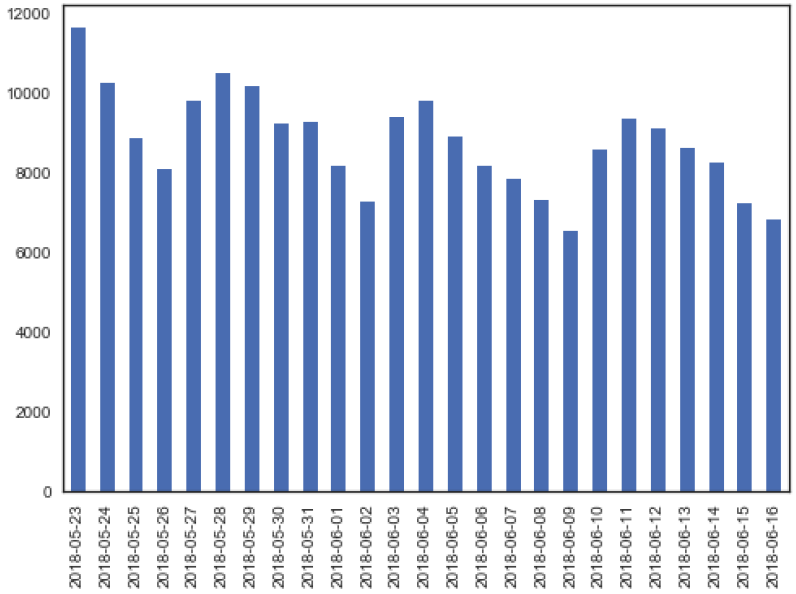
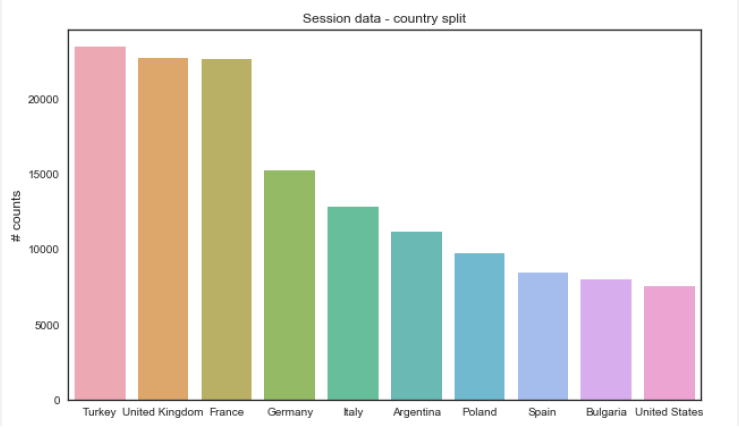
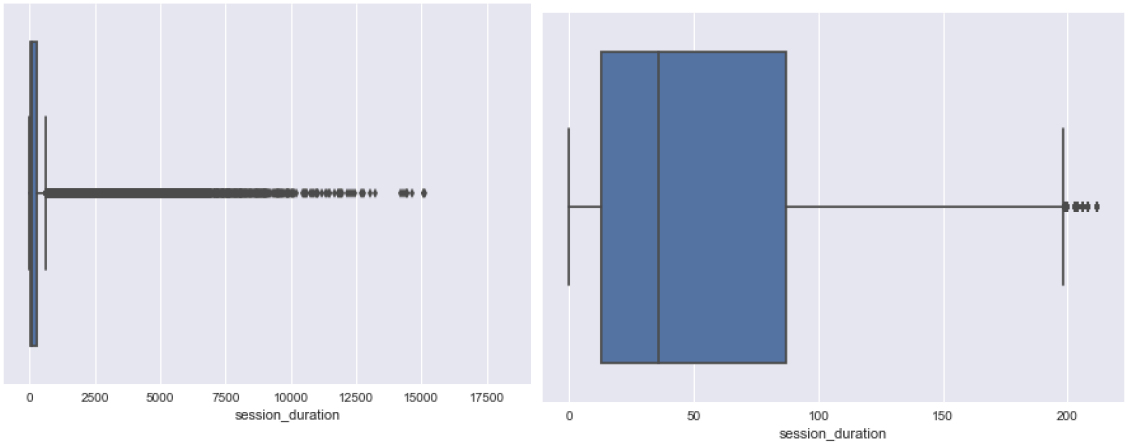
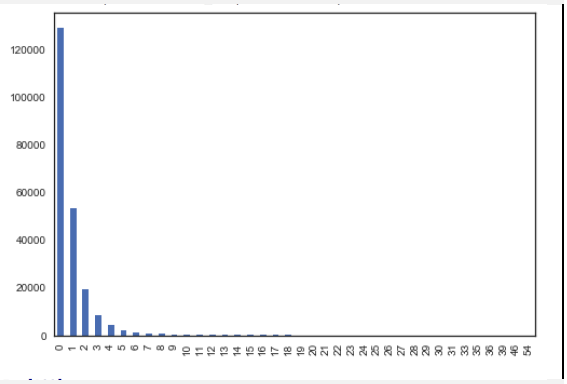
Task 1 - Session investigation

1. Descriptive analysis:

Data Frame	Data for this session consists of 219487 row entries in 17 columns Data types: float64(3), int64(11), object(3)
Date	From 23.05. 2018 - 16.06.2018 (25 unique values) Frequencies range: From 2018-05-23 (5% entries) to 2018-06-09 (3% entries) 
Is_repeat	0 0.606851 1 0.393149
Agent_id	array([18, 20, 4, 16, 14, 12, 2, 26, 28, 22, 24]) Top 5 frequencies account for 88 % of data: **20** 0.385736 **18** 0.229107 **2** 0.181109 **4** 0.048240 **16** 0.040463
Device	mobile 0.617248 desktop 0.312424 tablet 0.070328
Search_ type	**4** 0.666992 **1** 0.192617 **2** 0.125543 **3** 0.014848
Country_ name	**184** unique entries **171** missing values Top 30 countries account for 95% of session data ::::::::::::::::::::::::::: Turkey 0.106864 United Kingdom 0.103426 France 0.103216 Germany 0.069361 Italy 0.058345 
Entry_ page	Only 6 unique values with top 3 accounting for 99% of data **2113** 0.588208 **2111** 0.289188 **2114** 0.113656 2100 0.008821 2106 0.000077 2116 0.000050

Trivago session analysis by Iwona Smith

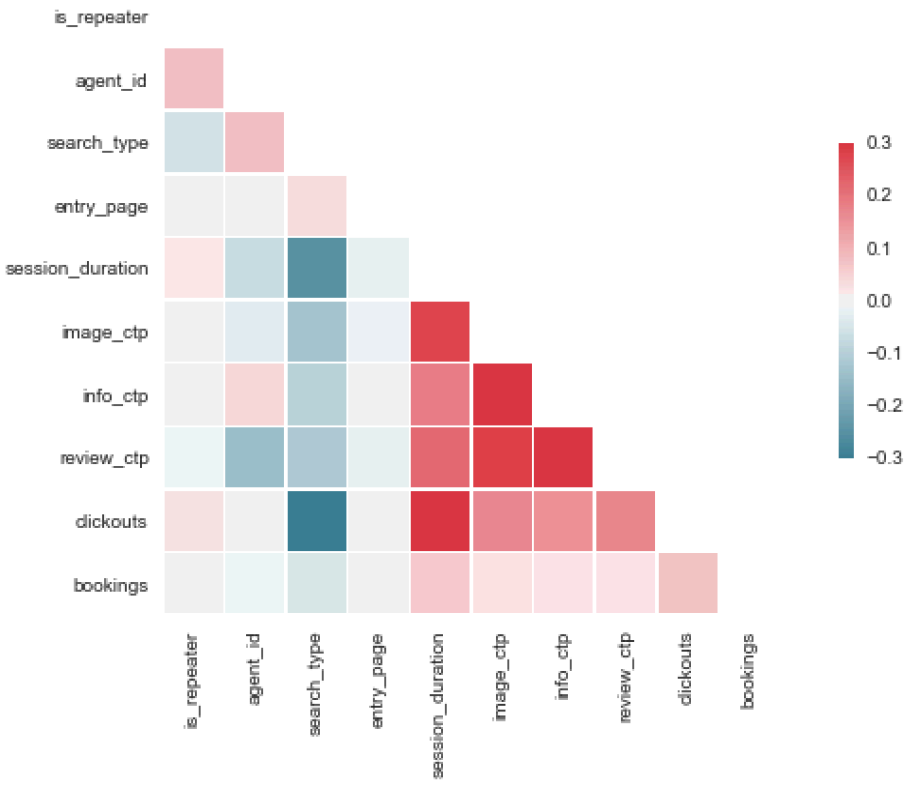
Session_ duration	Mean: 328.311 Median: 76 [1st Q: 18, 3rd : 258] Std: 753.233 Min, Max[0 ,19773] There is a lot of variation in the session duration data. The data are right skewed and there are a lot of outliers. Also we have to question the validity of 0 seconds sessions (4%) . Plot on the right had outliers removed and focuses on sessions that are no more than 200s long 
Ctp	% of null interaction with: image 67% , info 89% and review 89%
Clickouts	Highly skewed data with array ranging from 0 to 54. Top frequencies: 0 0.587114 1 0.242506 2 0.087677 3 0.037802 4 0.018684 
Bookings	The values range from 0:4, but we can ignore anything that is more than 1 as it is not representative. 0 0.997394 1 0.002515

The session data provided includes information on users visiting Trivago website for 25 consecutive days in May and June 2018. Users are more likely to be new to the website (60%) and they are mostly using mobiles (62%) to access the page. Again from looking at the search_type data we can identify 60% of users to be searching for similar type of hotel. There is very little interaction with the content, presented in form of images, information or reviews, related to hotels. Session_duration and clickouts data is very skewed, but we can see that the visit is likely to converge around 15s and most commonly a session will result in up to 2 clicks to partnering websites.

In general terms a user of the Trivago website wants an effective tool that can perform accurate search and display relevant information fast. That is why the website is very simple looking and it is easy to navigate helping the visitor to focus on the task ahead. There are only few actions that the potential user needs to perform to find desired hotel. It is important that the search results are accurate so the user can find what is needed. Knowing who the people are that are performing the hotel search is key to successful convergence.

All this session information on users aligns with Trivago's business model as the main focus is to channel the traffic to a partner booking page where transaction can finalised. The engagement with Trivago page should not be long as user expects to be presented with an accurate result that meets his/her needs.

When performing analysis, I like to look at the correlation between variables in data set hence the presentation of heat map of all the variables given.

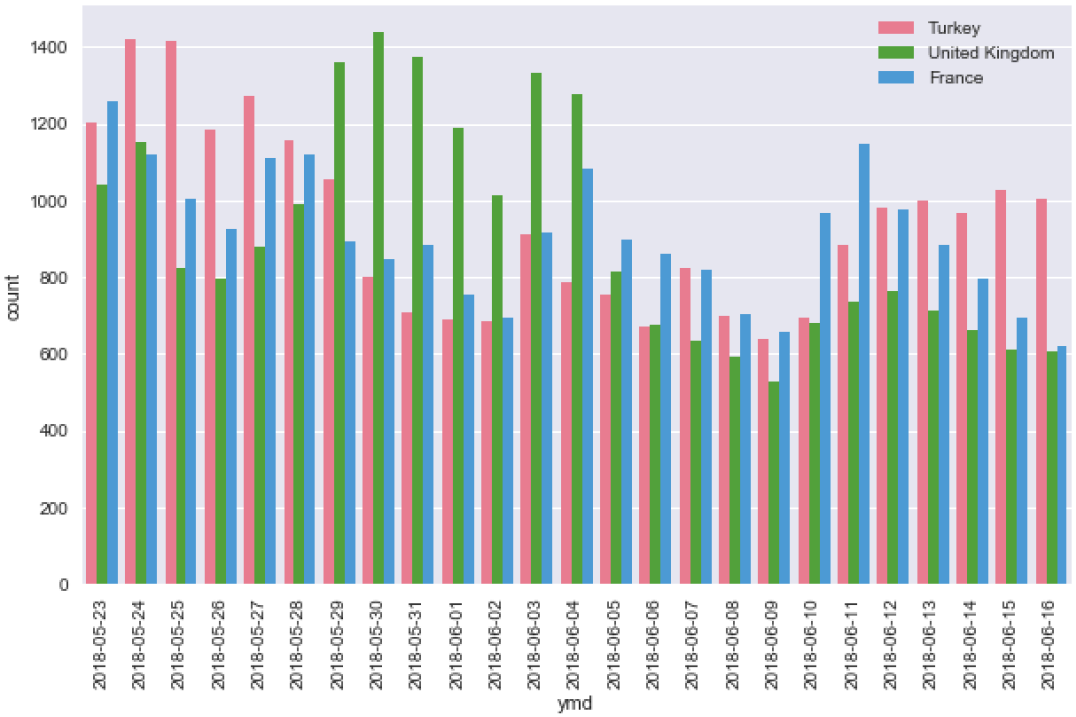


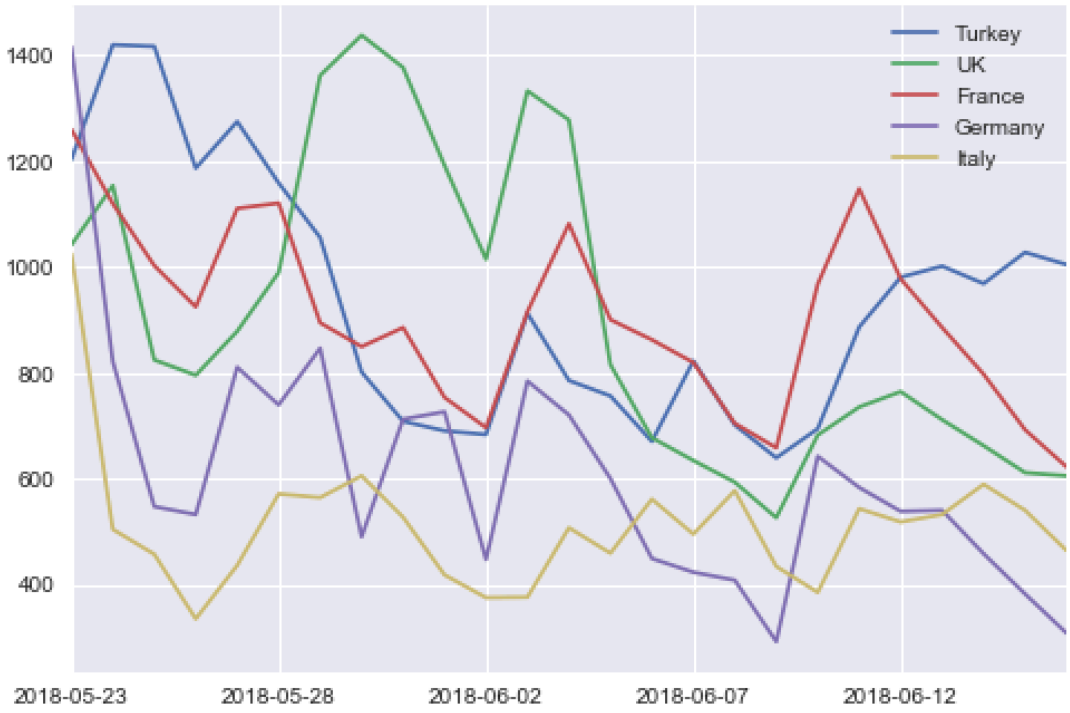
In terms of users, for me it would be important to know if a user has visited the page before. Having the information on returning user can enhance the performance and result in better outcome. Also the interaction with the website is very different depending on the device, operating system and web browser as this impacts the usability and influences user engagement. Understanding the category of user based on type of hotel they search for can be critical to achieving the convergence.

2. KPIs

(I) Engagement rate by country of origin

Thinking about the user the first thing I want to know is where is the user coming from. Looking at data set given we have 184 countries with most common visits paid by users from Turkey, United Kingdom and France. Top 30 countries account for account for 95% of data and if we focus on top 10 countries we account for 65% of data. Below we have a graph representing a count of users from Turkey, UK and France developed over duration of the session. We can see that in the first part of the session users from Turkey dominate, but this changes 6 days later and U.K users are more frequent. Finally in the last quarter French are more engaged with the website. Unfortunately the sample data given does not allow us to go deeper in understanding the trend. I would like to model this data by day of the week to see the trends.

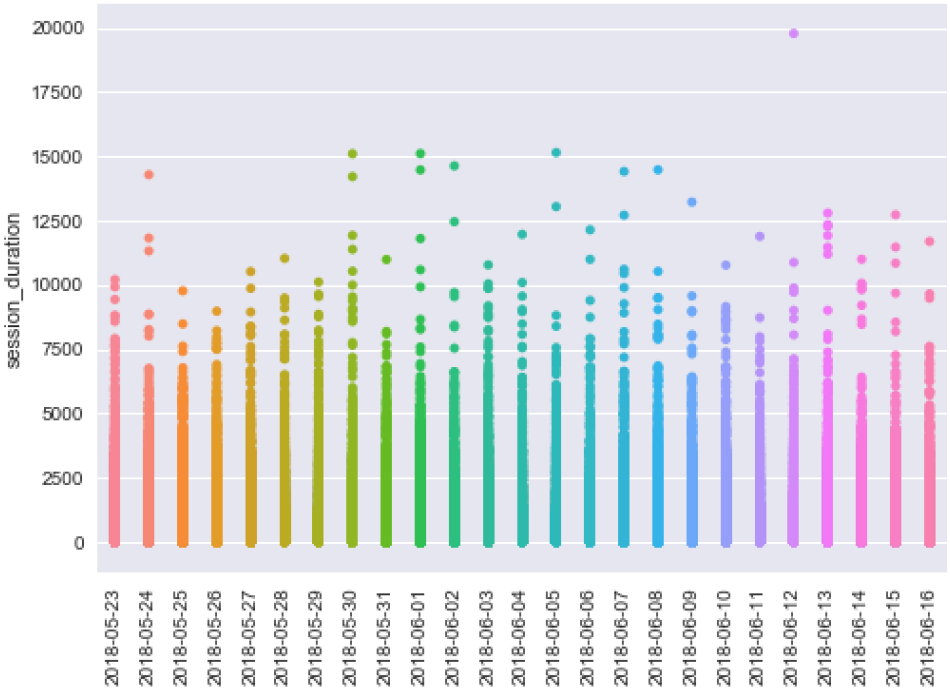




(II) Optimal search duration

Normally for web platforms we measure performance of our website by how long user engaged with it. And longer means better. For Trivago website this probably not a measure of success, but still can provide some interesting insight to keep a user engaged and at the same time channel him/her to converge.

Analysing the session data I noticed there a lot of outliers with session time much greater than expected. This probably indicates that there is an automated user viewing content on Trivago website. It would be good to have more information on consecutive pages that are visited by user with longer than the average session duration. If the search doesn't follow a certain pattern, but instead visitor enters every content possible this should be classified as a bot. It is important to distinguish between real users that will converge and someone/thing with other purposes.



(III) Classifying users

There is more than one quality/feature of a user that can improve our performance. Firstly, if there was previous interaction with the website is very valuable, but understanding the type of hotel our user is likely to book is critical for Trivago. Our session is dominated, by one category of user, but his/her presence on the website will probably vary depending on the time of the year. Knowing when this user is coming could allow Trivago to work with partners to prepare more suitable content.