

Airbnb Data Exploration

Dataset Source:

<https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata>

Goals:

With this exploration, I sought to better understand the New York City rental property market; specifically, which neighbourhoods have the most listings, the impact of keywords on price and availability, the distribution of room types and cancellation policies, and other impacts on price.

Query Results and Observations:



1. Data Cleaning:

Upon downloading the dataset, I observed cleaning scope before conducting any operations. This began with normalizing the naming convention for all the columns, as discrepancies included variance in capitalization, snake_case, and unclear names.

Next, I removed the host_id and license columns, given that the former yielded little purpose in my exploration—given that the id was serving as the primary key, and the latter was a fully blank column.

Finally, I fixed values to reduce confusion; this began with setting the blank country and country code values to be the United States and NYC, given that all properties were located in New York. I also observed some misspelled borough names when viewing the dataset, so thought it would be worth changing.

2. Which neighbourhood has the highest number of listings?

	neighbourhood character varying 	count bigint 
1	Bedford-Stuyvesant	7937

3. What is the percent distribution of different room types?

Given the vast number of properties in New York City, I thought that it would be interesting to understand the percent distribution of different room types.

	room_type character varying 🔒	room_info bigint 🔒	room_percent numeric 🔒
1	Entire home/apt	161103	52.3406660883634343
2	Private room	139668	45.3766605912338327
3	Shared room	6678	2.1696117895885925
4	Hotel room	348	0.11306153081414048870

4. What is the spread of cancellation types in different boroughs?

From the query, it was interesting to see how the strict cancellation policy, something that is commonly construed as the most restrictive, and hence the least desirable, is not the least common in Brooklyn and in Queens.

	cancellation_policy character varying 🔒	borough character varying 🔒	cancellation_info bigint 🔒
1	flexible	Manhattan	14723
2	moderate	Manhattan	14607
3	strict	Manhattan	14440
4	moderate	Brooklyn	14039
5	strict	Brooklyn	13976
6	flexible	Brooklyn	13786
7	strict	Queens	4475
8	moderate	Queens	4449
9	flexible	Queens	4335
10	flexible	Bronx	914
11	moderate	Bronx	901
12	strict	Bronx	896
13	moderate	Staten Island	338
14	flexible	Staten Island	308
15	strict	Staten Island	308
16	[null]	Bronx	1
17	[null]	Brooklyn	42
18	[null]	Manhattan	23
19	[null]	Queens	8
20	[null]	Staten Island	1

5. Is there a relationship between the construction year and the price?

My curiosity with this query stemmed from a conversation that I conducted with my mother regarding housing, in that years have a unique value in housing. For instance, in technology, as devices get older they lose value and any sense of desirability given their outdated functionality; the newer, by broad-stroker, the better. However, this is not the case with housing, as older houses have a certain mystique, intrigue, and comforting feel to them. They also can be more appropriate given the type of company and occasion that will be arranged in the Airbnb. On the other side of the spectrum, newer properties also have considerable value given their renovated and modern features.

The findings reflected and cemented the above contentions, demonstrating that the average price and the average price deviation by construction year remain relatively constant. For such a broad property market, it is surprising to see that New York does not have much price variance: whether one is going for a “rustic” or “new” vibe.

	construction_year integer	average_price numeric	standard_deviation numeric
1	2003	623.6022705030338618	333.21
2	2004	630.0935695799323114	324.38
3	2005	620.1932461448370096	333.64
4	2006	635.2245054734011907	332.80
5	2007	626.0231554160125589	332.75
6	2008	638.0030674846625767	332.99
7	2009	618.7993408297789841	331.36
8	2010	626.7378546443839876	331.41
9	2011	626.8072885719944543	331.64
10	2012	623.4436413361984763	328.66
11	2013	619.3118129122526484	329.65
12	2014	630.5583173996175908	332.52
13	2015	617.2270759543486816	335.40
14	2016	624.5972000000000000	336.48
15	2017	629.1030662710187933	324.84
16	2018	624.4419572107765452	334.56
17	2019	612.6132639290534027	327.92
18	2020	621.3109815354713314	331.99
19	2021	629.3143027650686294	335.46
20	2022	629.2305137722211369	330.91
21	[null]	599.7428571428571429	333.07

6. How do certain keywords impact property prices?

For this query, I wanted to better recognize the effects of certain words on the demand for a property, which is channelled through the price (especially in a competitive market such as New York. I consulted the top five words from a list from a study conducted by Istan Egresi. By comparing the data obtained by their studies with this dataset, I was of the perception that I could gain a more comprehensive view.

Word	Frequency	% Shown	No. of cases	% cases
location	350	9.17	342	34.20
apartment	334	8.75	245	24.50
clean	334	8.75	331	33.10
center	245	6.42	235	23.50
room	206	5.40	174	17.40

My findings while different in metric, opposed the findings of the study. Not only is the order of the words ranked by prices different, but it showed how the price of the listings with the inclusion of these keywords was generally slightly lower than the average price.

	word text	average_price numeric
1	center	653.2486772486772487
2	all properties	625.2935360325152415
3	apartment	624.4091236927515326
4	room	623.7413436139073320
5	location	622.1820191599115696
6	clean	612.4285714285714286