Author: Vanessa De Bellonia
Date: 2019-07-23

CAPSTONE PROJECT: Trending locations closest to bike-sharing docking stations

I. Introduction

As people are becoming more health and environmentally conscious, shared biking systems in metropolitan areas have become increasingly more popular over the years, with a steady increase in the number of users and docking stations.

Being a past user myself, I would always map out my routes ahead of time and made sure to check out what was nearby my destination station. Knowing the city, this task is easy. However, what if you are a visitor in a city and don't know where to go or what to do? Wouldn't it be great to have an easy way to find out the docking stations that have the highest number popular venues/activities in order to maximize your time within a new city you wish to discover?

For this capstone project, I will take a look at the most trending venues/activities within a 100 meter radius of a docking station in the city of Montreal. This information is useful for tourists visiting the city, who wish to discover the city using the shared-biking system Bixi, and are time-limited. However, the methodology used to discover the most trending locations in Montreal can be applied to other cities with a bike-sharing system such as Bixi. So the question to be answered through this project is:

What are the docking station locations in the city of Montreal that have the most trending venues/activities nearby?

II. Data

The dataset that will be required to answer the question stated in the introduction includes the locations of the docking stations, which is open data available on the Bixi website of Montreal in .csv format. The .csv file contains the most recent list of the docking stations in 2019 for the Bixi bikes, with details about each station (station code, station name, latitude and longitude). This information can be downloaded from the Bixi website: https://www.bixi.com/en/open-data.

Foursquare location data will be used as well to determine the most popular venues/activities in the city of Montreal. The location data will then be used to determine the most trending venues within a 100 m radius of each of the Bixi docking station locations provided in the Bixi dataset. The venues will then be analyzed by location in order to determine the docking station locations with the most trending venues. As a final result, the venues will be clustered by location using k-means clustering and will then be mapped by cluster along with the nearest Bixi docking station.

III. Exploratory Data Analysis

I begin exploring the data by importing the locations and docking station names from from the Bixi website and saving it to a dataframe named stations_df, using Pandas. The orginal file is in .csv format.

The resulting dataframe shows that there are a total of 615 docking stations for Bixi bikes in the Montreal area. I named it stations_df.

Next, I use geopy library to retrieve the latitude and longitude values of the city of Montreal. The geographical coordinates of the city of Montreal are 45.4972159, -73.6103642. This is important as these coordinates will be used to get the trending venues from Foursquare.

Next, we are going to start utilizing the Foursquare API to explore the trending venues in Montreal and segment them. After defining my Foursquare Credentials and Version, I was able to retrieve the top 500 trending venues within a 20 km radius of Montreal center (coordinates mentioned above).

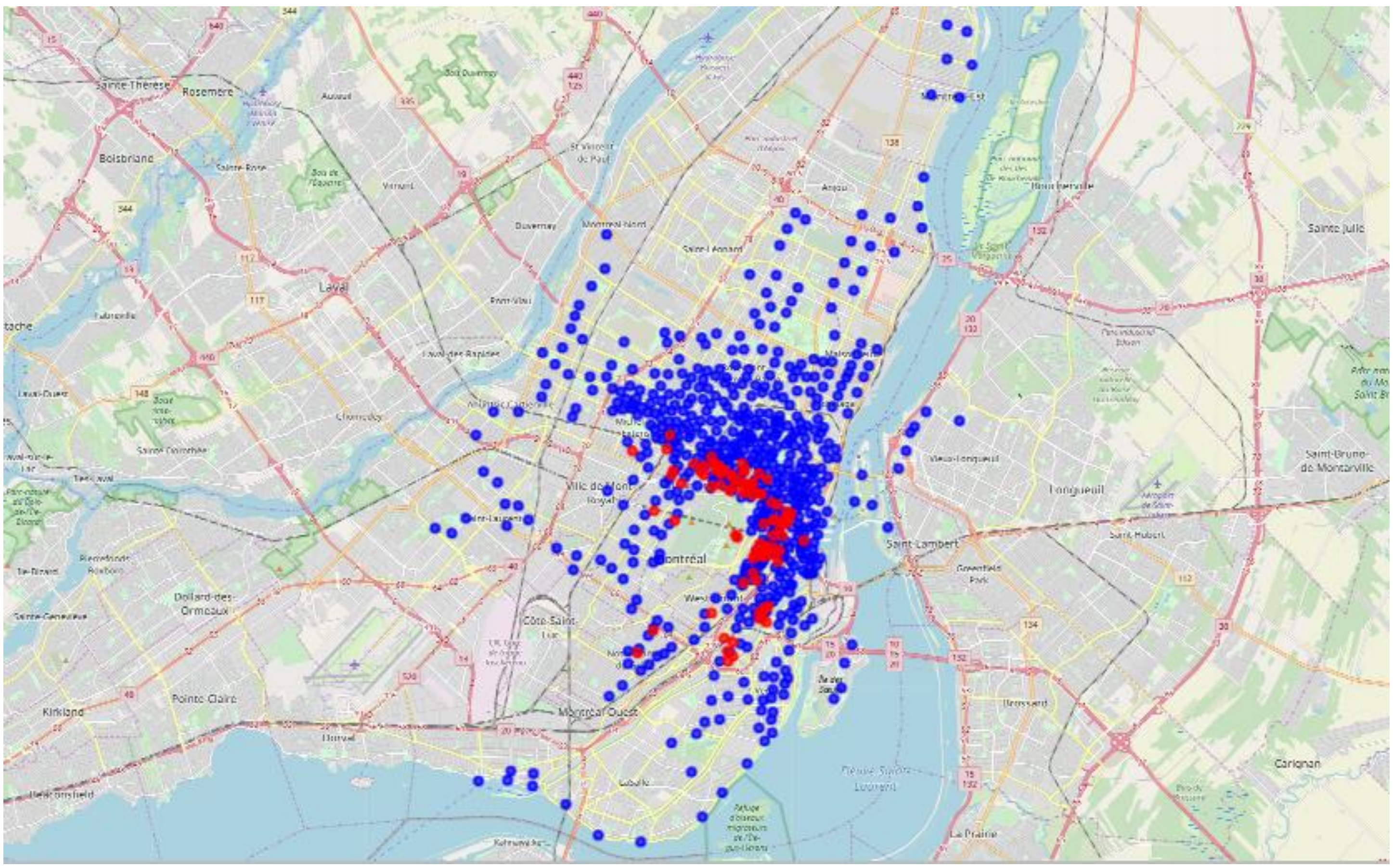
I then sent the GET request and examined the results by using a user-defined function called *get_category_type* function which I have borrowed from a previous lab. With this function I was able to extract the pertinent information from each venue. The result is a .json file.

I then used Pandas to clean up the .json file and put it in a dataframe called mtl_venues. The resulting dataframe shows 100 venues, their coordinates and the category type for each venue.

Despite the fact that I set a limit of 500 venues within a 20 km radius, Foursquare returned only 100 venues because that is the maximum supported amount of return values.

Now that we have all the data we need and it is cleaned up and stored in two different dataframes, let's start exploring it visually.

I then created a map using Folium to visualize the locations of the Bixi docking stations (in blue) and the trending venues (in red). The result is the following figure.



Based on the map, it is clear that that are a lot more bixi docking stations (615) than trending venues (100), and the trending venues are concentrated in Montreal's downtown core.

Therefore, it would be relevant to filter the docking stations and analyze the ones that are within 100 meters of a trending venue.

In order to do this, I first checked the data types for each dataframe, to ensure they are in the correct format. Then I created a user-defined function is called *haversine* and will be used to calculate the distance between two locations using latitude and longitude coordinates. It uses the Haversine formula, which considers the globe as a sphere.

I then created a process to determine the distance between each venue and docking station, return only the venues and stations within 100 meter radius, and store all this information into a new dataframe. The process iterates for all the venues and docking stations and calculates the distance between each one, returns only values within a distance of 100 meters, and stores all this information into a new dataframe called locations_df.

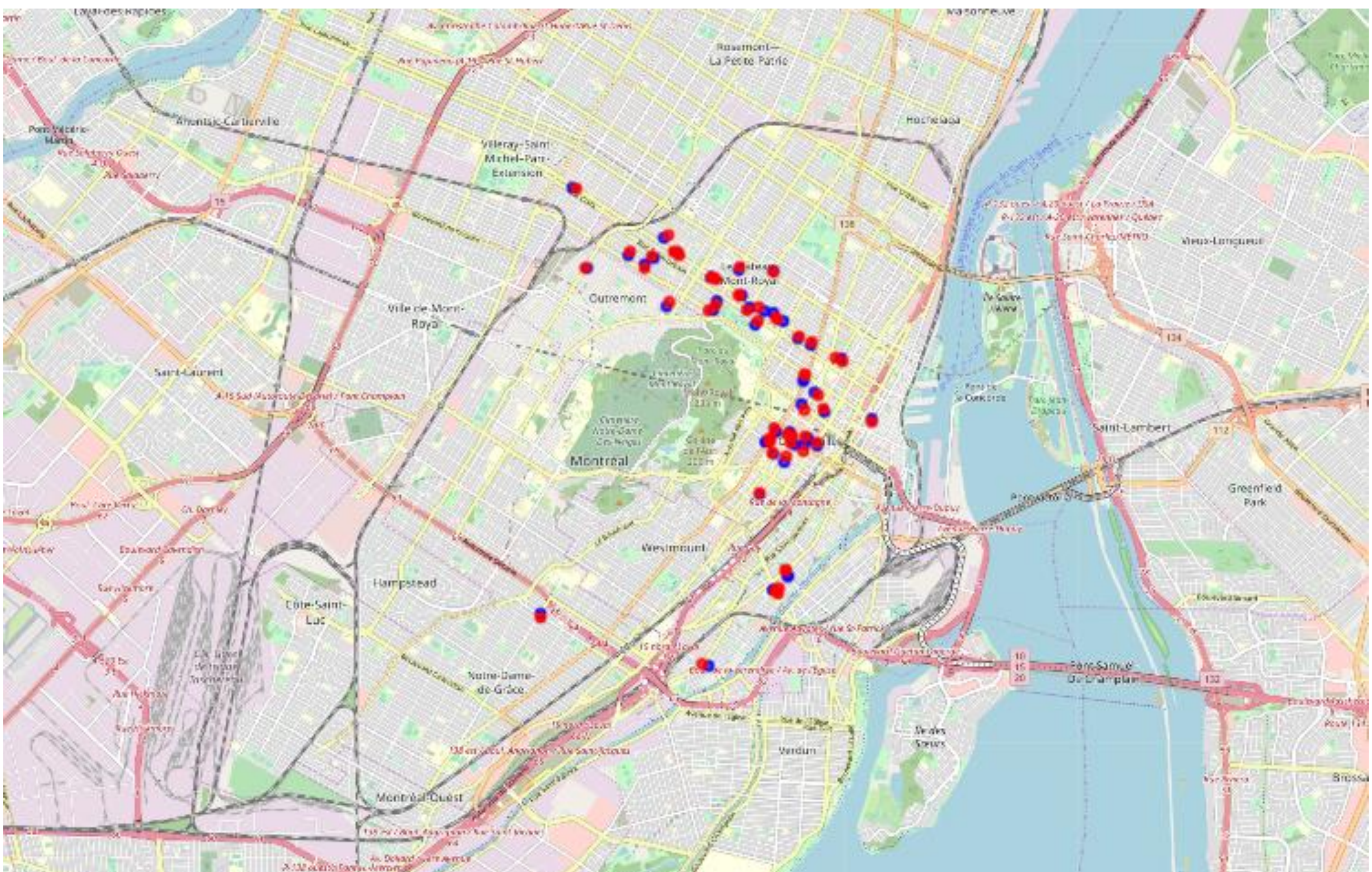
The resulting dataframe shows there are 65 venues that are within 100 meters of a docking station. This information also tells us that there are 35 venues that are not within 100 meters of a docking station. It also shows the following information:

- There are 37 unique categories.
- There are 55 unique venues.
- There are 48 unique docking station locations.

I then continued to further explore this information by filtering the dataframe in order to keep only unique venues. Out of the duplicates, only the one with the smallest distance will be kept. To do this, I first sort the dataframe by distance in ascending order, then drop the duplicates and only keep the first (which will have the smallest distance). We have now a dataframe with 55 unique venues, and the closest docking station to each venue.

After filtering and examining the data, it shows us that we 55 unique venues and 48 unique docking station locations. Some stations will have up to 3 nearby trending venues and some venues have several stations nearby. Our objective is to get the docking station locations with the most venues nearby so let's see if we can cluster the venues into a *K* number of clusters and determine the closest station to each cluster center. This will be the final result of our analysis.

The following figure show the current results of the filtered data.



First, I will use k-means clustering with HDBSCAN for this task in order to determine the cluster of venues by location. What's great about HDBSCAN is that it takes into account the distance between geo-locations, as the earth isn't flat. It considers the

Haversine function we used previously to determine the distance between venues and docking station locations. Also, I won't need to specify a *K* number of clusters, it will return the best number of clusters according to a minimum number of items in a cluster, that I will specify as 3.

Using *k-means* for clustering stations

What is important to note is that using HDBSCAN does not return cluster centers for each cluster. It returns only segmented clusters, which I will visualize using Folium map and then visually determine the most representative station location (based on location to each venue in the cluster, and the number of closest venues). This essentially will be the final result we are looking for: the location of docking stations that maximize the amount of trending venue nearby.

The result of the k-means algorithm returns a total of 8 clusters. Unsurprisingly, we end up with several points being flagged as noise (where the cluster label is -1). Since the minimum number of points for a cluster was set to 3, isolated locations are categorized as noise.

Our goal is to identify the docking station locations with the most nearby venues. Therefore, it makes sense to drop the isolated locations (venues where *Cluster = -1*).

We'll continue to further explore the data to see how we can merger clusters/locations in a meaningful way.

Let's explore each individual cluster.

Cluster 1

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
0	Liverpool House	45.482893	-73.575256	Restaurant	Duvernay / Charlevoix	45.482040	-73.574863	99.697293	0
1	Satay Brothers	45.480061	-73.577394	Asian Restaurant	Marché Atwater	45.480208	-73.577599	22.883836	0
2	Le Vin Papillon	45.482763	-73.575514	Wine Bar	Duvernay / Charlevoix	45.482040	-73.574863	95.059025	0
3	Marché Atwater	45.479790	-73.576775	Market	Marché Atwater	45.480208	-73.577599	79.278863	0
4	SAQ Sélection	45.480446	-73.576695	Liquor Store	Marché Atwater	45.480208	-73.577599	75.253797	0

Cluster 2

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
5	naada yoga	45.526864	-73.597479	Yoga Studio	St-Dominique / St-Viateur	45.526557	-73.598276	70.886785	1
6	Drawn & Quarterly	45.524748	-73.604614	Bookstore	Bernard / Jeanne-Mance	45.524286	-73.604973	58.393492	1
7	Salon de thé Cardinal / Cardinal Tea Room	45.524721	-73.596464	Tea Room	Maguire / St-Laurent	45.524628	-73.595811	51.894803	1
8	St-Viateur Bagel (La Maison du Bagel)	45.522573	-73.601878	Bagel Shop	Jeanne-Mance / St-Viateur	45.523026	-73.601840	50.538247	1
9	Pizzeria Magpie	45.524711	-73.595744	Pizza Place	Maguire / St-Laurent	45.524628	-73.595811	10.583038	1
10	Boucherie Lawrence	45.524232	-73.595292	Deli / Bodega	Maguire / St-Laurent	45.524628	-73.595811	59.768675	1
11	Café Olimpico	45.524116	-73.600442	Café	Waverly / St-Viateur	45.523856	-73.600127	37.976787	1

Cluster 3

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
12	Pikolo Espresso Bar	45.508606	-73.571772	Coffee Shop	Hutchison / Sherbrooke	45.507810	-73.572080	91.680589	2
13	Papeterie Nota Bene	45.508571	-73.571672	Paper / Office Supplies Store	Hutchison / Sherbrooke	45.507810	-73.572080	90.423304	2
14	Café Parvis	45.505817	-73.569302	Café	de Maisonneuve / City Councillors	45.506190	-73.569954	65.641906	2
15	Empire	45.503952	-73.571840	Sporting Goods Shop	du President-Kennedy / Robert-Bourassa	45.504627	-73.572325	84.019077	2
16	Il Focolaio	45.504009	-73.568213	Pizza Place	Square Phillips	45.503738	-73.568106	31.290720	2

Cluster 4

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
17	Cat's Corner	45.512893	-73.570403	Dance Studio	Milton / Clark	45.512541	-73.570677	44.618930	3
18	Bouillon Bilk	45.510845	-73.566017	Restaurant	Métro St-Laurent (de Maisonneuve / St-Laurent)	45.510660	-73.564970	84.116840	3
19	Café Nocturne	45.513566	-73.572849	Café	Clark / Prince-Arthur	45.513303	-73.572961	30.537609	3

Cluster 5

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
20	Tiffany & Co.	45.499782	-73.578403	Jewelry Store	de la Montagne / Sherbrooke	45.499745	-73.579034	49.303327	4
21	Mandy's	45.498286	-73.577895	Salad Place	Crescent / de Maisonneuve	45.498112	-73.577615	29.179987	4
22	Sofitel Montréal Le Carré Doré	45.501499	-73.577526	Hotel	Stanley / Sherbrooke	45.501041	-73.577178	57.728634	4
23	The Ritz-Carlton Montréal	45.500191	-73.578127	Hotel	de la Montagne / Sherbrooke	45.499745	-73.579034	86.358319	4
24	Maison Boulud	45.500062	-73.578324	French Restaurant	de la Montagne / Sherbrooke	45.499745	-73.579034	65.604786	4

Cluster 6

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
25	Louis Vuitton	45.497820	-73.575306	Boutique	Crescent / Ste-Catherine	45.497092	-73.575549	83.154334	5
26	Dominion Square Tavern	45.500405	-73.571636	Gastropub	Metcalfe / du Square-Dorchester	45.500208	-73.571138	44.564410	5
27	Club Sportif MAA	45.500700	-73.574817	Gym	de Maisonneuve / Peel	45.500951	-73.574578	33.489149	5
28	Ferreira Café	45.500434	-73.574060	Portuguese Restaurant	de Maisonneuve / Peel	45.500951	-73.574578	70.226903	5
29	Enso Yoga	45.500743	-73.574943	Yoga Studio	de Maisonneuve / Peel	45.500951	-73.574578	36.641021	5
30	Frank & Oak	45.499573	-73.574328	Men's Store	Stanley / Ste-Catherine	45.499344	-73.573760	51.069069	5

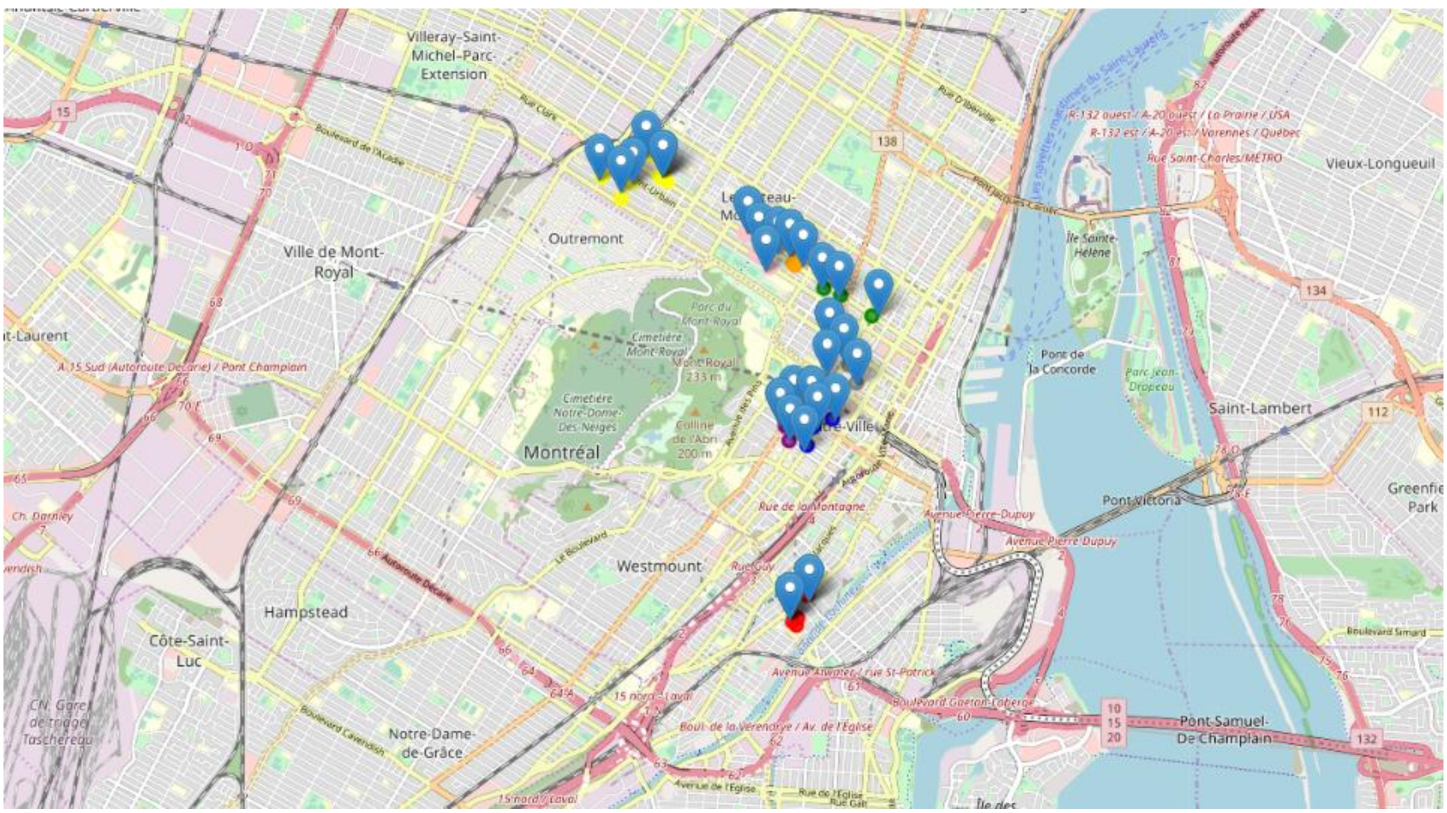
Cluster 7

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
31	Omnivore Comptoir Grill	45.519061	-73.584116	Middle Eastern Restaurant	Vallières / St-Laurent	45.518967	-73.583616	40.295889	6
32	Darling	45.518995	-73.584061	Lounge	Vallières / St-Laurent	45.518967	-73.583616	34.825866	6
33	Noren のれん	45.517026	-73.582608	Japanese Restaurant	Clark / Rachel	45.517354	-73.582129	52.149124	6
34	Le Majestique	45.517445	-73.580368	Restaurant	Duluth / St-Laurent	45.516876	-73.579460	94.824154	6
35	Café Santropol	45.515610	-73.580611	Café	Duluth / de l'Esplanade	45.515092	-73.581142	70.912305	6

Cluster 8

	Venue	Venue Latitude	Venue Longitude	Venue Category	Docking Station	Docking Station Latitude	Docking Station Longitude	Distance (m)	Cluster
36	La Vieille Europe	45.515891	-73.576982	Gourmet Shop	Roy / St-Laurent	45.515616	-73.575808	96.447711	7
37	Moksha Yoga Montreal	45.516057	-73.577188	Yoga Studio	Napoléon / St-Dominique	45.516745	-73.577658	84.827436	7
38	Jano	45.516066	-73.577361	Portuguese Restaurant	Napoléon / St-Dominique	45.516745	-73.577658	78.973763	7

The following figure represents the resulting 8 clusters by color, and the docking stations nearby (identified with the blue marker).



IV. Results and Discussion

Visualizing the clusters on the map by color makes it easy to select the optimal station for the cluster. We can immediately notice that there are several docking stations in one cluster, often right next to one venue. In all the clusters except for Cluster 4, we can observe the same docking station for at least 2 different venues.

However, the goal of this analysis was to determine only one docking station per cluster as it is more convenient for a tourist to be parking the bike in one docking station, and walking to the nearby venues.

I have visually selected the station that is most central to each venue in the cluster, or the station with the most venues nearby. Of course, there are several other ways to perform this final selection of docking stations, and in reality any of the stations within the cluster would be a good choice for a tourist to dock their bike, however visualizing the different clusters and stations is an intuitive and efficient way to make the final selection of docking stations.

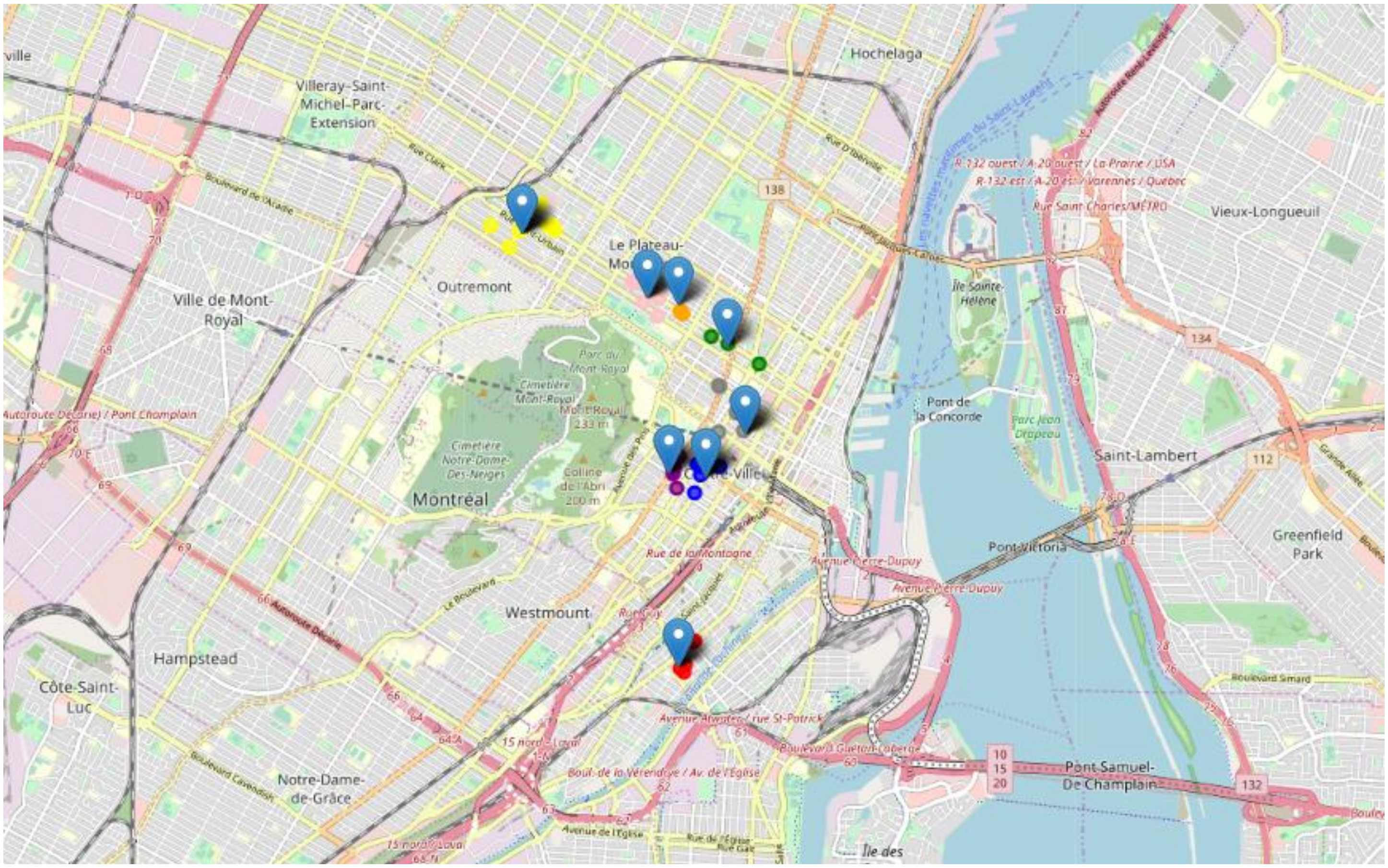
According to the map, the following stations seem to be the most representative for each cluster:

- *Cluster 1*: Marche Atwater, since it has 3 out the 5 venues nearby.
- *Cluster 2*: Waverly/St-Viateur, since it seems to be the most central in that cluster.
- *Cluster 3*: Square Philips, since it seems to be the most central in that cluster.
- *Cluster 4*: Milton/Clark, since it seems to be the most central in that cluster.
- *Cluster 5*: de la Montagne/Sherbrooke, since it seems to be the most central in that cluster and has the most venues nearby.
- *Cluster 6*: Stanley/St-Catherine, since it seems to be the most central in that cluster.
- *Cluster 7*: Clark/Rachel, since it seems to be the most central in that cluster.
- *Cluster 8*: Napoléon / St-Dominique, it seems to be the nearest to all venues in that cluster.

Now that the final selection of docking stations with the most trending venues nearby has been made, the results can now be incorporated into a new dataframe, where distances from each venue to the selected station are re-calculated. Some of the results can be following figure.

	Docking Station	Venue	Venue Category	Distance (m)
0	Marché Atwater	Liverpool House	Restaurant	350.005588
1	Marché Atwater	Satay Brothers	Asian Restaurant	22.885800
2	Marché Atwater	Le Vin Papillon	Wine Bar	327.262320
3	Marché Atwater	Marché Atwater	Market	79.287258
4	Marché Atwater	SAQ Sélection	Liquor Store	75.289608
5	Waverly / St-Viateur	naada yoga	Yoga Studio	378.948298
6	Waverly / St-Viateur	Drawn & Quarterly	Bookstore	37.964041
7	Waverly / St-Viateur	Salon de thé Cardinal / Cardinal Tea Room	Tea Room	51.894803
8	Waverly / St-Viateur	St-Viateur Bagel (La Maison du Bagel)	Bagel Shop	50.538247
9	Waverly / St-Viateur	Pizzeria Magpie	Pizza Place	10.583038
10	Waverly / St-Viateur	Boucherie Lawrence	Deli / Bodega	59.768675
11	Waverly / St-Viateur	Café Olimpico	Café	37.976787
12	Square Phillips	Pikolo Espresso Bar	Coffee Shop	91.680589
13	Square Phillips	Papeterie Nota Bene	Paper / Office Supplies Store	90.423304
14	Square Phillips	Café Parvis	Café	65.641906
15	Square Phillips	Empire	Sporting Goods Shop	84.019077
16	Square Phillips	Il Focolaio	Pizza Place	31.290720

The following map shows the results: 39 trending venues (colored by cluster) and 8 docking stations (one for each of the 8 clusters, identified with a blue marker).



V. Conclusion

The objective of this report was to analyse the trending venues in Montreal, and locate the nearest Bixi docking station, which would maximize a tourists time for visiting Montreal. Whilst exploring the data, it was clear that there are many Bixi docking stations around the city (615), a lot more than Foursquare had returned trending venues (100).

Of course, we could have fetched the trending venues around each docking station, but Foursquare would have returned too many values, which would have been filtered anyway in order to keep the most relevant. Choosing to stick with the top 100 trending venues in the city gave a more accurate representation of what actually is trending in the city, and not just because it is located near a docking station.

Using Folium to visualize the results at every step, meaning every time the data was filtered to keep the most relevant, gave insightful information about the venues and station locations. Since the objective of this project was to determine the docking stations with the most venues nearby, it made sense to use k-means algorithm with HBSCAN to cluster the trending venues by location. By setting the cluster minimum to 3 allowed this method determined the clusters of venues containing at least 3 venues next to one another. From that point on it would be easy to determine one docking station by cluster, simply by visualizing the venues by cluster, as well as the nearest stations within each cluster.

The final result yields a total of 8 clusters of trending venues, from a total of 39 unique venues, and a docking station within each cluster.