

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: Sree Ranga Vasudha Moda

Section: Tuesday Morning

Signature (if possible) M.S.R.Vasudha

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.

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.

Answer:

Having worked on business application and involved in decision making of business strategies, I believe retargeting the customers would definitely yield good results. I can justify this with the following reasons.

- When a customer calls up for the travel agency, that itself means he/she would have definitely had interest in any of the packages the company was offering. So, retargeting them would provoke their interest and gives them chance to know more details about the products which eventually may make them purchase.
- Also, this gives more space for the customers to negotiate on the deals they discussed in the previous call and if the deals are win-win, the talks would ultimately be gaining new customers.
- And also, it builds an image on company that the company values the customers most and their initiatives to make the new callers as actual customers and also that they care to know about their perspectives and concerns.

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.

Ans:

- First import two datasets into two tables in SQL Server
- Run the following query to get the data of customers who actually converted from visitors to customers.

Query:

```
select a.* from ABANDONED1 a , RESERVATION r where a.email=r.email or a.IPHONE=r.IPHONE or a.CPHONE=r.CPHONE
```

- Export the output to a new Excel sheet and save it.
- Now VLOOKUP the concatenated values of FirstName+Email+Contact_Phone in Abandoned dataset and Reservation dataset and set 1 for a match and 0 for a mismatch.
- Now name this new matched column as Test_Variable and save the abandoned dataset.
- Put another column for fetching reservation set Date using VLOOKUP and using NETWORKDAYS function filter days below 0. Save it.
- Now load this excel into R following below steps.

```
setwd("C:/MyData")
sqled_merge=read.csv("Cleaned SheetCSV.csv")
> summary(sqled_merge$Test_Variable)
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
0.0000 0.0000  1.0000  0.5053  1.0000  1.0000

> quantile(CleanedData$Test_Variable,c(0.05,0.95))
5% 95%
0  1
Hence, 5th quantile=0
      95th quantile=1

> sd(CleanedData$Test_Variable)
[1] 0.5000012
Standard Deviation=0.5000012
```


Overall Data summary:

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

State-Filtered Test_Variable Summary:

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

They are approximately the same in both the cases, but the mean is slightly closer to median in second case. So, it is more normalized data.

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.

Ans:

From the merging of the data, which came from intersection of Reservation data and abandoned data, filter the entire datasets accordingly to get four important numbers to fill the following table.

Total observations under Test: 4266

Total observations under Control: 4176

	Came Back	Dint come back
Test	335	3931
control	85	4091

Among these 3931 who were targeted and dint come back, 3576 persons' contact numbers or incoming number weren't logged which would give scope on the genuinity of the data.

Hence, it is contradicting the statement that “We do not observe some “outcomes” for some customers”.

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)

Yes , customers can be matched in both the datasets using some datakeys.
I have taken the datakeys like “EMAIL”,”Incoming_Phone”,”Contact_Phone” and run a query in SQL to get the joint set.

```
select a.* from ABANDONED1 a , RESERVATION r where a.email=r.email or  
a.IPHONE= r.IPHONE or a.CPHONE=r.CPHONE
```

Reason:

EmailId,Incoming_Phone and Contact_Phone would be unique and the highest means to validate the identity of the customer.

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:

- First import two datasets into two tables in SQL Server
- Run the following query to get the data of customers who actually converted from visitors to customers.

Query:

```
select a.* from ABANDONED1 a , RESERVATION r where a.email=r.email or  
a.IPHONE= r.IPHONE or a.CPHONE=r.CPHONE
```

- Export the output to a new Excel sheet and save it.
 - Now VLOOKUP the concatenated values of FirstName+Email+Contact_Phone in Abandoned dataset and Reservation dataset and set 1 for a match and 0 for a mismatch.
 - Now name this new matched column as Test_Variable and save the abandoned dataset.
 - Put another column for fetching reservation set Date using VLOOKUP and using NETWORKDAYS function filter days below 0.Save it.
 - Now select entire data and enable filter.
 - Filter Test_Variable and Test_Control variables for various combination to obtain following data.
-
- Total observations under Test: 4266
 - Total observations under Control: 4176

	Came Back	Dint come back
Test	335	3931
control	85	4091

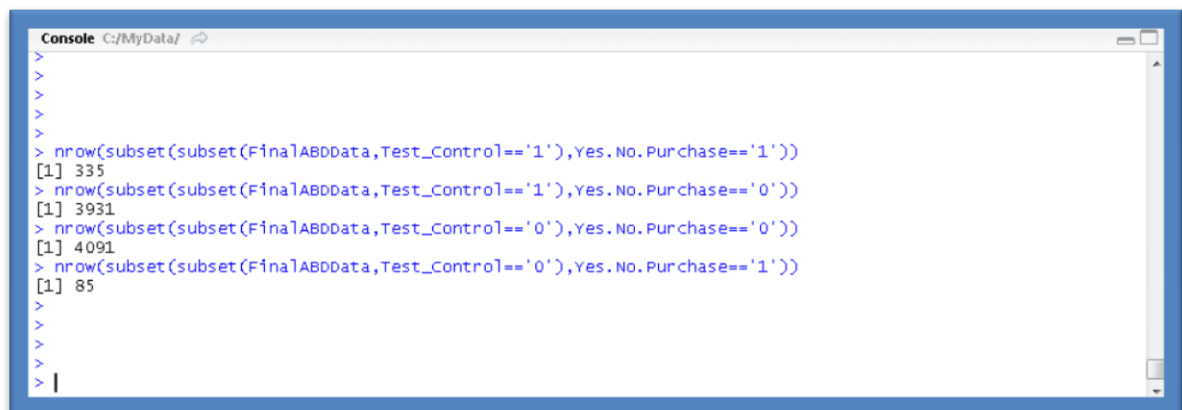
Q8: Are there problematic cases? i.e. data records not matchable? If so, provide a few examples and toss those cases out of the analysis.

I tried to match the data using EMAIL,Incoming_Phone and Contact_Phone data keys.I haven't faced any problem in the data match.

Q9: Complete the following cross-tabulation:

Group \ Outcome	Buy	No Buy
Treatment	335	3931
Control	85	4091

Answer:



```

Console C:/MyData/
>
>
>
>
> nrow(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='1'))
[1] 335
> nrow(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='0'))
[1] 3931
> nrow(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='0'))
[1] 4091
> nrow(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='1'))
[1] 85
>
>
>
>
> |

```

```

nrow(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='1'))
[1] 335
> nrow(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='0'))
[1] 3931
> nrow(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='0'))
[1] 4091
> nrow(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='1'))
[1] 85

```

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

State: AK

Group \ Outcome	Buy	No Buy
Treatment	3	26
Control	0	32

```

Console C:/MyData/
>
>
>
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='1'),Address=='AK'))
[1] 3
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='0'),Address=='AK'))
[1] 26
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='0'),Address=='AK'))
[1] 32
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='1'),Address=='AK'))
[1] 0
>
>
>
>
>
>
>
>
>
>

```

State: AL

Group \ Outcome	Buy	No Buy
Treatment	2	36
Control	0	42

```

Console C:/MyData/
>
>
>
>
>
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='1'),Address=='AL'))
[1] 0
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='1'),Address=='AL'))
[1] 2
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='0'),Address=='AL'))
[1] 36
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='0'),Address=='AL'))
[1] 42
>
>
>
>
>
>
>
>
>
>

```

State: DE

Group \ Outcome	Buy	No Buy
Treatment	3	31
Control	0	46


```

Console C:/MyData/
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='0'),Address=='AL'))
[1] 36
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='0'),Address=='AL'))
[1] 42
>
>
>
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='1'),Address=='DE'))
[1] 3
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='0'),Address=='DE'))
[1] 31
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='0'),Address=='DE'))
[1] 46
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='1'),Address=='DE'))
[1] 0
> |

```

State: FL

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

```

Console C:/MyData/
>
>
>
>
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='1'),Address=='FL'))
[1] 0
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='1'),Address=='FL'))
[1] 3
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='1'),Yes.No.Purchase=='0'),Address=='FL'))
[1] 35
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='1'),Address=='FL'))
[1] 0
> nrow(subset(subset(subset(FinalABDDData,Test_Control=='0'),Yes.No.Purchase=='0'),Address=='FL'))
[1] 37
>
>
> |

```

State: OH

Group \ Outcome	Buy	No Buy
Treatment	4	46
Control	0	39

Query:

```
select a.* from ABANDONED1 a , RESERVATION r where a.email=r.email or  
a.IPHONE= r.IPHONE or  
a.CPHONE=r.CPHONE
```

- Export the output to a new Excel sheet and save it.

2.) Data Matching with RESERVATION dataset:

- Now VLOOKUP the concatenated values of FirstName+Email+Contact_Phone in Abandoned dataset and Reservation dataset and set “1” for a match and “0” for a mismatch into a “OUTCOME” column.
- Now name this new matched column as Test_Variable and save the abandoned dataset.

3.) Computing difference between the days of two calls for matched entries.

- Put another column for fetching reservation set Date using VLOOKUP and using NETWORKDAYS function filter days below 0 and assign a value of 200 for all NA values. Save it.

Using only R:

Importing two datasets into R,

- Using match function, match the EMAIL ,Incoming_Phone and Contact_Phone column data and load the matched values to new dataset matches and bind cumulatively with everymatch.
- Remove duplicates in the new dataset using Caller_Id.
- Then matching caller_Id ,get the date column of RES into new dataset with new column
- Get the difference of the dates using PosixCT command and put in a column “Days in Between”
- Then assign all Test values to 1 and control values to 0.
- Similarly do for Outcome(Yes/No purchase).
- Create two dummy variables EMAIL_available and Address_Available accordingly for NA and non-NA values of the columns.

Matching the data:

```
match(ABD_data$ EMAIL, RES_data$ EMAIL, nomatch =0)  
similarly , for Incoming_Phone and Contact_Phone  
and merge using  
rbind(dataset1,dataset2) function
```

Creation of Dummy Variables:

```
> Dummy_Email=!is.na(New_Data$Email)
> as.numeric(Dummy_Email)

> Dummy_Address=!is.na(New_Data$Address)
> as.numeric(Dummy_ Address)
```

Put Na values in Days_in_between to 200

```
> for(i in 1:8442)
+ {
+   if(is.na(FinalABDDData$DAYS_IN_Between[i]))
+     FinalABDDData$DAYS_IN_Between=='200';
+ }
```

IV. Statistical Analysis

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

Q11: Run a Linear regression model for

$$\text{Outcome} = \alpha + \beta * \text{Test_Variable} + \text{error}$$

And Report the output.

Answer:

Outcome=0.020354 +058173* Test_Variable +0.2155

```
> LR_1=lm(CleanedData$Outcome~CleanedData$Test_Variable)
> summary(LR_1)
```

Call:lm(formula = CleanedData\$Outcome ~ CleanedData\$Test_Variable)

Residuals:

Min	1Q	Median	3Q	Max
-0.07853	-0.07853	-0.02035	-0.02035	0.97965

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.020354	0.003335	6.104	1.08e-09 ***
CleanedData\$Test_Variable	0.058173	0.004691	12.401	< 2e-16 ***

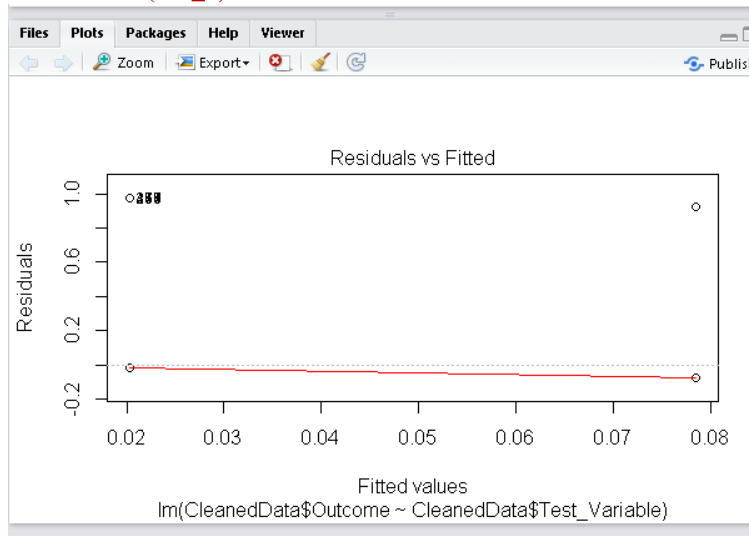
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2155 on 8440 degrees of freedom

Multiple R-squared: 0.01789, Adjusted R-squared: 0.01778

F-statistic: 153.8 on 1 and 8440 DF, p-value: < 2.2e-16

- Plot(LR_1)



Since we have very low R values, this doesn't predict a good linear regression model.

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:

In the Team Leada 4th module,

Hypothesis testing is defined as the process to which you can test a claim about a population parameter.

So correlating it with this project, we can say that the population parameter is the difference between the conversion rates of Tested and control groups.

So, we can state Null and alternate hypothesis as follows:

Null Hypothesis: $P1 - P2 = 0$

Alternate Hypothesis: $p1 - p2 > 0$

Where

p1: the conversion rate of TEST group.(Outcome=1)

p2: the conversion rate of CONTROL group.(outcome=1)

Hence from the data matching done above,

$P1 = 355$

$P2 = 85$

As $P1 - P2 = 355 - 85 \neq 0$.

And also here from linear regression result down, α is greater than P value which rejects null hypothesis.

Alternate Hypothesis is proven.

=> Retargeting customers is fruitful.

Argument for "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. Randomizing sampling in a proper way is very important due to following reasons:

- A random sample will be representative of the entire population
- Random sampling avoids to be biased by giving all individuals an equal chance to be chosen."
- The mathematical theorems which justify most frequent statistical procedures apply only to random samples.

Hence if randomization happens taking care about these factors, there is more chance of the result matching to the result when entire population is taken into picture.

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:

For a good regression model, R-Squared and Adjusted R-Squared would be high.

From the linear regression results of R,

Multiple R-squared: 0.01789, Adjusted R-squared: 0.01778

Which are very low, which indicates this is a poor linear regression model.

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)

```
> LR_2=lm(CleanedData$Outcome~CleanedData$Test_Variable+CleanedData$Email_Availability+CleanedData$Address_Availability)
> summary(LR_2)
```

Call:

```
lm(formula = CleanedData$Outcome ~ CleanedData$Test_Variable +
    CleanedData$Email_Availability + CleanedData$Address_Availability)
```

Residuals:

Min 1Q Median 3Q Max

-0.11857 -0.06601 -0.06136 -0.00880 0.99120

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.008801	0.003942	2.232	0.0256 *
CleanedData\$Test_Variable	0.057211	0.004681	12.221	< 2e-16 ***
CleanedData\$Email_Availability	0.035477	0.007336	4.836	1.35e-06 ***
CleanedData\$Address_Availability	0.017078	0.004823	3.541	0.0004 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

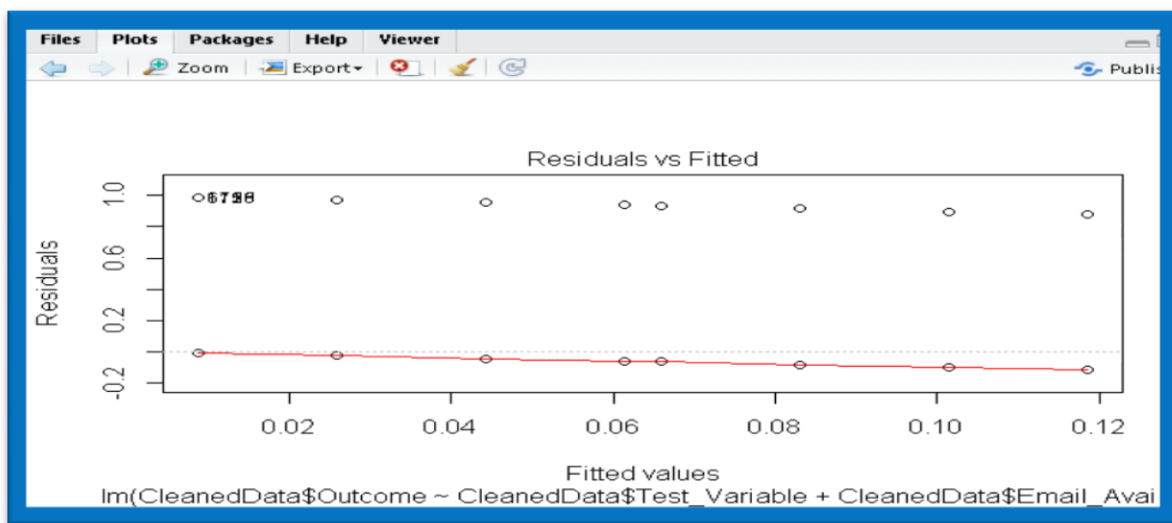
Residual standard error: 0.2149 on 8438 degrees of freedom

Multiple R-squared: 0.0232, Adjusted R-squared: 0.02285

F-statistic: 66.8 on 3 and 8438 DF, p-value: < 2.2e-16

Here R-Squared and Adjusted R-Squared are atleast bigger when compared to previous model covering more uncertainty, Hence this is a better model than the previous one.

Plot(LR_2):



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:

Here, I have taken two variables for Linear regression as the R-square values are high when two variables(outcome and Test_Variable) are taken when compared to only one variable(Test_Variable) which predicts good model.

```
> LR_4=lm(CleanedData$Days_In_Between~CleanedData$Test_Variable+CleanedData$Outcome)
> summary(LR_4)
```

Call:

```
lm(formula = CleanedData$Days_In_Between ~ CleanedData$Test_Variable +
    CleanedData$Outcome)
```

Residuals:

Min	1Q	Median	3Q	Max
-30.820	-0.044	0.042	0.042	35.180

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	199.95787	0.03931	5087.109	<2e-16 ***
CleanedData\$Test_Variable	0.08598	0.05567	1.544	0.123
CleanedData\$Outcome	-165.22407	0.12802	-1290.638	<2e-16 ***

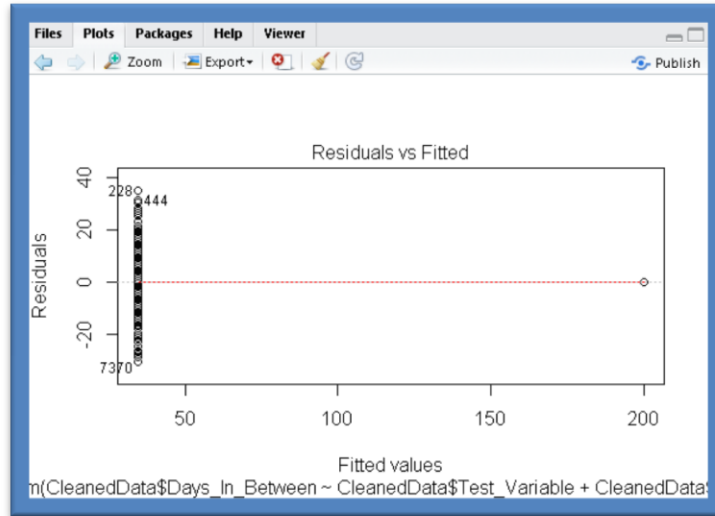
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.534 on 8439 degrees of freedom

Multiple R-squared: 0.995, Adjusted R-squared: 0.995

F-statistic: 8.478e+05 on 2 and 8439 DF, p-value: < 2.2e-16

With the high R-Squared values, covering most uncertainty, this seems to be like a good model.



Plot(LR_4)

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

- **To use R more efficiently than SQL and Excel.**
- **To best fit the data keys for Linear Regression analysis.**

I understood this analysis would definitely help to shoot the results in formulating better business strategies.

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:100%

Expected Performance:100%