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.

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

1 action will be taken:

- Access
  - Any rows with access as "private" will be removed and the feature will be dropped.
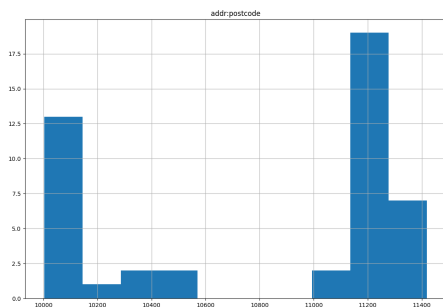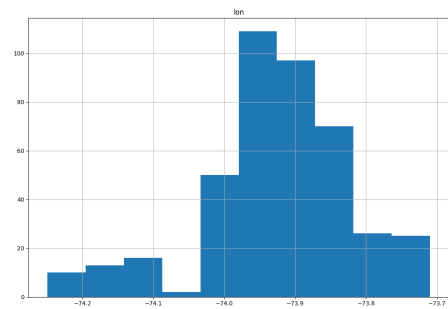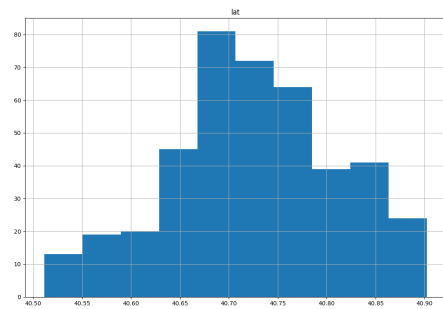

Appendix

Continuous Features

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max	%missing	card
lat	418.0	40.724541	0.086197	40.511315	40.668936	40.725117	40.782750	40.902500	0.000000	415
lon	418.0	-73.926336	0.104489	-74.249504	-73.972122	-73.921834	-73.857761	-73.709722	0.000000	410
addr:postcode	46.0	10806.173913	569.647388	10003.000000	10060.500000	11206.000000	11222.750000	11419.000000	88.995215	37

Histograms

Categorical Features

Descriptive Statistics

	count	unique	top	freq	second	second_freq	%missing	card
id	418	418	357545993	1	7788619719	1	0.000000	418
name	378	369	Chuck E. Cheese	5	Rose Garden	3	9.569378	369
addr:city	36	11	Brooklyn	13	New York	12	91.387560	11
addr:state	36	1	NY	36	None	0	91.387560	1
opening_hours	34	33	Mo-Su 18:00-04:00	2	"Temporarily closed"	1	91.866029	33
website	70	70	http://jupiterdisco.com/	1	https://www.haweedatavern.com	1	83.253589	70
addr:street	73	64	Flushing Avenue	3	3rd Avenue	3	82.535885	64
addr:housenumber	69	69	10	1	32-32	1	83.492823	69
phone	53	53	(718) 590-9000	1	+1-718-339-6977	1	87.320574	53
wheelchair	12	3	yes	8	no	3	97.129187	3
email	13	13	bookings@fiveirongolf.com	1	danceprojectofwhitestone@gmail.com	1	96.889952	13
access	21	3	yes	13	private	7	94.976077	3

Box Plots