Introduction/Business Problem

In this project we will try to find an optimal location for a restaurant. In particular for this project we will try to find an optimal location for an Italian Restaurant in Manhattan, NY.

Since there are a lot of restaurants in and around the many neighbourhoods of Manhattan, NY we will try to find locations where there aren't that many restaurants and in particular only a few Italian restaurants in the vicinity.

We will use our data science powers to generate a few promising locations based on these criteria. Advantages of each one will then be clearly expressed so that best possible final location can be chosen by stakeholders.

This problem will be particularly useful for anyone who intends to open a new place of business (be it a restaurant, shopping mall, gym, etc.) as the ideal location for all such enterprises is one where a lot of such businesses do not already exist in the vicinity.

Data

Based on definition of our problem, factors that will influence our decision are:

- number of existing restaurants in the neighbourhood (any type of restaurant).
- number of and distance to Italian restaurants in the neighbourhood, if any.
- distance from neighbourhood (By taking the Latitude and Longitude values of the neighbourhood as its centre).

Following data sources will be needed to extract/generate the required information:

- Longitude and latitude details of the different neighbourhoods in NY which is available from the lab assignment of week 3.
- centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using Google Maps API reverse geocoding.
- number of restaurants and their type and location in every neighbourhood will be obtained using Foursquare API.

Methodology

We first downloaded the data about the different neighbourhoods of NY and extracted the data pertaining to the Manhattan borough alone from it and stored it in a pandas DataFrame. We then also added the X and Y Cartesian coordinate values for all the neighbourhoods using their Latitude and Longitude details available in the DataFrame.

Using Foursquare API, we can then get the details of all the restaurants, tavernas, diners within a radius of 750 metres from the centre of every neighbourhood. We also keep track of the Italian restaurants among all the restaurants.

Now, since the number of neighbourhoods in Manhattan is 40 and they are all pretty close to each other, I decided to cluster them into 4 groups based on their proximity to each other using the k-means clustering algorithm.

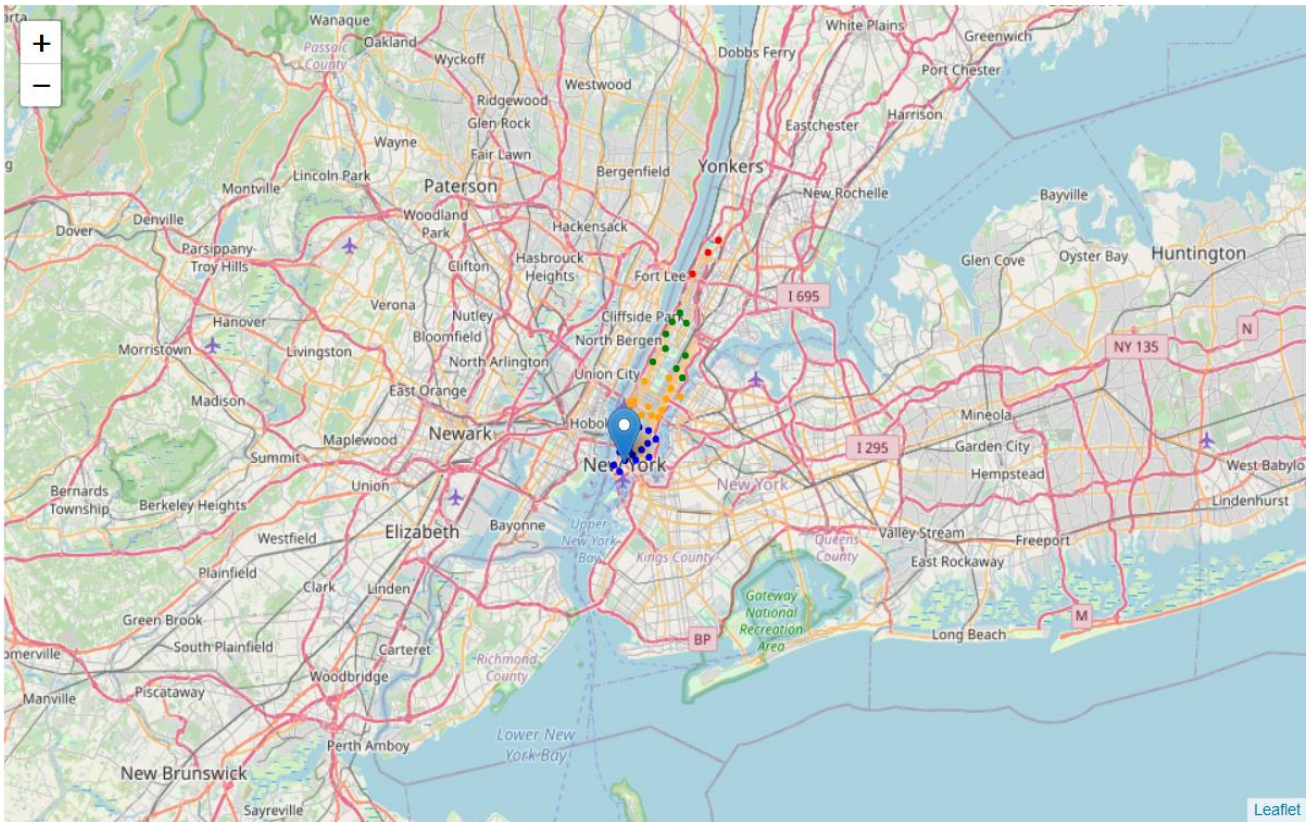

Fig 1. Cluster of neighbourhoods in Manhattan, NY.

From Fig 1, we can see that the red cluster has only 3 neighbourhoods and so I decided to not concentrate on that cluster, the green cluster also has only a few restaurants around and thus I concentrated only on 2 clusters of neighbourhoods closest to the Marker in Fig 1.

I decided to work with the two clusters separately and to come up with possible locations for each.

To get a better idea of the locations of the different restaurants, I decided to plot a heatmap of the different restaurants and the heatmap of the Italian restaurants and the comparison of the two can enable us to find regions of low dense Italian restaurants. We would ideally like regions of low dense Italian restaurants with a good density of restaurants since the presence of a good number of restaurants means that opening a new restaurant there could be profitable.

Once the region has been identified, I divided the region into different locations spaced 50 metres from each other and for each such location I calculated

- The number of restaurants within 250 metres.
- The distance of the closest Italian restaurant.

And from this data, I then decided to take only those locations that had less than 2 restaurants within 250 metres and the closest Italian restaurant was at least 400 metres away. Such locations are ideal because there would be less competition nearby. The result is the following with the blue dots denoting the possible location and the white circle as our region of interest,

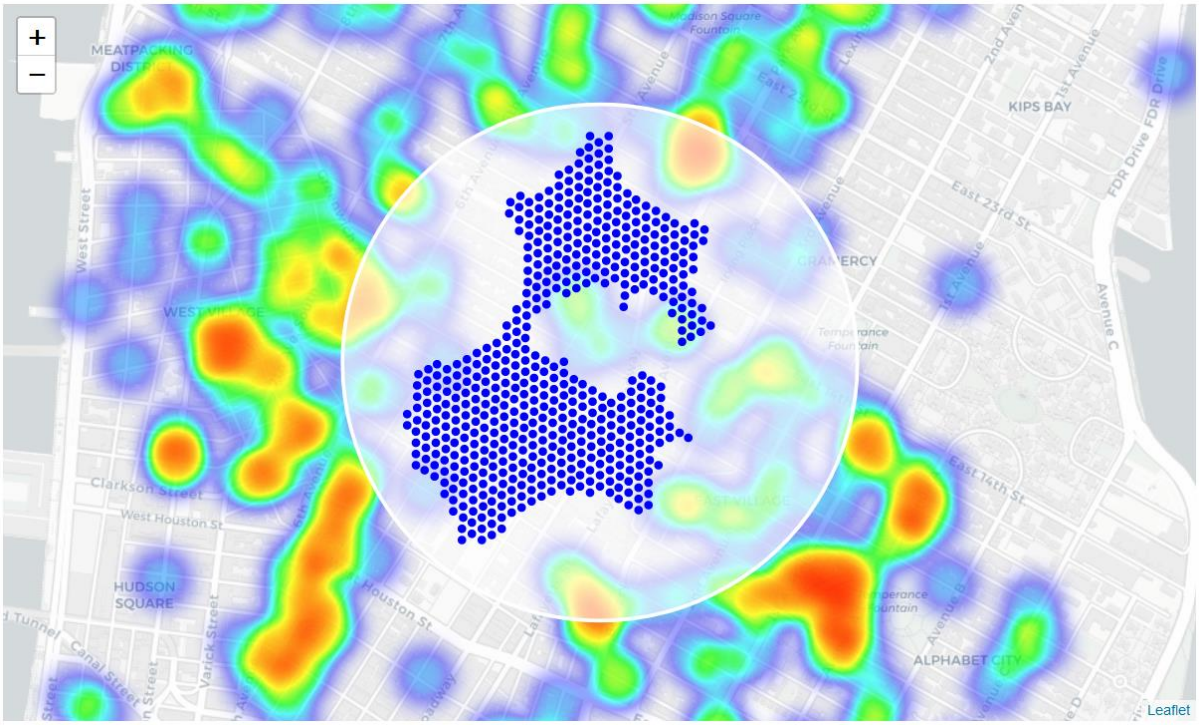

Fig 2. Map of Manhattan showing possible locations for an Italian Restaurant.

Since the number of such locations was around 500, I decided to group the locations into 10 clusters again using the k-means algorithm and get the cluster centers and reverse geocode these locations using Google API and give these 10 addresses as the output to the stakeholders. The stakeholders can then do the final street-level exploration around these addresses to find the best possible location according to their needs.

Results

I was able to come up with 10 different addresses situated within the regions formed by my 2 clusters of neighbourhoods (10 each) which can be seen in the following figure.

These are the 10 candidate addresses from our cluster 0 neighborhoods

```
for lon, lat in cluster_centers:
    addr = get_address(google_api_key, lat, lon)
    print(addr)
```

```
7 E 9th St, New York, NY 10003, USA
89 5th Ave, New York, NY 10003, USA
260 Greene St, New York, NY 10003, USA
780 Broadway, New York, NY 10003, USA
2 Union Square E, New York, NY 10003, USA
76 5th Ave, New York, NY 10011, USA
44 West 4th Street, New York, NY 10012, USA
18 Washington Square N, New York, NY 10011, USA
Union Square Park, 201 Park Ave S, New York, NY 10003, USA
13 Astor Pl, New York, NY 10003, USA
```

```
for lon, lat in cluster_centers:
    addr = get_address(google_api_key, lat, lon)
    print(addr)
```

```
341 Madison Ave, New York, NY 10017, USA
22 E 50th St, New York, NY 10022, USA
1180 6th Ave, New York, NY 10036, USA
66 E 46th St, New York, NY 10017, USA
33 W 42nd St, New York, NY 10036, USA
6 Av/W 48 St, 6th Ave, New York, NY 10020, USA
14 E 47th St, New York, NY 10017, USA
16 W 51st St, New York, NY 10111, USA
280 Park Ave # 27e, New York, NY 10017, USA
511 5th Ave, New York, NY 10017, USA
```

Fig 3. The resulting addresses

Fig 4.1

Fig 4.2

Fig 4. 4.1 shows the heatmap of restaurants and 4.2 shows the heatmap of Italian restaurants

Discussions

One of the observations that I made was that in one of the regions there was a fair few restaurants around but only a few Italian restaurants around, which could mean that the whole region as such could be a place to start while looking for a new place to open a restaurant (See Fig 4). From the addresses that has been generated it would be a good idea to do some exploration around these addresses to decide on the final location.

We could also use the same method if we are interested in opening a new mall, theatre, gym, diner, etc. and so that could be one of the improvements possible.

Conclusion

In this study, I looked at the problem of identifying an optimal location to open a new Italian restaurant in Manhattan, NY. I used the data about the various restaurants and Italian restaurants available in Foursquare to build a model that gives us locations where the number of restaurants and Italian restaurants in the vicinity is limited by pre-set conditions. This model can be very useful for those looking to open new businesses.