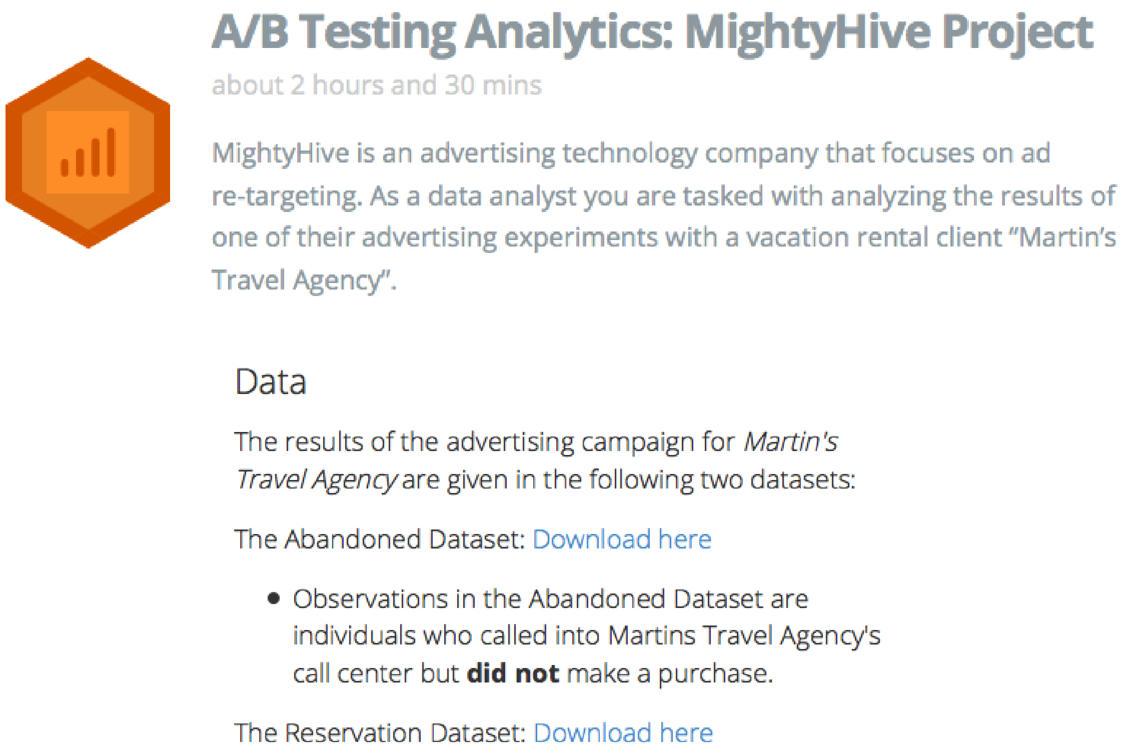
Jiten Punjabi



# A/B Testing Analytics: MightyHive Project

about 2 hours and 30 mins

MightyHive is an advertising technology company that focuses on ad re-targeting. As a data analyst you are tasked with analyzing the results of one of their advertising experiments with a vacation rental client "Martin's Travel Agency".

## Data

The results of the advertising campaign for *Martin's Travel Agency* are given in the following two datasets:

The Abandoned Dataset: Download here

- Observations in the Abandoned Dataset are individuals who called into Martins Travel Agency's call center but **did not** make a purchase.

The Reservation Dataset: Download here

## 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>

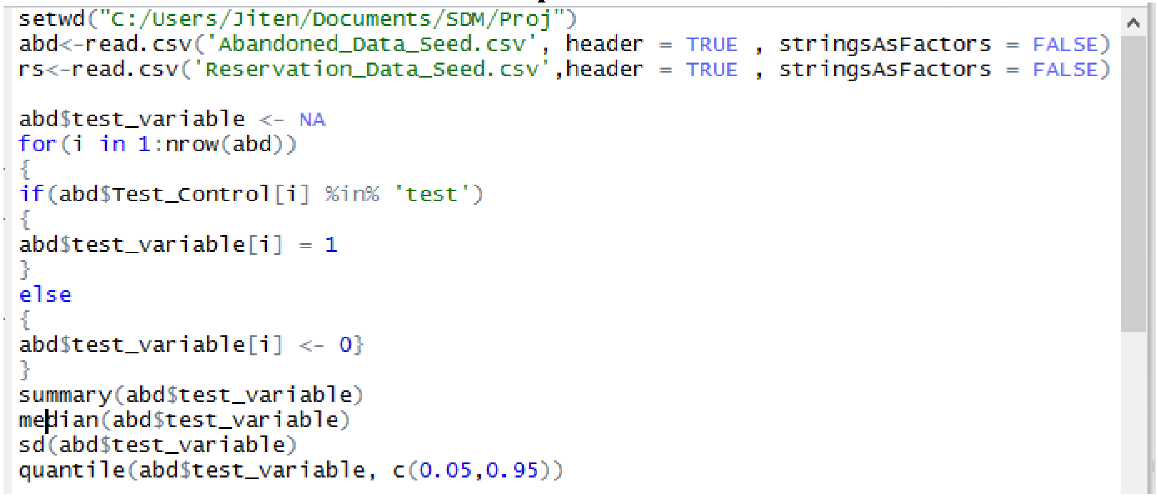
**Q1:** In light of your experience as a business woman/man, argue why this is a sensible business question.

**Retargeting allows you to showcase your product once again to the customer whom you would have already pursued but ended up not buying. Retargeting enables you to bring your brand back front and center and has proved to have high conversion rates. A customer who abandons the idea of buying a product, may end up buying the product eventually if that product is repeatedly advertised to the customer.**

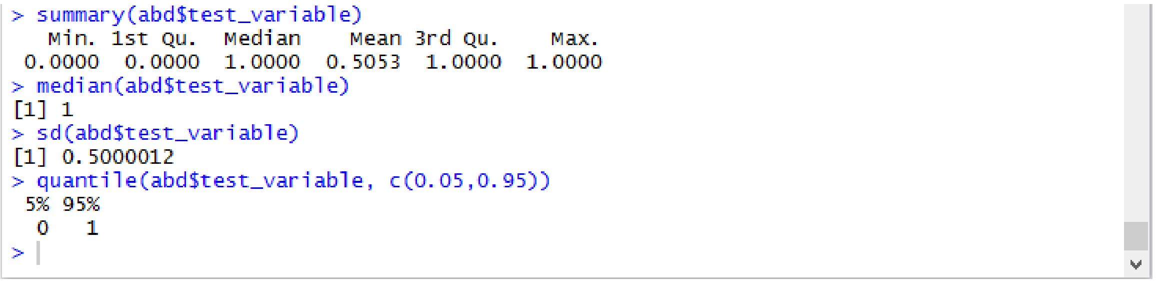
An experiment is run, where customers in the abandoned dataset are randomly placed in a treatment or in a control group (see column L in both files).
Those marked as "test" are retargeted (treated), the others marked as control are part of the control group.

**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.

**First created a new column test_variable, assigning 1 if in test group and 0 if in control group. Then calculated the summary statistics followed by median, standard deviation and 5th and 95th quantile.**



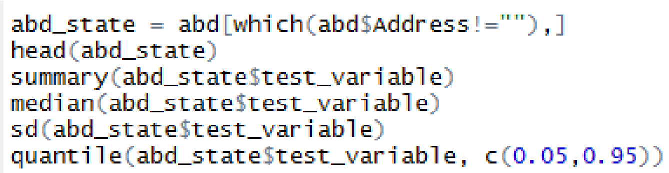
```
setwd("C:/Users/Jiten/Documents/SDM/Proj")
abd<-read.csv('Abandoned_Data_Seed.csv', header = TRUE , stringsAsFactors = FALSE)
rs<-read.csv('Reservation_Data_Seed.csv',header = TRUE , stringsAsFactors = FALSE)

abd$test_variable <- NA
for(i in 1:nrow(abd))
{
if(abd$Test_Control[i] %in% 'test')
{
abd$test_variable[i] = 1
}
else
{
abd$test_variable[i] <- 0}
}
summary(abd$test_variable)
median(abd$test_variable)
sd(abd$test_variable)
quantile(abd$test_variable, c(0.05,0.95))
```



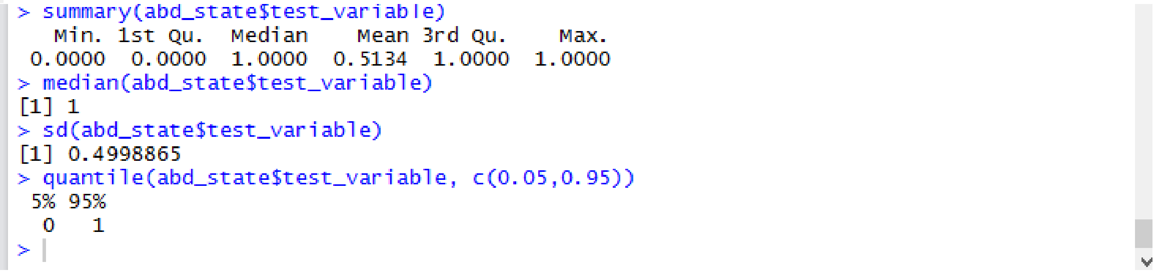
```
> summary(abd$test_variable)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  1.0000  0.5053  1.0000  1.0000 
> median(abd$test_variable)
[1] 1
> sd(abd$test_variable)
[1] 0.5000012
> quantile(abd$test_variable, c(0.05,0.95))
 5% 95% 
  0   1 
> |
```

**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.

**First created data frame called abd_state using which() function. It only considers those entries with known states. Then calculating the summary statistics followed by median, standard deviation and 5th and 95th quantile.**



```
abd_state = abd[which(abd$Address!=""),]
head(abd_state)
summary(abd_state$test_variable)
median(abd_state$test_variable)
sd(abd_state$test_variable)
quantile(abd_state$test_variable, c(0.05,0.95))
```



```
> summary(abd_state$test_variable)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  1.0000  0.5134  1.0000  1.0000 
> median(abd_state$test_variable)
[1] 1
> sd(abd_state$test_variable)
[1] 0.4998865
> quantile(abd_state$test_variable, c(0.05,0.95))
 5% 95% 
  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-only level?

**When considering all the entries of ABD file, the mean turns out to be 0.5053. Looks like almost equal number of assignments were done to treatment and control groups. In case of state-only level the mean turns out to be 0.5134. Here, the mean has increased, so it appears to be that number of entries in the treatment group is more than the number of entries in the control group.**

## 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.

**Not observing outcomes for some customers can be problematic for causal inference since we will not have proper results on how many customers bought the package and how many did not, and hence we will not be able to predict the effectiveness of the retargeting campaign.**
**However, the Reservation data file consists of all the customers that have purchased the vacation package, irrelevant of whether that person was in test or control. And we have the list of all the customers being retargeted in Abandon data file. Hence, by matching the list of customers in Abandon file present in Reservation data file, we will exactly know how many customers that were in Abandon file, bought the package.**

**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)

**Observing the two data files, it can be seen that the data can be matched across the three data keys. Most of the records have incoming phone information. Hence, maximum matching of the records between the two files can be done using incoming phone number data. Now there are many records with no incoming phone information. For these records we can use either contact phone number or Email ID of the customer to match the customer from ABD file to RS file. Hence, the data keys that we can use are**

1. **Incoming Phone**
2. **Contact Phone**
3. **E-mail ID**

**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.

**I used excel for data matching procedure. First, I found out the common entries between ABD file and RS file using VLOOKUP. This gave me the list of all the customers in ABD file, who abandoned the vacation package but ended up buying. However, this list consisted customers belonging to both the groups, treatment and control groups. The rest of the entries in the ABD file, i.e. the entries in the ABD file which were not present in RS file, ended up not buying the vacation package. With this I had the list of customers who bought and the list of customers who did not buy. Now, using the column test_control, I created another column test_variable with value 0 if in control group and 1 if in test group. In order to find the list of entries in the 4 categories, I did the following respectively,**

**(1) VLOOKUP on ABD file and RS file, and with value 1 in test_variable**
**(2) Rest of the entries in ABD file and with value 1 in test_variable**
**(3) VLOOKUP on ABD file and RS file, and with value 0 in test_variable**
**(4) Rest of the entries in ABD file and with value 0 in test_variable**

**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.

**Yes, there are problematic cases. There are many duplicate entries in the data files. Also, there are many entries where there is no information about the customer, except the incoming phone. There are entries with the same incoming phone number and different customer names. Also many customers have shared the same Contact Phone, hence tossing out all these entries which look insignificant or have little or no information.**

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

| **Group \ Outcome** | **Buy** | **No Buy** |
|---|---|---|
| **Treatment** | **290** | **3976** |
| **Control** | **77** | **4099** |

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

**State: CA**

| **Group \ Outcome** | **Buy** | **No Buy** |
|---|---|---|
| **Treatment** | **6** | **42** |
| **Control** | **0** | **37** |

**State: FL**

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

**State: GA**

| Group \ Outcome | Buy | No Buy |
|---|---|---|
| **Treatment** | **2** | **45** |
| **Control** | **0** | **33** |

**State: NJ**

| Group \ Outcome | Buy | No Buy |
|---|---|---|
| **Treatment** | **4** | **48** |
| **Control** | **3** | **33** |

**State: IL**

| Group \ Outcome | Buy | No Buy |
|---|---|---|
| **Treatment** | **2** | **35** |
| **Control** | **0** | **47** |

## 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.)

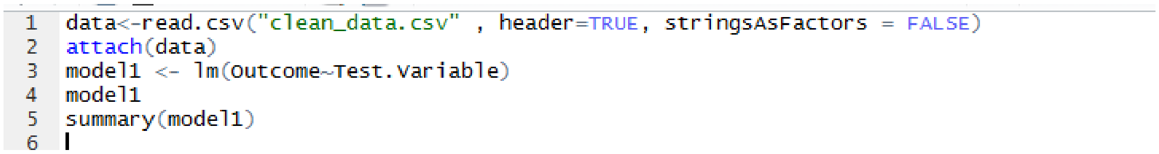
## IV. Statistical Analysis

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

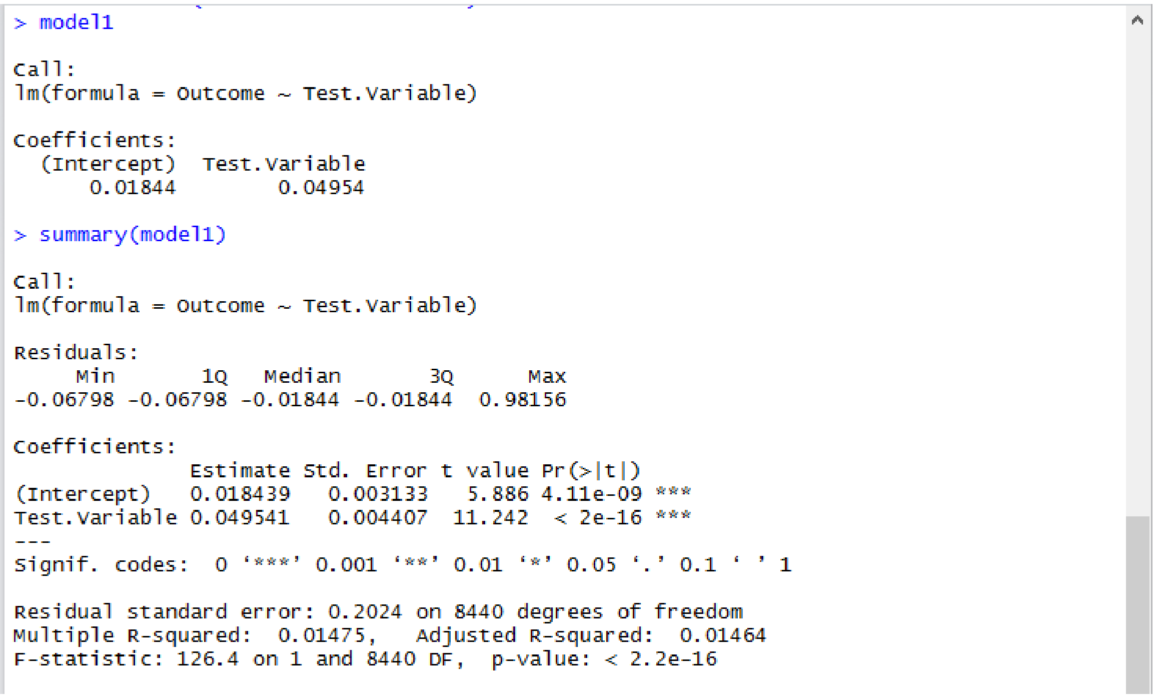
**Q11:** Run a Linear regression model for

Outcome = alpha + beta * Test_Variable + error

And Report the output.



```
data<-read.csv("clean_data.csv" , header=TRUE, stringsAsFactors = FALSE)
attach(data)
model1 <- lm(Outcome~Test.Variable)
model1
summary(model1)

```



```
> model1

Call:
lm(formula = Outcome ~ Test.Variable)

Coefficients:
  (Intercept)  Test.Variable
      0.01844        0.04954

> summary(model1)

Call:
lm(formula = Outcome ~ Test.Variable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06798 -0.06798 -0.01844 -0.01844  0.98156

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.018439   0.003133   5.886 4.11e-09 ***
Test.Variable 0.049541   0.004407  11.242  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2024 on 8440 degrees of freedom
Multiple R-squared:  0.01475,	Adjusted R-squared:  0.01464
F-statistic: 126.4 on 1 and 8440 DF,  p-value: < 2.2e-16
```

**So, after running the Linear Regression model on Outcome as the dependent variable and Test.Variable as the independent variable, we get the coefficients alpha = *0.0184* and beta = *0.0495* and error = *0.2024* .**

**Hence we have the equation,**

**Outcome = 0.01844 + 0.04954*Test.Variable + 0.2024**

**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 Teamleada module 4, for the landing page example, they have divided the data in two sets 'treatment' and 'control', just like our case here with Abandon data seed and reservation data see. In Leada, it is given that,
**Our null hypothesis:** the difference of conversion rates between landing page B (P-hat 1) and landing page A (P-hat 2) is equal to 0

**Our alternative hypothesis:** the difference of conversion rates between landing page B (P-hat 1) and landing page A (P-hat 2) is greater than 0.

Similarly, in our case,
Null hypothesis : difference between conversion rate between **test** group and **control** group is equal to 0
Alternate hypothesis : difference between conversion rate between **test** group and **control** group is greater than 0

For **null hypothesis** we have,
$H_0 : \mu_T - \mu_c = 0$
For **alternate hypothesis** we have,
$H_a : \mu_T - \mu_c > 0$

Hence, for $\mu_T$
Outcome = alpha + beta*(1) + error

And for $\mu_c$ ,
Outcome = alpha + beta*(0) + error = alpha + error
$\mu_T - \mu_c$= beta

Thus, for null hypothesis, **beta** should be equal to 0 , but clearly from the above model **beta = 0.04954** hence we can reject the null hypothesis and conclude that there is higher conversion rate in the **treatment** group compared to the **control** group.

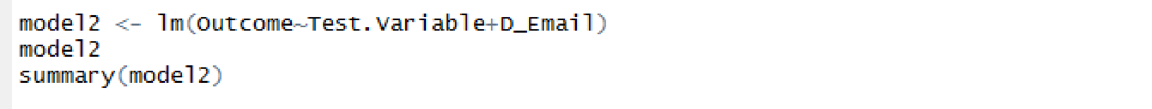
It is important to randomize the data for the experiment to provide accurate results. The test and control groups should be picked randomly. A random sample is a sample chosen that allows all subjects an equal probability of being selected.

**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?

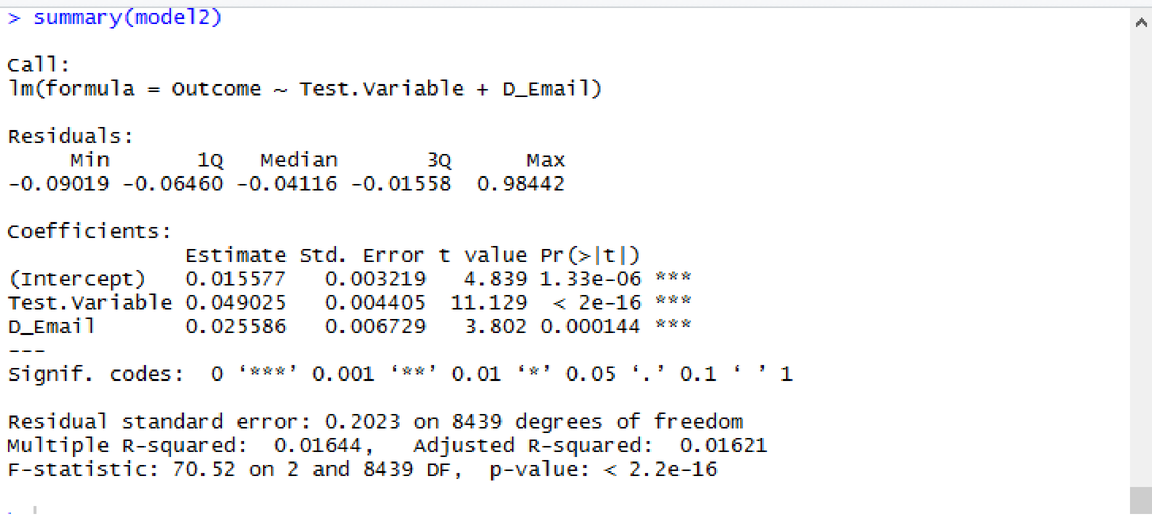
**According to the linear regression model run above, we can conclude that retargeting campaign helped in selling vacation package to the customers. However, we cannot infer much apart from that information. The result don't seem to be very statistically significant.**
**The Outcome in our case is either 0 or 1 which represents Buy or No buy. A linear regression model is the best predictor when the dependent or response variable is continuous. Though the linear regression model helps us identify that the campaign was effective, since beta ≠ 0, however our outcome is a binomial variable, 0 or 1, therefore using a linear probability model or logistic regression would provide us with better results. We would exactly know about the effectiveness of the retargeting campaign.**

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

**First, running the regression model on Dummy for E-mail. We have,**

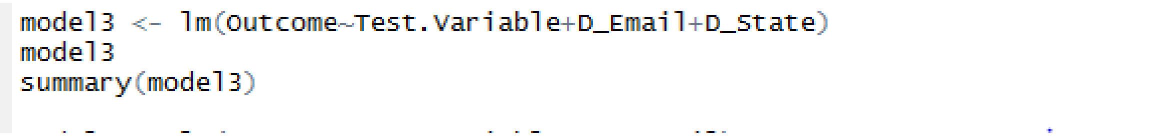


```
model2 <- lm(Outcome~Test.Variable+D_Email)
model2
summary(model2)
```

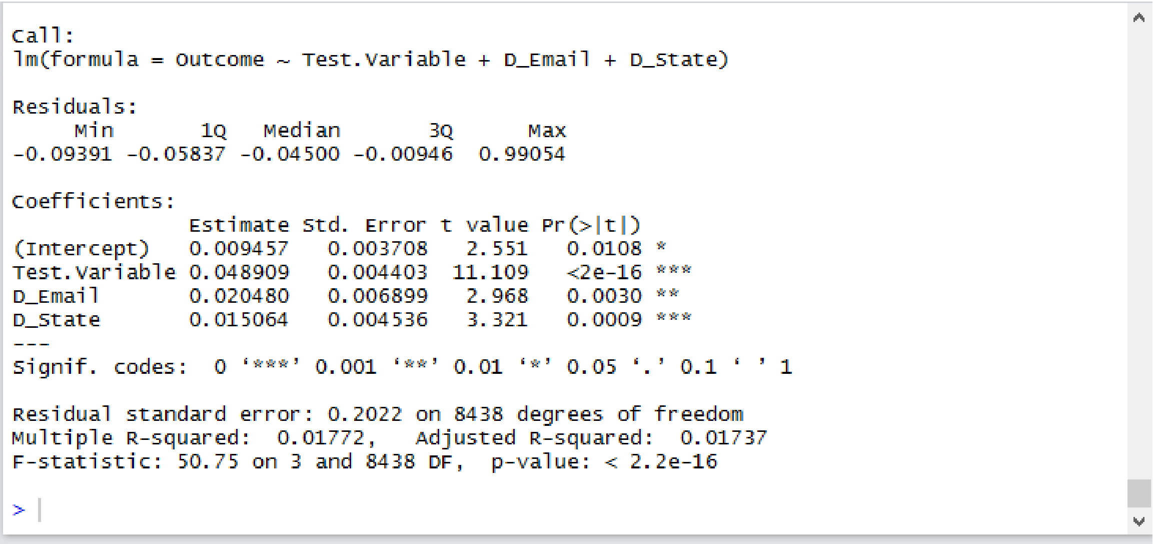


```
> summary(model2)

Call:
lm(formula = Outcome ~ Test.Variable + D_Email)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.09019 -0.06460 -0.04116 -0.01558  0.98442 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.015577   0.003219   4.839 1.33e-06 ***
Test.Variable 0.049025   0.004405  11.129  < 2e-16 ***
D_Email       0.025586   0.006729   3.802 0.000144 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2023 on 8439 degrees of freedom
Multiple R-squared:  0.01644,	Adjusted R-squared:  0.01621 
F-statistic: 70.52 on 2 and 8439 DF,  p-value: < 2.2e-16
```

**Here, we can see that the value of D_Email is quiet significant,**
**Now running the model with adding Dummy for state**



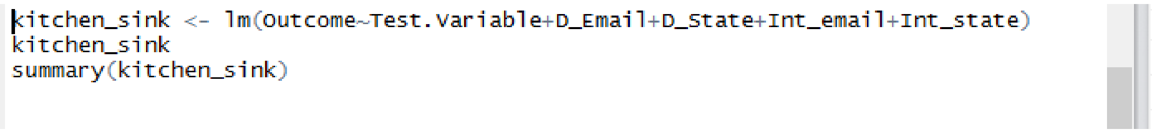
```
model3 <- lm(Outcome~Test.Variable+D_Email+D_State)
model3
summary(model3)
```



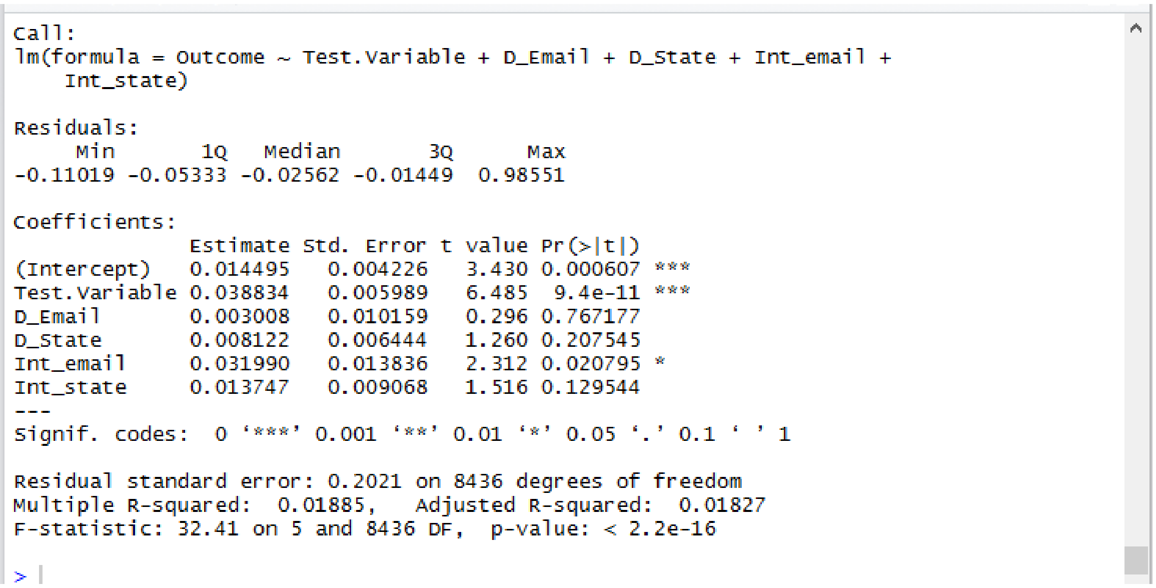
```
Call:
lm(formula = Outcome ~ Test.Variable + D_Email + D_State)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.09391 -0.05837 -0.04500 -0.00946  0.99054 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.009457   0.003708   2.551   0.0108 *  
Test.Variable 0.048909   0.004403  11.109   <2e-16 ***
D_Email       0.020480   0.006899   2.968   0.0030 ** 
D_State       0.015064   0.004536   3.321   0.0009 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2022 on 8438 degrees of freedom
Multiple R-squared:  0.01772,	Adjusted R-squared:  0.01737 
F-statistic: 50.75 on 3 and 8438 DF,  p-value: < 2.2e-16

> 
```

**We can see that D_Email and D_State are both significant variables in the model, we can also see an increase in R-squared values by adding Dummy for state.**

**We now try to run the kitchen sink model, by adding all the dummy variables along with interaction variables with interaction on State and Test variable and Email and test variable.**

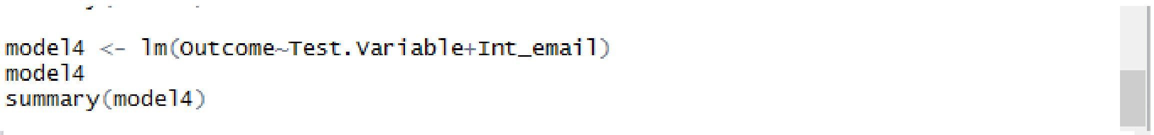


```
kitchen_sink <- lm(Outcome~Test.Variable+D_Email+D_State+Int_email+Int_state)
kitchen_sink
summary(kitchen_sink)
```

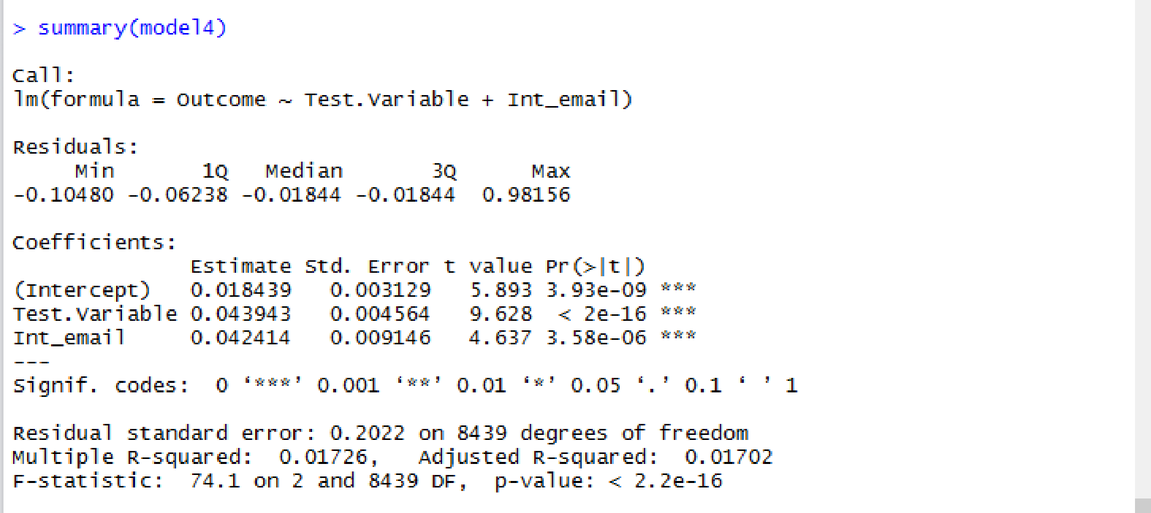


```
Call:
lm(formula = Outcome ~ Test.Variable + D_Email + D_State + Int_email + 
    Int_state)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.11019 -0.05333 -0.02562 -0.01449  0.98551 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.014495   0.004226   3.430 0.000607 ***
Test.Variable 0.038834   0.005989   6.485  9.4e-11 ***
D_Email       0.003008   0.010159   0.296 0.767177    
D_State       0.008122   0.006444   1.260 0.207545    
Int_email     0.031990   0.013836   2.312 0.020795 *  
Int_state     0.013747   0.009068   1.516 0.129544    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2021 on 8436 degrees of freedom
Multiple R-squared:  0.01885,	Adjusted R-squared:  0.01827 
F-statistic: 32.41 on 5 and 8436 DF,  p-value: < 2.2e-16

> 
```

**From the kitchen sink model, we can see that except for Int_email variable, all the other variables are insignificant. Hence, we toss out all the other variables and run the regression with just Test.Variable and Int_email.**



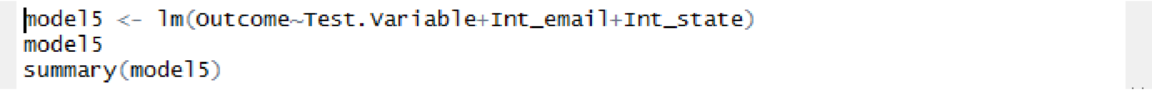
```
model4 <- lm(Outcome~Test.Variable+Int_email)
model4
summary(model4)
```



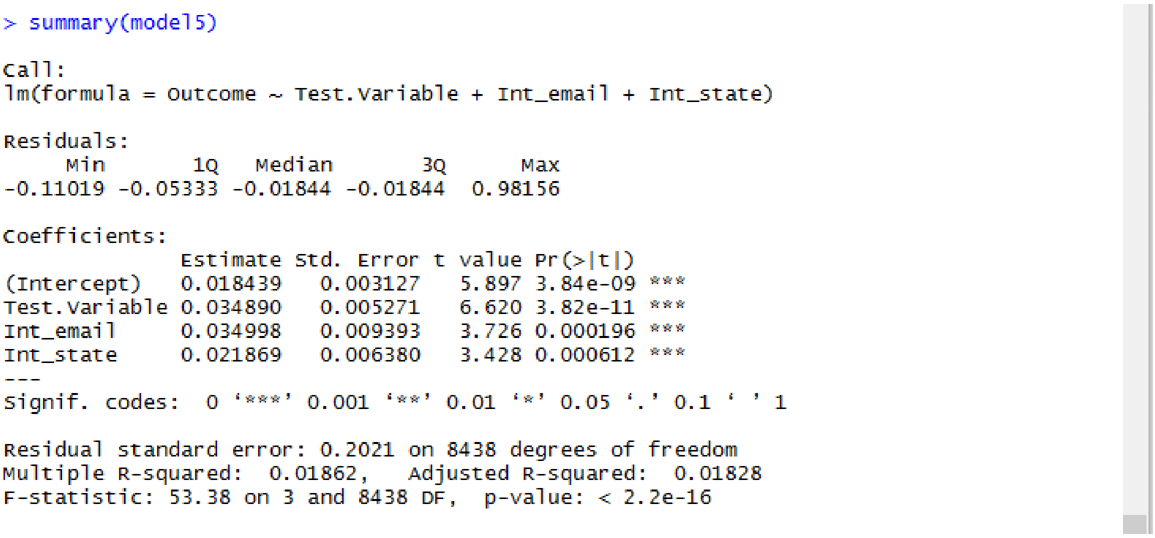
```
> summary(model4)

Call:
lm(formula = Outcome ~ Test.Variable + Int_email)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.10480 -0.06238 -0.01844 -0.01844  0.98156 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.018439   0.003129   5.893 3.93e-09 ***
Test.Variable 0.043943   0.004564   9.628  < 2e-16 ***
Int_email     0.042414   0.009146   4.637 3.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2022 on 8439 degrees of freedom
Multiple R-squared:  0.01726,	Adjusted R-squared:  0.01702 
F-statistic:  74.1 on 2 and 8439 DF,  p-value: < 2.2e-16
```

**There is a slight decrease in R-squared value, but we can see that both Test.variable and Int_email have become more significant compared to the previous model.**

**Also adding Int_state to the above model to see how it affects the significance,**



```
model5 <- lm(Outcome~Test.Variable+Int_email+Int_state)
model5
summary(model5)
```



```
> summary(model5)

Call:
lm(formula = Outcome ~ Test.Variable + Int_email + Int_state)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.11019 -0.05333 -0.01844 -0.01844  0.98156 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.018439   0.003127   5.897 3.84e-09 ***
Test.Variable 0.034890   0.005271   6.620 3.82e-11 ***
Int_email     0.034998   0.009393   3.726 0.000196 ***
Int_state     0.021869   0.006380   3.428 0.000612 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2021 on 8438 degrees of freedom
Multiple R-squared:  0.01862,	Adjusted R-squared:  0.01828 
F-statistic: 53.38 on 3 and 8438 DF,  p-value: < 2.2e-16
```

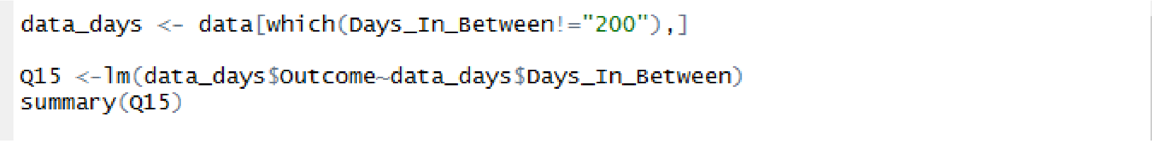
**Adding Int_state to the above model improves the R-squared value and all the three variables are quiet significant to the model. This model is the best predictor of the outcome.**

### V: Statistical Analysis: Response Times

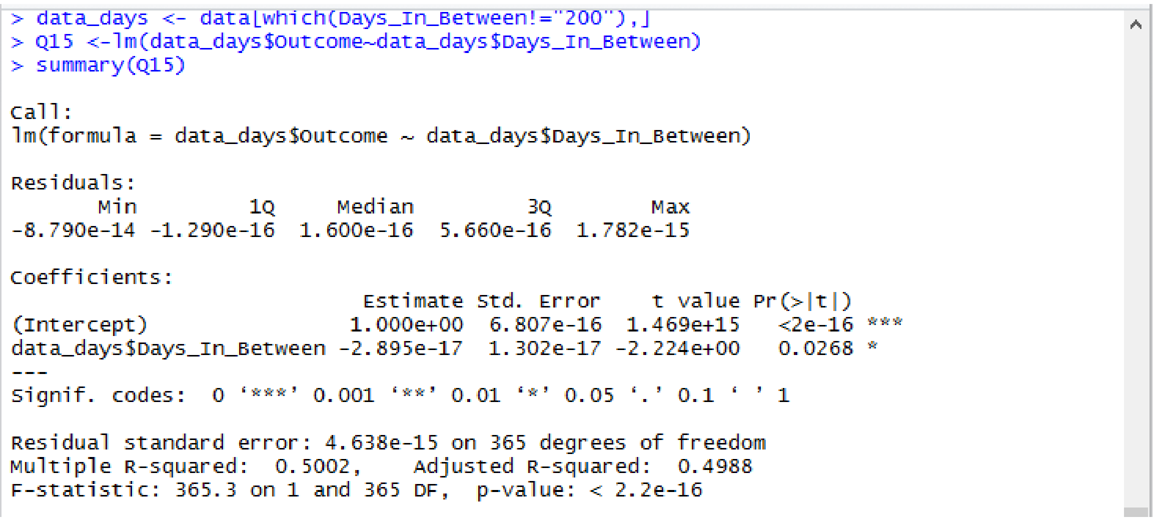
**RQ2: You want now to investigate whether the response time (time to make a purchase after the first contact) is influenced by the retargeting campaign.**

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?

**First separated the data entries with values not equal to 200. Then running the linear regression model on Outcome as the dependent variable and Days_In_Between as independent variable.**



```
data_days <- data[which(Days_In_Between!="200"),]

Q15 <-lm(data_days$Outcome~data_days$Days_In_Between)
summary(Q15)
```



```
> data_days <- data[which(Days_In_Between!="200"),]
> Q15 <-lm(data_days$Outcome~data_days$Days_In_Between)
> summary(Q15)

Call:
lm(formula = data_days$Outcome ~ data_days$Days_In_Between)

Residuals:
       Min         1Q     Median         3Q        Max 
-8.790e-14 -1.290e-16  1.600e-16  5.660e-16  1.782e-15 

Coefficients:
                           Estimate Std. Error    t value Pr(>|t|)    
(Intercept)               1.000e+00  6.807e-16  1.469e+15   <2e-16 ***
data_days$Days_In_Between -2.895e-17  1.302e-17 -2.224e+00   0.0268 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.638e-15 on 365 degrees of freedom
Multiple R-squared:  0.5002,	Adjusted R-squared:  0.4988 
F-statistic: 365.3 on 1 and 365 DF,  p-value: < 2.2e-16
```

**This model actually tells us about the effectiveness of retargeting the customers with respect to the response time of those customers. Beta here is negative and very small. We can also see that beta is significant.**

## 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**

**Retargeting is in fact effective, it can be seen from the results. People who were in the treatment group and who were retargeted bought more packages compared to people who were in control group.**

**A few things that I would have done differently would be, proper randomization of the data. There were almost equal number of entries in Treatment and Control groups, however, when considering the E-mail or State information, there was unbalanced distribution of data. And since the experiment is based on retargeting, such variables also need to be randomized equally. Also, when collecting data for Abandon and Reservation datasets, I would make sure that, as much as possible information would be**

**collected about the customers for better analysis. I would also make sure that the dataset has a mapping key for each customer, which will make the data cleaning and matching part easier, since it is the most time consuming part of the analysis.**