**TEAM LEADA PROJECT**

Important Note: It is assumed that each student will sign up for the TeamLeada modules at https://www.teamleada.com/courses/intro-to-ab-testing-in-r
**Not signing up will lead to an automatic score of zero in the project.**

This will give you access to two files, place in module five "A/B Testing Analytics: MightyHive Project"


Figure 1: The fifth module of the Leada Project

In the module at https://www.teamleada.com/projects/ab-testing-analytics-mightyhive-project/data-background/data-background, you will be prompted to download two files, the **abandoned data set (ABD hereafter)** and the **reservation dataset (RS hereafter)**

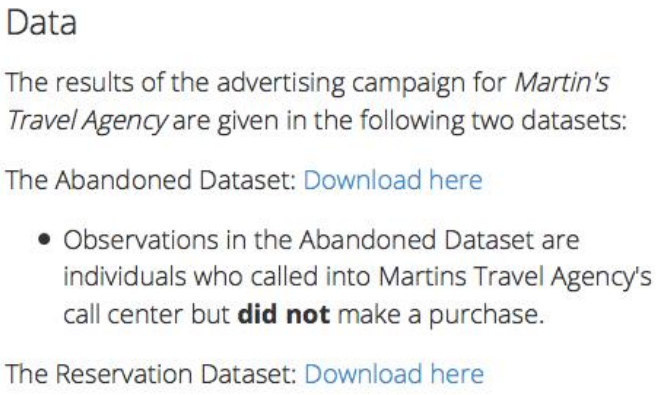


Figure 2: Where to download the two datasets

**EXAM**

**Feel free to use this document as a Template.**

**Name: Snehanshu Shankar**
**Section: Morning**
**Signature (if possible)**

**Did you work with someone else while cleaning or analyzing the data? Please disclose your teammates. Be forthcoming to avoid potential bad consequences.**

## 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.

Business Problem: Should we retarget those customers?

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

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.

Answer:
In my opinion, it is a sensible business question as there are many instances when people do not make a purchase in the first attempt, but after a few follow-ups with them they might buy. There are a plenty of reasons for this to happen and are as follows.
a) Customer might be interested in buying later.
b) Customer might be looking to check out other alternatives, and then compare and make a decision.
c) One might not be carrying their credits card at that moment of time.
d) One might be looking for their spouse or other family member consent.
Therefore, retargeting customers is one of the indispensible ways to bring more customers on the table.

**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.

Answer:
```
 summary(Abandoned_Data_Seed$D_Test_Variable)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5053  1.0000  1.0000
```

```
q95=quantile(Abandoned_Data_Seed$D_Test_Variable1,.95)
> q95
95%
  1
> q5=quantile(Abandoned_Data_Seed$D_Test_Variable1,.05)
> q5
5%
 0
> AbnData_SD <- sd(Abandoned_Data_Seed$D_Test_Variable1)
> AbnData_SD
[1] 0.5000012
```

**Q3:**compute the same summary statistics for this Test_variable by blocking on States (meaning considering only the entries with known "State"), wherever this information is available.

Answer:

```
Abandoned_Data_Seed["D_State"]<-NA
Abandoned_Data_Seed$D_State[Abandoned_Data_Seed$Address!=""]<-1
Abandoned_Data_Seed$D_State[is.na(Abandoned_Data_Seed$Address)]<-0
TestControl_State                                                                    <-
      Abandoned_Data_Seed$D_Test_Variable[which(Abandoned_Data_Seed$D_State==1)]
> summary(as.numeric(TestControl_State))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5134  1.0000  1.0000

> sd(as.numeric(TestControl_State))
[1] 0.4998865
```

**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-only level?

Answer:

It is evident from the summary values that the mean and standard deviation calculated for both the cases is nearly same. Hence, there is not any visible imbalance in the assignments to treatment and control when switching to the state-only level.

## 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.

Answer:

We have many unfilled values in columns like First Name, Last Name in the abandoned dataset. So in the first place it might seem difficult to inference about the totality of the customers who made a reservation.

However, after matching the two dataset on the basis of some of the key attributes like Incoming Phone, Contact Phone & Email would be a better way to fetch out customers present in Reserved Dataset from Abandoned dataset .

**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)

Answer:
After observing the two datasets, we can say that the customers can be matched on the basis of key columns like Incoming Phone, Contact Phone & 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.**

Answer:
Below are the mentioned steps:

1) Firstly, In both the data set, I have assigned NA for the blank fields.
2) Secondly, I have removed the dulplicate values from the dataset. In the code below, I have used dulplicated method in R to get TRUE in the result vector(logical) in the repeating positions and wanted to remove that row from the dataset. For all the three columns(Email, Incoming Phone & Contact Phone), I have filtered out non-duplicates logical vector using "&" operator among the three logical vector dataset.

```
#Code to remove duplicate from Purchased customers dataset.
Aban_Dupl_EmailRows <-duplicated(Abandoned_Data_Seed$Email,incomparables = NA)
Aban_Dupl_IncPhnRows <-duplicated(Abandoned_Data_Seed$Incoming_Phone,incomparables = NA)
Aban_Dupl_ContPhnRows <-duplicated(Abandoned_Data_Seed$Contact_Phone,incomparables = NA)
Aban_NonDuplicateRows <- !Aban_Dupl_EmailRows & !Aban_Dupl_IncPhnRows & !Aban_Dupl_ContPhnRows
Abandoned_Data_Seed <- Abandoned_Data_Seed[Aban_NonDuplicateRows,]
> nrow(Abandoned_Data_Seed)
[1] 8297
```

3) I have assigned 0 to NA fields in Abandoned dataset and 1 in Reserved dataset for Email, Incoming Phone & Contact Phone columns in order to avoid inconsistent comparison results with NA or blank.

```
Reservation_Data_Seed$Email[is.na(Reservation_Data_Seed$Email)]<-1
Abandoned_Data_Seed$Email[is.na(Abandoned_Data_Seed$Email)]<-0
Reservation_Data_Seed$Contact_Phone[is.na(Reservation_Data_Seed$Contact_Phone)]<-1
Abandoned_Data_Seed$Contact_Phone[is.na(Abandoned_Data_Seed$Contact_Phone)]<-0
Reservation_Data_Seed$Incoming_Phone[is.na(Reservation_Data_Seed$Incoming_Phone)]<-1
Abandoned_Data_Seed$Incoming_Phone[is.na(Abandoned_Data_Seed$Incoming_Phone)]<-0
```

4) I have taken out the logical vector for matched data in the two dataset on the basis of Email, Incoming Phone & Contact Phone using R "%in%" method.

```
matchesInPhone = Abandoned_Data_Seed$Incoming_Phone %in% Reservation_Data_Seed$Incoming_Phone
matchesEmail = Abandoned_Data_Seed$Email %in% Reservation_Data_Seed$Email
matchesContactPh = Abandoned_Data_Seed$Contact_Phone %in% Reservation_Data_Seed$Contact_Phone
```

```
5) In step 3, I have results in the 3 logical vector which matches on either of the
   columns. Hence, in this step I have taken "OR" operator among all the three vectors
   to derive a final logical vector set which includes common customers between two
   tables.
   RowsInAbdFromResTb = matchesInPhone | matchesContactPh | matchesEmail
   #Matched customers in the two dataset.
   ResDataFromAbandoned = Abandoned_Data_Seed[RowsInAbdFromResTb,]
   >   nrow(ResDataFromAbandoned)
   [1] 383

6) I have added a new column as an "Outcome" in the Abandoned dataset calculated in the
   2nd step to map the results with the matched customer dataset calculated in previous
   step. In outcome, it is 0 for not reserved & 1 for purchased.
   Abandoned_Data_Seed["Outcome"]<-NA
   OutcomeVector = Abandoned_Data_Seed$Caller_ID %in% ResDataFromAbandoned$Caller_ID
   OutcomeVector<-as.integer(OutcomeVector)
   Abandoned_Data_Seed$Outcome<-OutcomeVector
7) From above dataset, Customer in Treatment Group who do not made a purchase.
   Cust_in_Test_NotPurchased<-subset(Abandoned_Data_Seed,
   Abandoned_Data_Seed$Test_Control=="test" & Abandoned_Data_Seed$Outcome==0 )
   Cust_in_Test_NotPurchased
                  > nrow(Cust_in_Test_NotPurchased)
   [1] 3863

8) Customer in Treatment Group who made a purchase.
   Cust_in_Test_Purchased<-subset(Abandoned_Data_Seed,
   Abandoned_Data_Seed$Test_Control=="test" & Abandoned_Data_Seed$Outcome==1 )
   > nrow(Cust_in_Test_Purchased)
   [1] 301

9) Customer in Control Group who do not made a purchase.
   Cust_in_Control_NotPurchased<-subset(Abandoned_Data_Seed,
   Abandoned_Data_Seed$Test_Control=="control" & Abandoned_Data_Seed$Outcome==0 )
   Cust_in_Control_NotPurchased
   > nrow(Cust_in_Control_NotPurchased)
   [1] 4051
10)      Customer in Control Group who made a purchase.
   Cust_in_Control_Purchased<-subset(Abandoned_Data_Seed,
   Abandoned_Data_Seed$Test_Control=="control" & Abandoned_Data_Seed$Outcome==1 )
   Cust_in_Control_Purchased
   nrow(Cust_in_Control_Purchased)
   > nrow(Cust_in_Control_Purchased)
   [1] 82
```

**Q8: Are there problematic cases? i.e. data records not matchable? If so, provide a few examples and toss those cases out of the analysis.**

Answer:
Incoming Phone, Contact Phone & Email are the key attributes to match data between two datasets. However, there could be a scenario where data was not recorded for Incoming Phone , Contact Phone & Email. In this case, we can remove that tuple from the dataset. In this project, I have covered as much scenarios as possible and performed the efficient way to match the data.

**Q9: Complete the following cross-tabulation:**

| **Group \ Outcome** | **Buy** | **No Buy** |
|---|---|---|
| **Treatment** | 301 | 3863 |

| | | |
|---|---|---|
| **Control** | 82 | 4051 |

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

Answer:

```
For state: AK
Customer in Treatment Group who do not made a purchase and lives in state "AK".
Cust_in_Test_NotPurchased_AK<-subset(Abandoned_Data_Seed, Abandoned_Data_Seed$Test_Control=
="test" & Abandoned_Data_Seed$Outcome==0 & Abandoned_Data_Seed$Address=="AK")
Cust_in_Test_NotPurchased_AK
> nrow(Cust_in_Test_NotPurchased_AK)
[1] 25

Customer in Treatment Group who made a purchase and lives in state "AK"..
Cust_in_Test_Purchased_AK<-subset(Abandoned_Data_Seed, Abandoned_Data_Seed$Test_Control=="t
est" & Abandoned_Data_Seed$Outcome==1 & Abandoned_Data_Seed$Address=="AK")
Cust_in_Test_Purchased_AK
> nrow(Cust_in_Test_Purchased_AK)
[1] 3

Customer in Control Group who do not made a purchase and lives in state "AK".
Cust_in_Control_NotPurchased_AK<-subset(Abandoned_Data_Seed, Abandoned_Data_Seed$Test_Contr
ol=="control" & Abandoned_Data_Seed$Outcome==0 & Abandoned_Data_Seed$Address=="AK")
Cust_in_Control_NotPurchased_AK
> nrow(Cust_in_Control_NotPurchased_AK)
[1] 32

Customer in Control Group who made a purchase and lives in state "AK"..
Cust_in_Control_Purchased_AK<-subset(Abandoned_Data_Seed, Abandoned_Data_Seed$Test_Control=
="control" & Abandoned_Data_Seed$Outcome==1 & Abandoned_Data_Seed$Address=="AK")
Cust_in_Control_Purchased_AK
nrow(Cust_in_Control_Purchased_AK)
```

| Group \ Outcome | Buy | No Buy |
|---|---|---|
| **Treatment** | 3 | 25 |
| **Control** | 0 | 32 |

In same way as above, we can calculate for the below mentioned randomly selected states.

**For IA:**

| Group \ Outcome | Buy | No Buy |
|---|---|---|
| **Treatment** | 5 | 32 |
| **Control** | 0 | 34 |

**For OK:**

| Group \ Outcome | Buy | No Buy |
|---|---|---|
| **Treatment** | 0 | 31 |
| **Control** | 0 | 38 |

**For NV:**

| Group \ Outcome | Buy | No Buy |
|---|---|---|

| | | |
|---|---|---|
| **Treatment** | 3 | 36 |
| **Control** | 2 | 50 |

**For WI:**

| **Group \ Outcome** | **Buy** | **No Buy** |
|---|---|---|
| **Treatment** | 2 | 28 |
| **Control** | 1 | 42 |

## 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 with the following columns

Customer ID | Test Variable | Outcome | Days_in_Between | D_State | D_Email |

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". Note also we have two dummies to signal whether the State and Email information is available for the customer.

(Note that you should have as many rows as customers you were able to match across the two data sets. Be sure to attach this excel file to the submission for proper verification.)

Please find attached "CleanedData" Excel file for this section.

## V. Statistical Analysis

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

**Q11:** Run a Linear regression model for

$$\text{Outcome} = \text{alpha} + \text{beta} * \text{Test\_Variable} + \text{error}$$

And Report the output.

Answer:

```
model=lm(CleanedAbandonedData$Outcome~CleanedAbandonedData$Test_Variable)
summary(model)
```

```
Call:
lm(formula = CleanedAbandonedData$Outcome ~ CleanedAbandonedData$Test_Variable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07229 -0.07229 -0.01984 -0.01984  0.98016

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.019840   0.003239   6.126 9.43e-10 ***
CleanedAbandonedData$Test_Variable 0.052446   0.004572  11.472  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2082 on 8295 degrees of freedom
Multiple R-squared:  0.01562,  Adjusted R-squared:  0.0155
F-statistic: 131.6 on 1 and 8295 DF,  p-value: < 2.2e-16

Outcome = 0.019840   + 0.052446* Test_Variable + 0.003239
```

**Q12:** Argue this is statistically equivalent to the A/B test procedure described in Leada Module 4. And so argue why it's important to randomize the data properly.

In Team Leada module 4, a visitor is randomly assigned to landing page A or B, and either converts or does not convert to a user. In this experiment also the assignment of test and control group is totally randomized and the outcome is calculated. Both of the experiments the analysis is being done using the binary values. As P-value in A/B test mentioned below is less than the alpha(5%) , therefore the Null hypothesis is to be rejected. This implies that the result is statistically significant and hence the idea of retargeting the customers is effective. It is important to randomize the data properly as it will provide unbiased dataset in order to derive a better conclusion on Hypothesis test and efficient linear regression model.

Ho: P(test)-P(control)<=0
H-alt: P(test)-P(control) >0
P(test)=Purchase took in treatment data/total number of treatment data
P(control)=Purchase took in control data/total number of control data

```
> t.test(CusInTestGrp$Outcome,CustInControlGrp$Outcome,alternative='greater')

        Welch Two Sample t-test

data:  CusInTestGrp$Outcome and CustInControlGrp$Outcome
t = 11.495, df = 6400.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.04494045        Inf
sample estimates:
 mean of x  mean of y
0.07228626 0.01984031
```

**Q13:** 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?

Answer:
In linear regression model calculated above has Test_Variable coefficients (slope) equal to 0.052446. This implies that our dependent variable "Outcome" would increase by .05 times for every unit increase in Test_Variable which is very less. Also the Adjusted R-squared is 0.0155 which is again very less and hence not a significant model. We must include other parameters and check for the significance in order to derive an efficient linear regression model.

**Q14:** Now add to the regression model the dummiesfor State and Emails. Also consider including interactions with the treatment.Report the outcome and comment on the results. (You can compare with Q10)

Answer:
Outcome = Test Variable + D_Email + D_State
LRM:

```
> model<- lm(CleanedAbandonedData$Outcome~CleanedAbandonedData$Test_Variable + CleanedAband
onedData$D_State +CleanedAbandonedData$D_Email)
> summary(model)

Call:
lm(formula = CleanedAbandonedData$Outcome ~ CleanedAbandonedData$Test_Variable +
    CleanedAbandonedData$D_State + CleanedAbandonedData$D_Email)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11235 -0.06075 -0.05979 -0.00820  0.99180

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.008198   0.003833   2.139 0.032496 *
CleanedAbandonedData$Test_Variable 0.051594   0.004561  11.312  < 2e-16 ***
CleanedAbandonedData$D_State       0.017260   0.004700   3.672 0.000242 ***
CleanedAbandonedData$D_Email       0.035296   0.007158   4.931 8.33e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2076 on 8293 degrees of freedom
Multiple R-squared:  0.02132,  Adjusted R-squared:  0.02096
F-statistic: 60.21 on 3 and 8293 DF,  p-value: < 2.2e-16
```

From above model we can infer that, with the inclusion of D_Email & D_State in the linear regression model points the increase in outcome. For a known state and Email the outcome result increases by .017260 and 0.035296 respectively. It has slightly higher value for Adjusted R-squared than the previous model, hence a better version than previous one. However, change in the slope of the Test_Variables in both the models re mains almost same.

After adding interaction variables,
Outcome = Test Variable + D_Email + D_State + INT_TV_DState + INT_TV_DEmail

```
> model<- lm(CleanedAbandonedData$Outcome~CleanedAbandonedData$Test_Variable + CleanedAband
onedData$D_State +CleanedAbandonedData$D_Email +CleanedAbandonedData$INT_TV_DState +Cleaned
AbandonedData$INT_TV_DEmail)
> summary(model)

Call:
lm(formula = CleanedAbandonedData$Outcome ~ CleanedAbandonedData$Test_Variable +
    CleanedAbandonedData$D_State + CleanedAbandonedData$D_Email +
    CleanedAbandonedData$INT_TV_DState + CleanedAbandonedData$INT_TV_DEmail)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13798 -0.05384 -0.03092 -0.01410  0.98590

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.014102   0.004362   3.233  0.00123 **
CleanedAbandonedData$Test_Variable 0.039739   0.006196   6.414 1.50e-10 ***
CleanedAbandonedData$D_State       0.011587   0.006648   1.743  0.08138 .
CleanedAbandonedData$D_Email       0.005234   0.010454   0.501  0.61661
CleanedAbandonedData$INT_TV_DState 0.011267   0.009390   1.200  0.23024
CleanedAbandonedData$INT_TV_DEmail 0.056054   0.014331   3.911 9.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2074 on 8291 degrees of freedom
Multiple R-squared:  0.02365,  Adjusted R-squared:  0.02306
F-statistic: 40.16 on 5 and 8291 DF,  p-value: < 2.2e-16
```

For the above model, we have included interaction between Test Variable & D_State and Test Variable & D_Email. This model seems to be more efficient than the previous model as it has higher Adjusted R - Squared compared to the previous models. Also, retargeting a customer with known state and Email, the results (coefficients:positive slope) seems to be more plausible that it could convert it in the positive outcome as one might give Email if he/she is interested in purchase or for more detailed information.

## V: Statistical Analysis: Response Times

**RQ2: You wantnow to investigate whether the response time (time to make a purchase after the first contact) is influenced by the retargeting campaign.**

Q15: 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?

Answer:

Case 1:

```
Outcome_vs_DaysInBet <- subset(CleanedAbandonedData, CleanedAbandonedData$Outcome==1)
model<- lm(Outcome_vs_DaysInBet$Outcome ~ Outcome_vs_DaysInBet$Days_in_Between)
summary(model)
```

```
Call:
lm(formula = Outcome_vs_DaysInBet$Outcome ~ Outcome_vs_DaysInBet$Days_in_Between)

Residuals:
       Min         1Q     Median         3Q        Max
-1.230e-13  5.400e-17  3.730e-16  5.930e-16  1.402e-15

Coefficients:
                                     Estimate Std. Error   t value Pr(>|t|)
(Intercept)                          1.000e+00  1.027e-15 9.740e+14   <2e-16 ***
Outcome_vs_DaysInBet$Days_in_Between 2.450e-17  1.988e-17 1.232e+00    0.219
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.324e-15 on 381 degrees of freedom
Multiple R-squared:  0.4999,   Adjusted R-squared:  0.4986
F-statistic: 380.9 on 1 and 381 DF,  p-value: < 2.2e-16
```

In the model above, we can comment that the linear model seems to be significant due to high value of Adjusted R square. However, we cannot comment on the fitness of linearity model considering only “Days In Between” in independent variable list.

Case 2:

LRM:

```
> model<- lm(ResDataFromAbandoned$Days_in_Between ~ ResDataFromAbandoned$Test_Variable)
> summary(model)

Call:
lm(formula = ResDataFromAbandoned$Days_in_Between ~ ResDataFromAbandoned$Test_Variable)

Residuals:
   Min     1Q Median     3Q    Max
-45.12 -11.12  -1.12  11.47  46.88

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                       44.939      1.784  25.191   <2e-16 ***
ResDataFromAbandoned$Test_Variable    5.181      2.012   2.574   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.15 on 381 degrees of freedom
Multiple R-squared:  0.0171,   Adjusted R-squared:  0.01452
F-statistic: 6.628 on 1 and 381 DF,  p-value: 0.01042
```

From above linear regression model, we can infer that the time lapse increases when the customers are retargeted as the slope of the Test Variable is 5.181. It could be due to a number of reasons like customers might be looking to check out other alternatives, and then compare and make a decision or they are waiting for new product to be launched or they are waiting for some discount offer or sale.

## VI: Conclusion

**Q16: 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**

Answer:

I have learned how to perform Data Cleaning, Data Matching and A/B test with large dataset are derive linear regression models. I got hands-on experience in R as I have used it end to end for this project. In designing the experiment, I would have included other factors like age, marital status, income to derive a better and efficient model. Managerial implication that can be concluded from this study is that the quality of sample data plays an indispensible role in the outcome of any experiment. Data sample collected using Random selection is considered fit and unbiased in such type of experiments.

**Q17: Self evaluation. Please score your effort on a scale 0-100. Please score your expected performance on the same scale. Add comments if necessary**

Effort: 100%
Expected Performance: 99%
I have invested my whopping time in learning R and used it to do all the Data Cleaning , Data Matching and analysis using different scenarios very efficiently.