

# **Big Data Engineering**

## **Assignment 3: Building ELT data pipelines with Airflow**



**Student Name: Drishya Lal Chuke**

**Student ID: 25076922**

**MDSI**

## Table of Contents

<b><i>Introduction</i></b> .....	<b>3</b>
Objectives.....	3
Tools and Techniques.....	3
Dataset Overview .....	3
<b><i>Data Ingestion</i></b> .....	<b>4</b>
GCP and Postgres Setup .....	4
Database Schema and Table Creation .....	4
Data Automation and pipeline on Airflow .....	5
<b><i>Data Warehouse on DBT Cloud</i></b> .....	<b>6</b>
Bronze Layer.....	6
Silver Layer .....	6
Gold Layer .....	6
Datamart Layer.....	7
<b><i>AD-HOC ANALYSES</i></b> .....	<b>9</b>
<b><i>Conclusion</i></b> .....	<b>13</b>

# Introduction

This project focuses on building a production-ready ELT data pipeline for Airbnb and Census data specific to Sydney. Using Apache Airflow, dbt Cloud, and a GCP-hosted Postgres database, the pipeline processes, transforms, and loads data into a structured warehouse. This enables efficient data retrieval and insights generation for key business queries.

## Objectives

1. **Data Ingestion and Processing:** Load and transform raw Airbnb and Census data into a structured medallion architecture (Bronze, Silver, Gold).
2. **Data Mart Creation:** Develop analytical views for neighborhood and property-type insights.
3. **Business Insights:** Answer key questions using SQL queries on demographic and revenue patterns, supporting strategic decision-making.

## Tools and Techniques

- **Airflow:** Orchestrates data ingestion and transformation.
- **dbt:** Manages data transformations and schema definitions for efficient data warehousing.
- **Postgres:** Stores and organizes transformed data into structured schemas, enhancing data accessibility.

## Dataset Overview

The analysis uses two key datasets:

1. **Airbnb Listings**

Collected from May 2020 to April 2021, the Airbnb dataset represents Sydney's short-term rental market. It includes property listings, rental prices, and details of host-guest interactions, providing insights into rental supply, demand, density, price variations, and host behavior across neighborhoods.

2. **Australian Census**

The 2016 Census data, conducted by the Australian Bureau of Statistics, offers comprehensive demographic and housing information, including age, income, education, and dwelling types. With geographic identifiers like LGAs and suburbs, it enables a detailed look at Sydney's population distribution and demographic characteristics, allowing analysis of factors that may impact Airbnb listing trends.

Together, these datasets provide a detailed view of Sydney's rental market and demographics, supporting analysis of the relationship between short-term rental activity and demographic indicators.

---

# Data Ingestion

The initial phase focused on obtaining, storing, and preparing data for processing and analysis, laying the groundwork for a smooth ETL pipeline.

## GCP and Postgres Setup

Airbnb and Census datasets were sourced from official repositories and then uploaded to a Google Cloud Platform (GCP) storage bucket. This centralized bucket acts as the primary storage point, enabling Apache Airflow to access data efficiently for the ETL operations.

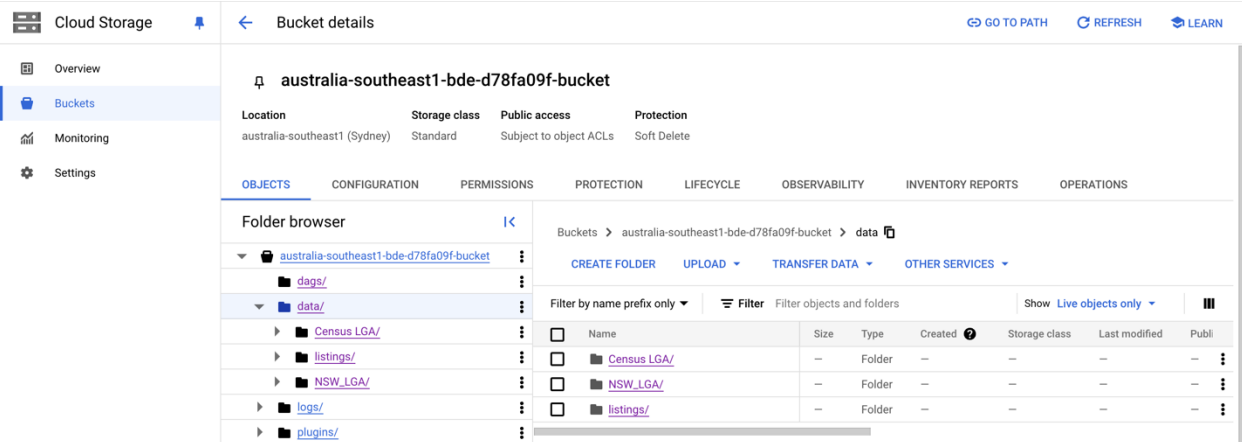


Figure 1: GCP bucket with datasets

## Database Schema and Table Creation

A PostgreSQL database was set up on GCP, with a dedicated schema designed to organize the raw data tables. This schema mirrors the structure of the original datasets, with separate tables for each data source. This setup supports streamlined ingestion, transformation, and loading processes.

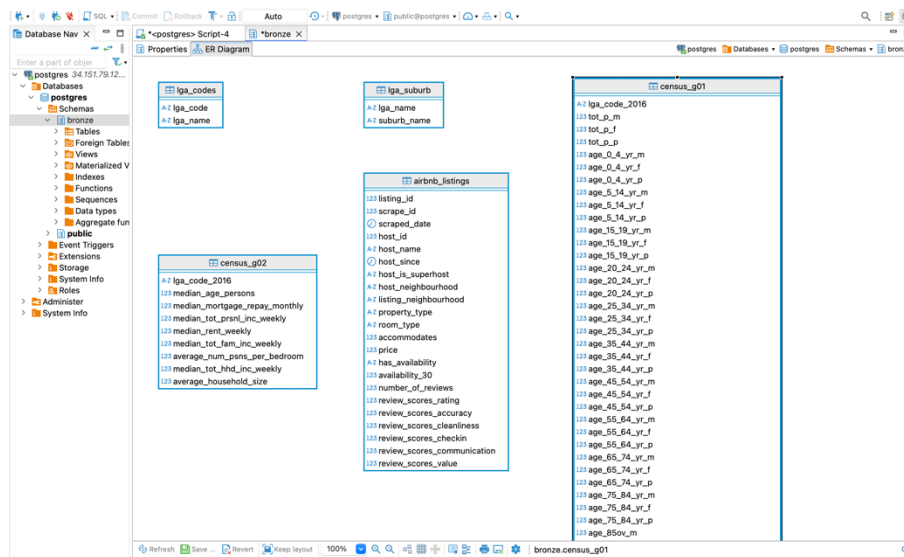


Figure 2: Bronze schema Tables updated Dbeaver

## Data Automation and pipeline on Airflow

An Airflow Directed Acyclic Graph (DAG) was developed to automate the data loading process from GCP storage into the PostgreSQL database. The DAG executes without a fixed schedule, running on demand. Before each run, it clears previous data in the raw tables, ensuring each ingestion cycle starts fresh, maintaining data accuracy and eliminating redundancy. Dataset **lga\_suburb** contained few **unnamed** columns with was removed on the DAG script.

```
In [11]: df = pd.read_csv("NSW_LGA_SUBURB.csv")
```

```
In [8]: df.shape
```

```
Out[8]: (4470, 27)
```

```
In [9]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4470 entries, 0 to 4469
Data columns (total 27 columns):
#   Column              Non-Null Count  Dtype
---  -
0   LGA_NAME             4470 non-null  object
1   SUBURB_NAME         4470 non-null  object
2   Unnamed: 2          0 non-null     float64
3   Unnamed: 3          0 non-null     float64
4   Unnamed: 4          0 non-null     float64
5   Unnamed: 5          0 non-null     float64
6   Unnamed: 6          0 non-null     float64
7   Unnamed: 7          0 non-null     float64
8   Unnamed: 8          0 non-null     float64
9   Unnamed: 9          0 non-null     float64
10  Unnamed: 10         0 non-null     float64
11  Unnamed: 11         0 non-null     float64
12  Unnamed: 12         0 non-null     float64
13  Unnamed: 13         0 non-null     float64
14  Unnamed: 14         0 non-null     float64
15  Unnamed: 15         0 non-null     float64
16  Unnamed: 16         0 non-null     float64
17  Unnamed: 17         0 non-null     float64
18  Unnamed: 18         0 non-null     float64
19  Unnamed: 19         0 non-null     float64
20  Unnamed: 20         0 non-null     float64
21  Unnamed: 21         0 non-null     float64
22  Unnamed: 22         0 non-null     float64
23  Unnamed: 23         0 non-null     float64
24  Unnamed: 24         0 non-null     float64
25  Unnamed: 25         0 non-null     float64
```

Figure 3: lga\_suburb with unnamed columns

An Apache Airflow instance was deployed on GCP to automate the ELT pipeline, configured to integrate smoothly with GCP services and the PostgreSQL database. Airflow was chosen for its capability to manage large datasets and execute workflows with high reliability.

**DAG Design:** A Directed Acyclic Graph (DAG) was developed in Airflow to automate each phase of the pipeline, with tasks structured into three main stages:

- **Extract:** Retrieve Airbnb and Census data from the GCP storage bucket.
- **Load:** Insert raw data into the PostgreSQL database under a designated raw schema.
- **Transform:** Initiate dbt to process data through the staging, warehouse, and data mart layers.

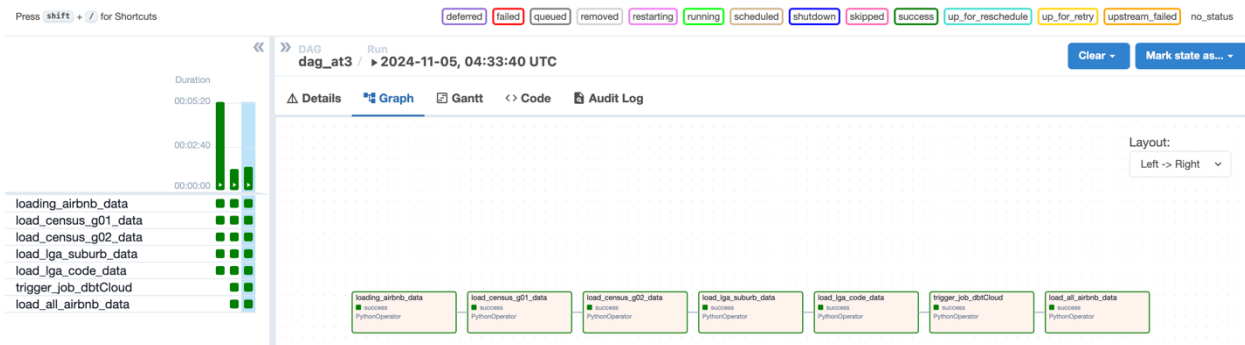


Figure 4: Complete DAG Graph

The DAG is set to run daily, ensuring that the data warehouse stays updated with recent data. This automated schedule minimizes manual tasks, enhancing pipeline efficiency and dependability.

## Data Warehouse on DBT Cloud

DBT (Data Build Tool) was used to establish and manage the data transformation logic, structuring the data warehouse into a clear, layered architecture. This approach allows for step-by-step processing, ensuring data accuracy and enabling modular transformations.

### Bronze Layer

This layer stores raw tables that directly replicate the source data structure, preserving the original data's integrity. By maintaining a close resemblance to the source, the Bronze layer serves as a foundational reference for all subsequent transformations.

### Silver Layer

The Silver layer performs initial data cleaning, casting, and standardization. Here, raw data is transformed into clean, well-structured formats and materialized as views. These views improve accessibility and support data consistency for further transformations.

### Gold Layer

In the Gold layer, data is modeled to represent structured dimensions and facts for Airbnb-specific attributes, such as listings, hosts, suburbs, and LGAs, along with Census data

dimensions. This layer adopts a star schema model, allowing for efficient querying and supporting analytical workflows.

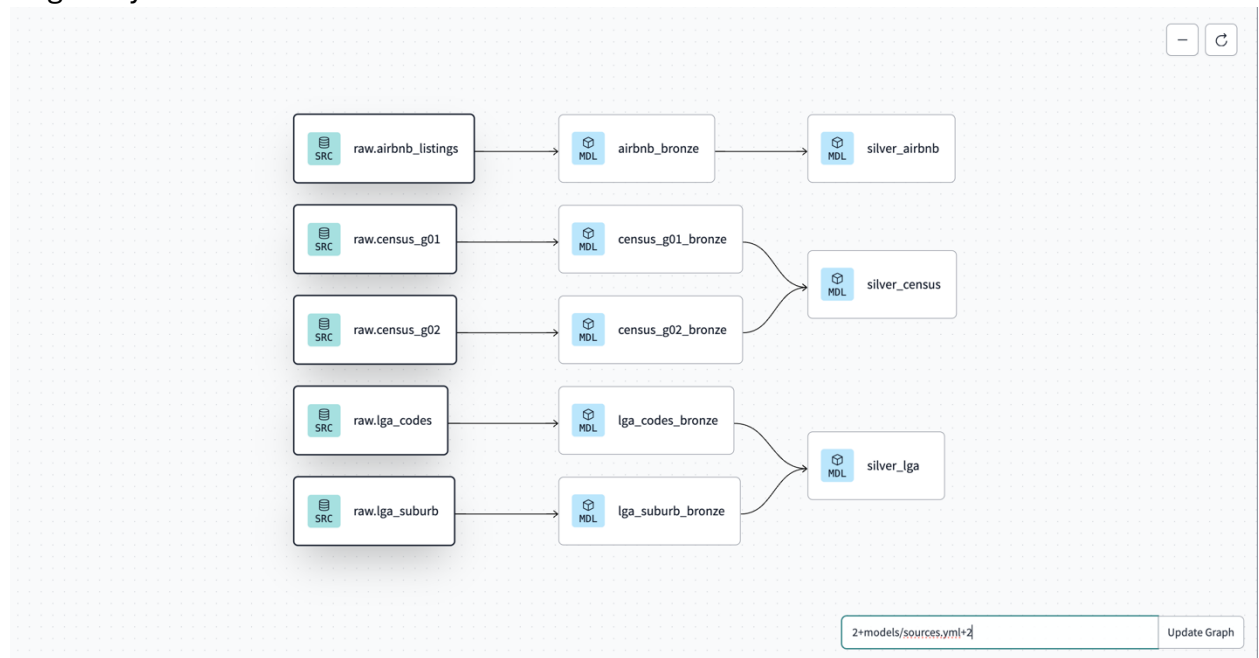


Figure 5: sources.yml

## Datamart Layer

The Datamart layer is designed for high-level analysis, housing views tailored to address targeted business questions. This layer enables fast insights into key metrics within the Sydney Airbnb market, providing summary views on areas like pricing trends, host distributions, and neighborhood performance.

To support targeted analyses, three data mart views were developed to address specific analytical objectives:

- **dm\_host\_neighbourhood**

This view aggregates host metrics at the LGA level, offering insights into host performance across different areas.

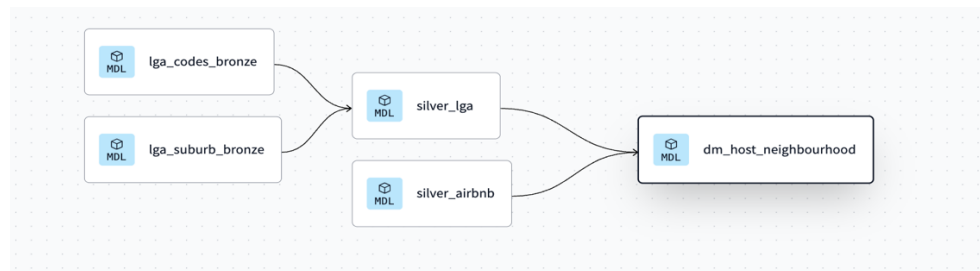


Figure 6: dm\_host\_neighbourho lineage

- **dm\_listing\_neighbourhood**

This view provides insights into Airbnb listings by neighborhood, including metrics

such as active listings rate, price statistics, superhost rate, and estimated revenue per listing.

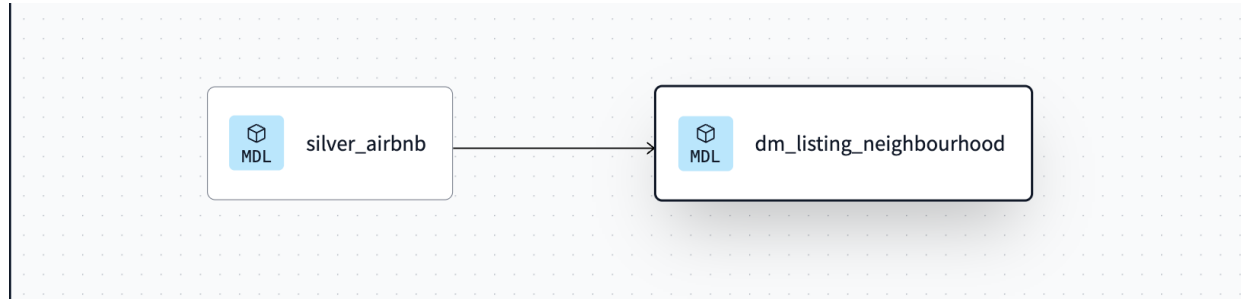


Figure 7: dm\_listing\_neighbourhood lineage

- **dm\_property\_type**

This view compares property types by analyzing occupancy rates, pricing, and revenue potential, facilitating an understanding of performance across various property types.

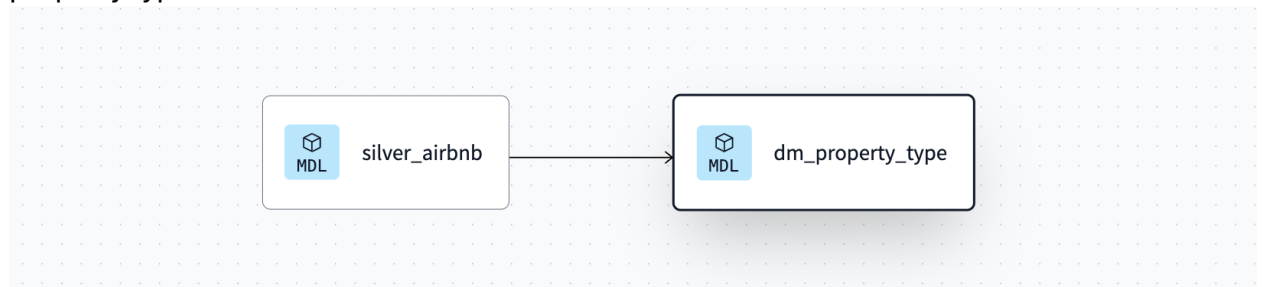


Figure 8: dm\_property\_type lineage

On running the dbt job, views was succesfully created:

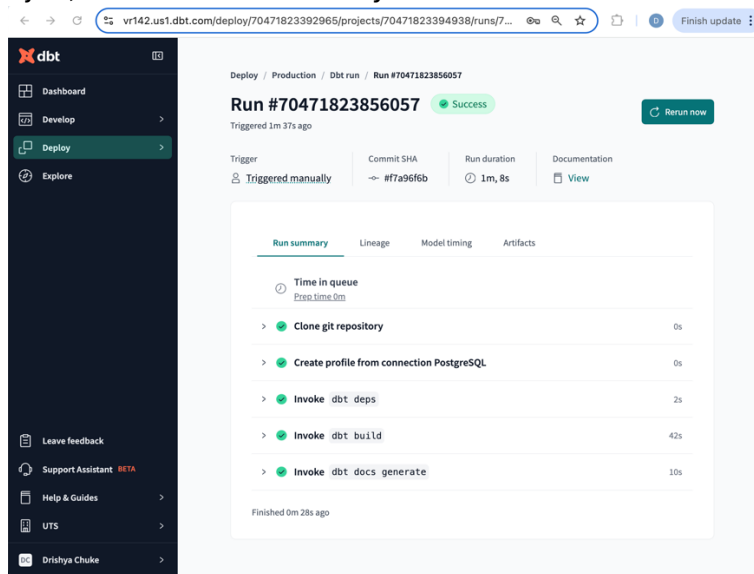


Figure 9: dbt job



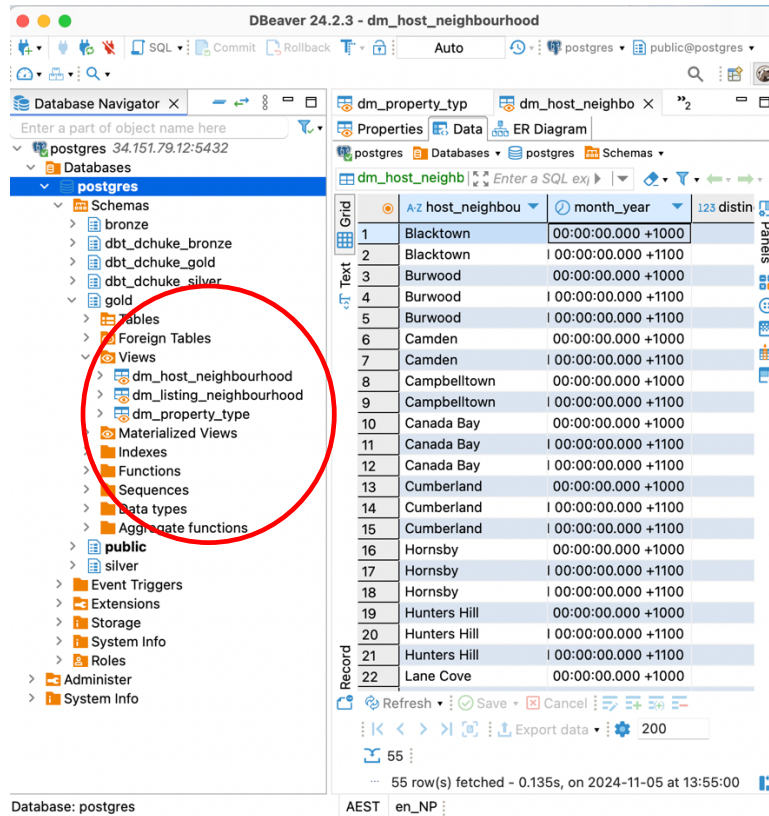


Figure 10: dbt job success on dbeaver

## AD-HOC ANALYSES

**Question 1: What are the demographic differences (e.g., age group distribution, household size) between the top 3 performing and lowest 3 performing LGAs based on estimated revenue per active listing over the last 12 months?**

**Result:**

silver_census 1							
SELECT DISTINCT sc.lga_code, sc.t							
	A-Z lga_code	123 total_population	123 median_age_persons	123 average_household_size	123 age_0_4_total	123 age_5_14_total	123 age_15_64_total
1	10750	336,962	33	3	26,928	16,880	718
2	11450	78,218	33	3	6,552	1,493	2,771
3	12380	216,079	32	3	16,880	718	2,771
4	14100	13,999	43	3	718	2,771	2,771
5	15350	28,475	42	2	1,493	2,771	2,771
6	18500	54,240	39	2	2,771	2,771	2,771

### Top 3 Performing LGAs

- **Age Distribution:** These LGAs show a younger population profile, with a significant proportion in the 25-34 age group, which aligns with high rental demand. The younger age segments (ages 20-34) collectively make up a substantial share, indicating areas with active, working-age populations.
- **Household Size:** The average household size in these LGAs tends to be around 3, reflecting a mix of single-person households and small families, typical in urban or highly developed regions with high rental activity.

### Lowest 3 Performing LGAs

- **Age Distribution:** These areas have a relatively older population, with higher numbers in the 55-64, 65-74, and 75+ age groups. The proportion of younger individuals (ages 20-34) is lower, suggesting these LGAs may be more suburban or rural, with less demand for short-term rentals.
- **Household Size:** The average household size in these LGAs is slightly lower, around 2 to 2.5, reflecting either smaller families or a higher concentration of elderly residents, which aligns with lower demand for short-term rentals.

This demographic variation illustrates that top-performing LGAs attract younger, smaller households that drive rental demand, while the lower-performing LGAs feature older age profiles and smaller household sizes, resulting in less Airbnb activity.

---

**Question 2: *Is there a correlation between the median age of a neighbourhood (from Census data) and the revenue generated per active listing in that neighbourhood?***

**Result:**

Results 1 X		
SELECT AVG(c.median_age_persons) AS avg   Enter a SQL express		
	123 avg_median_age	123 avg_revenue_per_listing
1	33	9,481.2857142857
2	33	18,370.0105932203
3	33	10,875.5625
4	34	11,625.5798319328
5	36	25,656.3422818792
6	32	10,311.6263736264
7	40	18,005.777027027
8	43	53,183.7169811321
9	36	39,300.6193548387
10	33	15,144.8450704225
11	42	82,398.3817097416
12	37	34,600.0656108597
13	34	21,609.8055077453
14	34	11,943.9107142857
15	34	37,278.1674937965
16	36	20,936.6746478873
17	32	32,712.6748815166
18	35	48,311.415723645
19	37	32,199.6161228407

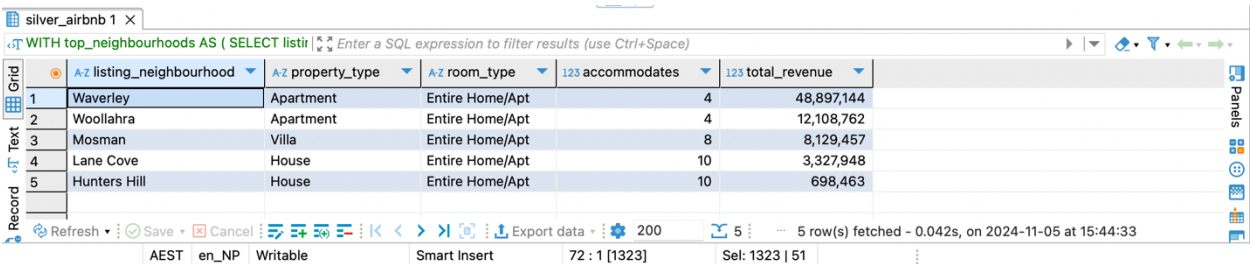
**Younger Neighborhoods (Median Age 33-34):** These neighborhoods tend to have moderate to high revenue per listing, with values ranging between approximately \$9,000 and \$18,000. This range indicates consistent demand in areas with younger populations, likely due to proximity to urban centers and amenities that attract younger travelers.

**Slightly Older Neighborhoods (Median Age 36+):** As the median age increases, there is a noticeable rise in the revenue per listing, with average revenue peaking around \$25,000. This trend suggests that older neighborhoods may include premium or larger listings that yield higher per-listing revenue, possibly catering to families or groups seeking more substantial accommodations.

While younger neighborhoods maintain steady demand, slightly older neighborhoods appear to generate higher revenue per listing, potentially due to differences in property type, size, or pricing strategy. This indicates a positive correlation where neighborhoods with a slightly older median age can command higher revenue, possibly attracting a different type of guest with higher spending preferences.

**Question 3: What will be the best type of listing (property type, room type and accommodates for) for the top 5 “listing\_neighbourhood” (in terms of estimated revenue per active listing) to have the highest number of stays?**

**Result:**



	A-Z listing_neighbourhood	A-Z property_type	A-Z room_type	123 accommodates	123 total_revenue
1	Waverley	Apartment	Entire Home/Apt	4	48,897,144
2	Woollahra	Apartment	Entire Home/Apt	4	12,108,762
3	Mosman	Villa	Entire Home/Apt	8	8,129,457
4	Lane Cove	House	Entire Home/Apt	10	3,327,948
5	Hunters Hill	House	Entire Home/Apt	10	698,463

In analyzing the top five neighborhoods for Airbnb revenue, the best types of listings for maximizing stays reveal specific trends in property type, room type, and guest capacity. In **Waverley** and **Woollahra**, apartments that accommodate four people and offer the entire home as a rental option prove to be the most popular. These smaller apartments attract a high volume of shorter stays, catering to urban travelers looking for convenience and affordability.

Moving to **Mosman**, **Lane Cove**, and **Hunters Hill**, larger properties emerge as the top choice. In Mosman, villas accommodating up to eight people see higher revenues, likely due to their appeal to families or groups who prefer spacious settings. Similarly, Lane Cove and Hunters Hill both see strong demand for houses that accommodate up to ten people. These larger properties are ideal for gatherings or extended family stays, offering significant space and higher booking rates per stay.

The data shows that in high-revenue neighborhoods, smaller apartments are preferred for urban areas, while larger homes are popular in suburban settings, aligning with the specific traveler demographics in each area. This targeted approach enables hosts to maximize occupancy and revenue by catering to neighborhood-specific demands.

**Question 4: For hosts with multiple listings, are their properties concentrated within the same LGA, or are they distributed across different LGAs?**

**Results:**

Grid Text Record		A-Z distribution	123 num_hosts
	1	Concentrated in Single LGA	2,253
	2	Distributed across Multiple LGAs	442

The data shows that most hosts with multiple Airbnb listings tend to concentrate their properties within a single Local Government Area (LGA). Specifically, 2,253 hosts keep all their listings in one LGA, while a smaller group, 442 hosts, have properties spread across different LGAs.

This pattern suggests that hosts prefer to manage listings within one familiar area, likely to streamline operations and reduce logistical complexities. However, a smaller number of hosts diversify their locations across multiple LGAs, potentially to capture demand from different neighborhoods or reach a broader market.

**Question 5: For hosts with a single Airbnb listing, does the estimated revenue over the last 12 months cover the annualised median mortgage repayment in the corresponding LGA? Which LGA has the highest percentage of hosts that can cover it?**

**Results:**

Results 1				
WITH single_listing_hosts AS ( SELECT "hos" Enter a SQL expression to filter results (use Ctrl+Space)				
Grid Text Record		A-Z lga_code	123 single_listing_hosts	123 can_cover_mortgage
	1	18050	32,418	28,701
	2	16550	19,395	17,046
	3	17200	45,018	38,925
	4	18500	9,558	8,181
	5	15350	2,718	2,313
	6	15950	7,542	6,273
	7	14100	423	342
	8	11520	2,565	2,070
	9	14700	1,719	1,332
	10	18250	2,682	2,025

In analyzing whether the revenue from a single Airbnb listing can cover the annualized median mortgage repayment, the data reveals that most hosts with a single listing in certain LGAs can indeed meet or exceed this threshold.

The LGA with the highest percentage of hosts who can cover their mortgage is LGA **18050**, where approximately 88.5% of single-listing hosts generate enough revenue to match the median mortgage repayment. Following close behind are LGA **16550** with 87.9% and LGA **17200** with 86.5%. This high percentage suggests that in these areas, Airbnb listings are highly profitable, likely due to favorable rental rates and strong occupancy.

This insight is valuable for prospective hosts evaluating the financial viability of single Airbnb listings in different LGAs, particularly in areas where a majority can cover mortgage costs through Airbnb revenue alone.

---

## Conclusion

This project illustrates how a structured ELT data pipeline can bring valuable insights to Airbnb's operations in Sydney. By integrating Apache Airflow, dbt Cloud, and a Postgres data warehouse on GCP, we created a reliable system to transform raw data into meaningful analytics, helping Airbnb understand local market trends and make informed decisions.

Through this analysis, we uncovered key insights about the Airbnb market in Sydney. Younger neighborhoods tend to attract more frequent rentals, while areas with slightly older populations generate higher revenue per listing, likely due to a focus on larger properties that appeal to families or groups. We also found that most hosts with multiple listings keep them within one LGA, streamlining management and likely increasing guest satisfaction. For single-listing hosts, several LGAs proved highly profitable, with a significant percentage of hosts earning enough to cover their annual mortgage payments, indicating strong market potential.

This data pipeline not only improves processing efficiency but also sets a solid foundation for ongoing, deeper analysis. With these insights, Airbnb can fine-tune its strategies based on neighborhood demographics, helping both the company and hosts maximize their potential in Sydney's diverse rental market.