The Battle of Neighborhoods
Shopping Mall Location Selection in St. Louis *

Ivan Yu

May 12, 2020

1 Background

St. Louis is an independent city in Missouri. It is the largest metropolitan area in the state of Missouri; however, the independent city of St. Louis is the second largest city in Missouri, behind Kansas City. It is situated along the western bank of the Mississippi River, which forms the state line between Illinois and Missouri.

For residents, people enjoy shopping at shopping malls, dining at restaurants, watching movies at cinemas, and doing many other activities in their spare time to relax themselves. For government, shopping malls are very important in a city's business systems and provide a lot of purchasing power. For property developers, shopping malls can be the most profitable investment. Running a shopping malls in appropriate location allows property developers to earn considerable rental income as well as administrative fee. However, an arbitrary location selection may result the poor operation of shopping malls and lead to bankrupt.

To sum up, opening a new shopping mall requires serious consideration, especially the location of the shopping mall. It is one of the most important decisions that will determine whether the mall will be a success or a failure.

2 Business Problem

In this project, we are going to analyze the distribution of shops, restaurants, cafe, etc. in St. Louis, and select the best locations to open a new shopping mall.

For a specific location, we will try to figure out questions like whether there are too much competitors, whether the business is flourishing in this area, etc. Finally, our project will provide suggestion to investors who are looking to open a new shopping mall on location selection.

*IBM Data Science Capstone Project

3 Description of Data

To solve the problem, we need the following data:

- **The list of neighbourhoods in St. Louis**: This defines the scope of this project which is confined to the city of St. Louis.
  - **Data Source**: St. Louis (Wikipedia)
- **Latitude and longitude coordinates of those neighbourhoods**: This is required in order to plot the map and also to get the venue data.
  - **Data Source**: Geocoder package
- **Venue data**: Data related to shopping malls, restaurants, and cafe. We will use this data to perform clustering on the neighbourhoods.
  - **Data Source**: Foursquare API

4 Methodology

For the list of neighbourhoods in St. Louis, we will use *requests* and *beautifulsoup* packages to help extract the data from the Wikipedia page. For Latitude and longitude coordinates of neighbourhoods we already get, we will use *Geocoder* package to help retrieve the information about the latitude and longitude. As for venue data, Foursquare API is a good data source.

Foursquare is a location data provider. Using the RESTful API to retrieve data from Foursquare database is pretty easy. We can just simply create a uniform resource identifier, or URI, and append it with extra parameters depending on the data that we are seeking from the database.

To analyze the data, we will perform clustering using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for different venues.

We will use the result to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Which neighbourhoods have higher concentration of restaurant and some other entertainment facilities while which neighbourhoods have fewer. With these information, we can

answer the question: where is the most suitable location to open a new shopping mall we raised at the beginning.

5 Results

We cluster all the neighborhoods according to:

- **Caterings**: Restaurant, Burger, etc.
- **Entertainments**: Bar, Club, Theater, etc.
- **Cafes**: Cafe, Breakfast, Dessert, etc.

The result is shown in the following map.

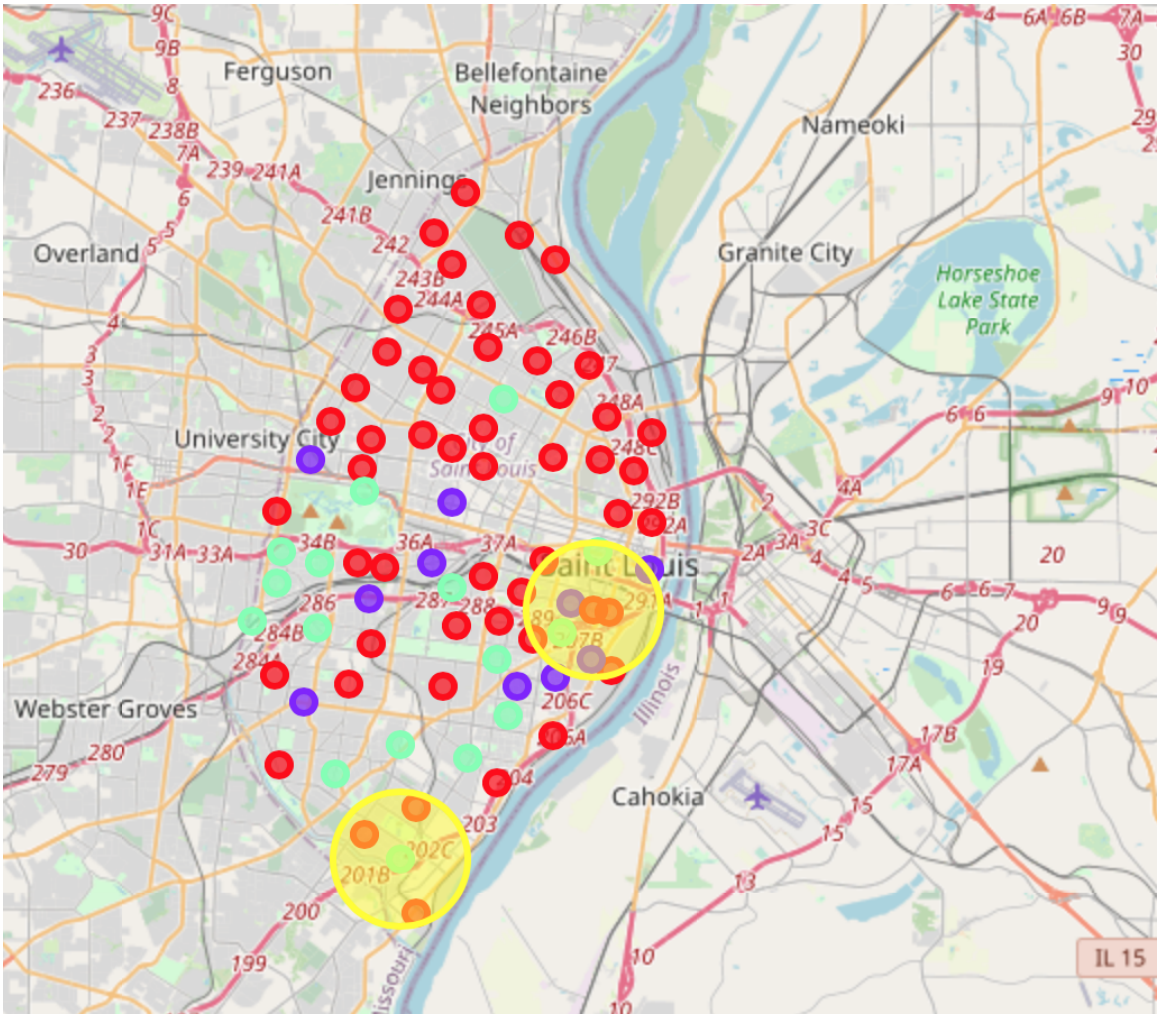

Figure 1: Map of St. Louis

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of Caterings, Entertainments and Cafes.

- **Cluster 0**: Least flourishing neighbourhoods
- **Cluster 1**: Most flourishing neighbourhoods
- **Cluster 2**: In the middle of **Cluster 0** and **Cluster 1**

In the map, red points represent those least flourishing(**Cluster 0**), purple points represent those most flourishing(**Cluster 1**), and mint green points represent those just middle(**Cluster 2**). We should open a new Shopping Mall at a flourishing location. So we should choose **Purple** point as our location.

Considering that we should avoid the potential competition with the existing Shopping Malls and Supermarkets, we mark out the existing Shopping Malls and Supermarkets in St. Louis on the map. And each has a 'Scope of Influence', which is represented by the larger yellow circle. We should choose the location outside these yellow circles.

In summary, we should choose Purple point outside the Yellow circles as the ideal location to open a new shopping mall.

6 Limitations and Suggestions for Future Research

In our research, we subjectivity define **Caterings**, **Entertainments** and **Cafes**. Besides, we manually filter venues to get these three parameters. Then we use these three parameters to cluster the neighborhoods. These can be problematic. The criterion we used to estimate where an area is flourishing may be unrobust. We can use some other criterion to exam whether our criterion can effectively cluster the neighborhoods. Also, we can refer to some research to find the proper criterion.

Another limitation in our project is that we do not do any research about the competition between our new shopping mall and the existing shopping malls. There is not available information about the 'Scope of Influence'. We just subjectivity choose a radius. In future research, this problem is deserved discussed in detail.

7 Conclusion

In this project, we go through the process of identifying the business problem, specifying the data required, extracting and processing the data, using machine learning algorithm (K-means clustering) to clustering the neighborhoods into 3 clusters based on their similarities in **Caterings**, **Entertainments** and **Cafes**. Finally, we provide our recommendation on new shopping malls location selection based the result we get, and visualize our result to help stakeholders better understand our recommendation. Our project will help property developers better make investment decision and help the government better establish city development plan.