Applied Data Science Capstone Project
Leslie Lui
2020-05-03

A Data Science-Driven Approach to Gastropub Restaurant Location Selection in Toronto

1. INTRODUCTION

1.1 Background

Around the world, there are few social institutions that typify societies as much as the restaurant. Members of every society desire places they can go to enjoy fine meals and socialize with friends, family, and strangers, and without the need to do any of the cooking and cleaning themselves. However, the restaurant business is a competitive industry to operate in. Every restaurant must compete with numerous other nearby restaurants for the attention and money of prospective customers, and provide a service that meets the cost constraints of their customers while still being profitable.

One recent trend in restaurant industry is the growing popularity of gastropubs. These restaurants fuse characteristics of restaurants with bars in a premium dining and entertainment experience. They offer the quality dining and comfort food menus of upscale and traditional neighborhood restaurants, a wide and premium variety of alcoholic beverages associated with some higher end bars, pubs, and microbreweries, and an upscale, yet comfortable and accessible environment for dining and socializing. When consumers go to a restaurant, they increasingly want more than just simply eating a prepared meal – they want to have a complete experience with dining, socializing, and entertainment, and they are willing to pay a premium for it. Gastropubs are one of the newer venues offering this combination. A new gastropub could be a lucrative business opportunity by offering customers an attractive alternative to traditional restaurants, and provide attractive profits by cashing in on the premium experience associated with gastropubs.

The increased popularity of gastropubs is an international trend. Toronto is a very international city, with a large and diverse population, so it is not surprising that there are already dozens of gastropubs in Toronto.

1.2 Problem

The presence of numerous gastropubs in Toronto proves that they are a viable restaurant concept in this market. However, selecting a suitable location for a gastropub is still important; a good location can allow a gastropub to thrive while a bad location can mean an early end. Instead of merely taking a guess, a data science-driven approach was used in this study to identify suitable locations for a gastropub, based on the features of locations in Toronto that currently have gastropubs.

2. METHODOLOGY

2.1 Overview

The study investigated existing gastropubs in the Toronto area, identified traits about the neighborhoods around them, built a model to determine if any traits are common to neighborhoods with gastropubs, and then used that model to identify other similar neighborhoods that could be good candidates for locating a new gastropub.

A cluster analysis was used to group the neighborhoods. When a cluster was found that effectively represented neighborhoods with gastropubs, it could be assumed that the neighborhoods in this cluster have features that generally make them favorable locations for gastropubs. Neighborhoods in this cluster would be the recommended locations for new gastropubs.

Regression and classification were not considered suitable methods to apply to this study for the following reasons:
1. The intent of this study was to determine which existing Toronto neighborhoods would be the most suitable for a gastropub. Regression and classification would use existing Toronto neighborhoods as training data, and so they would not be helpful for predicting the suitability of existing neighborhoods. They would be more useful for predicting whether new samples outside the training data set (ie. neighborhoods outside Toronto) would be appropriate for gastropubs.
2. The goal of regression and classification would be to create models that accurately predict whether a neighborhood should have a gastropub. If the model was very accurate, it would predict that the only neighborhoods that should have gastropubs are the neighborhoods that already have existing gastropubs, so regression and classification models would not be helpful for suggesting other potential neighborhoods.

2.2 Data Overview

The study data included data on the Toronto neighborhoods and business venues in them.

Two Toronto neighborhood datasets from the Toronto Open Data were used:

1. Toronto neighborhood profiles dataset [1]. This dataset contains demographic data on each neighborhood, taken from census reports. Data include general population features of each neighborhood, such as overall population and land area, as well as more specific population features such as distributions of age, income, family statuses, living conditions, ethnicities, languages spoken, etc. Only a select portion of this data was used for the analysis. This data was included to factor in the effects of a neighborhood's demographics on the likelihood of a gastropub to be located there.
2. Toronto neighborhood boundaries dataset [2]. This dataset contains geographic coordinates on the each of the 140 neighborhoods that comprise Toronto. The central coordinates were used for venue searches (described below), and the boundary coordinates were used for plotting neighborhood boundaries on a map.

The business venue information was provided by the Foursquare API service [3]. The information was obtained in queries sent to the API service through a Jupyter notebook. The information obtained included the following:

1. Gastropub venue search results for Toronto. A search was done in each Toronto neighborhood to identify as many gastropubs as possible. The critical information obtained from the results were the gastropub geographic coordinates, which were used to determine the nearest neighborhood for each one. This gastropub neighborhood information was used to evaluate the effectiveness of the cluster algorithm in generating a cluster representative of neighborhoods with gastropubs.
2. Exploration search results, starting at the central coordinates of each Toronto neighborhood. This provided information on popular venues around each neighborhood center, and thus a profile of the popular venues in each neighborhood.

2.3 Data Analysis Process

2.3.1 Data Analysis Procedures

1. Gather geographic information on the Toronto neighborhoods from the Toronto Open Data website. Edit and format the data to produce the following.
    a. A dictionary of neighborhood boundary coordinates, to allow the neighborhoods to be visualized on maps.
    b. A dataframe containing neighborhood names, neighborhood area numbers, and geographic coordinates of each neighborhood's central area.
2. Gather demographic information on the Toronto neighborhoods from the Toronto Open Data website. Edit and format the data to produce a dataframe containing the following primary demographic details on each neighborhood – population, land area, population age distribution, marital statistics, household counts, household size distribution, family structure, and average income.
3. Search for gastropubs in the Toronto area using the Foursquare API. The central coordinates of each neighborhood were used as starting points for this search. Edit and format the data to contain only venues categorized as gastropubs and only unique instances. Convert this into a dataframe containing gastropub identification information, geographic coordinates, and their nearest associated neighborhood.
4. Explore venues around each neighborhood using the Foursquare API. The central coordinates of each neighborhood were used as starting points for this search. Edit and format this information into a dataframe with venue search results sorted by neighborhood search area, and venue category details. The venue category details will be converted into dummy variables that will be appended to the end of the dataframe.
5. Combine the neighborhood venue and demographic data into a single dataframe for analysis.
6. Perform exploratory analysis on the neighborhood data to identify general trends and potential feature correlations.
7. Perform k-means clustering algorithm on the neighborhood data. Perform the algorithm for a range of cluster sizes.
8. Evaluate the quality of each resultant cluster set to determine a suitable cluster size to analyze.
9. Evaluate the results for the selected cluster size, including visually mapping the cluster neighborhoods.
10. Evaluate the differences within and between clusters.

3. DATA WRANGLING

3.1 Neighborhood Geographic Data

The data was downloaded from the Toronto Open Data website as a geojson file. It was read into the Jupyter notebook in string format and the **json.loads** function was used to convert into a dictionary named *tor_nh_coord_dict*. The beginning of this dictionary is shown to the right.


```
{'type': 'FeatureCollection',
 'crs': {'type': 'name',
  'properties': {'name': 'urn:ogc:def:crs:OGC:1.3:CRS84'}},
 'features': [{'type': 'Feature',
   'properties': {'_id': 4341,
    'AREA_ID': 25886861,
    'AREA_ATTR_ID': 25926662,
    'PARENT_AREA_ID': 49885,
    'AREA_SHORT_CODE': 94,
    'AREA_LONG_CODE': 94,
    'AREA_NAME': 'Wychwood (94)',
    'AREA_DESC': 'Wychwood (94)',
    'X': None,
    'Y': None,
    'LONGITUDE': -79.425514947,
    'LATITUDE': 43.6769192679,
    'OBJECTID': 16491505,
    'Shape__Area': 3217959.609375,
    'Shape__Length': 7515.779658331329},
   'geometry': {'type': 'Polygon',
    'coordinates': [[[-79.4359157087306, 43.68015339477749],
      [-79.4349150633973, 43.6803688699489],
      [-79.4339472722385, 43.6805785044903],
      [-79.433881624222, 43.6805899612147],
      [-79.4328154497888, 43.6808080444588],
      [-79.4326971769691, 43.6807965882232],
      [-79.4325306465987, 43.680827859628 71],
      [-79.4324594444965, 43.680858827103606],
      [-79.4324113272399, 43.6808979453726],
```


Figure 1. *tor_nh_coord_dict* Dictionary Structure and Data

Nested within the dictionary were neighborhood details, including neighborhood names, central coordinates, and boundary coordinates. The neighborhood names, neighborhood numbers, and central coordinates were output to a dataframe named *tor_nh*. The first few rows of this dataframe are shown below.

	Area Short Code	Neighborhood	Neighborhood Longitude	Neighborhood Latitude
0	94	Wychwood (94)	-79.425515	43.676919
1	100	Yonge-Eglinton (100)	-79.403590	43.704689
2	97	Yonge-St.Clair (97)	-79.397871	43.687859
3	27	York University Heights (27)	-79.488883	43.765736
4	31	Yorkdale-Glen Park (31)	-79.457108	43.714672

Figure 2. Head Rows for Dataframe *tor_nh*

While the neighborhood boundary coordinates could be exported to dataframe *tor_nh* as well, it could only be exported in a string format, which could not be directly used for generating a choropleth map. Therefore it was left out of dataframe *tor_nh*, and dictionary *tor_nh_coord_dict* was retained for mapping the neighborhood boundaries.

A Folium map was generated using the data in dictionary *tor_nh_coord_dict* and dataframe *tor_nh* to verify the fidelity of the data and structures. The dictionary provided the neighborhood boundaries and the dataframe provided the coordinates and descriptions for the markers in each neighborhood.

The map provided visual confirmation that the neighborhood geographic data were properly formatted and accurate.

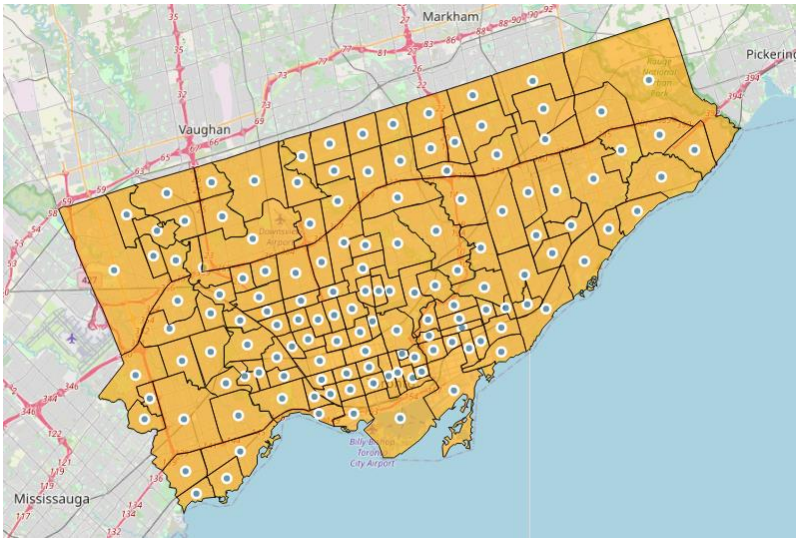

Figure 3. Folium Map of Toronto Neighborhoods, Data from *tor_nh_coord_dict* Dictionary and *tor_nh* Dataframe

3.2 Neighborhood Demographic Profile Data

The data was downloaded from the Toronto Open Data website as a csv file. It was read into the Jupyter notebook as a dataframe. An overview of the initial raw dataframe is shown below.

	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	...	Willowdale West	Willowridge-Martingrove-Richview	Woburn	Woodbine Corridor	Woodbine-Lumsden	Wychwood	Yonge-Eglinton	Yonge-St.Clair	York University Heights	Yorkdale-Glen Park
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	...	37	7	137	64	60	94	100	97	27	31
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	...	No Designation	No Designation	NIA	No Designation	No Designation	No Designation	No Designation	No Designation	NIA	Emerging Neighbourhood
2	3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016	2,731,571	29,113	23,757	12,054	30,526	...	16,936	22,156	53,485	12,541	7,865	14,349	11,817	12,528	27,593	14,804
3	4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011	2,615,060	30,279	21,988	11,904	29,177	...	15,004	21,343	53,350	11,703	7,826	13,986	10,578	11,652	27,713	14,687
4	5	Population	Population and dwellings	Census Profile 98-316-X2016001	Population Change 2011-2016	4.50%	-3.90%	8.00%	1.30%	4.60%	...	12.90%	3.80%	0.30%	7.20%	0.50%	2.60%	11.70%	7.50%	-0.40%	0.80%
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2378	2379	Mobility	Mobility status - Place of residence 5 years ago	Census Profile 98-316-X2016001	Migrants	400,950	3,170	3,145	925	6,390	...	3,765	2,270	7,260	985	620	1,350	2,425	2,310	4,965	1,345
2379	2380	Mobility	Mobility status - Place of residence 5 years ago	Census Profile 98-316-X2016001	Internal migrants	184,120	880	980	680	3,930	...	1,545	1,110	1,720	610	395	780	1,260	1,355	1,700	580
2380	2381	Mobility	Mobility status - Place of residence 5 years ago	Census Profile 98-316-X2016001	Intraprovincial migrants	141,135	735	760	615	2,630	...	1,070	960	1,400	350	320	570	970	1,025	1,490	445
2381	2382	Mobility	Mobility status - Place of residence 5 years ago	Census Profile 98-316-X2016001	Interprovincial migrants	42,985	135	220	70	1,310	...	475	150	335	250	85	210	290	325	195	135
2382	2383	Mobility	Mobility status - Place of residence 5 years ago	Census Profile 98-316-X2016001	External migrants	216,835	2,280	2,170	245	2,460	...	2,220	1,175	5,540	395	220	575	1,160	955	3,285	775

2383 rows × 146 columns

Figure 4. Raw Data from Toronto Neighborhood Demographic Profile Data Set

There were nearly 2400 different sets of statistics in the raw data set. Approximately 30 were of primary interest and were sliced out into a new dataframe *tor_nh_profiles*.

	0, Neighborhood	1, NH Number	2, Pop	3, Pop Change	4, Pop Density	5, Children	6, Youth	7, Working Age	8, Pre-Retire	9, Seniors	...	18, 4 Person Household	19, 5+ Person Household	20, Avg Household Size	21, Families	22, Couple Families	23, Single Families	24, Couples without Children	25, Couples with Children	26, Average Income	27, Land Area
0	Agincourt North	129	29,113	-3.90%	3,929	3,840	3,705	11,305	4,230	6,045	...	1750	1645	3.16	8425	6770	1655	2520	4250	427,037	7.41
1	Agincourt South-Malvern West	128	23,757	8.00%	3,034	3,075	3,360	9,965	3,265	4,105	...	1335	1175	2.88	6640	5275	1365	2095	3180	278,390	7.83
2	Alderwood	20	12,054	1.30%	2,435	1,760	1,235	5,220	1,825	2,015	...	795	390	2.6	3490	2815	665	1175	1645	168,602	4.95
3	Annex	95	30,526	4.60%	10,863	2,360	3,750	15,040	3,480	5,910	...	885	390	1.8	6700	5805	895	3830	1975	792,507	2.81
4	Banbury-Don Mills	42	27,695	2.90%	2,775	3,605	2,730	10,810	3,555	6,975	...	1515	675	2.23	7605	6350	1245	3015	3335	493,486	9.98
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
135	Wychwood	94	14,349	2.60%	8,541	1,860	1,320	6,420	1,595	3,150	...	750	330	2.24	3475	2730	750	1320	1405	239,484	1.68
136	Yonge-Eglinton	100	11,817	11.70%	7,162	1,800	1,225	5,860	1,325	1,600	...	690	220	2.08	3065	2640	425	1290	1350	222,648	1.65
137	Yonge-St.Clair	97	12,528	7.50%	10,708	1,210	920	5,960	1,540	2,905	...	395	135	1.76	3160	2775	390	1850	925	541,217	1.17
138	York University Heights	27	27,593	-0.40%	2,086	4,045	4,750	12,290	2,965	3,530	...	1495	1325	2.71	7000	4920	2075	1955	2960	302,358	13.23
139	Yorkdale-Glen Park	31	14,804	0.80%	2,451	1,960	1,870	5,860	1,810	3,295	...	780	590	2.62	3930	2935	1005	1245	1700	213,860	6.04

140 rows × 28 columns

Figure 5. Initial Version of Dataframe *tor_nh_profiles*

Much of the demographic data was recorded as absolute counts, not population proportions. For analytical and comparative purposes, normalizing the data relative to their relevant populations was necessary. The numeric data in dataframe *tor_nh_profiles* was initially in object format and needed to be converted to floating point values in order to perform normalization. Before this, extra commas and percent signs had to be removed from the string data in order to be converted to floating point values. The formatted, normalized data was output to a new dataframe named *tor_nh_profiles_stats*. The first few rows of this dataframe are shown below.

	0, Neighborhood	1, NH Number	2, Pop	3, Pop Change	4, Pop Density	5, Children	6, Youth	7, Working Age	8, Pre-Retire	9, Seniors	...	19, 5+ Person Household	20, Avg Household Size	21, Families	22, Couple Families	23, Single Families	24, Couples without Children	25, Couples with Children	26, Average Income	27, Land Area	Housing Density
0	Agincourt North	129	29113.0	-3.9	3929.0	0.131900	0.127263	0.388314	0.145296	0.207639	...	0.180274	3.16	8425.0	0.803561	0.196439	0.372230	0.627770	427037.0	7.41	1231.443995
1	Agincourt South-Malvern West	128	23757.0	8.0	3034.0	0.129436	0.141432	0.419455	0.137433	0.172791	...	0.144349	2.88	6640.0	0.794428	0.205572	0.397156	0.602844	278390.0	7.83	1039.591315
2	Alderwood	20	12054.0	1.3	2435.0	0.146010	0.102456	0.433051	0.151402	0.167164	...	0.084416	2.60	3490.0	0.806590	0.190544	0.417407	0.584369	168602.0	4.95	933.333333
3	Annex	95	30526.0	4.6	10863.0	0.077311	0.122846	0.492695	0.114001	0.193605	...	0.024490	1.80	6700.0	0.866418	0.133582	0.659776	0.340224	792507.0	2.81	5667.259786
4	Banbury-Don Mills	42	27695.0	2.9	2775.0	0.130168	0.098574	0.390323	0.128363	0.251851	...	0.055670	2.23	7605.0	0.834977	0.163708	0.474803	0.525197	493486.0	9.98	1214.929860

5 rows × 29 columns

Figure 6. Head Rows, Final Version of Dataframe *tor_nh_profiles_stats*

3.3 Gastropub Venue Data

The gastropub venue search was performed with the Foursquare API. Due to API search return limits, a single search from the center of the city, with a large search radius, would only provide a limited number of relevant results. Instead a distributed search was performed from the central coordinates of every neighborhood, with a shorter search radius, to reduce the likelihood of maxing out the search return limit for each.

Instead of assigning an arbitrary search radius, more specific values were calculated for consideration. It was desired for the radius to be small enough to avoid maxing out the search return limit, while being large enough to avoid gaps between neighborhood search areas. Candidate values were determined by calculating the distances between the centers of every neighborhood in Toronto, using the coordinate data in dataframe *tor_nh*. The **geodesic** function was used to calculate the distances between coordinate sets. The distances between adjacent neighborhoods (which is the smallest non-zero distance for each neighborhood) were the candidate values to consider for the search radius, as they wouldn't be excessively large but would prevent search gaps. The values ranged from 591 to 3455m, with an average of 1544m. The maximum of 3455m was selected since it did not seem to be a very large value.

The Foursquare API search type was set to a search by venue category ID. It was found in trial search runs that searching by the gastropub category ID (listed on the Foursquare API developer website [4]) instead of a keyword search returned more relevant results.

The search returned results in JSON format, which was converted into a dataframe. The raw data is shown below.

	id	name	categories	referralId	hasPerk	location.address	location.crossStreet	location.lat	location.lng	location.labeledLatLngs	location.distance	location.postalCode	location.cc	location.city	location.state	location.country	locatio
0	4cbf015897bc721e6fe08267	Chadwick's	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277898	False	268 Howland Ave	at Dupont St	43.673593	-79.412015	[{'label': 'display', 'lat': 43.67359329474282...	1148	M5R 3B6	CA	Toronto	ON	Canada	[Du
1	4f9b0e84e4b0d04dbf9767ee	The Oxley Public House	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277898	False	121 Yorkville Ave	Hazelton Ave	43.670537	-79.392977	[{'label': 'display', 'lat': 43.67053669383452...	2714	M5R 1C4	CA	Toronto	ON	Canada	[121 Yo
2	506f3aa3e4b0b0ed3e1dc405	Indie Alehouse	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277898	False	2876 Dundas St W	at Keele St	43.665475	-79.465290	[{'label': 'display', 'lat': 43.66547472315272...	3446	M6P 1Y9	CA	Toronto	ON	Canada	[2 Keele S
3	4e235b5cd22d0a3f5a0bc0a2	Folly Brewpub	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277898	False	928 College St	at Dovercourt Ave	43.653806	-79.425850	[{'label': 'display', 'lat': 43.65380564851785...	2573	M6H 1A4	CA	Toronto	ON	Canada	Dow
4	4ad4c05df964a5204df620e3	Rebel House	[{'id': '4bf58dd8d48988d116941735', 'name': 'B...	v-1588277898	False	1068 Yonge St.	btwn Roxborough St. & Gibson Ave.	43.677661	-79.389935	[{'label': 'display', 'lat': 43.67766092876885...	2865	M4W 2L4	CA	Toronto	ON	Canada	[Roxb
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
0	4e569dfd1495eb38e20ac5c5	Monk's Kettle	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277935	False	3073 Bloor St. W.	at Brentwood Rd. S.	43.646806	-79.513775	[{'label': 'display', 'lat': 43.64680589371333...	790	M8X 1C7	CA	Toronto	ON	Canada	[Brent
1	506f3aa3e4b0b0ed3e1dc405	Indie Alehouse	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277935	False	2876 Dundas St W	at Keele St	43.665475	-79.465290	[{'label': 'display', 'lat': 43.66547472315272...	3882	M6P 1Y9	CA	Toronto	ON	Canada	[2 Keele S
2	4ed549aa9911957ff17fbc4	Henry VIII Ale House	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277935	False	2469 Bloor Street West	Jane	43.649024	-79.484928	[{'label': 'display', 'lat': 43.64902448298273...	2125	NaN	CA	Toronto	ON	Canada	[24 (Jan
3	4d4c6033e1ec6dcb878bd875	Shakey's Original Bar & Grill	[{'id': '4bf58dd8d48988d155941735', 'name': 'G...	v-1588277935	False	2255 Bloor Street West	Runnymede	43.650771	-79.477581	[{'label': 'display', 'lat': 43.6507708866875,...	2675	M6S 1N8	CA	Toronto	ON	Canada	[22 (Runn
4	4ae3928cf964a520ce9621e3	My Place - a Canadian Pub	[{'id': '4bf58dd8d48988d11b941735', 'name': 'P...	v-1588277935	False	2448 Bloor st west	Jane st	43.648458	-79.485187	[{'label': 'display', 'lat': 43.64845840072295...	2121	NaN	CA	Toronto	ON	Canada	[244 st

1697 rows × 19 columns

Figure 7. Raw Gastropub Venue Search Results from Foursquare API

The **value_counts** method was applied to count how many results were returned from each neighborhood. Only two of the 140 neighborhoods returned the maximum 50 search limits, so the selected search radius value was acceptable.

The raw data contained 19 columns, and most were not needed. The first data cleanup step was to create a new dataframe *gastropub_search_results* containing only the columns of interest – Venue ID, name, location, venue category, and venue geographic coordinates.

The search returned a large number of results, but many individual venues were found multiple times because of overlapping search areas. The goal was to create a unique list of gastropubs in the Toronto area, so duplicate venue search results were removed from the dataframe using the **drop_duplicates** method. The unique venue ID's in the

Venue ID column were used to identify duplicates. Multiple venues may share the same name, but every venue has a unique venue ID.

While a search by category ID was requested and the ID for gastropub was used, the search results returned many venues which were not categorized as gastropubs in the *Venue Category* column. Some belonged to similar categories like pubs and bars, but it was desired that this analysis focus on venues that were specifically categorized as gastropubs. Therefore, venues with other venue categories were dropped from the dataframe.

The final edit to the dataframe was the addition of a column indicating the closest neighborhood to each gastropub. While the coordinates of the gastropub and the coordinates of the neighborhood boundaries were available, a method for determining which neighborhood boundary contained each gastropub was not available. Instead, the distances between each gastropub and every neighborhood's central coordinates were calculated, and the gastropub was assigned to the neighborhood with the closest central coordinates.

The formatted data was saved to a dataframe named *gastropub_list_*. The first few rows of are shown below.

	Venue ID	Venue Name	Venue Category	Venue Address	Venue Latitude	Venue Longitude	Nearest Neighborhood
0	4cbf015897bc721e6fe08267	Chadwick's	Gastropub	268 Howland Ave	43.673593	-79.412015	Annex (95)
1	4f9b0e84e4b0d04dbf9767ee	The Oxley Public House	Gastropub	121 Yorkville Ave	43.670537	-79.392977	Annex (95)
2	506f3aa3e4b0b0ed3e1dc405	Indie Alehouse	Gastropub	2876 Dundas St W	43.665475	-79.465290	Junction Area (90)
3	4e235b5cd22d0a3f5a0bc0a2	Folly Brewpub	Gastropub	928 College St	43.653806	-79.425850	Little Portugal (84)
4	4fdc0e98e4b05197cd14912b	The Abbot	Gastropub	508 Eglinton Ave W	43.703688	-79.413485	Yonge-Eglinton (100)

Figure 8. Head Rows, Dataframe *gastropub_list_*

The gastropub coordinates in *gastropub_list_* were plotted in a new Folium map to visually verify they were accurate. This map also contained the neighborhood geographic data in the previous Folium map.

The gastropubs are identified by circular markers outlined in blue. The map verified the data is valid.

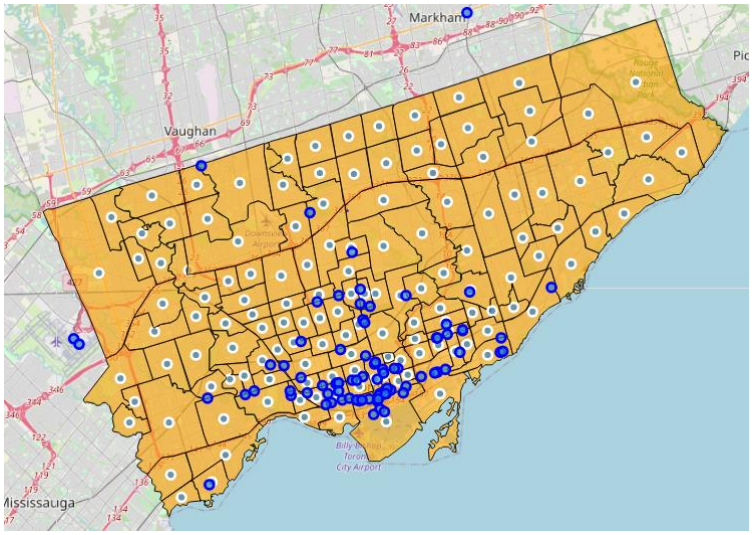

Figure 9. Folium Map of Gastropubs, Overlaid on Toronto Neighborhood Data

	Neighborhood	Number of Nearby GP	GP Nearby
0	Agincourt North (129)	0	0
1	Agincourt South-Malvern West (128)	0	0
2	Alderwood (20)	0	0
3	Annex (95)	2	1
4	Banbury-Don Mills (42)	0	0
...	...	...	...
135	Wychwood (94)	0	0
136	Yonge-Eglinton (100)	1	1
137	Yonge-St.Clair (97)	2	1
138	York University Heights (27)	0	0
139	Yorkdale-Glen Park (31)	0	0

140 rows × 3 columns

Figure 10. Dataframe *gp_nh_present*

A new dataframe *gp_nh_present* was also generated from the data in *gastropub_list_*. It contained three data columns – a list of all Toronto neighborhoods, a count of the gastropubs near each one, and a dummy variable indicating whether or not there is a gastropub near each neighborhood. This dataframe was used later on to evaluate the effectiveness of the clustering algorithm in grouping neighborhoods with gastropubs.

3.4 Neighborhood Venue Data

An exploratory search was performed around each neighborhood using the Foursquare API. Exploratory search returns the most popular venues, regardless of venue category, within the search radius of each neighborhood. The search radius was reduced to 1000m to reduce the likelihood of maxing out the search results limit.

As with the gastropub search, the neighborhood venue search results were returned in JSON format. This data was converted into a dataframe named *nh_explore_venues*. The dataframe data included names and coordinates of the neighborhoods where the searches originated, search result venue ID, name, coordinates, location, and venue category.

The search results included several venues that were categorized as “Neighborhood” and were actually other Toronto neighborhoods. These were removed from the dataframe because they were not considered valid venues and the venue category name would conflict with the primary *Neighborhood* column used to identify each row.

The information in the *Venue Category* column was one of the primary features considered in this analysis. However, they could not be used in the analysis as is, in a single column. Instead, the venue category information had to be transformed into a new dataframe of dummy variable columns using the **get_dummies** method. One column was created for each venue category; the value was set to 1 in rows for venues that match this category and 0 in all other rows. The first few rows of this dummy variable dataframe, named *nh_explore_venues_dummies*, are shown below.

```
nh_explore_venues_dummies=pd.get_dummies(nh_explore_venues['Venue Category'])
nh_explore_venues_dummies.head()
```

	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	...	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 347 columns

Figure 11. Head Rows, Dataframe *nh_explore_venues_dummies*

Dataframe *nh_explore_venues_dummies* was appended to the previous dataframe *nh_explore_venues* to create a single dataframe for manipulation. An additional column *Venue Count*, with an integer value of 1 in every row, was also added to keep track of venue counts in each neighborhood later on when the venues were grouped by neighborhood.

The **groupby** method was used on dataframe *nh_explore_venues* to create two summary dataframes. In the first dataframe *nh_area_venue_sum,* search results were grouped by neighborhood and their venue details were summed to create total counts of venues within each category for each neighborhood. The summation also calculated the total venue count in each neighborhood.

```
# Sum of venues by neighborhood and venue category.

nh_area_venue_sum=nh_explore_venues.groupby('Neighborhood').sum()
nh_area_venue_sum.reset_index(inplace=True)
nh_area_venue_sum
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Latitude	Venue Longitude	Venue Count	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	...	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	Agincourt North (129)	1883.633945	-3408.468623	1883.644489	-3408.558218	43	0	0	0	0	...	2	0	0	0	0	1	0	0	0	0
1	Agincourt South-Malvern West (128)	1707.757644	-3091.358860	1707.754242	-3091.469615	39	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Alderwood (20)	1046.518486	-1908.998662	1046.554876	-1909.068996	24	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Annex (95)	4367.158544	-7940.400062	4366.976440	-7940.264313	100	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Banbury-Don Mills (42)	1924.456908	-3491.387591	1924.385600	-3491.261795	44	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
135	Wychwood (94)	3887.245815	-7068.870830	3887.426974	-7068.840147	89	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
136	Yonge-Eglinton (100)	4370.468937	-7940.359017	4370.576932	-7940.021723	100	0	0	0	0	...	1	0	0	1	0	0	0	1	0	0
137	Yonge-St.Clair (97)	3451.340851	-6272.431791	3451.322364	-6272.143887	79	0	0	0	0	...	1	0	0	1	0	0	0	1	0	0
138	York University Heights (27)	919.080466	-1669.266542	919.067994	-1669.339542	21	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
139	Yorkdale-Glen Park (31)	2054.589562	-3734.484079	2054.679491	-3734.375296	47	0	1	0	0	...	1	0	0	0	0	0	0	0	0	0

140 rows × 353 columns

Figure 12. Head Rows, Dataframe *nh_area_venue_sum*

The **value_counts** method was applied to the *Venue Count* column of dataframe *nh_area_venue_sum* to evaluate how many search results were returned among the neighborhoods. Fourteen of the 140 neighborhoods reached the 100 results limit. While not desirable, they only accounted for 10% of all neighborhoods, so it was considered acceptable to proceed.

In the second dataframe *nh_area_venue_mean,* search results were grouped by neighborhood and mean values of their venue details were calculated. The mean value is the proportion of all venues in each neighborhood that each venue category represents.

Every value in the *Venue Count* column of this dataframe was 1.0 due to the mean calculation. These values were replaced by the ones in the *Venue Count* column of dataframe *nh_area_venue_sum*, since it actually contains the total venue count in each neighborhood. These values were then divided by 100 to normalize them to the highest venue count in all of the neighborhoods. The first few rows of the finished dataframe *nh_area_venue_mean* are shown below.

```
nh_area_venue_mean['Venue Count']=nh_area_venue_sum['Venue Count']/(nh_area_venue_sum['Venue Count'].max())
nh_area_venue_mean
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Latitude	Venue Longitude	Venue Count	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	...	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	Agincourt North (129)	43.805441	-79.266712	43.805686	-79.268796	0.43	0.0	0.000000	0.0	0.0	...	0.046512	0.0	0.0	0.000000	0.0	0.023256	0.000000	0.000000	0.0	0.0
1	Agincourt South-Malvern West (128)	43.788658	-79.265612	43.788570	-79.268452	0.39	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0
2	Alderwood (20)	43.604937	-79.541611	43.606453	-79.544542	0.24	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0
3	Annex (95)	43.671585	-79.404001	43.669764	-79.402643	1.00	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0
4	Banbury-Don Mills (42)	43.737657	-79.349718	43.736036	-79.346859	0.44	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.022727	0.000000	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
135	Wychwood (94)	43.676919	-79.425515	43.678955	-79.425170	0.89	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.011236	0.0	0.0
136	Yonge-Eglinton (100)	43.704689	-79.403590	43.705769	-79.400217	1.00	0.0	0.000000	0.0	0.0	...	0.010000	0.0	0.0	0.010000	0.0	0.000000	0.000000	0.010000	0.0	0.0
137	Yonge-St.Clair (97)	43.687859	-79.397871	43.687625	-79.394226	0.79	0.0	0.000000	0.0	0.0	...	0.012658	0.0	0.0	0.012658	0.0	0.000000	0.000000	0.012658	0.0	0.0
138	York University Heights (27)	43.765736	-79.488883	43.765143	-79.492359	0.21	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0
139	Yorkdale-Glen Park (31)	43.714672	-79.457108	43.716585	-79.454794	0.47	0.0	0.021277	0.0	0.0	...	0.021277	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0

140 rows × 353 columns

Figure 13. Head Rows, Dataframe *nh_area_venue_mean*

3.5 Neighborhood Venue and Geographic Data Merge

The dataframes *tor_nh_profiles_stats* and *nh_area_venue_mean* were merged to create a single neighborhood dataframe for the cluster analysis.

The two dataframes had different neighborhood identifier values, so their rows would be mismatched if they were directly merged. The *Neighborhood* column in dataframe *nh_area_venue_mean* specified each neighborhood's name along with its area number (eg. "Agincourt North (129)"). Dataframe *tor_nh_profiles_stats* had two neighborhood identifier columns, *Neighborhood* and *1, NH Number*, which respectively listed the neighborhood names and area numbers individually (eg. "Agincourt North" and "129"). To resolve this, dataframe *tor_nh* was used as a bridge; its *Neighborhood* column specified the combined neighborhood name and area number and its *Area Short Code* column specified the area number.

Dataframes *tor_nh* and *nh_area_venue_mean* were merged first. The resultant dataframe, was then merged with dataframe *tor_nh_profiles_stats* to create a new dataframe *nh_stats_venues* where data rows from the previous dataframes lined up by neighborhood. Several unnecessary and redundant neighborhood identification columns were dropped from the dataframe afterward.

The final formatting applied to dataframe *nh_stats_venues* was to re-sort the rows by neighborhood name in alphabetical order, to match the neighborhood order in dataframe *gp_nh_present*. These dataframes would be used together later so it was important they lined up.

3.6 Neighborhood Data Exploratory Analysis

In preparation for the clustering algorithm, correlation analysis was performed on the data to identify any close correlations early on. Close correlations would provide some initial insights, and the clustering could be simplified by eliminating some features that were closely correlated with others.

Dataframes *nh_stats_venues* and *gp_nh_present* were merged into a new dataframe *nh_stats_venues2*, to evaluate whether the number or presence of gastropubs near a neighborhood correlated to any neighborhood features.

The Pearson correlation between every pair of features in *nh_stats_venues2* was calculated. The Pearson correlation coefficients and p-values were output to new dataframes *nh_stats_venues_pearson_coef* and *nh_stats_venues_pearson_pval*, respectively.

Each dataframe contained 377 X 377 values, so manual evaluation was too difficult. To facilitate review, the correlation data was filtered to select only feature pairs with high correlation scores (coefficients <-0.8 or >+0.8, and p-value<0.05) and remove correlations between each feature and itself.

Filtering reduced the number of correlations to evaluate from 142129 to 238, and this was further cut in half to 119 because every feature pair was evaluated twice. The data for these highly correlated feature pairs were output to a new dataframe *nh_stats_venues_corr_features_* for closer evaluation.

	Feature 1	Feature 2	Pearson Correlation F-Test Score	Pearson Correlation p-value	Feature 1 Row	Feature 2 Column
0	11, Marriage Age Population	2, Pop	0.994533	2.689355e-137	9	1
1	14, Households	2, Pop	0.906656	1.406035e-53	12	1
2	21, Families	2, Pop	0.979924	1.558544e-98	19	1
3	Housing Density	4, Pop Density	0.982339	2.445181e-102	26	3
4	24, Couples without Children	5, Children	-0.845612	1.909128e-39	22	4
...	...	...	...	...	...	...
114	Street Art	Souvlaki Shop	1.000000	0.000000e+00	332	322
115	Warehouse Store	Spanish Restaurant	0.927212	1.013128e-60	366	324
116	Theme Park	Swiss Restaurant	0.864124	5.536086e-43	350	337
117	Tibetan Restaurant	Tattoo Parlor	0.945298	5.216987e-69	353	344
118	Zoo Exhibit	Zoo	0.969028	1.060107e-85	374	374

119 rows × 6 columns

Figure 14. Initial Format of Dataframe *nh_stats_venues_corr_features_*

The data was sorted, starting with feature pairs with the highest correlation coefficients. The transposed dataframe is shown below.

	74	104	32	112	41	94	95	101	69	34	...	4	16	25	22	11	12	14	27	28	9
Feature 1	Indie Theater	Shop & Service	Housing Development	Theme Restaurant	Doner Restaurant	Opera House	Tanning Salon	Tanning Salon	Government Building	Outdoors & Recreation	...	24, Couples without Children	25, Couples with Children	24, Couples without Children	24, Couples without Children	17, 3 Person Household	18, 4 Person Household	20, Avg Household Size	23, Single Families	25, Couples with Children	13, Not Married
Feature 2	Egyptian Restaurant	Peruvian Restaurant	Afghan Restaurant	Sake Bar	Belgian Restaurant	Monument / Landmark	Monument / Landmark	Opera House	Coworking Space	Amphitheater	...	5, Children	15, 1 Person Household	20, Avg Household Size	18, 4 Person Household	15, 1 Person Household	15, 1 Person Household	15, 1 Person Household	22, Couple Families	24, Couples without Children	12, Married
Pearson Correlation F-Test Score	1	1	1	1	1	1	1	1	1	1	...	-0.845612	-0.857722	-0.872863	-0.895952	-0.924707	-0.929787	-0.962486	-0.99977	-0.999823	-0.999968
Pearson Correlation p-value	0	0	0	0	0	0	0	0	0	0	...	1.90913e-39	1.05432e-41	7.70967e-45	1.72744e-50	9.59261e-60	9.22858e-62	4.68244e-80	3.8038e-232	5.11354e-240	2.20159e-291
Feature 1 Row	203	308	197	351	130	254	341	341	170	259	...	22	23	22	22	15	16	18	21	23	11
Feature 2 Column	135	268	32	299	60	241	241	255	114	35	...	4	14	19	17	14	14	14	21	23	11

6 rows × 119 columns

Figure 15. Dataframe *nh_stats_venues_corr_features_sorted*, Transposed View

Many feature pairs had perfect correlation (coefficients =+1, and p-value=0) or were nearly perfect. It turned out that all of these perfect/nearly perfect correlations were for pairs of venue categories where each one had only a single representative venue in all of Toronto, and both happened to be in the search area for one neighborhood or two adjacent neighborhoods. It was decided that all of these highly correlated venue categories would remain in the analysis data set because their limited representation prevented any conclusive determination of correlation. Also, their removal would affect the overall venue count in each neighborhood and skew the proportions of other venues.

Since the venue category features would be left alone, the focus shifted to the highly correlated demographic features.

Approximately 30 highly correlated feature pairs were found that related to demographic features. All involved pairs of demographic features; none of the demographic features were highly correlated to venue categories.

Several of the demographic features in this list were considered redundant features due to high correlation with other demographic features and were marked for removal from the analysis data set. These included the following - marriage age population, number of household, number of families, housing density, proportion of couples without children, older senior proportion of population, unmarried proportion population, and proportion of families with single parents.

	Feature 1	Feature 2	Pearson Correlation	F-Test Score	Pearson Correlation p-value	Feature 1 Row	Feature 2 Column
0	11, Marriage Age Population	2, Pop	0.994533	2.689355e-137		9	1
1	14, Households	2, Pop	0.906656	1.406035e-53		12	1
2	21, Families	2, Pop	0.979924	1.558544e-98		19	1
3	Housing Density	4, Pop Density	0.982339	2.445181e-102		26	3
4	24, Couples without Children	5, Children	-0.845612	1.909128e-39		22	4
5	25, Couples with Children	5, Children	0.846800	1.169445e-39		23	4
6	10, Older Seniors	9, Seniors	0.856877	1.538650e-41		8	8
7	14, Households	11, Marriage Age Population	0.938082	2.102460e-65		12	10
8	21, Families	11, Marriage Age Population	0.961671	2.005693e-79		19	10
9	13, Not Married	12, Married	-0.999968	2.201593e-291		11	11
10	21, Families	14, Households	0.822797	1.121671e-35		19	13
16	25, Couples with Children	15, 1 Person Household	-0.857722	1.054322e-41		23	14
15	24, Couples without Children	15, 1 Person Household	0.858121	8.809428e-42		22	14
14	20, Avg Household Size	15, 1 Person Household	-0.962486	4.682443e-80		18	14
12	18, 4 Person Household	15, 1 Person Household	-0.929787	9.228578e-62		16	14
11	17, 3 Person Household	15, 1 Person Household	-0.924707	9.592605e-60		15	14
13	19, 5+ Person Household	15, 1 Person Household	-0.833483	2.268352e-37		17	14
17	18, 4 Person Household	17, 3 Person Household	0.830155	7.869031e-37		16	16
18	20, Avg Household Size	17, 3 Person Household	0.862432	1.223783e-42		18	16
19	24, Couples without Children	17, 3 Person Household	-0.835494	1.055435e-37		22	16
20	25, Couples with Children	17, 3 Person Household	0.834052	1.828580e-37		23	16
21	20, Avg Household Size	18, 4 Person Household	0.886948	3.855555e-48		18	17
22	24, Couples without Children	18, 4 Person Household	-0.895952	1.727442e-50		22	17
23	25, Couples with Children	18, 4 Person Household	0.895624	2.121898e-50		23	17
24	20, Avg Household Size	19, 5+ Person Household	0.946973	6.462911e-70		18	18
25	24, Couples without Children	20, Avg Household Size	-0.872863	7.709671e-45		22	19
26	25, Couples with Children	20, Avg Household Size	0.872919	7.492165e-45		23	19
27	23, Single Families	22, Couple Families	-0.999770	3.803804e-232		21	21
28	25, Couples with Children	24, Couples without Children	-0.999823	5.113541e-240		23	23

Figure 16. Dataframe *nh_stats_venues_corr_features_sorted*, Demographic Feature Correlations Shown

There was no strong correlation between any neighborhood features and the presence of a gastropub near a neighborhood or the number of gastropubs near a neighborhood.

```
The maximum Pearson correlation coefficients for "Number of Nearby GP" is Number of Nearby GP    0.591175
dtype: float64
The minimum Pearson correlation coefficients for "Number of Nearby GP" is Number of Nearby GP   -0.578057
dtype: float64
The maximum Pearson correlation coefficients for "GP Nearby" is GP Nearby    0.564507
dtype: float64
The minimum Pearson correlation coefficients for "GP Nearby" is GP Nearby   -0.401695
dtype: float64
```

Figure 17. Maximum and Minimum Correlation Coefficients for Gastropub Presence Features

3.7 K-Means Clustering Analysis

A dataframe called *nh_stats_venues_norm_* was created for the clustering analysis. It contains most of the data in the primary neighborhood dataframe *nh_stats_venues*, with the exception of the statistically redundant demographic features listed above. The land area feature was excluded because it was only used for calculating housing density, which had already been dropped due to redundancy. The neighborhood identification columns were also excluded because neighborhood names and area numbers are not features that can be normalized or clustered.

Most importantly, the gastropub venue category feature was excluded so that the current existence of gastropubs near a neighborhood would not be a clustering factor. The intent of the cluster analysis was to group neighborhoods by general features, and find clusters with features that are more suitable for gastropubs. Including the gastropub venue category feature could cause overfitting - the clusters would be more influenced by the current existence of gastropubs and exclude neighborhoods without gastropubs but still have similar hospitable environments.

Dataframe *nh_stats_venues_norm_* was normalized using the **StandardScaler, fit, and transform** functions.

The sklearn **KMeans** function was used to perform the k-means clustering algorithm. Cluster sizes from 2 to 10 were evaluated, with 10 runs per cluster size.

The **KMeans** function generated k-means cluster labels for each cluster size. The function also output an **inertia_** value for each cluster size. The **inertia_** value is a calculation of the relative similarity of the neighborhoods within each cluster to the relative dissimilarity of the neighborhoods between clusters. This and the adjusted Rand metrics score were used to evaluate the quality of each cluster size. The adjusted Rand metrics score compared the cluster labels against a reference to evaluate the effectiveness of the clusters in separating the neighborhoods by the reference labels. The dummy variables in the *GP Nearby* column of dataframe *gp_nh_present* were used as the reference labels.

4. RESULTS AND DISCUSSION

When the gastropub locations were initially plotted on a Folium map, it was apparent that gastropubs were not ubiquitous through Toronto. Most were located in a few areas of Toronto, while most areas had no gastropubs. Therefore, there is likely some variables affecting the suitability of a neighborhood to host a gastropub.

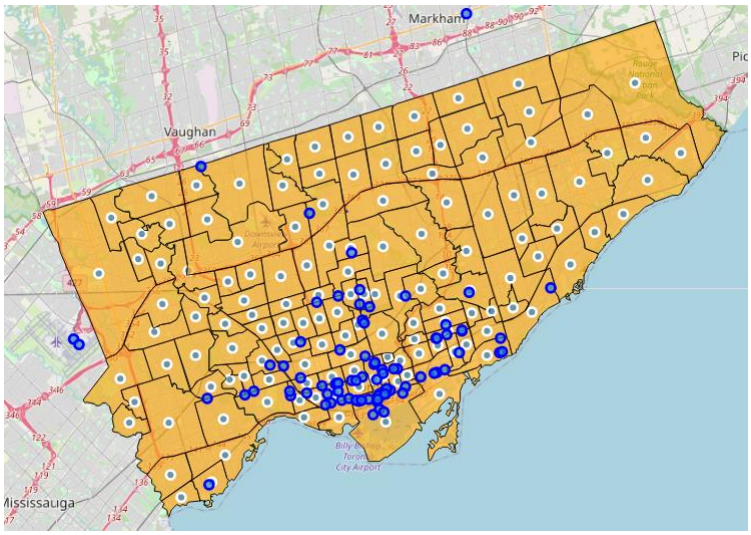

Figure 18. Folium Map of Gastropubs, Overlaid on Toronto Neighborhood Data

After the k-means clusters were generated, the metrics were reviewed.

The inertia values should generally decrease with increasing cluster size, as the neighborhoods within each cluster become more similar and the neighborhoods across clusters become increasingly dissimilar. The inertia values for the clusters in this analysis followed this trend.

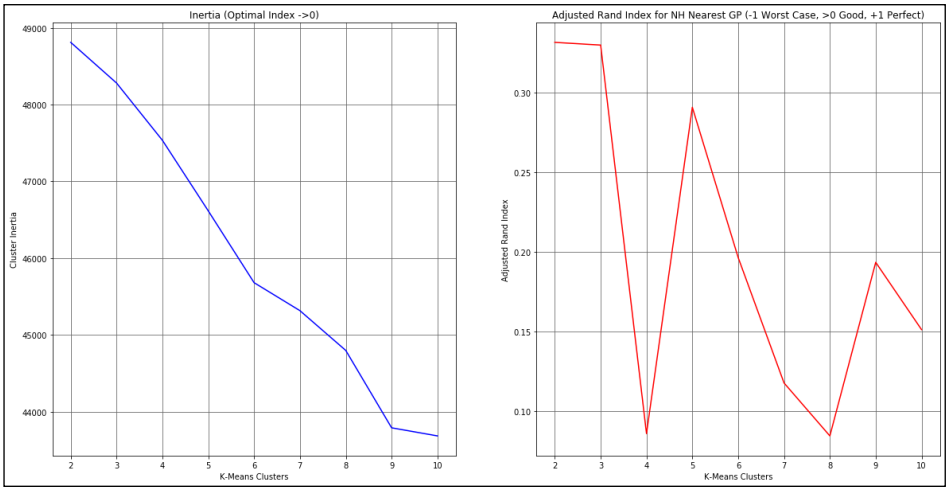

Figure 19. K-Means Cluster Evaluation Metrics

The optimal cluster size to consider would be the cluster size where the curve started leveling out and forming an elbow shape. This elbow would indicate that further increasing cluster size would provide diminishing returns in cluster quality improvement. There was no distinct elbow in this curve, so this metric did not provide a clear optimal cluster size to select.

The adjusted Rand index value for a quality cluster is +1, and decreases as quality drops. The highest index value is for a cluster size of 2, though the index value for a cluster size of 3 is only slightly lower. Both of these would be evaluated.

The cluster neighborhoods were visualized in Folium maps. The gastropub locations were included to visualize where they are located relative to the cluster neighborhoods. General statistics on each cluster were also generated.

Label	# of Neighborhoods in Cluster	# of Cluster Neighborhoods Near A Gastropub	# of Gastropubs Near A Neighborhood in the Cluster	Proportion of All Toronto Neighborhoods in Cluster	Proportion of All Toronto Gastropubs in Cluster	Proportion of Neighborhoods in Cluster Near A Gastropub
0	103	16	18	0.735714	0.219512	0.155340
1	37	25	64	0.264286	0.780488	0.675676

Figure 20. K-Means Cluster Size 2 General Statistics

A cluster size of 2 was effective in partitioning neighborhoods with and without nearby gastropubs. Cluster 1 contained only 26% of all Toronto neighborhoods, but 78% of all Toronto gastropubs were near these neighborhoods. Also, 68% of the cluster 1 neighborhoods were the nearest ones to these gastropubs.

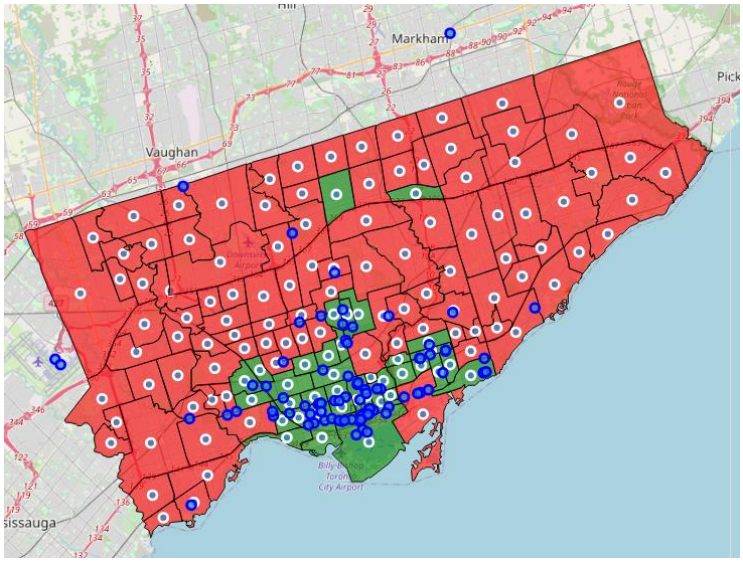

Figure 21. K-Means Cluster Size 2 Folium Map
Cluster 0 Neighborhoods – Red
Cluster 1 Neighborhoods – Green

Label	# of Neighborhoods in Cluster	# of Cluster Neighborhoods Near A Gastropub	# of Gastropubs Near A Neighborhood in the Cluster	Proportion of All Toronto Neighborhoods in Cluster	Proportion of All Toronto Gastropubs in Cluster	Proportion of Neighborhoods in Cluster Near A Gastropub
0	1	1	1	0.007143	0.012195	1.000000
1	103	16	18	0.735714	0.219512	0.155340
2	36	24	63	0.257143	0.768293	0.666667

Figure 22. K-Means Cluster Size 3 General Statistics

A cluster size of 3 was not significantly different from a cluster size of 2. A few neighborhoods in the 2 cluster set switched clusters, and the new cluster was comprised of only one neighborhood. The neighborhood and gastropub statistics for the two larger clusters are nearly identical to those in the 2 cluster set.

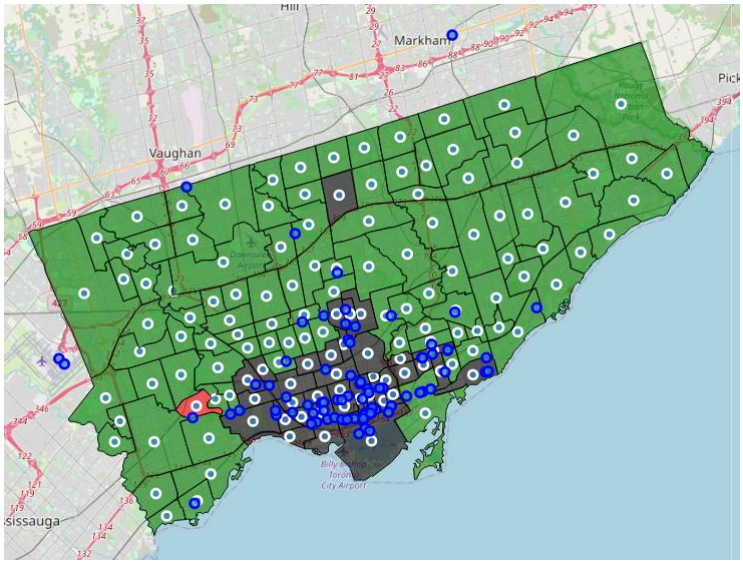

Figure 23. K-Means Cluster Size 3 Folium Map
Cluster 0 Neighborhoods – Red
Cluster 1 Neighborhoods – Green
Cluster 2 Neighborhoods - Black

A cluster size of 2 was selected as the model to continue evaluating. It had the slightly higher adjusted Rand index score, the neighborhood clusters did not change significantly between cluster sizes 2 and 3, and it was easier to compare two clusters than three.

The neighborhoods were grouped by label and the means of their feature data were calculated. The mean values for the demographic features are shown to the right.

There were noticeable differences in average values between the clusters. To determine if any features were statistically different between the clusters, an ANOVA study was performed on the neighborhood feature data, grouped by cluster label.

Two statistics were generated by the ANOVA study for each feature – an F-test score and a p-value. The clusters would be significantly different with regards to a feature if the F-test score for the feature exceeded an F-test score critical value. The F-test score critical value was determined by the number of clusters, total number of samples, and required confidence level. These were input into the **scipy.stats.f.ppf** function, which then output the critical value.

Label	Neighborhood Feature	0	1
none			
3	2, Pop	19570.873786	19345.162162
4	3, Pop Change	2.415534	8.564865
5	4, Pop Density	4614.349515	10845.432432
6	5, Children	0.158582	0.120822
7	6, Youth	0.125799	0.113797
8	7, Working Age	0.415342	0.521083
9	8, Pre-Retire	0.131238	0.110432
10	9, Seniors	0.168967	0.133960
11	10, Older Seniors	0.027145	0.018952
12	11, Marriage Age Population	16462.233010	17232.972973
13	12, Married	0.524551	0.472507
14	13, Not Married	0.475488	0.527456
15	14, Households	7339.271845	9648.783784
16	15, 1 Person Household	0.262774	0.408135
17	16, 2 Person Household	0.289360	0.317266
18	17, 3 Person Household	0.177857	0.131184
19	18, 4 Person Household	0.159469	0.097253
20	19, 5+ Person Household	0.110489	0.046156
21	20, Avg Household Size	2.639029	2.081351
22	21, Families	5324.805825	4603.243243
23	22, Couple Families	0.776083	0.824374
24	23, Single Families	0.223889	0.175939
25	24, Couples without Children	0.400107	0.550826
26	25, Couples with Children	0.600196	0.449480
27	26, Average Income	356788.737864	335930.216216
28	27, Land Area	5.393204	2.020000
29	Housing Density	1772.061737	5383.315628
30	Venue Count	0.246117	0.849730

Figure 24. K-Means Cluster Size 2 Group Demographic Feature Mean Values

	Neighborhood Feature	ANOVA F-test Score	ANOVA p-Value
0	Venue Count	394.602922	2.661151e-42
1	7, Working Age	126.884293	2.808697e-21
2	24, Couples without Children	97.239099	1.087054e-17
3	25, Couples with Children	97.130556	1.122584e-17
4	15, 1 Person Household	93.330417	3.494966e-17
5	20, Avg Household Size	88.230416	1.654219e-16
6	18, 4 Person Household	79.929569	2.247282e-15
7	Housing Density	75.686728	8.873876e-15
8	17, 3 Person Household	71.750646	3.254650e-14
9	4, Pop Density	66.305165	2.049869e-13
10	19, 5+ Person Household	62.776118	6.944225e-13
11	Yoga Studio	53.369496	2.010284e-11
12	Vegetarian / Vegan Restaurant	49.301675	9.100048e-11
13	Beer Bar	45.297620	4.164849e-10
14	8, Pre-Retire	43.322235	8.937917e-10
15	Gastropub	42.380497	1.290394e-09
16	Café	40.053273	3.226803e-09
17	Art Gallery	36.673644	1.250639e-08
18	5, Children	35.476355	2.035195e-08
19	16, 2 Person Household	27.102901	6.848318e-07

Figure 25. ANOVA Results for Cluster Size 2 Groups, Filtered and Sorted

The ANOVA values were filtered to extract only the features with F-test scores that exceeded the critical value and p-values below 0.05. These features were then sorted by F-test score.

Among the 375 features included in the clustering analysis, 127 features were statistically different between the clusters. The top F-test score features are shown in the dataframe to the right.

The feature with the greatest statistical difference between clusters was the average number of venues in each neighborhood. Several demographic features followed behind. The most significant venue category feature iwas “yoga studio” at the 11th position.

Gastropubs were also on this list, at the 15th position. This venue category feature was intentionally left out of the analysis feature data set to limit its effect on the clustering algorithm, but it appeared the other features in cluster 1 neighborhoods made this cluster distinctly more favorable to hosting gastropubs than cluster 0 neighborhoods.

Since there were so many statistically significant differences between cluster 0 and 1 neighborhood features, it was natural to consider statistical differences between cluster 1 neighborhoods with and without nearby gastropubs. Therefore, the ANOVA study and criteria filtering were repeated on cluster 1 neighborhoods, divided into two groups by gastropub presence.

The results showed far fewer statistically different features between the cluster 1 subgroups. Only ten features were statistically different between them (see Figure 26). One is gastropubs, which is the primary venue of interest. Another two, sports bars and lounges, could be considered venues related to gastropubs. There is no apparent connection between gastropubs and the other venue category features, though.

	Neighborhood Attribute	ANOVA F-test Score	ANOVA p-Value
162	Gastropub	14.209179	0.000605
227	Lounge	8.036879	0.007564
155	Frozen Yogurt Shop	7.038574	0.011913
199	Ice Cream Shop	6.542813	0.015018
326	Sports Bar	5.332434	0.026954
205	Intersection	4.826673	0.034735
95	Churrascaria	4.729586	0.036493
370	Wings Joint	4.691349	0.037212
237	Miscellaneous Shop	4.668005	0.037658
229	Market	4.338355	0.044635

Figure 26. Statistically Different Features for Cluster 1 Subgroups

A new Folium map was generated to differentiate the cluster 1 neighborhoods by current nearby gastropubs. Cluster 1 neighborhoods with nearby gastropubs are colored red, cluster 1 neighborhoods without nearby gastropubs are colored green. Cluster 0 neighborhoods are colored black.

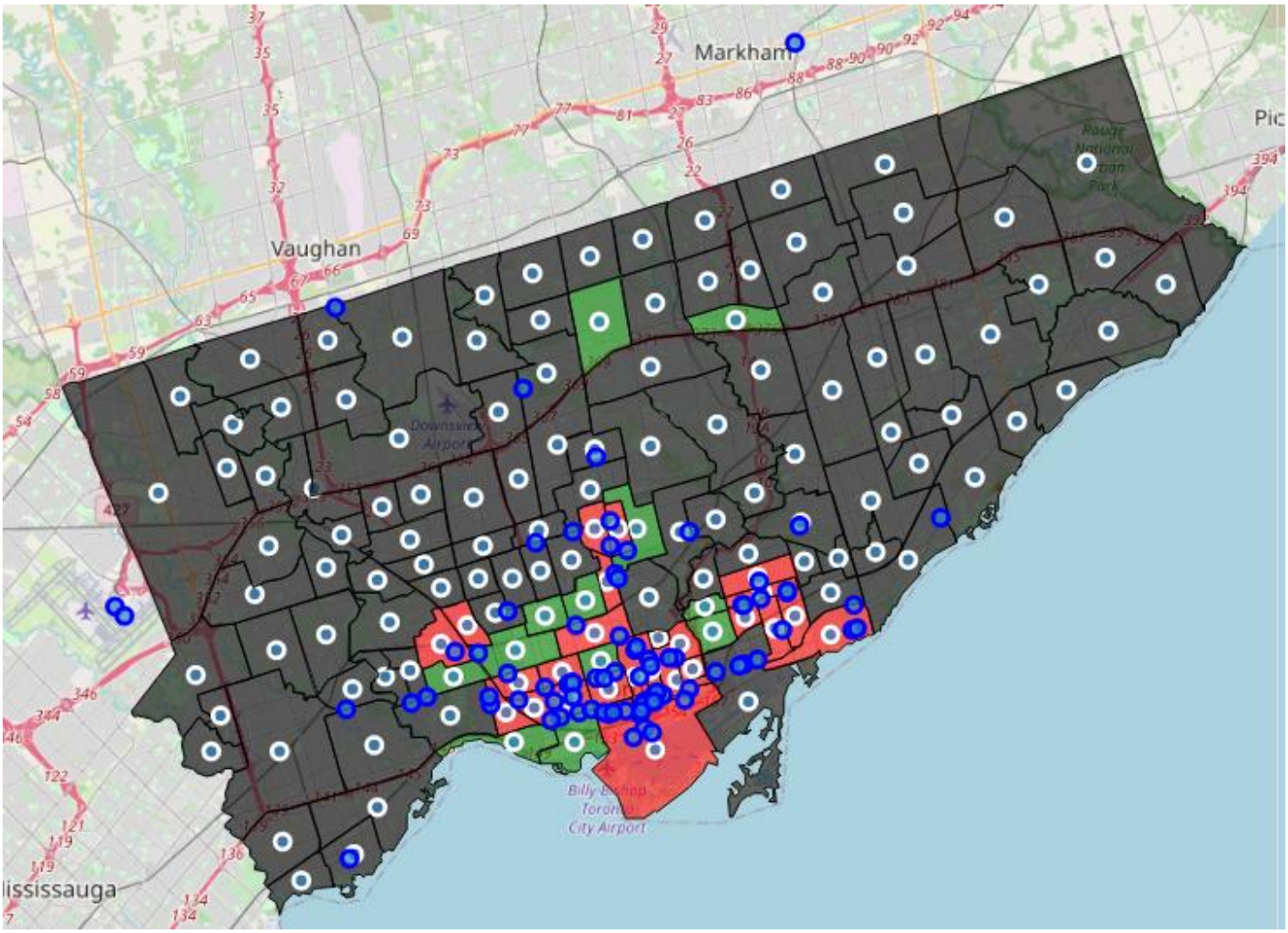

Figure 27. Cluster 1 Neighborhoods with and without Nearby Gastropubs

Around the Downtown area, the cluster 1 neighborhoods with and without nearby gastropubs are mixed together, with no clear geographical pattern. The neighborhoods to the north, well outside of Downtown, currently do not have nearby gastropubs.

The cluster 1 neighborhood subgroups are listed to the right. Since cluster 1 neighborhoods in general were popular for gastropubs, any from these cluster would be worth considering as locations for a new gastropub. Neighborhoods with nearby gastropubs would be a safe choice since the current existing gastropubs prove they are suitable locations, while neighborhoods currently without gastropubs may be emerging opportunities.

There are 12 cluster 1 neighborhoods without gastropubs nearby.

	Neighborhood
21	Casa Loma (96)
32	Dovercourt-Wallace Emerson-Junction (93)
48	Henry Farm (53)
49	High Park North (88)
82	Mount Pleasant East (99)
87	Niagara (82)
88	North Riverdale (68)
97	Playter Estates-Danforth (67)
109	South Parkdale (85)
120	University (79)
129	Willowdale East (51)
135	Wychwood (94)

Figure 28. Cluster 1 Neighborhoods without Nearby Gastropubs

There are 25 cluster 1 neighborhoods with gastropubs nearby.

	Neighborhood
3	Annex (95)
6	Bay Street Corridor (76)
14	Blake-Jones (69)
19	Cabbagetown-South St.James Town (71)
23	Church-Yonge Corridor (75)
28	Danforth (66)
29	Danforth East York (59)
34	Dufferin Grove (83)
46	Greenwood-Coxwell (65)
59	Junction Area (90)
62	Kensington-Chinatown (78)
71	Little Portugal (84)
79	Moss Park (73)
83	Mount Pleasant West (104)
89	North St.James Town (74)
94	Palmerston-Little Italy (80)
100	Regent Park (72)
103	Roncesvalles (86)
116	The Beaches (63)
119	Trinity-Bellwoods (81)
122	Waterfront Communities-The Island (77)
127	Weston-Pellam Park (91)
133	Woodbine Corridor (64)
136	Yonge-Eglinton (100)
137	Yonge-St.Clair (97)

Figure 29. Cluster 1 Neighborhoods with Nearby Gastropubs

	Neighborhood Feature	0	1
0	Venue Count	0.246117	0.849730
1	7, Working Age	0.415342	0.521083
2	24, Couples without Children	0.400107	0.550826
3	25, Couples with Children	0.600196	0.449480
4	15, 1 Person Household	0.262774	0.408135
5	20, Avg Household Size	2.639029	2.081351
6	18, 4 Person Household	0.159469	0.097253
7	Housing Density	1772.061737	5383.315628
8	17, 3 Person Household	0.177857	0.131184
9	4, Pop Density	4614.349515	10845.432432
10	19, 5+ Person Household	0.110489	0.046156
11	Yoga Studio	0.000384	0.007089
12	Vegetarian / Vegan Restaurant	0.000515	0.010145
13	Beer Bar	0.000000	0.006101
14	8, Pre-Retire	0.131238	0.110432
15	Gastropub	0.002125	0.018204
16	Café	0.019275	0.055312
17	Art Gallery	0.000000	0.005382
18	5, Children	0.158582	0.120822
19	16, 2 Person Household	0.289360	0.317266

Figure 30. Feature Means with the Highest Statistically Significant Difference between Cluster 0 and 1 Neighborhoods

Sorting the neighborhood feature means by the highest F-test score between clusters 0 and 1 provided insight into the top differentiating features for neighborhoods that are favorable to gastropubs. The following are the top general trends in cluster 1 neighborhoods based on the top 20 features:

- Higher venue density
- Higher proportion of population is working age, and smaller proportion are children or retired
- Higher proportion of couples without children
- Smaller average household size – greater proportion of households with 1 or 2 residents and fewer with 3 or more
- Higher population density and housing density
- More leisure and entertainment-type venues

5. CONCLUSION

The study showed that a data-driven approach has great potential for finding suitable new locations for a gastropub, or any other type of venue for that matter. K-means clustering was an effective method for identifying a cluster of Toronto neighborhoods near gastropubs, just based on neighborhood demographic and local venue information. At a cluster size of just 2, the algorithm was able to create a cluster (cluster 1) with a small percentage of the Toronto neighborhoods but near a large percentage of the gastropubs. Not every neighborhood in this cluster had a nearby gastropub, but the demographic and venue profile features of all the neighborhoods in this cluster are distinctly different from neighborhoods outside, and these differences may cause the former to be more favorable to supporting new gastropubs. Therefore the neighborhoods identified in cluster 1 are recommended as the top areas to initially investigate.

Among the entire group of Toronto neighborhoods, no individual feature significantly correlated to a neighborhood having a gastropub nearby. However, there were numerous statistically significant differences between cluster 0 and 1, including the presence of gastropubs. Demographic features accounted for the most significant differences between the clusters, but a few venue features also significantly differed. The top traits common to cluster 1 neighborhoods included higher venue, population, and housing densities, more working age residents and fewer children and retired residents, smaller household sizes and fewer couples with children, and more leisure and entertainment venues.

Between the cluster 1 neighborhoods with and without gastropubs, there were few statistically significant differences, and most of them were for venues that do not appear to be directly related to gastropubs.

A new gastropub could be located in one of the cluster 1 neighborhoods that already has a nearby gastropub, since they are proven to be viable locations, or another cluster 1 neighborhood without nearby gastropubs to potentially take advantage of less competition.

There are many other details to consider before ultimately choosing a location for a new gastropub. While all of the neighborhoods in the cluster have similar demographic and venue profiles, there may be other neighborhood aspects that may explain why some currently do not have gastropubs, such as zoning and other local restrictions, site availability and costs, and transportation considerations. None of these were factors in the study.

References

[1] https://open.toronto.ca/dataset/neighbourhood-profiles/
[2] https://open.toronto.ca/dataset/neighbourhoods/
[3] https://developer.foursquare.com/
[4] https://developer.foursquare.com/docs/build-with-foursquare/categories/