BUILDING AIRBNB ELT DATA PIPELINES WITH AIRFLOW AND DBT CLOUD

airbnb

Big Data Engineering | Spring 2024

By Huy Duc Vu

Table of Contents

Project Overview

The objective of this project aims to design and implement a production-graded ELT data pipeline for Airbnb and NSW Census data using Apache Airflow, Google Cloud Platform (GCP), and dbt Cloud. The pipeline adheres to the Medallion Architecture, comprising Bronze, Silver, and Gold layers, which progressively clean, transform, and model the data for analytical purposes. This architecture ensures scalability, data integrity, and transparency across all stages of data transformation.

The core stage is to extract raw Airbnb listings data for Sydney between May 2020 and April 2021, and combine it with demographic data from the Australian Bureau of Statistics at the Local Government Area (LGA) level. The final data warehouse enables business understanding, such as revenue performance by region, host characteristics across LGAs, or affordability comparison between Airbnb revenue and local mortgage repayments.

The project involves some key stages:

1. Apache Airflow orchestrated the ELT processes, automating ingestion from Google Cloud Storage (GCS) into a PostgreSQL database hosted on GCP.

2. dbt Cloud was used to manage data transformations, schema design, and warehouse Medallion architecture design.

3. Deployed a fully functional ELT pipeline with scalability, transforming raw datasets into analytical data marts, which support SQL analyses and provide actionable insights into Airbnb performance.

Data Source Analysis

This project extracts raw data from multiple sources, including 3 key sources:

- Airbnb listings information.
- Census data.
- Index lookup table between Local Government Area (LGA).

1. Content and Structure

- Airbnb Listings (CSV format)
    - Contains monthly snapshots of Airbnb property listings for Sydney between May 2020 and April 2021, with each file representing a month of data.
    - The data includes rich listing details such as listing ID, host ID, superhosts, listing or host neighbourhood, property type, room type, accommodates, prices or availability, with various other metrics.
- Census data (CSV format)
    - 2 census tables were extracted from the ABS 2016 General Community Profile, providing demographic information on LGA levels across NSW regions.
    - G01 (Selected Persons Characteristics by Sex): This table contains demographic distributions such as age, age group breakdowns, languages and gender composition.
    - G02 (Selected Medians and Averages): This table includes socioeconomic indicators such as average household size, median age, median mortgage repayments, and median weekly income.
- Index tables (CSV format)
    - This data includes 2 index tables bridging the Airbnb neighbourhoods with their corresponding LGA details (LGA name, LGA code, suburb name)
    - NSW_LGA_CODE: This file contains 2 key attributes, including LGA name and LGA code, establishing a unique code for each LGA.

- NSW_LGA_SUBURB: This file provides a many-to-one relationship between suburbs and LGAs, enabling spatial aggregation of Airbnb listings from the suburb level to the LGA level.
- It is worth noting that the host_neighbourhood field in Airbnb listings is recorded on the suburb-level, while the listing_neighbourhood attribute is recorded as LGA name.

2. Data Format

All datasets were provided as CSV files, uploaded to the Google Cloud Storage (GCS) bucket for ingestion by Airflow.

- Airbnb Listings:
  - 12 monthly CSVs, each containing 22 columns and thousands of rows.
  - Column headers were initially uppercase, requiring normalisation to lowercase before insertion into PostgreSQL.
- LGA–Suburb and LGA–Code index files:
  - Simple key–value mappings in flat CSV format.
- Census tables:
  - Structured CSVs that were converted into JSONB payloads during ingestion.
  - This approach preserved the original column metadata while simplifying schema evolution — each LGA's Census record is stored as a single row with all indicators accessible as JSON keys.

Within the PostgreSQL warehouse, the raw data was ingested into Bronze tables using Airflow's PostgresHook, while dbt handled transformations and casting into the Silver layer, ensuring type consistency (NUMERIC, BOOLEAN, and DATE) across datasets.

3. Data Quality

Several data quality considerations were addressed throughout the project:

- Column Inconsistencies: The Airbnb dataset contained varying column cases and naming patterns across months. These were standardised to lowercase using a preprocessing step in the Airflow DAG.
- Missing and Null Values: Multiple fields, such as review_scores_rating and price, occasionally contained nulls. Imputations were not applied directly; instead, downstream transformations excluded these records when calculating averages to avoid biasing metrics.
- Duplicate Records: Duplicates were prevented by implementing a `DELETE WHERE source_file = ...` operation before each load, ensuring idempotent ingestion in Airflow.
- Temporal Completeness: The Airbnb data covered a complete 12-month period (May 2020 – April 2021) without gaps, ensuring reliable time-based aggregations.

Through these steps, the project ensured high data reliability, consistency across schemas, and accurate integration between geographically and thematically diverse sources.

Overall Architecture/ Pipelines

This project implements a modern ELT pipeline on GCP with Airflow orchestration, PostgreSQL for storage, and dbt for transformation and modelling. The design follows the Medallion architecture (Bronze → Silver → Gold), adds SCD 2 snapshots for slowly-changing dimensions, and finishes with datamarts as views for most of the analysis. The result is a reproducible, auditable pipeline that turns raw CSVs into credible business-ready tables for data analysis. For this project, the initial load batch only ingests a single month of Airbnb listings to design the warehouse architecture with small-sized data before scaling to the full dataset, saving resources and time while building the pipeline.

1. High-level flow

- Storage (PostgreSQL)
    - Monthly Airbnb CSVs and static reference files (Census, LGA mappings) are uploaded to GCS under predefined prefixes (`data/airbnb/`, `data/census/`).
- Ingestion (Airflow to Postgres / Bronze)
    - Airflow DAG discovers monthly files, sorted by `MM_YYYY` format, then loads them sequentially into `bronze.*` tables.
    - The ingestion adds columns recording the load timestamp and the source file for traceability.
    - Idempotent loads: use `DELETE WHERE source_file = …` before insert, then archive the processed data file to `…/archive/` folder in each data folder, ensuring no duplication ingestion happens.
- Complete ELT pipeline (dbt → Bronze → Silver → Gold)
    - dbt implements a complete Medallion architecture, starting with initiating the sources to load raw data into Bronze tables.
    - dbt normalises types, fixes naming, and applies business logic to create cleaned, conformed Silver tables.
    - dbt snapshots maintain history for dimensions (SCD-2) in a dedicated `snapshots` schema.
    - dbt builds a Gold star schema (fact and SCD dims as tables) and datamarts as views for reporting business questions.

- Analysis
    - Utilise SQL and datamarts to conduct business analyses.

2. Environment and Schemas (dbt)

- Schemas
    - `bronze`: raw ingestion tables from Airflow.
    - `silver`: cleaned, typed, conformed models.
    - `gold`: star schema tables (`gold.star`) and views (`gold.mart`).
    - `snapshots`: dbt snapshot tables with `dbt_valid_from` and `dbt_valid_to`.
- dbt project configuration
    - `models.bronze` → `+schema: bronze`, materialised table.
    - `models/silver` → `+schema: silver`, materialised table.
    - `models/gold/star` → `+schema: gold`, materialised table.
    - `models/gold/mart` → `+schema: gold`, materialised view.
    - `snapshots/` → `+schema: snapshots` (timestamp strategy).

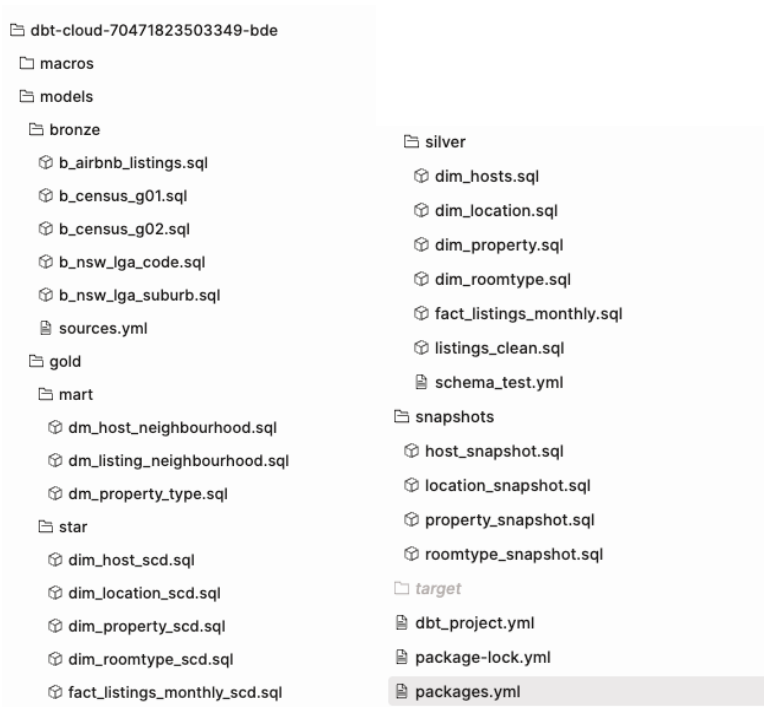

Figure 1. Medallion architecture in dbt Cloud.

3. Airflow orchestration (Bronze layer)

- DAG:
  - Discovery: lists `gs://<bucket>/data/airbnb/` files matching `.*/(\d{2})_(\d{4})\.csv$`, sorts chronologically.
  - Clean & align: lowercases headers, casts dates, fills missing expected columns with NULL.
  - Insert: `execute_values` bulk insert into `bronze.airbnb_listings_raw`.
  - Idempotency: `DELETE FROM bronze.airbnb_listings_raw WHERE source_file = <object>.`
  - Archiving: copies the processed file to `data/airbnb/archive/`, `data/census/archive/` and deletes the original.
  - Static loads: separate tasks/functions to load LGA code, LGA–suburb, and to ingest Census (G01 and G02) as one JSONB record per LGA to keep schema-drift friendly.
- Scheduling: `schedule_interval=None` (set to manual for this project), and strict concurrency = 1 to maintain order.
- Config via variables: bucket, prefixes (no dbt API needed per brief).

4. dbt load (Bronze layer)

- The Bronze layer in dbt mirrors the raw tables ingested by Airflow into PostgreSQL.
- Each raw dataset is registered as a dbt source and materialised as a pass-through table (`b_airbnb_listings`, `b_census_g01`, `b_census_g02`, `b_nsw_lga_code`, `b_nsw_lga_suburb`).
- No transformations are applied at this stage as the purpose is to establish a stable and documented interface for downstream layers.
- Each Bronze table retains ingestion metadata (`source_file`, `loaded_at`) to maintain full lineage and allow idempotent re-runs.

5. dbt transformations (Silver layer)

- Goals
    - Normalise names/cases, cast types (`NUMERIC`, `BOOLEAN`, `DATE`).
    - Produce a single, dependable `listings_clean` table (with `month_start` for the time grain, derive `month_start` from `scraped_date)`.
    - Build simple conformed dimensions at the "current state" level:
        - `dim_host` (host attributes, including `host_is_superhost`).
        - dim_location (maps `host_neighbourhood` and suburbs to LGA via `NSW_LGA_SUBURB` and `NSW_LGA_CODE`).
        - dim_property (`property_type`).
        - dim_roomtype (`room_type`).
- Key conventions
    - Consistent lower-snake-case field names.
    - Derived metrics ready for Gold layer:
        - `stays_in_month = 30 - availability_30`,
        - `est_revenue = price * stays` (for active listings only).
- Testing
    - Not-null / unique tests on keys (such as `listing_id`, `host_id`, `lga_code`).

6. dbt snapshots (SCD-2 history)

- Strategy: `strategy='timestamp'`, `updated_at='snapshot_ts'`.
- Scope: dimensions that may change over time
    - `host_snapshot` (capturing superhost status/history).
    - `location_snapshot` (host neighbourhood to LGA mapping history).
    - `property_snapshot` and `roomtype_snapshot`.
- Purpose: Gold fact joins require "as-of" dimension states at `month_start`, so historical reports reflect correct attributes.

7. dbt models (Gold layer)

Star schema (tables)

- Dimensions (SCD-2)
  - `dim_host_scd`, `dim_location_scd`, `dim_property_scd`, `dim_roomtype_scd`.
  - Each has a surrogate key (such as `host_sk = (host_id, dbt_valid_from)`, and `dbt_valid_from` / `dbt_valid_to`).
- Fact
  - `fact_listings_monthly_scd` at grain (listing_id x month_start).
  - Contains only IDs and metrics:
    - SKs to each SCD dimension
    - Metrics or flags needed downstream: `price`, `has_availability`, `availability_30`, `stays_in_month`, `est_revenue`, `review_scores_rating`.
  - "As-of" joins ensure the fact row links to the correct historical dim version.

Datamarts (views)

- `dm_listing_neighbourhood`: monthly KPIs by neighbourhood: active rate, superhost rate, price stats, MoM deltas, revenue per active listing.
- `dm_property_type`: slice by property_type x room_type x accommodates with the same KPIs.
- `dm_host_neighbourhood`: host-centric view aggregated to LGA using the suburb–LGA mapping, includes revenue per active listing and per host.

8. Run order

Run the dbt files in the following order:

- `dbt run --select silver.*`
- `dbt snapshot`
- `dbt run --select gold.star.* --full-refresh`
- `dbt run --select gold.mart.*`

Business Analysis

1. Demographic differences between top and bottom LGAs

This analysis compares the top three and bottom 3 performing LGAs in terms of average estimated revenue per active Airbnb listing over the last 12 months (May 2020 to April 2021). Using the aggregated results from `dm_host_neighbourhood` joined with Census demographics, the analysis highlights how population characteristics and household structures differ between high- and low-performing areas.

performance_group	lga_name	avg_revenue_per_active	median_age	pct_age_25_34	avg_household_size
Top 3	BATHURST REGIONAL	12,238.81	37	5,171	2.5
Top 3	NORTHERN BEACHES	9,400.85	40	29,575	2.7
Top 3	WOOLLAHRA	7,692.39	39	9,432	2.3
Bottom 3	PARRAMATTA	1,313.78	34	45,775	2.8
Bottom 3	CUMBERLAND	589.67	32	42,160	3.2
Bottom 3	MID-WESTERN REGION	228	42	2,677	2.4

Table 1. Demographic differences between top and bottom LGAs.

- The top-performing LGAs, including Bathurst Regional, Northern Beaches, and Woollahra, have a slightly older median age (37–40 years) and smaller household sizes (2.3–2.7). These areas also feature relatively lower proportions of the 25 to 34-year age group, indicating more established, higher-income residents who may operate or manage premium Airbnb listings.
- The lower-performing LGAs (Parramatta, Cumberland, and Mid-Western Regional) show a younger median age (32 to 34 years old) and larger households (2.8 to 3.2 persons). Parramatta and Cumberland, in particular, have very high counts of residents aged 25–34, which are consistent with denser, more affordable housing and higher rental competition rather than short-stay tourism demand.
- This suggests that Airbnb performance is positively associated with mature, smaller household populations typical of suburban or coastal areas catering to higher spending visitors, whereas the innerwest or outer-metro LGAs with younger, family-dense populations see lower yields per active listing.

→ Overall: High-revenue LGAs correspond to more mature, smaller households, typically wealthier, low-density suburbs where listings are more premium and occupancy rates are higher. In contrast, younger, family-dense LGAs have lower Airbnb performance, likely due to lower nightly rates and less tourist demand.

2. Correlation between median age and revenue per active listing

This analysis explores whether there is a relationship between the median age of residents in a neighbourhood (based on Census data) and the average estimated revenue per active Airbnb listing. By joining Airbnb neighbourhood performance with Census G02 data at the corresponding LGA level, the Pearson correlation (for linear association) and the regression slope (indicating the change in revenue with each additional year of median age) are measured.

n_neighbourhoods	pearson_corr	slope_rev_per_year	intercept
21	0.8237	679.8613	-20,399.77

Table 2. Correlation metrics between median age and revenue per active listing.

listing_neighbourhood	lga_code	median_age	avg_rev_active
Hunters Hill	14100	43	10,940.1184615385
Mosman	15350	42	9,311.7146153846
Woollahra	18500	39	6,807.7423076923
Waverley	18050	35	5,742.1207692308
Lane Cove	14700	36	4,612.4438461538
Willoughby	18250	37	4,527.2184615385
Randwick	16550	34	4,301.7323076923
North Sydney	15950	37	4,211.5484615385
Sydney	17200	32	3,925.7538461538
Canada Bay	11520	36	3,064.3169230769
Ryde	16700	36	2,761.0661538462
Hornsby	14000	40	2,614.7915384615
Penrith	16350	34	2,376.9330769231
Campbelltown	11500	34	2,197.6466666667
Liverpool	14900	33	2,182.9246153846
Parramatta	16260	34	1,834.1676923077
Burwood	11300	33	1,696.5646153846
Camden	11450	33	1,640.1646153846
Strathfield	11300	33	1,517.2723076923
Blacktown	10750	33	1,263.9430769231
Fairfield	12380	32	1,251.1207692308

Table 3. Median age and average revenue per active listing by listing neighbourhood.

- The Pearson correlation of 0.8237 shows a strong positive relationship between the median age of a neighbourhood and its Airbnb revenue performance. This means that as the median age increases, the average estimated revenue per active listing also tends to rise.
- The regression slope (around $680 per additional year of median age) suggests that, on average, listings in areas where residents are one year older in median age generate $680 more annual revenue per active listing.

- Neighbourhoods like Hunters Hill, with a median age of 43, Mosman of 42, and Woollahra of 39, show the highest Airbnb revenues (between $6,800 and $10,900 per active listing), indicating that affluent, older, and more established suburbs are more profitable.
- In contrast, younger and denser areas such as Blacktown (median age of 32), Fairfield (32), or Parramatta (33) record lower revenues (around $1,200 to $2,000 per active listing). These LGAs typically have lower property prices, higher rental occupancy, and less short-stay tourism demand.

→ Overall: The analysis reveals a clear socioeconomic gradient as older, high-income LGAs outperform younger, family-dense ones in Airbnb profitability. Airbnb listings in mature, wealthier suburbs generate substantially higher revenues, indicating that host performance aligns with broader demographic affluence.

3. Best type of listing for the top 5 neighbourhoods

This analysis identifies the optimal property type, including property type, room type, and number of accommodates, that achieves the highest average number of stays for each of the top 5 performing neighbourhoods by revenue per active listing. The analysis focuses on the last 12 months to determine which combination yields both high occupancy and strong revenue performance.

A-Z listing_neighbourhood	A-Z property_type	A-Z room_type	123 accommodates	123 avg_stays_active	123 avg_revenue_active
NORTHERN BEACHES	room in boutique hotel	hotel room	2	30	51,105
WAVERLEY	house	entire home/apt	14	30	45,000
MOSMAN	villa	entire home/apt	12	30	43,492.5
WOOLLAHRA	house	private room	6	30	36,000
HUNTERS HILL	townhouse	entire home/apt	4	30	13,500

Table 4. Top 5 neighbourhoods with the highest average number of stays.

- All five neighbourhoods achieve the maximum average stays (30 stays per active listing), indicating very high occupancy rates across premium Sydney suburbs.
- The property type and room type combination differ slightly by location:
    - Northern Beaches leads with boutique hotel-style listings, appealing to short-term tourists seeking coastal accommodation.

- Waverley and Mosman dominate with entire homes and villas, accommodating larger groups (12 to 14 people), which are ideal for family or group holidays.
- Woollahra stands out with private rooms in houses, suggesting strong demand for smaller, high-value stays in high-income residential areas.
- Hunters Hill has lower revenue despite full occupancy, reflecting smaller property sizes (4 accommodates), possibly lower nightly rates, and regional location.

- Across all top-performing suburbs, "Entire house/apartment" listings consistently rank among the highest-revenue categories, emphasising the value of offering full-space rentals in premium markets.

→ Overall: The most profitable and frequently booked listings are entire houses or villas in high-income, coastal neighbourhoods, accommodating 4 to 14 guests (wide range). These property types capture both longer stays and premium pricing, whereas smaller or shared-space listings, even with full occupancy, generate lower overall revenue.

4. Distribution of superhosts across LGAs

This analysis investigates whether Airbnb superhosts with multiple listings tend to concentrate their properties within the same Local Government Area (LGA) or diversify across multiple LGAs. The query examined all hosts with at least two distinct listings and calculated the number of unique LGAs (n_lgas) where their properties are located.

123 superhosts_single_lga	123 superhosts_multiple_lga
3,910	1,185

Table 5. Count of superhosts within single and multiple LGAs.

123 host_id	123 n_listings	123 n_lgas
194,230,296	496	24
175,128,252	356	23
279,001,183	210	19
235,137,306	188	19
189,112,177	108	16
113,180,379	31	14
301,753,450	67	13
7,409,213	206	12
15,739,069	124	12
36,410,227	108	11
137,278,159	33	11
224,249,958	24	11
251,905,190	21	11
41,215,678	65	10
102,664,004	43	10
185,783,910	31	10
178,729,132	23	10
208,408,190	17	10
216,159,830	64	9
270,132,187	52	9
357,566,424	38	9
329,529,761	24	9
11,914,644	75	8
293,526,132	46	8
382,207,272	30	8
94,377,709	30	8
38,478,183	22	8
267,733,940	22	8
4,298,915	21	8
10,441,624	18	8
1,106,924	14	8
227,947,343	12	8
44,869,453	11	8
288,743,418	105	7
101,139,031	61	7
283,812,868	44	7
103,069,716	36	7
189,309,820	34	7
302,271,656	31	7
21,058,208	31	7
198,996,712	24	7
169,795,871	20	7
292,913	18	7
246,661	16	7
346,620,688	15	7
153,195,376	12	7
16,357,713	144	6
2,450,066	120	6
293,274,101	65	6
226,913,216	52	6

Table 6. Count of superhosts with number of listings and number of LGAs spread.

- The majority of approximately 77% of superhosts manage all their listings within a single LGA, showing a high concentration of operations. This pattern suggests that most supershots specialise in a single market, which is likely their local area, with listing as their private property, where they can efficiently manage cleaning, maintenance, and guest turnover.
- The remaining 23% of hosts operate across two or more LGAs, with some large-scale operators managing listings in over 20 LGAs (particularly the top host with 496 listings across 24 LGAs). These hosts likely represent professional property management companies or agencies that oversee multiple short-term rentals under a central operation.
- The average host with multiple LGAs manages between 7 and 10 distinct regions, indicating a strategy of geographical diversification aimed at spreading occupancy risk and capturing demand from different travel hubs, such as between coastal and urban areas.

→ Overall: The Airbnb Sydney market is still dominated by localised hosting with their private property, although a growing number of professionalised hosts are expanding their reach across regions.

5. Mortgage repayments coverage by single listing hosts

This analysis examines whether Airbnb hosts with only one active listing earn enough estimated annual revenue to cover the median annual mortgage repayment for their corresponding LGA. The goal is to understand how sustainable single-listing hosting is relative to local housing costs, and which LGAs offer the strongest financial performance for small size hosts.

lga_name	total_hosts	hosts_covering_mortgage	pct_hosts_covering	avg_host_revenue	avg_annual_mortgage
NORTHERN BEACHES	4,327	2,561	59.19	75,932	33,600
MOSMAN	364	198	54.4	88,833	36,000
SUTHERLAND SHIRE	512	258	50.39	46,754	29,124
WAVERLEY	4,309	2,096	48.64	55,763	36,000
SYDNEY	5,854	2,803	47.88	39,180	29,988
RANDWICK	2,578	1,223	47.44	45,796	31,200
NORTH SYDNEY	1,047	482	46.04	45,323	31,200
HUNTERS HILL	68	30	44.12	50,236	36,396
INNER WEST	2,000	839	41.95	40,284	31,200
WOOLLAHRA	1,395	563	40.36	57,633	38,400
WILLOUGHBY	404	152	37.62	56,126	34,524
LANE COVE	260	97	37.31	49,865	31,200
CANADA BAY	389	145	37.28	33,197	30,000
RYDE	492	152	30.89	22,633	26,400
CUMBERLAND	404	124	30.69	43,787	24,000
CAMDEN	40	11	27.5	19,955	26,640
HORNSBY	364	100	27.47	28,868	28,800
BAYSIDE	1,179	323	27.4	23,589	28,800
THE HILLS SHIRE	249	67	26.91	29,900	30,000
STRATHFIELD	153	37	24.18	15,790	26,004
CANTERBURY-BANKST	473	113	23.89	17,801	24,000
GEORGES RIVER	295	68	23.05	20,028	26,004
PARRAMATTA	430	99	23.02	17,863	26,004
LIVERPOOL	106	24	22.64	22,268	25,476
PENRITH	120	27	22.5	17,285	24,000
BURWOOD	170	38	22.35	16,744	26,400
CAMPBELLTOWN	62	12	19.35	12,681	22,104
FAIRFIELD	47	9	19.15	12,351	21,600
BLACKTOWN	229	31	13.54	15,395	25,800

Table 7. Mortgage repayments coverage by single listing hosts.

- The Northern Beaches leads all regions, with 59.2% of non-superhosts generating enough annual Airbnb income to fully cover their median annual mortgage repayments. This high coverage rate reflects a combination of strong tourism demand, premium nightly rates, and relatively moderate mortgage costs compared to central Sydney.
- Mosman follows closely, with 54% of hosts covering mortgages, supported by the city's luxury rental market and high visitor appeal. Despite higher mortgages ($36,000), average Airbnb revenues in Mosman ($88,000) are the highest in Sydney.
- Areas such as Waverley, Sydney, and Sutherland Shire also demonstrate nearly 50% mortgage coverage, indicating strong host profitability potential for single-listing operators.

- In contrast, inner western LGAs such as Blacktown (13.5%), Campbelltown (19%), and Fairfield (18%) show limited profitability, reflecting both lower Airbnb demand and lower nightly rates, even though their mortgages are relatively cheaper.

→ Overall: Airbnb revenue can offset a significant portion of housing costs in premium, high-demand suburbs. For the majority of Sydney LGAs, about 40% to 60% of single-listing hosts can fully cover their median annual mortgage repayments, with the Northern Beaches standing out as the most financially advantageous location for hosts.

Potential Issues

- Data Alignment: Airbnb neighbourhoods and Census LGAs used different spatial identifiers, requiring careful mapping via the suburb–LGA index. Minor mismatches or missing suburbs could lead to incomplete joins.
- JSON Parsing in Census Data: Census G01 and G02 tables were stored as JSONB payloads. Extracting specific indicators relied on exact key names, which could break if the ABS schema changes.
- Null Airbnb fields: Some listings lacked price, rating, host neighbourhood, or availability data. Excluding them from calculations may slightly understate total market activity.
- Assumption in revenue estimation: Revenue metrics were derived from estimated formulas (stays x price), assuming consistent occupancy and pricing, which may not fully represent real earnings.

Conclusion

This project successfully built a production-ready ELT pipeline that integrates Airbnb listings and Census data for Sydney using Airflow, dbt Cloud, and PostgreSQL. Following the Medallion architecture, the workflow enabled efficient ingestion, transformation, and modelling of raw data into analytical data marts for business insights.

The analysis showed that Airbnb performance is closely tied to demographic and geographic factors. High-performing LGAs such as Northern Beaches, Mosman, and Woollahra have older populations, smaller households, and higher listing revenues. Most superhosts operate within a single LGA, while a smaller share manage large portfolios across multiple regions. For single-listing hosts, Airbnb income in premium suburbs can cover or exceed annual mortgage repayments, highlighting the profitability of these areas.

Overall, the project demonstrates how modern data engineering tools can transform raw data into actionable insights. The implemented pipeline is scalable, transparent, and provides a strong foundation for future extensions or deeper market analysis.