Airbnb Berlin Case Study: Project outline

Business case

Who
The client is a legal company which supports tenants in defending their right to a safe and affordable home.

Why
The client wants to see proof that short-term commercial rentals of apartments drive rental prices up and reduce the number of available apartments for permanent Berlin residents.

What
The storyboard will show the impact of commercial Airbnb hosts in Berlin on price and rental availability of apartments in Berlin.

When
It will be used when arguing with legislators for stricter regulations governing commercial short-term rentals with the aim to reduce and stabilize rental prices and availability for Berlin residents.

Where
Tableau Public.

Data Source

http://insideairbnb.com/get-the-data.html
Accessed on 16 September 2021.
Latest update of data set was on 12 July 2021.

The data is internal to Airbnb, therefore owned by them, but sourced from publicly available information from the Airbnb site. As such, it is as precise as it gets when it comes to data about Airbnb.

It is administrative data in the sense that it contains a directory of of information concerning rental rooms in Berlin 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.

The data contains 19095 observations of rooms for rent in Berlin, including
details on room description, location, prices, rental periods and reviews as of July 12, 2021.

Data profile

Columns: 16
Rows: 19095

Categorical variables

ID:	property ID
name:	property name
host_id:	host ID
host_name:	host name
room_type:	Entire place / private room / shared room

Location variables

neighbourhood_group:	City district the neighbourhood pertains to
neighbourhood:	Neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
latitude:	The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
longitude:	The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.

Quantitative variables

price:	Daily price in local currency
minimum_nights:	Minimum number of nights stay for the listing
number_of_reviews:	The number of reviews the listing has
last_review:	The date of the last / newest review
reviews_per_month:	The number of reviews the listing has over the lifetime of the listing
calculated_host_listings_count:	The number of listings the host has in the current scrape, in the city/region geography.
availability_365:	Availability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.

<u>Wrangling steps</u>

Columns dropped	Columns renamed	Columns's type changed	Comments
		ID from int64 to str	
		Host_id from int64 to str	
Id			Not needed
Name			Not needed
Host_id			Not needed
Host_name			Not needed
Last_review			Not needed
Neighbourhood			Not needed
	Neighbourhood_group		To match json

Consistency checks & cleaning

Dataset	Missing values	Missing values treatment	Dups	Duplicates treatment	Mixed type columns	Mixed type columns treatment	Outliers	Outliers treatment
Listing_ wrangled_ cleaned	Name (30)	No change				Replaced NaN with “missing”, changed type to str		
Listing_ wrangled_ cleaned	Host_name (12)	No change				Replaced NaN with “missing”, changed type to str		
Listing_ wrangled_ cleaned	Last_review (4155)	No change				Replaced NaN with “0”, changed type to int64		
Listing_ derived columns	Reviews_per_month (4155)	Imputed mean						
Listing_ wrangled_ cleaned							Price has 7 x value of 0	Replaced with mean, 73,30
Listing_ wrangled_ cleaned							Minimum_nights has 13 x over 365	Left them as they are.
Listing_ wrangled_ cleaned							Calculated_host_listings_count has 583 x over 20	Left them as they are.

Listing_ derived columns							Price has 3 x value of 8000	Imputed with mean.

Summary statistics

Before cleaning

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	19095.0	19095.0	19095.0	19095.0	19095.0	14940.0	19095.0	19095.0
mean	52.510215129313379	13.404654204095136	73.303221	9.105943964388583	21.63707776904949	0.7182737617135306	3.135847080387536	91.271694160777507
std	0.03239084494645433	0.06295250252312785	136.249622	33.63595600181032	48.67042696900742	1.4452721285482029	7.773246348000803	127.64533005331572
min	52.34007	13.09715	0.0	1.0	0.0	0.01	1.0	0.0
25%	52.48971	13.36716	35.0	2.0	1.0	0.09	1.0	0.0
50%	52.50995	13.41409	52.0	3.0	4.0	0.27	1.0	0.0
75%	52.53332	13.4389	81.0	5.0	17.0	0.83	2.0	175.0
max	52.65611	13.75737	8000.0	1124.0	620.0	94.35	76.0	365.0

After cleaning

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	19095.0	19095.0	19095.0	19095.0	19095.0	14940.0	19095.0	19095.0
mean	52.510215129313379	13.404654204095136	73.330092827712243	9.105943964388583	21.63707776904949	0.7182737617135306	3.135847080387536	91.271694160777507
std	0.03239084494645433	0.06295250252312785	136.24238966411178	33.63595600181032	48.67042696900742	1.4452721285482029	7.773246348000803	127.64533005331572
min	52.34007	13.09715	8.0	1.0	0.0	0.01	1.0	0.0
25%	52.48971	13.36716	35.0	2.0	1.0	0.09	1.0	0.0
50%	52.50995	13.41409	52.0	3.0	4.0	0.27	1.0	0.0
75%	52.53332	13.4389	81.0	5.0	17.0	0.83	2.0	175.0
max	52.65611	13.75737	8000.0	1124.0	620.0	94.35	76.0	365.0

Derived columns

Data set	New column	Column/s it was derived from	Conditions
listing_derivedcolumns	Price category	price	< 80 then “Low price”
			>= 80 and < 300 then “Middle price”
			>= 300 then “High price”
listing_derivedcolumns	Rental availability	availability_365	<= 90 then “Short term”
			> 90 and <= 180 then “Middle term”
			> 180 then “Long term”
listing_derivedcolumns	**host _type**		> 180 then “Long term”
			> 180 then “Long term”

Limitations

In terms of limitations, it’s important to note that the travel industry has slumped because of Covid-19 and that as a result many hosts might not have updated their Airbnb listing for a while.

On the other hand, it might show the start of a “post-Covid-19” private rentals scenario, the beginning of a new reality for the travel industry.

Only time will tell if the period analysed in this project will constitute a stumbling stone or the beginning of a new era.

Further, it’s important to remember that this data is only of Airbnb properties and Airbnb is just one provider out of many of 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 Berlin, as this would constitute a sample or exclusion bias.

Recommendations

- Conduct the same analysis with data from other short-term rental providers such as Wunderflats and 9flats,

- Compare, aggregate trends and aim to define the most precise number of misappropriated apartments in Berlin and what impact they have on the long-term rental market.