trivago

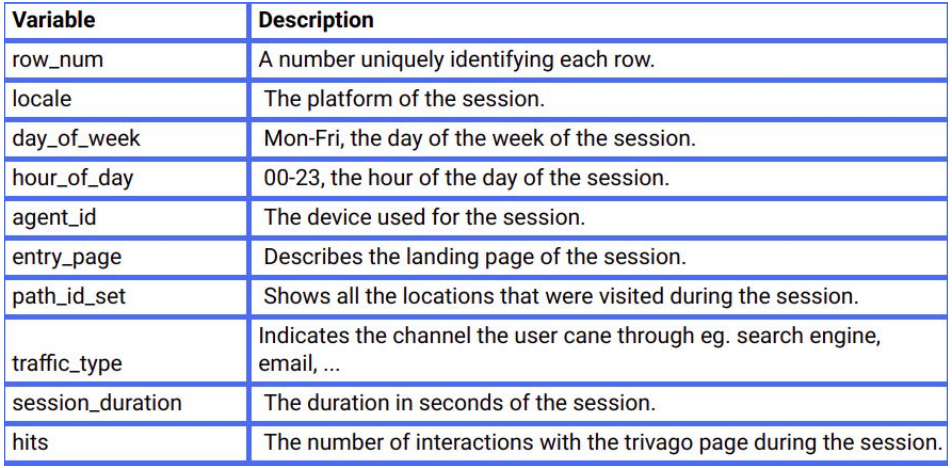
Building a “Hits” Predictive Model

Shelmith Nyagathiri Kariuki

Data Description

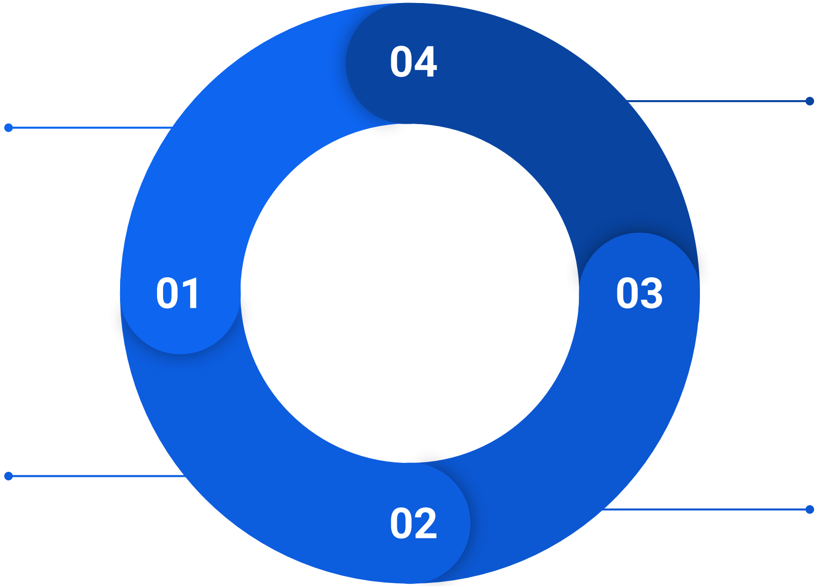
- The challenge was to build a predictive model that predicts the number of hits per session.

- The data consisted of 988,681 records and 10 variables.



Variable	Description
row_num	A number uniquely identifying each row.
locale	The platform of the session.
day_of_week	Mon-Fri, the day of the week of the session.
hour_of_day	00-23, the hour of the day of the session.
agent_id	The device used for the session.
entry_page	Describes the landing page of the session.
path_id_set	Shows all the locations that were visited during the session.
traffic_type	Indicates the channel the user cane through eg. search engine, email, ...
session_duration	The duration in seconds of the session.
hits	The number of interactions with the trivago page during the session.

Analysis Process



Cleaning

Converting string variables to factors, and numeric variables to actual numerics

Replacing “\N” with actual NAs

Feature Engineering

Generating additional variables from the ones that we already have

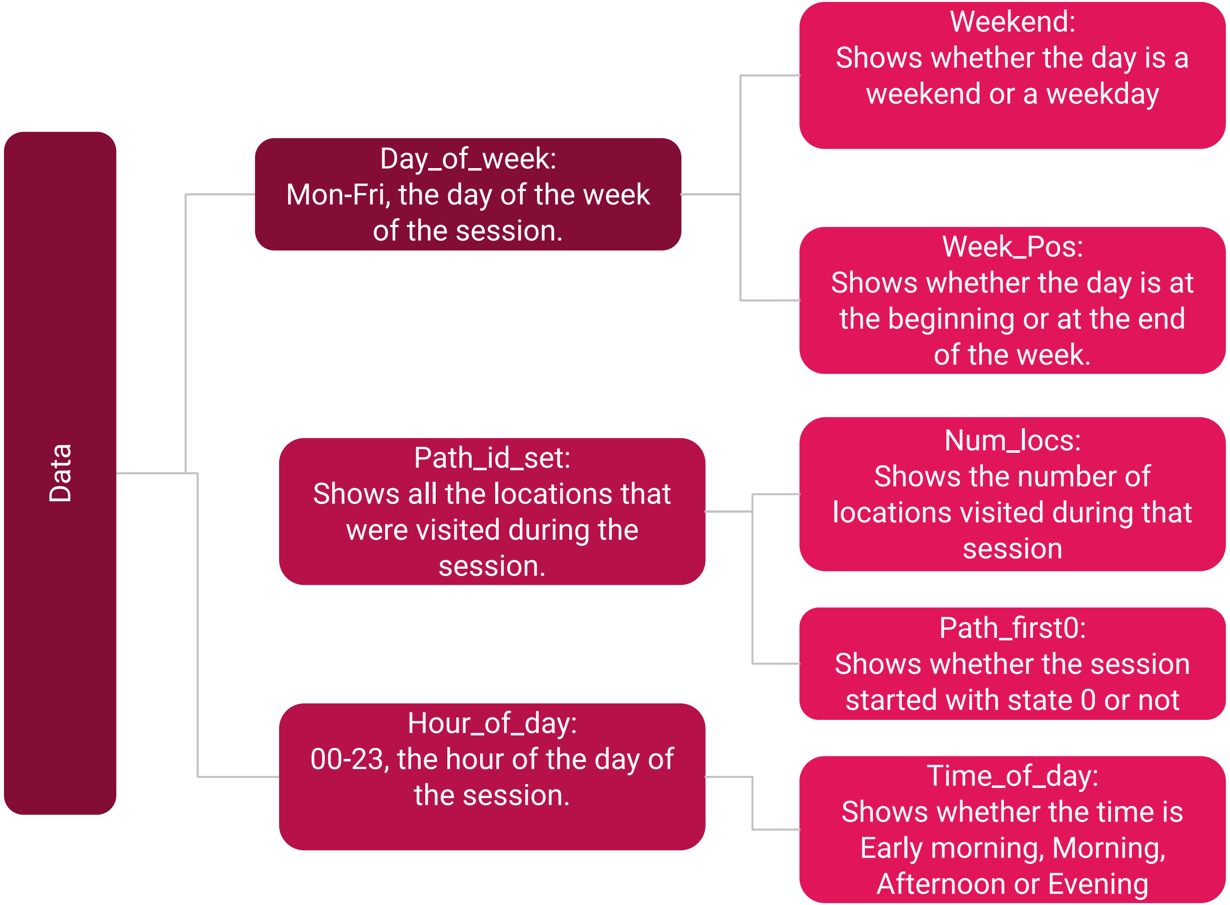
Modeling

Fitting the SVM and Random Forest models, and obtaining predicted values.

Data Exploration

Assessing relationships among the variables, and between the variables and the dependent variable

Feature Engineering



Data

Day_of_week:
Mon-Fri, the day of the week of the session.

Weekend:
Shows whether the day is a weekend or a weekday

Week_Pos:
Shows whether the day is at the beginning or at the end of the week.

Path_id_set:
Shows all the locations that were visited during the session.

Num_locs:
Shows the number of locations visited during that session

Path_first0:
Shows whether the session started with state 0 or not

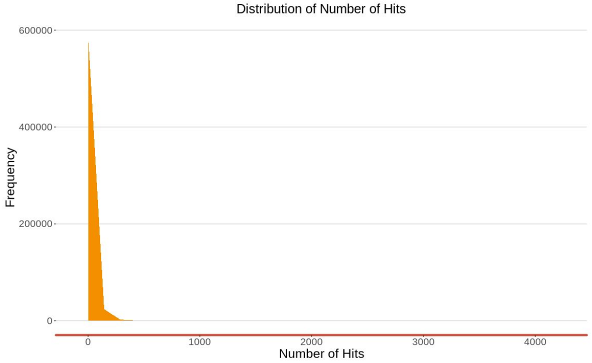
Hour_of_day:
00-23, the hour of the day of the session.

Time_of_day:
Shows whether the time is Early morning, Morning, Afternoon or Evening

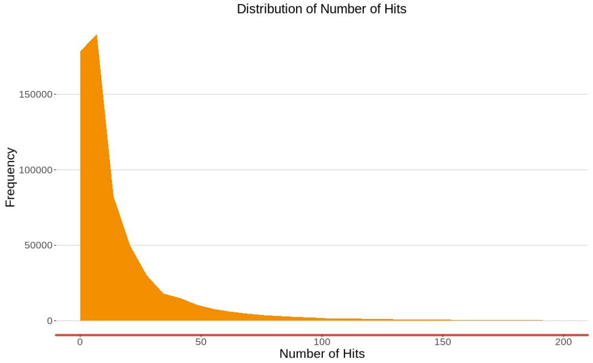
Data Exploration

Number of Hits

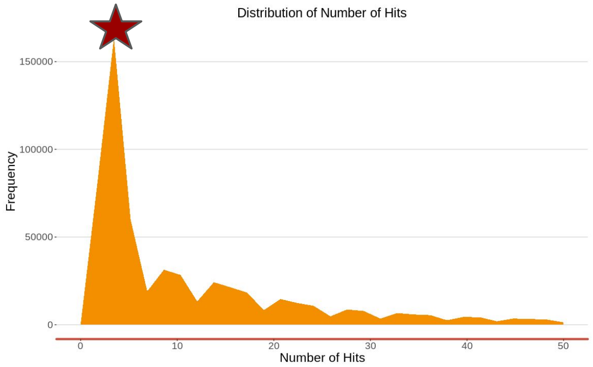
- A majority of visitors on the Trivago website make an average of about 5-10 hits per session, as seen on the third graph.


Distribution of Number of Hits

The whole dataset


Distribution of Number of Hits

Data truncated to only 200 hits

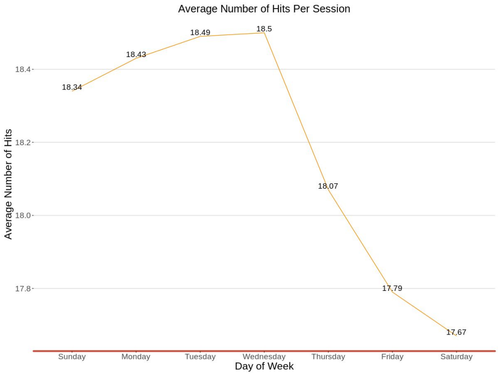

Distribution of Number of Hits

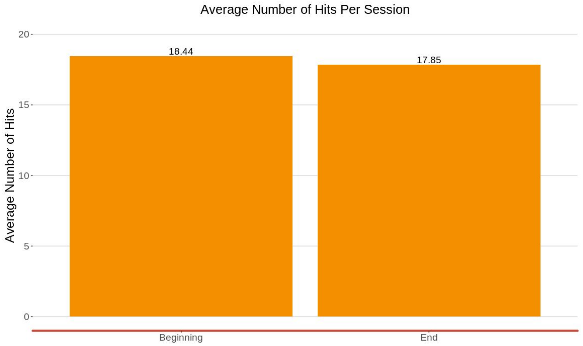
Data truncated to only 50 hits

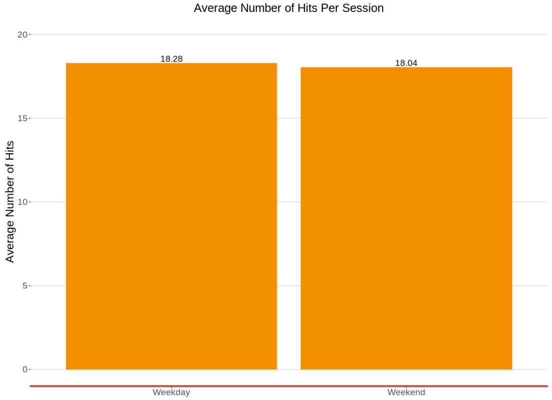
Days of the week

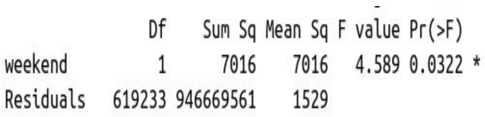
- There is an upward trend in the average number of hits as the week progresses, but this decreases as the week comes to an end.

- There is a **significantly** higher number of hits during the weekdays, as opposed to during the weekends.

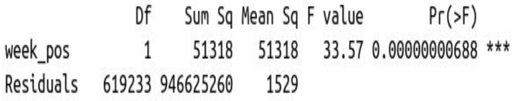








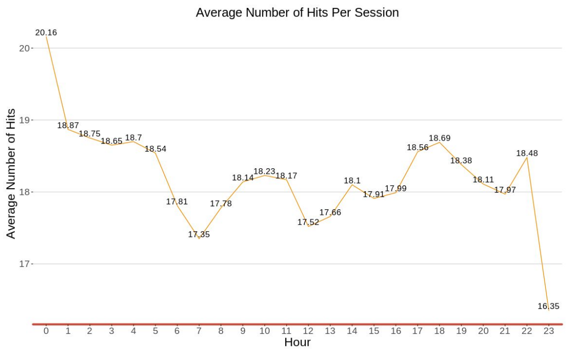
```
              Df     Sum Sq Mean Sq F value Pr(>F)
weekend        1       7016    7016   4.589 0.0322 *
Residuals 619233 946669561    1529
```

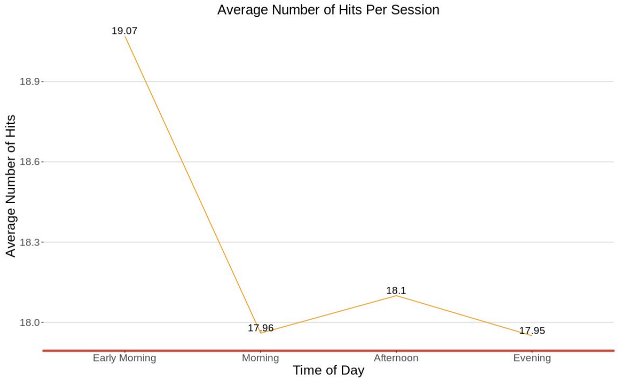


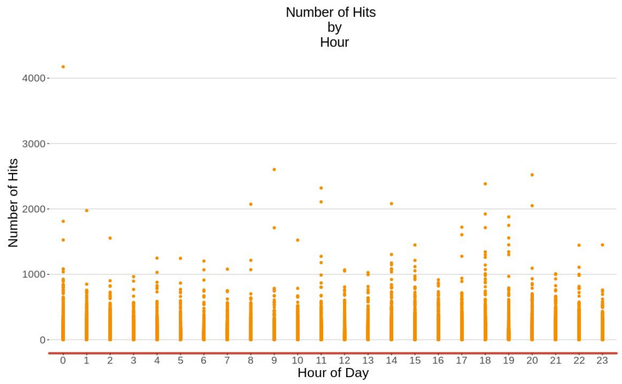
```
              Df     Sum Sq Mean Sq F value         Pr(>F)
week_pos       1      51318   51318   33.57 0.00000000688 ***
Residuals 619233 946625260    1529
```

Time of day

- The number of hits is relatively high at the beginning of each day.
- There is a very low negative correlation (0.00796667) between the number of hits and the time of day.

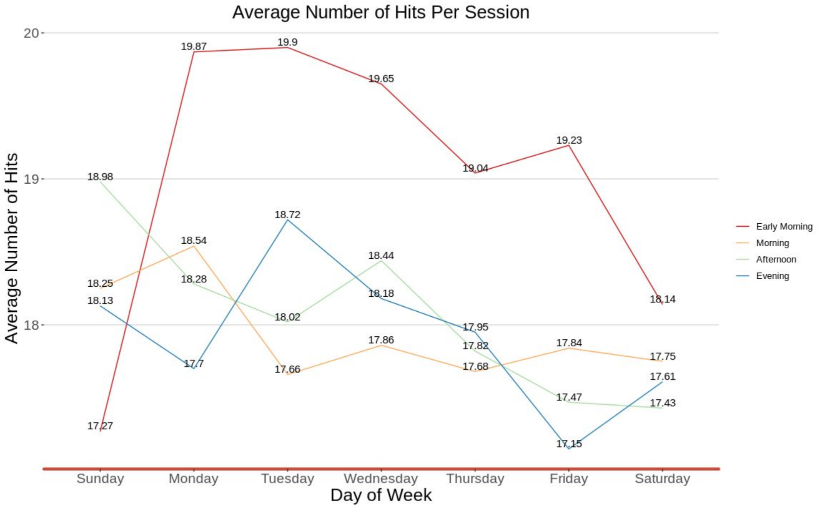


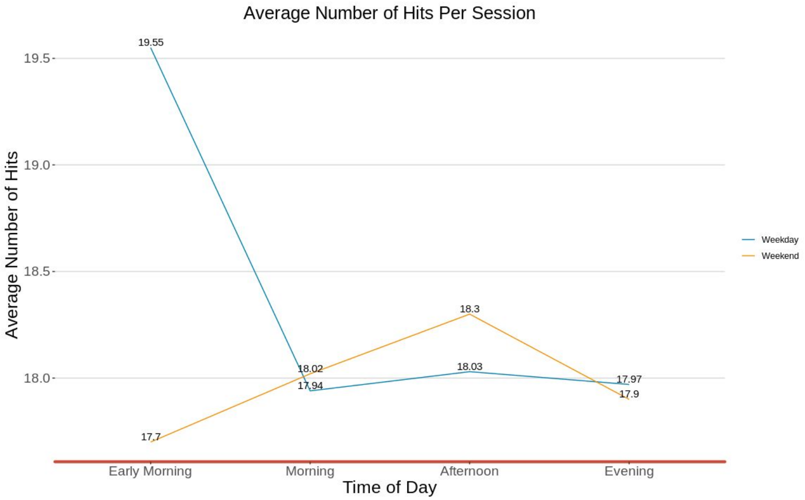




Time of day: Cont’d

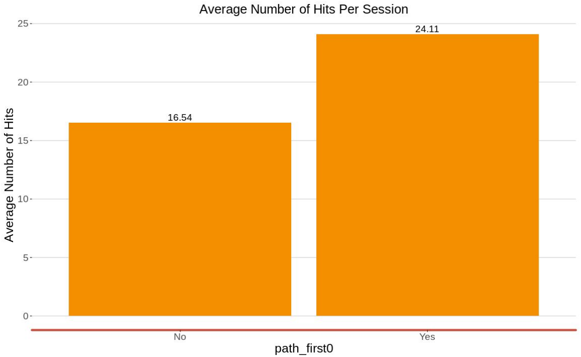
- Generally, there is higher traffic on the Trivago Website early in the morning (i.e between midnight and 5:00 am), as compared to other times.
- Traffic is higher during weekend afternoons as compared to weekday afternoons. This may be due to the fact that during weekdays, people are working during the day,that is why they visit the website either during early mornings, or in the evenings(i.e between 7:00 pm and 11:00pm).

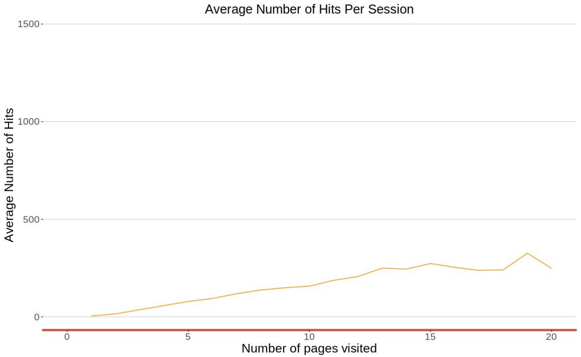


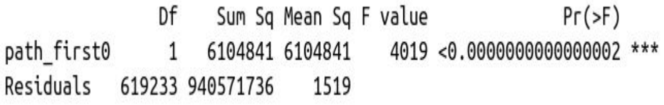


Locations visited

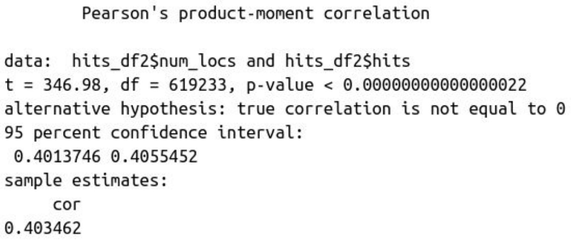
- There is a **moderate positive correlation(0.403462)** between the average number of hits and the number of pages visited per session .
- There is a higher number of hits for the sessions that start with path id 0, as compared to those that start with other path ids.







```
                Df    Sum Sq Mean Sq F value                Pr(>F)
path_first0      1   6104841 6104841    4019 <0.0000000000000002 ***
Residuals   619233 940571736    1519
```

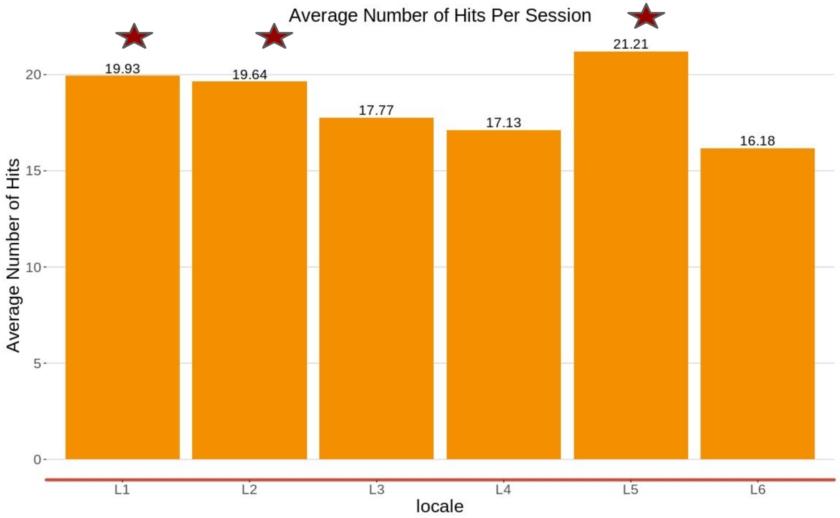


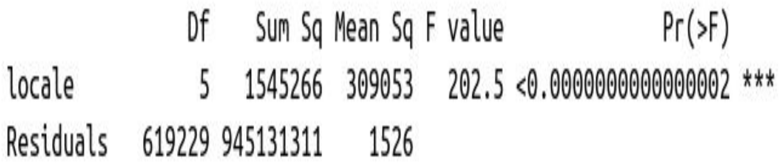
```
	Pearson's product-moment correlation

data:  hits_df2$num_locs and hits_df2$hits
t = 346.98, df = 619233, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4013746 0.4055452
sample estimates:
     cor
0.403462
```

Platforms

- L1, L2 and L5 result into a higher number of hits, as opposed to the rest of the platforms.

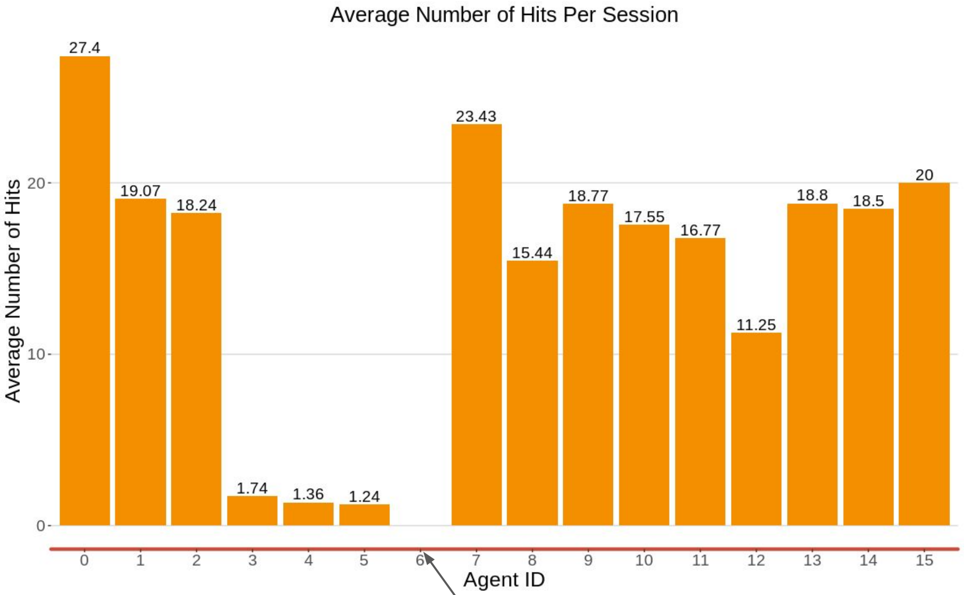




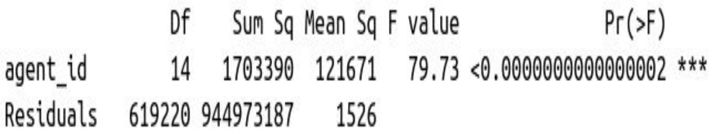
```
                Df    Sum Sq Mean Sq F value                Pr(>F)
locale           5   1545266  309053   202.5 <0.0000000000000002 ***
Residuals   619229 945131311    1526
```

Agent IDs

-



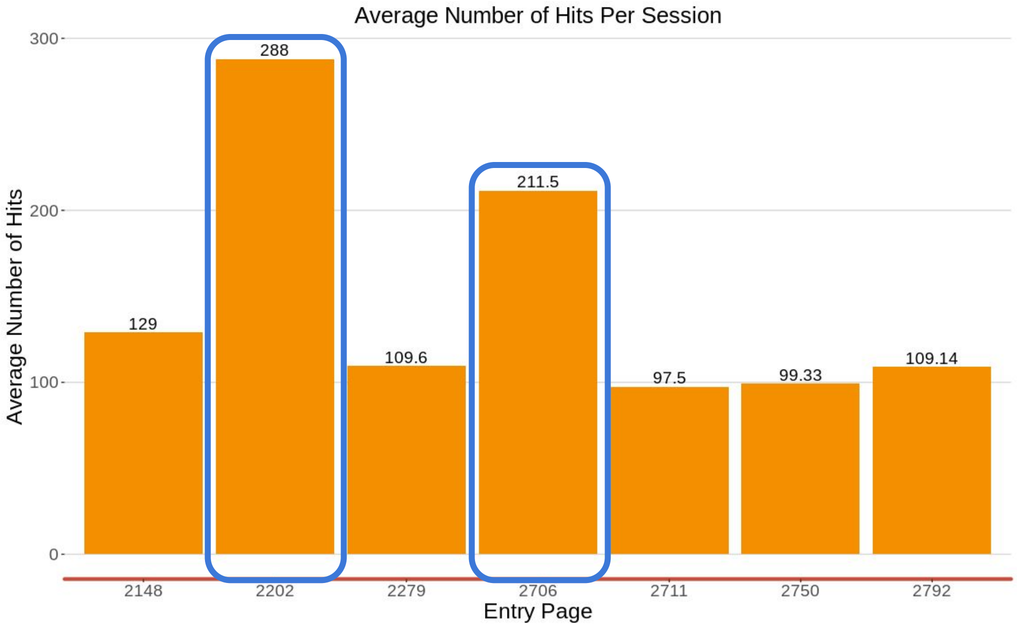
All the hits in the records of AgentId6 are missing



```
              Df    Sum Sq Mean Sq F value              Pr(>F)
agent_id      14   1703390  121671   79.73 <0.0000000000000002 ***
Residuals 619220 944973187    1526
```

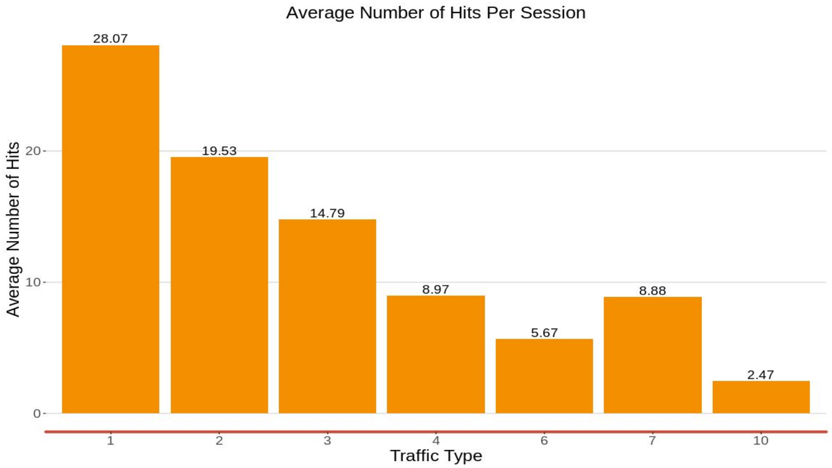
Entry Page

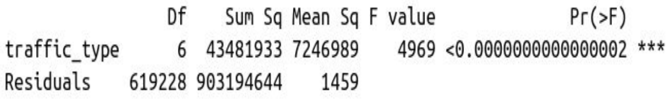
- The entry pages that result into a higher number of hits are 2202 and 2706.



Traffic

- The lesser the value of the traffic type, the higher the number of hits per session.
- I can confidently say that the higher the value of traffic type, the lesser the engagement on the website.



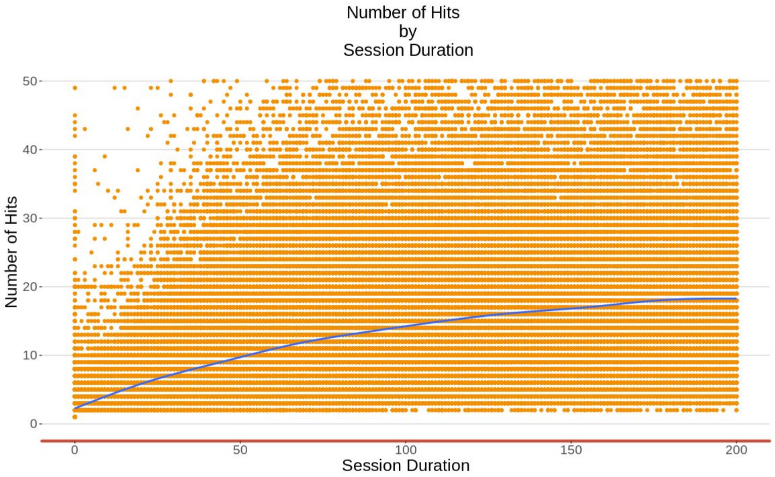


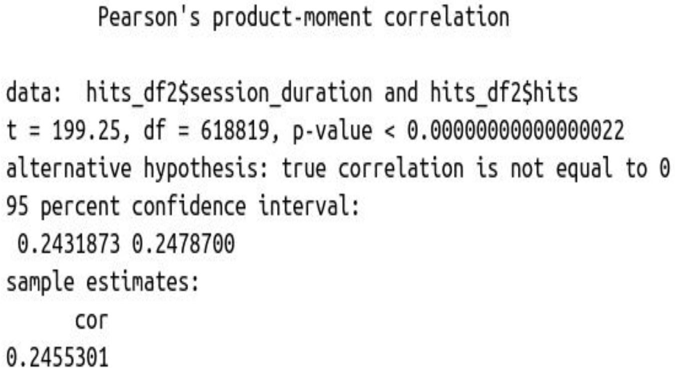
```
                Df    Sum Sq Mean Sq F value              Pr(>F)    
traffic_type      6  43481933 7246989    4969 <0.0000000000000002 ***
Residuals    619228 903194644    1459                                
```

Here, I assumed that the traffic type is coded in such a way that 1 represents higher traffic on the website, and 10 represents lesser traffic, as this behavior is depicted on the graph.

Session Duration

- There is a **moderate positive relationship(0.2455381)** between the session duration and the number of hits per session.



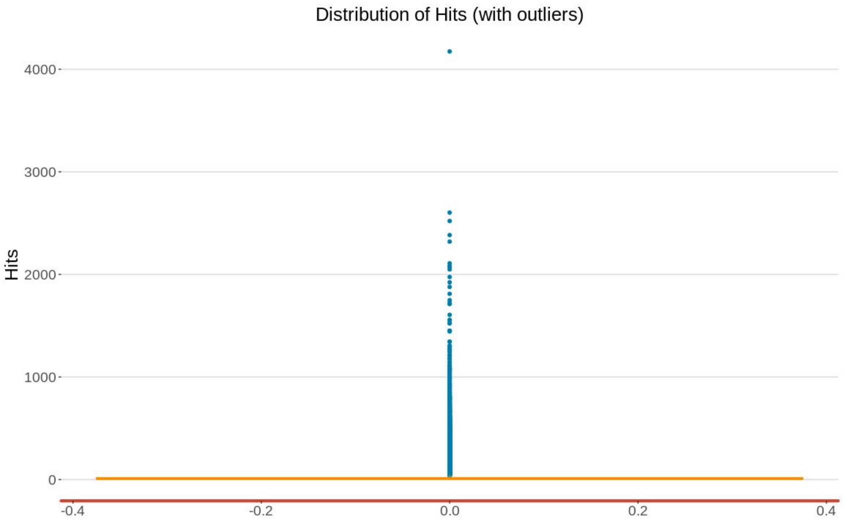


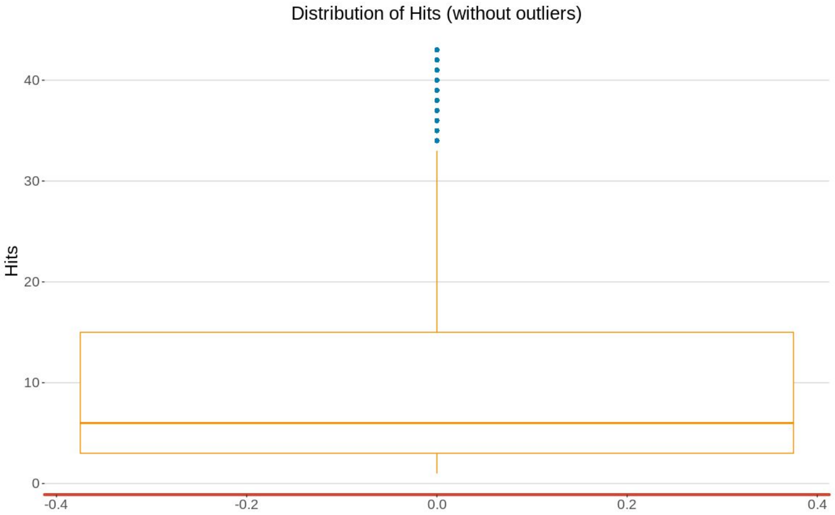
```
                Pearson's product-moment correlation

data:  hits_df2$session_duration and hits_df2$hits
t = 199.25, df = 618819, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2431873 0.2478700
sample estimates:
      cor
0.2455301
```

Predictive Model

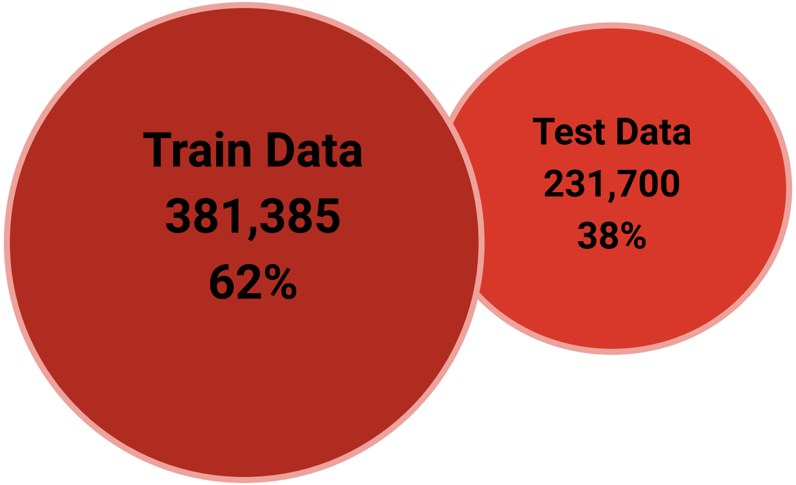
Examining Outliers





374,732 outliers were dropped at this point

Splitting the data into train and test

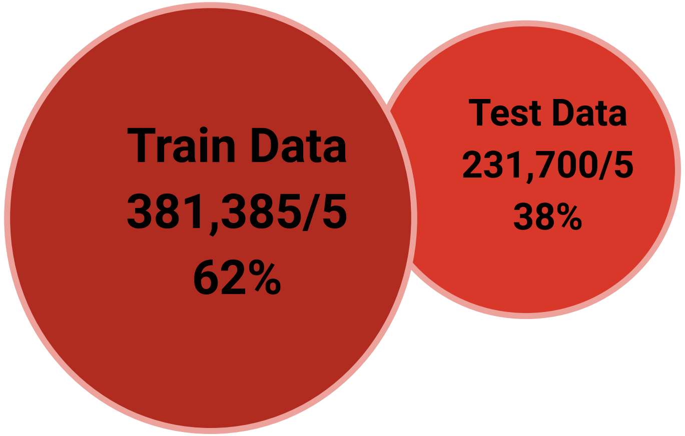


Train Data
381,385
62%

Test Data
231,700
38%

Alert!!

- My laptop’s computation power could not handle the whole dataset, so I reduced the size of my train and test dataset by **5**.



Train Data

381,385/5

62%

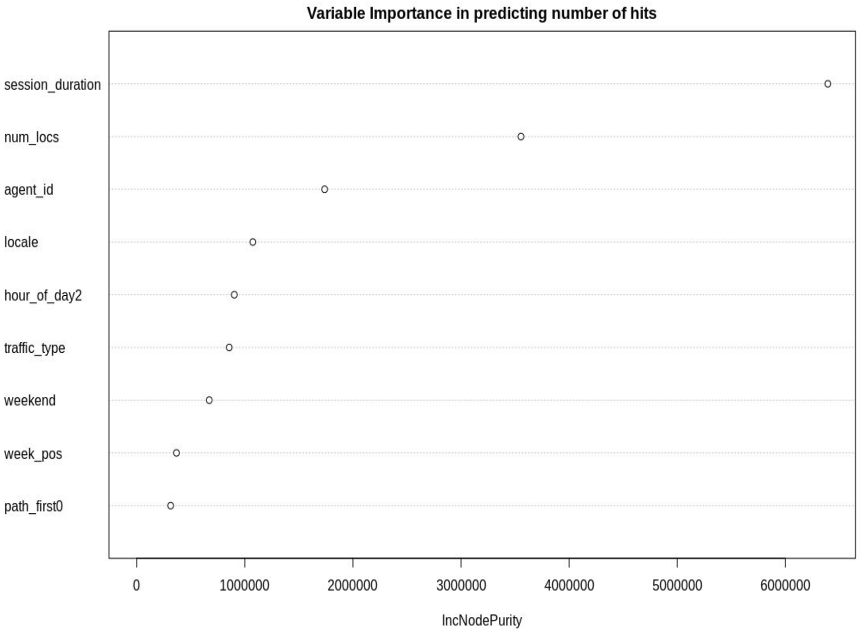
Test Data

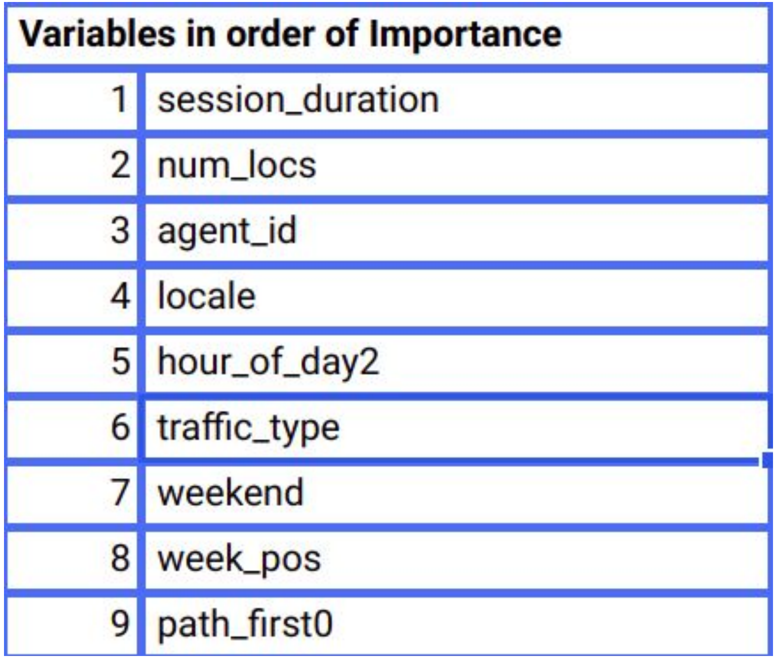
231,700/5

38%

Variable Importance

trivago





Variables in order of Importance	
1	session_duration
2	num_locs
3	agent_id
4	locale
5	hour_of_day2
6	traffic_type
7	weekend
8	week_pos
9	path_first0

Random Forest

- See data attached together with this presentation deck.

Support Vector Machines

- See data attached together with this presentation deck.

trivago

The End!