Template Data Quality Report

Overview

This report will outline the initial data quality findings on amenities data obtained from the overpass turbo API for the New York City area. 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.

Summary

This dataset consists of information on different amenities in New York City, including bars, restaurants, and cafés. There are significant amounts of missing values and most features will need to be dropped. There are no duplicate rows. Distribution of the data is consistent with expectations. While the missing data is not ideal, this data is still suitable for use as place data.

Review Logical Integrity

Test 1: No date in x is before Feb 2022 or after the upload date (22nd May).

- 0 instances.

Review Continuous Features

Descriptive Statistics

There are 3 continuous features in this dataset:

- ID
    - This is the Identification number of each place in OSM. Each row has a unique value. There are no missing values.
- Latitude
    - This is the latitude of each place in OSM. Each row has a unique value. There are no missing values. Latitude is centred around 40.72, roughly the centre of NYC.
- Longitude
    - This is the longitude of each place in OSM. Each row has a unique value. There are no missing values. Longitude is centred around -73.95, roughly the center of NYC.

Histograms

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

Review Categorical Features

There are 43 categorical features in the dataset. For brevity, only columns of interest will be listed. See the appendix for full descriptive statistics on all features.

- City
    - This has 71 unique values. The most common is New York. 66% of values are missing.
- House Number
    - This has 4269 unique values. The most common is 2. 34% of values are missing.
- State
    - This has 2 unique values. The most common is NY. The second most common (which needs to be removed) is NJ. 69% of values are missing, however these can all be imputed as NY.
- Amenity
    - The type of amenity. This has 24 unique values. The most common is restaurant. There are no missing values.
- Cuisine
    - This has 1303 unique values. The most common is pizza. 33% of values are missing.
- Name
    - This has 11435 unique values. 3% of values are missing.
- Opening Hours
    - This has 4405 unique values. The most common is 24/7. 53% of values are missing.
- Phone
    - This has 7598 unique values. There are some duplicate values, likely for large chains.
- Website
    - This has 6647 unique values. 54% of values are missing.
- Wheelchair
    - This has 6 unique values. The most common is yes. 87% of values are missing.
- Vegan
    - This has 4 unique values. The most common is yes. 96% of values are missing.
- Vegetarian
    - This has 4 unique values. The most common is yes. 96% of values are missing.
- Opening Hours
    - This has 4405 unique values. The most common is 24/7. 53% of values are missing.

Actions to take

x 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”.

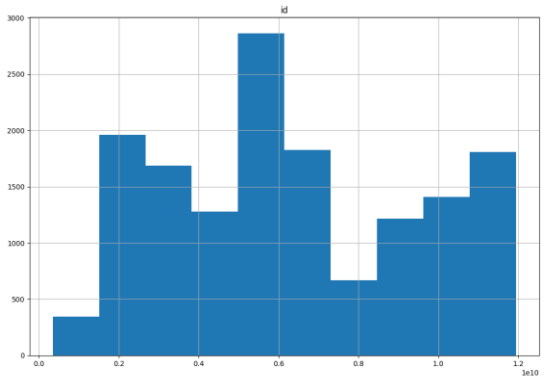
Appendix

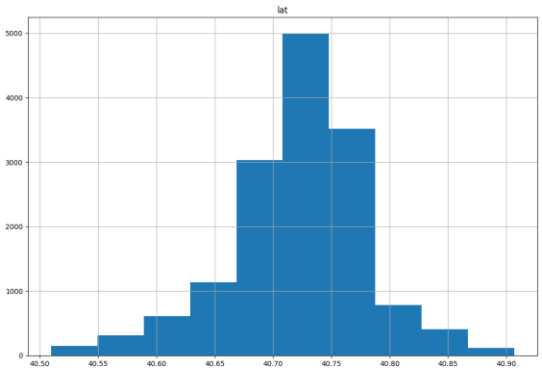
Continuous Features

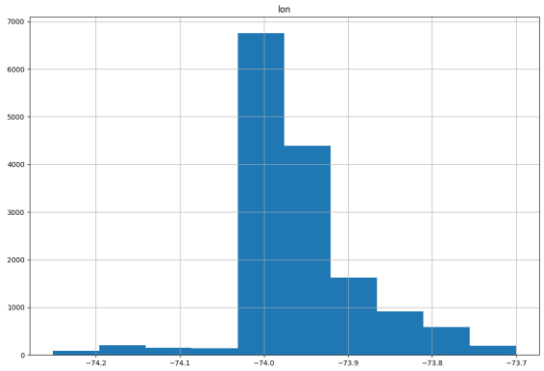
Descriptive Statistics

Table with statistics on continuous features.

Histograms







Categorical Features

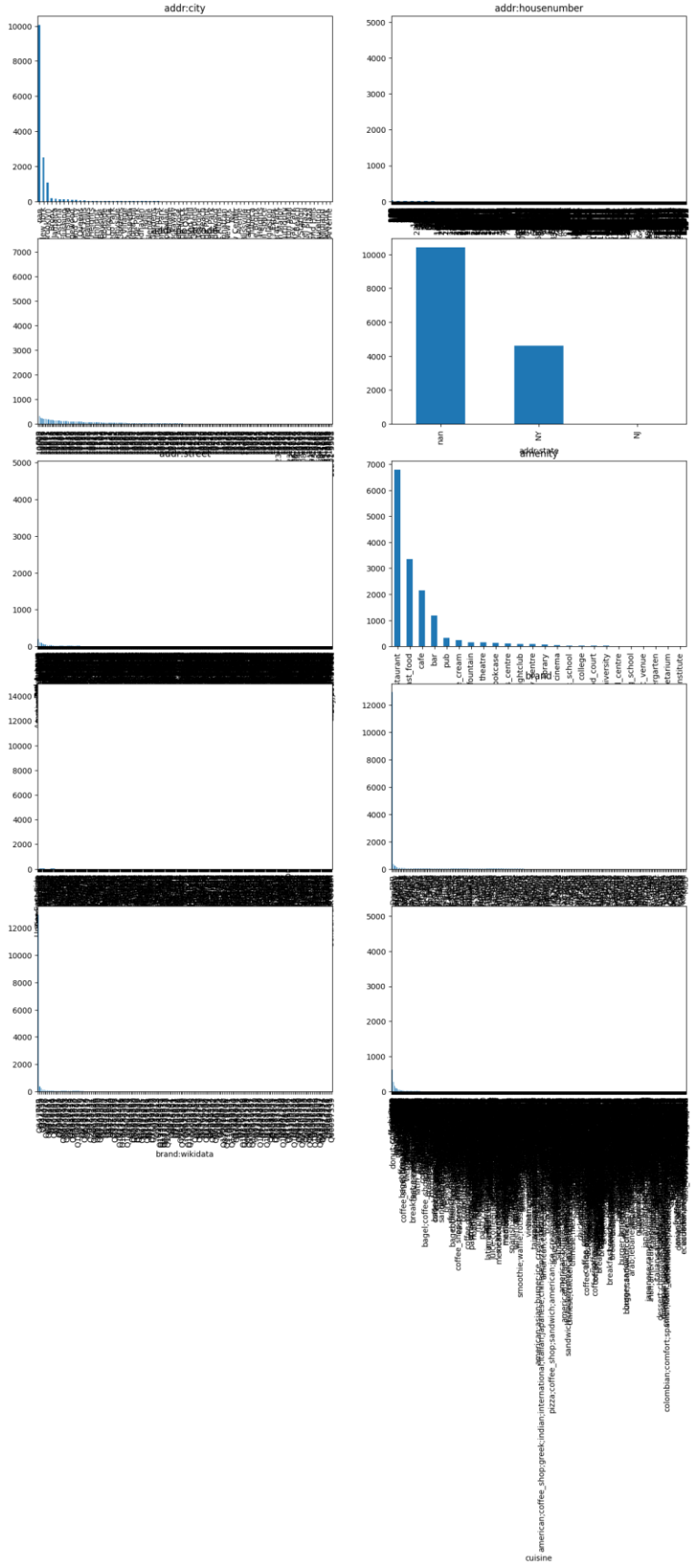
Descriptive Statistics

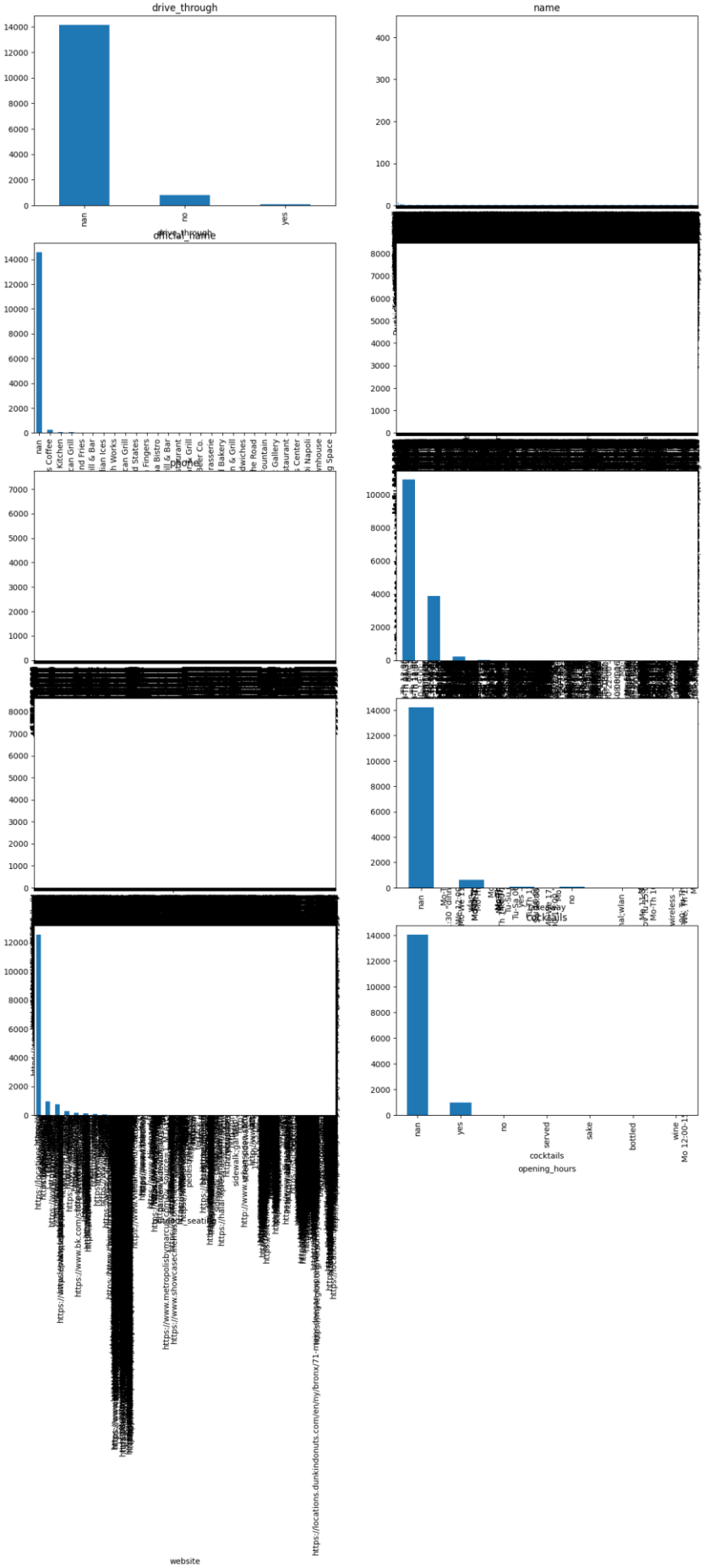
Table with statistics on categorical features.

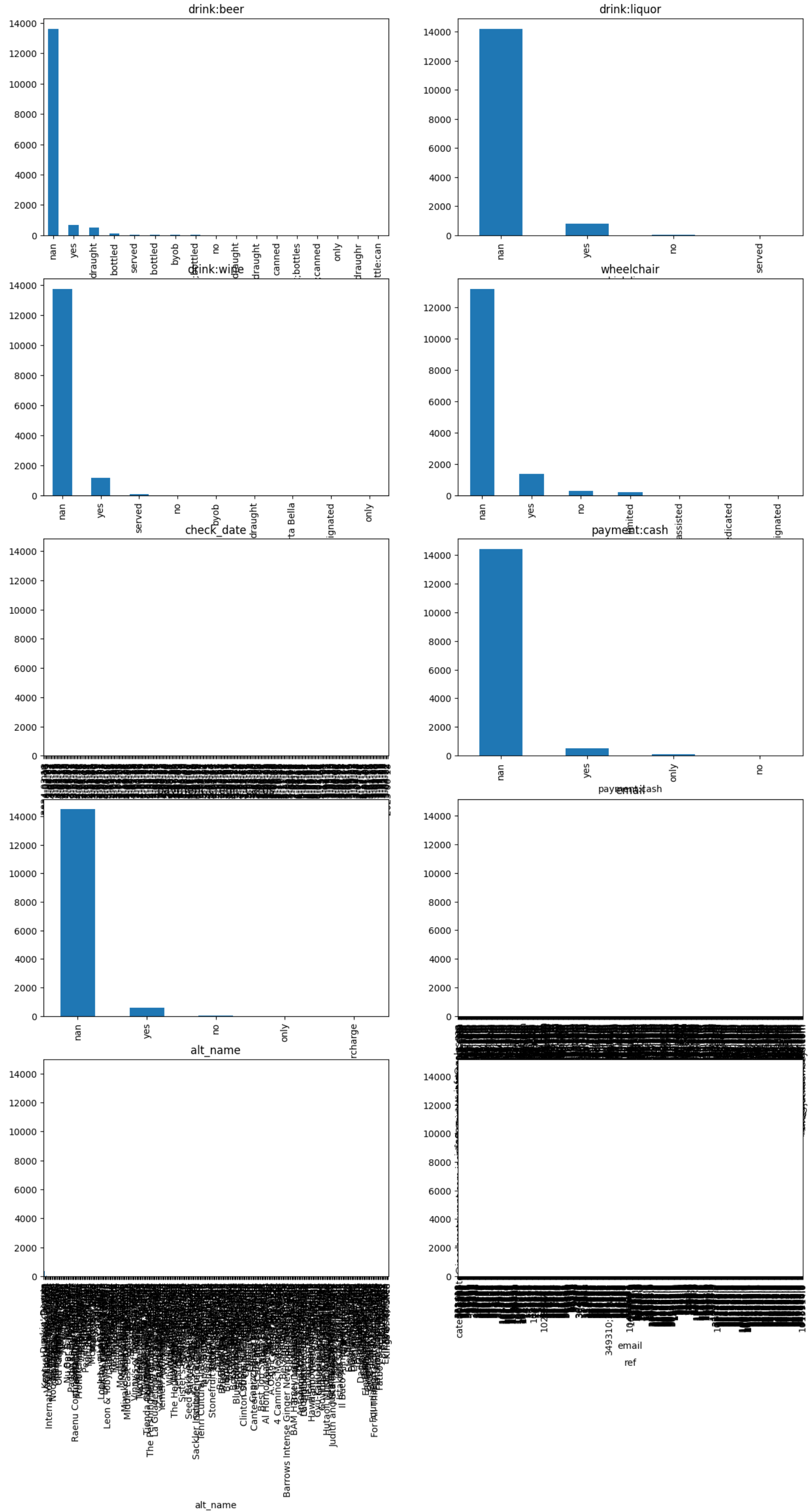
Feature	count	mean	std	min	25%	50%	75%	max	%missing	card
id	15055	6.33E+09	3.1E+09	3.49E+08	3.58E+09	5.87E+09	9.24E+09	1.19E+10	0	15055
lat	15055	40.72131	0.058308	40.50973	40.69008	40.72665	40.75634	40.90678	0	14986
lon	15055	-73.956	0.071642	-74.2514	-73.9917	-73.9739	-73.9291	-73.7003	0	14967

Feature	count	unique	top	freq	second	second_freq	%missing	card
addr:city	4991	71	New York	2493	Brooklyn	1083	66.84822	71
addr:housenumber	10134	4269	2	41	1	38	32.68682	4269
addr:postcode	7870	220	10003	327	10002	246	47.72501	220
addr:state	4633	2	NY	4632	NJ	1	69.22617	2
addr:street	10257	1243	Broadway	495	3rd Avenue	270	31.86981	1243
amenity	15055	24	restaurant	6783	fast_food	3355	0	24
branch	745	501	Upper East Side	43	Times Square	15	95.05148	501
brand	2136	212	Dunkin'	344	Starbucks	248	85.81202	212
brand:wikidata	2083	183	Q847743	346	Q37158	247	86.16407	183
cuisine	10020	1303	pizza	876	chinese	731	33.44404	1303
drive_through	905	2	no	812	yes	93	93.98871	2
name	14625	11435	Dunkin'	321	Starbucks	246	2.856194	11435
official_name	483	27	Starbucks Coffee	248	Popeyes Louisiana Kitchen	80	96.79176	27
opening_hours	7000	4405	24-Jul	207	Mo-Su 11:00-22:00	119	53.50382	4405
phone	7672	7598	-5453	7	+1 212-686-1444	4	49.04019	7598
takeaway	4156	11	yes	3892	only	230	72.39455	11
website	6856	6647	https://www.diginn.com/	17	https://www.xianfoods.com	8	54.46031	6647
internet_access	819	5	wlan	625	yes	100	94.55995	5
outdoor_seating	2520	31	yes	953	no	764	83.26137	31
cocktails	1004	6	yes	990	no	7	93.33112	6
drink:beer	1429	16	yes	694	draught	516	90.50814	16
drink:liquor	837	3	yes	810	no	26	94.44039	3
drink:wine	1309	8	yes	1167	served	107	91.30521	8
wheelchair	1883	6	yes	1373	no	297	87.49253	6
check_date	916	314	13/03/2024	33	26/04/2024	22	93.91564	314
payment:cash	658	3	yes	529	only	105	95.62936	3
payment:credit_cards	574	4	yes	557	no	15	96.18731	4
email	600	595	info@goles.org	3	cateringquote@jacobrestaurant.com;jacob373@jacobrestaurant.com	2	96.01461	595
alt_name	780	378	Dunkin' Donuts	345	Dig Inn	21	94.819	378
ref	586	586	3	1	331314	1	96.10761	586
level	979	32	0	770	1	81	93.49718	32
diet:vegan	484	4	yes	392	only	51	96.78512	4
diet:vegetarian	617	4	yes	570	only	25	95.90169	4
toilets	734	3	yes	642	no	77	95.12454	3
drink:coffee	974	4	yes	931	served	23	93.53039	4
drink:tea	478	4	yes	472	served	3	96.82498	4
smoking	521	7	no	454	outside	41	96.53936	7
bar	493	4	yes	425	no	58	96.72534	4
contact:instagram	461	452	https://instagram.com/vanleeuwenicecream/	4	https://www.instagram.com/courtstreetgrocers	4	96.93789	452
delivery	2144	45	yes	1991	no	107	85.75888	45
drink:espresso	519	2	yes	517	served	2	96.55264	2
reservation	622	11	yes	480	recommended	68	95.86848	11
indoor_seating	646	3	yes	599	no	46	95.70907	3

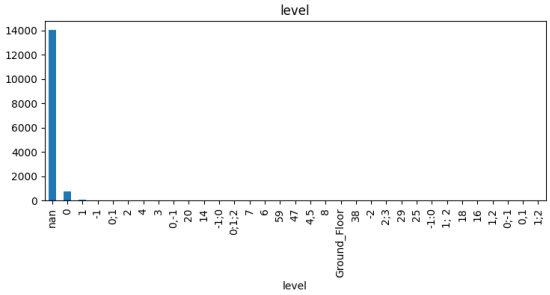
Box Plots

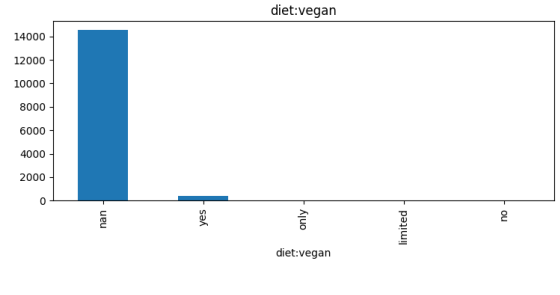

addr:city
addr:housenumber
addr:postcode
addr:state
addr:street
amenity
brand
brand:wikidata
cuisine

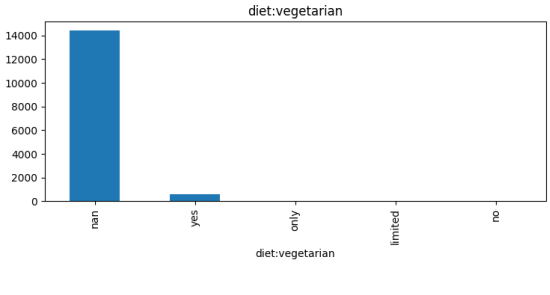

drive_through
name
official_name
phone
website
cocktails
opening_hours

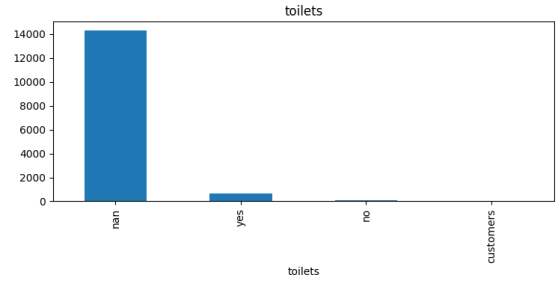


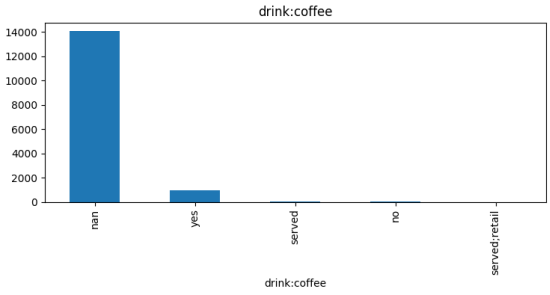
drink:beer
drink:liquor
drink:wine
wheelchair
check_date
payment:cash
email
alt_name
ref
0
2000
4000
6000
8000
10000
12000
14000
nan
yes
draught
bottled
served
byob
no
canned
only
assisted

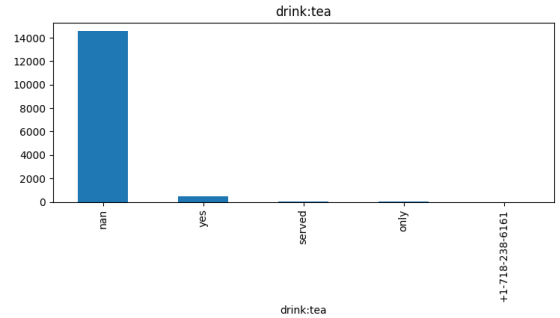

level
nan
0
1
-1
0;1
2
4
3
0;-1
20
14
-1;0
0;1;2
7
6
59
47
4;5
8
Ground_Floor
38
-2
2;3
29
25
-1;0
1;-2
18
16
1;2
0;-1
0;1
1;2
level

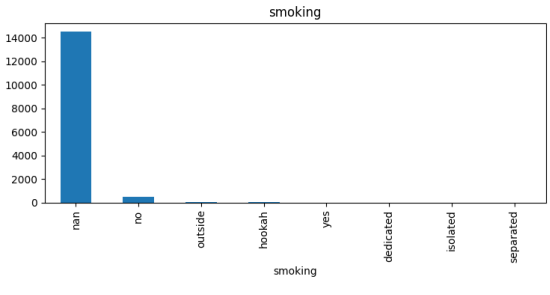

diet:vegan
nan
yes
only
limited
no
diet:vegan

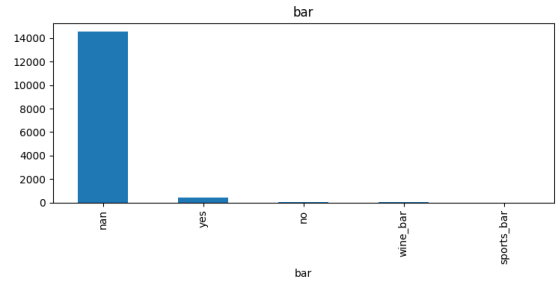

diet:vegetarian
nan
yes
only
limited
no
diet:vegetarian

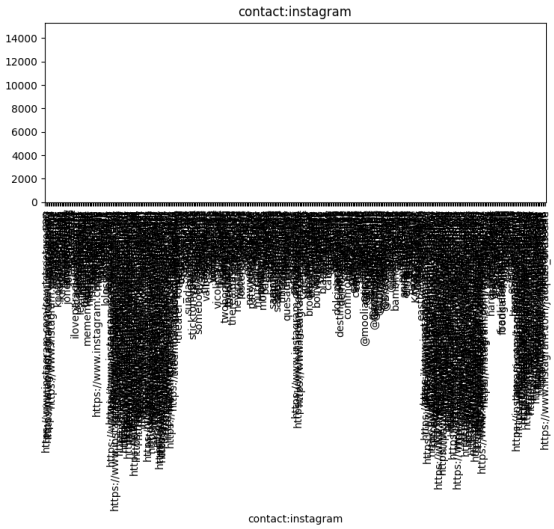

toilets
nan
yes
no
customers
toilets

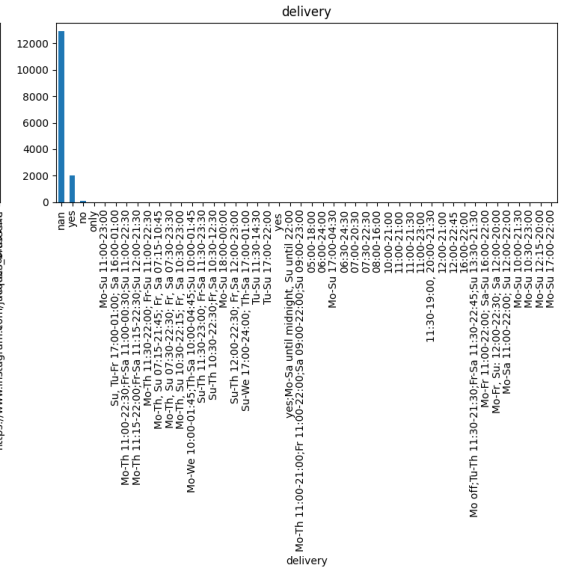

drink:coffee
nan
yes
served
no
served;retail
drink:coffee

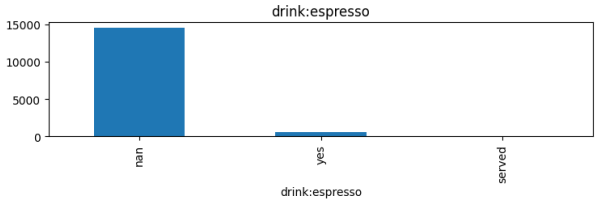

drink:tea
nan
yes
served
only
+1-718-238-6161
drink:tea

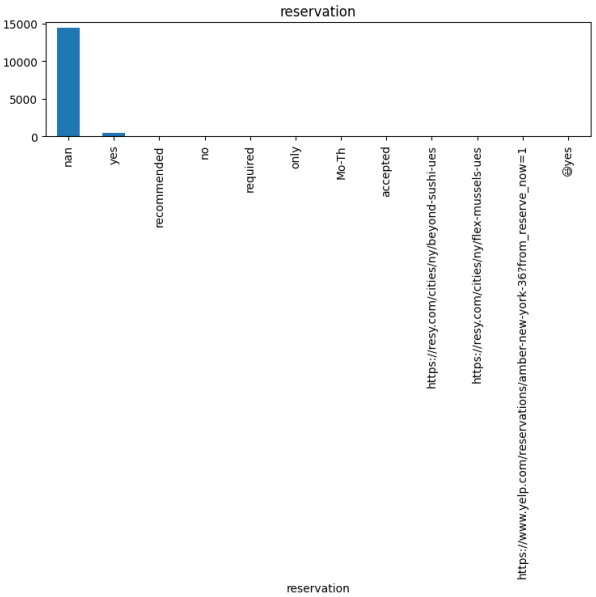

smoking
nan
no
outside
hookah
yes
dedicated
isolated
separated
smoking

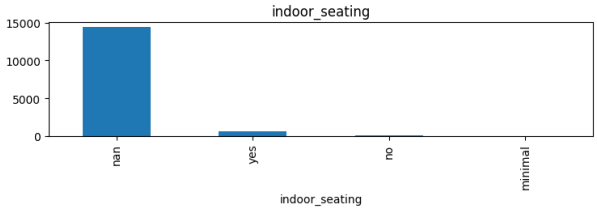

bar
nan
yes
no
wine_bar
sports_bar
bar


contact:instagram
contact:instagram


delivery
nan
yes
no
only
Mo-Su 11:00-23:00
Su,Th-Fr 17:00-01:00; Sa 16:00-21:00
Mo-Th 11:00-22:30; Fr-Sa 11:00-01:30
Mo-Th 11:15-22:00; Fr-Sa 11:15-22:30; Su 12:00-21:30
Mo-Th 11:30-22:00; Fr-Su 11:00-22:30
Mo-Th, Su 07:15-21:45; Fr, Sa 07:15-10:45
Mo-Th, Su 07:30-23:00; Fr, Sa 10:30-23:00
Mo-We 10:00-01:45;Th-Sa 10:00-04:45;Su 10:00-01:45
Su-Th 11:30-23:00; Fr-Sa 11:30-23:30
Su-Th 10:30-22:30; Fr-Sa 10:30-12:30
Mo-Su 18:00-00:00
Su-Th 12:00-22:30; Fr,Sa 12:00-23:00
Su-We 17:00-24:00; Th-Sa 17:00-01:00
Tu-Su 11:30-14:30
Tu-Su 17:00-22:00
yes
yes;Mo-Sa until midnight, Su until 22:00
Mo-Th 11:00-21:00;Fr 11:00-22:00;Sa 09:00-22:00;Su 09:00-23:00
05:00-18:00
06:00-04:00
Mo-Su 17:00-04:30
06:30-24:30
07:00-20:30
08:30-16:30
10:00-21:00
11:00-21:30
11:00-00:00
11:30-19:00, 20:00-21:30
12:00-21:00
12:00-22:45
13:30-21:30
Mo off;Tu-Th 11:30-21:30;Fr-Sa 11:30-22:45;Su 13:30-21:30
Mo-Fr 11:00-22:00; Sa-Su 16:00-22:00
Mo-Fr, Su: 12:00-22:30; Sa 12:00-20:00
Mo-Sa 11:00-22:00; Su 12:00-22:00
Mo-Su 10:00-21:30
Mo-Su 10:30-23:00
Mo-Su 12:15-20:00
Mo-Su 17:00-22:00
delivery


drink:espresso
15000
10000
5000
0
nan
yes
served
drink:espresso


reservation
15000
10000
5000
0
nan
yes
recommended
no
required
only
Mo-Th
accepted
https://resy.com/cities/ny/beyond-sushi-ues
https://resy.com/cities/ny/flex-mussels-ues
https://www.yelp.com/reservations/amber-new-york-36?from_reserve_now=1
@yes
reservation


indoor_seating
15000
10000
5000
0
nan
yes
no
minimal
indoor_seating