Architecture

Travel Data Analysis

(AirBNB Data Analysis)

Written By / Author	Shivank Singh
Document Version	LLD-V1.0
Last Revised Date	07/10/2022

Document Version Control:

Date	Version	Author	Comments
07/10/2022	V1.0	Shivank Singh	First Draft

Approval Status:

Version	Review Data	Reviewed By	Approved By	Comments
V1.0				

Contents

1 Introduction

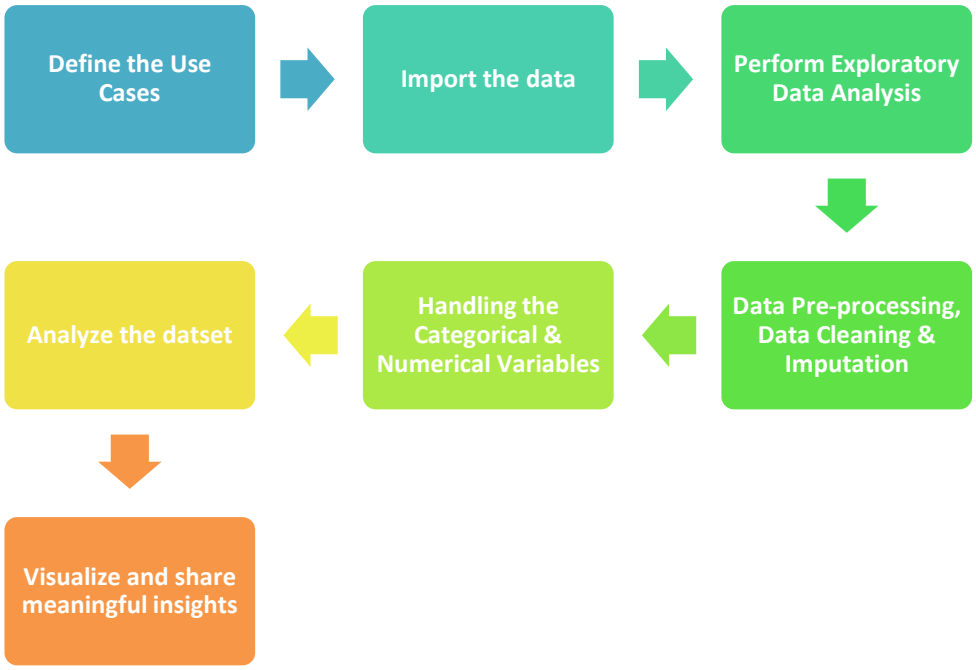
1.1 Why this Architecture design document?

The purpose of this document is to offer a detailed architecture design of the Airbnb Data Analysis Project by means of that specialize in each of the attributes of our structure. This report will cope with the background of this assignment, and the architecturally giant feature requirements. The intension of this record is to help the improvement crew to decide how the device can be dependent at the highest stage.

1.2 Scope

Architecture Design Document (ADD) is an architecture layout procedure that follows a step-by using-step refinement process. The manner may be used for designing records systems, required software program architecture, source code and ultimately, overall performance algorithms. Overall, the design principles may be defined at some stage in requirement evaluation after which delicate for the duration of architectural layout paintings.

2 Architecture



2.1 Architecture Description –

2.1.1 Data Description –

In this analysis project, our airbnb dataset have around 3.74 Lacs of records with 20 different features. Features are distributed as 11 Continuous features and 9 Categorical features and in our reviews dataset, these datasets are given in the form of Comma Separated Value (.csv) format having name airbnb price.

2.1.2 Define the Use Cases –

At this stage, primarily based on the given dataset and business issues we have described the numerous Use Cases to carry out the evaluation on and this can absolutely help out get the important thing insights from this information primarily based on which enterprise selections can be taken. Furthermore, It helps in not best understanding the significant relationships between attributes however it also allows us to do our personal research and are available-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 CSV FILE

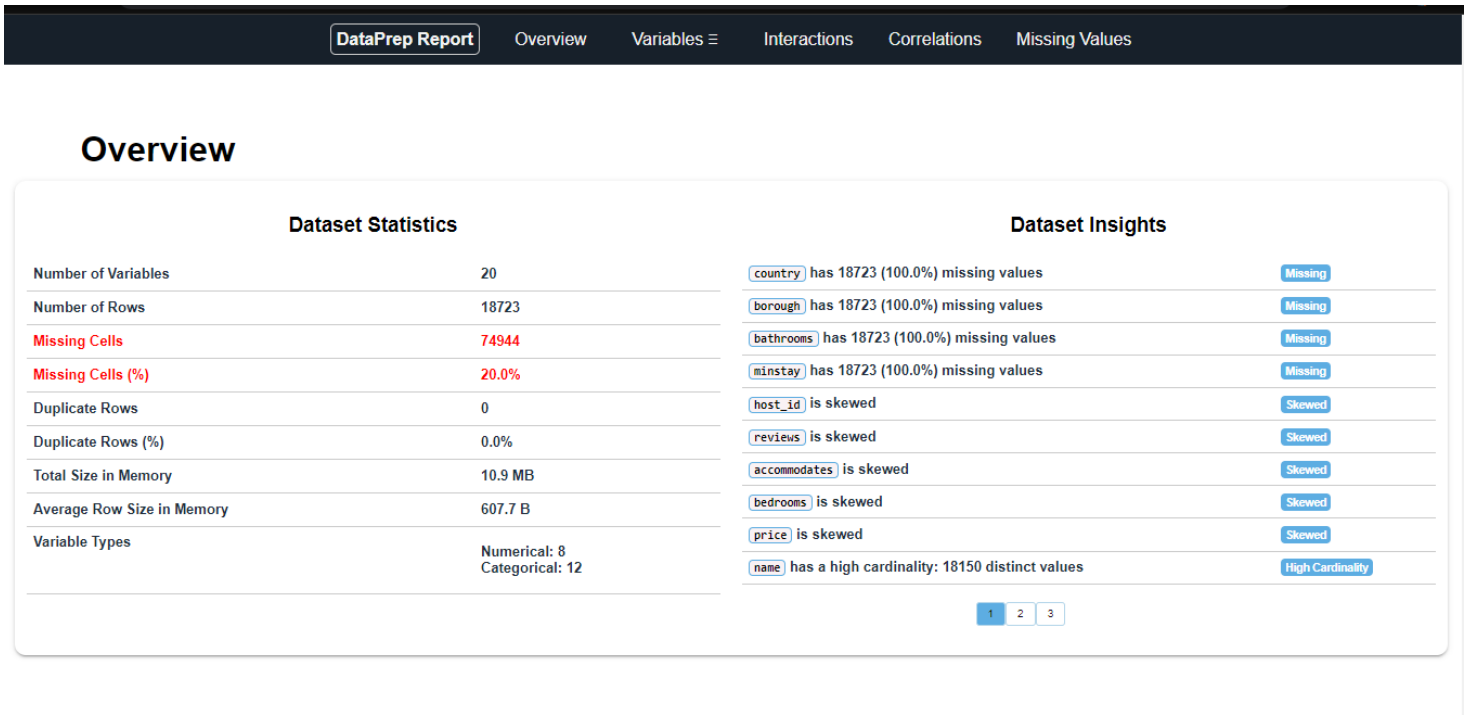
```
In [2]: df_airbnb = pd.read_csv('airbnb prices.csv')

In [3]: df_airbnb.shape
Out[3]: (18723, 20)

In [4]: df_airbnb.head()
Out[4]:
```

	room_id	survey_id	host_id	room_type	country	city	borough	neighborhood	reviews	overall_satisfaction	accommodates	bedrooms	bathroor
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

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 –

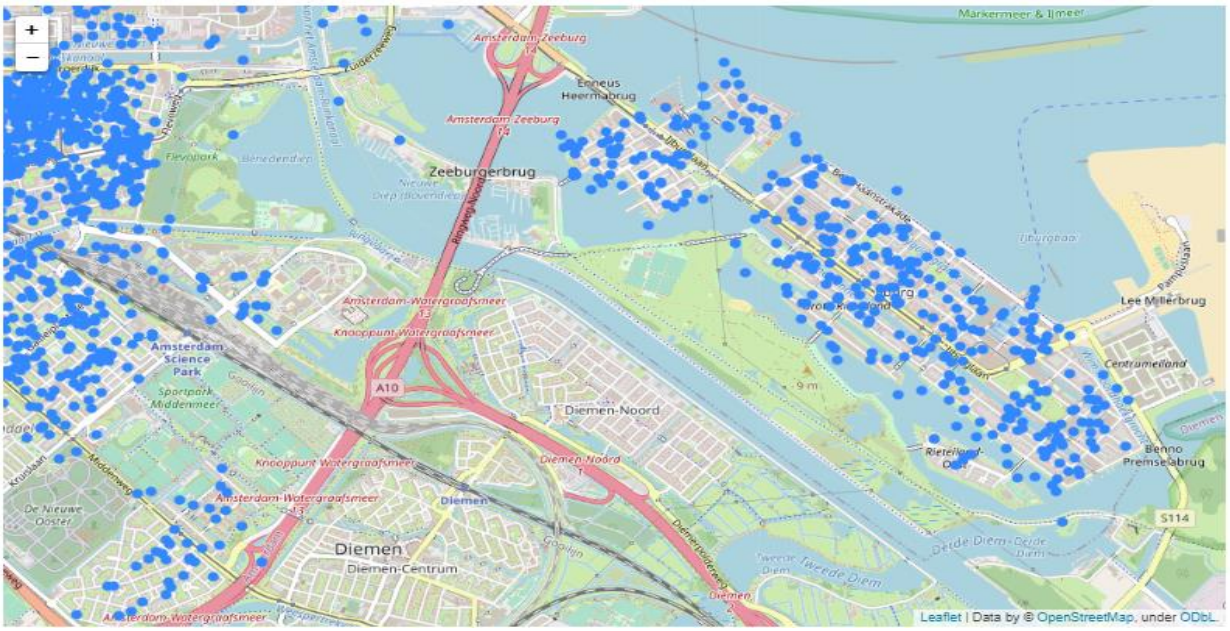
Once the pre-processing is done, we are good to go with our actual analysis where we write lines of codes and logics to prepare our data as per the defined use cases.

2.1.7 Visualize & Share Meaningful Insights –

Last but not the least , it is time to turn our data into some sort of visual representation. In short, Data visualizationis 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 two of our visuals are –

(1)



(2)

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