Yelp Restaurant Reviews and Toronto Neighborhood profiles

Contents

Description of the problem

Toronto is one of the biggest food cities in the world. We have several top restaurants and some amazing ethnic cuisine due to our multicultural makeup. I am looking to explore the food scene in Toronto using neighborhood profile information from the 2016 census and open data provided by Yelp about restaurants, particularly in the Toronto area. I am looking to identify if there's any trends related to neighborhood profile information (average income level, ethnic makeup, etc.) and distribution of restaurants by price and ratings.

Data Sources

This analysis and the creation of the visualizations uses the following data sources:

Data Set	Description of Data Set	Source	Reference Link
City of Toronto Open Data – Neighborhood Profile Information as of 2016 Census	This is neighborhood profile information provided by the City of Toronto. It includes attributes such as age, income and ethnic profiles of Toronto's 140 planning neighborhoods.	City of Toronto Open Data Catalog	https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#8c732154-5012-9afe-d0cd-ba3ffc813d5a
Yelp Restaurant Data	This is data derived from Yelp's large repository of data which includes reviews, ratings and other information about restaurants. The Toronto subset was pre-selected and augmented with Yelp's API.	Yelp	https://www.yelp.com/dataset
Census 2016 Data	Stats Canada Census 2016 data for the Toronto Metropolitan Area		http://www12.statcan.gc.ca/census-recensement/2016/dppd/prof/details/page.cfm?Lang=E&Geo1=CMACA&Code1=535&Geo2=PR&Code2=01&Data=Count&SearchText=toronto&SearchType=Begins&SearchPR=01&B1=Ethnic%20origin&TABID=1
Toronto Geocodes		Google	https://developers.google.com/maps/documentation/geocoding/intro
Toronto GeoJson information	Neighborhood ESRI Shape Files to determine Toronto Neighborhood borders which were converted to GeoJson using mapshape.org	City of Toronto Open Data Catalog	https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#a45bd45a-ede8-730e-1abc-93105b2c439f

In order to get the data to a usable level for visualizations, derivations had to be done and both manual and programmatic pre-processing which will be further explained in the data -preprocessing techniques section.

Technologies Used

List of Technologies

The following are the technologies being used in the analysis:

1. Python 3.6 with pandas – data transformation and pre-processing
2. Excel – manual data extraction and transformation
3. mapshape.org – ESRI conversion to GeoJSON
4. Folium – Geo-visualizations
5. Google Maps API – Extracting Geocodes and Geo-visualizations
6. D3.js – Creating World Clouds
7. Plot.ly – Bar, Pie, Box Plots and Heatmaps

Challenges

1) By far, the most challenging aspect of this work was data-preparation using Pandas. To tell the story I was looking to tell, it was hard to find the right data and bring it together, especially for the largest datasets. Pandas was excellent for this work but becoming proficient in the framework took some time.
2) Pandas is good for large scale automated data manipulation operations but most of the time it was more efficient to just extract data from csv's in Excel
3) Matplotlib requires too many parameters and Plotly ended up being better and easier to configure for most types of visualizations, except for Geo visualizations. That's where I found and used Folium
4) D3 is too low level, and likely needs an additional layer of abstraction. In general, to be productive in it for advanced visualizations you must heavily rely on example code

Data Pre-processing Techniques

Several data pre-processing techniques were required to find interesting trends or information within the data and to prepare many of the visualizations. They are as follows:

1. Extract Toronto ethnic and income data from the Census 2016 dataset to prepare Toronto level visualizations (this was done using Excel).
2. Convert ESRI Shape files to GeoJSON for use with Folium library using mapshape.org
3. Convert Toronto Neighborhoods names into decimal coordinates using Google Maps API Geolocation feature and append Yelp Restaurant's Toronto Neighborhood based on proximity
4. Extract top neighborhoods with over 100 restaurants for comparative analysis (this was done as it was impractical to create useful plots for all 140 neighborhoods as attributes. Further, neighborhoods with under 100 restaurants seemed to have less statistical significance in the comparison and the selected 21 neighborhoods served as a diverse enough sample in terms of income and ethnic profiles)

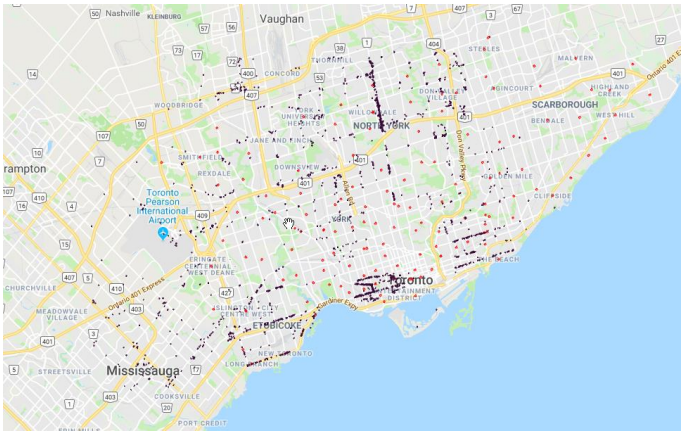
5. Apply filtering on the Neighborhood Profiles dataset to find ethnic and income information per neighborhood.
6. Apply filtering on the Yelp dataset for the selected 21 neighborhoods to create price and rating pie charts
7. Apply pivoting and aggregations to count and find means for the heatmaps

Visualizations

List of Visualizations and what they represent

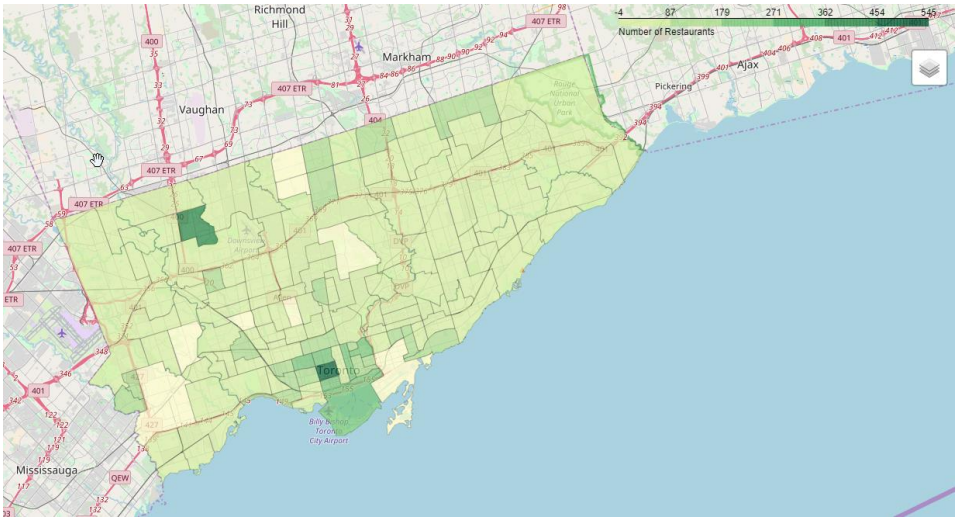
Scattermap

This is a scattermap of Toronto Neighborhoods (red points) and Yelp Restaurants (purple points) showing where the centroid of all 140 Toronto Neighborhoods are, and where the Yelp restaurants from the dataset are located. This was created using Google Maps API.



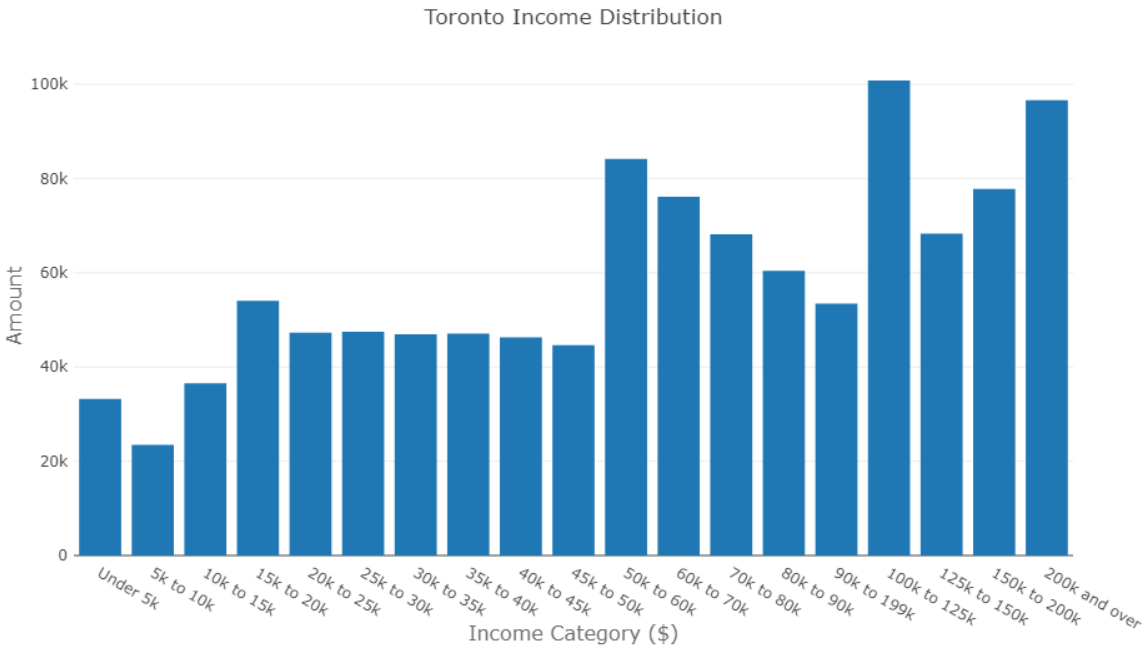
Choropleth of Toronto Neighborhoods and Yelp Restaurants.

In this choropleth, the darker polygons represent neighborhoods with more restaurants from the Yelp dataset. The two densest restaurant neighborhoods are Kensington-Chinatown and Glenfield-Jane Heights.



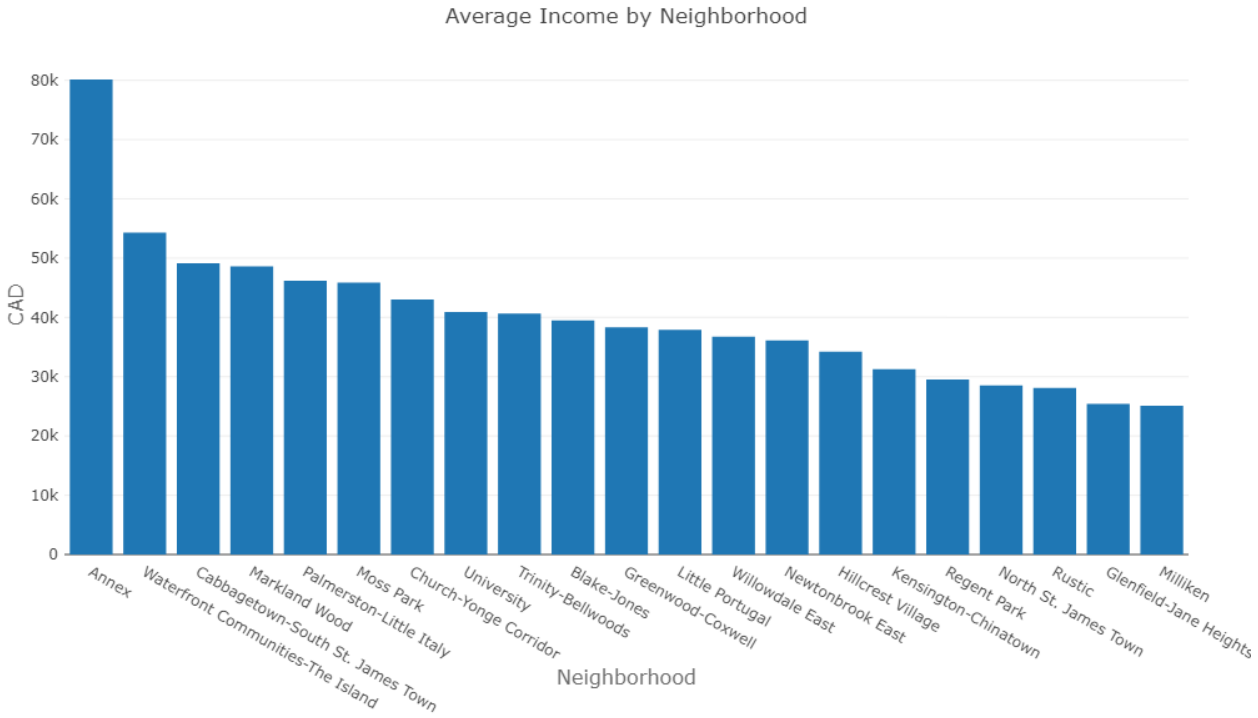
Histogram for Income distribution for Toronto

This visualization is showing the income distribution of Torontonians from the 2016 Census data across all income categories/brackets. Over ½ of the city is making over $50,000 per year.



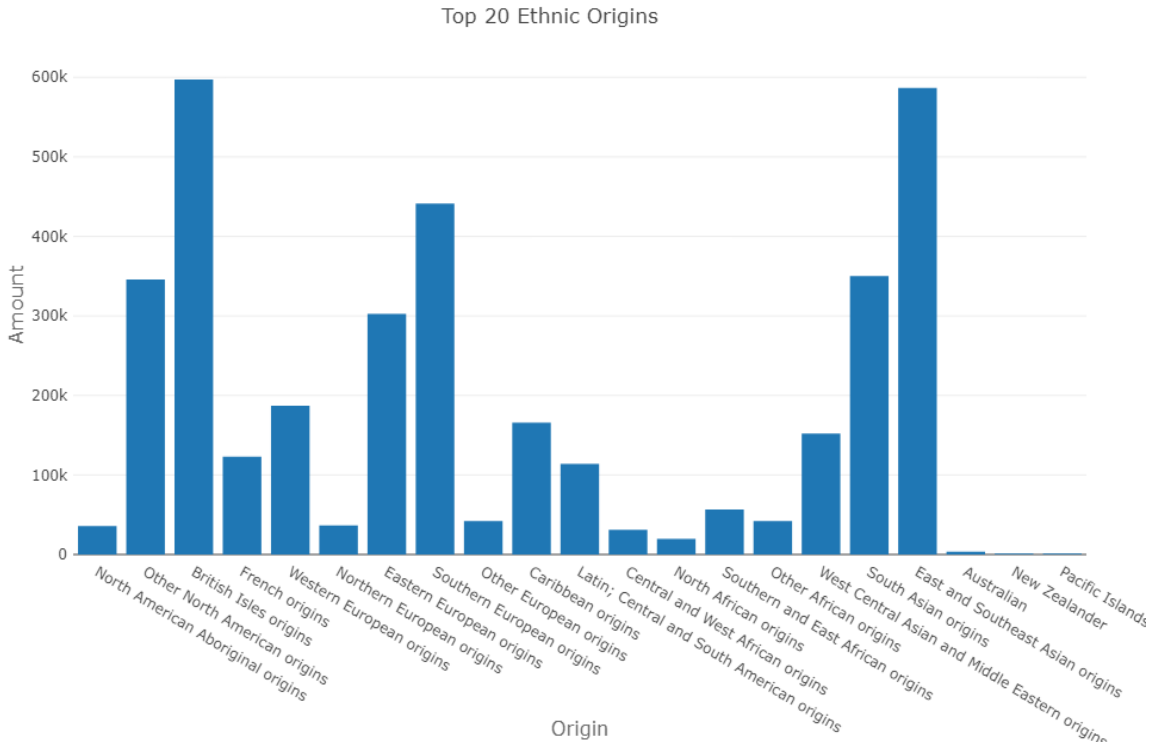
Bar chart for Average Income for Neighborhoods with greater than 100 restaurants

This visualization is showing average income of neighborhoods with more than 100 Yelp restaurants. As can be seen, the Annex is a clear leader here.



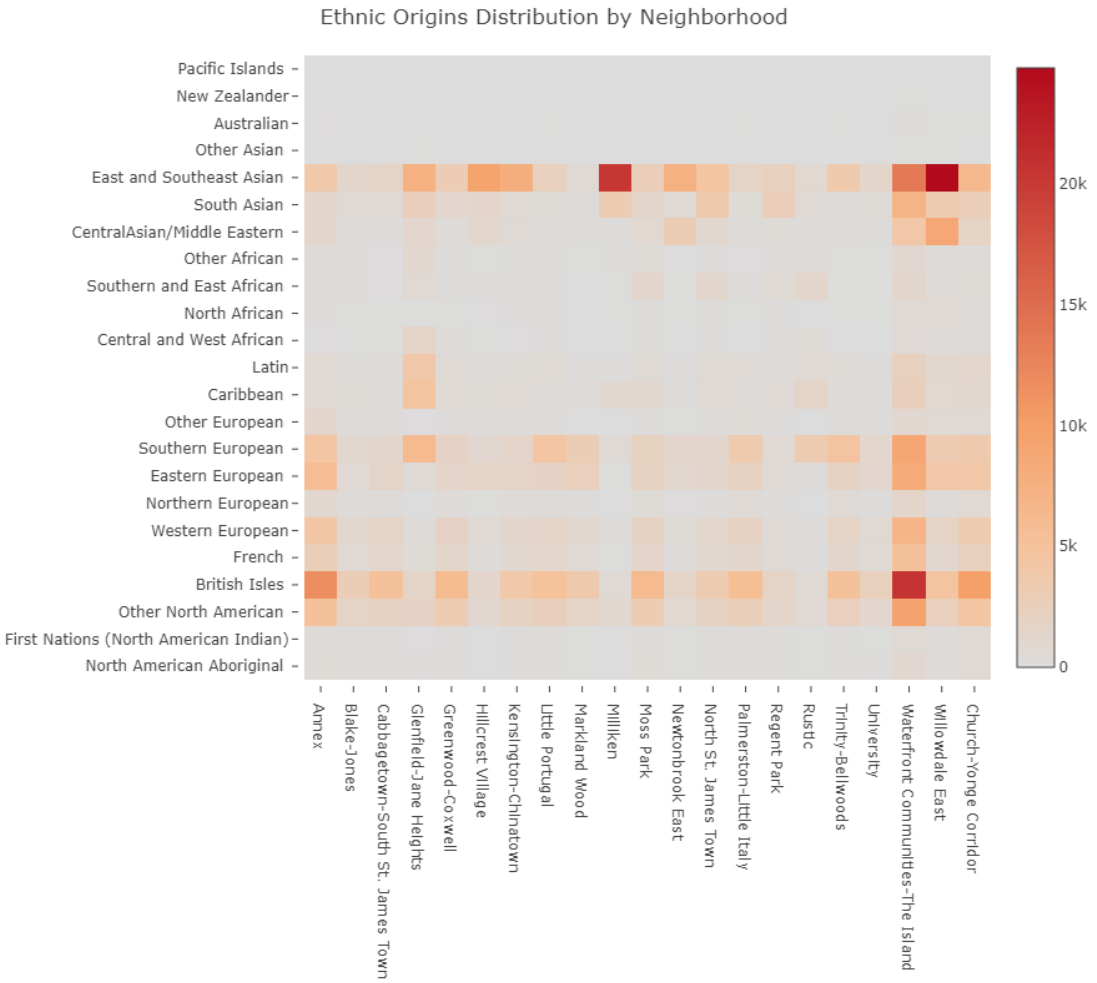
Bar chart for Ethnic Origins distribution for Toronto

This visualization is showing the the ethnic origins of all Torontonians from the 2016 Census data. This data is interesting, because people can have more than one response, and thus the total is greater than the population, but it still gives an interesting look at the distribution of ethnic origins. As ethnicity and food are closely related, I assumed there might be connection. The largest ethnic origins by far are East/Southeast Asian and British Isles.



Heatmap for Ethnic Origins distribution by Neighborhood for Neighborhoods with greater than 100 restaurants

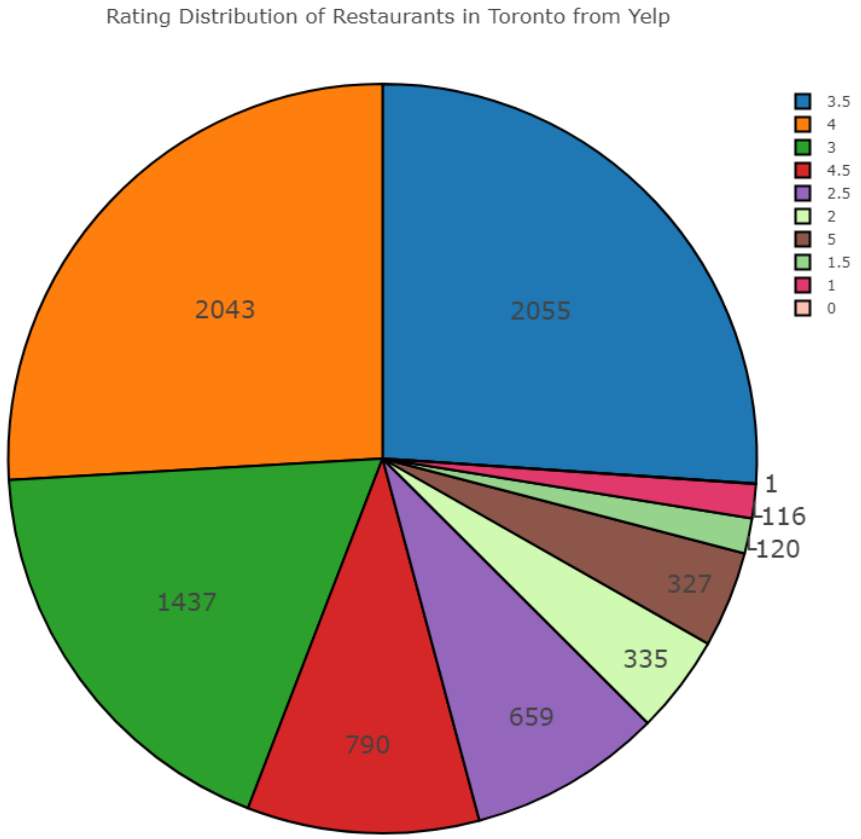
This heatmap shows us ethnic origins distribution by neighborhood. Some neighborhoods certainly have prevalent ethnicities, and this should in theory reflect in the restaurant category choices available as restauranteurs will cater to their customers



.

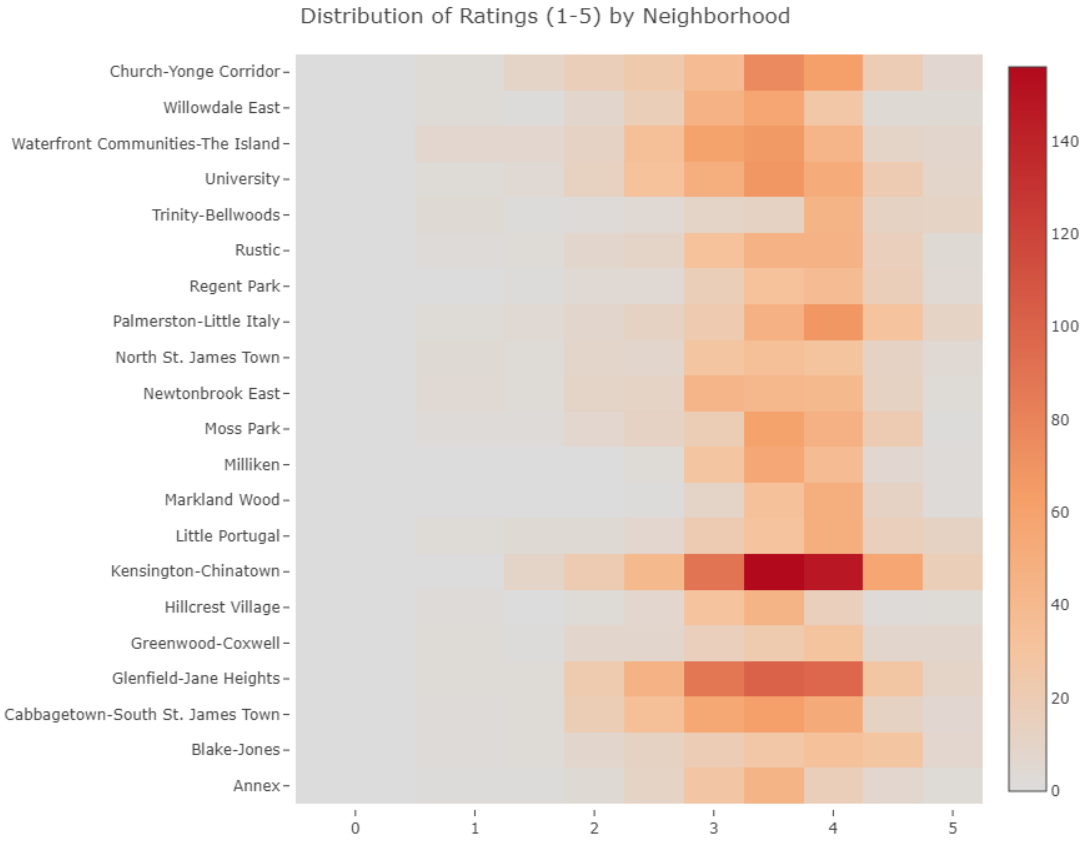
Pie chart for breakdown of overall Yelp ratings for Toronto restaurants

This is the overall ratings distribution/breakdown for Toronto for all Yelp restaurants in the dataset. 3.5-4 ratings are dominant which is expected.



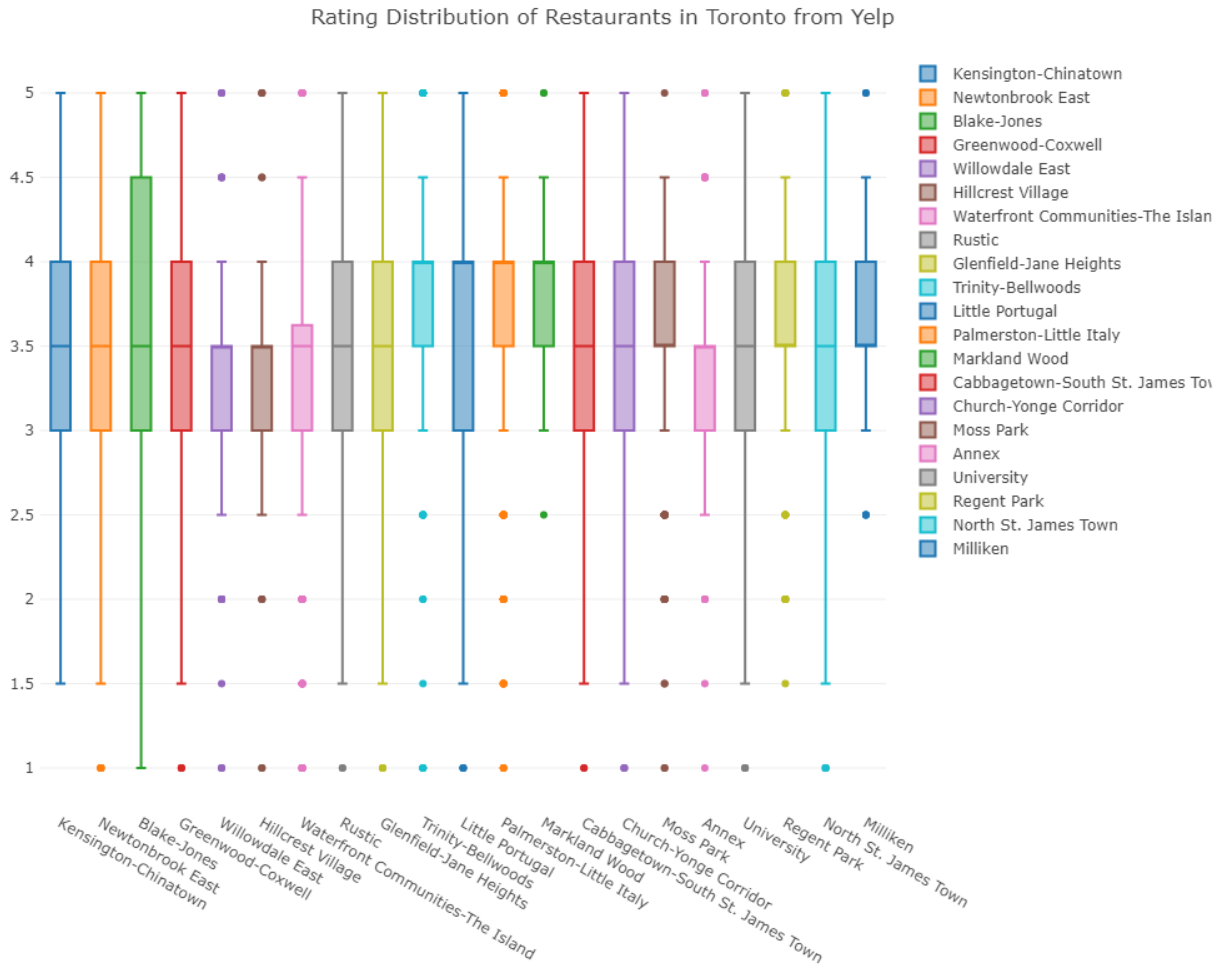
Heatmap for Rating by Neighborhoods for Neighborhoods with greater than 100 restaurants

This heatmap is showing us how the ratings are distributed amongst the 21 selected neighborhoods. The results here aren't terribly surprising, as the bulk of restaurants in the dataset are in Kensington-Chinatown and Glenfield-Jane Heights. What is interesting is how most ratings for all neighborhoods are in the middle (3-4).



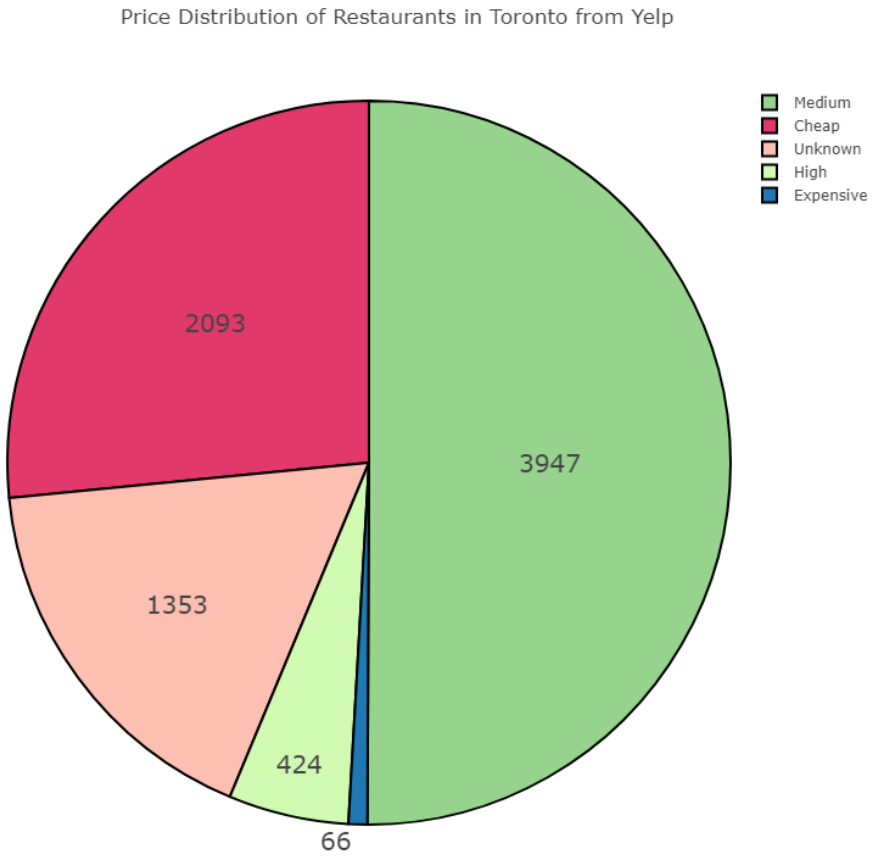
Boxplot for ratings distribution for Neighborhoods with greater than 100 restaurants

This boxplot is meant to show the statistical distribution of ratings for each of the selected neighborhoods. This view of the data is meant to normalize the distribution instead of the heatmap which will pronounce neighborhoods with more restaurants overall.



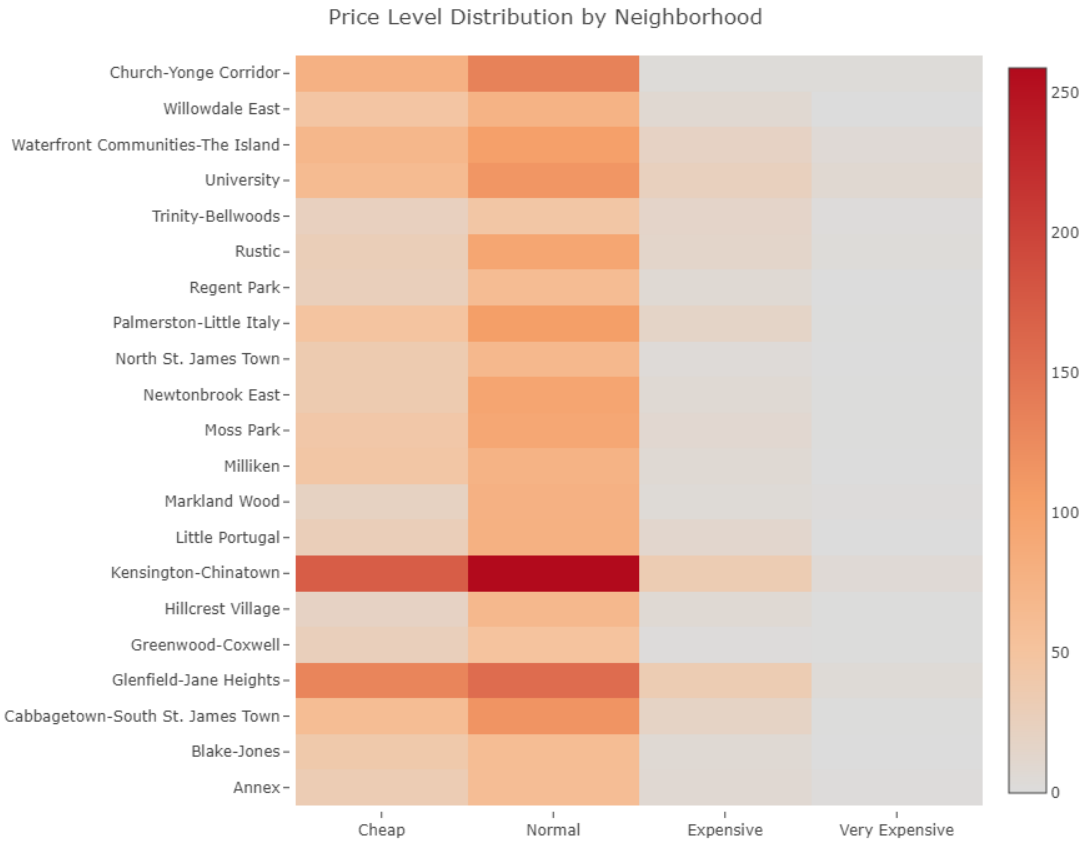
Pie chart for breakdown of overall Yelp price levels for Toronto restaurants

This is the overall price level distribution/breakdown in Toronto for all the Yelp restaurants. Not surprisingly, the bulk of the restaurants are in the “Medium” or “$$” price range.



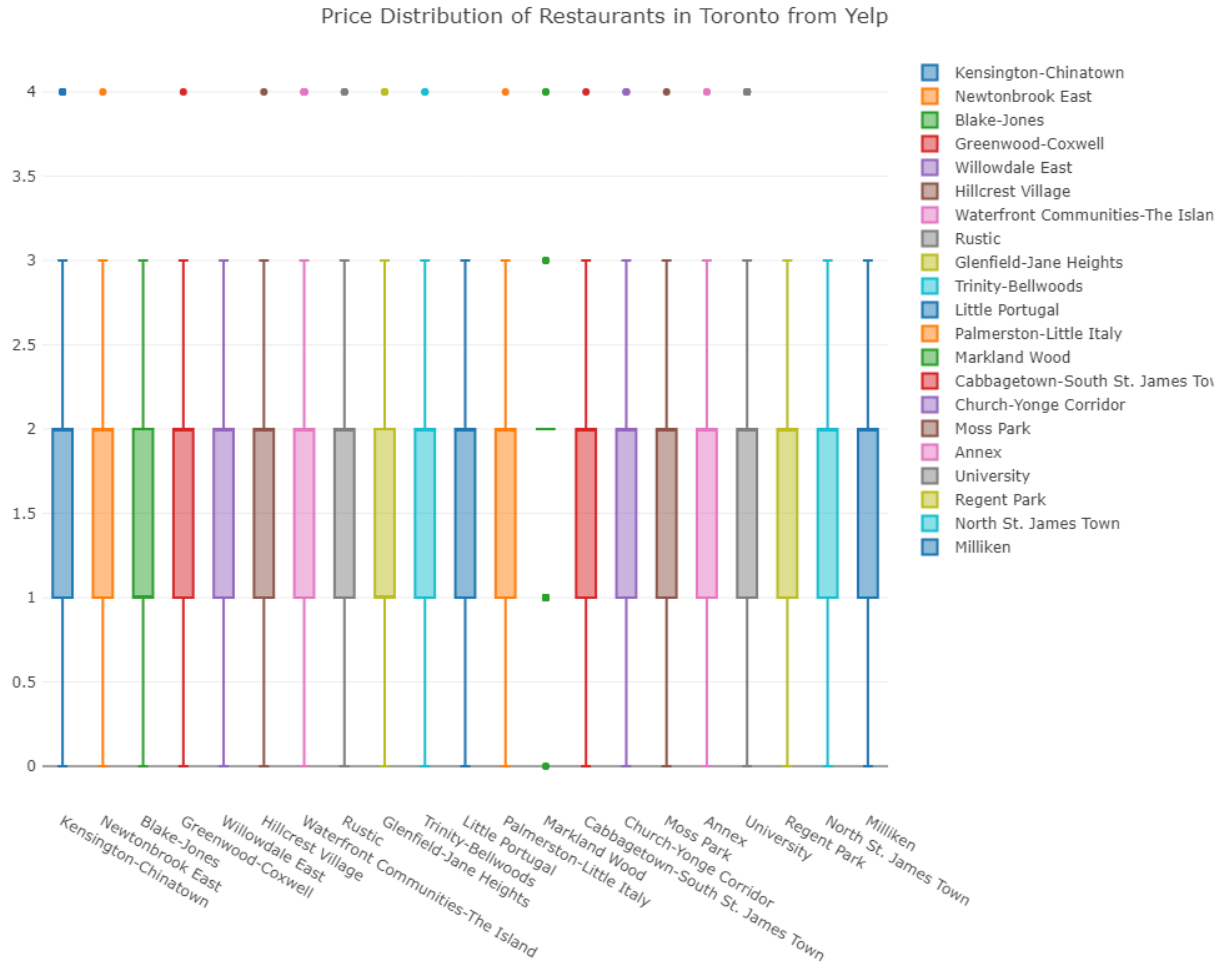
Heatmap for Price Level by Neighborhoods for Neighborhoods with greater than 100 restaurants

Similarly, to the ratings heatmap, this one is showing us the price distribution by the selected neighborhoods. Again, Kensington-Chinatown and Glenfield-Jane heights are clear leaders, and again the “Normal” or middle category is dominant for all neighborhoods.



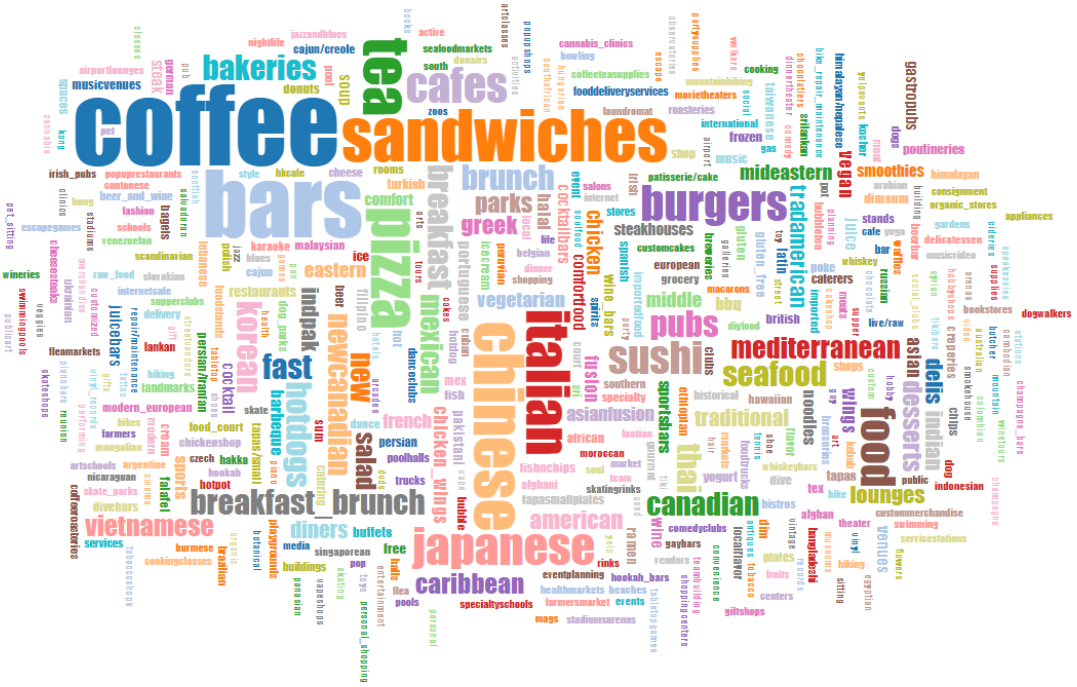
Boxplot for price distribution for Neighborhoods with greater than 100 restaurants

This is a boxplot showing the price distribution for the selected neighborhoods, serving as a similar purpose of normalization as with the ratings boxplot. This plot came as a bit of a surprise since the statistical distribution is almost equivalent for all the neighborhoods except for Markland Wood which is a bit different than the Rating boxplot.



Word clouds showing most popular Toronto Restaurant categories

This word cloud is showing the dominant categories for all restaurants in the Yelp dataset for Toronto. It's clear coffee and sandwiches are the dominant categories. It's clear that Chinese/Japanese and Italian food is also quite dominant which falls in line with the ethnic profile of Toronto.



Word clouds for Kensington-Chinatown vs. Glenfield-Jane Heights (two neighborhoods with greatest number of restaurants)

These are the world clouds for the above-mentioned neighborhoods for their categories. These two were chosen as they have the largest concentration of Yelp restaurants.

Kensignton-Chinatown

Glenfield-Jane Heights

Conclusions

1) Within the Yelp dataset, Kensington-Chinatown and Glenfield-Jane Heights are the two neighborhoods with the most restaurants
2) Within the Yelp dataset, there are 21 neighborhoods with over 100 restaurants
3) Toronto has a big income distribution but over half the population makes 50k per year or more, with over 100k people making over 100k per year
4) Some neighborhoods are wealthier by a wide gap, but out of the neighborhoods with 100 restaurants or more, the wealthiest neighborhood is the Annex with an average income of just under 80k per year
5) Toronto is quite multicultural and diverse, but the leading ethnic origin is East Asian, specifically Chinese, followed closely by British Isles origins
6) Of the Neighborhoods with over 100 Yelp restaurants, the Waterfront is heavily dominated by British Isles origins, whereas Kensington-China Town, Willowdale East, Milliken and Glenfield Jane Heights are dominated by East Asian origin
7) Overall, Toronto Yelp restaurants are rated in the middle (3.5-4)
8) Kensington-Chinatown has the most restaurants, therefore the highest concentration of 3.5-4 ratings
9) Most neighborhoods have a median rating of 3, but many outliers. Trinity-Bellwoods has the highest median rating at 4. Blake Jones has the largest distribution of ratings.
10) Overall, Toronto Yelp restaurants are dominated by Medium ($$) priced restaurants
11) Most neighborhoods have few if any Very Expensive ($$$$) restaurants. Most are Medium($$). Again, Kensington-Chinatown has the most restaurants overall and therefore has the most Medium($$) restaurants.
12) 20 out of the 21 neighborhoods with over 100 Yelp restaurants are evenly distributed in price statistics, except for Markland-Wood which overall has more restaurants in the 2 or above price level.
13) Most popular restaurant type/category is…Coffee & Sandwiches closely followed by Chinese/Japanese/Italian food.

Lessons Learned

Originally, I was intending on doing an analysis of Toronto price homes against income, but throughout my exploration of the datasets, I quickly came to the realization that this type of analysis requires much more time, effort and a strong background in economics. Further, I found it incredibly challenging to find the right data to answer that question.

Thus, I utilized some of the findings in some of the datasets I downloaded and became curious about Toronto's food scene. In particular I became interested in our neighborhood statistics and wanted to see if there were any noticeable trends.

Some of the lessons from this experience are:

1) Finding good quality data is important in solving any problem
2) Be flexible. I allowed the data to tell me it's story a decided to share it with the others.
3) Data can serve more than one purpose
4) Visualizing a large number of attributes/variables is decidedly quite hard to do effectively
5) Pre-processing is paramount to preparing the data to be visualized, and almost always the hardest part