Big Data Engineering

Assignment 3: Building ELT data pipelines with Airflow

Table of Contents

1. Project Overview

This project's main objective is to build ELT pipelines for processing Airbnb and Australian Census data for Sydney using Airflow, PostgreSQL, and dbt Cloud. Therefore, the task is to design a scalable Medallion Architecture (Bronze -> Silver -> Gold) that transforms raw data into an analytical data mart, enabling insights into Airbnb performance across suburbs and local government areas (LGAs).

Objectives -

- Develop data pipelines that extract and load raw Airbnb, Census and LGA mapping datasets into a PostgreSQL data warehouse using Airflow.
- Design and implement a Medallion data architecture:
  - Bronze Layer: Store raw ingested data from Airflow.
  - Silver Layer: Apply data cleansing, standardisation, and Slowly Changing Dimensions (SCD Type 2) for temporal accuracy.
  - Gold Layer: Build star-schema models with fact and dimension tables to support analytical queries.
- Create data marts (e.g., dm_listing_neighbourhood, dm_property_type, dm_host_neighbourhood) to answer business questions around Airbnb performance, pricing trends, and host behaviour.
- Answer business questions combining Airbnb and Census datasets to derive actionable insights on demographic and economic patterns across Sydney's LGAs.

2. About the Datasets

This project integrates two primary datasets — Airbnb and Australian Census — along with a reference LGA–Suburb mapping file to enable spatial and demographic analysis across Sydney.

2.1 Airbnb Dataset

The Airbnb dataset contains information about property listings in Sydney from May 2020 to April 2021.
Each monthly file includes details such as:

- Listing attributes: property type, room type, accommodates, price, and availability.
- Host details: host ID, host name, host neighbourhood, and superhost status.
- Performance metrics: number of reviews, review scores, and estimated revenue.

This dataset helps analyse how listing characteristics and host behaviour vary across suburbs and time, supporting insights on rental trends, occupancy rates, and revenue performance.

2.2 Census Datasets (G01 & G02)

The Australian Bureau of Statistics (ABS) 2016 Census data provides demographic and socio-economic indicators at the Local Government Area (LGA) level.
Two tables from the General Community Profile Pack were used:

- G01 – Selected Person Characteristics by Sex: includes population counts, age distributions, and family structures.
- G02 – Selected Medians and Averages: includes median age, household size, income, rent, and mortgage repayment values.

These datasets will help join Airbnb performance data with demographic context, enabling analyses such as the relationship between revenue and median age or mortgage affordability by LGA.

2.3 LGA–Suburb Mapping

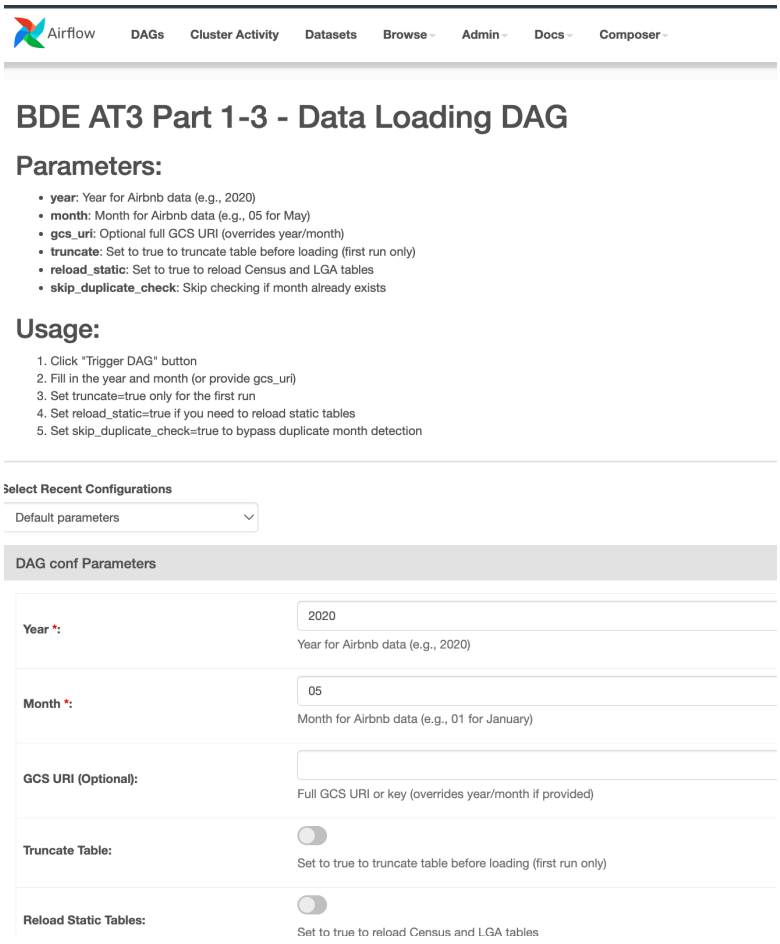
To map Airbnb listings with Census data, two reference mapping files were used to set the relationship between Local Government Areas (LGAs) and suburbs:

1. LGA Code–Name Mapping - This file contains the official LGA codes and corresponding LGA names.
It is the master reference for identifying and joining LGAs using a unique five-digit code.

2. LGA–Suburb Mapping - This file lists each LGA name alongside the suburbs within its LGA.It matches Airbnb listings or host neighbourhoods (captured as suburb names) to their corresponding LGAs, enabling the aggregation of Airbnb metrics at the LGA level.

3. Project Setup and Data Loading

The first step in this project involved setting up the data ingestion pipeline using Apache Airflow within Google Cloud Composer to load the raw datasets into a PostgreSQL instance hosted on Google Cloud SQL.



Dag Trigger Menu

3.1 Loading the Datasets

The first data to load was the May 2020 dataset, which was uploaded to the Airflow storage bucket along with the Census and LGA mapping files.

An Airflow DAG is used to load the datasets and has the following parameters to make it easier to load the datasets

Main Parameters

- year (int, required)

Calendar year of the Airbnb batch to load (e.g., 2020).

- month (int, required)

Calendar month of the Airbnb batch to load (1–12). For the initial loading, this was 5 (May).

- truncate (bool, default: false)

When true, raw (Bronze) target tables are truncated before loading. It's used only for the first month. For future loading, this was set to false.

- load_static (bool, default: false)

When true, loads Census G01, Census G02, LGA code–name, and LGA–suburb mapping datasets. This is set to true for the first month, and then disabled for further loading.

Upon clicking the trigger button on Airflow, the config menu will appear, where the user can input these parameters.

3.2 Data Modelling and Transformation (dbt Cloud)

Once the raw dataset is loaded into the bronze schema of the database using Airflow, the next stage involved transforming and modelling the data using dbt Cloud. The dbt (Data Build Tool) was used to structure the data warehouse following the Medallion Architecture, ensuring clean, consistent, and analysis-ready datasets.

3.2.1 Medallion Architecture

Bronze Layer

- This layer represents the raw ingestion layer, containing data loaded from Airflow.
- All original Airbnb, Census (G01 & G02), and LGA mapping tables are stored here without any transformations.
- This layer serves as the data source for all downstream transformations.

Silver Layer

This is where the transformations, cleaning and snapshots are executed. For each raw table loaded into the Bronze schema — Airbnb listings, hosts, Census (G01 & G02), and LGA mapping files — corresponding Silver models were created in dbt.

Data Cleaning and Standardisation

- Filtering rows with inconsistent scraped dates relative to the source month - Here, rows where the scraped_date did not align with the dataset's source month were dropped, ensuring each monthly dataset accurately represented the correct reporting period.

- Handling missing values—To maintain data completeness, Missing values in any category feature, such as host neighbourhood, listing neighbourhood, property type, etc., were marked as Unknown if null.

- Data type casting - Data types were cast for int, numeric,big int, timestamp and dates as appropriate for the columns.

- Cleaning LGA–Suburb Mapping - The lga_suburb table contained several extra columns and inconsistent text formats. These were removed or renamed, retaining only relevant columns (lga_name, lga_suburb) with clean, standardised naming.

- Cleaning Census Tables (G01 & G02) - In the Census datasets, LGA codes were prefixed with the text “LGA_” (e.g., LGA_16200). This prefix was removed in the Silver layer, and all codes were padded to a uniform five-digit format. This allowed it to join with other tables using the cleaned lga_code field and ensured compatibility with the reference LGA Code–Name mapping.

Snapshot Preparation (Feeds Folder Under Silver)

A dedicated feeds folder was created in dbt to act as a feeder for monthly snapshots and track historical changes in listings(property) and hosts.

- **silver_host_for_snapshot** view: It gives a monthly feed of host records by selecting the latest record per host for that month. This ensures that host attributes such as host_is_superhost, host_neighbourhood, and host_since are accurately captured at the correct time interval, serving as the input for the host_snapshot table.

- **silver_property_for_snapshot** view: This view extracts the latest record per listing/property for each month, preserving attributes such as room_type, accommodates, and price. This view feeds into the property_snapshot table, ensuring accurate monthly versioning of property attributes.

Snapshot Tables

The snapshot tables host_snapshot and property_snapshot were created using dbt’s timestamp strategy based on the scraped_date field.

- host_snapshot tracks changes in host-related attributes such as host_is_superhost, host_neighbourhood, and host_since, allowing historical tracking of a host’s profile over multiple months using host_id
- property_snapshot tracks changes in property-level details such as property_type room_type and accommodates using listing_id (since its listing level)

Gold Layer

This layer ensures all business entities are transformed into fact and dimension tables, ensuring historical accuracy through Slowly Changing Dimension (SCD Type 2) logic derived from the snapshot models.

Star Schema Design

The gold layer follows a star schema design, which separates descriptive data (dimensions) from numerical measures (facts).

Dimension Tables

- **dim_host** - Derived from the host_snapshot table using SCD Type 2 logic. It has host attributes such as host_name, host_neighbourhood, host_is_superhost, and host_since, along with validity date ranges (record_start_date, record_end_date). Each record represents a time-valid version of the host’s profile.

• **dim_property** - Derived from the property_snapshot table using the same timestamp-based SCD2 approach. It has property-specific details such as room_type, property_type, and accommodates. This allows historical tracking of property attribute changes over time.

• **dim_suburb**—This table contains a listing's listing neighbourhood information and its key, suburb_id(derived from the listing neighbourhood).

• **dim_date** - A date table which contains date_id and the scraped dates of listings.

These dimensions were designed following the standard of who, what, when, and where concepts.

Fact Table - fact_listings

The central fact table of the star schema stores monthly listing-level metrics, including price, availability_30, and number_of_reviews. Each record is linked to the relevant dimensions (dim_host, dim_property, dim_suburb, dim_date, and ref_lga) via keys (host_id, listing_id, date_id, and suburb_id), allowing accurate analysis across time, geography, and property attributes.

Reference Tables - ref_lga, ref_census_g01 and ref_census_g02
They link the Airbnb dimensions and facts to the LGAs and. Census data.

Data Mart Views

Data marts were created on top of the gold layer to support analytical insights into Airbnb data.

dm_listing_neighbourhood
Aggregates metrics by listing_neighbourhood and month/year, including:
- Active listings rate
- Minimum, maximum, median, and average price
- Distinct host count
- Superhost rate
- Average review score
- Month-to-month percentage change (active and inactive listings)
- Total number of stays and estimated revenue per active listing

dm_property_type

Summarises performance metrics by property_type, room_type, and accommodates, providing insights into which property configurations yield the highest occupancy and revenue. The same metrics that are computed for dm_listing_neighbourhood are done here, too.

dm_host_neighbourhood

Aggregates metrics at the host LGA level (derived by mapping host_neighbourhood to its corresponding LGA).
Includes:
- Number of distinct hosts per LGA
- Estimated revenue per active listing
- Estimated revenue per host

All marts use SCD joins to ensure that every record reflects dimension attributes as valid at the time, maintaining full temporal accuracy.

4. Ad-hoc Analysis

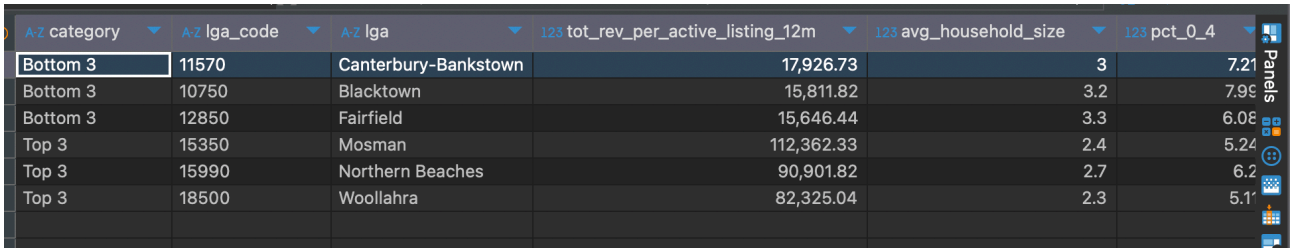
1) Demographic Differences Between Top and Bottom-Performing LGAs

The LGAs after ranking based on total estimated revenue per active listing across the 12 months found that, in terms of average household size, the bottom-performing LGAs, such as Canterbury-Bankstown, Blacktown and Fairfield, have much higher household size (3-3.3) compared to the top-performing LGAs, such as Mosman, Northern Beaches and Woollahra (2.3-2.7).

The age distribution also shows that bottom-performing LGAs have mostly a younger population, which we can see by comparing the percentage of age groups 15-19, 20-24,25-34 is larger than that of top-performing LGAs.

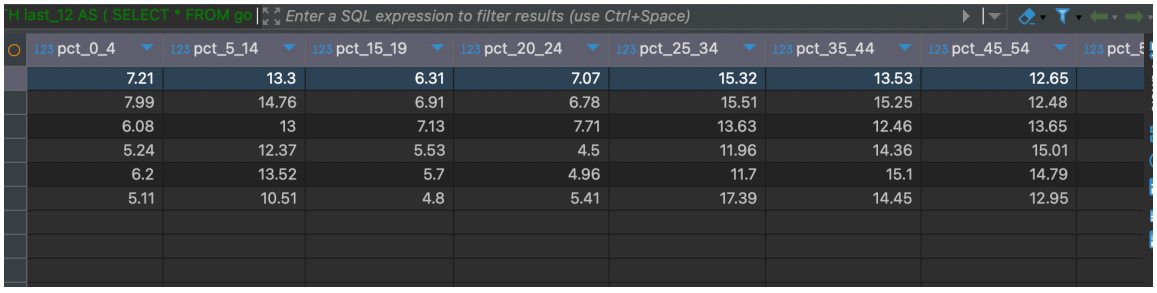
The top-performing LGAs are a bit more established financially in terms of family, whereas the bottom-performing LGAs represent young families with children and youth.

The analysis suggests that Airbnb revenue correlates with demographic and socio-economic characteristics — LGAs with smaller households, older populations, and higher affluence (e.g., Mosman, Woollahra) outperform more densely populated, family-centric areas such as Canterbury-Bankstown and Blacktown.



A-Z category	A-Z lga_code	A-Z lga	123 tot_rev_per_active_listing_12m	123 avg_household_size	123 pct_0_4
Bottom 3	11570	Canterbury-Bankstown	17,926.73	3	7.21
Bottom 3	10750	Blacktown	15,811.82	3.2	7.99
Bottom 3	12850	Fairfield	15,646.44	3.3	6.08
Top 3	15350	Mosman	112,362.33	2.4	5.24
Top 3	15990	Northern Beaches	90,901.82	2.7	6.2
Top 3	18500	Woollahra	82,325.04	2.3	5.11

Screenshot 1 of the Answer to Business Question 1



H last_12 AS (SELECT * FROM go | Enter a SQL expression to filter results (use Ctrl+Space)

123 pct_0_4	123 pct_5_14	123 pct_15_19	123 pct_20_24	123 pct_25_34	123 pct_35_44	123 pct_45_54	123 pct_5
7.21	13.3	6.31	7.07	15.32	13.53	12.65	
7.99	14.76	6.91	6.78	15.51	15.25	12.48	
6.08	13	7.13	7.71	13.63	12.46	13.65	
5.24	12.37	5.53	4.5	11.96	14.36	15.01	
6.2	13.52	5.7	4.96	11.7	15.1	14.79	
5.11	10.51	4.8	5.41	17.39	14.45	12.95	

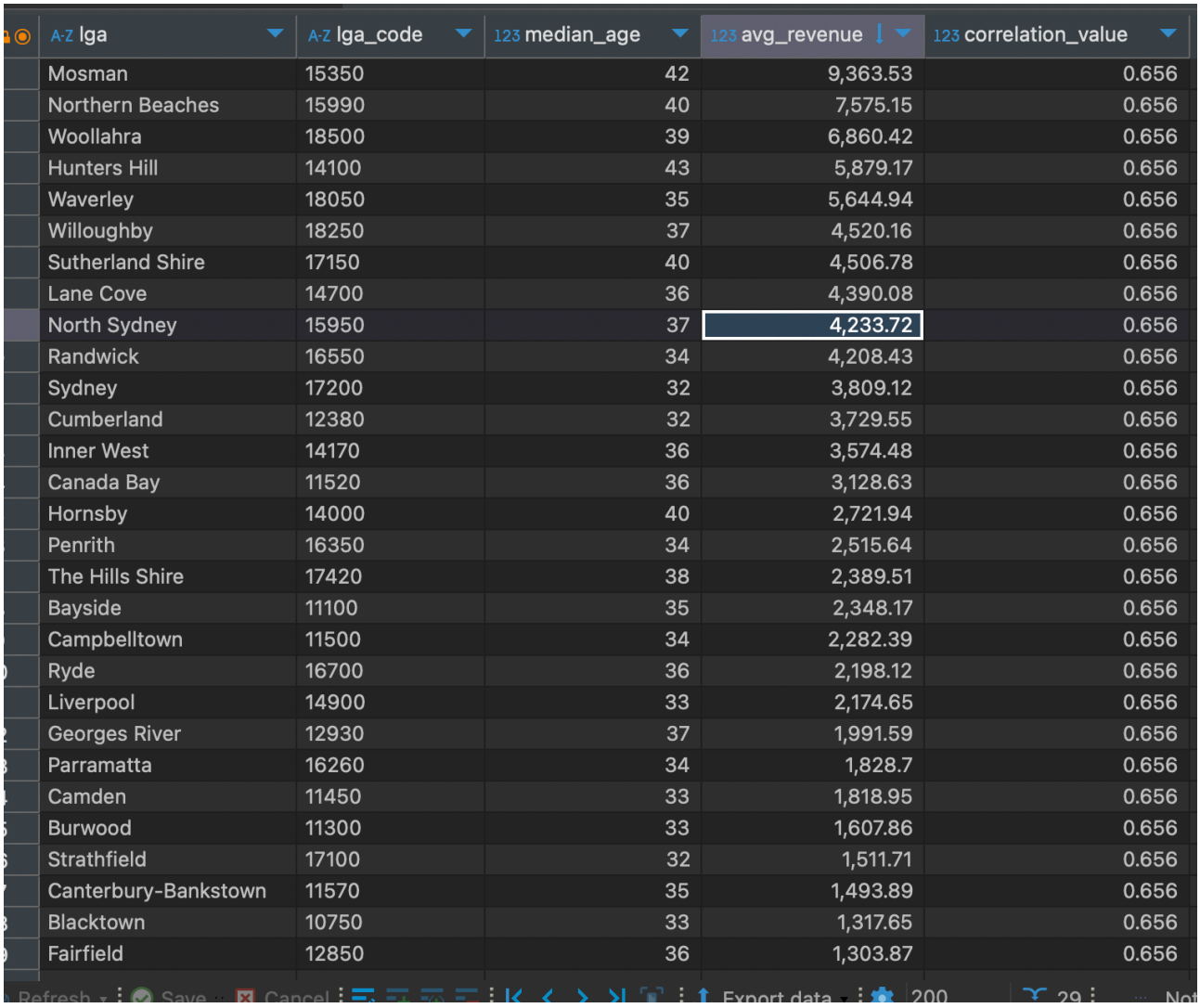
Screenshot 2 of the Answer to Business Question 1

	pct_25_34	pct_35_44	pct_45_54	pct_55_64	pct_65_74	pct_75_84	pct_85_plus
1	7	15.32	13.53	12.65	10.67	7.19	4.58	2.16
2	8	15.51	15.25	12.48	10	6.39	2.87	1.05
3	1	13.63	12.46	13.65	12.55	7.67	4.33	1.8
4	5	11.96	14.36	15.01	11.91	10.54	5.42	3.2
5	6	11.7	15.1	14.79	11.24	8.87	5.17	2.75
6	1	17.39	14.45	12.95	10.75	10.2	5.5	2.95

Screenshot 3 of the Answer to Business Question 2

2) Correlation Between Median Age and Airbnb Revenue

The analysis compared the median age of residents in each LGA (sourced from the 2016 Census)



lga	lga_code	median_age	avg_revenue	correlation_value
Mosman	15350	42	9,363.53	0.656
Northern Beaches	15990	40	7,575.15	0.656
Woollahra	18500	39	6,860.42	0.656
Hunters Hill	14100	43	5,879.17	0.656
Waverley	18050	35	5,644.94	0.656
Willoughby	18250	37	4,520.16	0.656
Sutherland Shire	17150	40	4,506.78	0.656
Lane Cove	14700	36	4,390.08	0.656
North Sydney	15950	37	4,233.72	0.656
Randwick	16550	34	4,208.43	0.656
Sydney	17200	32	3,809.12	0.656
Cumberland	12380	32	3,729.55	0.656
Inner West	14170	36	3,574.48	0.656
Canada Bay	11520	36	3,128.63	0.656
Hornsby	14000	40	2,721.94	0.656
Penrith	16350	34	2,515.64	0.656
The Hills Shire	17420	38	2,389.51	0.656
Bayside	11100	35	2,348.17	0.656
Campbelltown	11500	34	2,282.39	0.656
Ryde	16700	36	2,198.12	0.656
Liverpool	14900	33	2,174.65	0.656
Georges River	12930	37	1,991.59	0.656
Parramatta	16260	34	1,828.7	0.656
Camden	11450	33	1,818.95	0.656
Burwood	11300	33	1,607.86	0.656
Strathfield	17100	32	1,511.71	0.656
Canterbury-Bankstown	11570	35	1,493.89	0.656
Blacktown	10750	33	1,317.65	0.656
Fairfield	12850	36	1,303.87	0.656

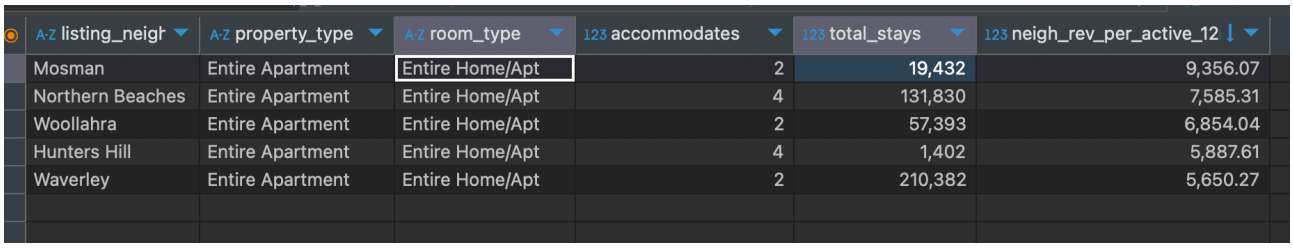
Screenshot from Query 2 Results

with the average estimated revenue per active Airbnb listing of the last 12 months (2020/05 - 2021/04).

The results show a moderate, significant positive correlation of 0.65 between the median age and the LGA's revenue. This means that LGAs with older median populations tend to generate higher Airbnb revenues, which tells the same story as the first analysis.

For example, high-performing LGAs such as Mosman (median age = 42) and Woollahra (39) recorded average revenues exceeding $6,000–$9,000 per active listing, while younger LGAs such as Blacktown (33) and Canterbury-Bankstown (35) showed substantially lower revenues (around $1,300–$1,500). This tells the story of affluent mature areas with older residents, which tend to attract higher-value Airbnb guests and premium property listings, whereas younger, family-oriented LGAs generate lower short-term rental income.

3) Best Performing Property Types for Top Listing Neighbourhoods



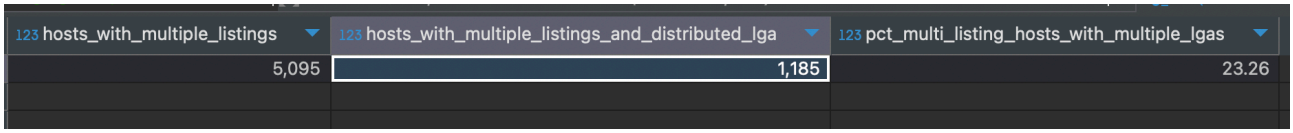
listing_neigh	property_type	room_type	accommodates	total_stays	neigh_rev_per_active_12
Mosman	Entire Apartment	Entire Home/Apt	2	19,432	9,356.07
Northern Beaches	Entire Apartment	Entire Home/Apt	4	131,830	7,585.31
Woollahra	Entire Apartment	Entire Home/Apt	2	57,393	6,854.04
Hunters Hill	Entire Apartment	Entire Home/Apt	4	1,402	5,887.61
Waverley	Entire Apartment	Entire Home/Apt	2	210,382	5,650.27

Result from Query 3

Here, an analysis of the top five neighbourhoods by estimated revenue per active listing — Mosman, Northern Beaches, Woollahra, Hunters Hill, and Waverley — reveals a consistent pattern in listing preferences and performance.

The best-performing property type for maximising stays in Sydney's top Airbnb neighbourhoods is an Entire Apartment/Home that accommodates 2–4 guests. This shows the appeal of high-quality, independent stays in affluent suburbs.

4) Distribution of Properties for Hosts with Multiple Listings



hosts_with_multiple_listings	hosts_with_multiple_listings_and_distributed_lga	pct_multi_listing_hosts_with_multiple_lgas
5,095	1,185	23.26

Result from Query 4

An analysis was performed to check whether hosts with multiple listings have their listings within the same LGA or are distributed across different LGAS, and also to find the percentage, if any. There are 5,095 hosts with multiple Airbnb listings in Sydney. Of these, 1,185 hosts (≈23.26%) had their listings distributed across different LGAs, while the remaining 76.7% managed various properties within the same LGA.

This indicates that most hosts who manage multiple listings focus on localised property management, where they have established expertise or market presence. A smaller but notable proportion (~23%) manage properties across different LGAs, showing more commercial or professional hosting operations that manage portfolios across multiple regions to diversify revenue streams.

5) Mortgage Coverage by Hosts with a Single Listing

	lga	hosts_total	hosts_can_cover	pct_can_cover
	Northern Beaches	4,221	2,841	67.31
	Mosman	325	199	61.23
	Sutherland Shire	492	292	59.35
	Waverley	4,055	2,283	56.3
	North Sydney	955	524	54.87
	Sydney	5,515	2,994	54.29
	Randwick	2,396	1,276	53.26
	Hunters Hill	52	26	50
	Woollahra	1,102	544	49.36
0	Inner West	1,831	877	47.9
1	Lane Cove	207	98	47.34
2	Willoughby	326	145	44.48
3	Canada Bay	325	136	41.85
4	Cumberland	326	135	41.41
5	Hornsby	307	114	37.13
6	Penrith	111	40	36.04
7	Camden	37	13	35.14
8	Ryde	408	133	32.6
9	The Hills Shire	216	69	31.94
0	Bayside	1,047	331	31.61
1	Canterbury-Bankst	415	128	30.84
2	Burwood	125	37	29.6
3	Parramatta	363	103	28.37
4	Liverpool	90	25	27.78
5	Georges River	239	66	27.62
6	Strathfield	115	31	26.96
7	Campbelltown	53	14	26.42
8	Fairfield	39	8	20.51
9	Blacktown	213	36	16.9

Refresh Save Cancel Export data 200

Result from Question 5

The analysis compared the estimated annual revenue per active listing against each LGA's annualised median mortgage repayment (sourced from the Census G02 dataset) to determine whether single-listing hosts could cover their housing costs through Airbnb income alone.

From the analysis, it shows that Northern Beaches has the highest percentage of single-listing hosts able to cover their annual mortgage repayments, with approximately 67.3% of hosts achieving full coverage. Other top-performing LGAs include Mosman (61.2%), Sutherland Shire (59.3%), and Waverley (56.3%).

In contrast, more affordable regions such as Blacktown (16.9%), Fairfield (20.5%), and Campbelltown (26.4%) recorded the lowest proportions of hosts capable of covering their mortgage costs through Airbnb income.
It can be inferred that the high-performing LGAs offer better financial sustainability for single-listing hosts, whereas outer suburban LGAs generate lower Airbnb returns relative to housing costs.

5. Conclusion

This project successfully demonstrated the design and implementation of a production-ready ELT data pipeline using Apache Airflow, PostgreSQL, and dbt Cloud to process and model Airbnb and Census data for Sydney. Using the Medallion architecture (Bronze -> Silver -> Gold), the pipeline ensured clean, consistent, and historically accurate data suitable for analytical reporting.

To summarise the findings,

The high revenue LGAs are characterised by smaller household sizes and older populations, reflecting affluent and mature markets. This is also established by the moderate positive correlation between median age and the revenue of LGAs, suggesting older and wealthier regions perform better. The most profitable listings were Entire Apartments/Homes accommodating 2–4 guests, while about 76% of hosts with multiple listings operated within a single LGA, reflecting localised management. Additionally, Northern Beaches recorded the highest proportion (67.3%) of single-listing hosts whose annual Airbnb revenue was sufficient to cover their annualised mortgage repayments.