Data Analysis and Data Visualization: A Case Study on Analysis of Best Home/Hotel Dataset

#1 Archies M. Shedge, #2 Prof. Sagar Kulkarni

#1 Student, PCACS, New Panvel, #2 Pillai College of Engineering, New Panvel

ABSTRACT

The agenda of this analysis is to find the best hotel or apartment based on Price, Reviews, and Ratings by using various techniques like Python, SQL, and data visualization tool MS Power BI. The dataset consists of 22,000 rows and 90 columns. The system uses UiPath Studio to get the data from the source and load it to the destination with the help of robotic process automation and used SQL operations to clean and enrich the data and also, used Python for analyzing the data and cleaning the customer feedback data. The system will detail the analysis to the questions of interest gain preliminary insights through exploratory data analysis and visualization. In addition to that, the system has covered spatial data to perform various variables from our dataset using spatial visualizations and has answered questions relating to variations in prices and ratings across different locations in Berlin. The spatial data has helped to form different clusters from hight to low rental prices and high to low ratings. The system would be helpful to make key business decisions and would act as a time-saver for the stakeholders and it can be used in various domains.

Keywords: Exploratory Data Analysis, Data Visualization, ETL, Power BI, Automation, Python, SQL.

I. INTRODUCTION

Dataset:

The system gathered the dataset from Kaggle (Berlin Airbnb Data | Kaggle), the dataset contains 22,000 rows and 96 columns. The system has performed data cleaning operations with the help of SQL Server and Python.

Once the data is cleaned, it would be further sent to data enrichment. Data enrichment contains removing blank spaces, removing unwanted characters, switching every character into lowercase, and replacing null or blank values.

Robotic Process Automation:

Once we are ready with the cleaned data, the system would be loading it into the data visualization tool using **UiPath Automation** tool. With the help of UiPath Studio, we can load the data automatically from MS Excel Spreadsheet to SSMS.

Whenever we make any changes in the existing spreadsheet, we do not need to manually load it into the SSMS, we can mitigate such steps using robotic process automation.

Exploratory Data Analysis:

Now that we are ready with our cleaned and transformed data, we are going to perform analysis on it and make reports out of it for better business decision-making. The system uses the MS Power BI tool for data visualization and using Python and SQL for EDA.

We can find answers to the following questions by using our data analysis techniques:

- What are the rental fee and cleaning fee offered by the host?
- How much security deposit compared with the rental fee you must give to the host?
- What are the most popular areas and highly-priced areas?
- Happy or Not Happy percent distribution using matplotlib.
- Get host details, their profile picture, room picture, room cleanliness rating in one go by using a custom tooltip.
- What are the different types of properties and beds available?
- Advantages of being a super-host.
- Do regular hosts and super hosts have different cancellation policies?
- Finding positive and negative ratings of feedback provided by the user.

II. PROBLEM STATEMENT

TABULAR FORMAT:

The existing decision-making is based on tabular and static chart formats. The problem with the tabular format is, we cannot fully understand the KPIs (key performing insights). The values shown in the table are on point yet unable to quickly grasp and make decisions out of it.

The tabular format will keep values in front of you but you cannot find what was the last six months or three months or three years revenue within a fraction of a second.

STATIC CHARTS:

Another problem of decision-making with the help of charts is presenting the data with static charts. Static charts are a good way to present the data insights but it is good until the stakeholders ask it on weekly basis.

What if stakeholders or management want numerous KPIs within a few seconds? Certainly, nobody could create a new chart within a few moments and present them flawlessly.

III. PROPOSED METHODOLOGY

The proposed methodology can simply remove existing problems by using data visualization tools like MS Power BI, Tableau, Google Analytics, and much more. With the help of MS Power BI, we can make such visuals where stakeholders can view the charts and insights dynamically.

The time require to make business decisions is minimized and would give effective outcomes flawlessly. In addition to that, we can categorize any feedback given by the user into two types Happy or Unhappy. This semantic analysis module will help stakeholders to understand what the particular customer is thinking about the property without even reading the full review.

The Python libraries would help us to reach our goal and they can even give an output to the live comment given by the user. It saves a lot of time as management does not need to read the full text given by the user. In both practices we can save a lot of time, and as we know time is money so, eventually we are saving a lot of money.

IV. EXPLORATORY DATA ANALYSIS

Preparing Python for EDA of Feedbacks

- Calculating Percentage and Count of Missing Values.

```
Count and Percentage of Missing Values:
```

	Count	Percentage
id	0	0.0
host_id	0	0.0
Feedback	0	0.0
Is_Response	0	0.0
User_ID	0	0.0
Browser_Used	0	0.0
Device_Used	0	0.0

Fig 1. Miss Values in %

- Using Matplotlib to represent Percentage Distribution by Review Type i.e Happy or Not Happy.

Preparing SQL Server Management Studio (SSMS) for EDA

Updating the NULL values with the help of the following queries. Updating NULL values from zipcode, square_feet, bedrooms, bathrooms, cleaning_fee, and many more columns.

```
Percentage for default

happy        68.19
not happy    31.81
Name: Is_Response, dtype: float64
```

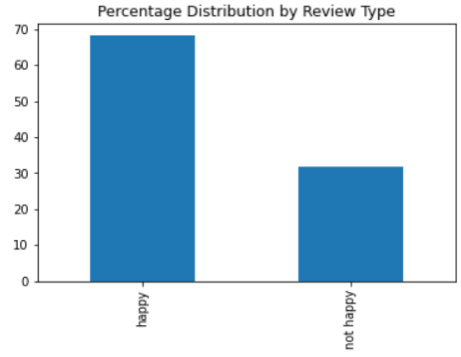

Fig 2. Review Type Distribution in %.

V. RESULT ANALYSIS

Data Visualization in MS Power BI Desktop

In the given report, we can observe the price comparison of Room Rent (Price) V/s Cleaning Fee by Host ID. It also uses a timebrush third-party visual for custom time slicing. Within the single report, we can switch between 'Price & Cleaning Fee' and 'Price &
Security Deposit' using the toggle switch button.

We have the host's full details that are, Profile Picture, Room Picture, Cleanliness Rating, and Maximum Accommodates Allowed per room and below the table, we can check for the highly-priced area as well as highly popular area. We can even filter out a few options like - How many beds we want, Price of it, City, and the Date. If we want to reset everything then we can do that too by clicking on the "clear filter" button present at the bottom.

Fig 3. Price V/s Cleaning Fee comparison

Fig 4. Price V/s Security Deposit comparison

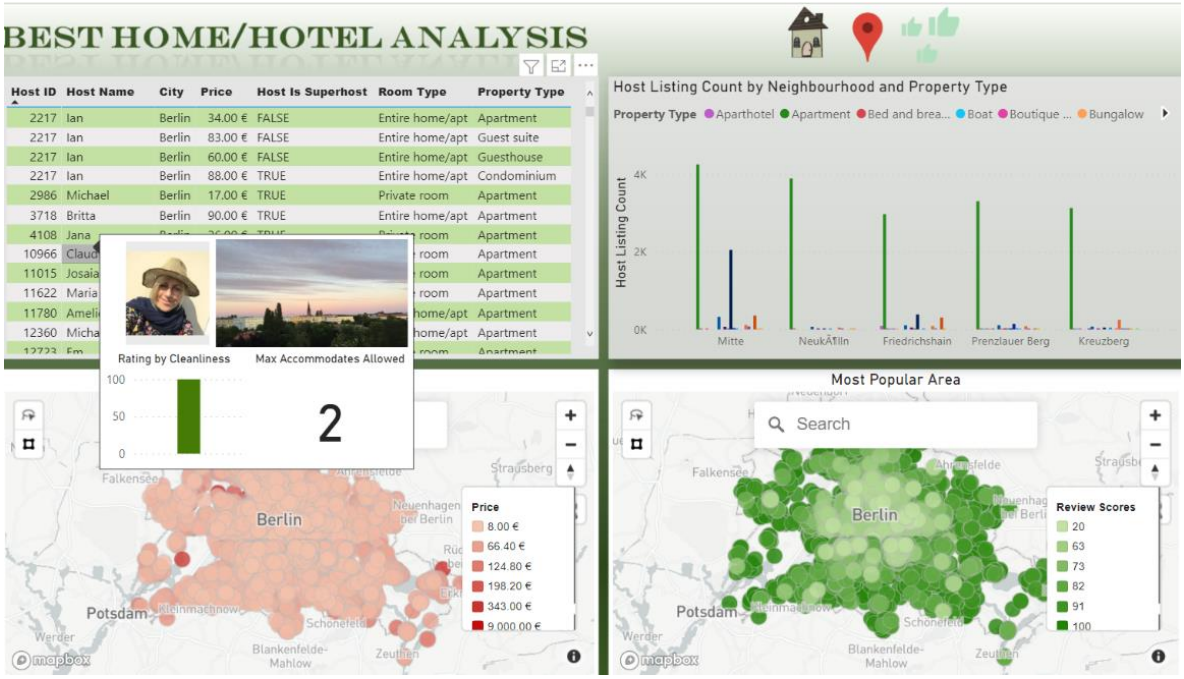

Fig 5. Host's details and global popularity rate

Fig 6. Filters for amenities

VI. CONCLUSION AND FUTURE SCOPE

Through this exploratory data analysis and data visualization project, we observed several outcomes and solutions with the help of Power BI, Python, and SQL. With the help of the proposed methodology, we can analyze the data more clearly and in a precise way. The data shown with some visualizations and reports are always better than the tabular format.

This analysis would help stakeholders or management of the AIRBNB to help understand what are the best rental places according to the customer, what are the customer's perception about the property.

The time require to make business decisions is minimized and would give effective outcomes flawlessly. In addition to that, we can categorize any feedback given by the user into two types Happy or Unhappy. The Python libraries would help us to reach our goal and they can even give an output to the live comment given by the user. It saves a lot of time as management does not need to read the full review given by the user. Based on the user's data we can analyze the business requirements and current trends. Data is everywhere and we can make use of the existing data to predict the future outcome and make better business decisions.

The dataset contains 22,000 rows and 90 columns and we may think this much data is sufficient but on the contrary, the big data consist of terabytes and even petabytes. In addition to that, the system has the data of Berlin and we must maximize the scope by capturing the data of the whole Germany and analyzing it accordingly. This would help us understand the business needs on a high scale. The reviews given by the user were in English language only, our system can perfectly analyze the sentiments based on the given inputs in English language but what if a user is unaware of the English and want to give feedback in German or Deutsch? The existing system would fail to analyze the sentiment as it was not trained in any other language than English. We can work on it in the future to perform sentiment analysis based on multiple languages.

VII. REFERANCES

[1] Chong Hu Yo, Azusa Pacific University - Exploratory Data Analysis in the Context of Data Mining and Resampling.

[2] Jack H Zheng, Kennesaw State University - Data Visualization for Business Intelligence

[3] Dr. Rohit Vishal Kumar, International Mgmt. Institute Bhubaneshwar - Exploratory Data Analysis Using R & RStudio

[4] Kiranbala Nongthombam and Deepika Sharma, Chandigarh University, Panjab - Data Analysis Using Python

[5] Nikos Bikakis NTU Athens & ATHENA R.C. Greece, Timos Sellis Swinburne Univ. of Technology Australia - Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art

[6] Nascif A. Abousalh-Neto and Sumeyye Kazgan, SAS Institute - Big Data Exploration through Visual Analytics

[7] Katarina Košmelj , Andrej Blejec , and Drago Kompan, Developments in Applied Statistics - Exploratory Data Analysis as an Efficient Tool for Statistical Analysis: A Case Study From Analysis of Experiments

[8] Mandava Geetha Bhargava, K. Tara Phani Surya Kiran & Duvvada Rajeswara Rao, Koneru Lakshmaiah Educational Foundation - Analysis and Design of Visualization of Educational Institution Database using Power BI Tool

[9] EDA and Visualization of Hotel Booking Project in Python – Towards AI — The Best of Tech, Science, and Engineering

[10] Conduct and report on exploratory data analysis (EDA) of housing. | PAPERS HOST

[11] Hospitality Industry Research Topics & Ideas 2021 (myresearchtopics.com)