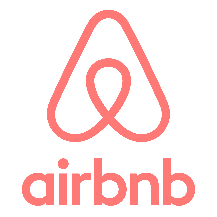
**Detailed Project Report (DPR)**

**Travel Data Analysis (Air BNB Case Studies)**



Last Revised Date: 10/08/2021

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author | Comments |
| 0.1 | 08-Oct-21 | Subhabrata Nath | Update Objective. |
| 0.2 |  |  |  |
|  |  |  |  |

**Reviews:**

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Reviewer | Comments |
|  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version | Review Date | Reviewed By | Approved By | Comments |
|  |  |  |  |  |

**Approval Status:**

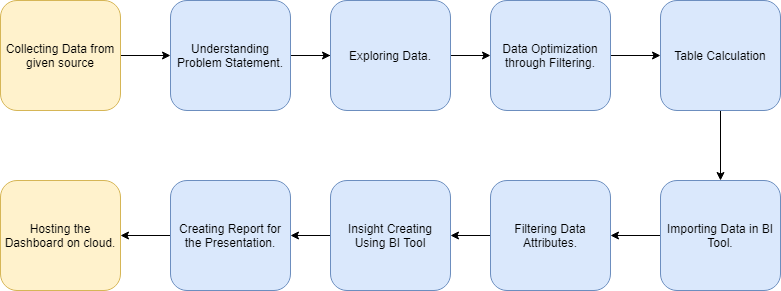
**1. Objective**

Perform data preprocessing through ETL, data cleaning, and perform data analysis to find Research Questions from the AirBNB travel dataset. Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in San Diego, California for 2019. Our primary goal is to find the relation between different attributes and find Various Business Questions and Answers from the Dataset by analyzing and plotting different attributes. This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

**1.1 Benefits:**

* Find Who is the top earned host from the different places.
* Get the understanding about location, from where AirBNB getting Maximum and Minimum Booking.
* Get an Understanding of Price Changes in different Months.
* Recognize the Costly Area for traveling.
* Know the relation between price and quality.

Architecture Design



**1.2 Architecture Description**

**1.2.1 Collecting Data:**

We have to collect the data for creating the report as needed, Dataset is already available on the project dashboard. We will use the given dataset for building the project.

**1.2.3 Understanding Problem Statement:**

We will perform exploring on AirBNB data and will find insightful information. Find information between different attributes.

**1.2.4 Exploring Data:**

The next steps will be exploring the data for becoming one with data. We will use the plotting library and python data manipulation library like pandas for these tasks.

**1.2.5 Data Optimization through Filtering:**

We will optimize the given dataset before starting the analysis. We have to drop duplicate columns and reduce the dataset dimensions as needed. We will also be encoding our data into numerical to categorical or vice versa as needed.

**1.2.6 Table Calculation:**

We will calculate some new fields that will represent more granular results than raw data. We will find the relationship between attributes and perform the calculation as required.

**1.2.7 Importing Data In BI Tool:**

We will use Power BI for creating the Visualization and Dashboard. Power BI is a very well-used application for creating Dashboard and it is very much efficient in handling a large amount of data. We can plot and show different types of relationships between attributes.

**1.2.8 Filtering Data Attributes:**

We will use different types of filters on our data set for plotting purposes and getting the insight information. Using filters can speed up information finding and also it can produce more variety and granular reports.

**1.2.9 Insight Creation Using BI Tools:**

This is the most crucial part of the complete process. Here we will find meaningful information by going deep into our dataset. We will plot different types of plots for every attribute on the Airbnb dataset to show the relationship between different attributes like location and price. This is our key finding on the dataset.

**1.2.10 Creating Report for the Presentation:**

After we find all the possible information that can be found from the dataset we will create a detailed report for presenting and publishing our works. We will include all the visual plots and will write key findings from the plots.

**1.2.11 Hosting the Dashboard on Cloud:**

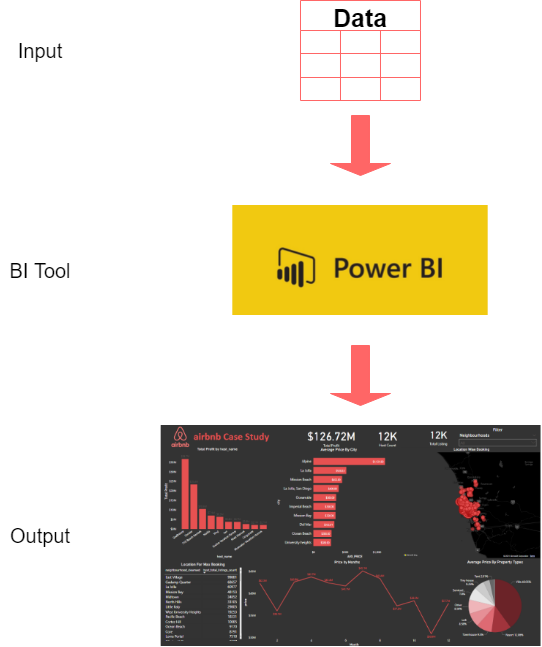
After creating a Detailed Project report and creating the dashboard we will host our dashboard on the cloud platform for global use. For this, we will use the Power BI service Platform. Here we can host our dashboard and it will create a sharable link for access globally.

**2. Data Manipulation**

The data is required for the analysis is provided and we have done data preprocessing to remove duplicate rows, impute missing values, and encoding numerical values to categorical values for producing a more readable report.

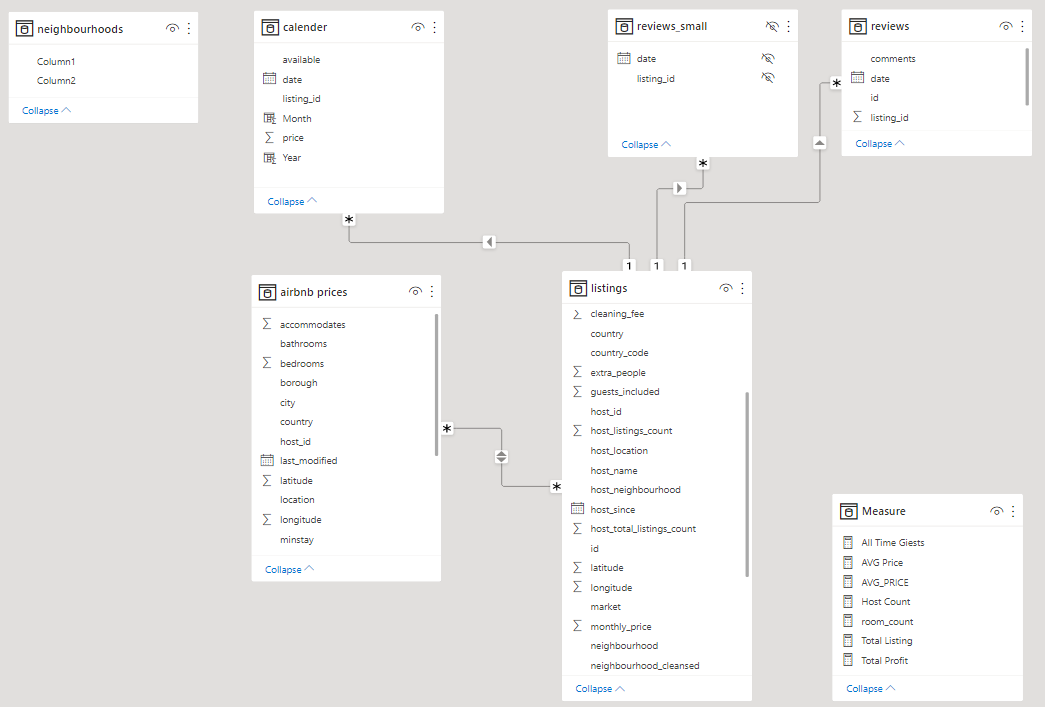
**3. BI Tools**

We have used Power BI for doing the analysis. We load our data into Power BI and perform ETL and various data analyses and try to find the relationship between different attributes.



**4. Entity-Relationship.**

As our dataset comes with different tables, We have created a relationship with matching columns for our analysis. The following diagram is showing the relationship.

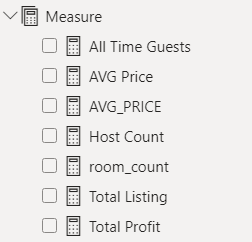


ER diagram help to explain the logical structure of the full dataset when there the multiple tables to work with. ER diagrams were created based on the matching columns from every table, from where the data types of the columns remain the same.

Here our Main Fact table is **`listing`** and we have created relationships with other tables based on the **`id`** column which contains every listing id.

**5. Measure**

We have created many custom measures for finding research questions. We take the help of the **`DAX`** function to write the custom measures. These helped us to produce more granular results. The following Custome Measure we have created.



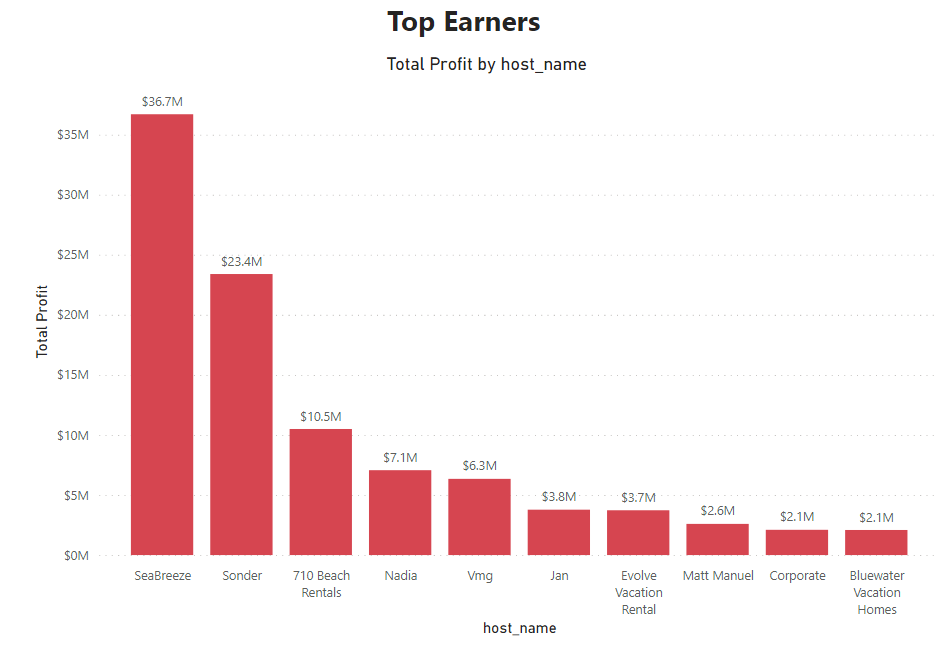
|  |  |
| --- | --- |
| **Custome Measures** | **Dax Function** |
| All Time Guests | = SUM(listings[guests\_included]) |
| AVG Price | = AVERAGE(listings[price]) |
| AVG\_PRICE | = AVERAGE('listings'[price]) |
| Host Count | = COUNT(listings[host\_id]) |
| room\_count | = COUNT(listings[room\_type]) |
| Total Listing | = COUNT(listings[id]) |
| Total Profit | = SUMX(listings, listings[price] \* listings[host\_total\_listings\_count]) |

Custome Measures do not include in the table, these are information that we have created by calculating different attributes from different tables by using pre-existing formulas.

**6. Key Findings**

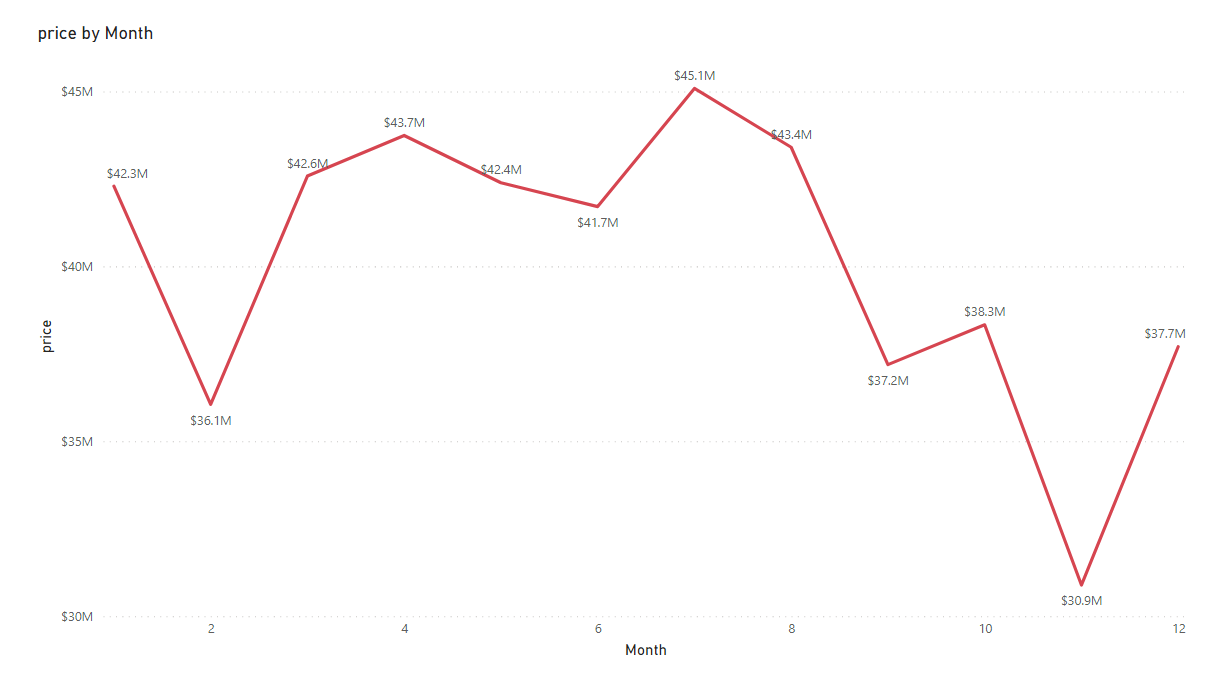
Here we will show different plots that we have come up with from the dataset and will describe the finding from the plot and will answering the Research Questions.

**6.1 Regarding The Host.**

**6.1.1 Who are the Top Earners.**

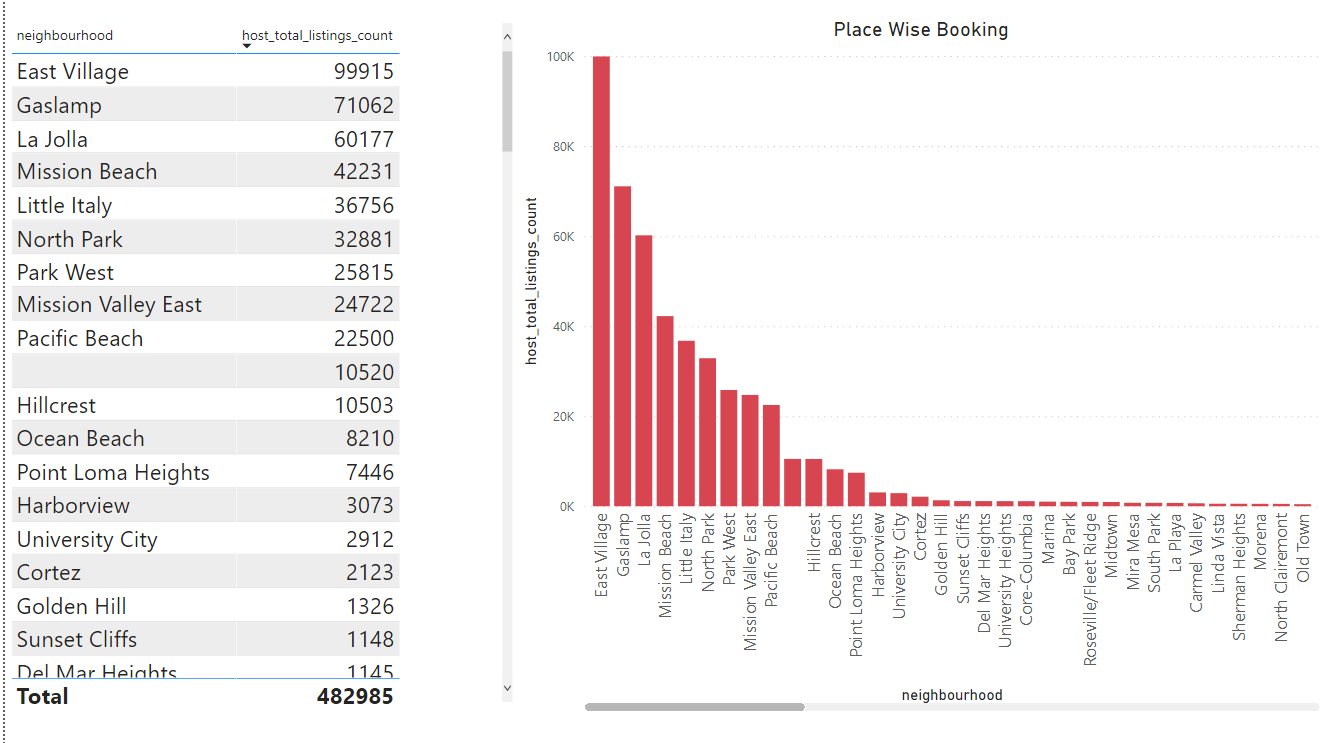
We have created a Bar Graph by taking **Hostname** and **Top Profit** measures to find the **Top Earners.**

From here we can see the top host by their total profit. So it is clear that the highest profitable hosts are top earners.

**6.1.2 Relationship Between Monthly earning and Price**

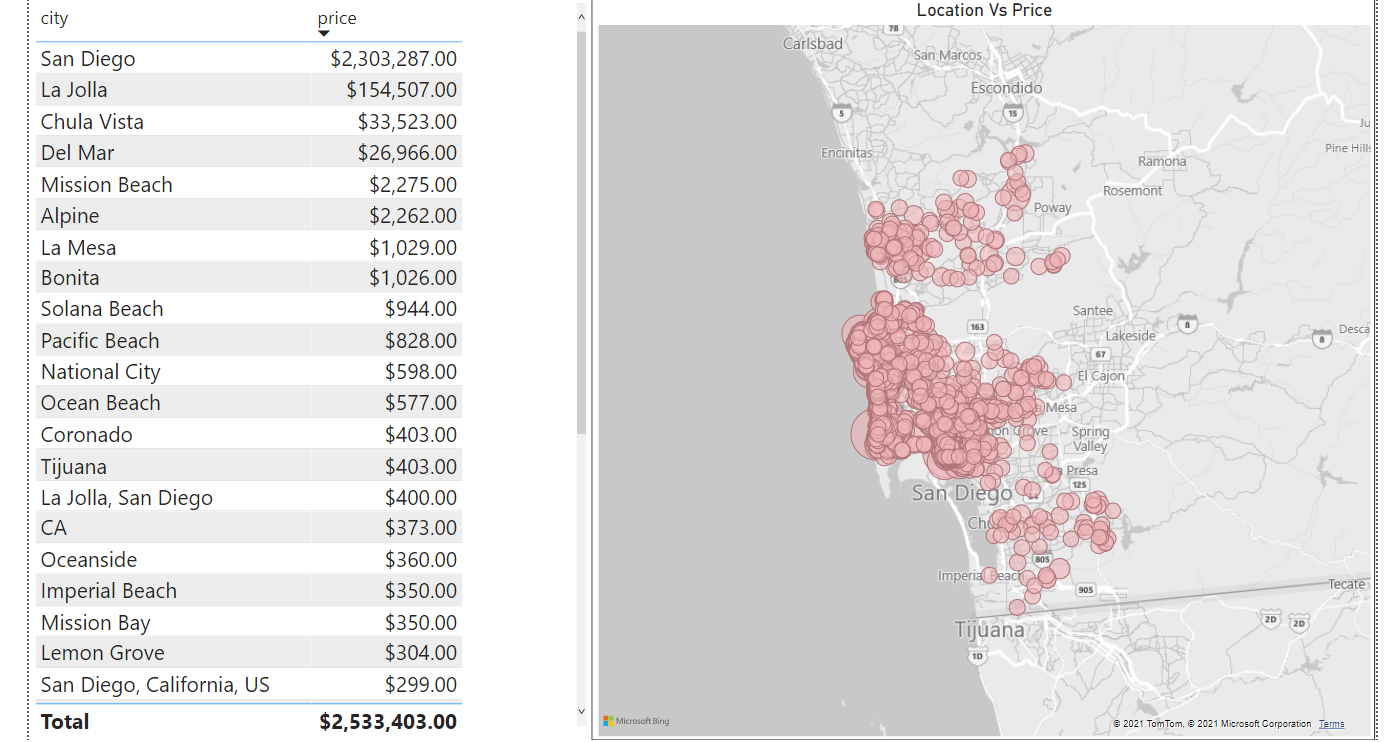
The price changes in months. By varying prices over months, the earning can vary also. From the above line chart is describe the Price goes up in **July-August** months and we can observe price drop in **February** and **November** months.

**6.2 Regarding Host**

**6.2.1 Any particular location getting the maximum number of bookings**

For finding the location from where we are getting the maximum number of bookings we have used **Bar Graph** to show the top location from where Airbnb is getting maximum numbers of bookings.

**6.2.2 Price Relation With respect to location**

****

From the above plot, we can observe the Expensive Places or top places where the price is high. As we have shown a table with location name and price with the descending format. And the map, where each pink circle is showing a specific location and the circle size, is depending on the price. The more is price, the bigger the circle. As we can observe the price is quite high in the sea beach area. It is normal. The hotel charges more those are near to the sea beach.

**6.3 Regarding the reviews.**

**6.3.1 Relationship between Quality and Price**

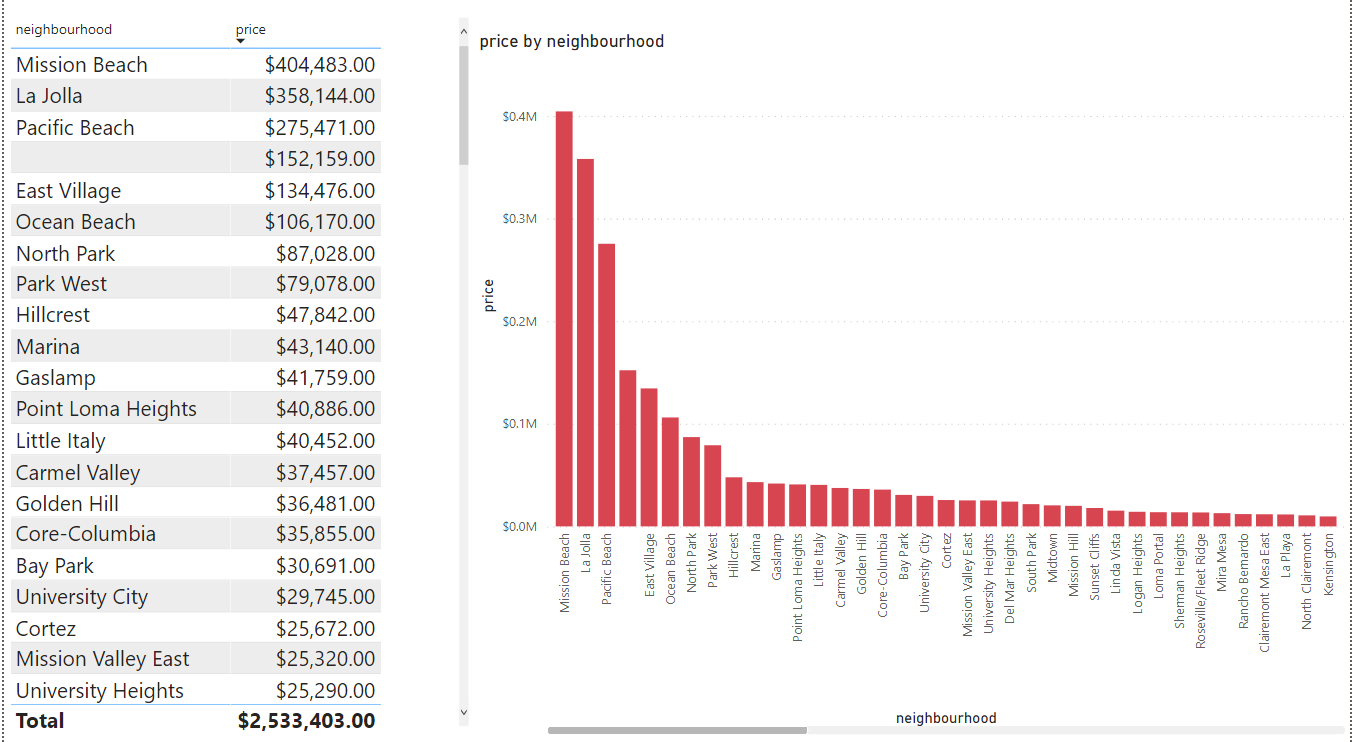
**Need to find out…**

**6.4 Regarding Price**

**6.4.1 Price Vs Amenities**

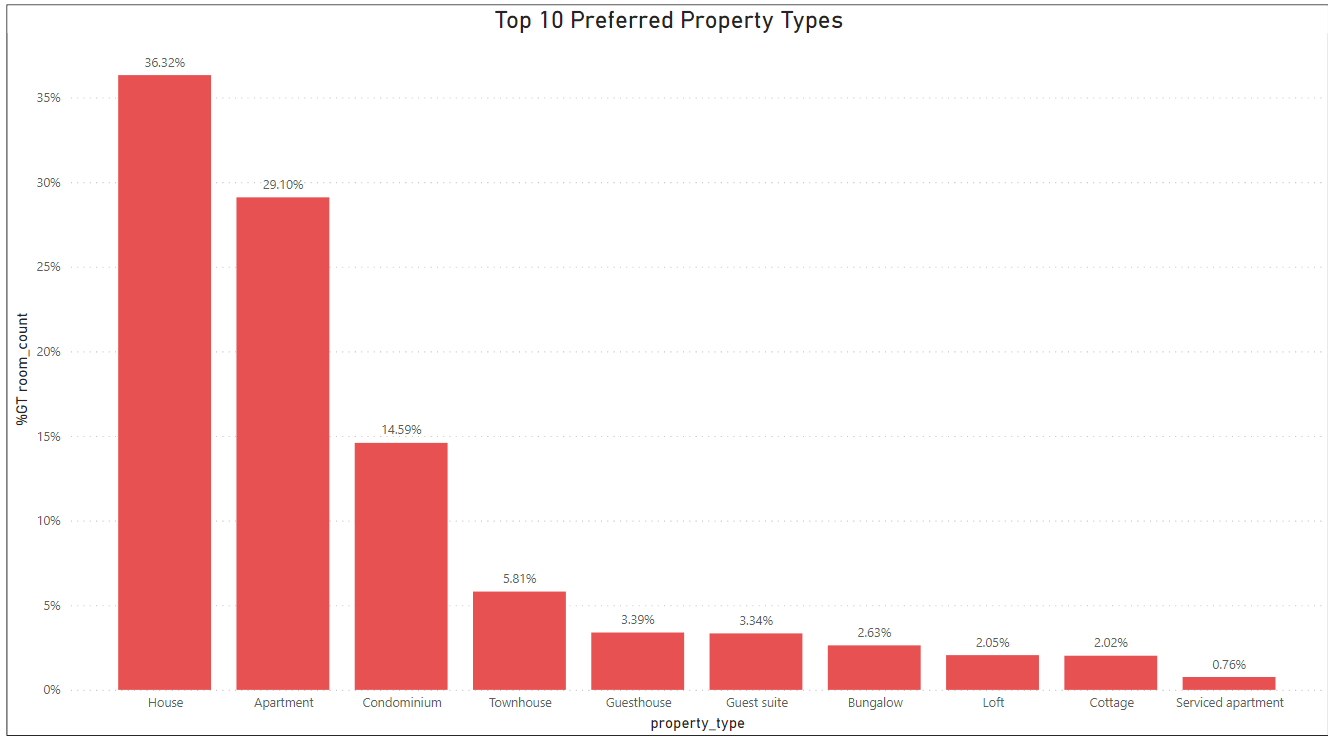
**Need to find out…**

**6.4.2 Price Vs Location**

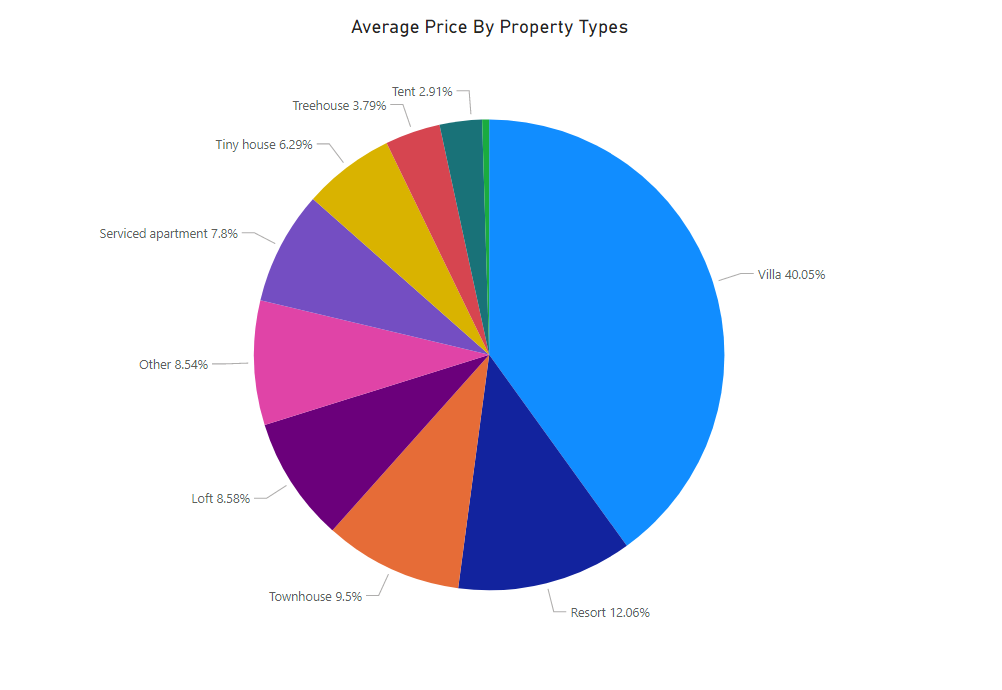


For finding the top expensive location we have created another **Bar Graph.** Where we have to use location columns `**neighborhood`** and `**Price`** columns to plot the graph. From the plot, we can observe **`Mission Beach`** is the most expensive area in California city. By following `**La Jolla`, `Pacific Beach`.**

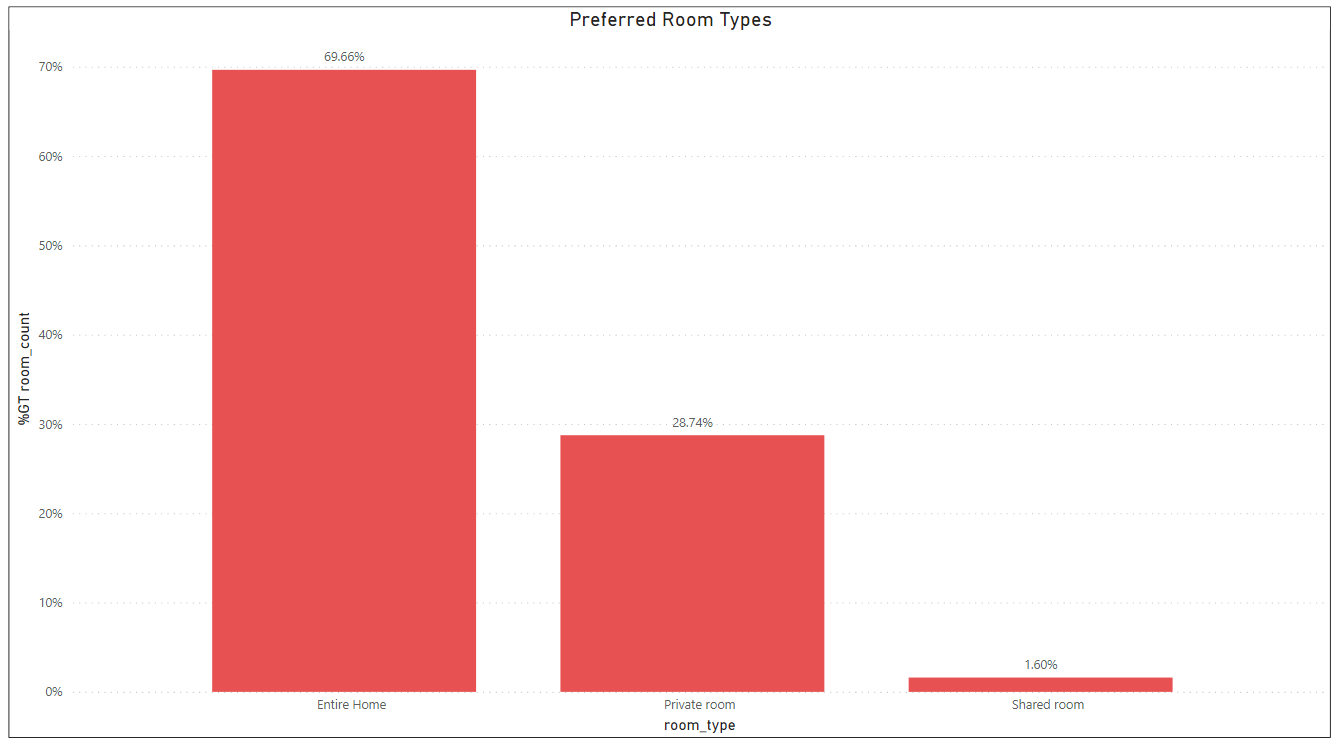
**7. Others Findings**

**7.1 Top Prefered type of Property**

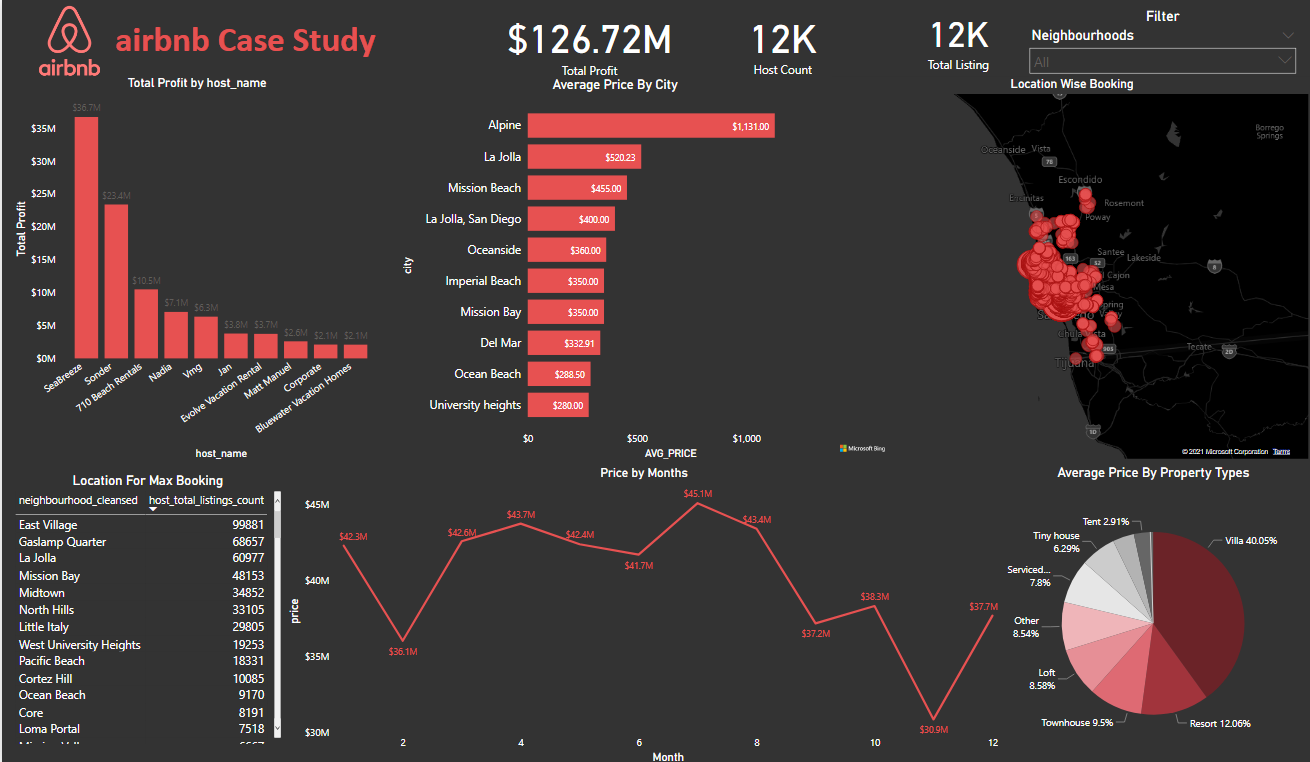
We have plotted a Bar Graph to find the most preferred property type that is tourists are looking for and they book. We have found that **Houses, Apartments, Condominiums** are the **top 3** property types that are a tourist looking for.

**7.2 Average Price by Property Type**

We have found that **`Villa`** is the most expensive property type. For finding the average price of every property we have created **`Avg Price`** custom measure. And Choose Pie Chart to show the descriptions. From the pie chart, we can observe the **`Villa(40%)`** is occupied most of the area. Then come **`Resort(12%)`** and **`Townhouse(9%)`.**

**7.3 Prefered Room Type**

From the above plot, we can find that most of the time tourists refer **`Entire Home`** and in second **`Private Room`** and last **`Shared Room`.** It can be also a general assumption most tourists will book an entire room for them, as most of them planed trips with a full family.

**8. Dashboard.**

We have created our dashboard by merging individuals' findings. We have included **`Top Earners`** and **`Location Vs Price`, `Price Vs Months`.** Also, we have included some text information to show the real number by using **`sticker`.** At the top right corner, we have included a filter based on Location. So we can change all plots by location. And can find different relations based on the specific location.

**9. Question and Answers (Q & A)**

**9.1 From Where Did You Collect The Dataset?**

The dataset is collected from the internship portal, It was given to us. The dataset can be found from the following link <https://drive.google.com/drive/folders/1ANkgtAT0Pdp2r86IxFKv9vKYmnsYjJDO?usp=s>

**9.2 What Is The Most Useful Table In Airbnb Travel Data?**

The **`listing`** table was the fact table for us, the most of the information we have received from the same table. This table consists of all the necessary information regarding the **hosts**, **price,** and **location.** And for other findings.

**9.3 What Are The Most Useful Attributes In Airbnb Travel Data?**

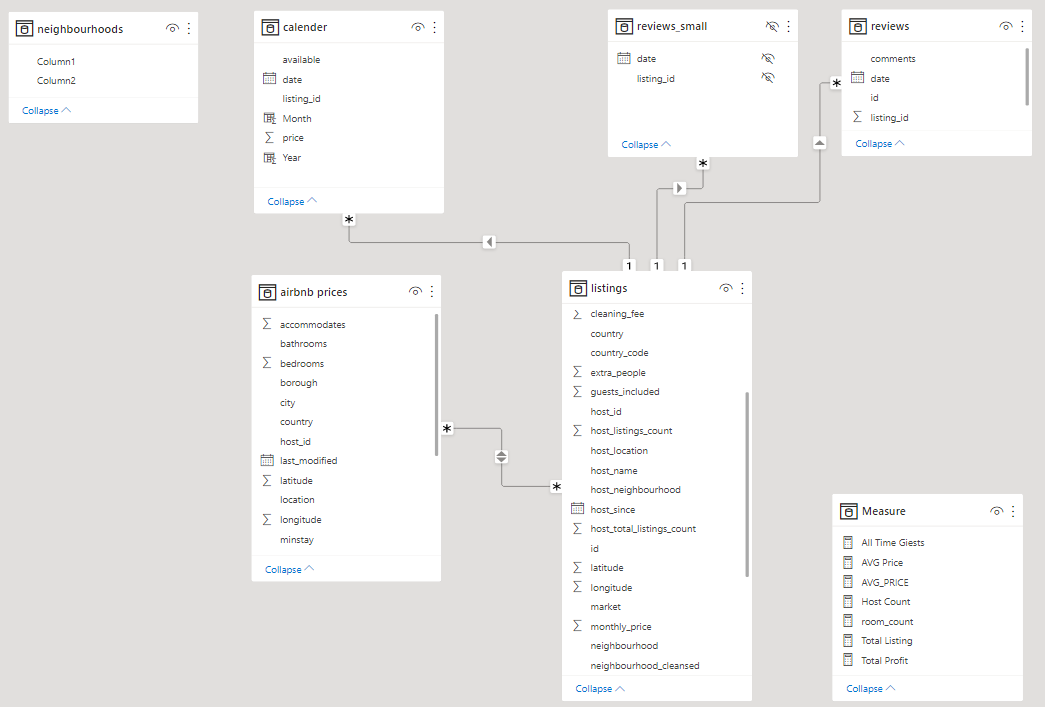
Attributes like **Host\_id, host\_name, Price, total\_listing\_count, city** are the most important feature that helps us to find the insight of travel data. From the neighborhood table, we take **`neighborhood name`** to plot the location to find the **`price vs location`.**

**9.4 Describe The Data Cleaning Process That You Performed On The Dataset.**

The dataset comes with in total 5 tables. We have removed many unnecessary columns from the **`listing`** table. And removes all the **`NaN`** or missing values from all tables. We have used **Power Query** for the data cleaning and make the data for further analysis.

**9.5 How Do You Create A Relationship Between Multiple Tables?**

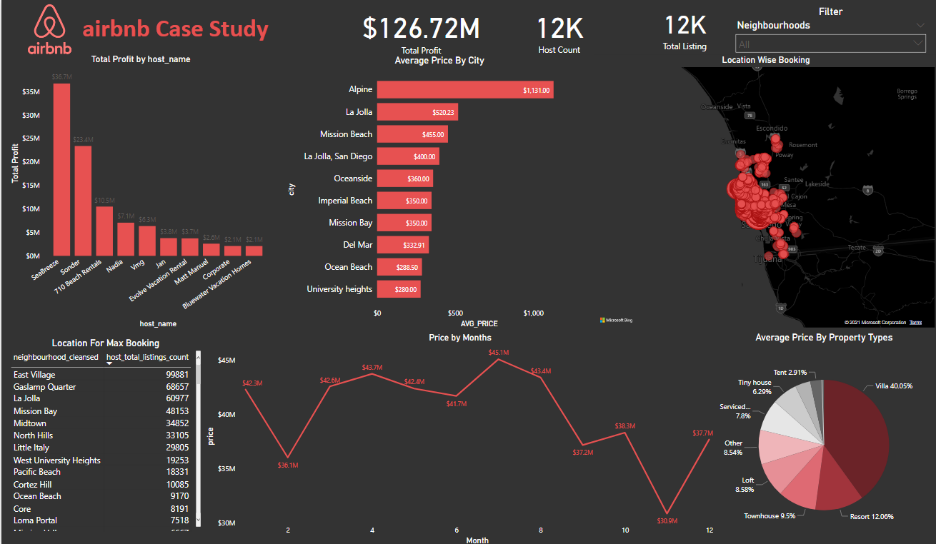
We create a relationship between different tables by taking one **Primary Key** and match that key with another table **Foreign Key**. Two columns from the different tables must be in the same data type. The main primary key was in most of the tables is **`host\_id`.** We match the **`id`** columns and create an entity diagram in Power BI.



**9.6 What Are The Important Tables Here To Find The Research Questions?**

The most important table is the **`listing`** table. This is our fact table and most of the plot and measure we have created from the same table. This table consists of most of the information like **`host\_name`, `price`, `neighborhood`, `location`,** and many more. We have created many custom measures for our analysis and we have used them on plotting.

**9.7 What Is The Final Outcome That You Are Showing?**

The outcome is Dashboard that we have built by merging the most important plots and present It for detailed analysis. We have deployed it on the **Power Bi Service** for global access.

**10. Conclusion**

Likhlo app lok. Kaya likhna haii…