

Retargeting Campaign Analysis

INTRODUCTION

To determine the statistical effectiveness of the retargeting campaign using the 'Abandoned.csv' and 'Reservation.csv' datasets I have performed these steps:

- Clean and integrate the data.
- Define key metrics (e.g., conversion rate).
- Compare test and control groups from 'Abandoned.csv.'
- Use ANOVA and regression analysis to assess the impact of the 'Test_Control' variable.
- Segment the data by different variables (e.g., state).
- Determine statistical significance.
- Conclude and provide recommendations for optimizing the campaign.

Task here is to : Establish whether the retargeting campaign was statistically effective.

1. Business Justification

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

Retargeting non-purchasing customers is cost-effective, leverages existing data, and enables personalized marketing, making it a practical and data-driven strategy for boosting sales and optimizing marketing efforts.

Here are reasons why retargeting customers makes business sense:

- **Cost-Efficiency:** Retargeting is cost-effective compared to acquiring new customers.
- **Personalization:** Tailor marketing to specific interests for higher conversions.
- **Learning Opportunity:** Gain insights into why customers aren't buying.
- **Existing Data:** Efficiently use data you already have.
- **Strategic Decisions:** Inform future business strategies based on data.

In this experiment, customers in the abandoned dataset were randomly assigned to either a treatment group or a control group. Customers labeled as "test" received retargeted marketing (treatment), while those labeled as "control" were not retargeted and served as the control group.

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

```
> summary(abddata$Test_Control)
```

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

0.0000	0.0000	1.0000	0.5053	1.0000	1.0000
--------	--------	--------	--------	--------	--------

```
> table(abddata$Test_Control)
```

0	1
---	---

4176	4266
------	------

The `Test_Control` column consists of two categories, most likely "test" and "control." The summary statistics show that roughly 50.53% of the observations are in the "test" category, while the remaining 49.47% are in the "control" category. This suggests a relatively balanced distribution between the two groups.

We can see from the abandoned dataset that under the test variable 4266 customers are marked and for the controlled 4176 customers are marked which are approximately equal indicating the experiment is successful.

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

```
> summary(abandoned.summary$Test_Control)
```

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

0.0000	0.0000	1.0000	0.5053	1.0000	1.0000
--------	--------	--------	--------	--------	--------

Interpretation:

Out of the known states, there are 1,856 entries in the control group and 1,958 entries in the test group. The mean value of 0.5134 indicates that, among the entries with known states, the distribution between the test and control groups is fairly balanced.

2. Data Alignment

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

The dataset allows us to perform matching between email and phone numbers in both the abandoned and reservation datasets. This matching is done in the following manner:

Abandoned Email to Reservation Email

Abandoned Incoming Phone to Reservation Incoming Phone

Abandoned Contact Phone to Reservation Contact Phone

Abandoned Incoming Phone to Reservation Contact Phone

Abandoned Contact Phone to Reservation Incoming Phone

5. 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.

To categorize the groups based on the "Test_Control" and "Purchase" parameters, we will perform the following matching:

Match Abandoned Email to Reservation Email.

Match Abandoned Incoming Phone to Reservation Incoming Phone.

Match Abandoned Contact Phone to Reservation Contact Phone.

Match Abandoned Incoming Phone to Reservation Contact Phone.

Match Abandoned Contact Phone to Reservation Incoming Phone.

If any of these conditions result in a match, we will classify the customers in the abandoned dataset as having purchased the product.

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

Even if we don't have all the information for every customer, we can still decide if they made a purchase or not by matching available data, like emails and phone numbers, in specific ways. This helps us label customers as buyers or non-buyers, even with incomplete information. We have done this step using the following code:

```
abddata$pur <- 0
```

```
abddata$pur <- 1 * (abddata$match_email | abddata$match_incoming |  
abddata$match_contact | abddata$match_incoming_contact |  
abddata$match_contact_incoming)
```

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

Group	Purchase	Not Purchased
Control	93	4083
Treatment	345	3921

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

CA

Group	Purchase	Not Purchased
Control	0	37
Treatment	6	42

LA

Group	Purchase	Not Purchased
Control	0	36
Treatment	2	37

FL

Group	Purchase	Not Purchased
Control	0	37
Treatment	4	34

OH

Group	Purchase	Not Purchased
Control	3	30
Treatment	4	46

UT

Group	Purchase	Not Purchased
Control	3	30
Treatment	4	23

3. Data Refinement

9. Generate a cleaned dataset with columns: Customer ID — Test Group — Outcome — State Available — Email Available. Each row should correspond to a matched customer from the datasets. (*Ensure you attach this cleaned dataset upon submission.*)

Create an Excel file with the following columns:

- *Customer ID*
- *Test Variable (indicating treatment or control group)*
- *Our (a binary variable indicating whether a vacation package was purchased)*
- *State (indicating the presence of state information)*
- *Email (indicating the presence of email information)*

Ensure that the Excel file has one row for each customer that could be matched across the two datasets. Please submit this Excel file for verification.

final_data_sheet.xlsx Attached on submission

4. Statistical Assessment

10. Execute a linear regression for the formula:
$$\text{Outcome} = \alpha + \beta * \text{Test Group} + \text{error}.$$

Share the results.

> #Linear regression for Pur based on Test_Control

> out_lm1 <- lm(pur ~ Test_Control, data = abddata)

> summary(out_lm1)

Output:

Call:

```
lm(formula = pur ~ Test_Control, data = abddata)
```

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_Control	0.058602	0.004786	12.244	< 2e-16 ***

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

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

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

```
> model <- aov( pur~Test_Control, data = abddata)
```

```
> summary(model)
```

Output:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Test_Control	1	7.2	7.247	149.9	<2e-16 ***

Residuals 8440 408.0 0.048

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

Both the Linear Regression (LM) and Analysis of Variance (ANOVA) tests show that 'Test_Control' significantly affects 'pur.' LM tells us the precise relationship, while ANOVA confirms the overall group differences. Both tests agree that there's a meaningful connection between the variables.

The small p-value of $2e-16$, which is less than the typical significance level of 0.05, allows us to reject the null hypothesis. This indicates a significant difference in the means of the groups.

Despite the p-values being the same ($2e-16$) in both tests, their interpretations differ. In the ANOVA test, it assesses whether the means of two or more groups are different. In contrast, the linear regression model examines if the regression, as a whole, is performing better than random chance by using the F-statistic. While the p-values match, their contexts and implications vary.

12. Debate the appropriateness of the regression model in making causal claims about the retargeting campaign's efficacy.

Based on the results of the linear regression analysis, we find that there is a strong and positive relationship between the test variable (often used to represent the retargeting campaign) and purchase behavior. In other words, when the test variable changes randomly, it tends to have a positive effect on whether customers make a purchase. The low p-value (p-value $< 2e-16$) indicates that this relationship is highly significant, meaning it's unlikely to have occurred by chance.

However, it's important to note that the linear regression model is not very powerful in explaining why people make purchases. It only accounts for a small portion of the reasons behind purchase decisions. In fact, the model explains just 1.7% of the variability in purchase behavior, which means that many other factors not included in the model influence whether a customer makes a purchase or not.

In simpler terms, while the test variable seems to play a role in influencing purchases, it's not the whole story. Many other factors, like pricing, product quality, or customer preferences, also impact purchase decisions. The model helps us understand part of

the picture, but there's much more to explore to fully comprehend why customers buy or don't buy.

13. **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.**

#Linear regression for Pur based on Test_control + Email + Address

```
> out_lm2 <- lm(pur ~ Test_Control + Email + Address , data = abddata)
```

```
> summary(out_lm2)
```

Call:

```
lm(formula = pur ~ Test_Control + Email + Address, data = abddata)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.12161	-0.06833	-0.06399	-0.01070	0.98930

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.010703	0.004023	2.661	0.007814 **
Test_Control1	0.057623	0.004777	12.064	< 2e-16 ***
Email	0.036416	0.007485	4.865	1.16e-06 ***
Address	0.016873	0.004921	3.429	0.000609 ***

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

Residual standard error: 0.2193 on 8438 degrees of freedom

Multiple R-squared: 0.02268, Adjusted R-squared: 0.02233

F-statistic: 65.26 on 3 and 8438 DF, p-value: < 2.2e-16

Analyzing the equation from the multiple linear regression model, we find that the relationship between the variables and the outcome can be described as follows:

$$\text{Our} = 0.057623 * \text{Test_Control} + 0.036416 * \text{Email} + 0.016873 * \text{Address} + 0.010703$$

While the overall performance of the model is not very strong (Adjusted R-squared: 0.02233), we can observe that the Test_Control variable (0.057623 with a p-value of $2e-16$), Email (0.036416 with a p-value of $1.16e-06$), and Address (0.016873 with a p-value of 0.000609) all have a positive impact on purchase behavior.

This model appears to be an improvement over the previous linear regression model, as it has a slightly better adjusted R-squared value (0.02233). This suggests that it explains a bit more of the variability in purchase behavior compared to the previous model.

However, it's still important to remember that many other factors could influence purchase decisions beyond these variables.

```
> #Linear regression for pur based on Test_control * Email + Test_control * Address
```

```
> out_lm3 <- lm(pur ~ Test_Control* Email + Test_Control* Address , data = abddata)
```

```
> summary(out_lm3)
```

Call:

```
lm(formula = pur ~ Test_Control * Email + Test_Control * Address,  
    data = abddata)
```

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

-0.14608 -0.06218 -0.03474 -0.01684 0.98316

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.016844	0.004583	3.675	0.000239 ***
Test_Control1	0.045338	0.006494	6.981	3.15e-12 ***
Email	0.007599	0.011016	0.690	0.490343
Address	0.010301	0.006987	1.474	0.140443
Test_Control1:Email	0.052981	0.015004	3.531	0.000416 ***
Test_Control1:Address	0.013011	0.009833	1.323	0.185810

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

Residual standard error: 0.2191 on 8436 degrees of freedom

Multiple R-squared: 0.02466, Adjusted R-squared: 0.02408

F-statistic: 42.65 on 5 and 8436 DF, p-value: < 2.2e-16

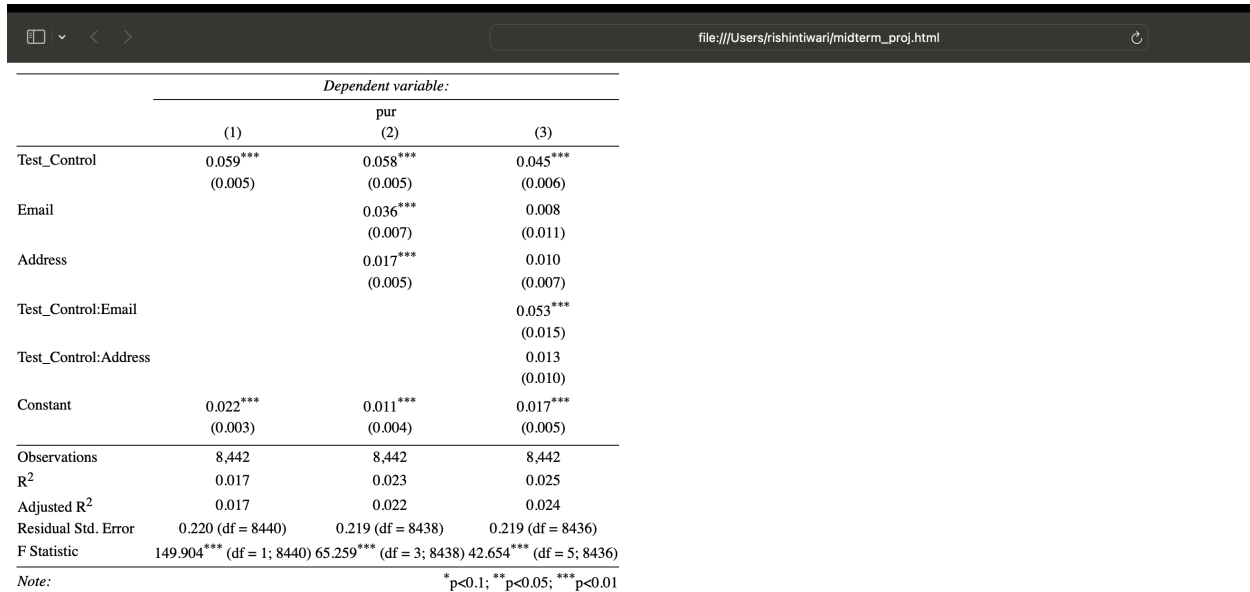
Examining the equation from the multiple linear regression model, we can express the relationship between the variables and the outcome as follows:

$$\text{`pur} = 0.052981 * \text{Test_Control} * \text{Email} + 0.013011 * \text{Test_Control} * \text{Address} + 0.045338 * \text{Test_Control} + 0.007599 * \text{Email} + 0.010301 * \text{Address} + 0.016844`$$

Despite the linear regression model having limited explanatory power (Adjusted R-squared: 0.02408), we can discern some important insights. Specifically, the interaction between Test_Control and Email (0.052981 with a p-value of 0.000416) exhibits a significant positive impact on purchase behavior. However, the interaction between Test_Control and Address (0.013011 with a p-value of 0.185810) does not appear to have a significant influence. This implies that when Test_Control and Email work together, it leads to more purchases.

Generated summary table using Stargazer

```
stargazer(out_lm1, out_lm2, out_lm3, type = "html", out = "midterm_proj.html")
```



Dependent variable:			
	(1)	pur (2)	(3)
Test_Control	0.059*** (0.005)	0.058*** (0.005)	0.045*** (0.006)
Email		0.036*** (0.007)	0.008 (0.011)
Address		0.017*** (0.005)	0.010 (0.007)
Test_Control:Email			0.053*** (0.015)
Test_Control:Address			0.013 (0.010)
Constant	0.022*** (0.003)	0.011*** (0.004)	0.017*** (0.005)
Observations	8,442	8,442	8,442
R ²	0.017	0.023	0.025
Adjusted R ²	0.017	0.022	0.024
Residual Std. Error	0.220 (df = 8440)	0.219 (df = 8438)	0.219 (df = 8436)
F Statistic	149.904*** (df = 1; 8440)	65.259*** (df = 3; 8438)	42.654*** (df = 5; 8436)
Note: *p<0.1; **p<0.05; ***p<0.01			

5. Reflections

14. Reflect on the project:

- Would you modify the experiment design if given a chance?
- Could alternative paths be taken with better-quality data?
- Are there actionable business implications from this analysis?

It's evident that the model lacks a unique identifier for each customer, which could lead to challenges in tracking and attributing conversions accurately. To address this, each customer should have a unique marker so that we can precisely determine which targeted customers are converting into actual buyers.

The available data in both scenarios can be invaluable for shaping our customer targeting strategies and assessing whether our advertising efforts have a positive impact. This insight allows us to refine our approaches in subsequent campaigns to achieve better results.

While the experiment design could be improved, the analysis still provides valuable insights for the business, and better data quality would only enhance these insights. The key takeaway is the importance of data-driven decision-making and the potential for cost-effective customer retargeting.

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

I've applied my analytical and interpretive skills to the fullest extent in this project. I've explored various matching character sets and their impact on the results. If I were to assess my effort, I would give myself a perfect score of 100. Just took few takeaways from Prof. Daniel Z. During the lectures and team meetings which helped me a lot to solve this project effectively.

Code :

```
#reloading the existing libraries
```

```
library("write1")
```

```
library("stringr")
```

```
library("dplyr")
```

```
library("readr")
```

```
library("rio")
```

```
library(stargazer)
```

```
library("moments")
```

```
#install.packages("writexl")
```

```
library("writexl")
```

```
#loading csv file
```

```
abddata <- read.csv("/Users/rishintiwari/Desktop/Fall 2023/Analytical Methods of  
Business(AMB)/Midterm Project/Abandoned.csv" , header=T, na.strings="")
```

```
resdata <- read.csv("/Users/rishintiwari/Desktop/Fall 2023/Analytical Methods of  
Business(AMB)/Midterm Project/Reservation.csv" , header=T, na.strings="")
```

```
# displaying the column names and dimensions
```

```
variable_names <- names(abddata)
```

```
print(variable_names)
```

```
dim(abddata)
```

```
dim(resdata)
```

```
# Matching data
```

```
match_email <- abddata$Email[complete.cases(abddata$Email)] %in%  
resdata$Email[complete.cases(resdata$Email)]
```

```
match_incoming <-  
abddata$Incoming_Phone[complete.cases(abddata$Incoming_Phone)] %in%  
resdata$Incoming_Phone[complete.cases(resdata$Incoming_Phone)]
```

```
match_contact <- abddata$Contact_Phone[complete.cases(abddata$Contact_Phone)]  
%in% resdata$Contact_Phone[complete.cases(resdata$Contact_Phone)]
```

```
match_incoming_contact <-  
abddata$Incoming_Phone[complete.cases(abddata$Incoming_Phone)] %in%  
resdata$Contact_Phone[complete.cases(resdata$Contact_Phone)]
```

```
match_contact_incoming <-  
abddata$Contact_Phone[complete.cases(abddata$Contact_Phone)] %in%  
resdata$Incoming_Phone[complete.cases(resdata$Incoming_Phone)]
```

```
# Create flags
```

```
# Create match_email and email columns
```

```
abddata$match_email <- 0
```

```
abddata$match_email[complete.cases(abddata$Email)] <- 1* match_email
```

```
# Create match_incoming and incoming columns
```

```
abddata$match_incoming <- 0
```

```
abddata$match_incoming[complete.cases(abddata$Incoming_Phone)] <- 1*  
match_incoming
```

```
# Create match_contact and contact columns
```

```
abddata$match_contact <- 0
```

```
abddata$match_contact[complete.cases(abddata$Contact_Phone)] <- 1*  
match_contact
```

```
# Create match_incoming_contact column
```

```
abddata$match_incoming_contact <- 0
```

```
abddata$match_incoming_contact[complete.cases(abddata$Incoming_Phone) ] <- 1*  
match_incoming_contact
```

```
# Create match_contact_incoming column
```

```
abddata$match_contact_incoming <- 0
```

```
abddata$match_contact_incoming[complete.cases(abddata$Contact_Phone) ] <- 1*  
match_contact_incoming
```

```
abddata$pur <- 0
```

```
abddata$pur <- 1 * (abddata$match_email | abddata$match_incoming |  
abddata$match_contact | abddata$match_incoming_contact |  
abddata$match_contact_incoming)
```

```
#Marking Test as 1 and Control as 0 for Treatment variables
```

```
abddata[abddata == 'test'] <- 1
```

```
abddata[abddata == 'control'] <- 0
```

```
#Cross Tabulation Value for the whole dataset
```

```
#test & purchase
```

```
a1 = nrow(abddata[abddata$pur == '1' & abddata$Test_Control == '1', ])
```

```
#test & not purchase
```

```
b1 = nrow(abddata[abddata$pur == '0' & abddata$Test_Control == '1', ])
```

```
# control & purchase
```

```
c1 = nrow(abddata[abddata$pur == '1' & abddata$Test_Control == '0', ])
```

```
#control & not purchase
```

```
d1 = nrow(abddata[abddata$pur == '0' & abddata$Test_Control == '0', ])
```

```
data.matrix = matrix(c(a1,b1,c1,d1),ncol=2,nrow=2,byrow=TRUE)
```



```
colnames(data.matrix) = c("Purchased", "Not Purchased")
```

```
rownames(data.matrix) = c("Treatment", "Control")
```

```
data.matrix.tb <- as.table(data.matrix)
```

```
data.matrix.tb
```

```
# Create a function to compute cross-tabulations for a given state
```

```
funct_cross_tab <- function(data, state) {
```

```
  subset_state <- subset(data, Address == state)
```

```
  cross_tab <- table(subset_state$Test_Control, subset_state$pur)
```

```
  return(cross_tab)
```

```
}
```

```
# Example usage for different states
```

```
cross_tab_CA <- funct_cross_tab(abddata, "CA")
```

```
print(cross_tab_CA, quote = FALSE)
```

```
cross_tab_LA <- funct_cross_tab(abddata, "LA")
```

```
print(cross_tab_LA, quote = FALSE)
```

```
cross_tab_FL <- funct_cross_tab(abddata, "FL")
```

```
cross_tab_OH <- funct_cross_tab(abddata, "OH")
```

```
cross_tab_UT <- funct_cross_tab(abddata, "UT")
```

```
# creating Dataframe with specific column for Exporting to our PC
```

```
final.data = select(abddata, c('Caller_ID', 'Test_Control', 'pur', 'Address', 'Email'), )
```

```
Final.data.sub = subset(final.data, pur == 1 )  
write_xlsx(Final.data.sub ,"/Users/rishintiwari/Desktop/final_data_sheet.xlsx")
```

```
# Marking Email and address as 1 in case of values available
```

```
abddata$Email<- 1*complete.cases(abddata$Email)
```

```
abddata$Address <- 1*complete.cases(abddata$Address)
```

```
#Analyzing Test-Control Division
```

```
abddata$Test_Control <- as.numeric(abddata$Test_Control)
```

```
summary(abddata$Test_Control)
```

```
table(abddata$Test_Control)
```

```
#Linear regression for Pur based on Test_Control
```

```
out_lm1 <- lm(pur ~ Test_Control, data = abddata)
```

```
summary(out_lm1 )
```

#The low R-squared value (0.01745) indicates that the 'Test_Control' variable explains only a small proportion of the variance in 'pur,' suggesting that there may be other factors not accounted for in this model.

```
#Intercept (Intercept): 0.022270
```

```
#Coefficient for Test_Control1 (Test_Control1): 0.058602
```

```
#Linear regression for Pur based on Test_control + Email + Address
```

```
out_lm2 <- lm(pur ~ Test_Control + Email + Address , data = abddata)
```

```
summary(out_lm2)
```

```
#The model is statistically significant and explains a small portion of the variance in 'pur' (2.268%).
```

```
#Intercept (Intercept): 0.010703
```

```
#Coefficient for Test_Control1 (Test_Control1): 0.057623
```

```
#Coefficient for Email (Email): 0.036416
```

```
#Coefficient for Address (Address): 0.016873
```

```
#Linear regression for pur based on Test_control * Email + Test_control * Address
```

```
out_lm3 <- lm(pur ~ Test_Control* Email + Test_Control* Address , data = abddata)
```

```
summary(out_lm3)
```

```
#Intercept (Intercept): 0.016844
```

```
#Coefficient for Test_Control1 (Test_Control1): 0.045338
```

```
#Coefficient for Email (Email): 0.007599
```

```
#Coefficient for Address (Address): 0.010301
```

```
#Coefficient for the interaction between Test_Control1 and Email (Test_Control1:Email): 0.052981
```

```
#Coefficient for the interaction between Test_Control1 and Address (Test_Control1:Address): 0.013011
```

```
#The R-squared value of 0.02466 indicates that the model explains only a small portion of the variance in the "pur" variable, suggesting that other factors not included in the model may influence purchase decisions.
```

```
#ANOVA
```

```
abddata$Test_Control = as.factor(abddata$Test_Control)
```

```
abddata$pur = as.numeric(abddata$pur)
```

```
model <- aov( pur~Test_Control, data = abddata)
```

```
summary(model)
```

```
#The ANOVA results demonstrate that the 'Test_Control' variable significantly affects  
the 'pur' variable.
```

```
#The small p-value (< 0.05) provides strong evidence to reject the null hypothesis,  
indicating that 'Test_Control' has a substantial impact on 'pur'.
```

```
#The high F-statistic value (149.9) further supports the significance of this  
relationship.
```

```
# Generate summary table
```

```
stargazer(out_lm1, out_lm2, out_lm3, type = "html", out = "midterm_proj.html")
```