AIRBNB PRICE PREDICTION

ANALYSIS

GUIDED BY: BHAVIK GANDHI

MOHNISH LAVANIA
RISHIK SHARMA
SARANG BAGUL
ATUL CHOUDHARY
RAHUL SINGH

EDA

Overview

Dataset info

Number of variables	89
Number of observations	55000
Missing cells	862480 (17.6%)
Duplicate rows	0 (0.0%)
Total size in memory	379.6 MiB
Average record size in memory	7.1 KiB

Variables types

CAT	45
NUM	36
URL	8

Toggle Reproduction Information
Toggle Warnings

Warnings

|		
`Access` has 24948 (45.4%) missing values		Missing
`Access` has a high cardinality: 27974 distinct values		Warning
`Amenities` has a high cardinality: 50263 distinct values		Warning
`Availability_30` has 19279 (35.1%) zeros		Zeros

`Availability_365` has 10894 (19.8%) zeros `Zeros`

`Availability_60` has 15957 (29.0%) zeros `Zeros`

`Availability_90` has 14030 (25.5%) zeros `Zeros`

`Bedrooms` has 4090 (7.4%) zeros `Zeros`

`Calendar_last_Scraped` only contains datetime values, but is categorical. Consider applying `pd.to_datetime()` `Type`

`Calendar_Updated` has a high cardinality: 72 distinct values `Warning`

`City` has a high cardinality: 1918 distinct values `Warning`

`Cleaning_Fee` has 19748 (35.9%) missing values `Missing`

`Description` has a high cardinality: 54670 distinct values `Warning`

`Extra_People` has 32494 (59.1%) zeros `Zeros`

`Features` has a high cardinality: 119 distinct values `Warning`

`First_Review` has 13500 (24.5%) missing values `Missing`

`First_Review` has a high cardinality: 2182 distinct values `Warning`

`Geolocation` has a high cardinality: 55000 distinct values `Warning`

`Has_Availability` has 54043 (98.3%) missing values `Missing`

`Host_About` has 21823 (39.7%) missing values `Missing`

`Host_About` has a high cardinality: 28594 distinct values `Warning`

`Host_Acceptance_Rate` has 50666 (92.1%) missing values `Missing`

`Host_Acceptance_Rate` has a high cardinality: 78 distinct values `Warning`

`Host_Location` has a high cardinality: 3187 distinct values `Warning`

`Host_Name` has a high cardinality: 14827 distinct values `Warning`

`Host_Neighbourhood` has 14016 (25.5%) missing values `Missing`

`Host_Neighbourhood` has a high cardinality: 2007 distinct values `Warning`

`Host_Response_Rate` has 12610 (22.9%) missing values `Missing`

`Host_Response_Rate` has 723 (1.3%) zeros `Zeros`

`Host_Response_Time` has 12610 (22.9%) missing values `Missing`

`Host_Since` has a high cardinality: 2770 distinct values	**Warning**
`Host_Verifications` has a high cardinality: 542 distinct values	**Warning**
`House_Rules` has 21668 (39.4%) missing values	**Missing**
`House_Rules` has a high cardinality: 31242 distinct values	**Warning**
`Interaction` has 26278 (47.8%) missing values	**Missing**
`Interaction` has a high cardinality: 26891 distinct values	**Warning**
`Jurisdiction_Names` has 39423 (71.7%) missing values	**Missing**
`Last_Review` has 13492 (24.5%) missing values	**Missing**
`Last_Review` has a high cardinality: 1198 distinct values	**Warning**
`Last_Scraped` only contains datetime values, but is categorical. Consider applying `pd.to_datetime()`	**Type**
`License` has 53367 (97.0%) missing values	**Missing**
`License` has a high cardinality: 1153 distinct values	**Warning**
`Market` has 653 (1.2%) missing values	**Missing**
`Market` has a high cardinality: 77 distinct values	**Warning**
`Maximum_Nights` is highly skewed ($\gamma_1 = 88.62314012$)	**Skewed**
`Medium_Url` has 10592 (19.3%) missing values	**Missing**
`Minimum_Nights` is highly skewed ($\gamma_1 = 160.5059403$)	**Skewed**
`Monthly_Price` has 44359 (80.7%) missing values	**Missing**
`Name` has a high cardinality: 54535 distinct values	**Warning**
`Neighborhood_Overview` has 22457 (40.8%) missing values	**Missing**
`Neighborhood_Overview` has a high cardinality: 31587 distinct values	**Warning**
`Neighbourhood` has 16867 (30.7%) missing values	**Missing**
`Neighbourhood` has a high cardinality: 1887 distinct values	**Warning**
`Neighbourhood_Cleansed` has a high cardinality: 1932 distinct values	**Warning**
`Neighbourhood_Group_Cleansed` has 43569 (79.2%) missing values	**Missing**
`Neighbourhood_Group_Cleansed` has a high cardinality: 68 distinct values	**Warning**
	Missing

`Notes` has 33022 (60.0%) missing values

`Notes` has a high cardinality: 20728 distinct values **Warning**

`Number_of_Reviews` has 13478 (24.5%) zeros **Zeros**

`Price` has 772 (1.4%) missing values **Missing**

`Review_Scores_Accuracy` has 14181 (25.8%) missing values **Missing**

`Review_Scores_Checkin` has 14227 (25.9%) missing values **Missing**

`Review_Scores_Cleanliness` has 14148 (25.7%) missing values **Missing**

`Review_Scores_Communication` has 14165 (25.8%) missing values **Missing**

`Review_Scores_Location` has 14222 (25.9%) missing values **Missing**

`Review_Scores_Rating` has 14076 (25.6%) missing values **Missing**

`Review_Scores_Value` has 14234 (25.9%) missing values **Missing**

`Reviews_per_Month` has 13500 (24.5%) missing values **Missing**

`Security_Deposit` has 31857 (57.9%) missing values **Missing**

`Smart_Location` has a high cardinality: 2045 distinct values **Warning**

`Space` has 16278 (29.6%) missing values **Missing**

`Space` has a high cardinality: 38324 distinct values **Warning**

`Square_Feet` has 53583 (97.4%) missing values **Missing**

`State` has 5307 (9.6%) missing values **Missing**

`State` has a high cardinality: 515 distinct values **Warning**

`Street` has a high cardinality: 22001 distinct values **Warning**

`Summary` has 1853 (3.4%) missing values **Missing**

`Summary` has a high cardinality: 52505 distinct values **Warning**

`Thumbnail_Url` has 10592 (19.3%) missing values **Missing**

`Transit` has 21214 (38.6%) missing values **Missing**

`Transit` has a high cardinality: 32971 distinct values **Warning**

`Weekly_Price` has 44072 (80.1%) missing values **Missing**

`XL_Picture_Url` has 10592 (19.3%) missing values **Missing**

`Zipcode` has 2127 (3.9%) missing values · **Missing**

`Zipcode` has a high cardinality: 11392 distinct values · **Warning**

`Availability_60` is highly correlated with `Availability_30` and 1 other fields · **High Correlation**

`Availability_30` is highly correlated with `Availability_60` · **High Correlation**

`Availability_90` is highly correlated with `Availability_60` · **High Correlation**

`Host_Total_Listings_Count` is highly correlated with `Host_Listings_Count` · **High Correlation**

`Host_Listings_Count` is highly correlated with `Host_Total_Listings_Count` · **High Correlation**

`Weekly_Price` is highly correlated with `Monthly_Price` · **High Correlation**

`Monthly_Price` is highly correlated with `Weekly_Price` · **High Correlation**

Variables

Access

Categorical

`MISSING`
`HIGH CARDINALITY`

Distinct count	27974
Unique (%)	50.9%
Missing	24948
Missing (%)	45.4%
Memory size	859.4 KiB

You have access to th…	94
Guests will have acces…	61
Everything	35
The house and its facili…	33
Entire apartment	29
Other values (27968)	29800

Toggle details

Accommodates

Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	17	**Mean**	3.333775846
Unique (%)	< 0.1%	**Minimum**	1.0
		Maximum	16.0
Missing	11	**Zeros**	0
Missing (%)	< 0.1%	**Zeros (%)**	0.0%
Infinite	0	**Memory size**	859.4 KiB
Infinite (%)	0.0%		
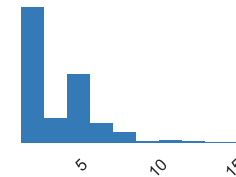
Toggle details

Amenities

Categorical

`HIGH CARDINALITY`

Distinct count	50263	translation missing: en....	86
		Family/kid friendly	52
Unique (%)	91.4%	Family/kid friendly,tran...	28
Missing	487	Internet,Wireless Inter...	27
Missing (%)	0.9%	Internet,Wireless Inter...	21
Memory size	859.4 KiB	Other values (50257)	54299

Toggle details

Availability_30

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS
HIGH CORRELATION

Distinct count	31	Mean	9.871618182
Unique (%)	0.1%	Minimum	0
Missing	0	Maximum	30
Missing (%)	0.0%	Zeros	19279
Infinite	0	Zeros (%)	35.1%
Infinite (%)	0.0%	Memory size	859.4 KiB
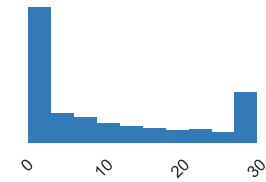
Toggle details

Availability_365

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct count	366	Mean	165.9412909
Unique (%)	0.7%	Minimum	0
Missing	0	Maximum	365
Missing (%)	0.0%	Zeros	10894
Infinite	0	Zeros (%)	19.8%
Infinite (%)	0.0%	Memory size	859.4 KiB
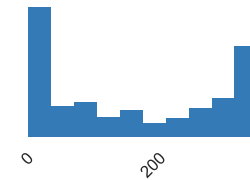
Toggle details

Availability_60

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS
HIGH_CORRELATION

Distinct count	61	Mean	22.81012727
Unique (%)	0.1%	Minimum	0
Missing	0	Maximum	60
Missing (%)	0.0%	Zeros	15957
Infinite	0	Zeros (%)	29.0%
Infinite (%)	0.0%	Memory size	859.4 KiB
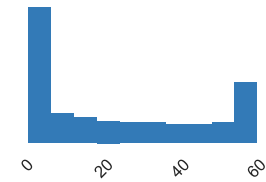
Toggle details

Availability_90

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS
HIGH_CORRELATION

Distinct count	91	Mean	37.69498182
Unique (%)	0.2%	Minimum	0
Missing	0	Maximum	90
Missing (%)	0.0%	Zeros	14030
Infinite	0	Zeros (%)	25.5%
Infinite (%)	0.0%	Memory size	859.4 KiB
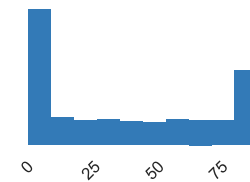
Toggle details

Bathrooms

Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	18	Mean	1.252484999
		Minimum	0.0

Unique (%)	< 0.1%	Maximum	8.0	
Missing	171	Zeros	149	
Missing (%)	0.3%	Zeros (%)	0.3%	
Infinite	0	Memory size	859.4 KiB	
Infinite (%)	0.0%			
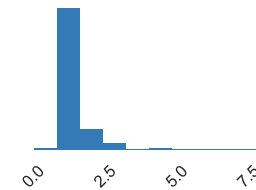
Toggle details

Bed_Type
Categorical

Distinct count	5
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	859.4 KiB

Real Bed	53182
Pull-out Sofa	1069
Futon	379
Couch	232
Airbed	138

Toggle details

Bedrooms
Real number ($\mathbb{R}_{\geq 0}$)

`ZEROS`

Distinct count	12	Mean	1.387498407	
Unique (%)	< 0.1%	Minimum	0.0	
		Maximum	10.0	
		Zeros	4090	
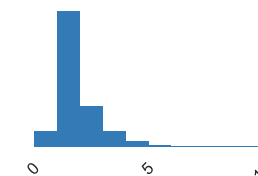
Missing	63	**Zeros (%)**	7.4%
Missing (%)	0.1%	**Memory size**	859.4 KiB
Infinite	0		
Infinite (%)	0.0%		

Toggle details

Beds
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	18	**Mean**	1.95422253	
Unique (%)	< 0.1%	**Minimum**	0.0	
		Maximum	16.0	
Missing	104	**Zeros**	1	
Missing (%)	0.2%	**Zeros (%)**	< 0.1%	
Infinite	0	**Memory size**	859.4 KiB	
Infinite (%)	0.0%			
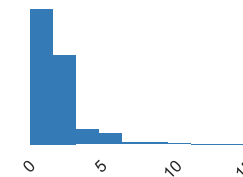
Toggle details

Calculated_ho…
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	113	**Mean**	6.541202186
Unique (%)	0.2%	**Minimum**	1.0
Missing	100	**Maximum**	752.0
Missing (%)	0.2%	**Zeros**	0
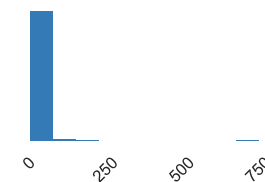
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	859.4 KiB

Toggle details

Calendar_last...
Categorical

TYPE DATE

Distinct count	49
Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	859.4 KiB

2017-05-03	4855
2017-04-02	4073
2017-05-09	4017
2017-04-05	3768
2017-04-08	3600
Other values (44)	34687

Toggle details

Calendar_Upd...
Categorical

HIGH CARDINALITY

Distinct count	72
Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.3 MiB

today	10446
a week ago	4394
2 weeks ago	4292
yesterday	3497
3 weeks ago	2682
Other values (67)	29689

Cancellation_...
Categorical

Distinct count	11
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.3 MiB

strict	20506
flexible	17207
moderate	13831
moderate_new	1166
strict_new	1045
Other values (6)	1245
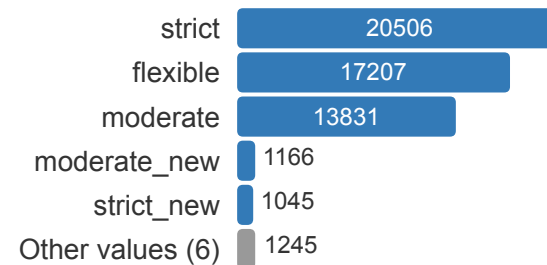
City
Categorical

`HIGH CARDINALITY`

Distinct count	1918
Unique (%)	3.5%
Missing	54
Missing (%)	0.1%
Memory size	859.4 KiB

Paris	5800
London	5115
Barcelona	3119
Amsterdam	3060
Los Angeles	2096
Other values (1912)	35756
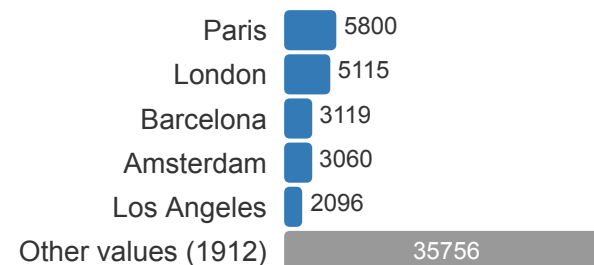
Cleaning_Fee
Real number (ℝ≥0)

MISSING

Distinct count	320
Unique (%)	0.6%
Missing	19748
Missing (%)	35.9%
Infinite	0
Infinite (%)	0.0%

Mean	59.47517871
Minimum	0.0
Maximum	950.0
Zeros	33
Zeros (%)	0.1%
Memory size	859.4 KiB
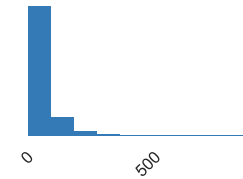
Toggle details

Country
Categorical

Distinct count	19
Unique (%)	< 0.1%
Missing	1
Missing (%)	< 0.1%
Memory size	3.3 MiB

United States	14676
United Kingdom	6518
Spain	6133
France	6015
Australia	4418
Other values (13)	17239

Toggle details

Country_Code
Categorical

Distinct count	19
Unique (%)	< 0.1%

US	14676
GB	6518
ES	6133

			FR	6015
Missing	0		AU	4418
Missing (%)	0.0%		Other values (14)	17240
Memory size	859.4 KiB			

Toggle details

Description
Categorical

`HIGH CARDINALITY`

Distinct count	54670		My place is good for co…	13
			brand new Sleepin…	12
Unique (%)	99.4%		My place is good for co…	11
			Mon logement est parf…	10
Missing	24		**brand new** Sleepin…	8
Missing (%)	< 0.1%		Other values (54664)	54922
Memory size	859.4 KiB			

Toggle details

Experiences_…
Categorical

Distinct count	5		none	54642
			business	138
Unique (%)	< 0.1%		family	93
			social	85
Missing	0		romantic	42
Missing (%)	0.0%			
Memory	859.4			

size	KiB

Toggle details

Extra_People
Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct count	164
Unique (%)	0.3%
Missing	3
Missing (%)	< 0.1%
Infinite	0
Infinite (%)	0.0%

Mean	11.4829718
Minimum	0.0
Maximum	999.0
Zeros	32494
Zeros (%)	59.1%
Memory size	859.4 KiB
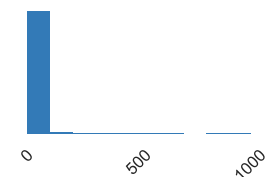
Toggle details

Features
Categorical

HIGH CARDINALITY

Distinct count	119
Unique (%)	0.2%
Missing	36
Missing (%)	0.1%
Memory size	859.4 KiB

Host Has Profile Pic,H…	13997
Host Has Profile Pic,Is …	9666
Host Has Profile Pic,H…	4156
Host Has Profile Pic,H…	3980
Host Has Profile Pic	3450
Other values (113)	19715

First_Review
Categorical

MISSING

HIGH CARDINALITY

Distinct count	2182
Unique (%)	4.0%
Missing	13500
Missing (%)	24.5%
Memory size	859.4 KiB

2017-01-01	256
2017-01-02	208
2016-01-03	187
2016-01-02	185
2016-01-05	170
Other values (2176)	40494

Geolocation
Categorical

UNIQUE

HIGH CARDINALITY

Distinct count	55000
Unique (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	3.3 MiB

41.3787907645,2.169…	1
49.2647979661,-123.1…	1
46.2012833839,6.130…	1
51.5018704962,-0.101…	1
32.7825202313,-117.2…	1
Other values (54995)	54995

Guests_Includ…
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	18	**Mean**	1.563763636
Unique (%)	< 0.1%	**Minimum**	0
Missing	0	**Maximum**	20
Missing (%)	0.0%	**Zeros**	187
Infinite	0	**Zeros (%)**	0.3%
Infinite (%)	0.0%	**Memory size**	3.3 MiB
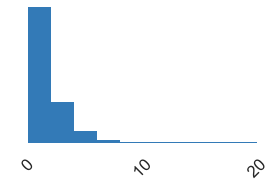
Toggle details

Has_Availability
Categorical

`MISSING`

Distinct count	2
Unique (%)	< 0.1%
Missing	54043
Missing (%)	98.3%
Memory size	3.3 MiB

t   957

Toggle details

Host_About
Categorical

`MISSING`
`HIGH CARDINALITY`

Distinct count	28594
Unique (%)	52.0%
Missing	21823
Missing (%)	39.7%
Memory size	3.3 MiB

.	100
My name is Angela an…	72
Greetings from London…	65
lamb&b is taking care …	43
…	37
Other values (28588)	32860

Toggle details

Host_Accepta…
Categorical

`MISSING`
`HIGH CARDINALITY`

Distinct count	78
Unique (%)	0.1%
Missing	50666
Missing (%)	92.1%
Memory size	3.3 MiB

100%	2408
0%	95
50%	92
67%	91
96%	88
Other values (72)	1560

Toggle details

Host_ID
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	49743
Unique (%)	90.4%

Mean	32225203.1
Minimum	678
Maximum	134820813

	Missing	0	**Zeros**	0
	Missing (%)	0.0%	**Zeros (%)**	0.0%
	Infinite	0	**Memory size**	3.3 MiB
	Infinite (%)	0.0%		
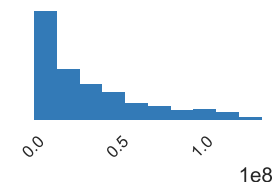
Toggle details

Host_Listings…
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct count	187	**Mean**	9.110886656
Unique (%)	0.3%	**Minimum**	0.0
Missing	52	**Maximum**	959.0
Missing (%)	0.1%	**Zeros**	50
Infinite	0	**Zeros (%)**	0.1%
Infinite (%)	0.0%	**Memory size**	859.4 KiB
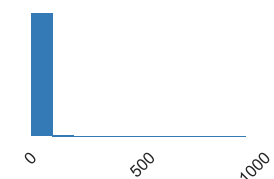
Toggle details

Host_Location
Categorical

HIGH CARDINALITY

Distinct count	**3187**
Unique (%)	5.8%
Missing	257
Missing (%)	0.5%

Paris, Île-de-France, Fr…	4759
New York, New York, …	3402
London, England, Unit…	2456
Rome, Lazio, Italy	2103
Barcelona, Catalonia, …	2076
Other values (3181)	39947

Memory size	3.3 MiB	

Toggle details

Host_Name
Categorical

`HIGH CARDINALITY`

Distinct count	14827
Unique (%)	27.0%
Missing	52
Missing (%)	0.1%
Memory size	3.3 MiB

David	340
Sarah	263
Michael	260
Laura	258
Maria	249
Other values (14821)	53578

Toggle details

Host_Neighbo…
Categorical

`MISSING`

`HIGH CARDINALITY`

Distinct count	2007
Unique (%)	3.6%
Missing	14016
Missing (%)	25.5%
Memory size	3.3 MiB

Oud-West	462
Montmartre	375
Williamsburg	344
République	299
Buttes-Chaumont - Bel…	281
Other values (2001)	39223

Toggle details

Host_Picture_…
URL

Distinct count	49538
Unique (%)	90.1%
Missing	52
Missing (%)	0.1%
Memory size	3.3 MiB

https://a0.muscache.co…	89
https://a0.muscache.co…	86
https://a0.muscache.co…	72
https://a0.muscache.co…	65
https://a0.muscache.co…	24
Other values (49532)	54612
(Missing)	52

Toggle details

Host_Respon…
Real number ($\mathbb{R}_{\geq 0}$)

MISSING

ZEROS

Distinct count	91
Unique (%)	0.2%
Missing	12610
Missing (%)	22.9%
Infinite	0
Infinite (%)	0.0%

Mean	93.59072895
Minimum	0.0
Maximum	100.0
Zeros	723
Zeros (%)	1.3%
Memory size	3.3 MiB
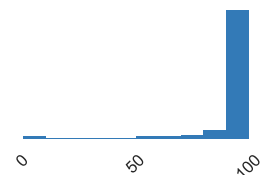
Toggle details

Host_Respon…
Categorical

MISSING

Distinct count	5
Unique (%)	< 0.1%
Missing	12610
Missing (%)	22.9%
Memory size	3.3 MiB

within an hour	22500
within a few hours	10666
within a day	7894
a few days or more	1330

Toggle details

Host_Since
Categorical

HIGH CARDINALITY

Distinct count	2770
Unique (%)	5.0%
Missing	52
Missing (%)	0.1%
Memory size	3.3 MiB

2015-05-21	100
2012-09-20	99
2015-03-30	71
2016-01-12	68
2015-04-13	64
Other values (2764)	54546

Toggle details

Host_Thumbn…
URL

Distinct count	49538
Unique (%)	90.1%

https://a0.muscache.co…	89
https://a0.muscache.co…	86
https://a0.muscache.co…	72
https://a0.muscache.co…	65

	Missing	52		https://a0.muscache.co…	24	
	Missing (%)	0.1%		Other values (49532)	54612	
	Memory size	3.3 MiB		(Missing)	52	

Toggle details

Host_Total_Li…
Real number ($\mathbb{R}_{\geq 0}$)

HIGH_CORRELATION

Distinct count	187	**Mean**	9.110886656	
Unique (%)	0.3%	**Minimum**	0.0	
Missing	52	**Maximum**	959.0	
Missing (%)	0.1%	**Zeros**	50	
Infinite	0	**Zeros (%)**	0.1%	
Infinite (%)	0.0%	**Memory size**	3.3 MiB	
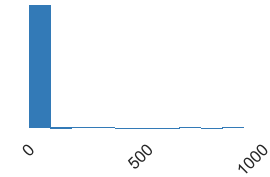
Toggle details

Host_URL
URL

Distinct count	49743	https://www.airbnb.co…	72	
		https://www.airbnb.co…	65	
Unique (%)	90.4%	https://www.airbnb.co…	31	
		https://www.airbnb.co…	24	
Missing	0	https://www.airbnb.co…	24	
Missing (%)	0.0%	Other values (49738)	54784	
Memory size	3.3			

MiB

Host_Verificati…
Categorical

`HIGH CARDINALITY`

Distinct count	542
Unique (%)	1.0%
Missing	32
Missing (%)	0.1%
Memory size	3.3 MiB

email,phone,reviews,ju…	11973
email,phone,reviews	11400
email,phone,facebook,…	3401
email,phone	2889
email,phone,facebook,…	2857
Other values (536)	22448

House_Rules
Categorical

`MISSING`

`HIGH CARDINALITY`

Distinct count	31242
Unique (%)	56.8%
Missing	21668
Missing (%)	39.4%
Memory size	3.3 MiB

- The lead guest will be…	63
No smoking	59
No smoking.	37
Please don't trash the …	35
.	33
Other values (31236)	33105

ID
Real number ($\mathbb{R}_{\geq 0}$)

UNIQUE

Distinct count	55000
Unique (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	9847567.556
Minimum	590
Maximum	19283762
Zeros	0
Zeros (%)	0.0%
Memory size	3.3 MiB
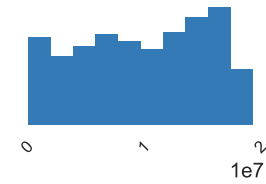
Toggle details

Interaction
Categorical

MISSING
HIGH CARDINALITY

Distinct count	26891
Unique (%)	48.9%
Missing	26278
Missing (%)	47.8%
Memory size	3.3 MiB

Iamb&b will let you in, …	81
If you have questions d…	63
We will be contactable …	42
Iambnb will let you in, …	22
We will be more than h…	20
Other values (26885)	28494

Toggle details

Jurisdiction_N…
Categorical

`MISSING`

Distinct count	21
Unique (%)	< 0.1%
Missing	39423
Missing (%)	71.7%
Memory size	3.3 MiB

Paris	5900
Amsterdam	3031
City of Los Angeles, CA	2245
SAN FRANCISCO	958
DISTRICT OF COLUM…	776
Other values (15)	2667

Toggle details

Last_Review
Categorical

`MISSING`

`HIGH CARDINALITY`

Distinct count	1198
Unique (%)	2.2%
Missing	13492
Missing (%)	24.5%
Memory size	3.3 MiB

2017-03-26	664
2017-03-19	548
2017-04-30	546
2017-01-01	541
2017-01-02	527
Other values (1192)	38682

Toggle details

Last_Scraped
Categorical

`TYPE DATE`

Distinct count	47
Unique (%)	0.1%

2017-05-09	4952
2017-05-03	4865
2017-04-08	4569
2017-04-02	4085

Missing	0	2017-04-05	■	4050
Missing (%)	0.0%	Other values (42)		32479
Memory size	3.3 MiB			

Latitude
Real number (ℝ)

`UNIQUE`

Distinct count	55000	**Mean**	37.8798182
Unique (%)	100.0%	**Minimum**	-38.1905409316
Missing	0	**Maximum**	55.9908313671(
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	3.3 MiB
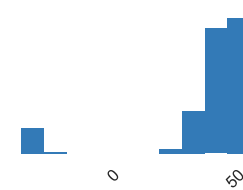
License
Categorical

`MISSING`
`HIGH CARDINALITY`

Distinct count	1153	City registration pending	436
Unique (%)	2.1%	145636	4
Missing	53367	0289599-0	4
Missing (%)	97.0%	154316	3
Memory size	3.3	14-241130-000-00-RS	3
		Other values (1147)	1183

MiB

Listing_Url
URL

`UNIQUE`

Distinct count	55000	https://www.airbnb.co…	1
		https://www.airbnb.co…	1
Unique (%)	100.0%	https://www.airbnb.co…	1
		https://www.airbnb.co…	1
Missing	0	https://www.airbnb.co…	1
Missing (%)	0.0%	Other values (54995)	54995
Memory size	3.3 MiB		

Longitude
Real number (ℝ)

`UNIQUE`

Distinct count	55000	**Mean**	-14.90010024
Unique (%)	100.0%	**Minimum**	-123.211214254
Missing	0	**Maximum**	153.6216900422
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	3.3 MiB
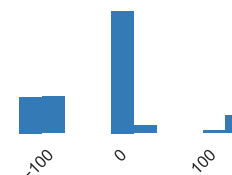
Market
Categorical

MISSING

HIGH CARDINALITY

Distinct count	**77**
Unique (%)	0.1%
Missing	**653**
Missing (%)	**1.2%**
Memory size	3.3 MiB

Paris	5999
London	5606
New York	4046
Los Angeles	3294
Barcelona	3145
Other values (71)	32257
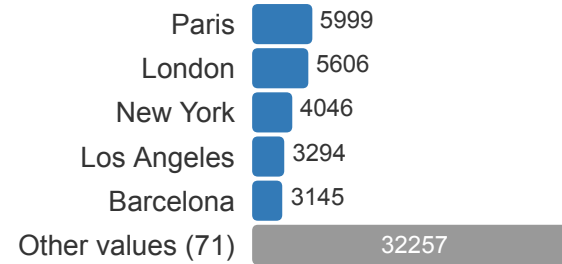
Toggle details

Maximum_Nig…
Real number ($\mathbb{R}_{\geq 0}$)

SKEWED

Distinct count	276	**Mean**	275053.4708
Unique (%)	0.5%	**Minimum**	1
Missing	0	**Maximum**	2147483647
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	3.3 MiB

Toggle details

Medium_Url

URL

`MISSING`

Distinct count	44409
Unique (%)	80.7%
Missing	10592
Missing (%)	19.3%
Memory size	3.3 MiB

https://a0.muscache.co…	1
https://a0.muscache.co…	1
https://a0.muscache.co…	1
https://a0.muscache.co…	1
https://a0.muscache.co…	1
Other values (44403)	44403
(Missing)	10592

Toggle details

Minimum_Nights

Real number (ℝ≥0)

`SKEWED`

Distinct count	70
Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	3.257090909
Minimum	1
Maximum	3888
Zeros	0
Zeros (%)	0.0%
Memory size	3.3 MiB
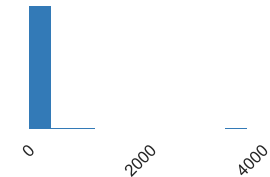
Toggle details

Monthly_Price

Real number (ℝ≥0)

`MISSING`

Distinct count	1572
Unique (%)	2.9%

Mean	2876.253642
Minimum	0.0

Missing	44359	**Maximum**	140008.0
Missing (%)	80.7%	**Zeros**	1
Infinite	0	**Zeros (%)**	< 0.1%
Infinite (%)	0.0%	**Memory size**	3.3 MiB
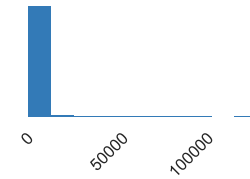
Toggle details

Name
Categorical

HIGH CARDINALITY

Distinct count	54535
Unique (%)	99.2%
Missing	53
Missing (%)	0.1%
Memory size	3.3 MiB

Single room	6
Chambre à louer	6
Double room	6
Charming apartment in…	6
Home Sweet Home	6
Other values (54529)	54917

Toggle details

Neighborhood…
Categorical

MISSING

HIGH CARDINALITY

Distinct count	31587
Unique (%)	57.4%
Missing	22457
Missing (%)	40.8%

(URL HIDDEN)	19
Amsterdam West is on…	14
Pollensa and its surrou…	11
Oud West is close to th…	9
Barcelona´s 19th centu…	8
Other values (31581)	32482

Memory size	3.3 MiB		

Toggle details

Neighbourhood
Categorical

MISSING

HIGH CARDINALITY

Distinct count	1887	Oud-West	437
Unique (%)	3.4%	Montmartre	380
		Williamsburg	339
Missing	16867	République	301
Missing (%)	30.7%	Buttes-Chaumont - Bel…	290
		Other values (1881)	36386
Memory size	3.3 MiB		

Toggle details

Neighbourhoo…
Categorical

HIGH CARDINALITY

Distinct count	1932	I Centro Storico	1242
Unique (%)	3.5%	Buttes-Montmartre	803
		Popincourt	783
Missing	0	Sydney	647
Missing (%)	0.0%	Tower Hamlets	600
		Other values (1927)	50925
Memory size	3.3 MiB		

Toggle details

Neighbourhoo…
Categorical

MISSING

HIGH CARDINALITY

Distinct count	68
Unique (%)	0.1%
Missing	43569
Missing (%)	79.2%
Memory size	3.3 MiB

Manhattan	2042
Brooklyn	1526
Eixample	1058
Centro	726
Ciutat Vella	689
Other values (62)	5390
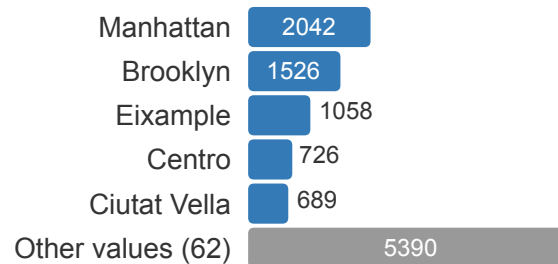
Toggle details

Notes
Categorical

MISSING

HIGH CARDINALITY

Distinct count	20728
Unique (%)	37.7%
Missing	33022
Missing (%)	60.0%
Memory size	3.3 MiB

Normal check-in and c…	27
Check-in: 4 pm Check-…	23
.	22
(PHONE NUMBER HI…	20
Guests we look for: - R…	19
Other values (20722)	21867
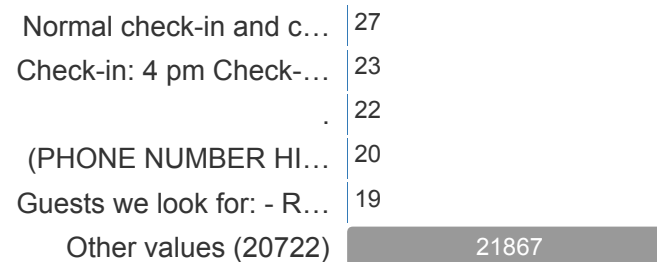
Toggle details

Number_of_R…
Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct count	335	Mean	17.23676364
Unique (%)	0.6%	Minimum	0
Missing	0	Maximum	735
Missing (%)	0.0%	Zeros	13478
Infinite	0	Zeros (%)	24.5%
Infinite (%)	0.0%	Memory size	3.3 MiB
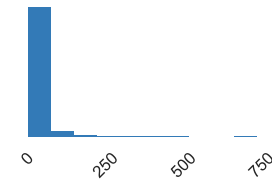
Toggle details

Picture_Url
URL

Distinct count	54940	https://public.opendata…	1
Unique (%)	99.9%	https://public.opendata…	1
Missing	61	https://public.opendata…	1
Missing (%)	0.1%	https://public.opendata…	1
Memory size	3.3 MiB	https://public.opendata…	1
		Other values (54934)	54934
		(Missing)	61

Toggle details

Price
Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct count	753	Mean	134.7691598
Unique (%)	1.4%	Minimum	0.0

Missing	772	**Maximum**	999.0
Missing (%)	1.4%	**Zeros**	1
Infinite	0	**Zeros (%)**	< 0.1%
Infinite (%)	0.0%	**Memory size**	3.3 MiB
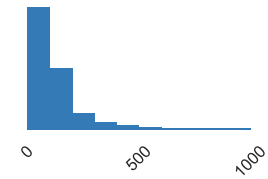
Toggle details

Property_Type
Categorical

Distinct count	35
Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.3 MiB

Apartment	39020
House	10507
Bed & Breakfast	1281
Condominium	1108
Townhouse	653
Other values (30)	2431

Toggle details

Review_Score...
Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	10	**Mean**	9.52595605
Unique (%)	< 0.1%	**Minimum**	2.0
		Maximum	10.0
Missing	14181	**Zeros**	0
		Zeros (%)	0.0%

Missing (%)	25.8%	**Memory size**	3.3 MiB
Infinite	0		
Infinite (%)	0.0%		

<div align="right">

[Toggle details](#)

</div>

Review_Score...
Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	10	**Mean**	9.687881686
Unique (%)	< 0.1%	**Minimum**	2.0
		Maximum	10.0
Missing	14227	**Zeros**	0
Missing (%)	25.9%	**Zeros (%)**	0.0%
Infinite	0	**Memory size**	3.3 MiB
Infinite (%)	0.0%		

<div align="right">

[Toggle details](#)

</div>

Review_Score...
Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	10	**Mean**	9.315847449
Unique (%)	< 0.1%	**Minimum**	2.0
		Maximum	10.0
Missing	14148	**Zeros**	0
Missing (%)	25.7%	**Zeros (%)**	0.0%

	Infinite	0	Memory size	3.3 MiB
	Infinite (%)	0.0%		

Review_Score…
Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	10	Mean	9.705644668	
Unique (%)	< 0.1%	Minimum	2.0	
		Maximum	10.0	
Missing	14165	Zeros	0	
Missing (%)	25.8%	Zeros (%)	0.0%	
Infinite	0	Memory size	3.3 MiB	
Infinite (%)	0.0%			

Review_Score…
Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	10	Mean	9.468389818	
Unique (%)	< 0.1%	Minimum	2.0	
		Maximum	10.0	
Missing	14222	Zeros	0	
Missing (%)	25.9%	Zeros (%)	0.0%	
Infinite	0	Memory size	3.3 MiB	

	Infinite (%)	0.0%			

Toggle details

Review_Score…
Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct count	58	**Mean**	92.84263513
Unique (%)	0.1%	**Minimum**	20.0
Missing	14076	**Maximum**	100.0
Missing (%)	25.6%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	3.3 MiB

Toggle details

Review_Score…
Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct count	10	**Mean**	9.311705833
Unique (%)	< 0.1%	**Minimum**	2.0
Missing	14234	**Maximum**	10.0
Missing (%)	25.9%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	3.3 MiB

Create PDF in your applications with the Pdfcrowd HTML to PDF API   PDFCROWD

Reviews_per_…
Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	973
Unique (%)	1.8%
Missing	13500
Missing (%)	24.5%
Infinite	0
Infinite (%)	0.0%

Mean	1.500886747
Minimum	0.01
Maximum	125.92
Zeros	0
Zeros (%)	0.0%
Memory size	3.3 MiB
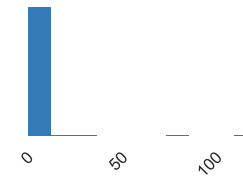
Room_Type
Categorical

Distinct count	3
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.3 MiB

Entire home/apt	35127
Private room	18870
Shared room	1003

Scrape_ID
Real number (ℝ≥0)

Distinct count	42	**Mean**	2.016940438e+1
Unique (%)	0.1%	**Minimum**	2015101123263
Missing	0	**Maximum**	2017061500270
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	3.3 MiB
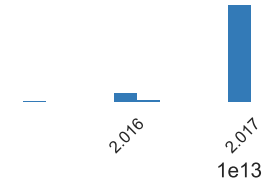
Toggle details

Security_Dep…
Real number (ℝ≥0)

`MISSING`

Distinct count	337	**Mean**	269.4072074
Unique (%)	0.6%	**Minimum**	0.0
Missing	31857	**Maximum**	999.0
Missing (%)	57.9%	**Zeros**	99
Infinite	0	**Zeros (%)**	0.2%
Infinite (%)	0.0%	**Memory size**	3.3 MiB
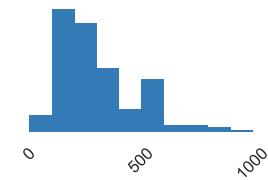
Toggle details

Smart_Location

Categorical

`HIGH CARDINALITY`

Distinct count	2045
Unique (%)	3.7%
Missing	0
Missing (%)	0.0%
Memory size	3.3 MiB

Paris, France	5798
London, United Kingdom	5093
Barcelona, Spain	3112
Amsterdam, Netherlands	3059
Los Angeles, CA	2091
Other values (2040)	35847

Toggle details

Space

Categorical

`MISSING`

`HIGH CARDINALITY`

Distinct count	38324
Unique (%)	69.7%
Missing	16278
Missing (%)	29.6%
Memory size	3.3 MiB

This gorgeous apartme…	26
Every booking is instan…	15
Discovery Melbourne i…	7
We are like a Europea…	6
The rates include brea…	6
Other values (38318)	38662

Toggle details

Square_Feet

Real number (ℝ≥0)

`MISSING`

Distinct count	239
Unique (%)	0.4%

Mean	561.8962597
Minimum	0.0
Maximum	6997.0

Missing	53583	**Zeros**	452
Missing (%)	97.4%	**Zeros (%)**	0.8%
Infinite	0	**Memory size**	3.3 MiB
Infinite (%)	0.0%		
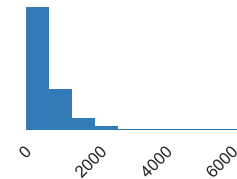
Toggle details

State
Categorical

MISSING

HIGH CARDINALITY

Distinct count	515
Unique (%)	0.9%
Missing	5307
Missing (%)	9.6%
Memory size	3.3 MiB

Île-de-France	5876
CA	5330
NY	4064
NSW	2679
Lazio	2545
Other values (509)	29199

Toggle details

Street
Categorical

HIGH CARDINALITY

Distinct count	22001
Unique (%)	40.0%
Missing	0
Missing (%)	0.0%
Memory size	3.3

Montmartre, Paris, Île-…	364
XI Arrondissement, Pa…	228
Alésia, Paris, Île-de-Fr…	223
Batignolles, Paris, Île-d…	212
République, Paris, Île-…	212
Other values (21996)	53761

MiB

Summary
Categorical

`MISSING`
`HIGH CARDINALITY`

Distinct count	52505
Unique (%)	95.5%
Missing	1853
Missing (%)	3.4%
Memory size	3.3 MiB

My place is good for co…	24
brand new Sleepin…	23
My place is good for co…	18
.	14
Unique apartment in vi…	13
Other values (52499)	53055

Thumbnail_Url
URL

`MISSING`

Distinct count	44409
Unique (%)	80.7%
Missing	10592
Missing (%)	19.3%
Memory size	3.3 MiB

https://a0.muscache.co…	1
https://a0.muscache.co…	1
https://a0.muscache.co…	1
https://a0.muscache.co…	1
https://a0.muscache.co…	1
Other values (44403)	44403
(Missing)	10592

Transit
Categorical

MISSING

HIGH CARDINALITY

Distinct count	32971
Unique (%)	59.9%
Missing	21214
Missing (%)	38.6%
Memory size	3.3 MiB

Ideal to discover the cit…	20
Yes	20
Convenient public tran…	15
Monastiraki metro stati…	7
Downtown Austin has …	7
Other values (32965)	33717

Toggle details

Weekly_Price
Real number ($\mathbb{R}_{\geq 0}$)

MISSING

HIGH CORRELATION

Distinct count	657
Unique (%)	1.2%
Missing	44072
Missing (%)	80.1%
Infinite	0
Infinite (%)	0.0%

Mean	494.693265
Minimum	0.0
Maximum	999.0
Zeros	1
Zeros (%)	< 0.1%
Memory size	3.3 MiB
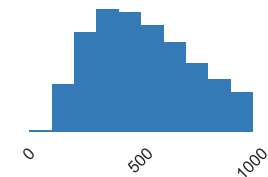
Toggle details

XL_Picture_Url
URL

MISSING

Distinct count	44409
Unique (%)	80.7%
Missing	10592
Missing (%)	19.3%
Memory size	3.3 MiB

https://a2.muscache.co…	1
https://a0.muscache.co…	1
https://a2.muscache.co…	1
https://a0.muscache.co…	1
https://a0.muscache.co…	1
Other values (44403)	44403
(Missing)	10592

Toggle details

Zipcode
Categorical

MISSING

HIGH CARDINALITY

Distinct count	11392
Unique (%)	20.7%
Missing	2127
Missing (%)	3.9%
Memory size	3.3 MiB

75018	788
75011	749
75020	431
75017	389
78704	378
Other values (11386)	50138

Toggle details

Correlations
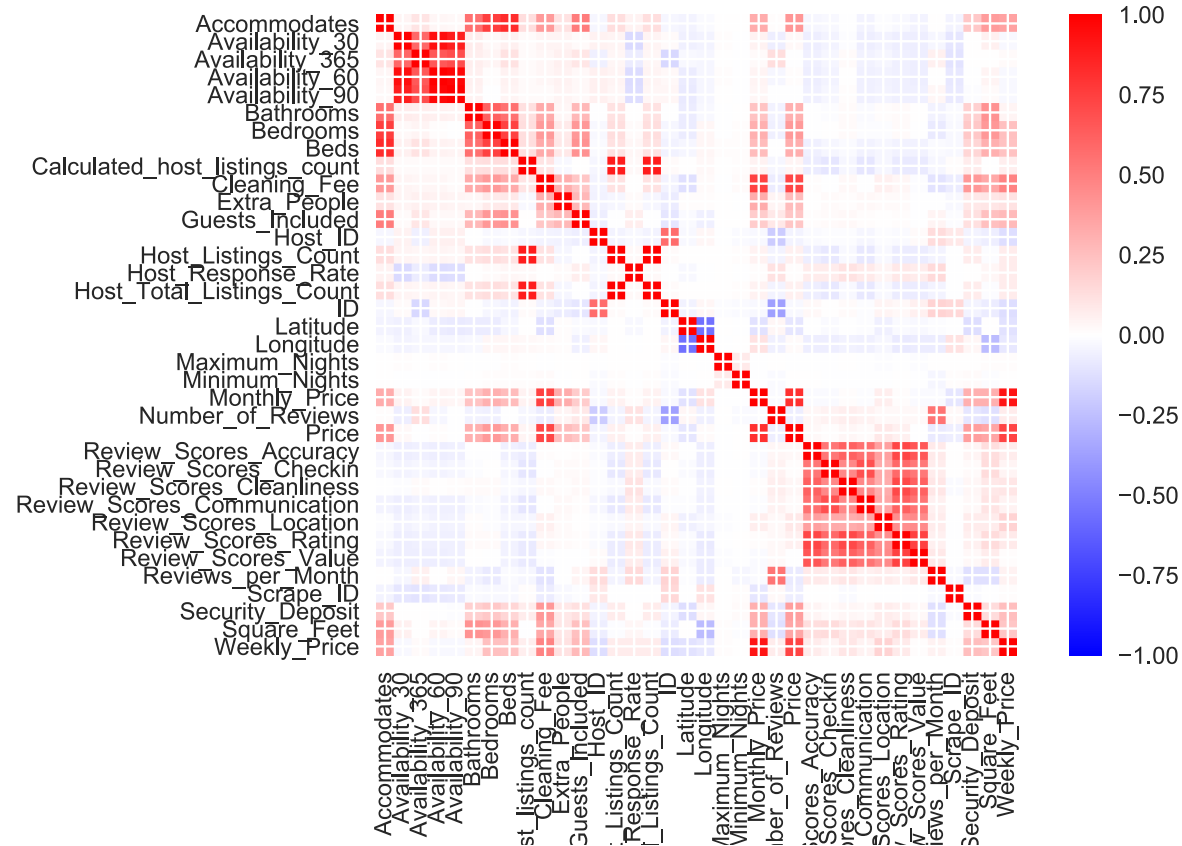
Missing values
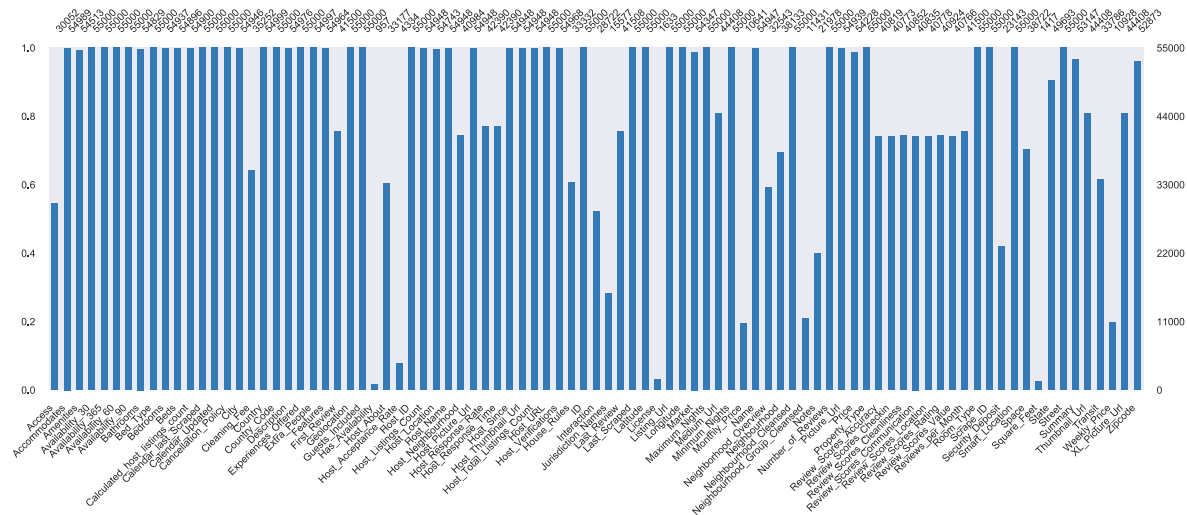
Sample

First rows

Access			Accommodates	Amenities

	Access	Accommodates	Amenities
0	Internet Private balcony The Kitchen ONLY for ...	2.0	Internet,Wireless Internet,Kitchen,S
1	NaN	4.0	TV,Kitchen,Heating,Family/kid frien
2	The loft and its contents are for you to enjoy...	4.0	TV,Internet,Wireless Internet,Air co
3	lavadora, secadora, wi-fi, toallas, sabanas....	2.0	Wireless Internet,Smoking allowed,
4	NaN	5.0	TV,Internet,Wireless Internet,Air co
5	NaN	3.0	Internet,Wireless Internet,Kitchen,S
6	We use the apartment ourselves for our holiday...	4.0	TV,Internet,Wireless Internet,Air co
7	NaN	5.0	TV,Internet,Wireless Internet,Air co
8	The apartment is located in the heart of the c...	4.0	TV,Internet,Wireless Internet,Air co
9	You can use the kitchen.	2.0	TV,Internet,Wireless Internet,Air co
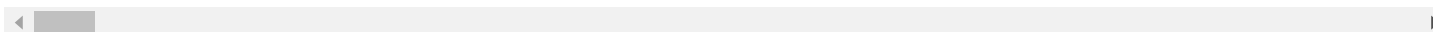
Last rows

	Access	Accommodates	Amenities
54990	A private apartment, wifi usage and all food f...	10.0	TV,Cable TV,Internet,Wireless

	Access	Accommodates	Amenities
54991	Amazing roof access! There is a beautiful roof...	2.0	TV,Internet,Wireless Internet,Ai
54992	You can use the bedroom, living room, dining r...	1.0	TV,Wireless Internet,Wheelcha
54993	To the living room with a tv, fully stock kitc...	2.0	TV,Internet,Wireless Internet,Ai
54994	Full kitchen, wifi	2.0	Internet,Wireless Internet,Kitch
54995	NaN	6.0	TV,Internet,Wireless Internet,Ai
54996	NaN	8.0	TV,Internet,Wireless Internet,Ai
54997	You will have access to Wifi, cable TV, and ba...	2.0	TV,Cable TV,Internet,Wireless
54998	There is a small backyard area that will be op...	4.0	TV,Cable TV,Internet,Wireless
54999	You will have your own room with a private bat...	2.0	TV,Wireless Internet,Air conditi
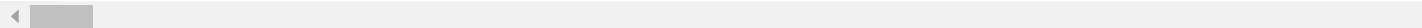
Report generated with pandas-profiling.

NULL VALUE IMPUTATION USING KNN IMPUTER

Each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors found in the training set. Two samples are close if the features that neither is missing are close. By default, a euclidean distance metric that supports missing values, nan_euclidean_distances, is used to find the nearest neighbors

CODE FOR KNN IMPUTER

```
from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=3)
df_filled = imputer.fit_transform(df1.select_dtypes(exclude="object"))
l=df1.select_dtypes(exclude="object").columns
df2=pd.DataFrame(df_filled,columns=l)
```

FEATURE HASHER

Implements **feature hashing**, aka the **hashing** trick. This class turns sequences of symbolic **feature** names (strings) into scipy. sparse matrices, using a **hash** function to compute the matrix column corresponding to a name. The **hash** function employed is the signed 32-bit version of Murmurhash3. It is used to reduce number of features.

```
from sklearn.feature_extraction import FeatureHasher
fh = FeatureHasher(n_features=10, input_type='string')
hashed_features = fh.fit_transform(df1['City'])
hashed_features = hashed_features.toarray()
hashed_features_state = fh.fit_transform(df1['State'])
hashed_features_state = hashed_features_state.toarray()
d=pd.concat([pd.DataFrame(hashed_features),pd.DataFrame(hashed_features_state)], axis=1)
```

MODELS:

LINEAR REGRESSION:

R-squared Training 0.6487852959512943
R-squared Testing 0.5865704373002176
Train RMSE : 86.78940238772637
Test RMSE : 92.01828247360002

RIDGE:

R-squared Training 0.6443076637572229
R-squared Testing 0.6060359775006094
Train RMSE : 87.34088877412083
Test RMSE : 89.82591305553078

LASSO:

R-squared Training 0.5825761561713922
R-squared Testing 0.5723513359965798
Train RMSE : 94.61695764501741
Test RMSE : 93.58730013192235

RANDOM FOREST REGRESSOR

R-squared Training 0.9534748956827677
R-squared Testing 0.656342443491584
Train RMSE : 31.588138297881606
Test RMSE : 83.89504430842655

 BY THE ABOVE ANALYSIS WE CAN SAY THAT RANDOM FOREST R
EGRESSOR IS THE BEST SUITABLE MODEL FOR OUR PRICE PREDICT
ION