OSM Shop Data Quality Report

Overview

This report will outline the initial data quality findings on shop 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. 13 features contain missing values.

Summary

This dataset consists of information on different shops in New York City, including charity, book, and pastry shops. The dataset has 29 features and 1933 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://

- 5 instances, all appear to be valid websites.

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

- 6 instances, all valid zip codes (10065 and 10075). This appears to be a data quality issue on the website where the valid zip codes were scraped from.

Review Continuous Features

There are 2 continuous features in this dataset:

- lat
    - This feature has a mean of 40.72, 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 -73.941, a min value of -74.249 and a max value of -73.711. 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 Categoric Features

There are 15 categoric features in this dataset:

- id
    - This has 1933 unique values. The most common is 357623896. There are no missing values.
- name
    - This has 1630 unique values. The most common is Paris Baguette. There are 140 missing values.
- opening_hours
    - This has 461 unique values. The most common is 24/7. There are 1331 missing values.
- shop
    - This has 11 unique values. The most common is deli. There are no missing values.
- website
    - This has 443 unique values. The most common is https://www.bookculture.com/. There are 1465 missing values.
- addr:city
    - This has 31 unique values. The most common is New York. There are 1552 missing values.
- addr:housenumber
    - This has 1074 unique values. The most common is 1. There are 675 missing values.
- addr:postcode
    - This has 134 unique values. The most common is 10003.0. There are 1181 missing values.
- addr:state
    - This has 1 unique values. The most common is NY. There are 1568 missing values.

- addr:street
    - This has 459 unique values. The most common is Broadway. There are 646 missing values.
- phone
    - This has 698 unique values. The most common is +1-212-633-2253. There are 1227 missing values.
- outdoor_seating
    - This has 11 unique values. The most common is yes. There are 1859 missing values.
- wheelchair
    - This has 3 unique values. The most common is yes. There are 1786 missing values.
- email
    - This has 64 unique values. The most common is 4cakes@roccospastry.com. There are 1869 missing values.
- drink:coffee
    - This has 1 unique values. The most common is yes. There are 1830 missing values.

Actions to Take

10 actions will be taken:

- Postcode
    - Change postcode type to int.

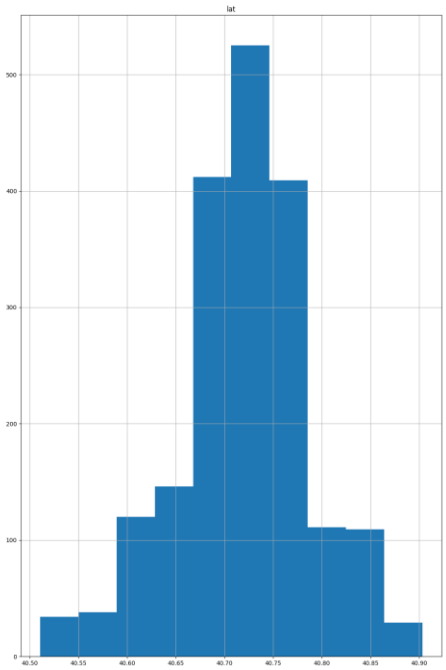
Appendix

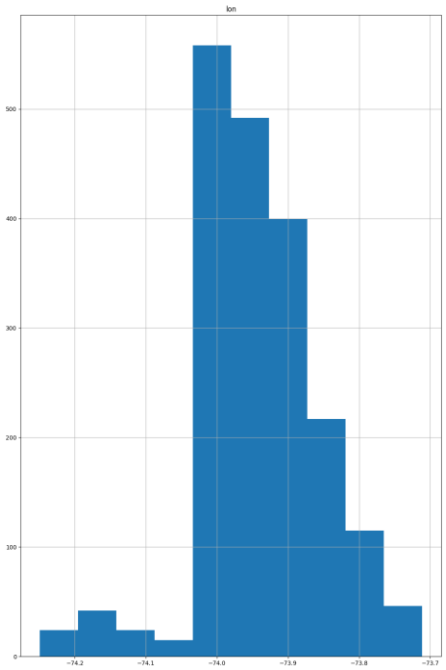
Continuous Features

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max	%missing	card
lat	1933.0	40.719953	0.067385	40.510539	40.684033	40.721509	40.758781	40.903393	0.0	1933
lon	1933.0	-73.940994	0.085146	-74.249339	-73.990213	-73.950633	-73.895647	-73.711323	0.0	1932

Histograms



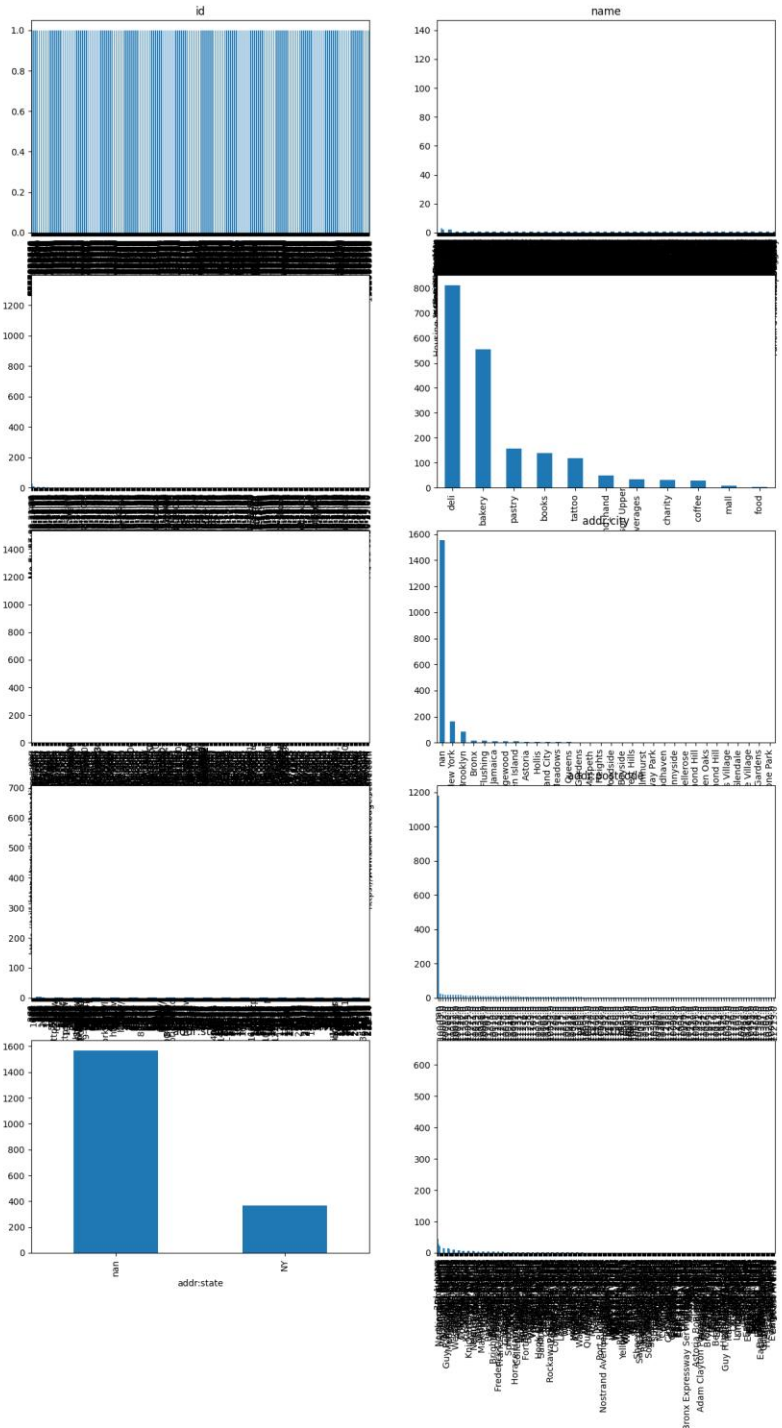


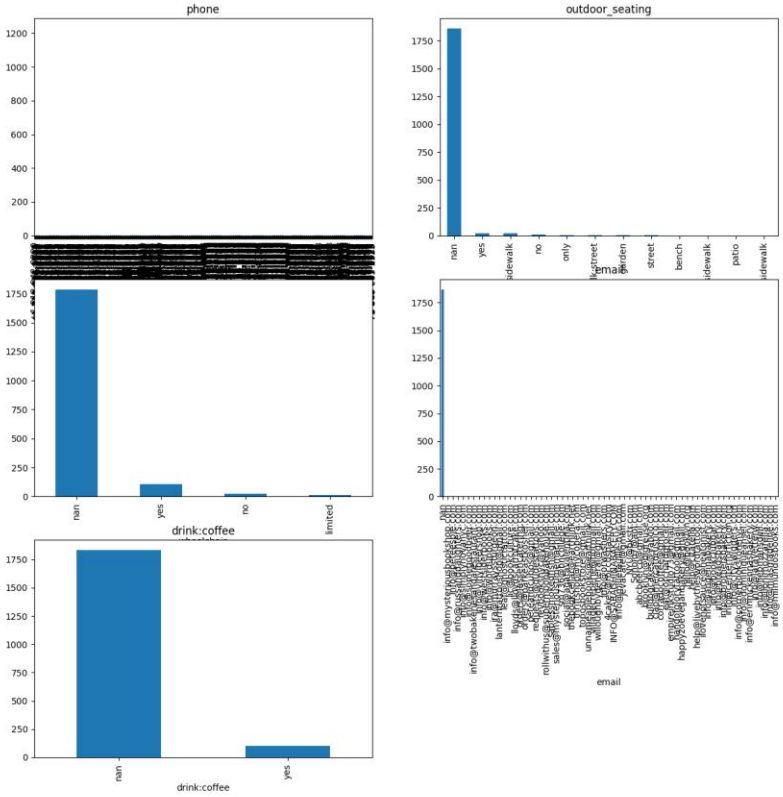
Categorical Features

Descriptive Statistics

	count	unique	top	freq	second	second_freq	%missing	card
id	1933	1933	357623896	1	8098089422	1	0.000000	1933
name	1793	1630	Paris Baguette	18	Insomnia Cookies	9	7.242628	1630
opening_hours	602	461	24/7	27	Mo-Su 07:00-21:00	10	68.856699	461
shop	1933	11	deli	812	bakery	556	0.000000	11
website	468	443	https://www.bookculture.com/	3	http://www.littlecupcakebakeshop.com	3	75.788929	443
addr:city	381	31	New York	165	Brooklyn	88	80.289705	31
addr:housenumber	1258	1074	1	5	44	5	34.919814	1074
addr:postcode	752.0	134.0	10003.0	24.0	10002.0	21	61.096741	134
addr:state	365	1	NY	365	None	0	81.117434	1
addr:street	1287	459	Broadway	45	3rd Avenue	29	33.419555	459
phone	706	698	+1-212-633-2253	2	+1 212-842-0220	2	63.476461	698
outdoor_seating	74	11	yes	22	sidewalk	19	96.171754	11
wheelchair	147	3	yes	109	no	25	92.395241	3
email	64	64	4cakes@roccospastry.com	1	INFO@SARAGHINABAKERY.COM	1	96.689084	64
drink:coffee	103	1	yes	103	None	0	94.671495	1

Box Plots


id
name
deli
bakery
pastry
books
tattoo
charity
coffee
mall
food
addr:city
addr:postcode
addr:state
nan
NY


phone
nan
yes
no
limited
outdoor_seating
nan
yes
sidewalk
no
only
street
garden
street
bench
sidewalk
patio
sidewalk
email
email
drink:coffee
nan
yes
drink:coffee