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

Applied Data Science Capstone Project

2020-05-03

Leslie Lui

Introduction

- Restaurant industry is exciting but competitive to be in
- New restaurants require innovation and attractiveness to thrive
- Innovate by following new restaurant trends
- Attract customers by selecting the right location

Introduction

Innovative idea? – Gastropub!

- Increasing international trend
- Offer unique dining and entertainment experiences to appeal to customers
- Premium atmosphere increases profitability

Location Selection

- Good locations attract customer attention or provide convenient access
- Bad locations can be overlooked or inaccessible, limiting business

Introduction

How to select a suitable location? – Data Science!

- Evaluates a problem objectively
- Finds insights hidden to casual observation
- Uses data to back up recommendations
- Easily analyze large amounts of data

Gastropub Location Selection Methodology

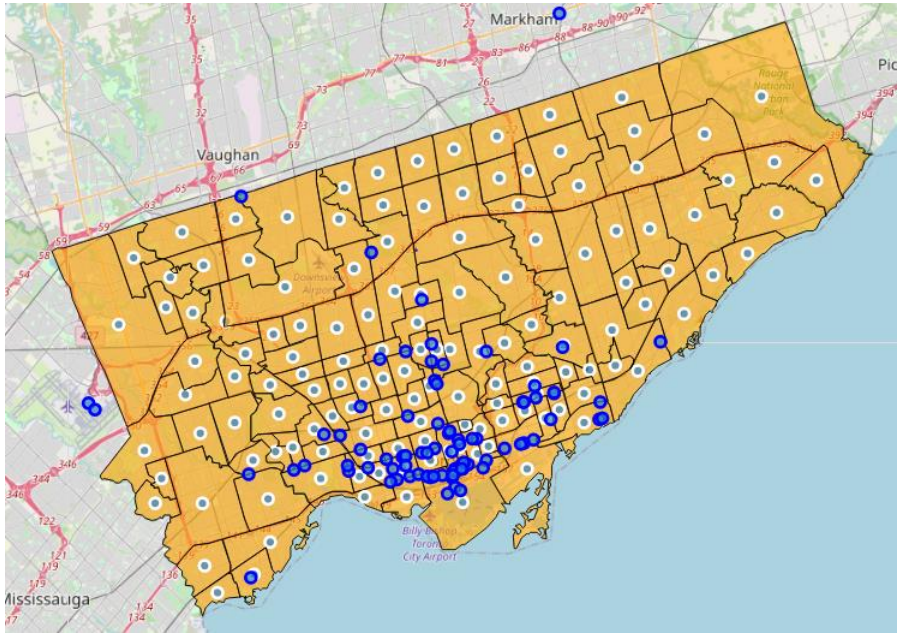
- Gather geographic and demographic information on Toronto neighborhoods
- Find existing gastropubs in Toronto, and determine their neighborhoods
- Gather venue profiles on every Toronto neighborhood
- Group the Toronto neighborhoods into clusters, based on similarities in their demographic, geographic, and venue features
- Find the clusters that effectively group together neighborhoods with existing gastropubs
- Evaluate the characteristics of the clusters to identify neighborhood traits that favor gastropubs
- Identify cluster neighborhoods as recommended locations for new gastropubs

Analysis Data

- Toronto neighborhood geographic data, from Toronto Open Data portal
  - Central and boundary coordinates for each neighborhood
- Toronto neighborhood demographic data, from Toronto Open Data portal
  - Population data for each neighborhood, including age, household size and composition, income, ethnicity, etc.
- Toronto gastropub data, from Foursquare API
  - Foursquare API service search results for gastropubs throughout Toronto
  - Key information include gastropub names and ID's and coordinates
- Toronto general venue data, from Foursquare API
  - Foursquare API service search results for popular venues in each neighborhood
  - Key information includes venue names and ID's and their venue category

Preliminary Data Review

- 140 neighborhoods in Toronto
- 5679 venues found around all Toronto neighborhoods
- 82 gastropubs located in Toronto, located near 41 neighborhoods



Gastropubs – Blue Circles
Neighborhood Centers – White Circles

Analysis Data Preparation

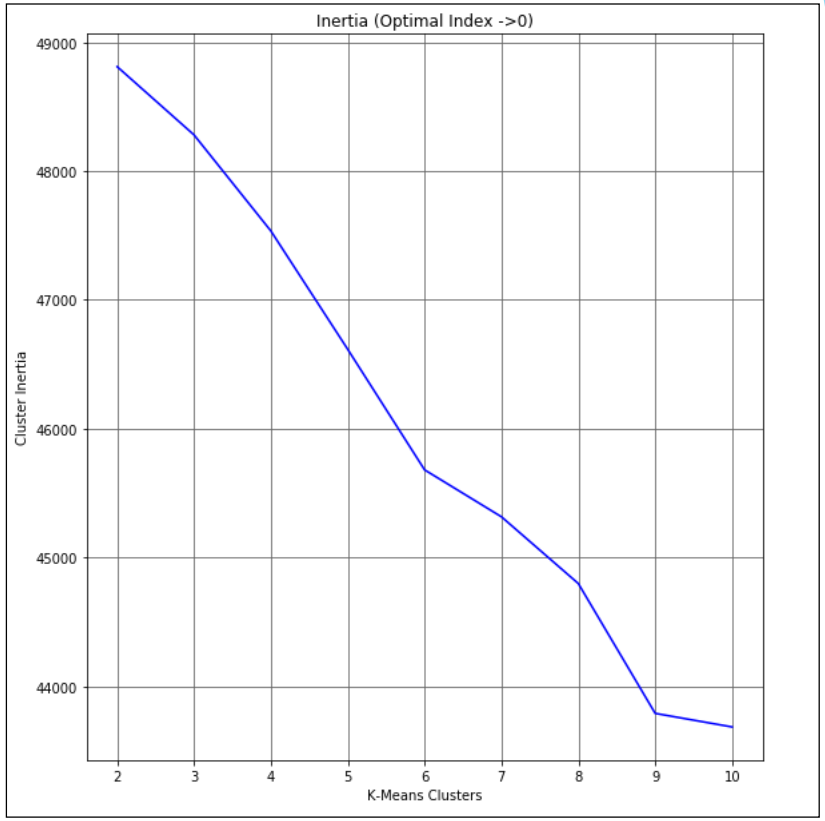
- Data for neighborhood venue and demographic information combined into a single dataset for cluster analysis
- Venue details included venue category and their proportion in each neighborhood; details involved 5679 venues returned in Foursquare search
- Neighborhood data on existing gastropubs excluded from analysis data set to prevent influencing cluster analysis
- Demographic information included population, age brackets, household counts, household sizes, family structure, and income
- Some highly correlated features in the neighborhood data set also removed to simplify analysis
- Other data included in the data files and search results were removed, including identification and coordinate data
- Data set normalized prior to cluster analysis

Clustering Algorithm

- K-Means clustering algorithm performed on neighborhood data
- No target cluster size specified, so K-Means clustering algorithm performed over range of cluster sizes 2-10
- For each cluster size, K-Means clustering algorithm assigns each neighborhood to a suitable cluster
- Metrics calculated for each cluster size to evaluate quality and select cluster sizes to evaluate
  - Inertia metric calculated similarity of neighborhoods in each cluster against dissimilarity to neighborhoods in other clusters
  - Adjusted Rand index calculated effectiveness of clusters to group neighborhoods by pre-assigned labels reference neighborhoods

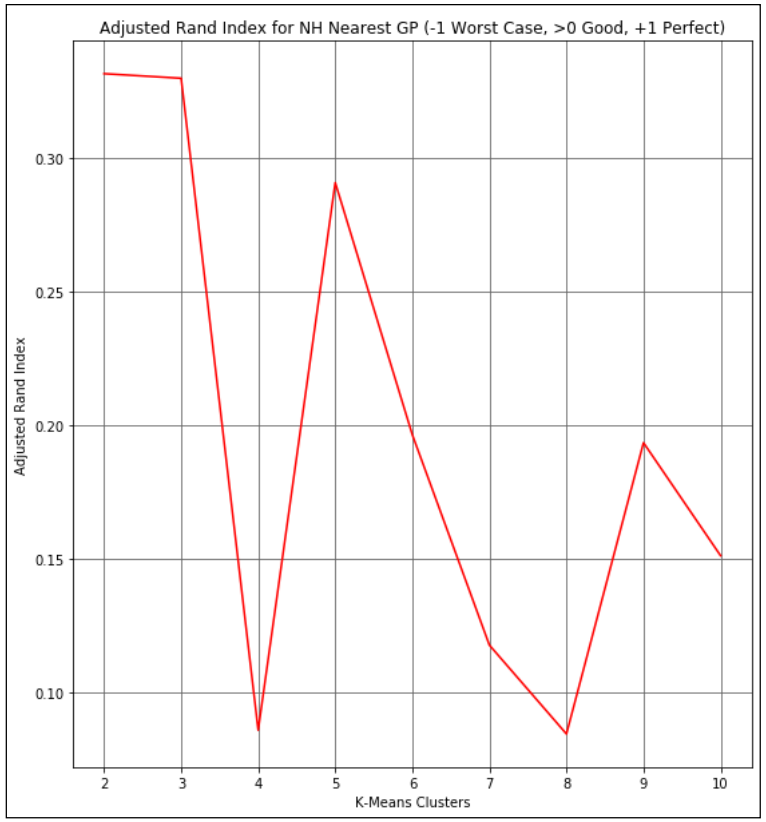
Cluster Size Selection

- Inertia metric
  - Lower inertia indicates neighborhoods within clusters are becoming more similar
  - Inertia generally decreases as cluster size increases, but usefulness of clusters in identifying general trends also decreases as cluster size increases
  - Optimal cluster size is typically where elbow develops in the cluster size-inertia curve, indicating diminishing rate of decrease in inertia as cluster size continues increasing
  - No clear elbow on this curve, so Adjusted Rand index considered instead



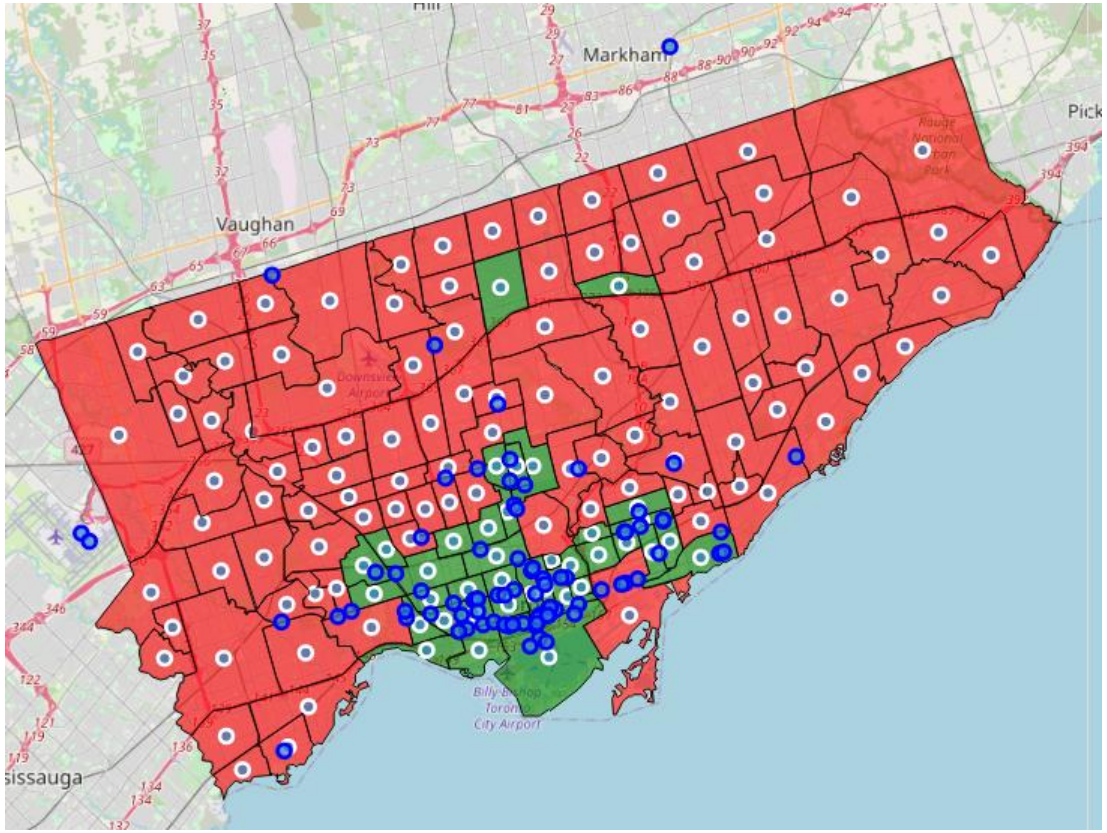
Cluster Size Selection

- Adjusted Rand index
  - Higher values indicate clusters effectively group neighborhoods together by pre-assigned labels
  - For this study, the pre-assigned labels used were based on whether each neighborhood had an existing gastropub nearby; the gastropub proximity status for each neighborhood was determined in earlier data preparation
  - Highest value exists at cluster size 2, so this cluster size was evaluated



Cluster Evaluation

Map of Toronto neighborhoods labeled by cluster



Cluster 0 Neighborhoods – Red
Cluster 1 Neighborhoods – Green
Gastropubs – Blue Circles
Neighborhood Centers – White Circles

Cluster Evaluation

- Cluster statistics

Cluster	# 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	0.74	0.22	0.16
1	37	25	64	0.26	0.78	0.66

- Cluster size 2 effective at grouping neighborhoods with nearby gastropubs
- Cluster 1 contained 26% of Toronto neighborhoods, but 68% had a nearby gastropub and 78% of all Toronto gastropubs were near these neighborhoods

Cluster Evaluation

- Comparison of average neighborhood data for each cluster showed distinct differences among some features

Cluster	Population	Population Change	Population Density	Children	Youth	Working Age	Pre-Retirement	Seniors	Older Seniors
0	19571	2.42	4614.4	0.159	0.126	0.415	0.131	0.169	0.027
1	19345	8.56	10845.4	0.121	0.114	0.521	0.110	0.134	0.019

Cluster	Marriage Age Population	Married	Not Married	Households	1 Person Household	2 Person Household	3 Person Household	4 Person Household	5+ Person Household
0	16462	0.525	0.475	7339	0.263	0.289	0.177	0.159	0.110
1	17233	0.473	0.527	9648	0.408	0.317	0.131	0.097	0.046

Cluster	Avg. Household Size	Families	Couple Families	Single Families	Couples without Children	Couples with Children	Average Income	Housing Density	Venue Count
0	2.639	5324	0.776	0.224	0.400	0.600	356789	1772	0.246
1	2.081	4603	0.824	0.176	0.551	0.449	335930	5383	0.850

Cluster Evaluation

- ANOVA performed to determine if any differences between clusters are statistically significant
- Statistically significant differences in 127 of 375 features, including both demographic and venues

Top features with the highest significant difference between clusters

Cluster	Venue Count	Working Age	Couples without Children	Couples with Children	1 Person Household	Avg. Household Size	4 Person Household	Housing Density	3 Person Household
0	0.246	0.415	0.400	0.600	0.263	2.639	0.159	1772	0.177
1	0.850	0.521	0.551	0.449	0.408	2.081	0.097	5383	0.131

Cluster	Population Density	5+ Person Household	Yoga Studio	Vegetarian/ Vegan Restaurant	Beer Bar	Pre-Retirement	Gastropub	Café	Art Gallery
0	4614.4	0.110	0.0004	0.0005	0	0.131	0.0021	0.0193	0
1	10845.4	0.046	0.0071	0.0101	0.0061	0.110	0.0180	0.0553	0.0054

Cluster	Children	2 Person Household	Cocktail Bar	Married	Not Married	Ramen Restaurant	New American Restaurant	Music Venue	Seniors
0	0.159	0.289	0	0.525	0.475	0.0004	0	0	0.169
1	0.121	0.317	0.0060	0.473	0.527	0.0066	0.0020	0.0025	0.134

Cluster Evaluation

General features in cluster 1 neighborhoods that differentiate it from Cluster 0

- Higher venue density
- Higher proportion of venues are leisure or entertainment-type
- Higher population and housing densities
- Smaller household sizes; higher proportions of 1 and 2 resident households
- Higher proportion of working age residents
- Lower proportions of children and retirement age residents

Cluster Evaluation

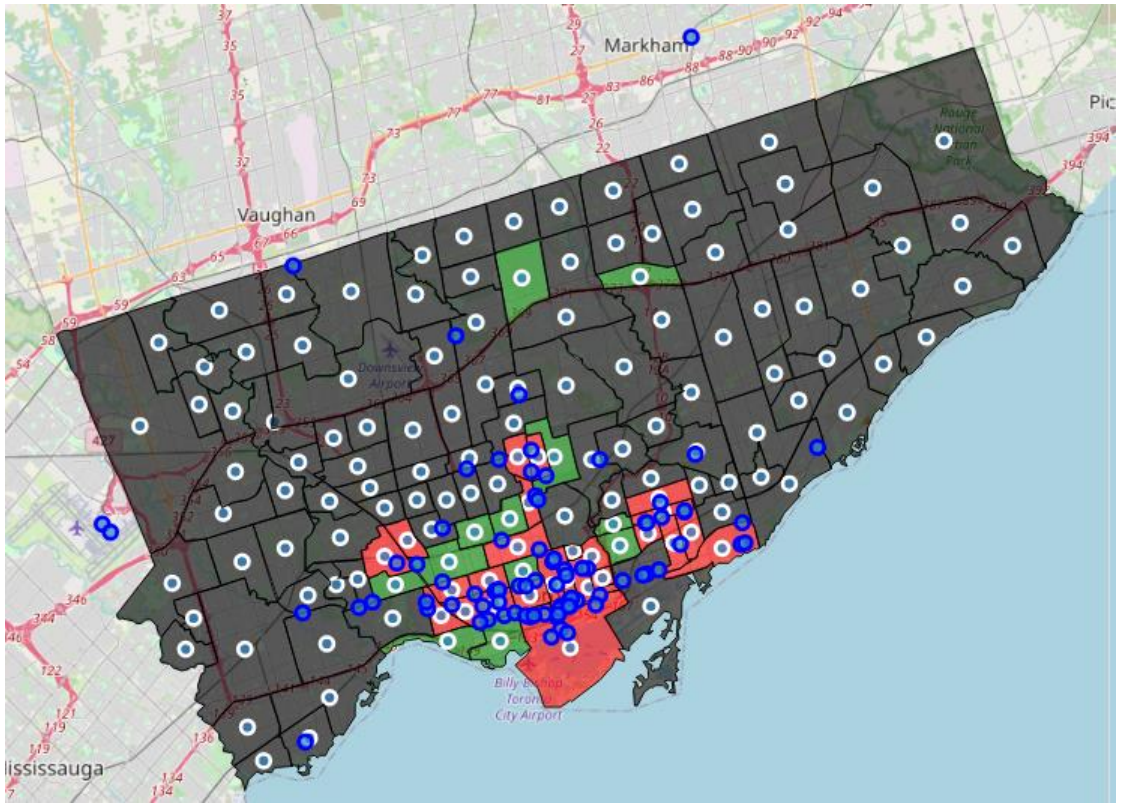
- Few statistically significant differences between cluster 1 neighborhoods with and without nearby gastropubs
- All statistically different features are venues
- Gastropubs are obvious difference
- Lounges and sports bars are slightly related types of venues to gastropubs
- Other venues have no clear relationship to gastropubs

Cluster 1 Subgroup Significant Differences
Gastropub
Lounge
Frozen Yogurt Shop
Ice Cream Shop
Sports Bar
Intersection
Churrascaria
Wings Joint
Miscellaneous Shop
Market

- The limited significant differences indicate all cluster 1 neighborhoods should be promising locations for gastropubs

Cluster Evaluation

Map of Cluster 1 Neighborhoods, Differentiated by Gastropub Presence



Cluster 1 Neighborhoods with Existing Gastropubs Nearby - Red
Cluster 1 Neighborhoods without Gastropubs Nearby - Green
Cluster 0 Neighborhoods - Black

Cluster Evaluation

Cluster 1 Neighborhoods with Gastropub (25 Total)	
Annex (95)	Mount Pleasant West (104)
Bay Street Corridor (76)	North St.James Town (74)
Blake-Jones (69)	Palmerston-Little Italy (80)
Cabbagetown-South St.James Town (71)	Regent Park (72)
Church-Yonge Corridor (75)	Roncesvalles (86)
Danforth (66)	The Beaches (63)
Danforth East York (59)	Trinity-Bellwoods (81)
Dufferin Grove (83)	Waterfront Communities-The Island (77)
Greenwood-Coxwell (65)	Weston-Pellam Park (91)
Junction Area (90)	Woodbine Corridor (64)
Kensington-Chinatown (78)	Yonge-Eglinton (100)
Little Portugal (84)	Yonge-St.Clair (97)
Moss Park (73)	

Cluster 1 Neighborhoods without Gastropubs (12 Total)
Casa Loma (96)
Dovercourt-Wallace Emerson-Junction (93)
Henry Farm (53)
High Park North (88)
Mount Pleasant East (99)
Niagara (82)
North Riverdale (68)
Playter Estates-Danforth (67)
South Parkdale (85)
University (79)
Willowdale East (51)
Wychwood (94)

Conclusion

- Cluster 1 neighborhoods with and without gastropubs have limited distinct differences, so all should be favorable locations for a new gastropub
- Cluster 1 neighborhoods are distinctly different from cluster 0 neighborhoods in many features
- Analysis was not exhaustive and other local variables not included in analysis (zoning, traffic, business costs) may be differentiating factors; separate investigation on this needed before making final location decision
- Data science-driven approach has potential for objectively assessing neighborhoods to choose locations suited for a venue