Ioana Rusu

AIRBNB Berlin Analysis

Data

1. Data Source

This public dataset is part of Airbnb, and the original source can be found here: http://insideairbnb.com/get-the-data/

2. Data Collection

The data was collected from AirBnB.com, one of the biggest holiday accommodation rental websites. The data has reviews from 2012-2023

3. Data Content

The data set contains information about listings on AirBnB. Including id, name, Host Name, Neighborhood, Price, Room Type, Availability, Reviews, etc.

4. Data Profile

1. 18 Columns, 12472 rows, after data quality check: 16 columns, 10101 rows.
2. Dropped host name due to Data Privacy and licence
3. No duplicates in the Dataset
4. Deleted some missing values in last_review
5. Changed price column to string

5. Data Types

COLUMN	DATA TYPE	DESCRIPTION	Time Variant/ Invariant	Data type
id	int64	Airbnb's unique identifier for the listing	Invariant	Qualitative
name	object	Name of the listing	Invariant	Qualitative
host_id	int64	Airbnb's unique	Invariant	Qualitative

		identifier for the host/ user		
neighbourhood_group	object	Neighbourhood Group of listing	Invariant	Qualitative
neighbourhood	object	Neighbourhood Of listing	Invariant	Qualitative
latitude	float64	Latitude of listing	Invariant	Qualitative
longitude	float64	Longitude of listing	Invariant	Qualitative
room_type	object	Type of the room	Invariant	Qualitative
price	object	Price per night		Qualitative
minimum_nights	int64	Minimum nights required to book	Invariant	Quantitative
number_of_reviews	int64	The nr of reviews the listing has	Variant	Quantitative
last_review	object	Date of last review	Invariant	Quantitative
reviews_per_month	float64	The nr of reviews the listing has over the lifetime of the listing	Variant	Quantitative
calculated_host_listings_count	int64	The nr of listings the host has in the curren t scrape, in the city/re gion geography	Variant	Quantitative
availability_365	int64	The availability of the listing x days in the fut ure as determined by the calendar. Note a li sting may not be avail able because it has be en booked by a guest or blocked by the host	Variant	Quantitative
number_of_reviews_ltm	int64	The number of review s the listing has (in the last 12 months)	Variant	Quantitative

6. Limitations and Ethics

- **Limitations:**

There 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.

7. Questions to explore

- What room types are available in Airbnb Berlin?
- What is the avg. price range for these room types?
- Which districts in Berlin are most popular?
- Which districts have higher Airbnb prices?
- What Influences price and demand of Airbnb in Berlin?

8. Why did I choose this dataset?

I chose this dataset because I live in Germany and I have chosen sometimes AirBnB as accommodation. By analyzing the dataset I want to implement my previous tourism experience with the data analysis one.