Stayzilla Property Analysis

Anupam, Deep, Sahil and Vaishnavi

December 17, 2017

Table of Contents

1. Project Summary

1.1 Introduction

StayZilla was a Bengaluru-based Indian homestay network founded in 2005 as "Inasra" and rebranded as "StayZilla" in 2010, which acted as a marketplace for homestays and alternate stays in India, with around 55,000 stay options across 4,500 towns in the country [1]. Although it had first-mover advantage in the space, its growth amped up only after it was rebranded. Its founder and CEO, Yogendra accepted that this was achieved at quite a high-cost and had a lot of pitfalls. He laid out a couple of reasons which led to the failure as follows: Supply-Demand Mismatch, Creating a Market and High Costs, Low Revenues.

Supply-Demand Mismatch:

There was supply – demand mismatch as the travel market in India does not experience a network effect. When they get a demand for the booking, they experienced bad supply for the need on the homestay.

Creating a Market:

The system of homestays is something very new in a country like India. For Stayzilla, there was no ready market to sell the product and, thus, it required investing in educating the market about the concept of homestay market place, how to use the product and even on how to use Internet to the users.

High Costs, Low Revenues:

In an industry where offers and promotions were the norm, Stayzilla's team was focusing on marketing ROI and getting bookings without any discounts. Discounting-based growth rampant in the travel industry was another reason what led to the fall of Stayzilla. They Forced to match prices, but could not even recoup what they put in, necessitating a very large capital requirement, simply to sustain growth.

1.2 Dataset description

We have Stayzilla property dataset to anylyze and apply machine learning algorithms. Stayzilla was a property aggregator in India and it provides budget hotels, apartments, homestays and much more.

We have 1207 observation and 33 features in the dataset. Most of the features are text data which need to be processed to make it usefull for performing various machine learning algorithms.

Below are few useful feaures:

`city`, `latitude`, `longitude`, `occupancy`, `property_name`, `property_type`, `room_price`, `room_types` etc.

1.3 Dataset screenshots

Raw dataset: File - stayzilla_com-travel_sample.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	additional_info	amenities	check_in_date	check_out_date	city	country	crawl_date	description	highlight_value	hotel_star_rating	image_count	image_urls	internet	landmark	latitude	longitude	occupancy
2	Acceptance Rate	Newspape	04-08-16	05-08-16	Kalir	India	20-07-16	Sangsey Valley Resort is Located at the wonderf			3	http://stay-imgs.stayzilla.com/r			27.0874	88.53136	3 Adults 2 K
3	Acceptance Rate	Parking|AC	04-08-16	05-08-16	Kanp	India	20-07-16	What should you know? Enjoy unmatched servi			0				26.4665	80.34745	2 Adults 2 K
4	Acceptance Rate	Pickup & D	04-08-16	05-08-16	Jodh	India	20-07-16	What should you know? A budget hotel, this acc			18	http://stay-imgs.stayzilla.com/r			26.279	73.01907	2 Adults 2 K
5	Acceptance Rate	WiFi|New:	04-08-16	05-08-16	Jalpa	India	20-07-16	What should you know? The Riverwood Forest I			9	http://stay-imgs.stayzilla.com/r			26.8085	88.8236	1 Adult 2 Ki
6	Acceptance Rate	Newspape	04-08-16	05-08-16	Kanp	India	20-07-16	What should you know? Located at a walking dis			5	http://stay-imgs.stayzilla.com/r			26.4882	80.32663	2 Adults 2 K
7	Acceptance Rate	WiFi|Free	04-08-16	05-08-16	Jam	India	20-07-16	What should you know? The hotel is situated or			60	http://stay-imgs.stayzilla.com/r			32.7387	74.86555	2 Adults 2 K

pageurl	property_address	property_id	property_name	property_type	qts	query_time_stamp	room_price	room_types	search_term	service_value	similar_hotel	sitename	things_to_do	things_to_note	unic
https://i	South Sikkim, Kalir	67368	Sangsey Valley R	Resort	201	2016-07-20 10:59:55	3167per nigl	3 Bedded Room		Not Verified		www.stayzilla.com	All taxes. |	Comp	52a
https://i	Ratanlal nagar, Kar	28733	Hotel Amantran	Hotel	201	2016-07-20 10:59:55	815per night	Double Non-A/C Rooms		Not Verified	Hotel Mangala	www.stayzilla.com	Kindly Note: |	Th	49a
https://i	Shakti Nagar, Jodh	53742	Hotel Krishna	Hotel	201	2016-07-20 10:59:55	1624per nigl	Deluxe AC Double		Not Verified	Gouri Heritage	www.stayzilla.com	All taxes. |	Comp	8b5
https://i	Dakshin Dhupjhora	15468	The Riverwood F	Hotel	201	2016-07-20 10:59:55	3509per nigl	Premium Single A/c		Not Verified		www.stayzilla.com	All taxes |	Compl	032
https://i	John Wallinger Ave	47032	Hotel Vijay Inter	Hotel	201	2016-07-20 10:59:55	5802per nigl	Deluxe AC Double		Not Verified	Hotel Royal Cl	www.stayzilla.com	Complimentary		ed5
https://i	Jammu, Jammu	61528	Hotel Red Rose	Hotel	201	2016-07-20 10:59:55	1400per nigl	Deluxe A/c Double (EP)		Not Verified	Hotel Natraj | H	www.stayzilla.com	CP - Complemer		9d2
https://i	PWD Road, Jodhpu	28525	Hotel Inn Seasor	Hotel	201	2016-07-20 10:59:55	4062per nigl	Luxury Suite AC Double		Not Verified	The Fern Resi	www.stayzilla.com	All taxes. |	Comp	b19

Processed Dataset for Exploratory Analysis: File - stayzilla_com-travel_sample1.csv

acceptanc	response_	amenities	check_in_	check_out	city	populatio	city_classi	country	crawl_dat	descriptio	highlight_	hotel_sta	image_co	image_url	internet	landmark	latitude	longitude	occupancy
100	6	Newspap	04-08-16	05-08-16	Kalimpon	141576	Urban	INDIA	20-07-16	Sangsey Valley Resort is Locate			3	http://stay-imgs.stayzilla.com/			27.08743	88.53136	3 Adults 2
100	6	Parking | A	04-08-16	05-08-16	Kanpur	3470334	Metropoli	INDIA	20-07-16	What should you know? Enjoy			0				26.46646	80.34745	2 Adults 2
100	6	Pickup & [	04-08-16	05-08-16	Jodhpur	1378224	Metropoli	INDIA	20-07-16	What should you know? A bud			18	http://stay-imgs.stayzilla.com/			26.27902	73.01907	2 Adults 2
100	6	WiFi | New	04-08-16	05-08-16	Jalpaiguri	323445	Urban	INDIA	20-07-16	What should you know? The Ri			9	http://stay-imgs.stayzilla.com/			26.80852	88.8236	1 Adult 2

pageurl	property_	property_	property_	property_	qts	query_tim	room_pric	room_typ	search_te	service_va	similar_ho	sitename	things_to_	things_to_	room_type
https://w	South Sikk	67368	Sangsey V	Resort	2016-07-2	2016-07-2	3167per n	3 Bedded Room		Not Verified		www.stayzilla.com	All taxes. |	3 Bedded Non AC	
https://w	Ratanlal n	28733	Hotel Ama	Hotel	2016-07-2	2016-07-2	815per nig	Double Non-A/C Roo		Not Verifi	Hotel Mar	www.stayzilla.com	Kindly No	Double Non AC	
https://w	Shakti Nag	53742	Hotel Kris	Hotel	2016-07-2	2016-07-2	1624per n	Deluxe AC Double		Not Verifi	Gouri Her	www.stayzilla.com	All taxes. |	Double AC	
https://w	Dakshin D	15468	The River	Hotel	2016-07-2	2016-07-2	3509per n	Premium Single A/c		Not Verified		www.stayzilla.com	All taxes |	Single AC	
https://w	John Wall	47032	Hotel Vija	Hotel	2016-07-2	2016-07-2	5802per n	Deluxe AC Double		Not Verifi	Hotel Roy	www.stayzilla.com	Complime	Double AC	

Data for model building: File - Stayzilla_clean.csv

	property_name	city_class	room_types	image_count	property_typ	service_val	Accept_Rat	Response_time	amenities_cnt	description_cnt	rm_price	similar_hotel_cnt	Kids	Adults	room_type_ca	city_class_cat
2	Sangsey Valley R	Urban	3 Bedded No	3	6	0	100	6	2	127	3167	2	2	3	4	4
3	Hotel Amantran	Metropoli	Double Non	0	3	0	100	6	2	112	815	6	2	2	15	1
4	Hotel Krishna	Metropoli	Double AC	18	3	0	100	6	4	112	1624	14	2	2	14	1
5	The Riverwood F	Urban	Single AC	9	3	0	100	6	7	129	3509	2	2	1	17	4
6	Hotel Vijay Inter	Metropoli	Double AC	5	3	0	100	6	2	136	5802	3	2	2	14	1

2. Methodolology and Terminology

2.1 Linear Regression

Linear Regression is a method to predict dependent variable (Y) based on values of independent variables (X). It can be used for the cases where we want to predict some continuous quantity. E.g., Predicting traffic in a retail store, predicting a user's dwell time.

2.2 Logistic Regression

Logistic Regression is a probabilistic classification model which is used to predict a binary response from a mix of the qualitative and quantitative predictors. The main advantage of using logistic regression in our analysis is it doesn't suffer from severe class imbalance problems [2].

Logistic Regression models the log odds of the event as a linear function:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P$$

$$p = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P)\right]}$$

This nonlinear function is a sigmoidal function of the model terms and constraints the probability estimates to between 0 and 1.

2.4 Random Forest

Random Forest is one of the classification methods which is good for prediction but a little bit difficult to interpret.

Random Forest grows a big tree without trimming and then, take majority vote of the results of all the trees.

The process of this method is [2]:

1. Take a sample of size n from the training dataset
2. Randomly choose p variables from all the variables available
3. Train a single big tree on the sample dataset and using p variables
4. Repeat the step above B times
5. Take a majority vote of the results for all of the B trees

2.3 Gower distance Clustering

Since the data set obtained after pre-preprocessing is mixed, i.e., categorical as well as continuous, we perform clustering using Gower distance, partitioning around medoids, and silhouette width [3].

A popular choice for clustering is Euclidean distance. However, Euclidean distance is only valid for continuous variables, and thus is not applicable here. In order for a clustering algorithm to yield sensible results, we have to use a distance metric that can handle mixed data types. In this case, we will use Gower distance.

The concept of Gower distance is actually quite simple. For each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1. Then, a linear combination using user-specified weights (most simply an average) is calculated to create the final distance matrix. The metrics used for each data type are described below:

- quantitative (interval): range-normalized Manhattan distance
- ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
- nominal: variables of k categories are first converted into k binary columns and then the Dice coefficient is used

Gower distance can be calculated using the daisy function. After calculating the Gower matrix, we can identify the most similar and dis-similar pairs in the dataset.

2.5 Naive Bayes Classification

"The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high."[4]

2.6 Akaike information criterion (AIC)

AIC is used for model selection. AIC helps to know the best model out of all other models used However if all models are imperfect, AIC will not let about that instead it gives us the best among all those poor models.

The preferred model is the one which has the minimum AIC value.

3. Stayzilla Data Analysis

3.1 Data Preprocessing & Exploratory Data Analysis

3.1.1 Data Preprocessing:

1. Imported data into R for data pre-processing and further analysis.
2. Derived new features from existing features and dropped features which are of no use like below ones:

   "additional_info", "check_in_date", "check_out_date", "country", "crawl_date", "highlight_value", "hotel_star_rating", "image_urls", "internet", "landmark", "pageurl", "property_address", "property_id", "qts", "query_time_stamp", "search_term", "sitename", "things_to_do", "things_to_note", "uniq_id"

3. Extracted numeric values for acceptance rate, response time, adults, kids and room price from string values.

additional_info	acceptance_rate	response_time
Acceptance Rate:100 percent|Response Time:< 6 hours	100	6
Acceptance Rate:100 percent|Response Time:< 6 hours	100	6

occupancy	Adults	Kids
3 Adults 2 Kids	3	2
2 Adults 2 Kids	2	2

room_price	rm_price
3167per night incl. tax	3167
815per night incl. tax	815

4. Downloaded some additional data from India Consensus website [5], like population and tier information which we used to classify the Stayzilla property by city tier.
5. Replace the blank values for amenities and description with NA.
6. Checked the count of missing data for all features:

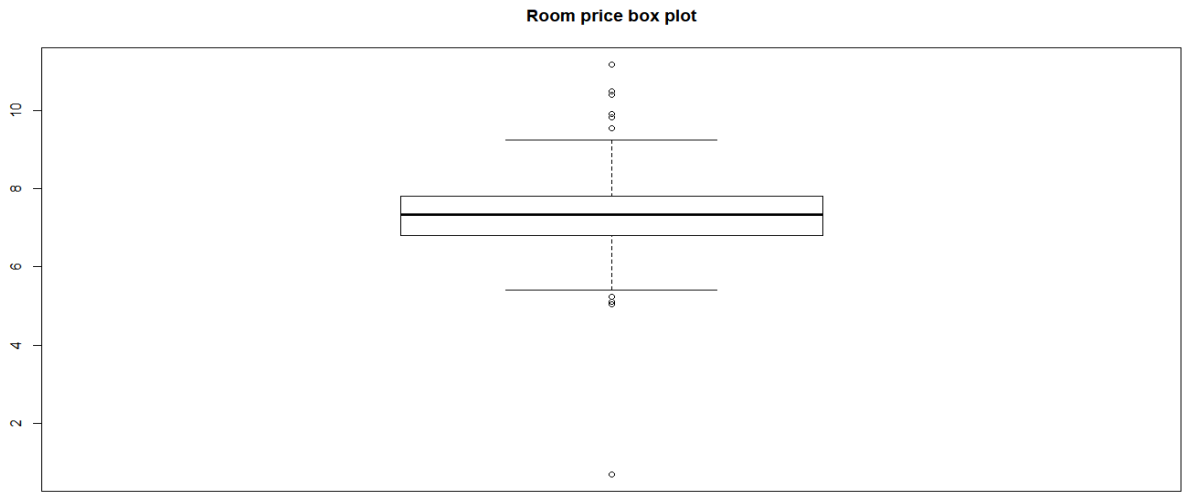
	missingCount	missingPercent
amenities	53	5%
description	220	19%

   We saw that only amenities and description field has missing values.

7. As the missing values are in amenities and description which are textual hence we just took the complete cases by skipping observation with NA values.

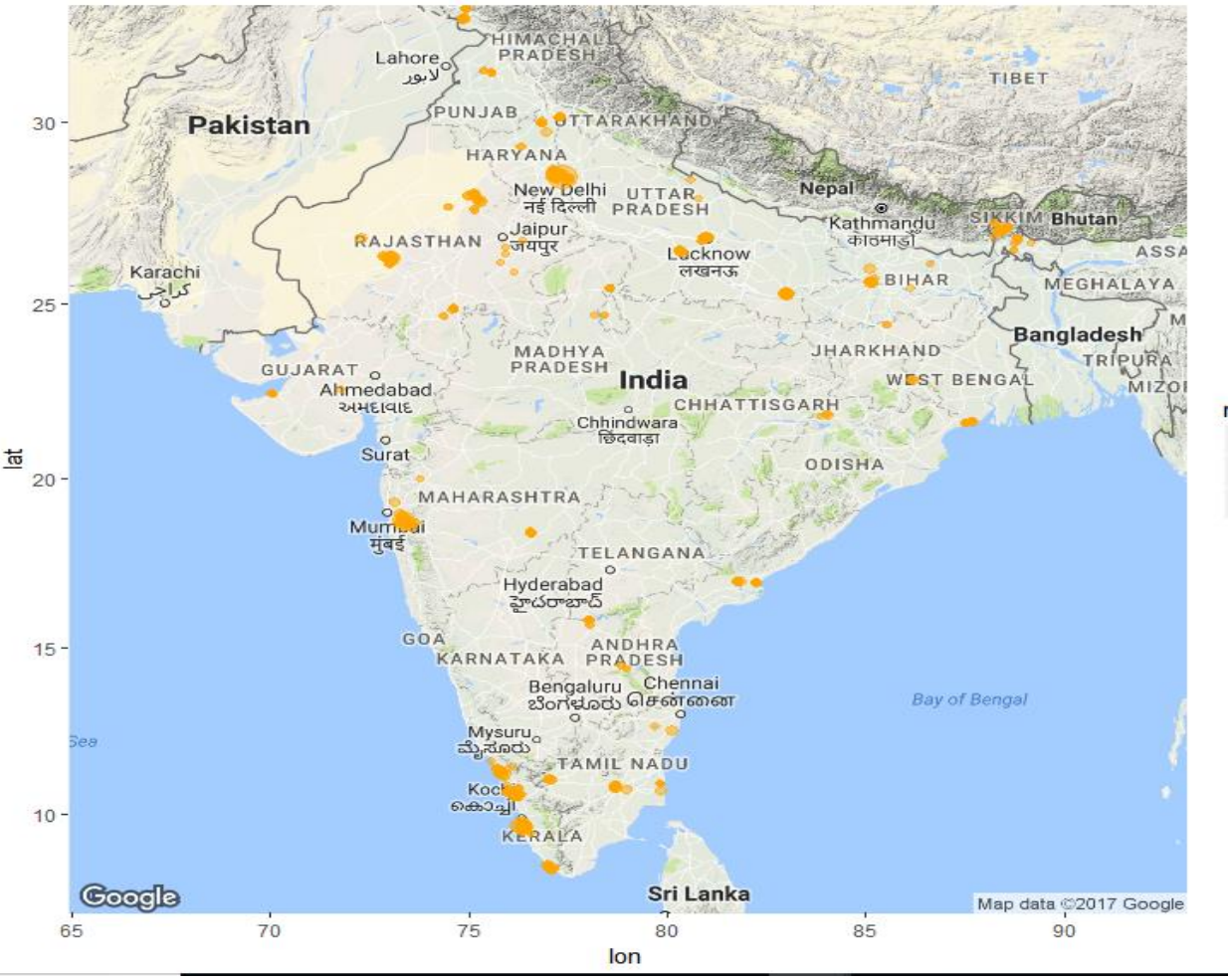
3.1.2 Exploratory Data Analysis:

1. Box plot showing some outliers in room prices:

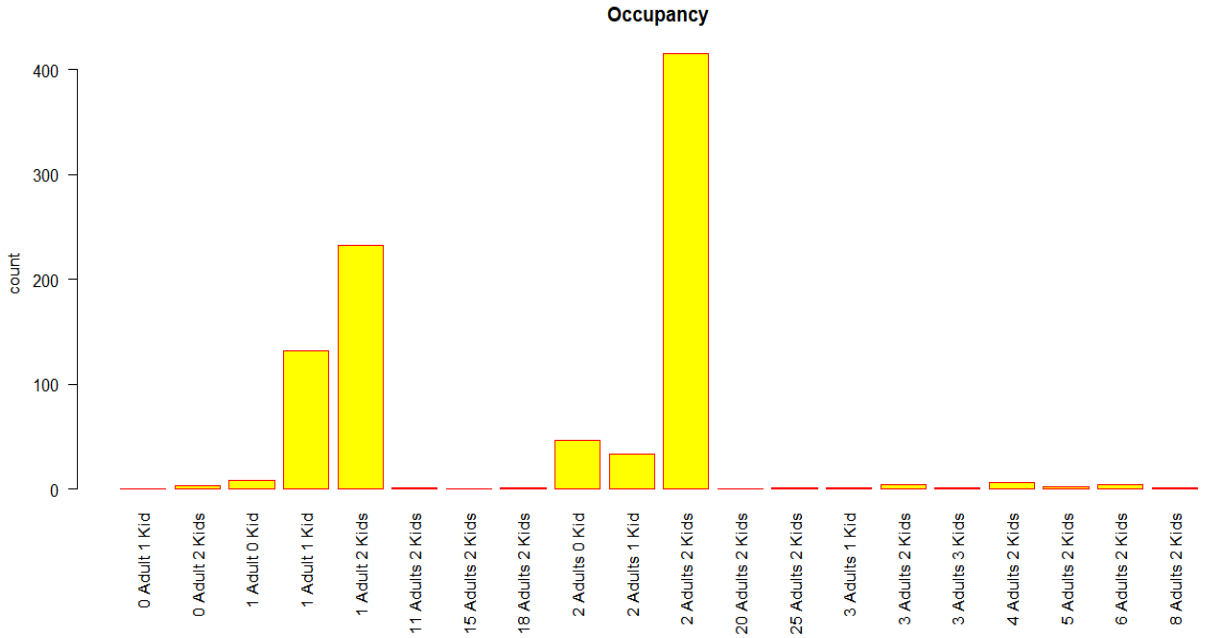


2. Demographic of Stayzilla property in India:

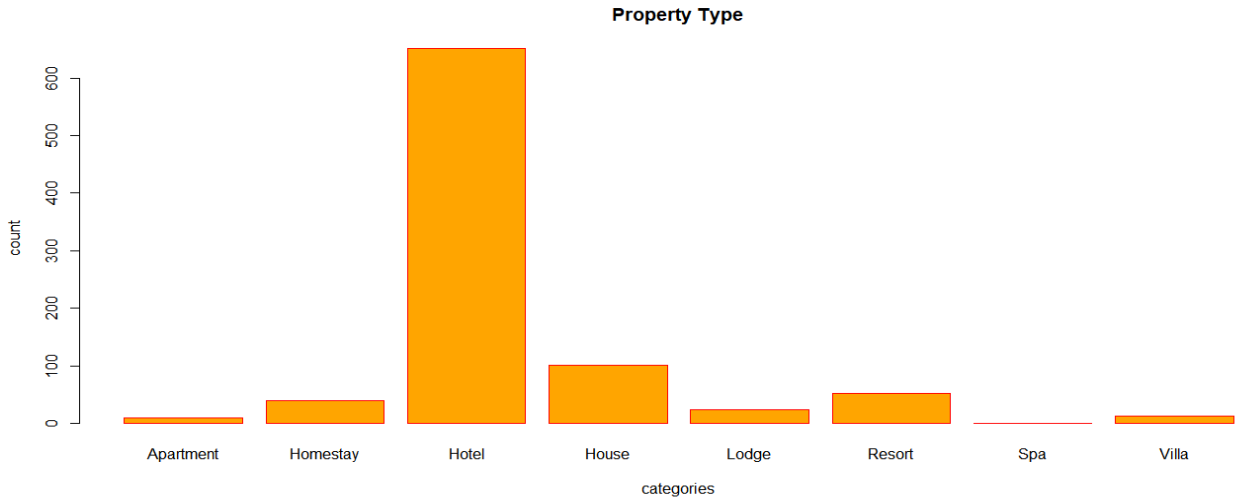
   We see that most of the Stayzilla properties were located in North and south rejoins like New Delhi, Noida, Kochi, Kovalam etc.



3. Most of the occupancy in Stayzilla was of 2 Adults and 2 kids meaning that most of the customers of Stayzilla properties are families.

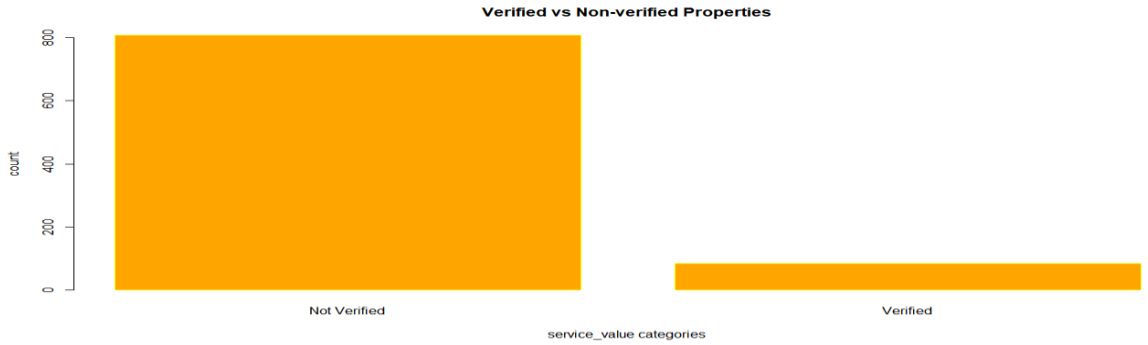


4. Major tie ups of Stayzilla was with Hotels only as we can see in below graph:



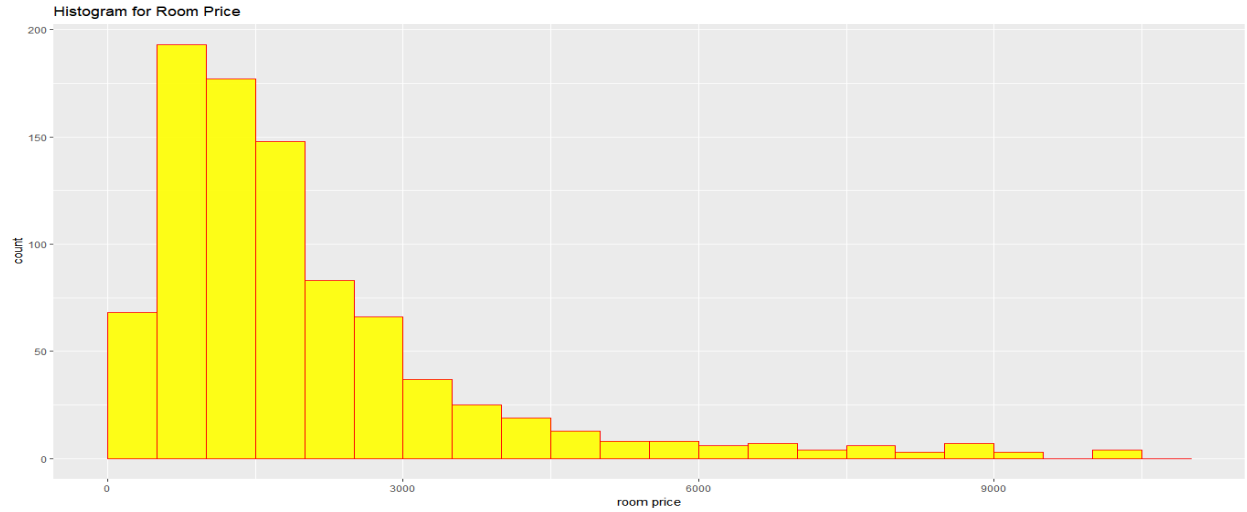
The vast majority of the properties on the site are not homes but (self-described) hotels. This is a prominent difference between StayZilla and the US equivalent, AirBnB, which does not allow (explicit) hotels in its listings. Hotels are subject to a raft of legal oversights that listing on a homestay platform sidesteps, a fact which has led to huge and very public fights between AirBnB and many of the US cities it operates in; evidently this is much less of or not even a concern in India.

5. We found that most of the properties registered on Stayzilla are non-verified.

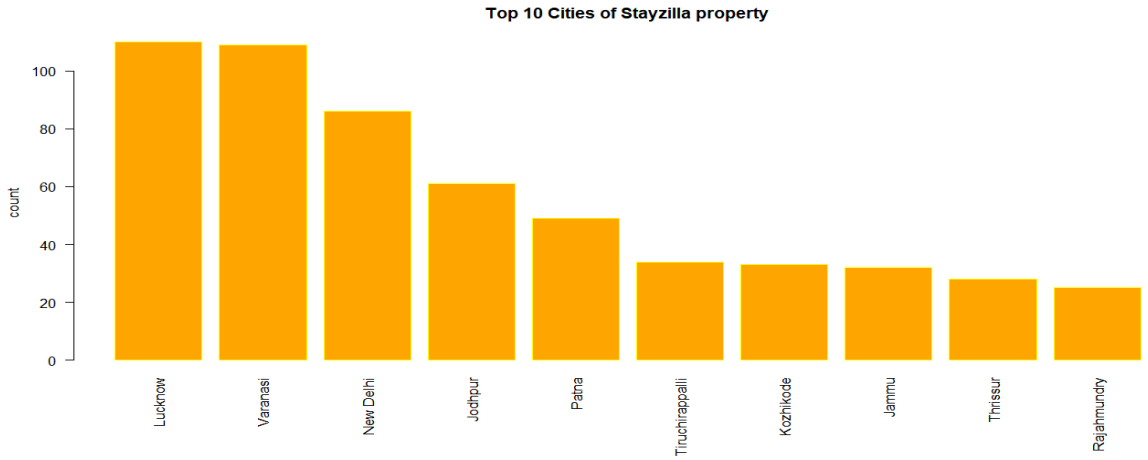


The vast majority of properties on the StayZilla platform seems to be unverified which is not trust worthy for the customers. **We can recommend them to make the properties verified to gain the trust based relationship from the customers.**

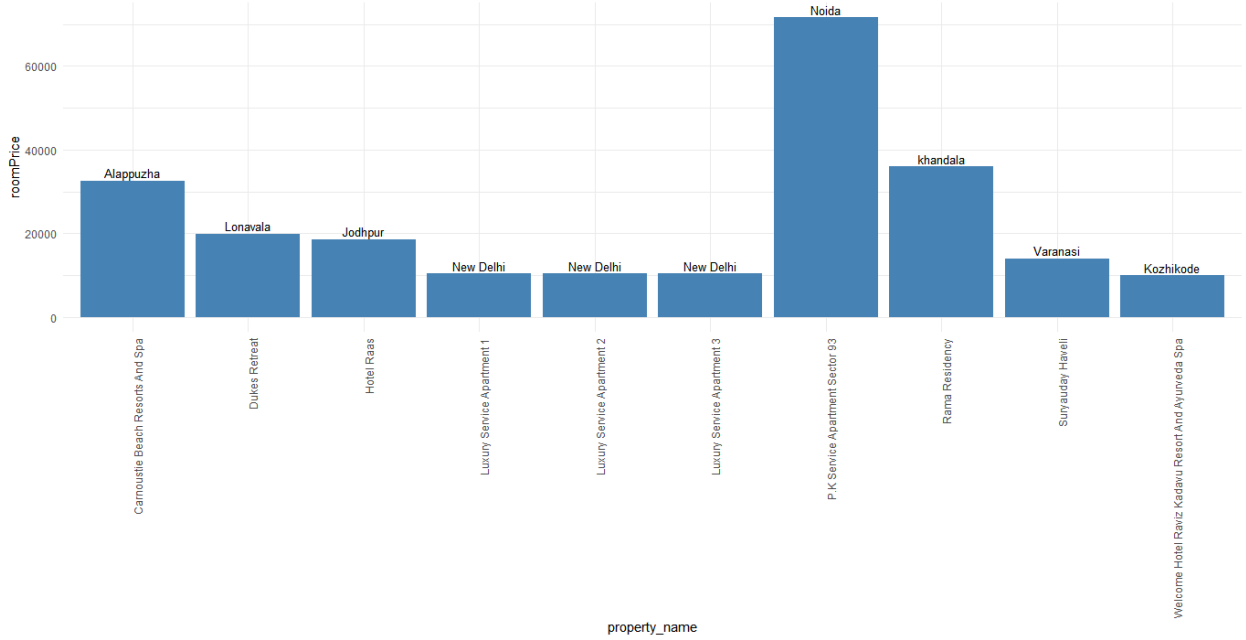
6. We also found that most of the Stayzilla property prices are below 3000



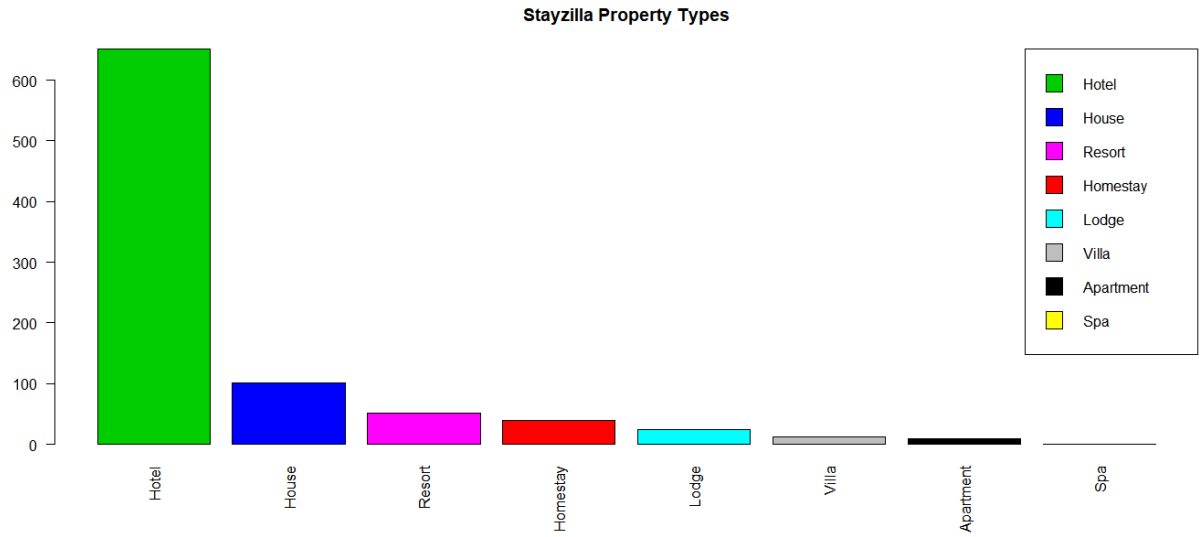
7. Top 10 cities for Stayzilla properties:



8. Top 10 most expensive of Stayzilla properties:



9. Most of the Stayzilla property types are Hotels only:



3.2 Predictive Analysis & Model Building

We divided data into test and training using random sampling method. We used 80% of the data for training and 20% for testing.

3.2.1 Regression

Linear Regression:

We run the Linear regression in created two models: (1) model1 - With all independent variables, (2) model2 - with important variables which are suggested by stepAIC() function which performs stepwise regression.

mod1 <- lm(rm_price ~., data = train[,c(-1)])

mod2 <- lm(rm_price ~ city_class + room_types + property_type + service_value + description_cnt + similar_hotel_cnt + Adults, data = train[,c(-1)])

We have got below RMSE for these two Linear regression models:

	Training RMSE	Test RMSE
model1	2584.529	2584.529
model2	2590.172	2590.172

As per RMSE values, model1 looks good as it has low RMSE.

We found that to predict room price, we have few significant variables like: city_class, room_type, property_type and similar_hotel_cnt. These variables have p-value less than 0.05.

Regression using SVM:

We also ran the regression using SVM to know if SVM works fine or not on the given dataset. We got below result from the SVM:

	Training RMSE	Test RMSE	Kernel	Cost	gamma	epsilon	support vectors
modSVM1	3039.239	2808.717	radial	1	0.03125	0.1	604
modSVM2	2934.162	2752.776	radial	3	0.030303	0.3	315

SVM model2 was run using cost function and it gave improved performance over the previous model however SVM is not good on the given data when compare with Linear regression as Linear model 2 gives lower RMSE means will get high accuracy in predicting the room price.

3.2.2 Classification

We performed the Multinomial Logistic Regression, Random Forest and Naïve Bayes classification to predict Stayzilla property types as property type is a categorical variable and has multiple classes.

Multinomial Logistic Regression:

Response Variable: property_type

Classes: Apartment, Homestay, Hotel, House, Lodge, Resort, Spa and Villa

Divided dataset into Training (75%) and testing (25%) data and predicted property type based on following features:

image_count, amenities_cnt, description_cnt, rm_price, similar_hotel_cnt

We got 72% accuracy with 95% CI: (0.6596, 0.7674)

Random Forest:

Random Forest is one of the best ensemble model for classification task. Below is the model formula used for Random Forest:

```
randomForest(formula = property_type ~ image_count + amenities_cnt +
description_cnt + rm_price + similar_hotel_cnt, data = train)
```

We got 73% accuracy with 95% CI: (0.6706, 0.7772)

Naïve Bayes Classification:

We also used Naïve Bayes to predict property types and got 66% accuracy with 95% CI: (0.5978, 0.7112)

3.2.3 Clustering

Gower Distance Clustering:

We used Gower distance clustering techniques to cluster most similar and dissimilar property. We used this techniques as we have mixed type of variables meaning that we have few categorical and few quantitative variable and in this case Gower clustering works well.

Below are the model highlights and results:

```
> # Output most similar pair
> View(Stayzilla_clean[
+   which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
+         arr.ind = TRUE)[1, ], ])
```

property_name	city_class	room_types	image_count	property_type	service_value	Accept_Rate	Response_time	amenities_cnt	description_cnt	rm_price	similar_hotel_cnt	Kids	Adults	room_type_cat	city_class_cat
Fariyas Holiday Resort	Semi-urban	Double Non AC	26	6	0	0	0	2	96	9412	2	2	2	15	3
Manaka Guest House	Metropolitan	Single Non AC	7	4	1	100	19	4	32	332	13	0	1	18	1

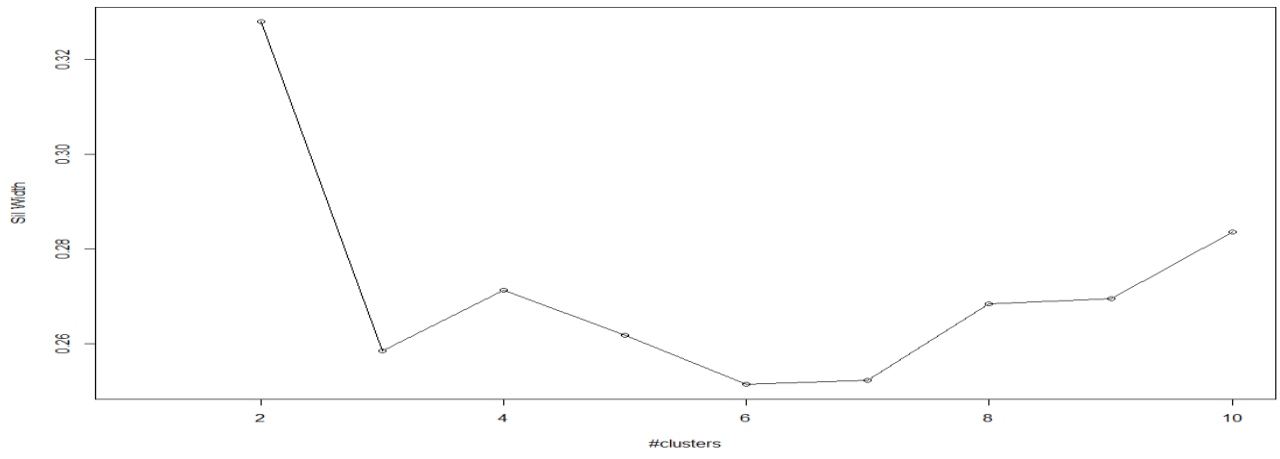
From the above result we can see that, Hotel Relax and New Garden is the most similar pair.

```
> # Output most dissimilar pair
> View(Stayzilla_clean[
+   which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
+         arr.ind = TRUE)[1, ], ])
```

property_name	city_class	room_types	image_count	property_type	service_value	Accept_Rate	Response_time	amenities_cnt	description_cnt	rm_price	similar_hotel_cnt	Kids	Adults	room_type_cat	city_class_cat
Hotel Relax	Urban	Single Non AC	0	3	0	100	6	2	2	389	2	1	1	18	4
New Garden Hotel	Urban	Single Non AC	0	3	0	100	6	2	2	378	2	1	1	18	4

We can see the dis-similarity between the records, for image count, amenities , response time, room price and similar hotel count.

Plot Silhouette width



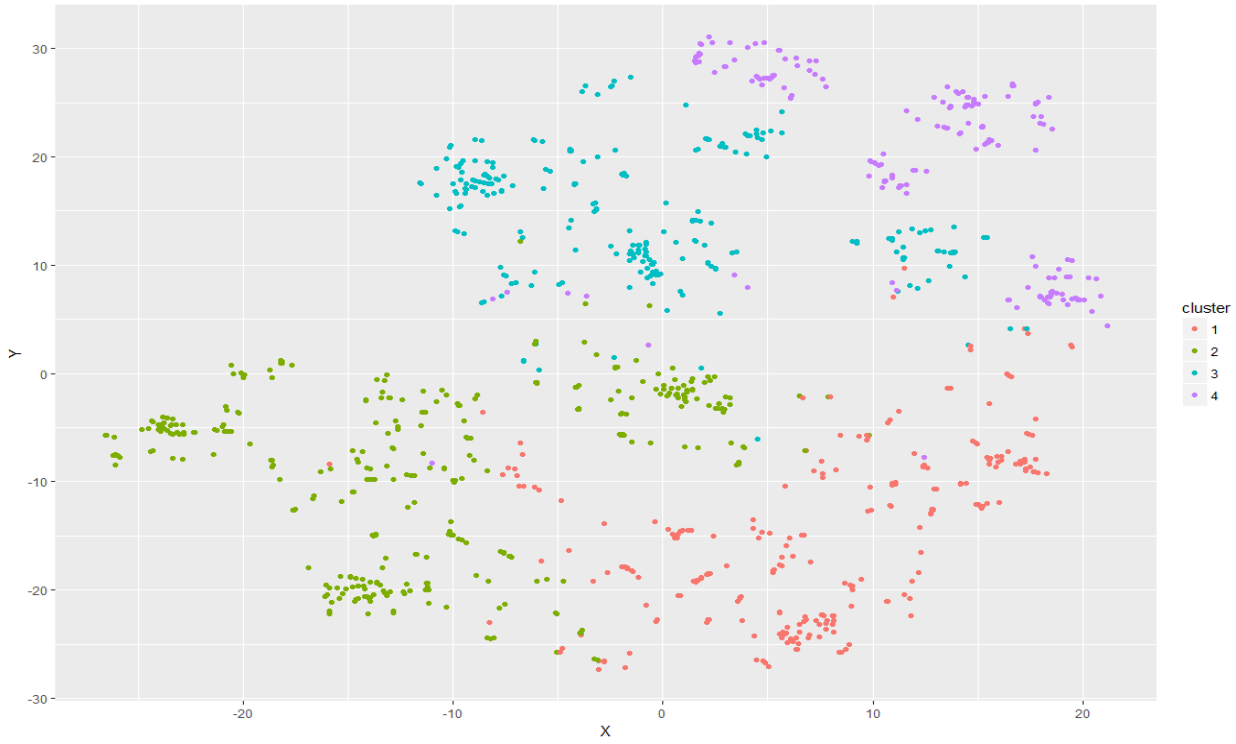
Performing the clustering by initializing 4 clusters.

```
> View(Stayzilla_final[pam_fit$medoids, ])
```

Below are the 4 medoids for the 4 clusters.

property_name	image_count	property_type	service_value	Accept_Rate	Response_time	amenities_cnt	description_cnt	rm_price	similar_hotel_cnt	Kids	Adults
Hotel Pushpa	4	3	0	100	6	2	115	473	3	2	2
Delight Inn	8	3	0	100	6	2	103	1778	11	2	2
Hotel Charans International	12	3	0	100	6	2	108	1959	11	2	1
Sri Iswarya Residency	4	3	0	100	6	2	129	620	3	1	1

Plotting the 4 clusters data:



3.2.4 Natural Language Processing

We used NLP technique to know the words mostly used in property description so we can understand the relation between those words and property. This particular machine-learning task inspects the entire "description" column in our dataset. After Inspection, Algorithm performs human-like understanding of text in the description column and spots various person entity and locations discussed while describing the properties on the website. By this, we get the location and name counts. Using location count, we could check for which particular locations were most discussed amongst the properties while describing themselves on Stayzilla.

4. Conclusion

In this study, we have done data preprocessing, exploratory analysis and applied many machine learning techniques to Stayzilla dataset and explored which are the techniques usefull on the given data. After analyzing the results from multiple machine learning algorithm in different cases we can

conclude that Linear regression is not working well on the given dataset as we do not have features which are good at predicting the room price. In case of classification of property types, among multinomial logistic regression, Random Forest and Naïve Bayes classification techniques, Random Forest working best to classify the property types as we got high accuracy (73%) in that. Below are the accuracy details of all three classification techniques used here:

	Multinomial LR	Random Forest	Naïve Bayes
Accuracy	72%	73%	66%
95% CI	(0.6596, 0.7674)	(0.6706, 0.7772)	(0.60, 0.72)

In case of clustering, Gower Distance clustering is doing a great job in clustering the similar and dissimilar properties using both categorical and quantitative variables.

5. References

1. https://inc42.com/buzz/stayzilla-shutdown/
2. http://www.columbia.edu/~jc4133/ADA-Project.pdf
3. https://www.r-bloggers.com/clustering-mixed-data-types-in-r/
4. http://www.statsoft.com/textbook/naive-bayes-classifier
5. http://www.censusindia.gov.in/2011census/population_enumeration.html