Dataset 1: Yelp Review Dataset

Each row in the Yelp Review dataset represents a single user-submitted review of a business on Yelp. Each observation corresponds to one review written by one user about one business at a specific point in time.

Column Name	Description	Example Values
review_id	Unique identifier for each review	Q1sbwvVQXV2734tPgoKj4Q
user_id	Unique identifier for the user who wrote the review	hG7b0MtEbXx5QzbzE6C_VA
business_id	Unique identifier for the business being reviewed	4JNXUYY8wbaaDmk3BPzlWw
stars	Star rating assigned by the reviewer (1–5 scale)	1, 3, 5
useful	Number of users who marked the review as useful	0, 2, 15
funny	Number of users who marked the review as funny	0, 1, 4
cool	Number of users who marked the review as cool	0, 3, 10
text	Full text content of the review	“Great food but slow service.”
date	Date the review was posted	2019-03-14 20:15:02

Dataset 2: Yelp Business Dataset

Each row in the Yelp Business dataset represents a single business listed on Yelp. Each observation corresponds to one business entity, including its identifying information, location, and category classifications.

Column Name	Description	Example Values

business_id	Unique identifier for each business	4JNXUYY8wbaaDmk3BPzlWw
name	Name of the business	Melt, Zaika, Dmitri’s
address	Street address of the business	123 Main St
city	City where the business is located	Las Vegas
state	U.S. state abbreviation	NV
postal_code	ZIP or postal code	89109
latitude	Geographic latitude coordinate	36.1147
longitude	Geographic longitude coordinate	-115.1728
stars	Average Yelp rating for the business (1–5 scale)	3.5, 4.0
review_count	Total number of reviews for the business	12, 245, 1023
is_open	Indicator of whether the business is currently open (1 = open, 0 = closed)	1, 0
attributes	Dictionary of business attributes (e.g., WiFi, OutdoorSeating)	{'WiFi': 'free'}
categories	Comma-separated list of business categories	“Mexican, Indian, Restaurants”
hours	Dictionary listing business operating hours by day	{'Monday': '9:00-17:00'}

Dataset 3: Final Analytical Dataset

Each row in the final analytical dataset represents a single Yelp review of a restaurant that has been merged with corresponding business information and processed for sentiment analysis. Each observation corresponds to one review–business pair, including the original star rating, extracted cuisine categories, and derived variables measuring the relationship between textual sentiment and numerical rating. This dataset reflects the cleaned and transformed data used for all statistical analyses in the study.

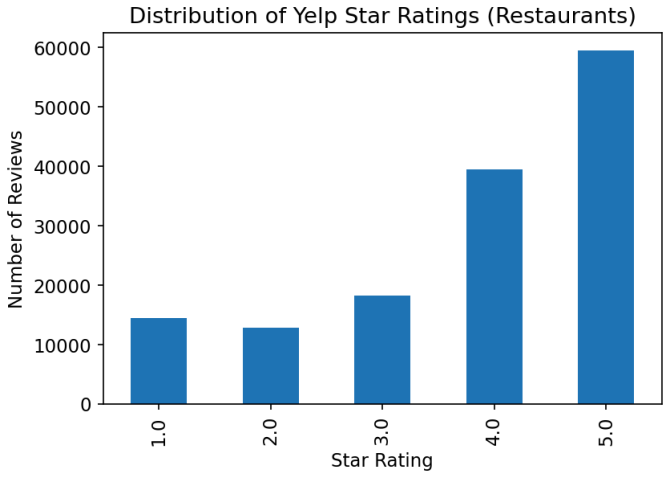
Column Name	**Description**	**Example Values**
business_id	Unique identifier for each business	4JNXUYY8wbaaDmk3BPzlWw
review_id	Unique identifier for each review	Q1sbwvVQXV2734tPgoKj4Q

text	Contains the full text of the Yelp review	“Great food but slow service.”
stars	Indicates the original Yelp star rating	1, 3, 5
cuisine	Indicates the extracted cuisine categories	[Mexican, Indian]
categories	Indicates the business category string from Yelp	“Mexican, Indian, Restaurants”
name	Indicates the business name	Melt, Zaika, Dmitri’s
stars_norm	Indicates the star rating rescaled to 0-1	0.2, 0.6, 1.0
mismatch	Indicates the magnitude of the difference between sentiment and star_norm	0.33
direction	Indicates the magnitude and sign of the difference between sentiment and star_norm	-0.33

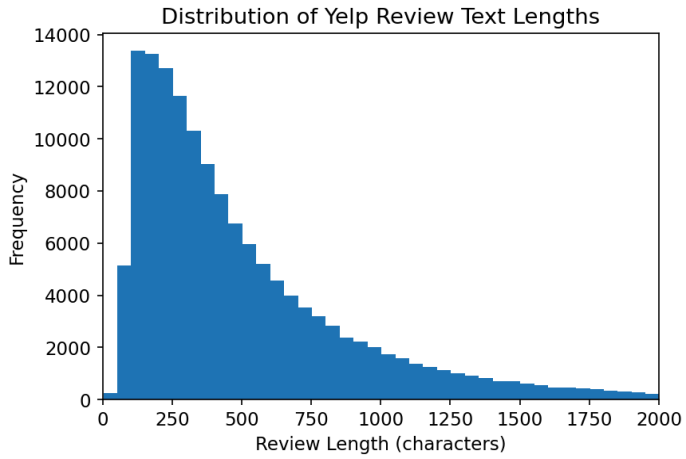
Exploratory Visualizations:

The following exploratory plots were created to establish familiarity with the dataset:

Distribution of Yelp Star Ratings

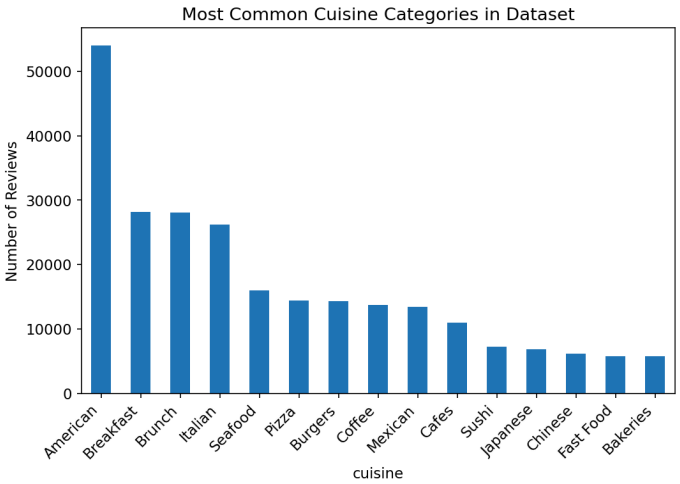


Review Text Length Distribution

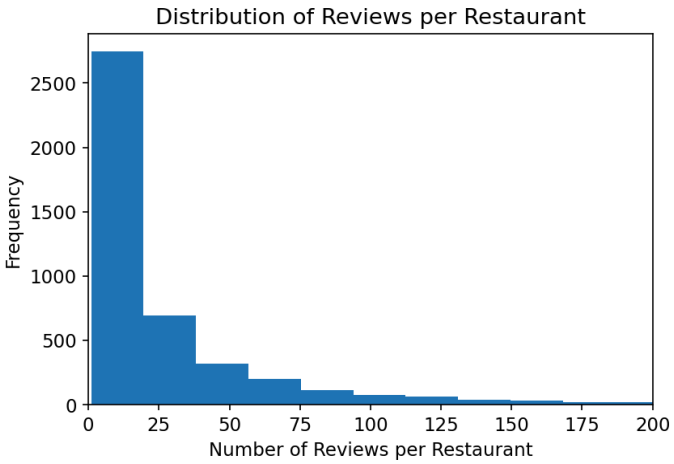


Number of Cuisine Tags per Review

Most Common Cuisine Categories



Reviews per Restaurant


Distribution of Reviews per Restaurant
Frequency
2500
2000
1500
1000
500
0
0
25
50
75
100
125
150
175
200
Number of Reviews per Restaurant