Data set

Berlin Airbnb Ratings - An Evaluation of Berlin Airbnb Properties Based on Host Ratings.

Data Source

The data is sourced from publicly available information from Inside Airbnb and scraped by Makeover Monday [1]. *Inside Airbnb* is a mission-driven project that provides data and advocacy about Airbnb's impact on residential communities. Their motto is "Adding data to the debate." They aim to empower communities with data and information to understand, decide, and control the role of renting residential homes to tourists. *Makeover Monday*, for its part, is an organization that shares charts and their data weekly to encourage everyone of all skills to speak about data and retell stories.

Data Profile

This dataset contains information on Airbnb listings in Berlin, including reviewer ratings and comments. It permits analyzing the property characteristics, host characteristics, and guest experience in the German capital.

The original data set has 47 columns and 456,961 entries.

Data cleaning

Columns dropped:

Column Name	Reason
Reviewer Name	Personally Identifiable Information (PII)
Listing URL	Information not necessary
Listing Name	Information not necessary. They are very descriptive names and don't add anything
Host URL	Information not necessary
Host Name	Personally Identifiable Information (PII)

[1] License: Dataset copyright by authors.

Neighborhood	Redundant with the 'Neighborhood Group' column and with lower quality
City	Many entries are neighborhoods and this column is not relevant for the analysis because our goal is the city of Berlin
Country Code	All the values are 'DE' except one 'GB' and one 'NL,' but these two rows are empty
Country	All the values are 'Germany' except one 'United Kingdom' and one 'Netherlands', but these two rows are empty

Renaming columns:

For consistency and tip errors, I have renamed the following columns:

- review_date: Review Date
- Accomodates: Accommodates

Finding Missing Values:

Considering that the data set was created to analyze the ratings, that the rows with no value in 'Value Rating' are almost completely blank, despite being useless, and only 1% of the data set, the best option is to delete these rows.

The column 'Square Feet' is empty in 94%, so I have decided to delete it.

I have imputed the remaining missing values in the rating columns with the median of each one.

Changing Data Types:

To exclude the columns with identification numbers while executing statistics, I have changed their type to string:

- Review ID

- Reviewer ID
- Listing Id
- Host ID

Values in ‘Review ID’ and ‘Reviewer ID’ had a decimal at the end because they were floats, so I have used the method .str.split() to remove ‘.0’.

On the other hand, I have changed the data type in ‘Price’ to float to include the column while executing statistics, removing the ‘,’ character when necessary.

I have converted the data to datetime in the columns:

- Review Date
- First Review
- Last Review
- Host Since

I have converted the Postal Code column to string and split the values to remove the final .0 where necessary.

I have mapped the values f/t as False/True and converted the data to boolean in the columns:

- Is Superhost
- Is Exact Location
- Instant Bookable
- Business Travel Ready

Comments and Host Response Time converted to string.

Converted Percentage string in Host Response Rate to numeric, removing first the % character.

Duplicates

No duplicates were found.

Subsetting

I have decided to create a new data frame with the columns 'Comments' and delete it in this one. Maybe it is useful at some point, but it is not for this analysis. Additionally, some comments have PII.

Other cleaning

I have corrected the name of the 'Neighborhood Groups' with false characters: Neukölln, Schöneberg and Köpenick.

The reviews are dated between 20.06.2009 and 14.05.2019.

The minimum in Accommodates, Bedrooms, Beds, and Bathrooms seems to be 0, which at first thought makes little sense in an Airbnb. However, I have done a little research and it is possible.

The maximum value for 'Min Nights' is 1,000 in four Listing IDs. However, there are different reviews from different reviewers in a shorter period of time, so it seems to be an error and I have imputed the value with the mode of 2 nights.

In the column 'Price', one listing is priced as free (0.00 €), something that seems to be an error. There are 144 rows priced under 8.00 €. There are listings with very high prices too, up to 9,000€. A quick research shows that it is an error, so considering the size of the data set and that we don't have a way to get the correct information, I have decided to delete those rows with extreme values in the column price (< 8 € & > 1000 €).

The clean data set has 25 columns and 451,600 entries.

Descriptive statistical analysis

	Accommodates	Bathrooms	Bedrooms	Beds	Price	Guests Included	Min Nights	Reviews	Overall Rating	Accuracy Rating	Cleanliness Rating	Checkin Rating	Communication Rating	Location Rating	Value Rating
count	451600	450664	450891	451529	451600	451600	451600	451600	451600	451600	451600	451600	451600	451600	451600
mean	3	1,09	1,22	1,92	67.73	2	5	104	94,57	9,50	9,50	9,80	9,81	9,64	9,43
min	1	0,00	0,00	0,00	9,00	1	1	1	20,00	2,00	2,00	2,00	2,00	2,00	2,00
25%	2	1,00	1,00	1,00	37,00	1	1	29	92,00	10,00	9,00	10,00	10,00	9,00	9,00
50%	2	1,00	1,00	1,00	53,00	1	2	76	96,00	10,00	10,00	10,00	10,00	10,00	9,00
75%	4	1,00	1,00	2,00	80,00	2	3	150	98,00	10,00	10,00	10,00	10,00	10,00	10,00
max	16	8,50	10,00	22,00	888,00	16	400	545	100,00	10,00	10,00	10,00	10,00	10,00	10,00
std	2	0,34	0.72	1,53	51,19	1	15	97	4,39	0,47	0,68	0,44	0,44	0,53	0,58

Limitations and Ethics

The main limitation of this data set is the high percentage of missing values and the impossibility of confirming the data in columns such as 'Square Feet', which could give good insights about the different prices of the listings. be, especially in 'Prize.'

We cannot be sure if some listings are fraudulent or if the reviewer is fudging data.

Questions to Explore

- How are the listings spread through the city of Berlin? Are neighborhoods with more than others?
- What does influence the different price ranges?
- Are there hosts with more than one listing? Has this had any influence on the overall ranting or on the prize?
- Which are the five best-rated listings, and what do they have in common?
- Which are the five worst-rated, and what do they have in common?
- Which are the five most expensive listings, and what do they have in common?
- Which are the five cheaper listings, and what do they have in common?
- Is there any period of the year with more reviews than others?