# Coursera Applied Data Science Capstone

# Identifying the best location in Chicago for opening a cleaning business

Dan Grigore

February 25, 2020

# Table of Contents

# 1. Description of the Problem and Background

## 1.1. Background

Chicago is the third most-populous city in The United States and the most populous city in the State of Illinois. Having an estimated population of over 3 million people, it is a central node of business, communication and transportation in the US.

According to Inc. magazine (https://www.inc.com/emily-canal/chicago-startup-city-talent-funding-inc-fast-growth-tour.html), Chicago is one of the most hospitable cities for entrepreneurs and startup businesses, contrary to its harsh winters and high competitive environment.

Among top reasons why one should open a business in this city are:

1. Diverse and healthy business climate
2. Major center for business in the Midwest
3. Easy finding fundraising opportunities
4. Reasonable cost of living
5. Highly valued educational system

## 1.2. Problem

With over 300.000 companies based in the city (https://www.census.gov/quickfacts/chicagocityillinois) in a 227 square miles area, it is real problem keeping the business clean. Keeping the business clean is usually a daily objective for most of the companies in order not to:

1. Turn customers away
2. Make employees more productive

3. Keep equipment and in good condition
4. Work in a Safe environment

Assigning this task to a team inside the company is an idea but outsourcing towards a professional cleaning company can reduce cost and increase productivity.

The **problem** would be finding a suitable location for the startup, keeping in mind the following conditions:

1. Must be close to venues in most need of daily cleaning
2. Must have a primary location, close to most popular venues in Chicago needing cleaning, where employees would be located and wait for requests, keep supplies and inventory
3. Can have other secondary locations, if needed, for helping reducing transportation costs to/from cleaning locations
4. The startup will concentrate mostly on restaurant, bar and hotel businesses

My objective for this analysis is to identify the best 2 places in Chicago for an entrepreneur in the commercial cleaning industry to open a business and offer its services.

### 1.3. Interest

Any entrepreneur willing to open a business in the cleaning industry in Chicago would be interesting in identifying the best locations for opening and work office in the nearby area of the most popular venues in the city. Also, venues would be interested in finding the cleaning company situated in closed location, thus getting the requested services in time, any time they would be interested in.

The advantages of opening a cleaning business are:

1. Can generate revenue very quickly
2. Low costs as employees can be kept part-time or from home
3. The commercial cleaning is a very flexible industry, offering the access to a variety of businesses with a variety of needs

There are many businesses in need for cleaning services, but for this research I decided to restrict the results to restaurants, bars and hotels.

# 2. Data and cleaning

## 2.1. Data sources

The data used in this report is the following:

1. Geonames.org zip file containing postal code information for all cities of The United States. More on Geonames.org can be found at the link https://en.wikipedia.org/wiki/GeoNames.

   The postal codes information file is in CSV format and it contains the following information:

   - Country Code
   - ***Postal Code - used in my future analysis***
   - ***Place Name - City - used in my future analysis***
   - ***State Name - State - used in my future analysis***
   - ***State Code - used in my future analysis***
   - County Name
   - County Code
   - Community Name
   - Community Code
   - ***Latitude - used in my future analysis***
   - ***Longitude -used in my future analysis***
   - Accuracy

The bolded columns represent the information used in the report.

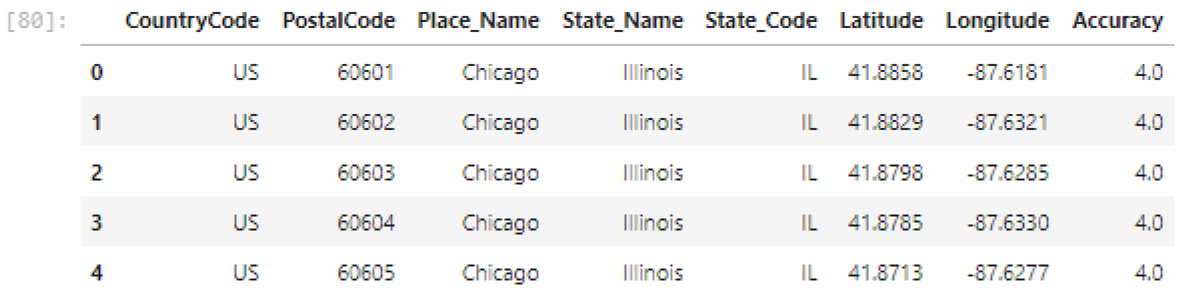
2. Foursquare API Venues endpoint, querying the database for the most popular venues in Chicago, IL and analyze their characteristics

## 2.2. Data cleaning and filtering

First, the Chicago postal codes data, the information was filtered for the Chicago area by the name of the city (Chicago) and code of state (IL).Several postal codes had the same geospatial coordinates so only one record was kept in the dataframe for consistency. The following columns were not used and dropped:

- County Name
- County code
- Community Name

- Community Code



[80]:

| | CountryCode | PostalCode | Place_Name | State_Name | State_Code | Latitude | Longitude | Accuracy |
|---|---|---|---|---|---|---|---|---|
| 0 | US | 60601 | Chicago | Illinois | IL | 41.8858 | -87.6181 | 4.0 |
| 1 | US | 60602 | Chicago | Illinois | IL | 41.8829 | -87.6321 | 4.0 |
| 2 | US | 60603 | Chicago | Illinois | IL | 41.8798 | -87.6285 | 4.0 |
| 3 | US | 60604 | Chicago | Illinois | IL | 41.8785 | -87.6330 | 4.0 |
| 4 | US | 60605 | Chicago | Illinois | IL | 41.8713 | -87.6277 | 4.0 |

Figure 1: Sample data from the Chicago are postal codes dataframe nomi2

In regard to the Foursquare venues information, data was filtered by restaurants, bars and hotels because, in our opinion, these are the locations which are in immediate need for cleaning services and will offer an immediate revenue for the business.

The following venues characteristics were included in analysis:

- Venue name
- Location by longitude and latitude
- Venue category

The following filters were used in querying the data with the Foursquare API:

- A radius of 500 meters from the geospatial coordinates
- A limit of up to 100 most popular venues for each postal code

All remaining postal codes were verified for consistency and for missing information before starting to explore the data.

# 3. Methodology

The data analysis is using different python libraries, but the most important ones are Pandas, Matplotlib and SciKit-Learn.

Pandas library is used for data loading, manipulation and analysis. Indexing, grouping, aggregation and filtering were used, and data reshaping and pivoting was essential for getting to the result.

Scikit-learn library is used to clustering the final data and assure the result is correct and consistent with the displayed information. Machine learning methods were used in this project in order to identify the best locations for opening a cleaning business. The k-means unsupervised learning algorithm was used in order to identify clusters.

The matplotlib library was used to plot the information in the bar graph to visualize the proportions for the top 15 postal codes.

We consider that choosing the top 15 postal codes by number of total venues (restaurants, bars and hotels) will offer the best competitive advantage to an entrepreneur, rather than analyzing the entire data set and clustering by using the whole dataframe information.

After verifying that the loaded data is correct, I applied the steps below for exploring the data, transform it in order to present the results to the audience.

1. Identified the geospatial coordinates of Chicago using Nominatim class in the geopy package to convert address into latitude and longitude address
2. Created the initial Chicago map using the geonames postal codes coordinates. This representation is helping us to visualize the locations.
3. Accessed Foursquare and retrieved using the API the most popular venues for each postal code coordinates. This was done by used specific functions:
    a. get_category_type – extracting category of the venue
    b. getNearbyVenues – return nearby venues for a specified latitude and longitude

The Foursquare API is a powerful tool which offers relevant information in searching for the most important venues. A sample postal code was tested, and information was displayed in order to analyze the data and identify the best features to take into consideration.

4. Created a dataframe for the most important 10 venues of Chicago, by postal code

A total of 2186 venues were returned by the API with the following information:

- Postal Code
- Postal Code Latitude
- Postal Code Longitude
- Venue Name
- Venue Latitude
- Venue Longitude
- Venue Category

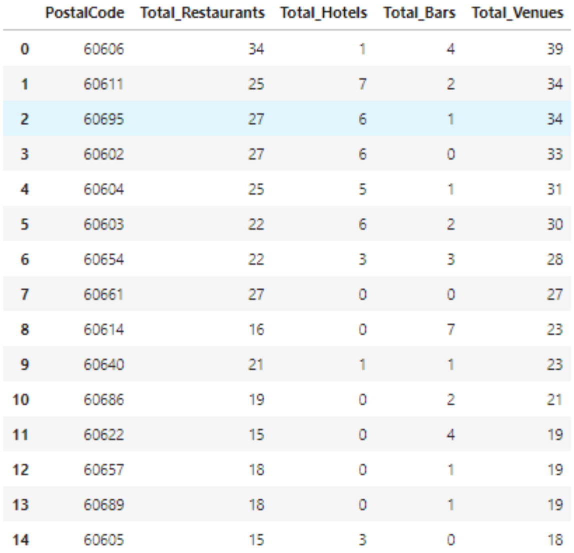
5. From all venues, restaurants, bars and hotels were selected for our analysis, as we considered these would bring immediate revenue to an entrepreneur and these would be in immediate need for cleaning services.
6. A new dataframe was created with totals of venues categories (restaurants, bars, hotels), grouped by postal code
7. At this stage, we have used data visualization by creating a bar chart of proportions for top 15 postal code
8. We have reviewed the top 10 most popular venues for each of the top 15 postal codes and apply k-means clustering to verify our results. For the resulted top 15 postal codes, we have clustered the results based on the similarities of venues categories, using the k-means algorithm from the Scikit-learn library. This step can be repetitive, as we already have chosen the top 15 postal codes with most popular venues as restaurants, bars and hotels, but we want to find out if the postal codes will be clustered respecting the requirements, and also identify the best postal code for opening the business (close location to the others, but visually in the center of the cluster).

We have used the Elbow method and the homogeneity chi-square test in order to verify that all postal codes are part of the same cluster and that our choice for top 15 records by total venues did not contain any significant error.

# 4. Results

The visual representation below offers a concise representation of proportion of venue categories in the most popular 15 zones in the Chicago area.

Table 1: Top 15 postal codes by total number of venues



| | PostalCode | Total_Restaurants | Total_Hotels | Total_Bars | Total_Venues |
|---|---|---|---|---|---|
| **0** | 60606 | 34 | 1 | 4 | 39 |
| **1** | 60611 | 25 | 7 | 2 | 34 |
| **2** | 60695 | 27 | 6 | 1 | 34 |
| **3** | 60602 | 27 | 6 | 0 | 33 |
| **4** | 60604 | 25 | 5 | 1 | 31 |
| **5** | 60603 | 22 | 6 | 2 | 30 |
| **6** | 60654 | 22 | 3 | 3 | 28 |
| **7** | 60661 | 27 | 0 | 0 | 27 |
| **8** | 60614 | 16 | 0 | 7 | 23 |
| **9** | 60640 | 21 | 1 | 1 | 23 |
| **10** | 60686 | 19 | 0 | 2 | 21 |
| **11** | 60622 | 15 | 0 | 4 | 19 |
| **12** | 60657 | 18 | 0 | 1 | 19 |
| **13** | 60689 | 18 | 0 | 1 | 19 |
| **14** | 60605 | 15 | 3 | 0 | 18 |

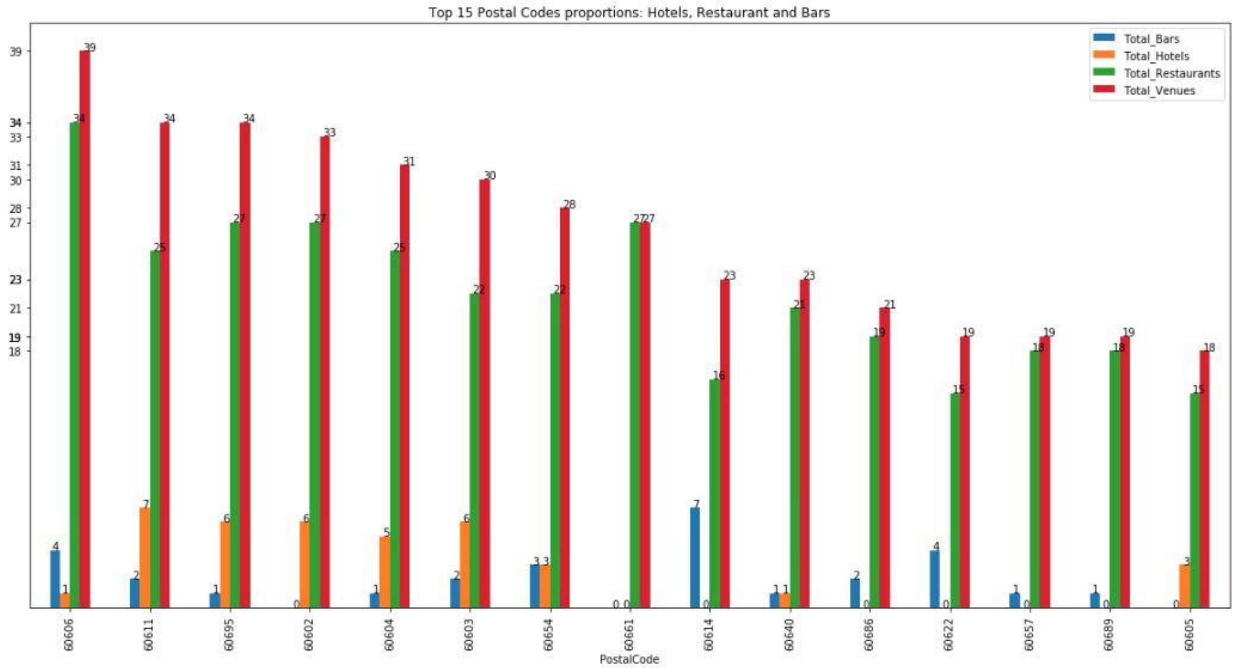


Figure 2: Categories proportions: Hotels, Restaurants, Bars

Postal code 60606 is the top location for opening a cleaning business, as it has 4 bars, 1 hotel and 34 restaurants, totaling 39 venues.

Figure 3: Overview of the 60606 Chicago postal code area

As a possible secondary location, postal code 60611 has a total of 34 venues, with 2 bar, 7 hotels and 25 restaurants. (we consider area 60611 as second place since it has 7 hotels and 2 bars, compared to area 60695).

Figure 4: Overview of the 60611 Chicago postal code area

For each of the top 15 postal codes, the most popular venues are displayed and presented.

The most popular venues are the restaurants, while the bars and hotels proportions differ from location to location.

If we would like to offer laundry services also, the zones 60611, 60695, 60602 60603 and 60604 would be of interest, since they do have most of the hotels.

Zones 60606, 60614 and 60622 have most of the bars from the top 15 areas, and these may require deep cleaning services, such as emptying trash bins, sweep and mop the floor behind the bar, clean refrigerators, remove empty bottles, clean freezers, polishing décor and light fixtures.

All the steps including accessing the data, data cleaning and preparation, processing and results are presented in the Jupyter Notebook Capstone_Project_Final at the link below:

https://github.com/datacrawlers/Coursera_Capstone/blob/master/Final_Project/Capstone_Final_Paper.pdf

# 5. Discussion

One of the advantages of the solution is that, by choosing the top 15 areas by total number of venues – from bars, restaurants and hotels – a cleaning business can expect to generate quick revenues with low costs.

The cleaning industry is very diverse, but for the core services, depending also on clients' needs, and choosing the right location of the point of work where to keep inventory and wait for service requests – it can be very rewarding and efficient for the business.

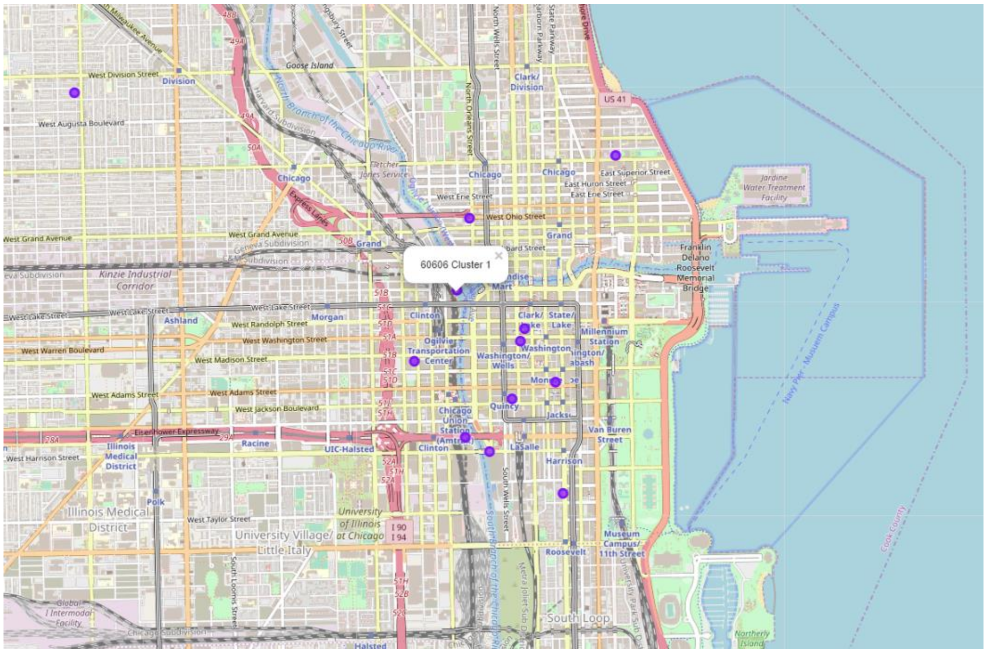


Figure 5: Cluster 1 optimal location 60606

As presented above, area 60606 – is the best for opening a possible cleaning business; not only that it has the most venues, but it is also situated in the close location of the other postal codes; near main transportation nodes, it will have a competitive advantage versus the others.

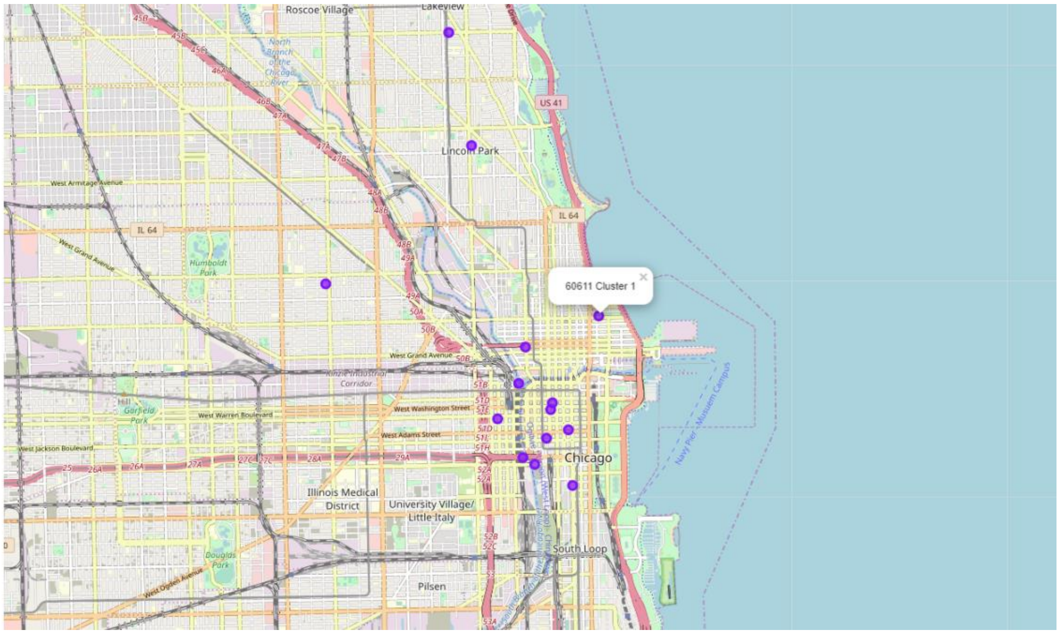
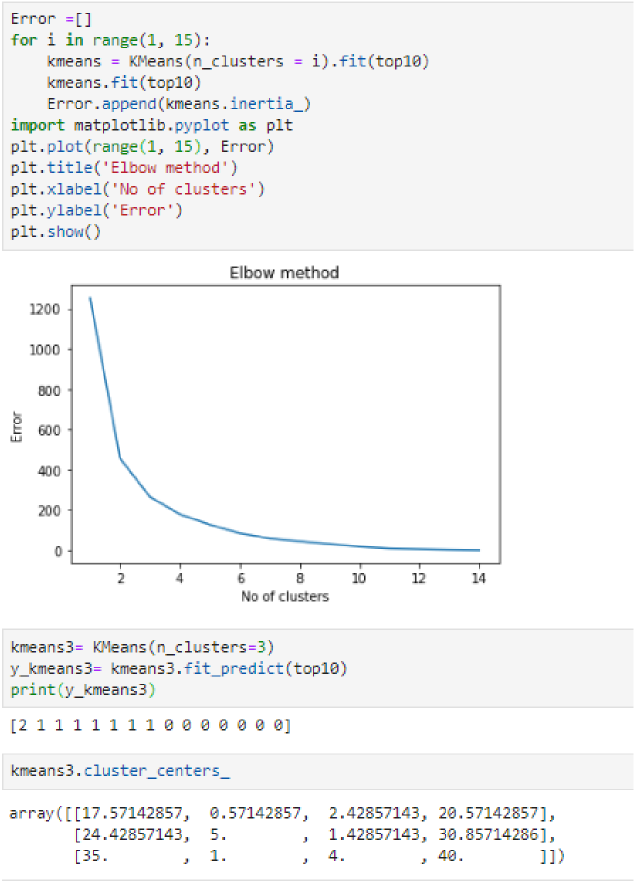


Figure 6: Cluster 1 location 60611 second choice

On the other hand, the second choice for opening 60611 is closed to two areas with similar number of venues, up north Uptown 60640 (23 venues) and Lincoln Park 60614 (23 venues). This location is close to the Chicago Downtown area and it can easily represent a backup location, when needing additional workers or materials.

While trying to cluster the final information, we have noticed that only one cluster is generated with all information, which is usually unusual for the k-means clustering algorithm. In this case, we can assume that the doing additional investigation of the data would offer information on the optimal clusters, either by using the Elbow, Silhouette or the Gap Statistics method.

Note: Even though the Elbow Method identified below the optimal number of clusters to be 3, we can definitely say that the chosen dataset is homogeneous enough in order to result only one cluster, containing all areas.



```
Error =[]
for i in range(1, 15):
    kmeans = KMeans(n_clusters = i).fit(top10)
    kmeans.fit(top10)
    Error.append(kmeans.inertia_)
import matplotlib.pyplot as plt
plt.plot(range(1, 15), Error)
plt.title('Elbow method')
plt.xlabel('No of clusters')
plt.ylabel('Error')
plt.show()
```

```
kmeans3= KMeans(n_clusters=3)
y_kmeans3= kmeans3.fit_predict(top10)
print(y_kmeans3)
```

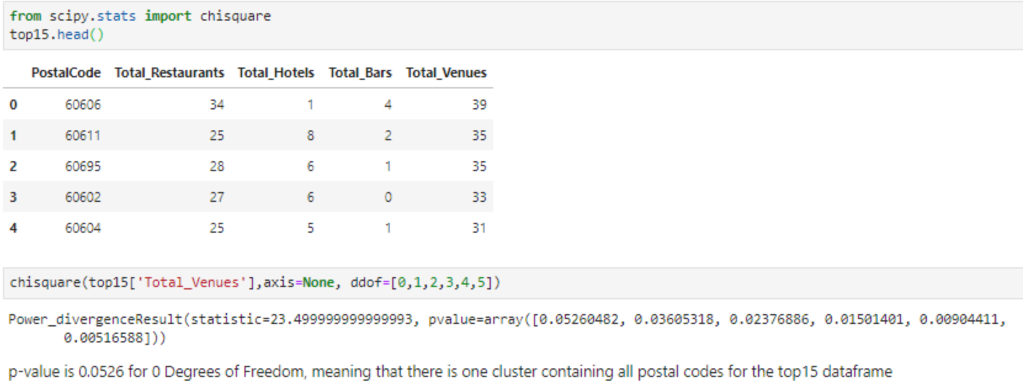
```
[2 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
```

```
kmeans3.cluster_centers_
```

```
array([[17.57142857,  0.57142857,  2.42857143, 20.57142857],
       [24.42857143,  5.        ,  1.42857143, 30.85714286],
       [35.        ,  1.        ,  4.        , 40.        ]])
```

Figure 7: Elbow method

On applying the homogeneity test chi-square, with a significance level of 0.05, we can confirm that all postal codes should be part of the same cluster 1.



```
from scipy.stats import chisquare
top15.head()
```

| | PostalCode | Total_Restaurants | Total_Hotels | Total_Bars | Total_Venues |
|---|---|---|---|---|---|
| **0** | 60606 | 34 | 1 | 4 | 39 |
| **1** | 60611 | 25 | 8 | 2 | 35 |
| **2** | 60695 | 28 | 6 | 1 | 35 |
| **3** | 60602 | 27 | 6 | 0 | 33 |
| **4** | 60604 | 25 | 5 | 1 | 31 |

```
chisquare(top15['Total_Venues'],axis=None, ddof=[0,1,2,3,4,5])
```

```
Power_divergenceResult(statistic=23.499999999999993, pvalue=array([0.05260482, 0.03605318, 0.02376886, 0.01501401, 0.00904411,
       0.00516588]))
```

p-value is 0.0526 for 0 Degrees of Freedom, meaning that there is one cluster containing all postal codes for the top15 dataframe

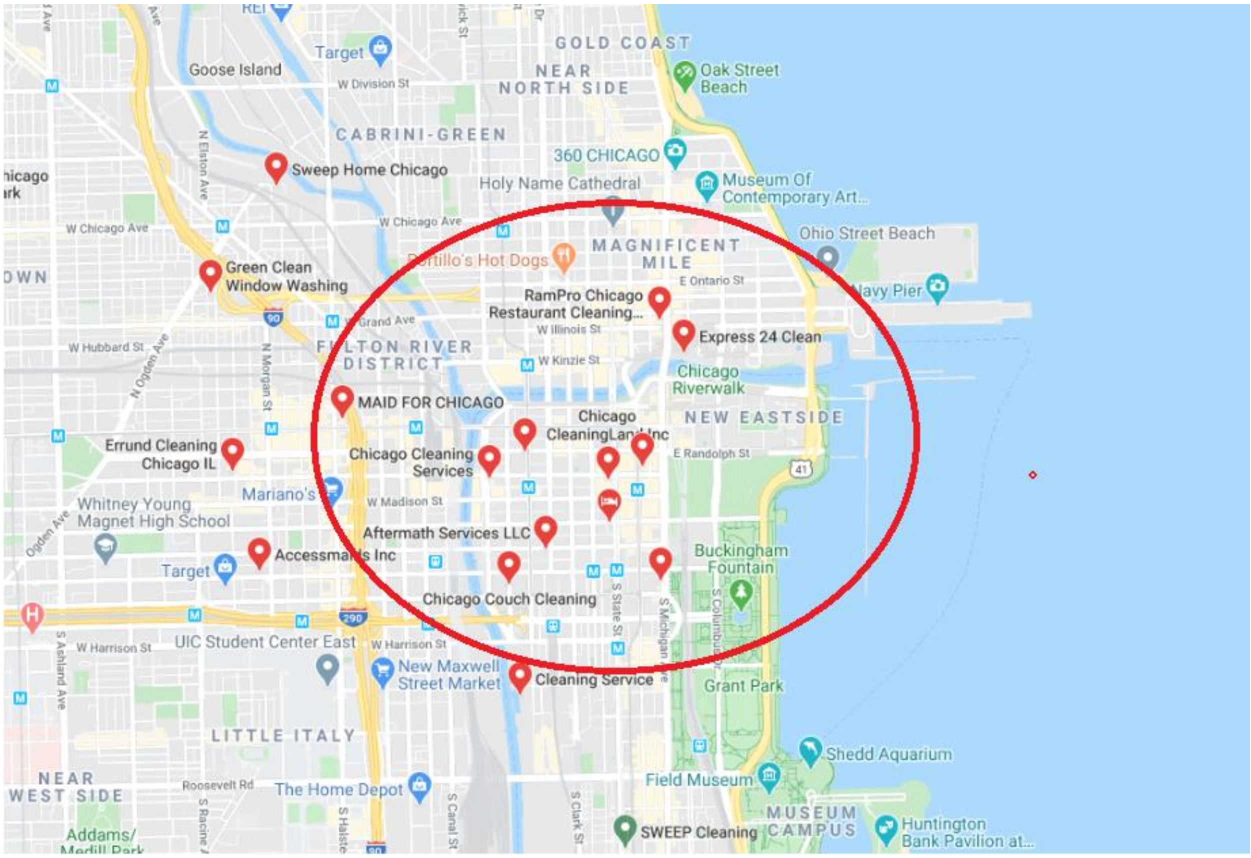
Figure 8: Chi-square method

# 6. Conclusion

We have presented in this report the identification of the best location to open a cleaning business in the Chicago area, by analyzing the Foursquare venue data and concluding our results based on top 15 areas by total number of venues (restaurants, bars, hotels).

I have built a model based on the provided venues information and mapped over the geospatial coordinates of the Chicago postal codes. The information was filtered by the total number of venues and presented in a bar chart and analyzed.

Two main locations were found as best candidates for opening a cleaning business in Chicago, based also on the fact that they are nearby most of the restaurants, hotels and bars from the area.



By doing a quick search on Google, we can find indeed many cleaning services business registered in the downtown area, where 60606 and 60611 are two of the most popular postal codes for existing companies in the domain.

As secondary sources of revenue for the business, laundry services can be provided in areas where hotels are in a high number, and deep clean services can be offered in areas where bars are located.

Future research can include additional services offered to other venue categories, based on costs and distance to potential clients.