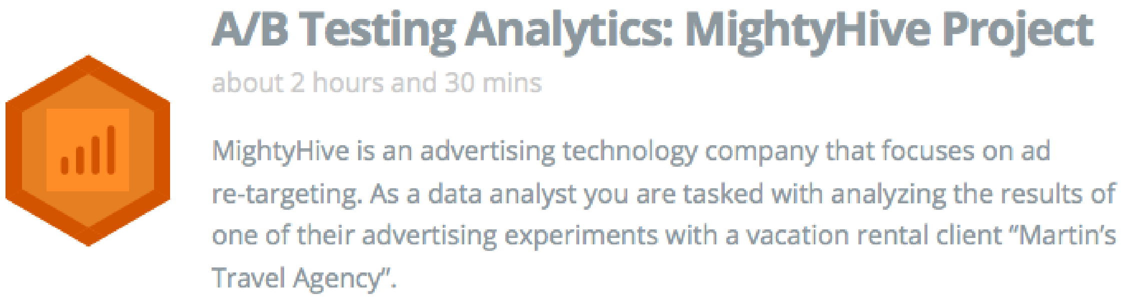
TEAM LEADA PROJECT

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"



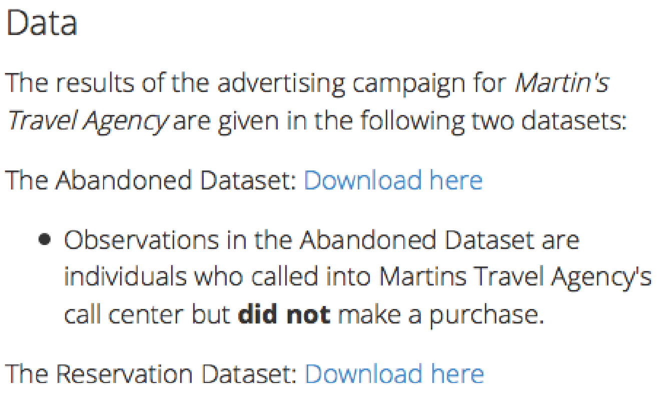
A/B Testing Analytics: MightyHive Project

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".

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)**



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

Figure 2: Where to download the two datasets

EXAM

Feel free to use this document as a Template.

Name: Gaurav Jetley
Section: ISM6137.001F15
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.

<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.

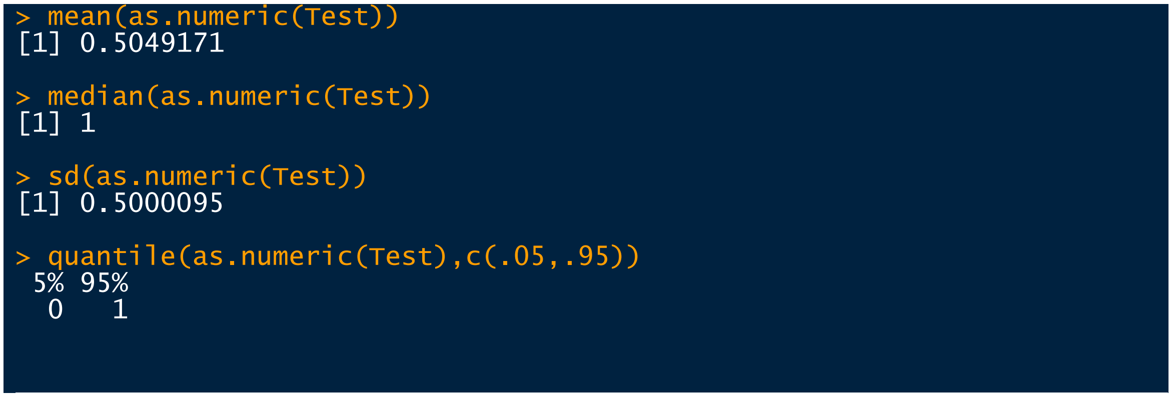
Retargeting customers who have already engaged with the business is an effective business and marketing strategy and is a must for any business. It's highly important for the following reasons:

1) *Builds brand awareness by capitalizing on initial audience*
2) *There are costs associated with expanding customer base thus retargeting known potential customers who ended up not buying makes sense*
3) *Producing relevant marketing strategy for customers based on their past interaction with the business produces positive results in terms of brand awareness*
4) *Converting potential customers on first visit is very challenging and retargeting the potential customers to change them to customers holds great business value*

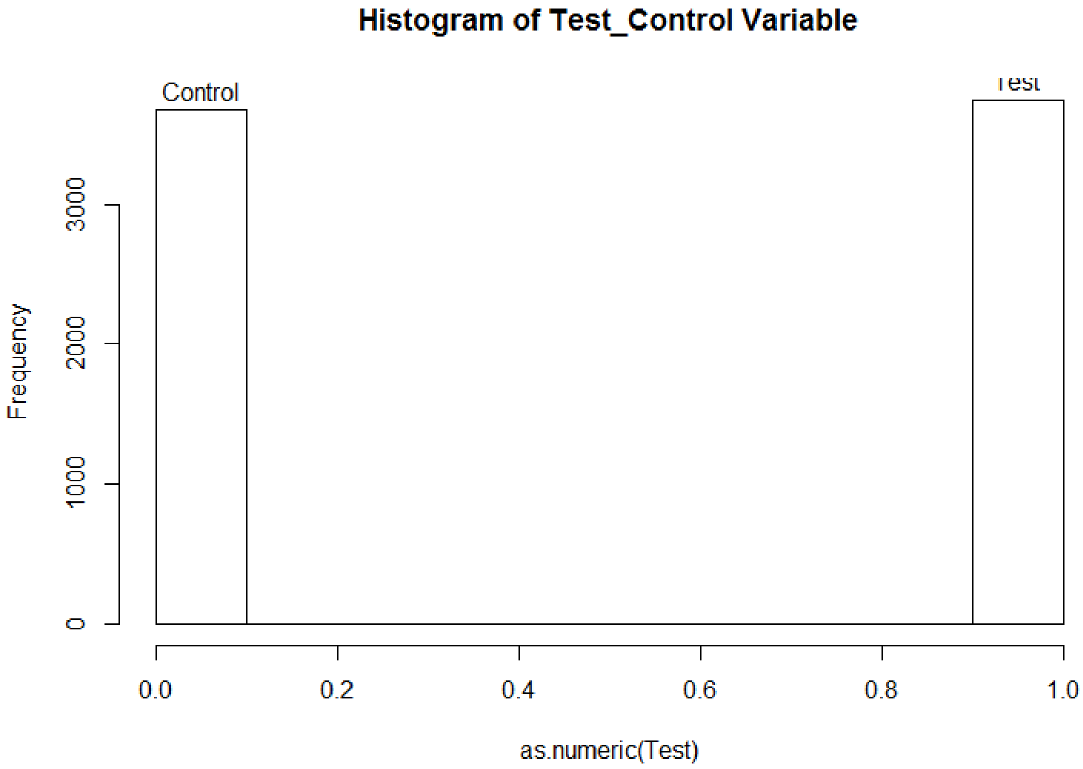
This is why "Should we retarget those customers?" is an important business question.

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.

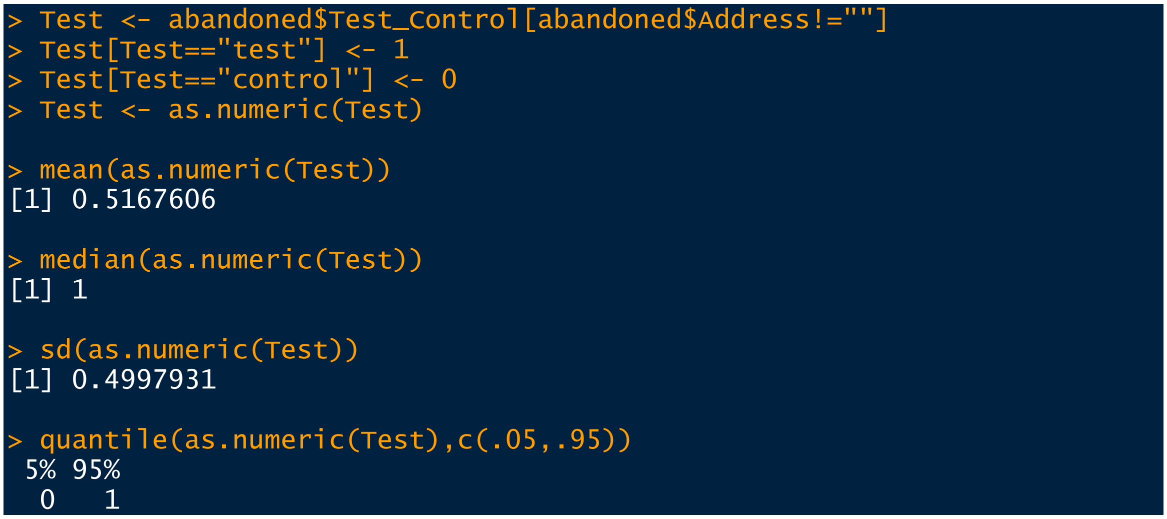
Q2: compute the summary statistics (mean, median, q5, q95, standard deviation) of the Test_variable: a dummy with a value of 1 if tested 0 if control in the ABD database.



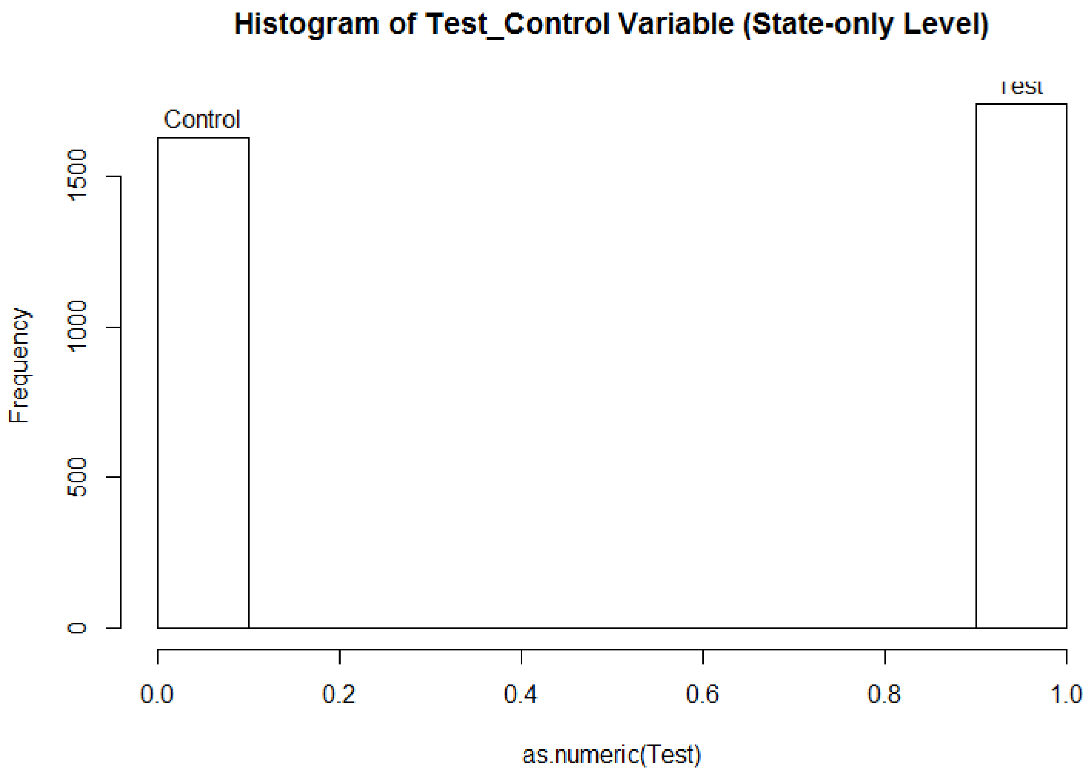
```
> mean(as.numeric(Test))
[1] 0.5049171

> median(as.numeric(Test))
[1] 1

> sd(as.numeric(Test))
[1] 0.5000095

> quantile(as.numeric(Test),c(.05,.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.



```
> Test <- abandoned$Test_Control[abandoned$Address!=""]
> Test[Test=="test"] <- 1
> Test[Test=="control"] <- 0
> Test <- as.numeric(Test)

> mean(as.numeric(Test))
[1] 0.5167606

> median(as.numeric(Test))
[1] 1

> sd(as.numeric(Test))
[1] 0.4997931

> quantile(as.numeric(Test),c(.05,.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?

The experiment seems to be executed properly. The summary statistics for Test variable at the "State-Only" level are almost the same as the entire dataset. No apparent imbalance in the assignments are visible.

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.

Missing data is always problematic. Causal Inference based on such problematic data can lead to potential problems like selection bias, non-randomization of treatment and control groups and missing important observations.

Such type of data collection also pose the problems in matching as some observations are not matched or give more than two matches. Some customers' "outcomes" are not present but only one outcome for customers is present who bought the service. Also a lot of observations in the RD dataset are not present in ABD dataset giving us no choice but to discard those observations. There is also the problem of duplicate observations and non-presence of the same Primary Key for both datasets which would have been very helpful in this situation.

This issue can be resolved by carefully matching ABD obs with RD obs. The obs which match can have a positive outcome (1) and the obs that didn't match can be labeled has having a negative outcome (0).

These problems could have been avoided with a better designed and well thought of database system.

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)

The non-presence of the same Primary Key in both datasets created an extra problem that could have been avoided.

Some Secondary Keys can be created by using some of the variables in both datasets. The Secondary Keys that I used in my data cleaning process are as follows:

1) *Concatenate(Incoming_Phone, Contact_Phone)*
2) *Concatenate(Incoming_Phone, Email)*

Both of the Secondary keys gave best results and lowered the amount of duplicate observations. Incoming_Phone was chosen as a variable in both the keys because it was the most populated of the three variables.
There was high variation and a higher degree of sparseness in other variables like Address, City, Zipcode etc and were unusable as part of the keys.

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.

Matching Procedure: ("Third Take")

1) *Created 2 primary keys in excel for both ABD and RS datasets:*
    a. *Concatenate(Incoming_Phone, Contact_Phone)*
    b. *Concatenate(Incoming_Phone, Email)*
2) *Removed all "Blank" Keys*
3) *Loaded the datasets into R*
4) *Matched datasets using both keys and an OR operator*
    a. *175 matches in ABD after deduplication*
    b. *171 matches in RS after deduplication*
5) *Created matched datasets for both ABD and RS*
6) *Created Indexes in the ABD dataset for corresponding RS dataset obs*
    a. *Indexinco => index for Incoming and Contact phone*
    b. *Indexinem => index for Incoming phone and Email*
7) *Added index in the matched ABD dataset*
8) *Exported datasets as CSV files for further processing in Excel*
9) *Followed the following documented steps in Excel:*

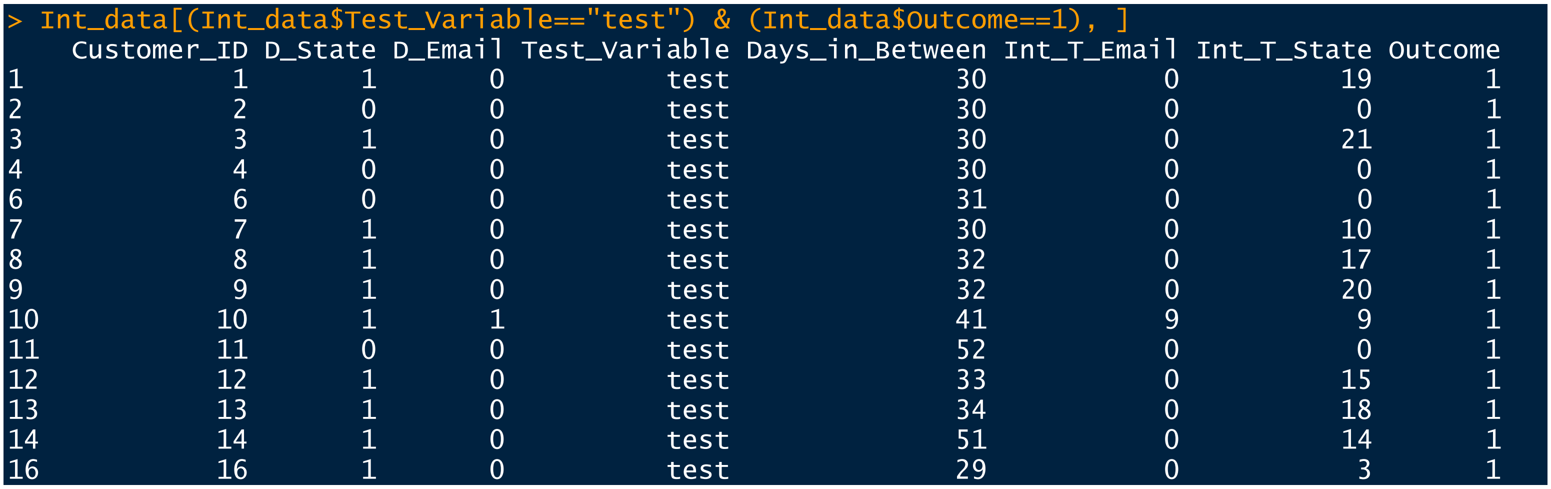
##### Done in Excel

#NON INTERACTION DATASET
#Creating Reservation_Index in excel in both datasets to match properly
#Sorting the datasets in excel according to Res_Index
#Converting the session variable to datetime in excel
#Subtracting Reservation session with Abandoned session to get days_in_between
#Creating days_in_between variable in excel
#subtituting session variable with days_in_between variable in the abandoned dataset
#Combining the abandoned dataset with newly created both_matched_v3 dataset
#removing duplicates with criteria (incoming and contact) first and then (email and incoming). (7325 obs left)
#replacing days_in_between variable values in abandoned dataset rows with 200. (abandoned dataset rows are above 171)
#deleting Caller_ID, Last_Name, Street, City, Zipcode variables in the both_matched_aban dataset
#Adding Customer_ID variable with index as ID from 1 to 7325
#Creating Variables D_Email (email given: 1), D_State (state given: 1), Outcome (reservation: 1), Test_Variable (test,control)
#Saving dataset as both_matched_aban_v3
#Removing Address, Email, Incoming_Phone, Contact_Phone
#Saving as both_matched_aban_v4
#deleting First_Name variable and saving dataset as both_matched_aban_v5
#Final dataset is both_matched_aban_v5 (no interactions)

#INTERACTION DATASET
#Saved in Interaction Folder
#both_matched_v3 includes matched abandoned and reservation obs with Date, Time, days_in_between, in_co, in_em
# Creating Interactions:
#Int_T_Email: 0: no email given, time value: email given
#Int_T_State: 0: no state given, 1: state given
#Saved as both_match_V4
#Combining abandoned with both_match_v4 and saving as both_match_aban_v1
#Populating Values for Time, Date, Days_in_Between, In_Co, In-Em, Int_T_Email, Int_T_State
#Saving as both_match_aban_v2
# Removing Duplicates using In_Co first and then In_Em (7325 obs left)
#Saving as both_match_aban_v3
# Converting both interaction variable values to "only hour" values
#Saving as both_match_aban_v4
#Created Customer_ID, D_Email, D_State, Outcome variables
#Saving as both_match_aban_v5
#Deleting Address, email variables and saving as both_match_aban_v6
#Deleting First_Name and saving as both_match_aban_v7
*#Adding Interactions variables Int_T_State_bin (binary), Int_T_Email_bin (binary), Int_Test_Email (Test*D_Email), Int_Test_State(Test*D_State), Test_Var (binary)*
#Readding First_Names for future analysis (Male/Female) using NLP
#Saving as both_matching_v8

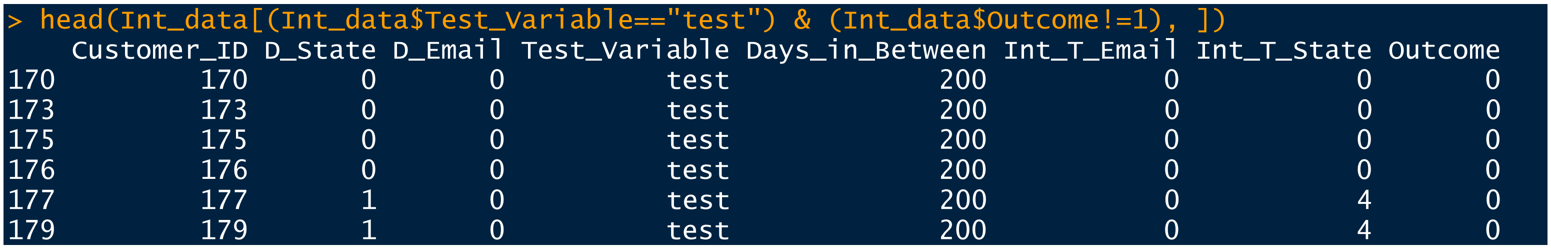
10) *Read the Final Cleaned Datasets (both_matching_aban_v7) back into R*
11) *All datasets, code and files can be found at*
    https://github.com/gauravjetley/MightyHive-Project

(1) Identification of Customers in the TREATMENT group who bought:



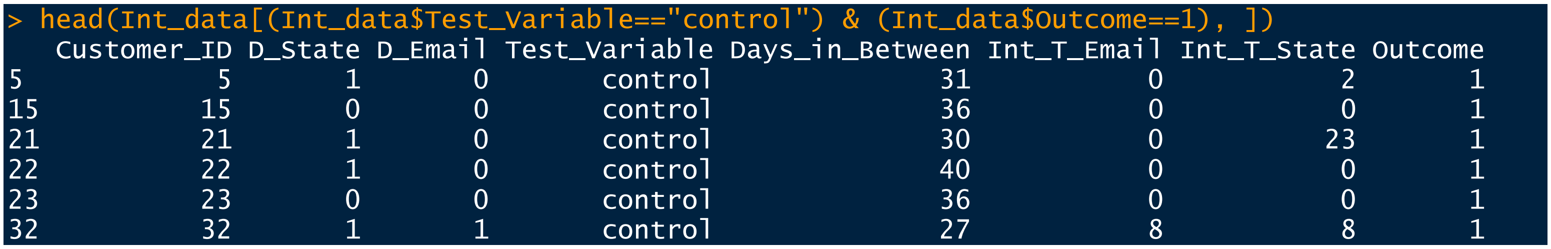
```
> Int_data[(Int_data$Test_Variable=="test") & (Int_data$Outcome==1), ]
   Customer_ID D_State D_Email Test_Variable Days_in_Between Int_T_Email Int_T_State Outcome
1            1       1       0          test              30           0          19       1
2            2       0       0          test              30           0           0       1
3            3       1       0          test              30           0          21       1
4            4       0       0          test              30           0           0       1
6            6       0       0          test              31           0           0       1
7            7       1       0          test              30           0          10       1
8            8       1       0          test              32           0          17       1
9            9       1       0          test              32           0          20       1
10          10       1       1          test              41           9           9       1
11          11       0       0          test              52           0           0       1
12          12       1       0          test              33           0          15       1
13          13       1       0          test              34           0          18       1
14          14       1       0          test              51           0          14       1
16          16       1       0          test              29           0           3       1
```

(2) Identification of Customers in the TREATMENT group who did not buy:



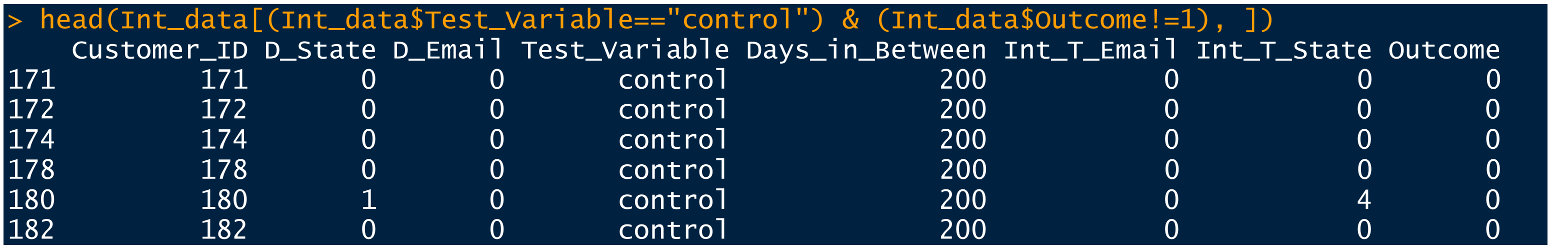
```
> head(Int_data[(Int_data$Test_Variable=="test") & (Int_data$Outcome!=1), ])
    Customer_ID D_State D_Email Test_Variable Days_in_Between Int_T_Email Int_T_State Outcome
170         170       0       0          test             200           0           0       0
173         173       0       0          test             200           0           0       0
175         175       0       0          test             200           0           0       0
176         176       0       0          test             200           0           0       0
177         177       1       0          test             200           0           4       0
179         179       1       0          test             200           0           4       0
```

(3) Identification of Customers in the Control group who bought:



```
> head(Int_data[(Int_data$Test_Variable=="control") & (Int_data$Outcome==1), ])
   Customer_ID D_State D_Email Test_Variable Days_in_Between Int_T_Email Int_T_State Outcome
5            5       1       0       control              31           0           2       1
15          15       0       0       control              36           0           0       1
21          21       1       0       control              30           0          23       1
22          22       1       0       control              40           0           0       1
23          23       0       0       control              36           0           0       1
32          32       1       1       control              27           8           8       1
```

(4) Identification of Customers in the Control group who did not buy



```
> head(Int_data[(Int_data$Test_Variable=="control") & (Int_data$Outcome!=1), ])
    Customer_ID D_State D_Email Test_Variable Days_in_Between Int_T_Email Int_T_State Outcome
171         171       0       0       control             200           0           0       0
172         172       0       0       control             200           0           0       0
174         174       0       0       control             200           0           0       0
178         178       0       0       control             200           0           0       0
180         180       1       0       control             200           0           4       0
182         182       0       0       control             200           0           0       0
```

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.

Even after data was matched and de-duplicated in R and loaded into Excel, there were a few exceptions in both datasets.

Abandoned Dataset Exceptions' Customer ID's

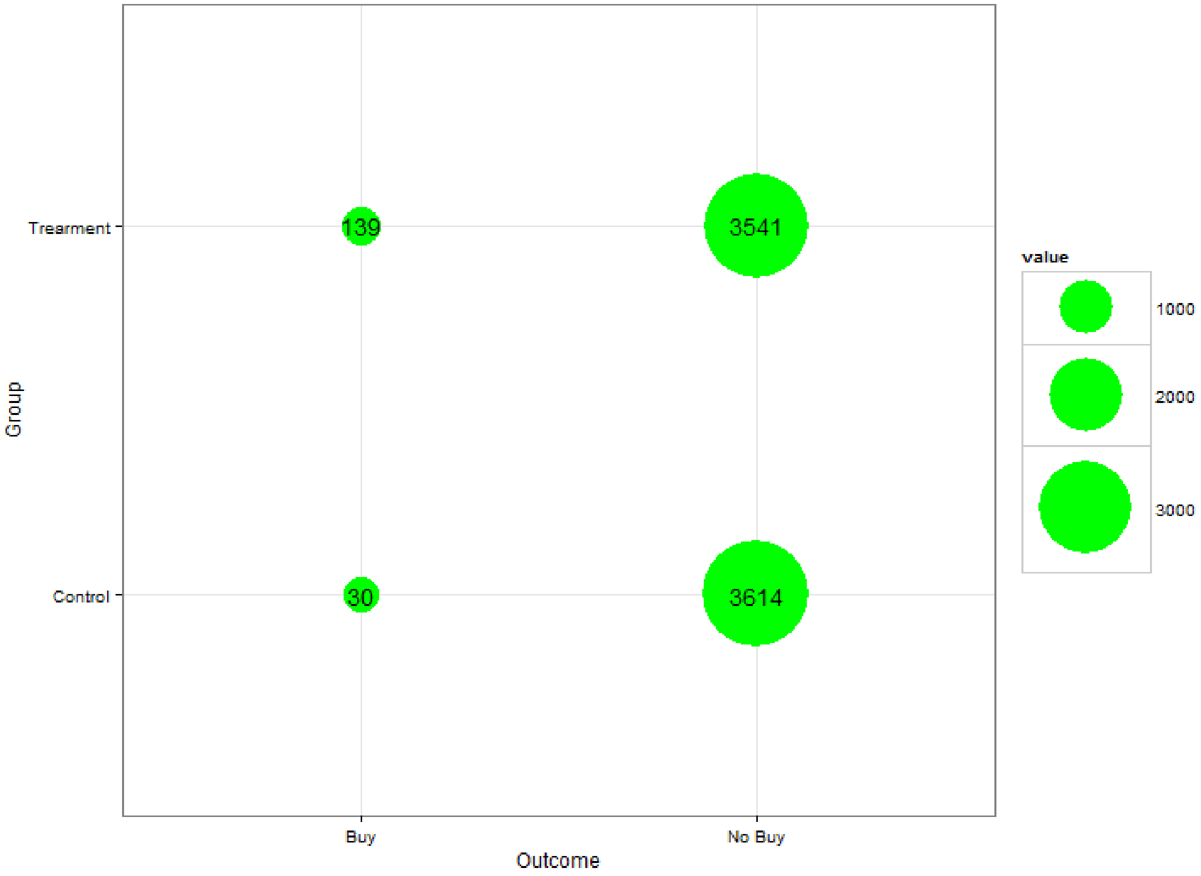
83670746PIZKTYAX
83009239LYSAJYVF
93089120PLIYTQGD
92413837GKMLBJGB
24311695KEJZXDAB
21933220BVVKJOYD

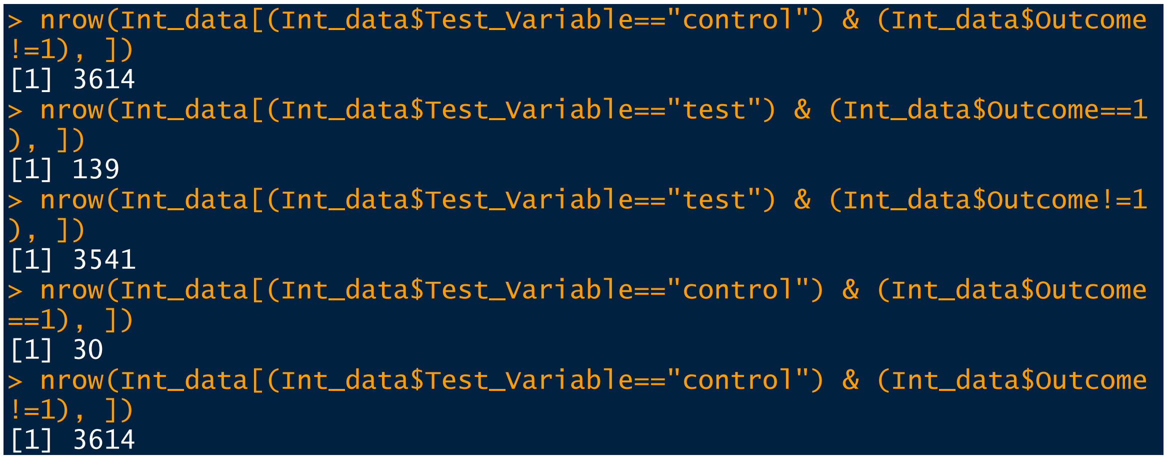
Reservation Dataset Exceptions' Customer ID's

51943963FYSKTXSE
66322605LMCXWYME
19552703COLNEOLT

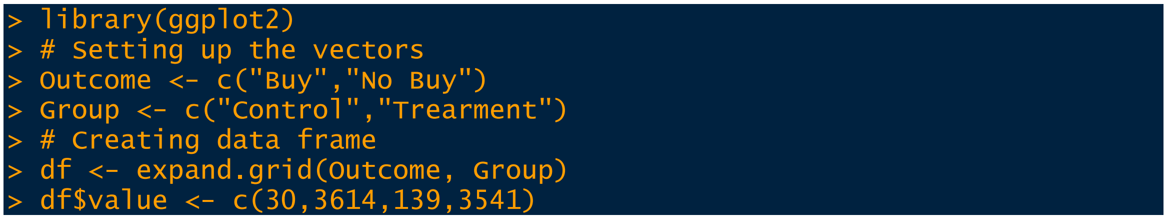
Q9: Complete the following cross-tabulation:

Group \ Outcome	Buy	No Buy
Treatment	139	3541
Control	30	3614





```
> nrow(Int_data[(Int_data$Test_Variable=="control") & (Int_data$Outcome
!=1), ])
[1] 3614
> nrow(Int_data[(Int_data$Test_Variable=="test") & (Int_data$Outcome==1
), ])
[1] 139
> nrow(Int_data[(Int_data$Test_Variable=="test") & (Int_data$Outcome!=1
), ])
[1] 3541
> nrow(Int_data[(Int_data$Test_Variable=="control") & (Int_data$Outcome
==1), ])
[1] 30
> nrow(Int_data[(Int_data$Test_Variable=="control") & (Int_data$Outcome
!=1), ])
[1] 3614
```



```
> library(ggplot2)
> # Setting up the vectors
> Outcome <- c("Buy","No Buy")
> Group <- c("Control","Trearment")
> # Creating data frame
> df <- expand.grid(Outcome, Group)
> df$value <- c(30,3614,139,3541)
> #Plotting Data
> g <- ggplot(df, aes(Var1, Var2)) + geom_point(aes(size = value), colo
ur = "green") + theme_bw() + xlab("Outcome") + ylab("Group")
> g + scale_size_continuous(range=c(10,30)) + geom_text(aes(label = val
ue))
```

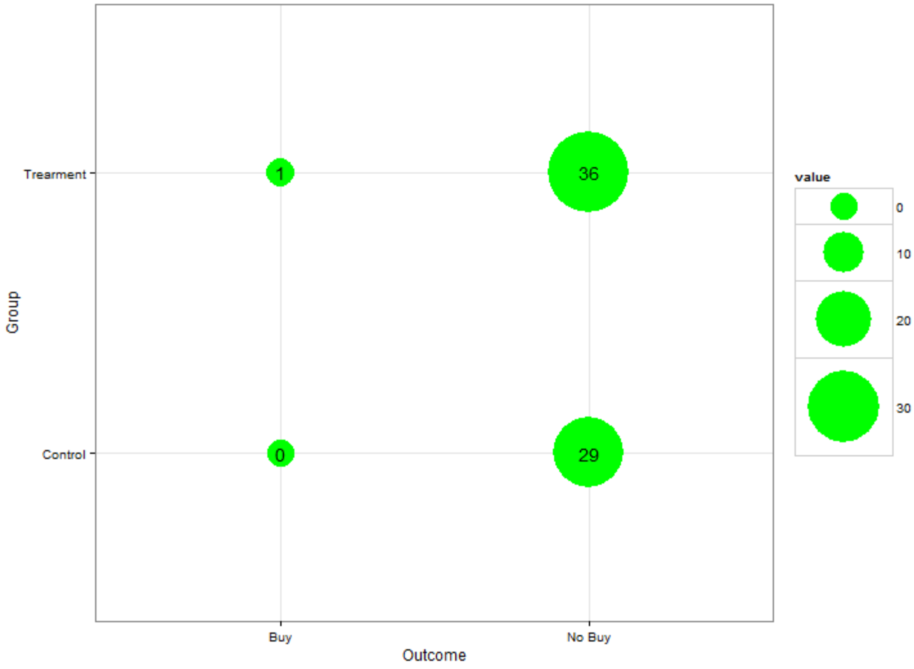
Q10: Repeat Q9 for 5 randomly picked states. Report 5 different tables by specifying the states you "randomly picked".

For this question, I used a previous version of my final dataset with had state names still present in it. Looking back at my documentation, both_match_aban_v5 in the Interaction Datasets folder would work for this questions purpose.

Using https://www.randomlists.com/random-us-states tool, the following states were picked:
Louisiana (LA)
North Dakota (ND)
Missouri (MO)
Wyoming (WY)
Oregon (OR)
Utah (UT)
Wisconsin (WI)

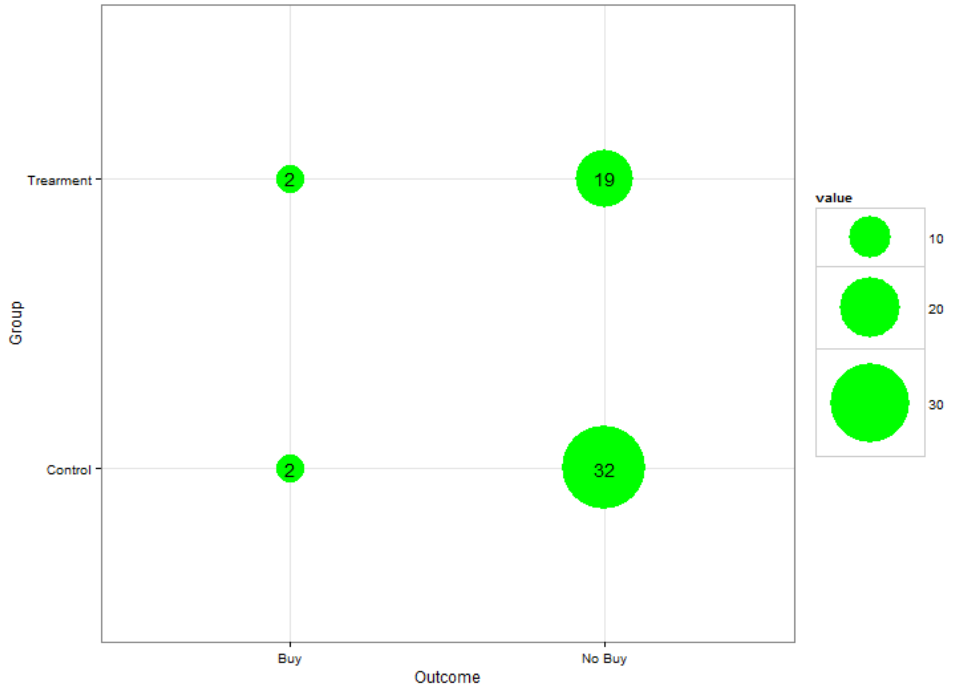
Louisiana (LA)

Group \ Outcome	**Buy**	**No Buy**
Treatment	**1**	**36**
Control	**0**	**29**



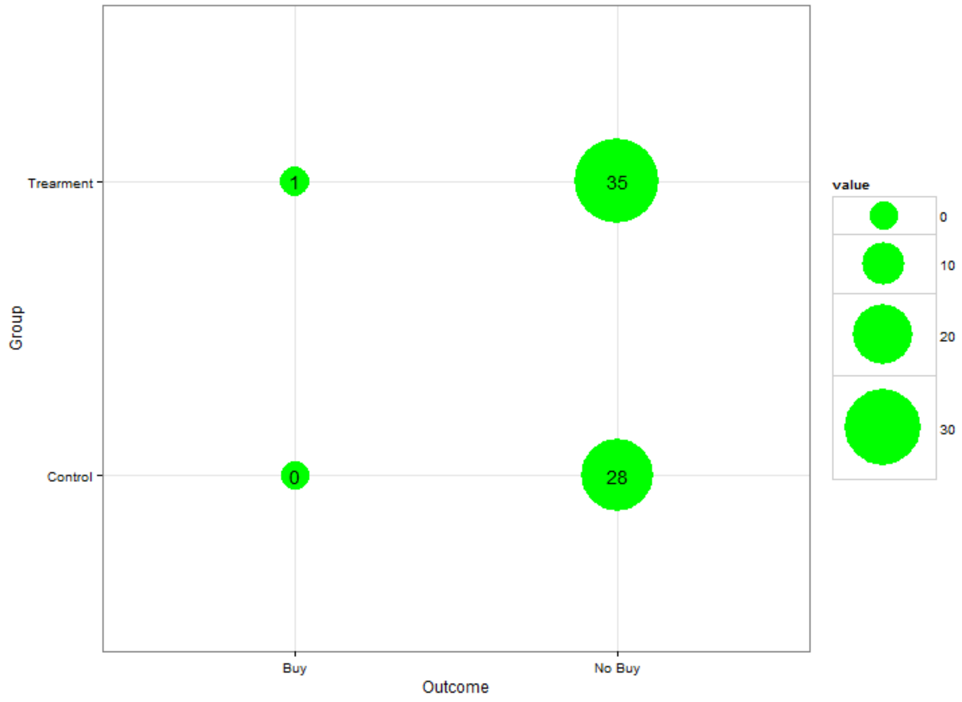
North Dakota (ND)

Group \ Outcome	Buy	No Buy
Treatment	2	19
Control	2	32



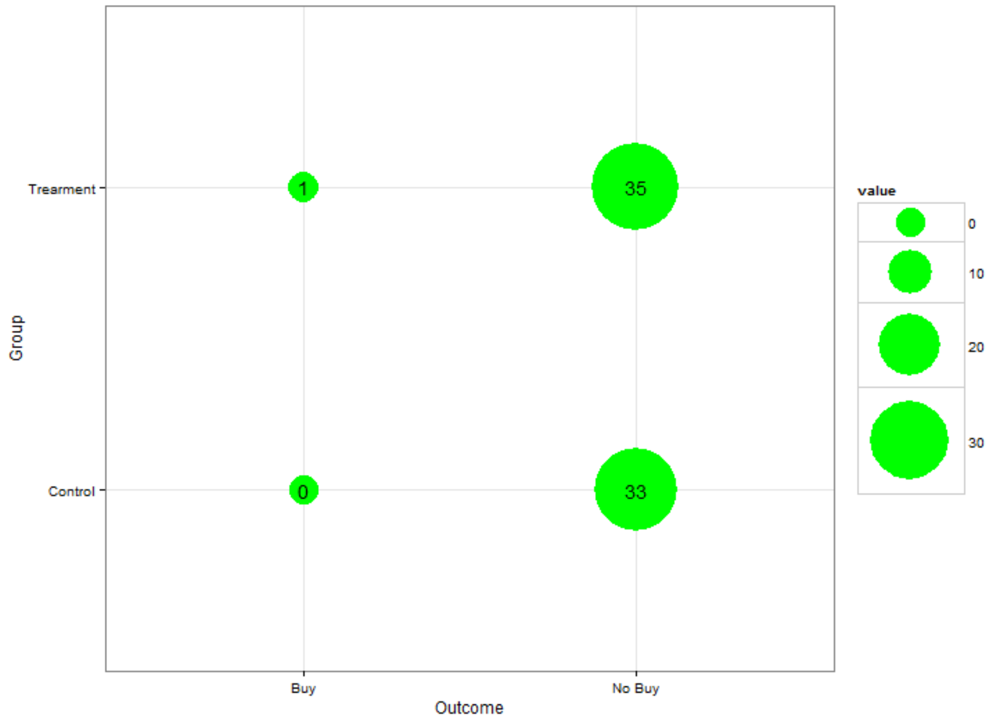
Missouri (MO)

Group \ Outcome	Buy	No Buy
Treatment	1	35
Control	0	28



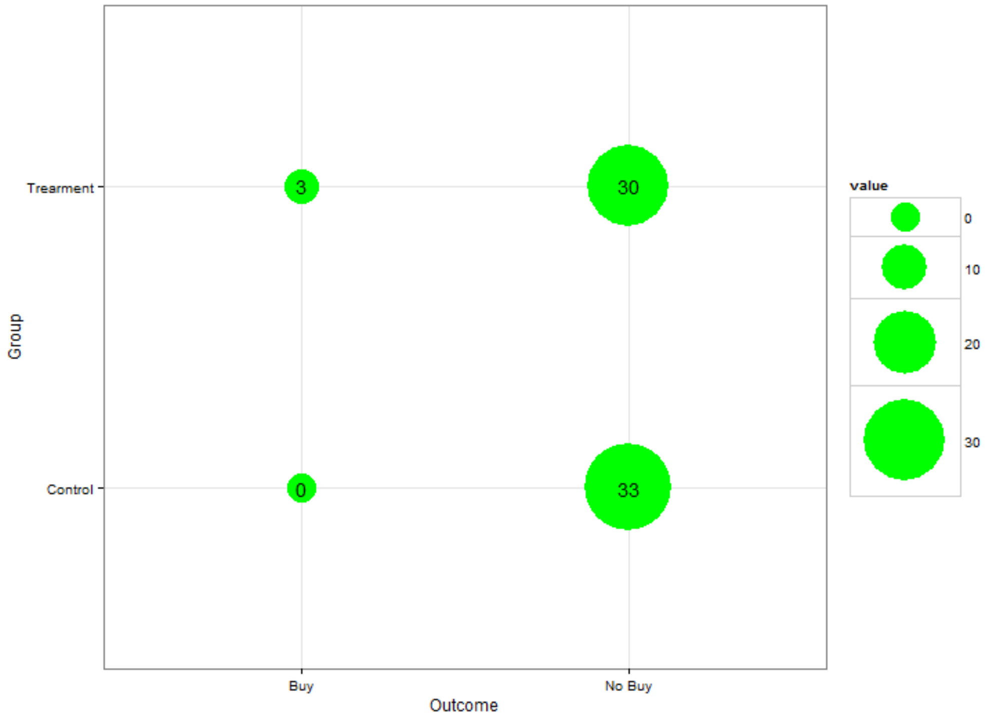
Wyoming (WY)

Group \ Outcome	Buy	No Buy
Treatment	1	35
Control	0	33



Oregon (OR)

Group \ Outcome	Buy	No Buy
Treatment	3	30
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 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.)

https://github.com/gauravjetley/MightyHive-Project/blob/master/Datasets%20for%20Interactions/both_matched_aban_v7.csv

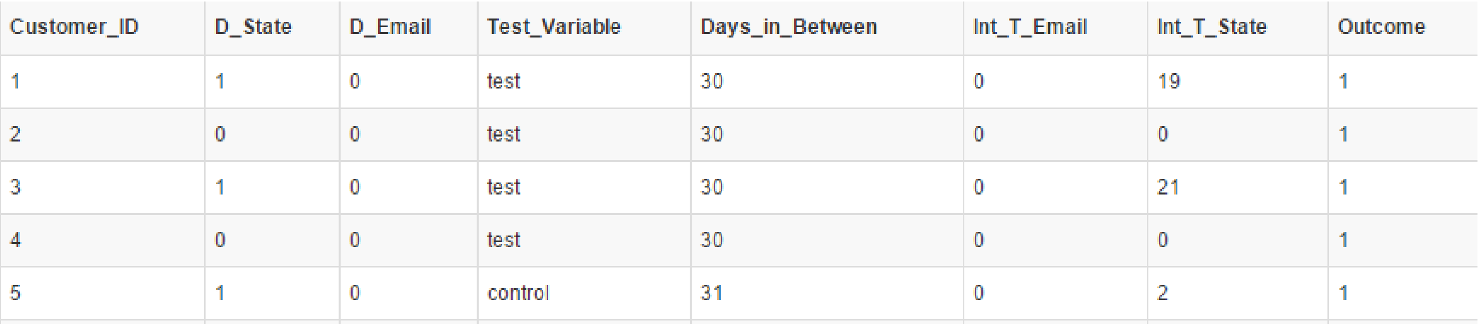
https://github.com/gauravjetley/MightyHive-Project/blob/master/Datasets%20for%20Interactions/both_matched_aban_v8.csv

Above are the links to the CSV files which contain the final cleaned datasets that I used.

The v7 dataset contains 2 interaction variables:

1) *Int_T_Email:*
    a. *The hour of day in which email address was provided*
    b. *0 is for no email provided*
2) *Int_T_State:*
    a. *The hour of day in which State was provided*
    b. *0 is for no State provided*

Subset of the both_matched_aban_v7 Dataset:

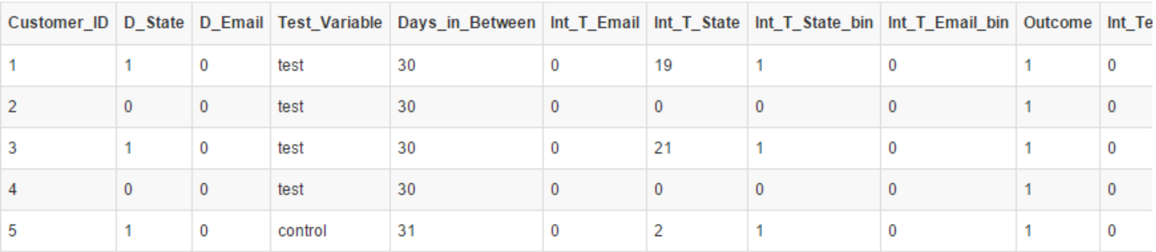


Customer_ID	D_State	D_Email	Test_Variable	Days_in_Between	Int_T_Email	Int_T_State	Outcome
1	1	0	test	30	0	19	1
2	0	0	test	30	0	0	1
3	1	0	test	30	0	21	1
4	0	0	test	30	0	0	1
5	1	0	control	31	0	2	1

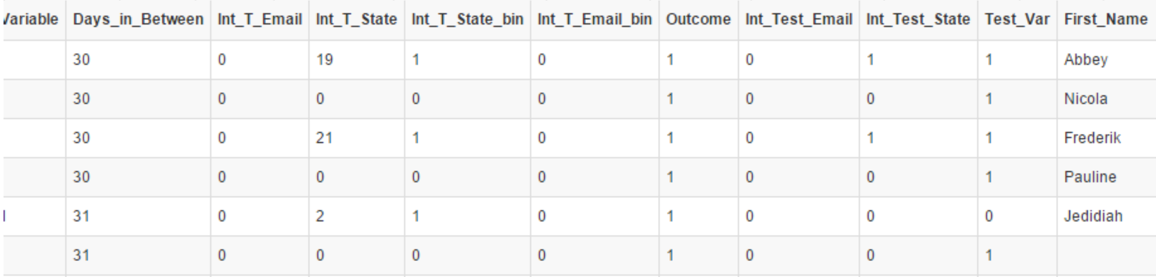
The v8 dataset contains 4 more interaction variables, 1 more binary and 1 categorical string variable:

1) *Int_Test_Email (Test*D_Email)*
2) *Int_Test_State(Test*D_State)*
3) *Test_Var (1: test, 0: control)*
4) *First_Name (For use with NLP to classify names as Male/Female)*
5) *Int_T_State_bin (TimeHour*D_State)(converted to binary)*
6) *Int_T_Email_bin (TimeHour*D_Email)(converted to binary)*

Subset of Dataset:



Customer_ID	D_State	D_Email	Test_Variable	Days_in_Between	Int_T_Email	Int_T_State	Int_T_State_bin	Int_T_Email_bin	Outcome	Int_Te
1	1	0	test	30	0	19	1	0	1	0
2	0	0	test	30	0	0	0	0	1	0
3	1	0	test	30	0	21	1	0	1	0
4	0	0	test	30	0	0	0	0	1	0
5	1	0	control	31	0	2	1	0	1	0



/ariable	Days_in_Between	Int_T_Email	Int_T_State	Int_T_State_bin	Int_T_Email_bin	Outcome	Int_Test_Email	Int_Test_State	Test_Var	First_Name
	30	0	19	1	0	1	0	1	1	Abbey
	30	0	0	0	0	1	0	0	1	Nicola
	30	0	21	1	0	1	0	1	1	Frederik
	30	0	0	0	0	1	0	0	1	Pauline
l	31	0	2	1	0	1	0	0	0	Jedidiah
	31	0	0	0	0	1	0	0	1	

IV. Statistical Analysis

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

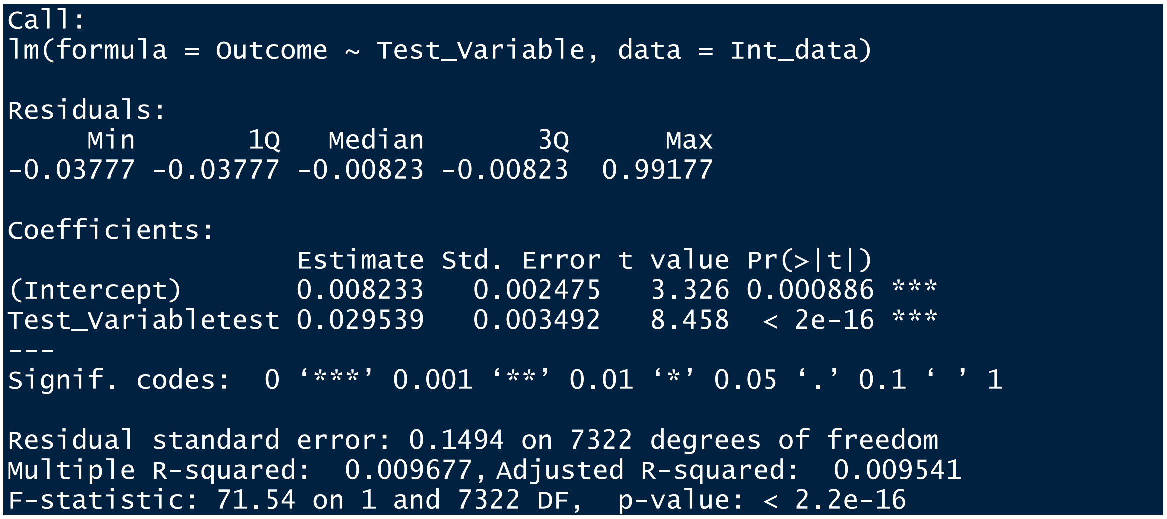
Q11: Run a Linear regression model for

Outcome = alpha + beta * Test_Variable + error

And Report the output.

Formula:

Outcome = β_0 + β1*Test_Variable + ε



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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.03777 -0.03777 -0.00823 -0.00823  0.99177 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.008233   0.002475   3.326 0.000886 ***
Test_Variabletest 0.029539   0.003492   8.458  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1494 on 7322 degrees of freedom
Multiple R-squared:  0.009677,	Adjusted R-squared:  0.009541 
F-statistic: 71.54 on 1 and 7322 DF,  p-value: < 2.2e-16
```

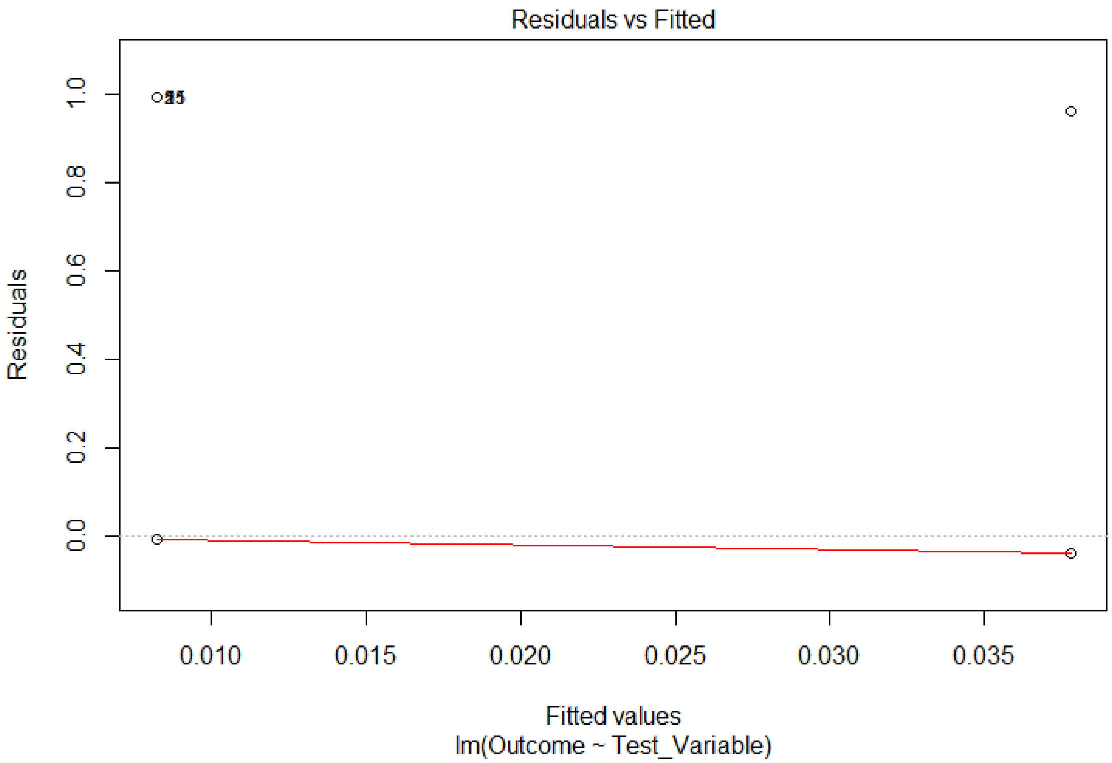
Outcome = 0.008233 + 0.029539*Test_Variable + ε

Dummy for Test_Variable:test has a coefficient of 0.029539 which is the increase in slope when the Test_variable is changed to the Treatment Group (test), keeping everything else constant.

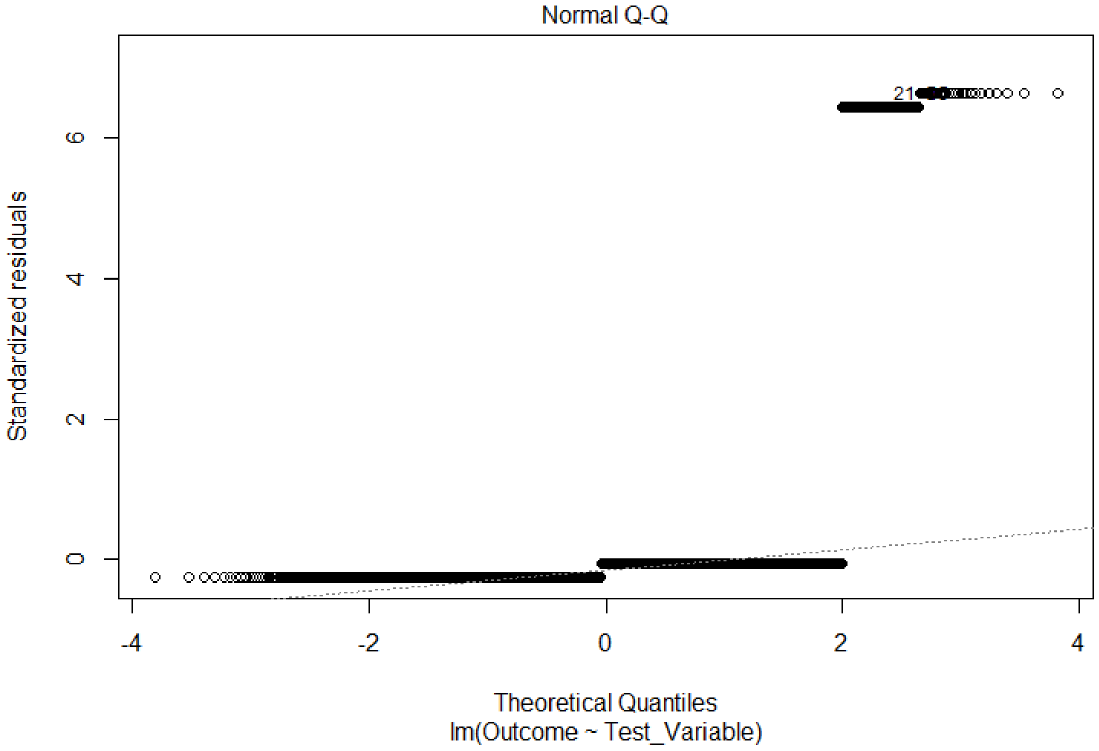
The coefficient for the Intercept also shows a significant p-value which suggests that the intercept by itself is also useful in predicting the value of Y (Outcome).

R^2 of 0.009677 states that the fitted model accounts for only 0.9451% of the variability present in the model. This is not a high number and shows that the model isn't very good in predicting the value of Y (Outcome). The p-Value of the F-Test is very significant tough, which suggests that the model is in-fact better at predicting the value of Y than the mean of Y. In any case, the strength of the relationship (R^2) is very low.

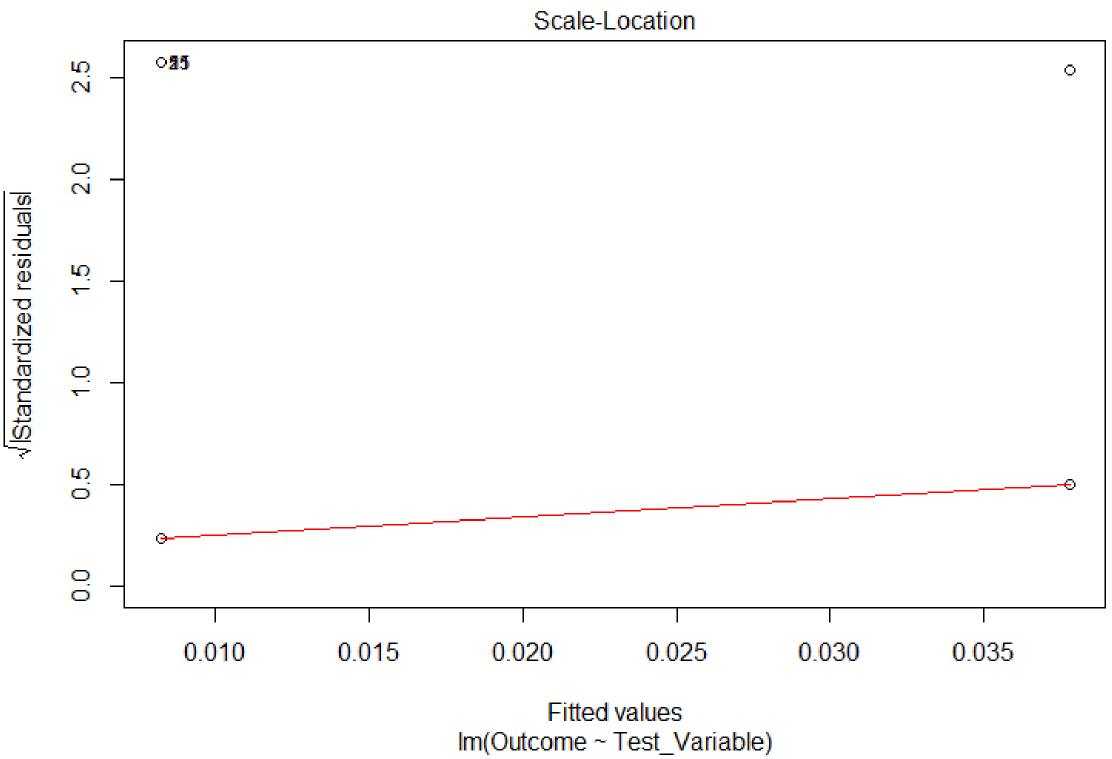
Checking for Assumptions and validity of the model:



There is a lack of pattern in the Residua vs Fitted Values plot but that will mostly be true for a Binary Regression Model. The residuals don't seem to have equal variance and has potential outliers marked. The fitted line is however very close to 0 and rather horizontal. Over all the plot shows that the residuals have are not centered around the mean of 0 with equal variance but the outliers shown in the plot are far less in number than the points around where Residuals = 0.



The QQ-Plot also shows that the residuals are rather not normally distributed with quite a few outliers. Outcome=1 are marked as outliers as there aren't many of them compared to all datapoints.



The Scale-Location plot also shows a slightly not horizontal fitted line which points towards heteroscedasticity in the data but the data gathering process suggests otherwise. This however is debatable.

Overall the model doesn't have a good fit. The residuals are not normal and the R^2 value is extremely small with suggest that the model does not have good predicting power. More variables should be added to the model to see if they increase the goodness of fit of the model or not.

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.

This experiment is statistically equivalent to A/B test procedure described in Leada Module 4.

Instead of showing two different versions of a websites' landing page, the customers are presented with 2 different scripts. One being a "control" script (default script) and one being the new "testing" script. The users which are given the control and test script is assigned randomly.

Then a test is carried out between the 2 groups to check if there is a significant difference between the 2 groups' results. The Regression Model we built is similar to an ANOVA or T-Test test we would have conducted in that module and checked for the validity of the test by checking the assumptions of the tests.

It's highly important to randomize the data properly and the assignment process to minimize the effect of variables we don't have control over, minimize bias and to achieve an overall better sample.

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?

This is not a very well fitted Regression Model as it fails to meet all the assumptions and has a very low goodness-of-fit R^2 value with fails to account for most of the variation in the model.

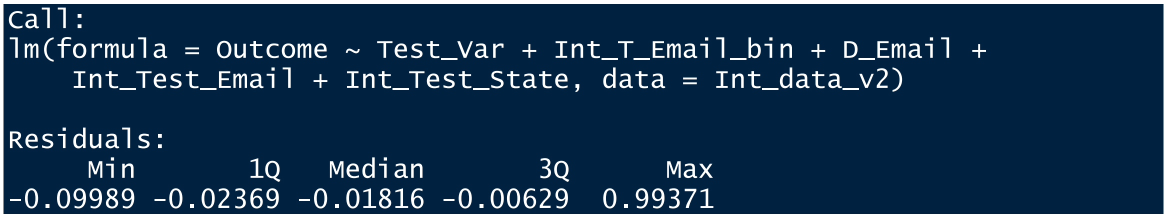
Because the Regression Model isn't very good, it won't make sense to make any Causal Statements about the effectiveness of the Retargeting Campaign. The results can be interpreted only as random covariance. There are many variables we haven't taken into account.

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)

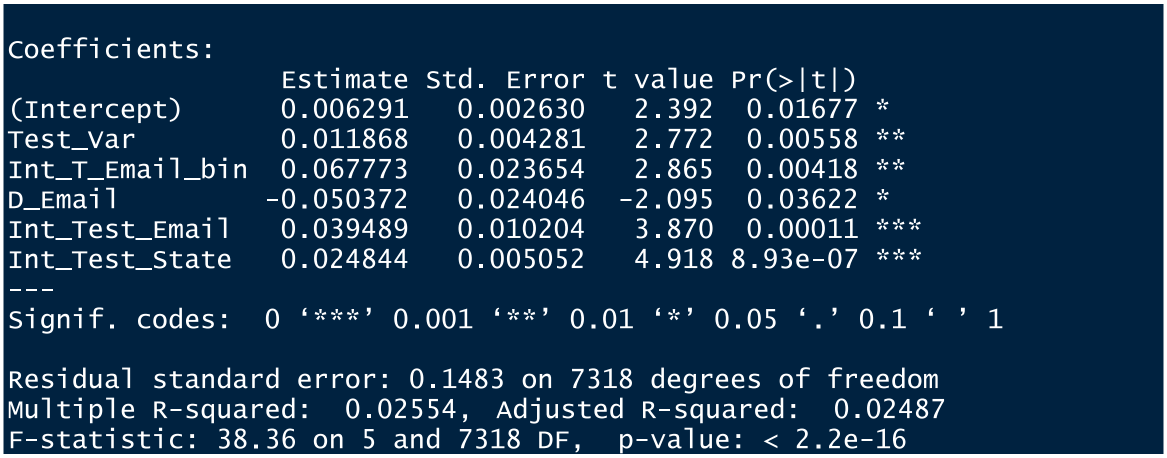
For this question, I used both_matched_aban_v8 dataset which also contains several extra interaction terms:

1) *Int_T_State_bin (TimeHour*D_State)(converted to binary)*
2) *Int_T_Email_bin (TimeHour*D_Email)(converted to binary)*
3) *Int_Test_Email (Test*D_Email)*
4) *Int_Test_State(Test*D_State)*
5) *Test_Var (1: test, 0: control)*

After a few attempts, the following model was chosen on the basis of highest Adjusted R^2 value and significant coefficients.



```
Call:
lm(formula = Outcome ~ Test_Var + Int_T_Email_bin + D_Email +
    Int_Test_Email + Int_Test_State, data = Int_data_v2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09989 -0.02369 -0.01816 -0.00629  0.99371
```



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.006291   0.002630   2.392  0.01677 *  
Test_Var          0.011868   0.004281   2.772  0.00558 ** 
Int_T_Email_bin   0.067773   0.023654   2.865  0.00418 ** 
D_Email          -0.050372   0.024046  -2.095  0.03622 *  
Int_Test_Email    0.039489   0.010204   3.870  0.00011 ***
Int_Test_State    0.024844   0.005052   4.918 8.93e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1483 on 7318 degrees of freedom
Multiple R-squared:  0.02554,	Adjusted R-squared:  0.02487 
F-statistic: 38.36 on 5 and 7318 DF,  p-value: < 2.2e-16
```

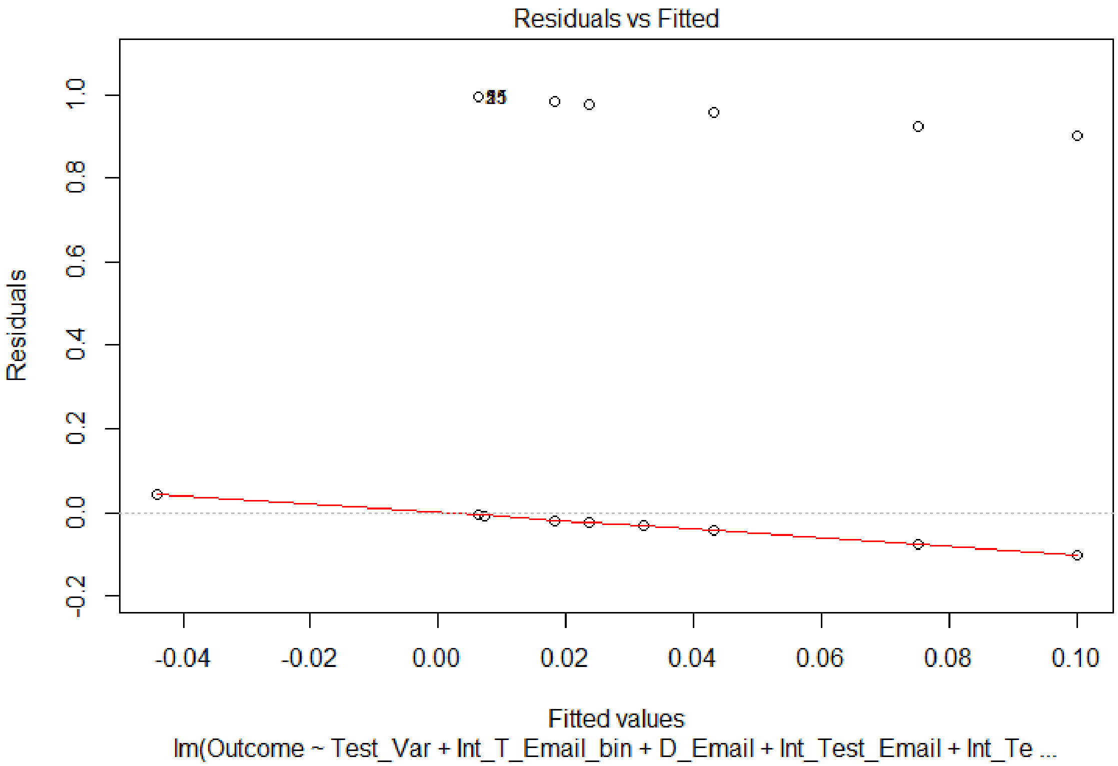
Equation:

> *Outcome = 0.006291 + (0.011868 * Test_Var) + (0.067773*TimeBinary*D_Email) + (-0.050372*D_Email) + (0.039489*Test_Var*D_Email) + (0.024844*Test_Var*D_State)*

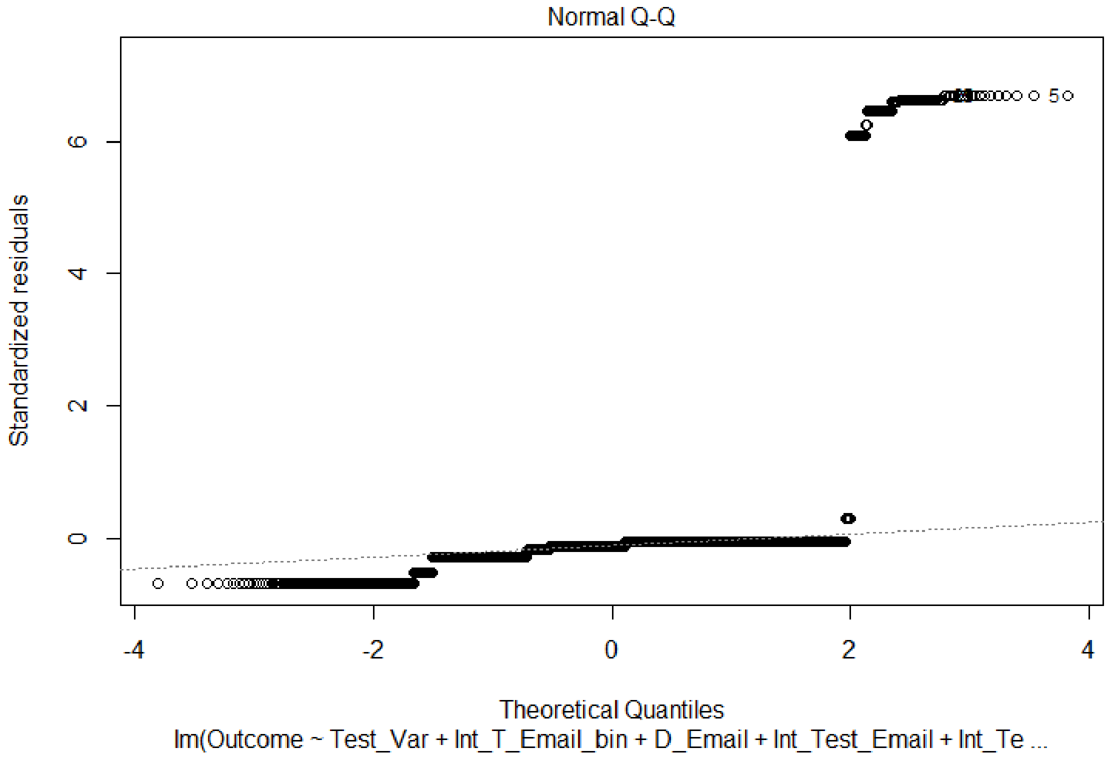
The F-Test of the model shows a very significant p-value which suggests that H0 should be rejected. In other words, it shows that the model is better at predicting the value of Y than just using the mean of Y. All of the p-Values of the individual coefficients are also very significant.

The Adjusted R^2 has increased to 2.487% but it's still not a very high number and is rather close to 0. But it's still an improvement to the model than just using Test_Variable.

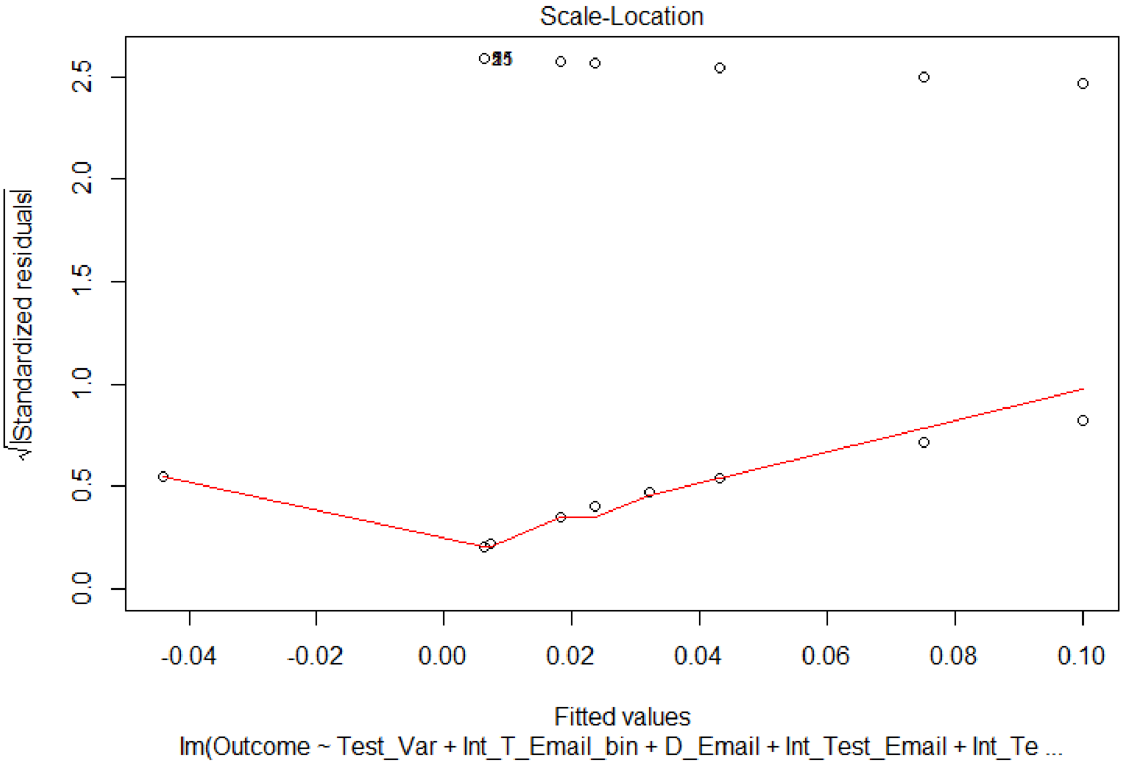
Checking for the validity of the model:



There seems to be a slight slanting pattern to the residuals but seem to be centered on 0. Some outliers are present like in the case of last model but generally the model looks better. The residuals which are around 1.0 are again not high in number when compared to the residuals which are around 0.0 thus they seem to have somewhat equal variance.



The QQ-Plot also shows the same pattern as in the last model.



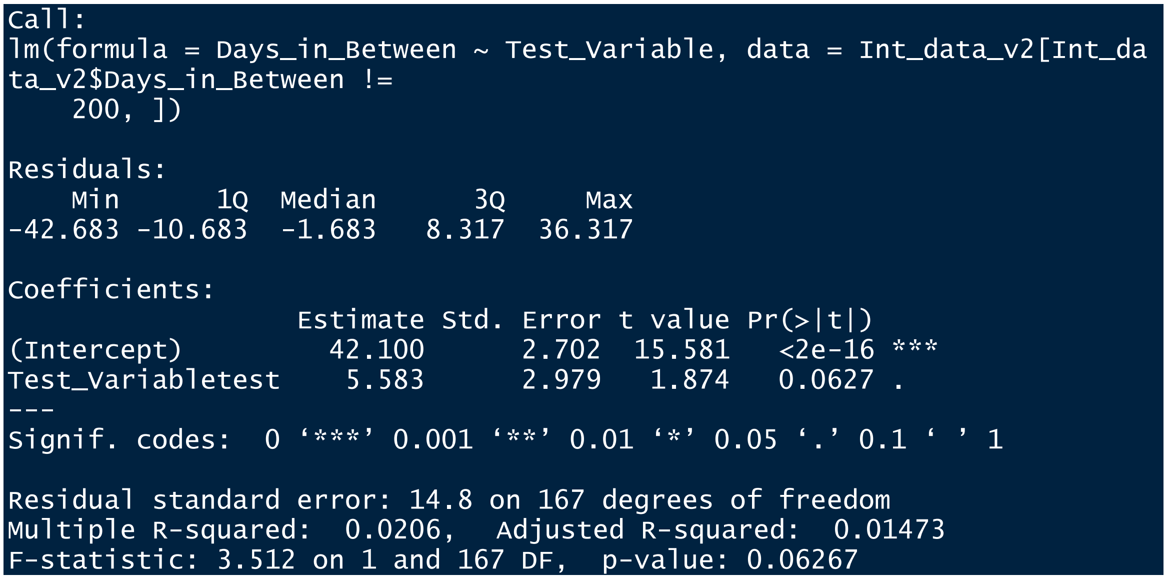
The Scale-Location plot looks better than the last models plot as the fitted line is somewhat equally plotted horizontally in both directions. With this we can say that the data is not heteroscedastic.

Overall, the model looks better than the previous model with a higher Adjusted R^2, Significant p-Value for the overall F-Test, Significant p-values for all coefficients, and better results when checking for the assumptions. But, the model is still not very good at predicting given the low Adjusted R^2. Some non-parametric classification models like Decision Trees, Support Vector Machines, Random Forests would be more useful in this situation and data.

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?



```
Call:
lm(formula = Days_in_Between ~ Test_Variable, data = Int_data_v2[Int_da
ta_v2$Days_in_Between !=
    200, ])

Residuals:
    Min      1Q  Median      3Q     Max 
-42.683 -10.683  -1.683   8.317  36.317 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         42.100      2.702  15.581   <2e-16 ***
Test_Variabletest    5.583      2.979   1.874   0.0627 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.8 on 167 degrees of freedom
Multiple R-squared:  0.0206,   Adjusted R-squared:  0.01473 
F-statistic: 3.512 on 1 and 167 DF,  p-value: 0.06267
```

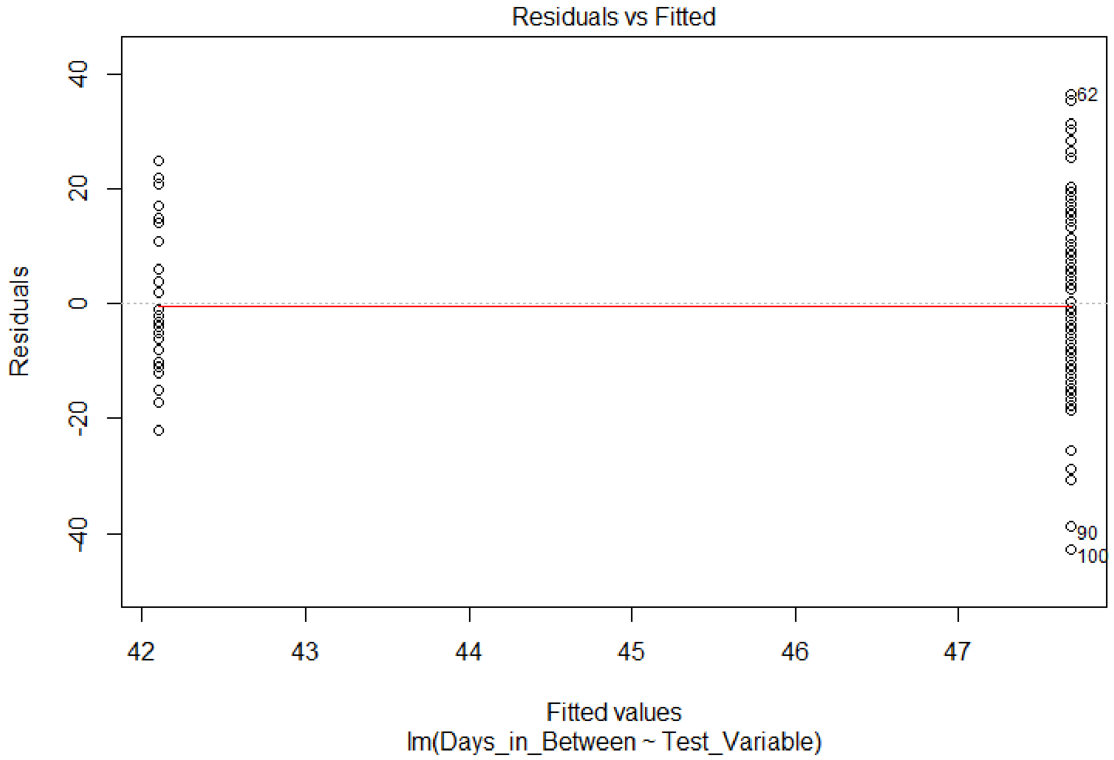
Equation:

*Days_in_Between = 42.100 + 5.583 * Test_Variable:Test + e*

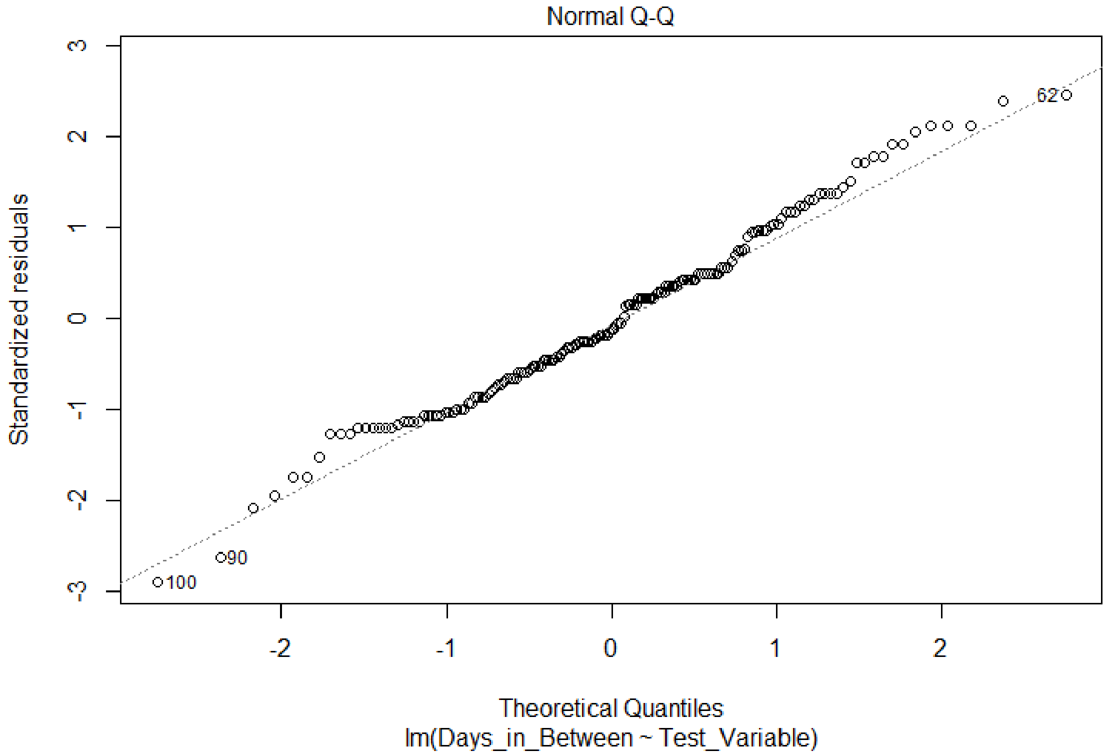
The overall F-Test on the model doesn’t have a great F-Value and corresponding p-Value is only significant at an Alpha of 0.1. The R^2 of the model is at 2% which is again low. The coefficient of the dummy for “Test_Variable: Test” shows only a significant p-Value at an alpha of 0.1 but not significant at an alpha of 0.05.

This shows that the response time is not influenced by the retargeting campaign when we keep an Alpha of 0.05.

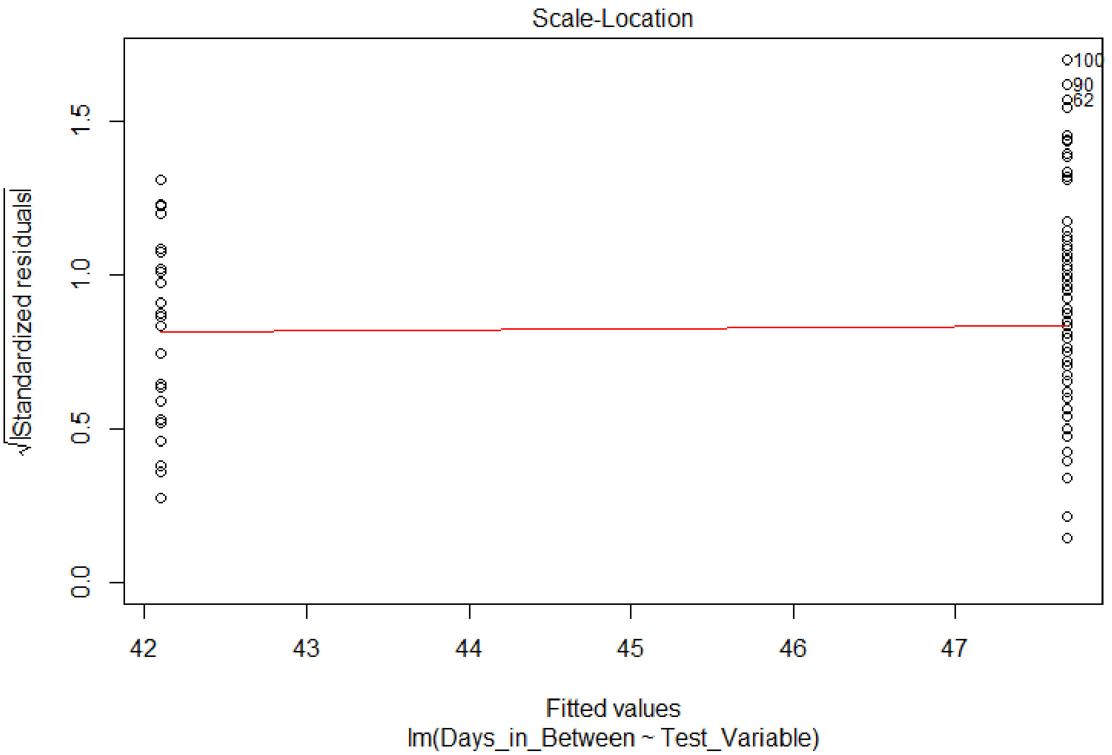
The coefficients should not be interpreted as causal in this case. The model seems too parsimonious and more variables should be added to see if they have any positive influence on the model.



The residuals are centered around mean of 0 but are not completely equally varied with an apparent conical pattern.



The QQ-Plot shows that the residuals are not completely normally distributed but are rather close to it.



Scale-Location plot just like the residual vs fitted value plot shows a conical pattern with the points on left side having more spread than on the right side. This challenges the assumption of homoscedasticity.

All datasets, R code and other files that I created/used related to this project can be found at https://github.com/gauravjetley/MightyHive-Project

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

The design of the experiment has a lot of issues regarding data that was captured. More variables should have been present (Male/Female, Age group, Education Level, Employment etc.) to make better inferences. There should have also been a mechanism to have a better primary key that would be per customer and not per call.

With better data, many new analysis could have been done like:

1) *Is the test script more effective to a specific gender and does the age group have any effect?*
2) *Does the education of a person has any relationship with buying the product?*

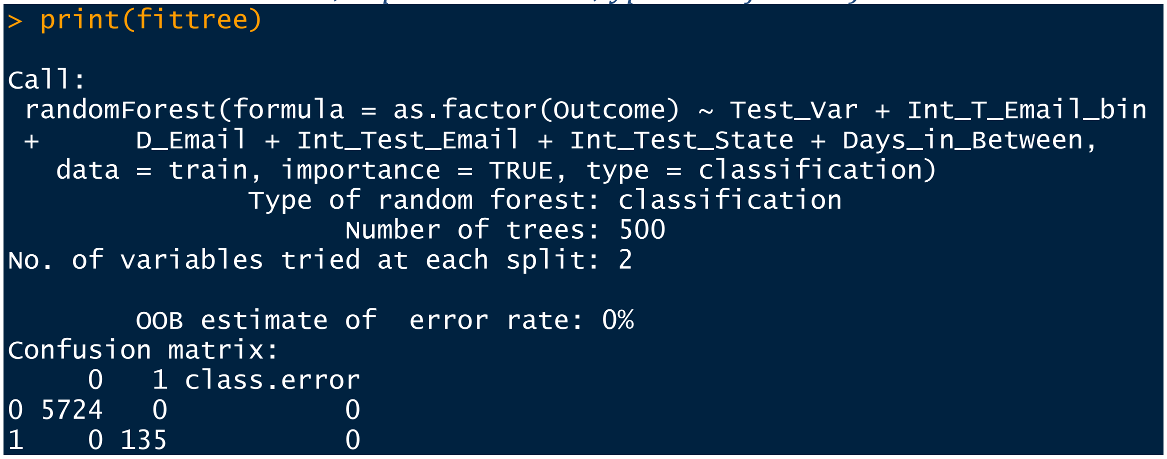
A better and more robust data capturing, validating and warehousing strategy could have improved the results of the analysis greatly.

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: *90*

```
########################################
#########      Using Random Forest      ############
########################################
library(randomForest)
#splitting dataset into 80:20
train <- Int_data_v2[c(1:135,170:5893),]
test <- Int_data_v2[c(136:169,5894:7324),]
fittree <- randomForest(as.factor(Outcome)~Test_Var+Int_T_Email_bin+
                   D_Email+Int_Test_Email+Int_Test_State+Days_in_Between,
                  data = train, importance=TRUE,type=classification)
```



```
> print(fittree)

Call:
 randomForest(formula = as.factor(Outcome) ~ Test_Var + Int_T_Email_bin
 +      D_Email + Int_Test_Email + Int_Test_State + Days_in_Between,
   data = train, importance = TRUE, type = classification)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0%
Confusion matrix:
     0   1 class.error
0 5724   0           0
1    0 135           0
```

```
test[,15] <- predict(fittree,newdata = test)
```

# 100% success rate at predicting "Outcome"

# NLP can be used with the "NAME" corpa to classify names as Male/Female