New_Facility_Location_Selection_Report_v2

December 30, 2019

1 Final Project:

1.1 New Facility Location Selection

1.1.1 by: Jeffrey Dupree

1.2 Table of contents

- Section **??**
- Section 1.4
- Section 1.5
- Section **??**
- Section 1.6
- Section 1.8

1.3 Introduction: Business Problem

The owner of several successful gyms wants to open a new facility in Tampa, FL. They want to ensure that the gym's location is in an area not already saturated with gyms and other businesses that might compete with a new gym. This initial analysis will be to determine to a neighborhood level, where to consider placing the new gym facility. Later analysis and research of available real estate will be required to select the final location. That is beyond the scope of this analysis.

In order to conduct this analysis, we must collect: * Zip Codes in Tampa, FL * Zip Code locations (latitude/longitude) * Zip Code boundaries * Business type and frequency

1.4 Data

1.4.1 Zip Codes

To begin with, the analysis will need specific Zip Code data for Tampa, FL.

Step one: Identify the list of Zip Codes that correspond to Tampa, FL. For that, this notebook will scrape information from a ZIP-CODES.COM page https://www.zip-codes.com/state/fl.asp#zipcodes to create a dataframe consisting of the Zip Code, the City name, County name and the Zip Code type. Use the BeautifulSoup package to scrape the information from the webpage. I used the lxml parsing method, but you can use any you like. Find the table using `soup.find` from BeautifulSoup. Initially, the analyst must display the structure and content of the table (a portion shown below). Once the analyst understands the structure, they can develop the logic required to extract the desired elements in the next steps.

```
<table border="0" cellpadding="0" cellspacing="0" class="statTable" id="tblZIP"
title="All Florida ZIP Codes, City, County, Classification, and Area Codes."
width="99%">
 <tr>
  <td class="label" title="All ZIP Codes for Florida">
   <strong>
    ZIP Code
   </strong>
  </td>
  <td class="info" title="The official city name as designated by the USPS.">
   <strong>
    City
   </strong>
  </td>
  <td class="info" title="The primary county or parish this ZIP Code serves.">
   <strong>
```

Now a pandas dataframe needs to be created. This will require looping through the elements from the table and assigning the elements to a list. The list can then be made into a dataframe using `pd.DataFrame`. The columns will need header names. I manually assigned these instead of pulling them from the BeautifulSoup object `table`. Next remove the rows where the type is "P.O. Box". The first five rows of the resulting dataframe look like this.

[7]:

	Zip_Code	City	County	Type
0	33602	Tampa	Hillsborough	Standard
1	33603	Tampa	Hillsborough	Standard
2	33604	Tampa	Hillsborough	Standard
3	33605	Tampa	Hillsborough	Standard
4	33606	Tampa	Hillsborough	Standard

Step two: The locations of the Zip Codes (latitude and longitude) will need to be collected. This will be accomplished through Nominatim in the Geopy library. This leverages the OpenStreetMap (OSM) dataset application programming interface (API) to geolocate each Zip Code. This adds two rows (location, point) to the dataframe. The first five rows are shown here.

```
C:\Users\JeffDupree\Anaconda3\lib\site-packages\tqdm\std.py:648: FutureWarning:
The Panel class is removed from pandas. Accessing it from the top-level
namespace will also be removed in the next version
  from pandas import Panel
```

[9]:

	Zip_Code	City	County	Type	\
0	33602	Tampa	Hillsborough	Standard	
1	33603	Tampa	Hillsborough	Standard	
2	33604	Tampa	Hillsborough	Standard	
3	33605	Tampa	Hillsborough	Standard	
4	33606	Tampa	Hillsborough	Standard	

	location	\
0	(Ybor City, Tampa, Hillsborough County, Florid…	

```
1  (Tampa, Hillsborough County, Florida, 33603, U…
2  (Sulphur Springs, Tampa, Hillsborough County, …
3  (East Ybor, Tampa, Hillsborough County, Florid…
4  (Hyde Park, Tampa, Hillsborough County, Florid…

                                      point
0               (27.9516574, -82.449638, 0.0)
1  (27.9823952329372, -82.4629461755015, 0.0)
2             (28.0127051, -82.4665599, 0.0)
3               (27.96589, -82.4209639, 0.0)
4             (27.9341317, -82.4680636, 0.0)
```

Now the latitude and longitude values for each of the postal codes are separated out into respective columns. Next we take the first portion of the location string, removing everything after the first comma, then renaming the column "Neighborhood". The dataframe now looks like this.

[12]:

	Zip_Code	City	County	Type	Neighborhood	Latitude	Longitude
0	33602	Tampa	Hillsborough	Standard	Ybor City	27.951657	-82.449638
1	33603	Tampa	Hillsborough	Standard	Tampa	27.982395	-82.462946
2	33604	Tampa	Hillsborough	Standard	Sulphur Springs	28.012705	-82.466560
3	33605	Tampa	Hillsborough	Standard	East Ybor	27.965890	-82.420964
4	33606	Tampa	Hillsborough	Standard	Hyde Park	27.934132	-82.468064

Step three: The last feature of Zip Code data needed are the boundaries of each Zip Code. These will be stored as latitudes and longitudes for the verices of polygons representing areas corresponding to each Zip Code. This data is downloaded as a GeoJSON file from https://opendata.arcgis.com/datasets/d356e19e0fb34524b54d189fafb0d675_0.geojson.

1.4.2 Business Data

Once the Zip Code data are collected, we need to collect the data on the surrounding businesses. We use the Foursquare API to collect data about the businesses near each Zip Code loaction.

1.5 Methodology

Locate Zip Codes Lacking Gyms We can start by visualizing the location of each zip code (based on the coordinates associated with it). The very first visualization is to plot the locations associated with each zip code to ensure that they fall within the intended area. These locations can also be labeled with information from the dataframe to make the graphic interactive. Selecting a point on the map reveals the Latitude, Longitude, and Neighborhood associated with that point.

[13]: <folium.folium.Map at 0x2ba6df071c8>

A query of the Foursquare API returns the top 150 venues within 1000 meters of the zip code locations. The query is passed as a url using the `get()` command and returns a json formatted response. After reviewing the structure of the JSON, a function must be created to extract the venue category types associated with each zip code. The venues can then be placed in a table with the venue name, category, latitude, and longitude as columns. The first five rows of the table are shown here.

[23]:

	name	categories	lat	lng
0	Pour House at Grand Central	Bar	27.951357	-82.447740
1	Crunch - Channelside	Gym / Fitness Center	27.951152	-82.447940
2	Cena	Italian Restaurant	27.951569	-82.447869
3	Publix - Channelside	Grocery Store	27.952128	-82.448741
4	City Dog Cantina	Mexican Restaurant	27.951118	-82.447726

We then create a function that uses the Foursquare API to find the nearby venues for all of the neighborhoods, by zip code. The `getNearbyVenues` function can then be applied to the dataframe to create a dataframe of the venues near the grid associated with each zip code. The first five rows of the resulting dataframe will look like this.

[27]:

	Zip_Code	Zip Latitude	Zip Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	33602	27.951657	-82.449638	Pour House at Grand Central	27.951357	-82.447740	Bar
1	33602	27.951657	-82.449638	Crunch - Channelside	27.951152	-82.447940	Gym / Fitness Center
2	33602	27.951657	-82.449638	Cena	27.951569	-82.447869	Italian Restaurant
3	33602	27.951657	-82.449638	Publix - Channelside	27.952128	-82.448741	Grocery Store
4	33602	27.951657	-82.449638	City Dog Cantina	27.951118	-82.447726	Mexican Restaurant

Once all of the venues have been associated with neighborhoods by proximity, the frequency of venue types can be determined. However, before the frequency of each venue can be calculated, a list of the unique venue categories must be created and evaluated. This can be a very long list of more than 100 categories. Many of these categories can be very similar. For example the categories "Gym / Fitness Center" and "Gym" appear in the list. These two categories could be considered to be the same. Another venue that would compete with a gym is 'Military Base'. Military bases have gyms and fitness centers for military members at no cost. This could reduce the need for another gym in the area. We will need to recode any gym-like categories with a common category name (i.e., gym). This will require examining the list of unique categories and creating a list of the categories that should be recoded. We use one-hot encoding to determine if a venue type exists in a neighborhood. One-hot encoding will create a column for each of the unique categories, and assign a value of 1 if that venue type exists in the neighborhood or 0 otherwise for each row. A portion of that table would look like this.

[31]:

	Zip_Code	Accessories Store	Airport	Airport Lounge	Airport Service	\
0	33602	0	0	0	0	
1	33602	0	0	0	0	
2	33602	0	0	0	0	
3	33602	0	0	0	0	
4	33602	0	0	0	0	

	American Restaurant	Antique Shop	Aquarium	Arcade	Art Gallery	…	\
0	0	0	0	0	0	…	
1	0	0	0	0	0	…	
2	0	0	0	0	0	…	
3	0	0	0	0	0	…	
4	0	0	0	0	0	…	

	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	\
0	0	0	0	
1	0	0	0	
2	0	0	0	
3	0	0	0	
4	0	0	0	

	Vietnamese Restaurant	Waste Facility	Wine Bar	Wings Joint	\
0	0	0	0	0	
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	

	Women's Store	Zoo	Zoo Exhibit
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0

[5 rows x 178 columns]

With the one-hot encoded data, we can determine the frequency with which each venue type occurs in each borough. This results in a dataframe with a column for each unique venue type and a row for each unique borough.

[32]:

	Zip_Code	Accessories Store	Airport	Airport Lounge	Airport Service	\
0	33602	0.0	0.0	0.0	0.0	
1	33603	0.0	0.0	0.0	0.0	
2	33604	0.0	0.0	0.0	0.0	
3	33605	0.0	0.0	0.0	0.0	
4	33606	0.0	0.0	0.0	0.0	

```
   American Restaurant  Antique Shop  Aquarium  Arcade  Art Gallery  …  \
0             0.033333      0.000000  0.011111     0.0     0.011111  …
1             0.000000      0.055556  0.000000     0.0     0.111111  …
2             0.025000      0.000000  0.000000     0.0     0.000000  …
3             0.083333      0.000000  0.000000     0.0     0.000000  …
4             0.040000      0.000000  0.000000     0.0     0.000000  …

   Vegetarian / Vegan Restaurant  Video Game Store  Video Store  \
0                          0.000               0.0         0.00
1                          0.000               0.0         0.00
2                          0.025               0.0         0.00
3                          0.000               0.0         0.00
4                          0.000               0.0         0.02

   Vietnamese Restaurant  Waste Facility  Wine Bar  Wings Joint  \
0                    0.0             0.0     0.000          0.0
1                    0.0             0.0     0.000          0.0
2                    0.0             0.0     0.025          0.0
3                    0.0             0.0     0.000          0.0
4                    0.0             0.0     0.020          0.0

   Women's Store    Zoo  Zoo Exhibit
0           0.00  0.000     0.011111
1           0.00  0.000     0.000000
2           0.00  0.025     0.250000
3           0.00  0.000     0.000000
4           0.02  0.000     0.000000

[5 rows x 178 columns]
```

Next we will determine the five most frequent venues within a borough to describe a neighborhood 'type', and group the borough by type symilarity. We begin by creating a function that will return the most common venues for each zip code.

[34]:
```
<pandas.io.formats.style.Styler at 0x2ba6fb6e508>
```

Now that we can see what the five most common venues are in each Zip Code, we can eliminate those Zip Codes with 'gym' type venues in the top five.

[35]:
```
    Zip_Code 1st Most Common Venue 2nd Most Common Venue 3rd Most Common Venue  \
5      33607        Scenic Lookout       Harbor / Marina            Food Truck
6      33609        Clothing Store         Women's Store        Sandwich Place
7      33610         Grocery Store        Discount Store            Restaurant
8      33611    Turkish Restaurant        Sandwich Place     Korean Restaurant
15     33618           Pizza Place   American Restaurant        Massage Studio
19     33625  Fast Food Restaurant            Nail Salon           Gas Station
20     33626      Insurance Office          Home Service           Zoo Exhibit
```

```
   4th Most Common Venue 5th Most Common Venue
5             Karaoke Bar     Mobile Phone Shop
6          Lingerie Store      Department Store
7             Video Store                   Spa
8                    Food                 Motel
15            Coffee Shop           Hobby Shop
19          Big Box Store          Coffee Shop
20          Event Service             Fountain
```

Now the list only includes Zip Codes where 'gym' type venues are not one of the five most frequent venue types. We can sort this list by descending frequency of gyms. Where the gym frequencies are equal, records are sorted by Zip_Code in ascending order.

[36]:
```
   Zip_Code       Gym
5     33607  0.000000
6     33609  0.000000
7     33610  0.000000
19    33625  0.000000
20    33626  0.000000
15    33618  0.025641
8     33611  0.066667
```

Now that we have the reduced list of zip codes, we join it to our location dataframe, rename the 'Gym' column as 'Gym Frequency', and reset the indeces.

[37]:
```
   Zip_Code  Gym Frequency   City        County      Type          Neighborhood  \
0     33607       0.000000  Tampa  Hillsborough  Standard                 Tampa
1     33609       0.000000  Tampa  Hillsborough  Standard            Palma Ceia
2     33610       0.000000  Tampa  Hillsborough  Standard              Ybor City
3     33625       0.000000  Tampa  Hillsborough  Standard  Hillsborough County
4     33626       0.000000  Tampa  Hillsborough  Standard  Hillsborough County
5     33618       0.025641  Tampa  Hillsborough  Standard           Mullis City
6     33611       0.066667  Tampa  Hillsborough  Standard            Palma Ceia

    Latitude  Longitude
0  27.973131 -82.585196
1  27.944813 -82.536276
2  27.977944 -82.442975
3  28.068327 -82.557302
4  28.057031 -82.610797
5  28.039589 -82.508293
6  27.880332 -82.498916
```

Now we can display the locations on a map. Selecting a marker on the map will display that zip code and the frequency of 'gym' type venues within 1km of the zip code central point.

[38]:
```
<folium.folium.Map at 0x2ba6e076608>
```

Using the GeoJSON file from https://opendata.arcgis.com/datasets/d356e19e0fb34524b54d189fafb0d675_0.geojso
polygons for the Zip Codes of intereset can be defined using the latitude and longitude coordinates.
Below we create a list of coordinates for both latitudes and longitudes, then place these lists at
the end of the dataframe.

1.6 Results

Now the polygons for the areas represented by the zip code can be overlaid on the map.

[41]:
```
<folium.folium.Map at 0x2ba700a94c8>
```

1.7 Discussion

Using this method the analyst is able to quickly gather and display location and venue information for the area of interest. With this data the analyst can categorize the areas by the types of venues in that are and the frequency with which they occur. This allows for a cursory analysis to narrow down the locations that may be good choices for a new gym facility.

There are some drawbacks to this application. Primarily that the search for venues is conducted in a circular area of radius 1km from the coordinates pulled from the website https://www.zip-codes.com/state/fl.asp#zipcodes. These coordinates do not alwasy correspond to the geographic center of the area. If the coordinates map to a location within the zip code area that is in a remote section, there may not be many venues within 1km of the point. Also, some of the points may be less than 1km from the boundary. This may result in some venues from other zip codes being included with multiple zip codes.

However, the strength of this methodology is that it is dynamic. As more venue infromation is added or modified within the FourSquare platform, the results of this analysis will take those changes into account when rerun.

1.8 Conclusion

[42]:
```
At the time of this model run, there were 7 zip codes that met the criteria for
the new location. The customer can now focus their location serach to a few zip
codes, saving time and money.
```