Aussie Pies

the great Australian bite. . .

. . .coming to Toronto

Prepared by: Andrew Huxham

Date: 28 Jun 2020

Commission: IBM Data Science – Coursera

Contents

INTRODUCTION

The report audience

This report is drafted primarily for an anonymous Australian investor at the embryonic phase of investigating prospects of a new venture. The report may also be of interest to potential investors or business associates considering a commitment to support the endeavour.

The business challenge

This analysis creates an initial market profile of metropolitan food outlets as preparatory research for a proposal to introduce a new product line of Australian meat pies into Canada. This research will focus on the City of Toronto, selected as a potential starter location due to being the most populous Canadian city (1).

Rather than compete with exiting businesses by opening another storefront, the proposed concept is to introduce a product line with a difference, the Australian meat pie. Ideally this will offer food retailers an additional sales opportunity while providing an agile, scalable market entry point for the new venture.

The meat pie is often referred to as national dish in Australia (2; 3). Impulsive snackers and regular fans from the Land Down Under consider the simple, savoury meat pie a staple stomach filler or craving crusher. The typical pie is a hand sized pastry bowl and lid filled with a dollop of diced or minced meat and gravy. As well as the ever-popular plain pie, the convenient hand sized meal comes in a wide range of delicious combinations including peas and mash potato, chicken, curry, mushroom, cheese and bacon or onion.

The super hungry can scale up with another pie or supplement with an equally healthy side dish such as deep-fried potato chips (fries). The venture anticipates that the delicacy from the land Down Under will carve out a cuisine niche among cosmopolitan Canadians with comparable taste traits.

Despite patriotic confidence that Canadians will share a wide appeal for the Aussie version of meat pie, the difficulty breaking into a new market is not underestimated. This analysis will create an initial profile of venues around Toronto that may support the concept. The analysis is preliminary in nature and is intended form a small but important part of any subsequent business case.

Image: (4)

DATA

Data priority

This preliminary analysis is concerned with identifying the profile of venues around the primary metropolitan area of Toronto Canada. The profile aims to identify:

Venue group	Description	Relevance
Casual eateries	Offers refreshments, snacks and meals either as takeaway or sit-down consumption. Includes cafes, coffee shops, diners and sandwich bars.	Initially anticipated as venues most likely to potentially distribute pies as an addition to their offering.
Restaurants	Generally, a place where diners would sit down to consume a main meal.	Not a likely distribution channel but indicates the range of cuisine in the area.
Bakeries	A venue producing or specialising in flour-based products.	This group of venue is considered the most likely source of direct competition with a comparable product.
Specialty eateries	Casual eateries other than restaurants that focus on a specific style of cuisine, from deserts and ice-creams to tacos.	Not a likely distribution channel but indicates the range of cuisine in the area.
Bars and pubs	Venues serving alcohol and may offer a wide range of food types, from snacks to full meals.	The venues contribute to drawing potential customers to the area. Some may prospective distribution channels.
Transportation	Places that facilitate transportation.	Transport users on the move are prospective consumers of convenient and fast savoury food such as the pie.
Household supplies	Venues cover a wide range of goods and services for household consumption, spanning from supermarkets, grocery stores to butchers.	Indicates the range of people who may be drawn to the area, particularly who may seek to purchase food while in the area.
Lifestyle	Venues cover anything from parks, churches, entertainment, recreation and sports ground.	Indicates the range of people who may be drawn to the area, particularly who may seek to purchase food while in the area.
Other	Any other venue group identified.	Adds to the profile of venues attracting potential consumers to the area.

Data source - target neighborhoods

The target area is Toronto Canada, within the postal sector 'M'.

Toronto neighborhoods within scope of consideration will be sourced from Wikipedia (5).

The data will be structured into a hierarchy of information categories by Postal Code, Borough and Neighborhood. The relevance of the hierarchy segments is as follows:

- **Borough**: is a Canadian municipal subdivision of a major city, which is Toronto for the purpose of this analysis.
- **Neighborhoods**: smaller communities that form part of the borough.
- **Postal Code**: Mail delivery area associated with a borough or neighborhood.

Wikipedia is used for pragmatic, cost effective access but is not a primary or official record source. There are known inconsistences such as borough, neighborhood or both classified as 'Not Assigned'. Treatment of these records is covered in the following section headed 'Methodology'.

Data source - spatial data

Location data and venue categories with related geographical coordinates will be sourced from foursquare.com.

The information categories utilised are:

- **Venues:** Categorical records of venues domiciled within respective neighborhoods will be imported, providing indications of popularity within neighborhoods.
- **Spatial data:** Correlated latitude and longitude coordinates will be imported to provide visual analysis of the venues and neighborhoods within scope.

Data solutions

Target neighbourhood information imported from Wikipedia (5) will be expanded by joining with records from Geospatial (6) and Foursquare (7), ultimately joined on common values.

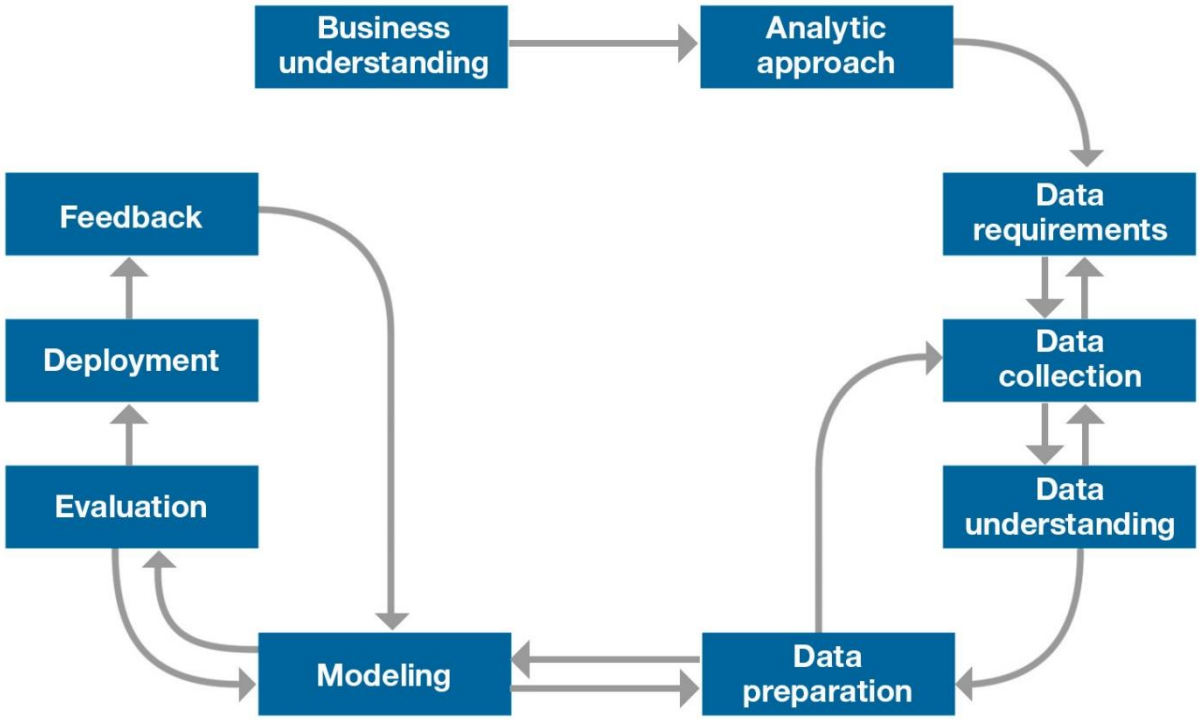
Application of the resulting dataset is intended to provide the following solutions:

- **Database:** Structured information set that defines the scope of information evaluated as part of this initial analysis.
- **Classification:** Numerical conversions will enable mathematical classification calculations from venue categorical records.
- **Clusters:** Consistent themes, classification clusters and outliers will be identified from the sample data.
- **Constraints:** The intent of analysis is an initial sample profile, so venue information will be limited to 100 records within a 500 metre radius.
- **Visualisation:** Additional context will be provided by generating visualisation of spatial proximity and classification clusters.

METHODOLOGY OF ANALYSIS

The creditability of analysis and associated results is commensurate with the reliability of the methods deployed to achieve the results. Understanding the structure and process used to generate outcomes is essential to the confidence and value attributed to results.

The analysis applied for this mandate guided by the Foundational Methodology for Data Science framework developed by John Rollins (8), as illustrated below:



A summary of how the framework has been incorporated is described in the following sections. It should be noted that the summary given in this report is intended only as an explanatory overview. Analysis commentary provided in conjunction with this report contains specific details of how data understanding techniques were applied throughout the analysis process. Application of methodology is discussed as follows:

1. Business understanding

The first requisite was to define the business purpose, creating opportunities to distribute Australian meat pies among eateries in in Toronto. The definition sets the purpose and context for subsequent analysis. Changes to the business definition would result in changing the analytic approach. Key elements that will shape the subsequent analysis:

- **Business purpose:** Aussie Pies will distribute meat pies through retail food outlets, for customers wanting to enjoy a convenient, hot, meat filled pastry savoury.
- **Target market:** Prospective distributors are casual eateries, pubs and bars, with initial focus on metropolitan Toronto.

- **Sales channel**: Product distribution is business to business rather than though a physical retail outlet, aiming to provide an agile, scalable market entry point for the new venture.

2. Analytic approach

The fundamental approach is to create information of value from available data, which is predominantly a profile of surrounding venues. The relevance of surrounding venues is described in the preceding section headed “Data priority”. In order to derive the profile, the analysis incorporates the approach described under the section headed “Data solutions”.

3. Data requirements

The data requirements to undertake the scope of analysis is set out in the sections headed “Data source - target neighborhoods” and “Data source - spatial data”.

4. Data collection

Data to meet the requirements identified has been collected from the following three sources:

- Wikipedia - List_of_postal_codes_of_Canada:_M (5)
- Cognitive Labs – Geospatial data (6)
- Foursquare – Venue names and categories (9)

The required fields from the respective datasets was sequentially combined throughout analysis, joining on a field common to respective sets. The following table illustrates unique and common data between the datasets, shading the intersection of data to the source:

Data type	Wikipedia - M Postal Codes - Canada	Cognitive - Geospatial Data	Foursquare	Total
Borough				**1**
Latitude				**2**
Longitude				**2**
Neighborhood				**1**
Postal Code				**2**
Venue Name				**1**
Venue Category				**1**
Total	**3**	**3**	**4**	**10**

Data collection was not a linear process but brought in progressively, triggered as additional information is required.

5. Data understanding

Building understanding of the data is not a discrete step but developed progressively and looping back to data collection or leading to subsequent data preparation. Relatively elementary but effective techniques of visualisation and data clustering were applied to derive data understanding. Those techniques are described as follows:

Output	Description	Application to gain understanding:
Dataframe (df)	The official definition is two-dimensional, size-mutable, potentially heterogeneous tabular data (10). A dataframe may be more commonly understood as a matrix or table, structuring discrete information horizontally, in vertically grouped columns.	Produce a simple visual table to evaluate how the data is organised in a range of variations as below: ▪ Partial= Sample of top (head()) or bottom (tail()) rows for a simple assessment. ▪ Complete = Full list for rare occasions where all the data is to be visualised, typically only small or filtered datasets.
Dimension shape	The df dimensions by columns and rows (11)	Enables quick assessment of applicable data volume, particularly useful to check context upon initial sourcing of the data or following a change.
Maps	Scalable street map ultimately sourced from OpenstreetMap (12).	Information is generated in multiple outputs to provide a circumspect context for interpretation. Maps provide visual context of spatial relationships and proximity clusters across a street map of Toronto. Maps produced are: ▪ **map_toronto**: simple map overlay depicting spatial proximity of neighborhoods across metropolitan Toronto. ▪ **map_clusters**: refined to neighborhoods with 'Toronto' as part of the name, including clusters of venue categories overlay the street map.
Statistical description	Compute descriptive statistics for: count, mean, standard deviation, minimum value, maximum value, percentiles.	Generate statistical description of venue categories by neighborhood, confirming the applicable range of counted values.
Data types	Report of the data characteristics such as text, integer or decimal.	Confirms: ▪ clustering labels produced prior application to neighborhoods are integers. ▪ Statistics of venue categories by neighborhood were produced as decimals (float64).
Clustering	A machine learning algorithm attempts to classify data in relevant groups (13).	Facilitates:

Output	Description	Application to gain understanding:
		▪ Dataframes and reports and of clustered neighborhoods and venue categories. ▪ Segmentation of clusters by colour ▪ Map overlays of clusters by neighborhood and venue category group.
Unique counts	Identifies unique data rows	Unique venue categories were counted

6. Data preparation

Data delivered from the respective sources is not automatically functional, but requires repeated processing to transform into meaningful, useful information. The iterative process preparing and transforming data is commonly referred to as 'data wrangling' (14; 15).

Examples of data preparation applied continuously throughout the analysis is tabled below:

Component	Description	Application examples
Tools	The software applications utilised to perform the functions data preparation.	▪ Python programming (16) language was used for enacting the transformation process. ▪ Jupyter notebooks (17) enabled creation and sharing of the coding.
Dataframes	Conversion of raw data into a dataframe	▪ Python code identifies the data in a wide range of formats and converts into a dataframe.
Combining data	The data collection process led to information in multiple dataframes that needed to be combined to be functional. Combining data can take the form of merge / joins or appends.	▪ Postal code data was joined with Geospatial records. ▪ Appending of columns was used to create 'neighborhoods_venues_sorted' ▪ 'toronto_merged' was developed through merging of 'neighborhoods_venues_sorted'.
Summarising	Grouping is the main form of summarisation used	▪ Post codes were grouped by Postal Code. ▪ 'toronto_venues_count' was grouped by venue category count
Dropping values	Some rows of data do not contain any data, contain incompatible data or duplicates. Such rows or columns are dropped (removed) from the dataframe.	▪ The postal code data from Wikipedia contained numerous rows of 'Not Assigned' These values were eliminated or 'dropped' from the dataframe. ▪ Duplicate rows and columns created through joining or appending various dataframes were subsequently dropped.

7. Modelling

The analysis has minimal modelling but has utilised K-Means (18) modelling to produce the clustering. Preparation for modelling required data manipulation referred to as 'get dummies (19)', converting the categorical (text) records of the venue categories into numeric values. This preparation enabled reports identifying the most common venues for respective neighborhoods.

Clustering is essentially a process of grouping data into clusters based upon data patterns and K-Means clustering applies a distanced algorithm that associates each cluster with a centroid (20). The clustering was used to visualise the patten of venue categories by neighborhood, overlaying a street map.

8. Evaluation

The current level of analysis is rudimentary, commensurate with the phase of business concept development. The analysis does incorporate basic evaluation in the form of tables and cluster reports used to cross reference the modelling outputs to the base data. Ideally progression of the concept would include sponsoring expanded evaluation and diagnostics.

9. Deployment

The submission of this report to the sponsors is pending. If the report is accepted, there should be some preliminary deployment to assess the validity before developing into an operational model. Validation should include variables such as multiple groups, skills and technologies (8).

10. Feedback

The submission of this report to the sponsors is pending. If the analysis is accepted and leads to progressing the business concept, then there should be sponsorship to rework and refine the analysis.

RESULTS

This section will present analysis outcomes that shape the conclusion. The steps described 'METHODOLOGY OF ANALYSIS' morph and recirculate throughout the analysis rather than following a simple linear flow. Results will be linked methodology steps incorporated but will focus on explaining results than precise correlation to the methodology.

Defining the Target Neighborhoods – part 1

Creating the first data frame

Notebook cell references: ([0] – [5])

The initial area of interest was based upon neighborhoods included within the Canada postal code sector 'M' as displayed below (21):

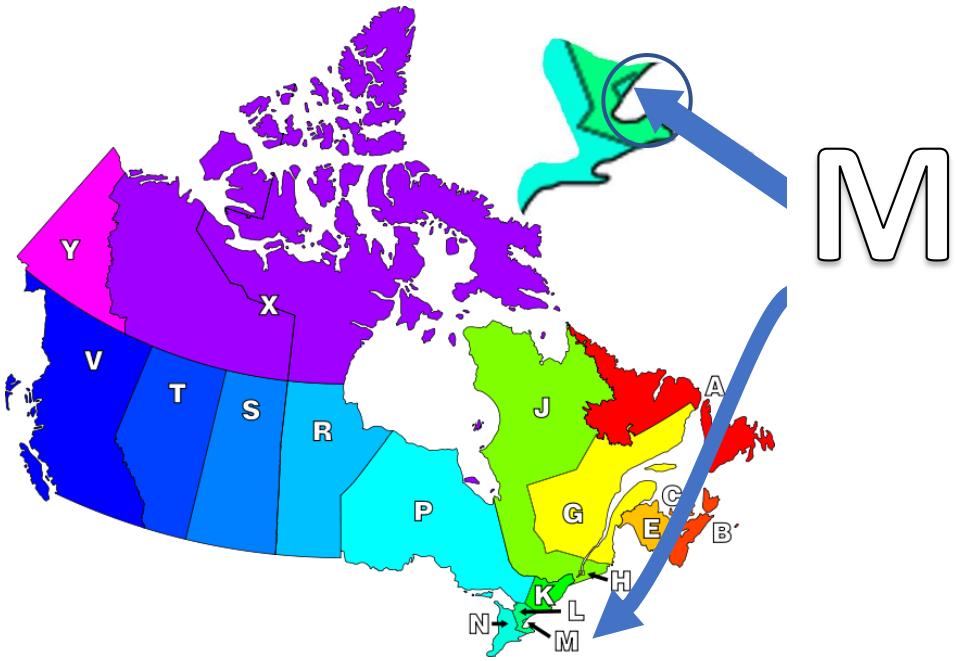


Image credit: (22)

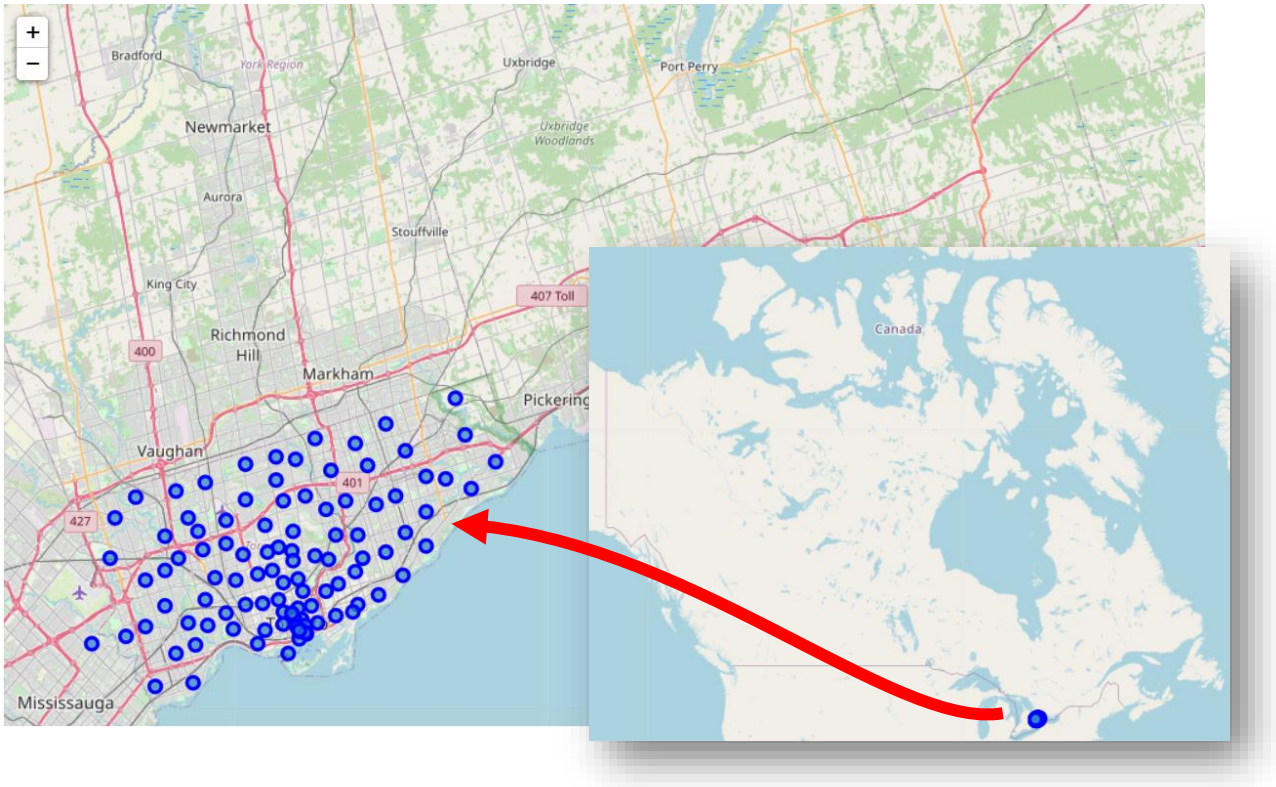
The initial dataframe sourced from Wikipedia (5) was 180 rows but included numerous records with values of 'Not assigned'. As discussed in the section "

Data preparation", these records were dropped from the dataframe. The resulting dataframe was 103 rows of boroughs and neighborhoods.

Integrating geospatial records

Notebook cell references: ([6] – [10])

The postal records were joined with geospatial records (6). The merging of longitude and latitude records enabled generation of a street map overlaid with blue markers centred on the spatial coordinates. The result enables visual appraisal of the proximity and density of centred neighborhoods. Following is the resulting map illustrating centred neighborhood surrounding metropolitan Toronto, as well as inset of the same map showing context to the rest of Canada:



Defining the Target Neighborhoods – part 2

Refining the scope

Notebook cell references: ([11])

The business concept is at the embryonic stage, so it was decided to filter the neighborhoods under consideration to produce a more refined sample around metropolitan Toronto. A new dataframe (toronto_data) was created by retaining only boroughs with 'Toronto' in the name. The refined dataframe narrowed the scope to 39 rows of boroughs and neighborhoods, which is the basis for the remainder of analysis. The entire 39 records were generated as a printed dataframe as for simple review as this is a very small data set.

Integrating venue categories

Notebook cell references: ([12] – [16])

After refining the areas under consideration, the additional information needs to be integrated to create a profile of surrounding venues. Venue information was sourced from Foursquare (9). For practical purposes, records were limited to 100 nearby venues within a radius of 500 metres from the centred neighborhood. The imported data was combined with 'toronto_data' to form a new dataframe 'toronto_venues'. Integrating the sample set is now 1,614 records, as multiple venues per neighborhood were integrated. The following sample head shows that venues have close but distinct spatial coordinates to the neighborhood centres they are linked to:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop

The dataframe output provides a simple but crucial check to understand the data. It would be evidence of some inaccuracy if all the venues shared the exactly the same coordinates as the neighborhood centre.

Quantify venues by category

Descriptive statistics

Notebook cell references: ([17] – [19])

The unique values among the 1,614 venues counted was reviewed in two ways.

- **Unique**: An algorithm was used to report the quantity of unique venue categories as a simple total of 236.
- **Grouping:** Venue categories were grouped, counting the number of applicable venues and producing a report as a new dataframe 'toronto_venues_count' sorted by quantity from highest to lowest. Following is a sample of the top five values:

Venue Category	Count
Coffee Shop	143
Café	89
Restaurant	54
Italian Restaurant	41
Hotel	37

As shown in the preceding table 'Restaurant' and 'Italian Restaurant' are deemed unique categories by the default Foursquare records. This has provides valuable insights to the diversity of tastes catered for but needs additional modification to derive the consolidated class of eatery of 'Restaurant' as nominated in 'Data priority'. In subsequent sections the venue categories will be consolidated further to be consistent with categories set out in 'Data priority'.

- **Descriptive statistics**: A summary report of the grouped venues (toronto_venues_count') was produced give descriptive statistics confirming the highest venue group was 'Coffee Shop', with 142 counted from a total of 236 unique values. The descriptive statistical summary produced is shown below:

Venue category	Value
Count of total categories	233.000000
Mean / category	6.927039
Standard deviation / category	13.058515
Minimum count / category	1.000000
25% quartile count / category	1.000000
50% quartile count / category	3.000000
75% quartile count / category	7.000000
Maximum count / category	143.000000

These statistics confirm that coffee shops are the dominant single venue category but also indicates that the categorisation is too narrow for the remainder of the records for the purposes of this mandate. This will be addressed in the following section.

Modelling

Report most common venues

Notebook cell references: ([20] – [24])

In order to create a richer profile of venues across neighborhood groups, categorical data was transformed into numeric values. The transformed data was then arranged and sorted by most common venue categories by neighborhood groups. Following is a cut down sample showing just five neighborhood groups and only the first and 10th most common venue categories.

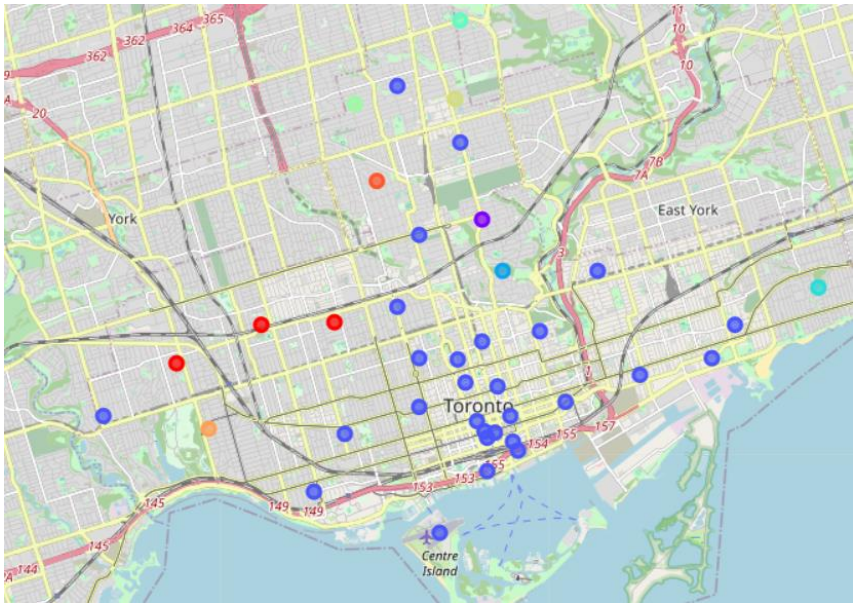
Neighborhood groups	1st Most Common Venue	2 – 9 Most Common Venue	10th Most Common Venue
Berczy Park	Coffee Shop	Removed for this illustration. Refer Jupyter Notebook for details	Shopping Mall
Brockton, Parkdale Village, Exhibition Place	Café		Furniture / Home Store
Business reply mail Processing Centre, South C...	Light Rail Station		Pizza Place
CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge		Airport Gate
Central Bay Street	Coffee Shop		Bubble Tea Shop

The modelling produced a full report of 10 most common venue categories across the 39 neighborhood groups. Such a report is useful as a ready reference on venue profile by neighborhoods.

Map of most common venues

Notebook cell references: ([25] – [27])

Transformation of categorical information in numeric text values also enabled production of a map showing clusters of venue categories across the target neighborhoods. The output is shown below:



The output clearly depicts the dominant cluster as purple markers, inviting additional investigation to understand distinction with clusters attributed to markers of other colours.

Investigating the clusters

Notebook cell references: ([28] – [39])

In order to understand the clusters further, reports were created grouping clusters around latitudes. The summary of computation is tabled below:

Cluster	Count
0	**3**
1	1
2	28
3	1
4	1
5	1
6	1
7	1
8	1
9	1

The output identified that Cluster 1 and 6 are the only clusters with a count greater than 1, reflecting the grouping by latitude. There is opportunity to continue refining the model, but Cluster 1 is arguably an adequate profile for the purposes of this mandate.

Grouping venue categories by data priority

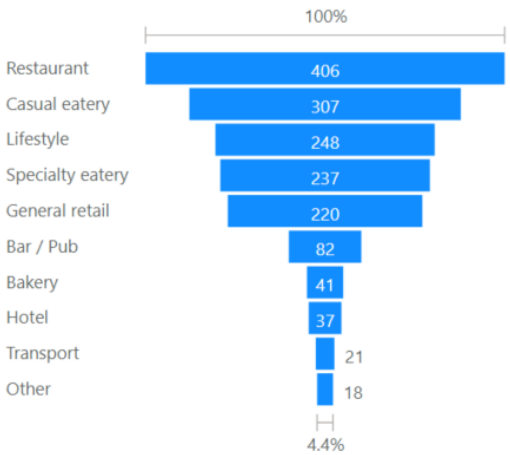
Reference: APPENDIX – National food themes

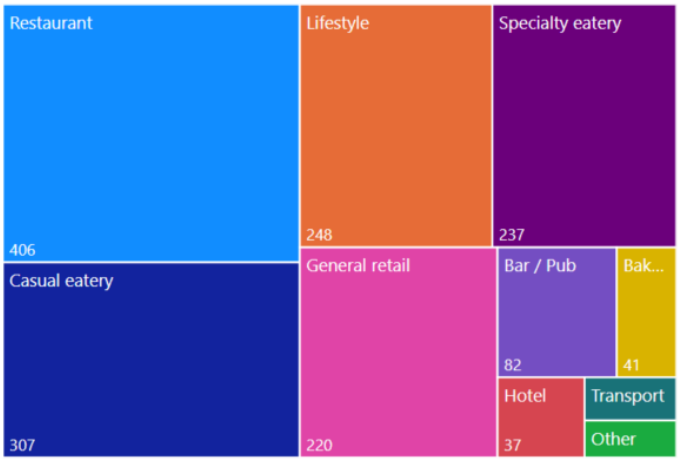
All the work so far has been computed using Python programming on Jupyter notebooks. Microsoft Power BI has been utilised to produce the supplementary analysis of venue categories. The dataframe of the 236 unique venue was re-categorised according to the venue groups specified in 'Data priority'. The new venue groups included identifying culinary venues that had a specific regional or national focus (National Foods). Default Microsoft Power BI functions was used to produce the following tables an visuals to supplement the Python computations.

Profile of venue categories by group

Group	Quantity	% of total Quantity	Nationality Foods	% of total Nationality Foods
Restaurant	406	25.11%	271	89.44%
Casual eatery	307	18.99%		
Lifestyle	248	15.34%		
Specialty eatery	237	14.66%	24	7.92%
General retail	220	13.61%		
Bar / Pub	82	5.07%		
Bakery	41	2.54%	8	2.64%
Hotel	37	2.29%		
Transport	21	1.30%		
Other	15	1.11%		
Total	**1614**		**303**	

Following is the same information shown as a funnel map that layers the data in order of the most common group and a tree map showing the comparative group proportions:





This analysis is important as the business premise is introducing food with an Australian theme. This table provides the following insights relating to the 1,614 venues:

- **1,073** or **66%** are clearly targeted to the culinary industry
- **389** or **24%** are aligned to the casual eatery or bar / pub group as potential distributors
- **303** or **18%** market their product according to a national food theme
- **41** or **2.5%** are bakeries, of which eight already have a dedicated focus

Venue groups with a national food theme

The recategorization was taken a step further investigating the range of culinary venues represented by a specific national food theme. The top 10 national food themes reflected out of 303 aligned to a national cuisine are:

Country	National Foods
Japan	59
Italy	41
United States	39
Thailand	30
France	22
Mexico	21
Greece	17
China	11
Middle East	8
India	7
Total	255

Geographical representation of food themes

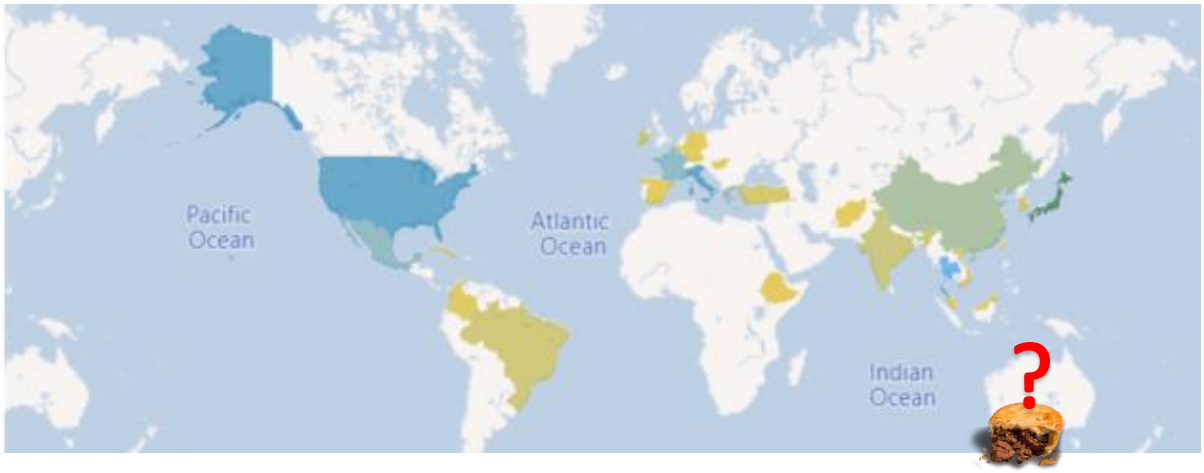
The regions represented by national food themes were displayed across world maps to provide a visual representation of the diversity of national food themes.

In order to generation the illustrations, the following refinements were made:

- **Exclusions:** Venues that are not food related or do not indicate a national food theme were not included. An example of and excluded food venues is ‘Coffee Shop’ or ‘Restaurant’.
- **Substitution**: Venues that indicate a food type rather than nationality were attributed to a country. Examples are ‘Burrito’ attributed to Mexico and ‘Ramen’ to China. Venues indicating a region rather than country were attributed to a country of that region. An example is ‘Middle East’ attributed to Saudi Arabia and ‘Eastern Europe’ to Hungary.

In cases where a specific a food type or a region is named rather than a specific nationality, a proxy has been used. Examples include using ‘Saudi Arab

Continents represented (other than Canada):



Nationalities represented by quantity of venues (other than Canada):

These visualisations identify two important factors in that the target neighborhoods:

- support a diverse range of international food interests
- Australia has negligible representation among the venues included

DISCUSSION OF RESULTS

The purpose of this section is summarise the analysis, reinforcing how steps processed contributed to the results.

The analysis has been divided into the following phases:

- Defining the Target Neighborhoods (Part 1 and 2)
- Quantify venues by category
- Grouping venue categories by data priorities
- Modelling

This section summarises the results link to the methodology framework.

Defining the Target Neighborhoods – part 1

Creating the first data frame

A dataframe has been produced defining the target neighborhoods as boroughs within the M postcode sector, eliminating rows with the value 'Not assigned'.

Output: Dataframe refined from 180 to 103 rows of boroughs and neighborhoods.

Primary methodology steps – creating the first dataframe:

- Step 1 - Business understanding: Focus on the target market area of metropolitan Toronto
- Step 2 - Analytic approach: Aligns focus to the purpose of creating a market profile
- Step 3 - Data requirements: Identifies data sources required
- Step 4 - Data collection: Sources some of the data required
- Step 5 - Data understanding: Identified data deficiencies
- Step 6 - Data preparation: Dropped data identified as deficient

Integrating geospatial records

Geospatial records were matched to the postal codes enabling analysis by spatial coordinates.

Output: Map of Toronto overlaid with blue markers indicating the centers of the shortlisted neighborhoods.

Primary methodology steps:

- Step 4 - Data collection: Source additional of the data required
- Step 5 - Data understanding: Use visualisation (map) to assess data
- Step 6 - Data preparation: Modified to generate mapping

Defining the Target Neighborhoods – part 2

Refining the scope

The foundational results of 'Defining the Target Neighborhoods – part 1' were refined to boroughs including 'Toronto' in the name.

Output: Refined dataframe narrowed 39 rows of boroughs and neighborhoods

Primary methodology steps:

- Step 5 - Data understanding: Used descriptive statistics (counts), review new dataframe
- Step 6 – Data preparation: Refined data to produce a more focused sample set

Integrating venue categories

Venue category information source from Foursquare was matched to the spatial coordinates.

Output: Dataframe identifying 1,614 venues across the 39 rows of boroughs and neighborhoods

Primary methodology steps:

- Step 4 - Data collection: Sources data from Foursquare
- Step 5 - Data understanding: Review outputs checking reasonableness of data
- Step 6 - Data preparation: Aggregated additional data

Quantify venues by category

Descriptive statistics

Outputs: New dataframe grouping venues to count quantities per neighborhood; summary statistics of unique categories per neighborhood.

Primary methodology steps:

- Step 5 - Data understanding: Use of descriptive statistics to understand content

- Step 6 - Data preparation: Transforming the data

Grouping venue categories by data priority

Reference: APPENDIX – National food themes

Profile of venue categories by group

The venue categories were allocated groups consistent with the descriptions outlined in the section 'Data priority'.

Outputs: Tables and visuals depicting the relative proportions of venue categories across the target neighborhoods.

Primary methodology steps:

- Step 5 - Data understanding: Use of visualisation to understand content
- Step 6 - Data preparation: Transforming the data

Venue groups with a national food theme

Culinary venues with a national food theme were identified to indicate the range of cuisines represented across the target neighborhoods.

Outputs: Tables and visuals depicting the international diversity of cuisines across the target neighborhoods.

Primary methodology steps:

- Step 5 - Data understanding: Use of visualisation to understand content

Modelling

Report of most common values

The categorical of information of venue was transformed to numerical values to identify the most common venues per neighborhood.

Output: Dataframe of most common venue categories by neighborhood groups.

Primary methodology steps:

- Step 5 - Data understanding: Use of descriptive statistics to understand content
- Step 6 - Data preparation: Transforming the data
- Step 7 - Modelling: Develop descriptive model using analytic approach
- Step 8 - Evaluation: Use tables or graphs for evaluation

Map of most common values

A form of machine learning was applied to produce a visualisation of the clusters across the target neighborhoods.

Output: Map of clustered venue categories across target neighborhoods

Primary methodology steps:

- Step 6 - Data preparation: Transforming the data
- Step 7 - Modelling: Develop descriptive model using analytic approach
- Step 8 - Evaluation: Use visualisation for evaluation

Investigating the clusters

The clusters represented on the on map were grouped by latitude converted into a dataframe for every cluster.

Outputs: Dataframes defining every cluster as well a summary confirming that only two clusters had more than 25 rows of most common venues assigned.

Primary methodology steps:

- Step 8 - Evaluation: Use tables for evaluation

Application of methodology

A secondary but useful insight generated from this analysis is the application of methodology across the process. The bulk of effort is concentrated to data preparation and understanding, as illustrated by the following table.

Methodology Step	Defining the Target Neighborhoods - part 1	Defining the Target Neighborhoods - part 2	Grouping venue categories by data priority	Modelling	Quantify venues by category	Grand Total
Step 1 - Business understanding	1					1
Step 2 - Analytic approach	1					1
Step 3 - Data requirements	1					1
Step 4 - Data collection	2	1				3
Step 5 - Data understanding	2	2	2		1	7
Step 6 - Data preparation	2	2	1	2	1	8
Step 7 - Modelling				2		2
Step 8 - Evaluation				3		3
Grand Total	**9**	**5**	**3**	**7**	**2**	**26**

Note: Steps 9 and 10 remain pending the outcome of this report by the business sponsor.

Additional analysis

An essential principle to any commercial element of data science is balancing the perpetual trade off between perfect analysis and effective application. Research is a business cost, so investing further in research needs to be balanced with the probable return on investment.

If sufficient information has been derived for the business sponsor, then other areas of research may present more compelling investment. For example, if at this point the business sponsor determines the data is suggesting competitive conditions are unpalatable then additional investment offers detrimental cost with no value. Conversely the sponsor may decide to change the analytic approach to address other opportunities or risk.

Before expending additional work, it is advisable to realign with to understanding the business purpose which shaped the analytic approach.

The project mandate is to create initial market profile for a proposal to introduce a new product line of Australian meat pies into Canadian city of Toronto. The conclusion at this point is to confirm with the business sponsor before progressing further.

CONCLUSION

The purpose of this analysis is to create a profile of metropolitan food venues for a proposal to introduce Australian pies as a new product line to Canada, starting in Toronto. This analysis aims to identify food outlets that may incorporate the product as an opportunity in a competitive market.

This analysis has aimed to demonstrate a verifiable approach to investigating to sourcing and refining available data to deliver information of value. Data priority, sources and solution were defined as foundational to the analysis. Toronto postal records was combined with geospatial coordinates and location data from Foursquare to create the profile. The subsequent analysis was aligned to a methodology that iteratively curates the data based upon the business understanding and approach. The analysis produced a range of perspectives that combine to provide an overall profile of venue categories that may support, complement or challenge the proposed venture. A synopsis of the analysis outputs that make up the overall profile are:

- **Target neighborhoods**: A shortlisted focal subset of **39** boroughs that include 'Toronto' in the name was systematically refined from an initial list of **180** postal codes.
- **Data priority:** The venue categories were grouped according to the data priorities, reshaping the perspective to show restaurants as the most prominent venue across the sample set.
- **Venue diversity:** The **1,614** venues across **236** unique categories identified, with coffee shops and café's as the most prominent.
- **Venue clusters:** The clusters of most common venues per neighborhood are represented by dataframe reports and street maps overlaid with markers indicating the clusters.
- **Venue grouping profile:** Approximately 24% of the 1,614 identified venues are either casual eateries or bar/ pubs, fitting the prospective market as potential distributors. Less than 3% are bakeries selling pastry products.
- **National food themes:** Venue categories indicate a diverse representation of international cuisines, of which Japan and United States dominate. A notable observation is that Australia in general and Australian pies in particular do not appear represented.

The profile of venues surrounding metropolitan Toronto appear congruent to the concept of food retailers an additional sales opportunity while providing an agile, scalable market entry point for the new venture. The profile created indicates that Australian meat pies would be a relatively distinct product line among Toronto eateries. There remains significant additional planning beyond the scope of this assessment that should take place as part of prudent business planning. This assessment may however present sufficient prospects for a keen Aussie entrepreneur to 'have a go' and 'try their luck'!

References

1. **Statistica.** Resident population of Canada in 2019, by metropolitan area. [Online] 09 Mar 2020. https://www.statista.com/statistics/443749/canada-population-by-metropolitan-area/.

2. **Lester, Kim.** Meat pie exhibition: How an ancient fast food became an Australian icon. *ABC NEWS.* [Online] ABC, 05 Apr 2016. https://www.abc.net.au/news/2016-04-04/meat-pie-exhibition-tracks-history-of-australia-iconic-fast-food/7297216.

3. **Connell, Jan.** The great Australian pie. *Australian food history timeline.* [Online] 2020. https://australianfoodtimeline.com.au/great-australian-pie/.

4. **Jchmrt.** File:Sunset Toronto Skyline Panorama Crop from Snake Island.jpg. *WIKIMEDIA COMMONS.* [Online] N.D. https://commons.wikimedia.org/wiki/File:Sunset_Toronto_Skyline_Panorama_Crop_from_Snake_Island.jpg.

5. **Wikipedia.** List of postal codes of Canada: M. [Online] n.d. . https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

6. **Cognitive Labs.** [Online] n.d. https://cocl.us/Geospatial_data.

7. **FOURSQUARE.** FOURSQUARE DEVELOPERS. [Online] 2020. https://developer.foursquare.com/.

8. **Rollins, John.** IBM Big Data & Analytics Hub. *Why we need a methodology for data science.* [Online] IBM, 24 Aug 2015. https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science.

9. **FOURSQUARE.** api. [Online] Jun 2020. 'https://api.foursquare.com/v2/venues/explore?&client_id.

10. **Pandas Development Team.** API reference >> pandas.DataFrame. *pandas.* [Online] 2014. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html.

11. —. API Reference > property DataFrame.shape. *pandas.* [Online] 2014. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.shape.html?highlight=shape#pandas.DataFrame.shape.

12. **OpenStreetMap.** About. *OpenStreetMap.* [Online] n.d. https://www.openstreetmap.org/about.

13. **Maklin, Cory.** K-means Clustering Python Example. [Online] Medium, 29 Dec 2018. https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203.

14. **Raza, Mohammed.** Data Wrangling With Pandas. [Online] Medium, 13 Nov 2018. https://towardsdatascience.com/data-wrangling-with-pandas-5b0be151df4e.

15. **Talend.** Data Wrangling: Speeding Up Data Preparation. *Resources.* [Online] Talend, 2020. https://www.talend.com/resources/data-wrangling/.

16. **Python Software Foundation.** About. *python.org.* [Online] 2020. https://www.python.org/about/.

17. **Project Jupyter.** About Us. *jupyter.org.* [Online] 25 Jun 2020. https://jupyter.org/index.html.

18. **Pedregosa et al.** sklearn.cluster.KMeans. *scikit learn.* [Online] 2011. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

19. **w3resource.** Pandas: Data Manipulation - get_dummies() function. *w3resource.* [Online] 04 May 2020. https://www.w3resource.com/pandas/get_dummies.php.

20. **Sharma, Pulkit.** The Most Comprehensive Guide to K-Means Clustering You'll Ever Need. *Analytics Vidhya.* [Online] 2019. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

21. **Wikipedia.** Postal codes in Canada. *Wikipedia.* [Online] n.d. https://en.wikipedia.org/wiki/Postal_codes_in_Canada.

22. **Denelson83.** File:Canadian postal district map.svg. [Online] WikiMedia Commons, 2007. https://commons.wikimedia.org/wiki/File:Canadian_postal_district_map.svg.

23. **User:Pmx.** File:Canada flag map.svg. *https://upload.wikimedia.org/wikipedia/commons/8/8d/Canada_flag_map.svg.* [Online] 2007. https://commons.wikimedia.org/wiki/File:Canada_flag_map.svg.

24. **Pandas Development Team.** pandas.get_dummies. *pandas.* [Online] 2014. pandas.get_dummies.

APPENDIX – National food themes

The following table defines how venues with a food theme were attributed to a country.

Country	Venue Category	Neighborhood count
Japan	Japanese Restaurant	33
	Sake Bar	1
	Sushi Restaurant	25
Japan Total		**59**
Italy	Italian Restaurant	41
Italy Total		**41**
United States	American Restaurant	21
	Burger Joint	9
	Cajun / Creole Restaurant	1
	New American Restaurant	8
United States Total		**39**
Thailand	Asian Restaurant	11
	Thai Restaurant	19
Thailand Total		**30**
Mexico	Burrito Place	11
	Mexican Restaurant	10
	Taco Place	1
Mexico Total		**22**
France	Creperie	8
	French Restaurant	10
	Modern European Restaurant	3
France Total		**21**
Greece	Greek Restaurant	13
	Mediterranean Restaurant	4
Greece Total		**17**
China	Chinese Restaurant	6
	Noodle House	1
	Ramen Restaurant	4
China Total		**11**
Saudia Arabia	Middle Eastern Restaurant	8
Saudia Arabia Total		**8**
Brazil	Brazilian Restaurant	3
	Latin American Restaurant	4
Brazil Total		**7**
India	Indian Restaurant	7
India Total		**7**
Vietnam	Vietnamese Restaurant	7
Vietnam Total		**7**
Ireland	Irish Pub	5
Ireland Total		**5**
South Caribbean	Caribbean Restaurant	4
South Caribbean Total		**4**

Country	Venue Category	Neighborhood count
Turkey	Doner Restaurant	1
	Falafel Restaurant	2
	Hookah Bar	1
Turkey Total		**4**
Hungary	Eastern European Restaurant	3
Hungary Total		**3**
Cuba	Cuban Restaurant	2
Cuba Total		**2**
Malaysia	Malay Restaurant	2
Malaysia Total		**2**
Morroco	Moroccan Restaurant	2
Morroco Total		**2**
Colombia	Colombian Restaurant	2
Colombia Total		**2**
South Korea	Korean Restaurant	2
South Korea Total		**2**
Afghanistan	Afghan Restaurant	2
Afghanistan Total		**2**
Belgium	Belgian Restaurant	2
Belgium Total		**2**
Taiwan		1
Philippines		1
Germany		1
Ethiopia		1
Grand Total		**303**

Image: (23)