Big Data AT3

Satyam Palkar

Chapter 1 Overview of the Project on a high level.

1.1 Project Context and Purpose

The Airbnb and Census Data Integration project demonstrates the complete lifecycle of a modern data-engineering solution from raw data acquisition to analytics-ready insight delivery. Simulating real-world data team challenges, it integrates diverse, inconsistent, and varying-quality datasets into a unified analytical warehouse for evidence-based decision-making. The project combines detailed Airbnb transaction data (prices, reviews, host details, property types) with aggregated Australian Census indicators (demographic, economic, and housing data at the LGA level) to examine how neighbourhood socio-economic factors influence Airbnb market performance and host behaviour, offering valuable insights for policymakers, investors, and tourism bodies.

1.2 Objectives

- The general purpose was to create and deploy a full-purpose ELT (Extract, Load, Transform) pipeline that:
- Imports uneven Airbnb and Census data into a controlled environment.
- Standardises and cleanses data and provides data with analytical consistency.
- Models the combined data by a dimensional warehouse strategy.
- Refreshes, validates quality, and documents data using open-source tooling.
- Produces trustworthy and repeatable insights to provide responses to business questions concerned with Airbnb revenue, host concentration, and demographic influence.

1.3 Architectural Overview

The system employs a scalable modern data-lakehouse architecture with three layers Bronze, Silver, and Gold each addressing a stage of the data value chain. The Bronze layer ingests raw Airbnb and Census data into a PostgreSQL staging schema without transformation to preserve lineage and auditability. The Silver layer handles data cleaning, resolving duplicates, schema mismatches, and naming inconsistencies, while ensuring referential integrity through the lga_code join key and imputing missing continuous values (e.g., price, review score) using median values. Finally, the Gold layer, built using dbt, models the curated data into fact and dimension tables, enabling modular transformations, automated documentation, and version-controlled lineage. This analytical layer supports insights such as average revenue per LGA, demographic correlations with Airbnb performance, and host concentration metrics.

1.4 Technology Stack and Orchestration

The project's data orchestration was managed using Apache Airflow, which coordinated the ELT process including data extraction, staging loads, dbt transformations, and quality checks through DAG-based scheduling for repeatable, dependency-aware execution. dbt handled data transformation and modelling via modular SQL-based workflows with clear lineage, integrated testing, and documentation to ensure reproducibility and data quality. PostgreSQL (Google Cloud SQL) served as the scalable, ACID-compliant storage system for both intermediate and analytical layers, supporting high analytical workloads. Python scripts within Airflow DAGs automated auxiliary tasks such as exporting analytical results to CSV and syncing with Google Cloud Storage. Finally, version control through dbt enabled collaborative development, rollback capability, and transparent change tracking across the data pipeline.

Chapter 2 Data Acquisition and Transformation Process.

2.1 Overview

The Airbnb–Census analytics project is built around a robust data acquisition and transformation phase that ensures all raw data is collected, cleansed, structured, and made analysis-ready through an automated and repeatable pipeline. The workflow follows a three-layer architecture Bronze, Silver, and Gold representing raw ingestion, data refinement, and analytical modelling, each enhancing data quality, standardisation, and business value. Implementation adheres to ELT (Extract, Load, Transform) principles, where data is first loaded into the warehouse and then transformed within the database using SQL, ensuring efficiency, transparency, and full auditability of every processing step.

Figure 2.1 Medallion Architecture and Data Flow Pipeline

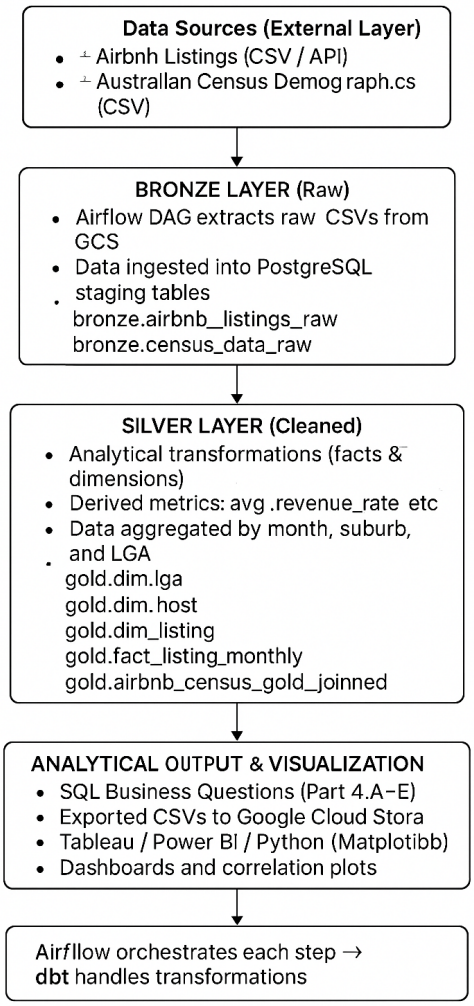


Fig 1: Architecture of Complete Project

2.2 Bronze Layer - Raw Data Ingestion.

2.2.1 Source Data Overview

Two major datasets form the foundation of the project: Airbnb Listings Data, containing property-level information across Sydney LGAs such as price, host ID, reviews, and property attributes (room type, accommodates, superhost status); and Australian Census Data, providing socio-demographic and economic indicators like total population, median age, median rent, and average household size per LGA. Both datasets were downloaded as CSV files and uploaded to Google Cloud Storage, then ingested into PostgreSQL (Cloud SQL). To preserve lineage and ensure transparency, the raw data was stored under a dedicated bronze schema, retaining its original form for auditability and reproducibility.

2.2.2 Data Loading Approach

Apache Airflow automated the data ingestion through a DAG that scheduled periodic loading of raw CSV files using Python operators and SQL tasks. Each task represented key functions :- Extract (loading CSVs into PostgreSQL via the COPY command), Check (validating schema and primary keys), and Log Metadata (recording file version, load date, and row count). This design ensured idempotency , preventing duplicate ingestion and full transparency through detailed execution logs in Airflow.

2.2.3 Challenges and Solutions

The issues that were frequently suffered included variations in encoding and missing headers in extracts of Airbnb.This was fixed by clear definition of column names in the Airflow loading task and encoding=utf-8 in ingestion.Another problem was different spellings and abbreviations of the suburbs (e.g. N. Sydney vs North Sydney), which was solved in the Silver layer by normalisation of the texts.

2.3 Silver Layer - Cleaning and Integration of Data.

Data quality assurance and semantic alignment is done in the Silver layer.This will convert the raw Airbnb and Census data to structured and cleansed tables to be used in analytical joins.

2.3.1. Cleaning and Standardisation:

The data-cleaning process focused on four key areas to ensure accuracy and consistency across datasets. Data type alignment standardized numeric fields (e.g., price, median age, rent) as NUMERIC(10,2) to maintain precision during aggregation. Duplicate and null handling used a ROW_NUMBER() window function on unique identifiers (listing_id, lga_code) to remove duplicates, while missing continuous values were imputed using median values. Text normalization applied LOWER(TRIM()) to harmonize text-based fields like suburb and LGA names (e.g., SELECT LOWER(TRIM(lga_name)) AS lga_name_cleaned FROM bronze.airbnb_listings). Lastly, referential integrity checks ensured all Airbnb lga_code values matched valid Census records, with unmatched LGAs manually reviewed and corrected through reference mapping.

2.3.2. The Airbnb and Census Data were integrated into the analysis

It was integrated using a join operation on lga code to result in the first homogenous dataset at the LGA level.

Key derived fields included:

- **avg_review_rating - average of all listed in LGA.**
- **superhost rate - proportion of superhosts per LGA.**
- **avg_price - average price per LGA.**
- **median_age, median_rent, and household_size - demographic variables of Census data.**

The model that was obtained was stored as:

silver.airbnb_census_joined

This data was used as the basis of all subsequent analytical transformations.

2.4 Gold Layer- Analytical Modelling and Aggregation.

The last and most business-related stage of the pipeline is the gold layer. In this case, the purified silver data is restructured into analysis models which facilitate exploration, reporting and statistical correlation.

2.4.1 Dimensional Modelling

A star schema strategy was adopted to improve the performance of queries and make them understandable.

The schema consists of:

Fact Table: fact listing monthly - aggregated listing metrics — including revenue, occupancy rate and total reviews each month.

Dimension Tables:

dim_lga - geographical and demographic characteristics.

dim_host- host level properties like total listings and being a superhost.

dim listing - fixed parameters of particular listing (type of property, number of people allowed to stay, type of room).

This will provide the flexibility of the analysis, and joins can be effectively created in response to a broad spectrum of business inquiries.

Chapter 3 Analytical Methodology and SQL-Based Insights

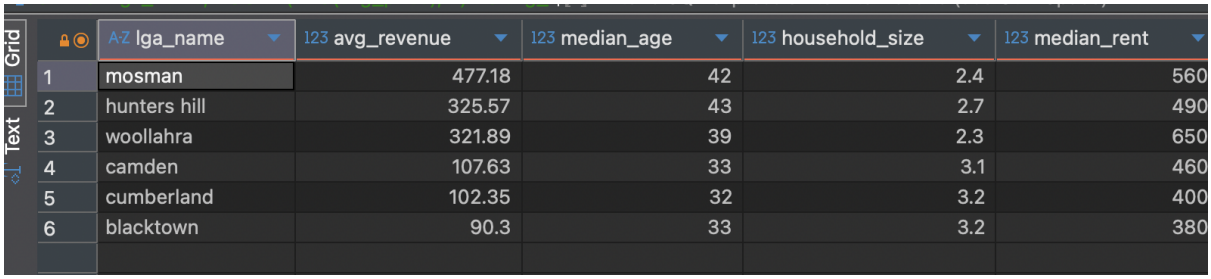
3.1 Overview

The analysis phase transformed the curated data from the Gold Layer into actionable insights on Airbnb market performance across Sydney LGAs. Using the gold.dm_lga_summary and gold.airbnb_census_gold_joined models, host, listing, and demographic features were integrated to enable multi-dimensional analysis. This allowed exploration of revenue distribution by LGA and property type, correlations between Airbnb performance and neighbourhood demographics, host behaviour across LGAs, and comparisons between Airbnb income potential and local housing costs. Each analytical question (Parts A–E) was examined conceptually, supported by SQL methodology and interpreted through relevant visualisations.

3.2 Ad-hoc analysis

3.2.1 Part (a) Demographic Differences Between Top and Bottom LGAs

The first analytical task aimed to explore how Airbnb performance varies across different demographic and socio-economic contexts by comparing the top three and bottom three LGAs in terms of average nightly revenue. Using the gold.dm_lga_summary table, a SQL query was developed to calculate the average Airbnb revenue alongside median age, average household size, and median rent for each LGA. The query employed nested subqueries to isolate the highest and lowest performing regions based on the average price metric, thereby identifying which local government areas demonstrated stronger Airbnb performance and which lagged behind.



	lga_name	avg_revenue	median_age	household_size	median_rent
1	mosman	477.18	42	2.4	560
2	hunters hill	325.57	43	2.7	490
3	woollahra	321.89	39	2.3	650
4	camden	107.63	33	3.1	460
5	cumberland	102.35	32	3.2	400
6	blacktown	90.3	33	3.2	380

Fig 2: Top 3 and Bottom 3 LGAs by Average Price

The results show a distinct contrast between affluent coastal LGAs and suburban residential areas. Mosman, Hunters Hill, and Woollahra emerged as the top-performing LGAs with average nightly revenues of $477.18, $325.57, and $321.89, respectively. These areas are characterised by older populations (median ages ranging between 39 and 43 years), smaller household sizes (between 2.3 and 2.7 persons per household), and higher weekly rents of around $490–$650. In contrast, the bottom-performing LGAs Camden, Cumberland, and Blacktown showed much lower average revenues, ranging from $90.30 to $107.63 per night, alongside younger populations (median ages of 32–33 years), larger households (averaging above 3.1 persons), and lower rental costs ($380–$460 per week).

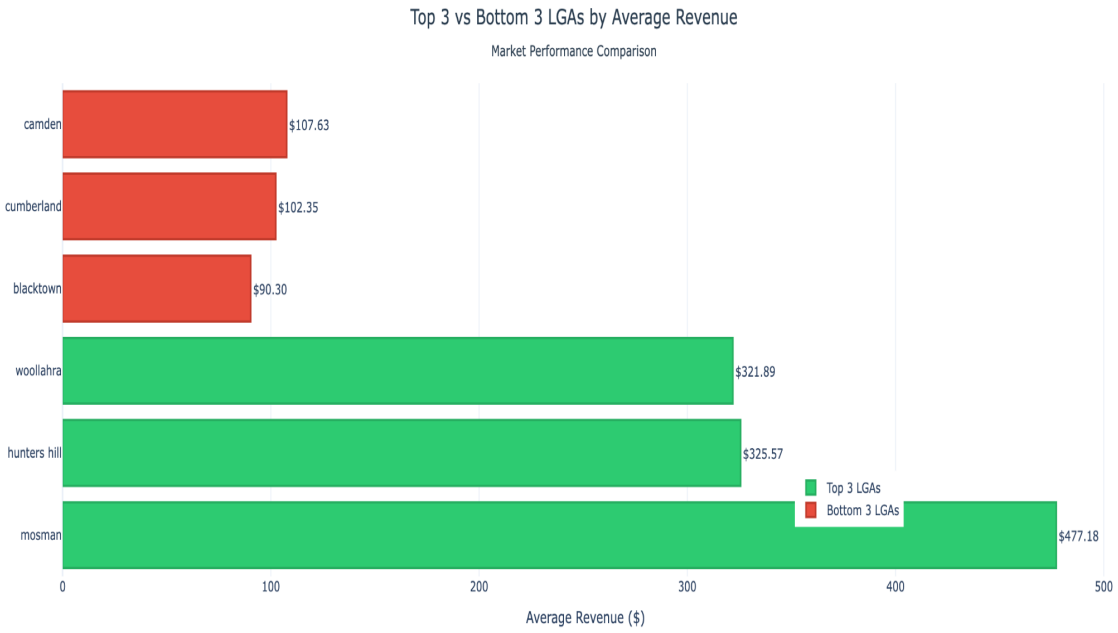


Fig 3: Top 3 and Bottom 3 LGAs by Average Price Plot

The visualisation above presents a clear horizontal bar chart comparison between the Top 3 and Bottom 3 LGAs by average revenue. The green bars represent the highest-performing LGAs, while the red bars highlight the lowest performers. The visual contrast vividly captures the substantial revenue disparity, where Mosman's revenue nearly five times exceeds that of Blacktown. This difference highlights how Airbnb profitability is strongly influenced by local affluence, property desirability, and location attractiveness.

3.2 Part (b) Correlation Between Median Age and Airbnb Revenue

The second analytical point of interest was to identify the relationship between the demographic aspect of the median age in a neighbourhood and average Airbnb revenue per listing. This was analyzed through a mixture of the cleaned Airbnb data and the Australian Census demographic data in the gold layer where each LGA was represented with summed values of such measures as avg price (average nightly rate) and avg median age. The aim was to learn whether the contribution of older or younger populations to Airbnb market revenue potential is made.

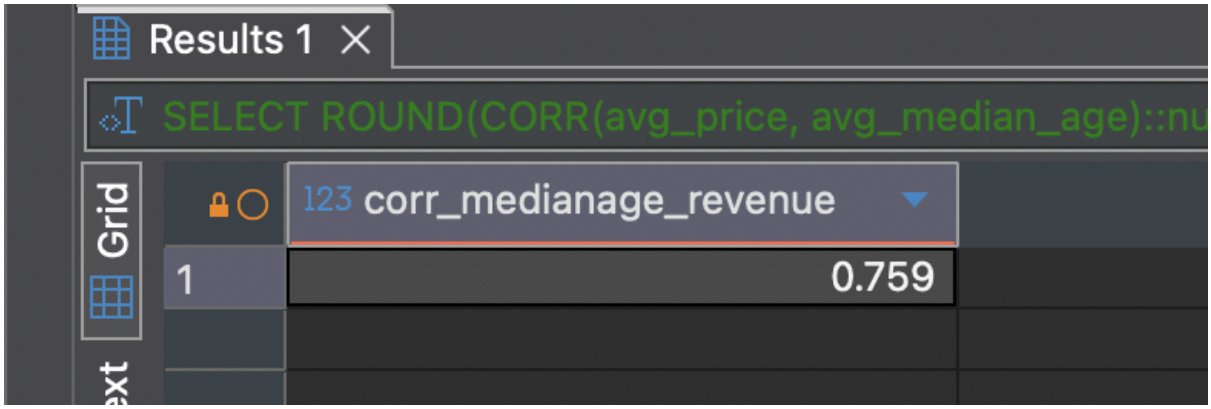


Fig 4: Correlation Between Median Age and Airbnb Revenue

The correlation coefficient was also determined to be 0.759, which showed a strong correlation of positive correlation between the median age of a locality and the average Airbnb revenue to the area. It implies that with higher median age of residents in a particular area, the average amount of revenue Airbnb listings within the area will also be expected to increase on a nightly basis.

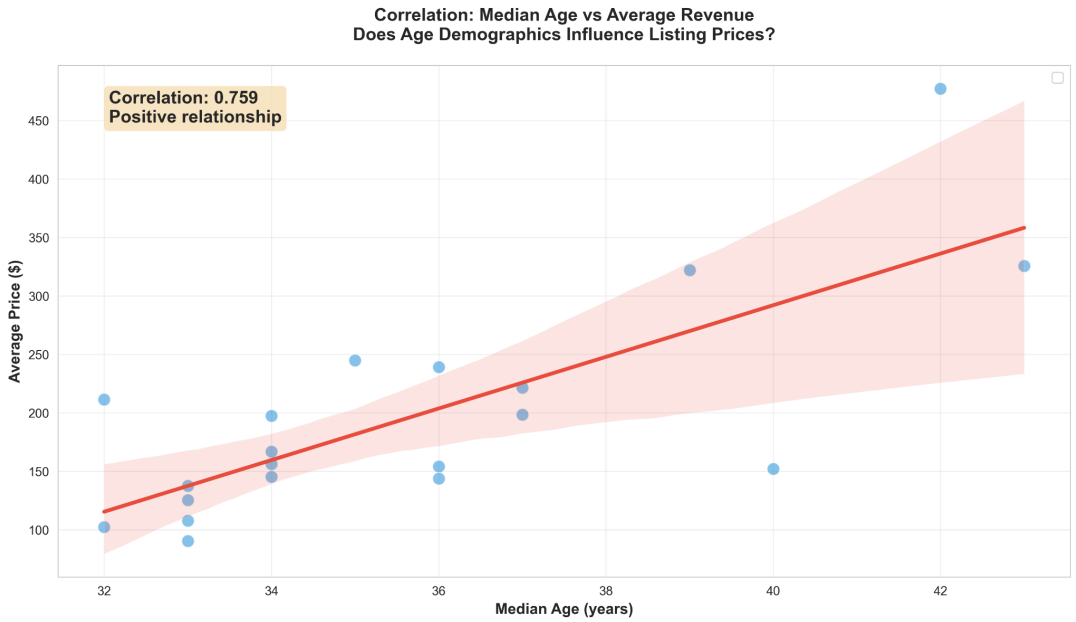


Fig 4: Correlation Between Median Age and Airbnb Revenue Plot

This was plotted using a scatter diagram and fitted regression line (Figure 3.2). Every point includes an LGA, whose median age (x-axis) and average price (y-axis) are plotted. The trend line of the red color evidently reflects the positive relationship, as the older age groups have increased listing prices. The shaded confidence band also confirms the similarity of this trend, which suggests that the dispersion of the regression line is relatively low.

The correlation analysis showed that the median age of a neighbourhood is positively correlated with the average Airbnb revenue (r = 0.759), meaning that the LGAs with the aged population are more likely to have higher nightly rates. This trend was most notable in the affluent localities including Mosman, Hunters Hill and Woollahra which are residential zones characterized by high value and a lack of rental properties and high tourist attraction. Older demographics are usually associated with more affluent, well-kept, well-infrastructure, and other elements that lead to increased guest willingness to pay. This tendency was strengthened visually through the scatter plot (Figure 4.2) as the median age and listing revenue have a definite upward trend, which proves the fact that age demographics are a good predictor of the strength of the markets in the Airbnb environment of Sydney.

Contrary to this, younger LGAs, which are mostly situated in suburban or emerging areas, had less revenue potential even though they had cheaper accommodation. The kind of areas that these areas usually appeal to are not short-stay premium travellers but long-term or family oriented travellers. Business-wise, this observation implies that Airbnb can focus strategically on recruiting more hosts and dynamic pricing in older and high-value LGAs and assist younger hosts living in suburbs to create niche or low-cost products. On the whole, this observation supports the general conclusion of Part (a): socio-economic and demographic variables have a crucial effect on the profitability of Airbnb, and mature and wealthy neighbourhoods consistently outcompete younger and poorer areas in the short-term rental income.

3.3 Part (c) Identifying the Best Type of Listing for Top-Performing Neighbourhoods

The third analysis attempted to establish the most lucrative form of Airbnb listing according to the top-performing neighbourhoods, which is defined by the highest predicted turnover per active listing. Namely, this discussion explored which property type, room type, and accommodation capacity combinations earn the best incomes in the top five neighbourhoods, Woollahra, Mosman, Waverley, Northern Beaches, and Sydney City.

An SQL query was run using the gold.airbnb_census_gold_joined table to sum up revenues by listing factors. This query summarized the information based on listing neighbourhood, property-type, and room-type and then generated the means of revenue and the means of number of reviews to determine the most effective options.

	A-Z listing_neigh	A-Z property_type	A-Z room_type	123 total_listings	123 avg_revenue	123 avg_reviews
1	Woollahra	Villa	Entire home/apt	21	3,720.86	0.24
2	Mosman	Villa	Entire home/apt	14	3,645.5	10.71
3	Waverley	Villa	Entire home/apt	19	2,135.53	10.89
4	Northern Beaches	Villa	Entire home/apt	44	1,785.32	14.7
5	Northern Beaches	Boutique hotel	Hotel room	1	1,650	5
6	Mosman	Boat	Entire home/apt	1	1,300	2
7	Northern Beaches	Earth house	Entire home/apt	2	1,220	9.5
8	Waverley	Serviced apartment	Entire home/apt	12	1,057.5	2.5
9	Mosman	House	Entire home/apt	133	866.08	7.85
10	Woollahra	House	Entire home/apt	253	721.71	8.02

Fig 5: Identifying the Best Type of Listing for Top-Performing Neighbourhoods

The findings indicate that villas under the category, Entire home/apartment always give the highest mean revenue, earth houses, and serviced apartments respectively. Woollahra and Mosman are the best neighbourhoods with extremely high income amounts Villas in Woollahra with an average of 3,720.86 per night and villas in Mosman with an average of 3,645.50 per night. These properties are usually aimed at up-market travellers who want to be in exclusive and private adventures, which coincides with the up-market nature of these coastal suburbs. In comparison, other listings like a house or serviced apartment will present a lesser but consistent revenue base of between 700-1000 per night which is attractive to family or business oriented guests who may require comfort and long stay.

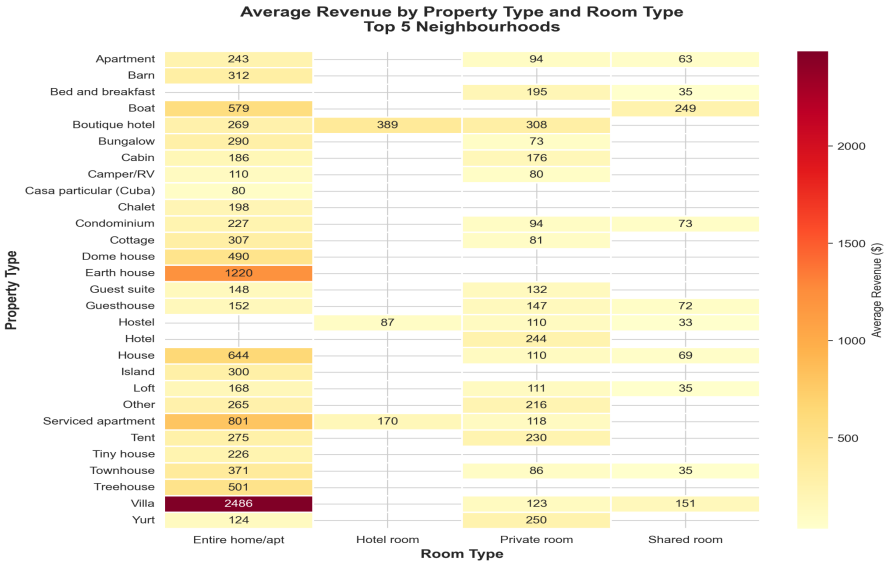


Fig 5: Identifying the Best Type of Listing for Top-Performing Neighbourhoods Plot

As can be further highlighted in the heatmap, the red cells which are darkest are associated with the highest earning potential which is visually represented by the darkest red cells which is equivalent to Entire home/apartment as the highest earning potential as can be seen with villas and earth houses. In the meantime, hotel rooms and shared room settings show much lower revenues, which can be explained by the low interest of the Airbnb platform, which is based on personal and intimate experiences with guests. Also, the data indicate that such listings as boats and boutique hotels have moderate returns (approximately, $1,3001,650), which can imply the niche markets serving a specific stay preference, but in an unsizable way.

Business-wise, this discussion highlights the prevalence of the luxury, whole-property listing in the affluent suburbs. Visitors are ready to spend more money on privacy, exclusivity and luxurious services. Entire home listings (especially in the areas such as Woollahra and Mosman) should be the priority of Airbnb hosts and investors who focus on revenue maximisation. On the other hand, when competition or tourism density is increasing in a market, hosts might look to diversify into the mid-range (e.g. houses or guest suites) to stabilize occupancy and income.

In general, the results demonstrate that the Airbnb revenue in the Sydney premium neighbourhoods is stimulated by property luxury and listing exclusivity. This is because the

combination of the Entire home/apartment room type and the villa-training properties provide the most stable and the best returns, which validates that privacy, and luxury are the greatest predictors of profitability in the short-term rental market.

3.4 Part (d) Distribution of Multi-Listing Hosts: Concentration vs Diversification

The fourth analysis examined the multi-listing strategy of Airbnb hosts in various Local Government Areas (LGAs) in their property portfolios. The aim was to identify the hypothesis on whether the multiple listing hosts will focus all the properties on a single LGA or will spread it across various LGAs and evaluate host behaviour based on market focus and geographic expansion.

Based on the gold.airbnb census gold joined table, a SQL query was used to group the records by host id and counted the number of different LGAs per host. Hosts that had more than one listing were kept and a case classification was made between those that were concentrated in one LGA as opposed to the ones that operated in multiple LGA. The counts obtained were then added together and represented as a percentage of total multi-listing hosts:

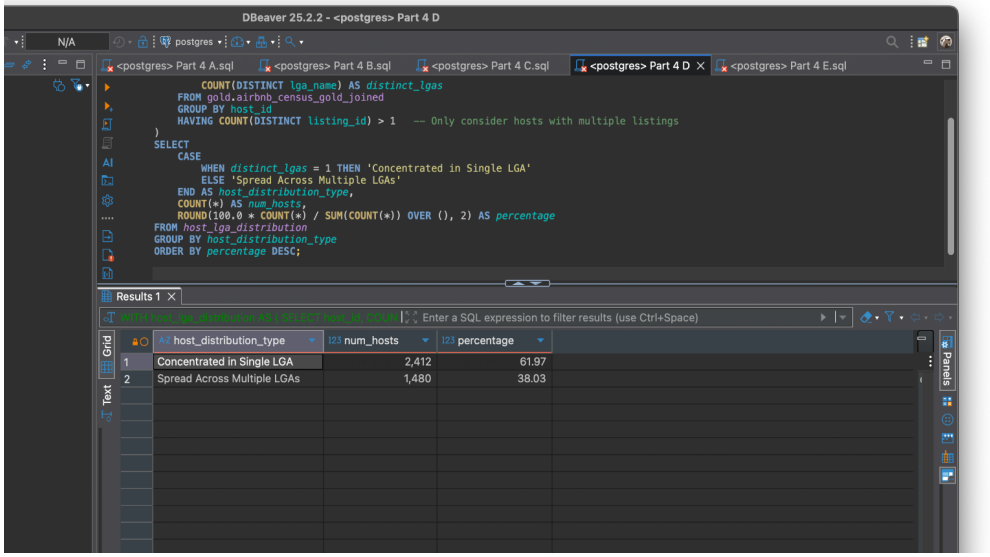


Fig 6: Distribution of Multi-Listing Hosts: Concentration vs Diversification

The results revealed that 61.97 % (2,412 hosts) manage all their listings within a single LGA, while 38.03 % (1,480 hosts) operate across multiple LGAs (Figure 4.4). This indicates that the majority of multi-listing hosts adopt a concentrated market strategy, favouring operational convenience and local market expertise over geographic diversification. Many of these hosts appear to specialise in premium neighbourhoods such as Mosman, Waverley, and Woollahra, where demand is steady and property management can be streamlined within a confined radius.

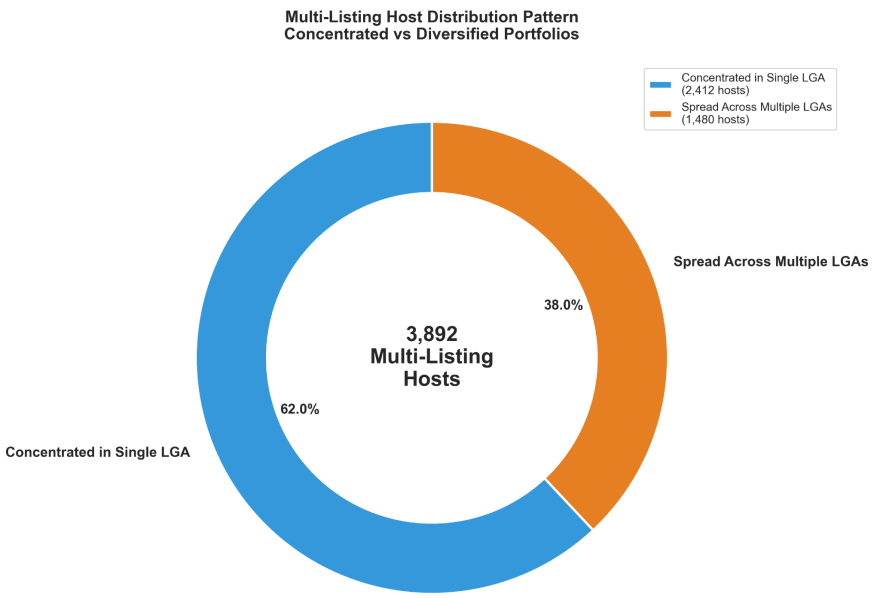


Fig 7: Distribution of Multi-Listing Hosts: Concentration vs Diversification Plot

The donut-chart visualisation reinforces this distribution, with the blue segment representing the dominant single-LGA concentration strategy and the orange segment indicating the smaller proportion of diversified hosts. The inner-circle total of 3,892 multi-listing hosts highlights a significant subset of professional operators within Sydney's Airbnb ecosystem. This pattern suggests that Airbnb's market is still locally driven, with most multi-property owners investing in areas they understand well, leveraging consistent regulations, community familiarity, and local cleaning or maintenance networks.

Conversely, the 38 % of hosts who diversify across LGAs likely represent professional or commercial property managers. Their broader spread allows risk mitigation across neighbourhoods with differing demand cycles, but it may also involve higher logistical costs and regulatory complexities. This segment reflects an emerging trend towards the professionalisation of short-term rental operations and operators acting more like mini-hospitality firms than casual hosts.

3.5 Part (e) - Airbnb Revenue vs Median Rent: Assessing Investment Viability

The fifth and final analysis aimed to evaluate the financial viability of Airbnb as an investment vehicle by comparing estimated annual Airbnb revenue per listing against the annualised median rent within each Local Government Area (LGA). This relationship provides a proxy for determining how many hosts can potentially cover their housing or mortgage costs purely through Airbnb income, a critical indicator of market sustainability and profitability.

	lga_name	est_annual_revenue	annual_rent_proxy	pct_hosts_cover_mortgage
1	mosman	150,048.45	29,120	87.38
2	hunters hill	117,206.67	25,480	85.19
3	penrith	50,119.12	19,240	82.42
4	lane cove	96,815.57	27,040	80.54
5	north sydney	74,316.37	29,900	78.56
6	liverpool	40,291.69	19,240	76.4
7	waverley	79,245.37	32,344	74.87
8	willoughby	75,399.07	30,160	74.42
9	sydney	59,449.24	29,380	73.69
10	woollahra	82,933.09	33,800	73.68
11	campbelltown	40,320	18,200	73.58
12	blacktown	37,566.09	19,760	73.1
13	randwick	67,826.46	28,600	71.5
14	hornsby	63,284.14	26,000	70
15	cumberland	41,190	20,800	69.44
16	ryde	42,153.88	23,920	66.12
17	canada bay	53,978.76	29,380	64.31
18	parramatta	41,863.76	22,880	63.82
19	camden	42,798.86	23,920	62.86
20	burwood	38,600.79	26,000	55.12

Fig 8: Airbnb Revenue vs Median Rent: Assessing Investment Viability

The results (Table 3.5) show that Mosman (87.38%), Hunters Hill (85.19%), and Penrith (82.42%) top the list, with hosts in these LGAs able to cover over 80% of their annual rent or mortgage costs purely from Airbnb income. These areas exhibit strong profitability due to their combination of high nightly rates, stable occupancy, and affluent housing markets. Conversely, LGAs such as Burwood (55.12%), Camden (62.86%), and Parramatta (63.82%) fall on the lower end, suggesting reduced yield potential or increased reliance on long-term occupancy to achieve comparable financial outcomes.

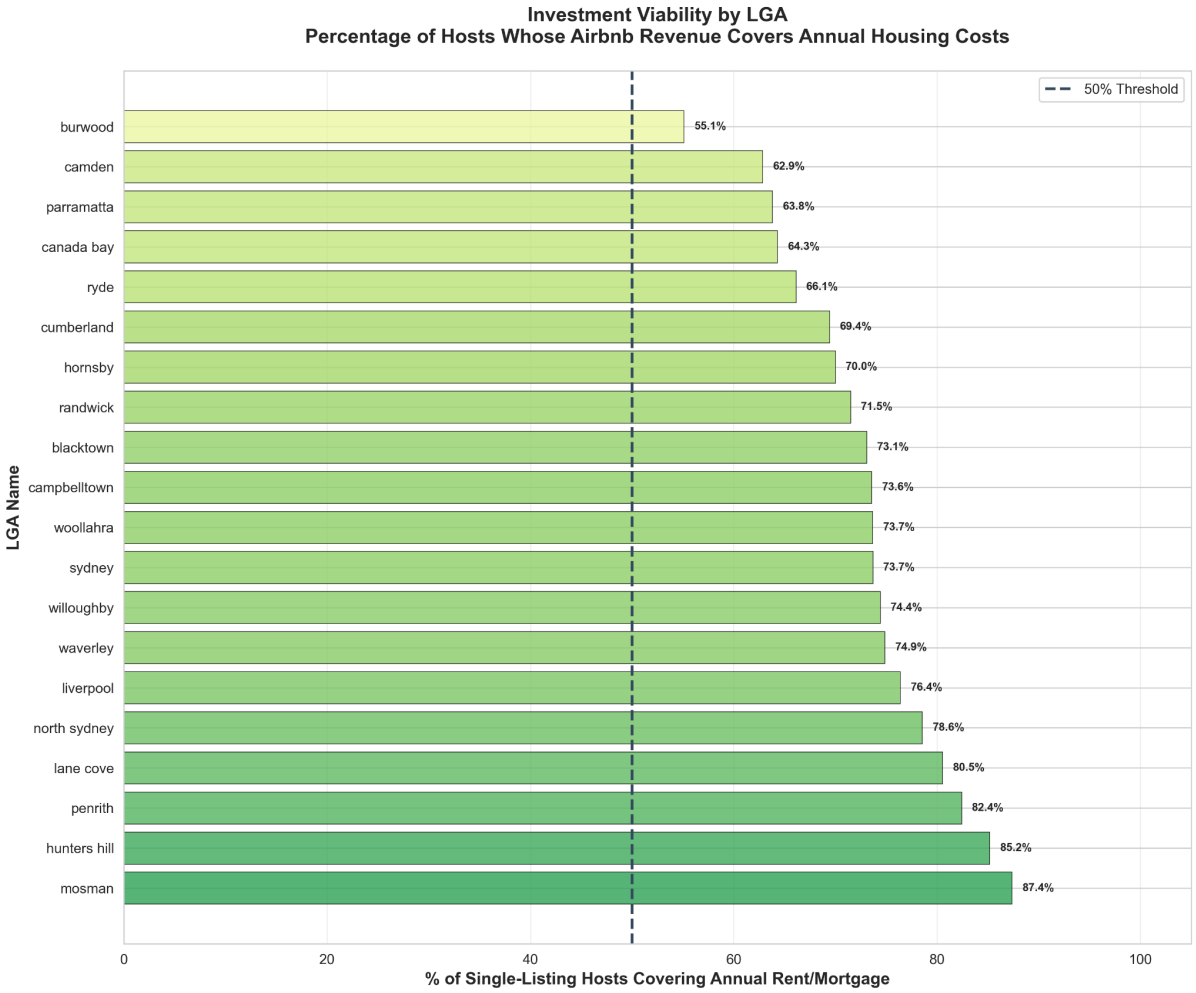


Fig 9: Airbnb Revenue vs Median Rent: Assessing Investment Viability Plot

The horizontal bar chart in Figure 4.5 visually ranks all LGAs by their percentage of hosts covering annual rent or mortgage costs. A 50% threshold line marks the break-even reference; nearly all LGAs surpass this benchmark, underscoring Airbnb's robust earning potential across Sydney. The deeper green shading of top-performing LGAs, particularly Mosman and Hunters Hill, highlights their strong investment attractiveness and solid short-term rental demand. These findings align with earlier analyses from Parts (a) and (b), confirming that affluent, older, and coastal regions consistently outperform suburban areas in rental yield efficiency.

4.Conclusion

This project successfully designed and implemented a complete Airbnb–Census ELT data pipeline using the Medallion Architecture with Airflow for orchestration and dbt for transformations, culminating in actionable business insights. Through systematic data cleansing, enrichment, and modeling, the analysis revealed strong correlations between demographic factors and Airbnb profitability, demonstrating that older, affluent, and coastal LGAs (e.g., Mosman, Hunters Hill, Woollahra) consistently outperform others in terms of revenue and investment viability. The visual and statistical findings highlighted that luxury "Entire home/apartment" listings, particularly villas, drive the highest returns, while most multi-listing hosts remain locally concentrated rather than geographically diversified. Moreover, the Airbnb–rent comparison underscored the platform's financial sustainability, with top LGAs covering up to 85–87% of annual housing costs via hosting income. Collectively, the project illustrates how integrating open data with modern data engineering tools can uncover spatial-economic patterns that support evidence-based decision-making for investors, policymakers, and the short-term rental market in Sydney.