U49253220

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

Retargeting abandoned customers will not be a problem until we can figure out why they decided not to buy even after calling. Generally, customers examine all the options before finally buying something. So even if the buyer did not buy, they might be interested in it to buy in the future.

So, even though these customers did not purchase this time, they may still be interested in the product in the future. Before retargeting, it would be wise to conduct a thorough analysis of both purchasers' and decliner's packages.

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

Yes, I have investigates the test/control variable and the experiment ran well.

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.

```
match_email=abd$Email[complete.cases(abd$Email)]%in%rsv$Email[complete.cases(rsv$Email)]
match_contact=abd$Contact_Phone[complete.cases(abd$Contact_Phone)]%in%rsv$Contact_Phone[co
mplete.cases(rsv$Contact_Phone)]
match_incoming_contact=abd$Incoming_Phone[complete.cases(abd$Incoming_Phone)]%in%rsv$Incomi
ng_Phone[complete.cases(rsv$Incoming_Phone)]

abd$match_email<-0
abd$match_email[complete.cases(abd$Email)]<-1*match_email
sum(abd$match_email)

[1] 75


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

[1] 185


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

[1] 327


abd$outcome<-0
abd$outcome<-1*(abd$match_email|abd$match_contact|abd$match_incoming_contact)
sum(abd$outcome)

[1] 420
```

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)

To match both abd and rsv some key attributes should be identified, also known as "data keys,"

In both data sets, certain keys can be used to specifically identify the customer. A few of these keys are:
1) [Email Address]
2) [Contact Number]
3. [Incoming Phone, Last Name]
4) [First Name, Last Name, Zip]

To identify the customers who have successfully converted to buy the package we should aggregate the data using the following keys.

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.

Data Matching Procedure :
Match the customers in both abandoned and reservation based on Email address
Match the customers in both abandoned and reservation based on Contact Phone
Match the customers in both abandoned and reservation based on combination of Incoming Phone and Contact phone

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

The primary attributes used to compare data between two datasets are:

1) [Email Address]
2) [Contact Number]
3. [Incoming Phone, Last Name]
4) [First Name, Last Name, Zip]

However, there are instances where the data of customers was not recorded. Then we should aggregate all customers after performing matches on each key.

Q8: Complete the following cross-tabulation:

```
control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0])


library(knitr)
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
kable(out)
```

```
| test_buy| test_no_buy| control_buy| control_no_buy|
|--------:|-----------:|-----------:|--------------:|
|      335|        3931|          85|           4091|
>
```

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

```
control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "CA"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"CA"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "CA"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"CA"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)
```

```
| test_buy| test_no_buy| control_buy| control_no_buy|
|--------:|-----------:|-----------:|--------------:|
|      152|        2205|          36|           2322|
```

```
control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "NY"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"NY"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "NY"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"NY"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)
```

```
| test_buy| test_no_buy| control_buy| control_no_buy|
|--------:|-----------:|-----------:|--------------:|
|      149|        2200|          37|           2320|
```

```
control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "FL"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"FL"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "FL"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"FL"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)
```

```
| test_buy| test_no_buy| control_buy| control_no_buy|
|--------:|-----------:|-----------:|--------------:|
|      149|        2198|          36|           2322|
```

```
control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "TX"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"TX"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "TX"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"TX"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)
```

```
| test_buy| test_no_buy| control_buy| control_no_buy|
|--------:|-----------:|-----------:|--------------:|
|      149|        2204|          36|           2318|
```

```
control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "NV"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"NV"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "NV"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"NV"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)
```

```
| test_buy| test_no_buy| control_buy| control_no_buy|
|--------:|-----------:|-----------:|--------------:|
|      151|        2203|          38|           2336|
```

Cleaning Data:

You have now identified all the relevant customers for the analysis and their outcome, and you also know if they are in a treated or in a control group.

Produce an Excel File with the following columns

Customer ID | Test Variable | Outcome | D_State | D_Email |

Where Test Variable indicates the treatment or the control group, the Outcome is a binary variable indicating whether a vacation package was ultimately bought. D_State and D_Email identify whether the information is present on file.

(Note that you should have as many rows as customers you were able to match across the two data sets. Be sure to attach this excel file to the submission for proper verification.)

```
abd$treatment <- NULL
abd$treatment <- ifelse(abd $Test_Control=="test",1,0)
summary(abd$treatment)
abd$d_email <- 1*complete.cases(abd$Email)
abd$d_state <- 1*complete.cases(abd$Address)

data_frame<- abd[c(1,16,17,18,19)]

colnames(data_frame) <- c('CustomerID','Outcome','TestVariable','D_Email','D_State')
```

Excel File Attached

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

```
install.packages("writexl")
library("writexl")
write_xlsx(df,"keep file location")

d1 = lm(abd$outcome~abd$treatment)
summary(d1)

Call:
lm(formula = abd$outcome ~ abd$treatment)

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.020354   0.003335   6.104 1.08e-09 ***
abd$treatment 0.058173   0.004691  12.401  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2155 on 8440 degrees of freedom
Multiple R-squared:  0.01789,    Adjusted R-squared:  0.01778
F-statistic: 153.8 on 1 and 8440 DF,  p-value: < 2.2e-16

>
```

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

Yes, It is statistically equivalent to ANOVA/t-test. This can be explained because both explain the same distribution of means.

Q12:Argue whether this is a properly specified linear regression model, and if so if we can draw any causal statement about the effectiveness of the retargeting campaign. Is this statistically significant?

As the test variable coefficient is equal to 0.3 and the dependent variable would only increase by 3% for every additional customer in the test group. This linear regression is not properly specified and also the R-squared value is only 0.01. So, to make a proper model one should include more variables.

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 Q9)

```
d2 = t.test(abd$outcome~abd$treatment)
summary(d2)
d3 = lm(abd$outcome~abd$treatment*abd$d_state*abd$d_email)
summary(d3)
```

```
> d2 = t.test(abd$outcome~abd$treatment)
> summary(d2)
            Length Class  Mode
statistic   1      -none- numeric
parameter   1      -none- numeric
p.value     1      -none- numeric
conf.int    2      -none- numeric
estimate    2      -none- numeric
null.value  1      -none- numeric
stderr      1      -none- numeric
alternative 1      -none- character
method      1      -none- character
data.name   1      -none- character

> d3 = lm(abd$outcome~abd$treatment*abd$d_state*abd$d_email)
> summary(d3)

Call:
lm(formula = abd$outcome ~ abd$treatment * abd$d_state * abd$d_email)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16274 -0.06313 -0.02654 -0.01453  0.98547
```

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              0.014532   0.004572   3.178  0.00149 **
abd$treatment                            0.048601   0.006490   7.489 7.66e-14 ***
abd$d_state                              0.012011   0.007173   1.674  0.09409 .
abd$d_email                              0.019081   0.020193   0.945  0.34471
abd$treatment:abd$d_state                0.003134   0.010134   0.309  0.75715
abd$treatment:abd$d_email               -0.017467   0.027571  -0.634  0.52641
abd$d_state:abd$d_email                 -0.019762   0.023887  -0.827  0.40808
abd$treatment:abd$d_state:abd$d_email  0.102605   0.032582   3.149  0.00164 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2146 on 8434 degrees of freedom
Multiple R-squared:  0.02719,    Adjusted R-squared:  0.02638
F-statistic: 33.68 on 7 and 8434 DF,  p-value: < 2.2e-16
```

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

Overall, it was a positive experience because it included data cleaning, in-depth analysis, statistical modeling, and inferences. Given that each customer has a unique ID assigned by the agency, matching could have been done only using this key, cutting down on the time needed for data matching and cleaning.

Managerial Implications: Retargeting certainly helped and can be seen. We can understand from the last a regression model that customers should be retargeted only in certain states and before a certain time to achieve a maximum probability of conversion.

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

I self evaluate myself between 90-95. I had to work on this assignment for the whole weekend to get this result. It was though for me and I used google in many instances to successfully complete the assignment.

Code:

```
rm(list = ls())

setwd("//Users//vivekvarma//Desktop//Mid")

abd<- read.csv("Abandoned.csv", header=T,na.strings = "")
rsv<- read.csv("Reservation.csv", header=T,na.strings = "")

match_email=abd$Email[complete.cases(abd$Email)]%in%rsv$Email[complete.cases(rsv$Email)]
match_contact=abd$Contact_Phone[complete.cases(abd$Contact_Phone)]%in%rsv$Contact_Phone[co
mplete.cases(rsv$Contact_Phone)]
match_incoming_contact=abd$Incoming_Phone[complete.cases(abd$Incoming_Phone)]%in%rsv$Incomi
ng_Phone[complete.cases(rsv$Incoming_Phone)]

abd$match_email<-0
abd$match_email[complete.cases(abd$Email)]<-1*match_email
sum(abd$match_email)

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

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

abd$outcome<-0
abd$outcome<-1*(abd$match_email|abd$match_contact|abd$match_incoming_contact)
sum(abd$outcome)

control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0])


library(knitr)
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
kable(out)


control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "CA"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"CA"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "CA"])
```

```
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"CA"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)

control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "NY"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"NY"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "NY"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"NY"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)

control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "FL"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"FL"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "FL"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"FL"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)

control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "TX"])
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"TX"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "TX"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"TX"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)

control_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 1 & abd$Address
== "NV"])
```

```
test_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 1 & abd$Address ==
"NV"])
control_no_buy <- length(abd$outcome[abd$Test_Control == "control" & abd$outcome == 0 &
abd$Address == "NV"])
test_no_buy <- length(abd$outcome[abd$Test_Control == "test" & abd$outcome == 0 & abd$Address ==
"NV"])
out = data.frame(test_buy, test_no_buy, control_buy,
                 control_no_buy)
library(knitr)
kable(out)

abd$treatment <- NULL
abd$treatment <- ifelse(abd $Test_Control=="test",1,0)
summary(abd$treatment)
abd$d_email <- 1*complete.cases(abd$Email)
abd$d_state <- 1*complete.cases(abd$Address)

data_frame<- abd[c(1,16,17,18,19)]

colnames(data_frame) <- c('CustomerID','Outcome','TestVariable','D_Email','D_State')

install.packages("writexl")
library("writexl")
write_xlsx(df,"keep file location")

d1 = lm(abd$outcome~abd$treatment)
summary(d1)

d2 = t.test(abd$outcome~abd$treatment)
summary(d2)
d3 = lm(abd$outcome~abd$treatment*abd$d_state*abd$d_email)
summary(d3)
```