Contoso

INT 353 EDA PROJECT

BY RAJNISH BHARTI

REG. NO. – 12015883

ROLL NO. – B68

SECTION – K20RU

Contoso

HOTELS ON MAKEMYTRIP

- DIL TO ROAMING HAI

https://github.com/rajnish1602/EDA-Project-on-MMT-Dataset.git

Overview

- More over Indians are now booking tickets and hotels online than ever before.
- You can check out the prices and compare them to get the best out of the deal in MMT.
- A company that holds a major share in the Indian online travel industry is MakeMyTrip. Since 2000, MakeMyTrip is helping millions of Indians book railway tickets, airlines tickets, bus tickets, reserve hotel rooms, and buy holiday packages.
- Founded in 2000 and headquartered in Gurugram, Haryana, MakeMyTrip is one of the most popular and dependable Indian travel company that helps the people of India avail of online travel services that includes airline tickets, domestic and international holiday packages, reservations of hotels, railways, and bus tickets.

COMPANY HIGHLIGHTS

- Headquarters: Gurugram
- Founder: Deep Kalra, Keyur Joshi, Rajesh Magow, Sachin Bhatia (resigned as Co-founder and CMO)
- Sector: Travel tech
- Founded: 2000
- Total Funding: $748mm (March 2022)
- Website: https://www.makemytrip.com

make my trip
INDIA'S FAVOURITE HOMESTAYS Awards
RURAL HOMESTAY OF THE YEAR
otaram Prajapats Homestay, Jo

Contoso

WHY I CHOOSE THIS DATA SET ?

- Already In my Digital Marketing course I had research and analysis the marketing strategy of MakeMyTrip and Yatra which help me to take this project and work easily

- MakeMyTrip use some different strategy in comparison to other platform like e-wallet, free cancelation, .etc.,

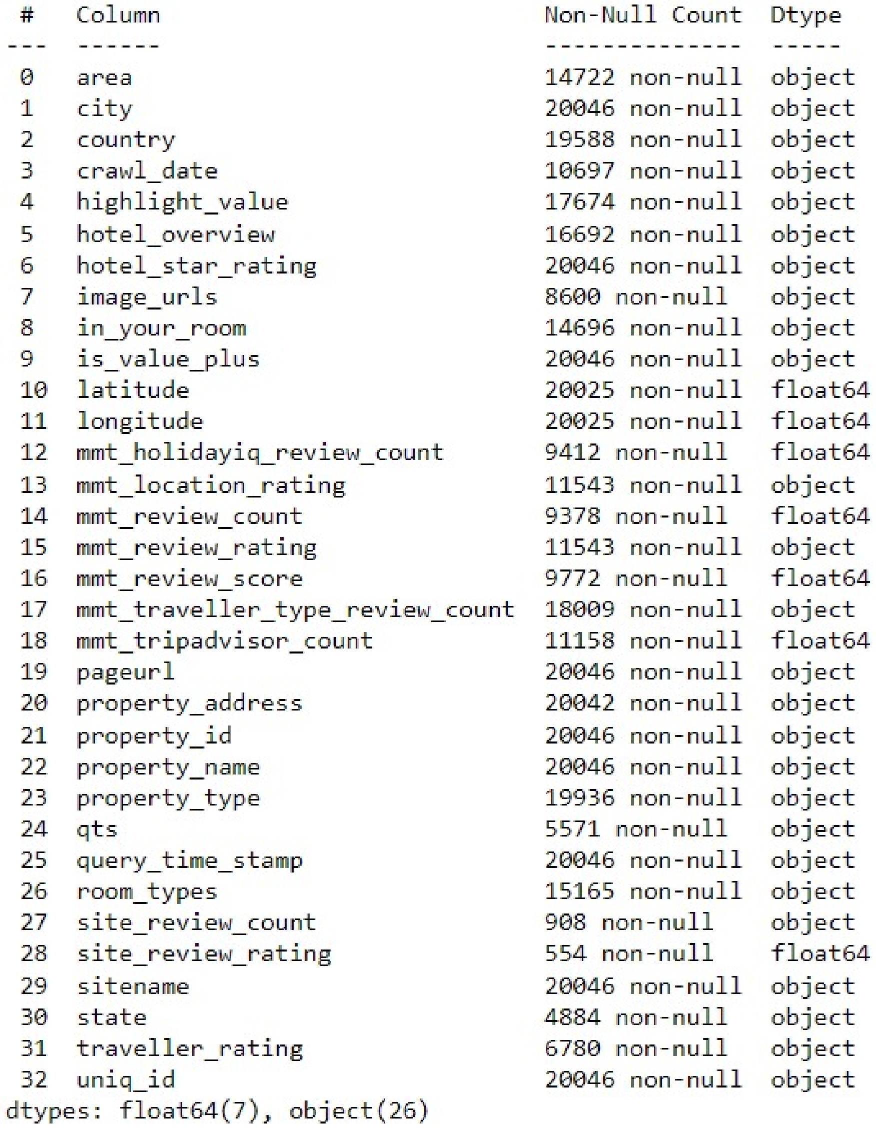
PROJECT GOALS

- Exploratory Data Analysis
- Understanding what type content is available in different hotels and its review
- No. of hotels present in different location.
- Hotel Business are too much volatile business they can be depends upon different factors like rating of the hotel, facilities, location and many more.
- The main objective behind this project to visualization and analyse MakeMyTrip dataset to discover various factors that govern the booking and insight the management so that they can perform various campaigns to boost the business.

DATASET INFO ?

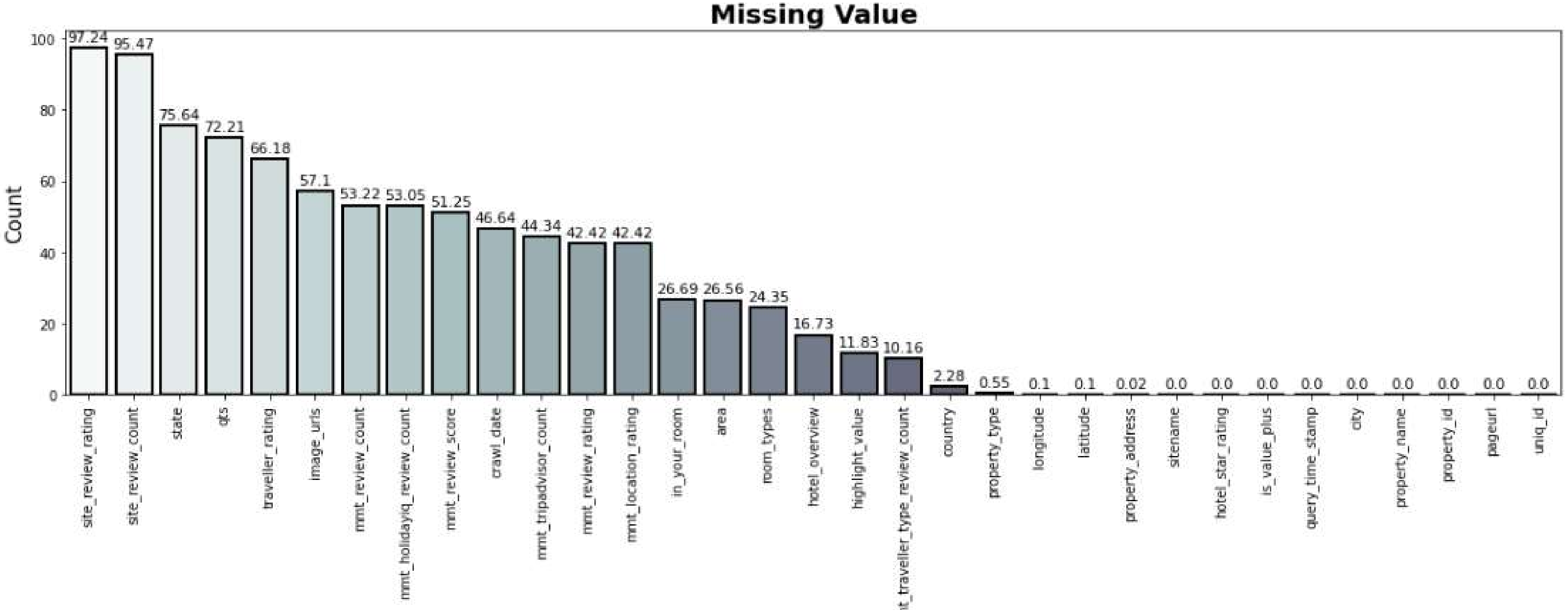
- There are 20046 rows and 33 columns in the dataset

  Here we can see that there are columns are of float type - 26 columns are of object type



```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   area                            14722 non-null  object
 1   city                            20046 non-null  object
 2   country                         19588 non-null  object
 3   crawl_date                      10697 non-null  object
 4   highlight_value                 17674 non-null  object
 5   hotel_overview                  16692 non-null  object
 6   hotel_star_rating               20046 non-null  object
 7   image_urls                      8600 non-null   object
 8   in_your_room                    14696 non-null  object
 9   is_value_plus                   20046 non-null  object
 10  latitude                        20025 non-null  float64
 11  longitude                       20025 non-null  float64
 12  mmt_holidayiq_review_count      9412 non-null   float64
 13  mmt_location_rating             11543 non-null  object
 14  mmt_review_count                9378 non-null   float64
 15  mmt_review_rating               11543 non-null  object
 16  mmt_review_score                9772 non-null   float64
 17  mmt_traveller_type_review_count 18009 non-null  object
 18  mmt_tripadvisor_count           11158 non-null  float64
 19  pageurl                         20046 non-null  object
 20  property_address                20042 non-null  object
 21  property_id                     20046 non-null  object
 22  property_name                   20046 non-null  object
 23  property_type                   19936 non-null  object
 24  qts                             5571 non-null   object
 25  query_time_stamp                20046 non-null  object
 26  room_types                      15165 non-null  object
 27  site_review_count               908 non-null    object
 28  site_review_rating              554 non-null    float64
 29  sitename                        20046 non-null  object
 30  state                           4884 non-null   object
 31  traveller_rating                6780 non-null   object
 32  uniq_id                         20046 non-null  object
dtypes: float64(7), object(26)
```

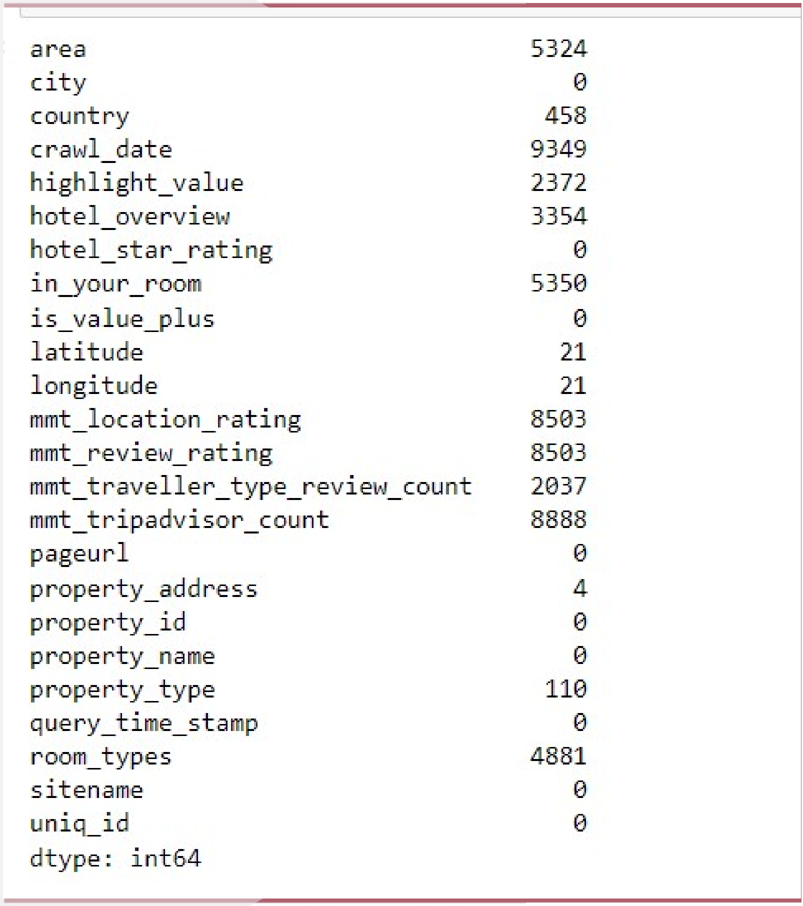
GRAPH FOR MISSING VALUES



•Hai contains max 97.24% missing values.
•We remove the column having null value more than 50 percent by taking threshold.
•And for rest column we treat the missing values by each columns.

Here we observe that,

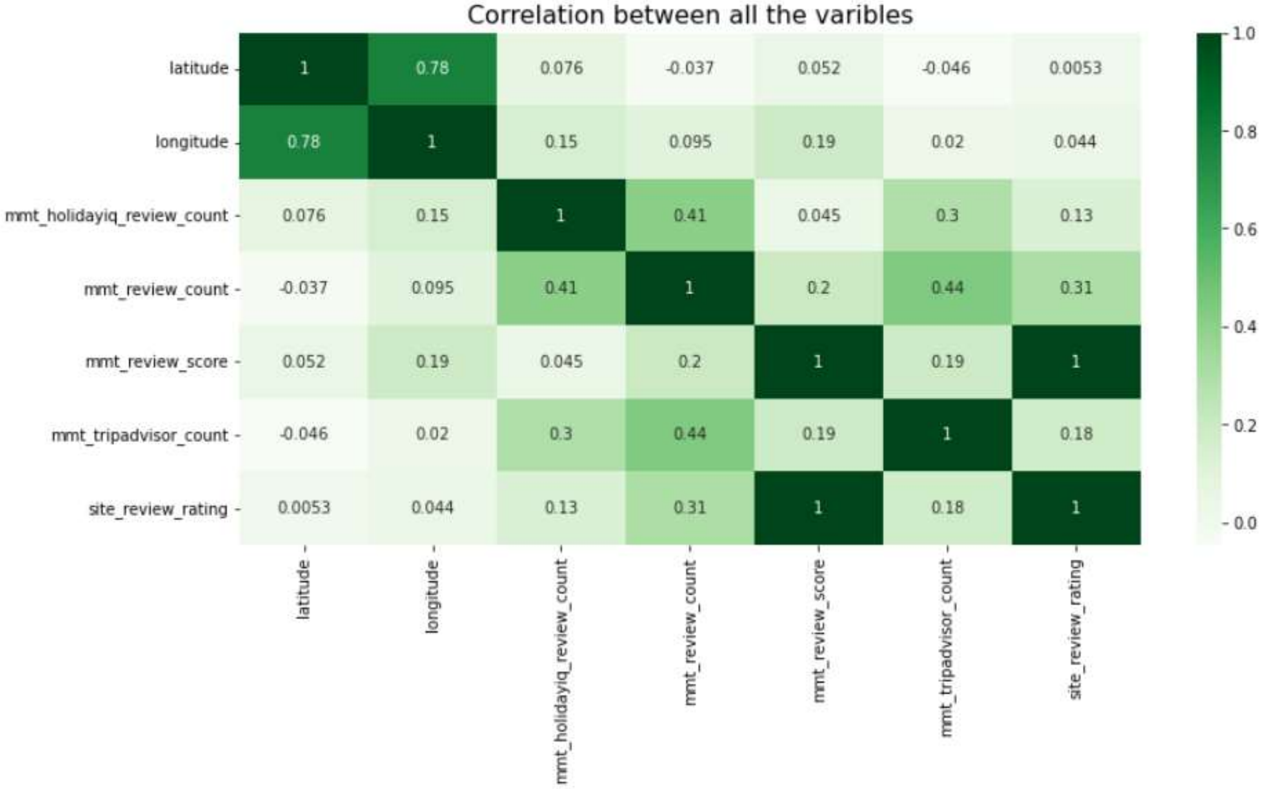
- area column having 5328 null values
- Country column having 458 null value
- And maximum null values in crawl_date column
- From here we can visualize the column one by one and clean the null values.
- Also we do further univariant and bivariant analysis



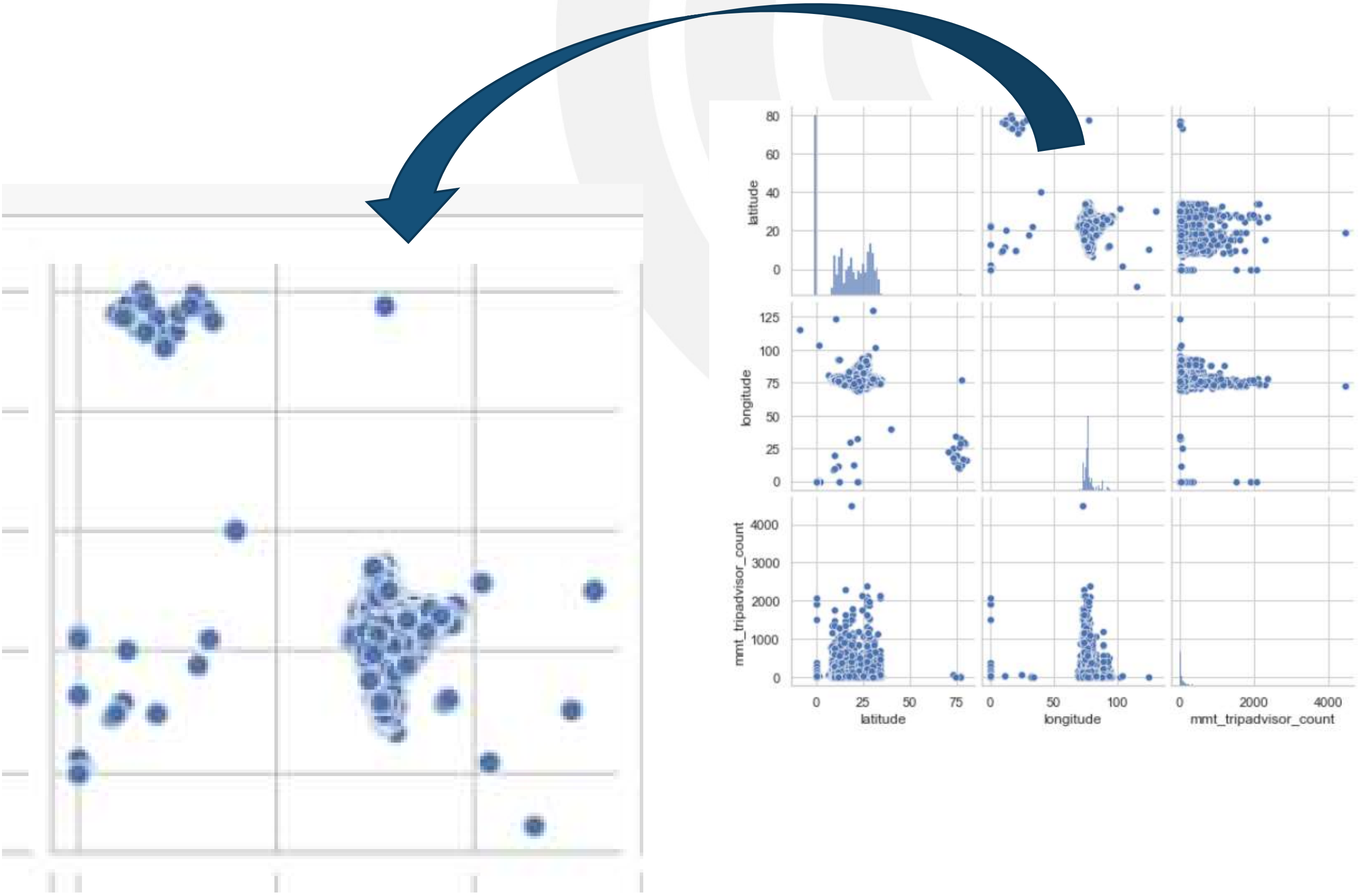
```
area                               5324
city                                  0
country                             458
crawl_date                         9349
highlight_value                    2372
hotel_overview                     3354
hotel_star_rating                     0
in_your_room                       5350
is_value_plus                         0
latitude                             21
longitude                            21
mmt_location_rating                8503
mmt_review_rating                  8503
mmt_traveller_type_review_count    2037
mmt_tripadvisor_count              8888
pageurl                               0
property_address                      4
property_id                           0
property_name                         0
property_type                       110
query_time_stamp                      0
room_types                         4881
sitename                              0
uniq_id                               0
dtype: int64
```

CO- RELATION BETWEEN ALL COLUMNS

- Latitude and longitude as 0.78 co- related
- mmt_review_count and mmt_tripadvisor_count as 0.44
- Site_review_rating and mmt_review_score having most dependent.



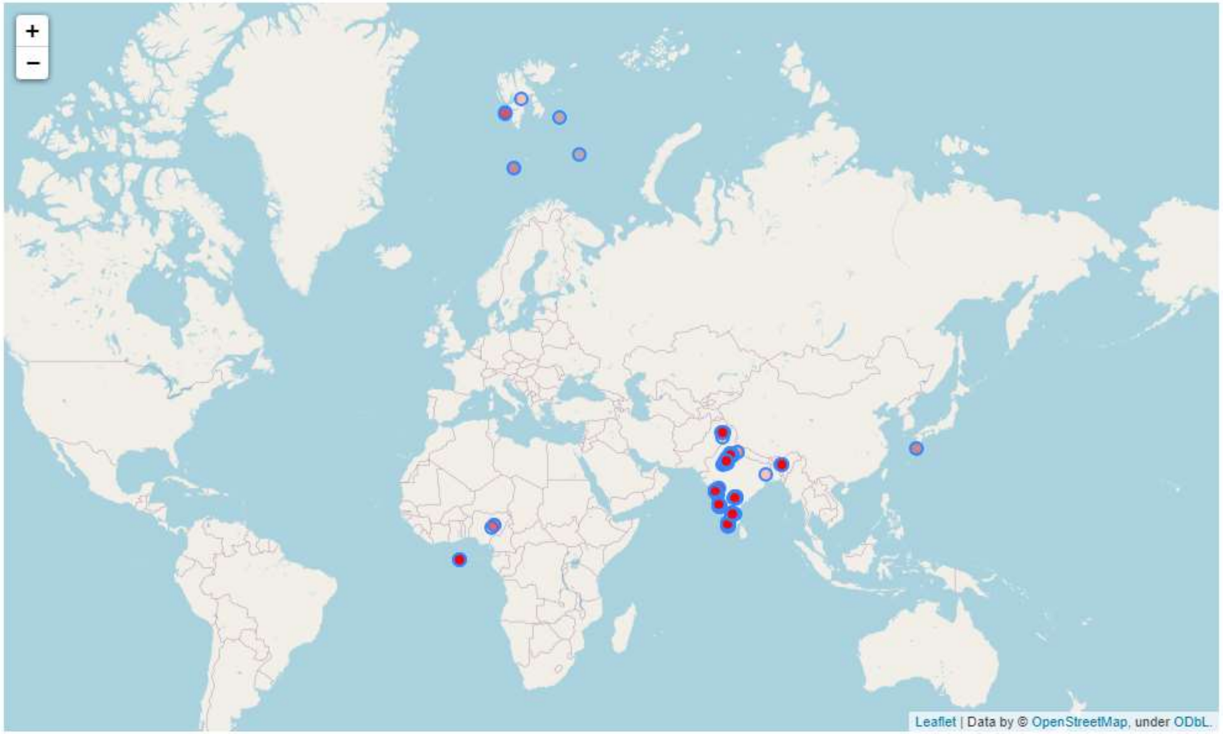
- Here I plot pair plot of MMT dataset. Where I get to know longitude and latitude are highly correlated
- Clusters showing no. of values present in the graph.
- So I decided to take both column and visualize further.



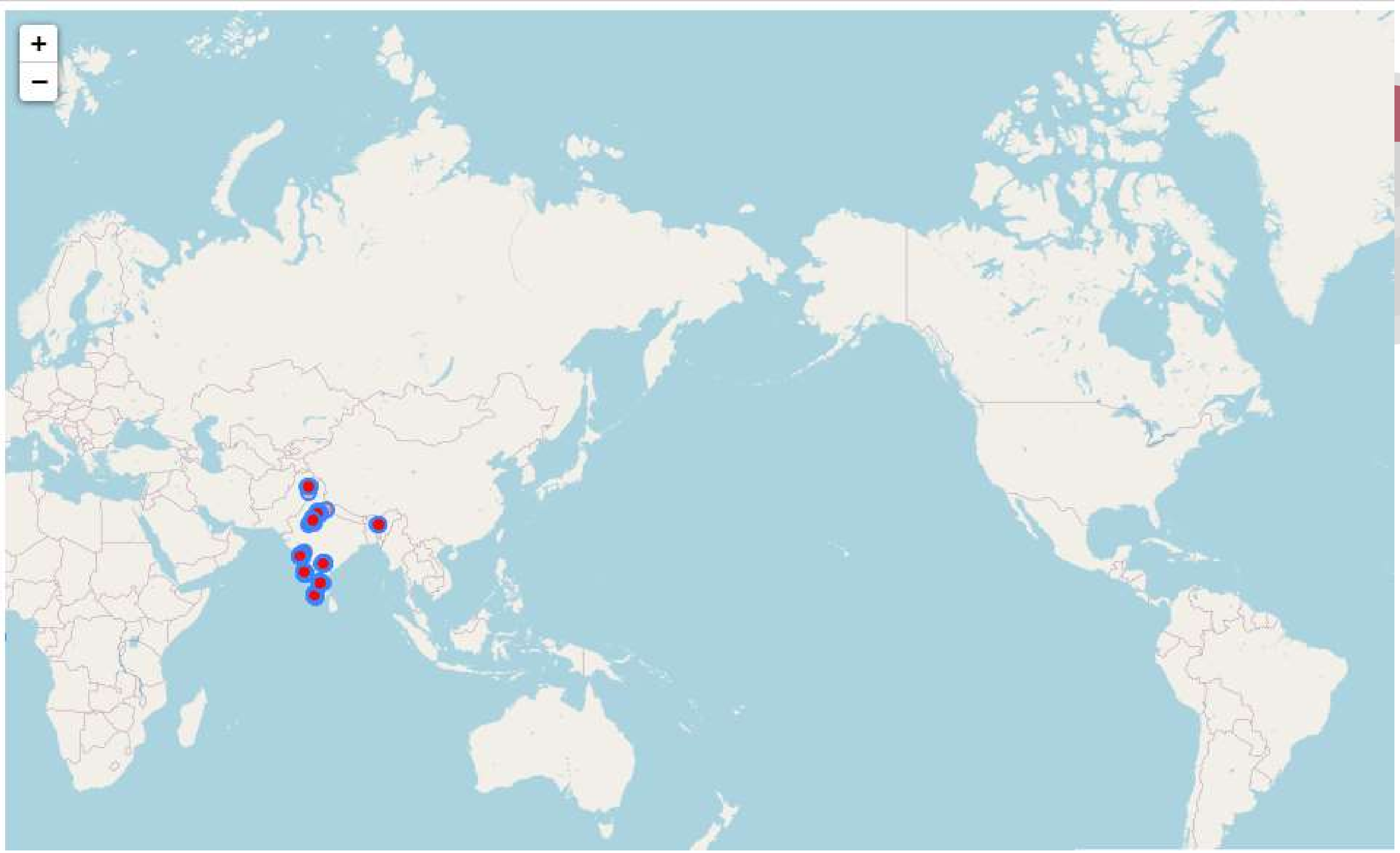
LETS VISUALIZE IN MAP WHERE HOTELS ARE PRESENT

- I had install folium library for plot map of world with the help of longitude and latitude columns
- I drop all the rows were null value are present

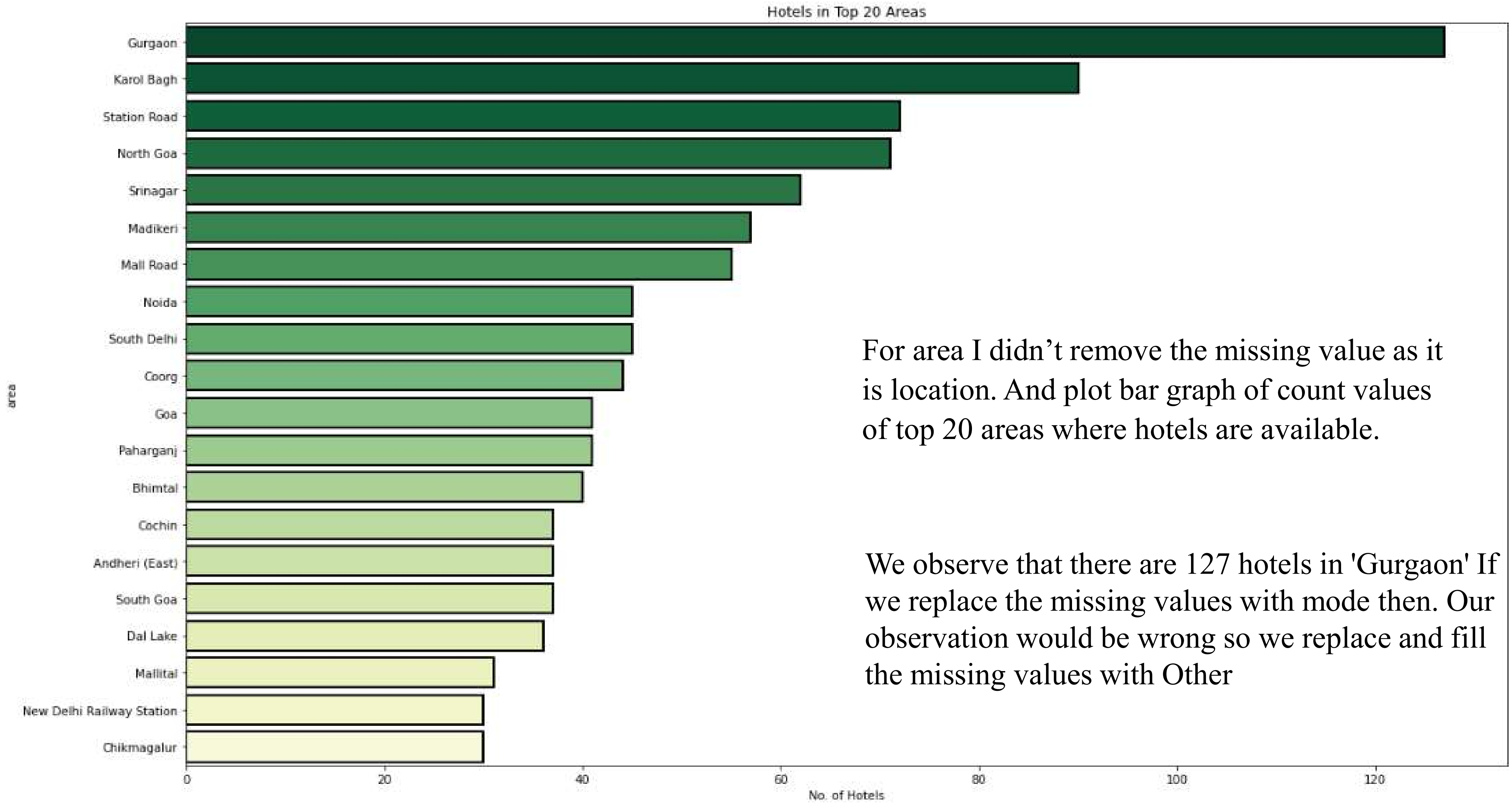
- Here we can find that some of the location are fetching out side the India.
- As Hotels on MMTdataset having only hotel related to India so we Drop that columns having outside location
- Some places are RMV Ext , Sanjaynagar, cochin, golden lake dal lake Srinagar Boulevard Road, Calangute, Boulevard Road, Dal Lake, MA Road, Besides Peddamma Temple, 201301, Dollars, Candolim, west Extension



CORRECT MAP OF HOTEL LOCATION



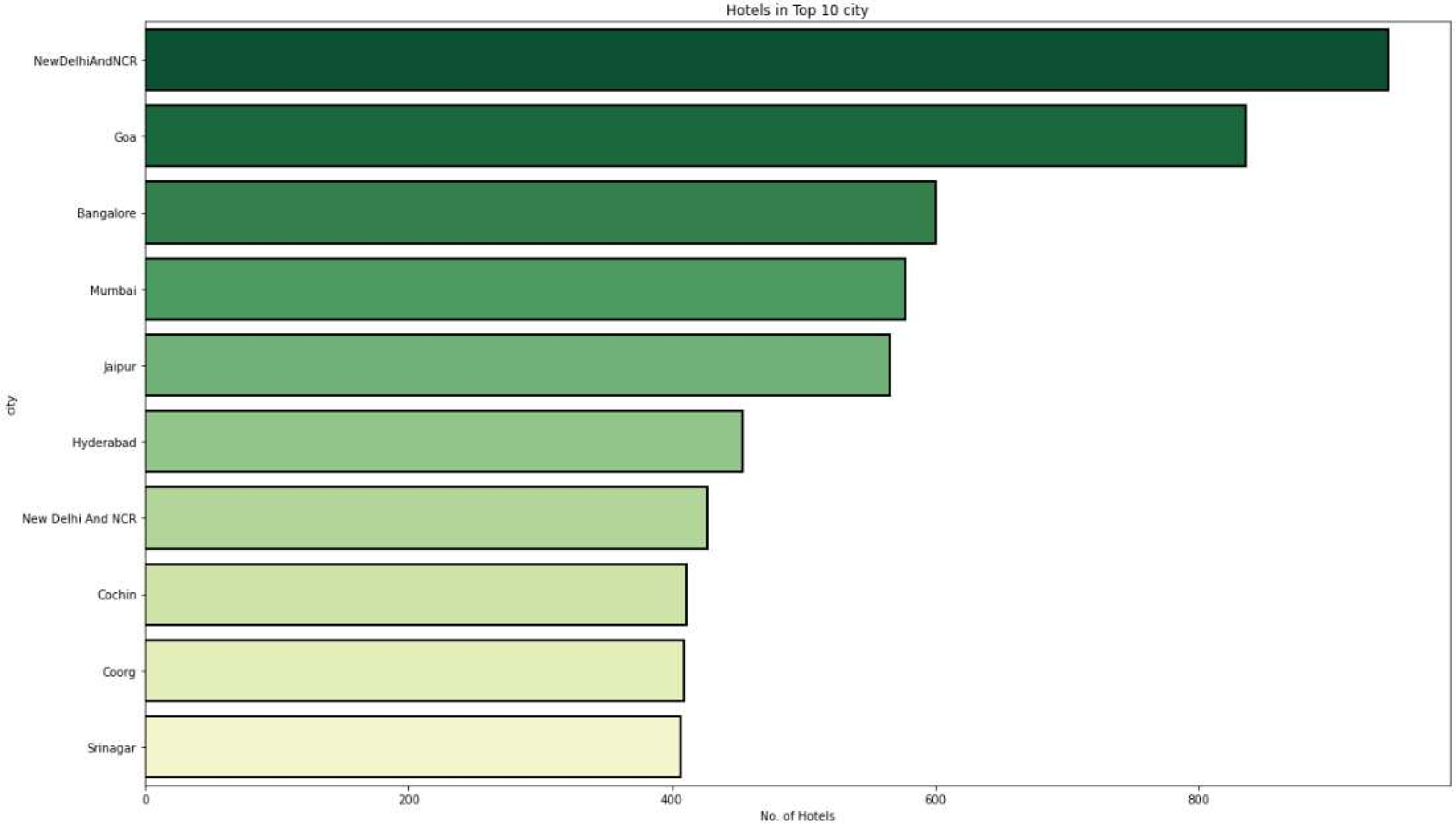
HOTELS ON TOP 20 AREAS



For area I didn’t remove the missing value as it is location. And plot bar graph of count values of top 20 areas where hotels are available.

We observe that there are 127 hotels in 'Gurgaon' If we replace the missing values with mode then. Our observation would be wrong so we replace and fill the missing values with Other

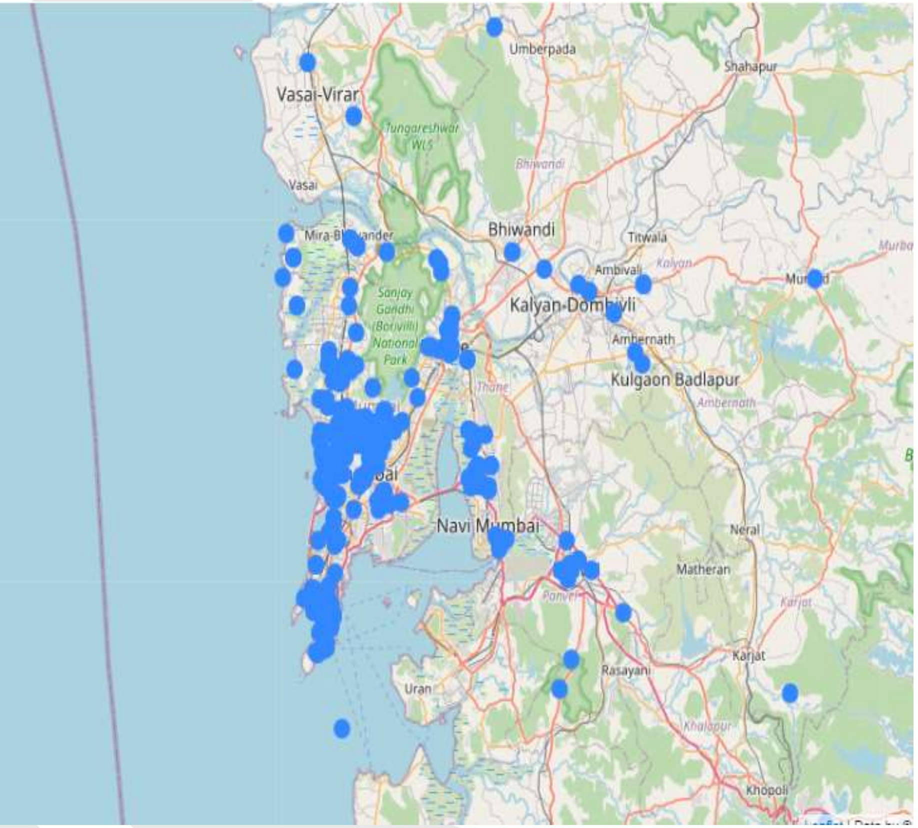
HOTELS ON TOP 10 CITIES



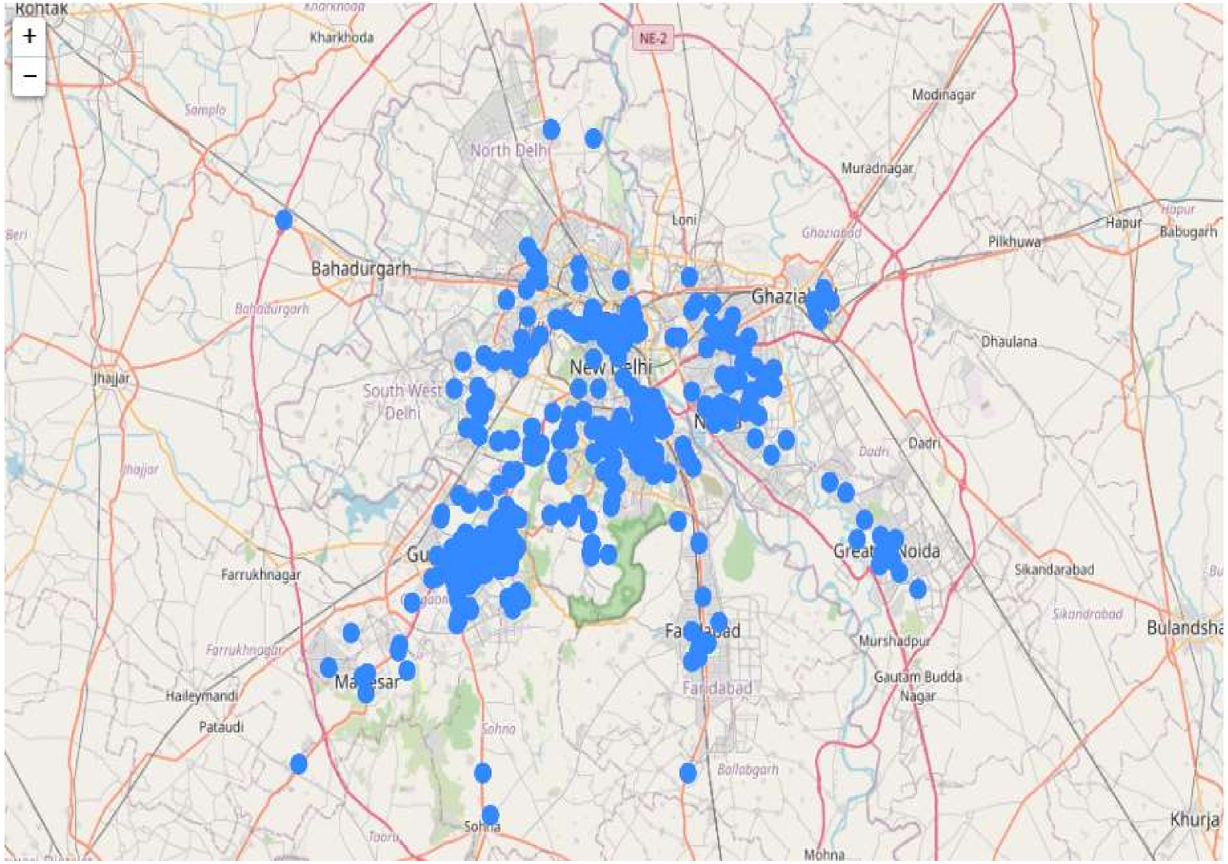
- NewDelhiAndNCR having max hotels 944
- Then Goa having 944 hotels

HOTEL LOCATION IN TOP CITY

- Mumbai



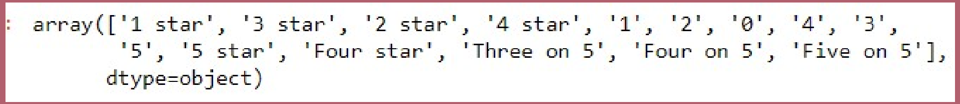
- NewDelhiAndNCR



We observe that 2 hotels are out side Mumbai but its Area fetching to Mumbai so, I removed that.

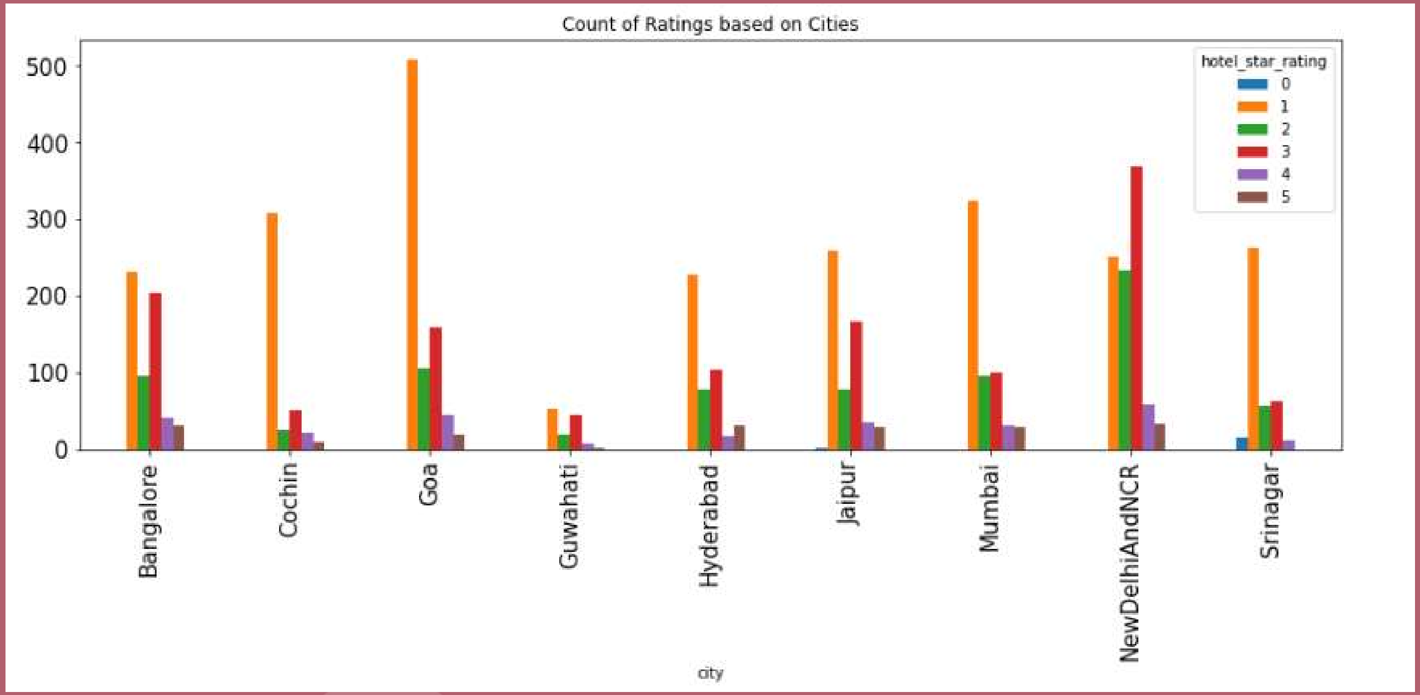
HOTELS RATING BASED ON CITIES

- We observe that hotel_star_rating are in the form of integer and string both. So we convert into integer.
- Also from above we get to know hotels in top 10 cities.
- Syntax: df[‘column_name']=df[‘column_name'].replace(‘value ‘,1 to 5).astype(str)
- NewDelhiAndNCR having overall max hotels having 5 star.
- Goa having too much 1stars rating hotel. Here MMT can improve their facilities to grow their business.

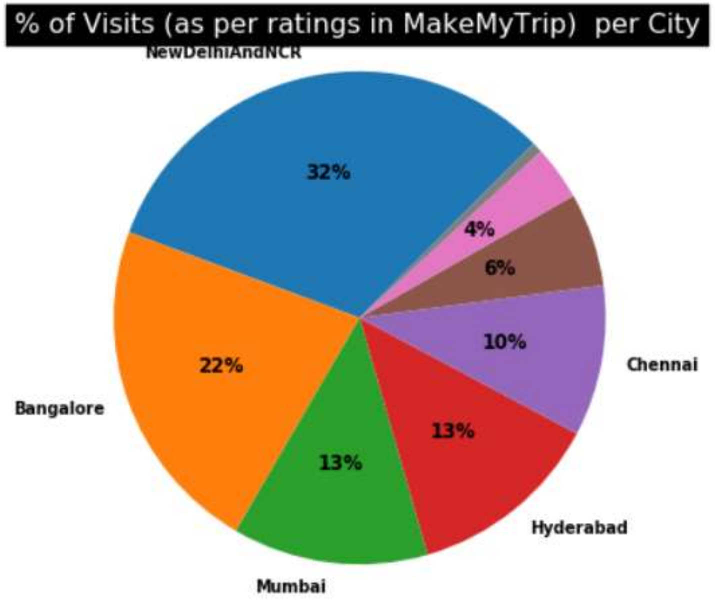


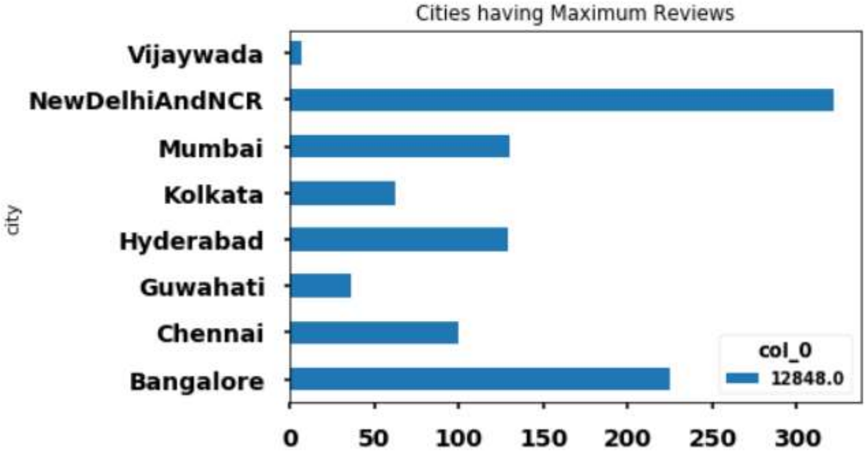
```
: array(['1 star', '3 star', '2 star', '4 star', '1', '2', '0', '4', '3',
         '5', '5 star', 'Four star', 'Three on 5', 'Four on 5', 'Five on 5'],
       dtype=object)
```

1,2,3,4,5



- Here we visualize that maximum people reviews hotels of NewDelhiAndNCR





Some Other Column Manipulation

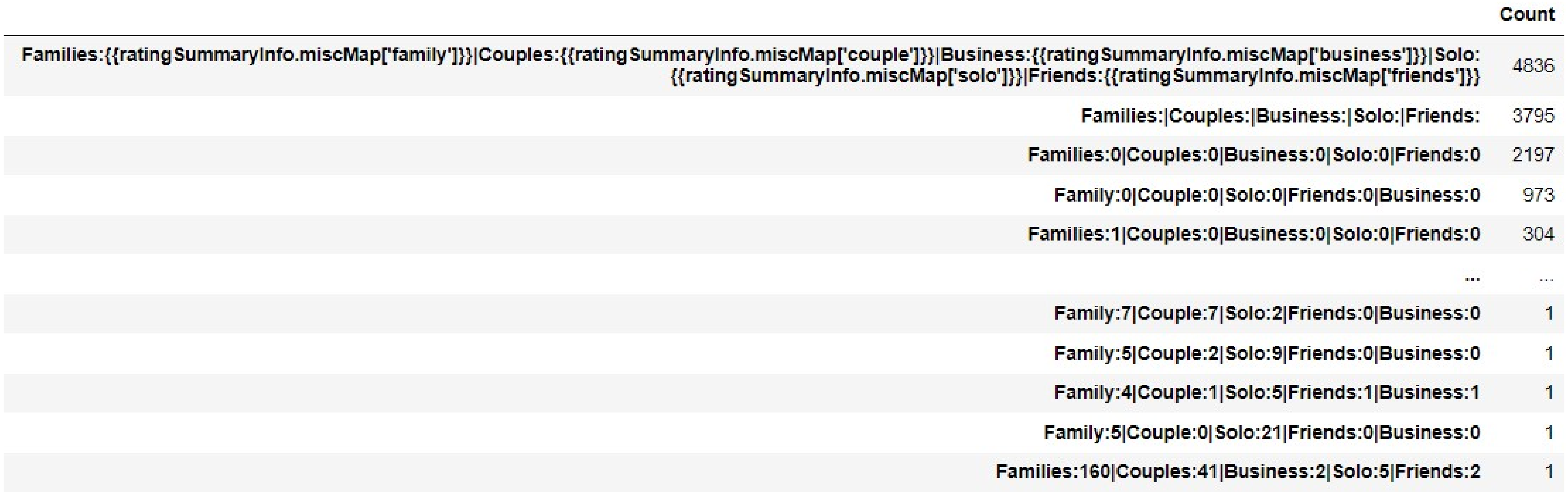
- **Crawl_date** : This columns having date format so replace the null values as mode. It just showing the no. of peoples are revisited the site. On 05-06-2016 max people are revisited

- In **highlight_value** column we observe that maximum row are filled with {{facilities}}. So I decided to fill the null values and replace {{facilities}} into the second facility max. i.e., **doctor on call.**

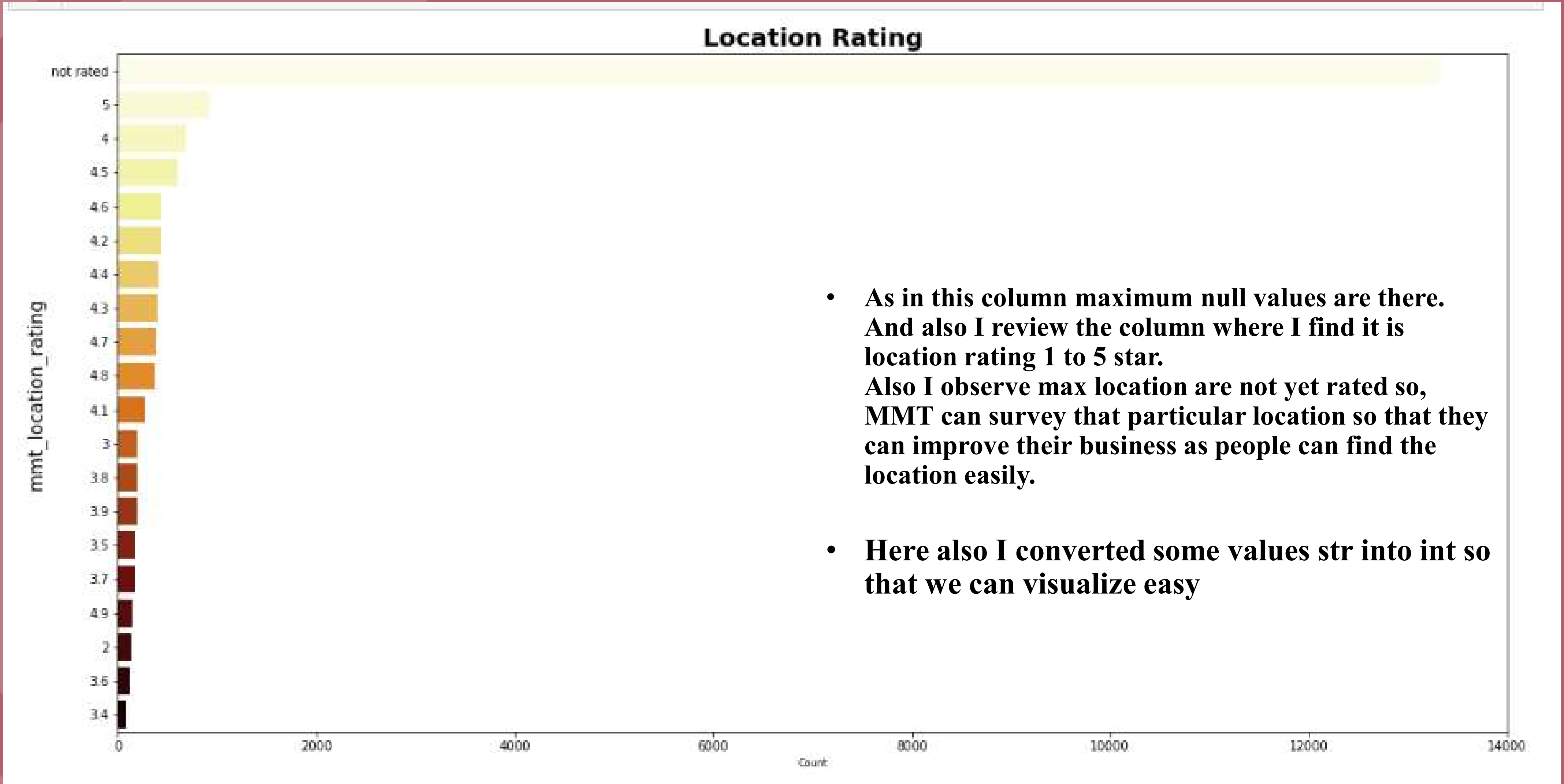
- In **Hotel_overview** column hotels overview is mentions as the max row of hotel_overview is ||less so in order to improve business MMT need to ask the hotel to mention their overview so that person can easily find the location and their needed hotels.

- In **column in_your_room :** The equipment's which are present in the rooms. As Here also "{{value}}" present in the max row so I replace it with needs of equipment present in a room.

- In mmt_traveller-_type_review_count column there are description of type of visitors review of hotel.
- i.e., with family, friends, solo, Business, Couples.
- Here I decided to drop this column as this is of no use.

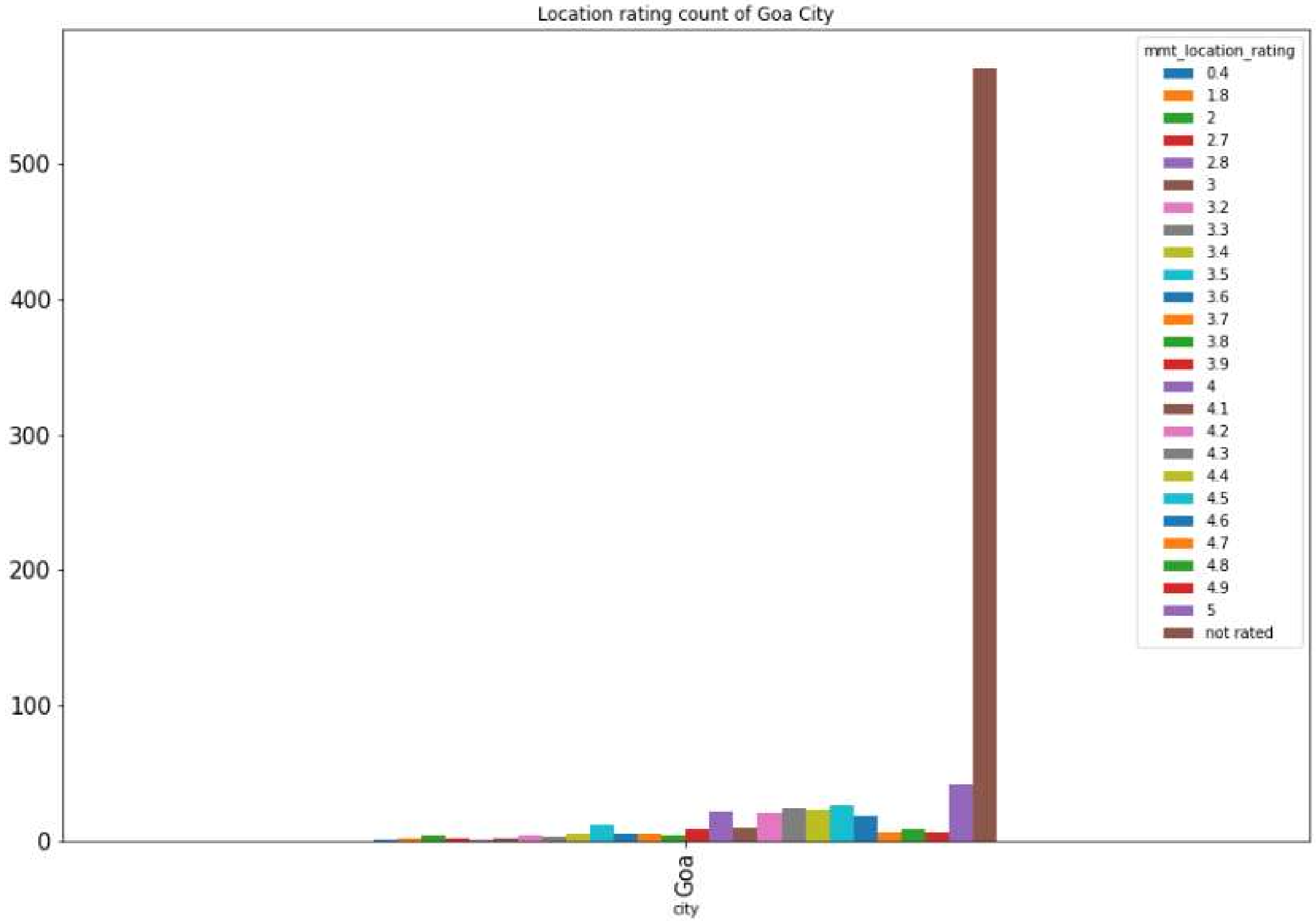


	Count
Families:{{ratingSummaryInfo.miscMap['family']}}|Couples:{{ratingSummaryInfo.miscMap['couple']}}|Business:{{ratingSummaryInfo.miscMap['business']}}|Solo:{{ratingSummaryInfo.miscMap['solo']}}|Friends:{{ratingSummaryInfo.miscMap['friends']}}	4836
Families:|Couples:|Business:|Solo:|Friends:	3795
Families:0|Couples:0|Business:0|Solo:0|Friends:0	2197
Family:0|Couple:0|Solo:0|Friends:0|Business:0	973
Families:1|Couples:0|Business:0|Solo:0|Friends:0	304
...	...
Family:7|Couple:7|Solo:2|Friends:0|Business:0	1
Family:5|Couple:2|Solo:9|Friends:0|Business:0	1
Family:4|Couple:1|Solo:5|Friends:1|Business:1	1
Family:5|Couple:0|Solo:21|Friends:0|Business:0	1
Families:160|Couples:41|Business:2|Solo:5|Friends:2	1



- **As in this column maximum null values are there. And also I review the column where I find it is location rating 1 to 5 star. Also I observe max location are not yet rated so, MMT can survey that particular location so that they can improve their business as people can find the location easily.**

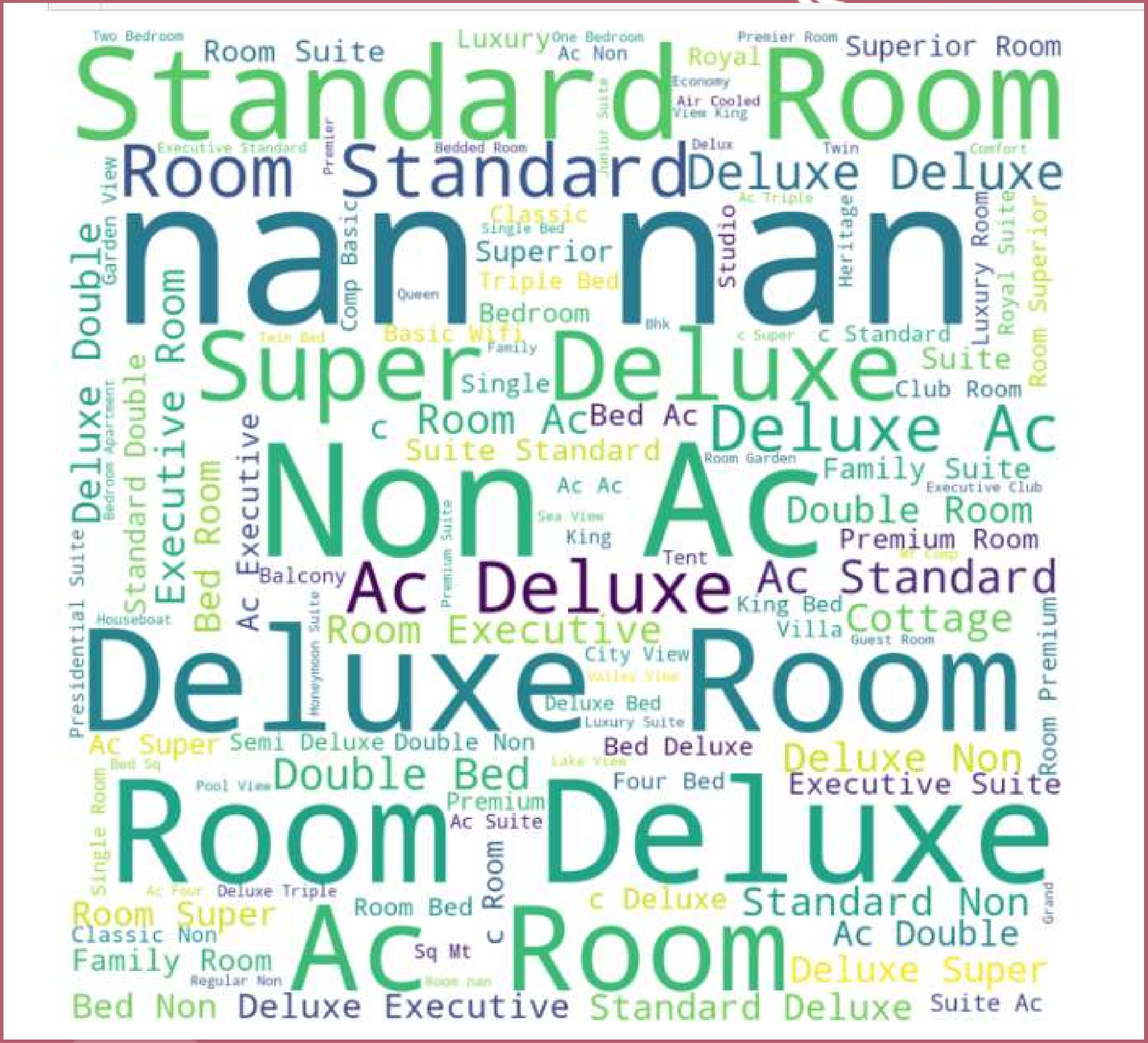
- **Here also I converted some values str into int so that we can visualize easy**



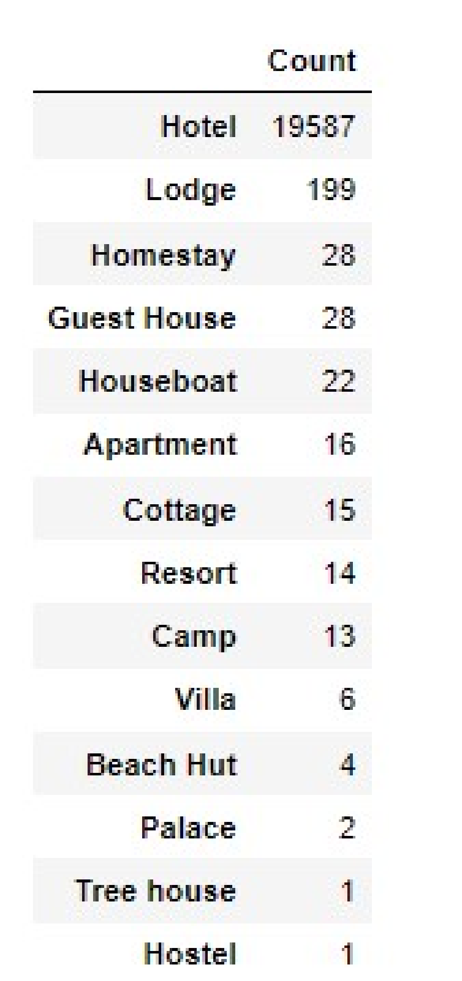
- here we can observe that many locations are not yet rated in Goa
- after that 5 rated place are max
- to improve business in Goa all the locations are to be rate so that people can visit in that location and book thier room

Room Types

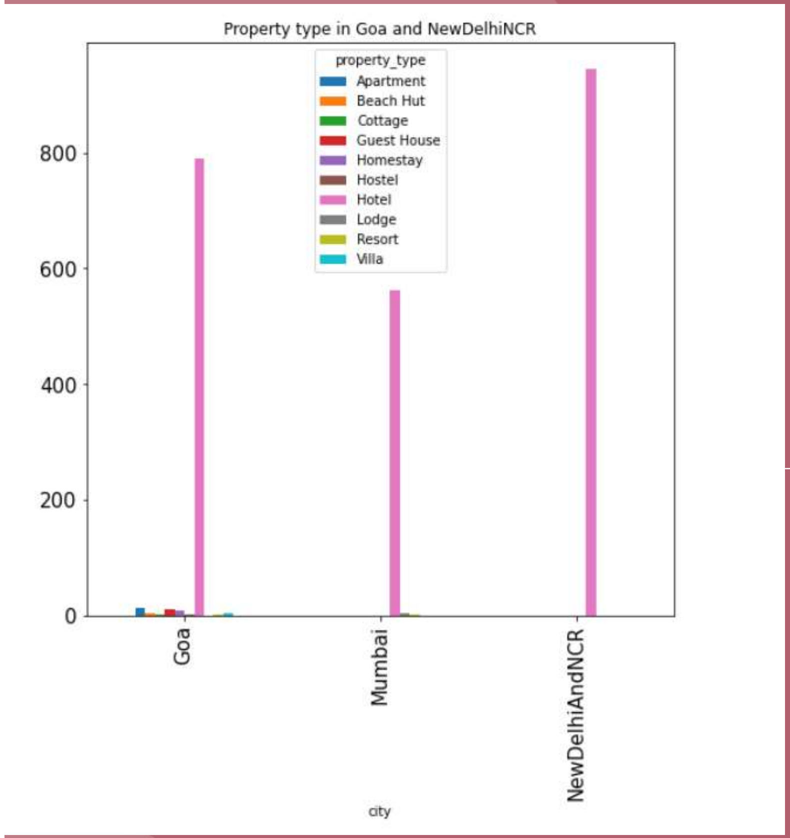
- Maximum Hotels are having Standard Room, Non Ac room, Deluxe Room, Ac Room.
- Here nan showing the null values so we replace that as Other.



HOTELS PROPERTY TYPE



	Count
Hotel	19587
Lodge	199
Homestay	28
Guest House	28
Houseboat	22
Apartment	16
Cottage	15
Resort	14
Camp	13
Villa	6
Beach Hut	4
Palace	2
Tree house	1
Hostel	1



WE CAN OBSERVE HERE THAT

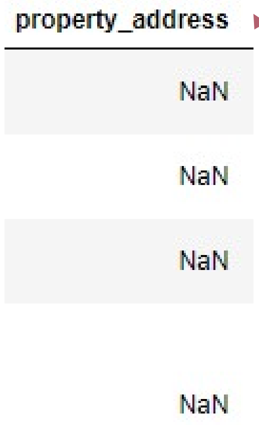
- MAXIMUM PROPERTY ARE “HOTEL”
- ONLY ONE PROPERTY IS HOSTEL
- NEWDELHIANDNCR HAVING ALL PROPERTY TYPE ARE “HOTELS”

- We observe that after all analysis. There are some null values in latitude and longitude. Here we already drop the rows where null value are present. And observer the location of hotels in map.

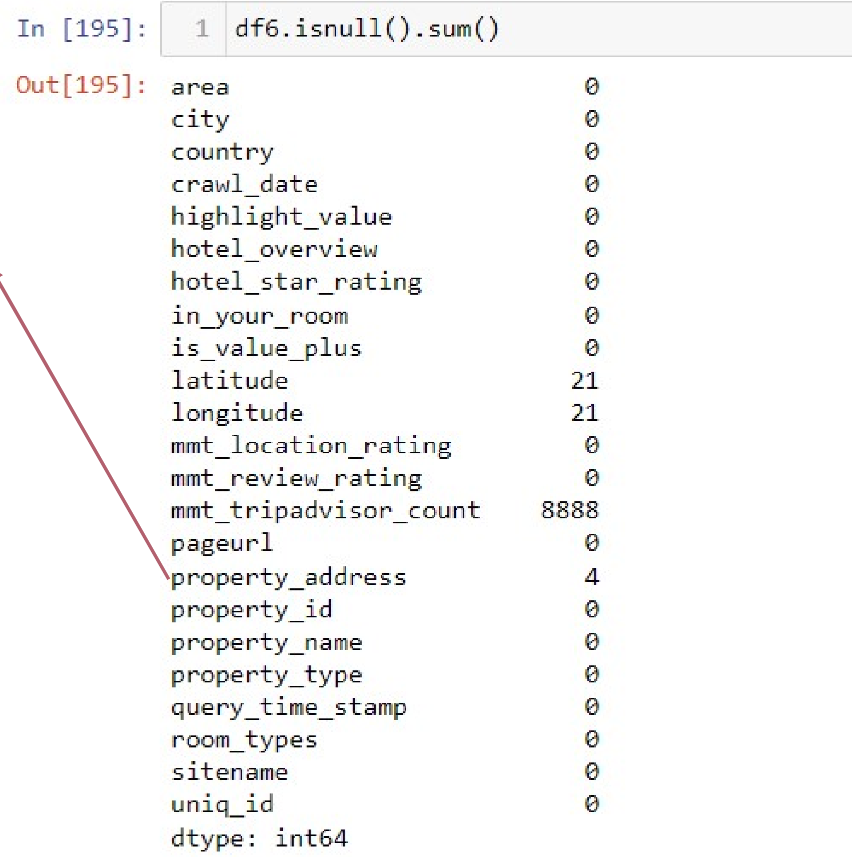
- In property_address there are 4 null values. So we drop that row having address.

- By the help of “isna” we find all four rows i.e., 686, 2896,15035, 19964.

- Or MMT ask that hotel address In order to customers will more book that hotel too.



property_address
NaN
NaN
NaN
NaN



```
In [195]:  df6.isnull().sum()

Out[195]:  area                      0
           city                      0
           country                   0
           crawl_date                0
           highlight_value           0
           hotel_overview            0
           hotel_star_rating         0
           in_your_room              0
           is_value_plus             0
           latitude                 21
           longitude                21
           mmt_location_rating       0
           mmt_review_rating         0
           mmt_tripadvisor_count  8888
           pageurl                   0
           property_address          4
           property_id               0
           property_name             0
           property_type             0
           query_time_stamp          0
           room_types                0
           sitename                  0
           uniq_id                   0
           dtype: int64
```

Problem I faced

- In Dataset too much null values are there in any columns.
- Many location data are fake in the dataset
- 4-5 columns are not of use like unique id, image URL
- Also before this I changed 1 dataset because I can not understand the data

Solution for Problem

- For null values I remove the columns by taking threshold value.
- I verified the fake location in the column by plotting map and remove that data
- Guidance of my TA and support of my friend Vikash
- I face some issue in Jupiter notebook then visualised the data in Tableau and get it easily

Contoso

CONCLUSION

- We find that DelhiAndNCR having maximum no. hotels are rated.
- Goa having maximum number of hotels which having rating 1 star
- Maximum Property is of Hotel Type
- One Property is hostel
- Maximum hotels are having Standard AC and Non-Ac, Dulux rooms.
- Many places in different location still not rated yet.
- For Business Advices : As I found we can work on that location where people are most visiting and most popular places like Goa, New Delhi, Bangalore. We can improve facilities and accommodation and also take survey of all the areas so that people can book their tickets and hotels accordingly

What I learnt from this project ?

- Visualisation of data. i.e., EDA
- Become more confidence what I'm researching
- Many visualisation I did from Tableau too that help me to explore more about Tableau

Contoso

✉