

EXPLORATORY DATA ANALYSIS

METHODS FOLLOWED

- Created ER diagram to understand how the different sources of data are connected.
- After looking at the ER diagram came to understand that for the lodging and accommodation business there are several factors which influence the ratings , price. The 2 most important variables to be considered for any listing.
- Removed all the listings which were not available.
- Removed the rows which had blanks,NA's for the variable under consideration.
- Observed patterns in variables and developed insights from those and substantiated the insights with visual charts developed using Excel, ggplot in R, and python.
- Used packages in python for word cloud, cleaning and calculating sentiment scores for the text data.

OVERVIEW OF LEARNINGS FROM EDA

There are a lot of factors influencing the accommodation trend. When we dig deep and analyse in a granular level using the data from listings.csv we are able to derive the following

Using business logic we can state that factors like number of bedrooms, price of rooms etc also play an important role which are proven by observing behaviour in the charts. Using these charts we can also explore and understand the data even more

Deriving insights such as sentiment etc from the reviews and text data that is provided

Ratings play an important part in maintain business. Review rating , host rating are all measure of how much the customer likes the services and facilities provided. If the ratings are high we can expect that more people will come to the resorts and hence an increasing trend can be expected.

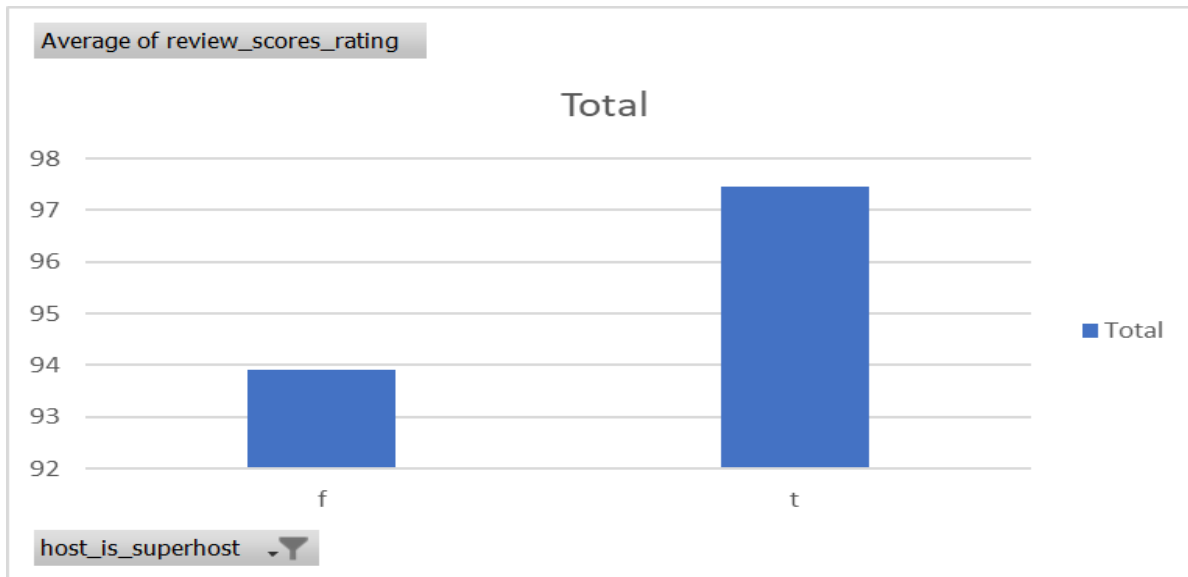
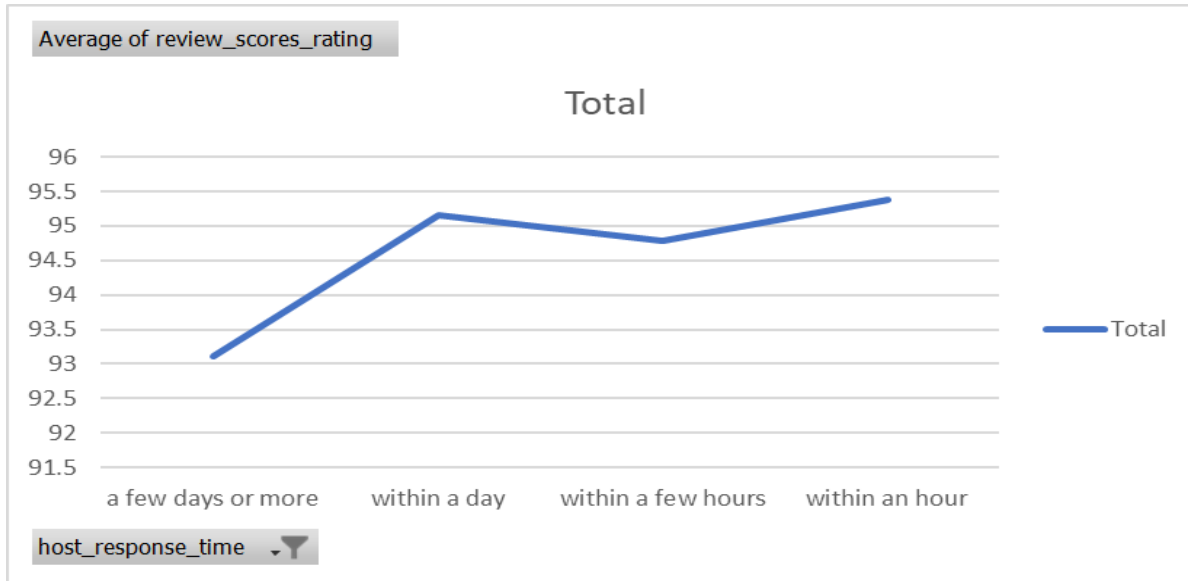
RATING ,CANCELLATION_ POLICY AND PRICE ANALYSIS



Inferences:

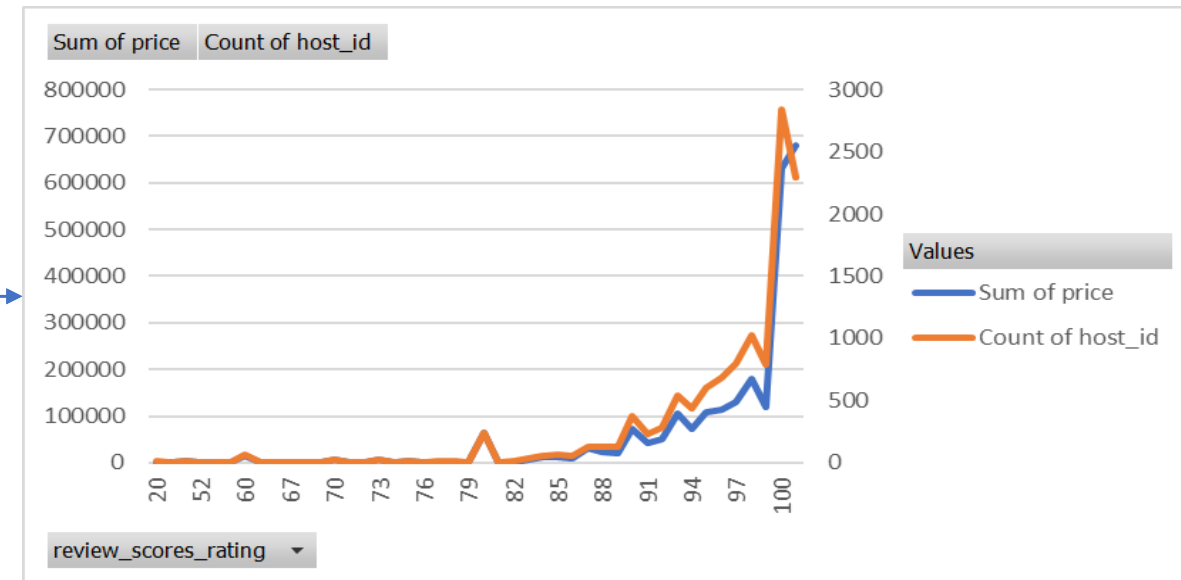
1. Lowest rated is category “strict” with the least sum of price meaning that not many customers prefer that category.
2. As the average price increases we can see that the rating is reduced relatively
3. Category “strict_14” has highest prices with rating almost equal to the top rating implying that between 2018-2019 most listings were in this category and hence is very popular.
4. Categories “Flexible” and “moderate” have the highest ratings with considerable price implying they are very reliable options for customers and hosts

HOST RATING ANALYSIS

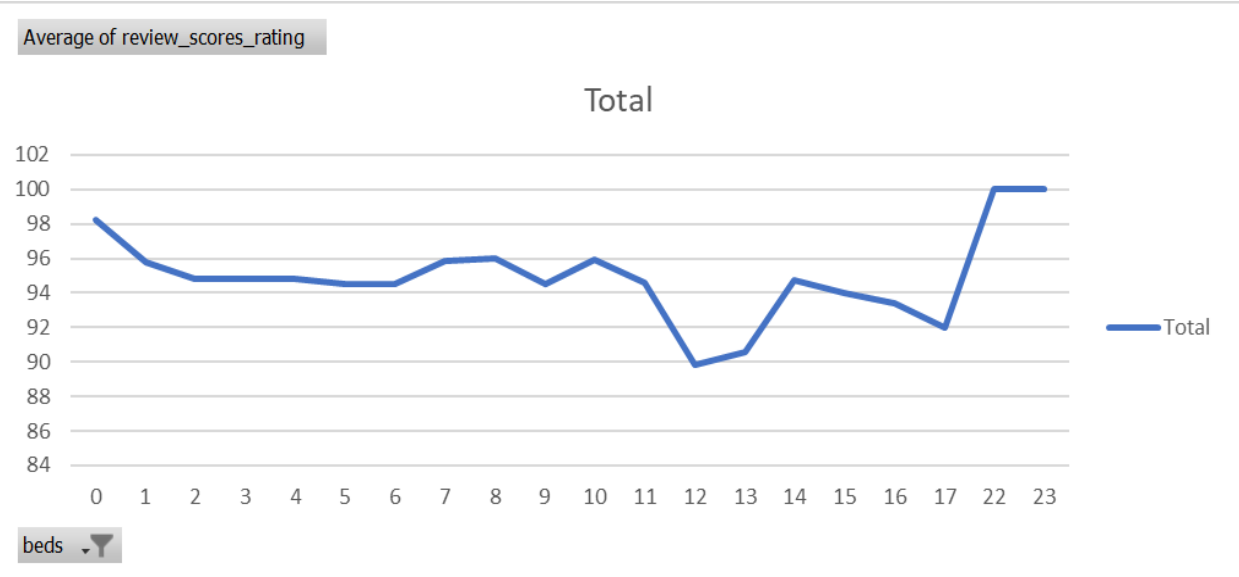
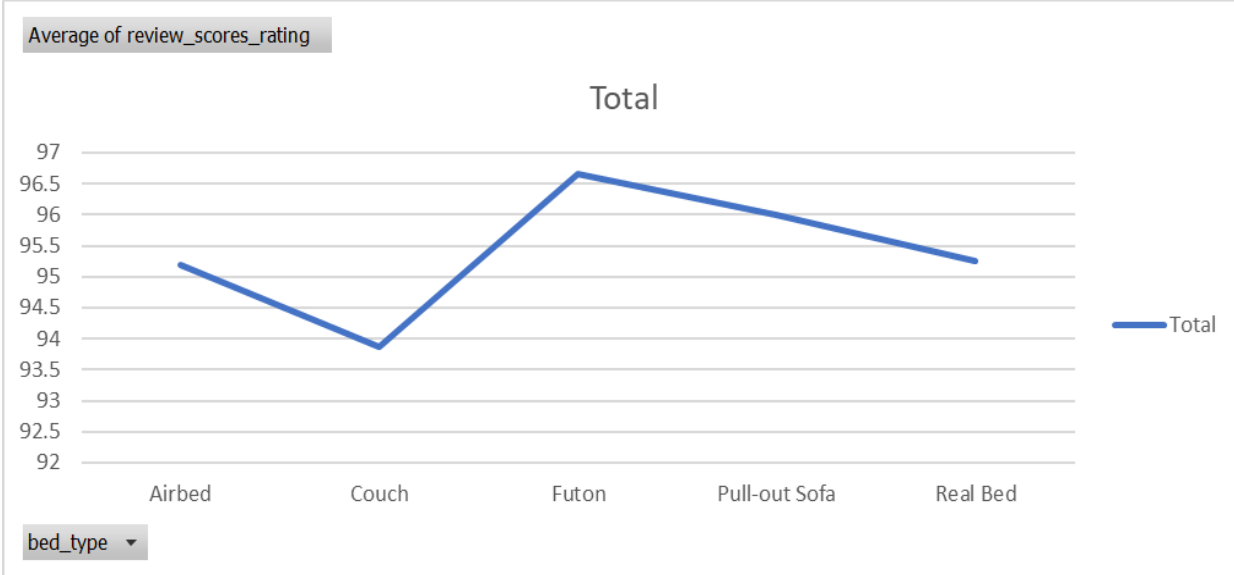
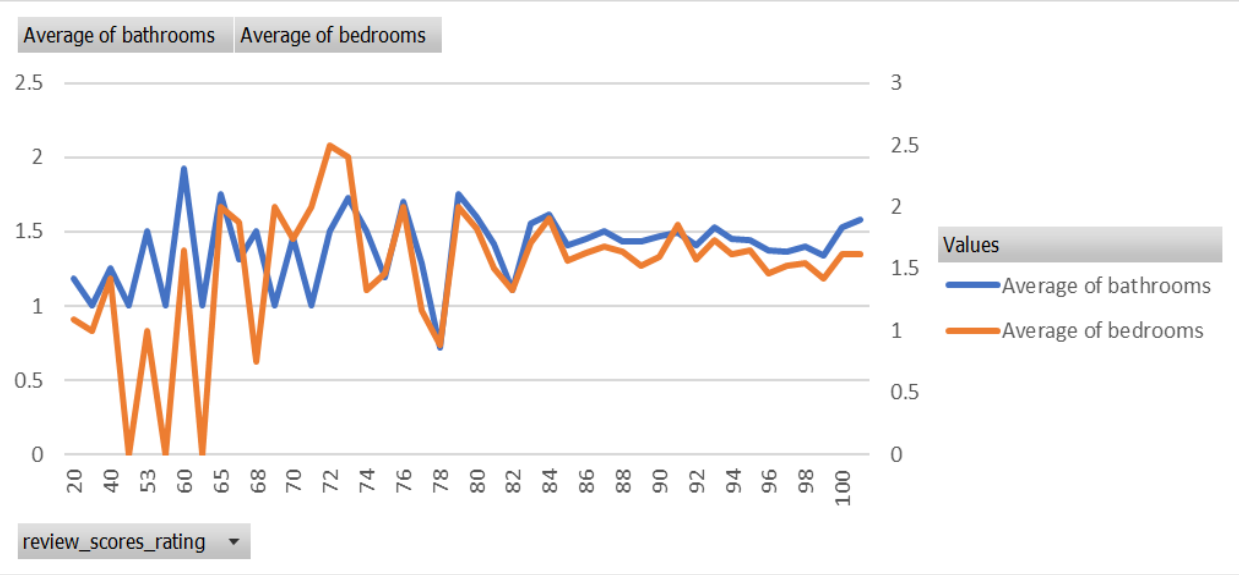


Inferences:

1. Hosts with minimum response time have the highest average rating.
2. Average rating of super_hosts is higher than that of non-super_hosts
3. Higher the rating higher the price the host charge.

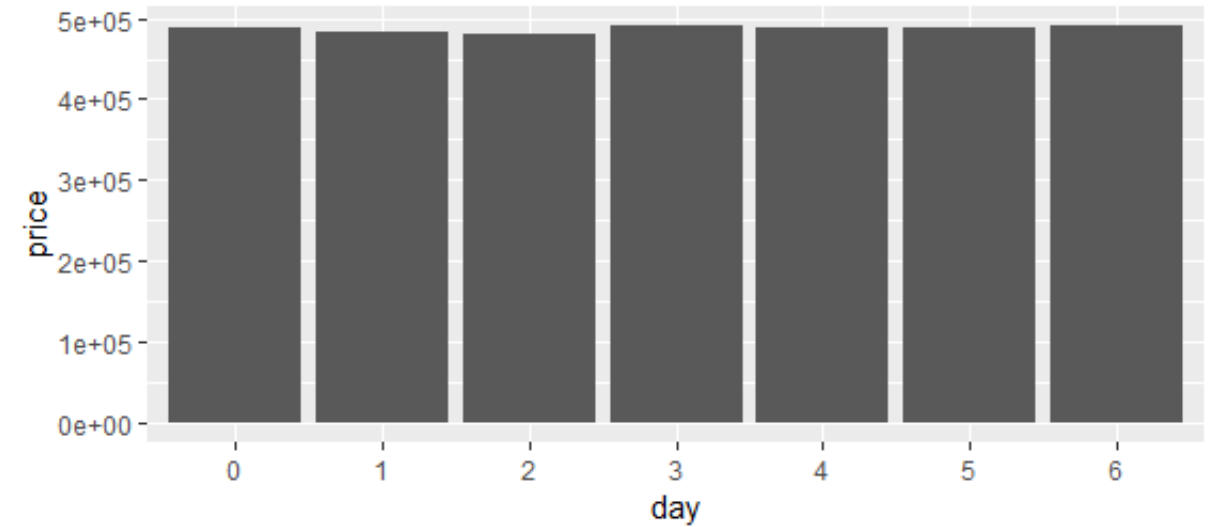
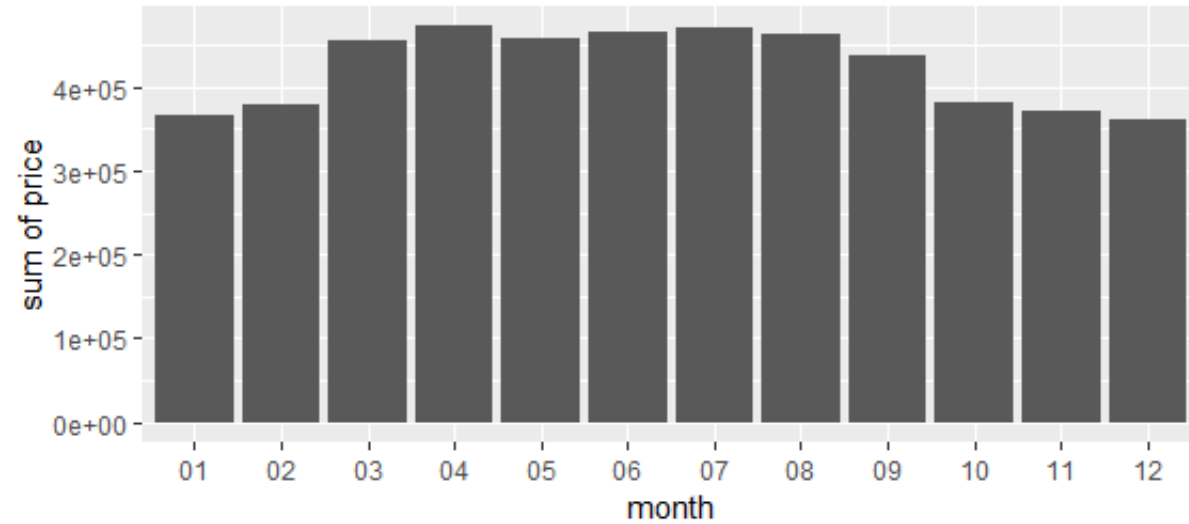


BATHROOM,BEDREOOM AND BED TYPES



1. From bed and bedroom charts its visible that the number of bedroom/beds/bathroom is not majorly affecting the average rating but there are other factors. Ex-maintenance of bed/bathrooms is more important than number of beds/bathrooms
2. From the bed_type chart its visible that all types of bed have a very good rating with “Futon” having the highest.

Analysis on calender.csv



We can see that the aggregated price across months is similar with very low seasonality

We can see that the aggregated price across days of the week is similar with very low seasonality

From this we can decide that there is no seasonality component to the number of listings/room bookings

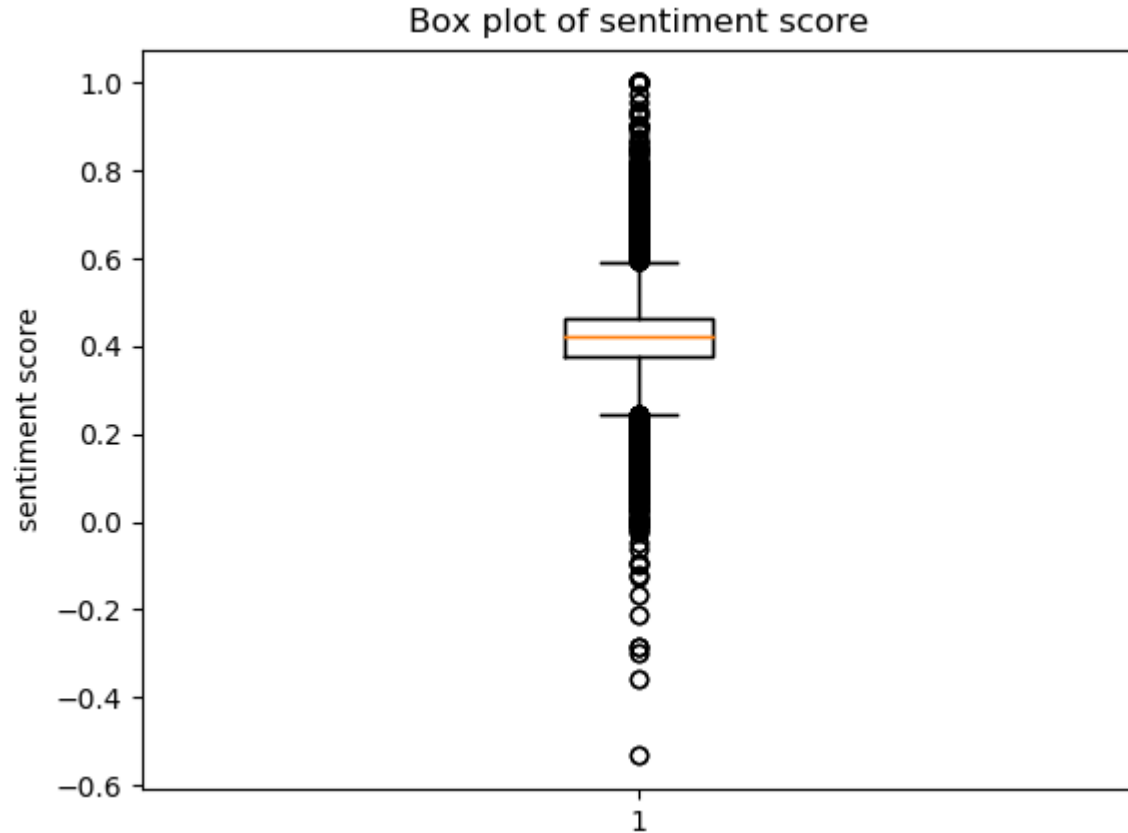
A word cloud visualization of customer feedback for the San Diego Marriott Hotel. The words are arranged in a circular pattern, with 'great' and 'value' being the most prominent. Other visible words include 'comfortable', 'space', 'good', 'neighborhood', 'within', 'beats', 'bed', 'hotels', 'affordable', 'everything', 'minutes', 'stay', 'way', 'diego', 'awesome', and 'san'.

This is the word cloud one specific review.
This is to show how the word cloud works.



This is the word cloud of all the reviews.
We can look at the most commonly used words
and infer that the overall reviews highlight the
positives of the places

SENTIMENT ANALYSIS OF REVIEWS



Sentiment is calculated for all the listings. Listings with same listing_id are grouped together and then sentiment score is calculated.

1. It is clear from the box plot that the major chunk of the reviews have a very positive sentiment meaning customers are satisfied with the services and facilities provided.
2. The number of listings with negative sentiment are less in number compared to the positive ones. We can focus on these listings and see what had caused this thereby improving the ratings, price etc for the hosts.

THANK YOU