SOCIAL MEDIA TOURISM

CAPSTONE PROJECT

NOTES - I

Submitted To:

Concerned Faculty
At
Great Learning
The University of Texas at Austin

Submitted By:

Rachit Mittal
PGPDSBA online July E 2020

Problem Statement

An aviation company that provides domestic as well as international trips to the customers have collaborated with a social networking platform to **a)** Learn digital and social behaviour of the customers, **b)** Provide the digital advertisement on the user page of the targeted customers and **c)** Target those customers that have high propensity to take up the product separately for both mobiles and laptops.

Need of the Study

Instead of tele-calling each and every customer from its database, company can target those customers that are interested in purchasing the product leading to effective time and cost management. This will also help the company to understand the behaviour of the customers that lead to their conversion. This will also result in better digital presence among its customers which will result in better revenue generation for the company.

Understanding Business/Social Opportunity

The leading trends towards the Social Networking has drawn high public attention from past 'two' decades. For both small businesses and large corporations, social media is playing a key role in brand building and customer communication. Apart from social networking sites like Facebook, Twitter, Instagram, Snapchat etc, other categories like news, Communication, Commenting, Marketing, Banking, Entertainment etc. are also generating huge social media content every minute.

The understanding of the customer's behaviours on a social media platform will result in targeting advertisements according to the needs and wants of the specific set of customers that can result in high propensity to take up the product. Apart from this company can understand the problems associated with the customers that have posted bad reviews. Then, instead of calling each and every customer company can utilize its resources to improve revenue.

Data Collection

Data is collected through social media monitoring and online marketing analytics of the company's page as well as various travelled related pages along with the monitoring of the customer's account throughout the year on daily basis.

Data Inspection

	UserID	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins
0	1000001	Yes	307.0	iOS and Android	38570.0	1
1	1000002	No	367.0	iOS	9765.0	1
2	1000003	Yes	277.0	iOS and Android	48055.0	1
3	1000004	No	247.0	iOS	48720.0	1
4	1000005	No	202.0	iOS and Android	20685.0	1

Fig 1

This is the head of the data containing first 5 rows of the data.

#	Column	Non-Null Count	Dtype
0	UserID	11760 non-null	int64
1	Taken_product	11760 non-null	object
2	Yearly_avg_view_on_travel_page	11179 non-null	float64
3	preferred_device	11707 non-null	object
4	total_likes_on_outstation_checkin_given	11379 non-null	float64
5	yearly_avg_Outstation_checkins	11685 non-null	object
6	member_in_family	11760 non-null	object
7	preferred_location_type	11729 non-null	object
8	Yearly_avg_comment_on_travel_page	11554 non-null	float64
9	total_likes_on_outofstation_checkin_received	11760 non-null	int64
10	week_since_last_outstation_checkin	11760 non-null	int64
11	following_company_page	11657 non-null	object
12	montly_avg_comment_on_company_page	11760 non-null	int64
13	working_flag	11760 non-null	object
14	travelling_network_rating	11760 non-null	int64
15	Adult_flag	11760 non-null	int64
16	Daily_Avg_mins_spend_on_traveling_page	11760 non-null	int64

dtypes: float64(3), int64(7), object(7)

Fig 2

This shows the number of columns in the data and data type of each and every column. The entire dataset consists of 3 float type variables, 7 integer type variables and 7 object or string type variables.

The number of rows in the data are 11760. The number of columns in the data are 17.

UserID	0
Taken_product	0
Yearly_avg_view_on_travel_page	581
preferred_device	53
total_likes_on_outstation_checkin_given	381
yearly_avg_Outstation_checkins	75
member_in_family	0
preferred_location_type	31
Yearly_avg_comment_on_travel_page	206
total_likes_on_outofstation_checkin_received	0
week_since_last_outstation_checkin	0
following_company_page	103
montly_avg_comment_on_company_page	0
working_flag	0
travelling_network_rating	0
Adult_flag	0
Daily_Avg_mins_spend_on_traveling_page	0

Fig 3

The total number of **null values** in the data are 1430. The number of **duplicate values** is 0.

Description	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins	member_in_family	preferred_location_type	Yearly_avg_comment_on_travel_page	total_likes_on_outofstation_checkin_receive	week_since_last_outstation_checkin	following_company_page	montly_avg_comment_on_company_page	working_flag	travelling_network_rating	Adult_flag	Daily_Avg_mins_spend_on_traveling_page
count	11760	11179	11707	11379	11685	11760	11729	11554	11760	11760	11657	11760	11760	11760	11760	11760
unique	2	-	10	-	30	7	15	-	-	-	4	-	2	-	-	-
top	No	-	Tab	-	1	3	Beach	-	-	-	No	-	No	-	-	-
freq	9864	-	4172	-	4543	4561	2424	-	-	-	8355	-	9952	-	-	-
mean	-	280.83	-	28170.48	-	-	-	74.79	6531.70	3.20	-	28.66	-	2.71	0.79	13.82
std	-	68.18	-	14385.03	-	-	-	24.03	4706.61	2.62	-	48.66	-	1.08	0.85	9.07
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

Fig 4

It clearly shows that there are high number of customers that **have not purchased the product** of the company. The **average weeks since last outstation check-in** is 2.62 and have **below average travel rating** of 2.71 meaning close friends of customers who also like travelling are very less.

Most of the customers **do not follow the company page** as well and **prefer "Tab"** as the operating device. Most of the customers have a **family size** of 3 and **travel 1 time in a year**.

The customers spend an average of 13.82 minutes on a **travelling page on daily basis**.

Univariate Analysis

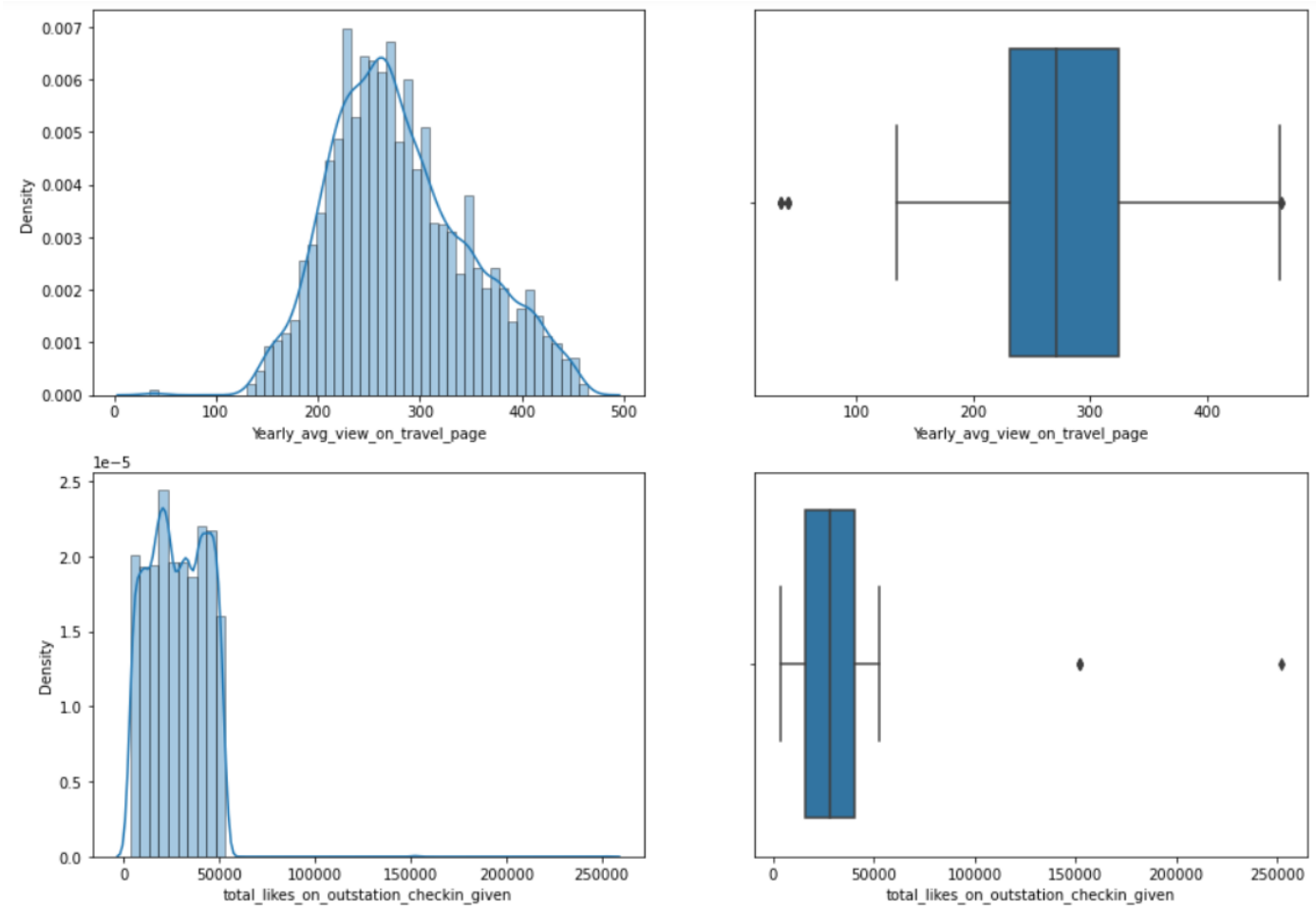


Fig 5

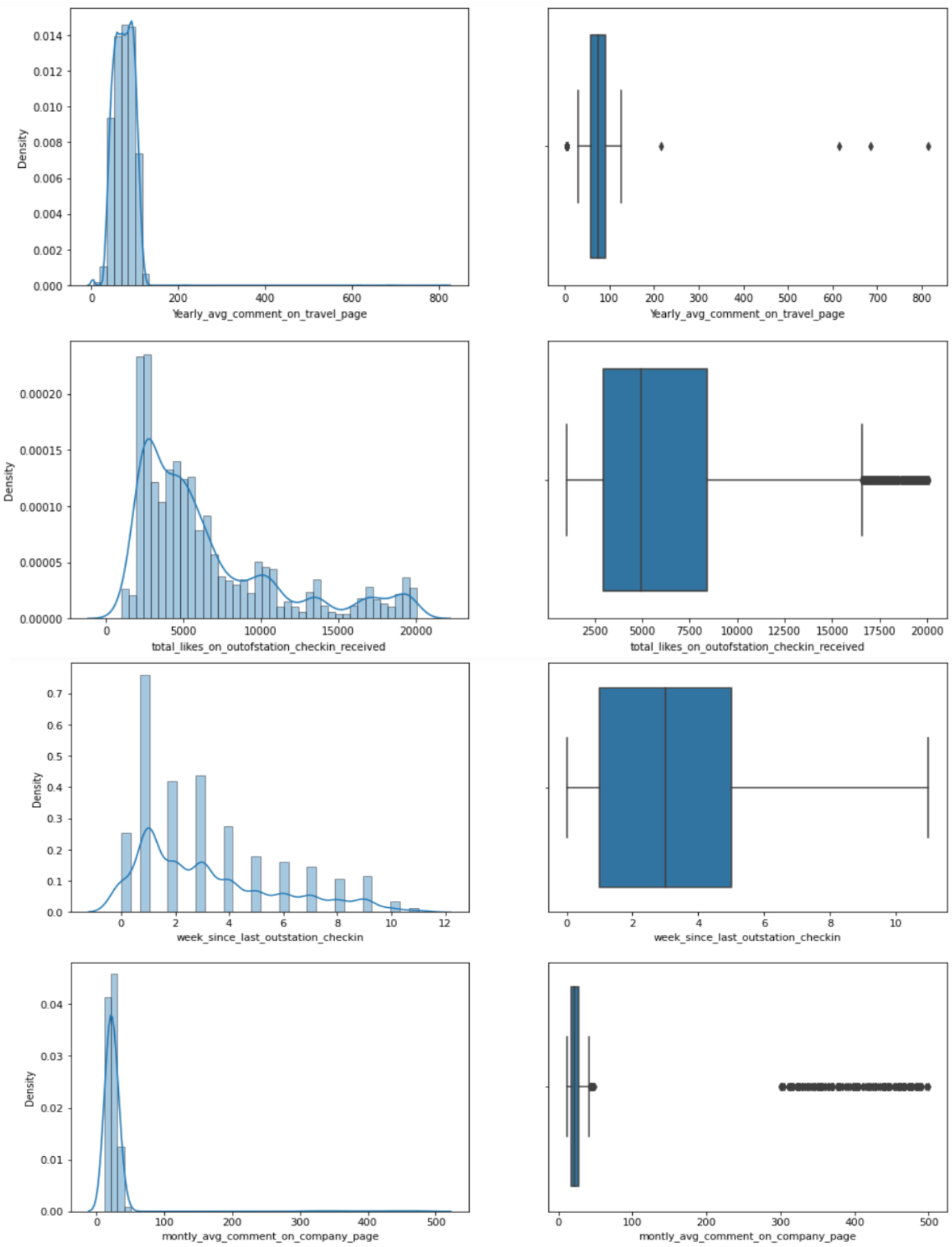

Density
Yearly_avg_comment_on_travel_page
Yearly_avg_comment_on_travel_page
Density
total_likes_on_outofstation_checkin_received
total_likes_on_outofstation_checkin_received
Density
week_since_last_outstation_checkin
week_since_last_outstation_checkin
Density
montly_avg_comment_on_company_page
montly_avg_comment_on_company_page

Fig 6

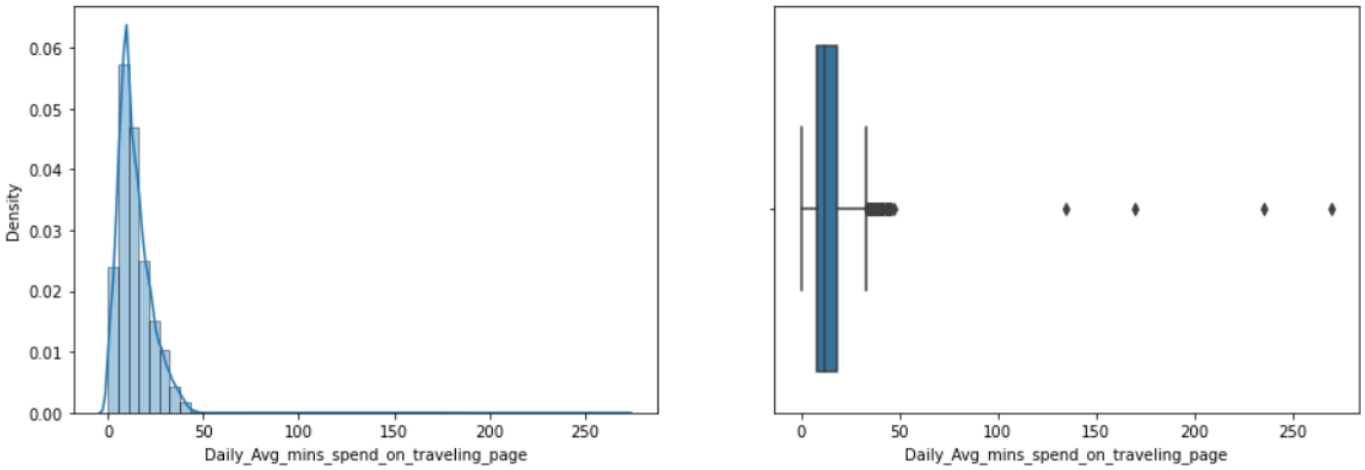


Fig 7

The figures 5, 6 and 7 provides the distribution and box plots of the numerical columns in the data. Most of the numerical columns in the data are rightly skewed and have large number of outliers in the data set.

The variable “Yearly average view on Travel page” is somewhat normally distributed but still have outliers on both sides of the distribution whereas “week since last outstation check-in” variables has no outlier in it despite showing somewhat right skewness.

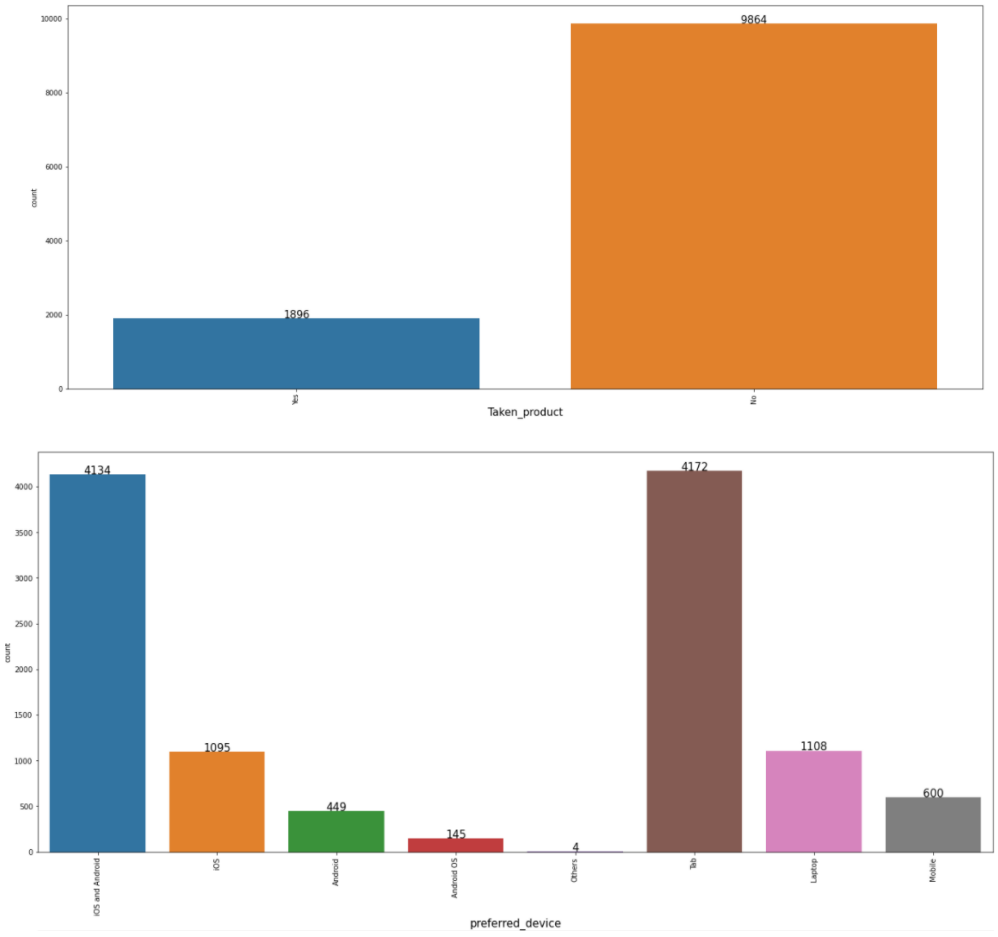


Fig 8

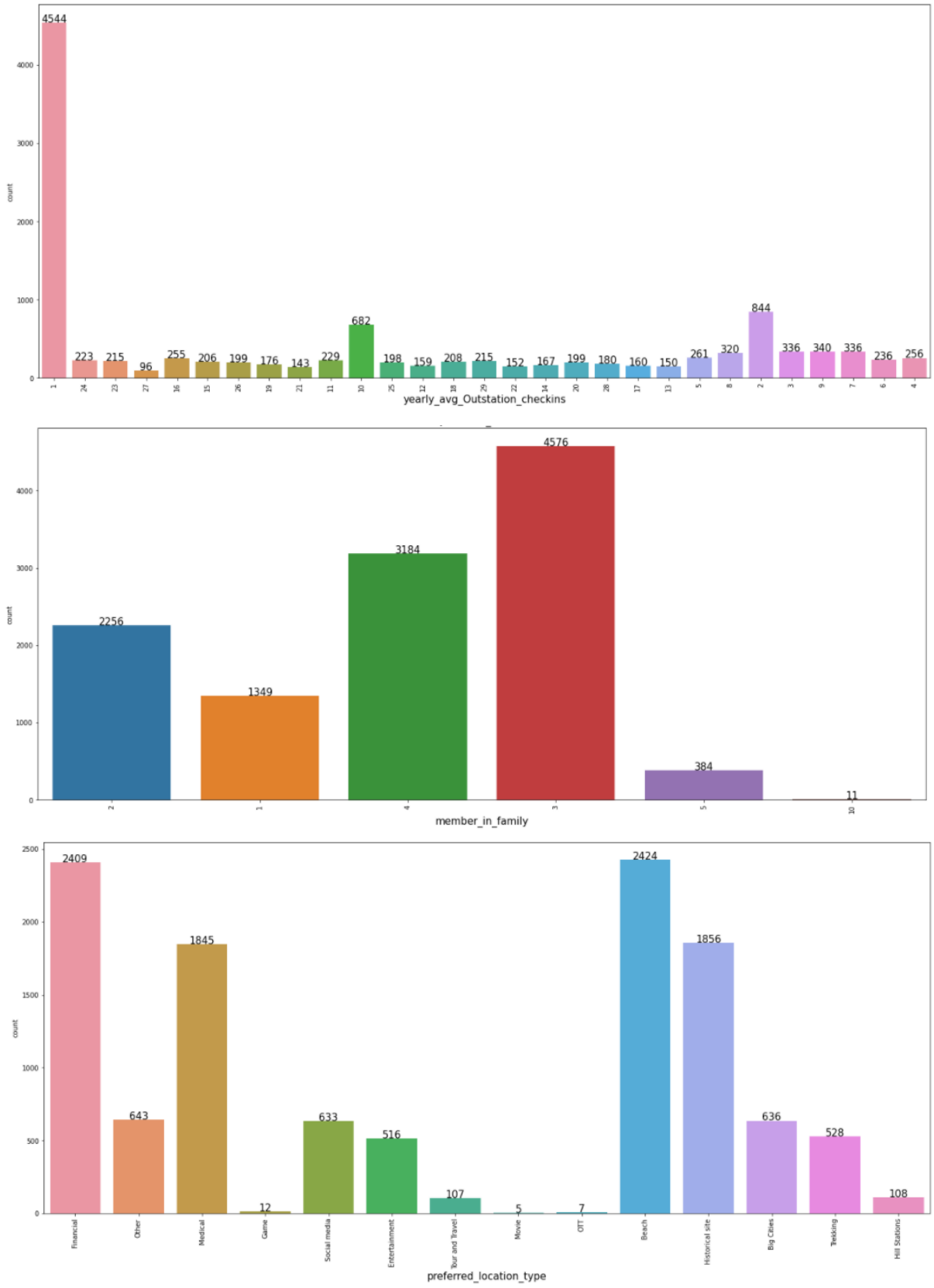
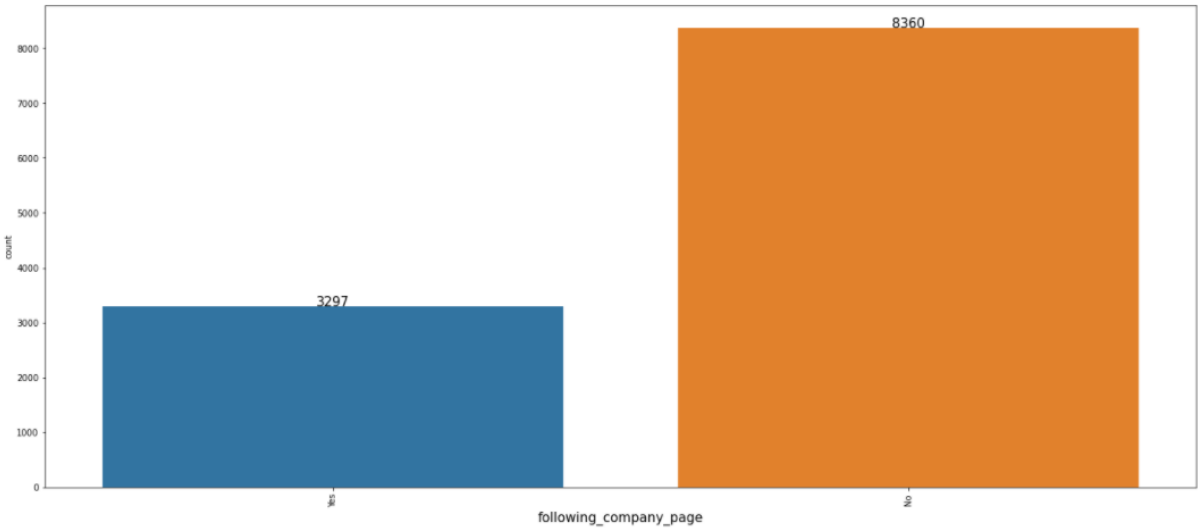
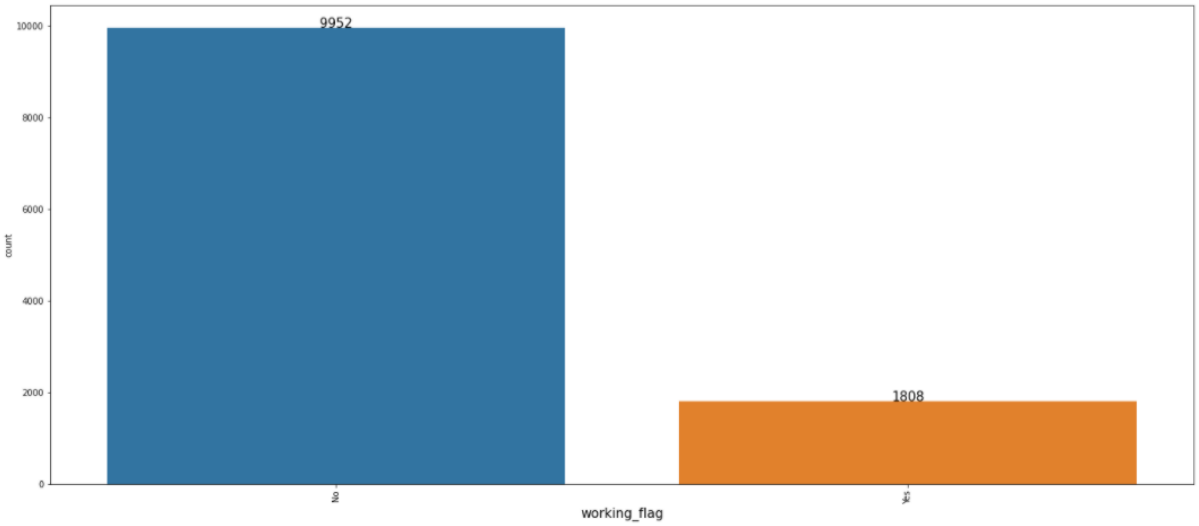

count
yearly_avg_Outstation_checkins
4544
223
215
96
255
206
199
176
143
229
682
198
159
208
215
152
167
199
180
160
150
261
320
844
336
340
336
236
256
member_in_family
2256
1349
3184
4576
384
11
preferred_location_type
2409
643
1845
12
633
516
107
5
7
2424
1856
636
528
108
Financial
Other
Medical
Game
Social media
Entertainment
Tour and Travel
Movie
OTT
Beach
Historical site
Big Cities
Trekking
Hill Stations

Fig 9


8360
3297
count
following_company_page
Yes
No


9952
1808
count
working_flag
No
Yes

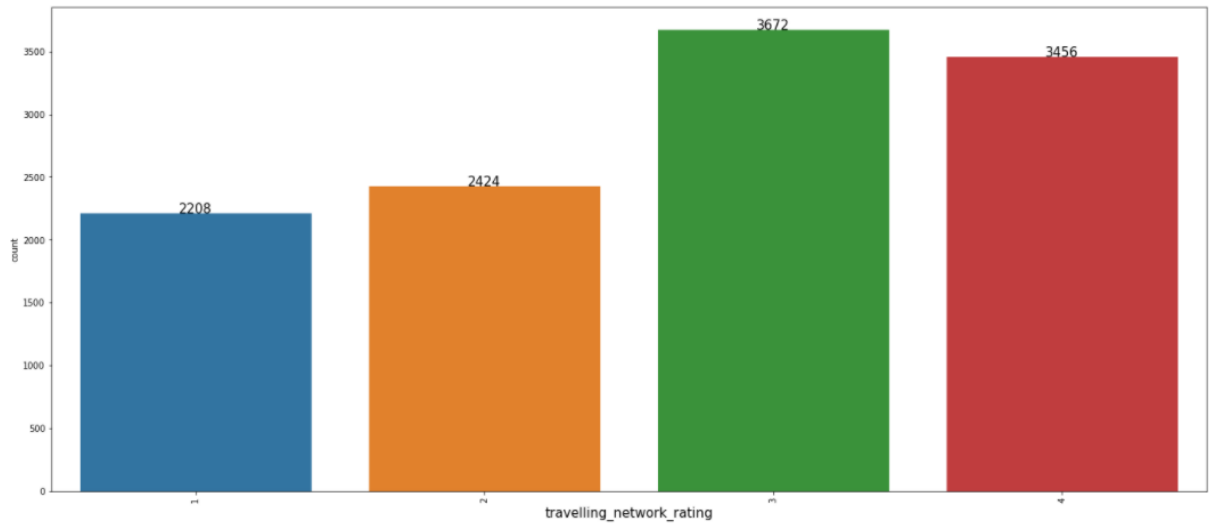

3672
3456
2424
2208
count
travelling_network_rating

Fig 10

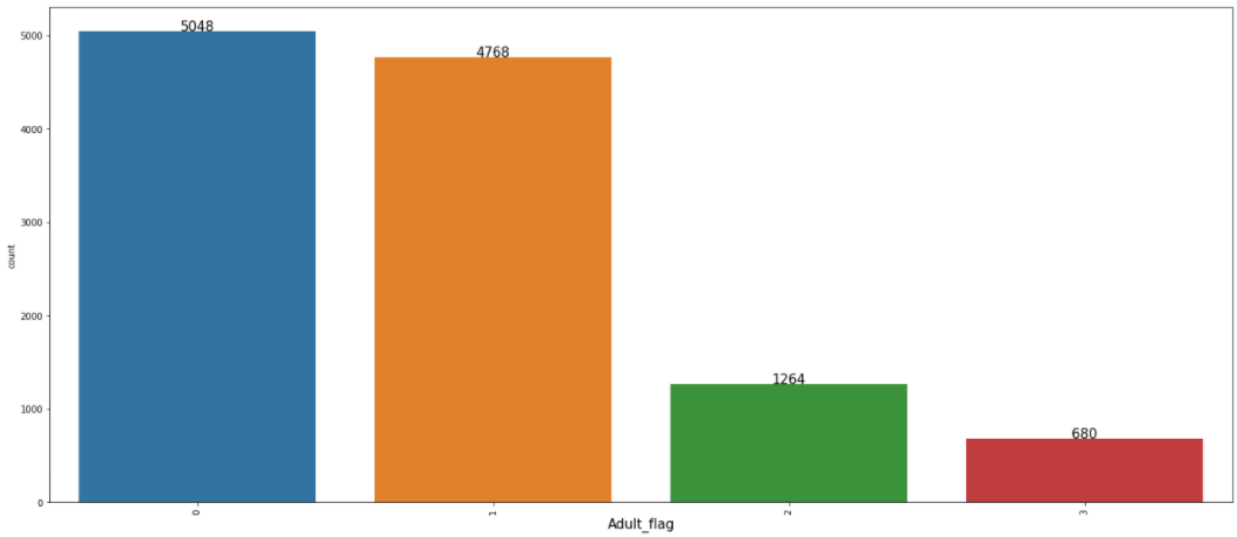


Fig 11

The figures 8, 9,10 and 11 provides the count plots of the categorical columns in the data.

1. The figure 8 shows that there are 11760 customers in the data out which only 1896 customers have purchased the product whereas 9864 customers **did not purchase the product**.
2. The figure 8 also shows that large **customer base prefers "Tab" devices**. Collectively, Mobile devices are preferred as only 1108 customers prefer **"Laptop" devices**.
3. The figure 9 shows that most of the customers travel **once per year** (4544 customers) followed by **twice per year** visits (844).
4. It also shows that most customers have **3 members** (4576 customers) in the family followed by **4 members** (3184 customers). Also, most of the customers prefer "**Beach**" (2424 customers) as their location closely followed by location for "**Financial**" purpose (2409 customers).
5. The figure 10 shows that most of the people are **not following the company page** (8360 customers) and also most of the customer base is a **non-working class** (9952 customers).
6. The figure 10 also shows that most of the customers have very few close friends that like travelling with highest being **rated 3** (3672 customers) followed by **4 rating** (3456 customers).

Bivariate Analysis

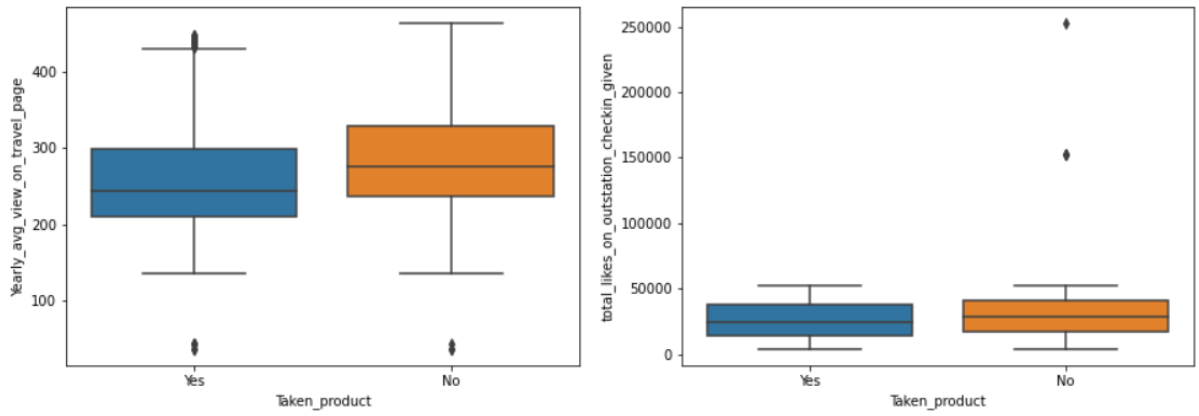


Fig 12

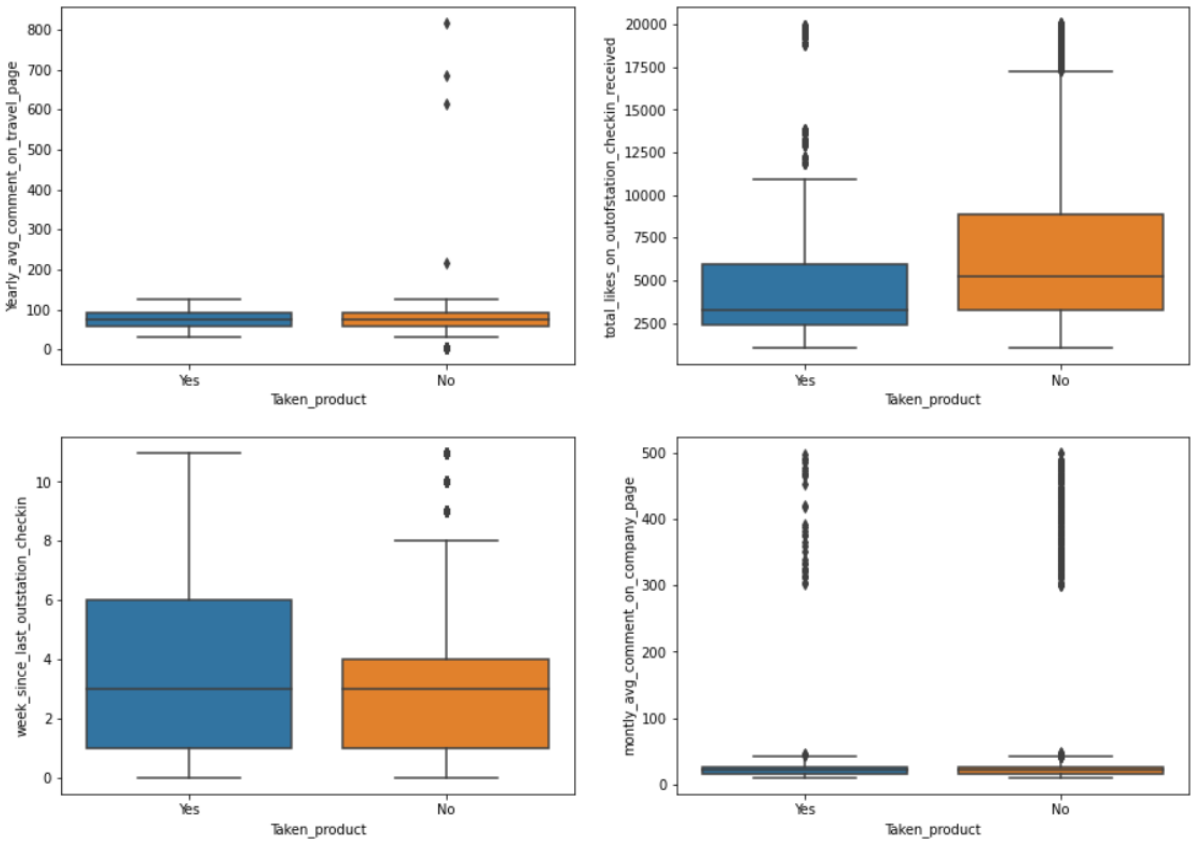


Fig 13

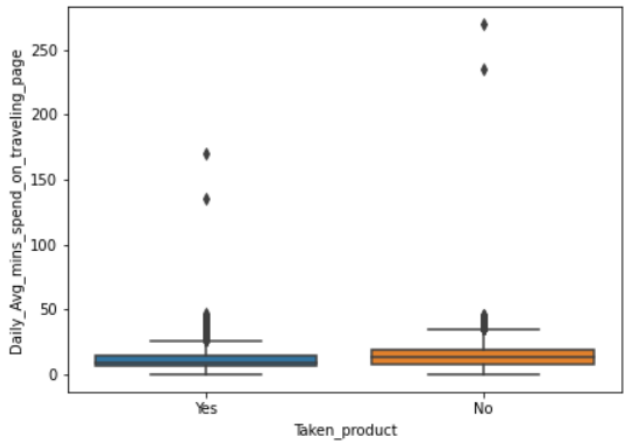


Fig 14

The figures 12, 13 and 14 provides the box plots of the numerical columns in the data.

The figure 12 suggests that despite a greater number of "**yearly average view on travel page"** the customers have not purchased the product while customers that have spent less on travel page viewing have high tendency to purchase the product.

In most of the cases **No cases** have large number outliers. In figure 13 fewer people have put **likes in outstation check-in** have purchased product whereas less people have purchased the product despite large number of likes on outstation check-in.

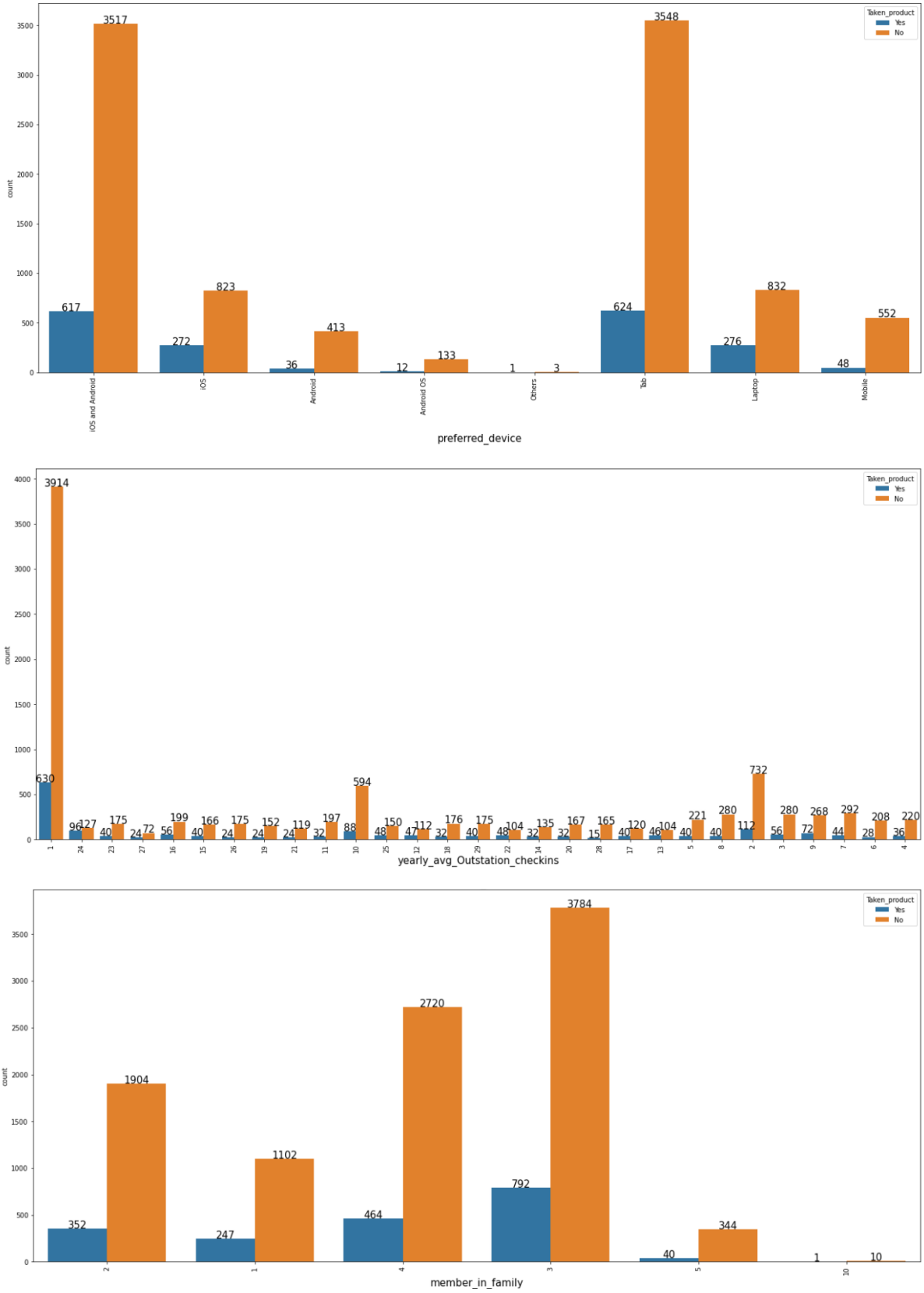
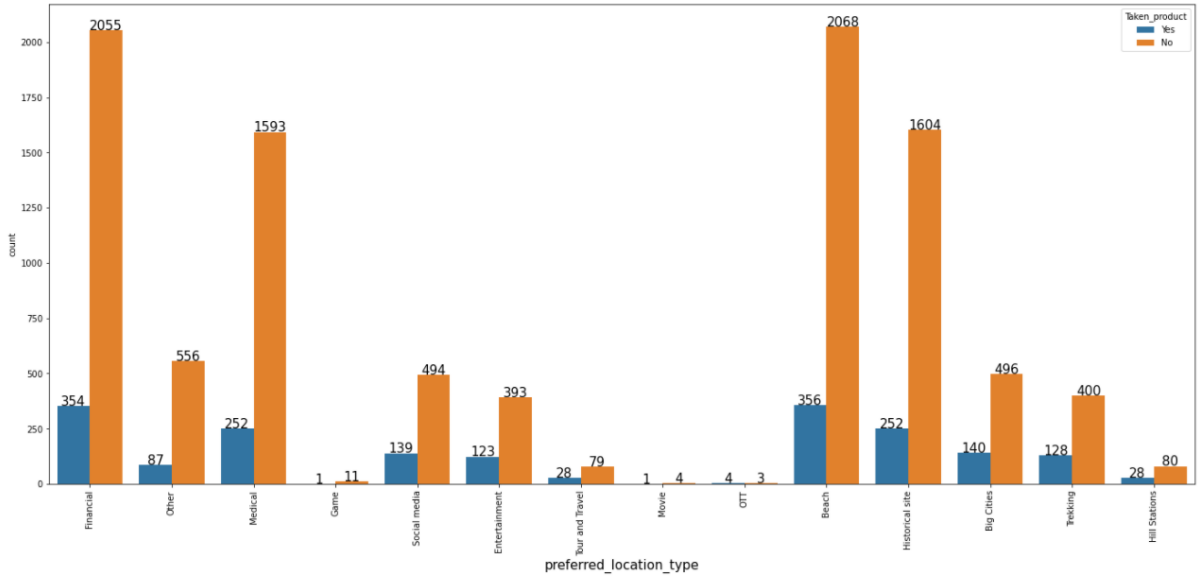
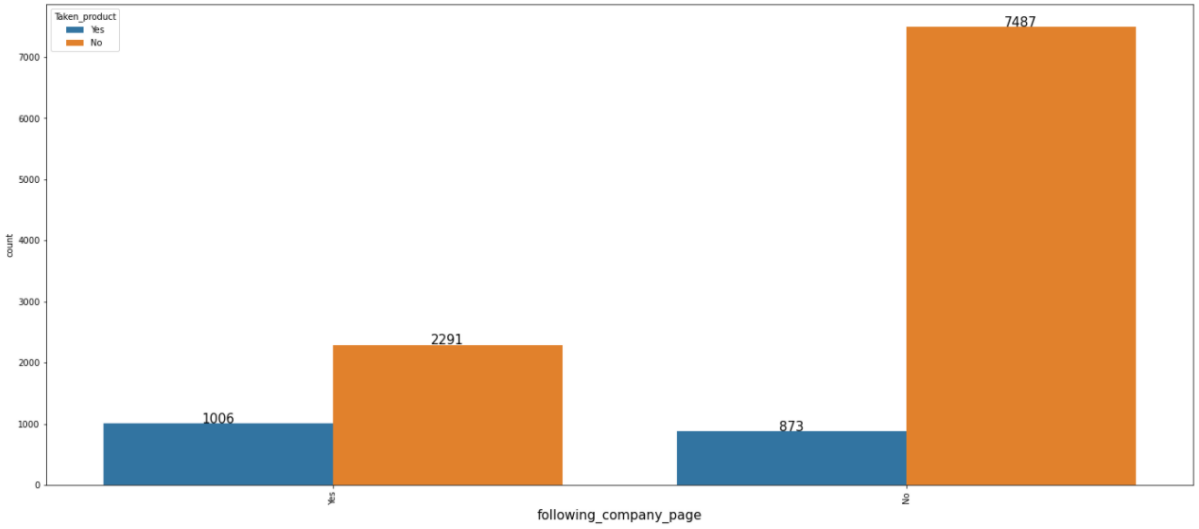

Taken_product
Yes
No
count
preferred_device
iOS and Android
iOS
Android
Android OS
Others
Tab
Laptop
Mobile
3517
617
823
272
413
36
133
12
3
1
3548
624
832
276
552
48
yearly_avg_Outstation_checkins
3914
630
member_in_family
3784
2720
1904
1102
792
464
352
247
344
40
10
1

Fig 15


Taken_product
Yes
No
count
2055
354
556
87
1593
252
1
11
494
139
393
123
79
28
1
4
4
3
2068
356
1604
252
496
140
400
128
80
28
Financial
Other
Medical
Game
Social media
Entertainment
Tour and Travel
Movie
OTT
Beach
Historical site
Big Cities
Trekking
Hill Stations
preferred_location_type


Taken_product
Yes
No
count
7487
2291
1006
873
Yes
No
following_company_page

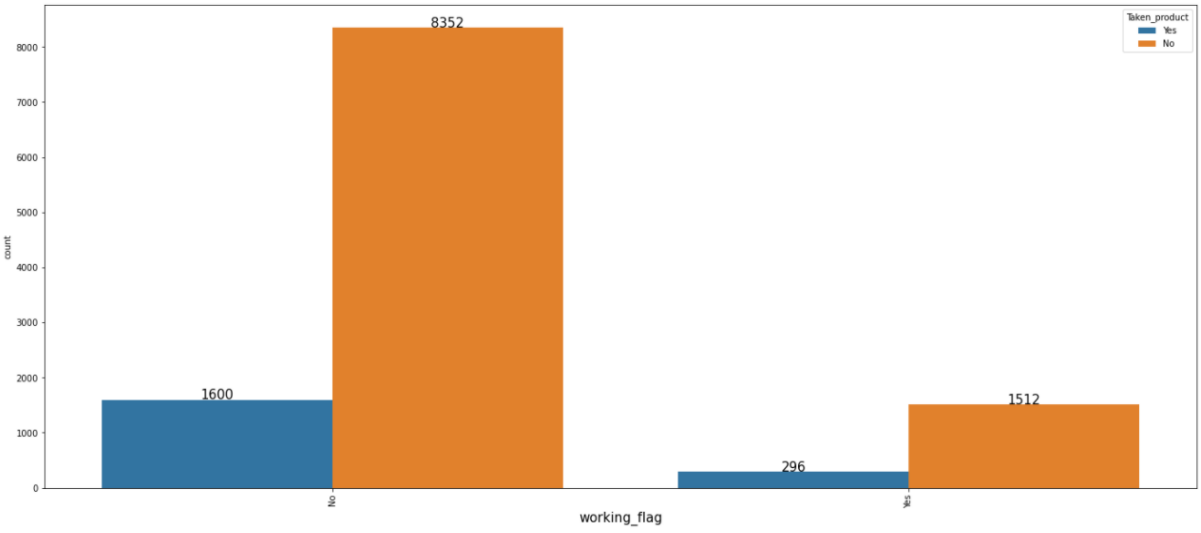

Taken_product
Yes
No
count
8352
1600
1512
296
No
Yes
working_flag

Fig 16

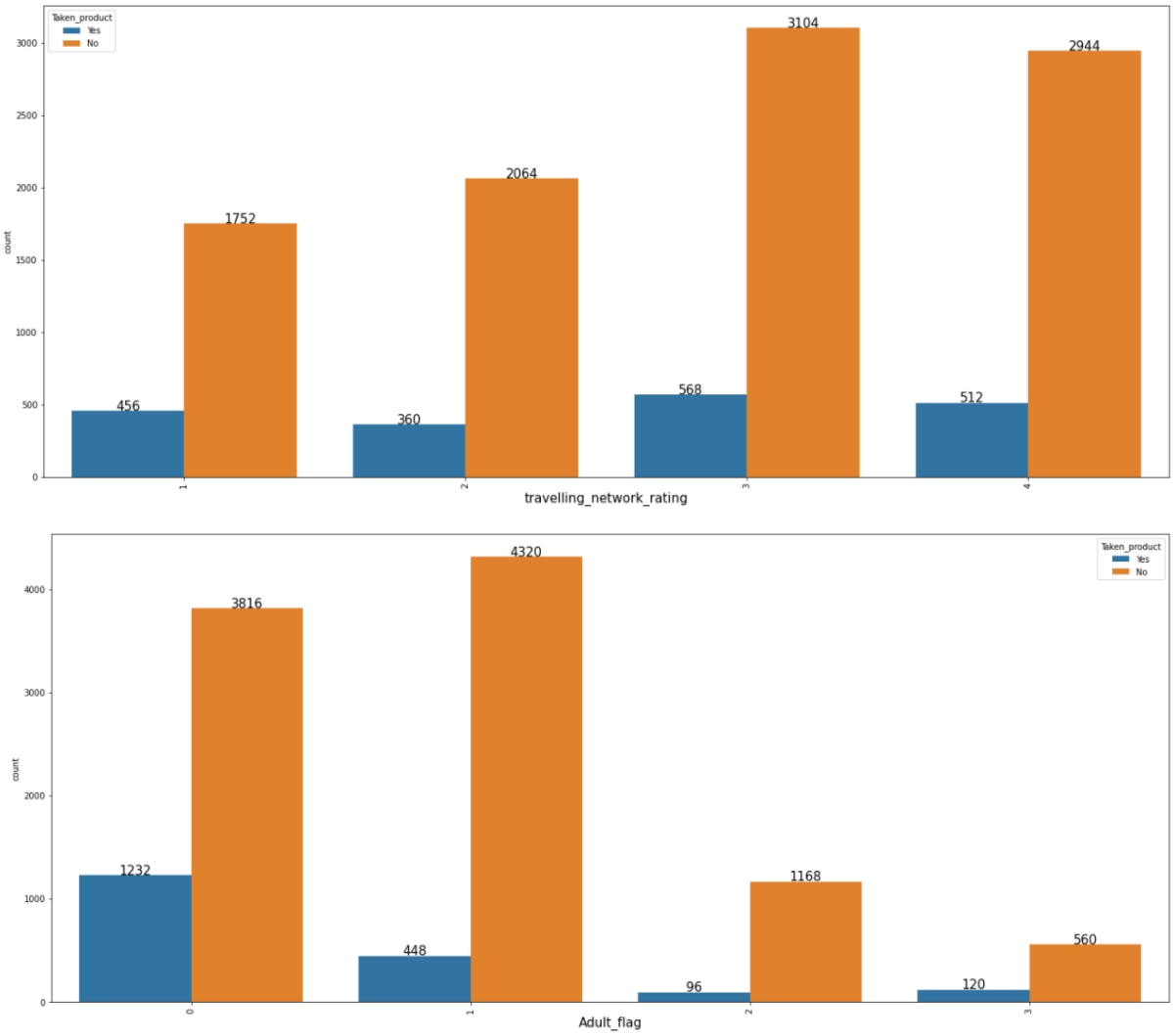


Fig 17

The figures 15, 16 and 17 provides the count plots of the categorical columns in the data.

1. The figure 15 shows that despite any preferred devices there is very high attrition rate amongst the customers as they are not interested in purchasing the product at all.
2. Also, customers that visits once in a year does not like to purchase the ticket from the company. Also, as the family size of the customer increases the chances of purchasing the product decreases.
3. The people that travels for beaches and financial purposes which are two major reasons for travel are less likely to buy the company's products. The figure 16 also shows that customers following the company's page have high chances of purchasing the products. The working category have high chances of purchasing the product as compared to the people that are not working.
4. In figure 17, we can see that customer that have close friends which love to travel have high chances of purchasing the product but as the rating increases the likelihood of purchasing a product decline.

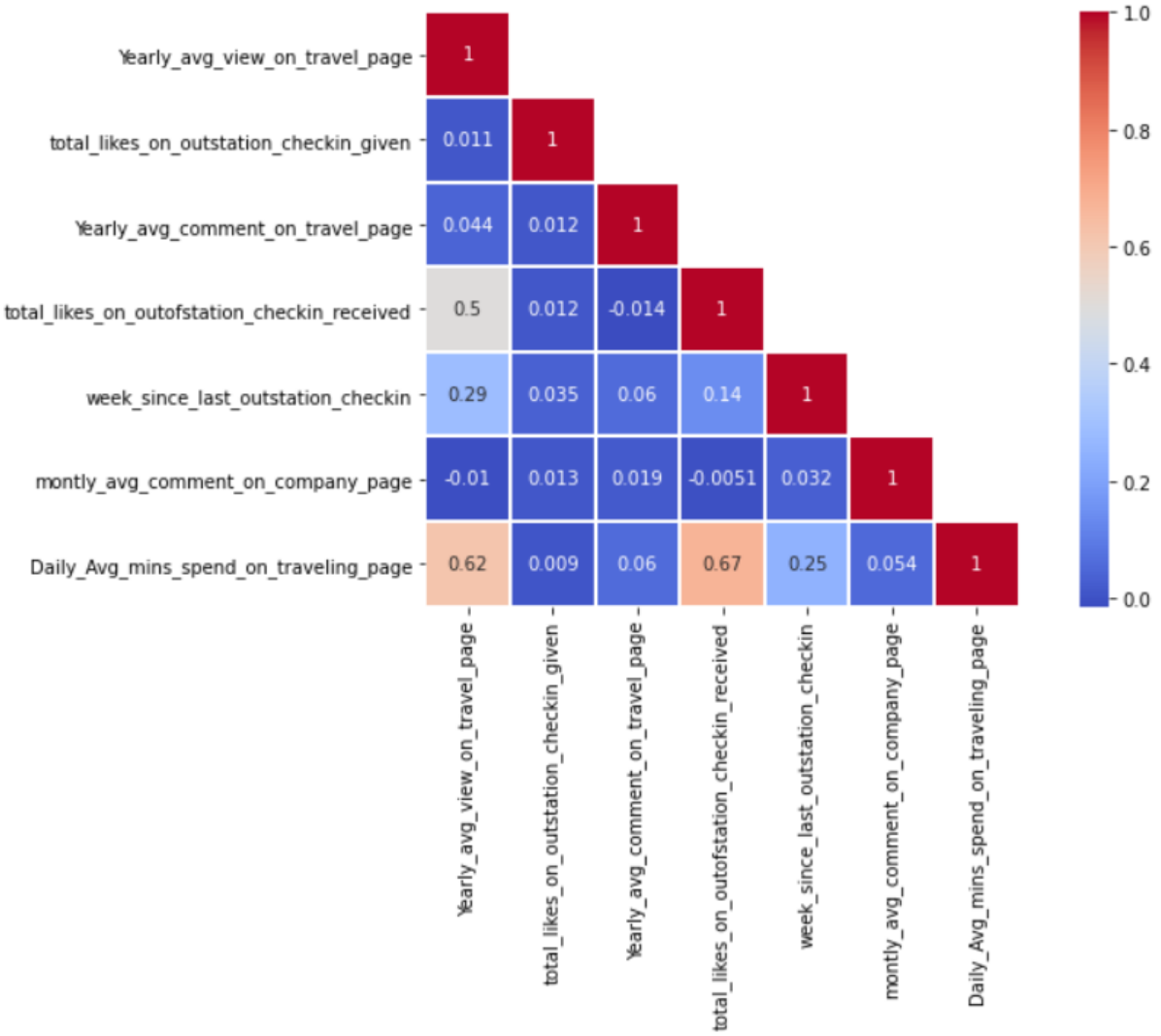


Fig 18

1. Although, there is very weak correlation amongst the but there are some variables like "**total likes on outstation received**" and "**yearly average view on travel page**" that have a moderate correlation of 0.5 between them.
2. Also, variables like "**Daily average minutes spend on travelling page**" and "**yearly average view on travel page**" also have a moderate correlation of 0.62 and "**Daily average minutes spend on travelling page**" and "**total likes on outstation received**" of moderate correlation 0f 0.67 amongst them.

Missing Value Treatment

Before the treatment there are 1430 null values in the data which are present in both numerical variables as well as in object or string variables. In order to treat the null values in numerical variables we replace the null values with median of that particular column.

Also, in order to treat the null values in categorical values we replace the null values with the mode of that particular column. After treating there are no null values left in the data.

Outlier Treatment

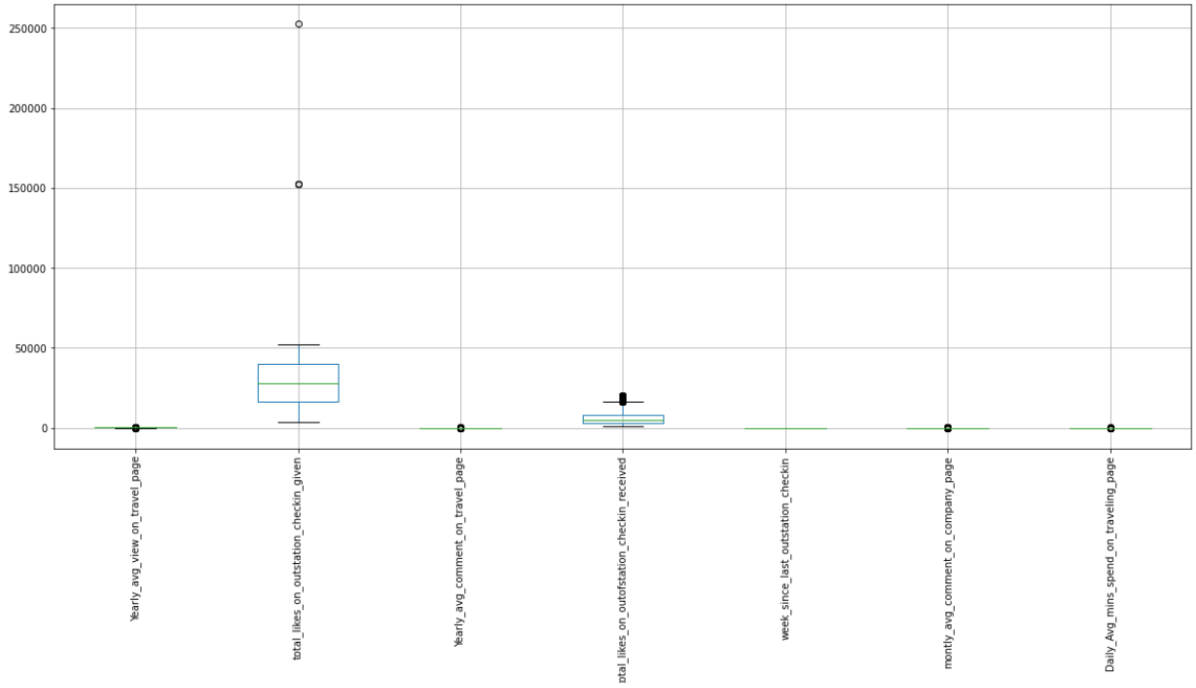


Fig 19

Before, the treatment of outliers every variable except "**week since last outstation check-in**" has outliers in the data. In order to treat those outliers in the data we replaced those outliers with the upper limit and lower limit of the particular columns.

The values in the column that are greater than the upper limit are replace with its upper limit of that column and values that are lower than lower limit are replaced with the lower limit of that column.

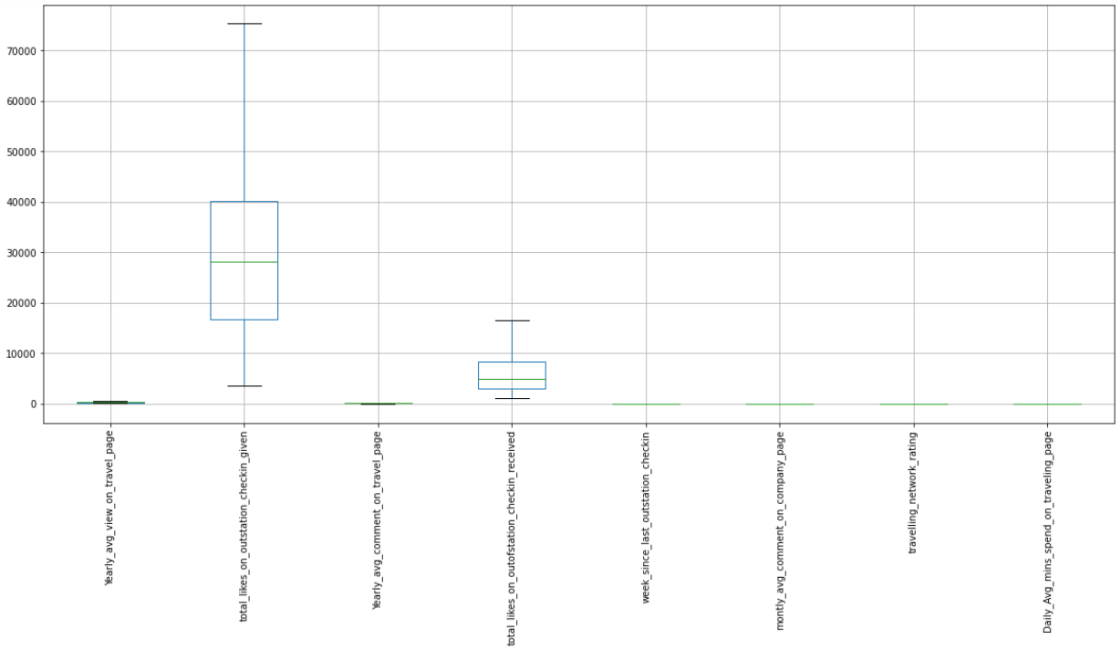


Fig 20

Variable Transformation / Addition of New Variables

Assuming that 1 is equal to "Yes" and 0 is equal to "No" and vice-versa. Several, variables are transformed.

1. The target variable named "Taken product" is transformed where "Yes" is turned to 1 and "No" is turned to 0 with the variable type of float.
2. The new variables like "**working flag label**" and "**following company label**" are added where "Yes" is turned to 1 and "No" is turned to 0 from variables like "**working flag**" and "**following company label**" respectively with the variable type of float.
3. Another new variable that is "**Laptop/ Mobile**" is added where all the other prefer devices except for Laptops were marked as Mobiles. This further leads to the addition of new variable named "**Laptop/Mobile label**" where "Laptop" is turned to 0 and "Mobile" is turned to 1 with the variable type of float.
4. Some variables like "**member in the family**", "**yearly average outstation check-in**" and "**Adult flag**" are converted to float variable type. Also, "**travelling network rating**" is converted to category variable.
5. Also, "**preferred location type label**" where Location is arranged from 1-14 with 14 being marked as the most preferred location and 1 as least preferred location from the "**preferred location type** ".

Removal of Unwanted Variables

	variables	VIF
1	Yearly_avg_view_on_travel_page	27.707664
15	preferred_location_type_lbl	17.408071
8	montly_avg_comment_on_company_page	13.595628
5	Yearly_avg_comment_on_travel_page	12.837435
11	Daily_Avg_mins_spend_on_traveling_page	10.764774
14	Laptop/Mobile_lbl	8.878148
4	member_in_family	8.687737
6	total_likes_on_outofstation_checkin_received	6.864445
9	travelling_network_rating	6.858929
2	total_likes_on_outstation_checkin_given	4.935385
7	week_since_last_outstation_checkin	2.869913
3	yearly_avg_Outstation_checkins	1.927082
10	Adult_flag	1.913456
13	following_company_page_lbl	1.500105
12	working_flag_lbl	1.455599
0	Taken_product	1.346810

Fig 21

All though most of the variables have their variance inflation factor less than 10. There are some variables like "**Yearly average view on travel page**", "**monthly average comment on company page**" and "**preferred location type label**" which have inflation greater than 10.

So, we drop these variables one by one till we have all the variables whose variance inflation factor less than 10. Hence after dropping the variables one by one we obtain the following variables.

	variables	VIF
4	Yearly_avg_comment_on_travel_page	9.963561
9	Daily_Avg_mins_spend_on_traveling_page	8.911769
12	Laptop/Mobile_lbl	8.124978
3	member_in_family	7.706776
5	total_likes_on_outofstation_checkin_received	6.663508
7	travelling_network_rating	6.304482
1	total_likes_on_outstation_checkin_given	4.713784
6	week_since_last_outstation_checkin	2.807432
2	yearly_avg_Outstation_checkins	1.903982
8	Adult_flag	1.899842
11	following_company_page_lbl	1.493347
0	Taken_product	1.342935
10	working_flag_lbl	1.176109

Fig 22

Data Insights

```
0.0    9864
1.0    1896
Name: Taken_product, dtype: int64


Normalized Score is
 0.0    0.838776
1.0    0.161224
Name: Taken_product, dtype: float64
```

Fig 23

The data is highly imbalanced as out of 11760 customers there are 9864 customers that are not interested in purchasing our product which constitutes around 83.9%. This can be treated with the help of random based over sampling or random based under sampling. It can also be treated using K- fold cross validation.

This shows that there are customers that may purchasing the product of some other company rather than preferring the product of this company. It can also be attributed to the fact that customers are not satisfied with company's service and ae switching to other companies.

Despite performing clustering and taking the cluster value minimum as 2 we observed that silhouette width was coming to negative. Despite increasing the clusters, the negative score increased which can be attributed to the fact that data is highly imbalanced and due to which the boundaries are very less. Even after continuing with this no conclusive conclusion can be drawn from it.