# **New Restaurant Location Suggestion**

May 2020

1. **Introduction** 
   1. **Background**

The restaurant business is an ever growing industry that comes with its own levels of uncertainty and risk, among the many things one has to consider in preparation to embark on a new restaurant opening, the main concerns in my opinion are the type and location of the restaurant. Therefore any additional insight one could garner to inform this decision would be advantageous.

* 1. **Problem**

The data around location and restaurant type is not a readily available data set to come by, using location, type and ratings this project will help to identify areas where popular restraint types have yet to find a footing.

* 1. **Interest**

The most interested parties would be individuals or businesses looking to open a new restaurant with no definable pre-requisites as to where and what they would like to set up, or those looking to take advantage of gaps in the market

1. **Data collection and cleaning**
   1. **Data Sources**

The data had to be pulled from 2 sources and combined to form the base data set for analysis, the suburb list was collected from [here](http://geo.mycyberict.com/south_africa/johannesburg/), which was the main source for getting a full itemized listing of the suburbs in my province, the second source as prescribed by the course outline was Foursquare, which was used to pull the nearby restaurants and their ratings.

In addition to this the data collection also involved the use of the Geopy Package in python that was used to source the Coordinates for all suburbs.

* 1. **Data Cleaning**

Data was 1st pulled for all suburbs and their PO and street postal code, these were run through Geopy in order to gather the coordinates for each suburb, the cleaning on this required that areas with the exact same coordinates be removed, this was to account for the issue of area extensions that common in my city, in where ‘Bryanston’ and ‘Bryanston EXT 21’ would not result in duplicated venues, as they are essentially the same area. Any suburb that were polled through geopy with no coordinates were also dropped from the data set this accounted for less than 1% of the data after consideration for Extensions(EXT) is included.

On this dataset the below locations (table 1) were removed using an exclusion of points above the equator, due to their placements being incorrect, additional amendments were attempted - including the country however made no change in their polled location and therefor the 2 locations were removed.

|  |  |
| --- | --- |
| **Table 1. Location Removed from Dataset due to incorrect polling of coordinates** | |
| **Excluded location** | **Reason** |
| Edenburg, Johannesburg | Latitude & Longitude not correctly polled through Geopy |
| Lotto, Johannesburg | Latitude & Longitude not correctly polled through Geopy |

A similar methodology was adopted once the venue within each suburb were collected, on this front the location data was collected straight through foursquare as it was provided with the basic calls to the API. Due to the set radius of 2Km and the proximity of smaller suburbs it was necessary to remove the duplication of venues, this was done by dropping Venues who’s Latitude and Longitude were identical.

The Venue data also required cleaning, using Foursquare all venues within a 2Km radius were collected for each suburb together with their type, scrutiny of the list showed that there were key words or part thereof that could be used to identify only eateries table 2. Illustrates the list of Types were kept, the *Appendices ‘A’* Illustrates the full list of dropped categories.

* 1. **Feature Selection**

The feature selection for this project is fairly small the Final data table Consists of the below headers.

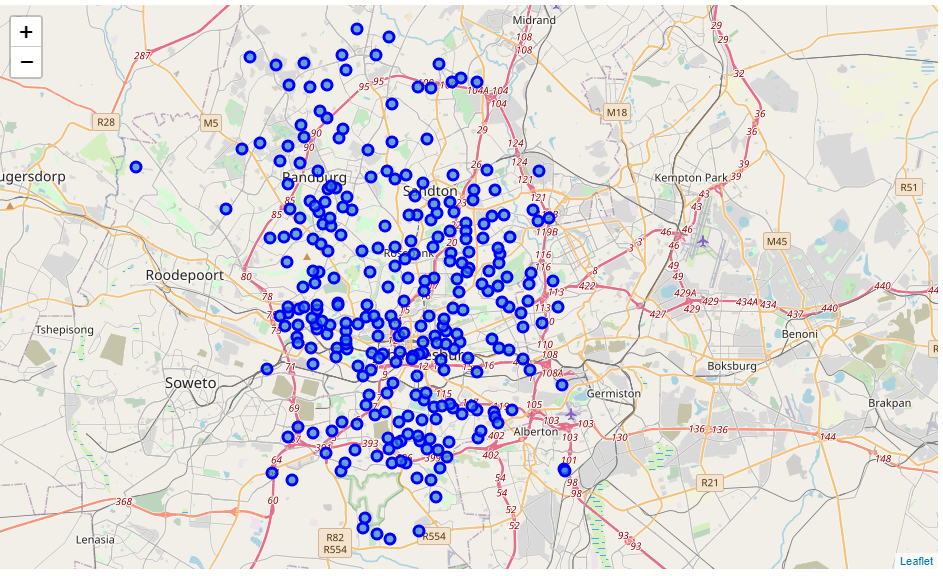
|  |  |
| --- | --- |
| **Final Feature Selection list** | |
| **Header** | **Description** |
| Suburb | Suburb name |
| Venue ID | Foursquare Venue ID, used for parsing queries |
| Venue Name | Name as per Foursquare of Venue ID |
| Category | Category as per Foursquare |
| Latitude | Latitudinal Coordinate of venue |
| Longitude | Longitudinal Coordinate of venue |
| Rating | Rating out of 10 for venue as per foursquare |

It was decided that for the ease of use for myself that the below headers were dropped during the cleaning process

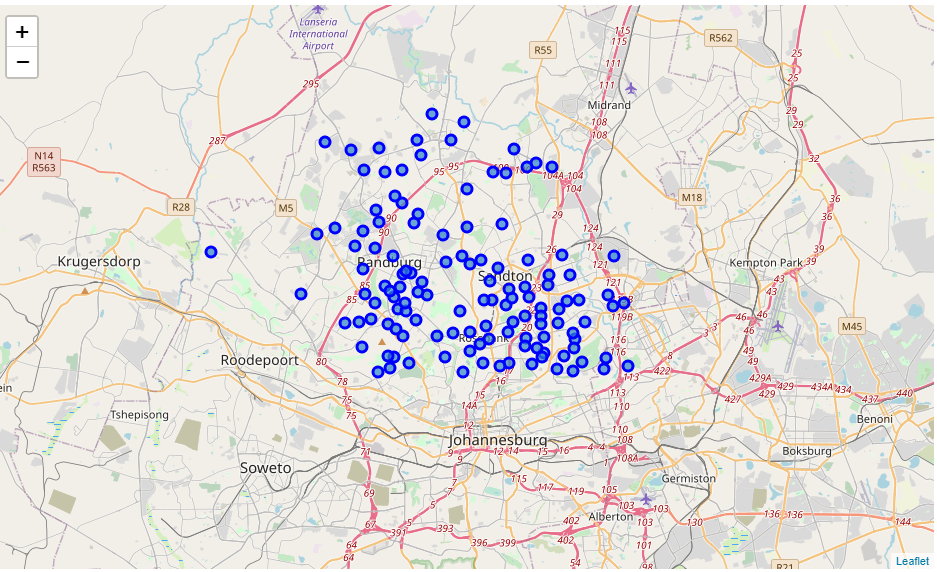
|  |  |
| --- | --- |
| **Dropped Features** | |
| **Header** | **Description** |
| Postal Code | The postal code has no value for this exercise |
| Street Code | The postal code has no value for this exercise |
| Latitude | This was the Latitude for the Suburb |
| Longitude | This was the Longitude for the Suburb |

1. **Exploratory Data analysis**
   1. **Identify northern suburb Line**

The process began with plotting all the suburbs I had accumulated for the city to determine the line that I would use to separate the more residential north from the CBD and commercial south.

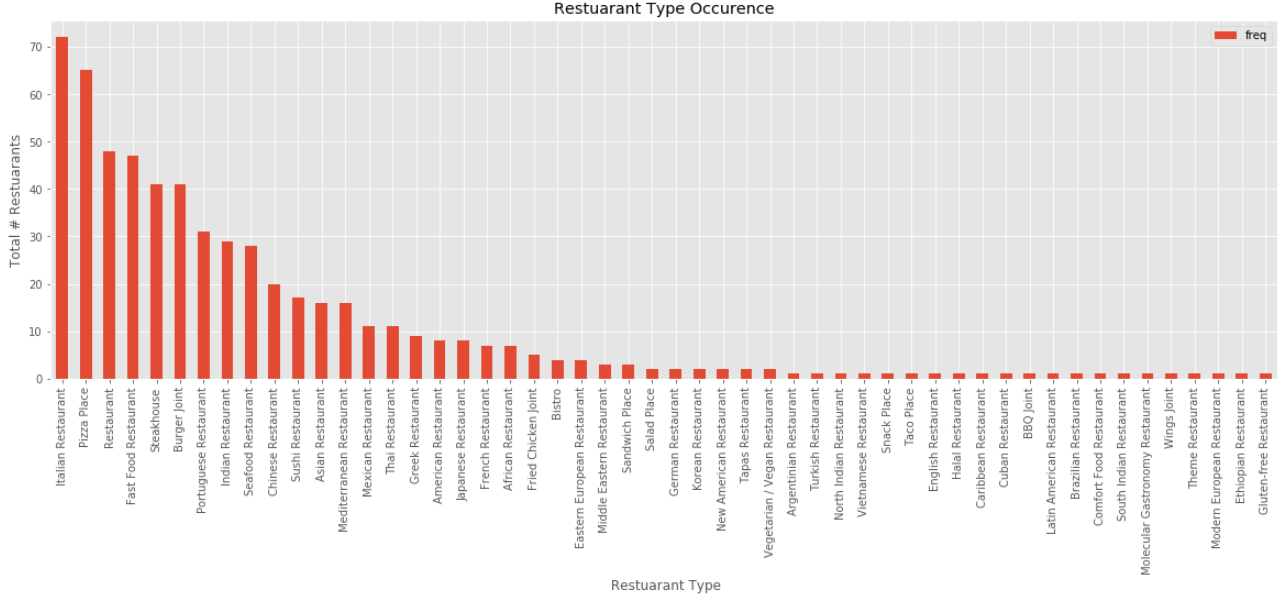


Eventually the Latitude of -26.166239 was used as the separator for my points of reference this resulted in a reduction of the points as shown below:

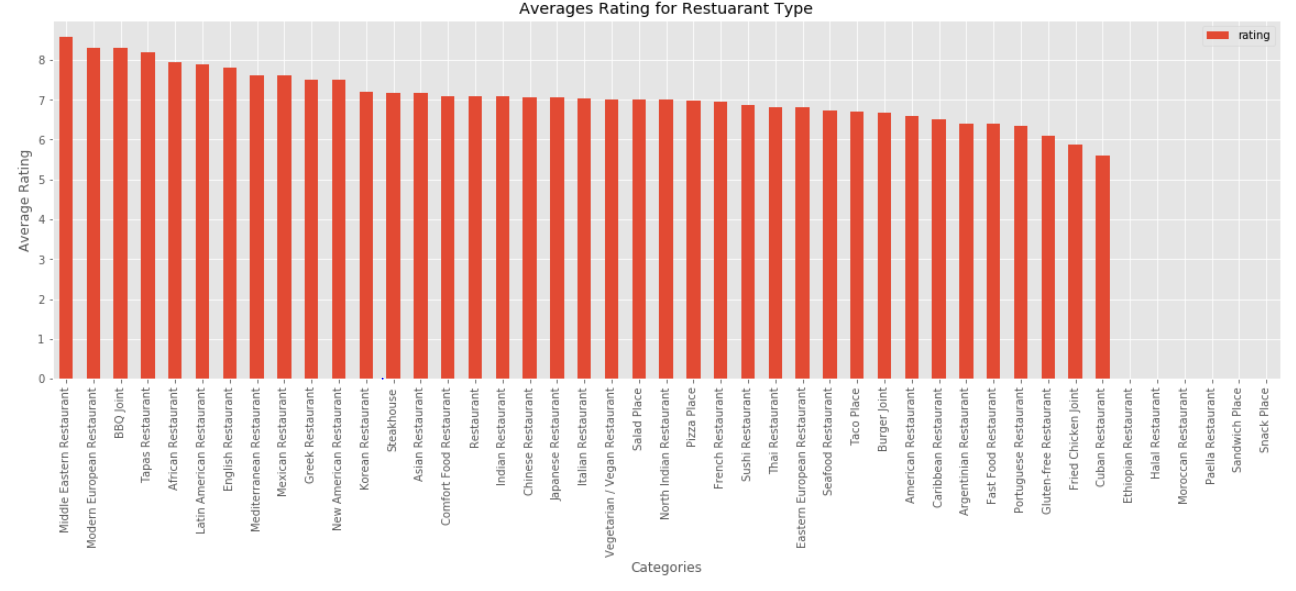


* 1. **Identification of Potential Restaurant Types**

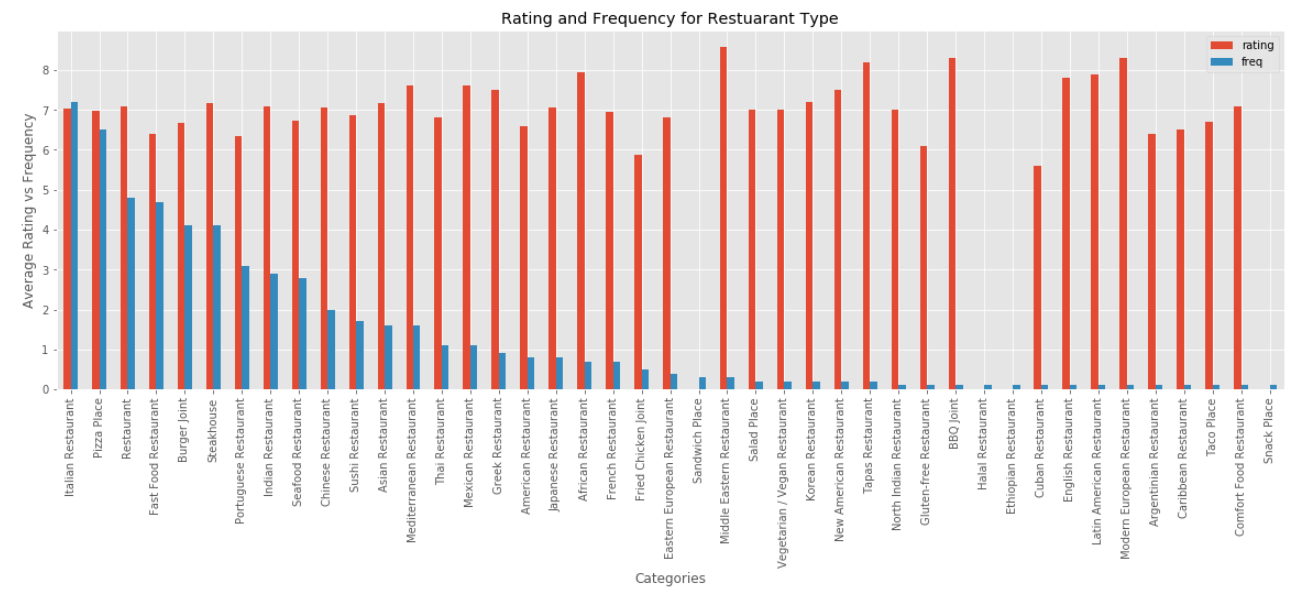
To try and identify a viable restaurant type to select I thought it best to do an analysis on the current mix of the market, Starting with the number of occurrences the different types of restaurant, I took the full venue list and did a count of the types found in the dataset. We see a distinct swing with Italian & pizza stores being the most frequently found restaurant type by some degree.



In addition to the occurrence of the restaurants its important to identify the average rating of the restaurant as the frequency is not a clear indication of its popularity, I though it best to measure two axis of reference rather than just one.



It was then decided that in order for selection it would be best to visually evaluate the rating and frequency side by side the objective is to further identify whether the frequency of a type had any correlation to its rating.



The above graph confirm that there is very real trend with frequency and rating, the types with increased frequency that still show a high mean rating is indicative that the type is popular, we also see that there is no downward trend as the frequency of restaurants decreases. There seems to be a strong set of outliers in cases where the type of Restaurant frequency is low, I chose to exclude these as possible options as they serve a niche segment of the market, although there ranking are high the low volume of these store would indicate that there isn’t a strong basis of comparison for the success that could be achieved by tackling a niche segment.

1. **Clustering of Suburbs**
   1. **Criteria for clustering**

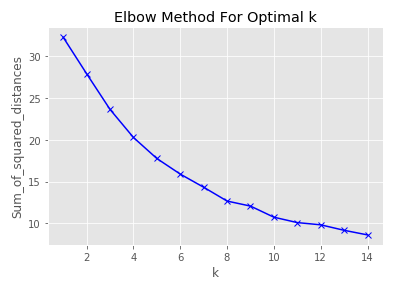
The clustering of suburbs needs to be such that it would be easy to identify potential locations for the selected restaurant type, identifying the top 5 most frequently found restaurant types in a specific area would be indicative of the makeup of available options.

Based on this criteria suburb’s clustered together whose makeup did not include the selected type or whose selected type fell into a very low position on its occurrence.

* 1. **Clustering Problems and Solution**

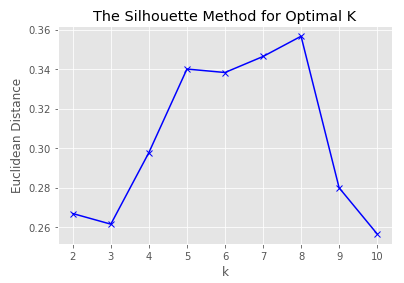
The main issue with the clustering method is identifying the most efficient number of clusters to use to categorize and group your data set, initially It was deemed best to use the elbow method to identify the best *K*.

The Elbow method: an analysis for *K* = 1 to 14 was generated, as per the methodology on should follow with the elbow method a clear distinct change in the graphs trajectory to a horizontal movement is often deemed to be the ‘elbow’ indicating the most viable *K* for the model being tested;



As on can see with the Elbow method, a distinct issue occurs – no discernible elbow is found the viable range could generally be expressed within the range of K = 6 to 10, but nothing is distinct enough to be used.

The Silhouette method: due to the unsuccessful usage of the elbow method this method would then be the next most likely method to reliably determine the correct *K* to use, using the method we would look for the peak in the Euclidean distance to identify the best *K* for your K-means clustering.



* 1. **Final Clusters**

The resulting clusters were as follows:

***Cluster 1: Mixed cluster***

This cluster seemed to be an outlier with no matching suburbs



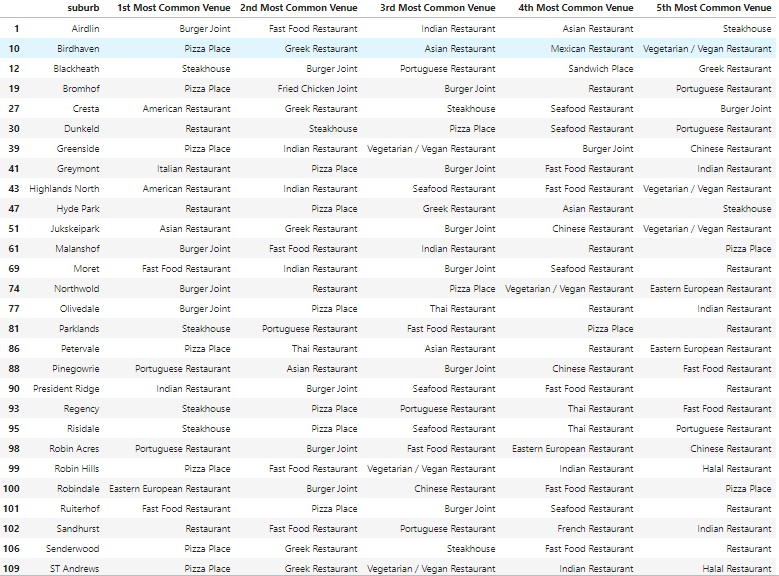
***Cluster 2: Portuguese and Chicken***

This Cluster can be classified as having a heavy presence of Portuguese and Fried chicken Restaurants.



**Cluster 3: High Density Entertainment Zone**

This has grouped the high density entertainment suburbs, we can see that the top rated and most frequently found restaurants in our analysis are all present as top 5’s in the below areas, most instances contain multiple instances of the high rated restaurants,



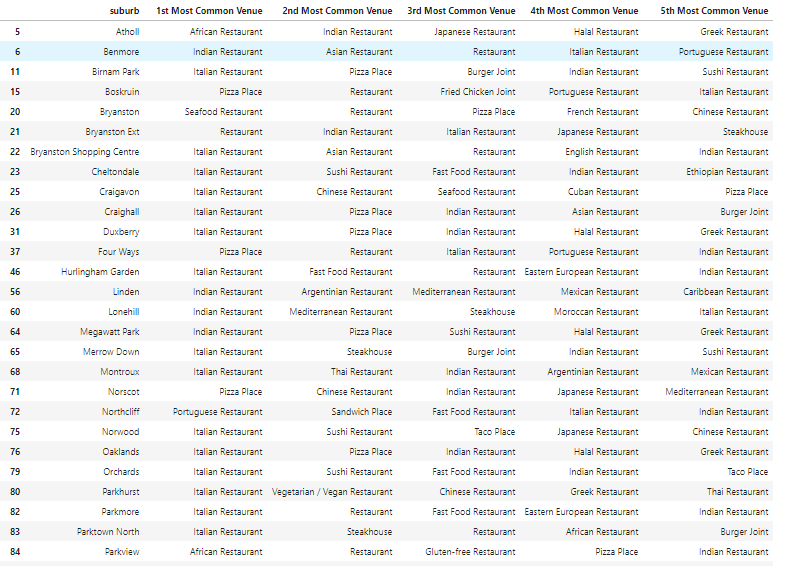
**Cluster 4: Generic & Vegan/Vegetarian Restaurants**

This cluster has a high focus on generic family restaurants and vegetarian places,



**Cluster 5: Italian, Pizza Places & Indian**

This cluster has a strong presence of the Italian and Pizza Places with a small secondary weighting of Indian food, closer analysis showed that in most cases the frequency of restaurants found to be 4th or 5th most common in the suburb did not appear more than twice.



**Cluster 6: Pizza and vegan/vegetarian**

This cluster surprisingly follows a very distinct pattern of occurrences of restaurants despite being separated by a fair distance.



**Cluster 7: Italian and Asian foods**

This cluster includes 1 outlier that does not seem to match the general format for the other suburbs, but in either case a strong Italian, Japanese and Indian restaurant presence can be found in these suburbs



**Cluster 8: Steakhouse**

This Cluster has a strong presence of steakhouses.



1. **Conclusions**

Based on the clustering and the top rated / frequently found restaurants and working off of the premise that a strongly rated restaurant type will be success given low levels of competitors then the below recommendations can be made for the top 3 frequently found restraunts;

### **Opening an Italian restaurant:**

The recommended clusters in descending order are as follows

**Recommend 1 - Cluster 4: Generic & Vegan/Vegetarian Restaurants**

This cluster has a high focus on generic family restaurants and vegetarian places,



***Recommend 2 - Cluster 2: Portuguese and Chicken***

This Cluster can be classified as having a heavy presence of Portuguese and Fried chicken Restaurants.



**Recommend 3 - Cluster 8: Steakhouse**

This Cluster has a strong presence of steakhouses.



### **Opening a Pizza Place**

The recommended clusters in descending order are as follows

**Recommend 1 - Cluster 7: Italian and Asian foods**

This cluster includes 1 outlier that does not seem to match the general format for the other suburbs, but in either case a strong Italian, Japanese and Indian restaurant presence can be found in these suburbs



**Recommend 2 - Cluster 4: Generic & Vegan/Vegetarian Restaurants**

This cluster has a high focus on generic family restaurants and vegetarian places,



**Recommend 3 - Cluster 8: Steakhouse**

This Cluster has a strong presence of steakhouses.

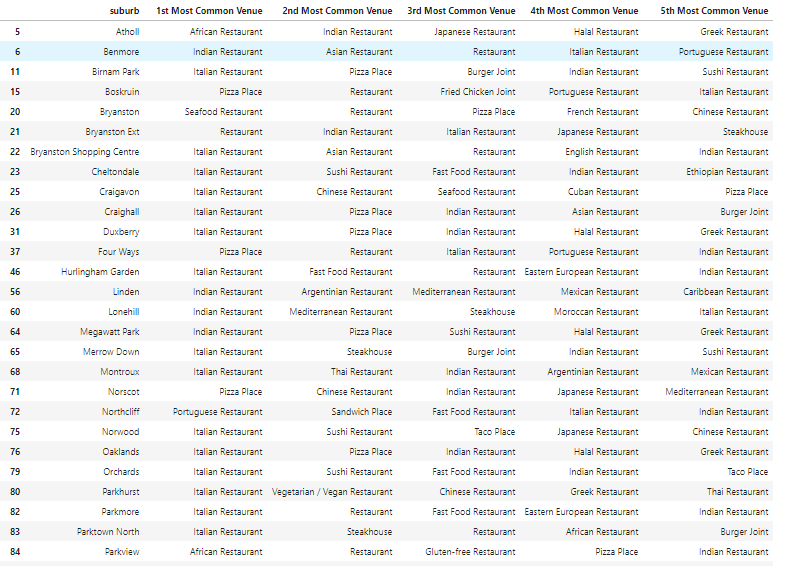


### **Opening a generic Family Restaurant**

The recommended clusters in descending order are as follows

**Recommend 1 - Cluster 5: Italian, Pizza Places & Indian**

This cluster has a strong presence of the Italian and Pizza Places with a small secondary weighting of Indian food, closer analysis showed that in most cases the frequency of restaurants found to be 4th or 5th most common in the suburb did not appear more than twice.



**Recommend 2 - Cluster 7: Italian and Asian foods**

This cluster includes 1 outlier that does not seem to match the general format for the other suburbs, but in either case a strong Italian, Japanese and Indian restaurant presence can be found in these suburbs



**Recommend 3 - Cluster 8: Steakhouse**

This Cluster has a strong presence of steakhouses.



1. **Final thoughts and Future considerations**

On this project I took a different route than just comparing neighbourhoods and using that to recommend where a restraint should be opened mainly based on what is currently popular and where these places would experience the least amount of competition. This approach is sound in theory but a few things fall short due to the nature of the data I have available.

Basing a restraints popularity on its frequency only serves to prove that specifics types of restaurants can be found in abundance it’s not indicative in popularity, also using the ratings for each type although aggregated, its more indicative of the quality of the restraint type found in my city, it’s not indicative that a new restraint of that type will be highly rated but for the purposes of analysis in this capstone I think it helped guide the analysis.

In future I would hope that it could be possible to get further details on the restaurant such as its grade (Low cost, High end), this would better help separate out the demographic that a restaurant serves, it would be interesting to see if high-end restaurants have a different rating or frequency. Also I think it would be helpful to have further details on the makeup of each suburb. I attempted to narrow down the list by separating the CBD and commercial suburbs out from more residential ones, however even the remaining suburbs would contain a level of commercial vs residential areas, being able to quantify this might help better narrow down viable restraint areas. This in conjunction with restaurants grading and the inclusion of earning data for each suburb could help enhance the usability of the report.

# Appendices

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Appendices A - Categories dropped from foursquare venue List** | | | | |
| Airport Terminal | Comedy Club | Gay Bar | Moving Target | Shopping Plaza |
| Art Gallery | Concert Hall | Gift Shop | Multiplex | Skating Rink |
| Arts & Crafts Store | Construction & Landscaping | Golf Course | Museum | Ski Area |
| Arts & Entertainment | Convenience Store | Golf Driving Range | Music Store | Soccer Field |
| Athletics & Sports | Convention Centre | Gourmet Shop | Music Venue | Soccer Stadium |
| Auto Garage | Cosmetics Shop | Grocery Store | Nail Salon | Spa |
| Automotive Shop | Cricket Ground | Gym | Nightclub | Sporting Goods Shop |
| Baby Store | Cupcake Shop | Gym / Fitness Centre | Noodle House | Sports Bar |
| Bakery | Dance Studio | Gym Pool | Office | Sports Club |
| Bar | Deli / Bodega | Hardware Store | Optical Shop | Stables |
| Baseball Field | Department Store | Health Food Store | Other Nightlife | Supermarket |
| Basketball Court | Dessert Shop | Hobby Shop | Paper / Office Supplies Store | Tea Room |
| Bed & Breakfast | Diner | Hostel | Park | Tennis Court |
| Beer Garden | Dog Run | Hotel | Pet Store | Theatre |
| Bike Shop | Donut Shop | Hotel Bar | Pharmacy | Theme Park |
| Boarding House | Electronics Store | Housing Development | Platform | Toy / Game Store |
| Bookstore | Event Space | Ice Cream Shop | Playground | Track Stadium |
| Boutique | Farm | Indie Movie Theatre | Plaza | Trail |
| Bowling Alley | Farmers Market | Indie Theatre | Pool | Train Station |
| Breakfast Spot | Fish Market | Juice Bar | Pool Hall | Video Game Store |
| Brewery | Flea Market | Lake | Pub | Video Store |
| Building | Flower Shop | Lawyer | Racetrack | Warehouse Store |
| Business Service | Food | Liquor Store | Radio Station | Whisky Bar |
| Butcher | Food & Drink Shop | Lounge | Record Shop | Wine Bar |
| Cafes© | Frozen Yogurt Shop | Market | Rental Service | Winery |
| Cafeteria | Fruit & Vegetable Store | Martial Arts Dojo | Resort | Women's Store |
| Casino | Furniture / Home Store | Miscellaneous Shop | Road | Yoga Studio |
| Climbing Gym | Garden | Mobile Phone Shop | Rock Climbing Spot | Zoo |
| Clothing Store | Garden Centre | Monument / Landmark | Scenic Lookout |  |
| Cocktail Bar | Gas Station | Motorcycle Shop | Shop & Service |  |
| Coffee Shop | Gastropub | Motorsports Shop | Shopping Mall |  |
|  |  |  |  |  |
|  |  |  |  |  |