Airbnb Listings Data Pipelines

Name: Wenying Wu
Date: 17/05/2022

PROJECT OBJECTIVE

This project is to build production-ready ELT pipelines using Airflow for Airbnb listings. The ELT pipeline includes processing and cleaning two provided datasets and loading the data into a data warehouse and data mart for analysis.

PROJECT DESCRIPTION

Datasets Provided

One of the provided datasets is the 12 months of Airbnb listing data in Sydney, the other is the General Community Profile Pack from the 2016 census at the LGA level. A CSV file with the LGA name and corresponding code is also provided for the purpose of joining two datasets.

Project Overview

This project aims to build a data mart and gain insights into the Airbnb business in the Sydney area in combination with census data. The ELT pipeline will be built by python scripts, SQL scripts, Airflow, GCP Cloud Composer, and Snowflake. The raw data will be extracted from the provided CSV files, loaded into the Snowflake database, then transformed into star schema in the data warehouse. After that, a datamart will be designed and populated through an ETL pipeline from the data warehouse. The last part of this project is to perform some ad-hoc analysis that will be presented in a later section.

It is expected that this project will satisfy all the requirements listed in the project brief document.

STEPS OF THE PROJECT

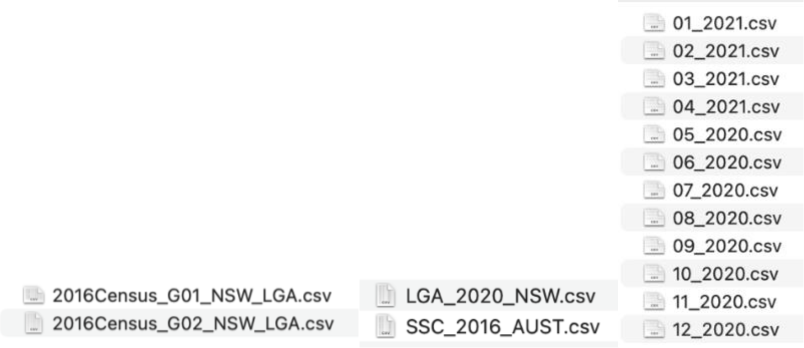
1. Data understanding

There are three groups of datasets. Please see the appendix for detailed data sources.

1. Airbnb listing data:
   - 12 months of Airbnb listing data in Sydney from May 2020 to April 2021
   - The data was provided by Inside Airbnb in 12 CSV files, one for each month
   - The Airbnb data contain data about the listing information of the property listing on the Airbnb platform for each month, including room information (bedroom number, bathroom number, location, price, etc.) and host information (name, location, etc.).
2. Census data:
   - Tables G01 (“Selected Person Characteristics by Sex”) and G02 (“Selected Medians and Averages”) of the General Community Profile Pack from the 2016 census at the LGA level of New South Wales/
   - Tables G01 and G02 contain census data of each LGA regarding population, income, education, etc.
   - The data was provided by the Australian Bureau of Statistics.

3. Location data (external):

- Local Government Area (LGA) data, “New South Wales Local Government Area ASGS Edition 2020 in .csv Format”
- State Suburb (SSC) data “State Suburbs ASGS Edition 2016 in .csv Format” contains the information of the suburb names and codes under each LGA.
    - The two data sets are sourced from the Australian Bureau of Statistics
    - The name and code of lga and suburb from these datasets would be used to map the location of listing properties and Airbnb hosts that appeared in the Airbnb data. The given dataset lacks the mapping of suburbs to LGA, so external datasets are used.



2016Census_G01_NSW_LGA.csv
2016Census_G02_NSW_LGA.csv

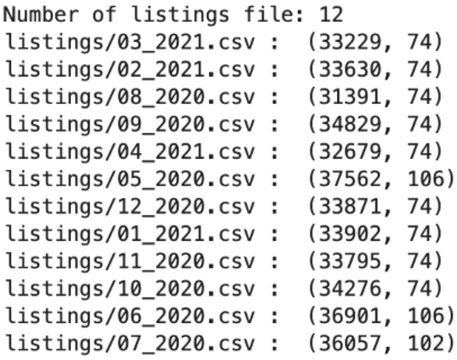
LGA_2020_NSW.csv
SSC_2016_AUST.csv

01_2021.csv
02_2021.csv
03_2021.csv
04_2021.csv
05_2020.csv
06_2020.csv
07_2020.csv
08_2020.csv
09_2020.csv
10_2020.csv
11_2020.csv
12_2020.csv

Figure 1. Datasets Used

2. Initial data preparation

Before performing any actual work on building the ELT pipeline, it is critical to ensure all the CSV files of Airbnb listing data are consistent in the data structure. Otherwise, the pipeline will not work as expected. Python was used to view all the provided CSV files to preprocess the data. Data structure here means the columns in each CSV file. If there are columns that exist inside one file only but not in other files, it is considered redundant columns and will not be pushed into the database. On the other hand, if there are columns in all CSV files except for one or two files, these columns are deemed valuable and will be populated to the database and fill the file(s) without these columns with NULL values in these columns. There are 74 columns selected after this step, as most of the datasets have 74 columns shown in Figure 1. Details of the process are shown in “workfile_preprocess.ipynb”.



```
Number of listings file: 12
listings/03_2021.csv :  (33229, 74)
listings/02_2021.csv :  (33630, 74)
listings/08_2020.csv :  (31391, 74)
listings/09_2020.csv :  (34829, 74)
listings/04_2021.csv :  (32679, 74)
listings/05_2020.csv :  (37562, 106)
listings/12_2020.csv :  (33871, 74)
listings/01_2021.csv :  (33902, 74)
listings/11_2020.csv :  (33795, 74)
listings/10_2020.csv :  (34276, 74)
listings/06_2020.csv :  (36901, 106)
listings/07_2020.csv :  (36057, 102)
```

Figure 2. Shape of listing files

3. Design and populate a data warehouse using ELT pattern

Connections among Airflow, GCP Cloud Composer, and Snowflake are established following the official documentation. After the connections are set, two months of listing data, two of the provided census datasets and the two external datasets are uploaded into Snowflake from GCP Cloud Composer. Basically, the star schema of the data warehouse is described below:

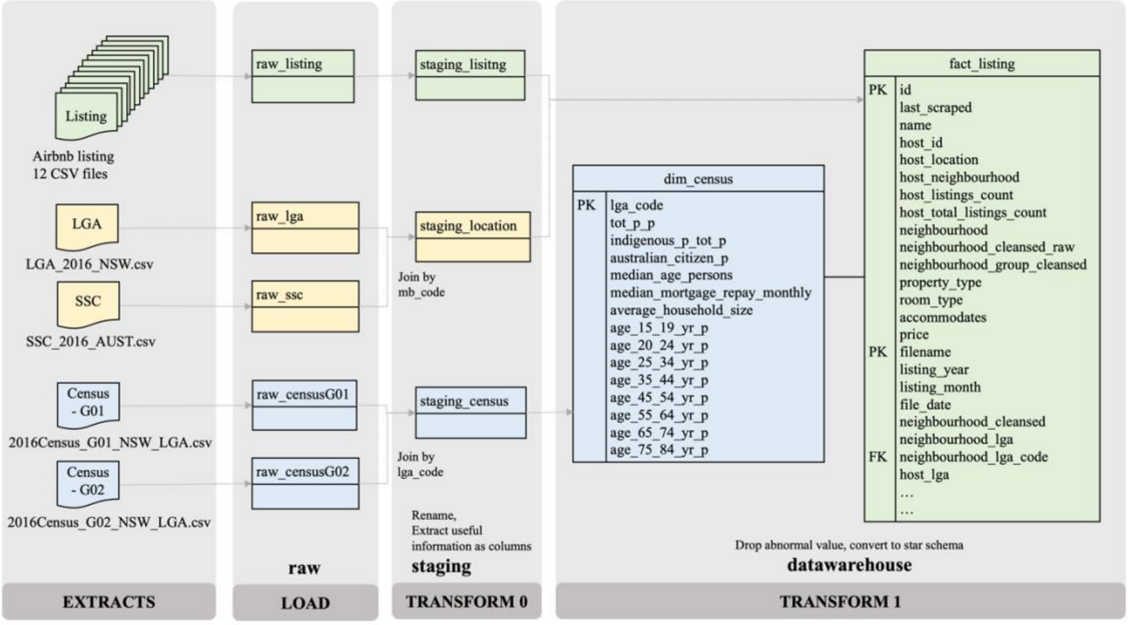


Figure 3: Overview of the project

Dimension table dim_census:

- Selected columns from the two provided census datasets which may be useful for future analysis.
- The dataset are merged by the common column.
- The primary key is lga_code.

Dimension table dim_location:

- Only selected columns from the two external datasets (LGA and SSC), including "lga_code", "lga_name" and "suburb_name".
- The primary key is lga_code.
- There are numerous meshed blocks "mb_code" under one "suburb_name". SSC table are joined with LGA table based on the same "mb_code". There are meshed blocks (locations) in the same suburbs but are belong to different LGAs. For example, some meshed blocks in the suburb New Town belong to Inner West LGA while the others belong to Sydney LGA. In this case, I map the suburb to the LGA that owns the largest area of this suburb. For example, if 60% of area of Newtown belong to Inner West LGA and 40% area of Newtown belong to Sydney LGA, then I map Newtown to Inner West LGA.

Fact table fact_listing:

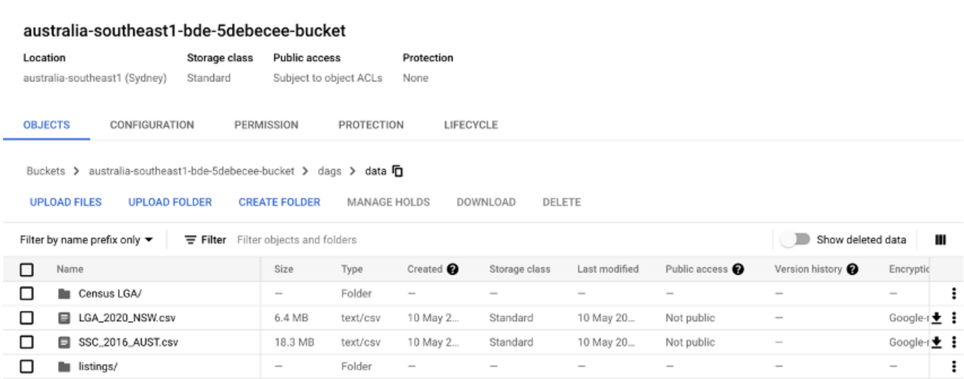
- datawarehouse.fact_listing is storing the minor details of the listing
  - 74 columns from raw data are selected. Below are the added key columns, intermediate columns are not mentioned here, details are shown in part2.sql.

    - Added “file_name” the indicated the source of data, and evaluated whether the data is valid or not, the data
    - Added “year”, and “month” for analysis purposes using “last_scaped” (The date and time this listing was scraped)
    - Only includes data if “id”, “host_id”, and “price” are not null for analysis purposes.
    - Added columns “neighbourhood_lga” and “host_lga” based on the suburb from location tables to be in line with official names in LGA data for the datamarts requirement.
- Table fact_listing should be built after dim_location because of the addition of columns “neighborhood_lga” and “host_lga”.
- The composite primary key is: “file_name”, “id”.
- The foreign key is “neighbirhood_lga_code” references datawarehouse.dim_census “lga_code”

Text cleaning was performed in staging and data warehouse tables such as extracting the necessary information from columns. For example.

- Upper case location variables to match each other
- Trim leading and trailing space and replace double space with a single space.
- Change substring like “Saint Peters” to “St Peters”
- Change “lga_code” such as “LGA10050” to “10050”
- Change “lga_name” such as “Albury (C)” to “Albury”
- Change “ssc_name” such as “St Peters (SA)” to “St Peters”
- Based on the data dictionary provided by Inside Airbnb, “neighbourhood_cleansed” is the neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. It represents the listing property’s local government area. But it contains some errors. For example, “Ashfield” is a suburb, but it appears in “neighbourhood_cleansed” which should only contain LGA names. Therefore, I created “neighbourhood_lga” by mapping “neighbourhood” to LGA names for later analysis.

After the star schema is set, all remaining listing data are uploaded to Google Cloud Platform to form the final data warehouse.



australia-southeast1-bde-5debecee-bucket

Location	Storage class	Public access	Protection
australia-southeast1 (Sydney)	Standard	Subject to object ACLs	None

Buckets > australia-southeast1-bde-5debecee-bucket > dags > data

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryptic
□	Census LGA/	–	Folder	–	–	–	–	–	–
□	LGA_2020_NSW.csv	6.4 MB	text/csv	10 May 2...	Standard	10 May 20...	Not public	–	Google-r
□	SSC_2016_AUST.csv	18.3 MB	text/csv	10 May 2...	Standard	10 May 20...	Not public	–	Google-r
□	listings/	–	Folder	–	–	–	–	–	–

Figure 4. files uploaded to Google Cloud Platform

4. Design the Datamart schema

Three tables are established in the datamart. Namely “datamart.kpi_neighbourhood_month”, “datamart.kpi_property_month” and “datamart.kpi_host_neighbourhood_month”. Each one is corresponding to 1 KPI requirement in the brief document and the “datamart.kpi_neighbourhood_month” is created using “neighbourhood_lga” (self-cleaned version of “neighbourhood_cleansed”).

Additional table “datamart.kpi_neighbourhood_month_raw” has also been created to show the original result by using “neighbourhood_cleansed” (provided column). Details are shown in “workfile_populate_data_warehouse.py” under section Query for KPI.

The process is shown below by DAG structures in Airflow.

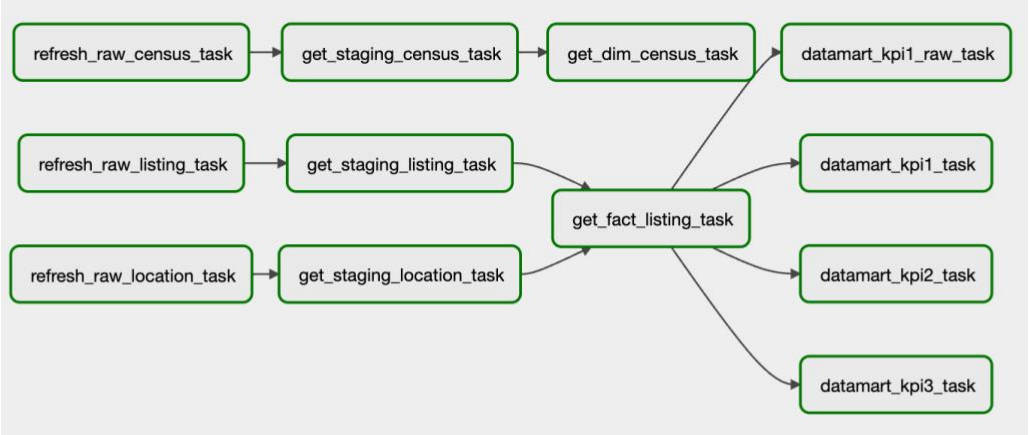


Figure 5. The DAG

AD-HOC ANALYSIS

a. What are the main differences from a population point of view (i.g. higher population of under 30s) between the best performing “neighbourhood_cleansed” and the worst (in terms of estimated revenue per active listings) over the last 12 months?

During the 12 months from May 2020 to April 2021, Mosman was the best performing neighbourhood with the highest estimated revenue per active listing, it reached $7,213.16 and it is more than seventh times then the worst performing neighbourhood LGA (Cumberland). The total population in Mosman is much less than in Cumberland, only about 13% of the population in Cumberland. Only 0.2% of the people in Mosman are indigenous, while the percentage of indigenous in Cumberland is 3 times more than in Mosman. The median age person for Mosman is 42, which is much higher than Cumberland. The percentage of people under 35 in Mosman is 22%, which is less than the percentage in Cumberland. The percentage of people age over 65 is 16% in Mosman, which is more than 1.5 times the percentage in Cumberland.

The table below shows the result from the SQL query.

Neighbourhood_lga	**Mosman**	**Cumberland**
Estimated_revenue_per_active_listings	7,213.16	1,024.49
Median_age_persons	42	32
Tot_p_p	28,475	216,079

Indigenous_p_tot_p	60	1,394
Percent_indigenous_p_tot_p	0.21%	0.65%
Australian_citizen_p	22,660	156,354
Percent_australian_citizen_p	79.58%	72.36%
Age_under_35	6,262	71,846
Percent_age_under_35	21.99%	33.25%
Age_35_64	7,478	51,037
Percent_age_35_64	26.26%	23.62%
Age_above_65	4,542	20,834
Percent_age_above_65	15.95%	9.64%
Age_above_35	12,020	71,871
Percent_over_35	42.21%	33.26%

Table 1. Result for Question 1

b. What will be the best type of listing (property type, room type and accommodates for) for the top 5 "neighbourhood_cleansed" (in terms of estimated revenue per active listing) to have the highest number of stays?

Mosman, Central Coast, Northern Beach, Wollondilly, and Eurobodalla are the top 3 listing local government area in terms of estimated revenue per active listing. The table shows the best type of listing (property type, room type and accommodates) to have the highest number of stays. As we can see from the table, a lot of types of listing in Mosman and Northern Beach are reached the highest average number of stays, 30 days. 34 types of listing in Northern Beach and 24 types of listing in Mosman reached the highest average number of stays, 30 days. As we can see from the simple graph made from Snowflake, Entire home/apt are more likely to get a high number of stays. 4 is the best maximum capacity of the listing, followed by 2, 6.

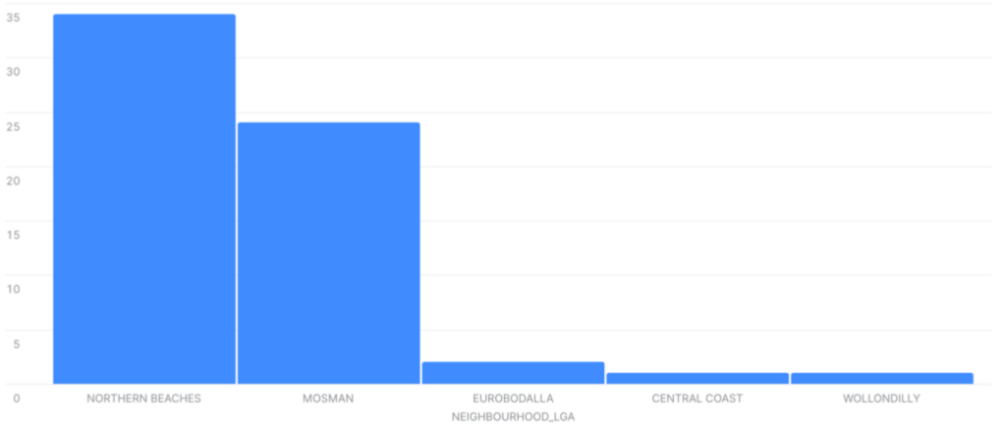


Figure 6. Plot made in Snowflake – count of LGA

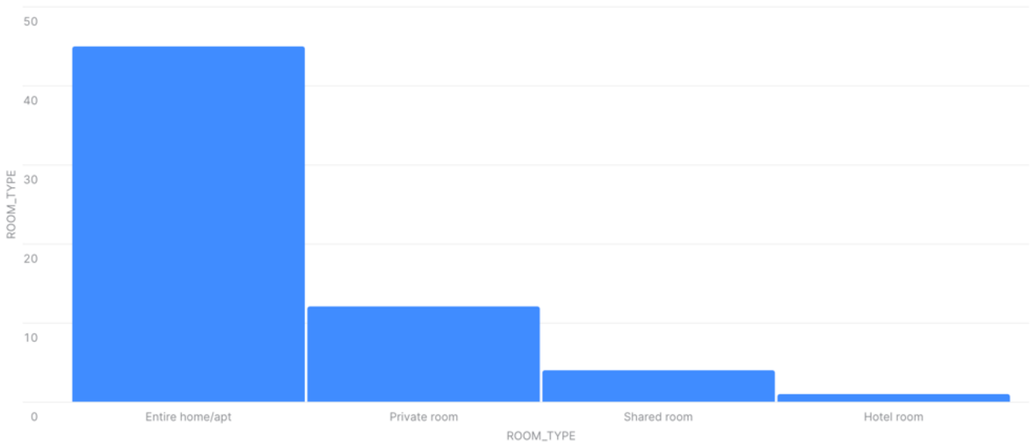


Figure 7. Plot made in Snowflake – count of room types

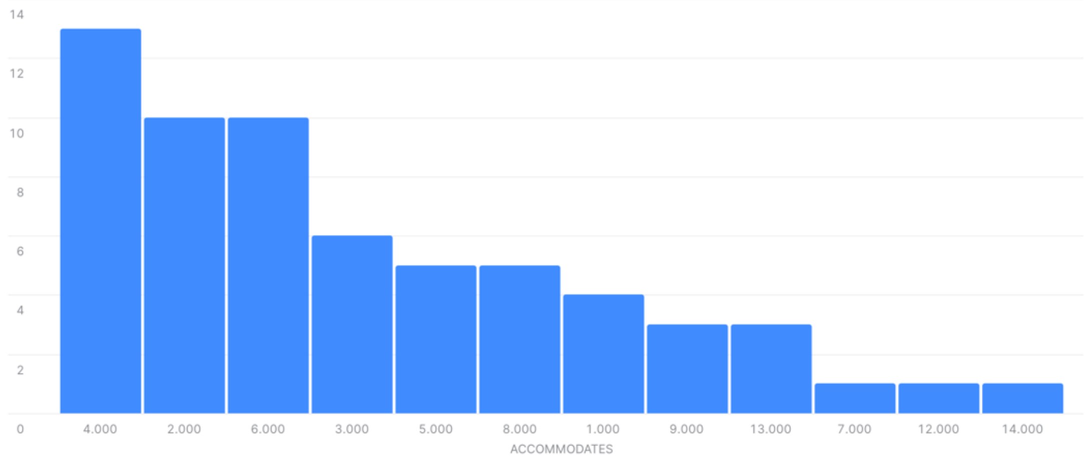


Figure 8. Plot made in Snowflake – count of accommodates

The table below shows the result from the SQL query.

NEIGHBOURHOOD_LGA	PROPERTY_TYPE	ROOM_TYPE	ACCOMMODATES	AVG_NUMBER_STAYS
MOSMAN	Apartment	Private room	4	30
MOSMAN	Apartment	Shared room	2	30
MOSMAN	Bungalow	Entire home/apt	6	30
MOSMAN	Condominium	Entire home/apt	3	30
MOSMAN	Entire apartment	Entire home/apt	1	30
MOSMAN	Entire apartment	Entire home/apt	8	30
MOSMAN	Entire bungalow	Entire home/apt	6	30
MOSMAN	Entire floor	Entire home/apt	5	30

MOSMAN	Entire guesthouse	Entire home/apt	4	30
MOSMAN	Entire house	Entire home/apt	14	30
MOSMAN	Entire townhouse	Entire home/apt	4	30
MOSMAN	Entire villa	Entire home/apt	2	30
MOSMAN	Entire villa	Entire home/apt	6	30
MOSMAN	Guest suite	Entire home/apt	8	30
MOSMAN	Guesthouse	Entire home/apt	4	30
MOSMAN	House	Entire home/apt	13	30
MOSMAN	House	Private room	6	30
MOSMAN	Private room in apartment	Private room	3	30
MOSMAN	Private room in apartment	Private room	6	30
MOSMAN	Room in boutique hotel	Hotel room	4	30
MOSMAN	Townhouse	Entire home/apt	5	30
MOSMAN	Townhouse	Private room	2	30
MOSMAN	Villa	Entire home/apt	12	30
MOSMAN	Villa	Entire home/apt	2	30
CENTRAL COAST	Entire cottage	Entire home/apt	6	20
NORTHERN BEACHES	Apartment	Entire home/apt	9	30
NORTHERN BEACHES	Apartment	Private room	3	30
NORTHERN BEACHES	Apartment	Shared room	1	30
NORTHERN BEACHES	Barn	Entire home/apt	4	30
NORTHERN BEACHES	Bungalow	Entire home/apt	5	30
NORTHERN BEACHES	Bungalow	Entire home/apt	8	30
NORTHERN BEACHES	Bungalow	Private room	2	30
NORTHERN BEACHES	Cabin	Entire home/apt	3	30
NORTHERN BEACHES	Cabin	Entire home/apt	4	30
NORTHERN BEACHES	Cabin	Entire home/apt	6	30
NORTHERN BEACHES	Condominium	Entire home/apt	4	30
NORTHERN BEACHES	Cottage	Entire home/apt	7	30

NORTHERN BEACHES	Entire apartment	Entire home/apt	9	30
NORTHERN BEACHES	Entire bungalow	Entire home/apt	5	30
NORTHERN BEACHES	Entire chalet	Entire home/apt	2	30
NORTHERN BEACHES	Entire house	Entire home/apt	13	30
NORTHERN BEACHES	Entire place	Entire home/apt	3	30
NORTHERN BEACHES	Entire place	Entire home/apt	4	30
NORTHERN BEACHES	Entire place	Entire home/apt	6	30
NORTHERN BEACHES	Entire townhouse	Entire home/apt	8	30
NORTHERN BEACHES	Guest suite	Entire home/apt	1	30
NORTHERN BEACHES	Hostel	Private room	6	30
NORTHERN BEACHES	House	Entire home/apt	13	30
NORTHERN BEACHES	House	Shared room	2	30
NORTHERN BEACHES	Loft	Entire home/apt	5	30
NORTHERN BEACHES	Other	Entire home/apt	3	30
NORTHERN BEACHES	Other	Entire home/apt	8	30
NORTHERN BEACHES	Private room	Private room	2	30
NORTHERN BEACHES	Private room in camper/rv	Private room	4	30
NORTHERN BEACHES	Shared room in apartment	Shared room	1	30
NORTHERN BEACHES	Tiny house	Entire home/apt	4	30
NORTHERN BEACHES	Townhouse	Entire home/apt	9	30
NORTHERN BEACHES	Townhouse	Private room	4	30
NORTHERN BEACHES	Yurt	Entire home/apt	2	30
WOLLONDILLY	Entire house	Entire home/apt	6	14
EUROBODALLA	Entire house	Entire home/apt	4	30
EUROBODALLA	Private room in house	Private room	2	30

Figure 2. Result for Question b

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

This question is assumed to not include records if neighbourhood LGA or host LGA is "OTHER" or "MISSING". Because there are uncertainties in these hosts and listings, some might be having a listing in their LGA while others do not. I noticed that there are many hosts having missing values in location data. Thus, in the group of hosts with multiple listings, only 5,044 hosts over 29,425 hosts are in this analysis. If the host has multiple listings, there are 32% that at least one listing is in the same LGA, and 50% of all their listings are in the same LGA.

Percentage_in_same_lga	**100%**	**50% - 99%**	**<50%**
Number_of_host_same_lga_per_range	821.00	611.00	206
Total_number_of_host_same_lga	1,638	1,638	1,638
Percentage_of_host_with_same_lga_mutiple_listings	50%	37%	13%
Total_number_of_host_with_mutiple_listings	5,044	5,044	5,044
Percentage_of_host_with_mutiple_listings	16%	12%	4%

Table 3. Result for Question c

d. 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 "neighbourhood_cleansed"?

For hosts with a unique listing, there are about 20% of the host can cover all annualised mortgage repayment, and about 43% of the host cannot cover the mortgage.

Total_number_of_host	30,305
Total_number_of_host_can_cover_all	5,902
Total_number_of_host_can_cover_half	9,615
Total_number_of_host_can_cover_20per	13,819
Total_number_of_host_cannot_cover	12,894
Percentage_of_host_can_cover_all	19.48%
Percentage_of_host_can_cover_half	31.73%
Percentage_of_host_can_cover_20per	45.6%
Percentage_of_host_cannot_cover	42.55%

Table 4. Result for Question d

CHALLENGES DURING THE PROJECT

Implementing ELT pipeline

The first challenge I met was the confusion about ELT and ETL, I was not sure if I can drop columns or do any cleaning before loading the raw data into the database. And did not know if I perform any cleaning on the dataset before uploading it to Snowflake will be considered as an ETL pipeline instead of an ELT pipeline. Furthermore, I did extensive research on the internet but there are very few resources about how to implement an ELT pipeline, most of the online resources are about ETL pipelines.

I decided to do the initial python cleaning (selecting common columns) because otherwise, I do not know how to upload all CSV files without creating extra tables (dimension tables in the data warehouse).

Working on neighbourhood location

Identifying correct LGA names for “neighbourhood_lga” and “host_lga”columns is another major challenge I faced. First of all, the “neighbourhood cleansed” column contains many suburb names instead of LGA names. So I decided to build a new one by mapping “neighborhood” to LGA names as mentioned in the Steps of Project section. I also thought about finding a suburb-to-lga name with latitude and longitude data on the internet to map the address into LGA names by using python packages. But I do not know how to achieve this output by using SQL. Furthermore, I found that not all areas in one suburb belong to the same LGA, as in the abovementioned Newtown example. In the end, I solved this challenge by mapping the suburb to the LGA with the largest suburb area in that suburb. However, this solution will reduce the accuracy of any later analysis because it is approximating all areas in one suburb into one LGA, which is different from the real-world situation.

The tool – SQL instead of python

I found it hard to use SQL to perform data cleaning, data cleaning here means extracting and transforming data to create the data mart schema. I’m confident to finish the same task in python in a much shorter time, for example, I believe cleaning the “host_neighbourhood” column will be much easier in python, whether by using re functions or extracting the geolocations and using the mapping library. But I understand that python is not good at these big datasets and SQL queries much faster than python on large datasets.

I also struggled to define functions in snowflake, I believe if I know how to use functions in Snowflake it will save a lot of time. And Snowflake SQL is somewhat different from MySQL and PostgreSQL, and I found Snowflake a bit harder to use or adapt to. The way I tackle this problem is like a brute force solution, devote more time and effort to this project and work on the complicated logic using Snowflake SQL.