Assignment 3 Handover Report

Author: Chengyang Zhou (24873630)
Big Data Engineering
Date: 31 Oct 2023

Census

Introduction

This research delves deeply into the intricacies of data pipelines, harnessing the capabilities of Airflow to orchestrate data workflows. By integrating Airflow with Postgres, the project ensures a smooth transition and manipulation of data, utilizing DBeaver as the primary interface for database management. Moving further into data warehousing, dbt within the VS Code environment becomes the cornerstone. This setup facilitates efficient data transformations, priming the data for analytical insights. To encapsulate the entire process, all structured data models and workflows are diligently documented and housed in a GitHub repository via git, providing a comprehensive and accessible record for all stakeholders.

Dataset Descriptions

Airbnb Listings (Sydney) dataset:
Overview: Airbnb is an online platform that revolutionized the hospitality sector by enabling property owners to list their spaces for short-term or long-term rentals. Given its worldwide reach and vast user base, Airbnb amasses a wealth of data on its listings and user interactions. For this project, we focus on Airbnb's operations in Sydney, using monthly data snapshots spanning May 2020 to April 2021.
Attributes:

1. Listing Details: Unique identifiers, property type, room type, accommodation capacity, and rental price.
2. Host Information: Host identifiers, names, status as a superhost, and their associated neighborhood.
3. Reviews: Metrics related to guest reviews, such as the number of reviews, average scores, and specific ratings for aspects like cleanliness and communication.

NSW Local Government Areas (LGA) dataset:
Overview: This dataset provides a mapping between suburbs and their corresponding Local Government Areas (LGAs) in New South Wales (NSW).
Attributes:

1. LGA Codes: A list of LGAs in NSW and their associated codes.
2. Suburb Mapping: Indicates which suburb belongs to which LGA.

Census LGA Dataset:
Overview: The Census of Population and Housing, conducted by the Australian Bureau of Statistics (ABS), is an extensive statistical undertaking that captures the demographic and socio-economic facets of Australia's populace. The dataset used herein provides granular data at the LGA level.
Attributes:

1. Demographics: Gender distribution, age groups, marital status, etc.
2. Education: Highest year of school completed, types of qualifications, and more.
3. Employment: Employment status, occupation types, industry of employment, etc.
4. Housing: Dwelling structures, tenure types, and mortgage repayments, among others.

Issues/Bugs Faced and Resolutions:
Issue: Failed to load listings data to Postgres through Airflow due to missing values.

- Resolution: Filled missing values with placeholders for columns like HOST_SINCE, HOST_IS_SUPERHOST, and HAS_AVAILABILITY in the DAG file to ensure data loading.

Issue: Encountered errors during ‘dbt run’, with indications of missing target folders and some staging files, even though files were correctly placed with the right syntax.

- Resolution: Realized that dbt was installed using pip install dbt instead of the correct pip install dbt-core. Reinstalling with the correct command resolved the issues.

Issue: Faced challenges matching lga_code_2016 with lga_code. Performing a join operation on lga_code resulted in empty tables.

- Resolution: Upon investigating the original file, it was discovered that lga_code_2016 was formulated using "LGA" + lga_code. Addressed the issue by using the 'like' clause during code matching.

Part 0: Dataset Acquisition

The foundational step involves sourcing the relevant datasets:

Airbnb Listings: A 12-month dataset detailing Airbnb listings in Sydney, sourced from monthly snapshots ranging from May 2020 to April 2021.
NSW LGA Data: Essential mapping data that provides:
LGA codes correspond to each Local Government Area in NSW.
Mapping of each suburb to its associated LGA.
Census LGA Data: Comprehensive demographic and socio-economic data at the LGA level, sourced from the Australian Census.

Part 1: Data Ingestion using Airflow and Postgres

Upon acquiring the datasets, they are transferred into the AirFlow storage bucket. Simultaneously, within the Postgres environment, a foundational structure termed 'raw’ schema is established. This schema acts as a blueprint for data storage, ensuring an organized and methodical data repository.
To create tables that mirror the raw schema specifications, DBeaver is employed. Tailored specifically for this project, these tables are geared to house the raw datasets. This foundation sets the stage for subsequent data processing and analysis.
The core of this project revolves around the orchestration of tasks using an Airflow Directed Acyclic Graph (DAG). This DAG defined in the dag_assignment3.py file, is carefully crafted to automate and sequence data operations. The purpose of this DAG is multifaceted:

1. Environment Setup: Necessary Python libraries are imported, including libraries specific to data manipulation such as pandas and numpy, and those specific to Airflow and database operations. The DAG is set up with specifics such as concurrency limits, retries, and scheduling intervals.
2. Data Loading Logic: Custom Python functions are developed to load CSV data into the Postgres database. For instance, the load_csv_to_postgres function handles the intricacies of reading data from the CSV files, preprocessing it, and finally inserting it into the appropriate tables in Postgres. Special attention is paid to handling specific columns that might have missing data or require transformation.

3. Task Definition: Individual tasks are then defined using the PythonOperator in Airflow. These tasks correspond to the functions that handle loading data from specific datasets. For instance:
➔ load_NSW_LGA_task loads data from the New South Wales Local Government Area (NSW LGA) dataset.
➔ load_listings_task processes and loads Airbnb listings data for specific months and years.
➔ load_census_task is dedicated to loading data from the Australian census datasets.
4. Task Sequencing: The tasks are then sequenced in a specific order to ensure data is loaded in the correct sequence, maintaining data integrity and dependencies.
5. Data Warehousing: Using dbt in the Visual Studio Code environment, the data is then structured and transformed for analytical purposes. This step ensures data is organized efficiently for queries and analyses.

Part 2: Data Warehousing using dbt
This transformation process is achieved through the creation of a data warehouse in Postgres, which consists of distinct layers, each serving its unique purpose.

a. Raw layer

Within the raw layer, snapshots capture the essence of the source data at specific intervals. These snapshots play a pivotal role, especially when understanding the evolution of data over time, such as with Slowly Changing Dimensions (SCDs). The project includes snapshots like census_g01_snapshot.sql and listings_snapshot.sql, which represent different facets of the source data.

b. Staging layer

The staging layer serves as the first significant checkpoint in the data's journey, acting as a bridge between raw data and its more refined state. Here, raw data is subjected to thorough cleaning, necessary transformations, and systematic renaming. The files integral to this layer include:

➔ census_g01_stg.sql: This processes data from the G01 table of the Census dataset, preparing it for subsequent analytical exploration.
➔ census_g02_stg.sql: Similarly, this file refines the G02 table, ensuring the data is primed for integration with other datasets.
➔ listing_stg.sql: This file is dedicated to the Airbnb listings, transforming raw listing data into a format that facilitates deeper analysis.
➔ nsw_lga_code_stg.sql and nsw_lga_name_stg.sql: These files process the NSW LGA data, which plays a crucial role in mapping listings to specific geographic regions.

c. Warehouse layer

In the warehouse layer, the data's structure further evolves, embracing the star schema model. This layer establishes a robust backbone for the entire data warehouse, ensuring optimal organization. The files within this layer are:

Dimension Tables:

- ➔ dim_census.sql: A comprehensive table capturing the demographics and key characteristics of the Australian population.
- ➔ dim_host.sql: This table sheds light on the diverse array of Airbnb hosts, their attributes, and their interactions within the platform.
- ➔ dim_lga.sql: Focusing on geography, this table details the Local Government Areas, essential for spatial analysis.
- ➔ dim_property.sql: As the name suggests, this table dives into the different property types listed on Airbnb.

Fact Table:

- ➔ facts_listings.sql: This central table combines information, providing a holistic view of Airbnb listings, and integrating data points from various dimensions.

d. Datamart layer

In the datamart layer, where the data is finessed into specific views tailored for precise analytical queries. Materialized for swift querying, these views encapsulate the essence of the project's analytical goals:

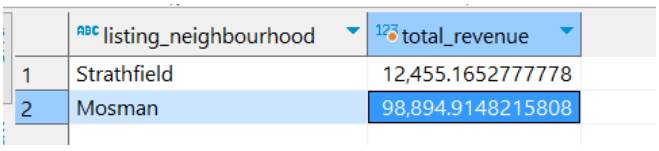
- ➔ dm_host_neighbourhood.sql: This view offers insights into the connection between hosts and the Local Government Areas of their listings.
- ➔ dm_listing_neighbourhood.sql: Focusing on temporal and spatial dimensions, this view highlights listing trends across different neighborhoods over time.
- ➔ dm_property_type.sql: This multi faceted view provides insights based on property characteristics, room types, accommodation capacities, and more, all juxtaposed against temporal attributes.

PART3:Ad-hoc Analysis

a. Population Differences between Best and Worst Performing Neighbourhoods:

Method:

To discern the main demographic differences between the top and bottom neighborhoods in terms of estimated revenue per active listing over the past year, we initially identified Mosman as the best-performing neighborhood and Strathfield as the least performing. Following this identification, these neighborhoods were mapped to their respective Local Government Area codes, a vital step given that our demographic data is categorized by LGA codes. With these LGA codes in hand, we queried the dim_census table to extract the demographic insights, which include age group distributions, gender demographics, and overall population counts, thus offering a detailed view of the population landscape of these neighborhoods.



	listing_neighbourhood	total_revenue
1	Strathfield	12,455.1652777778
2	Mosman	98,894.9148215808

	suburb_name	lga_code
1	STRATHFIELD	11300
2	MOSMAN	15350

lga_code_2016	male_population	female_population	total_population	
LGA11300	17,785	19,028	36,809	
LGA11300	17,785	19,028	36,809	
LGA11300	17,785	19,028	36,809	
LGA11300	17,785	19,028	36,809	
LGA15350	13,189	15,290	28,475	

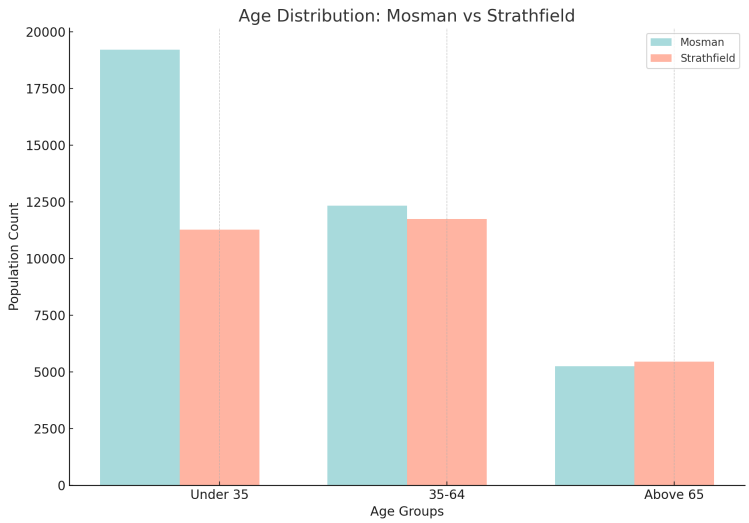
Result:

Mosman (Best Performing)

- Total Population: 36,809
- Male Population: 17,785
- Female Population: 19,028
- Population Under 35: 19,218 (52.21%)
- Population Between 35 and 64: 12,337 (33.52%)
- Population Above 65: 5,258 (14.28%)

Strathfield (Worst Performing)

- Total Population: 28,475
- Male Population: 13,189
- Female Population: 15,290
- Population Under 35: 11,276 (39.60%)
- Population Between 35 and 64: 11,751 (41.27%)
- Population Above 65: 5,452 (19.15%)



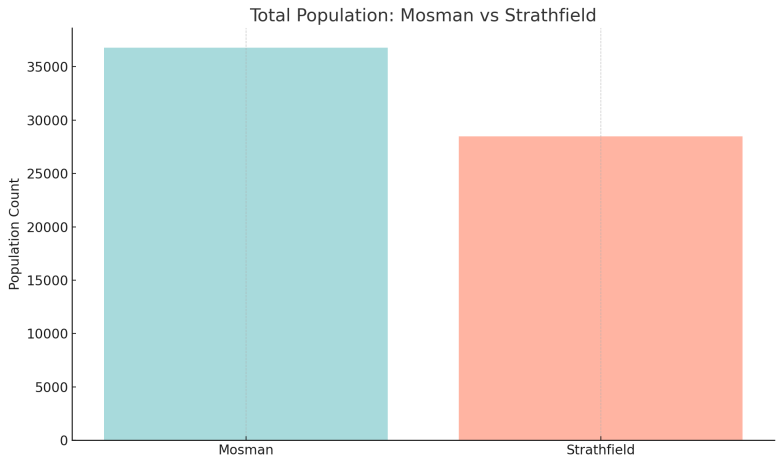
From the visualization, we can make several observations regarding the age distribution in Mosman and Strathfield:

➔ Young Population (Under 35): Mosman exhibits a notably larger young population when compared to Strathfield. This might indicate a more vibrant and dynamic environment in Mosman, potentially attracting younger professionals or students. On

the other hand, Strathfield has a relatively smaller young population, which could be indicative of a more family-oriented or settled environment.

➔ Middle-aged Population (35-64): Both neighborhoods display a substantial middle-aged population, with Strathfield having a slightly larger count in this category. This age group typically represents settled professionals, families, and individuals in their prime working years.

➔ Senior Population (Above 65): Strathfield possesses a more pronounced senior population when compared to Mosman. Such a demographic might imply a longer-term resident population in Strathfield, perhaps individuals or families who've resided in the area for many years.



From the visualization, we see a stark difference in the total populations of Mosman and Strathfield. Mosman boasts a notably higher population count compared to Strathfield.

b. Optimal Type of Listing for Highest Number of Stays:

Method:
The research began by extracting data for the top 5 neighborhoods based on their estimated revenue per active listing. Using the gathered data, and then sought to determine the most preferred type of listing, characterized by property type, room type, and capacity to accommodate guests. The results demonstrated a clear preference among guests for apartments that offer the entire home or apartment, especially those that can house 2 to 4 individuals.

WITH Top5Neighbourhoods AS (SELECT listi | Enter a SQL expression to filter results (use Ctrl+Space)

	listing_neighbourhood	property_type	room_type	accommodates	total_stays
1	Hunters Hill	Apartment	Entire home/apt	4	135
2	Hunters Hill	House	Private room	2	120
3	Hunters Hill	House	Entire home/apt	8	96
4	Hunters Hill	Apartment	Entire home/apt	2	90
5	Hunters Hill	House	Entire home/apt	10	67
6	Hunters Hill	Apartment	Private room	2	62
7	Hunters Hill	Townhouse	Entire home/apt	6	61
8	Hunters Hill	House	Entire home/apt	4	60
9	Hunters Hill	House	Entire home/apt	7	60
10	Hunters Hill	Entire house	Entire home/apt	6	45
11	Hunters Hill	House	Entire home/apt	6	44
12	Hunters Hill	Entire house	Entire home/apt	8	40
13	Hunters Hill	House	Entire home/apt	5	37
14	Hunters Hill	Townhouse	Entire home/apt	4	30

Result:

Hunters Hill:

- Property Type: Apartment
- Room Type: Entire home/apt
- Accommodates: 4 people
- Total Stays: 135

Mosman:

- Property Type: Apartment
- Room Type: Entire home/apt
- Accommodates: 2 people
- Total Stays: 2378

Northern Beaches:

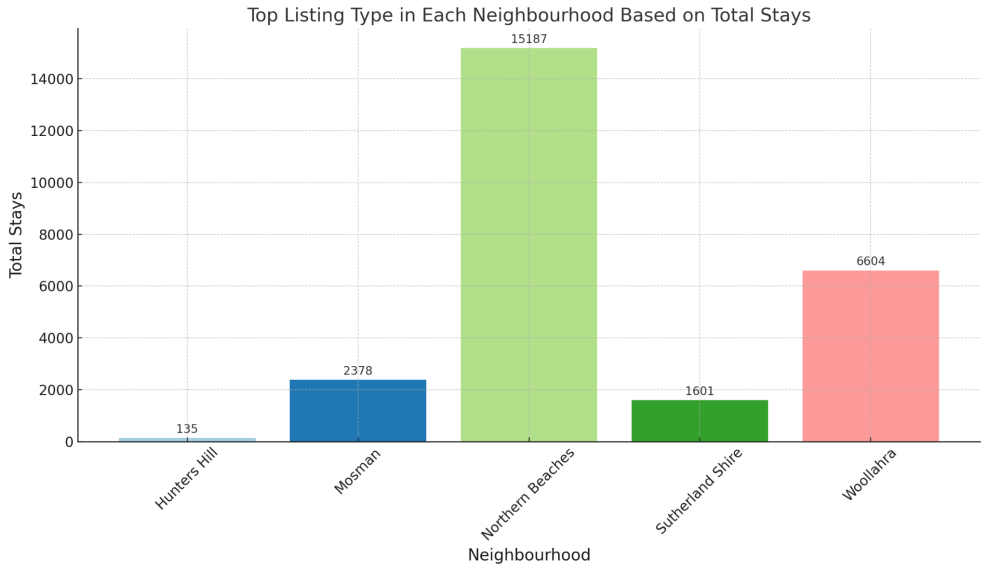
- Property Type: Apartment
- Room Type: Entire home/apt
- Accommodates: 4 people
- Total Stays: 15187

Sutherland Shire:

- Property Type: Apartment
- Room Type: Entire home/apt
- Accommodates: 4 people
- Total Stays: 1601

Woollahra:

- Property Type: Apartment
- Room Type: Entire home/apt
- Accommodates: 2 people
- Total Stays: 6604



The visual representation clearly indicated the dominance of Northern Beaches as a preferred neighbourhood. With its top listing type, Northern Beaches has significantly higher stays, marking it as a prime hotspot among the top 5 neighbourhoods. Woollahra and Mosman, too, have carved out their appeal, registering a considerable number of stays for their top listings. Conversely, Hunters Hill, despite its revenue-based ranking in the top 5, observed fewer stays for its prime listing. The visualization underscored a uniform preference across all neighbourhoods: apartments categorized as "Entire home/apt" emerged as the top choice. This highlights the current trend where guests are gravitating towards listings that offer the complete apartment experience, replete with privacy and all essential amenities.

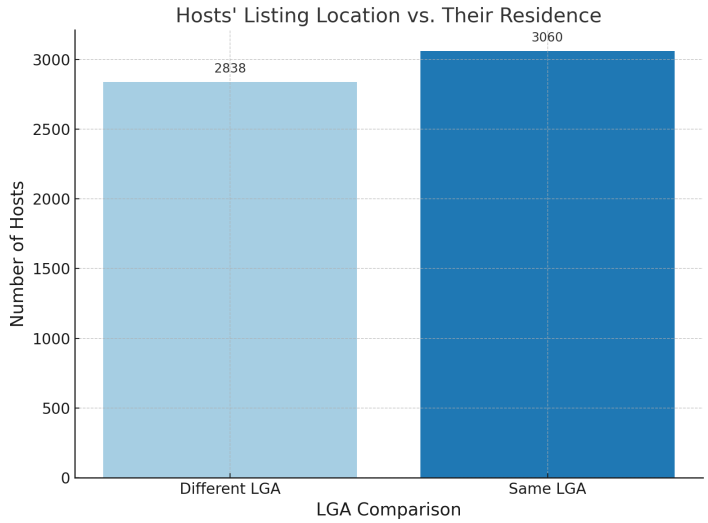
c. Hosts' Preference for Multiple Listings:

Method:
The research starts with identifying hosts who have multiple listings. This is achieved using the MultipleListingsHosts Common Table Expression (CTE). Here, the database is queried to select host_id from the public_warehouse.facts_listings table. The query groups the results by host_id and filters out hosts that have only a single listing, focusing exclusively on those with more than one. The next CTE, HostNeighbourhoodComparison and then establishes a comparison between where a host's listing is located (listing_lga_code) and where the host actually resides (host_lga_code). It joins the results of the previous CTE with the public_warehouse.facts_listings table using the host_id as the key.

Finally, the main query classifies the results into two categories: ‘LGA' where the listing's location matches the host's living location and 'Different LGA' where the listing's location is different from where the host lives. The results are then grouped by these categories, and the number of distinct hosts in each category is counted.

	lga_comparison	host_count
1	Different LGA	2,838
2	Same LGA	3,060

Result:



Same LGA: 3060
Different LGA: 2838

From the visualization, we observe that 3,060 hosts have their listings in the same LGA as their residence, while 2,838 hosts have their listings in a different LGA. This indicates a slightly higher inclination among hosts with multiple listings to list properties within their own LGA.
In conclusion, hosts with multiple listings seem more inclined to have their listings in the same LGA as where they live, although the difference between the two categories is not stark. This could be attributed to the convenience, familiarity, and ease of management when the property is closer to the host's own residence.

d. Estimated Revenue vs Mortgage Repayment:

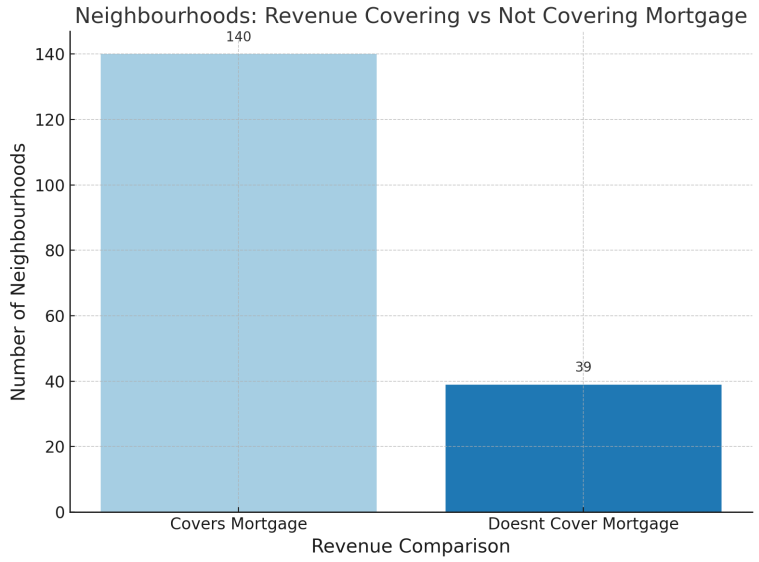
Method:
The research starts with using AnnualRevenue CTE, which computes the total estimated revenue for each host_neighbourhood_lga over the past year. It aggregates the estimated_revenue from the public_datamart.dm_host_neighbourhood table and groups the results by host_neighbourhood_lga.
Subsequently, the LGA_Mapping CTE links this revenue data with the public_warehouse.dim_lga table to map the neighbourhood to its corresponding LGA code. This CTE also serves to convert the host_neighbourhood_lga to uppercase for compatibility in the joining process.
The main query then associates the revenue data from the LGA_Mapping CTE with the median monthly mortgage repayment details from the public_warehouse.dim_census table. The mortgage repayment is annualized by multiplying the median monthly mortgage

repayment by 12. A comparison is then drawn between the annualized revenue and the annualized median mortgage repayment to determine if the revenue covers the mortgage. The results are categorized as either 'Covers Mortgage' or 'Doesn't Cover Mortgage'.

	host_neighbourhood_lga	total_estimated_revenue_per_lga	annual_mortgage_repayment	revenue_comparison
1	ABBOTSFORD	97,879	30,000	Covers Mortgage
2	ALEXANDRIA	547,260	29,988	Covers Mortgage
3	ALLAWAH	27,198	26,004	Covers Mortgage
4	ANNANDALE	1,169,284	31,200	Covers Mortgage
5	ARNCLIFFE	303,769	28,800	Covers Mortgage
6	ARTARMON	99,792	34,524	Covers Mortgage
7	ASHBURY	6,700	24,000	Doesnt Cover Mortgage
8	ASHFIELD	220,858	31,200	Covers Mortgage
9	AUBURN	35,848	26,004	Covers Mortgage
10	AVALON BEACH	2,885,186	33,600	Covers Mortgage
11	BALGOWLAH	2,121,877	33,600	Covers Mortgage
12	BALMAIN	5,375,951	31,200	Covers Mortgage
13	BALMORAL	3,261	20,796	Doesnt Cover Mortgage
14	BANKSIA	24,225	28,800	Doesnt Cover Mortgage
15	BANKSTOWN	170,518	24,000	Covers Mortgage
16	BARDWELL PARK	4,616	28,800	Doesnt Cover Mortgage
17	BARDWELL VALLEY	9,300	28,800	Doesnt Cover Mortgage
18	BEACONSFIELD	146,169	29,988	Covers Mortgage
19	BELFIELD	23,100	24,000	Doesnt Cover Mortgage
20	BELLEVUE HILL	2,974,913	38,400	Covers Mortgage
21	BELMORE	26,074	24,000	Covers Mortgage
22	BERALA	56,890	24,000	Covers Mortgage
23	BEVERLY HILLS	50,672	24,000	Covers Mortgage
24	BEXLEY	210,240	28,800	Covers Mortgage
25	BEXLEY NORTH	7,556	28,800	Doesnt Cover Mortgage
26	BIRCHGROVE	240	31,200	Doesnt Cover Mortgage
27	BLAKEHURST	30,000	26,004	Covers Mortgage

Result:

The majority of hosts with a unique listing can cover the annualized median mortgage repayment of their neighbourhood with their estimated listing revenue.



From the visualization:

- ➔ Covers Mortgage: In 140 neighbourhoods, hosts with a unique listing have their estimated revenue covering the annual mortgage repayment. This showcases that in these areas, hosts can potentially offset their mortgage expenses with their Airbnb earnings.
- ➔ Doesn't Cover Mortgage: In 39 neighbourhoods, the estimated revenue doesn't meet the annual mortgage repayment. For hosts in these areas, relying solely on Airbnb revenue may not be sufficient to handle their mortgage costs.

The significant difference between the two categories underscores the fact that in a majority of the neighbourhoods, Airbnb can be a viable financial venture for hosts with unique listings, at least in terms of offsetting mortgage costs.

Conclusion

In this research, the primary aim was to construct a robust, production-ready data pipeline, leveraging the capabilities of Airflow and Postgres via DBeaver. The focus was on meticulously processing and refining data from two distinct datasets. These datasets underwent transformation within a data warehouse using dbt in the VS Code environment and then transitioned into a data mart designed for in-depth analytics.

During the investigation, significant insights emerged regarding Sydney's Airbnb landscape. By comparing Airbnb data against Census demographics, the research unveiled intricate patterns and relationships potentially influencing the performance of Airbnb listings.

Conclusively, this project underscores the transformative potential of strategic data analysis. It emphasizes that with the right methodologies and tools, unprocessed data can be molded into invaluable insights, paving the way for informed decision making and strategic planning.