QMB: 6304 (FINAL PROJECT)

The Advertising Campaign Analysis

Deepali Rajput

(U70271982)

Table of Contents

Background: A prominent travel agency has conducted an advertising campaign. Two datasets - 'Abandoned.csv' (ABD) and 'Reservation.csv' (RES) - provide insights into customer interactions and outcomes from this campaign.

Objective: Determine the statistical success of a retargeting campaign by matching and analyzing the datasets.

Introduction

'Abandoned.csv' contains data about customers who engaged but didn't purchase a vacation package. Notice the potential missing data and duplicates. These customers were divided into test and control groups for a retargeting campaign. 'Reservation.csv' documents customers who eventually purchased vacation packages.

Task: Establish whether the retargeting campaign was statistically effective.

Business Justification

1. **Explain Why Retargeting Customers Who Initially Didn't Buy A Package Makes Business Sense.**

   Justification: Retargeting aims to recapture customers who initially showed interest but did not make a purchase. In marketing, individuals who already interacted with a product or service are considered "warm leads." These leads have a higher probability of converting compared to entirely new leads. For the travel agency, retargeting those who abandoned their purchase can increase conversion rates while minimizing advertising costs, as these customers are familiar with the brand and may only need a slight nudge to complete the purchase. An effective retargeting strategy could include personalized ads, reminders, or incentives to overcome any previous hesitation. This method is a strategic business decision because it focuses resources on customers who are likely to respond positively, thus improving the campaign's cost-efficiency and return on investment (ROI).

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

```r
# Deepali Rajput (U70271982)

# Loading necessary libraries
library(dplyr)
library(stargazer)

# Reading data
abnd <- read.csv("Abandoned_Data.csv", header = TRUE, na.strings = "")
res <- read.csv("Reservation_Data.csv", header = TRUE, na.strings = "")

# Viewing data
View(abnd)
View(res)
```

Abandoned data (abnd):

Filter

	Caller_ID	Session	First_Name	Last_Name	Street	City	Address	Zipcode	Email
1	99374869QDSERAAM	2014.01.14 13:59:57	Marlen	Jacobs	432 Cassandra Stravenue	New Priscilla	WV	91357	awardprocessingoffice@yahoo.co.uk
2	41069893LJBAJSXY	2014.01.10 14:21:02	Humberto	NA	NA	NA	NA	NA	NA
3	44770091IMOCGIFM	2014.02.04 05:39:02	Eldon	NA	NA	NA	NY	NA	NA
4	46075824GHWLITOB	2014.01.07 23:41:25	Felicita	NA	NA	NA	NA	NA	NA
5	19329061DKRDHXSG	2014.01.12 05:27:14	Zita	McCullough	NA	NA	NA	NA	NA
6	68404344HPMADDDM	2014.02.02 08:40:36	Alexanne	NA	NA	NA	NA	NA	NA
7	77262308FQQJDWEL	2014.01.22 22:14:22	Serenity	NA	NA	NA	NA	NA	NA
8	91322765NVELDXFK	2014.01.11 12:05:17	Vaughn	Donnelly	NA	NA	NA	NA	NA
9	69018590HMPKOSRM	2014.01.22 16:00:47	Winifred	NA	NA	NA	NA	NA	NA
10	22686703CLSIJIMO	2014.01.13 03:22:07	Anahi	NA	NA	NA	VA	NA	NA
11	47134279RJJHDZJC	2014.01.21 19:44:00	Clementina	NA	NA	NA	DE	NA	NA
12	10939041COTKLMAN	2014.01.18 23:01:47	Marcelo	NA	NA	NA	NA	NA	NA
13	45501194CKLPUWDD	2014.02.02 18:44:21	Chad	NA	NA	NA	WA	NA	NA
14	43706992SMKUOANM	2014.01.31 13:20:09	Jessie	Bayer	NA	NA	NA	NA	NA
15	30179343RBGHAPOR	2014.01.21 18:52:28	Bert	NA	NA	NA	MT	NA	NA
16	52950182CRDKTNKA	2014.01.16 19:05:23	Dane	NA	NA	Botsfordshire	IA	14716	NA
17	94146871TDFNVSBT	2014.01.12 17:06:56	Lizzie	NA	NA	NA	NA	NA	NA

Showing 1 to 17 of 8,442 entries, 12 total columns

Incoming_Phone	Contact_Phone	Test_Control
(201)-050-5120	(201)-050-5120	test
(201)-114-9817	(201)-114-9817	test
(201)-131-2383	(201)-131-2383	control
(201)-158-0060	(201)-158-0060	control
(201)-244-9836	(201)-244-9836	test
(201)-259-4230	(201)-259-4230	test
(201)-319-0408	(201)-319-0408	control
(201)-351-7247	(201)-351-7247	test
(201)-358-6788	(201)-358-6788	control
(201)-490-1365	(201)-490-1365	test
(201)-522-7496	(201)-522-7496	control
(201)-565-0729	(201)-565-0729	test
(201)-593-4842	(201)-593-4842	test
(201)-650-7479	(201)-650-7479	control
(201)-669-7166	(201)-669-7166	test
(201)-772-3074	(201)-772-3074	control
(201)-847-5254	(201)-847-5254	control

Reservation data (res):

	Caller_ID	Session	First_Name	Last_Name	Street	City	Address	Zipcode	Email
1	34014222NEBBRWHK	2014.02.04 10:07:05	Chad	Gaylord	634 Wyman Lane	East Cory	CO	17314-1301	johnuwalaka2008@live.com
2	17286517QAFGZEHE	2014.02.04 10:10:35	Estell	Littel	7170 Yost Valley	Wilfredoview	AL	16060	mrphilllipeoa@yahoo.co.jp
3	96550704BWMSBBYI	2014.02.04 10:14:32	Jo	Grant	23315 Bogisich Knoll	East Werner	MO	69115-6141	meh.edwige1984@yahoo.com
4	95079589IXXPLXAT	2014.02.04 10:18:33	Taya	Koelpin	93838 Hazel Meadows	Windlerburgh	WV	15945	robert.clarkson73@yahoo.com.hk
5	77216124MZPECQVI	2014.02.04 10:22:34	Enoch	Johnson	136 Susan Locks	Wolffort	FL	50389	moler1940za@myway.com
6	03752891GNQXFONQ	2014.02.04 10:26:35	Damian	Cronin	848 Jewell Divide	Boganfort	ID	73748	mr.benjamind@live.com
7	22933340LROUBSFR	2014.02.04 10:30:40	Marcia	Langworth	NA	NA	NA	NA	claimservice01@9.cn
8	22302738NSTTNEHG	2014.02.04 10:35:07	Camille	Paucek	767 Armstrong Turnpike	West Juston	VA	37333-2107	julen_kown3@yahoo.cn
9	07191808ZEDKFSCV	2014.02.04 10:41:13	Michaela	Kirlin	581 Cassin Fords	West Luisaton	KS	24613-7833	atmswiftcardpaymentcentcbn1@gmail.com
10	28937780DYNDQEVD	2014.02.04 10:47:13	Lisa-Test-ID-1223	WebTest	9805 Leonardo Ranch	Boscoport	CA	84043	paulsonrichard77@yahoo.com.ph
11	26311496PHYAZNTE	2014.02.04 10:51:45	Sandrine	Kuphal	190 Oberbrunner Track	Hadleyfurt	ND	31719-4292	webkano25@att.net
12	78908033UBFUWDYW	2014.02.04 10:56:12	Cheyenne	Schowalter	4915 Ida Mount	Bauchside	NM	28703	frontarols@terra.es
13	92604443GSWKHKBM	2014.02.04 11:01:12	Caroline	O'Hara	3580 Marquardt Rue	Handshire	DE	99093-8260	info_christianaidenquiryunit06@yahoo.co.uk
14	58323926BCNGJMTR	2014.02.04 11:05:48	Marilie	Collins	50539 Moen Terrace	South Vesta	NY	29394-7138	philipezaoma@ibibo.com
15	41928070LFULYLCL	2014.02.04 11:10:47	Elroy	Hahn	84151 Schultz Green	East Celestino	CO	17698-3071	joymary7777@msn.com
16	69043303FAVYGYJG	2014.02.04 11:17:12	Elsie	Crist	14402 Weissnat Drive	O'Reillyview	PA	31928	mike.mullen303@rocketmail.com
17	08328108QSFIYAZJ	2014.02.04 11:20:59	Dora	Casper	926 Kozey Overpass	East Lucile	NV	68247-2047	jessica3_oflove@yahoo.com

Showing 1 to 17 of 20,814 entries, 12 total columns

Incoming_Phone	Contact_Phone	Test_Control
(614)-714-8068	NA	test
(262)-184-3193	NA	test
(248)-367-3066	NA	test
(717)-048-9487	NA	test
(830)-234-7472	NA	test
(212)-925-8173	NA	test
(401)-919-1169	NA	test
(229)-599-5178	NA	test
(479)-427-1300	NA	test
(937)-313-9577	NA	test
(272)-054-5943	NA	test
(234)-751-0870	NA	test
(973)-556-6908	NA	test
(262)-819-8744	NA	test
(323)-434-6911	NA	test
(408)-274-7340	NA	test
(786)-304-5397	NA	test

```
# Analyze Test/Control Division
# Step 1: Checking counts of each group (Test and Control)
test_control_counts <- table(abnd$Test_Control)
print("Counts of each group (Test and Control):")
print(test_control_counts)

# Step 2: Calculating proportions for each group
test_control_proportions <- prop.table(test_control_counts)
print("Proportion of each group:")
print(test_control_proportions)

# Step 3: Performing a chi-square test to assess if the division is approximately 50/50
# Assuming a 50/50 split is intended
expected_proportions <- c(0.5, 0.5)
chi_square_test <- chisq.test(test_control_counts, p = expected_proportions)
print("Chi-square test results for Test/Control balance:")
print(chi_square_test)
```

```
> print(test_control_counts)

control    test
   4176    4266

> print(test_control_proportions)

  control      test
0.4946695 0.5053305

> print(chi_square_test)

        Chi-squared test for given probabilities

data:  test_control_counts
X-squared = 0.95949, df = 1, p-value = 0.3273
```

Observation: Yes, the code seems to be well-executed.

1. Control group has 4176 counts and Test group has 4266 counts. These counts are very close, indicating an almost equal division between the test and control groups.
2. The calculated proportions are: Control group: 0.4947 (49.47%) and Test group: 0.5053 (50.53%). These proportions are also very close to 50/50, suggesting a well-balanced division.
3. The chi-square test results show a chi-squared statistic of **0.9595** with **1 degree of freedom** and a p-value of **0.3273**. Since the p-value is greater than 0.05, there is no statistically significant difference between the observed distribution of test and control groups and the expected 50/50 split. This means we cannot reject the null hypothesis, which assumes that the groups are balanced.

Overall, the test/control split seems well-executed and fair, as the proportions for each group are close to the expected 50%. The chi-square test supports that any small difference observed is not statistically significant, indicating an unbiased allocation between the test and control groups.

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

```
# Q3: Compute summary statistics for the test variable, segmented by available State data

# Checking distribution of test and control groups
table(abnd$Test_Control)

# Summary statistics segmented by State. Using "Address" as a proxy for state
test_control_stats <- abnd %>%
  group_by(Address, Test_Control) %>%
  summarise(count = n())
test_control_stats
```

```
   Address Test_Control count
   <chr>   <chr>        <int>
 1 AK      control         32
 2 AK      test            29
 3 AL      control         42
 4 AL      test            38
 5 AR      control         46
 6 AR      test            38
 7 AZ      control         44
 8 AZ      test            54
 9 CA      control         37
10 CA      test            48
# ℹ 92 more rows
```

Observation: The count shows the number of observations in each test/control group for each state (address).

Data Alignment

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

```r
# Task 2
# Q4: Assuming 'Email', 'Incoming_Phone', 'Contact_Phone' as matching keys
# Match based on different keys and create logical vectors for each condition
match_email <- abnd$Email[complete.cases(abnd$Email)] %in% res$Email[complete.cases(res$Email)]
match_incoming <- abnd$Incoming_Phone[complete.cases(abnd$Incoming_Phone)] %in% res$Incoming_Phone[complete.cases(res$Incoming_Phone)]
match_contact <- abnd$Contact_Phone[complete.cases(abnd$Contact_Phone)] %in% res$Contact_Phone[complete.cases(res$Contact_Phone)]
match_incoming_contact <- abnd$Incoming_Phone[complete.cases(abnd$Incoming_Phone)] %in% res$Contact_Phone[complete.cases(res$Contact_Phone)]
match_contact_incoming <- abnd$Contact_Phone[complete.cases(abnd$Contact_Phone)] %in% res$Incoming_Phone[complete.cases(res$Incoming_Phone)]
```

```r
# Creating flags for matches using the specified pattern
abnd$match_email <- 0
abnd$match_email[complete.cases(abnd$Email)] <- 1 * match_email

abnd$match_incoming <- 0
abnd$match_incoming[complete.cases(abnd$Incoming_Phone)] <- 1 * match_incoming

abnd$match_contact <- 0
abnd$match_contact[complete.cases(abnd$Contact_Phone)] <- 1 * match_contact

abnd$match_incoming_contact <- 0
abnd$match_incoming_contact[complete.cases(abnd$Incoming_Phone)] <- 1 * match_incoming_contact

abnd$match_contact_incoming <- 0
abnd$match_contact_incoming[complete.cases(abnd$Contact_Phone)] <- 1 * match_contact_incoming
```

```r
# Logical selection for matching records
abnd$pur <- 1 * (abnd$match_email | abnd$match_incoming | abnd$match_contact |
                 abnd$match_incoming_contact | abnd$match_contact_incoming)

# Create additional columns for analyses
abnd$email <- 1 * complete.cases(abnd$Email)
abnd$state <- 1 * complete.cases(abnd$Address) # Using 'Address' as proxy for state or location data
abnd$treat <- 1 * (abnd$Test_Control == "test") # 1 for treatment (test) group, 0 for control group
```

Incoming_Phone	Contact_Phone	Test_Control	match_email	match_incoming	match_contact	match_incoming_contact	match_contact_incoming
(864)-004-6354	(864)-004-6354	test	0	0	0	0	0
(703)-220-0148	(703)-220-0148	control	0	0	0	0	0
(559)-299-7745	(559)-299-7745	control	0	0	0	0	0
(636)-611-4439	(636)-611-4439	test	0	0	0	0	0
(253)-461-5118	(253)-461-5118	control	0	0	0	0	0
(407)-910-9280	(407)-910-9280	test	0	0	0	0	0
(803)-853-6182	(803)-853-6182	test	0	0	0	0	0
(631)-808-0736	(631)-808-0736	test	0	0	0	0	0
(918)-738-2706	(918)-738-2706	control	0	0	0	0	0
(440)-480-7247	(440)-480-7247	test	0	0	0	0	0
(929)-150-0791	(929)-150-0791	control	0	0	0	0	0
(813)-434-7170	(813)-434-7170	test	0	0	0	0	0
(530)-629-3863	(530)-629-3863	control	0	0	0	0	0
(757)-759-5303	(510)-985-8923	test	0	0	1	0	1
(430)-237-2099	(430)-237-2099	control	0	0	0	0	0
(857)-002-7905	(857)-002-7905	test	0	0	0	0	0
(303)-643-3078	(303)-643-3078	test	0	0	0	0	0
(713)-652-3296	(713)-652-3296	control	0	0	0	0	0

Showing 1 to 18 of 8,442 entries, 21 total columns

pur	email	state	treat
0	0	0	1
0	0	0	0
0	0	0	0
0	0	0	1
0	0	0	0
0	0	0	1
0	0	0	1
0	0	1	1
0	0	0	0
0	0	1	1
0	0	1	0
0	0	0	1
0	0	0	0
1	0	1	1
0	0	0	0
0	0	1	1
0	0	0	1
0	0	0	0

Observation:

1. Email Matching: Email is used as the primary match key due to its uniqueness and stability across records. A binary column is created to indicate email matches between the abandoned and reservation datasets, where a value of 1 signifies a match and 0 indicates no match. This approach efficiently flags corresponding records based on email.

2. Phone Number Matching: Phone number matching is performed across different fields (e.g., "Incoming_Phone" and "Contact_Phone") to account for potential variations in labeling between the two datasets. Separate binary columns are created for each type of match, where 1 represents a match and 0 indicates no match. This ensures that phone-based matching is comprehensive, identifying customers even when phone numbers are recorded in different fields.

3. Aggregated Matching Indicator: A consolidated column, referred to as the "pur" flag, is established to indicate if any match exists based on email or phone number fields. If any of the match conditions are met, the pur flag is set to 1 (indicating a match), and if none are met, it is set to 0 (no match). This aggregation simplifies the classification of records as either matched or unmatched for further analysis.

4. Additional Variables for Analysis:

   Email and State Flags: Binary columns are created to show the availability of email and state information, where 1 indicates that the data is available, and 0 represents missing data. This helps in segmenting and analyzing records based on data completeness.
   Treatment Flag (treat): A binary column is created to identify the group assignment, with 1 representing the treatment group (customers targeted by the campaign) and 0 representing the control group. This flag is essential for analyzing the impact of the retargeting campaign on customer outcomes.

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

    a. **Treatment group who purchased.**

    b. **Treatment group who didn't purchase.**

    c. **Control group who purchased.**

    d. **Control group who didn't purchase.**

```r
# Q5: Identify customers in each group
# Treatment group who purchased
treatment_purchased <- abnd %>% filter(Test_Control == "test", pur == 1) # 1 for matched rows

# Treatment group who didn't purchase
treatment_not_purchased <- abnd %>% filter(Test_Control == "test", pur == 0) # 0 for unmatched rows

# Control group who purchased
control_purchased <- abnd %>% filter(Test_Control == "control", pur == 1) # 1 for matched rows

# Control group who didn't purchase
control_not_purchased <- abnd %>% filter(Test_Control == "control", pur == 0) # 0 for unmatched rows
```

For treatment_purchased:

	match_contact_incoming	pur	email	state	treat
1	1	1	0	1	1
2	0	1	0	1	1
3	0	1	0	0	1
4	1	1	0	1	1
5	1	1	0	0	1
6	1	1	1	1	1
7	1	1	0	1	1
8	0	1	0	0	1
9	0	1	0	0	1
10	1	1	0	0	1
11	1	1	0	0	1

For treatment_not_purchased:

match_incoming_contact	match_contact_incoming	pur	email	state	treat
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	0	1	1
0	0	0	0	1	1
0	0	0	0	0	1
0	0	0	0	1	1
0	0	0	0	0	1
0	0	0	0	1	1
0	0	0	0	1	1

For control_purchased:

	match_contact_incoming	pur	email	state	treat
1	1	1	0	1	0
2	0	1	1	1	0
3	0	1	0	1	0
4	1	1	0	0	0
5	1	1	1	1	0
6	1	1	1	0	0
7	1	1	0	1	0
8	1	1	0	0	0
9	1	1	0	1	0
10	0	1	0	1	0
11	1	1	0	1	0

For control_not_purchased:

match_contact_incoming	pur	email	state	treat
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	1	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Observation:

1. Treatment Group Who Purchased: The dataset was filtered to identify customers in the treatment group (Test_Control == "test") who made a purchase (pur == 1). This selection isolates customers who were exposed to the retargeting intervention and subsequently converted, providing a critical group for evaluating the campaign's effectiveness on those who made a purchase.
2. Treatment Group Who Didn't Purchase: Filtering criteria for Test_Control == "test" and pur == 0 were applied to identify customers in the treatment group who did not make a purchase. Here, pur == 0 indicates no match with the reservation data, signifying non-conversion. This segment represents treatment group customers who did not respond to the intervention, allowing for an analysis of potential factors that may have hindered conversion.
3. Control Group Who Purchased: The dataset was filtered to include customers in the control group (Test_Control == "control") who made a purchase (pur == 1). This group serves as a natural baseline for purchase behavior without exposure to the treatment, enabling a comparison with the treatment group to assess the incremental effect of the campaign on conversions.
4. Control Group Who Didn't Purchase: Filtering was applied for Test_Control == "control" and pur == 0, capturing control group customers who did not make a purchase. This group provides a baseline for non-conversion without the campaign's influence, essential for understanding natural non-conversion rates. Comparing this baseline to the treatment group helps assess the true impact of the retargeting

intervention on purchase behavior.

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

```r
# Q6: Identify unmatched records and examples
unmatched <- abnd % >% filter(pur == 0)
print(unmatched)

# Remove unmatched records for further analysis
abnd <- abnd %>% filter(pur == 1)
View(abnd)
```

```
    Contact_Phone Test_Control match_email match_incoming match_contact match_incoming_contact match_contact_incoming pur
1  (864)-004-6354         test           0              0             0                      0                      0   0
2  (703)-220-0148      control           0              0             0                      0                      0   0
3  (559)-299-7745      control           0              0             0                      0                      0   0
4  (636)-611-4439         test           0              0             0                      0                      0   0
5  (253)-461-5118      control           0              0             0                      0                      0   0
6  (407)-910-9280         test           0              0             0                      0                      0   0
7  (803)-853-6182         test           0              0             0                      0                      0   0
8  (631)-808-0736         test           0              0             0                      0                      0   0
9  (918)-738-2706      control           0              0             0                      0                      0   0
10 (440)-480-7247         test           0              0             0                      0                      0   0
11 (929)-150-0791      control           0              0             0                      0                      0   0
12 (813)-434-7170         test           0              0             0                      0                      0   0
13 (530)-629-3863      control           0              0             0                      0                      0   0
14 (430)-237-2099      control           0              0             0                      0                      0   0
15 (857)-002-7905         test           0              0             0                      0                      0   0
16 (303)-643-3078         test           0              0             0                      0                      0   0
17 (713)-652-3296      control           0              0             0                      0                      0   0
```

After removing unmatched records (pur =0)

Incoming_Phone	Contact_Phone	Test_Control	match_email	match_incoming	match_contact	match_incoming_contact	match_contact_incoming	pur
(757)-759-5303	(510)-985-8923	test	0	0	1	0	1	1
(402)-153-4684	(619)-074-3663	test	0	1	0	0	0	1
(703)-986-0864	(518)-375-2652	test	0	0	1	0	0	1
(830)-998-3332	(830)-998-3332	test	0	1	1	1	1	1
(775)-329-0338	(775)-329-0338	test	0	1	1	1	1	1
(814)-861-7221	(814)-861-7221	test	0	1	0	0	1	1
(248)-549-8764	(248)-549-8764	test	0	1	1	1	1	1
(385)-720-2094	(423)-618-4176	test	0	1	0	0	0	1
(956)-919-2793	(270)-358-4201	test	0	0	1	0	0	1
(916)-824-7278	(916)-824-7278	test	0	1	1	1	1	1
(878)-091-4844	(878)-091-4844	control	0	1	1	1	1	1
(484)-705-2967	(484)-705-2967	test	0	1	0	0	1	1
(210)-420-3060	(210)-420-3060	test	0	1	0	0	1	1
(803)-748-6444	(803)-748-6444	test	0	1	1	1	1	1
(270)-976-8746	(270)-976-8746	test	0	1	0	0	1	1
(615)-690-7091	(615)-690-7091	test	1	1	0	0	1	1
(785)-484-8767	(785)-484-8767	test	0	1	1	1	1	1

Showing 1 to 17 of 438 entries, 21 total columns

Observation: The original dataset contained 8,442 records. After excluding unmatched records, only 438 records remain, indicating that a substantial majority of records (8,004) did not have a corresponding match in the reservation data.

1. Unmatched Records: The initial step identifies records where pur == 0, representing customers who did not have any matches in the reservation dataset. Specifically, these unmatched records lack any matching Email, Incoming_Phone, or Contact_Phone, or fail cross-matching conditions between these fields across datasets.

   For instance, the first row with Contact_Phone "(864)-004-6354" in the treatment group (Test_Control == "test") and the second row with Contact_Phone "(703)-220-0148" in the control group (Test_Control == "control") both have pur == 0, indicating no match. Similarly, other records with varying contact phone numbers do not satisfy any of the matching conditions (match_email, match_incoming, match_contact, etc.), resulting in pur remaining 0.

2. Matched Records: Filtering for records where pur == 1 isolates matched records, resulting in 438 out of the initial 8,442 records. These matched entries represent customers from the abandoned dataset who were successfully identified in the reservation dataset through matching on fields such as Email, Incoming_Phone, or Contact_Phone. This subset of 438 matched records is now prepared for further analysis, representing a smaller but crucial segment for evaluating the campaign's impact.

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

```
# Q7: Cross-tabulation of outcomes for treatment and control groups
cross_tab <- abnd %>%
  group_by(treat, pur) %>%
  summarise(count = n()) %>%
  ungroup()
print(cross_tab)
```

```
  treat   pur count
  <dbl> <dbl> <int>
1     0     1    93
2     1     1   345
```

Observation:

1. The table provides a cross-tabulation of outcomes for two groups: a control group (where treat = 0) and a treatment group (where treat = 1).
2. In the control group, there were 93 instances where the outcome (pur) was 1.
3. In the treatment group, there were 345 instances where the outcome (pur) was 1.

The treatment group shows a significantly higher count of the positive outcome (pur = 1) compared to the control group. The large difference in outcome counts between the treatment and control groups hints that the treatment could be effective.

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

```r
# Q8: Cross-tabulation for five randomly chosen addresses (for state)
set.seed(1982)  # last 4 digits of U number to set seed for Reproducibility
selected_addresses <- sample(unique(abnd$Address), 5)
address_cross_tab <- abnd %>%
  filter(Address %in% selected_addresses) %>%
  group_by(Address, treat, pur) %>%
  summarise(count = n())
print(address_cross_tab)
```

	Address	treat	pur	count
	<chr>	<dbl>	<dbl>	<int>
1	AR	0	1	1
2	AR	1	1	3
3	ID	1	1	3
4	LA	1	1	2
5	MN	0	1	1
6	MN	1	1	3
7	NJ	0	1	3
8	NJ	1	1	5

Detailing the Selections:

A seed (set.seed(1982)) was set for reproducibility to ensure that the same states are selected every time the code is run. The seed value is based on the last four digits of my U number, adding a personalized element to the selection process.

Sampled States: The sample function was used to select five unique states randomly from the Address column in the dataset. In this output, the selected states are AR, ID, LA, MN, and NJ.

Reason for choosing these states: Since the selection is random, these five states represent a subset of the larger dataset without any systematic bias. This provides a cross-sectional view of treatment and control group outcomes across different geographical areas which allows us to observe any potential variation in treatment effects across locations.

Observation: The randomly chosen states in this cross-tabulation are AR, ID, LA, MN, and NJ.

Treatment and Control Counts:

AR: The control group (treat = 0) has 1 count, while the treatment group (treat = 1) has 3 counts.

ID: Only the treatment group (treat = 1) is observed, with 3 counts.

LA: Only the treatment group (treat = 1) is observed, with 2 counts.

MN: The control group (treat = 0) has 1 count, and the treatment group (treat = 1) has 3 counts.

NJ: The control group (treat = 0) has 3 counts, and the treatment group (treat = 1) has 5 counts.

For states where both control and treatment groups are present, the treatment group consistently has a higher count than the control group. This suggests that the treatment group may be more likely to achieve the outcome.

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.)***

```r
# Task 3:
# Q9: Generating cleaned dataset
cleaned_data <- abnd %>%
  select(Caller_ID, treat, pur, state, email)

write.csv(cleaned_data, "cleaned_data.csv", row.names = FALSE)
View(cleaned_data)
```

	Caller_ID	treat	pur	state	email
1	03241649AHZKPWYH	1	1	1	0
2	85080592TEFIPACV	1	1	1	0
3	83559451LHCUAFYT	1	1	0	0
4	18086538MZFGFFTH	1	1	1	0
5	38297698NQJIEDHS	1	1	0	0
6	36854393GIZMEDRD	1	1	1	1
7	05334034DMHRGBJP	1	1	1	0
8	72535168IUDJYABX	1	1	0	0
9	32597460SCPZKXYI	1	1	0	0
10	56895604BZVXIOOY	1	1	0	0
11	99131886JEWYGEJQ	0	1	1	0
12	26694082KLLWJQTW	1	1	0	0
13	30796839YRFDNQMX	1	1	1	0
14	12735352AZTUHXTW	1	1	0	0
15	81997214OOEKQCCZ	1	1	1	0
16	29788310DXSYOYWT	1	1	1	1
17	92486801SHVFQQPV	1	1	1	0
18	47678843KOIPKWGW	0	1	1	1
19	02097717MEMMKLLQ	1	1	1	1
20	41778285YYJBRGED	0	1	1	0

howing 1 to 20 of 438 entries, 5 total columns

Observation: The cleaned dataset had 438 entries and each row corresponds to a matched customer (pur = 1) from the datasets. Here, Customer ID - Caller ID, Test Group - treat, Outcome - pur, State Available - state, Email Available - email. The data is stored in csv file which is attached during the submission.

Statistical Assessment

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

```r
# Task 4
# Q10: Execute a linear regression for Outcome = α + β * Test Group + error
# Model 1: Basic linear regression with only the treatment group as the predictor
linear_model <- lm(pur ~ treat, data = cleaned_data)
summary(linear_model)
```

```
Call:
lm(formula = pur ~ treat, data = cleaned_data)

Residuals:
       Min         1Q     Median         3Q        Max 
-6.940e-17 -6.940e-17 -6.940e-17 -6.940e-17  2.386e-14 

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)    
(Intercept) 1.000e+00  1.187e-16 8.427e+15   <2e-16 ***
treat1      6.936e-17  1.337e-16 5.190e-01    0.604    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.144e-15 on 436 degrees of freedom
Multiple R-squared:  0.4995,	Adjusted R-squared:  0.4983 
F-statistic: 435.1 on 1 and 436 DF,  p-value: < 2.2e-16
```

Observation: The intercept value is 1.000 and highly significant (p-value < 2e-16). This indicates that the baseline outcome for the control group is estimated with high precision and is reliably close to 1.

The treatment effect coefficient is 6.936e-17, a value very close to zero, with a high p-value (0.604). This result suggests that the effect of the treatment group is not statistically significant. Therefore, there is no strong evidence that the treatment has an impact on the outcome.

The model's R-squared is 0.4995, and adjusted R-square is 0.4983 shows that about 50% of the variance in the outcome is explained by the model. However, this explanatory power largely comes from the intercept (the control group outcome), rather than the treatment effect.

F-statistic: The F-statistic is high, with a very low p-value (< 2.2e-16) shows that the model as a whole is statistically significant. However, this significance does not extend to the treatment variable itself.

Overall, this model does not provide support for a significant effect of the treatment on the outcome. While the overall model fit is statistically robust, the treatment effect is negligible which suggests that additional factors or approaches might need to be considered.

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

```r
# Q11: Justification of this regression as statistically comparable to an ANOVA/t-test
# Question 11: Perform an independent t-test and ANOVA
# Independent T-test
t_test_result <- tryCatch({
  t.test(pur ~ treat, data = cleaned_data)
}, error = function(e) { "T-test not applicable due to lack of outcome variation" }) #since the outcome variable only has matched values pur = 1
print(t_test_result)


# ANOVA
anova_model <- aov(pur ~ treat, data = cleaned_data)
anova_result <- summary(anova_model)
print("ANOVA result:")
print(anova_result)
```

```
> print(t_test_result)
[1] "T-test not applicable due to lack of outcome variation"
> # ANOVA
> anova_model <- aov(pur ~ treat, data = cleaned_data)
> anova_result <- summary(anova_model)
> print("ANOVA result:")
[1] "ANOVA result:"
> print(anova_result)
             Df    Sum Sq   Mean Sq F value Pr(>F)
treat         1 4.000e-31 3.524e-31   0.269  0.604
Residuals   436 5.709e-28 1.309e-30
```

Observations and justification:

1. The previous regression output shows a t-value of 0.519 and a p-value of 0.604 for the treat coefficient, indicating that any difference between the treatment and control groups is statistically insignificant.
2. In the ANOVA output, the F-value for treat is 0.269 with a p-value of 0.604, which aligns with the regression results that show no significant effect of the treatment on the outcome.
3. The T-test is not applicable in this case due to the lack of variation in the outcome (pur), as all values of pur are identical (pur = 1), leaving no variability to compare.

Statistical Comparability: All three methods—linear regression, ANOVA, and t-test—are conceptually testing the same hypothesis: whether the mean outcome (pur) differs between the treatment and control groups.

In each method, the null hypothesis is the same: there is no difference in the outcome (pur) between the treatment and control groups. The F-statistic from the ANOVA (0.269) and the t-value for treat in the regression (0.519) correspond to the same lack of significance. Both statistics have a p-value of 0.604, indicating that the treatment effect is not statistically different from zero.

This lack of significance is consistent in both the regression and ANOVA results which confirms that neither method finds a meaningful difference between groups. Because pur has no variation (all values are 1), the t-test cannot perform a meaningful comparison between groups, as there is no actual difference to measure. The linear regression and ANOVA gives output, but both show that any effect of treat on pur is statistically insignificant. Although the t-test could not be performed due to lack of variation in `pur`, the results from regression and ANOVA still clearly show that the treatment had no meaningful effect on the outcome. This further shows that, despite using different methods, the interpretation remains consistent: there is no meaningful difference in the outcome between treatment and control groups.

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

The observations from the regression model show that it's not suitable for making strong causal claims about the effectiveness of the retargeting campaign:

1. High Intercept but No Significant Treatment Effect: The intercept value is 1.000 and is highly significant, which means the outcome for the control group is accurately estimated and close to 1. But the treatment effect (6.936e-17) is extremely close to zero and not significant (p-value = 0.604). This means there's no evidence that the treatment (or retargeting campaign) had any real impact on the outcome. To make a causal claim, we'd need a clear, significant effect from the treatment, which isn't present in the model.
2. R-squared and adjusted R-squared: The R-squared (0.4995) and adjusted R-squared (0.4983) show

that about 50% of the outcome's variation is explained by the model. However, this is mostly due to the intercept, not the treatment effect. Since the treatment isn't explaining much of the outcome, it weakens any argument for causation. Basically, the model's explanatory power is coming from the control group outcome, not from the treatment's impact.

3. F-statistic and Overall Model Significance: The F-statistic is high with a very low p-value, meaning the overall model seems statistically significant. However, this significance doesn't apply to the treatment effect itself. In other words, while the model as a whole may look strong, the treatment variable doesn't have a significant effect, so it doesn't support any causal claims about the treatment.
4. Unable to Prove Causation: To make a strong causal claim, we'd need a significant treatment effect and a well-designed experiment (with things like randomization and control for other factors). This regression model doesn't provide that – it lacks a significant treatment effect, and it mostly explains the control outcome, not the treatment effect.

Therefore, this regression model doesn't support making causal claims about the campaign's effectiveness. Even though the model as a whole is statistically solid, it doesn't show any meaningful impact from the treatment, suggesting we'd need a different approach to make any causal conclusions.

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.

```
# Q13: Integrate State and Email dummies into the regression and add interaction terms
# Model 2: Adding state and email as predictors to control for their effects
out2 <- lm(pur ~ treat + state + email, data = cleaned_data)
summary(out2)
```

```
Residuals:
       Min         1Q     Median         3Q        Max 
-1.554e-16 -1.554e-16 -3.070e-17 -2.490e-17  2.377e-14 

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)    
(Intercept)  1.000e+00  1.343e-16  7.444e+15   <2e-16 ***
treat        7.917e-17  1.341e-16  5.900e-01    0.555    
state        1.305e-16  1.163e-16  1.122e+00    0.262    
email       -1.247e-16  1.408e-16 -8.860e-01    0.376    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.145e-15 on 434 degrees of freedom
Multiple R-squared:  0.5008,	Adjusted R-squared:  0.4974 
F-statistic: 145.2 on 3 and 434 DF,  p-value: < 2.2e-16
```

```
# Model 3: Adding interaction terms between treat and both state and email to explore potential moderating effects
out3 <- lm(pur ~ treat * state + treat * email, data = cleaned_data)
summary(out3)
```

```
Residuals:
       Min         1Q     Median         3Q        Max 
-1.833e-16 -1.833e-16 -9.700e-18 -9.700e-18  2.375e-14 

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)    
(Intercept)  1.000e+00  1.843e-16  5.427e+15   <2e-16 ***
treat        9.682e-18  2.066e-16  4.700e-02    0.963    
state       -7.520e-29  2.428e-16  0.000e+00    1.000    
email       -4.192e-30  3.268e-16  0.000e+00    1.000    
treat:state  1.736e-16  2.769e-16  6.270e-01    0.531    
treat:email -1.635e-16  3.627e-16 -4.510e-01    0.652    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.147e-15 on 432 degrees of freedom
Multiple R-squared:  0.501,	Adjusted R-squared:  0.4953 
F-statistic: 86.76 on 5 and 432 DF,  p-value: < 2.2e-16
```

```r
# Model 4: Full interaction model with all main effects and interactions among treat, state, and email
out4 <- lm(pur ~ treat * state * email, data = cleaned_data)
summary(out4)
```

```
Residuals:
       Min         1Q     Median         3Q        Max 
-1.945e-16 -1.945e-16  0.000e+00  0.000e+00  2.373e-14 

Coefficients:
                    Estimate Std. Error    t value Pr(>|t|)    
(Intercept)        1.000e+00  1.915e-16  5.221e+15   <2e-16 ***
treat             -7.475e-29  2.143e-16  0.000e+00    1.000    
state             -8.784e-29  2.610e-16  0.000e+00    1.000    
email             -7.349e-29  6.057e-16  0.000e+00    1.000    
treat:state        1.945e-16  2.968e-16  6.550e-01    0.513    
treat:email        9.095e-29  7.231e-16  0.000e+00    1.000    
state:email        9.807e-29  7.200e-16  0.000e+00    1.000    
treat:state:email -1.945e-16  8.391e-16 -2.320e-01    0.817    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.149e-15 on 430 degrees of freedom
Multiple R-squared:  0.5011,	Adjusted R-squared:  0.493 
F-statistic:  61.7 on 7 and 430 DF,  p-value: < 2.2e-16
```

```r
# Output summaries for each regression model
stargazer(linear_model, out2, out3, out4, type = "text", title = "Regression Analysis Results", out = "analysis_output.html")
```

Regression Analysis Results

	Dependent variable: pur (1)	(2)	(3)	(4)
treat	0.000	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
state		0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)
email		-0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)
treat:state			0.000	0.000
			(0.000)	(0.000)
treat:email			-0.000	0.000
			(0.000)	(0.000)
state:email				0.000
				(0.000)
treat:state:email				-0.000
				(0.000)
Constant	1.000***	1.000***	1.000***	1.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	438	438	438	438
R2	0.499	0.501	0.501	0.501
Adjusted R2	0.498	0.497	0.495	0.493
Residual Std. Error	0.000 (df = 436)	0.000 (df = 434)	0.000 (df = 432)	0.000 (df = 430)
F Statistic	435.099*** (df = 1; 436)	145.154*** (df = 3; 434)	86.763*** (df = 5; 432)	61.704*** (df = 7; 430)
Note:				*p<0.1; **p<0.05; ***p<0.01

Observation and Insights:

Model 1: Baseline Model

The first (previous) regression model includes only the treat variable, which shows a non-significant effect (p-value = 0.604). This suggests that there is no meaningful difference in the outcome (pur) between the treatment and control groups. The R-squared for this model is 0.499, and the adjusted R-squared is 0.498, meaning the model explains about 49.9% of the variation, but this is likely due to the intercept since there's no real impact from treat on the outcome.

Model 2: Adding State and Email Dummies

In the second model, state and email are added as predictors. Both state (p-value = 0.262) and email (p-value = 0.376) remain non-significant which shows that they don't add any meaningful explanation to the outcome. The R-squared increases slightly to 0.501, and the adjusted R-squared to 0.497, but this minor change doesn't suggest any real impact of state or email on pur.

Model 3: Including Interaction Terms (treat * state and treat * email)

This model introduces interaction terms between treat and both state and email. These interactions are also non-significant (p-values of 0.531 and 0.652, respectively) which shows that state and email don't influence the effect of treat. The R-squared remains 0.501 with a slight decrease in adjusted R-squared to 0.495 also suggests that these interactions don't add explanatory power to the model or affect the outcome.

Model: Full Interaction Model (treat * state * email)

In the final model, all interactions, including treat, state, email, are included. None of these interactions are significant. The R-squared is 0.5011, and the adjusted R-squared is 0.493, indicating no meaningful change in the model's ability to explain the outcome, and no sign of any impact from treat, state, email, or their interactions.

Insights:

No Significant Effects: Across all models, none of the predictors (treat, state, email) or their interactions show any significant impact on pur, meaning the treatment, state, and email do not explain changes in the outcome.

Minimal Changes in Model Fit: Adding state, email, and interaction terms results in only slight changes in R-squared and adjusted R-squared values, suggesting these additions don't improve the model's explanatory power.

Therefore, the consistent lack of significance and minor R-squared changes confirm that pur (the outcome) does not vary with treatment, state, or email status. This means that none of these factors have any measurable impact on the outcome, which further shows that the treatment itself does not affect pur. And, even after adding additional variables like state and email (along with their interactions with the treatment group) to the regression model, the overall conclusion remains the same: there is still no significant evidence that the retargeting campaign had any effect on the outcome. In other words, including these extra factors did not change the results – the campaign still doesn't appear to have impacted whether people made a purchase.

Reflections

14. Reflect on the project:

a. Would you modify the experiment design if given a chance?

Here are a few changes that could improve the analysis:

1. Collect data on both purchasers and non-purchasers: Including both types within the treatment and control groups would create more variability in the outcome (pur), making it easier to assess the treatment's impact.
2. Increase Sample Size: Expanding the sample size for both treatment and control groups would improve the statistical power of the analysis, making it easier to detect even small effects of the retargeting campaign. A larger sample would provide more reliable and generalizable insights into the campaign's effectiveness.
3. Segment participants by specific characteristics: Dividing customers by factors like browsing behavior, location, or purchase history could show if certain groups respond better to retargeting, helping to create more targeted and effective strategies.
4. Gather baseline metrics: Collecting initial data such as interest level or past engagement would help understand differences between groups and control for pre-existing factors, leading to more accurate insights on the treatment effect.

b. Could alternative paths be taken with better-quality data?

Here are a few alternative ways to improve the analysis:

1. Collect data on customer interactions: Track metrics such as time spent on the site, pages viewed, and cart abandonment reasons. This would allow an analysis of engagement patterns

that might influence retargeting effectiveness, highlighting behaviors linked to higher conversion likelihood.

2. Consider additional outcome measures: Beyond purchase data, track re-engagement rates, click-through rates on retargeting messages, and repeat purchases. These metrics provide a fuller picture of the campaign’s impact, showing areas of success and needed adjustments.
3. Gather demographic and psychographic data: Collecting information on customer demographics (e.g., age, location) and psychographics (e.g., interests, spending habits) enables more personalized analysis, helping identify which profiles respond best to retargeting for tailored marketing.
4. Analyze timing of retargeting interactions: Track data on the timing of retargeting efforts (e.g., time since cart abandonment, seasonality) to understand how timing affects purchase decisions, offering insights into the best times for retargeting messages.

c. Are there actionable business implications from this analysis?

Here are some actionable business implications from the analysis:

1. Refine the retargeting approach: The lack of significant impact suggests the current campaign may not be as effective as intended. The business could experiment with more personalized or segmented retargeting messages rather than a one-size-fits-all strategy.
2. Improve data collection practices: Gaps in data, such as limited outcome variation and customer profile information, indicate a need for better data collection. Gathering more detailed engagement, demographic, and behavioral data would support more insightful and reliable analyses for future campaigns.
3. Implement a multi-touch strategy: Single-touch retargeting may not be enough to convert abandoned cart customers. A multi-touch strategy across channels (e.g., emails, social media ads) could increase engagement and improve conversion rates.
4. Optimize timing of retargeting efforts: Testing different intervals for retargeting messages (e.g., immediately vs. a few days after abandonment) could reveal the optimal timing for re-engagement, increasing the likelihood of conversions.

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

I would rate my effort at 100%. Initially, I would have rated it at 99% to account for missing the original deadline due to final exams in other classes. However, since the professor co-operated really well and extended the deadline, I was able to fully focus on the project and put my best effort into every part of the analysis. With this extra time, I was able to ensure thoroughness and accuracy, so I feel confident in giving myself a full score of 100%.