Yelp_Businesses

March 9, 2019

JSC270 Assignment 2: Yelp Business Types and Relationships

Shirley Wang

Feb 24, 2019

---

1 Introduction

> Yelp is a local-search service powered by crowd-sourced review forum run by an American multinational corporation. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations. The Yelp.com automated filter algorithm removes many positive reviews from companies, in view to avoid fake reviews, which has caused controversy. (see Yelp's fake review problem) Yelp was founded in 2004. By 2010 it had $30 million in revenues and the website had published more than 4.5 million crowd-sourced reviews. From 2009 to 2012, Yelp expanded throughout Europe and Asia. During the fourth quarter of 2017, Yelp has 77 million unique visitors via desktop computer and 64 million unique visitors via mobile website on a monthly average basis. By the end of 2017, Yelp had 148 million reviews.

-- Adapted from Wikipedia

In this report I will analyze the Yelp dataset and use it to investigate the cities and most popular business types on Yelp. I will investigate what kind of businesses tend to have bike parking, and if reviews are correlated with ratings. I will also investigate Canadian businesses and businesses in the Greater Toronto Area specifically, and analyze Tim Hortons and Starbucks distances and reviews. I will specifically be answering these questions:

1.1 Questions

1. Loading the Yelp Dataset Challenge (Round 13)

   1.1. Permissions. Before downloading, read the Dataset License and explain in a few sentences what you can and cannot do with this data using your own words.

   1.2. The data. Before loading the data, get yourself acquainted with the data. How is the dataset structured and what relationships are present between the different files?

2. All businesses

   2.1. What cities does this dataset encompass?

   2.2. What are the most frequent business categories overall?

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

   2.4. An article recently claimed that having more yelp reviews lead to a higher rating, and hence increased sales. Do the data support this claim?

3. Canadian businesses

   3.1. What cities does this dataset encompass?

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

4. GTA businesses

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

   4.2. What are the top franchises in the city?

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

   4.4. Is it true that for every Tim Hortons in the GTA there is a Starbucks nearby? Calculate distances between establishments of the two groups and assess distance patterns. Plot the two types of establishments on a map.

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

2 Question 1: Loading the Yelp Dataset Challenge (Round 13)

2.1 1.1. Permissions

This Yelp dataset is distributed with the purpose of us participating in Yelp's Dataset Challenge. We are allowed to analyze this data for academic purposes only, to learn things from it. But Yelp still owns the data and as a result has the right to know what we do with the data. We aren't allowed to profit from it, publicly display it, use it to compete with Yelp, or do anything illegal with the data.

2.2 1.2. Data

	business	checkin	tip	photo	review	user
0	address	business_id	business_id	business_id	review_id	user_id
1	attributes	date	compliment_count	caption	user_id	name
2	business_id		date	label	business_id	review_count
3	categories		text	photo_id	stars	yelping_since
4	city		user_id		useful	useful
5	hours				funny	funny
6	is_open				cool	cool
7	latitude				text	elite
8	longitude				date	friends
9	name					fans
10	postal_code					average_stars
11	review_count					compliment_hot
12	stars					compliment_more
13	state					compliment_profile
14						compliment_cute
15						compliment_list
16						compliment_note
17						compliment_plain
18						compliment_cool
19						compliment_funny
20						compliment_writer
21						compliment_photos

The dataset is stored in the JSON file format, with each line representing an entry in its dataframe.

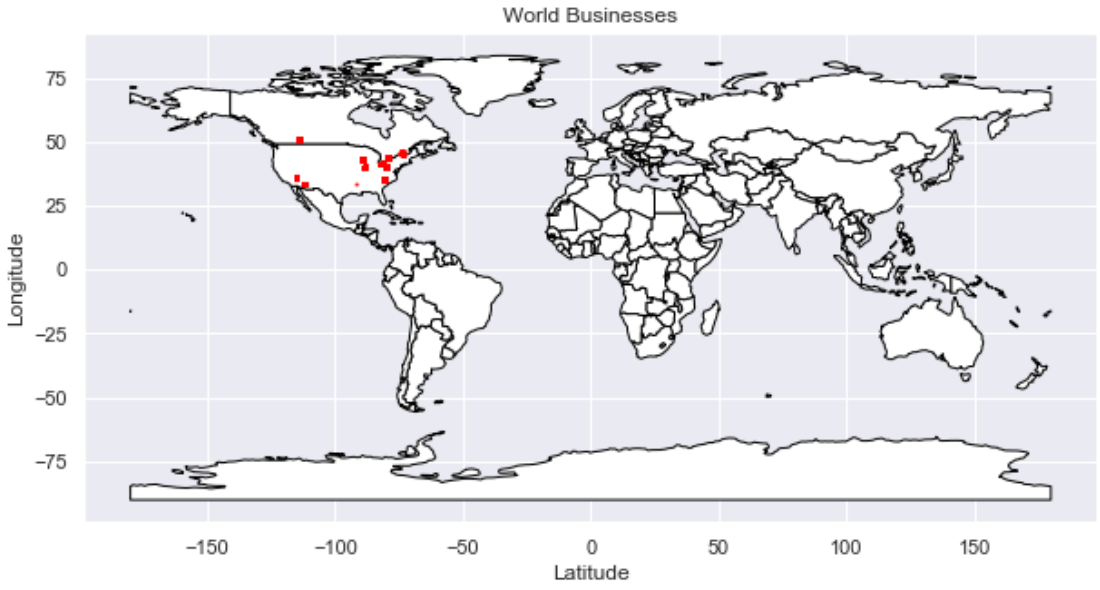
The business dataset contains information about every business in Yelp, corresponding to its own unique business id. The checkin dataset contains information about the number of times each business has checked in with Yelp. The review dataset contains information about reviews on Yelp, with each one having the business id of the business, the user id of the user who provided the review, and its corresponding unique review id. The photo dataset contains information about every photo on Yelp and the business id of the business that the photo belongs to. The tip dataset contains information about tips on Yelp, with each one having the business id of the business, and the user id of the user who provided the tip. The user dataset contains information about every user on Yelp, the number of reviews and statistics of reviews they've made, and their corresponding user id.

The main thing to note is that many of these datasets share business ids and user ids, which each correspond to a specific user or business that uses Yelp and identify what belongs to who.

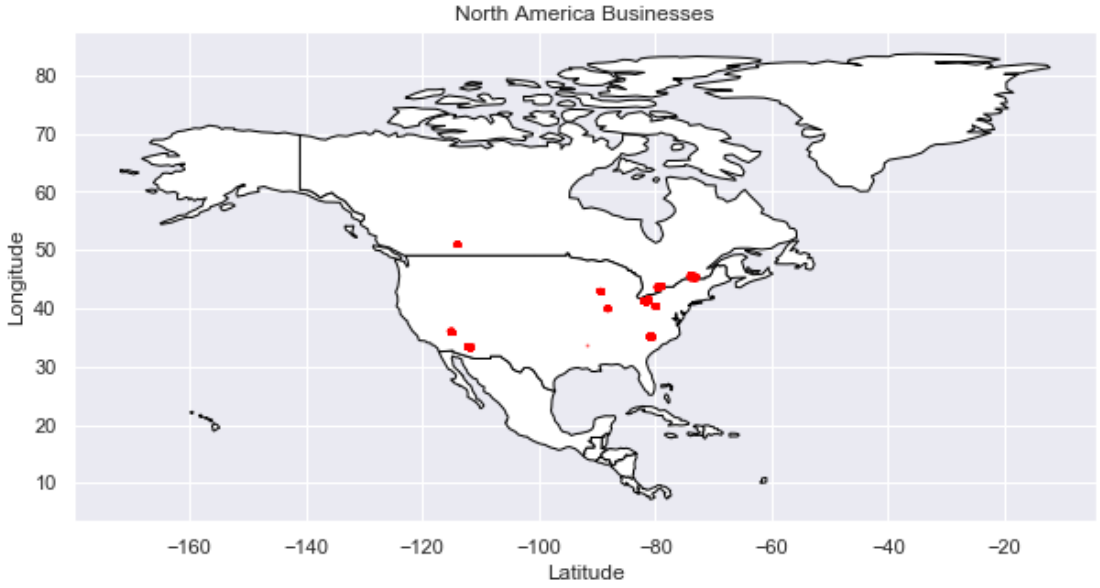
3 Question 2: All businesses

3.1 2.1. What cities does this dataset encompass?

We will graph all of the cities in this dataset on a map of the world and North America, and see which cities have the most businesses within them in this dataset.



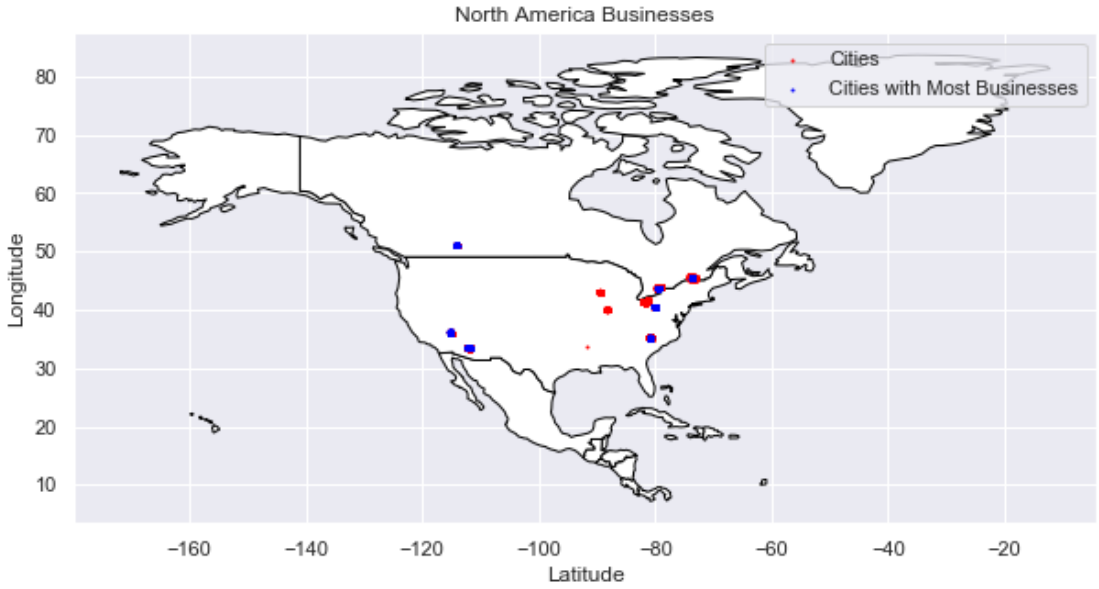
We can see that all of the businesses in the dataset are located in North America.



We can see that all of the businesses in this dataset are concentrated in certain regions of North America. It is not a good representation of businesses across North America, or even USA or Canada. From this, we can infer that the given dataset is most likely not a random sample of businesses on Yelp, but rather the businesses from a few selected big cities across North America.

The 10 cities with the most businesses present within the dataset are:

	City	Business Count	Population
0	Las Vegas	29370	623747
1	Toronto	18906	2826498
2	Phoenix	18766	1563025
3	Charlotte	9509	827097
4	Scottsdale	8837	249950
5	Calgary	7736	1230915
6	Pittsburgh	7017	302407
7	Montreal	6449	1753034
8	Mesa	6080	496401
9	Henderson	4892	302539

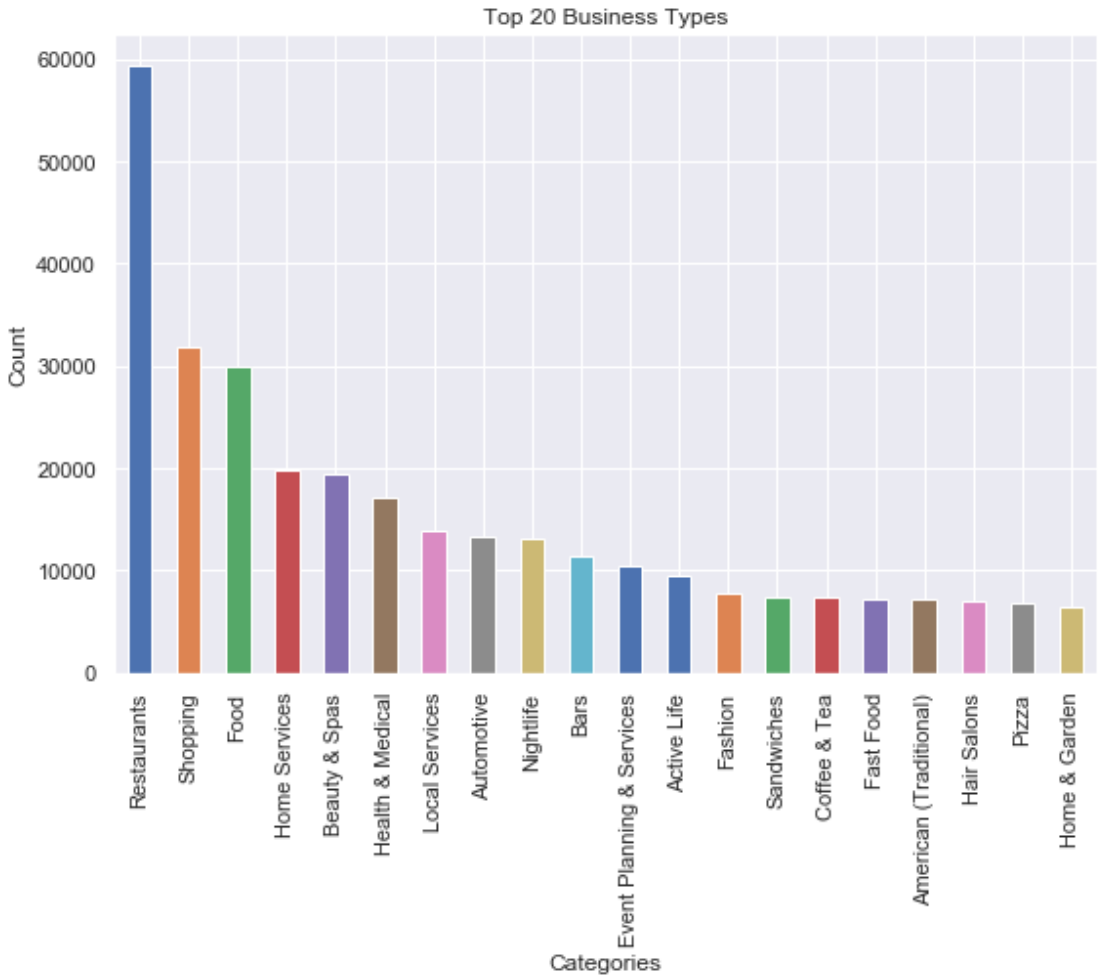


The cities with the most businesses are in the middle of all the concentrated regions of businesses in this dataset. But it doesn't seem like a larger number of businesses is correlated with a higher population in the city, just looking at the cities with the most businesses. The reason for this is Las Vegas, which has an extremely high number of businesses but in comparison to the other top 10 cities, not a very high population. This is mainly just due to Las Vegas being a very tourist oriented city. Therefore higher population does not mean a higher number of Yelp businesses in the city.

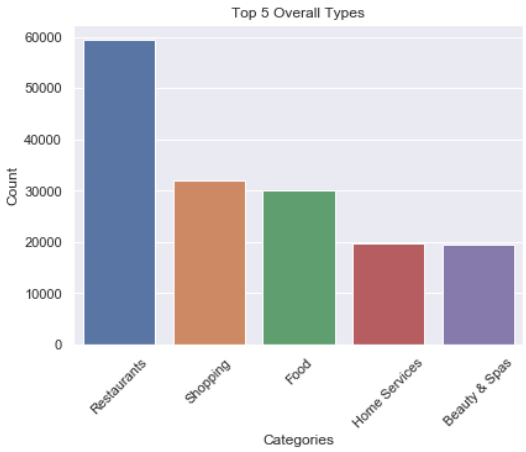
3.2 2.2. What are the most frequent business categories overall?

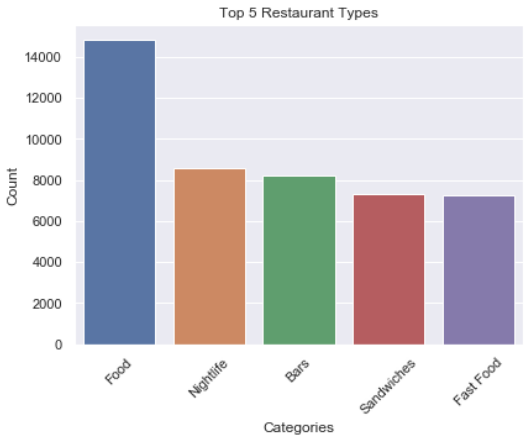
We will observe the top 20 most frequent business categories, and then look at the top five of them in more depth to understand what kind of businesses they detail.

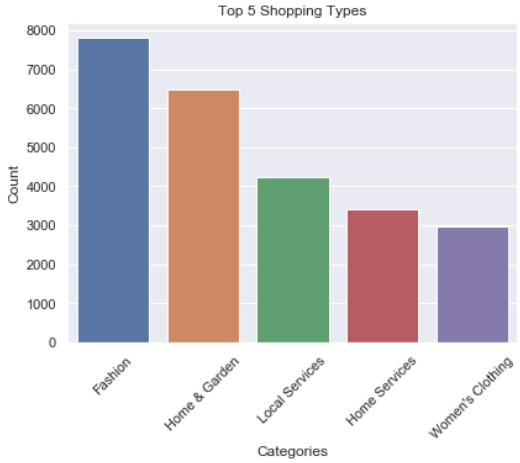
The top 20 most frequent business categories are:

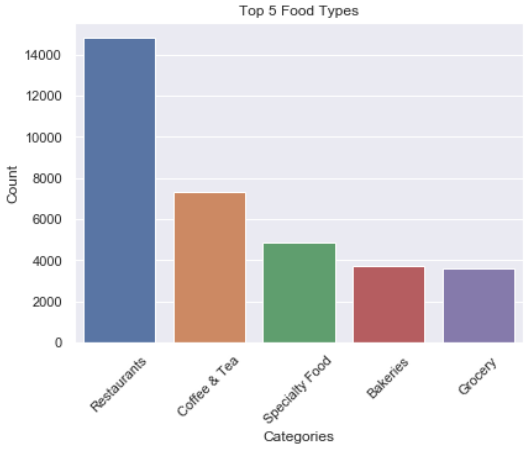


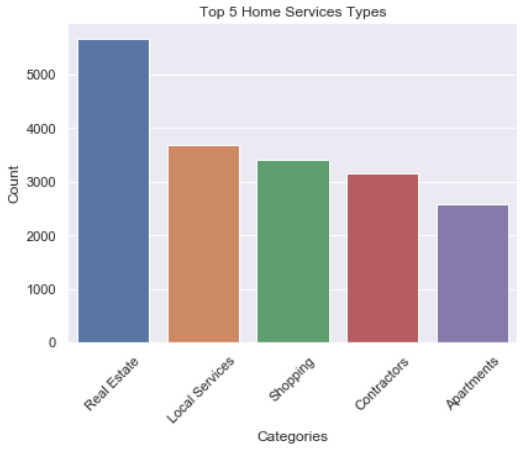
Restaurants are the top business type for businesses in this dataset, with Shopping as second, Food as third, Home Services as fourth, and Beauty and Spas as fifth. These categories are pretty broad, and there is the mystery of why it appears that only half the Restaurants in this dataset serve food. Therefore we will look at these top five categories in some more depth:

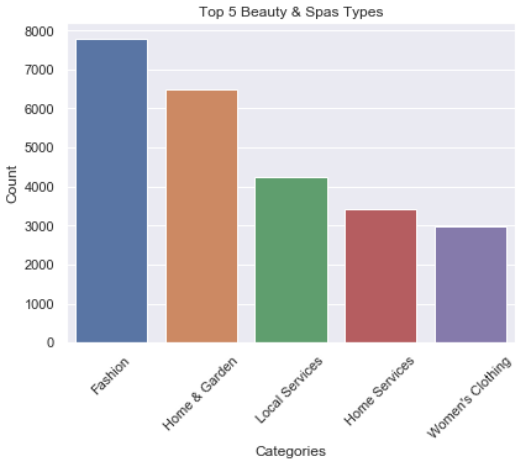

Top 5 Overall Types
Count
Categories
Restaurants
Shopping
Food
Home Services
Beauty & Spas

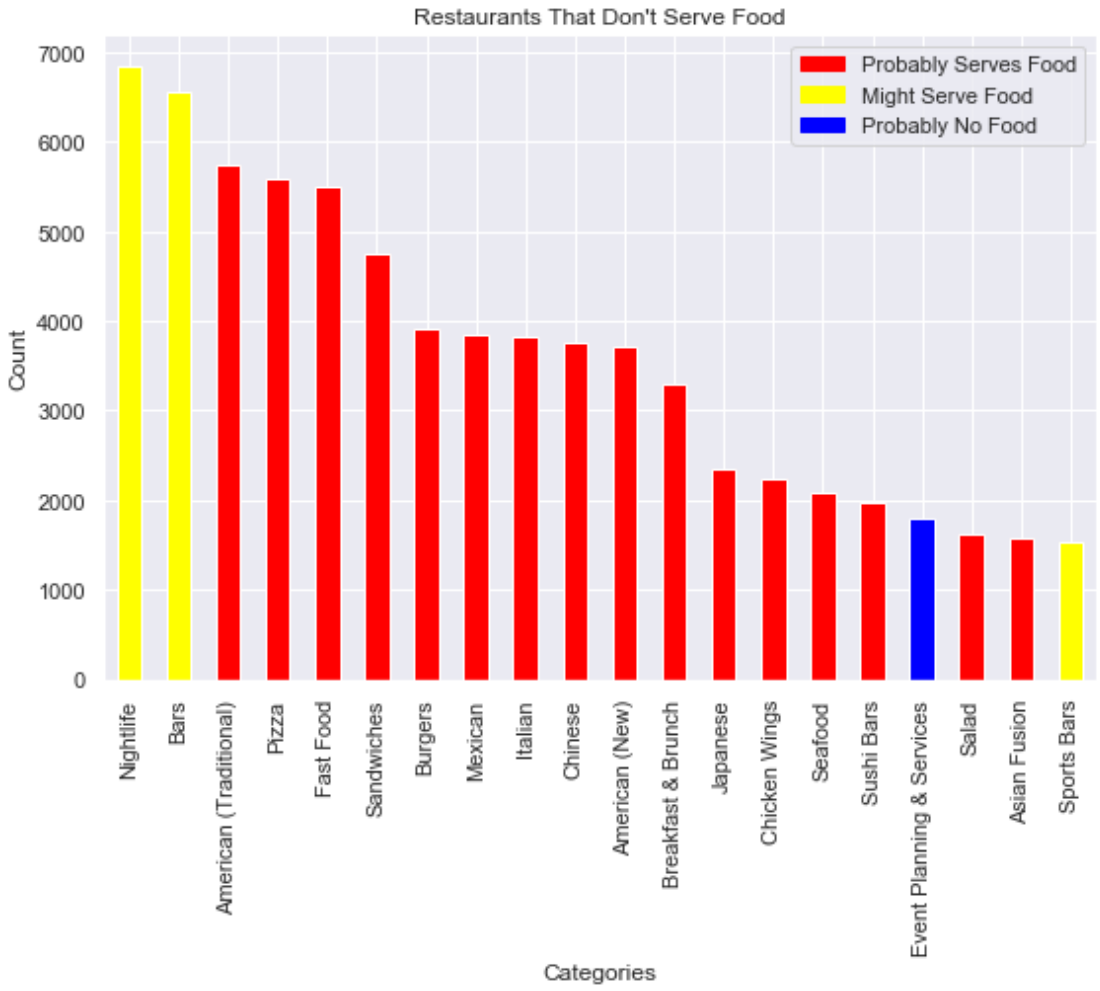

Top 5 Restaurant Types
Count
Categories
Food
Nightlife
Bars
Sandwiches
Fast Food

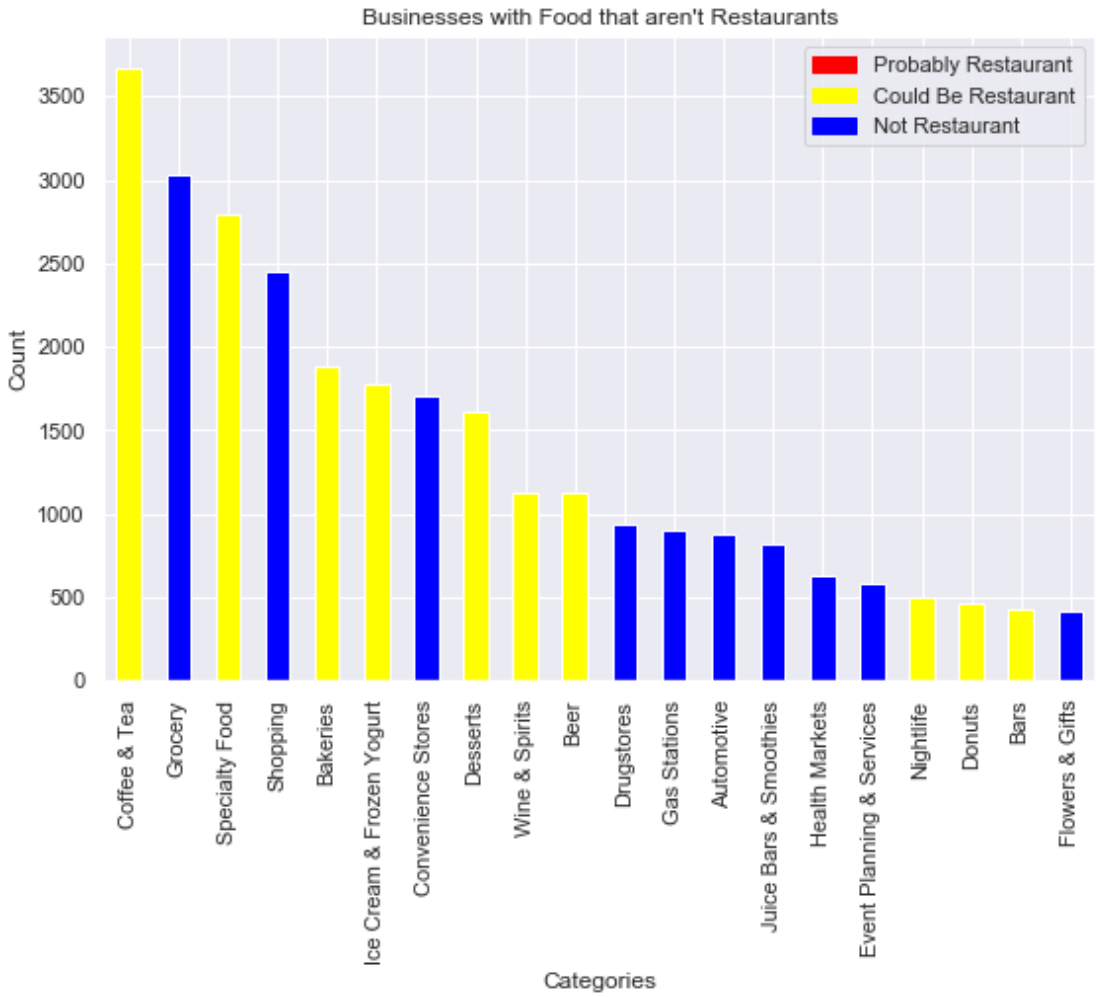

Top 5 Shopping Types
Count
Categories
Fashion
Home & Garden
Local Services
Home Services
Women's Clothing


Top 5 Food Types
Count
Categories
Restaurants
Coffee & Tea
Specialty Food
Bakeries
Grocery


Top 5 Home Services Types
Count
Categories
Real Estate
Local Services
Shopping
Contractors
Apartments


Top 5 Beauty & Spas Types
Count
Categories
Fashion
Home & Garden
Local Services
Home Services
Women's Clothing


Restaurants That Don't Serve Food
Probably Serves Food
Might Serve Food
Probably No Food
Count
0
1000
2000
3000
4000
5000
6000
7000
Nightlife
Bars
American (Traditional)
Pizza
Fast Food
Sandwiches
Burgers
Mexican
Italian
Chinese
American (New)
Breakfast & Brunch
Japanese
Chicken Wings
Seafood
Sushi Bars
Event Planning & Services
Salad
Asian Fusion
Sports Bars
Categories



The categories 'Restaurant' and 'Food' are closely related. The most frequent restaurant types are Nightlife, Bars, Sandwiches, and Fast Food. The most frequent 'Food' types are Coffee & Tea, Specialty Food, Bakeries, and Groceries. We can see that there are definitely places that sell food that are not restaurants, like cafes and grocery stores. However, the distinction for restaurants that don't serve food seems limited to possibly just bars, and even then most bars have a food menu. Then it seems logical to conclude that Restaurant is most likely a subset of Food, and not the other way around.

However, it seems that many business owners who use Yelp haven't realized this distinction for the categories, and many restaurants that do serve food just tag their business with the more specific type of food, rather than the 'Food' category itself. Meanwhile, it would be inaccurate for many places such as grocery stores to tag their business as 'Restaurant', but they definitely qualify under 'Food'. As a result, it may be more accurate to say that 'Food' is the most frequent business category, while Restaurant is second.

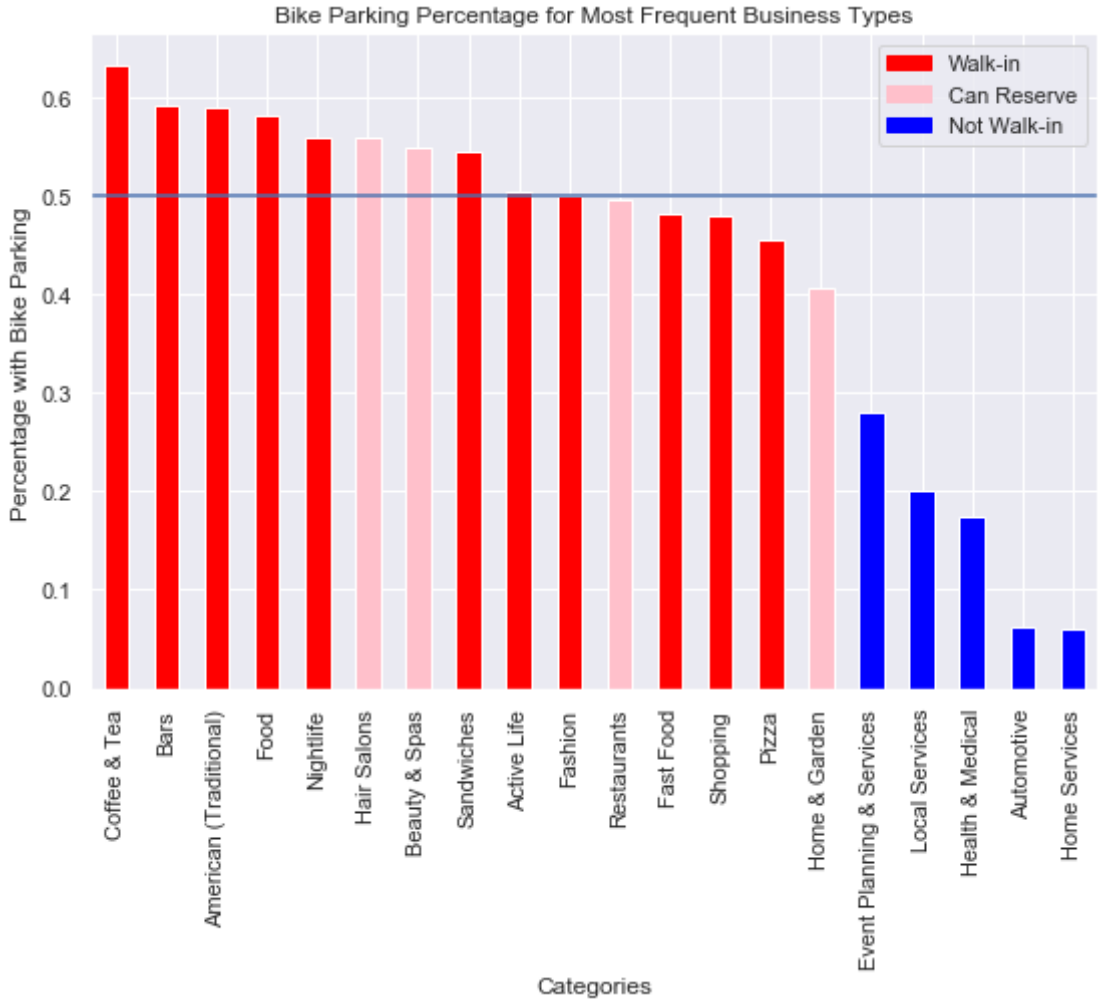
The most frequent 'Shopping' types are Fashion, Home & Garden, Local Services, Home Services, and Women's Clothing.

The most frequent 'Home Services' types are Real Estate, Local Services, Shopping, Contractors, and Apartments. The most frequent type of Home Services seems to refer to finding housing.

The most frequent 'Beauty & Spas' types are Hair Salons, Nail Salons, Hair Removal, Skin Care, and Health & Medical.

3.3 2.3. What types of establishments tend to have bike parking?

We will calculate the percentage of businesses that have bike parking for each category, then look at the top 20 most frequent business categories and compare their percentages. Not all businesses in this dataset have the attribute "Bike Parking", so we assume any place that does not explicitely mention it does not have bike parking. We are looking at the top 20 because if we looked at all categories, the business categories that have very few businesses have a much higher skewed percentage.



We label the categories that most people just walk into to use their services as "Walk-in", the categories that people can walk into, but it's considered good manners to call and reserve beforehand "Can Reserve", and places that most people will call ahead of time to use their services as "Not Walk-in".

Here we can see that among the top 20 most frequent business categories, businesses with categories 'Coffee & Tea', 'Bars', and 'Food' have the highest percentage of bike parking. Over 50% of Coffee & Tea, Bars, American (Traditional), Food, Nightlife, Hair Salons, Beauty & Spas, Sandwiches, and Active Life places in this dataset have bike parking.

As well, we can see that businesses that we can just "Walk-in" to have a much higher percentage of places with bike parking than places without. Places that we "Can Reserve" but don't always do so also have bike parking but at a slightly lower percentage. And places that most people will have to reserve services for beforehand have the lowest percentage of bike parking among these categories. This is probably because if you have to call ahead of time to reserve services, the place may be farther away from where they live so they have to drive to the business, or it's a more formal business that people drive their vehicles to.

3.4 2.4. An article recently claimed that having more yelp reviews lead to a higher rating, and hence increased sales. Do the data support this claim?

We will see if there is an increasing relationship between review count and star rating. We will try a linear regression model, and we will compare the means of the review count for each category of star rating. We use the mean because we do want know which categories tend to have the very large review counts, so we want to take into account large outliers.

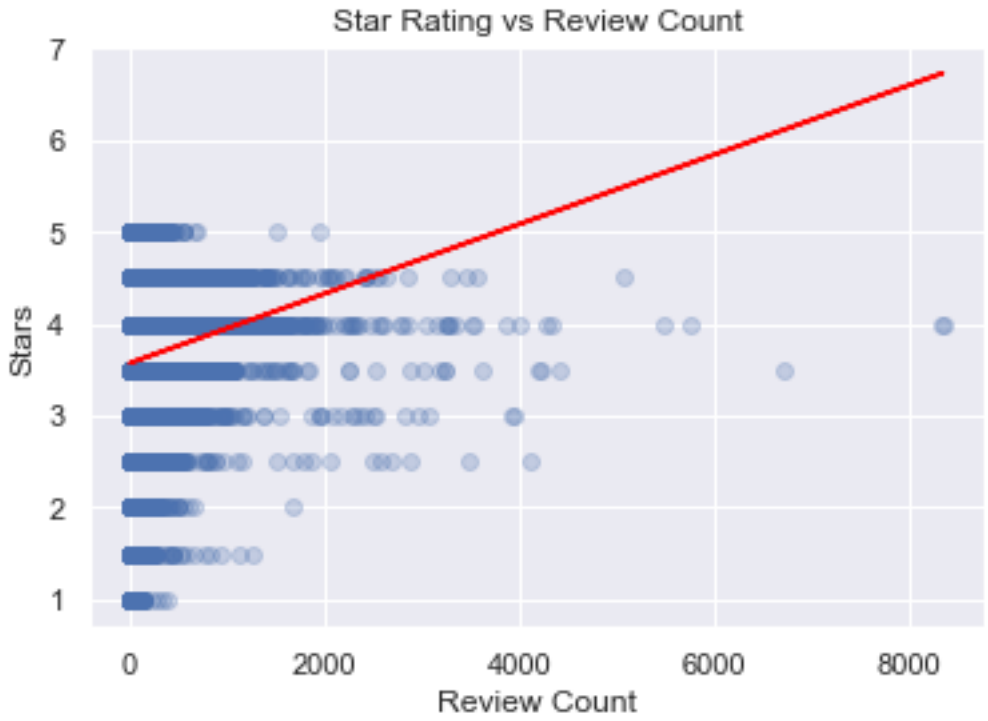
OLS Regression Results

Dep. Variable:	stars	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	321.6
Date:	Sat, 09 Mar 2019	Prob (F-statistic):	7.47e-72
Time:	00:04:20	Log-Likelihood:	-2.7666e+05
No. Observations:	192609	AIC:	5.533e+05
Df Residuals:	192607	BIC:	5.533e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.5730	0.002	1474.101	0.000	3.568	3.578
review_count	0.0004	2.11e-05	17.933	0.000	0.000	0.000

Omnibus:	8390.789	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8539.085
Skew:	-0.481	Prob(JB):	0.00
Kurtosis:	2.630	Cond. No.	120.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



We can see that there is somewhat of a positive relationship between Review Count and Stars. However, a linear regression model doesn't work very well to examine the correlation between review count and star rating for businesses. The line barely fits the data points, and the R-squared value is only 0.002, extremely small. It does imply a positive correlation, but it isn't very trustworthy.

We examine the means of category of star rating, since star rating is finite.

```
stars
1.0     5.815552
1.5    15.596664
2.0    15.108874
2.5    20.910630
3.0    30.857286
3.5    40.681130
4.0    56.523228
4.5    43.444453
5.0    12.113942
Name: review_count, dtype: float64
```

We test to see if the mean of review counts is significantly different for different star ratings, by testing the means of each category and the one 0.5 stars higher.

```
t-test p-value for stars 1.0 and 1.5: 5.03200542978842e-50
t-test p-value for stars 1.5 and 2.0: 0.4827022519507418
```

```
t-test p-value for stars 2.0 and 2.5: 1.4649101598717048e-20
t-test p-value for stars 2.5 and 3.0: 1.1008381059083963e-36
t-test p-value for stars 3.0 and 3.5: 8.088016636257389e-33
t-test p-value for stars 3.5 and 4.0: 1.3890578175346252e-48
t-test p-value for stars 4.0 and 4.5: 1.5014503528297853e-28
t-test p-value for stars 4.5 and 5.0: 0.0


t-test p-value for mean of businesses with stars below 3.5 and above 3.5: 0.0
```

By our t-test results, we conclude that there is a significant increase in the mean of review counts between increasing star ratings for businesses, except between star rating 1.5 and 2.0, between star ratings 0 to 4. Then a higher review count leads to a higher star count when the star rating is between 1 to 4. Above 4.0, having a higher review count seems to lead to a lower star rating. This could be thought of as if the business has such a high rating that people are no longer concerned with leaving reviews for the business.

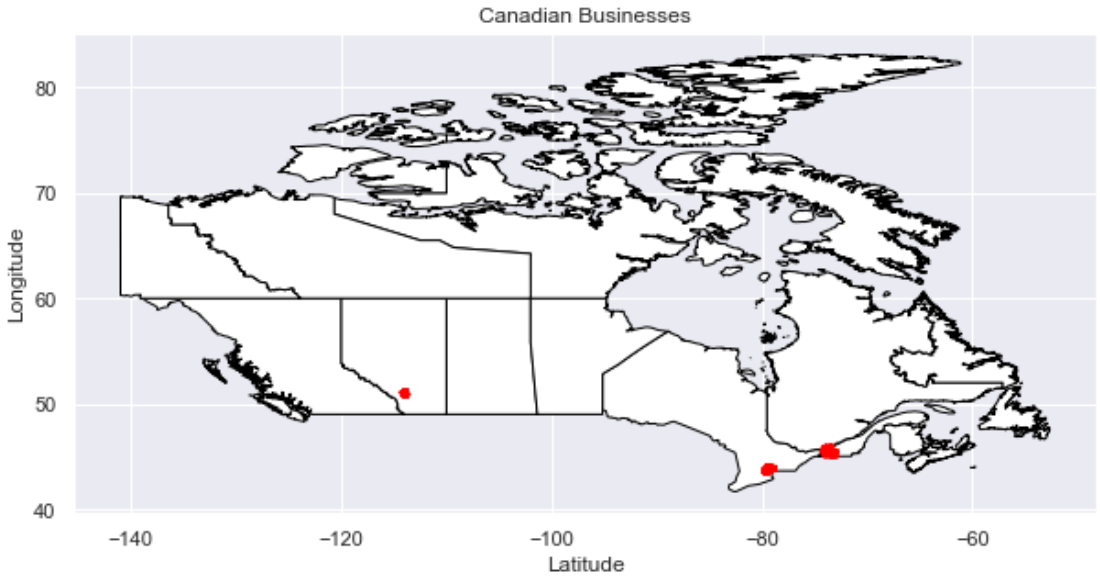
As well, there is a difference in the mean of review counts between businesses that have a star count larger or equal to 3.5, and businesses that have a star count below 3.5. Then we can conclude that businesses with higher review counts tend to have higher star ratings.

4 Question 3: Canadian businesses

4.1 3.1. What cities does this dataset encompass?

We will graph all the cities in Canada present in this dataset, as well as note the cities in each province that have the most businesses present in them.

Provincial map data taken from here



	Alberta Cities	Alberta Counts	Ontario Cities	Ontario Counts
0	Calgary	7745	Toronto	19705
1	Airdrie	168	Mississauga	3117
2	Chestermere	31	Markham	1768
3	Rocky View	20	North York	1210
4	Balzac	11	Scarborough	1106
5	Rocky View County	9	Richmond Hill	1028
6	Rocky View No. 44	4	Brampton	1002
7	Division No. 6	3	Vaughan	922
8	Rockyview	2	Thornhill	397
9	Edmonton	2	Oakville	362

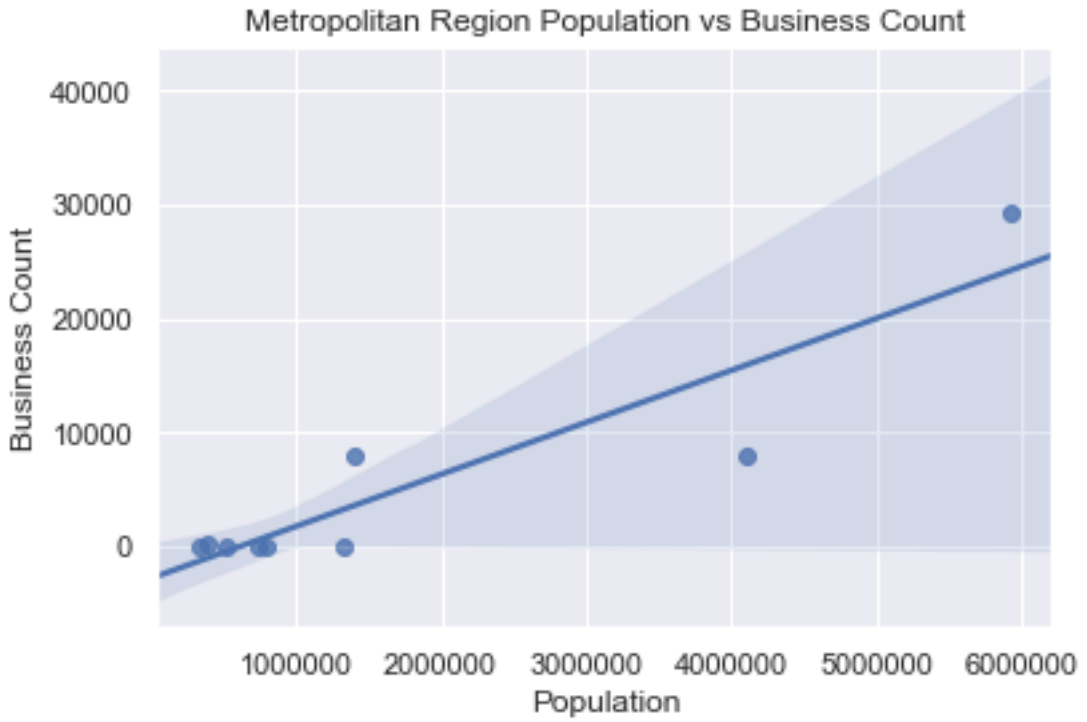
	Quebec Cities	Quebec Counts
0	Montreal	6601
1	Laval	394
2	Brossard	177
3	Verdun	153
4	Saint-Laurent	136
5	Pointe-Claire	102
6	Dorval	97
7	Westmount	97
8	Dollard-des-Ormeaux	87
9	Longueuil	78

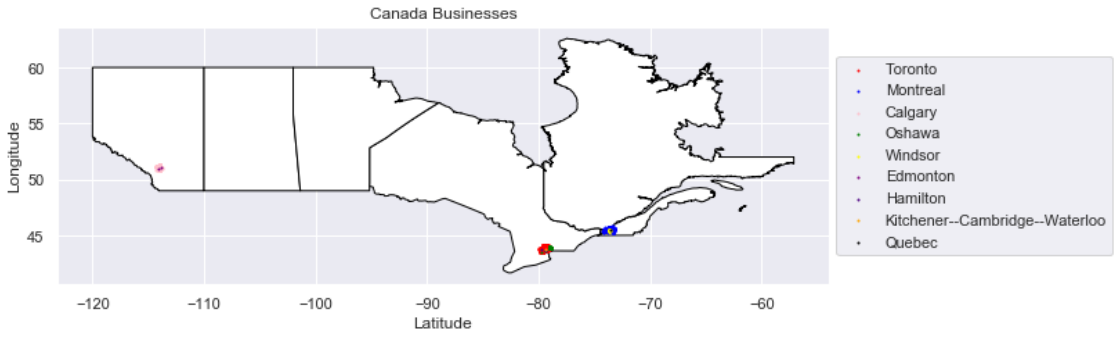
From the map, we can see that the Canadian businesses in this dataset are all from three concentrated regions in Alberta, Ontario, and Quebec. The table shows the top 10 cities with the most businesses in the dataset for each of the three provinces. The overwhelming majority of businesses are from Ontario, and then Quebec, and then Alberta.

4.2 3.2. Identify the larger metropolitan regions that these cities belong to.

We will gather information about the metropolitan regions of canada from the Statistics Canada website, then use it to create a new variable in the dataframe that represents the region the business is in.

	Area	Business Count	Population
0	Toronto	29315	5928040
1	Montreal	8051	4098927
2	Calgary	7953	1392609
3	Oshawa	279	379848
4	Windsor	19	329144
5	Edmonton	2	1321426
6	Hamilton	2	747545
7	Kitchener–Cambridge–Waterloo	1	523894
8	Quebec	1	800296





There are 9 metropolitan regions that can be identified in this dataset, spread across three provinces. The largest is the Greater Toronto Area, and the smallest is the Quebec and Waterloo regions.

Ontario has five metropolitan regions in this dataset: the GTA, Oshawa, Windsor, Hamilton, and Kitchener--Cambridge--Waterloo.

Quebec has two metropolitan regions in this dataset: Montreal and Quebec.

Alberta has two metropolitan regions in this dataset: Calgary and Edmonton.

While Toronto, Montreal, and Calgary have the overall largest number of businesses and largest population in their area, the remaining areas don't have a visible trend between the number of businesses and the population in their area. It is likely then that Yelp chose to include businesses in the Toronto, Montreal, and Calgary areas, along with more businesses close to these areas for this dataset.

5 Question 4: GTA businesses

5.1 4.1. What are the most frequent business categories? How do they compare against the trends listed in 2.2?

We will count the most frequent business categories in the GTA region, and compare them with the overall most frequent business categories

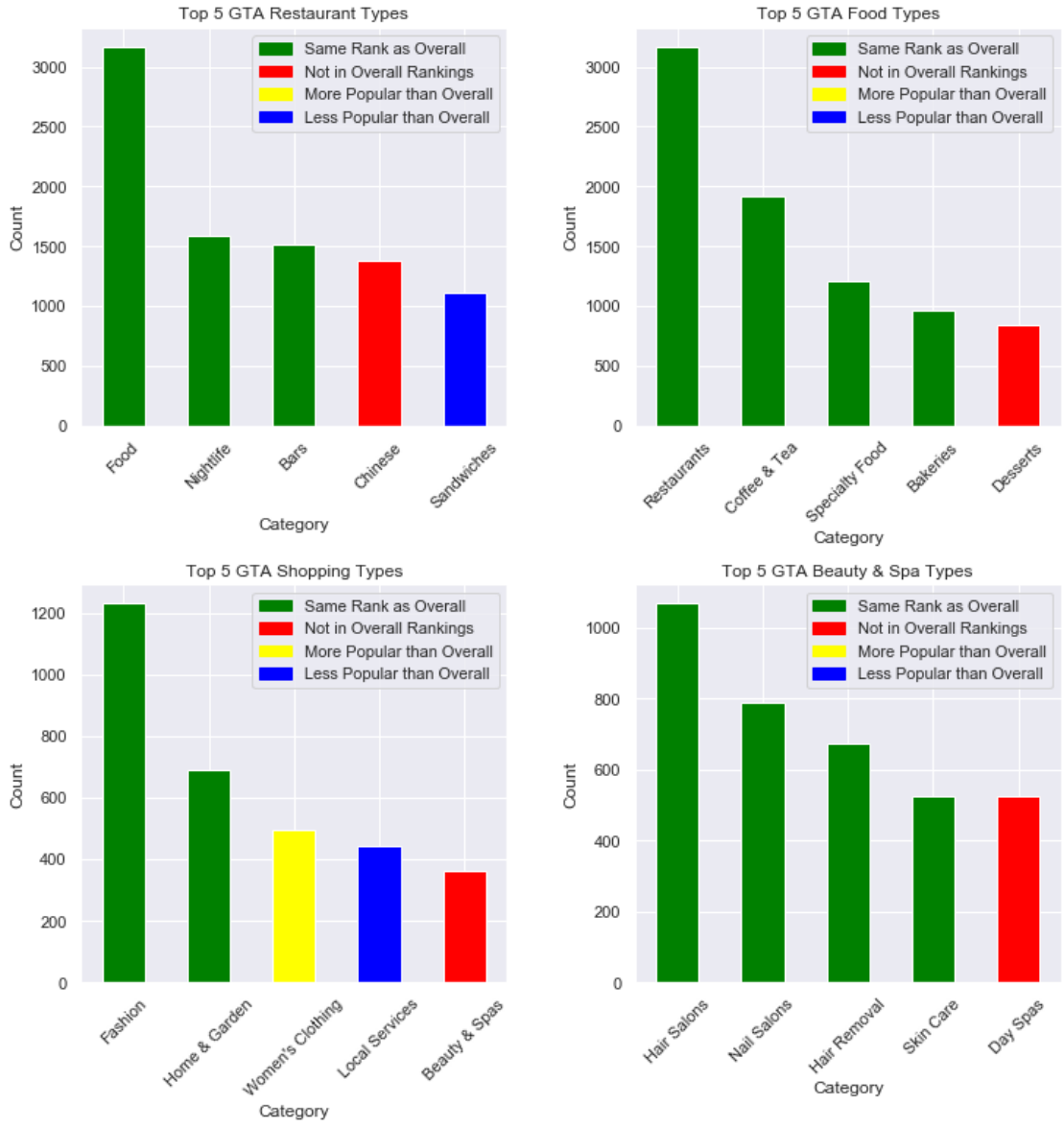
	GTA Top Categories	Overall Top Categories
0	Restaurants	Restaurants
1	Food	Shopping
2	Shopping	Food
3	Beauty & Spas	Home Services
4	Nightlife	Beauty & Spas
5	Bars	Health & Medical
6	Coffee & Tea	Local Services
7	Health & Medical	Automotive
8	Event Planning & Services	Nightlife
9	Chinese	Bars

We can see that Restaurants, Food, Shopping, Beauty & Spas, Health & Medical, Nightlife, and Bars are in the top 10 most frequent business categories for both the GTA and overall, although Nightlife and Bars are more frequent than Health & Medical in the GTA, while the opposite holds overall.

In the GTA, Coffee & Tea, Event Planning & Services, and Chinese are some of the most frequent business categories that aren't part of the top 10 overall.

Overall, Home Services, Local Services, Automotive are more prominent than in the GTA.

We will look at the most popular subcategories for Restaurants, Food, Shopping, and Beauty & Spas to see if there is a difference in their distribution from the overall businesses.

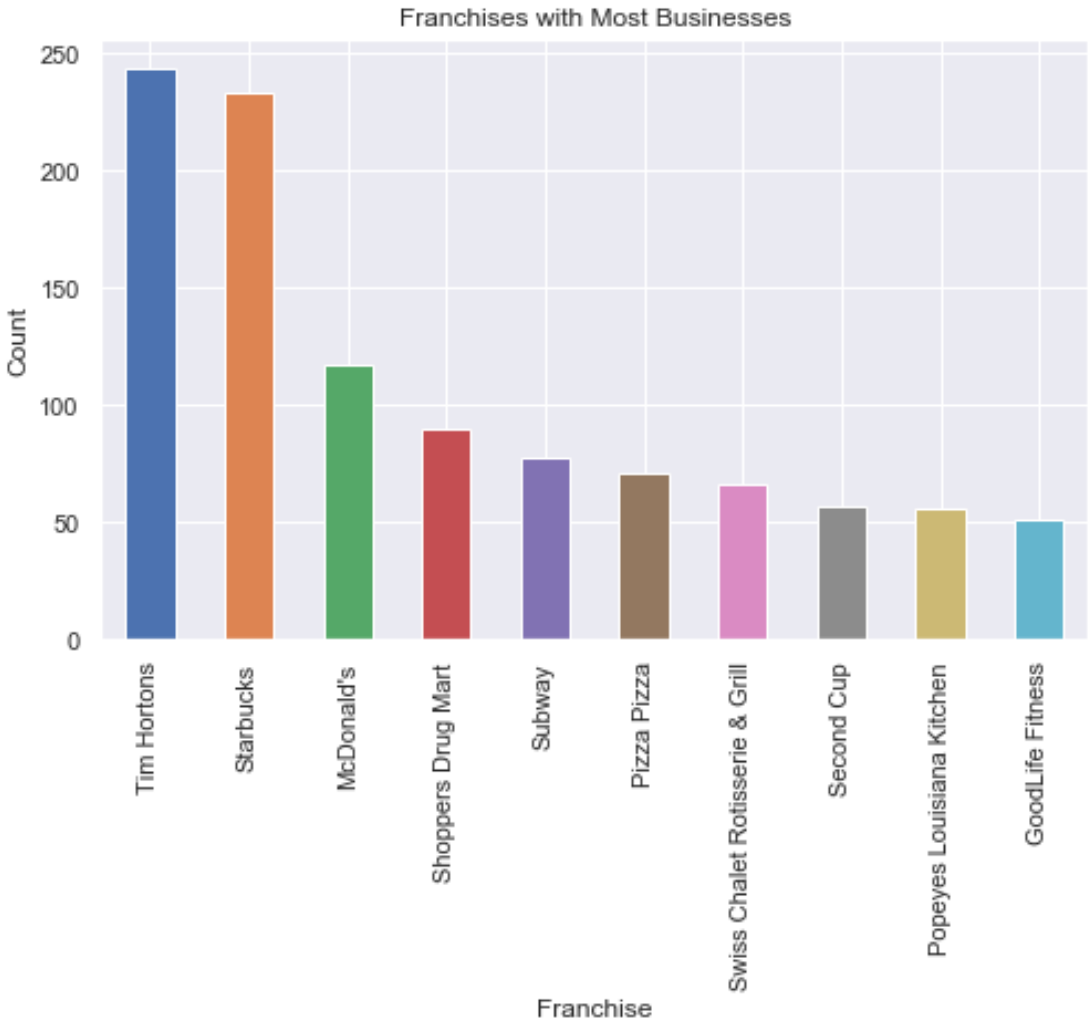



```
The number of businesses overall is: 192609
The number of businesses in the GTA is: 29315
The proportion of businesses that are in the GTA is: 0.1521995337704884
```

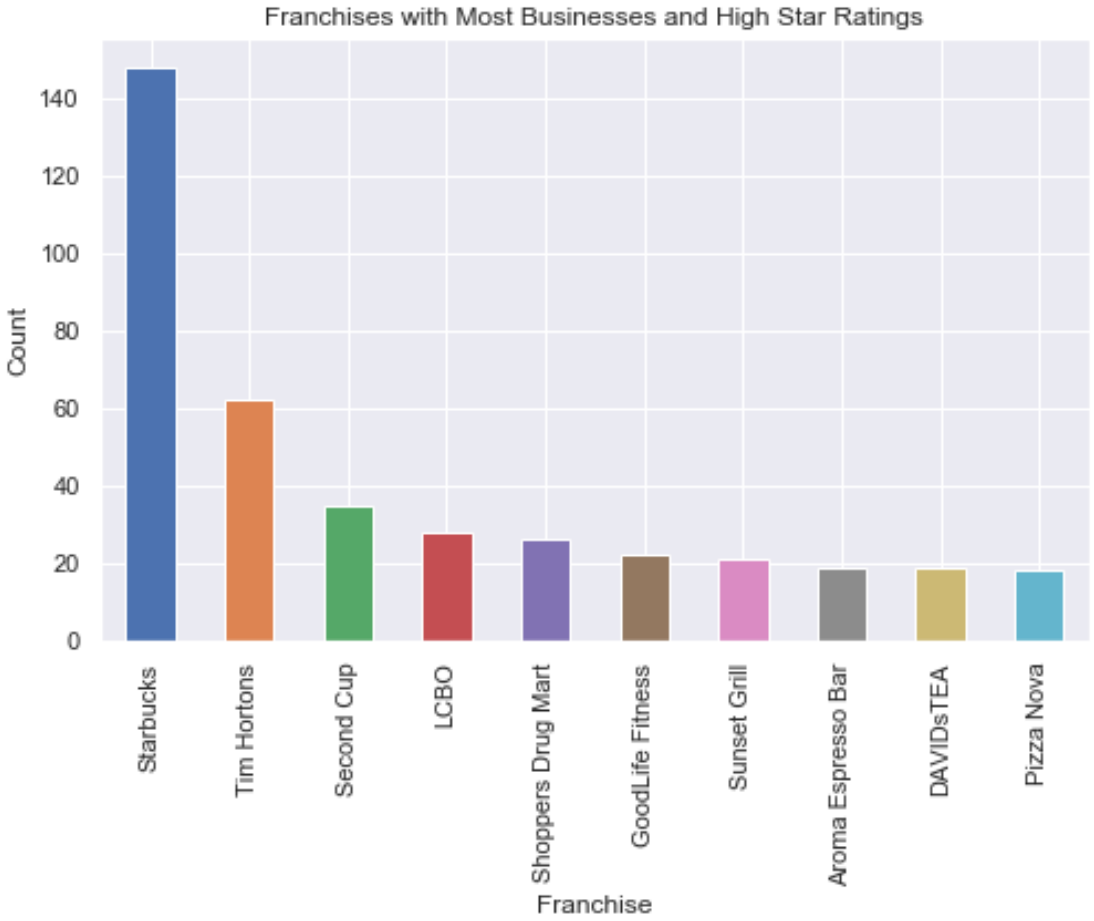
Looking at the graph of the top subcategories in each of the four largest categories in the GTA, we can see that the most popular subcategories are very similar, with the lower ranked categories varying slightly. The first and second most popular subcategories are identical for all four of these categories. The lower three are slightly different from the Overall most popular subcategories, and there's always one subcategory in the top 5 for GTA that isn't in the Overall top 5.

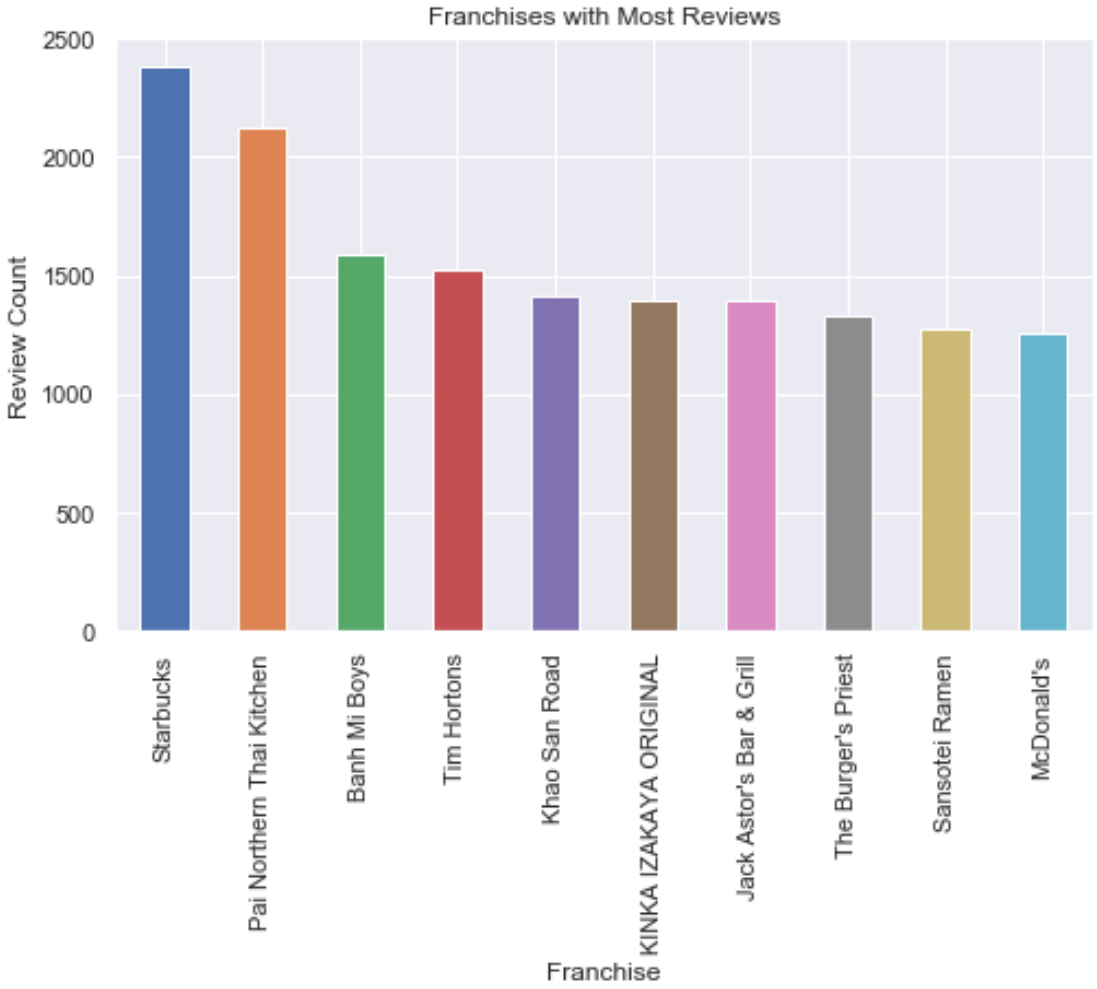
5.2 4.2. What are the top franchises in the city?

We will count the top 10 franchises that have the most businesses present in the GTA, the top 10 franchises that have the most businesses and also have star ratings 3.5 or above, and the top 10 franchises with the most reviews overall.



Notably, all of these businesses except for Shoppers Drug Mart, Swiss Chalet, and GoodLife Fitness are fast food/drink places.


Franchises with Most Businesses and High Star Ratings
Count
0
20
40
60
80
100
120
140
Starbucks
Tim Hortons
Second Cup
LCBO
Shoppers Drug Mart
GoodLife Fitness
Sunset Grill
Aroma Espresso Bar
DAVIDsTEA
Pizza Nova
Franchise



	Top By Amount	Top Rated By Amount	Top By Reviews
0	Tim Hortons	Starbucks	Starbucks
1	Starbucks	Tim Hortons	Pai Northern Thai Kitchen
2	McDonald's	Second Cup	Banh Mi Boys
3	Shoppers Drug Mart	LCBO	Tim Hortons
4	Subway	Shoppers Drug Mart	Khao San Road
5	Pizza Pizza	GoodLife Fitness	KINKA IZAKAYA ORIGINAL
6	Swiss Chalet Rotisserie & Grill	Sunset Grill	Jack Astor's Bar & Grill
7	Second Cup	Aroma Espresso Bar	The Burger's Priest
8	Popeyes Louisiana Kitchen	DAVIDsTEA	Sansotei Ramen
9	GoodLife Fitness	Pizza Nova	McDonald's

When looking at the top 10 franchises just by the amount of businesses that they have in the Greater Toronto Area, Tim Hortons is first with Starbucks second by a small difference, and both of these two franchises have much more businesses than the other eight following franchises. The large majority of these businesses are also fast food/drink locations.

When we look at the top 10 franchises that also have a star rating equal or above 3.5, Starbucks

has an overwhelming lead over Tim Hortons and all other franchises.

As well, when looking at the top 10 franchises by number of reviews for all their businesses, Starbucks is still in the lead, and many new franchises get second and third, while Tim Hortons is only fourth.

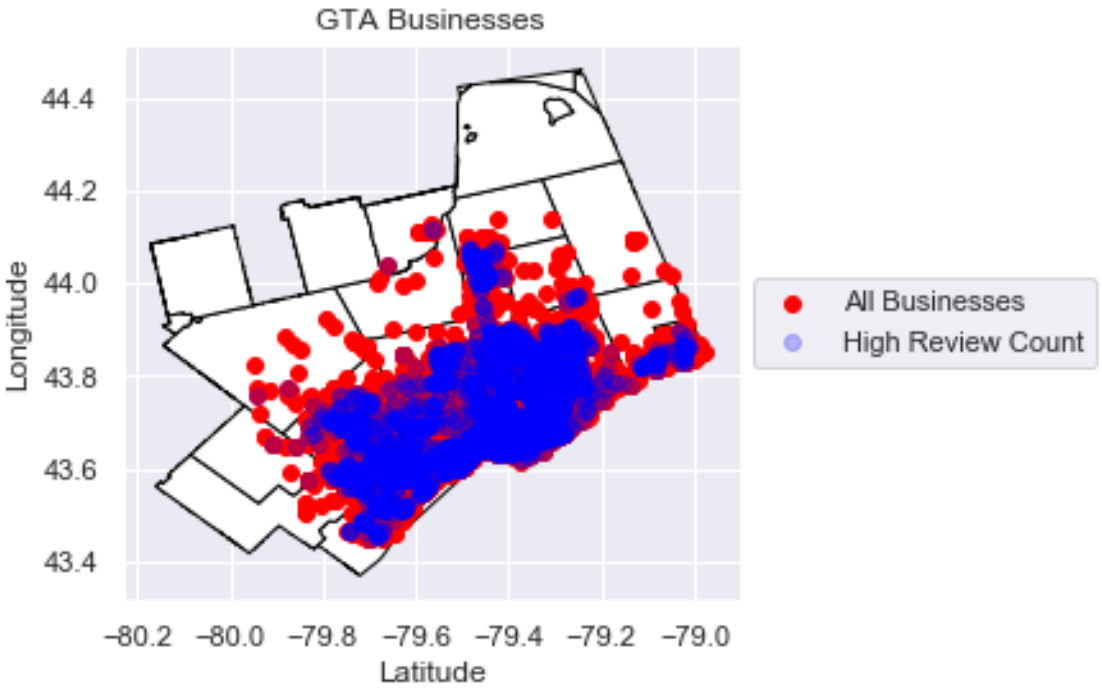
Starbucks seems to be the most consistently popular across these three ranking methods. Tim Hortons has more franchises, but it seems to be of lesser quality than the Starbucks in the GTA, since it has less higher ranked businesses and less reviews.

McDonalds also has many franchises, but isn't even in the top ten for businesses with high ratings, and is ranked tenth for review count. This does seem to imply that while McDonalds is a large franchise, it isn't of very good quality, service or foodwise.

5.3 4.3. Does business location play an important role in reviews?

We will investigate where businesses with high review counts are located in relation to all businesses, and businesses with high ratings are located in relation to all the businesses, since high ratings are correlated with high review count. We will also investigate where businesses with popular reviews are located in relation to all the businesses.

Greater Toronto Area map data taken from here

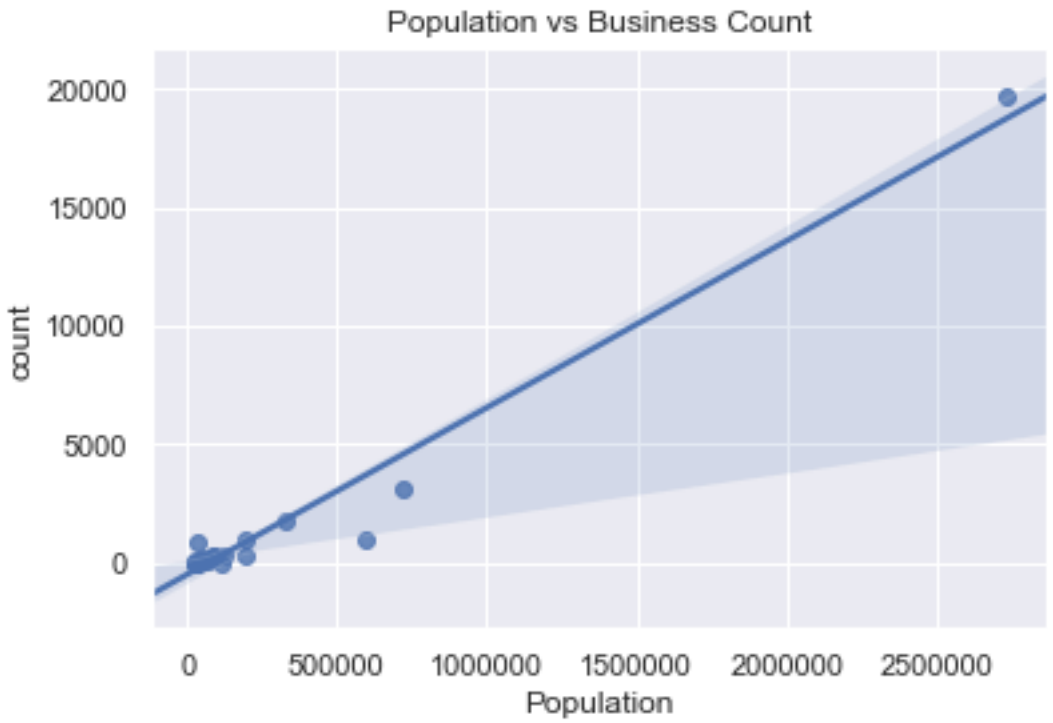


We can see that businesses with high review counts are concentrated in areas with lots of businesses.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  count   R-squared:                       0.965
```

```
Model:                            OLS   Adj. R-squared:                  0.963
Method:                 Least Squares   F-statistic:                     491.3
Date:                Sat, 09 Mar 2019   Prob (F-statistic):           1.62e-14
Time:                        00:39:56   Log-Likelihood:                -162.06
No. Observations:                  20   AIC:                             328.1
Df Residuals:                      18   BIC:                             330.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -498.1047    208.216     -2.392      0.028    -935.550     -60.660
Population     0.0070      0.000     22.165      0.000       0.006       0.008
==============================================================================
Omnibus:                       21.475   Durbin-Watson:                   2.167
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.823
Skew:                          -1.949   Prob(JB):                     9.08e-07
Kurtosis:                       7.266   Cond. No.                     7.24e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.24e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```



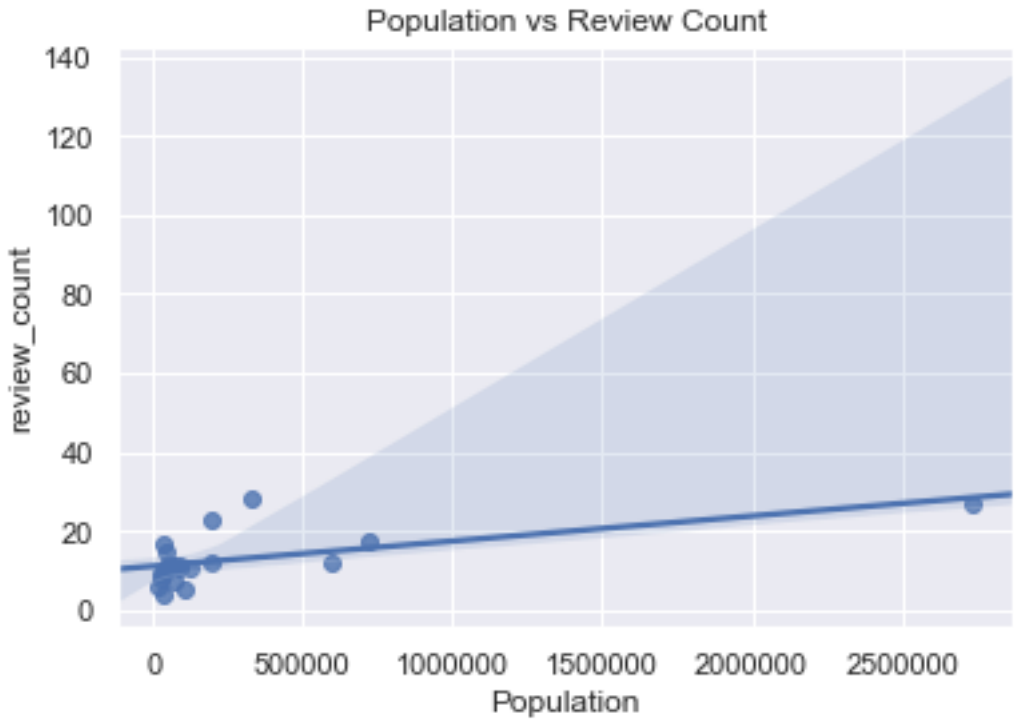
OLS Regression Results

Dep. Variable:	review_count	R-squared:	0.340
Model:	OLS	Adj. R-squared:	0.304
Method:	Least Squares	F-statistic:	9.285
Date:	Sat, 09 Mar 2019	Prob (F-statistic):	0.00694
Time:	00:39:57	Log-Likelihood:	-61.526
No. Observations:	20	AIC:	127.1
Df Residuals:	18	BIC:	129.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.0408	1.366	8.081	0.000	8.171	13.911
Population	6.356e-06	2.09e-06	3.047	0.007	1.97e-06	1.07e-05

Omnibus:	10.490	Durbin-Watson:	1.048
Prob(Omnibus):	0.005	Jarque-Bera (JB):	7.860
Skew:	1.325	Prob(JB):	0.0196
Kurtosis:	4.552	Cond. No.	7.24e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.24e+05. This might indicate that there are
strong multicollinearity or other numerical problems.



OLS Regression Results

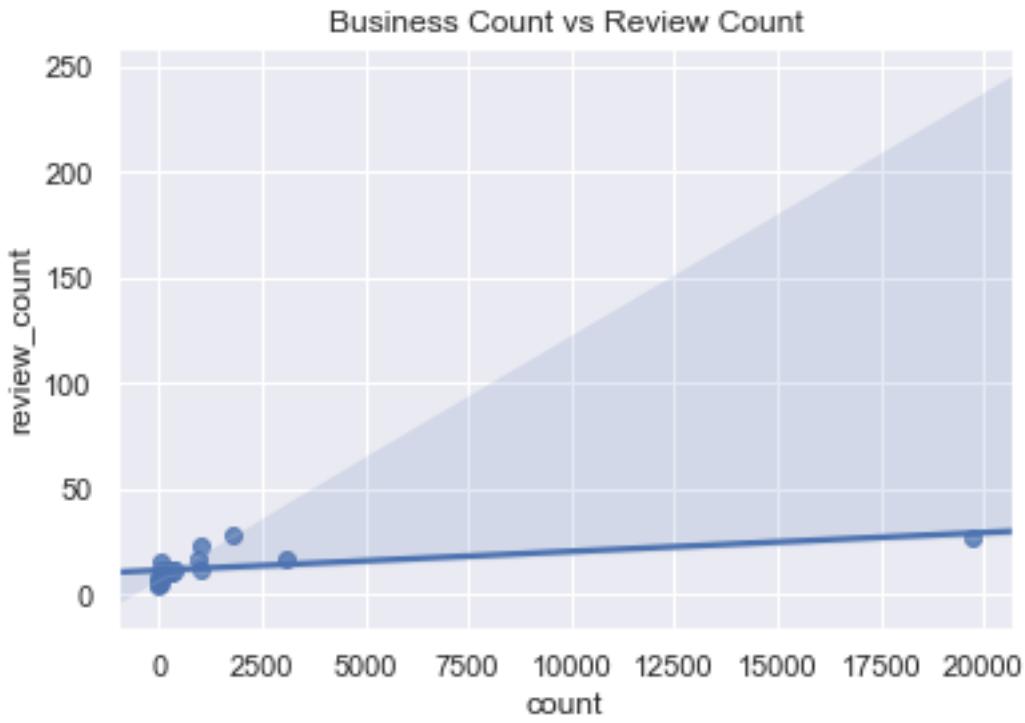
Dep. Variable:	review_count	R-squared:	0.344
Model:	OLS	Adj. R-squared:	0.308
Method:	Least Squares	F-statistic:	9.445
Date:	Sat, 09 Mar 2019	Prob (F-statistic):	0.00655
Time:	00:39:57	Log-Likelihood:	-61.467
No. Observations:	20	AIC:	126.9
Df Residuals:	18	BIC:	128.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.5063	1.304	8.825	0.000	8.767	14.246
count	0.0009	0.000	3.073	0.007	0.000	0.002

Omnibus:	10.312	Durbin-Watson:	1.011
Prob(Omnibus):	0.006	Jarque-Bera (JB):	7.675
Skew:	1.290	Prob(JB):	0.0215
Kurtosis:	4.599	Cond. No.	4.76e+03

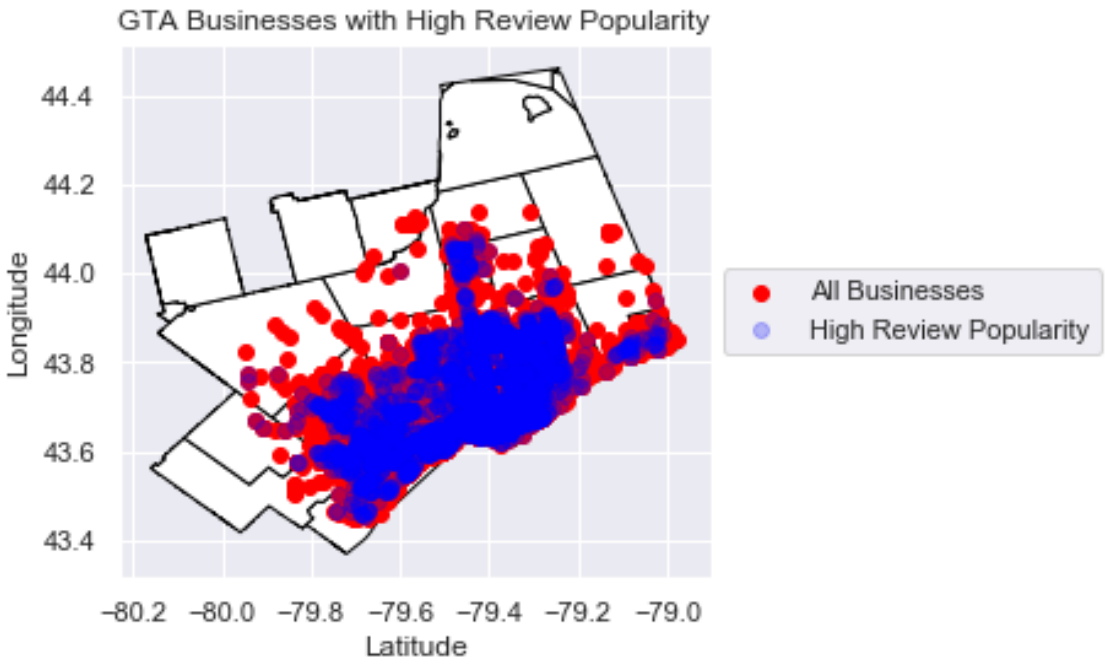
Warnings:

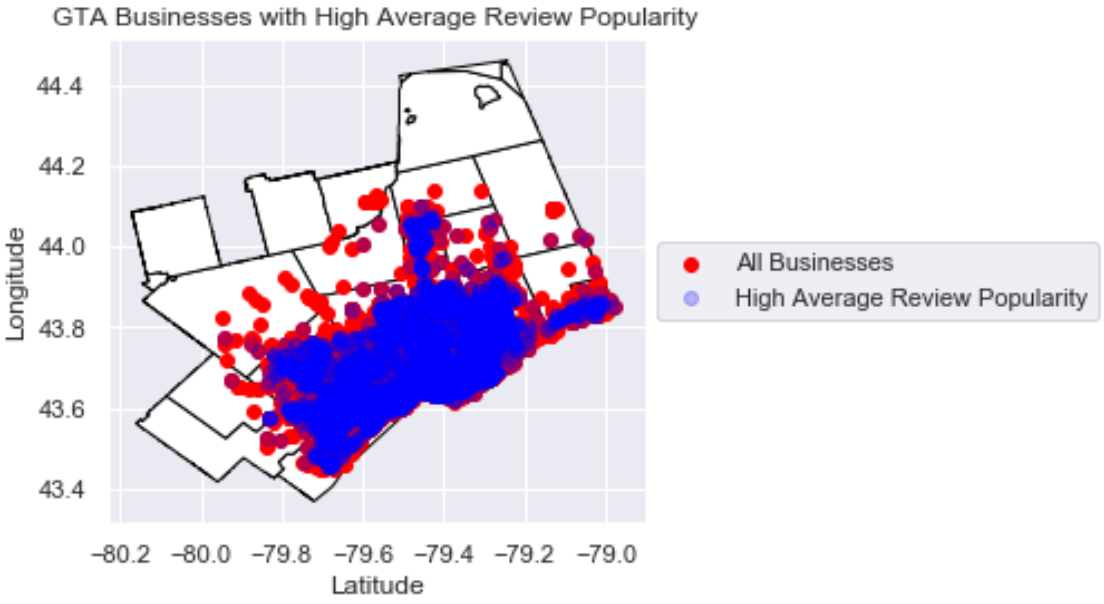
```
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.76e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



While there is a very good correlation between the population of a city and the amount of businesses in this dataset, the correlation between average review count and number of businesses is much weaker. However, even though its R-squared value is just 0.340 and 0.344 for the linear regression between amount of businesses and review count, and population and review count respectively, the confidence interval still implies a very strong possibility of there being a positive correlation between them.

Therefore location most likely does affect the review count of businesses. Cities with higher populations and higher number of businesses tend to have slightly higher average review counts on their businesses. This could be implying that a higher population in a city leads to more people reviewing the businesses there.





OLS Regression Results

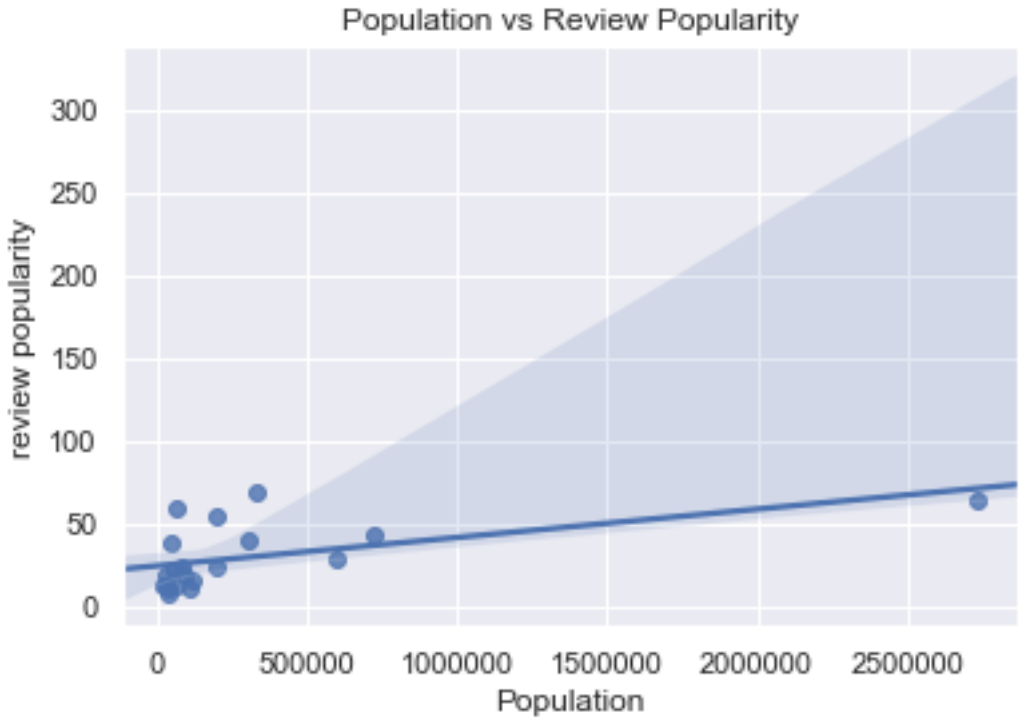
Dep. Variable:	review popularity	R-squared:	0.279
Model:	OLS	Adj. R-squared:	0.238

Method:	Least Squares	F-statistic:	6.949
Date:	Sat, 09 Mar 2019	Prob (F-statistic):	0.0168
Time:	00:47:51	Log-Likelihood:	-84.099
No. Observations:	20	AIC:	172.2
Df Residuals:	18	BIC:	174.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	24.4521	4.265	5.733	0.000	15.491	33.413
Population	1.708e-05	6.48e-06	2.636	0.017	3.47e-06	3.07e-05

Omnibus:	6.284	Durbin-Watson:	0.879
Prob(Omnibus):	0.043	Jarque-Bera (JB):	4.607
Skew:	1.171	Prob(JB):	0.0999
Kurtosis:	3.208	Cond. No.	7.35e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.35e+05. This might indicate that there are
strong multicollinearity or other numerical problems.



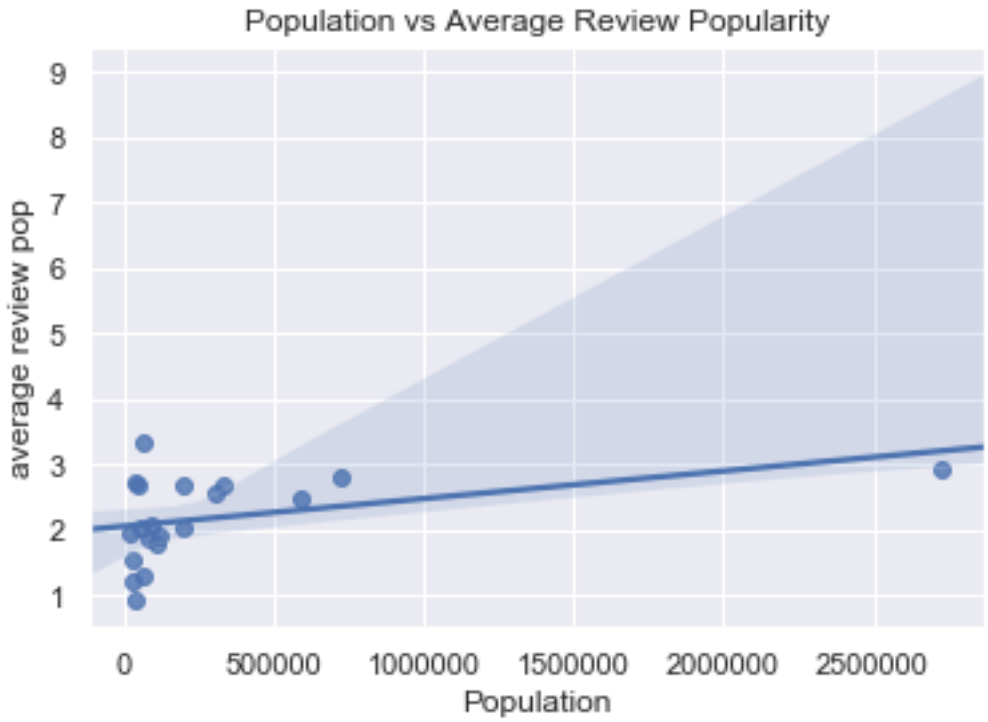
OLS Regression Results

Dep. Variable:	average review pop	R-squared:	0.160
Model:	OLS	Adj. R-squared:	0.114
Method:	Least Squares	F-statistic:	3.434
Date:	Sat, 09 Mar 2019	Prob (F-statistic):	0.0803
Time:	00:47:51	Log-Likelihood:	-17.122
No. Observations:	20	AIC:	38.24
Df Residuals:	18	BIC:	40.23
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.0622	0.150	13.765	0.000	1.747	2.377
Population	4.216e-07	2.28e-07	1.853	0.080	-5.64e-08	9e-07

Omnibus:	0.070	Durbin-Watson:	1.431
Prob(Omnibus):	0.966	Jarque-Bera (JB):	0.071
Skew:	0.045	Prob(JB):	0.965
Kurtosis:	2.723	Cond. No.	7.35e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.35e+05. This might indicate that there are
strong multicollinearity or other numerical problems.



We define "Review Popularity" as the number of 'useful', 'funny', and 'cool' votes given to the review, and the business's overall review popularity as the sum of all its reviews' popularities. This is a measure of how often users will seriously read other user's reviews given to this business. Looking at the graph of the locations of businesses that have high review popularities, and the graph of the locations of businesses that have a high average review popularity, we see that these businesses are clustered in the middle of where all the businesses are located.

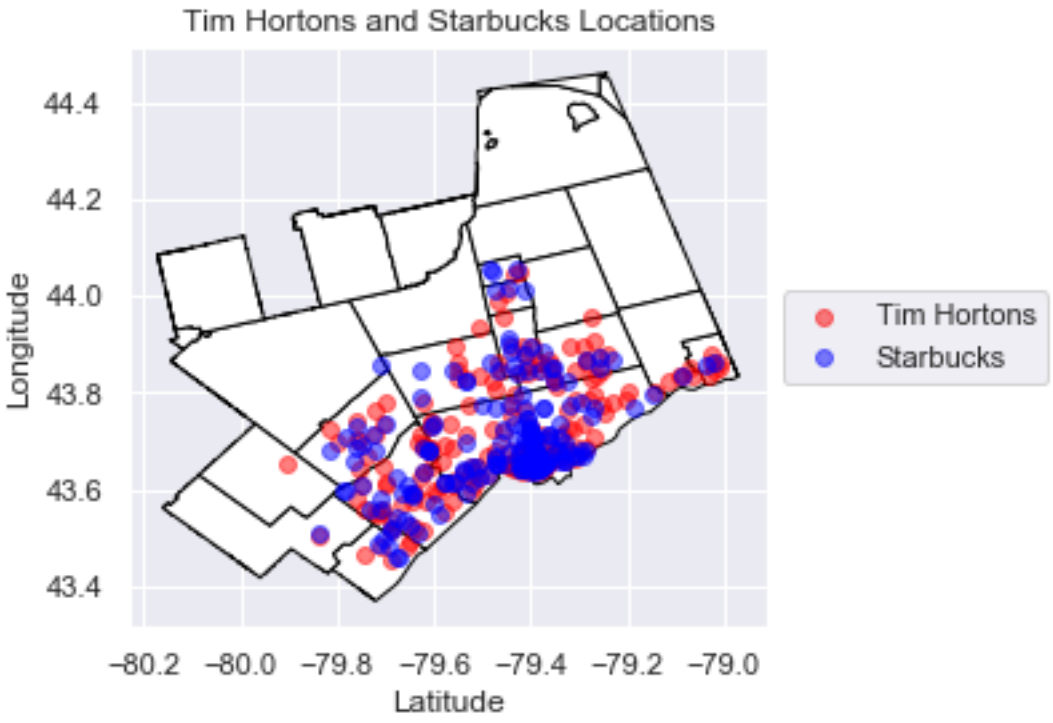
We also conducted a linear regression for the population of a city and the review popularity of a business, as well as the population of a city vs the average review popularity of a business. From the linear regression, we see that there is a somewhat positive correlation between the review popularity of a business and the population of the city its in.

However, there doesn't seem to be a conclusive correlation between the average review popularity and the population of the city the business is in. The R-squared value is very small at 0.160, we can see that the regression line doesn't fit the data very well, and the confidence interval for the slope has both negative and positive values in it, so we cannot conclude a positive correlation. Therefore we cannot conclude that businesses located in higher populated cities get more votes on their reviews per review, just that they get more votes on their reviews overall. And since from our previous results we concluded that businesses in higher populated cities tend to have more reviews, we can conclude that while the average ratings per review doesn't seem to be correlated with city population, the businesses located in higher populated cities will have more reviews, and as a result the over ratings for reviews for those businesses are higher than those businesses with less reviews.

From this we conclude that businesses located in areas with more businesses and more populated areas have higher review counts, and higher amounts of people who read and rate the business's reviews.

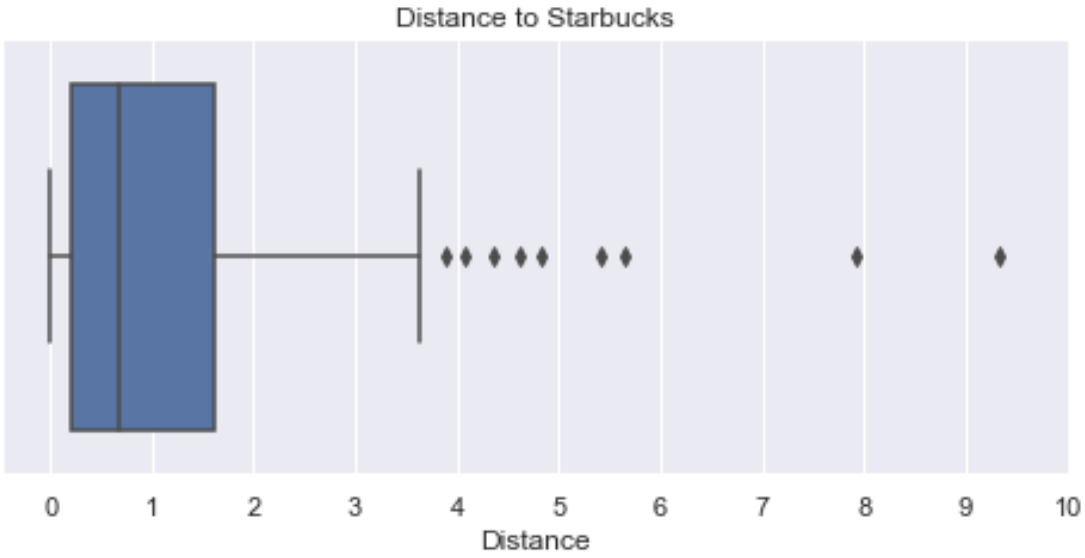
5.4 4.4. Is it true that for every Tim Hortons in the GTA there is a Starbucks nearby? Calculate distances between establishments of the two groups and assess distance patterns. Plot the two types of establishments on a map.

We will define 'nearby' as within 1 kilometre, and 'kind of near' as within 2 kilometres. Then we will calculate the distance to the nearest Starbucks for every Tim Hortons.




```
Percentage of Starbucks within 1km: 59.01639344262295%
Percentage of Starbucks within 2km: 79.09836065573771%
```

	city	count	mean	std	min	50%	max	population
0	Halton Hills	1.0	7.925562	NaN	7.925562	7.925562	7.925562	61161
1	King	1.0	5.405239	NaN	5.405239	5.405239	5.405239	24512
2	Milton	1.0	0.173407	NaN	0.173407	0.173407	0.173407	110128
3	Whitchurch-Stouffville	1.0	9.335472	NaN	9.335472	9.335472	9.335472	45837
4	Aurora	2.0	1.874816	0.297910	1.664161	1.874816	2.085470	55445
5	Newmarket	2.0	1.192519	0.220572	1.036551	1.192519	1.348486	84224
6	Pickering	5.0	1.430915	1.083371	0.158614	2.033908	2.535236	91771
7	Ajax	6.0	1.523778	0.868459	0.416458	1.679034	2.798948	119677
8	Oakville	6.0	1.661013	1.333108	0.169667	1.628548	3.301628	193832
9	Brampton	11.0	1.860293	1.348020	0.292594	1.521695	4.830057	593638
10	Richmond Hill	12.0	1.202582	1.263348	0.124725	0.897923	4.619392	195022
11	Vaughan	12.0	2.099633	1.567460	0.437299	1.644272	5.649957	306233
12	Markham	15.0	1.519032	1.337110	0.208783	0.878668	4.084931	328966
13	Mississauga	33.0	1.227786	0.950753	0.000000	0.936822	3.171094	721599
14	Toronto	136.0	0.716261	0.869021	0.016560	0.289110	3.640957	2731571



We can see from the map that while both Tim Hortons and Starbucks are crowded in the more concentrated regions of the GTA, there are a few locations out of downtown Toronto that are spread out further away.

Looking at the boxplots of distances from a Tim Hortons to its nearest Starbucks, we can see that outliers begin a little below 4 kilometres, and the 75th percentile is below 2km and very left skewed, so the majority of Tim Hortons have a Starbucks nearby. 59% of Tim Hortons locations have a Starbucks within 1 km, and 79% have a Starbucks within 2 km.

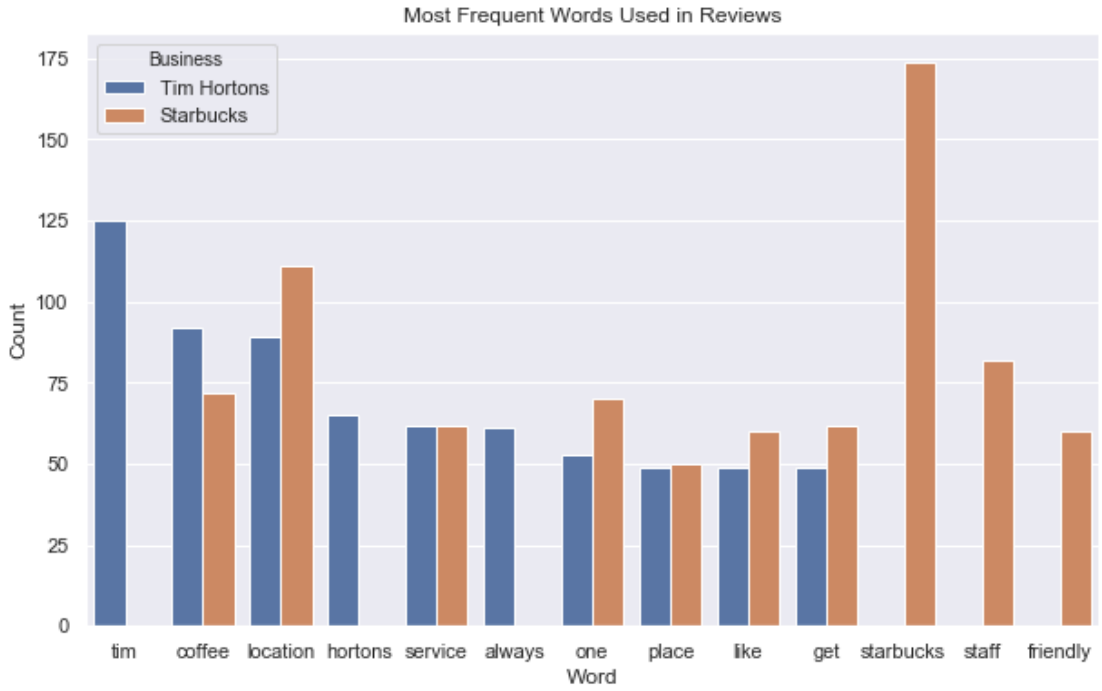
Looking at the chart of distances for each individual city, only five out of the sixteen cities have a median below 1 kilometre. However, three of these five cities have the highest count of businesses in them. As well, looking at the boxplot of mean distances per city, the majority of cities where we would have to walk more than 2 kilometres to find a Starbucks are all outliers.

So we conclude that there usually is a Starbucks within 2km of a Tim Hortons for all the Tim

Hortons in this dataset, mostly for Tim Hortons in cities with many businesses.

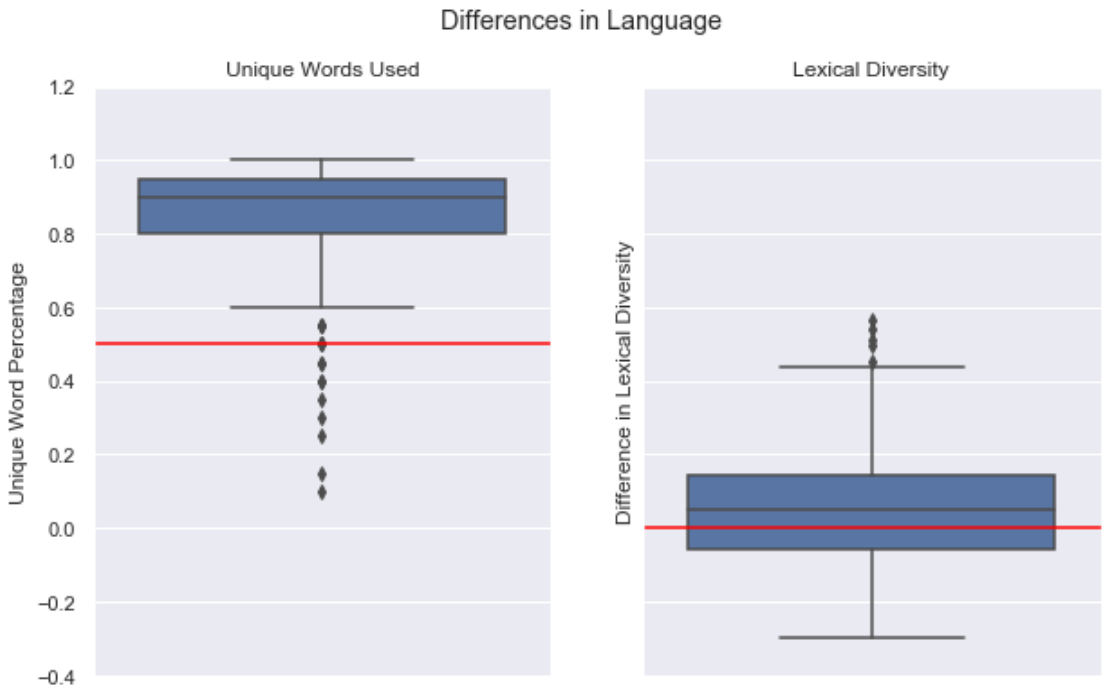
5.5 4.5. Do Yelp reviewers use similar language in their reviews of GTA's Tim Horton's and Starbucks?

We will investigate the users who have made reviews for both Tim Hortons and Starbucks, and see if their is a significant difference in the language they use in their reviews. We will investigate their language through checking if the most frequently used words in their reviews are relatively the same for both franchises, not including stopwords. We will also investigate it through the lexical diversity used in Tim Hortons reviews versus Starbucks reviews.



We look at the words used in Tim Hortons and Starbucks reviews by the same user, and count the most frequent words used for each business, not counting stop words. We have graphed the top 10 most frequently used words for Tim Hortons and Starbucks. Obviously the names of the business are some of the most frequently used for each respective business.

We can see that reviews about Tim Hortons tend to use the word 'always' much more than ones about Starbucks, and that reviews about Starbucks tend to use the words 'staff' and 'friendly' much more than ones about Starbucks. The other words appear at relatively similar counts. From just the most frequent words used, as well as our previous conclusion from 4.3 that Starbucks in the GTA tend to have more businesses with high reviews that Tim Hortons, we could infer that Starbucks locations tend to have friendlier service than Tim Hortons locations.



We compare for each user, the most frequently used words they use in their Tim Hortons reviews and Starbucks reviews, and note the percentage of those words that are unique to the franchise. From the comparison of the most frequently used words in Tim Hortons reviews and Starbucks reviews, the percentage of those words that are unique to the franchise are relatively high. The median is over 0.8, and the lower bound is above 0.5. All of the users who have a unique word usage per franchise below 0.5 are considered outliers. This means for most people, the most frequent words used in reviews differ depending on if it's for Tim Hortons or Starbucks.

From the comparison of lexical diversity, we can see that the median is above zero. This means that over fifty percent of users use more kinds of words when reviewing Tim Hortons in comparison to Starbucks.

Therefore we conclude that people do use different language when reviewing Tim Hortons and Starbucks. People tend to use more kinds of words when reviewing Tim Hortons compared to Starbucks. From our previous analysis in 4.3, we also know that Tim Hortons gets less reviews and also lower ratings than Starbucks does. Maybe when people are describing mediocrity in Tim Hortons, they use more colourful language than their Starbucks reviews.

6 Conclusion

6.1 Question 2

The Yelp dataset contains businesses clustered in specific regions of North America, rather than an even spread of businesses across the continent. This makes it extremely unlikely that this is a random independant sample of businesses. The cities with the most businesses present in this dataset are Las Vegas and Toronto.

The most frequent business categories are Food, Restaurants, Shopping, Home Services, and Beauty & Spa. Businesses on Yelp that are restaurants tend not to tag their restaurants also as a part of the 'Food' category and go for the more specific tag of what kind of food they serve, so it is likely that Food is the actual most frequent business category.

Establishments with bike parking tend to be of the categories Coffee & Tea, Bars, American (Traditional), Food, Nightlife, Hair Salons, Beauty & Spas, Sandwiches, and Active Life. Businesses that people can just walk into to use their services tend to have bike parking more than businesses that you can't walk into.

Having more Yelp reviews is somewhat correlated to a higher star rating, but only up to star rating 4.0. Above 4, review count no longer matters, and businesses with lower review counts have a higher star rating. This can be see as if a business gets enough high star ratings, everyone automatically assumes that it's a very good establishment and no longer bothers with leaving reviews.

6.2 Question 3

The Canadian cities contained in the Yelp dataset are all from Ontario, Quebec, and Alberta, and mostly concentrated in the Toronto, Montreal, and Calgary metropolitan regions, and a few businesses from neighbouring regions.

6.3 Question 4

In the Greater Toronto Area, notably, Coffee & Tea, Event Planning & Services, and Chinese are much more frequent business categories than overall. Restaurants, Food, Shopping, Beauty & Spas, Health & Medical, Nightlife, and Bars are of around the same high frequency, and Home Services, Local Services, and Automotives are more prominent overall than in the GTA.

The top franchises are Tim Hortons, Starbucks, and McDonalds by number of businesses. If we rank the franchises by number of businesses that have a star rating above 3.5, the top franchises are then Starbucks, Tim Hortons, and Second Cup. While Tim Hortons and McDonalds both have a large number of businesses, they tend to have lower reviews and ratings. This could be because people care less about lower quality as long as the business is quick and accessable. This ties in with how the majority of franchises with a large number of businesses are fast food/drink places, so they're very accessable for people to spend their money at.

We conclude that business location plays a role in the review count of a business, as well as the popularity of the review. Businesses that are located in higher concentrated places with many businesses, like Downtown Toronto, tend to have higher review counts, and the reviews are more popular. This can be interpreted as more people living in places with more businesses, and more people leads to more people reviewing and more people liking reviews to increase popularity.

It is somewhat true that for every Tim Hortons in the GTA there is a Starbucks nearby. Over 50% of the Tim Hortons in the dataset have a Starbucks within 1km of it. The locations that don't have a Starbucks with 1km of them all tend to be further from Toronto city, in the areas of the GTA with less businesses, and so less Tim Hortons and Starbucks.

We conclude that Yelp reviewers do use somewhat different language in their reviews of GTA's Tim Horton's and Starbucks. From our analysis we see that over 50% of the more frequently used words in reviews for the two franchises are different. As well, reviewers tend to use more lexical diversity when reviewing Tim Hortons than Starbucks.

6.4 Limitations

Some limitiations of this analysis is that this is the data provided by the Yelp website. Yelp chose business data from specific regions, rather than data spread out across the country. Since this dataset is very concentrated in certain regions, it is not a very good indicator of overall businesses, both in the GTA and overall. The Canadian businesses are only from three different provinces, and not even spread out across the provinces. Especially looking at the map of the businesses in the GTA, it's very concentrated in the south, while there are little to no observations for other cities spreading outwards. So, for example, while we make the conclusion that that the majority of Tim Hortons have a Starbucks nearby, the majority of the data that we used to make that conclusion is from downtown Toronto, and we have little basis for this conclusion in the more northern parts of the GTA.

As well, there are some typos in city and state names which would have resulted in some data getting left out when conducting our investigations. Another problem is with special characters in the alphabet for names of cities, especially in Quebec. When reading from the JSON file, many of the special characters such as 'é' got messed up when loaded. This resulted in less data from Quebec being used in the analysis for Question 3 than actually present, as well as some businesses getting left out because they don't correspond to the exact spelling of a city in the Greater Toronto Area in Question 4.