**PROJECT REPORT**

**MIS 6392 Causal Analytics and A/B Testing**

**CUSTOMER RETARGETING- MIGHTY HIVE**

**Members :**

Dhivakar Rangarajan - DXR180020

Naveenkumar Jayakumar - NXJ180010

Mukilraj Kannan- MXK180040

Murugan Krishnakumari- MXK180051

**ABSTRACT:**

The purpose of this experiment is to test the effectiveness of retargeting the customers who have not really shown interest to buy the campaigns( vacation package ) through an agency. There are many uses of retargeting as a strategy and also has its own share of disadvantages. In this experiment, we propose the idea that there will be a positive effect when we introduce the strategy of retargeting. So as to find the true effect of retargeting the customers, they are randomly split into control and treatment group and the experiment was conducted.

It was hypothesized that the phenomenon of retargeting of customers may or may not be effective and we have to find statistical evidence that it does for our case of analysis. We ran multiple regression analysis with multiple outcome variables to find the statistical significance and also the size of the effect ultimately.

Data was downloaded from **MightHive**, a leading data and digital media consultancy that helps marketers take control and boost sales through retargeting the customers. We need to know if there is a statistical difference between the customers that reserved from the treatment group and the ones in the control group. We need to conduct a A/B test and our Null and Alternate Hypothesis are given below:

**Null**: The difference of Reservation rates between treatment and control group is equal to 0.

**Alternate**: The difference of Reservation rates between treatment and control group is greater than 0.

We could find the evidence that the retargeting has been proved effective based on the data that we have. A much robust data would have helped us to do more statistical tests which will help attain to a very solid conclusions regarding the effect of retargeting formulated on various aspects.

**INTRODUCTION:**

As a business analyst giving answers to question is not enough, the goal of an analyst is to discover useful information, informing conclusions and supporting decision-making. Hence without analyzing data we just have an opinion of what’s better. The problem statement for the Martin’s travel agency is having a bad conversion rate of turning leads into customers and the higher-level management decided to proceed with Mighty Hive as a source to retarget customers to improve the conversion rate. Again as an analyst the task is to find statistical significance and find the cause of a boost in sales due to retargeting the customers

Retargeting customers who have already engaged with the business is an effective business and marketing strategy. The reasons for retargeting customers is to:

1)Builds brand awareness by capitalizing these audience.

2)Producing relevant marketing strategy for customers based on their past interaction with the business produces positive results in terms of brand awareness.

3)Converting potential leads on first visit is very challenging and retargeting the potential customers to convert them to a sale would be more sensible in a business point of view.

The difference between retargeting and remarketing is in its actual strategies. Retargeting is mostly about serving ads to potential customers based on cookies while remarketing is usually based on email. Remarketing works by collecting the information of users and creating lists, which are used later to send sales emails. Hence Mighty Hive’s retargeting strategy is chosen here.

As the company had just a limited data of people who made calls to the call center and abandoned the calls retargeting was a better strategy which was selected to be implemented. But before we implement its traditional do to tests from the obtained data to conclude whether the strategy is worth implementing or not. As the problem stated we felt that A/B testing on the data would be right choice as we wanted to see the effects of retargeting.

The data from mighty hive was divided into two sets.

1. The Abandoned Dataset: Observations in the Abandoned Dataset are individuals who called into Martins Travel Agency's call center but did not make a purchase.
2. The Reservation Dataset: Observations in the Reservation Dataset are the customers who called into Martins Travel Agency's call center and made a reservation.

Now our task is to match the fields in the two files and try to understand a causal inference between retargeting and not using any ad-algorithms of mighty hive. Also, to analyze various interaction effects using the given data.

**EXPERIMENT DESIGN:**

**MightyHive** is an advertising technology company that uses retargeting methods to send ads to users online. One product, Call Center Remarketing, uses call center log data to retarget those consumers online whom did not make a purchase. The results of the advertising campaign for Martin's Travel Agency are given in the following two datasets:

The Abandoned Dataset: 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: Observations in the Reservation Dataset are the customers who called into Martins Travel Agency's call center and made a reservation. The schema for both of the datasets is provided below:

C Caller\_ID - A unique ID generated for every incoming phone call to the call center

Session - The Year/Month/Day/Time of each incoming phone call to the call center

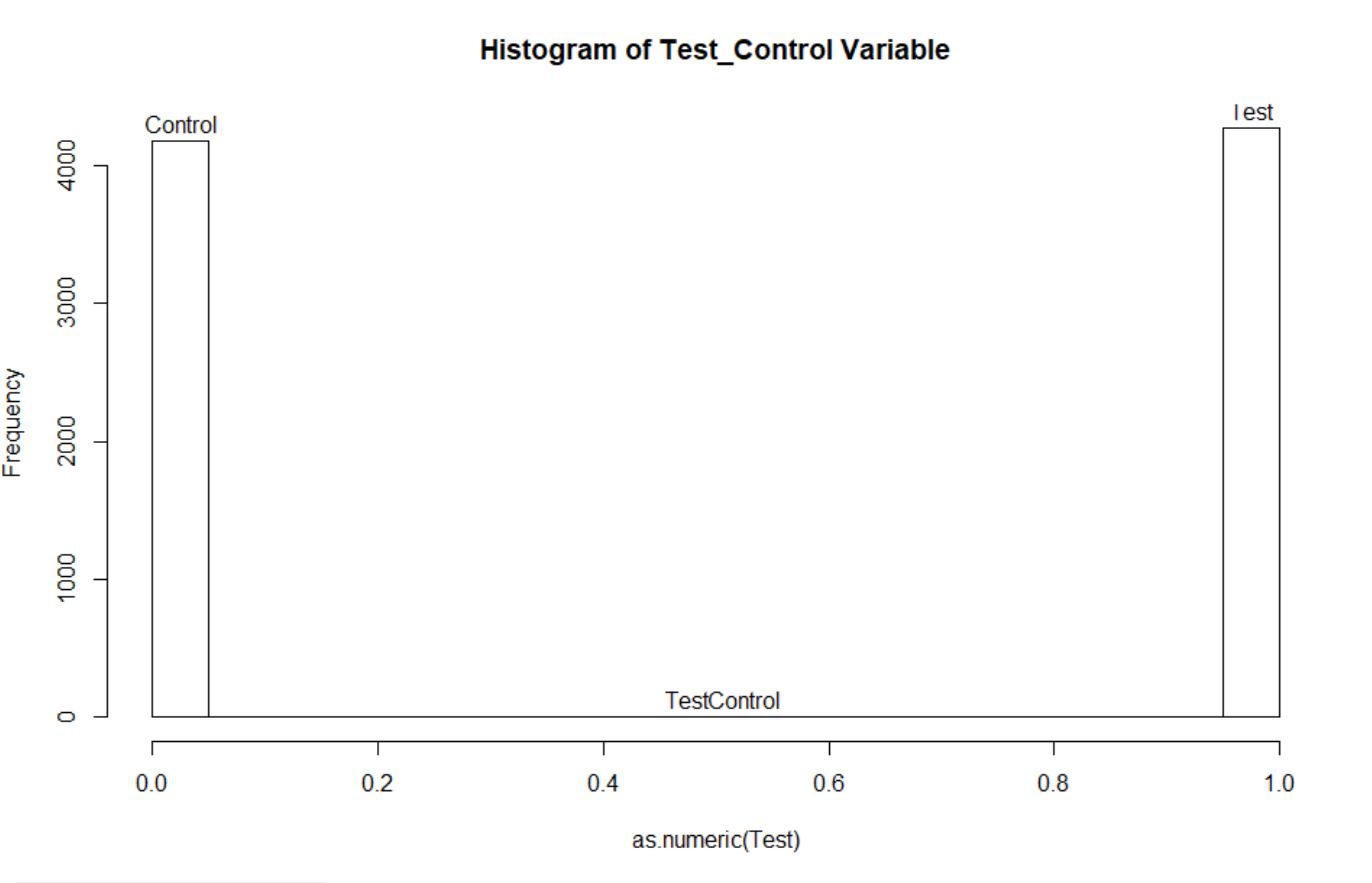
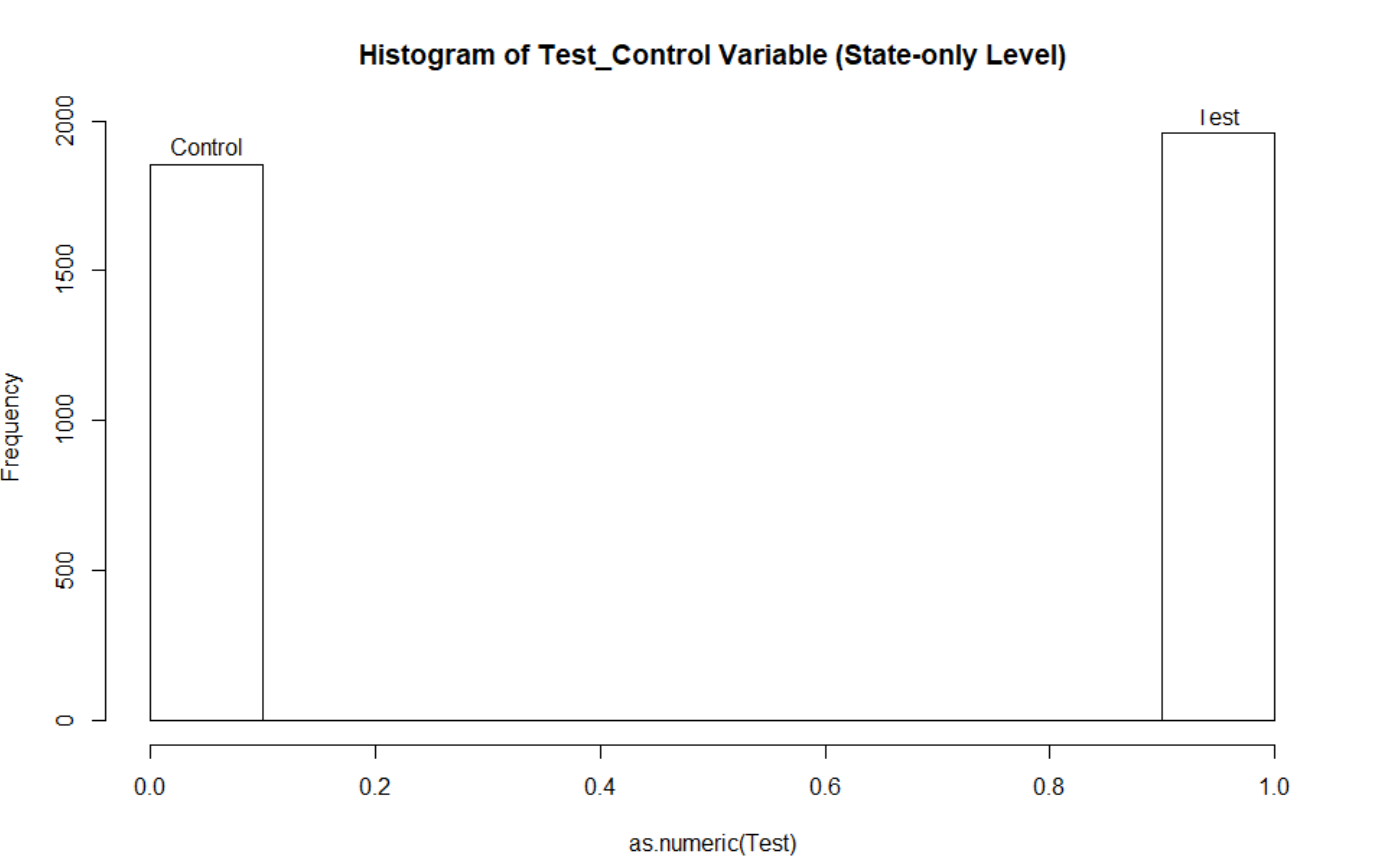
Incoming\_Phone - Phone number identified using caller indentification

Contact\_Phone - Phone number the caller submits

Test\_Control - Experiment tag

**To verify the proper distribution of data**

In order to verify the proper execution of the experiment i.e. to check the test and control variables are distributed randomly and uniformly in the entire dataset we took summary statistics for a test variable “State-only” and compared this with the entire dataset.

From the above results we can be observe that the summary statistics for the test variable at the state-only level are almost same as the entire dataset hence we don’t have any obvious imbalance in the dataset.

**DATA COLLECTION STRATEGY**

The experiment was conducted by MightyHive based on the data gathered from Martin’s Travel 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 thorough analysis would be good to carry out between the ones those bought the package and the ones who turned down before we can retarget.

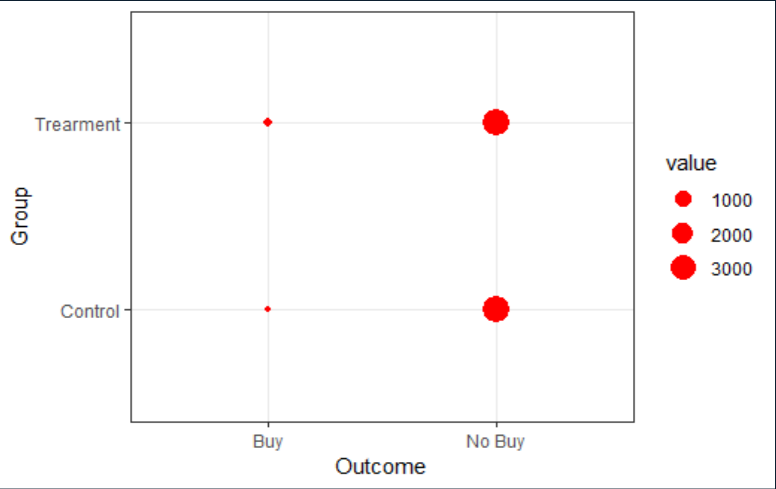
An experiment is run, where customers in the abandoned dataset are randomly placed in a treatment or in a control group (see column L in both files). Those marked as “test” are retargeted (treated), the others marked as control are part of the control group.

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.

**Data Matching**

In order to obtain the outcome variable of buy/notbuy we followed the following matching procedure:

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

Final dataset:

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

**DATA DESCRIPTION AND STATISTICAL ANALYSIS** - 2 pages

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

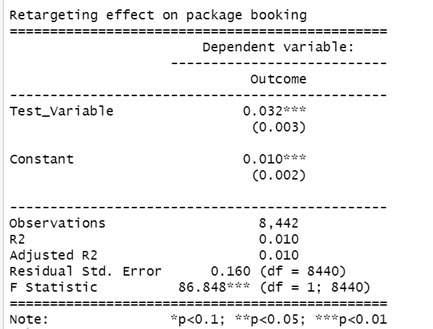
Such type of data collection also poses the problems in matching as some observations are not matched or give more than two matches. Some customers’ “outcomes” are not present but only one outcome for customers is present who bought the service. Also, a lot of observations in the RD dataset are not present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

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

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

**ANALYSIS**

1. **Regression model Outcome = alpha + beta \* Test\_Variable + error**

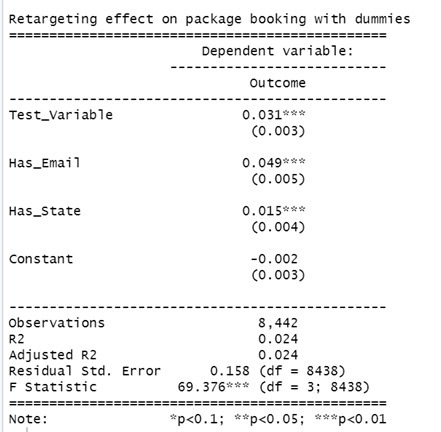


The above regression model is not well specified. As a test we ran this regression and find out how much we can rely just on the test variables. Here we get a coefficient value of 0.03 that is the dependent variable “Outcome” would increase 3% for every addition of a customer in the Test group which is very less.

The R-sqaured values is around 0.01 which is also a extremely small number and would'nt really explain the changes to the outcome variable. To make the test more robust we need to add more variables to figure out the significance of this model.

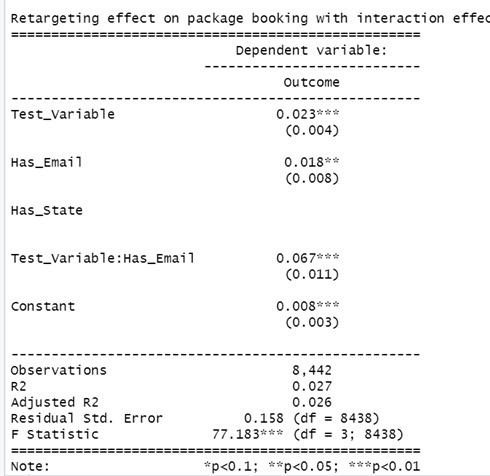
1. Here we applied regression model the dummies for State and Emails. Also included interactions with the treatment, namely between email and retargeting

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



The adjusted R-squared has now increased to 0.024 after using the dummy variables - Has Email and Has State. Hence compared to the first model, this performs better. But this would be not enough to come to a conclusion on a causal effect as there are a lot of other variables which are not controlled by this model. Incorporating all the other features along with this model would capture the other effects related to the conversion variable.

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

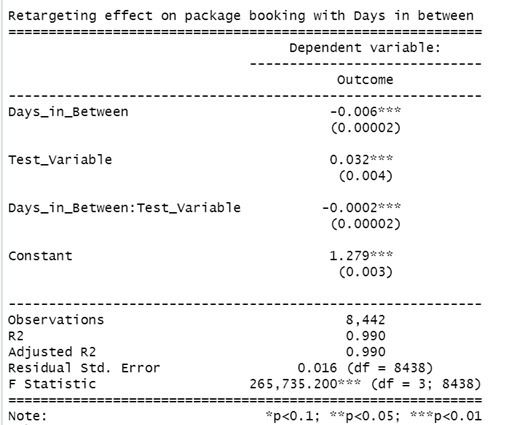


When interactions are included with the treatment group the Test Variable and Has Email, also including the State variable gives a better explanation for the error term. With the available variables even though the error has not been completely explained we are coming towards a better explanation for the outcome variable.

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.

1. Investigating whether the response time (time to make a purchase after the first contact) is influenced by the retargeting campaign. For obtaining this term we have created a calculated field to obtain the days in between abandoned call and purchase.

**Model-4: Outcome = alpha + beta \* Days in between: Test Variable + error**

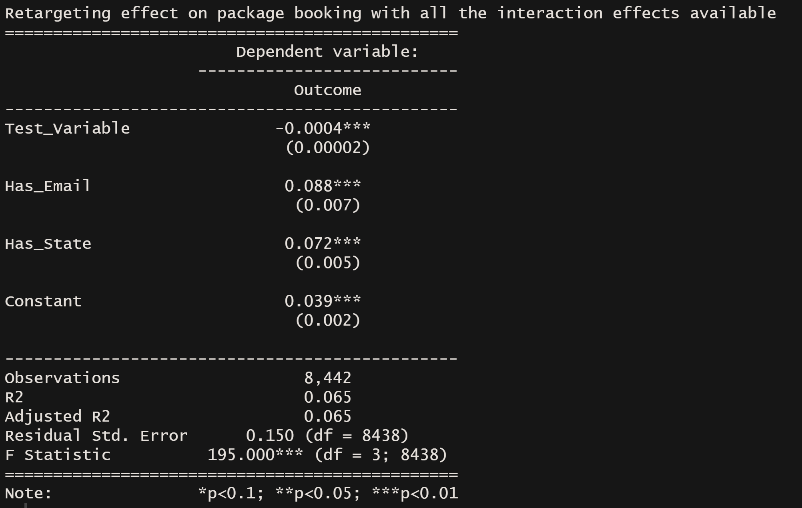
Here a R-squared is observed to be 0.99, this linear regression model fits well. Also, the error term has also significantly reduced to come to good conclusion about the effect of days in between’s interaction with the test variable when regressed with the outcome variable.

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 negative coefficient of Days in Between means that as the days in between the abandoned call and calling again increases for each customer, the chances of making a sale reduces. Hence it is inversely proportional to outcome.

1. For the final regression a model we calculated the total retargeting effect on package booking with all the interaction effects between the variables and the test variable

**Model-5: Outcome = alpha + beta1 \* Test Variable \* Has Email \*+ beta2 \* Has State: Test Variable + beta3\*Test\_variable +error**

In this model we find a very low error as we can see that the interaction effect of days in-between and test variable has a negative effect as more days are taken between the next call the more chances of not getting the call converted into a sale. While other interaction effects has a positive effect along with the test variable we can conclude that the retargeting is effective in converting an abandoned call into a sale.

**CONCLUSION:**

From the above test results conducted for various outcome variables, we could provide evidence and thereby conclude that the retargeting has indeed proved itself effective and also efficient in improving the rate of customers in buying the vacation packages who have previously declined to buy. We could show statistical significance and also the effect of retargeting was captured and formulated.

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

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

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

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