New York Airbnb listings

Data source:

The data used for this project is an open-source data set collected from Kaggle ([Dataset](#))

The reason I chose this dataset:

I choose this project because I always have being interested in the real estate business and technology and Airbnb is a company that merge this 2 topics perfectly, making it interesting to understand how the listings work and how customers and hosts behaviors affect the business.

Data cleaning:

541 duplicate values were removed.

2 misspelling fixed in "neighborhood group"

5 values homogenized in "neighborhood."

131 values were added to blank spaces as US in the "country code."

Thirteen values homogenized in "minimum nights."

36 values that were over 365 days were changed to missing in "minimum nights."

12 values that were after 3/4/2022 were changed to missing due to data being collected on 3/3/2022

4311 values homogenized in "availability 365."

"House rules" and "license" were removed for the analysis

Limitations and ethics:

- Most of the entries were entered by hand by the hosts, this give a big window for errors that can include; mistype, non-homogenized entries, duplicates, etc.
- The data is just focus in New York which in a whole business analysis would represent just a small part of the complete business operations
- Some entries may look like they are duplicated but they are actually listings that were in the app once, took down and the they were put up again with minimum variation
- The data was collected on 3/3/2022 which means that all the listings that were placed or removed after this time wouldn't be contemplated in this analysis

Data profiling:

Variable	Time variant / invariant	Structured / unstructured	Qualitative / quantitative
id	invariant	structured	qualitative
NAME	invariant	unstructured	qualitative
host id	invariant	structured	qualitative
host_identity_verified	invariant	structured	qualitative
host name	invariant	structured	qualitative
neighbourhood group	invariant	unstructured	qualitative
neighbourhood	invariant	unstructured	qualitative
lat	invariant	structured	quantitative
long	invariant	structured	quantitative
country	invariant	structured	qualitative
country code	invariant	structured	qualitative
instant_bookable	invariant	structured	qualitative
cancellation_policy	invariant	structured	qualitative
room type	invariant	structured	qualitative
Construction year	invariant	structured	quantitative
price	variant	structured	quantitative
service fee	variant	structured	quantitative
minimum nights	variant	structured	quantitative
number of reviews	variant	structured	quantitative
last review	variant	structured	quantitative
reviews per month	variant	structured	quantitative
review rate number	variant	structured	quantitative
calculated host listings count	variant	structured	quantitative
availability 365	variant	structured	quantitative

Questions to explore:

- Does the price of the rental affect the service fee charged in the properties?
- Does the construction year matters when listing the price of the properties?
- Does the price of the rental affect the average rating in the properties?
- Is there any relationship between the listing price and the neighborhood group?
- Is there any neighborhood that are more popular than the rest? Does this affect it price?