AirBnB Price Prediction Case Study

Introduction and Problem Overview

Data Summary

Variable Creation/ Feature engineering and Interesting findings

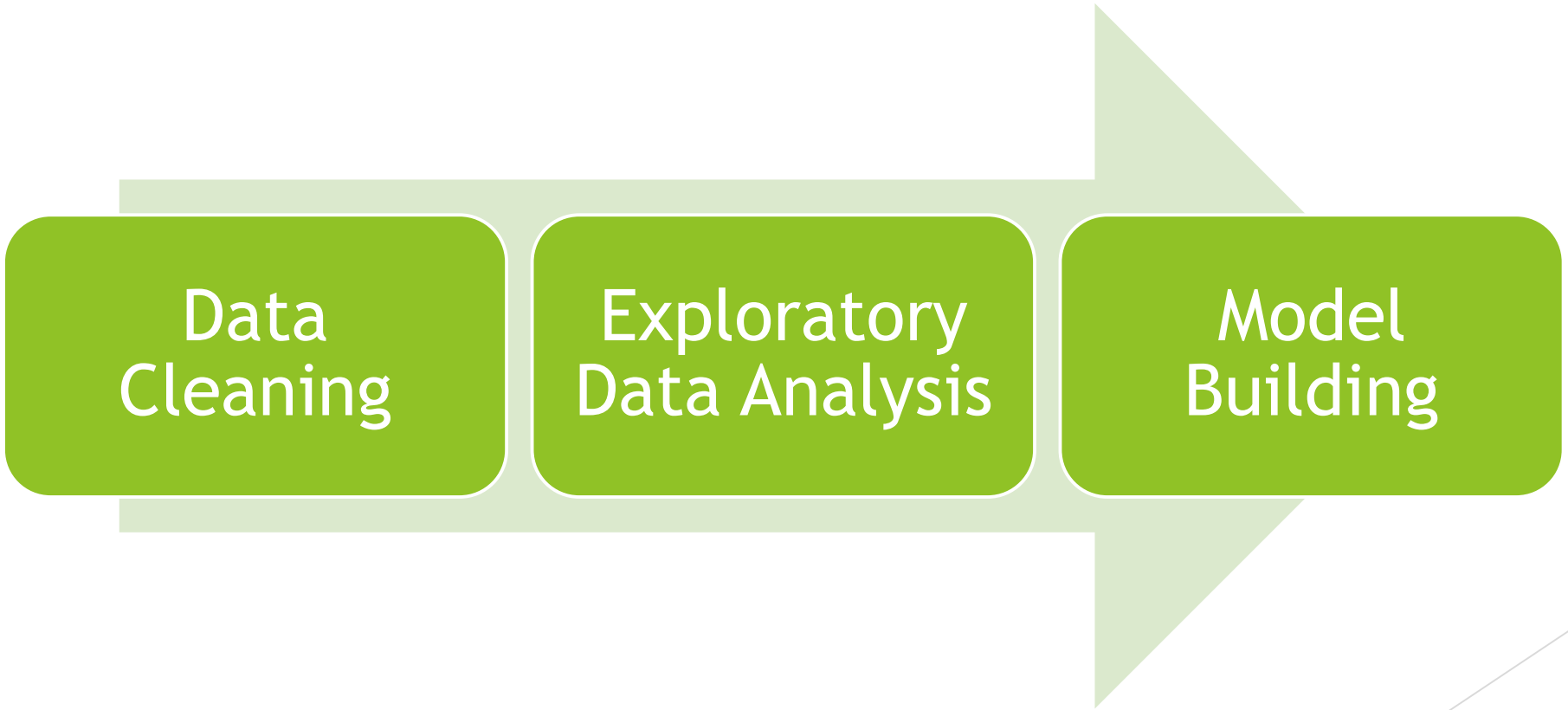
Modelling Approach and results

Conclusion and final thoughts

Introduction & Problem overview

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in that locale. The challenge statement was AirBNB price prediction.
Some of the factors were property, room type, amenities, accommodates, bathroom, bed type, etc.
Using machine learning technique, proceeding with all factors leading to the price prediction to develop the model.

Introduction & Problem overview



Data Summary

- The raw data had 29 columns which include 1 dependent variable column which depends on other 28 variable columns.
- 28 columns are 'id', 'property type', 'room type', 'amenities', 'accommodates',
- 'bathrooms', 'bed type', 'cancellation policy', 'cleaning fee', 'city',
- 'description', 'first review', 'host_has_profile_pic',
- 'host_identity_verified', 'host_response_rate', 'host since',
- 'instant_bookable', 'last review', 'latitude', 'longitude', 'name',
- 'neighborhood', 'number_of_reviews', 'review_scores_rating',
- 'thumbnail_url', 'zip code', 'bedrooms', 'beds', 'log price'

Data Summary

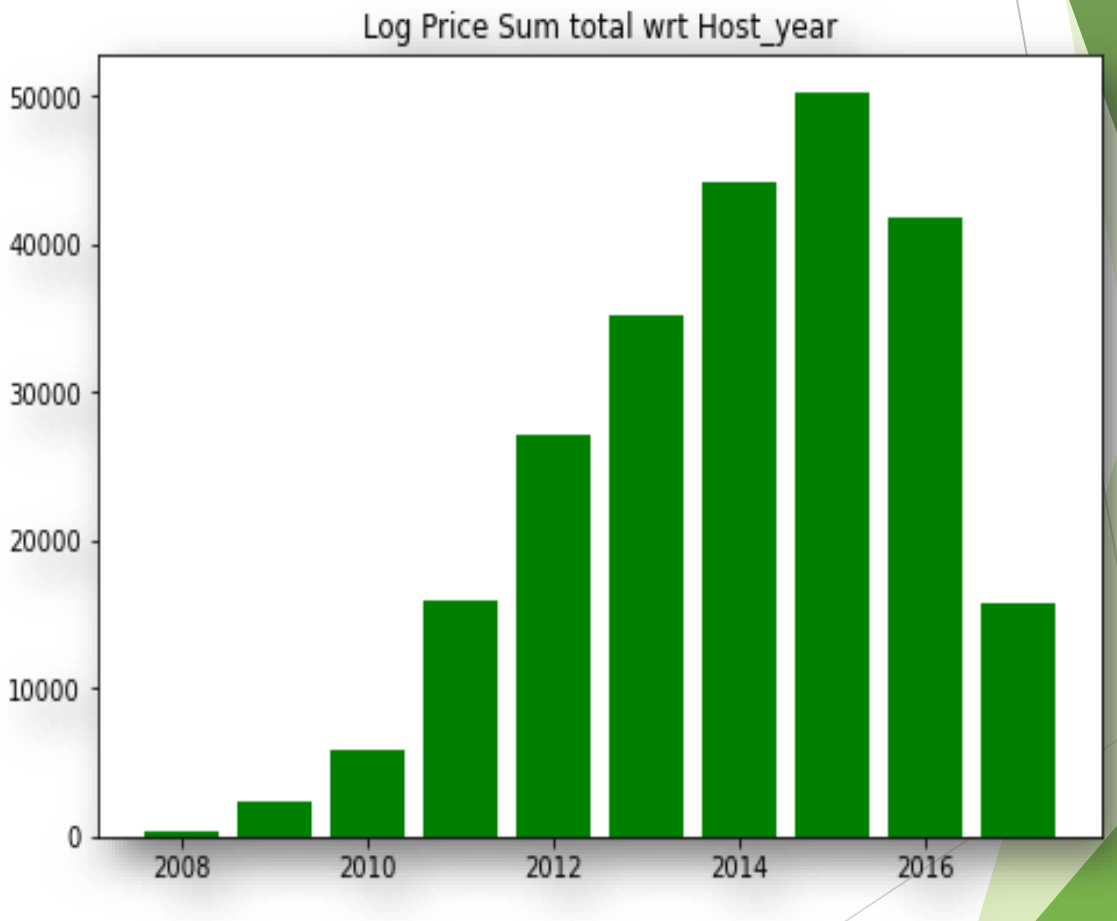
- 'property type‘ -> 'Apartment', 'House', 'Condominium‘ , various types of categories
- 'room type‘ -> Entire room, Shared room and private room
- 'amenities‘ -> amenities they provide in that property
- 'accommodates‘ -> Capacity of that property to hold occupants
- 'bathrooms' -> number of bathrooms
- 'bed type'‘ -> Real bed, Funston
- ‘cancellation policy‘ -> strict, moderate, very strict
- 'cleaning fee’ -> cleaning fee
- 'city’ -> NYC, SF, DC, LA, Chicago, Boston
- 'host since', -> hosted year
- 'name', -> name of the property
- 'neighborhood‘ -> popular neighborhood places
- 'number_of_reviews', 'review_scores_rating‘ -> regarding ratings
- 'bedrooms', 'beds‘ -> number of bedrooms
- 'log price‘ -> price of the property …& many more related variables

Variable Creation/ Feature engineering and Interesting findings

- **Converting "host since" column into proper date time format, filling up an values with forward fill and extracting only year.**
- **One hot encoding categorical variables, as it will increase the model prediction accuracy ->'city Chicago', 'city DC', 'city LA', 'city_NYC', 'city_SF',**
- **'property_type_Bed & Breakfast', 'property_type_Boat',**
- **'property_type_Boutique hotel', 'property_type_Bungalow',**
- **'property_type_Cabin', 'property_type_Camper/RV',**
- **'property_type_Casa particular', 'property_type_Castle',**
- **'property_type_Cave', 'property_type_Chalet',**
- **'property_type_Condominium', 'property_type_Dorm',**
- **'property_type_Earth House', 'property_type_Guest suite',**
- Etc. were the new columns generated in the dataset.

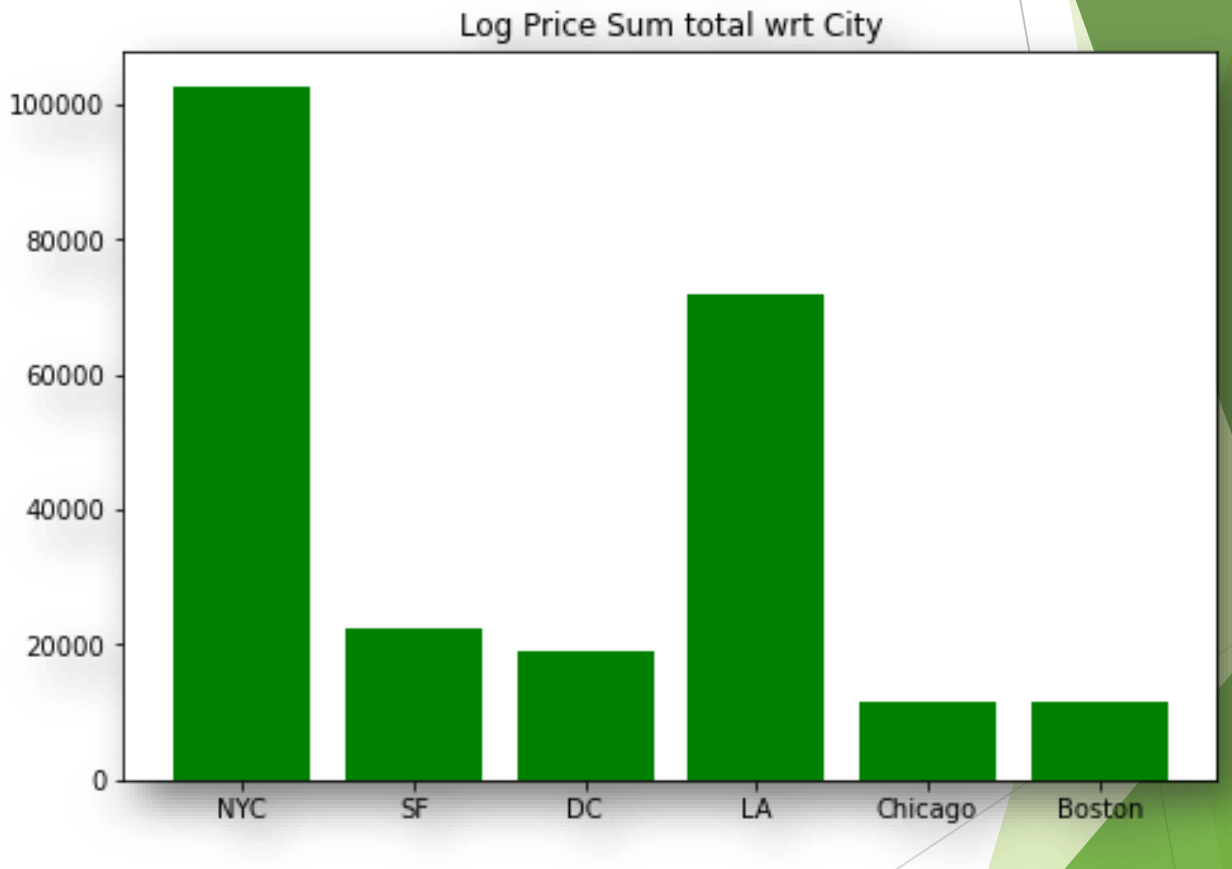
Variable Creation/ Feature engineering and Interesting findings

- **Insights 1: Which host year's property had the good economy log_prices?**
- The host year 2015 has the highest log_prices followed by the year 2014 and the year 2016. The least mean log_price value can be seen for the host year 2011 and the host year 2017.



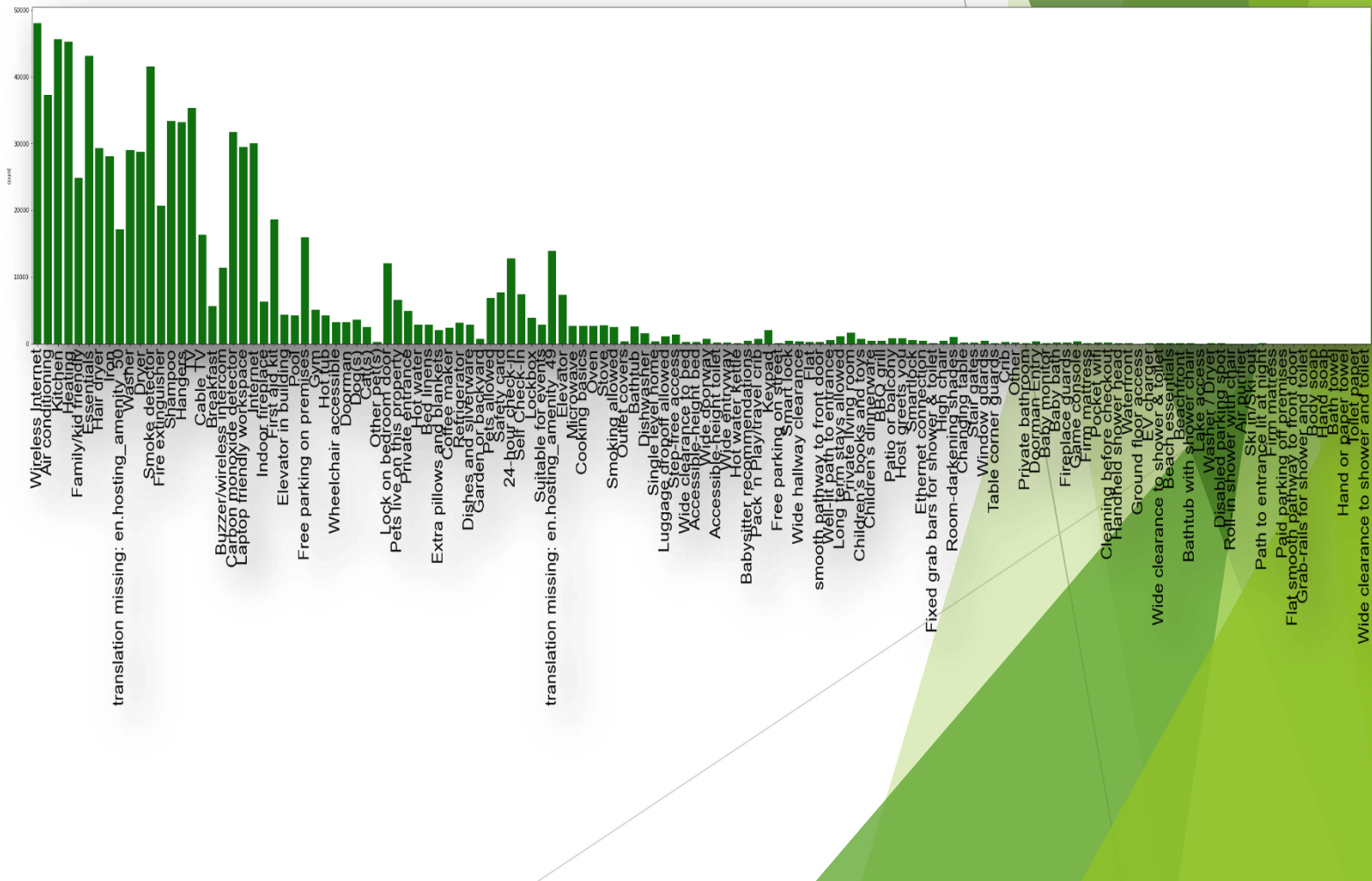
Variable Creation/ Feature engineering and Interesting findings

- **Insights 2 : Which city among NYC, SF, DC, LA, Chicago, Boston were in the good list for better price or having more occupancy of tenants?**
- Found to be better in pricing, NYC and LA leads from most of the cities. Probably most accurate reasons according to the data would be the popularity of these cities. We can also derive there is more occupancy of tenants in these cities.



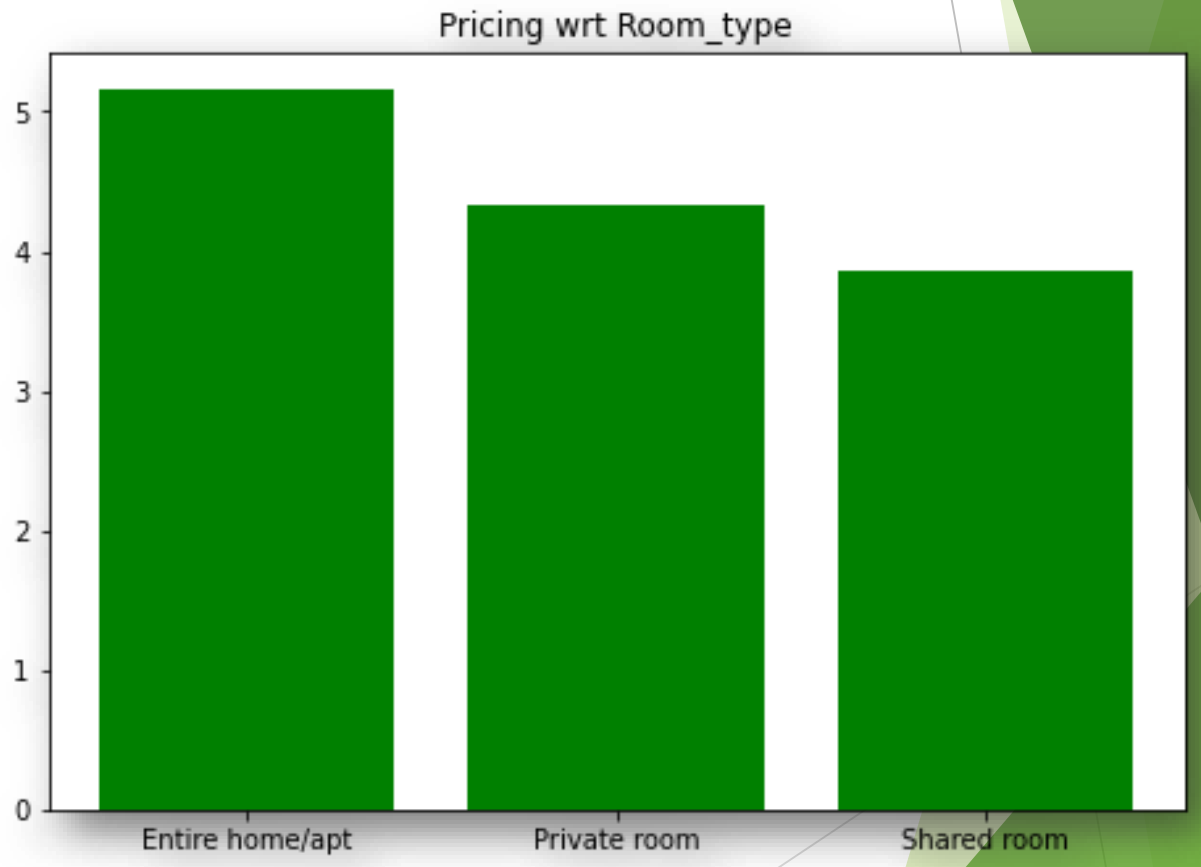
Variable Creation/ Feature engineering and Interesting findings

- **Insights 3: What are the most common facilities offered to the tenants of the house or which all amenities most common in all property places?**
- It can be seen from the graph that most common amenities were Wireless Internet, Air Conditioning, Kitchen, Heating, Family-friendly apartments, pet-friendly apartments, Smoke Detector, Carbon Monoxide Detector were the most common of them all. Can also be seen from the respective of pricing, these facilities were dependent upon the pricing as well.



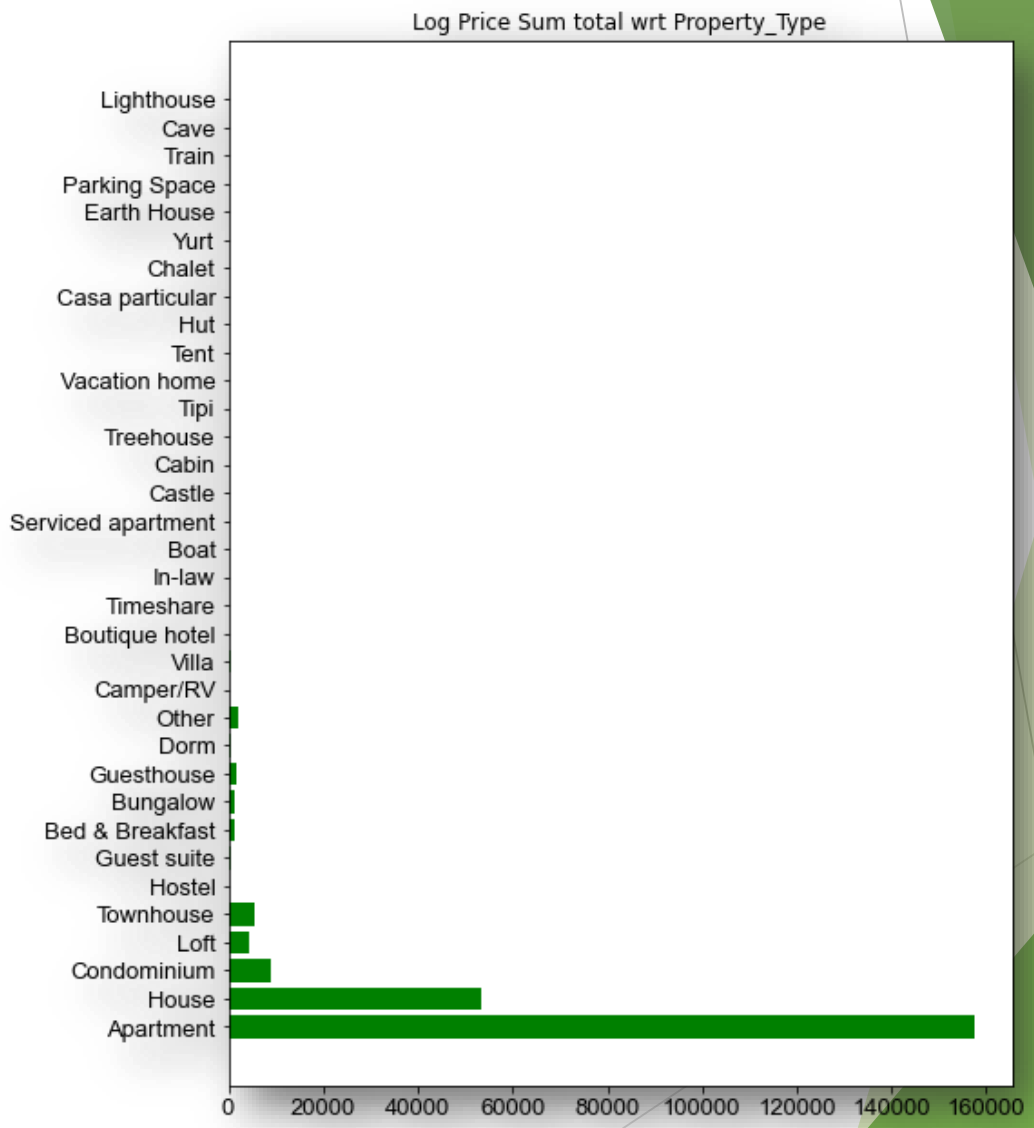
Variable Creation/ Feature engineering and Interesting findings

- **Insights 4: Which room_type has the good Price for occupancy?**
- We can see the cost of stay in Entire Apartment is greatest followed by the private room and shared room.



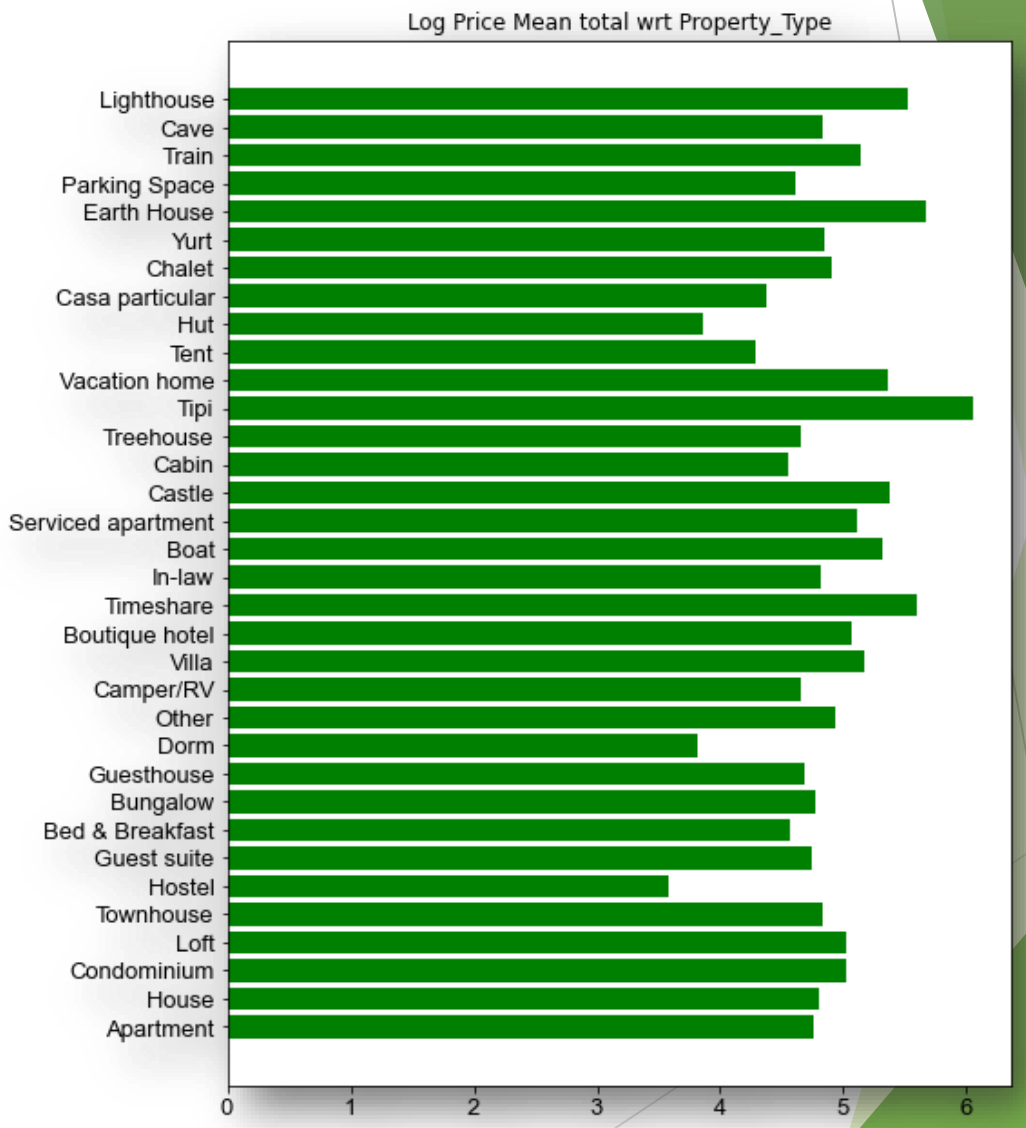
Variable Creation/ Feature engineering and Interesting findings

- **Insights 5: Which property type has the most occupancy?**
- Apartments has the most occupancy followed by the Independent Houses and Condominium.



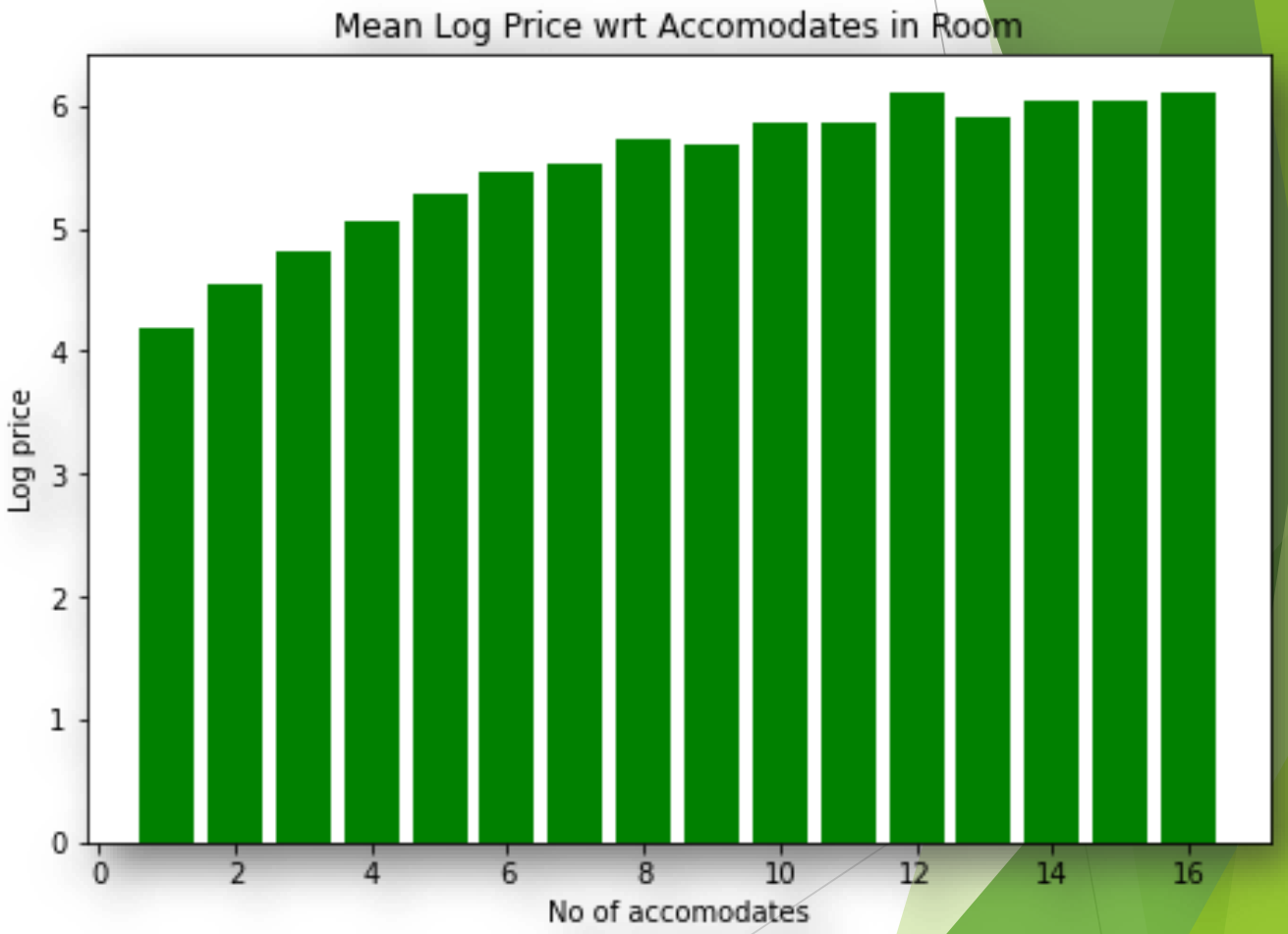
Variable Creation/ Feature engineering and Interesting findings

- **Insights 6 : Which all Property type has the good price range?**
- Hiring a Tipi, vacation home, light house, Earth house would be best and expensive options for the ones who are planning to spend their vacation in. Cheapest options can be seen as in Hostel, Dorm, Hut, Tent etc.



Variable Creation/ Feature engineering and Interesting findings

- **Insights 7 : Is there any correlation between number of accommodates and the price mentioned?**
- We can see clear relation with increasing number of accomodates and log price.

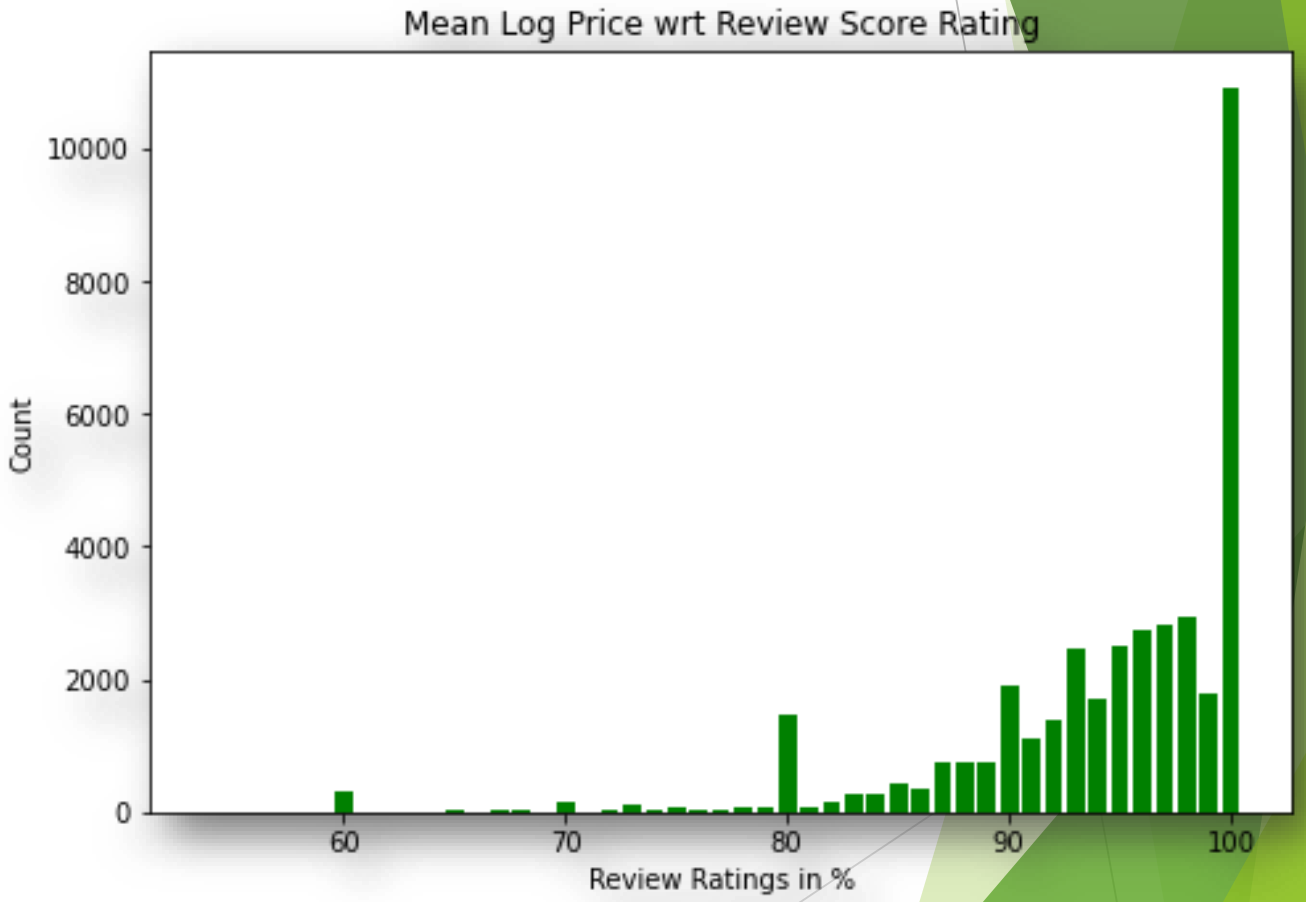


Variable Creation/ Feature engineering and Interesting findings

- **Insights 8 : Is there any correlation between number of bedrooms and the price mentioned?**
- We can see somewhat clear relation with increasing number of bedrooms and log price.

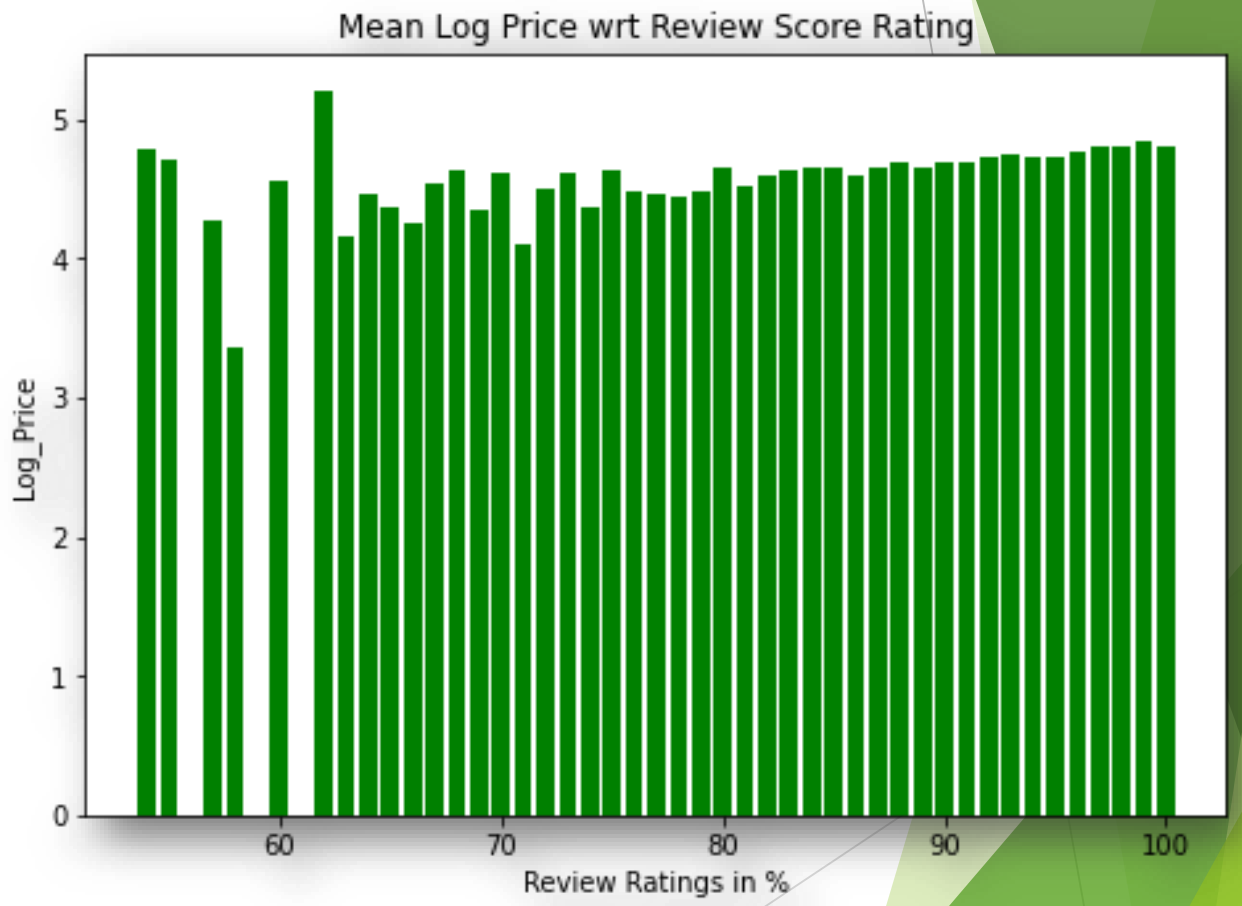
Variable Creation/ Feature engineering and Interesting findings

- **Insights 9 : Any insightful regarding review ratings?**
- Surprisingly, we can see good number of people are satisfied by the service.



Variable Creation/ Feature engineering and Interesting findings

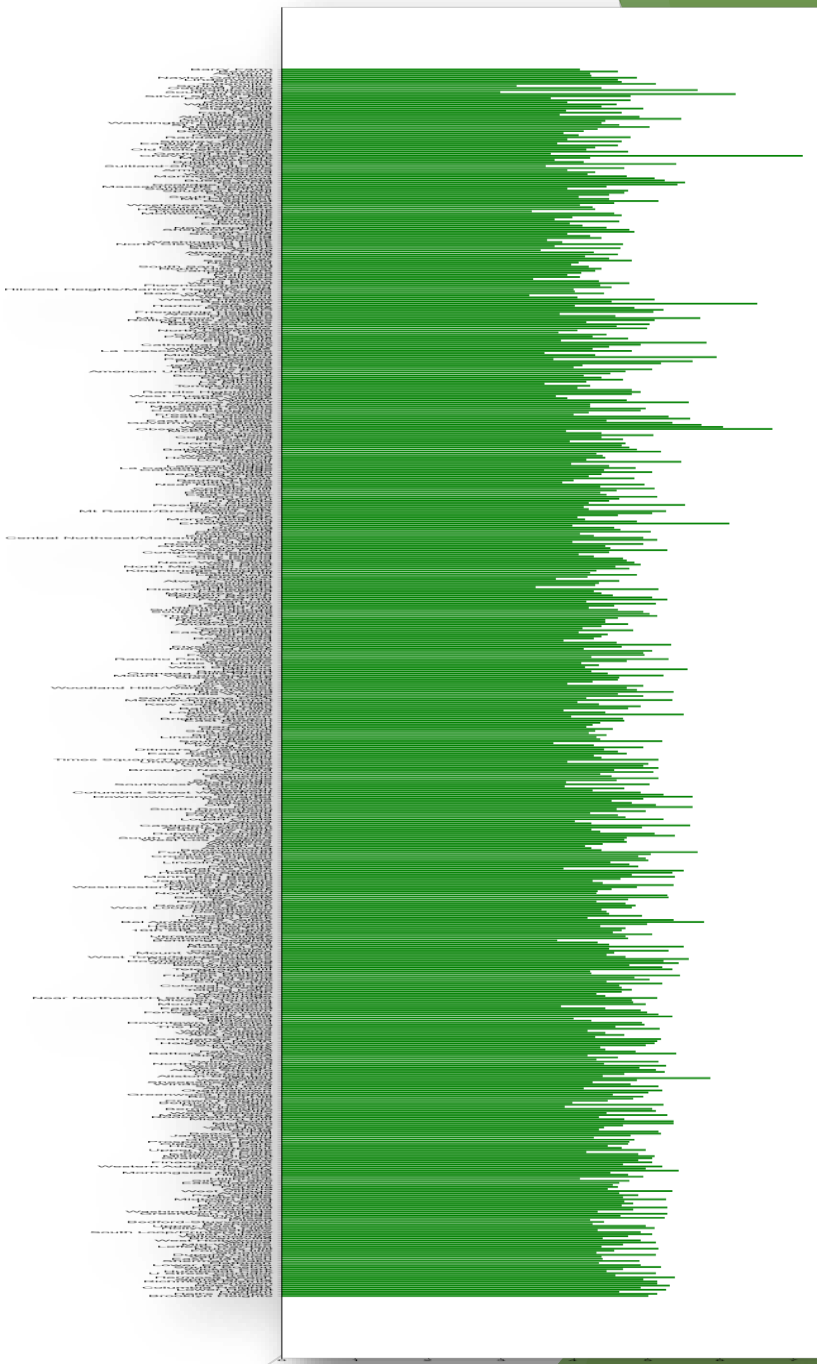
- **Insights 10 : Any improvements by judging other aspects of review rating?**
- The services should also be provided good where the property ratings are little lower.



Variable Creation/ Feature engineering and Interesting findings

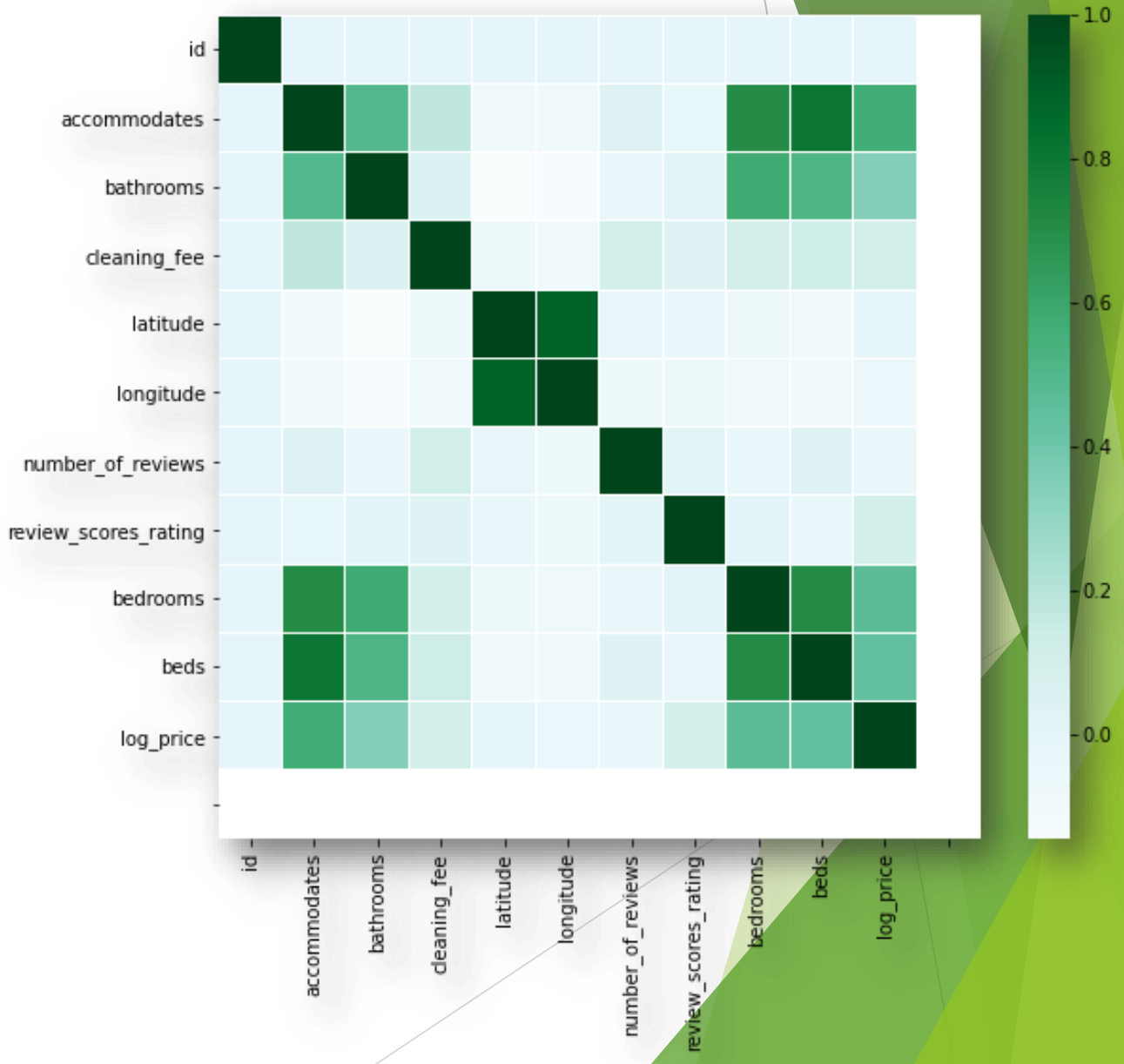
Insights 11 : Do neighborhood places influence the property price?

- Yes, we can see the popular neighborhoods influence the price with a greater extent. Some of the well-liked neighborhoods are South Chicago, Chevy Chase MD, Sea Cliff, Observatory Circle, Emerson Hill and many more.
- **Image can be seen properly in the code book.**



Variable Creation/ Feature engineering and Interesting findings

- **Insights 12 & Conclusion : What are some mostly seen correlations with the price and other aspects?**
- Speaking about the price, this variable in the dataset is likely having more influenced by the count of accommodates, beds, bedrooms and bathrooms.
- If we take the decision of cleaning fee, they are likely influenced by the number of the accommodates.
- If we take decision for making increase of accommodates, the decision would be close to promoting more Family Friendly apartments.



Modelling Approach and results

Taking Next Step into Variable Creation and Feature Engineering

- The 'Amenities' Column had be taken further to making only certain important following amenities: Internet, AirConditioning, Kitchen, FamilyFamily, Essentials, TV, Pets Friendly, Breakfast and Smoke Detector.
- Host response rate from percentage character to converting it into float and filling up null values in the column by its mean
- Converting "host since" column into proper date time format, filling up na values with forward fill and extracting only year and variable creation

Model Preparation

After analysing models, Random Forest Regressor can be considered.

- Random Forest Model
- MSE = 0.198, Accuracy = 61.54%

- Improving model by tuning its parameter
- Using grid Search CV
- MSE = 0.196 Accuracy = 61.90% (IMPROVED)

Conclusion and final thoughts

- The best performing model was able to predict 62% of the variation in price with an MSE of 0.196. Which means we still have a remaining 38% unexplained. This could be due to several other features that are not part of our dataset or the need to analyze our features more closely.
- Highlighting accessibility and location benefits of staying with them could perhaps benefit them and how much they can ask for their listing.