# Datasheet for dataset “add dataset name here”

Questions from the [Datasheets for Datasets](https://arxiv.org/abs/1803.09010) paper, v7.

Jump to section:

* [Motivation](#motivation)
* [Composition](#composition)
* [Collection process](#collection-process)
* [Preprocessing/cleaning/labeling](#preprocessingcleaninglabeling)
* [Uses](#uses)
* [Distribution](#distribution)
* [Maintenance](#maintenance)

## Motivation

### For what purpose was the dataset created?

*We are building a ML model that predict the prices for an Airbnb listing based on the location in NY City.*

### Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

*Airbnb*

*http://insideairbnb.com/*

### Who funded the creation of the dataset?

*Veralyn Williams and Mark Winston Griffith, community organizers and investigative journalists from the Brooklyn Movement Center / Brooklyn Deep*

*Noel Hidalgo from Beta NYC,*

*Dennis George, the Principal of George Architect*

*Paula Z. Segal, Esq.*

*Samantha Box, Jason Brathwaite, Chris Carter, Andrew Cox, Juny E. François, Esq., Mark Winston Griffith, Noel Hidalgo, Chris Hwong, Clarisa James, Harvir Kaur, Joanne Miguel, Nicole Salzano, Paul Smullen and Veralyn Williams.*

*Freepik from www.flaticon.com.*

*John Morris*

*Samantha Box*

## Composition

### What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

* *Users*
* *Ratings*
* *People*
* Prices
* Locations
* Coordenates
* Reviews
* Types of listings

### How many instances are there in total (of each type, if appropriate)?

*There are 10,019 entries of listings*

*Data columns (total 16 columns):*

*# Column Non-Null Count Dtype*

*--- ------ -------------- -----*

*0 listing\_id 10019 non-null int64*

*1 name 10014 non-null object*

*2 host\_id 10019 non-null int64*

*3 host\_name 10017 non-null object*

*4 neighbourhood\_full 10019 non-null object*

*5 coordinates 10019 non-null object*

*6 room\_type 10019 non-null object*

*7 price 9781 non-null object*

*8 number\_of\_reviews 10019 non-null int64*

*9 last\_review 7944 non-null object*

*10 reviews\_per\_month 7944 non-null float64*

*11 availability\_365 10019 non-null int64*

*12 rating 7944 non-null float64*

*13 number\_of\_stays 7944 non-null float64*

*14 5\_stars 7944 non-null float64*

*15 listing\_added 10019 non-null object*

### Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

*This data is a sample just for the month of July 2019 for Airbnb in New York.*

### What data does each instance consist of?

*“Raw” data consist of the following features:*

* *listing\_id: The unique identifier for a listing*
* *description: The description used on the listing*
* *host\_id: Unique identifier for a host*
* *host\_name: Name of host*
* *neighbourhood\_full: Name of boroughs and neighbourhoods*
* *coordinates: Coordinates of listing (latitude, longitude)*
* *Listing added: Date of added listing*
* *room\_type: Type of room*
* *rating: Rating from 0 to 5.*
* ***price: Price per night for listing(Which is the label)***
* *number\_of\_reviews: Amount of reviews received*
* *last\_review: Date of last review*
* *reviews\_per\_month: Number of reviews per month*
* *availability\_365: Number of days available per year*
* *Number of stays: Total number of stays thus far*

### Is there a label or target associated with each instance?

*Price is our label.*

### Is any information missing from individual instances?

*Yes, there is missing information for the following features:*

*Out of the 10019 entries, name is missing 5 values, host\_name 2 values, price 238 values, and incidentally there is the same number of missing values 2075 for the following features: last review, reviews\_per\_month, raiting, number\_of\_stays and 5\_stars. Here is an evident correlation between these instances.*

### Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?

*Yes because the coordinates of their home are made publicly available*

### Are there recommended data splits (e.g., training, development/validation, testing)?

*We will split the data set into 70% for training and 30% for testing in order to evaluate our model*

### Are there any errors, sources of noise, or redundancies in the dataset?

*Yes, there is errors and potential problems with the raw data listed below:*

1. *We have many instances with missing values*
2. *We also found that the instances ‘price’ and ‘coordinates’ are objects and should be numerical values.*
3. *Likewise Is evident here that for ‘Entire home/apt’ and ‘home’ is one type of room but we have 2 types, also for ‘Private room ‘ we have ‘Private’ and ‘PRIVATE ROOM’ which all 3 appears to be the same type. And lastly for ‘Shared room’ we have ‘ Shared room’ in which evident that the spacing used when typing may be causing this redundancy.*
4. *We have encountered that we have 40 instances that are duplicated entries, of which 26 are identical duplicates and 14 others have some different values in ‘price’ and the ‘listing\_added’ instances.*
5. *We have instances where the star rating is higher than the maximum rating of 5 stars.*
6. *The neighborhoods consist of 2 parts separated by ,* *and can be separated and consider them as individual features*

### Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals’ non-public communications)?

*Yes, the dataset contains private and sensitive information of Airbnb hosts like:*

* *Name*
* *Host ID*
* *Host name*
* *Coordinates*

### Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

*No*

### Does the dataset relate to people?

*Yes, it relates to Airbnb hosts and users*

### Does the dataset identify any subpopulations (e.g., by age, gender)?

*No*

### Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

*Yes, as one of the instances is the host name and the host ID, and you also have the coordinates and the neighborhood in which the Airbnb is located.*

### Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

*No*

## Collection process

### How was the data associated with each instance acquired?

*Trough Airbnb’s listing. The dataset was sourced directly by Airbnb.*

### What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

### *The data utilizes public information compiled from the Airbnb web-site including the availabiity calendar for 365 days in the future, and the reviews for each listing. Data is verified, cleansed, analyzed and aggregated.*

### Who was involved in the data collection process (e.g., students, contractors) and how were they compensated (e.g., how much these people were paid)?

*The dataset was sourced directly by Airbnb.*

### Over what timeframe was the data collected?

*July 2019*

### Were any ethical review processes conducted (e.g., by an institutional review board)?

*NA*

### Does the dataset relate to people?

*Yes*

### Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

*Third-Part (Airbnb)*

### Were the individuals in question notified about the data collection?

*Yes, is part of terms and conditions*

### Did the individuals in question consent to the collection and use of their data?

*Yes, when they enroll to the platform is part of terms and conditions*

### If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

*No*

### Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

*No*

## Cleaning

*The questions in this section are intended to provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks involving word order.*

### Was any cleaning of the data done (e.g, removal of instances, processing of missing values)?

### *We have many instances with missing values: name is missing 5 values, host\_name 2 values, price 238 values, and incidentally there is the same number of missing values 2075 for the following features: last review, reviews\_per\_month, raiting, number\_of\_stays and 5\_stars.*

### *We also found that the instances ‘price’ and ‘coordinates’ are objects and should be numerical values, this is a source of noise. For price we need to remove the $ sign, for the coordinates we removed the “()”and then split it into two columns Latitude and Longitude using the “,” and then removing the ‘coordinates’ column*

### *Likewise Is evident here that for ‘Entire home/apt’ and ‘home’ is one type of room but we have 2 types, also for ‘Private room ‘ we have ‘Private’ and ‘PRIVATE ROOM’ which all 3 appears to be the same type. And lastly for ‘Shared room’ we have ‘ Shared room’ in which evident that the spacing used when typing may be causing this redundancy. We need to clean the names.. We renamed the room types to: ’’Private Room’, ‘Shared Room’ and ‘Entire House’*

### *We have encountered that we have 40 instances that are duplicated entries, of which 26 are identical duplicates and 14 others have some different values in ‘price’ and the ‘listing\_added’ instances.*

### *We have instances where the star rating is higher than the maximum rating of 5 stars. We will need to put all the higher values of stars equals to 5.*

### *The neighborhoods we can simplify only by the main ones without going to the suburb part as we can use the coordinates feature for that. We split the neighborhood into suburb and borough creating 2 different columns so we can see the 5 suburbs and make it easier to make correlations later on with the coordinates, we then removed the “neighbourhood\_full” column*

### *We created a new feature named price\_bucket so we can group by price ranges like this: bins=[0,50,100,500,1000,8000]*

### *I have dropped the following columns that can be consider unethical 'name' 'host\_id' and 'host\_name'*

### *Check how many duplicates we have in our data based on the listing ID and we have 40 so I decided to delete those rows*

### *We have 238 null values for the price so I will use the mean price for each of the three room types to replace those null values.*

### *For the 4 listings with a rating higher than five stars I will assume this equals to five and replace those values with 5*

### *There are NaN values in 2075 rows in all have this columns last\_review, reviews\_per\_month, rating, number\_of\_stays and 5\_stars with NaN values so there is a correlation probably no raiting was given and so the other fields were left blank so I decided to delete those rows.*

### *I need to convert the values in the following columns as categories so the model can use this data: room\_type, borough and suburb. The model won’t allow me to use the 5\_star column as it stats with a numerical value so I created a copy of that column named fivestars*

### Was the “raw” data saved in addition to the cleaned data (e.g., to support unanticipated future uses)?

*Yes, there is a separate file with the cleaned data here: /content/cleaned\_airbnb.csv* *https://github.com/adelnehme/python-for-spreadsheet-users-webinar/blob/master/datasets/airbnb.csv?raw=true*

## Uses

*These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.*

### Has the dataset been used for any tasks already?

*No*

### Is there a repository that links to any or all papers or systems that use the dataset?

*No*

### What (other) tasks could the dataset be used for?

*Rate of new Airbnb hosts/frequency of usage/popular areas*

### Is there anything about the composition of the dataset or the way it was collected and cleaned that might impact future uses?

*We made assumption for the room type private we also considered is one private room and not a private 2 bedroom etc.*

### Are there tasks for which the dataset should not be used?

*NA*

## Distribution

### Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

*Yes, The dataset is publicly available and made for consumption*

### How will the dataset will be distributed (e.g., zip file, website, GitHub)?

*It will be shared via Google Drive*

### Will the dataset be distributed under a copyright or other intellectual property (IP) license ?

## Maintenance

*These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.*

### Who is supporting/hosting/maintaining the dataset?

### How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

### Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

*If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

### If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

*If so, please describe these limits and explain how they will be enforced.*

### Will older versions of the dataset continue to be supported/hosted/maintained?

*If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

### If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

*If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*