Analysis of the most influential factors on Online Hotel Aggregators in India

Database And Analytical Programming(H9DAP)

Sai Kaushik Kodali   
*School of Computing*   
*National College of Ireland* Dublin, Ireland  
x19194064@student.ncirl.ie

Akanksha Ashok Thorat  
*School of Computing*  
*National College of Ireland*Dublin, Ireland  
x19195443@student.ncirl.ie

Tej Rup Sai Munagala   
*School of Computing*  
*National College of Ireland*Dublin, Ireland  
x19196628@student.ncirl.ie

*Abstract*— As a traveller, one should do proper planning for their trips and take informed decisions in the process. There are certain web portals which help us make comparisons between different hotels by displaying their prices, locations and all the facilities that they provide. We have considered such four leading websites from India namely Booking.com, Stayzilla.com and Cleartrip.com. The customer interest is evaluated based on the hotels that they selected. Random forest and gradient boost machine learning algorithms are applied to check for the most influential factors. Also, the state in which the hotel is in the country and the facilities provided by the hotels are found to be the most influential factors from both these algorithms respectively. Random forest algorithm is found to be better than gradient boost algorithm in this research.

Keywords— Online hotel booking, gradient boost algorithm, random forest algorithm, booking, cleartrip, stayzilla

# Introduction (*Heading*

The usage of the internet has been exponentially increasing in the last few years. Everything that we do is related to the internet these days. Be it ordering food, planning a vacation or booking hotels. The hotel industries are flourishing very well. According to statistics, every year 148.3 million travel bookings are done online out of which hotel bookings account to 58.83 million that is 39% [1]. To increase the efficiency of the bookings and to make better marketing strategies it becomes necessary to analyse all the factors that go into the online hotel booking process. Also, the consumer needs and expectations while booking a hotel play an important role here. Our aim in this project is to ‘**analyze the most influential factors on the online hotel aggregators in India for the three chosen web portals**’. Some of the factors considered in the research are customer reviews, facilities that the hotels provide and the location of the hotels.

# Related Work

## Related Papers

This paper focuses on the hotel bookings done by the Irish people which was validated by a survey that was conducted on 12 people in groups of two. It was observed that around 3.8 million Irish residents travel abroad in the holiday season and this count has an exponential rise of 4.3% each year. Getting to know the consumer psychology behind booking the hotels, that is the factors that they consider the most while booking is the key to developing a good marketing strategy. The findings of this research stated that the location of the hotel was the most influencing factor in consumer psychology.

In this paper, the authors have explored the browsing preferences of the citizens of Malaysia and some international people surfing on the web as well. This study helps in understanding the pattern and likings of the people through their browsing details which will give the hoteliers a proper insight in planning a good business strategy for the hotel and tourism website. The analysis for the same was carried out using some tests like Levene’s test, Pearson’s test and independent-sample t test[2] .

This paper presents an automatic textual review system that was created to analyze the reviews given on the various web portals for hotel booking. Understanding the reviews is an integral part of the system as it lets us know the public opinion towards the facilities that they received at the hotel, which will prove to be useful to the industry. Natural language processing was used to understand the reviews. This system’s performance was satisfactory[3].

In this paper the study on hotel reviews was carried out for Skopje, a tourist destination in the Republic of Macedonia. The reviews of 80 hotels were considered which were posted on tripadvisor.com and booking.com. ArcGIS software was used for graphical analysis. Some facilities like sauna, soundproof, spa, fitness center and free parking was identified which affect the review score for the hotels. Amongst these the sound proofing and fitness centers were proved to be the most influential facilities which will help the hotel managers to get better reviews[4].

The authors for this paper examined the factors that affect people in choosing the online or offline channels in booking the hotels. A survey on 421 Hong Kong residents was carried out for analyzing the behavior of preferred channel adoption. It was observed that the people who chose the online medium for booking hotels get higher benefits, while the one’s choosing the offline medium receive a poor service quality[5].

Authors [Razavi](https://www.emerald.com/insight/search?q=Rouzbeh%20Razavi) and [Israeli](https://www.emerald.com/insight/search?q=Aviad%20A.%20Israeli) examined the prices of the hotels depending on the customer ratings and the stars the hotel has from perspective of supply and demand side in this research paper. Advanced regression models in machine learning were applied on hotel prices data, available from trivago.com and for combining this data Amazon Mechanical Turk portal was used. The customer ratings on the hotels was observed to be the most influential factor on the demand side and age, education and income were some other valuable characteristics which were needed to be considered[6].

This paper deals with the luxury brand hotels and the opinion that the consumers have of them. Through application of certain machine learning algorithms on wound 7000 consumer posted images of the hotel, it was observed that the specific interior of the hotels proves to put a good impact on the hotel brand[7].

The data of hotels in Tehran available on tripadvisor.com was analyzed by applying 8 different supervised machine learning algorithms like decision tree, SVM, logistic regression, KNN, Naive Bayes to name a few. This research aimed at providing a new user’s rate to hotel depending on the details taken from the user and the hotel. KNN emerged to be the best fitted model amongst all the models for this research[8].

## Our project relation with the related papers

Eight high-quality papers were studied for our research as can be seen in the above section. These papers were validated with various surveys conducted on Irish, Malaysian, Skopje and Hong Kong citizens (as mentioned in papers 1,2,4,5). These papers mostly focus on the improvement of the hotel booking websites by analyzing the factors like hotel details, location, facilities, reviews etc. mentioned on the website, which is also the part of our research. As supervised machine learning models were applied in the eighth paper, in our research as well we implemented the gradient boost and random forest algorithm to check for results. For our research, the random forest algorithm is found to be more efficient on the data that we have used.

# Methodology

**Data Source and Attributes:**

The datasets chosen for this analysis are public domain data for three major online hotel aggregators in India. The first three datasets for ‘Cleartrip’, ‘Stayzilla’ and ‘booking.com’ are available in csv format[9][10][11].

The attributes of each data sets are as follows:

1. **Cleartrip:**

Records: 5000 rows

Variables: 33 columns

1. **Stayzilla**

Records: 1200 rows

Variables: 33 columns

1. **Booking**:

Records: 6000

Variables: 29

All the mentioned datasets above have some common variables which can be used for comparative analysis these attributes include Hotel Names, Location of hotels, States, Price of rooms, Room facilities, Hotel Rating, Customer Review etc.

**Process Flow:**

A close up of a logo

Description automatically generatedThe analysis of the hotels takes place in five main phases. The first three phases comprise on an Extract, Transform, Load (ETL) Process. The fourth phase involves the visualization of the data while the final phase is based on analysis using Machine Learning algorithms.

**Process Implementation:**

A screenshot of a cell phone

Description automatically generatedEach dataset follows a well-structured Extract, Transform, Load (ETL) process. The structure of the ETL Process is given below:

1. **Data Extraction:**

The datasets used in the analysis have been sourced from Kaggle in csv format. The sourced csv files are then preprocessed into JSON format using the pandas library and built-in function in python.

The converted data is then loaded onto MongoDB into three different collections.

1. **Data Transformation:**

In the second phase of the process flow, the data in MongoDB collections is loaded into a python data frame. This data is in an unstructured raw format and requires some prerequisite preprocessing. The following preprocessing steps are applied on the dataset:

1. **Data Reduction:**

In this step the data in filtered to discard irrelevant and unwanted variables from the data. This is achieved using the ‘pandas.drop()’ function. The datatypes of certain variables are converted into suitable datatypes to facilitate loading into PostgreSQL.

1. **Deleting Missing Values:**

The missing (NA) values in each dataset is mitigated by either dropping the corresponding rows or by imputing the missing values with median/mean values of that particular variable.

1. **Handling Special Characters:**

The special characters in few columns which are deemed useful for further analysis are handled by using ‘encode ()’ and ‘decode ()’ functions.

1. **Data Load:**

This is the third phase of the process flow. In this phase the pre-processed data is converted into a structured database by loading the data frames into PostgreSQL. The pysog2 library was used of facilitate this process. Tables for each database have been created using SQL commands which would alter facilitate easy access to structured data for creating insights and visualizations.

1. A picture containing screenshot

   Description automatically generated**Data Visualization:**

The fourth stage of the process flow the structured data is queried from PostgreSQL into python to develop the required insights and visualizations. These visualizations aim at highlighting important factors for each individual hotel individually as well as comparative analysis of common variables of all aggregators pooled together.

The data from different tables in the database has been consolidated by implementing SQL join operations.

1. **Machine Learning**

The final phase of the process flow involves using the highlighted factors in the data visualizations to implement machine learning algorithms like Gradient Boost and Random forest algorithm.

**Process Automation:**

The process flow is encapsulated into one single script where all phases of the process flow are executed through this script.

The stepwise process automation is described below:

**Step 1:** ‘extract\_data.py’ file is executed to convert the CSV data acquired from the source to be converted into an semi structured JSON file.

**Step 2:** ‘loaddb\_mongo.py’ is executed to load the semi structured JSON files to MongoDB collections. With a database name ‘Hotels’.

**Step 3:** ‘etl\_process.py’ file is executed to extract the unstructured data from MongoDB and process it into a structured database format in PostgreSQL. This step also filters and cleans the data to facilitate smooth loading into Postgres.

**Step 4:** ‘visualizations.py’ is used to create insights and identify the important variables and factors in the data. The factors highlighted in this step are further used for implementing machine learning algorithms.

**Step 5:** ‘ml.py’ is executed to carry out the machine learning algorithms on the most influential factors identified in the visualizations.

# Results

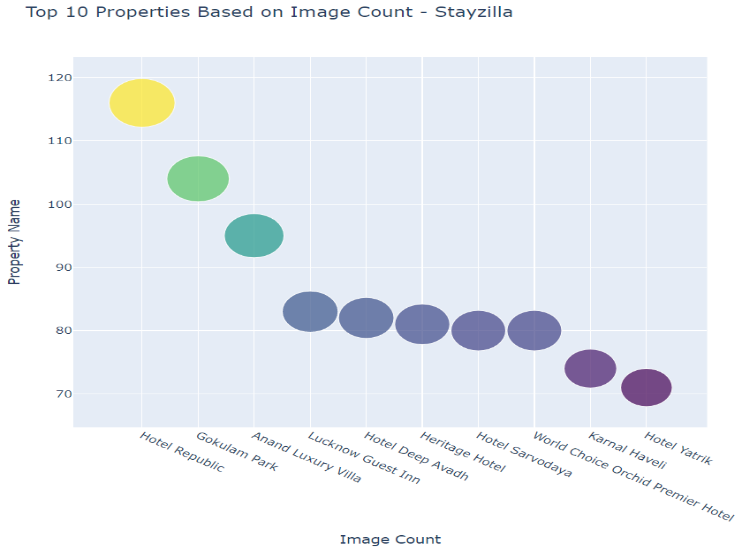
By using SQL commands, we are retrieving the data from the PostgreSQL. we are selecting the columns that are necessary for the visualization of our data and with the help of ‘Plotly’ library we are visualizing our data. We have total of four different dataset, and we are visualizing each of them individually and combinedly.

## Stayzilla.com

Our first dataset is Stayzilla, we have inferred following insights:

Plot of Top 10 expensive hotels:

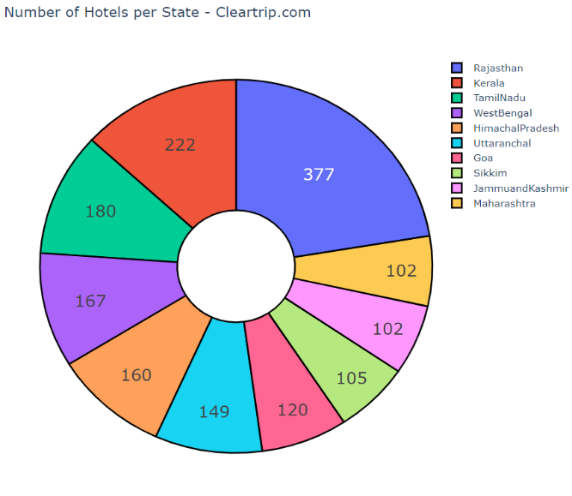
From the Fig: 3, tells about the top 10 expensive hotels on basis of price per night. The Luxury service apartment and P.K service apartment sector 93 are the pricey hotel from the Stayzilla website.

Plot of 10 hotels on bases of image count:

The Fig: 4, illustrates about the Top 10 properties based on the Image count. Image count given in the customer review is much important variable to for customers to use certain website. The hotel Republic and Gokulam Park are the top 2 hotels with 120 and 100 image reviews respectively.

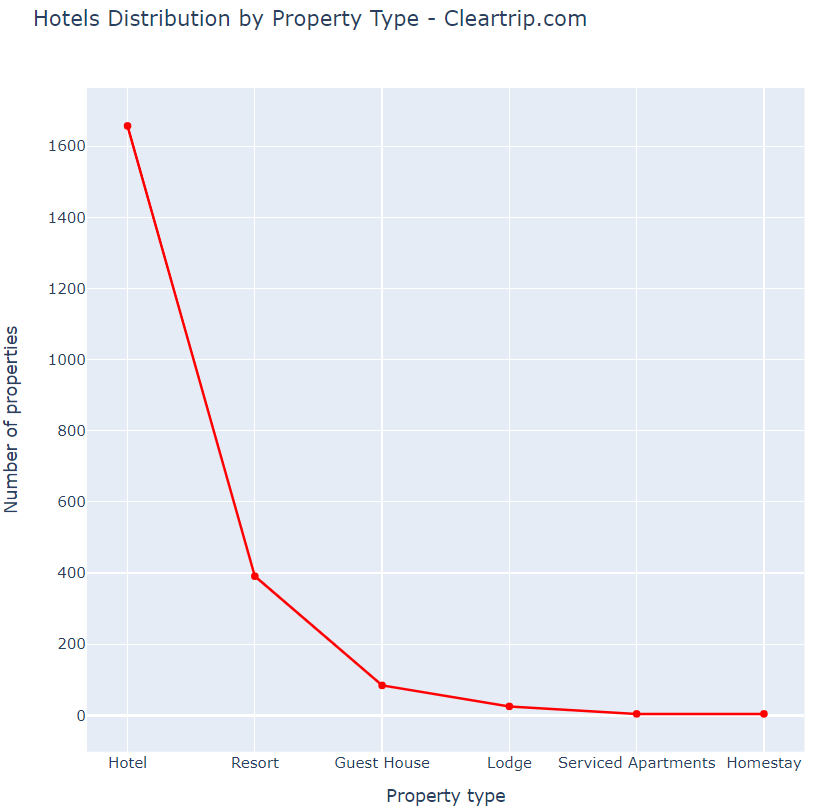
## Cleartrip.com

The second dataset is Cleartrip hotel data, following are the plots we have visualized here:

Plot of hotel count in each state:

Above fig.5 is the pie chart tells about the count of hotels in each state from the Cleartrip website. We can observe states like Rajasthan, Kerala and Tamil Nadu are the top three states with highest hotel counts from Cleartrip hotel website.

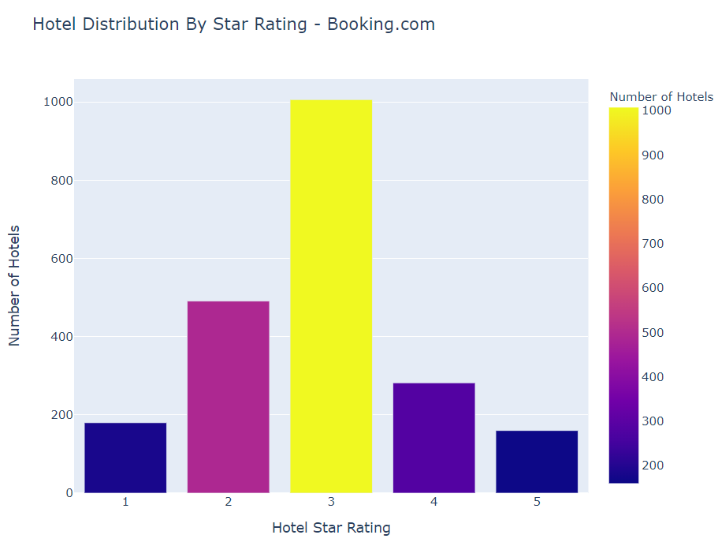
Plot of different property types:



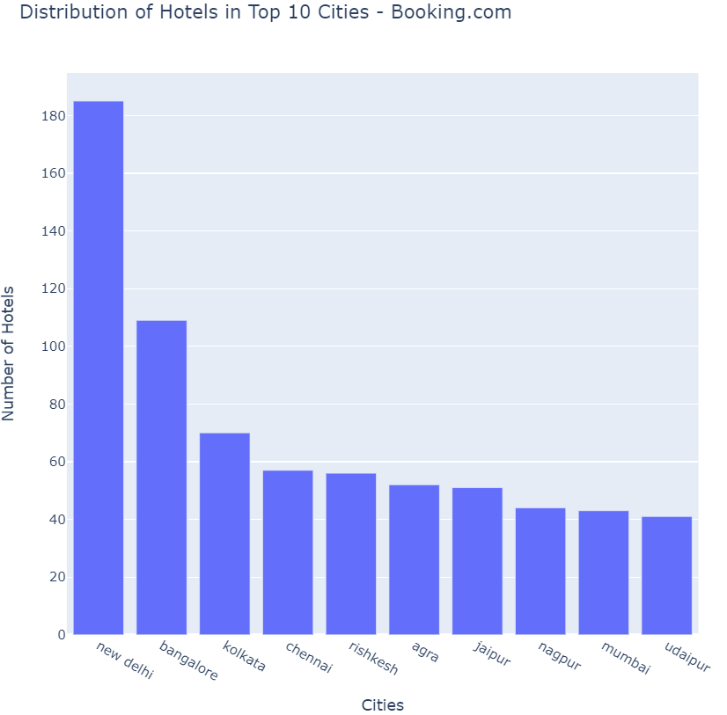
From the Fig:6, tells about all the count of all hotel types from the Cleartrip website. Here the top and least hotels that are recorded, homestay and serviced apartments ate the properties that have least count. However, properties like Hotel has the highest count of 1600.

## Booking.com

The third dataset that we used here is the Booking.com and the following insights are taken:

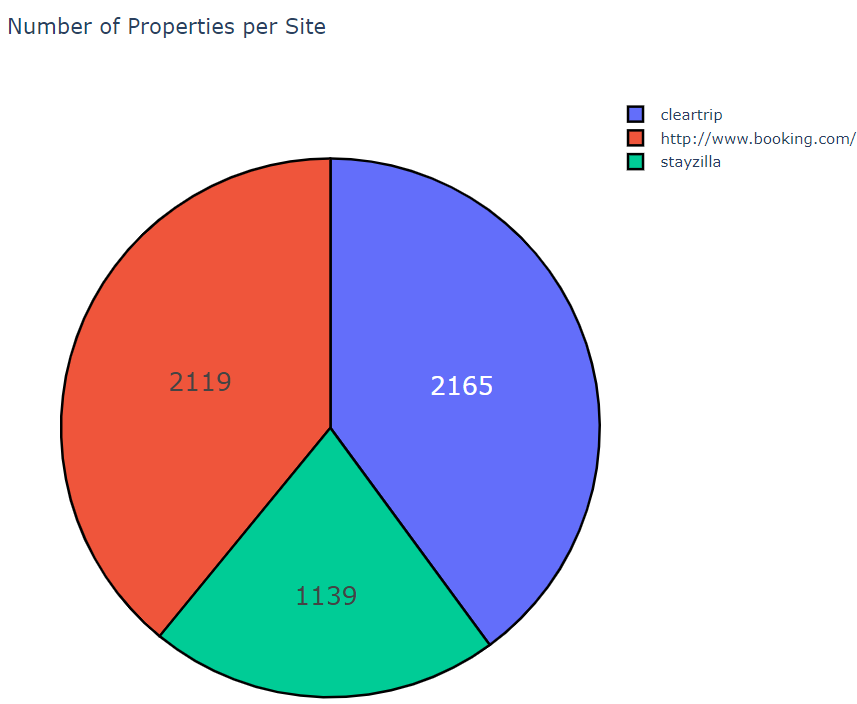
Plot based on hotel star rating:

The fig:7, tells about the distribution of different types of hotel based on the hotel star rating. We can infer that most of hotels from the booking.com website comes under three-star rating and the count of these three-star rating is about 1000.

Plot of top 10 cities on bases of hotel count:

The fig: talks about the count of all hotels in individual cities, cities like New Delhi, Bangalore and Kolkata have a greater number of properties with a count of 180, 100 and 80 respectively.

By combining all the three datasets on bases of a primary key that is ‘number\_of\_properties’ and took the count of all these unique properties and made the visualization of pie chart which shows the count of each websites hotel properties.



The Fig: 9, tells about the count of each individual hotel and shows that the cleartrip.com and booking.com has equal number of hotels.

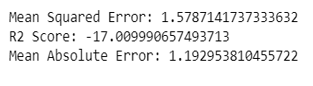
## Machine Learning implimentation:

Random forest and gradient boosting algorithms were chosen for project because of the efficiency they provide with data when the parameters to be considered are more. As in our datasets we are considering a lot of variables like star ratings, review count, hotel facilities, hotel location to name a few. Random forest and gradient boosting both work by creating trees, but the computation methodology of results for both is different. Thus, we will be using these two algorithms to know which are the influential factors that a customer is looking for while booking on an online hotel aggregator.

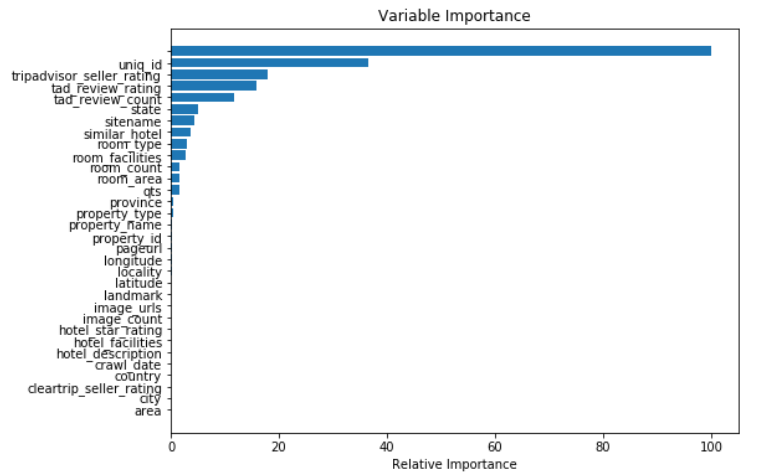
Gradient boosting algorithm:

After implying the gradient boost algorithm following are the results that are obtained.

Model Results:



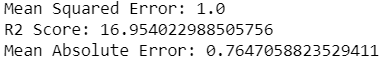
Influential Variables:



After applying the model, we came to know that seller\_rating, review\_rating, review\_count are the most influential factors in the dataset. The accuracy of prediction of this algorithm is 17%.

Random Forest Algorithm:

The following are the results after applying the random forest to know the influential factors from this algorithm:

Model results:

Influential Variables:

After applying the model, the model shows an accuracy value of 16%. The most influential variables are property\_name, pageurl and hotel\_facilites.

Out of the both algorithms we came to know what the most influential factors from the dataset are:



# Conclusion

In this system we have studied the overall working of three datasets namely booking.com, stayzilla.com and cleartrip.com based on their common key that is the count of the properties that each website has. Also, a relative study is done on the factors most influencing these websites by applying two machine learning algorithms random forest and gradient boost. It was found that the review counts and the ratings were of higher importance according to the gradient boost algorithm. While the property names and the facilities the hotels provided were more important according to random forest algorithm. Considering the mean square errors, the random forest algorithm seems to have given a more accurate result.

# Future Work

More websites like tripadvisor.com, trivago.com, Makemytrip.com can be analyzed and compared in this project further to get a better insight on this topic. Also supervised machine learning algorithms like KNN, decision tree and Naïve Bayes can be applied and their results can be checked and compared with our existing system.

# Reference:

[1] “Online Travel Hotel booking Statistics | Visual.ly.” [Online]. Available: https://visual.ly/community/Infographics/travel/online-travel-hotel-booking-statistics. [Accessed: 06-May-2020].

[2] P. Donald, “An exploratory study of the key factors driving hotel booking behaviour among consumers living in Ireland,” no. August, 2018.

[3] A. Y. Cheng and N. R. A. Hamid., “Behavior and preferences in browsing the travel and tourism websites,” in *2011 IEEE Colloquium on Humanities, Science and Engineering*, 2011, pp. 270–273, doi: 10.1109/CHUSER.2011.6163731.

[4] I. Perikos, A. Tsirtsi, K. Kovas, F. Grivokostopoulou, I. Daramouskas, and I. Hatzilygeroudis, “Opinion Mining and Visualization of Online Users Reviews: A Case Study in Booking.com,” in *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2018, pp. 1–5, doi: 10.1109/IISA.2018.8633597.

[5] M. Stojchevska, A. Naumoski, and K. Mitreski, “Modelling the Impact of the Hotel Facilities on Online Hotel Review Score for City of Skopje,” in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018, pp. 1–5, doi: 10.1109/ISMSIT.2018.8567317.

[6] Q. Wei, C. Cheung, and R. Law, “Tourist Perception of Online Hotel Bookings,” in *2010 International Conference on E-Business and E-Government*, 2010, pp. 2273–2276, doi: 10.1109/ICEE.2010.574.

[7] R. Razavi and A. A. Israeli, “Determinants of online hotel room prices: comparing supply-side and demand-side decisions,” *Int. J. Contemp. Hosp. Manag.*, 2019, doi: 10.1108/IJCHM-09-2018-0707.

[8] S. Giglio, E. Pantano, E. Bilotta, and T. C. Melewar, “Branding luxury hotels: Evidence from the analysis of consumers’ ‘big’ visual data on TripAdvisor,” *J. Bus. Res.*, 2019, doi: 10.1016/j.jbusres.2019.10.053.

[9] “Indian Hotels on Cleartrip | Kaggle.” [Online]. Available: https://www.kaggle.com/PromptCloudHQ/indian-hotels-on-cleartrip. [Accessed: 06-May-2020].

[10] “Indian hotels on Booking.com | Kaggle.” [Online]. Available: https://www.kaggle.com/PromptCloudHQ/indian-hotels-on-bookingcom. [Accessed: 06-May-2020].

[11] “Properties on StayZilla | Kaggle.” [Online]. Available: https://www.kaggle.com/PromptCloudHQ/properties-on-stayzilla#stayzilla\_com-travel\_sample.csv. [Accessed: 06-May-2020].

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**