**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”

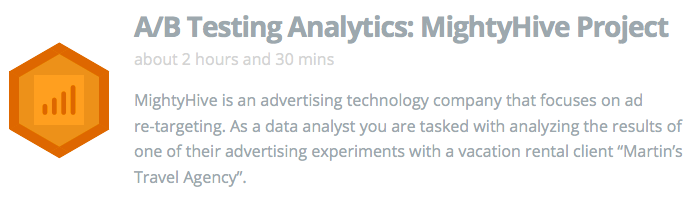


Figure 1: The fifth module of the Leada Project

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)**

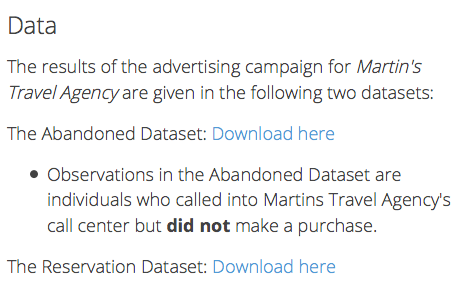


Figure 2: Where to download the two datasets

**EXAM**

**Name: Harshita Srivastava**

**Section: SDM 901**

**Signature (if possible) : Harshita Srivastava**

**Did you work with someone else while cleaning or analyzing the data? Please disclose your teammates. Be forthcoming to avoid potential bad consequences.**

**I did my assignment individually.**

**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.

This is a very sensible business question as the customers in the ABD dataset or the customers who already advertised but ended up not buying the vacation package are the customers who are potential customers or buyers of the Martins’s Travel Agency.

These customers were/are interested in buying the package at the first place but may have not purchased the deal depending on various reasons at that particular point of time. The reasons may account to some of the following explanations - the customer was interested in going for a holiday but at a later time, the customers were comparing the rates with a different travel agency , others.

This customer group should be re-targeted and it is seen that many a times customers purchase a deal when re-targeted.

Executives of the Martin's may find this data useful to deploy any kind of advertising or marketing strategies for the company and it might help them to understand whether customers actually turn in when re-targeted and in how much response time.

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.

summary(ABD\_test$Test\_control)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.0000 1.0000 0.5053 1.0000 1.0000

We calculate the 95% CI manually using the following code,

upper\_bound = b\_proportion - a\_proportion + 1.96\*SE

lower\_bound = b\_proportion - a\_proportion - 1.96\*SE

Confidence intervals are a measure of uncertainty that comes with the sampling method. In repeated samples, some intervals would include the true population parameter and some would not, 95% of those intervals calculated would include the true population parameter.

**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.

The summary statistics for the Test Variable is as under -

*summary(CleanData\_State$Test.Variable)*

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.0000 0.0000 1.0000 0.5134 1.0000 1.0000 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?

In the light of the summaries in Q3,Q4 the experiment appears to not have executed properly because we are not able to deduce the outcome of our business problem whether the experiment is executed properly or not.

As per my results, the treatment data when *switching to one state only* level is 3184 and zero state level is 7114 - 3184 = 3930.The treatment data *without switching* the one state level is 3327 and for the control group was 3477.There is an imbalance in the assignment of treatment as the value decreased from 3327 to 3184 and the value increased for the control group from to 3477 to 3930. Therefore there is a difference as it is not an ideal 50% -50% split in this case.

**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.

If “We do not observe some “outcomes” for some customers”, then we might not deduce the outcome very well from our model. More clean and simple the data will be, it will become easier for us to run the linear regression taking into account the various parameters which might affect the result.

If we match the ABD and RS dataset appropriately than we have a fair estimate of the customers who purchased the vacation package and we can find out the customers who did not purchase. We can then deduce from our model whether the customers in the ABD dataset who did not purchase earlier should be retargeted or not.

**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)

Customers can be matched across various data keys such as incoming phone, contact phone, address, email address etc.

For example, a data record may have a customer with different first name but same address, contact information as this person can be from the same household.

In the example demonstrated below, Bradley Little from the Abandoned Data Seed and Dillon Quitzon from the Reservation Seed have the same email and contact number and so they should be considered one record and not two as probably they are two different people from the same household.

Data from ABD file -

Data from RES file -

**Q7: EXTREMELY CAREFULLY DESCRIBE YOUR DATA MATCHING PROCEDURE IN ORDER TO IDENTIFY: (1) Customers in the TREATMENT group who bought (2) Customers in the TREATMENT group who did not buy (3) Customers in the Control group who bought, and (4) Customers in the Control group who did not buy. Be as precise as possible.**

I used the R script to perform the data matching procedure as described in the business problem.Before starting the data matching process I eliminated the duplicates from the ABD file by applying the duplicated function on the incoming phone and contact phone column of the ABD data seed. I received 7114 records in the ABD file on performing null checks and removing the duplicates.

**(1) Customers in the TREATMENT group who bought**

I then matched the customers which were present in both the ABD and RS files on the basis of Email and/or incoming phone to verify the customers present in both and these customers are the ones who actually bought the vacation package.I further categorized then in treatment and control groups based on the values in the test\_control variable.Therefore, customers in the treatment group who bought came to be 408 customers as per my solution.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dillon | Quitzon |  |  | AK | 46695-9066 | antifraudefcc\_ng@aol.in | (321)-492-4457 |  | test |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bradly | Little |  |  | NE |  | antifraudefcc\_ng@aol.in | (321)-492-4457 | (321)-492-4457 | test |

**(2) Customers in the TREATMENT group who did not buy**

The customers who did not buy are simply the difference of total customers in the ABD file and the matching customers in the RS file.Again we can categorize the information in treatment and control based on the variable test\_control and as per my solution the customers in the treatment group who did not buy are 3327

**(3) Customers in the Control group who bought**

Similarly, I found the difference on the set with total number of customers who bought and customers in the treatment group who bought and found out the value to be 145.

**(4) Customers in the Control group who did not buy.**

It is the difference of the total customers who did not buy and the customers in the treatment group who did not buy and this comes out to be 3477.

Code snippet -

#If the Contact Phone is duplicated in the data set

ABD\_Contact\_Phn\_Duplicates = duplicated(ABD\_Control\_Phn\_Duplicates\_Removed$Contact\_Phone)

#The number of rows in the data set is now 7114

ABD\_Contact\_Phn\_Duplicates\_Removed=ABD\_Control\_Phn\_Duplicates\_Removed[!ABD\_Contact\_Phn\_Duplicates,]

View(ABD\_Contact\_Phn\_Duplicates\_Removed)

#Storing the abandoned data after removing the duplicates in CleanData data frame

CleanData = ABD\_Contact\_Phn\_Duplicates\_Removed

CleanData <- CleanData[, c(2,7,9,10,11,12)]

#Matching whether the Incoming Phone from the Abandoned set data matched the same field in the reservation data

CleanData$Y <- ABD\_Contact\_Phn\_Duplicates\_Removed$Incoming\_Phone %in% RES$Incoming\_Phone

**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.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Kolby | Hermiston | 635 Brown Squares | Port Noeltown | RI | 78231-6267 | godwinchukwu@sohu.com | (234)-704-7890 |  | test |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Casper | Goldner |  |  |  |  |  | (234)-704-7890 | (234)-704-7890 | test |

Yes there are problematic cases of data records not matchable in the ABD and the RS file as per the example below -

From this example, if we compare the two set of records, the incoming phone value matches in both the rows but no other value is present in row 2 , so it is difficult to analyze whether the incoming phone number belongs to the person from the same household or is it some other person. Here email address is also not provided for the confirmation. We should toss out such cases from our analysis.

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

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **Number** | **Number** |
| **Control** | **Number** | **Number** |

**Result obtained after Data Cleaning MightyHive project -**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **408** | **3327** |
| **Control** | **145** | **3477** |

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

I picked the states AZ,AK,CA,AL and AR.

State = AK,

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy 1** | **No Buy 0** |
| **Treatment 1** | **0** | **21** |
| **Control 0** | **0** | **20** |

State = AL,

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **0** | **51** |
| **Control** | **0** | **38** |

State = AR,

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **0** | **31** |
| **Control** | **0** | **29** |

State = AZ

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **0** | **41** |
| **Control** | **0** | **33** |

State = CA

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy 1** | **No Buy 0** |
| **Treatment 1** | **0** | 35 |
| **Control 0** | **0** | **30** |

**III. Data Cleaning:**

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

Produce an Excel File with the following columns

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

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

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

**IV. Statistical Analysis**

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

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

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

And Report the output.

**Call**:

lm(formula = CleanData$Outcome ~ CleanData$Test.Variable)

**Residuals**:

Min 1Q Median 3Q Max

-0.06831 -0.06831 -0.01863 -0.01863 0.98137

**Coefficients**:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.018634 0.003405 5.472 4.6e-08 \*\*\*

CleanData$Test.Variable 0.049676 0.004806 10.337 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2027 on 7112 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.0148, Adjusted R-squared: 0.01466

F-statistic: 106.9 on 1 and 7112 DF, p-value: < 2.2e-16

Output -

alpha = 0.018634

Beta = 0.049676

**Outcome** = 0.018634 + 0.049676 \* Test\_Variable + error

Interpreting the alpha = 0.018634 - his coefficient as the *intercept* of the regression line

It means that in the absence of any Test Variable (Test or control), we can still expect an outcome (buy/no buy) of 0.018634 on average.

Adjusted R-squared: 0.01466 - *R*-squared measures the overall quality of a regression model. Our model explains 1.47% of variality in the 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.

In the ideal world, all statistical hypotheses would be tested on entire populations. However, this is often impractical or impossible, so you typically examine a random sample from the population.

Hypothesis testing is defined as the process to which you can test a claim about a population parameter.Because we don't have data for every user that will visit their two landing pages (which would enable us to calculate the true population parameter), we take a sample, calculate a sample statistic, and then determine the likelihood that our sample statistic could have occured, assuming our null hypothesis regarding the population parameter is true.

The null hypothesis:

• The statement about the population parameter that is assumed to be true, it's usually the hypothesis that the sample statistic observed was purely from chance.

The alternative hypothesis:

• The alternative hypothesis directly contradicts the null and is usually the hypothesis that the sample statistic was influenced by some cause.

**Our null hypothesis:** the difference of averages between test variable T (mu T) and control variable C (mu C) is equal to 0

**Our alternative hypothesis**: the difference of conversion rates between test variable T (mu T) and control variable C (mu C) is greater than 0.

Null hypothesis : mu(T) - mu(C) = 0

= (Avg. Y/ X1) - ( Avg Y / X0)

= (alpha + Beta) - Beta = 0

= alpha = 0

alpha = 0.018634

p-value: < 2.2e-16

In this case, alpha > p-value

If your p-value is less than your alpha, that means you reject your null hypothesis and conclude that the test statistic observed was most likely not due to chance or, within the accepted low probability of it being chance.

**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?

**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)

**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?

**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**

**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**

**Expected Performance:**