SOULPAGE

IT SOLUTIONS

San Francisco Airbnb Price Prediction Analysis

airbnb

San Francisco Airbnb Price Prediction

Problem statement:-

Pricing a rental property on Airbnb is a challenging task for the owner as it determines the number of customers for the place. On the other hand, customers have to evaluate an offered price with minimal knowledge of an optimal value for the property. We have several thousand records of data available from Airbnb with tens of Features that help us to predict the price of the property.

Opportunity:-

As we have several thousands of records and a few features, we have the opportunity to bring out some interesting answers with the Data.

- Who are the Top10 hosts in the san Francisco Airbnb?
- What are the Top5 host locations?
- How fast All the Hosts are responding?
- What are the Top 10 host neighbourhood locations?
- What are the Available rooms count for different groups of people?
- What are the Top5 property types that are easily available?
- What are the different room types available in Different properties?
- How are the prices concerning Properties?
- Price ranges for Number of Accommodates concerning Room types?
- What are the different types of Amenities provided by hosts?
- Do Bed types make a difference in price?

And many more in the document.

Tools used:-

- Python 3.6
- Pandas
- Matplotlib
- Variance Inflation Factor
- XGBRegressor
- Tpot Regressor

Dataset:-

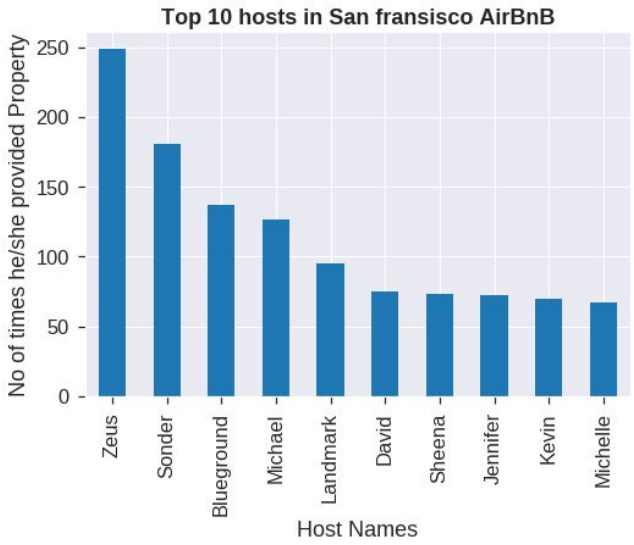
We have datasets available from InsideAirBnb here is up to 8th July 2019. This has approximately 7000 rows/records with 106 columns/features of file size 37.2MB.

Process:-

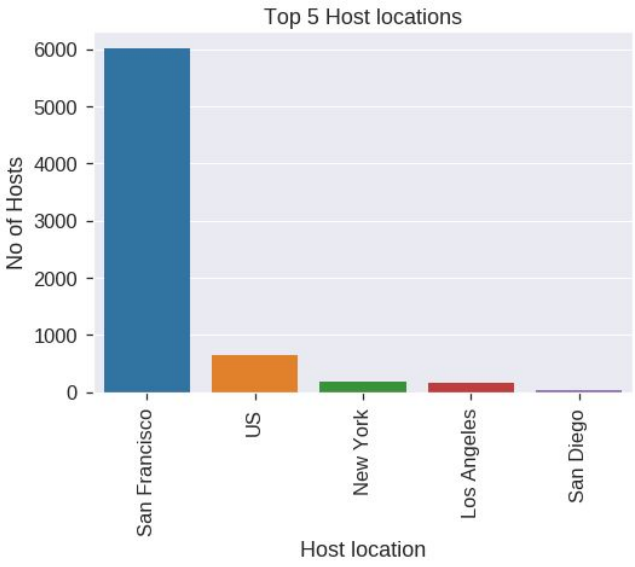
We used all the above mentioned tools to get answers to our questions and then we used Machine learning models like Linear regression, Lasso regression, Ridge regression, Random forest, XGboost, and Tpot Regressor to predict the price of the property using features like No of Bedrooms, No of Beds, Reviews, Host response time, Host response rate and many more.

Who are the Top10 hosts in Sanfrancisco Airbnb?

We can see from the insight below, the top 10 hosts that provided property/space for people to book are Zeus, Sonder, Blueground and others.



What are the Top 5 Host locations?



The Top5 locations we have hosted are from San Francisco, Newyork, Los Angeles, San Diego. As we have the Head office of Airbnb in San Francisco, so we have more hosts from the same location.

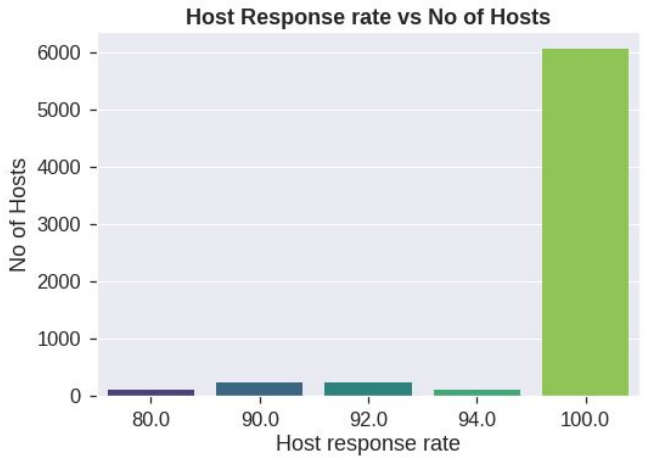
How fast the Host response to Guests?

It would be interesting to know how fast Hosts are responding to Guests call to have their property so that it helps for New guests to take their Decision as per host response. We almost have 5000 number of hosts that respond within an Hour, and others you can see from the graph below.

.

Number of hosts with Response rate 100%

From the graph below we will be able to understand that we have almost 6000 Number of hosts with the Response rate of 100%, which is 90% of Sanfrancisco Airbnb hosts. As we have more Number of Response rate from the hosts, Guests would likely to make more bookings respectively.

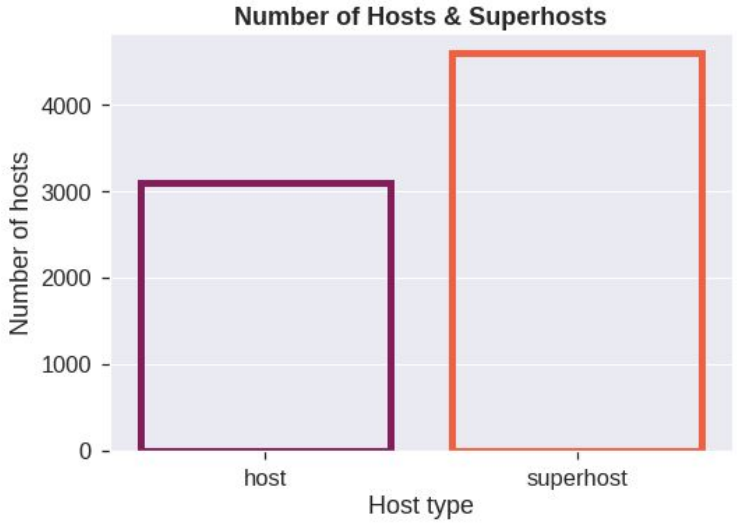


Number of Superhosts in San Francisco Airbnb?

Airbnb provides super host tag only to the hosts who are with them for 10 years. Also

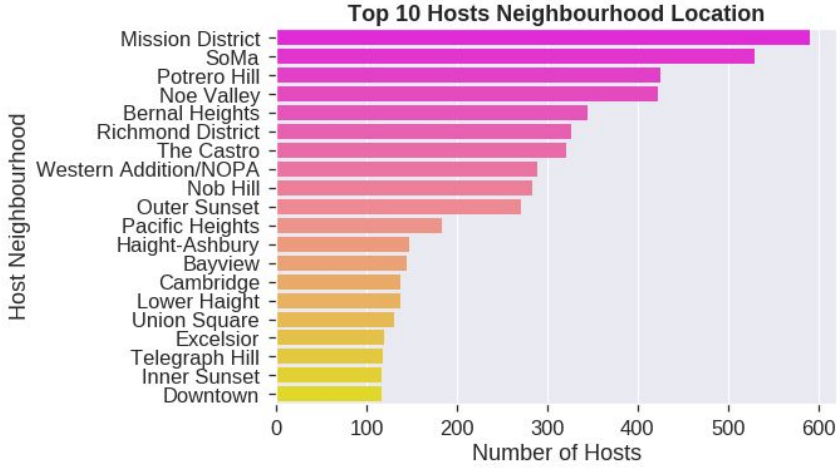
- Respond to guests quickly and maintain a 90% response rate or higher
- Have at least 80% 5-star reviews
- Honour confirmed reservations (meaning hosts should rarely cancel)

From the graph below we can see that we have more than 4000 hosts with the Superhost tag which is 70% of the total number of hosts. This makes Guests have Good Experience when staying in Superhosts property.



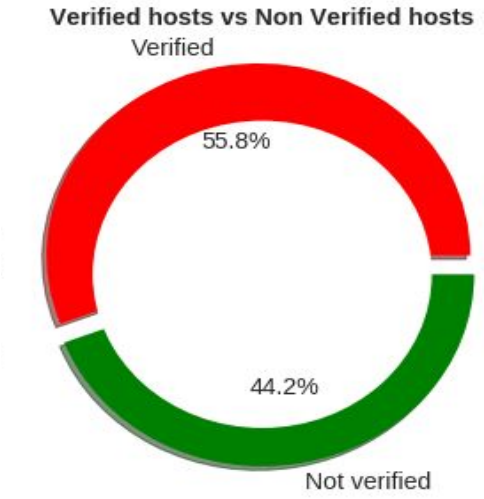
What are the Top Host Neighbourhood Locations?

We have Mostly the Hosts from the location Mission District, SoMa, Potrero Hill, and other locations are as shown in the figure below.



How Many of the Hosts Identity is verified?

Airbnb verifies host Identity using Government ID, Phone, Email, Reviews, and many other ways. From the graph below we can see that we have 55.8% of hosts verified till 8th July 2019.

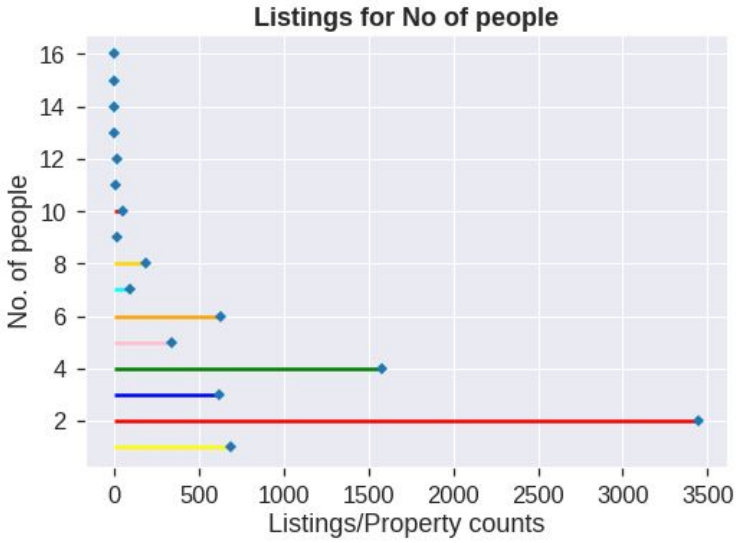


Do all the hosts provide a Correction location?

We have More than 6000 hosts with Exact Location mentioned about their property on the Airbnb website. So More than 1000 hosts have provided wrong information about their property location that can mislead Guests whoever wants to book that property.

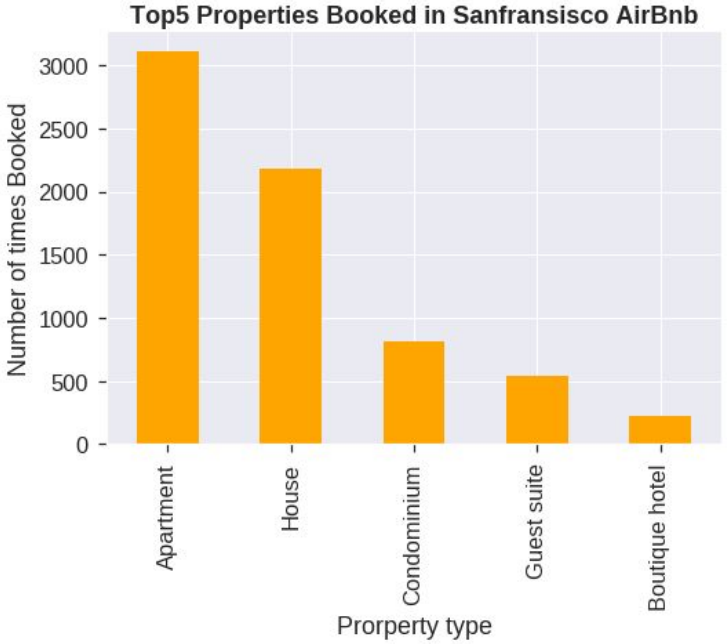
How many Properties/Listings available for Different groups of People?

We have Different Groups of people/Accommodates that book properties to have their accommodation are from 2 people to 16 people are shown in the graph below. From the graph below we have almost 3500 property/spaces available for 2people. We have more than 1500 spaces available for 4 people to live.



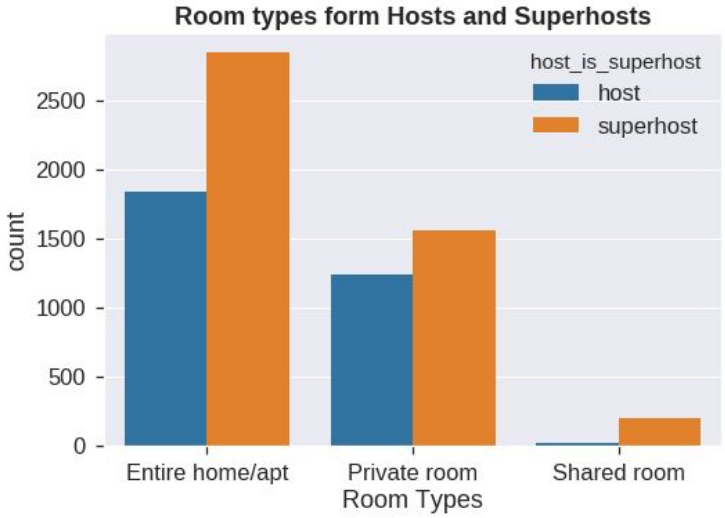
What are the Top 5 Properties getting Bookings?

We have several property types Apartments, Hostel, Guesthouse, Villa, Boat and Many more. Out of all, We can see that Apartments having More bookings, followed by House, Condominium, Guest suite, and Boutique hotel.



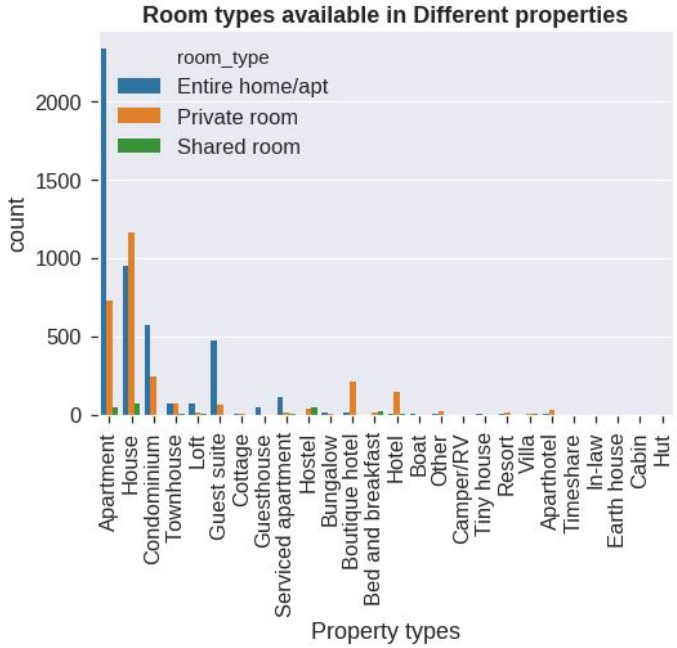
The number of Hosts and Superhosts providing different Room types?

We have different types of Rooms available provided by hosts and super hosts. Down below we can see that in every Room Type super hosts are more in the count.



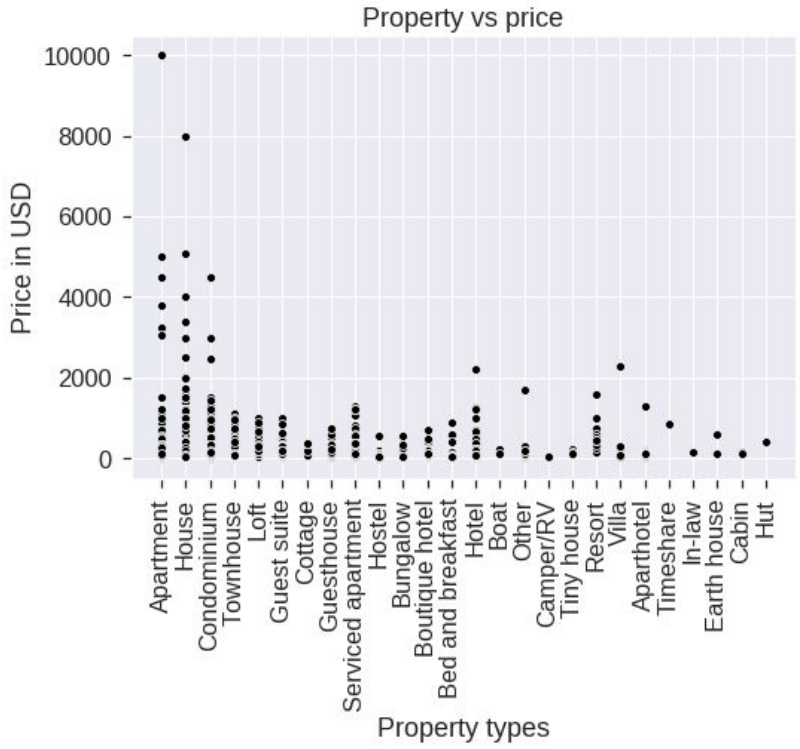
What is the Most taken Room type in Every Property?

We can see that we have an Entire home/Apartment as room type more in Apartments. We also can see that we have Private room types more in houses, Entire homes in Condominium, and Entire home in Guest suite.



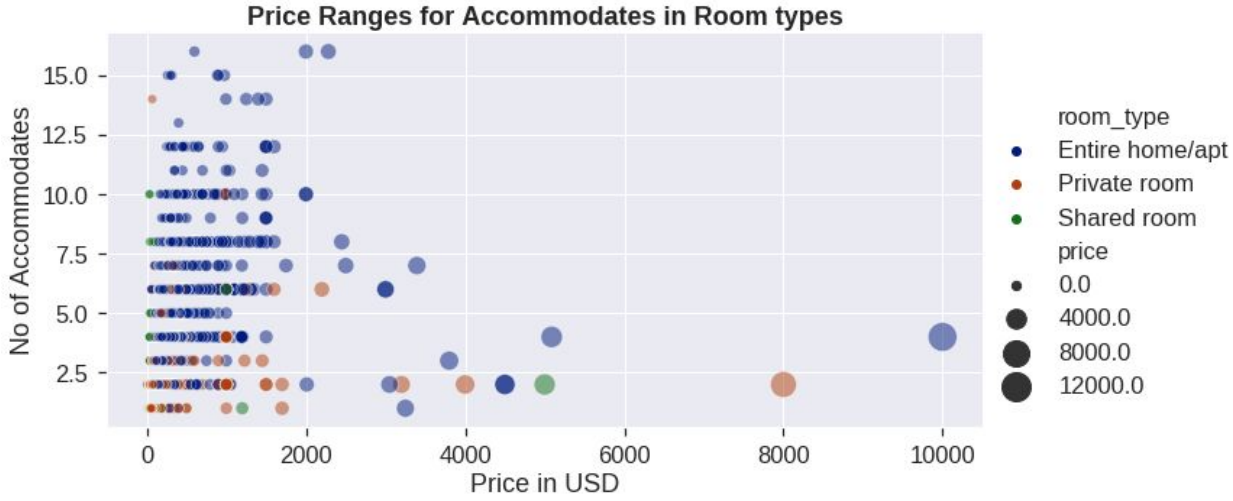
What are the Different Prices for different Properties?

From the graph below we can understand that most of every property is in the Price range below 2000$. We have the highest price for the property type Apartment with 10000$.



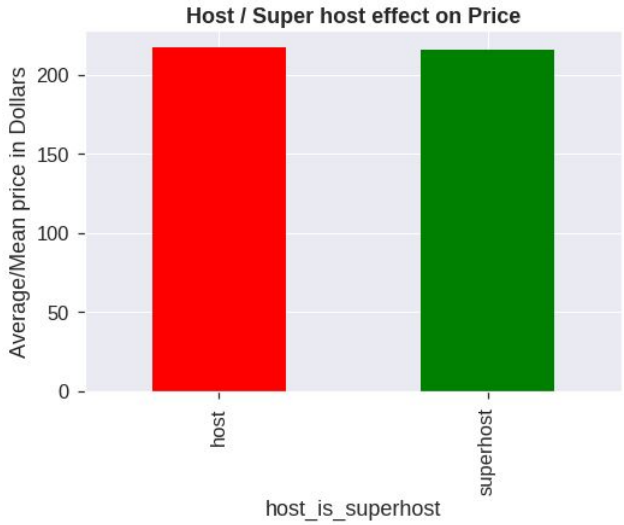
Price ranges for Accommodates in Different Room types:-

This graph below tells that Room types with the Price range for Different sets of Accommodates. We can see the Blue colour bubble at the price of 1000USD for 4 Accommodates. We can visualize like same for others.



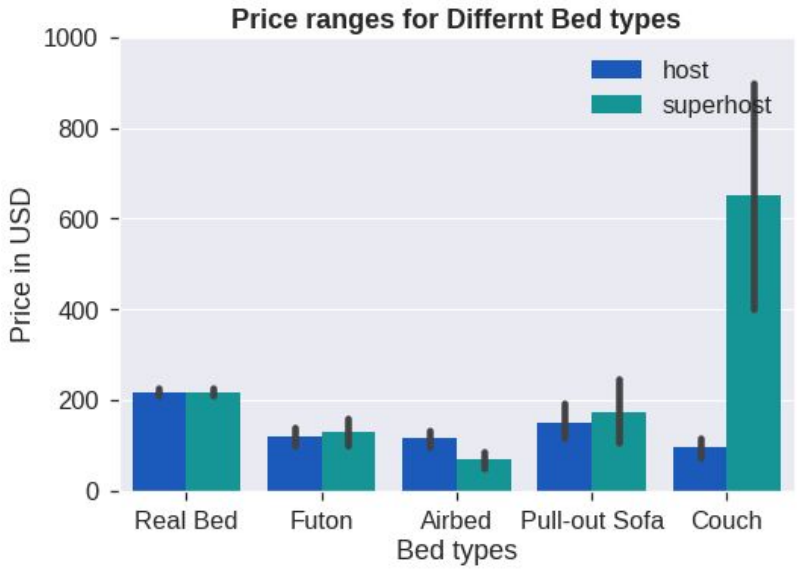
Does Being Host and Superhost show effect on price?

We can infer from the graph below that being Host or Super host has no effect on the price. The plot below tells us the same. Both host and super host are charging almost the same price which is little more than 200 USD of Average/Mean price.



Do bed types make a difference in price?

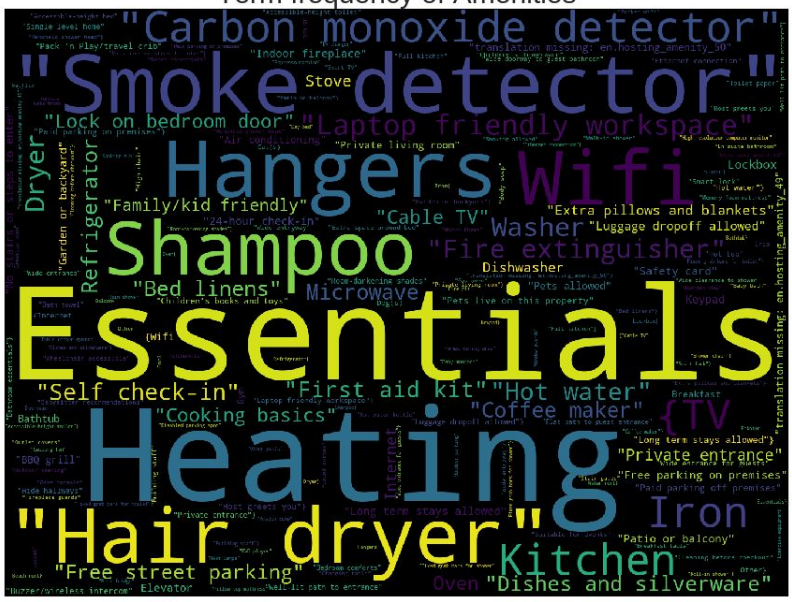
From the graph below we see that we have hosts and Superhosts providing Real Bed for just 200 USD of Average price. But when it comes to Couch, we have Superhosts charging more with the price of more than 600 USD.



What are the More provided Amenities for Guests?

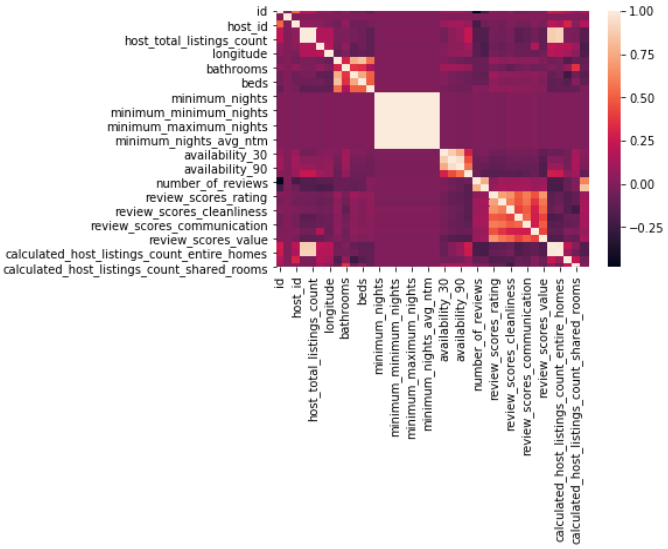
The below Wordcount plot says that we have Heating, Hairdryers, Hangers, Shampoo, Smoke detectors and some essentials along with wifi.

Term frequency of Amenities



Model Building: Price Prediction

We have several columns/Features available in our dataset. There are some Preprocessing steps before you head to model building like **Filling/Removing Null values, Column renaming, Changing Data types, Cleaning text, Outlier Removals,**



Removing Highly correlated variables, Label encoding, One- hot encoding and many more.

VIF:-

We did all the preprocessing steps on the Dataset and finally, we got 34 columns or features. We performed the Variance Inflation Factor which detects Multicollinearity in our Dataset by estimating the variance of Regression coefficient inflated. Generally, VIF values within the range of 1 to 5 are considered as our variables are moderately correlated.

We did Data manipulation steps like Standardising the data which brings all our data into the same scale range that helps our model to run fast by optimizing the computation and could also increase accuracy.

Feature importance:-

Few Tree-Based models and XGB has Feature importance parameter, that can show the importance to each variable in building the model. Features with less importance can be removed. We also have a Feature importance chart for our Data shown below

We Used Regression models like linear regression, Lasso regression, Ridge regression, Random Forest Regression, XGBoost Regression on our data. With the **XGBoost Regression model**, we were able to achieve an accuracy of **61.56%.**

Parameter tuning:-

We used TPOT Regressor that can automate our Machine learning workflow and finds out the best Model with Best parameters by doing feature engineering, feature scaling, null value imputation, one-hot encoding and stacking on its own.

After performing this We got XGB Regressor with parameters like

- learning_rate=0.1
- max_depth=9
- min_child_weight=7
- n_estimators=100
- nthread=1

Increased the accuracy to **65.41%**.

The results are as shown below

S No	Actual Price in USD	Predicted Price in USD
1.	307.0	312.78
2.	86.0	90.44
3.	225.0	234.85
4.	375.0	363.99
5.	50.0	219.89
6.	199.0	179.12
7.	143.0	181.27
8.	225.0	194.01
9.	120.0	175.34
10.	220.0	279.80

Conclusion:

Out of all the Regression models we had XGBRegressor giving high accuracy of 65.41%. Further, the accuracy of the model can be increased by doing Sentimental Analysis/Natural Language processing on the text data we have like reviews, summary, and Description columns that can reach up to 69% which is benchmark prediction of now.