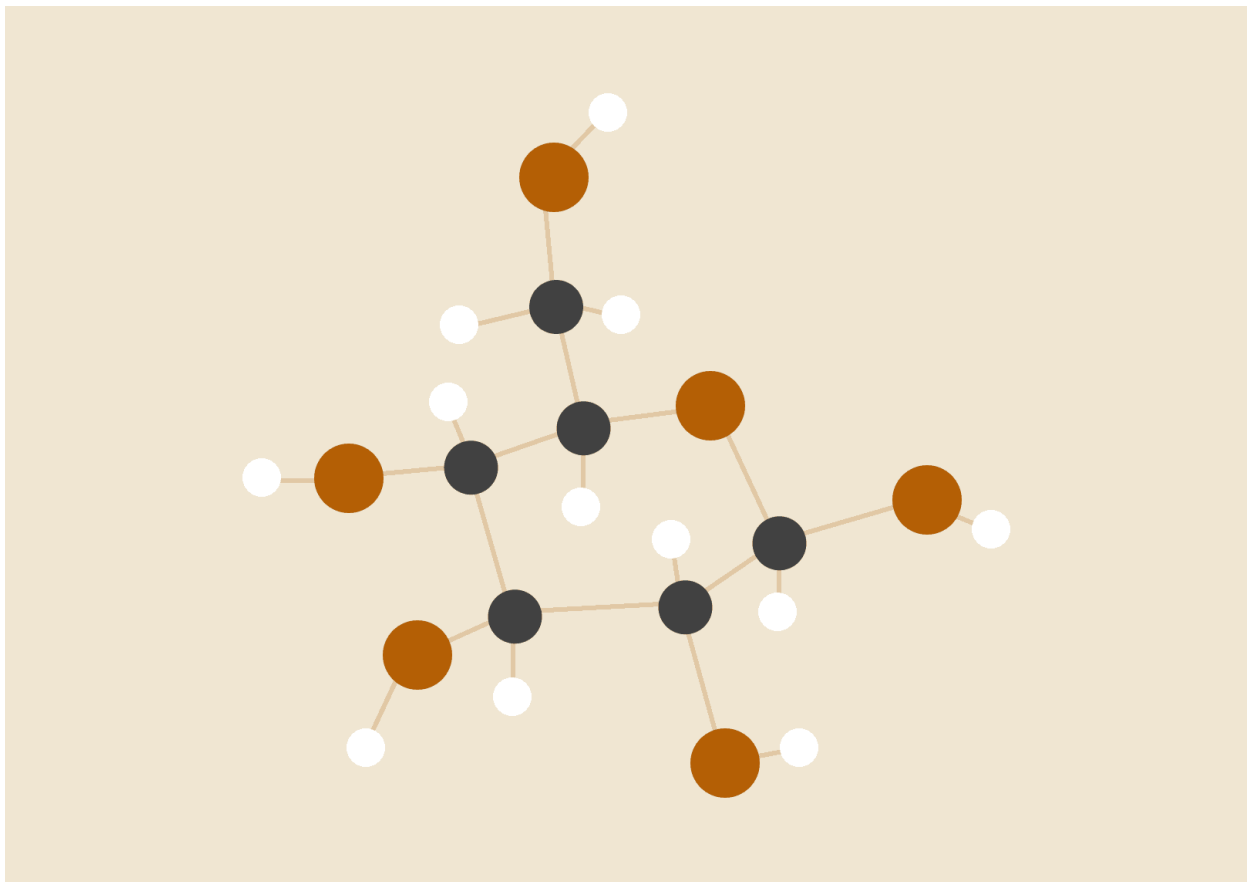


Airbnb Analysis: Leveraging Data Insights for Strategic Decision-Making



Trung Kien Hoang

07.11.2023

Table of content:

INTRODUCTION.....	2
Steps.....	2
Problem Identification.....	5
Business Question.....	5
Question 1.....	5
Question 2.....	6
Question 3.....	7
Question 4.....	7
Conclusion.....	8

INTRODUCTION

The aim of this task is to construct data pipelines that are ready for deployment and can be used in a production environment using Airflow and Google Cloud Platform (Cloud Composer and SQL instance). The primary objective of this project is to process and clean the provided datasets obtained from Airbnb and the government's census. The second objective of this project is to load the valuable data into a data warehouse using ELT pipelines and a data mart for the purpose of analysis. This project examines the host characteristics, including the host's name or address, as well as the listing elements such as room type, property type, and accommodation. Additionally, the project analyzes the reviews features, including the review score. Other significant factors, such as pricing, duration of stay, revenue, age demographics, and housing arrangements, are also utilized to analyze the activities, incomes and mortgage of hosts using the Airbnb platform.

Steps

1. **Data loading:** The first phase of this project entails importing unprocessed data into Postgres using Airflow. The provided datasets are uploaded to the AirFlow storage bucket and subsequently imported into the raw schema on Postgres. The raw schema consists of the necessary raw tables that store the raw data, utilizing DBeaver.

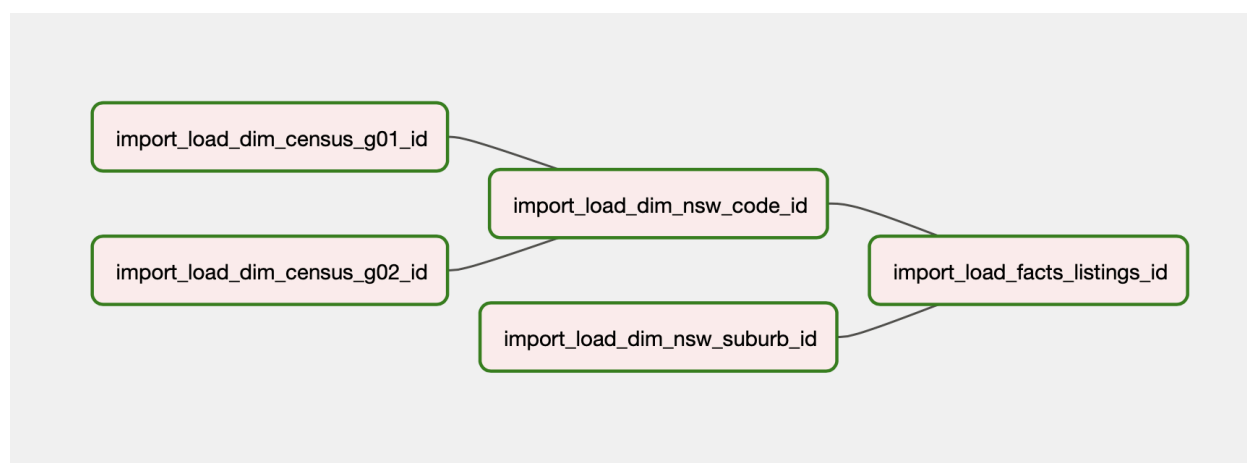


Figure 1: The graph tree in importing data to raw schema.

The diagram above illustrates a DAG structure that represents the logic of operators. The tasks "import_load_dim_census_g01_task" and "import_load_dim_census_g02_task" can

be executed simultaneously for the dependency structure. After both tasks have successfully finished, the tasks "import_load_dim_nsw_code_task" and "import_load_dim_nsw_suburb_task" will run concurrently. Ultimately, once both sets of activities are finished, the import_load_facts_listings_task will be executed.

The import function established all potential columns and inserted each record into a newly formed schema. In addition, the schema has a limitation on the number of pages that can be uploaded due to the huge file size. Therefore, the facts files are read and imported separately.

2. **Data warehouse:** The second phased of this project is to design the architecture of a data warehouse on Postgres with 4 layers:
 - 2.1. **Raw:** It contains the raw tables that store unprocessed data. It also includes snapshots of dimensions that specify properties such as the type of room, type of property, and host entities.
 - 2.2. **Staging:** It contains transformation, cleaning and rename processes of the raw and snapshot data. There are 8 stage files are created in this project:
 - 2.2.1. G01_stg: stores census data, including the LGA code, as well as statistical information such as the population count for different age groups. A transformation process is performed to change the LGA code into a numerical representation. An unknown record is also defined for future works.
 - 2.2.2. G02_stg: contains census data, including the LGA code, as well as statistics information such as the monthly median mortgage repayment. A transformation process is performed to change the LGA code into a numerical representation. An unknown record is also defined for future works.
 - 2.2.3. Lga_stg: stores NSW LGA data, including the LGA code and LGA name. An uppercase transformation method is executed to convert all LGA names into capital letters.
 - 2.2.4. Suburb_stg: stores NSW LGA data, including the LGA name and LGA suburb.
 - 2.2.5. Host_stg: contains the host attributes, such as host identification, host name, host address, and superhost status. The cleaning technique involves replacing null entries with the value "unknown" and applying an uppercase translation method to turn all host addresses into capital letters. An unknown record is also established for future works.

- 2.2.6. Room_stg: holds information about rooms, including their types and identification numbers. The renaming process is carried out in order to change the label of a room. An unknown record is also defined for future works.
- 2.2.7. Property_stg: holds information about property, including their types and identification numbers. The renaming process is carried out in order to change the label of a property. An unknown record is also defined for future works.
- 2.2.8. Listing_stg: contains data related to listings, including the listing identification, scraped date, address, accommodation details, price, and length of stay. It also includes attributes related to reviews, such as the number of reviews, rating scores, accuracy scores, cleanliness scores, check-in scores, communication scores, and value scores. In addition, it includes dimension identification for the host, room, and property. The cleaning methodology entails substituting null entries with the value "unknown" and implementing an uppercase translation algorithm to convert all list addresses into capital letters. Additional transformations are executed to convert the data type of certain columns to integer and date formats.
- 2.3. **Warehouse:** star schema consisting of seven dimension tables and one fact table. Dimension tables provide a description of the different properties or characteristics of the data, whereas the fact table describes the measurable aspects and is used to establish connections with other datasets. In this project, the fact table is combined with dimensions such as host, property, room and LGA.
- 2.4. **Datamart:** The final layer is usually optimised specifically for reporting and analytical purposes.
 - 2.4.1. The first data mart provides comprehensive information on active listings in each combination of neighbourhood of list and month in year, encompassing key statistical measures such as the minimum, maximum, median, and average prices, the count of distinct hosts, the proportion of Superhosts, the average rating for current listings, percentage fluctuations in active and inactive listings, the total number of stays, and the average projected revenue per current listing.
 - 2.4.2. The second data mart provides comprehensive information on active listings in each combination of property type, room type,

accommodates and month in year, encompassing key statistical measures such as the minimum, maximum, median, and average prices, the count of distinct hosts, the proportion of Superhosts, the average rating for current listings, percentage fluctuations in active and inactive listings, the total number of stays, and the average projected revenue per current listing.

- 2.4.3. The third data mart provides a thorough overview of host information categorised by host suburb and month in a given year. It includes important statistical metrics such as the number of unique hosts, estimated revenue, and estimated revenue per unique host.

Problem Identification

The study faced technical difficulties due to the database timing out for connection. The fundamental cause was the intermittent change in the home IP address. Consequently, it is necessary to refresh the home network connection on the Google Cloud Platform.

Alternatively, there are some ethical considerations regarding this research due to the inclusion of a dataset including personal information about hosts, including their names and addresses. Consequently, this may give rise to problems around data privacy.

Safeguarding this data is vital to thwart unauthorized access and data breaches. The dataset may be susceptible to unauthorized access, perhaps leading to malicious activities such as phone scams or mail scams. Moreover, the purpose of this study is to analyze the income generated from listing and hosting. Therefore, using this study as a pricing guideline may result in unfairness to other hosts.

Business Question

Question 1

What are the main differences from a population point of view between the best performing “listing_neighbourhood” and the worst (in terms of estimated revenue per active listings) over the last 12 months?

	ASC listing_neighbourhood	123 total_revenues	123 percentage_0_34	123 percentage_35_64	123 percentage_over_65
1	CAMDEN	1,105,395	52.1734127694	37.2932061674	10.546165844
2	NORTHERN BEACHES	439,134,554	42.0728572671	41.1316919621	16.791891742

Screenshot 1: The most exceptional and least desirable neighbourhood for listing performances.

The first screenshot illustrates that over a span of 12 months, the cumulative revenues for all listings in Camden amount to approximately 1 million dollars, whereas listings on the Northern Beaches are 400 times more. Furthermore, Camden has a significant proportion of individuals under the age of 35, whilst the Northern Beaches exhibit a more even distribution between the age groups under 35 and between 35 and 64. Thus, the presence of a larger young population and a smaller senior population may result in decreased listing revenues. In addition, it is advisable for new hosts to commence their listing in the northern beach area.

Question 2

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?

	ASC listing_neighbourhood	ASC property_description	ASC room_description	123 accommodates	123 total_revenue
1	NORTHERN BEACHES	Private room in guest suite	Private room	3	439,134,554
2	RANDWICK	Loft	Entire home/apt	2	155,351,375
3	SYDNEY	Villa	Shared room	1	383,078,635
4	WAVERLEY	House	Shared room	2	323,211,378
5	WOOLLAHRA	Bed and breakfast	Private room	2	117,354,701

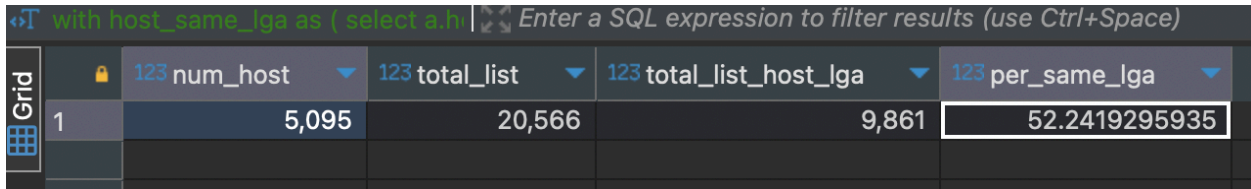
Screenshot 2: Top 5 highest total number of revenue per active listing.

The second screenshot demonstrates that the Northern Beaches has the greatest total revenue per active listing, followed by Sydney, Waverley, Randwick, and Woollahra. The most popular combinations of property type, room type, and number of guests that have the biggest number of stays in the Northern beaches area are guest private rooms, private rooms, and accommodations for three people, respectively. In the Sydney area, the most popular combinations of property type, room type, and number of guests that have the highest number of stays are villas, shared rooms, and accommodations for one person, respectively. In the Waverley region, the property type that has the most number of stays is a house, the room type with the biggest number of stays is a shared room, and

the number of accommodations that is most commonly chosen is two. In the Randwick area, the most favourable options for property type, room type, and number of accommodates that yield the highest frequency of stays are loft, complete home or flat, and a capacity of two individuals, respectively. The most popular combinations of property type, room type, and number of guests in the Woollahra region are Bed and breakfast, private room, and accommodation for two people, respectively.

Question 3

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



The screenshot shows a SQL query interface with a table of results. The query is: `with host_same_lga as (select a.h`. The table has four columns: `num_host`, `total_list`, `total_list_host_lga`, and `per_same_lga`. The first row of data shows values 5,095, 20,566, 9,861, and 52.2419295935 respectively.

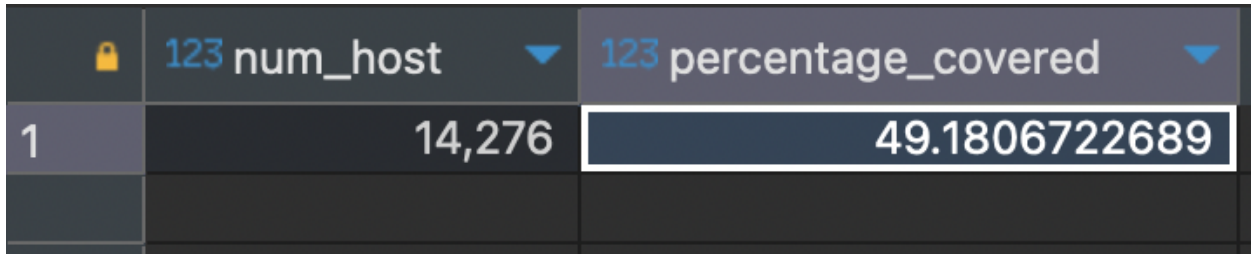
	123 num_host	123 total_list	123 total_list_host_lga	123 per_same_lga
1	5,095	20,566	9,861	52.2419295935

Screenshot 3: Total number host, list and their relation information.

The third screenshot indicates that out of a total of 5095 hosts who possess multiple listings, there are 52 percent of them who have their listings located in the same Local Government Area (LGA) as their place of residence.

Question 4

For hosts with a unique listing, does their estimated revenue over the last 12 months can cover the annualized median mortgage repayment of their listing’s “listing_neighbourhood”?



The screenshot shows a SQL query interface with a table of results. The query is: `with host_unique_listing as (select a.h`. The table has two columns: `num_host` and `percentage_covered`. The first row of data shows values 14,276 and 49.1806722689 respectively.

	123 num_host	123 percentage_covered
1	14,276	49.1806722689

Screenshot 4: Total number of hosts and the percentage of hosts can cover the annualized median mortgage repayment.

According to snapshot 4, among the 14276 hosts who have only one listing, only 49.2

percent of them have earned enough money from their listing in the past 12 months to cover the annualised median mortgage repayment. A single listing by the host results in a negative income trend.

Conclusion

To summarize, the objective of the project is to construct a data pipeline that is ready for production in order to conduct valuable analysis on the Airbnb dataset. A recommended approach for new hosts is to begin their first listing in the Northern Beaches of Sydney, namely in the areas of Waverley, Randwick, or Woollahra. The ideal property type for this listing would be guest private rooms, with the room type specified as private rooms, and the accommodations should be suitable for three people. On the contrary, it is not recommended to start the listing at Camden.