BOTN_Report

June 12, 2019

by Faeez Safedien

1 I. Introduction & Business Problem:

This report is for the final course of the Data Science Specialization. A 9- course series created by IBM, hosted on Coursera platform. The problem and the analysis approach are left for the learner to decide, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

1.0.1 Problem Background:

The City of New York, is the most populous city in the United States. It is diverse and is the financial capital of USA. It is multicultural. It provides lots of business oppourtunities and business friendly environment. It has attracted many different players into the market. It is a global hub of business and commerce. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States.

This also means that the market is highly competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analysed carefully. The insights derived from analysis will give good understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk. And the Return on Investment will be reasonable.

1.0.2 Problem Description:

A restaurant is a business which prepares and serves food and drink to customers in return for money, either paid before the meal, after the meal, or with an open account. The City of New York is famous for its excelllent cuisine. It's food culture includes an array of international cuisines influenced by the city's immigrant history.

- Central and Eastern European immigrants, especially Jewish immigrants - bagels, cheesecake, hot dogs, knishes, and delicatessens
- Italian immigrants - New York-style pizza and Italian cuisine
- Jewish immigrants and Irish immigrants - pastrami and corned beef
- Chinese and other Asian restaurants, sandwich joints, trattorias, diners, and coffeehouses are ubiquitous throughout the city

- mobile food vendors - Some 4,000 licensed by the city
- Middle Eastern foods such as falafel and kebabs examples of modern New York street food
- It is famous for not just Pizzerias, Cafe's but also for fine dining Michelin starred restaurants.The city is home to "nearly one thousand of the finest and most diverse haute cuisine restaurants in the world", according to Michelin.

So it is evident that to survive in such competitive market it is very important to startegically plan.

We will tackle this problem by segmenting neighborhoods in clusters that have a similar profiles of food establishments. Within each cluster we will see if there is a gap in the market, a venue that is highly frequent in the cluster but not in the neighborhood. This will provide us with possible locations and venue types. How final recommendation will take into account the average house price as a proxy for spending power of residents in the neighborhood to better understand the our price target for the venue

1.0.3 Target Audience:

To recommend the correct location, XYZ Company Ltd has appointed me to lead of the Data Science team. The objective is to locate and recommend to the management which neighborhood of New York city will be best choice to start a restaurant. The Management also expects to understand the rationale of the recommendations made.

This would interest anyone who wants to start a new restaurant in New York city.

1.0.4 Success Criteria:

The success criteria of the project will be a good recommendation of borough/Neighborhood choice to XYZ Company Ltd based on Lack of such restaurants in that location and nearest suppliers of ingredients.

2 II. Data description:

New York city neighborhoods were chosen as the observation target due to the following reasons: - The availability of real estate prices. Though very limited. - The diversity of prices between neighborhoods. For example, a 2- bedrooms condo in Central Park West, Upper West Side can cost 4.91 million USD on average; while in Inwood, Upper Manhattan, just 30 minutes away, it's only 498 thousand USD. - The availability of geo data which can be used to visualize the dataset onto a map. The type of real estate to be considered is 2-bedroom condo, which is common for most normal nuclear families. The dataset will be composed from the following two main sources: - CityRealty which provides the neighborhoods average prices. https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804

2.1 Getting and Cleaning Data

Scraping Data from CityRealty

```
(54, 3)
```

```
Out[6]:          Area        Neighborhood  AvgPrice
        0   Brooklyn  Bedford-Stuyvesant    750000
        1   Brooklyn         Boerum Hill  1.69e+06
        ..       ...                 ...       ...
        8   Brooklyn   Downtown Brooklyn  1.79e+06
        9   Brooklyn               DUMBO  2.24e+06

        [10 rows x 3 columns]
```

2.2 Get the neighborhoods coordinate:

Free geodata is available free at: https://geo.nyu.edu/catalog/nyu_2451_34572 A copy has been downloaded and stored in IBM cloud

```
Data downloaded!
```

```
Out[12]:        Borough  Neighborhood   Latitude  Longitude
         0    Manhattan   Marble Hill  40.876551 -73.910660
         1     Brooklyn     Bay Ridge  40.625801 -74.030621
         ..         ...           ...        ...        ...
         108   Brooklyn       Erasmus  40.646926 -73.948177
         109  Manhattan  Hudson Yards  40.756658 -74.000111

         [110 rows x 4 columns]
```

2.3 Combine the two dataframes:

There are three problems here causing the number of neighborhoods doesn't match: * First, avg price data isn't available to all neighborhoods. * Second, some neighborhoods name scrapped from the website is not same as their corresponding ones in the geo dataset. * Third, real estate market names some neighborhoods differently, or make up of new names. All for the purpose of sale.

Each line of price data will be considered, and suitable action will be performed: * If the names is different, decide which one to use after searching on the internet. * If the neighborhood is missing from the geo datafram, add it's coordinate. * If the neighborhoods is makeup, combine them into the larger neighborhood which exist in the geo dataframe.

```
Out[14]:            Neighborhood  AvgPrice   Latitude  Longitude
         0    Bedford-Stuyvesant    750000  40.687232 -73.941785
         1           Boerum Hill  1.69e+06  40.685683 -73.983748
         ..                  ...       ...        ...        ...
         48       Lincoln Square  2.52e+06  40.773529 -73.985338
         49  Morningside Heights  2.52e+06  40.808000 -73.963896

         [50 rows x 4 columns]
```

```
Out[16]: <folium.folium.Map at 0x7f6bd83f5470>
```

FourSquare API which provides the surrounding venues of a given coordinates and returns the following: * Venue ID * Venue Name * Coordinates : Latitude and Longitude * Category Name

The process of collecting and clean data: - Scrap the CityRealty webpage for a list of New York city neighborhoods and their corresponding 2-bedroom condo average price. - Find the geographic data of the neighborhoods. Both their center coordinates and their border. - For each neighborhood, pass the obtained coordinates to FourSquare API. The "explore" endpoint will return a list of surrounding venues in a pre-defined radius. - Count the occurrence of each venue type in a neighborhood. Then apply one hot encoding to turn each venue type into a column with their frequency as the value. - Standardize the average price by removing the mean and scaling to unit variance. - A latitutde column and a longitude column - Each row represents a neighborhood.

The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time.

3 III. Methodology:

3.0.1 1. Data Cleanup and re-grouping.

The retrieved table contains some un-wanted entries and needs some cleanup. The following tasks will be performed: * Drop/ignore cells with missing data. * Use most current data record. * Fix data types. Post Processed Singapore towns list with and median residential rental prices * Adding geographical coordinates of each town location.

3.0.2 2. Retrieving Neighborhood coordinates

Free geodata is available free at: https://geo.nyu.edu/catalog/nyu_2451_34572

The coordinates will be used in retrieval of Foursquare API location data.

```
Out[17]: {'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]}}
```

4 IV. Segmenting and Clustering Neigborhoods in New York City

Retrieving FourSquare Places of interest. Using the Foursquare API, the **explore** API function was be used to get the most common venue categories in each neighborhood, and then used this feature to group the neighborhoods into clusters. The k-means clustering algorithm was used for the analysis. Finally, the Folium library is used to visualize the recommended neighborhoods and their emerging clusters.

In the ipynb notebook, the function **getNearbyVenues** extracts the following information for the dataframe it generates: * Venue ID * Venue Name * Coordinates : Latitude and Longitude * Category Name

The function **getVenuesByCategory** performs the following: 1. **Category** based venue search to simulate user venue searches based on certain places of interest. This search extracts the following information: * Venue ID * Venue Name * Coordinates : Latitude and Longitude * Category Name

2. For each retrieved **venueID**, retrieve the venues category rating.

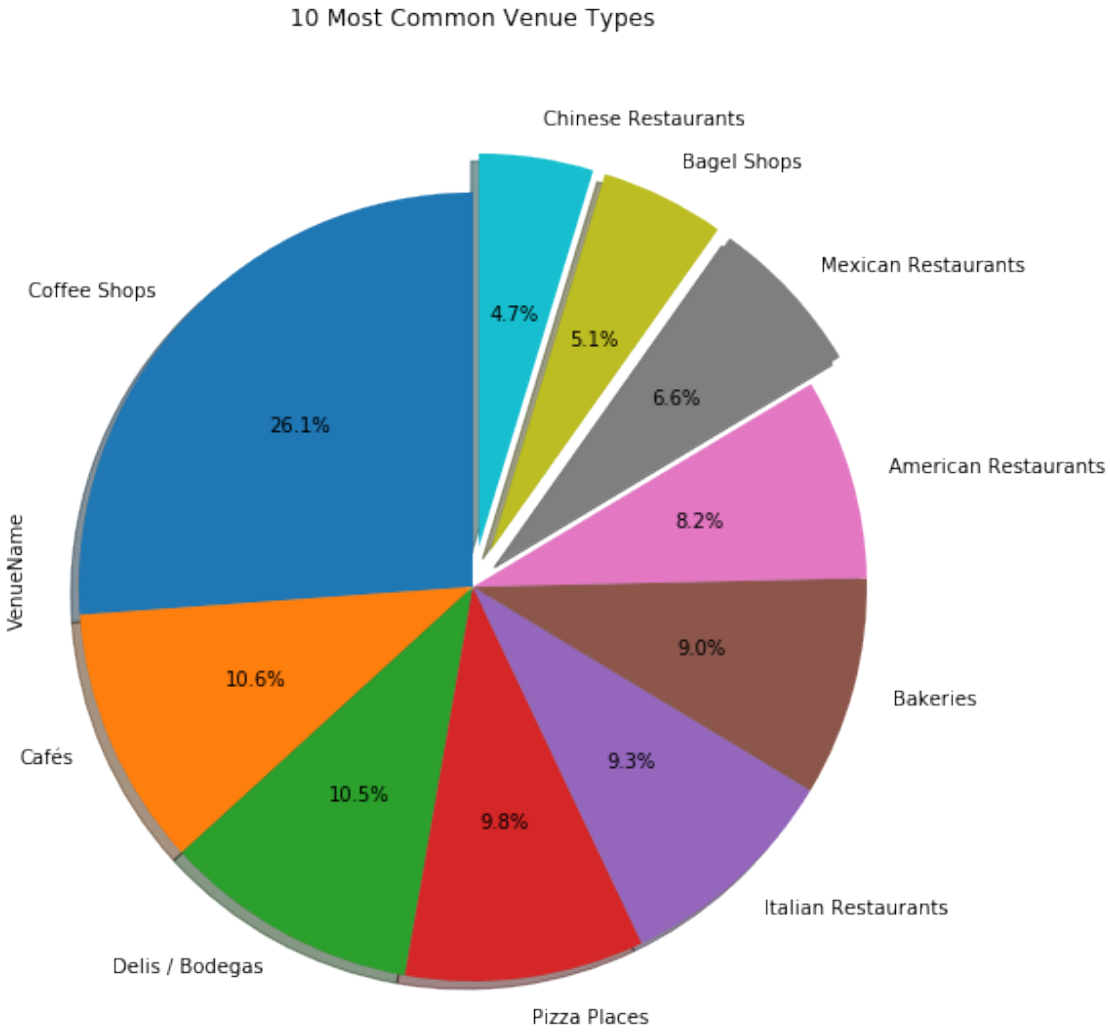
```
[#Start getVenuesByCategory]
Bedford-Stuyvesant ,Boerum Hill ,Brooklyn Heights ,Bushwick ,Carroll Gardens ,Clinton Hill ,Cobb
[#Done getVenuesByCategory]
```

4.0.1 RESULTS: How many unique categories can be curated from all the returned venues?

There are 150 uniques categories.

4.0.2 What are the top 10 most common venue types?

category	VenueName
Coffee Shops	317
Cafés	129
Delis / Bodegas	127
Pizza Places	119
Italian Restaurants	113
Bakeries	109
American Restaurants	100
Mexican Restaurants	80
Bagel Shops	62
Chinese Restaurants	57



Coffee Shops are the most common venue type with it's 317 stores making up for more than a quarter of all food establisments.

4.1 Analyse Each Neighborhoods' nearby recommended venues

- Technique: **One Hot Encoding**

```
One hot encoding returned "2483" rows.
One hot encoding re-group returned "50" rows.
```

Each rows reprents the frequency of the different venue types in a given neighborhood

4.2 Analysis of New York City's most visited venues

4.3 RESULTS: Categorized Result

Out[33]:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Battery Park City	Coffee Shops	Burger Joints	American Restaurants	BBQ Joints	Salad Places	Sandwich Places	French Restaurants	Italian Restaurants	Food Courts	Bakeries
Bedford-Stuyvesant	Delis / Bodegas	Chinese Restaurants	Pizza Places	Coffee Shops	Cafés	Fried Chicken Joints	Sandwich Places	Bagel Shops	Food	Brazilian Restaurants
...	...	...	...	...	...	...	...	...	...	...
Windsor Terrace	Delis / Bodegas	Chinese Restaurants	Italian Restaurants	French Restaurants	Cafés	Diners	Pizza Places	Vegetarian / Vegan Restaurants	Mexican Restaurants	Sushi Restaurants
Yorkville	Coffee Shops	Pizza Places	Diners	Ice Cream Shops	Italian Restaurants	Bagel Shops	Bars	Bakeries	New American Restaurants	Cafés

[50 rows x 10 columns]

4.4 RESULTS : k-means Cluster Results

Clustered results for k-means to cluster with 5 clusters. The results of grouping the neighborhoods into clusters with similar frequency of food venues are displayed on the New York City map

```
Out[37]: <folium.folium.Map at 0x7f6bd8a35828>
```

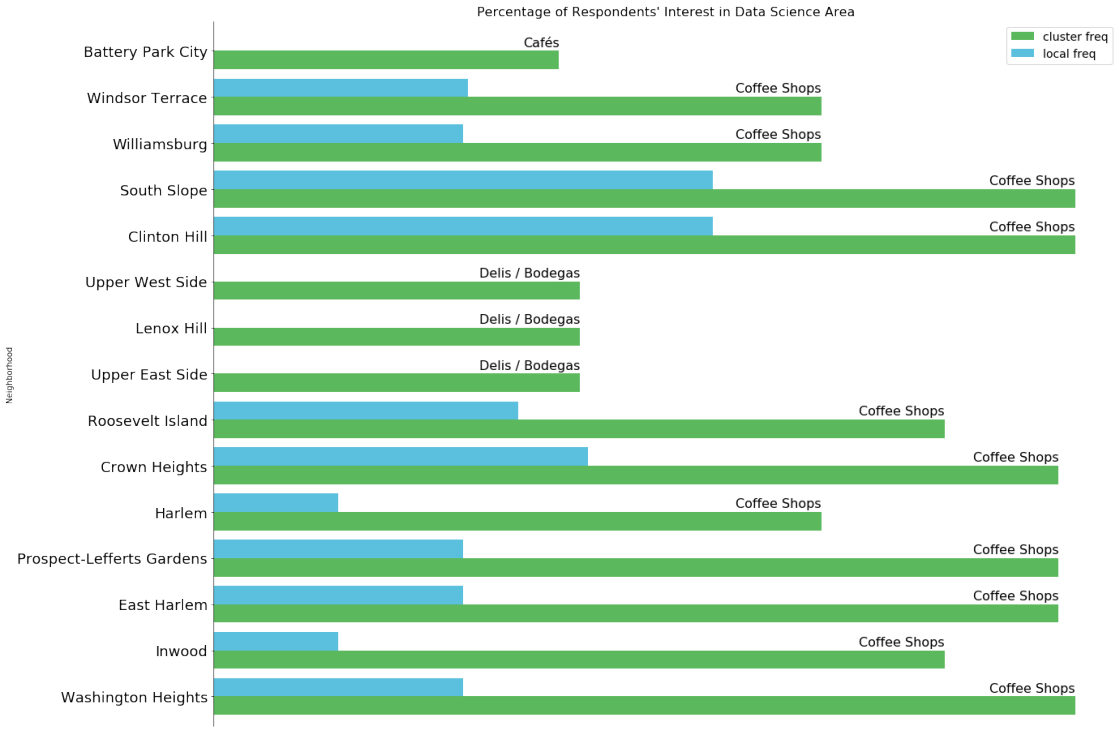
The top ten occuring venues for each cluster is tabled below

```
Out[40]:                1st Most Common Venue 2nd Most Common Venue  \
         Cluster Labels
         0                       Coffee Shops         Burger Joints
         1                    Delis / Bodegas   Chinese Restaurants
         2                       Coffee Shops       Delis / Bodegas
         3                       Coffee Shops          Pizza Places
         4                       Pizza Places          Coffee Shops

                        3rd Most Common Venue 4th Most Common Venue  \
         Cluster Labels
         0                American Restaurants            BBQ Joints
         1                       Pizza Places          Coffee Shops
         2                    Sandwich Places   Mexican Restaurants
         3                           Bakeries                Diners
         4                    Delis / Bodegas   Mexican Restaurants

                        5th Most Common Venue 6th Most Common Venue  \
         Cluster Labels
         0                       Salad Places       Sandwich Places
         1                              Cafés  Fried Chicken Joints
         2                               Bars  American Restaurants
         3                American Restaurants   Italian Restaurants
         4                              Cafés              Bakeries

                             7th Most Common Venue       8th Most Common Venue  \
         Cluster Labels
         0                       French Restaurants         Italian Restaurants
         1                          Sandwich Places                Bagel Shops
         2                Middle Eastern Restaurants       Fast Food Restaurants
         3                          Delis / Bodegas         Mexican Restaurants
         4                              Bagel Shops  Latin American Restaurants

                         9th Most Common Venue 10th Most Common Venue
         Cluster Labels
         0                         Food Courts              Bakeries
         1                                Food  Brazilian Restaurants
         2                   French Restaurants    Italian Restaurants
         3                           Juice Bars       Thai Restaurants
         4                Fast Food Restaurants    Italian Restaurants
```

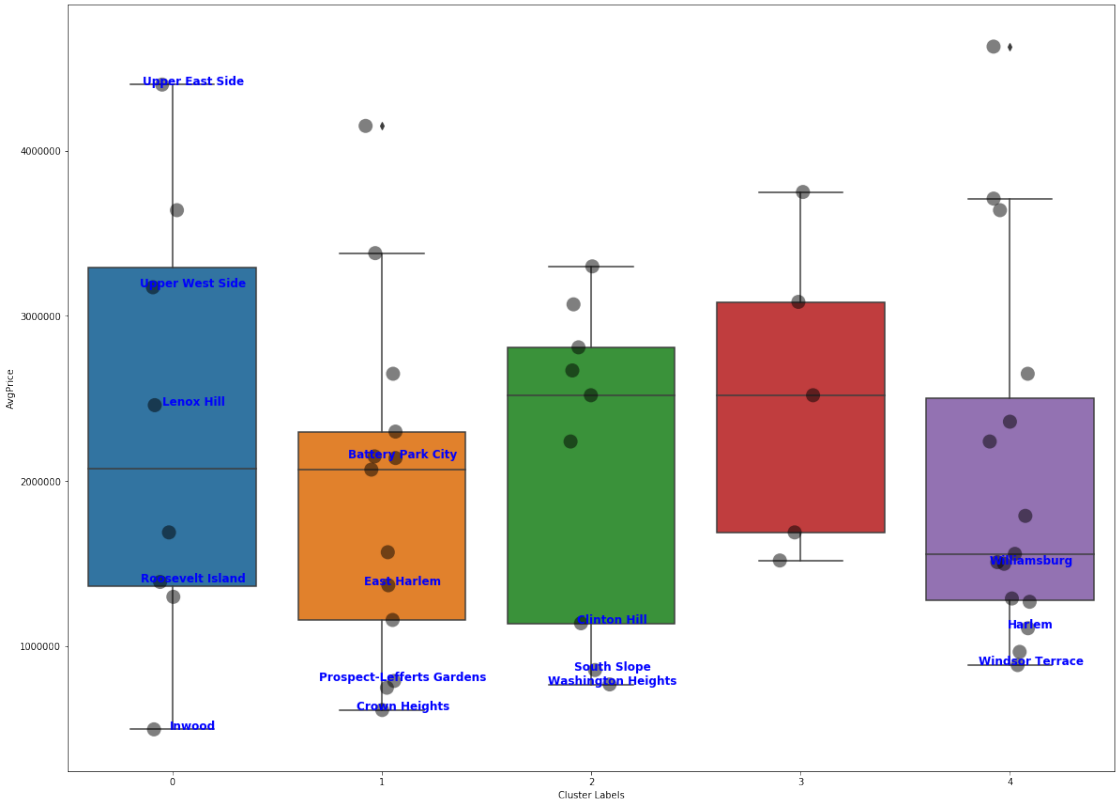
4.5 RESULTS: Most underrepresented venues

We create a dataframe with the difference between the cluster frequency of a venue and the local neighborhood frequency. This gives an indication if a venue is underrepresnted in a neighborhood.

For each neighborhood the most underrepresented venue is found and illustrated in the bar chart bellow



To help us better understand these business opportunities we use a boxplot to see the distribution of **average house prices** for each cluster and see how far the neighborhoods are from the median value



Both neighborhoods that are in their respective upper quartile, Upper East Side and Greenwich Village do not have delis/bodegas. This is a gap in the market for higher end delis to be opened in these neighborhoods. All neighborhoods in their respective bottom quartile do not have coffee shops. Coffee shop franchises looking to expand should consider looking into these neighborhoods with lower property value. Or alternatively it's an opportunity to rebrand and market coffee shops for lower income neighborhoods

5 V. Discussion and Conclusion

In this notebook, Analysis of business oportunities based on Food venue category has been presented. Recommendations are based on other neighborhoods that are in the same cluster based on the occurance of venue types.

Using the Foursquare API, we have collected a good amount of venues in New York City but since we used a free account we don't have a complete list of venues . Sourcing from the venue recommendations from FourSquare has its limitation; The list of venues is not exhaustive list of all the available venues is the area. The results therefore may significantly change, when more information are collected on those with missing data.

The generated clusters from our results shows that **average house prices** vary much between clusters, this shows that food preference isn't related to income background. However, when there is lack of food venue in the neighborhood, relative to other neighborhoods in the same cluster, then it's usually because that particular neighborhood varies greatly form the median house price in the cluster. This is a clear opportunity to open those venues but adjust for the income group

appropriately. The results show the **Coffee Shops** are the most common venue type but are not present in some lower income neighborhoods. Alternatively Upper East Side and Greenwich Village are relatively expensive neighborhoods and lack delis whih=ch are present in their food venue cluster. A clear opportunity for an upmarket deli concept exists.

I will be providing a other supplementary Inferential Statics in the future about on these data collected and also update in a new notebook using other categories. For now, this completes the requirements for this task.

Thank you. Faeez Safedien email: faeezs@gmail.com linkedin: https://www.linkedin.com/in/faeez-safedien-66285b165/ Created For: COURSERA IBM Applied Data Science Capstone Project