Airbnb Booking Analysis

Problem:

The objective of this capstone is to do EDA on the given dataset and find out the insights from it.

In this dataset, we are provided with 16 features and around 49k data instances.

Now the basic thing to start with is exploring the features and finding out the meaningful insights from them.

I started with doing NA value handling, univariate analysis, multivariate analysis and conclusion.

Step 1: NA value replacement

In this dataset we are having 4 columns with 'NA' values namely:

1) Name column: this describes the information regarding the property.

2) Host_name: this describes the name of the host or we can say an individual person's name.

3) Last_review: this shows us the last date of the review.

4) Reviews_per_month: this shows the reviews got per month.

Name column we replaced with the corresponding value of the room_type column.

Host_name column we dint do anything.

Last_Review we replaced 'NA' with 0 value.

Reviews_per_month we converted it into categorical data type and replaced 'NA' with 'Never' string type value.

Step 2: univariate analysis

In this step, we started with a univariate analysis of individual features.

1) For the name column we generated word_cloud.

    Then we take a look at the top 50 common words based on their frequency of them this way we came to know which words are useful for keywords.

2) similarly we plotted the count plot for other features like neighbourhood, neighbourhood_group and found out which neighbourhood_group and neighbourhood is most popular in terms of staying and which are least preferred.

3) We plotted a scatter plot for latitude and longitude features this shows us the density of rooms/in each of the regions.

4) For the room type feature we plotted the count plot and looked at each type.

As this room_type is having categories we can do a similar type of segmentation in the price column and do our analysis, so we divided the price into a range of prices like from 0-80$ cheap category from 80-500$ affordable price range and above 500$ expensive. so affordable range was the most preferable category among the people followed by cheap and expensive price range.

5) From minimum_nights columns after analysis we concluded that people try to spend 1-4 days.

6) Number_of_reviews tells us that the average rating is 23 times.

7) Calculated_host_listing_counting tells how many times the host_id is listed this shows the most famous host and least famous host.

8) Last review column tells us that 75% of the times rating given is around 1.5-2. on rating scale.

Step 3: Multivariate Analysis

1) relation between neighbourhood_group and median price.

From this, we can say that Manhattan is having highest mean price and also high price properties are also available in this region followed by Brooklyn and Queens.

2) Relationship between neighbourhood and median price.

From this, we can say that Battery Park City id having the highest median price.

3)Relationship between price and room_type

From this, we can say that if the customer wants to book an entire apartment then definitely they have to pay more. followed by a private room and shared room.

4) Relationship between room_type and neighbourhood_group.

From this, we can say that Manhattan is having highest booking of entire apartments followed by Brooklyn and Queens.

Similarly for private rooms, Brooklyn is having highest booking followed by Manhattan and Queens.

5) Which neighbourhoods are generating maximum, minimum, revenues from room types are as follows:

- Entire_home/apartment.
  - Williamsburg is having maximum revenue from Entire_home/apartment. Which is around 389724$.
  - New_Drop is the least or having a minimum share of income from Entire_home/apartment.

- Maximum revenue from Private Room.
  - Williamsburg is having maximum revenue generation from Private Rooms around 171265$
  - Graniteville is having minimum revenue generation from Private Rooms around 20$.

- Maximum revenue from shared Room.
  - Hell's Kitchen is having maximum revenue generation from Private Rooms at around 9488$.
  - Randoll Manor is having minimum revenue generation from Private Rooms at around 13$.

Step 4: Conclusion

In this simple yet powerful way we had done the EDA on the Airbnb dataset. Certainly, this is not the end rather this this the start we can say as per business requirement changes we need to find the insights in that direction and justify the business problems. There can be n-number of questions and n-number of dimension to explore the dataset and find the insight from them, this there is no limit unless the business constrain is solved.

Contributing Member:

Name	Email
Bharath P	Bharath0924.bp@gmail.com