PREDICTING AND ANALYSING AIRBNB DATASET

Prof. Guide: Dr. Alok Kumar

Contributors: Prachika Kanodia, Ishika Gupta, Akshi Agarwal

ABSTRACT

Airbnb has become increasingly popular among travelers for accommodation across the world. In this project, we aim to predict Airbnb listing price, to find the spike in accommodation price during peak and off-peak seasons and to find the review score with the help of sentimental analysis in four different cities- Boston, Amsterdam, Hong Kong and Athens with various machine learning approaches. After doing price prediction using various ML algorithms it was noticed that Random Forest and Naive Bayes Classification algorithm give the highest accuracy when compared with actual price. After finding the review score and doing sentimental analysis of all the cities we can say that the Airbnb reviews are almost similar across different cities. Most tourists leave positive reviews and use similar positive words to describe the Airbnb houses. After determining that the peak season for these cities is October, with the exception of Hong Kong, which has a busiest time in April. In addition, there is a significant price difference between off-season and peak-season hotel rates.

INTRODUCTION

Airbnb is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking. Airbnb's arrival is without a doubt one of the most significant and revolutionary recent developments in the global tourism industry. Despite the fact that Airbnb has only been around for about ten years, the firm has released a timely innovation by changing the age-old practice of peer-to-peer accommodation with a new technology-driven distribution network.

OBJECTIVE

1. To predict and validate the price of different cities of different continents and compared it to recommend the best according to the need of Airbnb users and non-users.

1.1 To visualize that where to invest in a property to get the maximum number of returns from Airbnb.

1.2 To visualize that which Room Type is most and least expensive and come under which Property Type and Neighbourhood of Boston.

1.3 To visualize that which listing id has good and bad Review Score Ratings on the basis of Neighbourhood, Property Type, Room Type and Bedrooms available in the individuals.

2. To apply sentimental analysis on Airbnb dataset of different cities.

2.1 To analyse the sentiments on the dataset i.e., positive, negative or neutral.

2.2 To find the Top Hosts based on User Reviews and Top Hosts' neighbourhood.

3. To predict the spike in accommodation prices during peak and off-peak seasons of different cities.

3.1 To find the most common amenities present in hotels.

3.2 To find out trends of visitors in particular cities.

3.3 To find out max and min range of particular property type.

LITERATURE REVIEW

According to author [1], the emergence of Airbnb is unquestionably one of the most significant and transformative recent developments within the worldwide tourism sector. Airbnb had 140,000 guest arrivals in 2010; 800,000 in 2011; three million in 2012; six million in 2013; 16 million in 2014; 40 million in 2015; 80 million in 2016; an estimated 115 million in 2017; and an estimated 164 million in 2018. To accommodate these guests, at the time of writing the company boasted over five million active worldwide listings, which was higher than the room capacity of the top five worldwide hotel companies combined. Furthermore, it recently was estimated that if Airbnb were to go public, its market capitalization would be around $60 billion which is significantly higher than even Marriott International. Also, price is clearly an important factor as Airbnb guests assess their options, several researchers have instead examined the more general concept of value. Also, Airbnb users willing to pay a premium or to make an investment based on perceived functional and social value early on in the buying process also depending upon reviews of the neighbourhood cities.

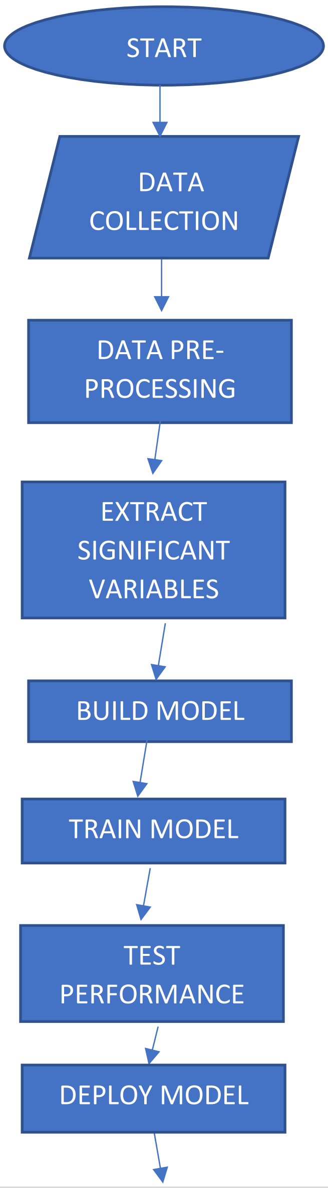
According to author [2], reviews are fundamental to the success of Airbnb, and the overall smart tourism ecosystem. Prospective guests seeking to make the optimal accommodation choice have to gather the relevant information necessary from comments and reviews by previous guests. Hence, the pressure to write reviews after staying is higher compared to staying at hotels. Unlike numeric ratings, where the scores can be easily understood and compared, albeit subjected to the various amount of biases , text reviews offer a richer and deeper set of information. Text reviews are usually presented in an unstructured manner creating the inherent issues of making it challenging to locate relevant information, also known as uncertainty, and creating confusion when there is conflicting information, or equivocality . Increasing the number of reviews can help decrease the degree of uncertainty while users can learn to recognize which information to trust. In other words, the participation in writing and sharing travel experiences is a core component in not just enhancing the Airbnb ecosystem, but the overall smart tourism ecosystem.

According to Author [1], existing Airbnb research was categorised into six subject categories: Airbnb guests, Airbnb hosts, Airbnb supply and its impacts on destinations, Airbnb regulation, Airbnb's impacts on the tourist sector, and the Airbnb corporation in the existing literature. This work fills a significant research vacuum by examining this vast, recent body of information. It's also identified a few areas of Airbnb understanding that are starting to mature as similar studies come to an agreement. The relevance of money in driving both Airbnb hosts and guests, the value of variables like room types and visitor capacities in deciding listing costs, and the geographical clustering of Airbnb listings in numerous city centres, for example, have all been discovered repeatedly. This review of the literature has both conceptual frameworks. In terms of theory implications, this review adds a new layer to conceptions of tourism lodging choice and the different elements (e.g., perceived originality) that influence such choice, as well as fresh perspectives for thinking about creativity and value co-creation. In terms of practical implications, this paper provides a valuable formulation of Airbnb knowledge that should be useful to Airbnb and other tourism accommodation providers in their competition for guests, Airbnb, and other peer-to-peer quick lease platforms in their efforts to attract and retain hosts, destination marketing organisations in their efforts to better cater to new tourism interests, and policymakers in their efforts to better manage the Airbnb trend.

METHODOLOGY

1. To predict and validate the price of different cities of different continents and compared it to recommend the best according to the need of Airbnb users and non-users.

We first consider different traditional ML methods using continuous and categorical features in multiclass classification. We used Multi-linear Regression, K-Nearest Neighbor (KNN) Classification, Naive Bayes Classification, Random Forest Classification, Decision Tree Classification and Ensembling (Voting Classification) Algorithms to compare the actual and predicted prices. We fit models for the above methods using only selected features which can be strongly correlated with the price. The detailed process is explained below before applying the algorithms.



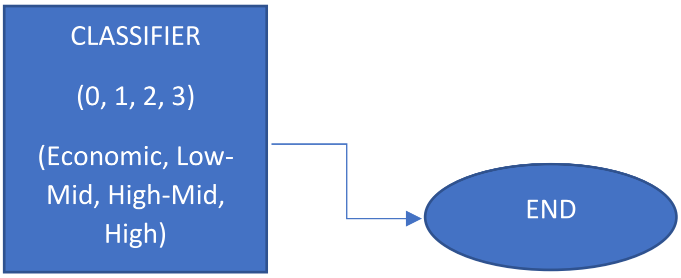


Fig 1. Process Flow Block Diagram

DATA PRE-PROCESSING:

It involves data loading, understanding, cleaning, feature engineering, etc. to get the useful data for the price predictions.

1. **Data Loading and Understanding:** Firstly, all the data has been loaded after which all the columns are analysed manually whether they are useful or not for the fulfilment of objectives.
2. **Data Cleaning:** Firstly, columns which are not strongly correlated with the price are dropped after which the null values in integer type column filled with '0' and null values in string type column filled with 'None'.
3. **Feature Engineering:** Some of the column values not able to give the proper visualization for which some conversion has been done i.e., take log values, convert dummy variables trap into binary classification and labelling continuous values into different classes using LabelEncoder().
4. **Data Reduction:** Dataset is very big so using train_test_split it has been divided into 20-80% ratio respectively in testing and training set.
5. **Data Transformation:** Converting probability values in NumPy array in 0's and 1's as per the lowest and highest value form to get the accuracy score using testing set.

ALGORITHMS:

1. **Multi-Linear Regression:** It is used to model the relationship between two or more features and a response by fitting a linear equation to observed data. We can use it to find out which factor has the highest impact on the predicted output i.e., price having continuous values and how different variables relate to each other.
2. **K-Nearest Neighbor Classification:** It is basically a classification algorithm which belongs to the supervised learning category. In this K is specified i.e., labels of multi-classification. It predicts the result on the basis of the majority.
3. **Naive Bayes Classification:** It is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation. This algorithm is a fast and good fit for multi-class prediction. It can be built using Gaussian distribution. This algorithm is scalable and easy to implement for a large data set.
4. **Random Forest Classification:** It is based on supervised learning which can be used for both regression and classification problems. It is viewed as a collection of multiple decision trees algorithm with random sampling. It includes the random selection of features. The idea is to make the prediction precise by taking the average or mode of the output of multiple decision trees. It is he shortcomings of Decision Tree algorithm i.e., greater the number of decision trees is considered; the more precise output will be.
5. **Decision Tree Classification:** It is a part of classification algorithm which also provides solutions to the regression problems using the classification rule (starting from the root to the leaf node) where each leaf node is used to represent the class label (results that need to be computed after taking all the decisions) and the branches represents conjunctions of features that lead to the class labels. It is a tree-like graph where sorting starts from the root node to the leaf node until the target is achieved. It is the most popular one for decision and classification based on supervised algorithms.
6. **Ensembling (Voting Classification):** It is defined as the multimodal system in which different classifier are strategically combined into a predictive model. It also helps to reduce the variance in the predicted data, minimize the biasness in the predictive model and to classify and predict the statistics from the complex problems with better accuracy.

FORMULAS AND FUNCTIONS:

1. **R2 Score, Neigh Score, Accuracy Score:** It tells about the variation in target variable means accuracy of the model used for price prediction i.e., the ratio of correct prediction to total prediction.
2. **Mean Absolute Error, Mean Square Error, Root Mean Square Error:** It tells about the variation in target variable means error of the model used for price prediction.
3. **Classification Report:** It contains the following values:

- Precision = (TP)/(TP+FP)
- Sensitivity (Recall) = (TP)/(TP+FN)
- Specificity (Support) = (TN)/(TN+FP)
- F1-Score (Precision and Recall) = (2PrecisionRecall)/(Precision+Recall)
- Accuracy and Average i.e., Macro and Weighted.

4. **AUC ROC Score and Curve:** AUC i.e., Area Under ROC Curve is a measure of ability of model to discriminate positives and negatives correctly. ROC Curve i.e., Receiver Operating Characteristics which is a plot between Sensitivity (TP rate) and 1-Specificity (FP rate) in vertical and horizontal axis respectively. If the area under ROC is:

- 0.5 No discrimination
- 0.7 <= ROC area < 0.8 Acceptable discrimination
- 0.8 <= ROC area < 0.9 Excellent discrimination
- ROC area >= 0.9 Outstanding discrimination

EXPLORATORY DATA ANALYSIS:

In this step, Data Visualization is done using different types of plots like Bar Graph, Donut Chart, Pie Chart, Correlogram, Heatmap, Scatter Plot, Box Plot, Frequency Curve, World Map, etc. to get visuals output of the objectives mentioned above.

In most of the plots, target value i.e., price is compared with different features which are strongly correlated with it which are depicted in the result section.

1.1 To visualize that where to invest in a property to get the maximum number of returns from Airbnb.

In this plotting is done for comparing the Neighbourhood, Room Type and Property Type with the Price so that one can able to find the property in particular city according to the facilities in demand to make the investment in it to get the maximum returns in the near future.

1.2 To visualize that which Room Type is most and least expensive and come under which Property Type and Neighbourhood of Boston.

In this plotting is done to know that which Room Type is most and least expensive and come under which category of Property Type and in which Neighbourhood of Boston.

1.3 To visualize that which listing id has good and bad Review Score Ratings on the basis of Neighbourhood, Property Type, Room Type and Bedrooms available in the individuals.

In this plotting is done to know the Review Score Ratings depending on various factors like Neighbourhood, Property Type, Room Type and Bedrooms so that it is easy to validate that whether Price matches according to Review Score Ratings or not.

After which various algorithms of regression and classifications are applied for which Confusion Matrix and ROC Curve is generated along with some of different kinds of plots correlated with price and used for its predictions.

2. To apply sentimental analysis on Airbnb dataset of different cities.

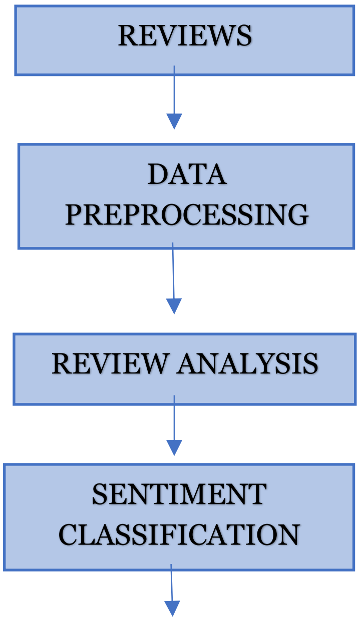


Fig 2: Sentimental Analysis Model

DATA PRE-PROCESSING:

1. Merging the reviews dataset with the listing dataset and dropping columns which are unnecessary.
2. Checking for null and empty values
3. **Word Frequency Analysis:** One of the key steps in NLP or Natural Language Process is the ability to count the frequency of the terms used in a text document or table. In this step words are split to count the frequency.

	words	count
0	the	524864
1	and	494980
2	a	337239
3	to	304281
4	was	222760

4. Removing punctuations, special characters and numbers.
5. **Removing short words:** To remove all the words having length 3 or less. For example, terms like “hmm”, “oh” are of very little use. It is better to get rid of them.
6. **Tokenization:** Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP.

```
0    [Daniel, really, cool, place, nice, clean, Ver...
1    [Daniel, most, amazing, host, place, extremely...
2    [such, great, time, Amsterdam, Daniel, excelle...
3    [Very, professional, operation, Room, very, cl...
4    [Daniel, highly, recommended, provided, necess...
5    [Daniel, great, host, made, everything, easy, ...
6    [Daniele, amazing, host, provided, everything,...
7    [have, nicer, start, Amsterdam, Daniel, such, ...
8    [Daniel, fantastic, host, place, calm, clean, ...
9    [Daniel, great, couldn, enough, gone, trouble,...
```

7. **Stemming:** Stemming is a method of removing the suffix of the word and bringing it to a base word. Stemming is the normalization technique used in Natural language processing that reduces the number of computations required.
8. **Stop words:** Stop words are the most commonly occuring words which are not relevant in the context of the data and do not contribute any deeper meaning to the phrase. In this case contain no sentiment. NLTK provide a library used for this.
9. **Language Detection:** To detect the language used in the comments we import langdetect.
10. **Sentimental Analysis:** VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. Vader not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

    The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).

    positive sentiment : (compound score >= 0.05)
    neutral sentiment : (compound score > -0.05) and (compound score < 0.05)
    negative sentiment : (compound score <= -0.05)

ALGORITHM:

Natural Language Processing (NLP):

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Sentiment analysis, also refers as opinion mining, is a sub machine learning task where we want to determine which is the general sentiment of a given document. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinion about a selling objects, or predict stock markets for a given company like, if most people think positive about it, possibly its stock markets will increase, and so on. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- With the help of EDA we found the neighbourhood with the most number of listings.
- Compared the rating scores in terms of different aspects like Location, Check In, Cleanliness and Communication.
- Visualized the most frequent words being used in the reviews
- Found the most frequent language being used in the reviews section.
- Visualized the most spoken English comments with the help of WordCloud.

3. To predict the spike in accommodation prices during peak and off-peak seasons of different cities.

EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- With the help of EDA, we found the months with the most number of visitors.
- With the help of EDA, we found the average price in particular month.
- Find out price range of particular property type.
- Visualized the most common amenities.

RESULT

1. To predict and validate the price of different cities of different continents and compared it to recommend the best according to the need of Airbnb users and non-users.

In the Regression algorithm price has continuous values but in Classification algorithms price values using LabelEncoder() converted into multi-classes using ranges from continuous value i.e., price_range_category for the better prediction i.e., **0 = economic (< 100.0), 1 = low-mid (>= 100.0 & < 250.0), 2 = high-mid (>= 250.0 & < 600.0), 3 = high (>= 600.0).**

BOSTON (USA)

The correlation of Price with the Final Features considered for its Prediction and also after doing feature engineering with the number_of_reviews and room_type column because they are not giving proper visualization and correlation with the price because of the variation in their values.

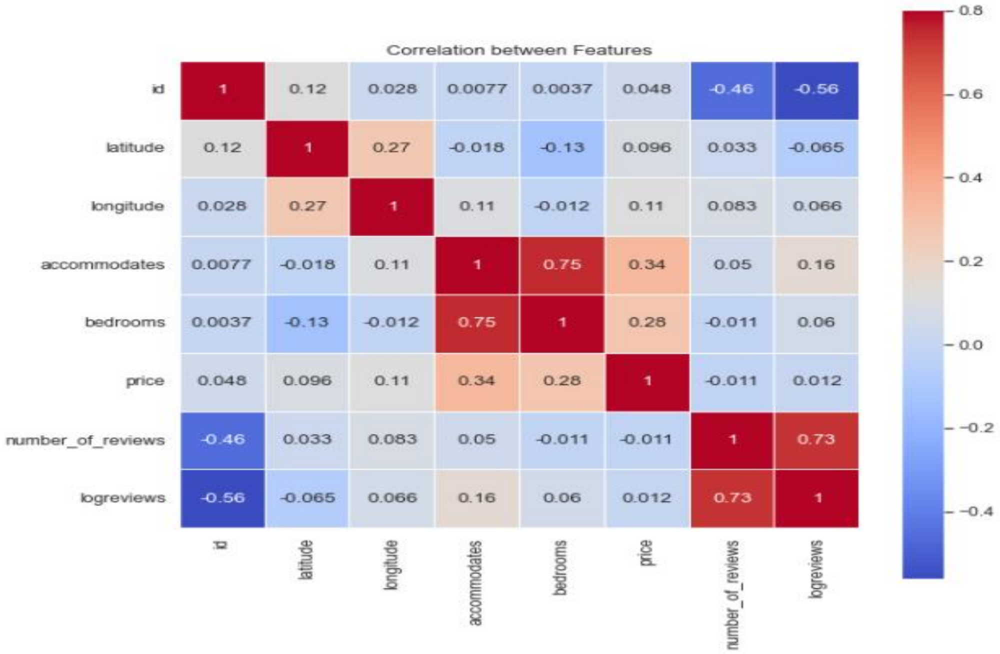
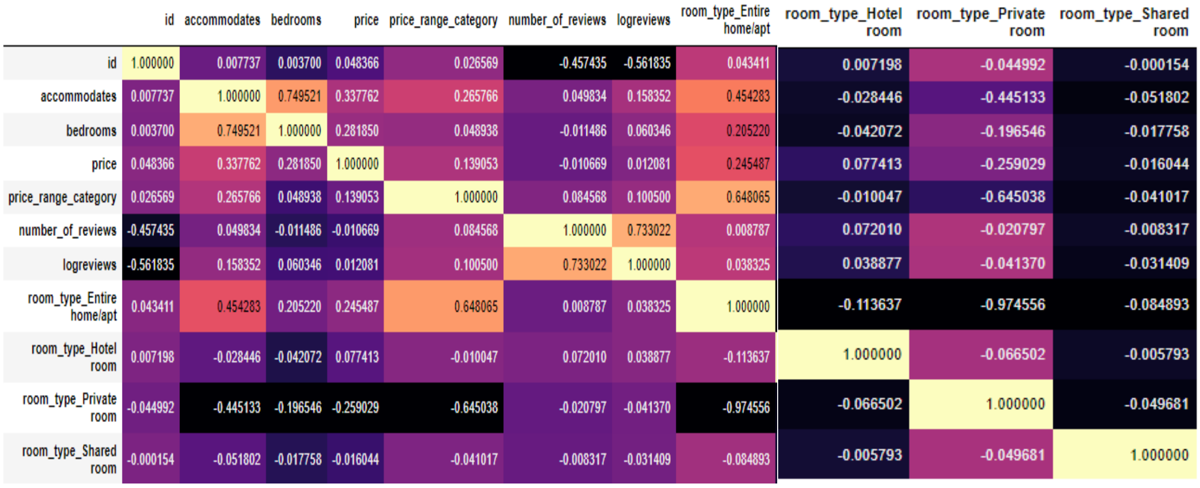


Fig. 3 Correlation between Features



	id	accommodates	bedrooms	price	price_range_category	number_of_reviews	logreviews	room_type_Entire home/apt	room_type_Hotel room	room_type_Private room	room_type_Shared room
id	1.000000	0.007737	0.003700	0.048366	0.026569	-0.457435	-0.561835	0.043411	0.007198	-0.044992	-0.000154
accommodates	0.007737	1.000000	0.749521	0.337762	0.265766	0.049834	0.158352	0.454283	-0.028446	-0.445133	-0.051802
bedrooms	0.003700	0.749521	1.000000	0.281850	0.048938	-0.011486	0.060346	0.205220	-0.042072	-0.196546	-0.017758
price	0.048366	0.337762	0.281850	1.000000	0.139053	-0.010669	0.012081	0.245487	0.077413	-0.259029	-0.016044
price_range_category	0.026569	0.265766	0.048938	0.139053	1.000000	0.084568	0.100500	0.648065	-0.010047	-0.645038	-0.041017
number_of_reviews	-0.457435	0.049834	-0.011486	-0.010669	0.084568	1.000000	0.733022	0.008787	0.072010	-0.020797	-0.008317
logreviews	-0.561835	0.158352	0.060346	0.012081	0.100500	0.733022	1.000000	0.038325	0.038877	-0.041370	-0.031409
room_type_Entire home/apt	0.043411	0.454283	0.205220	0.245487	0.648065	0.008787	0.038325	1.000000	-0.113637	-0.974556	-0.084893
room_type_Hotel room	0.007198	-0.028446	-0.042072	0.077413	-0.010047	0.072010	0.038877	-0.113637	1.000000	-0.066502	-0.005793
room_type_Private room	-0.044992	-0.445133	-0.196546	-0.259029	-0.645038	-0.020797	-0.041370	-0.974556	-0.066502	1.000000	-0.049681
room_type_Shared room	-0.000154	-0.051802	-0.017758	-0.016044	-0.041017	-0.008317	-0.031409	-0.084893	-0.005793	-0.049681	1.000000

Fig. 4 Correlation between Features after Feature Engineering

Now the various algorithms are applied for which some important parameters are calculated included in tables below.

Table 1: Price Predicted values compared with Actual values calculated using different algorithms.

Here, A – Actual and P – Predicted.

Sr. No.	**Multi-Linear Regression**		**K-Nearest Neighbor Class.**		**Naive Bayes Class.**		**Random Forest Class.**		**Decision Tree Class.**		**Ensembling (Voting Class.)**	
	A	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**
1.	231	158	3	3	3	3	3	3	3	3	3	2
2.	118	212	3	3	3	3	3	3	3	1	3	3
3.	128	275	3	3	3	3	3	3	3	1	3	3
4.	150	252	3	3	3	3	3	3	3	1	3	3
5.	196	202	3	3	3	3	3	3	3	1	3	3
6.	110	249	3	3	3	3	3	3	3	1	3	3
7.	615	590	1	1	1	1	1	1	1	1	1	1
8.	192	321	3	3	3	3	3	3	3	1	3	3
9.	163	218	3	3	3	3	3	3	3	1	3	3
10.	218	199	3	3	3	3	3	3	3	1	3	2

Table 2: Score and Error related parameters are compared generated from different algorithms for Price Predictions.

Sr. No.	Algorithm and Type	Function	Accuracy Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	ROC AUC Score
1.	Multi-Linear Regression	LinearRegression()	0.128	87.703	72832.591	269.871	-
2.	K-Nearest Neighbor Classification	StandardScaler()	0.984	0.016	0.020	0.141	0.990
3.	Naïve Bayes Classification	GaussianNB()	1.0	0.0	0.0	0.0	1.0
4.	Random Forest Classification	RandomForestClassifier()	1.0	0.0	0.0	0.0	1.0
5.	Decision Tree Classification	DecisionTreeClassifier()	0.272	1.424	3.203	1.789	1.0
6.	Ensembling (Voting Classification)	VotingClassifier()	0.658	0.630	1.458	1.207	0.98

Table 3: Classification Report generated from different algorithms used for Price Predictions.

<table>
<tr><th>Sr. No.</th><th>Algorithm and Type</th><th>Classification Report</th></tr>
<tr><td>1.</td><td>Multi-Linear Regression</td><td>-</td></tr>
<tr><td>2.</td><td>K-Nearest Neighbor Classification</td><td><pre>
              precision   recall f1-score   support

           0   1.00   1.00   1.00   203
           1   1.00   0.89   0.94   18
           2   0.96   0.96   0.96   125
           3   0.98   0.99   0.99   304

    accuracy   0.98   650
   macro avg   0.99   0.96   0.97   650
weighted avg   0.98   0.98   0.98   650
</pre></td></tr>
</table>

<table>
<tr><td>3.</td><td>Naive Bayes Classification</td><td>
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>0</td><td>1.00</td><td>1.00</td><td>1.00</td><td>203</td></tr>
<tr><td>1</td><td>1.00</td><td>1.00</td><td>1.00</td><td>18</td></tr>
<tr><td>2</td><td>1.00</td><td>1.00</td><td>1.00</td><td>125</td></tr>
<tr><td>3</td><td>1.00</td><td>1.00</td><td>1.00</td><td>304</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>1.00</td><td>650</td></tr>
<tr><td>macro avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>650</td></tr>
<tr><td>weighted avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>650</td></tr>
</table>
</td></tr>
<tr><td>4.</td><td>Random Forest Classification</td><td>
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>0</td><td>1.00</td><td>1.00</td><td>1.00</td><td>203</td></tr>
<tr><td>1</td><td>1.00</td><td>1.00</td><td>1.00</td><td>18</td></tr>
<tr><td>2</td><td>1.00</td><td>1.00</td><td>1.00</td><td>125</td></tr>
<tr><td>3</td><td>1.00</td><td>1.00</td><td>1.00</td><td>304</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>1.00</td><td>650</td></tr>
<tr><td>macro avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>650</td></tr>
<tr><td>weighted avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>650</td></tr>
</table>
</td></tr>
<tr><td>5.</td><td>Decision Tree Classification</td><td>
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>0</td><td>0.88</td><td>0.31</td><td>0.46</td><td>203</td></tr>
<tr><td>1</td><td>0.05</td><td>0.94</td><td>0.09</td><td>18</td></tr>
<tr><td>2</td><td>0.00</td><td>0.00</td><td>0.00</td><td>125</td></tr>
<tr><td>3</td><td>0.41</td><td>0.32</td><td>0.36</td><td>304</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.27</td><td>650</td></tr>
<tr><td>macro avg</td><td>0.33</td><td>0.39</td><td>0.23</td><td>650</td></tr>
<tr><td>weighted avg</td><td>0.47</td><td>0.27</td><td>0.31</td><td>650</td></tr>
</table>
</td></tr>
<tr><td>6.</td><td>Ensembling (Voting Classification)</td><td>
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>0</td><td>0.78</td><td>0.74</td><td>0.76</td><td>203</td></tr>
<tr><td>1</td><td>0.35</td><td>0.33</td><td>0.34</td><td>18</td></tr>
<tr><td>2</td><td>0.49</td><td>0.54</td><td>0.51</td><td>125</td></tr>
<tr><td>3</td><td>0.67</td><td>0.67</td><td>0.67</td><td>304</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.66</td><td>650</td></tr>
<tr><td>macro avg</td><td>0.57</td><td>0.57</td><td>0.57</td><td>650</td></tr>
<tr><td>weighted avg</td><td>0.66</td><td>0.66</td><td>0.66</td><td>650</td></tr>
</table>
</td></tr>
</table>

The Confusion Matrix and ROC Curve generated from different algorithms based on parameters mentioned above are displayed.

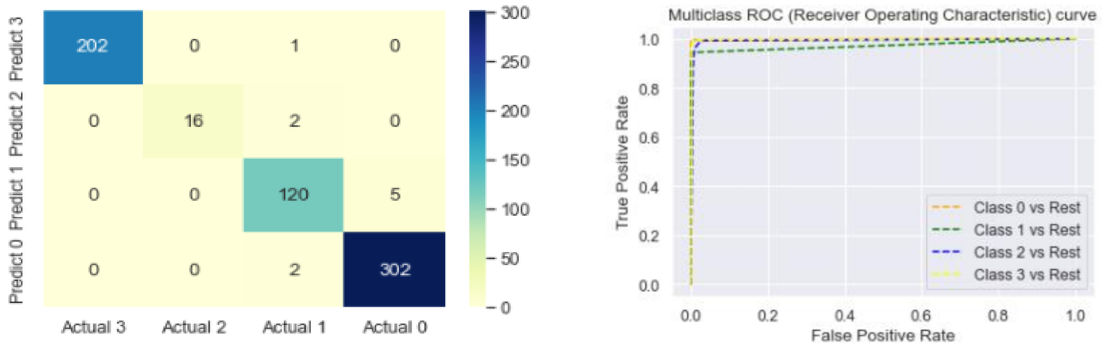


Fig. 5 Confusion Matrix and ROC Curve in K-Nearest Neighbor Classification

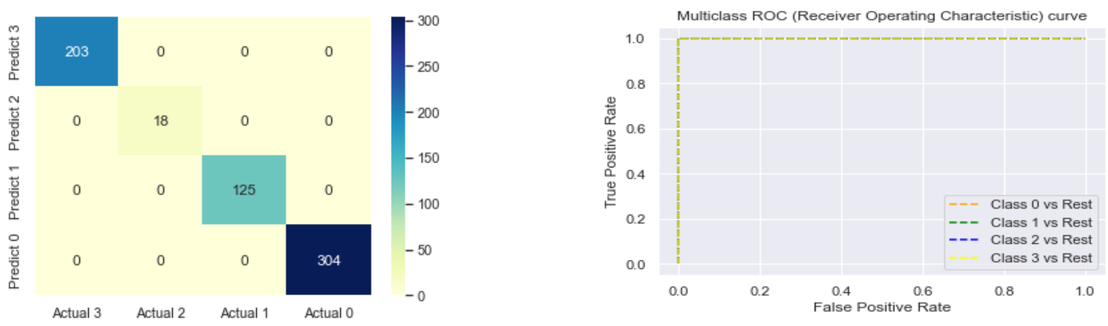


Fig. 6 Confusion Matrix and ROC Curve in Naive Bayes Classification

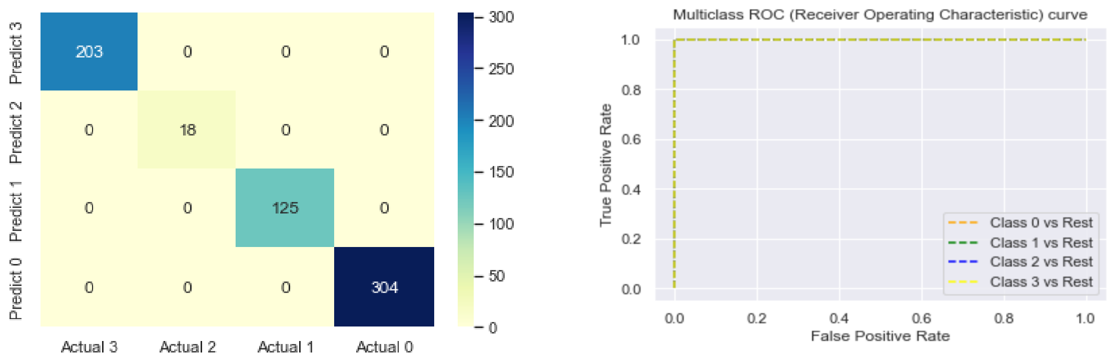


Fig. 7 Confusion Matrix and ROC Curve in Random Forest Classification

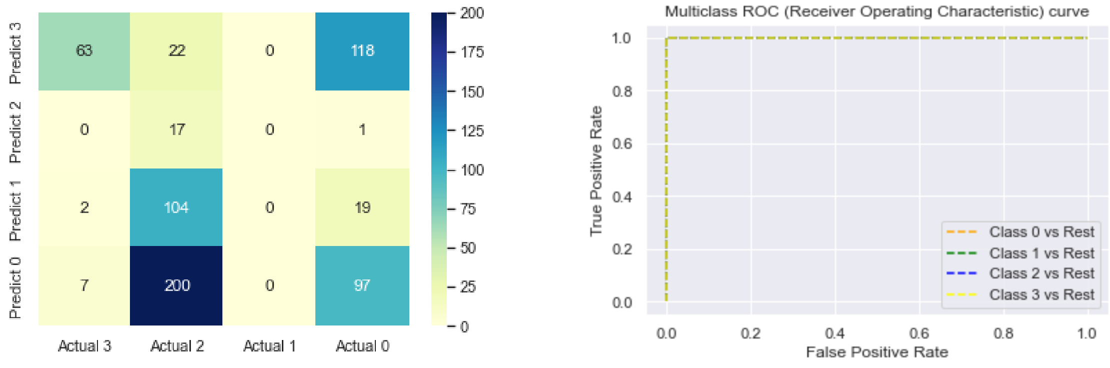


Fig. 8 Confusion Matrix and ROC Curve in Decision Tree Classification

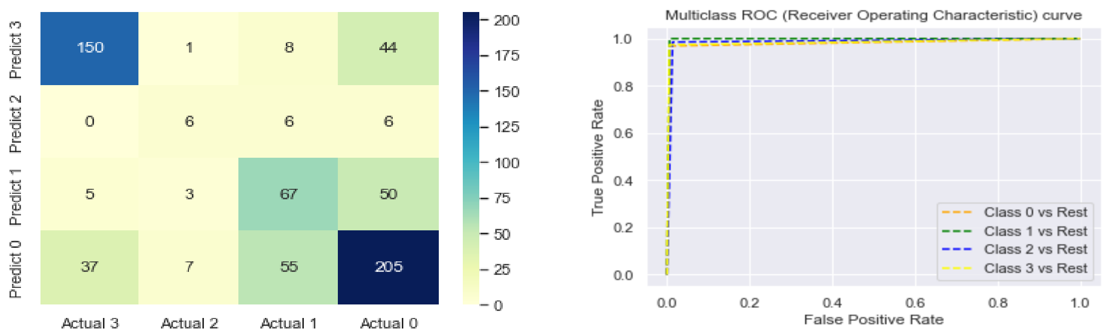


Fig. 9 Confusion Matrix and ROC Curve in Ensembling (Voting Classification)

AMSTERDAM (NETHERLAND)

The correlation of Price with the Final Features considered for its Prediction and also after doing feature engineering with the number_of_reviews and room_type column because they are not giving proper visualization and correlation with the price because of the variation in their values.

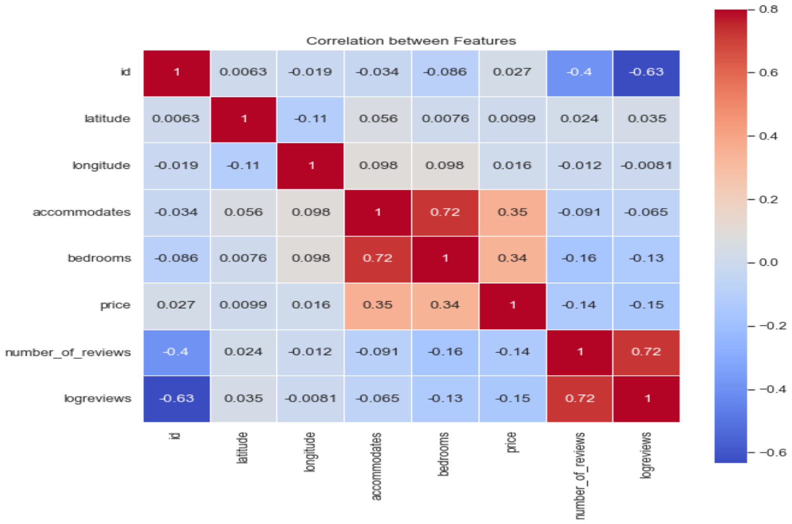


Fig. 10 Correlation between Features

	id	accommodates	bedrooms	price	price_range_category	number_of_reviews	logreviews	room_type_Entire home/apt	room_type_Hotel room	room_type_Private room	room_type_Shared room
id	1.000000	-0.033858	-0.086076	0.027316	0.012117	-0.396302	-0.633907	-0.080770	0.036610	0.069570	0.019722
accommodates	-0.033858	1.000000	0.724159	0.352095	0.177402	-0.091468	-0.064605	0.267582	-0.069316	-0.251836	-0.008614
bedrooms	-0.086076	0.724159	1.000000	0.344000	0.178828	-0.164017	-0.132286	0.334627	-0.070536	-0.316915	-0.029752
price	0.027316	0.352095	0.344000	1.000000	0.141866	-0.136287	-0.147404	0.200361	-0.027251	-0.193440	-0.020899
price_range_category	0.012117	0.177402	0.178828	0.141866	1.000000	-0.221056	-0.171781	0.404195	-0.027513	-0.395614	-0.058601
number_of_reviews	-0.396302	-0.091468	-0.164017	-0.136287	-0.221056	1.000000	0.722738	-0.354139	0.021957	0.349760	0.031272
logreviews	-0.633907	-0.064605	-0.132286	-0.147404	-0.171781	0.722738	1.000000	-0.290475	0.022917	0.287963	0.006678
room_type_Entire home/apt	-0.080770	0.267582	0.334627	0.200361	0.404195	-0.354139	-0.290475	1.000000	-0.180683	-0.955961	-0.083416
room_type_Hotel room	0.036610	-0.069316	-0.070536	-0.027251	-0.027513	0.021957	0.022917	-0.180683	1.000000	-0.087616	-0.007645
room_type_Private room	0.069570	-0.251836	-0.316915	-0.193440	-0.395614	0.349760	0.287963	-0.955961	-0.087616	1.000000	-0.040450
room_type_Shared room	0.019722	-0.008614	-0.029752	-0.020899	-0.058601	0.031272	0.006678	-0.083416	-0.007645	-0.040450	1.000000

Fig.11 Correlation between Features after Feature Engineering

Now the various algorithms are applied for which some important parameters are calculated included in tables below.

Table 4: Price Predicted values compared with Actual values calculated using different algorithms.

Here, A – Actual and P – Predicted.

Sr. No.	Multi-Linear Regression		K-Nearest Neighbor Class.		Naive Bayes Class.		Random Forest Class.		Decision Tree Class.		Ensembling (Voting Class.)	
	A	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**
1.	64	138	0	0	0	0	0	0	0	2	0	0
2.	105	157	3	3	3	3	3	3	3	2	3	3
3.	101	171	3	3	3	3	3	3	3	2	3	3
4.	100	191	3	3	3	3	3	3	3	2	3	3
5.	160	231	3	3	3	3	3	3	3	2	3	3
6.	105	144	3	3	3	3	3	3	3	2	3	0
7.	55	164	0	0	0	0	0	0	0	2	0	3
8.	131	129	3	3	3	3	3	3	3	2	3	3
9.	85	96	0	0	0	0	0	0	0	2	0	0
10.	140	130	3	3	3	3	3	3	3	2	3	0

Table 5: Score and Error related parameters are compared generated from different algorithms for Price Predictions.

Sr. No.	Algorithm and Type	Function	Accuracy Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	ROC AUC Score
1.	Multi-Linear Regression	LinearRegression()	0.354	55.408	8428.421	91.806	-
2.	K-Nearest Neighbor Classification	StandardScaler()	0.996	0.003	0.003	0.060	0.991
3.	Naïve Bayes Classification	GaussianNB()	1.0	0.0	0.0	0.0	1.0
4.	Random Forest Classification	RandomForestClassifier()	1.0	0.0	0.0	0.0	1.0
5.	Decision Tree Classification	DecisionTreeClassifier()	0.146	1.148	1.782	1.335	1.0

6.	Ensembling (Voting Classification)	VotingClassifier()	0.609	0.821	2.083	1.443	0.982

Table 6: Classification Report generated from different algorithms used for Price Predictions.

Sr. No.	Algorithm and Type	Classification Report
1.	Multi-Linear Regression	-
2.	K-Nearest Neighbor Classification	precision recall f1-score support 0 1.00 1.00 1.00 266 1 1.00 0.83 0.91 12 2 0.99 0.99 0.99 142 3 1.00 1.00 1.00 661 accuracy 1.00 1081 macro avg 1.00 0.95 0.97 1081 weighted avg 1.00 1.00 1.00 1081
3.	Naive Bayes Classification	precision recall f1-score support 0 1.00 1.00 1.00 266 1 1.00 1.00 1.00 12 2 1.00 1.00 1.00 142 3 1.00 1.00 1.00 661 accuracy 1.00 1081 macro avg 1.00 1.00 1.00 1081 weighted avg 1.00 1.00 1.00 1081
4.	Random Forest Classification	precision recall f1-score support 0 1.00 1.00 1.00 266 1 1.00 1.00 1.00 12 2 1.00 1.00 1.00 142 3 1.00 1.00 1.00 661 accuracy 1.00 1081 macro avg 1.00 1.00 1.00 1081 weighted avg 1.00 1.00 1.00 1081

5.	Decision Tree Classification	precision recall f1-score support 0 0.00 0.00 0.00 266 1 0.03 0.17 0.06 12 2 0.13 0.87 0.22 142 3 0.55 0.05 0.09 661 accuracy 0.15 1081 macro avg 0.18 0.27 0.09 1081 weighted avg 0.35 0.15 0.08 1081
6.	Ensembling (Voting Classification)	precision recall f1-score support 0 0.55 0.52 0.54 266 1 0.12 0.08 0.10 12 2 0.32 0.32 0.32 142 3 0.70 0.72 0.71 661 accuracy 0.61 1081 macro avg 0.42 0.41 0.42 1081 weighted avg 0.61 0.61 0.61 1081

The Confusion Matrix and ROC Curve generated from different algorithms based on parameters mentioned above are displayed.

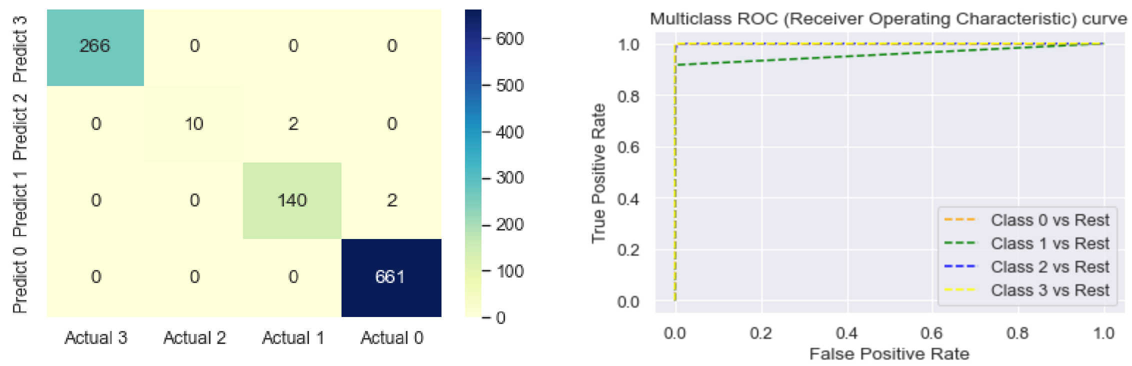


Fig. 12 Confusion Matrix and ROC Curve in K-Nearest Neighbor Classification

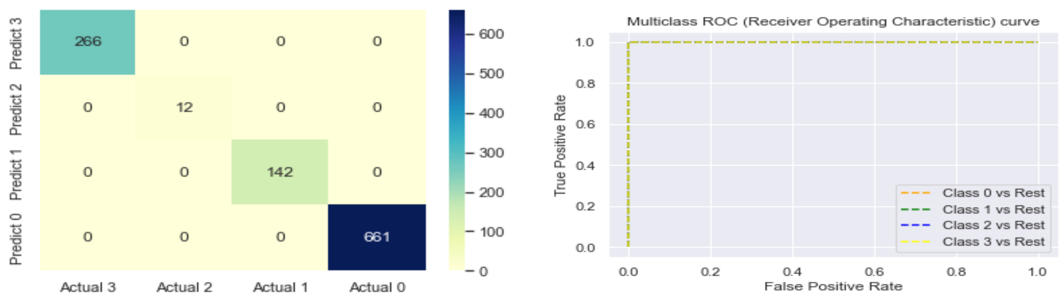


Fig. 13 Confusion Matrix and ROC Curve in Naive Bayes Classification

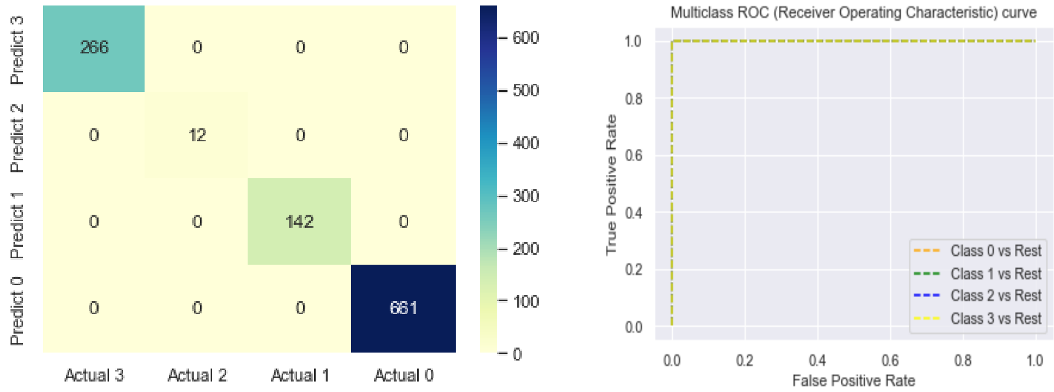


Fig. 14 Confusion Matrix and ROC Curve in Random Forest Classification

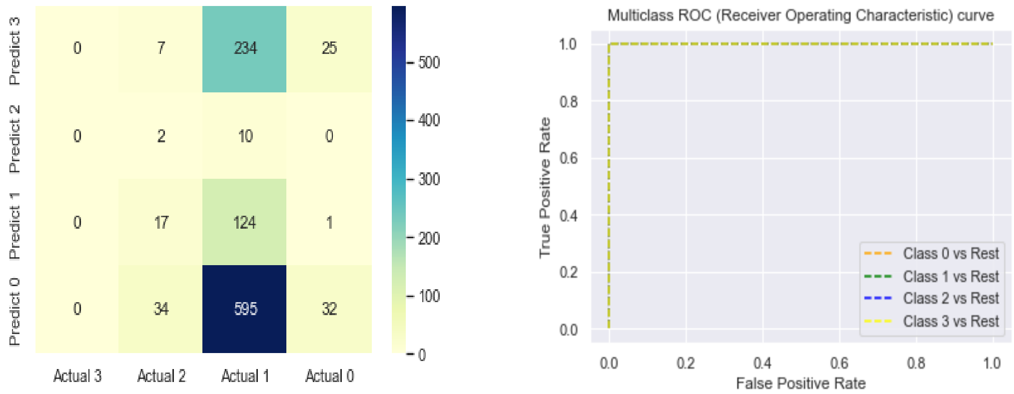


Fig. 15 Confusion Matrix and ROC Curve in Decision Tree Classification

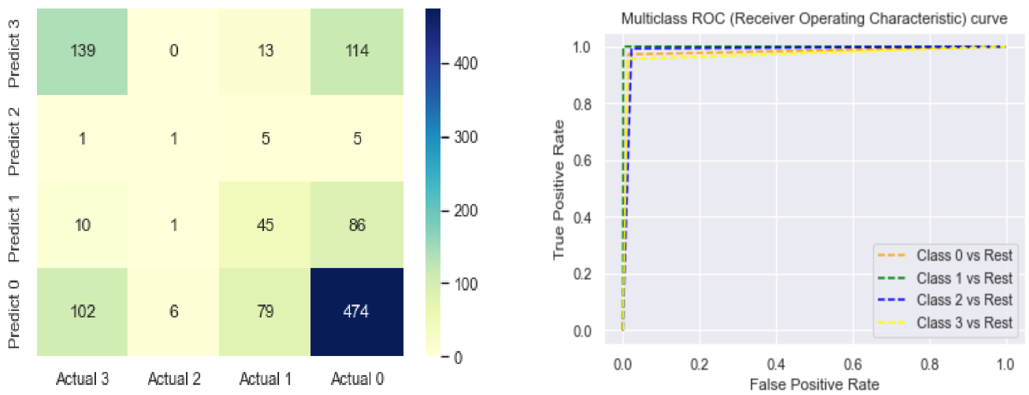


Fig. 16 Confusion Matrix and ROC Curve in Ensembling (Voting Classification)

HONGKONG (CHINA)

The correlation of Price with the Final Features considered for its Prediction and also after doing feature engineering with the number_of_reviews and room_type column because they are not giving proper visualization and correlation with the price because of the variation in their values.

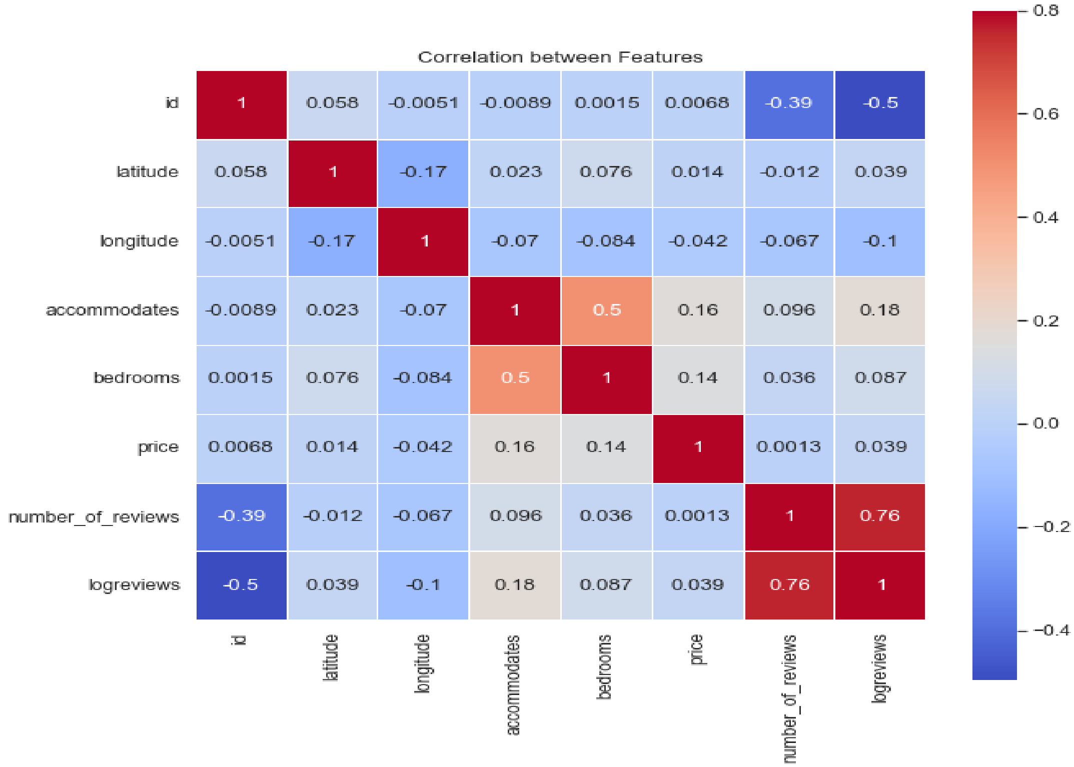


Fig. 17 Correlation between Features

	id	accommodates	bedrooms	price	price_range_category	number_of_reviews	logreviews	room_type_Entire home/apt	room_type_Hotel room	room_type_Private room	room_type_Shared room
id	1.000000	-0.008895	0.001484	0.006811	-0.010686	-0.388603	-0.497266	0.154619	-0.064882	-0.134530	-0.000204
accommodates	-0.008895	1.000000	0.501812	0.161799	-0.343598	0.095945	0.181396	0.225412	-0.014896	-0.248789	0.069142
bedrooms	0.001484	0.501812	1.000000	0.141533	-0.231241	0.036130	0.086596	0.068629	-0.010139	-0.059912	-0.009876
price	0.006811	0.161799	0.141533	1.000000	-0.290858	0.001310	0.039306	0.101726	-0.010857	-0.093056	-0.007798
price_range_category	-0.010686	-0.343598	-0.231241	-0.290858	1.000000	-0.122620	-0.248356	-0.494191	-0.043668	0.414101	0.172195
number_of_reviews	-0.388603	0.095945	0.036130	0.001310	-0.122620	1.000000	0.762111	0.061715	0.075265	-0.061607	-0.038484
logreviews	-0.497266	0.181396	0.086596	0.039306	-0.248356	0.762111	1.000000	0.108104	0.104930	-0.112894	-0.041991
room_type_Entire home/apt	0.154619	0.225412	0.068629	0.101726	-0.494191	0.061715	0.108104	1.000000	-0.103088	-0.859920	-0.196421
room_type_Hotel room	-0.064882	-0.014896	-0.010139	-0.010857	-0.043668	0.075265	0.104930	-0.103088	1.000000	-0.139399	-0.031841
room_type_Private room	-0.134530	-0.248789	-0.059912	-0.093056	0.414101	-0.061607	-0.112894	-0.859920	-0.139399	1.000000	-0.265608
room_type_Shared room	-0.000204	0.069142	-0.009876	-0.007798	0.172195	-0.038484	-0.041991	-0.196421	-0.031841	-0.265608	1.000000

Fig. 18 Correlation between Features after Feature Engineering

Now the various algorithms are applied for which some important parameters are calculated included in tables below.

Table 7: Price Predicted values compared with Actual values calculated using different algorithms.

Here, A – Actual and P – Predicted.

Sr. No.	Multi-Linear Regression		K-Nearest Neighbor Class.		Naive Bayes Class.		Random Forest Class.		Decision Tree Class.		Ensembling (Voting Class.)	
	A	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**
1.	1000	1541	1	1	1	1	1	1	1	2	1	1
2.	1650	1885	1	1	1	1	1	1	1	2	1	1
3.	200	939	3	3	3	3	3	3	3	2	3	1
4.	180	562	3	3	3	3	3	3	3	2	3	3
5.	464	1213	2	2	2	2	2	2	2	2	2	1
6.	4000	1528	1	1	1	1	1	1	1	2	1	1
7.	660	657	1	1	1	1	1	1	1	2	1	2
8.	290	618	2	2	2	2	2	2	2	1	2	2
9.	140	583	3	3	3	3	3	3	3	3	3	2
10.	180	555	3	3	3	3	3	3	3	2	3	3

Table 8: Score and Error related parameters are compared generated from different algorithms for Price Predictions.

Sr. No.	Algorithm and Type	Function	Accuracy Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	ROC AUC Score
1.	Multi-Linear Regression	LinearRegression()	0.031	650.95	81477	650.95	-
2.	K-Nearest Neighbor Classification	StandardScaler()	0.990	0.009	0.009	0.099	0.941
3.	Naïve Bayes Classification	GaussianNB()	1.0	0.0	0.0	0.0	1.0
4.	Random Forest Classification	RandomForestClassifier()	1.0	0.0	0.0	0.0	1.0

5.	Decision Tree Classification	DecisionTreeClassifier()	0.361	0.691	0.796	0.892	1.0
6.	Ensembling (Voting Classification)	VotingClassifier()	0.647	0.414	0.544	0.738	0.943

Table 9: Classification Report generated from different algorithms used for Price Predictions.

Sr. No.	**Algorithm and Type**	**Classification Report**
1.	Multi-Linear Regression	-
2.	K-Nearest Neighbor Classification	precision recall f1-score support 0 1.00 0.50 0.67 2 1 0.99 0.99 0.99 419 2 0.99 0.98 0.99 422 3 0.99 1.00 1.00 374 accuracy 0.99 1217 macro avg 0.99 0.87 0.91 1217 weighted avg 0.99 0.99 0.99 1217
3.	Naive Bayes Classification	precision recall f1-score support 0 1.00 1.00 1.00 2 1 1.00 1.00 1.00 419 2 1.00 1.00 1.00 422 3 1.00 1.00 1.00 374 accuracy 1.00 1217 macro avg 1.00 1.00 1.00 1217 weighted avg 1.00 1.00 1.00 1217
4.	Random Forest Classification	precision recall f1-score support 0 1.00 1.00 1.00 2 1 1.00 1.00 1.00 419 2 1.00 1.00 1.00 422 3 1.00 1.00 1.00 374 accuracy 1.00 1217 macro avg 1.00 1.00 1.00 1217 weighted avg 1.00 1.00 1.00 1217

<table>
<tr><td>5.</td><td>Decision Tree Classification</td><td><pre>
              precision   recall f1-score   support

           0   0.00   0.00   0.00   2
           1   0.40   0.39   0.40   419
           2   0.30   0.53   0.38   422
           3   0.71   0.15   0.24   374

    accuracy   0.36   1217
   macro avg   0.35   0.27   0.26   1217
weighted avg   0.46   0.36   0.34   1217
</pre></td></tr>
<tr><td>6.</td><td>Ensembling (Voting Classification)</td><td><pre>
              precision   recall f1-score   support

           0   0.00   0.00   0.00   2
           1   0.65   0.65   0.65   419
           2   0.57   0.62   0.59   422
           3   0.76   0.67   0.71   374

    accuracy   0.65   1217
   macro avg   0.50   0.49   0.49   1217
weighted avg   0.66   0.65   0.65   1217
</pre></td></tr>
</table>

The Confusion Matrix and ROC Curve generated from different algorithms based on parameters mentioned above are displayed.

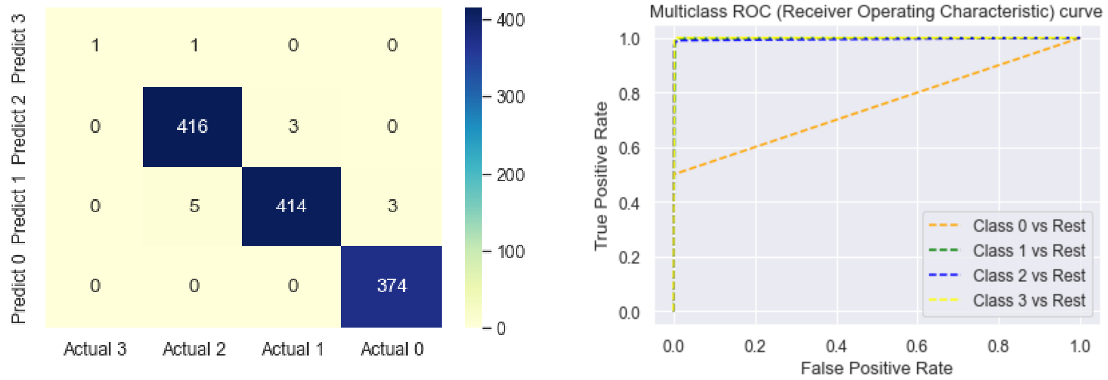


Fig. 19 Confusion Matrix and ROC Curve in K-Nearest Neighbor Classification

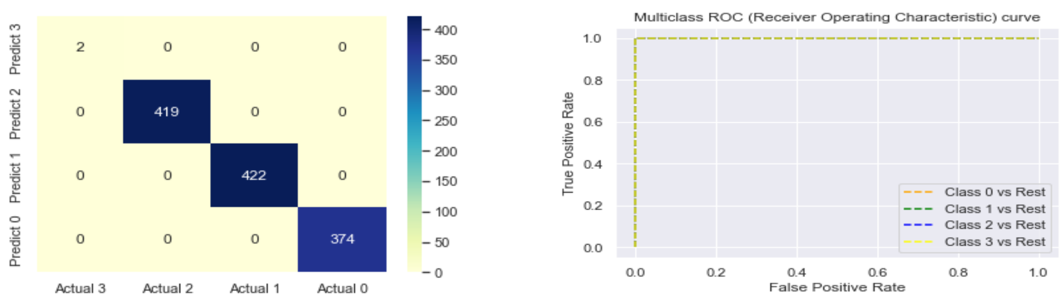


Fig. 20 Confusion Matrix and ROC Curve in Naive Bayes Classification

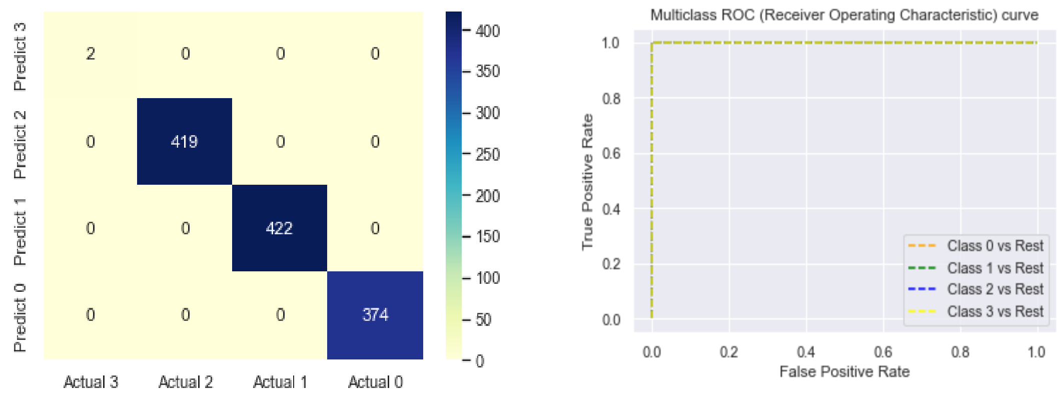


Fig. 21 Confusion Matrix and ROC Curve in Random Forest Classification

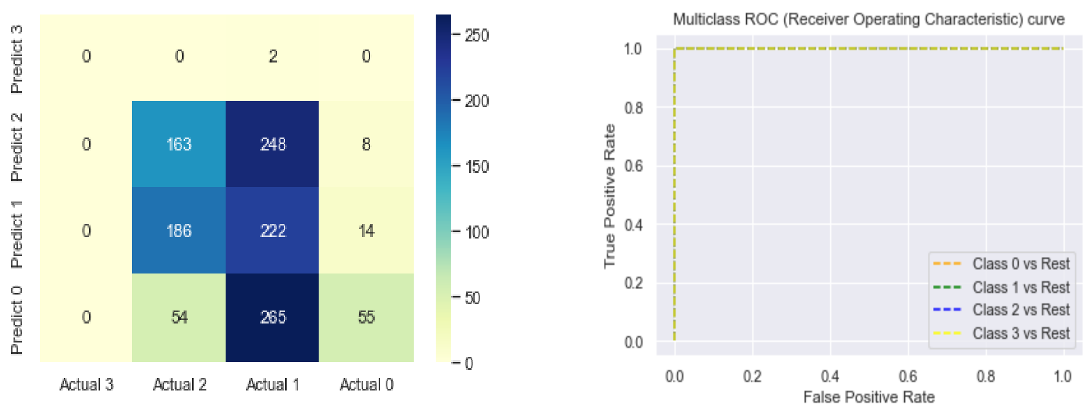


Fig. 22 Confusion Matrix and ROC Curve in Decision Tree Classification

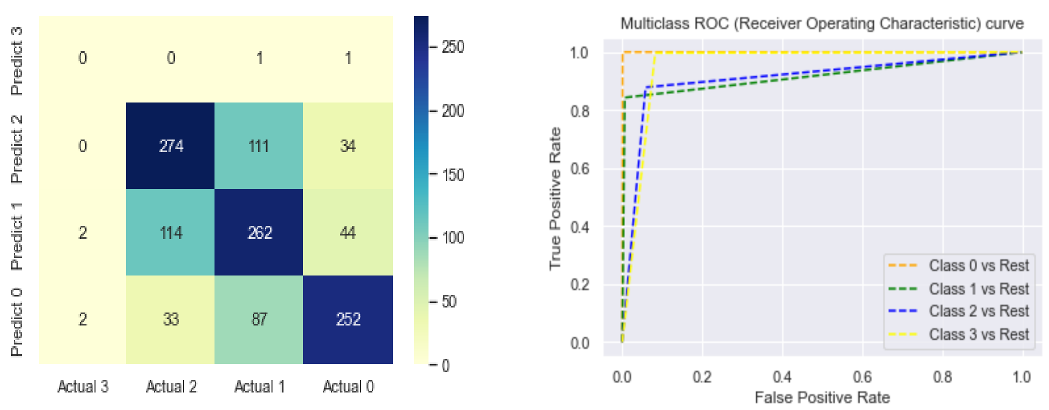


Fig. 23 Confusion Matrix and ROC Curve in Ensembling (Voting Classification)

ATHENS (GREECE)

The correlation of Price with the Final Features considered for its Prediction and also after doing feature engineering with the number_of_reviews and room_type column because they are not giving proper visualization and correlation with the price because of the variation in their values.

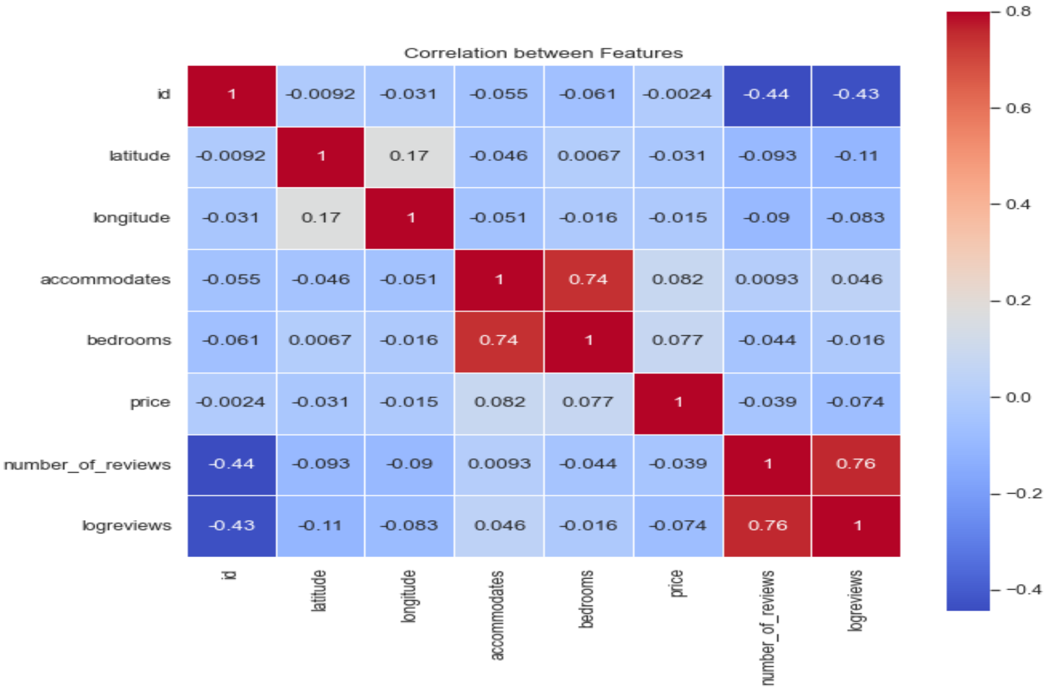


Fig. 24 Correlation between Features

	id	accommodates	bedrooms	price	price_range_category	number_of_reviews	logreviews	room_type_Entire home/apt	room_type_Hotel room	room_type_Private room	room_type_Shared room
id	1.000000	-0.054826	-0.061231	-0.002356	0.005613	-0.444224	-0.430305	-0.045396	0.002993	0.031605	0.052294
accommodates	-0.054826	1.000000	0.744313	0.081760	0.280865	0.009286	0.045710	0.236077	-0.037733	-0.220722	-0.074870
bedrooms	-0.061231	0.744313	1.000000	0.077357	0.274092	-0.043978	-0.016401	0.122875	-0.042096	-0.104920	-0.041229
price	-0.002356	0.081760	0.077357	1.000000	0.156219	-0.038618	-0.074442	-0.033151	0.041606	0.020305	-0.002147
price_range_category	0.005613	0.280865	0.274092	0.156219	1.000000	-0.092429	-0.162533	-0.093280	0.113225	0.063513	-0.020495
number_of_reviews	-0.444224	0.009286	-0.043978	-0.038618	-0.092429	1.000000	0.758148	0.107040	-0.059432	-0.078389	-0.046971
logreviews	-0.430305	0.045710	-0.016401	-0.074442	-0.162533	0.758148	1.000000	0.225318	-0.110102	-0.181385	-0.068098
room_type_Entire home/apt	-0.045396	0.236077	0.122875	-0.033151	-0.093280	0.107040	0.225318	1.000000	-0.332436	-0.880544	-0.267787
room_type_Hotel room	0.002993	-0.037733	-0.042096	0.041606	0.113225	-0.059432	-0.110102	-0.332436	1.000000	-0.038705	-0.011771
room_type_Private room	0.031605	-0.220722	-0.104920	0.020305	0.063513	-0.078389	-0.181385	-0.880544	-0.038705	1.000000	-0.031178
room_type_Shared room	0.052294	-0.074870	-0.041229	-0.002147	-0.020495	-0.046971	-0.068098	-0.267787	-0.011771	-0.031178	1.000000

Fig. 25 Correlation between Features after Feature Engineering

Now the various algorithms are applied for which some important parameters are calculated included in tables below.

Table 10: Price Predicted values compared with Actual values calculated using different algorithms.

Here, A – Actual and P – Predicted.

Sr. No.	Multi-Linear Regression		K-Nearest Neighbor Class.		Naive Bayes Class.		Random Forest Class.		Decision Tree Class.		Ensembling (Voting Class.)	
	A	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**
1.	38	78	0	0	0	0	0	0	0	1	0	0
2.	139	197	3	3	3	3	3	3	3	1	3	3
3.	35	71	0	0	0	0	0	0	0	2	0	0
4.	65	82	0	0	0	0	0	0	0	1	0	0
5.	48	41	0	0	0	0	0	0	0	1	0	0
6.	98	80	0	0	0	0	0	0	0	1	0	0
7.	28	28	0	0	0	0	0	0	0	1	0	0
8.	45	49	0	0	0	0	0	0	0	2	0	0
9.	35	32	0	0	0	0	0	0	0	1	0	0
10.	53	42	0	0	0	0	0	0	0	1	0	0

Table 11: Score and Error related parameters are compared generated from different algorithms for Price Predictions.

Sr. No.	Algorithm and Type	Function	Accuracy Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	ROC AUC Score
1.	Multi-Linear Regression	LinearRegression()	0.012	54.331	83362.507	288.725	-
2.	K-Nearest Neighbor Classification	StandardScaler()	0.995	0.004	0.004	0.068	0.984
3.	Naïve Bayes Classification	GaussianNB()	1.0	0.0	0.0	0.0	1.0
4.	Random Forest Classification	RandomForestClassifier()	1.0	0.0	0.0	0.0	1.0

5.	Decision Tree Classification	DecisionTreeClassifier()	0.009	1.200	1.646	1.283	1.0
6.	Ensembling (Voting Classification)	VotingClassifier()	0.853	0.401	1.145	1.070	0.950

Table 12: Classification Report generated from different algorithms used for Price Predictions.

Sr. No.	Algorithm and Type	Classification Report
1.	Multi-Linear Regression	-
2.	K-Nearest Neighbor Classification	precision recall f1-score support 0 1.00 1.00 1.00 1610 1 1.00 0.80 0.89 10 2 0.95 0.85 0.90 47 3 0.97 1.00 0.99 250 accuracy 1.00 1917 macro avg 0.98 0.91 0.94 1917 weighted avg 1.00 1.00 1.00 1917
3.	Naive Bayes Classification	precision recall f1-score support 0 1.00 1.00 1.00 1610 1 1.00 1.00 1.00 10 2 1.00 1.00 1.00 47 3 1.00 1.00 1.00 250 accuracy 1.00 1917 macro avg 1.00 1.00 1.00 1917 weighted avg 1.00 1.00 1.00 1917
4.	Random Forest Classification	precision recall f1-score support 0 1.00 1.00 1.00 1610 1 1.00 1.00 1.00 10 2 1.00 1.00 1.00 47 3 1.00 1.00 1.00 250 accuracy 1.00 1917 macro avg 1.00 1.00 1.00 1917 weighted avg 1.00 1.00 1.00 1917

<table>
<tr><td>5.</td><td>Decision Tree Classification</td><td><pre>
              precision   recall f1-score   support

           0   0.62   0.00   0.01   1610
           1   0.01   1.00   0.01   10
           2   0.00   0.00   0.00   47
           3   0.12   0.01   0.02   250

    accuracy   0.01   1917
   macro avg   0.19   0.25   0.01   1917
weighted avg   0.54   0.01   0.01   1917
</pre></td></tr>
<tr><td>6.</td><td>Ensembling (Voting Classification)</td><td><pre>
              precision   recall f1-score   support

           0   0.86   0.99   0.92   1610
           1   0.50   0.20   0.29   10
           2   0.44   0.09   0.14   47
           3   0.66   0.15   0.25   250

    accuracy   0.85   1917
   macro avg   0.62   0.36   0.40   1917
weighted avg   0.82   0.85   0.81   1917
</pre></td></tr>
</table>

The Confusion Matrix and ROC Curve generated from different algorithms based on parameters mentioned above are displayed.

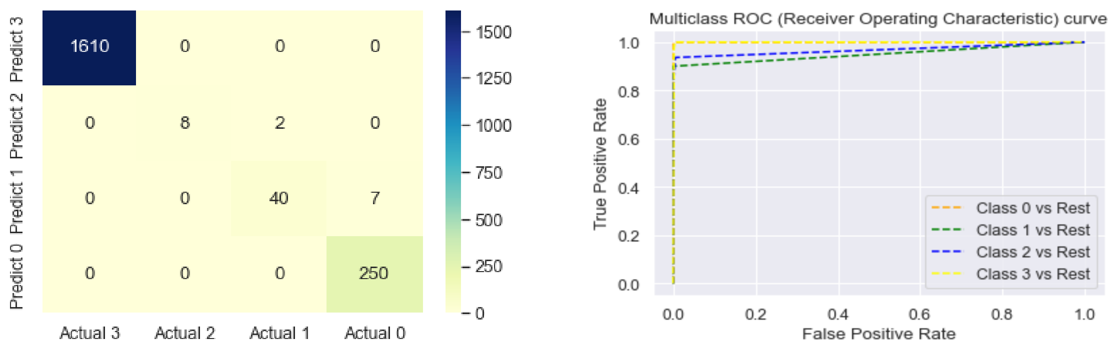


Fig. 26 Confusion Matrix and ROC Curve in K-Nearest Neighbor Classification

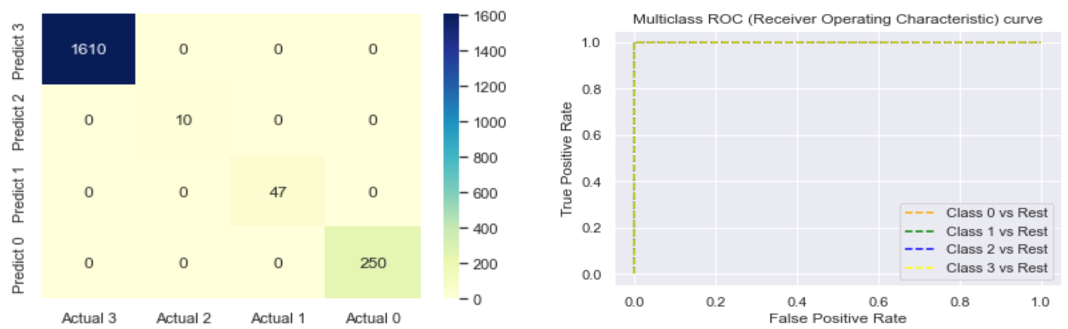


Fig. 27 Confusion Matrix and ROC Curve in Naive Bayes Classification

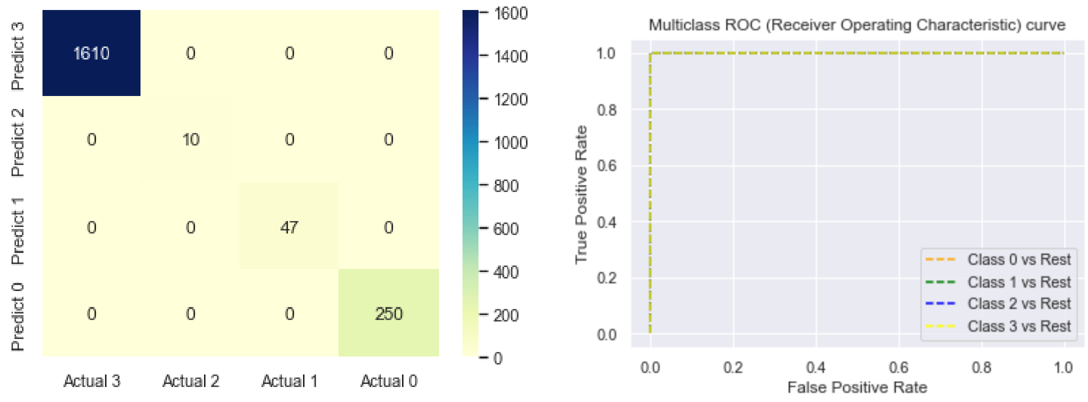


Fig. 28 Confusion Matrix and ROC Curve in Random Forest Classification

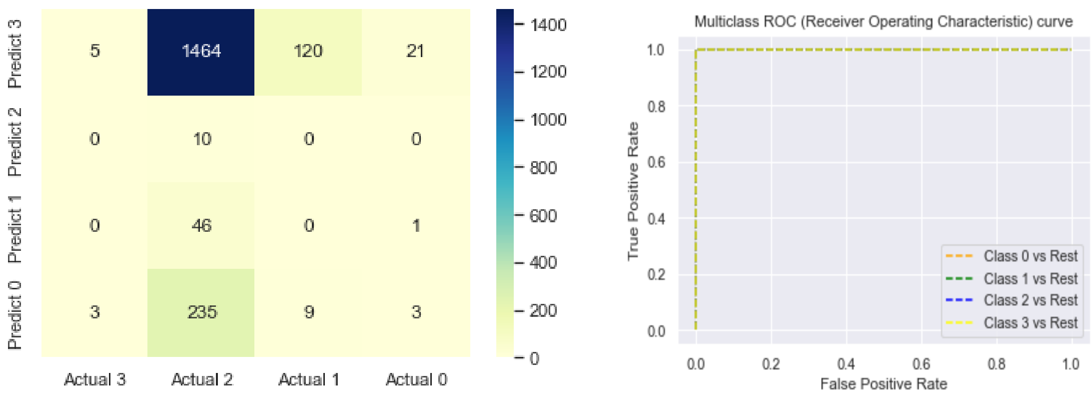


Fig. 29 Confusion Matrix and ROC Curve in Decision Tree Classification

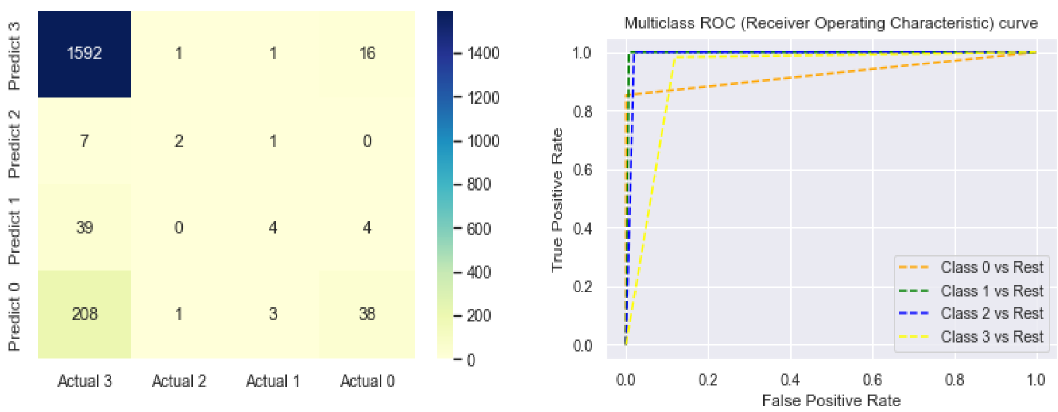


Fig. 30 Confusion Matrix and ROC Curve in Ensembling (Voting Classification)

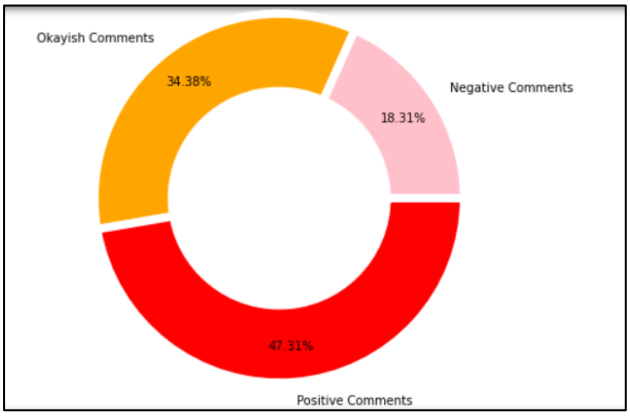
So, from the above validations in different cities of different continents it can be seen that Multi Linear Regression shows the huge difference in Predicted and Actual Price as it takes the continuous values so after converting into class labels i.e., economic, low-mid, high-mid and high then Naive Bayes and Random Forest Classification Predictions are very accurate followed by K-Nearest Neighbor Classification and Decision Tree Classification Algorithm.

2. To apply sentimental analysis on Airbnb dataset of different cities.

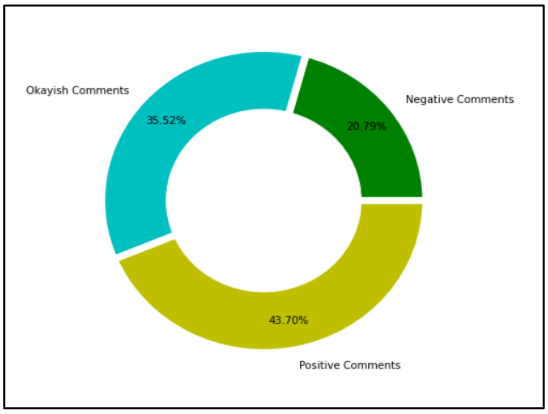
Table 13: To analyse the sentiments on the dataset i.e. positive, negative or neutral.

CITY	REVIEWS	POSITIVE	NEUTRAL	NEGATIVE
BOSTON	1,26,679	59,931(47.31%)	43,552 (34.38%)	23,194 (18.31%)
AMSTERDAM	2,66,861	11,668 (43.70%)	94,789 (35.52%)	55,480 (20.79%)
HONG KONG	1,06,538	65,542 (61.52%)	40,995 (38.48%)	0 %
ATHENS	4,06,607	1,78,947(44.01%)	1,42,556 (35.06%)	85,102 (20.93%)

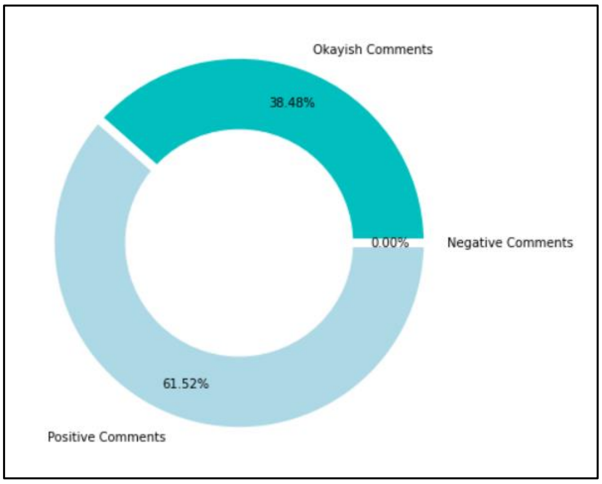
(a) Boston, USA



(b) Amsterdam, Netherlands



(c) Hong Kong, China



(d) Athens, Greece

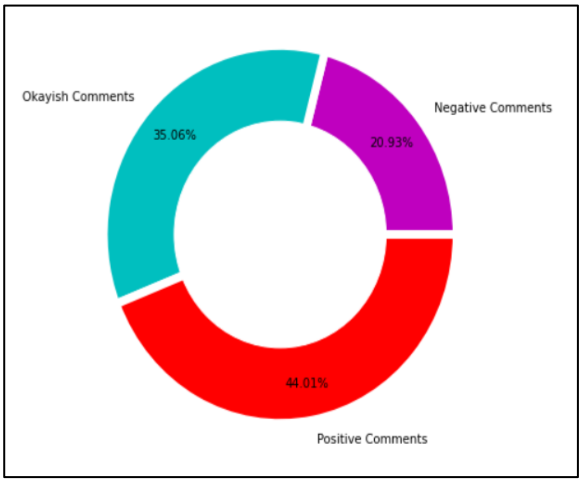


Fig.31 Donut Graph for all cities to analyse the sentiments on the dataset i.e. positive, negative or neutral

From this table we can conclude that Athens had the most number of reviews i.e. 4,06,607. But from the graph and the table we can clearly conclude that Hong Kong had the most number of positive reviews i.e. 61.52%. As it can be seen in the table, Hong Kong have almost 0% percentage of negative reviews compared to cities such as Amsterdam and Athens. The former city have a reputation of being friendly and welcoming.

We can say that the Airbnb reviews are almost similar across different cities. Most tourists leave positive reviews and use similar positive words to describe the Airbnb houses.

BOSTON (USA)

Fig.32 Word Cloud for positive comments in Boston

These graphs gave us an idea of the positive words frequently being used in the reviews in Boston. The keywords highlighted in the word cloud are clean, helpful, stay, great time. The graph shows the most frequent words being used in the positive reviews which are great, stay, place, boston and clean.

AMSTERDAM (NETHERLANDS)

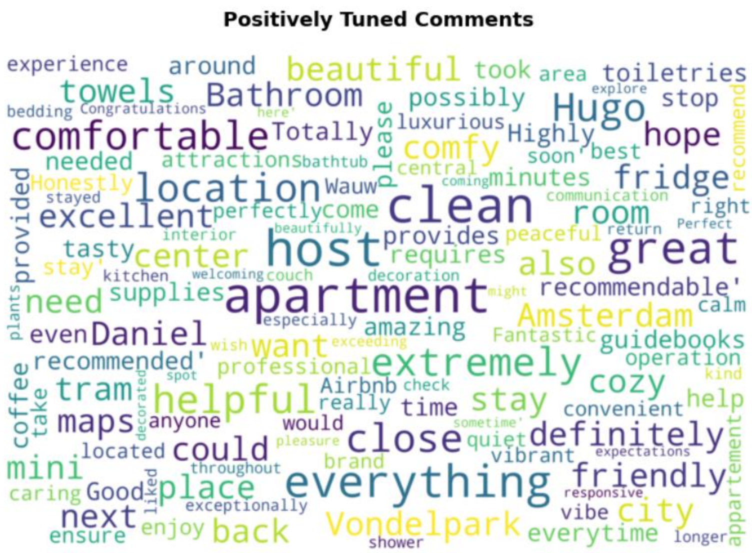


Fig. 33 Word Cloud for positive comments in Amsterdam

These graphs gave us an idea of the positive words frequently being used in the reviews in Amsterdam. The keywords highlighted in the word cloud are clean, helpful, host, comfortable, great . The graph shows the most frequent words being used in the positive reviews which are great, stay, place, Amsterdam, apartment.

HONG KONG (CHINA)

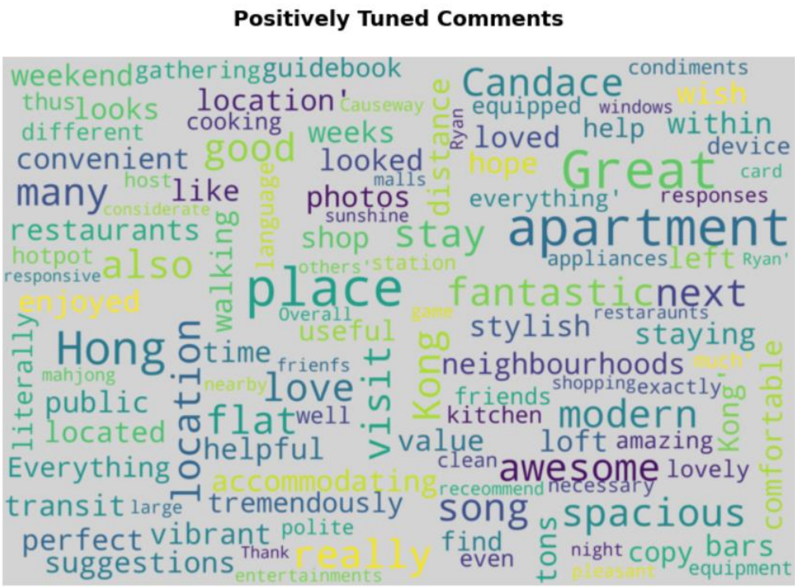


Fig. 34 Word Cloud for positive comments in Hong Kong

These graph gave us an idea of the positive words frequently being used in the reviews in Hong Kong. The keywords highlighted in the word cloud are great, apartment, awesome, spacious.

ATHENS (GREECE)

Fig. 35 Word Cloud for positive comments in Athens

These graph gave us an idea of the positive words frequently being used in the reviews in Athens . The keywords highlighted in the word cloud are recommend, place, Athen, apartment, definitely, perfect.

2.2 Top Hosts based on User Reviews and Top Hosts' neighbourhood.

BOSTON (USA)

```
# finding the names of top hosts' property

top_hosts.host_name
```

```
0    Justin
1      Nina
2     Huggy
3      Paul
4       Leo
Name: host_name, dtype: object
```

```
top_hosts.neighbourhood_cleansed
```

```
0     South End
1       Roxbury
2       Roxbury
3    Dorchester
4    Roslindale
Name: neighbourhood_cleansed, dtype: object
```

From the table obtained we can clearly identify that the Top Host in Boston with the best review sentimental score is Justin and his neighbourhood is South End.

Now we can look at the review people wrote about the Top Host with a compound sentimental score of 0.9978.

```
The room was much prettier and better equiped comparing to photos av
ailable on airbnb (was for sure refreshed, more furniture added, mak
ing this small place very practical). On-site staff was very friendl
y, helpful and kept common spaces very clean.Common kitchen and loun
dry space is very clean and well equiped, with good quality amenitie
s (including loundry powder, basic food supplies etc).Location is gr
eat.Place felt very safe and well taken care of, on-site staff makes
this place very "homely”. Justin answers messages very fast
```

AMSTERDAM (NETHERLAND)

```
# finding the names of top hosts' property

top_hosts.host_name
```

```
0     Sevi
1    Jelle
2      Bas
3     Asma
4    Marco
Name: host_name, dtype: object
```

```
list(top_hosts.neighbourhood_cleansed)
```

```
['Oostelijk Havengebied - Indische Buurt',
 'Westerpark',
 'Zuid',
 'Noord-West',
 'Centrum-West']
```

From the table obtained we can clearly identify that the Top Host in Amsterdam with the best review sentimental score is Sevi and the neighbourhood is Oostelijk Havengebied - Indische Buurt .

Now we can look at the review people wrote about the Top Host with a compound sentimental score of 0.9974.

```
If you are scrolling by now, just reserve this property and come bac
k to fully read this review. The location is wonderful and safe. Wit
hin a 4 block circle, you are surrounded by grocery stores, amazing r
estaurants, bars, cafes, barbers, salons, and anything else you coul
d need. I love the accessibility to everything and you don’t have to
worry about going miles to get to places. You won’t need a car becau
se of how close the trains are. The 14 tram is about 3 min away and t
akes you right into Amsterdam in like 20 minutes. You don’t need to c
hange trains. The other trains and Sprinter are about 7-10 min walk a
way and give you access to other places you need to go. It was so ea
sy The apartment was wonderful. The pictures don’t do it justice. Th
e decorations are simple but yet amazing. It had a very homey feel t
o it that made it feel like you lived there.
```

HONG KONG (CHINA)

```
# finding the names of top hosts' property

top_hosts.host_name
```

```
0      Maria
1      Brian
2        Mrs
3    Crystal
4        Jov
Name: host_name, dtype: object
```

```
# finding the neighbourhood of top hosts' property

top_hosts.neighbourhood_cleansed
```

```
0         Sha Tin
1    Yau Tsim Mong
2     Kowloon City
3       Yuen Long
4    Yau Tsim Mong
Name: neighbourhood_cleansed, dtype: object
```

From the table obtained we can clearly identify that the Top Host in Hong Kong with the best review sentimental score is Maria and the neighbourhood is Sha Tin.

Now we can look at the review people wrote about the Top Host with a compound sentimental score of 0.9976.

```
We are so thankful and blessed to have Maria as our host!! She is su
per kindness, humble and faithful to God! Our family plan to retreat
and rest and her place is perfect for us to stay away from city!!! T
he environment is like jungle forest with lots of mountain and we ca
n breathe fresh air!!! Here is very quiet, good for rest and retreat
!! Our family was so blessed and we also would like to thanks her ma
id Shirely even make a nice dinner for us and help us clean our clot
hes!!! In the morning, we heard bird singing and Maria’s friend wors
hip, love to join and worship God together!!!We would happy to keep i
n touch with her and support her missionary work too. Thanks so much
for your kindness hospitality
```

ATHENS (GREECE)

```
# finding the names of top hosts' property

top_hosts.host_name
```

```
0    Andreas
1     Argyro
2      Liana
3      Mania
4     Rosina
Name: host_name, dtype: object
```

```
top_hosts.neighbourhood_cleansed
```

```
0            ΑΝΩ ΠΑΤΗΣΙΑ
1               ΠΑΓΚΡΑΤΙ
2              ΠΕΤΡΑΛΩΝΑ
3    ΚΟΥΚΑΚΙ-ΜΑΚΡΥΓΙΑΝΝΗ
4              ΛΥΚΑΒΗΤΤΟΣ
Name: neighbourhood_cleansed, dtype: object
```

From the table obtained we can clearly identify that the Top Host in Athens with the best review sentimental score is Andreas and the neighbourhood is ΑΝΩ ΠΑΤΗΣΙΑ

Now we can look at the review people wrote about the Top Host with a compound sentimental score of 0.9956.

```
What a lovely apartment!  My husband, baby, toddler and I stayed the
re for 2 weeks and found it a perfect home away from home. The place
was beautiful: great location (right in front is a church with large
trees surrounding it, with a view of the entire city in front as it\
's on a slight hill, with a gorgeous balcony with a sofa and patio f
urniture to enjoy it on), beautiful furnishings, well-equipped kitch
en, and generally everything you could think of for a comfortable st
ay.  Although it\'s a 1 bedroom apartment, it\'s a huge bedroom (rea
lly 2, with a sliding glass door between if you like), and a comfy c
ouch in the living room, so fit us all.  The location is out of the d
owntown core of Athens, in a residential neighbourhood (so not touri
sty), with real neighbourhood shops, taverns, cafes, plazas, etc.  I
t\'s perfect in that it\'s only a 20 min direct bus downtown, but al
l the benefits of being in a "real" neighbourhood, with parks (great
for families), and very quite (as quiet as you can get for Athens, w
ithout getting into far away suburbs.  Andrea and Elena were great h
osts, having the place nicely prepared for us (we needed milk for wh
en we arrived late at night for the toddlers), and even took an emer
gency trip out at 4am as we were leaving for the airport and we real
ized we\'d forgotten our cell phone in the apartment as we closed th
e door on our way out.  Sorry!  Thanks again for such a lovely stay.
We will stay with you again on our next trip through Athens :-)'
```

3. To predict the spike in accommodation prices during peak and off-peak seasons of different cities.

BOSTON (USA)

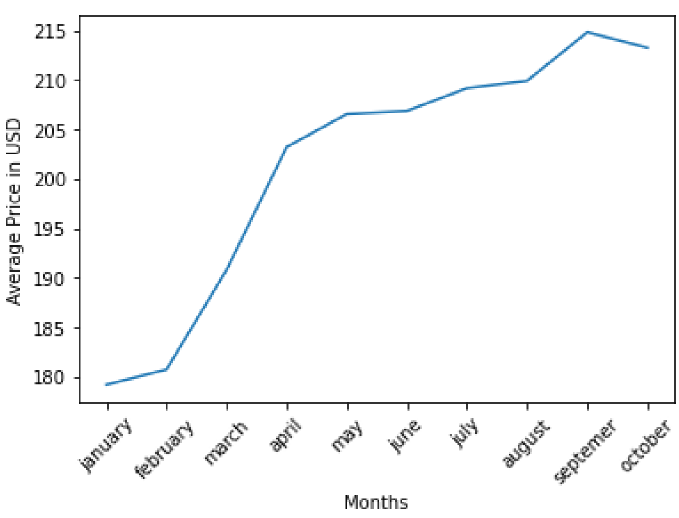
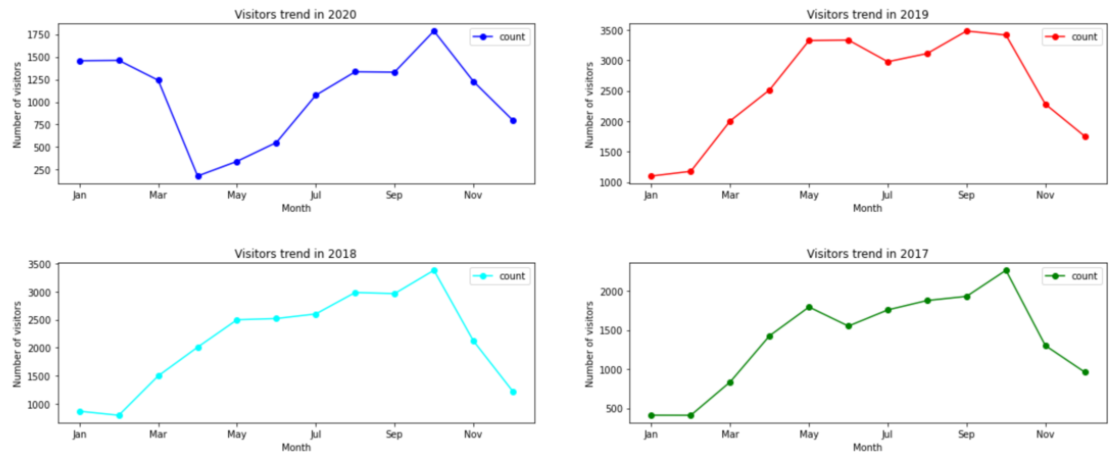


Fig. 36 Graph between Average Price in USD v/s Months

So, from the graph we can conclude that in peak season that is from April-October price range from 203 USD to 214USD. And in off season that is January to February price ranged from 179 to181.



we can see october is the most demanded season and also in most cases 2019-2017 januray,feb and dec is the off season excpet in 2020 april is the off season

Fig. 37 Number of visitors per months from year 2017-20

Above graph we can conclude that October is the month having most number of visitors from year 2017-2020 . January to February is off peak or people like to visited least in year 2017-2019 whereas as in 2020 April was month having least number of visitors .

Table 13: Maximum and Minimum price of specific property type in Boston.

Out[325]:

	property_type	price		
		min	max	mean
0	Boat	266	278	269.321429
1	Entire bed and breakfast	200	200	200.000000
2	Entire condominium (condo)	64	975	217.307785
3	Entire guest suite	49	294	127.885393
4	Entire guesthouse	99	200	147.094737
5	Entire loft	79	425	163.144295
6	Entire place	177	185	184.567568
7	Entire rental unit	49	5000	227.675236
8	Entire residential home	99	1052	316.236010
9	Entire serviced apartment	200	353	341.966102
10	Entire townhouse	157	5000	588.745238
11	Private room in bed and breakfast	55	285	192.396648
12	Private room in bungalow	44	66	58.510638
13	Private room in condominium (condo)	26	220	77.688073
14	Private room in guest suite	60	157	94.395833
15	Private room in guesthouse	50	50	50.000000
16	Private room in loft	33	116	88.560209
17	Private room in rental unit	26	159	80.263276
18	Private room in residential home	26	350	86.610047
19	Private room in townhouse	50	190	105.595745
20	Room in bed and breakfast	75	75	75.000000
21	Room in boutique hotel	249	285	261.727273
22	Room in hotel	255	255	255.000000
23	Shared room in rental unit	115	115	115.000000
24	Shared room in residential home	49	49	49.000000

We can conclude that Airbnb in Boston have 24 different types of property available for rent having average price range between 49 USD to 588 USD.

Table 14: Common amenities in Boston.

```
to_1D(airbnb["amenities"]).value_counts().head()
```

```
Smoke alarm                 3146
Wifi                        3116
Long term stays allowed     3075
Carbon monoxide alarm       2944
Kitchen                     2909
dtype: int64
```

So from the above graph we can conclude that above mentioned amenities is the most common amenity that is generally available in Airbnb hotels.

AMSTERDAM (NETHERLAND)

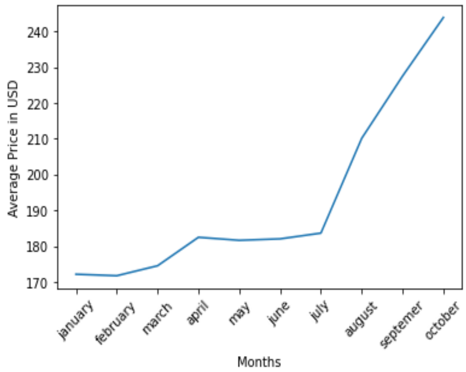


Fig. 38 Graph between Average Price in USD v/s Months

We can see august to October is peak season having price range from 210 USD to 245USD. And in off-season its 171 USD to 183 USD.

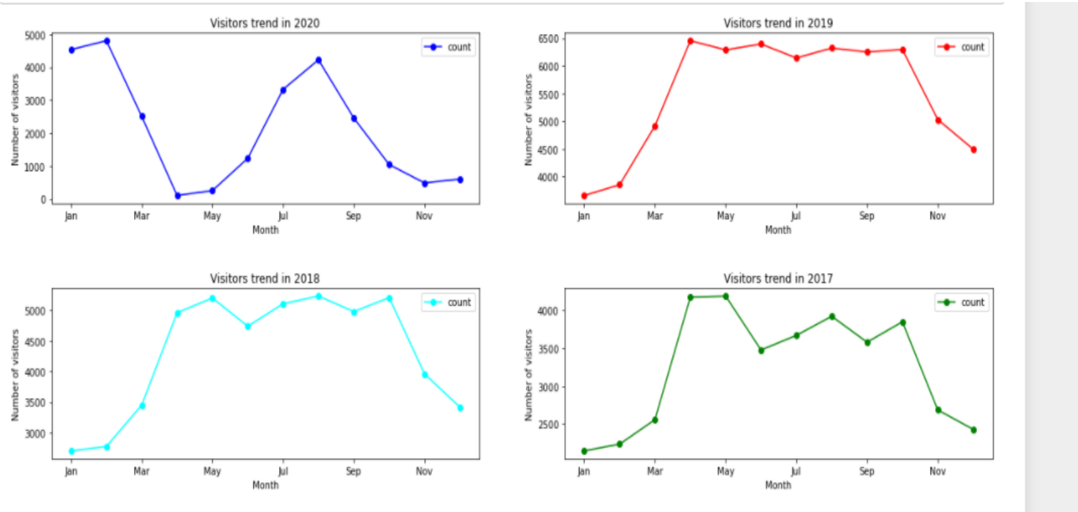


Fig. 39 Number of visitors per months from year 2017-20

Above graph we can conclude that October is the month having most number of visitors from year 2017-2019 and for 2020 its February . January to February is off peak or people like to visited least in year 2017-2019 whereas as in 2020 April was month having least number of visitors.

Table 15: Maximum and Minimum price of specific property type.

Out[119]:

	property_type	price		
		min	max	mean
0	Barn	85	85	85.000000
1	Boat	19	843	266.125000
2	Bus	50	50	50.000000
3	Entire cabin	88	200	174.153846
4	Entire chalet	90	125	117.125000
5	Entire condominium (condo)	51	600	171.411326
6	Entire cottage	181	324	216.750000
7	Entire guest suite	75	344	116.446777
8	Entire guesthouse	80	302	129.254717
9	Entire loft	70	1160	228.861111
10	Entire place	98	299	171.384615
11	Entire rental unit	45	857	174.502241
12	Entire residential home	34	850	221.543091
13	Entire serviced apartment	160	729	252.417031
14	Entire townhouse	85	810	240.834795
15	Entire villa	165	448	320.727273
16	Houseboat	100	1190	209.524252
17	Private room	95	115	95.814815
18	Private room in bed and breakfast	26	500	113.024203
19	Private room in boat	73	797	122.933602
20	Private room in bungalow	95	95	95.000000
21	Private room in cabin	96	117	102.681818
22	Private room in condominium (condo)	27	224	89.308054
23	Private room in farm stay	82	103	87.260870
24	Private room in guest suite	52	399	109.085944
25	Private room in guesthouse	51	533	77.203008
26	Private room in hostel	125	198	178.186782
27	Private room in houseboat	50	327	104.174441
28	Private room in island	75	75	75.000000
29	Private room in loft	55	200	113.904847
30	Private room in rental unit	9	600	90.028188
31	Private room in residential home	26	231	79.006810
32	Private room in serviced apartment	180	328	203.172414
33	Private room in tiny house	143	143	143.000000
34	Private room in townhouse	30	205	88.816395
35	Private room in villa	65	175	85.120000
36	Room in aparthotel	269	499	378.550000
37	Room in bed and breakfast	79	289	128.353268
38	Room in boutique hotel	53	325	106.689922
39	Room in hostel	25	167	47.439394
40	Room in hotel	85	123	106.660000
41	Room in serviced apartment	152	900	165.664093
42	Shared room in bed and breakfast	145	145	145.000000
43	Shared room in hostel	32	40	36.997033
44	Shared room in houseboat	205	336	277.050000
45	Shared room in rental unit	45	100	78.557692
46	Shared room in residential home	50	60	57.500000
47	Tower	326	326	326.000000

.

Amsterdam has 47 different types of properties available for lease. And having mean rent price range between 47 USD to 378 USD.

Table 16: Common amenities in Amsterdam.

```
In [12]: to_1D(airbnb["amenities"]).value_counts().head()

Out[12]: Wifi           5270
         Essentials     5192
         Heating        4812
         Smoke alarm    4746
         Hangers        4343
         dtype: int64
```

So, from the above graph we can conclude that above mentioned amenities is the most common amenity that is generally available in Airbnb hotels.

HONG KONG (CHINA)

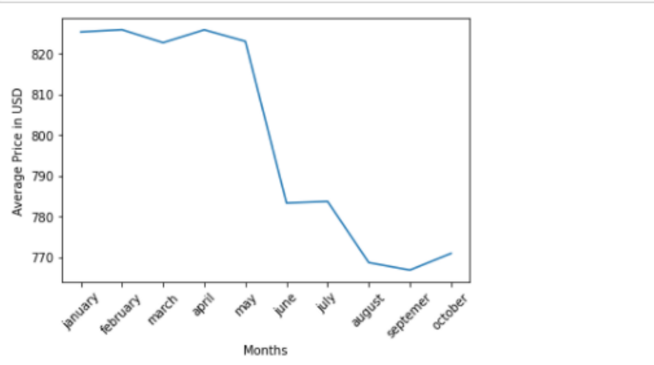


Fig. 40 Graph between Average Price in USD v/s Months

It is evident that January to May is the peak season with having price range from 784 USD to 825 USD and August to October have comparatively low average price which is 766USD to 771.

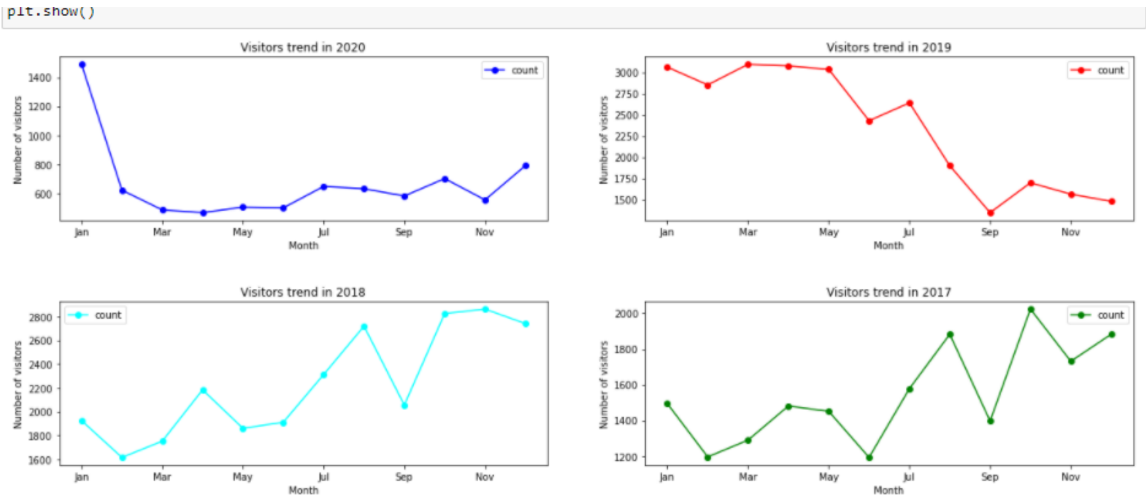


Fig. 41 Number of visitors per months from year 2017-20

Above graph we can conclude that October is the month having most number of visitors in year 2017 and 2018 . and in 2019 its March to May whereas in 2020 JANUARY is the peak season

February to March is off peak or people like to visited least in year 2017 and 2018 whereas as in 2019 its September and in 2020 its from March to June.

Table 17: Maximum and Minimum price of specific property type in Hong Kong.

	property_type	price		
		min	max	mean
0	Boat	266	278	269.076923
1	Entire bed and breakfast	200	200	200.000000
2	Entire condominium (condo)	53	975	213.560606
3	Entire guest suite	49	294	124.730201
4	Entire guesthouse	99	404	164.655172
5	Entire loft	79	425	216.951648
6	Entire place	185	185	185.000000
7	Entire rental unit	49	5000	221.514521
8	Entire residential home	80	2589	326.296474
9	Entire serviced apartment	139	469	289.060811
10	Entire townhouse	157	1014	387.789593
11	Houseboat	212	212	212.000000
12	Private room in bed and breakfast	50	285	135.584270
13	Private room in bungalow	44	66	57.918367
14	Private room in condominium (condo)	26	220	85.527668
15	Private room in guest suite	60	157	97.167442
16	Private room in guesthouse	50	50	50.000000
17	Private room in loft	33	116	97.721805
18	Private room in rental unit	25	200	82.216338
19	Private room in residential home	26	1000	87.208152
20	Private room in townhouse	39	190	104.958840
21	Room in bed and breakfast	75	75	75.000000
22	Room in boutique hotel	119	10000	581.979021
23	Room in hotel	0	431	168.083333
24	Shared room in rental unit	115	115	115.000000
25	Shared room in residential home	49	110	61.200000
26	Shared room in townhouse	27	27	27.000000

Hong-Kong has 26 different types of properties available for lease. And having rent price range between 27 USD to 582 USD.

Table 18: Common amenities in Hong-Kong.

```
: to_1D(airbnb["amenities"]).value_counts().head()
```

```
: Long term stays allowed     5855
  Air conditioning            5855
  Wifi                        5759
  Essentials                  4220
  Hangers                     4050
  dtype: int64
```

So from the above graph we can conclude that above mentioned amenities is the most common amenity that is generally available in Airbnb hotels of the city.

ATHENS (GREECE)

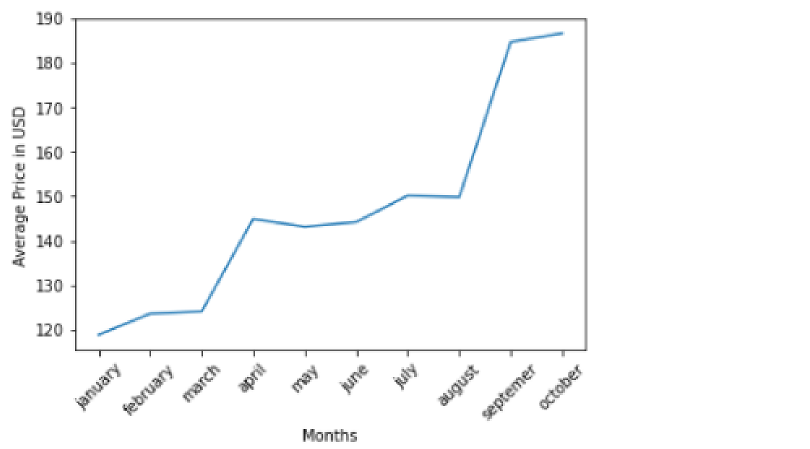


Fig. 42 Graph between Average Price in USD v/s Months

It is evident that September to October is the peak season with having price range from 185 USD to 187 USD and January and February have comparatively low average price which range between 119 USD to 124 USD.

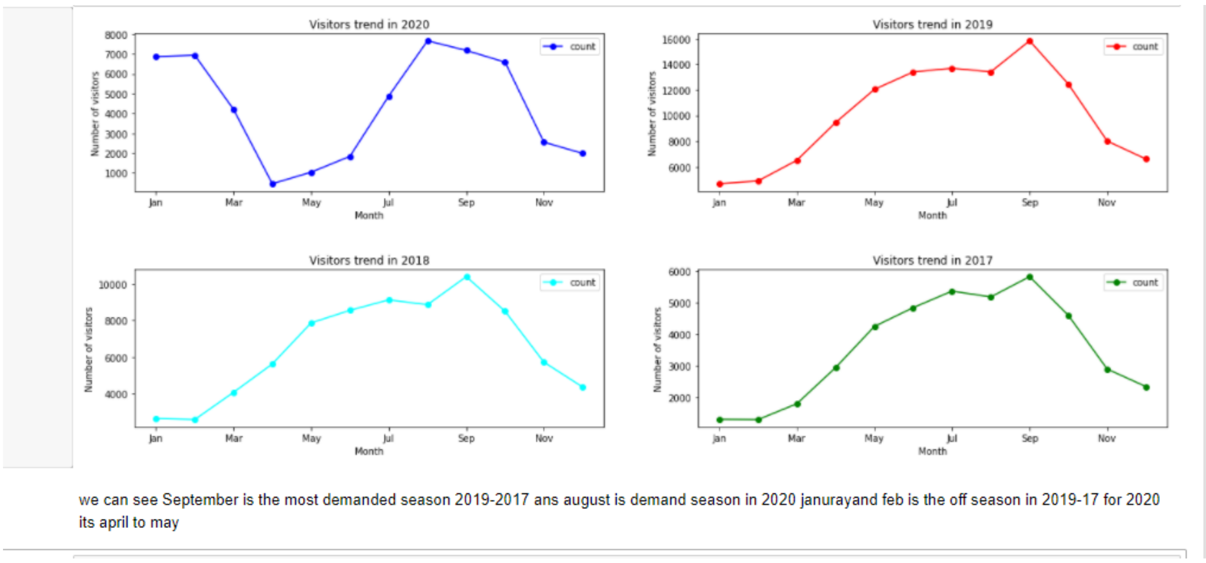


Fig. 43 Number of visitors per months from year 2017-20

From above graphs we can conclude that September is the month having most number of visitors in years 2017-2019. whereas in 2020 August is the peak season.

January and February is off peak or people visited least in year 2017-19 whereas as 2020 its from April to May.

Table 19: Maximum and Minimum price of specific property type in Athens.

	property_type	price		
		min	max	mean
0	Camper/RV	30	30	30.000000
1	Earth house	35	60	35.265957
2	Entire condominium (condo)	15	413	50.935763
3	Entire guest suite	24	116	44.882653
4	Entire guesthouse	21	48	32.341772
5	Entire loft	15	322	71.685322
6	Entire place	80	347	291.375000
7	Entire rental unit	10	8000	63.734452
8	Entire residential home	16	1500	97.036824
9	Entire serviced apartment	33	287	155.535270
10	Entire townhouse	32	145	59.488189
11	Entire villa	136	518	312.604167
12	Floor	267	267	267.000000
13	Private room in bed and breakfast	17	70	48.000000
14	Private room in condominium (condo)	13	40	26.444828
15	Private room in floor	20	20	20.000000
16	Private room in guest suite	43	45	44.255814
17	Private room in guesthouse	40	46	40.285714
18	Private room in hostel	15	15	15.000000
19	Private room in loft	30	30	30.000000
20	Private room in rental unit	12	5000	45.169910
21	Private room in residential home	16	72	26.933566
22	Private room in resort	45	45	45.000000
23	Private room in serviced apartment	13	90	16.558824
24	Room in aparthotel	39	990	81.525692
25	Room in boutique hotel	100	365	310.630769
26	Room in hotel	35	150	69.823529
27	Room in serviced apartment	30	8000	105.896657
28	Shared room in hostel	12	25	18.111111
29	Shared room in nature lodge	12	12	12.000000
30	Shared room in rental unit	11	18	15.882353
31	Shared room in residential home	10	500	479.148936
32	Tiny house	25	40	35.312903

Athens has 32 different types of properties available for lease. And having mean rent price range between 12 USD to 479 USD.

Table 20: Common amenities in Athens.

```
to_1D(airbnb["amenities"]).value_counts().head()
```

```
Essentials              9078
Hair dryer              8774
Wifi                    8756
Long term stays allowed 8696
Air conditioning        8670
dtype: int64
```

So, from the above graph we can conclude that above mentioned amenities are the most common amenity that is generally available in Airbnb hotels present in this city.

DISCUSSION

Table 21: To visualize that where to invest in a property in different cities of different continents to get the maximum number of returns from Airbnb.

Here, L – Least Expensive, M – Most Expensive.

Sr. No.	**Cities**	**Neighbourhood**		**Room Type**		**Property type**	
		L	**M**	**L**	**M**	**L**	**M**
1.	**Boston**	Hyde Park ($91.540)	Back Bay ($324.589)	Private Room ($97.639)	Hotel Room ($438.560)	Room in Hostel ($0.0)	Entire townhouse ($548.588)
2.	**Amsterdam**	Gaasperdam – Dreimond ($99.428)	Centrum - Oost ($212.909)	Shared Room ($111.368)	Entire home/apt ($192.383)	Shared room in hostel ($39.750)	Shared room in boat ($500.0)
3.	**HongKong**	Wong Tai Sin ($500.833)	Tsuen Wan ($5142.650)	Private Room ($600.920)	Entire home/apt ($1080.944)	Tent ($140.333)	Entire villa ($10800.0)
4.	**Athens**	ΠΕΝΤΑΓΩΝΟ ($33.0)	ΑΓΙΟΣ ($560.095)	Shared Room ($76.833)	Hotel Room ($186.533)	Shared room in serviced apartment ($11.500)	Private room in bed and breakfast ($469.638)

The Neighbourhood, Room Type and Property Type compared with the Average Mean Price so that one can able to find the property in particular city in particular continent according to the facilities in demand to make the investment in it to get the maximum returns in the near future.

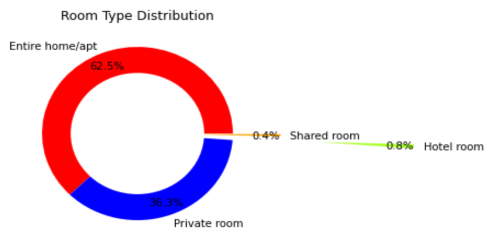
Also, from above mentioned we can see that investment will be beneficial in HongKong as per the average prices founded there.

Table 22: To visualize that which Room Type is most and least expensive and come under which Property Type and Neighbourhood in different cities of different continents.

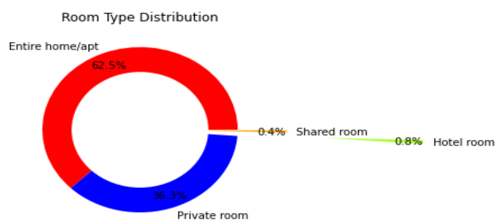
Here, L – Least Expensive, M – Most Expensive.

Sr. No.	**Cities**		**Entire home/apt**		**Hotel Room**		**Private Room**		**Shared Room**	
			L	**M**	**L**	**M**	**L**	**M**	**L**	**M**
1.	**Boston**	**Property Type**	Entire home/apt ($75)	Entire townhouse ($548)	Room in hostel ($0)	Room in hotel ($597)	Private room in guesthouse ($50.00)	Rook in boutique hotel ($563)	Shared room in bed and breakfast ($20)	Shared room in boutique hotel ($203)
		Neighbourhood	Longwood Medical Area ($106)	Mattapan ($331)	South End and Fenway ($0)	Downtown ($888)	Mattapan ($62)	Back Bay ($371)	Mission Hill ($20)	Fenway ($750)
2.	**Amsterdam**	**Property Type**	Bus ($50)	Tower ($421)	Room in hostel ($61)	Room in casa particular ($270)	Private room in island ($75)	Private room in serviced apartment ($361)	Shared room in hostel ($39)	Shared room in boat ($500)
		Neighbourhood	Gassperdam ($140)	Centrum – West ($242)	Bos en ($0)	Centrum – Oost ($237)	Bijlmer ($63)	Centrum – Oost ($188)	Oud – Oost ($39)	Oostel jik ($500)
3.	**HongKong**	**Property Type**	Tent	Entire villa	Room in	Room in boutiq	Private room in minus	Private room in	Shared room in hostel	Shared room

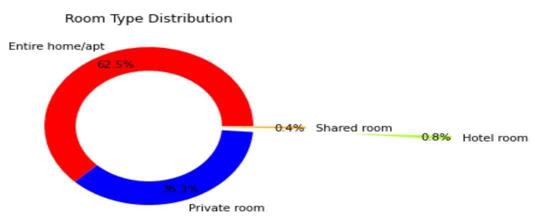
			($140)	($11049)	aparthotel ($500)	ue hotel ($1534)	($257)	townhouse ($9155)	($220)	in tiny house ($2850)
		Neighbourhood	Sha Tin ($673)	Sai Kung ($5557)	Yau Tism Mong ($440)	Central and Western ($2043)	Sham Shui Po ($254)	Tsuen Wan ($9228)	North ($85)	Sham Shui Po ($4241)
4.	**Athens**	**Property type**	Tiny House ($30)	Boat ($450)	Room in hotel ($207)	Room in serviced apartment ($278)	Private room in condominium ($26)	Private room in bed and breakfast ($469)	Shared room in condominium ($10)	Shared room in residential home ($427)
		Neighbourhood	-	-	-	-	-	-	-	-



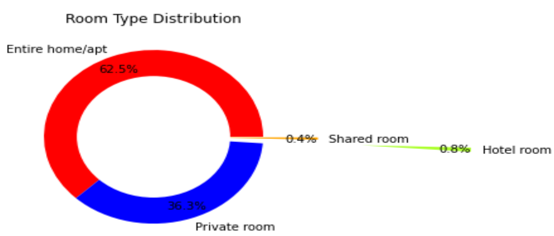
(a) Boston



(b) Amsterdam



(c) Hong Kong



(d) Athens

Fig.44 Room type Distributions

Table 23: To visualize that which listing id has good and bad Review Score Ratings on the basis of Neighbourhood, Property Type, Room Type and Bedrooms available in the individuals.

Here, Review Score Ratings – 5>4>3>2>1 i.e., good to bad ratings sequence.

Sr. No.	Cities	Neighbourhood	Property Type	Room Type	Bedrooms
1.	**Boston**	Back Bay (5)	Shared room in Condonium (5)	Entire home/apt (5)	13 (5)
2.	**Amsterdam**	Gaasperdam – Dreimond (5)	Tiny House (5)	Private Room (4.875)	10 (4.875)
3.	**HongKong**	Sham Sui Po (5)	Hut (5)	Entire home/apt (5)	6 (5)
4.	**Athens**	-	Boat (5)	Hotel Room (5)	4 (5)

From above mentioned ratings Airbnb can easily suggest the best provision to their users if they are looking or searching facilities in the terms of review score ratings.

2. To apply sentimental analysis on Airbnb dataset of different cities.

BOSTON (USA)

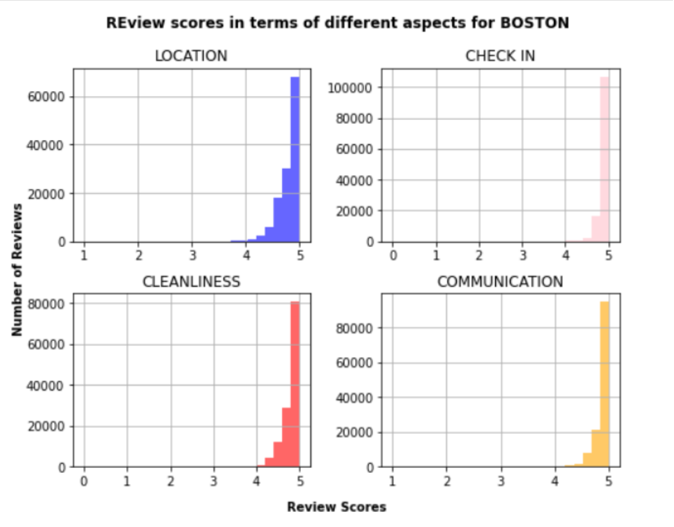


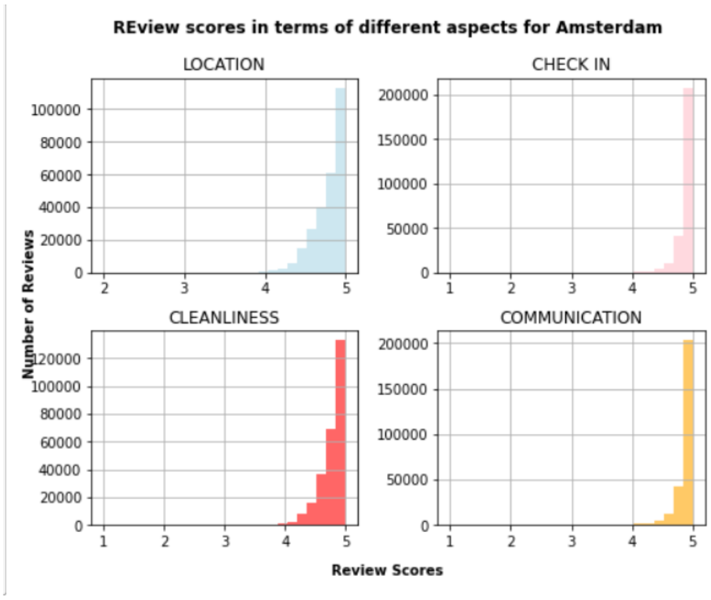
Fig. 45 Review Score in terms of different aspects for Boston

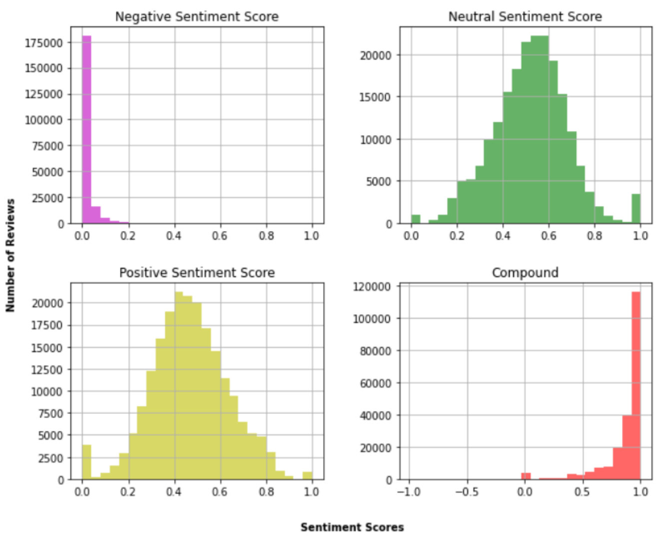
This graph gives relation between the review score given by the guest and number of reviews in terms of different aspects like Location, Check In, Cleanliness and Communication. From this graph we can conclude that most of the review scores for all the aspects were between 4-5 which is considered to be a high score.

Fig. 46 Sentiment Analysis of Airbnb reviews for Boston

This is a combined graph of sentimental analysis of Airbnb Reviews. The graph was plotted between the number of reviews and negative, positive, neutral and compound sentimental score. From the graph we can visualised that the value of negative sentimental score was very less, the highest value of neutral sentimental score lie between 0.6 – 0.8. The highest value of positive sentimental score lies between 0.2- 0.4, and the compound score for every review was almost between 0.5 to 1.0.

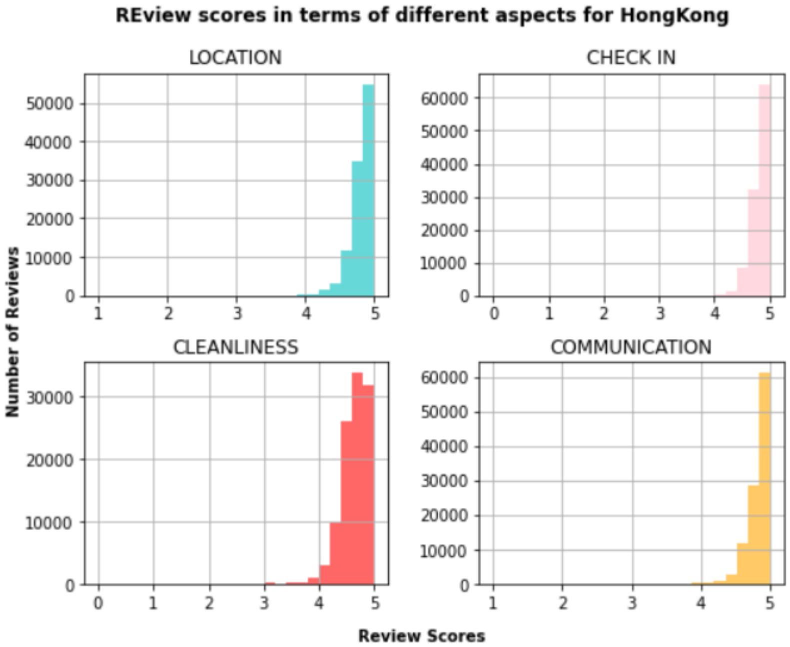
AMSTERDAM (NETHERLANDS)

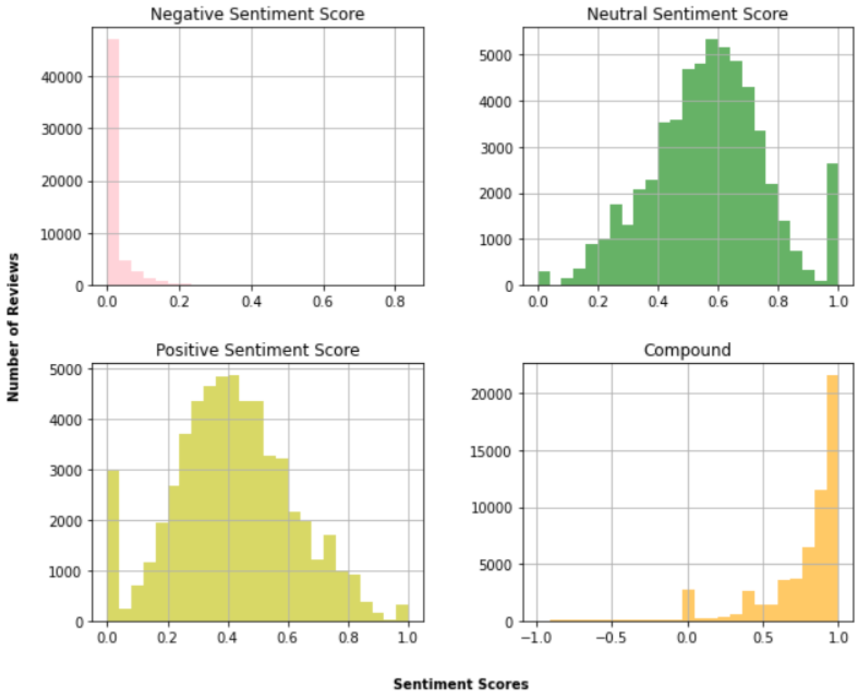




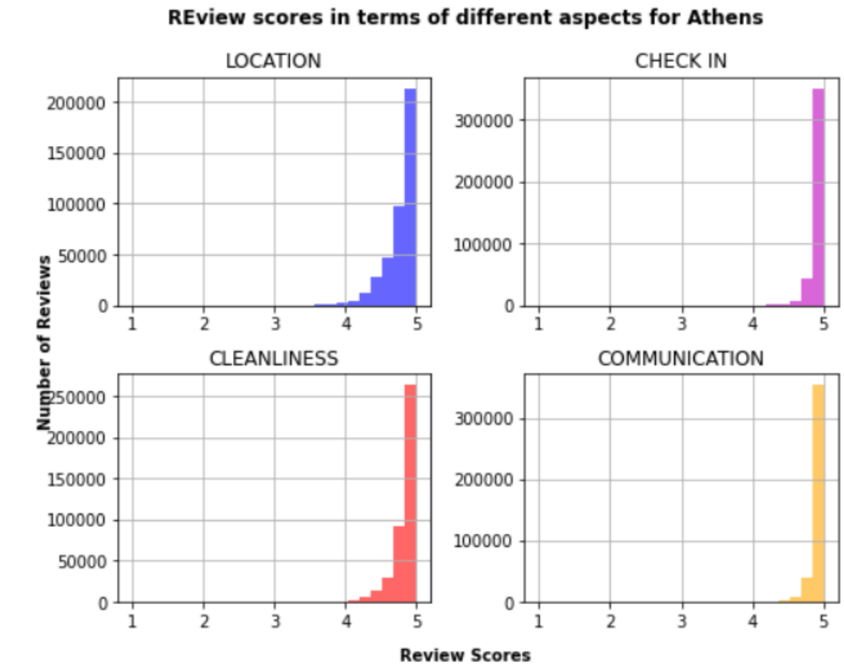
9

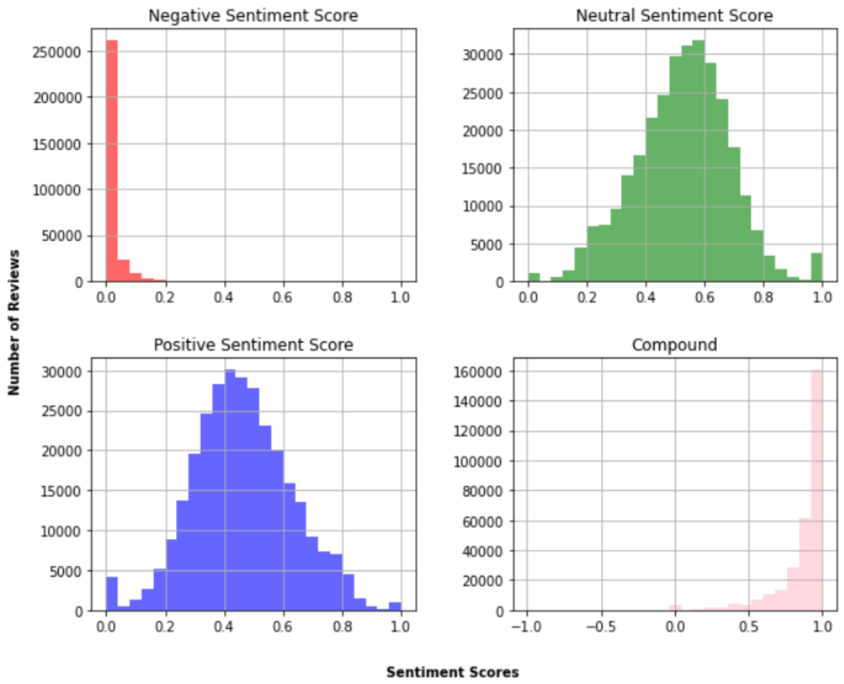
HONG KONG (CHINA)





ATHENS (GREECE)





After visualising all the graphs, we could say that the review score in terms of Location, Check In, Cleanliness and Communication is almost same for all the cities. The review rating is almost between the range of 4-5 which is a high score.

Also, after looking at the sentimental scores of all the cities we could conclude that most of the cities compound sentimental score lies between 0.5 to 1.0 which according to VADER if (compound score >= 0.05) the review is positive.

So, almost all the tourists have given a positive review for these Airbnb.

- The neighbourhood with the highest review score in **Boston, USA** is South End.
- The neighbourhood with the highest review score in **Amsterdam, Netherlands** is Oostelijk Havengebied - Indische Buurt.
- The neighbourhood with the highest review score in **Hong Kong, China** is Sha Tin.
- The neighbourhood with the highest review score in **Athens, Greece** is ΑΝΩ ΠΑΤΗΣΙΑ

3. To predict the spike in accommodation prices during peak and off-peak seasons of different cities.

(A) From the average price graphs of different cities, we found out that

1. **In Boston** the peak season is from April-October where price range from 203 USD to 214USD. And in off season that is January to February price range from 179 to181USD. So, if you want to spend less you can visit here in off season as we can see there is a good difference in price in peak and off season.

2. **In Amsterdam** the peak season is in August and October with price range from 210 USD to 245USD. And in off-season its 171 USD to 183 USD. So, if you want to spend less you can visit here in off season as we can see there is a good difference in price in peak and off season.

3. **In Hong-Kong** January to May is the peak season with having price range from 784 USD to 825 USD and August to October have comparatively low average price which is 766 USD to 771.So, if you want to spend less you can visit here in off season as we can see there is a good difference in price in peak and off season.

4. **In Athens** September to October is the peak season with having price range from 185 USD to 187 USD and January and February have comparatively low average price

which range between 119 USD to 124 USD. So, if you want to spend less you can visit here in off season as we can see there is a good difference in price in peak and off season.

From average price range we can observe that Hong-Kong is quite expensive and out of all Athens is least costly. And we can all observe that expect Hong-Kong other cities have peak season in October.

(B) From Visitors Trend Graph we can conclude that

1. **The city Boston** October is the busiest month, with the most visitors, however if you don't like crowds, go in the early months of the year.

2. **The city Amsterdam** from 2017 - 2019, the month with the most visitors was October, and in 2020, it was February. If you don't want to visit when it's crowded, go during the first few months of the year.

3. **In Hong-Kong** in both 2017 and 2018, the month of October had the highest number of visits. In 2019, the busiest time was from March to May, whereas in 2020, the peak season was from January to February. In the years 2017 and 2018, the off-season is February to March, but the off-season is September in 2019 and March to June in 2020. Hong-Kong visitors trend is quite unpredictable from past visitors' history.

4. **In Athens** the months of August and September are the busiest, with the highest number of visitors. In the years 2017-19, the months of January and February were off-peak, with the least number of visitors, however in 2020, the months of April and May was off. If you don't want to visit when it's crowded, go during the first few months of the year.

From results from average price and Visitor's trend we can validate that In cities expect Hong-Kong the month of October will be having highest price as its the peak season having most number of visitors.

(C) From Property Type and Price table we can see that:

1. **Boston** has 24 different types of property available for rent having price range between 49 USD to 266 USD.
2. **Amsterdam** has 47 different types of properties available for lease. And having mean rent price range between 47 USD to 378 USD.
3. **Hong-Kong** has 26 different types of properties available for lease. And having rent price range between 27 USD to 582 USD.
4. **Athens** has 32 different types of properties available for lease. And having mean rent price range between 12 USD to 479 USD.

(D) From top or common amenities table we can see that

In all 4 cities Wi-Fi , smoke alarm , long-term stay, heating and essentials are the top most common amenities available in Airbnb hotels .

AUTHOR’s CONTRIBUTION

Task	Prachika Kanodia	Ishika Gupta	Akshi Agarwal
Objective	✔	✔	✔
Literature Review	✔	✔	✔
Methodology	✔	✔	✔
Result preparation	✔	✔	✔
Result interpretation (Discussion)	✔	✔	✔
Report	✔	✔	✔

REFERENCES

[1] Guttentag, D. (2019). Progress on Airbnb: a literature review. Journal of Hospitality and Tourism Technology, 10(4), 814–844. https://doi.org/10.1108/jhtt-08-2018-0075

[2] RICHARD DIEHL MARTINEZ, ANTHONY CARRINGTON, TIFFANY KUO, LENA TARHUNI, NOUR ADEL ZAKI ABDEL-MOTAAL The Impact of an AirBnB Host's Listing Description 'Sentiment' and Length On Occupancy Rates. avi.org. https://arxiv.org/ftp/arxiv/papers/1711/1711.09196.pdf

[3] sklearn.metrics.roc_auc_score. (n.d.). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html#sklearn.metrics.roc_auc_score

[4] sklearn.metrics.auc. (n.d.). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html

[5] sklearn.naive_bayes.GaussianNB. (n.d.). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

[6] Pedamkar, P. (2021, April 30). Ensemble Methods in Machine Learning. EDUCBA. https://www.educba.com/ensemble-methods-in-machine-learning/?source=leftnav

[7] Chawla, R. (2019b, March 14). Analyzing the AirBnB Dataset for trends using Data Visualizations and Modeling. Medium. https://medium.com/ml2vec/data-analysis-on-the-airbnb-dataset-e0be9254eeb9

[8] Jones, A. B. (2018, June 20). Sentiment analysis of reviews: Text Pre-processing – Anna Bianca Jones. Medium. https://medium.com/@annabiancajones/sentiment-analysis-of-reviews-text-pre-processing-6359343784fb

[9] Richard Diehl Martinez, Anthony Carrington, Tiffany Kuo, Lena Tarhuni, Nour Adel Zaki Abdel-Motaal The Impact of an AirBnB Host's Listing Description 'Sentiment' and Length On Occupancy Rates

[10] Gupta, S. (2019b, January 5). Airbnb Rental Listings Dataset Mining - Towards Data Science. Medium. https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec

[11] Chen, B. (2021, March 13). All Pandas groupby() You Should Know for Grouping Data and Performing Operations. Medium. https://towardsdatascience.com/all-pandas-groupby-

you-should-know-for-grouping-data-and-performing-operations-2a8ec1327b5

[12] Hilsdorf, M. (2020, September 6). Dealing with List Values in Pandas Dataframes - Towards Data Science. Medium. https://towardsdatascience.com/dealing-with-list-values-in-pandas-dataframes-a177e534f173

[13] Sharma, A. (2020, June 25). What Are Lambda Functions | Lambda Function In Python. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/what-are-lambda-functions-in-python/

[14] Aprilliant, A. (2021, June 16). The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical). Medium. https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

[15] When-airbnb-listings-in-a-city-increase-so-do-rent-prices. (2019, April). https://hbr.org/2019/04/research-when-airbnb-listings-in-a-city-inc

ANNEXURE

Dataset:

Inside Airbnb. Adding data to the debate. (n.d.-a). Inside Airbnb. http://insideairbnb.com/get-the-data.html

GitHub Link:

Kanodia, P., Agarwal, A., & Gupta, I. (n.d.). Build software better, together. GitHub. https://github.com/prachikakanodia2507/Predicting-and-Analysing-Airbnb-Dataset