AIRBNB PIPELINE USING AIRFLOW

Author: Nitish K

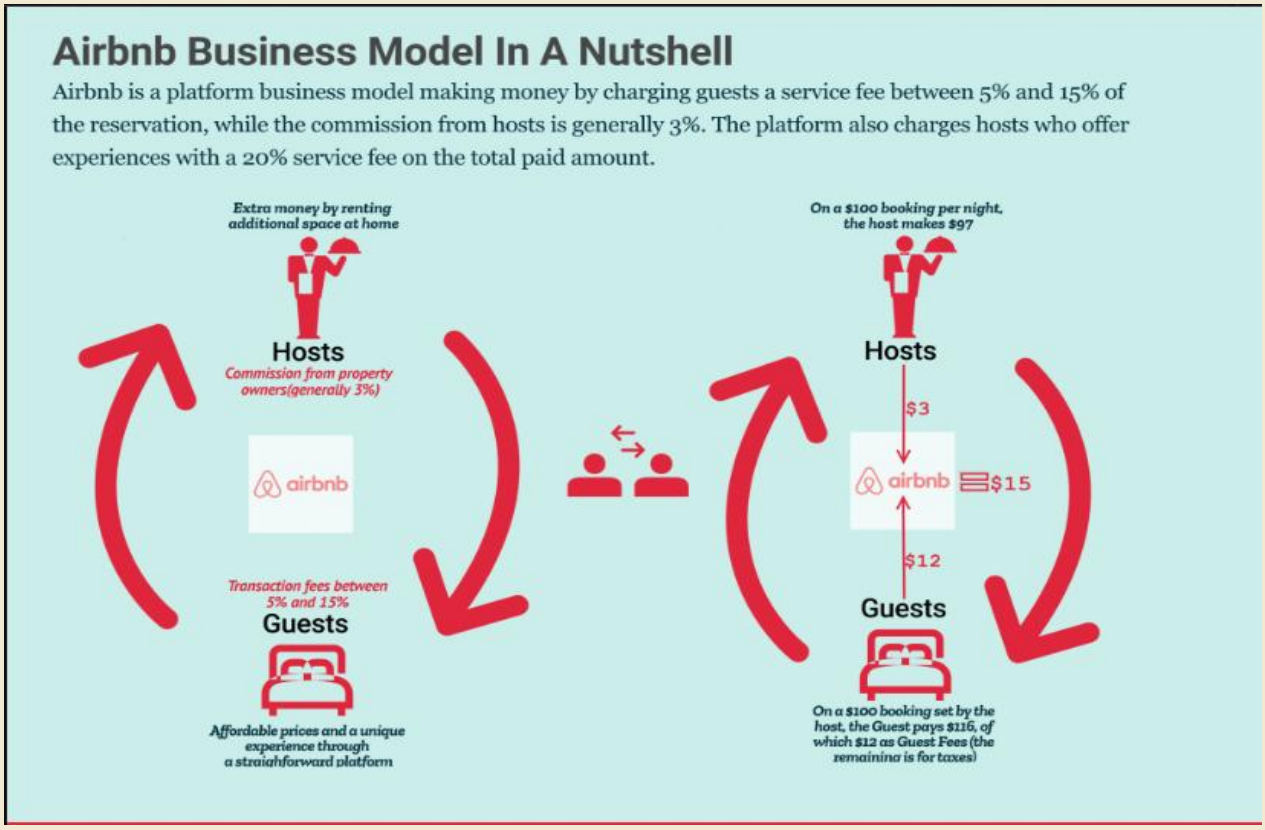
Intern ID: NLI-534

Introduction

Airbnb has emerged as a pioneering force, redefining the way people explore the world and find places to stay. Founded in 2008 by Brian Chesky, Joe Gebbia, and Nathan Blecharczyk, this innovative platform has transformed the concept of hospitality, ushering in a new era of personalized and immersive travel experiences.

Airbnb's journey from a humble startup to a global hospitality giant is nothing short of remarkable. Airbnb has successfully bridged the gap between travelers seeking unique, local accommodations and hosts eager to share their spaces and culture. Today, Airbnb boasts a diverse and expansive network of hosts and guests in virtually every corner of the globe.

Airbnb is based on an aggregator business model. Aggregator Business Model is a network model where the company collects the data about a specific good/service provider, make the providers their partners, and sell their services under its brand. People having vacant space at their home and willing to earn some extra money can list their rooms on the website and earn some extra bucks. Airbnb also has personal profiles and rating/reviewing systems that help the travelers. They visit a property and get high-definition photographs of the property. These high-definition photographs improve the click-rate and help in getting more responses.



My Research Work

I referred the Airbnb official website for information such as the following there around 190 countries and 34000 plus cities where Airbnb's are located. Over 150 million users make use of Airbnb for staying purposes. Around 1.9 million plus listings are available on their site. Places which were considered for analysis were from US, Europe, Asia, Australia. The descriptions are real as per mentioned on their site.

The building types include the following Apartment, Shared room, Entire room, Private room. The properties which were considered were Condominium, Loft, Townhouse, Cabin, Apartment, House, Bed & Breakfast, Unconventional. Most of the houses are instant bookable. The cancellation policies depend on the nature of the host. Most of them impose strict cancellation policy while others are moderate or flexible. Even the host experience can also be taken into account. Most of the hosts are experienced, few of them maybe Moderately experienced or Inexperienced.

There are various amenities which are provided out of which the compulsory ones are Wi-Fi, hot water bath, fully equipped kitchen, proper double bed, first aid kits, Smoke alarm, Fire Extinguisher. Additional amenities depending on the hosts include Air conditioning, Pool, Free parking, Iron and iron board, Washer and Dryer, Smart locks for self check-in, LED TV, Coffee machine, Heaters. The hosts do provide raw materials for preparing food such as Cereals, Fruits, Vegetables. The other additional food depending the host include Yogurt, Milk, OJ, Eggs, Juices, Muffins, Bread and Jam. All of these do have their respective costs and the costs are set based on the host.

Columns for the dataset

1. id ---> generate a 7-digit number
2. Name ---> Name of the house
3. Description --> Generate description of the house take it from the csv file
4. Building type --> ENTIRE HOUSE, Apartment, Private Room, Shared Room
5. property type --> Condominium, Loft, Townhouse, Cabin, Apartment, House, Bed & Breakfast, Unconventional
6. host id --> generate a 7-digit number
7. host name --> generate names

8. host_identity_verified --> confirmed not confirmed
9. neighborhood --> generate valid names
10. country --> Take from usa, Europe, Asia, Australia
11. zip codes --> generate random 4 digit
12. currency --> generate currency
13. instant bookable --> true, false
14. cancellation policy --> strict, moderate flexible
15. Construction year --> starting from 2008
16. Accommodation year --> consider from 2009 (construction year + 1)
17. commission --> generate random commission between 10% to 15%
18. Booking total --> generate a descent amount
19. status --> confirmed or not confirmed
20. bookings --> generate proper values
21. service fee --> 20% (if confirmed Service fee on booking total else don't put)
22. food cost --> between 20 to 40 usd
23. minimum nights --> generate based on user
24. reviews --> generate based on user
25. number of reviews -->
26. reviews per month --> between 1 to infinity
27. review rate number --> between 1 to 10
28. availability 365 -->
29. house_rules --> consider from the dataset

KPI Column

1. Total revenue : If host is experience then 20% of booking total(bt)+food cost+airbnb mag ($15) else 3% of bt+service fee*booking+food cost+airbnb mag
2. Host revenue = BT*Host fee+10% commission*BT
3. Platform revenue = CTR*Impressions*Cost per click*Rating factor
    a. CTR = 0.3
    b. Impressions = Reviews per month
    c. CPC = $0.76
    d. Rating = Review Rate number

Technologies used

1. Language: Python
2. Tools: Pycharm, Google Sheets
3. APIs: Google drive API, Google Sheets API
4. Visualization: Google Data Studio
5. ETL: Airflow

Implementations

The data was generated as follows

names = ["Black Lake Cabin", "Radcliff Refuge", "Rockaway Beach Villa", "The Red Bungalow of Deerfield"]

descriptions = ["Clean & quiet apt home by the park", "Skylit Midtown Castle"]

building_types = ["Entire house", "Apartment", "Private room", "Shared room"]

property_type = ["Condominium", "Loft", "Townhouse","Cabin","Apartment", "House","Bed & Breakfast", "Unconventional"]

Above I have given u a glimpse of the data that is going to be fed into google sheets

Random.choice method from random class was made use of to randomly select a value from a list of values. Faker library was made use of for ids, zip codes and also for the dates.

The same method was used to fill up the columns which I have mentioned above.

This entire thing which we have done above is a task. Now we have to schedule this task so that it gets feed into the google sheet. To do so we need to create a pipeline which will extract this data and load to the sheets. We make use of Airflow for this.

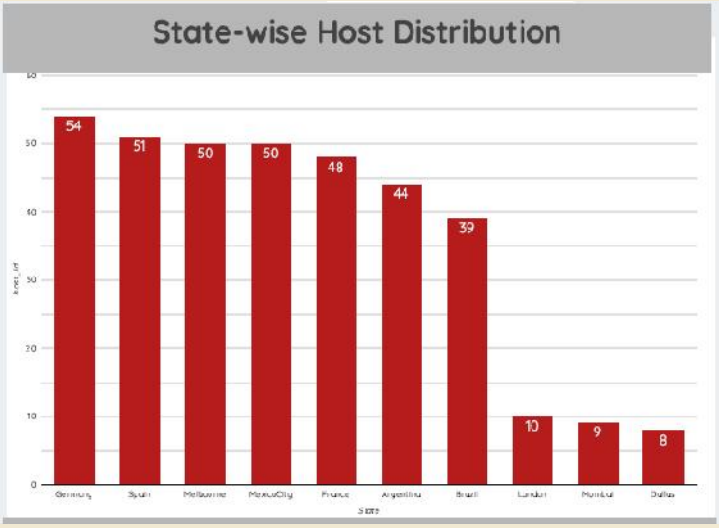
I created a DAG named “airbnb_end_to_end” for the same which has two tasks in it

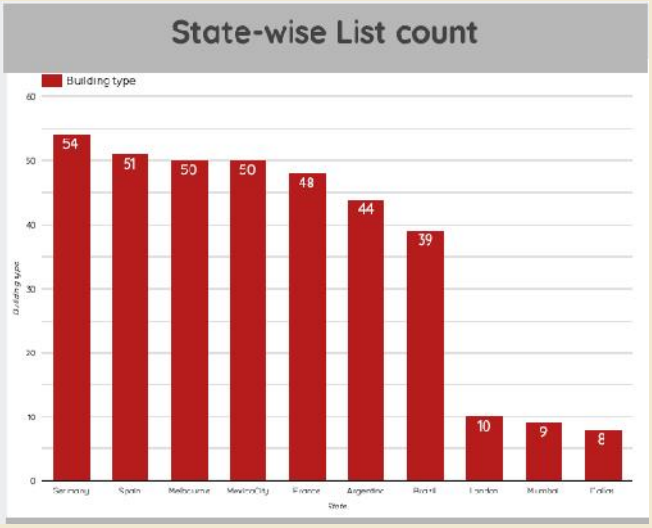
1. generate_data_task
2. task_try

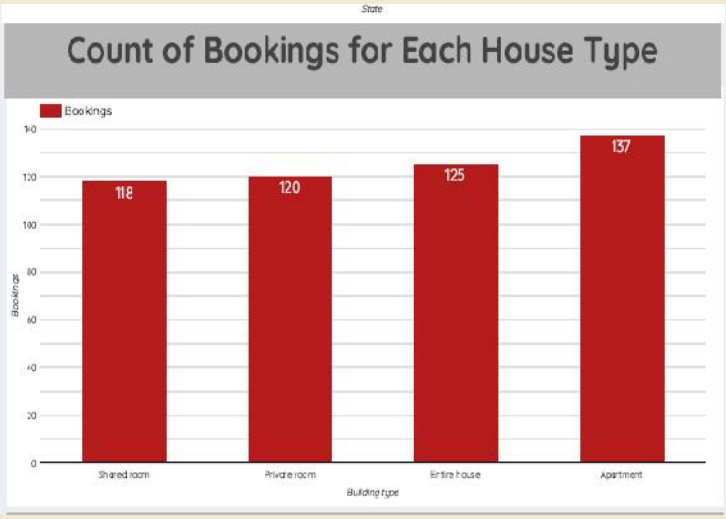
Generate_Data_Task will generate the data from the code based on the conditions specified and Task_Try is responsible for pushing the data to the Google sheets. This DAG is scheduled to run for every 15 minutes since Data studio also has a refresh time of 15 minutes.

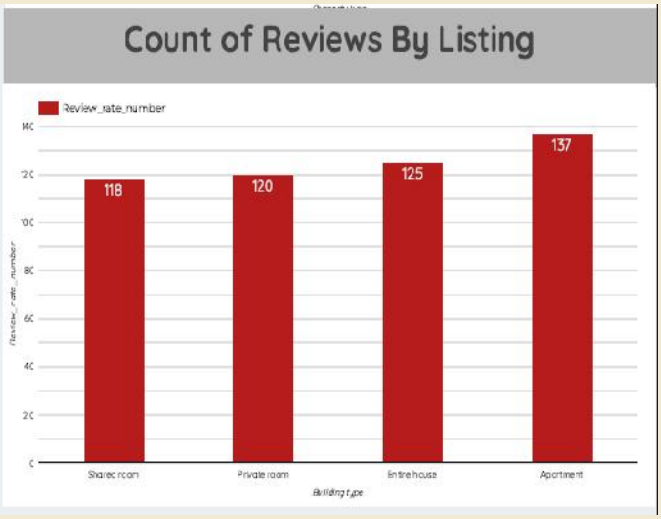
Visualizations

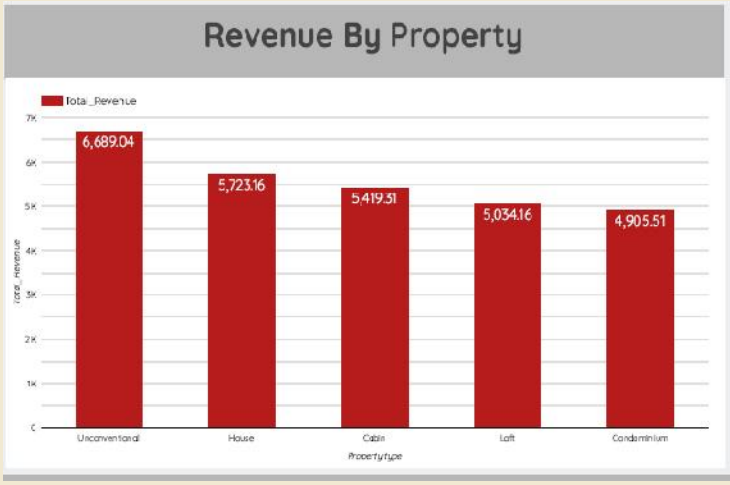
1. Comparisons

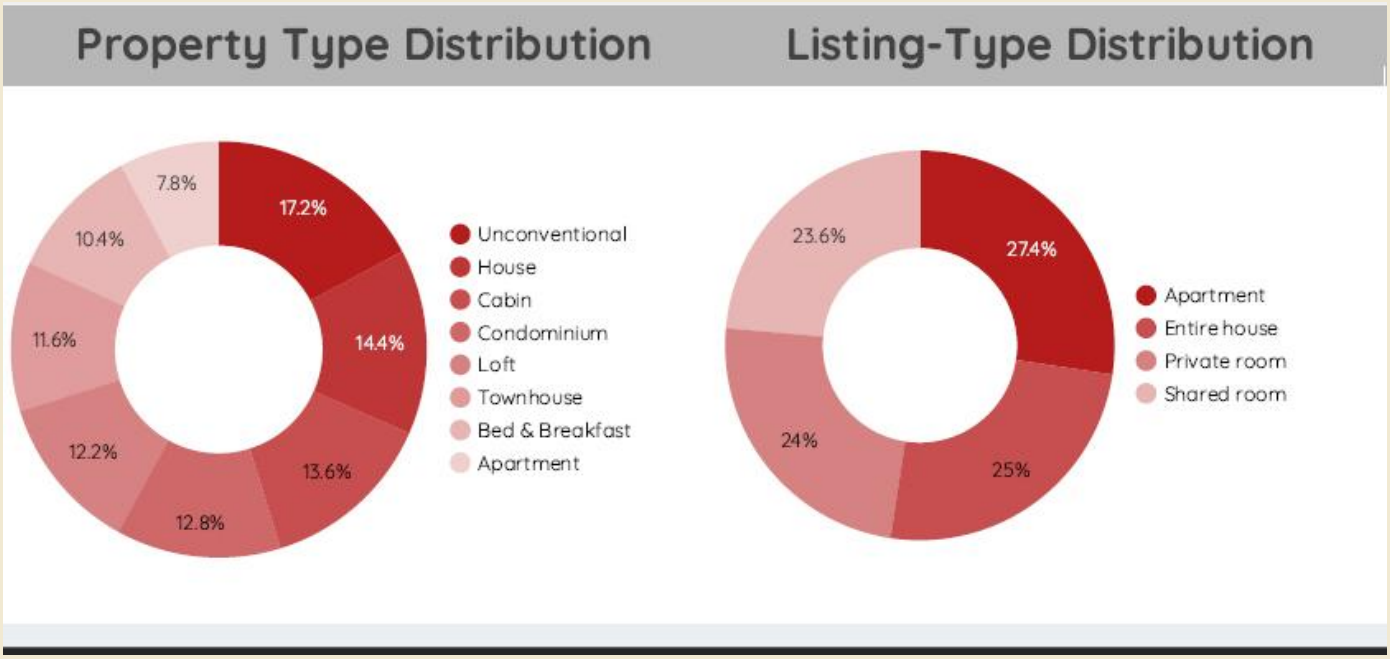




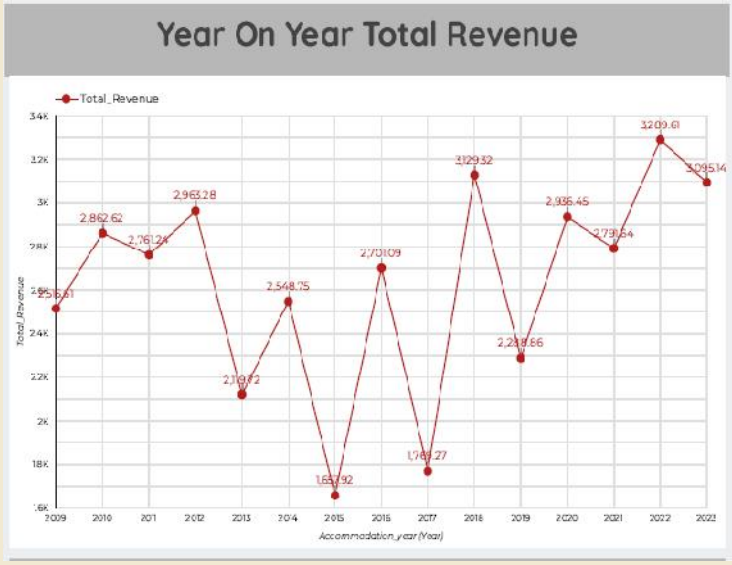


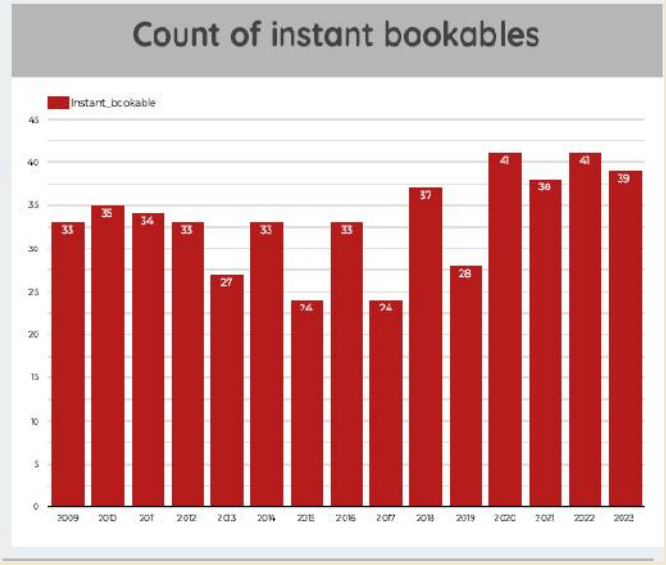


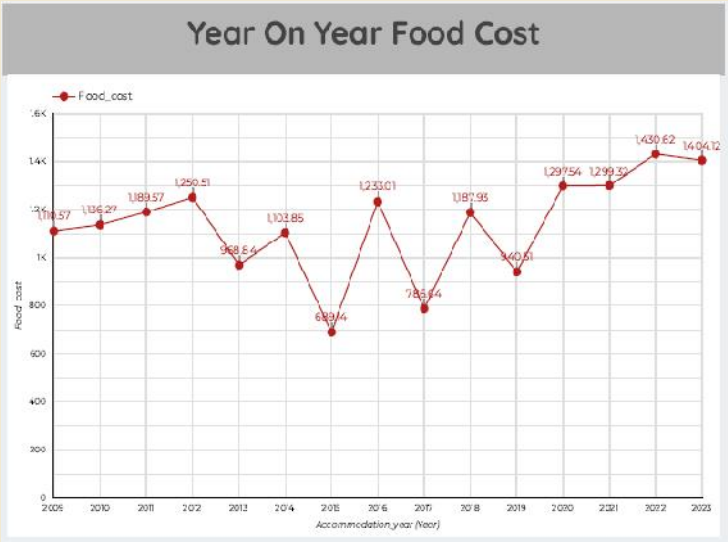


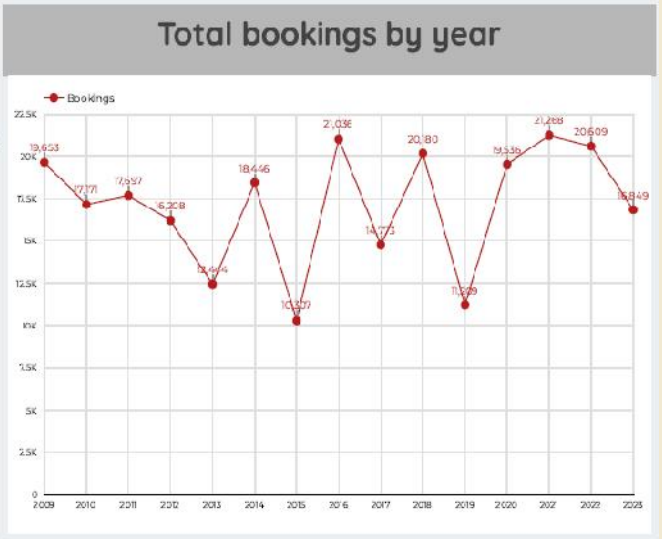


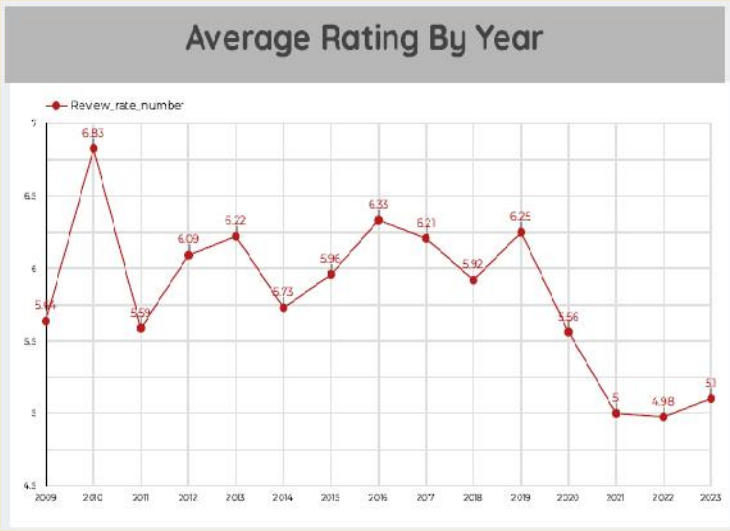
2. Seasonal Trends

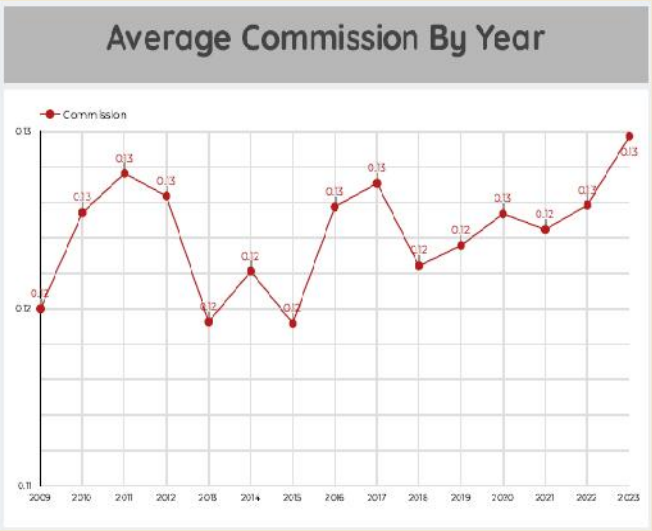












Insights

1. Each host owns a particular property
2. Highest count of hosts can be seen in Germany (54)
3. Highest list count can be seen in Germany (54)
4. Highest food cost is in Germany ($1,839.78) lowest is in Brisbane ($266.09)
5. Highest revenue is earned from property type unconditional ($6,689.04)
6. Count of bookings for apartments is the highest (137).
7. Count of reviews for apartments is the highest (137).
8. Correspondingly revenue earned from apartments is the highest ($10,951.59).
9. Highest revenue was earned in 2022 ($3,289.61) since the count of instant bookable were 41
10. Highest revenue earned by the host was in 2020 ($1,147.62).
11. Highest revenue earned through platform was in 2019 ($336.53)
12. Highest bookings were done in 2021.