Contents

List of Figures

List of Tables

1 Introduction

1.1 Brief Description of Yelp

Yelp is an American multinational corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations. The company also trains small businesses in how to respond to reviews, hosts social events for reviewers. For the context of this paper however, the value of *Yelp* is that it provides all sorts of data about businesses, from health inspection scores to what genre of business they conduct.

The *Yelp* platform operates in 219 cities across the world and has been collecting data in a sizeable capacity since 2006. Additionally, yelp operates on an ordinal review system, where consumers can leave stars with their written reviews. The star rank is assumed to indicate satisfaction, 1 being very dissatisfied and 5 being very satisfied. This qualifies Yelp as a potential measure of a business ability to satisfy their customers.

Also, as proven in the *Bell Labs* assessment of the Netflix rating system, the very act of choosing to leave stars is an indication of polarized interest, which means that counting the number of reviews left for a particular business is a feature which makes the Yelp Platform an accurate reflection of customer interest (Koren, 2009).

1.2 Motivation of Investigation

With this project I wanted to do something related to the distribution of small to medium business success in North America. But how do we quantify business success? That is why a measuring scale that is consistent and relevant, is both required and necessary. One could attempt an observation study; however I found the idea of sampling to be a rather tedious task for the short time that we had, and additionally the sample produced would be very unlikely to correctly represent the whole population.

This is why I resorted to secondary data. I initially thought to use Statistics Canada; however, three major problems arose: **First** it provides data from only 2017 and 2018, which would limit my ability to analyze the change over time, **second** it only provides statistics but not raw data in my field and **most importantly**, it only represents business in Canada but not the US. As a result it was not a proper choice for this study.

As I mentioned in 1.1, although the platform has limitations (Which Ill be discussing in the conclusion), it is a possible choice for estimating the desired distribution. This is because Yelp provides enough of the right kind of data to solve the three issues I mentioned with Statistics Canada.

1.3 Purpose

The purpose of this study is to interpret what Yelp can tell us (Regional differences, trends in interest etc.), and whether or not these statistics play an appropriate role as an **accurate reflection** of the distribution of business success in North America.

1.4 Ultimate Question

To reduce what I've said in this introduction into a single thesis we have the question: "What is the **Statistical Interpretation** of *Yelp* data from **2004 to 2017**. And do these statistics reflect the *actual* statistical characteristics of the **population of interest** accurately?"

The four sub-questions that will aid in the answering of the ultimate question are listed below:

1. How are business ratings changing over time?

2. How is user engagement changing over time?
3. Are the ratings of Business in Toronto consistent with the overall?
4. Do all Yelp regions hold the same level of interest?

2 Plan

2.1 Identification of Samples and Populations

For all the sub questions, the sample is the pool of businesses registered on the Yelp platform and the population is the pool of all registered retail/entertainment/restaurant businesses in North America. In sub-question 3 we also take a look at a smaller sample of all businesses registered on Yelp based in Toronto.

2.2 Identification of Variable Types

All variable types that appear in this investigation will be summarized in the following table.

Table 1: Variable Types

#	Dependent Var.	Independent Var.	Dependent V. Type	Independent V. Type
1	Mean Rating	Time by Year	Quantitative Continuous	Quantitative Discrete
2	Number of Reviews	Time by Year	Quantitative Discrete	Quantitative Discrete
3	N/A	N/A	N/A	N/A
4	Number of Reviews	Region by Major City	Quantitative Discrete	Quantitative Nominal

2.3 Evaluation of Secondary Data Reliability

As mentioned before the data mentioned in this study is indeed secondary, and drawn from one source only. That being Kaggle (*Kaggle Datasets*, n.d.).

This data can be retrieved from the following website:

https://www.kaggle.com/yelp-dataset/yelp-dataset/data

More information about the dataset can be found here:

https://www.yelp.com/dataset/challenge

For Yelp (*Yelp Dataset Challenge*, n.d.), the data is entered into the platform by millions of reviewers every year, although this cant be regulated using traditional means, the data is handled by some of the worlds most capable data scientists. And although this does not explicitly mean that the data is reliable, the shear volume of the data coupled with the corroboration of facts by millions of parties (*crowd-sourcing*) means that the dataset can be effectively regarded as highly reliable.

However no Data is without some implicit bias, which will be discussed now.

2.4 Evaluation of Bias

2.4.1 Volunteer Bias

Volunteer bias will occur because the act of leaving a review on Yelp is volunteer-based. Thus the very act of leaving a review is an indication of opinion. In my experience this means leaving higher

ratings on average. Therefore the rating distribution would be more left-skewed than the population. (More on this in sub-question 3)

2.4.2 Response Bias

For similar reasons to the aforementioned, certain reviewers may be more vocal in their online presence to others. This can be because of a variety of factors; however I hypothesize that it could be a result of regional difference in reviewing habits and patterns. This could potentially prevent Yelp from being an accurate reflection of the distribution of interest across space. (More on this in sub-question 4)

3 Analysis

All data listed come from Yelp and Kaggle (note that some tables only display the first few rows for brevity):

Table 2: Mean Yelp Ratings, 2004-2017

Year	Mean Score
2004	4.269230769
2005	4.002807801
2006	3.828529017
2007	3.842335567
2008	3.731093123
2009	3.62931951
2010	3.629479401
2011	3.597490852
2012	3.573333251
2013	3.571030618
2014	3.568498149
2015	3.583754714
2016	3.60635927
2017	3.610622227

Table 3: Total Populations in 2017 by City (Top 14)

Location	Las Vegas	Phoenix	Toronto	Scottsdale	Charlotte	Henderson	Pittsburgh
Population	632912	1615000”	5928040	246645	842051	292969	303625
Location	Mesa	Tempe	Chandler	Montreal	Gilbert	Glendale	Cleveland
Population	484587	182498	247477	1750000	237133	200831	385809

Table 4: Number of Reviews written in 2017 by Cities (Top 14)

Location	Las Vegas	Phoenix	Toronto	Scottsdale	Charlotte	Henderson	Pittsburgh
# of Reviews	332286	122082	81752	63748	53228	40814	35920
Location	Mesa	Tempe	Chandler	Montreal	Gilbert	Glendale	Cleveland
# of Reviews	34164	32512	28483	27438	25780	18219	17378

Table 5: Toronto Score Distribution

Bin	Frequency	Cumulative Frequency	Relative Frequency
1.5	430	2.546%	2.546%
2	980	8.349%	5.803%
2.5	1707	18.456%	10.107%
3	2861	35.396%	16.940%
3.5	3769	57.712%	22.316%
4	3617	79.128%	21.416%
4.5	2088	91.492%	12.363%
5	1437	100.000%	8.508%

Table 6: North American Score Distribution

Bin	Observed Frequency	Cumulative Frequency	Relative Frequency
1	3788	2.170%	2.170%
1.5	4303	4.635%	2.465%
2	9320	9.974%	5.339%
2.5	16148	19.224%	9.250%
3	23141	32.481%	13.256%
3.5	32038	50.834%	18.353%
4	33491	70.020%	19.186%
4.5	24796	84.224%	14.205%
5	27539	100.000%	15.776%

3.1 Sub-question 1: How are business ratings changing over time?

3.1.1 Motivation

The motivation here is quite obvious. The rating system is what people come to Yelp for. It is their main consumer service and the primary reason why people choose to use Yelp in the first place.

For the context of this study we want to determine if a business's yelp rating is analogous to the level of success that a business has at doing what it does.

Because of this, its stability is a crucial parameter in determining the credibility of the Yelp platform, as a stable system that facilitates the *direct* comparison of reviews from year to year without the use of z-scores or weight functions (variable can be regarded as approximately normally distributed since the sample size is massive and **Central Limit Theorem** applies) (Rosenblatt, 1956).

For the context of this paper the stability of a rating system over time decides whether it can be an accurate measure of the distribution studied.

3.1.2 Analysis

The measure of centre, the **mean** and the measure of spread, the **interquartile range** (IQR), can be used to define the **degree of challenge** here. After these two measures are calculated, a **trend line analysis** from different years can be compared without additional calculations.

The means from 2004 to 2017 are listed in Table 2, the table is reproduced as below:

Table 7: Mean Yelp Ratings, 2004-2017

Year	2004	2005	2006	2007	2008	2009	2010
Mean	4.269230769	4.002807801	3.828529017	3.842335567	3.731093123	3.62931951	3.629479401
Year	2011	2012	2013	2014	2015	2016	2017
Mean	3.597490852	3.573333251	3.571030618	3.568498149	3.583754714	3.60635927	3.610622227

The scatter plot with trend line, and a residual plot of the linear regression model constructed are displayed in Figure 1 and Figure 2:

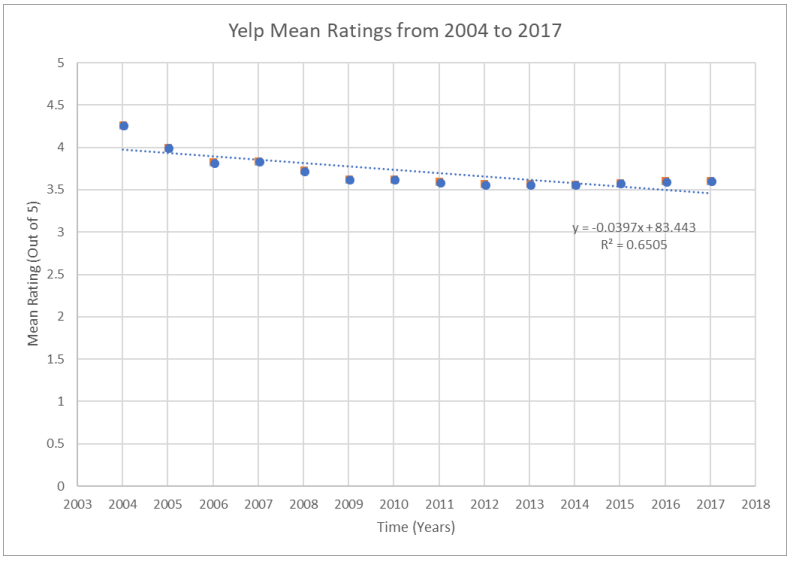

Figure 1: Yelp Mean Rating over 2004-2017

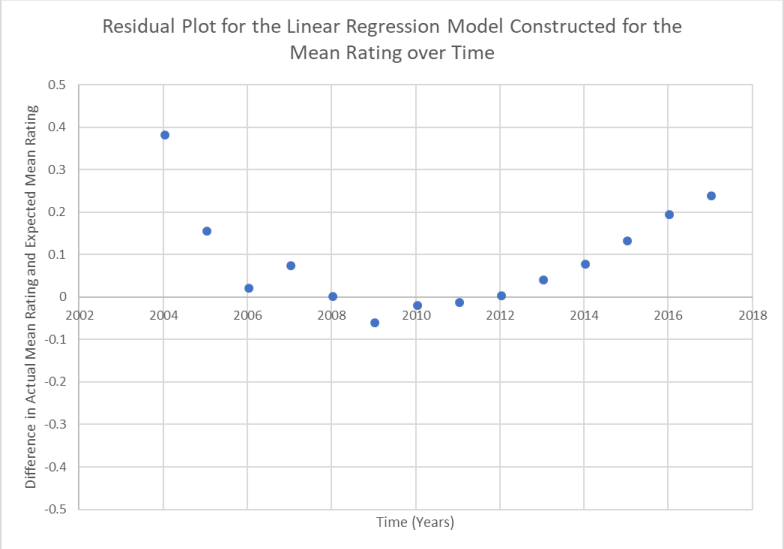

Figure 2: Residual Plot for Yelp Mean Rating over 2004-2017

Note the coefficient of determination, the R^2 value ($0.6505 < 0.8$ and $r = 0.8065$) indicates a weak correlation between the rating and time, and a negative slope demonstrates that the mean score is decreasing over the course of time. This implies that the difficulty to get a high score is increasing slowly (the difficulty is increasing since the mean rating is decreasing), but only 65.05% of its variance is accounted by the variance in time (be aware that although there seems to be an **association**, there is still room for *no* relationship to exist).

The residual plot presents a non-random pattern, which suggests the linear regression is not the most appropriate model for the data. Therefore, a **polynomial regression of degree 2** can be applied for higher accuracy and its graph is given by Figure 3. We can assume **degree 2** because of the *parabolic* nature of the residual curve. The new model gives a higher R^2 value ($0.9506 > 0.8$ and $r = 0.9750$), as expected. Possible outliers are data from 2004 to 2006 as the participation rate was much lower then, but it is unlikely.

From the analysis of centre before, we conclude that the overall rating difficulty of the Yelp platform is increasing over the years. This makes perfect sense because as competition arises, businesses are forced to adapt and improve. So a slight increase in difficulty to get higher ratings is expected.

Then, to study the spread of the data, we graph the IQR for Yelp ratings from 2004-2017 in Figure 4.

The IQR plot, which is not affected by outliers, implies a very stable spread around the centre, from which we conclude that this slight decrease in score that was found earlier has an *extremely small* effect on the overall rating distribution. Therefore, although the challenge required to get high ratings is increasing slightly over time, businesses' performance in different years can still be compared by calculating z-score. But since the shifting is so small, we can compare two ratings from year to year directly for the context of this paper.

3.1.3 Sub-conclusion

In general we preformed this analysis because regardless of our conclusions to the following sub-questions regarding this distribution, if the rating system is found to be unstable, then we must scrutinize or even go as far as to disregard those conclusions.

We find that the mean yelp rating over time is decreasing, this indicates and increase in the challenge for businesses to succeed. This makes perfect sense because as time goes on, competition arises, and businesses are forced to adapt and improve. So a slight increase in difficulty to get higher ratings is expected. Although the overall degree of challenge of getting a high Yelp score is *generally* increasing over time, it has a very small effect on the final ratings distribution. This stability means that *we can safely make direct comparisons from year to year in future analysis.*

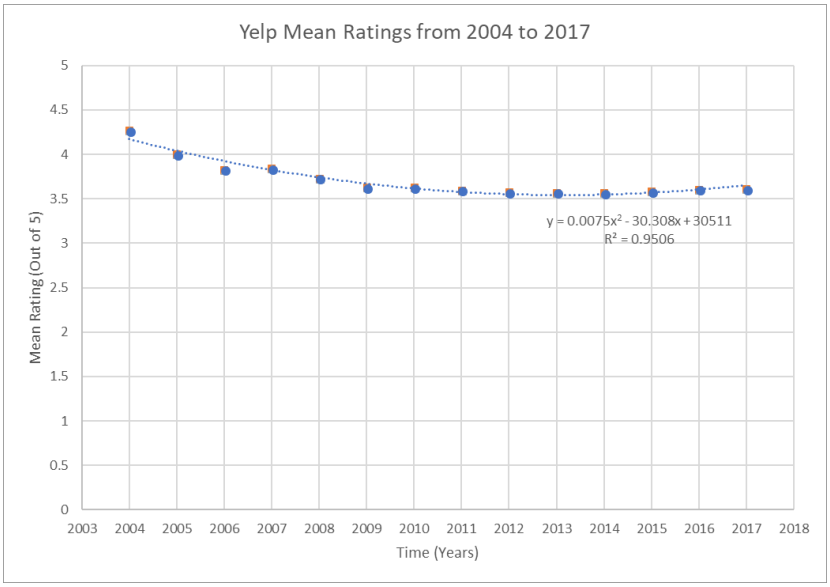

Figure 3: Polynomial Regression for Yelp Mean Rating over 2006-2017

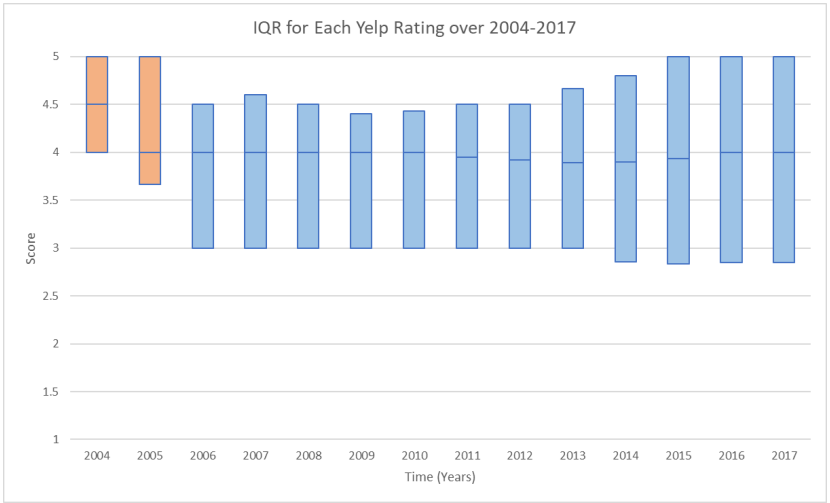

Figure 4: IQR for Yelp Mean Rating over 2004-2017

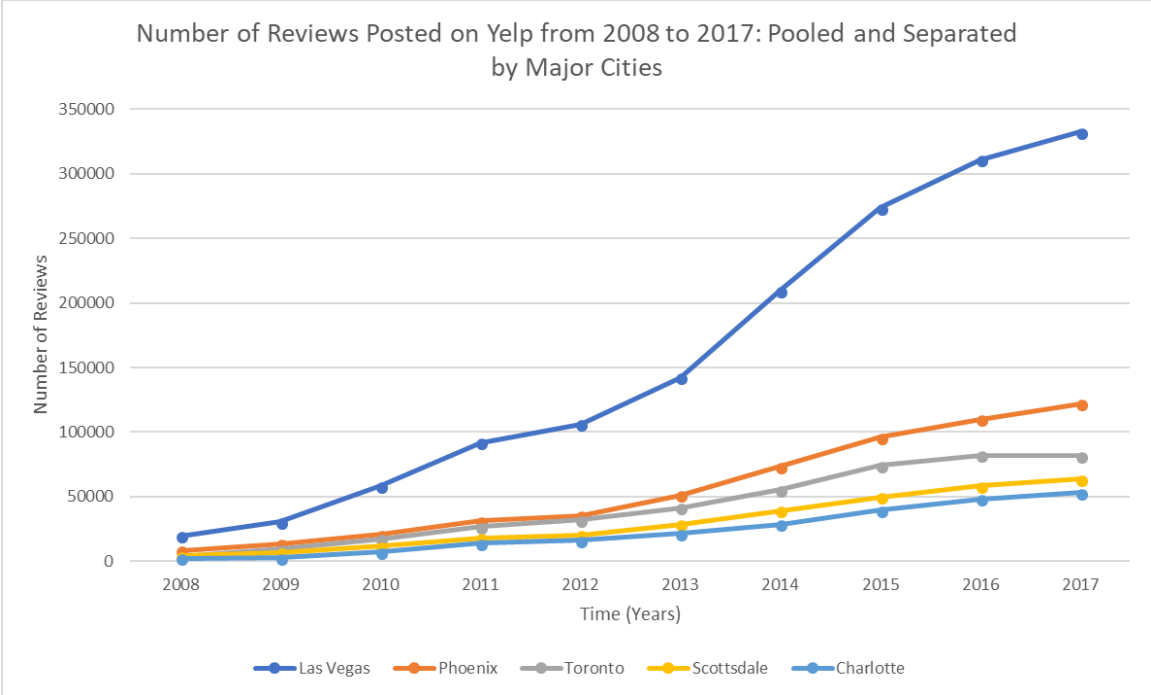

Figure 5: Engagement Counts by Major City over 2008-2017 (Top 5 Cities)

3.2 Sub-question 2: How is user engagement changing over time?

3.2.1 Motivation

The **engagement count** is exactly the **sample size** in this context. The larger the counts, the larger the sample size, and the more accurate our inference of the population from the sample will be. Furthermore, the participation count also measures the popularity among city folk that is generally affected by business success, accessibility, outcome, cost and so on. For example, for a review platform like Yelp we get a feedback loop in that, if people use the platform it is a good platform, therefore more people will use it and the cycle continues.The engagement count directly corresponds to the popularity of the service, knowing if people are continuing to use the platform after 10 years after its release will be a big indicator of its accuracy.

3.2.2 Analysis

Prior to all, we first graph the scatter plot of engagement counts in the top 5 major cities in which Yelp is used, and then connect each point. It can be observed that the number of reviews is fairly dependent on time, but is very likely subject to **confounding variables** (advertisements, social media, general buzz about the platform, number of participating businesses). Thus, trend line analysis is not appropriate and simple **direct observation with some explanation is required**.

From direct observation we notice that in general Las Vegas experiences several times the amount of user engagement than any of the other major cities; however this gets even more impressive when the scale is examined. After 2013 we see a noticeable increase in all major cities engagement counts. An explanation for this increase, is likely a result of the mobile boom hitting its peak in 2013 in which 90% of the world's data was being uploaded and consumed on mobile traffic. Additionally in 2013, Yelp received a large financial investment from the head venture capital partner. This influx

of cash resulted in a marketing campaign by Yelp and also likely contributes to this increase in user engagement.

From the Analysis above we could regard the participation counts to be steadily increasing, and claim that in most circumstances, the value of the Yelp platform is becoming more apparent as time goes on.

3.2.3 Sub-conclusion

Over time the Yelp platforms user engagement is steadily increasing suggesting that the platform has merit in its primary operative (produce business ratings) which suggests it is a credible source for distributional analysis. However; because of the discrepancy between Las Vegas and the other major cities, *the dataset could be subject to some regional bias.* Which will be explored in the next sub question.

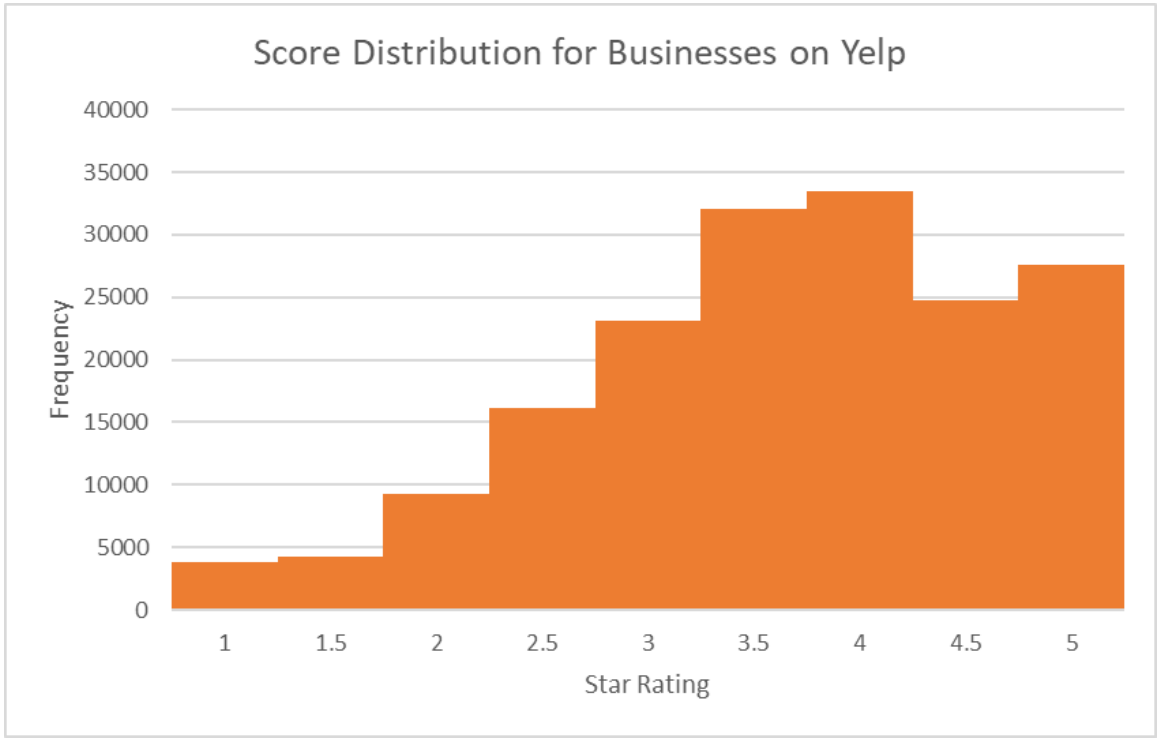

Figure 6: Ratings Distribution for 2017 Yelp

3.3 Are the Ratings of Businesses in Toronto Consistent with the Overall Ratings Across North America?

3.3.1 Motivation and Directions

To see whether Yelp can be an accurate reflection of overall business success in North America, it is helpful to see if consistency exists in the scores between any given region and the overall (only helpful not determining). If consistency exists, the given city is a **homogeneous subset** of the population and **stratified sampling** can be applied to estimate the statistics of the population and given a sample similar to Toronto, the platform is generalizable. However; if the consistency does not exist, then the given city is a **heterogeneous subset** of the population and thus, and

thus not an appropriate **stratified level** and given a sample similar to Toronto, the platform is not generalizable.

3.3.2 Analysis

First, a **normal approximation model** is constructed using Table 6 to describe the overall Yelp ratings in 2016 of North America. A normal distribution for the ratings distribution is appropriate in this case because the population is large, and a yelp rating is approximately a random variable. The actual rating distribution is graphed in a histogram, Figure 6. And the construction of the normal distribution is calculated in Table 7 and Table 8, and graphed in Figures 7-9.

$$X \sim N(\bar{X}, \sigma^2) \sim N(3.632189, 1.003738)$$

Both Figure 7 and 8 display the comparison as **continuous data**, while Figure 9 displays the comparison as **discrete data**. From the graphs, we could see the tail on the right is "fatter" than the tail on the left (since the rating distribution is under the model on the left and above the predicted on the right) and the centre (**sample mean**) is to the left of the **normal mean**. This gives a slight left-skewed distribution. In order to apply a two-tail-z-test to judge the consistency between Toronto's results with the overall performance in North America, it is necessary for the sample to satisfy the conditions of the significance test: the sample size is *greater* than 5 (16889 ¿ 5); the population distribution is *large*, and *approximately normal* (recall the data displayed in Table 6, which is graphed in Figure 10); all data are *independent* of each other (data scientists at Yelp are responsible for this). These conditions are verified by Figure 12.

Table 8: Normal Distribution Model Construction for Yelp 2017

Bin	z-score	Relative CDF [SNorm]	Predicted Cfreq.	Relative PDF [SNorm]	Predicted PFreq.
1	-2.622387341	0.437%	762.1126521	1.276%	2228.1338
1.5	-2.124249289	1.682%	2936.978527	4.163%	7267.443187
2	-1.626111237	5.196%	9070.868613	10.595%	18495.05093
2.5	-1.127973185	12.967%	22634.94964	21.038%	36725.0948
3	-0.629835133	26.440%	46154.93603	32.595%	56898.89432
3.5	-0.131697081	44.761%	78136.93196	39.402%	68782.54427
4	0.366440971	64.298%	112241.5053	37.165%	64876.31193
4.5	0.864579023	80.637%	140762.3094	27.351%	47744.96483
5	1.362717075	100.000%	159466.6846	15.705%	27415.90017

Table 9: Normal Distribution Model Compared with Actual Statistics for 2017 Yelp

Bin	Predicted Cfreq.	Actual Cfreq.	Predicted Pfreq.	Actual Pfreq.
1	762.1126521	3788	2228.1338	3788
1.5	2936.978527	8091	7267.443187	4303
2	9070.868613	17411	18495.05093	9320
2.5	22634.94964	33559	36725.0948	16148
3	46154.93603	56700	56898.89432	23141
3.5	78136.93196	88738	68782.54427	32038
4	112241.5053	122229	64876.31193	33491
4.5	140762.3094	147025	47744.96483	24796
5	159466.6846	160598.88	27415.90017	27539

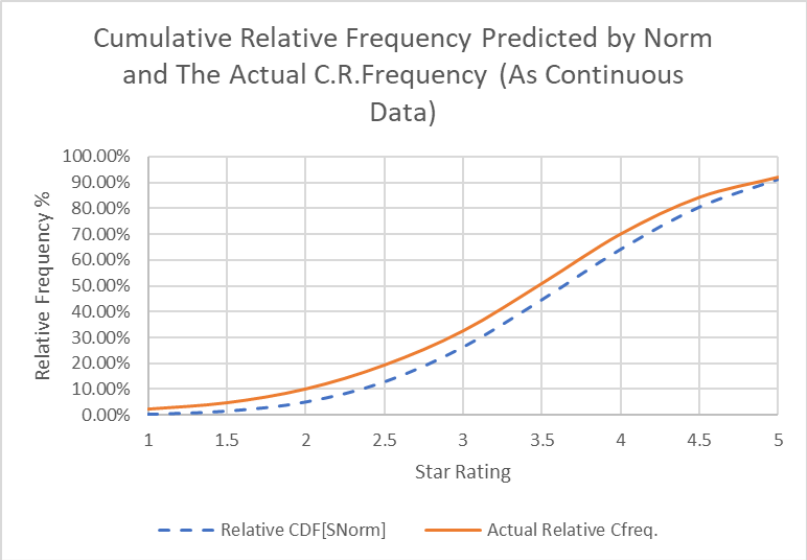

Figure 7: Cumulative Frequency Distribution Compared to S.Norm

To apply the two-tail-z-test, we are interested in whether consistency exists between Toronto's mean score and the overall mean score in North America. First of all, the **null hypothesis** and the **alternative hypothesis** are stated (where $\bar{X}$ represents the **sample mean** and μ_0 is the **population mean**):

$$H_0 : \bar{X} = \mu_0$$

$$H_a : \bar{X} \neq \mu_0$$

Before runnint the test, it is practical to determine the **rejection rule**: if the calculated z-score corresponds to a probability greater than 0.05, we do *not* have enough **evidence** to reject the null hypothesis (in which case the difference is regarded as **sampling variations**); if the critical z-score has a **probability density** which is less than 0.05 but greater than 0.025, in this specific context, we have decent evidence to *reject* the null hypothesis and *accept* the alternative hypothesis with 90% confidence; similarily, if the probability density is less than 0.025, we have 95% confidence to *reject* the null and *accept* the alternative hypothesis (Rosenblatt, 1956).

The calculation of z-score is:

$$z* = \frac{\bar{X} - \mu_0}{\sigma} = \frac{3.487332946 - 3.63218934}{1.001867165} = 0.144316964$$

$$z* = 0.144316964 \sim p* = 0.885422 > 0.05$$

Therefore, from the calculation above, we conclude that we *do not* have enough evidence to reject the null hypothesis at a 90% confidence level and state that Toronto's ratings is *consistent* with the overall ratings distribution in North America.

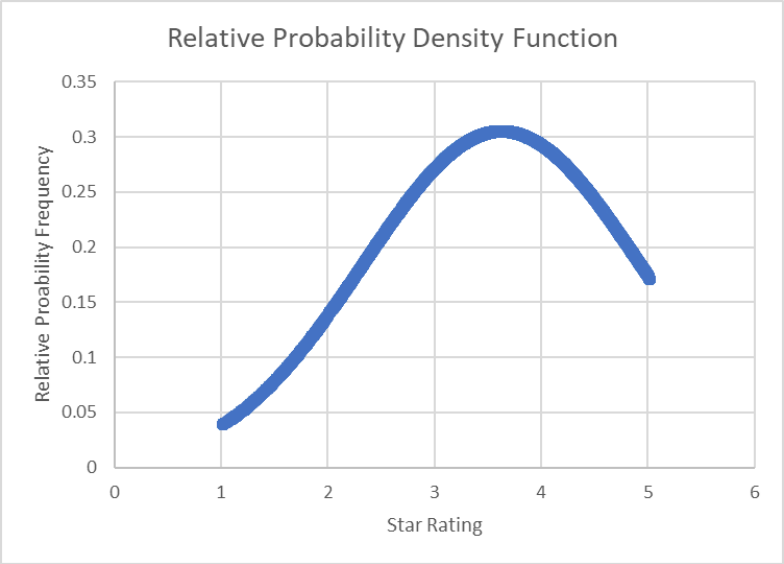

Figure 8: Relative Frequency Distribution Compared to S.Norm

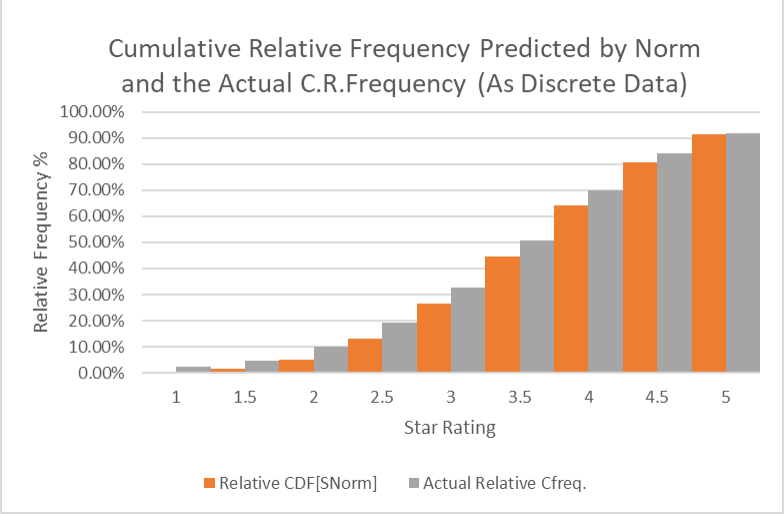

Figure 9: Cumulative Frequency Distribution Compared to S.Norm

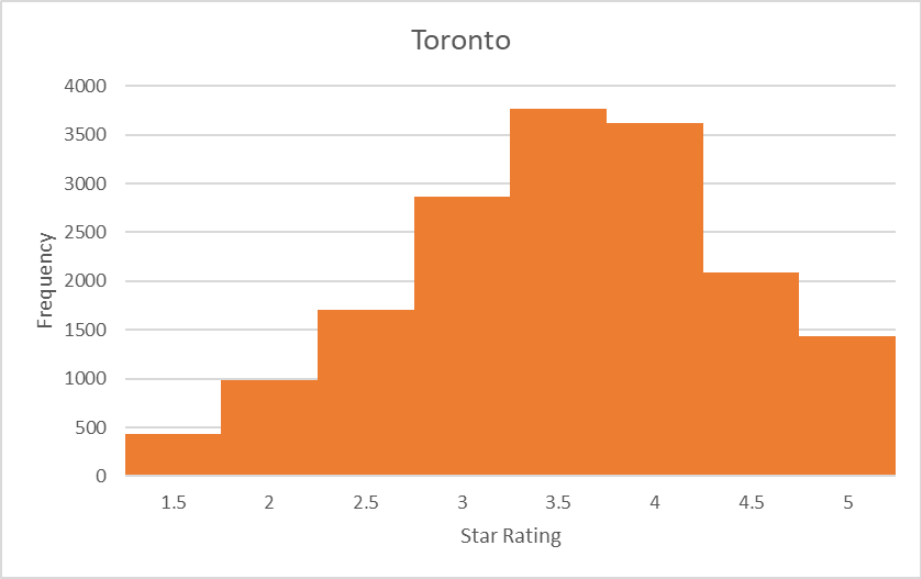

Figure 10: Ratings Distribution of Toronto (As Discrete Data)

3.3.3 Sub-conclusion

Toronto's ratings in 2017 are consistent with the overall ratings distribution in North America. Thus, *the normal approximation model can be generalized to specific cases* (i.e. Toronto) as a rough approximation of business success in most cities. The question remains, can we generalize this discrepancy in regional consistency?

3.4 Do all Yelp Regions Hold the Same Level of Interest for the Platform?

3.4.1 Motivation

This question is intended to study the influence and degree of value of the Yelp Platform in various cities. This is primarily assessed by examining the varied participation from region to region. A service that is used by more people could imply that it is a valuable tool to guage business success. Differences in *popularity* (engagement rate) by region could indicate a potential **regional bias** in the platform's influence on consumers decisions. The degree of popularity can be described by the engagement rate in each city, while the differences are shown by **direct comparison** and the application of proper **significance tests**. In this context, a strong non-uniform distribution of engagement accross regions will imply that the yelp platform is *biased* as a measure of business sucess in North America, while a uniform distribution will justify that Yelp is a non-biased estimator of business success and population distribution.

3.4.2 Analysis

Because we are expecting a uniform distribution it is reasonable to compare actual counts to the expected regional distributions, with respect to the population of the city. For the scope of this analysis, I examined the 14 largest cities by actual reviews. The actual participation and the expected are calculated, and listed in Table 10 below. Since the data are divided into 14 categories, a χ^2 **test of goodness of fit** is employed to do the comparison. But before the test, there are three conditions to be satisfied (Rosenblatt, 1956):

1. The examined data are **absolute frequencies**, not proportions or means;
2. The sample size is large enough that a χ^2 distribution is approximately followed;
3. All frequencies in the table are *at least 5*.

All three conditions are verified by Table 10.

Table 10: Engagement Counts by City Compared to Expected Counts Based on Population

Location	Population	Expected Proportion	Expected Freq.	Actual Freq.	Delta Freq.
Las Vegas	632,912	6.845%	43324.03321	332286	1927314.058
Phoenix	1,615,000	6.845%	110549.8294	122082	1202.995602
Toronto	5,928,040	6.845%	405785.6413	81752	258751.8877
Scottsdale	246,645	6.845%	16883.32054	63748	130086.8615
Charlotte	842,051	6.845%	57639.99653	53228	337.7119116
Henderson	292,969	6.845%	20054.28667	40814	21489.95398
Pittsburgh	303,625	6.845%	20783.71019	35920	11023.4057
Mesa	484,587	6.845%	33170.90414	34164	29.73206239
Tempe	182,498	6.845%	12492.33608	32512	32082.62578
Chandler	247,477	6.845%	16940.27253	28483	7864.959509
Montreal	1,750,000	6.845%	119790.8368	27438	71199.48984
Gilbert	237,133	6.845%	16232.206	25780	5616.01856
Glendale	200,831	6.845%	13747.26489	18219	1454.574063
Cleveland	385,809	6.845%	26409.36169	17378	3088.506832

First we state the null hypothesis and the alternative hypothesis:
H_0: The participation distribution is presented as:

n1	n2	n3	n4	n5	n6	n7
43324	110549	405785	16883	57639	20054	20783

n_8	n_9	n_{10}	n_{11}	n_{12}	n_{13}	n_{14}
33170	12492	16940	119790	16232	13747	26409

H_0: The participation distribution is presented as:

n1	n2	n3	n4	n5	n6	n7
43324	110549	405785	16883	57639	20054	20783

n_8	n_9	n_{10}	n_{11}	n_{12}	n_{13}	n_{14}
33170	12492	16940	119790	16232	13747	26409

Generally, if the agreement between the actual and the expected is perfect, then all $O_i = E_i$ and thus $\chi^2 = 0$, while the worse the fit, the larger χ^2 will be. The rejection rule for this test is: if the χ^2 value is smaller than 0.05, then with 95% confidence, we have enough evidence to reject the **null hypothesis** and accept the **alternative hypothesis**; and if the χ^2 value is greater than 0.05, there is not enough evidence to reject the null. The χ^2 value for the χ^2 **test of goodness of fit** for determining the statistical significance of the regional deviations from the expected value is calculated as:

$$\chi^2(14) = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} = 2471542.781 \sim p * (14) \approx 0 \ll 0.05$$

This χ^2 value 2471542.781 corresponds to a probability extremely close to zero on the χ^2 distribution of **degree of freedom 13** ($df = 11 - 1 = 10$) and thus, far smaller than 0.05. We reject the null hypothesis and accept the alternative hypothesis instead. From the calculation above, we conclude that the regional bias difference is **statistically significant** and not solely caused by **sampling variations**. Factors affecting the difference in engagement from city to city can be complex, but some are obvious. For example it makes sense that a city like Las Vegas would have a disproportionate amount of user engagement to the overall because of all the entertainment in that city. Additionally, many rural cities that only have a few businesses on Yelp would also experience the adverse effect. Therefore this regional bias makes practical sense as well as mathematical sense.

3.4.3 Sub-conclusion

From the discussion above, we conclude that Yelp cannot be an accurate measure of the overall level of Business success as it is **biased** with respect to the region in question.

4 Conclusion

4.1 Summary of Sub-conclusions

4.1.1 Sub-question 1

Comment 1: We see that although the overall challenge in getting a higher Yelp rating is increasing (as demonstrated by the declining mean rating), it has a very small effect on the overall rating distribution.

Explanation 1: Although the challenge and mean rating is varying over time, the IQR is not shifting.

4.1.2 Sub-question 2

Comment 2: In most time, both the overall and city based engagement counts are dependent on time and indicate widespread support for the platform, suggesting Yelp is credible.

Explanation 2: The participation counts show steady increase upon observation.

4.1.3 Sub-question 3

Comment 3: The overall Yelp ratings can be generalized to act as a rough approximation of the business success on a **city level**. This can be generalized on a case by case basis.

Explanation 3: The overall ratings distribution in North America is consistent with Toronto's ratings distribution and status in 2017.

4.1.4 Sub-question 4

Comment 4: Yelp cannot be an accurate measure of the overall business success level in North America since its engagement is biased with respect to the city regions it is prevalent in.

Explanation 4: The regional engagement rate is far different with the expected proportions of engagement based on the populations of the regions. Proven by the significance test.

4.2 Ultimate Conclusion

When used as a measure of the overall distribution of business success in North America, Yelp can be an accurate parameter over *time*, and *regional subsets* but not all of space.

Explanation 4: The Yelp results can be compared over years since the rating distribution is independent of varying challenge (from Comment 1); the participation is constant over time, both over the whole distribution and city distributions (from Comment 2). The Yelp results cannot reflect the business success metrics regionally mainly because of the differences in the degree of non-response bias in each city (from Comment 4). Additionally, the Yelp distribution from any given regional subset (city) can be used to determine the business success relative to the overall rating distribution in North America.

4.3 Limitations

4.3.1 Limitation 1

The one and only limitation of this study is really that there is a massive difference between those cities which widely use Yelp and those who are indifferent to the service. As a result we see huge variations from region to region.

4.3.2 Limitation 2

Although Yelp's database is several terabytes large, it still doesn't come close to representing every single business in North America, so conclusions drawn from Yelp data; albeit accurate and generally true, doesn't come close to being a one size fits all solution to quantifying business success.

4.4 Code

As I mentioned earlier the dataset I scraped through was truly massive, I had to write several SQL scripts to query the terabytes of data. This had to be done because there is no way Excel would be able to handle opening a file sized 1 GB let alone a terabyte.

The following is just one of the sample scripts I wrote:

```
select yelp_business.city, yelp_review._stars_ as 'Stars'
from yelp_review
inner join yelp_business ON yelp_review._business_id_ = yelp_business.business_id
where year(yelp_review._date_) = 2017 AND city = 'Toronto'
```

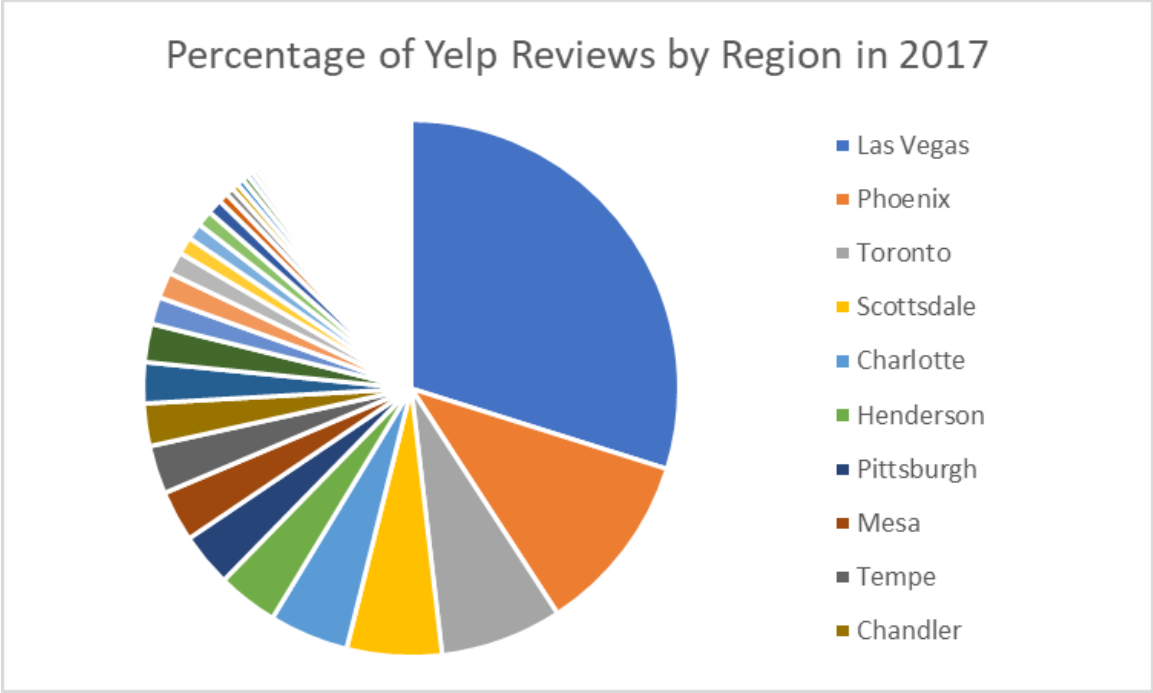

Figure 11: Pie chart for Engagement Contributed by Each City

References

Kaggle datasets. (n.d.). Retrieved 2018-05-10, from `https://www.kaggle.com/yelp-dataset/yelp-dataset/data`

Koren, Y. (2009). The bellkor solution to the netflix grand prize. *Netflix prize documentation*, *81*, 1–10.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, *42*(1), 43–47.

Yelp dataset challenge. (n.d.). Retrieved 2018-05-10, from `https://www.yelp.com/dataset/challenge`