**ReadMe**

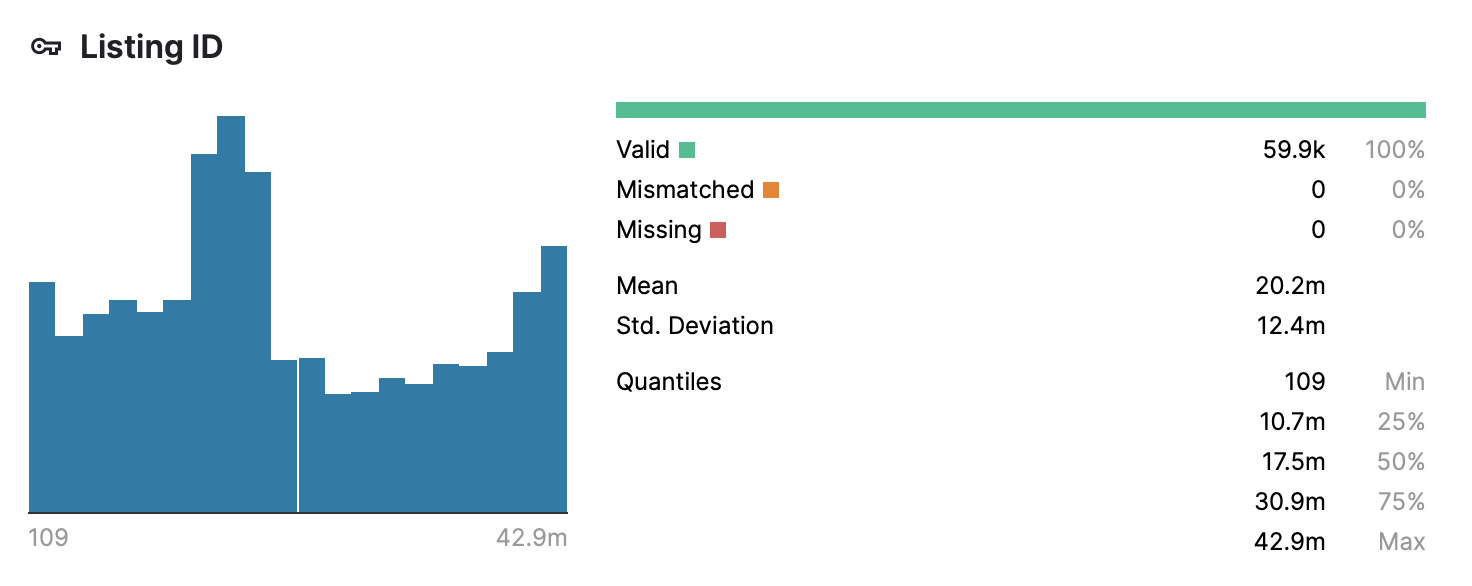
**Data Set Overview**

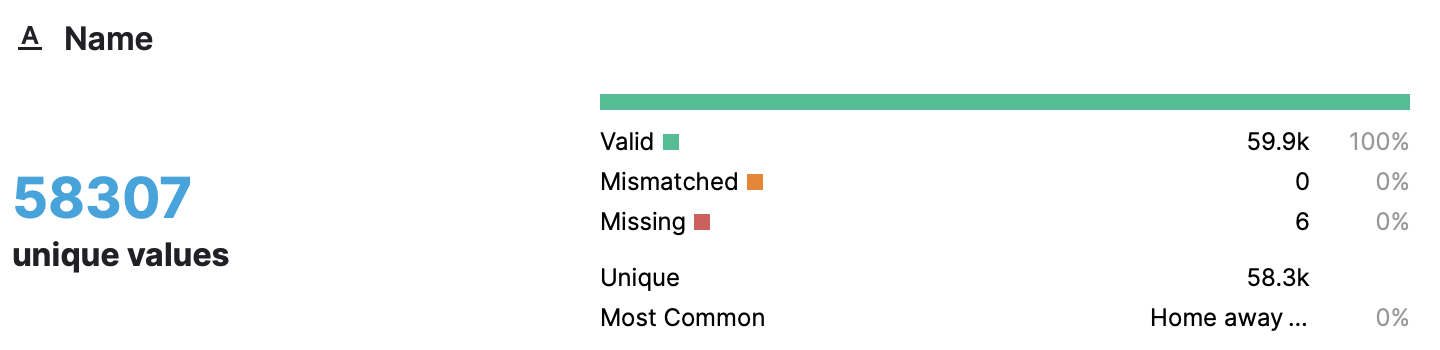
Data source: <https://www.kaggle.com/samyukthamurali/airbnb-ratings-dataset>

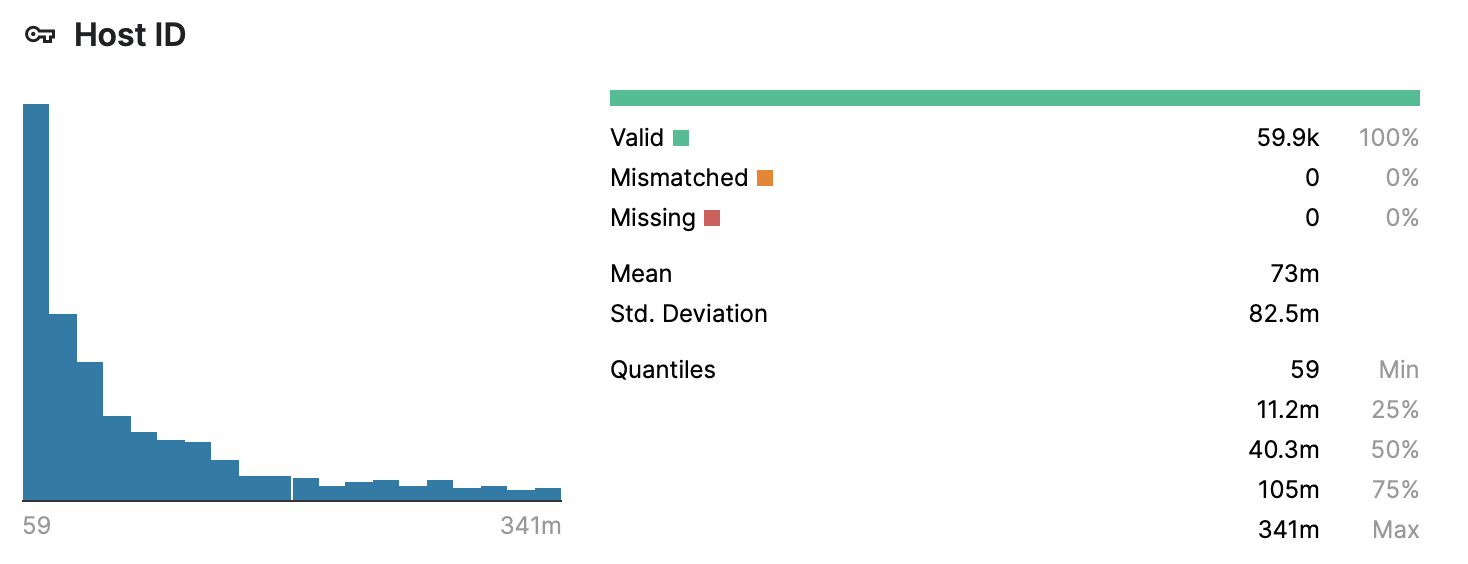
The original dataset contains four sub-datasets: **LA\_Listings**, **NY\_Listings**, **airbnb\_ratings\_new** and **airbnb-reviews**. The first three datasets contain 59.9k instances and 35 attributes, including **customer ID**, **host ID**, **locations**, **layouts**, **furnishings**, **prices of the residences**, **review scores**, etc. The last dataset contains 1325 instances and 6 attributes, which are **customer ID**, **host ID**, **review ID**, **reviewer name**, **date** and **comments**.

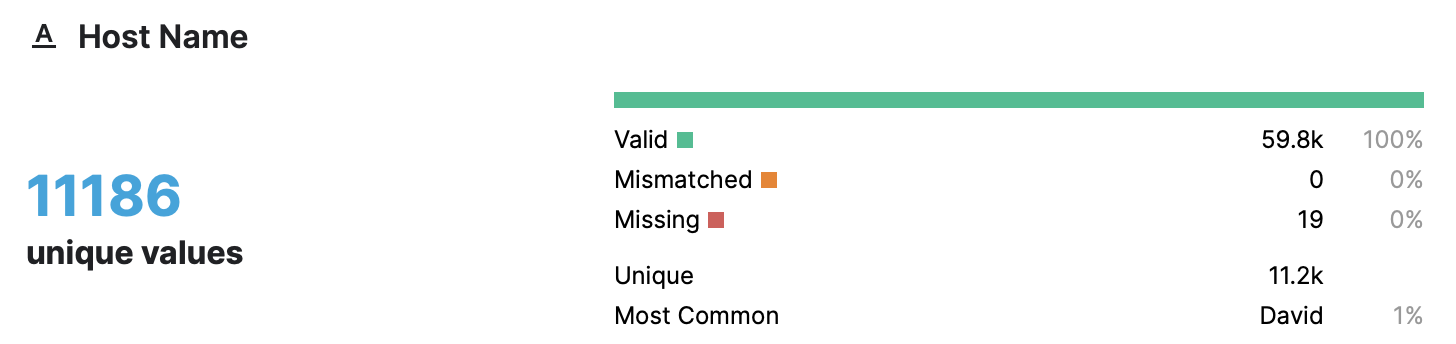


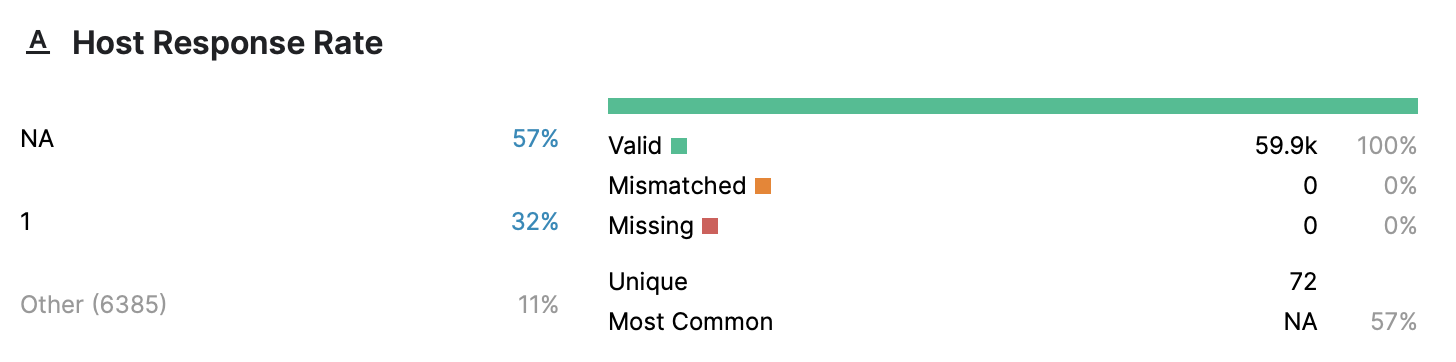
**COLUMNS**

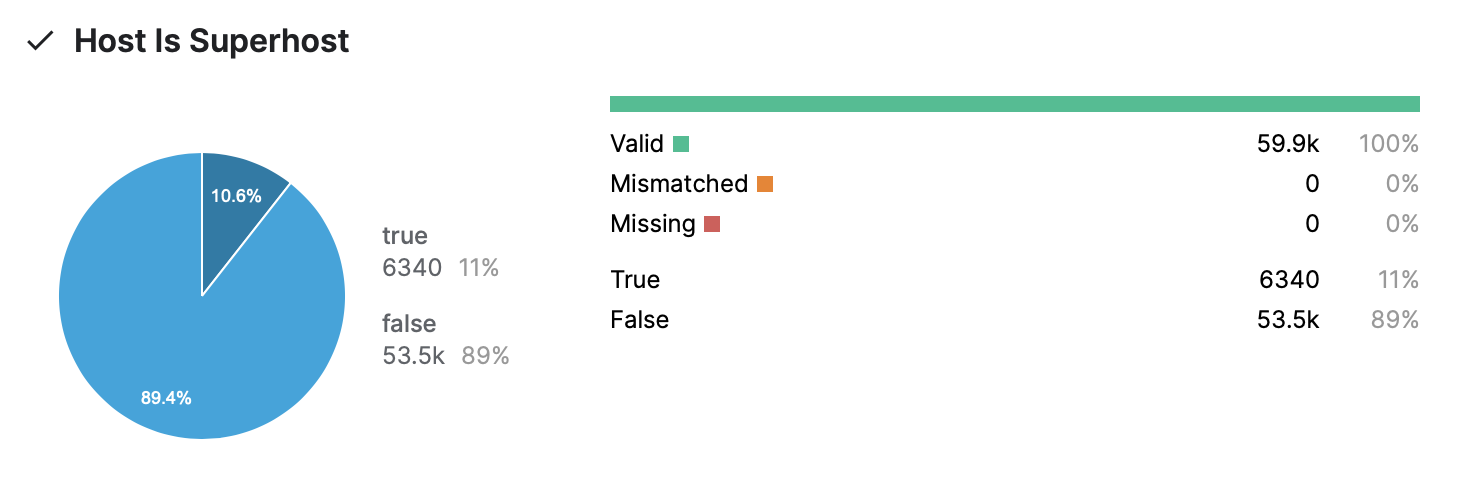


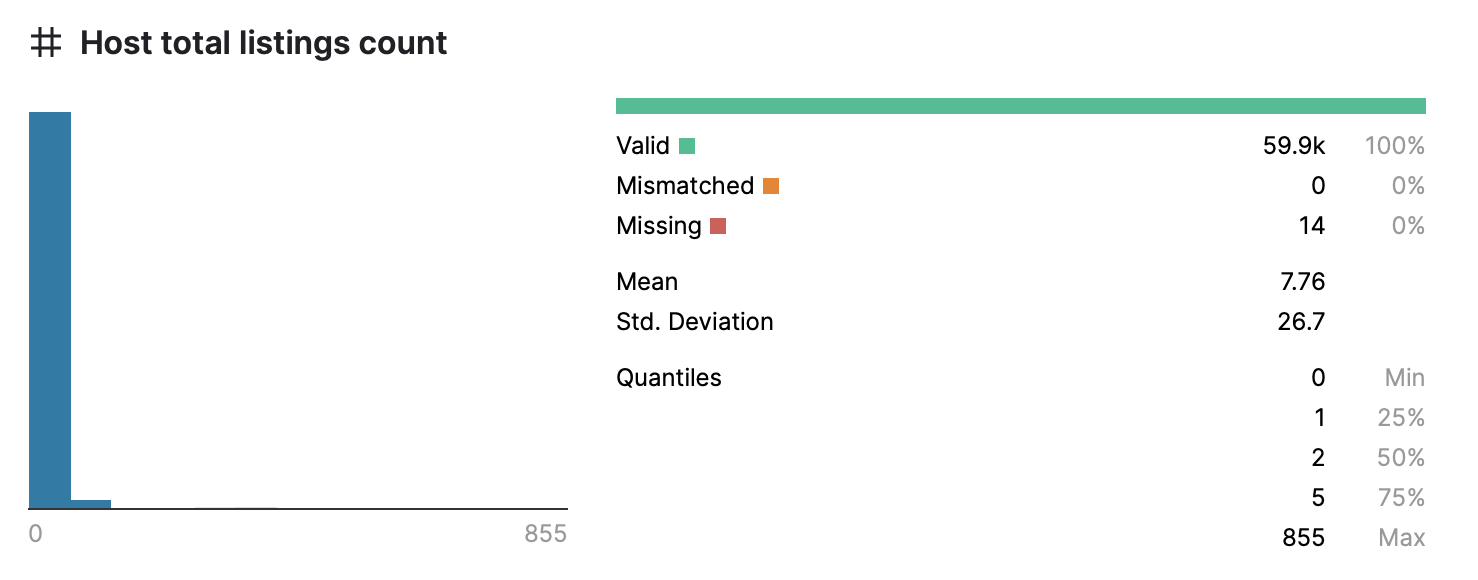


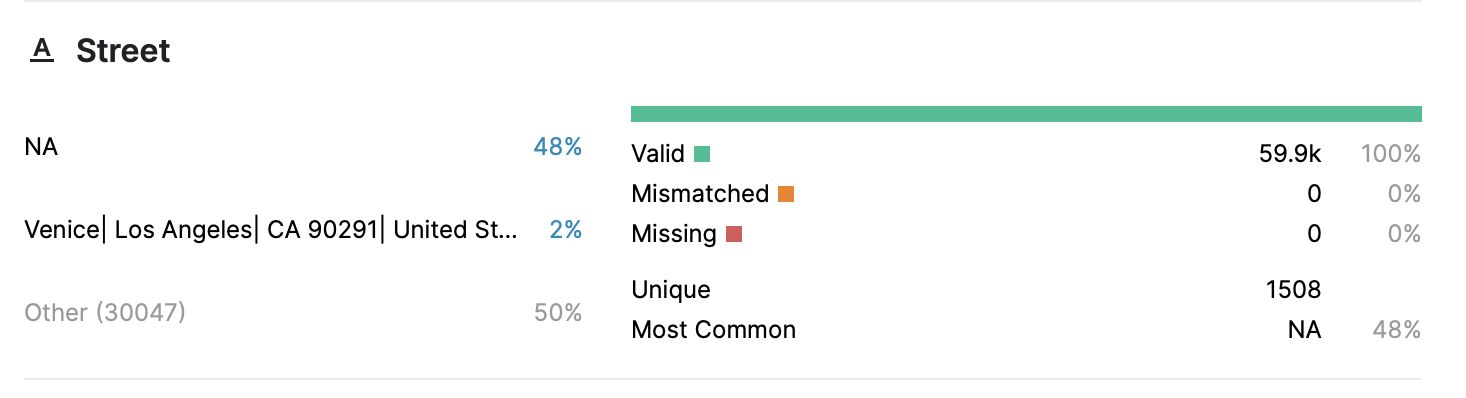


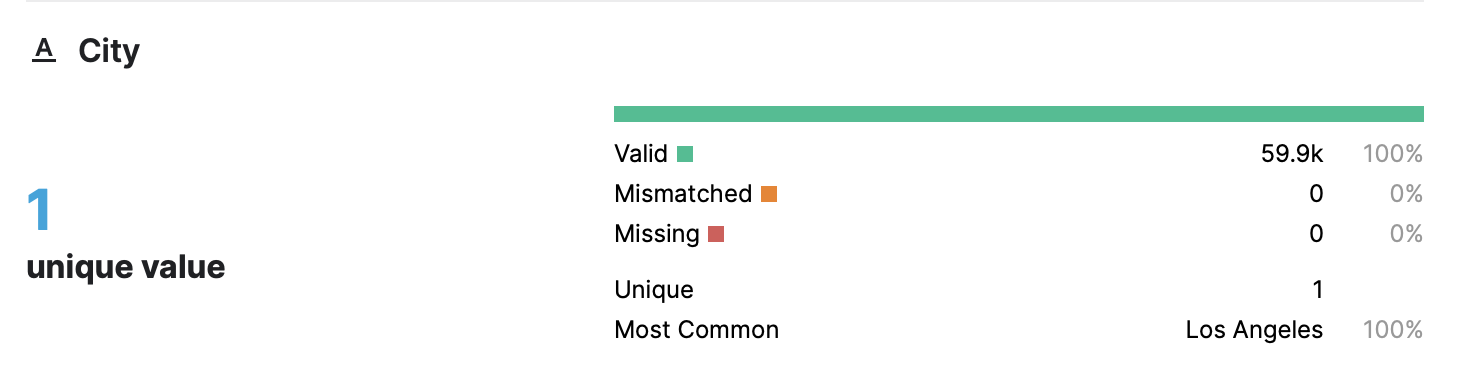


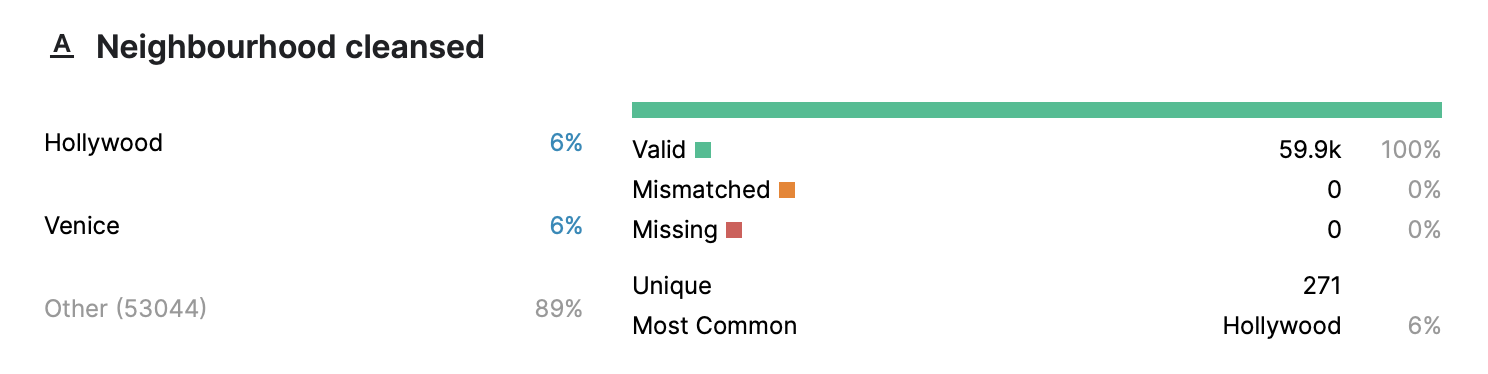




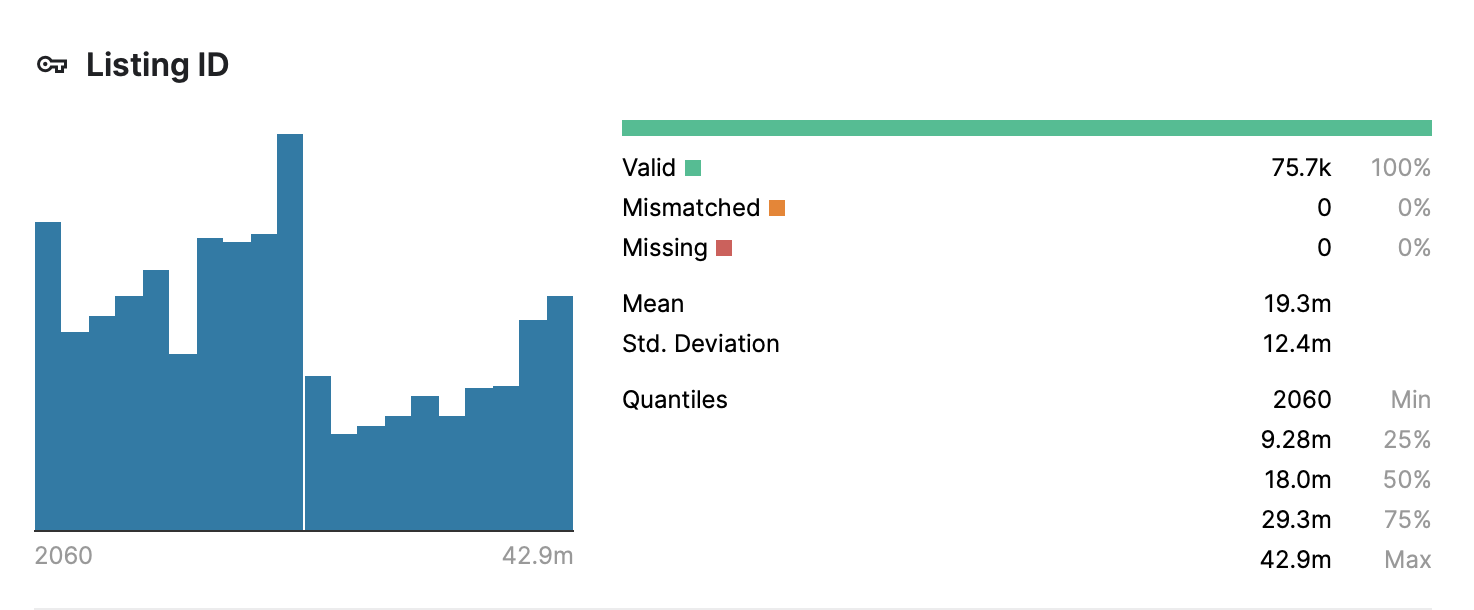


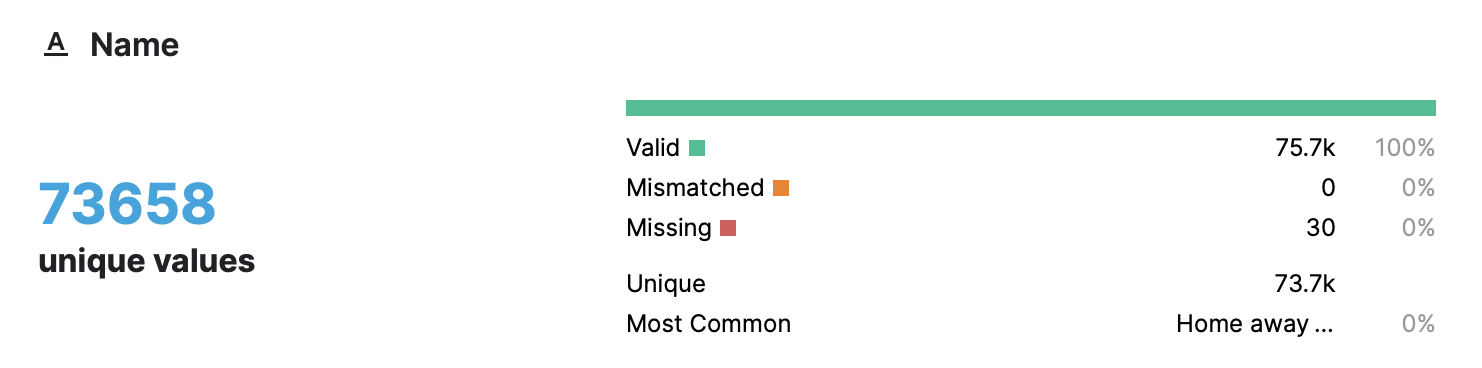


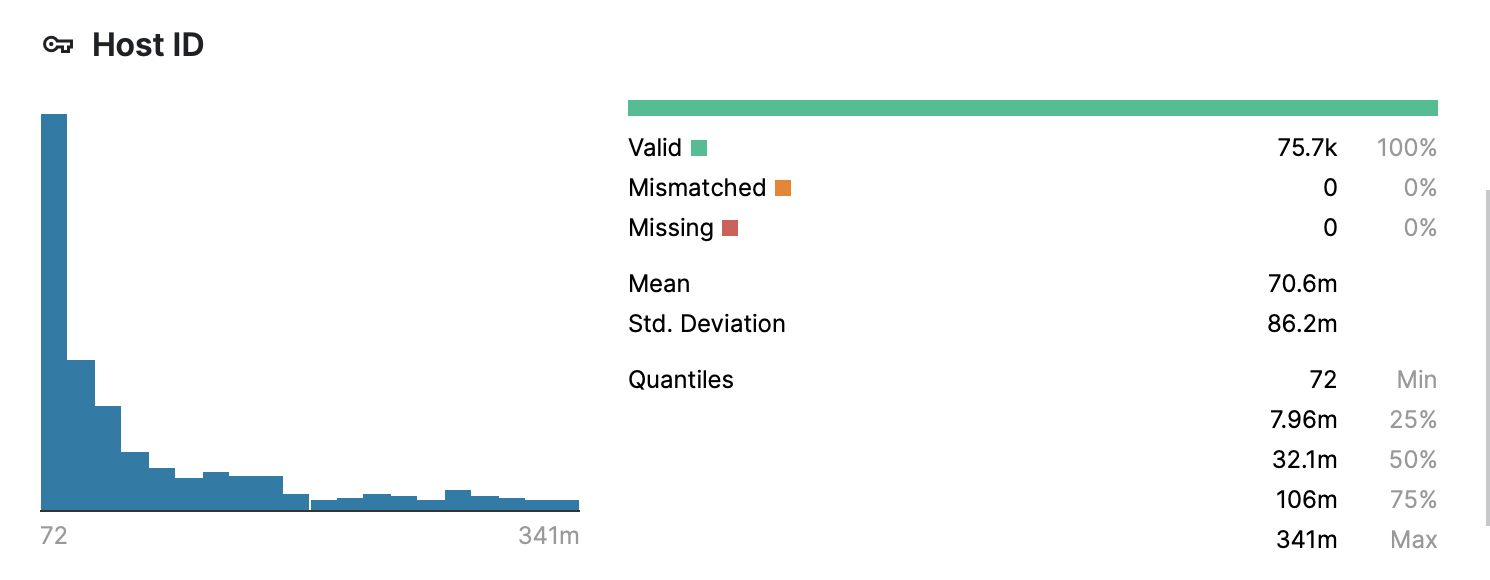


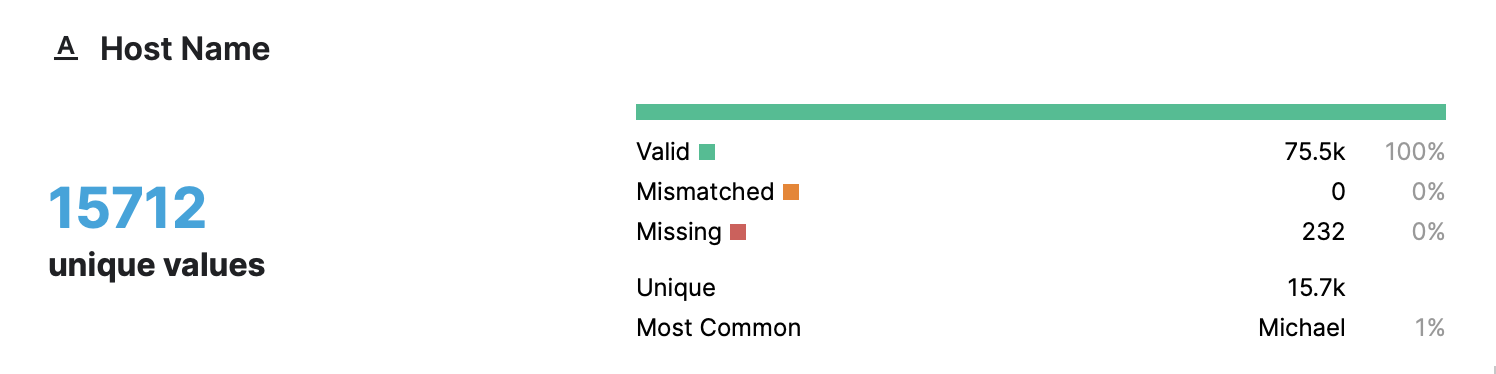


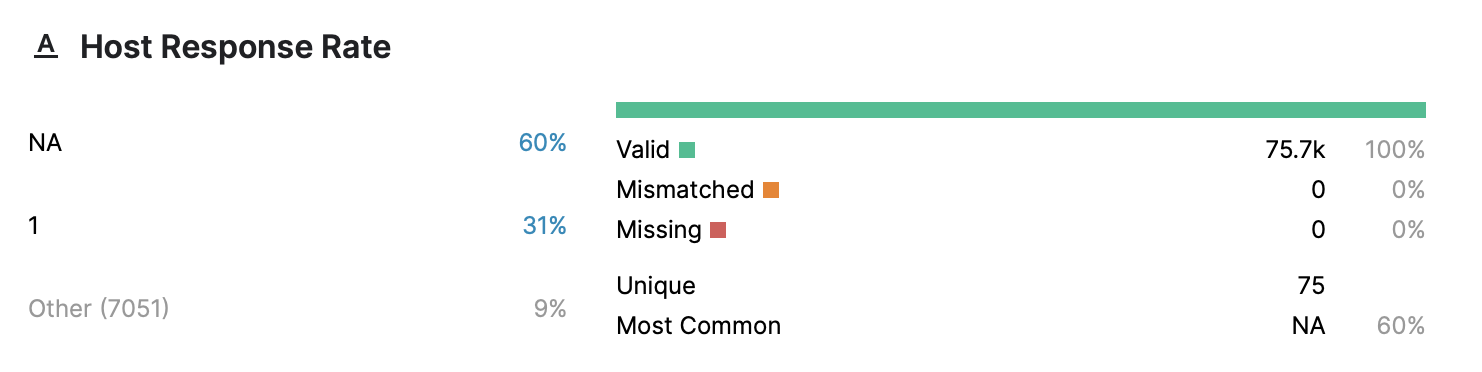


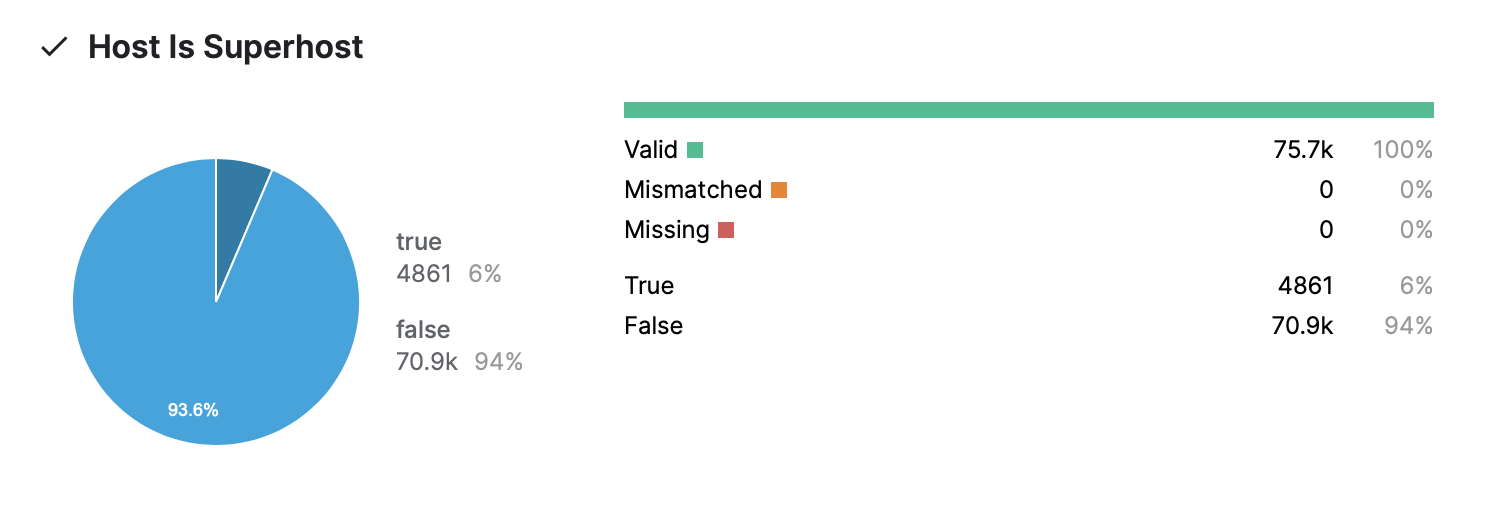


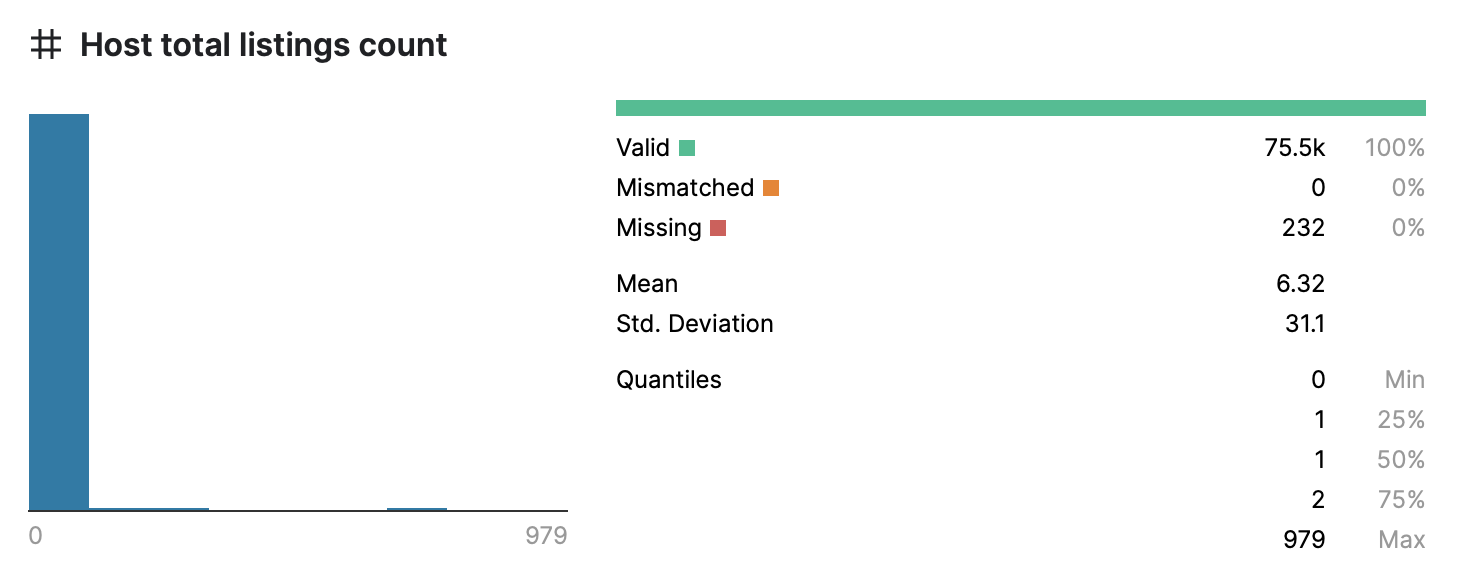


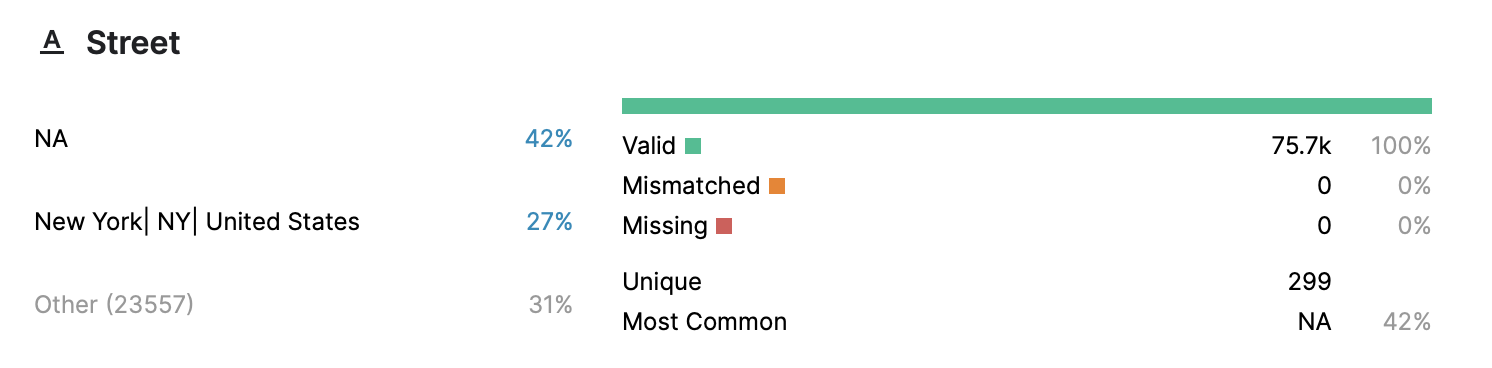


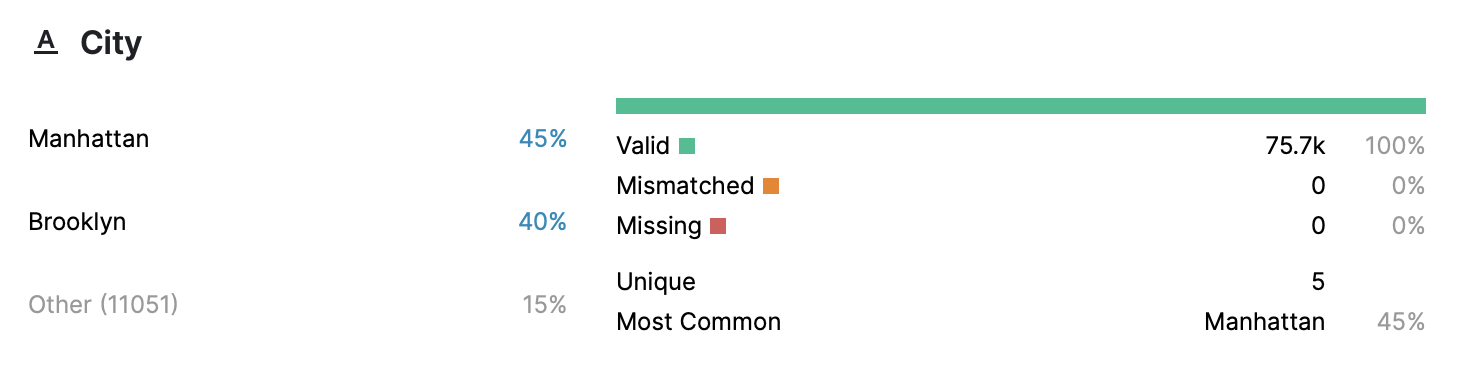


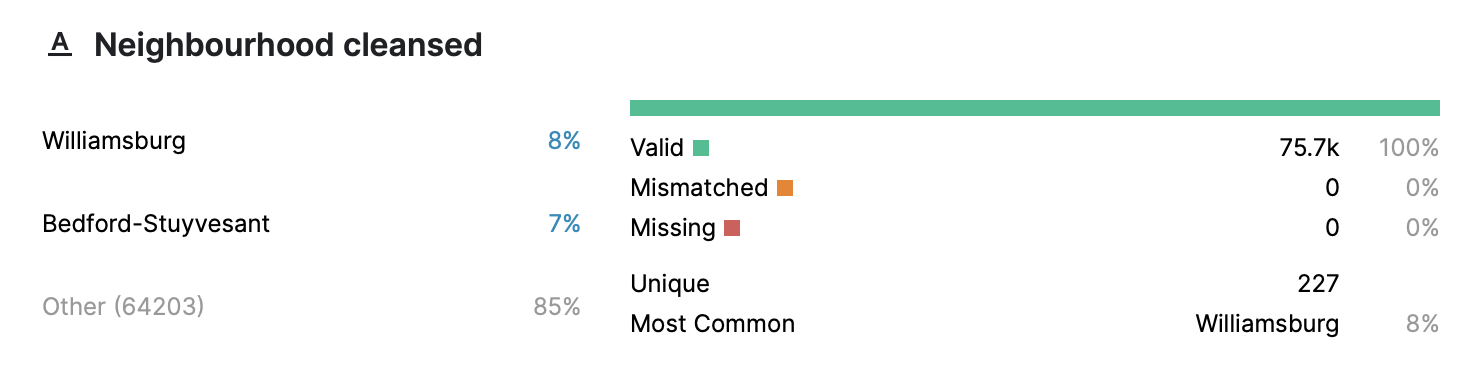




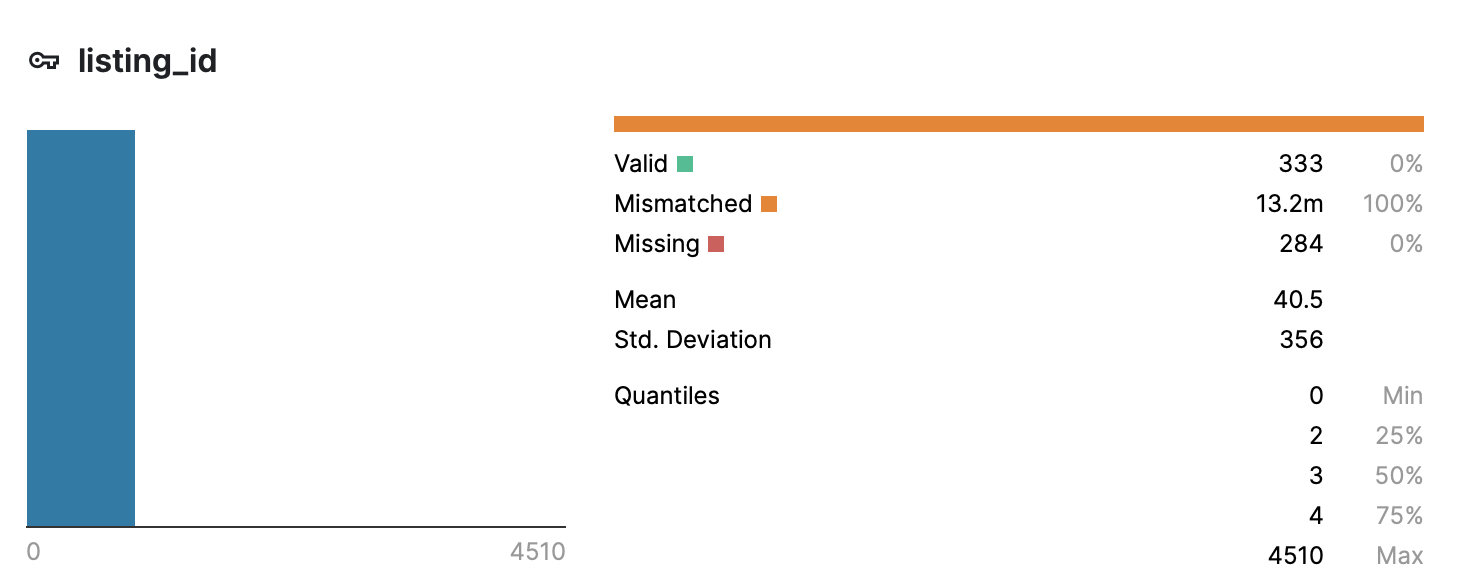


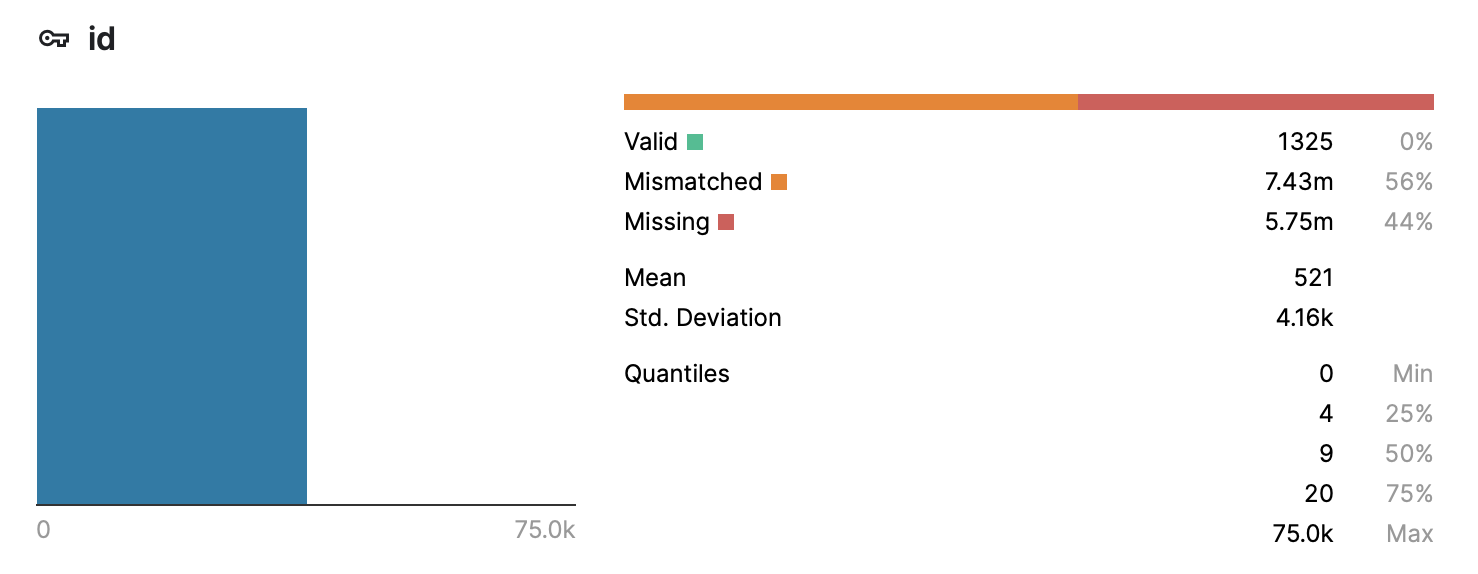


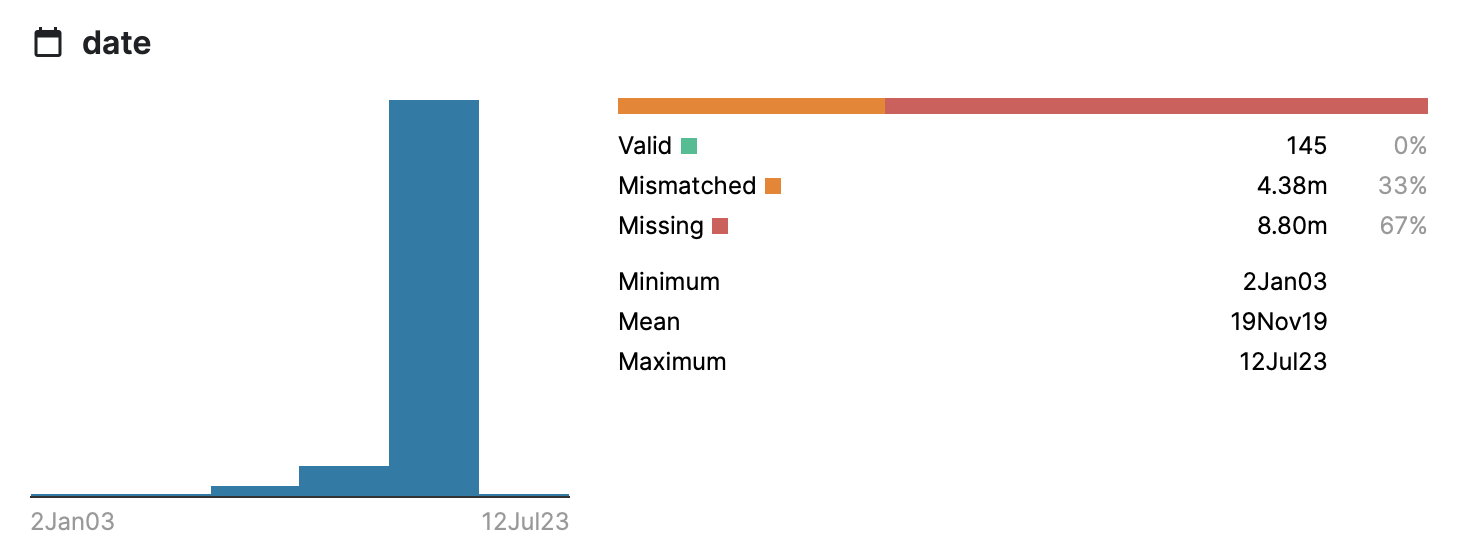


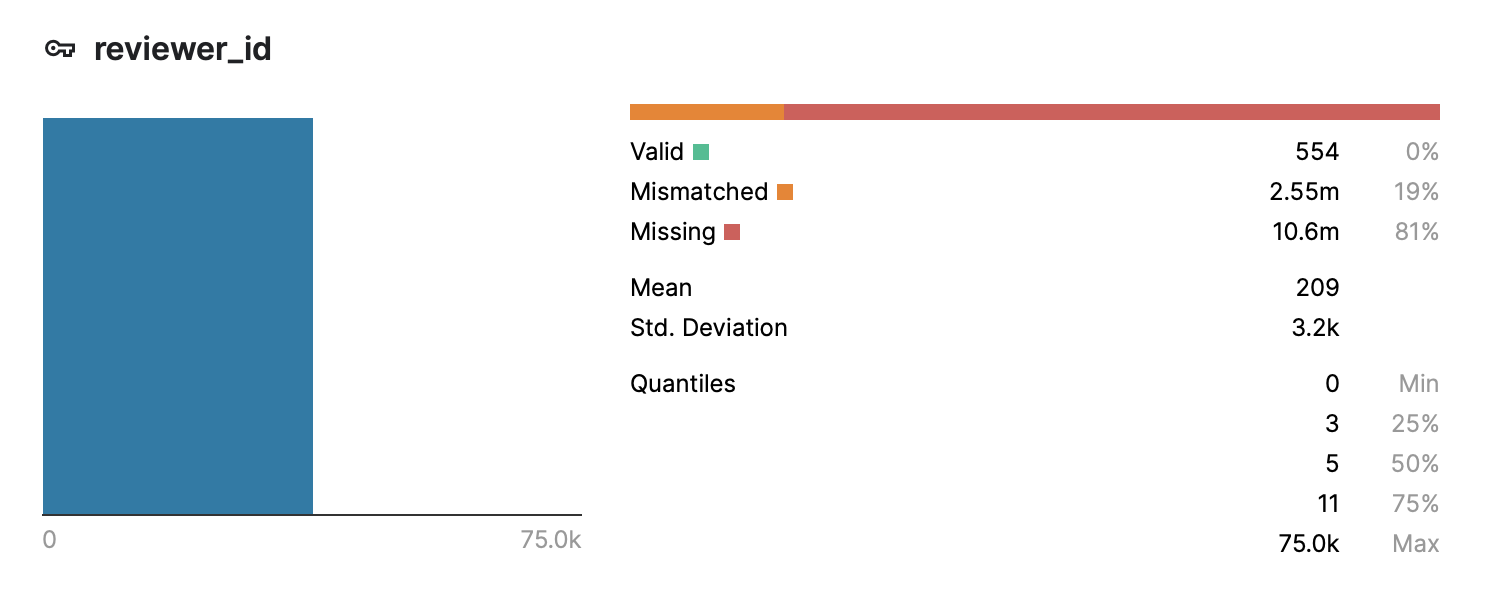


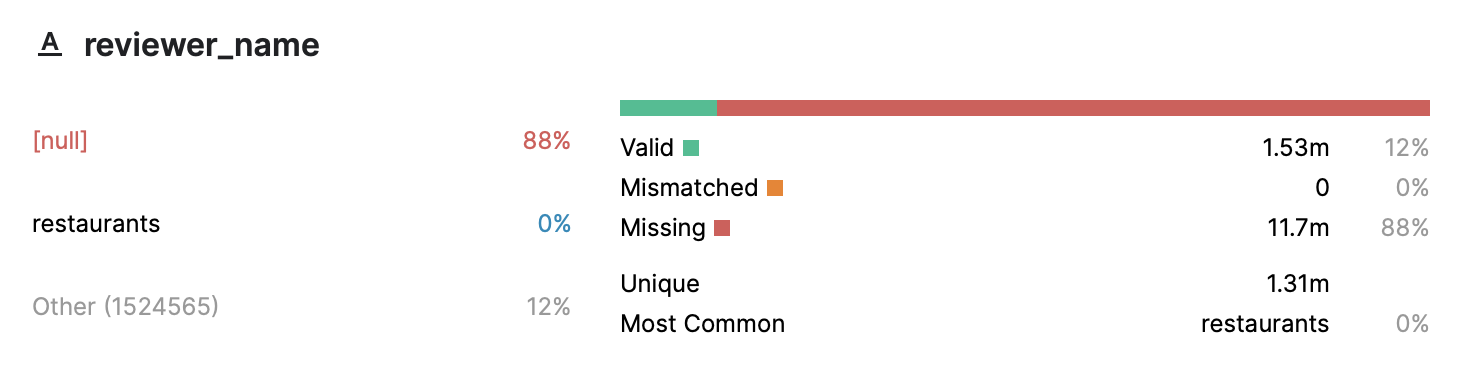


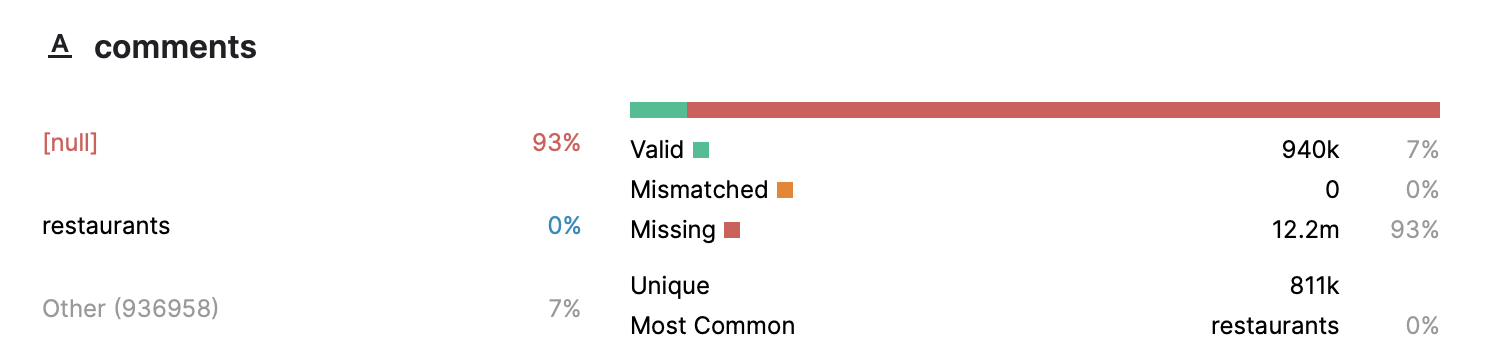




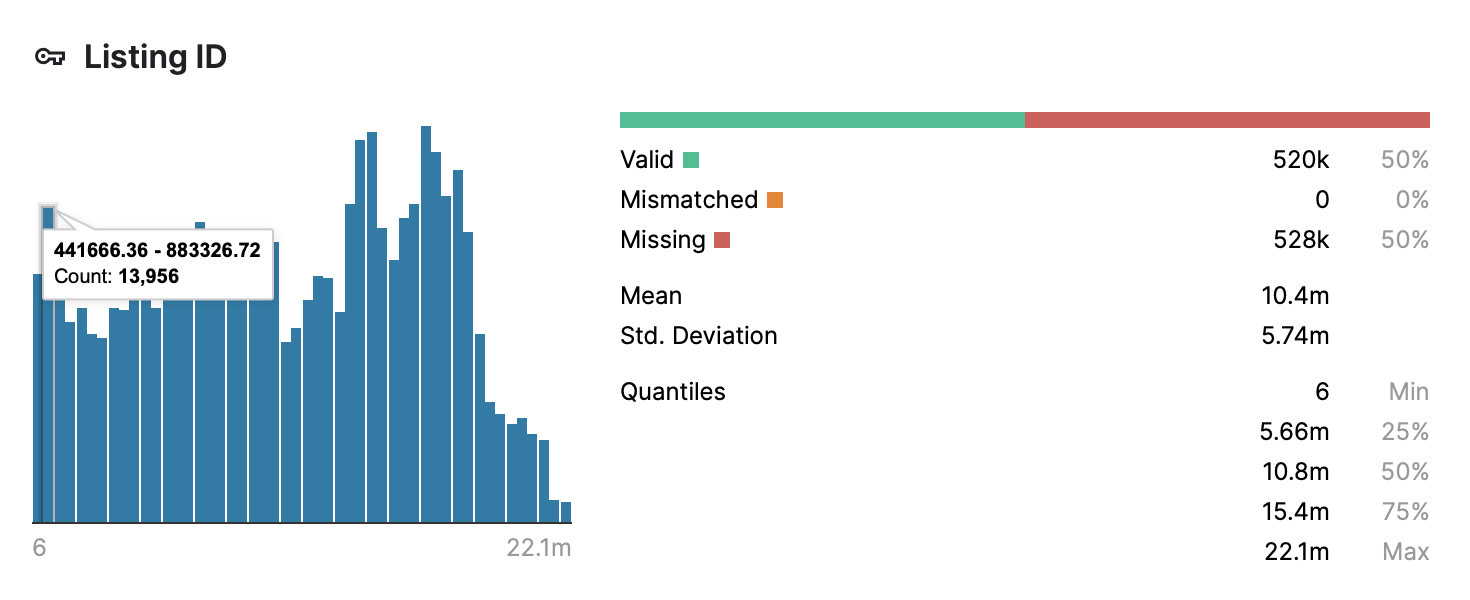


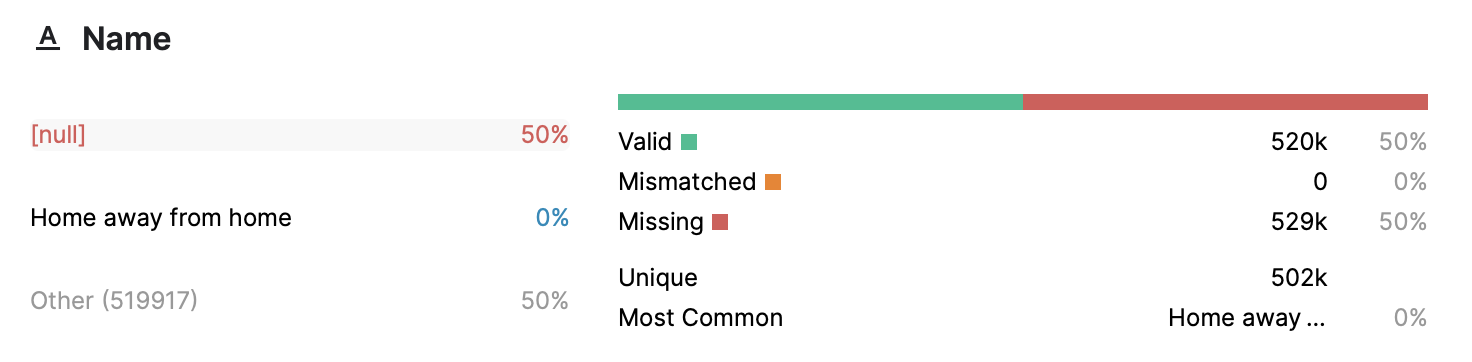


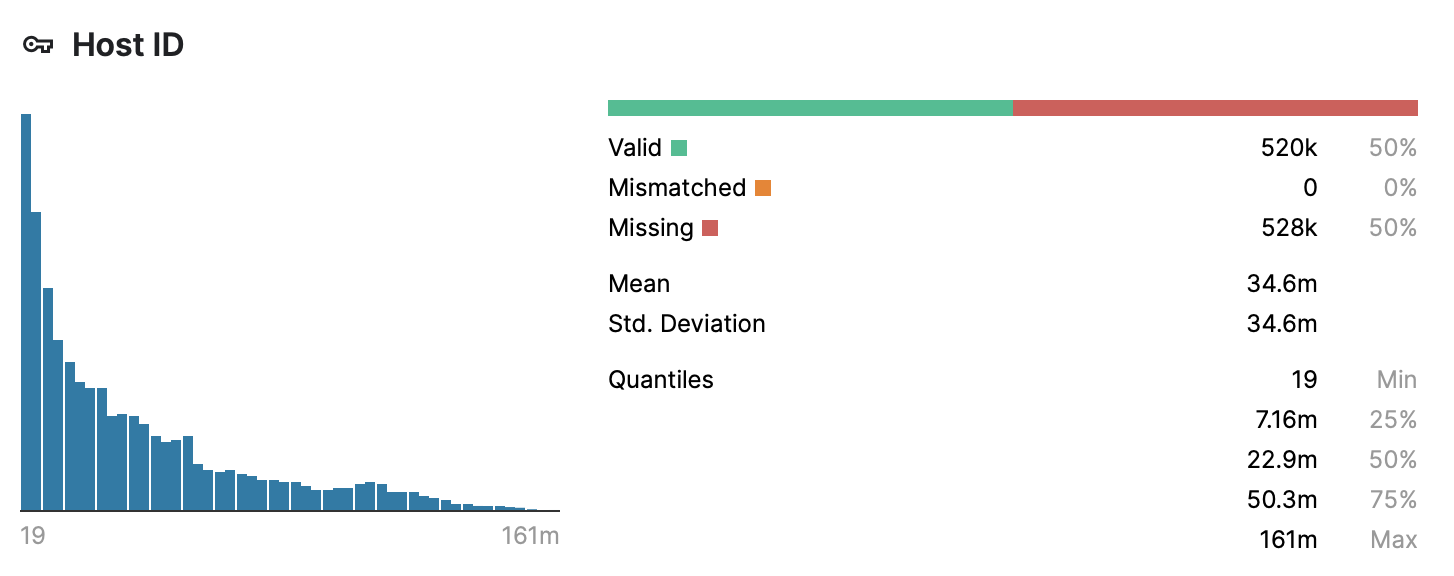


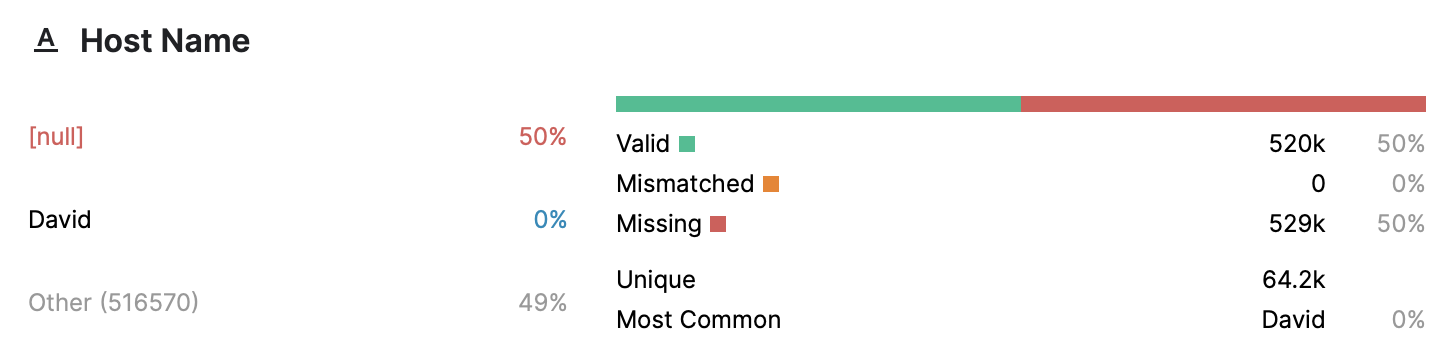


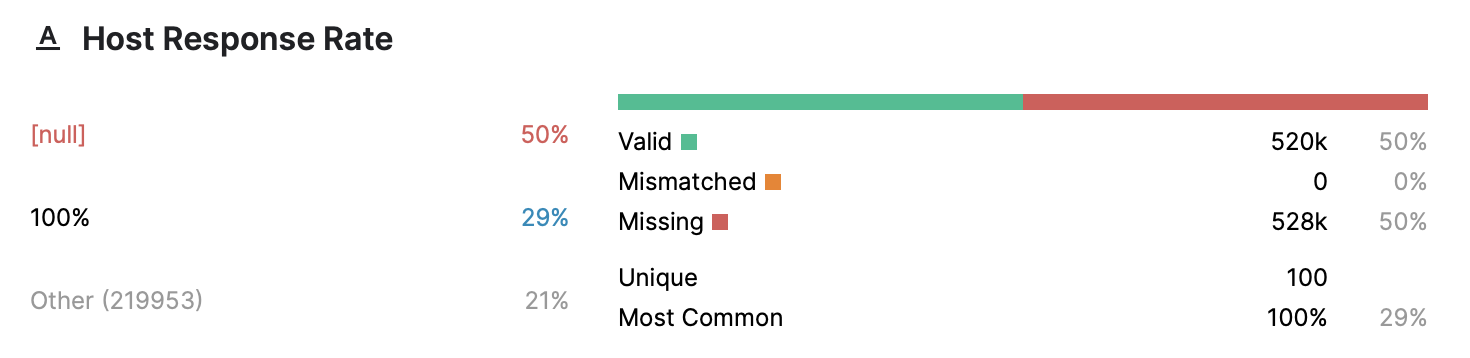


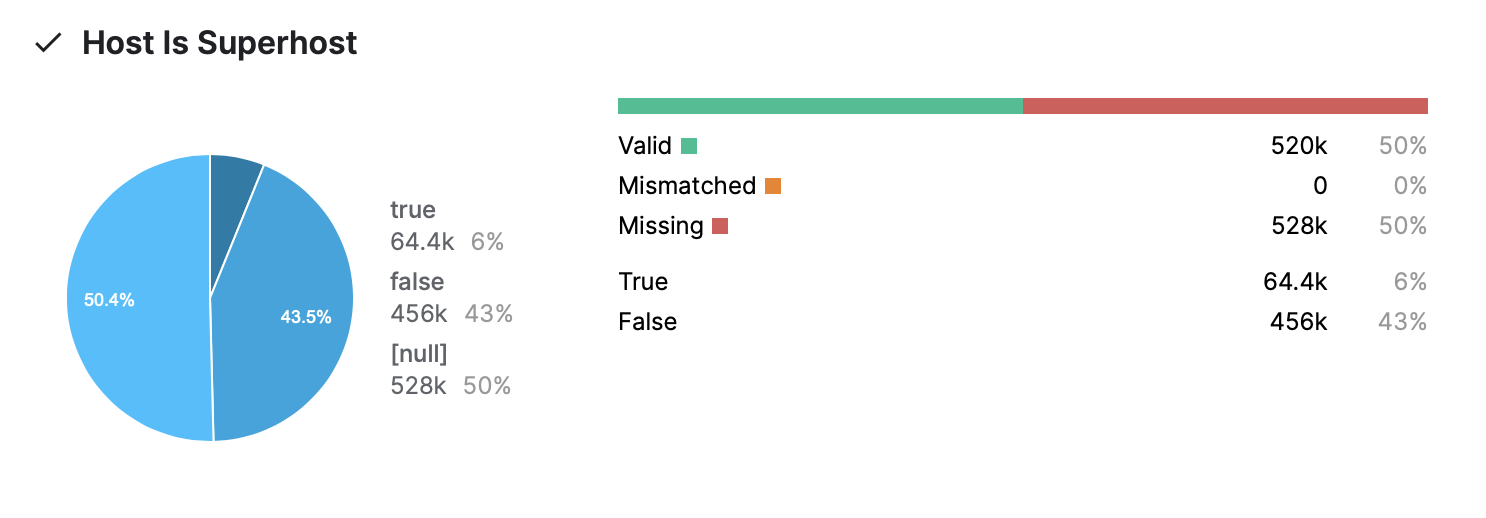


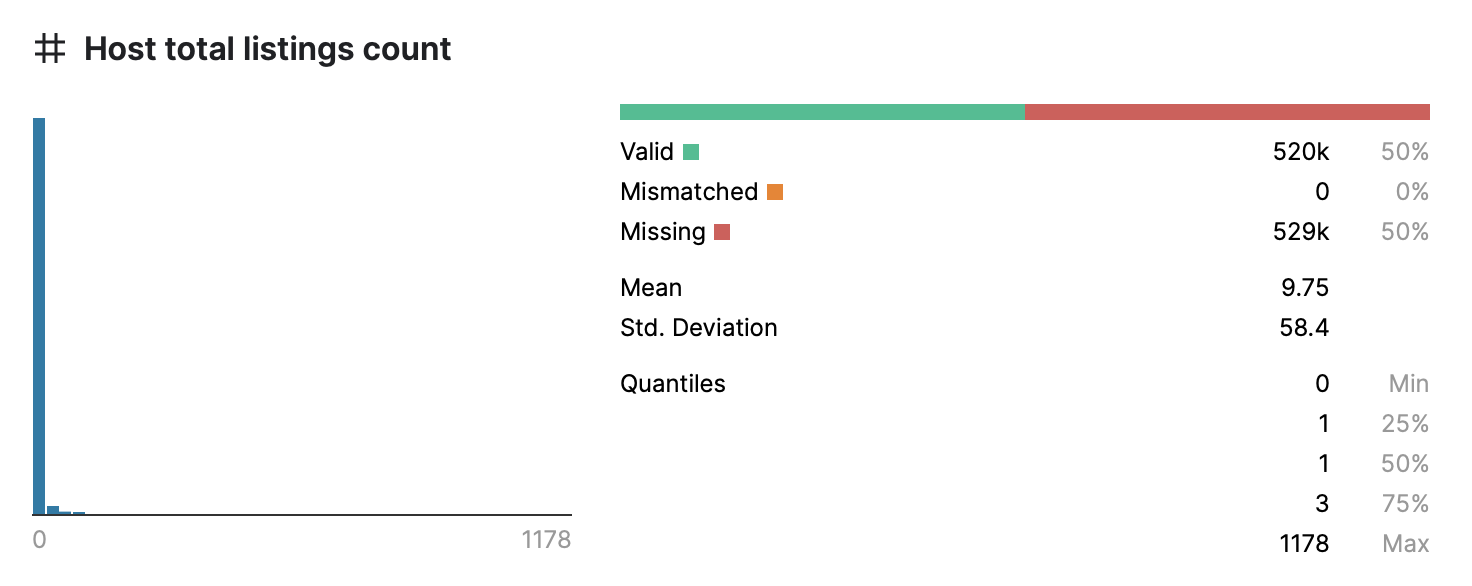


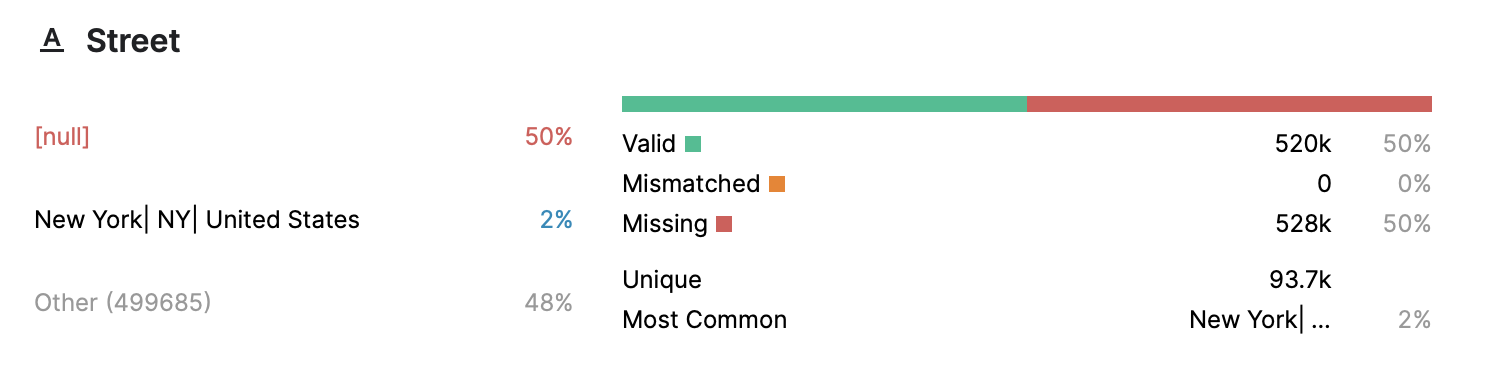


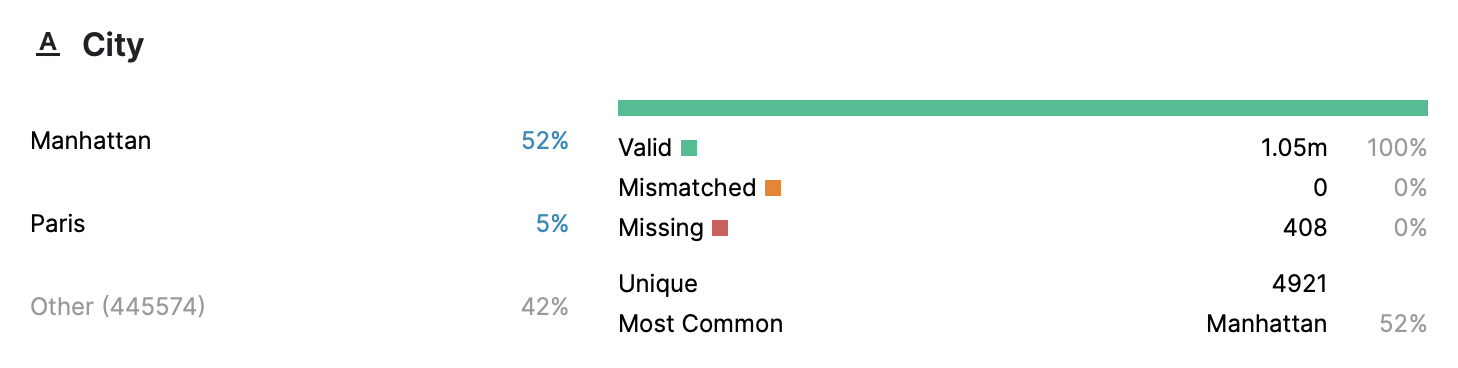


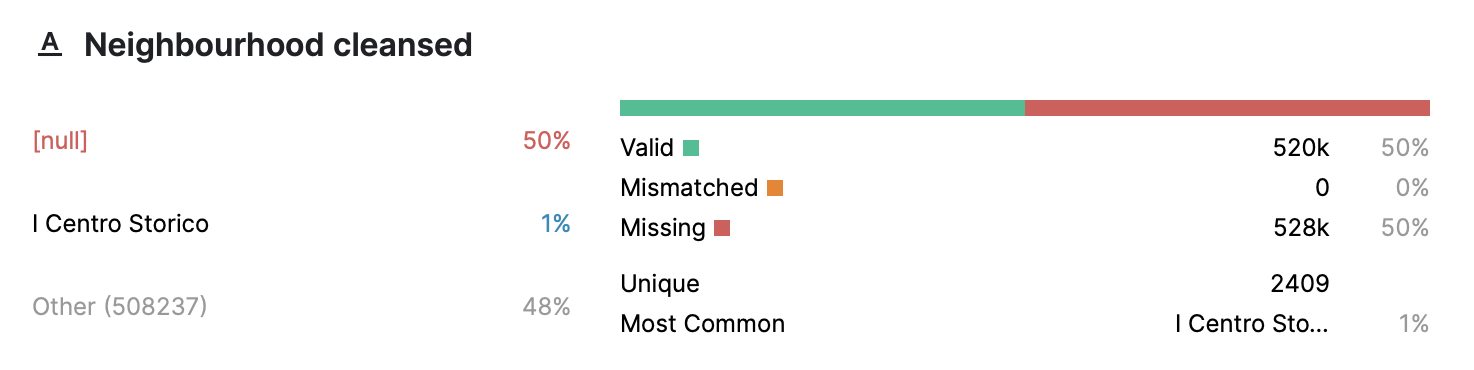












**Road Map**

**// look details in EDA files**

For our case, we picked attributes that are meaningful to our analysis from the four datasets and form our own dataset by filtering and combining the data (which will be discussed in the data preparation section). This new dataset has 295,452 instances and 19 attributes. The description of each attribute is listed below:

**Listing ID**: the ID number of an Airbnb

**Host ID** the ID of the host

**Host total listings count**: the total number of host listings

**Longitude**: the longitude of the Airbnb

**Accommodates**: the number of people an Airbnb can accommodate

**Bathrooms**: number of bathrooms

**Bedrooms**: number of bedrooms

**Price**: price of an Airbnb per day

**Minimum nights**: the minimum number of nights a guest stay

**Maximum nights**: the maximum number of nights a guest stay

**Availability 365**: the number of days available in a year

**Number of reviews**: the total number of reviews

**Review Scores Accuracy**: how accurately did the listing page represent an Airbnb

**Review Scores Cleanliness**: how clean and tidy did the guests feel about an Airbnb

**Review Scores Checkin**: how smoothly did check-in go

**Review Scores Communication**: how well did the guests communicate with the hosts before and during the stay

**Review Scores Location**: how did guests feel about the neighborhood

**Review Scores Value**：did the guest feel that the listing provided good value for the price

**Reviews per month**: the number of reviews a host receives per month

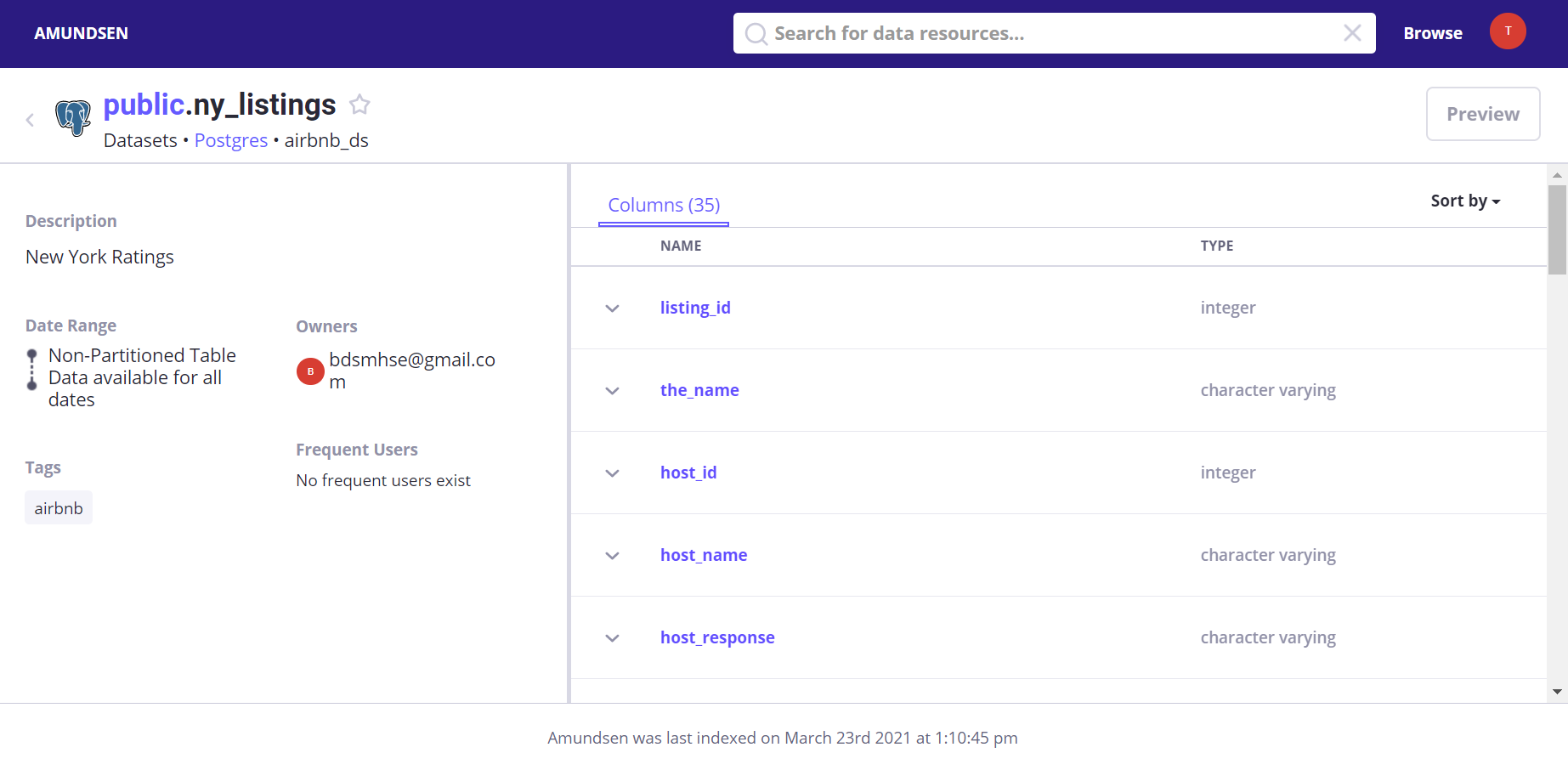
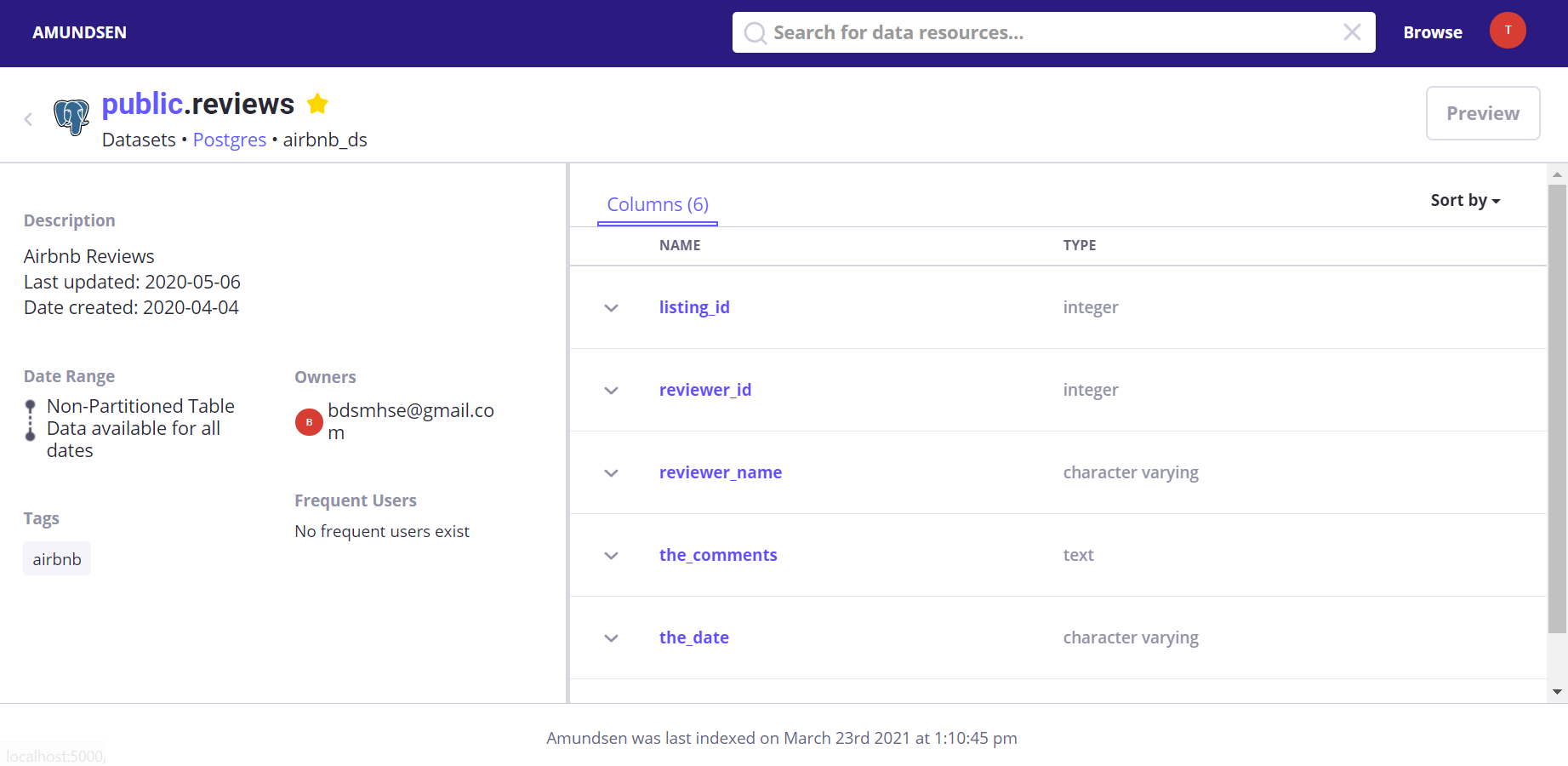
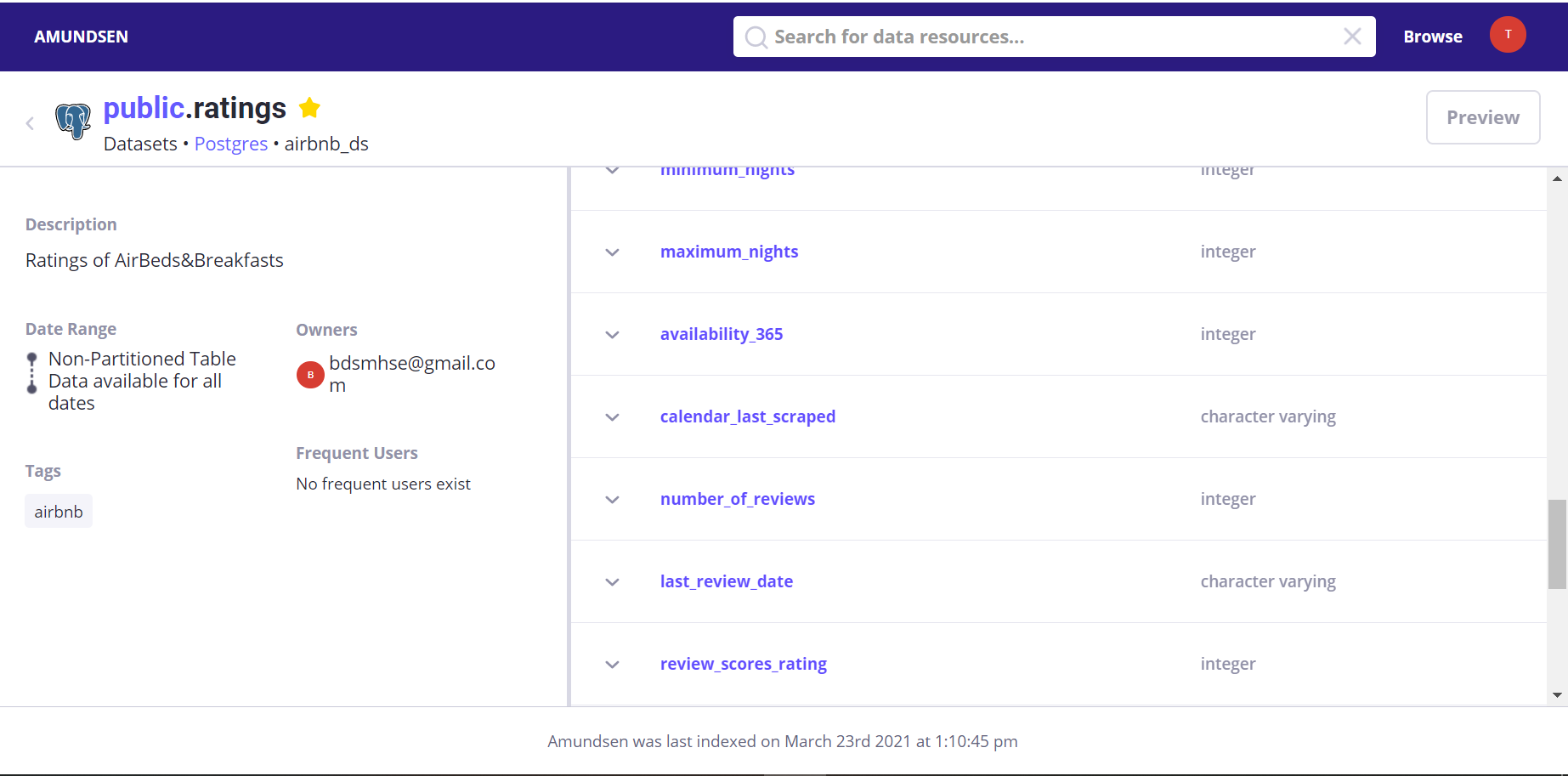
1. Exploratory Data Analysis on AirBnb ratings & reviews dataset
2. As the prices varies a lot in different countries It is reasonably to consider only USA listings
3. Combine those three datasets into one
4. 'df\_final' has 295,452 lines of data and ready to use.
5. Do exploratory data analysis by examine the correlation between Price with number of bedrooms, bathrooms and review scores.
6. Becaue of the large range of 'Price', we need to filter those unnessary data which could influence our analysis. After observation, we found set the range from 0 to 500 is appropriate.
7. See the distribution of numbers of Bedrooms:
8. Most houses have bedrooms from 0 to 6, filter the data
9. The shape of the distributions of 'Numbers of Bedrooms' and the distributions of 'Price' are very similar, which indicates the possibilities between them, and we will do further investigations later. Before that, let's do more distribution graph on other variables.
10. Distribution with more varibles
11. Filter the Varible
12. Filter the data
13. Most listing have 0 - 10 reviews.
14. Do the correlation test to find out the potential correlation with Spearman

**From the result table, we found that *'Price'* and *'Accommodates'* have a correlation coefficient of 0.55, which indicates they are *moderately* correlated, and *'number of Bedrooms'* has a correlation coefficient of 0.46 with *'Price'*, which is the second highest value in all variables, which can be understand, because more bedrooms a house has, the higher the price can be, and more people a house can accommodates, more expensive it will be.**

**MetaData**

**// look details in metadata-2 file**

1. Metadata extraction
2. Define the quality criteria
3. Access data quality

**Amundsen Screenshots**

**Principal Scheme**

