Yelp Data Set Analysis & Report

Coursera, University of California – SQL for Data Science
Stephen Doan – Nguyen
10/26/2023

Table of Contents

Introduction

The Yelp Data Set Analysis report presents an exploratory and comprehensive examination of the Yelp's extensive dataset. The analysis delves into the different layers of data provided by Yelp which includes business reviews, user interactions, and a wealth of associated attributes. By employing SQL queries and data analysis techniques, the report analysis aims to discover:

- Entities insights
- Relationships between entities
- Distribution patterns

By venturing into this set of data, the report reaches out to offer a narrative that not only highlights the quantitative aspects of the dataset but also includes the qualitative nuances that define the Yelp ecosystem. The goal is to provide a narrative that is as insightful and actionable for businesses as it is enlightening for users and researchers alike.

Exploratory Data Analysis

Exploratory Data analysis (EDA), will determine the data analysis workflow by taking the dataset given by the course and breaking them down into their fundamental attributes, relationships, and correlations between them. Commonly the key components of EDA are 1.Data Collection, 2.Understandinng the Data, 3.Cleaning the Data, 4. Identify Correlations, 5. Statistical Methods, 6. Data Visualization. Given the limited amount of resources, we can utilize the resources given to the utmost satisfaction of most of the components and the given requirements from the SQL for Data Science course. The type of EDA is Multivariate EDA and

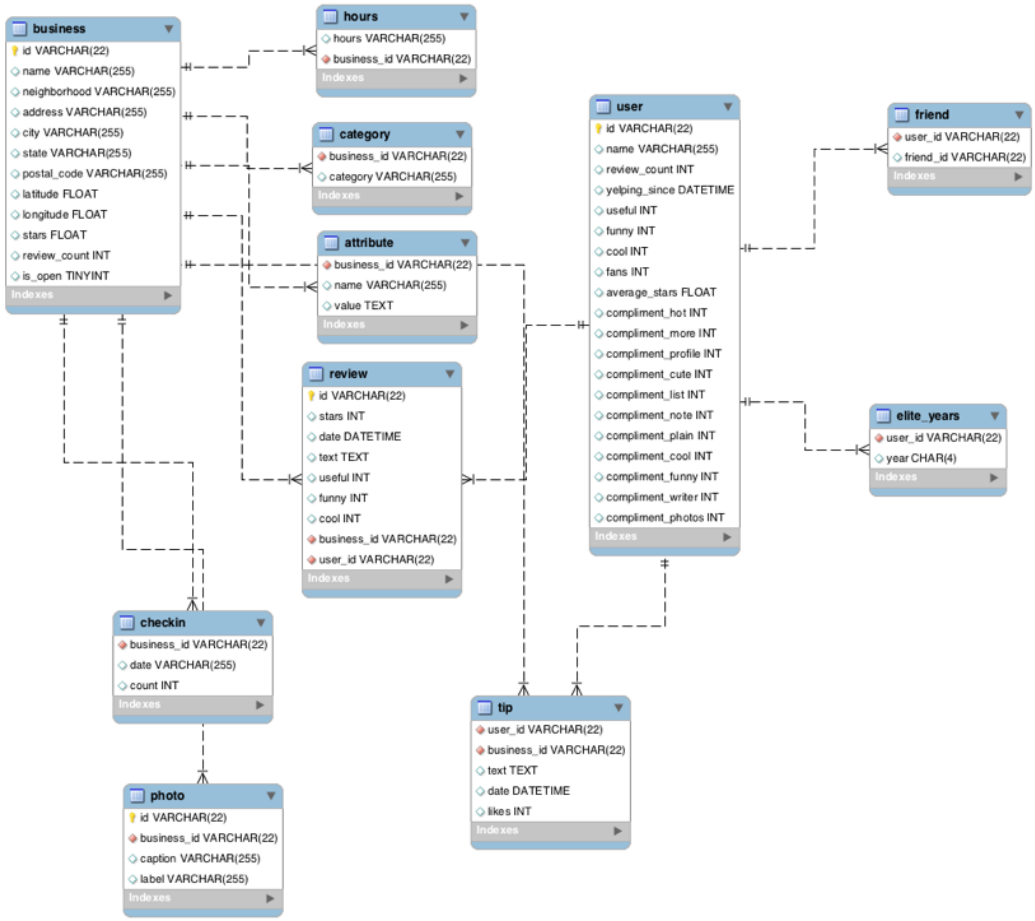
will satisfy the methods of Multivariate EDA with the limited resources and conditions given. Multivariate EDA is a type of EDA that will explore the relationships between two or more entities, with the goal of understanding how those entities react to one another. The principal objective of the EDA is to obtain distinguishable correlations that can help in decision-making.

Dataset

Description

The dataset was provided by Coursera, SQL for Data Science. Although the dataset was not made public due to privacy reasons from Yelp, the results from the queries and entity-relationship diagram were made public. The results from the queries will be used to meet requirements and components. The Yelp dataset offers a large collection of information and attributes containing reviews about a variety of organizations - businesses, restaurants, health clubs, hospitals, local governmental offices, and charitable organizations. The dataset serves as a foundation and complementary resource for the reviews of organizations.

Entity Relationship Diagram



Data Type Identification, Entity - attributes

The entity-relationship diagram that was provided outlines the schema for the review platform Yelp, in detailing the various interconnected entities and their respective attributes within their relationships.

Business - The central entity which encapsulates the core information of the listed organizations in the dataset. Each business is identified uniquely by an ID and then further on by the business's name, neighborhood, address, city, state, postal code, geographical coordinates, average rating stars, count of reviews, and operational status.

Hours and Attributes – Linked to the Business entity are two other entities which are the Hours and the Attributes. The Hours entity contains the operating times of a business

defined by the business ID, the day of the week, and the hours of opening/closing. The Attribute entity carries the specific characteristics of a business such as parking and wifi availability, which is described by a name and as well as a value for each business ID.

Review, Checkin, Photo – The interactions of users with the businesses are stored within the Review, CheckIn, Photo entities. The Review entity entails the textual feedback from the user's interaction, where each review bears its own ID, the business ID it pertains to, user's ID, a star rating, the date of submission, and count of helpfulness, funny, or coolness. The Checkin entity records the frequencies of a check-in of the user at a business, logging the instances with a business ID, date, and count. Photos that have correlations to the business are stored within the Photo entity, where each of the photo is identified by the ID and which contains a caption and labels with the associated business ID.

Users - The profiles/users of the Yelp platform are represented by the User entity each containing a unique ID and personal details as well such as name, review count, date joined, and activity metrics. An average star rating is also represents the average rating of all of the user's review. Users also is able to receive various types of compliments that would each be stored as a count against types like profile, cute, list, note, plain, cool, funny, writer, photos, and more.

Tips - Users who leave a short message or advice are recorded within the Tip entity which are linked to the user and business entailing the text of the tip, the submission of the text, and the amount of likes accumulated.

Friend - The social connections between the users are represented by the Friend entity. Indicating user relationship through user and friend IDs

Elite Years – tracks the years in which a user has achieved the 'elite' status which is recognition often for highly active or influential members, storing the user ID and the corresponding years of having the elite status.

Relationship Analysis, Entity – relationships

The relationships of each diagram that was provided with a series of interlinked entities depict how different the components and entities are related and signifies how they interact with one another. The Business entity are the central of the schema forming several other relation ships with several other entities.

Business – Hours relationship is a one-to-many indicating that a singular business can have multiple operating hours entries typically for different days during the week, although the operating hours is specific to a singular business allowing the platform of the structure to reflect on the multiple operation times for each of the business.

Business – Attribute relationship is a one-to-many. Each business may have multiple of varieties of attributes that describe its feature, which the following could be "accept credit cards", "wifi available", "parking", and more. A business can have multiple attributes but each of the attributes is only linked to one business.

Business – Review relationship is considered as a one-to-many. This is a vital relationship for the review platform, allowing numerous reviews from the different users to a single business. Each review is linked to one business which reflects the direct feedback from the customer/users.

Business – Checkin relationship is established as a one-to-many as well following a similar pattern as the business-review relationship. This relationship allows to have

multiple check-in records for each business which helps indicate the popularity or customer engagement levels. Each check-in instance is related back to one business entity.

User – Review relationship is a one-to-many, where a user has the option to write multiple reviews for different businesses although each of the review is written by a single user.

User – Tip relation is also a one-to-many, similarly to the user-review, users could provide multiple tips across numerous of business, but each tip is written by a single user

User – Friend relationship is a many-to-many, allowing users to have a numerous amounts of friendships across the platform, and each friend can be friends with multiple other users which creates a network of connections.

User – Elite Years relationship is a one-to-many, highlighting the user's status within the community. This relationship implies that users can be recognized as elite over the multiple of years, indicating the sustained contributions or the influence on the review platform

These relationships are established by the use of the primary and foreign keys, securing the data integrity that is maintained within the database. The relationships of the Entity-Relationship Diagram is denoted by different connections which is indicated by crow's feet notation revealing the type of relationship between the entities.

Profiling the Dataset

We will be profiling the dataset by satisfying the following requirements using SQL queries. The profiling of the dataset will give us more of an indepth understanding of the range and size of the data set for the following queries in the future

1. **Profile the data by finding the total number of records for each of the table below**

i. Attribute table = 10000

```
SELECT COUNT(*)
FROM attribute
```

ii. Business table = 10000

```
SELECT COUNT(*)
FROM business
```

iii. Category table = 10000

```
SELECT COUNT(*)
FROM category
```

iv. Checkin table = 10000

```
SELECT COUNT(*)
FROM checkin
```

v. elite_years table = 10000

```
SELECT COUNT(*)
FROM elite_years
```

vi. friend table = 10000

```
SELECT COUNT(*)
FROM friend
```

vii. hours table = 10000

```
SELECT COUNT(*)
FROM hours
```

viii. photo table = 10000

```
SELECT COUNT(*)
FROM photo
```

ix. review table = 10000

```
SELECT COUNT(*)
FROM review
```

x. tip table = 10000

```
SELECT COUNT(*)
FROM tip
```

xi. user table = 10000

```
SELECT COUNT(*)
FROM user
```

2. **Find the total records by either the foreign key or primary key for each table. If two foreign keys are listed in the table, please specify which foreign key**

i. Business = 10000

```
SELECT COUNT(distinct id)
FROM business
```

ii. Hours = 1562

```
SELECT Count(distinct id)
FROM business
```

iii. Category = 2643

```
SELECT COUNT (distinct business_id)
FROM category
```

iv. Attribute =1115

```
SELECT COUNT (distinct business_id)
FROM attribute
```

v. Review = 10000(id), 8090 (business_id), 9581(user_id)

```
SELECT
   (SELECT COUNT(distinct id FROM review) AS unique_review_count,
   (SELECT COUNT (distinct business_id FROM review) AS unique_review_count
   (SELECT COUNT (distinct user_id FROM review) AS unique_review_count,
```

vi. Checkin = 493

```
SELECT COUNT (distinct business_id)
FROM checkin
```

vii. Photo = 10000 (id), (business_id) = 6493

```
SELECT
 (SELECT COUNT (distinct id)FROM photo) AS photo_count,
 (SELECT COUNT (distinct business_id) FROM photo) As businessid_count
```

viii. Tip = (user_id) 537, (business_id) = 3979

```
SELECT
 (SELECT COUNT (distinct user_id)FROM tip) AS id_count,
(SELECT COUNT (distinct business_id) FROM tip) As businessid_count
```

ix. User = 10000

```
SELECT COUNT (distinct id)
FROM user
```

x. Friend = (user_id) 11

```
SELECT COUNT (distinct user_id)
FROM friend
```

xi. Elite_years = (user_id) 11

```
SELECT COUNT (distinct user_id)
FROM friend
```

Data Quality Assessment

The Data Quality Assessment will assess the quality of the data is the data pertains any values that do not hold up to the technical standard. Within the database, there are no values within the columns in the user tables that contain a null value. We can identify this by the SQL query:

```
SELECT count(*)
FROM user
WHERE id IS NULL
```

The data set contains no null values therefore no missing values are within this dataset. If a null value were to exist within the data set, we can use the `DELETE` to delete the column that would contain a null value. The following query would be used :

```
DELETE
FROM user
WHERE id IS NULL
```

Value Distribution

In terms of Value Distribution, we will look at the distribution, frequencies, and outliers among the data sets. We can determine the success by answering the following objectives that were given:

1) **Find the distribution of star ratings to the business in the following cities: Avon**

The SQL code that was used to query the database and extract the distribution of the star ratings of businesses in the city of Avon:

```
SELECT stars as [Star Rating], count(stars) as [Count]
FROM business b
WHERE city = 'Avon'
GROUP by stars;
```

Star Ratings of businesses in the city of Avon

Star Ratings	Count
1.5	1
2.5	2
3.5	3
4.0	2
4.5	1
5.0	1

The distribution of star ratings is formed in a discrete frequency distribution, where the data points are distinct and separate, and the frequency of each point is counted. The results of the query show a total of six different ratings that were given to the businesses. From the query, we can deduce the mode, skewness, and range of the distribution. The data illustrates the range of ratings from 1.5 to 5 stars, with varying counts of businesses at each rating level. The mode of the distribution is 3.5 stars with the highest frequency of 3, with the concentration of businesses in the middle with a slight skew towards the lower ratings indicating that the businesses within the city of Avon show a moderate level of customer satisfaction.

2) **Find the distribution of star ratings to the business in: Beachwood**

The SQL code that was used to query the database and extract the distribution of the star ratings of businesses in the city of Beachwood:

```
SELECT stars as [Star Rating], count(stars) as [Count]
FROM business b
WHERE city = 'Beachwood'
GROUP BY stars;
```

Star Ratings	Count
2.0	1
2.5	1
3.0	2
3.5	2
4.0	1
4.5	2
5.0	5

The result of the query shows an overview for businesses in the city of Beachwood, presented in a frequency distribution table with a range of ratings from 2.0 to 5.0 stars. At the lower end of the scale, we can see singular businesses with low star ratings from the range of 2.0 to 2.5 stars, indicating the underperformance of customer satisfaction compared to other businesses. Progressing towards the mid-range we see the ratings have become denser: 2 businesses have received a 3.0 star rating and another following 2 have received a 3.5 star rating. This signifies a modest level of customer satisfaction. The distribution begins to peak at the 5.0 star rating, with the mode of the distribution being 5.0 showing an indication of exceptional service and customer experience. With the type of distribution, we can observe the skewness is towards the higher ratings because of the dense and clustered amount of rating counts. In other words, while there is a representation across the distribution, it mainly leans towards the higher ratings which could indicate that these business have distinguishing features.

3) Find the top 3 users based on their total reviews:

The query used to determine the results of the top 3 users based on their total reviews:

```
SELECT name, review_count
FROM user
ORDER BY review_count DESC
LIMIT 3;
```

Name	Review_counnt
Gerald	2000
Sara	1629
Yuri	1339

The results of the SQL query provide an ordered list of the top three users from the 'users' database, ranked inn descending order of their total number of reviews. The distribution of the review counts decreases with each rank showing a "power-law" distribution. We can suggest that the distribution is a power-law due to decreasing frequency and with the increase of counts. For instance there might be one user with 2000 counts but several users with 1000 counts. This creates a distribution where the frequency of users decreases as the number of reviews increases. If we were to plot the number of users against the number of reviews on a log-log scale, we would likely be able to see a rough linear relationship for the portion of the distribution indicating that when users decrease the number of reviews increases, following a specific pattern.

4) Are there more reviews with the word "love" or with the word "hate" in them?

There are more reviews the word "love" and the query used to determine the result:

```
SELECT
  SUM(CASE WHEN LOWER(text) LIKE "%love%" THEN 1 ELSE 0 END) as love_text,
  SUM(CASE WHEN LOWER(text) LIKE "%hate%" THEN 1 ELSE 0 END as hate_text
FROM review;
```

love_text	hate_text
1780	232

Given the results, it is evident that the word “love” appears in reviews significantly more often then “hate”. This sort of distribution between the two texts suggests that customers are more inclined and willing to express positive experiences than negative reviews.

5) Find the top 10 users with the most fans:

The SQL query results provide a ranking of the top 10 users by the number of fans they have on the platform. The query used to determine the distribution with the top 10 users with the most fans:

```
SELECT u.name, u.fans
FROM user AS u
ORDER BY u.fans DESC
LIMIT 10;
```

name	fans
Amy	503
Mimi	497
Harald	311
Gerald	253
Christine	173
Lisa	159
Cat	133
William	126
Fran	124
Lissa	120

This distribution could indicate the power law or the Pareto distribution, given the small number of individuals (users) account for a large portion of total outcomes (fans). This type of distribution is common within social dynamics, where popularity are quite often concentrated within the small segments of the population. The steep decline from the top users to those with fewer fans is also a typical characteristic of the power-law distribution. The distribution also shows that the popularity among these users is not uniform, and that there is a

variation from one user to another especially within the top two ranks. This indicates and emphasizes the disparities of the population within the platform.

6) List the cities with the most reviews in descending order:

The SQL code that was used to query the database and extract the list of cities with the most reviews in descending order

```
SELECT city, sum(review_count)
FROM business
GROUP BY city
ORDER BY sum(review_count) DESC
```

city	Sum(reviewcount)
Las Vegas	82854
Pheonix	34503
Toronto	24113
Scottsdale	20614
Charlotte	12523
Henderson	10871
Tempe	10504
Pittsburgh	9798
Montreal	9448
Chandler	8112
Mesa	6875
Gilbert	6380
Cleveland	5593
Madison	5265
Glendale	4406
Missisauga	3814
Edinburgh	2792
Peoria	2624
North Las Vegas	2438
Markham	2352
Champaign	2029
Stuttgart	1849
Surprise	1520
Lakewood	1465
Goodyear	1155

The distribution of the review counts among the cities represents a class long-tail distribution. A few numbers of the cities account for a disproportionately large number of reviews which can be from the characteristics of having larger populations, a greater number of

businesses, or highly active reviewing communities. As the list goes down, the cities become to have fewer counts of reviews; reflecting the nature of many real-world datasets, where the few items(cities) dominate the frequency(reviews). Similarly again the pattern follows the power-law distribution which is often related to the Pareto principle (80/20) rule, where roughly 80% of the effects come from the 20% of causes. In this scenario, its likely that few numbers of cities contribute to a large portion of the reviews while the other cities have more of a smaller presence within the dataset

Summary Statistics

The summary statistics will summarize the properties of the data, giving insight into the shape, size, and spread of the data. The measure of central tendency and the following requirements will be the success criteria in satisfying the summary statistics.

7) **For each table and column listed below, display the smallest (minimum), largest (maximum), and average (mean) values for the following fields:**

   a. **Table: Review, Column: Stars**

      min: 1 max: 5 avg: 3.7082

```
SELECT
     MIN(stars) AS min _stars,
     MAX(stars) AS max _stars,
     AVG(stars) AS avg _stars
FROM Review;
```

   b. **Table: Business, Column: Stars**

      min: 1 max: 5 avg: 3.6549

```
SELECT
     MIN(stars) AS min _stars,
     MAX(stars) AS max _stars,
     AVG(stars) AS avg _stars
FROM Business;
```

c. Table: Tip, Column: Likes

min: 0 max: 2 avg:0.0144

```
SELECT
     MIN(likes) AS min _ likes,
     MAX(likes) AS max _ likes,
     AVG(likes) AS avg _ likes
FROM Tip;
```

d. Table: Checkin, Column: Count

min: 1 max: 53 avg:1.9414

```
SELECT
     MIN(count) AS min _ count,
     MAX(count) AS max _ count,
     AVG(count) AS avg _ count
FROM Checkin;
```

e. Table: User, Column: Review_count

min: 1 max: 2000 avg: 24.2995

```
SELECT
     MIN(review_count) AS min _ review_count,
     MAX(review_count) AS max _ review_count,
     AVG(review_count) AS avg _ review_count
FROM User;
```

The data result from the various of tables suggest a general positive inclination among the users of the platform. Reviews and business ratings have averages above the midpoint of the scale as well with review ratings being slightly higher on average than business ratings; indicating that the individual experiences might be rated more favourably than overall business performance. The tip likes are used sparingly by the user base, as it is reflected in the low average like count, this could imply that users are not engaged with the features or that tips left are not often found to be useful by others. Check-ins have a wide range, but the average suggests

that they are relatively infrequent, which could indicate that either the majority of places do not compel repeated visits or users do not regularly use the check-in feature. Finally, the review activity by users shows a vast range, from none to 2000 reviews per user, with the average being 24 reviews; suggesting that there is a small subset of highly active users against a backdrop of a much larger number of minimally active or inactive users. Combined, the data points illustrate a picture of an active and generally positive community with different degrees of engagement across the different features of the platform

Inferences and Analysis

The first two sections are the inferences sections, creating inferences on the results of the query to satisfy the criteria. The last section is the Yelp analysis, analyzing the dataset. The inferences and analysis will be based off the given criteria from the course by choosing from the following givens and using SQL queries results to analyze and create inferences to give further insight of the data set; thus answering the questions given.

Section 1: Pick one city and category of your choice and group the businesses in that city or category by their overall star rating. Compare the businesses with 2-3 stars to the businesses with 4-5 stars and answer the following questions.

1. **Do the two groups you choose to analyze have different distribution hours?**

The two groups chosen to be analyzed for the different distribution hours are “Las Vegas” as the city and “Shopping” as the category. The query below shows that the two groups chosen have slightly different distribution hours:

```
SELECT  CASE WHEN stars >= 4 THEN "4-5 stars"
                                  WHEN stars >= 2 THEN "2-3 stars"
                                  ELSE "below 2"
                                  END star_rank,
                               city,
                               c.category,
```

```
        count(distinct business.id) AS company_count,
        count(h.hours) AS working_days
FROM business
JOIN hours h ON business.id = h.business_id
JOIN category c ON business.id = c.business_id
WHERE city = "Las Vegas" AND c.category = "Shopping"
GROUP BY star_rank
```

Results:

star_rank	city	category	company_count	working_days
2-3 stars	Las Vegas	Shopping	2	13
4-5 stars	Las Vegas	Shopping	2	12

The group with 2-3 stars has companies that are operational for a total of 13 days, whereas the '4-5' group has companies operational for a total of 12 days. It is difficult to make a definite and confident analysis with the limited amount of information given that there are only 2 companies within the research target, thus not providing an accurate statement for the shopping businesses in Las Vegas. Although with the limited amount of information, I can assume that the 2-3 stars would require a slightly more working day to generate more profit as 4-5 stars have more customers and attraction

2. **Do the two groups you chose to analyze have a different number of reviews?**

The two groups I chose to analyze have a different number of reviews from the 4-5 stars is doubled compared to the 2-3 stars, as we mentioned earlier in question 4.(Value distribution) customers seem to be more inclined to make positive reviews then negative reviews. The SQL code used to support the analysis:

```
SELECT  CASE WHEN stars >= 4 THEN "4-5 stars"
             WHEN stars >= 2 THEN "2-3 stars"
             ELSE "below 2"
             END star_rank,
        city,
        c.category,
        count(distinct business.id) AS company_count,
        sum(review_count) AS total_review
FROM business
JOIN category c ON business.id = c.business_id
WHERE city = "Las Vegas" AND c.category = "Shopping"
```

```
GROUP BY star_rank
```

Results:

star_rank	city	category	company_count	working_days
2-3 stars	Las Vegas	Shopping	2	17
4-5 stars	Las Vegas	Shopping	2	36

3. Can you infer anything from the location data provided by these two groups?

From the locational data that was provided from the query, there is a noticeable difference in the geographical locations and distribution of the stores within the two groups; '2-3 stars' and '4-5 stars'. Both of the stores rated within the '2-3 stars' have the same postal codes areas which is associated with the Southeast and Eastside of Las Vegas. This could indicate that a cluster of stores that are similar are potentially serving a specific local market or demographic, or it might suggest that it serves a particular competitive or economic environment in that area. On the other side, the stores with '4-5 stars' ratings have different postal codes showing that they are geographically separated. The dispersion of the higher rated stores could show that their target market is more of a diverse customer base, suggesting that they draw customers from a wider area or marketing in more of a higher economic are. The preliminary analysis leads to different types of hypothesis and inferences such that the lower-rated stores in the same area could be. The SQL code used to support the inference:

```
SELECT  CASE WHEN stars >= 4 THEN "4-5 stars"
                WHEN stars >= 2 THEN "2-3 stars"
                ELSE "below 2"
                END star_rank,
            address,
            neighborhood,
            city,
            postal_code
    FROM business
    JOIN category c ON business.id = c.business_id
    WHERE city = "Las Vegas" AND c.category = "Shopping"
    ORDER BY star_rank
```

Results:

star_rank	address	neighborhood	postall_code
2-3 stars	3421 E Tropicana Ave, Ste I	Southeast	89121
2-3 stars	3808 E Tropicana Ave	Eastside	89121
4-5 stars	1000 Scenic Loop Dr	2	89161
2-3 stars	3555 W Reno Ave, Ste F		89118

Section 2: Group business based on the ones that are open and the ones that are closed. What differences can you find between the ones that are still open and the ones that are closed? List at least two differences and the SQL code that you used to arrive at your answer.

1. **Difference 1: Total review is noticeably higher between open and closed business**
   The businesses that are still open have a significantly higher cumulative number of reviews (269,300) compared to those that have closed (35,261). The disparity is mainly because the business has closed down thus not being able to receive any more reviews.

2. **Difference 2: Average stars are given very close to each other, we can infer which business were closed solely to poor service or quality**
   The average stars for open businesses is 3.68, where for closed business its 3.52. The close range indicates that open businesses have the advantage to still increase their ratings still they are able to still receive ratings.

```
SELECT  CASE    WHEN is_open = 1 THEN "STILL OPEN"
                    WHEN is_open = 0 THEN "CLOSED"
                    END status,
            count(distinct id) AS num_company,
            sum(review_count) AS total_review,
            round(avg(review_count),2) AS avg_review,
            round(avg(stars),2) AS avg_stars
    FROM business
    GROUP BY is_open
    ORDER BY status DESC
```

Results:

Status	num_company	total_review	avg_review	avg_stars
STILL OPEN	8480	269300	31.76	3.68
CLOSED	1520	35261	23.2	3.52

Yelp Analysis:

In this analysis, the aim is to identify the most successful businesses across the different categories. To achieve this, the required data is needed from the business table; business ID, star ratings, and review counts, coupled with categorical data from the category table. It is essential to consider the number of businesses within each category and as well the average star ratings assigned by customers, alongside the total number of reviews to ensure the robustness and lack of bias in our analysis. To refine the data as well as enhance the relevance of our findings, the focus of the categories will contain at least ten businesses and boast an average customer rating of 3.5 stars of higher. The specific approach ensures that we only analyze the most relevant and pertinent data, leading to a better and clearer understanding of the attributes that characterize the most successful businesses within each category. The SQL code to achieve this:

```
SELECT  category,
        count(distinct id) AS num_companies,
        round(avg(stars),2) AS avg_stars,
        sum(review_count) total_reviews
FROM business
JOIN category ON business.id =
category.business_id
GROUP BY category
HAVING avg_stars >= 3.5 AND num_companies >= 10
ORDER BY avg_stars DESC
```

Results:

category	num_categories	avg_stars	total_reviews
Local Services	12	4.21	100
Active Life	10	4.15	131
Health & Medical	17	4.09	203
Home Services	16	4.0	94
Shopping	30	3.98	977
Beauty & Spas	13	3.88	119
American (Traditional)	11	3.82	1128
Food	23	3.78	1781
Bars	17	3.5	1322

It is evident that success is measured by customer ratings and isn't uniform across all categories. What measures as success in local services might not directly apply to food or bars. Each category has its own dynamic contributing characteristic to customer satisfaction and its reviews. It also shows within the relationship between total reviews and the average stars, that the relationship isn't linear indicating that the relationship does not determine success but rather other quantitative and qualitative factors that depend on each category

Final Results

The analysis of the Yelp dataset provides a dive into the fabric of user engagement and business success across different cities and categories. The analysis illustrates a picture of a community where positive expression like "love" significantly differs from the negative sentiments such as "hate" in reviews, indicating a general pattern and tendency of users to share positive experiences.

The data reveals a distribution of star ratings, where certain cities like Avon and Beachwood display a range of customer satisfaction levels, highlighting a discrete frequency distribution of ratings. This indicates that businesses within these cities experience a spectrum of performance, with Avon displaying a moderate level of customer satisfaction and Beachwood showing an inclination towards higher ratings, which may indicate distinguishing features or exceptional service within some businesses.

The profiling of user activity through reviews unveils a power-law distribution, typical of social platforms where few users account for a large portion of interactions. In this case, users like Gerald, Sara, and Yuri dominate in terms of the number of reviews they have contributed, which could be indicative of their significant influence or engagement on the platform. When focusing on the cities with the most reviews, a long-tail distribution becomes apparent, with a

small number of cities like Las Vegas, Phoenix, and Toronto amassing a disproportionately large number of reviews. This likely reflects the dense populations, a greater number of businesses, or more active reviewing communities present in these areas.

Finally, a look into the summary statistics offers insights into the general behavior of the Yelp community, with overall positive ratings and a significant spread in user review activity, ranging from highly active individuals to a majority who are less involved. Through this analysis, the report underscores the nuanced and complex nature of the data within Yelp's platform, revealing patterns and distributions that speak to the intricate interplay of user engagement, business performance, and social dynamics.

References

"SQL for Data Science." *Coursera*, Coursera, www.coursera.org/learn/sql-for-data-science?specialization=learn-sql-basics-data-science&utm_medium=sem&utm_source=gg&utm_campaign=B2C_NAMER_learn-sql-basics-data-science_ucdavis_FTCOF_specializations_country-US-country-CA&campaignid=20743946861&adgroupid=158296001907&device=c&keyword=&matchtype=&network=g&devicemodel=&adposition=&creativeid=679555118314&hide_mobile_promo&gad_source=1&gclid=CjwKCAjw17qvBhBrEiwA1rU9w9wUSAoTis037fYuYdUeCR6auXBEg32iXQbdrVFHmWjpBALnTphmDRoC5DIQAvD_BwE. Accessed 11 Mar. 2024.

Power Law Distributions, rinterested.github.io/statistics/power_law.html. Accessed 11 Mar. 2024.

"Exploratory Data Analysis." *EPA*, Environmental Protection Agency, www.epa.gov/caddis/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,step%20in%20any%20data%20analysis. Accessed 11 Mar. 2024.