A/B Testing Analytics: Mighty Hive Project

Vishal Punjabi

March 17, 2016

I. The Business Problem

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

Business Problem: Should we retarget those customers?

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

It takes no harm to retarget the abondoned customers to understand why they did not buy the vacation package in the first attempt even after calling the agency.

A customer may not always buy the product in the first attempt because 1) He might more time to buy 2) He might consider other alternatives, compare them and then decide. 3) He might be interested to buy the product in future.

Also, even though these customers did not buy it this time, they might be interested in future, hence retargeting them might help for the next time.

However, a thourough analysis would be good to carry out between the ones those bought the package and the ones who turned down before we can retarget.

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.

```
abandoned_data$Test  <- NA
abandoned_data$Test[abandoned_data$Test_Control == "test"] <- 1
abandoned_data$Test[abandoned_data$Test_Control == "control"] <- 0

summary(abandoned_data$Test)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5053  1.0000  1.0000
```

Standard Deviation = 0.5000012
q5 = 0
q95 = 1

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

```
abandoned_data$Has_State <- 0
abandoned_data$Has_State[abandoned_data$Address != ""] <- 1

summary(abandoned_data$Test[abandoned_data$Has_State == 1])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5134  1.0000  1.0000
```

Standard Deviation = 0.4998865
q5 = 0
q95 = 1

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

The mean and the standard deviation of the test variable is very near to 0.5 for in both the cases. i.e considering all samples and only State level ones. Hence there is no imbalance in the assignements to treatment and control and the experiment appear to be executed properly.

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.

Attributes like First Name and Last Name are missing in the datasets hence it cannot be used as the only way of determining the outcomes of the test. So in order to match the datasets correctly some key attributes about the customers need to be identified and matched in both the datasets. Only then it can be said that a match is present, and the customer who was abandoned has be converted in the reservation category.

Hence robust matching of the datasets is required which can backout the information correctly.

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)

In order to match the ABD dataset with the RS dataset robustly, some key attributes or "data keys" need to be identified. Some of these keys that can be used to uniquely identify

the customer in both data sets are -
1) [Email Address]
2) [Contact Phone]
3) [Incoming Phone,Last Name]
4) [First Name, Last Name, Zip]

An aggregation of unique data using these 4 keys to match should be enough in identifying the customers that have been converted to buy the package successfully.

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.

```
abandoned_data[abandoned_data == ""] <- NA
reservation_data[reservation_data == ""] <- NA

#Email Matches
Email_Matches_Abandoned <-
ifelse(!is.na(abandoned_data$Email),abandoned_data$Email %in%
reservation_data$Email,FALSE)

#Contact Phone Matches
ContactPhone_Matches_Abandoned <-
ifelse(!is.na(abandoned_data$Contact_Phone),abandoned_data$Contact_Phone %in%
reservation_data$Contact_Phone,FALSE)

#Last Name, Incoming Phone Matches
LastName_Incoming_Matches_Abandoned <-
ifelse(!is.na(abandoned_data$Last_Name) &
!is.na(abandoned_data$Incoming_Phone),paste0(abandoned_data$Last_Name,abandon
ed_data$Incoming_Phone) %in%
paste0(reservation_data$Last_Name,reservation_data$Incoming_Phone),FALSE)

#First Name, Last Name, Zip Matches
Names_Zip_Matches_Abandoned <- ifelse((!is.na(abandoned_data$First_Name) &
!is.na(abandoned_data$Last_Name)) & !is.na(abandoned_data$Zipcode)
,paste0(abandoned_data$First_Name,abandoned_data$Last_Name,abandoned_data$Zip
code) %in%
paste0(reservation_data$First_Name,reservation_data$Last_Name,reservation_dat
a$Zipcode),FALSE)

# Combine all Matches
All_Matches_Abandoned = Email_Matches_Abandoned |
ContactPhone_Matches_Abandoned | LastName_Incoming_Matches_Abandoned |
Names_Zip_Matches_Abandoned
abandoned_data_matches <- abandoned_data[All_Matches_Abandoned,]
```

```
#Remove Duplicates based on the keys
abandoned_data_matches <-
abandoned_data_matches[!duplicated(abandoned_data_matches[,c("Email")],incomp
arables = NA),]
abandoned_data_matches <-
abandoned_data_matches[!duplicated(abandoned_data_matches[,c("Contact_Phone")
],incomparables = NA),]
abandoned_incoming_dup <-
duplicated(abandoned_data_matches[,c("Incoming_Phone")],incomparables = NA)
abandoned_lastname_dup <-
duplicated(abandoned_data_matches[,c("Last_Name")],incomparables = NA)
abandoned_firstname_dup <-
duplicated(abandoned_data_matches[,c("First_Name")],incomparables = NA)
abandoned_zipcode_dup <-
duplicated(abandoned_data_matches[,c("Zipcode")],incomparables = NA)
abandoned_data_matches <- abandoned_data_matches[!(abandoned_incoming_dup &
abandoned_lastname_dup),]
abandoned_data_matches <- abandoned_data_matches[!(abandoned_firstname_dup &
abandoned_lastname_dup & abandoned_zipcode_dup),]

# Store Outcome in original dataset
abandoned_data$Outcome <- 0
abandoned_data$Outcome[as.numeric(row.names(abandoned_data_matches))] <- 1
```

Detailed Data Matching Procedure :

- Assign all missing values in both Abandunt and Reservation dataset as NA
- Match the customers in both AB and RS based on Email address only
- Match the customers in both AB and RS based on Contact Phone only
- Match the customers in both AB and RS based on combination of Last Name and Incoming Phone
- Match the customers in both AB and RS based on combination of First Name, Last Name and Zip Code.
- Combine all the customer matches in a single list.
- Remove duplicate customers based on each of the Keys i.e. Email, Contact Phone, Last Name, Incoming Phone, First Name, Zip Code.
- Total 223 matched customers are marked Buy = 1 in the original dataset.

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.

Contact Phone, Email, Incoming Phone, First Name, Last Name are the key attributes to match data between two datasets. However, there are scenarios where data was not recorded for Incoming Phone , Contact Phone & Email for some customers. In that case, matches has to be done on each key and then all customers need to be agregated.

Q9: Complete the following cross-tabulation:

Group Outcome Buy No Buy Treatment Number Number Control Number Number

```
library(knitr)

## Warning: package 'knitr' was built under R version 3.2.4

treatments <- nrow(abandoned_data[abandoned_data$Test == 1,])
controls <- nrow(abandoned_data[abandoned_data$Test == 0,])

treatment_buy <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 &
abandoned_data$Test == 1])
treatment_nobuy <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0
& abandoned_data$Test == 1])
control_buy <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 &
abandoned_data$Test == 0])
control_nobuy <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0 &
abandoned_data$Test == 0])

conv_rate_treatment <- treatment_buy/treatments*100
conv_rate_control <- control_buy/controls*100
cross_tab <-
data.frame(treatment_buy,treatment_nobuy,control_buy,control_nobuy)
kable(cross_tab)
```

treatment_buy	treatment_nobuy	control_buy	control_nobuy
181	4085	42	4134

Conversion Rate for Treatment Group is 4.2428504 %.
Conversion Rate for Control Group is 1.0057471 %.

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

State: New York

```
#NY
treatment_buy_NY <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1
& abandoned_data$Test == 1 & abandoned_data$Address == "NY"])
treatment_nobuy_NY <- length(abandoned_data$Outcome[abandoned_data$Outcome ==
0 & abandoned_data$Test == 1 & abandoned_data$Address == "NY"])
control_buy_NY <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 &
abandoned_data$Test == 0 & abandoned_data$Address == "NY"])
control_nobuy_NY <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0
& abandoned_data$Test == 0 & abandoned_data$Address == "NY"])
cross_tab_NY <-
data.frame(treatment_buy_NY,treatment_nobuy_NY,control_buy_NY,control_nobuy_N
Y)
kable(cross_tab_NY)
```

treatment_buy_NY	treatment_nobuy_NY	control_buy_NY	control_nobuy_NY
64	2285	16	2341

State: Ohio

```
#OH
treatment_buy_OH <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1
& abandoned_data$Test == 1 & abandoned_data$Address == "OH"])
treatment_nobuy_OH <- length(abandoned_data$Outcome[abandoned_data$Outcome ==
0 & abandoned_data$Test == 1 & abandoned_data$Address == "OH"])
control_buy_OH <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 &
abandoned_data$Test == 0 & abandoned_data$Address == "OH"])
control_nobuy_OH <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0
& abandoned_data$Test == 0 & abandoned_data$Address == "OH"])
cross_tab_OH <-
data.frame(treatment_buy_OH,treatment_nobuy_OH,control_buy_OH,control_nobuy_O
H)
kable(cross_tab_OH)
```

treatment_buy_OH	treatment_nobuy_OH	control_buy_OH	control_nobuy_OH
64	2295	15	2345

State: Arizona

```
#AZ
treatment_buy_AZ <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1
& abandoned_data$Test == 1 & abandoned_data$Address == "AZ"])
treatment_nobuy_AZ <- length(abandoned_data$Outcome[abandoned_data$Outcome ==
0 & abandoned_data$Test == 1 & abandoned_data$Address == "AZ"])
control_buy_AZ <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 &
abandoned_data$Test == 0 & abandoned_data$Address == "AZ"])
control_nobuy_AZ <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0
& abandoned_data$Test == 0 & abandoned_data$Address == "AZ"])
cross_tab_AZ <-
data.frame(treatment_buy_AZ,treatment_nobuy_AZ,control_buy_AZ,control_nobuy_A
Z)
kable(cross_tab_AZ)
```

treatment_buy_AZ	treatment_nobuy_AZ	control_buy_AZ	control_nobuy_AZ
63	2300	16	2349

State: Illinois

```
#IL
treatment_buy_IL <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1
& abandoned_data$Test == 1 & abandoned_data$Address == "IL"])
treatment_nobuy_IL <- length(abandoned_data$Outcome[abandoned_data$Outcome ==
0 & abandoned_data$Test == 1 & abandoned_data$Address == "IL"])
control_buy_IL <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 &
abandoned_data$Test == 0 & abandoned_data$Address == "IL"])
control_nobuy_IL <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0
& abandoned_data$Test == 0 & abandoned_data$Address == "IL"])
cross_tab_IL <-
```

```
data.frame(treatment_buy_IL,treatment_nobuy_IL,control_buy_IL,control_nobuy_IL)
kable(cross_tab_IL)
```

treatment_buy_IL	treatment_nobuy_IL	control_buy_IL	control_nobuy_IL
62	2284	15	2353

State: California

```
#CA
treatment_buy_CA <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 & abandoned_data$Test == 1 & abandoned_data$Address == "CA"])
treatment_nobuy_CA <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0 & abandoned_data$Test == 1 & abandoned_data$Address == "CA"])
control_buy_CA <- length(abandoned_data$Outcome[abandoned_data$Outcome == 1 & abandoned_data$Test == 0 & abandoned_data$Address == "CA"])
control_nobuy_CA <- length(abandoned_data$Outcome[abandoned_data$Outcome == 0 & abandoned_data$Test == 0 & abandoned_data$Address == "CA"])
cross_tab_CA <-
data.frame(treatment_buy_CA,treatment_nobuy_CA,control_buy_CA,control_nobuy_CA)
kable(cross_tab_CA)
```

treatment_buy_CA	treatment_nobuy_CA	control_buy_CA	control_nobuy_CA
64	2293	15	2343

III. Data Cleaning

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

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

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

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

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

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

```
#Getting the corresponding match indices in the reservation dataset
reservation_email_matches <- match(abandoned_data_matches$Email,
reservation_data$Email, nomatch = 0, incomparables = NA)
```

```
reservation_phone_matches <- match(abandoned_data_matches$Contact_Phone,
reservation_data$Contact_Phone, nomatch = 0, incomparables = NA)
reservation_name_incoming_matches <-
ifelse(!is.na(abandoned_data_matches$Last_Name) &
!is.na(abandoned_data_matches$Incoming_Phone),match(paste0(abandoned_data_mat
ches$Last_Name,abandoned_data_matches$Incoming_Phone),
paste0(reservation_data$Last_Name,reservation_data$Incoming_Phone), nomatch =
0, incomparables = NA),0)
reservation_name_zip_matches <-
ifelse(!is.na(abandoned_data_matches$First_Name) &
!is.na(abandoned_data_matches$Last_Name) &
!is.na(abandoned_data_matches$Zipcode),match(paste0(abandoned_data_matches$Fi
rst_Name,abandoned_data_matches$Last_Name,abandoned_data_matches$Zipcode),
paste0(reservation_data$First_Name,reservation_data$Last_Name,reservation_dat
a$Zipcode), nomatch = 0, incomparables = NA),0)
reservation_all_matches <- reservation_email_matches
reservation_all_matches <- ifelse(reservation_all_matches ==
0,reservation_all_matches+reservation_phone_matches,reservation_all_matches)
reservation_all_matches <- ifelse(reservation_all_matches ==
0,reservation_all_matches+reservation_name_incoming_matches,reservation_all_m
atches)
reservation_all_matches <- ifelse(reservation_all_matches ==
0,reservation_all_matches+reservation_name_zip_matches,reservation_all_matche
s)
abandoned_all_matches <- as.numeric(row.names(abandoned_data_matches))

#Session Calculation, Days in Between
abandoned_data$Days_in_between <- 200
abandoned_data$Days_in_between[abandoned_data$Outcome == 1] <-
as.numeric(as.Date(reservation_data$Session[reservation_all_matches],"%Y.%m.%
d %H:%M:%S") -
as.Date(abandoned_data$Session[abandoned_all_matches],"%Y.%m.%d %H:%M:%S"))

cleaned_abandoned_data <-
data.frame(c(1:nrow(abandoned_data)),abandoned_data$Test,abandoned_data$Outco
me,abandoned_data$Days_in_between,abandoned_data$Address)
colnames(cleaned_abandoned_data) <-
c("Customer_ID","Test_Variable","Outcome","Days_in_Between","State")

write.csv(cleaned_abandoned_data,file = "Cleaned_Abandoned_Data_Seed.csv")
```

Cleaned_Abandoned_Data_Seed.csv has been attached with the submission for review.

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.

Model-1: Outcome = alpha + beta * Test Variable + error*

```
lmodel1 <- lm(cleaned_abandoned_data$Outcome ~
cleaned_abandoned_data$Test_Variable)
kable(summary(lmodel1)$coef, digits=3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.010	0.002	4.073	0
cleaned_abandoned_data$Test_Variable	0.032	0.003	9.319	0

Outcome = 0.01 + 0.032 * Test_Variable + 0.002
Adjusted R-squared = 0.0100679

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

No, This is not a properly specified linear regression model because the Test_Variable coefficient (beta) is eual to 0.03 i.e. the dependent variable "Outcome" would increase 3% for every addition of a customer in the Test group which is very less. Also the R-Sqaured value is 0.01 which is also very less to properly incorporate the variability of the dataset. We need to add more variables to get the significance out of this model.

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

Adding Has_Email as a Dummy Variable in the original dataset.

```
abandoned_data$Has_Email <- 0
abandoned_data$Has_Email[!is.na(abandoned_data$Email)] <- 1
```

Model-2: Outcome = alpha + beta1 * Test Variable + beta2 * Has Email * beta3 * Has_State + error*

```
lmodel2 <- lm(cleaned_abandoned_data$Outcome ~
cleaned_abandoned_data$Test_Variable + abandoned_data$Has_Email +
abandoned_data$Has_State)
kable(summary(lmodel2)$coef, digits = 3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.002	0.003	-0.635	0.525
cleaned_abandoned_data$Test_Variable	0.031	0.003	9.034	0.000
abandoned_data$Has_Email	0.049	0.005	8.997	0.000
abandoned_data$Has_State	0.015	0.004	4.092	0.000

Outcome = -0.001 + 0.031 * Test Variable + 0.048 * Has Email * 0.014 * Has_State + 0.002
Adjusted R-squared = 0.0237249

The adjusted R-squared has increased to 0.023 after using the dummy variables - Has Email and Has State. Hence compared to the output from the first model, this fits better.

Model-3: Outcome = alpha + beta1 * Test Variable * Has Email * beta2 * State + error*

```
lmodel3 <- lm(cleaned_abandoned_data$Outcome ~
cleaned_abandoned_data$Test_Variable*abandoned_data$Has_Email +
cleaned_abandoned_data$State)
kable(summary(lmodel3)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0149849	0.0244899	0.6118808	0.5406536
cleaned_abandoned_data$Test_Variable	0.0277829	0.0069233	4.0129629	0.0000611
abandoned_data$Has_Email	0.0134249	0.0113205	1.1858864	0.2357421
cleaned_abandoned_data$StateAL	-0.0218481	0.0321386	-0.6798098	0.4966668
cleaned_abandoned_data$StateAR	-0.0153992	0.0318182	-0.4839731	0.6284332
cleaned_abandoned_data$StateAZ	-0.0164691	0.0308588	-0.5336912	0.5935868
cleaned_abandoned_data$StateCA	-0.0041644	0.0317363	-0.1312173	0.8956104
cleaned_abandoned_data$StateCO	0.0092925	0.0324195	0.2866340	0.7744084
cleaned_abandoned_data$StateCT	-0.0291119	0.0326076	-0.8927944	0.3720244
cleaned_abandoned_data$StateDE	-0.0140307	0.0321555	-0.4363397	0.6626154
cleaned_abandoned_data$StateFL	-0.0247889	0.0326016	-0.7603577	0.4470885

cleaned_abandoned_data$StateGA	-0.0092125	0.0321608	-0.2864522	0.7745476
cleaned_abandoned_data$StateHI	0.0088478	0.0322358	0.2744722	0.7837369
cleaned_abandoned_data$StateIA	0.0144741	0.0328046	0.4412226	0.6590772
cleaned_abandoned_data$StateID	0.0084242	0.0343876	0.2449790	0.8064860
cleaned_abandoned_data$StateIL	-0.0249158	0.0318118	-0.7832261	0.4335437
cleaned_abandoned_data$StateIN	-0.0343514	0.0338375	-1.0151870	0.3100820
cleaned_abandoned_data$StateKS	-0.0158384	0.0323383	-0.4897722	0.6243237
cleaned_abandoned_data$StateKY	-0.0296647	0.0336035	-0.8827861	0.3774083
cleaned_abandoned_data$StateLA	-0.0318149	0.0326176	-0.9753909	0.3294292
cleaned_abandoned_data$StateMA	0.0014909	0.0331310	0.0450015	0.9641085
cleaned_abandoned_data$StateMD	0.0032727	0.0323168	0.1012687	0.9193425
cleaned_abandoned_data$StateME	0.0165227	0.0327075	0.5051670	0.6134711
cleaned_abandoned_data$StateMI	0.0065946	0.0325213	0.2027797	0.8393182
cleaned_abandoned_data$StateMN	-0.0185023	0.0332539	-0.5563968	0.5779727
cleaned_abandoned_data$StateMO	-0.0200306	0.0327176	-0.6122265	0.5404250
cleaned_abandoned_data$StateMS	-0.0110067	0.0337185	-0.3264300	0.7441172

cleaned_abandoned_data$StateMT	-0.0072416	0.0332289	-0.2179295	0.8274959
cleaned_abandoned_data$StateNC	0.0135130	0.0331345	0.4078218	0.6834277
cleaned_abandoned_data$StateND	0.0294907	0.0341037	0.8647365	0.3872386
cleaned_abandoned_data$StateNE	0.0132709	0.0323303	0.4104769	0.6814795
cleaned_abandoned_data$StateNH	0.0217146	0.0329126	0.6597651	0.5094450
cleaned_abandoned_data$StateNJ	0.0131561	0.0315131	0.4174798	0.6763513
cleaned_abandoned_data$StateNM	0.0066873	0.0324256	0.2062365	0.8366174
cleaned_abandoned_data$StateNV	-0.0159720	0.0308381	-0.5179314	0.6045366
cleaned_abandoned_data$StateNY	0.0050642	0.0325314	0.1556698	0.8763017
cleaned_abandoned_data$StateOH	-0.0109494	0.0314464	-0.3481934	0.7277144
cleaned_abandoned_data$StateOK	-0.0383080	0.0330217	-1.1600881	0.2460866
cleaned_abandoned_data$StateOR	0.0087460	0.0323314	0.2705107	0.7867822
cleaned_abandoned_data$StatePA	0.0014746	0.0313042	0.0471064	0.9624309
cleaned_abandoned_data$StateRI	-0.0054873	0.0332555	-0.1650033	0.8689503
cleaned_abandoned_data$StateSC	0.0481968	0.0328207	1.4684866	0.1420558
cleaned_abandoned_data$StateSD	-0.0098074	0.0328009	-0.2989978	0.7649582
cleaned_abandoned_data$StateTN	0.0015273	0.0320938	0.0475883	0.9620469
cleaned_abandoned_data$StateTX	-	0.0324	-	0.3662

	0.0292922	172	0.9035997	656
cleaned_abandoned_data$StateUT	0.0216456	0.0344109	0.6290341	0.5293649
cleaned_abandoned_data$StateVA	-0.0073390	0.0320689	-0.2288521	0.8189964
cleaned_abandoned_data$StateVT	0.0196138	0.0320624	0.6117390	0.5407474
cleaned_abandoned_data$StateWA	-0.0024484	0.0330332	-0.0741191	0.9409196
cleaned_abandoned_data$StateWI	-0.0130084	0.0327160	-0.3976173	0.6909349
cleaned_abandoned_data$StateWV	0.0569159	0.0308969	1.8421265	0.0655354
cleaned_abandoned_data$StateWY	-0.0230951	0.0325368	-0.7098169	0.4778617
cleaned_abandoned_data $Test_Variable: abandoned_data$Has_Email	0.0864081	0.0154452	5.5945083	0.0000000

Outcome = 0.014 + 0.086 * Test Variable * Has Email * beta * State + 0.024
Adjusted R-squared = 0.0360938

When interactions are included with the treatment group i.e. Test Variable and Has Email, also including the State variables, a much better adjusted R-squared is obtained.

Hence, from the managerial perspective, customers who have a recorded email address and state on file from the treatment group grouping per state have more chances of converting to a reservation category.

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. Make sure you describe carefully how you compute response times (there is no clear answer, so make any sensible assumption).

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

Model-4: Outcome = alpha + beta * Days in Between + error

```
lmodel4 <- lm(cleaned_abandoned_data$Outcome ~
cleaned_abandoned_data$Days_in_Between)
kable(summary(lmodel4)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3045088	0.0014490	900.2612	0
cleaned_abandoned_data$Days_in_Between	-0.0065211	0.0000073	-888.8747	0

Outcome = 1.305e+00 + -6.521e-03 * Days in Between + 1.449e-03
Adjusted R-squared = 0.9894294

As the R-squared is 0.98, this linear regression model fits well. It can be inferred that the outcome is drastically dependent on the "Days in Between" i.e. the session time between the calls for the customer.

A negetive coefficient of Days in Between means as the Days in Between increases for each customer, the chances of Outcome being 1 reduces. Hence it is inversely proportional to outcome.

Model-5: Outcome = alpha + beta1 * Days in Between + beta2 * State + error

```
lmodel5 <- lm(cleaned_abandoned_data$Outcome ~
cleaned_abandoned_data$Days_in_Between + cleaned_abandoned_data$State)
kable(summary(lmodel5)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2999813	0.0032006	406.1693916	0.0000000
cleaned_abandoned_data$Days_in_Between	-0.0065083	0.0000106	-614.9843770	0.0000000
cleaned_abandoned_data$StateAL	0.0003449	0.0032515	0.1060729	0.9155302
cleaned_abandoned_data$StateAR	0.0042551	0.0032178	1.3223793	0.1861223
cleaned_abandoned_data$StateAZ	0.0029336	0.0031195	0.9403861	0.3470799
cleaned_abandoned_data$StateCA	0.0015947	0.0032097	0.4968294	0.6193384
cleaned_abandoned_data$StateCO	0.0057830	0.0032787	1.7637763	0.0778508
cleaned_abandoned_data$StateCT	0.0013843	0.0032981	0.4197351	0.6747029

cleaned_abandoned_data$StateDE	0.0025944	0.0032514	0.7979154	0.4249700
cleaned_abandoned_data$StateFL	0.0010372	0.0032981	0.3144909	0.7531657
cleaned_abandoned_data$StateGA	0.0038676	0.0032514	1.1895184	0.2343108
cleaned_abandoned_data$StateHI	0.0044432	0.0032603	1.3628051	0.1730255
cleaned_abandoned_data$StateIA	0.0052948	0.0033182	1.5956676	0.1106470
cleaned_abandoned_data$StateID	-0.0008251	0.0034780	-0.2372423	0.8124817
cleaned_abandoned_data$StateIL	0.0012605	0.0032179	0.3917181	0.6952887
cleaned_abandoned_data$StateIN	0.0016750	0.0034231	0.4893155	0.6246469
cleaned_abandoned_data$StateKS	0.0059555	0.0032694	1.8215753	0.0685988
cleaned_abandoned_data$StateKY	0.0034155	0.0033975	1.0053125	0.3148111
cleaned_abandoned_data$StateLA	0.0024257	0.0032981	0.7354656	0.4621018
cleaned_abandoned_data$StateMA	-0.0002820	0.0033504	-0.0841771	0.9329201
cleaned_abandoned_data$StateMD	0.0013372	0.0032694	0.4089986	0.6825640
cleaned_abandoned_data$StateME	0.0039266	0.0033080	1.1870056	0.2353004
cleaned_abandoned_data$StateMI	0.0012983	0.0032883	0.3948327	0.6929888
cleaned_abandoned_data$StateMN	0.0067025	0.0033617	1.9937950	0.0462466
cleaned_abandoned_data$StateMO	0.0027568	0.0033079	0.8334042	0.4046697
cleaned_abandoned_data$StateMS	0.0026063	0.0034098	0.7643387	0.4447134
cleaned_abandoned_data$StateMT	0.0014204	0.0033616	0.4225322	0.6726607

cleaned_abandoned_data$StateNC	0.0015450	0.0033505	0.4611121	0.6447449
cleaned_abandoned_data$StateND	0.0002685	0.0034499	0.0778380	0.9379611
cleaned_abandoned_data$StateNE	0.0010577	0.0032695	0.3234997	0.7463348
cleaned_abandoned_data$StateNH	0.0017294	0.0033288	0.5195223	0.6034272
cleaned_abandoned_data$StateNJ	-0.0011168	0.0031871	-0.3504052	0.7260543
cleaned_abandoned_data$StateNM	0.0002044	0.0032788	0.0623483	0.9502888
cleaned_abandoned_data$StateNV	0.0022263	0.0031196	0.7136343	0.4754976
cleaned_abandoned_data$StateNY	0.0024972	0.0032883	0.7594281	0.4476441
cleaned_abandoned_data$StateOH	0.0054009	0.0031795	1.6986787	0.0894625
cleaned_abandoned_data$StateOK	0.0016750	0.0033398	0.5015258	0.6160305
cleaned_abandoned_data$StateOR	-0.0038653	0.0032696	-1.1821935	0.2372037
cleaned_abandoned_data$StatePA	0.0010743	0.0031653	0.3394178	0.7343139
cleaned_abandoned_data$StateRI	0.0065752	0.0033616	1.9559485	0.0505452
cleaned_abandoned_data$StateSC	0.0036857	0.0033195	1.1102981	0.2669416
cleaned_abandoned_data$StateSD	0.0011669	0.0033181	0.3516684	0.7251067
cleaned_abandoned_data$StateTN	-0.0026155	0.0032428	-0.8065546	0.4199741
cleaned_abandoned_data$StateTX	0.0015609	0.0032788	0.4760595	0.6340596
cleaned_abandoned_data$StateUT	0.0016318	0.0034783	0.4691291	0.6390046
cleaned_abandoned_data$StateVA	0.0020728	0.0032427	0.6392303	0.5227120

cleaned_abandoned_data$StateVT	0.0029005	0.0032430	0.8943934	0.3711687
cleaned_abandoned_data$StateWA	0.0032288	0.0033394	0.9669069	0.3336527
cleaned_abandoned_data$StateWI	0.0024930	0.0033079	0.7536422	0.4511113
cleaned_abandoned_data$StateWV	-0.0007503	0.0031272	-0.2399158	0.8104086
cleaned_abandoned_data$StateWY	0.0006731	0.0032882	0.2047115	0.8378086

Outcome = 1.300e+00 + -6.508e-03 * Days in Between + beta * State + error
Adjusted R-squared = 0.9901341

When State as a dependent variable is included in the regression model, adjusted r-squared is almost 99% i.e. outcome for the treatment group is dependent on Days in Between and also dependent on each state

Hence, This is the best linear regression model for this dataset

Q15: Lesson Learned. What would you have done differently in designing the experiment? Any other directions you could have taken with better data? Are there any prescriptive managerial implications out of this study? Please answer briefly

It was a good experience overall as it involved everything from data cleaning to rigorous analysis and statistical modeling and inferences.

The experiment was fairly designed as matching turned out to be a challenge as there was no unique identity that could determine a customer who called. Hence if every customer is assigned a unique customer ID for every successive calls, it could be very easy in matching and understanding that the same customer has called again and has bought the package.

Having a customer ID as a unique identifier for each customer for the agency, matching could have been done solely on this key reducing the time required in Data Matching and Cleaning.

Managerial Implications : It can be seen that "Retargeting certainly helped". However, customers should be retargeted only in certan states and before a certain time to achieve maximum probablity of conversion. This can be understood from the last regression model.