

The merged dataset (interaction_df_cleaned) contains columns from all five original files:

- 1. **review.json**
- 2. **user.json**
- 3. **business.json**
- 4. **tip.json**
- 5. **checkin.json**

Attributes Overview

Attribute	Source	Description
review_id	review.json	Unique identifier for each review. Null for tips and check-ins.
user_id	All	Unique identifier for each user. Present in reviews, tips, check-ins.
business_id	All	Unique identifier for each business. Present in reviews, tips, check-ins.
user_rating	review.json	Rating given by the user to the business (1 to 5). 0 if from tip or check-in.
text	review.json, tip.json	Text content of the interaction. Full review text from review.json or short tip from tip.json. Null for check-ins.
date	All	Date of interaction (review, tip, or check-in).
business_name	business.json	Name of the business. Null if the business is not found in business.json.
categories	business.json	List of categories associated with the business (e.g., Restaurants, Cafes).

Attribute	Source	Description
attributes	business.json	Business-specific attributes (e.g., Wi-Fi availability, Parking, Alcohol served).
business_average_rating	business.json	Average star rating of the business. 0 if the business is not found in business.json.
business_review_count	business.json	Number of reviews the business has received. 0 if not found in business.json.
is_open	business.json	Indicator if the business is open (1 = Open, 0 = Closed). Null if business is not found.
address	business.json	Street address of the business. 'Unknown' if not found.
city	business.json	City where the business is located. 'Unknown' if not found.
state	business.json	State where the business is located. 'Unknown' if not found.
postal_code	business.json	Postal code of the business. 'Unknown' if not found.
latitude	business.json	Latitude of the business location. 0 if not found.
longitude	business.json	Longitude of the business location. 0 if not found.
average_stars	user.json	Average rating given by a user across all their reviews. Null if not found.
friends	user.json	List of user IDs representing the user's friends. Null if not found.
yelping_since	user.json	Date when the user joined Yelp. Null if not found.

Attribute	Source	Description
<code>source</code>	Merged Dataset	The source of the interaction: <code>review</code> , <code>tip</code> , or <code>checkin</code> . Added during merging.

Attributes Related to Reviews

- **review_id**: Unique ID for each review. Only exists if the interaction comes from `review.json`.
- **user_rating**: Rating given by the user (1 to 5). 0 if the interaction is a tip or check-in.
- **text**: Text of the review or tip. If null, it's likely a check-in interaction.

Attributes Related to Businesses

- **business_id**: Unique ID for the business. Common across all datasets.
- **business_name**: Name of the business. null if the business entry is not found in `business.json`.
- **categories**: Categories describing the business (e.g., Restaurants, Cafes).
- **attributes**: Detailed attributes (e.g., Wi-Fi availability, Parking options).
- **business_average_rating**: Average rating given to the business by all users.
- **business_review_count**: Number of reviews received by the business.
- **is_open**: Whether the business is open or closed (1 = Open, 0 = Closed).
- **address, city, state, postal_code**: Location details.
- **latitude, longitude**: Geographical coordinates of the business.

Attributes Related to Users

- **user_id**: Unique identifier for the user.
- **average_stars**: The average rating the user has given across all reviews.
- **friends**: List of friends (user IDs) connected to this user. Useful for social-based recommendation systems.

- **yelping_since:** When the user joined Yelp.

Attributes Related to Tips and Check-Ins

- **source:** Indicates whether the interaction came from:
 - review (Full review from review.json)
 - tip (Short recommendation from tip.json)
 - checkin (Interaction from checkin.json without text or rating)
- **date:** Date when the interaction occurred. Important for **sequential modeling**.

Importance of These Attributes

- **User-Item Interaction Modeling:** Combines various types of user interactions with businesses to create a robust dataset.
- **Sequential Modeling:** The date attribute allows for building **time-aware models**.
- **Location-Based Recommendations:** city, state, latitude, and longitude provide valuable information for local recommendations.
- **Social-Based Recommendations:** friends attribute helps in building **social-aware models**.