Contents

[**Introduction** 2](#_Toc107755862)

[Defining problem statement 2](#_Toc107755863)

[Need of the study/project 2](#_Toc107755864)

[Understanding business/social opportunity 2](#_Toc107755865)

[**EDA and Business Implication** 3](#_Toc107755866)

[Data Report 3](#_Toc107755867)

[Data type: 4](#_Toc107755868)

[Duplicates: 4](#_Toc107755869)

[Null Values : 4](#_Toc107755870)

[insigts about numerical data 4](#_Toc107755871)

[Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) 5](#_Toc107755872)

[Categorial Variables 5](#_Toc107755873)

[Continuous variables 7](#_Toc107755874)

[Bivariate analysis (relationship between different variables, correlations) 13](#_Toc107755875)

[Categorical variables 13](#_Toc107755876)

[Continuous variables 15](#_Toc107755877)

[Correlation heat map 18](#_Toc107755878)

[**Data Cleaning and Pre-processing** 19](#_Toc107755879)

[Removal of unwanted variables (if applicable) 19](#_Toc107755880)

[Missing Value treatment (if applicable) 19](#_Toc107755881)

[Outlier treatment (if required) 20](#_Toc107755882)

[Variable transformation (if applicable) 20](#_Toc107755883)

[Encoding 20](#_Toc107755884)

[Categorical variables 21](#_Toc107755885)

[Continuous variable into categorical 21](#_Toc107755886)

[**Model building and validation** 22](#_Toc107755887)

[Data split 22](#_Toc107755888)

[Train Test Split 22](#_Toc107755889)

[df\_1 23](#_Toc107755890)

[df\_2 23](#_Toc107755891)

[Model building 23](#_Toc107755892)

[Decision tree(DT) 23](#_Toc107755893)

[For mobile devices(df\_1) 23](#_Toc107755894)

[For laptops(df\_2) 23](#_Toc107755895)

[Random Forest(RF) 24](#_Toc107755896)

[For mobile devices(df\_1) 24](#_Toc107755897)

[For Laptops(df\_2) 25](#_Toc107755898)

[Artificial Neural Network(ANN) 26](#_Toc107755899)

[For mobile devices(df\_1) 26](#_Toc107755900)

[For Laptops(df\_2) 27](#_Toc107755901)

[Logistic Regression(LR) 28](#_Toc107755902)

[For mobile devices(df\_1) 29](#_Toc107755903)

[For Laptops(df\_2) 30](#_Toc107755904)

[Linear Discriminant Analysis(LDA) 31](#_Toc107755905)

[For mobile devices(df\_1) 31](#_Toc107755906)

[For Laptops(df\_2) 32](#_Toc107755907)

[Model tuning 33](#_Toc107755908)

[Model comparision 33](#_Toc107755909)

[For mobile 33](#_Toc107755910)

[For laptop 33](#_Toc107755911)

[**Final interpretation / recommendation** 33](#_Toc107755912)

# **Introduction**

## 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.

# **EDA and Business Implication**

## **Data Report**

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’

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 |

### 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

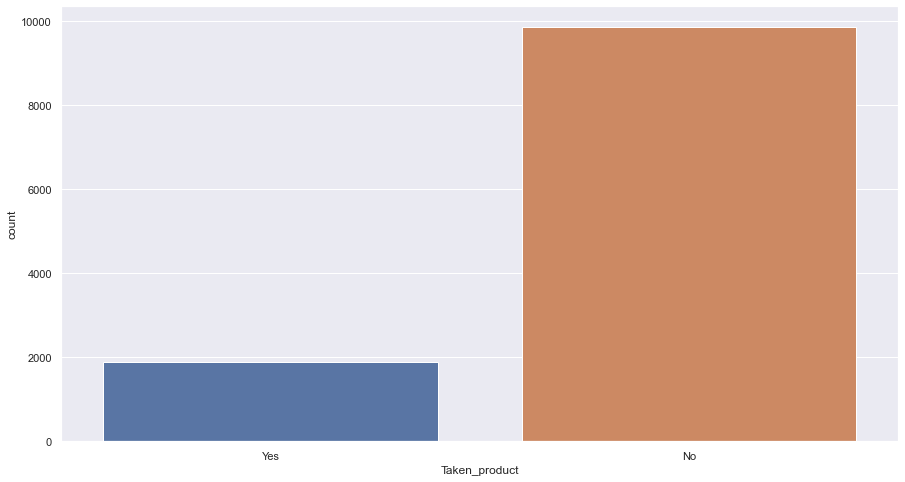
### insigts about numerical data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Yearly\_avg\_view\_on\_travel\_page** | **total\_likes\_on\_outstation\_checkin\_given** | **Yearly\_avg\_comment\_on\_travel\_page** | **total\_likes\_on\_outofstation\_checkin\_received** | **week\_since\_last\_outstation\_checkin** | **montly\_avg\_comment\_on\_company\_page** | **travelling\_network\_rating** | **Adult\_flag** | **Daily\_Avg\_mins\_spend\_on\_traveling\_page** |
| **count** | 11179 | 11379 | 11554 | 11760 | 11760 | 11760 | 11760 | 11760 | 11760 |
| **mean** | 280.830844 | 28170.4818 | 74.790029 | 6531.69907 | 3.203571 | 28.661565 | 2.712245 | 0.793878 | 13.817432 |
| **std** | 68.182958 | 14385.0321 | 24.02665 | 4706.61379 | 2.616365 | 48.660504 | 1.080887 | 0.851823 | 9.070657 |
| **min** | 35 | 3570 | 3 | 1009 | 0 | 11 | 1 | 0 | 0 |
| **25%** | 232 | 16380 | 57 | 2940.75 | 1 | 17 | 2 | 0 | 8 |
| **50%** | 271 | 28076 | 75 | 4948 | 3 | 22 | 3 | 1 | 12 |
| **75%** | 324 | 40525 | 92 | 8393.25 | 5 | 27 | 4 | 1 | 18 |
| **max** | 464 | 252430 | 815 | 20065 | 11 | 500 | 4 | 3 | 270 |

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

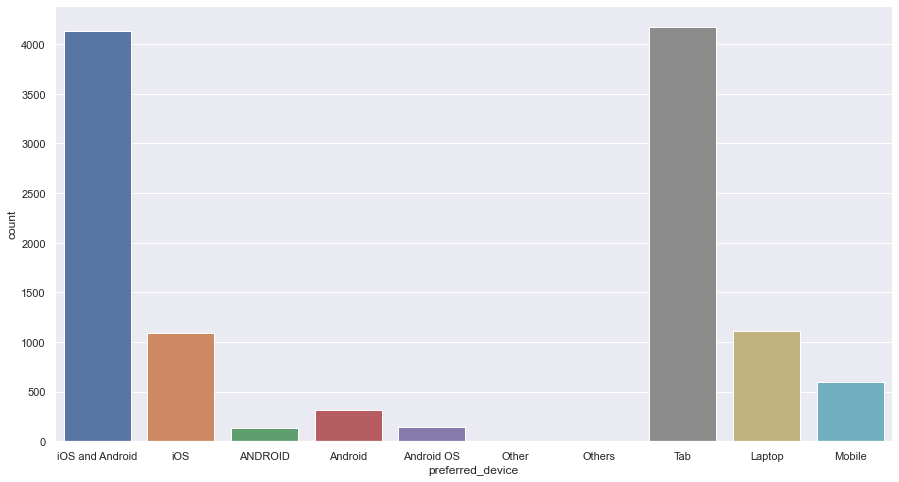
### Categorial Variables

#### Taken product



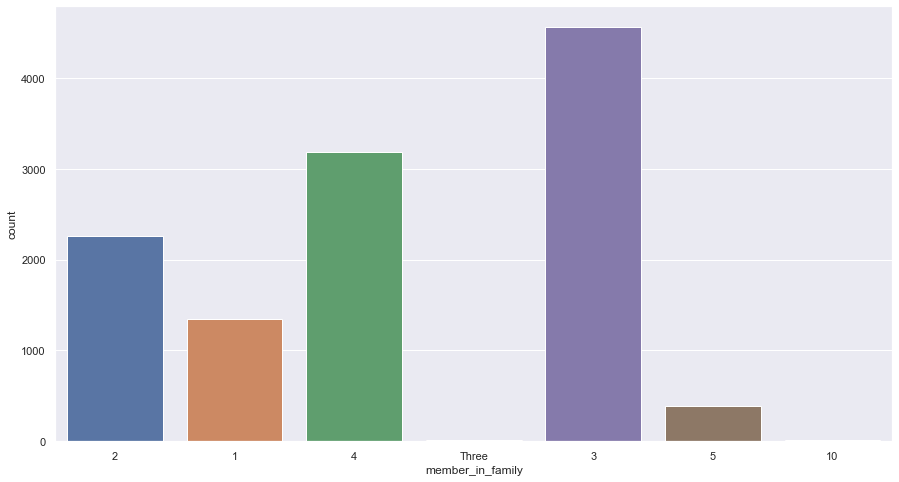
The distribution is unequal.

#### Preferred device



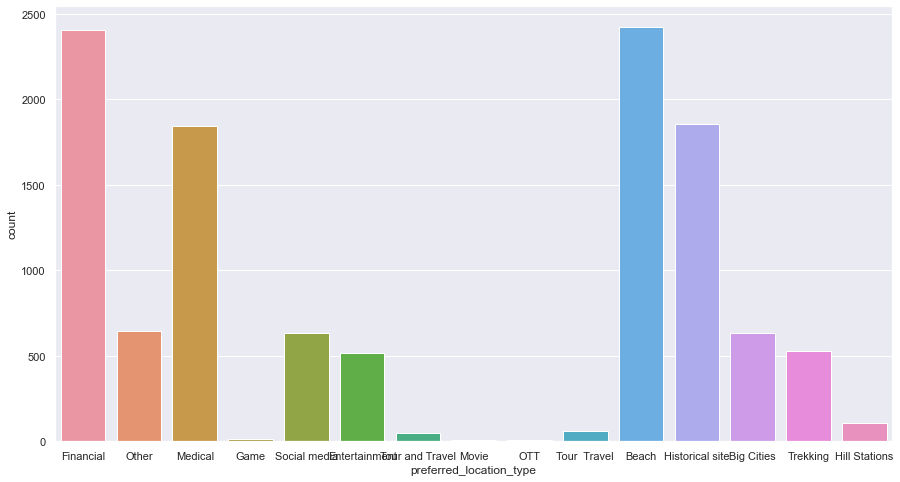
Laptop users are low. Will be clubbing all other types into one

#### Member in family



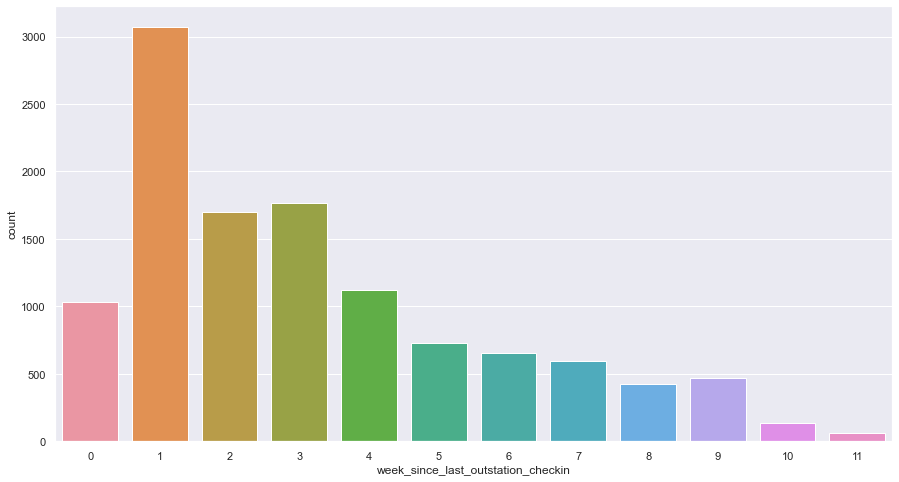
2-3 no. of children is the most common

#### Preferred location type



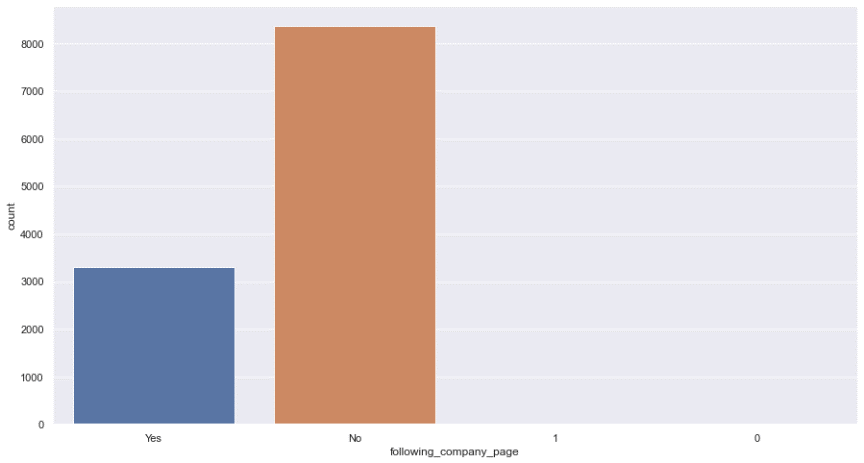
Financial, beach, historic, medical cover most data

#### Week since last outstation checkin



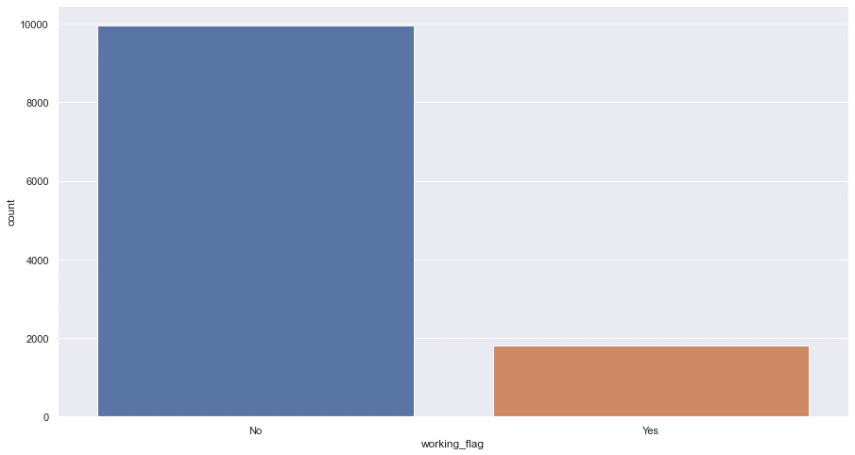
Most data Is of people who travel every month i.e. 1-4 weeks since last checkin

#### Following company page



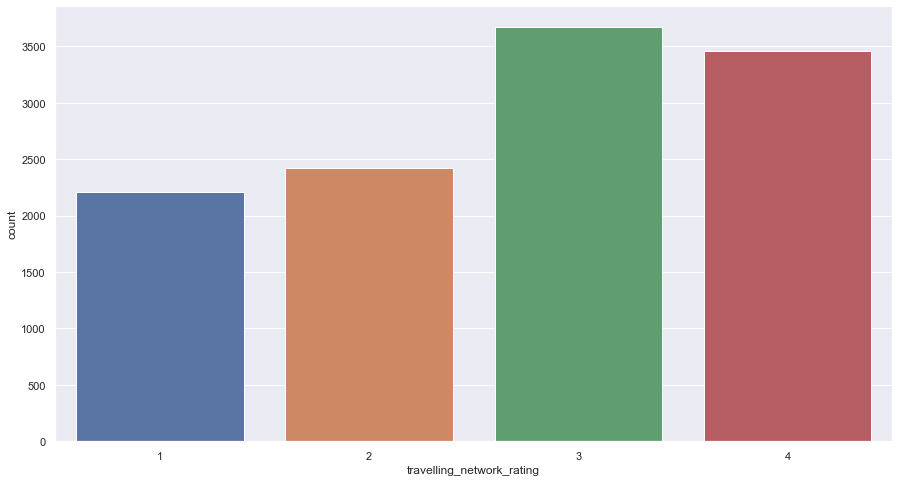
Those who don’t follow company page are potential customers

#### Working flag



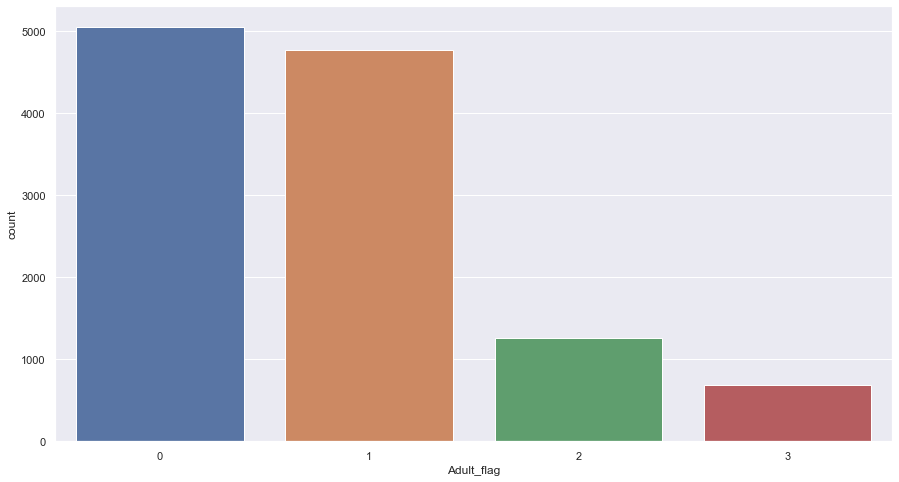
People who do not work are higher in no.

#### Travelling network rating



Most common rating is 3 but not by a huge margin

#### Adult flag

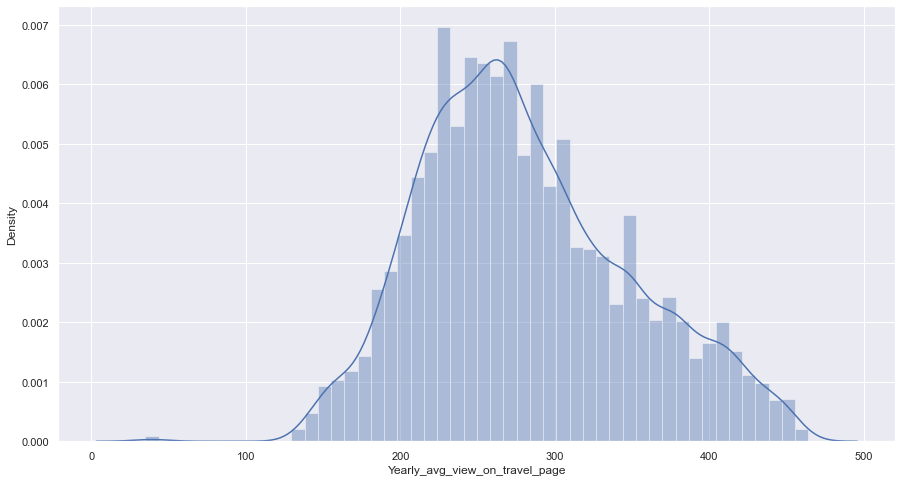


Adult to children ratio is not very high

### Continuous variables

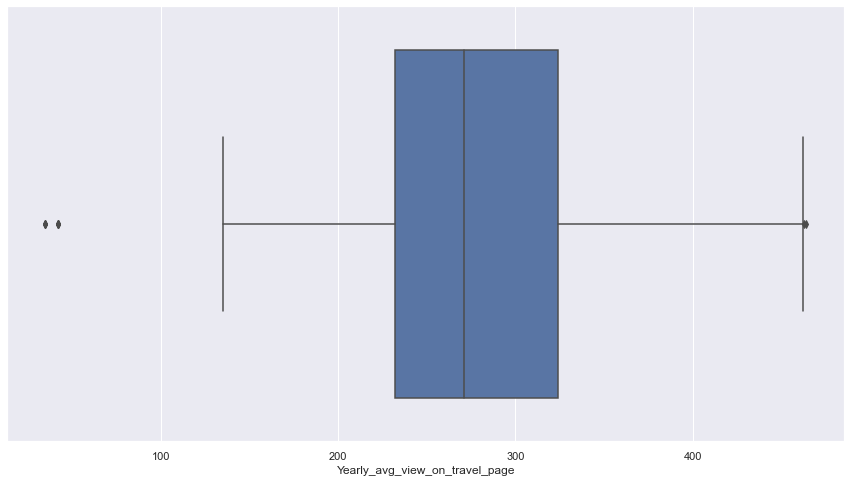
#### Yearly avg. view on travel page

plot



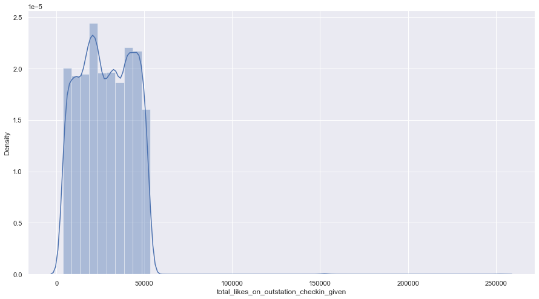
200-350 is the most commin views

Box plot

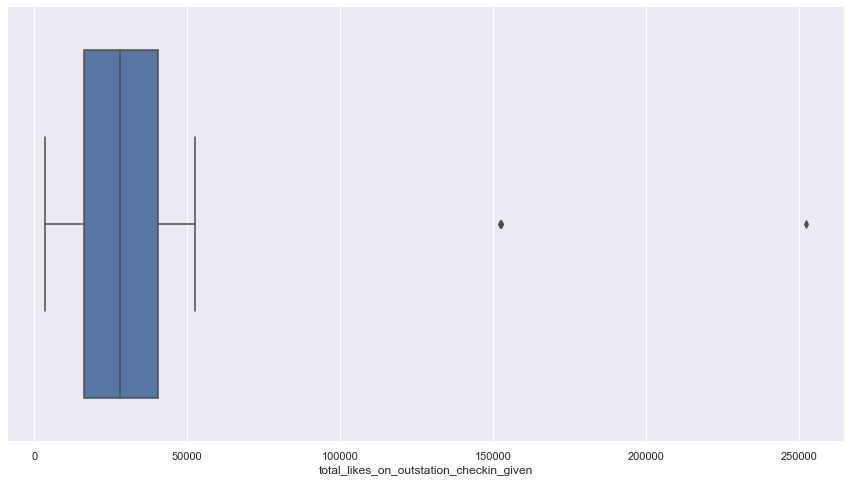


#### total likes on outstation checkin given

plot

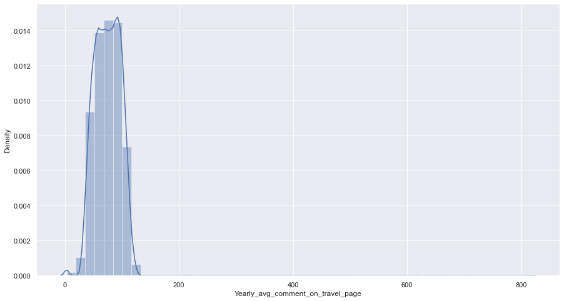


Box plot



#### Yearly avg. comment on travel page

plot

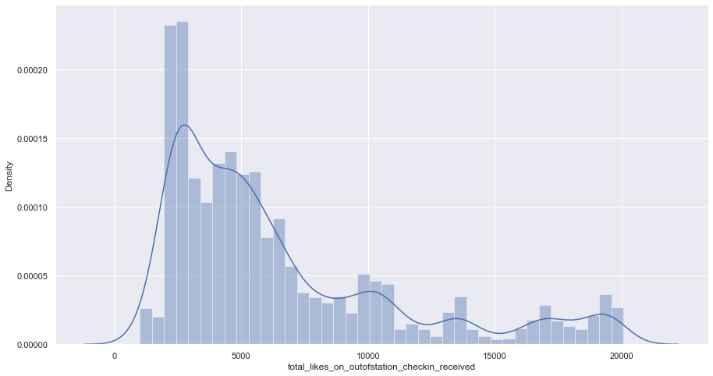


Box plot

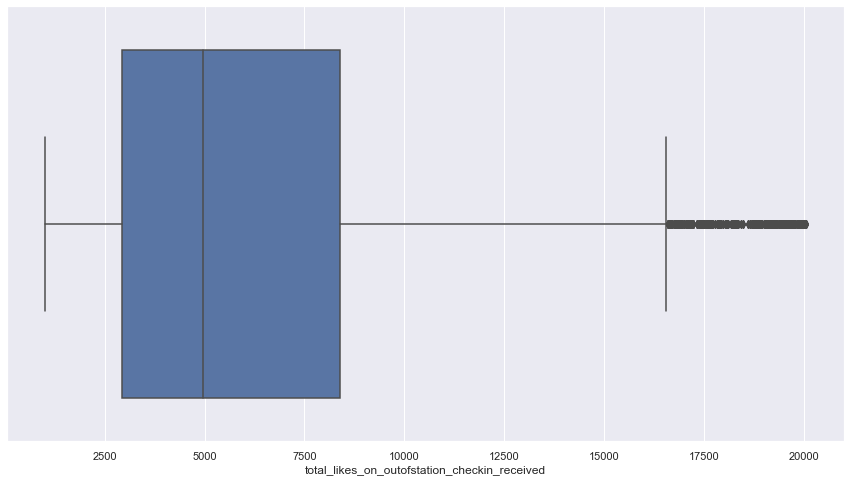


#### total likes on out of station checkin received

plot

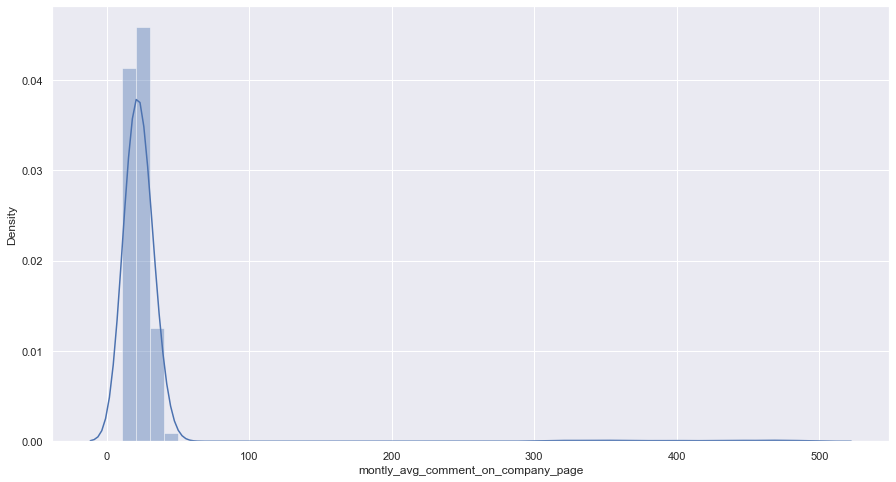


Box plot

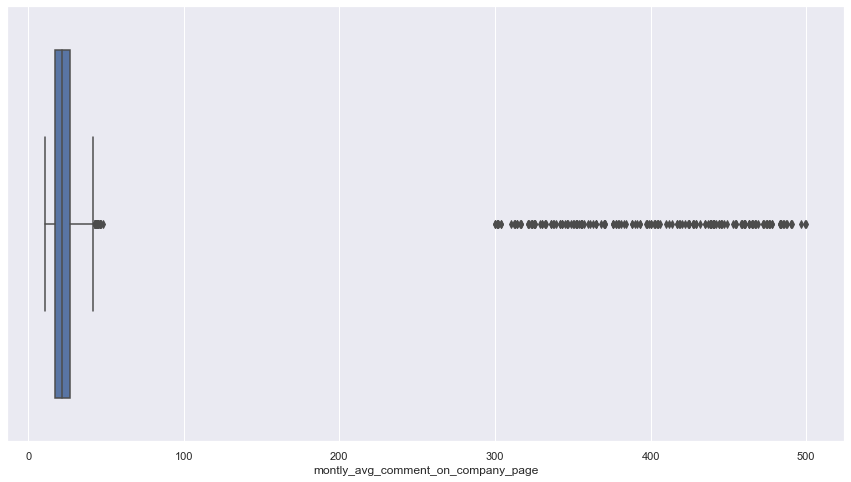


#### monthly avg. comment on company page

plot

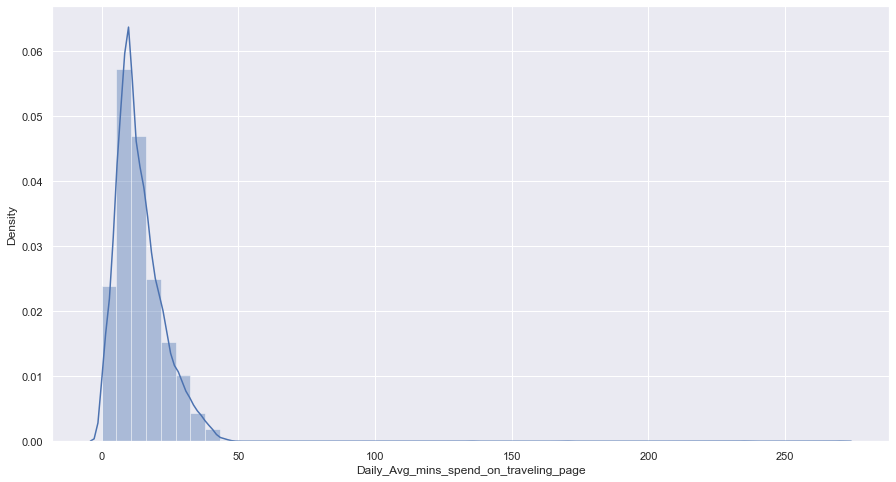


Box plot



#### Daily Avg mins spend on traveling page

plot

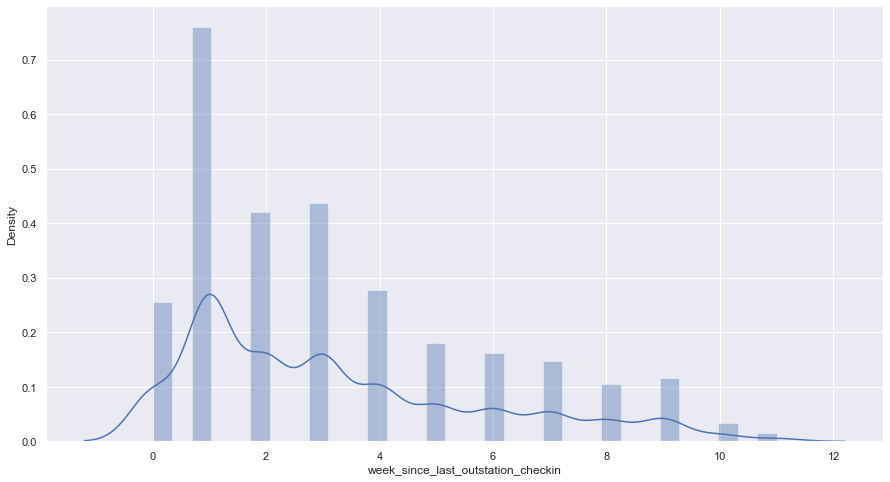


Box plot

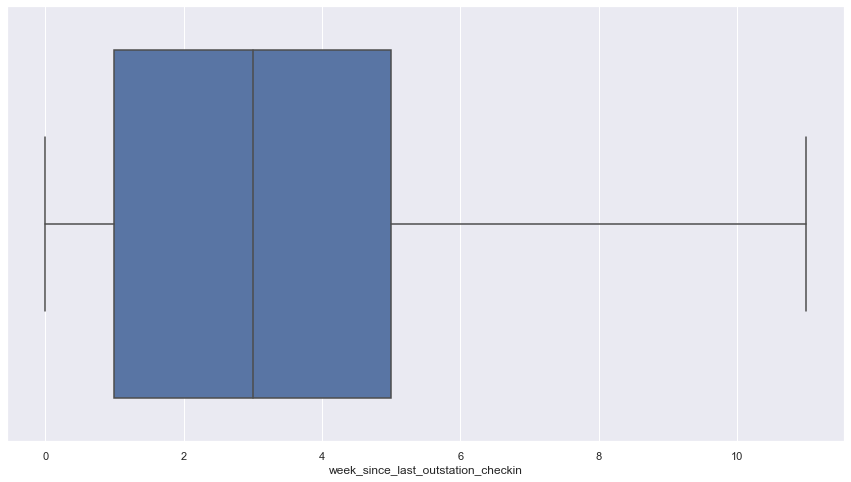


#### week since last outstation checkin

plot

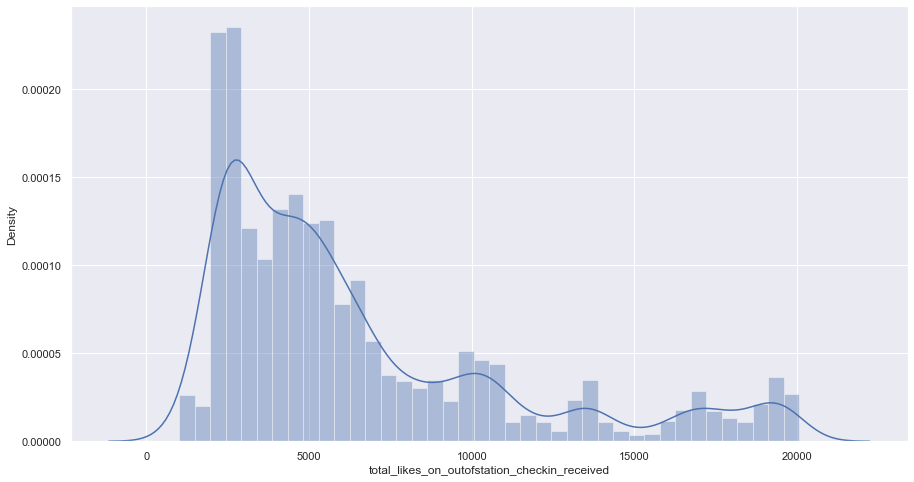


box plot

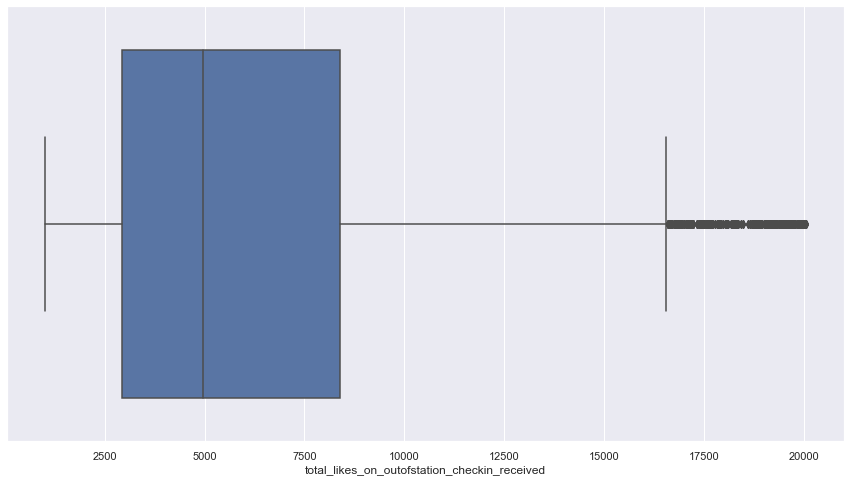


#### total likes on out of station checkin received

plot



box plot

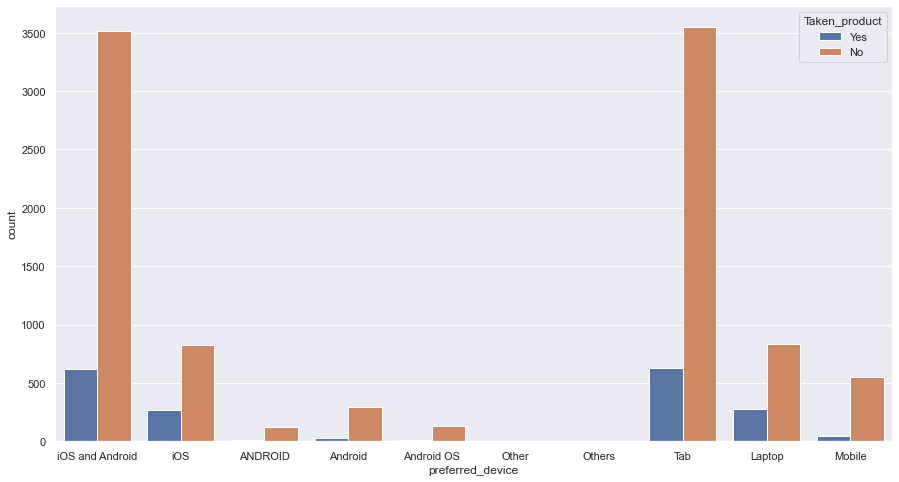


## Bivariate analysis (relationship between different variables, correlations)

We are analysing dependant variable with other variables.

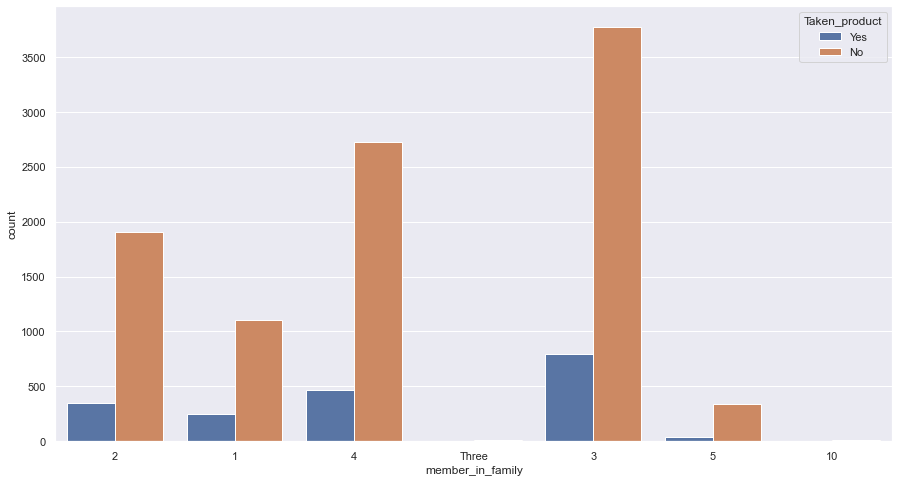
### Categorical variables

Preferred device



The distribution of target variable seems equal and random

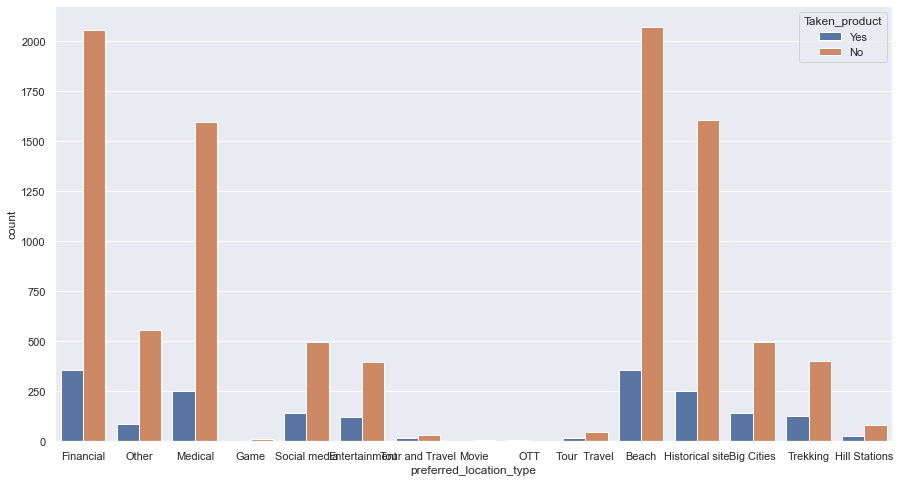
Member in family



The distribution of target variable seems equal and random

Member in family

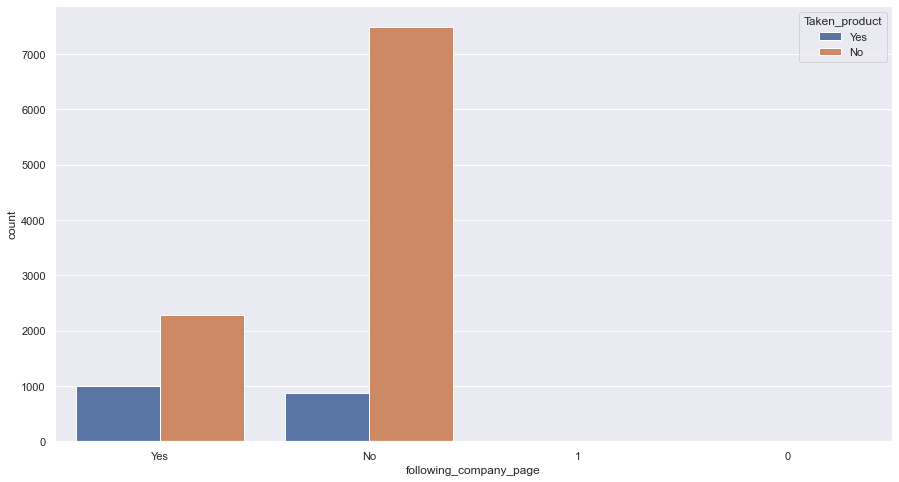
Preferred location type



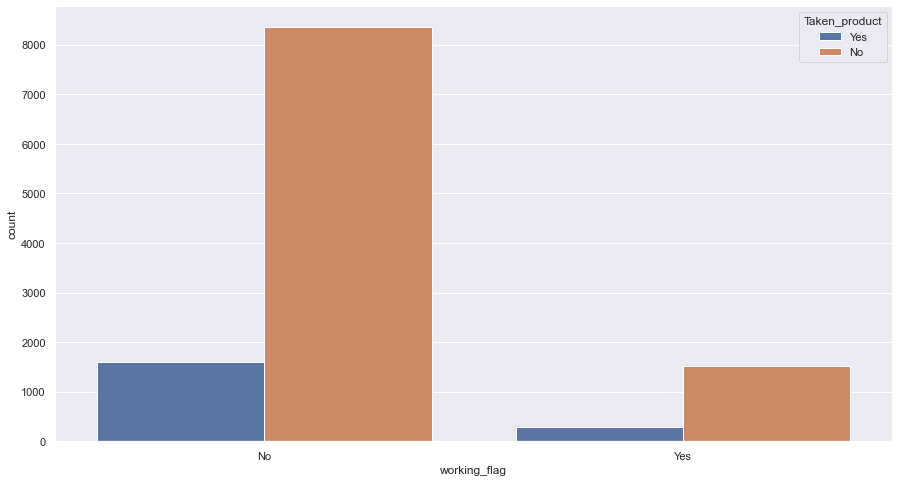
The distribution of target variable seems equal and random

Member in family

Following company page

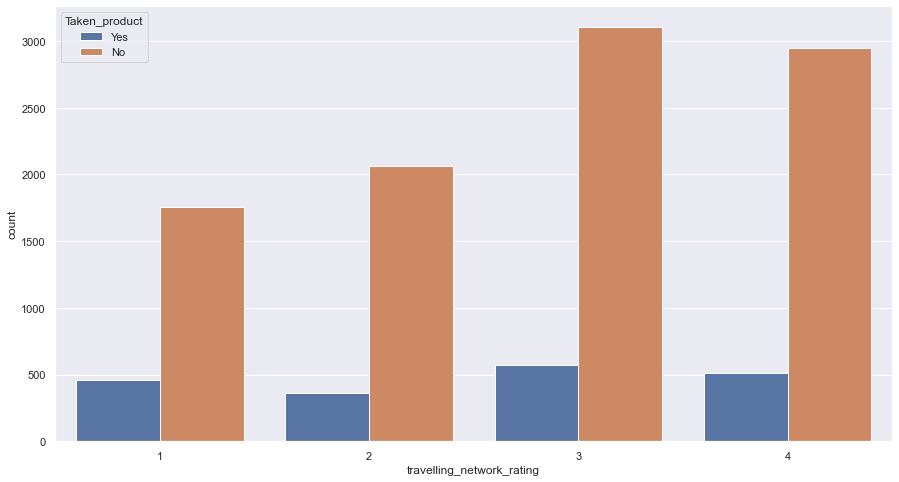


Working flag

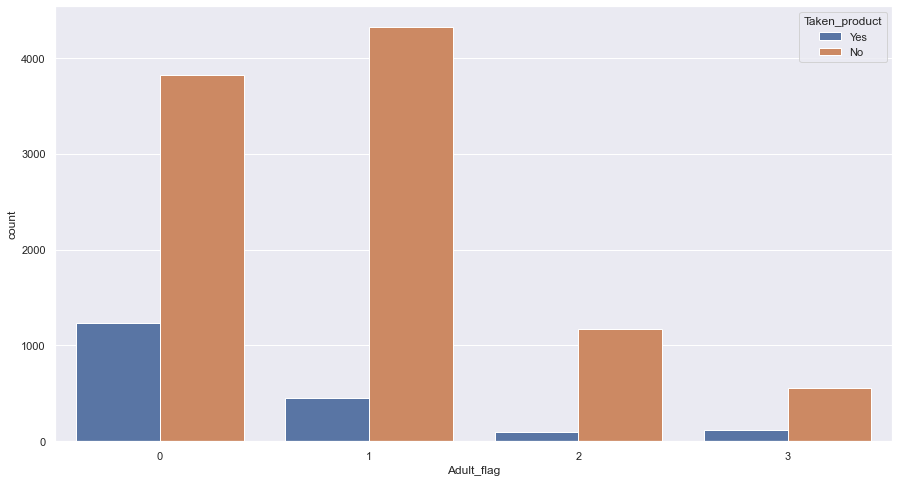


More no of people who do not work seem to take the product

Travelling network rating



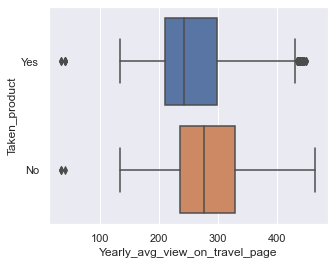
Adult flag



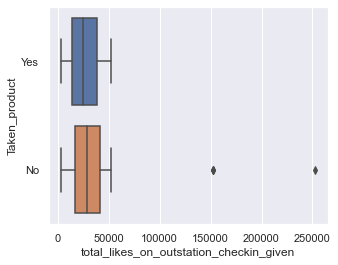
Children tend to take the product more the adults.

### Continuous variables

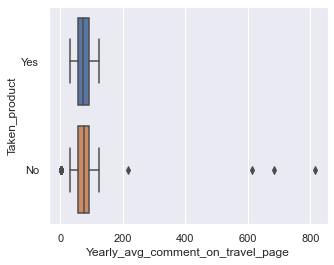
Yearly avg view on travel page



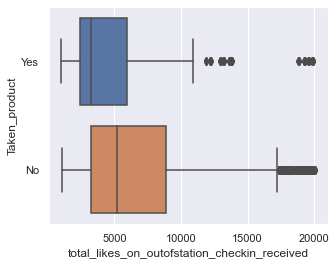
total likes on outstation checkin given



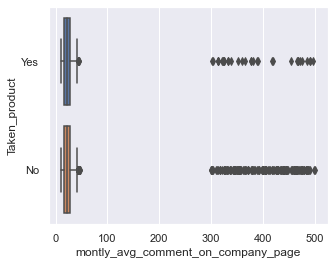
Yearly avg. comment on travel page



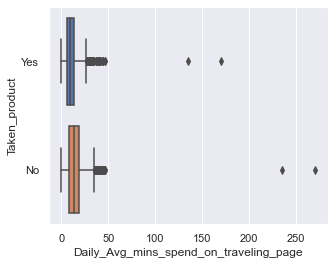
total likes on out of station checkin received



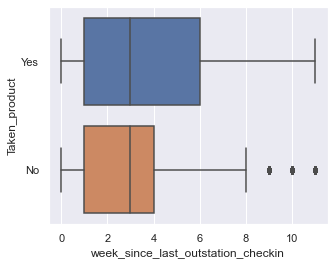
monthly avg. comment on company page



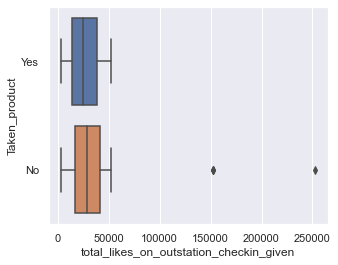
Daily Avg mins spend on traveling page



week since last outstation checkin

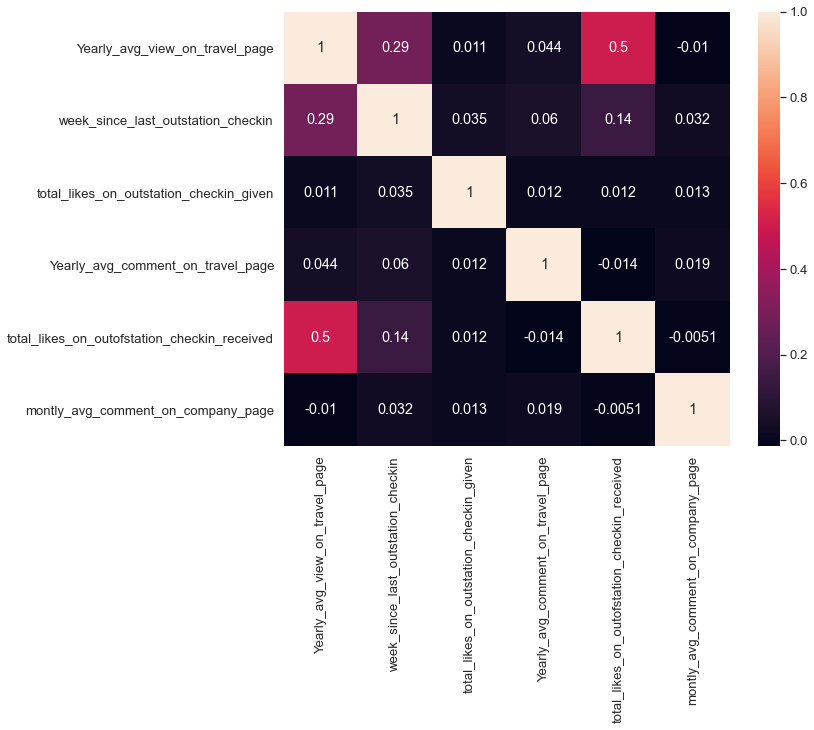


total likes on outstation checkin given



## Correlation heat map

To better understand the corelation between various variables



# **Data Cleaning and Pre-processing**

## 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’

## Encoding

We have encoded the following columns

### Categorical variables

Taken Product:

Yes=1

No=0

Preferred device:

Tab=0

IOS and android=1

Laptop=2

IOS=3

Mobile=4

Android=5

Others=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 city=9

Trekking=10

Hill station=11

Follow company page:

Yes=1

No=0

Working flag:

Yes=1

No=0

### Continuous variable into categorical

Yearly avg. view on travel page

0-100=0

100-200=1

200-300=2

300-400=3

400-500=4

Total likes on outstation checkin given

0-10000=0

10000-20000=1

20000-30000=2

30000-40000=3

40000-50000=4

50000-60000=5

Yearly avg. comment on travel page

0-40=0

40-60=1

60-80=2

80-100=3

100-140=4

Total likes on outstation checkin received

0-2500=0

2500-5000=1

5000-7500=2

7500-10000=3

10000-12500=4

12500-15000=5

15000-20000=6

Monthly avg. comment on company page

0-10=0

10-20=1

20-30=2

30-40=3

Daily avg. mins spent on traveling page

(-5)-10=0

10-20=1

20-30=2

30-40=3

# **Model building and validation**

## Data split

We have split the data based on the preferred devise by the user into laptop and other devices.

This is done because the behaviours of user is very different on different devices and must be delt with separately

Thus, the new data being

df\_1 = All devices except laptop

df\_2 = Laptop

As df\_2 has only one type of entry in preferred device column. We have dropped the column

df\_1 has 9347 rows and 16 columns

df\_2 has 1108 rows and 15 columns

## Train Test Split

We split the data into train and test at a ratio of 70-30.

We also drop the user id row.

### df\_1

train data has 6542 rows and 15 columns

test data has 2805 and 15 rows

### df\_2

train data has 775 rows and 14 columns.

test data has 333 and 14 rows.

## Model building

We have build 5 models with might help us predict the consumer behaviour.

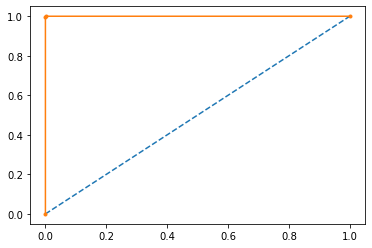
## Decision tree(DT)

### For mobile devices(df\_1)

#### The area under the curve

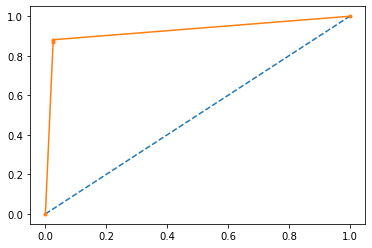
training data

AUC=1



Test data

AUC=0.928



#### Classification report

Train data

precision recall f1-score support

0 1.00 1.00 1.00 5554

1 1.00 1.00 1.00 988

accuracy 1.00 6542

macro avg 1.00 1.00 1.00 6542

weighted avg 1.00 1.00 1.00 6542

test data

precision recall f1-score support

0 0.98 0.97 0.98 2376

1 0.86 0.87 0.87 429

accuracy 0.96 2805

macro avg 0.92 0.92 0.92 2805

weighted avg 0.96 0.96 0.96 2805

#### confusion Metrix

train data

[5554 0]

[ 4 984]

Test data

[2316 60]

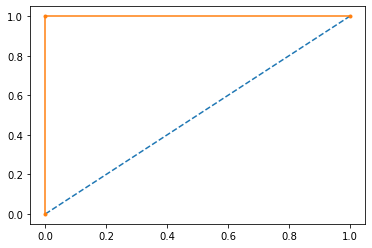
[ 55 374]

### For laptops(df\_2)

#### The area under the curve

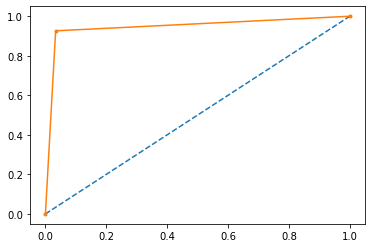
training data

AUC=1.0



Test data

AUC=0.946



#### Classification report

Train data

precision recall f1-score support

0 1.00 1.00 1.00 594

1 1.00 1.00 1.00 181

accuracy 1.00 775

macro avg 1.00 1.00 1.00 775

weighted avg 1.00 1.00 1.00 775

Test data

precision recall f1-score support

0 0.97 0.97 0.97 238

1 0.92 0.93 0.92 95

accuracy 0.95 333

macro avg 0.94 0.95 0.94 333

weighted avg 0.96 0.95 0.96 333

#### confusion Metrix

train data

[594 0]

[ 0 181]

test data

[230 8]

[ 7 88]

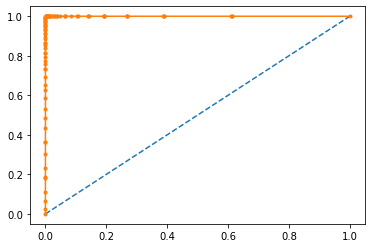
## Random Forest(RF)

### For mobile devices(df\_1)

#### The area under the curve

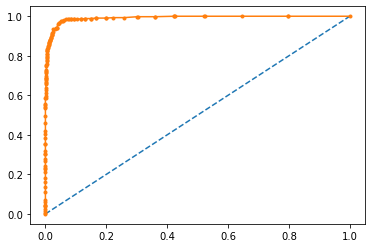
training data

AUC= 1.0



Test data

AUC=0.992



#### Classification report

Train data

precision recall f1-score support

0 1.00 1.00 1.00 5554

1 1.00 1.00 1.00 988

accuracy 1.00 6542

macro avg 1.00 1.00 1.00 6542

weighted avg 1.00 1.00 1.00 6542

Test data

precision recall f1-score support

0 0.96 0.99 0.98 2376

1 0.97 0.79 0.87 429

accuracy 0.96 2805

macro avg 0.96 0.89 0.92 2805

weighted avg 0.96 0.96 0.96 2805

#### confusion Metrix

train data

[5554 0]

[ 4 984]

test data

[2364 12]

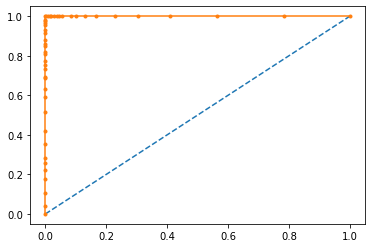
[ 89 340]

### For Laptops(df\_2)

#### The area under the curve

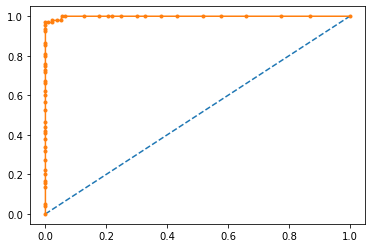
training data

AUC=1.0



Test data

AUC=0.999



#### Classification report

Train data

precision recall f1-score support

0 1.00 1.00 1.00 594

1 1.00 1.00 1.00 181

accuracy 1.00 775

macro avg 1.00 1.00 1.00 775

weighted avg 1.00 1.00 1.00 775

Test data

precision recall f1-score support

0 0.96 1.00 0.98 238

1 1.00 0.91 0.95 95

accuracy 0.97 333

macro avg 0.98 0.95 0.97 333

weighted avg 0.97 0.97 0.97 333

#### confusion Metrix

train data

[594 0]

[ 0 181]

test data

[238 0]

[ 9 86]

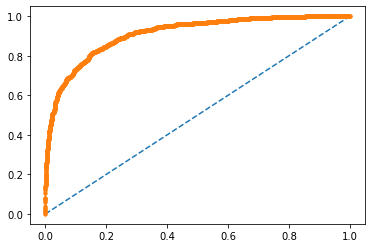
## Artificial Neural Network(ANN)

### For mobile devices(df\_1)

#### The area under the curve

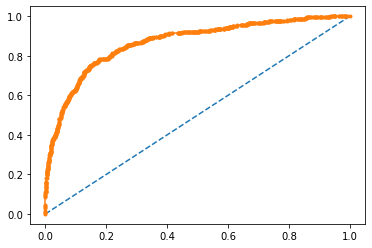
training data

AUC=0.911



Test data

AUC=0.868



#### Classification report

Train data

precision recall f1-score support

0 0.90 0.99 0.94 5554

1 0.87 0.35 0.50 988

accuracy 0.89 6542

macro avg 0.88 0.67 0.72 6542

weighted avg 0.89 0.89 0.87 6542

Test data

precision recall f1-score support

0 0.89 0.99 0.93 2376

1 0.81 0.30 0.43 429

accuracy 0.88 2805

macro avg 0.85 0.64 0.68 2805

weighted avg 0.87 0.88 0.86 2805

#### confusion Metrix

train data

[5504 50]

[ 643 345]

test data

[2347 29]

[ 302 127]

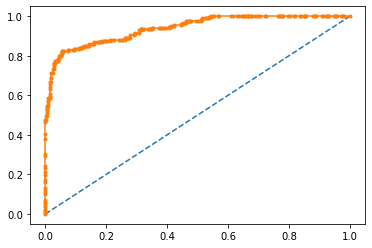
In [38]:

### For Laptops(df\_2)

#### The area under the curve

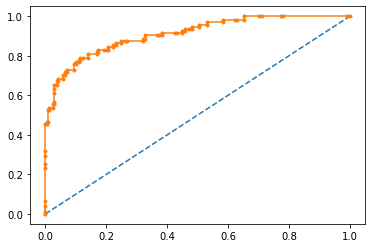
training data

AUC=0.939



Test data

AUC=0.910



#### Classification report

Train data

precision recall f1-score support

0 0.92 0.98 0.95 594

1 0.91 0.71 0.80 181

accuracy 0.92 775

macro avg 0.92 0.85 0.87 775

weighted avg 0.92 0.92 0.91 775

Test data

precision recall f1-score support

0 0.88 0.95 0.91 238

1 0.84 0.68 0.76 95

accuracy 0.87 333

macro avg 0.86 0.82 0.84 333

weighted avg 0.87 0.87 0.87 333

#### confusion Metrix

train data

[582 12]

[ 52 129]

test data

[226 12]

[ 30 65]

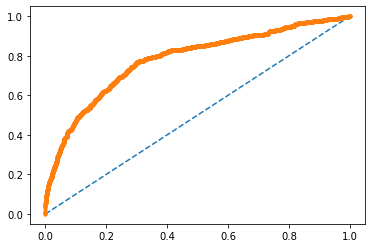
## Logistic Regression(LR)

### For mobile devices(df\_1)

#### The area under the curve

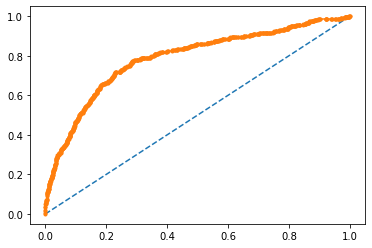
training data

AUC=0.781



Test data

AUC=0.788



#### Classification report

Train data

precision recall f1-score support

0 0.87 0.98 0.92 5554

1 0.64 0.18 0.28 988

accuracy 0.86 6542

macro avg 0.76 0.58 0.60 6542

weighted avg 0.84 0.86 0.83 6542

Test data

precision recall f1-score support

0 0.87 0.98 0.92 2376

1 0.64 0.17 0.27 429

accuracy 0.86 2805

macro avg 0.75 0.58 0.60 2805

weighted avg 0.83 0.86 0.82 2805

#### confusion Metrix

train data

[5454 100]

[ 808 180]

test data

[2334 42]

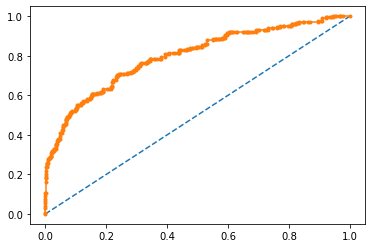
[ 355 74]

### For Laptops(df\_2)

#### The area under the curve

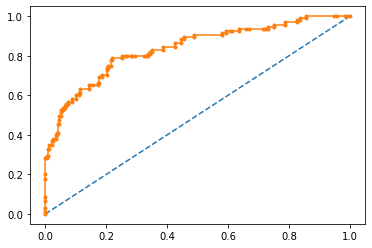
training data

AUC=0.799



Test data

AUC=0.838



#### Classification report

Train data

precision recall f1-score support

0 0.83 0.95 0.89 594

1 0.70 0.38 0.49 181

accuracy 0.82 775

macro avg 0.77 0.67 0.69 775

weighted avg 0.80 0.82 0.80 775

Test data

precision recall f1-score support

0 0.81 0.96 0.88 238

1 0.81 0.44 0.57 95

accuracy 0.81 333

macro avg 0.81 0.70 0.73 333

weighted avg 0.81 0.81 0.79 333

#### confusion Metrix

train data

[565 29]

[112 69]

test data

[228 10]

[ 53 42]

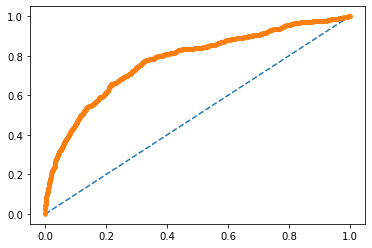
## Linear Discriminant Analysis(LDA)

### For mobile devices(df\_1)

#### The area under the curve

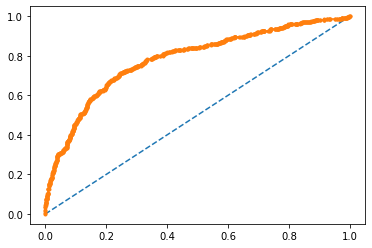
training data

AUC=0.778



Test data

AUC=0.782



#### Classification report

Train data

precision recall f1-score support

0 0.88 0.98 0.92 5554

1 0.63 0.22 0.33 988

accuracy 0.86 6542

macro avg 0.75 0.60 0.62 6542

weighted avg 0.84 0.86 0.83 6542

Test data

precision recall f1-score support

0 0.87 0.97 0.92 2376

1 0.60 0.21 0.31 429

accuracy 0.86 2805

macro avg 0.74 0.59 0.62 2805

weighted avg 0.83 0.86 0.83 2805

#### confusion Metrix

train data

[5424 130]

[ 770 218]

test data

[2316 60]

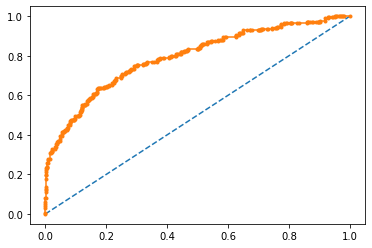
[ 338 91]

### For Laptops(df\_2)

#### The area under the curve

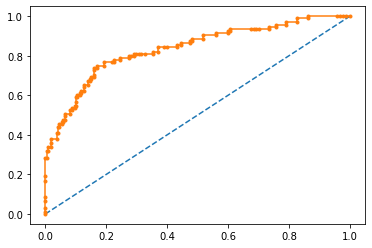
training data

AUC= 0.794



Test data

AUC=0.837



#### Classification report

Train data

precision recall f1-score support

0 0.84 0.95 0.89 594

1 0.71 0.39 0.51 181

accuracy 0.82 775

macro avg 0.77 0.67 0.70 775

weighted avg 0.81 0.82 0.80 775

Test data

precision recall f1-score support

0 0.81 0.96 0.88 238

1 0.80 0.43 0.56 95

accuracy 0.81 333

macro avg 0.81 0.69 0.72 333

weighted avg 0.81 0.81 0.79 333

#### confusion Metrix

train data

[565 29]

[110 71]

test data

[228 10]

[ 54 41]

### Model tuning

We tried various model tuning techniques such as ensambling and bragging and random forest

Out of all the machine learning technique Random forest gave the best results.

With a model score of 0.999 for train and 0.963 for test.

This is because the entire data columns are in discrete values (1,2,3…)

Also, the customer behaviour is very easily predictable using a decision tree.

## Model comparision

### For mobile

|  |  |  |  |
| --- | --- | --- | --- |
| Model | accuracy | F1 score | AUC |
| Decision tree | 0.96 | 0.959 | 0.930 |
| Random forest | 0.96 | 0.964 | 0.992 |
| ANN | 0.891 | 0.89 | 0.883 |
| LR | 0.86 | 0.858 | 0.788 |
| LDA | 0.86 | 0.858 | 0.782 |

### For laptop

|  |  |  |  |
| --- | --- | --- | --- |
| Model | accuracy | F1 score | AUC |
| Decision tree | 0.96 | 0.96 | 0.951 |
| Random forest | 0.97 | 0.972 | 1.0 |
| ANN | 0.88 | 0.88 | 0.92 |
| LR | 0.81 | 0.810 | 0.838 |
| LDA | 0.81 | 0.807 | 0.837 |

Thus for both the device users we have selected Random forest as the best model to predict the target variable.

# **Final interpretation / recommendation**

* As decision tree is a very well-fitting model after categorising. And thus, can say that variable affect the target variable directly.
* For mobile and laptop users the most important column is ‘yearly avg. Outstation checkin’ with direct relation. This means people who have a high outstation checkin avg. tend to prefer to take the product.
* Families with 3-4 or more members tend to take the package the most which means ‘family trips’ is a good product.
* Special attention must be given to the trips which are for financial, beach travel, historical site visit, medial reasons.
* A huge part of people does not follow the company page and thus are a very good potential customer.
* People who do not work which may include financially dependant or stable (do not wish to work anymore) tend to take the product more.
* based on columns such as ‘total likes on outstation checkin given’, ‘total likes on out of station checkin received’, ‘Daily Avg. mins spend on traveling page’, ‘total likes on outstation checkin given’, ‘Yearly avg. view on travel page’ we can assume the high use of social media.
* People who are active on social media are less likely to take the product. Thus, social media engagement can be kept low while focusing on lead generation.