Data Source

Data

I am using 4 data sets in total, representing Airbnb listings information. These data sets consist of *“listings summary”* and *“reviews detailed”.* There are 2 versions of each data set, one scraped from June 2023 and one from March 2024. The purpose of this is to be able to compare and contrast Airbnb activity across a 9 month period.

Data Source

My data was externally sourced from Inside Airbnb (http://insideairbnb.com/), a site whose mission is to provide “data and advocacy about Airbnb's impact on residential communities” through the collection and publication of Airbnb listing information.

Data Collection

This data has been scraped from the official Airbnb website which provides full information regarding listing details.

Data Contents

- **Listings:** These data sets contain details regarding the Airbnb properties and hosts available in Dublin. Such information includes: *property id, property name, host id, host name, neighbourhood, property latitude & longitude, property type, property price, minimum stay, number of reviews,* and *host total listings.*
- **Reviews:** These data sets contain details regarding the reviews left on Airbnb properties in Dublin. Such information includes: *listing id, date of review, reviewer id, name of reviewer,* and *comments.*

Statistical analysis - listings

```
[85]: listings_23.describe()
```

[85]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	calculated_host_listings_count
count	8.440000e+03	8.440000e+03	8440.000000	8440.000000	8440.000000	8440.000000	8440.000000	8440.000000
mean	2.733412e+17	1.395853e+08	53.345449	-6.255111	181.120498	6.184716	31.159005	5.366588
std	3.738527e+17	1.507540e+08	0.046388	0.062639	562.541185	29.167667	68.372446	12.511049
min	4.407700e+04	4.398400e+04	53.207663	-6.497588	9.000000	1.000000	0.000000	1.000000
25%	1.933922e+07	2.754379e+07	53.327961	-6.278417	70.000000	1.000000	1.000000	1.000000
50%	3.598778e+07	7.335573e+07	53.344160	-6.259215	115.000000	2.000000	6.000000	1.000000
75%	6.970116e+17	2.029741e+08	53.357097	-6.233293	200.000000	4.000000	27.000000	3.000000
max	9.101728e+17	5.191105e+08	53.623490	-6.052910	46513.000000	1125.000000	1218.000000	76.000000

```
[86]: listings_24.describe()
```

[86]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	calculated_host_listings_count
count	9.020000e+03	9.020000e+03	9020.000000	9020.000000	9020.000000	9020.000000	9020.000000	9020.000000
mean	3.534469e+17	1.541454e+08	53.345827	-6.255699	168.425831	6.250000	32.381596	6.209091
std	4.264895e+17	1.646165e+08	0.046197	0.063016	553.539762	30.776392	72.620441	14.600766
min	1.484200e+04	4.398400e+04	53.207663	-6.498946	15.000000	1.000000	0.000000	1.000000
25%	2.038340e+07	2.959183e+07	53.327808	-6.278165	110.000000	1.000000	1.000000	1.000000
50%	4.158785e+07	8.055937e+07	53.344154	-6.258841	124.000000	2.000000	6.000000	1.000000
75%	8.388324e+17	2.264616e+08	53.357780	-6.233503	140.000000	4.000000	28.000000	3.000000
max	1.045709e+18	5.504933e+08	53.623240	-6.051490	45703.000000	1125.000000	1466.000000	84.000000

Data Profile

Data Cleaning

- Listings 2023
    - Dropped columns: host_name, neighbourhood_group, last_review, reviews_per_month, availability_365, number_of_reviews_ltm, licence.
- Listings 2024
    - Dropped columns: host_name, neighbourhood_group, last_review, reviews_per_month, availability_365, number_of_reviews_ltm, licence.
    - Missing values: ‘Price’ column contained 3793 missing values. These missing values were imputed using the column’s median value.
- Reviews 2023
    - Dropped columns: reviewer_name.
    - Missing values: ‘Comments’ column contained 26 missing values. These rows were removed as a comment is required for any sentiment or textual analysis.
- Reviews 2024
    - Dropped columns: reviewer_name.
    - Missing values: ‘Comments’ column contained 30 missing values. These rows were removed as a comment is required for any sentiment or textual analysis.

Data Limitations and Ethics

There are some potential limitations to these data sets, such as:

- Data completeness: some listings may not be included due to scraping limitations caused by changes in website structure or restrictions imposed by Airbnb's terms of service.
- Temporal impact: The dataset may not reflect the most recent version of Airbnb listings due to the dynamic nature of the platform. Listings may be added, removed, or updated frequently which could not be represented in the event that the data is not regularly updated.
- Market fluctuation: The datasets may not capture market dynamics, such as seasonal variations in pricing or changes in supply and demand.

There are minimal ethical concerns within these datasets, primarily as all information has been publicly posted on Airbnb’s website by users who consent to their names and other information being publicised. Despite this, I still chose to remove host and reviewer names from the dataset to further secure their privacy.

Questions to Explore

1. What properties get the most reviews?
2. What vocabulary occurs most frequently in positive/negative reviews?
3. How have host profiles changed over the last 9 months? I.E do hosts, on average, own more properties in 2024 than in 2023? Are there more hosts?
4. Do any properties/hosts appear to be in conflict with Airbnb’s policies? I.E are there any potentially illegal listings?
5. Has the average price per night changed?

6. How does pricing vary with location?
7. How does pricing vary with property-type?