

May, 2019

ORLANDO BENCHMARK

1. INDEX

To get started right away, just tap any placeholder text (such as this) and start typing.

- A. Introduction
- B. Crawling Wikipedia with Beautiful Soup
- C. Geocoder and Folium use of the Location Data
- D. Segmentation and Clustering of the Orlando Data
 - i. By Neighborhood
 - ii. By Venue Category
- E. Conclusions

A. Introduction

The Orlando area is explored for business opportunities using the Foursquare API. To do that, Wikipedia is scraped for the neighborhood names of the area and Geocoder is used to extract the coordinates of each neighborhood.

Then with the coordinates for the neighborhoods the Foursquare API is used to get the most common categories of venues in these neighborhoods and thus bring some conclusions as to whether there are any unmet needs in the area that could be satisfied through an entrepreneurship project.

The Foursquare API is leveraged through its different features in order to explore the area and find new insights and Geocoder is used to find the coordinates of some neighborhoods in Orlando that were scraped from Wikipedia.

B. Crawling Wikipedia with BeautifulSoup

The BeautifulSoup package is used by creating a soup object containing html text. Normal text can be extracted using a 'parser' so the html code becomes regular written text. Wikipedia contained a list of neighborhoods for Orlando. This list was contained in a table so a loop was used to extract the 'td' elements in the html. Further preparation of the data was done in excel and the clean data was uploaded through a CSV.

```
import requests
import lxml
! pip install BeautifulSoup
from bs4 import BeautifulSoup

wikipedia_link='https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Orlando,Florida'

raw_wikipedia_page = requests.get(wikipedia_link)
page= raw_wikipedia_page.text
soup= BeautifulSoup(page,'lxml')

orlando_data_saved=""
for table in soup.find_all('tr'):
    orlando_data=""
    for data in soup.find_all ('td'):
        orlando_data=orlando_data+", " + data.text
    orlando_data_saved=orlando_data_saved + "\n" + orlando_data[1:]
    print(orlando_data_saved)
```

C. Geocoder and Folium use of the Location Data

With a clean list of the neighborhoods we proceed to start the geocoder and mapping processes of the neighborhoods. We add the string ", Florida" to the neighborhoods in order to improve accuracy of the location.

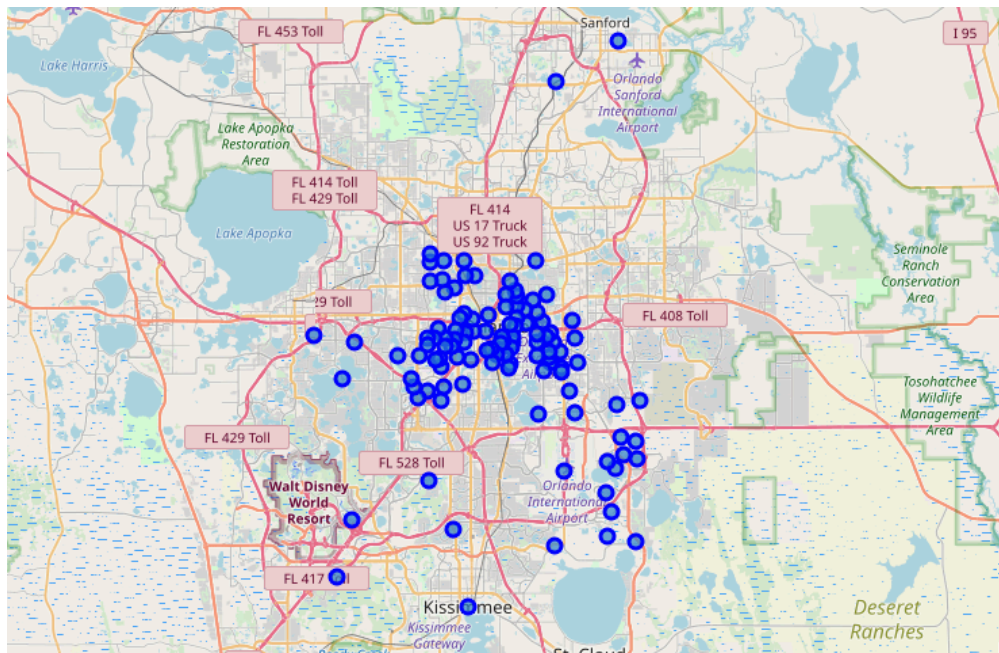
Then we start using a loop to obtain the coordinates of each neighborhood making sure to use a time.sleep of 0.45 so the geocoder does not kick our calls out.

for address in addressess:

```
geolocator = Nominatim(user_agent="FL_explorer")  
location = geolocator.geocode(address)  
latitude = location.latitude  
longitude = location.longitude  
time.sleep(.45)  
print(address, '{}, {}'.format(latitude, longitude), " " )
```

After having the appropriate latitude and longitude variables we proceed to map the neighborhoods using folium adding a circle marker to each neighborhood point on the map.

With Folium we are able to map the neighborhoods that we extracted from Wikipedia:



D. Segmenting and Clustering Using Foursquare API

We start by using a Foursquare API calls function. In this function we are interested in finding the categories of each venue so we can generalize what types of venues there are

in the area and process a lot more information by grouping all the venues according to their category.

The Foursquare API uses the latitude and longitude variables that we geocoded earlier, as well as limit which assigns the maximum amount of requests, and the radius of each call according to coordinates which is in meters.

```
LIMIT = 500
```

```
radius = 1000
```

```
url =
```

```
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&l={}&radius={}&limit={}'.format(
```

```
    CLIENT_ID,
```

```
    CLIENT_SECRET,
```

```
    VERSION,
```

```
    latitude,
```

```
    longitude,
```

```
    radius,
```

```
    LIMIT)
```

```
results = requests.get(url).json()
```

By using a radius of 1000 meters to look on the Foursquare API we are able to extract around 3000 venues which gives a more robust dataset, and with that quantity of information the analysis becomes more interesting.

In order to segment or cluster the data we convert the categorical values into numerical values. We achieve this by applying one hot (get_dummies) encoding to the Venue Category column, thus creating a dataframe that measures how often there is a specific category of venue in each neighborhood.

However, we need to process the data further in order to see some concrete conclusions.

We can visualize the top 5 venue categories in each neighborhood. It gives us a first glance into the data organized in a way that allows for some initial conclusions about the dataset.

If we look at the top 5 venue categories for the neighborhood Lake Buena Vista, we can confirm that it's top venues are clothing stores, resorts, gift shops, restaurants and theme parks. This is aligned with the reality of the neighborhood because it is the neighborhood adjacent to the Disney Theme Parks.

Thus we can assume that there is good credibility to the data we have and we can trust the processes we performed so far.

```

----Lake Buena Vista----
      venue  freq
0      Clothing Store 0.08
1      Resort 0.07
2      American Restaurant 0.04
3  Theme Park Ride / Attraction 0.04
4      Gift Shop 0.04

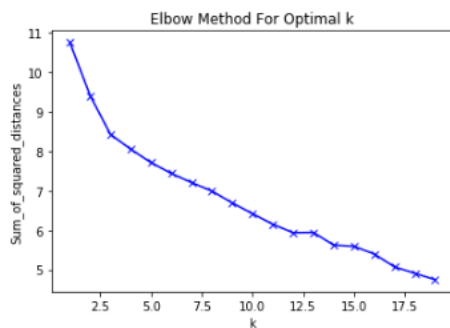
```

Also, displaying a table of the 10 most common venues will help when we try to cluster the neighborhoods by similar venues. However, as we will see clustering helps partially. Mostly to get more familiar with the data, however, it will not help us much in drawing conclusions.

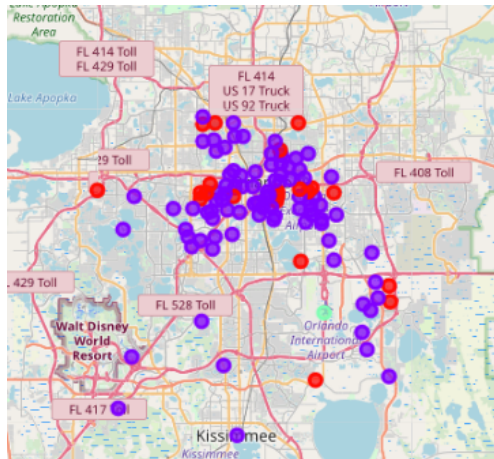
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Audubon Park	Bakery	Clothing Store	Sushi Restaurant	Karaoke Bar	Department Store	Park	Coffee Shop	Gas Station	Accessories Store	Record Shop
1	Azalea Park	Cuban Restaurant	Construction & Landscaping	Arepa Restaurant	Latin American Restaurant	Automotive Shop	Office	Baseball Field	South American Restaurant	Convenience Store	Fish Market
2	Bal Bay	Yoga Studio	Cosmetics Shop	Dog Run	American Restaurant	Steakhouse	Art Gallery	Pool	Office	Fast Food Restaurant	Ethiopian Restaurant
3	Baldwin Park	Park	Mexican Restaurant	Light Rail Station	Pharmacy	Restaurant	Sandwich Place	Candy Store	Dog Run	Burger Joint	Smoothie Shop
4	Bel Air	Baseball Field	Park	Baseball Stadium	Breakfast Spot	Discount Store	Fast Food Restaurant	Ethiopian Restaurant	Event Space	Fabric Shop	Farmers Market

i. By Neighborhood

We run some clustering tests iteratively, in order to find the best number of k clusters. After using the elbow method for inertia, we see that a set of 3 clusters is likely the optimal amount.



The Elbow Method does not show a strong elbow; however, it seems to happen at $k=3$.



At technically 3 clusters we can see that neighborhoods are distributed in mainly two clusters. 22 neighborhoods that have Parks as their main Venue, and 88 that have stores, restaurants and pretty much most types of venues.

However, experimenting with different amount of clusters anywhere from 10 or more clusters allows us to see that there are some neighborhoods with similar characteristics, and allows us to see that restaurants and stores are very common like a regular city. Some neighborhoods also have parks, yoga studios, and gyms as their main venues.

This suggests that in the Orlando area, customers are interested mostly in food venues, but less intuitively they are also interested in fitness-health venues, which directs us to look at the possibility of exploring a food-health business project.

In addition, there are is very few amount of vegetarian-vegan restaurants locally which would also provide an interesting niche for a restaurant. The neighborhoods where we can find these vegetarian/vegan restaurants are: Dover Manor, East Central Park, Colonialtown North, Colonialtown South, Park Lake/Highland, Lake Fairview, and Millenia.

Neighborhood	Neighborhood	Neighborhood	Venue	Venue Category
Dover Manor	28.5280201	-81.321284	Julia's Hippy House	Vegetarian / Vegan Restaurant
East Central Park	28.5443922	-81.347801	Market on South	Vegetarian / Vegan Restaurant
East Central Park	28.5443922	-81.347801	Dixie Dharma	Vegetarian / Vegan Restaurant
Colonialtown North	28.5604652	-81.358061	The Sanctum Cafe	Vegetarian / Vegan Restaurant
Colonialtown South	28.5490633	-81.35819	The Sanctum Cafe	Vegetarian / Vegan Restaurant
Colonialtown South	28.5490633	-81.35819	Veggie Garden	Vegetarian / Vegan Restaurant
Park Lake/Highland	28.5586222	-81.369055	Veggie Garden	Vegetarian / Vegan Restaurant
Lake Fairview	28.6007094	-81.411018	Dajen Eats Cafe And Creamery	Vegetarian / Vegan Restaurant
Millenia	28.4882989	-81.432327	Maoz Vegetarian	Vegetarian / Vegan Restaurant

ii. By Venue Category

Due to the limited amount of insight provided by the segmentation and clustering by neighborhood, we proceed to segment the data by venue category thus showing what are the most common neighborhoods for each kind of venue.

This new arrangement of the data allows us to look at the categories we are now interested in e.g. Gyms and restaurants. As a highlight, the Metrowest neighborhood has

the highest amount of Gyms / Fitness Center category in the city as do, Wadeview Park, Colonial Town Center and Coytown.

----Gym / Fitness Center----			----Gym----		
	hood	freq		hood	freq
0	Metrowest	0.06	0	Wadeview Park	0.06
1	Park Central	0.04	1	Colonial Town Center	0.06
2	Airport North	0.04	2	Coytown	0.06
3	Lake Fredrica	0.04	3	Williamsburg	0.04
4	South Orange	0.04	4	Lake Formosa	0.04
5	Kirkman South	0.04	5	Lake Holden	0.04
6	South Semoran	0.04	6	Colonialtown South	0.04
7	Mariners Village	0.04	7	South Orange	0.04
8	Windhover	0.04	8	Camellia Gardens	0.04
9	Lake Holden	0.04	9	Rowena Gardens	0.02

In addition, if we explore which neighborhoods have the highest amount of the venue category 'Yoga Studio', we find East Central Park, Audubon Park, Colonial Town Center, Williamsburg, Park Lake/ Highland and Lake Formosa.

----Yoga Studio----		
	hood	freq
0	East Central Park	0.17
1	Audubon Park	0.08
2	Colonial Town Center	0.08
3	Williamsburg	0.08
4	Park Lake/Highland	0.08
5	Lake Formosa	0.08
6	Kirkman South	0.08
7	Dover Shores West	0.08
8	Coytown	0.08
9	Lawsona/Fern Creek	0.08

This is relevant because these would be potential neighborhoods where we could introduce our Healthy-Food-Vegan restaurant concept, given that there seem to be markets for fitness and health enthusiasts.

E. Conclusions

Overall, in Orlando the most common venues:

0	Park	132
1	Pizza Place	89
2	Sandwich Place	85
3	Convenience Store	84
4	Fast Food Restaurant	71
5	American Restaurant	71
6	Grocery Store	61
7	Cosmetics Shop	60
8	Clothing Store	59
9	Pharmacy	52
10	Coffee Shop	52
11	Gym	51
12	Intersection	49
13	Gym / Fitness Center	47
14	Mexican Restaurant	45
15	Discount Store	44
16	Hotel	42
17	Bakery	42
18	Bar	41
19	Vietnamese Restaurant	38

So initially, a Vegan Restaurant would be competing with Pizza Places, Sandwich Places, and Fast Food Restaurants, and American Restaurants, which are the top 5 venue categories in Orlando.

However, gyms are also ubiquitous, with almost 100 in the entire city. As well as parks with 132 parks in the area.

A restaurant that represents a healthy option to pizzas, sandwiches and fast food, could be an interesting business model. Metrowest and Wadeview Park in particular seem to be interesting areas to further explore given that they have a high number of gyms.

This additional exploration could include demographic information of the neighborhoods such as median income or population per area, as well as real state or rent prices for the business' location. Also, a regression could be made regarding the correlation between gyms and healthy food venues in order to see if in other cities people in neighborhoods with a high number of gyms demand healthy food places.

The value in this model is that it is flexible and can be used to explore different locations. Leveraging tools like geocoder and Foursquare API is very useful since there is a big amount of information available.

The problem of finding a business opportunity in Orlando has been narrowed down throughout the project to opening a Vegan-Vegetarian Restaurant in an optimal location.

The data that was generated in the project allowed to both define the problem and to point into a solution.

We could conclude that the project has given an initial solution to the problem:

What is an interesting business venture in the Orlando Area?

However, we only used the Foursquare API for benchmarking by location. The Foursquare API could be further leveraged to do sentiment analysis surrounding venues in Orlando by analyzing people's review and thus having direct contact with the customers. However, the call limits in Foursquare API don't allow to obtain a large amount of this data but, this is where the most valuable information is at.

We thus conclude, that according to this benchmarking study, the Vegetarian/Vegan Restaurant has limited competition in its niche. However, in a more global sense it is competing against fast food and American food which is ubiquitous in Orlando. Thus, an option for the business model would be to adopt fast food and American dishes and transform them into vegan dishes, in order to appeal to a larger market since the Vegetarian / Vegan Restaurant niche is presently small.