Business Intelligence Project Report

Vaibhav Sabhahit- B00759606

Introduction:

Objectives of the project:

- Through the Boston Airbnb data provided how can we best gauge the interests and needs of the customer and use this to maximize sales.
- Through the Airbnb Listing's reviews provided, perform sentiment analysis and understand the needs and demand of the customer

Ex-ante strategy: I believe the future will be based on demand based dynamic pricings. Hence the Idea is to analyse the demand of various listing, listing characteristics and come up with a dynamic pricing strategy, whereby we charge more for greater demand features/ characteristics.

Some assumptions to fill in the gaps of the data.

1. It is assumed that every customer who has made an Airbnb reservation has also left a review for the listing. Hence each review is taken as one reservation. We use number of reviews as a proxy for reservations as we do not have the reservation number per listing provided directly.
2. Since we do not have information of the revenue generated by each Listing we make the assumption that every customer has stayed at the listing for at least the minimum number of nights specified. Hence **revenue generated= (Price of listing * minimum nights* number of Reviews(Reservation))**

Using the above two assumptions we are able to formulate the number of reservations and revenue generated for each Listing.

Strategy for objective one:

- As in introduction we want to see how Airbnb has been performing since its inception. Another important measure to check initially is the number of Listings getting added to Airbnb's offering over the years. These are two key metrics which would indicate to us how Airbnb has been performing in the last few years.
- We then move on to gauging the demand of various aspects of the listings. Some key characteristics we want to analyse:
  1. Location: Can we determine if for more popular neighbourhoods (where revenue is the highest / where the location rating highest) we can charge more. Has this strategy already been implemented?
  2. Property type: Is there a specific property that keeps getting booked more often than other property types. Is there a similarity for this trend across all the neighbourhoods?

3. Demand anticipating: Is there a pattern followed by the reservations every year(seasonality) .If year which months have highest demand and which months have the lowest.
4. Number of people accommodated in each listing: Is there an ideal sweet spot that customers prefer? Can we target this number and increase our offerings accordingly?
5. How man optional guest's additions must each listing have? Again is there an optimal number?

Based on this above key metrics we draw conclusions and make recommendations.

Strategy for objective two:

- Looking at the data one key problem that comes to light is the imbalance in the Sentiment conveyed through reviews. 95% of the listings have net positive reviews. Another 4% of the listings have net neutral reviews and less than 1% of the listings have net negative reviews. Hence our strategy is aimed at looking at what makes the customers provide a positive review. And if we can charge a premium for this.
- We check the correlation of sentiments portrayed through the reviews with different ratings Ex: Cleanliness, Communication, Location.
- Lastly we check if the 'Superhost' status provided by Airbnb in general has a more positive response than 'non superhost' listings.

Description of the data:

1) Listings.csv – Contains data above each of the listings present in Boston.
   - Listing id- Primary key unique to each listing
   - Property type- Type of the listing, house/apartment/hotel room etc
   - Latitude/ Longitude- The exact location of each Listing in Boston
   - Accommodates: Number of people who can stay in the Airbnb
   - Bedrooms: Number of rooms present in the Listing
   - Guests included: Number of additional guest who can be accommodated based on request.
   - Review_score_cleanliness/communication/checkin/location- The rating the customer has provided for various characteristics of the Airbnb
2) Reviews: Contains the reviews left by the customer for each of the listings.
   - ID: unique identifies to each review present n the table
   - Listing ID: indicated which listing the review is for
   - Reviews: The textual reviews from customers
   - Customer ID: Identification number of each customer
3) Generated fields: As per assumption
   - Compound: The sentiment score generated for each review
   - Sentiment: The category each sentiment belongs to(pos/neg/neu)
   - Estimated revenue: Calculated as per the formula mentioned earlier

Data Modelling and preparation:

Initial data engineering in Python.

Some data cleaning steps

- All the null values in the numerical fields are replaced by zero. And in all the character fields we replace the null fields with '0'.
- In the Superhost and is_location_excat field we replace the true and false values with 1 and 0
- In the cleaning_fee and price field we remove the $ symbol preceding the numeric dollar values.

Listings table:

1) Calculating the Sentiment of each comment: The Reviews were present in the reviews table. We had to generate the net sentiment associated to each Listing in the Listings table.
   For this we use the Vader Sentiment library which is a pre trained library which can generate the compound sentiment for a text that is passed to it.
   We Initially merge both the Reviews and Listings tables then calculate the compound sentiment for each review. Since we require the mean sentiment for each listing, we group by the listing ID and take the mean of the compound sentiment.

2) We also calculate the Revenue for each in a similar manner. Once we merge the the Listings and review tables we calculate the Estimated_revenue field= Minimum_nights*price.

Date_count table:

1) We need to generate the number of bookings each date had. For this we take the reviews table and group by date and take count of the ID. ID is the unique identifier in the reviews field.

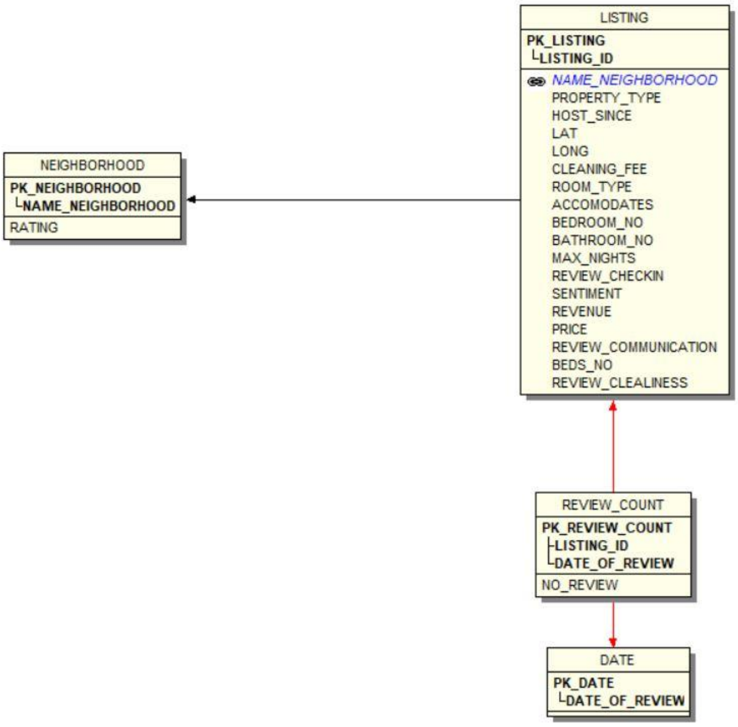
Hence the information from Reviews table has been incorporated into the Date_count and the Listings_detailsall table.

We generate two different tables

1) Listings_detailsall.xlsc: Contains Listing data consolidated over all the years
2) Date_count.xlsc :table which contains the time series data of the number of reservations made over the years.

Schema design:

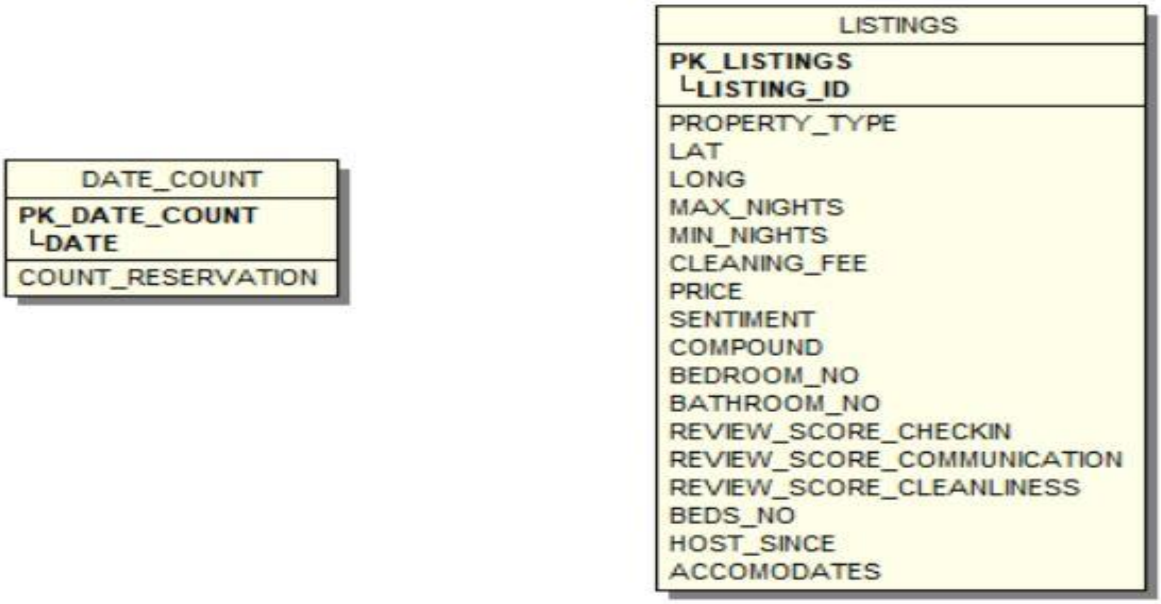
Initial proposed Schema :



Problems Faced with this design:

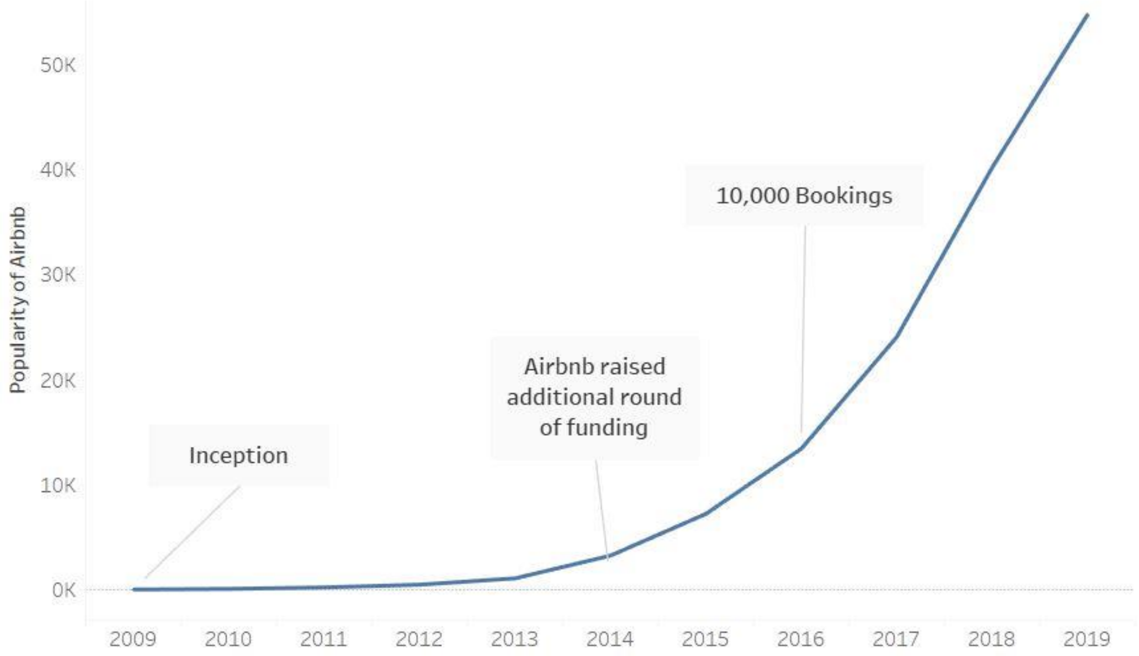
The Listings table had contain data only with respect to the listing in a consolidated manner. So when we create a hierarchy with the Review_count table the listings values incorporated the date field multiplies and the data we receive is not for each listing consolidated.

Hence the proposed model:



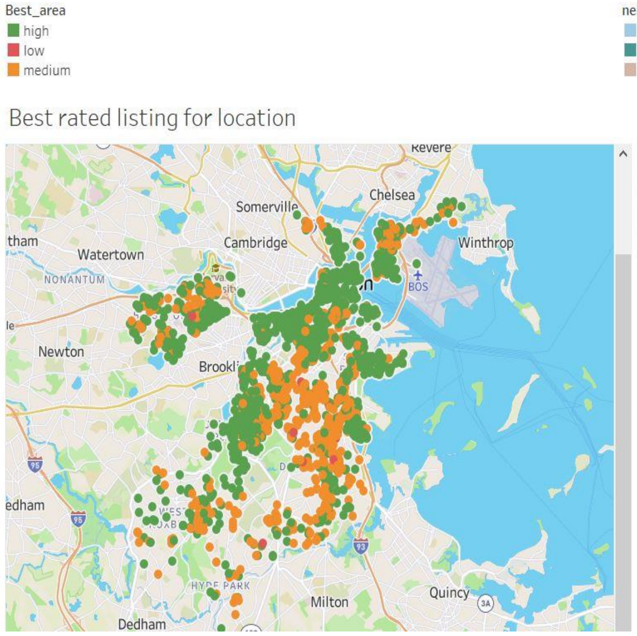
Application and findings:

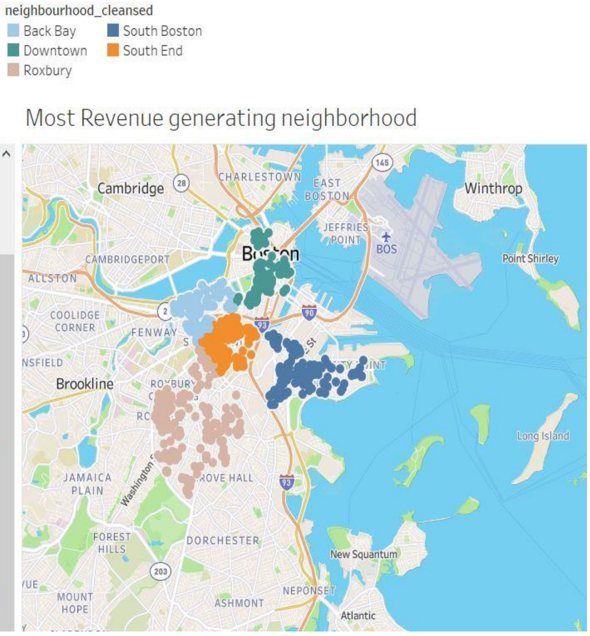
1) Popularity of Airbnb:



The number of Reservations in Airbnb have been increasing through its inception.

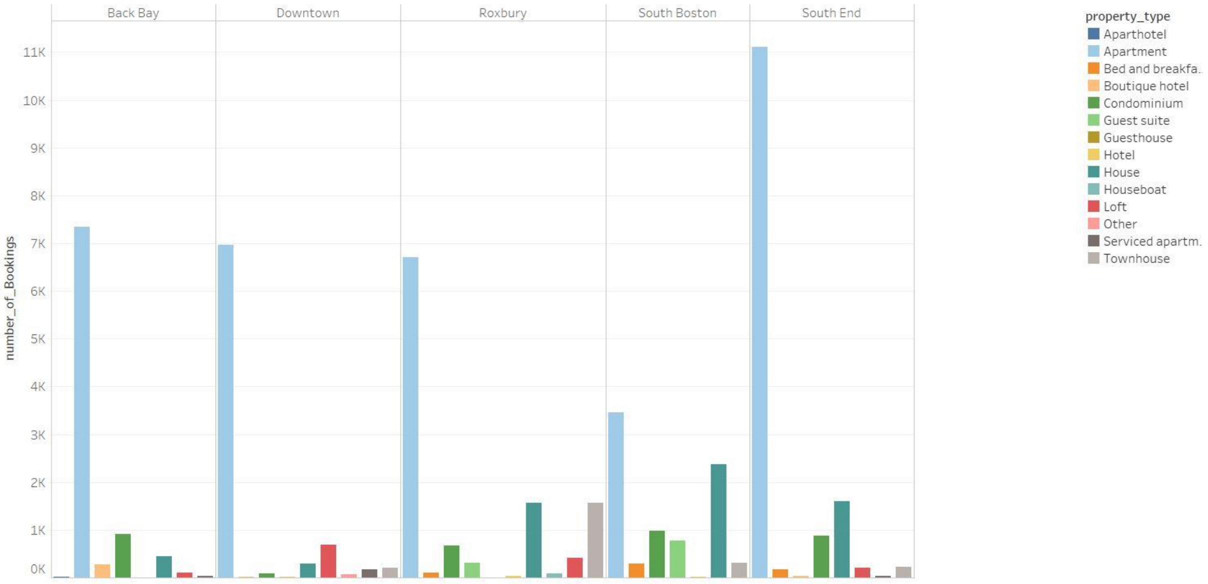
2) Location contribution to demand:





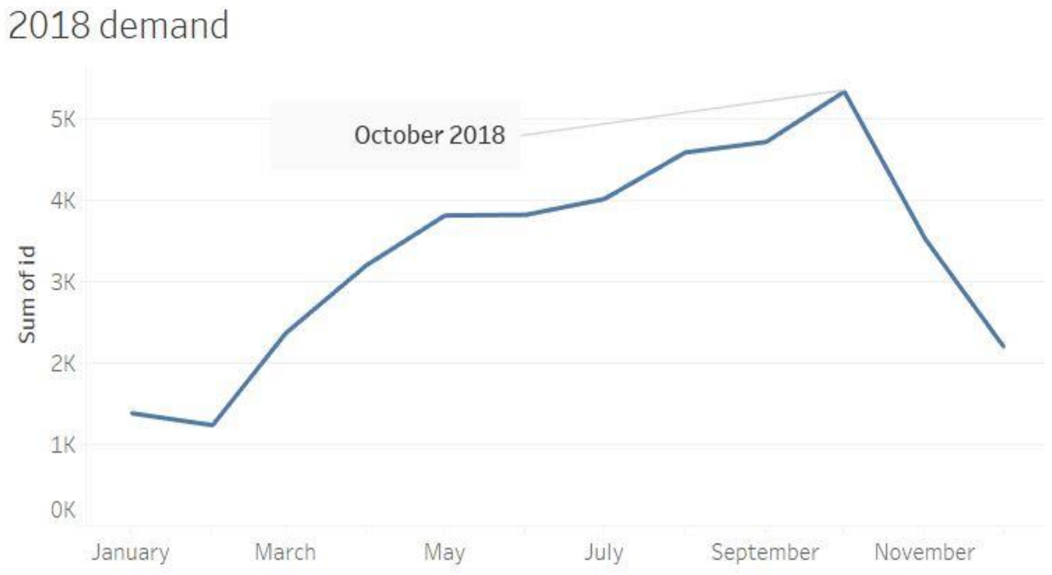
On the left we have the Listings represented according to their location ratings. And on the right we see the top five highest revenue generating neighbourhood. From this we can see that some neighbourhoods are more popular than others and people are willing to pay a premium to stay in them. As seen in the tableau file this is indeed true in the pricing structure of the popular neighbourhood listings. Hence we concluded that location is key for dynamic pricing as people are willing to pay more for popular locations.

3) Type of Property:



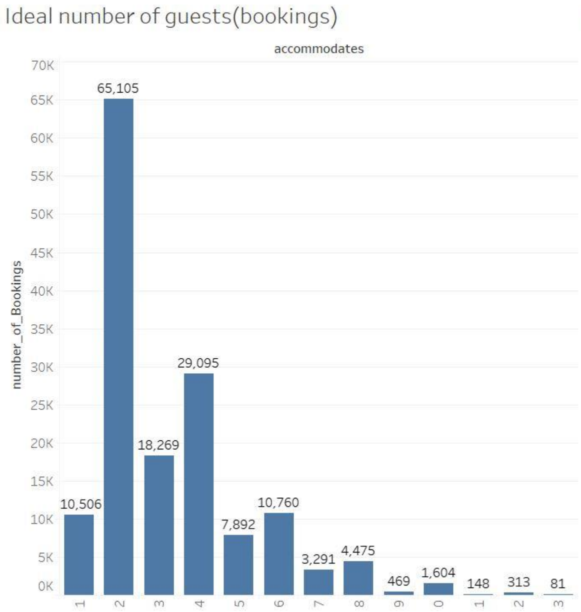
Above we have the property type count for the top 5 highest performing neighbourhoods in terms of revenue. One common pattern we notice across all the property types is that apartments is the highest sort after. The demand for apartments is significantly higher across neighbourhoods. This can be leveraged in our dynamic pricing strategy.

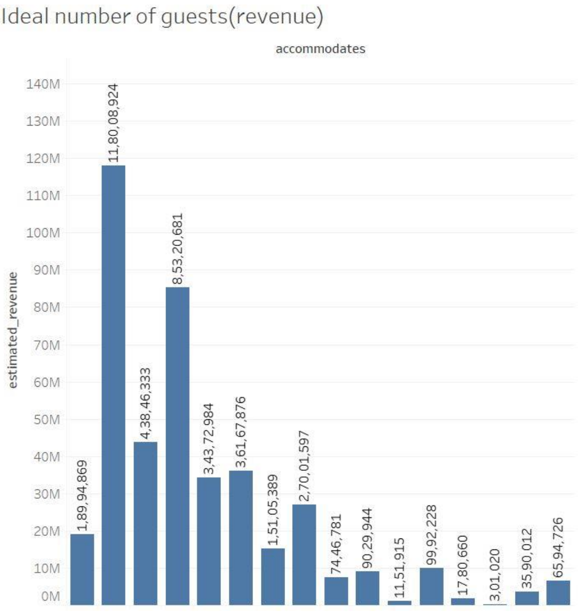
4) Seasonal demand:
We plot the reservations made in a year and analyse the trend.



In the tableau workbook we see the a similar pattern for all years- 2014,2016,2019. Hence we clearly see that June-October is when the demand for reservations is the highest. October is peak season. January, February, November, December are specific months where the demand is low. This is an interesting pattern as intuitively we would assume November, December and January being the holiday season would have high demand. Again this pattern of demand can be leveraged in our strategy.

4)Number of people accommodated in a listing:



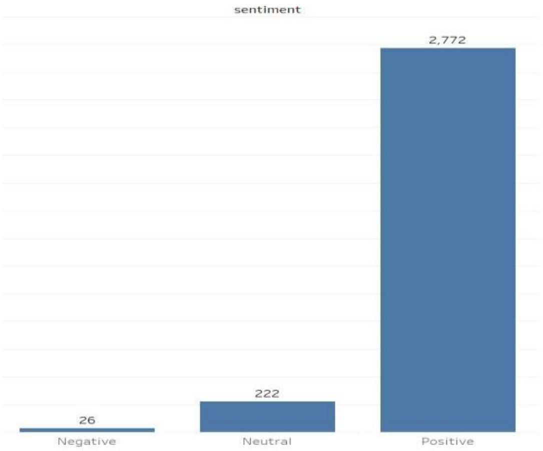


On the left hand side we can see the popularity of the count of people that can be accommodated in a listing. And on the right we see the count that generates the highest revenues. Now Based on these numbers we can see that popularity of listings accommodating 2-4 people is the highest. Hence we can conclude that apartments which accommodates 2-4 people is popular and we can charge a premium or this.

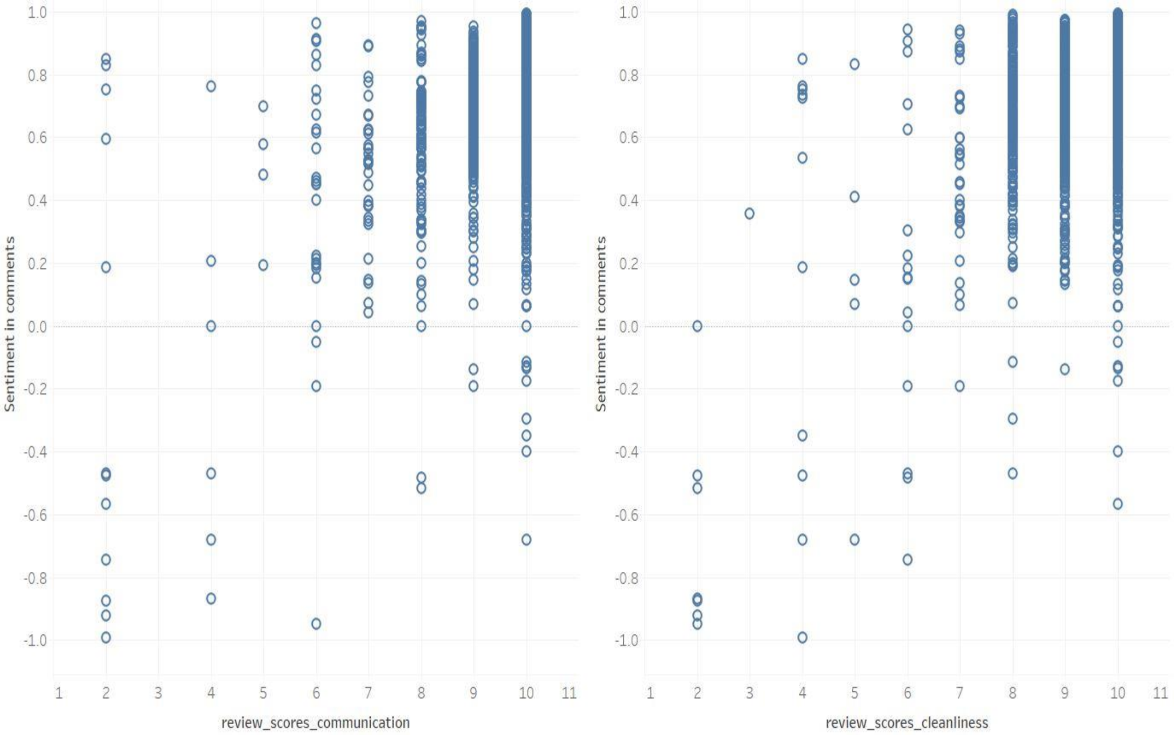
Additional guests: The same process is implemented even for Additional guests and we see that having 1-2 additional guests is the optimal condition.

Sentiment analysis:

Since the sentiment obtained from the review analysis for most listings was highly skewed towards positive/neutral , we stick to looking at what exactly made the customers give positive reviews.



We check the relation of sentiment with cleanliness and communication of the Listing and found a very high correlation with sentiment.



We also want to see if there is any relation between the location the listing is located and the review given by the customer.

Map representing the distribution of Negative Sentiment review listing in Boston.

From the above map we can see that there is no specific relation between negative sentiment of a listing and its location. All the listings are located evenly throughout the map.

Conclusions:

Demand Based Pricing model: Customers pay a premium price for when the Listing/ Location/ Characteristic of Listing is in high demand.

Some useful insights for this model:

- Location is key, **customers are willing to pay a significant premium for a popular location** as compared to the locations which are not well known.
- **Seasonality**: June- October is the peak season Listings can charge a premium for these months. November to December are off season months hence this time can be used for either maintenance of the property of listing on Airbnb for a low price.
- Apartments are the highest sort after property, hence we can charge a premium for this type of property when compared to the others.
- The ideal accommodation capacity for each listing is 3-4 people, such offerings across property type must be significantly increased.
- Ideally each property must be able to accommodate 1-2 guests extra. There can be a premium charged for this.

Sentiment Analysis:

- There is a high correlation between the sentiment generated and the cleanliness and communicating of the listing.
- Locating does not have an impact on the reviews given by the customer.
- Airbnb has already set a standard measure called superhost. The following are the requirements.

  -Completed at least 10 trips OR completed 3 reservations that total at least 100 nights
  -Maintained a 90% response rate or higher
  -Maintained a 1% percent cancellation rate
  -Maintained a 4.8 overall rating (Cleanliness, communication)
- Most people would prefer a staying in a location where the listing has obtained superhot status. Hence a premium can be charged even with respect to this.

Improvement suggestions:
1)Perform topic/aspect analysis of the reviews data that is present.