# Dataset description

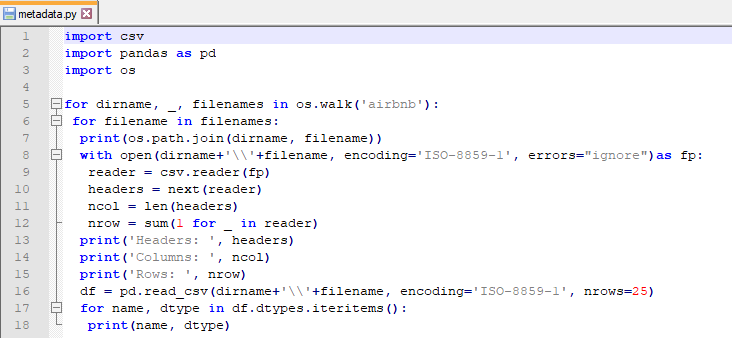
The initial dataset contains four files: airbnb-reviews.csv, airbnb\_ratings\_new.csv, LA\_Listings.csv and NY\_Listings.csv. The first file (the biggest one) contains 6 attributes and about 14.8KK entries. The second file has 35 attributes and 1KK entries. The last two have the same number of attributes (35) and 59K and 75K entries accordingly.

Data exploration revealed the following meaningful attributes:

* For the first file (airbnb-reviews.csv):
  + **Listing ID:** the ID number of Airbnb listing
  + **ID:** the ID number of the review
  + **Date:** Review Date
  + **Reviewer ID:** the ID number of the reviewer
* Other three files:
* **Listing ID**: the ID number of Airbnb listing
* **Host ID** the ID number of the host
* **Host total listings count**: the total number listings the host has
* **Latitude**: the latitude of the Airbnb listing location
* **Longitude**: the longitude of the Airbnb listing location
* **Accommodates**: the number of people an Airbnb listing can accommodate
* **Bathrooms**: number of bathrooms in the Airbnb listing
* **Bedrooms**: number of bedrooms in the Airbnb listing
* **Price**: price of an Airbnb listing per night
* **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

# Metadata extraction

By creating and running a python script:



We were able to retrieve the following metadata:

**airbnb\airbnb-reviews.csv**

**Headers:** ['listing\_id', 'id', 'date', 'reviewer\_id', 'reviewer\_name', 'comments']

**Columns:** 6

**Rows:** 14856578

listing\_id int64

id int64

date object

reviewer\_id int64

reviewer\_name object

comments object

**airbnb\airbnb\_ratings\_new.csv**

**Headers:** ['Listing ID', 'Name', 'Host ID', 'Host Name', 'Host Response Rate', 'Host Is Superhost', 'Host total listings count', 'Street', 'City', 'Neighbourhood cleansed', 'State', 'Country', 'latitude', 'longitude', 'Property type', 'Room type', 'Accommodates', 'Bathrooms', 'Bedrooms', 'Amenities', 'Price', 'Minimum nights', 'Maximum nights', 'Availability 365', 'Calendar last scraped', 'Number of reviews', 'Last Review Date', 'Review Scores Rating', 'Review Scores Accuracy', 'Review Scores Cleanliness', 'Review Scores Checkin', 'Review Scores Communication', 'Review Scores Location', 'Review Scores Value', 'Reviews per month']

**Columns:** 35

**Rows:** 1048575

Listing ID int64

Name object

Host ID int64

Host Name object

Host Response Rate object

Host Is Superhost bool

Host total listings count int64

Street object

City object

Neighbourhood cleansed object

State object

Country object

latitude float64

longitude float64

Property type object

Room type object

Accommodates int64

Bathrooms float64

Bedrooms int64

Amenities object

Price float64

Minimum nights int64

Maximum nights int64

Availability 365 int64

Calendar last scraped object

Number of reviews int64

Last Review Date object

Review Scores Rating float64

Review Scores Accuracy float64

Review Scores Cleanliness float64

Review Scores Checkin float64

Review Scores Communication float64

Review Scores Location float64

Review Scores Value float64

Reviews per month float64

**airbnb\LA\_Listings.csv**

**Headers:** ['Listing ID', 'Name', 'Host ID', 'Host Name', 'Host Response Rate', 'Host Is Superhost', 'Host total listings count', 'Street', 'City', 'Neighbourhood cleansed', 'State', 'Country', 'latitude', 'longitude', 'Property type', 'Room type', 'Accommodates', 'Bathrooms', 'Bedrooms', 'Amenities', 'Price', 'Minimum nights', 'Maximum nights', 'Availability 365', 'Calendar last scraped', 'Number of reviews', 'Last Review Date', 'Review Scores Rating', 'Review Scores Accuracy', 'Review Scores Cleanliness', 'Review Scores Checkin', 'Review Scores Communication', 'Review Scores Location', 'Review Scores Value', 'Reviews per month']

**Columns:** 35

**Rows:** 59865

Listing ID int64

Name object

Host ID int64

Host Name object

Host Response Rate float64

Host Is Superhost bool

Host total listings count int64

Street object

City object

Neighbourhood cleansed object

State object

Country object

latitude float64

longitude float64

Property type object

Room type object

Accommodates int64

Bathrooms float64

Bedrooms int64

Amenities object

Price int64

Minimum nights int64

Maximum nights int64

Availability 365 int64

Calendar last scraped object

Number of reviews int64

Last Review Date object

Review Scores Rating int64

Review Scores Accuracy int64

Review Scores Cleanliness int64

Review Scores Checkin int64

Review Scores Communication int64

Review Scores Location int64

Review Scores Value int64

Reviews per month float64

**airbnb\NY\_Listings.csv**

**Headers:** ['Listing ID', 'Name', 'Host ID', 'Host Name', 'Host Response Rate', 'Host Is Superhost', 'Host total listings count', 'Street', 'City', 'Neighbourhood cleansed', 'State', 'Country', 'latitude', 'longitude', 'Property type', 'Room type', 'Accommodates', 'Bathrooms', 'Bedrooms', 'Amenities', 'Price', 'Minimum nights', 'Maximum nights', 'Availability 365', 'Calendar last scraped', 'Number of reviews', 'Last Review Date', 'Review Scores Rating', 'Review Scores Accuracy', 'Review Scores Cleanliness', 'Review Scores Checkin', 'Review Scores Communication', 'Review Scores Location', 'Review Scores Value', 'Reviews per month']

**Columns:** 35

**Rows:** 75749

Listing ID int64

Name object

Host ID int64

Host Name object

Host Response Rate float64

Host Is Superhost bool

Host total listings count float64

Street object

City object

Neighbourhood cleansed object

State object

Country object

latitude float64

longitude float64

Property type object

Room type object

Accommodates int64

Bathrooms float64

Bedrooms int64

Amenities object

Price int64

Minimum nights int64

Maximum nights int64

Availability 365 int64

Calendar last scraped object

Number of reviews int64

Last Review Date object

Review Scores Rating int64

Review Scores Accuracy int64

Review Scores Cleanliness int64

Review Scores Checkin int64

Review Scores Communication int64

Review Scores Location int64

Review Scores Value int64

Reviews per month float64

# Define the quality criteria

As a data quality assessment system, we will use following categories:

|  |  |
| --- | --- |
| Category | Description |
| Accuracy (0-100%) | The degree to which the data tested is in argument with a standard source, which is assumed to be correct. It is actually a count of number of times an attribute was incorrect. |
| Integrity (0-100%) | The data is not inadvertently deleted, and changes made are applied. You can measure integrity by tracking the number of issues logged or records affected from a call center. |
| Secure (yes/no) | Users and systems that should not have access rights are properly controlled. You may measure this by logging incidents, tracking the number of individuals with write access, or measuring conformance to security standards by role. |
| Completeness (0-100%) | The degree to which all of the required data values appear in the data record or business object. You can measure this by tracking the percentage of records with one or more missing required attributes. |
| Validity (high/low) | The degree to which the tested data conforms to the data validation requirements. This is a count of the number of rules in conformance. |
| Timeliness (0-100%) | The degree to which the data is provided at the time required or specified. You can track this by comparing planned versus actual availability. |
| No redundancy (yes/no) | The degree to which there are multiple databases. To measure this, you can track the number of copies of that record throughout your landscape. |
| No duplicates (yes/no) | The degree to which there are duplicate records for the same item. |
| Relevance (yes/no) | The data should be tied to a business process, metric or objective. |
| Accessibility (high/low) | The ease of use for data consumers to access and consume the data. |
| Availability (high/low) | The degree to which data is available for access to the users. |
| Consistency (high/low) | The ability to use or interpret the data accurately across multiple domains, for example multiple product descriptions for the same product across the application. |

And also, will added the data requirements. To each data is described the format.

|  |  |
| --- | --- |
| **Name** | Requirements |
| Listing ID | Digits without special symbols and letters. Length to 50 symbols. |
| Host ID | Digits and letters without special symbols. Length to 50 symbols. |
| Host total listings count | Digits. Length to 50 |
| Longitude | Digits with comma. Float number. Length to 35 symbols. |
| Accommodates | Digits without special symbols and letters. Length to 50 |
| Bathrooms | Digits without special symbols and letters. Length to 35 |
| Bedrooms | Digits without special symbols and letters. Length to 35 |
| Price | Digits without special symbols and letters. Length to 50 |
| Minimum nights | Digits without special symbols and letters. Length to 2 |
| Maximum nights | Digits without special symbols and letters. Length to 10 |
| Availability 365 | Digits without special symbols and letters. Length to 3 |
| Number of reviews | Digits without special symbols and letters. Length to 3 |
| Review Scores Accuracy | Digits without special symbols and letters. Length to 3 |
| Review Scores Cleanliness | Digits without special symbols and letters. Length to 3 |
| Review Scores Checkin | Digits without special symbols and letters. Length to 3 |
| Review Scores Communication | Digits without special symbols and letters. Length to 3 |
| Review Scores Location | Digits without special symbols and letters. Length to 3 |
| Reviews per month | Digits without special symbols and letters. Length to 3 |

# Access data quality

listing\_id int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

id int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

date object

|  |  |
| --- | --- |
| Accuracy | 80 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

reviewer\_id int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

reviewer\_name object

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

comments object

|  |  |
| --- | --- |
| Accuracy | 20 |
| Integrity | 100 |
| Secure | no |
| Completeness | 50 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | Yes |
| Availability | Yes |
| Consistency | yes |

Listing ID int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

Name object

|  |  |
| --- | --- |
| Accuracy | 80 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | no |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

Host ID int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

Host Name object

|  |  |
| --- | --- |
| Accuracy | 80 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 80 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

Host Response Rate object

|  |  |
| --- | --- |
| Accuracy | 80 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

Host Is Superhost bool

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

Host total listings count int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | yes |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

Street object

|  |  |
| --- | --- |
| Accuracy | 50 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | no |
| No duplicates | no |
| Relevance | Yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

City object

|  |  |
| --- | --- |
| Accuracy | 50 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Neighbourhood cleansed object

|  |  |
| --- | --- |
| Accuracy | 50 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

State object

|  |  |
| --- | --- |
| Accuracy | 50 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

Country object

|  |  |
| --- | --- |
| Accuracy | 50 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

latitude float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

longitude float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

Property type object

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

Room type object

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | no |
| Relevance | yes |
| Accessibility | no |
| Availability | yes |
| Consistency | yes |

Accommodates int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | yes |
| No duplicates | yes |
| Relevance | yes |
| Accessibility | no |
| Availability | no |
| Consistency | yes |

Bathrooms float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Bedrooms int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Amenities object

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Price float64

|  |  |
| --- | --- |
| Accuracy | 80 |
| Integrity | 100 |
| Secure | no |
| Completeness | 80 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Minimum nights int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Maximum nights int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Availability 365 int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | high |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Calendar last scraped object

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Number of reviews int64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Last Review Date object

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Review Scores Rating float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Review Scores Accuracy float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Review Scores Cleanliness float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Review Scores Checkin float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Review Scores Communication float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Review Scores Location float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Review Scores Value float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |

Reviews per month float64

|  |  |
| --- | --- |
| Accuracy | 100 |
| Integrity | 100 |
| Secure | no |
| Completeness | 100 |
| Validity | low |
| Timeliness | 100 |
| No redundancy | No |
| No duplicates | No |
| Relevance | yes |
| Accessibility | No |
| Availability | yes |
| Consistency | yes |