AirBnB Berlin Analysis

by Kilian Loeffner

Data

a. Data Source

This data is publicly available open source data, downloaded from KAGGLE.com. **Source**

Link: Berlin Airbnb Ratings

b. Data Collection

The data was collected from AirBnB.com, one of the biggest holiday accommodation rental websites. The data was collected in 2019.

c. Data Content

The data set contains information about listings on AirBnB. Including Listing Name, Host Name, Host Response Time, Is Superhost, neighboorhood, Price, City, Room Type, Reviews, Overall Rating, Cleanliness Rating etc.

d. Data Profile

1. 46 Columns, 456961 rows, after data quality check 39 columns, 452805 rows.
2. Dropped all rows without rating ID, cause the rating is essential for my analysis. Dropped 7 unnecessary columns not needed for analysis.
3. No duplicates in the Dataset
4. Some missing values couldn't be imputed and are in the Dataset as NaN values.

Index	*Column Name*	*Column Description*	*Time Variant/ Invariant*	*Data type*
1	Review ID	Internal ID for review	Invariant	Qualitative
2	Review_Date	Date of the review	Invariant	Quantitative
3	Reviewer ID	Internal ID for Reviewer	Invariant	Qualitative
4	Rating_Comment	Comment on Rating	Invariant	Qualitative
5	Listing ID	Internal ID for Listing	Invariant	Qualitative
6	Listing Name	Name of the listing	Invariant	Qualitative
7	Host ID	Internal ID for Host	Invariant	Qualitative

8	Host Since	Date since when Host	Invariant	Quantitative
9	Host Response Time	Usual response time from host	Variant	Qualitative
10	Host Response Rate	How often the host gives a response	Variant	Quantitative
11	Is Superhost	Superhost True/False	Variant	Qualitative
12	Neighborhood	Neighborhood of listing	Invariant	Qualitative
13	Neighborhood Group	Neigboorhood group of listing	Invariant	Qualitative
14	City	City of listing	Invariant	Qualitative
15	Postal Code	Postal code of listing	Invariant	Qualitative
16	Country	Country of listing	Invariant	Qualitative
17	Latitude	Latitude of listing	Invariant	Qualitative
18	Longitude	Longitude of listing	Invariant	Qualitative
19	Property Type	Type of the property (Apartment, House etc.)	Invariant	Qualitative
20	Room Type	Type of the room (Private Room, Entire home/apt etc.)	Invariant	Qualitative
21	Accommodates	Number of people the listing can accommodate	Invariant	Quantitative
22	Bathrooms	Number of Bathrooms	Invariant	Quantitative
23	Bedrooms	Number of Bedrooms	Invariant	Quantitative
24	Beds	Number of Beds	Invariant	Quantitative
25	Price	Price per night		Quantitative
26	Guests Included	Number of guests included in the price	Invariant	Quantitative
27	Min Nights	Minimum nights required to book	Invariant	Quantitative
28	Reviews	Number of reviews	Variant	Quantitative
29	First Review	Date of first review	Invariant	Quantitative
30	Last Review	Date of last review	Invariant	Quantitative
31	Overall Rating	Listings Overall rating	Variant	Quantitative
32	Accuracy Rating	Rating how accurate the listing is	Variant	Quantitative
33	Cleanliness Rating	Rating how clean the accommodation is	Variant	Quantitative
34	Checkin Rating	Rating for Check-In	Variant	Quantitative

35	Communication Rating	Rating for communication with host	Variant	Quantitative
36	Location Rating	Rating for listing location	Variant	Quantitative
37	Value Rating	Rating for the value of listing	Variant	Quantitative
38	Instant Bookable	Instant Bookable True/False	Invariant	
39	Business Travel Ready	Business Travel Ready True/False	Invariant	

e. Limitations and ethics:

- **Limitations:**
  - **T**here are a few limitations in this dataset:
    - Missing values in several columns
    - I don't know for sure that this dataset contains all listings from AirBnB, possible sampling bias.
    - Possible measurement bias as I don't know how the data was collected
- **Ethical Issues:** I removed all personal data for privacy reasons, otherwise I don't see ethical issues. All data are openly available on AirBnB website so there shouldn't be other ethical issues.

f. Questions:

- Have some neighborhoods far better ratings than others?
- Which neighborhood is most pricey and least pricey?
- Are the ratings from Superhosts higher in general?
- Are there certain criteria that are outstanding on listings with very good ratings
- Is there a correlation between the overall rating and the price?
- Are there certain neighborhoods where the cleanliness rating is unusually low?
- Are there some outstanding criteria that are crucial for good overall ratings.

I choose this dataset because I’m born and raised near Berlin and therefore I have a connection to Berlin, also I like AirBnB for travels. By analyzing the dataset I want to find out if there are specific criteria contributing to a positive stay.