PGP - Data Science and Business Analytics

CAPSTONE PROJECT FINAL REPORT

Akshaya Nallathambi

3rd April, 2022

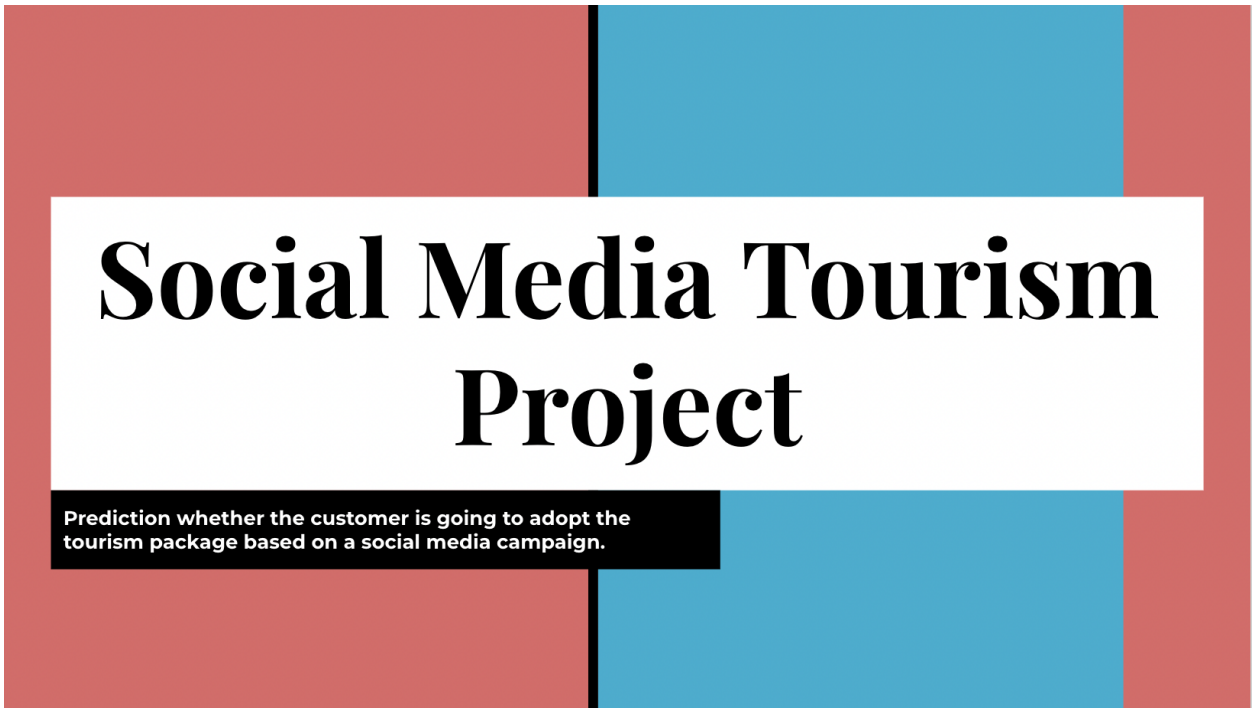


Table Of Contents

List of Figures

List of Graphs

List of Tables

Problem

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 behavior 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.

Propensity of buying tickets is different for different login devices. Hence, you have to create 2 models separately for Laptop and Mobile. [Anything which is not a laptop can be considered as mobile phone usage.]

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

Data Description-

Variable	Description
UserID	Unique ID of user
Taken_product	Whether they bought the product or not
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 relationship mentioned by user in the account
preferred_location_type	Preferred type of the location for traveling of user
Yearly_avg_comment_on_travel_page	Average yearly comments on any travel related page by user
total_likes_on_outofstation_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	Whether the customer is following company page (Yes or No)
montly_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	Whether 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

1) Introduction

-Brief introduction about the problem statement and the need of solving it.

Digital marketing is a coordinated marketing effort to reinforce or assist with a business goal using one or more social media platforms. Campaigns differ from everyday social media efforts because of their increased focus, targeting and measurability. We have an aviation company who is trying to make use of a digital marketing platform to come up with a plan for the targeted customers. They have collaborated with a social networking platform to understand the behavior of the customers. The purchasing behavior is device specific – Laptop and Mobile (Models are run separately for each device type). The models used should be accurate as the advertisements for digital platforms are expensive. Now my goal for this project is to predict whether the customer is going to adopt the tourism package based on the social media campaign. Social media lets you reach out to a greater population with information about your business. If you can use social media well, you can improve awareness of your brand, increase the number of visitors to your website, and earn more money. Without it, you may not be able to do well in your business as we are currently living in the tech era.

This project will help the aviation company get a clearer picture about their customers and help them according to their expectation. Targeted advertisement not only helps the company's revenue grow but also bring satisfaction to the customers as their individual need is met. Once the advertisements are digitized, it also helps us eradicate the mundane task of calling each and every customer to check if they want to adopt a tour package. If the company has a clear idea on their customers they can come up with a strategy to improve business. Various tour packages can be framed and sent to the specific group of people.

The business opportunity here is that it helps the company give attention to targeted audiences. This will in turn attract more customers to buy the product. Eventually, this will lead to an increase in revenue. Once the company revenue increases, the company size will also grow and the company will expand. A company with good growth will have various employment opportunities. This can be considered the social opportunity of this project. Not only job generation, from the customer's perspective, the company can give discounts on the tour packages or tickets and it can deliver the products on time to the customer without any delay. This will improve the customer experience which will again increase the revenue. Thus, the company can have better business opportunities as well as better social opportunities.

2) EDA and Business Implication

-Univariate / Bi-variate / Multivariate analysis to understand relationship b/w variables. How is your analysis impacting the business?
-Both visual and non-visual understanding of the data.

The data given has 11760 entries with 17 variables. The year when the data set was collected is not mentioned in the given data. There are 3 variables ("Yearly_avg_view_on_travel_page", "yearly_avg_Outstation_checkins" and "Yearly_avg_comment_on_travel_page") that have data points taken for a year. There is 1 variable ("montly_avg_comment_on_company_page") that has the data points taken for a month. There is some data taken on a daily basis and some on a weekly basis. Then the average of these data is taken and added as a part of the dataset.

The below figures give the first 5 rows of sample data.

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

FIGURE 1

member_in_family	preferred_location_type	Yearly_avg_comment_on_travel_page	total_likes_on_outofstation_checkin_received	week_since_last_outstation_checkin
2	Financial	94.0	5993	8
1	Financial	61.0	5130	1
2	Other	92.0	2090	6
4	Financial	56.0	2909	1
1	Medical	40.0	3468	9

FIGURE 2

following_company_page	montly_avg_comment_on_company_page	working_flag	travelling_network_rating	Adult_flag	Daily_Avg_mins_spend_on_traveling_page
Yes	11	No	1	0	8
No	23	Yes	4	1	10
Yes	15	No	2	0	7
Yes	11	No	3	0	8
No	12	No	4	1	6

FIGURE 3

The image below gives the basic information of the data set. There are 17 variables, out of which 3 variables are of type float, 7 variables are of type int and 7 variables are of type object. The data given is for 11760 individuals. There are null values that require processing.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11760 entries, 0 to 11759
Data columns (total 17 columns):
 #   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)
memory usage: 1.5+ MB
```

FIGURE 4

The below image gives the five point summary of the continuous variables in the data set.

	count	mean	std	min	25%	50%	75%	max
UserID	11760.0	1.005880e+06	3394.963917	1000001.0	1002940.75	1005880.5	1008820.25	1011760.0
Yearly_avg_view_on_travel_page	11179.0	2.808308e+02	68.182958	35.0	232.00	271.0	324.00	464.0
total_likes_on_outstation_checkin_given	11379.0	2.817048e+04	14385.032134	3570.0	16380.00	28076.0	40525.00	252430.0
Yearly_avg_comment_on_travel_page	11554.0	7.479003e+01	24.026650	3.0	57.00	75.0	92.00	815.0
total_likes_on_outofstation_checkin_received	11760.0	6.531699e+03	4706.613785	1009.0	2940.75	4948.0	8393.25	20065.0
week_since_last_outstation_checkin	11760.0	3.203571e+00	2.616365	0.0	1.00	3.0	5.00	11.0
montly_avg_comment_on_company_page	11760.0	2.866156e+01	48.660504	11.0	17.00	22.0	27.00	500.0
travelling_network_rating	11760.0	2.712245e+00	1.080887	1.0	2.00	3.0	4.00	4.0
Adult_flag	11760.0	7.938776e-01	0.851823	0.0	0.00	1.0	1.00	3.0
Daily_Avg_mins_spend_on_traveling_page	11760.0	1.381743e+01	9.070657	0.0	8.00	12.0	18.00	270.0

FIGURE 5

UNIVARIATE ANALYSIS

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and its major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. [1]

(i) Continuous Variables

Skewness

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. [2]

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

TABLE 1

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to each variable in the given data set.

	Skewness
Yearly_avg_view_on_travel_page	0.446079
total_likes_on_outstation_checkin_given	0.498350
yearly_avg_Outstation_checkins	0.977120
Yearly_avg_comment_on_travel_page	4.910321
total_likes_on_outofstation_checkin_received	1.368404
week_since_last_outstation_checkin	0.915217
montly_avg_comment_on_company_page	7.683170
travelling_network_rating	-0.302518
Daily_Avg_mins_spend_on_traveling_page	4.480111

FIGURE 6

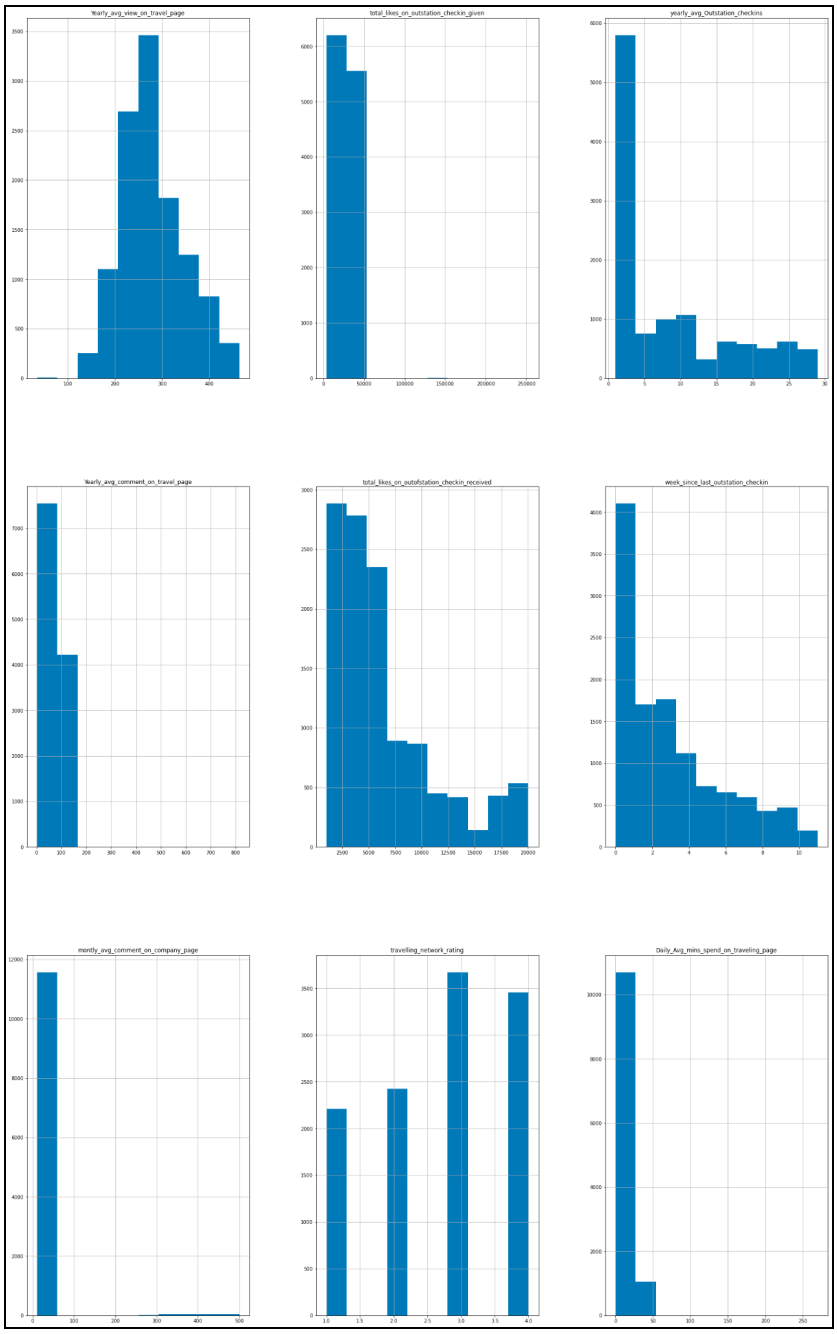
Right skewed variables
Yearly_avg_view_on_travel_page
Total_likes_on_outstation_checkin_given
Yearly_avg_Outstation_checkins
Yearly_avg_comment_on_travel_page
Total_likes_on_outofstation_checkin_received
Week_since_last_outstation_checkin
Montly_avg_comment_on_company_page
Daily_Avg_mins_spend_on_traveling_page
Left skewed variables
Travelling_network_rating

TABLE 2

Histogram

The **histograms** are used for **numerical variables** to perform univariate analysis.

It is clear from the graph (*Graph 1*) that all the numerical variables are skewed.

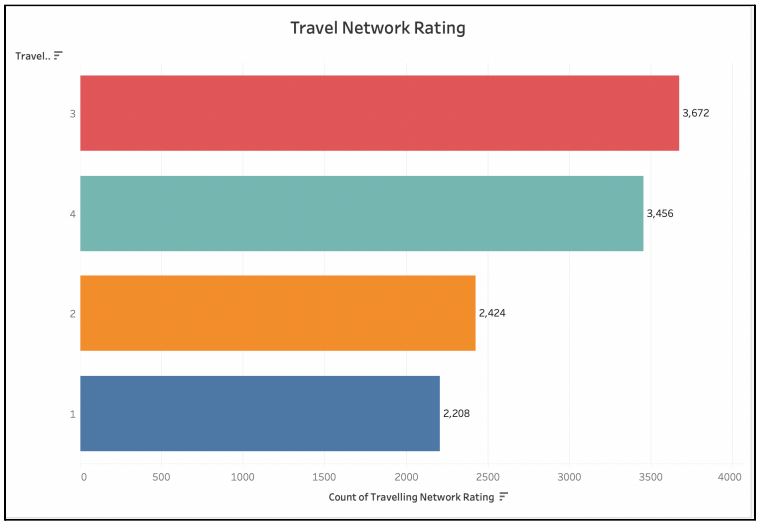


GRAPH 1

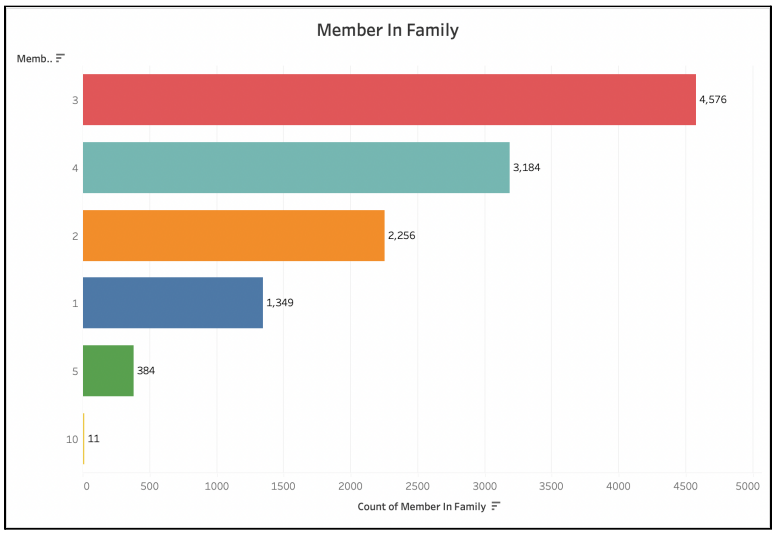
(ii) Categorical Variables

Count Plot

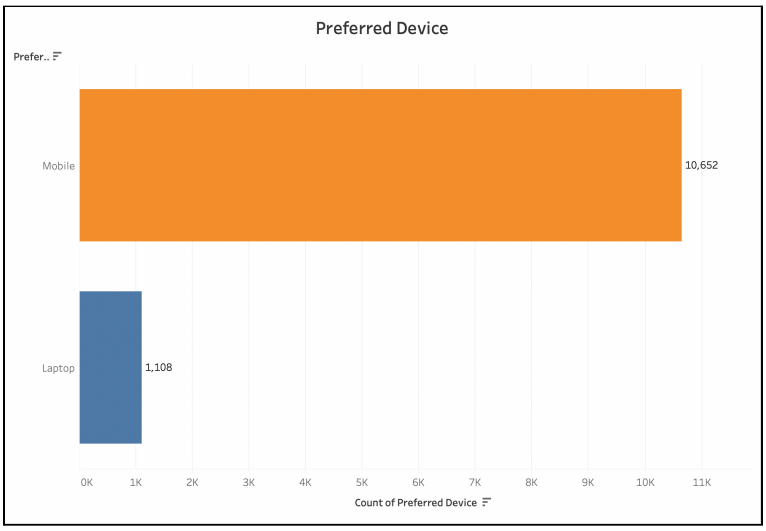
The **count plots** are used for **categorical variables** to perform univariate analysis. It gives the count of each category in a particular variable. I have created it for four categorical variables and sorted them in descending order.



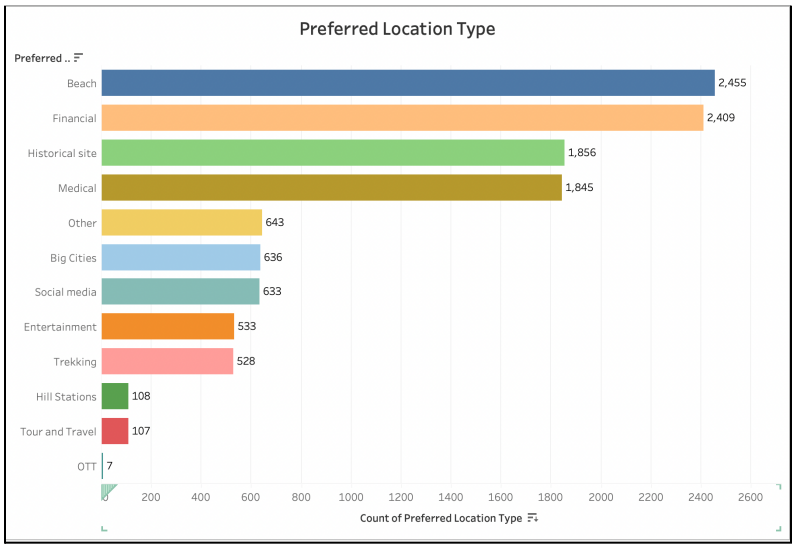
GRAPH 2



GRAPH 3



GRAPH 4



GRAPH 5

BIVARIATE AND MULTIVARIATE ANALYSIS

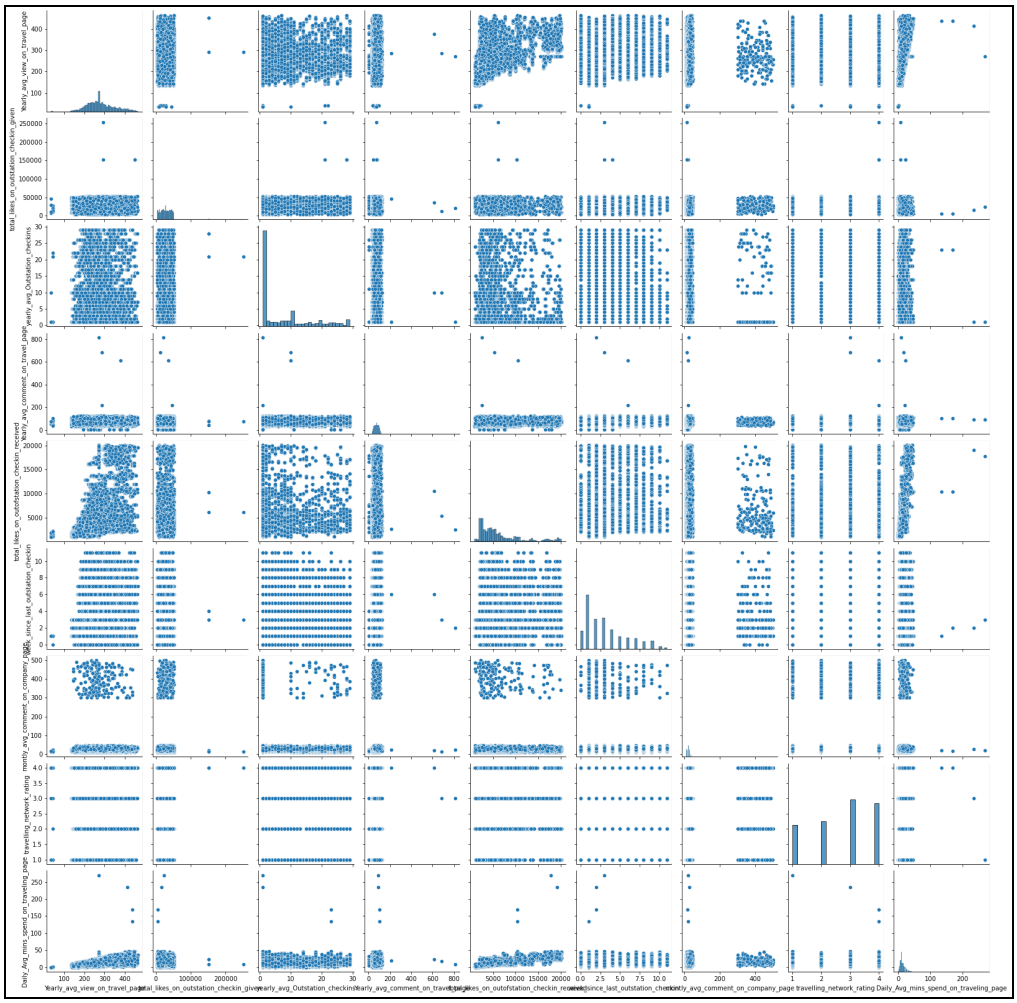
Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. [3]

The pairplot is generally used for numerical variables and box plots are used for categorical with numerical variables to perform bivariate analysis.

(i) Continuous variables

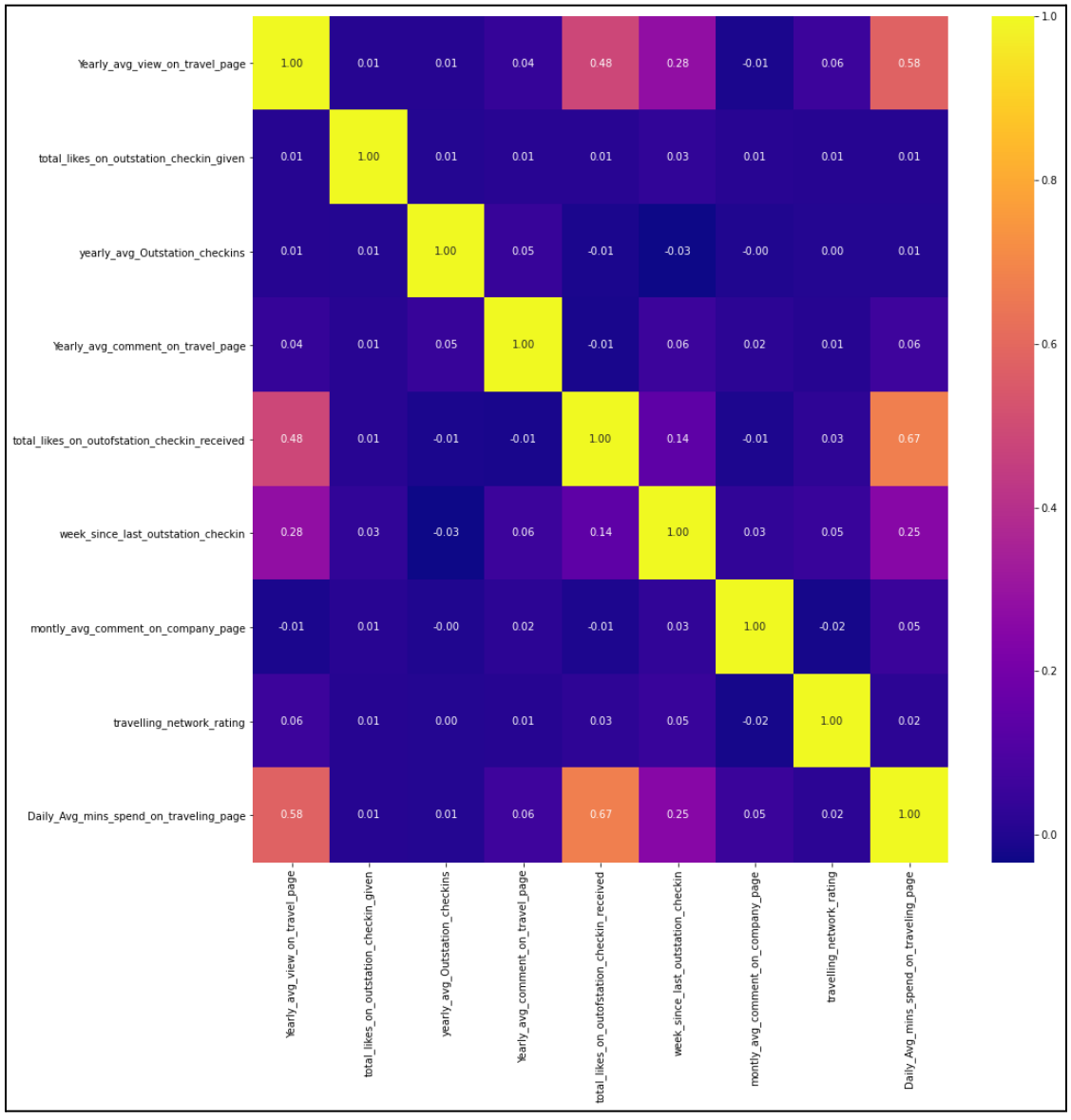
Pair Plot and Heat Maps

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. [6]



GRAPH 6

The heat map can also be used to check the association between two numeric variables. All the boxes with a value higher than 0.8 are highly correlated. But in the given data set none of the variables have a value 0.8 or more. The heat map for all the numerical variables is below-



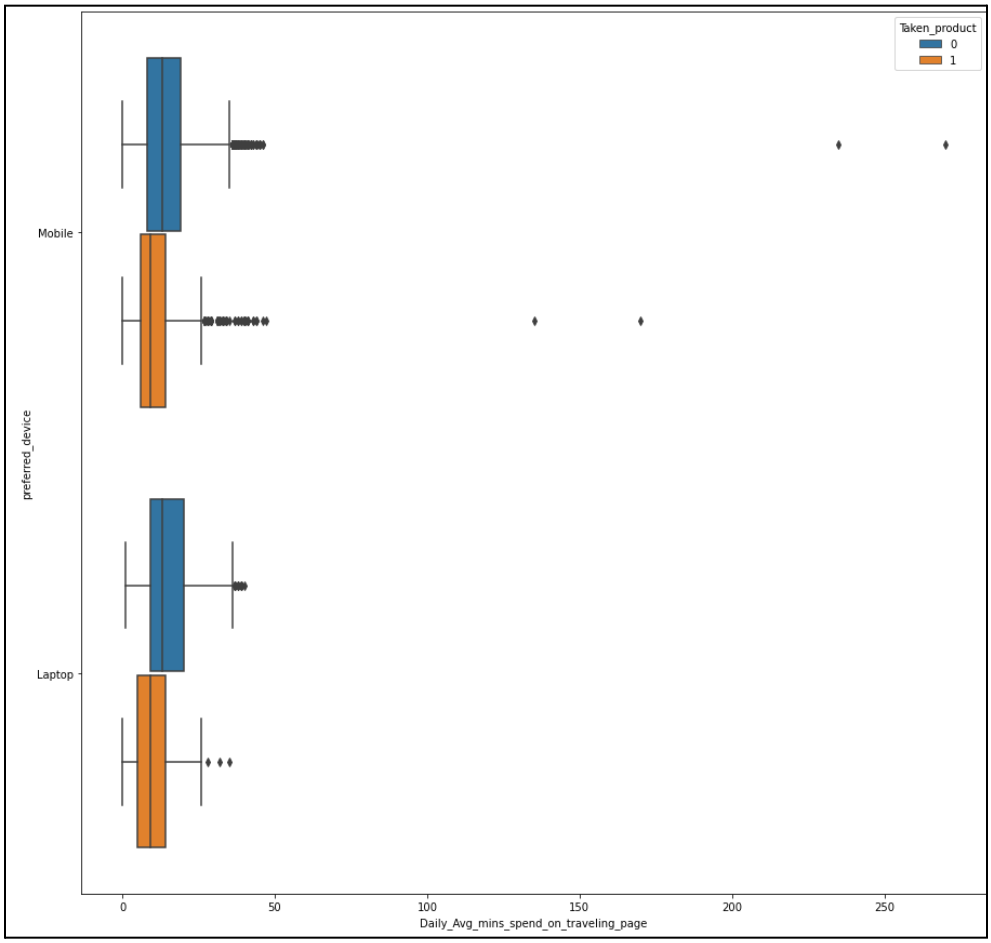
GRAPH 7

The pair plot and the heat map doesn't show much correlation between the variables.

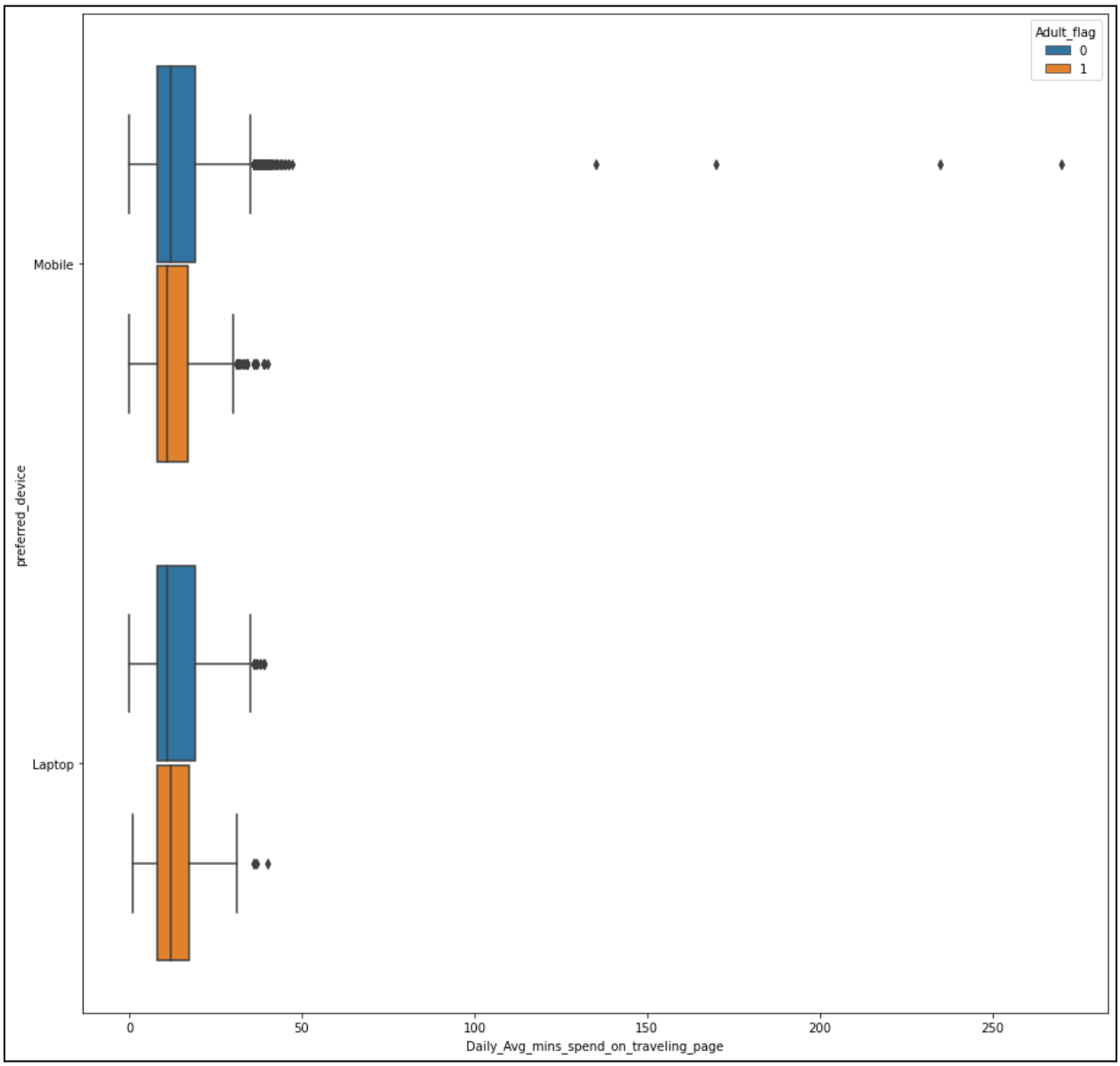
(ii) Categorical variables

Box plot

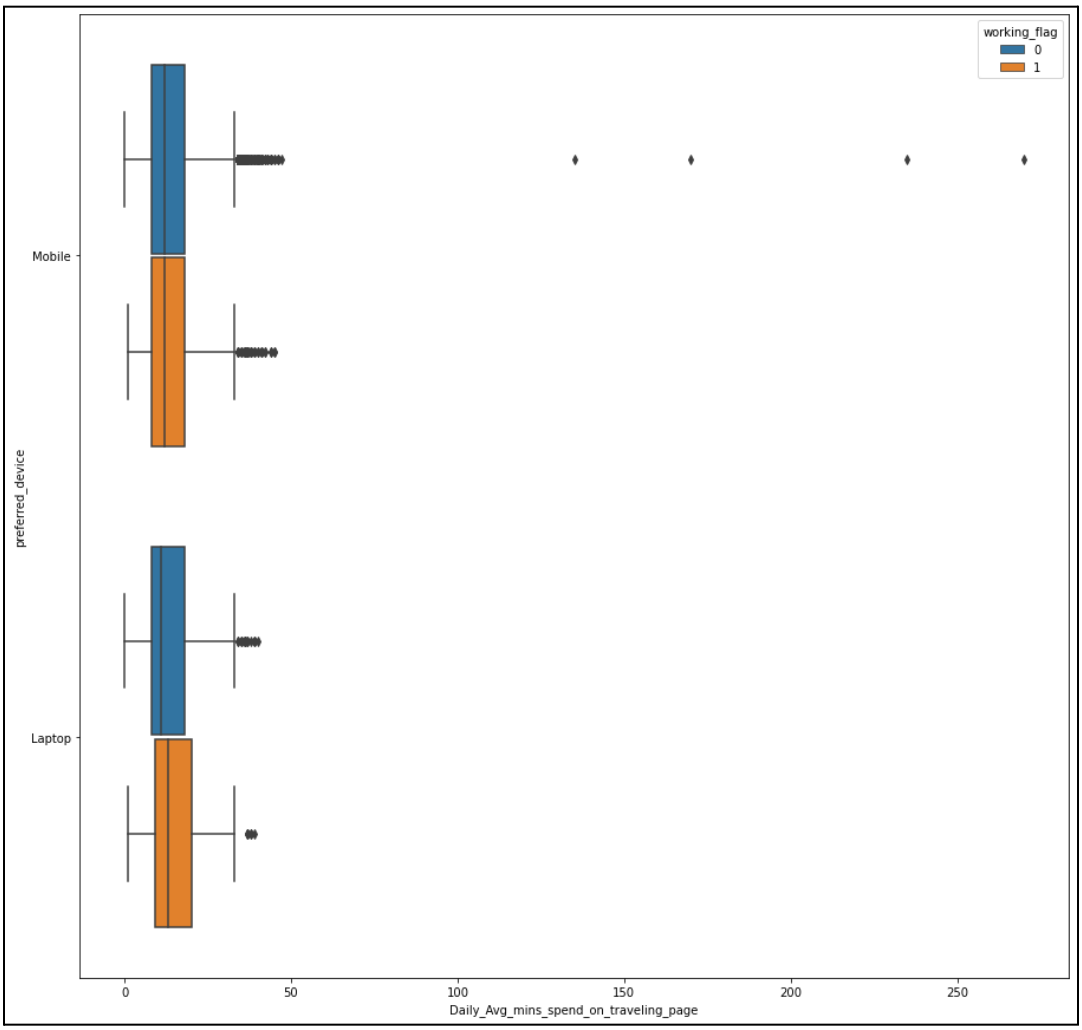
A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed. [4]



GRAPH 8


Adult_flag
0
1
Mobile
Laptop
preferred_device
0
50
100
150
200
250
Daily_Avg_mins_spend_on_traveling_page

GRAPH 9



GRAPH 10

Findings from the graphs -

- The device mobile and laptop almost have the same median when it comes to users taking the product.
- For adults, both the devices are used majorly.
- There are almost equal number of Mobile users who have taken the product and those who have not taken the product.

3) Data Cleaning and Preprocessing

- Approach used for identifying and treating missing values and outlier treatment (and why)

There were a lot of missing variables in the given data set. It can be seen in the below image where the unhighlighted variables have null values. The missing values were treated with imputation. It is not advisable to drop the missing values as there were only a few row values missing and they can be imputed.

Column	Non-Null Count	Dtype
UserID	11760 non-null	int64
Taken_product	11760 non-null	object
Yearly_avg_view_on_travel_page	11179 non-null	float64
preferred_device	11707 non-null	object
total_likes_on_outstation_checkin_given	11379 non-null	float64
yearly_avg_Outstation_checkins	11685 non-null	object
member_in_family	11760 non-null	object
preferred_location_type	11729 non-null	object
Yearly_avg_comment_on_travel_page	11554 non-null	float64
total_likes_on_outofstation_checkin_received	11760 non-null	int64
week_since_last_outstation_checkin	11760 non-null	int64
following_company_page	11657 non-null	object
montly_avg_comment_on_company_page	11760 non-null	int64
working_flag	11760 non-null	object
travelling_network_rating	11760 non-null	int64
Adult_flag	11760 non-null	int64
Daily_Avg_mins_spend_on_traveling_page	11760 non-null	int64

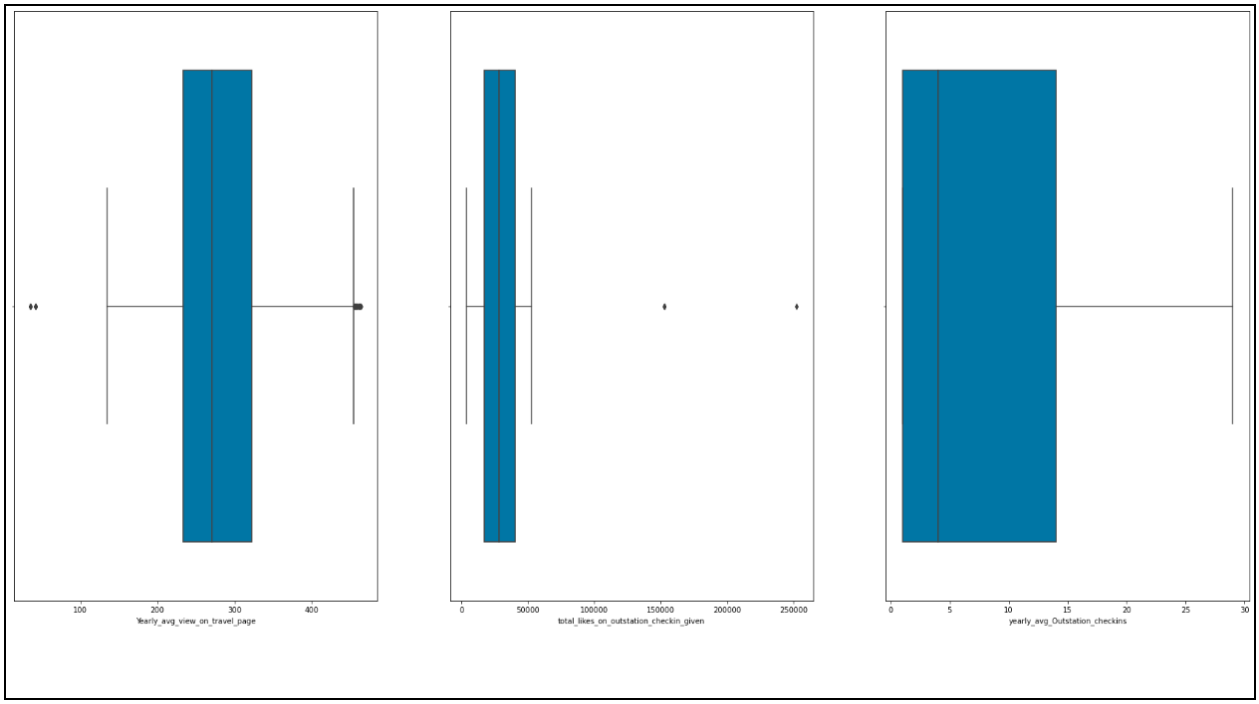
FIGURE 7

The variables ***Yearly_avg_view_on_travel_page, preferred_device, total_likes_on_outstation_checkin_given, yearly_avg_Outstation_checkins, preferred_location_type, Yearly_avg_comment_on_travel_page*** and ***following_company_page*** had to be treated for missing values. The object type variables were imputed with mode function and the numerical variables were imputed with median function. As the mode and median are resistant to outliers, these were chosen for imputation.

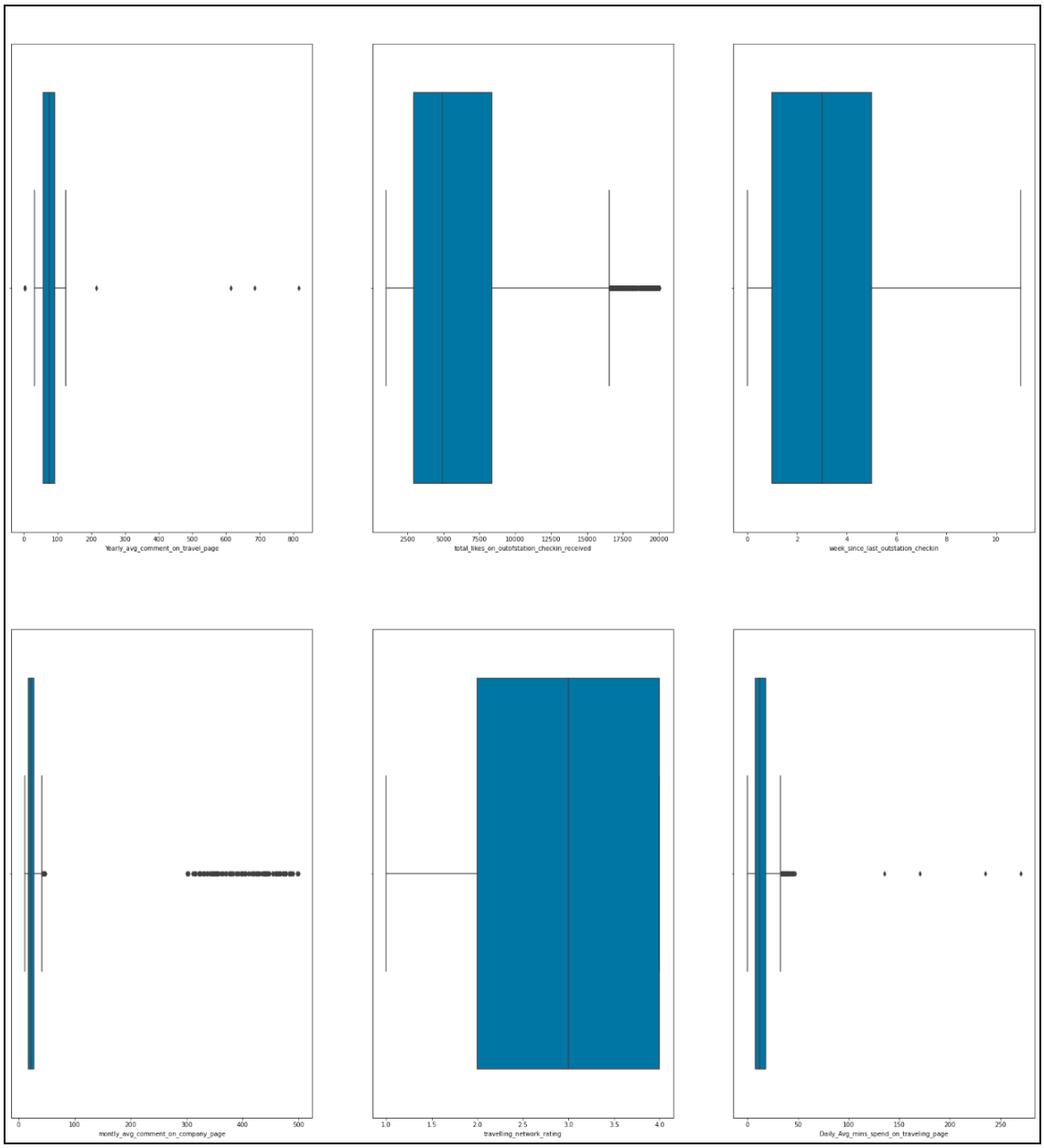
There are outliers in the below mentioned variables. This is evident from the box plots. (Graph 11 & 12).

Yearly_avg_view_on_travel_page
Total_likes_on_outstation_checkin_given
Yearly_avg_comment_on_travel_page
Total_likes_on_outofstation_checkin_received
Montly_avg_comment_on_company_page
Daily_Avg_mins_spend_on_traveling_page

Before outlier treatment-



GRAPH 11



GRAPH 12

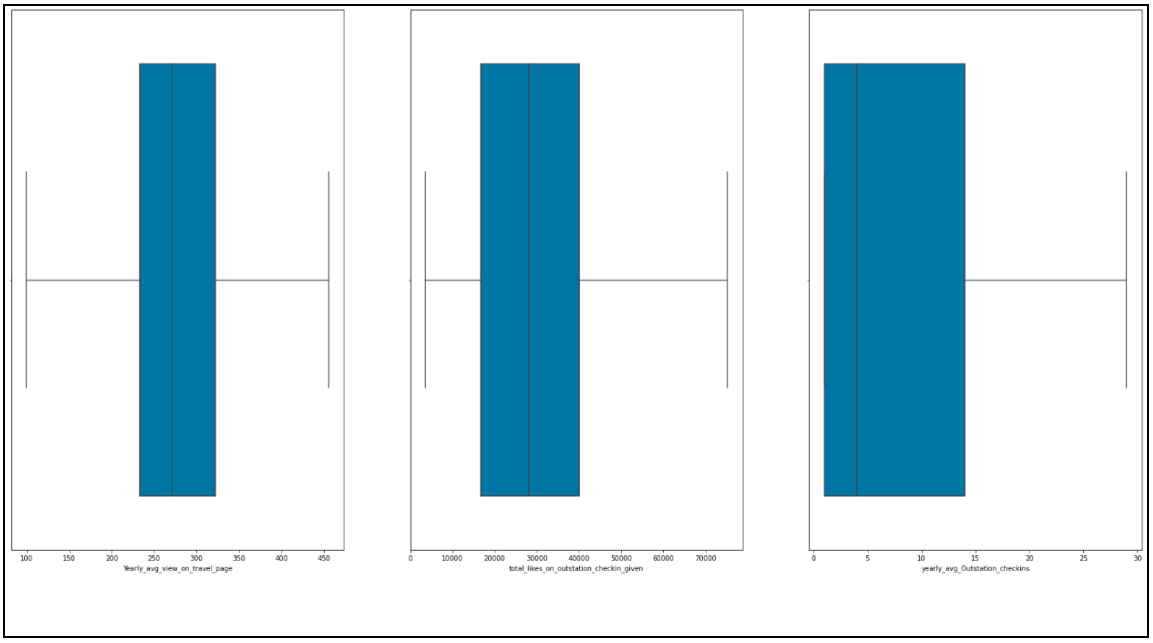
The outlier in the data set is treated using the IQR method.

Inter quartile range (IQR) method -

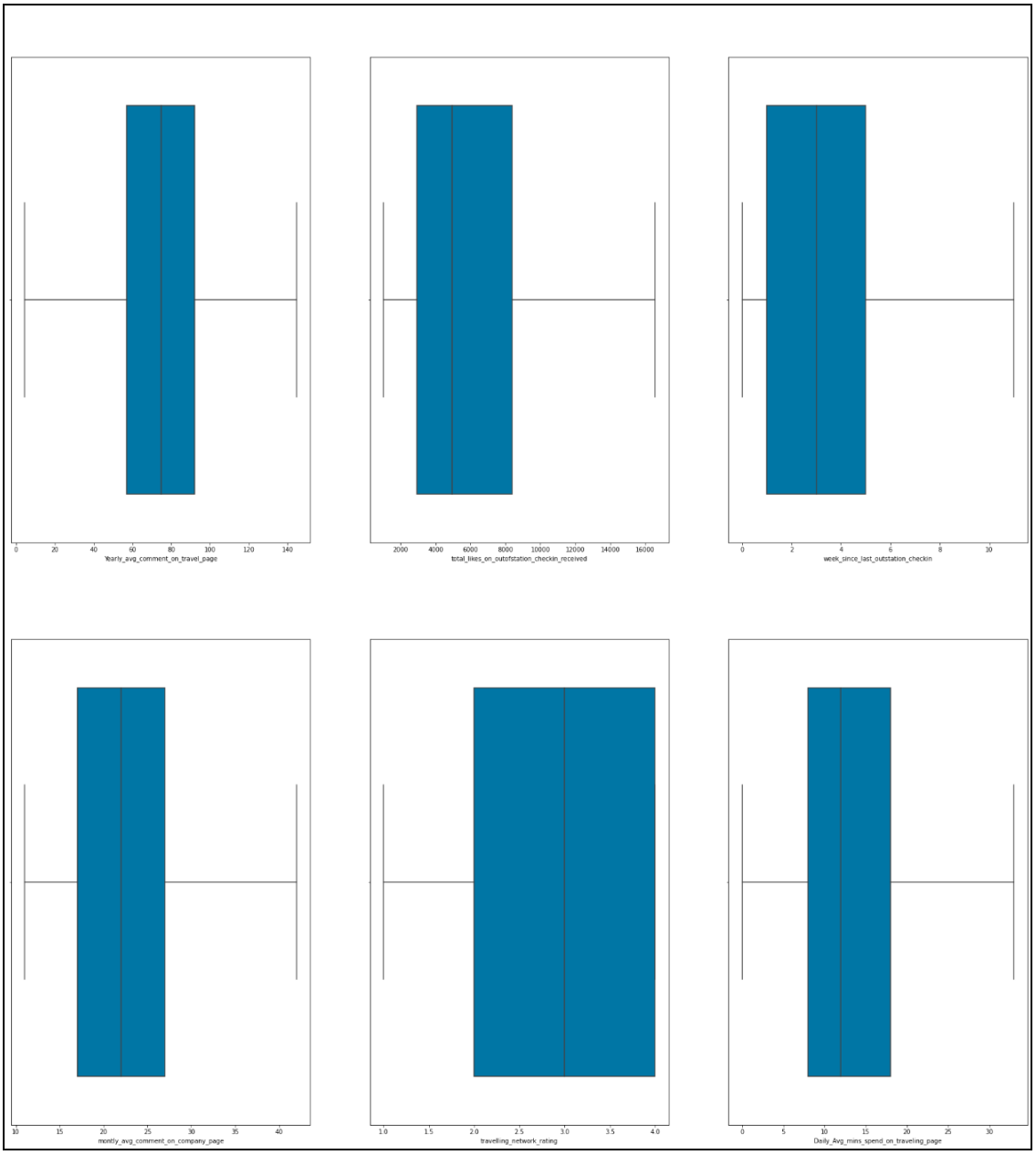
Each dataset can be divided into quartiles. The first quartile point indicates that 25% of the data points are below that value whereas the second quartile is considered as the median point of the dataset.

The inter quartile method finds the outliers on numerical datasets by following the procedure below,
Find the first quartile, Q1.
Find the third quartile, Q3.
Calculate the IQR. IQR= Q3-Q1.
Define the normal data range with lower limit as Q1–1.5*IQR and upper limit as Q3+1.5*IQR.
Any data point outside this range is considered an outlier and should be removed for further analysis.
The concept of quartiles and IQR can best be visualized from the boxplot. It has the minimum and maximum point defined as Q1–1.5*IQR and Q3+1.5*IQR respectively. Any point outside this range is outlier. [5]

After outlier treatment-



GRAPH 13



GRAPH 14

- Need for variable transformation (if any)

There are multiple transformations done in the data set.

The variable "***member_in_family***" had both numerical and string values. (i.e) it had 3 and three as the value in the rows. That has been changed to a numerical value.

The binary valued data has been transformed to 0 and 1 where 0 being No and 1 being Yes.

The variable "***Adult_flag***" has been transformed to a binary data, where 0 & 1 being 0 and 2 & 3 being 1. This change was done assuming the data is talking about whether the user is an adult or not. In that case, 2 & 3 will not have any meaning to it.
The variable "***preferred_location_type***" has many redundant choices that has been reduced.

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set "***total_likes_on_outstation_checkin_given***" is having values in thousands and "***yearly_avg_Outstation_checkins***" is just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale. StandardScaler normalizes the data using the formula (x-mean)/standard deviation. We will be doing this only for the numerical variables.

Before Scaling:

	count	mean	std	min	25%	50%	75%	max
Yearly_avg_view_on_travel_page	11760.0	280.385587	66.347859	99.5	233.00	271.0	322.00	455.50
total_likes_on_outstation_checkin_given	11760.0	28132.657058	13883.783884	3570.0	16697.25	28076.0	40115.25	75242.25
yearly_avg_Outstation_checkins	11760.0	8.176871	8.663686	1.0	1.00	4.0	14.00	29.00
Yearly_avg_comment_on_travel_page	11760.0	74.649320	21.526694	4.5	57.00	75.0	92.00	144.50
total_likes_on_outofstation_checkin_received	11760.0	6387.709439	4345.180379	1009.0	2940.75	4948.0	8393.25	16572.00
week_since_last_outstation_checkin	11760.0	3.203571	2.616365	0.0	1.00	3.0	5.00	11.00
montly_avg_comment_on_company_page	11760.0	22.854422	7.354454	11.0	17.00	22.0	27.00	42.00
travelling_network_rating	11760.0	2.712245	1.080887	1.0	2.00	3.0	4.00	4.00
Daily_Avg_mins_spend_on_traveling_page	11760.0	13.633673	7.980341	0.0	8.00	12.0	18.00	33.00

FIGURE 8

After Scaling:

	count	mean	std	min	25%	50%	75%	max
Yearly_avg_view_on_travel_page	11760.0	-7.799883e-16	1.000043	-2.726437	-0.714230	-0.141466	0.627242	2.639450
total_likes_on_outstation_checkin_given	11760.0	3.301403e-17	1.000043	-1.769237	-0.823687	-0.004081	0.863101	3.393282
yearly_avg_Outstation_checkins	11760.0	4.188539e-15	1.000043	-0.828421	-0.828421	-0.482133	0.672159	2.403598
Yearly_avg_comment_on_travel_page	11760.0	-5.660438e-16	1.000043	-3.258852	-0.819915	0.016291	0.806042	3.244978
total_likes_on_outofstation_checkin_received	11760.0	-7.714917e-17	1.000043	-1.237909	-0.793317	-0.331349	0.461575	2.343913
week_since_last_outstation_checkin	11760.0	-8.924267e-16	1.000043	-1.224488	-0.842262	-0.077810	0.686642	2.979997
montly_avg_comment_on_company_page	11760.0	-1.224425e-14	1.000043	-1.611938	-0.796071	-0.116182	0.563707	2.603374
travelling_network_rating	11760.0	-1.930617e-16	1.000043	-1.584178	-0.658973	0.266233	1.191438	1.191438
Daily_Avg_mins_spend_on_traveling_page	11760.0	-4.843366e-16	1.000043	-1.708480	-0.705974	-0.204721	0.547159	2.426857

FIGURE 9

- Variables removed or added and why (if any)

There were no new variables added as we already have sufficient variables to proceed with the analysis. The variable “UserID” was dropped from the data set as it doesn't contribute to our analysis.

4) Model building

- Clear on why a particular model(s) was chosen.

- Effort to improve model performance.

The models used in the project were CART, Random Forest, Logistic Regression, LDA and KNN. These models were chosen as they perform well for a classification type problem. Cart model tends to be useful for large data sets. Random forest is an ensemble technique and I have seen it performing well for classification type problems. Logistic regression is easier to implement, interpret, and very efficient to train. It is a simple, fast and portable algorithm as it beats some algorithms when its assumptions are met. KNN is very easy to implement. There are only two parameters required to implement KNN i.e. the value of K and the distance function.

Each model’s performance was improved by various tuning parameters.

The CART model was tuned by pruning the decision tree. Pruning a decision tree means to remove a subtree that is redundant and not a useful split and replace it with a leaf node. The models Random Forest used random search to find the best hyperparameters. Logistic Regression and LDA used the grid search to find the best hyperparameters. The optimum value for k is 3, which was obtained from the scree plot for the KNN model.

5) Model validation

- How was the model validated? Just accuracy, or anything else too?

The various performance metrics can be seen below at various stages-
Laptop:

Train data set	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	100 %	100 %	79 %	82 %	99 %
Recall	1.00	1.00	0.20	0.37	0.98
Precision	1.00	1.00	0.71	0.71	0.99
F1 score	1.00	1.00	0.32	0.49	0.99

TABLE 3

Test data set	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	91 %	97 %	75 %	80 %	95 %
Recall	0.81	0.88	0.25	0.41	0.91
Precision	0.87	1.00	0.65	0.81	0.93
F1 score	0.84	0.94	0.36	0.55	0.92

TABLE 4

Before Tuning	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	91 %	97 %	76 %	80 %	95 %
Recall	0.81	0.88	0.25	0.41	0.91
Precision	0.87	1.00	0.65	0.81	0.93
F1 score	0.84	0.94	0.36	0.55	0.92

TABLE 5

After Tuning	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	81 %	98 %	80 %	80 %	95 %
Recall	0.48	0.94	0.41	0.41	0.91
Precision	0.75	1.00	0.81	0.81	0.93
F1 score	0.59	0.97	0.55	0.55	0.92

TABLE 6

Before Smote	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	81 %	98 %	80 %	80 %	95 %
Recall	0.48	0.94	0.41	0.41	0.91
Precision	0.75	1.00	0.81	0.81	0.93
F1 score	0.59	0.97	0.55	0.55	0.92

TABLE 7

After Smote	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	81 %	98 %	67 %	74 %	98 %
Recall	0.48	0.97	0.77	0.76	1.00
Precision	0.75	0.97	0.46	0.53	0.92
F1 score	0.59	0.97	0.57	0.62	0.96

TABLE 8

Mobile:

Train data set	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	100 %	100 %	85 %	86 %	99 %
Recall	1.00	1.00	0.00	0.12	0.97
Precision	1.00	1.00	0.50	0.63	0.99
F1 score	1.00	1.00	0.00	0.20	0.98

TABLE 9

Test data set	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	97 %	97 %	85 %	85 %	98 %
Recall	0.89	0.82	0.00	0.14	0.91
Precision	0.89	1.00	0.00	0.63	0.95
F1 score	0.89	0.90	0.00	0.22	0.93

TABLE 10

Before Tuning	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	97 %	97 %	85 %	85 %	98 %
Recall	0.89	0.82	0.00	0.14	0.91
Precision	0.89	1.00	0.00	0.63	0.95
F1 score	0.89	0.90	0.00	0.22	0.93

TABLE 11

After Tuning	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	88 %	99 %	85 %	85 %	98 %
Recall	0.47	0.91	0.11	0.14	0.91
Precision	0.68	1.00	0.62	0.63	0.95
F1 score	0.56	0.95	0.19	0.23	0.93

TABLE 12

Before Smote	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	88 %	99 %	85 %	85 %	98 %
Recall	0.47	0.91	0.11	0.14	0.91
Precision	0.68	1.00	0.62	0.63	0.95
F1 score	0.56	0.95	0.19	0.23	0.93

TABLE 13

After Smote	CART	Random Forest	Logistic Regression	LDA	KNN
Accuracy	88 %	99 %	60 %	68 %	98 %
Recall	0.47	0.98	0.69	0.77	0.92
Precision	0.68	0.99	0.23	0.29	0.91
F1 score	0.56	0.95	0.35	0.42	0.94

TABLE 14

The model is evaluated based on the accuracy. Also, since the original problem statement is to predict the users who will take the product, Type 2 error should be reduced. Therefore, recall is an important metric to consider. In both mobile and laptop the model Random Forest has the best accuracy and recall.

6) Final interpretation / recommendation

- Detailed recommendations for the management/client based on the analysis done.

Analysis:

- Based on the VIF factor the below variables are important in the same order of importance-
    - Total_likes_on_outofstation_checkin_received
    - Yearly_avg_comment_on_travel_page
    - Total_likes_on_outstation_checkin_given
    - Yearly_avg_view_on_travel_page
    - Daily_Avg_mins_spend_on_traveling_page
    - Montly_avg_comment_on_company_page
- Using clustering techniques, the customers were divided into three segments. All the customers belonging to cluster 2 have taken the product. All the customers belonging to cluster 1 have not taken the product. The customers of cluster 0 have divided choices. Here 0 being no and 1 being yes.

Recommendations:

- The customers with 3 or 4 family members are more likely to go on trips. Therefore, we can focus on these customers to get a good response.
- The customers belonging to cluster 0 and cluster 2 should be prioritized as they have taken the product before.
- The customers opting for the location “beach” and “financial” contribute to the maximum count. We can come up with a separate tour package for these users.
- The customers following the company page and who belong to cluster 0 & 1 are highly likely to go on a trip. There can be targeted ads for these users.
- A tour plan can be suggested to the users based on the variables identified as important using VIF. The total_likes_on_outstation_checkin_given has come out to be the first in priority. This means that the user uses social media platforms highly and their interest is on tours. This can be used to come up with a tour package for those users. Similarly, total_likes_on_outofstation_checkin_received variable shows that the users are much interested in travel. The variables such as the views of a travel page and comments of the users in the same travel page and company page will give us more insights on what the users like and based on that a special tour package can be introduced. The variable week_since_last_outstation_checkin will also be very useful to assess when these users will go on their next trip.

- The total likes and comments given by the users indicate they are active users in the media platform. The highly active customers should be given high priority as there is a strong chance that they will take the product.
- We can come up with a travel combo for the users who are the frequent travelers.
- We could request a testimony from the loyal customers to promote the brand.
- Targeted advertisements to a specific destination can be sent to the users based on their interest.
- There can be discounts offered to the customers who did not take the product to attract them.

References

Websites-

[1] https://www.statisticshowto.com/univariate/

[2] https://www.spss-tutorials.com/skewness/

[3] https://en.wikipedia.org/wiki/Bivariate_analysis

[4] https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd5

[5] https://www.geeksforgeeks.org/interquartile-range-to-detect-outliers-in-data/

End of Project