Capstone Project - Clustering Bixi Stations Based on Nearby Venues

1. Introduction

Few days ago, I read that the City of Montreal has reached an agreement with the not-for-profit corporation Bixi Montreal to keep the bikes on city streets for the next 10 years and to extend the bike-sharing system to cover all 19 boroughs by 2028. Sixty new docking stations and 1,000 bicycles will be added to the network when the season begins in April. Reading this, I thought why not study the use patterns by clustering the end stations according to the nearby venues. This could come very handy to Bixi when deciding the locations of the new docking stations.

2. Data

Bixi Montreal made all the network use data available on their website here, so if you are curious about how many purchases and trips are taken by members and occasional users you can always check it out there in monthly format or even using the live feed. For this project, we acquired the trips data for last year, 2018, as well as the stations information as for names, codes and locations. The data is available directly in csv format. Trips are described through start and end stations, times, duration and whether the user is member or not. It is available in monthly format so we merge it all and keep only relevant information for our project before joining it to the stations csv to have the final dataframe.

For the venues available nearby docking stations, we are using the Foursquare API which allows us to explore and acquire venues locations and categories for a given location.

3. Methodology

It is a fair to assume that the main use of the Bixi bikes is related mainly to some activities, like shopping, visiting restaurant, coffees, commuting to work and so on... Since there are about 12000 docking stations covering almost all the city of Montreal, users tend to end their bike trips just close to their initial destination. This allows us to scan the surroundings of the end stations, in terms of venues, to have an idea about the use patterns and reasons of the Bixi users.

For this project, we are focusing the analysis on one particular station, the one closest to where I live now, near the corner St-André / Cherrier, with identifying code of 6175 (and obviously, this can be extended afterwards to all the others). Thus, we keep only the trips that started from our station and to analyze the most important and recurrent trips patterns, we focus on the most popular ones, i.e., we select the trips that went to the top 100 destination stations from our station. The Foursquare API comes in handy here to know what is around the end stations: surroundings venues, in a 100m radius, as well as their locations and categories are provided for each end station. We found 139 categories covering a wide range of activities like Arts & Entertainment, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport and others.

For each end station, top 20 most commons venues were selected and will be used as features for next step: clustering using K-means, expecting to capture use patterns in the trips' most common end stations.

4. Results

Using k=5 for k-mean clustering algorithm, we have the following distribution of stations per cluster:

```
Cluster 0: 15 stations
Cluster 1: 18 stations
Cluster 2: 62 stations
Cluster 3: 2 stations
Cluster 4: 2 stations
```

We can see here the most common venues for each cluster:

```
Most commom venues in cluster 0 are: ['Yoga Studio' 'Hot Spring' 'Dog Run' 'Intersection']
Most commom venues in cluster 1 are: ['Bakery' 'Supermarket' 'Gourmet Shop' 'Liquor Store' 'Grocery Store']
Most commom venues in cluster 2 are: ['Restaurant' 'Bar' 'Coffee Shop' 'Food & Drink Shop' 'Sandwich Place']
Most commom venues in cluster 3 are: ['Yoga Studio' 'Bus Station']
Most commom venues in cluster 4 are: ['French Restaurant' 'Dumpling Restaurant']
```

And finally, the following map illustrates how those cluster are distributed geographically:

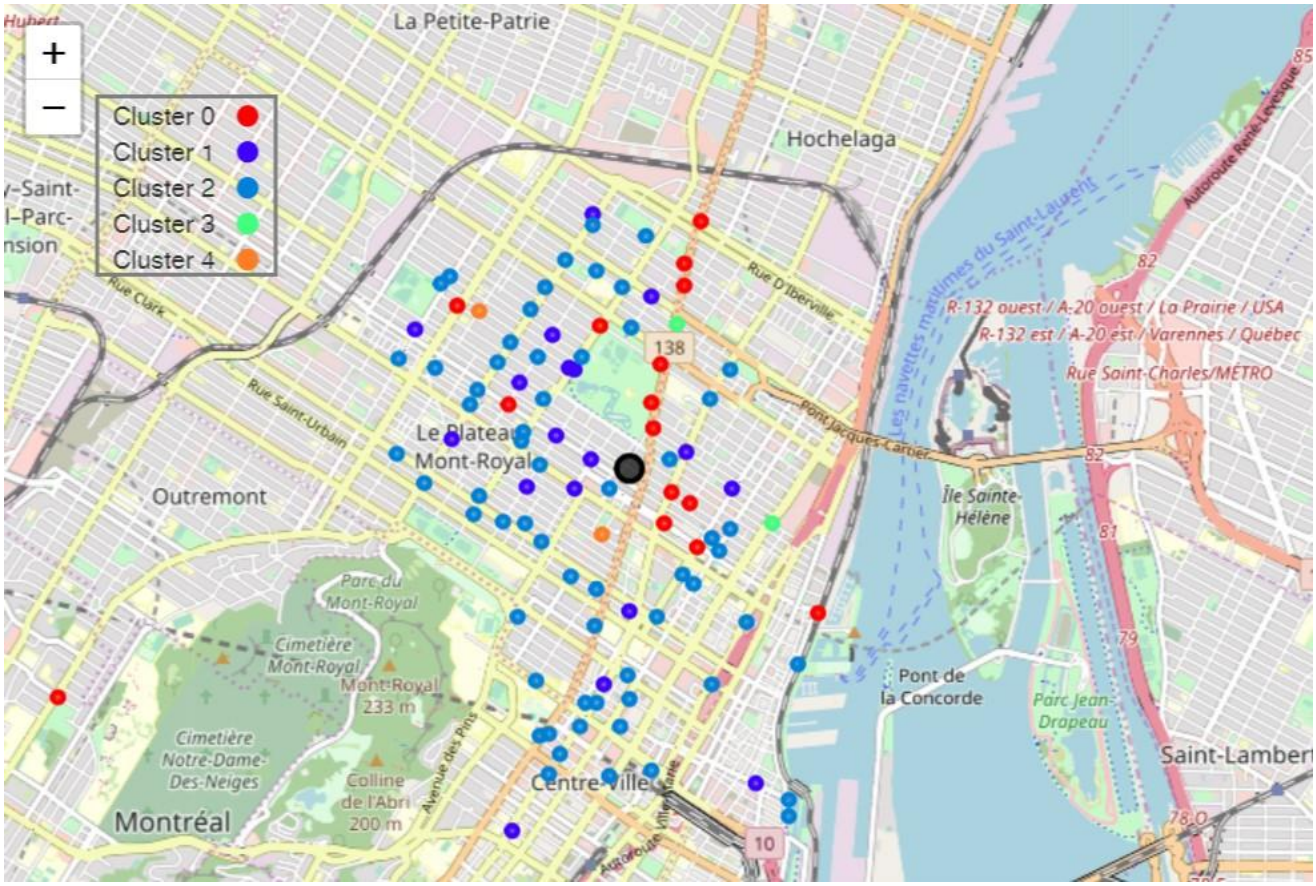

Figure 1: Stations' clusters distribution

5. Discussion

As we can see the biggest cluster, 2, is related to food and drinks shops. Cluster 1 is for markets and stores. Yoga studios and Hot spring are mostly in cluster 0. It is not that clear for the last two small cluster, since indeed, they're too small and have like random venues. I tried out other values for the k parameter but the same behavior keeps reoccurring: there is always one or two of these 'not so clear' cluster in addition to two or three clusters where the venues' category is quite discernable. All in all, the main clusters correspond to what we initially assumed, that bike use is mainly related to food& drinks, shopping and sporty activities. The clustering we got could be useful when planning for new deployments, the size of docking stations (how many bikes) and their locations could be then optimized based on the surrounding venues and their categories/ clusters.

6. Conclusion

This project is a quite interesting one since the idea come out while reading an online news article and I felt like I could make use of the Foursquare API venues for some interesting and potentially useful application: The results are quite promising and could certainly be generalized to all stations of the network while studying extension plans to draw insights regarding the location and size of new dockings stations based on and optimized to match the venues and their categories in each neighborhood/ street.