Data Appendix

Restaurant Dataset

The unit of observation for this dataset is “restaurant,” that is, each row corresponds to one restaurant in Philadelphia, PA.

business_id

This variable is a unique string identifier for each restaurant in the dataset. There are 816 observations total, with no missing values.

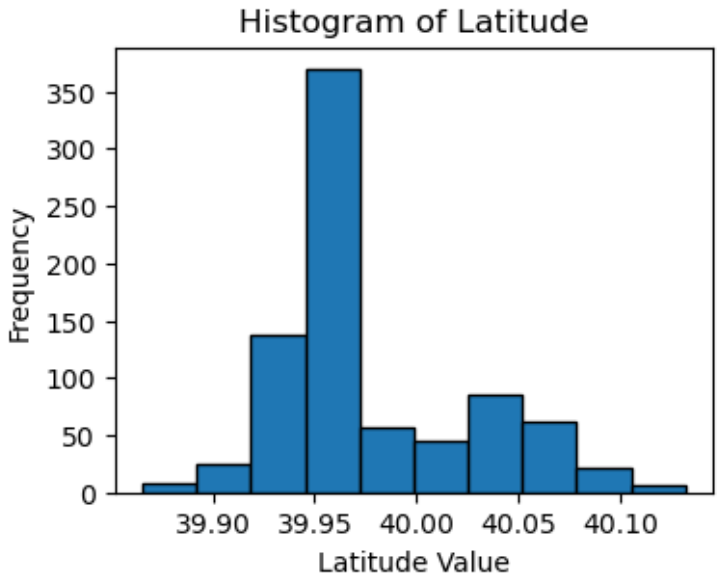
name

This variable is a unique string containing the name of each restaurant in the dataset. There are 816 observations total, with no missing values.

latitude

This variable is the latitude value for each restaurant in the dataset as a floating point number. There are 816 observations total, with no missing values.

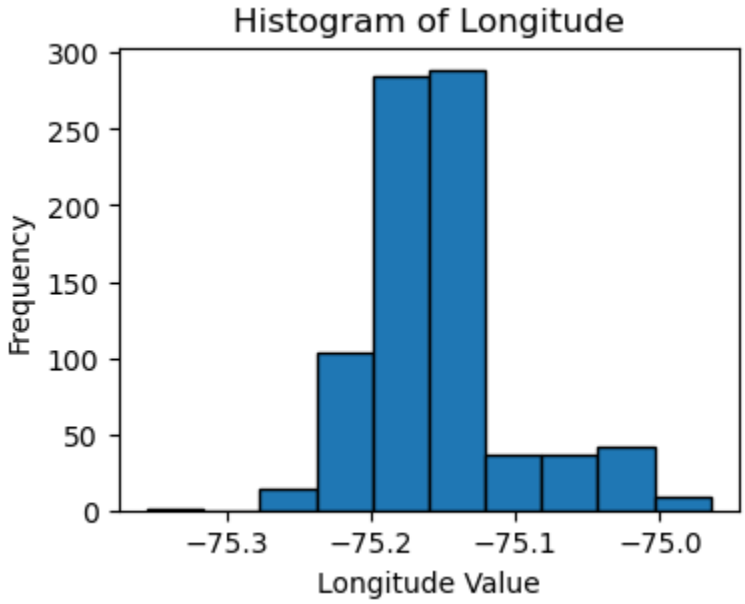
Statistic	Value
mean	39.976225
standard deviation	0.048560
minimum	39.865466
25th percentile	39.948201
median	39.954226
75th percentile	40.007555
maximum	40.131959



longitude

This variable is the longitude value for each restaurant in the dataset as a floating point number. There are 816 observations total, with no missing values.

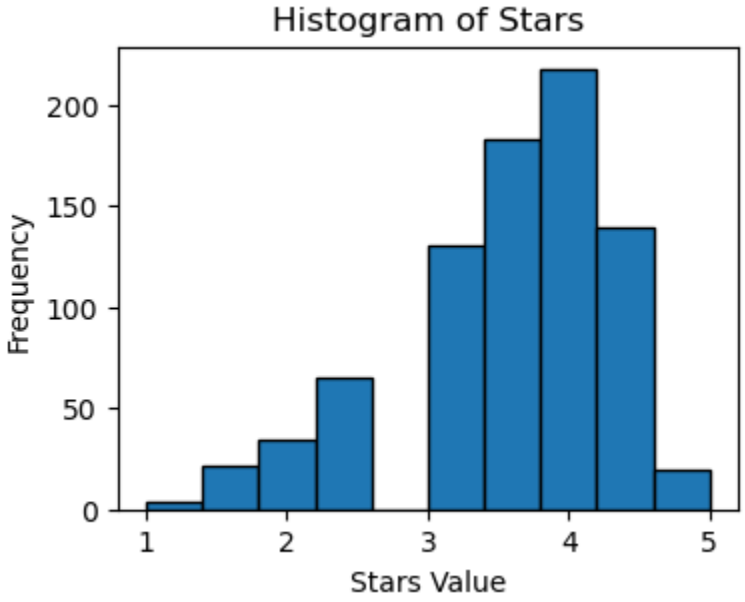
Statistic	Value
mean	-75.153262
standard deviation	0.050905
minimum	-75.355294
25th percentile	-75.175469
median	-75.159771
75th percentile	-75.144479
maximum	-74.964808



stars

This variable is the number of stars for each restaurant, determined by Yelp, which is an average of stars given in reviews. There are 816 observations total, with no missing values.

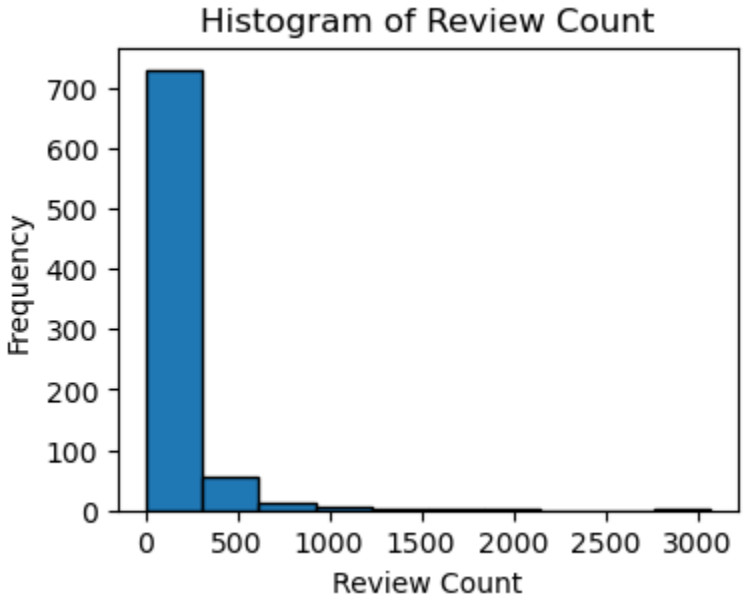
Statistic	Value
mean	3.552083
standard deviation	0.801484
minimum	1
25th percentile	3
median	3.5
75th percentile	4
maximum	5



review_count

This variable is an integer representing the number of reviews each restaurant has in the reviews dataset. There are 816 observations total, with no missing values.

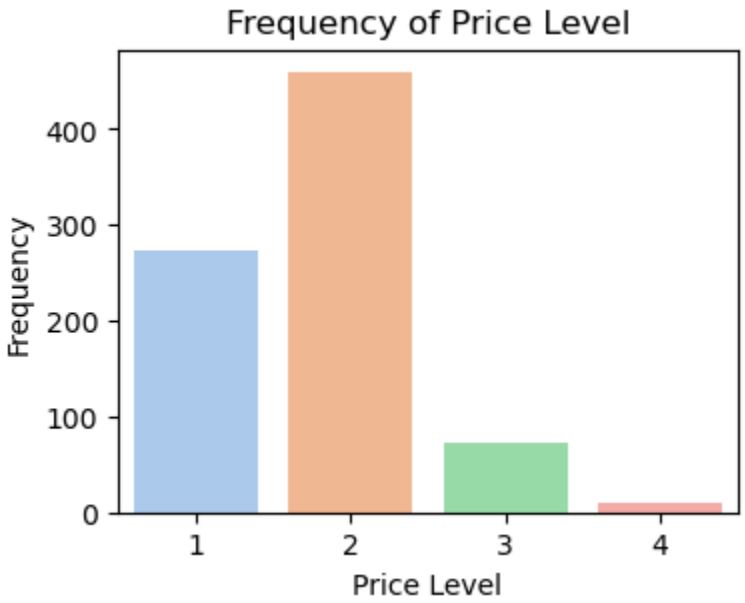
Statistic	Value
mean	127.626225
standard deviation	259.711300
minimum	5
25th percentile	14
median	40.5
75th percentile	127.25
maximum	3065



price_level

This variable is the price level for each restaurant, determined by Yelp. It ranges from 1 representing the least expensive restaurants, to 4 which are the most expensive restaurants. There are 816 observations total, with no missing values.

Value	Count
1	274
2	460
3	72
4	10



full_address

This variable is a string containing the address for each restaurant in the dataset. It was created by combining the 'address', 'city', 'state', and 'postal_code' columns from the raw data. There are 816 observations total, with no missing values.

Reviews Dataset

The unit of observation for this dataset is “review,” that is, each row corresponds to one review that a person left on Yelp.

review_id

This variable is a unique string identifier for each review in the dataset. There are 107600 observations total, with no missing values.

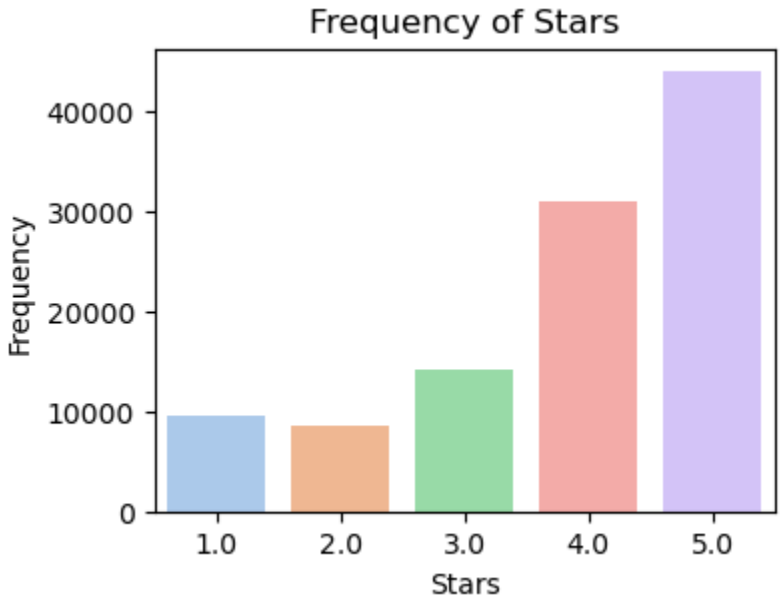
business_id

This variable is a unique string identifier for each restaurant in the dataset, which corresponds to the business_id in the restaurant dataset. There are 107600 observations total, with no missing values.

stars

This variable is the number of stars given by a customer for a restaurant review. There are 107600 observations total, with no missing values.

Value	Count
1	9614
2	8640
3	14208
4	31015
5	44123



text

This variable is the text of the review given by a customer for a restaurant. There are 107600 observations total, with no missing values.

date

This variable is the date and time each review was left on Yelp, in ‘yyyy-mm-dd hh:mm:ss’ format. The dates range from ‘2005-07-26 23:42:06’ to ‘2022-01-19 19:03:21.’