01 Sourcing Open Data
-Ray Rusli Junior-

A. Summary

Airbnb Munich Case Study

- Data Source: It is an internal data from Airbnb and owned by them. Airbnb made the information available on the Airbnb site. It is an administrative data that contains information concerning rental rooms in Munich as published on the Airbnb website.

  It is licensed under Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication" license and is therefore free to use.

- Data Collection: This dataset was published by Airbnb, covering all the listing in the Munich with summary information and metrics.

- Data Contents: This data contains information about all the listing at Airbnb in Berlin. The data contains 4995 observations of rooms for rent in Berlin, including details on room description, location, prices, rental periods and reviews as of December 12, 2021.

- Resources: http://insideairbnb.com/get-the-data.html , accessed on 17.03.2022. Latest datasets were released on 17.12.2021

- Why I chosen this data: I spent plenty of hours and days searching for an open dataset that fits my project. The most challenging part was searching for the dataset that provides the project's requirements and my interest. I have chosen this set since I feel like there are many possibilities to play around with, and it seems to be relatively straightforward, and it comes from a trustworthy source.

B. Data Profile

Data cleaning and consistency check:

The original data set contains 18 columns and 4995 rows, since I am just on the first stage and still don't have a clear picture of what I am going to do with the analysis, On the first stag ed I decided to remove 4 Columns.

Renaming columns:

No columns have been renamed, since the name has been clear enough to understand

Dropping columns:

I dropped 4 column which has missing value and irrelevant for my analysis, these are: 'neighb ourhood group', 'last_review', review_per_month' and 'license'

Missing values:

After deleting the 4 columns which have a missing value, the remaining columns do not have a missing value, except "name" and "host_name" - which is also irrelevant for my analysis, however I inputted the value with "missing".

Duplicates values:

There is no duplicate value in our data set.

Basic descriptive statistics

Rows: 4.995
Columns: 14
Total record counts: 4.995

Descriptive statistics were calculated for every column in the Jupyter notebook.

	Min	Max	Mean	Count	Frequency Table
price	0	9999	127,85	4995	see Jupyter notebook
minimum_nights	1	1000	9	4995	see Jupyter notebook
number_of_reviews	0	765	22,34	4995	see Jupyter notebook
Calculated_host_listings_count	1	33	2,56	4995	see Jupyter notebook
Availability_365	0	365	135,47	4995	see Jupyter notebook
Number_of_reviews_ltm	0	450	3,48	4995	see Jupyter notebook

After adjusting some of the row price due to irregularity.

	Min	Max	Mean	Count	Frequency Table
price	0	8000	118,45	4995	see Jupyter notebook

Derived Columns

Data set	New column	column/s it was derived from	conditions
munich_clean2	price_category	price	low: < 80
			medium: >= 80 & < 200
			high: >= 200
munich_clean2	rental_term	availability_365	short term: <= 90
			medium term: > 90 & <= 180
			long term: > 180
munich_clean2	host_type	calculated_host_listing_count	private host: <= 1
			commercial host: > 1

Limitation and ethical consideration

Regarding limitations, By the time the data was released (December 2021), Germany was on partial lockdown due to the COVID 19 pandemic. At the beginning of 2021, there was a lockdown for over 4 months. It is also to note that travel has been started to pick up again by the end of 2021. The listing situation on Airbnb only presents what is on their data by the time it was released.

This data is only of Airbnb properties, and Airbnb is just one provider out of many such services.

For these reasons, this data set couldn't be used to extrapolate results to other cities or the entire private property rental market in Munich. This would constitute a sample or exclusion bias.

C. Define Questions

- What findings could we get from the Airbnb Data from the city of Munich?
- Is there an equal amount of Airbnb listings in all 25 neighborhoods in Munich?
- What is e the average price of an Airbnb listing in Munich? If a place is available for a longer time, could the price be significantly higher?