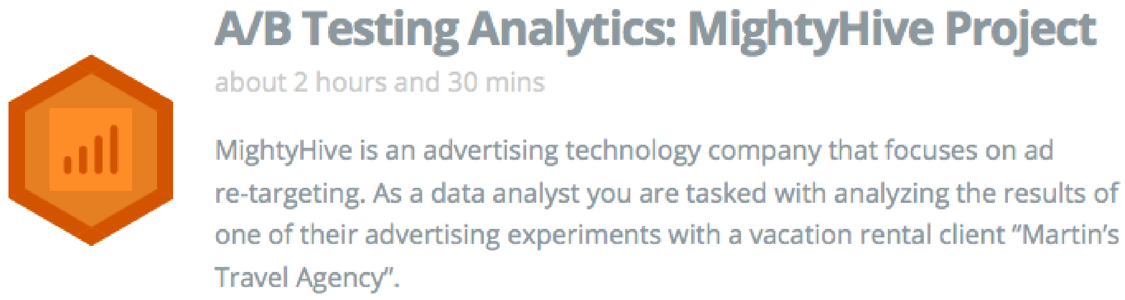
TEAM LEADA PROJECT

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"



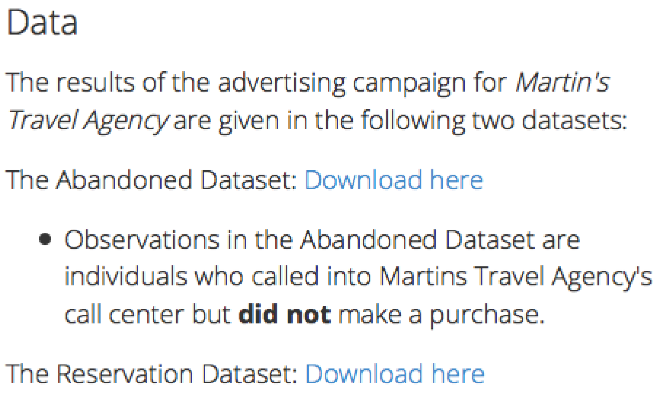
A/B Testing Analytics: MightyHive Project

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: **SACHIN KANT MISRA**
Section: **EVENING**
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.
Answer: Yes, in my opinion this is a very sensible and genuine question in business terms. Since the number of customers who were advertised last year is huge and out of them, a significant amount of people (20,814) did make a reservation and another (8442) have been treated (advertised) last year and so there is a high chance that people listening to the offer again over the call may opt for it this time. There may be numerous reasons for these (8442) set of people to abandon last year's call viz-a-viz most of these targeted people might have been farmers and due to drought last year, they weren't in a good state of finance to go for the offer and henceforth. Like this scenario, there may be good number of people who were attracted to the offer but failed to make an reservation and could possibly benefit from it this year.

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.

Answer:

```
> clean_dataset <- read.csv("MightyHive Cleaned Data.csv", header = T, stringsAsFactors = FALSE)
> clean_dataset$Test_Variable <- ifelse(clean_dataset$Test_Variable=="test",1,0)
> summary(clean_dataset$Test_Variable)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  1.0000  0.5053  1.0000  1.0000 
> sd(clean_dataset$Test_Variable)
[1] 0.5000012
> mean(clean_dataset$Test_Variable) - 1.96*sd(clean_dataset$Test_Variable)/sqrt(nrow(clean_datase
t))
[1] 0.4946644
> mean(clean_dataset$Test_Variable) + 1.96*sd(clean_dataset$Test_Variable)/sqrt(nrow(clean_datase
t))
[1] 0.5159966
```

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:

```
> clean_dataset$Address <- ifelse(clean_dataset$Address!="",1,0)
> clean_dataset_states <- clean_dataset[which(clean_dataset$Address=='1'),]
> summary(clean_dataset_states$Test_Variable)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  1.0000  0.5134  1.0000  1.0000 
> sd(clean_dataset_states$Test_Variable)
[1] 0.4998865
> mean(clean_dataset_states$Test_Variable) - 1.96*sd(clean_dataset_states$Test_Variable)/sqrt(nro
w(clean_dataset_states))
[1] 0.4975097
> mean(clean_dataset_states$Test_Variable) + 1.96*sd(clean_dataset_states$Test_Variable)/sqrt(nro
w(clean_dataset_states))
[1] 0.5292479
```

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: As we can see clearly from the facts in Answer 3 and 4, there is no experimental deviation in the statistical summary on Test_Variable for our complete cleaned dataset and for State only filtered values. Hence, yes, the experiment appears to be executed properly.
Also, below statistical evidence showcase that there is almost no imbalance in the assignments to treatment and control when switching to the State-only level.The only difference that can be seen is the slight statistical difference in the data for Days_in_between and Test_Variable. And hence can be ignored.

```
> clean_dataset <- read.csv("MightyHive Cleaned Data.csv", header = T, stringsAsFactors = FALSE)
> clean_dataset$Test_Variable <- ifelse(clean_dataset$Test_Variable=="test",1,0)
> clean_dataset$Address <- ifelse(clean_dataset$Address!="",1,0)
> clean_dataset_states <- clean_dataset[which(clean_dataset$Address=='1'),]
> clean_dataset_test_states <- clean_dataset_states[which(clean_dataset_states$Test_Variable=='1'),]
> clean_dataset_control_states <- clean_dataset_states[which(clean_dataset_states$Test_Variable=='0')
,]
> summary(clean_dataset_test_states)
  Caller_ID         Test_Variable   Outcome          Days_in_Between    Address      Email          
 Length:1957        Min.   :1     Length:1957        Min.   :  7.0   Min.   :1   Length:1957       
 Class :character   1st Qu.:1     Class :character   1st Qu.:200.0   1st Qu.:1   Class :character  
 Mode  :character   Median :1     Mode  :character   Median :200.0   Median :1   Mode  :character  
                    Mean   :1                        Mean   :186.5   Mean   :1                     
                    3rd Qu.:1                        3rd Qu.:200.0   3rd Qu.:1                     
                    Max.   :1                        Max.   :200.0   Max.   :1                     
> summary(clean_dataset_control_states)
  Caller_ID         Test_Variable   Outcome          Days_in_Between    Address      Email          
 Length:1855        Min.   :0     Length:1855        Min.   : 11     Min.   :1   Length:1855       
 Class :character   1st Qu.:0     Class :character   1st Qu.:200     1st Qu.:1   Class :character  
 Mode  :character   Median :0     Mode  :character   Median :200     Median :1   Mode  :character  
                    Mean   :0                        Mean   :196     Mean   :1                     
                    3rd Qu.:0                        3rd Qu.:200     3rd Qu.:1                     
                    Max.   :0                        Max.   :200     Max.   :1                     
```

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: As we can clearly read from the fact above that the campaign went over for three months. Thus by looking at the abandon sheet and promptly concluding that “We do not observe some “outcomes” for some customers” would be wrong. This is because there may be scenarios wherein the customer rejected the vacation package offered for the first time due to numerous personal or circumstantial reasons but later on he called back and made the reservation. However, this is also proved by the data cleaning activity wherein we got a match of around **385 customers from abandon dataset with reservation dataset on the unique phone number key as matching index** and around **75 customers matched on unique email id key as matching index**. So, the fact that there are no potential outcomes for some customers can also be attributed to so many duplicate values on Phone and Email as indexing field.

```
> abandon_data<-read.csv("Abandoned_Data_Seed.csv", header=T, stringsAsFactors = F)
> reservation_data<-read.csv("Reservation_Data_Seed.csv", header=T, stringsAsFactors = F)
> abandon_data$Empty_IncomingPhone <- ifelse(abandon_data$Incoming_Phone=="",1,0)
> table(abandon_data$Empty_IncomingPhone)

   0    1 
7262 1180 
> abandon_data$Phone <- ifelse(abandon_data$Incoming_Phone=="", abandon_data$Contact_Phone, abandon_da
ta$Incoming_Phone)
> abandon_data$Empty_Phone <- ifelse(abandon_data$Phone=="",1,0)
> table(abandon_data$Empty_Phone)

   0 
8442 
> abandon_data_IP_DUP_ENTRY <- duplicated(abandon_data$Incoming_Phone)
> summary(abandon_data_IP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical    7166    1276       0 
> abandon_data_CP_DUP_ENTRY <- duplicated(abandon_data$Contact_Phone)
> summary(abandon_data_CP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical    8272     170       0 
> test_abandon_data=abandon_data[abandon_data$Test_Control=="test",]
> reservation_data$Empty_Incoming_Phone <- ifelse(reservation_data$Incoming_Phone=="",1,0)
> table(reservation_data$Empty_Incoming_Phone)

    0     1 
19406  1408 
> reservation_data$Phone <- ifelse(reservation_data$Incoming_Phone=="", reservation_data$Contact_Phone
, reservation_data$Incoming_Phone)
> reservation_data_IP_DUP_ENTRY <- duplicated(reservation_data$Incoming_Phone)
> summary(reservation_data_IP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical   18282    2532       0 
> reservation_data_CP_DUP_ENTRY <- duplicated(reservation_data$Contact_Phone)
> summary(reservation_data_CP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical    9713   11101       0 
> abandon_reservation_phone_matched <- abandon_data$Phone %in% reservation_data$Phone
> summary(abandon_reservation_phone_matched)
   Mode   FALSE    TRUE    NA's 
logical    8057     385       0 
> abandon_data$test_email <- ifelse(abandon_data$Email=="", "BLANK", abandon_data$Email)
> abandon_reservation_email_matched <- abandon_data$test_email %in% reservation_data$Email
> summary(abandon_reservation_email_matched)
   Mode   FALSE    TRUE    NA's 
logical    8367      75       0 
```

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 looking at the columns of both files, and the reasoning that the customer who might have abandoned the offer on the first go could have taken the offer later within the three month period of time this experiment extended, I have found below script to match the two sheets for such customers.

For such index matching, we need to make sure there are no Empty or Blank or Space value in any column, so we filled the respected columns. For eg. I have created a new column named "Phone" which takes value from Incoming_Phone column from the sheet and in case Incoming_Phone is Empty or Spaces, it takes value from Contact_Phone column of the same sheet. After doing this, the sheet is filtered for unique "Phone" Column added in the sheet after removing duplicates as that means the reservation was done by same person who was contacted earlier and it might just that be he called back from some other number like office or friend's number to make the reservation. The same technique is applied on the Email column which is also used as secondary match index after Phone to match the two sheets.

For the data cleaning activity wherein I got a match of around **385 customers from abandon dataset with reservation dataset on the unique phone number key as matching index** and around **75 customers matched on unique email id key as matching index**.

So, Finally I have used **two column keys** as matching index-

(1) Phone **(2) Email**

Below is the script and output:

```
> abandon_data<-read.csv("Abandoned_Data_Seed.csv", header=T, stringsAsFactors = F)
> reservation_data<-read.csv("Reservation_Data_Seed.csv", header=T, stringsAsFactors = F)
> abandon_data$Empty_IncomingPhone <- ifelse(abandon_data$Incoming_Phone=="",1,0)
> table(abandon_data$Empty_IncomingPhone)

   0    1 
7262 1180 
> abandon_data$Phone <- ifelse(abandon_data$Incoming_Phone=="", abandon_data$Contact_Phone, abandon_d
ata$Incoming_Phone)
> abandon_data$Empty_Phone <- ifelse(abandon_data$Phone=="",1,0)
> table(abandon_data$Empty_Phone)

   0 
8442 
> abandon_data_IP_DUP_ENTRY <- duplicated(abandon_data$Incoming_Phone)
> summary(abandon_data_IP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical    7166    1276       0 
> abandon_data_CP_DUP_ENTRY <- duplicated(abandon_data$Contact_Phone)
> summary(abandon_data_CP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical    8272     170       0 
> test_abandon_data=abandon_data[abandon_data$Test_Control=="test",]
> reservation_data$Empty_Incoming_Phone <- ifelse(reservation_data$Incoming_Phone=="",1,0)
> table(reservation_data$Empty_Incoming_Phone)

    0     1 
19406  1408 
> reservation_data$Phone <- ifelse(reservation_data$Incoming_Phone=="", reservation_data$Contact_Phon
e, reservation_data$Incoming_Phone)
> reservation_data_IP_DUP_ENTRY <- duplicated(reservation_data$Incoming_Phone)
> summary(reservation_data_IP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical   18282    2532       0 
> reservation_data_CP_DUP_ENTRY <- duplicated(reservation_data$Contact_Phone)
> summary(reservation_data_CP_DUP_ENTRY)
   Mode   FALSE    TRUE    NA's 
logical    9713   11101       0 
> abandon_reservation_phone_matched <- abandon_data$Phone %in% reservation_data$Phone
> summary(abandon_reservation_phone_matched)
   Mode   FALSE    TRUE    NA's 
logical    8057     385       0 
> abandon_data$test_email <- ifelse(abandon_data$Email=="", "BLANK", abandon_data$Email)
> abandon_reservation_email_matched <- abandon_data$test_email %in% reservation_data$Email
> summary(abandon_reservation_email_matched)
   Mode   FALSE    TRUE    NA's 
logical    8367      75       0 
```

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: The below script elaborates complete steps followed for data cleaning activity:

#Details of filteration
#Two Datasets, Abandoned and Reservation, Matched on the basis of Email And Phone (Either (1)Incoming or (2)Contact)
#Duplicated Emails,Phones removed
#Dates split into date and time (Done Through MS-EXCEL)

```
abandon_data<-read.csv("Abandoned_Data_Seed.csv", header=T, stringsAsFactors = F)
reservation_data<-read.csv("Reservation_Data_Seed.csv", header=T, stringsAsFactors = F)

abandon_data$Empty_IncomingPhone <- ifelse(abandon_data$Incoming_Phone=="",1,0)
table(abandon_data$Empty_IncomingPhone)
# This Table has 1180 Missing Incoming phone number in abandon dataset

abandon_data$Phone <- ifelse(abandon_data$Incoming_Phone=="", abandon_data$Contact_Phone, abandon_data$Incoming_Phone)
#Created a new field "Phone" , by Copying Contact Number from Incoming Phone fields or Contact Phone in case Incoming is Empty.

abandon_data$Empty_Phone <- ifelse(abandon_data$Phone=="",1,0)
table(abandon_data$Empty_Phone)
#This Table shows there is no empty field in Abandon dataset for PHONE Attribute created above.
#Data contained is either Incoming_Phone or Contact_Phone

abandon_data_IP_DUP_ENTRY <- duplicated(abandon_data$Incoming_Phone)
summary(abandon_data_IP_DUP_ENTRY)
#Removing duplicates on Incoming Phone field, found 1276 Phones are duplicate in Incoming Phone field.

abandon_data_CP_DUP_ENTRY <- duplicated(abandon_data$Contact_Phone)
summary(abandon_data_CP_DUP_ENTRY)
#Removing duplicates on Contact Phone field, found 170 Phone are duplicate in Contact Phone field.

#Filtered data from abandon dataset where Test_Control field is test
test_abandon_data=abandon_data[abandon_data$Test_Control=="test",]


reservation_data$Empty_Incoming_Phone <- ifelse(reservation_data$Incoming_Phone=="",1,0)
table(reservation_data$Empty_Incoming_Phone)
#Checking for Empty Incoming Phone field doesn't have data. #1408 Rows found.

reservation_data$Phone <- ifelse(reservation_data$Incoming_Phone=="", reservation_data$Contact_Phone, reservation_data$Incoming_Phone)
reservation_data_IP_DUP_ENTRY <- duplicated(reservation_data$Incoming_Phone)
summary(reservation_data_IP_DUP_ENTRY)
#Checking for duplicate Incoming Phone entry.#2532 duplicated Incoming_Phone exists in reservation dataset

reservation_data_CP_DUP_ENTRY <- duplicated(reservation_data$Contact_Phone)
summary(reservation_data_CP_DUP_ENTRY)
#Checking for duplicate Contact Phone entry.#11101 duplicate Contact_Phone exists

abandon_reservation_phone_matched <- abandon_data$Phone %in% reservation_data$Phone
summary(abandon_reservation_phone_matched)
#385 Common enteries are matched between two datasets with Created Attribute "Phone" as parameter

abandon_data$test_email <- ifelse(abandon_data$Email=="", "BLANK", abandon_data$Email)
#Empty Email fields are filled with BLANK text and rest of the rows are kept as it is for matching the datasets on Email.

abandon_reservation_email_matched <- abandon_data$test_email %in% reservation_data$Email
summary(abandon_reservation_email_matched)
#75 Email address matched from test_abandon_data dataset and reservation_data.
#These are the people who bought the product after abandoning it for the first time.

abandon_data$Purchase_Status_basedon_Email <- abandon_data$test_email %in% reservation_data$Email
summary(abandon_data$Purchase_Status_basedon_Email)
#Now Dataset test_abandon_data contains all those who purchased the product based on the email id match condition.
```

```
#Further on, we compare the two dataset on the basis of Phones
abandon_data$Purchase_Status_basedon_Phone <- abandon_data$Phone %in% reservation_data$Phone
summary(abandon_data$Purchase_Status_basedon_Phone)
#385 Phone numbers matched which means these are also the people who bought the product later on.

#So in All we found 385(Email)+75(Phone) rows in our test_abandon_data whom we can say 460 people purchased the product after they had abandoned .
#There might be duplicate entries on Email & Phone, which we'll remove next.

abandon_data$Purchase_Status <- ifelse(abandon_data$Purchase_Status_basedon_Email=="FALSE" & abandon_data$Purchase_Status_basedon_Phone=="FALSE",FALSE,TRUE)
summary(abandon_data$Purchase_Status)
#If either of Email or Phone matches, it signifies customer has bought the product.
#399 customer has bought the product.

for (i in 1:nrow(abandon_data)){
  if ((abandon_data$Purchase_Status[i])=="TRUE"){
    abandon_data$Purchase_Status[i]<- "BUY"
  }else { abandon_data$Purchase_Status[i] <- "NO BUY"}
}
#To make the column values signify: TRUE -->BUY, FALSE-->NOT BUY

abandon_data$Purchase_Status_basedon_Email <- NULL
abandon_data$Purchase_Status_basedon_Phone <- NULL
abandon_data$Empty_IncomingPhone <- NULL
abandon_data$Empty_Phone <- NULL
#Deleted Temporary columns which we had earlier created for data filteration.

#Filtering the data again based on the Phone Attribute to further concise our data for uniqueness.
filtered_abandon_data_DUP_Phone <- duplicated(abandon_data$Phone)
summary(filtered_abandon_data_DUP_Phone)
#We got around 130 more rows again for customers who have duplicated Phone number on our final dataset which we cannot eliminate as it represents big part of regression testing.


# Date and Time has been separated.
abandon_data$AB_temp <- strptime(abandon_data$Session, "%Y.%m.%d %H:%M:%S")
abandon_data$AB_Date <- strftime(abandon_data$AB_temp,"%Y-%m-%d")
abandon_data$AB_Time <- strftime(abandon_data$AB_temp,"%H:%M:%S")
reservation_data$RS_temp <- strptime(reservation_data$Session, "%Y.%m.%d %H:%M:%S")
reservation_data$RS_Date <- strftime(reservation_data$RS_temp,"%Y-%m-%d")
reservation_data$RS_Time <- strftime(reservation_data$RS_temp,"%H:%M:%S")
# The calculation of the date difference is done through excel VLOOKUP Formula.

#Empty State/Address and Email Fields are filled with value '0' in the final dataset.
abandon_data$Email <- ifelse(abandon_data$Email=="", "0", abandon_data$Email)
abandon_data$Address <- ifelse(abandon_data$Address=="", "0", abandon_data$Address)

#Removing Temporary and redundant columns from the dataset.
abandon_data$AB_temp <- NULL
abandon_data$Session <- NULL
abandon_data$First_Name <- NULL
abandon_data$Last_Name <- NULL
abandon_data$Street <- NULL
abandon_data$City <- NULL
abandon_data$Zipcode <- NULL
abandon_data$Incoming_Phone <- NULL
abandon_data$Contact_Phone <- NULL
abandon_data$Empty_IncomingPhone <- NULL
abandon_data$Phone <- NULL
```

```
abandon_data$Diff <- NULL
reservation_data$RS_temp <- NULL


#Writing final CSV file to system.
write.csv(abandon_data,"Test Final Abandon Dataset.csv")
write.csv(reservation_data,"Test Final Reservation Dataset.csv")
```

After this, the RS_Phone,RS_Date,RS_Time columns have been copied into Test Final Abandon Dataset and below matchup functions have been used for two match keys
i.e. (1) Phone (2) Email and the difference of the dates (in days) has been calculate in the column named as "Days_in_Between" as:

```
#This means if the Phone from Abandoned dataset matches with that from Reservation Phone, then take the second column i.e.RS_Date field for that particular row
where A2=Phone from Abandoned data set
=IFERROR(VLOOKUP(A2,Reservation_check_phone,2,FALSE),"200")
```

```
#Similarily,if the Email from Abandoned dataset matches with that from Reservation Phone, then take the second column i.e.RS_Date field for that particular row
where C2=Email from Abandoned data set
=IFERROR(VLOOKUP(C2,Reservation_check_email,2,FALSE),"200")
```

```
#This is the column "Days_in_Between", which matched the AB_Date Values with RS_Date values pulled in the above steps for Phone and email. Then
calculates the difference in days where J2=RS_Date with Phone match, K2=RS_Date with Email match and D2=AB_Date.
=IF(J2=200,IF(K2<>200,D2-K2,200),J2)
```

The final sheet is the cleaned data sheet where in all column added from Reservation datasets have been removed and only the relevant columns like CallerID,Test_Variable,Outcome,Days_in_between,Address, and Email been retained.

The sheets contains below statistics-

(1) Customers in the TREATMENT group who bought - 313
(2) Customers in the TREATMENT group who did not buy - 3953
(3) Customers in the Control group who bought - 86
(4) Customers in the Control group who did not buy - 4090

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: Yes, when we did match from abandoned dataset with the reservation dataset for some of the customers to get the reservation date(RS_Date) matching on Phone Key and Email Key, there were some cases which had Phone match but email mismatch and vice –versa. For such scenarios, I gave preference to the RS_Date obtained on the basis of Phone key match and in case this value is absent then pull out the RS_date for the same callerID on the basis of Email key match.

Below are few examples –

Caller_ID	Test_Variable	Outcome	Days_in_Between	Address	Email
57824008HGUXWDLO	0	1	20	1	0
30572700NMTAWKNZ	0	1	20	1	0
81947658KLKFQCCT	1	1	9	0	0
81801918BMCTGPZK	1	1	5	0	0
18169453RQNJDGRU	0	1	17	1	0
52867482KJPNFZNV	1	1	17	1	1

Q9: Complete the following cross-tabulation:

Group \ Outcome	Buy	No Buy
Treatment	313	3953
Control	86	4090

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

Answer:

AR (Arizona):

Group \ Outcome	Buy	No Buy
Treatment	2	36
Control	1	45

FL (Florida):

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

CA (California):

Group \ Outcome	Buy	No Buy
Treatment	6	42
Control	0	37

NY (New York):

Group \ Outcome	Buy	No Buy
Treatment	2	38
Control	1	35

TX (Texas):

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 with the following columns

Answer: Attached is the excel → MightyHive Cleaned Data.csv MightyHive Cleaned Data 0-1.csv

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

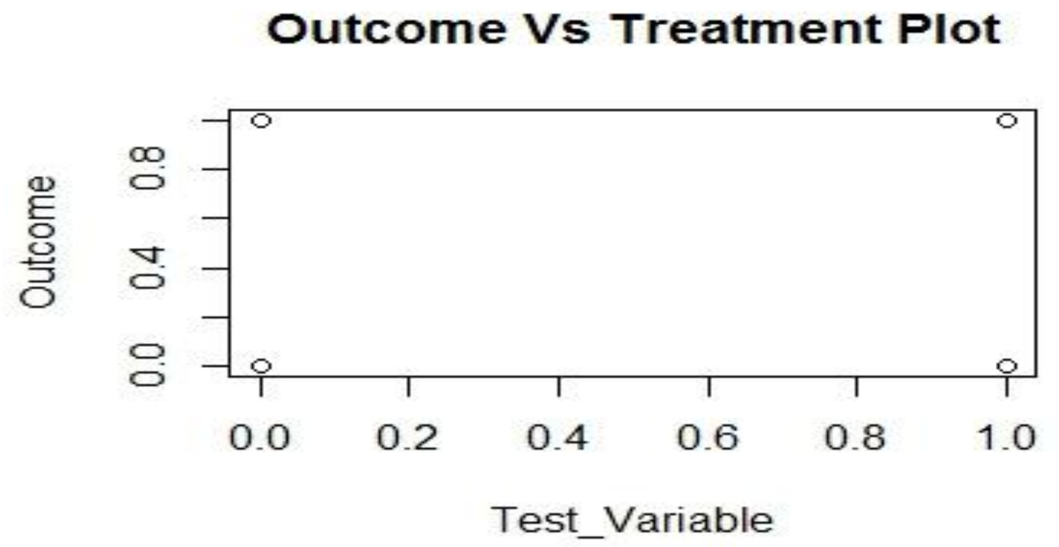
And Report the output.

Answer:

NULL HYPOTHESIS: THERE IS NO SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE i.e. $H_0 : \mu_1 - \mu_2 = 0$

ALTERNATE HYPOTHESIS: THERE IS A SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE i.e. $H_a : \mu_1 - \mu_2 \neq 0$

```
> clean_dataset <- read.csv("MightyHive Cleaned Data.csv", header = T, stringsAsFactors = FALSE)
> clean_dataset$Test_Variable <- ifelse(clean_dataset$Test_Variable=="test",1,0)
> clean_dataset$Outcome <- ifelse(clean_dataset$Outcome=="BUY",1,0)
> lm_clean_dataset <- lm(Outcome~Test_Variable)
> summary(lm_clean_dataset)

Call:
lm(formula = Outcome ~ Test_Variable)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.07337 -0.07337 -0.02059 -0.02059  0.97941 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.020594   0.003259    6.32 2.75e-10 ***
Test_Variable 0.052777   0.004584   11.51  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2106 on 8440 degrees of freedom
Multiple R-squared:  0.01546,	Adjusted R-squared:  0.01535 
F-statistic: 132.6 on 1 and 8440 DF,  p-value: < 2.2e-16

> plot(Outcome~Test_Variable, main = "Outcome Vs Treatment Plot")
```



Hence, we can conclude that,

alpha(Intercept) = **0.020594**

beta (Slope) = **0.052777**

And as the p-value is much less than 0.05, we reject the null hypothesis that *β(beta)* = 0. Hence there is a significant relationship between the variables in the linear regression model for Outcome and Test_variable.

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.

Answer: The below observation of the 95% confidence interval for the Outcome based on Test_Variable clearly indicates the p-value for this test, **2.2e-16 < 0.05** and hence rejects the Null Hypothesis. In other words, Alternate Hypothesis is TRUE. This means there exists a significant value for covariance which is not equal to 0 strongly suggesting relationship between the Outcome (offer taken or not) and the Test_Variable (treated or control group).

NULL HYPOTHESIS: THERE IS NO SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE i.e. $H_0 : \mu_1 - \mu_2 = 0$

ALTERNATE HYPOTHESIS: THERE IS A SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE i.e. $H_a : \mu_1 - \mu_2 \neq 0$

```
> t.test(Test_Variable)

        One Sample t-test

data:  Test_Variable
t = 92.86, df = 8441, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.4946631 0.5159979
sample estimates:
mean of x
0.5053305
```

It is important to randomize the sample data properly because to analysis a business decision based on statistical data requires sample collection uniformly from the range of values and not just the extreme values. In here within our experiment, we have taken samples of the data from both "Treated" and "Control" groups from last year's sample abandoned dataset and we have found there exists 313 customers from Treated group who bought the offer after calling and 86 customers who bought the offer without calling. This signifies, the data not only considers focusing on the treated group for statistical analysis rather it takes into consideration control group statistical data as well. This strengthens our data value and predicts more reliable data analysis.

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: Not Exactly because the R-Squared value is very small ~ 1.6%, But the linear regression model specified above **is statistically significant** with the assumption that the data does correlates with the Outcome and Test_Variable after

it rejects the null hypothesis based on A/B Testing. This is because as shown below, the experiment indicates the p-value for this test, **2.2e-16 < 0.05** and hence rejects the Null Hypothesis. In other words, Alternate Hypothesis is TRUE. This means there exists a significant value for covariance, which is not equal to 0, strongly suggesting relationship between the Outcome (offer taken or not) and the Test_Variable (treated or control group).
Also, In here within our experiment, we have taken samples of the data from both "Treated" and "Control" groups from last year's sample abandoned dataset and we have found there exists 313 customers from Treated group who bought the offer after calling this year and 86 customers who bought the offer without calling this year. This concludes, calling people and offering them again does convince them to buy it this year and statistical evidence shows the effectiveness of the retargeting campaign.

NULL HYPOTHESIS: THERE IS NO SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE i.e. $H_0 : \mu_1 - \mu_2 = 0$

ALTERNATE HYPOTHESIS: THERE IS A SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE i.e. $H_a : \mu_1 - \mu_2 \neq 0$

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.07337 -0.07337 -0.02059 -0.02059  0.97941 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.020594   0.003259    6.32 2.75e-10 ***
Test_Variable 0.052777   0.004584   11.51  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2106 on 8440 degrees of freedom
Multiple R-squared:  0.01546,	Adjusted R-squared:  0.01535 
F-statistic: 132.6 on 1 and 8440 DF,  p-value: < 2.2e-16
```

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)

Answer:

```
> lm_dummy_clean_dataset_new <- lm(Outcome~(Test_Variable+Address+Email))
> summary(lm_dummy_clean_dataset_new)

Call:
lm(formula = Outcome ~ (Test_Variable + Address + Email))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.11176 -0.06185 -0.05990 -0.01000  0.99000 
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.009996   0.003853   2.594  0.00949 ** 
Test_Variable 0.051856   0.004575  11.334  < 2e-16 ***
Address       0.015094   0.004713   3.202  0.00137 ** 
Email         0.034812   0.007169   4.856 1.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2101 on 8438 degrees of freedom
Multiple R-squared:  0.02044,     Adjusted R-squared:  0.02009 
F-statistic:  58.7 on 3 and 8438 DF,  p-value: < 2.2e-16
```

The linear regression model for above statistics is:

Outcome = alpha + $\beta0$(Test_Variable) + $\beta1$*(Address) + $\beta2$*(Email) + error*

NULL HYPOTHESIS: THERE IS NO SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE, ADDRESS(STATE) and EMAIL i.e. $H_0 : \mu_1 - \mu_2 = 0$

ALTERNATE HYPOTHESIS: THERE IS A SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE, ADDRESS(STATE) and EMAIL i.e. $H_a : \mu_1 - \mu_2 \neq 0$

Hence, we can conclude that,

alpha(Intercept) = **0.00996**
$\beta0$ (Slope1) = **0.051856**
$\beta1$ (Slope2) = **0.015094**
$\beta2$ (Slope1) = **0.034812**

And as the p-value (for all data variables) is much less than 0.05, we reject the null hypothesis that $\beta0, \beta1, \beta2$ = 0 and our alternate hypothesis is TRUE. Hence there is a significant relationship between the variables in the linear regression model for Outcome and Test_variable, Address(State), Email.

Now, **Including interactions with the treatment:**
The linear regression model for statistics is:

Outcome = alpha + $\beta0$(Test_Variable)*(Address) + $\beta1$*(Test_Variable)*(Email) + error*

```
> lm_dummy_clean_dataset_new <- lm(Outcome~(Test_Variable*Address+Test_Variable
*Email))
> summary(lm_dummy_clean_dataset_new)

Call:
lm(formula = Outcome ~ (Test_Variable * Address + Test_Variable * 
    Email))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13693 -0.05533 -0.02982 -0.01652  0.98348 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           0.016520   0.004389   3.764 0.000169 ***
```

```
Test_Variable         0.038807   0.006220   6.240  4.6e-10 ***
Address               0.007781   0.006692   1.163 0.244934    
Email                 0.005518   0.010550   0.523 0.600950    
Test_Variable:Address 0.014478   0.009417   1.537 0.124246    
Test_Variable:Email   0.053829   0.014369   3.746 0.000181 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2098 on 8436 degrees of freedom
Multiple R-squared:  0.02275,	Adjusted R-squared:  0.02217 
F-statistic: 39.28 on 5 and 8436 DF,  p-value: < 2.2e-16
```

Hence, we can conclude that,

alpha(Intercept) = **0.016520**

β0 (Slope1) = **0.014478**

β1 (Slope2) = **0.053829**

and **Multiple R-squared: 0.02275 ~23% which denotes that the interaction model is relatively better than the above model.**

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?

Answer:

The linear regression model for above statistics is:

Outcome = alpha + β0(Test_Variable) + β1*(Days_in_between) + error*

NULL HYPOTHESIS: THERE IS NO SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE, DAYS_IN_BETWEEN i.e. H_0 : $\mu_1 - \mu_2 = 0$

ALTERNATE HYPOTHESIS: THERE IS A SIGNIFICANT RELATIONSHIP BETWEEN OUTCOME AND TEST_VARIABLE, DAYS_IN_BETWEEN i.e. H_a :$\mu_1 - \mu_2 \neq 0$

```
> lm_dummy_clean_dataset_DIB <- lm(Outcome~Test_Variable+Days_in_Between)
> summary(lm_dummy_clean_dataset_DIB)

Call:
lm(formula = Outcome ~ Test_Variable + Days_in_Between)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.272654 -0.001438  0.000406  0.000406  0.294504 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
```

```
(Intercept)      1.3034051  0.0014948  871.96  < 2e-16 ***
Test_Variable    0.0018439  0.0004789    3.85 0.000119 ***
Days_in_Between -0.0065191  0.0000074 -881.00  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02184 on 8439 degrees of freedom
Multiple R-squared:  0.9894,     Adjusted R-squared:  0.9894
F-statistic: 3.942e+05 on 2 and 8439 DF,  p-value: < 2.2e-16
```

Hence, we can conclude that,

alpha(Intercept) = **1.3034051**
$\beta0$ (Slope1) = **0.0018439**
$\beta1$ (Slope2) = **-0.0065191**

And as the p-value (for all data variables) is much less than 0.05, we reject the null hypothesis that $\beta0, \beta1$ = 0 and our alternate hypothesis is TRUE. Hence there is a significant relationship between the variables in the linear regression model for Outcome and Test_variable, Days_in_Between.
Also, The Adjusted R-Squared ~ 99% which indicates the regression model is highly significant for prediction when Time response is included as data value.

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: The data values considered for the experiment are too narrow. And most of the data in the abandon and reservation datasets like email, state/address, etc are empty or space or blank. This causes lot many complications while matching the sheets to make a statistical inference if the customer was responsive after the treatment. So, the first point I would have made sure in designing the experiment is there were all data fields filled with data values. Also, there could have been more data values like reason for abandoning, work_sector, salary_range, etc which would have made us more clearly analyze the experiment and get to know why had people abandoned last year and is it going to be beneficial to treat them again.
There were so many redundant column values like Incoming_Phone, Contact_Phone which were containing duplicates. This should have been considered to contain a unique value and saved lot of time in data cleaning.

Also, the abandon dataset (8442 rows) and reservation dataset (20814 rows) were unevenly distributed and hence a randomized sample set with equal distribution could have given more profound results.

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: 95
Comment – Devoted lot of time in figuring out the script and matching indexes. Also found so many redundant column rows based on Phone and Email. So filtering the data and final set of rows is very precise and compact in my set of cleaned data values. I have also tried my best to elaborate each and every significant conclusion withdrawn along with assumptions and every set of operation performed over data cleaning.

Expected Performance: 95