Analysis of Yelp data

Haoda Li

March 10, 2019

Part I Loading the Yelp Dataset Challenge

1.1. Permissions

The data obtained from the Yelp Dataset Challenge will only be used for JSC270 Course project. I am permitted use these data to create statistical summaries and visualizations in electronic form. I don't have the permission to publicly display these data set or use them for any commercial purposes. I will not use these data for any purpose that may against law or competitive in nature with Yelp. I will not transfer or manipulate any part of the data.

1.2. Data summary

The data is stored in 5 tables. Each of them stores the following information.(1)

Business

Give all information about businesses on Yelp and stores. Each line gives information about one business

- Referenced by **business-id**
- Address (actual address, city, latitude, longitude, postal code, state),
- Name of the store
- Number of reviews
- Rating stars
- Categories that classify the business
- Attributes: a nested dictionary that stores the different aspects of the restaurant.

Checkin

Give the checkin date time of businesses. Each line gives information about one business

- business-id: reference to the business being checked in
- date: a list of datetime, separated by comma

Review

Give all reviews and rating made by a user. Each line gives information about one review

- referenced by **review-id**

- business-id: reference to the business being reviewed
- cool, funny, useful: number of users that rate this comment as cool, funny, or useful, respectively.
- stars: the integer variable scaled from 0-5 representing the rating
- text: the content of the review
- user-id: reference to the user wrote the review
- date: the date of the review posting

Tip

Give all tips (key information about a business). Each line gives information about one tip

- business-id: reference to the business being tipped
- compliment-count: the count of compliments by other users
- date: the date of the tip posting
- text: the content of the tip
- user-id: reference to the user wrote the tip

User

Give all information about users. Each line gives information about one user. - reference by **user-id** - average_stars: average of rating stars given by the user - complement: there are 11 types of complements, and each stores the count of compliments in that type given by other users. - cool, funny, useful: the counts of cool, funny, useful given by other users - elite: the years when the user is an elite of Yelp - fans: the number of fans - friends: a list of user-id referred to the friends of the user - name: the name of the user - review_count: the total number of reviews written by the user - yelping_since: the datetime start being a Yelp user.

1.3. Additional Data

In this report, I'll use the additional data for plotting maps. Since the map's shapefile uses meta data, I'll extract the data into the root directory. The data used is included in the Reference at the bottom.

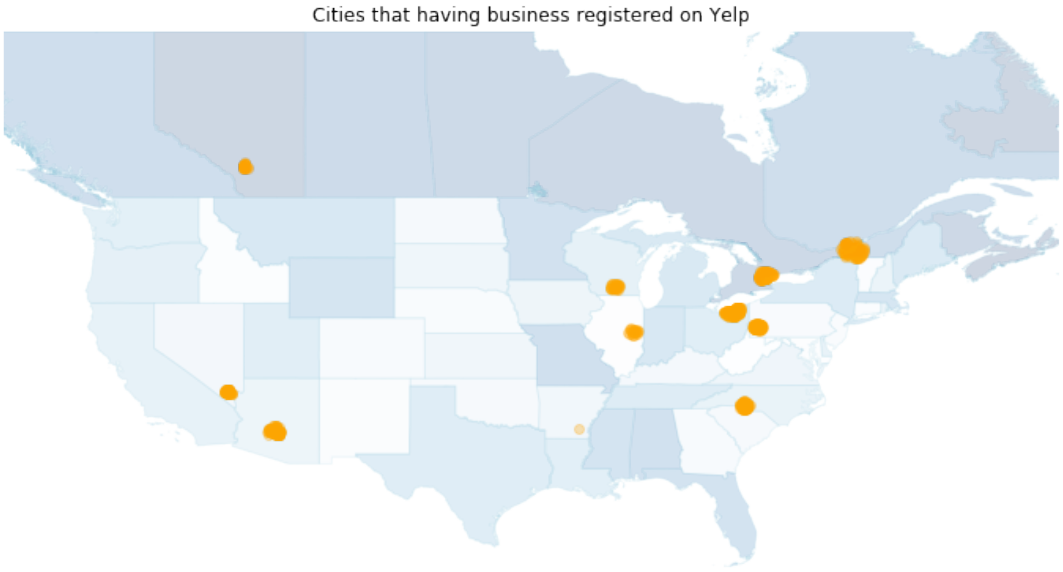
Part II All businesses

There are about 200,000 business in the U.S. and Canada in this dataset.

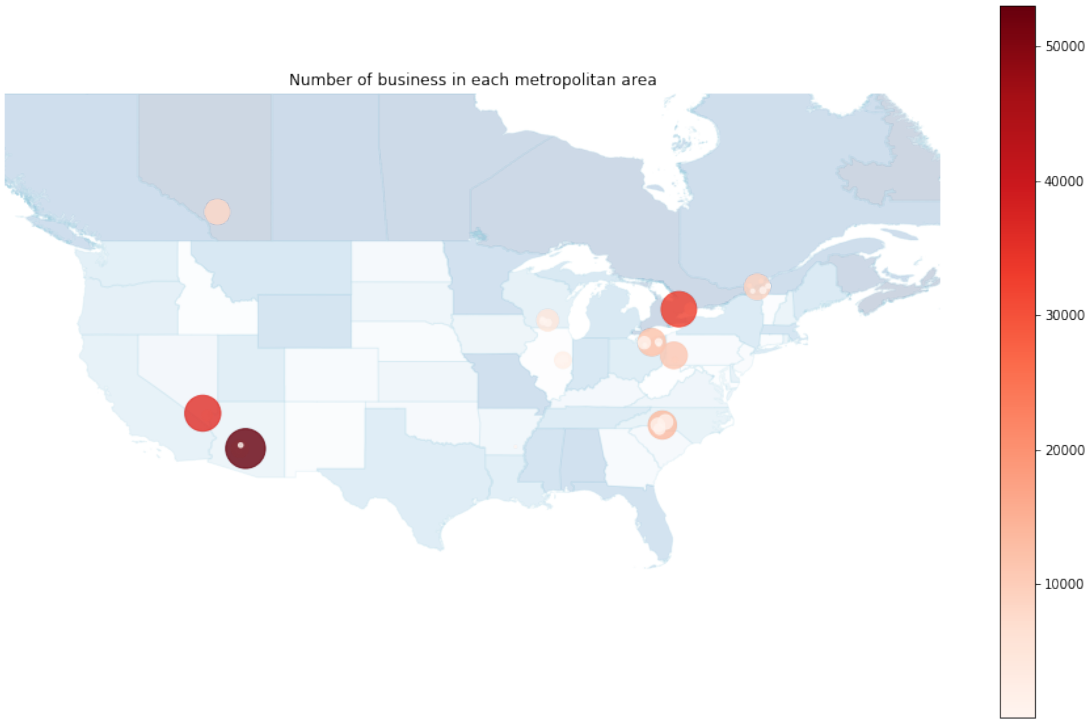
2.1. What cities does this dataset encompass?

Investigation

Since the total number of business it too large, it's unwise to plot all of them on the map. I grouped the business into cities that they belonged to and plot these cities on the North America map(2)(3)
.

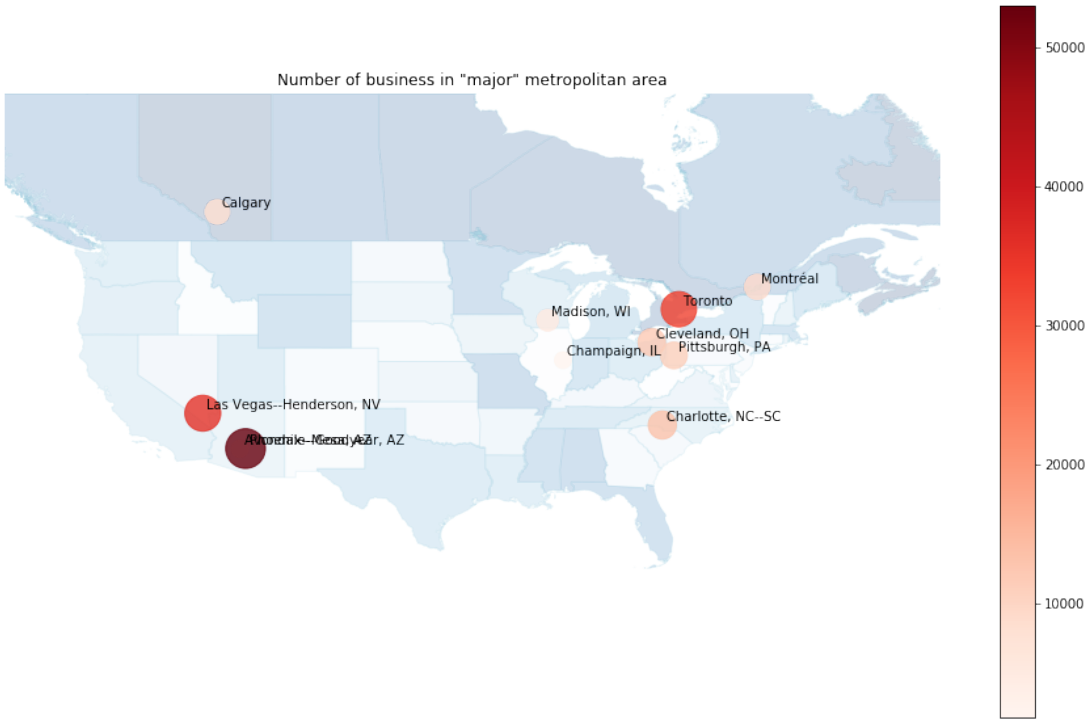


I indicated that these cities are grouped near some metropolitan areas. For example, Toronto, Montreal, Los Vegas. Therefore, I obtained metropolitan areas data from Canada Census (4) and U.S. Census (5). Then, I group the business into their metropolitan areas by their longitude and latitude, and plot the number of business in each metropolitan area.



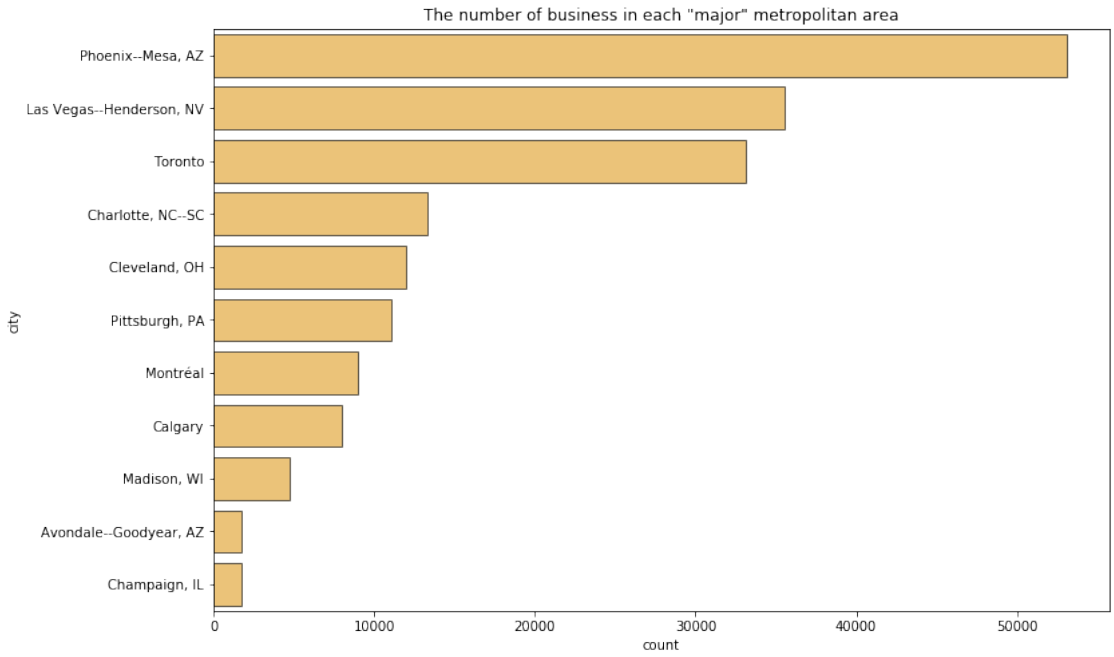
From the plot, I also noticed that most of the metropolitan areas with fewer business are located near some large metropolitan areas. I divide the metropolitan areas into two categories. The "minor" metropolitan areas are those having fewer than 1700 business, and "major" metropolitan areas area those having more than 1700 business. I chose 1700 so that 95% of the business are in some "major" metropolitan areas, and the "minor" metropolitan areas are much insignificant in number and became safe to ignore in this investigation.

Then, I plot all the "major" metropolitan areas.



Conclusion

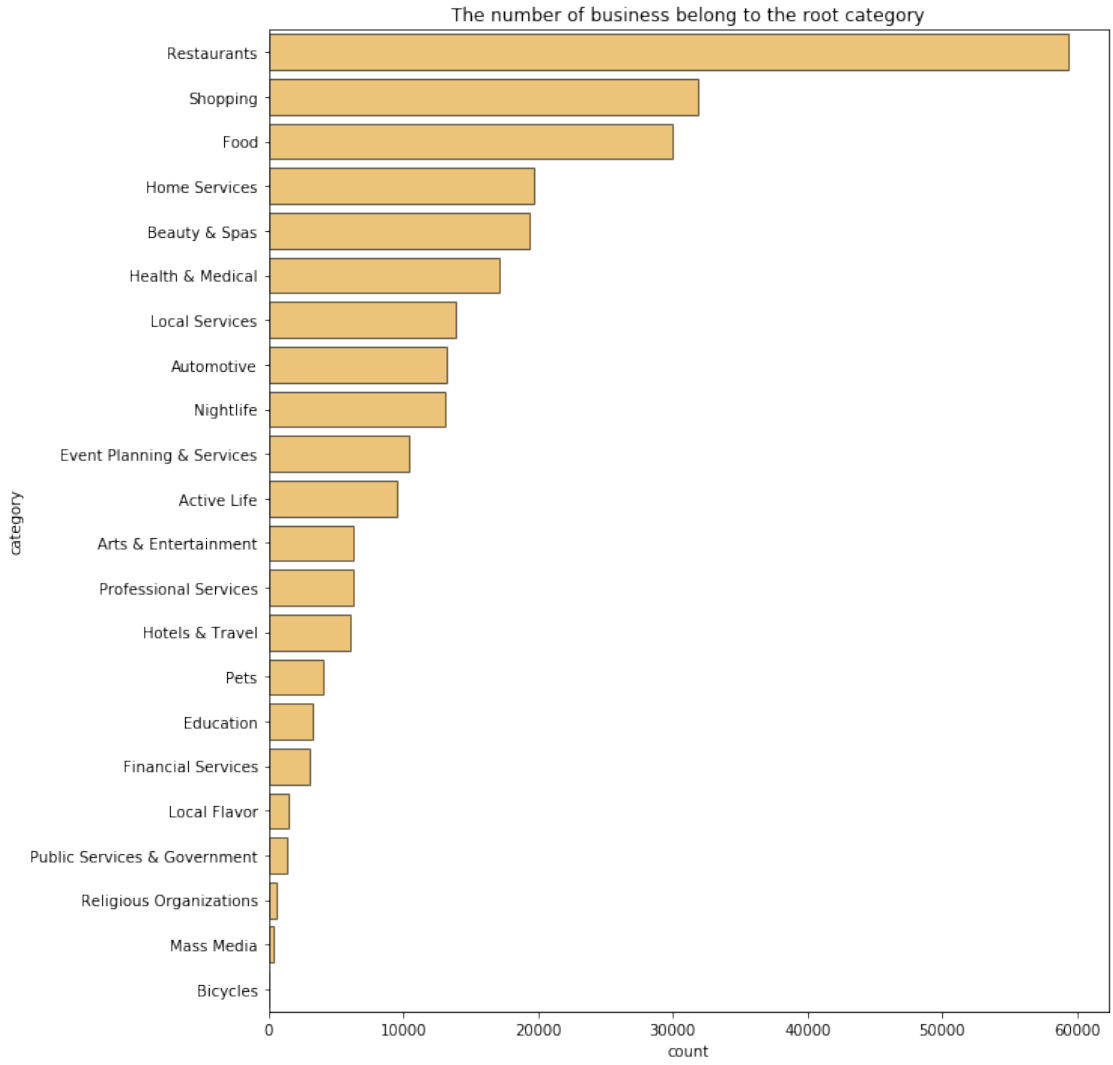
The dataset includes about 200,000 business and encompasses 1258 cities within 36 states/provinces in the U.S. and Canada. These cities are in several large metropolitan areas, as listed below. About 98% of the business are located in some metropolitan areas. The majority of the business are located near Toronto, Las Vegas-Henderson, and Phoenix-Mesa.



2.2 What are the most frequent business categories overall

Investigation

Each business in the dataset is categorized by several key words. In addition, I found the category listing of *Yelp's ALL Category List* (6). Indicated in the category list, each category has a parent attribute and the root categories have empty parent attribute. The categories are managed in a tree-shaped relationship. Therefore, I plot the number of business in each root category (categories without parent category) to examine the distribution of business categories.



Without doubt, "Restaurants" is the most frequent business category overall.

Then, I'm interested in the most frequent restaurant category. Since there are too many restaurant categories, I will look at the 10 most frequent categories for the general pattern.

The 10 most frequent restaurant categories

category	count
Sandwiches	7332
Fast Food	7257
American (Traditional)	7107
Pizza	6804
Burgers	5404
Breakfast & Brunch	5381
American (New)	4882
Italian	4716
Mexican	4618
Chinese	4428

The restaurant categories are divided into ethnic style and the type of food they serve. Among the type of food being served, sandwiches, fast food, piazza, burgers are the most frequent, the number of business in each of such categories does not vary a lot. Among the ethic style, American style shows a definite win. However, I noticed that sandwiches, piazza, burgers are usually served in fast food restaurant. Therefore, I'll say that sandwiches and fast food are the most frequent restaurant categories.

Conclusion

The most frequent business category overall is Restaurants, fast food and sandwiches restaurant are the most popular restaurant category.

Discussion and Further research

An interesting observation from the categories is that there are many overlapping such as "Food" and "Restaurant''. The many overlapping categories are important for the business owners to best describe their business, while it may cause trouble when investigating the proportions of different categories of business. Therefore, further investigation can be performed on the relations and correlations among these categories. A a more concise categorization can be proposed to group business.

2.3. What types of establishments tend to have bike parking?

Investigation

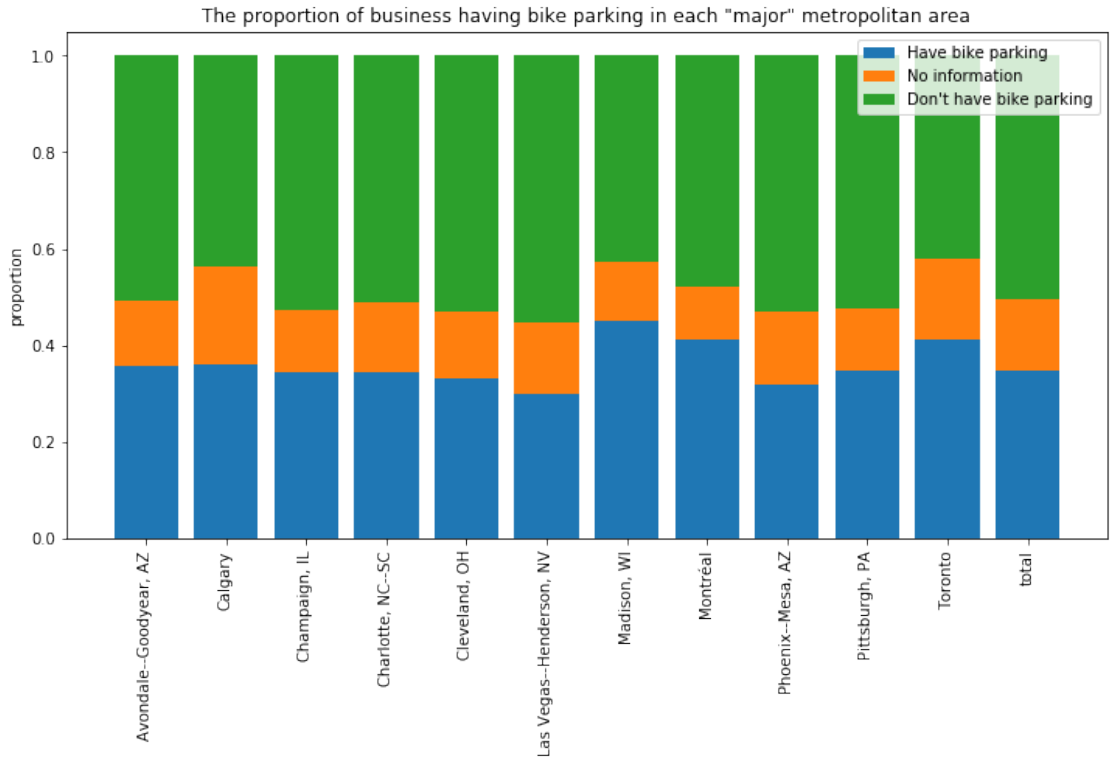
The information about bike parking is stored in the business attributes. A business may have bike parking, have no bike parking, or have no information about bike parking.

To define the "type of establishment", I'm interested in

1. The city the business is belonged to.
2. The category of the business.

whether having bike parking is dependent on the city the business is belonged to.

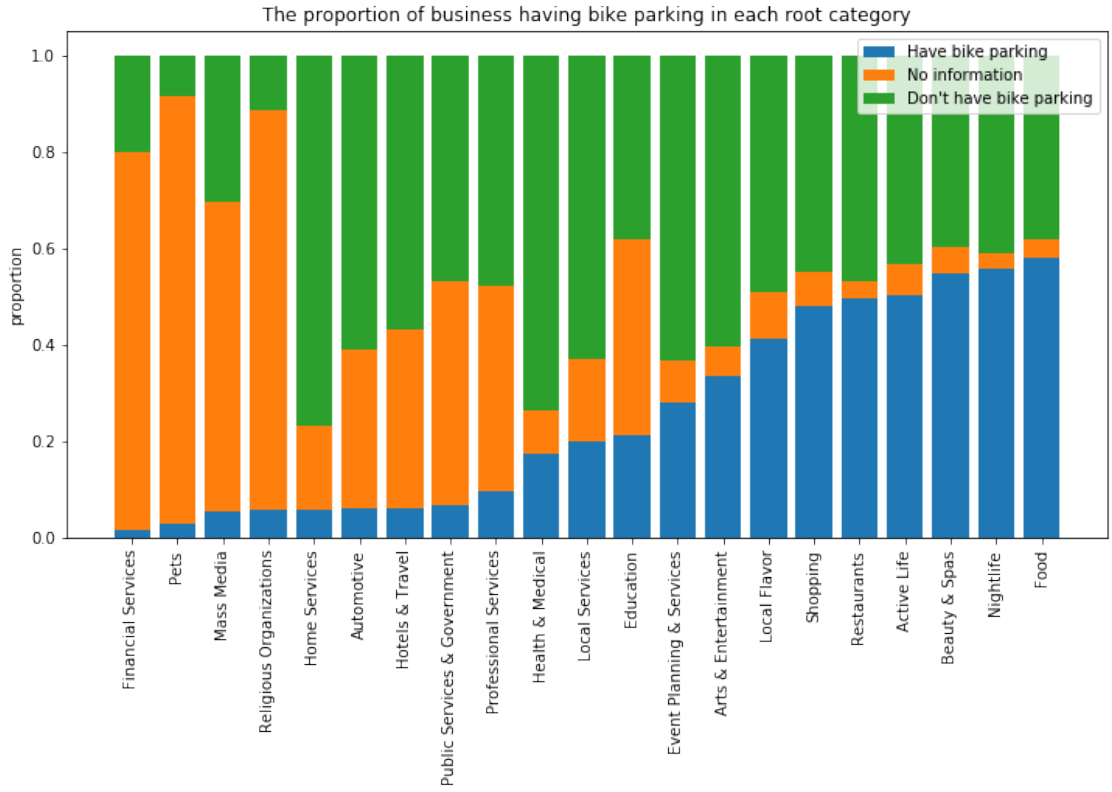
First, I look at the proportion of business that have bike parking in each "major" metropolitan area.



From the plot, business in Madison, Toronto, and Montreal tends be more likely to have bike parking. However, the difference among the proportion is not significant.

whether having bike parking is dependent on the category of the business.

First, I plot the proportion of business that have bike parking for each root category.



From the proportion plot, we notice that there are significant differences among the proportions for different categories.

Then, I performed chi square contingency test. The null hypothesis is that having bike parking is independent on the category of the business, and the alternative hypothesis is that having bike parking is dependent on the category of the business. The test statistic is 94046, the distribution of the test statistic follows χ^2_{40} and p-value is 0.0. The extreme small p-value gives strong evidence that having bike parking is dependent on the category of the business.

From the observations, food and nightlife are most likely to have bike parking.

Conclusion

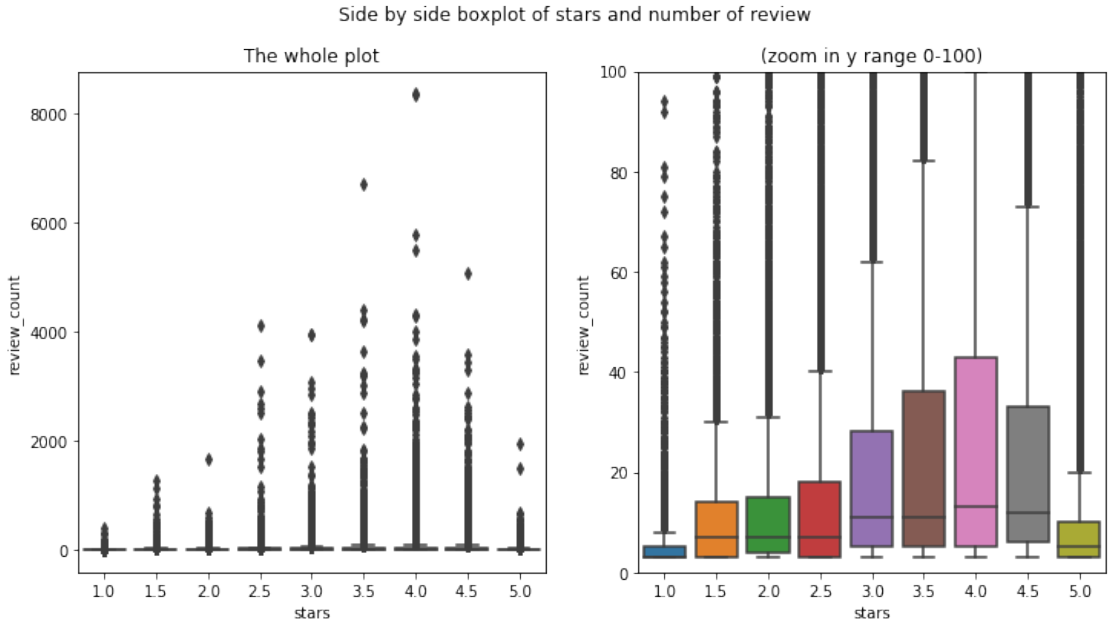
There is some evidence that the business location and the likelihood of having bike parking are independent. On the other hand, there are strong evidence of a relationship between the category of the business and the likelihood of having bike parking. Food, nightlife, and Beauty&Spa have the highest proportion of having bike parking.

Discussion and Further research

Consider the categories with extremely large proportion of lacking bike parking information, namely financial services, pets, and religious organizations. The numbers of business in such categories are significantly smaller than those with smaller proportion of lack bike parking information. Further research may look at whether the popularity of the business category is related to the proportion of business that misses no descriptive attributes (such as bike parking).

2.4.Does more yelp reviews lead to a higher rating, and hence increased sales?

First, I look at the how the numbers of reviews are distributed to each level of stars by side by side boxplots.



There are lots outliers for each box plot, and for the majority of the business (> 75%), they have fewer than 100 reviews. For the outliers, the density of outliers decreases as the review count increases.

From the observations, the minimum of each group is the same, first quartile, median, third quartiles, and maximum all reach peak at star 4.0.

Then, I looked at the mean of the number of reviews and the count of business in each group of rating star

The count and mean of business in each group of rating star

stars	count	mean
1.0	4874	5.815552
1.5	4976	15.596664
2.0	11426	15.108874
2.5	18843	20.910630
3.0	25996	30.857286
3.5	35008	40.681130
4.0	35969	56.523228
4.5	27301	43.444453
5.0	28216	12.113942

Conclusion

Although the question asks whether more yelp reviews lead to a higher rating, I cannot draw any conclusion about the causation without further information. I can only state that there is a correlation between the two variables. From all the statistical summaries, there is a relationship between the number of yelp reviews and the rating. Overall there is no evidence that the relationship is positive. Instead, the number of reviews is higher when the rating is closer to 4.0.

Discussion

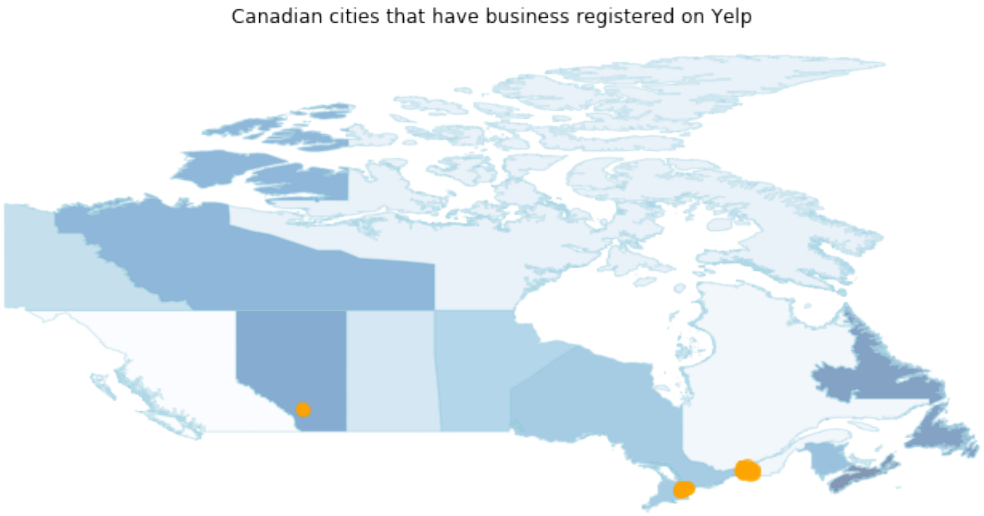
I notice that the average rating of all reviews on Yelp is about 3.7. The correlation may not be meaningful because the user might just be more likely to give a rating of 4, and it is not related to how popular the business is.

Part III Canadian Business

There are 50,644 Canadian business in this dataset.

3.1. What cities does this dataset encompass?

I plot all the Canadian cities that have Yelp business registered on the map.



When I checked the number of cities that have Yelp business. I notice the city in British Columbia, which does not appear on the map plot above. I checked the city and it's Richmond Hill and its geometry point is in Ontario. I decide to remove the error data point.

Conclusion

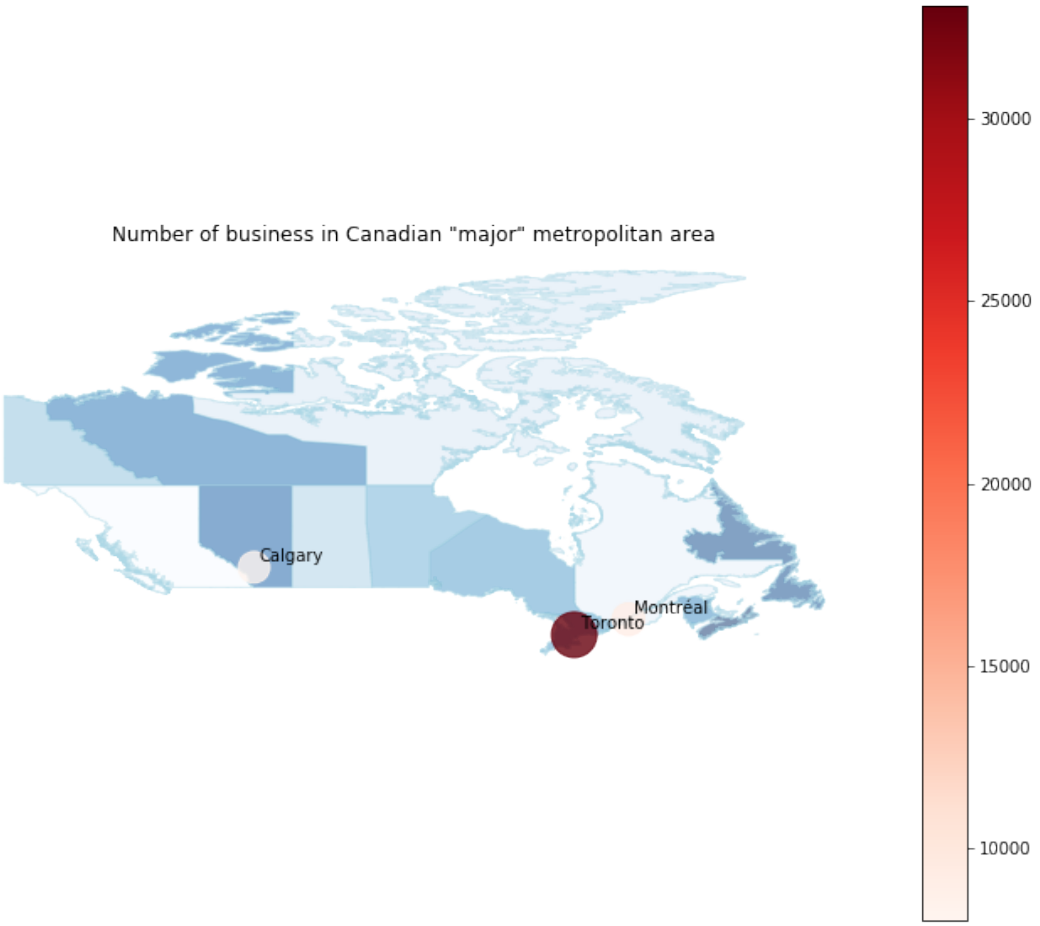
The dataset encompasses 391 cities in 3 provinces, namely Alberta, Ontario, and Quebec.

The number of cities that have Yelp business in each province

Province	number of cities
AB	33
ON	139
QC	219
Total	391

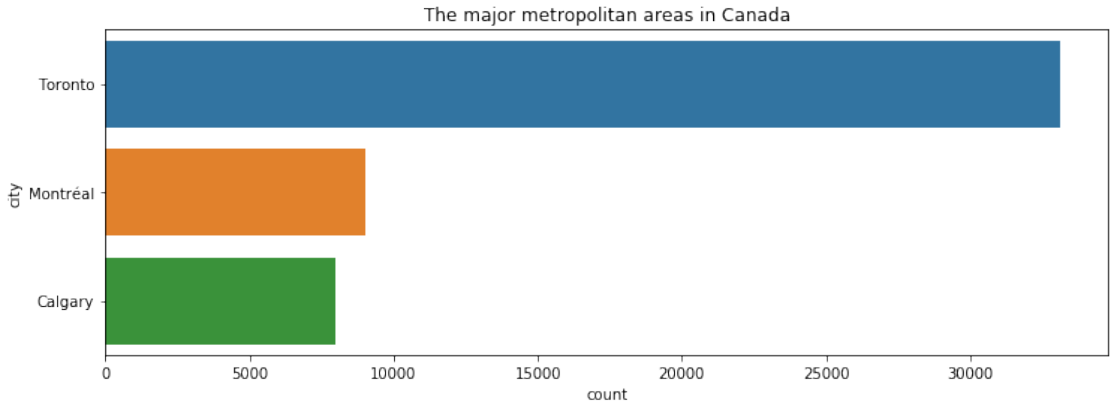
3.2. Identify the larger metropolitan regions that these cities belong to.

Since I have done the similar investigation in question 2.1, I will just plot the Canada portion and draw the conclusion.



Conclusion

These cities belongs to the three major metropolitan areas: Toronto, Montreal, and Calgary. Among the three areas, GTA has the most business.



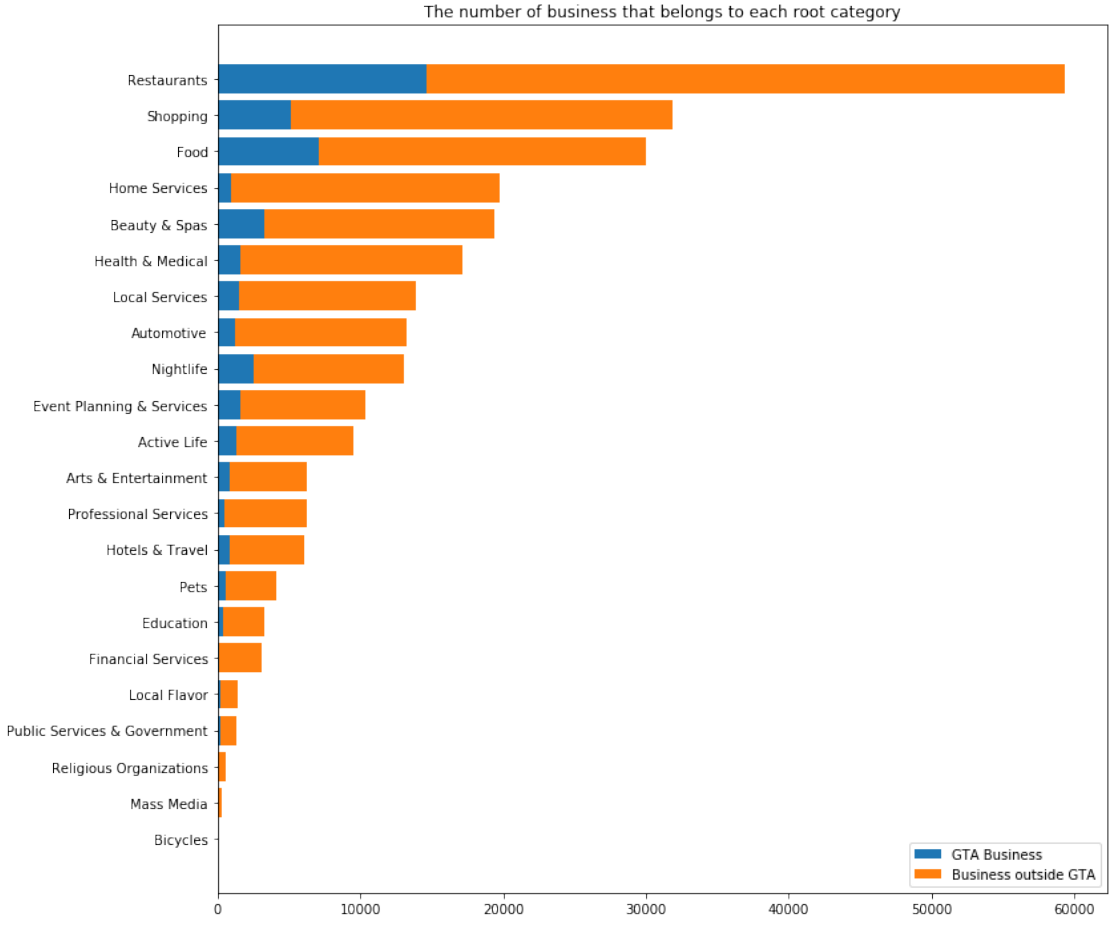
Part IV GTA businesses

The number of business in GTA is 33112.

4.1. What are the most frequent business categories? How do they compare against the trends listed in 2.2?

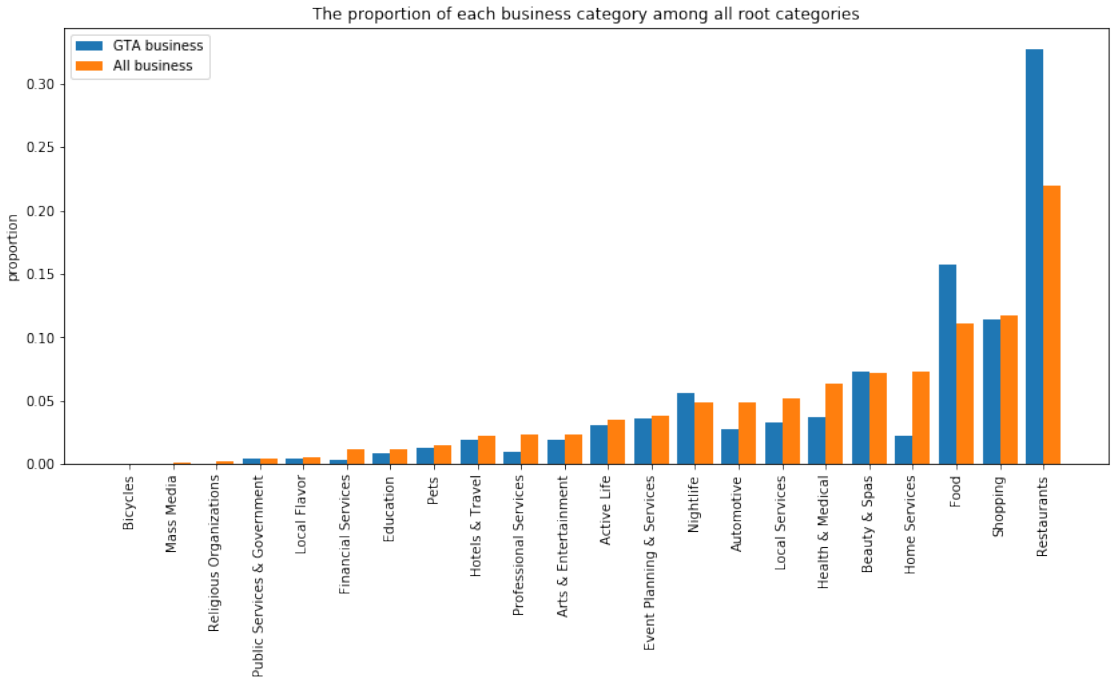
Investigation

To find the general trend, I plot the number of all GTA business in each root category against all business.



Obviously, "restaurants" is the most frequent business category then follows food.

Also, the bar plot shows that the trend of GTA business categories is different from that of all business. To further analyze the trend, I looked at the proportion of the count of each category in all the count.



From the plot, I observe that restaurants, food, and nightlife take a larger proportion in GTA business than in all business. On the other side, home service, health & medical, local services, and automotive, professional services take a smaller proportion in GTA business than in all business.

Conclusion

Restaurants is the most frequent category in GTA area. Compare to the trend of all business on Yelp, GTA business on Yelp favors more about Restaurants, food, nightlife; and favors less about home service, health & medical, local services, and automotive, professional services.

Discussion and further analysis

Notice that the categories are not independent of each other, hence the proportion of each category is influenced by correlation.

Also, I'm interested whether Toronto actually favors more about restaurants, food, and nightlife compared to other metropolitan area. Alternatively, whether Yelp service focus more on the restaurant in GTA, even Canada, than in the U.S.

4.2. What are the top franchises in the city?

Investigation

First, I group the business by their name to find the frequency distribution of the names. I expect to see some well known brands appears at the head. I put the 20 most frequent names in a table to examine which are the top franchises.

The 20 most frequent names in GTA business

name	count
Starbucks	259
Tim Hortons	233
McDonald's	146
Shoppers Drug Mart	101
Pizza Pizza	95
Subway	88
Swiss Chalet Rotisserie & Grill	76
GoodLife Fitness	68
Popeyes Louisiana Kitchen	67
Second Cup	63
LCBO	56
Tim Horton's	55
Pizza Nova	51
Pizza Hut	51
Domino's Pizza	48
Wild Wing	44
Sunset Grill	43
KFC	41
Wendy's	40
Aroma Espresso Bar	38

From the table, Starbucks seems to be the top franchise. However, from my own experience, there are definitely more Tim Horton's in the city than Starbucks. When I looked closed at the data, I notice that both "Tim Hortons" and "Tim Horton's" are in the table.

The business names may vary while the brands are iconic. Therefore, I propose 10 potential franchises from the table above and find the count of these brands in the data set. To maximize the accuracy, I remove the apostrophe and space and lower case the letters. Because each brand name is long enough, I assume that there is no coincidence that some other business may include the keywords.

The 10 most frequent brand names

proposed brand	searched keyword	count
Starbuck's	starbucks	260
Tim Hortons	timhorton	299
McDonamd	mcdonald	160
Shoppers Drug	shoppersdrug	103
Pizza Pizza	pizzapizza	95
Subway	subway	120
Swiss Chalet	swisschalet	95
Good Life	goodlife	74
Popeyes Louisiana Kitchen	popeyeslouisiana	67
Second Cup	secondcup	88

Conclusion

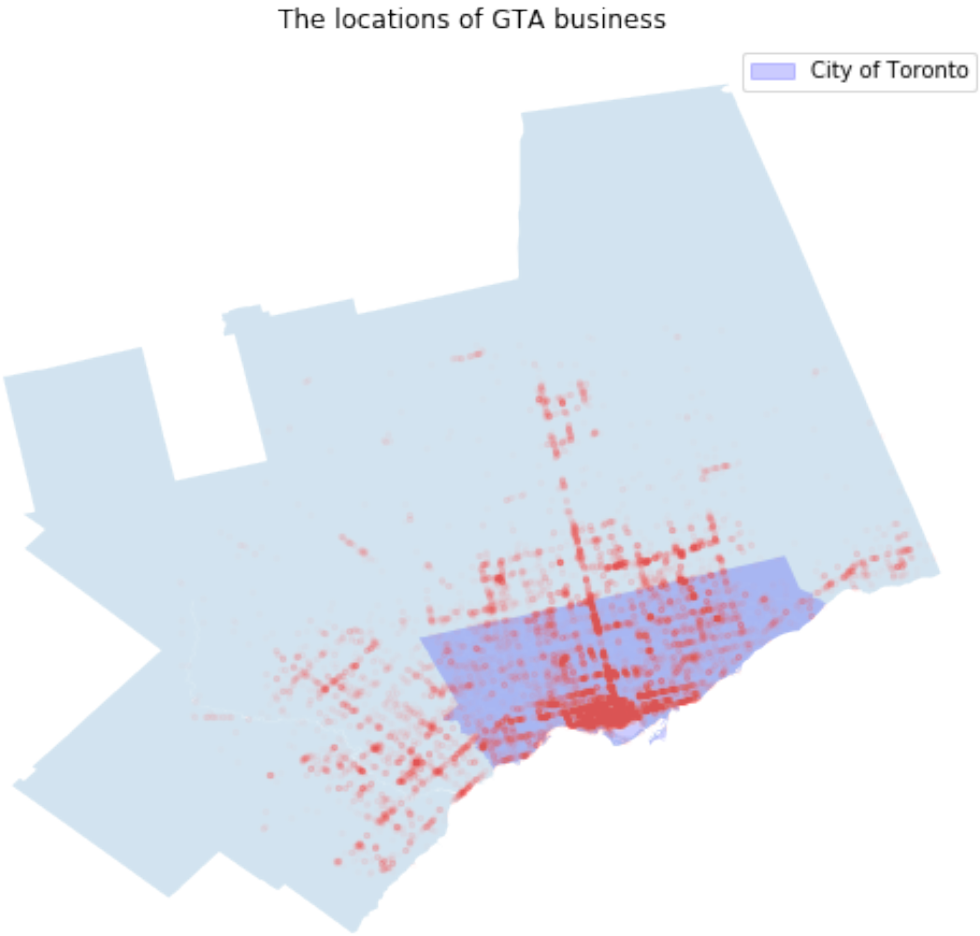
Tim Horton's is the top franchise in the city, which has near 300 locations. The second place is

Starbuck's, with 260 shops.

4.3. Does business location play an important role in reviews?

Investigation

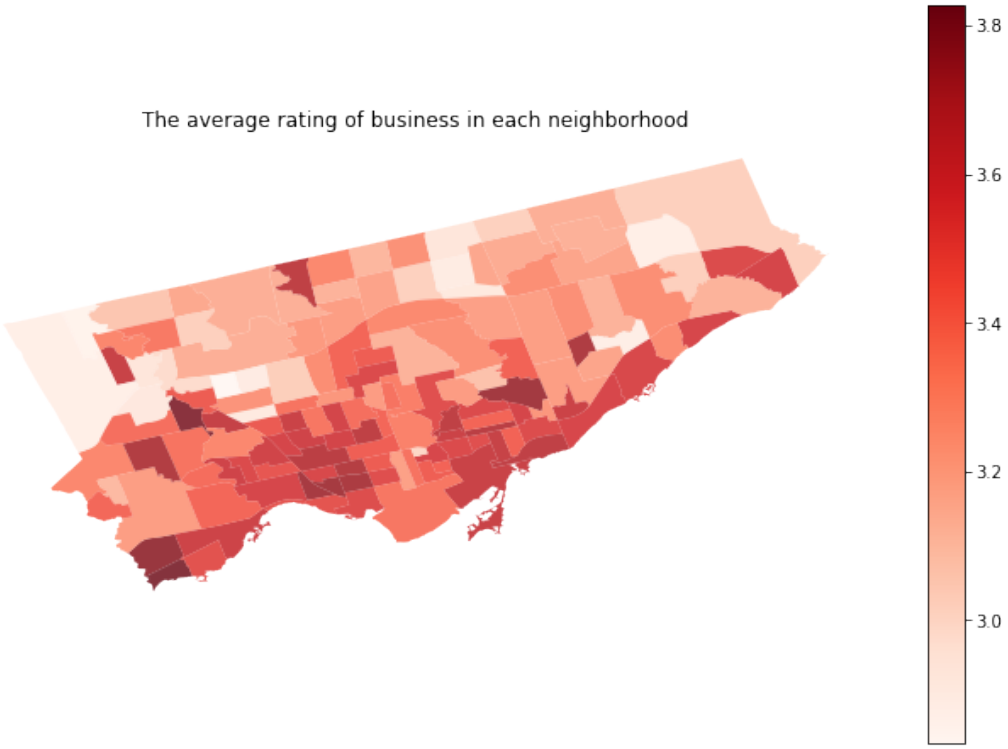
To investigate the relationship between the business and reviews, I first plot all GTA business on the map to see how they are distributed.



Notice that the majority of the points are located near the city of Toronto. I decide to take a closer look at the Toronto portion. To look at the relationship between location and reviews, I looked at reviews from two aspects: the rating and the number of reviews.

how does business location influence the rating

I'm interested in whether the rating depends on the neighborhood the business is located. For the neighborhoods in Toronto7, I color them by their rating.



Observed from the plot, the business near the lake shore tends to have higher rating.

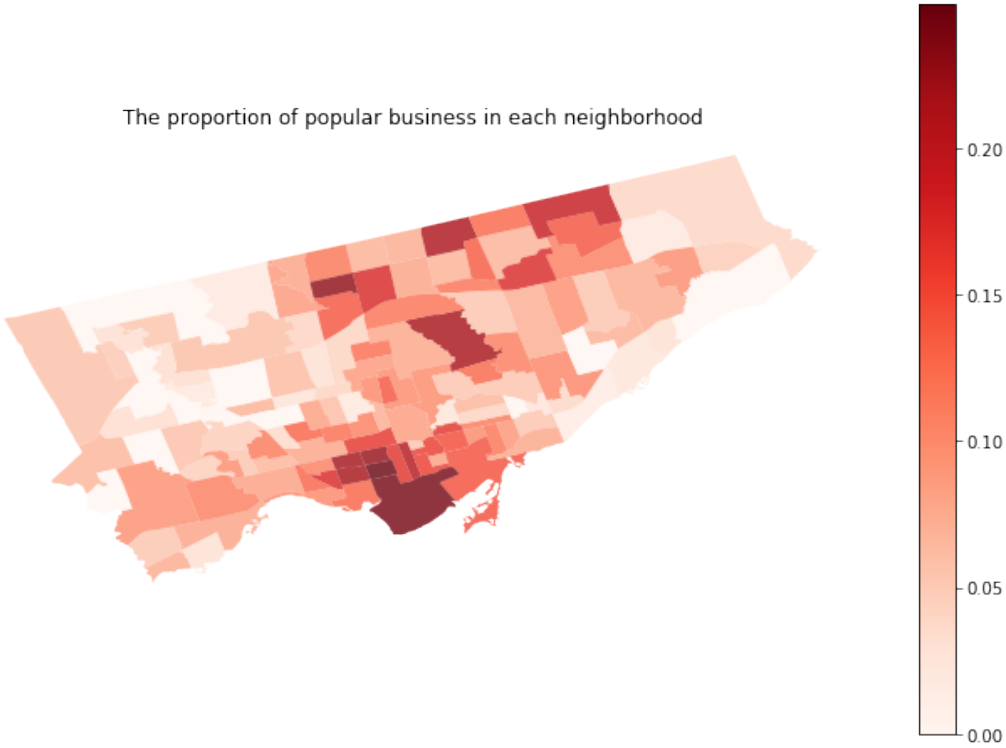
how does the business location influence the count of reviews

From the previous investigation, I know that the count of reviews is not evenly or normally distributed among business, as shown in the boxplot below.

Finding mean review count of each area is insignificant: the business that have plenty of reviews has too much influence on the mean. Also, I'm more interested in whether the location is associated with the "popular" business.

Therefore, I propose that the "popular" business are those have more reviews than 44 reviews and the rest are "unpopular''. 44 is from the boxplot whiskers, and we can obtain the business that are far more popular with this categorization.

Then, I can calculate the proportion of popular business in each neighborhood, and color the neighborhood on the plot.



I don't observe any obvious pattern on the influence of business location that influence the proportion of popular business.

Conclusion

There is evidence that business location and reviews are related. The business location plays an important role in review ratings. Specifically, the closer to the lake shore is correlated with the higher rating. However, there is no obvious pattern between the business location and the proportion of business with more reviews.

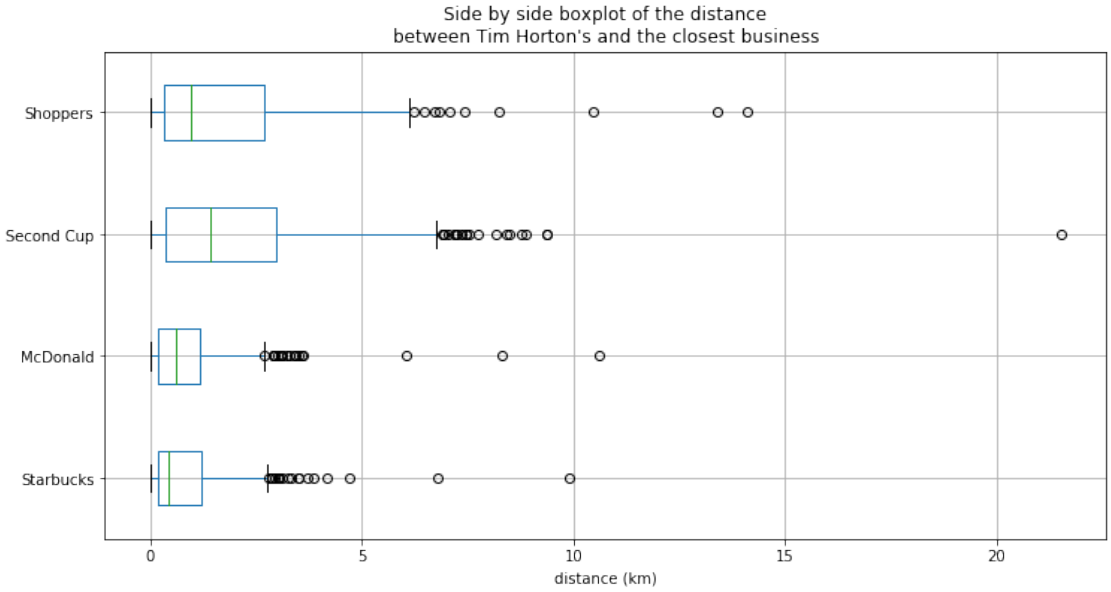
4.4. Is it true that for every Tim Hortons in the GTA there is a Starbucks nearby?

Investigation

First, I calculate the closest Starbucks for each Tim Horton's location and access the distribution of the distance. I plot all the Starbucks and Tim Horton's on GTA map, along with the line connecting each Tim Horton's and its closet Starbucks.

From the plot, if you are in the center of Toronto, there's definitely a Starbucks near Tim Horton's. However, it is not true when you are away from the city center. Tim Horton's seems to cover more areas than Starbucks.

I further do the side by side boxplot to examine the distribution of distance between Tim Horton's and its closest Starbucks compare to the distance between Tim Horton's and other franchises, namely McDonald, Second Cup, and Shoppers. I choose these three because McDonald is one of the top three franchises, Second Cup is another Coffee franchises, and Shoppers is the top non-food franchise.



In addition, I find the proportion of Tim Horton's that the closest Starbucks, McDonald, Second Cup, and Shoppers that is within 1km. I choose 1km because it's about the sight distance and notice that the distance is the length of the line between two points, the actual distance is usually greater.

The proportion of Tim Horton's that the closest brand business is within 1km

brand	proportion
Starbucks	0.69
McDonald	0.68
Second Cup	0.43
Shoppers	0.51

Conclusion

It's not true that for every Tim Horton's in the GTA there is a Starbucks nearby. However, it is usually the case, more specifically, in 70% of time and when you are located in the city of Toronto. Another interesting discovery is that the distance distribution of Tim Horton's and closest Starbucks is similar to that of Tim Horton's and closest McDonald. Therefore, maybe there is no "special connection" between Tim Horton's and Starbucks.

4.5. Do Yelp reviewers use similar language in their reviews of GTA's Tim Horton's and Starbucks?

Investigation

To approach this question, I will also use Second Cup as the reference group. I choose Second Cup as a reference since it's another popular coffee brand in GTA.

First, I find all reviews about GTA's Tim Horton's, Starbucks, and Second Cup.

The number of reviews of each brand

name	Total reviews	Reviews per shop
Tims	1842	6.16
Starbucks	2940	11.31
second cup	832	9.45

I extract all the noun phrases, since they contributes to the majority of a text's meaning. I do a word count on the extracted phrases. I remove phrases thats related to the business brand itself, since they are highly dependent on the brand the reviewer went. Note that this means there's no phrase like "Starbucks" in Starbucks' reviews but "Tim Horton's" will still exist in Starbucks' reviews.

After the processing, there are 11606 phrases describing Tim Horton's, 19666 phrases describing Starbucks, and 6504 phrases describing Second Cup.

To analyze the phrases, I first do the word cloud to see the general languages and word distributions.

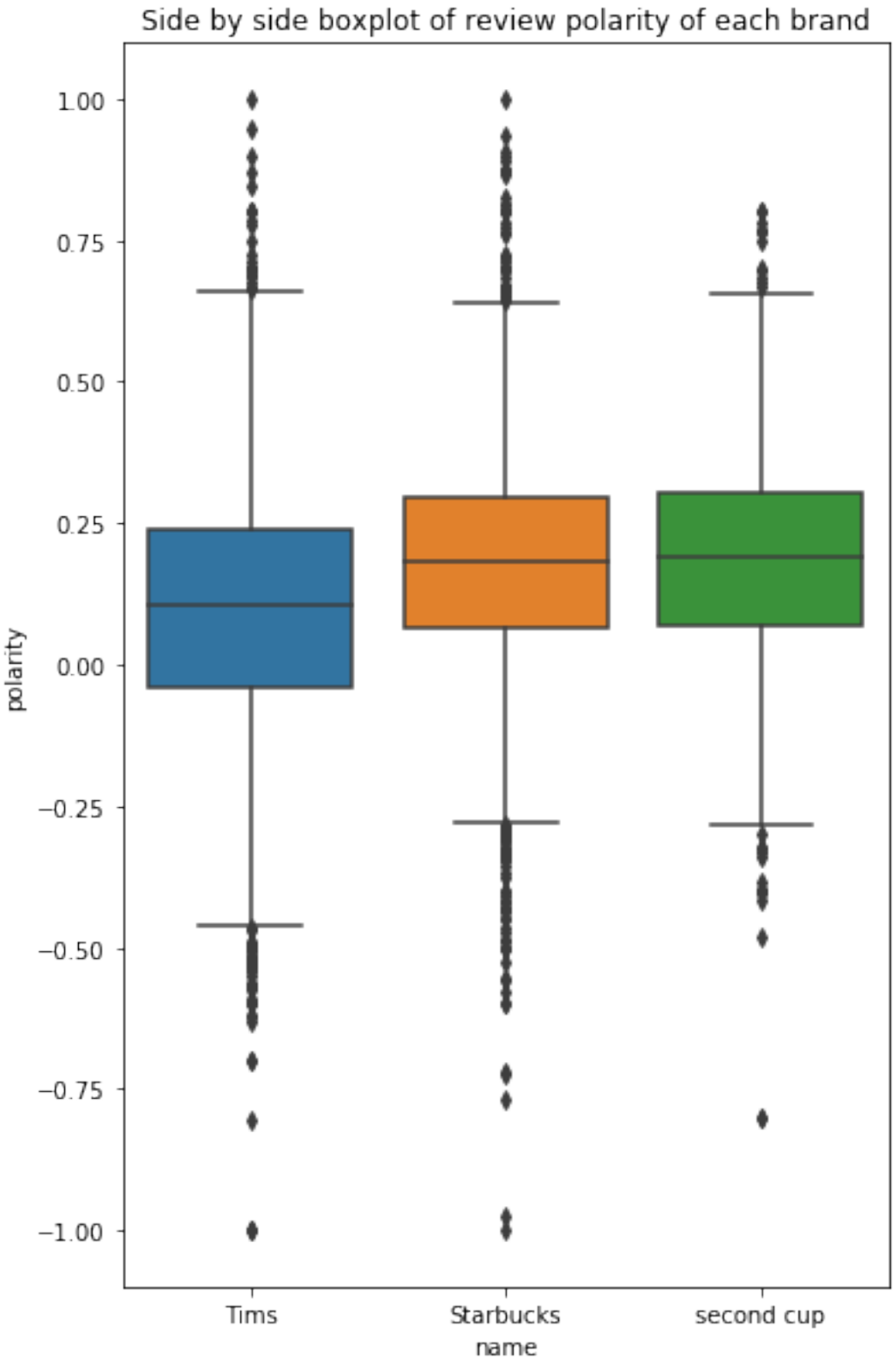
Then, I look at the 10 most frequent phrases for the three franchises.

The 10 most frequent phrases for each business

index	Tims rank	Starbucks rank	Second Cup rank
canada	1		
toronto	2	3	2
starbucks	3		1
ca n't	4	2	8
coffee	5		4
customer service	6	5	
staff	7	1	
wendy	8		
wifi	9	6	9
overall	10	7	
nice		4	5
love		8	
great place		9	
great location		10	
coffee shop			3
cup location			6
hot chocolate			7
free wifi			10

"Toronto", "can't", and "wifi" appear in all three cases. Also, Tim Horton's and Starbucks have 6 phrases, Tim Horton's and Second Cup share only 5 phrases, Starbucks and Second Cup share 4 phrases. Overall, reviews on Tim Horton's and Starbucks share more similarities compares to reviews on Second Cup.

ALso, a very interesting observation is that reviews on both Tim Horton's and Second Cup mentioned "Starbucks" frequently, while no reviews on Starbucks mentioned "Tims" or "Second Cup''. Also, Starbucks reviewers have more positive phrases such as "nice", "love", "great place", and "great location''. Therefore, I did a sentiment test to evaluate their polarity of each review. Polarity is a measurement between [-1, 1], from -1 being the most negative to 1 being the most positive.


Side by side boxplot of review polarity of each brand
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
polarity
Tims
Starbucks
second cup
name

From the boxplot, I notice that the Tim Horton's has significantly smaller number for all quartiles and lower whisker than Starbucks and Second Cup. Starbucks and Second Cup have similar statistics.

I also calculate the proportion of positive reviews (polarity > 0) and the mean polarity of reviews for each coffee brand.

The proportion of positive reviews and the mean of polarity

brand	proportion of positive reviews	mean polarity of reviews
Tims	0.676982	0.098226
Starbucks	0.836054	0.179703
second cup	0.837740	0.188114

Again, Tim Horton's has lower proportion and mean than the other two brands. From all observed statistic above, I have enough evidence that Tim Hortons' reviews have more negative language than Starbucks' and Second Cup's reviews; Starbucks and Second Cup have similar languages in their reviews.

Conclusion

Generally, Yelp reviewers use different language in their reviews of GTA's Tim Horton's and Starbucks. Although some phrases are mentioned in both franchises' reviews, Starbucks customers tend to have more positive languages in their reviews than Tim Horton's customers.

Discussion and Further Research

Starbucks' and Second Cup's products often have high price tag than Tim Horton's product, and they have more positive reviews. Also, in question 4.3 I conclude that the business near the lake shore have high ratings, and we know that most luxurious business are located near the shore. Therefore, further research can be done on whether there is a positive relationship between business pricing and the review ratings / languages.

Reference

1. Yelp. *Yelp Dataset JSON*.

https://www.yelp.com/dataset/documentation/main

2. 5. U.S. Census Bureau. *Cartographic Boundary Shapefiles - States*

https://www.census.gov/geo/maps-data/data/cbf/cbf_state.html

3. 4. Statistics Canada. *Boundary Files, 2011 Census. Catalogue no. 92-160-X*.

https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2011-eng.cfm

6. Yelp Fusion. *All Category List*.

https://www.yelp.com/developers/documentation/v3/all_category_list

7. City of Toronto. *Locations & Mapping – Data Catalogue*.

https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/locations-and-mapping/#a45bd45a-ede8-730e-1abc-93105b2c439f