Applied Data Science Capstone Project: Battle of the Neighborhoods

By Magdalena Sapag

February 2, 2019
Santiago, Chile

Table of Contents

1. Introduction

Data Science is an important tool for decision making in almost every business. It allows managers and analysts to take advantage of the data available to make intelligent decisions. The data used for these decisions can come either from within the company and its records or can be obtained from an outside provider. This project will focus specifically on location data made available by Foursquare and the analysis that can be made based on it to help a business grow its profits.

Objective
The objective of this project is to identify the best corners in the city of Santiago, Chile to place on-field sellers for a flower shop named Crocus. According to the company, the average daily sales of a flower shop show seasonality within a week, increasing by 200% on the weekends. Additionally, profits increase to 9 times the average daily profit for special dates, like Valentine's Day or Mother's Day. The company is interested in capturing more of the market in those dates, and to achieve that they plan on placing on-field sellers in 5 corners of certain neighborhoods of Santiago for weekends and special dates.

To decide which corners are better than others, Crocus has provided a list of 30 potential corners of the city that have shown high foot traffic. The company also requested that the choice is made based on nearby complementary products, such as chocolate shops, restaurants and movie theaters, and competitor location. The final criteria for the selection of corners will be explained in detail in the methodology section.

Audience
The specific client for this project is a flower shop named Crocus that has online operations on the city of Santiago and has been in operation for 5 years. The company is a small business; hence the scope of this specific project is small. However, the basic logic and methodology can be scaled to satisfy similar needs for larger companies.

There is a large amount of businesses in the city of Santiago that rely on on-field sales force or promoters that are located on specific streets or corners. The results of this project have the potential to help these types of businesses make decisions that would grow their profits and help them capture a higher percent of the market. Any business interested in deciding the placement of on-field sales force can replicate this model and add to it if necessary, in order to make the decision based on Foursquare location data, or any other location data provider.

2. Data

For the purpose of this project there are two types of data that will be used: The table containing the potential corners and their coordinates and Foursquare location data.

Potential corners coordinates
The company requesting this project has provided a list of 30 potential corners in the city of Santiago, specifically the upper west side of the city, that have high foot traffic. The data set has been delivered as a csv file with 31 rows and 5 columns. The columns of the data set are: an id number for each corner, the address, the neighborhood, the latitude and the longitude of its location. The complete data set can be seen in the following table.

Table 1: Potential Corners in Santiago

Corner	Address	Neighborhood	Latitude	Longitude
1	Plazuela Los Leones	Providencia	-33.419869	-70.605912
2	Avenida Nueva Providencia 2200	Providencia	-33.422751	-70.60956

3	Avenida Pedro de Valdivia 101	Providencia	-33.424649	-70.611985
4	Avenida Suecia 780	Providencia	-33.427532	-70.605505
5	Avenida El Bosque 963	Providencia	-33.428427	-70.596128
6	Avenida Nueva Providencia 1398	Providencia	-33.428874	-70.618359
7	Metro Francisco Bilbao	Providencia	-33.430808	-70.586816
8	Latadia 4141	Las Condes	-33.431166	-70.578576
9	Avenida Ossa 1552	Ñuñoa	-33.439474	-70.572611
10	El Alcalde 15	Las Condes	-33.416337	-70.595227
11	Avenida Apoquindo 3898	Las Condes	-33.415298	-70.590077
12	Avenida Apoquindo 4483	Las Condes	-33.413435	-70.583082
13	Centro Comercial Apumanque	Las Condes	-33.409423	-70.567761
14	Parroquial Los Dominicos	Las Condes	-33.407847	-70.541797
15	Avenida Presidente Kennedy 5413	Las Condes	-33.402279	-70.578125
16	Avenida Cuarto Centenario 1001	Las Condes	-33.417755	-70.55362
17	Avenida Padre Hurtado Sur 875	Las Condes	-33.416215	-70.53963
18	Avenida Francisco Bilbao 8464	Las Condes	-33.429683	-70.54508
19	Avenida Larrain 5862	La Reina	-33.453282	-70.570271
20	Salvador Izquierdo 1777	La Reina	-33.438958	-70.556324
21	Plaza Pedro de Valdivia 1731	Providencia	-33.438743	-70.608079
22	Santa Isabel 400	Providencia	-33.446353	-70.626189
23	Avenida Salvador 42	Providencia	-33.433757	-70.626389
24	Avenida Ricardo Lyon 1146	Providencia	-33.430869	-70.606272
25	Avenida Suecia 181	Providencia	-33.418879	-70.609346
26	Avenida Cristobal Colon 4431	Las Condes	-33.423352	-70.578923
27	Parque Bicentenario	Vitacura	-33.399924	-70.602398
28	Avenida Vitacura 3744	Vitacura	-33.402006	-70.594212
29	Mall Alto Las Condes	Las Condes	-33.390066	-70.547686
30	Manuel de Salas 39	Ñuñoa	-33.454969	-70.593477

Foursquare location data

Foursquare is a provider of accurate location data, currently used by over 100,000 developers, including services like Apple Maps, Uber, Twitter, among others. Foursquare allows access to information about venues and users in different locations. Each venue has different attributes, such as name, category, address, working hours, menu, tips and images. There is also the possibility of getting information about the users, such as name, profile picture, tips and images she has uploaded.

For the purpose of this project only information about venues will be taken into consideration and for each venue, only the category and exact location will matter. An example of how Foursquare categories work can be seen bellow, where a search for Flower Shops in Santiago is carried out and the results given by Foursquare are shown.
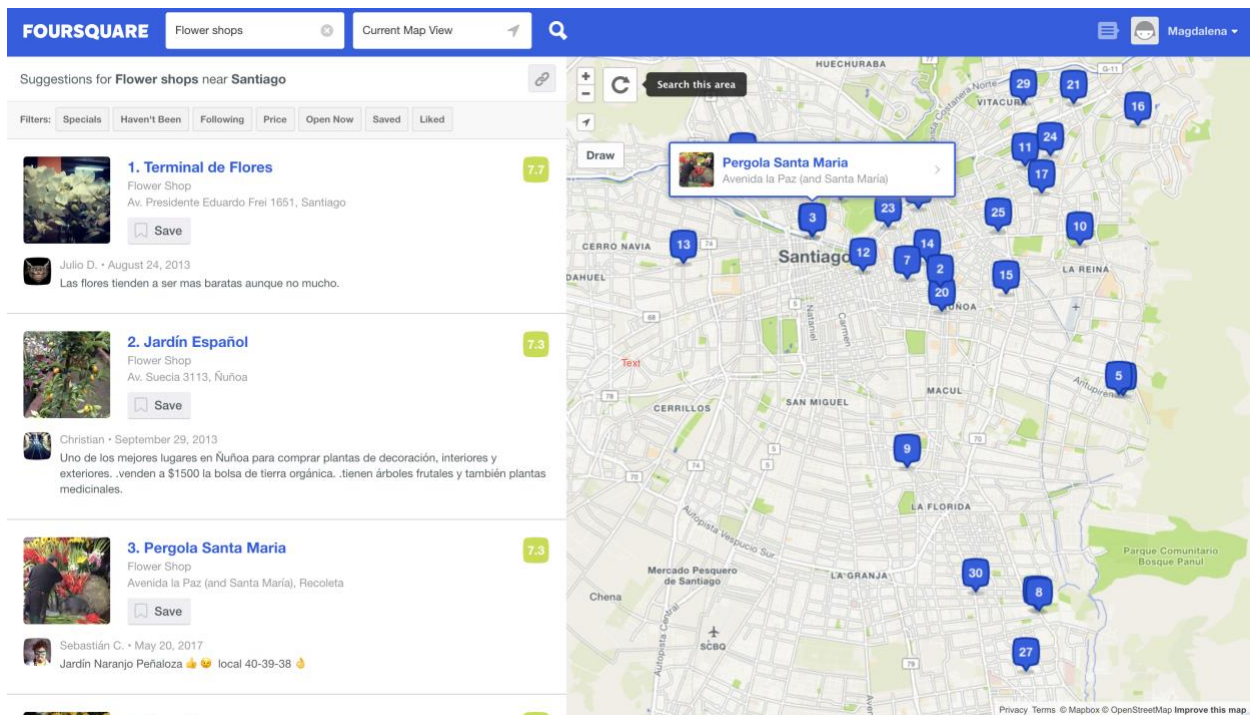
Figure 1 Foursquare search for flower shops in Santiago

Having an account in Foursquare will provide credentials which allow Foursquare information about venues to be accessed through a Python script. This will be the second main source of information for the project, which will have to be cleaned and processed first, in order to use it for the analysis required.

Specifically, the information that will be used for this project is obtained through a search request in Foursquare. For the request a query will be given to find each type of venues, such as 'Flower Shop' or 'Movie Theater'. Another argument that will be used in order to narrow the search is categoryID, which allows the user to give a specific category in which the result must be, for example the category ID for flower shops is '4bf58dd8d48988d11b951735', so that will be given as an argument when the request for flower shops is made. The result of this request is a json file with all the venues that were found such as the one shown in the following figure, this example comes from a different project in which the search was made in New York for Italian food.

```
{'meta': {'code': 200, 'requestId': '5c3e51aa4c1f671cc4a58cc8'},
 'response': {'venues': [{'id': '4fa862b3e4b0ebff2f749f06',
    'name': "Harry's Italian Pizza Bar",
    'location': {'address': '225 Murray St',
     'lat': 40.71521779064671,
     'lng': -74.01473940209351,
     'labeledLatLngs': [{'label': 'display',
       'lat': 40.71521779064671,
       'lng': -74.01473940209351}],
     'distance': 58,
     'postalCode': '10282',
```

Figure 2 Example of json file obtained from a search request in Foursquare

This json file will be transformed into a data frame using the keys for name and category of the resulting venues. No other key will be used since the final analysis only requires a count of resulting venues for each category. The data frame obtained for this project will have 3 columns: Corner, Venue and Venue Category. The resulting data frame will look like the example shown in the next figure, in this example the data frame shows the result for the search for movie theaters for different corners in Santiago.

	Corner	Venue	Venue Category
0	1	Cineplanet	Multiplex
1	1	Sala Prime Cineplanet	Multiplex
2	1	Cine Planet Sala 3D	Movie Theater
3	1	Cine Arte Tobalaba	Movie Theater
4	1	pasillo 6-8 Cine Hoyts Maipú	Multiplex
5	1	Cineplanet Costanera Center	Movie Theater
6	2	pasillo 6-8 Cine Hoyts Maipú	Multiplex
7	7	Sale De Cine Chack	Indie Movie Theater
8	13	cine Edificio Urbano Plus	Indie Movie Theater
9	14	Cine tiempo	Movie Theater
10	15	Cine Hoyts	Multiplex
11	15	Cine Hoyts Sala 4	Movie Theater
12	15	Sala 6 - Cine Hoyts	Movie Theater
13	15	Cine Hoyt Sala 4 DX	Movie Theater
14	15	Cine Hoyts Sala 10	Multiplex
15	15	Hoyts Premium Class	Multiplex

Figure 3 Example for data frame resulting from Foursquare search

For this project a search will be carried out for the following types of venues: Flower shop, chocolate shop, candy shop (since some chocolate shops are under this category), restaurants and movie theaters. Each request will result in a different data frame. The 5 data frames will then be combined to carry out the analysis. These steps will be explained in the next section of the report.

3. Methodology

In order to do the required analysis and select the 5 corners, a series of 3 main steps must be followed:

1. Extract the information
2. Process the information to count number of venues by category on each corner
3. Sort corners according to criteria

Every step taken and the python code used in the notebook will be explained in this section, as will be the criteria used for decision making at the end.

1. Extract the information
The data comes from two sources as was explained above. First, there is a csv file containing every corner, their addresses and coordinates. Second, there is the location data from Foursquare. The logical sequence is to first import the data from the csv file, since then it will be used to search for the venues in Foursquare.
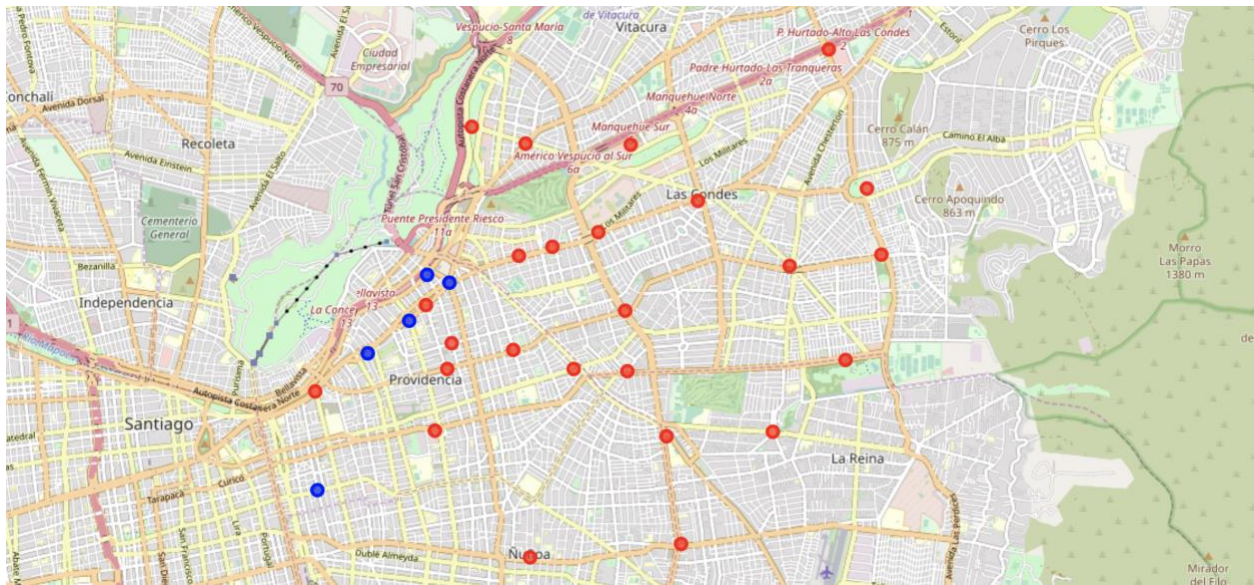
CSV File
In order to get the information from the csv file, the pandas library needs to be imported into the notebook. Then, the pandas.read_csv() function can be used. When looking at the notebook on github with the code, refer to inputs 1 and 2 for the exact code. The resulting data frame contains the following columns: Corner, Address, Neighborhood, Latitude, Longitude. The last two columns will then be used to search Foursquare information for venues relevant to the analysis.

Foursquare location data
The first step to access Foursquare data is to define the credentials that will be used, this includes client id, client secret and foursquare version. It is also useful to define at the beginning the radius for the search

and the limit of maximum requests made. For the purpose of this project, the radius was set to 300m, since the business is looking to attract people who stumble upon them near either a complimentary product, such as chocolate, or a final destination such as a restaurant or movie theater. The company considered that 300m is the maximum distance a person would

4. Results

5. Discussion

6. Conclusion