**MIDTERM PROJECT**

**The data results from an advertising campaign of a large travel agency.**

**You have access to two datasets**

**Adandoned.csv (ABD)**

**Reservation.csv (RES)**

**Introduction**

The first contains demographic information about customers who engaged with the travel agency, but that did not buy a vacation package. You may notice the sparseness/missingness of the data and also the presence of potential duplicates.

These customers were randomly assigned to a retargeting campaign; you can see their status in the “Test\_Control” column.

Sometime later, data about reservations were pulled from the system, so have obtained the “Reservation.csv” File. Note that you will just see those who bought a package, whether in the treatment or control group, or neither.

You are generally tasked with matching the data across the files to test whether the retargeting campaign has worked statistically.

In doing so, you will have to make assumptions about reconciling the records (missing info, matching rules, duplicates, etc.)

There is no solution, but the instructor knows the reasonable range of results we should expect.

Delivery: you can use this doc as a template, you are also expected to upload the Rscript that runs all the analyses top to bottom, and the cleaned data (csv format preferred) that you used to answer the questions.

Note: this project is individual based, you can discuss with your peers, but you are expected to disclose all the people you collaborated with.

**I. The Business Problem**

ABD contains data for all the customers in the dataset that were already pursued (advertised) but did not buy a vacation package.

Business Problem: Should we retarget those customers?

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

I would believe that as a Businessman, it is a very sensible question to have to retarget the group who did not buy a vacation package; as there are many possibilities for which the customer might have not chosen to purchase the package. Maybe the reason could be the price of the package, the destination, the trip duration and also based on the season of the trip, existing offers, quality of service etc. However, it is in best practice to be able to consider these customers using certain strategies based on different criteria. For example, one of the criteria of selecting the group of customers and advertising the packages to them is based on their current status, i.e, single or couple, business travel or family gateways etc. It is always a necessity that the agency knows the customer and attends to their needs based on their interest and what works the best for the company. One way to be able to do this is by increasing the feedbacks and the user ratings of the historical trips and also retargeting the rest of the customers with smart advertising and attractive deals.

An experiment is run where customers in the abandoned dataset are randomly placed in a treatment or a control group (see column L in both files).

Those marked as “test” are retargeted (treated); the others marked as control are part of the control group.

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

> table(abd$Test\_Control)

control test

4176 4266

#From the above code, it is determined that both the test and control are having almost equal number of data.

**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.

**> known\_states <- abd[complete.cases(abd['Address']),]**

**> table(known\_states$Test\_Control)**

**control test**

**1855 1957**

#From the above code, it is determined that both the test and control are having almost equal number of data.

**II. Data Matching**

About three months later, the experiment/retargeting campaign is over.

Customers, presented in the ABD excel file, who bought vacation packages during the time frame, are recorded in the RS excel file.

**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)

#Match on key. Returns a logical vector

#First Name and Last Name

> rs$First\_Name[rs$First\_Name==""] <- 0

> rs$Last\_Name[rs$Last\_Name==""] <- 0

> rs$First\_Last\_Name <- paste(rs$First\_Name,rs$Last\_Name,sep = "\_")

> match\_name=abd$First\_Last\_Name %in% rs$First\_Last\_Name

#Match Email

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

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

>

#Match Incoming Phone

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

#Match Contact

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

#Match based on Incoming and contact phone

> match\_incoming\_contact= abd$Incoming\_Phone[complete.cases(abd$Incoming\_Phone)] %in% rs$Contact\_Phone[complete.cases(rs$Contact\_Phone)]

#Match based on contact phone and incoming.

> match\_contact\_incoming= abd$Contact\_Phone[complete.cases(abd$Contact\_Phone)] %in% rs$Incoming\_Phone[complete.cases(rs$Incoming\_Phone)]

>

#Creating the flags for the above match cases:

> #Email

> abd$match\_email <- 0

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

> sum(match\_email)

[1] 75

As per the above output it can be concluded that for the email variable, there are 75 matches for the data in the reservation file, when both the reservation and the abandoned files are taken into consideration.

> #Incoming

> abd$match\_incoming <-0

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

> sum(match\_incoming)

[1] 327

As per the above output it can be concluded that for the email variable, there are 327 matches for the data in the reservation file, when both the reservation and the abandoned files are taken into consideration.

> #Contact

> abd$match\_contact <-0

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

> sum(match\_contact)

[1] 185

As per the above output it can be concluded that for the email variable, there are 185 matches for the data in the reservation file, when both the reservation and the abandoned files are taken into consideration.

> #Incoming Contact

> abd$match\_incoming\_contact <-0

> abd$Incoming\_Phone[complete.cases(abd$Incoming\_Phone)] <- 1\* match\_incoming\_contact

> sum(match\_incoming\_contact)

[1] 129

As per the above output it can be concluded that for the email variable, there are 129 matches for the data in the reservation file, when both the reservation and the abandoned files are taken into consideration.

> #Contact Incoming

> abd$match\_contact\_incoming <-0

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

> sum(match\_contact\_incoming)

[1] 344

As per the above output it can be concluded that for the email variable, there are 344 matches for the data in the reservation file, when both the reservation and the abandoned files are taken into consideration.

> #First Name Last Name

> sum(match\_name)

[1] 996

As per the above output it can be concluded that for the email variable, there are 996 matches for the data in the reservation file, when both the reservation and the abandoned files are taken into consideration.

As per the above output it can be concluded that for the

**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.

#outcome variable

> abd$Outcome <- 0

> abd$Outcome <- 1\*(abd$match\_email | abd$match\_incoming | abd$match\_contact | abd$match\_incoming\_contact | abd$match\_contact\_incoming)

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

> table(abd$Outcome , abd$treat)

0 1

0 4083 93

1 3921 345

From the above it can be concluded that 93 customers from the control group have purchased the package while the rest of the 4083 did not. Same as for the 345 customers from the treatment group has purchased the package while the 3921 from the same group did not.

**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.

**If we are to analyze the columns such as email, incoming phone, it can be noticed that there are a lot of null values and duplicate values. Hence, I had to use the complete.cases() function to eliminate any such null values in the dataset.**

**Q8: Complete the following cross-tabulation:**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **345** | **3921** |
| **Control** | **93** | **4083** |

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

**> state1 <- data.frame(subset(abd,abd$Address == "FL"))**

**> table(state1$Outcome, state1$treat)**

**0 1**

**0 37 34**

1. **0 4**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **4** | **0** |
| **Control** | **34** | **37** |

**> state2 <- data.frame(subset(abd,abd$Address == "KS"))**

**> table(state2$Outcome, state2$treat)**

**0 1**

**0 41 32**

1. **0 5**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **5** | **0** |
| **Control** | **32** | **41** |

**> state3 <- data.frame(subset(abd,abd$Address == "AZ"))**

**> table(state3$Outcome, state3$treat)**

**0 1**

**0 43 51**

1. **1 3**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **3** | **1** |
| **Control** | **51** | **43** |

**> state4 <- data.frame(subset(abd,abd$Address == "UT"))**

**> table(state4$Outcome, state4$treat)**

**0 1**

**0 30 23**

1. **3 4**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **4** | **3** |
| **Control** | **23** | **30** |

**> state5 <- data.frame(subset(abd,abd$Address == "NY"))**

**> table(state5$Outcome, state5$treat)**

**0 1**

**0 35 37**

**1 1 3**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** | **3** | **1** |
| **Control** | **37** | **35** |

**III. Data Cleaning:**

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.)

**new\_df<-subset(abd,abd$Outcome==1)**

**new\_df$D\_Email<- ifelse(new\_df$Email != 'NA',1,0)**

**new\_df["D\_Email"][is.na(new\_df["D\_Email"])] <- 0**

**new\_df$D\_State<- ifelse(new\_df$Address != 'NA',1,0)**

**new\_df["D\_State"][is.na(new\_df["D\_State"])] <- 0**

**new\_file<- data.frame(**

**Customer\_ID=new\_df$Caller\_ID,**

**Test\_Variable =new\_df$Test\_Control,**

**Outcome = new\_df$Outcome,**

**D\_Email = new\_df$D\_Email,**

**D\_State = new\_df$D\_State**

**)**

**> new\_file**

**Customer\_ID Test\_Variable Outcome D\_Email D\_State**

**1 03241649AHZKPWYH test 1 0 1**

**2 85080592TEFIPACV test 1 0 1**

**3 83559451LHCUAFYT test 1 0 0**

**4 18086538MZFGFFTH test 1 0 1**

**5 38297698NQJIEDHS test 1 0 0**

**6 36854393GIZMEDRD test 1 1 1**

**7 05334034DMHRGBJP test 1 0 1**

**8 72535168IUDJYABX test 1 0 0**

**9 32597460SCPZKXYI test 1 0 0**

**10 56895604BZVXIOOY test 1 0 0**

**11 99131886JEWYGEJQ control 1 0 1**

library(writexl)

write\_xlsx(new\_file,"C:/Users/Srinidhi/Downloads/Matched.xlsx")

**IV. Statistical Analysis**

We are finally in a condition to try to answer the relevant business question.

**Q10:** Run a Linear regression model for

Outcome = alpha + beta \* Test\_Variable + error

And Report the output.

> linear\_model<-lm(abd$Outcome~abd$Test\_Control, data = abd)

> summary(linear\_model)

Call:

lm(formula = abd$Outcome ~ abd$Test\_Control, data = abd)

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 \*\*\*

abd$Test\_Controltest 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

As per the above output it can be seen that the p value is less than 0.05 and hence the Null Hypothesis is rejected.

Also the outcome is as below:

Outcome = 0.022270+0.058602\*Test\_variable

It can also be seen that the Multiple R-squared value is 0.01745 which means that the total of 1.74% of variation of the outcome is being explained by the Test\_variable.

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

> anova\_out=aov(abd$Outcome ~ abd$Test\_Control, data = abd)

> summary(anova\_out)

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

abd$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

As per the above output it can be seen that the p value is less than 0.05 and hence the Null Hypothesis is rejected. This signifies that both Anova and Linear regression tests are same. Anova and Linear regression are the same relationship wise but one of the major differences between the Anova and linear regression is that regression is used to predict a continuous outcome based on one or more continuous predictor variables. In case of ANOVA it predicts the outcome based on one or more categorical predictor variables.

**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?

Given the outcome of our analysis for Linear regression, the Multiple R-squared value is 0.017 which is close to zero. Since it falls under the binary distribution we can choose to do our analysis using the **logistic regression** or poison distribution which will provide more effective results.

**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)

Performing multiple linear regression with state and email

> out <- lm(new\_file$Outcome~new\_file$Test\_Variable\*new\_file$D\_State+ new\_file$Test\_Variable\*new\_file$D\_Email)

> summary(out)

Call:

lm(formula = new\_file$Outcome ~ new\_file$Test\_Variable \* new\_file$D\_State +

new\_file$Test\_Variable \* new\_file$D\_Email)

Residuals:

Min 1Q Median 3Q Max

-1.780e-16 -1.780e-16 -9.400e-18 -9.400e-18 2.306e-14

Coefficients:

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

(Intercept) 1.000e+00 1.789e-16 5.589e+15 <2e-16 \*\*\*

new\_file$Test\_Variabletest 9.401e-18 2.006e-16 4.700e-02 0.963

new\_file$D\_State -7.720e-29 2.357e-16 0.000e+00 1.000

new\_file$D\_Email -5.445e-30 3.173e-16 0.000e+00 1.000

new\_file$Test\_Variabletest:new\_file$D\_State 1.686e-16 2.689e-16 6.270e-01 0.531

new\_file$Test\_Variabletest:new\_file$D\_Email -1.588e-16 3.522e-16 -4.510e-01 0.652

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.114e-15 on 432 degrees of freedom

Multiple R-squared: 0.5017, Adjusted R-squared: 0.4959

F-statistic: 86.99 on 5 and 432 DF, p-value: < 2.2e-16

As per the above output it can be seen that the p value is less than 0.05 and hence the Null Hypothesis is rejected.

Performing multiple linear regression with

> multiple\_lm\_out <- lm(new\_file$Outcome~new\_file$Test\_Variable\*new\_file$D\_State)

> summary(multiple\_lm\_out)

Call:

lm(formula = new\_file$Outcome ~ new\_file$Test\_Variable \* new\_file$D\_State)

Residuals:

Min 1Q Median 3Q Max

-1.204e-16 -1.204e-16 0.000e+00 0.000e+00 2.311e-14

Coefficients:

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

(Intercept) 1.000e+00 1.759e-16 5.685e+15 <2e-16 \*\*\*

new\_file$Test\_Variabletest -6.567e-29 1.977e-16 0.000e+00 1.000

new\_file$D\_State -7.703e-29 2.330e-16 0.000e+00 1.000

new\_file$Test\_Variabletest:new\_file$D\_State 1.204e-16 2.624e-16 4.590e-01 0.647

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.112e-15 on 434 degrees of freedom

Multiple R-squared: 0.5026, Adjusted R-squared: 0.4992

F-statistic: 146.2 on 3 and 434 DF, p-value: < 2.2e-16

**Stargazer**

install.packages("stargazer")

library(stargazer)

stargazer(out,multiple\_lm\_out,type = "html",out="midterm.htm")

> stargazer(out,multiple\_lm\_out,type = "html",out="midterm.htm")

<table style="text-align:center"><tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"></td><td colspan="2"><em>Dependent variable:</em></td></tr>

<tr><td></td><td colspan="2" style="border-bottom: 1px solid black"></td></tr>

<tr><td style="text-align:left"></td><td colspan="2">Outcome</td></tr>

<tr><td style="text-align:left"></td><td>(1)</td><td>(2)</td></tr>

<tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Test\_Variabletest</td><td>0.000</td><td>-0.000</td></tr>

<tr><td style="text-align:left"></td><td>(0.000)</td><td>(0.000)</td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td style="text-align:left">D\_State</td><td>-0.000</td><td>-0.000</td></tr>

<tr><td style="text-align:left"></td><td>(0.000)</td><td>(0.000)</td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td style="text-align:left">D\_Email</td><td>-0.000</td><td></td></tr>

<tr><td style="text-align:left"></td><td>(0.000)</td><td></td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td style="text-align:left">D\_State</td><td>0.000</td><td>0.000</td></tr>

<tr><td style="text-align:left"></td><td>(0.000)</td><td>(0.000)</td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td style="text-align:left">D\_Email</td><td>-0.000</td><td></td></tr>

<tr><td style="text-align:left"></td><td>(0.000)</td><td></td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td style="text-align:left">Constant</td><td>1.000<sup>\*\*\*</sup></td><td>1.000<sup>\*\*\*</sup></td></tr>

<tr><td style="text-align:left"></td><td>(0.000)</td><td>(0.000)</td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Observations</td><td>438</td><td>438</td></tr>

<tr><td style="text-align:left">R<sup>2</sup></td><td>0.502</td><td>0.503</td></tr>

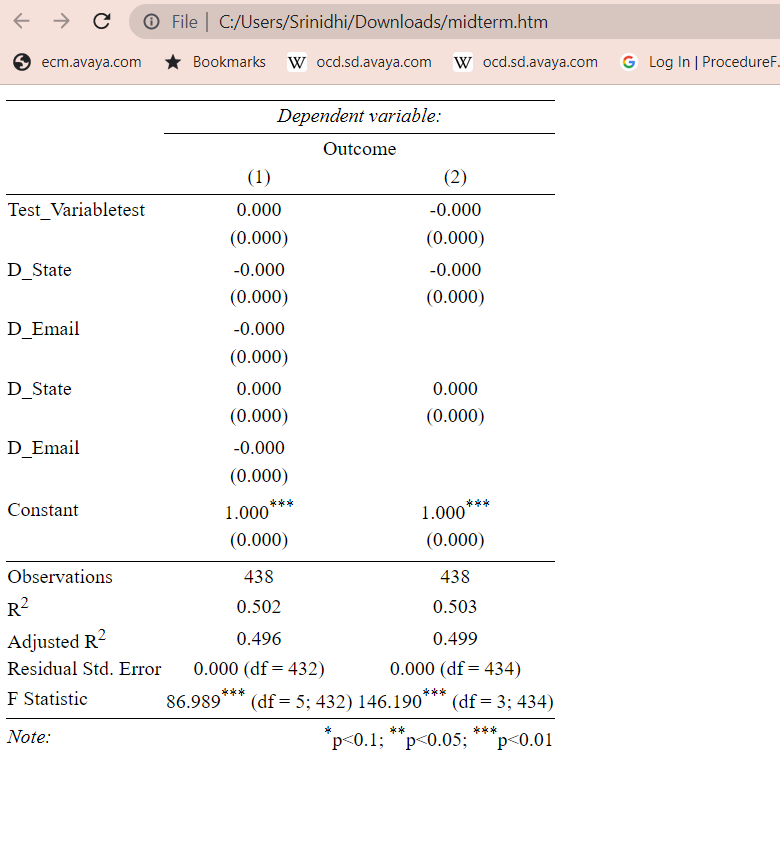
<tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.496</td><td>0.499</td></tr>

<tr><td style="text-align:left">Residual Std. Error</td><td>0.000 (df = 432)</td><td>0.000 (df = 434)</td></tr>

<tr><td style="text-align:left">F Statistic</td><td>86.989<sup>\*\*\*</sup> (df = 5; 432)</td><td>146.190<sup>\*\*\*</sup> (df = 3; 434)</td></tr>

<tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"><em>Note:</em></td><td colspan="2" style="text-align:right"><sup>\*</sup>p<0.1; <sup>\*\*</sup>p<0.05; <sup>\*\*\*</sup>p<0.01</td></tr>

</table>



From the above screencap, the number of stars indicates the significance of a variable.

Also based on the addition of new variables, the value for the Multiple R-squared will increase.

**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**

**Due to this project, I was able to learn how to analyze and fix the missing values by matching different datasets having missing/ duplicates and Nulls effectively. Also, I have learnt in how to successfully to do the predictive analysis of the data using the test/treatment and the control groups. I have also learnt how to accomplish and analyze the cross validation.**

**Since this is a basic dataset/s, this might not be providing the accurate type for the real world scenarios where more unstructured data has to be formalized. In doing so, we might have to use more effecting pre processing procedures to be able to accomplish the desired outcomes. However, it is always a necessity to structure the data and add important nodes or leaves as much as we can to be able to make it as formalized as possible and find the hidden pattens for the algorithms to work efficiently and predict more accurate results.**

**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 would honestly conclude that through this project I was able to stay more focused and I was able to achieve the desired results while I was going through the learning curve. It was as if everything has fallen into place through this project and I was able to learn new information as I was solving the problems. I would like to use other techniques than what were already provided in the class in spare time with different data sets to be able to do more deeper analysis into the subject. However, I would like to leave my expected performance score at 95% for the same reason while my efforts were major and would score myself a 100%. I have definitely collaborated to take suggestions with my peers during this project to gain more insights into the pre processing of the data. I was able to coordinate with Akshay Ramesh and Manikanta during the initial attempts of this project. I was able to gain more insights from Professor Ron Satterfield’s videos posted on canvas to better understand the theory.**