Northeastern

COURSE

COLLECTING, STORING AND RETRIEVING DATA
DA5020

TERM

Full Summer 2017

PROJECT TITLE

TRIPADVISOR VS AIRBNB: A COMPARATIVE STATISTICAL STUDY

Project done by:

Khozema Lokhandwala (NUID: 001621853) &
Pramod Nashipudi

Table of Contents

OVERVIEW

TripAdvisor and Airbnb are two of the most well-known vacation home rental sites and each have unique components to them. So, how do you choose the right one for your vacation? Also, could we get some hints from price trends for helping Airbnb hosts adjusting their listing price?

The primary objective of this project was to study how the vacation home rental services, specifically Airbnb and TripAdvisor, compare against each other for the same user input. The statistical study done can also potentially help a host in deciding to choose a correct rent price for his listing to maximize his revenue. It was decided to query a search for 2-night stay for 2+ guests in Washington, DC. We decided to choose Washington, DC because it was a familiar place and thus it would help in interpreting the data analyzed in the best possible way.

The purpose of this report is to document both the implemented database design and all corresponding data collection, data cleaning and exploratory data analysis conducted as part of this comparative study.

For this project, the main focus was on a method to collect the data from these two services using web scraping, clean the raw data that was scraped, store this dataset and retrieve it using an analytical package.

SPECIFIC GOAL

1. Web scraping Airbnb data & TripAdvisor Vacation Rentals data for search query:
    a. Location: "Washington DC"
    b. Check in: 1st September 2017; Check Out: 3rd September 2017
    c. 2 guests
2. Data cleaning using R
3. Database storage in SQL Server and Retrieval using RSQLite
4. Exploratory data analysis
5. Random forest model for estimating which features affect prices

TOOLS USED

- Import.io - Web Scraping API: *For extracting data elements out of the Airbnb and TripAdvisor websites.*
- Microsoft Excel: *Spreadsheet used to store the scraped raw data and perform initial data cleaning.*
- R (for data manipulation, visualization and data modelling): *tidyverse, tidyr, openxlsx, base, randomForest, dplyr, RSQLite.*

DATA COLLECTION

1. WEBSCRAPING DATA FROM Airbnb.com & TripAdvisor.com

The data to be scraped from the websites are shown below:

DATA REQUIRED	
Name	Reviews
Bedrooms	Ratings
Bathrooms	Location
Accommodates	Price/night
Amenities	House rules
Fees	Safety features
Tax	Host details
Property type	Cancellation policy
Room type	

NOTE: Exact location for any listing from Airbnb and TripAdvisor could not be obtained as it is not made available until booking is confirmed.

To achieve the data extraction (web scraping), the web scraper Import.io was used. Each step is described below in detail. The steps shown below illustrate the scraping procedure for Airbnb data.
NOTE: Similar procedure was used for scraping data from TripAdvisor.com which is not shown in the report.

1.1. Getting a full list of apartments/houses from Airbnb.com and TripAdvisor.com

a) The first step was to get a full list of vacation rentals from both websites before we get the detailed information of each listing. The name of the listing with URL along with price, reviews, #beds, property type details was selected in the extractor.

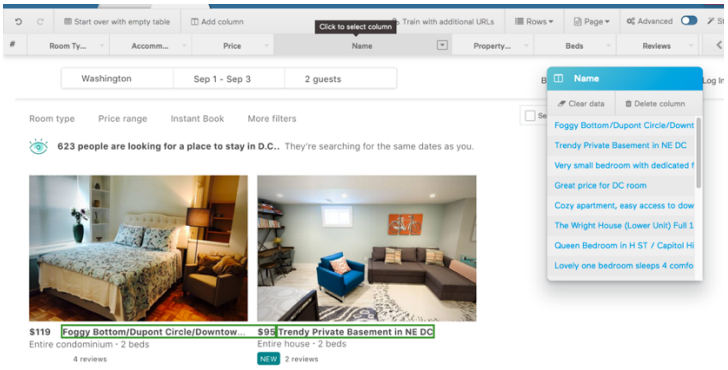


Figure: 1

b) The extractor was then trained for all the pages containing the listings (Airbnb-17 pages; TripAdvisor-5 pages) and saved. The URLs for all the pages are entered in the settings menu of the extractor as shown in the figure 2 below. The result obtained after extracting the above details was previewed inside the extractor as shown in Figure 3.

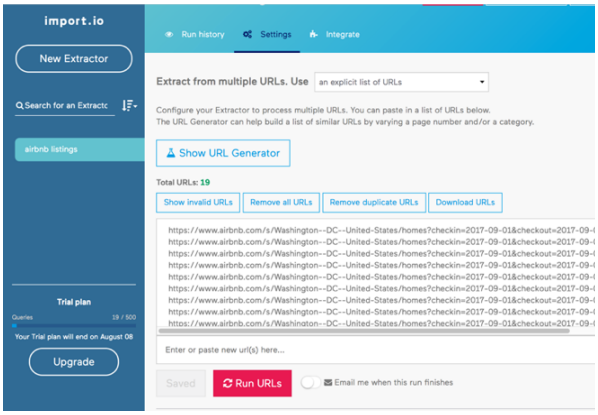


Figure: 2

Figure: 3

1.2. Getting details of apartments/houses from Airbnb.com and TripAdvisor.com

1. After the extractor was run for all the listings it was time to extract the details from each individual listing like ratings, amenities, house rules, safety features, fees, bathrooms etc. An example of extracting the name of property from the detail page is shown in the screenshot below. Extracting the ratings of that property was a challenge as it was an embedded attribute inside the stars and could not be selected directly by the extractor. Thus, for ratings we had to manually select the XPath for the ratings attribute by inspecting the HTML source and entering it in the manual XPath setting of that Ratings column. It is demonstrated in the Figure 5 below. Another issue was extracting the amenities from the detail page on

Airbnb. The top amenities were shown directly on the page but for the rest of the amenities the data was hidden for which we had to convert the page to a JavaScript page.

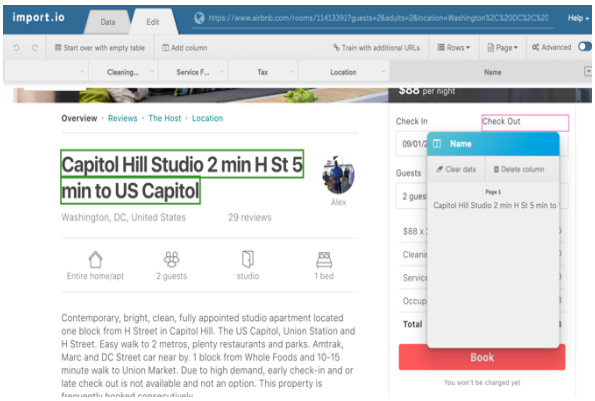


Figure: 4

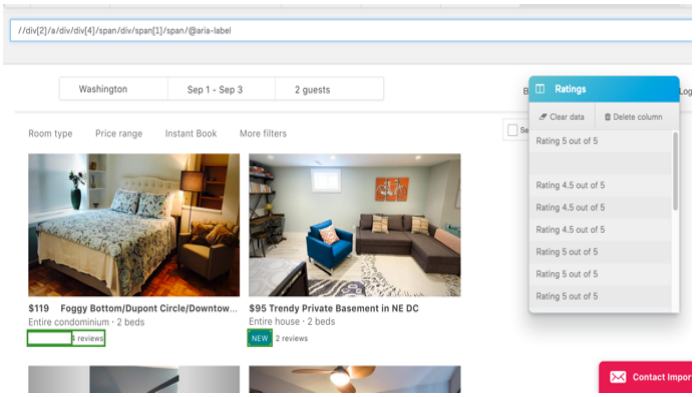


Figure: 5

2. Next step was to link the detail page extractor to the extractor which contained all of the listings. The "URLs from another extractor" option in the settings provides that functionality as demonstrated in the figure 6 below. The extractor was then run and the final result obtained was exported as an ".xlsx" format file which is shown in figure 7 below.

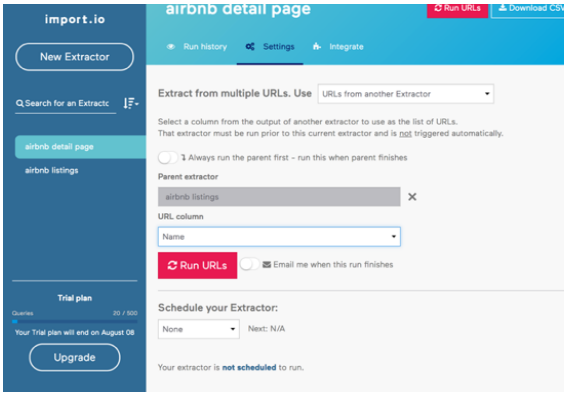


Figure: 6

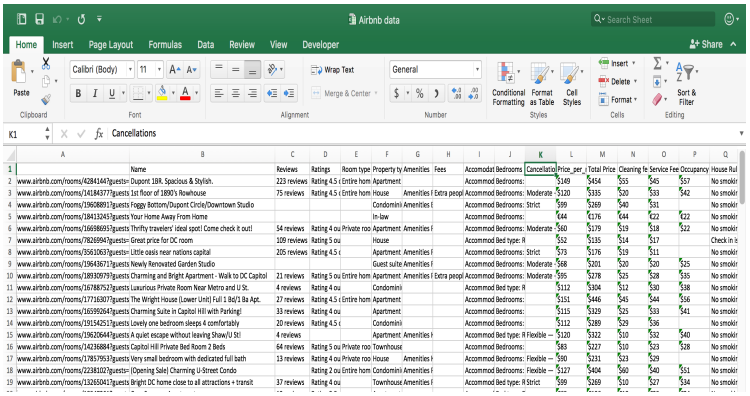


Figure: 7

2. TIDYING AND CLEANING DATA IN R

The scraped data was then imported in R using the *read.xlsx ()* function.
The data was in the raw form and needed cleaning and manipulation to standardize the format in each column. This data cleaning and manipulation was done in R using concepts learnt in the DA5020 course. Columns like Ratings, Reviews, etc. were cleaned of any text and converted into numeric variables. Other columns like Room type, Property type, locations were also standardized into consistent format across all rows of the dataset using *stringr* functions in R.

Initially the Amenities Column contained all amenities as one string concatenated using ";" and thus we had to separate the Amenities into different columns which would make it easier for us to do some analysis on it. Thus new columns for each amenity were created using *stringr* function.

Another challenge was to deal with NA values in the dataset. For columns such like Reviews, Host Reviews, Cleaning fees, service fees and tax, all the NA values were replaced by zeros as it made the most sense. For the Ratings column for both services and Bookings column for TripAdvisor we used a prediction model to fit in the values for NA using the other records using ANOVA analysis. This model will be discussed later in section 4.6.

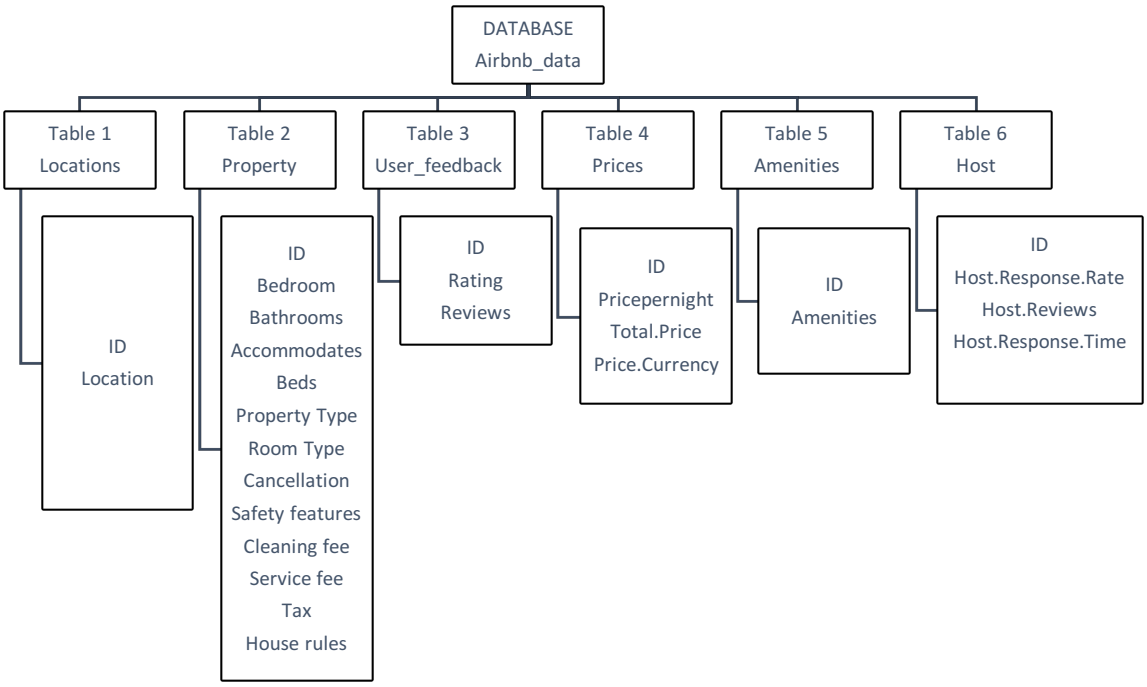
DATA STORAGE

3. DATABASE STORAGE USING SQLite

After the dataset was cleaned and formatted it was time to use a database server for storing this data. For our database design, we chose to use SQL as our database server because of the relationships between variables that exist in our data. A NoSQL database such as MongoDB is better meant for datasets having embedded documents. and no relationships.

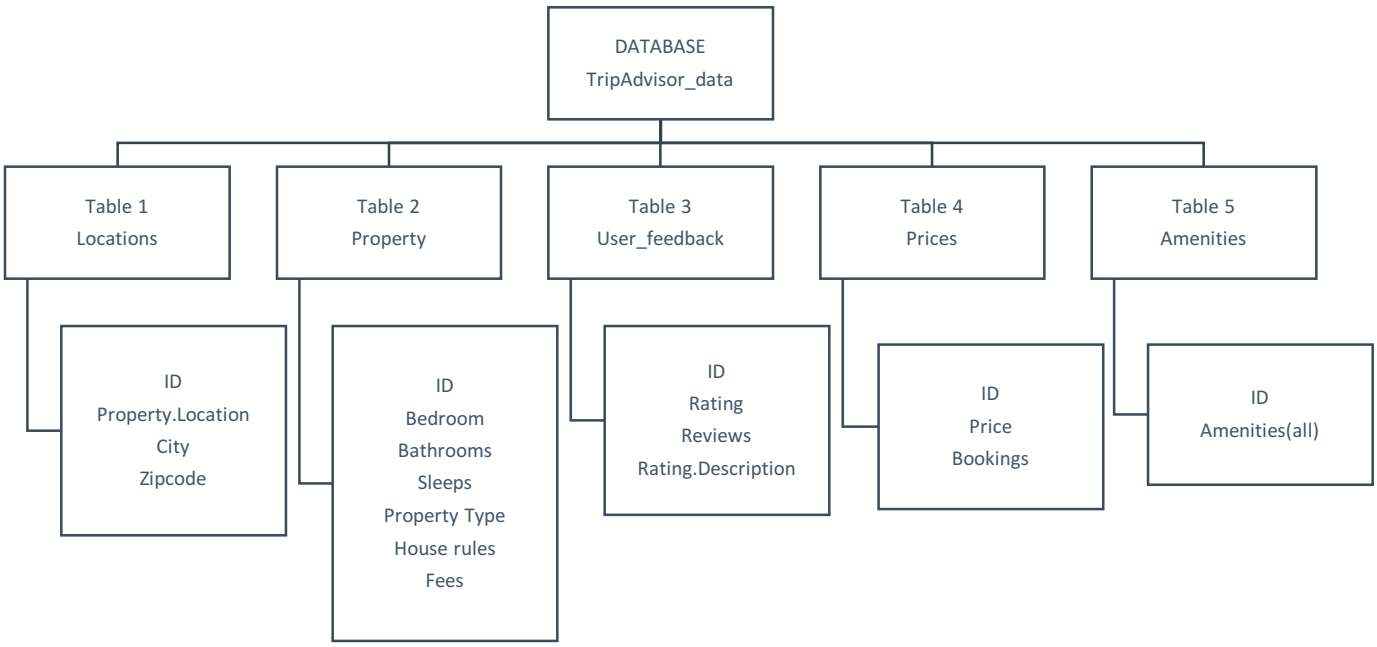
Thus, for our project we decided to use the RSQLite package available in R to write and read data from the SQLite server.

For Airbnb, the database schema designed and used is as shown below:



SCHEMA FOR AIRBNB

Similarly, for TripAdvisor the database schema designed and used is as shown below:



SCHEMA FOR TRIPADVISOR

Data stored in the SQL server had to be retrieved using SQL query. For getting different data columns, joins across different tables was required for which we used INNER JOIN to get data based on the matching ID values. The screenshot below shows an excerpt of the R code used to write data and retrieve data:

```
dbWriteTable(conn=database, name="Locations", value=location_data,row.names = FALSE, header = TRUE,overwrite=TRUE)
dbWriteTable(conn=database, name="Property", value=Property_data,row.names = FALSE, header = TRUE,overwrite=TRUE)
dbWriteTable(conn=database, name="User_feedback", value=Userfeedback,row.names = FALSE, header = TRUE,overwrite=TRUE)
dbWriteTable(conn=database, name="Prices", value=priceinfo,row.names = FALSE, header = TRUE,overwrite=TRUE)
dbWriteTable(conn=database, name="Amenities", value=amenitiesinfo,row.names = FALSE, header = TRUE,overwrite=TRUE)
dbWriteTable(conn=database, name="Host", value=Hostdata,row.names = FALSE, header = TRUE,overwrite=TRUE)
```

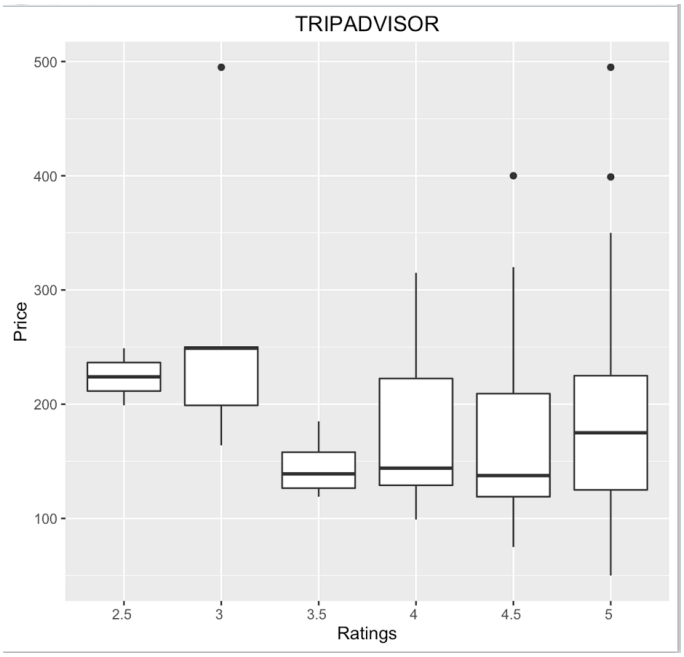
```
c<- dbGetQuery(database, "SELECT * FROM Locations
                INNER JOIN Property ON Property.ID = Locations.ID
                INNER JOIN Prices ON Prices.ID=Locations.ID
                INNER JOIN Amenities ON Amenities.ID=Locations.ID
                INNER JOIN User_feedback ON User_feedback.ID=Locations.ID")
```

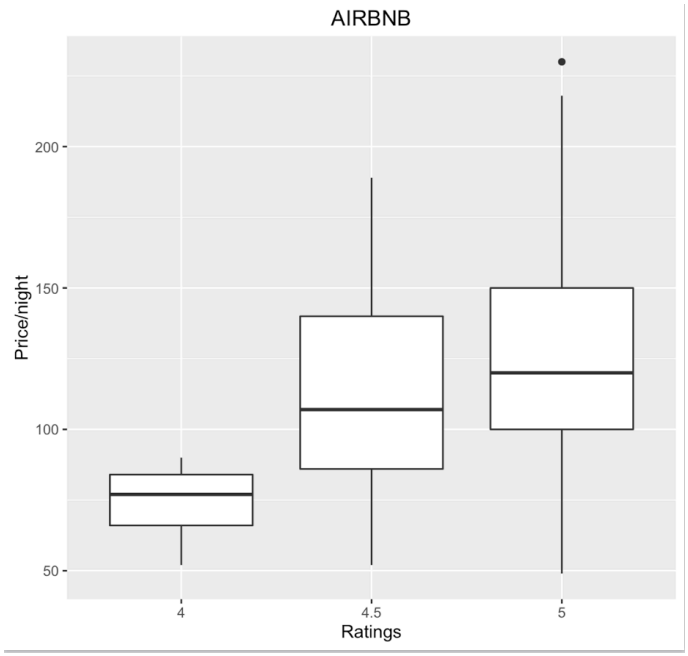
EXPLORATORY DATA ANALYSIS

For this comparative study, we wanted to answer the following questions:

4.1. Do high ratings mean high prices and vice versa?

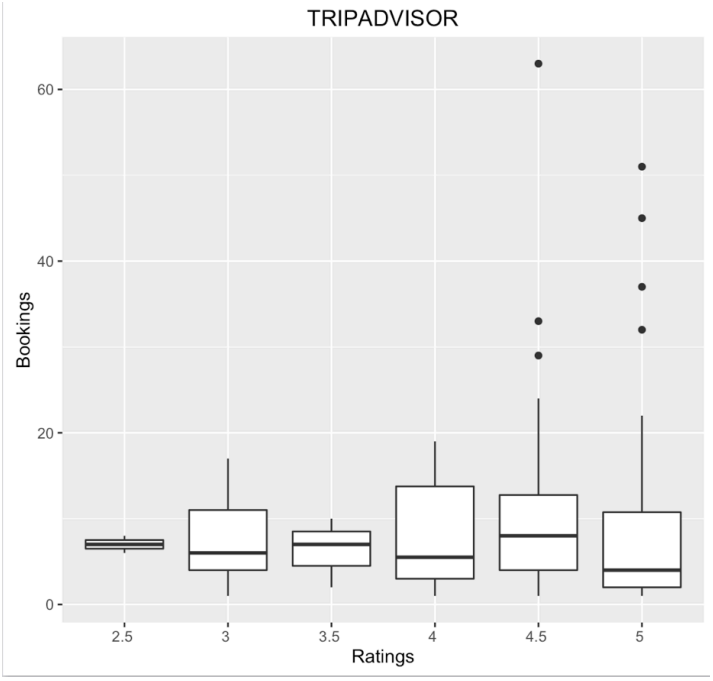
We see from the graphs below that for TripAdvisor the rentals that are rated high have a low price per night as compared to Airbnb where an opposite case is seen. For Airbnb, we see that the listings do not have ratings below a 4 star for our dataset. This may be biased and also vary with how big the dataset is. Our dataset has 306 records and in it we see that for listings with 4.5 stars and 5 stars rating the price per night is also high.
We also observe from the TripAdvisor graph that there are some few records with 5 stars rating whose price is really high. These records might indicate the properties in prime locations.





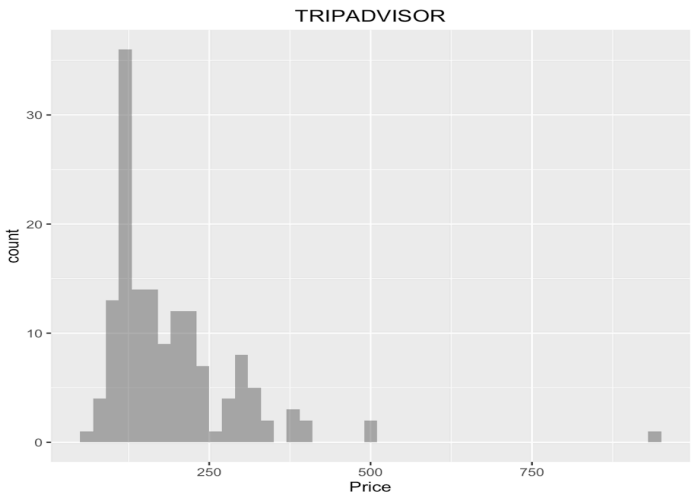
Concluding we see that listings with ratings above 4-star have **better value for money** than other for TripAdvisor data.

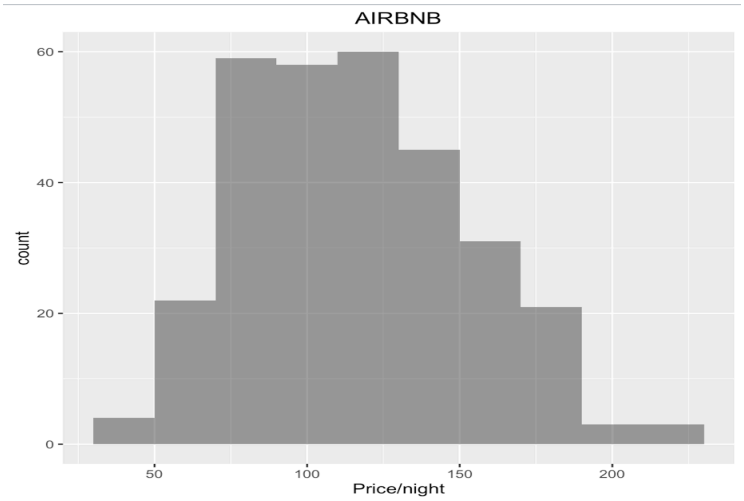
4.2. Do high ratings impact the number of bookings?



For TripAdvisor, we see that from the graph above that on an average 4 star rated rentals are booked the most. Also, the outliers of the box plot indicate that there are few 4.5 star and 5 star listings that are really popular and have high bookings.
We could not get the bookings data for Airbnb as it was not available on their website.
In conclusion, we see that there is **positive correlation** between ratings and bookings.

4.3. What is the general price trend across the dataset?

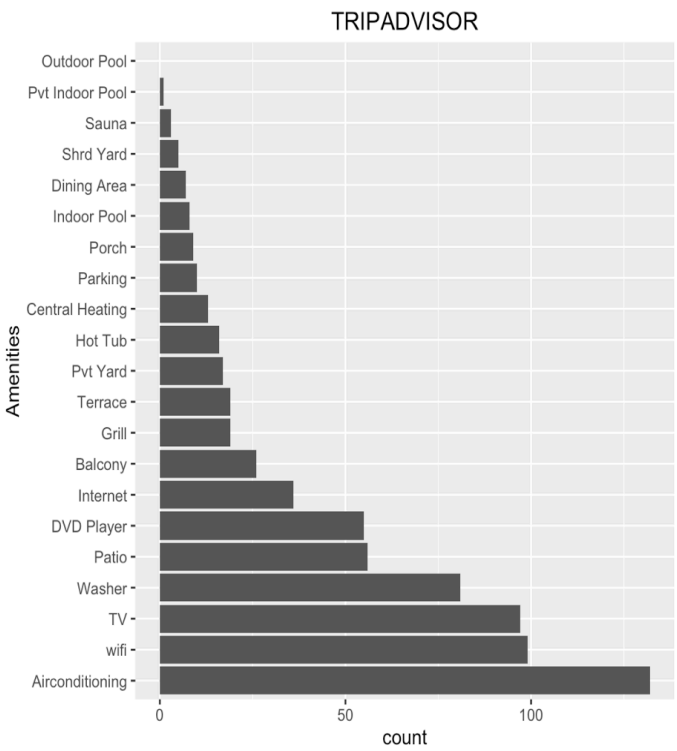


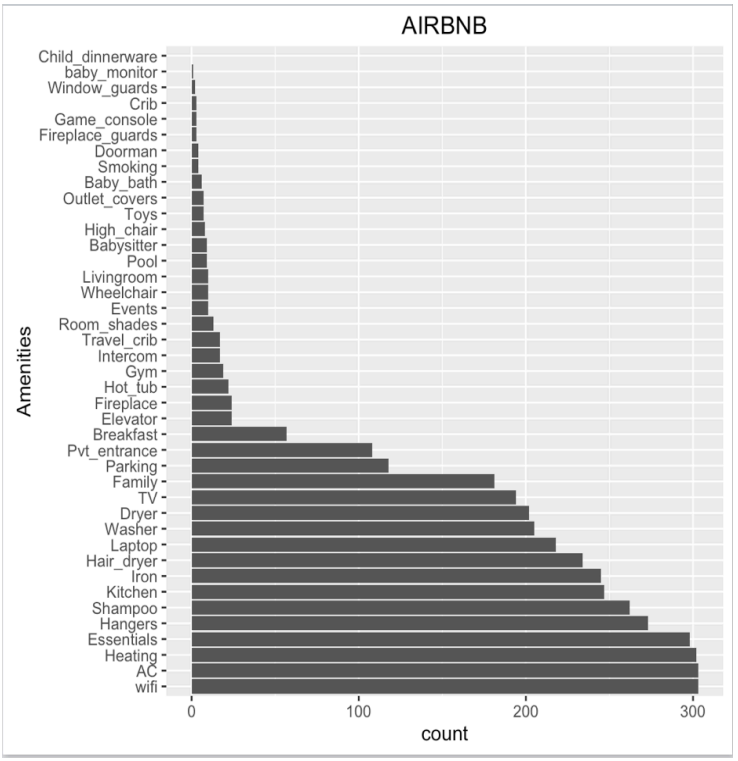


To study the general price trend across the dataset, we plotted a histogram for the Price/night. As can be seen from the graphs above, majority of listings for TripAdvisor are priced less that $375 but there are some which are priced at around $500/night and very few priced at around $1000/night.

The Airbnb dataset shows that all their listings are more economically priced as compared to TripAdvisor. Their highest priced listing is about $250/night.
Thus, it is seen that Airbnb provides more **budget friendly** options as compared to TripAdvisor.

4.4. What are the amenities provided? Good insight for Airbnb hosts?





For visualizing the top amenities that are most common across listings, we summarized the amenities columns and plot a histogram/bar chart for better understanding.

We see that for both Airbnb and TripAdvisor listings the top two amenities provided are Air conditioning, Wi-Fi, TV and Washer which is obvious as these are essentials that every customer looks for during their stay. We

also observe that Airbnb listings have a wider range of amenities as compared to TripAdvisor. These statistics validate the fact that Airbnb services have progressed over the years and its popularity has significantly increased in the recent times.

Looking at the amenities chart we also mention that this can be used as good insight for Airbnb hosts in determining what services to offer with their listing.

4.5. Which are amenities that are less common for a vacation home rental?

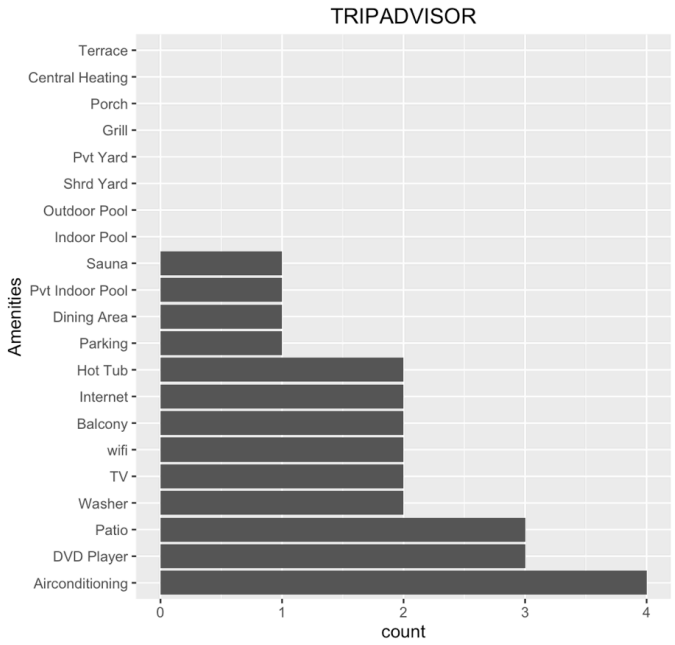
We also see from the amenities chart that that Parking is less common for TripAdvisor listings as compared to Airbnb. Also, amenities like swimming pool, hot tub, private yard are less common for both the services.

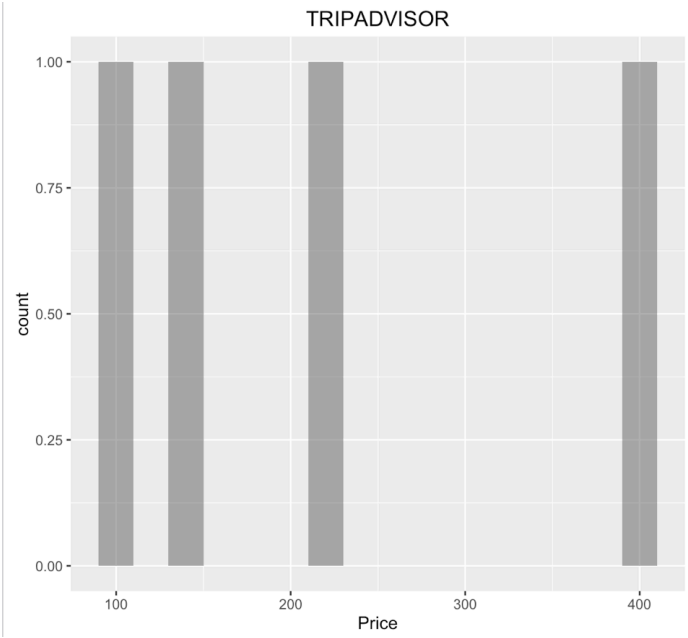
4.6. What features are offered by listings with high ratings and reviews?

We also wanted to analyze what are the characteristics of listings which have been given high ratings and which have a lot of reviews. So, for TripAdvisor, we queried to retrieve data for listings which have Ratings > 4 stars and Reviews > 20. The output was 4 listings which are shown below:

TRIPADVISOR

	Property.Location	City	Zipcode	Bedrooms	Bathrooms	Sleeps	Property.type	Price	Bookings	Ratings	Reviews	Rating.Descriptio
1	Capitol Hill	Washington DC	20003	1	1	4	Rental Home	220	19	5.0	30	Excellent
2	Penn Quarter, Upper Northwest	Washington DC	20004	2	2	5	Condo/Apartment	148	1	4.5	61	Very good
3	*NA*	Washington DC	20007	5	4	12	Rental Home	399	9	5.0	23	Excellent
4	Upper Northwest	Washington DC	20005	2	2	5	Condo/Apartment	109	9	4.5	88	Very good





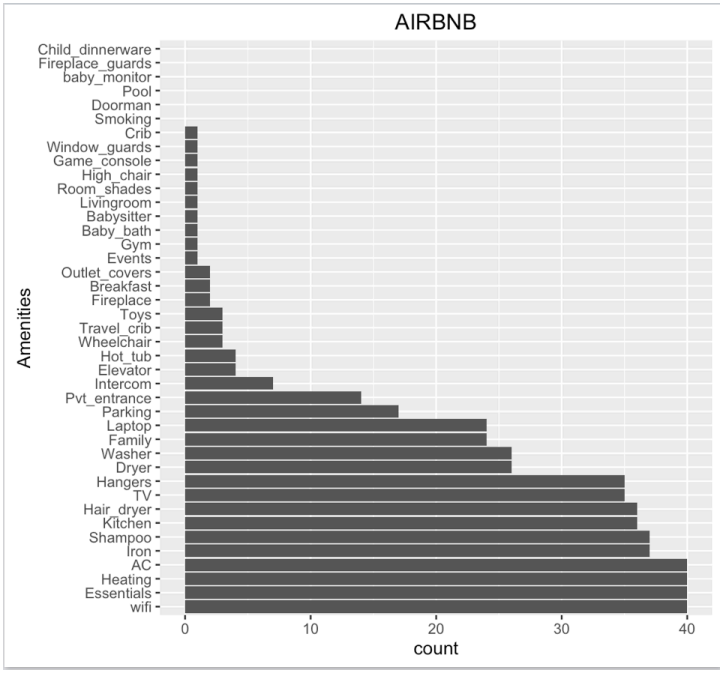
It was observed that for the above obtained listings two of them were rental homes which had 5 stars rating and two were condos which had 4.5 stars rating.

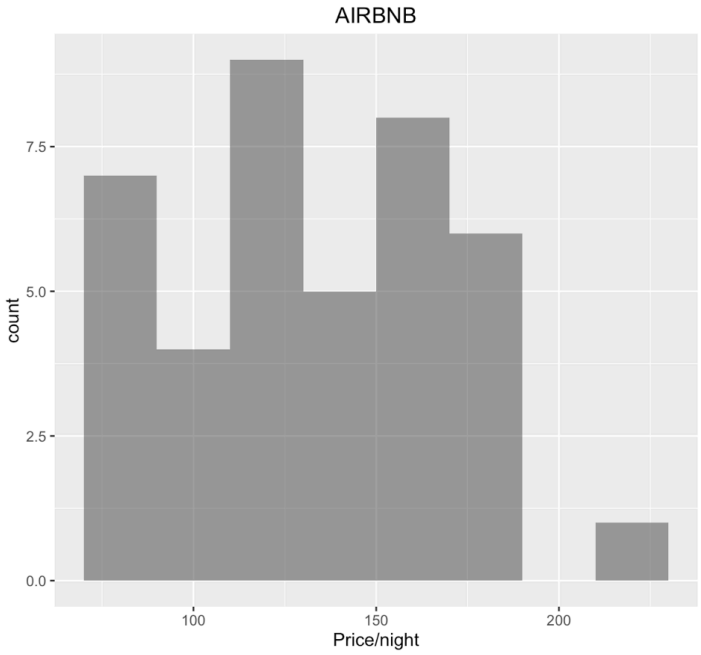
Also, the first 5 stars listing on Capitol Hill had high number of bookings which can also be contributed to its prime location. It is also seen that these listings have a **higher price range** compared to the average price across the dataset.

For Airbnb, a similar query was fired for getting listings with Ratings =5 and Reviews >100. the output from that query is shown below:

AIRBNB

	Bedrooms	Bathrooms	Accommodates	Beds	Property.Type	Room.Type	Pricepernight	Total.Price	Price.Currency	Amenities	Ratings	Reviews	Host.Response.rate	Host.Reviews
1	1	1.0	3	2	Apartment;	Entire home/apt	171	440	USD	Free parking on premises; Family/kid friendly; Internet...	5	224	NA	241
2	1	1.0	2	1	Apartment	Entire home/apt	133	342	USD	Internet; Kitchen; Wireless Internet; Cable TV; Laptop f...	5	147	Response rate: 100%	5
3	1	1.0	4	2	House;	Entire home/apt	177	456	USD	Pets allowed; Family/kid friendly; Internet; Kitchen; Wir...	5	111	NA	13
4	1	1.0	4	2	Apartment	Entire home/apt	173	446	USD	Elevator; Internet; Kitchen; Buzzer/wireless intercom; ...	5	125	Response rate: 100%	77
5	1	1.0	2	1	House	Private room	113	290	USD	Pets allowed; Family/kid friendly; Kitchen; Wireless Int...	5	121	Response rate: 100%	398
6	1	1.0	2	1	House	Private room	85	219	USD	Family/kid friendly; Internet; Kitchen; Wireless Internet...	5	255	NA	33
7	1	1.0	2	1	House;	Private room	74	190	USD	Free parking on premises; Internet; Buzzer/wireless int...	5	111	NA	80
8	1	1.0	2	1	Apartment;	Entire home/apt	115	297	USD	Family/kid friendly; Kitchen; Wireless Internet; Cable T...	5	291	NA	51
9	1	1.5	2	1	House	Private room	90	232	USD	Free parking on premises; Kitchen; Wireless Internet; C...	5	194	NA	472
10	1	1.0	2	1	Condominium;	Private room	104	267	USD	Free parking on premises; Family/kid friendly; Internet...	5	182	Response rate: 85%	203
11	1	1.0	2	1	House	Private room	85	218	USD	Internet; Kitchen; Wireless Internet; Cable TV; Dryer; T...	5	275	NA	398
12	1	1.0	2	1	House	Entire home/apt	97	250	USD	Family/kid friendly; Internet; Wireless Internet; Cable T...	5	142	Response rate: 100%	33
13	1	1.0	2	1	Apartment;	Entire home/apt	160	412	USD	Elevator; Internet; Kitchen; Buzzer/wireless intercom; ...	5	131	Response rate: 100%	30
14	1	1.0	2	1	Apartment	Entire home/apt	119	308	USD	Pets allowed; Internet; Wheelchair accessible; Kitchen; ...	5	150	Response rate: 100%	597
15	1	1.0	4	2	House	Entire home/apt	164	423	USD	Family/kid friendly; Internet; Kitchen; Wireless Internet...	5	122	NA	6
16	1	1.0	4	2	Apartment;	Entire home/apt	172	443	USD	Family/kid friendly; Internet; Kitchen; Wireless Internet...	5	146	NA	38
17	1	1.5	2	1	House	Private room	72	185	USD	Free parking on premises; Internet; Kitchen; Wireless I...	5	120	Response rate: 100%	72
18	1	1.0	2	1	Apartment;	Entire home/apt	154	396	USD	Family/kid friendly; Internet; Kitchen; Wireless Internet...	5	142	Response rate: 94%	147
19	1	1.0	4	2	Apartment;	Entire home/apt	155	400	USD	Pets allowed; Elevator; Free parking on premises; Inter...	5	109	NA	9
20	1	1.0	4	2	Apartment;	Entire home/apt	117	301	USD	Free parking on premises; Internet; Kitchen; Wireless I...	5	176	NA	358
21	1	1.0	2	1	House;	Entire home/apt	119	308	USD	Free parking on premises; Family/kid friendly; Internet...	5	133	NA	16
22	1	1.0	2	1	House	Private room	98	253	USD	Internet; Kitchen; Wireless Internet; Cable TV; Dryer; T...	5	192	NA	159
23	1	1.0	2	1	Apartment;	Private room	112	289	USD	Family/kid friendly; Internet; Kitchen; Wireless Internet...	5	104	NA	17
24	2	1.5	3	1	House	Private room	83	215	USD	Free parking on premises; Internet; Kitchen; Wireless I...	5	132	Response rate: 100%	156
25	1	1.0	2	1	Townhouse;	Entire home/apt	148	381	USD	Family/kid friendly; Internet; Wheelchair accessible; Kit...	5	138	NA	194
26	1	1.0	4	2	Apartment	Entire home/apt	215	555	USD	Free parking on premises; Family/kid friendly; Internet...	5	153	NA	15
27	1	2.0	5	2	Apartment;	Entire home/apt	161	414	USD	Internet; Kitchen; Wireless Internet; Cable TV; Laptop f...	5	148	NA	22
28	1	1.0	4	2	House;	Entire home/apt	183	471	USD	Family/kid friendly; Internet; Kitchen; Wireless Internet...	5	121	Response rate: 100%	122
29	1	1.0	2	1	Guesthouse;	Entire home/apt	158	406	USD	Family/kid friendly; Internet; Kitchen; Buzzer/wireless ...	5	108	NA	110
30	1	1.0	3	2	House;	Entire home/apt	135	349	USD	Family/kid friendly; Internet; Hot tub; Kitchen; Suitable...	5	106	Response rate: 100%	60
31	1	1.0	4	1	Apartment;	Entire home/apt	177	456	USD	Pets allowed; Elevator; Family/kid friendly; Internet; Kit...	5	114	NA	0
32	2	1.0	3	1	Apartment;	Entire home/apt	138	354	USD	Free parking on premises; Family/kid friendly; Internet...	5	104	NA	541
33	1	1.0	3	2	Apartment;	Entire home/apt	169	436	USD	Free parking on premises; Family/kid friendly; Internet...	5	124	NA	10
34	1	1.0	2	1	House	Private room	72	186	USD	Internet; Kitchen; Wireless Internet; Iron; Hangers; TV; ...	5	119	NA	74





40 records for Airbnb listings were obtained from the above-mentioned query which is much greater in number than TripAdvisor and is obvious as the Airbnb dataset is double that of TripAdvisor. It is seen that maximum number of listings with high ratings and high reviews have their price around $125-$170 which is also the average price across the dataset. Thus, overall it is seen that **Airbnb listings have better feedback and user experience as compared to TripAdvisor.**

7. What factors affect price?

Assuming the differences in price is caused by the amenities difference between Airbnb and TripAdvisor service. Based on this assumption, a random forest model was built, to check what features affects hotel price and Airbnb price most.

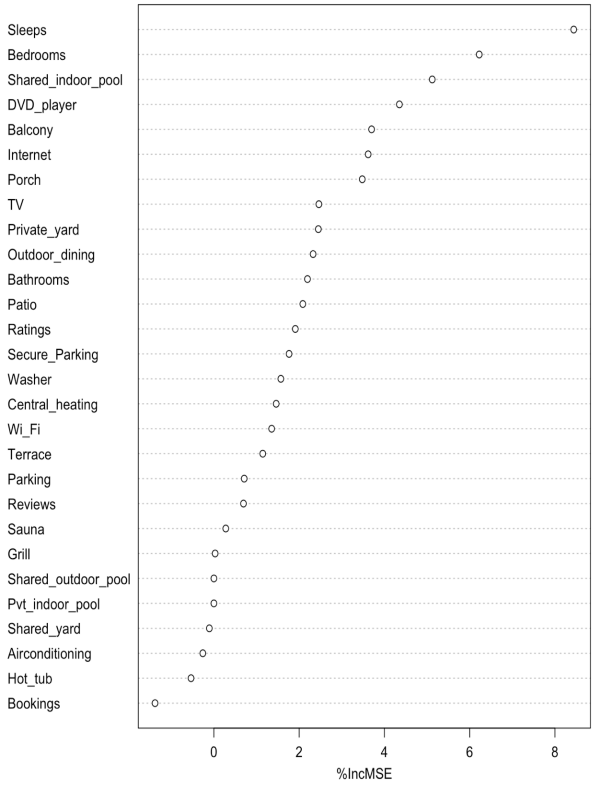
A pre-requisite in building the random forest model was to remove all NA's in the dataset for Ratings column and Bookings column (for TripAdvisor data). In order to fill all NA values, we built a prediction model to best predict the Ratings and Bookings using the ANOVA analysis method. The R code for the TripAdvisor prediction model is shown in the screenshot below.

```
tripadvisor$Ratings<- as.numeric(tripadvisor$Ratings)
Ratingfit<- rpart(Ratings~ .,
                  data=c1[!is.na(c1$Ratings),],
                  method="anova")
c1$Ratings[is.na(c1$Ratings)] <- predict(Ratingfit, c1[is.na(c1$Ratings),])
Bookingfit<-rpart(Bookings~ .,
                  data=c1[!is.na(c1$Bookings),],
                  method="anova")
c1$Bookings[is.na(c1$Bookings)] <- predict(Bookingfit, c1[is.na(c1$Bookings),])
```

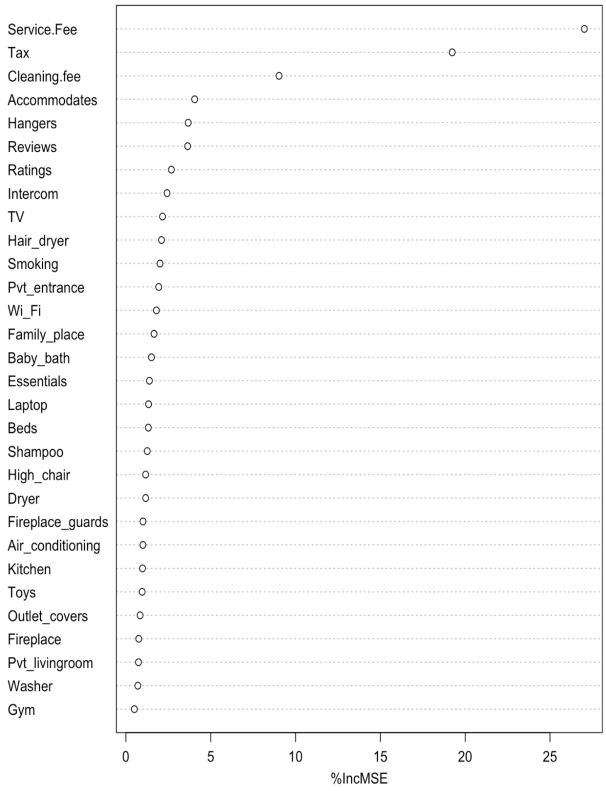
After filling in the NA values with the predicted values, a random forest model was generated considering all the numeric variables like bedrooms, bathrooms, accommodates, ratings, reviews, all amenities and bookings. The R code for the random forest model is shown in the screenshot below:

```
set.seed(415)
`Factors affecting Price` <- randomForest(Price~ .,
                                          data=c1,
                                          importance=TRUE, proximity=TRUE,
                                          ntree=200)
varImpPlot(`Factors affecting Price`)
```

TRIPADVISOR



AIRBNB



Though this model is not completely reliable as we do not have the location data but hosts could get an idea about what price to set for their listing if it has certain features.

It is seen from the result above that the top 7 factors affecting the price of a listing are:

TRIPADVISOR	AIRBNB
Accommodates	Service Fee
Bedrooms	Tax
Indoor Pool	Cleaning fee
DVD Player	Accommodates
Balcony	Hangers
Internet	Reviews
Porch	Ratings

LEARNINGS AND FUTURE WORK

1. Data transformation and manipulation required most amount of efforts which we tried to accomplish using the concepts taught in class.
2. A lot of debugging was required in the process of coding for this project for which we utilized stack overflow and other aid on the internet.
3. The scope of project can be extended by doing season wise analysis (i.e. how the prices vary per different seasons of the year).
4. Prices of houses also majorly depend on the crime rate for a particular geographic location and thus crime data can be included as part of the analysis.
5. Another prospect of future work is creating a recommendation model for setting price based on analysis.