Boston
Area
Research
Initiative

Boston Area Research Initiative ● Northeastern & Harvard Universities
E-Mail: bari@northeastern.edu ● Web: www.bostonarearesearchinitiative.net

A Guide to Yelp Data for Boston

Overview

This dataset contains information about 2,664 restaurants in Boston that were reviewed on yelp.com between October 13th, 2004 to August 17th, 2020. Using a Python script, we searched Yelp by ZIP code to create a list of all Boston restaurants listed on the website. We then scraped each restaurant page to collect information about general restaurant characteristics and reviews. These data have been processed by the Boston Area Research Initiative (BARI) to generate metrics at three analytic levels:

- *YELP.Reviews.csv* contains information about Yelp reviews posted for restaurants in our sample. The file also contains contextual information about the user who posted each review.
- *YELP.Restaurant.csv* is a file that contains information about the restaurants listed on Yelp.
- *YELP.CT.csv* contains aggregate measures (i.e. ecometrics) that describe neighborhoods (in the form of census tracts) in terms of reviewer activity and pricing. These variables are provided in a spreadsheet format (*.csv*) and in mappable shapefiles (*.shp*).

This documentation describes the contents and variables of this dataset.

Table of Contents

1. Reviews

1.1. Summary of Yelp Review Data

To create a database of reviews, we collected all reviews posted on the Yelp pages of all restaurants in Boston (see Section 2.1 for further information on restaurant identification process). In reviews, Yelp users rate restaurants on a 5-point numeric system and post a text passage describing their experience. Limited information about the user who authored the post is listed along with each review. BARI gathered and processed these features of Yelp pages to generate the variables included in this review-level datafile. Aggregate measures based on these reviews were also generated to describe restaurants (see Section 2).

1.2. Description of Variables

Review variables are split into two categories: review information and user information. Review information includes the review text, the date the review was posted, the rating associated with that review. User information includes contextual variables about the user who posted the review

1.2.1. Review Information

- *restaurant_name* is the restaurant name posted on Yelp.
- *restaurant_ID* is a unique identifier for each restaurant.
- *comment* is a string variable containing the text from the posted review.
- *rating* indicates the rating the reviewer left, with a 5 representing the best possible experience and a 0 representing the worst possible experience.
- *review_date* is the date the review was posted.

1.2.2. User Information

- *reviewer_name* is the username of the user who posted the review.
- *reviewer_origin* is the home city and state identified by the user.
- *reviewer_profile* is a URL linking to the reviewer's Yelp profile.
- *history_1* is the number of 1-star reviews the user has posted on Yelp in the past.
- *history_2* is the number of 2-star reviews the user has posted on Yelp in the past.
- *history_3* is the number of 3-star reviews the user has posted on Yelp in the past.
- *history_4* is the number of 4-star reviews the user has posted on Yelp in the past.

- *history_5* is the number of 5-star reviews the user has posted on Yelp in the past.
- *reviewer_friends* is the number of friends the user has on Yelp.
- *reviewer_reviews* is the number of reviews the user has posted on Yelp
- *reviewer_photos* is the number of photos the user has posted on Yelp.
- *reviewer_elite* is a binary variable where “1” indicates that the user has achieved “elite” status on Yelp, and “0” indicates the user has not achieved elite status.

2. Restaurants

2.1. Summary of Yelp Restaurant Data

To organize user reviews, Yelp creates a webpage for each restaurant that includes general information about the restaurant along with the posted reviews. To create a list of every Boston restaurant listed on Yelp, we searched all Boston zip codes through yelp and collected the URLs that were yielded. To further limit this sample to restaurants exclusively within the city of Boston, we geocoded the address of each restaurant (as stated by Yelp) using here.com and then overlayed the geographic coordinates over BARI's Boston Tax Assessment database shapefile to link each restaurant to the land parcel in which it is located. We then scraped the Yelp pages for each restaurant in Boston to generate the variables included in this restaurant-level datafile.

2.2. Description of Variables

Restaurant variables are split into three categories: restaurant characteristics, longitudinal review patterns, and geographical information. Restaurant characteristics include variables regarding the basic identity and attributes of the business according to Yelp. To facilitate research about how restaurants have been impacted by COVID-19, we have generated longitudinal review pattern metrics, which include monthly totals of restaurant reviews during 2020 (January-August) and the corresponding months of 2019. For longitudinal variables, variable suffixes (MMM)_(YY) indicate the month and year of measurement, respectively. Geographical information provides further detail on the location of the restaurant.

2.2.1. Restaurant Characteristics

- *restaurant_name* is the restaurant name posted on Yelp.
- *restaurant_ID* is a unique identifier for each restaurant.
- *restaurant_address* is the restaurant's postal address according to Yelp
- *restaurant_tag* indicates the tags used to describe features of the restaurant (e.g. seafood, Chinese).
- *rating* indicates the average rating based on user reviews.
- *price* indicates the cost of food at the restaurant based on Yelp's rough classification system. Values include: 1 for $, 2 for $$, 3 for $$$, and 4 for $$$$
- *review_number* is the number of total reviews the restaurant has received

- *unique_reviewer* is the number of unique reviewers who reviewed the restaurant.
- *URL* is the URL of the restaurant's page on Yelp.

2.2.2. Longitudinal Review Patterns

- *reviews_MMM_YY* is the number of reviews the restaurant received in the measured month.
    - *Note*: The August 2020 value is based only on reviews from August 1st to August 20th.

2.2.3. Geographical Information

- *restaurant_neighborhood* indicates which neighborhood the restaurant is in according to Yelp.
- *GIS_ID* is the identifier for the land parcel the restaurant is in.
- *CT_ID_10* is the 2010 Census Tract ID number.

3. Neighborhoods

3.1. Summary of Neighborhood-level Yelp Data

A neighborhood-level dataset was created that describes aggregate features of neighborhood restaurants. Neighborhood-level variables are provided at the census tract level. For longitudinal variables, variable suffixes (MMM)_(YY) indicate the month and year of measurement, respectively. Neighborhood-level measures are provided in both standard format (.csv) and as mappable shapefiles (.shp).

3.2. Description of Variables

- *CT_ID_10* is the 2010 Census Tract ID number.
- *NUM_REST* is the number or restaurants located in the tract.
- *RATE_REVIEWS* represents the frequency of reviews per restaurant in the tract.
- *RATE_REVIEWERS* represents the frequency of unique reviewers per restaurant in the tract.
- *AVG_RATING* indicates the average rating assigned to restaurants in the tract.
- *PCT_DLRS_1* is the percentage of restaurants in the tract assigned a 1-dollar sign price-value by Yelp.
- *PCT_DLRS_2* is the percentage of restaurants in the tract assigned a 2-dollar sign price-value by Yelp.
- *PCT_DLRS_3* is the percentage of restaurants in the tract assigned a 3-dollar sign price-value by Yelp.
- *PCT_DLRS_4* is the percentage of restaurants in the tract assigned a 4-dollar sign price-value by Yelp.
- *PCT_DLRS_NA* is the percentage of restaurants in the tract that have not been assigned a dollar sign price-value by Yelp.
- *reviews_MMM_YY* is the number of reviews the restaurant received in the measured month.
    - *Note*: The August 2020 value is based only on reviews from August 1st to August 20th.