Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Shashwat Kumar Pani, shashwatpani1998@gmail.com, Individual Project.
Please paste the GitHub Repo link.
Github Link: - https://github.com/Shashwat-spyder/airbnb-data-analysis
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words) **Airbnb, Inc.** is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com. 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. This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

The dataset has 16 columns, namely-

ID: Booking ID
NAME: Residence Type
HOST_ID: Customer ID
HOST_NAME: Customer Name
NEIGHBOURHOOD_GROUP: Neighbourhood Group name
NEIGHBOURHOOD: neighbourhood name
LATITUDE: latitude
LONGITUDE: longitude
ROOM_TYPE: Type of room-- Private/Shared/Entire_home
PRICE: Price of home per night
MINIMUM_NIGHTS: minimum nights guest stayed
NUMBER_OF_REVIEWS: Count of cutomer revies
LAST_REVIEW: Date of last review submitted by customer
REVIEWS_PER_MONTH: average reviews per month
CALCULATED_HOST_LISTINGS_COUNT: number of time property is listed
AVAILABILITY_365: Number of days rooms available in a year

As the first step, drop the columns which are not necessary and may increase the complexity. It is important to use variables which provide information and patterns within data and are focused on the scope of project.

The data may contain some null values hence rechecking the data and handling it with imputed values. Mainly measures of central tendency is used in place of missing values.

Now all the steps required for Data Wrangling on the current dataset are finished. I have considered following columns for performing an exploratory data analysis:
'neighbourhood_group','host_id', 'host_name', 'neighbourhood', 'room_type','price', 'minimum_nights', 'number_of_reviews', 'reviews_per_month', 'calculated_host_listings_count' and 'availability_365.

Further, the project is divided into 4 parts.

In the first part, Graphical and statistical techniques are implemented to learn about different hosts and areas. The major columns taken for analysis are 'host_name''neighbourhood_group''neighbourhood'calculated_host_listings_count'. A group by operation is performed on former three columns w.r.t the latter one column. This gave a idea that host Sonder from NYC has most number of listing counts from Manhattan neighbourhood group. Also, the Manhattan area is most popular location among hosts.

In the second part, insights are drawn from prediction values like locations, prices, reviews etc. The following part is fulfilled by mainly focusing on price column and plotting it against different parameters.
A plot between price and neighbourhood group gives us the understanding of the costliest locations. Price and reviews reveal that most of the guest prefer cheap accommodations rather than luxury ones.

In the third part, the trends on host data are of concern. Number of reviews and room type columns are considered. A statistical table and a graph show that the demand for private rooms and entire home/apt are more and the respective host can be considered as busy.

Finally, the traffic among different areas is analysed. Use of the minimum night's stay count of guests and group by function, we can say that Manhattan and Brooklyn are visited by almost twice more guests that any other place.

The number of people staying for long run does not use review feature often, rooms are available at all locations round the clock and all types of room are available are some of more points that can be concluded.