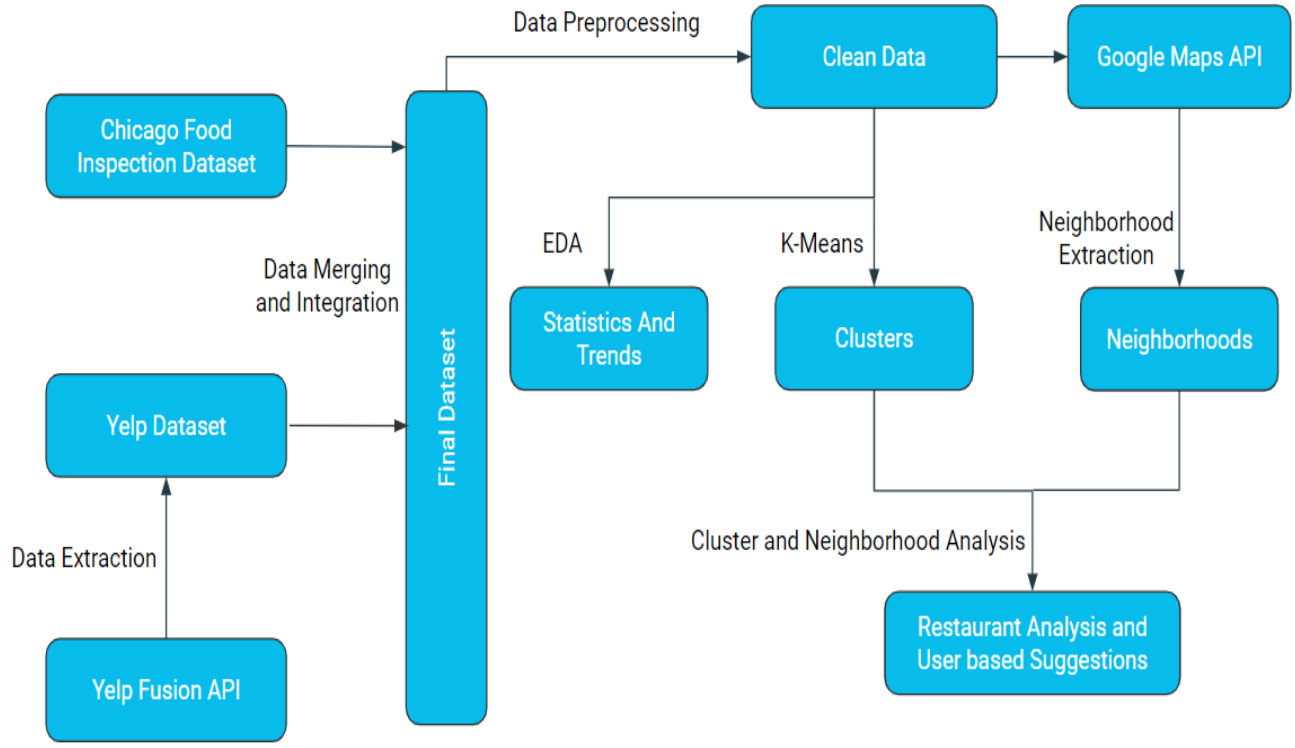


Fig 1. System Process Flow

2. Data

a. Chicago Food Inspection Dataset

This dataset is derived from inspections of restaurants and other food establishments in Chicago from January 1, 2010 to the present. Inspections are performed by staff from the Chicago Department of Public Health's Food Protection Program using a standardized procedure. This is a public dataset available at:
https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5
This dataset has all inspection data for food establishments and contains multiple entries for food establishments for inspection held in various years. The data had 224,799 records and each record has 17 attributes. Some key attributes are:

- DBA Name: Name of the facility
- Facility Type: Type of facility for ex: Restaurant, Grocery store
- Risk: Risk of the restaurant Risk 1 (High), Risk 2 (Medium), Risk 3 (Low).
- Address: The address of the restaurant
- Violations: Violations if any found
- Results: The result of the inspection - Pass, Fail, etc
- Latitude
- Longitude

Format: CSV

Dataset Size: 224799 x 17 (254 MB)

Dataset:

https://drive.google.com/file/d/1W4g4MfnGSTnIh7c1RJOOLwAEArbZizNh/view?usp=sharing

b. Data from Yelp

We used Yelp Fusion API's /businesses/{id} endpoint which returns detailed business content. The API returns a JSON response, which we collected into a JSON file, thus compiling the responses of all the businesses we requested.

Few important attributes are:

- categories: A list of category title and alias pairs associated with this business
- display_phone: Phone number of the business formatted nicely to be displayed to users. The format is the standard phone number format for the business's country.
- hours: Opening hours of the business.
- location.address1: Street address of this business.
- name: Name of this business.

- rating: Rating for this business (values range from 1, 1.5, … , 4.5, 5).
- transactions : A list of Yelp transactions that the business is registered for. Current supported values are "pickup", "delivery", and "restaurant_reservation".
- review_count : Number of reviews for this business.

Format: JSON response

Dataset Size: 18427 response (29.5 MB)

Dataset:

https://drive.google.com/file/d/13nxb1n2_gDzzSQTEv_Ug3Zdr78Q-85ci/view?usp=sharing

c. Final Dataset

We merged both the datasets to get our final dataset which will be used for our analysis.

Format: CSV

Dataset Size: 15266 x 30 (26.7 MB)

Dataset:

https://drive.google.com/file/d/1dHbgAaWBVuOLm2B8MiT4ckZ-_0LrJXQY/view?usp=sharing

3. Methods

For our task, we needed to collect data using the Yelp Fusion API and then merge it with the Chicago Food Inspection dataset.

a. Data Collection

The Yelp Fusion API provides only the first 1000 entries (20 results per API call and 50 offsets available), if you search businesses by keywords using the Business Search at https://www.yelp.com/developers/documentation/v3/business_search . This would severely limit our dataset and prevent us from conducting any solid analysis. To tackle this situation we performed the following steps:

- We first used the Chicago Food Inspection Dataset to identify and extract a list of unique businesses, which came up to 33,855.
- This was a key step, since we were then able to search for the business ID of our food facilities using its name and address, via Yelp's Business Match API at https://www.yelp.com/developers/documentation/v3/business_match .
- Once we had the business ID, we used it to get all the details of the businesses we had in the Chicago Food Inspection Dataset, for which Yelp had a corresponding data entry, using Yelp's Business Details API at https://www.yelp.com/developers/documentation/v3/business .
- This resulted in a dataset obtained from Yelp, having a total of 18,427 entries, in the form of a JSON file.

b. Data Integration

Now that we had our two datasets, the biggest challenge was merging them together using data integration.

- The business name and the address parameters from the Chicago Food Inspection Dataset were passed to the two Yelp APIs mentioned above, and the corresponding JSON responses returned by the API were stored in a JSON file. The names and addresses of the restaurants returned by the Yelp API were not exactly the same as the one passed to it from the Chicago Food Inspection Dataset (since using a strict matching would cause us to miss a lot of data). Thus, the post collection merging of the two datasets became a challenge.
- We tried out three libraries, to match the similarity of the name and address in the two datasets, which were SequenceMatcher, Jellyfish and Fuzzywuzzy.
- Fuzzywuzzy gave us the best results, and thus, the names and address matching was done using this library.

c. **Data Cleaning / Preprocessing**

- Extraction of the Year from Inspection Date column in the Chicago Food Inspection Dataset, so that we could keep the latest record of each business.
- Extraction of YelpAddress from the location column's dictionary in the Yelp dataset, to use it for similarity matching.
- Cleaning up and imputing NaN values for some critical columns.
- Dropping records of the Chicago Food Inspection Dataset, for which there were no corresponding records in the Yelp dataset, to use it for merging. After this process, we ended up with a total dataset size of 15,266 records. This is because:
  1. Yelp did not have the data for some businesses from the Chicago Food Inspection Dataset, like churches, school canteens, etc.
  2. Yelp stores data for only those businesses that have reviews. Businesses without reviews are not available on Yelp
  3. The Chicago Food Inspection Dataset had repeating columns for the same business, depending on the number of inspections performed for that business.
  4. Some of the businesses from the Chicago Food Inspection Dataset are now shut and thus, have no record on Yelp.
  5. The similarity matching technique failed to correctly match some entries (very few), and thus, these records got dropped.
- Dropping unnecessary columns from the merged dataset, like id, image_url, City, State, phone, AKA Name, as well as other unnecessary or repeated columns.
- Parsing the strings having list and dictionary syntax into their respective data types (list/dict).
- Converting the price column from the number of '$' symbols to a numeric value.
- Adding an extra column for the number of transaction types, to perform EDA.
- In the future, we also plan to create more columns via data extraction from converted lists/dictionaries, and any encoding if needed.

d. **Google Maps API for fetching neighborhood**

- Used Google Maps' reverse geocoding API to retrieve neighborhoods of each restaurant, using their latitude and longitude.
- Found the neighborhood of each cluster by assigning the neighborhood with the highest frequency, to get a rough estimate of the neighborhood of the cluster.
- Obtained a rough estimate of the neighborhood, instead of a cluster label for making our suggestions more meaningful and insightful.

e. Preliminary EDA and Insights/Observations

Descriptive Analysis

- The mean review count is 117.2
- The average Yelp rating is 3.54
- Analysis on 'Results' attribute showed that about 5456 businesses fell out of business between 2010 and 2021.
- About 10,000 establishments have a high risk of one or more violations.
- Of all the establishments, about 11,000 are restaurants
- The zip code 60614 has the most number of establishments inspected, with about 839 businesses.
- The Chicago Department of Public Health's Food Protection Program inspected 4150 establishments in 2021 which is the highest ever in a year.

Trend Analysis

- The average number of reviews increases with the total rating till it reaches the maximum value for the rating 4.0, and after that it decreases sharply.
- The average rating for restaurants was found to be directly proportional to the number of food facilities till the rating of 4.0; and above the rating of 4.0, the number of food facilities was found to decrease with increase in rating.
- The average number of reviews for a restaurant is directly proportional to the number of transaction methods of that restaurant.
- The average rating of a restaurant was also found to be directly proportional to the number of transaction methods of that restaurant.
- Surprisingly, the average rating was found to be higher for restaurants with higher risk factors than those of the lower risks!
- The number of open businesses is always higher than the number of closed businesses, with respect to the average rating, even for those with a rating of only 1.0!

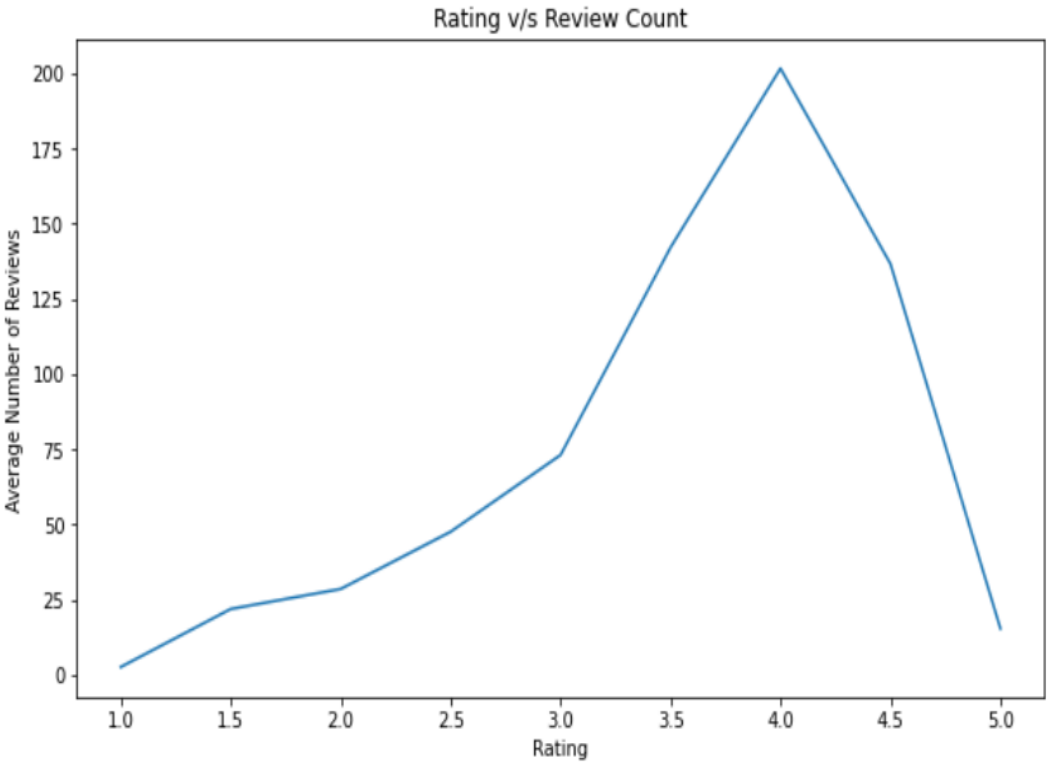


Fig 2. EDA: Rating v/s Review Count

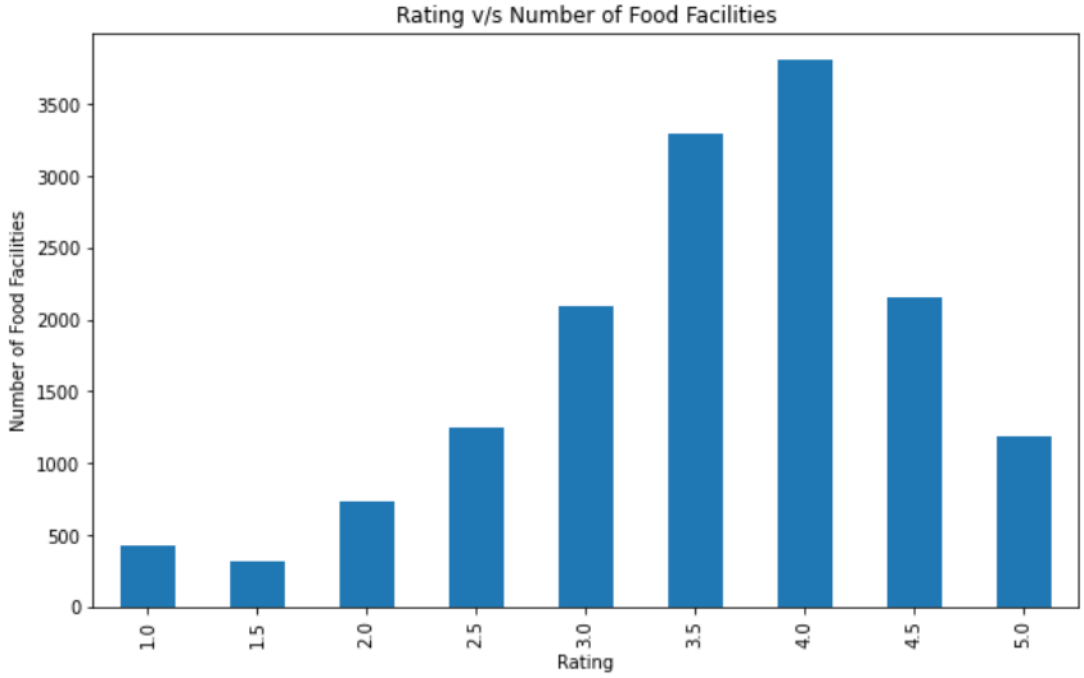


Fig 3. EDA: Rating v/s Number of Food Facilities

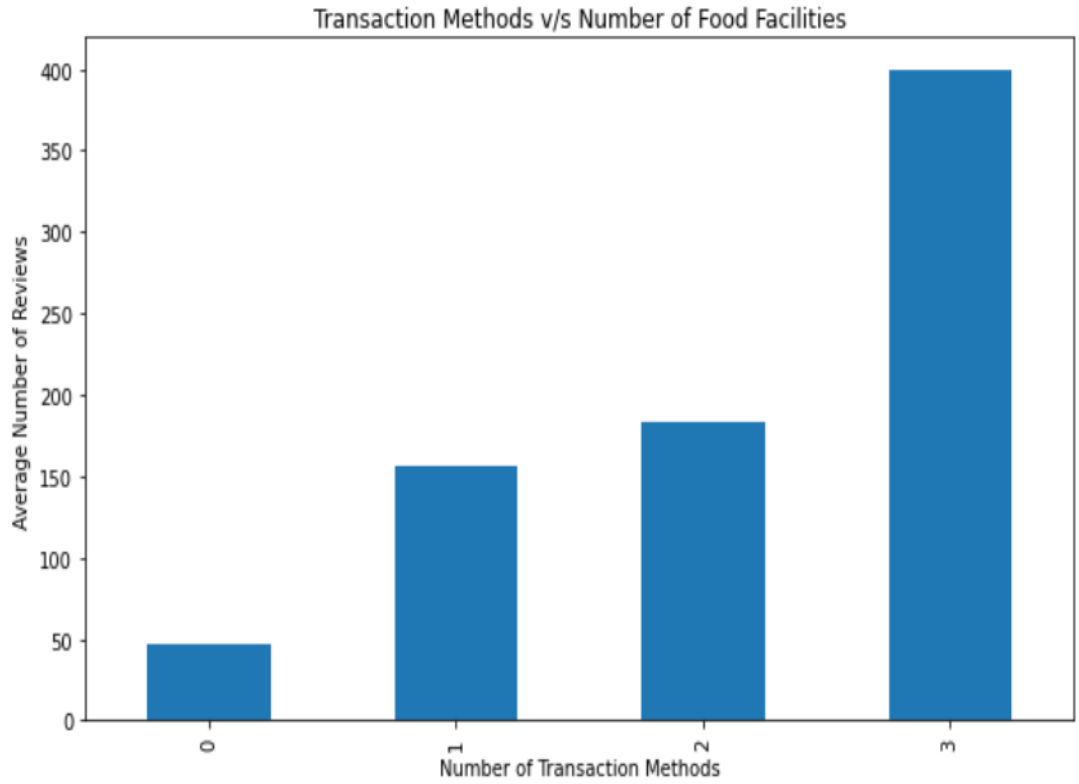


Fig 4. EDA: Transaction Methods v/s Number of Food Facilities

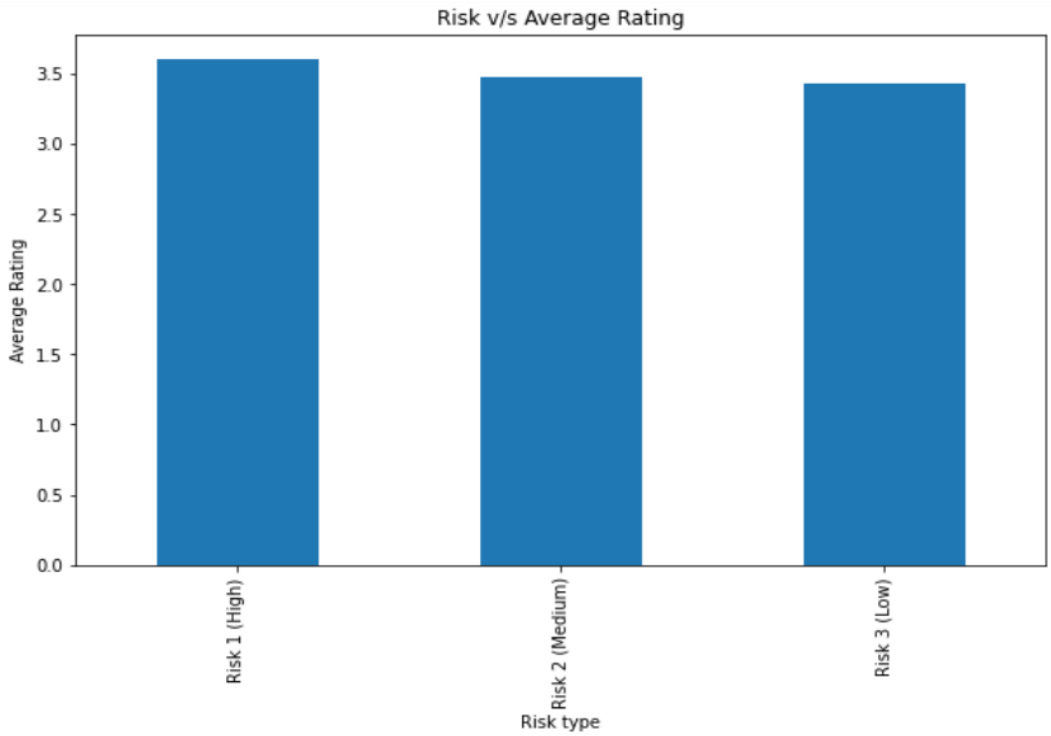


Fig 5. EDA: Risk v/s Average Rating

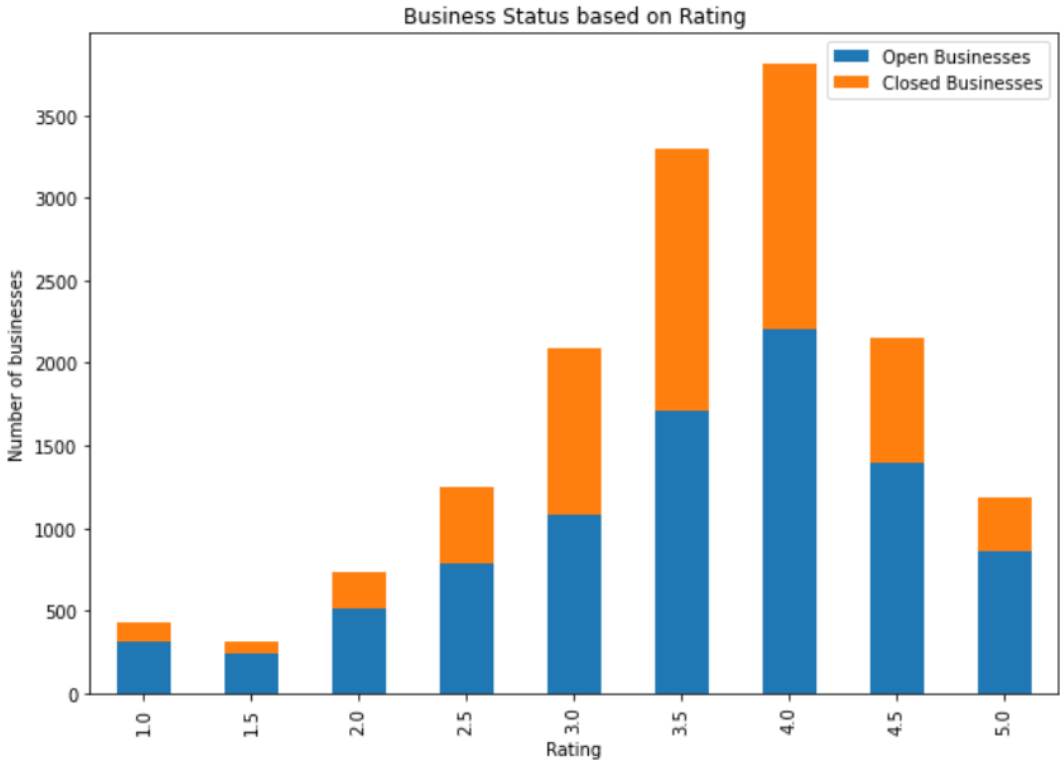


Fig 6. EDA: Business Status based on Rating

4. Analysis

a. <u>Data Modeling: Clustering</u>

To proceed with clustering, we had to find the best clustering algorithm for our data, based on the latitude and longitude of the restaurants.

We experimented with different clustering techniques, such as:

i. K-Means
ii. DBSCAN
iii. HDBSCAN

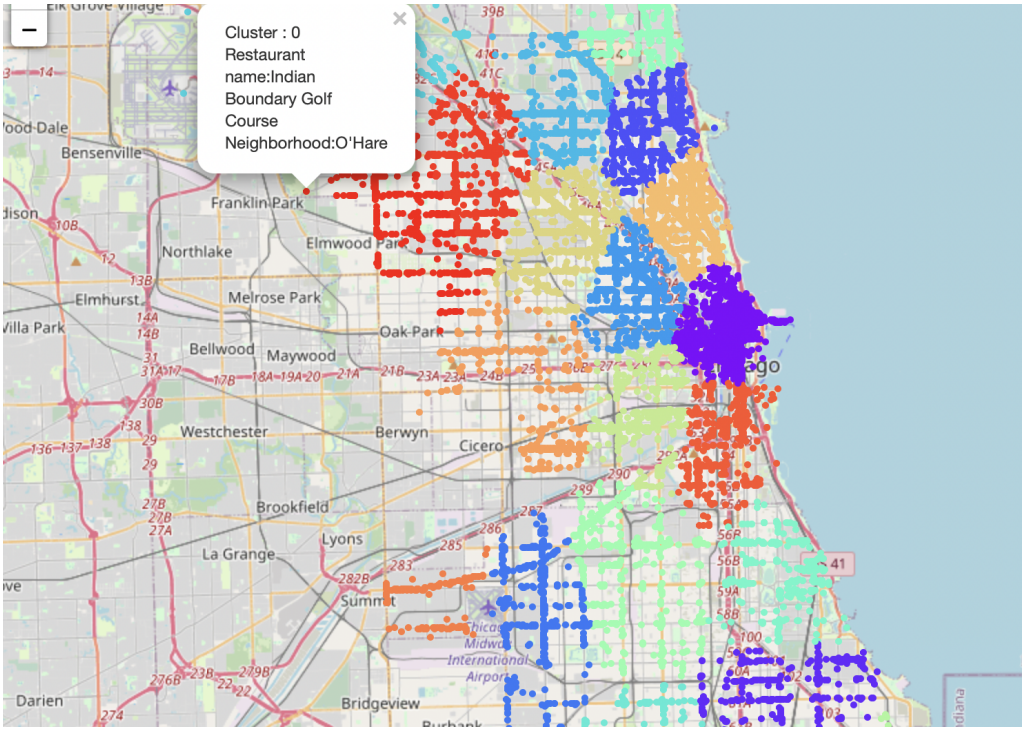
- Algorithms like DBSCAN, HDBSCAN, OPTICS perform density based clustering.
- In general, these algorithms used with Haversine distance perform much better on geospatial data, as compared to a non density based clustering algorithm like K-Means.
- However, in our case, we have a large number of restaurants within a single, much smaller geographical entity - the city of Chicago, instead of the data being spread across multiple cities in a country.
- This makes it difficult for density based algorithms to form multiple clusters of a meaningful cluster size.
- We observed that such algorithms usually formed 1 to 3 very large clusters and

treated the other points as noise or single point clusters.

- Thus, we decided to proceed with K-Means, since this algorithm gives us the flexibility of selecting the number of clusters, without being affected by the high density of most data points in a single region.

b. Cluster Analysis: Neighborhoods

- We used the neighborhoods obtained from Google Maps API along with our clusters and plotted them using folium. The figure below displays the clusters we obtained, each having a unique color, based on its neighborhood.



0:Austin
1:Chicago Loop
2:Greater Grand Crossing
3:North Center
4:West Lawn
5:West Town
6:Albany Park
7:Norwood Park
8:Beverly
9:Hyde Park
10:Rogers Park
11:Brighton Park
12:East Side
13:Lower West Side
14:Logan Square
15:Lake View East
16:South Lawndale
17:Garfield Ridge
18:Armour Square
19:Portage Park

Fig 7. Clustering: Neighborhoods

- Our project plots all the restaurants on the map with each color representing a cluster.
- There is also a pop-up of every restaurant that displays the corresponding cluster number, restaurant name and the neighborhood it belongs to.

c. Cluster Analysis: Business Opportunity Ranking

We define Business Opportunity Ranking for a particular cluster 'i' as:

$$= \frac{AverageRating_i \; X \; AverageRisk_i \; X \; ClusterSize_i}{TotalAverageRating \; X \; TotalAverageRisk \; X \; TotalAverageSize} \; X \; 100 \; X \; 100 \; X \; 100$$

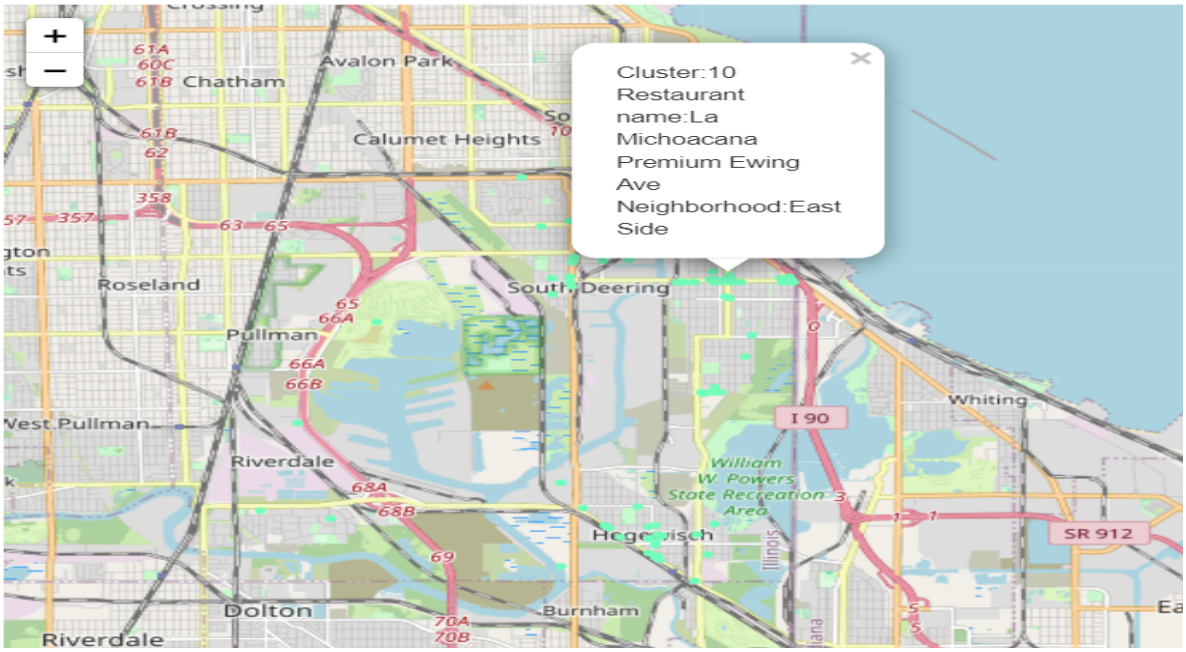
In general, the chances of a business being successful is:
(i) Inversely proportional to the cluster size (greater the size more dense will be the cluster and thus more will be the competition).

(ii) Inversely proportional to the cluster average rating (greater the rating lesser will be the chance of new business to succeed among the successful competition).

(iii) Inversely proportional to the risk rating of a cluster (lesser the average risk rating, better is the inspection record of that cluster, more will be the competition from good inspection records restaurants).

Smaller the value of Business Opportunity Ranking for a cluster, more will be the scope of success of new business in that area!

```
The best cluster to open restaurant is 10:East Side
```



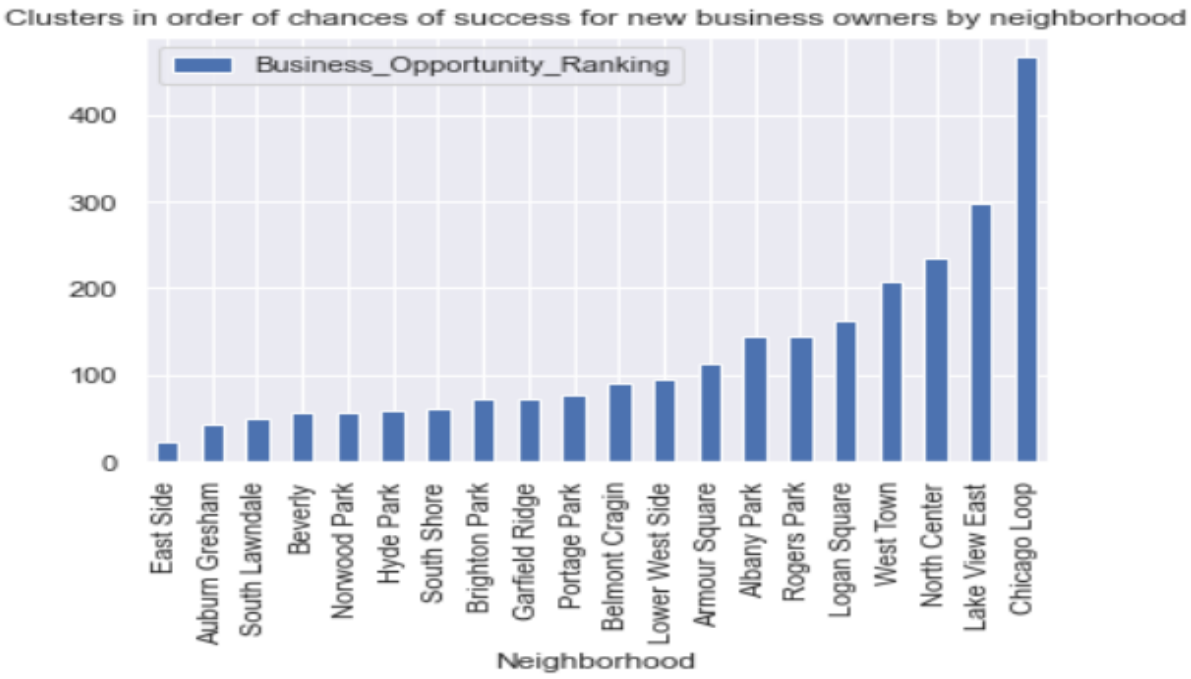


Fig 8. Business Opportunity Ranking

- Business opportunity ranking will give us those neighborhoods, where the overall chances of competition will be less for a new business.
- Lesser the value of business opportunity ranking, more will be the chances of a new business to succeed, as all the restaurants in that cluster will overall have less average rating, poor inspection records and less restaurants.
- The bar graph will give the order of neighborhoods that will have a higher scope of success for a new business.

d. Cluster Analysis: Miscellaneous

We also performed clusterwise analysis on other factors like average risk, average rating, business failure rate, etc.

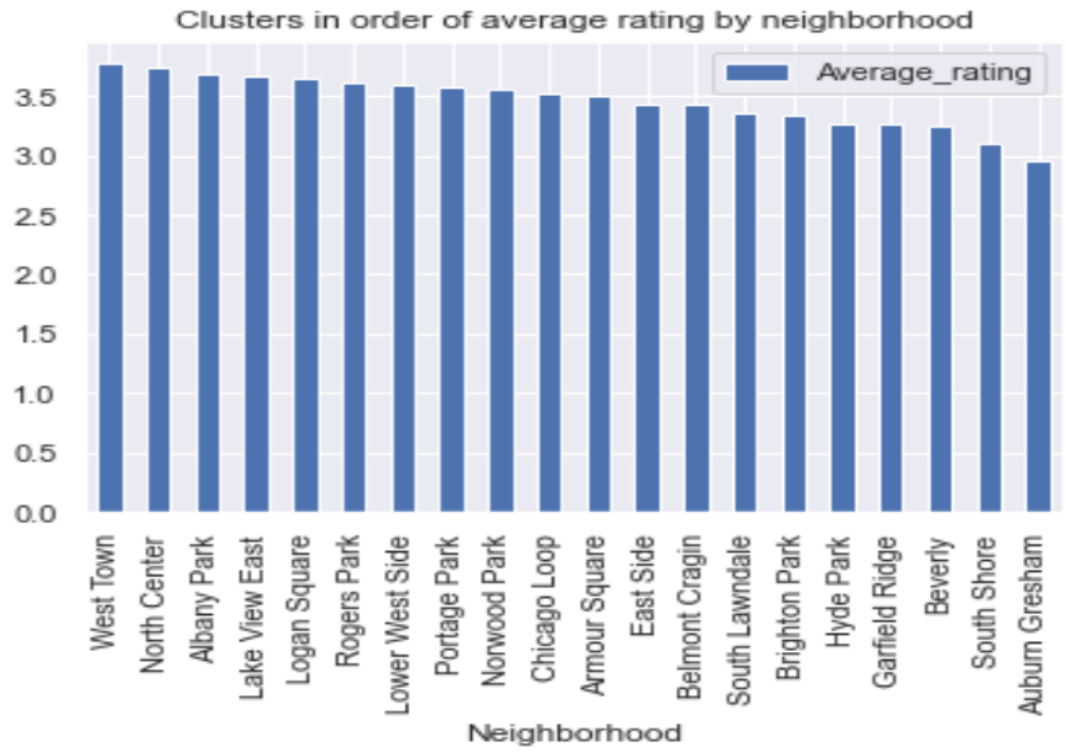


Fig 9. Clusters in order of average rating by neighborhood

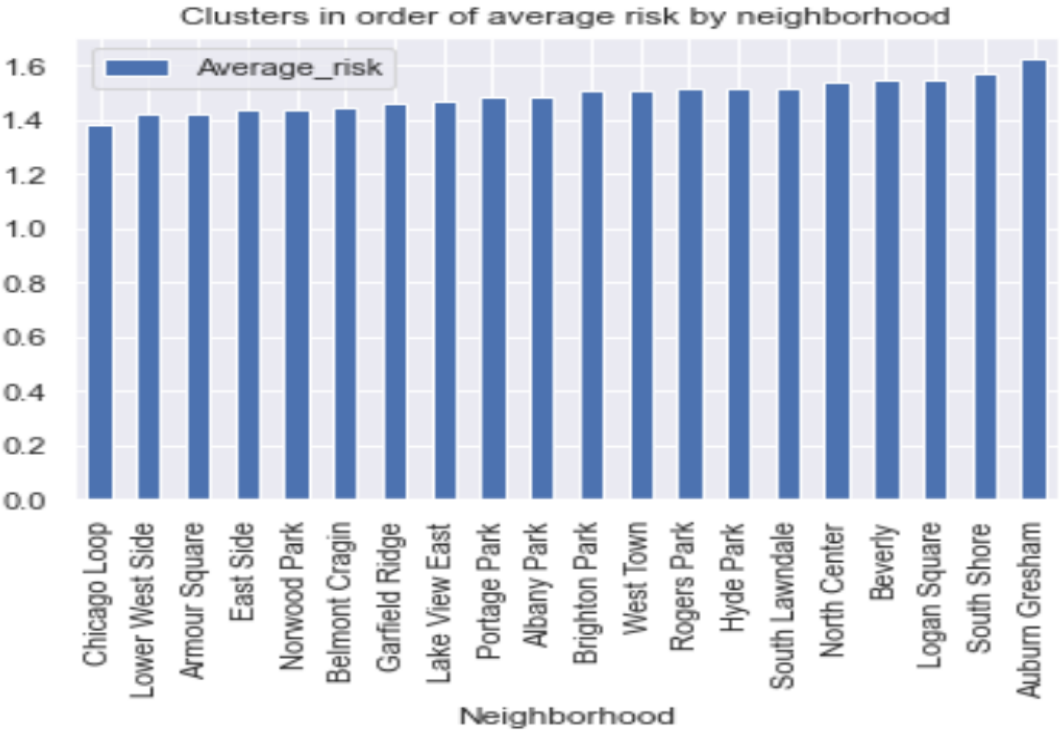


Fig 10. Clusters in order of average risk by neighborhood

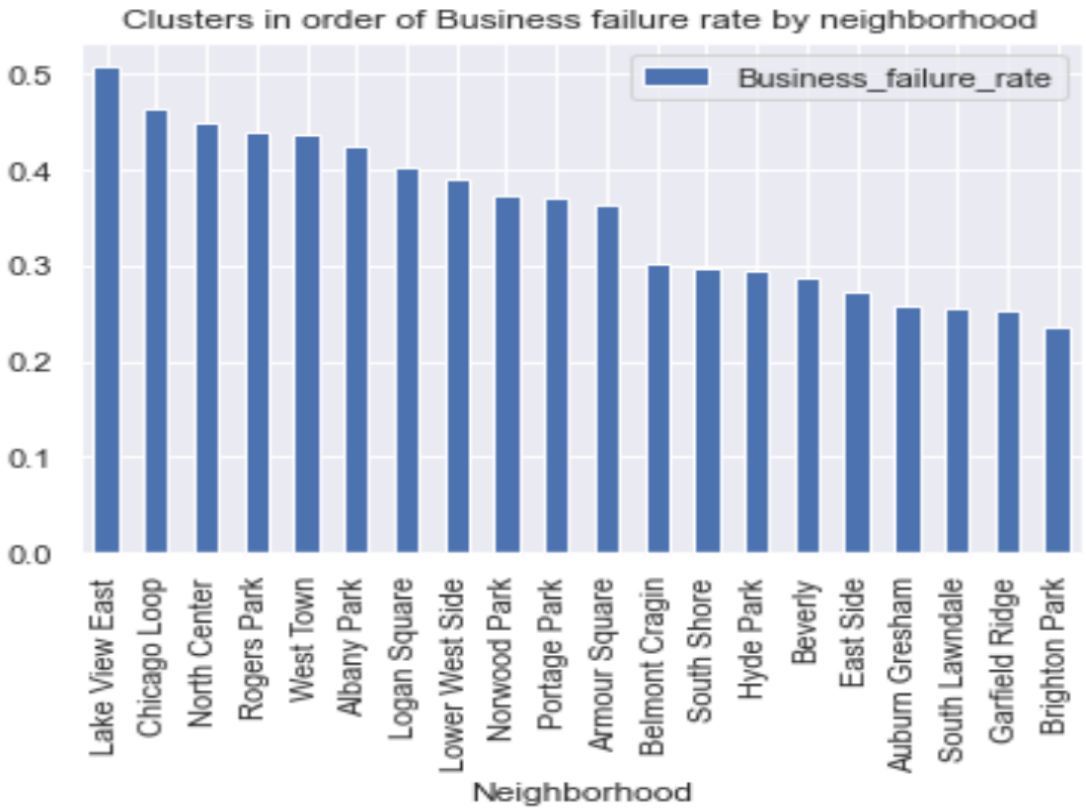


Fig 11. Clusters in order of business failure rate by neighborhood

5. Results

Cuisine Dominant Cluster

- Taking a particular cuisine from the user and returning the cluster where that particular cuisine is dominant.
- This data can help in finding the spread of various cuisines over the map of Chicago.
- This can also help a user find the neighborhood which would be perfect for exploring a particular cuisine type.
- New business owners can use this analysis to study the market for their target cuisine to better choose the neighborhood for opening their business.

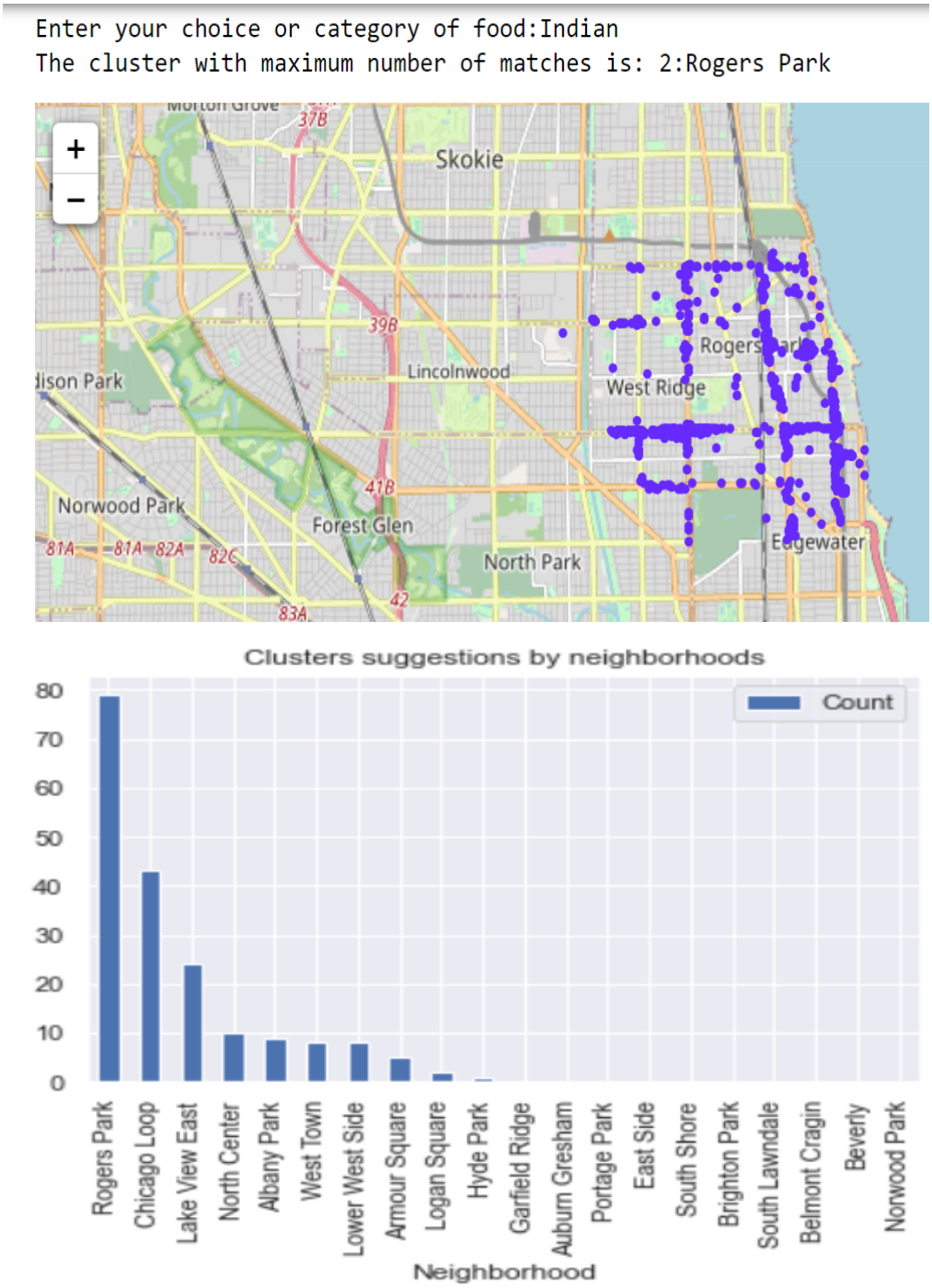
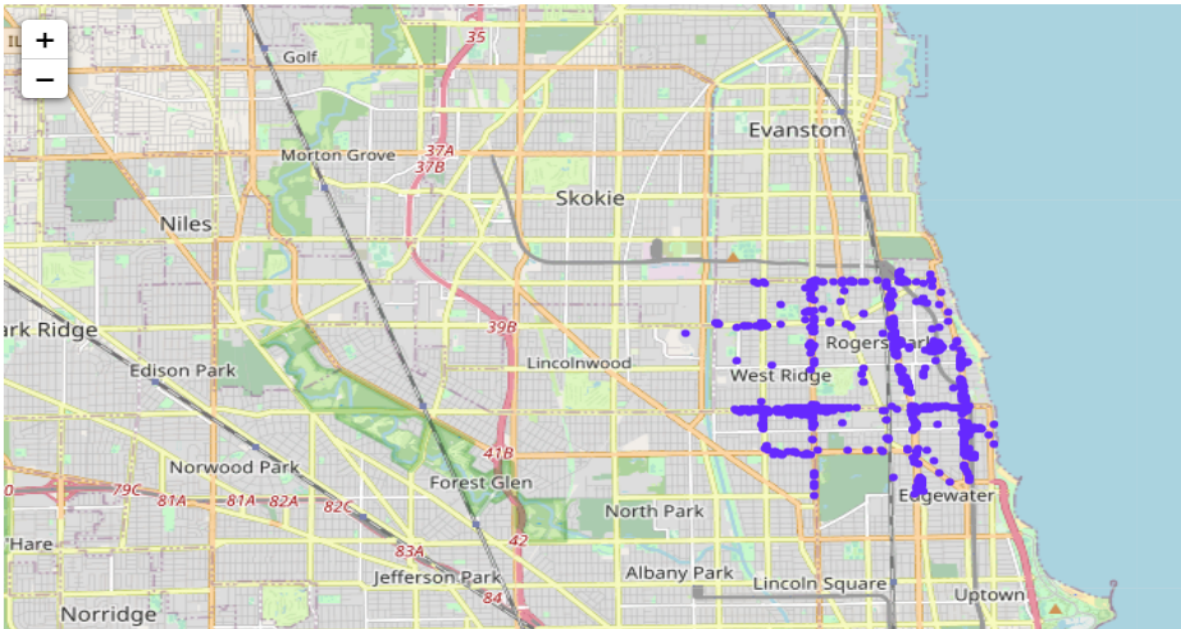


Fig 12. Cuisine Dominant Cluster

- The bar graph will help to better visualize the neighborhood by the dominance of the user entered cuisine.
- We can also find out those neighborhoods where a particular cuisine is not available and thus serve as the potential neighborhoods for new business planners, not seeking business competition.

```
Enter your choice or category of food:Indian
The cluster with maximum number of matches is: 2:Rogers Park
```



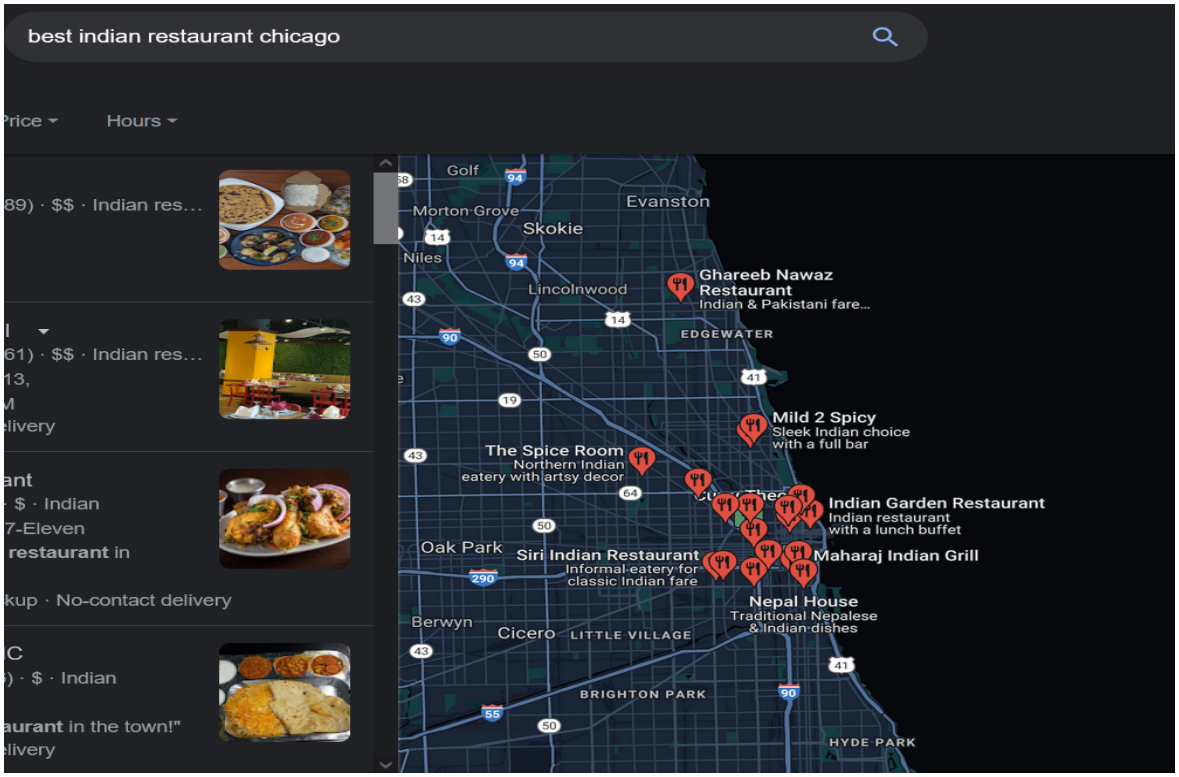
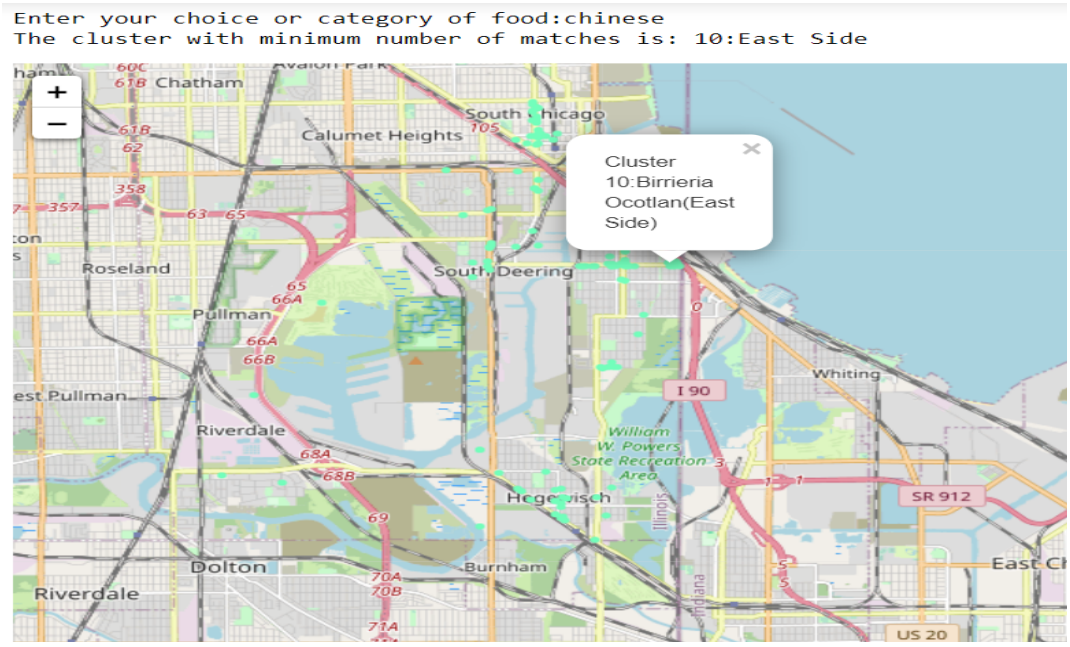


Fig 13. Comparing our suggestions with Google results

Cluster with least restaurants of a particular cuisine (Competition)

- Taking a particular cuisine from the user and returning the cluster where that particular cuisine is least popular.
- This data set will mainly help new business owners find those neighborhoods, where businesses serving a particular cuisine are either not available or are less in number as compared to other neighborhoods.
- These clusters could imply less competition, and therefore, higher chances of the business being successful.



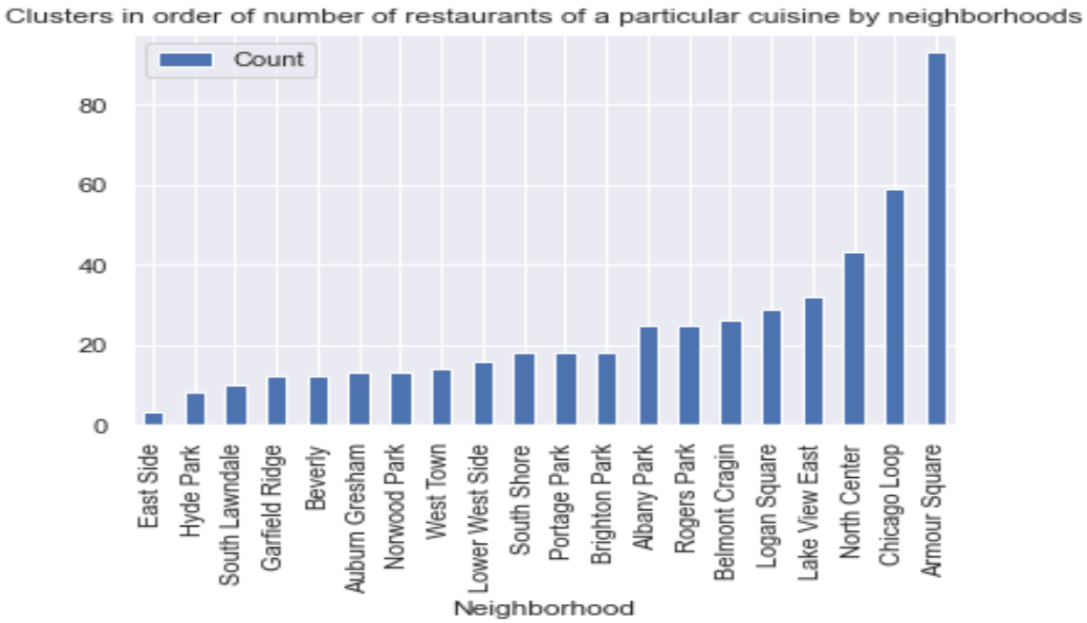


Fig 14. Cuisine Competition Cluster

- The bar graph will help to better visualize the neighborhoods by the least number of restaurants of the user entered cuisine.
- We will get the order of neighborhoods by the number of restaurants of a particular cuisine and thus serve as the potential neighborhoods for new business planners, not seeking business competition.

6. Conclusion

- Our project explored the restaurants in Chicago and clusters them based on their location.
- We then performed clusterwise analysis on the clusters formed.
- This analysis was used to explore various cuisines in Chicago, based on the user input.
- New business planners can also use the insights from our analysis to make better decisions which will increase their chances of success.