CityBnB: Project Overview

Dataset Used: Airbnb Open Data – San Francisco, Kaggle

Project Type: IDIS (Information and Data Insight Systems)

Background

San Francisco—home of the Golden Gate Bridge, innovation, and culture—has always attracted travelers from around the world. In 2008, amidst the financial crisis, Airbnb emerged with a bold promise: let everyday people rent out rooms and help others experience cities like locals. It seemed revolutionary—affordable stays, extra income for residents, and a more connected world.

But over time, the model shifted. The spirit of the sharing economy gave way to a pursuit of profit. What started as couch-surfing and spare bedrooms became a landscape dominated by commercial operators. The friendly face of Airbnb was replaced by a complex network of investors running short-term rentals at scale.

San Francisco, already strained by limited housing and rising rents, became a prime target. Today, data reveals that nearly 70% of Airbnb listings in the city are operated by multi-listing hosts—essentially unlicensed hotel chains operating under the pretense of home-sharing. Worse still, ghost listings—properties that don't exist—are being used to falsely inflate demand and drive up prices. Behind the scenes, shell companies are buying up homes and listing them en masse.

This transformation has real consequences. Families are priced out. Communities become fragmented. Long-term housing becomes scarce. The platform that once connected people now threatens the fabric of local neighborhoods.

CityBnB, a data consultancy rooted in social impact, is joining the effort to restore balance—not through politics, but through data.

Using SQL, we aim to:

- Map Airbnb listing density across neighborhoods to reveal saturation patterns.
- Analyze host behavior to uncover those running unlicensed hotel chains.
- Track pricing trends and compare them to traditional rental markets.

- Dig into reviews to surface issues like noise, over-occupancy, and community disruption.

This isn't just analysis—it's an investigation. Our goal is to help reshape regulation, support fair hosting, and protect San Francisco's identity as a place to live—not just visit.

Problem Statement

You are a team of Data Analysts at **CityBnB**, a local consultancy that partners with government agencies, real estate stakeholders, and tourism boards to examine short-term rental trends—particularly the rise and impact of Airbnb.

San Francisco faces a housing crisis worsened by the unchecked growth of short-term rentals. What started as a way for residents to earn extra income has evolved into a complex and sometimes exploitative system. Through data analysis, your team is tasked with revealing how Airbnb is reshaping the city's housing landscape.

Your mission:

Use SQL to analyze Airbnb data and uncover the truth behind the listings.

Key Questions:

- Who are the hosts operating in the city?
- Where are short-term rentals most heavily concentrated?
- Are prices rising beyond market norms?
- What do reviews reveal about guest experiences and community impact?

This investigation will highlight patterns in host behavior, identify pricing and availability trends, and offer insights into how Airbnb may be affecting long-term residents.

The Method and Data

Context

This project uses the **San Francisco Airbnb Open Data**. Originally curated for the Airbnb Inside initiative and later expanded by Arian Azmoudeh, the dataset is designed to simulate real-world analysis challenges, including inconsistencies ideal for data cleaning practice.

The dataset includes three main components:

- **Listings** – Property details, host info, geolocation, and review scores
- **Reviews** – Full comment texts from guests
- **Calendar Data** – Daily prices and availability for each listing

Key Fields

- id: Unique listing ID
- host_id: Unique host ID
- host_name: First name of the host
- host_identity_verified: Whether the host is verified
- neighbourhood: Specific district within San Francisco
- latitude / longitude: Geolocation coordinates
- price, availability_365: Indicators of listing behavior
- review_scores_rating, comments: Performance and sentiment data

This mix of structured and unstructured data allows for comprehensive analysis using SQL techniques like joins, filters, aggregations, and window functions.

Project Phases

Phase 1: Mapping San Francisco's Airbnb Terrain

- Use SQL to identify verified vs. anonymous hosts
- Detect clusters of listings by host
- Analyse Airbnb density by neighbourhood to see where the shift is most dramatic
- Study review text to reveal guest sentiment

Phase 2: Exposing the Operators

- Use window functions to uncover hosts with lots of properties
- Compare occupancy to vacancy to flag ghost listings (Nonexistent, Duplicated and Inactive listings)
- Analyse price spikes — are Airbnb rates inflating the long-term market?

Phase 3: Accountability through Integration

- Join Airbnb listings with city tax records to trace hidden landlords
- Audit top property owners for legal violations or exceeding rental limits
- Flag listings managed by shell companies (businesses that exist only on paper, they have no real operations)

Phase 4: Recreating Regulation

- Use insights to draft smarter policies: enforce limits, redistribute Airbnb revenue, and support neighborhood renewal
- Propose real-time dashboards so communities can monitor Airbnb activity themselves

---

Hypothesis

Airbnb activity in San Francisco has transitioned away from traditional home-sharing. We hypothesize that approximately **70% of hosts manage multiple listings**, and many are exploiting regulatory loopholes. These operators may be contributing to **housing scarcity, community disruption, and rising rental prices**.

Goals

CityBnB's mission is to **restore balance** to the city through ethical, data-driven analysis.

- **For communities**: Empower leaders to enforce fair housing policies
- **For local government**: Inform policy using evidence, not speculation
- **For businesses**: Support ethical hosting and improve guest experience
- **For analysts**: Transform raw data into insight that impacts lives

Predictions and Exploration Areas

To understand the true scope of Airbnb's impact on San Francisco, our analysis will explore five major focus areas. Each area is driven by a central goal: to uncover patterns that affect housing availability, affordability, and community dynamics.

---

1. Neighborhood Performance

Goal: Assess how Airbnb listings are distributed across San Francisco and understand how different neighborhoods are being affected in terms of saturation, tourism pressure, and rental diversity.

Exploration Areas:

- How are Airbnb listings distributed across neighborhoods, and which areas are experiencing the highest concentration?
- In which parts of the city are guest reviews most frequent, and what does this suggest about traveler behavior or listing quality?
- Are premium-priced listings concentrated in traditionally touristic areas, and how does that correlate with neighborhood gentrification?
- Do certain room types (entire homes vs. shared rooms) dominate specific neighborhoods, and what might that mean for local housing dynamics?

---

2. Pricing Patterns

Goal: Investigate pricing strategies and trends to determine how short-term rental pricing compares to long-term housing markets, and whether certain pricing behaviors reflect profit-driven or community-focused intentions.

Exploration Areas:

- What are the average nightly prices across different neighborhoods, and which areas have the widest price variations?
- How does the type of accommodation (entire home, private room, shared room) influence average price levels?
- Are certain neighborhoods disproportionately expensive when adjusted for review scores and availability?
- How do booking lengths (short vs. extended stays) affect pricing strategies, and what does this suggest about target demographics?

---

3. Host Activity

Goal: Understand the structure of Airbnb hosting in the city, with a focus on identifying dominant hosts, commercial behavior, and possible misuse of the platform.

Exploration Areas:

- Who are the most active or high-volume hosts, and how many properties are they managing across the city?
- Is there a difference in guest satisfaction between verified and unverified hosts, and what role does verification play in host reputation?
- Are commercial hosts (with multiple listings) shaping or distorting the market, and to what extent is the platform used for full-scale rental businesses?

---

4. Availability Trends

Goal: Explore how frequently listings are available and whether there are patterns suggesting ghost listings, underutilization, or commercial-style full-time rentals.

Exploration Areas:

- Which listings are consistently available throughout the year, and what does that suggest about whether they're residential or commercial in nature?
- How does availability differ between neighborhoods, and what patterns emerge across room types and host identities?
- Is there a relationship between year-round availability and other indicators such as high pricing or high review frequency?

---

5. Review Insights

Goal: Analyze guest feedback to identify satisfaction levels, recurring complaints, and neighborhood-specific concerns that may affect local communities.

Exploration Areas:

- Which listings attract the most guest interaction, and what makes them stand out in terms of service, pricing, or location?
- How does the frequency and recency of reviews correlate with rating scores—are more active listings consistently better rated?
- Are there shifts in review sentiment over time, and do they point to increasing concerns around issues like cleanliness, noise, safety, or host reliability?