**Midterm Project : The Advertising Campaign Analysis**

Rithika Kandimalla

**1 Business Justification**

1. **Explain why retargeting customers who initially didn’t buy a package makes business sense.**

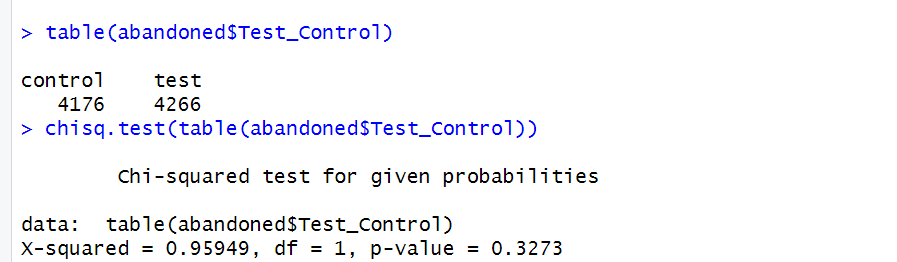
* These people have already shown the interest in the package but they didn’t buy, but there's still a good chance they might change their mind if reminded.
* It's usually cheaper to try to sell to these interested customers again than to find entirely new ones.
* By studying why they didn't buy initially, we can learn more about what they're looking for and what might convince them to purchase in the future.
* Even if they didn't buy right away, they may think of our agency when they're ready to book a vacation.
* If competitors are also targeting similar customers, retargeting can help maintain or regain a competitive edge.

1. **Analyze the test/control division. Does it seem well-executed?**

**Input:**

****

**Output:**

****

From the above test results we can observe that

* The observed distribution of customers between the "control" and "test" groups is as follows:

Control group: 4,176 customers

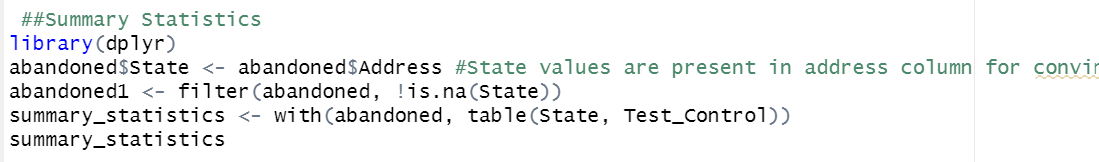
Test group: 4,266 customers

* Chi-squared Value: X-squared = 0.95949
* Degrees of Freedom (df): 1
* P-value: 0.3273

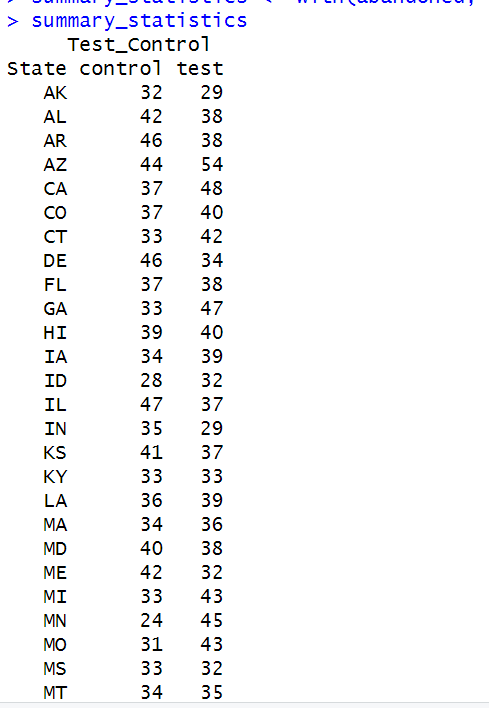
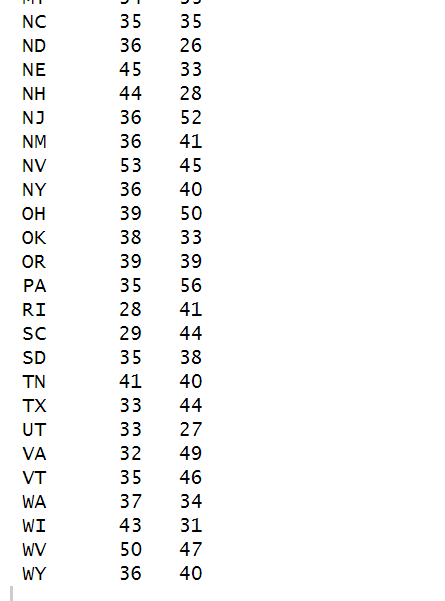
Here the p-value is 0.3273, which is greater than 0.05. This means we don’t have strong evidence to reject the null hypothesis. The null hypothesis in this case states that there is no significant difference in the distribution of customers between the test and control groups. The distribution seems to be reasonably well-executed, and the differences observed are not statistically significant

1. **Compute summary statistics for the test variable, segmenting by available State data**

**Input:**

****

**Output:**

**** ****

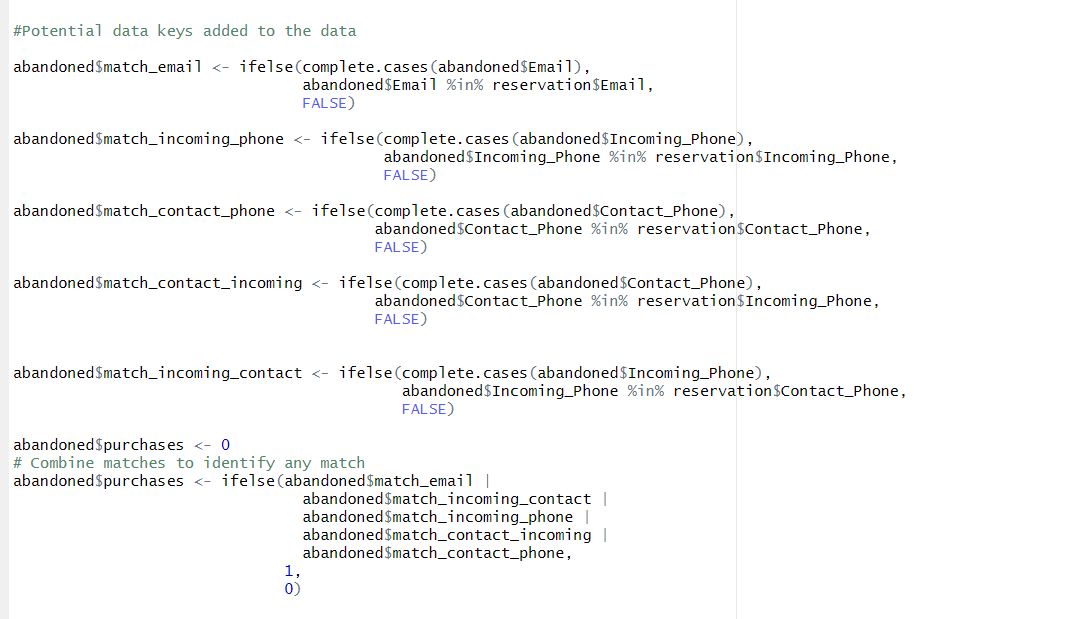
**2.Data Alignment**

1. **From your examination of both files, propose potential data keys to match customers.**

* Email Id
* Incoming\_Phone
* Contact\_Phone are the potential data keys to match the customers.

Below is the code for the match based on above keys and logical vectors for each condition.

**Input:**

****

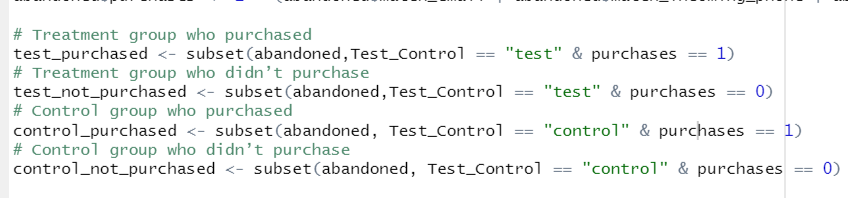
1. **Detail your procedure to identify customers in:**

* Treatment group who purchased.
* Treatment group who didn’t purchase.
* Control group who purchased.
* Control group who didn’t purchase.

After the matches were found evaluate adandoned$purchases based on matches i.e assign 1 if the record in abandoned dataset matches with the records in the reservation dataset ,assign 0 other wise.

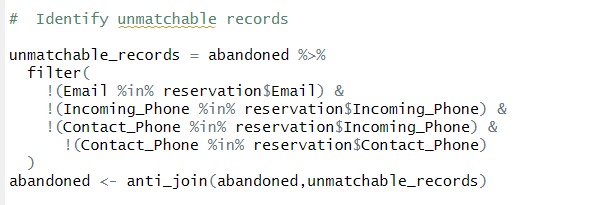
Now subset the abandoned data set based on test and control values from the Test\_Control column and purchases column which contains value 1 for match customers and 0 for others.Below is the Code

**Input:**



1. **Are there unmatchable records? If yes, provide examples and exclude them from the analysis.**

Yes there are unmatchable records, I . e the values with no data in the columns below is the code to check the unmatched records and duplicates in reservation dataset(multiple records matching with single record in the abandoned dataset-a customer may purchase multiple records)

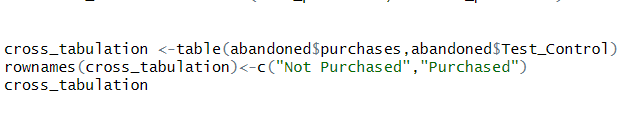


There are 782 unmatchable records in abandoned data set.

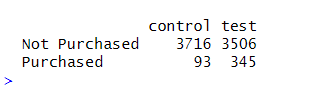
By removing these records there are 7660 records in abandoned dataset which we will be considering for further analysis

1. **Provide a cross-tabulation of outcomes for treatment and control groups.**

**Input:**

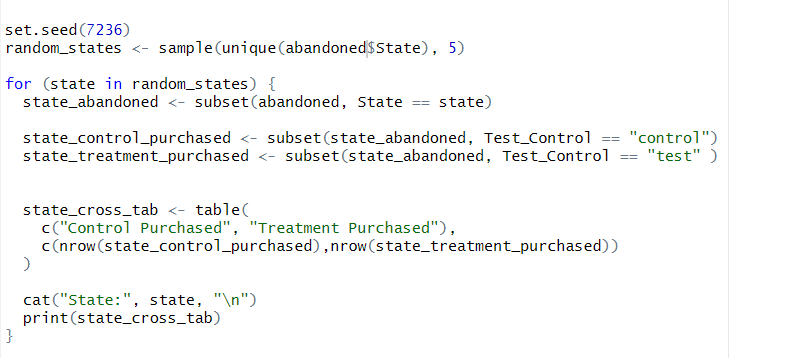
****

**Output:**

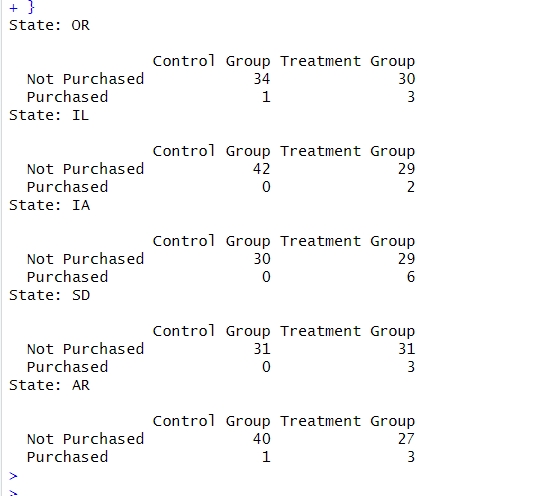
****

1. **Replicate the cross-tabulation for five randomly chosen states, detailing your selections**.

Input:

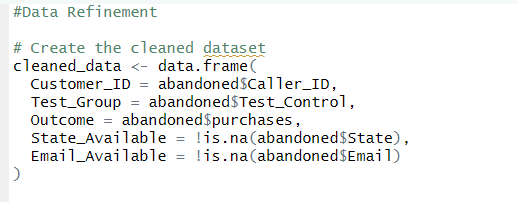


Output:



**3.Data Refinement**

**Input :**

****

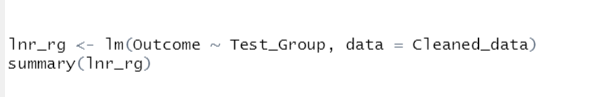
**Output:**

Attached the document

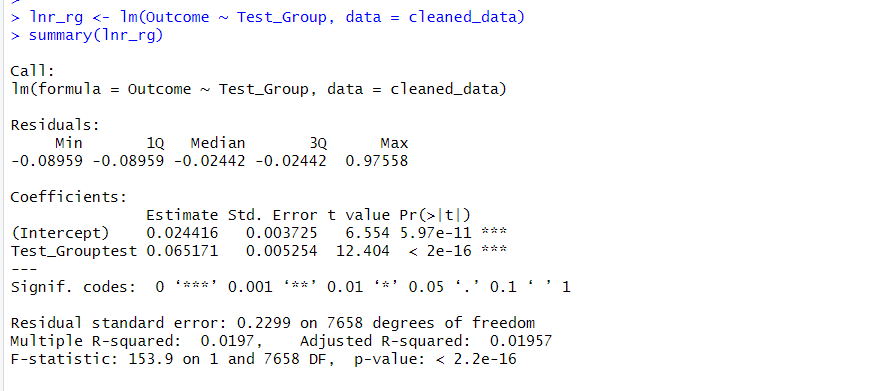
**Statistical Assessment**

1. **Execute a linear regression for the formula: Outcome = α + β \* Test Group + error. Share the results.**

**Input:**

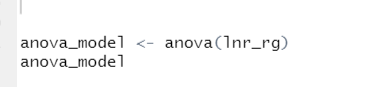
****

**Output:**

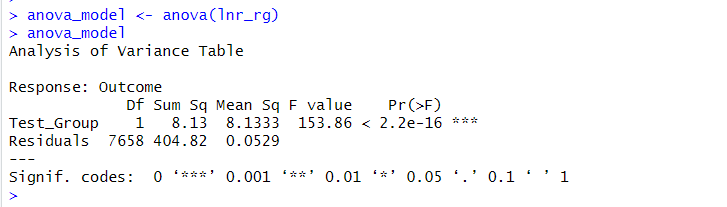
****

1. **Justify that this regression is statistically comparable to an ANOVA/t-test.**

**Input:**

****

**Output:**

****

The p-value associated in the regression test and the p-value in Anova test is same. This confirms that the regression result is statistically comparable to the ANOVA test.

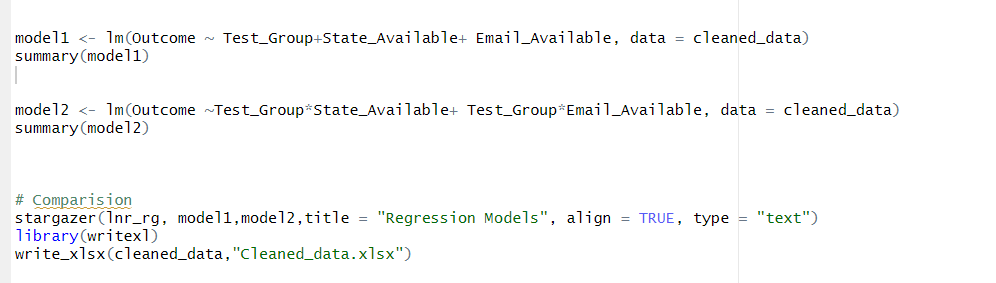
In summary, both the regression result and the ANOVA test indicate that Test\_Group is statistically significant in explaining the variation in Outcome. The p-values are consistent between the two tests**.**

1. **Debate the appropriateness of the regression model in making causal claims about the retargeting campaign’s efficacy.**

* The model indicates that the 'Test\_Group' variable is statistically significant (p-value < 0.05), suggesting that there is some evidence of an association between being in the test group and the likelihood of making a purchase.
* The adjusted R-squared value even its low(0.01957) explains that about 1.957% of the variation in the outcome variable is "Test\_Group."
* The coefficient for Test\_Grouptest is 0.0652. This implies that being in the "test" group is associated with an average increase in the outcome by 0.0652 units, compared to the control group.
* The F-statistic is 153.9 with a very low p-value (< 2.2e-16), suggesting that the model as a whole is statistically significant.
* Even though the "Test\_Group" coefficient is statistically significant, we cannot confirm that the retargeting campaign caused the observed changes

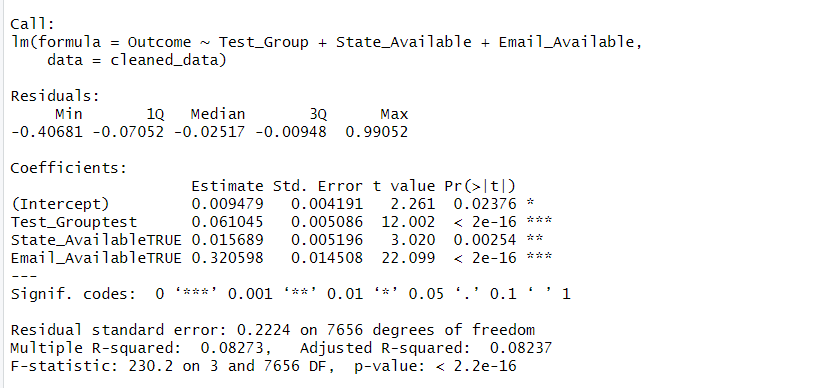
1. **Integrate State and Email dummies into the regression. Also consider interactions with the treatment group. Compare these results to the previous regression and provide insights.**

**Input:**

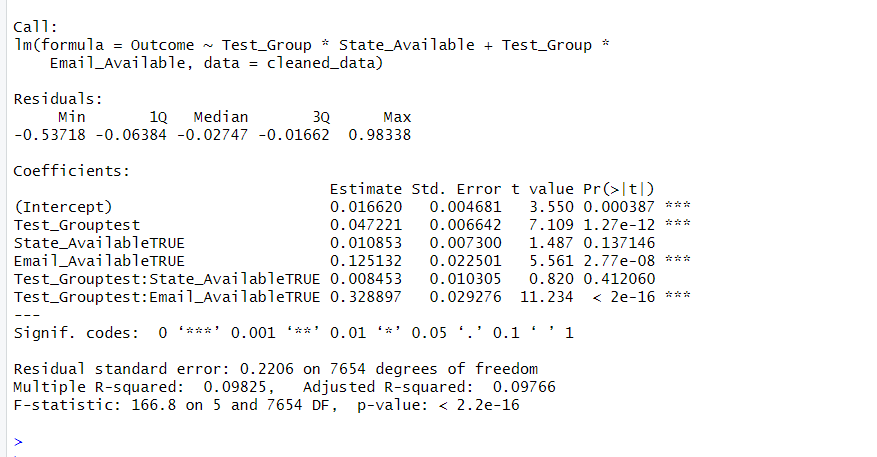
****

Output:

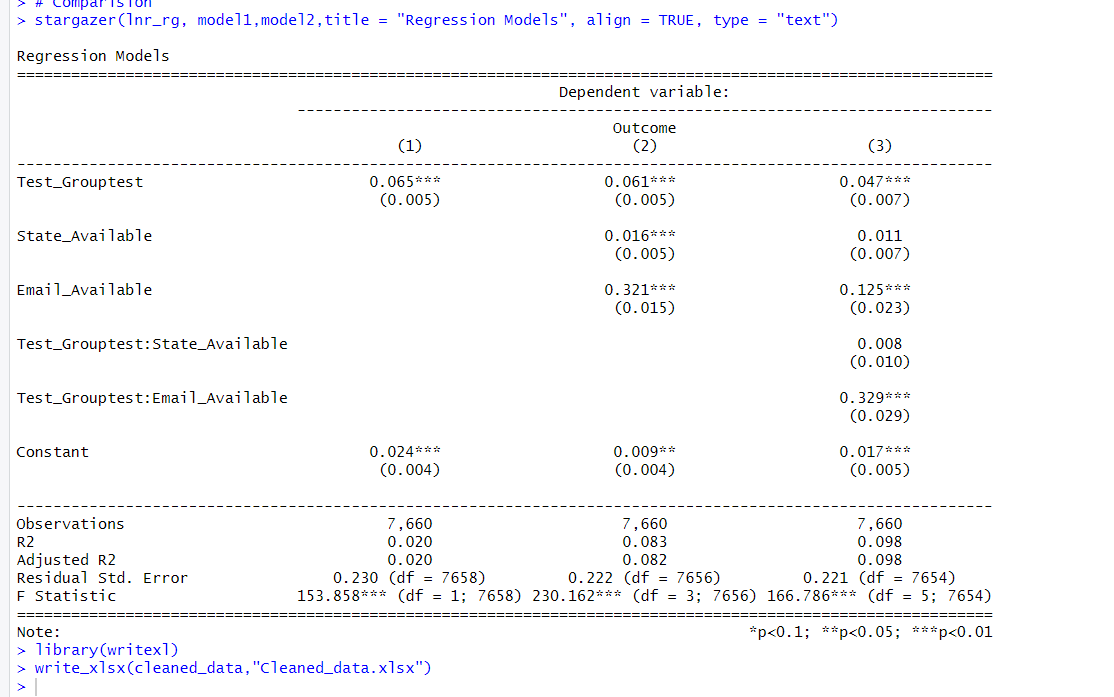
Model 1



Model 2



Stargazer result



lnr\_rg:(column-1)

* Variables : Test\_Group
* R-squared (R2) is 0.02 .

This model considers only the Test\_Group variable. It suggests that being in the test group is associated with an increase in the outcome variable. The coefficient for Test\_Grouptest is 0.065, i.e. the test group has a higher outcome value compared to the control group.

Model 1: (column-2)

* Variables : Test\_Group, 'State\_Available', 'Email\_Available
* R-squared (R2) is 0.083 .
* Test\_Group coefficient remains similar
* Having state information available is associated with an increase in the outcome variable by 0.016 units. This indicates that customers with state information available tend to have higher outcomes.
* Having email information available is associated with a significant increase in the outcome variable by 0.321 units. Customers with email information available tend to have substantially higher outcomes.

Model 2: (column-3)

* Variables : Test\_Group, 'State\_Available', 'Email\_Available and interactions
* R-squared (R2) is 0.098
* The coefficient for Test\_Grouptest decreases slightly compared to Model 1, indicating a slightly weaker association.
* The coefficient for State\_Available remains similar to Model 1, suggesting that having state information available is associated with an increase in the outcome variable.
* The coefficient for Email\_Available decreases compared to Model 1 but remains significant, indicating that having email information available is associated with an increase in the outcome variable.
* The interaction term Test\_Grouptest:Email\_Available is statistically significant. This suggests that the impact of being in the test group depends on the availability of email information.

**Summary :**

* Adding information about whether the state and email details are available helps us understand the outcome better. Among the three models the last model, which considers these additional factors, provides the clearest picture of what affects the outcome
* The interaction terms in Model 2 further improved the model's performance, indicating that the effect of being in the test group may vary based on the availability of State and Email.

**5.Reflections**

**4. Reflect on the project:**

• **Would you modify the experiment design if given a chance?**

Here the groups weren't chosen completely randomly, I might adjust that. It's important for the groups to be as fair and balanced as possible. Also, making sure that each state is represented fairly in both groups would help us understand how location affects the response to our campaign.

• **Could alternative paths be taken with better-quality data?**

If we had more detailed info about our customers, like what they look at on our site or what they've bought before, we could dig deeper. This could help us find specific groups of customers who are more likely to respond to our retargeting. It's like finding the best strategy for different types of customers.

• **Are there actionable business implications from this analysis?**

Depending on what we find, there could be some practical things we can do. For example, if we notice that certain areas really like our retargeting, we might focus more on advertising there. Also, understanding how email availability matters can help us decide how to talk to our customers effectively.

**15. Self-assessment: Rate your effort (0-100) and anticipated performance. Elaborate if needed, mentioning any collaborations.**

* I would like to rate 95% for my effort and 90% for my anticipated performance.I have put significant efforts and time to thoroughly understand the assignment and performed the detailed analysis. I have referred to the solution trace which was immensely valuable. It allowed me to tap into collective knowledge, gain different perspectives, and validate my approach.

R-Code:

library("rio")

library("moments")

library("stargazer")

abandoned <-read.csv("C:/Users/NISHANTH REDDY/Documents/abandoned.csv", header = TRUE, na.strings = "")

reservation<- read.csv("C:/Users/NISHANTH REDDY/Documents/reservation.csv", header = TRUE, na.strings = "")

sum(duplicated(abandoned ))

sum(duplicated(reservation))

#Analyse the test/control using chi sq test

table(abandoned$Test\_Control)

chisq.test(table(abandoned$Test\_Control))

##Summary Statistics

library(dplyr)

abandoned$State <- abandoned$Address #State values are present in address column for convinience added state column

summary\_statistics <- with(abandoned, table(State, Test\_Control))

summary\_statistics

#Potential data keys added to the data

abandoned$match\_email <- ifelse(complete.cases(abandoned$Email),

abandoned$Email %in% reservation$Email,

FALSE)

abandoned$match\_incoming\_phone <- ifelse(complete.cases(abandoned$Incoming\_Phone),

abandoned$Incoming\_Phone %in% reservation$Incoming\_Phone,

FALSE)

abandoned$match\_contact\_phone <- ifelse(complete.cases(abandoned$Contact\_Phone),

abandoned$Contact\_Phone %in% reservation$Contact\_Phone,

FALSE)

abandoned$match\_contact\_incoming <- ifelse(complete.cases(abandoned$Contact\_Phone),

abandoned$Contact\_Phone %in% reservation$Incoming\_Phone,

FALSE)

abandoned$match\_incoming\_contact <- ifelse(complete.cases(abandoned$Incoming\_Phone),

abandoned$Incoming\_Phone %in% reservation$Contact\_Phone,

FALSE)

abandoned$purchases <- 0

# Combine matches to identify any match

abandoned$purchases <- ifelse(abandoned$match\_email |

abandoned$match\_incoming\_contact |

abandoned$match\_incoming\_phone |

abandoned$match\_contact\_incoming |

abandoned$match\_contact\_phone,

1,

0)

sum(abandoned$purchases)

# Treatment group who purchased

test\_purchased <- subset(abandoned,Test\_Control == "test" & purchases == 1)

# Treatment group who didn’t purchase

test\_not\_purchased <- subset(abandoned,Test\_Control == "test" & purchases == 0)

# Control group who purchased

control\_purchased <- subset(abandoned, Test\_Control == "control" & purchases == 1)

# Control group who didn’t purchase

control\_not\_purchased <- subset(abandoned, Test\_Control == "control" & purchases == 0)

# Identify unmatchable records

unmatchable\_records = abandoned %>%

filter(

!(Email %in% reservation$Email) &

!(Incoming\_Phone %in% reservation$Incoming\_Phone) &

!(Contact\_Phone %in% reservation$Incoming\_Phone) &

!(Contact\_Phone %in% reservation$Contact\_Phone)

)

abandoned <- anti\_join(abandoned,unmatchable\_records)

cross\_tabulation <-table(abandoned$purchases,abandoned$Test\_Control)

rownames(cross\_tabulation)<-c("Not Purchased","Purchased")

cross\_tabulation

set.seed(7236)

random\_states <- sample(unique(abandoned$State), 5)

for (state in random\_states) {

state\_abandoned <-abandoned[abandoned$State == state,]

state\_control\_purchased <- subset(state\_abandoned, Test\_Control == "control")

state\_treatment\_purchased <- subset(state\_abandoned, Test\_Control == "test" )

state\_cross\_tab <- table( state\_abandoned$purchases, state\_abandoned$Test\_Control)

rownames( state\_cross\_tab ) <- c("Not Purchased", "Purchased")

colnames( state\_cross\_tab ) <- c("Control Group", "Treatment Group")

cat("State:", state, "\n")

print(state\_cross\_tab)

}

#Data Refinement

# Create the cleaned dataset

cleaned\_data <- data.frame(

Customer\_ID = abandoned$Caller\_ID,

Test\_Group = abandoned$Test\_Control,

Outcome = abandoned$purchases,

State\_Available = !is.na(abandoned$State),

Email\_Available = !is.na(abandoned$Email)

)

lnr\_rg <- lm(Outcome ~ Test\_Group, data = cleaned\_data)

summary(lnr\_rg)

anova\_model <- anova(lnr\_rg)

anova\_model

model1 <- lm(Outcome ~ Test\_Group+State\_Available+ Email\_Available, data = cleaned\_data)

summary(model1)

model2 <- lm(Outcome ~Test\_Group\*State\_Available+ Test\_Group\*Email\_Available, data = cleaned\_data)

summary(model2)

# Comparision

stargazer(lnr\_rg, model1,model2,title = "Regression Models", align = TRUE, type = "text")

library(writexl)

write\_xlsx(cleaned\_data,"Cleaned\_data.xlsx")