Architecture

Travel Data Analysis

(AirBnB Data Analysis)

Written By / Author	Lokesh Attarde
Document Version	LLD-V1.0
Last Revised Date	15/10/2021

Document Version Control:

Date	Version	Author	Comments
15/10/2021	V1.0	Lokesh Attarde	First Draft

Approval Status:

Version	Review Data	Reviewed By	Approved By	Comments
V1.0				

Contents

1 Introduction

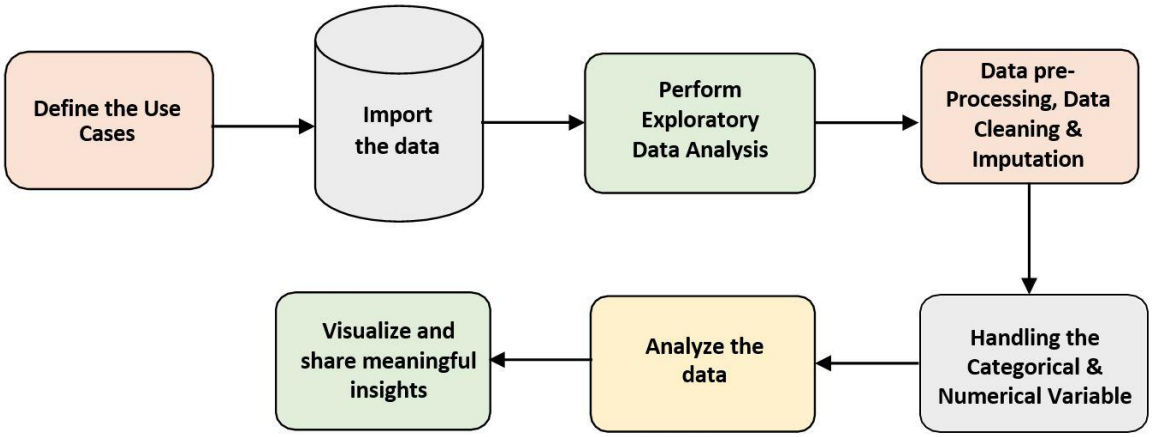
1.1 Why this Architecture design document?

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.

1.2 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.

2 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.

Reading Data

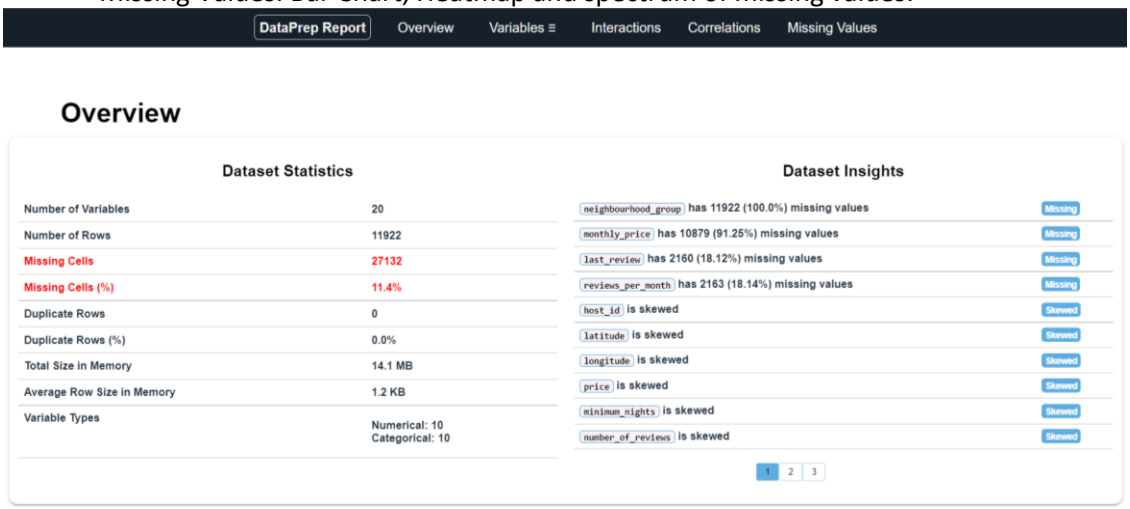
```
In [2]: df_Listings = pd.read_csv('listings.csv')
        df_Listings.head()
```

Out[2]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	city	latitude	longitude	property_type	room_type	
0	6	Large Craftsmen w/ yard ~ Perfect for families	29	Sara	NaN	North Hills	San Diego	32.753993	-117.129705	House	Entire home/apt	
1	5570	Ocean front condo on the sand	8435	Jef Karchin'S MISSIONBEACHRETREAT	NaN	Mission Bay	San Diego	32.784304	-117.252578	Condominium	Entire home/apt	TV
2	8095	Sunset Cliffs Studio	270	Marin	NaN	Ocean Beach	San Diego	32.735170	-117.243793	Guesthouse	Entire home/apt	
3	39516	Art Studio Retreat/Rustic Cabin	169649	Chris And Jean	NaN	North Hills	San Diego	32.731884	-117.119180	Tiny house	Entire home/apt	{Inte
4	45429	OB cottage SD--view on waterway	197919	Melissa	NaN	Loma Portal	San Diego	32.748768	-117.229371	House	Entire home/apt	{T

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 gives you a better Understanding of the "Variables" and the "Relationships" between them.
- Here, we make use of dataprep 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.



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.

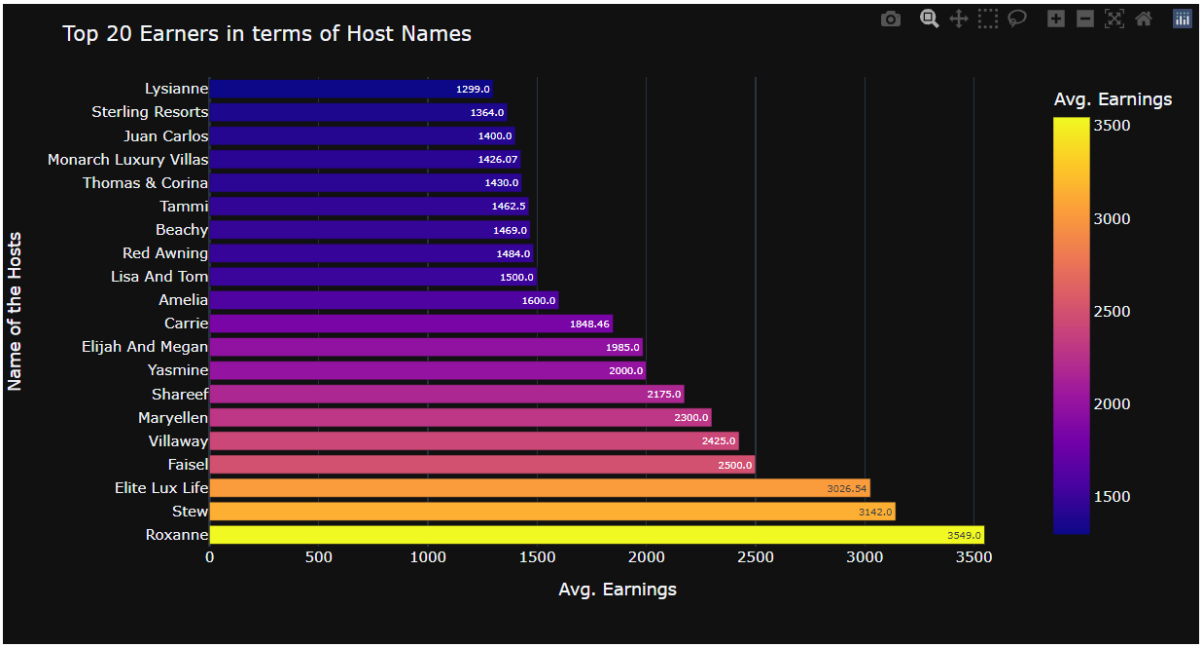
2.1.6 Analyse the Data –

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.

2.1.7 Visualize & Share Meaningful Insights –

Finally, it's 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 –



All those 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.