Yelp Dataset Analysis

Overview

Yelp is a platform for users to rate and review businesses. It was founded in 2005 and has business listing from major cities in over 20 countries.

Businesses across categories can list themselves on Yelp. Users can add other users as friends or follow them as fans. Users who leave thoughtful reviews, upload awesome photos, send compliments and up-vote reviews can apply for elite status for a given year.

Objective

Our objective is to explore the data and try and find insights that might be useful for the platform or the businesses using it. We also want to achieve the goal of learning about different countries in regions based on their business type and rating data.

Data

The Yelp dataset consists of six tables:

1. Business
2. Reviews
3. Users
4. Checkin
5. Tip
6. Photo

After reviewing the documentation for the dataset, we decided to focus our analysis on two tables – business and user. The following tables describe the variables in these table along with the review table. We have to use the review table and it contains both the business_id and the user_id fields which help us link the business and user tables.

Table: business
Contains business data including location data, attributes and categories

Variable	Type	Description
Business_id	String	22-character unique string business id
Name	String	the business's name
Address	String	the city
State	String	2-character state code, if applicable
Postal code	String	the postal code
Latitude	Float	latitude
Longitude	Float	longitude
Stars	Float	star rating, rounded to half-stars
Review_count	Integer	number of reviews
Is_open	Integer	0 or 1 for closed or open respectively
Categories	String Array	An array of strings of business categories

Table: review
Contains full review text data including the user_id that wrote the review and the business_id that the review is written for.

Variable	Type	Description
Review_id	String	22-character unique review id
User_id	String	22-character unique user id, maps to user in user table
Business_id	String	22-character business id, maps to in business table
Stars	Integer	star rating
Date	String	date formatted YYYY-MM-DD
Text	String	the review itself
Useful	Integer	number of useful votes received
Funny	Integer	number of funny votes received
Cool	Integer	number of cool votes received

Table: user
User data including the user's friend mapping and all the metadata associated with the user.

Variable	Type	Description
User_id	String	22-character unique user
Name	String	the user's first name
Review_count	Integer	the number of reviews they've written
Yelping_since	String	When the user joined Yelp, formatted like YYYY-MM_DD
Friends	String Array	An array of the user's friends as user_ids
Useful	Integer	Number of useful votes sent by the user
Funny	Integer	Number of funny votes sent by the user
Cool	Integer	Number of cool votes sent by the user
Fans	Integer	Number of fans the user has
Elite	Intger Array	The years the user was elite
Average_stars	Float	Average rating of all reviews
Compliment_hot	Integer	Number of hot compliments received by the user
Compliment_more	Integer	Number of more compliments received by the user
Compliment_profile	Integer	Number of profile compliments received by the user
Compliment_cute	Integer	Number of cute compliments received by the user
Compliment_list	Integer	Number of list compliments received by the user
Compliment_note	Integer	Number of note compliments received by the user
Compliment_plain	Integer	Number of plain compliments received by the user
Compliment_cool	Integer	Number of cool compliments received by the user
Compliment_funny	Integer	Number of funny compliments received by the user
Compliment_writer	Integer	Number of writer compliments received by the user
Compliment_photos	Integer	Number of photo compliments received by the user

More details about this dataset can be found on Yelp's documentation page for their dataset.[1]

1 Source (Yelp Documentation)

Importing data into SAS

We downloaded the dataset from the Yelp website. According to Yelp, each file is composed of a single object type, one JSON-object per line. However, we were unable to read the data into SAS. We encountered the following error.

```
ERROR: Invalid JSON in input near line 2 column 2: Unexpected
characters found after valid JSON text.
```

From multiple posts on communities.sas.com we identified that this issue occurs when the JSON file contains a line-break at the end of each line. SAS cannot currently read files in this format.[2]

As a workaround, we found an upload by the Yelp team on Kaggle that contained this dataset in the CSV format.[3]

The size of the three CSV files:

File	Records	Size (GB)
Business	174,567	0.03
Review	~5,200,000	3.53
User	1,326,100	1.27

Due to the large size of the review and user tables, trying to import the review and user file causes SAS University Edition to crash. Breaking the CSV file into chunks[4] of 120 MB allowed us to import them in and confirm if there were specific variables which were responsible for the large file size.

User

The results of PROC CONTENTS showed us that the friends column was responsible for the large size of the user data. It contains the user id of all friends of each user and has a length of 6312 characters.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
11	average_stars	Char	6	$6.	$6.
19	compliment_cool	Char	3	$3.	$3.
15	compliment_cute	Char	3	$3.	$3.
20	compliment_funny	Char	3	$3.	$3.
12	compliment_hot	Char	3	$3.	$3.
16	compliment_list	Char	3	$3.	$3.
13	compliment_more	Char	3	$3.	$3.
17	compliment_note	Char	3	$3.	$3.
22	compliment_photos	Char	3	$3.	$3.
18	compliment_plain	Char	3	$3.	$3.
14	compliment_profile	Char	3	$3.	$3.
21	compliment_writer	Char	3	$3.	$3.
8	cool	Char	3	$3.	$3.
10	elite	Char	6	$6.	$6.
9	fans	Char	3	$3.	$3.
5	friends	Char	6312	$6312.	$6312.
7	funny	Char	3	$3.	$3.
2	name	Char	9	$9.	$9.
3	review_count	Char	4	$4.	$4.
6	useful	Char	4	$4.	$4.
1	user_id	Char	24	$24.	$24.
4	yelping_since	Char	12	$12.	$12.

2 Source (SAS Communities)
3 Source (Kaggle)
4 Split CSV

We didn't want to delete this variable as we felt that this column could potentially help us explain user behaviour. We decided to add a column that counted the number of friends (friendscount) from this column and then drop it. We did this for all the 12 files that contained our data and merged the resulting datasets.

Dropping the friends columns reduced the dataset size in SAS from 720 MB to 10 MB for each file.

Review

The large size of the review data is because of the text variable which contains the complete text of the reviews. Our primary objective is using the review table to join the business and user table. We decided to drop the column. This would have been time consuming in SAS as we had 22 files, each containing 120 MB of data.

We decided to perform this operation using the read.csv function in R. This reduced the CSV file size from to 3.5 GB to 536 MB and the final SAS file to 604 MB.

Data pre-processing

The list of variables and data types in the business, user and review table are presented below. Several of the numeric fields were imported as character. We converted them to back to numeric in a data step.

business

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
1	business_id	Char	22	$22.	$22.
11	categories	Char	95	$95.	$95.
3	city	Char	14	$14.	$14.
10	is_open	Num	8	BEST12.	BEST32.
6	latitude	Num	8	BEST12.	BEST32.
7	longitude	Num	8	BEST12.	BEST32.
2	name	Char	32	$32.	$32.
5	postal_code	Char	7	$7.	$7.
9	review_count	Num	8	BEST12.	BEST32.
8	stars	Num	8	BEST12.	BEST32.
4	state	Char	2	$2.	$2.

User

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
2	TotComp	Num	8		
5	average_stars	Num	8		
7	cool	Num	8		
11	elite_ever	Num	8		
10	fans	Num	8		
3	friendcount	Num	8		
8	funny	Num	8		
9	review_count	Num	8		
6	useful	Num	8		
1	user_id	Char	24	$24.	$24.
4	using_since	Num	8	YYMMDD10.	

review

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
3	business_id	Char	22	$22.	$22.
8	cool	Num	8	BEST12.	BEST32.
5	date	Num	8	DDMMYY10.	DDMMYY10.
7	funny	Num	8	BEST12.	BEST32.
1	review_id	Char	22	$22.	$22.
4	stars	Num	8	BEST12.	BEST32.
6	useful	Num	8	BEST12.	BEST32.
2	user_id	Char	22	$22.	$22.

All the variables are correctly formatted in the business and review table. The user table has multiple numeric fields that have been assigned character data type. We fixed this using the input statement in SAS.

Selecting Data to Analyze – Primary Analysis

Locations

The business table contains data about 174,567 businesses. Most of these cities are in 4 countries – the United States, Canada, Germany and the UK. The user table consists of users from these locations.

Country	Number of businesses	Regions / States
USA	128504	25
Canada	38420	3
Germany	3177	5
UK	4536	2

We consider comparing businesses and users across countries. However, we were not sure if these comparisons would be meaningful as the number of businesses listed, active users and engagement might be different between the four countries.

We used data from SimilarWeb to compare usage patterns and the popularity of the Yelp platform across these countries in Feb 2020.[5] [6]

Metric	USA	Canada	Germany	UK
Total Visits (million)	120.41	6.77	4.89	2.60
Visit per capita	0.36	0.18	0.04	0.03
Avg. Visit Duration (seconds)	190	129	90	80
Pages Per Visit	7.19	4.41	2.68	2.80

The user engagement and penetration of the platform is much lower in Germany and the UK compared to the USA and Canada. Based on this, we decided to limit our analysis to users and businesses in the USA and Canada.

The number of businesses for which we have information per state/province is presented below for all states with at least 500 businesses. We selected the top 7 for analysis.

Obs	state	No_of_business
1	AZ	52214
2	NV	33086
3	ON	30208
4	NC	12956
5	OH	12609
6	PA	10109
7	QC	8169
8	WI	4754
9	ED	3795
10	BW	3118
11	IL	1852
12	SC	684

5 US and Canada

6 Germany and UK

Business Type

The categories for a business are stored as semicolon separated values in the categories variable. Businesses have between 1 and 12 categories. The distribution of the number of categories is presented for the selected states is presented below. 70% of businesses have four or fewer categories.

The FREQ Procedure

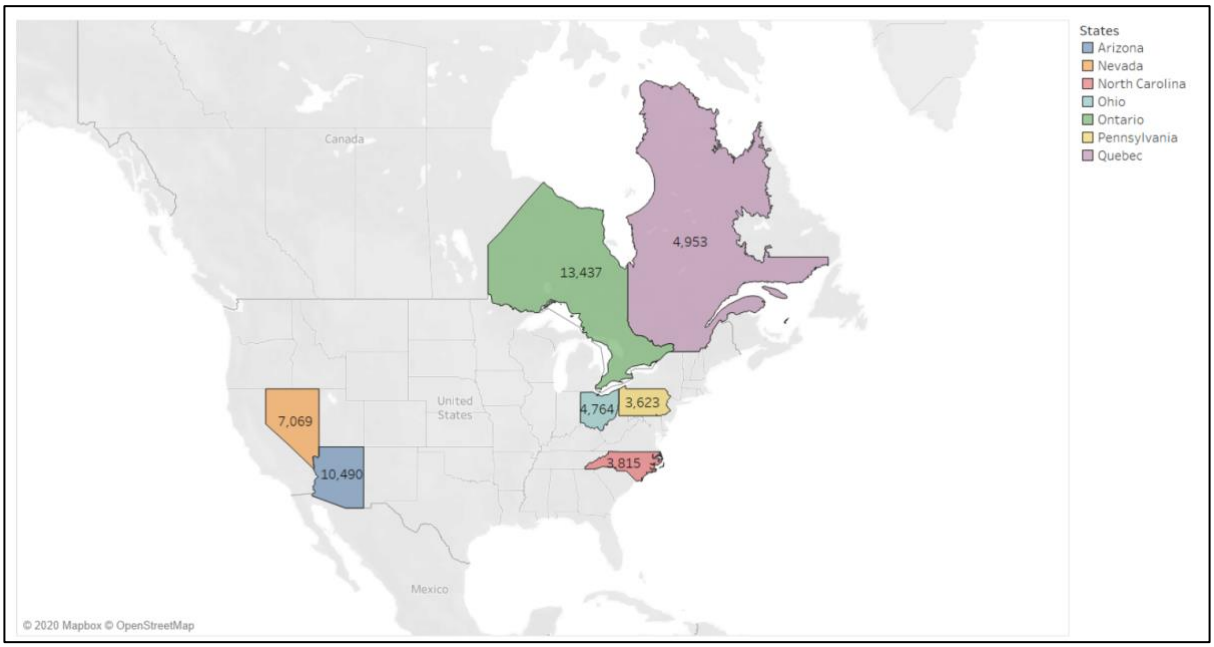
categorycount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1774	1.02	1774	1.02
2	51472	29.49	53246	30.50
3	37947	21.74	91193	52.24
4	31359	17.96	122552	70.20
5	23340	13.37	145892	83.57
6	15348	8.79	161240	92.37
7	8641	4.95	169881	97.32
8	3397	1.95	173278	99.26
9	1016	0.58	174294	99.84
10	232	0.13	174526	99.98
11	37	0.02	174563	100.00
12	4	0.00	174567	100.00

Frequency analysis confirmed that Restaurants is the most popular category. Based on this, we further decided to limit our analysis to businesses classified as restaurants.

Eliminating locations and non-restaurant businesses leaves us with 48159 restaurants across 7 states in the US and Canada.

Business Table: Exploratory Analysis

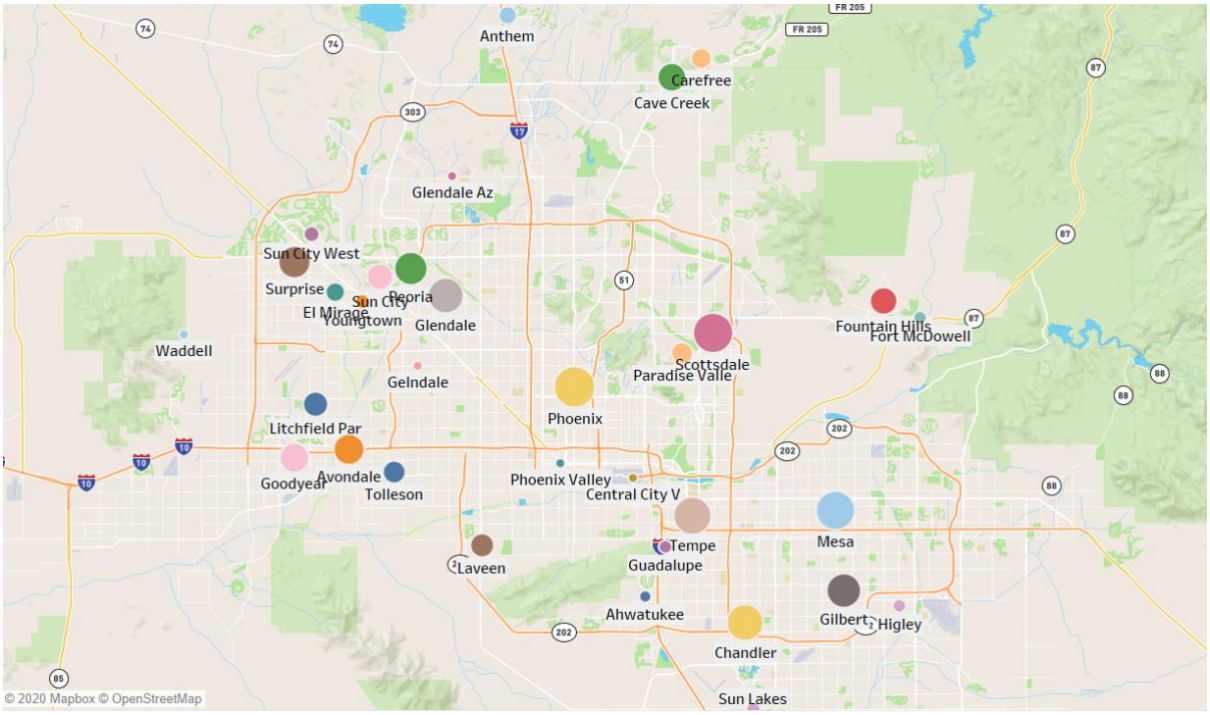
We started by visualizing the business distribution at a state and city and level. The plots for the same are presented below. These plots were generated using Tableau.



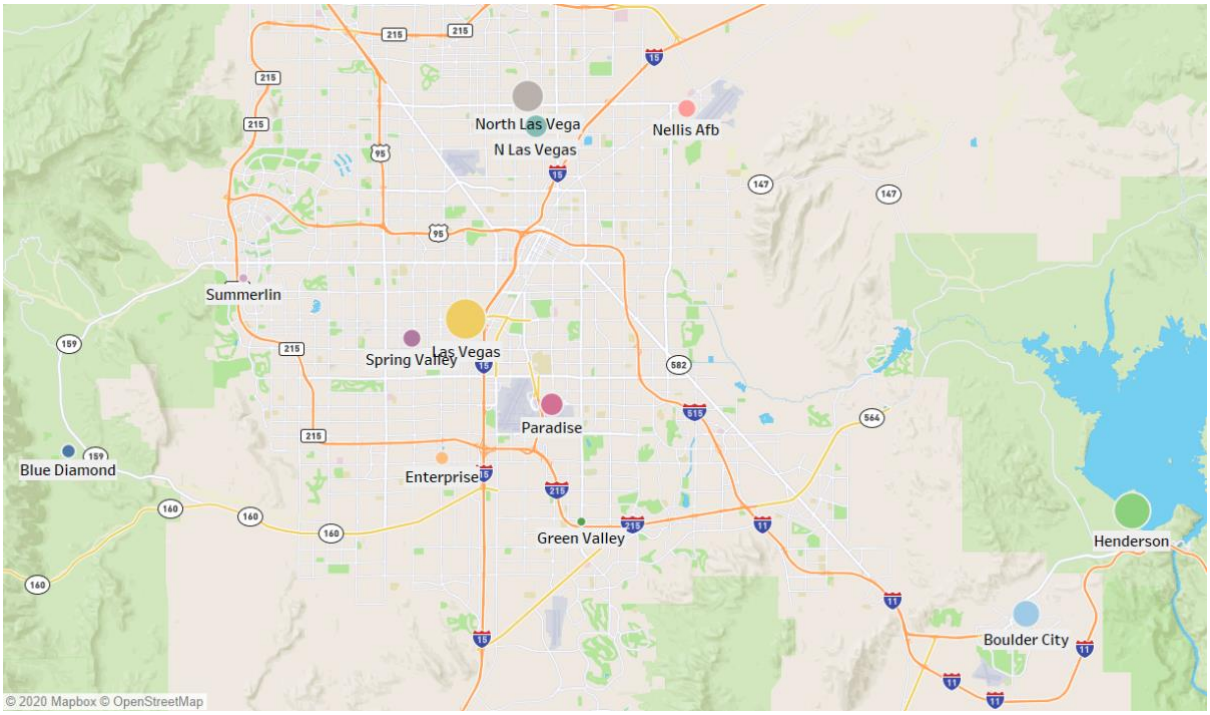
While plotting the city level data, we identified that there were errors in city names (missing letters). Additionally, there was some incorrect zip code data. For instance, in Nevada, the city for a business was mentioned as “North Las Vega”. Another business in Nevada was assigned a zip code 93013 which is actually a California zip code.

The distribution of cities within each state is as below. The US states have many more suburbs, and more restaurants in the suburbs compared to the Canadian provinces.

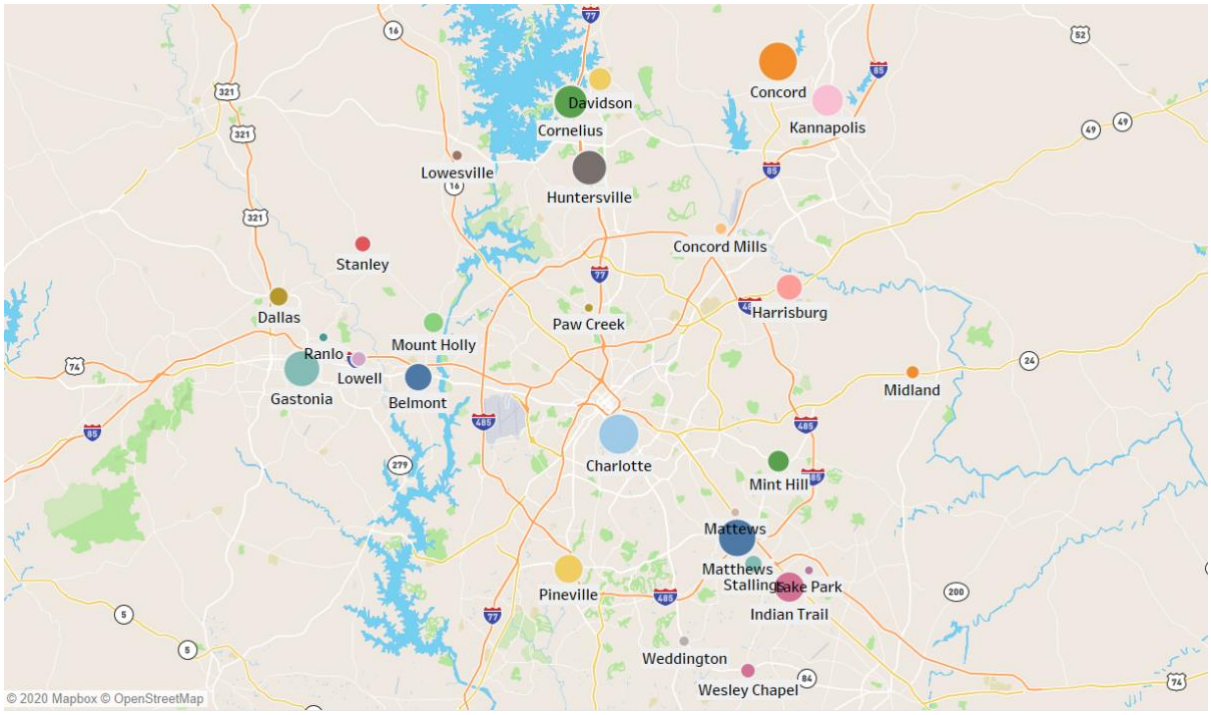
1. Arizona



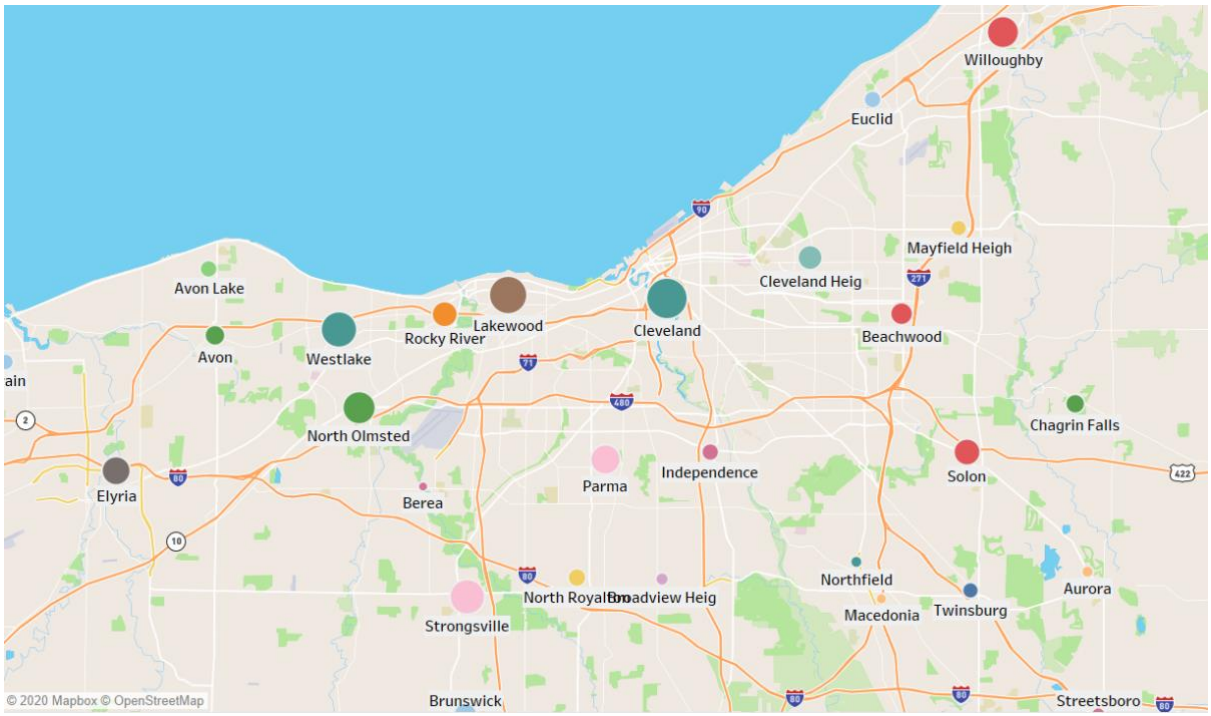
2. Nevada



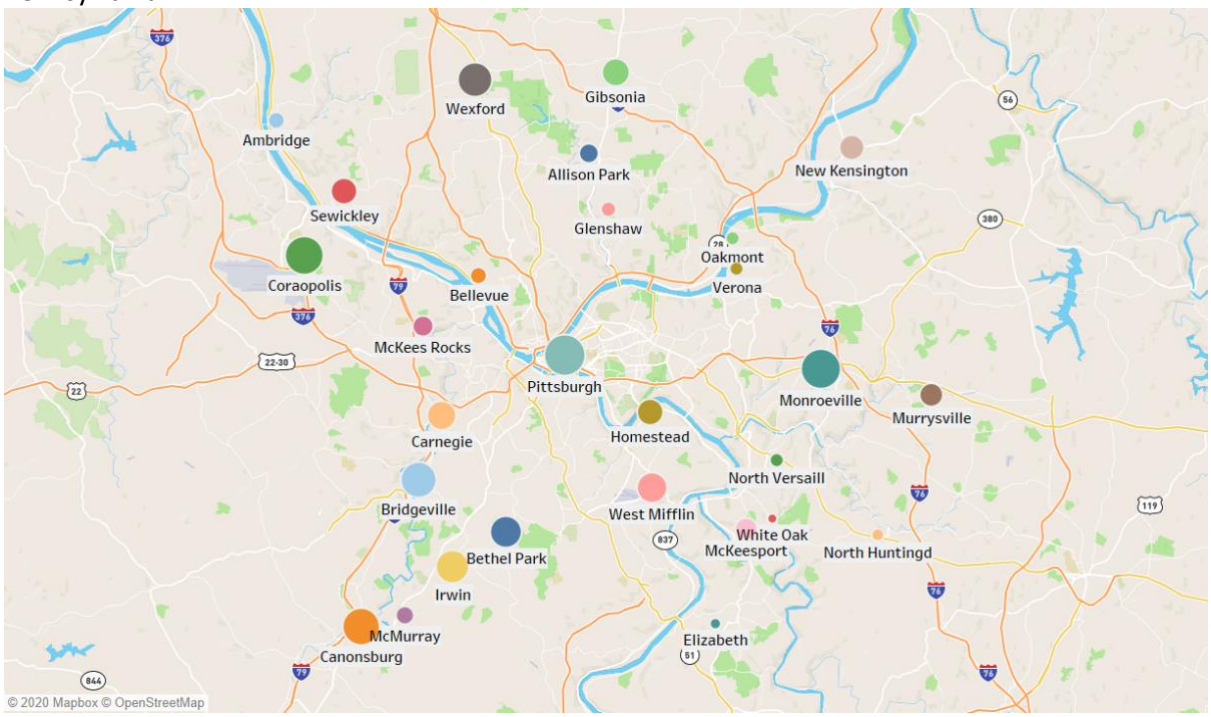
3. North Carolina



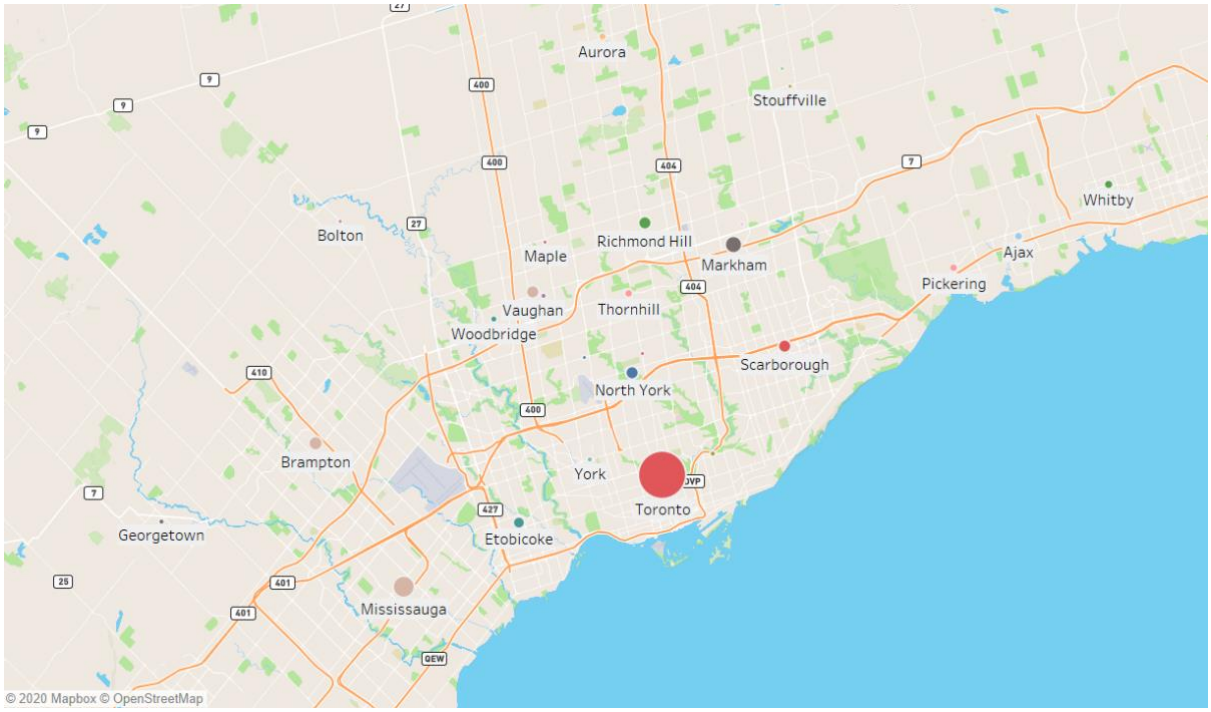
4. Ohio



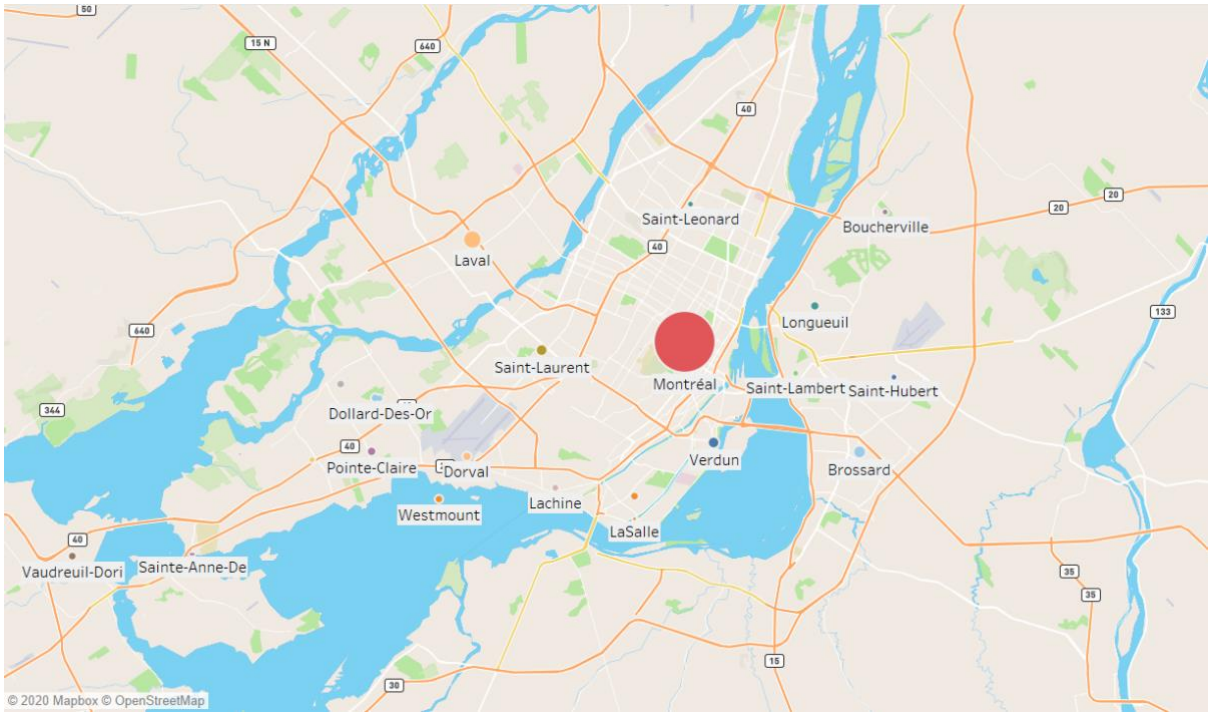
5. Pennsylvania



6. Ontario

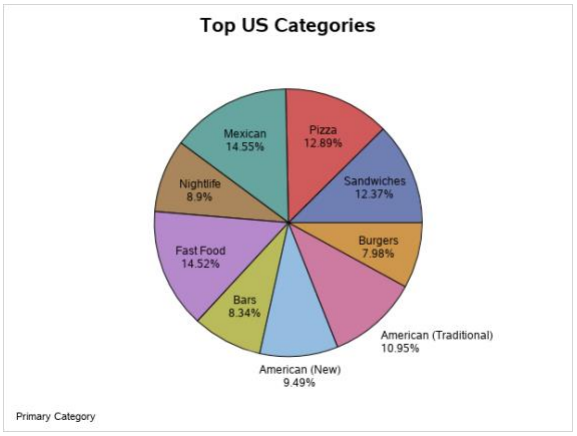


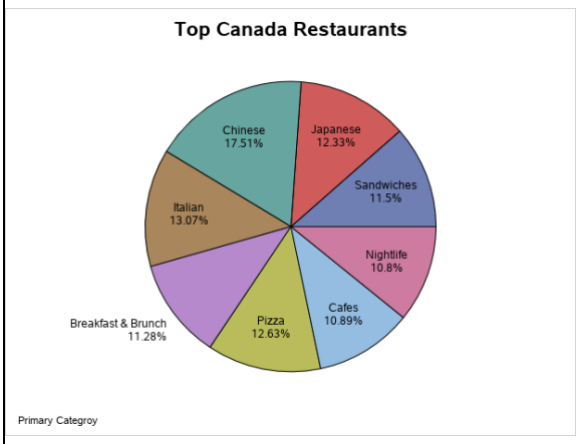
7. Quebec


Saint-Leonard
Boucherville
Laval
Longueuil
Saint-Laurent
Montréal
Saint-Lambert
Saint-Hubert
Dollard-Des-Or
Pointe-Claire
Dorval
Verdun
Brossard
Lachine
Westmount
LaSalle
Vaudreuil-Dori
Sainte-Anne-De
© 2020 Mapbox © OpenStreetMap

Popular Food Categories

The top categories for the US and Canada restaurants are presented below.





The top 3 food categories for each region in the US is as follows

State	Top 3 Categories	Percentage of restaurants
Arizona	American, Mexican, Pizza	46%
Nevada	American, Mexican, Pizza	41%
Pennsylvania	American, Pizza, Italian	53%
Ohio	American, Pizza, Mexican	48%
North Carolina	American, Pizza, Mexican	40%

The top 3 food categories for each Canadian province is as follows

State	Top 3 Categories	Percentage of restaurants
Ontario	Chinese, Italian, Pizza	24%
Quebec	Pizza, French, Italian	26%

- The top 3 categories account for at least 40% of the restaurants in the US. This number is around 25% for Canada. We combined American (Traditional) and American (New) for this analysis.
- Pizza is the only food category that is in the top 3 in all seven regions.
- We expected French food to be more dominant in Quebec. However, they actually have one more than restaurant classified as pizza (425) than French (424).
- *This analysis shows that the most popular categories make up a significantly smaller percentage of the total restaurants in Canada than in the US.*

Business Table: Hypothesis Analysis

Question

Is it possible to predict if a restaurant is open or closed based on data made available on Yelp?

Feature Selection

- We identified average rating as a variable that might help identify if businesses are open or closed. We expect them to be quite different for open and closed businesses.
- We created a new variable called bus_count which is a count of the number of businesses for each zip code. This variable acts as a measure for competition within the zip code.
- We decided to consider review_count in our model. However, we're not sure if it would be significant as newer restaurants will always have fewer reviews.
- The following table shows the difference in means for the variables considered.

Variable	Is_open = 0	Is_open = 1
Review_count	32	73
Stars	3.40	3.44
Bus_count	119	93

- All three variables show some difference in the mean values for open and closed restaurants.
- *There is hardly any difference in rating between open and closed restaurants.*
- This is somewhat counter-intuitive to what we believe about the power of Yelp.

Logistic Regression

- We performed binary logistic regression in SAS. The output is presented below.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0864	0.0463	551.4178	<.0001
stars	1	-0.0464	0.0131	12.4900	0.0004
review_count	1	0.00707	0.000208	1154.9253	<.0001
bus_count	1	-0.00225	0.000076	874.2899	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
stars	0.955	0.930	0.980
review_count	1.007	1.007	1.008
bus_count	0.998	0.998	0.998

- Logistic regression indicates that all three factors considered were statistically significant.
- However, the effect size for all variables is really small.
- Additionally, there is not much difference in the deviance between the null model and our model.
- This indicates that while the variables we've used our significant, they only explain a small proportion of the reason behind restaurant closers.

Limitations

- The two variable model has some predictive power and can act as a base for better analysis.
- Variables which capture how long a business has been open and the number of reviews by year would increase the accuracy of the model.
- Additionally, the analysis could potentially be more accurate if we measured competition by restaurant category and also built separate models for each region.

User Table

Feature Creation
We added a few additional variables to the user table for our analysis.

- elite_ever
  We created a variable called elite_ever using the elite variable which contains a string of years that a user has been elite. Elite_ever measures if a user was classified as elite in year. The purpose of this variable is to identify users who have at any point of time been more invested in the platform than regular users.

- City and State
  We wanted to identify the city and state for users to check if there are any differences in the other variables for users across different states. The table provided by Yelp did not provide any location information for the users.

  We joined the business and review table to the user data to add the city of the businesses for which users have left a review to the user table. However, there were several users who had reviewed businesses in multiple cities.

  We looked for a mode function that would enable us to select the most common city for each user but were unable to find any such function either in SAS or PROC SQL. We found some documentation for a workaround for this situation and implemented it using subqueries in PROC SQL.[7]

  Additionally, we created a table which maps cities to states and used that to add the State data to user.

Exploratory Data Analysis

Friends and Fans

The following table presents the average number of friends, fans and the ratio between the two. The data is arranged by descending number of users.

State	Users	Avg. Friends	Avg. Fans	Friends / Fans
Nevada	340772	45.2	2.5	18.1
Arizona	236675	28.0	0.8	33.7
Ontario	92367	20.4	1.0	20.6
North Carolina	58833	28.3	1.1	26.9
Ontario	52745	24.5	1.0	25.5
Pennsylvania	47464	27.2	1.2	22.7
Quebec	32669	28.2	2.1	13.7

- The average number of friends is between 20 and 29 for all states apart from Nevada where the number is 45.2.

7 Source (SAS Communities)

- The number of fans is between 13 and 33 times smaller across all regions. This makes sense as we're considering all users (elite and non-elite). We don't expect non-elite users to have a large number of fans.
- *For average friends and average fans there is more difference between the states within a country than between countries.*

Friends and Fans - Elite Users

The distribution of elite users across regions is presented below. We built this result by using the elite_ever variable which captures who have been classified as elite in at least one year.

State	Users	Elite_ever	Elite %
Nevada	340772	26877	8%
Arizona	236675	6986	3%
Ontario	92367	5384	6%
North Carolina	58833	3091	5%
Ohio	52745	2904	6%
Pennsylvania	47464	3018	6%
Quebec	32669	3118	10%

- The percentage of elite users by region varies from 3% in Arizona to 10% in Quebec.
- Similar to the friends and fans count distribution there is no clear difference between the USA and Canada.
- *The percentage of elite users is negatively correlated with the total number of users with the exception of Nevada.*

We decided to look at the distribution of friends and fans separately for elite users as they are a small percentage of total user base in each region.

State	Elite Users	Avg. Friends	Avg. Fans	Friends / Fans
Nevada	26877	172.5	21.9	7.9
Arizona	6986	132.3	15.3	8.7
Ontario	5384	101.5	12.4	8.2
Quebec	3118	114.7	16.3	7.0
North Carolina	3091	116.6	13.7	8.5
Pennsylvania	3018	114.1	13.2	8.6
Ohio	2904	103.0	11.8	8.8

- The ratio of friends to fans is more consistent for this group compared to the overall user group.
- The average number of friends and fans to be higher for this group as expected.
- *We thought that the friends to fans ratio would be lower for elite users than non-elite users. However, the opposite is true for all regions.*

Average Ratings and Rating Count

The average rating and average rating count by region are presented below.

State	Users	Avg. Review Count	Avg. Stars
Nevada	340772	39.4	3.8
Arizona	236675	17.8	3.7
Ontario	92367	20.0	3.5
North Carolina	58833	22.7	3.6
Ohio	52745	21.4	3.6
Pennsylvania	47464	25.6	3.7
Quebec	32669	36.4	3.7

- The average review count is highest for Nevada and Quebec (39.4 and 36.4 respectively). It's between 17 and 26 for the other regions.
- This seems to be related to the fact that Nevada and Quebec have the highest percentage of elite users.
- The average star rating is consistent across regions.

Average Ratings and Rating Count – Elite Users

The table for elite users is presented below.

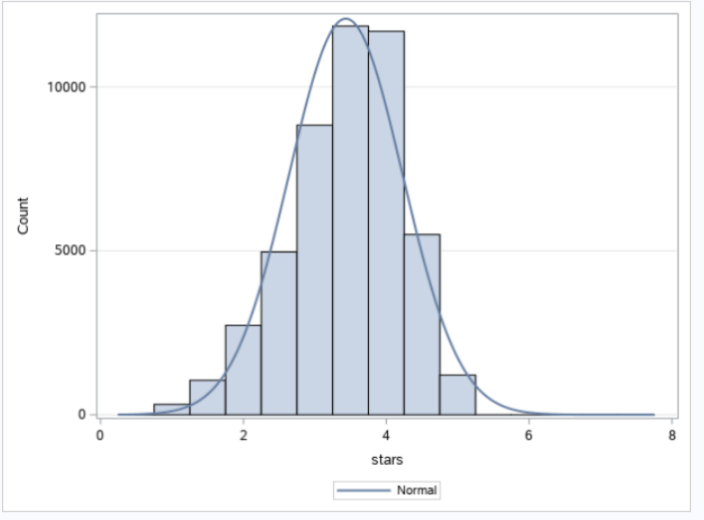
State	Elite Users	Avg. Review Count	Avg. Stars
Nevada	26877	249.1	3.8
Arizona	6986	194.9	3.9
Ontario	5384	171.9	3.8
Quebec	3118	214.6	3.8
North Carolina	3091	198.0	3.9
Pennsylvania	3018	195.1	3.9
Ohio	2904	176.8	3.9

- Our hypothesis that it was the higher proportion of elite users in Nevada and Quebec that is responsible for the higher average review count is incorrect.
- *Elite and non-elite users in Nevada and Quebec appear post significantly more reviews than users from other regions.*

Additional Analysis: Average Ratings - A look at the distribution

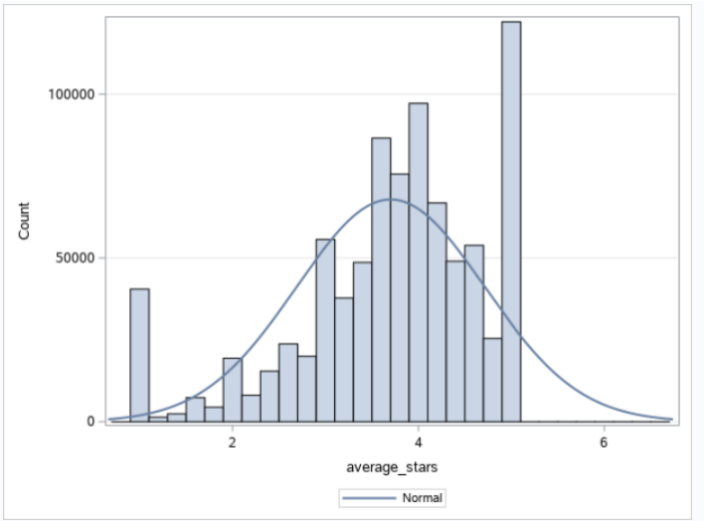
The distribution of average ratings for the businesses considered is presented below.

- The ratings are almost normally distributed.
- Most restaurants have a rating between 3.5 and 4 stars.



The distribution of average ratings given by the users is presented below.

- The most common average rating for users is 5. This is driven by a large number of users who have given only a few ratings.
- There are also a lot of users who have given an average star rating of 1.
- Outside of these extremes, the ratings distribution is reasonably normal.



Additional Analysis: Users who leave only one review

This table shows the distribution of reviews for users who have left only one review.

Average Rating	Percent of total
1	27%
2	6%
3	4%
4	8%
5	55%

- 5 stars is the most common rating as expected.
- *However, we were surprised to find that 1 star is the second most common rating given by users who have left only one review.*
- This percentage could be due to a combination of angry non-users who were only bothered to rate a restaurant and fake reviews.

Additional Analysis: Ratings for Restaurants vs Other Businesses

The table below shows the average rating for restaurants and other business that are open and closed.

Open	Restaurants	Other businesses
1	3.46	3.72
0	3.42	3.60

- Restaurants have a lower average rating than other businesses whether they are open or not.
- The difference between average rating between open and closed restaurants is much closer than for other businesses. There is only a .04 difference in star average of open and closed restaurants and a .12-star difference in open and closed non-restaurants.
- These differences could be evidence that restaurants have an inherent disadvantage in the nature of their business, resulting in lower star averages.
- On the other hand, rating don't mean as much to the open status of their business as it does to non-restaurants.

Takeaways

- There is more difference in user behaviour within states in the US and Canada than between the two countries.
- Less popular categories (outside the top 3) have a larger market share in Canada than in the USA.
- Elite and non-elite users in Nevada and Quebec are significantly more engaged than in other regions – they post more reviews and have a higher percentage of elite users. This can be investigated more deeply by Yelp to identify what's driving engagement in these regions and try and replicate the same elsewhere.

Limitations

- We were unable to predict whether a business is open or not accurately with the features available.
- We did not analyse the text of the reviews. That would have given us a better understanding of differences in user expectations in different regions.
- We did not plot our geo-visualizations in SAS University Edition.