Data Source

Summary**:**

Airbnb Listings: Data Sourcing

This dataset is sourced from **Inside Airbnb**, an independent open-data initiative that collects and publishes publicly available Airbnb listing information from major cities worldwide. The data is external to this project and is derived from Airbnb's public platform. Inside Airbnb is widely recognized within the data science community as a trustworthy and transparent source for short-term rental market research.

Because the dataset is compiled using publicly available Airbnb listing information, it can be considered a reliable and authoritative source for analyzing short-term rental markets, host behavior, and neighborhood-level pricing trends.

Airbnb Listings: Data Collection

The data is collected automatically by scraping publicly available Airbnb listing pages at regular intervals. It represents **usage data**, capturing listing-level information such as property characteristics, host attributes, availability, and pricing. The dataset reflects real-world platform activity and may include a time lag depending on the scrape cycle, but overall provides a comprehensive snapshot of current market conditions.

Airbnb Listings: Data Contents

The dataset contains detailed information for thousands of Airbnb listings within a major U.S. city; in this case, Columbus, Ohio. Each row represents an individual property listing.

Key variables include:

Geographic Variables

- Latitude
- Longitude
- Neighbourhood
- City
- State

Continuous Variables

- Price
- Minimum nights
- Number of reviews
- Reviews per month
- Availability (365 days)

Categorical Variables

- Room type
- Property type
- Host is superhost
- Instant bookable
- Neighbourhood group

These variables enable analysis of spatial pricing patterns, availability trends, host behavior, and neighborhood market differences.

Airbnb Listings: Limitations

- Listings are self-reported by hosts and may contain reporting bias.
- Some properties may be inactive or seasonally unavailable.
- Pricing may vary dynamically, meaning the dataset represents a snapshot rather than continuous real-time changes.
- Short-term rental data may reflect market conditions influenced by regulation, tourism, and local housing policies.

Airbnb Listings: Data Relevance

The dataset provides strong geographic and pricing information suitable for analyzing urban rental market behavior. It enables geospatial analysis, regression modeling, clustering, and time-series exploration of pricing and availability trends. These characteristics make the dataset highly relevant for investigating how neighborhood location, property features, and host characteristics influence Airbnb pricing and demand.

Why This Dataset Was Chosen

This dataset was selected because it provides rich geographic detail, a large volume of observations, and a diverse mix of continuous and categorical variables that are well-suited for advanced exploratory and predictive analysis. The Airbnb listings data allows for meaningful investigation into how location, property characteristics, and host behavior influence pricing, availability, and demand within urban rental markets.

The dataset includes precise latitude and longitude coordinates, enabling geospatial analysis and neighborhood-level mapping. It also contains multiple continuous variables such as price, availability, and review activity, which support regression modeling and time-series analysis, along with categorical variables such as room type and host attributes that support segmentation and clustering.

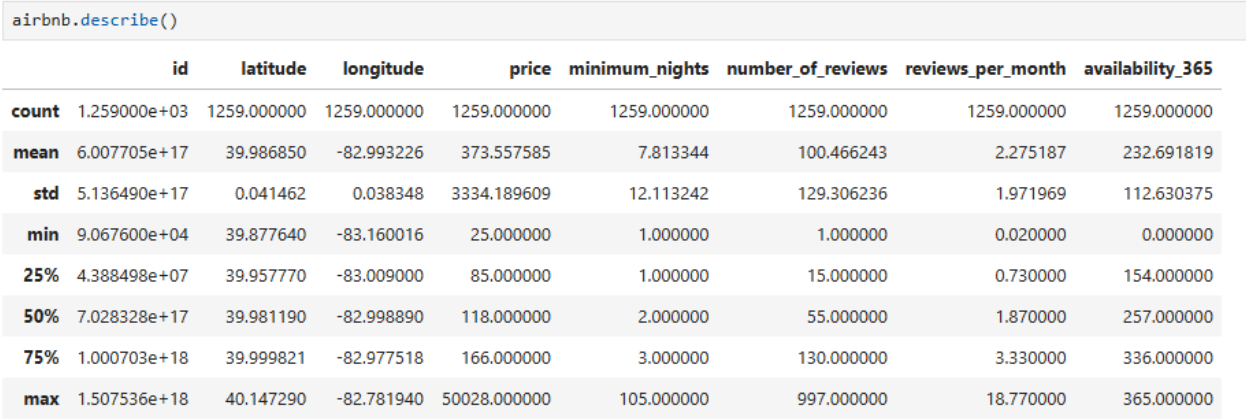
In addition, short-term rental markets represent a real-world business problem with direct economic and housing implications, making the findings both analytically valuable and practically relevant. The scale and structure of the data allow for the application of machine learning techniques while producing insights that mirror real industry use cases in pricing strategy, market segmentation, and demand forecasting.

Data Profile

Data Understanding & Descriptive Statistics:

Data Cleaning Summary

The dataset was cleaned by converting pricing values to numeric format, subsetting to key analytical variables, and removing rows with missing values across geographic, pricing, and availability fields. This resulted in a clean master dataset of **1,259 listings across 14 variables**, suitable for regression modeling, clustering, and geospatial visualization.



```
airbnb.describe()
```

	id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	availability_365
count	1.259000e+03	1259.000000	1259.000000	1259.000000	1259.000000	1259.000000	1259.000000	1259.000000
mean	6.007705e+17	39.986850	-82.993226	373.557585	7.813344	100.466243	2.275187	232.691819
std	5.136490e+17	0.041462	0.038348	3334.189609	12.113242	129.306236	1.971969	112.630375
min	9.067600e+04	39.877640	-83.160016	25.000000	1.000000	1.000000	0.020000	0.000000
25%	4.388498e+07	39.957770	-83.009000	85.000000	1.000000	15.000000	0.730000	154.000000
50%	7.028328e+17	39.981190	-82.998890	118.000000	2.000000	55.000000	1.870000	257.000000
75%	1.000703e+18	39.999821	-82.977518	166.000000	3.000000	130.000000	3.330000	336.000000
max	1.507536e+18	40.147290	-82.781940	50028.000000	105.000000	997.000000	18.770000	365.000000

Limitations & Ethical Considerations

The Airbnb listings dataset represents self-reported information published by hosts on the Airbnb platform and therefore may contain reporting bias, inaccuracies, or outdated information. Because pricing and availability are dynamic, the dataset reflects a snapshot in time and may not fully capture seasonal fluctuations or real-time market changes.

Additionally, short-term rental data may be influenced by local housing regulations, tourism patterns, and enforcement differences across neighborhoods, which may limit the generalizability of findings beyond the specific city studied.

From an ethical perspective, short-term rental markets can impact housing affordability and long-term rental availability in urban areas. While the dataset contains no personally identifiable information, analyses should be interpreted with awareness of their potential social and economic implications on local communities and housing supply.

Questions to Explore

Clarifying Questions (What is happening?)

1. Which neighborhoods have the highest concentration of Airbnb listings?
2. What is the average nightly price of Airbnb listings across different neighborhoods?
3. What property types and room types are most commonly listed?
4. How many nights per year are most listings available?

Adjoining Questions (How does this compare / what influences it?)

5. How do average listing prices differ between neighborhoods?
6. How does listing availability relate to pricing?
7. Are superhosts more likely to have higher prices and more reviews?
8. Do instant-bookable listings differ in price and availability compared to non-instant listings?

Funneling Questions (Drilling into causes and drivers)

9. Which features (location, room type, property type, reviews, availability) most strongly influence price?
10. Do listings with more reviews tend to command higher nightly prices?
11. Are entire homes priced significantly higher than private rooms or shared rooms?
12. Does minimum stay length affect nightly price?

Elevating Questions (Big-picture & impact)

13. Which neighborhoods show signs of being saturated with short-term rentals?
14. How might Airbnb pricing and availability patterns reflect tourism demand across the city?
15. Could high concentrations of short-term rentals be associated with reduced long-term housing availability in certain neighborhoods?

Privacy & Ethical Considerations

16. Does this dataset contain personally identifiable information about hosts or guests?
17. Could findings be misused to promote short-term rental expansion in areas with housing shortages?
18. How should pricing recommendations be responsibly framed to avoid contributing to housing affordability issues?