A

Project Report

On

Capstone Project – Data Analysis: Snacks on Wheels - A mobile snack shop

Submitted by

Binesh Jose PhD*

Table of content

1. Introduction / Business Problem:

An investor in New York would like to start a new business “Snacks On Wheels”, a mobile snack shop on trailer which supply Coffee, Ice Cream and other Snacks on various locations within city of New York and its neighborhood. The Snacks on Wheels will park at different locations in New York and provide its services to customers depend on time to time demands.

This business need a data analysis support to identify the best sales spot based on most trending venue of this hour so that the sales can be maximized. Also the trailer need to move different location depends on demand, even though such moment results some additional set up expenses. The business owner likes to have following data so that he can select the next sales spot.

- Top 10 Most Trending Venue
- Nearest (min distance) from current sales spots
- Venue Rating
- Number of tips / tips details
- Consulting/contacting user who gave most valuable tip (if needed)
- Other popular spots near the venue (Explore the location)

The business owners also have some additional requirements including

- Exclude restaurants from trending venues, because they consider restaurants are the competitor for their business.
- Interested only in trending venues within next 2000 meters
- Also need details of popular spots near by

2. Data Used

The “Snacks On Wheels” project need location data , trending venues nearby information , Venue Rating , tips details , User data (who gave tips) , and nearby popular spot data for its analysis.

For the New York neighborhood data New York University spatial data repository is used. This raw data scraped, wrangle, clean, and then read it into a *pandas* dataframe so that it is will be in a structured format for further analysis.

Foursquare API is used to get the trending venues nearby information. These are venues that have the highest foot traffic when the call to the database is made. Therefore the results vary depending on when the call is made. So in the morning we might find that trending venues are coffee shops or office spaces whereas in the evening trending venues can be malls, museums or parks. And to get the trending venues, we simply use the trending endpoint along with the credentials and the latitude and the longitude coordinates of the place of interest. When we make the call to the foursquare database, we get a JSON file of the trending venues that are nearby. In the JSON file, for each trending venue, we get mostly its name, unique ID, location, and category.

For further analysis of trending venue, we used Foursquare nearby API calls which give venue rating and number of tips. Tips again further analyzed with an API call using keyword tips. Details of the user also obtained fourscore followed by explore popular spots near by.

3. Methodology

New York neighborhood data obtained from New York University spatial data repository.

- Source: geo.nyu.edu/catalog/nyu_2451_34572
- After preprocess the data , it loaded to a data frame.
- Then created a new dataframe with only required fields ('Borough', 'Neighborhood', 'Latitude', 'Longitude')
- The data is further limited to borough = Manhattan

Once the current location (sales spot) information of "Snacks on wheel trailer" is available (User will input this data or use GPS) , it used for the "trending venues nearby" analysis. For the testing of project , we randomly selected current location.

Foursquare API is used to get the trending venues nearby information. These are venues that have the highest foot traffic when the call to the database is made. Therefore the results vary depending on when the call is made. So in the morning we might find that trending venues are coffee shops or office spaces whereas in the evening trending venues can be malls, museums or parks. And to get the trending venues, we simply use the trending endpoint along with the credentials and the latitude and the longitude coordinates of the place of interest. When we make the call to the foursquare database, we get a JSON file of the trending venues that are nearby. In the JSON file, for each trending venue, we get mostly its name, unique ID, location, and category.

The Number of results to return is limited by top 10, with in the radius of 2000 meters. The distance between current location and trending location is calculated for further analysis. If distance is not available in Foursquare response we uses the 'haversine' formula to calculate the great-circle distance between two points – that is, the shortest distance over the earth's surface – giving an 'as-the-crow-flies' distance between the points (ignoring any hills they fly over, of course!).

Haversine formula: $a = \sin^2(\Delta\varphi/2) + \cos\varphi1 \cdot \cos\varphi2 \cdot \sin^2(\Delta\lambda/2)$

$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$

$d = R \cdot c$

Where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km);

Note that angles need to be in radians to pass to trig functions!

Foursquare Response sample:

{ "meta": { "code": 200, "requestId": "5ac51dde351e3d4df64064f8" }, "response": { "venues": [{ "id": "5735dc3f498e1ac6a088f324", "name": "Union Fare", "location": { "address": "5 E 17th St", "crossStreet": "btwn 5th Ave & Union Sq W", "lat": 40.737697, "lng": -73.991402, "labeledLatLngs": [{ "label": "display", "lat": 40.737697, "lng": -73.991402 }], "distance": 1802, "postalCode": "10003", "cc": "US", "city": "New York", "state": "NY", "country": "United States", "formattedAddress": ["5 E 17th St (btwn 5th Ave & Union Sq W)", "New York, NY 10003", "United States"] }, "categories": [{ "id": "4bf58dd8d48988d157941735", "name": "New American Restaurant", "pluralName": "New American Restaurants", "shortName": "New American", "icon": { "prefix": "https://ss3.4sqi.net/img/categories_v2/food/newamerican_", "suffix": ".png" }, "primary": **true** }] }] } }

Once the trending venue detail is available, after removing the venues which fall under restaurant category, it used for finding the most nearby trending location based on statistical minimum distance calculation

Initial preferred Venue = Minimum [Trending Data. Distance] ……. (1)

Initial preferred venue is further analyzed for venue rating and number of tips using following URLs

'https://api.foursquare.com/v2/venues/{}/tips?client_id={}&client_secret={}&v={}&limit={}'

It's also required to find the details of user who gave the top rated tips so that the business owner have an option to contact such user if need arise. Finally this work explore the near by location of preferred venue using following URL.

https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}

4. Result

The raw data from New York University spatial data repository scraped, wrangle, cleaned , and then read it into a *pandas* dataframe as follows.


```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 '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]}}
```


Figure 1: Sample data format

This data frame is combined with Geospatial Data and obtained the final data set.

	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

Figure 2: Final data set

For testing purpose, we randomly selected a current location Geospatial Data from the final data set and send to foursquare API.

```
cur_loc=manhattan_data ['Neighborhood'].sample()
latitude=manhattan_data['Latitude'].loc[manhattan_data['Neighborhood'].isin(cur_loc)] .values[0]
longitude=manhattan_data['Longitude'].loc[manhattan_data['Neighborhood'].isin(cur_loc)].values[0]
cur_Borough=manhattan_data['Borough'].loc[manhattan_data['Neighborhood'].isin(cur_loc)].values[0]
print(latitude, longitude)
print (cur_Borough)
```

```
40.77352888942166 -73.98533777001262
Manhattan
```

Figure 3: Random selection of current location

The return result from Foursquare further filtered to satisfy the criteria of exclude "restaurants" from trending venue data set.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng
0	sweetgreen	Salad Place	100 Kenmare St	US	New York	United States	btwn Cleveland & Mulberry St	664	[100 Kenmare St (btwn Cleveland & Mulberry St)...	[{'label': 'display', 'lat': 40.7211844, 'lng'...	40.721184	-73.997111
1	Brooklyn Bridge	Bridge	Brooklyn Bridge	US	New York	United States	NaN	1093	[Brooklyn Bridge, New York, NY 10038, United S...	[{'label': 'display', 'lat': 40.70596749893083...	40.705967	-73.996707
2	Whole Foods Market	Grocery Store	270 Greenwich Street	US	New York	United States	at Warren St	1538	[270 Greenwich Street (at Warren St), New York...	NaN	40.715877	-74.012514

Figure 4: Raw data – Trending Venues

The result presented to business owner as table and map.

**********TRENDING VENUS **********

	name	categories	address	city	distance	lat	lng
0	sweetgreen	Salad Place	100 Kenmare St	New York	1940	40.721184	-73.997111
1	Brooklyn Bridge	Bridge	Brooklyn Bridge	New York	1184	40.705967	-73.996707
2	Whole Foods Market	Grocery Store	270 Greenwich Street	New York	988	40.715877	-74.012514

Figure 5: Trending Venues

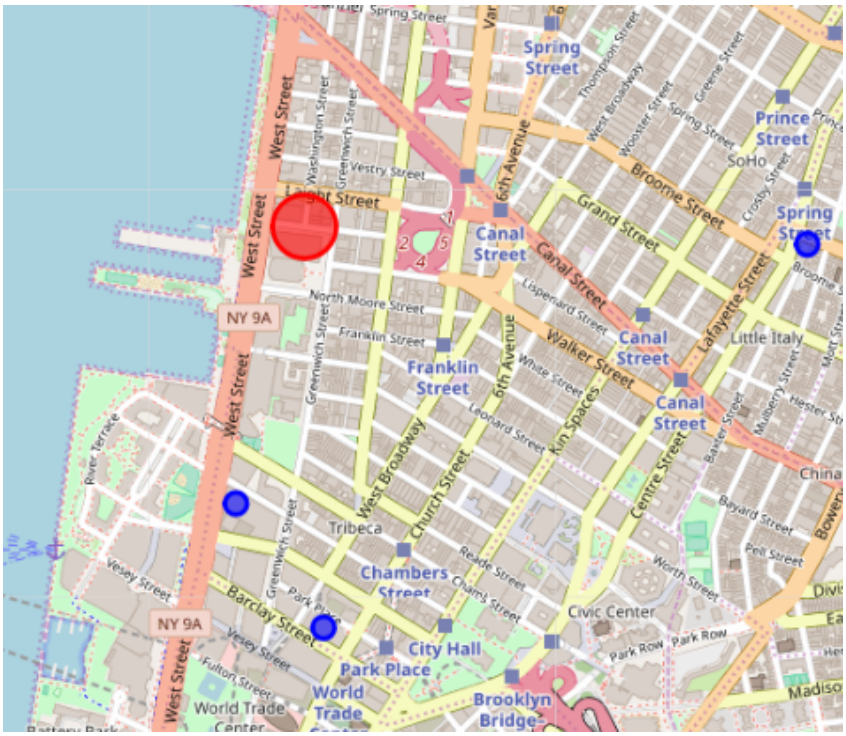

Figure 6: Trending Venues Maps

5. Discussion

The statistical function min () is used to find the minimum distanced trending venue and then it assigned as initial preferred venue for venue rating and tips analysis. This method is shows good response.

Figure 7: Venues Analysis

It's also important to know what other saying about the trending location since it may though some useful information. These details are obtained from Foursquare by sending venue ID as an input. The detail of the user who provide such feedback is also find useful for cross validation and credibility check.

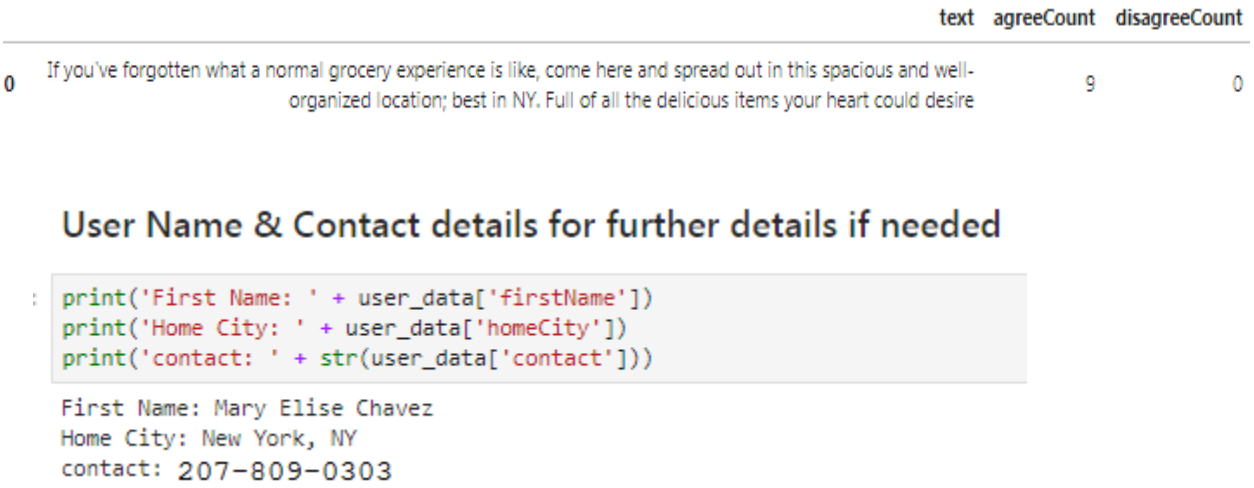

Figure 8: Venues tips & User Detals

Finally, the decision was made after exploring the nearby popular spots since it can act as a potential catalyst for further sale..

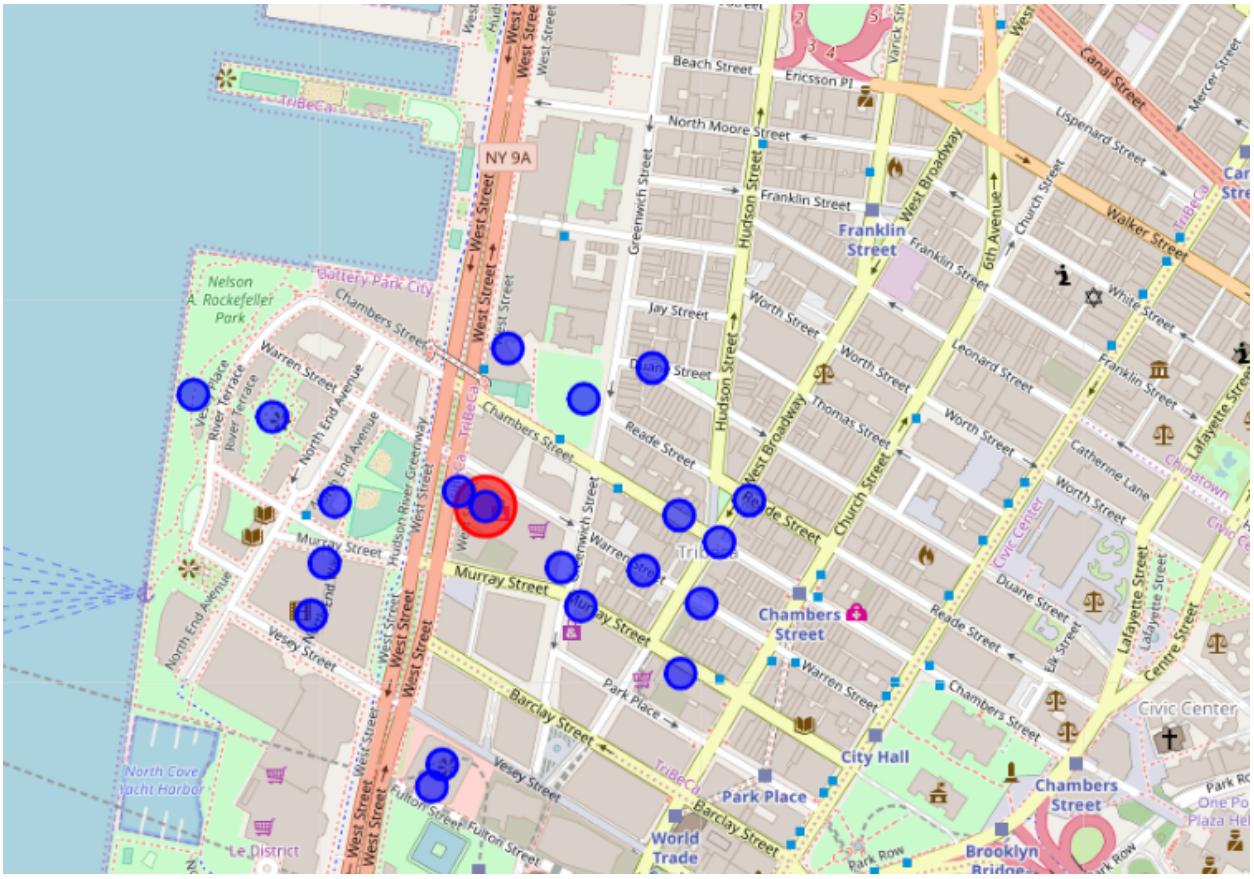

Figure 9 : Nearby Popular Spots

6. Conclusion

We developed a data analysis platform which helps Snacks on wheel mobile snack shop to find a sales spot based on trending venue details. Fourscore API is used for trending venue analysis , venue rating , number of tips , tip details , detail of the user who gave the most rated tips and finally explore the nearby location. The results indicates Snacks on wheel business owners can effectively use this data analysis methodology since the result from Fourscore is reliable and dynamic with respect to time.

7. Limitation

We identified many areas in this work which need further improvement. The current trend analysis is possible within the radius of 2000 meter , with maximum of 50 results. It may need to further improve for better venue selection. Further we filtered the category only for the key word “restaurants” , but there are many other potential filtering which helps to fine tune the outcome.

8. Reference

a. Y. Li, M. Steiner, L. Wang, Z. Zhang and J. Bao, "Exploring venue popularity in foursquare," *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Turin, 2013, pp. 205-210. doi: 10.1109/INFCOMW.2013.6562896

b. X. Long, L. Jin and J. Joshi, "Understanding venue popularity in Foursquare," *9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, Austin, TX, 2013, pp. 409-418.

c. C. Robles and J. Benner, "A Tale of Three Cities: Looking at the Trending Feature on Foursquare," *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, Amsterdam, 2012, pp. 566-571

d. Yanhua Li, Moritz Steiner, Jie Bao, Limin Wang, Ting Zhu, "Region sampling and estimation of geosocial data with dynamic range calibration", *Data Engineering (ICDE) 2014 IEEE 30th International Conference on*, pp. 1096-1107, 2014

e. https://developer.foursquare.com/docs/api (Accessed on sep 28th 2018)