1. Introduction

1.1. Background

Living in Singapore for 1 year and a half, I recently moved out my apartment to settled at Newton, in a better apartment (with a balcony), closer to my workplace and located in a better condo. Although I was very happy with that choice (and with the money I saved on rent), it entailed a vey sad consequence. I was moving away from the infamous restaurant Swee Choon, labeled (by myself) best Dim Sum restaurant in Singapore. BUT moving closer to Newton meant having a minute away from my flat the delicious hawker's food (featured in the movie Crazy Rich Asian). My heart was at a cross-road.

But what if… data science could settle this debate for me? Was it worth it leaving the fabulous Swee Choon neighborhood near Kallang for the Newton/Novena area?

And even better, what if I could apply some machine learning to determine where I should head for my next apartment?

1.2. Problem

Is it then possible to harness the different open source data APIs and data science methodology to determine the characteristic of each Singaporean neighborhood (based on a clustering approach) and then use a collaborative-filtering approach to build a recommender system that will point out the best choices of neighborhood (to live their or just to head for your next foodie adventure)?

1.3. Interest

The quick answer is yes, it is absolutely possible. But the ways to do it are very interesting. I will walk you through my methodology and code I used to build this model. What we are trying to achieve is basically the Netflix for food (which are two very important hobbies for Singaporeans (?)): what can we recommend you based on experience of similar users in different neighborhood of Singapore, food-wise. I am not saying this is a business model for the next billion-dollar start-up, but I'm not saying it's not, either.

Let's be honest, food is a big deal in Singapore, and foodies want to know more than just where to find the best chicken rice. With this model, I hope to bring some relevant insights and metrics to the topic. With this tool, not only will you have the hotspot of food-scene in Singapore, you would know where to go exactly to satisfy you cravings.

2. Building the Clusters

2.1. Data where you describe the data that will be used to solve the problem and the source of the data.

2.1.1. Data source

The source of the data is twofold:

First, we use the Airbnb data, easily accessible on their website, to build the neighborhood of Singapore.

Then, to source our data for clustering, we use the API of Foursquare. This gives us information about Singaporean neighborhood and the features (types of restaurants, etc).

2.1.2. Data cleaning

We keep only the relevant data from Airbnb: Neighborhood, latitude, longitude.

From the Foursquare API, we keep the venue, category, the latitude, the longitude.

2.1.3. Features selections

The features that were selected were all related to Food and Beverage industry. Thus, we only keep the category value that contains a "restaurant" in the name.

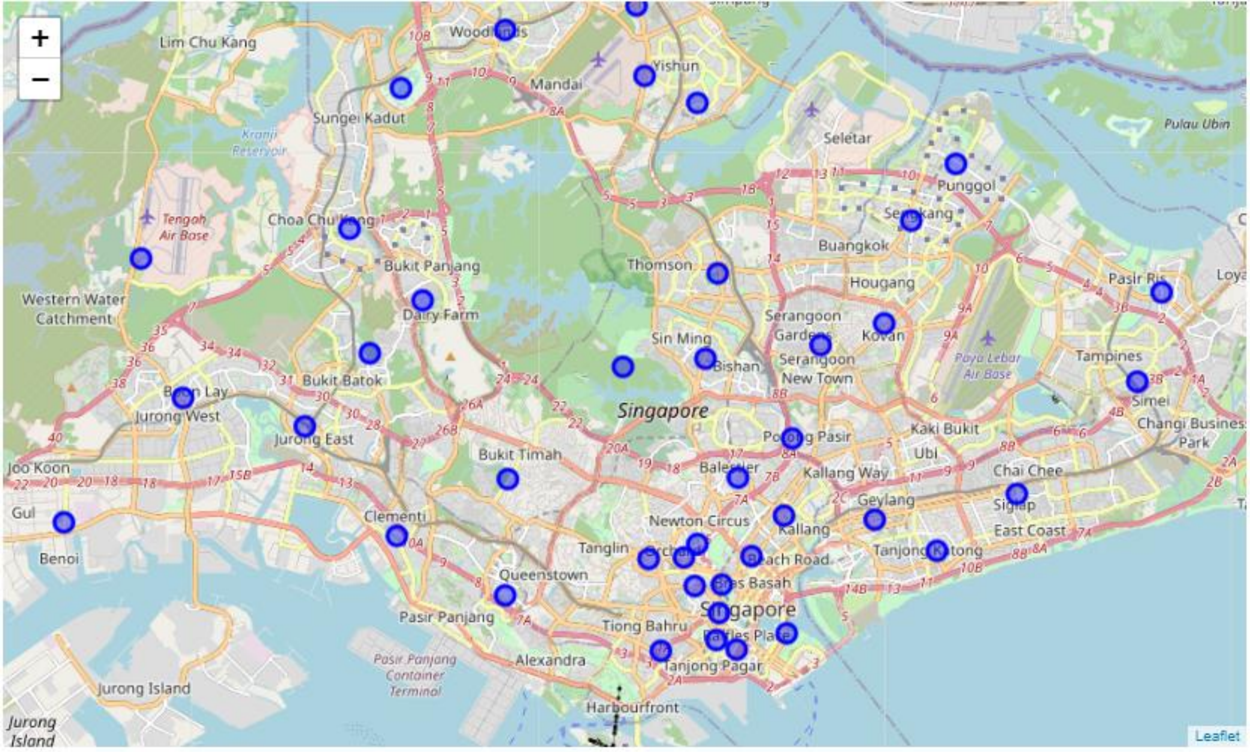
2.2. <u>Methodology section.</u>

2.2.1. Exploratory

We merge the different pieces of information, in order to build a data frame containing the neighborhood, latitude, longitude, and the number of occurrences for each category.

We explore the Ang Mo Kio neighborhood to see what the data frame looks like. Given the high number of categories (even after filtering the non-restaurant related), it seems wise to only select the 5 top venues of each neighborhood.

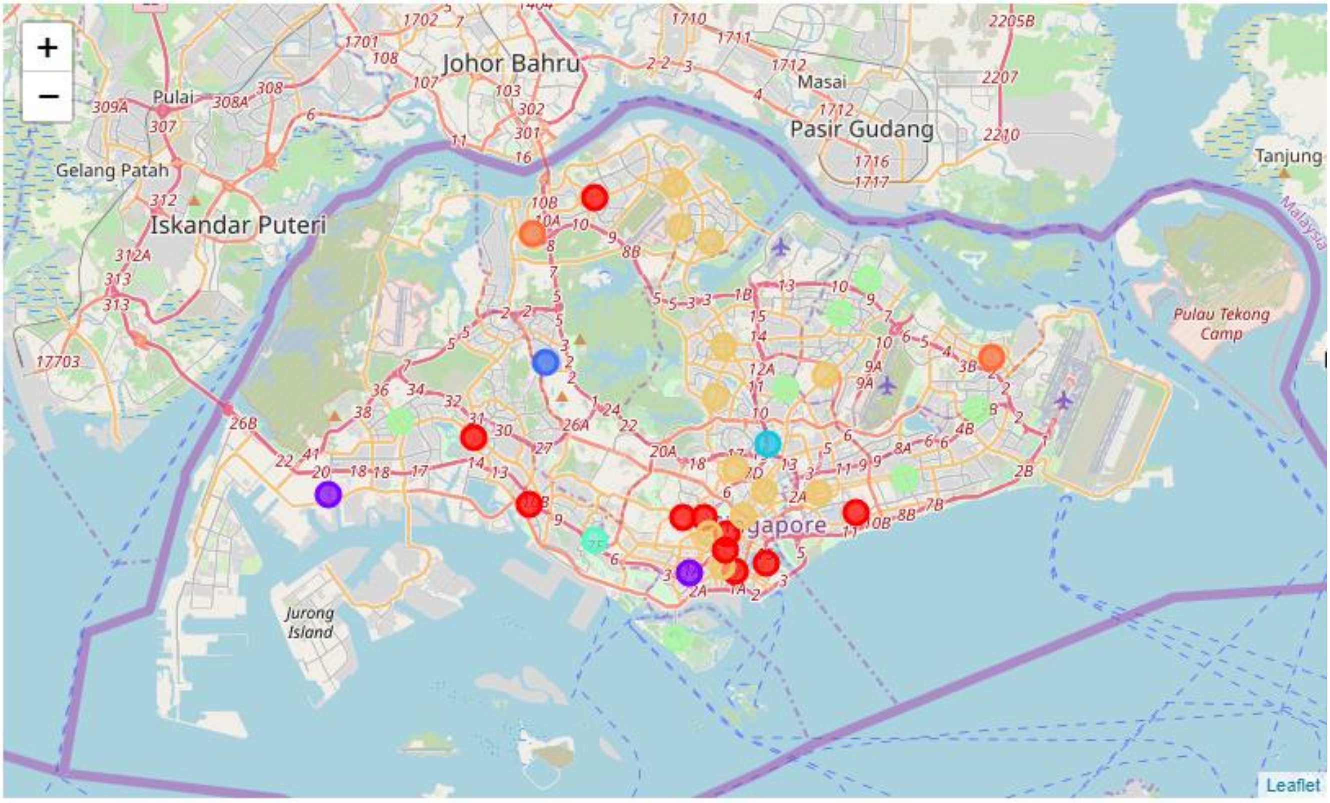
We also control the creation of neighborhood in Singapore:



2.2.2. Modeling

We use the K-means clustering method, creating 8 clusters (for next to 48 neighborhood) and the standard 300 iteration to determine the distribution of those clusters.

Once created, we applied those 8 clusters to the map, and create a color distinction (see below).



3. Building my recommender system

3.1. Data where you describe the data that will be used to solve the problem and the source of the data.

3.1.1. Data source

The data used in the second part relies on the clusters identified in Part I (and their characteristic), and a new data frame (user input) to capture my personal preferences (necessary to build the content-based recommender system).

3.1.2. Data cleaning

No data cleaning is needed as we rely on a clean data frame from Part I.

3.2. Methodology section

3.2.1. Exploratory

We have a detailedl view of the distribution of each characteristic for each clusters (sample below):

Cluster Labels	Chinese Restaurant	French Restaurant
0.0	1	0.0
1.0	1	0.0
2.0	0	0.0
3.0	0	0.0

3.2.2. Modeling

We transpose the personal rating matrix with the cluster characteristics matrix to perform the content-based recommendation.

We have the results as follow:

```
Cluster Labels
0.0    1.000000
6.0    0.942857
5.0    0.800000
1.0    0.594286
2.0    0.400000
7.0    0.308571
4.0    0.308571
3.0    0.000000
```

4. Results section where you discuss the results.

The results are twofold. The clustering enables to cluster the different neighborhood of Singapore together (hence knowing which neighborhood are similar or dissimilar).

Then, based on this clustering, and after inputting my own preferences of clusters that I know, the algorithm able to predict which cluster (i.e which neighborhood) I don’t know already will fit my taste.

5. Conclusion section where you conclude the report.

As a conclusion, we can answer the many questions I raised at the beginning of this report.

- Yes, it is possible for data science to bring me a satisfying answer to my existential problem: the new cluster (cluster number 5, incl. Novena) is better rated (score of 0.800000) by the algorithm than the previous cluster (cluster number 4, incl. Kallang, score of 0.308571). Hence, it was a wise choice – foodwise – to change apartment.

- Yes, the model did recommend different neighborhood in the new well rated clusters, which will be good choices of settlement.
- At a larger and crowdsourced level, it would be totally possible to build a “Netflix for restaurants” (i.e. “Nestaurant”) which would recommend toward which neighborhood/type of neighborhood a user should head, based on its previous experiences (if you are a Venture Capitalist, please contact me, this is a billion dollar start up idea).