MA678 Midterm Project: Identification of Hotel Cluster Choice

https://github.com/KratosMC/Expedia-Hotel-Recommendations | Chenghao Meng

1 Abstract

As age of Internet evolves, more and more people are using online travel platform to book hotels. Expedia, as a famous online travel shopping company, are interested in using the formatted user event data to identify users' hotel choice. In the following chapters, a multinomial model is established to identify users' hotel choice given a series of user events based on the cleaned data and engineered features. Model fitness is relatively good, and suggestions are given based on the analysis as a whole, which are also divided into user, hotel and company management perspectives.

2 Introduction

This project is using dataset from Expedia downloaded from Kaggle[1]. Expedia is an American online travel shopping company for consumer and small business travel, and this dataset from Expedia contains the hotels' geolocation and type information, as well as logs of customer behavior, including what customers searched for, how they interacted with search results (click/book), whether the search result was a travel package, etc.[2] In this project, we will establish a model using the information of user event to identify which hotel cluster those users will book. Based on the exploratory analysis and the model, we will give some suggestions to Expedia regrading refinement of hotel booking service.

3 Method

3.1 Data Extraction and Processing

3.1.1 Data Extraction

Since the original dataset on Kaggle has 37,670,293 rows in a .csv file, which is very large for the local machine to proceed, we will select the 200,000 rows randomly without replacement by using ***sample_n*** function in ***dplyr*** package to conduct the following exploratory data analysis.

3.1.2 Data Processing

- *Data Transformation*

First, we will look at the dataset to prepare for further data processing. After looking at the summary, we will conduct data transformation. Columns **date_time**, **srch_ci** and **srch_co** should be in date-time format. Columns **site_name**, **posa_continent**, **user_location_country**, **user_location_region**, **user_location_city**, **user_id**, **channel**, **srch_destination_id**, **srch_destination_type_id**, **hotel_continent**, **hotel_country**, **hotel_market** and **hotel_cluster** should be in categorical format.

- *Imputation*

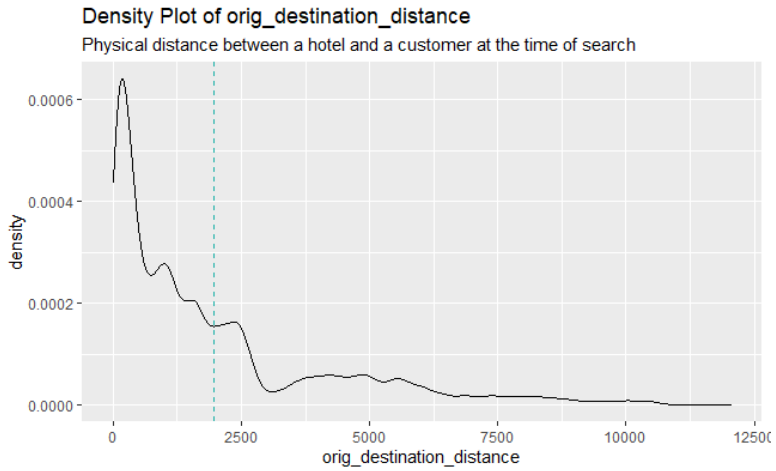
There are 35.84% of the data in column **orig_destination_distance** are missing values. Since the missing of data may be due to users' unwillingness to give Expedia access to their geographic locations, deleting the rows with NAs is not a good approach to address this problem. After checking the distribution of the data, since all the relevant columns are categorical, we will use the CART[3] in the ***mice*** function to fill the missing values. Meanwhile, because ***mice*** function cannot process all

[1] Overview: https://www.kaggle.com/c/expedia-hotel-recommendations/overview

[2] Data Description: https://www.kaggle.com/c/expedia-hotel-recommendations/data

[3] Abbreviation of Classification and Regression Trees method

200,000 columns at once, we divided the dataset into 20 patches, and there will be 10,000 rows in each patch by using the ***For-loop***.



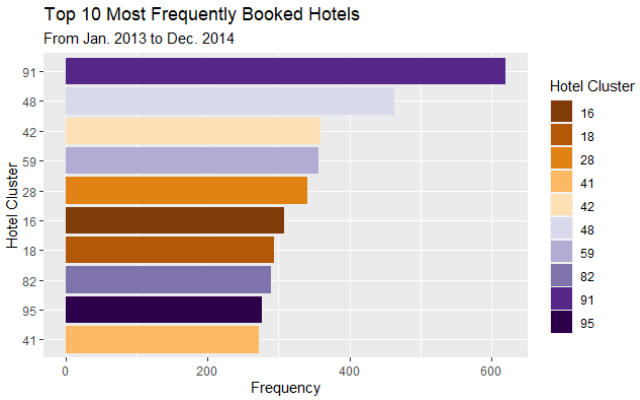
Since there are only 177 NAs in **srch_ci** and 177 NAs in **srch_co**, we have dropped those rows with NAs.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Hotel-related Exploration

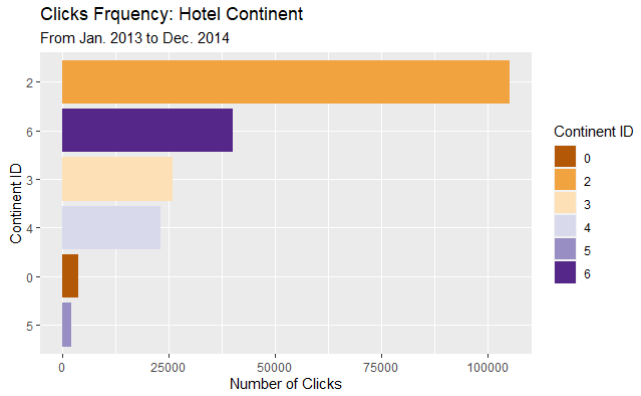
- *Most Popular Hotel Cluster*

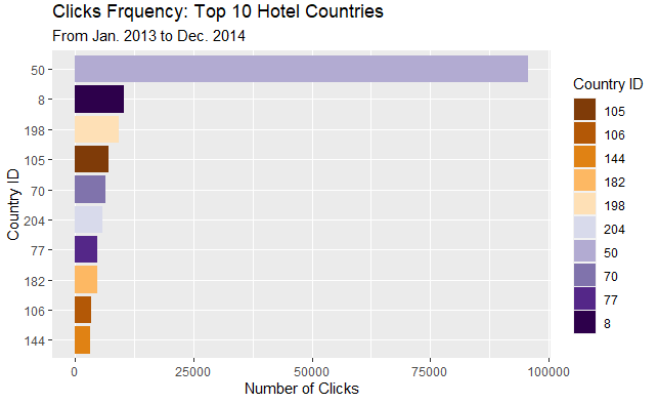
When exploring hotel-related feature in the dataset, the first thing of interest would be to know which hotel cluster is most frequently booked. The graph below shows that Hotel Cluster 91 is most frequently booked in the given time window.



- *Most Popular Hotel Continent and Country*

Then, we would also like to know which continent and country are most popular when the users are searching their hotels. The plots below show that Hotel Continent 2 and Hotel Country 50 are the most popular continent and country.



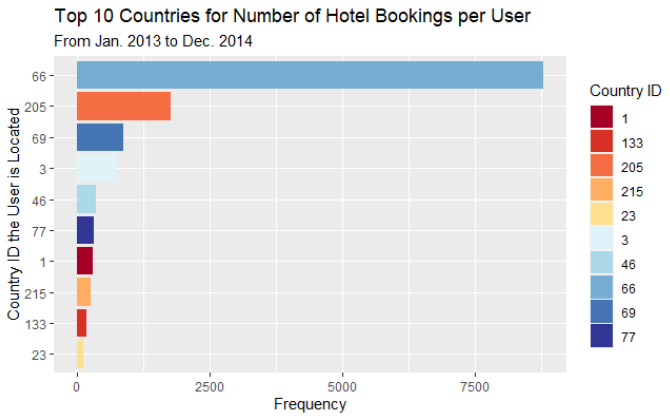


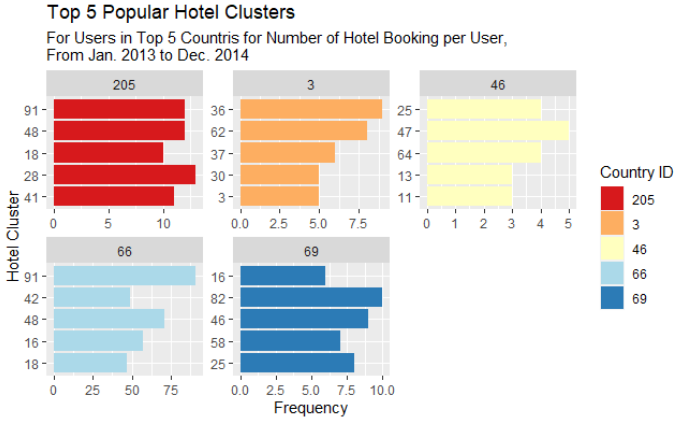
3.2.2 User-related Explorations

- *Which country's users have the most hotel bookings & Which hotel clusters are popular in each country?*

When exploring the user-related features, we would like to know which country's users book hotels in the most frequent

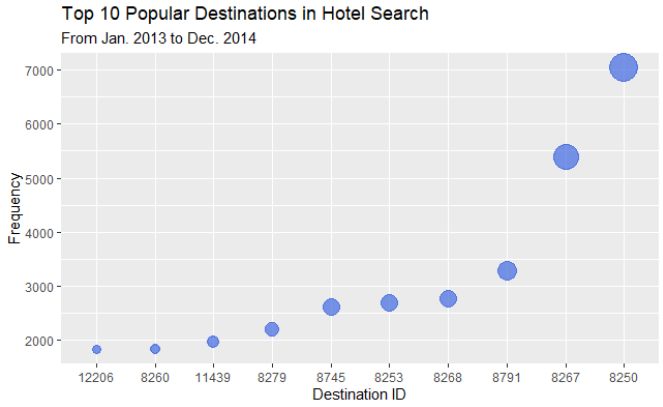
way. The plot below indicates that User Location Country 66 has the most bookings among other countries, which might indicate that the average income of Expedia users in User Location Country 66 is relatively high Note that, in User Location Country 66, the Hotel Cluster 91 is the most popular hotel cluster, which coincided with the fact that Hotel Cluster is the most popular hotel cluster in general as indicated above during this particular time window.





- *Which destination the users like most?*

After exploring the hotel cluster, we would also want to know which destination is the most popular one when the users of Expedia are searching for their hotel. The plot below denotes that Srch (Search) Destination 8250 is the most popular destination when users are looking for hotels.

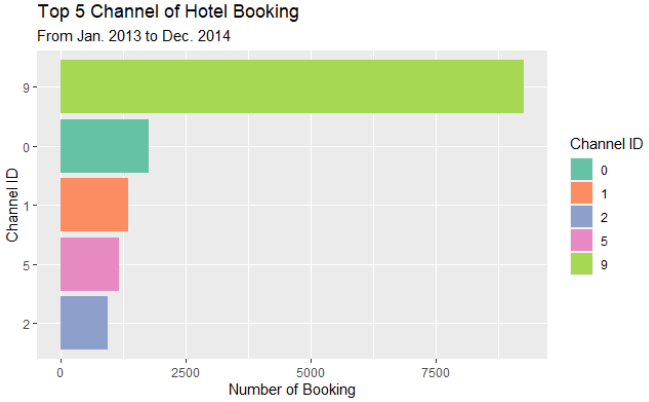


3.2.3 Expedia-related Explorations

After exploring the information relating to user and hotel, we want to dig further to the information relating to the Expedia website.

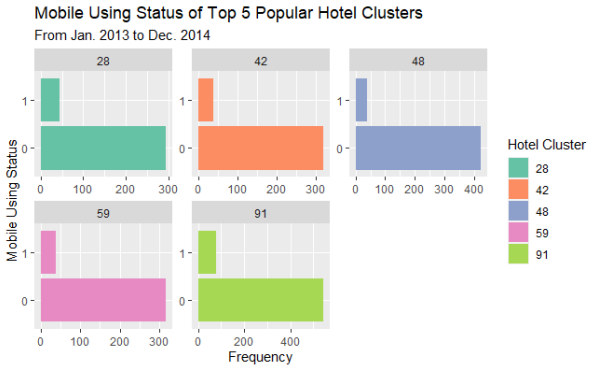
- *Which channel is most frequently used when users are booking hotels?*

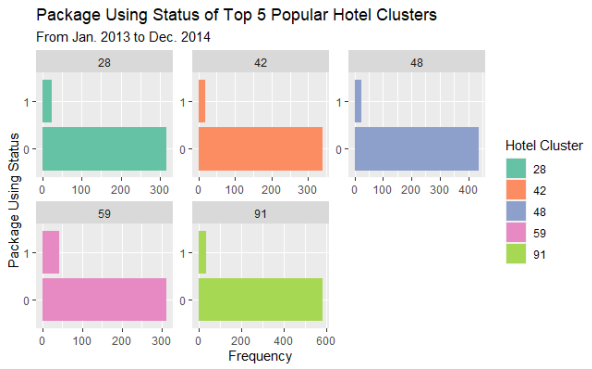
Users are using different channels to book hotel on Expedia, so knowing which channel is most frequently used when users are booking hotel is very important for the company to optimize their consumer experience. The plot above indicates that Channel 9 is mostly used when users are booking hotels.



- *Do users use package and mobile phone to book hotel?*

What's more, the using status of package and mobile can be another important indicator of user event when users are booking hotel. By understanding the relationship between the status and hotel cluster booking, the company can have a better understanding of their user. Since 100 hotel clusters are too many, we will only select the top 5 popular hotel clusters as indicated in 3.2.1, and the plots below shows that most of the users are not using package and mobile when they are booking those top 5 popular hotel clusters.





3.3 Feature Engineering

3.3.1 Feature Establishment

- *Duration of Stay: stay_duration*

Since **srch_ci** denotes check-in date and **srch_co** denotes check-out date, the duration of stay would be a feature of interest to explore users' hotel booking behavior.

- *Hotel Rank: hotel_rank*

Meanwhile, since the number of hotel cluster is relatively large (100 hotel clusters), we would like to transform it into a 10-point scale based on the number of booking for the simplicity of the analysis.

If hotel cluster ID equals to top 10 frequently booked hotel, then it would be ranked as 10. If hotel cluster ID equals to top 11-20 frequently booked hotel, it would be ranked as 9. If hotel cluster ID equals to top 11-30 frequently booked hotel, it would be ranked as 8. The rest of the ranking will do the same as above.

3.3.2 Feature Selection

Since **site_name** and **posa_continent** in the dataset all reflect which continent the users are based, and **posa_continent** has similar structure as **hotel_continent**, the **posa_continent** will be chosen for the modelling part.

Meanwhile, if the user has children that are needed to be placed in a separate room, it will reflect in the **srch_rm_cnt**, so we will choose this column and drop **srch_adults_cnt** and **srch_children_cnt**.

Because the meanings of **hotel_market** and **srch_destination_type_id** are not specified in the data description part on Kaggle, and **srch_destination_id** cannot match with continent and country information, those columns will also be dropped for interpretability of the variable.

What's more, user has city-level feature **user_location_city** and region-level feature **user_location_region**, however, hotel only have continent feature **hotel_continent** and country-level feature **hotel_country**, so **user_location_city** and **user_location_region** will be dropped for the correspondence of the features. Moreover, **date_time** will also be dropped for the simplicity of the model.

3.3 Model

3.3.1 Data Sub-setting

Since we are interested in using the information of user event to identify which hotel cluster those users will book, the

data indicating a series of user event for not booking a hotel should not be considered in the model.

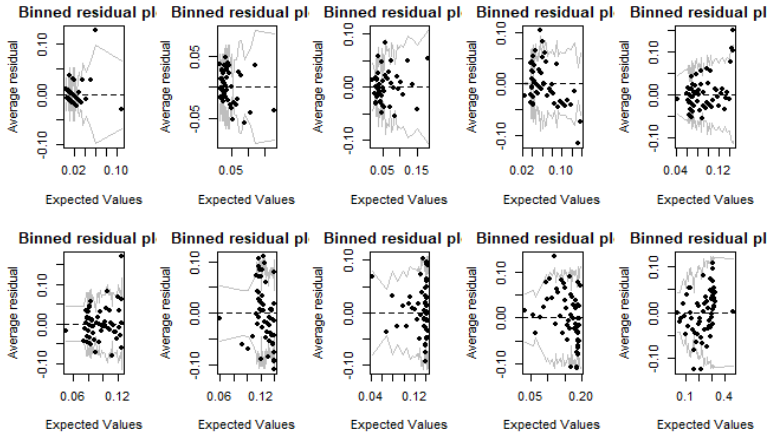
3.3.2 Model Establishment

The outcome variable of interest in **hotel_rank**, which is an ordered categorical data based on the popularity of the hotel cluster, so we will fit a multinomial model. Model is established as follows:

```r
# fit the model
fit <- polr(ordered(hotel_rank)~user_location_country +
posa_continent + orig_destination_distance  + hotel_continent+
is_mobile + is_package  + srch_rm_cnt  + cnt + channel+
stay_duration,data=df_final,Hess = T)
```

3.3.3 Model Checking

After fitting the model, binned plots should be drawn to visualize the fitness of the model. The binned plots are showed as below. The binned residual plots indicate that there are only a few outliers, most of the estimations are be in the range. And the result of the model in the Appendix shows that the standard errors are relatively small.



4 Result

According to the EDA, User Location Country 66's frequency of booking a hotel is much larger than users in the rest of user groups. Moreover, Top 10 Frequently Booked Hotel Clusters also frequently appeared in countries with large quantity of hotel booking per user. Besides, Channel 9 takes the main position among all channels.

According to the model, we believe the fit is relatively good. So, the established model may conduct the identification of hotel cluster booking with given user events.

5 Discussion

According to the EDA, **from the user side**, we can see that users from User Location Country 66 are the group of users that Expedia should target at. **From hotel side**, Expedia may want to establish more closer relationship with Top 10 Frequently Booked Hotel Clusters. **From the Expedia itself**, Expedia should also improve the use of other channels to better integrate all channels, and promoting the user of mobile phone to book hotel is also in great need since only a relatively small amount of users are using mobile phone to book a hotel.

As for the next steps, a more cautious feature selection should be conducted based on the combination of statistical method and more abundant literature review. Time series factors might need to be added to the model to better identify the booking of hotel clusters with given user event.

Reference

[1] Pantelic, Vladan. (2017). Factors influencing hotel selection: Decision making process.

[2] Baber, Ruturaj. (2015). Criteria for Hotel Selection: A Study of Travellers. Pranjana: The Journal of Management Awareness. 18. 52-59. 10.5958/0974-0945.2015.00012.6.

Appendix

Model: Result of the Model

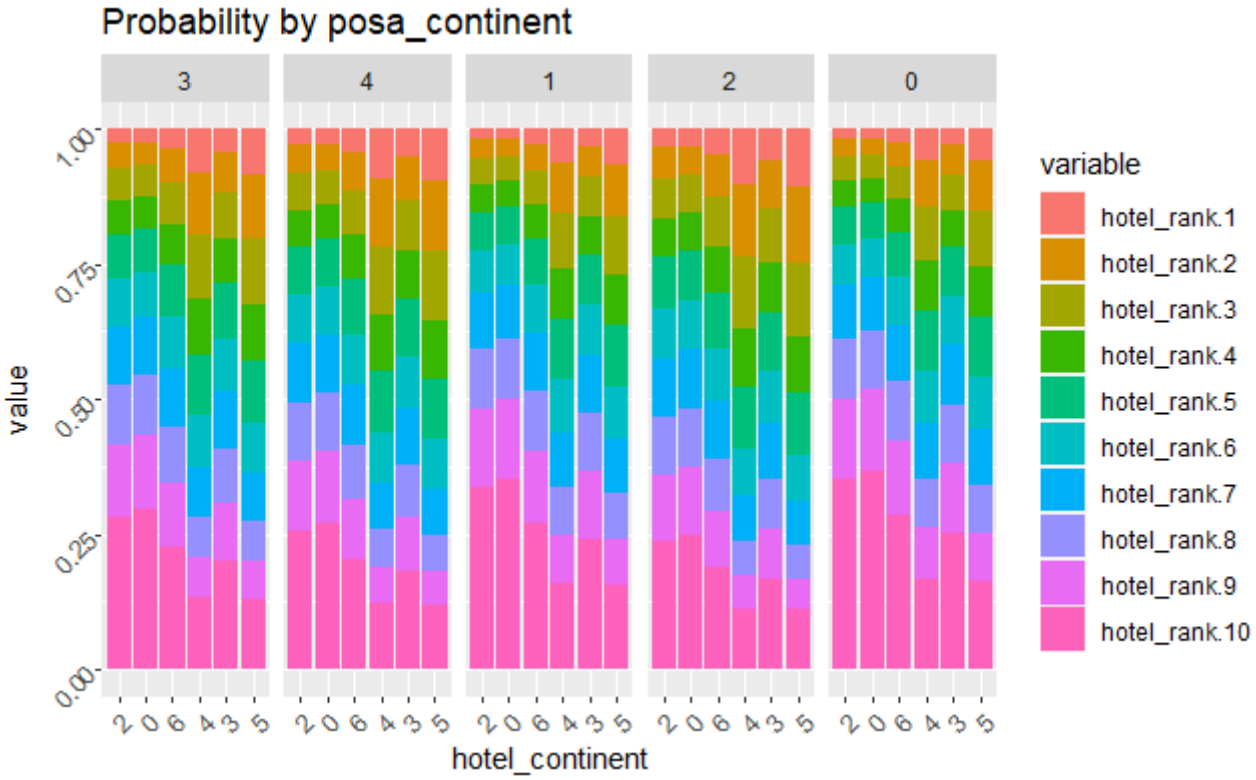
	coef.est	coef.se
user_location_country1	-0.10	0.01
user_location_country100	-0.57	0.00
user_location_country103	-0.94	0.00
user_location_country104	-1.45	0.00
user_location_country105	0.09	0.00
user_location_country11	14.07	0.00
user_location_country114	0.07	0.00
user_location_country115	-0.17	0.00
user_location_country117	0.48	0.00
user_location_country119	0.25	0.00
user_location_country12	-0.35	0.00
user_location_country123	0.24	0.00
user_location_country126	0.85	0.00
user_location_country129	-0.24	0.00
user_location_country130	-1.22	0.00
user_location_country131	1.46	0.00
user_location_country133	-0.23	0.00
user_location_country134	0.22	0.00
user_location_country136	2.45	0.00
user_location_country141	-0.43	0.00
user_location_country143	0.35	0.00
user_location_country146	-0.18	0.00
user_location_country148	0.53	0.00
user_location_country149	-1.57	0.00
user_location_country152	-1.26	0.00
user_location_country154	0.83	0.00
user_location_country162	0.00	0.00
user_location_country163	-0.54	0.00
user_location_country166	-0.22	0.00
user_location_country167	0.36	0.00
user_location_country173	-3.43	0.00
user_location_country178	0.02	0.00
user_location_country179	0.83	0.00
user_location_country181	-0.80	0.00
user_location_country182	0.47	0.00
user_location_country188	0.37	0.00
user_location_country189	15.29	0.00
user_location_country19	0.83	0.00
user_location_country190	-2.37	0.00
user_location_country191	-0.43	0.00
user_location_country193	0.13	0.00
user_location_country194	0.40	0.00
user_location_country195	0.09	0.01
user_location_country197	-0.09	0.00
user_location_country198	-0.48	0.00
user_location_country202	0.60	0.00
user_location_country205	0.41	0.05
user_location_country206	1.43	0.00
user_location_country208	-0.67	0.00
user_location_country209	0.05	0.00
user_location_country210	2.27	0.00
user_location_country214	1.66	0.00
user_location_country215	0.30	0.01
user_location_country217	-0.21	0.00
user_location_country218	-0.71	0.00
user_location_country219	-0.85	0.00
user_location_country221	-0.12	0.00
user_location_country222	1.05	0.00
user_location_country225	14.57	0.00
user_location_country228	0.53	0.00
user_location_country229	-0.13	0.00
user_location_country23	0.76	0.00
user_location_country230	0.95	0.00
user_location_country231	-0.16	0.00
user_location_country235	-0.71	0.00
user_location_country24	-2.75	0.00
user_location_country27	14.25	0.00
user_location_country28	0.73	0.00
user_location_country29	-0.05	0.00
user_location_country3	0.32	0.04
user_location_country30	-1.12	0.00
user_location_country31	0.01	0.00
user_location_country32	0.18	0.00
user_location_country39	0.81	0.00
user_location_country4	0.02	0.00
user_location_country44	15.81	0.00
user_location_country46	0.03	0.01
user_location_country48	-0.33	0.00
user_location_country49	0.39	0.00
user_location_country5	0.82	0.00
user_location_country50	1.27	0.00
user_location_country51	0.24	0.00
user_location_country52	-0.89	0.00
user_location_country55	0.15	0.00
user_location_country57	-1.20	0.00
user_location_country6	1.71	0.00
user_location_country62	0.13	0.00
user_location_country63	0.44	0.00
user_location_country64	-0.17	0.00
user_location_country66	0.35	0.05
user_location_country68	-0.01	0.00
user_location_country69	0.08	0.04
user_location_country70	-1.59	0.00
user_location_country71	-0.10	0.00
user_location_country72	-2.61	0.00
user_location_country75	0.90	0.00
user_location_country77	0.19	0.02
user_location_country8	1.65	0.00
user_location_country80	-0.37	0.00
user_location_country82	0.37	0.00
user_location_country83	0.14	0.00
user_location_country85	-1.20	0.00
user_location_country87	-0.02	0.00
user_location_country91	-0.54	0.00
user_location_country93	-0.39	0.00
user_location_country95	-2.47	0.00
user_location_country99	2.55	0.00

posa_continent1	-0.09	0.06
posa_continent2	-0.75	0.05
posa_continent3	-0.43	0.05
posa_continent4	-0.60	0.03
orig_destination_distance	0.00	0.00
hotel_continent2	-0.09	0.06
hotel_continent3	-0.72	0.08
hotel_continent4	-1.46	0.06
hotel_continent5	-1.53	0.00
hotel_continent6	-0.50	0.06
is_mobile	-0.06	0.09
is_package	-0.39	0.08
srch_rm_cnt	0.02	0.06
cnt	-0.14	0.03
channel1	-0.07	0.09
channel10	-1.38	0.00
channel2	0.04	0.08
channel5	-0.31	0.09
channel6	1.17	0.00
channel7	-0.22	0.00
channel8	-0.11	0.00
channel9	0.05	0.05
stay_duration	-0.07	0.01
1|2	-5.18	0.02
2|3	-3.94	0.08
3|4	-3.13	0.08
4|5	-2.57	0.08
5|6	-1.98	0.08
6|7	-1.47	0.09
7|8	-0.91	0.09
8|9	-0.33	0.09
9|10	0.48	0.09

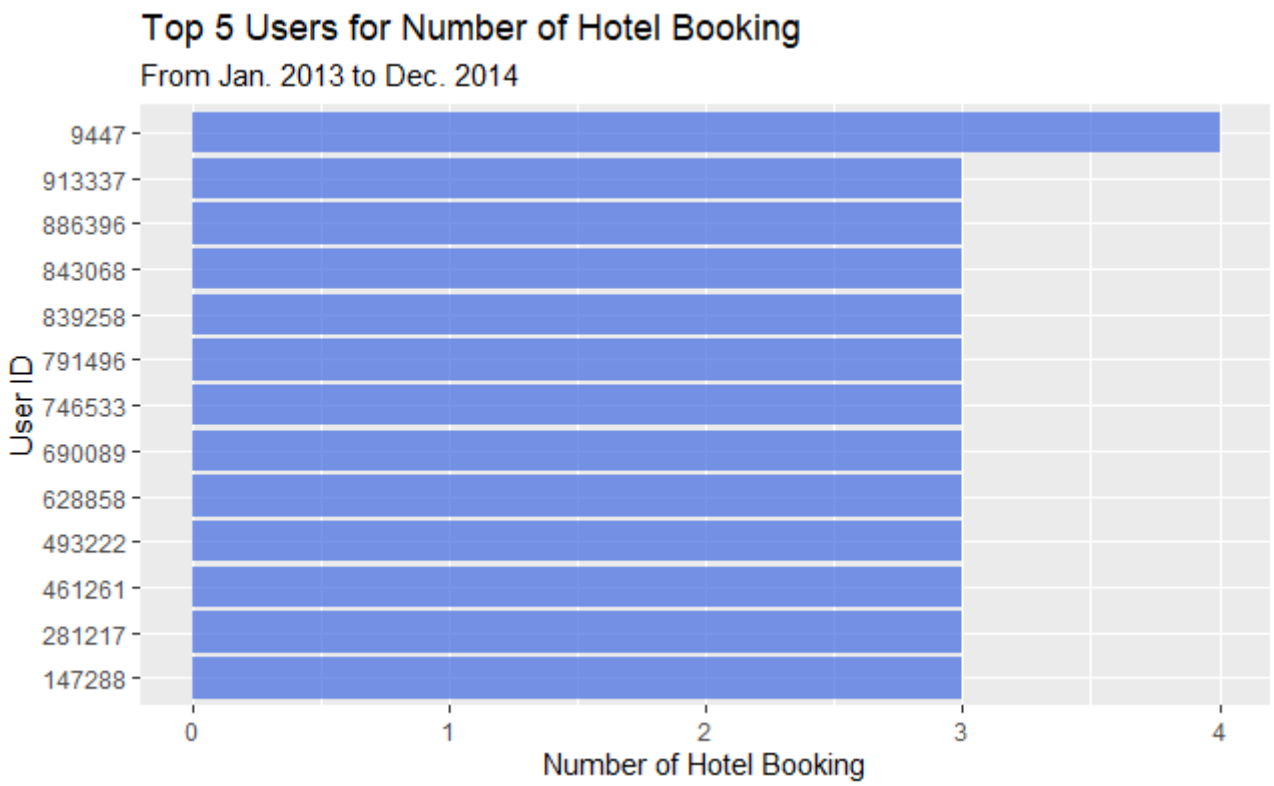
---

n = 3942, k = 141 (including 9 intercepts)
residual deviance = 16170.0, null deviance is not computed by polr

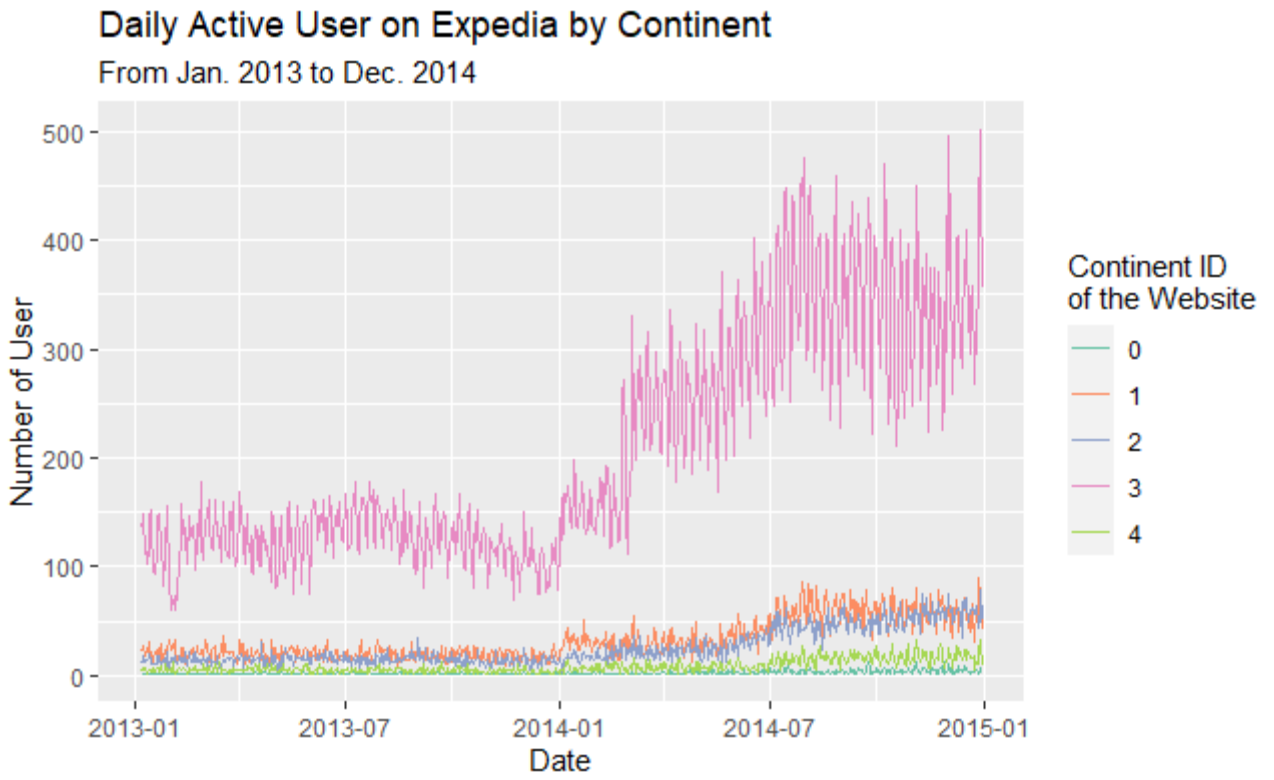
Model: Visualization of Probabilities



EDA: Top 5 Users for Number of Hotel Booking



EDA: Daily Active Users



Supplement

Code: Imputation by patches

```r
# Divide the dataset into patches
steps <- 10000
for (i in 1:20){
  imp <- mice(dt2[seq((steps*i)-9999,steps*i),],
method="cart",seed = 1,printFlag=F)
  hotel_temp <- complete(imp)
  hotel <- rbind(hotel,hotel_temp)
}
```