Contents

[I. **Part 0. Overview of the project** 2](#_Toc150241121)

[II. **Part 1. Use Airflow to load raw data into Postgres** 2](#_Toc150241122)

[III. **Part 2. Design a data warehouse using dbt** 4](#_Toc150241123)

[IV. **Part 3. Ad-hoc analysis** 8](#_Toc150241124)

1. **Part 0. Overview of the project:**

The primary goal of this data engineering assignment is to build production-ready data pipelines with Apache Airflow, create a data warehouse using dbt, and perform ad-hoc analysis on two different datasets. The project focuses on working with Airbnb data for Sydney and the 2016 Australian Census data to derive insights and address specific business questions. The dataset represents Airbnb’s marketing platform connecting hosts and guests for property rentals. The project focuses on the Sydney region and analyses data from May 2020 to April 2021.

1. **Part 1. Use Airflow to load raw data into Postgres**
2. Upload the dataset into the AirFlow storage bucket

A screenshot of a computer

Description automatically generated

1. Create a raw schema on Postgres and the relevant raw tables which will contain the raw data using DBeaver.

First I will create schema schema in Dbeaver:

A screenshot of a computer

Description automatically generated

It includes four schemas, namely raw, staging, warehouse, datamart. Raw layer will store raw tables and snapshot of dimensions. Staging layer is where data begins to undergo basic cleaning, transformations, and standardization. Warehouse Layer is the heart of the data warehouse, where data is organized. It includes both dimension tables and fact tables that are optimized for analytical querying. Finally, Datamart Layer is where data is presented in a format tailored to specific business needs. It consists of answers of business questions, which are created to support specific analytical queries and reporting.

A screenshot of a computer

Description automatically generated

In raw schema, I will create five blank table, namely census\_g01, census\_g02, listings, nsw\_lga\_code, nsw\_lga\_suburb. In next step data will be imported to those table from google cloud storage.

1. Create an one-off Airflow Dag (set the schedule\_interval to None) which will read the data from the storage bucket and load the raw data into the raw schema.

First I will upload data to data folder in the bucket.

A screenshot of a computer

Description automatically generated

Next, due to the limited memory and the speed of airflow, I will create 5 different dags for importing data from storage bucket to dbeaver

A screenshot of a computer

Description automatically generated

Then airflow is triggered to import data to dbeaver

A screenshot of a computer

Description automatically generated

1. **Part 2. Design a data warehouse using dbt**
2. Design the architecture of a data warehouse on Postgres with 4 layers
3. Raw: contains the raw tables + snapshots of dimensions with strategy based on timestamp

A screenshot of a computer

Description automatically generated

Raw layer contains raw tables, which are imported from google storage and its snapshot of dimensions with strategy based on timestamp. In the Raw layer, only datetime format data is handled, and it is imported without any transformations.

1. Staging: Cleaning/transformations and renaming of raw/snapshot data. A screenshot of a computer

   Description automatically generated

In the "Staging" layer, data undergoes cleaning (for example, missing value), transformations, and renaming processes, specifically focusing on raw and snapshot data. This layer is responsible for preparing the data for further downstream analysis and reporting by ensuring it is standardized, cleaned of any inconsistencies.

1. Warehouse: Star schema with dimensions and fact tables. A screenshot of a computer

   Description automatically generated

In the "Warehouse" layer, a star schema is implemented, comprising dimensions and fact tables. This architecture is used for structuring and organizing data, with dimensions representing the various attributes or characteristics of the data, and fact tables containing the numerical measures or metrics. This schema design is commonly employed for efficient querying and analysis, making it easier to retrieve insights from the data.

1. Datamart : This is where the answers to the following questions will live. Needs to be materialised as views.

A screenshot of a computer

Description automatically generated

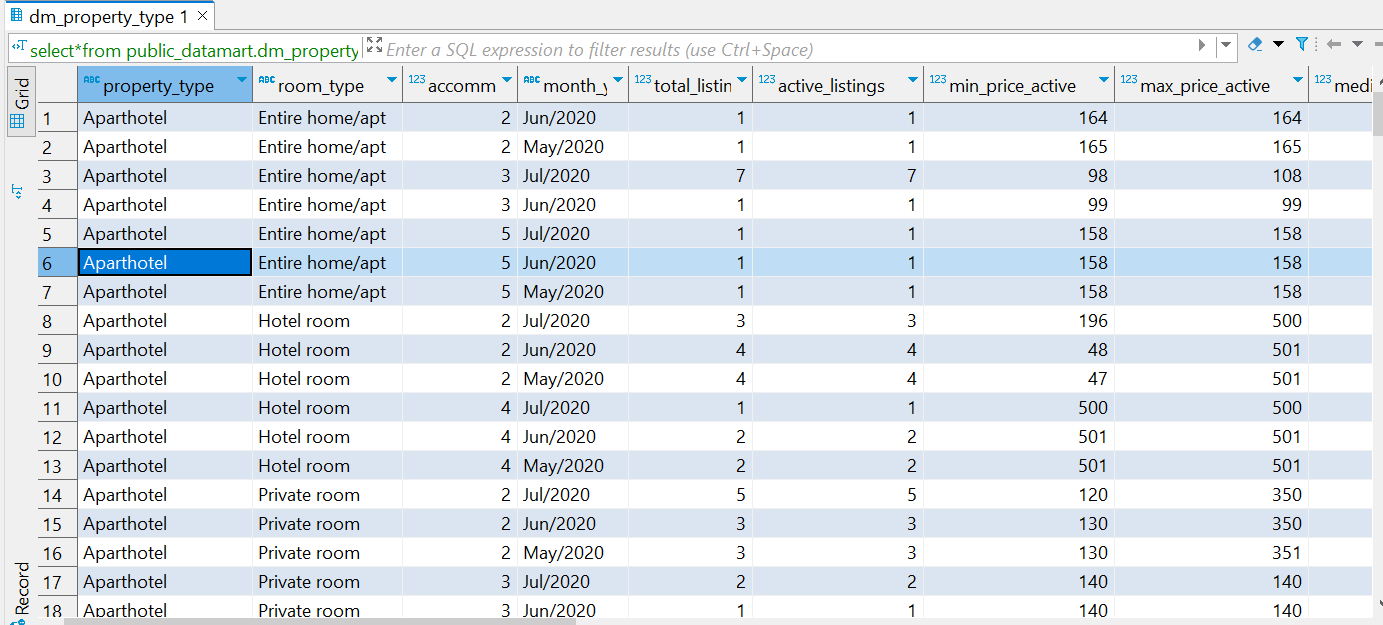
In the "Datamart" layer, answers to specific questions are stored. The data is materialized as views. This layer will provide the necessary data and insights for any analytical requirements.

+ Per “listing\_neighbourhood” and “month/year:

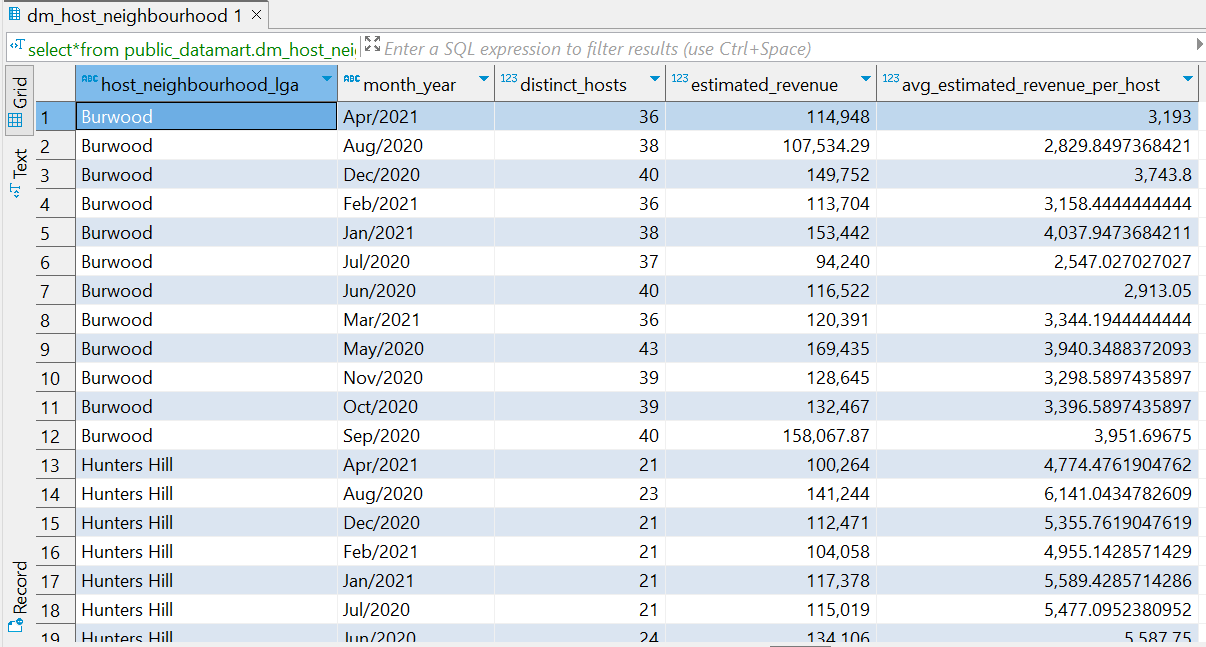
A screenshot of a computer

Description automatically generated

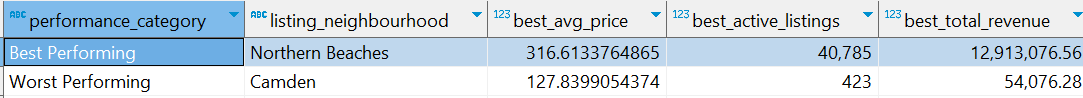
+ Per “property\_type”, “room\_type” ,“accommodates” and “month/year” :



+ Per “host\_neighbourhood\_lga” which is “host\_neighbourhood” transformed to an LGA (e.g host\_neighbourhood = 'Bondi' then you need to create host\_neighbourhood\_lga = 'Waverley') and “month/year”



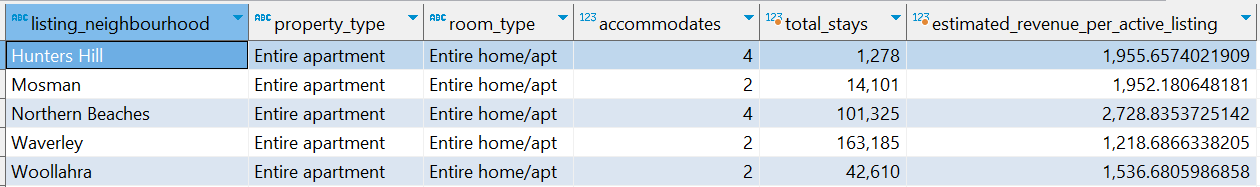
1. **Part 3. Ad-hoc analysis**
2. What are the main differences from a population point of view (i.g. higher population of under 30s) between the best performing “listing\_neighbourhood” and the worst (in terms of estimated revenue per active listings) over the last 12 months?



The key differences between these two neighborhoods are related to the average price per listing, the number of active listings, and the total revenue generated.

Northern Beaches has a significantly higher average price per listing compared to Camden, which indicates that the listings in this neighborhood are priced higher on average. Northern Beaches also has a much larger number of active listings compared to Camden, suggesting that there are more properties available for rent in this neighborhood. The total revenue generated in Northern Beaches is substantially higher than in Camden, indicating that Northern Beaches is generating much more revenue from its active listings. The differences may be attributed to several factors, including property types, demand, or location.

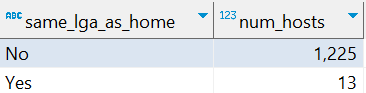
1. What will be the best type of listing (property type, room type and accommodates for) for the top 5 “listing\_neighbourhood” (in terms of estimated revenue per active listing) to have the highest number of stays?



Northern Beaches stands out as the top-performing neighborhood in terms of estimated revenue per active listing, with a significantly higher value of $2,728.84, driven by Entire apartments accommodating 4 guests. Mosman also performs well with a value of $1,952.18 for Entire apartments accommodating 2 guests.

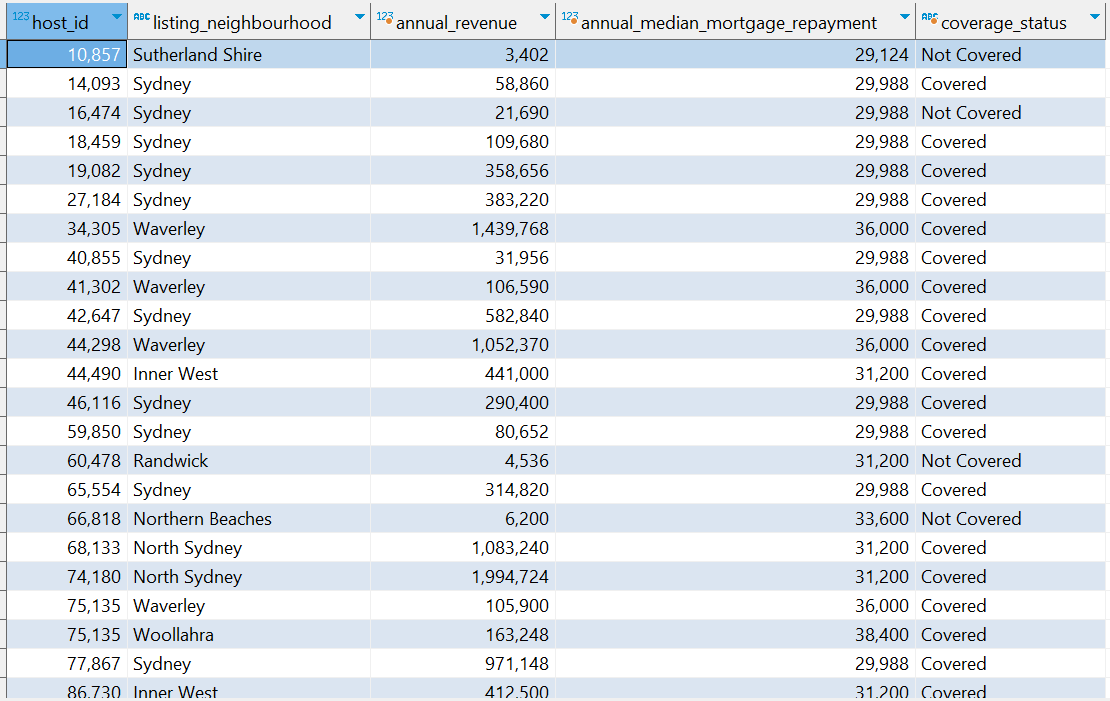
Waverley and Woollahra have lower estimated revenue per active listing compared to the top two neighborhoods, with values of $1,218.69 and $1,536.68, respectively. For both Mosman and Woollahra, Entire apartments accommodating 2 guests are the preferred listing types for maximizing revenue. Another striking point is that Hunters Hill stands out for its high number of total stays relative to other neighborhoods.

1. Do hosts with multiple listings are more inclined to have their listings in the same LGA as where they live?



The result indicates that a significant number of hosts (1225) with multiple listings have not chosen to list all their properties in the same LGA as their residence. This suggests that that most hosts with multiple listings are not limited by the location of their residence when choosing where to list their properties. They are more inclined to diversify their listings across different neighborhoods and LGAs, indicating a degree of flexibility in their listing locations.

1. For hosts with a unique listing, does their estimated revenue over the last 12 months can cover the annualised median mortgage repayment of their listing’s “listing\_neighbourhood”?



The table below summarizes basic statistics for above table:

|  |  |  |  |
| --- | --- | --- | --- |
| Row Labels | Count of coverage\_status | Average of annual\_revenue | Average of annual\_median\_  mortgage\_repayment |
| Covered | 16,050 | 564,821 | 31,844 |
| Not Covered | 3,493 | 9,059 | 31,177 |

Hosts who generate a significantly higher average annual revenue comfortably cover their mortgage repayments. On the other hand, hosts in the "Not Covered" category struggle to generate sufficient income from their listings to cover their mortgage obligations, and their numbers are smaller. It highlights the importance of financial planning and understanding the income potential of Airbnb listings in their specific neighborhoods.