JSC270, Winter 2020 - Prof. Chevalier

Assignment 2 - Yelp: Analysis on Businesses and Reviews

Karl Hendrik Nurmeots, February 19, 2020

Abstract

The Yelp Challenge dataset encompasses businesses located in east USA, the Greater Toronto Area (GTA), and the southwestern parts of the US and Canada. Most of these businesses are in the food & drinks, or services industries; retail businesses are noticeably less common.

Bike parking is most available at businesses providing bike or active lifestyle services, and at businesses focused on serving alcohol.

There is very little association between the number of reviews a business has recieved, and the rating the business has.

Businesses in the GTA are similar to those in the entire dataset, but ethnic cuisine businesses are more relavant in the GTA. The two most common chains in the GTA are Starbucks and Tim Hortons, who on average are at most 600m away from each other. Starbucks has placed their locations in a way that allows them to dominate over Tim Hortons in a greater part of the GTA. On average, businesses in downtown Toronto, Mississauga and Markham recieve more reviews than elsewhere. Downtown Toronto, coastal areas of Lake Ontario, and areas to the west of King City have relatively greater average ratings.

The majority of Yelp reviews are written by users who write just a few reviews in total. Looking at the language used in reviews, reviewers' opinions about Tim Hortons and Starbucks are positive, though it appears that users prefer Starbucks more.

Introduction

Yelp (https://www.yelp.com/) is a business directory service and crowd-sourced review forum. This means that users are able to share their experiences at businesses with everyone in the world. Yelp reviews are extremely valuable to businesses: having a better rating means showing up further up the list, and hence leads to more customers visiting the business.

This report aims to answer the following questions about all businesses in the dataset:

- What sort of businesses are present in the dataset, i.e. where are they located and what sort of industry do they belong to? Which of them tend to have bike parking available?
- Does a higher review count lead to a greater review rating?

For businesses located in the Greater Toronto Area (GTA), we will attempt to answer:

- What sort of businesses of the dataset are in the GTA? Which franchises are most frequent?
- Are review count and rating related to the location of businesses?
- Is it true that for every Tim Hortons there is a Starbucks nearby? Which one of these chains has an advantage in the way they've placed their shops with respect to the other?
- Are most reviews written by a small group of users?
- Do people use similar language when writing reviews about Tim Hortons and Starbucks? Is this the case among users who have reviewed both chains?
- Can we detect fake reviews?

The Yelp Dataset

For this report we will be using the Yelp dataset (https://www.yelp.com/dataset/challenge). **For this report to compile, you must download the data from the website and place it in extracted form in the 'data' directory!** This requires you to provide your email and name.

Per the Yelp dataset license (https://s3-media1.fl.yelpcdn.com/assets/srv0/engineering_pages/06cb5ad91db8/assets/vendor/yelp-dataset-agreement.pdf), one is allowed to create a report of the data such as this one only for academic purposes. Any sort of disclosure or sharing of the dataset itself is strictly prohibited - this is a private dataset. Any use of the dataset must not disparage Yelp.

A more detailed overview of the dataset can be found on the dataset documentation page (https://www.yelp.com/dataset/documentation/main). Due to the comprehensive extent of the dataset we will only provide a short insight into what the dataset contains. The dataset consists of 6 JSON files:

`business.json` contains information about each establishment in the dataset such as the business' name, ID, location (address and coordinates), Yelp rating and review count, attributes (e.g. if the business provides a takeout service), category, and opening hours.

`review.json` contains information about each review, namely the author ID, business reviewed ID, rating, date, full review text, and the number of "useful", "funny" and "cool" votes the review has recieved. Author ID maps to the corresponding user in user.json, and business ID maps to the business in business.json.

`user.json` contains information about users' IDs, first names, their friends' user IDs, and info about the feedback they have recieved on their reviews (e.g. total "useful" votes).

`checkin.json` contains each business' ID and a collection of timestamps when users have checked in at that business. Again, the business ID maps to business.json.

`tip.json` contains information about short tips that users have posted, and is structured similarly to review.json. Tips are essentially very short reviews, but do not give a numeric rating about the business. Instead of tracking the number of different votes, the number of "compliments" the tip has recieved is tracked.

`photo.json` contains metadata about photos users have posted: the business' ID the picture is about, the author's user ID, the picture's caption and label (e.g. "food").

For this analysis, we will only use `business.json` and `review.json`. `review.json` is an extremely big file, so we will only use a subset of 2 million reviews for reasonable computing times

Analysis

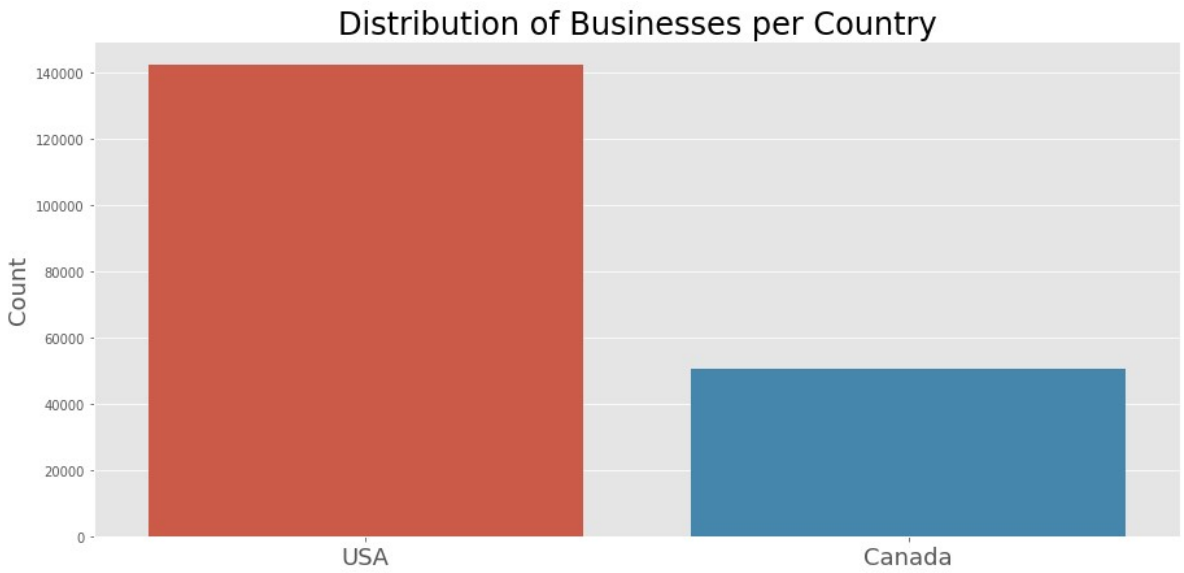
1. Businesses in the Dataset

1.1. Location of Businesses

Oddly enough, the dataset does not contain information about what country each business is located in, so to start off, let's visualize all of our businesses on a world map. For this and all following map visualizations we will be using the Mercator projection.

All of the businesses in the dataset are located either in the US or Canada. We have information about the businesses' postal codes, and we can notice that USA and Canada have different postal code formats: US postal codes include only numbers, while Canadian ones always include alphabetic letters, so we can easily create a new column `country` to distinguish the country the business is located in based on their postal code.

Let's see how the businesses in the dataset distribute in terms of the country they are located in:



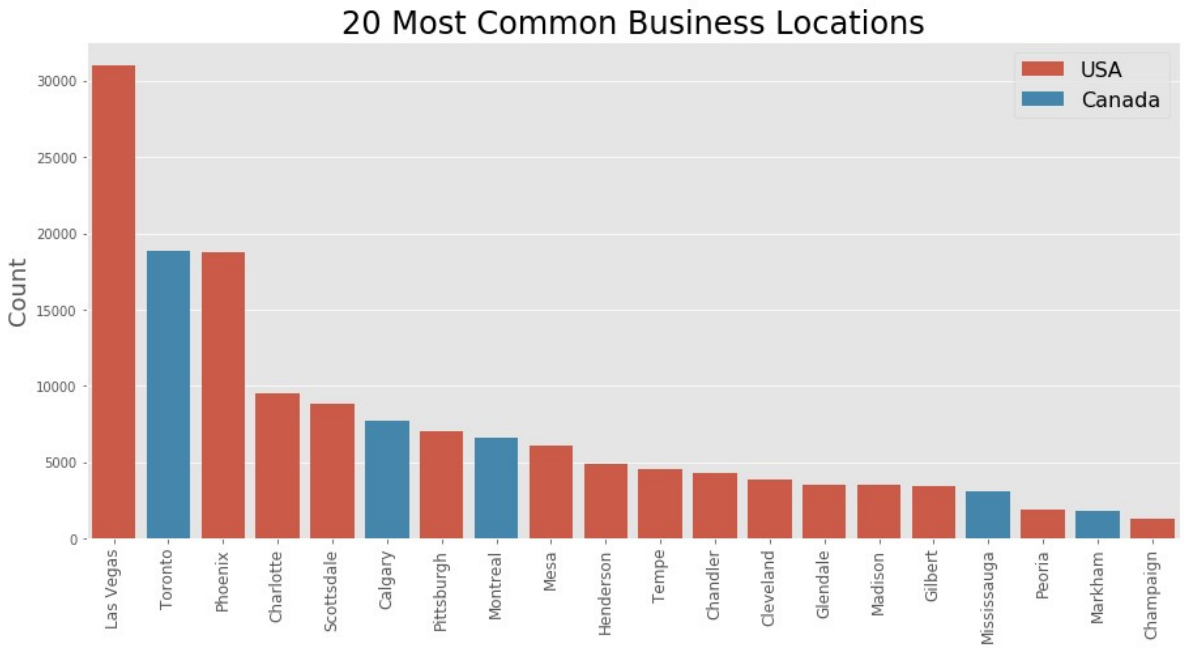
We can see that over 70% of the businesses in the dataset are located in the US.

Moving on, we would like to see how the businesses distribute across cities in the dataset. When we take a look at the different cities present in the dataset, however, we will realise that there are many inconsistencies in the city names: for example, both "Montréal" and "Montreal" show up.

We would like to fix this issue, so we will use FuzzyWuzzy, a fuzzy string matching package. Our businesses dataset is very large, and this type of string fixing is very demanding, so we will have to compromise: we will only try to fix the names of the largest 100 cities in the US and Canada (separately). A quick glimpse into the data tells us that most businesses are located in these cities, so it is also more likely that there are inconsistencies in the city names for these businesses.

More precisely, we will look at the biggest cities by population based on Wikipedia. Here's our reference data: Canada (https://en.wikipedia.org/wiki/List_of_the_100_largest_municipalities_in_Canada_by_population) and USA (https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population).

With the city names corrected, we can take a look at how the businesses distribute based on the city they are located in:



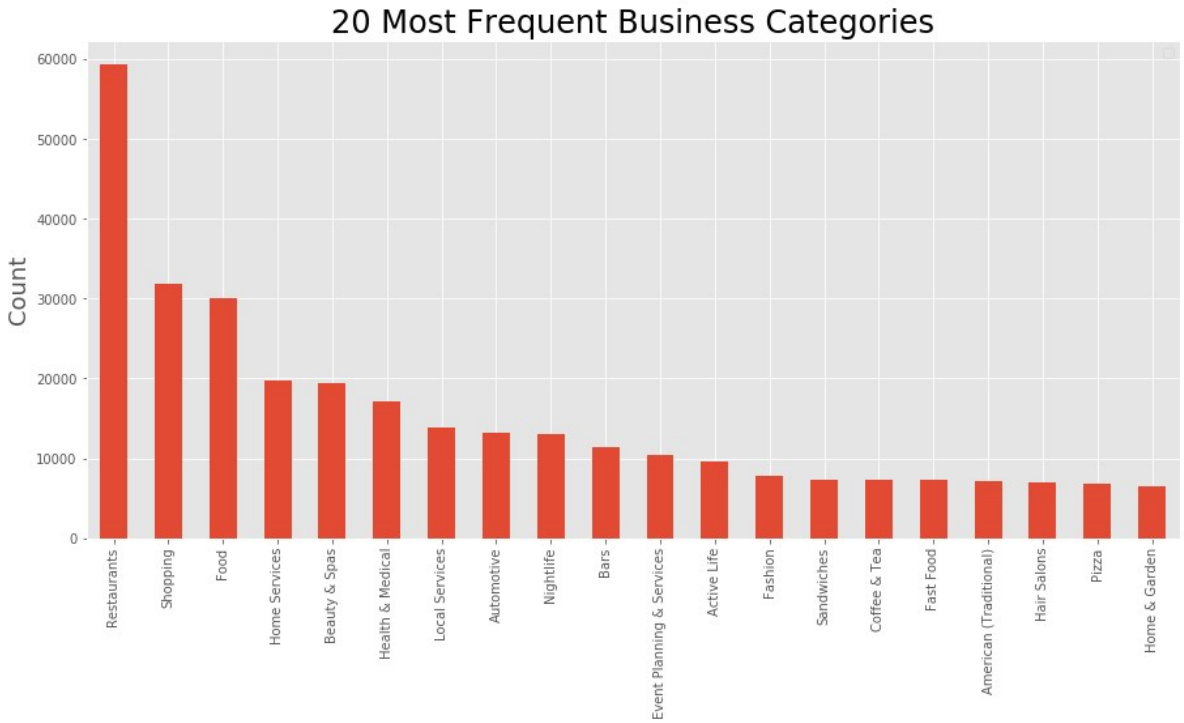
In total, there are just over 1000 different cities in the dataset, however the businesses in the top 20 most common cities account for about 78% of all the businesses in the dataset. As evident, this list is dominated by cities in the US, accompanied by some of the biggest cities in Canada. The most common city in the dataset is Las Vegas, which accounts for about 15% of all businesses.

When we map the location of businesses, we can see that the dataset contains businesses from very few regions: most of them are clustered together in the east US and GTA area, with some others situated in the southwestern parts of the US and Canada.

1.2. Business Categories

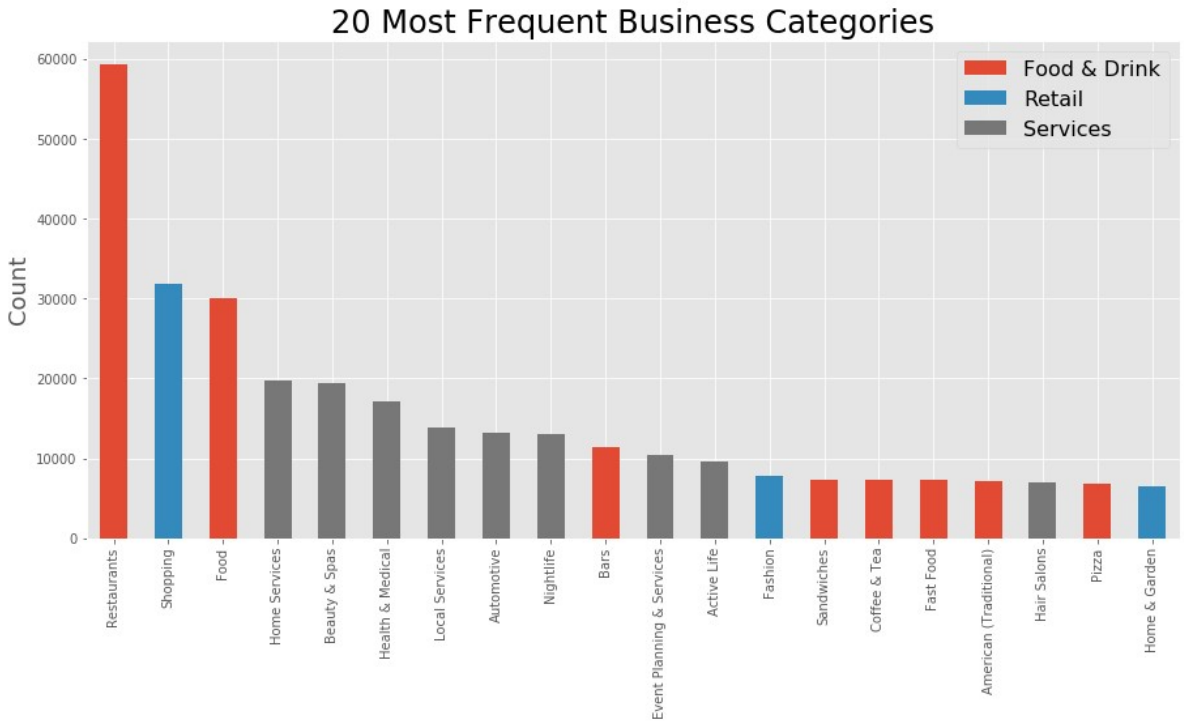
We would like to analyse what categories of businesses are present in the dataset. The dataset has a column `categories`, which contains usually multiple tags for each business. The data is not well formatted: for example, the most common value in `categories` is "Restaurants, Pizza", while the third most common value is "Pizza, Restaurants". Because of all of the different possible tags and their permutations, we need to reduce the number of categories to get a better overview of what's going on.

To do this, we will extract the separate category tags from the categories column (e.g. 'Pizza' and 'Restaurants' from 'Pizza, Restaurants'), and find the total count for how many times each such category shows up in the dataset. There are many different tags used in the dataset, so we will only look at the 20 most frequent ones.

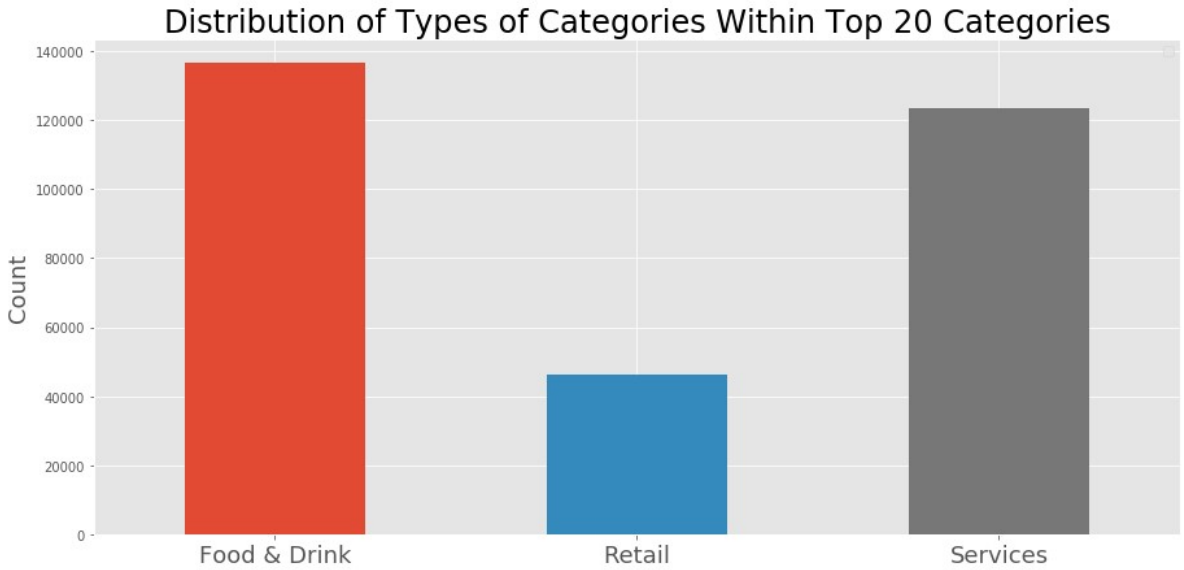


For the most part, this sort of splitting works: most of the categories we are left with are "general" as they describe a very broad category of business. However, we can still see that there's likely some overlap: "Food", "Sandwiches", "Fast Food" and others could be considered subsets of "Restaurants". At the same time, not every business with the tag "Fast Food" is tagged with "Restaurants", so in many cases there is no overlap.

To get a slightly better idea of what's going on, let's divide these categories into three simpler ones: Food & Drink, Retail, and Services. This sort of categorizing will not be perfect because some of the tags can be very vague: for example, notice "American (Traditional)". Presumably this denotes businesses that serve American food, however we can't exatly be sure either. The way we will categorize our business types is definitely subjective and debatable, but it will still give us a simplified overview of the situation.



Let's also combine this with the total counts within the top 20 for our three simple categories:

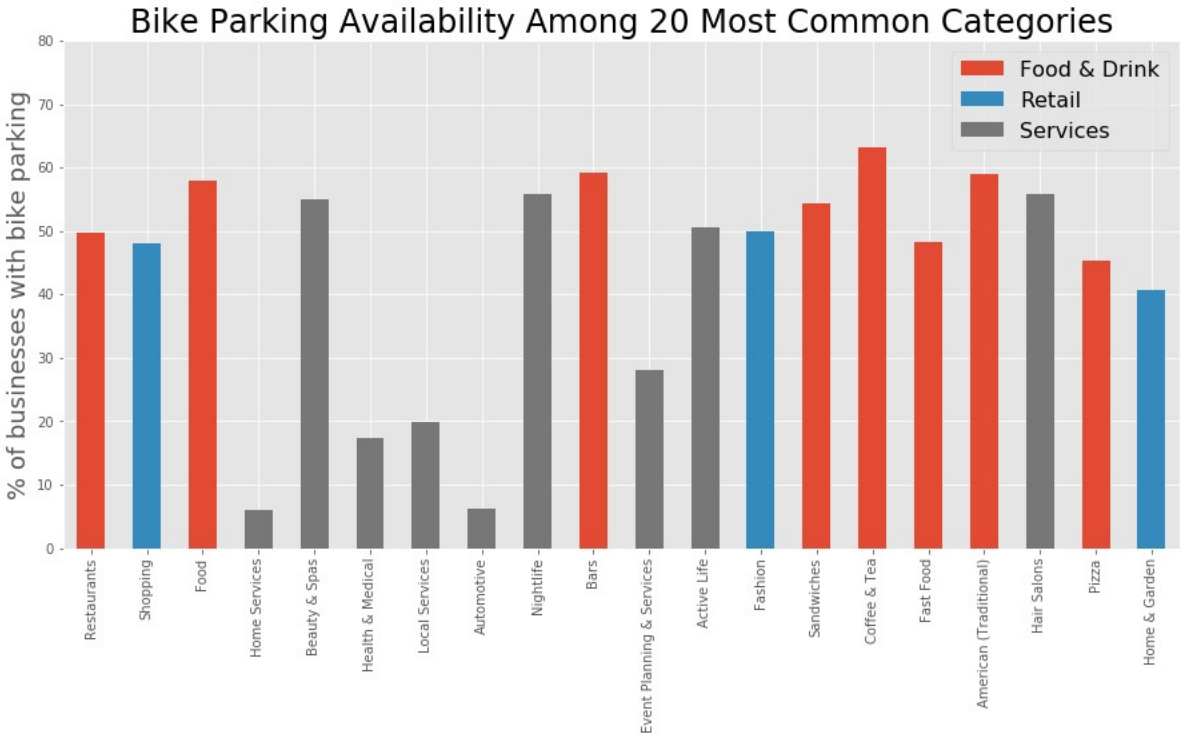


From these two visualizations we can see that food and drink categories are most common due to "Restaurants" and "Food" being so common, but they are closely followed by services categories, which consist of many categories with relatively smaller counts. Businesses dealing in retail are much less common among the most popular categories.

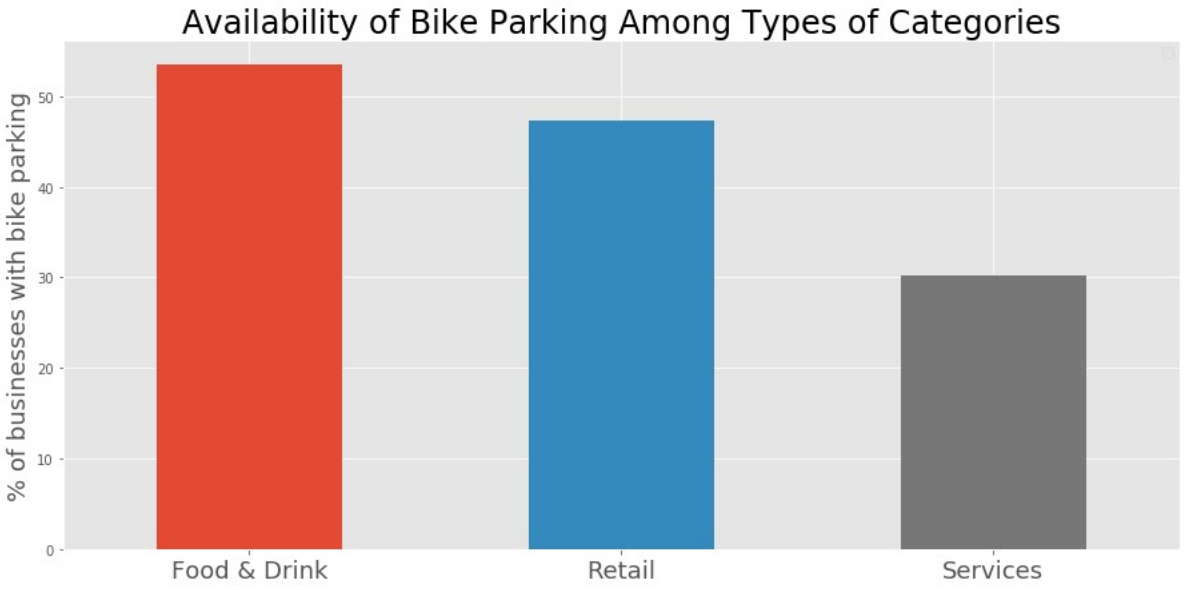
1.3. Bike Parking Availability

We would like to know what types of businesses tend to have bike parking. In the dataset, the column `attributes` contains various attributes relating to each business. One of them is 'BikeParking', which can evaluate to true or false. We have found the proportion of businesses per category that have bike parking available and stored it in the column `prop_bike_parking`.

First, let's take a look at how common bike parking is among the 20 most common business categories we saw in the previous section.



We can see that generally, about 40-60% of the businesses in each category among this top 20 list have bike parking available, with the exception of a few services categories which have noticeably lower proportions.



Looking at our simplified types, we can see that bike parking is most common for food & drink businesses, over half of which have bike parking available, closely followed by retail businesses. Bike parking is less common among businesses falling in the services category.

When we start looking at the list of categories where bike parking is most or least common among the entire dataset, on both ends we get lots of categories with very few businesses recorded as having that category tag. This does not help us answer us the question of which types of businesses tend to have bike parking, so we will only look at categories that show up in the dataset at least 15 times. This cutoff is fairly arbitrary, but is still based on a bit of investigation into the data: as we will see, most categories in the following list have much more than 15 observations.

Out[19]:

	category	count	bike_parking	prop_bike_parking
203	Beer Gardens	80	69	86.2
705	Basketball Courts	17	14	82.4
205	Bike Repair/Maintenance	284	233	82.0
786	Rock Climbing	33	27	81.8
697	Ethical Grocery	21	17	81.0
724	Used Bookstore	30	24	80.0
670	Vinyl Records	140	112	80.0
466	Mountain Biking	82	65	79.3
734	Threading Services	279	220	78.9
561	Whiskey Bars	50	39	78.0

Above are shown the business categories where bike parking is most common. This list consists of businesses focused on serving alcohol, businesses providing bike services, businesses that provide active lifestyle services (basketball courts, rock climbing), and also some businesses that don't belong to any of these categorizations, and don't really form a group among themselves either.

We won't show this, but if we continue to go down the list, businesses serving alcohol, businesses providing bike services, and businesses providing active lifestyle services all show up frequently, so it seems that these types of categories tend to have bike parking most commonly.

Intuitively, this makes a lot of sense. It's no surprise that businesses providing bike services have bike parking. People that use the active lifestyle services are likely very interested in biking to the locations of these services, so again it is no surprise that such businesses frequently offer bike parking.

Perhaps the only somewhat surprising insight is that businesses focused on serving alcohol are so high up the list. Riding a bike while intoxicated is illegal both in the US and in Canada, but it seems that this is not frowned upon as much as driving under the influence, so businesses provide bike parking to direct customers to choose what is considered the lesser of two evils if they really have to. This is a very subjective perspective, and could very much not be actual explanation behind it. Based on the data we have, it is impossible to actually explain this phenomenon.

There are 24 categories where at least 75% of businesses provide bike parking.

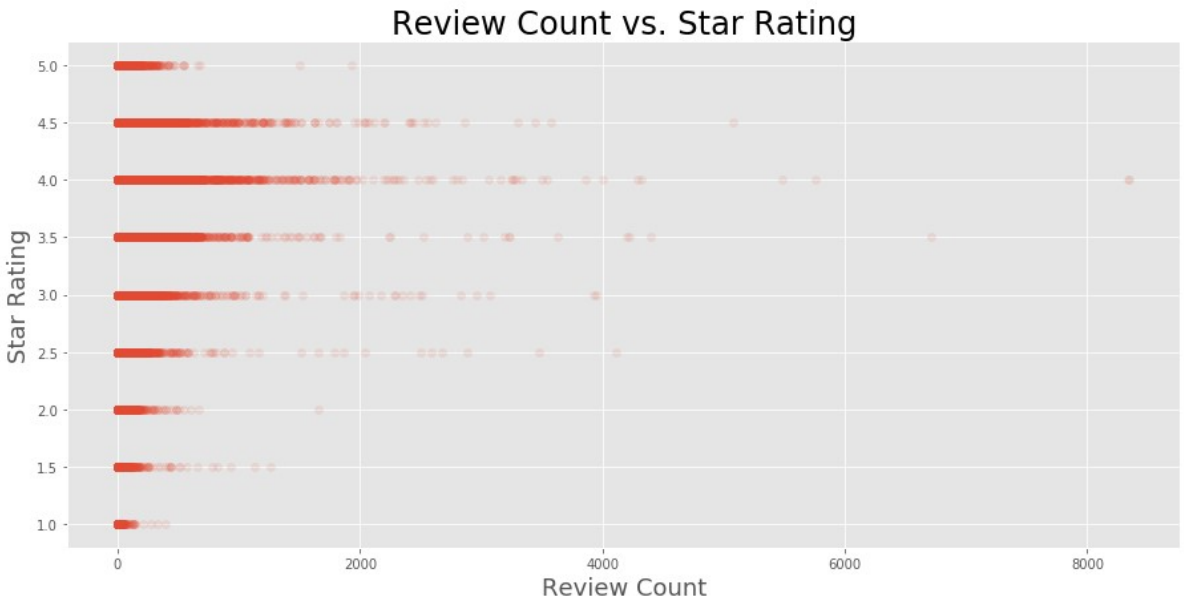
The bottom of the list is much more packed: there are 93 categories where not a single business has bike parking available. Keep in mind that we are still looking at categories with at least 15 observations. There does not appear to be a common characteristic between these categories: they range from religious schools to car brokers. Perhaps the most general insight we can provide is that a lot of these businesses provide services, which matches what we noticed when looking at the 20 most frequent categories.

When we look at the categories that have at least one business with bike parking, the bottom of the list consists of categories that have hundreds or thousands of observations, but only in one to three cases does such a business have bike parking. Again, this list generally consists of businesses providing services.

1.4. Relationship Between Rating Count and Star Rating

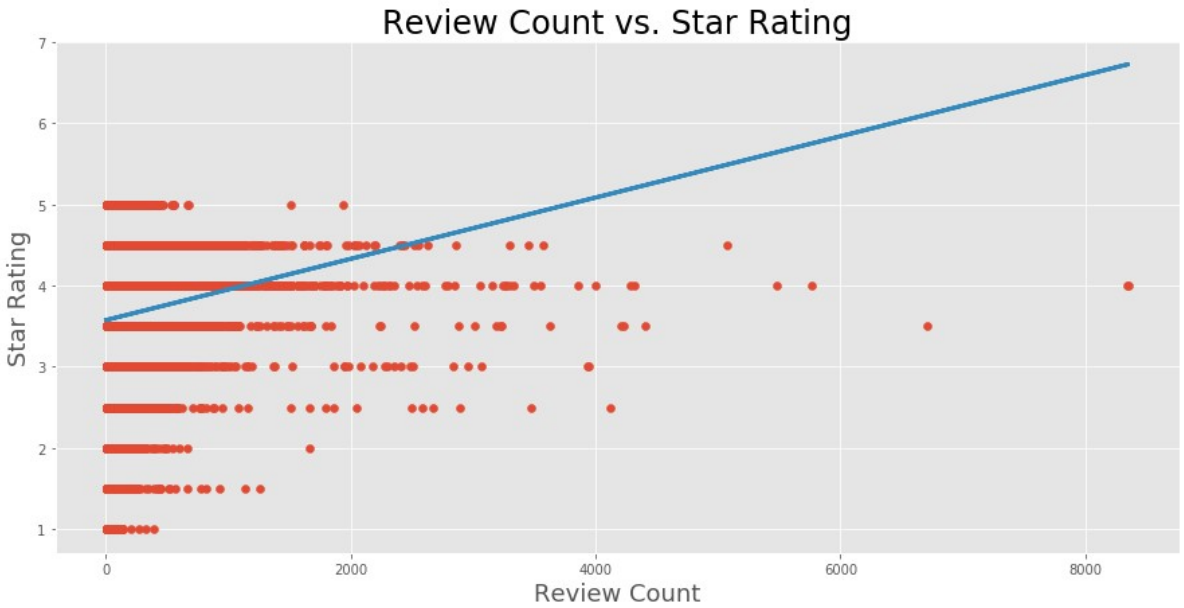
We want to investigate whether having a greater review count leads to a better rating. Recall that each business' rating is given in the column `stars`, which ranges from ranges from 0-5 with a step of 0.5.

Let's plot our two columns on a scatterplot:



There are a lot of observations in our dataset, so we've made each point transparent to see the actual trend better. It does not seem like there is much going on - the data has a lot of variance, but we can still notice that higher ratings are more common for businesses with higher review counts.

We can fit a linear regression model to our data to get a more clear idea of the relationship.



Whilst our linear regression model does show a positive relationship between the variables, even visually we can see that it does not fit the model well at all. This is confirmed by the model's R^2 value of 0.002, i.e. the model is able to explain only 0.2% of the variance in the data.

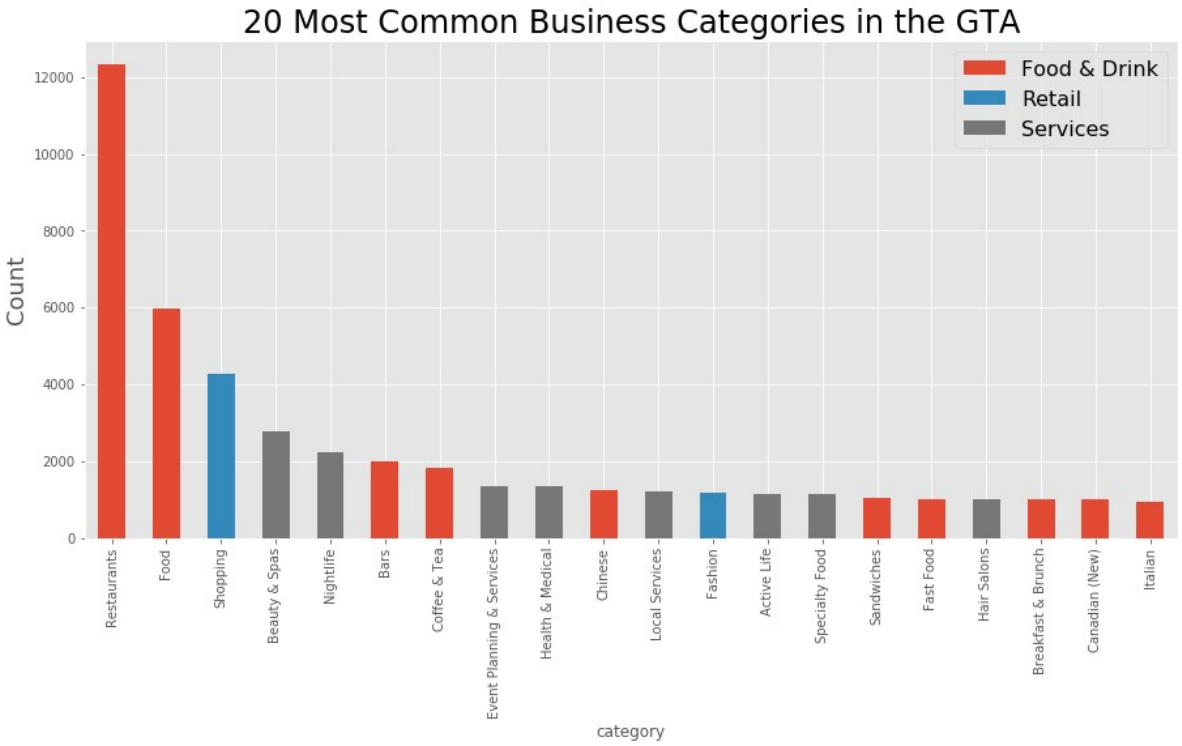
Hence we can fairly certainly claim that there is very weak association between review count and star rating. The insight that higher ratings are more common for businesses with higher review counts is likely caused by how the ratings themselves distribute: most ratings fall in the 4.5-3 range.

2. Businesses in the GTA

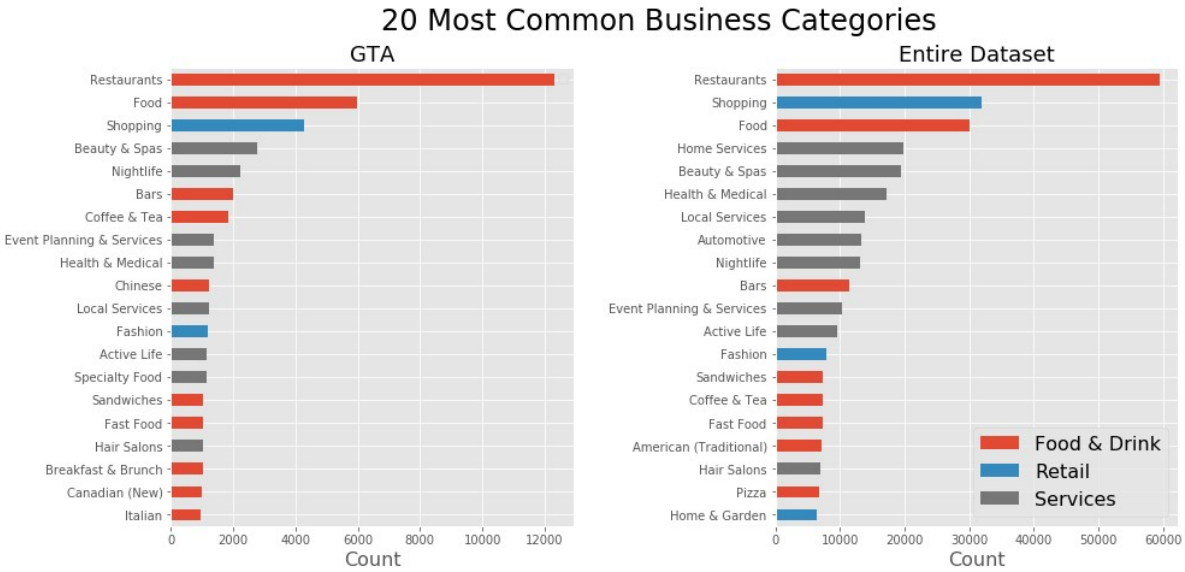
For this section we are interested in businesses situated in the Greater Toronto Area (GTA). People's definition of which municipalities are contained within the GTA varies, so we will be using the list of cities given in the relevant Wikipedia article (https://en.wikipedia.org/wiki/Greater_Toronto_Area).

2.1. Business Categories

We will take a look at what are the most popular business categories in the GTA, and compare it to our previous results regarding the entire dataset. To make comparisons easier, we will use the same methodology to classify our categories into three simpler ones: Food & Drink, Retail, and Services.

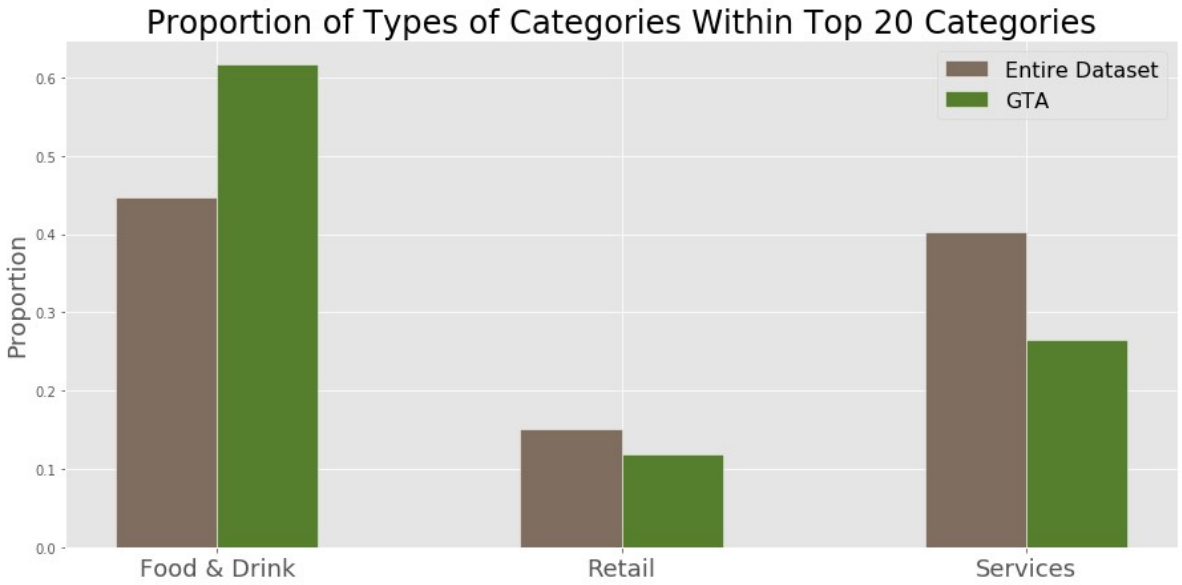


At first glance this distribution looks fairly similar to what we saw previously with the entire dataset included, but when we take a closer look we start to notice differences. Let's place this visualization side-by-side with the equivalent graph for the entire dataset for better comparison:



Of course, count-wise the numbers are much greater when the entire dataset is included. Restaurants, Food, and Shopping are the 3 most common categories in both datasets. Home Services, the 4th most popular category in the entire dataset is not even present in the GTA top 20 list. In the overall list, American (Traditional) is the only category referring to cuisine from a certain country or ethnicity, but in the GTA dataset there are multiple such categories: Chinese, Canadian (New), and Italian.

Let's also see how the three simple categories compare across the entire dataset and its subset in terms of proportions:



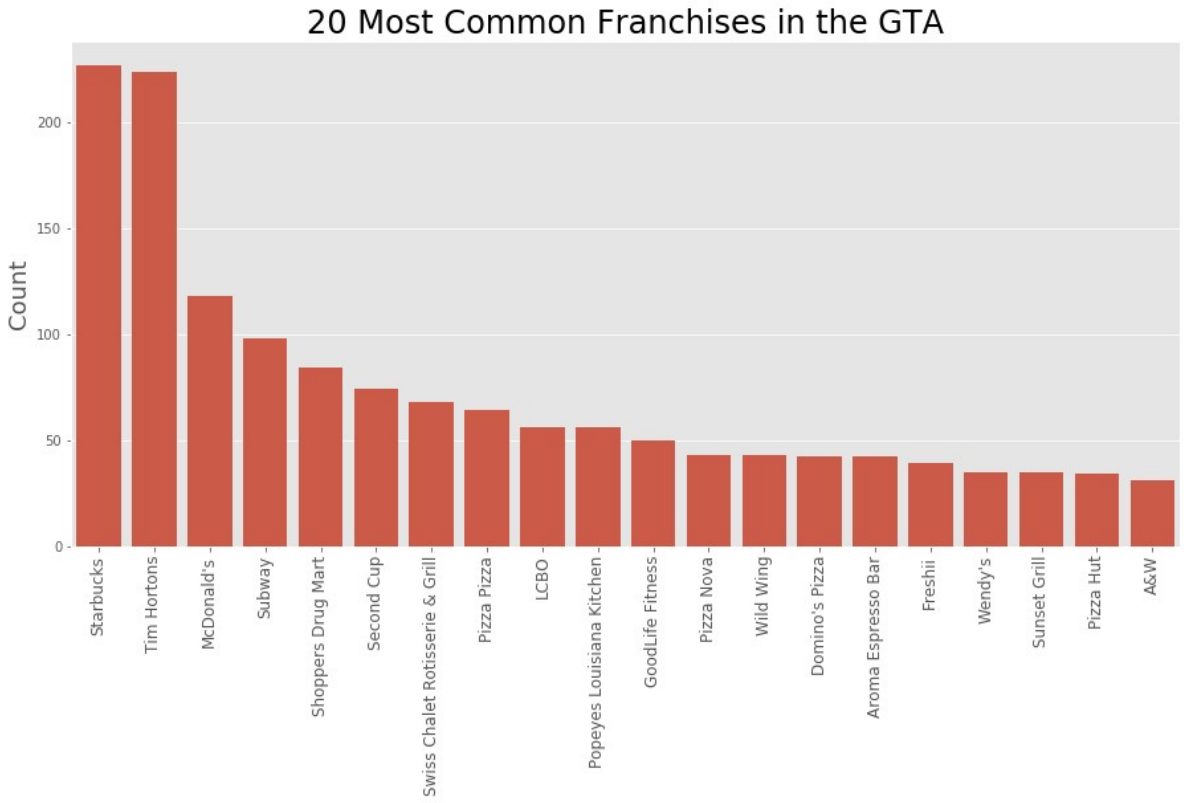
We can see that Food & Drink categories make up a much greater proportion in the GTA than in the entire dataset, and this mainly comes from Services categories having a noticeably lower propotion. Just like in the overall case, retail categories make up the smallest proportion of categories.

2.2. Top Franchises in the GTA

We would like to know what are the most common franchises in the GTA. Before we can start this analysis, we realise that the same issue we had with city names shows up again - there are inconsistencies in business names. For example, both "Tim Hortons" and "Tim Horton's" are present in the dataset.

Again, we will use Fuzzywuzzy to fix this, but we will only do so for the 20 most popular business names before any sort of corrections. While it is possible that if we fixed all business names the top 20 list would change, a quick glimpse into the data shows that it would only affect the tail of the list, if at all. This is necessary for reasonable computing times.

There is one exception to this: we will exclude Pizza Pizza from being fixed. Because lots of pizza restaurants have "Pizza" in their names, they will frequently get matched with "Pizza Pizza". This means that potentially Pizza Pizza could be higher up the list, but we don't have an easy way of fixing this unless we had more computing power for a more exhaustive string correction method.



As we saw previously, most businesses in the GTA were in the food and drink industry, so it is no surprise that this top 20 list consists of businesses in that industry. Only two businesses, GoodLife Fitness and Shoppers Drug Mart, do not belong to this industry. Starbucks and Tim Hortons are the two big chains that have about double the locations than those that immediately follow them in ranking.

2.3. Effect of Business Location on Reviews

We would like to analyse whether the location of a business can affect their reviews: namely, we will look at if business location can affect the star rating and the number of reviews a business gets.

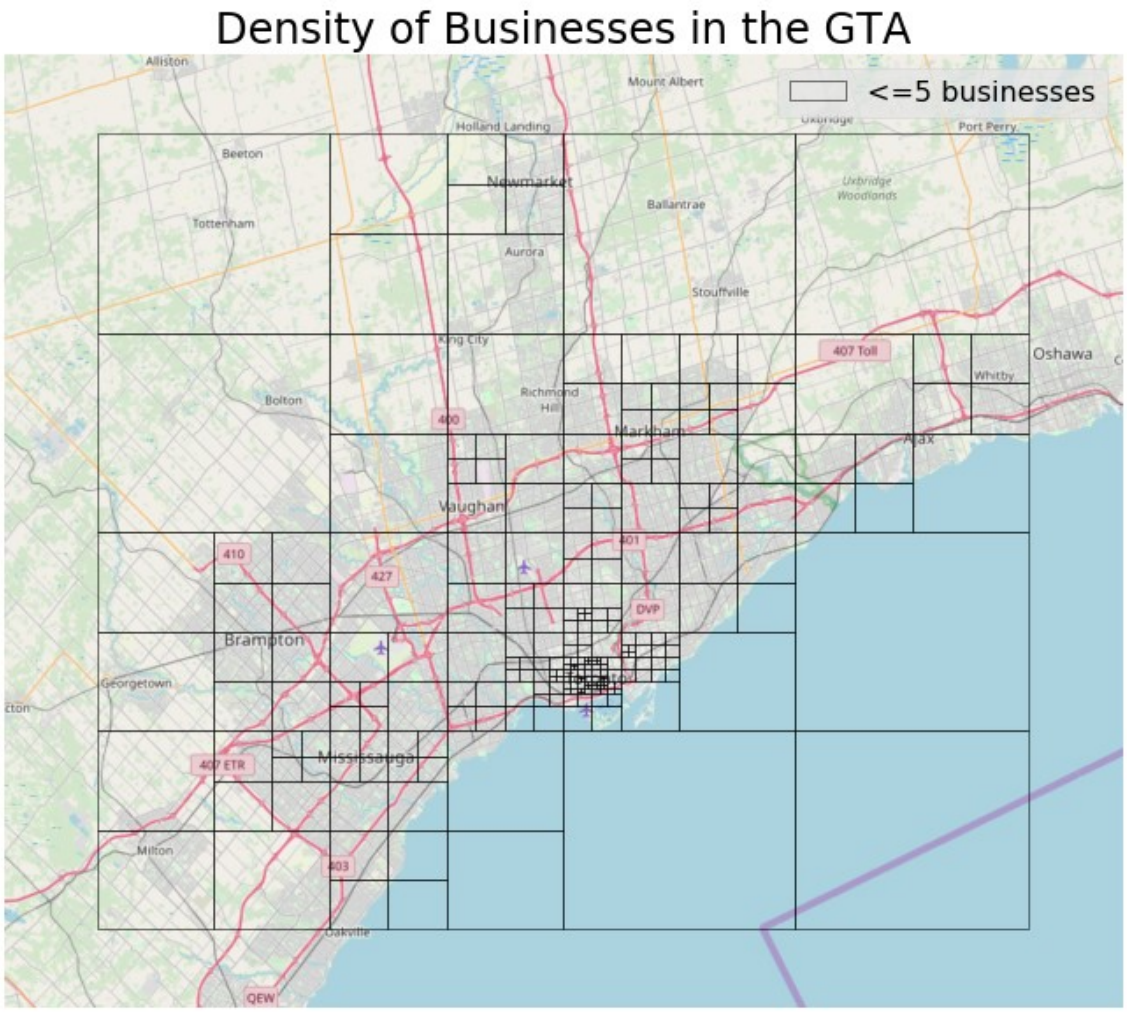
Before we jump into those questions, let's take a look at how the businesses in our GTA dataset are located. There are a lot of observations in the data, so a simple point plot would not work: everything would just be really cluttered. Instead, we want to use our points to define areas that are similar to each other, and then see what we can find out about each area.

To do this, we will use Quadtree plots (https://residentmario.github.io/geoplot/plot_references/plot_reference.html#quadtree) that come as a part of the geoplot (https://residentmario.github.io/geoplot/index.html) Python package. In our case, this algorithm creates rectangles that contain no more than 5 businesses each by recursively dividing the space our businesses are located on.

Before we can create these plots, we have to adjust our data a bit. For some reason, the coordinates provided in the dataset have an extreme degree of precision (https://xkcd.com/2170/), which means that on the scale the quadtree algorithm operates on, many points fall on the same location and hence cause the algorithm to fail since it cannot divide the entire area into the rectangles we desire.

Thus, we will need to "jitter" the points. That is, we will shift all points in our dataset by a tiny random amount. This will, of course, introduce some imprecision, but it is on a scale of centimeters - unnoticeable for the scale we are doing our analysis on.

Now, we can finally create our quadtree plot.

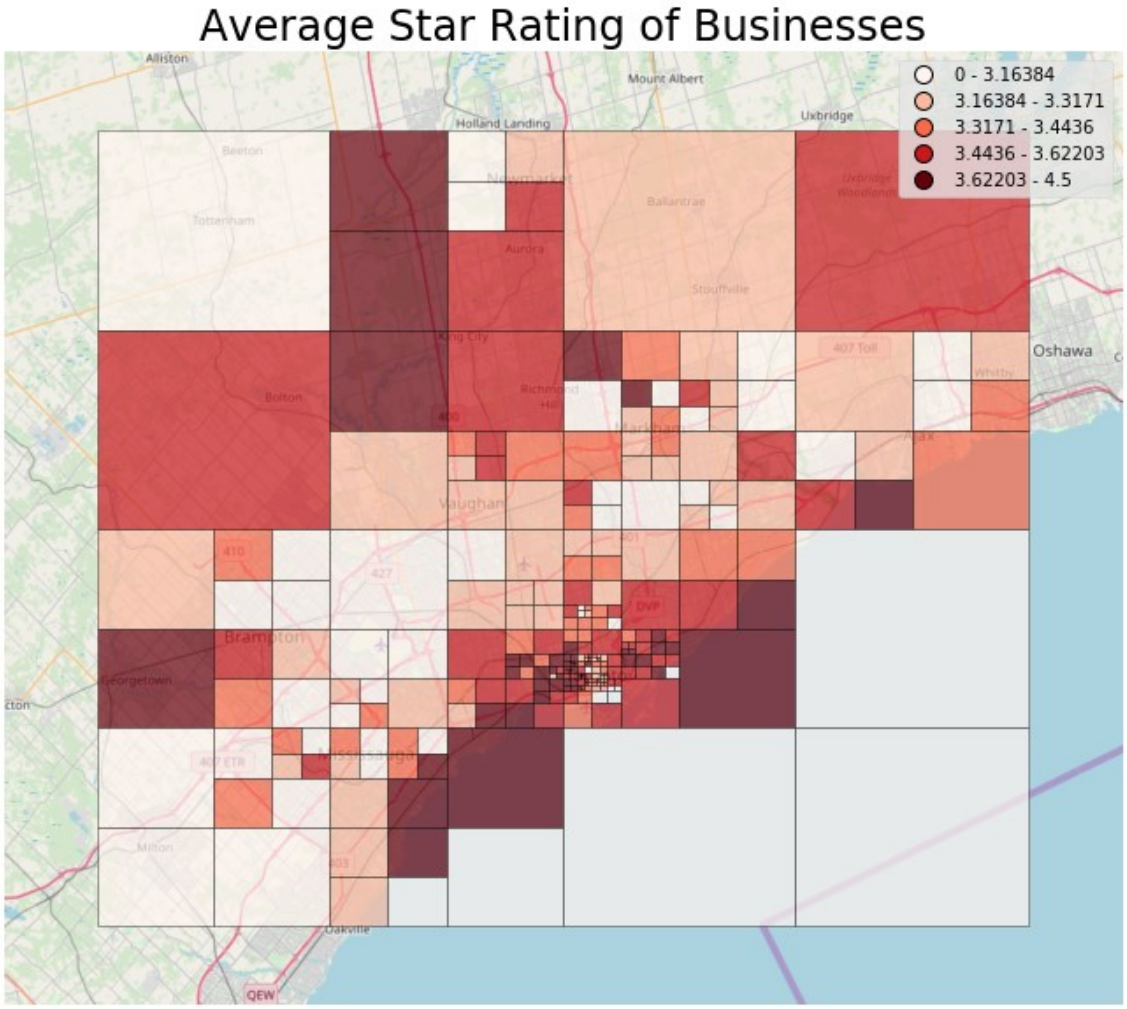


Recall that each rectangle contains at most 5 businesses from our dataset. Thus, these rectangles visualize density really well: clusters of small rectangles such as in the downtown Toronto area show that there are a lot of businesses in that region. Markham and Mississauga are the other locations where businesses are densely packed.

Immediately one should notice that we have also created some rectangles entirely inside Lake Ontario. This, of course, does not mean that there are businesses there, but since this algorithm starts from a large rectangle bounding all of the points and then divides it down, as a by-product we are also left with some rectangles that don't have any businesses within them. This can also be the case for some of the other rectangles that are on the edges of the big rectangle, i.e. outside of our area of study, or for rectangles mainly covering highways.

So, as we proceed with this analysis it is important to keep in mind that not all rectangles actually represent information about businesses. Because we are using an open-source package for these visualizations, the package does not have a way of avoiding this, and no better alternative methods exist. Really large rectangles should be a warning that they may contain few or no businesses.

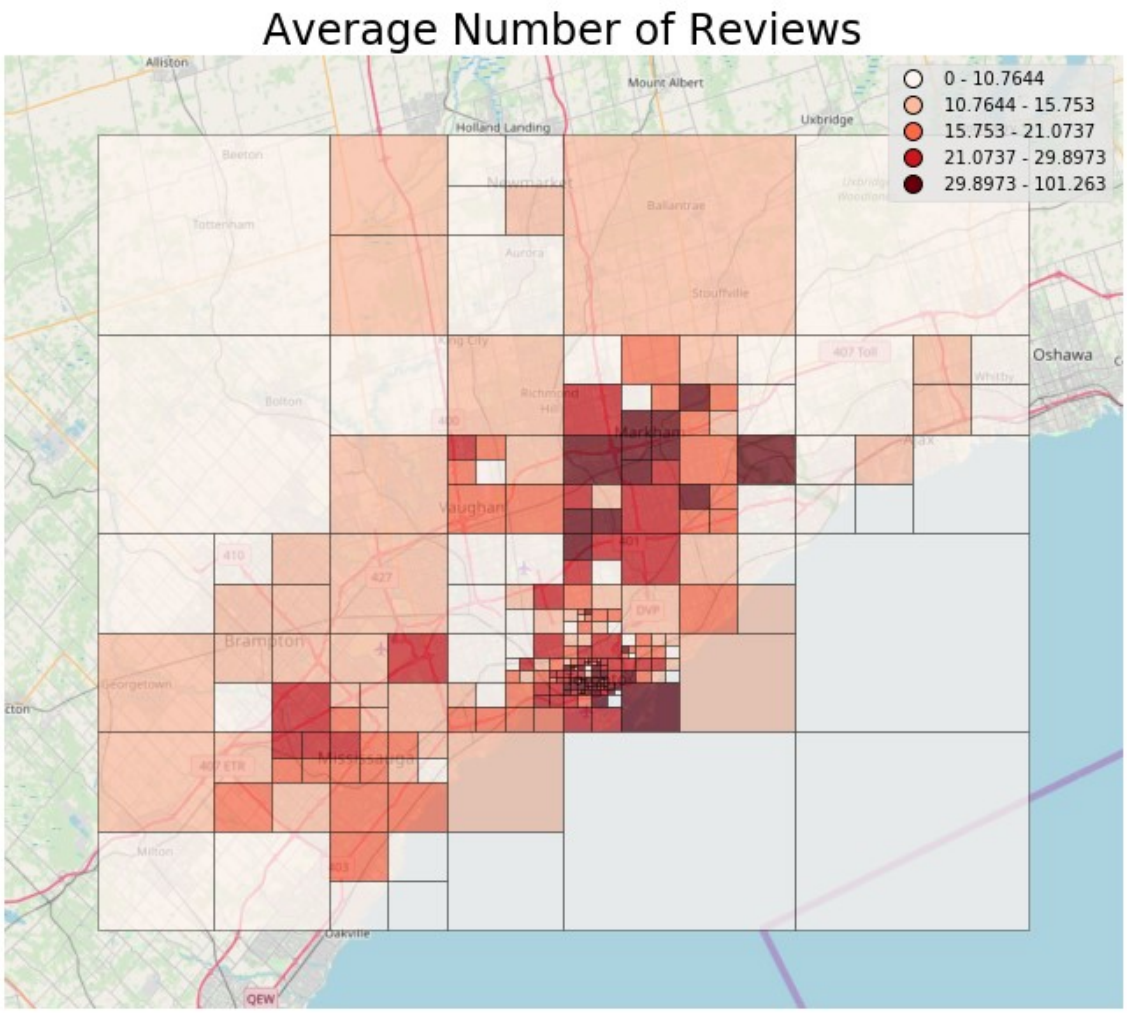
Now that we have divided the GTA into these rectangles, we can give them a hue based on the data we care about. First, let's colour each rectangle based on the average star rating of the businesses within them:



Rather than a gradual scale, we have coloured the rectangles by 20% quantiles since most businesses have a rating in the 3-4 range. We can really easily see that businesses in downtown Toronto and businesses along the coast of Lake Ontario have greater ratings on average than elsewhere. Also, some businesses to the west of King City along highway 400 have really high ratings on average.

Recall our discussion about rectangles that may not contain any businesses: this is likely the case for most rectangles that have the lowest average rating value - a lot of these rectangles fall outside of our area of study or on areas that consist largely of highways. Of course, it is possible that in some of these cases there actually are businesses with a really low average rating. Again, we are unfortunately limited by the tools available to us - we don't have a way to ignore the rectangles that don't actually hold any value to us.

We can also follow the same methodology for review count:



Businesses in the downtown Toronto and Markham areas tend to have more reviews than elsewhere on average. Reviews are also frequent for businesses located in Mississauga and Vaughan.

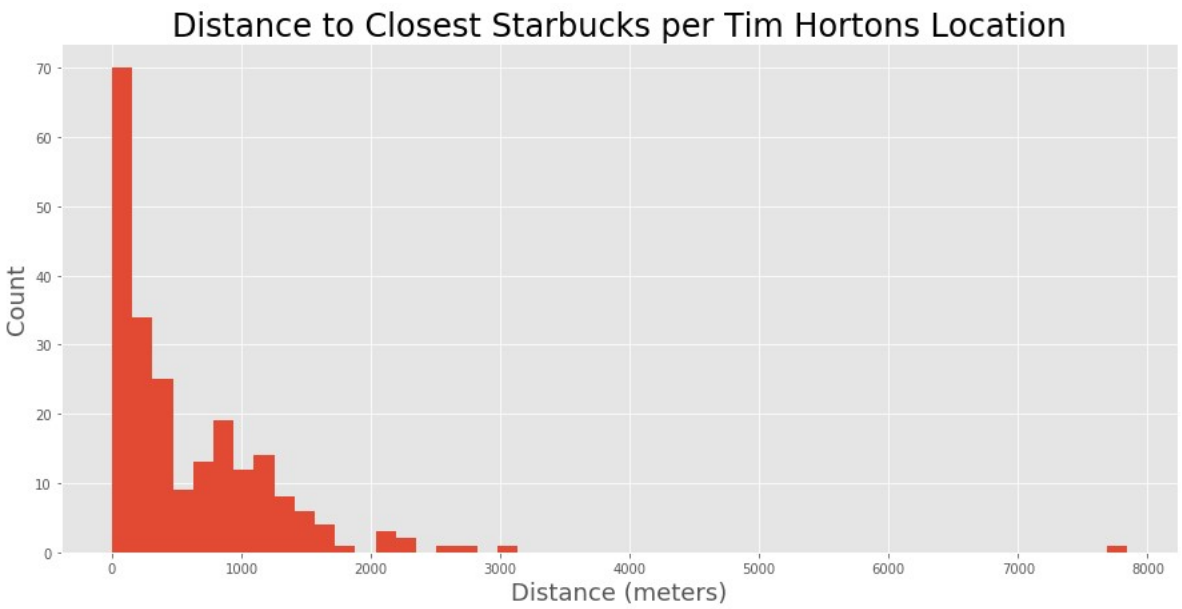
If we cross-examine the rectangles with very few reviews with those with really low average ratings, we can make a good guess on which of them contain no businesses - as hypothesised, it appears that these are mostly businesses on the edges of the big rectangle.

2.4. Tim Hortons and Starbucks

When we analysed the most common franchises in the GTA, we saw that this list was dominated by the two coffee giants Tim Hortons and Starbucks. In fact, it seems that where ever you walk, these two chains are always around every corner, and nearly always they are located right next to each other. We would like to investigate whether this is an actual phenomenon or just a case of confirmation bias.

To do so, we can compute the distance between every Tim Hortons location and their closest Starbucks neighbour. Since our dataset contains spherical coordinates, we can use the Haversine formula (https://en.wikipedia.org/wiki/Haversine_formula) to compute these distances.

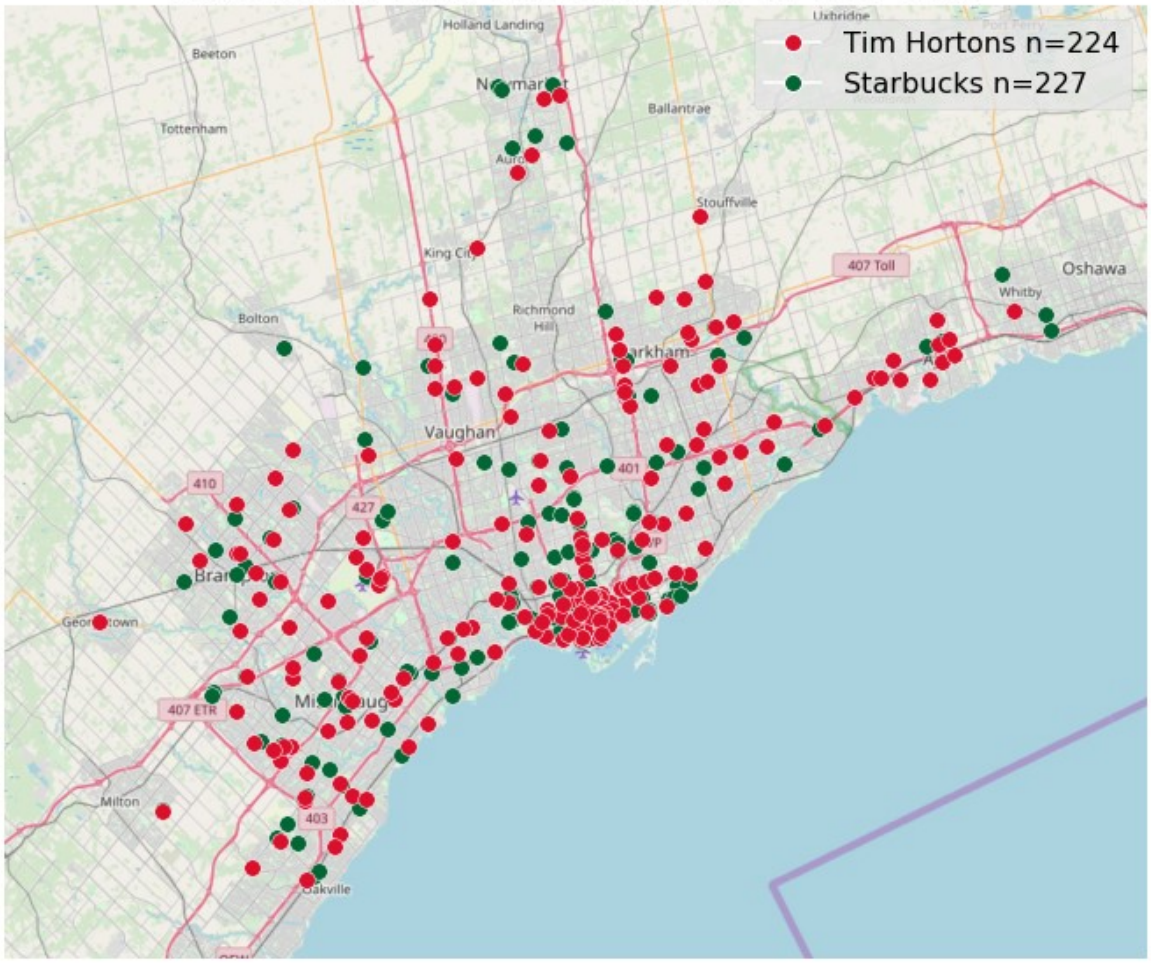
We can visualize the distribution of these distances on a histogram:



Out[40]:

	Distance to Closest Starbucks (meters)
count	224.0
mean	609.3
std	760.9
min	0.0
25%	113.3
50%	351.3
75%	919.7
max	7838.8

Of course, the interpretation of "right next to" is subjective, but it certainly does not seem like every Tim Hortons has a Starbucks right next to it. Only in fewer than 25% of the cases is there a Starbucks within a 100m radius from a Tim Hortons location. On average, the distance between the two chains is just over 600 meters. In a curious outlying case, a Tim Hortons location is nearly 8km away from its closest neighbouring Starbucks.


Starbucks and Tim Hortons in the GTA
Tim Hortons n=224
Starbucks n=227
Holland Landing
Beeton
Tottenham
Ballantrae
Stouffville
King City
Richmond Hill
Bolton
Vaughan
Oshawa
Whitby
407 Toll
410
427
401
407 ETR
Milton
403
Oakville

When we visualize our coffee shops on a map we can come up with an explanation for why the distance distribution is bimodal: the competing shops are really close together in downtown areas, and further apart outside of them.

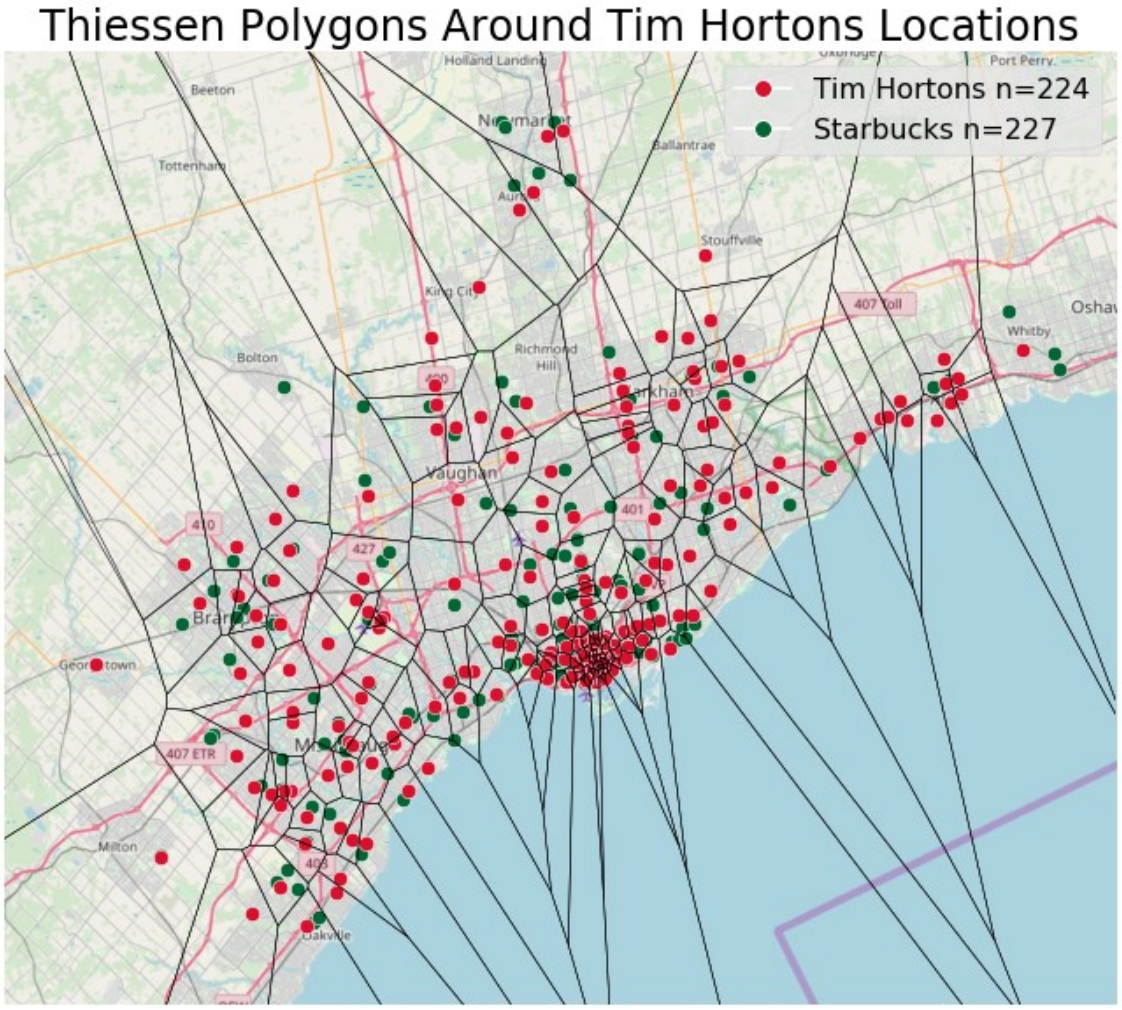
Surely there are a lot of factors that are taken into consideration when these chains choose the locations for their shops. Since these chains are the two most common businesses in the GTA and they operate in the exact same industry, it is reasonable to say that Tim Hortons and Starbucks are each other's main competitors, and as such the way they choose the locations for their shop should depend on how their competitor does it.

So naturally, it would be interesting to see if there's a way we can measure which of these two companies has been more successful in choosing their locations. "Success" is another very vague term. We certainly don't have the data on how much revenue or profit each shop generates, but perhaps there is a way to measure "success" spatially.

Consider this: the businesses are located in the GTA, our area of interest. Every Tim Hortons shop's main competitors are the Starbucks locations closest to it. Hence, if we could somehow divide the GTA into smaller areas based on clusters of Tim Hortons locations and the Starbucks shops right around them, we could analyse which chain is dominant in each such area by counting the number of shops each chain has.

We will achieve exactly this by using Thiessen (Voronoi) polygons (https://en.wikipedia.org/wiki/Voronoi_diagram). We will create a Thiessen polygon for each Tim Hortons location in our dataset, which means that for every such polygon, any other point (a Starbucks location, for example) in this polygon is guaranteed to be closest to the Tim Hortons the polygon was created for.

The algorithm for generating such polygons is quite complicated, but thankfully we can compute them using the Voronoi tool in the SciPy package, and even visualize them using geoplot:



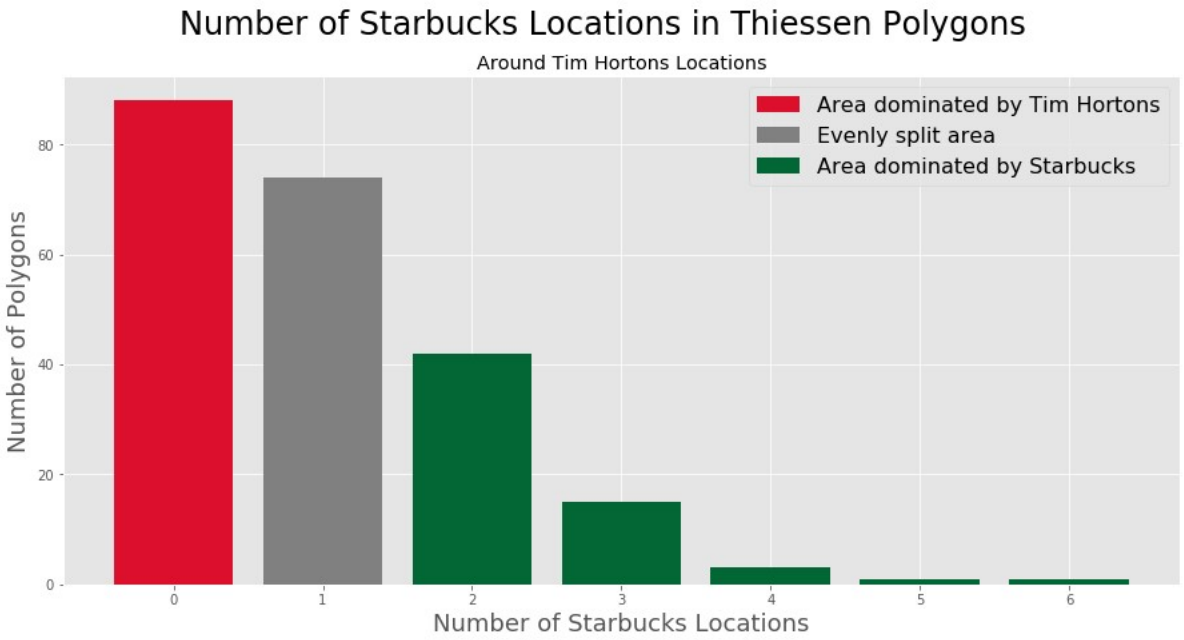
Besides helping us understand the concept, visualizing Thiessen polygons does not help us much. Convince yourself that each polygon has just a single Tim Hortons location, and that any Starbucks within a polygon is closer to that Tim Hortons location than to any other.

Also notice the polygons on the edges of the map - since there are no further Tim Hortons locations, these polygons are infinite in size. We will attempt to fix this by manually bounding the polygons, but this will not be perfect: unless we trace the exact city boundaries, it is difficult to decide where the boundaries of the GTA lie. Furthermore, some of the polygons extend into Lake Ontario, which we certainly don't want to take into account. Again, we will attempt to bound the polygons, but we will still capture the lake in many cases. We could fix these issues more successfully if we had access to specialized and optimized geographic information system (GIS) software, but due to the free Python software we are using this level of error will have to suffice.

We have created a dataframe `pol_df`, where each row corresponds to a polygon. We have not included any polygons that are infinite in size; after our corrections very few valid polygons had to be removed. We will count the number of Starbucks locations in each polygon, and compute the area size for each polygon, adding these statistics to the dataframe.

Recall that the coordinate data in our dataset was given in spherical coordinates - this does not work when computing areas. Hence, we will project our polygons using the Albers Equal Area conic projection - a projection that preserves size, and is designed to be used for locations in temperate areas such as Canada. Projections that maintain size distort shape, but this is not a problem for our analysis since each Starbucks is still contained within the same polygon as before, only visually would it look different (likely unnoticeably due to the scale we are working on).

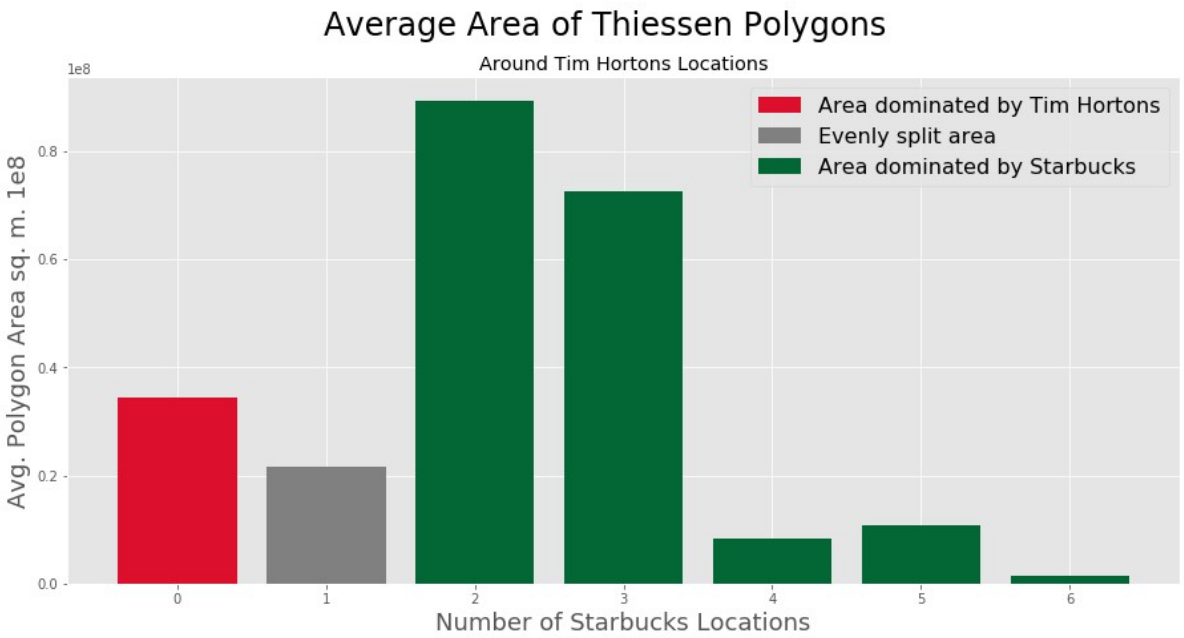
We can now visualize the number of Starbucks locations within our polygons using a barplot:



We have grouped the polygons based on the number of Starbucks locations they contain. If there are no Starbucks locations in a polygon, it means that the Tim Hortons shop within it "has the entire area to themselves", i.e. Tim Hortons dominates in the area. If there is one Starbucks in a polygon, the area is split evenly. If there are more than one, it means that area is dominated by Starbucks.

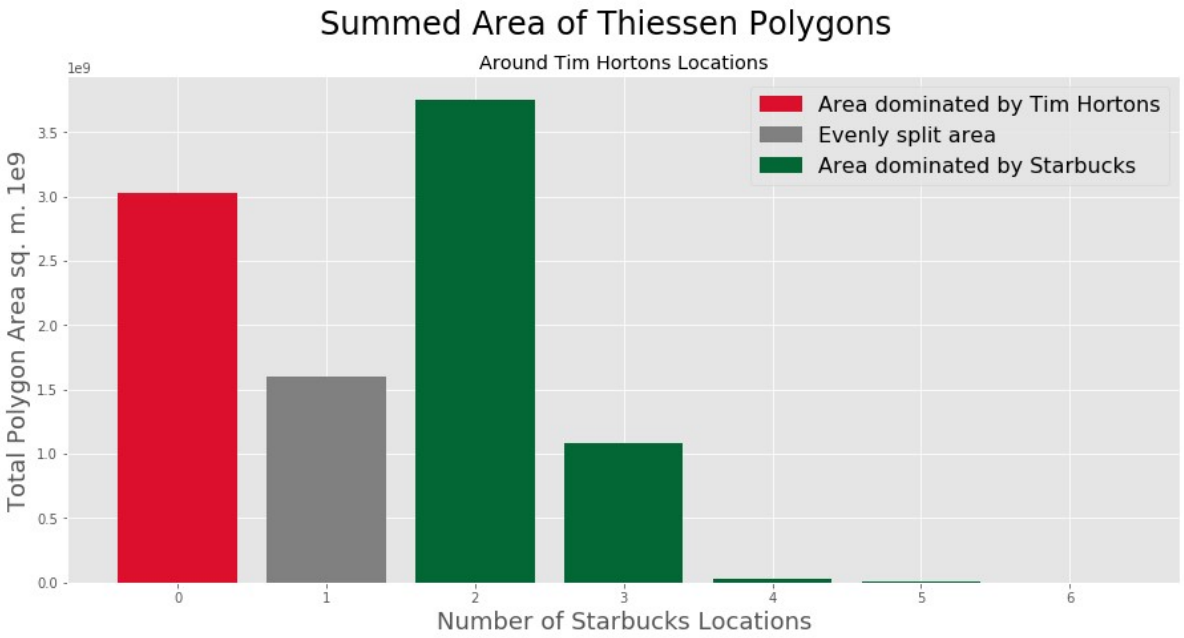
Most Tim Hortons locations are situated in a way that they get to dominate their polygon, or have split the area evenly with Starbucks. This is favourable for Tim Hortons, since it means that other Tim Hortons locations deal with the competing Starbucks shops, and these Tim Hortons shops can operate more freely in their polygon. In fewer cases is there a situation such that Starbucks dominates in the area, and in most such cases Starbucks only dominates by one or two shops.

However, this sort of counting does not paint the entire picture. Sure, Tim Hortons dominates in more areas, but the properties of the areas involved matter too. Dominating in an area of greater size is more valuable because it is likely to capture more people - potential customers. Hence, we should look at the average sizes of the polygons involved.



Immediately we notice a different pattern: the average sizes of the polygons where Starbucks shops dominate are significantly greater than the sizes of polygons where Tim Hortons dominates. The areas where Starbucks dominates by 3 or more shops are on average small in size: these are probably polygons in downtown areas.

Finally, we should look at the summed areas these polygons cover:



We see that even though Tim Hortons has placed their locations so that they have more polygons to themselves, the total areas of the polygons show that Starbucks has the upper hand. This domination is certainly not crushing for Tim Hortons, since in the large majority of cases Starbucks only dominates by a shop or two.

Remember our initial motivation for this methodology: we wanted to split up the GTA into smaller areas. If we summed up the bars in this visualization we would get the entire area size for our area of interest. This means that Starbucks has placed their shops relative to Tim Hortons locations in a way that allows them to dominate in most of the GTA.

Thiessen polygons can be a very powerful tool. Recall that these polygons mean that any location within this polygon is guaranteed to be closest to the point we created the polygon around. If we assume that people are rational beings (which is certainly a bold assumption), it is very likely that a person decides which coffee shop to visit based on the polygon they are situated in. This is why the dominance we described can be ever so valuable.

There are so many ways to branch out this analysis. For example, if we combined our polygon data with population, income, or commute data, we could evaluate how many people a polygon is likely to capture, and how valuable these people could be in terms of the revenue they can generate. Or instead, if we could analyse the road infrastructure and accessibility of each polygon, we could see where coffee shops are most convenient to visit - the business model of these coffee chains is to serve people as quickly as possible, and many customers take their beverages with them instead of staying in the shop.

It should also be noted that our choice to create the polygons around Tim Hortons shops was arbitrary. Since the number of Tim Hortons and Starbucks locations is nearly the exact same, we would expect the results of this analysis to be the same if we created our polygons around Starbucks locations instead.

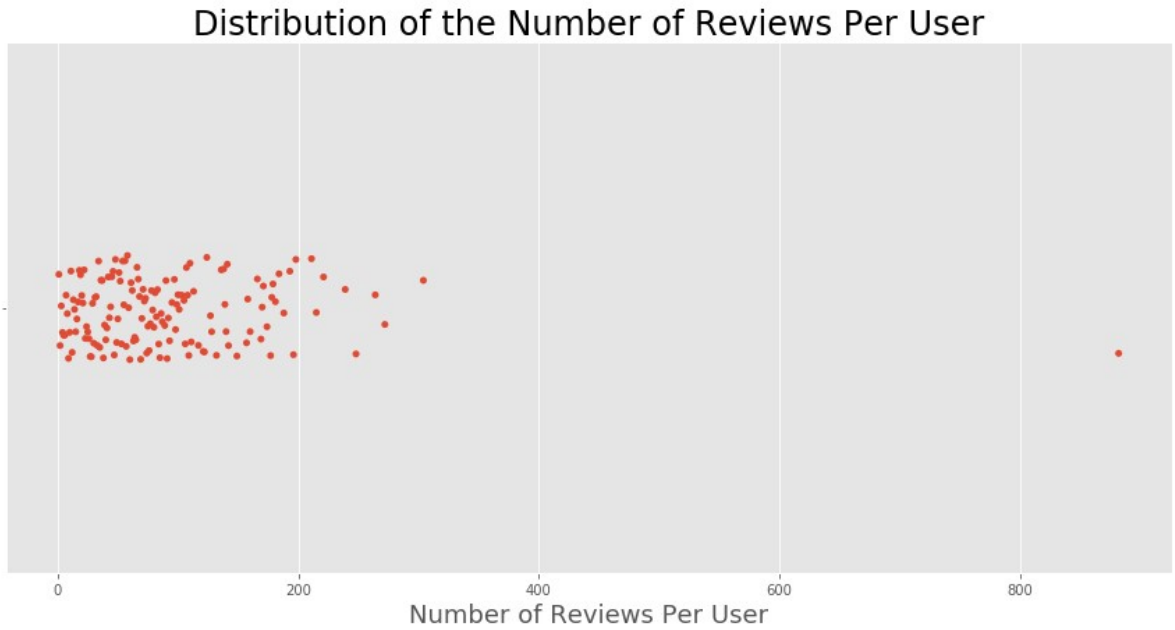
3. Business Reviews in the GTA

3.1. User Review Frequency

We would like to analyse whether a small group of users is responsible for the majority of reviews. Recall that we restricted ourselves to two million reviews from our dataset. When we filter out only reviews about businesses located in the GTA, we are left with about 200 000 reviews.

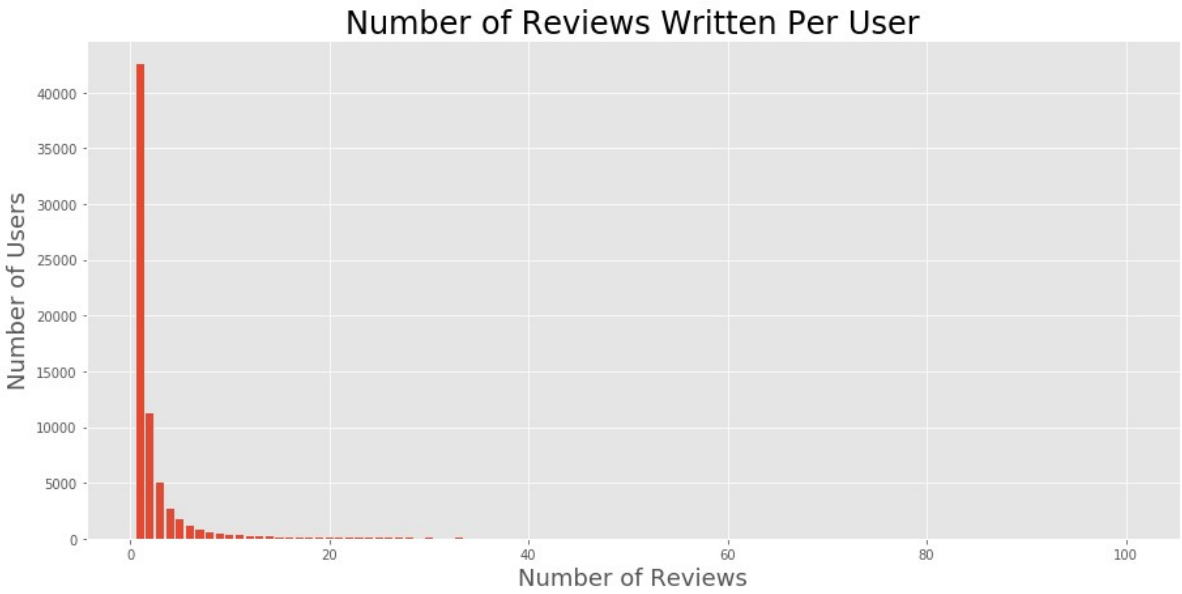
We can count the number of reviews each user has made about businesses in the GTA (we are not looking at users who have made no such reviews), and visualize the distribution.

First, if we do so using a strip plot, we immediately see that there are many outliers in the data:



The majority of users have written fewer than 100 reviews, but in the most extreme case a user has written over 800 reviews about businesses in the GTA.

Let's restrict ourselves to the main part of the distribution of users who have written fewer than 100 reviews to get a better grasp on how this distribution looks like:



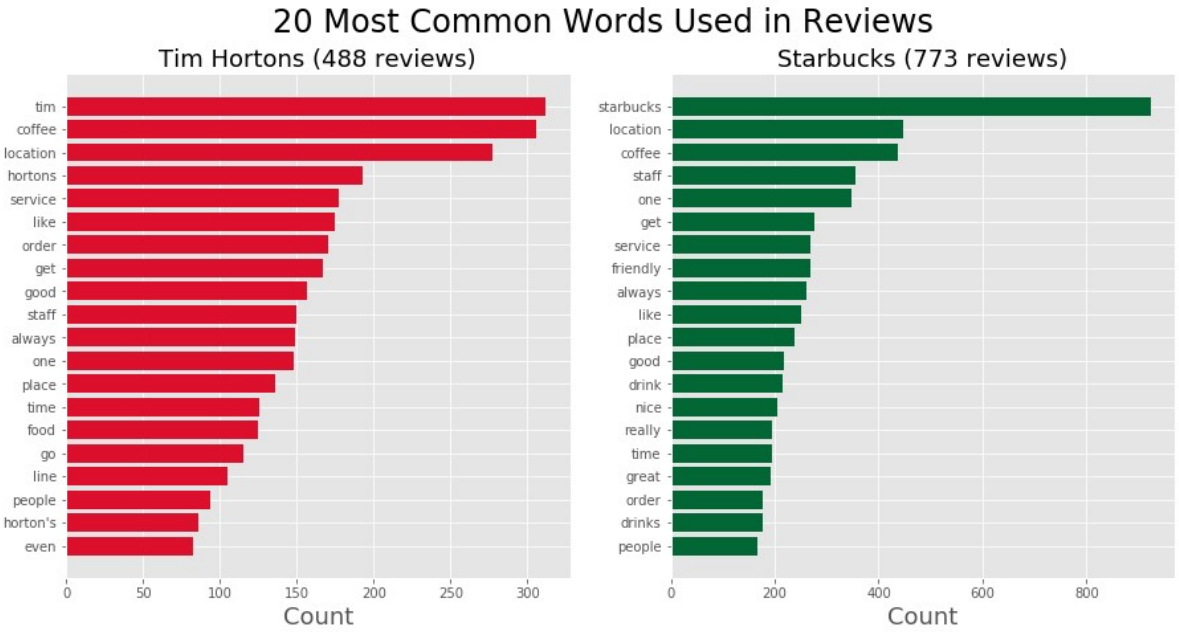
Even with this restriction we can see that the great majority of users have written just a very few reviews. Only on a handful of cases have users written more than 20 reviews - such users are so scarce that they can't all even be seen on the graph.

Thus, it appears to be very clear that it is not true that a small group of users is responsible for most reviews. Instead, most users write just a couple of reviews, and since there are so many such users, their reviews form the great majority of all reviews.

3.2. Language Use in Reviews for Tim Hortons and Starbucks

Previously we investigated the battle between the two coffee giants Tim Hortons and Starbucks. We concluded that Starbucks has been more successful in how they have selected the locations of their shops with respect to their main competitor Tim Hortons. With the review data available, we can analyse how pleased consumers are with these two chains by looking at the reviews written about them.

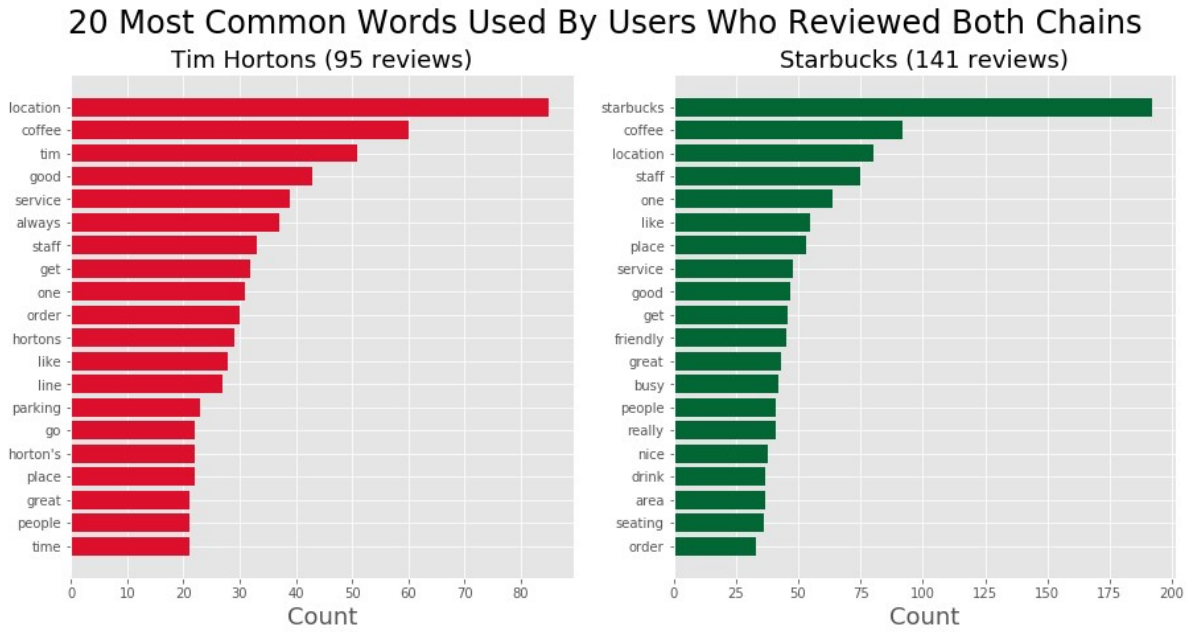
We will filter out reviews written about the shops of the two chains in the GTA, and look at the language used in the reviews.



For both chains one of the most frequently used words is "location", which suggests that the location of coffee shops is significant to customers. Words with positive meaning such as "like" (depending on context, of course), "good", "nice" and "friendly" show up. Words expressing negative feelings and experiences like "busy" are less frequent, which suggests that overall users are satisfied with their experiences in both chains.

However, notice that the list for Starbucks words such as "always" and "really" show up higher up the list - these words express stronger feelings and thoughts, which indicates that perhaps customers are more pleased with Starbucks than Tim Hortons.

For a more effective comparison, we can also look at the reviews of users who have reviewed each chain at least once.



These top lists slightly differ from what we saw when we considered all of the reviews we are using. Words with positive meanings are generally higher up the list for Starbucks than Tim Hortons, suggesting that this subset of customers tend to prefer Starbucks over Tim Hortons.

At the same time, the only word with a very likely negative meaning in these two lists is "busy", which shows up only in the list for Starbucks. Recall that "busy" appeared at the bottom of the top 20 list for Starbucks when we considered all users, but now the word has moved up to the middle of the pack, suggesting that these users are more bothered by this aspect of their experience at Starbucks.

There are 56% more reviews written about Starbucks by these users, which can also suggest that these users prefer Starbucks over Tim Hortons: more reviews can suggest more visits, or that these users are more keen to share their experience at Starbucks. As discussed, it appears that these experiences tend to be positive, so the difference in review counts can suggest that these users' experiences at Starbucks left a stronger positive impression on them.

We have gone through this analysis by only looking at the words used. Without context, we cannot be sure what the actual meaning and sentiment behind these words is, so we cannot draw hard conclusions. More reliable results could be achieved by using more sophisticated sentiment analysis methods, or other natural language processing techniques.

However, our discussion is backed up by actual numbers: the average star rating for Tim Hortons locations in the GTA is 2.6, while for Starbucks the rating is noticeably higher at 3.6. The same is true for the reviews written by users who have reviewed both chains: Starbucks has an average rating of 3.4, but Tim Hortons has a rating of 2.9.

So, it seems obvious that Yelp users prefer Starbucks over Tim Hortons, but we can't say for sure if it is for the reasons we outlined above. "Coffee" is one of the most popular words used in reviews about these businesses, and of course is the main product the two companies sell, so it would make sense to analyse what the users' opinions about each chain's coffee is. With our analysis, we only see that coffee is mentioned, and nothing else.

3.3. Detecting Fake Reviews

Yelp has a massive fake review problem (https://fortune.com/2013/09/26/yelps-fake-review-problem/), to solving which they've allocated a lot of resources. Despite this, no system is impenetrable and most likely there are fake reviews present in our dataset. Yelp is a massive platform, and ranking higher than your competitors on Yelp's website is bound to bring in more customers, so businesses could definitely be motivated to pursue this illegal avenue of purchasing fake reviews. We would like to see if we can detect any reviews we suspect to be fake.

Fake reviews don't necessarily have to be positive. In 2017, the CNN app was bombarded with 1-star reviews, tanking its ranking on phone app stores (https://www.forbes.com/sites/ryanerskine/2017/07/15/how-trump-supporters-tanked-cnns-app-with-1-star-ratings/#5fce16b87e45). However, a common characteristic among fake reviews seems to be that the ratings they provide are extreme - either extremely high, or extremely low.

Users can rate each business as 1, 2, 3, 4, or 5 stars, so let's take a look at reviews that have given the business a rating of either 1 or 5 stars. There are over 85 000 such reviews - over 40% of the dataset we are working with, so this is clearly not enough to find fake reviews.

We previously saw that most users have written just a very few reviews. Writing fake reviews requires creating fake accounts, and this process can be very difficult - there are many systems in place to avoid letting anyone do this. Hence, we would expect that anyone posting fake reviews would use the same account multiple times. So, let's only look at users who have posted at least 30 reviews with a rating of 1 or 5. This cutoff is arbitrary, but is selected such that we have enough users to look at, but we don't end up with too many reviews to go through.

This leaves us with 33 unique users.

Now, we will count the number of extreme ratings (either 1 or 5) and non-extreme ratings (anything inbetween) each such user has given. It would look extremely suspicious if a user only gives out extreme ratings, so it's possible that fake accounts may try to mask this by also giving out ratings with non-extreme values. With this data available, let's look at users who have given out at least 1.5 times (another arbitrary cutoff to simplify analysis) as many extreme ratings as non-extreme ratings.

This methodology leaves us with just two "suspects".

It appears that suspect A is just a very passionate reviewer. They have reviewed 264 unique businesses in 304 reviews, and nearly all of their reviews are fairly long and colourfully written.

Take for example this user's review about Triple M, an accountant's office:

> I've needed a proper accountant for quite some time, but until some other pressing paperwork required I figure out all of my money ish, I ignored this need with all of my might. Enter Triple M. Michael was kind, efficient and did everything for me at a very reasonable cost. His office is located in a really sleek building on King W, and it turns out combing through back taxes was completely painless. Don't be like me and delay, Triple M is here!

This user has only left reviews with a rating of 5, 4, 3, or 2. Finding a shorter and more dull review from this user's list is difficult, so it is fairly certain that this user is not a fake reviewer.

It appears that suspect B is off the hook, too. They have left 178 reviews for 176 unique businesses. Again, it seems that this is just a user who is passionate about sharing their experiences: most reviews are very long and exhaustive. Their longest review is over 800 words long. This no-longer-suspect is more critical in their reviews than on previous occasions, dishing out star ratings of 1 on 18 different occasions.

Curiously enough, this user has posted a bunch of negative reviews on Tim Hortons, and sometimes recommends Starbucks instead. One can only wonder why they keep going back...

All in all, it turns out that detecting fake reviews isn't all that easy. As mentioned, Yelp has invested a lot of time and effort into avoiding fake reviews, so it's reasonable that the way we searched for fake reviews proved to be fruitless - Yelp's algorithms have probably already caught all such reviews, or it was just a case of our own algorithm not being sophisticated enough.

The latter is definitely true to great extent. Yelp is very secretive about how they detect fake reviews because they don't want to give businesses the knowledge on how to avoid getting caught. This system is very complicated, and is definitely not reproducible in the context of this report. Though it isn't unreasonable to think that some of the aspects we pointed out are incorporated in Yelp's algorithms.

Conclusion

The Yelp dataset consists of businesses located mainly in eastern parts of the US, the GTA, and the southwestern parts of the US and Canada. 78% of businesses are situated in 20 cities, the most common of which is Las Vegas. The data has issues with the way city names are entered, so it is possible our data is flawed despite our best attempts to correct it.

Each business has a list of categories corresponding to the type of the business. Looking at the most used categories, the list is dominated by categories falling in the food & drink industry such as “restaurants” and “food”. Such categories are closely followed by categories representing businesses offering services: “home services”, “beauty & spas”. Retail categories such as “shopping” and “fashion” are much less common.

Within these three simplified categories, bike parking is available for over 50% of the businesses in the food & drink industry, but least commonly available for around 30% of businesses in the services industry. More specifically, bike parking tends to be available most commonly for businesses that provide bike or active lifestyle services, and for businesses that are focused on serving alcohol. Whilst the former two types of businesses offer no surprise, it may be that businesses serving alcohol make bike parking available to discourage customers from driving cars while under the influence of alcohol, even though riding a bicycle while intoxicated is also illegal. A case-by-case analysis would have to be conducted to come up with a more certain explanation for this phenomenon.

There is very little association between review count and star rating. Our fitted linear regression model was able to explain just 0.2% of the variance in the data. Even though higher star ratings are more common for businesses with more reviews, it appears to be a product of the underlying distribution of ratings – most ratings fall in the 4.5-3 range. Generally, businesses in the GTA are similar to those present in the entire dataset. However, we noticed that ethnic food category businesses are more prominent in the GTA. Food & drink industry businesses form a greater proportion of all businesses in the GTA than was the case for the entire dataset. This mostly comes at the expense of businesses providing services, of which there are proportionally fewer in the GTA.

The two most common franchises in the GTA are Tim Hortons and Starbucks, which have nearly double the locations as the 3rd most popular franchise, McDonald’s. Apart from two businesses, the 20 most common businesses in the GTA are in the food & drink industry. Just like city names, there are inconsistencies in business names. Again, we attempted to fix these issues as well as possible, but we were limited by hardware, and the small aspects of the data itself, potentially skewing our analysis.

Most businesses in the GTA are located in downtown Toronto, Mississauga and Markham. Businesses in those areas tend to receive more reviews on average, too. The businesses with the highest average reviews are generally situated in downtown Toronto, along the coast of Lake Ontario, and to the west of King City. Well-designed free GIS tools are scarce, so our visual analysis was tampered with by unavoidable artifacts. Even though a critical eye can catch these artifacts, it is possible that they caused us to misinterpret some of these results. These errors could greatly be mitigated by using specialized GIS software.

It cannot be claimed that there is a Starbucks right next to every Tim Hortons location. Only on fewer than 25% of the cases is the closest Starbucks to a Tim Hortons within a 100m radius. On average, a Tim Hortons shop’s closest neighbouring Starbucks is about 600m away. We divided the GTA into Thiessen polygons around Tim Hortons locations and saw that Starbucks has placed their shops in relation to Tim Hortons locations in such a way that they have a dominant position in most of the GTA. This analysis could benefit a lot when combined with population, income, commute, or infrastructure data to better understand what the “value” of each such polygon is to these coffee chains. Our analysis introduced a measurement error on the scale of centimeters – insignificant to our analysis. A much bigger error was caused by how we defined the boundaries for our area of interest – instead of eyeballing it, a more complicated solution would provide considerably more reliable results.

Generally, most Yelp users write just a couple of reviews, forming the great majority of all reviews. When looking at the types of words reviewers most commonly use when talking about Tim Hortons and Starbucks, we saw that the overall mood appears to be positive for both businesses. However, it seems that users have stronger and more passionate feelings about Starbucks. The same phenomenon seems to be the case when comparing reviews for users who have reviewed both businesses at least once. This is backed up by the actual ratings these businesses have: Starbucks has an average rating of 3.6, but Tim Hortons falls short with a rating of 2.6. Without proper context, we lose out on a lot of information by just looking at the words used. Instead, a more sophisticated sentimental analysis or similar natural language processing technique would