# OSM Leisure Data Quality Report

## Overview

This report will outline the initial data quality findings on leisure 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. 12 features contain missing values.

## Summary

This dataset consists of information on different leisure places in New York City, including arcades, beach resorts, parks and stadiums. The dataset has 27 features and 418 rows. 10 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:

* lat
* This feature has a mean of 40.725, a min value of 40.511 and a max value of 40.903. There are no missing values.
* lon
* This feature has a mean of -73.926, a min value of -74.25 and a max value of -73.71. There are no missing values.
* addr:postcode
* This feature has a mean of 10806, a min value of 10003 and a max value of 11419. There are 372 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:

* id
* This has 418 unique values. There are no missing values.
* name
* This has 369 unique values. The most common is Chuck E. Cheese. There are 40 missing values.
* addr:city
* This has 11 unique values. The most common is Brooklyn. There are 382 missing values.
* addr:state
* This has 1 unique values. The most common is NY. There are 382 missing values.
* opening\_hours
* This has 33 unique values. The most common is Mo-Su 18:00-04:00. There are 384 missing values.
* website
* This has 70 unique values. The most common is http://jupiterdisco.com/. There are 348 missing values.
* addr:street
* This has 64 unique values. The most common is Flushing Avenue. There are 345 missing values.
* addr:housenumber
* This has 69 unique values. The most common is 10. There are 349 missing values.
* phone
* This has 53 unique values. The most common is (718) 590-9000. There are 365 missing values.
* wheelchair
* This has 3 unique values. The most common is yes. There are 406 missing values.
* email
* This has 13 unique values. The most common is bookings@fiveirongolf.com. There are 405 missing values.
* access
* This has 3 unique values. The most common is yes. There are 397 missing values.

## Actions to Take

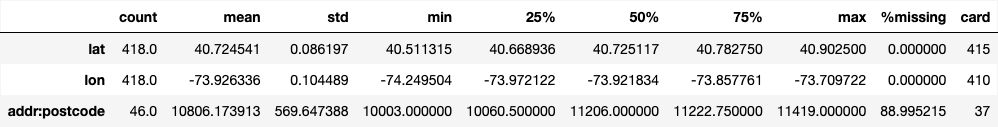
2 actions will be taken:

* Access
  + Any rows with access as “private” will be dropped and the feature will be dropped.
* Postcode
  + Change postcode type to int

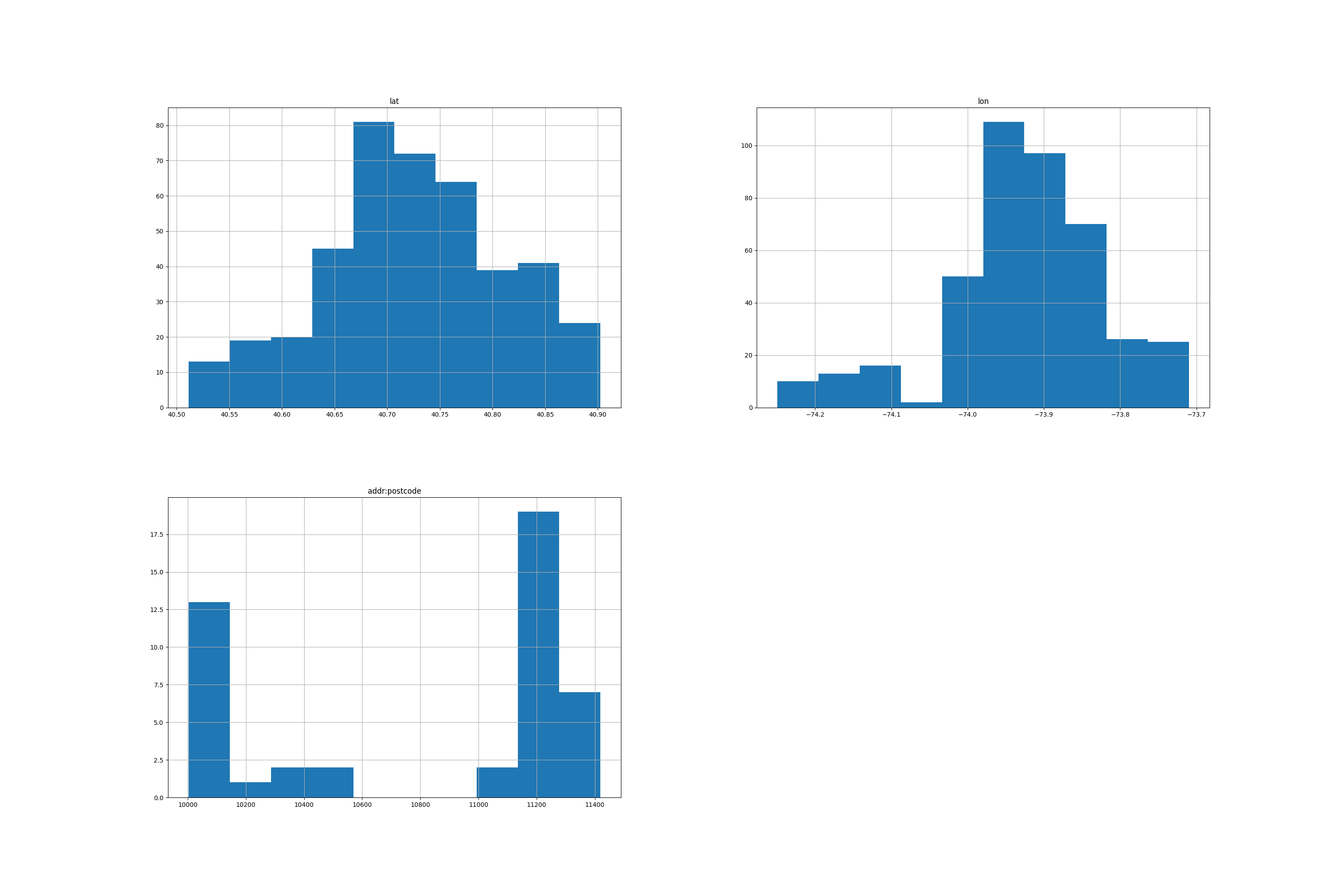
# Appendix

## Continuous Features

### Descriptive Statistics



### Histograms



## Categorical Features

### Descriptive Statistics



### Box Plots

