Low Level Design (LLD)

AirBNB Travel Data Analysis

Author	Hemant Sharma
Document Version	LLD-V1.0
Last Revised Date	20/08/2023

Document Version Control

Date Issued	Version	Description	Author
20/08/2023	LLD-V1.0	First Version of Complete LLD	Hemant Sharma

Contents

Abstract

AirBNB began in 2008 when two designers who had space to share hosted three travelers looking for a place to stay. Now, millions of hosts and travelers choose to create a free AirBNB account so they can list their space and book unique accommodation anywhere in the world. In addition, AirBNB experience hosts share their passions and interests with both travelers and locals.

AirBNB helps make sharing easy, enjoyable, and safe. We verify personal profiles and listings, maintain a smart messaging system so hosts and guests can communicate with certainty, and manage a trusted platform to collect and transfer payments.

AirBNB has provided many travelers a great, easy and convenient place to stay during their travels. Similarly, it has also given an opportunity for many to earn extra revenue by listing their properties for residents to stay. However, with so many listings available with varying prices, how can an aspiring host know what type of property to invest in if his main aim is to list it in AirBNB and earn rental revenue? Additionally, if a traveler wants to find the cheapest listing available but with certain features, he prefers like 'free parking' etc., how does he know what aspects to look intoto find a suitable listing? There are many factors, which influence the price of a listing. Which is why we aim to find the most important factors that affect the price and more importantly the features that is common among the most expensive listings. This will allow an aspiring AirBNB hostto ensure that his listing is equipped with those important features such that he will be able to charge a higher price without losing customers. Moreover, a traveler will also know the factors to look into to get the lowest price possible while having certain features he prefers.

In the arena of rising new generation and innovation, Travel enterprise is advancing with the function of Data Science and Analytics. Data analysis can assist them to understand their businessin a quiet distinct way and helps to improve the exceptional of the provider by using identifying the vulnerable areas of the business. This examine demonstrates the how distinct analysis assist outto make higher business choices and help examine customer tendencies and pride, that may lead to new and higher products and services. Different evaluation accomplished along with ExploratoryData Analysis and Descriptive Analysis on type of use instances to get the important thing insights from these records primarily based on which enterprise decisions might be taken.

1 Introduction

Why this Low-Level design document?

The purpose of this LLD or a Low-Level Design (LLD) document is to give the internal logical design of the actual program code for Airbnb Data Analysis project. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document. This document is intended for both the stakeholders and the developers of this project and will be proposed to the higher management for its approval.

The main objective of the project is to analyze the various aspects with different use cases which covers many aspects of AirBNB listings. It helps in not only understanding the meaningful relationships between attributes but it also allows us to do our own research and come-up with our findings.

Scope

Low-level design (LLD) is a component-level design process that follows a step-by step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work.

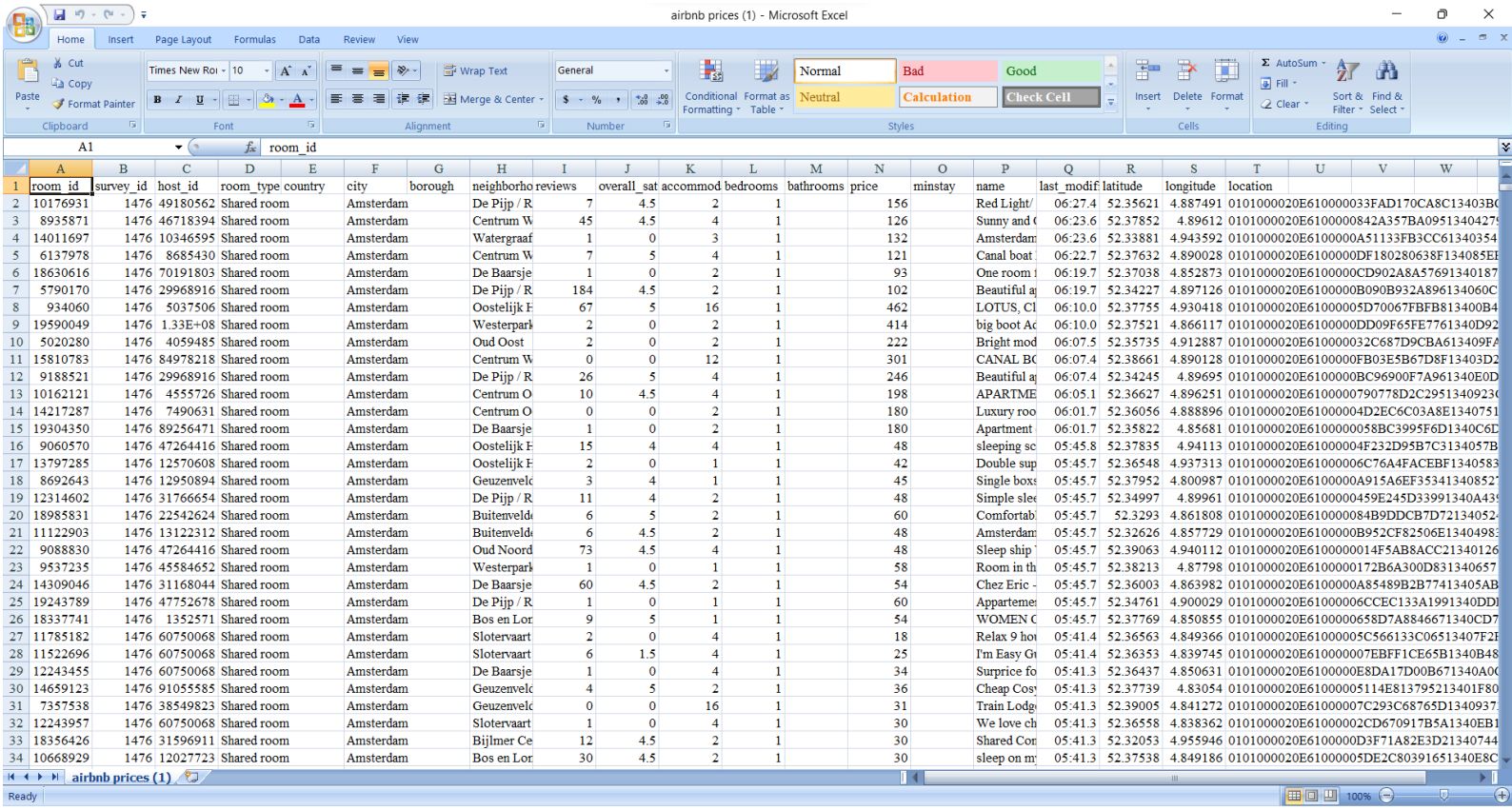
This study demonstrates the how different analysis help out to make better business decisions and help analysis customer trends and satisfaction, which can lead to new and better products and services.

Constraints

The analysis must be user-friendly, code must be neat & clean, EDA must be automated as much as possible because it will save huge amount of time. Moreover, users should not be required to have any of the coding knowledge as the insights they are looking for are mentioned in-detail with respective visuals.

2 Technical Specifications

Listings Dataset



room_id	survey_id	host_id	room_type	country	city	borough	neighborho	reviews	overall_sat	accommod	bedrooms	bathrooms	price	minstay	name	last_modif	latitude	longitude	location
10176931	1476	49180562	Shared room		Amsterdam		De Pijp / R	7	4.5	2	1		156		Red Light/	06:27.4	52.35621	4.887491	0101000020E610000033FAD170CA8C13403B
8935871	1476	46718394	Shared room		Amsterdam		Centrum W	45	4.5	4	1		126		Sunny and	06:23.6	52.37852	4.89612	0101000020E6100000842A357BA0951340427
14011697	1476	10346595	Shared room		Amsterdam		Watergraaf	1	0	3	1		132		Amsterdam	06:23.6	52.33881	4.943592	0101000020E6100000A51133FB3CC61340354
6137978	1476	8685430	Shared room		Amsterdam		Centrum W	7	5	4	1		121		Canal boat	06:22.7	52.37632	4.890028	0101000020E6100000DF180280638F134085E
18630616	1476	70191803	Shared room		Amsterdam		De Baarsje	1	0	2	1		93		One room	06:19.7	52.37038	4.852873	0101000020E6100000CD902A8A57691340187
5790170	1476	29968916	Shared room		Amsterdam		De Pijp / R	184	4.5	2	1		102		Beautiful a	06:19.7	52.34227	4.897126	0101000020E6100000B090B932A896134060C
934060	1476	5037506	Shared room		Amsterdam		Oostelijk F	67	5	16	1		462		LOTUS, Cl	06:10.0	52.37755	4.930418	0101000020E61000005D70067FBFB813400B4
19590049	1476	1.33E+08	Shared room		Amsterdam		Westerpark	2	0	2	1		414		big boot Ac	06:10.0	52.37521	4.866117	0101000020E6100000DD09F65FE7761340D92
5020280	1476	4059485	Shared room		Amsterdam		Oud Oost	2	0	2	1		222		Bright mod	06:07.5	52.35735	4.912887	0101000020E610000032C687D9CBA613409FA
15810783	1476	84978218	Shared room		Amsterdam		Centrum W	0	0	12	1		301		CANAL BC	06:07.4	52.38661	4.890128	0101000020E6100000FB03E5B67D8F13403D2
9188521	1476	29968916	Shared room		Amsterdam		De Pijp / R	26	5	4	1		246		Beautiful a	06:07.4	52.34245	4.89695	0101000020E6100000BC96900F7A961340E0D
10162121	1476	4555726	Shared room		Amsterdam		Centrum O	10	4.5	4	1		198		APARTME	06:05.1	52.36627	4.896251	0101000020E6100000790778D2C2951340923
14217287	1476	7490631	Shared room		Amsterdam		Centrum O	0	0	2	1		180		Luxury roo	06:01.7	52.36056	4.888896	0101000020E61000004D2EC6C03A8E1340751
19304350	1476	89256471	Shared room		Amsterdam		De Baarsje	1	0	2	1		180		Apartment	06:01.7	52.35822	4.85681	0101000020E6100000058BC3995F6D1340C6D
9060570	1476	47264416	Shared room		Amsterdam		Oostelijk F	15	4	4	1		48		sleeping sc	05:45.8	52.37835	4.94113	0101000020E61000004F232D95B7C3134057B
13797285	1476	12570608	Shared room		Amsterdam		Oostelijk F	2	0	1	1		42		Double sup	05:45.7	52.36548	4.937313	0101000020E61000006C76A4FACEBF1340583
8692643	1476	12950894	Shared room		Amsterdam		Geuzenveld	3	4	1	1		45		Single boxs	05:45.7	52.37952	4.800987	0101000020E6100000A915A6EF353413408527
12314602	1476	31766654	Shared room		Amsterdam		De Pijp / R	11	4	2	1		48		Simple slee	05:45.7	52.34997	4.89961	0101000020E6100000459E245D33991340A43
18985831	1476	22542624	Shared room		Amsterdam		Buitenveld	6	5	2	1		60		Comfortab	05:45.7	52.3293	4.861808	0101000020E61000084B9DDCB7D72134052
11122903	1476	13122312	Shared room		Amsterdam		Buitenveld	6	4.5	2	1		48		Amsterdam	05:45.7	52.32626	4.857729	0101000020E6100000B952CF82506E1340498
9088830	1476	47264416	Shared room		Amsterdam		Oud Noord	73	4.5	4	1		48		Sleep ship	05:45.7	52.39063	4.940112	0101000020E6100000014F5AB8ACC21340126
9537235	1476	45584652	Shared room		Amsterdam		Westerpark	1	0	1	1		58		Room in th	05:45.7	52.38213	4.87798	0101000020E6100000172B6A300D831340657
14309046	1476	31168044	Shared room		Amsterdam		De Baarsje	60	4.5	2	1		54		Chez Eric -	05:45.7	52.36003	4.863982	0101000020E6100000A85489B2B77413405AB
19243789	1476	47752678	Shared room		Amsterdam		De Pijp / R	1	0	1	1		60		Apartemer	05:45.7	52.34761	4.900029	0101000020E61000006CCEC133A1991340DDI
18337741	1476	1352571	Shared room		Amsterdam		Bos en Lor	9	5	1	1		54		WOMEN C	05:45.7	52.37769	4.850855	0101000020E6100000658D7A8846671340CD7
11785182	1476	60750068	Shared room		Amsterdam		Slotervaart	2	0	4	1		18		Relax 9 hot	05:41.4	52.36563	4.849366	0101000020E61000005C566133C06513407F2
11522696	1476	60750068	Shared room		Amsterdam		Slotervaart	6	1.5	4	1		25		I'm Easy Gu	05:41.4	52.36353	4.839745	0101000020E610000007EBFF1CE65B1340B48
12243455	1476	60750068	Shared room		Amsterdam		De Baarsje	1	0	4	1		34		Surprice fo	05:41.3	52.36437	4.850631	0101000020E6100000E8DA17D00B671340A0
14659123	1476	91055585	Shared room		Amsterdam		Geuzenveld	4	5	2	1		36		Cheap Cosy	05:41.3	52.37739	4.83054	0101000020E61000005114E813795213401F80
7357538	1476	38549823	Shared room		Amsterdam		Geuzenveld	0	0	16	1		31		Train Lodge	05:41.3	52.39005	4.841272	0101000020E61000007C293C68765D1340937
12243957	1476	60750068	Shared room		Amsterdam		Slotervaart	1	0	4	1		30		We love ch	05:41.3	52.36558	4.838362	0101000020E61000002CD670917B5A1340EB1
18356426	1476	31596911	Shared room		Amsterdam		Bijlmer Ce	12	4.5	2	1		30		Shared Cor	05:41.3	52.32053	4.955946	0101000020E6100000D3F71A82E3D21340744
10668929	1476	12027723	Shared room		Amsterdam		Bos en Lor	30	4.5	2	1		30		sleep on m	05:41.3	52.37538	4.849186	0101000020E61000005DE2C80391651340E8C

Listings Dataset Overview –

The Listings dataset consists of a table with 18722 records and 20 features. Features are distributed as 11 Continuous features and 9 Categorical features. There are a total 20% of records having Missing values.

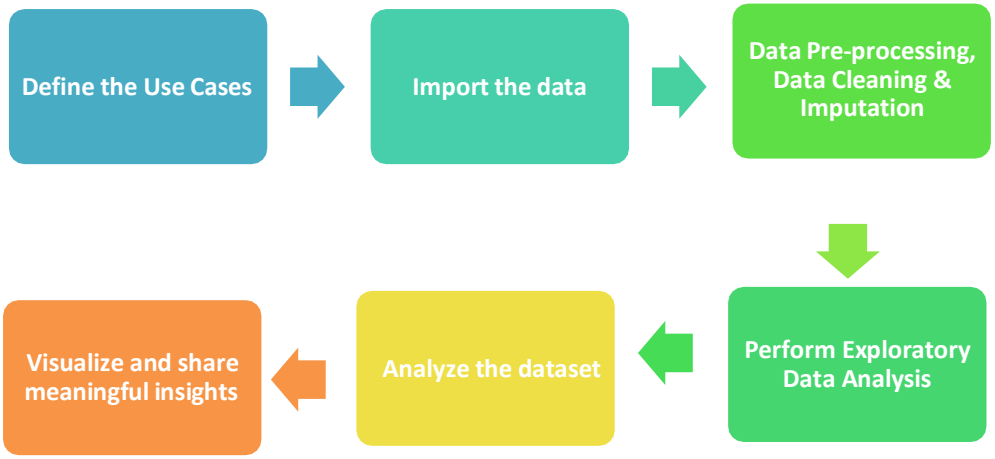
```
In [4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18723 entries, 0 to 18722
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   room_id               18723 non-null  int64
 1   survey_id             18723 non-null  int64
 2   host_id               18723 non-null  int64
 3   room_type             18723 non-null  object
 4   country               0 non-null      float64
 5   city                  18723 non-null  object
 6   borough               0 non-null      float64
 7   neighborhood          18723 non-null  object
 8   reviews               18723 non-null  int64
 9   overall_satisfaction  18723 non-null  float64
 10  accommodates          18723 non-null  int64
 11  bedrooms              18723 non-null  float64
 12  bathrooms             0 non-null      float64
 13  price                 18723 non-null  float64
 14  minstay               0 non-null      float64
 15  name                  18671 non-null  object
 16  last_modified         18723 non-null  object
 17  latitude              18723 non-null  float64
 18  longitude             18723 non-null  float64
 19  location              18723 non-null  object
dtypes: float64(9), int64(5), object(6)
memory usage: 2.9+ MB
```

```
In [6]: df.isnull().sum()

Out[6]: room_id                    0
        survey_id                  0
        host_id                    0
        room_type                  0
        country                18723
        city                       0
        borough                18723
        neighborhood               0
        reviews                    0
        overall_satisfaction       0
        accommodates               0
        bedrooms                   0
        bathrooms              18723
        price                      0
        minstay                18723
        name                      52
        last_modified              0
        latitude                   0
        longitude                  0
        location                   0
        dtype: int64
```

3 Architecture



Architecture Description –

Data Description –

As we have seen earlier, in our listings dataset, we have around 1.19 Lacs of records with 20 different features. Features are distributed as 10 Continuous features and 10 Categorical features and in our reviews dataset, we have around 3.44 Lacs of records with 6 different features among them there are 3 Continuous features and 3 Categorical features. These datasets are given in the form of Comma Separated Value (.csv) format.

Define the Use Cases –

At this stage, based on the given dataset and business problems we have defined the several Use Cases to perform the analysis on and this will definitely help out get the key insights from this data based on which business decisions will be taken. Furthermore, it helps in not only understanding the meaningful relationships between attributes but it also allows us to do our own research and come-up with our findings.

Import the Dataset –

As we have received the dataset in the form of Comma Separated Value (.csv) format, therefore we can import the same using Pandas read_csv( ) function.

In [2]:
```
#loading dataset
data = pd.read_csv("airbnb prices.csv")
df = data.copy()
df.head()
```

Out[2]:

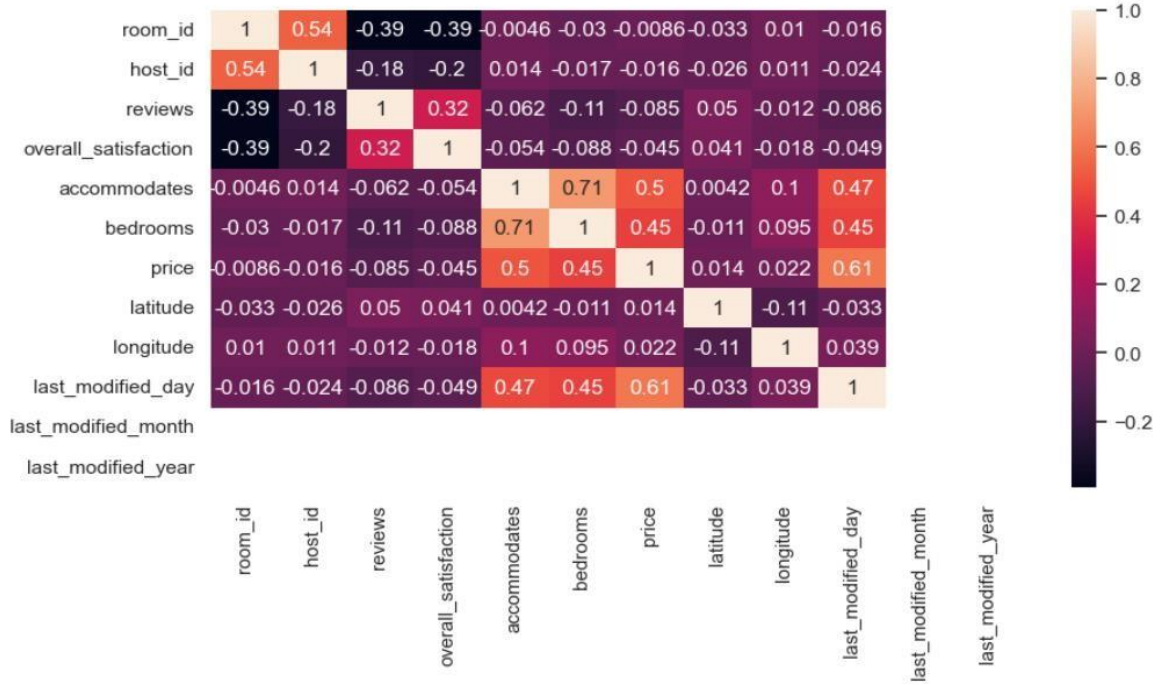
	room_id	survey_id	host_id	room_type	country	city	borough	neighborhood	reviews	overall_satisfaction	accommodates	bedrooms	bathroom
0	10176931	1476	49180562	Shared room	NaN	Amsterdam	NaN	De Pijp / Rivierenbuurt	7	4.5	2	1.0	Na
1	8935871	1476	46718394	Shared room	NaN	Amsterdam	NaN	Centrum West	45	4.5	4	1.0	Na
2	14011697	1476	10346595	Shared room	NaN	Amsterdam	NaN	Watergraafsmeer	1	0.0	3	1.0	Na
3	6137978	1476	8685430	Shared room	NaN	Amsterdam	NaN	Centrum West	7	5.0	4	1.0	Na
4	18630616	1476	70191803	Shared room	NaN	Amsterdam	NaN	De Baarsjes / Oud West	1	0.0	2	1.0	Na

Exploratory Data Analysis (EDA)–

- "Exploratory Data Analysis" (EDA) is a "Data Exploration" step in the Data Analysis Process where a number of techniques are used to better understand the dataset being used.
- Understanding the Dataset can refer to a number of things including but not limited to…
  - Extracting Important "Variables".
  - Identifying "Outliers", "Missing Values", or "Human Error".
  - Understanding the Relationships between variables.
  - Ultimately, maximizing our insights of a dataset and minimizing potential "Error" that may occur later in the process.
- In other words, it will give you a better Understanding of the "Variables" and the "Relationships" between them.
- Here, we make use of data preprocessing module to automate our EDA process.
- It provides the following information:
  - Overview: detect the types of columns in a DataFrame.
  - Variables: variable type, unique values, distinct count, missing values
  - Quartile statistics like minimum value, Q1, median, Q3, maximum, range, interquartile range
  - Descriptive statistics like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness.
  - Correlations: highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
  - Missing Values: Bar Chart, Heatmap and spectrum of missing values.

- Below we can visualize correlation with Heatmap:

```
In [62]: #heatmap for correlation
         plt.figure(figsize=(10,5))
         sns.heatmap(df_cleaned.corr(),annot=True)
Out[62]: <Axes: >
```



Data Pre-processing, Data Cleaning & Imputation (Handling the Categorical & Numerical Variables) –

Data pre-processing is a process of preparing the raw data and making it suitable for our analysis purpose, where we have to do lot of Data Cleaning, handle the missing values by using appropriate imputation techniques and based on that variable nature i.e. either of Categorical & Numerical variable. Here, in this project, we have done the substitution/imputation of missing values using either mean, median or mode according to the nature of those variables. Moreover, we also removed the columns which are does not participate in our analysis.

Analyze the Data –

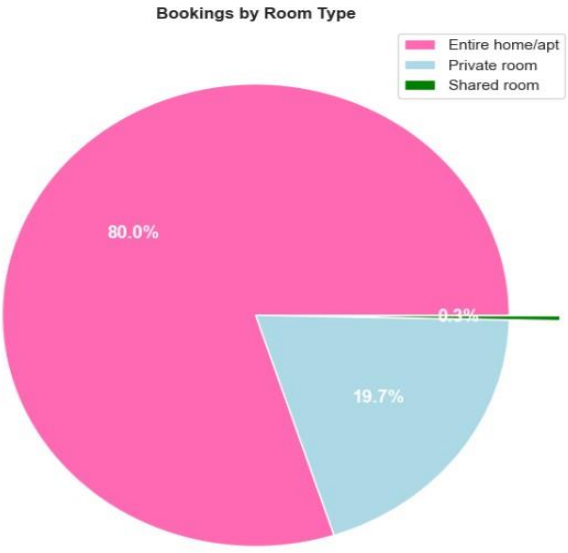
Once the pre-processing is done, we are good to go with our actual analysis where we write lines of codes and logics to prepare our data as per the defined use cases.

Visualize & Share Meaningful Insights –

Finally, it is time to turn our data into some sort of visual representation. In short, Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals such as Bar Plot, Pie Chart, Heat map, Box Plot, Scatter Plot, and many more. The resulting visual representation of data makes it easier to identify and share insights about the information represented in the data.

Here is the beautiful glimpse of one of our visuals are –

>> Which Room Type is Mostly liked?



Insights:

Above Pie Chart Shows

1. Entire Home/Apt are mostly booked, it cover 80% of the area.
2. Secondaly Private rooms are booked by more than 19%.
   However Shared rooms are less booked or less prefered by cutomers.

All those different analysis help out to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.

4 Technology Stack

Data Manipulation & Mathematical Computation Library	Pandas, NumPy
Visualization Library	Matplotlib, Seaborn, Plotly
Dataset	.CSV Format
IDE	Jupyter Notebook