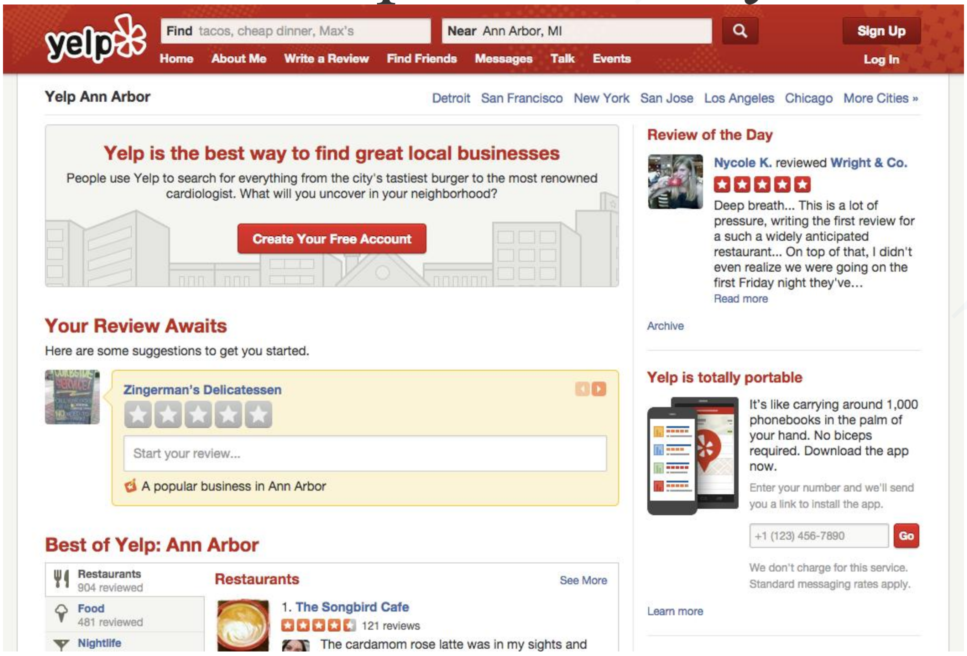
Author: Zhi (Katy) Gui

Yelp Data Analysis



Find tacos, cheap dinner, Max's

Near Ann Arbor, MI

Sign Up

Home About Me Write a Review Find Friends Messages Talk Events

Log In

Yelp Ann Arbor

Detroit San Francisco New York San Jose Los Angeles Chicago More Cities »

Yelp is the best way to find great local businesses

People use Yelp to search for everything from the city's tastiest burger to the most renowned cardiologist. What will you uncover in your neighborhood?

Create Your Free Account

Your Review Awaits

Here are some suggestions to get you started.

Zingerman's Delicatessen

Start your review...

A popular business in Ann Arbor

Best of Yelp: Ann Arbor

Restaurants
904 reviewed

Food
481 reviewed

Nightlife

Restaurants

See More

1. The Songbird Cafe

121 reviews

The cardamom rose latte was in my sights and

Review of the Day

Nycole K. reviewed Wright & Co.

Deep breath... This is a lot of pressure, writing the first review for a such a widely anticipated restaurant... On top of that, I didn't even realize we were going on the first Friday night they've...

Read more

Archive

Yelp is totally portable

It's like carrying around 1,000 phonebooks in the palm of your hand. No biceps required. Download the app now.

Enter your number and we'll send you a link to install the app.

+1 (123) 456-7890

Go

We don't charge for this service. Standard messaging rates apply.

Learn more

What is Yelp?

Yelp is an online platform that allows users to search for and browse information about businesses, including their address, phone number, hours of operation, and other details. It also enables users to view reviews and ratings of businesses written by other users, and to write and publish their own reviews and comments. Yelp's review system leverages the power of social networking, encouraging users to share their experiences and opinions about businesses, which helps other users to find better businesses. Yelp was founded in 2004 in the United States and has expanded to other countries including Canada, the United Kingdom, Australia, France, Germany, Italy, and Switzerland, among others.

Background

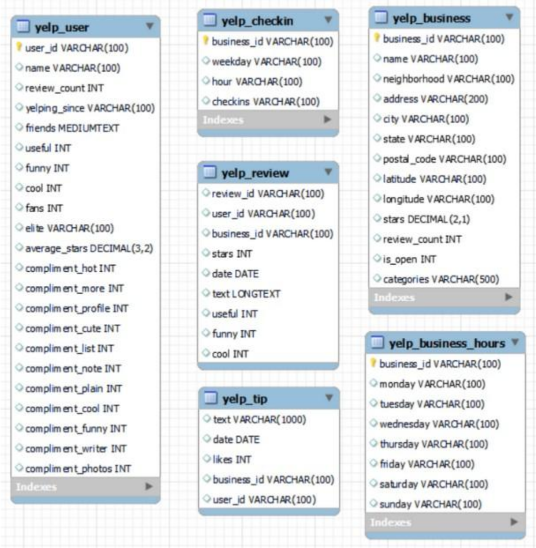
The Yelp dataset is a collection of data related to businesses, reviews, users, and other interactions on the Yelp platform. The dataset includes information from several cities across the United States, covering a variety of business categories and user demographics.

The purpose of this dataset is to enable researchers, data analysts, and data scientists to explore and analyze the dynamics of the Yelp platform and gain insights into user behavior, business performance, and market trends. With the increasing popularity of online review platforms like Yelp, this dataset provides a valuable resource for understanding the factors that influence customer satisfaction, business success, and platform growth.

In addition, the Yelp dataset is a widely used benchmark dataset for evaluating the performance of machine learning algorithms, particularly in the areas of natural language processing, sentiment analysis, and recommendation systems. By providing a large and diverse dataset that reflects real-world interactions and behaviors, the Yelp dataset allows researchers to develop and test new algorithms and models that can be applied to other online platforms and domains.

Overall, the Yelp dataset is a valuable resource for anyone interested in exploring the dynamics of online review platforms and understanding the factors that influence user behavior, business success, and platform growth.

About Dataset



yelp_user
user_id VARCHAR(100)
name VARCHAR(100)
review_count INT
yelping_since VARCHAR(100)
friends MEDIUMTEXT
useful INT
funny INT
cool INT
fans INT
elite VARCHAR(100)
average_stars DECIMAL(3,2)
compliment_hot INT
compliment_more INT
compliment_profile INT
compliment_cute INT
compliment_list INT
compliment_note INT
compliment_plain INT
compliment_cool INT
compliment_funny INT
compliment_writer INT
compliment_photos INT

yelp_checkin
business_id VARCHAR(100)
weekday VARCHAR(100)
hour VARCHAR(100)
checkins VARCHAR(100)

yelp_review
review_id VARCHAR(100)
user_id VARCHAR(100)
business_id VARCHAR(100)
stars INT
date DATE
text LONGTEXT
useful INT
funny INT
cool INT

yelp_tip
text VARCHAR(1000)
date DATE
likes INT
business_id VARCHAR(100)
user_id VARCHAR(100)

yelp_business
business_id VARCHAR(100)
name VARCHAR(100)
neighborhood VARCHAR(100)
address VARCHAR(200)
city VARCHAR(100)
state VARCHAR(100)
postal_code VARCHAR(100)
latitude VARCHAR(100)
longitude VARCHAR(100)
stars DECIMAL(2,1)
review_count INT
is_open INT
categories VARCHAR(500)

yelp_business_hours
business_id VARCHAR(100)
monday VARCHAR(100)
tuesday VARCHAR(100)
wednesday VARCHAR(100)
thursday VARCHAR(100)
friday VARCHAR(100)
saturday VARCHAR(100)
sunday VARCHAR(100)

- **Customer**

Number of reviews, good/bad reviews, length of reviews, number of photos, number of followers, elite users

- **Restaurant**

Number of restaurant distribution, star rating, rating, location, days of operation, closed restaurants

- **Platform**

Number of registered users, number of user reviews, percentage of elite customers, retention rate

This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. In the most recent dataset you'll find information about businesses across 8 metropolitan areas in the USA and Canada.

1. `yelp_business.csv`: This file contains information about businesses on the Yelp platform, including their ID, name, address, latitude and longitude, star rating, categories, and city.
2. `yelp_business_attributes.csv`: This file contains attributes of businesses, such as whether they offer Wi-Fi or have a parking lot.
3. `yelp_business_hours.csv`: This file contains information about the hours of operation of businesses, including the opening and closing times for each day of the week.
4. `yelp_checkin.csv`: This file contains check-in information for businesses, such as the time and date when users check in.
5. `yelp_review.csv`: This file contains reviews of businesses, including the user ID, business ID, rating, review text, and date of the review.
6. `yelp_tip.csv`: This file contains tips and suggestions from users about businesses, such as recommended dishes or things to watch out for.
7. `yelp_user.csv`: This file contains information about Yelp users, including their ID, name, registration time, and average rating.

When we use the Yelp dataset for data analysis, we can analyze the data from the perspectives of the platform, users, and merchants, and obtain the following benefits:

Platform perspective

From the platform's perspective, we can use the Yelp dataset to analyze user behavior and merchant operations, thereby guiding platform operations and marketing strategies. For example, we can analyze popular merchants and user preferences on the Yelp platform to provide more accurate targeting services to merchants and advertising clients. In addition, we can use data to analyze the development of cities to discover new business opportunities and market trends. For example, we can analyze the number and rating of merchants in different cities to understand which cities have more intense market competition and which cities have more promising business opportunities.

User perspective

From the user's perspective, we can use the Yelp dataset to analyze user behavior, taste preferences, and opinions and suggestions to provide better services and experiences for merchants. For example, we can analyze user check-ins and reviews on the Yelp platform to understand the types of merchants and service quality that users prefer. In addition, we can use data to analyze user suggestions and opinions to provide better products and services to merchants. For example, we can analyze user suggestions for different merchants on the Yelp platform to help merchants improve their products or services.

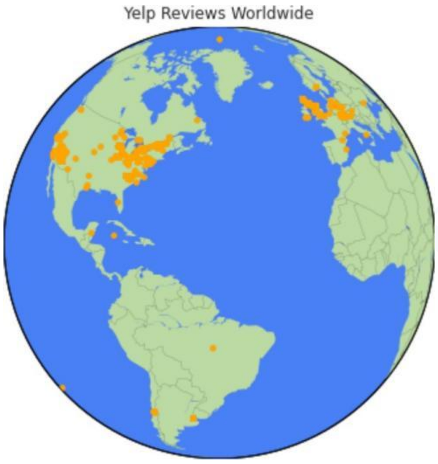
Business perspective

From the merchant's perspective, we can use the Yelp dataset to analyze merchant reviews, check-ins, and business hours to improve service quality and optimize business models. For example, we can analyze the ratings and reviews of merchants on the Yelp platform to improve their products and services. In addition, we can use data to analyze merchant check-ins and business hours to optimize their business strategies and increase customer traffic. For example, we can analyze customer traffic during different time periods and workdays to help merchants arrange more reasonable working hours and manpower resources.

In summary, using the Yelp dataset for data analysis can provide insights into the operation of the Yelp platform and user behavior from multiple perspectives, thereby providing better services to the platform and better operational strategies to merchants.

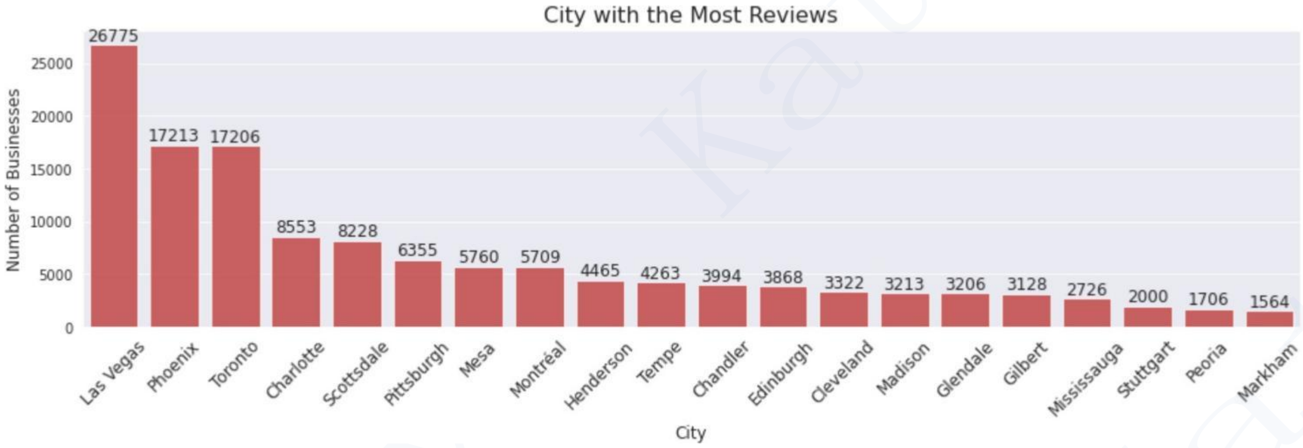
Analysis Perspectives

Business Types and Locations



We can analyze the number and distribution of businesses in different cities or regions, and understand which types of businesses are most popular and which areas have more businesses. This can help the platform and businesses understand market demand and competition. For example, we can analyze the business category and location information in the Yelp Business dataset, and determine the most popular business types and the most popular business districts in a city or region.

User Reviews and Preferences



We can analyze user ratings and reviews of businesses on the Yelp platform, and understand information such as user preferences for business types, service quality, and food taste, which can help businesses better optimize their products and services. For example, we can analyze a business's food taste, service quality, and user feedback through the user reviews and ratings data in the Yelp Review dataset, and help businesses improve or optimize their products and services.

Business Operations and Revenue

We can analyze a business's operating conditions and revenue situation, including information such as the business's business hours, customer flow, revenue, and expenses, which can help businesses better formulate operating strategies and increase revenue. For example, we can analyze the distribution of business traffic and operating hours through the check-in and business information data in the Yelp Checkin and Business datasets, and help businesses better arrange work hours and human resources.

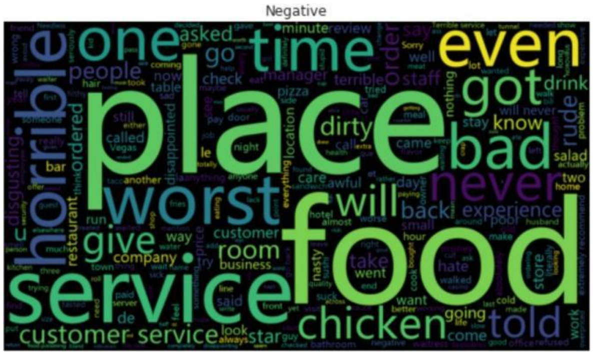
User Behaviors and Trends

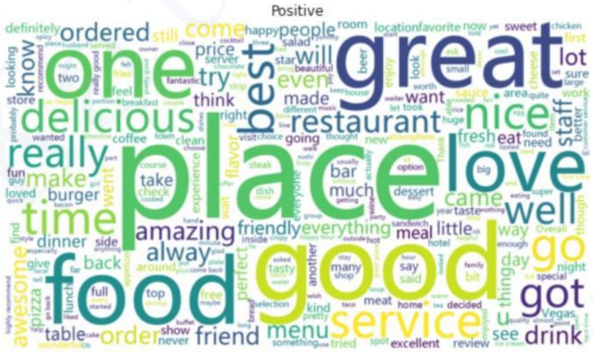
We can analyze user behavior and trends on the platform, including user reviews of businesses, check-ins, favorites, and likes, which can help businesses better understand user needs and behavior habits. For example, we can analyze user suggestions and preferences for businesses through the user comments and likes data in the Yelp Tip and Review datasets, and help businesses improve or optimize their products and services.

Market Competition and Trends

We can analyze market competition and trends, including information such as the number of businesses, reviews, market share, and trends in different cities. This can help businesses and the platform better understand market development trends and opportunities. For example, we can analyze the number of businesses and review situation in a city or region through the business and review data in the Yelp Business and Review datasets, understand the level of market competition and development trends, and help businesses and the platform better formulate development strategies.

Sentiment Analysis



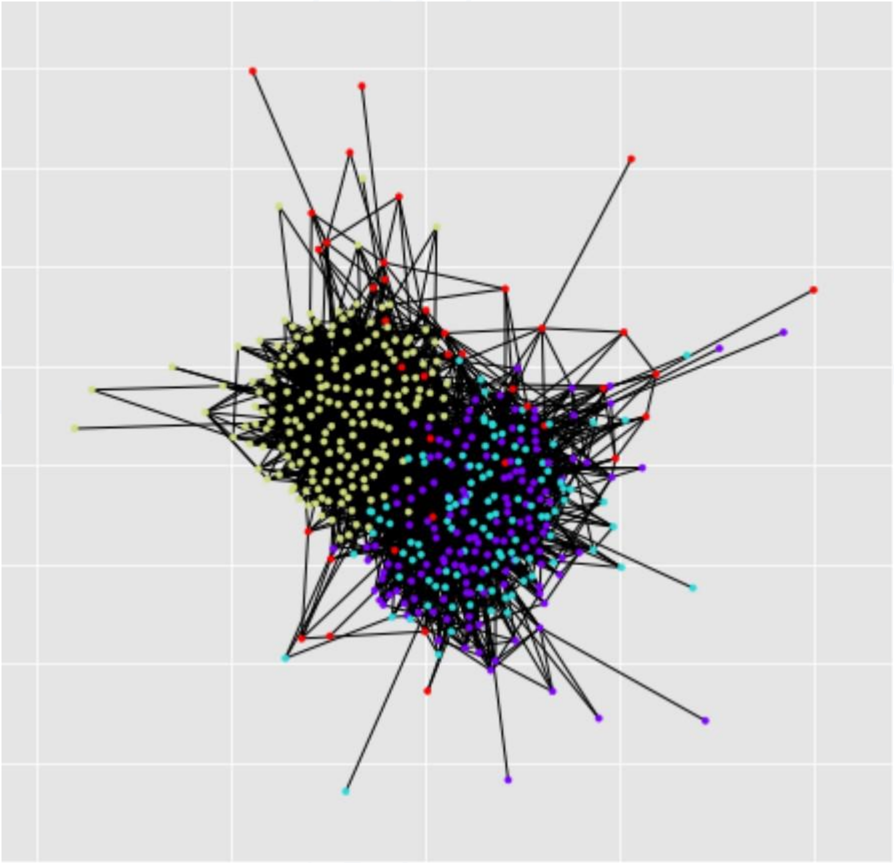


It involves using natural language processing techniques to analyze the opinions and emotions expressed in user reviews and ratings. By doing so, we can identify patterns and trends in user sentiment towards different businesses and their products or services, and help businesses improve their products or services to meet customer expectations

For example, we can use sentiment analysis to identify common issues or complaints mentioned in user reviews, such as long wait times, poor customer service, or low quality food. Businesses can then use this information to address these issues and improve their overall customer satisfaction. The platform itself can also benefit from sentiment analysis by identifying trends in user satisfaction and dissatisfaction across different categories of businesses, and use this information to improve the platform's features and services.

In addition, sentiment analysis can also be used to predict future user behavior and preferences, which can help businesses and the platform stay ahead of market trends and respond proactively to changing customer needs.

Network & Community Detection



It involves using network analysis techniques to identify groups of businesses or users that are closely related to each other based on their interactions on the platform. By doing so, we can identify communities of businesses or users that share similar interests, preferences, or behaviors, and use this information to better understand the dynamics of the Yelp platform.

For example, we can use community detection to identify groups of businesses that are competing with each other in the same market segment or location. By analyzing the interactions and relationships between these businesses, we can better understand their competitive landscape and help them develop more effective marketing strategies.

Similarly, we can use community detection to identify groups of users that are interacting with each other on the platform, such as users who frequently review or recommend the same types of businesses. By analyzing these user communities, we can better understand their preferences and behaviors, and use this information to improve the platform's recommendation algorithms and personalized services.

Overall, community detection is a powerful tool for understanding the complex network of relationships between businesses and users on the Yelp platform, and can help businesses and the platform better understand their customers and improve their services.

Challenges & Pain Points

Data volume and complexity

The Yelp dataset contains millions of rows of data, which can make it challenging to extract meaningful insights and patterns. One solution to this challenge is to use sampling techniques to reduce the size of the dataset while still maintaining the integrity of the analysis. For example, a data analyst may choose to focus on a subset of businesses or reviews that are most relevant to the research question at hand. However, the drawback of this solution is that it may overlook important patterns and trends that are only visible in the full dataset.

Data quality

The Yelp dataset is user-generated, which means that the quality of the data can vary widely. One solution to this challenge is to use data cleaning and preprocessing techniques to ensure that the data is accurate and reliable. For example, a data analyst may remove duplicates, correct misspellings, and standardize formats to ensure consistency. However, the drawback of this solution is that it can be time-consuming and may require manual intervention to correct errors that cannot be easily automated.

Bias

The Yelp dataset may be subject to bias and manipulation, which can make it difficult to draw accurate conclusions. One solution to this challenge is to use statistical techniques to account for bias and adjust for confounding variables. For example, a data analyst may use regression analysis to control for factors that may be influencing the outcome of interest, such as the number of reviews or the location of the business. However, the drawback of this solution is that it may not fully account for all sources of bias, and may require extensive data exploration to identify and address potential confounders.

Interpretation and communication

Before performing data analysis, a comprehensive understanding of Yelp's platform and the restaurant industry is required to better understand the data and develop appropriate analysis plans. For example, understanding user and merchant behavior and feedback on Yelp's platform, as well as the characteristics and business models of different types of restaurants, is necessary to perform an insightful analysis.

Data analysis is only valuable if the results are interpreted correctly and communicated effectively to stakeholders. One solution to this challenge is to use data visualization and storytelling techniques to convey the key insights and findings in a compelling and accessible way. For example, a data analyst may use interactive charts and graphs to highlight key trends and patterns, or use narrative techniques to explain the significance of the results in plain language. However, the drawback of this solution is that it may require specialized skills and expertise in data visualization and communication, which may not be available to all data analysts.

Gain

Data analysis skills

In the Yelp data analysis project, you need to clean, transform, aggregate and analyze large-scale and complex datasets, such as data type conversion, missing value handling, outlier detection and removal, data standardization, and group aggregation. At the same time, you also need to use different modeling and analysis methods, such as regression analysis, classification analysis, clustering analysis, and time-series analysis, to interpret the data and extract valuable information based on the analysis purpose.

Business analysis skills

In the Yelp data analysis project, you need to analyze the data from different perspectives, such as the platform, users, and merchants, to understand the operational status of the Yelp platform. For example, you can analyze user reviews for different types of restaurants to understand the popularity of different types of restaurants on Yelp; you can also analyze merchant feedback to understand their satisfaction with Yelp's services and use their feedback to improve the platform.

Programming skills

In the Yelp data analysis project, you need to use programming languages to process and analyze data. For example, you can use Python to write data cleaning and analysis scripts, use SQL to perform data querying and aggregation analysis, and use R for statistical analysis and data visualization.

Visualization skills

In the Yelp data analysis project, you need to use visualization tools to display the analysis results to non-technical stakeholders. For example, you can use tools such as Tableau, Power BI to perform visualization analysis, and create charts such as bar charts, scatterplots, line charts, and heatmaps to show the results of Yelp data analysis.

Teamwork skills

In the Yelp data analysis project, you need to collaborate with team members to develop analysis plans and interpret the analysis results. For example, you can work with team members to discuss the analysis purpose and research questions, design analysis plans and experiment workflows, and collaborate on data cleaning, analysis, and visualization tasks.