# Data Cleaning outcome and process - MightyHive project

Result obtained after Data Cleaning MightyHive project -

|  | CONTROL | TEST |
| --- | --- | --- |
| NO BUY | 3477 | 3327 |
| BUY | 145 | 408 |

I used R script to clean the data primarily and then Apple numbers for other basic manipulations to the clean data file.

Attached please find the R script -

#Setting the path

setwd("/Users/harshitasrivastava/Desktop/SDM/TeamLeada/MightyHive")

#Loading data into Abandoned and Reserved data seed

ABD <- read.csv("Abandoned\_Data\_Seed.csv", header = TRUE, stringsAsFactors = F)

RES <- read.csv("Reservation\_Data\_Seed.csv", header = TRUE, stringsAsFactors = F)

#Checking the structure of the data frame

str(ABD)

#Total rows in the Abandon data set - 8442

nrow(ABD)

#Total rows in the Abandon data set - 20814

nrow(RES)

#Omitting the rows which are empty from the intial dataset.

#The complete.cases function detects rows in a data.frame that do not contain any missing value

#None of the row is empty in our data set

ABD\_Complete <- complete.cases(ABD)

is.na(ABD\_Complete)

#The na.omit function resulting logical can be used to remove incomplete records from the data.frame.

ABD\_Complete1 <- na.omit(ABD\_Complete)

#The row.names of the removed records are stored in an attribute called na.action.

na.action(ABD\_Complete1)

#Abandoned Data Seed

#Removing the duplicates

#If the Incoming Phone is duplicated in the data set

ABD\_Control\_Phn\_Duplicates = duplicated(ABD$Incoming\_Phone)

#Removing the phone duplicates in the ABD control data file

ABD\_Control\_Phn\_Duplicates\_Removed = ABD[!ABD\_Control\_Phn\_Duplicates,]

#The number of rows in the data set now is reduced to 7166

nrow(ABD\_Control\_Phn\_Duplicates\_Removed)

View(ABD\_Control\_Phn\_Duplicates\_Removed)

#If the Contact Phone is duplicated in the data set

ABD\_Contact\_Phn\_Duplicates = duplicated(ABD\_Control\_Phn\_Duplicates\_Removed$Contact\_Phone)

#The number of rows in the data set is now 7114

ABD\_Contact\_Phn\_Duplicates\_Removed=ABD\_Control\_Phn\_Duplicates\_Removed[!ABD\_Contact\_Phn\_Duplicates,]

View(ABD\_Contact\_Phn\_Duplicates\_Removed)

#Storing the abandoned data after removing the duplicates in CleanData data frame

CleanData = ABD\_Contact\_Phn\_Duplicates\_Removed

CleanData <- CleanData[, c(2,7,9,10,11,12)]

#Matching whether the Incoming Phone from the Abandoned set data matched the same field in the reservation data

CleanData$Y <- ABD\_Contact\_Phn\_Duplicates\_Removed$Incoming\_Phone %in% RES$Incoming\_Phone

nrow(CleanData)

#Exporting data for control group with option of Buy

write.csv(CleanData,file = "/Users/harshitasrivastava/Desktop/SDM/TeamLeada/MightyHive/Clean\_data\_Final.csv")