**TEAM LEADA PROJECT**

Important Note: It is assumed that each student will sign up for the TeamLeada modules at <https://www.teamleada.com/courses/intro-to-ab-testing-in-r>

**Not signing up will lead to an automatic score of zero in the project.**

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

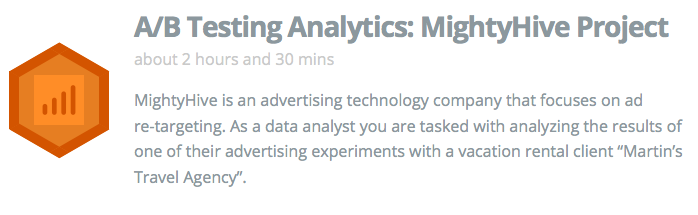


Figure 1: The fifth module of the Leada Project

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

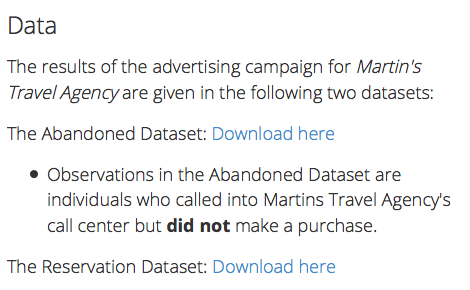


Figure 2: Where to download the two datasets

**Name: Sandeep Kumar Patra**

**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.

Reason :

This is a sensible business question because of many reasons. Firstly, for a successful business we need to have a good business strategy. In this case, strategy is to sell more vacation packages. With this analysis the previously abandoned customers were retargeted for the vacation package. As retargeting will increase the cost, so we will clean the data of abandon customers who has bought the package.

**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.

Solution –

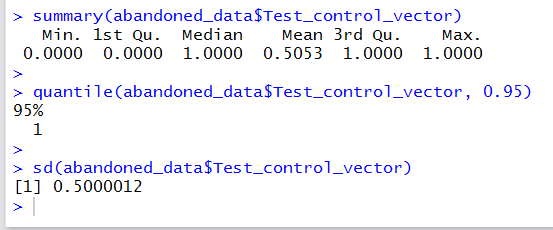
Mean - .5053

Median - 1

Q5 - 0

Q95 - 1

SD - .5



**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.

Solution –

Mean - .5134

Median - 1

Q5 - .0

Q95 - .1

SD - .4998

SCRIPT-

state = which(z$Address != '0')

state\_test\_vector\_state = z$Test\_control\_vector[state]

summary(state\_test\_vector\_state)

quantile(state\_test\_vector\_state, 0.05)

quantile(state\_test\_vector\_state, 0.95)

sd(state\_test\_vectorzz\_state)

**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?

Solution –

Mean (state-only level) - Mean (all) = .5134 - .5053 = .0081

The above difference in Mean is very small this suggest that no imbalance when switching to the state-only level .

**II. Data Matching**

About three months later, the experiment/retargeting campaign is over.

Customers, presented in the ABD excel file, who bought a vacation packages during the time frame, are recorded in the RS excel file.

**Q5:** Argue that for proper causal inference based on experiments this is potentially problematic: “We do not observe some “outcomes” for some customers”. Argue that, however, matching appropriately the ABD with the RS dataset can back out this information.

Reason5-

For some of the customers there is no outcome as the data is insufficient for them.

In this experiment we can make multiple data set by matching the Abandoned and Reservation dataset on the basis variables such as Emails, Contact etc. The new match data set of Abandoned and Reservation has customers called at least two times as they were retargeted, buying or making the reservation in the second time. Thus matching appropriately ABD and RS dataset we can back this information.

**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)

Reason 6-

Looking at the two dataset we can match the two files Abandoned and Reservation

On the basis of Incoming Phone, Contact, Email, First Name and Last Name etc. I had chosen Incoming Phone and Contact Phone as they uniquely identify individuals and we can easily remove duplicates without errors.

**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.**

Reason-

1. I combined both the sheets in one to make the process easier.

2. In abandoned\_data\_seed I removed all the rows which had null values for Incoming\_Phone. Created another column for test/control variable and assigned the value 1 to test and 0 to control.

3. Following data keys were most promising set of uniquely identifiable data to link callers across the two data seeds.

Incoming\_Call, Contact\_Phone, and Email.

4. To use the VLOOKUP to match the data in two sheets I set the incoming\_phone column to the first in the reservation data seed and I matched the cells till four columns from top to end. This operation was performed in Abandoned data seed.

5. I was able to retrieve some 259 rows by performing the above procedure. I then removed the rows where there were duplicates of Incoming\_Phone present. Which gave me total rows of the cleaned data around 220.

6. These records were pulled from the reservation data seed and one more important aspect of the analysis was to have the date difference between the forst and second call so I performed the trim function in the excel and got the difference of days between the two call sessions. .

7. I had already removed the rows where incoming\_phone was null so I was left with some 7090 rows and when combined this data seed in abandoned I was left with some 6870 number including test and control who did not buy the package.

**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.**

Reason – Yes , there are few problematic cases in the data sets, like for some of the records we didn’t had proper information of contact Email etc.

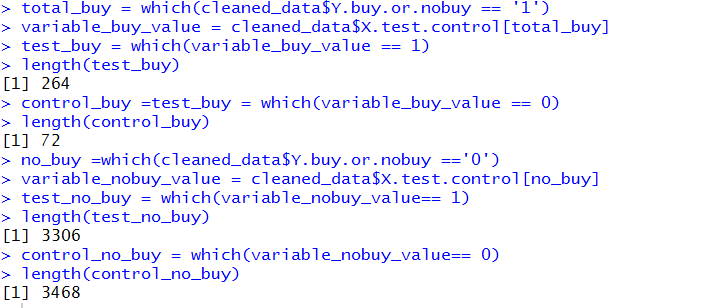
|  |
| --- |
| 86506709TBYMGXTQ- In this case there was no information available except the Incoming Phone |

64469846ZZVNVNL and 11719081PUQAOWSV – The two cases has same incoming phone number and no other fields on the basis of which it can be identified.

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

**Solution-**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **264** | **3306** |
| **Control** | **72** | **3468** |



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

**State : DE**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **2** | **22** |
| **Control** | **0** | **41** |

**State: NY**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **2** | **26** |
| **Control** | **1** | **32** |

**State: FL**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **1** | **24** |
| **Control** | **1** | **35** |

**State: IA**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **2** | **30** |
| **Control** | **1** | **26** |

**State: WV**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **7** | **38** |
| **Control** | **1** | **36** |

**III. Data Cleaning:**

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

Produce an Excel File with the following columns

Customer ID | Test Variable | Outcome | Days\_in\_Between | D\_State | D\_Email |

Where Test Variable indicates, again, the treatment or the control group, Outcome is a binary variable indicating whether a vacation package was ultimately bought, Days in between is the (largest) difference between the dates in the ABD and RS dataset (Columns B). If no purchase, set “Days\_in\_between” as “200”. Note also we have two dummies to signal whether the State and Email information is available for the customer.

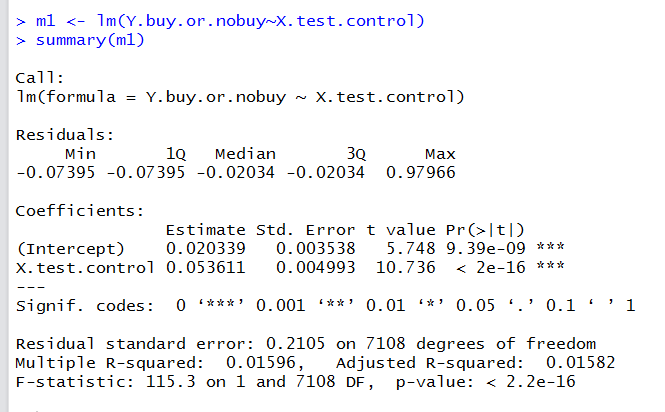
(Note that you should have as many rows as customers you were able to match across the two data sets. Be sure to attach this excel file to the submission for proper verification.)

**IV. Statistical Analysis**

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

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

Outcome = alpha + beta \* Test\_Variable + error



**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.

Solution –

t.test(Y.buy.or.nobuy,X.test.control)

Welch Two Sample t-test

data: Y.buy.or.nobuy and X.test.control

t = -70.607, df = 9589.2, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.4674800 -0.4422246

sample estimates:

mean of x mean of y

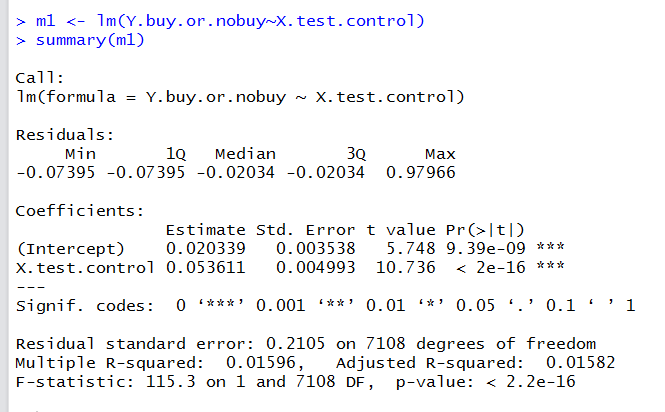
0.04725738 0.50210970

The output obtain from Linear Regression model is approximately similar to the output from A?B test procedure.

It is important to randomize the data properly .to eliminate any bias of the correlations on the BUY or NO BUY results for Treatment and Control.

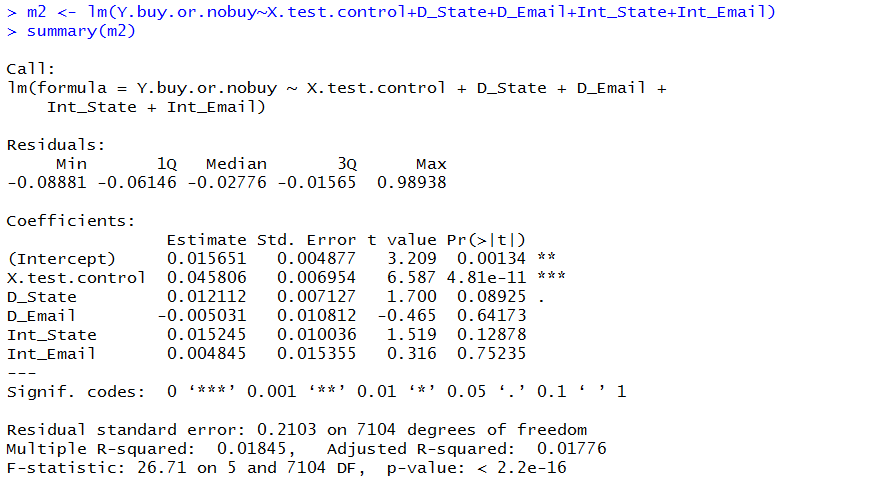
**Q13:** Argue whether this is a properly specified linear regression model, if so, if we can draw any causal statement about the effectiveness of the retargeting campaign. Is this statistically significant?

Reason –



The above model has very small Adjusted R2 and very small P-value which is equivalent to 0. So we can run another regression model by adding other independent variables such as email, state.

**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)

Solution –

The value of adjusted R- squared is1.7 % which is more than the Adjusted R2 of question 11 ie 1.5 percent hence this is a better model. This indicate that the outcome would be more efficient if we include all the variables and people with given email and state are more likely to buy the vacation package.

**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?

