

Tourist pattern analysis and hotel recommendation

Submitted By:

Namratha Channabasavaiah

Nikanshi Yadav

Rajat Jain

Varun Muralidharan

Shivani Ramdas

TABLE OF CONTENTS

Introduction	2
About Expedia	2
A. Services	2
B. Global reach and customer base	2
Data Analysis	3
A. Types of Data	3
B. Choosing the dataset	4
C. Data Cleansing	5
D. Challenges faced with the data analysis	5
E. Descriptive Analysis	6
Problem Statement	7
Algorithms used for Data Analysis	7
A. Frequent Itemset Mining	9
B. To predict the seasonal booking traffic for each country	14
C. Importance of Marketing Channel	19
D. Predict the purchase of flight packages along with a hotel booking based on user search data.	25
Output & Analysis	26
For the year 2013	26
Recommendations	30
Appendix	31
References	50

Introduction

The data set chosen by us was taken from Kaggle and is a random selection from Expedia and is not representative of the overall statistics. The data collection process for Expedia is through their online website (Assumption). The data is collected basis the time stamp and the source (marketing channel) through which the users have reached the website.

About Expedia

Expedia: “Bringing the world within the reach”

The Expedia group is the world’s travel platform that allows its customers to choose a travel plan ranging from modest to luxury. Collectively it permits researching, planning, booking airline tickets, from choosing which hotels to check into, to letting one plan what they can do in that destination when they arrive.

A. Services

It is one of the world’s leading full-service online travel brands helping travelers choose from its wide range of services. Keeping its primary focus on services like the widest selection of vacation packages, flights, and hotels, it ensures that the customer experience is enhanced by its other services like easy booking of rental cars, rail, cruises, browsing activities and major attractions in and around the destination.

B. Global reach and customer base

According to its website, Expedia has a global presence with 200+ booking sites in 75 countries and 150+ mobile websites in nearly 70 countries and 35 languages with an international revenue of 45%. Having such a vast reach, it aims at targeting a wide variety of customers narrowed down to three sectors: Business, Vacation, and Hot Spots.

The **business** sector targets all sizes of business. It encourages the business or its traveling employees to become a member allowing them to accumulate points. As a strategy, Expedia understands that the customers in this sector have high value for time and money and hence extends offers of cheaper

and faster flights and accommodation. They also make use of targeted channels to attract these customers and give an advantage during deals. With everything being digitized, it allows to check and get the boarding pass directly through its application.

The **vacation** sector is the huge in size owing to the extensive travel plans made by the families. Hence this sector is focussed towards leisure and luxury. Keeping in mind that the families would like to spend more time in their destination city, layovers are kept minimal and cheaper deals for bulk travel are provided. Along with booking flights, Expedia also permits these families an easy access to the hotels promoting a one-stop access to all the facilities. To target this sector, Expedia tracks the activities of the users on other websites and attracts people already looking for deals onto their website. They then email these customers.

The **Hot Spots** sector, **on the other hand**, includes events hosted around the world. Expedia allows its customers to get special discounted and inclusive packages. According to The Squad, an article published on September 15, 2016, stated that "Expedia has partnered with the Olympics to give out huge bundles and deals to travelers heading to the games." It continues to say that "And by using banners that use phrases like "You're Invited," they are pulling on the notion that people hate to miss out."

Data Analysis

A. Types of Data

We were provided with the logs of customer behavior about the hotel booking data across multiple continents, which was further subdivided at countries, city and region level. These include what customers searched for, how they interacted with search results (click/book), whether or not the search result was a travel package.

Expedia has provided three datasets that are split based on time :

1. train.csv - the training dataset contains information from the year 2013 and 2014. It includes logs on all the users and their clicks / booking events.
2. test.csv - the test dataset contains only the booking information from the year 2015.
3. destinations.csv - consists of features extracted from hotel reviews text.

Column name	Description	Data type
date_time	Timestamp	string
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent	ID of continent associated with site_name	int
user_location_country	The ID of the country the customer is located	int
user_location_region	The ID of the region the customer is located	int
user_location_city	The ID of the city the customer is located	int
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	double
user_id	ID of user	int
is_mobile	1 when a user connected from a mobile device, 0 otherwise	tiny int
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	int
channel	ID of a marketing channel	int
srch_ci	Check-in date	string
srch_co	Checkout date	string
srch_adults_cnt	The number of adults specified in the hotel room	int
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room	int
srch_rm_cnt	The number of hotel rooms specified in the search	int
srch_destination_id	ID of the destination where the hotel search was performed	int
srch_destination_type_id	Type of destination	int
hotel_continent	Hotel continent	int
hotel_country	Hotel country	int
hotel_market	Hotel market	int
is_booking	1 if a booking, 0 if a click	tiny int
cnt	Number of similar events in the context of the same user session	big int
hotel_cluster	ID of a hotel cluster	int

B. Choosing the dataset

Although Kaggle provided training and test dataset separately, we decided to work with training dataset only. The reason for this is the size of the training dataset. With 24 variables, 37 million records and 4GB in size, it is large enough to split this dataset into training and validation sets according to the problem statement and still obtain effective results.

C. Data Cleansing

We audited the dataset with the use of statistical methods with the goal of identifying anomalies and contradictions. Later, we carried out exploratory data analysis on the data set to identify possible relationships between the variables and how they impact the outcome, which is the booking status. We intend to use only those variable that would make it feasible for us to derive solutions to our problem statements and benefit further analysis.

The data was presented in raw form which required necessary extraction of information like the week, month, count of booking etc. Multiple transformations of columns were conducted to convert into time series data, log transform to aid the smooth flow of processing. Each model required specialized transformation. The trend analysis required melting and extracting the count of bookings and channel booking which gave us an insight into the best medium for advertisements and the expected trend in the upcoming months for the top 3 destinations. Imputation of the data was carried out and certain data fields which had missing values were omitted out of the dataset.

D. Challenges faced with the data analysis

The data we received was above 4 GB in size. Due to this we initially faced difficulties to load the dataset onto our machines and processing any algorithm on this data would consume a large amount of time. For the purpose of the initial study, we only imported about 1000 rows. On receiving help from the UT Dallas Tech support team, we were able to load the entire dataset and run exploratory analysis/ refine the dataset for our predictive algorithms.

One of the variables in the dataset “orig_destination_data” which depicts the distance between the user and the destination they wish to travel had a lot of null values.

Most of the columns are integers or floats, so we can't do a lot of feature engineering. For example, user_location_country isn't the name of a country, it's an integer. This makes it harder to create new features because we don't know exactly which each value means.

E. Descriptive Analysis

	vars	n	mean	sd	min	max	range	se
site_name	1	3000693	9.37	11.92	2.00	53.00	51.00	0.01
posa_continent	2	3000693	2.71	0.73	0.00	4.00	4.00	0.00
user_location_country	3	3000693	87.62	59.02	0.00	239.00	239.00	0.03
user_location_region	4	3000693	312.91	204.50	0.00	1027.00	1027.00	0.12
user_location_city	5	3000693	27880.12	16731.00	0.00	56507.00	56507.00	9.66
orig_destination_distance	6	1985514	1688.68	2157.34	0.01	12199.17	12199.16	1.53
user_id	7	3000693	606620.88	349317.42	5.00	1198784.00	1198779.00	201.66
is_mobile	8	3000693	0.10	0.30	0.00	1.00	1.00	0.00
is_package	9	3000693	0.14	0.34	0.00	1.00	1.00	0.00
channel	10	3000693	6.18	3.61	0.00	10.00	10.00	0.00
srch_adults_cnt	11	3000693	1.87	0.91	0.00	9.00	9.00	0.00
srch_children_cnt	12	3000693	0.28	0.66	0.00	9.00	9.00	0.00
srch_rm_cnt	13	3000693	1.13	0.48	0.00	8.00	8.00	0.00
srch_destination_id	14	3000693	15379.85	11639.09	1.00	65104.00	65103.00	6.72
srch_destination_type_id	15	3000693	2.88	2.21	0.00	9.00	9.00	0.00
is_booking	16	3000693	1.00	0.00	1.00	1.00	0.00	0.00
hotel_continent	17	3000693	3.01	1.62	0.00	6.00	6.00	0.00
hotel_country	18	3000693	80.54	54.91	0.00	212.00	212.00	0.03
hotel_market	19	3000693	621.26	492.19	0.00	2117.00	2117.00	0.28
hotel_cluster	20	3000693	47.69	29.04	0.00	99.00	99.00	0.02

Problem Statement

The Expedia dataset contains their customer's travel information like what they searched for, whether or not the search result was a travel package, the check in and check out date. After analyzing this data, a set of problem statements are identified which will give an overview of the customer's behavior and requirements.

Given the complexity of the dataset, we could come up with more than one business case:

1. To contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups.
2. Predict the purchase of flight packages along with a hotel booking based on user search data.
3. Identifying & Predicting Growth Across the Marketing Channels for Expedia for their Hotel Booking Service.
4. To predict the seasonal booking traffic for each country.

Algorithms used for Data Analysis

Time Series (ARIMA):

Autoregressive Moving Average algorithm is the predominant time series model used for identifying the short-term trend of variables based on date or date-time. The algorithm is built under the package "forecast" and works efficiently for relatively stable past trend and lesser outliers.

Time series data analysis means analyzing the available data to find out the pattern or trend in the data to predict some future values which will, in turn, help more effective and optimize business decisions. Trend analysis was a significant business case in this project. We utilized the ARIMA algorithm for the following problem statements:

- Identifying & Predicting Growth Across the Marketing Channels for Expedia for their Hotel Booking Service.
- To predict the seasonal booking traffic for each country.

ARIMA algorithm has been performed for time series analysis where we identified trends across channels and the top 3 destinations for Hotel Bookings in 2014 and 2013.

ARIMA models are a popular and flexible class of forecasting model which utilize historical data to make predictions. This model is a forecasting that can be used as a foundation for more complex models.

Random Forest Classifier:

Random Forest is a supervised learning algorithm. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Random Forests is a learning method for classification based on generating a large number of decision trees. A large number of decision trees are created and every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model. Averaging the Trees helps us to reduce the variance and also improve the Performance of Decision Trees on Test Set and eventually avoid Overfitting. The idea is to build lots of Trees in such a way to make the Correlation between the Trees smaller.

In our case we had a binary classification, the logistic model seemed to be convincing, but the huge number of factor variables in each column will definitely choke the logistic model and cause resource constraint. Therefore, we decided to go on with random forest classifier and condense it into a binary classifier.

Apriori:

Apriori algorithm is a association rule mining classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It runs based on the support and confidence value which are the number of appearance of items divided by the total number of transactions and the ratio of

the number of joint occurrence of the antecedent and consequent divided by the number of occurrence of the antecedent.

One of our business cases was to identify the preferred destination type and hotel cluster which frequent customers preferred. We had the data for each trip identified by the user id. Each record was condensed into transactions to make it suitable for apriori algorithm. The algorithm was expected to produce the frequently visited destination type and hotel clusters and suggest probable destination type and hotel cluster in next visit for recommendation for similar customers.

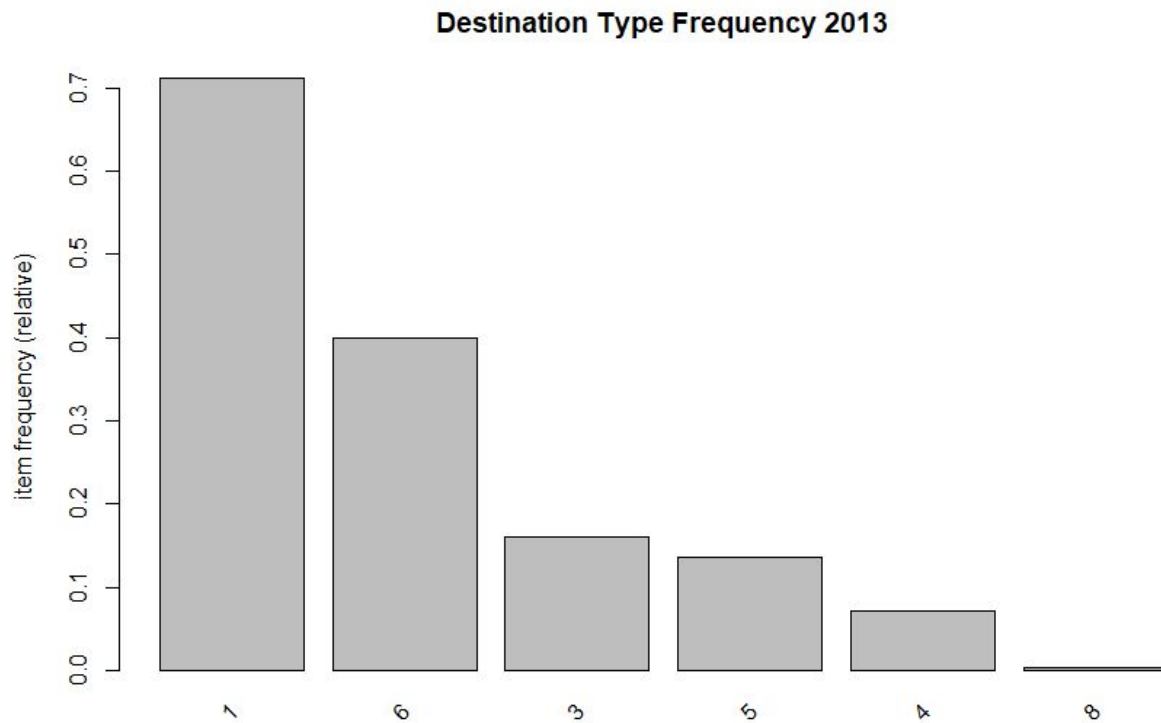
A.Frequent Itemset Mining

In recent years, the travel industry is growing rapidly. There are numerous companies entering the market. With this kind of competition, to stay on top of the market, it is necessary for the company to understand their customer's requirements and attract them with deals and discounts.

The first step to retaining the existing customers and attracting new one is by analyzing the customer's behavior. With Expedia's dataset, we would like to study the preference of each customer and their travel history and accordingly predict their likelihood of traveling to a particular destination type. Using this information, we can recommend different destinations that will be of interest to the customers. The data consist of 2013 and 2014 data, we initially separated the data into 2013 and 2014, performing the analysis individually.

To analyze the travel history of the customers, only the booking data from the dataset is separated. The data contains information about the same places the user has visited multiple times. To remove this redundancy, the ID of the user was considered to remove redundant records of the ID of the destination they visited. This data was then converted into transactions.

The below figure represents the item-frequency plot for the destination types. We can infer that destination type 1 is the most preferred relative to other destination types, whereas destination type 8 is the least preferred. The transactions were then fed to association rule-based algorithm (APRIORI).



The purpose of APRIORI is to generate association rules using the past transactions and sequence mining of visiting a destination type. The rules depict the probable next destination type that most users will prefer.

	lhs	rhs	support	confidence	lift	count
[1]	{1,3,4}	=> {5}	0.004311212	0.3882954	2.841126	1672
[2]	{1,4,5}	=> {3}	0.004311212	0.4381551	2.731039	1672
[3]	{3,4}	=> {5}	0.004832064	0.3342251	2.445498	1874
[4]	{1,3,6}	=> {5}	0.017752807	0.3330270	2.436732	6885
[5]	{1,4,6}	=> {5}	0.007258410	0.3306319	2.419207	2815
[6]	{4,5}	=> {3}	0.004832064	0.3818256	2.379934	1874
[7]	{1,4,6}	=> {3}	0.008238231	0.3752643	2.339037	3195
[8]	{1,5,6}	=> {3}	0.017752807	0.3723836	2.321082	6885
[9]	{3,6}	=> {5}	0.020431843	0.2879360	2.106805	7924
[10]	{4,6}	=> {5}	0.008550226	0.2781413	2.035137	3316
[11]	{5,6}	=> {3}	0.020431843	0.3227830	2.011919	7924
[12]	{4,6}	=> {3}	0.009674442	0.3147123	1.961614	3752
[13]	{1,3}	=> {5}	0.024018503	0.2571429	1.881494	9315
[14]	{1,3,4}	=> {6}	0.008238231	0.7419879	1.856375	3195
[15]	{1,4}	=> {5}	0.009839464	0.2532351	1.852901	3816
[16]	{1,3,5}	=> {6}	0.017752807	0.7391304	1.849226	6885
[17]	{1,4,5}	=> {6}	0.007258410	0.7376834	1.845605	2815
[18]	{1,5}	=> {3}	0.024018503	0.2912576	1.815420	9315
[19]	{1,4}	=> {3}	0.011102917	0.2857522	1.781105	4306
[20]	{3,5}	=> {6}	0.020431843	0.6829857	1.708757	7924

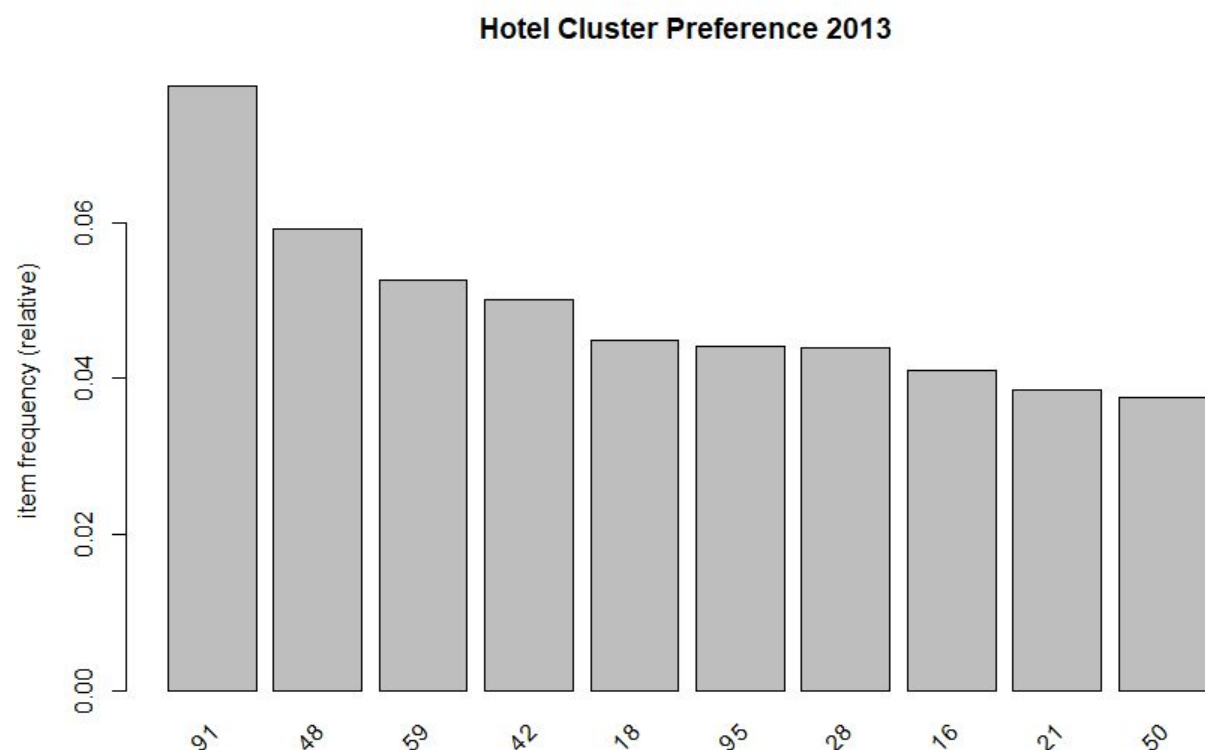
We have taken the top 20 rules (2013) based in the on the lift ratio, where the minimum support and confidence is 0.004 and 0.2 respectively.

For eg: The first rule generated explains that a user who has visited destination type 1,3 and 4 is likely to visit destination type 5. Likewise the second tells that a user who has visited destination type 1,4 and 5 is likely to visit destination type 3. The LHS represents the past bookings and the represents the most probable destination which may be preferred by customers.

Limiting to destination type was not sufficient, we also needed to find out the most probable preference of hotel clusters. The hotel cluster consists of similar hotels grouped into 108 clusters.

Hotel recommendation based on suggested rules is imperative and is highly likely to yield a better turnover. Therefore it would be necessary to obtain the most probable hotel that customers would prefer.

Apriori model was again useful when it comes to frequent pattern mining. To analyze the preference of hotel pattern we considered the past bookings for each and identify each unique hotel cluster visited and convert this into transactions.



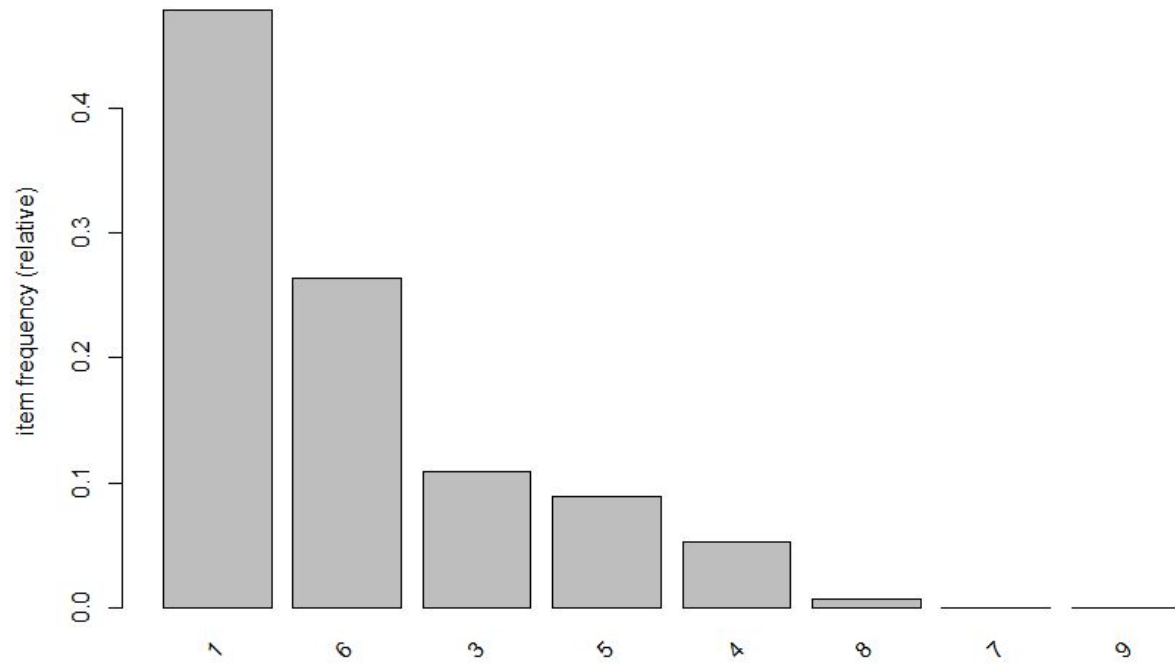
The figure describes the frequency of top 10 clusters, we can infer that cluster number 91 has the highest relative frequency. The transaction was applied to the APRIORI algorithm and the frequent itemsets were generated. We again consider the top 20 association rules with minimum support of 0.001 and confidence of 0.2.

	lhs	rhs	support	confidence	lift	count
[1]	{46, 97}	=> {64}	0.001077803	0.4523810	13.886742	418
[2]	{47, 50}	=> {39}	0.001095852	0.2428571	12.548137	425
[3]	{39, 50}	=> {47}	0.001095852	0.3509496	11.231836	425
[4]	{46, 64}	=> {97}	0.001077803	0.2084788	10.531914	418
[5]	{32, 50}	=> {47}	0.001266032	0.3262458	10.441213	491
[6]	{47, 50}	=> {32}	0.001266032	0.2805714	10.396799	491
[7]	{64, 82}	=> {46}	0.001095852	0.3480753	10.322908	425
[8]	{32, 47}	=> {50}	0.001266032	0.3869188	10.291282	491
[9]	{39, 47}	=> {50}	0.001095852	0.3811659	10.138266	425
[10]	{46, 58}	=> {64}	0.001137108	0.3252212	9.983319	441
[11]	{25, 46}	=> {64}	0.001010763	0.3236994	9.936604	392
[12]	{46, 64}	=> {58}	0.001137108	0.2199501	9.510801	441
[13]	{58, 64}	=> {46}	0.001137108	0.3181818	9.436353	441
[14]	{47, 48}	=> {7}	0.001106166	0.2289221	9.397898	429
[15]	{16, 48, 91}	=> {42}	0.001150000	0.4689800	9.358510	446
[16]	{28, 48, 91}	=> {42}	0.001080381	0.4655556	9.290175	419
[17]	{25, 64}	=> {46}	0.001010763	0.3125997	9.270802	392
[18]	{59, 82}	=> {29}	0.001206727	0.2494670	9.207249	468
[19]	{16, 42, 91}	=> {48}	0.001150000	0.5360577	9.076891	446
[20]	{29, 59}	=> {82}	0.001206727	0.3383948	9.055914	468

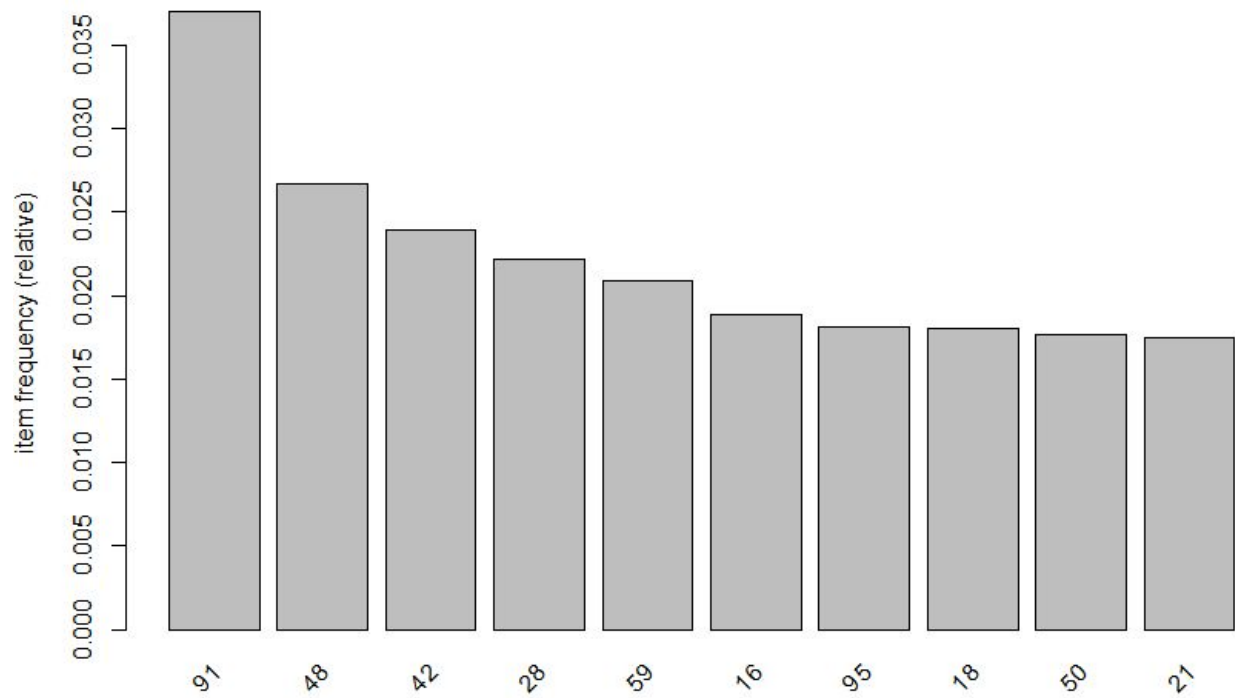
The figure depicts the association rules for 2013 hotel preference, LHS represents the past hotels booked and the RHS represents the probable hotel cluster which most users may prefer in the next vacation.

The same analysis is performed for year 2014, but there were very few data with multiple travels per user, therefore the minimum confidence and significance were required to be set to an extremely low value. This proved unworthy of the rules since there was no significant number of multiple travel pattern by customers.

Destination Type Frequency 2014



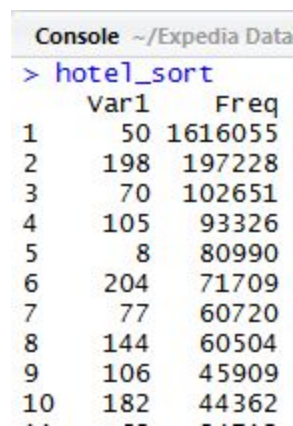
Hotel Cluster Preference 2014



B. To predict the seasonal booking traffic for each country

Another business case that is relevant and important to the travel industry is predicting the expected number of customers in the future. In this problem statement, we want to find the number of bookings each hotel will be anticipating in the upcoming months. However, since the dataset does not contain information about each hotel, the next relevant variable 'hotel_country' is selected.

First, the booking data is separated from the dataset. The frequency of each of the hotel country that had a booking is calculated. The top three countries most frequently visited by the users are considered for this problem statement. The reason is that there are approximately 2000 different hotel countries in the dataset and we do not have the time or resources to run the algorithm for all of them.



	Var1	Freq
1	50	1616055
2	198	197228
3	70	102651
4	105	93326
5	8	80990
6	204	71709
7	77	60720
8	144	60504
9	106	45909
10	182	44362

The table above shows the hotel country and the total number of bookings they had in the year 2013 and 2014. The top 3 countries: 50, 198 and 70 are selected for further analysis.

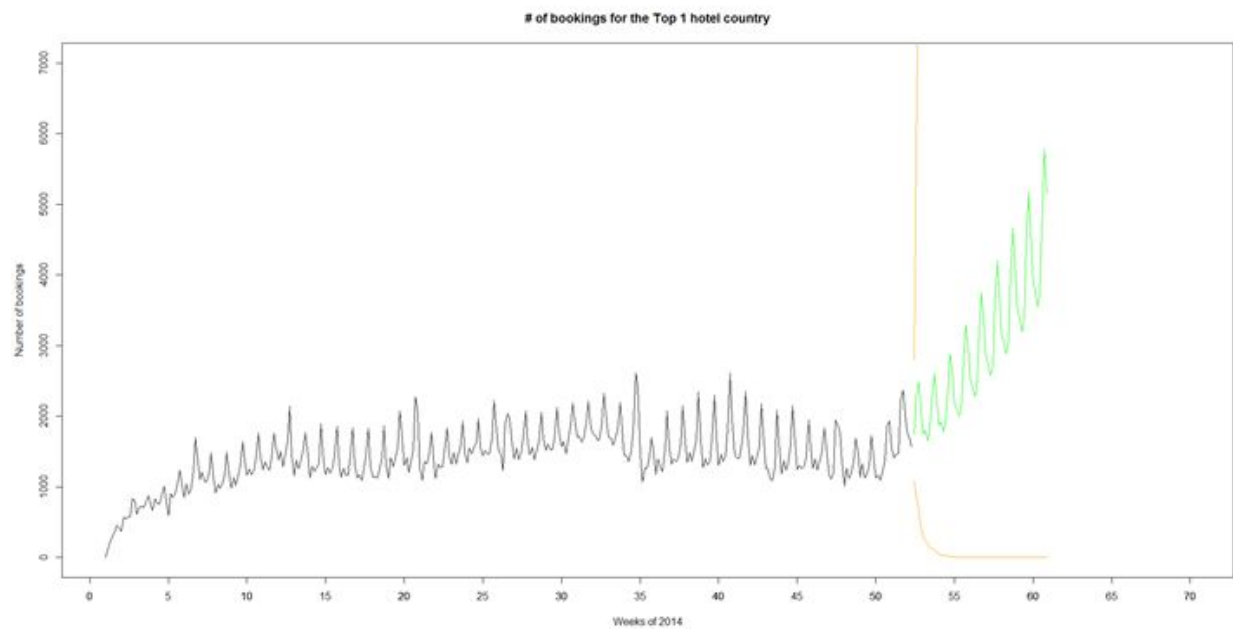
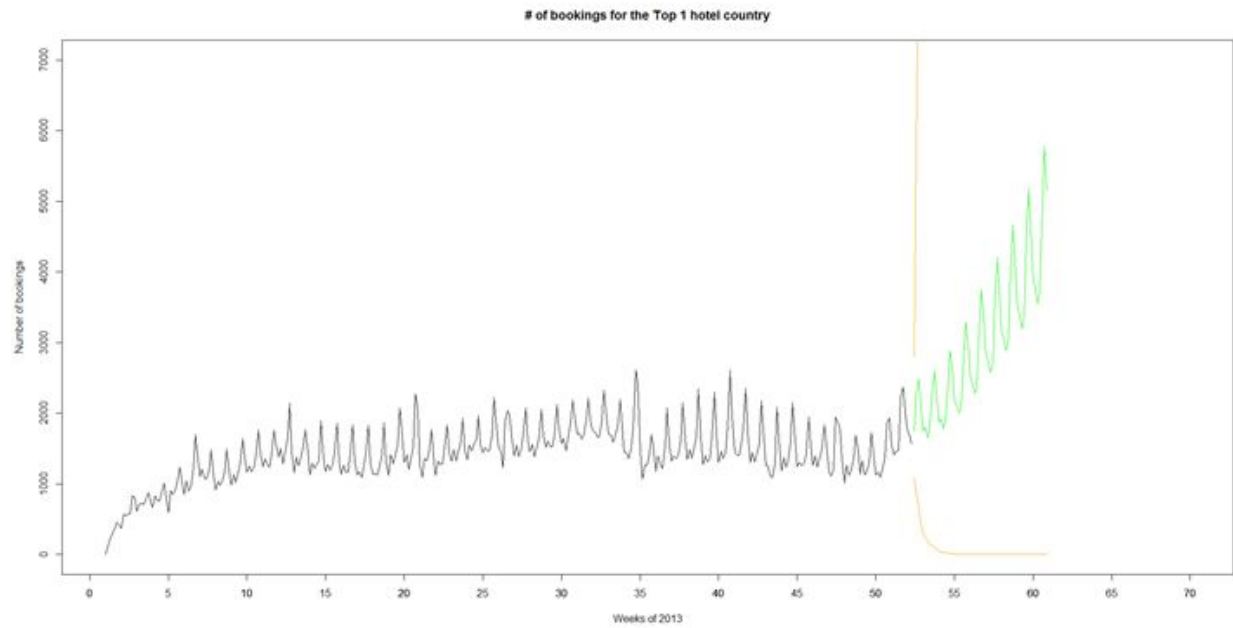
Using the check-in date of the customer for the above-mentioned hotel countries, AutoRegressive Integrated Moving Average (ARIMA) model is run against the frequency of the hotel country to plot the number of bookings the hotel country had for each week of a year. The parameters of the ARIMA model are then used as a predictive model for making forecasts for future values of the time series.

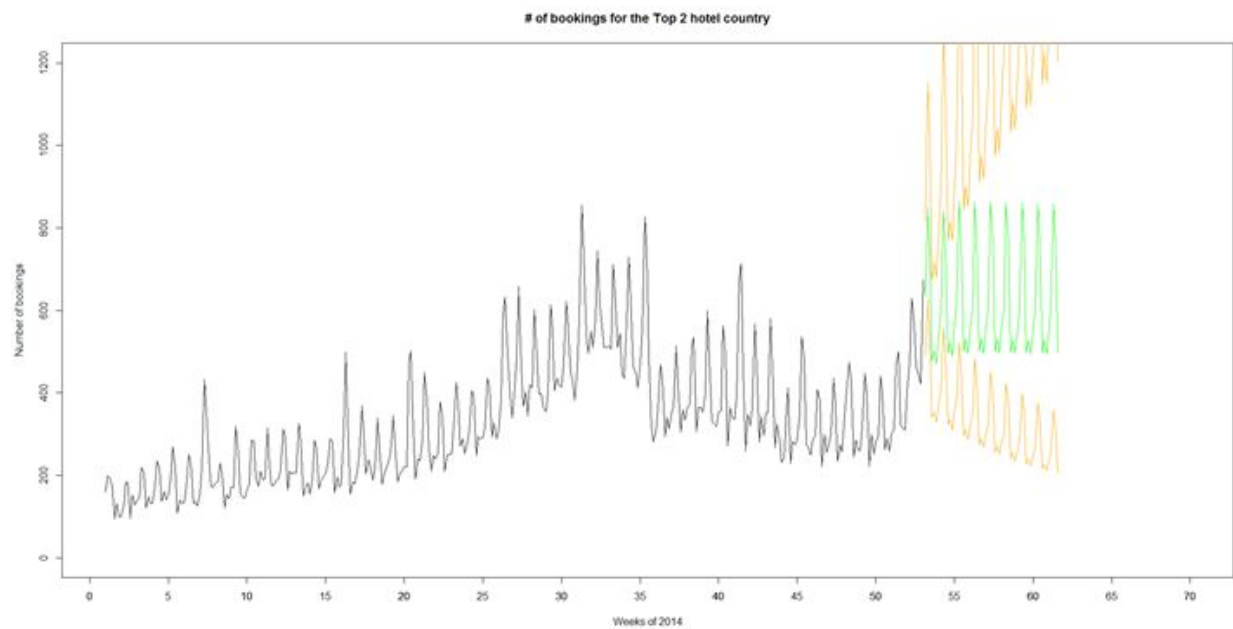
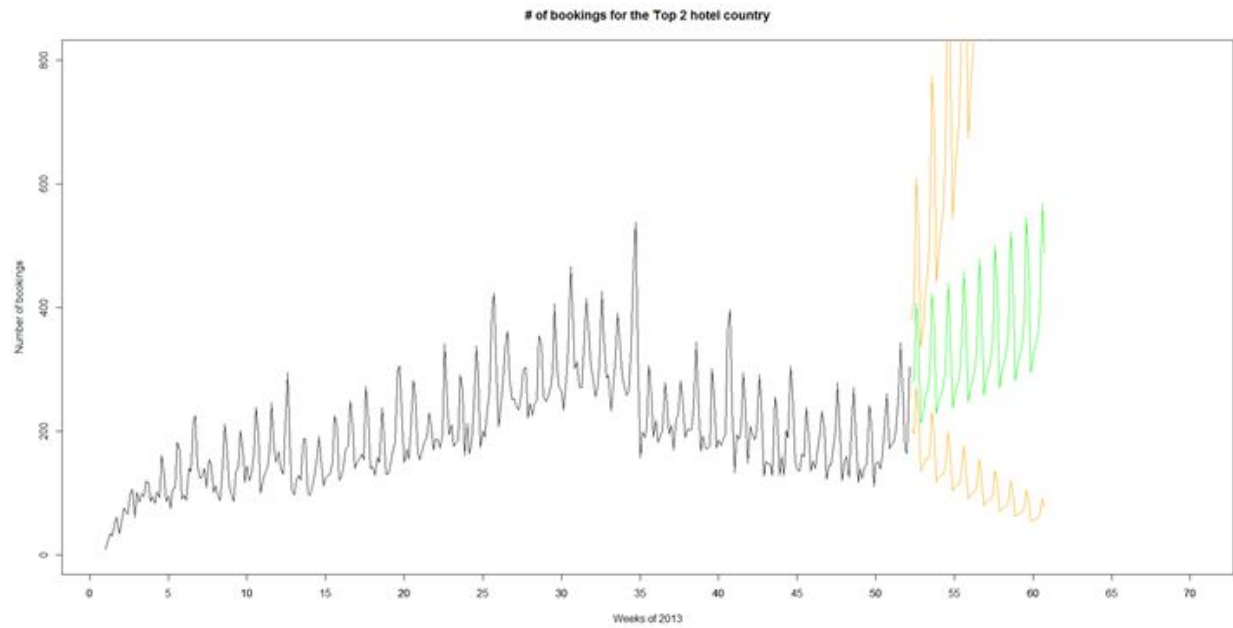
The graphs below show the time series analysis and forecast of the top 3 countries for the year 2013 and 2014 respectively. The x-axis represents each week of the current year and few weeks of the

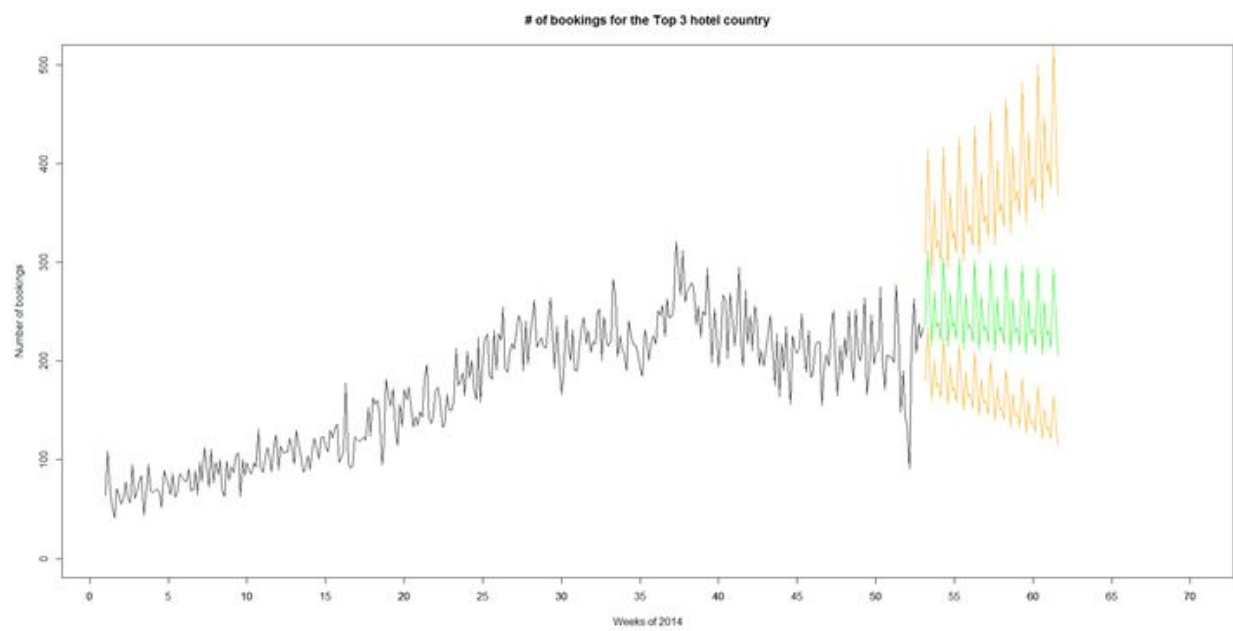
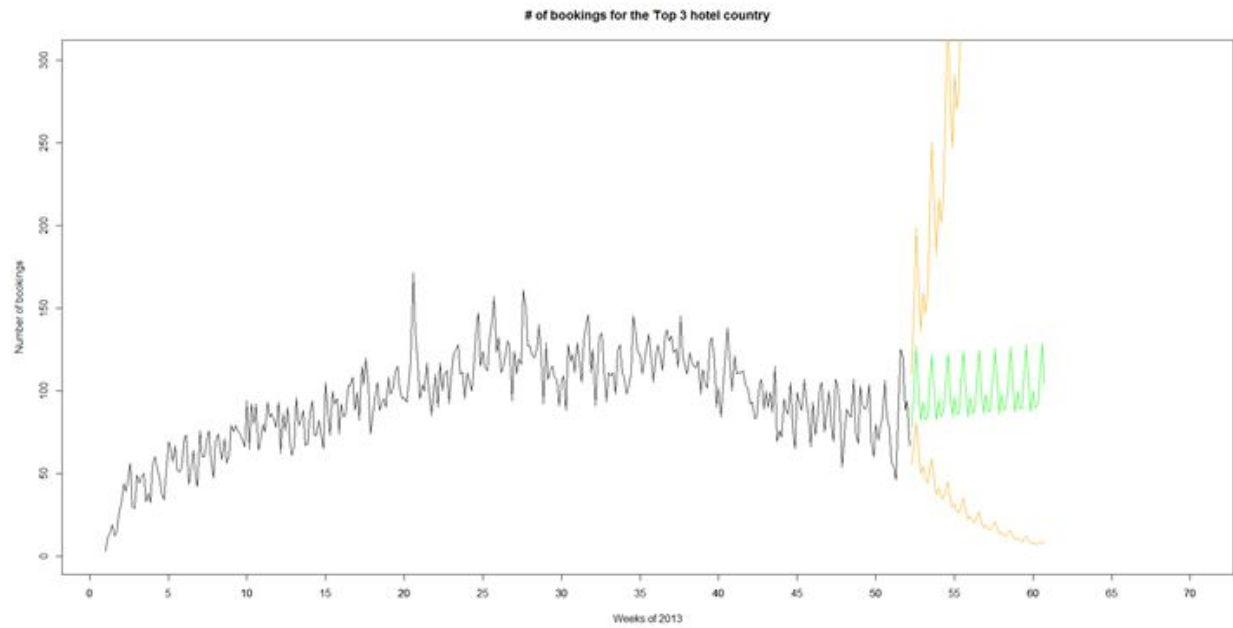
next year. Y-axis represents the number of bookings. After the 52nd year, the green trend shows the predicted number of bookings for the country. The orange line above and below the green line shows the upper and lower bound of number of bookings calculated by taking 2 standard deviations. This means the maximum and the minimum number of bookings that the country can reach for each upcoming week.

Output

Year	Top 1	Top 2	Top 3
2013	Maximum bookings in the 34 th and 41 st week	Maximum bookings in the 34 th week.	Maximum bookings in the 21 st week
2014	Maximum bookings in the 34 th and 41 st week	Maximum bookings in the 31 st and 36 th week	Maximum bookings in the 37 th week







C. Importance of Marketing Channel

Understanding the user conversion journey across the Marketing channel is one of the most important aspects for the business, knowing the traffic on the website from different marketing channels and the conversion rates across these channels can help the decision makers to prioritize the marketing efforts to certain identified channels basis the performance.

In our case we have identified the top marketing channels for the year 2013 & 2014 and for these channels we have further created a 2-month prediction basis the data model.

The process of creating predictive model involved segmenting the time series into the weekly cycle and allow the algorithm to identify the pattern for each week and extrapolate to the future weeks. The top marketing channel is the channel 9 with the maximum 0.6M conversions in the year 2013 & 1.1M Conversions in the year 2014, channel 0 follows the channel 9 in the number of bookings with 0.1M in 2013 & 0.23M in 2014

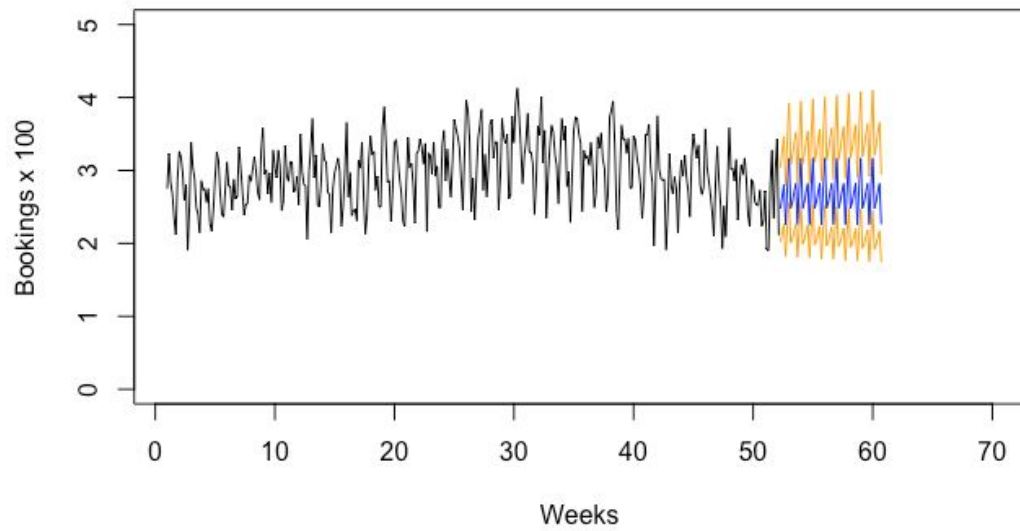
Number of Bookings across channel in 2013:

```
> table(data2013$channel)
```

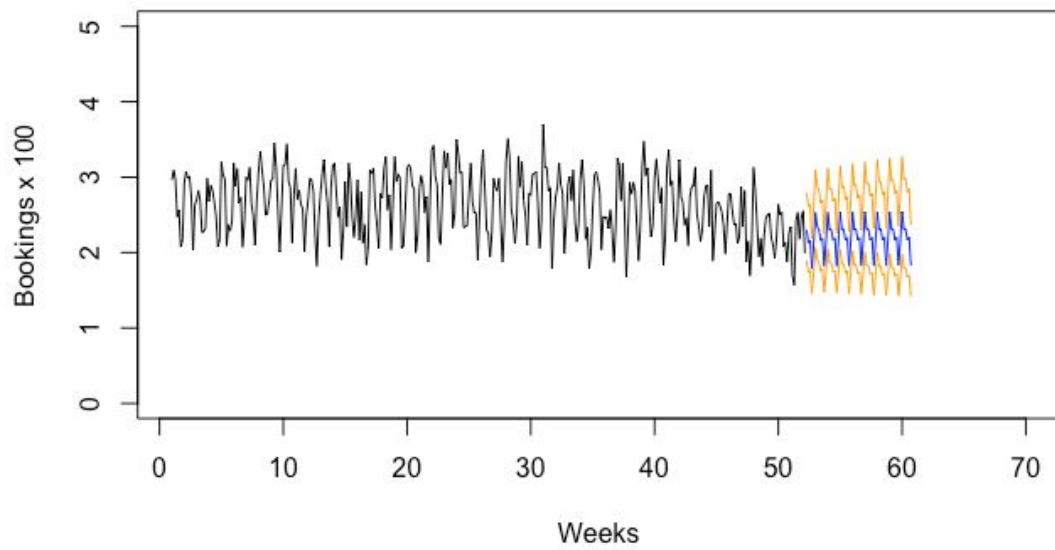
0	1	2	3	4	5	6	7	8	9	10
106214	94488	49092	25609	38664	49990	1018	4186	3509	651396	223

Channel level Bookings & Predictions Graphs for 2013

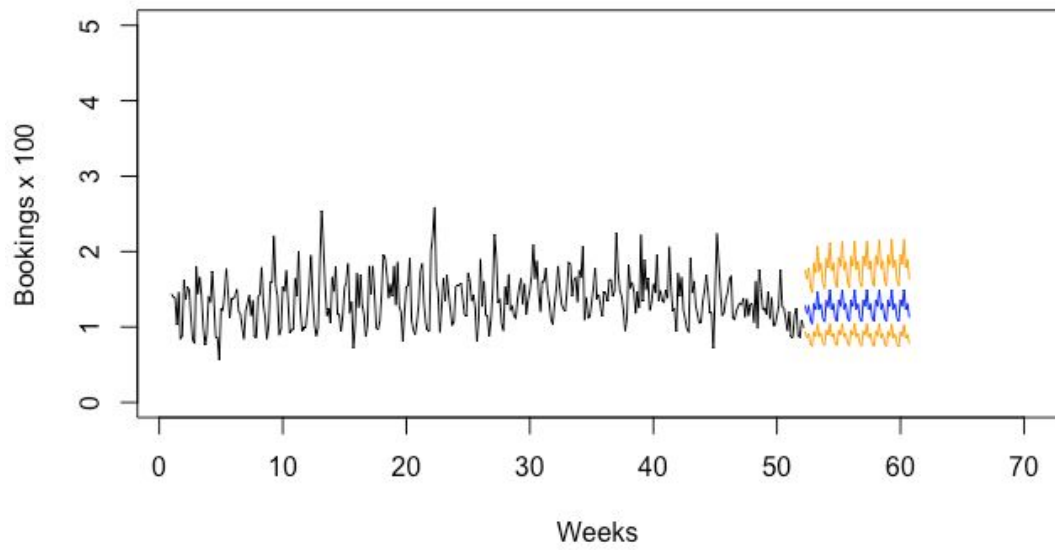
2013 Hotel Booking Data for Channel 0



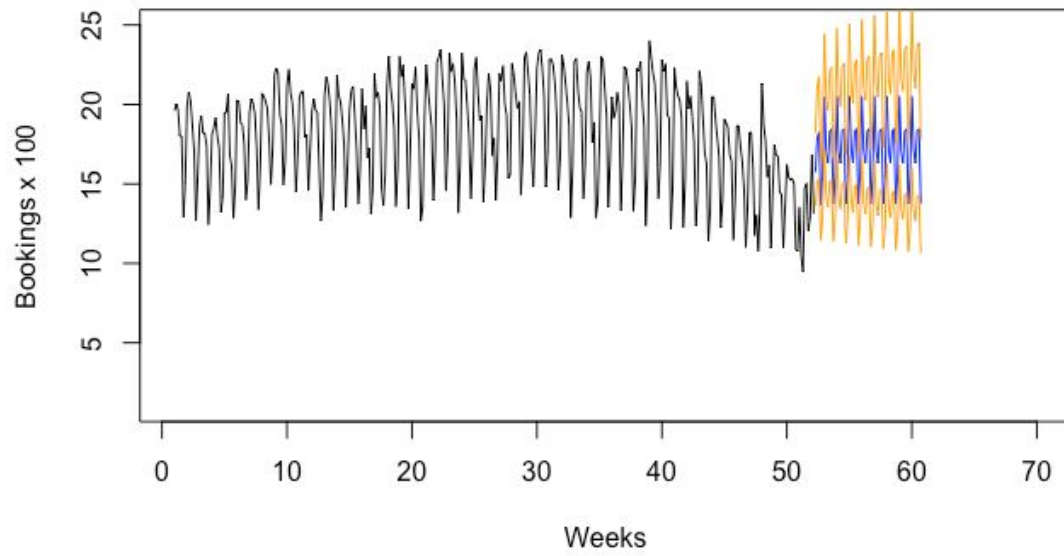
2013 Hotel Booking Data for Channel 1



2013 Hotel Booking Data for Channel 2



2013 Hotel Booking Data for Channel 9



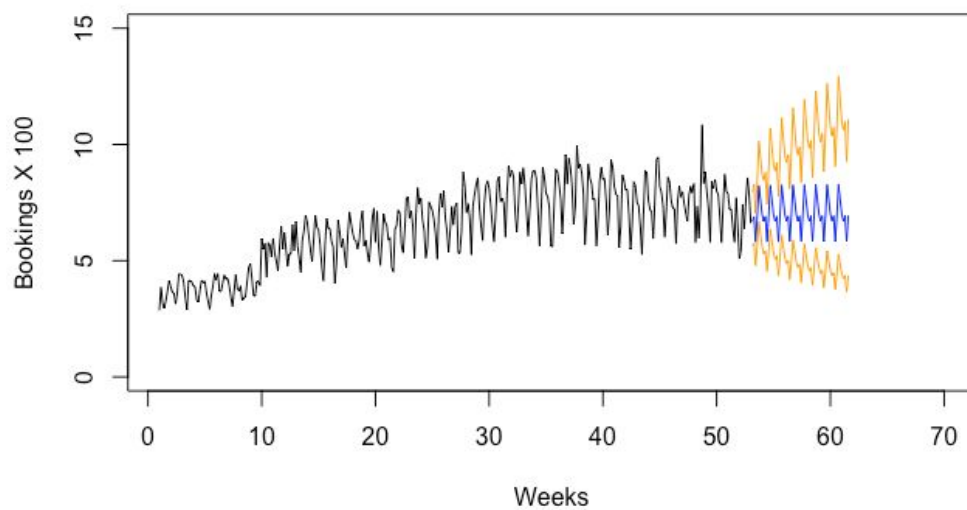
Number of Bookings across channel in 2014:

```
> table(data2014$channel)
```

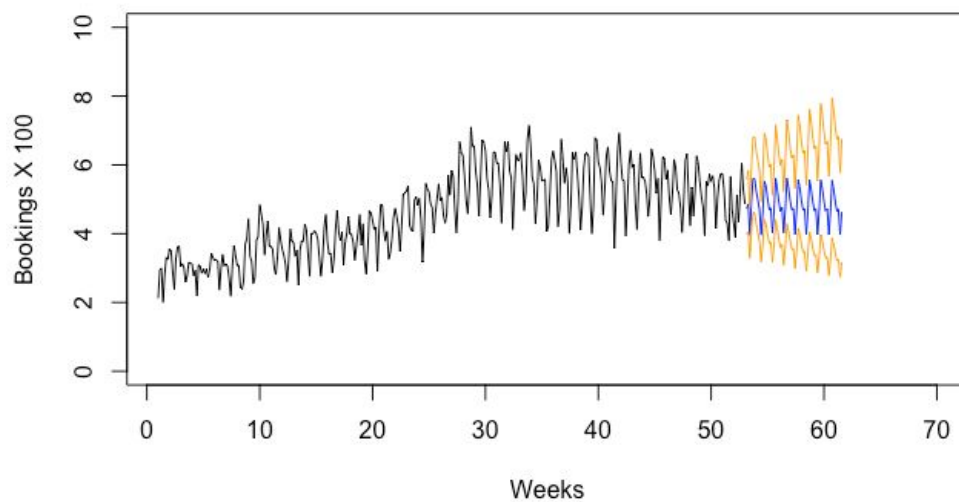
0	1	2	3	4	5	6	7	8	9	10
235144	167766	130094	69316	71797	168634	1868	8277	5806	1117389	213

Channel level Bookings & Predictions Graphs for 2013

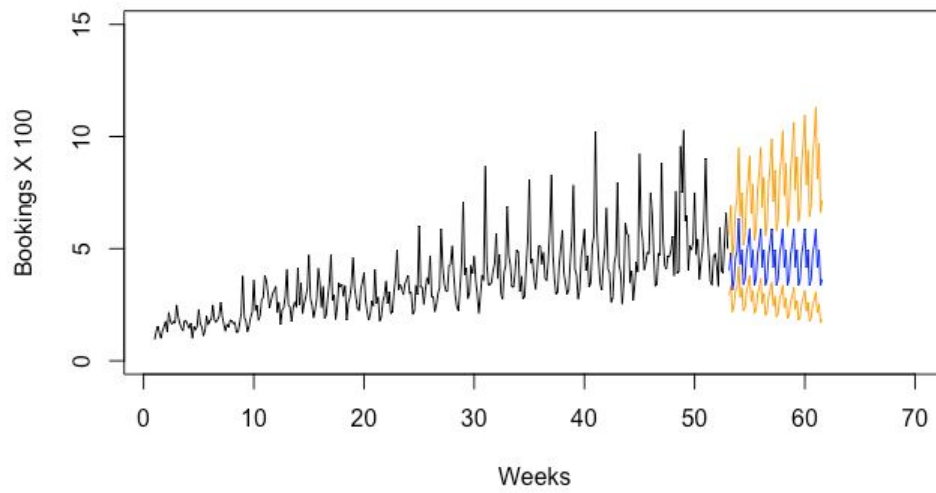
2014 Hotel Booking Data for Channel 0



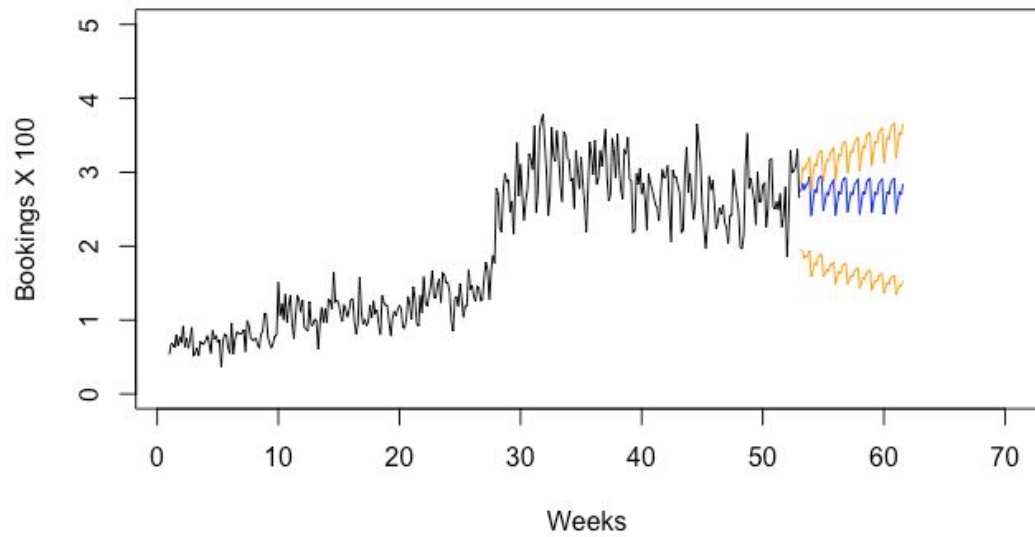
2014 Hotel Booking Data for Channel 1



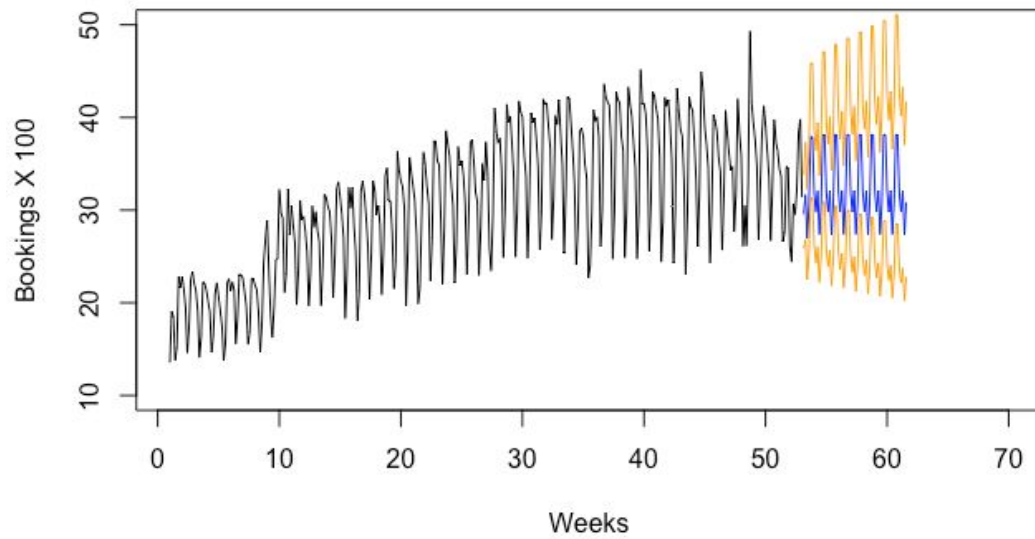
2014 Hotel Booking Data for Channel 2



2014 Hotel Booking Data for Channel 3



2014 Hotel Booking Data for Channel 9

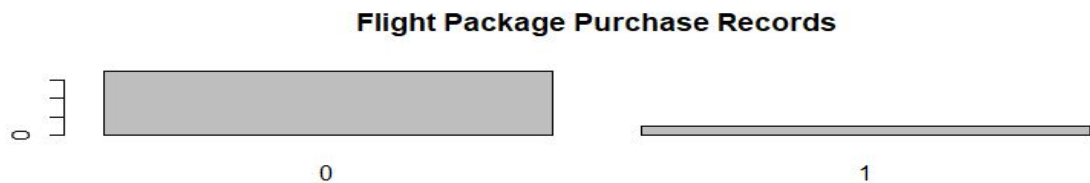


D. Predict the purchase of flight packages along with a hotel booking based on user search data.

The variable `is_package` tells us whether a customer booked a flight (1) or not (0) through Expedia along with their hotel booking.

Following changes were made to variables to optimize the model and reduce the number of dimensions.

- If `is_booking` is 0, it represents a click, and a 1 represents a booking. Here we are only considering booking events.
- `Srch_ci` and `srch_co` were replaced by the column no. of days representing the duration of the hotel stay.
- The variables `month_of_travel`, `month_of_booking`, `week_of_travel`, `week_of_booking` were derived from `date_time` and `Srch_ci` and `srch_co` to account for the seasonal variability of hotel bookings



- Since the number of records with 0 values for `is_package` is a lot higher we took a random sample with the equivalent number of 0 & 1 records to avoid the bias in prediction.

List of Predictors considered for Random Forest

posa_continent	:	int	3	3	3	3	3	1	3	3	1	2	...
is_mobile	:	int	0	0	0	0	0	0	0	0	0	0	...
is_package	:	int	1	1	1	1	1	1	1	1	1	1	...
channel	:	int	9	4	1	1	1	5	9	9	0	2	...
srch_adults_cnt	:	int	2	2	3	2	2	2	1	2	2	0	...
srch_children_cnt	:	int	0	0	0	0	0	0	0	0	0	0	...
srch_rm_cnt	:	int	1	1	2	1	1	1	1	1	2	1	...
srch_destination_type_id	:	int	1	1	1	1	1	1	1	1	1	1	...
hotel_continent	:	int	2	2	2	2	2	2	2	2	2	3	...
no_of_days	:	int	5	3	6	6	4	6	3	4	4	3	...
week_of_travel	:	int	2	2	2	2	2	3	2	1	4	3	...
week_of_booking	:	int	4	1	1	1	2	1	3	4	1	4	...
month_of_booking	:	int	8	4	2	2	2	10	11	1	5	9	...
month_of_travel	:	int	8	2	1	1	1	9	10	10	4	9	...

Output & Analysis

For the year 2013

Confusion Matrix and Statistics

		Reference	
Prediction		0	1
0	38074	5816	
1	11128	43386	

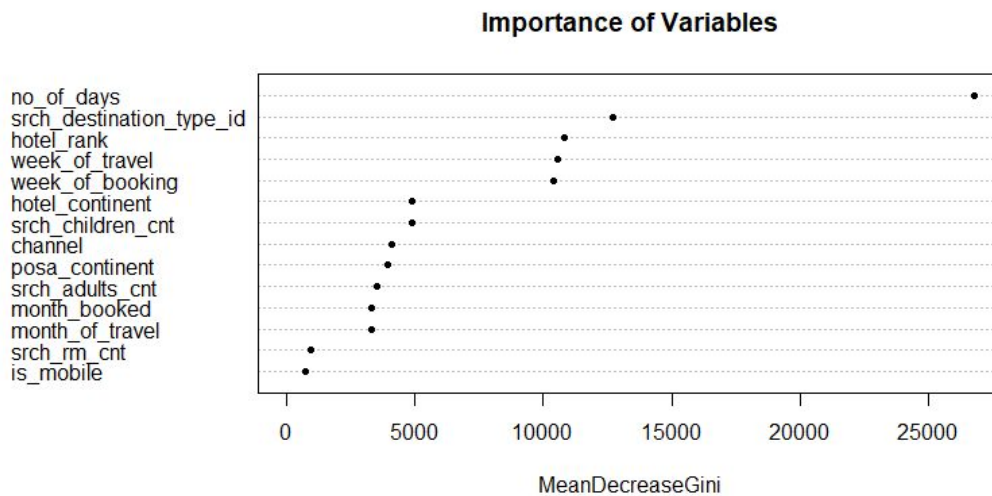
Accuracy : 0.8278
95% CI : (0.8254, 0.8302)
No Information Rate : 0.5
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6556
McNemar's Test P-Value : < 2.2e-16

Sensitivity : 0.7738
Specificity : 0.8818
Pos Pred Value : 0.8675
Neg Pred Value : 0.7959
Prevalence : 0.5000
Detection Rate : 0.3869
Detection Prevalence : 0.4460
Balanced Accuracy : 0.8278

'Positive' Class : 0

Records were classified with 82.78% accuracy



no_of_days is the most significant contributor. Followed by srch_destination_type_id, hotel_rank, week_of_travel and week_of_booking. srch_rm_cnt and is_mobile are the least contributing factors.

For the year 2014

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	58712	11059
1	15143	62796

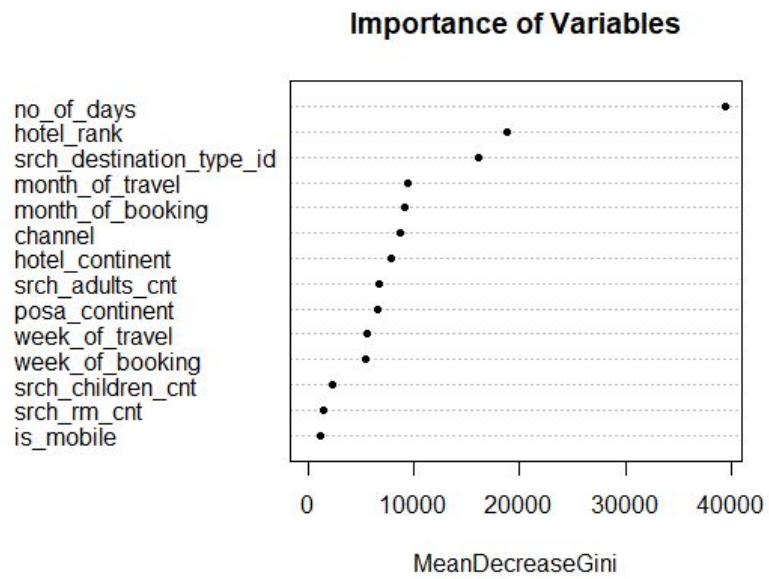
Accuracy : 0.8226
 95% CI : (0.8207, 0.8246)
 No Information Rate : 0.5
 P-Value [Acc > NIR] : < 2.2e-16

 Kappa : 0.6452
 Mcnemar's Test P-Value : < 2.2e-16

 Sensitivity : 0.7950
 Specificity : 0.8503
 Pos Pred Value : 0.8415
 Neg Pred Value : 0.8057
 Prevalence : 0.5000
 Detection Rate : 0.3975
 Detection Prevalence : 0.4724
 Balanced Accuracy : 0.8226

 'Positive' Class : 0

>> We were able to classify the records with 82.26% accuracy



>> We can see that `no_of_days` is the most significant predictor. Followed by `hotel_rank` and `srch_destination_type_id`. `srch_rm_cnt`, `srch_children_cnt`, and `is_mobile` are the least contributing factors

Recommendations

1. Expedia can explore cross-selling opportunities for their flight packages by using the Random Forest Classifier. They can also target their flight deals through emailers, social media advertisements to the customers that are likely to purchase the flight package.
2. The top marketing channels for Expedia for Hotel Bookings are Channel 0, 1, 2, 5 & 9 where most bookings have been from channel 9, marketing efforts should be planned in a way to allow maximum exposure to this channel and marketing budgets of non-performing channels should be shifted to this channel.
3. Most number of booking has been made to three countries 50, 198 and 70. More research can be done on these countries to get a better idea of the type of hotels they contain, their tourist attractions and marketing strategies. Better deals and offers can be provided to the remaining destinations to increase their traffic.
4. Most people who have visited destination type {1,3,4} has a high chance of visiting a destination of type 5. Similarly, people who have visited hotel cluster {46,97} are highly likely to visit hotels which come under cluster 64. More rules have been found which seems to boost profit if concentrating significant part of revenue thereby investing in these areas.

References

“Vision, Purpose, Strategic Imperatives & Guiding Principles.”

<https://www.expediagroup.com/about/vision-purpose-strategy-principles/>

>> Newsroom

<https://newsroom.expedia.com/home/fast-facts>

R. Narayanan, D. Honbo, G. Memik, A. Choudhary, and J. Zambreno, “An fpga implementation of decision tree classification,” in Design, Automation Test in Europe Conference Exhibition, 2007. DATE 07, April 2007

R. J. Prenger, B. Y. Chen, D. M. Merl, T. D. Lemmond, and W. G. Hanley, “Fast map search for compact additive tree ensembles,” LLNL, Tech. Rep., 2012

Box G E P, Jenkins G M and Reinsel G C 1994 Time Series Analysis: Forecasting and Control 3rd edn (Englewood Cliffs, NJ: Prentice-Hall)

Miller R G 1997 Beyond ANOVA: Basics of Applied Statistics (New York: Chapman and Hall)

Jiawei Han, Jian Pei, Yiwen Yin: Mining Frequent Patterns without Candidate Generation. SIGMOD Conference 2000: 1-12

Zijian Zheng, Ron Kohavi, and Llew Mason, Real World Performance of Association Rule Algorithms, KDD 2001.