Capstone Project

Battle of the Neighbourhoods: Toronto

Problem Statement

Toronto is a bustling city full of opportunity for new businesses. While there is opportunity to be had, this has attracted many to start a business already and to unlock opportunity a new business will need to compete with the existing market.

Joe, a 35 old from Toronto has saved up working in a law firm and now would like to start his own business. It is crucial to understand where to locate his business to ensure he is not marginalised by the competition. Joe has a wide array of skills so happy to consider a range of different industries to enter.

The key questions Joe needs to answer are:

- Which industry should he consider?
- Where should Joe locate his first venue?
- Are there any further expansion opportunities for the future?

Data which can solve this problem

Data set	Source	Key features	How this will be used
Toronto post code	https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M	Postcode district and area information	To segment Toronto geographically into areas (postcodes) and return basic information of the area (postcode)
Geo-coordinates	Geocoder API	Location data	To enhance Toronto postcode data with geo co-ordinates
Foursquare venue API	https://api.foursquare.com/v2/venues/explore?<API call>	Venue types and locations	Identify existing business in Toronto, their location and type to help answer Joe's questions

Methodology

Data methodology

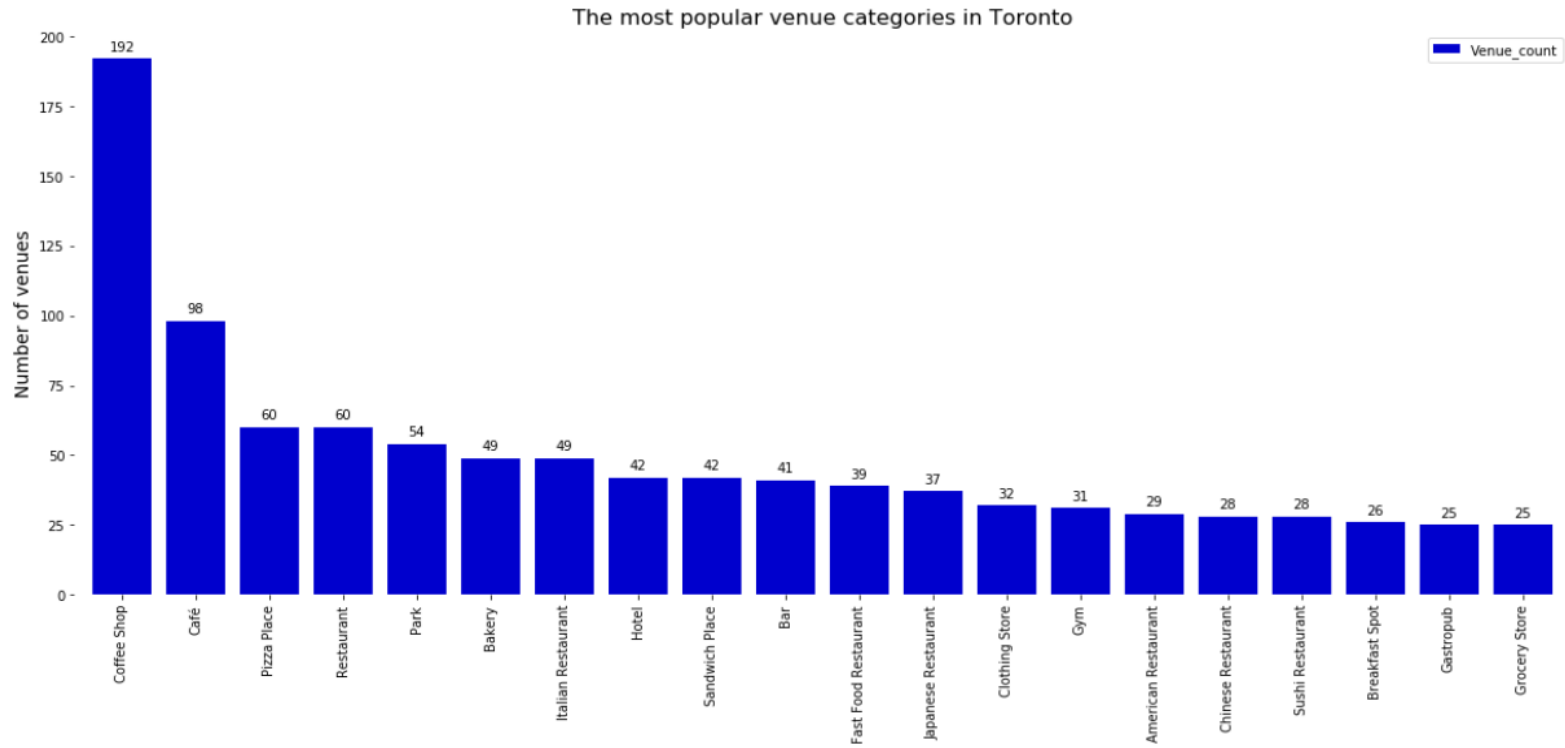
- Import postcode data from Wikipedia and enhance with geo-coordinates from geocoder
- Use the Foursquare API to gather the venue data for all postcodes
- Group the data into postcode level dataframe and another at venue level for the analysis

Analysis key steps

- Analyse which type of business is the most popular
- Understand which districts have high / low penetration of the industry selected and how the venue types correlate with each other
- Cluster to understand which is best location to start
- Use K-means algorithm to understand the distinct groups of neighbourhoods.

Results – Venue Category Rank

Coffee shops appear to be the most popular venue by far, with similar venue types also appearing in the top 20. This suggests a strong market for further venues.



There are roughly double the number of coffee shops than any other venue. There are a large number of eateries otherwise.

Results – Related Venues

Analysing the correlation between the top venue categories shows that cafes and restaurants are highly correlated with coffee shops. Whereas there is low correlation with parks, grocery stores and fast food venues

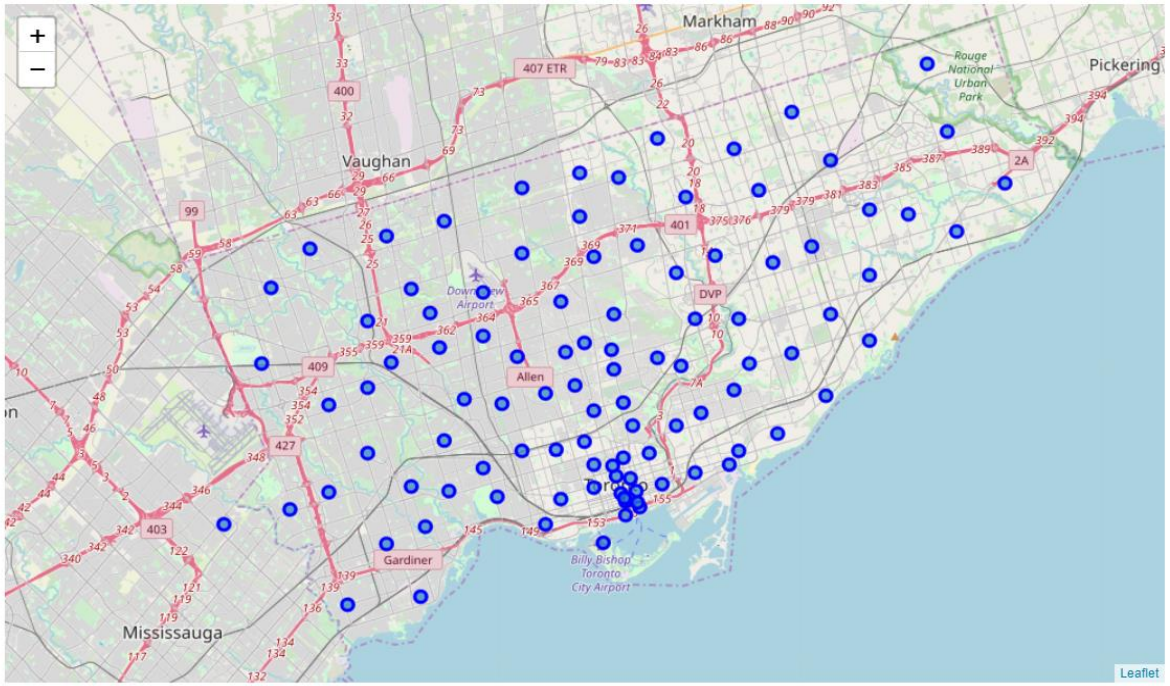
	Venue_count	Venue_category_count	Coffee_shop_count	Cafe_shop_count	Pizza_count	Restaurant_count	Park_count	Italian_Rest
Venue_count	1	0.988443	0.892716	0.853133	0.47576	0.840252	0.319523	
Venue_category_count	0.988443	1	0.865866	0.812004	0.474346	0.824408	0.323955	
Coffee_shop_count	0.892716	0.865866	1	0.74984	0.367581	0.754921	0.302048	
Cafe_shop_count	0.853133	0.812004	0.74984	1	0.445629	0.754378	0.29501	
Pizza_count	0.47576	0.474346	0.367581	0.445629	1	0.517424	0.0937003	
Restaurant_count	0.840252	0.824408	0.754921	0.754378	0.517424	1	0.167491	
Park_count	0.319523	0.323955	0.302048	0.29501	0.0937003	0.167491	1	
Italian_Restaurant_count	0.758906	0.745474	0.771432	0.682214	0.448574	0.662209	0.316515	
Bakery_count	0.715979	0.688132	0.594795	0.731727	0.231706	0.55807	0.38161	
Bar_count	0.602735	0.574915	0.466451	0.630021	0.262461	0.425347	0.0891153	
Hotel_count	0.756788	0.690411	0.773834	0.727035	0.356431	0.790685	0.201426	
Sandwich_Place_count	0.369833	0.389893	0.39049	0.373186	0.492871	0.280324	0.119853	
Fast_Food_count	0.375211	0.412564	0.303715	0.115211	0.134962	0.36769	0.042779	
Japanese_Restaurant_count	0.739934	0.739905	0.652035	0.578889	0.263478	0.722938	0.139803	
Gym_count	0.601319	0.576105	0.581524	0.597227	0.316684	0.667223	0.0293167	
American_Restaurant_count	0.679639	0.633487	0.653108	0.6887	0.271461	0.705807	-0.00701849	
Clothing_Store_count	0.394235	0.427456	0.326673	0.115317	0.0425003	0.276543	-0.0262382	
Sushi_Restaurant_count	0.462172	0.493218	0.394203	0.363276	0.434001	0.469244	0.0140005	
Breakfast_Spot_count	0.465456	0.466187	0.443401	0.399495	0.154527	0.572036	0.198331	
Chinese_Restaurant_count	0.468576	0.459354	0.341247	0.443209	0.131939	0.124622	0.187559	
Grocery_Store_count	0.105515	0.129043	0.0170996	0.121162	3.90999e-17	0.110883	0.0972684	
Gastropub_count	0.719081	0.681939	0.647191	0.751882	0.355215	0.765041	0.115419	

Results – Geographic understanding the neighbourhoods (1/2)

Plotting the clusters of neighbourhoods by using a geographical lens results in two highly concentrated neighbourhoods being returned – one near union station and another slightly north of this.

This appears to be a key area to consider for business setup given the density of these neighbourhoods.

Out[33]:

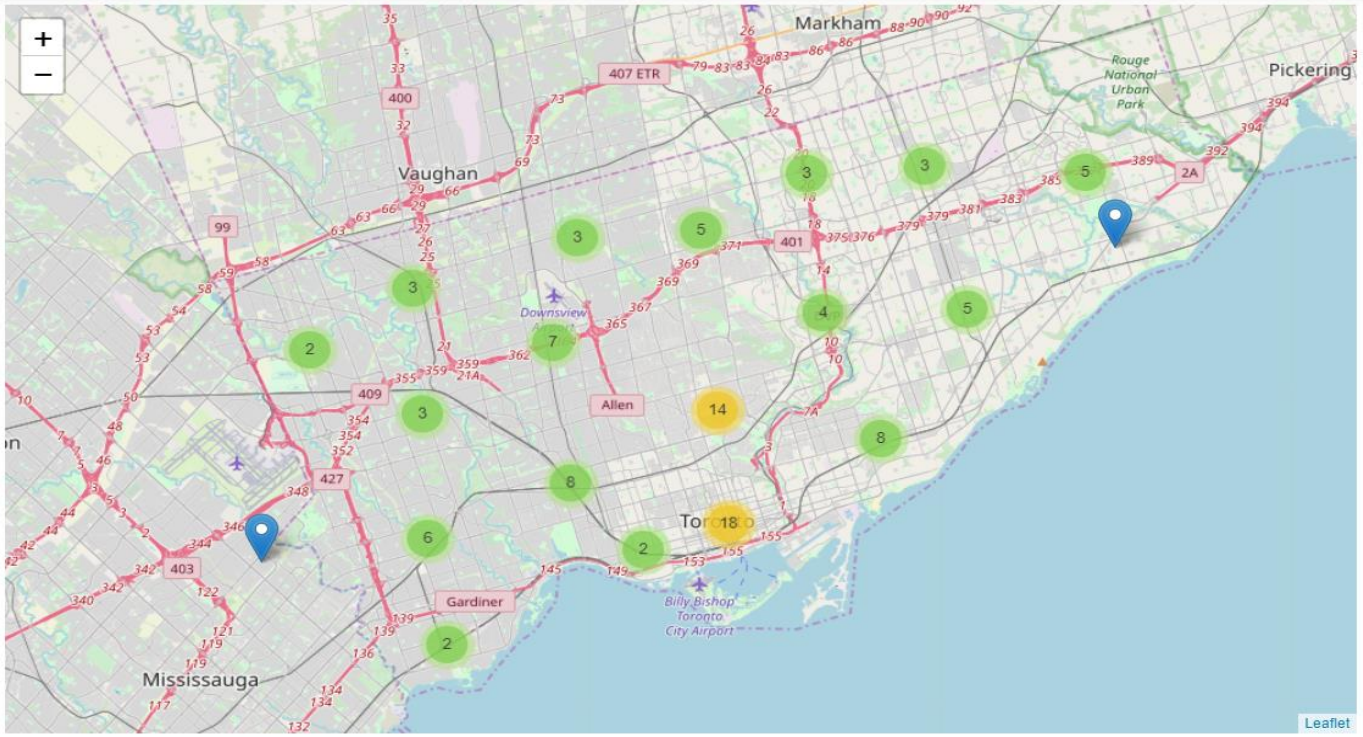


By mapping the neighbourhoods, it is possible to see a concentrated cluster by Toronto Union station.

Results – Geographic understanding the neighbourhoods (2/2)

This becomes more clear when we plot the neighbourhoods by geographical clusters.

Out[36]:



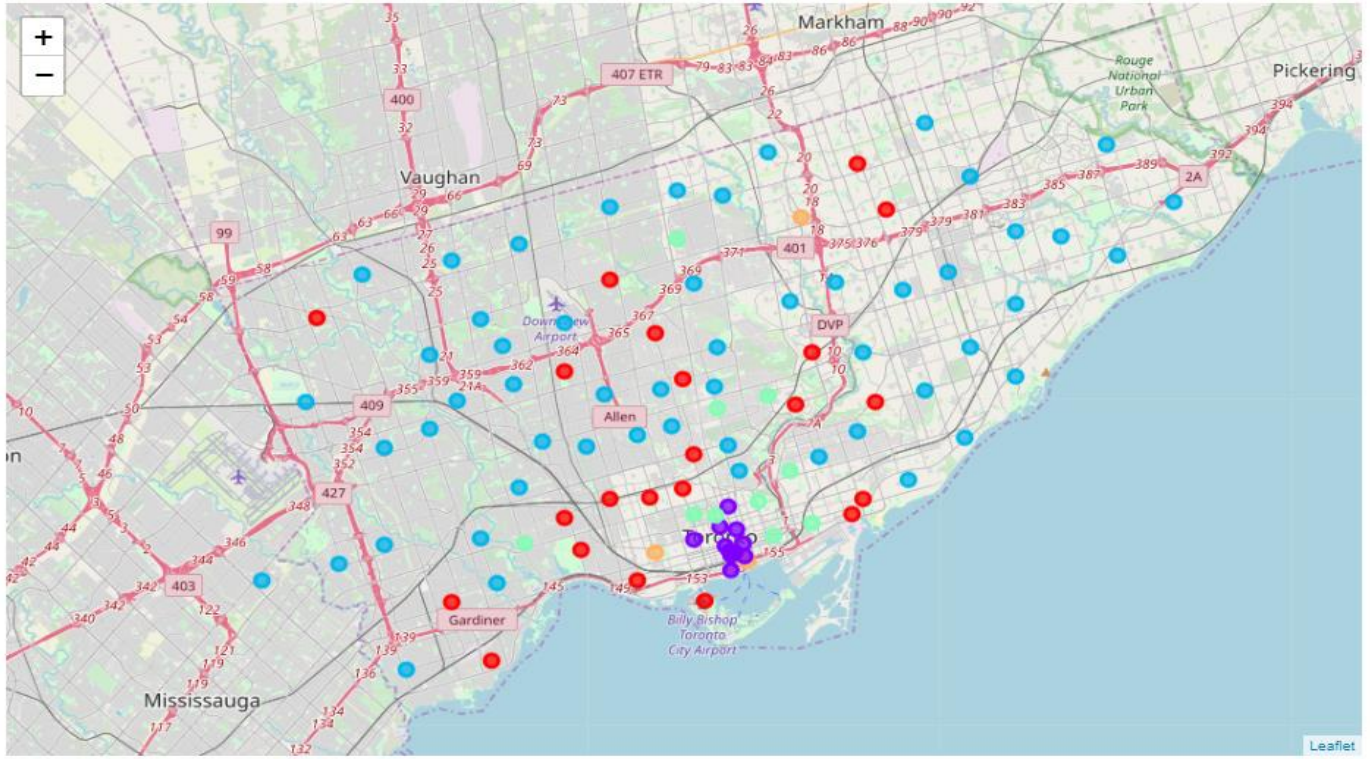
Plotting the markers into clusters we can see that there are two very high concetrations of neighbourhoods - one by the union station and another by Davisville station (slightly north of union station)

Additionally, some medium concentration clusters appear south of the airport and also east of union station.

Results – Venue analysis for Toronto Neighbourhoods

The clustering of neighbourhood by venue characteristics validates the downtown clusters previously seen (purple). It also reveals the distinct suburb groups which are shown in blue, red and orange. The green segments highlight neighbourhoods with inner suburbia characteristics.

Out[160]:

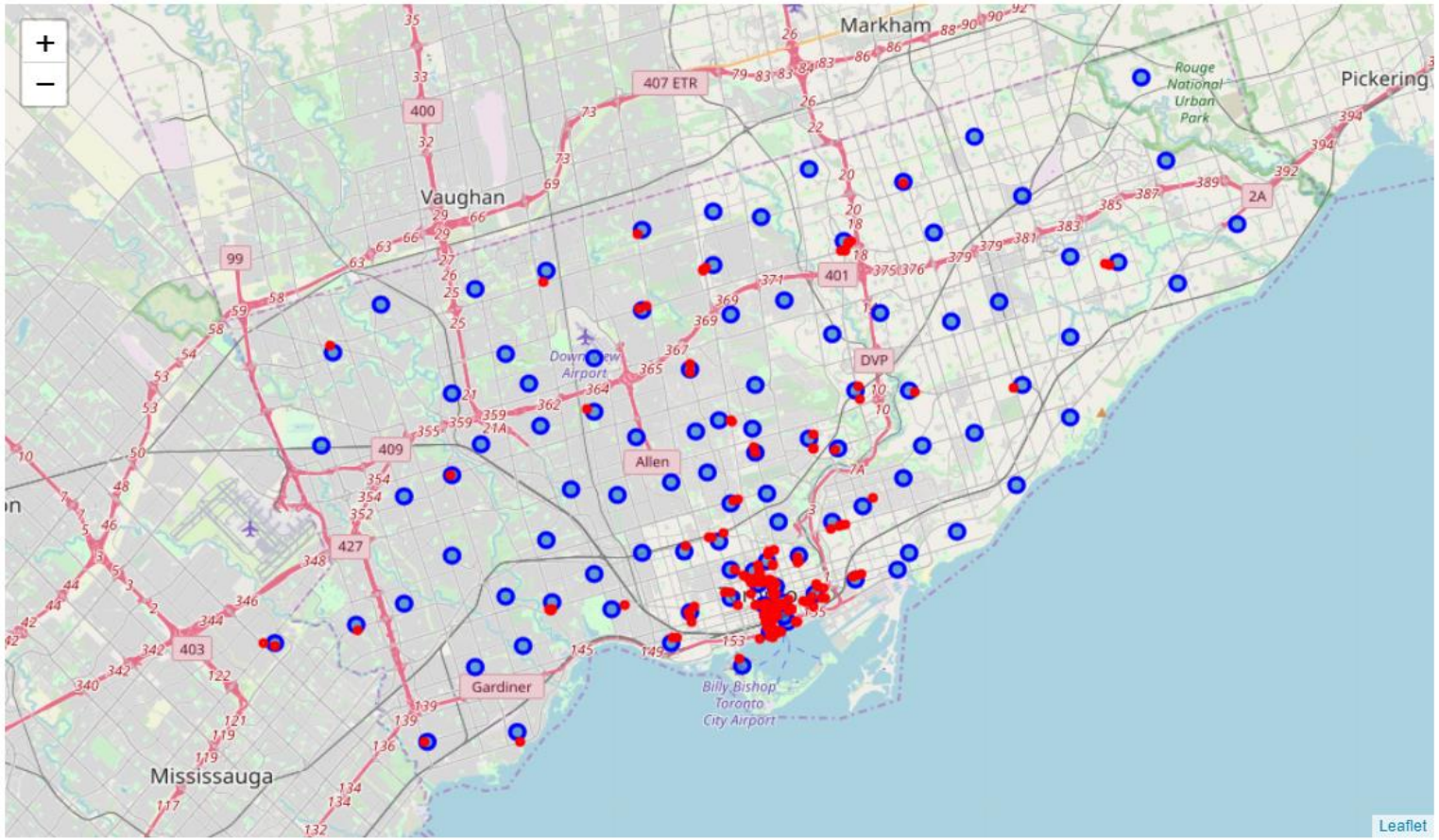


The k means clustering returned a clear 'downtown' group (in purple). There is an orange group which appears to refer to non geographic places (eg sorting office ect). The remaining three groups are clusters of suburbia.

Results – Understanding the location of existing coffee shops and similar venues (1/2)

Existing coffee shops appear concentrated around the downtown area.

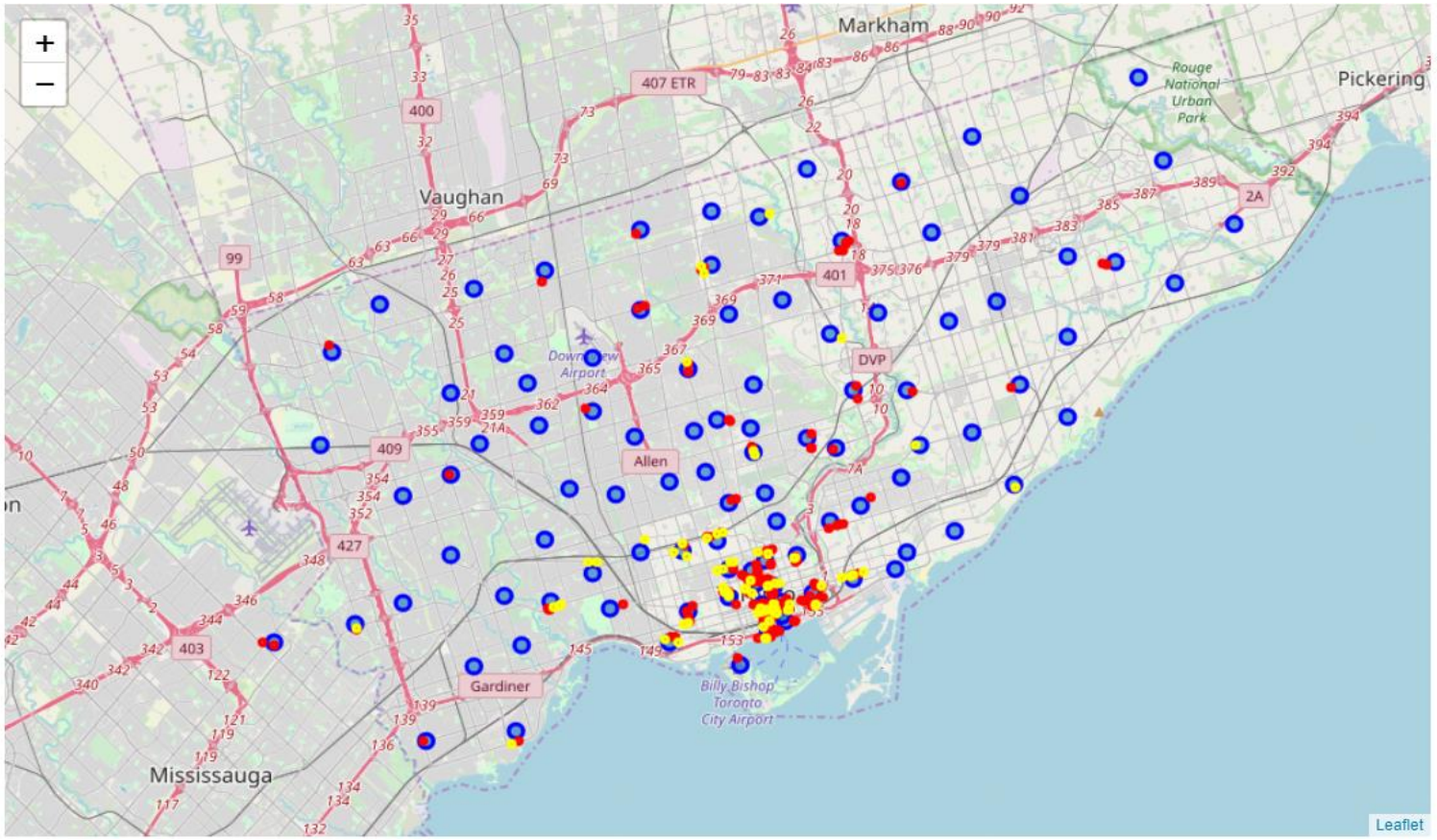
Out[137]:



Results – Understanding the location of existing coffee shops and similar venues (2/2)

When adding related businesses to this picture, the same distribution is seen. The yellow dots represent café venues.

Out[139]:



Discussion & Recommendations

- The most popular venue type by far is the coffee shop, and this is the recommended market for Joe to enter. Existing businesses are concentrated geographically, so there remains opportunity in this market.
- Location wise, setting up downtown initially where there is high demand would be the recommendation from the analysis on the data. Joe should focus on where there is less competition, so locations where there are higher numbers of park, grocery store or fast food venues.
- The future market opportunities look good – once Joe has cracked the downtown market, there is also the inner suburbia and multiple suburbia markets to tap into.

Conclusion

- Joe now has a recommendation to focus on from his original 3 wide ranging questions. Opening a business is a major undertaking and as such, he needs to undertake further analysis and financial modelling to validate the robustness of the opportunity surfaces. Further investigation of the competitor service needs to be completed – ie price points and product offering. Joe should then decide whether to compete on product, price or service depending on the opportunities and competition levels.
- Data analysis which could inform this would be looking at reviews and business revenue / turnover.