Christopher Yang
Sameer Zubairi
Alan Deng
Jacob Evans

Project 1 Analysis
Feb 14, 2023

Toronto vs Vancouver Restaurants Analysis

Using the Yelp API and Geoapify API, data was extracted to analyze the restaurants in Toronto vs Vancouver. Analysis was made using the data of restaurant ratings, proximity to busier areas, restaurant preferences, types of restaurant, and hours of operation.

Usage of Yelp Fusion API for Base Dataset Creation:

Using Yelp Fusion API, pull a sample population of 1000 restaurants from Toronto and Vancouver. This is done using the Yelp Fusion API python example repository, which defined the use of a function get_businesses. Inputs used consisted of Toronto and Vancouver for location, Restaurant for queries, and use of individual api_keys. This resulted in two .CSV outputs that were used by the team for individual analysis.

The Yelp API only allows for a 50 limit search at time, requiring the use of an offset to continue collecting data. The API also has a total query cap of 1000 at a time before returning an error.

Usage of GeoApify API:

The GeoApify API was used as a secondary source for answering more data from the Yelp API. This was done by using the latitude and longitude data of each restaurant, previously extracted from Yelp, and entering into the GeoApify API to return driving directions and distance using GeoApify's "middle optimised" driving algorithm. This was done by using iterrows() in Pandas to retrieve route information for each restaurant. Only the distance portion of the returned API output is used for further analysis.

Question: Toronto Vs. Vancouver: Who has Better Restaurants?

The purpose of this analysis is to provide a broad general overview of the "quality" of restaurants available in each city, quality being determined by average reviews within each city.

Taking the average rating for restaurants in each of the Toronto and Vancouver datasets, a histogram can be plotted to highlight the number of restaurants that fall into each ratings category. It should be noted that Yelp's API only returns restaurant ratings in half star increments: there are no "4.8" star restaurants, only 4.5 and 5 star restaurants.

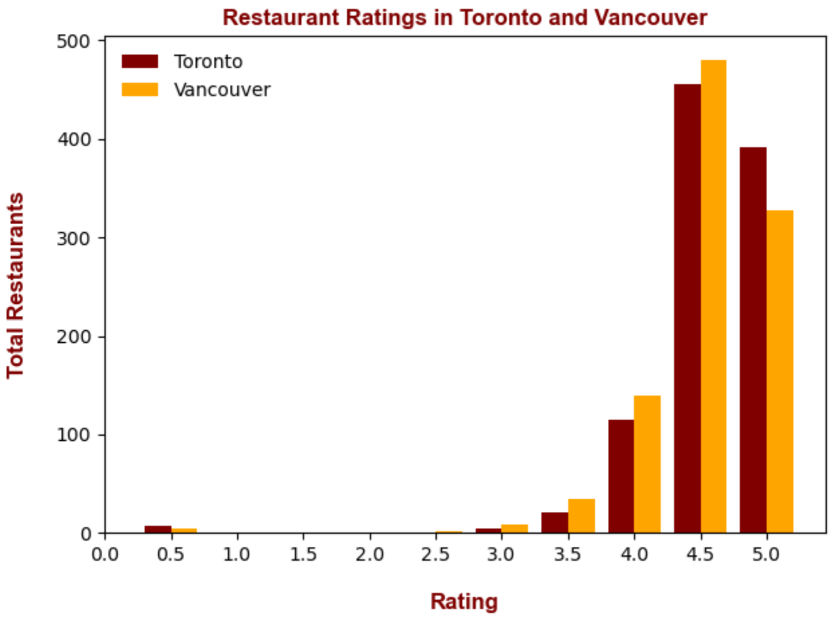


Fig 1. Restaurant Rating in Toronto vs Vancouver

Looking at the histogram, it becomes immediately apparent that the majority of restaurants in both cities fall into the 4 star and up columns. Because of this, the graph is heavily skewed towards the right (better ratings), and indeed performing a normal test (null hypothesis: the distribution shows a null hypothesis) shows that this is the case, with the results shown below of the alpha value being much greater than the p-value rejecting the null hypothesis:

```
Toronto NormaltestResult(statistic=710.5969813838792, pvalue=4.96393474763195e-155)
```

```
Vancouver NormaltestResult(statistic=589.8227297653107, pvalue=8.348752392074616e-129)
```

On first glance, the data seems to imply that a greater share of restaurants in Toronto are 5 stars compared to Vancouver. Vancouver on the other hand has a larger number of 3.5-4.5 star restaurants. Does the higher number of 5 star restaurants in Toronto mean that food quality is better?

This answer requires another look with deeper statistical comparison. To that end, a box and whisker plot is used to look deeper into restaurant ratings.

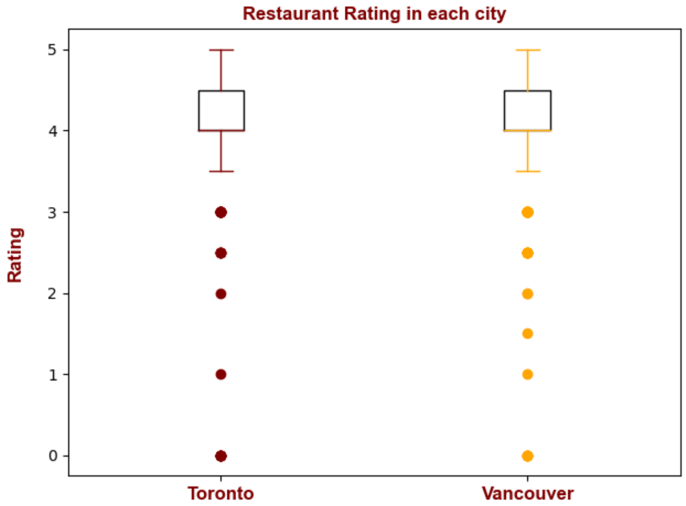


Fig 2. Boxplot of Restaurant Rating in Toronto and Vancouver

Some accompanying statistics with the box and whisker plot include:

For the city of Toronto

- The lower quartile is: 4.0
- The upper quartile is: 4.5
- The interquartile range is: 0.5
- The median is: 4.0
- Values below 3.25 could be outliers.
- Values above 5.25 could be outliers, which is not possible.

For the city of Vancouver

- The lower quartile is: 4.0
- The upper quartile is: 4.5
- The interquartile range is: 0.5
- The median is: 4.0
- Values below 3.25 could be outliers.
- Values above 5.25 could be outliers, which is not possible.

It becomes more apparent when comparing the two box and whisker plots that statistically, both cities perform very similarly as far as rated restaurants go. Both cities have an interquartile range of 4-4.5 with a median rating of 4, and both cities (via the 1.5IQR test) show that restaurants with below 3.5 stars are considered outliers. There is very little statistical difference in restaurant ratings between the two cities.

Questions arise from the nearly identical results shown in the box and whisker plots as well as the non-normalized histograms. It is currently unknown how Yelp chooses which restaurants to return for the given queries (parameters again being “Toronto/Vancouver, Restaurants”). The data is not seemingly randomised as multiple searches of the database seem to return the exact same restaurants. There could be implicit bias in the sample population as it is predicated on what Yelp decides to feature for the queries. For example, the strong ratings shown by all restaurants could be indicative that Yelp is returning either best rated or most popular restaurants within each city. It could also be hypothesised that the sample population is in fact indicative of the total population and that restaurants are generally very good in both Toronto and Vancouver, although this is hard to prove with the given dataset since again multiple queries seem to return the same set of restaurants.

Given the dataset, the answer to the question “Which City has better restaurants” is “given the sample population, both cities offer similarly good restaurants”.

Question: Does Restaurant Distance from a Landmark affect its Rating?

The purpose of this analysis is to determine if restaurant distance from well-known city landmarks affects restaurant rating. Intuitively, one could make the argument that restaurants that are farther away from landmarks have less stature and popularity and therefore be associated with lower ratings.

Obtaining distances requires defining a landmark. While this is subjective to an extent, a reasonable argument can be made that these include the CN Tower for Toronto and Canada Place for Vancouver. Using these landmarks, the driving distance to each restaurant is calculated using GeoApify’s “route” option, and only the driving distance is retained and used for the purposes of this study. It should be noted that in the case of Toronto, the data had to be cleaned with two restaurants excluded because they are located on the Toronto Islands- which you cannot drive to. Plotting the rating vs. Distance for Toronto yields the following:

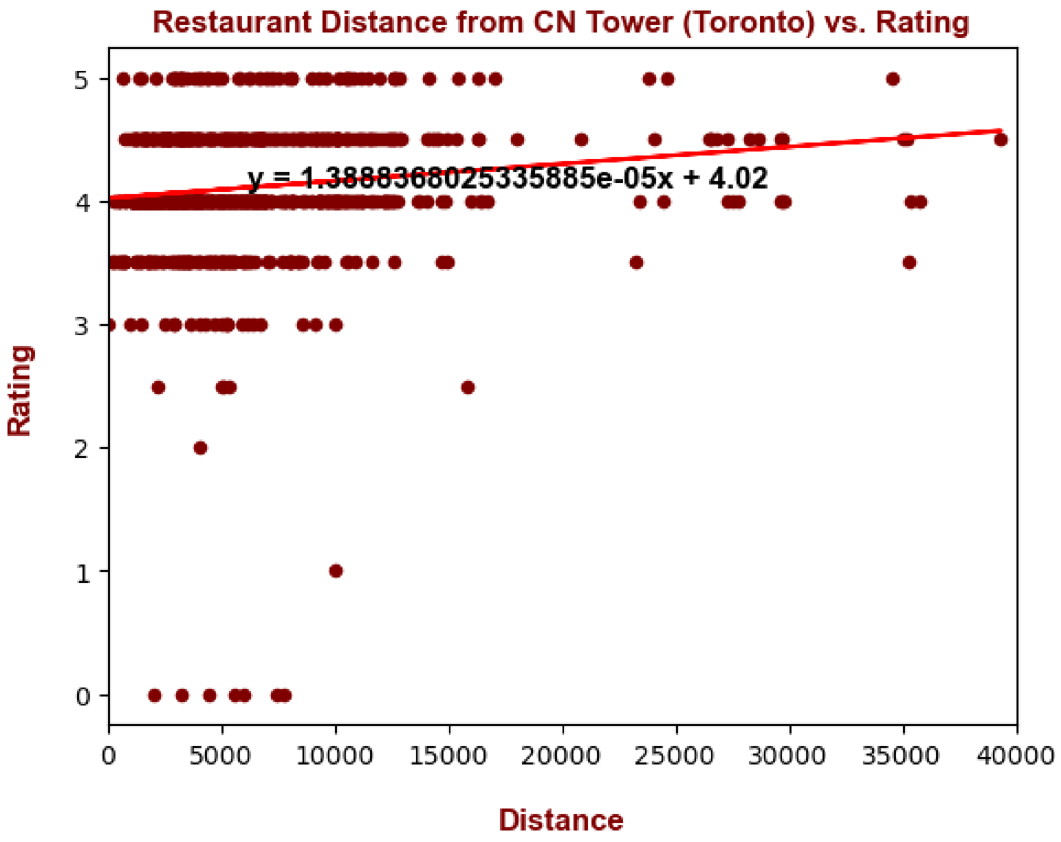


Fig 3. Restaurant Distance from CN Tower vs Rating

There does not appear to be a good correlation between the distance of a restaurant from a landmark vs its rating. Both the R-squared value of the associated linear regression, as well as the Pearson correlation test results are shown below:

```
The r-squared is: 0.016820481368420633
PearsonRResult(statistic=0.12969379849638393, pvalue=3.966754766151683e-05)
```

With the R-squared value being so low, and with the Pearson Correlation coefficient R being much lower than the P-value, it is clear that there is no correlation between restaurant distance from the CN Tower and restaurant rating.

Similarly, when the analysis was performed for Vancouver:

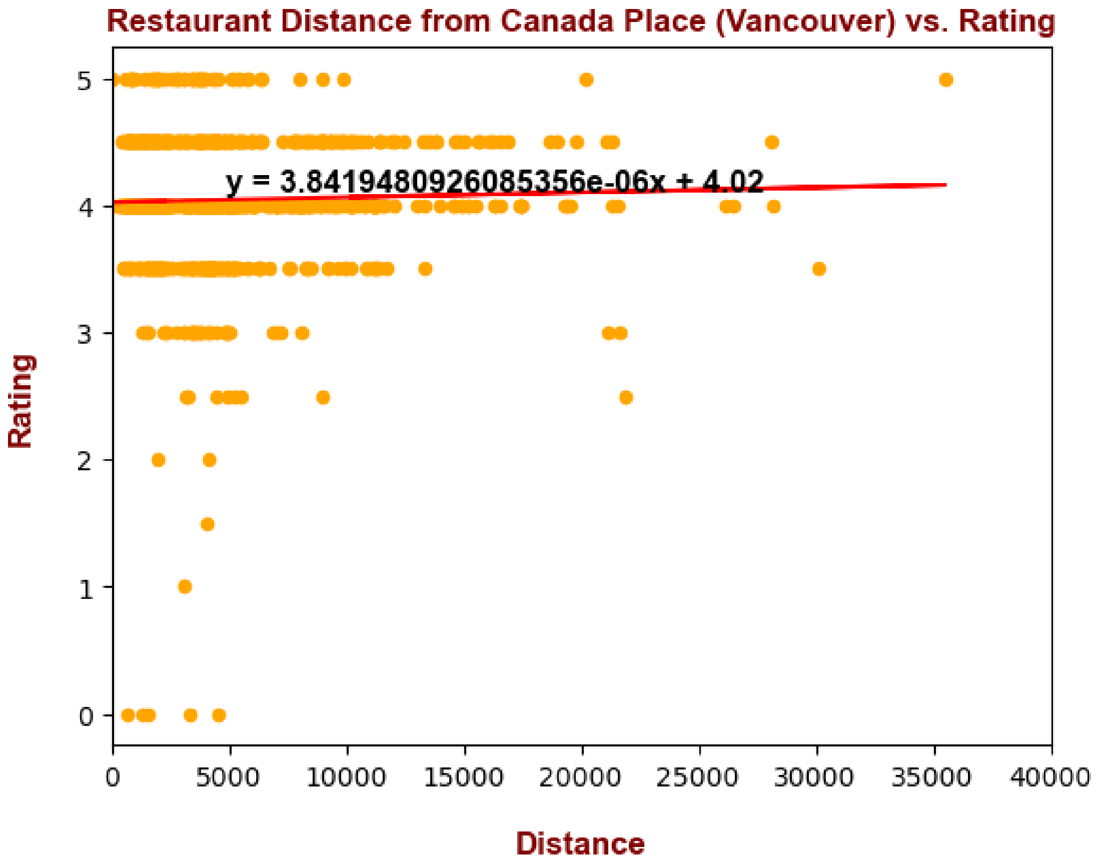


Fig 4. Restaurant Distance from Canada Place vs Rating

The R-squared and Pearson Correlation results:

```
The r-squared is: 0.000881199587707198
PearsonRResult(statistic=0.02968500610926668, pvalue=0.34837038519649094)
```

Show again that there is no correlation between restaurant rating vs. distance from Canada place.

Again, questions arise from this data, based on the method of population sampling from the Yelp API. As before, it is currently unknown how Yelp chooses which restaurants to return for each city. The previous question seems to imply that Yelp might choose to bias sample population data by providing a greater share of popular/highly rated restaurants. What this analysis shows is that if the sample population is biassed, it is not biassed based on restaurant location.

With the given dataset, the answer to the question “Does restaurant distance from well-known city landmarks affect restaurant rating” is “No, restaurant distance does not”.

Question: What type of restaurants do each city have the most of?

After cleaning the data of both Toronto and Vancouver Yelp data, the data was sorted by price range. By doing this, we can see that most of the restaurants in both cities have a price range of two dollar signs($$). After doing some research on the internet, yelp categorises their price range with dollar signs. These dollar signs indicate a price range of the restaurant's menu items: $ = less than 10 dollars, $$ = between 11-30 dollars, $$$ = between 31-60 dollars, and $$$$ = more than 60 dollars.

In both cities most restaurants are in the mid range category of between 11-30 dollars. By cleaning the data even more, the top ten categories of restaurants are used for this analysis. From the data, we can safely say that Vancouver has the most mid-range Japanese restaurants.Where as in Toronto, there are two categories of restaurants, Italian and Japanese.

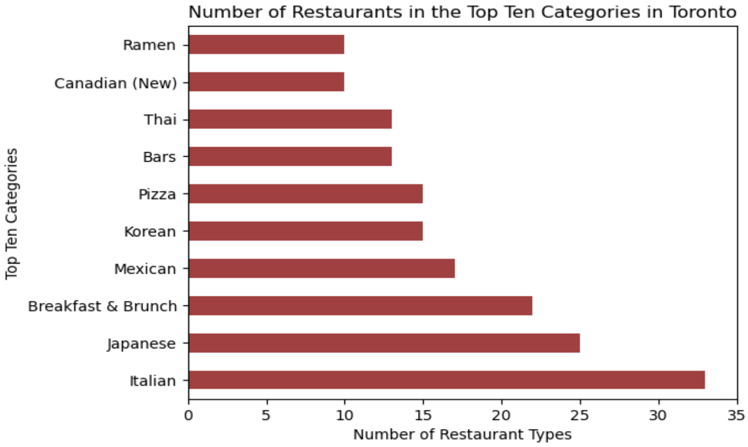


Fig 5. Top Ten Categories of Restaurants in Toronto

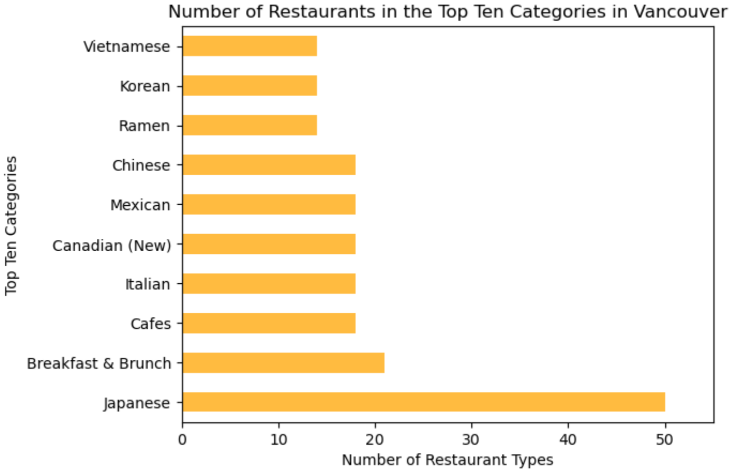


Fig 6. Top Ten Categories of Restaurants in Vancouver

Question: Do the number of reviews impact the ratings of these restaurants?

The null hypothesis: The number of reviews a restaurant has has no impact on the ratings of the restaurant.

From the generated box plots of both cities, we can see that the average number of reviews in Toronto is around 250 and in Vancouver is a little bit below 250. For restaurants that have a rating of 4, both cities have the most amount of reviews within that rating. We can also see that the whiskers of all the boxes within both cities are either in the middle or lower than the median, meaning that most of the restaurants either have around 250 reviews or less. We can also see all the outliers within the data. These outliers that have a lot of reviews and high ratings indicate that these restaurants are very popular within their city. Since the f-test score is low and the p-value is 0.733 which is significantly high, this means that the evidence for the question "Does the number of reviews affect the rating of a restaurant?" is very weak. In conclusion, the number of reviews a restaurant does not affect the rating a restaurant has on Yelp.

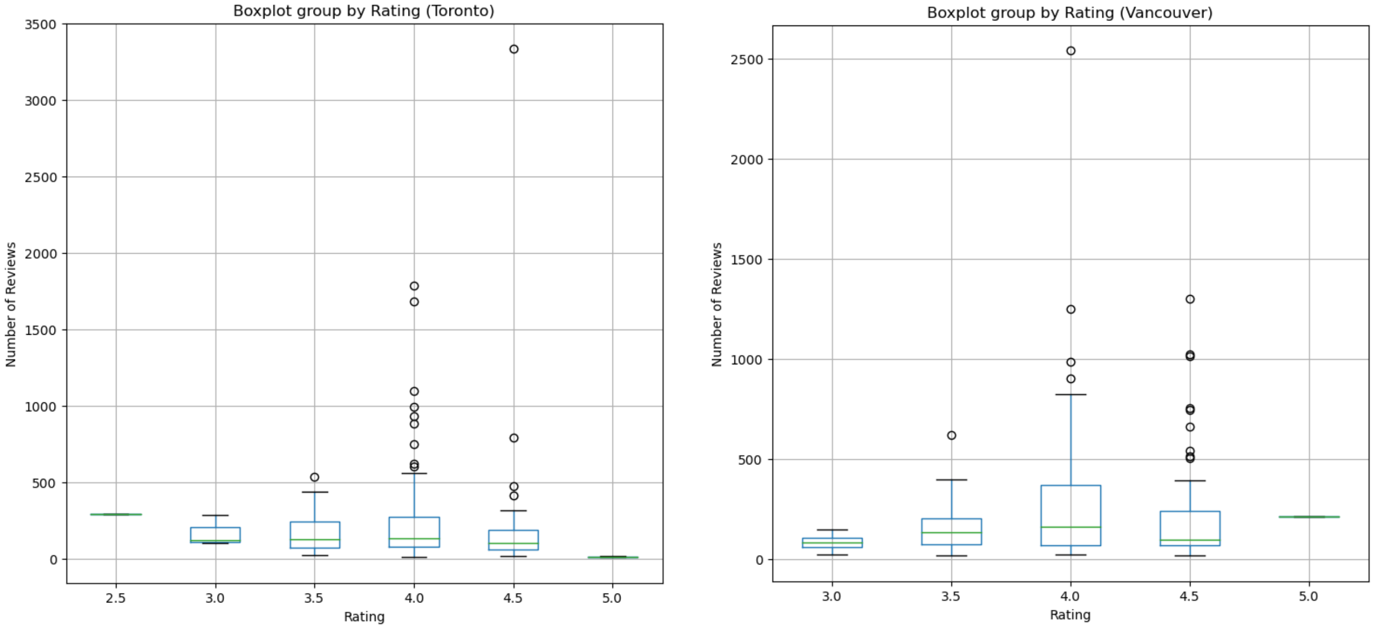
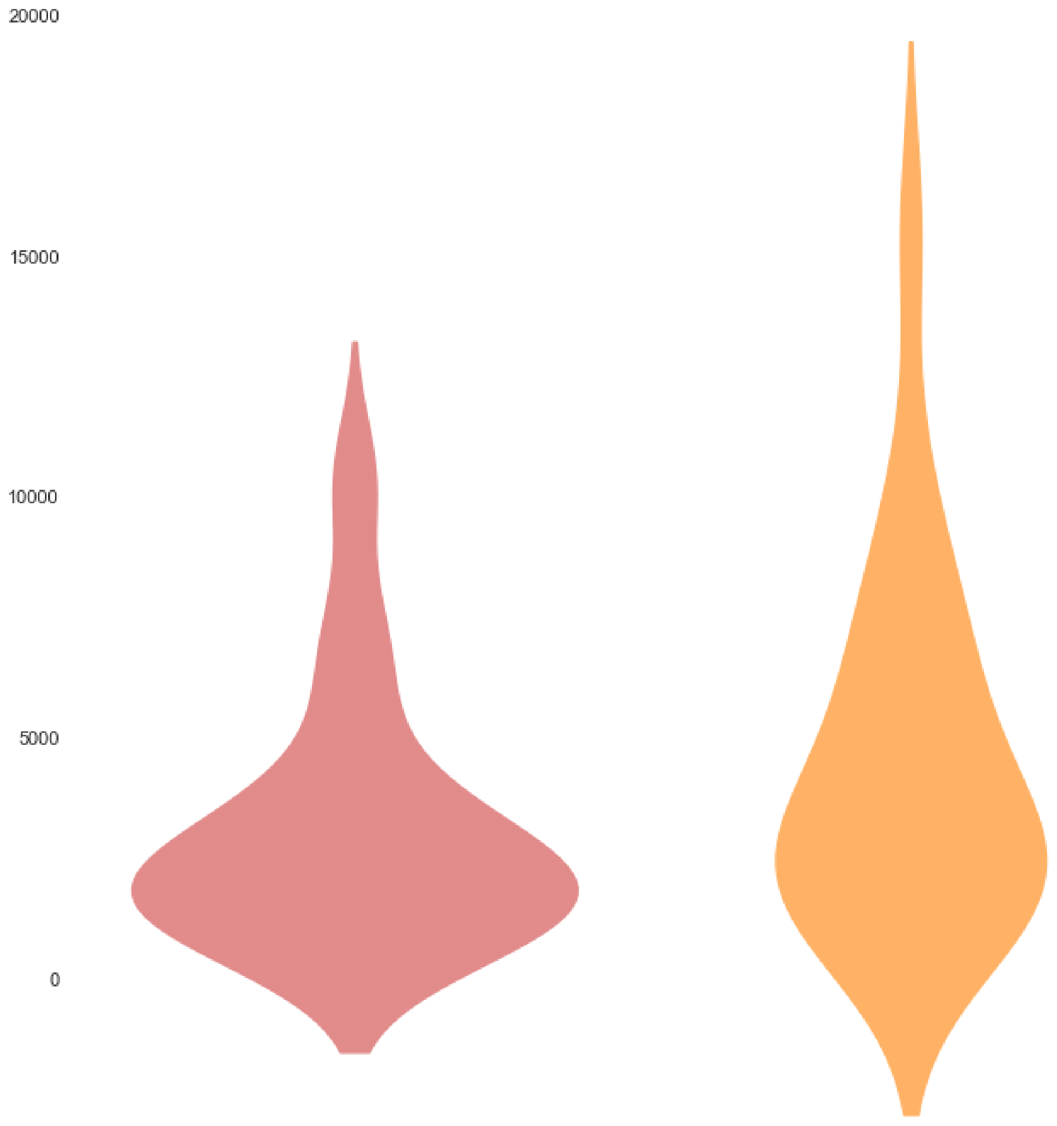


Fig 7. Boxplot Grouped by Rating for Toronto Restaurants (Left) and Boxplot Grouped by Rating for Vancouver Restaurants (Right)

Question: What unique restaurant options exist in each city? From the types of restaurants that exist in both cities, what is preferred in each city?

The purpose of this analysis is to compare a sample of 1000 restaurants located in Toronto to a sample of 1000 restaurants located in Vancouver gathered from Yelp’s Fusion API, and determine any differences in the diversity of options and tastes between the two cities. This was done on a qualitative basis to study the unique options found in either city and a quantitative basis to study the weighted ratings of categories found in both cities.

First off, the number of unique categories in the sample for each city was calculated to be 122 in Toronto and 119 in Vancouver. Then, the distrution of the number of reviews per category with at least a total of 1000 reviews was analyzed as shown in Fig 8 below. The distribution for Vancouver was more spread out than Toronto indicating that Vancouver has categories that are more highly reviewed than Toronto. This could be due to the implicit bias in Yelp's business search API or the possibility that restaurants in Toronto are being categorized more granularly than Vancouver.



Next, the types of restaurants that were unique to each city were analyzed and narrowed down from the original samples. The data was then cleaned up by removing restaurants with less than 75 reviews or belonging to an ambiguous category like 'Food Court' or 'Organic Store'. The data was then geoplotted on the map for each city showing unique restaurants in both Toronto and Vancouver.

The analysis identified a number of restaurants with a rating above 3.5 and at least 75 reviews. These were then geoplotted on a map of each city as shown in fig. 9 and fig. 10 below:

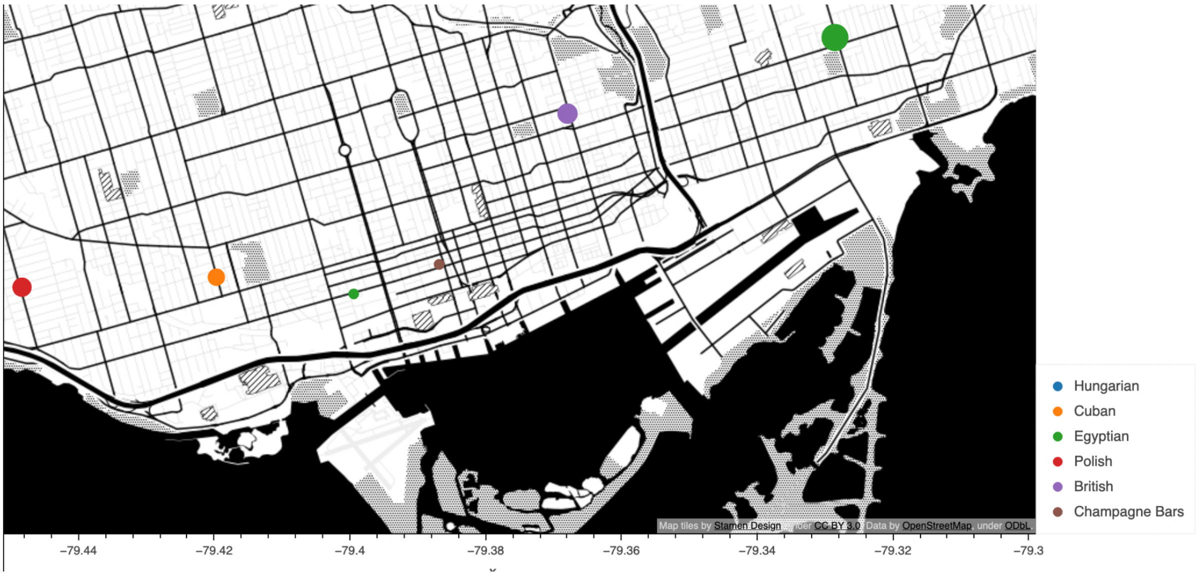


Fig 9. Unique restaurants by type in Toronto. Size of point represents the number of reviews for that restaurant.

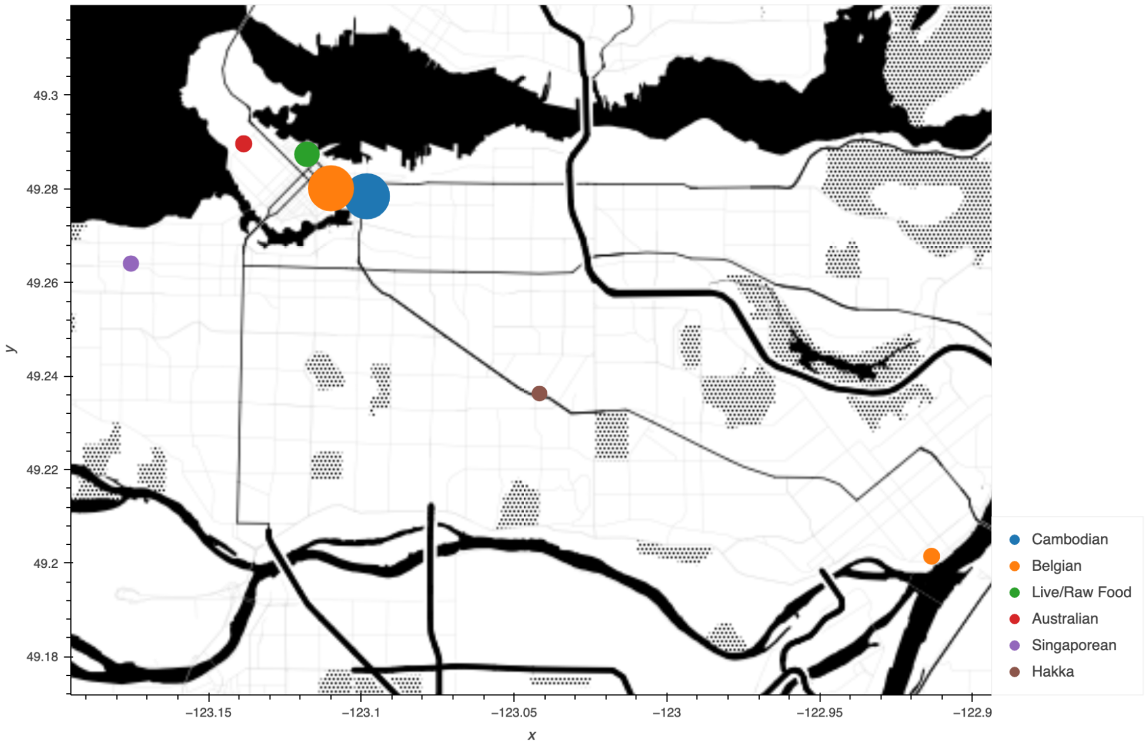


Fig 10. Unique restaurants by type in Vancouver. Size of point represents the number of reviews for that restaurant.

According to the sampled data, Toronto offers less specialty restaurants with a high rating combined with lots of reviews, whereas Vancouver has a number of specialty stand-out restaurants. Another clear observation from these visualizations is these unique specialty restaurants are more spread out across Toronto, while the standout options in Vancouver are clustered together near the downtown core.

It's important to note that this analysis has major limitations since Yelp's Fusion business search API has its own algorithm for showing the most relevant options when searching by city or keyword. Therefore, certain options that seemingly only exist in one city are bound to also exist in the other, but just did not appear in our sample set due to biases.

The second part of this analysis grouped the sample data based on categories that exist in both cities and narrowed this data down to restaurants with more than 250 reviews. Then, to get a weighted average of the rating for a particular category, each restaurant's number of reviews were divided by the total number of reviews for the category and multiplied by the rating for that particular restaurant. Then, these weighted scores were summed within each category to give the average weighted rating within that category.

These scores were then mapped for each city as a single bar graph. This graph was then combined with a line graph of the number of restaurants within each category for each city. The resulting plot can be seen in Fig 11. below:

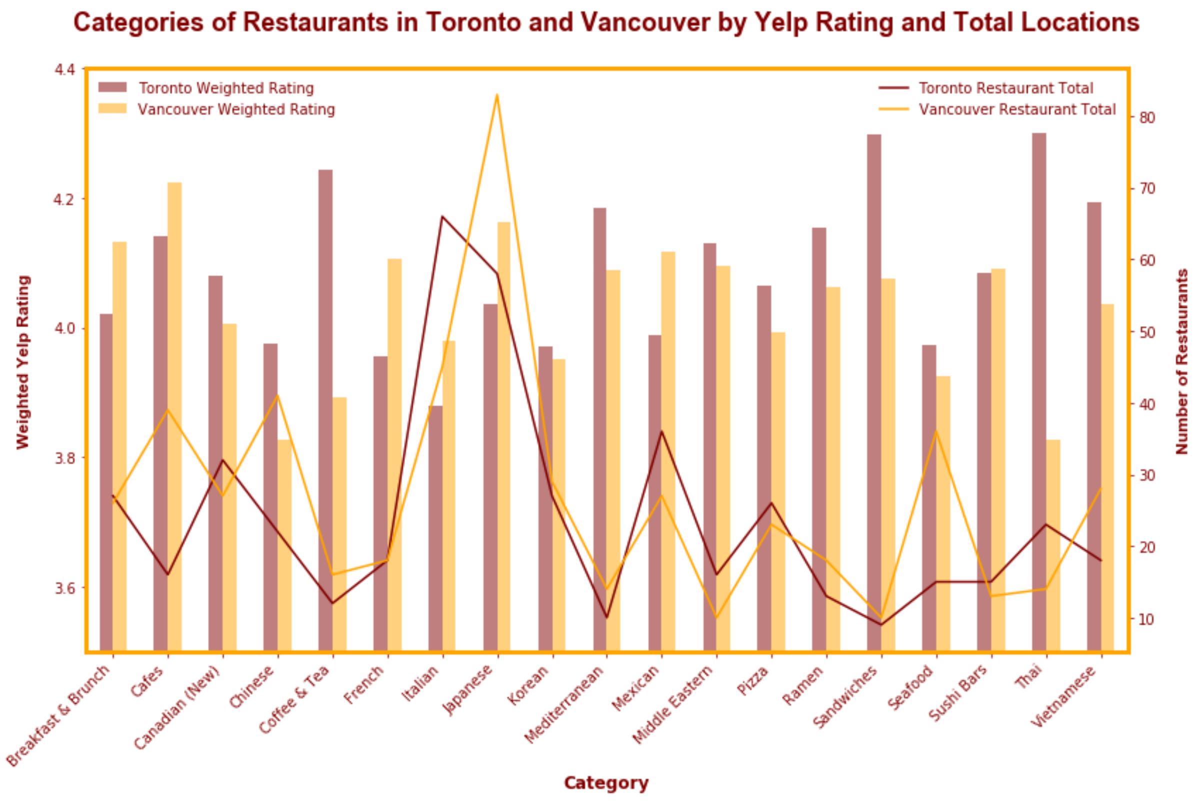
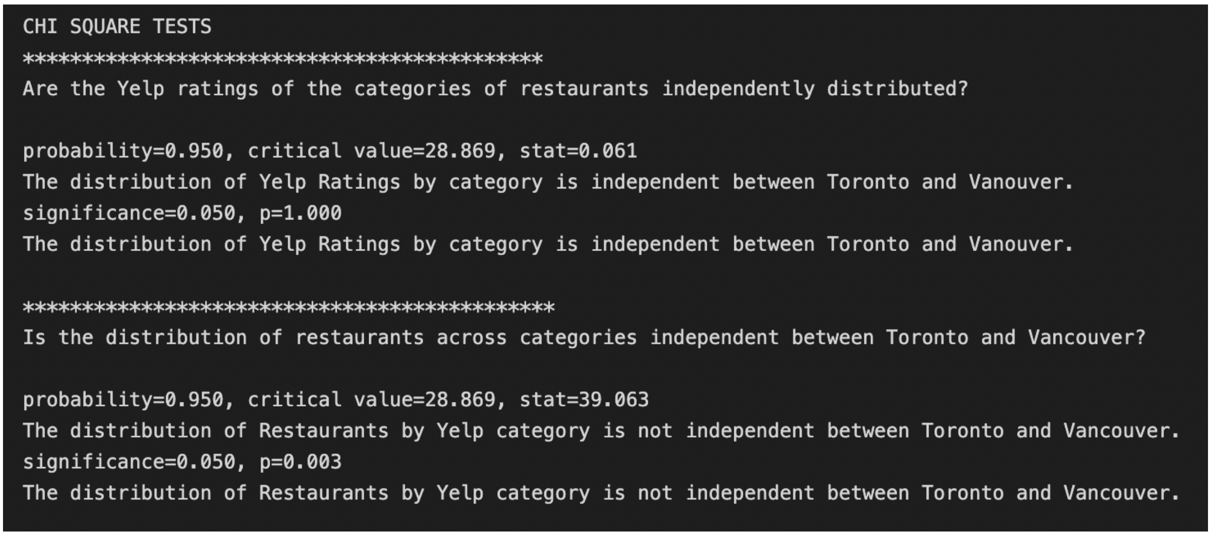


Fig 11. Dual axis plot mapping out the categories of restaurants found in both Toronto and Vancouver by Yelp Rating and Total Location.

Two Chi-Square tests were run to determine whether the distribution of weighted ratings across common categories was independent between cities, and whether the distribution of restaurants across common categories was independent between cities. The result's for both questions are as follows:



```
CHI SQUARE TESTS
********************************************
Are the Yelp ratings of the categories of restaurants independently distributed?

probability=0.950, critical value=28.869, stat=0.061
The distribution of Yelp Ratings by category is independent between Toronto and Vanouver.
significance=0.050, p=1.000
The distribution of Yelp Ratings by category is independent between Toronto and Vanouver.

*********************************************
Is the distribution of restaurants across categories independent between Toronto and Vancouver?

probability=0.950, critical value=28.869, stat=39.063
The distribution of Restaurants by Yelp category is not independent between Toronto and Vancouver.
significance=0.050, p=0.003
The distribution of Restaurants by Yelp category is not independent between Toronto and Vancouver.
```

Therefore, the distribution of restaurants across common categories is NOT independent between Toronto and Vancouver. The restaurants types that exist in both tend to exist in similar frequencies. However, the distribution of weighted ratings of these restaurants is independent between the cities, meaning Toronto and Vancouver Yelp users have different preferences across these common categories.

To verify if the differences in ratings observed in our graph were meaningful, an independent T-test was run on the weighted scores across all restaurants within each common category. The values from this independent T-test analysis can be found in Fig 11. below:

Category	Statistic	P-Value
Breakfast & Brunch	-0.163857	0.870648
Cafes	1.934556	0.069992
Canadian (New)	-0.500657	0.618556
Chinese	1.754715	0.090114
Coffee & Tea	1.180503	0.249528
French	-0.116881	0.907673
Italian	-1.844154	0.069229
Japanese	1.196994	0.233897
Korean	0.267902	0.789803
Mediterranean	0.479953	0.640299
Middle Eastern	-1.210344	0.246001
Ramen	0.747504	0.464260

Sandwiches	0.292427	0.773512
Seafood	2.333075	0.032520
Sushi Bars	-0.371531	0.713441
Thai	-0.767846	0.448096
Mexican	-0.994258	0.324237
Pizza	-0.312086	0.756364
Vietnamese	1.094509	0.286801

Fig. 12 Independent T-test measuring the statistical significance of differences in weighted average ratings for every restaurant within a certain category between Toronto and Vancouver.

If our α value is set to 0.1, setting the confidence level to 0.9, then the ratings difference between Toronto and Vancouver when it comes to Cafes, Seafood, Chinese and Italian restaurants is significant. When this value is dropped to the standard 0.05, then only the Seafood category can reject the null hypothesis of the difference in ratings between Toronto and Vancouver as statistically significant.