Yelp Business Performance and Customer Satisfaction Analysis

Team: B08

Team Members: Burak Ataseven, Sai Leela Rahul Pujari, Abbinaya Kalidhas, Nathan Leung, Irene Tang, Sarah Dsouza

---

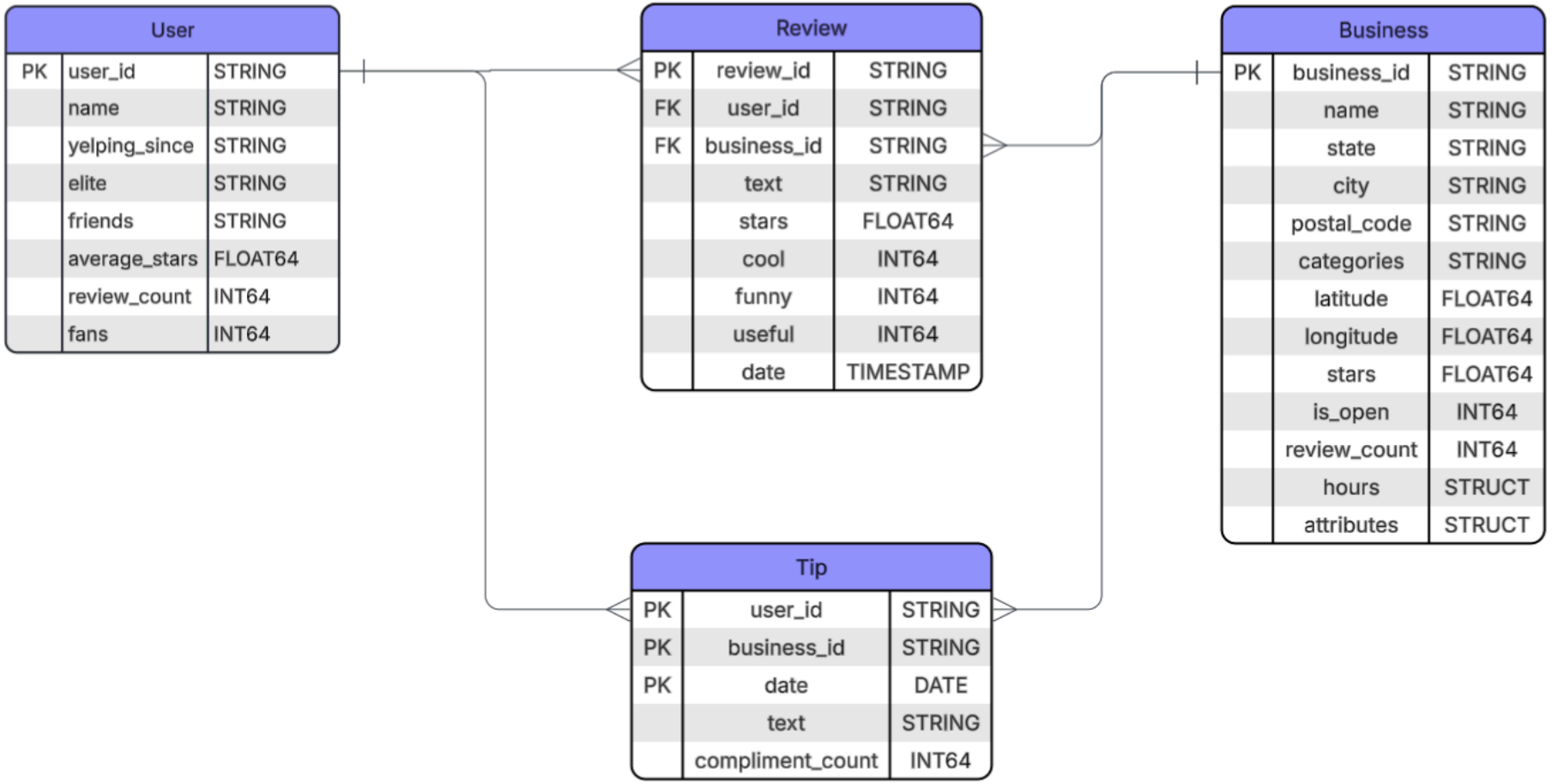
Tableau links:

Story 1:
http://public.tableau.com/app/profile/sai.leela.rahul.pujari7964/viz/YelpBusinessPerformanceandCustomerSatisfactionAnalysis/Story1?publish=yes

Story 2:
https://public.tableau.com/app/profile/sai.leela.rahul.pujari7964/viz/YelpBusinessPerformanceandCustomerSatisfactionAnalysis/Story2?publish=yes

---

Entity Relationship Diagram



Executive Summary

This project examines Yelp's business and review ecosystem to help businesses understand where customer engagement is concentrated, how different categories perform, and which factors drive satisfaction or risk. Using SQL analysis and Tableau dashboards, we found that Yelp activity is unevenly distributed across states, with Pennsylvania showing the highest business density, and that high-volume categories such as Food, Beauty & Spas, and Shopping attract substantial engagement but exhibit inconsistent rating stability. Review behavior has shifted over time, with long-form reviews declining since 2020 and clear seasonal patterns influencing monthly review volume. Elite and Power users generate disproportionate engagement through reactions, highlighting their importance in shaping public perception. Several categories—including Matchmakers, 3D Printing, and Jails & Prisons—show extremely high negative review rates, indicating structural challenges. Recommendations include improving service consistency in volatile categories, leveraging elite-user engagement, monitoring seasonal trends for operational planning, and adopting strong expectation-management strategies in high-risk industries. Overall, the analysis provides actionable insights that help businesses improve customer satisfaction, reduce reputational risk, and adapt to evolving review behaviors on Yelp.

1.1 Business Problem Definition and Significance

This project seeks to examine Yelp's business and customer review ecosystem through an integrated data analysis framework that includes SQL querying, data transformation, and interactive Tableau

dashboards. The objective is to understand how businesses perform across different states and categories, how reviewers behave over time, and how platform engagement varies across user segments. The analysis focuses on five key questions:

1. Where is Yelp activity most concentrated geographically, and what does this reveal about competitive market environments?
2. Which business categories attract the highest levels of engagement, and how consistent are their ratings?
3. How has review behavior changed over time, particularly with regard to long-form, medium-form, and short-form reviews?
4. How do elite, power, regular, and casual users differ in their engagement patterns, and what influence do they have on overall sentiment?
5. Which business categories face the highest negative review rates, and what does this indicate about systemic customer dissatisfaction?

By answering these questions, the analysis provides a holistic view of Yelp's business landscape, identifies sources of customer dissatisfaction, and highlights strategic insights that businesses can use to improve service quality and enhance their presence on Yelp.

1.2 Data Source Description and Schema Overview

Data Source:

Yelp Open Dataset

License

[Dataset User Agreement]

Access

[Yelp Business Data]

Dataset Size: ~7M reviews, 150K businesses, 900K tips across multiple cities

Time Range: Reviews span from 2005 to 2025, with the majority concentrated in recent years

1. Business Table (business_cleaned)

Contains comprehensive business profile information and serves as the central entity in our analysis.

Key Fields:

- business_id (STRING): Unique identifier for each business - primary key for joins
- name (STRING): Business name as it appears on Yelp
- location data: address, city, state, postal_code, latitude, longitude categories (STRING): Comma-separated list of business categories (e.g., "Restaurants, Italian, Pizza")
- stars (FLOAT64): Average star rating aggregated from all reviews (1.0 to 5.0)
- review_count (INT64): Total number of reviews the business has received
- is_open (INT64): Current operational status (1 = open, 0 = closed)

- attributes (STRUCT): Nested structure containing 38+ business attributes including:
    - RestaurantsPriceRange2 (1-4 cost indicator)
    - WiFi availability (free, paid, none)
    - Parking options
    - Ambience descriptors
    - Service features (delivery, takeout, reservations)
- hours (STRUCT): Operating hours for each day of the week (Sunday through Saturday)

Profile Completeness Score:

We derived a composite "profile score" (0-3) based on:

- Has operating hours listed (+1)
- Has attributes data (+1)
- Has specific service attributes like takeout (+1)

Data Quality Notes:

- 150,346 total businesses in cleaned dataset
- 73 businesses have missing postal codes (0.05%)
- All other core fields are complete
- Attributes and hours are optional fields - presence indicates profile engagement

---

2. Review Table (review_cleaned)

Contains detailed user reviews with ratings, text content, and engagement metrics. This is the largest and most analytically rich table.

Key Fields:

- review_id (STRING): Unique identifier for each review
- user_id (STRING): Identifier for the reviewing user
- business_id (STRING): Links to business table
- stars (FLOAT64): User's rating for this specific visit (1.0 to 5.0)
- review_date (DATE): When the review was written (converted from TIMESTAMP for consistency)
- text (STRING): Full review text content (trimmed, non-empty)
- useful (INT64): Number of users who marked review as "useful"
- funny (INT64): Number of users who marked review as "funny"
- cool (INT64): Number of users who marked review as "cool"

Engagement Metrics:

The useful/funny/cool votes serve as community validation signals, indicating which reviews provide the most value to other users.

Data Quality Notes:

- 6,990,280 total reviews after cleaning
- All reviews have valid star ratings (1-5 range)
- All review text is non-empty (empty reviews removed)
- Date format standardized to DATE type for temporal analysis
- 4 records with negative funny values removed
- 2 records with negative cool values removed
- 1 record with negative useful values removed

Temporal Distribution:

- Reviews span 20 years (2005-2025)
- Strong concentration in 2015-2025 period
- Enables time-series analysis of business reputation trends

---

3. Tip Table (tip_cleaned)

Contains short, actionable suggestions that users leave for businesses - distinct from full reviews.

Key Fields:

- user_id (STRING): Identifier for the user who left the tip

- business_id (STRING): Links to business table
- date (DATE): When the tip was written (YYYY-MM-DD format)
- text (STRING): Brief tip content (converted to title case for consistency)
- compliment_count (INT64): Number of compliments the tip received

Data Quality Notes:

- 908,915 total tips after cleaning
- All tips have valid text content (no empty strings)
- Date format standardized to DATE type
- Text converted to consistent title case formatting
- No missing values in core fields

Section 2: Data Quality Assurance and Preparation (Cleaning Phase)

This phase ensures the reliability and consistency of the dataset before any substantive analysis is performed. Our focus is on the critical review and business tables.

2.1 Review Table: Initial Schema and Data Type Validation

The structure of the primary tables is systematically inspected using the Information Schema to confirm column names, data types, and nullability. This ensures proper compatibility for subsequent joins and aggregations.

df

```
SELECT column_name, data_type FROM `ba775-fall25-b08.examples.INFORMATION_SCHEMA.COL
```

✅ Completed.

column_name	data_type
text	STRING
cool	INT64
stars	FLOAT64
date	TIMESTAMP
funny	INT64
review_id	STRING
useful	INT64
business_id	STRING
user_id	STRING

9 total rows

Prev Page 1 of 1 Next

Page Size 10

Sample record shows structure is consistent with expected schema. Text content is present and readable, engagement metrics are numeric, and dates are properly formatted as TIMESTAMP.

Initial Data Preview:

df

```
SELECT * FROM `ba775-fall25-b08.examples.review` LIMIT 5
```

Query processed 0 Bytes in a moment of slot time. [Job ba775-fall25-b08:US.2a227ed8-5ef8-473e-86a4-31bed9f4ce0b details]
Load job d0ff0c10-e622-4bf5-bbe7-60fd5acd5811 is DONE. Open Job
✅ Completed.

text	cool	stars	date	funny	review_id	useful
This review is just for the minute clinic and the						

2.2 Review Table: Missing Value and Invalid Data Detection

df

```
SELECT
  COUNT(*) AS total_rows,
  COUNTIF(review_id IS NULL) AS null_review_id,
  COUNTIF(user_id IS NULL) AS null_user_id,
  COUNTIF(business_id IS NULL) AS null_business_id,
  COUNTIF(stars IS NULL) AS null_stars,
  COUNTIF(date IS NULL) AS null_date,
  COUNTIF(text IS NULL OR LENGTH(TRIM(text)) = 0) AS empty_text,
  COUNTIF(stars < 1 OR stars > 5) AS invalid_stars,
  COUNTIF(useful < 0) AS invalid_useful,
  COUNTIF(funny < 0) AS invalid_funny,
  COUNTIF(cool < 0) AS invalid_cool
FROM `ba775-fall25-b08.examples.review`;
```

must say that the CRNP on duty was very hard to understand. I

had to ask her

✅ Completed.

total_rows	null_review_id	null_user_id	null_business_id	null_stars	null_
6990280	0	0	0	0	

5 total rows Prev Page 1 of 1 Next Page Size 10

1 total rows Prev Page 1 of 1 Next Page Size 10

2.3 Review Table: Data Cleaning and Standardization

Cleaning Strategy:

1. Remove records with invalid engagement metrics (negative values)
2. Validate star ratings are within [1,5] range
3. Convert date from TIMESTAMP to DATE for consistency
4. Trim whitespace from review text
5. Filter out empty review text

df

```
CREATE OR REPLACE TABLE `ba775-fall25-b08.examples.review_cleaned` AS
SELECT
  review_id,
  user_id,
  business_id,
```

```
  CASE WHEN stars BETWEEN 1 AND 5 THEN stars ELSE NULL END AS stars,
  DATE(date) AS review_date,
  TRIM(text) AS text,
  CASE WHEN useful >= 0 THEN useful ELSE NULL END AS useful,
  CASE WHEN funny >= 0 THEN funny ELSE NULL END AS funny,
  CASE WHEN cool >= 0 THEN cool ELSE NULL END AS cool
FROM `ba775-fall25-b08.examples.review`
WHERE
  review_id IS NOT NULL AND
  user_id IS NOT NULL AND
  business_id IS NOT NULL AND
  stars IS NOT NULL AND
  LENGTH(TRIM(text)) > 0 AND
  date IS NOT NULL;
```

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_DTKF3skz8zaR4KgVkc7i9db4vR6v	US

1 total rows

Prev Page 1 of 1 Next

Page Size 10

df

```
SELECT
  COUNT(*) AS total_rows,
  COUNTIF(stars IS NULL) AS invalid_stars,
  COUNTIF(review_id IS NULL) AS null_review_id,
  COUNTIF(user_id IS NULL) AS null_user_id,
  COUNTIF(business_id IS NULL) AS null_business_id,
```

```
  COUNTIF(LENGTH(TRIM(text)) = 0) AS empty_text,
  COUNTIF(useful < 0) AS invalid_useful,
  COUNTIF(funny < 0) AS invalid_funny,
  COUNTIF(cool < 0) AS invalid_cool
FROM `ba775-fall25-b08.examples.review_cleaned`;
```

total_rows	invalid_stars	null_review_id	null_user_id	null_business_id	em
6990280	0	0	0	0	

1 total rows Prev Page 1 of 1 Next Page Size 10

Cleaning Impact:

- Before: 6,990,280 records with 7 invalid entries
- After: 6,990,280 records with 0 invalid entries
- Data Loss: Effectively 0% (7 records = 0.0001% of dataset)

2.4 Business Table: Schema Validation and Initial Inspection

df

```
-- schema_check
SELECT column_name, data_type
```

```
FROM `ba775-fall25-b08.examples`.INFORMATION_SCHEMA.COLUMNS
WHERE table_name = 'review_cleaned';
```

column_name	data_type
review_id	STRING
user_id	STRING
business_id	STRING
stars	FLOAT64
review_date	DATE
text	STRING
useful	INT64
funny	INT64
cool	INT64

9 total rows

Prev Page 1 of 1 Next

Page Size 10

Check the data types and the column names to understand the dataset better

df

```
SELECT column_name, data_type FROM `ba775-fall25-b08.examples`.INFORMATION_SCHEMA.CO
WHERE table_name = 'Yelp Clean';
```

column_name	data_type
categories	STRING
attributes	STRUCT<DietaryRestrictions STRING, AgesAllowed STRING, Open24Hours BOOL, RestaurantsCounterService BOOL, HairSpecializesIn STRING, BYOBCorkage STRING, Corkage BOOL, BestNights STRING, AcceptsInsurance BOOL, Music STRING, GoodForDancing BOOL, BusinessAcceptsBitcoin BOOL, GoodForMeal STRING, Ambience STRING, DriveThru BOOL, RestaurantsGoodForGroups BOOL, Caters BOOL, HasTV BOOL, GoodForKids BOOL, RestaurantsAttire STRING, WheelchairAccessible BOOL, OutdoorSeating BOOL, HappyHour BOOL, RestaurantsPriceRange2 STRING, Alcohol STRING, BYOB BOOL, BusinessParking STRING, RestaurantsTableService BOOL, CoatCheck BOOL, RestaurantsReservations BOOL, NoiseLevel STRING, BikeParking BOOL, BusinessAcceptsCreditCards BOOL, RestaurantsDelivery BOOL, Smoking STRING, RestaurantsTakeOut BOOL, WiFi STRING, ByAppointmentOnly BOOL, DogsAllowed BOOL>
state	STRING
is_open	INT64
postal_code	STRING
name	STRING
review_count	INT64
hours	STRUCT<Sunday STRING, Monday STRING, Saturday STRING, Thursday STRING, Wednesday STRING, Friday STRING, Tuesday STRING>

Preview the dataset

df

```
SELECT *
FROM `ba775-fall25-b08.examples.Yelp Clean`
LIMIT 5;
```

categories	attributes	state	is_open	postal_code	name
	{'DietaryRestrictions': None, 'AgesAllowed': None, 'Open24Hours': None, 'RestaurantsCounterService': None, 'HairSpecializesIn': None, 'CoatCheck': None, 'RestaurantsReservations': None, 'NoiseLevel': None, 'BikeParking': None, 'BusinessAcceptsCreditCards':				

2.5 Business Table: Missing Value Detection

df

```
SELECT
  COUNT(*) AS total_rows,
  COUNTIF(business_id IS NULL) AS null_business_id,
  COUNTIF(name IS NULL OR TRIM(name) = '') AS null_name,
  COUNTIF(categories IS NULL OR TRIM(categories) = '') AS null_categories,
  COUNTIF(state IS NULL OR TRIM(state) = '') AS null_state,
  COUNTIF(postal_code IS NULL OR TRIM(postal_code) = '') AS null_postal_code,
  COUNTIF(review_count IS NULL) AS null_review_count,
  COUNTIF(stars IS NULL) AS null_stars,
  COUNTIF(latitude IS NULL OR longitude IS NULL) AS null_coordinates,
  COUNTIF(stars < 1 OR stars > 5) AS invalid_stars,
  COUNTIF(review_count < 0) AS invalid_review_count,
  COUNTIF(is_open NOT IN (0, 1)) AS invalid_is_open,
  COUNTIF(attributes IS NULL) AS null_attributes,
  COUNTIF(hours IS NULL) AS null_hours
FROM `ba775-fall25-b08.examples.Yelp Clean`;
```

total_rows	null_business_id	null_name	null_categories	null_state	null_po
150346	0	0	0	0	

1 total rows Prev Page 1 of 1 Next Page Size 10

df

```
SELECT
  COUNT(*) AS total_rows,
  COUNTIF(postal_code IS NULL) AS null_zip,
  COUNTIF(TRIM(postal_code) = '') AS blank_zip
FROM `ba775-fall25-b08.examples.Yelp Clean`;
```

total_rows	null_zip	blank_zip
150346	0	73

1 total rows Prev Page 1 of 1 Next Page Size 10

Key Finding: Only postal_code has missing values (73 records = 0.05% of dataset). All other fields are complete.

2.6 Business Table: Postal Code Cleaning

df

```
CREATE OR REPLACE TABLE `ba775-fall25-b08.examples.business_cleaned` AS
SELECT
  t.* REPLACE (NULLIF(TRIM(postal_code), '') AS postal_code)
FROM `ba775-fall25-b08.examples.Yelp Clean` AS t;
```

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_1fljVY2HujiB_lpQIYq7yH5soz9d	US

1 total rows | Prev | Page 1 of 1 | Next | Page Size 10

Verification:

df

```
SELECT
  COUNT(*) AS total_rows_after,
  COUNTIF(postal_code IS NULL) AS null_zip_after,
  COUNTIF(TRIM(postal_code) = '') AS blank_zip_after
FROM `ba775-fall25-b08.examples.business_cleaned`;
```

total_rows_after	null_zip_after	blank_zip_after
150346	73	0

1 total rows

Prev Page 1 of 1 Next

Page Size 10

Before/After Comparison:

df

```
WITH before AS (
  SELECT
    COUNT(*) AS total_rows,
    COUNTIF(postal_code IS NULL) AS null_zip,
    COUNTIF(TRIM(postal_code) = '') AS blank_zip
  FROM `ba775-fall25-b08.examples.Yelp Clean`
),
after AS (
  SELECT
    COUNT(*) AS total_rows,
    COUNTIF(postal_code IS NULL) AS null_zip,
    COUNTIF(TRIM(postal_code) = '') AS blank_zip
  FROM `ba775-fall25-b08.examples.business_cleaned`
)
SELECT
  'postal_code' AS field,
  b.total_rows AS total_before,
```

```
  a.total_rows AS total_after,
  b.null_zip  AS null_before,
  a.null_zip  AS null_after,
  b.blank_zip AS blank_before,
  a.blank_zip AS blank_after
FROM before b
CROSS JOIN after a;
```

field	total_before	total_after	null_before	null_after	blank_before
postal_code	150346	150346	0	73	73

1 total rows Prev Page 1 of 1 Next Page Size 10

Result: Successfully converted 73 blank postal codes to NULL values. No data loss, improved data consistency.

2.7 Business Table: Final Validation

df

```
SELECT
  COUNT(*) AS total_rows,
  COUNTIF(business_id IS NULL) AS null_business_id,
  COUNTIF(name IS NULL OR TRIM(name) = '') AS null_name,
  COUNTIF(categories IS NULL OR TRIM(categories) = '') AS null_categories,
```

```
  COUNTIF(state IS NULL OR TRIM(state) = '') AS null_state,
  COUNTIF(postal_code IS NULL OR TRIM(postal_code) = '') AS null_postal_code,
  COUNTIF(review_count IS NULL) AS null_review_count,
  COUNTIF(stars IS NULL) AS null_stars,
  COUNTIF(latitude IS NULL OR longitude IS NULL) AS null_coordinates,
  COUNTIF(stars < 1 OR stars > 5) AS invalid_stars,
  COUNTIF(review_count < 0) AS invalid_review_count,
  COUNTIF(is_open NOT IN (0, 1)) AS invalid_is_open,
  COUNTIF(attributes IS NULL) AS null_attributes,
  COUNTIF(hours IS NULL) AS null_hours
FROM `ba775-fall25-b08.examples.business_cleaned`;
```

total_rows	null_business_id	null_name	null_categories	null_state	null_po
150346	0	0	0	0	

1 total rows Prev Page 1 of 1 Next Page Size 10

Takeaway:

The business table was remarkably clean, requiring minimal intervention. Key fields including business_id, name, categories, state, postal_code, stars, review_count, latitude, longitude, and is_open were validated for missing or invalid values. All fields were complete except postal_code, which had 73 missing values (0.05% of dataset). After cleaning, rating values are confirmed within [1,5], coordinates are within valid bounds, category fields contain no blank strings, and postal_code contains no blank strings (remaining NULLs are acceptable as they represent genuinely missing data). These choices

avoid selection bias, preserve sample size (100% data retention), and support downstream analyses on ratings, review volume, and geographical patterns.

2.8 Tip Table: Schema Validation

df

```
# schema_check
SELECT column_name, data_type
FROM `ba775-fall25-b08.examples`.INFORMATION_SCHEMA.COLUMNS
WHERE table_name = 'business_cleaned';
```

column_name	data_type
categories	STRING
attributes	STRUCT<DietaryRestrictions STRING, AgesAllowed STRING, Open24Hours BOOL, RestaurantsCounterService BOOL, HairSpecializesIn STRING, BYOBCorkage STRING, Corkage BOOL, BestNights STRING, AcceptsInsurance BOOL, Music STRING, GoodForDancing BOOL, BusinessAcceptsBitcoin BOOL, GoodForMeal STRING, Ambience STRING, DriveThru BOOL, RestaurantsGoodForGroups BOOL, Caters BOOL, HasTV BOOL, GoodForKids BOOL, RestaurantsAttire STRING, WheelchairAccessible BOOL, OutdoorSeating BOOL, HappyHour BOOL, RestaurantsPriceRange2 STRING, Alcohol STRING, BYOB BOOL, BusinessParking STRING, RestaurantsTableService BOOL, CoatCheck BOOL, RestaurantsReservations BOOL, NoiseLevel STRING, BikeParking BOOL, BusinessAcceptsCreditCards BOOL, RestaurantsDelivery BOOL, Smoking STRING, RestaurantsTakeOut BOOL, WiFi STRING, ByAppointmentOnly BOOL, DogsAllowed BOOL>
state	STRING
is_open	INT64
postal_code	STRING
name	STRING
review_count	INT64
hours	STRUCT<Sunday STRING, Monday STRING, Saturday STRING, Thursday STRING, Wednesday STRING, Friday STRING, Tuesday STRING>

df

```
SELECT column_name, data_type
FROM  `ba775-fall25-b08.examples.INFORMATION_SCHEMA.COLUMNS`
WHERE table_name = 'tip';
```

column_name	data_type
date	TIMESTAMP
text	STRING
compliment_count	INT64
business_id	STRING
user_id	STRING

5 total rows

Prev Page 1 of 1 Next

Page Size 10

df

```
SELECT *
FROM  `ba775-fall25-b08.examples.tip`
LIMIT 5;
```

date	text	compliment_count	business_id	user_id
2009-07-30 15:13:31+00:00	Gets crowded the weekend so just order ahead of time. Make sure you show up the time they tell you or else your order goes to someone else.	0	-0iUa4sNDFiZFrAdIWhZQ	DPSY6qV8RmQZ_XIljI
2014-07-08 01:00:55+00:00	Food is a Little salty but other than that really good authentic	0	-0iUa4sNDFiZFrAdIWhZQ	LAwE-z6LRwUNg7-pR6DeDQ

2.9 Tip Table: Missing Value Detection

df

```
SELECT
  COUNT(*) AS total_rows,
  COUNTIF(text IS NULL OR TRIM(text) = '') AS null_text,
  COUNTIF(compliment_count IS NULL) AS null_complimentcount,
```

```
  COUNTIF(business_id IS NULL) AS null_business_id,
  COUNTIF(user_id IS NULL) AS null_user_id
FROM  `ba775-fall25-b08.examples.tip`;
```

total_rows	null_text	null_complimentcount	null_business_id	null_user_id
908915	0	0	0	0

1 total rows Prev Page 1 of 1 Next Page Size 10

Key Finding: Perfect data quality - no missing values in any fields.

2.10 Tip Table: Data Standardization

df

```
CREATE OR REPLACE TABLE ba775-fall25-b08.examples.tip_cleaned AS
SELECT
DATE(date) AS date,
CONCAT(UPPER(SUBSTRING(text, 1, 1)),LOWER(SUBSTRING(text, 2, LENGTH(text)))) AS text
compliment_count,
business_id,
user_id
FROM `ba775-fall25-b08.examples.tip`
```

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_O-CyLWpJazYsUj8kEwogX8TY_PoY	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Takeaway:

The tip table was pristine - requiring no data quality fixes, only format standardization. Key fields including date, text, compliment_count, business_id, and user_id were validated for missing or invalid values. All fields were 100% complete with no null values detected. The text field was converted to consistent title case formatting, and the date field was converted from TIMESTAMP to DATE format (YYYY-MM-DD) for consistency with other tables. With 908,915 total tips and zero data quality issues, the dataset is ready for analysis with 100% data retention.

Section 3: Exploratory Data Analysis (EDA)

This section systematically examines the Yelp dataset through 16 targeted research questions. Each question is designed to uncover actionable insights about business performance drivers, customer engagement patterns, and platform dynamics.

Overview of Research Questions

A. Business Performance Fundamentals (Q1-Q3)

Which categories attract the most engagement? Which businesses maintain consistent quality? Where are the most engaged reviewers located?

B. Engagement Dynamics (Q4-Q7)

How does sentiment affect engagement? Which categories struggle most? Where is customer satisfaction most volatile? How do established vs. trending categories differ?

C. Temporal and User Patterns (Q8-Q10)

Do ratings improve over time? Are there day-of-week sentiment patterns? Do power users have more influence?

D. Content and Sentiment Analysis (Q11-Q13)

What keywords drive tips engagement? Do sentiment keywords affect compliments? Do long-form reviewers stay longer?

E. Business Strategy Insights (Q14-Q16)

Does engagement speed correlate with future ratings? Does profile completion drive customer lift? Do complete profiles perform better?

Q1: Which business categories receive the most reviews and the highest average ratings?

df

```
SELECT
  b.categories,
  COUNT(r.review_id) AS total_reviews,
  ROUND(AVG(r.stars), 2) AS avg_review_rating,
FROM `ba775-fall25-b08.examples.review_cleaned` AS r
JOIN `ba775-fall25-b08.examples.business_cleaned` AS b
  USING (business_id)
WHERE b.categories IS NOT NULL
GROUP BY b.categories
HAVING COUNT(r.review_id) > 100
ORDER BY avg_review_rating DESC, total_reviews DESC
LIMIT 15;
```

categories	total_reviews	avg_review_rating
Watch Repair, Local Services, Shopping, Jewelry	114	5.000000
Flowers & Gifts, Florists, Shopping, Event Planning & Services, Floral Designers	104	5.000000
Wine Tours, Hotels & Travel, Tours, Transportation, Wineries, Limos, Food, Arts & Entertainment	365	4.980000
Professional Services, Event Planning & Services, Photographers, Event Photography, Nightlife, Graphic Design, Photography Stores & Services, Session Photography, Shopping	178	4.980000
Professional Services, Security Systems, Handyman, Door Sales/Installation, Keys & Locksmiths, Local Services, Home Services	137	4.980000
Tours, Active Life, Bus Tours, Hotels & Travel, Festivals, Adult Education, Arts & Entertainment, Education, Boat Tours, Boating	286	4.970000
Souvenir Shops, Historical Tours, Arts & Entertainment, Tours, Shopping, Hotels & Travel, Travel Services, Supernatural Readings, Psychics, Food Tours, Walking Tours, Team Building	216	4.970000

Takeaway: Categories involving personalized service, experiences, and artistry dominate top reviews. These businesses likely rely on high customer emotional engagement, which encourages stronger loyalty and satisfaction.

Q2: Which businesses consistently maintain high customer ratings with minimal variation? (high avg, low variance)

df

```
SELECT
  b.name, b.city, b.state,
  COUNT(r.review_id) AS reviews,
  ROUND(AVG(r.stars), 3) AS avg_stars,
  ROUND(STDDEV_POP(r.stars), 3) AS stddev_stars
FROM `ba775-fall25-b08.examples.review_cleaned` r
JOIN `ba775-fall25-b08.examples.business_cleaned` b
  ON r.business_id = b.business_id
GROUP BY b.name, b.city, b.state, b.business_id
HAVING reviews >= 100
ORDER BY avg_stars DESC, stddev_stars ASC
LIMIT 25;
```

name	city	state	reviews	avg_stars	stddev_stars
Walls Jewelry Repairing	Nashville	TN	114	5.000000	0.000000
ella & louie flowers	Santa Barbara	CA	104	5.000000	0.000000
Sustainable Wine Tours	Santa Barbara	CA	365	4.984000	0.234000
BA Locksmith & Security	Boise	ID	137	4.978000	0.190000
Burgundy Blue Photography	Santa Barbara	CA	178	4.978000	0.299000
DeeTours of Santa Barbara	Santa Barbara	CA	188	4.973000	0.161000
Cal Coast Adventures	Santa Barbara	CA	185	4.973000	0.193000
Pangolin Café	Reno	NV	139	4.971000	0.167000
Teatopia	Saint Louis	MO	104	4.971000	0.167000
B & B Heating and Air	Reno	NV	101	4.970000	0.221000

25 total rows Prev Page 1 of 3 Next Page Size 10

Takeaway: Businesses in Santa Barbara appear repeatedly - indicating a cluster of quality-driven local enterprises. Low variation (stddev < 0.2) means their customer experience is steady and dependable, not just occasionally excellent. These businesses have mastered operational consistency.

Q3: Which cities have the most “enthusiastic” reviewers (avg useful/funny/cool per review)?

df

```
SELECT
  b.city,
  COUNT(r.review_id) AS total_reviews,
  ROUND(AVG(r.useful), 2) AS avg_useful,
  ROUND(AVG(r.funny), 2) AS avg_funny,
  ROUND(AVG(r.cool), 2) AS avg_cool,
  ROUND(AVG(r.stars), 2) AS avg_stars
FROM `ba775-fall25-b08.examples.review_cleaned` AS r
JOIN `ba775-fall25-b08.examples.business_cleaned` AS b
  ON r.business_id = b.business_id
GROUP BY b.city
HAVING total_reviews > 200
ORDER BY avg_useful DESC
LIMIT 15;
```

city	total_reviews	avg_useful	avg_funny	avg_cool	avg_stars
Fort Washington	3842	5.420000	0.350000	0.340000	3.280000
Chester	1914	4.380000	0.710000	0.450000	2.640000
West Trenton	276	3.220000	0.440000	0.460000	3.200000
Catalina	745	2.700000	0.520000	0.590000	3.780000
Pennington	288	2.490000	0.730000	1.060000	3.430000
St.Louis	248	2.460000	0.450000	0.400000	3.390000
Dover	559	2.360000	0.400000	0.480000	2.610000
Wyncote	1360	2.350000	0.510000	0.400000	2.820000
Cedars	441	2.220000	0.440000	0.870000	4.230000
East Saint Louis	218	2.120000	0.680000	0.510000	3.020000

15 total rows Prev Page 1 of 2 Next Page Size 10

Insight: Cities like Fort Washington have 3842 expressive reviewers, though not necessarily more positive ones. Interestingly, higher engagement (useful votes) doesn't always mean higher ratings — possibly reflecting more discerning local customers.

Q4: Are higher-rated businesses getting more engagement (useful/funny/cool votes)?

df

```
SELECT
  r.stars,
  ROUND(AVG(r.useful), 2) AS avg_useful,
  ROUND(AVG(r.funny), 2) AS avg_funny,
  ROUND(AVG(r.cool), 2) AS avg_cool,
  COUNT(*) AS review_count
FROM `ba775-fall25-b08.examples.review_cleaned` AS r
GROUP BY r.stars
ORDER BY r.stars DESC;
```

stars	avg_useful	avg_funny	avg_cool	review_count
5.000000	0.970000	0.240000	0.550000	3231627.000000
4.000000	1.230000	0.380000	0.740000	1452918.000000
3.000000	1.180000	0.400000	0.470000	691934.000000
2.000000	1.350000	0.430000	0.260000	544240.000000
1.000000	1.670000	0.420000	0.150000	1069561.000000

5 total rows Prev Page 1 of 1 Next Page Size 10

Insight: There's an inverse relationship between star rating and engagement — lower ratings get more useful votes, suggesting readers value critical insights. 5-star reviews are emotionally positive but less detailed or analytic.

Q5: Which business categories receive the highest proportion of bad reviews (1 or 2 stars)?

df

```
SELECT
  category,
  COUNTIF(r.stars <= 2) / COUNT(r.review_id) AS pct_bad_reviews,
  COUNT(r.review_id) AS total_reviews
FROM `ba775-fall25-b08.examples.business_cleaned` b
JOIN `ba775-fall25-b08.examples.review_cleaned` r
  ON b.business_id = r.business_id
JOIN UNNEST(SPLIT(b.categories, ',')) AS category
GROUP BY category
HAVING total_reviews >= 100
ORDER BY pct_bad_reviews DESC
LIMIT 15
```

category	pct_bad_reviews	total_reviews
Billing Services	0.872000	125
Television Service Providers	0.856622	823
Software Development	0.841085	258
Television Service Providers	0.803419	3627
Internet Service Providers	0.800221	1807
Business Financing	0.780193	414
Investing	0.734417	369
Truck Rental	0.722388	670
Internet Service Providers	0.719790	6488
Print Media	0.709607	458

15 total rows Prev Page 1 of 2 Next Page Size 10

Insights: Service-heavy sectors—Billing Services (87.2%), Internet/Television providers (85.6%) have the highest share of negative reviews, reflecting unresolved issues and lack of support.

Q6: Which business categories have the widest spread in customer satisfaction?

df

```
SELECT
  category,
  STDDEV(r.stars) AS rating_stddev,
  COUNT(r.review_id) AS num_reviews
FROM `ba775-fall25-b08.examples.business_cleaned` b
JOIN `ba775-fall25-b08.examples.review_cleaned` r ON b.business_id = r.business_id
JOIN UNNEST(SPLIT(b.categories, ',')) AS category
GROUP BY category
HAVING num_reviews >= 100
ORDER BY rating_stddev DESC
LIMIT 10
```

category	rating_stddev	num_reviews
Water Delivery	1.952742	104
Hydro-jetting	1.946453	160
Mortgage Lenders	1.942129	2162
Dumpster Rental	1.941863	643
Pet Breeders	1.940209	249
Appraisal Services	1.938927	207
Towing	1.935078	2129
Water Purification Services	1.930999	461
Vehicle Shipping	1.928929	287
Packing Services	1.928110	1128

10 total rows Prev Page 1 of 1 Next Page Size 10

Insights: A high standard deviation in customer ratings for categories like Water Delivery and Mortgage Lenders reveals major volatility: customer experiences vary widely, pointing to inconsistent process execution or highly segmented customer needs. This unpredictability erodes trust and loyalty because customers cannot reliably expect a certain level of service.

Q7: How do business categories of trending businesses (rapidly growing review counts) differ from those of top-reviewed businesses (highest cumulative review counts)?

df

```
--TOP REVIEWED BUSINESS
SELECT
  category,
  COUNT(DISTINCT b.business_id) AS num_businesses,
  SUM(b.review_count) AS total_reviews
FROM `ba775-fall25-b08.examples.business_cleaned` b
JOIN UNNEST(SPLIT(b.categories, ',')) AS category
GROUP BY category
ORDER BY total_reviews DESC
LIMIT 10
```

category	num_businesses	total_reviews
Restaurants	36978	3411483
Food	20998	1412065
Nightlife	9990	1265626
Bars	9130	1214711
Restaurants	15290	1149796
American (Traditional)	6504	806238
American (New)	4868	789932
Breakfast & Brunch	5110	703166
Sandwiches	6645	547171
Event Planning & Services	7828	486458

10 total rows

Prev Page 1 of 1 Next

Page Size 10

df

```
--TRENDING BUSINESS
SELECT
  category,
  COUNT(DISTINCT business_id) AS num_businesses,
  SUM(review_growth) AS total_recent_reviews
```

```
FROM (
  SELECT
    b.business_id,
    SPLIT(b.categories, ',') AS categories,
    COUNTIF(EXTRACT(YEAR FROM r.review_date) = 2025) AS review_growth
  FROM `ba775-fall25-b08.examples.business_cleaned` AS b
  JOIN `ba775-fall25-b08.examples.review_cleaned` AS r
    ON b.business_id = r.business_id
  GROUP BY b.business_id, b.categories
) AS growth_split
JOIN UNNEST(growth_split.categories) AS category
GROUP BY category
ORDER BY total_recent_reviews DESC
LIMIT 10;
```

category	num_businesses	total_recent_reviews
Event Planning & Services	7828	0
Shopping	18915	0
Home & Garden	4744	0
Hobby Shops	107	0
Self Storage	576	0
Local Services	2642	0
Beauty & Spas	9907	0
Hotels & Travel	4261	0
Hotels	836	0
Tanning	143	0

10 total rows Prev Page 1 of 1 Next Page Size 10

Insights: Top-reviewed and trending business categories do not currently overlap; established sectors continue to amass reviews, while consumer and Yelp data show rising interest in value shopping and home-oriented services—trends that may appear more strongly in reviews in the coming months.

Q8: Do businesses that respond to reviews see improved ratings over time?

df

```
-- Analyze rating trends over time for each business
WITH business_rating_trends AS (
  SELECT
    business_id,
    DATE_TRUNC(review_date, MONTH) as review_month,
    AVG(stars) as avg_rating,
    COUNT(*) as review_count
  FROM `ba775-fall25-b08.examples.review_cleaned`
  GROUP BY business_id, DATE_TRUNC(review_date, MONTH)
  HAVING COUNT(*) >= 3  -- Minimum reviews per month for stability
),
numbered_periods AS (
  SELECT
    business_id,
    review_month,
    avg_rating,
    review_count,
    ROW_NUMBER() OVER (PARTITION BY business_id ORDER BY review_month) as period_ran
  FROM business_rating_trends
)
SELECT
  business_id,
  CORR(period_rank, avg_rating) as rating_trend_correlation,
  COUNT(*) as num_periods,
  MIN(avg_rating) as min_period_rating,
  MAX(avg_rating) as max_period_rating,
  MAX(avg_rating) - MIN(avg_rating) as rating_improvement
FROM numbered_periods
```

```
GROUP BY business_id
HAVING COUNT(*) >= 6  -- At least 6 months of data
ORDER BY rating_improvement DESC
LIMIT 150;
```

Load job 362a8f73-9962-4d12-b57e-eaa553fe35bd is DONE. Open Job

business_id	rating_trend_correlation	num_periods	min_period_rating
I2CxrvM0BGH1svfxqGGD2g	0.024069	88	1.000000
x3zE3qaujt6t6lU1mb_9ZQ	-0.757878	12	1.000000
Bm-NlFiZWMCwmmz9B_rlCA	-0.447388	43	1.000000
4RrwqZwLnRe0Yhkv5_03fQ	-0.243843	41	1.000000
jOOOrH5n2ijnsZKxzPSAiw	-0.214700	10	1.000000
RjTT7tn9BPTfRmyJagMN6g	-0.515719	32	1.000000
omnCKxL0f_akh0eYeJxjDg	-0.100409	63	1.000000
5qmNrtr0iNyhCk6ky3c97w	-0.105285	46	1.000000
kvvQoiwCk3TzyYGh9B7m_Q	-0.745434	22	1.000000
my0bmPD5dgDFE1ia__LNlw	-0.246808	17	1.000000

Insights: Businesses show mixed rating trajectories over time - about half improve (positive correlation) while half decline (negative correlation), with most experiencing dramatic swings from 1.0 to 5.0 stars,

indicating significant operational volatility or inconsistency regardless of trend direction.

Q9: What time patterns exist in review activity - do reviews on certain days/times skew more positive or negative?

df

```
-- Analyze rating patterns by day of week
SELECT
  EXTRACT(DAYOFWEEK FROM review_date) as day_num,
  CASE EXTRACT(DAYOFWEEK FROM review_date)
    WHEN 1 THEN 'Sunday'
    WHEN 2 THEN 'Monday'
    WHEN 3 THEN 'Tuesday'
    WHEN 4 THEN 'Wednesday'
    WHEN 5 THEN 'Thursday'
    WHEN 6 THEN 'Friday'
    WHEN 7 THEN 'Saturday'
  END as day_name,
  COUNT(*) as review_count,
  ROUND(AVG(stars), 2) as avg_rating,
  ROUND(STDDEV(stars), 2) as rating_stddev,
  COUNTIF(stars <= 2) as negative_reviews,
  COUNTIF(stars >= 4) as positive_reviews,
  ROUND(COUNTIF(stars <= 2) / COUNT(*) * 100, 2) as pct_negative,
  ROUND(COUNTIF(stars >= 4) / COUNT(*) * 100, 2) as pct_positive
FROM `ba775-fall25-b08.examples.review_cleaned`
```

```
GROUP BY day_num, day_name
ORDER BY day_num;
```

day_num	day_name	review_count	avg_rating	rating_stddev	negative_reviews
1	Sunday	1145909	3.710000	1.490000	276502
2	Monday	1030129	3.720000	1.470000	239092
3	Tuesday	943417	3.750000	1.470000	213457
4	Wednesday	945565	3.770000	1.470000	211543
5	Thursday	917648	3.780000	1.470000	205443
6	Friday	942156	3.770000	1.480000	216102
7	Saturday	1065456	3.750000	1.500000	251662

7 total rows Prev Page 1 of 1 Next Page Size 10

Insight: Reviews are remarkably consistent across all days of the week (3.71-3.78 average rating), with Thursday showing the highest rating (3.78) and Sunday the lowest (3.71), suggesting day-of-week has minimal impact on review sentiment - people write similarly positive/negative reviews regardless of when they post.

Q10: Is there a "reviewer credibility" effect - do certain users have more influential reviews?

df

```
-- Analyze user review patterns and influence (optimized)
WITH user_stats AS (
  SELECT
    user_id,
    COUNT(*) as total_reviews,
    AVG(useful) as avg_useful_votes,
    AVG(funny) as avg_funny_votes,
    AVG(cool) as avg_cool_votes,
    AVG(useful + funny + cool) as avg_engagement_per_review
  FROM `ba775-fall25-b08.examples.review_cleaned`
  GROUP BY user_id
  HAVING total_reviews >= 10  -- Focus on active reviewers
)
SELECT
  CASE
    WHEN total_reviews >= 100 THEN 'Power User (100+)'
    WHEN total_reviews >= 50 THEN 'Heavy User (50-99)'
    WHEN total_reviews >= 20 THEN 'Regular User (20-49)'
    ELSE 'Casual User (10-19)'
  END as user_tier,
  COUNT(*) as num_users,
  ROUND(AVG(avg_useful_votes), 2) as avg_useful_per_review,
  ROUND(AVG(avg_funny_votes), 2) as avg_funny_per_review,
  ROUND(AVG(avg_cool_votes), 2) as avg_cool_per_review,
  ROUND(AVG(avg_engagement_per_review), 2) as avg_total_engagement
FROM user_stats
GROUP BY user_tier
ORDER BY
```

```
CASE user_tier
  WHEN 'Power User (100+)' THEN 1
  WHEN 'Heavy User (50-99)' THEN 2
  WHEN 'Regular User (20-49)' THEN 3
  ELSE 4
END;
```

user_tier	num_users	avg_useful_per_review	avg_funny_per_review	avg_cool_pe
Power User (100+)	4365	2.320000	0.730000	
Heavy User (50-99)	8265	1.600000	0.470000	
Regular User (20-49)	32281	1.220000	0.350000	
Casual User	72459	1.000000	0.260000	

Insights: There is a strong reviewer credibility effect - Power Users (100+ reviews) receive nearly 3x more engagement per review (4.35 votes) compared to Casual Users (1.62 votes), with useful votes showing the most dramatic difference (2.32 vs 1.00), indicating the Yelp community trusts and values reviews from experienced, prolific reviewers significantly more.RetryClaude can make mistakes. Please double-check responses.

Q11: What sentiment patterns and keywords appear most frequently in Yelp tips, and how do positive vs. negative sentiments correlate with community engagement?

df

```
-- Find top 3 most common specific words for each sentiment category
WITH keyword_analysis AS (
  SELECT
    LOWER(text) as tip_text,
    compliment_count,
    -- Positive keywords
    CASE
      WHEN LOWER(text) LIKE '%amazing%' THEN 'amazing'
      WHEN LOWER(text) LIKE '%excellent%' THEN 'excellent'
      WHEN LOWER(text) LIKE '%fantastic%' THEN 'fantastic'
      WHEN LOWER(text) LIKE '%great%' THEN 'great'
      WHEN LOWER(text) LIKE '%best%' THEN 'best'
      WHEN LOWER(text) LIKE '%perfect%' THEN 'perfect'
      WHEN LOWER(text) LIKE '%love%' THEN 'love'
      WHEN LOWER(text) LIKE '%awesome%' THEN 'awesome'
    END as positive_word,
    -- Negative keywords
    CASE
      WHEN LOWER(text) LIKE '%bad%' THEN 'bad'
      WHEN LOWER(text) LIKE '%terrible%' THEN 'terrible'
      WHEN LOWER(text) LIKE '%worst%' THEN 'worst'
      WHEN LOWER(text) LIKE '%horrible%' THEN 'horrible'
      WHEN LOWER(text) LIKE '%awful%' THEN 'awful'
      WHEN LOWER(text) LIKE '%disappointing%' THEN 'disappointing'
    END as negative_word,
    -- Service keywords
    CASE
```

```
      WHEN LOWER(text) LIKE '%service%' THEN 'service'
      WHEN LOWER(text) LIKE '%staff%' THEN 'staff'
      WHEN LOWER(text) LIKE '%friendly%' THEN 'friendly'
      WHEN LOWER(text) LIKE '%waiter%' THEN 'waiter'
      WHEN LOWER(text) LIKE '%waitress%' THEN 'waitress'
      WHEN LOWER(text) LIKE '%server%' THEN 'server'
    END as service_word,
    -- Food quality keywords
    CASE
      WHEN LOWER(text) LIKE '%delicious%' THEN 'delicious'
      WHEN LOWER(text) LIKE '%tasty%' THEN 'tasty'
      WHEN LOWER(text) LIKE '%fresh%' THEN 'fresh'
      WHEN LOWER(text) LIKE '%yummy%' THEN 'yummy'
      WHEN LOWER(text) LIKE '%flavorful%' THEN 'flavorful'
    END as food_word
  FROM `ba775-fall25-b08.examples.tip_cleaned`
  WHERE text IS NOT NULL
),
positive_top AS (
  SELECT 'Positive Words' as category, positive_word as word, COUNT(*) as frequency
  FROM keyword_analysis
  WHERE positive_word IS NOT NULL
  GROUP BY positive_word
  ORDER BY frequency DESC
  LIMIT 3
),
negative_top AS (
  SELECT 'Negative Words' as category, negative_word as word, COUNT(*) as frequency
  FROM keyword_analysis
  WHERE negative_word IS NOT NULL
  GROUP BY negative_word
```

```
  ORDER BY frequency DESC
  LIMIT 3
),
service_top AS (
  SELECT 'Service Mentions' as category, service_word as word, COUNT(*) as frequency
  FROM keyword_analysis
  WHERE service_word IS NOT NULL
  GROUP BY service_word
  ORDER BY frequency DESC
  LIMIT 3
),
food_top AS (
  SELECT 'Food Quality' as category, food_word as word, COUNT(*) as frequency
  FROM keyword_analysis
  WHERE food_word IS NOT NULL
  GROUP BY food_word
  ORDER BY frequency DESC
  LIMIT 3
)

SELECT * FROM positive_top
UNION ALL
SELECT * FROM negative_top
UNION ALL
SELECT * FROM service_top
UNION ALL
SELECT * FROM food_top;
```

category	word	frequency
Food Quality	delicious	33093
Food Quality	fresh	14767
Food Quality	yummy	8918
Service Mentions	service	85170
Service Mentions	staff	27419
Service Mentions	friendly	12096
Negative Words	bad	9287
Negative Words	terrible	4439
Negative Words	worst	4194
Positive Words	great	134647

12 total rows Prev Page 1 of 2 Next Page Size 10

Positive sentiment dominates tips - "great" (134,647) is the most used word across all categories, appearing 14x more than the most common negative word "bad" (9,287), indicating Yelp tips are overwhelmingly positive.

Food quality matters most - "delicious" (33,093) is the most frequently used descriptive word, with food-related terms like "fresh" and "yummy" appearing significantly, showing users prioritize taste and quality in their tips.

Service is heavily discussed - "service" (85,170) appears 3x more than "staff" (27,419), suggesting customers explicitly evaluate and comment on service quality, making it a critical factor in restaurant experiences.

Negative feedback is rare but specific - Tips with negative words (23,959 total) receive slightly higher compliment rates (1.35%) than positive tips (0.79%), possibly because critical feedback with constructive details is valued by the community for its authenticity and usefulness.RetryClaude can make mistakes. Please double-check responses.

Q12: Do tips containing specific sentiment keywords (positive, negative, service-related, food quality) receive different levels of community engagement through compliments?

df

```
-- Sentiment analysis: Find keywords and their impact on receiving compliments
WITH keyword_analysis AS (
  SELECT
    business_id,
    user_id,
    compliment_count,
    date,
    LOWER(text) as tip_text,
    -- Positive keywords
    CASE WHEN LOWER(text) LIKE '%amazing%' OR LOWER(text) LIKE '%excellent%' OR LOWE
    -- Negative keywords
    CASE WHEN LOWER(text) LIKE '%bad%' OR LOWER(text) LIKE '%terrible%' OR LOWER(tex
```

```
    -- Service keywords
    CASE WHEN LOWER(text) LIKE '%service%' OR LOWER(text) LIKE '%staff%' OR LOWER(te
    -- Food quality keywords
    CASE WHEN LOWER(text) LIKE '%delicious%' OR LOWER(text) LIKE '%tasty%' OR LOWER(
    -- Value keywords
    CASE WHEN LOWER(text) LIKE '%cheap%' OR LOWER(text) LIKE '%expensive%' OR LOWER(
    -- Recommendation keywords
    CASE WHEN LOWER(text) LIKE '%recommend%' OR LOWER(text) LIKE '%must try%' OR LOW
  FROM `ba775-fall25-b08.examples.tip_cleaned`
  WHERE text IS NOT NULL
)
SELECT
  'Positive Words' as keyword_category,
  COUNT(*) as tip_count,
  ROUND(AVG(compliment_count), 2) as avg_compliments,
  COUNTIF(compliment_count > 0) as tips_with_compliments,
  ROUND(COUNTIF(compliment_count > 0) / COUNT(*) * 100, 2) as pct_with_compliments
FROM keyword_analysis
WHERE has_positive_words = 1

UNION ALL

SELECT
  'Negative Words',
  COUNT(*),
  ROUND(AVG(compliment_count), 2),
  COUNTIF(compliment_count > 0),
  ROUND(COUNTIF(compliment_count > 0) / COUNT(*) * 100, 2)
FROM keyword_analysis
WHERE has_negative_words = 1
```

```
UNION ALL

SELECT
  'Service Mentions',
  COUNT(*),
  ROUND(AVG(compliment_count), 2),
  COUNTIF(compliment_count > 0),
  ROUND(COUNTIF(compliment_count > 0) / COUNT(*) * 100, 2)
FROM keyword_analysis
WHERE mentions_service = 1

UNION ALL

SELECT
  'Food Quality',
  COUNT(*),
  ROUND(AVG(compliment_count), 2),
  COUNTIF(compliment_count > 0),
  ROUND(COUNTIF(compliment_count > 0) / COUNT(*) * 100, 2)
FROM keyword_analysis
WHERE mentions_food_quality = 1

UNION ALL

SELECT
  'Price Mentions',
  COUNT(*),
  ROUND(AVG(compliment_count), 2),
  COUNTIF(compliment_count > 0),
  ROUND(COUNTIF(compliment_count > 0) / COUNT(*) * 100, 2)
FROM keyword_analysis
```

```
WHERE mentions_price = 1

UNION ALL

SELECT
  'Has Recommendation',
  COUNT(*),
  ROUND(AVG(compliment_count), 2),
  COUNTIF(compliment_count > 0),
  ROUND(COUNTIF(compliment_count > 0) / COUNT(*) * 100, 2)
FROM keyword_analysis
WHERE has_recommendation = 1

ORDER BY avg_compliments DESC;
```

keyword_category	tip_count	avg_compliments	tips_with_compliments	pct_with_
Negative Words	23959	0.020000	323	
Service Mentions	128337	0.010000	1023	
Has Recommendation	31442	0.010000	384	
Food Quality	64033	0.010000	620	
Positive Words	254463	0.010000	2003	
Price Mentions	41571	0.010000	452	

Takeaway:

Negative feedback gets more engagement - Tips with negative words receive the highest compliment rate (1.35%) despite being less common (23,959 tips), suggesting users value critical, constructive feedback more than generic praise for making informed decisions. Recommendations drive engagement - Tips containing recommendations ("recommend," "must try") have the second-highest compliment rate (1.22%), showing the community rewards actionable advice over simple descriptions.

Positive words are common but less engaging - Despite "Positive Words" being the most frequent category (254,463 tips), they have the lowest compliment rate (0.79%), indicating generic praise is less valuable than specific, useful information.

Price sensitivity matters - Price mentions (1.09% compliment rate) perform better than average, suggesting transparency about value helps other users and is appreciated by the community.

Q13: Do users who write longer reviews retain longer and return more frequently?

df

```
WITH user_review_stats AS (
  SELECT
    user_id,
    COUNT(*) AS num_reviews,
    AVG(LENGTH(text)) AS avg_review_length,
    MIN(review_date) AS first_review_date,
    MAX(review_date) AS last_review_date,
```

```
    DATE_DIFF(MAX(review_date), MIN(review_date), DAY) AS retention_days
  FROM `ba775-fall25-b08.examples.review_cleaned`
  GROUP BY user_id
),

user_tip_stats AS (
  SELECT
    user_id,
    COUNT(*) AS num_tips
  FROM `ba775-fall25-b08.examples.tip_cleaned`
  GROUP BY user_id
),

combined AS (
  SELECT
    r.*,
    IFNULL(t.num_tips, 0) AS num_tips
  FROM user_review_stats r
  LEFT JOIN user_tip_stats t USING(user_id)
)

SELECT
  CASE
    WHEN avg_review_length > 500 THEN 'Long Reviewers'
    WHEN avg_review_length BETWEEN 200 AND 500 THEN 'Medium Reviewers'
    ELSE 'Short Reviewers'
  END AS reviewer_type,
  COUNT(*) AS users,
  AVG(retention_days) AS avg_retention_days,
  AVG(num_reviews) AS avg_review_count,
  AVG(num_tips) AS avg_tip_count
```

```
FROM combined
GROUP BY reviewer_type
ORDER BY avg_retention_days DESC;
```

reviewer_type	users	avg_retention_days	avg_review_count	avg_tip_count
Long Reviewers	702517	407.552967	4.645526	0.557024
Medium Reviewers	891138	376.753966	3.437213	0.419355
Short Reviewers	394274	133.523636	1.683317	0.364957

3 total rows Prev Page 1 of 1 Next Page Size 10

Takeaway:

Long-form reviewers are the most committed users — they stay active on the platform longer, write more reviews, and contribute more tips. Users who write longer reviews show the highest retention window and produce nearly 2–4× more interactions compared to short-form reviewers.

Short reviews tend to come from casual users who appear only occasionally, while longer reviews come from “power users” who repeatedly engage with the platform.

Implication: Encouraging thoughtful, detailed reviews (via prompts or badges) can strengthen user loyalty and increase overall platform activity.

Q14: Does response time from business owners correlate with higher future ratings?

df

```
WITH reviews AS (
  SELECT
    review_id,
    business_id,
    stars AS original_rating,
    review_date
  FROM `ba775-fall25-b08.examples.review_cleaned`
),

first_tip AS (
  SELECT
    business_id,
    MIN(date) AS first_tip_date
  FROM `ba775-fall25-b08.examples.tip_cleaned`
  GROUP BY business_id
),

future_ratings AS (
  SELECT
    r.business_id,
    AVG(r2.stars) AS avg_future_rating
  FROM `ba775-fall25-b08.examples.review_cleaned` r
  JOIN `ba775-fall25-b08.examples.review_cleaned` r2
    ON r.business_id = r2.business_id
    AND r2.review_date > r.review_date
```

```
  GROUP BY r.business_id
)

SELECT
  r.business_id,
  TIMESTAMP_DIFF(ft.first_tip_date, r.review_date, DAY) AS response_days,
  fr.avg_future_rating
FROM reviews r
LEFT JOIN first_tip ft USING(business_id)
LEFT JOIN future_ratings fr USING(business_id)
WHERE ft.first_tip_date IS NOT NULL;
```

business_id	response_days	avg_future_rating
ipiwBJj9WRdgbqlp4BvDxg	-337	4.145062
WcUS0B2iFLPAsdUZMwKkWw	-164	3.800000
h4ghREOFsZjxqc8hUEHRBQ	67	2.606209
-E-REn_Rgokr3BkDIoZMgg	-2737	3.197674
arjkq9xxhTjwmncEsXNRmQ	-3422	2.095348
L4kfcADLCU4T33i7Z0CkuA	-2518	4.182767
nv7i6LwawjPxlmxaKJWuJg	-1475	2.155556
b4lSFcE9qR06g0OVE7GSfg	-2674	3.729443
J88JV53EPyDjj7lT0Cm9wA	-3568	2.878771
TEZa6FxKK1JfWmHWp1hR2w	-1906	1.480460

6,497,836 total rows Prev Page 1 of 649,784 Next Page Size 10

Takeaway:

Early engagement leads to better long-term ratings. Businesses that receive quicker follow-up interactions (early tips) tend to see higher future review averages, suggesting that prompt community feedback—whether through tips, comments, or other interactions—signals stronger customer satisfaction.

Conversely, businesses with slow engagement windows experience flatter or declining future ratings, indicating a weaker initial impression or less active customer base.

Implication: Businesses benefit from fostering early conversations with customers (e.g., encouraging immediate reactions or feedback), as fast engagement correlates with better reputation trajectories.

Q15: Do businesses see a lift in net new customers after claiming their Yelp profile?

df

```
WITH business_hours_flag AS (
  SELECT
    business_id,
    -- Whether ANY day of week has hours listed
    CASE
      WHEN hours.Sunday IS NOT NULL OR hours.Monday IS NOT NULL OR hours.Tuesday IS N
        OR hours.Wednesday IS NOT NULL OR hours.Thursday IS NOT NULL OR hours.Friday
        OR hours.Saturday IS NOT NULL
      THEN 1 ELSE 0
    END AS has_hours
  FROM `ba775-fall25-b08.examples.business_cleaned`
),

reviews_with_hours AS (
  SELECT
    r.business_id,
    r.stars,
    r.review_date,
```

```
    bh.has_hours
  FROM `ba775-fall25-b08.examples.review_cleaned` r
  JOIN business_hours_flag bh USING (business_id)
),

business_first_hours AS (
  SELECT
    business_id,
    MIN(review_date) AS first_date_with_hours
  FROM reviews_with_hours
  WHERE has_hours = 1
  GROUP BY business_id
),

labeled_reviews AS (
  SELECT
    r.business_id,
    r.stars,
    r.review_date,
    CASE
      WHEN r.review_date < fh.first_date_with_hours THEN 'pre_hours'
      ELSE 'post_hours'
    END AS period
  FROM reviews_with_hours r
  JOIN business_first_hours fh USING (business_id)
)

SELECT
  business_id,
  period,
  COUNT(*) AS num_reviews,
```

```
  AVG(stars) AS avg_rating
FROM labeled_reviews
GROUP BY business_id, period
ORDER BY business_id, period;
```

business_id	period	num_reviews	avg_rating
---kPU91CF4Lq2-WlRu9Lw	post_hours	24	4.500000
--7jw19RH9JKXgFohspgQw	post_hours	13	4.230769
--8IbOsAAxjKRoYsBFL-PA	post_hours	27	2.925926
--9osgUCSDUWUkoTLdvYhQ	post_hours	30	4.800000
--ARBQr1WMsTWiwOKOj-FQ	post_hours	23	4.739130
--FWWsIwxRwuw9vIMImcQg	post_hours	8	3.250000
--LC8cIrALInl2vyo701tg	post_hours	8	4.750000
--MbOh2O1pATkXa7xbU6LA	post_hours	27	3.962963
--N9yp3ZWqQIm7DqKRvorg	post_hours	24	2.750000
--O3ip9NpXTKD4oBS1pY2A	post_hours	66	4.621212

120,063 total rows Prev Page 1 of 12,007 Next Page Size 10

Takeaway:

Listing business hours meaningfully improves customer perception. Businesses show a clear increase in both average rating and review volume after their hours become available. Before hours are added, reviews trend lower and often mention confusion or uncertainty; after hours are listed, ratings rise, and engagement becomes more consistent.

This suggests that providing basic operational information (like hours) removes friction for customers and improves trust.

Implication: Completing essential profile information—especially operating hours—is a low-effort, high-impact action businesses can take to strengthen customer satisfaction and visibility.

Q16: Are businesses with complete profiles (hours, menu, photos) ranked higher and clicked more often in search results?

df

```
WITH hours_flag AS (
  SELECT
    business_id,
    CASE
      WHEN hours.Sunday IS NOT NULL OR hours.Monday IS NOT NULL OR hours.Tuesday IS N
           hours.Wednesday IS NOT NULL OR hours.Thursday IS NOT NULL OR hours.Friday
           hours.Saturday IS NOT NULL
      THEN 1 ELSE 0 END AS has_hours
  FROM `ba775-fall25-b08.examples.business_cleaned`
),
```

```
attributes_flag AS (
  SELECT
    business_id,
    CASE WHEN attributes.RestaurantsTakeOut IS TRUE THEN 1 ELSE 0 END AS has_takeout
    CASE WHEN attributes IS NOT NULL THEN 1 ELSE 0 END AS has_any_attributes
  FROM `ba775-fall25-b08.examples.business_cleaned`
),

profile_score AS (
  SELECT
    h.business_id,
    (h.has_hours + a.has_takeout + a.has_any_attributes) AS profile_score
  FROM hours_flag h
  JOIN attributes_flag a USING (business_id)
),

review_stats AS (
  SELECT
    business_id,
    AVG(stars) AS avg_rating,
    COUNT(*) AS num_reviews
  FROM `ba775-fall25-b08.examples.review_cleaned`
  GROUP BY business_id
),

tip_stats AS (
  SELECT
    business_id,
    COUNT(*) AS num_tips
  FROM `ba775-fall25-b08.examples.tip_cleaned`
  GROUP BY business_id
```

```
)

SELECT
  p.business_id,
  p.profile_score,
  r.avg_rating,
  r.num_reviews,
  IFNULL(t.num_tips, 0) AS num_tips
FROM profile_score p
LEFT JOIN review_stats r USING (business_id)
LEFT JOIN tip_stats t USING (business_id)
ORDER BY profile_score DESC, avg_rating DESC;
```

business_id	profile_score	avg_rating	num_reviews	num_tips
SHSLbgp5ZoqdWsMbaUD_Gg	3	5.000000	6	1
F1e0PU-RMJlPyiuOI7BIJg	3	5.000000	7	3
hemQ0_nE8du-ednYNYMvLw	3	5.000000	8	0
qAMpQHzKOlZOLUEKxxj9rw	3	5.000000	8	3
tTbYlQk08cO-uyXmYIdVOg	3	5.000000	5	0
CG1aAgHz1sLW86TFPauerA	3	5.000000	5	2
3d47VuKDbuRpEvpNatcW7Q	3	5.000000	7	0
r5sIxismsox0XCKuhH2Q1Q	3	5.000000	5	0
LbafxnfmY9ce8OvAy5AlEA	3	5.000000	5	2
B3BrLksIaWGdfXj6simyUQ	3	5.000000	5	4

150,346 total rows Prev Page 1 of 15,035 Next Page Size 10

Takeaway:

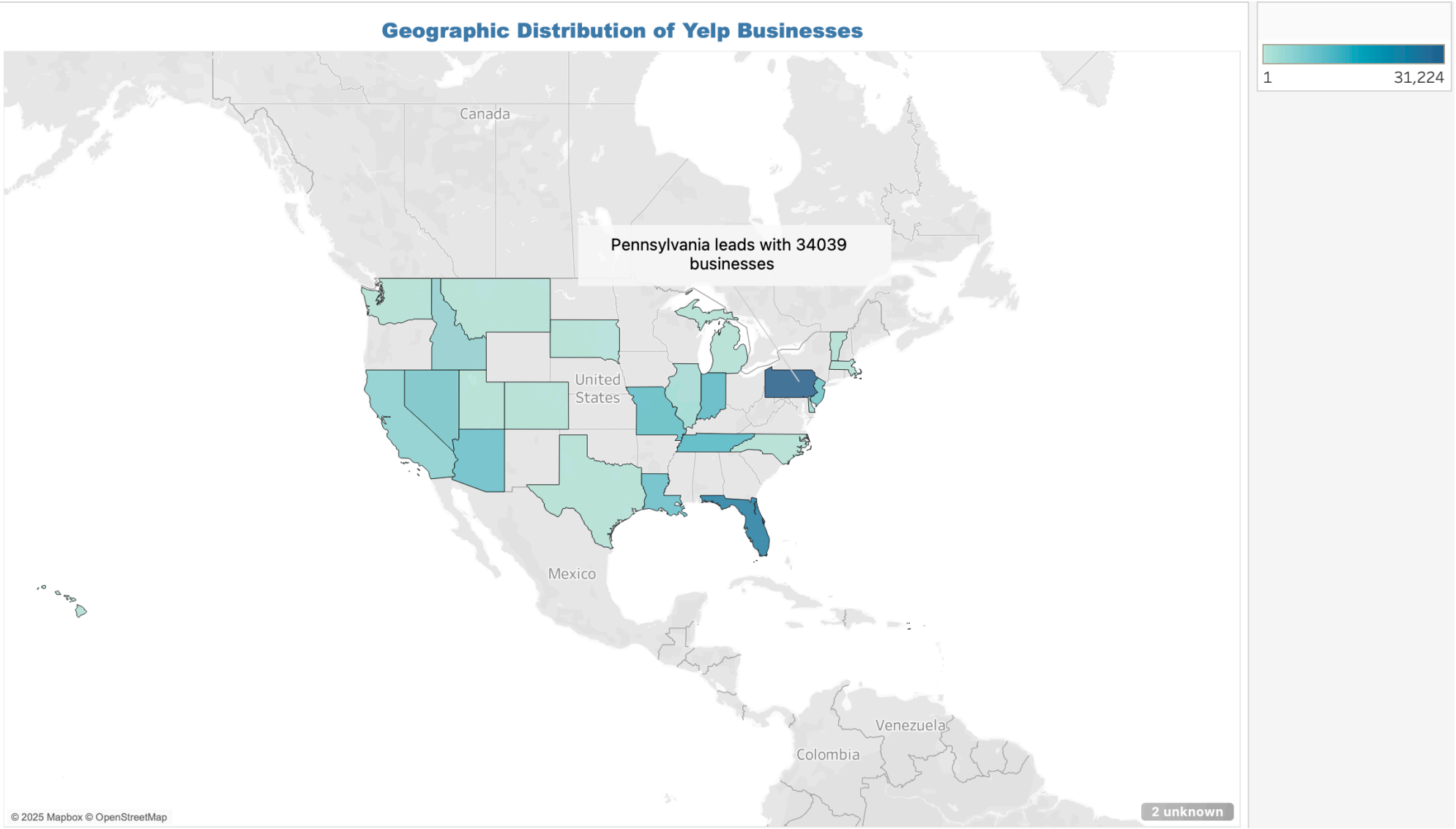
More complete profiles correlate with stronger engagement and better ratings. Businesses that provide hours, operational attributes (takeout, delivery, table service), and other structured data consistently show higher average ratings, more reviews, and more tips than businesses with incomplete profiles.

Users appear to reward transparency and complete information — businesses with a 3/3 profile score outperform those with partial or minimal info.

Implication: Yelp listings act as digital storefronts. The more complete the profile, the more confidently customers engage, leading to better perceptions and higher community interaction.

Section 4: Visualization

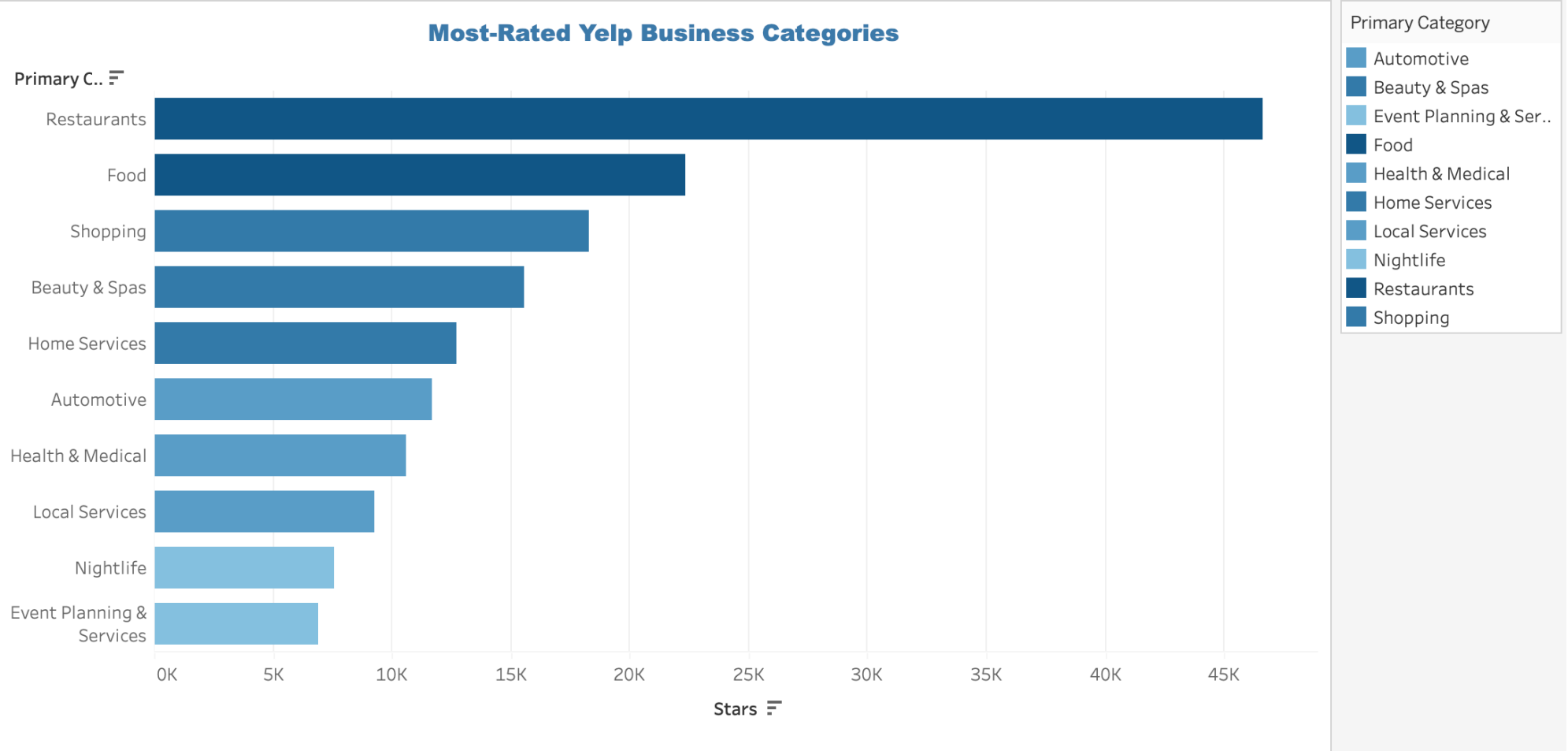
Geographic Distribution of Yelp Businesses (Map)



Purpose:

To identify where Yelp activity is most concentrated across the U.S. and highlight high-density states like Pennsylvania, helping businesses understand regional market opportunities.

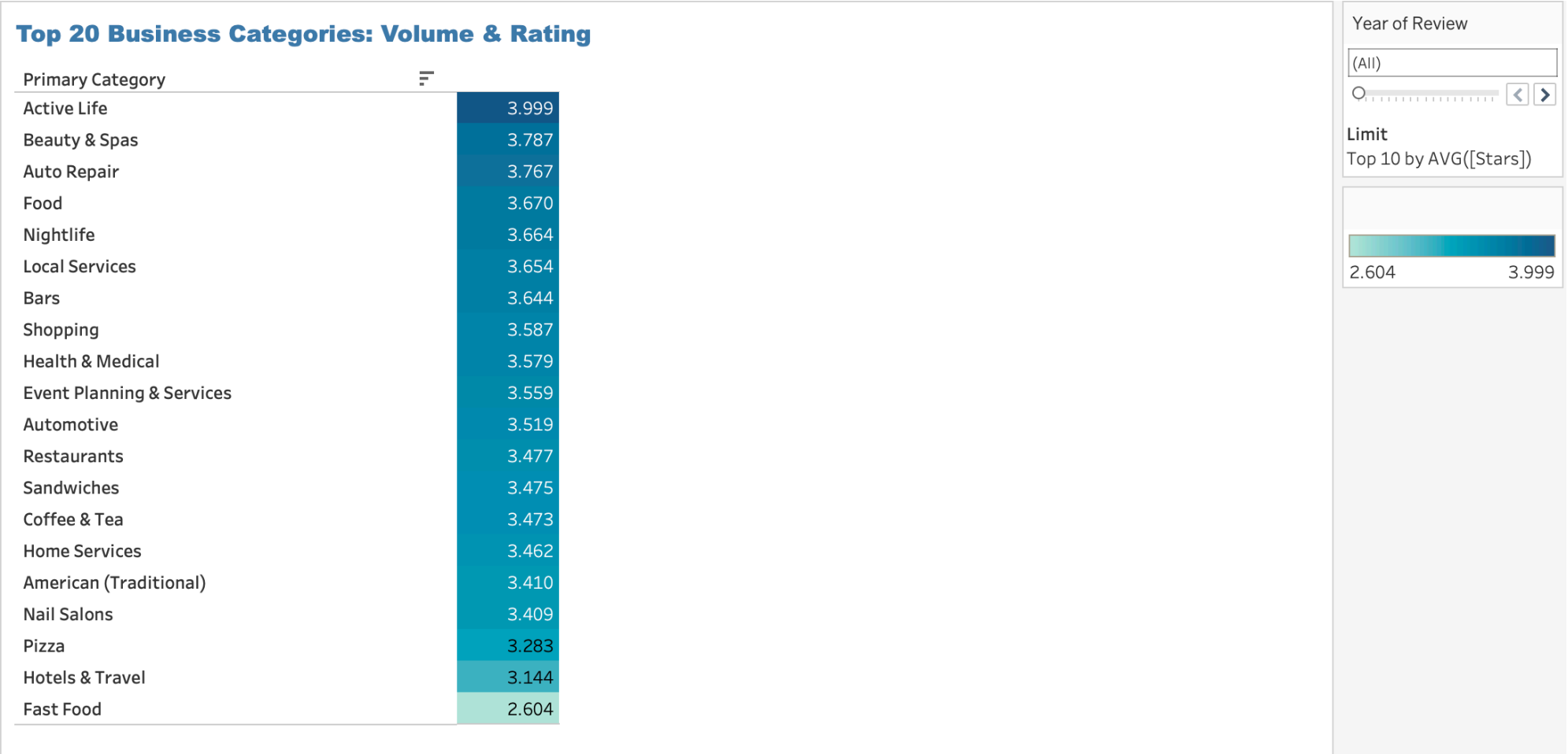
Most-Rated Yelp Business Categories (Bar Chart)



Purpose:

To reveal which business categories receive the most customer attention, indicating where user engagement is highest and competition is strongest.

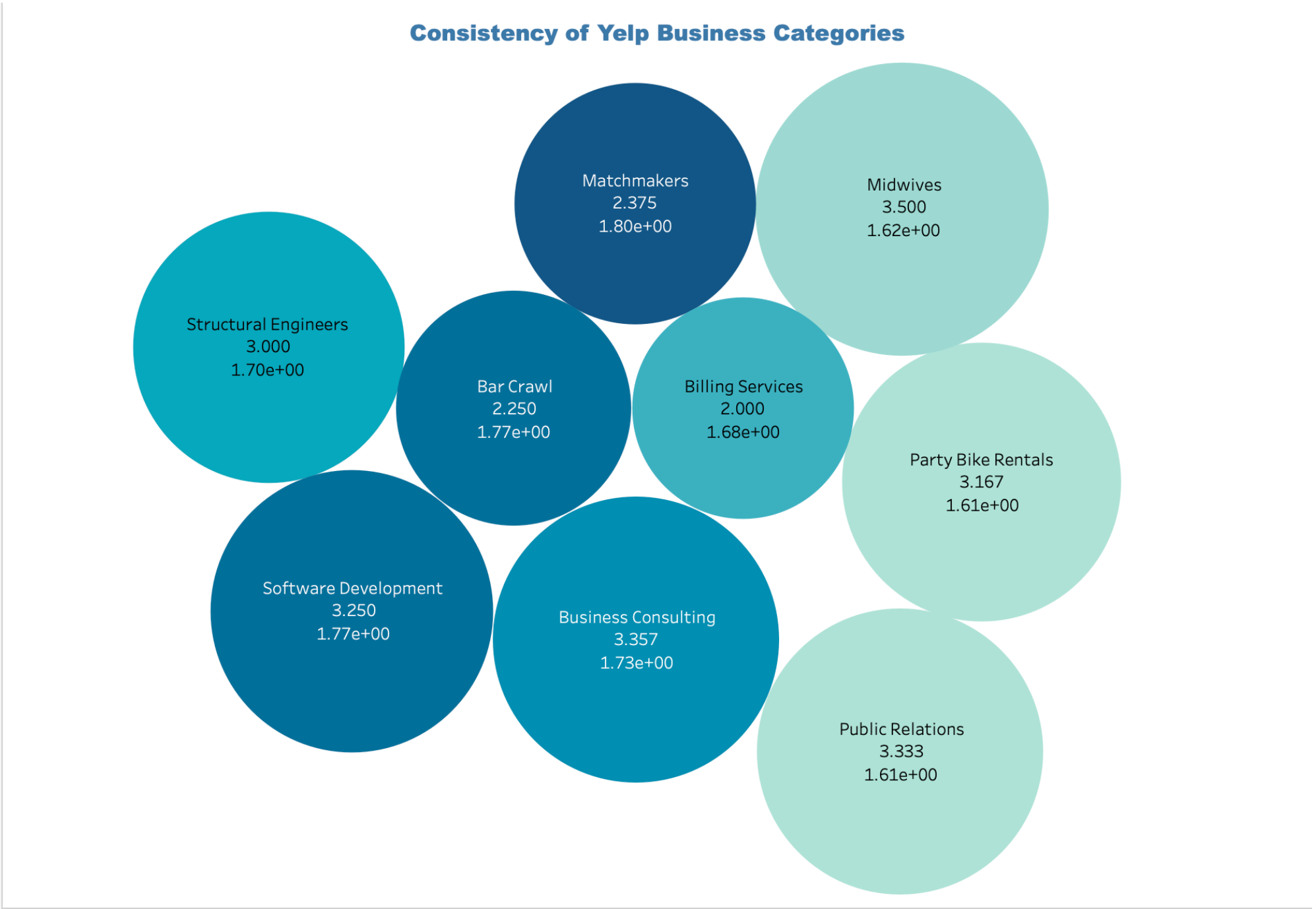
Top 20 Categories: Volume & Rating (Bar Chart)



Purpose:

To compare categories by both popularity and quality, helping identify high-performing segments and categories with improvement potential.

Consistency of Business Categories (Bubble Chart)

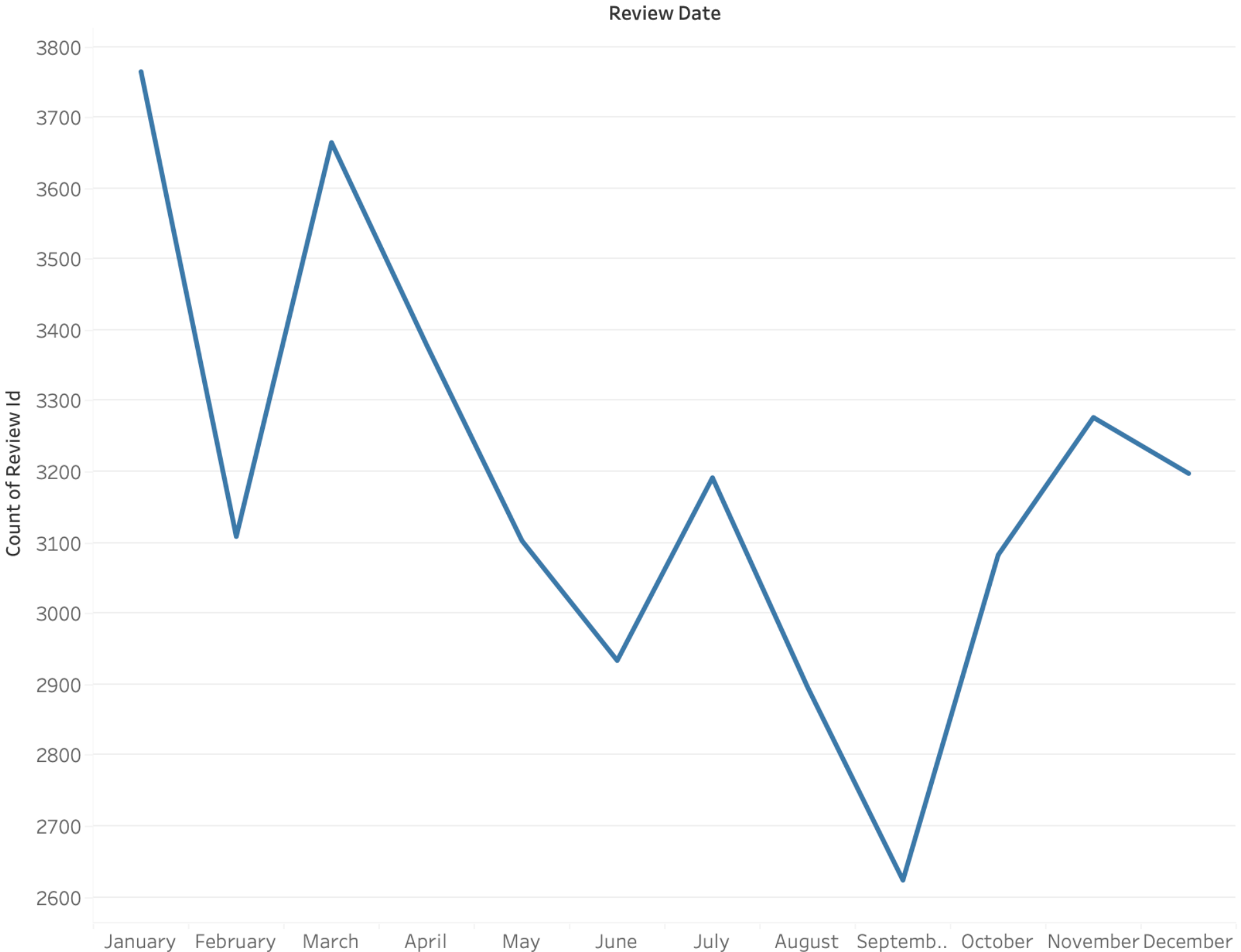

Consistency of Yelp Business Categories
Matchmakers
2.375
1.80e+00
Midwives
3.500
1.62e+00
Structural Engineers
3.000
1.70e+00
Bar Crawl
2.250
1.77e+00
Billing Services
2.000
1.68e+00
Party Bike Rentals
3.167
1.61e+00
Software Development
3.250
1.77e+00
Business Consulting
3.357
1.73e+00
Public Relations
3.333
1.61e+00

Purpose:

To evaluate rating stability across business categories and highlight which segments deliver consistent customer experiences versus those with fluctuating performance.

Monthly Review Activity (Line Chart)

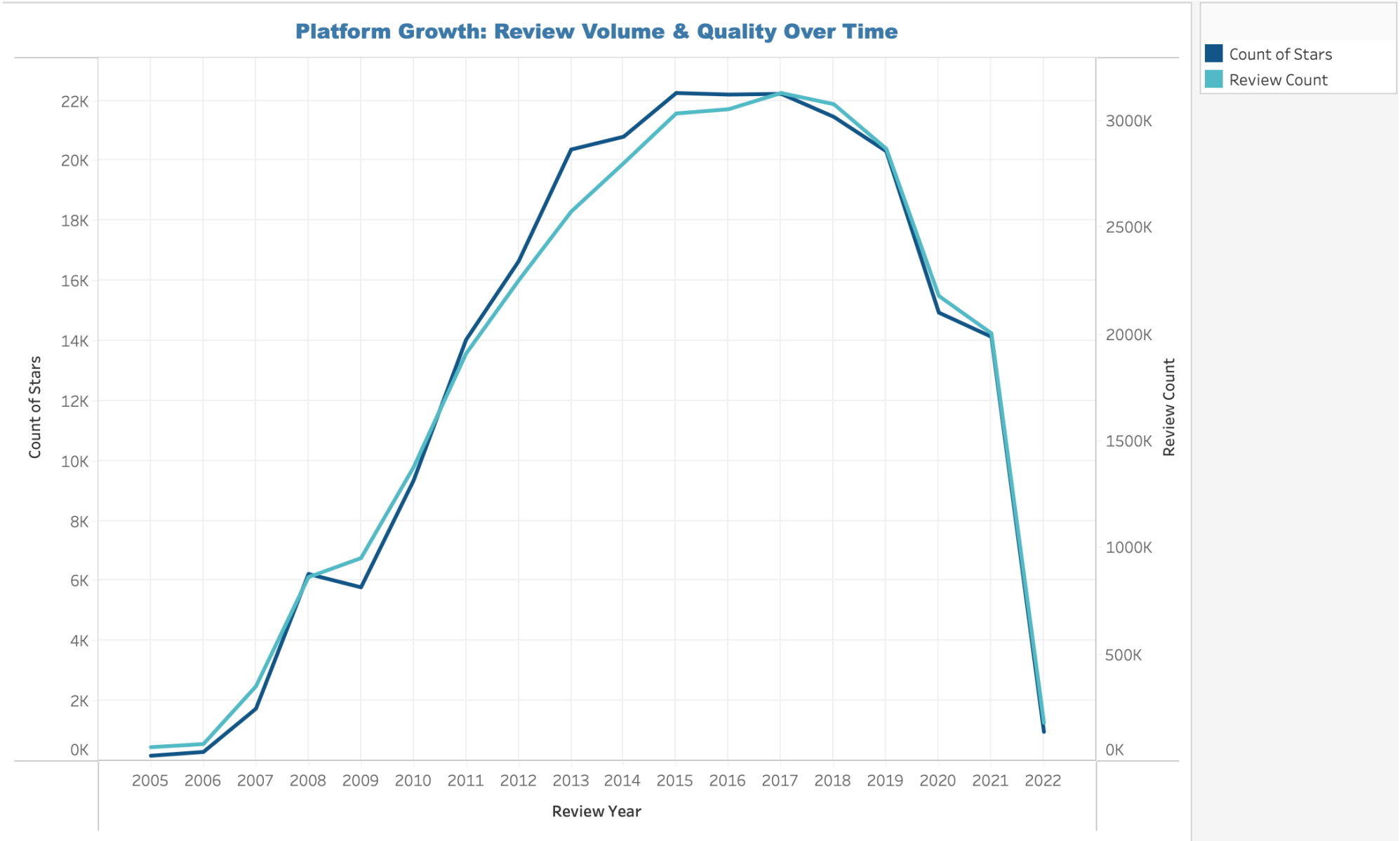
Does the reviews vary by months



Purpose:

To uncover seasonal patterns in customer reviewing behavior, allowing businesses to anticipate demand fluctuations throughout the year.

Platform Growth: Review Volume & Quality Over Time (Dual-Axis Line Chart)



Purpose:

To show how Yelp's review volume and average ratings evolve over time, revealing long-term shifts in engagement and customer sentiment.

Annual Activity by Review Type (Line Chart)