a

Contents

[Introduction of the business problem 2](#_Toc96895739)

[Defining problem statement 2](#_Toc96895740)

[Need of the study/project 2](#_Toc96895741)

[Understanding business/social opportunity 2](#_Toc96895742)

[Data Report 2](#_Toc96895743)

[Understanding how data was collected in terms of time, frequency and methodology 2](#_Toc96895744)

[Visual inspection of data (rows, columns, descriptive details) 3](#_Toc96895745)

[Understanding of attributes (variable info, renaming if required) 3](#_Toc96895746)

[Data type: 3](#_Toc96895747)

[Duplicates: 4](#_Toc96895748)

[Null Values : 4](#_Toc96895749)

[Exploratory data analysis 4](#_Toc96895750)

[Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) 4](#_Toc96895751)

[Categorial Variables 4](#_Toc96895752)

[Taken product 4](#_Toc96895753)

[Preferred device 5](#_Toc96895754)

[Member in family 5](#_Toc96895755)

[Preferred location type 6](#_Toc96895756)

[Continuous variables 8](#_Toc96895757)

[Bivariate analysis (relationship between different variables, correlations) 13](#_Toc96895758)

[Categorical variables 13](#_Toc96895759)

[Continuous variables 16](#_Toc96895760)

[Correlation heat map 19](#_Toc96895761)

[Removal of unwanted variables (if applicable) 19](#_Toc96895762)

[Missing Value treatment (if applicable) 19](#_Toc96895763)

[Outlier treatment (if required) 20](#_Toc96895764)

[Variable transformation (if applicable) 20](#_Toc96895765)

[Addition of new variables (if required) 21](#_Toc96895766)

[Business insights from EDA 21](#_Toc96895767)

[Is the data unbalanced? If so, what can be done? Please explain in the context of the business 21](#_Toc96895768)

[Any business insights using clustering  (if applicable) 22](#_Toc96895769)

[Any other business insights 22](#_Toc96895770)

# **Introduction of the business problem**

## Defining problem statement

An aviation company that provides domestic as well as international trips to the customers now wants to apply a targeted approach instead of reaching out to each of the customers. This time they want to do it digitally instead of tele calling. Hence, they have collaborated with a social networking platform, so they can learn the digital and social behaviour of the customers and provide the digital advertisement on the user page of the targeted customers who have a high propensity to take up the product.

## Need of the study/project

The advertisements on the digital platform are a bit expensive; hence, you need to be very accurate while creating the models.

2 different models will be created in order to cater to People who use different devices Mainly Laptop and Mobile as primary categories.

## Understanding business/social opportunity

The model can help us predict with high accuracy the capability and the willingness of an individual to book an international flight ticket. This will help the companies target these customers with a higher chance of success and spent the marketing budget more effectively. This data can also be used for other integrated industries such hotels, restaurant booking, travel packages etc.

# **Data Report**

## Understanding how data was collected in terms of time, frequency and methodology

The data is mostly collected from the user’s social media account. It is either purchased directly from the app or a third party.

The name and personal details of the user are not disposed. Assuming the reason of privacy

The data must be over a span of a year as many variables are regarding ‘avg. over the year’.

## Visual inspection of data (rows, columns, descriptive details)

The data contains 11760 rows and 17 columns

Description of the variables used are as follows

(Columns names have been cleaned for purpose of reading)

|  |  |
| --- | --- |
| User ID | Unique ID of user |
| Taken product | Buy ticket in next month |
| Yearly avg. view on travel page | Average yearly views on any travel related page by user |
| preferred device | Through which device user preferred to do login |
| total likes on outstation checkin given | Total number of likes given by a user on out of station checkings in last year |
| yearly avg. Outstation checkins | Average number of out of station check-in done by user |
| member in family | Total number of relationships mentioned by user in the ac |
| preferred location type | Preferred type of the location for travelling of user |
| Yearly avg. comment on travel page | Average yearly comments on any travel related page by user |
| total likes on out of station checkin received | Total number of likes received by a user on out of station checkings in last year |
| week since last outstation checkin | Number of weeks since last out of station check-in update by user |
| following company page | Weather the customer is following company page (Yes or No) |
| monthly avg. comment on company page | Average monthly comments on company page by user |
| working flag | Weather the customer is working or not |
| travelling network rating | Does user have close friends who also like travelling. 1 is highs and 4 is lowest |
| Adult flag | Weather the customer is adult or not |
| Daily Avg mins spend on traveling page | Average time spend on the company page by user on daily basis |

## Understanding of attributes (variable info, renaming if required)

### Data type:

At initial glance the data types of variables are as follows

User ID – int64

Taken product - object

Yearly avg. view on travel page – float64

preferred device - object

total likes on outstation checkin given - float64

yearly avg. Outstation checkins - object

member in family - object

preferred location type - object

Yearly avg. comment on travel page - float64

total likes on out of station checkin received - int64

week since last outstation checkin - int64

following company page - object

monthly avg. comment on company page - int64

working flag - object

travelling network rating - int64

Adult flag - int64

Daily Avg mins spend on traveling page - int64

### Duplicates:

The data does not contain any duplicate

### Null Values :

The variables with null values are as follows

Yearly avg. view on travel page - 581

preferred device - 53

total likes on outstation checkin given - 381

yearly avg. Outstation checkins - 75

preferred location type - 31

Yearly avg. comment on travel page - 206

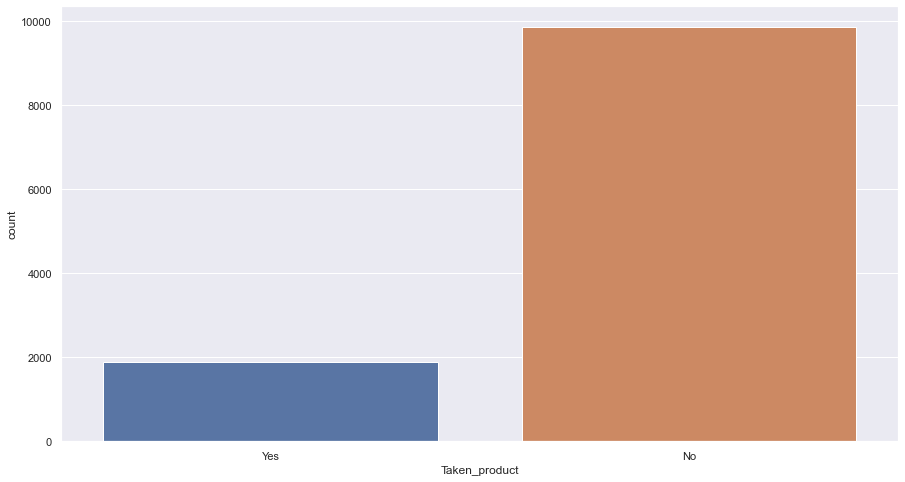
following company page - 103

# **Exploratory data analysis**

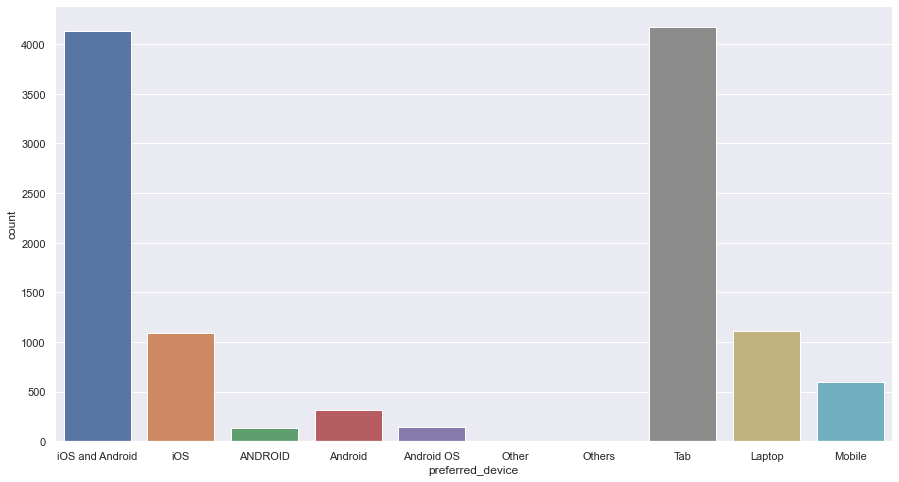
## Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

## Categorial Variables

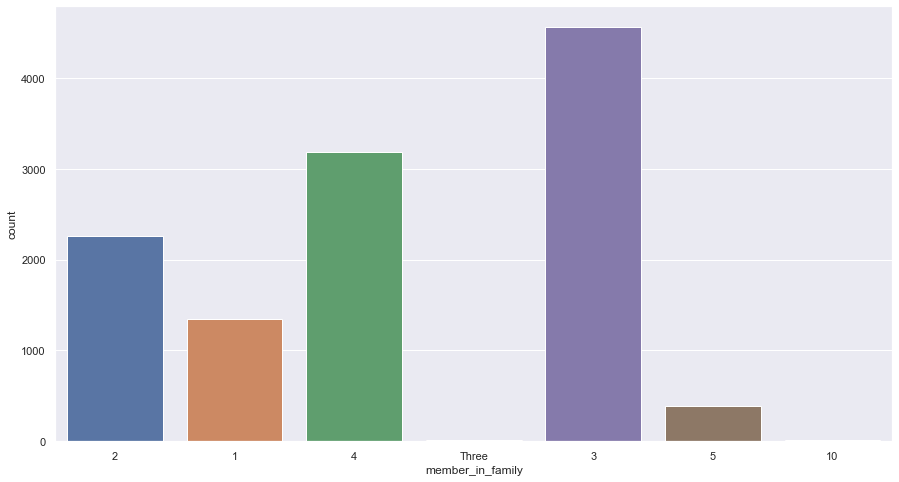
### Taken product



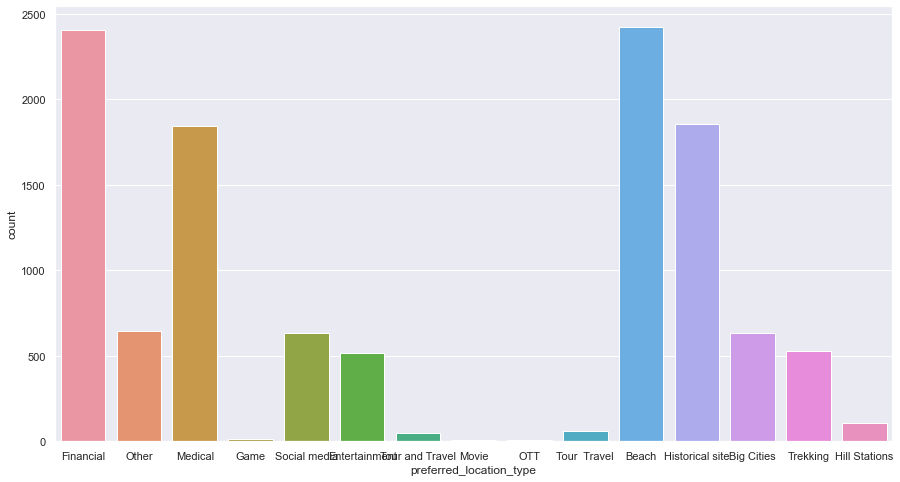
### Preferred device



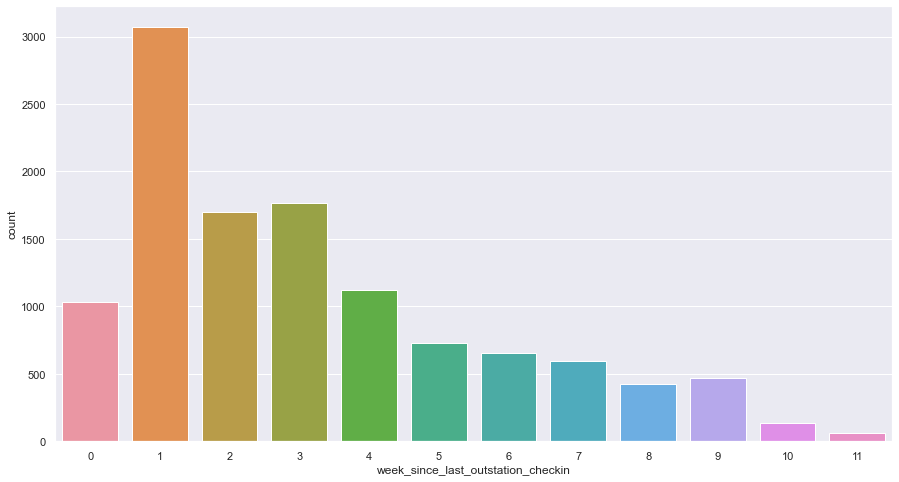
### Member in family



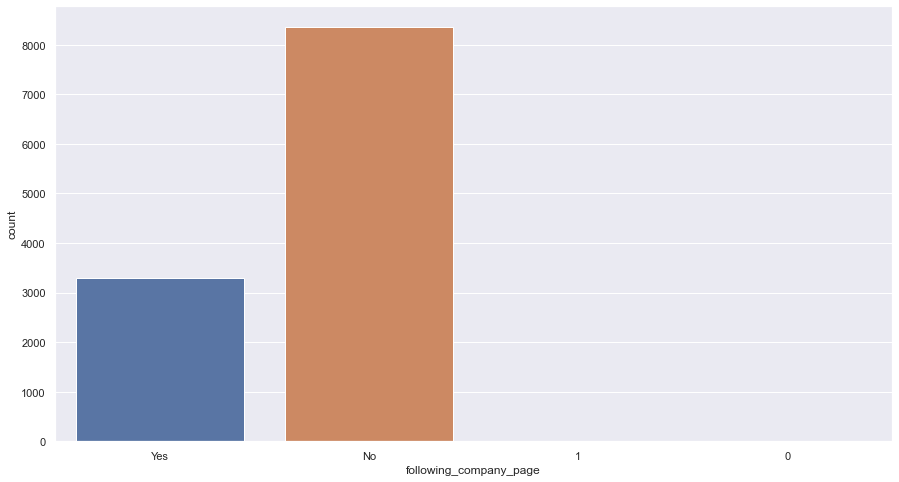
### Preferred location type



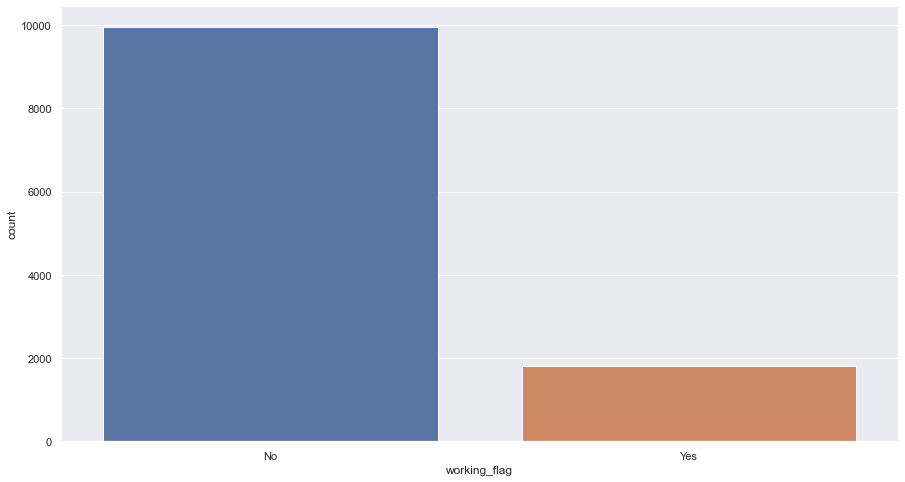
Week since last outstation checkin



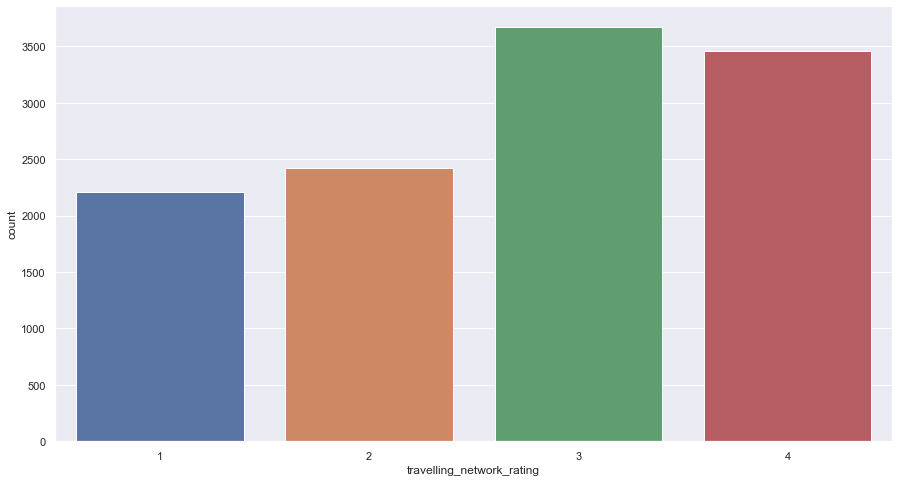
Following company page



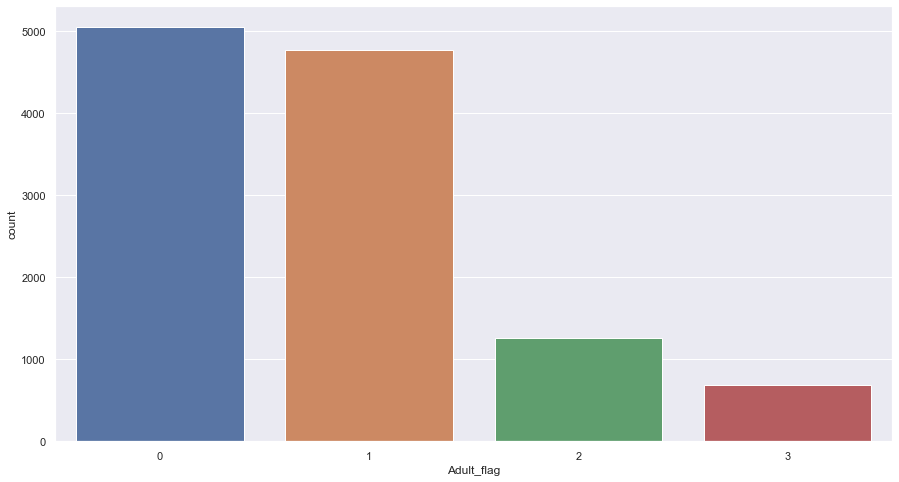
Working flag



Travelling network rating



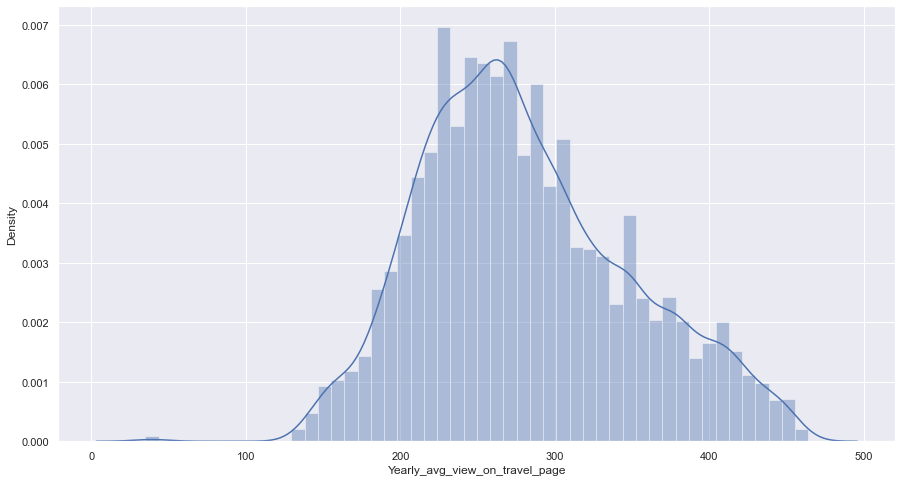
Adult flag



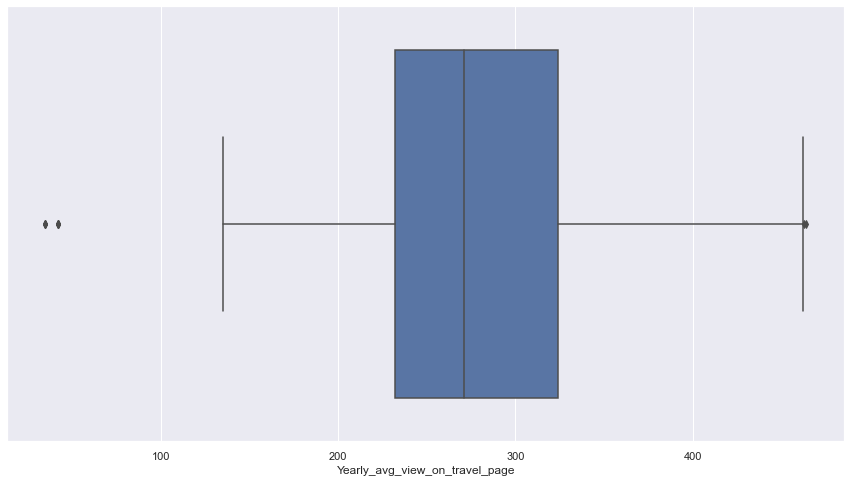
## Continuous variables

Yearly avg. view on travel page

plot

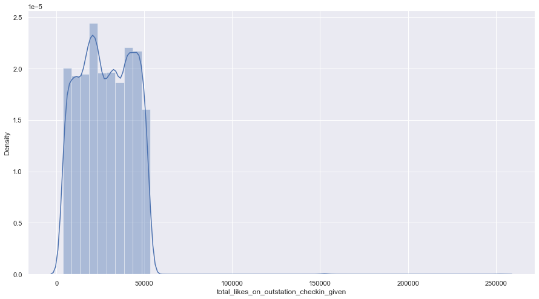


Box plot

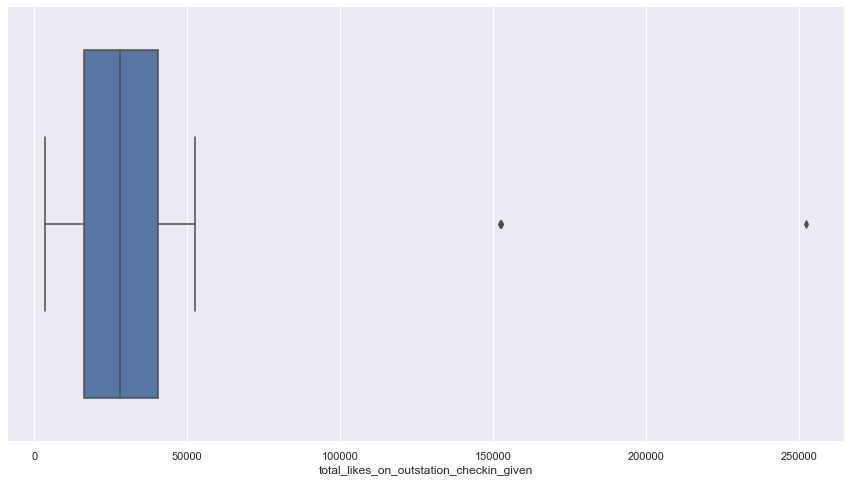


total likes on outstation checkin given

plot

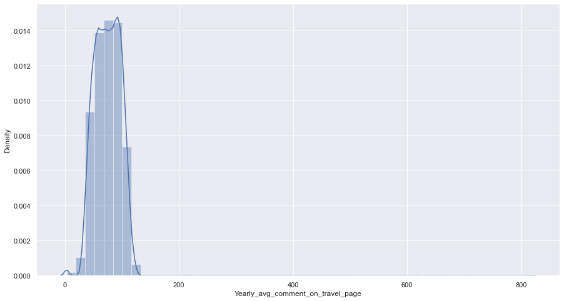


Box plot



Yearly avg. comment on travel page

plot

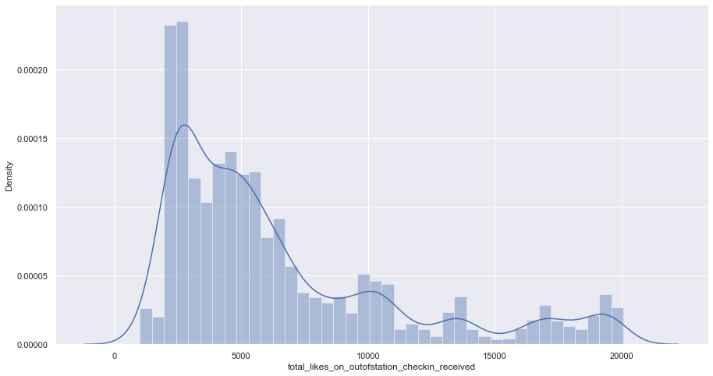


Box plot

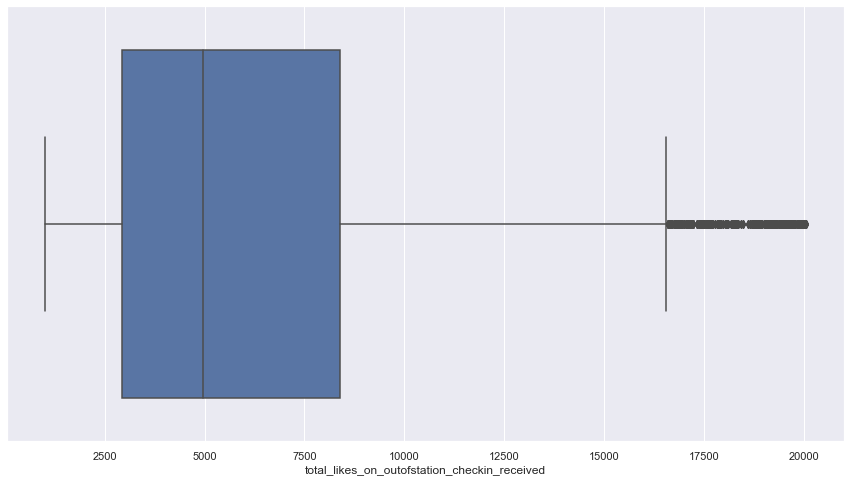


total likes on out of station checkin received

plot

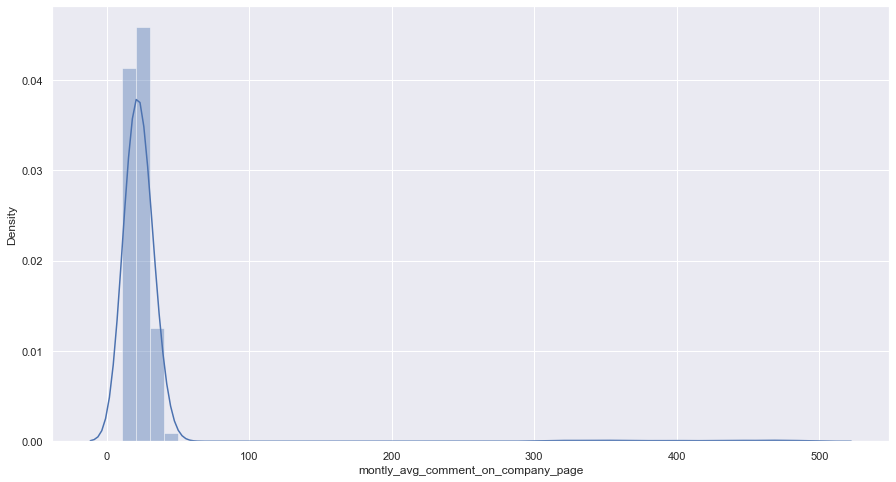


Box plot

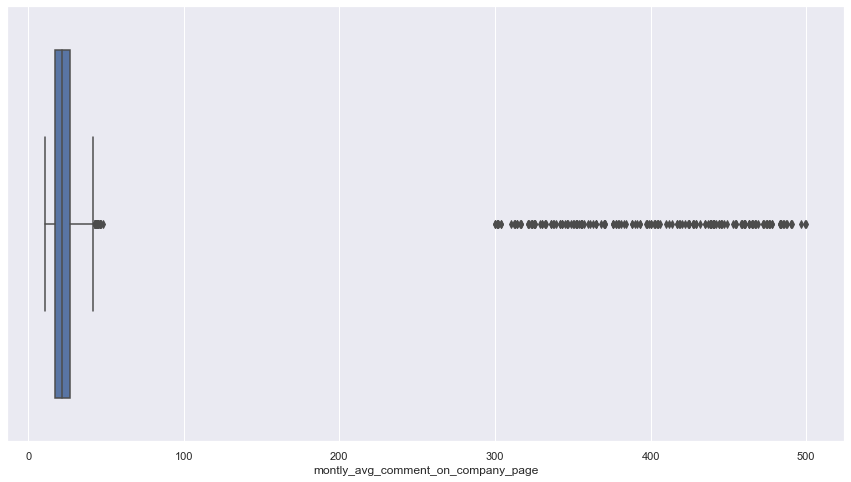


monthly avg. comment on company page

plot

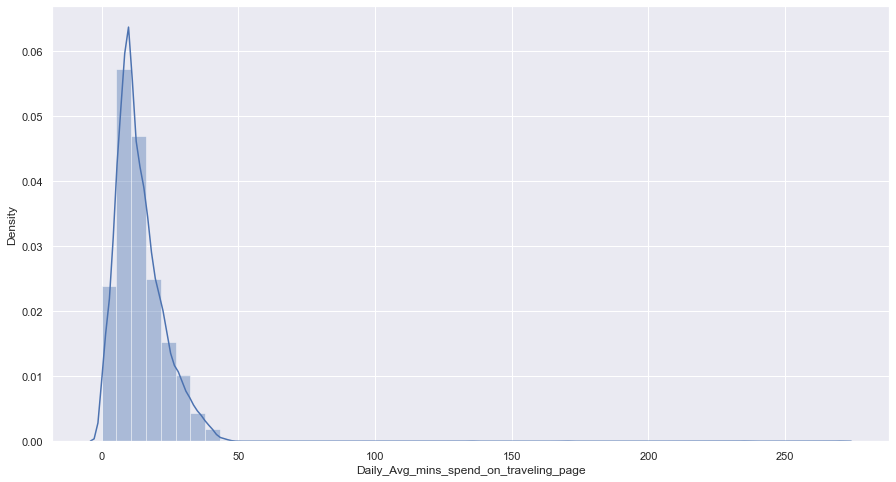


Box plot



Daily Avg mins spend on traveling page

plot

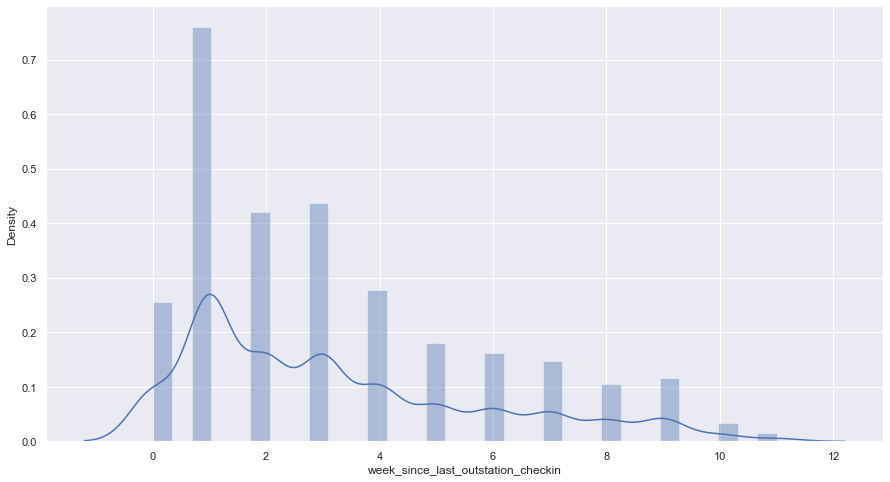


Box plot

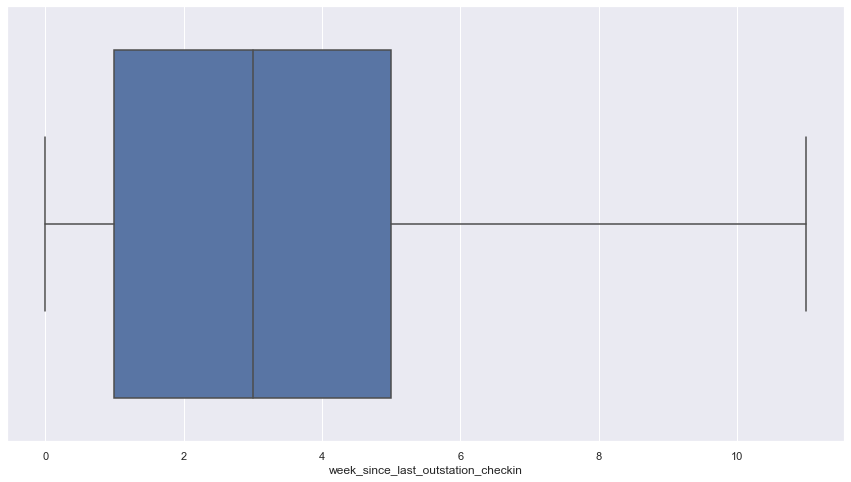


week since last outstation checkin

plot

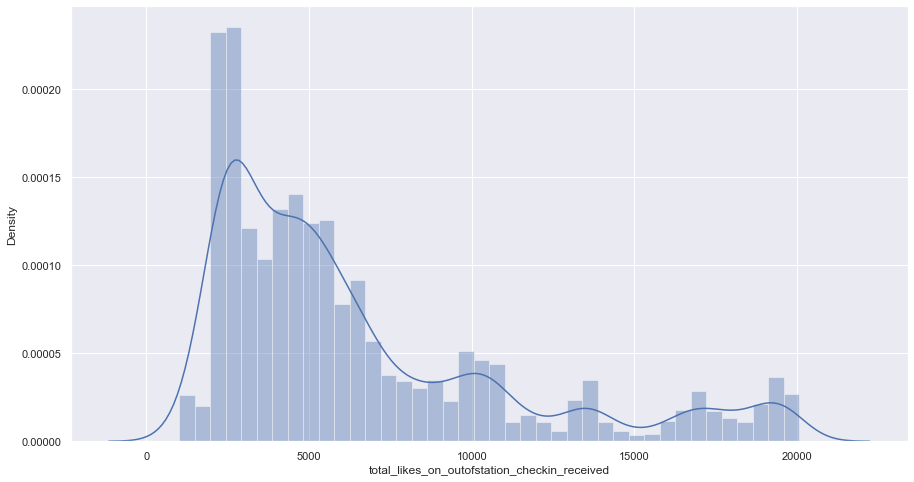


box plot

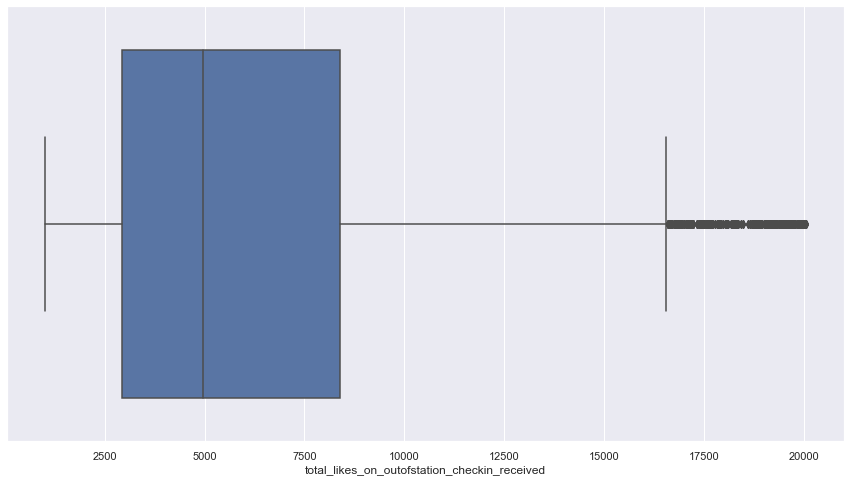


total likes on out of station checkin received

plot



box plot

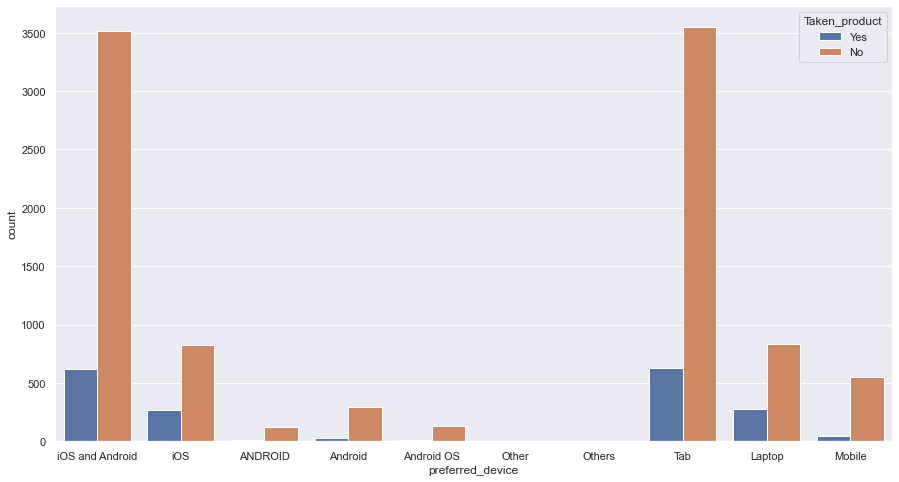


## Bivariate analysis (relationship between different variables, correlations)

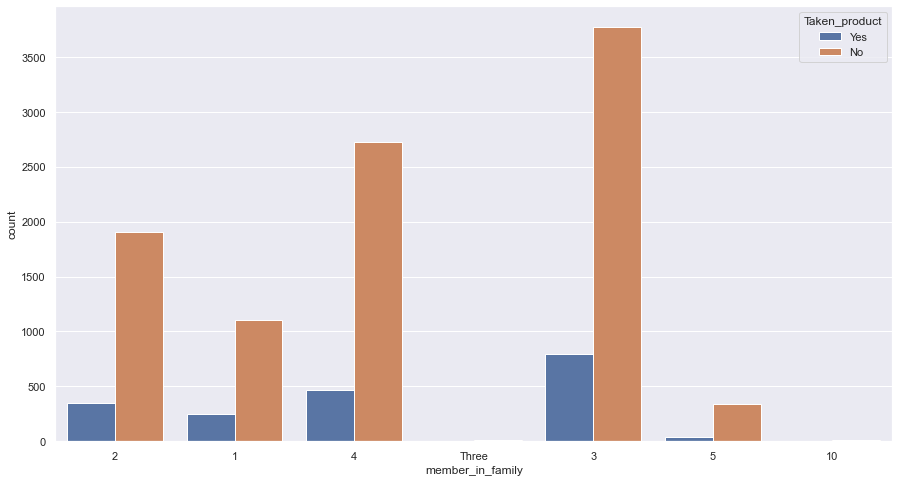
We are analysing dependant variable with other variables.

### Categorical variables

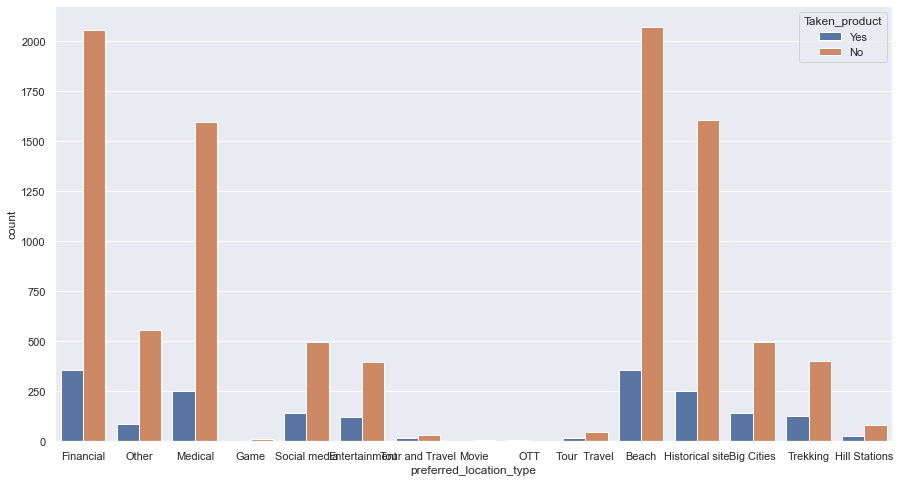
Preferred device



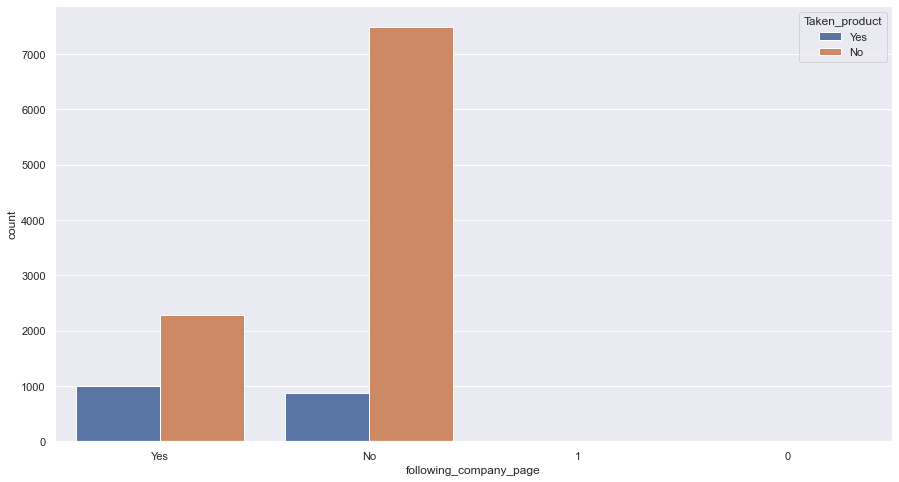
Member in family



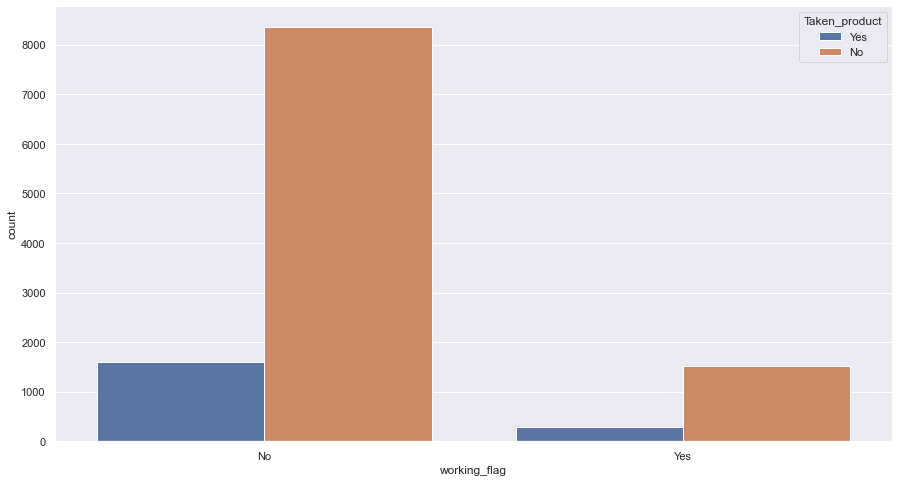
Preferred location type



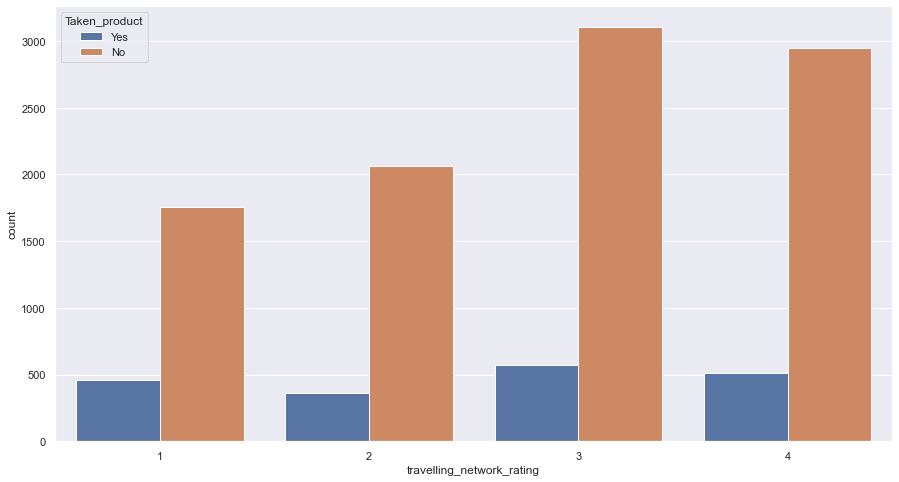
Following company page



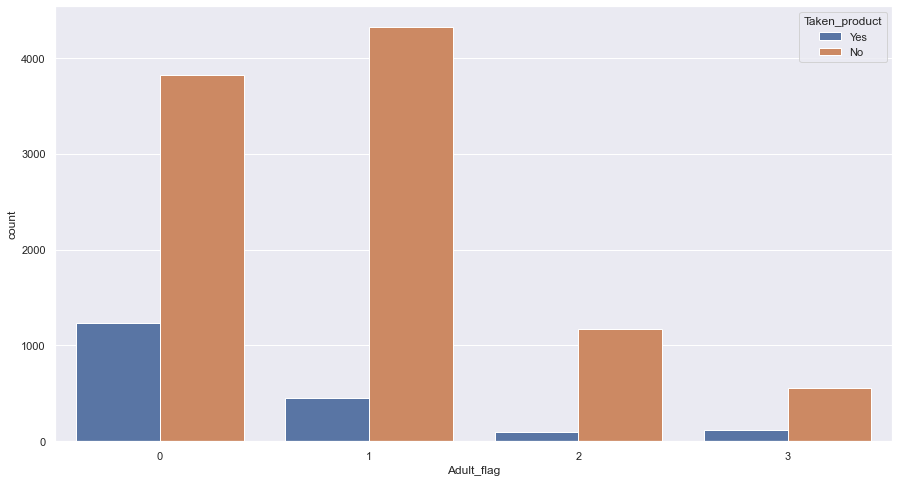
Working flag



Travelling network rating

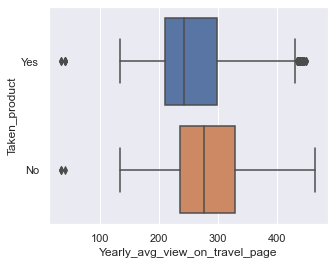


Adult flag

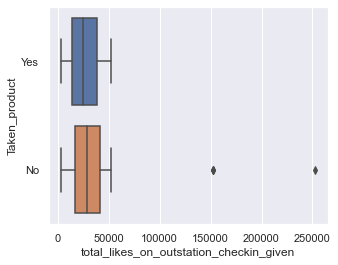


### Continuous variables

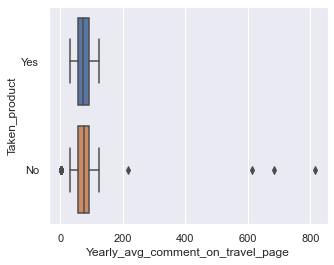
Yearly avg view on travel page



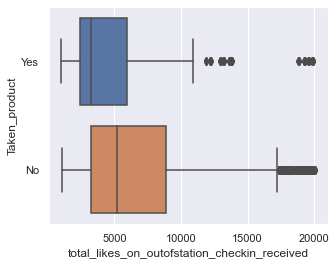
total likes on outstation checkin given



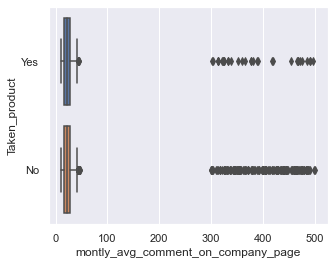
Yearly avg. comment on travel page



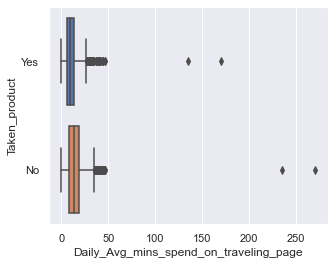
total likes on out of station checkin received



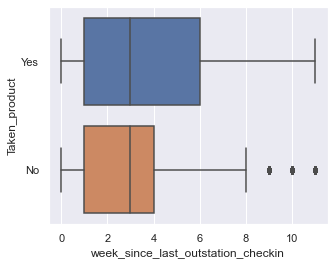
monthly avg. comment on company page



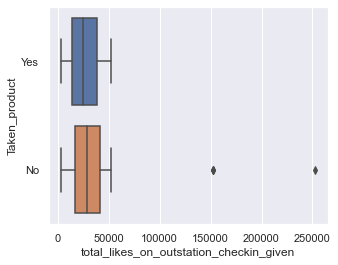
Daily Avg mins spend on traveling page



week since last outstation checkin

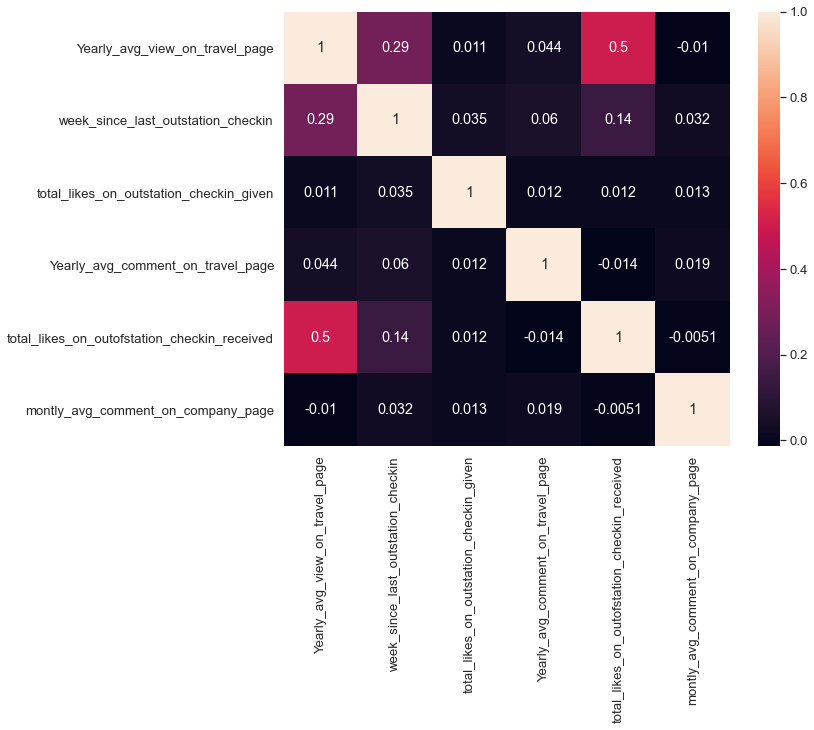


total likes on outstation checkin given



## Correlation heat map

To better understand the corelation between various variables



## Removal of unwanted variables (if applicable)

There are no variables with very high missing values, or irrelevant to the target variable.

Also, none of the variables have high correlation between then.

The User id is kept for labelling purpose and will be removed during model building.

There are no duplicates in the data.

Thus, we will not remove any variable at this stage.

## Missing Value treatment (if applicable)

As discussed earlier the value of null entries are not very high and thus, we can directly drop them.

The new shape of the data is 10456 rows and 17 columns.

## Outlier treatment (if required)

Following are the variable names and the outlier treatment provided to them.

total likes on outstation checkin given

we have caped the max value to 52229

Daily avg. mins spend on traveling page

we have caped the max value to 33.00

monthly avg. comment on company page

we have caped the max value to 37.00

total likes on out of station checkin received

we have caped the max value to 16428

Yearly avg. comment on travel page

we have caped the max value to 119.00

Yearly avg. view on travel page

we have caped the min value to 114.00

## Variable transformation (if applicable)

There are many entries in various columns with do no fit well into the variable description and need to be treated before we can build any model.

The following are the variable and the treatments provided to them.

Following company page

1 were replaced by ‘Yes’ and zero were replaced by ‘No’

Preferred location type

‘Tour and Travel’ was merged with ‘Tour Travel’

‘OTT’ and ‘Movie’ were merged with ‘Others’

Member in family

‘Three’ was replaced with ‘3’

Preferred device

‘Android OS’ and ‘ANDROID’ were replaced with ‘Android’

‘Others’ was replaced with ‘other’

After correcting and merging the categorical variables we will be encoding them.

The key for the encoded variables is as follows

Taken product:

Yes – 1

No – 0

Following company page:

Yes – 1

No – 0

Working flag:

Yes – 1

No – 0

Preferred device:

Tab – 0

iOS and Android – 1

Laptop – 2

iOS – 3

Mobile – 4

Android – 5

Other – 6

Preferred location type:

Financial – 0

Other – 1

Medical – 2

Game – 3

Entertainment – 4

Social media – 5

Tour Travel – 6

Beach – 7

Historical site - 8

Big Cities – 9

Trekking – 10

Hill Station – 11

## Addition of new variables (if required)

No additional Variables have been needed to be added at this stage

# **Business insights from EDA**

## Is the data unbalanced? If so, what can be done? Please explain in the context of the business

The data is not extremely balanced as our target variable is at a ratio of 84% no and 16% yes.

There are almost equal no. of categorial variables as continuous variables.

Most Continuous variables do resemble a bell curve and thus have a high predictive chance using normal distribution curve methods

The data is not particularly as bad data with good EDA and a deep understanding of model building, we can gain a high accuracy model.

## Any business insights using clustering  (if applicable)

## Any other business insights

The data is of their social media usage and thus will inherently have a little bias towards their usage of social media in daily life.

Many such factors will have to be considered for the interpretation of the model results and during the planning of the business model of advertising during deployment.