ANALYZING AIRBNB DATASET FOR NYC: APPROACH AND METHODOLOGY

1. Introduction:

1.1. The methodology document provides a detailed overview of the analytical approach used to analyze the Airbnb dataset. It outlines the step-by-step process followed, from data preparation and cleaning to exploratory data analysis and modeling techniques, to derive meaningful insights.

1.2. The document ensures transparency and reproducibility by showcasing the methods, techniques, and tools employed in the analysis. It serves as a valuable resource for stakeholders, enabling a clear understanding of the analytical process and the basis for the conclusions and recommendations presented in the presentations.

2. Background:

The COVID-19 pandemic has significantly impacted the travel industry, leading to a major decline in revenue for Airbnb. As travel restrictions are gradually lifted and people regain confidence in traveling, Airbnb aims to rebound and maximize revenue opportunities.

3. Problem Statement:

3.1. Address the revenue decline faced by Airbnb during the pandemic by leveraging data insights to identify key areas of improvement and revenue growth opportunities.

3.2. Airbnb wants to gain a deeper understanding of customer preferences, including room types, pricing ranges, and neighborhood preferences, to tailor its offerings and optimize the order of property listings for maximum traction and customer satisfaction.

4. Objective:

4.1. **Customer Insights and Revenue Optimization:** Our objective is to uncover valuable customer insights through room type analysis and understand their preferences and needs. By identifying high-demand neighborhoods for host acquisition and property expansion, we can offer personalized offerings. Furthermore, we aim to optimize pricing strategies based on booking distribution and price analysis to drive revenue growth for Airbnb.

4.2. **Strategic Property Acquisition and User Experience Enhancement:** The objective is to develop a comprehensive strategy for host acquisitions and operations that focuses on optimal property selection, negotiation, and adjustments to align with customer preferences. Additionally, we aim to enhance user experience by optimizing the property listing order in targeted neighborhoods to increase traction and improve customer satisfaction.

5. Data Collection:

The dataset used in this analysis was procured from the university only and *there were no additional data collection procedures involved*. It includes relevant information related

to Airbnb listings in New York, such as property details, booking information, customer preferences, and neighborhood data.

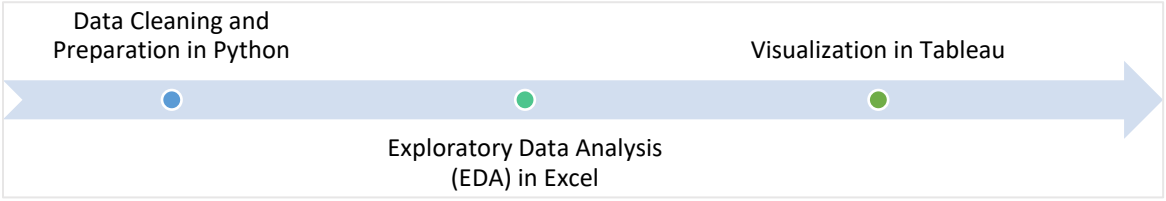

Fig: Sequential Approach

6. Data Cleaning and Preprocessing:

6.1. Data Loading and Initial Overview

- Imported necessary libraries and loaded the dataset using read_csv().
- Checked the original shape of the dataset and observed the number of rows and columns. It had 48895 rows and 16 columns. There are no duplicate rows in the data.
- Obtained an overview of the data by using info() and describe() functions, which provided information on data types, non-null values, and descriptive statistics of numeric columns.

6.2. Handling Missing Values:

- After checking null values, we found Columns 'last_review', 'reviews_per_month', host_name', 'name' have missing values.
- The missing values in 'reviews_per_month' are replaced with '0' which means no reviews were received which makes no missing values.
- Replaced the Nan values in new columns with 'Not reviewed' value, in turn handling all missing values.
- Removed the rows in 'host_name' and 'name' that has missing values as the missing value percentage is very less.

6.3. Data Transformation & Feature Engineering:

- Converted the 'last_review' column to datetime format and extracted three new columns 'review_day', 'review_month', 'review_month' and removed the actual column.

6.4. Outliers Treatment:

- Checked for any outliers in the numerical columns ['price', 'minimum_nights', 'number_of_reviews', 'reviews_per_month', 'calculated_host_listings_count' by using boxplot method.
- Calculated the 10th percentile (Q1) and 90th percentile (Q3) for each column.
- Used the interquartile range (IQR) calculated as Q3 - Q1.

- Removed outliers by selecting values within the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR.
- Reset the index of the Data Frame to maintain a sequential order of rows.

6.5. **Final Dataset Summary:** Descriptive statistics, including percentiles, are generated for the cleaned dataset.

6.6. **Exporting Cleaned Dataset:** The cleaned Data Frame is exported to a new CSV file using the to_csv() function.

7. Exploratory Data Analysis in Excel:

7.1. The dataset in *.csv* format was converted to *.xlsx* format for analysis due to the limited functionality of .csv files and the need to preserve formulas used in Excel for data manipulation and exploration.
7.2. Explored the cleaned dataset in Excel after preprocessing.
7.3. Utilized pivot tables to analyze the data and extract valuable insights.
7.4. Created various pivot tables to summarize data based on different attributes such as neighborhood, property type, pricing, and host-related information.
7.5. Calculated 2 new columns- “Min Price per Booking” and “Customer Segments”.
7.6. “Min Price per Booking” is calculated by multiplying “price” with “minimum_nights”.
7.7. “Customer Segments” was calculated based on booking percentage and avg. price of room type. (Refer graph 1.1 for more reference)
7.8. Analyzed key metrics such as average prices, occupancy rates, customer preferences, and host performance.
7.9. Examined the distribution of prices, customer reviews, and booking patterns through histograms, bar charts, and line graphs.
7.10. Identified popular neighborhoods, property types, and price ranges based on the data analysis.
7.11. Used the findings from the Excel analysis to generate informative visualizations and graphs in Tableau for further exploration and presentation.

8. Limitations and Assumptions:

8.1. The only assumption made in the analysis was regarding the total bookings of each property. Since the exact number of bookings was not provided, we considered the number of reviews as a proxy for bookings. It is important to note that not all bookings may have received reviews, but we assumed that the presence of reviews indicates that bookings were made for the respective properties.
8.2. As a result, throughout the analysis, the terms "number of reviews" and "number of bookings" have been used interchangeably, considering that the presence of reviews implies the occurrence of bookings. This assumption allows us to make inferences about the level of activity and popularity of each property based on the available review data.

ANALYZING AIRBNB DATASET FOR NYC: APPROACH AND METHODOLOGY

9. Methodology Document Conclusion:

9.1. The insights derived from the EDA in Excel and the visualizations generated in Tableau were then used to create a comprehensive PowerPoint presentation. The presentation included a clear and concise overview of the analysis, key findings, and actionable recommendations for the stakeholders.

9.2. The combination of Python, Excel, Tableau, and PowerPoint allowed for a robust analysis, ensuring that all aspects of the data were thoroughly explored and presented in a visually appealing and informative manner.

9.3. The findings and recommendations presented in the PowerPoint presentation were based on a rigorous analysis of the dataset, supported by visualizations and data-driven insights.

10. Appendix:

10.1. Original Data Source: Provides information on the sources of the data used in the analysis.

10.2. Data Dictionary: A comprehensive guide documenting the meaning, structure, and relationships of the data variables used in the analysis.

10.3. Python Code File: Python code that outlines the steps taken to clean, transform, and prepare the raw data for analysis.

10.4. Cleaned Data for EDA in Excel: Includes various calculations, charts, and insights derived from the data for further exploration and decision-making.

10.5. Tableau Analysis: Interactive visualizations and dashboards created in Tableau for data exploration and presentation purposes.