

Analytics on YELP

Data Technology Solutions





*“In God we trust. All
others must bring data.”*

Professor W. Edwards Deming

Engineer, Statistician

Developed the sampling techniques



GRATITUDE

We like to thank our teacher Mr. Bhavik Gandhi for sharing his knowledge and ideas. We are fortunate to have him as our teacher for this course

VP, Product and Engineering, HaikuJAM
Data Science Manager, Shopify
Associate Vice President, Service & Analytics, Pepperfry
Program Manager and Software Development Engineer, Microsoft

"Are marriages made in heaven or do algos decide who we love?"

Talk about 3 different use cases of data science at Shaadi.com
<https://www.youtube.com/watch?v=sJIfHYDFtMk>

Midpoint OBJECTIVES

Objective 1

The objective of this project is to help Yelp through analyzing the Yelp dataset and gain insights into various aspects of businesses, users, and reviews.

Objective 2

The analysis aims to extract valuable information such as top-rated businesses, user behaviors, popular business categories, geographic trends, and more.

Objective 3

The project aims to use data exploration, visualization, and statistical analysis techniques to uncover patterns and trends within the Yelp dataset.



WHAT IS YELP

Yelp is a user-generated review platform that operates in the local business review and recommendations industry. Its primary focus is on providing a platform for users to discover and review local businesses, while also offering businesses an opportunity to engage with their customers and potentially attract new ones.

HOW YELP WORKS

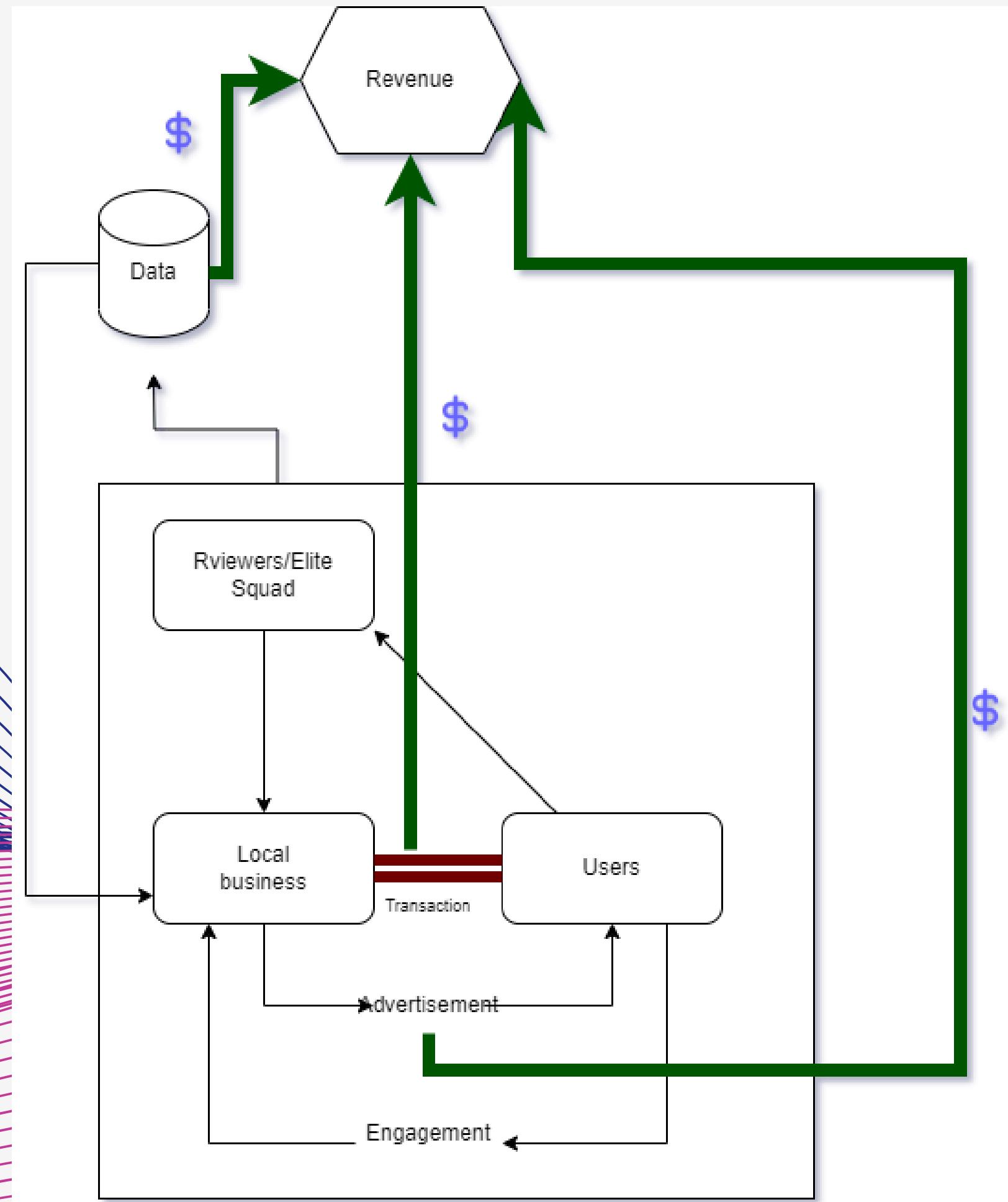


- **User-Generated Content:** The core of Yelp's platform is driven by user-generated reviews and recommendations, providing valuable insights for other users to make informed decisions.
- **Advertising Revenue Model:** Yelp's significant revenue stream comes from businesses paying for advertising services, increasing their visibility on the platform.
- **Mobile App Optimization:** Yelp's mobile app plays a crucial role in its success, enabling users to access reviews and discover businesses while on the go, leveraging location-based services for real-time recommendations.
- **Community Engagement:** Yelp fosters an active community of reviewers, known as "Yelpers," contributing to the platform's credibility and attracting new users through genuine user experiences.

UNDERLYING REASONS FOR YELP SUPREMACY

- Large and Active Community: One of the primary reasons to choose Yelp is its large and active user community. With millions of users leaving genuine reviews and ratings, businesses can tap into a vast audience and increase their visibility.
- Authentic and Trustworthy Reviews: Yelp's user-generated reviews are authentic and unbiased, giving businesses and users confidence in the credibility of the feedback. This authenticity helps users make informed decisions about where to spend their time and money.
- All-in-One Business Information: Yelp provides comprehensive business information, including contact details, photos, menus, and operating hours. Having all this data in one place makes it convenient for users to find everything they need about a business quickly.
- User-Friendly Mobile App: Yelp's mobile app is easy to use and feature-rich, making it a preferred choice for users searching for local businesses on the go. The app's convenience ensures that users can access reliable information wherever they are.
- Refined Search Filters: Yelp's advanced search filters allow users to narrow down their search results based on specific criteria, such as location, category, price range, and ratings. This feature helps users find businesses that suit their preferences.

Business Process MAPPING



BUSINESS MODEL CANVAS



Key Partners



- Content Contributors (Active reviewers)
- Partnerships
- Extended Partners

Key Activities



- Network Effects
- Customer Experience
- Ad Selling
- Platform Growth
- Grow local communities
- Raise brand awareness
- Manage day to day operation

Key Resources



- Employees
- High-Quality Content
- Brand Reputation
- Engaged Communities
- Partnerships

Value Propositions



local Businesses:

- exposure and visibility
- Increase traffic
- Reviews to promote businesses
- Targeted advertisements

Users:

- browse local businesses
- discover new places
- save time and money

Content Contributors:

- Engage in fun social interaction
- Support the community and local businesses

Customer Relationships



- Self Service Platform
- Brand Reputation
- Engagement
- Dedicated Sales
- Engaged Community
- Communication

Channels



- website or app
- Ads channel
- webpages for the business owner and developers
- Social media pages

Customer Segments



- Local Businesses
- Users
- Content contributors (Active reviewers)

Cost Structure



- Cost of Revenue (6%)
- Sales and Marketing (51%)
- Product Development (23%)
- General and Administrative (13%)
- Depreciation (5%)

Revenue Streams



- Advertisement (97% of revenue)
- Transaction revenue (1% of revenue)
- Other Services (2% of revenue)



Business Strategy Hub

Business REGISTRATION

Before registration process there are three requirements that are to be fulfilled:

- **Identify business structure:** There are several types of business structures like sole, partnership, each with their own costs, risks, and tax implications.
- **Select the business locations:** The registration process for businesses varies from state to state, and so do the business taxes must pay. As such, selecting business location is a significant decision to make.
- **Pick the business name:** A business name that want to stand out from the crowd. A creative, descriptive, or fun business name can make a lasting impression on potential customers.



How Yelp MAKES BUSINESS FROM BUSINESSES

01

Businesses can offer deals and gift certificates through Yelp's platform. These promotions attract customers and drive sales, with Yelp taking a percentage of the revenue generated from these offerings.

02

Yelp offers a reservation and waitlist management system for restaurants. Participating restaurants pay a fee for using these services, which help streamline the reservation process for customers and generate revenue for Yelp.

05

Yelp offers a reservation and waitlist management system for restaurants. Participating restaurants pay a fee for using these services, which help streamline the reservation process for customers and generate revenue for Yelp.

03

Businesses can pay for sponsored placements on Yelp's platform to increase visibility and reach a larger audience.

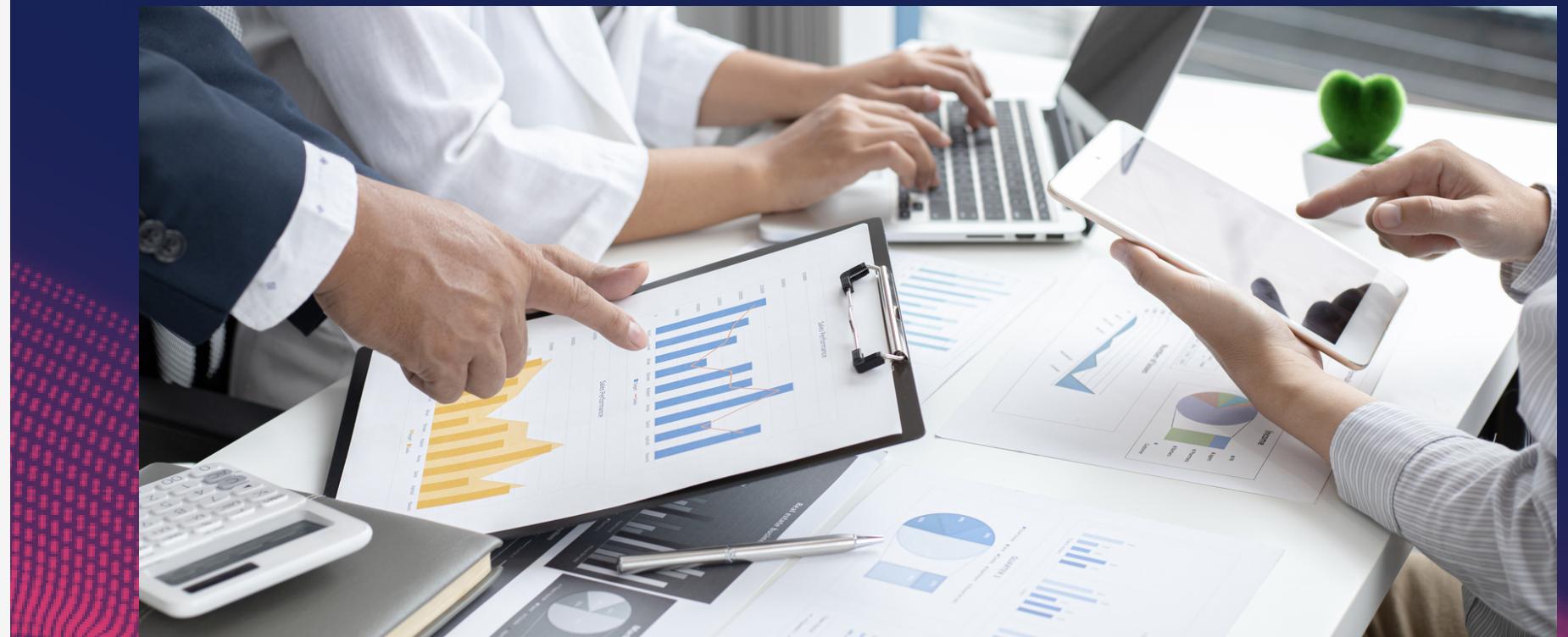
04

Businesses can use Yelp Connect to share updates with their followers, and Yelp may charge for this feature.

Opportunities We Are Addressing

VALUE PROPOSITION

- **Business Optimization:** The actionable insights derived from our analytics project enable Yelp to identify inefficiencies and bottlenecks within their processes, allowing them to optimize operations and improve overall productivity.
- **Customer Behavior Analysis:** Through the analysis of customer data, we can provide actionable insights on consumer behavior, preferences, and pain points, enabling Yelp to tailor their products and services to better meet customer needs.
- **Product Innovation:** Our analytics project highlights areas for potential product improvements and innovation.
- **Risk Mitigation:** Actionable insights from our analytics efforts help the organization to proactively identify potential risks and vulnerabilities, allowing them to implement strategies to mitigate these risks and ensure long-term stability and success.



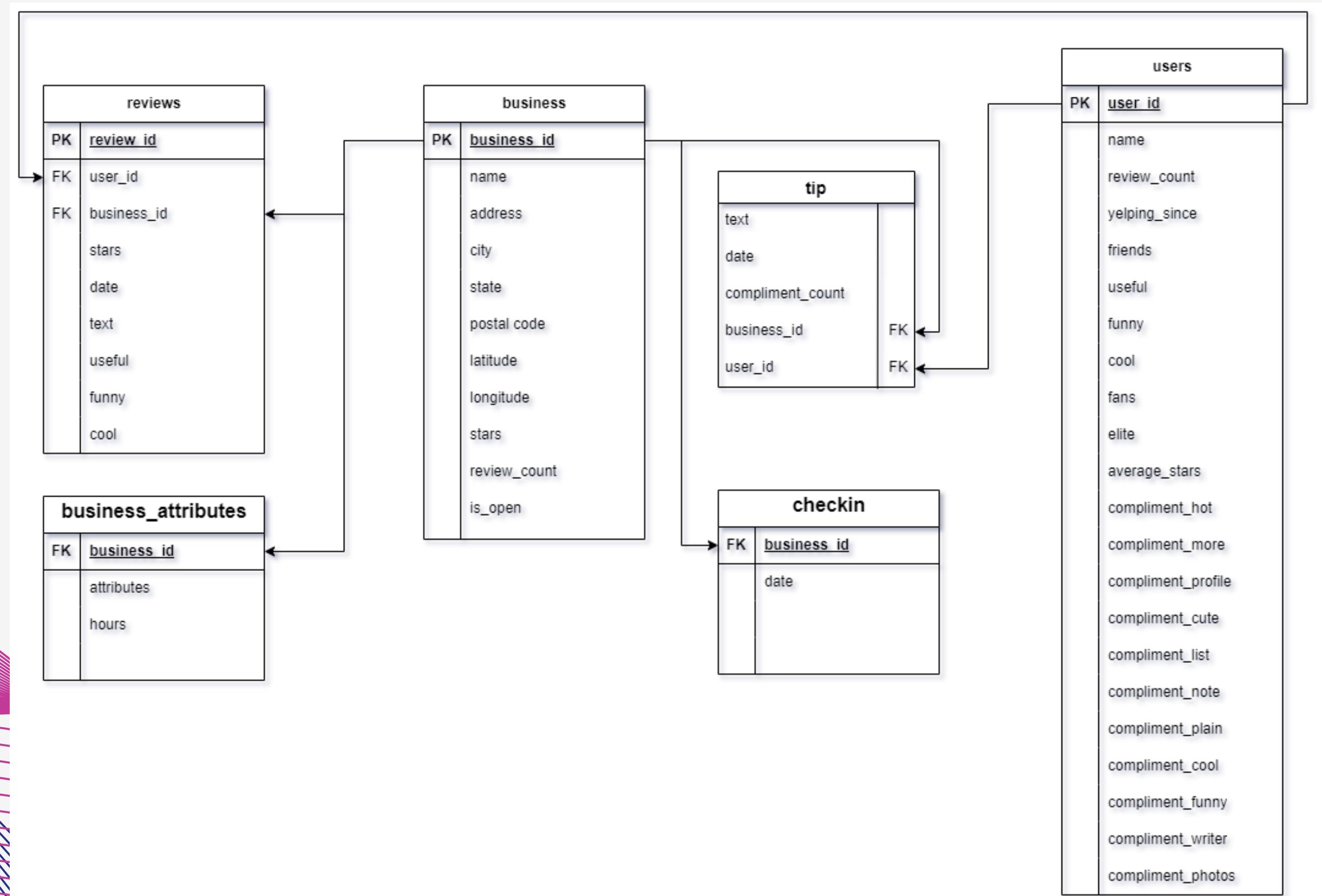
Data Model & Analytics

Data SOURCE

- The data source for this project is the Yelp dataset, which is a collection of data related to businesses, users, and reviews from the Yelp platform.
- The dataset includes information about businesses such as their names, categories, locations, and star ratings. It also contains user profiles with details like user IDs, names, review counts, and average star ratings. Additionally, the dataset includes individual reviews that users have left for businesses, along with associated star ratings.
- Link : <https://www.yelp.com/dataset>
- Size : 9.29 GB



Data Model



PROJECT CHALLENGES



LARGE DATASET

The Yelp dataset is quite large, and handling big data efficiently is crucial. We tackled this challenge by using pandas for reading and cleaning the data and SQLite database for storing the data and run SQL queries.

DATA CLEANING

Datasets may contain missing or inconsistent data, requiring careful data cleaning and preprocessing. We used pandas to handle missing values and ensure data quality.

JSON FILES

All of the original files were in JSON formats and those were not parsed by Postgres and Bigquery. We moved to SQLite through Python to push the data to the database and running queries

PROCESSING AND STORAGE

Due to our processing power limitations for this big database, we shifted to a cloud notebook. Fortunately, Kaggle hosted the whole database in their memory and we leveraged their storage and processing power

Insights on Data

Yelp Database Analytics

File Edit View Run Add-ons Help

+ Run All Code

0990280 rows = 9 columns

```
#To Identify Most Reviewed Businesses
query = """
SELECT name, review_count, stars
FROM business
ORDER BY review_count DESC
LIMIT 5;
"""

result = pd.read_sql_query(query, conn)
result
```

	name	review_count	stars
1	Aging Oyster House	7560	4.5
2	Omeza Grill	3400	4.5
3	Hattie B's Hot Chicken - Nashville	6291	4.5
4	Reading Terminal Market	5271	4.5
5	Ruby Slipper - New Orleans	5190	4.5

+ Code + Markdown

This query exemplifies Yelp's ability to identify the most reviewed businesses, which can be crucial for businesses seeking to improve their reputation and customer engagement.

This is explained as follows :

1. Yelp's platform encourages user-generated reviews, resulting in businesses with high foot traffic accumulating numerous reviews, contributing to their visibility and popularity.
2. Yelp uses star ratings and sorting algorithms to display businesses with high average ratings and recent, engaging reviews at the top of search results.
3. Yelp's "Popular" and "Top" categories highlight businesses with substantial positive reviews, providing increased exposure and attracting potential customers.
4. Engaging with Yelp's community, including the Yelp Elite Squad, can further enhance a business's reputation and customer engagement on the platform.

The screenshot shows a Jupyter Notebook interface with the title "Yelp Database Analytics". The code cell contains the following Python code:

```
#To Calculate Average Ratings per Category
query = """
SELECT categories, AVG(stars) AS avg_rating
FROM business
GROUP BY categories
ORDER BY avg_rating DESC;
"""

result = pd.read_sql_query(query, conn)
result
```

The resulting DataFrame is displayed below the code cell:

	categories	avg_rating
0	Dance, Arts & Entertainment, Performing Arts, A...	5.0
1	Ziplining, Team Building Activities, Kids Activi...	5.0
2	Yoga, Trainers, Gyms, Fitness & Instruction, A...	5.0
3	Yoga, Trainers, Active Life, Health & Medical...	5.0
4	Yoga, Trainers, Active Life, Gyms, Fitness & I...	5.0
...
83156	Active Life, Amateur Sports Teams, Sports Clubs	1.0
83157	Accountants, Professional Services, Property M...	1.0
83158	Accessories, Fashion, Shopping, Women's Clothi...	1.0
83159	Accessories, Fashion, Shopping, Shopping Cente...	1.0
83160	Accessories, Fashion, Shopping, Bikes, Sportin...	1.0

83161 rows × 2 columns

At the bottom of the notebook, there are buttons for "+ Code" and "+ Markdown".

This query showcases Yelp's capacity to calculate and display the average rating for each business category, enabling users to identify the most popular and well-rated industries.

This is explained as follows:

1. Yelp calculates and displays the average rating for businesses in each specific industry or category based on user-generated reviews and star ratings.
2. The platform aggregates reviews for businesses within a category, allowing users to identify the most popular and well-rated industries.
3. Yelp's sorting and ranking algorithms prioritize higher-rated businesses within a category, making it easier for users to find the best-reviewed establishments.
4. Users can use filters and search options to refine their results and make informed decisions when choosing businesses, restaurants, or services in a particular category.

The screenshot shows a Jupyter Notebook interface with a dark theme. The top bar displays the title "Yelp Database | Kaggle" and the URL "kaggle.com/code/arunabharmann11/yelp-database/edit". Below the title, there are several tabs: "Lambton - Self Serve...", "Lambton - Data Service...", "Lambton - Services...", "mylambton", "Lambton - Moodle", "Spotify", and "GitHub".

The main area contains a code cell with the following content:

```
#Q1 Identify the city where most shops opened
query = """
SELECT city, COUNT(*) AS open_shop_count
FROM business
WHERE is_open = 1
GROUP BY city
ORDER BY open_shop_count DESC
LIMIT 10;
"""

result = pd.read_sql_query(query, conn)
result
```

Below the code cell is a table titled "city open_shop_count" showing the top 10 cities with the highest number of shop openings:

city	open_shop_count
Philadelphia	1000
Tucson	750
Tempe	700
Indianapolis	550
Bangkok	500
Denver	450
New Orleans	400
Edmonton	350
Saint Louis	300
Santa Barbara	250

At the bottom of the code cell, there are two buttons: "Code" and "Markdown".

By identifying the city where most shops were opened, Yelp can use this information to improve its services and offerings in several ways:

Targeted Marketing: Yelp can focus its marketing efforts on the city where most shops were opened. This includes targeted advertising campaigns, sponsored listings, and promotions to attract more users and businesses to the platform in that city.

Expansion Planning: If a particular city shows significant growth in the number of shops being opened, Yelp can prioritize its expansion efforts in that area. This could involve increasing resources, customer support, and local partnerships to enhance the Yelp experience in that city.

Localized Content: Yelp can curate and promote localized content, including featured businesses, trending spots, and user-generated content from the city with the highest number of shop openings. This can help engage users in that city and encourage more reviews and interactions.

Business Insights: By analyzing the data from shops that have been opened in the identified city, Yelp can gain insights into the types of businesses that are thriving in that area. This information can be valuable for business owners and potential entrepreneurs looking to start new ventures.

Overall, identifying the city where most shops were opened can provide Yelp with valuable insights to tailor its strategies, offerings, and services to cater to the specific needs and preferences of users and businesses in that area.

Yelp Database Open saved

File Edit View Run Add-ons Help

+ X Run All Code Draft Session Close

query: SELECT user_id, name, review_count FROM user ORDER BY review_count DESC LIMIT 5;

result = pd.read_sql_query(query, conn)

result

	user_id	name	review_count
1	H100eGZhvQf0HvWfZ71eaA	Pax	11473
2	He3uO-mPyAbI5HJwzASmA	Victor	94978
3	EWQDyvJxPLSeDfGcmtt	Bruce	76567
4	PGQgD9wBQoobUqzrA	Sara	72968
5	PAwULnqz-2t8KmMqQ	Kim	89471

+ Code + Markdown

(14):

Search

File Edit View Insert Tools Window Help

Yelp can use the insights gained from the top 5 users with the highest number of reviews on its platform to enhance various aspects of its service. Some are:

User Engagement: Yelp can create incentive programs or loyalty rewards for users who write a good number of reviews. This could encourage more users to actively engage with the platform and share their experiences.

Quality Control: One of the major problems Yelp faced was fake reviews. Yelp can analyze the reviews from top users to identify patterns of high-quality content. This analysis can be used to develop guidelines and standards for other users to follow, improving the overall quality of reviews on the platform.

Social Media Promotion: Yelp can collaborate with top users to promote their reviews and experiences on social media channels. This cross-promotion can increase user engagement and attract new users to the platform.

Yelp can strengthen its community, enhance the quality of user-generated content, and provide more valuable information to users and businesses alike. This can lead to increased user engagement, improved business visibility, and a more vibrant and thriving Yelp ecosystem.

The screenshot shows a Jupyter Notebook environment. At the top, there's a toolbar with various icons and a tab bar with several open notebooks. Below the toolbar is a menu bar with File, Edit, View, Run, Address, and Help. The main area contains a code cell and a results cell.

```
query = """
SELECT user_id, username AS username, COUNT(r.useful) AS total_useful_votes
FROM user u
JOIN reviews r ON u.user_id = r.user_id
GROUP BY user_id, username
ORDER BY total_useful_votes DESC
LIMIT 5;
"""

result = pd.read_sql(query, conn)
result
```

	user_id	username	total_useful_votes
1	4572174000000000000	Craig	14974
2	1770470797000000000	Michelle	11472
3	1770470797000000000	Michael	11293
4	1770470797000000000	Angela	11202
5	1770470797000000000	Christy	10794

At the bottom, there's a toolbar with various icons for file operations and a search bar.

Yelp can capitalize on the valuable feedback and influence of the top 5 users with the most useful votes received for their reviews to improve its platform and user experience.

Moderation and Quality Control: Reviews from top users with useful votes can be given higher visibility and priority during content moderation. Yelp can use their reviews as exemplars of high-quality content and use them to set standards for other users.

Community Building: Yelp can organize meetups or virtual events for the top users to connect and share their experiences. These events can foster a sense of community and encourage more engagement among influential users.

Getting Suggestions: Yelp can designate these top users providing them with a direct channel to share their feedback, concerns, and suggestions with Yelp's team. This can foster a sense of ownership and partnership between Yelp and its influential users.

Yelp can foster a stronger user community, improve content quality, and enhance the overall platform experience. This, in turn, can attract more users and businesses, making Yelp an even more valuable resource for both consumers and local enterprises.

The screenshot shows a Jupyter Notebook environment with several tabs at the top: 'Yelp Business', 'Yelp Dataset', 'Data Technical', 'Yelp Dataset', '2020-71_ED...', 'Yelp Revenue', and 'The Best 10'. The main area has a title 'Database Analytics' and a subtitle 'Draft saved'. A toolbar below includes 'Edit', 'View', 'Run', 'Add-ons', and 'Help'. The code cell contains a SQL query:

```
SELECT attributes FROM business_attributes JOIN business  
WHERE business_attributes.business_id = business.business_id AND business.rating > 5.0 AND business_attributes.  
result = pd.read_sql_query(query, conn)  
result
```

The output cell displays a table titled 'attributes' with 14 rows of data:

	attributes
0	{"ByAppointmentOnly": "True"}
1	{"ByAppointmentOnly": "True"}
2	{"BusinessParking": "garage", "lot", "street"}
3	{"BusinessAcceptsCreditCards": "True", "Wi-Fi": "True"}
4	{"BusinessAcceptsCreditCards": "True"}
...	...
14353	{"ByAppointmentOnly": "True"}
14354	{"BusinessAcceptsCreditCards": "True", "BusinessParking": "garage", "lot", "street"}
14355	{"GoodForDogs": "True"}
14356	{"ByAppointmentOnly": "True", "BusinessAcceptsCreditCards": "True"}
14357	{"BusinessParking": "garage", "lot", "street"}

A note at the bottom states: 'Kaggle uses cookies to deliver our services, analyze web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies.'

With the help of the above query Yelp can discover the amenities that have been top rated by customers and can suggest those amenities to other businesses which have low rating to pick up their businesses.

Preferred Amenities :

Attributes like "outdoor seating," "free Wi-Fi," and "good for groups" are popular among top-rated businesses.

Yelp can emphasize these amenities to encourage other businesses to adopt them and enhance customer satisfaction.

Pet-Friendly Environment :

Top-rated businesses often have attributes like "pet-friendly," indicating a pet-welcoming environment.

Yelp can promote pet-friendly businesses, catering to users who wish to bring their pets along.

Exceptional Customer Service :

Attributes such as "friendly staff," "excellent service," and "knowledgeable staff" are associated with top-rated businesses.

Yelp can highlight the importance of outstanding customer service for businesses seeking higher ratings.

Unique and Attractive Ambiance :

Top-rated businesses are frequently praised for their "cozy ambiance," "stylish decor," and "unique atmosphere."

Yelp can encourage businesses to create distinct atmospheres to stand out and attract positive reviews.

The screenshot shows a Jupyter Notebook interface with several tabs at the top: 'Yelp Business', 'Yelp Dataset', 'Data Notebook', 'Yelp Database', '2020-11_0C', 'Yelp Revenue', and 'THE BEST 10'. The main area contains the following code:

```
query = """
SELECT
    strftime("%H", date) AS review_hour,
    COUNT(review_id) AS review_count
FROM
    reviews
GROUP BY
    review_hour
ORDER BY
    review_count DESC;
"""

result = pd.read_sql_query(query, conn)
result
```

A table titled 'result' is displayed below the code, showing the following data:

	review_hour	review_count
0	14	49433
1	19	46377
2	00	45376
3	01	452189
4	02	452124
5	23	448409
6	17	447387
7	03	447387

At the bottom of the notebook, there is a note: 'on Kaggle to deliver our services, analyse web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies.'

with the query Yelp can Analyze the peak review times which can help Yelp identify when users are most active, allowing them to optimize the platform's content, push notifications, or marketing efforts during these hours to increase user engagement.

Busy Hours

The query reveals peak review times during specific hours, such as 6 PM and 8 PM, suggesting high user activity during these periods.

Yelp can optimize content, push notifications, and marketing efforts during these busy hours to engage users effectively.

Lunchtime Reviews

Peak review times around 12 PM indicate that users are actively posting reviews during lunchtime.

Yelp can use this information to promote lunch deals and specials to attract users during these hours.

Evening Engagement

Reviews surge around 7 PM, indicating higher user engagement during the evening hours.

Yelp can consider featuring businesses with evening events or happy hours to capture this engaged audience.

kaggle.com/khanalabs/yelp-dataset/edit

Yelp Dataset Draft saved

File Edit View Run Add-ons Help

+ X

```
#To find the top 5 businesses with stars equal to 1.0 and the highest review count, along with their names and cities.
```

```
query = """
SELECT business_id, name, city, stars, review_count
FROM business
WHERE stars = 1.0
ORDER BY review_count DESC
LIMIT 5;"""

result = pd.read_sql_query(query, conn)
result
```

	business_id	name	city	stars	review_count
0	0NuqbnSeuUWumgjU_4DQ	Sears Home Services	Fenton	1.0	579
1	xDLh8Igh1nLxZm7wYQBA	Defender Security Company	Indianapolis	1.0	413
2	WNgzr5V9mzPjHgjKA	Frontier Communications	Tempe	1.0	397
3	VqgVB9mDmE2NEKOSnxA	Center	Ewing	1.0	394
4	dpdDk4hOGxGvTBjRQEm	Express Scripts	Maryland Heights	1.0	306

Type here to search

In this query we find the businesses with the minimum star rating with the most number of review counts to get the desired output.

<https://kaggle.com/shanilabs/yelp-dataset/edit>

Dataset Draft saved

Edit View Run Add-ons Help

Code Run All Draft Session (42m)

```
query = """
SELECT business_id, name, city, stars, review_count
FROM business
WHERE stars = 5.0
ORDER BY review_count DESC
LIMIT 5;
"""

result = pd.read_sql_query(query, conn)
result
```

[41]:

	business_id	name	city	stars	review_count
0	_s07TPOnecW_YiaX0IpCA	Blues City Deli	Saint Louis	5.0	991
1	8QqyRpm-QxGyjDNuu0E5TA	Carlitos Cocina	Sparks	5.0	799
2	zxFrbaI+eDsn87yu7A	Free Tours By Foot	New Orleans	5.0	769
3	DvBIRvmCpkqaYMeHiseMg	Tumericos	Tucson	5.0	703
4	qP_cW7yka2Roch_GurCWNQ	Yata	Franklin	5.0	623

+ Code + Markdown

here to search

2

The below query is to find the top performing businesses with most 5 star ratings. This helps Yelp to identify the companies which have the most customer satisfaction, this can be used to give bonuses to the employees of the top business.

baggle.com/shanilabs/yelp-dataset/edit

Yelp Dataset Draft saved

File Edit View Run Add-ons Help

+ X Run All Code Draft Session (42m)

```
query = """
SELECT business_id, name, city, stars, review_count
FROM business
WHERE stars = 5.0
ORDER BY review_count DESC
LIMIT 5;"""

result = pd.read_sql_query(query, conn)
result
```

[41]:

	business_id	name	city	stars	review_count
0	_j6t7POnacW_VicfXhpCA	Blues City Deli	Saint Louis	5.0	991
1	8QgnRpM-QxGyjDNuu0ESTA	Carlitos Cocina	Sparks	5.0	799
2	zxF-iwai-eOzn87yuTA	Free Tours By Foot	New Orleans	5.0	769
3	DVB/RvnCpkqaYl6nHrpaMg	Tumerico	Tucson	5.0	705
4	gP_oWlykA2RocL_GurkWQ	Vato	Franklin	5.0	623

+ Code + Markdown

Type here to search

22°C

2. The below query is to find the top performing businesses with most 5 star ratings. This helps Yelp to identify the companies which have the most customer satisfaction, this can be used to give bonuses to the employees of the top business.

kaggle.com/shanilabs/yelp-dataset/edit

Yelp Dataset Draft saved

File Edit View Run Add-ons Help

+ X Run All Code

Draft Session (1h:7m)

```
query3 = """
    SELECT categories, COUNT(*) AS category_count
    FROM business
    GROUP BY categories
    ORDER BY category_count DESC
    LIMIT 10;
"""

result3 = pd.read_sql_query(query3, conn)
result3
```

[55]:

	categories	category_count
0	Beauty & Spas, Nail Salons	5012
1	Restaurants, Pizza	4955
2	Nail Salons, Beauty & Spas	4944
3	Pizza, Restaurants	4223
4	Restaurants, Mexican	728
5	Restaurants, Chinese	708
6	Mexican, Restaurants	472
7	Chinese, Restaurants	451

Type here to search

The query is to find the most number of Business Categories

→ C kaggle.com/shanilabs/yelp-dataset/edit

Yelp Dataset

Draft saved

File Edit View Run Add-ons Help

+ | X | ☰ | 📁 | ▶ | ▶▶ | Run All | Code | Draft Session (1h:12m) | ⚙ | ⌂ | ⌂ | ⌂

```
query8 = """
    SELECT city, AVG(stars) AS avg_rating, COUNT(*) AS total_businesses
    FROM business
    GROUP BY city
    ORDER BY avg_rating DESC, total_businesses DESC;
"""

result8 = pd.read_sql_query(query8, conn)
result8
```

[62]:

	city	avg_rating	total_businesses
0	Santa Barbara	5.0	5
1	Catalina Foothills	5.0	2
2	Reno	5.0	2
3	Town N Country	5.0	2
4	Aliso Viejo	5.0	1
...
1411	Peerless Park	1.0	1
1412	Pentlyn	1.0	1

Type here to search

The below query finds the location wise businesses count and their average rating. By this Yelp can find which location is best to start any new business or improve the business in the location where the average rating is low

Yelp Dataset Draft saved

File Edit View Run Add-ons Help

+ X Run All Code

Draft Session (1h55m)

query5 = """
SELECT strftime('%Y-%m', yelping_since) AS registration_month, COUNT(*) AS active_users
FROM `user`
WHERE yelping_since <= DATE('now')
GROUP BY registration_month;
""";

result5 = pd.read_sql_query(query5, conn)
result5

[65]:

	registration_month	active_users
0	2004-10	53
1	2004-11	14
2	2004-12	23
3	2005-01	24
4	2005-02	39
..
160	2021-09	2985
204	2021-10	2793

Type here to search

5. The below query is to find the active users by each month. By this Yelp can identify active user growth over the year and make sure the user base increases in the future.

Final Submission PLAN

Overall, the final submission aims to provide a comprehensive and insightful analysis of the Yelp dataset, drawing meaningful conclusions and recommendations for businesses and users alike.

- We plan to further deepen the analysis and apply more advanced data analysis techniques and technologies like Hadoop and Spark.
- Conducting sentiment analysis on reviews to understand the overall sentiments of users towards businesses.
- Analyzing geographic trends, such as the distribution of highly rated businesses across different locations.
- Creating interactive data visualizations to present the findings in a more user-friendly and informative way.
- Providing actionable recommendations to businesses based on the analysis to improve their services and ratings.

Conclusion

Through the SQL queries and understanding the business process of Yelp, we wanted to find out how we as analysts can contribute to their business growth. While this midpoint submission is just scratching the surface, we hope to produce greater insights using more big data technologies at the time of the final presentation.

404 !

Questions ?

THANK YOU !