Prepared for submission to JHEP

Students Need Their Coffee: Post Pandemic

Dylan Linthorne

Ottawa-Carleton Institute for Physics, Carleton University,
1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada

Contents

1 Introduction

There's one thing most students can't go without, and that's their local coffee shops. Whether it's the productivity boost from a cup of coffee or the shop itself as the socializing hub, it's hard to deny the importance these businesses have on the micro-economies of urban (student dense) populations. With new coffee shops appearing yearly it's an exciting market to penetrate even with higher competition. A new business owner would benefit from being strategic in where they open up their shop. Naively, the business owner might assume that opening up closer to any highly populated university district would benefit their business the most. This might not be the case when larger chained coffee shops, such as, Starbacks and Second Cup have larger shares in student populated areas. On top of the saturation of chained coffee shops, any modern business owner has to take into account the affects the pandemic has had on the market within their proposed areas of interest. Within the same area, franchised shops have a higher tendency of sustaining themselves during covid-19 shutdown orders, whereas smaller shops are more than likely to fail.

2 Data

In this report, I attempt to survey the Greater Toronto Area (GTA) coffee shop market for a potential new coffee shop owner, as well as, current coffee shop stakeholders. This survey will consider multiple factors: proximity to a university, franchise coffee shop occurrences, and the site's historical pandemic case information. Alongside specific predetermined features, a k-means clustering algorithm will be employed to locate any hidden features within the dataset.

This survey considers only data within the (GTA) but this is merely an example and the analysis should be general enough to apply to any urban populous. With that in mind, Toronto neighbourhoods are divided by postal code status, or FSA. Postal code data was

PostalCode	Latitude	Longitude	Universities	Uni Pop	Covid Cases
M5B	43.657162	-79.378937	Ryerson	39471	0.003853
M1C	43.784535	-79.160497	UofT	12980	0.011873
M1G	43.770992	-79.216917	Contennial	35000	0.017357
M2J	43.778517	-79.346556	Seneca	97500	0.020655
M3J	43.767980	-79.487262	York	49905	0.019874
M2M	43.789053	-79.408493	Tysdale	1361	0.011074
M5R	43.672710	-79.405678	George Brown	32117	0.003836
M5S	43.662696	-79.400049	UofT	93,081	0.002586
M5T	43.653206	-79.400049	OCAD	6072	0.004513
M9W	43.706748	-79.594054	Humber	83000	0.025967

Table 1. Example subset of the refined dataset used for the survey.

webscraped from `Wikipedia` and a geographic coordinate is assigned to each postal code using the `Geocoder` API.

(GTA) Is a great example of a highly student populated area, with over ten universities and colleges with populations greater than a few thousand. In this study, only universities with student enrolments greater than 1000 were considered. universities and colleges with narrow fields and predominantly graduate diplomas were ignored. The reason being is that these schools usually cater to an older student population with differing consumer habits to that of an average undergraduate. University data was collected from each of the individual schools official websites. The data include; postal code, university name, and last recorded student population.

The final source of data comes from the Toronto Open Data Catalogue[1] which hosts the city's cumulative covid-19 cases to date. This data was scrapped using their own API. The dataset was later refined to only include cases that are considered "Community Spread" as this is a better indicator on the affects to the local community. The mean of the positive cases by postal code was normalized using,

$$||\text{Cases}||_i = \frac{\text{Cases}_i}{\sum_i \text{Cases}_i} \tag{2.1}$$

Where i denotes the associated postal code, and the sum is over only postal codes considered in the analysis. A subset of the dataset is shown in Table. 1, where only postal codes hosting one of the universities is shown and scaled covid cases equals $||\text{Cases}||_i$

3 Methodology

Three individual datasets were merged to give data structured in the form of Table. 1. To accomplish this all three datasets were refined to ensure they have the correct shape

[1] https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

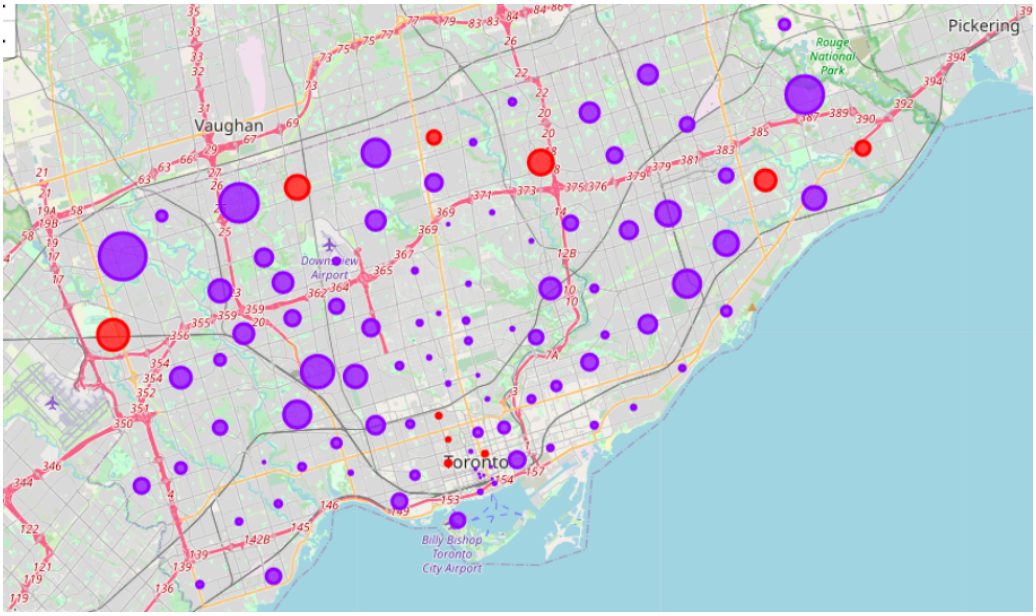

Figure 1. Follium geographical representation of GTA postal codes. Red (purple) markers represent areas with (no) universities. The radii are scaled by the normalized covid cases.

and data types. In the case of the covid data, there were a few postal codes missing data. One option would have been to estimate the missing data by using the mean case size of the surrounding neighbourhoods. After noticing the large non-linear swings in covid cases between neighbourhoods the missing data rows were instead dropped.

`Follium` was used to overlay our data over a geographical representation of the Toronto neighbourhoods. In Fig. 1 each postal code is given its own marker. The marker colours distinguish whether the postal code hosts a university (red) or not (blue). Each marker is given a size that is proportional to the normalized covid cases from Eq. 2.1. It's easy to notice that the outer suburban areas have larger community spread than that of the downtown core, which has been an observed global correlation.

to access the coffee shop statistics for each neighbourhood the `Foursquare` API was used to query the number of venues associated with the *coffee* keyword. A bias might appear when non-coffee shop related results are accessed under the keyword, or the opposite and coffee shops are missing due to inefficient keywords. To remedy some of these biases, only categories with coffee related wording was kept. Search queries were restricted to a radius of $r = 500$ m and the result's latitude lat_c, longitude lng_c, distance d_c, and it's name were considered. The venues names were grouped together and the top four coffee shops were calculated for each postal code. An example of of the top coffee shop per postal code is shown in Table. 2. The idea is to see highest share holders of the neighbourhood market. The cafe distance information was used to calculate the mean distance $\bar{d}_c$ between coffee shops for a given area.

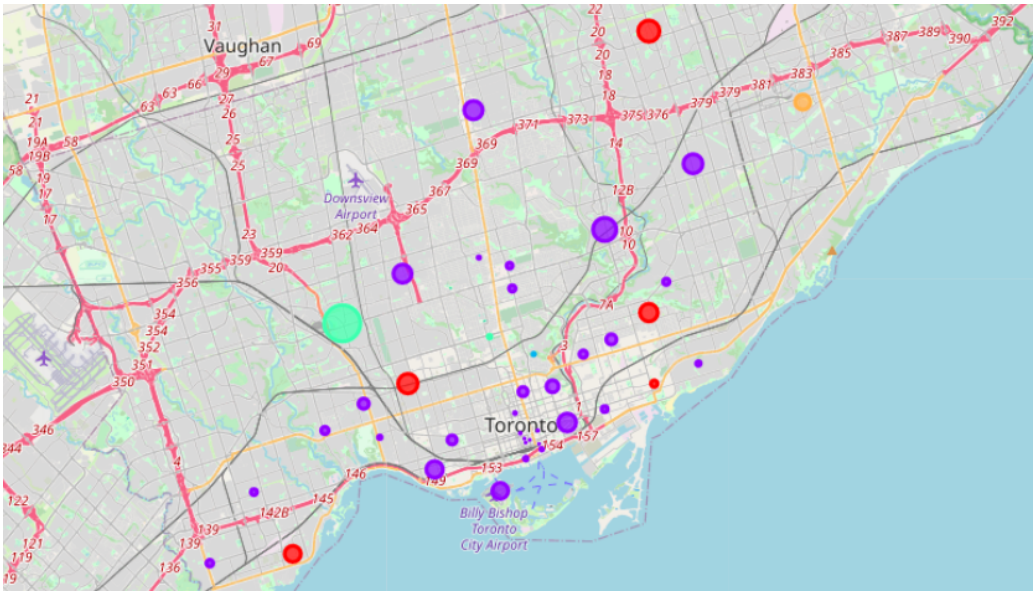

Figure 2. k-Means clusters shown in colour (k=5) for Toronto neighbourhoods. Clusters are trained on top leading coffee shop market shares for a given area. Red = 0, purple = 1, blue = 2, green = 3, orange = 4. Radii are scaled by normalized covid cases.

The dataset was split into to subsets, one set with all of the above information for postal codes with universities and the other set without which will be refereed to as the public data set. The closest university was found for each row (postal code) of the public dataset using the minimum of the euclidean distance metric,

$$d_{\min} = \min(\sqrt{(lat_u - lat_p)^2 + (lng_u - lng_p)^2} \tag{3.1}$$

Where lat_i is the latitude for either the university (u) or public (p) coordinates, similarly for lng_i and their longitudes. Keep in mind that we're using the distance metric on angular coordinates, but since we're looking at a small local region any corrections are negligible. We'll assume a units are dimensionless for consistency. For each postal code, the above equation found the closest university, its population. and its distance.

The next step of the analysis was to use the k-Means algorithm to search for any hidden

PostalCode	1st Most Common Cafe	2nd Most Common Cafe
M1H	Coffee Culture Cafe & Eatery	135 Ossington
M1R	Sam's Coffee Truck	Reunion Island Coffee
M1W	Coffee Time	Coffee Here
M2N	Second Cup Coffee Co.	Little Italy Coffee Shop
M3C	Pauls coffee shop	Reunion Island Coffee
M4B	Nostalgia Coffee Co.	135 Ossington

Table 2. Example of most populated coffee chains

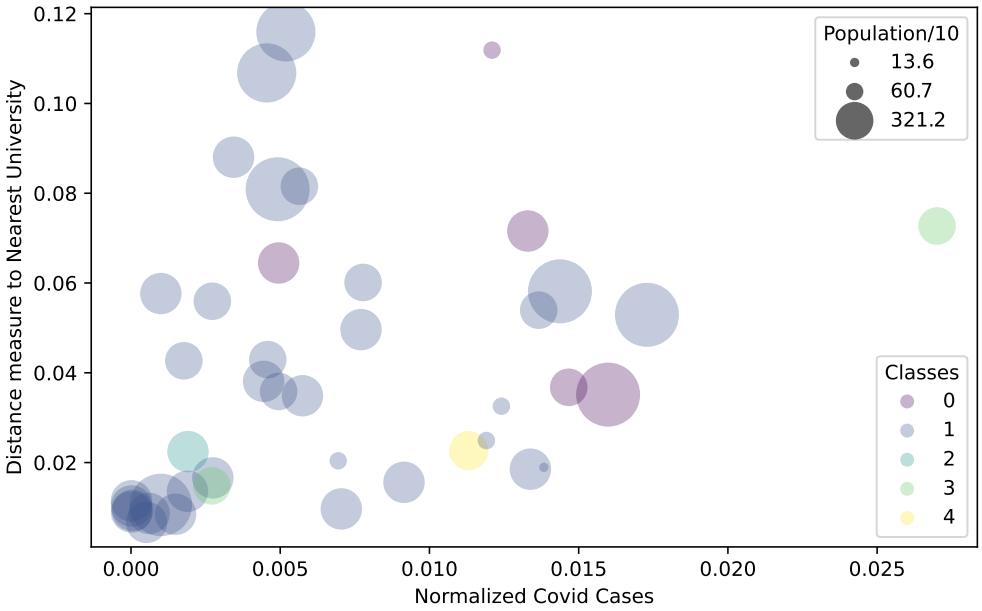

Figure 3. Scatter plot of clusters from the k-Means analysis. The radii are scaled by the scaled university population.

Figure 4. k-Means clusters shown in colour (k=5) for Toronto neighbourhoods. Clusters are trained on top leading coffee shop market shares for a given area.

features to assist in the inference. The ML k-means method was given the examples of the top four coffee shops for each postal code. A $k = 5$ was taken throughout. The results are overlayed with cluster label colours using `Follium` In Fig. 2.

A regression analysis was employed to find any correlation between covid cases and distance from a university. The regression model is of a single two parameter linear function $y = a_1 \cdot x + a_2$. The r^2 and the p-value (assume a null hypothesis of zero slope) were also estimated using the `scipy` libraries.

4 Results

Considering the clusters visually shown in Fig. 2 the various labels can be further investigated. This is done by looking deeper into what are the shares that coffee chains have in these clusters. Looking closer leads to:

- Cluster 0: Coffee time is the leading chain, showing up as the most common 80% of the time

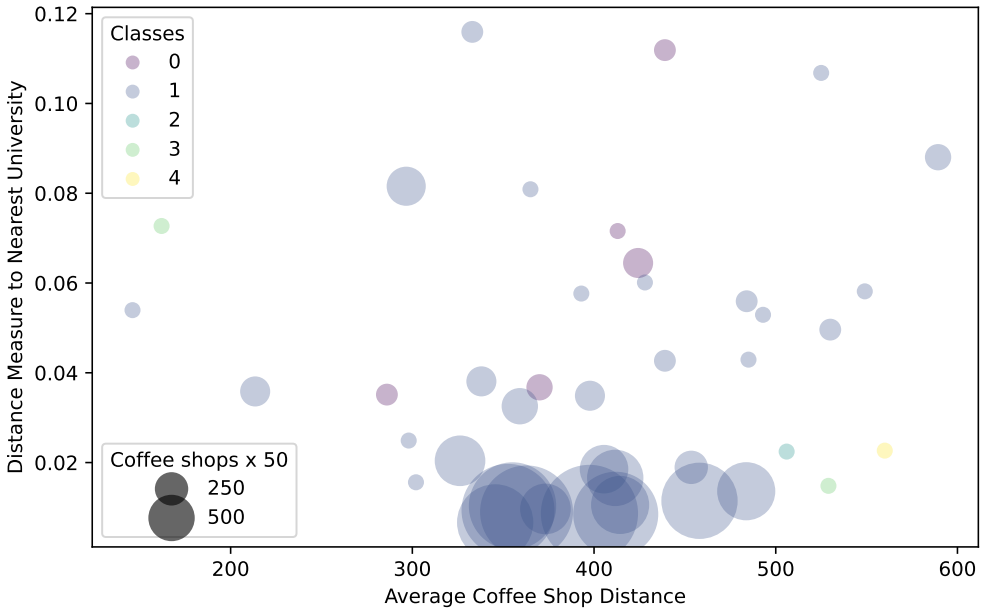

Figure 5. Scatter plot of clusters from the k-Means analysis. The radii are scaled by the scaled coffee shop density

Figure 6. k-Means clusters shown in colour (k=5) for Toronto neighbourhoods. Clusters are trained on top leading coffee shop market shares for a given area

- Cluster 1: These areas are highly dense with coffee shops, with the leading chains commonly showing up such as Starbucks, Second Cup, Timothy's World Cofee.
- Cluster 2: This is an isolated area with no major coffee chains in the down town core.
- Cluster 3: These areas, however sparse, are dominated by Timothy's World Coffee.
- Cluster 4: This isolated area have no large chains and away from the down town core.

Now looking at the distribution of clusters as a function of normalized covid cases and distance to nearest university. This is seen in Fig. 3. The sizes of the circles are proportional to the population size of the nearest university. It can be seen that clusters 2 and 4, which aren't dominated by large chains, have both low covid cases and relatively close to their local university with a sufficient student population size.

Correlation Coefficient	P-value
0.2860	0.0629

Table 3. Regression analysis statistics

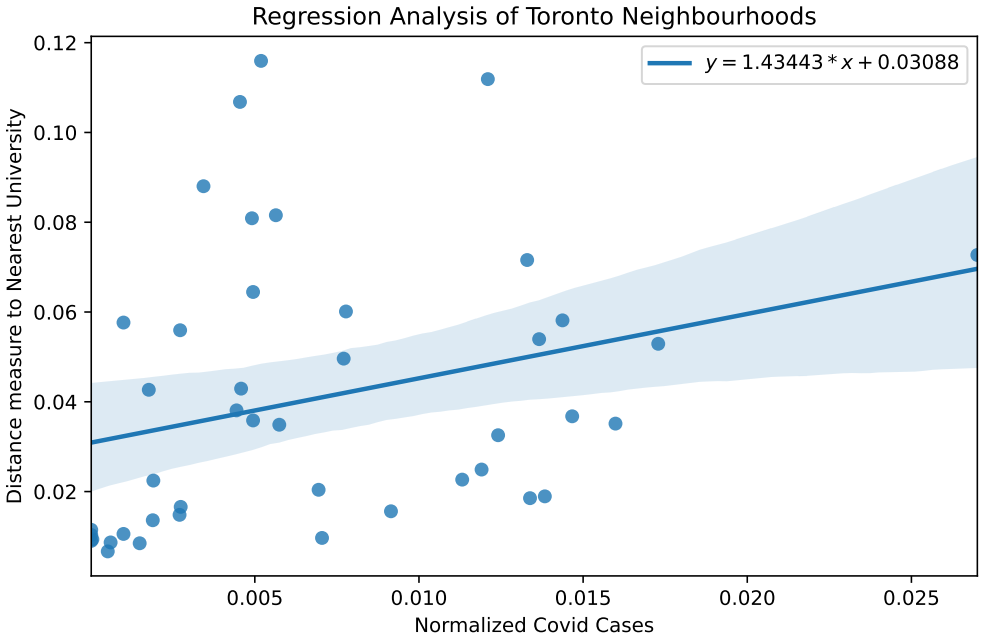

Figure 7. k-Means clusters shown in colour (k=5) for Toronto neighbourhoods. Clusters are trained on top leading coffee shop market shares for a given area

instead of looking at covid effects, the university distance is plotted again the average coffee shop distance for each cluster. This is seen in Fig. 4. The sizes of the circles are now proportional to the number of coffee shops within a 500m radius. Cluster 1's high density of coffee shops is realized in the plot. Looking again at the candidate clusters 2 and 4, it can seen that both have low coffee shop densities as well as a large separation between shops.

Finally, after visually inspecting Fig. 3 a linear trend is witnessed. If a trend does exist between the covid cases and distance to a university then shop owners can use it as a gauge on effects from future covid waves. To verify this hypothesis a linear regression was applied onto the dataset, as discussed in the methodology.

The goodness of fit statistics are given Table. 3. The correlation Coefficient gives insight on a small possible trend within the data. Whereas the p-value, which assumes a zero slope null hypothesis, dos not have $p < 0.5$, but does give a Gaussian equivalent value within one standard deviation.

5 Summary and Conclusions

In summary, the modern coffee shop owner or stackholder will want to invest in the most profitable area. Many factors can influence the consumer's habits, such as university location, choices in the market, and pandemic shop closures. In this survey neighbourhoods

were clustered into segments which provided a boundary between areas hosting a high density of chained coffee shops ,and ones with out (clusters 2 and 4). Multiple dimensions of information were used to seek out the ideal locations for an up coming coffee shop in the GTA. The recommendation to the shop owner would be to open up a shop either near a university further from the down town core (such as in cluster 4) as it benefits from a thriving university population with minimal competition. An alternative would be to choose a location similar to cluster 2 which is closer to the down town core but benefits from a lack of chained shops and a large inter-shop distance. Both these areas seem to also represent a smaller portion of the covid community spread cases, give the owner a sense of relative stability during future lockdown policies.

A two-parameter linear regression analysis was employed on the data summarized in Fig. 3, as a trend was visually suspected. The model did not fit the data with enough statistical signifigance to deem it reliable to the stackholders. Future studies could use differing models from that of the linear model and reassess their significance. In the end, although recommendations have been made this survey would benefit from additional data from similar city municipalities.