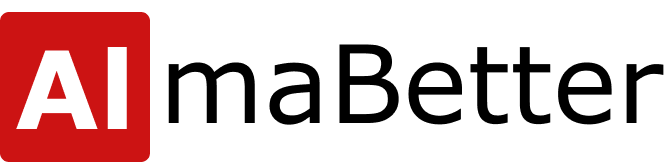
**“Airbnb Bookings Analysis”**

**EDA Capstone Project**

**Submitted To**



**AlmaBetter**

**Submitted By:**

**Vikas panchal**

**Email id: -** [**panchalvicky501@gmail.com**](mailto:panchalvicky501@gmail.com)

**Naveen Kumar batta**

**Email id: -** [**naveenbatta4587@gmail.com**](mailto:naveenbatta4587@gmail.com)

Abstract

Airbnb is an American company that facilitates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. It basically connecting traveller’s with local hosts who want to rent out their homes with people who are looking for accommodations in that locality. On the other hand, this platform enables host to list their available space and earn extra income in the form of rent and it also enables traveller’s to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. In the world of rising new technology and innovation, Travel industry is advancing with the role of Data Science and Analytics. Data analysis can help them to understand their business in a quiet different manner and helps to improve the quality of the service by identifying the weak areas of the business. This study demonstrates the how different analysis help out to make better business decisions and help analyse customer trends and satisfaction, which can lead to new and better products and services. Different analysis performed such as Exploratory Data Analysis and Descriptive Analysis on variety of use cases to get the key insights from this data based on which business decisions will be taken

Table of Contents

**1 Introduction1**

1.0 Introduction 1

1.1. Scope 1

1.2. Problem statement 1

1.3 Objective 2

1.4 Ethics2

1.5 Delimitations2

**2. Architecture Description3**

2.1. Data Description3

2.2. Define the Use Cases 3

2.3. Import the Dataset4

2.4. Exploratory Data Analysis (EDA) 4

2.5. Data Pre-processing, Data Cleaning & Imputation 5

2.6. Analyse the Data 5

**3. Technology Used4**

3.1. Python5

3.2. Jupyter Notebook 5

3.3. Pandas6

3.4. NumPy 6

3.5. Matplotlib6

**4. Overview 7**

4.1 General overview of dataset7

4.2 Understanding the Data7

4.3 Missing value Handling7

4.4 Overview of Related Work8

4.4.1 Top 25 most common words used in listing names8

4.4.2 Top 25 Used Words for Listing Names9

4.4.3 Neighbourhood group Frequency9

4.4.4 Top 15 highest listing Neighbourhood10

4.4.5 Total count of each room types as per listing11

4.4.6 Minimum no’s of nights stayed and preferred rooms12

4.4.7 Value count of Neighbourhood13

4.4.8 Area vs Number of reviews14

4.4.9 Room types and their Relation with availability in different neighbourhood groups14

4.4.10 Find total no. nights spend as per location15

**5.Conclution17**

5.1 Challenges Faced17

5.2 We defined some points which can help Airbnb in their business17

5.3 Limitation17

5.4 Scope of Improvement18

5.5 Conclusion18

Chapter 1

# Introduction

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

The purpose of this document is to provide a detailed architecture design of the Airbnb Data Analysis Project by focusing on each of the attributes of our architecture. This document will address the background of this project, and the architecturally significant function requirements. The intension of this document is to help the development team to determine how the system will be structured at the highest level

## 1.1 Scope

Architecture Design Document (ADD) is an architecture design process that follows a step-by-step refinement process. The process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the design principles may be defined during requirement analysis and then refined during architectural design work.

**1.2 Problem statement:**

Our biggest problem as hosts is with guests who don’t READ the listing. We make it very clear - as in, repeating it several times in the listing - that these are two private ROOMS in the boss’ home. She lives their full time. And yet we still have people show up thinking they booked the entire house for $80/night.

The other biggest problem is the last minute “Do you mind if we add one more person” requests. Again, our listing is VERY clear that there are three beds, with a total guest potential of 6 people. That’s all. And yet, at least once a month, the day before the guests are scheduled to arrive, we get an email asking if it’s okay if they bring one, two, or even three more people! The boss does have a blow-up mattress, and she’s willing to add one more person sometimes, but not two or three. We had one person book the rooms once, and 6 adults and 3 children showed up! She accommodated them but sent AirBnb a message about this and advising the company that in future, if someone pulled this crap, she’d be sending them away. Airbnb agreed with her that she would be perfectly within her rights to do so.

Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

We need to explore and analyze the data to discover key understandings (not limited to these) such as:

* What can we learn about different hosts and areas?
* What can we learn from predictions? (ex: locations, prices, reviews)
* Which hosts are the busiest and why?
* Is there any noticeable difference of traffic among different areas and what could be the reason for it?

**1.3 Objective:**

In the world of rising new technology and innovation, Travel industry is advancing with the role of Data Science and Analytics. Data analysis can help them to understand their business in a quiet different manner and helps to improve the quality of the service by identifying the weak areas of the business.

**Benefits:**

* Help out to make better business decisions.
* Help analyze customer trends and satisfaction, which can lead to new and better products and services.
* Gives better insight of customers base.
* Helps in easy flow for managing resources.

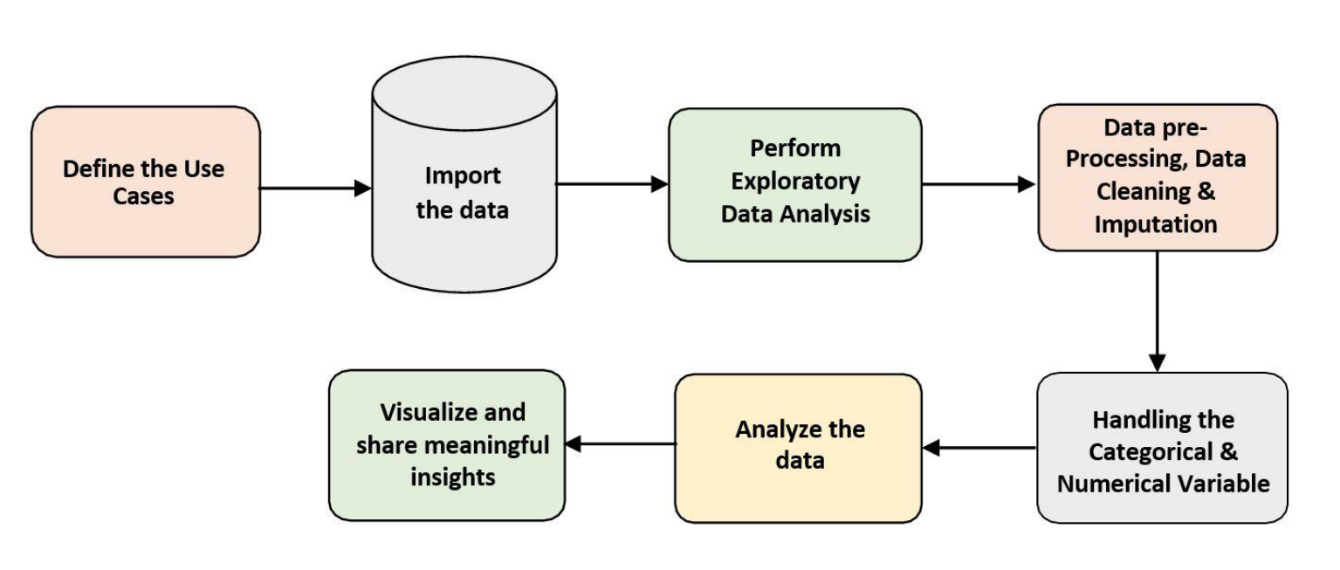
## 1.4 Ethics

Movie information is the only part that may have ethics problem. However, all the information we get for research is from public database such as Wikipedia and own public database. So there are no data confidentiality and user privacy problems.

## 1.5 Delimitations

In this project, we will not deep in how to collect and generate the data, which is the business of other teams . In the validation part, the data is divided into training and testing parts. But how to split data will not presented in detail neither.

Chapter 2

**Architecture**

**Fig. 1 Architecture**

**2.1 Architecture Description –**

**2.1.1. Data Description –**

In this analysis project, our listings dataset 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.

**2.1.2 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.

**2.1.3 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.How to calculate the similarity between two movies.

## Graphical user interface, text, application, email Description automatically generated

Fig 2. import data set

**2.1.4 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 prep module to automate our EDA process.
* It provides the following information:
* Overview: detect the types of columns in a Data Frame.
* 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.

## 2.1.5 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.

**2.1.6 Analyse 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.

Chapter 3

**Technology Used**

**3.1.Python**

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

**3.2. Jupyter Notebook**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Fig 3. Technology used

**3.3. Pandas**

Pandas is a Python library for data analysis. **Started by Wes McKinney in 2008** out of a need for a powerful and flexible quantitative analysis tool, pandas has grown into one of the most popular Python libraries.

**3.4. NumPy**

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. **NumPy was created in 2005 by Travis Oliphant**. It is an open source project and you can use it freely.

**3.5. Matplotlib**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Chapter 4

**4.1 General overview of dataset.**

❖ This dataset has around is mix between categorical and numeric values.

❖ Price is a dependent column.

❖ Total 16 columns are present in the dataset.

❖ Total observations are 48895.

❖ Min of Price variable is 0, max is 10000$.

❖ Mean price is 152$

❖ On an average people stay 7 days in a room.

❖ 75 Percentage of times minimum nights stayed is 5.

❖ Mean reviews given to Room/apartment is 23

**4.2 Understanding the Data:**

* There are 49,000 observations with various types of field in our dataset.
* List of field:
* Id
* Name
* Host\_id
* Host\_name
* Neighbourhood\_group
* Neighborhood
* Latitude
* Longitude
* Room\_type
* Price
* Minimum\_nights
* Number\_of\_reviews
* Last\_review
* Reviews\_per\_month
* Calculated\_host\_listing\_count
* availabilty\_365

**4.3 Missing value Handling:**

* ‘name’ = 16 null values.
* ‘host\_name’ = 21
* ‘last\_review’ = 10052
* ‘Reviews\_per\_month’ = 10052
* ‘name’ column we will replace the ‘nan’ values with corresponding ‘room\_type’ values.
* ‘host\_name’ will not use as those are names of individuals.
* 'reviews\_per\_month' we will replace with 0 for NA values.
* In ‘last\_review’ We will convert its data type to categorical and replace 'NA' with 'never'.

# **4.4 Overview of Related Work**

* Top 25 most common words used in listing names
* Top 25 Used Words for Listing Names
* Neighbourhood group Frequency.
* Top 15 highest listing Neighborhood.
* Total count of each room types as per listing.
* Minimum no’s of nights stayed and preferred rooms
* Value count of Neighborhood.
* Area vs Number of reviews
* Room types and their relation with availability in different neighborhood groups? (contd.)
* Price vs Number of Reviews
* Find total no. nights spend as per location
* Challenges Faced
* We defined some points which can help Airbnb in their business
* Conclusion

## 4.4.1 Top 25 most common words used in listing names

* As we can see most of the listing names include words related to property type such as ‘bedroom’, ‘cozy’, ‘private’, ‘apartment’ and ‘spacious’.
* It is interesting to see that words related to proximity or connection to public places such as ‘park’, ‘near’, ‘village’ and ‘heart’ rank lower in chart

A picture containing graphical user interface

Description automatically generated

Fig 4. Top 25 most common words used in listing names

**4.4.2 Top 25 Used Words for Listing Names:**

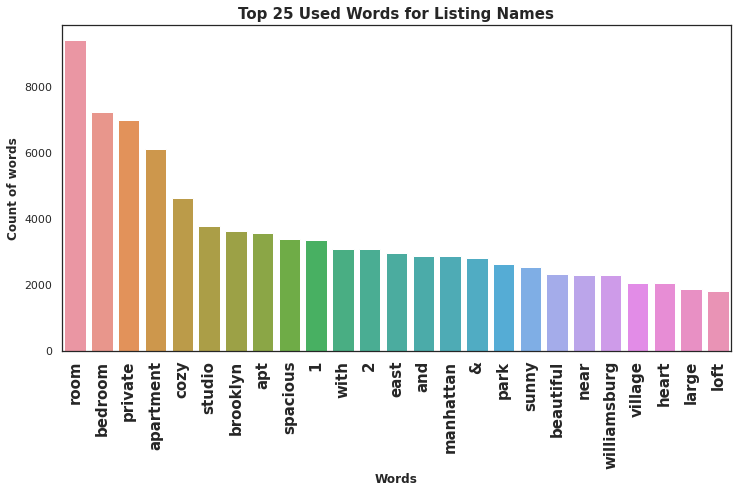
****

Fig 5. Words count

**4.4.3 Neighbourhood group Frequency:**

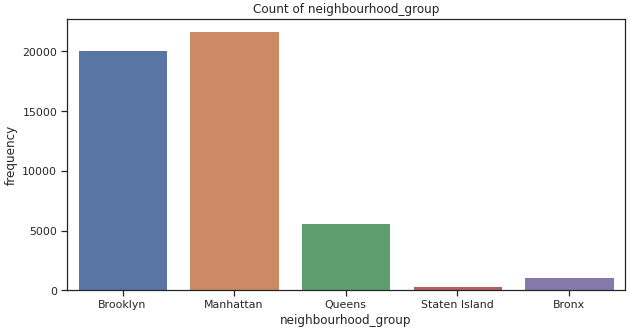


Fig 6. Neighbourhood group Frequency count

**4.4.4 Top 15 highest listing Neighbourhood:**

* Here we can state that Williamsburg has the highest listing.
* And almost same in Bedford-Stuyvesant.
* it’s a tourist attraction hence the number of listings could be highe

Chart, bar chart

Description automatically generated

Fig 7. Top 15 highest listing Neighbourhood:

**4.4.5 Total count of each room types as per listing:**

* Entire home/apt has the highest number of listing of 52% among other room.
* Private room has 45.7% of listing among other room types.
* Shared Room is the least listed room type at only 2.4% in total.

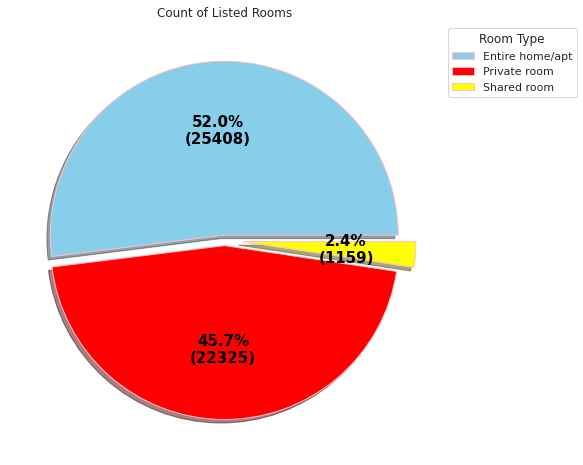


Fig 8. Count of listed rooms

**4.4.6 Minimum no’s of nights stayed and preferred rooms:**

Entire home/apt has more than 50% proportion in New York city and it too has highest avg price also. Shared room are the cheapest, but only has 2.4% proportion. No wonder New York life is of high standard.

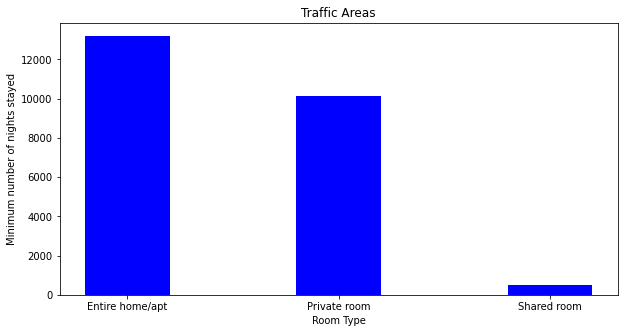


Fig 9. Minimum no’s of nights stayed and preferred rooms

**4.4.7 Value count of Neighbourhood:**

Chart, bar chart

Description automatically generated

Fig 10. Value count of Neighborhood

1.There are three types of rooms

2. Namely:

  1.Private room

  2.Entire home/apt room type

3.Shared room.

3.People mostly preferred to take whole apartment on rent followed by Private room.

4. very few people preferred to have shared rooms.

5. We will try to categories the 'price' like cheap, affordable and expensive and then analyze the 'price' for 'room\_type' as per it.

**4.4.8 Area vs Number of reviews:**

Chart, bar chart

Description automatically generated

Fig 11. Area vs Number of reviews

**4.4.9 Room types and their Relation with availability in different neighbourhood groups? (contd.)**

* Brooklyn and Manhattan have the least availability of rooms overall, as low as 0 days.
* Staten Island and Bronx has the highest availability rate overall at around 300 days.
* Form this analysis we can say that people stay for longer duration of time in Private rooms in Brooklyn and Manhattan.

Chart, bar chart, box and whisker chart

Description automatically generated

Fig 12.Relation between neighbourhood group and availability per day

**4.4.10 Find total no. nights spend as per location :**

* The locations are categorized based on its total number of minimum nights spend by customer.
* From here we can get an overall idea of which location the customer prefers more.
* We can state that more customers prefer Manhattan and Brooklyn for night stay as compared to other locations

Chart, waterfall chart

Description automatically generated

Fig 13. total no. nights spend as per location

Chapter 5

**5.1 Challenges Faced:**

* Reading the dataset and understanding the meaning of some columns.
* For answering some of the questions we had to understand
* the business model of Airbnb that how they work.
* Handling Nan values, null values and duplicates.
* Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader

**5.2 We defined some points which can help Airbnb in their business:**

* Manhattan is the most focused place in New York for hosts to do their business
* Customers pay highest amount in Brooklyn, Queens and Manhattan that is $10,000 and lowest amount is $10.
* For the three types of room type (i.e. Entire home, Shared room, & Private room) average price of entire home is around $157, for Shared room is around $60, and for private room is around $75.
* Top three host base on their turnover are Sonder(nyc),Red awning, Henry and best host is Sonder(NYC).

**5.3 Limitation:**

Though the dataset is very feature rich and shares less correlation and contains enough sample to perform regression on price prediction, but the correlation with target price is also low. So, it will result as high MSE. Also, the features dataset provides in terms of modern world, are of very poor quality in deciding the valuation of a property. Also, as the features are positively skewed, we need to treat it before prediction.

In order to have a better analysis regarding the quality of the properties, it would be interesting if we had an analysis of sentiments with property valuations.

User ratings of hosts aren’t available, it would’ve been better to rank our hosts based on user satisfaction and ratings. Also, in those cases further analysis can be done to view how guests tends to rate in terms of price or room type or is rating decides the valuation of property. Normally a low rated property tends to lower their price.

The exact number of guests counts also missing; it is just assumed that the guests by col: last\_review. A new host may never have been rated, that doesn’t mean no guest has ever stayed there.

**5.4 Scope of Improvement:**

As dataset has few qualifying attributes to value a property, more features can be added like bedroom, bathroom, property age (it might be one of the most important one), tax rate applicable, distance to nearest airport, hospital or schools.

In presence of ratings, hosts can be classified and ranked, special discount or offer can be given to highest rated hosts following marketing strategy.

Time series analysis can be done to make prediction on occupancy rate based on tourist season.

**5.5 Conclusion:**

From the entire analysis, it can be concluded that,

* Most visitors don’t prefer shared rooms, they tend to visit private room or entire home.
* Manhattan and Brooklyn are the two distinguished, expensive & posh areas of NY
* Some properties are having Minimum Nights to stay is more than 365 Days which can be favorable among Students, Low-Income Employees & Immigrants.
* Though location of property has high relation on deciding its price, but a property in popular location doesn’t mean it will stay occupied in most of the time.
* Performing a regression on this dataset may result in high error rate, as the features given in this dataset, are of very poor quality in deciding the property valuation. We can see this by looking at correlation heatmap. We would need more features like bedrooms, bathroom, property age (guessed it’d be a very important one), tax rate applicable on land, room extra amenities, distance to nearest hospital, stores or schools. These features might have a high relation with price.
* Time series analysis is possible to make prediction related to occupancy rate at particular time of a month, or particular time of a season.
* It’d be a better if we had avg guest ratings of a property, that would be beneficial in understanding the property more and could also be a factor in deciding price. A low rated property tends to lower their price.