Feel free to use this document as a Template.

Name: Ritik Gupta

Also please upload this document, together any script needed to verify any step of your solutions

I. The Business Problem

ABD contains data for all the customers in the dataset that were already pursued (advertised) but ended up not buying a vacation package.

<u>Business Problem: Should we retarget those customers?</u>

Q1: In light of your experience as a business woman/man, argue why this is a sensible business question.

When you are in charge of a business you want to serve your customers according to their needs. In doing so you need to understand your customers. It is like customer need is a lock and you have to find key for that lock. Hence, if you target the customers and most of the times they will churn then you need to know why. In order to prevent competitors to get their hold, you need to devise a strategy to increase customer base and make business flourish. Therefore we need to retarget those customers with some innovation about our product that fits their buying criteria after understanding why did not they buy first.

An experiment is run, where customers in the abandoned dataset are randomly placed in a treatment or in a control group (see column L in both files).

Those marked as “test” are retargeted (treated), the others marked as control are part of the control group.

Q2: compute the summary statistics (mean, median, q5, q95, standard deviation) of the Test_variable: a dummy with a value of 1 if tested 0 if control in the ABD database.

```
#Creating dummy variable
Test_Variable = as.numeric(as.factor(abd$Test_Control)) -1
```

General summary statistics:

```
> summary(Test_Variable)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5053  1.0000  1.0000

#median of test variable
> median_test_variable
[1] 1

#Standard Deviation
> sd_test_variable
[1] 0.5000012

#q5 and q95

> quantile(Test_Variable, c(.05, .95))
 5% 95%
  0   1
```

Q3: compute the same summary statistics for this Test_variable by blocking on States, wherever this information is available.

```
#state_dummy having state with test or control as 0 or 1
state_dummy = data.frame(abd$Address,Test_Variable)

#SUMMARY STATS FOR BLOCKING ON EACH STATE
#state_dummy having state with test or control as 0 or 1

state_dummy = data.frame(abd$Address,Test_Variable)
aggregate(state_dummy, list(abd.Address=Test_Variable), mean)
mean_each_state= aggregate(state_dummy[, 2], list(state_dummy$abd.Address), mean)
mean_each_state

> mean_each_state
   Group.1         x
1          0.4987041
2       AK 0.4754098
3       AL 0.4750000
4       AR 0.4523810
```

5	AZ	0.5510204
6	CA	0.5647059
7	CO	0.5194805
8	CT	0.5600000
9	DE	0.4250000
10	FL	0.5066667
11	GA	0.5875000
12	HI	0.5063291
13	IA	0.5342466
14	ID	0.5333333
15	IL	0.4404762
16	IN	0.4531250
17	KS	0.4743590
18	KY	0.5000000
19	LA	0.5200000
20	MA	0.5142857
21	MD	0.4871795
22	ME	0.4324324
23	MI	0.5657895
24	MN	0.6521739
25	MO	0.5810811
26	MS	0.4923077
27	MT	0.5072464
28	NC	0.5000000
29	ND	0.4193548
30	NE	0.4230769
31	NH	0.3888889
32	NJ	0.5909091
33	NM	0.5324675
34	NV	0.4591837
35	NY	0.5263158
36	OH	0.5617978
37	OK	0.4647887
38	OR	0.5000000
39	PA	0.6153846
40	RI	0.5942029
41	SC	0.6027397
42	SD	0.5205479
43	TN	0.4938272
44	TX	0.5714286
45	UT	0.4500000
46	VA	0.6049383
47	VT	0.5679012
48	WA	0.4788732
49	WI	0.4189189
50	WV	0.4845361

```
51      WY 0.5263158
```

```
#median each state
median_each_state= aggregate(state_dummy[, 2], list(state_dummy$abd.Address), median)
median_each_state
```

```
> median_each_state
   Group.1   x
1          0.0
2       AK 0.0
3       AL 0.0
4       AR 0.0
5       AZ 1.0
6       CA 1.0
7       CO 1.0
8       CT 1.0
9       DE 0.0
10      FL 1.0
11      GA 1.0
12      HI 1.0
13      IA 1.0
14      ID 1.0
15      IL 0.0
16      IN 0.0
17      KS 0.0
18      KY 0.5
19      LA 1.0
20      MA 1.0
21      MD 0.0
22      ME 0.0
23      MI 1.0
24      MN 1.0
25      MO 1.0
26      MS 0.0
27      MT 1.0
28      NC 0.5
29      ND 0.0
30      NE 0.0
31      NH 0.0
32      NJ 1.0
33      NM 1.0
34      NV 0.0
35      NY 1.0
36      OH 1.0
37      OK 0.0
38      OR 0.5
```

```
39      PA 1.0
40      RI 1.0
41      SC 1.0
42      SD 1.0
43      TN 0.0
44      TX 1.0
45      UT 0.0
46      VA 1.0
47      VT 1.0
48      WA 0.0
49      WI 0.0
50      WV 0.0
51      WY 1.0
```

```
#sd each state
sd_each_state= aggregate(state_dummy[, 2], list(state_dummy$abd.Address), sd)
sd_each_state
```

```
> sd_each_state
  Group.1         x
1            0.5000523
2       AK 0.5035394
3       AL 0.5025253
4       AR 0.5007166
5       AZ 0.4999474
6       CA 0.4987379
7       CO 0.5028966
8       CT 0.4997297
9       DE 0.4974619
10      FL 0.5033223
11      GA 0.4953901
12      HI 0.5031546
13      IA 0.5022779
14      ID 0.5030977
15      IL 0.4994259
16      IN 0.5017331
17      KS 0.5025741
18      KY 0.5038315
19      LA 0.5029642
20      MA 0.5034046
21      MD 0.5030708
22      ME 0.4987953
23      MI 0.4989463
24      MN 0.4797698
25      MO 0.4967499
26      MS 0.5038315
```

```
27      MT 0.5036102
28      NC 0.5036102
29      ND 0.4974818
30      NE 0.4972452
31      NH 0.4909191
32      NJ 0.4944837
33      NM 0.5022165
34      NV 0.5008934
35      NY 0.5026247
36      OH 0.4989775
37      OK 0.5023086
38      OR 0.5032363
39      PA 0.4891996
40      RI 0.4946431
41      SC 0.4927171
42      SD 0.5030349
43      TN 0.5030770
44      TX 0.4981168
45      UT 0.5016921
46      VA 0.4919099
47      VT 0.4984544
48      WA 0.5031090
49      WI 0.4967499
50      WV 0.5023570
51      WY 0.5026247
```

```
#q5q95 each state
q5q95_each_state = do.call("rbind", tapply(state_dummy$Test_Variable,
state_dummy$abd.Address, quantile, c(0.05, 0.95)))
q5q95_each_state
```

```
> q5q95_each_state
   5% 95%
    0   1
AK  0   1
AL  0   1
AR  0   1
AZ  0   1
CA  0   1
CO  0   1
CT  0   1
DE  0   1
FL  0   1
GA  0   1
HI  0   1
IA  0   1
```

ID 0 1
IL 0 1
IN 0 1
KS 0 1
KY 0 1
LA 0 1
MA 0 1
MD 0 1
ME 0 1
MI 0 1
MN 0 1
MO 0 1
MS 0 1
MT 0 1
NC 0 1
ND 0 1
NE 0 1
NH 0 1
NJ 0 1
NM 0 1
NV 0 1
NY 0 1
OH 0 1
OK 0 1
OR 0 1
PA 0 1
RI 0 1
SC 0 1
SD 0 1
TN 0 1
TX 0 1
UT 0 1
VA 0 1
VT 0 1
WA 0 1
WI 0 1
WV 0 1
WY 0 1

Q4: In light of the summaries in **Q3, Q4** does the experiment appear to be executed properly? Any imbalance in the assignments to treatment and control when switching to the State level? What would you have done differently?

I think Experiment appears to be executed properly as we can see that mean median mode and standard deviation are approximately the same so there is equal distribution in the states and no bias. Although there are some differences in median like some states have more treatment test group while others have more on the control side

Although homogeneous should be an experiment, I would want to try distributions based on some metrics like want to know what states people want to travel often and have some predictions on this and try to figure out some different plan for the strategy.

II. Data Matching

About three months later, the experiment/retargeting campaign is over.

Customers, presented in the ABD excel file, who bought a vacation packages during the time frame, are recorded in the RS excel file.

Q5: Argue that for proper causal inference based on experiments this is potentially problematic: “We do not observe some “outcomes” for some customers”. Argue that, however, matching appropriately the ABD with the RS dataset can back out this information.

After the retargeting is done, we don’t know if customers who were tested went for purchasing or not, In that case we need to determine if customer purchased or not. If we don’t have proper data or consistent data then our analysis is like an arrow aimed at a no mark and we are really interested in the question, if retargeting would be efficient or not.

We have a reservation data set which has the details of customers who actually purchased. By matching that reservation data with abandoned data we would be able to determine customers who were retargeted and who purchased. We can then match both the data files and merge that data into abandoned data set for analysis

Q6: After observing the data in the both files, argue that customers can be matched across some "data keys" (columns labels). Properly identify all these data keys (feel free to add a few clarifying examples if needed)

Both the data files are similar, In order to match them it is important to note unique columns that can be matched and merged upon.

Data keys that are deemed unique in both the data files are :

Incoming_Phone

Contact_Phone

Email.

So now we would run an algorithm that will match both the data sets based on these keys then merge them later for the use.

CODE USED TO MATCH 2 KEYS:

Specimen code to merge by their email_id:

```
#The merge() function of r is used to merge 2 datafiles based on some attributes.

mergeCols_email <- c("Email")

inner_email <- unique(merge(abd_nonull_email, res_nonull_email, by = mergeCols_email))
```

Q7: EXTREMELY CAREFULLY DESCRIBE YOUR DATA MATCHING PROCEDURE IN ORDER TO IDENTIFY: (1) Customers in the TREATMENT group who bought (2) Customers in the TREATMENT group who did not buy (3) Customers in the Control group who bought, and (4) Customers in the Control group who did not buy. Be as precise as possible.

#DESCRIPTION OF DATA MATCHING

> **Matched the data on three keys namely Contact_Phone, Incoming_Phone and Email.**

 # RETRIEVE THE ORIGINAL DATA FILES

```
library(readxl)

abd <- read.csv(file.choose())
```

res <- read.csv(file.choose())

#MATCHING THEM BY EMAIL

 # removing missing email data from both tables

 # I created 2 variables abd_nonull_email and res_nonull_email which contains data from original dataset excluding missing data

abd_nonull_email <- abd[-which(abd$Email == ""),]

res_nonull_email <- res[-which(res$Email == ""),]

#merging abd and res no null data by email

 # I merged abd_nonull with res_nonull on matching emails to find matching data items. Got output as inner_email

mergeColsEmail <- c("Email")

inner_email <- merge(abd_nonull_email, res_nonull_email, by = mergeColsEmail)

#Creating a dataframe that only contains Email and Session.y from inner_email

#from Inner_email I created a dataframe that only contains email and session.y information. So after matching email I could get session.y information as well on the matched abandoned data set.

e= data.frame(inner_email$Email,inner_email$Session.y)

#Changing column names to Email and Session.y

#changed column names so that i could match e with abd_no_null with Email.

colnames(e)[colnames(e)=="inner_email.Email"] <- "Email"

colnames(e)[colnames(e)=="inner_email.Session.y"] <- "Session.y"

#Merging the e dataset(EMAIL And Session.y) and abandoned no null data

 # final_df_1 contains data on email matching

final_df_1 = merge(e, abd_nonull_email, by = mergeColsEmail)

#Eliminating duplicate values.

final_df_1=distinct(final_df_1)

Found 90 Observations after removing duplicates

MATCH BY INCOMING_PHONE

 # removing missing email data from both tables

 # I created 2 variables abd_nonull_incoming and res_nonull_incoming which contains data from original dataset excluding missing data

```
abd_nonull_incoming <- abd[-which(abd$Incoming_Phone == ""), ]

res_nonull_incoming <- res[-which(res$Incoming_Phone == ""), ]
```

#merging abd and res no null data by incoming phone

 # I merged abd_nonull with res_nonull on matching Incoming_Phone to find matching data items. Got output as inner_phone

```
mergeColsIncoming <- c("Incoming_Phone")

inner_phone <- merge(abd_nonull_incoming, res_nonull_incoming, by = mergeColsIncoming)
```

#Creating a dataframe that only contains Incoming_phone and Session.y from inner_phone

#from Inner_phone I created a dataframe that only contains email and session.y information. So after matching Incoming_Phone I could get session.y information as well on the matched abandoned data set.

```
p= data.frame(inner_phone$Incoming_Phone,inner_phone$Session.y)
```

#Changing column names to Incoming_Phone and Session.y

#changed column names so that i could match p with abd_no_null with Incoming_Phone.

```
colnames(p)[colnames(p)=="inner_phone.Incoming_Phone"] <- "Incoming_Phone"
```

colnames(p)[colnames(p)=="inner_phone.Session.y"] <- "Session.y"

#Merging the p dataset(Incoming_Phone And Session.y) and abandoned no null data

 # final_df_2 contains data on Incoming_Phone matching.i.e those customers mathced on incoming_phone when retargeted

final_df_2 = unique(merge(p, abd_nonull_incoming, by = mergeColsIncoming))

<u>#Found 368 observations by matching on incoming_phone</u>

<u>**# MATCH BY CONTACT_PHONE**</u>

#Followed the same Procedure as for matching email and incoming_phone

abd_nonull_contact <- abd[-which(abd$Contact_Phone == ""),]

res_nonull_contact <- res[-which(res$Contact_Phone == ""),]

mergeColsContact <- c("Contact_Phone")

inner_contact <- merge(abd_nonull_contact, res_nonull_contact, by = mergeColsContact)

c= data.frame(inner_contact$Contact_Phone,inner_contact$Session.y)

colnames(c)[colnames(c)=="inner_contact.Contact_Phone"] <- "Contact_Phone"

colnames(c)[colnames(c)=="inner_contact.Session.y"] <- "Session.y"

final_df_3 = merge(c, abd_nonull_contact, by = mergeColsContact)

final_df_3=distinct(final_df_3)

<u>#Found 232 observation matched on contact_phone</u>

Generating data to analyze

#the final variable contains all the data from email, contact_phone and incoming phone matching. It contains information of all matching reservation and abandoned data of the users who purchased after retargeting

final = rbind(final_df_1,final_df_2,final_df_3)

#we get 690 observations and a lot of matching attributes are duplicated like email, incoming_phone and contact_phone and we need to seperate key duplicated data.

#Used duplicated function to remove duplicates that contain same email,phone and contact_phone as this are our problematic cases and we would then have trouble finding predictions so it is better to remove them from analysis.

duplicates = duplicated(final[,c("Email","Incoming_Phone", "Contact_Phone")])

final =final[!duplicates,]

#408 observations are recorded.

 # finally merging the data with abd to find list of customers who purchased after retargeting and who did not.

data= merge(abd,final,all.x = TRUE)

#Creating Outcome column depicting if person has purchased or not.

#Binary Outcome variable is created. Is.na(data$Session.y) denotes if there is null, Outcome would be TRUE or if there is no null. If there is some value, the Outcome would be FALSE.

Outcome = is.na(data$Session.y)

#if Outcome is true we label is as No Buy. If Outcome is False we label it as Buy

Outcome[Outcome == TRUE] <- "No Buy"

Outcome[Outcome == FALSE] <- "Buy"

#Binding data with outcome we created

data = cbind(data, Outcome)

Q9: Complete the following cross-tabulation:

data_test_buy = data[which(data$Test_Control=='test'& data$Outcome=='Buy'),]

dim(data_test_buy)

> dim(data_test_buy)

[1] 328 14

>data_test_control=data[which(data$Test_Control=='control'&data$Outcome=='Buy'),]

> dim(data_test_control)

[1] 80 14

data_control_nobuy = data[which(data$Test_Control=='control'& data$Outcome=='No Buy'),]

dim(data_control_nobuy)

> dim(data_control_nobuy)

[1] 4096   14

Group \ Outcome	Buy	No Buy
Treatment	328	3938
Control	80	4096

Q10: Repeat Q9 for 5 randomly picked states. Report 5 different tables by specifying the states you "randomly picked".

#FLORIDA

```
> data_FL = data[which(data$Address=='FL'),]

> data_test_buy_FL = data_FL[which(data_FL$Test_Control=='test'& data_FL$Outcome=='Buy'),]

> dim(data_test_buy_FL)

[1]  3 14

>data_control_buy_FL=data[which(data_FL$Test_Control=='control'&
data_FL$Outcome=='Buy'),]

> dim(data_control_buy_FL)

[1]  0 14

>data_test_nobuy_FL=data_FL[which(data_FL$Test_Control=='test'&
data_FL$Outcome=='No Buy'),]

> dim(data_test_nobuy_FL)

[1] 35 14

>data_control_nobuy_FL=data_FL[which(data_FL$Test_Control=='control'&
data_FL$Outcome=='No Buy'),]

> dim(data_control_nobuy_FL)

[1] 37 14
```

Group \ Outcome	**Buy**	**No Buy**
Treatment	**3**	**35**
Control	**0**	**37**

#NEW JERSEY

```
> data_NJ = data[which(data$Address=='NJ'),]
> data_test_buy_NJ = data_NJ[which(data_NJ$Test_Control=='test'&
data_NJ$Outcome=='Buy'),]
> dim(data_test_buy_NJ)
```

```
[1]  5 14
> data_control_buy_NJ= data_NJ[which(data_NJ$Test_Control=='control'&
data_NJ$Outcome=='Buy'),]
> dim(data_control_buy_FL)
[1]  0 14
> data_test_nobuy_NJ = data_NJ[which(data_NJ$Test_Control=='test'&
data_NJ$Outcome=='No Buy'),]
> dim(data_test_nobuy_NJ)
[1] 47 14
> data_control_nobuy_NJ = data_NJ[which(data_NJ$Test_Control=='control'&
data_NJ$Outcome=='No Buy'),]
> dim(data_control_nobuy_NJ)
[1] 33 14
```

Group \ Outcome	**Buy**	**No Buy**
Treatment	**5**	**47**
Control	**0**	**33**

New York

```
> data_NY = data[which(data$Address=='NY'),]

>data_test_buy_NY=data_NY[which(data_NY$Test_Control=='test'&
data_NY$Outcome=='Buy'),]

> dim(data_test_buy_NY)

[1]  3 14

>data_control_buy_NY=data_NY[which(data_NY$Test_Control=='control'&
data_NY$Outcome=='Buy'),]

> dim(data_control_buy_NY)

[1]  1 14
```

>data_test_nobuy_NY=data_NY[which(data_NY$Test_Control=='test'& data_NY$Outcome=='No Buy'),]

> dim(data_test_nobuy_NY)

[1] 37 14

>data_control_nobuy_NY=data_NY[which(data_NY$Test_Control=='control'& data_NY$Outcome=='No Buy'),]

> dim(data_control_nobuy_NY)

[1] 35 14

Group \ Outcome	Buy	No Buy
Treatment	**3**	**37**
Control	**1**	**35**

#CALIFORNIA

> data_CA = data[which(data$Address=='CA'),]

>data_test_buy_CA=data_CA[which(data_CA$Test_Control=='test'& data_CA$Outcome=='Buy'),]

> dim(data_test_buy_CA)

[1] 6 14

>data_test_nobuy_CA=data_CA[which(data_CA$Test_Control=='test'& data_CA$Outcome=='No Buy'),]

> dim(data_test_nobuy_CA)

[1] 42 14

>data_control_buy_CA=data_CA[which(data_CA$Test_Control=='control'& data_CA$Outcome=='Buy'),]

```
> dim(data_control_buy_CA)

[1]  0 14

>data_control_nobuy_CA=data_CA[which(data_CA$Test_Control=='control'&
data_CA$Outcome=='No Buy'),]

> dim(data_control_nobuy_CA)

[1] 37 14
```

Group \ Outcome	Buy	No Buy
Treatment	**5**	**42**
Control	**0**	**37**

#TEXAS

```
> data_TX = data[which(data$Address=='TX'),]
> data_test_buy_TX = data_TX[which(data_TX$Test_Control=='test'&
data_TX$Outcome=='Buy'),]
> dim(data_test_buy_TX)
[1]  3 14
> data_test_nobuy_TX = data_TX[which(data_TX$Test_Control=='test'&
data_TX$Outcome=='No Buy'),]
> dim(data_test_nobuy_TX)
[1] 41 14
> data_control_buy_TX= data_TX[which(data_TX$Test_Control=='control'&
data_TX$Outcome=='Buy'),]
> dim(data_control_buy_TX)
[1]  0 14
> data_control_nobuy_TX = data_TX[which(data_TX$Test_Control=='control'&
data_TX$Outcome=='No Buy'),]
> dim(data_control_nobuy_TX)
[1] 33 14
```

Group \ Outcome	Buy	No Buy

Treatment	**3**	**41**
Control	**0**	**33**

III. Data Cleaning:

You have now identified all the customers who are relevant for the analysis and their outcome and you also know if they are in a treated or in a control group.

Produce an Excel File (or CSV) with the following columns

Customer ID | Test Variable | Outcome | Days_in_Between | State |

Where Test Variable indicates, again, the treatment or the control group, Outcome is a binary variable indicating whether a vacation package was ultimately bought, Days in between is the (largest) difference between the dates in the ABD and RS dataset (Columns B). If no purchase, set "Days_in_between" as "200".

To be perfectly clear, you should have as number of rows all the customers you were able to match across the two data sets. Be sure to attach this excel file to the submission for proper verification.

Produce a script (R or SQL) detailing the entire data cleaning procedure, from loading and attaching the original data file to saving the pos-processed one, for reproducibility purposes. Bonus points may be applied.

#In order to Produce Date difference, we would use POSIXlt Function in R which converts strings into data and time object and then we can calculate difference directly for Days_In_Between

```
data$Session = as.POSIXlt(as.character(data$Session), tz="GMT",format="%Y.%m.%d
%H:%M:%S")
data$Session.y = as.POSIXlt(as.character(data$Session.y), tz="GMT",format="%Y.%m.%d
%H:%M:%S")
data$Difference <- round(data$Session.y - data$Session, digits = 0)
```

#Created new data frame reg_data which contains all the important data for analysis also includes if Days_In_Between is null then value is 200

```
reg_data = data.frame(Customer_ID = data$Caller_ID, Test_Variable = data$Test_Control,
Outcome = data$Outcome, Days_in_Between = data$Difference, State = data$Address,
Email = data$Email)
reg_data$Days_in_Between[is.na(reg_data$Days_in_Between)] = 200
reg_data$Days_in_Between=as.numeric(reg_data$Days_in_Between)

#Produced to excel file
library("xlsx")

write.xlsx(x = reg_data, file = "midterm_data_analysis.xlsx",
        sheetName = "Customer_Data", row.names = FALSE)
```

IV. Statistical Analysis

We are finally in a condition to try to answer the relevant business question.

Q11: Run a Linear regression model for

Outcome = alpha + beta * Test_Variable + error

And Report the output.

#The code below changes Outcome of Buy to 0 and No Buy to 1

```
reg_data$Outcome = as.numeric(as.factor(reg_data$Outcome)) -1
```

Regression Output

```
> reg.out=lm(Outcome~Test_Variable,data=reg_data)
> summary(reg.out)

Call:
lm(formula = Outcome ~ Test_Variable, data = reg_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98012  0.01988  0.01988  0.07689  0.07689

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.980125   0.003301  296.89   <2e-16 ***
Test_Variabletest -0.057012   0.004644  -12.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2133 on 8440 degrees of freedom
Multiple R-squared:  0.01754,	Adjusted R-squared:  0.01743
F-statistic: 150.7 on 1 and 8440 DF,  p-value: < 2.2e-16
```

Q12: Argue whether this is a properly specified linear regression model, if so, if we can draw any causal statement about the effectiveness of the retargeting campaign. Is this statistically significant?

Outcome = 1.98 - 0.05Test_Variabletest

By looking at the p values and statistical significance, the model is quite significant(<0.05) but our slope only explains 1% of the variation in Test_variable(R square .01), Questioning its effectiveness.

The model implies for every test we can expect Outcome to be 0.98-0.05(0 Being Buy, 1 being not). so we can not really determine if a person who is tested will actually buy it or not.

However, Model is suggesting for every test probability of buying will increase because it has -.05 as its coefficient on test and 1 being No buy and 0 means buy. So it is coming closer to 0 to some extent.

Overall it is not a good model to determine.

Q13: Now add to the regression model the dummies for State and Emails. Also consider including interactions with the treatment, namely between email and retargeting. Report the outcome and comment on the results. (You can compare with Q11). You should see something interesting appearing, if possible, provide a managerial interpretation)

#Creating State and Email dummy

```
has_state = as.numeric(as.factor(reg_data$State)) -1

has_state[has_state!=0]<-1



has_email = as.numeric(as.factor(reg_data$Email)) -1

has_email[has_email!=0]<-1

reg_data$has_state = has_state

reg_data$has_email = has_email
```

#REGRESSION MODEL FOR DUMMIES OF STATE AND EMAILS

```
reg.out3=lm(Outcome~Test_Variable +has_email+has_state,data=reg_data)

summary(reg.out3)
```

Call:

lm(formula = Outcome ~ Test_Variable + has_email + has_state,

data = reg_data)

Residuals:

Min	1Q	Median	3Q	Max
-0.99150	0.00850	0.06090	0.06455	0.11695

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.991502	0.003903	254.056	< 2e-16	***
Test_Variabletest	-0.056049	0.004634	-12.094	< 2e-16	***
has_email	-0.035802	0.007262	-4.930	8.37e-07	***
has_state	-0.016600	0.004774	-3.477	0.00051	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2128 on 8438 degrees of freedom

Multiple R-squared: 0.02291, Adjusted R-squared: 0.02256

F-statistic: 65.95 on 3 and 8438 DF, p-value: < 2.2e-16

According to this model :

Outcome = 0.98 - 0.03*has_email -0.01*has_state

As we Know out Outcome for Buy is denoted by 0 in the model and Outcome for Not buy is 1. If wee have the customer buy it our model should predict 0.

As seen by this model if campaign is done by email and state is known then we can expect our outcome to drop by -.01- .03 (i.e Outcome = 0.98(Intercept) -0.03-0.01)

The fit is very bad which is 2%. However Coefficients and slope are statistically significant.00

By looking at the coefficients on has_email and has_state, We can have a general idea that when we include state and email we can have higher chance of customers buying the deal.

 # Regression model for Interaction variable:

reg.out4=lm(Outcome~Test_Variable +has_email*Test_Variable,data=reg_data)
summary(reg.out4)

```
Call:
lm(formula = Outcome ~ Test_Variable + has_email * Test_Variable,
    data = reg_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98113  0.01887  0.02784  0.06778  0.13677

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.981127   0.003493 280.913  < 2e-16 ***
Test_Variabletest       -0.048910   0.004941  -9.898  < 2e-16 ***
has_email               -0.008964   0.010444  -0.858    0.391
Test_Variabletest:has_email -0.060020   0.014201  -4.227  2.4e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2127 on 8438 degrees of freedom
Multiple R-squared:  0.02358,	Adjusted R-squared:  0.02323
F-statistic: 67.91 on 3 and 8438 DF,  p-value: < 2.2e-16
```

Now, when we include interaction variable between retargeting * email we can observe that if advertisement is done on email of those who were going to be retargeted we can observe drop going by 0.06 in favour of buy..
I.e Outcome = 0.98 -0.04*Test_Variabletest-0.008*has_email
-0.06*Test_Variabletest*has_email
(Outcome for BUY is 0)

As we can see Interaction term has statistical significant p value (<0.05) we can conclude that if advertisement is done by email on retargeted customers we can have a general idea that it will turn to favour of buy. Although our fit is bad.

Managerial Interpretation:
If we compare this model of interaction to previous model we can observe that when our model was predicting Outcome based on just if it was retargeted or not, Our results were not that great although significant we could only predict that if test was done then decrease in Outcome was 0.98-0.02(Outcome 0 is buy).
Now when we added our interaction term i.e has_email and If they were retargeted we can have a higher decrease in outcome and model was performing better although fit was bad but we had that general idea of model performing better

Reason Might be if customers were retargeted using emails then they had time to think without being disturbed on phone that might have created the possibility for them to buy the vacation package.
Other reason might be the emails were descriptive and graphically interactive giving them better expectations and to go through the details taking their time.
While on Phone details could be vague or not properly illustrated.

RQ2: You want now to investigate whether the response time (time to make a purchase after the first contact) is influenced by the retargeting campaign. Make sure you describe carefully how you compute response times (there is no clear answer, so make any sensible assumption).

Q14: Set up an appropriate linear regression model to address the RQ2 above. Make sure to select the appropriate subset of customers. Report output analysis with your interpretation. Can the coefficients be interpreted as causal in this case? Is there evidence of any interactions effect?

#Selecting customers with purchase after a span of time

response_time = reg_data[-which(reg_data$Days_in_Between==200),]

#Regression model between Days_in_between and Test_Variable

reg_response=lm(Days_in_Between~Test_Variable,data=response_time)

```
summary(reg_response)
Call:
lm(formula = Days_in_Between ~ Test_Variable, data = response_time)

Residuals:
   Min     1Q Median     3Q    Max
-44.86 -10.87  -0.88  11.12  47.14

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        44.880      1.752   25.62   <2e-16 ***
Test_Variabletest    4.980      1.961    2.54   0.0115 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.96 on 409 degrees of freedom
Multiple R-squared:  0.01553,           Adjusted R-squared:  0.01312
F-statistic: 6.451 on 1 and 409 DF,  p-value: 0.01146
```

This Model estimates after how many days the customer in the treatment group purchases the package.

According to the model :

Days_In_Between = 44.880 +4.980*Test

According to our model, we can expect the purchase of a retargeted customer in the test group to be after 50 days

Our model only explains 1% of the variation so we are not certain.

```
# Regression model predicting days_in_between(response time) using dummy email
variable and retargeting variable

reg_response1=lm(Days_in_Between ~ has_email+Test_Variable,data=response_time)
summary(reg_response1)

Call:
lm(formula = Days_in_Between ~ has_email + Test_Variable, data = response_time)

Residuals:
    Min      1Q  Median      3Q     Max
-45.631 -11.631  -1.394  10.369  46.369

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        45.394      1.773  25.607   <2e-16 ***
```

```
has_email          -3.286      1.904  -1.725   0.0852 .
Test_Variabletest    5.237      1.962   2.670   0.0079 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.92 on 408 degrees of freedom
Multiple R-squared:  0.02266,           Adjusted R-squared:  0.01787
F-statistic:  4.73 on 2 and 408 DF,  p-value: 0.009318
```

This model predicts Days in between when customer who was retargeted used email.
Days_In_Between = 45.94 -3.286* has_email +5.237*Test_Variable

We can expect to predict Days_In_Between: if he has and was contacted by email we can expect to decrease Days_in_between by 3.286 and if he was tested, days_in_between to be increased by 5.237

Our model explains 1% of the variation and since it is statistically significant we cannot have a causal inference based on this model.

```
# Model Including Interaction term

#Including Interaction terms for has_email and Test_Variable

reg_response2=lm(Days_in_Between ~ has_email*Test_Variable,data=response_time)
summary(reg_response2)

Call:
lm(formula = Days_in_Between ~ has_email * Test_Variable, data = response_time)

Residuals:
    Min      1Q  Median      3Q     Max
-45.606 -11.606  -1.486  10.394  46.394

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              45.4857     1.9051  23.876   <2e-16 ***
has_email                 -3.8703     4.8137  -0.804   0.4219
Test_Variabletest          5.1199     2.1544   2.376   0.0179 *
has_email:Test_Variabletest   0.6933     5.2425   0.132   0.8949
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.94 on 407 degrees of freedom
Multiple R-squared:  0.0227,  Adjusted R-squared:  0.0155
F-statistic: 3.151 on 3 and 407 DF,  p-value: 0.0249
```

The Model :
Days_In_Between = 45.4857
-3.8703*has_email+5.11*Test_Variabletest+0.69*has_email*Test_Variabletest.

In this Model we used Interaction term of has_email*Test_Variable for predicting Days_In_between. As we can see in our model Interaction Term is completely insignificant(p-value =.8949) which is much greater than the significance level(.05). Additionally we have r value of .0227 that explains only 2% of the variation.

Hence, We can conclude that there is no evidence of interaction term.

```
#Interaction term including has_state and Test_Variable
reg_response2=lm(Days_in_Between ~ has_state*Test_Variable,data=response_time)
summary(reg_response2)
```

```
> summary(reg_response2)

Call:
lm(formula = Days_in_Between ~ has_state * Test_Variable, data = response_time)

Residuals:
    Min      1Q  Median      3Q     Max
-44.147 -11.147  -1.411  10.589  48.188

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 47.714      2.696  17.698   <2e-16 ***
has_state                   -4.902      3.545  -1.383    0.168
Test_Variabletest            1.433      3.008   0.476    0.634
has_state:Test_Variabletest  6.166      3.965   1.555    0.121
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.95 on 407 degrees of freedom
Multiple R-squared:  0.02134,	Adjusted R-squared:  0.01413
F-statistic: 2.959 on 3 and 407 DF,  p-value: 0.0322
```

By looking at the figures this is also way insignificant by looking at p values
Hence, We cannot find any interaction effect on days in between

Q15: Lesson Learned. What would you have done differently in designing the experiment? Any other directions you could have taken with better data? Are

there any prescriptive managerial implications out of this study? Please answer briefly

I would have tried Classification using Logistic Regression or any other classification methods like trees and use the best one.
Additionally, I would be Interested in knowing the actual pattern of how customers are buying and would like to mine that and create insights.
As we can see retargeting was not that successful Instead of targeting the customers back, As a manager I would just choose a little subset by identifying the interest ratio. Would try to do some text mining on what customers have to say and build a predictive model using ensembling.
By looking at our regression models, we can use that interaction effect of email and test on Outcome and Focus on the quality of packages being offered.