Coursera Applied Data Science Capstone

Identifying the best location in Chicago for opening a cleaning business

Dan Grigore

February 2020

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.

Chicago is one of the most hospitable cities for entrepreneurs and startup businesses, contrary to its harsh winters and high competitive environment. .

Problem

With over 300.000 companies based in the city with 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:

- ✓ Turn customers away
- ✓ Make employees more productive
- ✓ Keep equipment and in good condition
- ✓ Work in a Safe environment

Problem

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

- ✓ Must be close to venues in most need of daily cleaning
- ✓ 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
- ✓ Can have other secondary locations, if needed, for helping reducing transportation costs to/from cleaning locations
- ✓ 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.

Interest

Any entrepreneur willing to open a business in the cleaning industry in Chicago

The advantages of opening a cleaning business are:

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

Data and cleaning

Data Sources

1. geonames.org zip file containing postal code information for all cities of The United States.

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

Data cleaning and filtering

1. Geonames: the columns not used were dropped

[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

2. Foursquare: 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.

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
- ✓ Data reshaping and pivoting
- ✓ Machine learning methodology
  - ✓ Scikit-learn library is used to clustering the final data
  - ✓ k-means unsupervised learning algorithm was used in order to identify clusters
  - ✓ Bar graph for data representation
  - ✓ 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.

Results

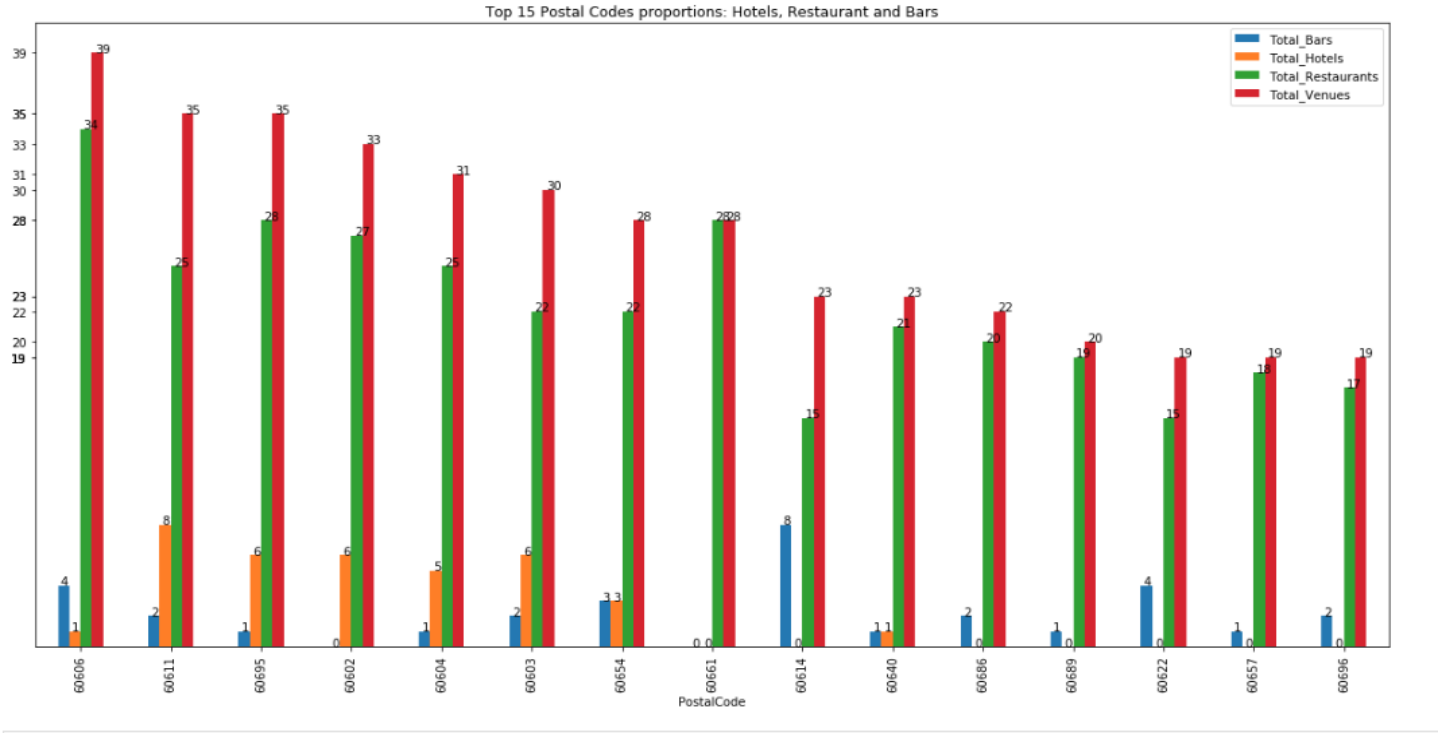
The visual representation below offers a concise representation of proportion of venue categories in the most popular

15 zones in the Chicago area

[24]:

	PostalCode	Total_Restaurants	Total_Hotels	Total_Bars	Total_Venues
0	60606	35	1	4	40
1	60611	25	9	2	36
2	60695	27	6	1	34
3	60602	26	6	0	32
4	60603	22	6	2	30
5	60604	24	5	1	30
6	60654	22	3	3	28
7	60661	25	0	1	26
8	60614	15	0	8	23
9	60640	21	1	1	23
10	60686	20	0	2	22
11	60689	19	0	1	20
12	60622	15	0	4	19
13	60657	18	0	1	19
14	60605	15	3	0	18

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



Top 15 Postal Codes proportions

Results

The Postal code 60606 is the top location for opening a cleaning business,

as it has 4 bars, 1 hotel and 35 restaurants, totaling 40 venues.

As a possible secondary location, postal code 60611 has a total of 36 venues,

with 2 bar, 9 hotels and 25 restaurants.

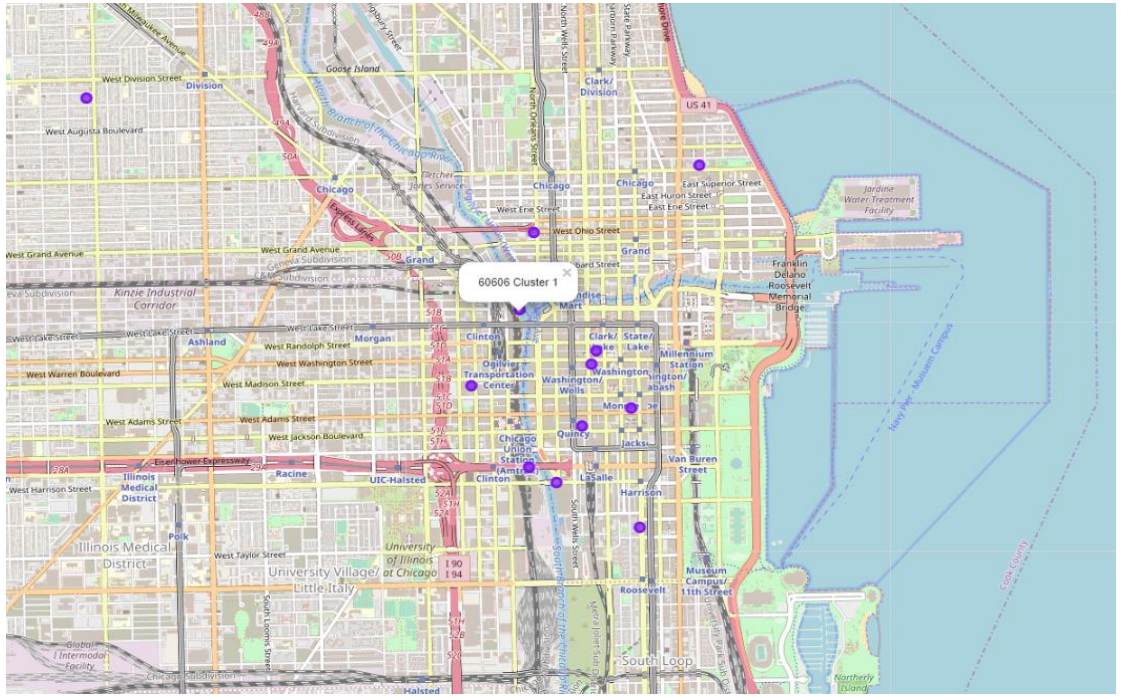
Results

Reviewing the bar graph and table above we can also observe the following:

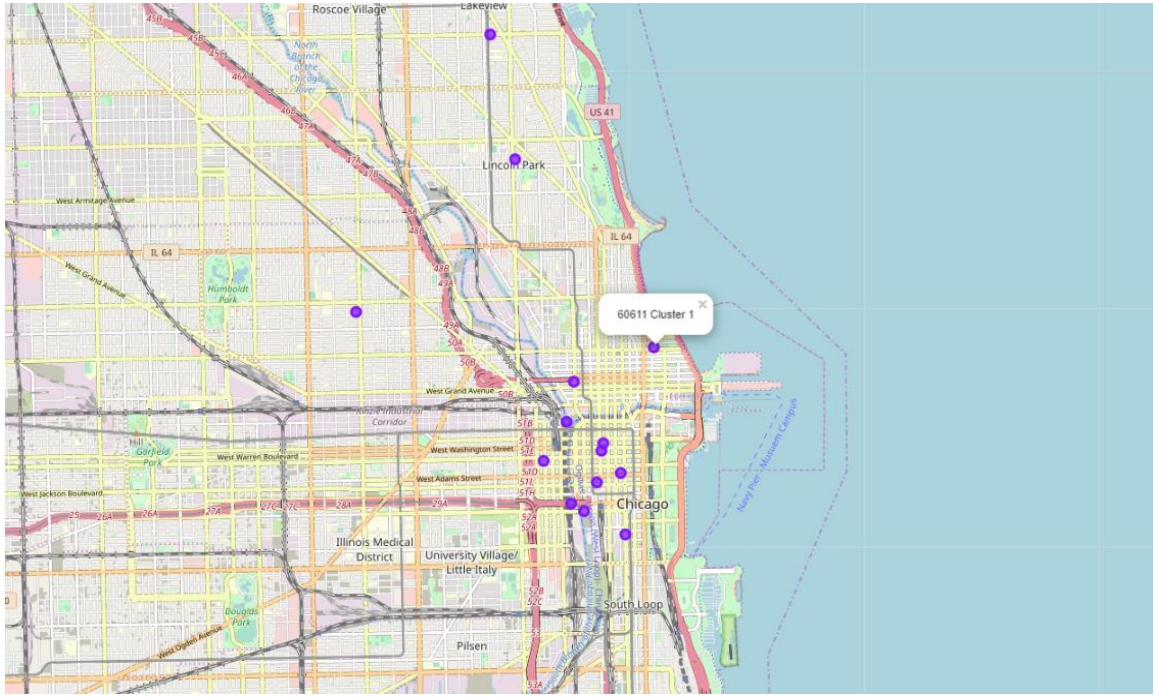
- ✓ The most popular venues are the restaurants, while the bars and hotels proportions differ from location to location.
- ✓ If we would like to offer laundry services also, the zones 60611, 60695, 60602 60603 and 60604 would be of interest, since they do have most of the hotels.
- ✓ Areas 60606, 60614 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.
- ✓ Area 60611 is closed to two other postal codes 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.

Results

Clustering the top 15 postal codes



60606
Primary location



60611
Secondary location

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.

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.

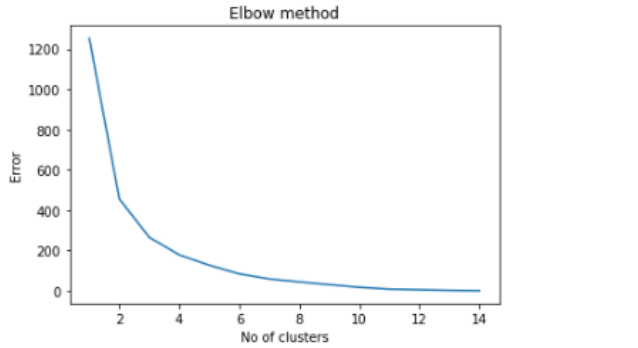
As secondary sources of revenue for the business, laundry services can be provided in areas where hotels are in a high number, and deep clean services can be offered in areas where bars are located.

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

Discussion

Note: Even though the **Elbow Method** identified below the optimal number of clusters to be 3, we can conclude that the chosen dataset is homogeneous, and all data can be included into 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 0]
```

```
kmeans3.cluster_centers_
```

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

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

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).

The built model is 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.

Conclusion

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.