A New Store for Toronto

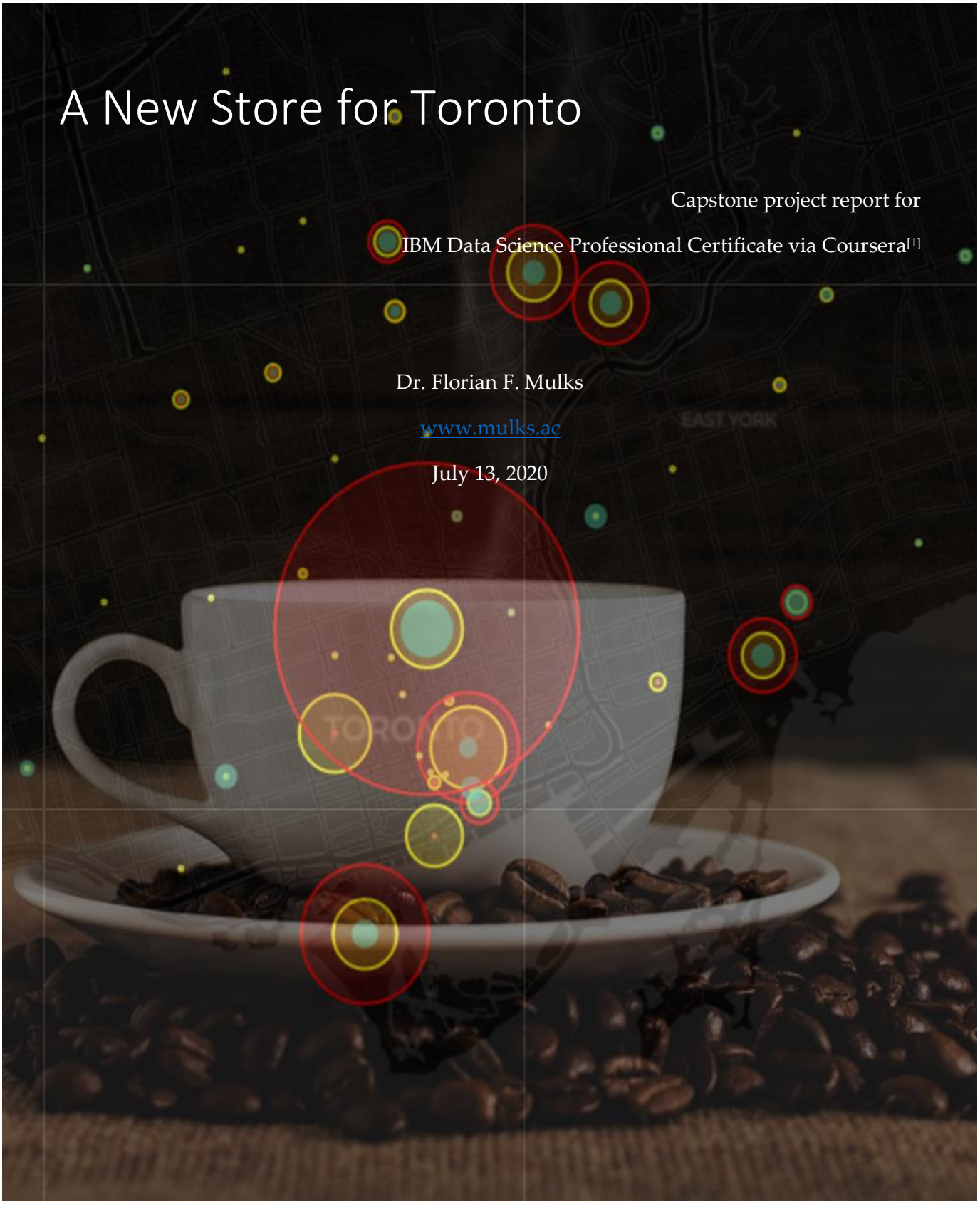
Capstone project report for

IBM Data Science Professional Certificate via Coursera[1]

Dr. Florian F. Mulks

www.mulks.ac

July 13, 2020



ABSTRACT

A non-biased approach was used to isolate the best place and type for launching a venue in Toronto. Clustering and linear regression methods were used to deduce these conclusions for data on current existing venues in Toronto. A neighborhood with a strong lack in a category of venues describing coffee shops, tea rooms, and several others was found. In this neighborhood, Church and Wellesley, an extreme lack of coffee shops was deduced and it was shown that little competition of other related venues can be expected.

TABLE OF CONTENTS

1 INTRODUCTION

Toronto is the most populated city of Canada with 6.5 million citizens as of 2019. Canada is the country with the second largest area after Russia. This makes Toronto as metropolitan among the top 10 most populated cities of North America a valuable target when aiming to succesfully open a shop. This capstone project analyzes the structure of Toronto's venues in order to identify a promising venue category and location for launching a new store.

This is the capstone project of the IBM via Coursera Data Science Professional Certificate which entails that it has been analyzed to great depths with various approaches before. Valuable tools that were identified before are clustering approaches based on the amount of venues of certain categories, rents, population denisity and alike. Most analyses worked with a discrete aim that might be finding the best location of launching a certain venue type,[3] finding a good place to rent for a personal flat,[4] or simply identifying the "best" neighborhood.[5]

Herein, we will employ Foursquare data on existing venues of different categories in Toronto to deduce a recommendation on which a type of venue at which location in Toronto can safely be assumed to become successful. This is useful information in three scenarios: planning to move to Toronto and into independence, planning to open a new store as a company, or for city planning purposes.

2 METHODOLOGY

Wikipedia is used to import postal codes starting with M in Canada, which include neighborhoods in Toronto, Canada. This data set contains 103 postal codes attributed to different neighborhoods in different boroughs. Geocodes of these neighborhoods are retrieved and then used to request data from the Foursquare API about surrounding venues. This data set contains the names, IDs, addresses, categories, and more attributes of these venues. We use the category of venues to predict healthy neighborhood structures. 2130 venues were retrieved which are separated into 271 categories. These are assigned to the 103 postal codes. Unsupervised clustering algorithms, Principal Component Analysis, and ridge regression are used to structure the data and to build a predictive model. See chapter 3.8 “Building a Predictive Model” and Figure 9 for a more detailed discussion of the methods used in the final model.

3 RESULTS AND DISCUSSION

3.1 RETRIEVING NEIGHBORHOOD LIST

We first import the necessary libraries requests and pandas for downloading the table with data from Wikipedia. As the homepage structure is quite simple, we can use pandas to extract the table with postal codes of Canada. Numpy is going to be used for treating NaN values and might be useful for handling arrays later. This creates a dataframe with postal codes assigned to boroughs and neighborhoods in Toronto.

3.2 GEOCODE RETRIEVAL OF COORDINATES

With the dataframe at hand, we now require the coordinates of our neighborhoods for querying details about them from Foursquare later on. We try to use the geocoder library for this. Python lists for the latitudes and longitudes are populated with a for-loop over the rows of our data. They are efficient data storage structures for later appending them to our dataframe.[1] We will attempt to retrieve data from OpenStreetMap (OSM). OSM seems to only have very limited functionality in the desired region in Toronto. We will have to use a pre-compiled csv-file with geocodes of the postal codes that is offered on the course website. Only 25 locations are successfully retrieved from OSM (even while allowing for 5 attempts each), only a single of which is similar to the corresponding entry from the csv-file. The data from the csv-file is appended to our dataframe (Table 1).

3.3 IMPORTING FOURSQUARE DATA

We now use the data to access Foursquare and do some initial clustering to achieve an understanding of the data before we go into the capstone research. Foursquare is a web service that collects venue names, location, categories, and user-entered tips, photos, and ratings of the venues. The postal code locations are passed to the Foursquare API to re quest lists of venues in the vicinity in 500 m radii with a limit of max. 100 venues per request. We retrieve 2130 venue names, locations, and categories this way. Individual venues are not interesting for the evaluation of their impact on our postal code areas, thus, we append our postal code dataframe (Table 1) only with pre-processed data that is meaningful for the whole area rather than only for a single venue.

Table 1. Extract from the dataframe assigning postal codes to boroughs, neighborhoods, and geolocation data.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

3.4 FEATURE CONSTRUCTION

For constructing our features, we one-hot encode the venue categories so that we can use the types of venue in order to group neighborhoods later. A set is created that contains all venue categories in order to count the numbers of venues of certain categories for each postal code area (Table 2).

Two additional features were collected: the mean venue distances in neighborhoods and the total number of venues. These might be more interesting for differentiating neighborhoods than just venue categories. The first corresponds to a venue density, implying that many venues are found in close vicinity if a low mean distance is found. The Haversine distances were calculated to find the distance between venues on the earth surface. The total number of venues promises similar information by adding the total number of features in reach of a certain area to our data set.

3.5 DESCRIPTIVE STATISTICS

The vast majority of Neighborhoods has below 43 venues (maximum: median(8 venues) + 1.5 * 3rd quartile(23)), with the interquartile range (IQR) spanning 4 to 23 venues (Figure 1(*a*)). With only a few outliers we can assume that most neighborhoods are captured representatively. All data points apart from 7 are also captured in the IQR for the mean distance of venues to other venues.

The distribution plot shows that the data set is skewed towards the lower end (Figure 1(*b*)). Only a few Toronto areas lean towards higher number of venues. Some outliers even contained more than the 100 venues that were the limit of our Foursquare request but as the boxplot has

Table 2. Extract from the dataframe assigning postal codes to boroughs, neighborhoods, and geolocation data which now is appended with the mean venue distance in the area, the total number of venues, and the one-hot encoded venue categories..

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Distance to others	Number of venues	Intersection	Bike Shop	Antique Shop	...	Recording Studio	Soup Place	Creperie
0	M3A	North York	Parkwoods	43.7	-79.3	140.0	3	0	0	0	...	0	0	0
1	M4A	North York	Victoria Village	43.7	-79.3	162.0	5	1	0	0	...	0	0	0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6	-79.3	473.0	44	0	0	1	...	0	0	0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.7	-79.4	344.0	15	0	0	0	...	0	0	0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.6	-79.3	521.0	34	0	0	0	...	0	0	1

shown before, the majority of areas is well captured as any neighborhood with more than 43 venues is classified as outlier. Results of our models regarding highly populated areas are, thus, to be handled carefully.

A joint distribution plot of the venue numbers and their mean distance over the areas shows a strongly favored average in both features (Figure 2). The average postal code area can be expected to contain around 8 (median) venues that are roughly 350 m apart. The plot also shows that the distance distribution is broad in low venue count areas, some venues are very close to each other, some are very far apart.

We then attempt to model the correlation of venue counts with their mean distances. Linear regression shows a poor correlation with an R^2-score of 0.01 (Figure 3(*a*)). A cubic polynomial fit is following the general observable trends nicely but also only reaches an R^2-score of 0.06 (Figure 3(*b*)). This means that the two models can only explains 1% and 6% of the variation of our data in regard to these two variables, respectively. This shows that we cannot build simple models based on these two variables but we will keep them together with the 271 categories as they still convey valuable information independently, especially giving low venue count areas two additional degrees of freedom to separate from each other.

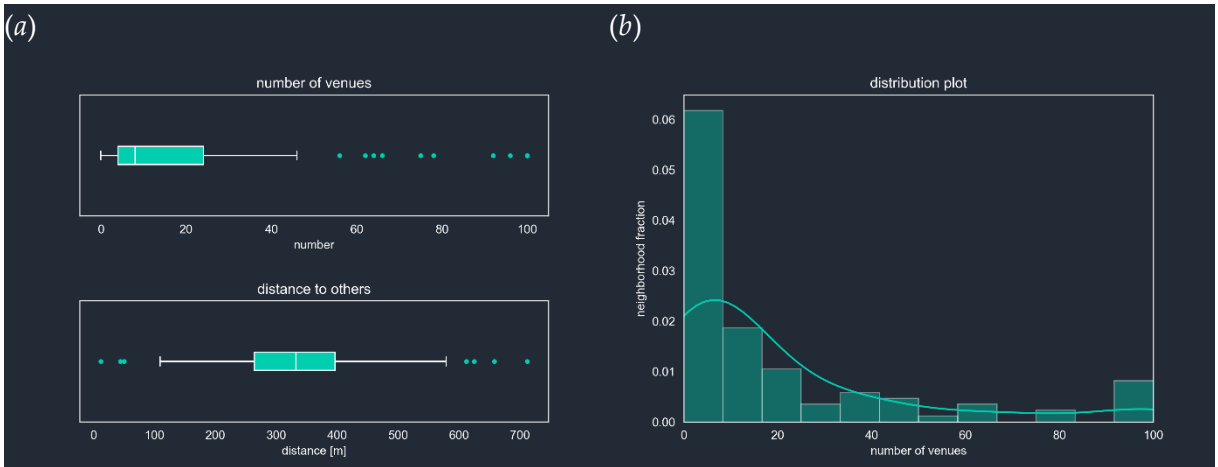


Figure 1. (*a*) Box plots showing the data distributions regarding the total number of venues and regarding the mean venue distances in the 103 postal code areas. (*b*) Distribution plot showing the distribution of the total number of venues over the different neighborhoods.

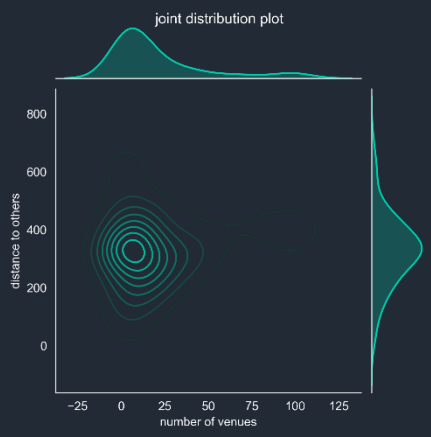


Figure 2. Joint distribution plot showing the distribution of venue counts and their mean distances in the different postal code areas in relation to each other.

Interestingly, an increasing amount of registered venues is not clearly correlated with decreasing distance between the venues. Neighborhoods with low amounts of recorded venues sometimes have them very close to each other, sometimes far from each other. With increasing amount of recorded venues, mainly the distribution becomes narrower rather than showing a decreasing mean distance. They rather seem to be closely clustered, which also makes sense. This indicates that Toronto is heavy in "High Street" and "Shopping Mall" arrangements where even with lower venue amounts they gather quite closely, while the average distance largely remains with increasing venue counts.

3.6 VISUALIZATION OF CATEGORIES AND NEIGHBORHOODS BY 2D-PCA

We now reduced the dimensionality of our categories to find some basic underlying principles of the data structure. A 2-dimensional principal component analysis (PCA) was performed to accomplish this (Figure 5(*a*)). Looking at some of the data points that appear on the extremes of the generated principal components (PC) shows that the large cluster of data points represents low density neighborhoods. The x-axis seems to correspond to higher venue counts with increasing values. The y-axis captures a certain feature distinguishing different venue categories. The total count rises both with positive and negative y-values but with different category structures of the data.

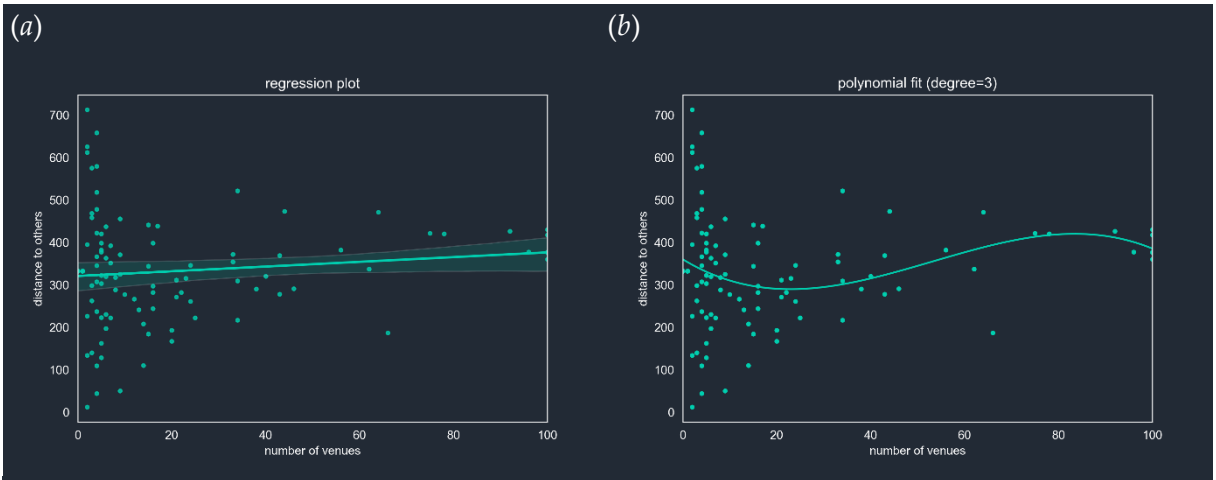


Figure 3. Plots showing regressions over the variables number of venues and distance to others. (*a*) Linear regression, (*b*) cubic (polynomial) regression.

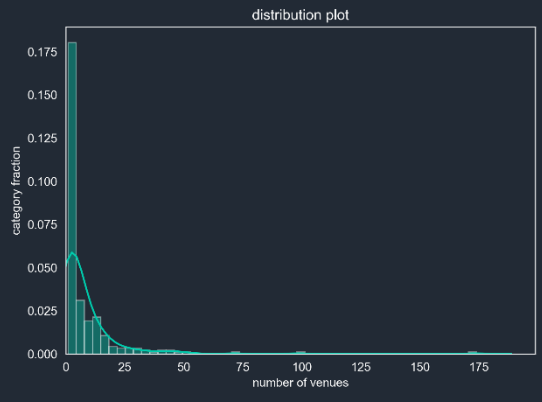


Figure 4. Distribution plot showing the distribution of our venues over the different categories assigned to them in the Foursquare database.

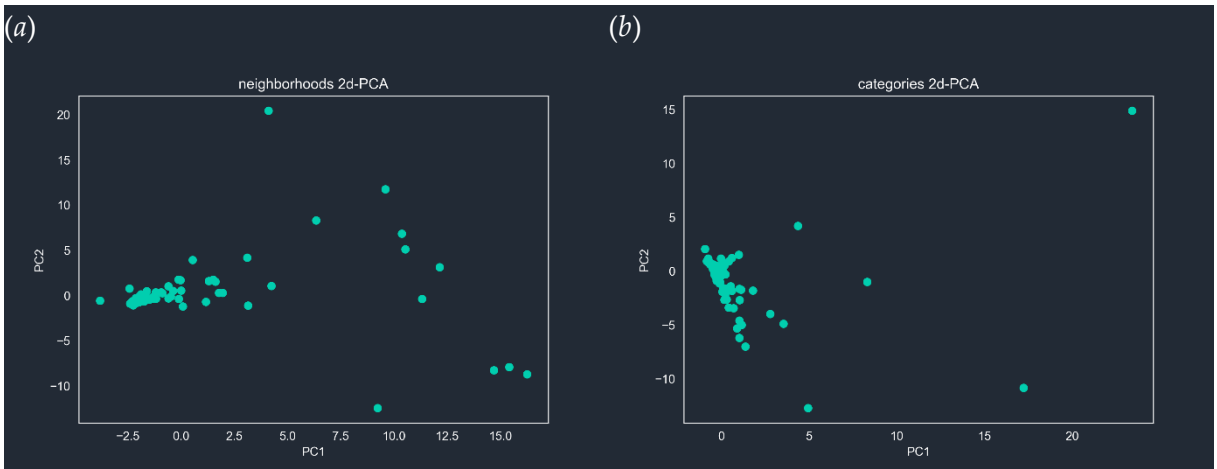


Figure 5. 2D-PCA representation of neighborhoods (*a*) and categories (*b*) to explore key features of the data set.

We then transpose the dataframe so that the different categories are seen as samples of a data set collecting their appearance in different postal code areas (Figure 5(*b*)). While the overview that we can achieve here is lackluster, it seems that the x-axis correlates to the probability of high venue density. Venues with high x-values mainly appear in neighborhoods with many venues. An example for such a venue category is the convenience store. Higher y-values with low x-values represent venues that mainly appear in low venue density neighborhoods such as baseball fields. Our most extreme example, the playground has high values of both and appears mainly in neighborhoods that contain a few venues of different categories resulting in intermediate venue density.

No meaningful grouping of venue categories was found, but 2d-PCA analysis has shown that there clearly are underlying similarities and codependences of venue types of certain categories. This reaffirms that meaningful statistical dependencies can be found in the data we collected.

3.7 Clustering Toronto Neighborhoods

We then created an interactive map in order to explore clusters of the different postal code areas (Figure 6). We chose to look at the algorithms "Agglomerative Clustering", "DBSCAN", and "K-Means Clustering". The interactive map tool allowed us to visualize the results of these algorithms. It also allows to chose if we want the total number and mean distance of venues included in the calculations. Furthermore, we look at using either relative or absolute venue numbers. This is an important difference depending on what we want to achieve with our analysis in later parts of this project. Relative numbers give an insight into the structure of existing venues, but absolute numbers are also important for making a call on where starting a new business might be important. A neighborhood lacking a Chinese restaurant but being filled with dozens of others is not a smart place to open one.

Employing absolute venue counts, generally, no meaningful structure in the data points is found (Figure 7). All algorithms deliver one large cluster containing low venue count areas and identify high venue count outliers as single node clusters. Using relative venue counts instead improves this behavior. Agglomerative Clustering still shows the same behavior and DBSCAN only identifies two clusters. Interestingly, DBSCAN is the only algorithm that

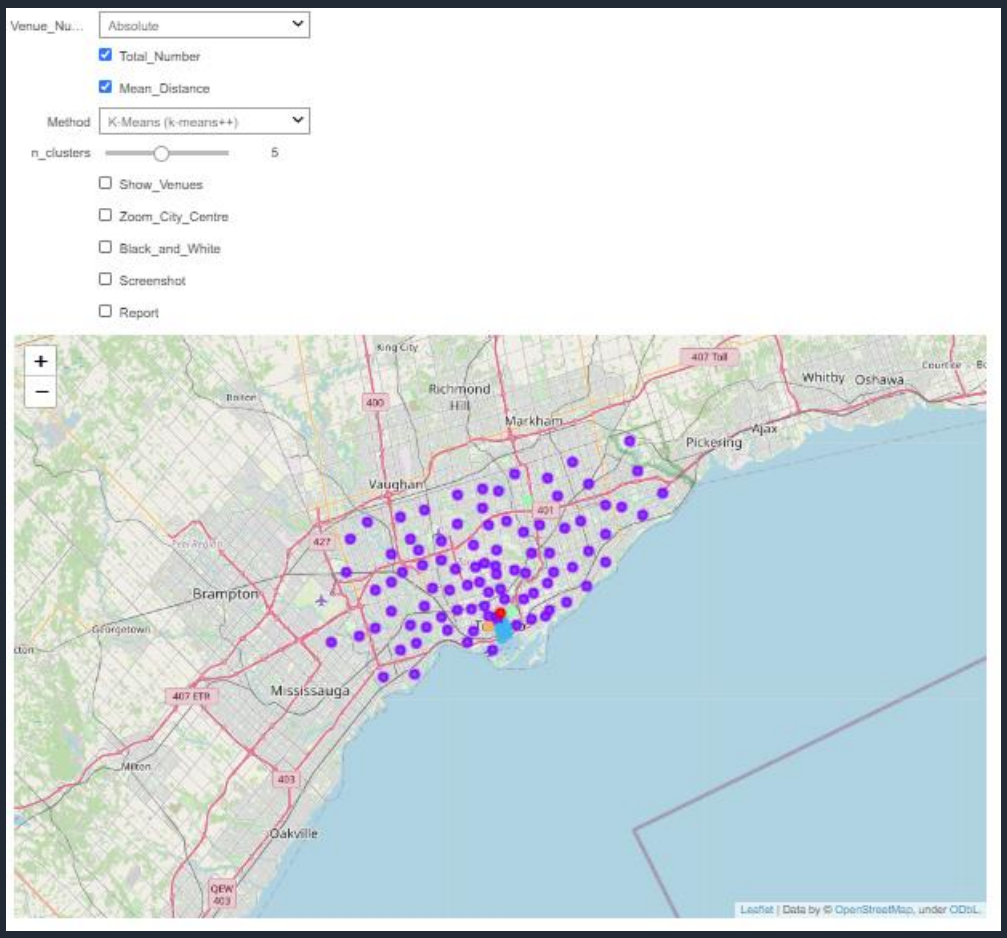


Figure 6. Screenshot of the project Jupyter notebook to demonstrate the created interactive map which was used for exploring different clustering algorithms. Venue_Number: controls if absolute or relative venue numbers are used in the calculations; Total_Number and Mean_Distance: controls the employment of the two additional features that we created; Method: 5 different clustering algorithms/options; Show_Venues: labels the total number of venues in the map; Zoom_City_Centre: switches to a downtown view; Black_and_White: employs a dark mode tile layer; screenshot: uses a webdriver to render the map in Google Chrome and produce reproducible screenshots; report: prints the number of neighborhoods in each cluster and the indices of all neighborhoods for each venue.

identifies two clusters within the suburban areas rather than assigning a second cluster to high venue count areas in downtown Toronto. K-Means is the only algorithm capable of both identifying high venue count areas and identifying some structure in the lower venue count areas, we looked into it a bit more. The different initialization methods k-means++ and random initialization deliver very similar results. The attempt to create more meaningful features by using a 50-component PCA (reducing the number features from 273 to 50) reduces the sensitivity towards the discussed observables. With the number of clusters k = 5, 4 single node clusters are found and one cluster containing the remaining 99 neighborhoods.

An exemplary look at K-Means with random initialization shows that the clustering quality only slightly improves with different numbers of clusters k (Figure 8(*a*–*d*)). The lack of an elbow in the elbow plot showing the error/distortion for k = 1–40 confirms the notion that there is no number of clusters delivering a desirable clustering quality for unsupervised clustering (Figure 8(*e*)). The postal code areas cannot clearly be separated into different groups but rather act as continuum.

Figure 7. Visualization of the results of different clustering algorithms for clustering our postal code areas (with k = 5 apart from DBSCAN which decides on the number of clusters itself). (*a*) Agglomerative clustering, (*b*) DBSCAN (only 2 clusters found), (*c*) K-Means with k-means++ initialization, (*d*) K-Means with random initialization, and (*e*) K-Means with random initialization and 50 component PCA.

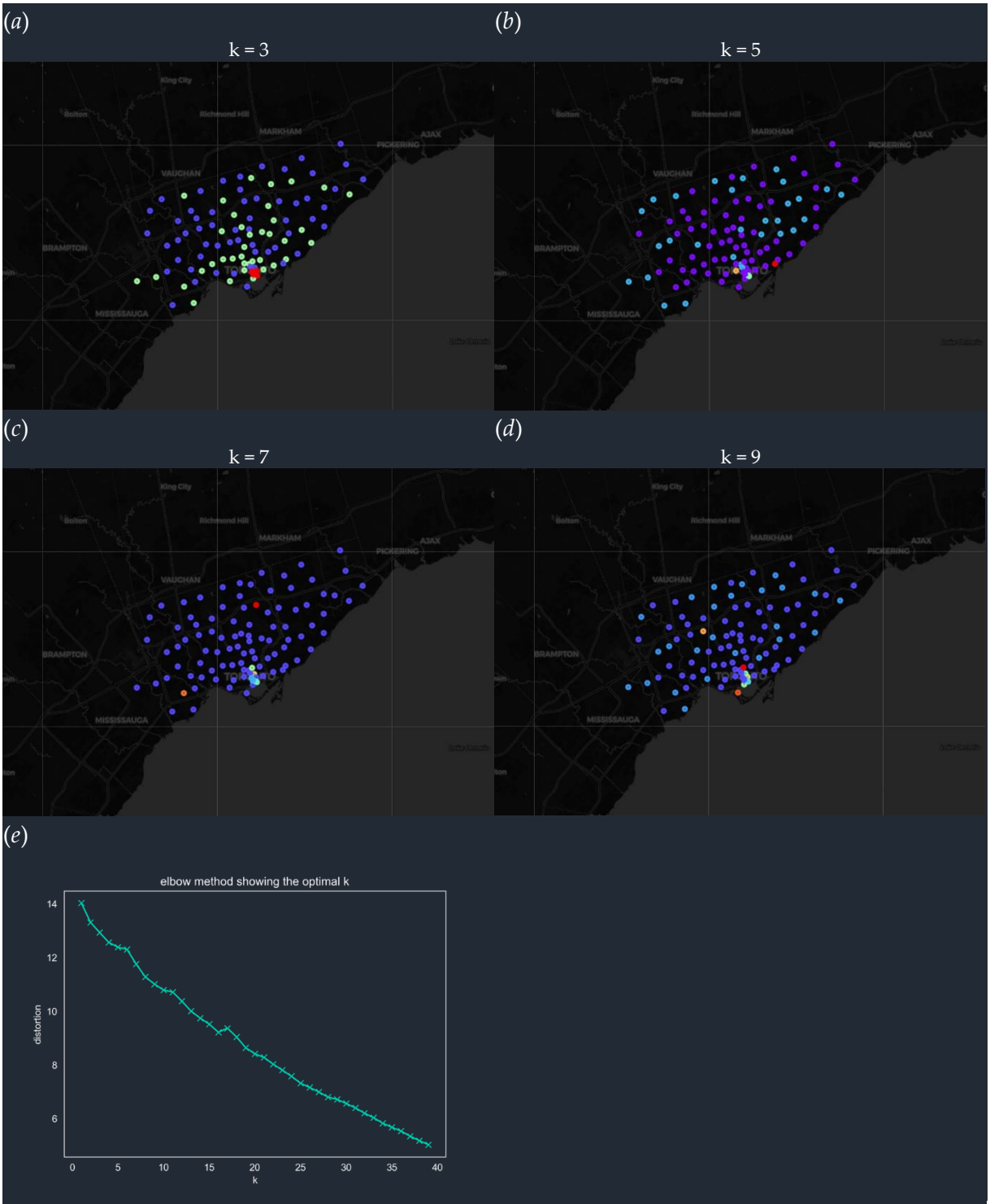


Figure 8. Visualization of the results of neighborhood K-Means clustering (random initialization) with k = 3 (*a*), k = 5 (*b*), k = 7 (*c*), and k = 9 (*d*). Elbow plot showing the distortion/error for k = 1–40 (*e*).

3.8 Building a Predictive Model

We then proceed to build a predictive model for the demand of venues in Toronto (Figure 9). Multiple linear regression models are employed to predict the number of venues in each category in each postal code area. This models healthy neighborhood structures, predicting how many venue types of a certain category should be there based on the amount and types of other venues in the neighborhood. These predictions are then compared to the true venue numbers to identify market demand/saturation.

With the highly detailed categories contained in the data set retrieved from Foursquare, we need to apply means of correcting for the saturation of market demand due to similar types of venues. The categories e.g. contain both "cafès" and "coffee shops" which clearly serve very similar customer demands. We transpose the data matrix and employ K-Means (random initialization) with k = 30 to identify 30 groups of similar venues. These groups are used as comparison to our predictions regarding single venue categories. This effectively reduces the dimensionality of our original data set from 273 to 30 for a second analysis. Our approach is very similar to PCA with the difference that we do not judge the impact of categories on the outcome of our predictions, i.e. the category clusters are not formed for maximized prediction accuracy but for maximum category similarity. We use these clusters to correct our computed market demands for market saturation due to similar venues. We can identify the highest demands by disregarding oversaturated markets (positive true venue counts minus predicted venue counts) and then multiplying the demand of a single venue category in an area with the demand of venues in the whole category cluster.

The resulting product shall be considered true market demand and will guide our decision on which venue category is in the direst demand in which exact area in Toronto. Further details on the built model and its evaluation are discussed in the following.

3.9 Ridge Regression

273 multiple linear regression models are trained, assigning each feature as dependent variable for one model. The remaining 272 variables are used in training the models. As we

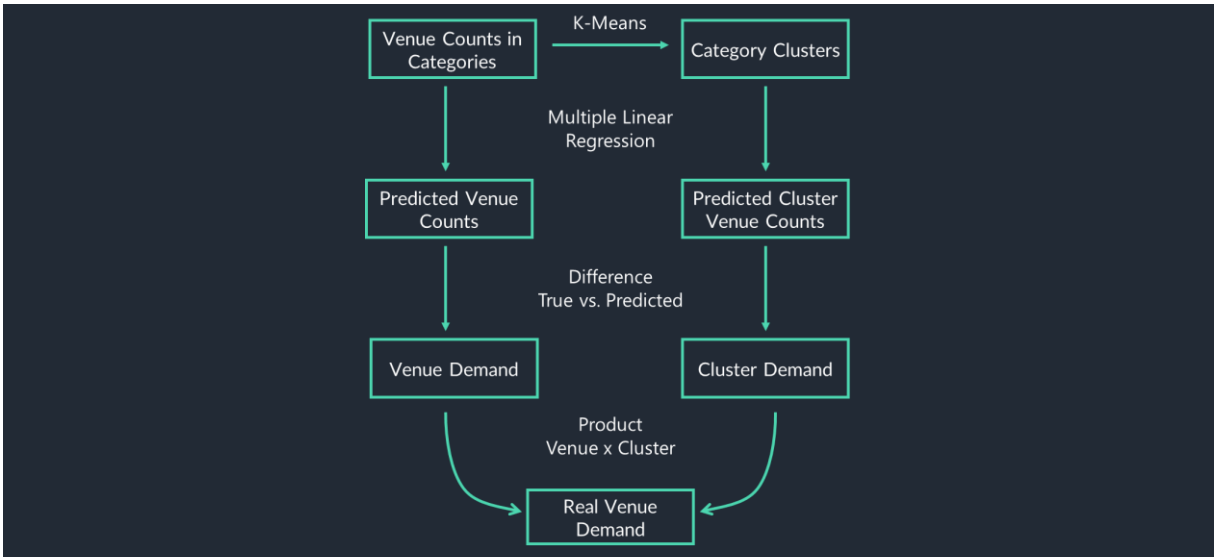


Figure 9. Schematic overview of the model developed for finding the best location and type of venue for a venue to launch in Toronto.

have far more features than training samples, Thikonov regularization was applied automatically within a ridge regression algorithm. The models were trained with 5-fold cross-validation to further counter overfitting. This can be used to identify which venue categories are underrepresented so that we can isolate good places to install new venues.

We now have a dataframe containing predicted values for each of our numeric columns. An exemplary look at the R^2-scores for some variables shows that the models expectedly perform very well the number of venues (R^2-score 1) while failing to model the distance to others (R^2-score 0.14). Within the venue categories, some can be modelled well and others only poorly (examples: dog run (0.86), smoke shop (0.69), mobile phone shop (0.27)).

Next, we computed the differences between true and predicted venue counts as measure for market saturation/demand (Figure 10). This considers the predicted venue counts as the healthy amount of venues of a certain category that should be in an area based on how often it appears in other neighborhoods that have a similar network of different venues in the vicinity.

We can easily identify that certain neighborhoods have a high demand for coffee shops and cafés, but the diagram also brings to mind again that the Foursquare categories are too detailed for simply basing decisions on single category observations. Cafés and coffee shops, obviously, serve almost identical customer demands. The University of Toronto area e.g. has zero coffee shops but five cafés. Three coffee shops are predicted to be in demand there, while an oversaturation of three cafés is computed. This results in a well-balanced coffee demand in the area.

3.10 Category Demand Correction By Category Clustering

Having seen that similar venue types catering the same customer needs need to be accounted for, we went ahead to cluster the venue categories. K-Means clustering with random initialization and a number of clusters k = 30 was used. This is also expected to increase the accuracy of our predictions, as we have been working with 273 dimensions and only 103 samples in our earlier calculations.

This delivered groups of many rather obvious interdependencies and some quite surprising ones which potentially gives interesting insights into the structure of customer groups. We

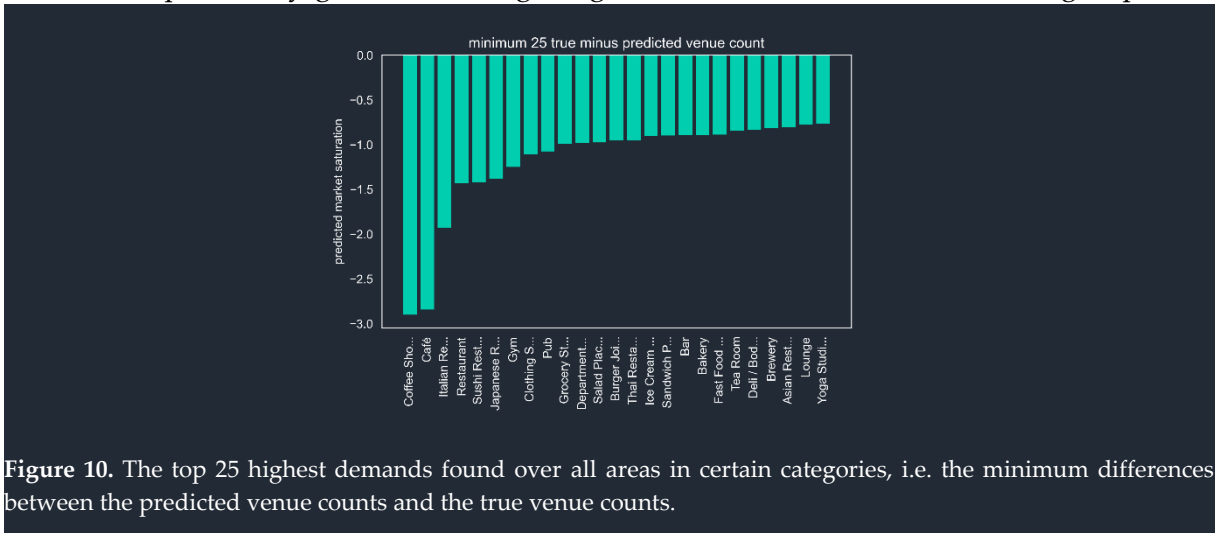


Figure 10. The top 25 highest demands found over all areas in certain categories, i.e. the minimum differences between the predicted venue counts and the true venue counts.

have seen before that coffee shops have the highest overall demand in any location Toronto. Coffee shops were assigned to a cluster containing the following categories:

> General Travel, Modern European Restaurant, Steakhouse, Restaurant, Plaza, Cuban Restaurant, Gift Shop, New American Restaurant, Brazilian Restaurant, Japanese Restaurant, Smoke Shop, French Restaurant, Pub, Art Gallery, Shopping Mall, Speakeasy, Tea Room, Italian Restaurant, Salon/Barbershop, Food Court, Vegetarian/Vegan Restaurant, Seafood Restaurant, Concert Hall, Nightclub, Gluten-free Restaurant, Soup Place, American Restaurant, Bakery, Department Store, Gastropub, Hotel, Coffee Shop, Opera House, Food Truck, Lounge, Asian Restaurant, Art Museum, Cupcake Shop, Train Station, Beer Bar, Colombian Restaurant, Café, Record Shop, Bookstore, Deli/Bodega, Building, Men's Store, Fast Food Restaurant, Wine Bar, Dog Run, Monument/Landmark, Museum, Thai Restaurant, Salad Place.

Many of these can be imagined to serve coffee, e.g. tea rooms and ice cream shops, but some of the appearing categories such as concert halls, general travel, or art galleries seem surprising. Based on our data, however, they appear in similarly structured areas and cater similar needs as cafés and coffee shops. The predicted total market demand of the coffee cluster is the highest of all 30 clusters (Figure 11). This category especially shows demand in downtown areas of Toronto. Coffee shops and cafés are demanded the most (Figure 12).

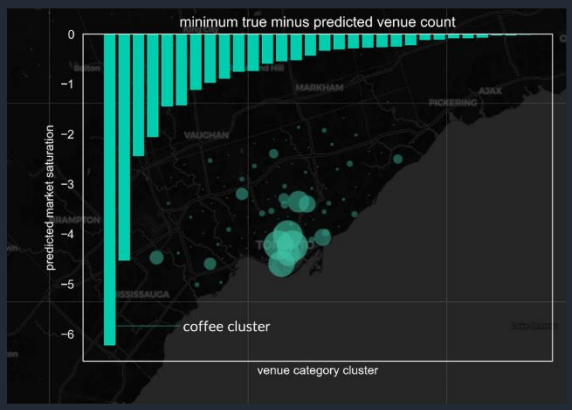


Figure 11. The highest demands found over all areas for our 30 category clusters, i.e. the minimum differences between the predicted venue counts and the true venue counts in each whole cluster. The map in the background shows the demands in all postal code areas for the coffee cluster with the highest found demand.

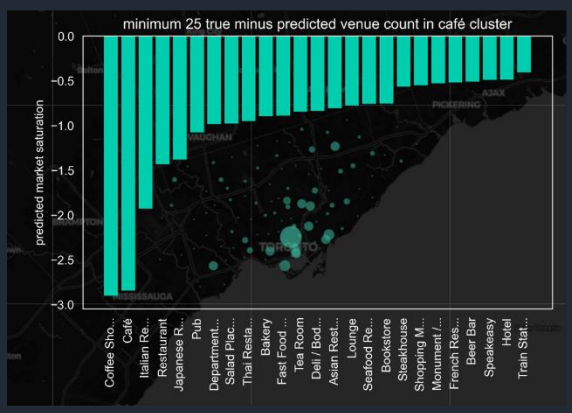


Figure 12. The top 25 highest demands found over all areas in certain categories, i.e. the minimum differences between the predicted venue counts and the true venue counts.

Within this category cluster, there is also a lack found for many food-serving venues and other coffee-serving venues such as tea rooms and bakeries. Combining coffee shops and cafés, a demand for an additional 4.5 venues is found in Church and Wellesley. The whole cluster shows a total demand for 6 venues. This indicates that this area would easily support several additional venues of these types because the infrastructure of other venues in the area is favorable for attracting customers for these venues.

The category demand was multiplied with the demand for venues in the whole category cluster to account for market saturation by venues of similar categories (Figure 13). This highlights demands that are high both in the category combination "coffee shop/café" and in the category cluster that summarizes all categories that are found to cater similar market needs. Church and Wellesley area by far shows the best conditions for launching a coffee shop/café. With 26.7 it has a demand product more than double of any other postal code area. The next best follow-ups have demand products of 10.9, 8.8, and 7.5, respectively.

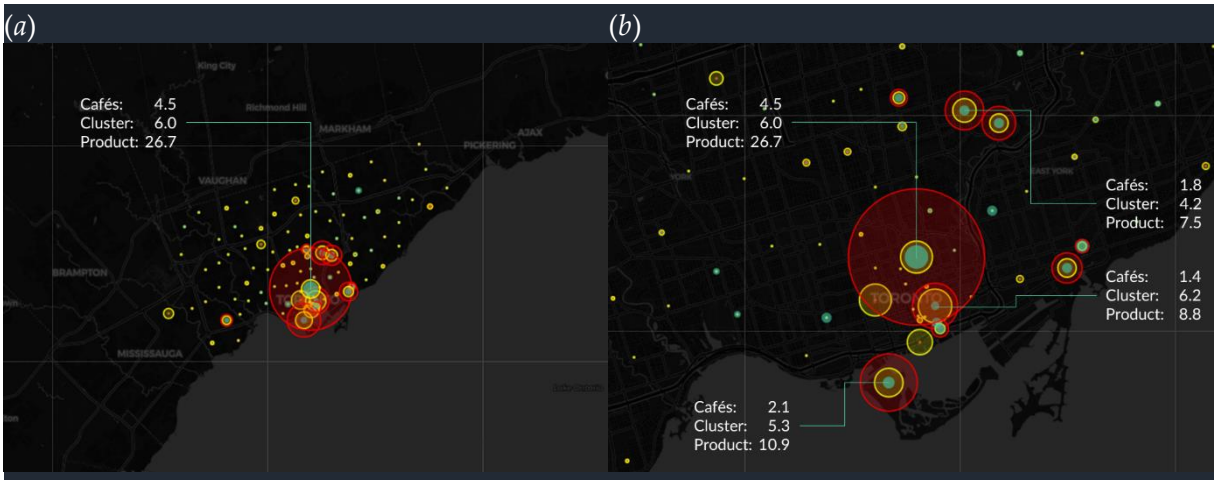


Figure 13. Visualization of the predicted demand for coffee shops/cafès (green), for venues of the whole coffee cluster (yellow) and of the product of both demands (red). (*a*) Overview of Toronto, (*b*) city center extract of the map.

4 CONCLUSION

Toronto shows a diverse culture of stores, restaurants, etc. With an ever-changing city, the market saturation of different categories of venues fluctuates. We found that a timeless classic is in extreme demand right now and found it strongly recommendable to launch coffee shops and cafés in the Church and Wellesley neighborhood in Toronto. This decision is based on data on the amount, local density, and type of existing venues in the area. Clustering algorithms found this area to be extraordinary, four out of five different approaches assigned the area to an exclusive cluster. 273 ridge regression models were used for predicting the venue demand in all categories. This neighborhood was shown to stand out among Toronto's neighborhoods with a strong lack of any type of similar venue so that little competition is to be expected. The venue structure within this neighborhood would even support four to five additional coffee shops.

Figure 14. Coffee can be beautiful.[2]

5 REFERENCES

[1] Course material: https://www.coursera.org/professional-certificates/ibm-data-science, July 12, 2020.

[2] Mike Kenneally via https://unsplash.com, July 12, 2020.

[3] Selected examples: https://towardsdatascience.com/exploring-toronto-neighborhoods-to-open-an-indian-restaurant-ff4dd6bf8c8a, https://capstoneprojectcoursera.wordpress.com/, July 12, 2020.

[4] Selected examples: https://github.com/gnavia007/Coursera_Capstone/, http://roshangrewal.com/capstone-project-the-battle-of-neighborhoods-finding-a-better-place-in-scarborough-toronto/, July 12, 2020.

[5] Selected examples: https://medium.com/@dougm_9851/the-battle-of-neighborhoods-coursera-ibm-capstone-project-52b4292ef410, https://www.linkedin.com/pulse/capstone-project-battle-neighborhoods-rohitaksh-gs/, July 12, 2020.