Midterm Project Code:

library(readr)

abd<-read.csv("~/Downloads/Abandoned.csv", header=T, na.strings="")

rs<-read.csv("~/Downloads/Reservation.csv", header=T, na.strings="")

#checking randomness of Test\_control variable

abd$TestingVariable <- NA

abd$TestingVariable

abd$TestingVariable[abd$Test\_Control == "test"] <- 1

abd$TestingVariable[abd$Test\_Control == "control"] <- 0

summary(abd$TestingVariable)

mean(abd$TestingVariable)

sd(abd$TestingVariable)

#summary statistics for this Test\_variable by blocking on States

abd$AddressVariable <- 0

abd$AddressVariable[abd$Address != ""] <- 1

summary(abd$TestingVariable[abd$AddressVariable == 1])

mean(abd$TestingVariable[abd$AddressVariable == 1])

sd(abd$TestingVariable[abd$AddressVariable == 1])

#match Email

match\_email=abd$Email[complete.cases(abd$Email)] %in% rs$Email[complete.cases(rs$Email)]

abd$match\_email <-0

abd$match\_email[complete.cases(abd$Email)] <- 1\* match\_email

sum(abd$match\_email)

#match Incoming Number

match\_Incoming\_Number=abd$Incoming\_Phone[complete.cases(abd$Incoming\_Phone)] %in% rs$Incoming\_Phone[complete.cases(rs$Incoming\_Phone)]

abd$match\_Incoming\_Number <-0

abd$match\_Incoming\_Number[complete.cases(abd$Incoming\_Phone)] <- 1\* match\_Incoming\_Number

sum(abd$match\_Incoming\_Number)

#Match\_Contact

match\_Contact\_Phone=abd$Contact\_Phone[complete.cases(abd$Contact\_Phone)] %in% rs$Contact\_Phone[complete.cases(rs$Contact\_Phone)]

abd$match\_Contact\_Phone <-0

abd$match\_Contact\_Phone[complete.cases(abd$Contact\_Phone)] <- 1\* match\_Contact\_Phone

sum(abd$match\_Contact\_Phone)

#Match\_Incoming\_contact

match\_Incoming\_Contact=abd$Contact\_Phone[complete.cases(abd$Contact\_Phone)] %in% rs$Incoming\_Phone[complete.cases(rs$Incoming\_Phone)]

abd$match\_Incoming\_Contact <-0

abd$match\_Incoming\_Contact[complete.cases(abd$Contact\_Phone)] <- 1\* match\_Incoming\_Contact

sum(abd$match\_Incoming\_Contact)

#match contact Incoming

match\_Contact\_Incoming=abd$Incoming\_Phone[complete.cases(abd$Incoming\_Phone)] %in% rs$Contact\_Phone[complete.cases(rs$Contact\_Phone)]

abd$match\_Contact\_Incoming <-0

abd$match\_Contact\_Incoming[complete.cases(abd$Incoming\_Phone)] <- 1\* match\_Contact\_Incoming

sum(abd$match\_Contact\_Incoming)

All\_Matches<-abd$match\_email | abd$match\_Incoming\_Number | abd$match\_Incoming\_Contact | abd$match\_Contact\_Phone | abd$match\_Contact\_Incoming

abd\_noduplicates <- abd[All\_Matches,]

abd\_noduplicates

#Removed duplicates using the variables

abd\_noduplicates <- abd\_noduplicates[!duplicated(abd\_noduplicates[,c("Email")],incomparables = NA),]

abd\_noduplicates <- abd\_noduplicates[!duplicated(abd\_noduplicates[,c("Contact\_Phone")],incomparables = NA),]

abd\_noduplicates <- abd\_noduplicates[!duplicated(abd\_noduplicates[,c("Incoming\_Phone")],incomparables = NA),]

abd\_FirstName\_nodupe <- duplicated(abd\_noduplicates[,c("Last\_Name")],incomparables = NA)

abd\_LastName\_nodupe <- duplicated(abd\_noduplicates[,c("First\_Name")],incomparables = NA)

abd\_Address\_nodupe<- duplicated(abd\_noduplicates[,c("Address")],incomparables = NA)

abd\_Zipcode\_nodupe<-duplicated(abd\_noduplicates[,c("Zipcode")],incomparables=NA)

abd\_noduplicates<-abd\_noduplicates[!(abd\_FirstName\_nodupe & abd\_LastName\_nodupe & abd\_Address\_nodupe & abd\_Zipcode\_nodupe),]

abd\_noduplicates

# Purchased Outcome in original dataset

abd$Purchased <- 0

abd$Purchased[as.numeric(row.names(abd\_noduplicates))] <- 1

#logical dummies for variables

abd$email1 <- 1\*complete.cases(abd$Email)

abd$Address1 <- 1\*complete.cases(abd$Address)

abd$Zipcode1<-1\*complete.cases(abd$Zipcode)

abd$TestControl<-1\*(abd$Test\_Control=='test')

table(abd$Purchased,abd$TestControl)

#Analysis using Address variable for 5 states

abd\_states<-data.frame(subset(abd,abd$Address!=""))

table(abd\_states$Purchased, abd\_states$TestControl)

abd\_OH<-data.frame(subset(abd,abd$Address=="OH"))

table(abd\_OH$Purchased, abd\_OH$TestControl)

abd\_NY<-data.frame(subset(abd,abd$Address=="NY"))

table(abd\_NY$Purchased, abd\_NY$TestControl)

abd\_FL<-data.frame(subset(abd,abd$Address=="FL"))

table(abd\_FL$Purchased, abd\_FL$TestControl)

abd\_IL<-data.frame(subset(abd,abd$Address=="IL"))

table(abd\_IL$Purchased, abd\_IL$TestControl)

abd\_AZ<-data.frame(subset(abd,abd$Address=="AZ"))

table(abd\_AZ$Purchased, abd\_AZ$TestControl)

#Forming the excel sheet all data

install.packages("writexl")

library("writexl")

uid <- paste0("ASDFG", formatC(1:8442, width = 10, format = "d", flag = "0"))

head(uid)

abd\_Excel<-data.frame(uid,abd$Purchased,abd$email1,abd$Address1,abd$TestControl)

colnames(abd\_Excel)<-c("CustomerID","Outcome","D\_Email","D\_State","Test\_Variable")

abd\_Cleaned<-write\_xlsx(abd\_Excel, "~/Downloads/midtermproject.xlsx")

library("readxl")

abd\_Cleaned<-read\_excel("~/Downloads/midtermproject.xlsx")

#Excel sheet for 401 records(only purchased customers)

Purchased\_data <- subset(abd\_Cleaned,abd\_Cleaned$Outcome=="1")

write\_xlsx(Purchased\_data, "~/Downloads/purchasedcustomer\_midterm.xlsx")

#T Test

Anova<-aov(Outcome~Test\_Variable, data=abd\_Cleaned)

summary(Anova)

coef(Anova)

Anova2<-aov(Outcome~Test\_Variable+D\_Email+D\_State, data=abd\_Cleaned)

summary(Anova2)

coef(Anova2)

Anova3<-aov(Outcome~Test\_Variable+ Test\_Variable\*D\_Email, data=abd\_Cleaned)

summary(Anova3)

coef(Anova3)

t.test(abd\_Cleaned$Outcome,abd\_Cleaned$Test\_Variable)

#linear Regression

out1<- lm(Outcome~Test\_Variable, data=abd\_Cleaned)

summary(out1)

out2<-lm(Outcome~Test\_Variable+D\_Email+D\_State, data=abd\_Cleaned)

summary(out2)

out3<-lm(Outcome~Test\_Variable+ Test\_Variable\*D\_Email, data=abd\_Cleaned)

summary(out3)

out4<-lm(Outcome~Test\_Variable+Test\_Variable\*D\_State, data=abd\_Cleaned)

summary(out4)

out5<-lm(Outcome~Test\_Variable+ Test\_Variable\*D\_Email+ Test\_Variable\*D\_State, data=abd\_Cleaned)

summary(out5)

install.packages("stargazer")

library("stargazer")

stargazer(out1,out2,out3,out4,out5, type="html", out="Output.htm")

AIC(out1,out2,out3,out4,out5)

#logistic regression analysis code

install.packages("corrplot")

library("corrplot")

c = cor(abd\_Cleaned[,2:5])

corrplot(c)

c

out6= glm(Outcome~Test\_Variable,abd\_Cleaned, family=binomial())

out7= glm(Outcome~Test\_Variable+D\_State+D\_Email,abd\_Cleaned, family = binomial())

out8= glm(Outcome~Test\_Variable+Test\_Variable\*D\_State + Test\_Variable\*D\_Email,abd\_Cleaned, family =binomial())

library(stargazer)

stargazer(out6,out7,out8, type="html",out="gmlregression.htm")

AIC(out6,out7,out8)

#compute confusion matrix

Y\_pred1 = 1\*(predict(out6, type="response")>0.05)

Y\_pred2 = 1\*(predict(out7, type="response")>0.05)

Y\_pred3 = 1\*(predict(out8, type="response")>0.05)

table(Y\_pred1,abd\_Cleaned$Outcome)

table(Y\_pred2,abd\_Cleaned$Outcome)

table(Y\_pred3,abd\_Cleaned$Outcome)

I. The Business Problem

Business Problem: Should we retarget those customers?

Q1: In light of your experience as a businessperson/man, argue why this is a sensible business question.

It is better to retarget the customers, because Retargeting is a marketing strategy, and it helps us in increasing the customer base.

The customers in the abandoned dataset are the ones who engaged with the   
travel agency, if we can reach them again it may help us, because some customers need that product as a future purchase and are just searching other agencies for the best purchase.

Then Retargeting after some days might help us in persuading the customer to buy our product.

Q2: Investigate the test/control variable. Does the experiment seem to be run   
properly?

Code:

#checking randomness of Test\_control variable

abd$TestingVariable <- NA

abd$TestingVariable

abd$TestingVariable[abd$Test\_Control == "test"] <- 1

abd$TestingVariable[abd$Test\_Control == "control"] <- 0

summary(abd$TestingVariable)

mean(abd$TestingVariable)

sd(abd$TestingVariable)

Output:

The customers are randomly assigned to ‘test’ and ‘control group.’

summary(abd$TestingVariable)

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

0.0000 0.0000 1.0000 0.5053 1.0000 1.0000

> sd(abd$TestingVariable)

[1] 0.5000012

> mean(abd$TestingVariable)

[1] 0.5053305

Here we could see that mean and sd values are very much near to 0.5, From this we could day that nearly 50 percent of data for Test\_control variable is assigned with value “test” and the other data with “Control’.

So from this we can say that Treatment group and Control group of customers are randomly picked and the test runs properly as that data is not imbalanced.

Q3: compute the same summary statistics for this Test\_variable by stratifying on   
States (meaning considering only the entries with known “State”), wherever this   
information is available.

#summary statistics for this Test\_variable by blocking on States

abd$AddressVariable <- 0

abd$AddressVariable[abd$Address != ""] <- 1

summary(abd$TestingVariable[abd$AddressVariable == 1])

mean(abd$TestingVariable[abd$AddressVariable == 1])

sd(abd$TestingVariable[abd$AddressVariable == 1])

Output:

summary(abd$TestingVariable[abd$AddressVariable == 1])

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

0.0000 0.0000 1.0000 0.5134 1.0000 1.0000

> mean(abd$TestingVariable[abd$AddressVariable == 1])

[1] 0.5133788

> sd(abd$TestingVariable[abd$AddressVariable == 1])

[1] 0.4998865

Here we checked Randomness of Test\_Control variable for only the customers who has an address value.

Here also we could see that mean and sd values are very much near to 0.5, From this we could day that nearly 50 percent of data for Test\_control variable is assigned with value “test” and the other data with “Control’.

So from this we can say that Treatment group and Control group of customers are randomly picked and the test runs properly as that data is not imbalanced.

II. Data Matching

Q5)

After observing the data in both files, argue that customers can be matched across   
some “data keys” (column labels). Correctly identify all these data keys (feel free to   
add a few clarifying examples if needed).

Output:

Data keys that can be unique identifiers of customer between both the datasets are:

Email, Incoming Phone, Contact Phone.

First Name and Last name can also be used to matched data but combinedly.

Q6: EXTREMELY CAREFULLY DESCRIBE YOUR DATA MATCHING PROCEDURE 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 have matched data present in abandoned data and that are also present in reservation data.

After collecting matched data I have checked for duplicates based on Email, Incoming number, Phone Number, and Last name+First Name+ state+ Zipcode and removed them from matched data.

I took all the outputs in dummy values. I have assigned purchased value a 1 for all the matched customers. Then from the following table. I got the outputs.

> table(abd$Purchased,abd$TestControl)

0 1

0 4086 3955

1 90 311

From the above matrix we can see that.

Customers in the TREATMENT group who bought are 311

Customers in the TREATMENT group who did not buy are 3955

Customers in the Control group who bought are 90

Customers in the Control group who did not buy 4086

Q7: Are there problematic cases? i.e. data records not matchable? If so, provide a few examples and toss those cases out of the analysis.

As I have considered Email, Incoming\_Phone and Contact\_Phone to match data, there are many customers for whom one or more data is missing for the above variables. So we had to use more variables to match the records.

8: Complete the following cross-tabulation:  
Group \ Outcome Buy No Buy  
Treatment 311 3955  
Control 90 4083

Q9: Repeat Q8 for 5 randomly picked states. Report 5 different tables by   
specifying the states you “randomly picked”

For “OH” state

> abd\_OH<-data.frame(subset(abd,abd$Address=="OH"))

> table(abd\_OH$Purchased, abd\_OH$TestControl)

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

For “NY” state

> abd\_NY<-data.frame(subset(abd,abd$Address=="NY"))

> table(abd\_NY$Purchased, abd\_NY$TestControl)

Group \ Outcome Buy No Buy  
Treatment 3 37  
Control 1 35

For “FL” state

> abd\_FL<-data.frame(subset(abd,abd$Address=="FL"))

> table(abd\_FL$Purchased, abd\_FL$TestControl)

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

For “IL” state

> abd\_IL<-data.frame(subset(abd,abd$Address=="IL"))

> table(abd\_IL$Purchased, abd\_IL$TestControl)

Group \ Outcome Buy No Buy  
Treatment 2 35  
Control 0 47

For “AZ” state

> abd\_AZ<-data.frame(subset(abd,abd$Address=="AZ"))

> table(abd\_AZ$Purchased, abd\_AZ$TestControl)

Group \ Outcome Buy No Buy  
Treatment 43 51  
Control 1 3

Q10: Run a Linear regression model for   
Outcome = alpha + beta \* Test\_Variable + error  
And Report the output.

Output:

out1<- lm(Outcome~Test\_Variable, data=abd\_Cleaned)

> summary(out1)

Call:

lm(formula = Outcome ~ Test\_Variable, data = abd\_Cleaned)

Residuals:

Min 1Q Median 3Q Max

-0.08087 -0.08087 -0.02227 -0.02227 0.97773

Coefficients:

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

(Intercept) 0.022270 0.003402 6.545 6.28e-11 \*\*\*

Test\_Variable 0.058602 0.004786 12.244 < 2e-16 \*\*\*

Residual standard error: 0.2199 on 8440 degrees of freedom

Multiple R-squared: 0.01745, Adjusted R-squared: 0.01733

F-statistic: 149.9 on 1 and 8440 DF, p-value: < 2.2e-16.

Q11: Argue this is statistically equivalent to an ANOVA/t-test.

Linear Regression is statistically equivalent to ANOVA/t-test in this scenario as we are using dummy variables to perform the test.

The F static (a ratio of the means) and P value and co efficients are same in case of anova and linear regression so we can say that they are statistically equivalent.

Anova Test:

Anova<-aov(abd\_Cleaned$Outcome~abd\_Cleaned$Test\_Variable, data=abd\_Cleaned)

> Anova<-aov(Outcome~Test\_Variable, data=abd\_Cleaned)

> summary(Anova)

> coef(Anova)

(Intercept) Test\_Variable

0.02155172 0.05135029

Df Sum Sq Mean Sq F value Pr(>F)

Test\_Variable 1 5.6 5.564 124.8 <2e-16 \*\*\*

Residuals 8440 376.4 0.045

P value is very less, null hypothesis is statically significant, so we reject the null hypothesis indicates that outcome variable is not very much related with Test\_Variable.

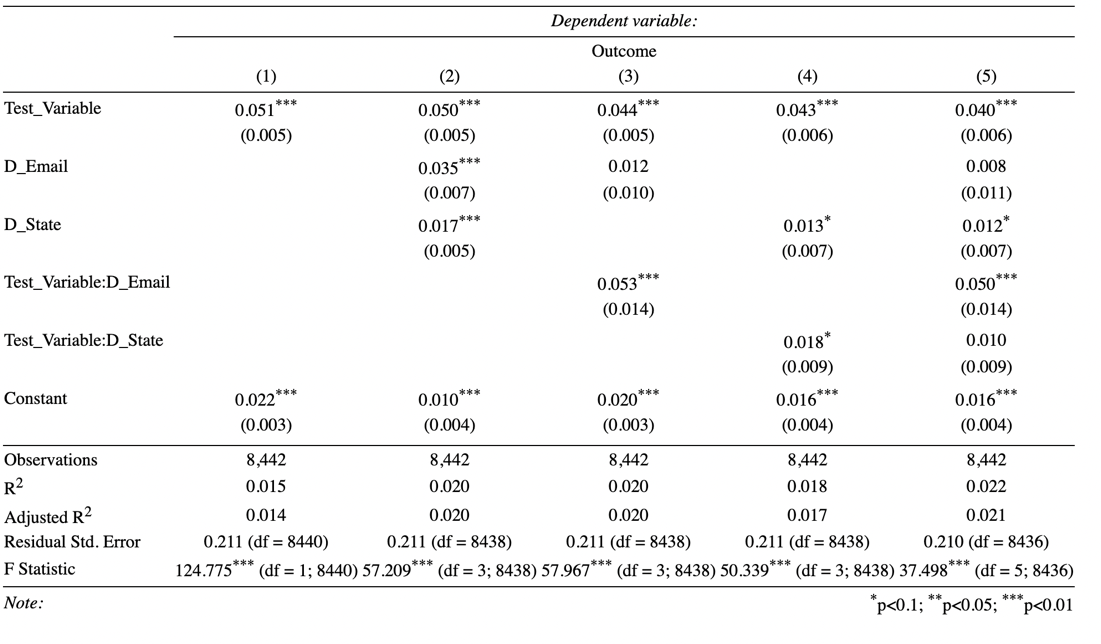
ANOVA and Linear regression are statistically equivalent when more variables are used in linear regression and anova test . Output is same for both.

Q12: 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?

Output:

It is not a properly specified linear regression model as the R-squared: 0.015 values is very less which indicates that the model needs more variables to improve the significance of the model. So, It is statistically not significant to determine the outcome from above linear regression model.

Q13: Now add the dummies for State and Emails to the regression model. Also   
consider including interactions with the treatment. Report the outcome and comment   
on the results. (You can compare with Q10)



> AIC(out1,out2,out3,out4,out5)

df AIC

out1 3 -2294.254

out2 5 -2336.348

out3 5 -2338.578

out4 5 -2316.116

out5 7 -2347.933

The AIC value is less for out 1 compared to others which indicates first linear regression model is a best fit compared to others.

From above report we could see that Test\_Variable has more significance in determining the Outcome variable as the coefficient value is more for Test\_Variable compared to other independent Variables.

So Outcome variation will be more when Test\_variable is changed as of other variables .

outcome changes by 5.1% when a customer is added to “test “ in Test\_Variable.

R squared value had increased when interactions between variables is increased. So, the customers with more variables like Email, State, has better outcome. But R square value is still low to determine the dependent variable from the given inputs.

VI: Conclusion

Q14: 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 of this study? Please answer   
briefly.

I have learnt about matching the data between two different data sets, data cleaning and statistical analysis.

I would add an extra variable “Positive Response” where the values are in 0 ‘s and 1’s. Value 1 indicates that there was some positive response when the customer has engaged with the agency.

There is so much missing data, so the statistical analysis is not 100 percent accurate. To match customer data, we are using so many different variables, instead of that it would have been easier if we have a unique ID for every customer, so that we could match data in both datasets using that ID.

Here the number customers in test group who purchased the product is thrice compared to the customers in control group. So retargeting works, but there is large amount to people who didn’t but the product, so it is better to retarget only those customers who were more interested in the product during the first call or interaction. Or target only the customers for those states for which purchase rate is high compared to others.Or target only the customers for those states for which purchase rate is high compared to others.etc

Q15: Self-evaluation. Please score your effort on a scale of 0-100. Please score   
your expected performance on the same scale. Add comments, if necessary,   
including whether you collaborate with your peers.

I would rate my effort to be 90 on a scale of 100, as you have already provided us some code syntax for the project, so it took a bit less effort. I think I have given my 100 percent. I have not collaborated with anyone for the project.