A2 Report - Wu, Sixuan

Exploring Yelp

Author: Sixuan Wu

Date: February 15, 2020

Introduction

Yelp is a business directory service and crowd-sourced review forum, and a public company of the same name that is headquartered in San Francisco, California. The company develops, hosts, and markets the Yelp.com website and the Yelp mobile app, which publish crowd-sourced reviews about businesses. It also operates an online reservation service called Yelp Reservations.

--adapted from Wikipedia

The development of Yelp improves people's quality of life in many ways. For example, citizens can find a nice restaurant to enjoy delicious food during weekends or they can find a fitness center nearby where they can do daily exercise. Therefore, the importance of Yelping data analysis is getting more essential which can provide public more high-quality entertainment places.

In this report, conclusions will be from three main aspects: all business on Yelp, business in GTA(Great Toronto Area (https://en.wikipedia.org/wiki/Greater_Toronto_Area)) and users reviews. The results of this report can be helpful to improve the user experience and know more about establishments in the North America and more detailed in GTA.

Data Description

The dataset I used is from Yelp Dataset Challenge (https://www.yelp.com/dataset/challenge) which is provided by Yelp officially. The json format dataset which is compressed in tar format can be downloaded by providing email and name. Following are related permissions (https://s3-media1.fl.yelpcdn.com/assets/srv0/engineering_pages/06cb5ad91db8/assets/vendor/yelp-dataset-agreement.pdf) of the dataset:

Cannot Dos:

- Use the data to create or update my own business.
- Give the data to the third party without permission of Yelp.
- Give the data to others to make profit.

Can Dos:

- Use information from data to do academic project in for this course.

Extract Data

The downloaded data is in tar compressed file, the first thing is to extract data to json format data. There are several json formatted datasets after extraction:

- business.json
- checkin.json
- photo.json
- review.json
- tip.json
- user.json

Observe Data

By observing the datasets listed above, here is the structure and relationship of those datasets:

Dataset	Structure	Relationship
business.json	**business_id**: id to represent the merchant **name:** name of the merchant **address:** specific address of the merchant **city:** city where the merchant located **state:** state where the merchant located **postal_code:** the postal code of the merchant **latitude:** latitude of the merchant **longitude:** longitude of the merchant **stars:** stars given by customers **review_count:** number of reviews provided by users **is_open:** whether the shop/restautant open, 1 for yes and 0 for no **attributes:** some special sevices provided by the merchant **categories:** some features of the merchant **hours:** opening hours of the restaurant	Detailed information of each merchants registered on yelp.
checkin (https://blog.yelp.com/2018/12/perfect-yelp-check-in-offer).json	**business_id:** id of the merchant **date:** exact time the bussiness check in	The exact check in time of the merchants which share the same business_id in business dataset.
user.json	**user_id:** id of the user **name:** name of the user **review_count:** number of reviews given by the user **yelping_since:** time the user join in Yelp **useful:** number of useful recieved by the user **funny:** number of funny recieved by the user **cool:** number of cool recieved by the user **elite (https://www.yelp-support.com/article/What-is-Yelps-Elite-Squad?l=en_US):** the elite year of the user **friends:** id of friends of the user **fans:** number of fans of the user **average_stars:** average stars given by the user **compliment_hot:** number of recommandations the user compliment hot **compliment_more:** number of recommandations the user compliment more **compliment_profile:** number of profiles the user compliment **compliment_cute** number of recommandations the user compliment cute **comliment_list:** number of lists the user compliment **compliment_note:** number of notes the user write for compliments **compliment_plain:** number of recommandations the user compliment plain **compliment_cool:** number of recommandations the user compliment cool **compliment_funny:** number of recommandations the user compliment funny **compliment_writers:** number of writers the user compliment **compliment_photos:** number of photos the user compliment	Detailed information of each user registered on Yelp.

Dataset	Structure	Relationship
photo.json	**caption:** comments provided by users **photo_id:** id of the photo **business_id:** id of the merchant the photo indicates **label:** inside or outside of the restaurant/shops the photo shows	Photos provided by users to describe the shop/restaurant. Share the same business_id in business.json and same user_id in user.json
review.json	**review_id:** id of the review **user_id:** id of the user who write this review **business_id:** id of the business which is the described by this review **stars:** stars of the review **useful:** number of users think the review is useful **funny:** number of users think the review is funny **cool:** number of users think the review is cool **text:** detailed content of the review	Reviews provided by users to describe the shop/restaurant. Share the same business_id in business.json and same user_id in user.json
tip.json	**user_id:** id of user who gives the tip **business_id:** id of merchant who recieve the tip **text:** some comments given by users **date:** the date when the tip is given **compliment_count:** number of compliments	Tip information between customers and merchants. Share the same business_id in business.json and same user_id in user.json

Note: The datasets are stored in json format, however each line in the dataset is a json object. Therefore I will parse the dataset for each line.

All Business

What cities does this dataset encompass?

Overview

Out[5]:
```
count        192609
unique         1204
top       Las Vegas
freq          29370
Name: city, dtype: object
```

There are 192609 restaurants locates in 1204 cities in North America use Yelp to attract more customers. In Las Vegas, there are 29370 restaurants decide to use Yelp which is an amazing quantity and is also the city with most restaurants number use Yelp among North America.

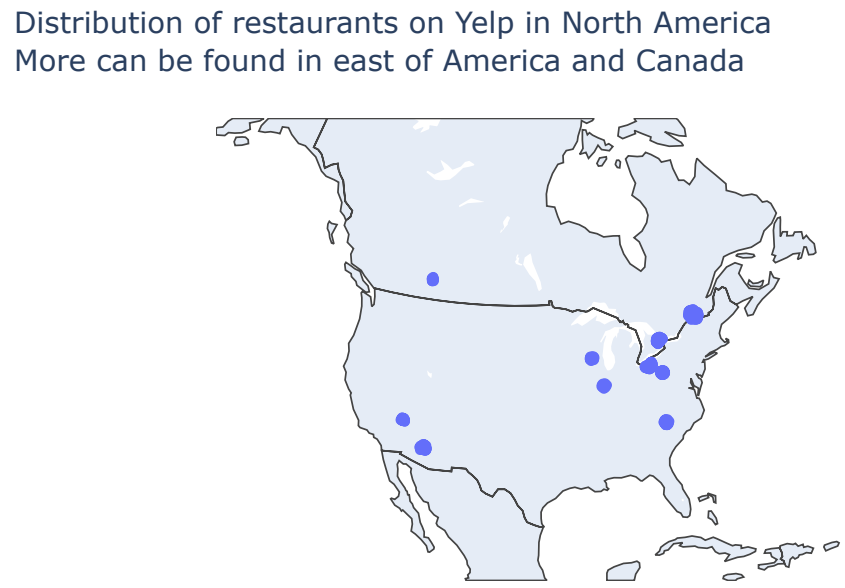
Distribution Visualization

In order to gain more insight into the cities that the dataset encompass, cities with restaurants use Yelp will be plot on the North America map.

Method

1. Unite the longitude and latitude of restaurants located in the same city.
2. Calculate the number of restaurants using Yelp in each city.
3. Plot those cities on the North America map.

Out[9]:



Each point on the map represents a city.
City name and number of restaurants using Yelp in the city can be found by point to the point.

Conclusion and Discussion

By observing the distribution of the cities, it is interesting to find there are two characters:

- Cities with Yelp applied gather in some regions.
- More cities in the east of the US and Canada use Yelp than in the west of the US and Canada.

According to the first observation, the development routine of Yelp can be known. When Yelp is used by a restaurant in a city, after other restaurants finding the impact of Yelp, then other restaurant will start using Yelp. After some time, business in surrounding cities will use Yelp to propaganda. This can explain why the cities use Yelp gather together. In this way, Yelp can only do propaganda in some big cities, and then satellite cities will be influenced by the big city. Therefore the scale of Yelp can be expanded in the area. In this way, other business similar to Yelp can use same way to develop.

The other interesting point is that Yelp is a company in San Francisco which locates in the west of the US, however, Yelp seems more popular in the east of both the US and Canada by the second observation. Therefore, further research will be done to figure out whether Yelp is more prevalent in the east than the west.

However the limitation of this method is if there is a typo in the dataset, then the business in the same city will be considered as business in two cities. The limitation will not influence the result too much because the business are showed on the map according to the exact position provided by latitude and longitude. Therefore the visualization of the distribution of business will not have a huge difference.

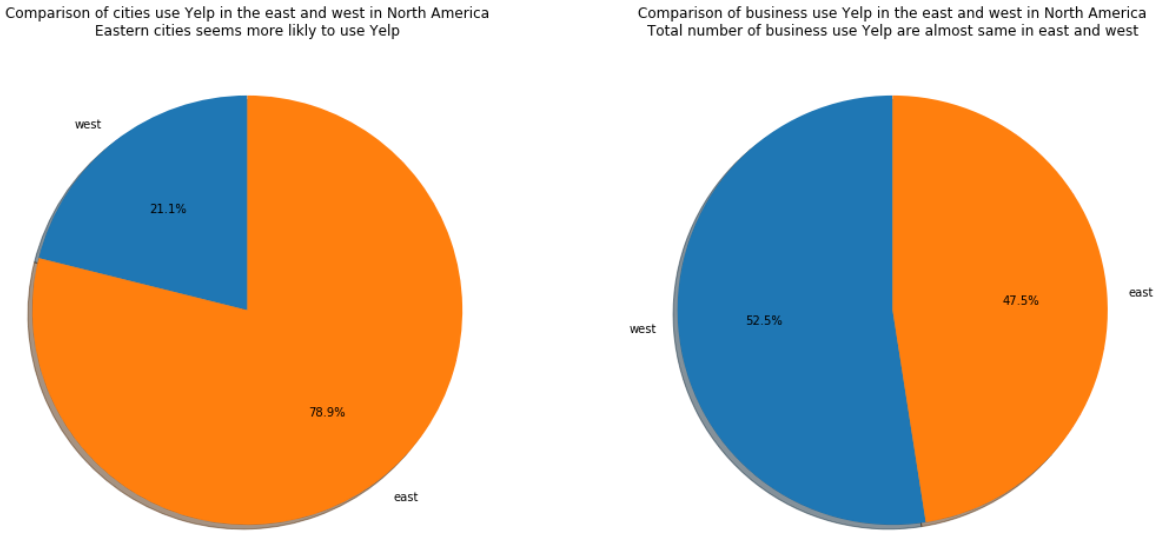
Compare Popularity of Yelp

We have had an intuitively recongnition of the distribution of business on Yelp in the last part. In this part, popularity of Yelp in western North America and in eastern North America will be analysised.

Method

1. Classify cities into two regions: west and east. The benchmark of classification is if the longitude of a city is less than -100^o, then the city is considered as western city, otherwise, the city will be considered as eastern city.
2. Count total number of cities in the west and east respectively.
3. Count total number of restaurants/shops in the west and east respectively.
4. Visualize those data by pie charts.

The dividing line of western and eastern should be Mississippi River actually. However I did not use this dividing line because the dataset still contain the business located in Canada and it is unnecessary to use Mississippi River as the dividing line. -100^o longitude line can also divide western cities and eastern cities as required.



Conclusion and Discussion

The pie char in the left represents the comparison of the number of cities in the east of North America and the number of cities in the west. On the other hand, the pie chart in the right above indicates the proportion of restaurants in the east and west seperatly.

1. Number of cities use Yelp in the east of North America is more than the number of cities in the west.
2. Number of business use Yelp in the east and west are almost same.

From the observation, results can be gained:

1. More cities would like to use Yelp in the east than in the west although Yelp is located in San Francisco, Yelp is more prevalent in eastern part of North America.
2. Compare to the popularity of cities use Yelp, total number of restaurants in east and west are almost same, which means the density of restaurants use Yelp is higher in the west.

The conclusion is **Yelp is more popular among cities in the east than in the west while it is more prelavent in each of cities of the west than of the east**. Therefore, Yelp can concentrate more on the develop in the western cities in the North America.

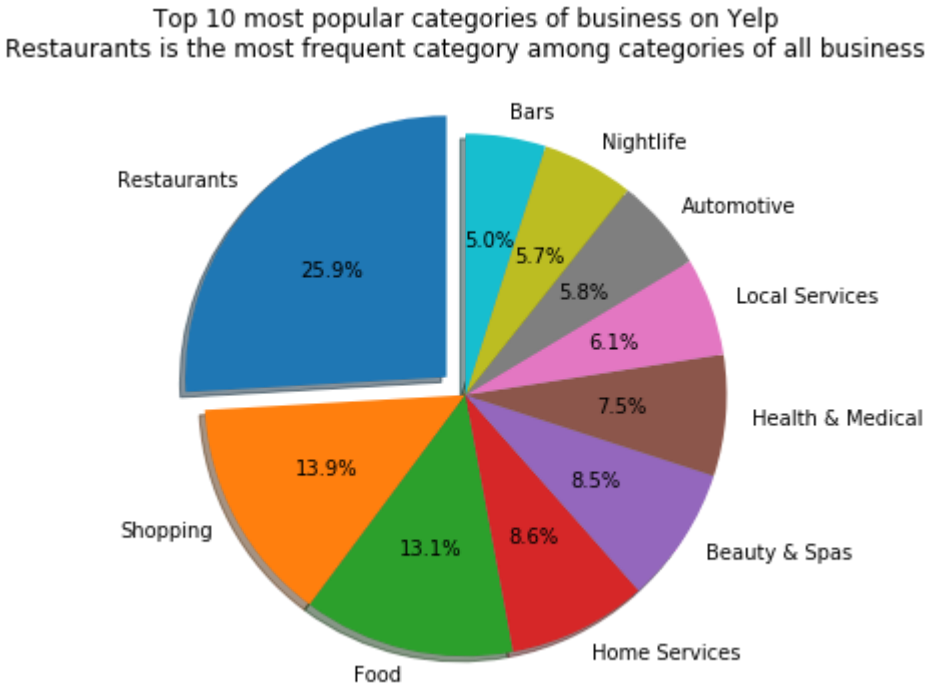
What are the most frequent business categories overall?

Overview

According to the data we have, there are total 1300 categories. It will be interesting to find what are the most frequent business categories. Then we can know more about the demand of public for each categories of business which can be helpful for those people who want to open their own business.

Method

1. Extract categories in each business.
2. For each category, count how many business in the category.
3. Visualize top 10 frequent business categories in bar plot.



Conclusion and Discussion

The bar plot above indicates the top 10 categories in among all business in North America registered on Yelp. There are 59371 business are classified as restaurants which is also the most frequent business which takes up 25.9% among all categories of business.

According to this result, preference of public can be found which is related to **eating and shopping**. Therefore, the hypothesis can be come up with that people are fond of spending their spare time on delicious food and shopping. More business related eating and shopping can be opened in the future because the huge demand of these two categories. On the other hand, large quantity of these two categories of business also indicate the intense competition.

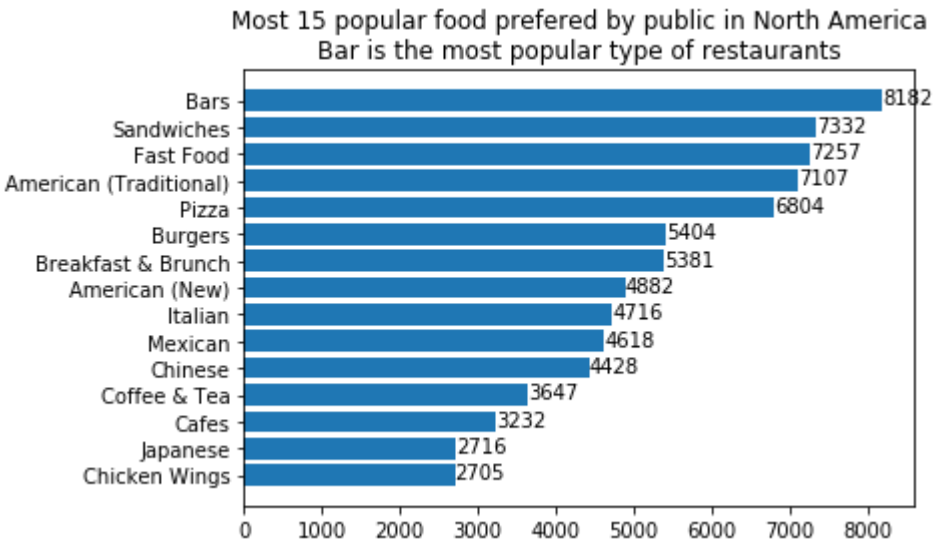
Nevertheless, the limitation of this method is there are thousand number of categories in total, some categories may have relationship to each other, which means some categories may be the subset of other category or they may have a strong reltion. For example, business with category "Chinese" have huge probability of possession of category restaurants. In this way, some categories are hard to be extracted. However, the limitation does not influence the result of the most frequent business too much.

Which food do public prefer?

According to the result above, restaurants is the most frequent business registered on Yelp which implies the huge demands of public, therefore it is necessary to do more research into which kind food public prefer.

Method

1. Filter restaurants from all business.
2. Count each categories form restaurants.
3. Visualize the result by bar plot.



Conclusion and Discussion

Based on the bar plot above, it is not hard to find there are 8182 bars on Yelp in total which is the most prevalent type of restaurant. This phenomenon implies a high demand for bars of people.

In addition, sandwiches and fast food take the second and third places respectively.

Most data in the dataset are belongs to the US, the result indicates that American people prefer sandwiches and fast food a lot. Therefore, a new American restaurant had better to choose to provide public more sandwiches and fast food which are the most popular food in the US. Then if a person open a bar or a sandwiches restaurant, it might be a better choice.

Limitation of this method is similar to the last one. Categories are complex and messy, it is easy to have relation between one or more categories. For example bars often represents nightlife, cages may indicate Coffee & Tea, in this way, some also may be missed which means some restaurants are counted as cafe but not coffee & tea. This may have influence on our result.

What types of establishments tend to have bike parking?

Bicycle is a convenient transportation tool and lots of people prefer riding bikes to walking or driving. However, there may be some business do not offer bike parking. Thus, it is important to find out those business are likely to offer bike parking service. If the business with same categories but there is no information on bike parking, people can use this result to determine the transportation they use.

Method

1. Extract the business with information of bike parking.
2. Classify business with bike parking and without bike parking into two dataframes.(none bike parking is considered as business without bike parking because it can be interpreted as there is none bike parking of the business and those business without bike parking information are disgarded)
3. Find the number of each business in each category in each dataframe.
4. Define parking coefficient of a category: parking_coefficient = $\frac{\frac{number\ of\ category\ with\ bike\ parking}{number\ of\ business\ with\ bike\ parking}}{\frac{number\ of\ category\ without\ bike\ parking}{number\ of\ business\ without\ bike\ parking}}$
5. Visualze the result.

Explanation to the method

It is rational to use this method to measure the trend to have bike parking for a category because it is easy to come up with the idea to compare the number of the bussiness in the category with bike parking and without bike parking.

However the base is different which means the total number of business with bike parking and without bike parking are different. Thus the total number of business without bike parking and with bike parking divide the number of business with or without bike parking in the category respectively which is to standardize the number of business in the category.

Finally the division of these two values can represent the tendency of the business in the category is whether with bike parking or without bike parking.

Therefore, 1 can be used as the dividing point. If the coefficient is greater than 1, the trend will be providing bike parking and vice versa. The higher result represents the type of establishment is more likely to have bike parking area.

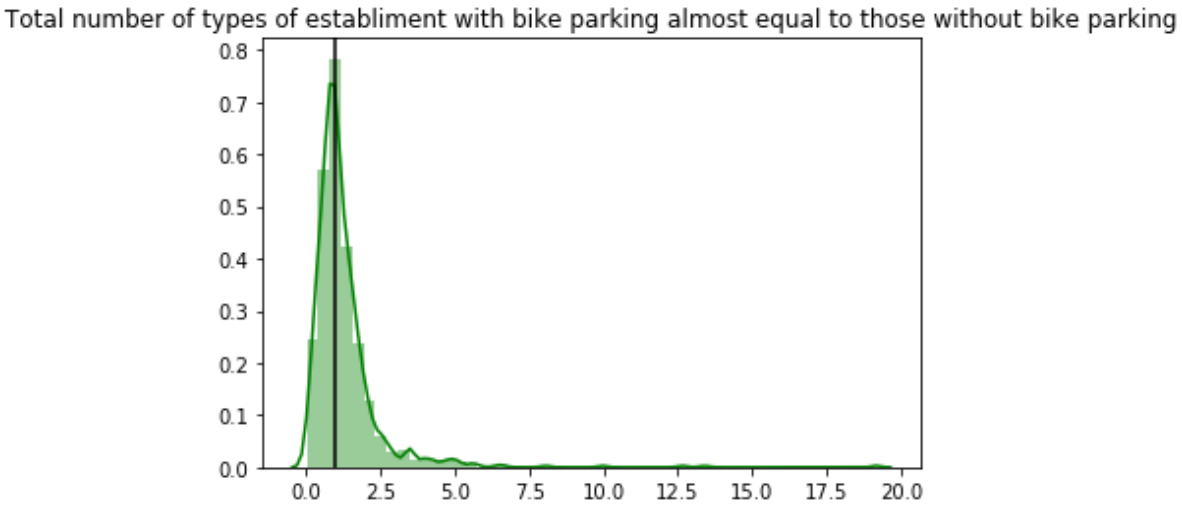
Advantage

There is another method to define this coefficient: $\frac{number\ of\ category\ with\ bike\ parking}{number\ of\ category\ without\ bike\ parking}$. It is easy to understand that the higher the coefficient is, the high probability that the business providing bike parking. However, this coefficient can only be used to compare which business is more likely to have bike parking. If there is only one category, this method cannot show the trend of providing bike parking.

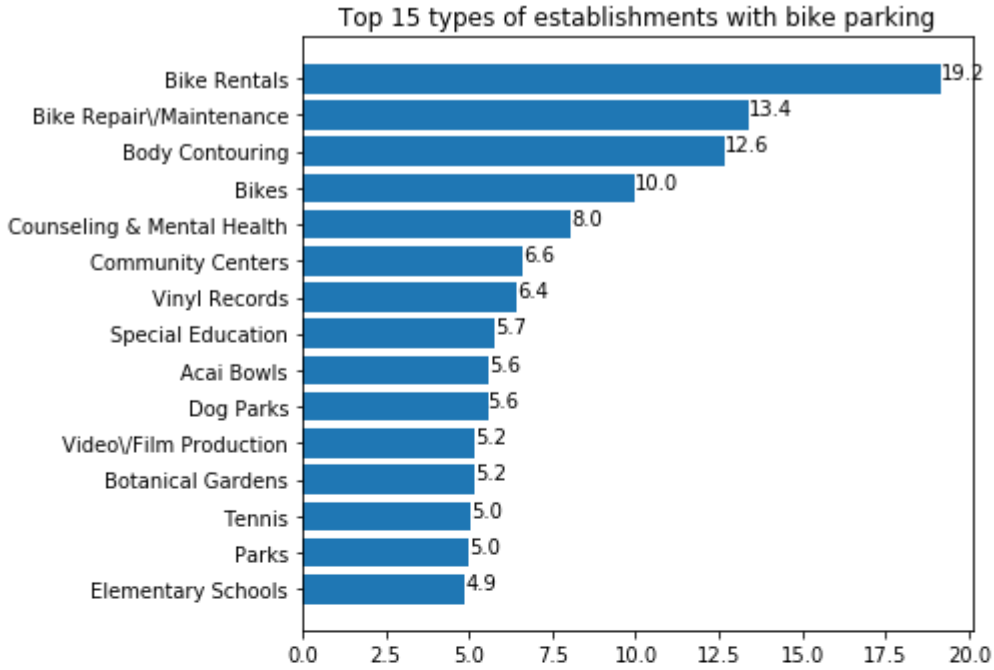
Therefore the advantage of this method is we can use 1 as the dividing point which provides us with a benchmark to measure the trend of a category of business providing bike parking. Even if there is only one category, this coefficient can also show us the tendency.

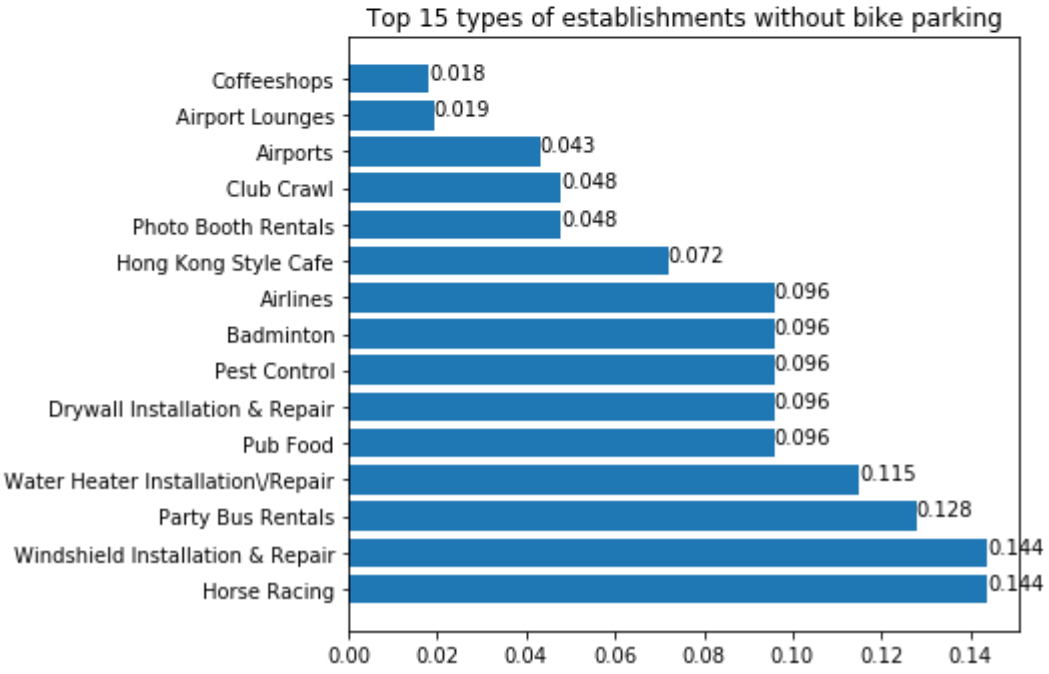
Limitation

- Information of several categories is limited and the denominator of the formula above sometimes can be 0 which means all business in such category are with bike parking and vice versa. In this way, those categories with this phenomenon will be analysised seperately.
- The range of this coefficient is [0, $+\infty$) which is not symmetic. Therefore it is hard to show the extent of tendency of providing bike sharing. If use subtraction to subsitute division which solve this problem. However, the result of subtaction is not easy to be visualized because the coefficient of each category is extremely near.



According to this histogram, we know the total number of business with bike parking almost equals to the number of business without bike parking area.





Conclusion and Discussion

Business related to bikes are most likely to have bike parking are while coffee shops and places related to airports are the places most impossible to provide bike parking area which makes a lot sense.

The type of business are all with and without parking area are analysised as following.

There are two possibilities for this phenomenon:

- The number of business of that type is extremely small, for this case, the information can be discarded because the result wll be meaningless.
- There are a lot of sample, however, all the sample in such type have the same attitude on bike parking. In this case, the result need to be considered.

However, after analysising these two datasets, the largest number of business in the all business with bike parking dataset is 14 which are baseball fields, while the largest number in the other dataset is 3 which are airport shuttles. Compared to the total business with bike parking information dataset, these numbers are too small which means these business can be discarded because even if analysis them, the result cannot be accurate enough.

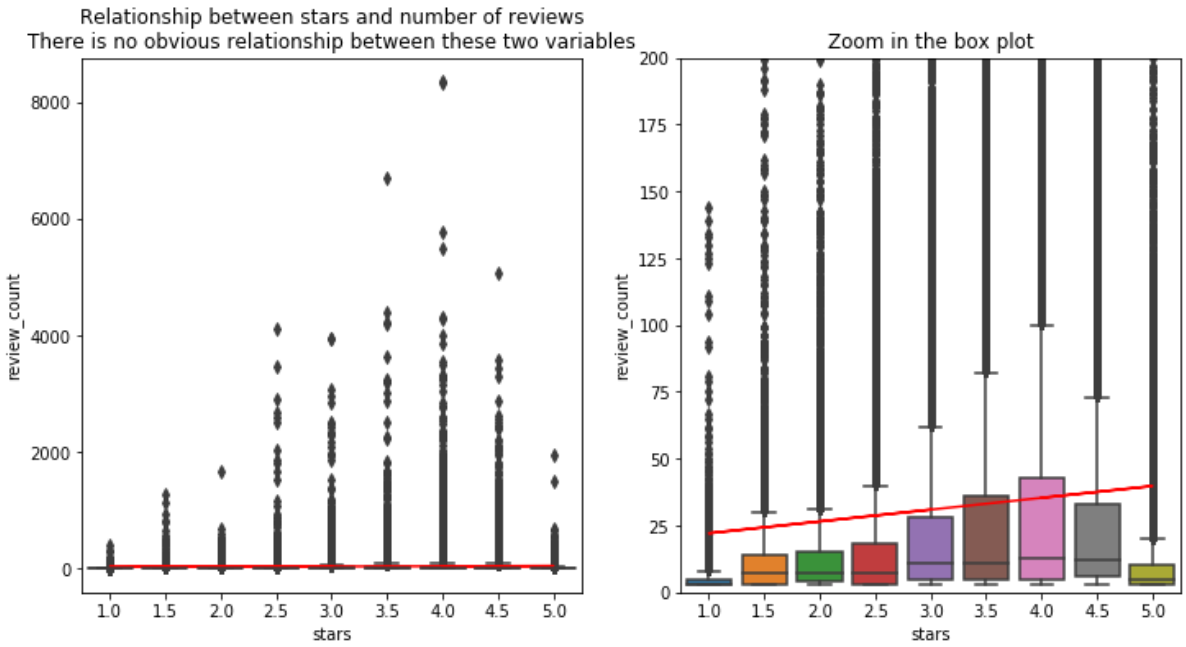
An article recently claimed that having more yelp reviews lead to a higher rating, and hence increased sales. Do the data support this claim?

Overview

Yelp provides review function which allow users to show the feeling to each business. As for business, they may more concern about how many customers will give them feedback and how are the comments. If a business gains many yelp reviews, the business can know their advantages and disadvantages, therefore they may improve themselves to gain a high rating. In this part, I will find whether there is a relationship between number of reviews a business received and the rating.

Method

1. Do linear regression of review_count and stars.
2. Visualize the resualt in box plot and do the linear regression between stars and review_counts.




```
The correlation coefficient is 0.04
```

Conclusion and Discussion

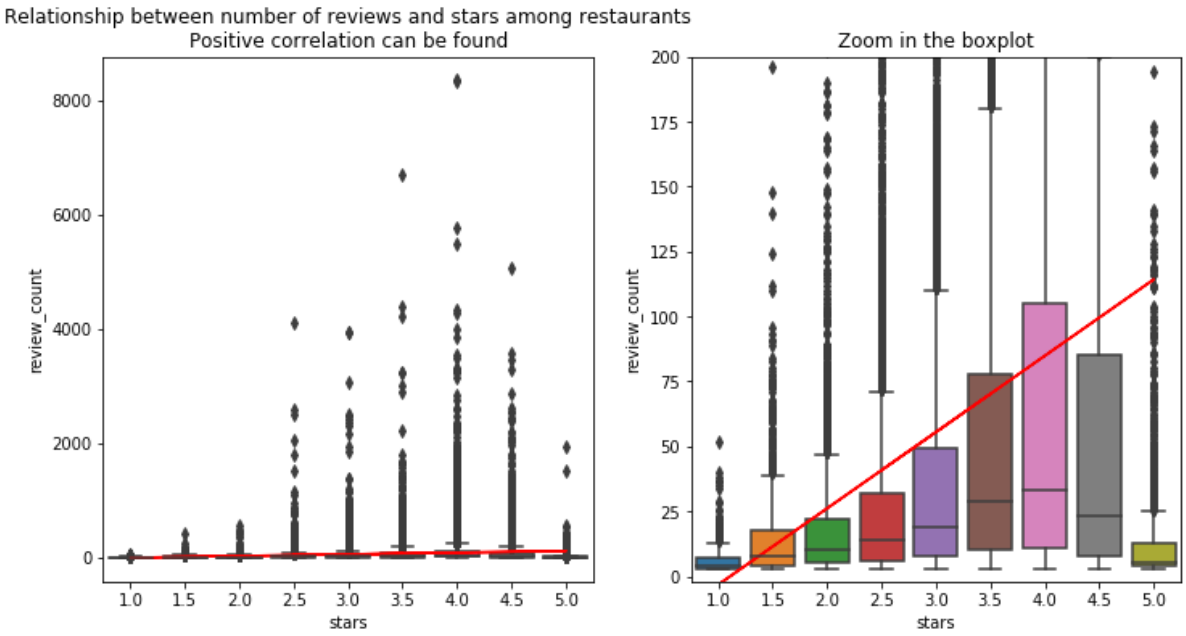
According to the boxplot above, the median of each star category have no big difference. By the result of linear regression, **there is no obvious relation between stars and number of reviews.** Therefore, more reviews of a business cannot imply higher rate it can obtain. I will do more research in restaurants data to see whether there is a relationship between these to vaiables which may provide new idea to help users select restaurants.

Relationship between number of Yelp reviews and rating for restaurants

According to all of the business on Yelp, **more reviews does not imply higher rates**. However, it is still necessary to do more research into the same question but narrow the range of business to restaurants. The previous result indicates public have most demand for restaurant resources because of high quantity of restaurants on Yelp. Thus, this research may provide a new method to predict the quality of a restaurant.

Methods

1. Do linear regression between review_count and stars.
2. Visualize the result by boxplot.
3. Calculate correlation coefficient.




```
The correlation coefficient is 0.13
```

Conclusion and Discussion

By observing the boxplot and the line of linear regression, **a positive correlation can be found between number of reviews and stars among restaurants dataset**.

To guarantee the existence of positive correlation, correlation coeffcient is calculated by

$$r = \frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma_X}\right)\left(\frac{Y_i-\bar{Y}}{\sigma_Y}\right)$$

Then the correlation coefficient is 0.13 which means there is a weak positive relation between review_count and stars these two variables in the restaurants dataset.

This results is quite amazing compared to the last result. Positive relation between number of reviews and rating cannot be found when the dataset contains all business while can be found when the dataset only contains restaurants. Therefore the statement in the article is partially true beacause the phenomenon exists when for business relate to restaurants. In this way, public also can select restaurant not only by the stars but also by the number of feedback of other customers. However, this may also cause business hire people to provide positive reviews to themselves which can attract more customers.

I did not remove any outliers in this part because there really exists business with more than 8000 reviews which is true. Therefore, there is not any reason to remove outliers when doing data analysis. However, there may be some restaurants may hire "paid reviewers" which represents those people given money by business write positive comments. In this way, the number of reviews may increase a lot while it can still have a low star rating.

Another concern on this question is although we can observe a positive relationship between number of reviews and star ratings, there is no data for us to analysis whether this can increase the profit for business. Therefore, the further research should be done if we can have the dataset which provide us with the profit information for each business. After we do the further research, the most accurate conclusion can be gained.

GTA businesses

Data Cleaning

Business only in GTA(Great Toronto Area) will be analysised in this part, then before analysis, data need to be cleaned which means business in GTA need to be selected and those business are not in GTA need to be dropped.

Method

1. Filter business locate in Ontario.
2. According to the postal_code, filter out business in GTA.

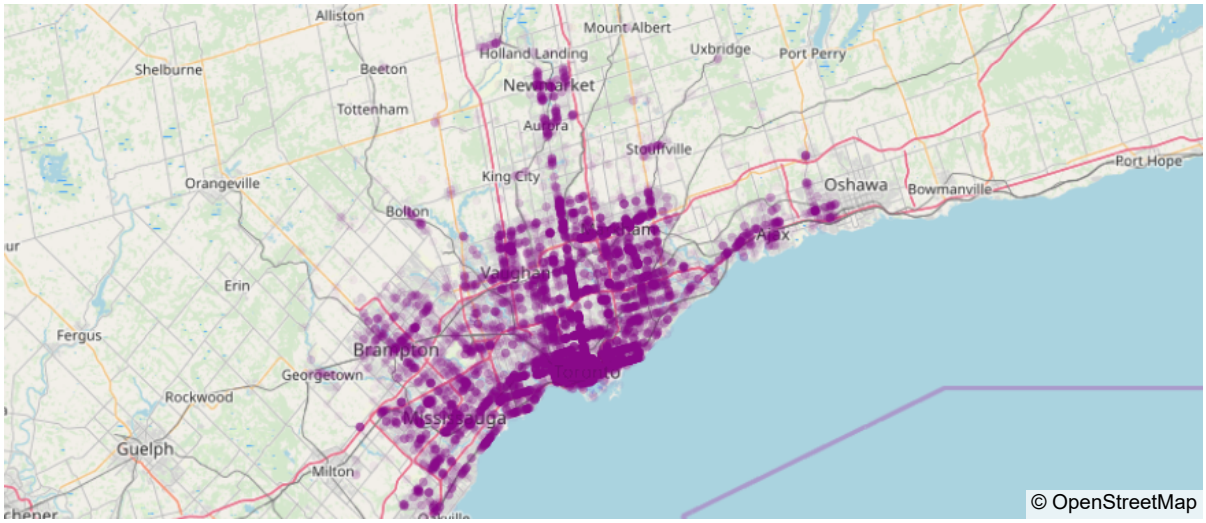
Explanation

This method is reasonable because according to the Forward Sortation Area (https://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br03396.html), cities in GTA are with postal code started with "M" or "L". Therefore, business in GTA can be selected in this way.

Overview

In order to the GTA business, I will first show all GTA business on the map which will provide us a perceptual conginition of the distribution of GTA business. Then I will do further research in more detailed questions.

Out[26]:



The plot above indicates all business in the Great Toronto Area, each purple point represents a busines. Then the area with more business will be shown in a darker color on the map above. We can see the density of business is the highest in downtown area and there are a lot of business on Yonge street.

What are the most frequent business categories?

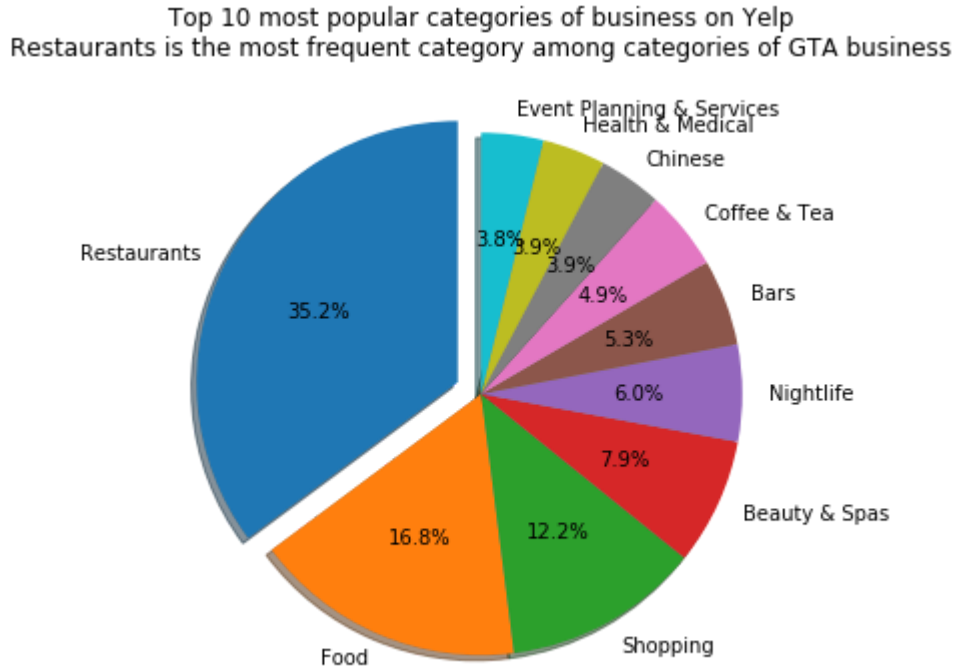
Overview

For all business in the North America, the frequent business category is restaurant. However is the answer same if we narrow the restriction to Great Toronto Area? In this part, I will find the most frequent business category in GTA.

Method

The method used here is similar to the method in the last part where the most frequent category is found for all business on Yelp.

1. Discard those business which did not declare the categories.
2. Find the top 10 popular categories from the dataset.
3. Visualize the results by bar plot.



Conclusion and Discussion

1. Compared to the categories among all business, the same point is restaurants is still the most prevalent category on Yelp.
2. The different point is public in **GTA seems to have a larger demand for restaurants**. The number of restaurants is nearly 3 times of the number of shopping while this is nearly 2 times in all business analysis. In addition, there are only 4 categories which is not related to eating in top 10 GTA business while the number is 6 among all business.

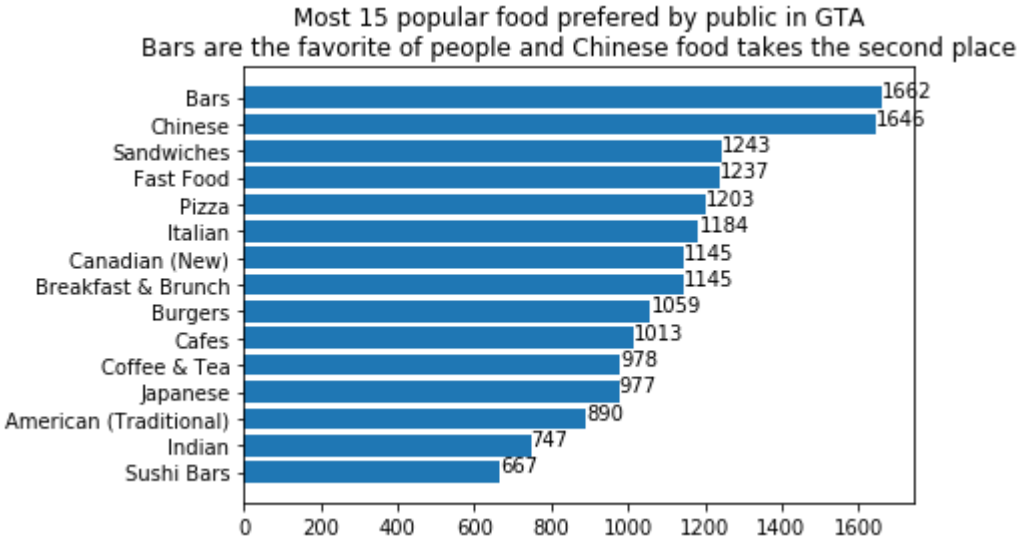
Therefore, restaurants seems to be the most important and popular business in GTA. Then restaurants data will be done further research to find the difference of flavors between people in GTA and whole North America.

Most frequent restaurant categories in GTA

Restaurants is the most popular category among all business in no matter GTA or North America which indicates the extremely large demand of people on restaurants. However, there lots of kinds of restaurants such as Chinese, Italian and American. Will the flavor of public change because of different restriction area and which kind of restaurant is the most popular one? I will analysis further in this part.

Method

1. Select restaurants from all gta business.
2. Extract each catefories from each restaurants.
3. Visualize and analysis the result.



Conclusion and Discussion

This result is interesting, the bar plot can be interpreted as following:

1. Compared to the business among all North America, bars is still the most frequent category in GTA.
2. The different point is that rather than sandwiches, Chinese food take the second place. This result implies that **Chinese food is more prefered in GTA and the density of Chinese people in GTA is much higher than in whole North America**. According to this distribution of restaurants, density of each ethnicities also can be predicated in the area.

The correctness of this result can be partially proven by the rank of tranditional American food among business and the rank of new Canadian food in GTA. In this way, more business in GTA can be opened related to Chinese culture to earn more profit.

It is hard to find the most accurate category for each restaurant. For example, there may be an Italian food restaurant is only labeled as restaurant but not as Italian restaurant. In this way, the number of restaurants in each category maybe a little bit lower than the real value.

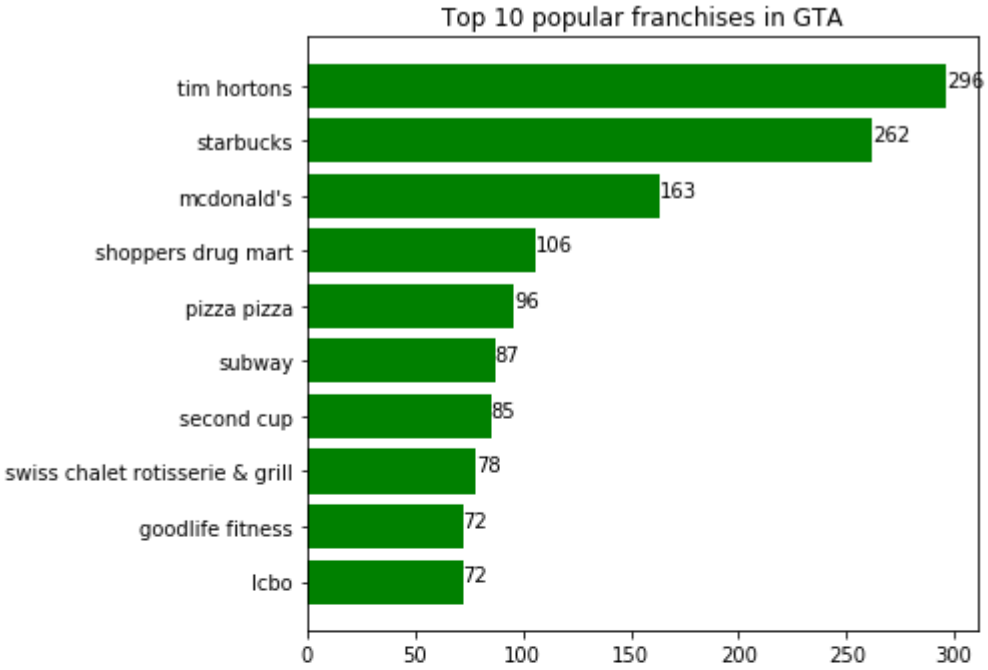
What are the top franchises in the city?

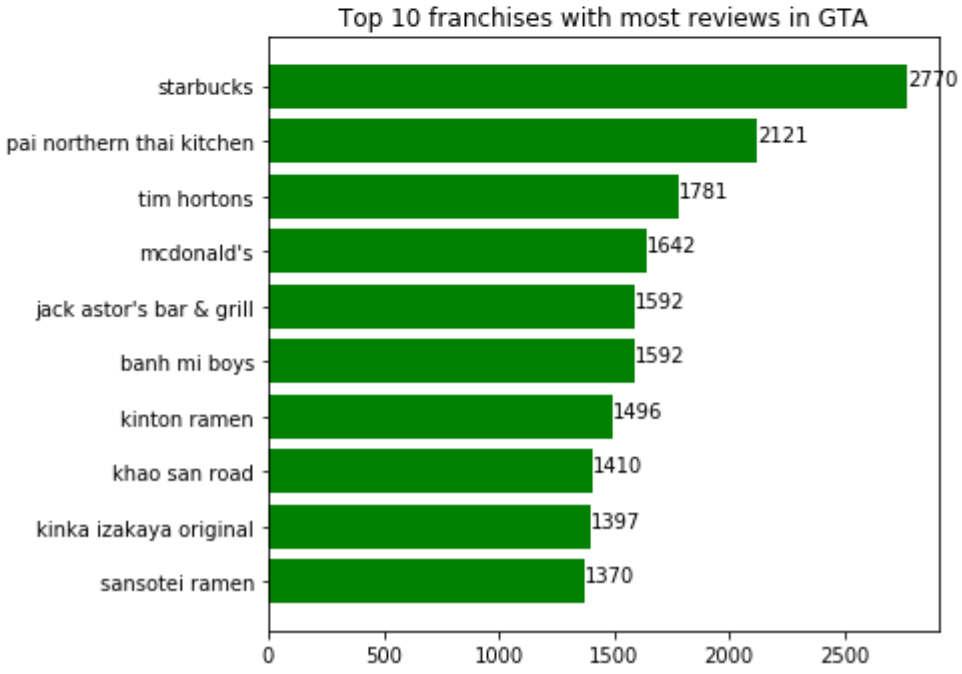
In this part, top 10 popular franchies in the city will be figured out. However, some fanchise in the city are with different names but they are actually same. Therefore, some data cleaning will be done as following method.

Method

There are two benchmarks to defind top framchises: popularity(density i.e. number of the franchises in GTA) and number of reviews. It is necessary to analysis this question in these two respects seperately and analysis the relationship between these two benchmarks.

1. Find top 20 popular franchies in the city.
2. Find the proper name by using fuzzywuzzy
3. Clean the name for these franchies manully because this will not have big impact on the final result.
4. Find top 10 popular franchies in the city.





Conclusion and Discussion

Tim Hortons is the most popular franchises in GTA and Starbucks take the second price and Goodlife Fitness is the franchise with largest density in GTA which is not related to food & restaurant while Starbucks is the franchises with most reviews. It is interesting to find that more popular franchises does not mean more reviews of the franchises and more reviews does not mean the franchises is popular.

However, each franchises must want to earn more profit, therefore they need feedback to improve themselves. In the future part, I will analysis the distribution of these two franchises and the relationship between Starbucks and Tim Hortons to show the feedback of customers on these most two prevalent franchises in Great Toronto Area.

Does business location play an important role in reviews?

As we known, the rent at downtown is higher than the surrounding area. However, is it true that the business at central Toronto are more likely to receive more reviews than the business at suburbs which can make the business be known by more people and is helpful for the establishment to gain more profit? In this part I will do analysis on this question.

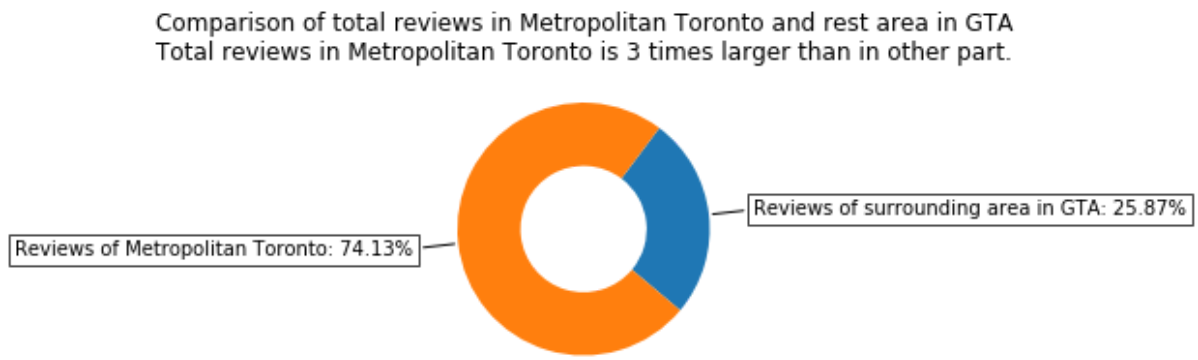
Relation between total reviews and location

First I will analysis the relationship between total reviews received by all business and location. I will sperate GTA into parts and find the result.

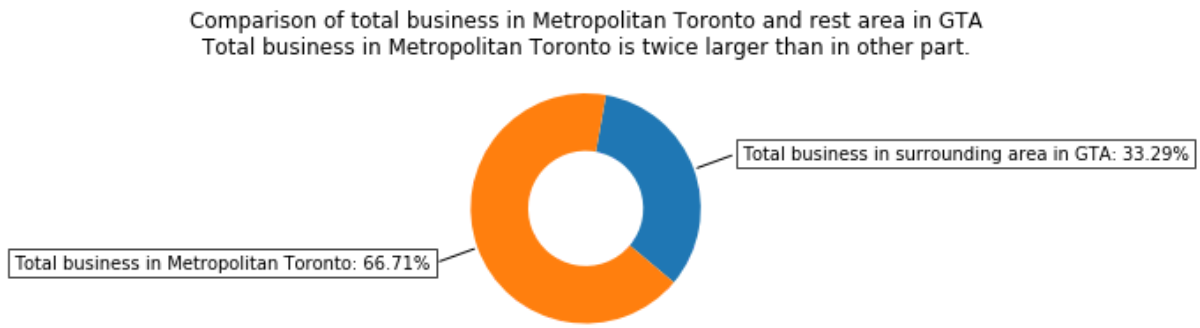
Method

There are two main parts in GTA: Metropolitan Toronto and surrounding area of GTA. In order to seperate these two areas based on Forward Sortation Area (https://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br03396.html), business with postal codes start with M are the business located at the Metropolitan Toronto while postal codes start with L are in surrounding area of GTA.

First, total number of reviews in these two areas are calculated respectively and it is visualized in a donut pie chart as following:




```
<Figure size 864x864 with 0 Axes>
```




```
<Figure size 864x864 with 0 Axes>
```

Conclusion and Discussion

Most reviews are given for establishments locate in Metropolitan Toronto which is almost 3 times larger than the number of reviews given to business in the surrounding area in GTA. The initial conclusion can be gained that business in Metropolitan Toronto seems to have more reviews.

However, the main limitation of this method is that the total number of business in Metropolitan Toronto and surrounding area of GTA is different, which may have negative influence on the result. In the second pie chart above, business in Metropolitan Toronto is twice larger than business in surrounding area of GTA.

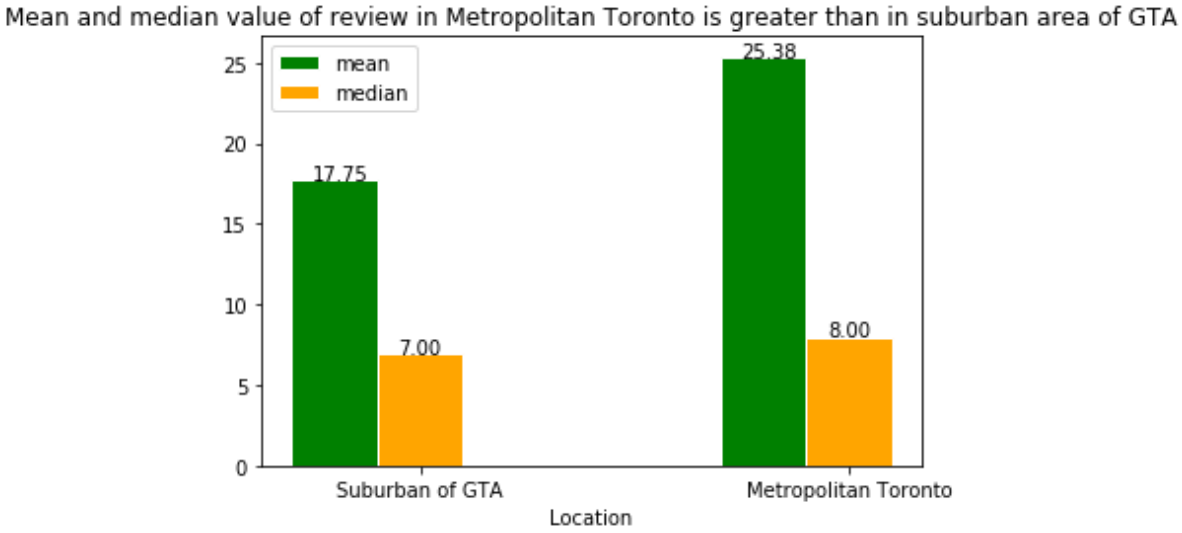
Therefore, the mean and median of reviews in each area will be analysised which will provide a more accurate result.

Relation between mean/median reviews and business location

Although we know business in central Toronto area receive more reviews than those in surrounding area in GTA, however, the total number of business in central Toronto is greater than the number of business in suburbs. We cannot say to much about the result we get above, therefore I will analysis the mean and meadian reviews per business receive regarding to the location.

Method

The median and mean value of reviews in Metropolitan Toronto and surrounding area are calculated seperately. Then the results are visalized in the bar plot after which the result can be concluded.



Conclusion and Discussion

Based on the visualization above, the conclusion is **business in Metropolitan Toronto are more likely to gain reviews.**

In the last part, total reviews for business in Metropolitan Toronto gained much more reviews than those in suburban of GTA and in this part.

In this part, each business in Metropolitan Toronto can get 25.38 reviews while the expectation of reviews of business in suburban is 17.75. The median for business in central Toronto is also greater than those in the surrounding. Therefore, location is important to obtain more reviews. In this way, if a business is eager to have better future, it had better develop in central Toronto but is not just limited in the suburban of GTA.

In addition, we have known there is a positive relationship between count of reviews and the star rating for restaurants type business in the North America from the previous research. Therefore, it might be more important for restaurants to select a better location although the rent is higher in central Toronto.

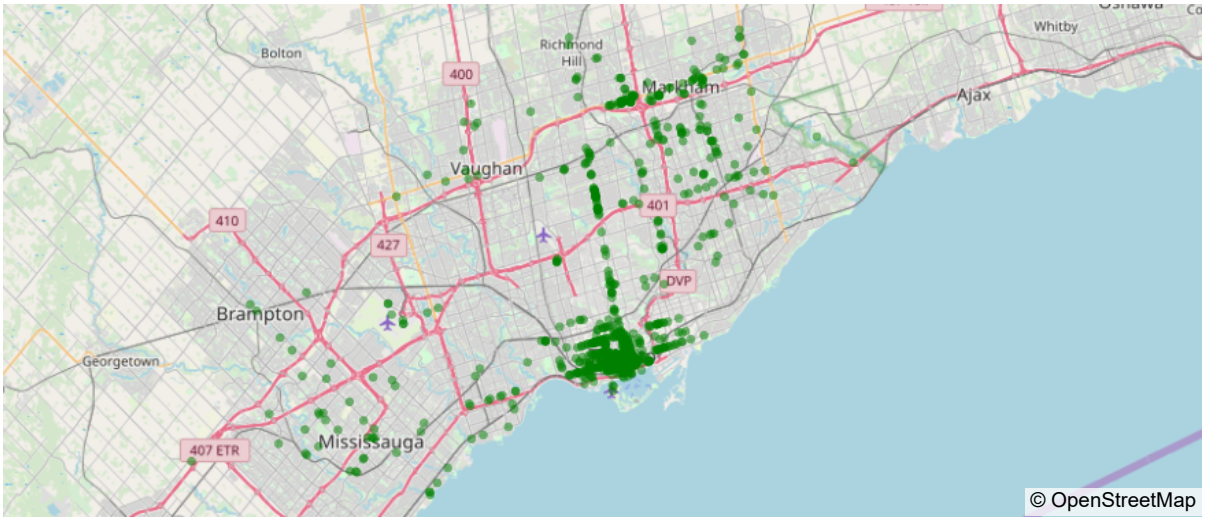
How to select location for a newly opened business?

It is interesting to find that the business locate in central Toronto tend to gain more reviews which may indicate more profit. However, the cost of a newly opened business locates in the central Toronto is also extremely high. Therefore, it is interesting to find whether there any other place can receive reviews but not in downtown Toronto.

Method

1. Find top 1000 business in GTA sorted by number of reviews received.
2. Visualize the distribution of those business on the map.

Out[39]:



Conclusion and Discussion

According to the distribution of top 1000 business on the map, there are some concentrations areas: downtown Toronto, Youge street and Markham. In addtion, some area with lower density can be find in Missisauga and Vaughan. Therefore we can conclude that a newly opened restaurant can locate on Yonge street or Markham rather than just in downtown Toronto. In addition, Vaugahn and Mississauge are also good choice.

The limitation of this method is the total number of business at central Toronto is larger than the number of business in other areas. This plot can only show most of top 1000 business locate at central Toronto but it does not indicate that the establishment at central Toronto are easier to get a higher star rating.

In addition, business owners may more concerned about at which location is more likely to gain more profit. However, stars and count of reviews cannot represents profit. More central area the business is, the higher cost it means. Therefore the data we have currently is not enough to find the direct relationship between profit and business locations which need to be done further research with more valid data.

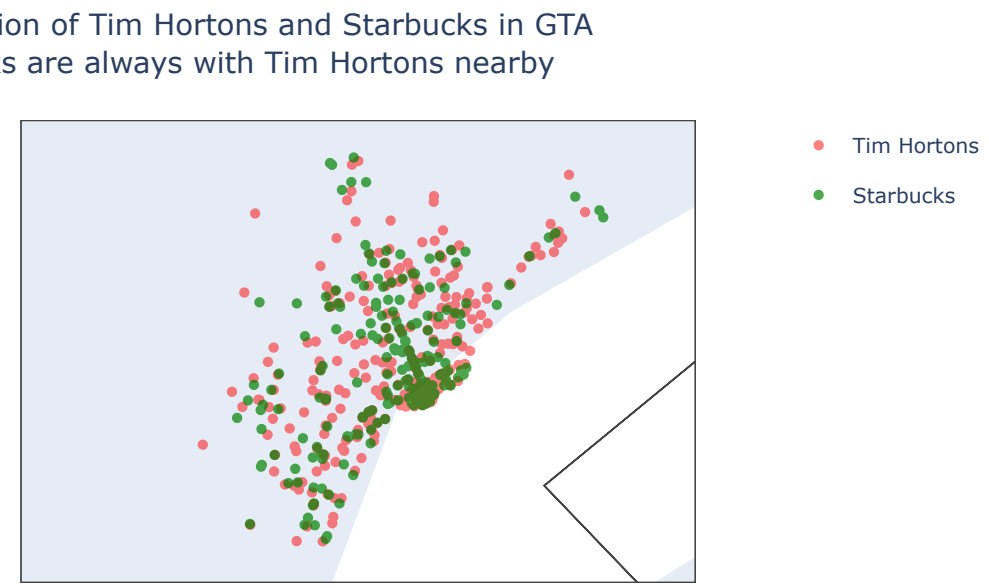
Is it true that for every Tim Hortons in the GTA there is a Starbucks nearby?

Tim Hortons is one of the most popular franchise in Canada while Starbucks is the most popular coffee shop in the world. It will be interesting to find is that true for every Tim Hortons in the GTA there is a Starbucks nearby which is similar to that we can always find a Burgerking near a Mcdonald. In this way, we may know more about which one is more prelavent in GTA and the compatitive pattern between these two prestigious franchises.

Method

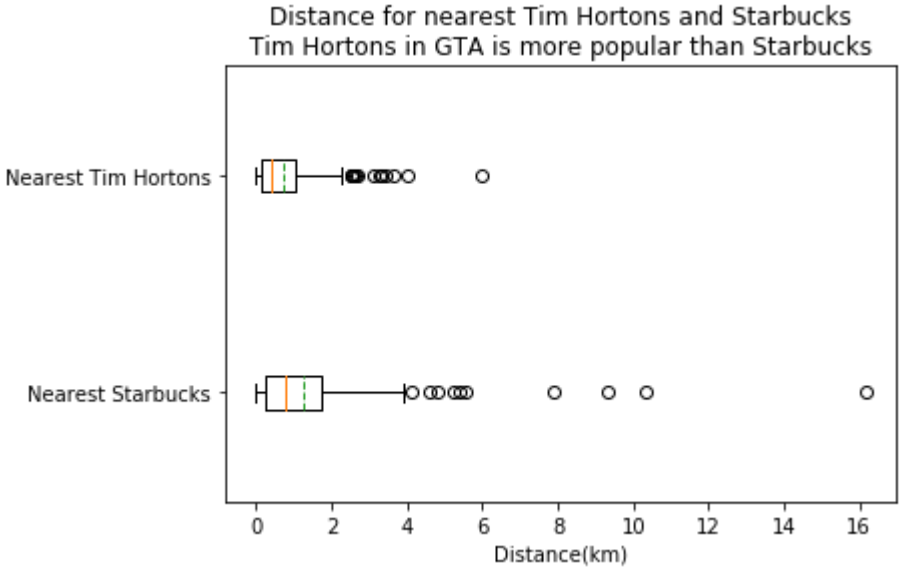
1. Visualize the plot Tim Hortons and Starbucks on GTA map directly, which provides a overview result.
2. For each Tim Hortons, find the nearest Starbucks and record the distance. For each Starbucks, find the nearest Tim Hortons and recode the distance.
3. Visualize the result in boxplot and compare the distances.

Out[42]:



According to the visualization on the GTA map, it is not hard to observe the number of red points(Tim Hortons) is almost same to the number of green ponts(Starbucks), which indicates the Tim Hortons and Starbucks are same prevalent in GTA.

However, some Tim Hortons locate at the edge of the GTA, Starbucks cannot be found nearby. Therefore Tim Hortons is a little bit more common in the edge part of GTA, indicating that Tim Hortons is the more popular one in the edge of GTA than Starbucks.



Conclusion and Discussion

The green line represents the mean of all distance and the orange line indicate the median of the distance set. According to the boxplt above, both median and mean of nearest Starbucks for Tim Hortons is greater than the the nearest Tim Hortons for Starbucks.

This implies that there may be Tim Hortons with Starbucks far away while for all Starbucks are not hard to find a Tim Hortons nearby. Thus, this result indicates the idea from the map is true and shows that a person is easier to find Tim Hortons than Starbucks in GTA which also indicates Starbucks should open more branches if Starbucks want to compete to Tim Hortons.

However, we do not know if higher density of Tim Hortons represents a more positive reviews it can gain. Therefore, I will find which one seems to be easier to get a positive reviews. This will be helpful to find the advantages and disadvantages for each of these two franchise so that they can provide better service in the future.

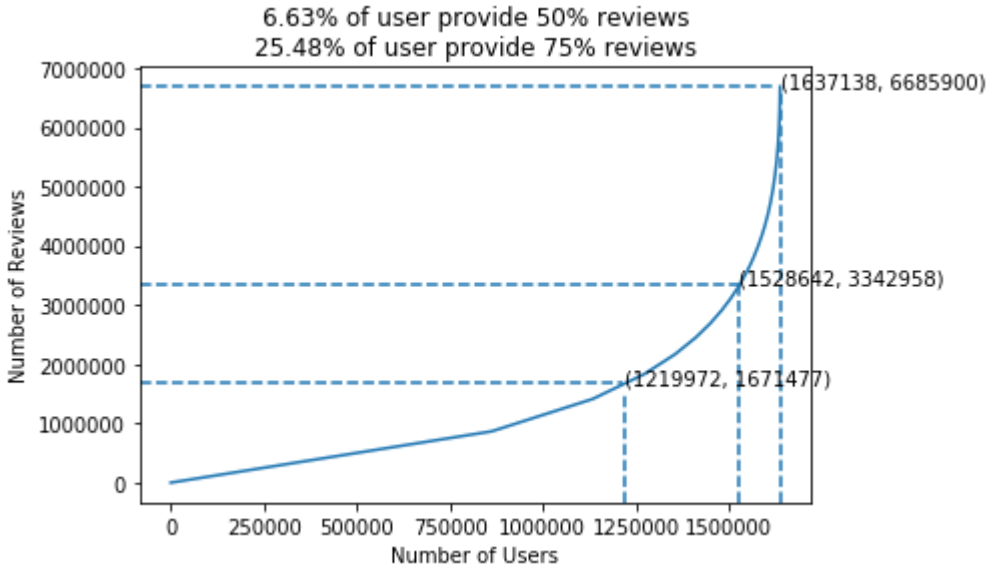
Reviews

Is there a small group of users responsible for most reviews?

Someone says that there only a small group of elite make the progress for the human. We can also find a lot of same things in our real lives. As for Yelp reviews, is that mean there is only a small group users reponsible for most reviews? In this part I will find the distribution of users make reviews.

Methods

1. Calculate how many reviews given for each user.
2. Visualize the relation between total number of user and total number of reviews.



Explanation

X-axis represents the total number of users while y-axis represents the cumulative total number of reviews. For example, the points labeled as (1219972, 1671477) indicates that first 1219972 users provided 1671477 reviews.

Conclusion and Discussion

The plot implies 6.63% of users provide 50% reviews while 25.48% provide 75% of total reviews. According to this results, **there is a small group of users responsible for most reviews.**

However, it is necessary to be aware of the reliability of the reviews of those users because some of them may be paid user which means those used are paid for giving good comments for some business.

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

Starbucks is one of the most famous coffee franchise in the US while Tim Horton's is the most popular franchise in Canada. All of these two franchises can be considered as coffee shops. Therefore, it is interesting to investigate whether customers for both Tim Hortons and Starbucks use similar language in their reviews.

Second Cup is also a relavent coffee shop in Canada. In order to analysis the language in reviews of GTA's Tim Horton's and Starbucks, the second cup is used to be compared group. Because its style is more similar to Starbucks which are both coffee shops while second cup and Tim Horton's are all Canadian business. Thus second cup will be an effective compared group.

To find the language using in reviews, I will analysis it in two parts:

- Top noun phrases describing each of the franchise.
- Reviews sentiment score of each brand.

Top noun phrases decribe Tim Hortons, Starbucks and Second Cup

Method

1. Clean the data. For each review, I extract all noun phrase in the review text and collect them.
2. In Tim Horton's data, the tim hortons can be represented in multiple ways in different reviews such as Tim Hortons, Tim Horton's, Timmies, etc. Then I uniform the name and sum all these words to Tim Hortons.
3. Find the top frquences of noun phrases for each franchise.
4. Visualize the result in word cloud.

After extract all noun phrases for each franchise, we find that there are 10706 different noun phrases to describe Starbucks which is the most while there are only 4019 different phrases to label the second cup which is also the least one. 6741 noun phrases are used to describe Tim Hortons. Then I will show the top word to describe those franchise.

Noun Phrases Describing Tim Hortons

Noun Phrases Describing Starbucks

Noun Phrases Describing Second Cup

In order to gain a better visualization, Second Cup is put at the bottom of Tim Hortons and Starbucks which is convenient to compare each of those two word clouds.

Conclusion and Discussion

It is interesting to find that the most popular phrase to describe Starbucks and Second Cup are all starbucks which indicates that the customers in Second Cup would like to use Second Cup to compare with Starbucks. The most popular phrase for Tim Hortons reviews is Tim Horton's. However we still can find Starbucks in Tim Hortons reviews. Then I suppose that Starbucks is still the benchmark of coffee shop industry because even if people in other shops, they still like to use Starbucks as the standard level.

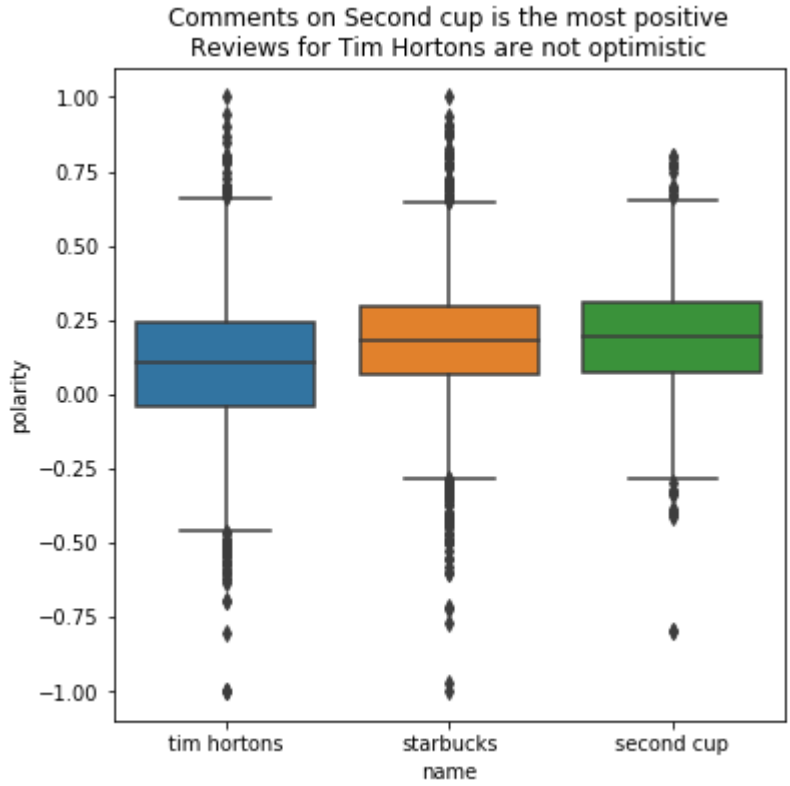
In addition, the other popular phrase to describe these two business is coffee shop, which implies the property of these two business while it is hard to find coffee shop phrase in Tim Hortons reviews. Though we can find it, notice that the word is really tiny. However, we can find donuts, food court, breakfast sandwich to describe Tim Hortons. Therefore Tim Hortons is more likely to restaurant while second cup and starbucks are literally coffee shops.

Besides, we can find noun phrases such as long line, wrong order and horrible to describe Tim Hortons while phrases such as nice, friendly stuff, friendly service can be found to describe Starbucks which indicates Starbucks provide a better atmosphere for customers while Tim Hortons need to make more improvement.

Sentiment of reviews for Tim Hortons, Starbucks and Second Cup

Method

1. Find the sentiment score of each review text according to each franchise.
2. Visualize the distribution of polarity score of each business in boxplot.



Based on the boxplot above, the observation that the second cup's polarity score is the highest which indicates the people tend to give second cup positive reviews. On the other hand, the meadian of Tim Hortons is the lowest. This shows that Tim Hortons need to do more in order to catch up with Starbucks and Second Cup.

Conclusion and Discussion

Based on the result we get, we can conclude that people would like to use similar word to describe those three franchise. We can see customers often compare those three most popular coffee shops to each other. In addition, people always give feedback on the service, environment, flavor and location aspects which share a large number of similar words.

Nevertheless, we also can find which franchise have a better quality regarding to the polarity analysis. We can notice that the polarity of starbucks and second cup are higher than the polarity of Tim Hortons obviously which might indicate Tim Hortons need to make more progress.

The visualization of polarity distribution indicates that the number of comments received by Second Cup is less than Tim Hortons and Starbucks, then this may cause an error when we doing the analysis. In addition, when we find the noun phrases, since I used API of TextBlob directly, I do not know the algorithm of this NLP tool. Therefore, when it finds noun phrases, it might make some mistakes such as it may consider second cup as cup. Then the error might occor. Therefore, when we analysis the reuslt, we had better consider what kind of aspects the comments of customers relating to. Then in each aspects, we can find some similar words.

Can we automatically detect "paid reviewers" (i.e. people who are paid to write positive reviews)?

There are huge number of paid reviewers in real business industry. It is important to find those paid reviewers so that customers can have better experience. However, it is difficult to detect paid reviewers. The method is provided as following, however there exists a lot of limitaions and restrictions to the method which will be discussed after the method.

Method

1. For each business find the reviews with abnormal star which is defined as the outliers of the score distribution.
2. Find all the reviews that this reviewer given, calculate the number of the reviews with abnormal star that the user made, the more reviews represents higher probability the user is paid reviewer.
3. If the user only have several reviews given this can be set to less than 3 and every reviews are extremely vivid, then the user also can be considered as paid reviewer.

Reason

By common sense, no paid reviewers will want others know they are paid, then they will try their best to be normal. However, the star might be abnormal which can have impact the average star rating. In addition, paid reviewers will write fantastic essays to attract others. But paid reviewers are not likely to use their own account which means they may create other account which is used to write "fake reviews". Therefore, the method above indicates these aspects.

Conclusion and Discussion

There are lots of restriction to this method.

First it is hard to define what are paid reviewers even if we know the idea of paid reviewers are those people who are paid to write positive reviews. However there are multiple ways to pay, including paid directly, offering free dessert after having meal or offering coupons. Besides, there are multiple kinds of paid reviewers, some business may paid users for providing positive reviews to themselves while some other merchants might paid users to give negative feedback to their components.

Besides that, what paid reviewers' "fake reviews" can be also different. Some of them can provide an extremely high score reviews or extremely low scores to make the average star change. Or they can write a fantastic essay to the reviews but with a normal score even with photos inserted, then other customers will come to the business after reading the review. There also are some paid reviewers can make them seems like the person had been to the business although they have not.

In addition, the methods I come up with to detect paid reviewers cannot work well because I do not know whether the user is actually a paid reviewers, then I cannot measure effectiveness of the method. Even if I use some machine learning methods or supervised learning method, beacuse of lack of answers.

Not to say automatically detect "paid reviewers", I look into the reviews given by some users, I cannot say whether the user is a paid reviewer. What I can do is to give the idea that whether I think the user is a paid reviewer or to say the probability of the user is a paid reviewer, which is really subjective. In this way, some people may disagree on me.

On the other hand, the criterias to judge if a reviewer is a paid reviewer, there still exists the possibility that I consider the user is paid reviewer but actually the reviewer is not. In addition, the scoring criterias are different for from people to people. Some people will always give 5 stars and some of them may never give 1 star. There

are lots of people do not like to write reviews, therefore only some reviews can be found among all reviews provided by the user. In this way, those normal reviewers will have influence on the judgement of paid reviewers.

The method I given is based on my personal understanding of paid reviewers this may be not accurate or this is subjective. To be more rigorous, the method I given can only filter those "weird" reviewers but I cannot prove they are paid reviewers. Because paid reviewers can be normal and normal reviewers can be weird and we even cannot tell the proportion of these two types of people in the whole population.