Suggestions for TOP venues in Lisbon!

Final Report - Table of Contents

1. Introduction
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

1. Introduction

Hello everyone!

I am Solange, a citizen from Lisbon, in Portugal.

As you may know, Lisbon is blooming with tourists.

This is a major opportunity for many business owners and entrepreneurs, mostly in the **food & leisure industry**. Nonetheless, without proper planing and market insights, some of them don't succeeded, close their business or even go bankrupt.

This situation raises a huge economic and social concern. Luckily there are great examples that we can learn from, such as New York city and Toronto. Being one of the top destinations for tourists worldwide since years, they still provide diversity and quality for all tastes.

Therefore this project aims to provide insights from these two cities, which will definitely aid **business owners and all the stakeholders**, including local **authorities and the citizens**, to succeed and hence, make **Lisbon a great destination** for all.

2. Data

2.1 Sources and format

To explore the two cities we start by creating two data frames which include at least Borough and Neighborhoods information. New York (NY) data is available as a json file in the following link. It includes features such as Borough, Neighborhood, Latitude and Longitude. https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

```
[113]: neighborhoods_data = newyork_data['features']
       neighborhoods_data

[113]: [{'type': 'Feature',
         'id': 'nyu_2451_34572.1',
         'geometry': {'type': 'Point',
          'coordinates': [-73.84720052054902, 40.8947051766 1]},
         'geometry_name': 'geom',
         'properties': {'name': 'Wakefield',
          'stacked': 1,
          'annoline1': 'Wakefield',
          'annoline2': None,
          'annoline3': None,
          'annoangle': 0.0,
          'borough': 'Bronx',
          'bbox': [-73.84720052054902,
           40.89470517661,
           -73.84720052054902,
           40.89470517661]}},
        {'type': 'Feature',
         'id': 'nyu_2451_34572.2',
         'geometry': {'type': 'Point',
          'coordinates': [-73.82993910812398, 40.87429419303012]},
         'geometry_name': 'geom',
         'properties': {'name': 'Co-op City',
          'stacked': 2,
          'annoline1': 'Co-op',
          'annoline2': 'City',
          'annoline3': None,
          'annoangle': 0.0,
          'borough': 'Bronx',
          'bbox': [-73.82993910812398,
           40.87429419303012,
           -73.82993910812398,
           40.87429419303012]}},
        {'type': 'Feature',
```

For Toronto, the data is in a table on the Wikipedia page url. The BeautifulSoup was selected among the different website scraping libraries and packages in Python, to transform the data in the table on the Wikipedia page into a pandas dataframe. It includes the PostalCode, the Borough and Neighborhood.

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF

Article Talk

List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are local the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartp validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directo

Toronto - FSAs [edit]

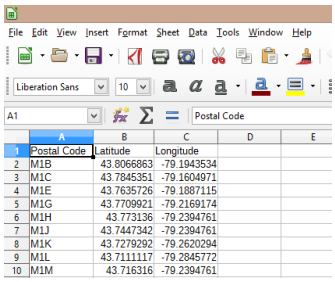
Note: There are no rural FSAs in Toronto, hence no postal codes start with M0.

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue
M1B	Scarborough	Rouge

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The latitude (lat) and longitudes (long) will be added directly from url:

http://cocl.us/Geospatial_data.



	A	B	C	D	E
1	Postal Code	Latitude	Longitude		
2	M1B	43.8066863	-79.1943534		
3	M1C	43.7845351	-79.1604971		
4	M1E	43.7635726	-79.1887115		
5	M1G	43.7709921	-79.2169174		
6	M1H	43.773136	-79.2394761		
7	M1J	43.7447342	-79.2394761		
8	M1K	43.7279292	-79.2620294		
9	M1L	43.7111117	-79.2845772		
10	M1M	43.716316	-79.2394761		

2.2 Dependencies

Several libraries are required for data handling and analysis, such as pandas, numpy and json; for cluster analysis, kmeans from sklearn will be imported as well as matplotlib, requests and geocoders to map the reults.

3. Methodology

In this project we will convert NY and Toronto addresses into their equivalent lat and long values. Then, use the Foursquare API to explore neighborhoods in both cities to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. The machine learning algorithm k-means clustering will be especially useful to quickly segment the neighborhoods based on the top venues categories. In parallel with the Folium library we will visualize the neighborhoods in both cities and their emerging clusters.

3.1 Download and create New York city dataset

In order to explore NY city top venues, first we need a dataset that essentially contains the neighborhoods and respective coordinates. So, with the 'wget' command will load and the data and then loop to extract the borough, neighborhood, lat and long data from the 'Features' in the json file (figure on the right).

Finaly, a pandas data frame will be created with the Borough, the Neighborhood and their lat and long coordinates (Preview displayed in the picture bellow).

```
neighborhoods.head()
```

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

3.2 Download and create Toronto dataset

Toronto data is in a wikipedia page, so we used the BeautifulSoup package to transform the table in the Wikipedia page into a pandas dataframe. Then, cells with a borough that is Not assigned were ignored. Rows with duplicate PostalCode were grouped and their Neighborhoods separated with a comma. Then to get their lat and the long coordinates this dataset was merged with Geospatial data.

```
df_m=pd.merge(df1, c, on='PostalCode', how='inner')
print('shape',df_m.shape)
df_m.head()
```

```
shape (103, 5)
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

3.3 Explore Neighborhoods

In order to search for the top venues in Foursquare, one needs to convert New York city and Toronto addresses into their equivalent lat and long with geolocator, using the wget command.

Given the extension of NY city we will explore the top venues only in **Manhattan** neighborhoods. Given it's multi-cultural population, this Borough area should cover enough diversity to get a snapshot of the different types of venues and hence really interesting results.

So first we'll slice the original data frame and create a new data frame that focus on the neighborhoods in Manhattan.

Similarly, in Toronto, we create a new data frame which focus only in Boroughs that contain "Toronto" in their names.

With this filter we are able to simplify the search, and still make it representative of the diversity of both cities considering that approximately 40 neighborhoods is a reasonable number to explore.

So now we will use the Foursquare API to explore the neighborhoods and segment them based on the top venues.

3.4 Top venues per neighborhood

The function get request was created to extract only relevant data for each venue (name, lat, long and category) within 500m search radius and a limit of 100 venues. So applying this function to the Manhattan and Toronto dataset, we will create a data frame with the neighborhoods data together with the explore data (venue name, lat, long and category).

Then we can determine the frequency of occurrence of each category grouped by neighborhood and with a function to sort the venues in descending order we will create a data frame with the top 3 venues in Manhattan and Toronto.

3,5 Cluster Neighborhoods

Finally we cluster the neighborhoods to find the top venues in each cluster. The top 3 venues per neighborhood will be grouped with k-means clustering algorithm in a final data frame. The parameters for the number of clusters (kclusters) was 3.

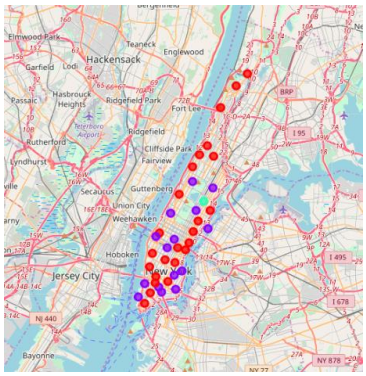
4. Results - Examine Clusters

NY city is spread across 5 Boroughs and 306 neighborhoods. Within Manhattan Borough, where there are 40 neighborhoods, in a search radius of 500m, we found 3313 venues, belonging to 332 categories.

For Toronto there are 11 Boroughs and 103 neighborhoods. Four out of those 11 Borough's have "Toronto" in the name, and comprise 38 neighborhoods. In the same search radius, 500m, we found 1707 venues spread in 238 unique categories.

After filtering, a reasonable and comparable number of neighborhoods remained in both cities, as mentioned, approximately 40.

Manhattan Cluster Map



Toronto Cluster Map

In Manhattan there are 24 Neighborhoods in the 1st cluster, 15 in the 2nd 1 in the 3rd.

In Toronto there are 2 Neighborhoods in the 1st cluster, 35 in the 2nd 1 in the 3rd.

After careful examination of each neighborhood we determine the common and the discriminating venue categories that distinguish each cluster.

To facilitate the interpretation a data frame with the top 3 venues per cluster was created and combined in the end to get the TOP 3.

Manhattan Clusters

The next 3 tables represent the top 3 most common categories for each cluster.

Cluster 1	1st most common	2nd most common	3rd most common	Total
Italian Restaurant	9.0	2.0	2.0	13.0
Coffee Shop	5.0	2.0	2.0	9.0
Sushi Restaurant	0.0	3.0	1.0	4

Cluster 2	1st most common	2nd most common	3rd most common	Total
Coffee Shop	4.0	1.0	0.0	5.0
Park	1.0	1.0	1.0	3.0
Café	0.0	2.0	1.0	3.0

Cluster 3	1st most common	2nd most common	3rd most common	Total
Coffee Shop	0.0	0.0	1.0	1.0
Cosmetics Shop	0.0	1.0	0.0	1.0
Pizza Place	1.0	0.0	0.0	1.0

Combining this nformation we finaly get the top 3 categories. depicted in this final Manhattan_top3 data frame:

Manhattan TOP 3	Total
Italian Restaurant	16.0
Coffee Shop	15.0
Café	6.0

Toronto Clusters

The next 3 tables represent the top 3 most common categories for each cluster.

Cluster 1	1st most common	2nd most common	3rd most common	Total
Airport Lounge	1.0	0.0	0.0	1.0
Airport Service	0.0	0.0	1.0	1.0
Airport Terminal	0.0	1.0	0.0	1.0

Cluster 2	1st most common	2nd most common	3rd most common	Total
Park	3.0	2.0	2.0	7.0
Coffee Shop	3.0	3.0	1.0	7.0
Café	2.0	2.0	2.0	6.0

Cluster 3	1st most common	2nd most common	3rd most common	Total
Café	0.0	0.0	1.0	1.0
Coffee Shop	1.0	0.0	0.0	1.0
Restaurant	0.0	1.0	0.0	1.0

Combining this nformation we finaly get the top 3 categories. depicted in this final Toronto_top3 data frame:

Toronto TOP 3	Total
Coffee Shop	9
Cafe	8
Park	7

Finally, we can combine the information from both cities to get the ultimate top 3 categories:

TOP 3
Italian Restaurant
Coffee Shop/Café
Park

5. Discussion

More than 5000 venues spread in a 500m radius in Manhattan and Toronto, were explored to determine the most frequent categories and hence those among more than 300 categories with highest probability to succeed in Lisbon.

The request was limited to 100 venues per neighborhood and among the different categories there were Pharmacy, coffees, restaurants,bars, parks and stores.

The most frequent venue categories were selected for each of the nearly 40 neighborhoods in each city dataset.

The cluster analysis allowed to merge neighborhoods based on their venues. This way we were able to understand which venues were common within and between clusters.

The cluster number 3 in both cities had a single neighborhood. Still we kept those clusters.

In Manhattan there were 24 Neighborhoods in the 1st cluster, 15 in the 2nd 1 in the 3rd. The top categories between those clusters were Coffee Shop followed by Italian Restaurant and Café.

In Toronto there were 2 Neighborhoods in the 1st cluster, 35 in the 2nd 1 in the 3rd and the top categories were again Coffee Shopand Café and a new one, the Park.

Therefore, the TOP venues are mostly related to food and drinks, Italien restauranta, Coffes shop and Café. Luckily Park was also on the TOP list.

After combining the information from both cities, the bottom categories from the list included mostly services such as stores, markets, pool, pharmacy. There were also venues which are less frequent given their specificity, such as airport lounge, bus line and so on.

Interestingly the food industry was also in the bottom of the most frequent, and included specific cuisines, such as African, Vegan, New American and Thai food. Therefore, maybe it is better to deeply understand the local market and their interests before diving in one of these businesses.

6. Conclusion

The aim of this project was to provide a list with the top venues in two of the number one destination for travelers around the world, New York and Toronto.

With the list we hope to provide some guidance to entrepreneurs, business owners, the government and all the stakeholders, so that they can make informed decision which will ultimately lead them to success, improving the touristic experience with top offers and also benefit Lisbon citizens.

The top category venues were Italian Restaurant, Coffee Shop/Café and Park.

No wonder there are so many sayings about coffee and food.I leave you with one of my favorite even though it is the end!

Aknowledgements

Thank you for reading my final assignment!

This report was created by Solange Pacheco, Lisbon, Portugal.

This work is part of a course on Coursera called Applied Data Science Capstone.