# OSM Amenities Data Quality Report

## Overview

This report will outline the initial data quality findings on amenities data obtained from Overpass Turbo API which can be found at <https://overpass-turbo.eu/>. This report will include an overview of the dataset, and a review of the continuous and categorical features, including histograms and bar charts. On initial review, this dataset contains a lot of missing data for most features. The data that is present appears to be reasonable and logical, however a number of columns will need to be dropped. 42 features contain missing values.

## Summary

This dataset consists of information on different Amenities in New York City, including bars, restaurants, and cafes. The dataset has 46 features and 14647 rows. 36 features will have to be dropped due to missing values. There are no duplicate rows. Distribution of the data is consistent with expectations.

## Review Logical Integrity

Test 1: Address state is not “NY”

* 0 instances.

Test 2: Website does not start with http:// or https://

* 2 instances, both starting with “www”

Test 3: Zip code not in list of NYC zip codes

* 0 instances.

## Review Continuous Features

There are 3 continuous features in this dataset:

* id
* This feature has a mean of 6.2947e+09, a min value of 3.4932e+08 and a max value of 1.1951e+10. There are no missing values.
* lat
* This feature has a mean of 40.722, a min value of 40.51 and a max value of 40.907. There are no missing values.
* lon
* This feature has a mean of -73.956, a min value of -74.251 and a max value of -73.7. There are no missing values.

### Histograms

All Histograms can be found in the appendix as a summary sheet. All features show a plausible distribution.

## Review Continuous Features

There are 3 continuous features in this dataset:

* addr:city
* This has 71 unique values. The most common is New York. There are 9684 missing values.
* addr:housenumber
* This has 4257 unique values. The most common is 2. There are 4565 missing values.
* addr:postcode
* This has 220 unique values. The most common is 10003. There are 6817 missing values.
* addr:state
* This has 2 unique values. The most common is NY. There are 10039 missing values.
* addr:street
* This has 1237 unique values. The most common is Broadway. There are 4442 missing values.
* amenity
* This has 16 unique values. The most common is restaurant. There are no missing values.
* branch
* This has 495 unique values. The most common is Upper East Side. There are 13909 missing values.
* brand
* This has 207 unique values. The most common is Dunkin'. There are 12575 missing values.
* brand:wikidata
* This has 178 unique values. The most common is Q847743. There are 12627 missing values.
* cuisine
* This has 1304 unique values. The most common is pizza. There are 4620 missing values.
* drive\_through
* This has 2 unique values. The most common is no. There are 13740 missing values.
* name
* This has 11305 unique values. The most common is Dunkin'. There are 212 missing values.
* official\_name
* This has 25 unique values. The most common is Starbucks Coffee. There are 14169 missing values.
* opening\_hours
* This has 4407 unique values. The most common is 24/7. There are 7649 missing values.
* phone
* This has 7567 unique values. The most common is +1-718-636-4100. There are 7006 missing values.
* takeaway
* This has 11 unique values. The most common is yes. There are 10489 missing values.
* website
* This has 6598 unique values. The most common is https://www.diginn.com/. There are 7841 missing values.
* outdoor\_seating
* This has 31 unique values. The most common is yes. There are 12124 missing values.
* cocktails
* This has 6 unique values. The most common is yes. There are 13642 missing values.
* drink:beer
* This has 16 unique values. The most common is yes. There are 13217 missing values.
* drink:liquor
* This has 3 unique values. The most common is yes. There are 13809 missing values.
* drink:wine
* This has 8 unique values. The most common is yes. There are 13337 missing values.
* wheelchair
* This has 6 unique values. The most common is yes. There are 12770 missing values.
* check\_date
* This has 308 unique values. The most common is 2024-03-13. There are 13717 missing values.
* payment:cash
* This has 3 unique values. The most common is yes. There are 13988 missing values.
* payment:credit\_cards
* This has 4 unique values. The most common is yes. There are 14071 missing values.
* email
* This has 591 unique values. The most common is info@goles.org. There are 14051 missing values.
* alt\_name
* This has 375 unique values. The most common is Dunkin' Donuts. There are 13871 missing values.
* ref
* This has 548 unique values. The most common is 3.0. There are 14099 missing values.
* internet\_access
* This has 5 unique values. The most common is wlan. There are 13834 missing values.
* level
* This has 31 unique values. The most common is 0. There are 13652 missing values.
* diet:vegan
* This has 4 unique values. The most common is yes. There are 14163 missing values.
* diet:vegetarian
* This has 4 unique values. The most common is yes. There are 14031 missing values.
* toilets
* This has 3 unique values. The most common is yes. There are 13913 missing values.
* drink:coffee
* This has 4 unique values. The most common is yes. There are 13673 missing values.
* drink:tea
* This has 4 unique values. The most common is yes. There are 14169 missing values.
* smoking
* This has 7 unique values. The most common is no. There are 14126 missing values.
* bar
* This has 4 unique values. The most common is yes. There are 14154 missing values.
* contact:instagram
* This has 450 unique values. The most common is https://instagram.com/vanleeuwenicecream/. There are 14188 missing values.
* delivery
* This has 45 unique values. The most common is yes. There are 12502 missing values.
* drink:espresso
* This has 2 unique values. The most common is yes. There are 14128 missing values.
* reservation
* This has 11 unique values. The most common is yes. There are 14025 missing values.
* indoor\_seating
* This has 3 unique values. The most common is yes. There are 14001 missing values.

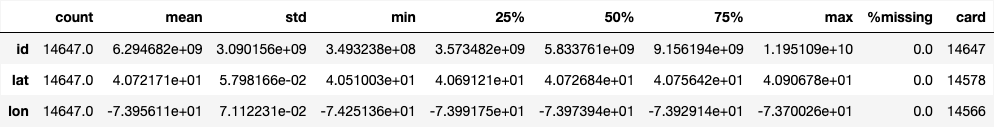
## Actions to Take

4 actions will be taken:

* Missing name
  + Drop rows with missing name.
* Vegan/Vegetarian
  + Change “only” and “limited” values to “yes”.
* State
  + Drop rows with state as “NJ”
* Wheelchair
  + Change “designated” and “dedicated” to “yes”.

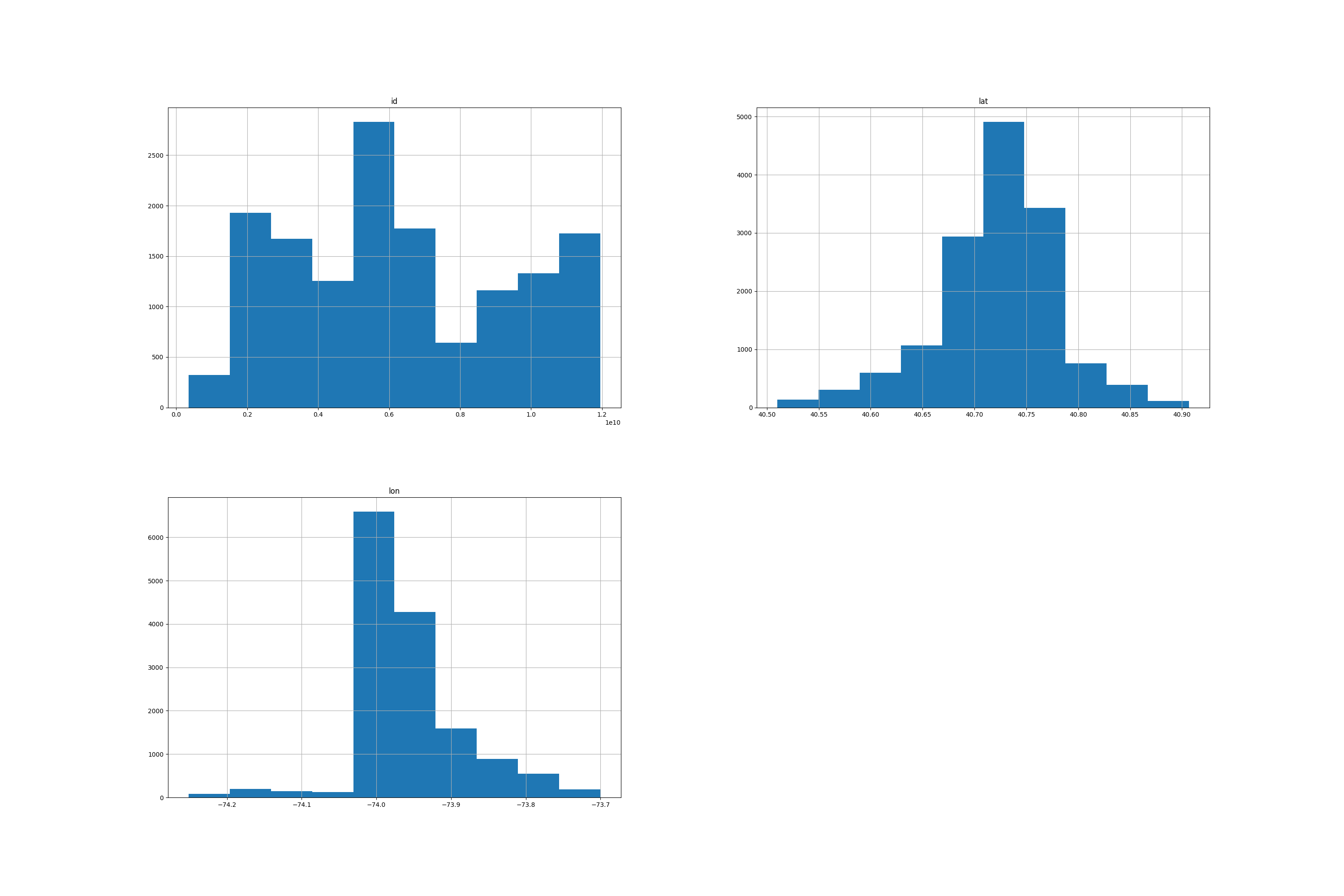
# Appendix

## Continuous Features



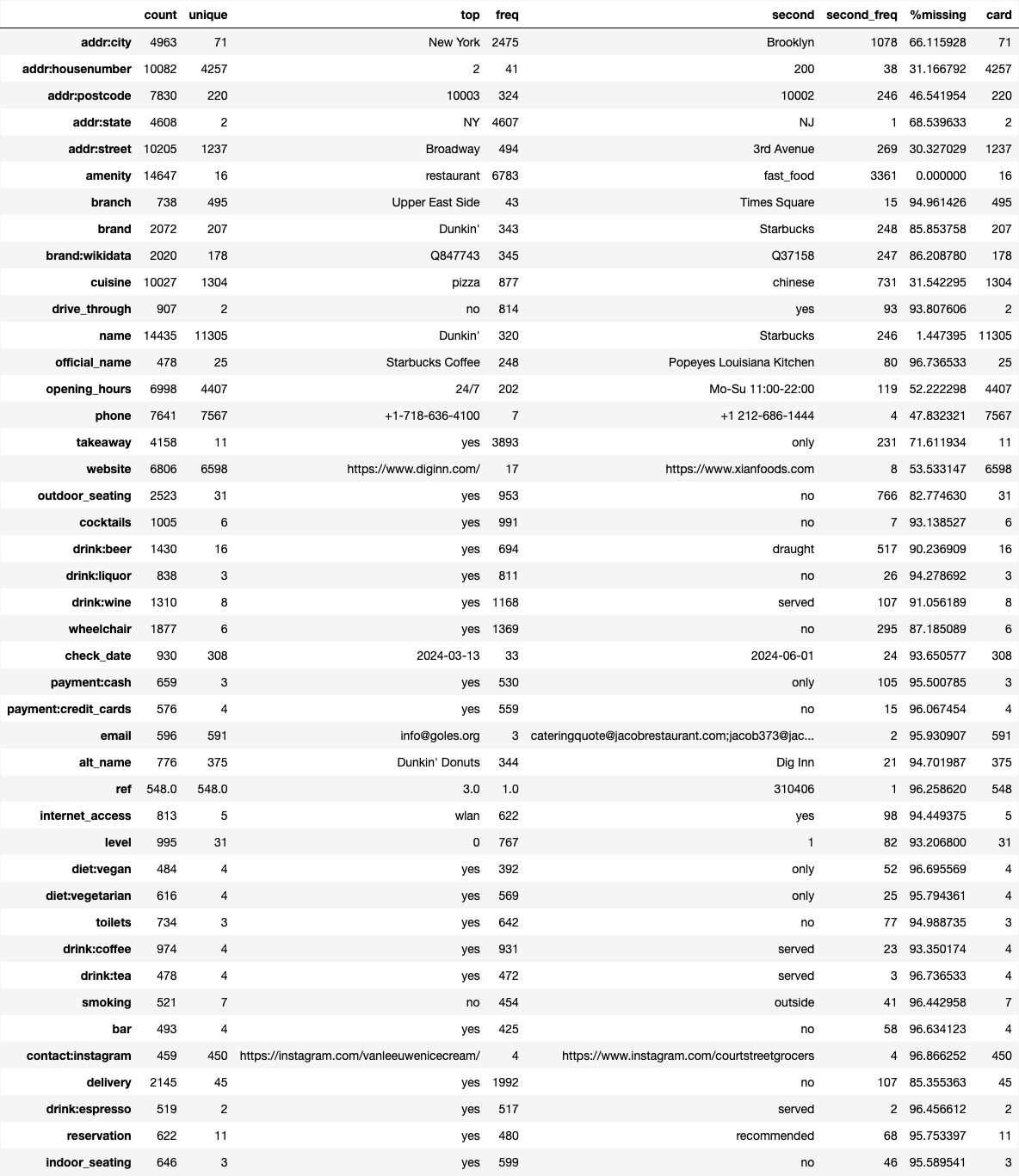
### Descriptive Statistics

### Histograms



## Categorical Features

## Descriptive Statistics



## Box Plots

