Analysis on Venue Ratings

Introduction

Background

Foursquare is a platform where users can visit different venues and provide feedback on these venues. The venues that are analysed can include restaurants, tourist attractions, gyms and various other retailers and service providers. This gives other users an idea of the quality of each venue.

In this investigation, we will be analysing the venue ratings in the area of Queens, New York. With rating data joined with location data, we hope to identify areas of interest with groupings of low, mid and high levels of venue qualities.

Interest

The findings from this investigation will be of interest to various parties. This could be people looking for an area to live, tourists looking for good places to visit, or people looking to open a business in a particular location based on competition or the reputation of the surrounding venues. In this report, we will be focusing on the use case of deciding on a place to live.

Data

Data collection

In order to draw conclusions on this problem, we will require the following information:

- Neighborhood information for Queens, New York, **provided by Coursera**
- Venue locations **provided by Foursquare**
- Ratings for each venue **also provided by Foursquare**

To obtain this information, we used the New York neighborhood information and obtained up to 20 venues within a 500m radius of each neighborhood in Queens. Neighborhoods in Queens are shown in the image below.

Figure 1: Map representation of neighborhoods in Queens, New York

Data processing

From the venues collected, we only used venues which had been rated on Foursquare. On top of this, we have the limitation from Foursquare, where there is a quota for the number of calls that can be made per day (I was unable to get around the bug of creating an app using a verified personal account, so have used the Sandbox account). I could only obtain the rating data for a very limited number of venues.

Due to the lack of data, the scope must be reworked to accommodate for what is available. From the image below, we can see that there is a huge discrepancy between venues available (left) and venues with collected data (right). The new scope will only be related to the areas with data points on the right image.

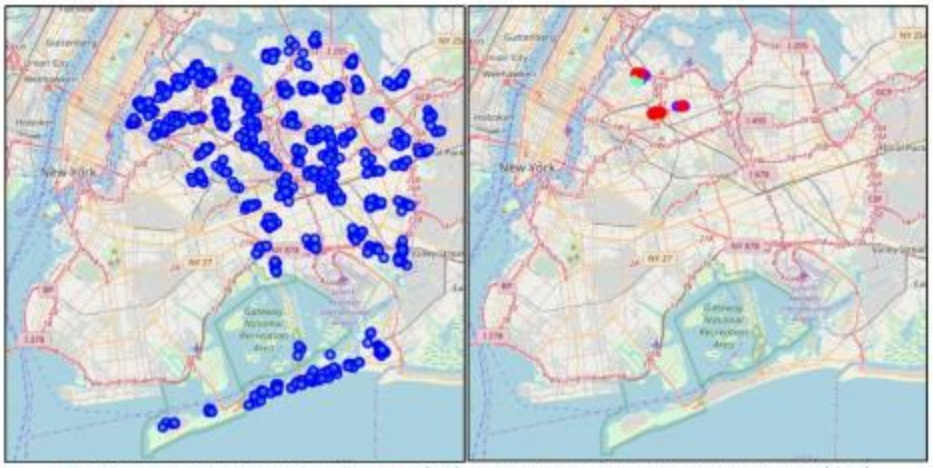

Figure 2: Venues found for each neighborhood (left). Venues where ratings data could be collected (right)

A request was sent for each neighbourhood to get the top 20 venues. Each venue then had a request sent to obtain the rating. The resulting dataframes were then merged together based on the unique venue ID. These two dataframes are shown below.

	Neighborhood	Venue	Venue ID	Venue Latitude	Venue Longitude
0	Astoria	Favela Grill	4bdf502a89ca76b062b75d5e	40.767348	-73.917897
1	Astoria	Orange Blossom	52c580e8498eddd52d925dd9	40.769856	-73.917012
2	Astoria	Titan Foods Inc.	4a9c0105f964a520b03520e3	40.769198	-73.919253
3	Astoria	CrossFit Queens	4c94d26d58d4b60c40fc2b29	40.769404	-73.918977
4	Astoria	Simply Fit Astoria	4d7ce85486cfa14365a2d2a0	40.769114	-73.912403

	Venue_ID	rating
0	4bdf502a89ca76b062b75d5e	8.6
1	52c580e8498eddd52d925dd9	8.1
2	4a9c0105f964a520b03520e3	9.3
3	4c94d26d58d4b60c40fc2b29	8.9
4	4d7ce85486cfa14365a2d2a0	8.5

Figure 3: Dataframes containing all values that were collected

Methodology

In this project, we aim to identify the general quality of venues in specific locations. The area of interest was Queens. A 500m radius around each neighborhood was used to search for venues, with a maximum of 20 venues per neighborhood. However, even with the limit, only a small proportion of data was able to be collected. Therefore, in the analysis, we need to take into account unavailable ratings and restrict the scope to only the areas where data was collected.

In order to carry out the investigation, we acquired the name, coordinates and ratings for each venue. This data was then clustered using k-means clustering method on the ratings to divide the venues into three tiers of quality, low medium and high.

We then use a map to visualise the clusters and their locations. From this, we can make conclusions regarding which areas have a higher density of venues at different quality levels and explore which locations would be best to live in.

We then found the mean for each cluster to better understand what kind of quality each cluster represents. The output of cluster means is shown below.

```
Cluster Labels
0    7.800000
1    8.447368
2    9.000000
```

Figure 4: Output of mean ratings for each cluster

From the results, we can see that:

- Cluster 0: low rated venues
- Cluster 1: mid rated venues
- Cluster 2: high rated venues

And from the map, we can see the cluster distribution (High rated venues: blue, mid rated venues: purple, low rated venues: red).

Figure 5: Distribution of clusters on a map. Highly rated (blue), mid rated (purple), lowly rated (red)

Results and Discussion

From the data that is available, we can see that there are three distinct areas where data was able to be collected. When we compare this with the raw number of venues found, this investigation does not contain enough data to make any meaningful conclusions based on all of Queens. If we limit our investigation to these three areas with data, we can only rank the areas in order of preference, relative to each other.

Figure 6: Ratings data grouped into three areas for comparison

In Area 1, we have the most number of venues belonging to cluster 2, high quality venues, whereas area 2 has the highest concentration of venues, but most of them are belonging to cluster 0, indicating that venues in this area are not of a very high standard. The third area has a combination of mid and low quality venues.

From these observations, we can conclude that the order of preference for locations to live around would be area 1, area 3, and then area 2 if we base our decision on the quality of venues in the area.

Conclusion

From this investigation, we found:

- With the limited data obtained from Foursquare, venues found were clustered into three distinct areas (indicated on the map)
- Clustering this data into groups of low, medium and high quality venues based on user ratings, the mean of each cluster was:
  0: 7.800000
  1: 8.447368
  2: 9.000000
- Area 1 contained the highest proportion of high quality venues, while still containing a mixture of low to mid-range venues in the northern side.
- Area 2 contained the highest density of venues, most of them being low range venues.
- Area 3 contained a mixture of low and mid-range venues, with no high range venues.
- The order of preference for locations to live based on venue quality would be area 1, area 3, then area 2.