SPEARS SCHOOL OF BUSINESS

MANAGEMENT SCIENCE & INFORMATION SYSTEMS

MSIS 5223 - PROGRAMMING FOR DATA SCIENCE & ANALYTICS

SUBMITTED BY:

- ISHAN MALPOTRA (A20104861)

BIKE MS (MULTIPLESCLEROSIS) DATA ANALYSIS

SPRING 2018

Table of Contents

General Project Information

This project is a part of the Teradata competition which is organized annually. This is the fourth annual Data Challenge competition in which we are provided with the data sets by a non-profit organization: National Multiple Sclerosis Society (NMSS). We have analyzed the provided data and business questions related to their Bike MS program and presented their findings and results.

Executive Summary

Multiple sclerosis (MS) is an unpredictable, often disabling disease of the central nervous system that disrupts the flow of information within the brain, and between the brain and body. We are working on the dataset for Teradata's Student Competition which mainly focuses on the Bike MS campaign which is the National Multiple Sclerosis Society's largest fundraising campaign, engaging over 70,000 participants to raise $68 million in over 75 rides across the country. It is the largest charity cycling series in the United States. The events are team-focused, with teams responsible for 87% of fundraising. Bike MS participation and revenue have seen a steady decline since the peak in 2012. While retention is relatively high – over 50% – there are not enough new participants joining the series to reverse the damage caused by attrition. We mainly focus on increasing new participant acquisition and finding out factors which can help in improving the campaign and improve awareness for it. More details can be found in the link provided below:

http://www.teradatauniversitynetwork.com/Community/Student-Competitions/2018/Data-Challenge

Project Schedule, Duration and Estimates

GANTT Chart

The whole project was supposed to be completed in 2 and half months. To focus on the Teradata University Competition as well, a team project meeting was kept every week to discuss on the completed tasks, action items for the next week and the work breakdown structure for the whole team for the future. The meeting was usually of around 4 hours duration. Since we started working on this project during our spring break, we did not have any major holidays to deal with.

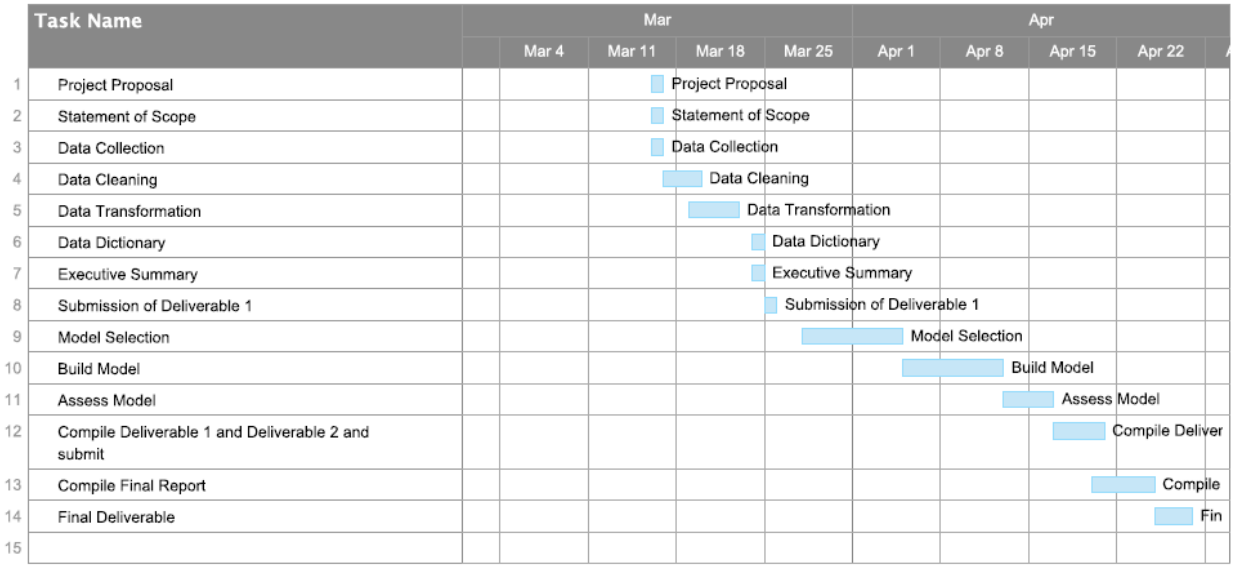

Fig1. GANTT Chart showing the timeline of different project tasks

Tasks Assignment

The table containing the resource assignment is shown below which contains the tasks performed by the whole team.

Lists of Tasks	Duration	Start Date	End Date	Description of Work
Business Problem	1	16-Mar	17-Mar	Identifying the business problem
Data Collection	1	16-Mar	17-Mar	Collecting the data for the analysis from the competition's website
Data Walkthrough	1	16-Mar	17-Mar	Giving a walkthrough of the data obtained for the analysis to the team members
Explanatory Analysis	1	17-Mar	18-Mar	Generating Summary statistics of the data
Data Visualization	1	17-Mar	18-Mar	Creating visual graphs for the current data to understand about it
Data Consolidation	2	17-Mar	19-Mar	Combining all the datasets required for the analysis into a common one
Data Access & Cleaning	3	19-Mar	22-Mar	Cleaning the data after accessing it by filling in the missing values and removing some insignificant values
Data Transformation	2	23-Mar	24-Mar	Transforming the variables by creating some new variables and converting continuous to categorical variables
Data Reduction	1	23-Mar	24-Mar	Reducing the data to reduce its size and make it less bulkier
Data Dictionary	1	24-Mar	25-Mar	Creating data dictionary which contains the information about the variables with their abbreviation
Identify Models	13	25-Mar	07-Apr	Understood the different models to be built
Create Models	8	08-Apr	16-Apr	Built the identified models
Validating the Model	5	17-Apr	22-Apr	Validated the results of the models built
Reporting the results	4	23-Apr	27-Apr	Reported the results to validate the models built
Model Selection	4	28-Apr	02-May	Selected the best model by comparing all the models built
Documentation	2	03-May	05-May	Preparing the report for deliverable 2

Statement of Scope

The purpose of this project is to analyze on increasing new participant acquisition and finding out other factors which would help in improving the Bike MS campaign and improve awareness for Multiple sclerosis. Below are the questions that we are analyzing:

- What industries have had the strongest involvement in Bike MS in the last five years?
- What occupations were responsible for most of our fundraising?

The target variables used for both the business questions are "NoofParticipant" for business question 1 and "ParticipantOccupation" for the business question 2. The predictor variables include all the significant variables found in both PCA and FA apart from the dummy variables created in data transformation.

Data Preparation

Data Access

We have obtained the data from the Teradata University Competition website mentioned in the above sections. All the data consists details about the Bike MS competition from 2013 to 2017. There were 8 data sets, but we chose the only relevant 2 datasets, viz. Participants and Bike Teams. These two datasets contain all the relevant columns which is required to analyze the questions that has been mentioned. The first dataset, "Participants" contains all the information about the participants who participated in the Bike MS competition. It has 13,121 rows and 26 variables. The second dataset, Bike Teams contains all the information about the teams participated in the Bike MS competition from 2013 to 2017, like number of participants in a team,

total donations accumulated by a particular team within those years. It has 29,937 rows and 15 variables.

Below is the code for accessing the dataset:

```
library("readxl")
#load datasets
participants <- read_excel("participantsV2.xlsx")
biketeam <- read_excel("biketeamsV2.xlsx")

#check number of rows
nrow(participants)
nrow(biketeam)

#check colnames
colnames(participants)
colnames(biketeam)
```

Below is the screenshot of the output:

```
> #check colnames
> colnames(participants)
 [1] "SecurityCategoryName"          "Fiscal Year"
 [3] "InternalEventName"             "EventDate"
 [5] "ParticipationTypeName"         "TeamName"
 [7] "TeamCreationDate"              "TeamDivision"
 [9] "TeamID"                        "ContactID"
[11] "MemberID"                      "RegistrationDate"
[13] "IsPriorParticipant"            "TotalofAllConfirmedGifts"
[15] "TotalFromParticipant"          "TotalNotFromParticipant"
[17] "NumberFromParticipant"         "NumberNotFromParticipant"
[19] "ParticipantEmployer"           "ParticipantOccupation"
[21] "ParticipantConnectiontoMS"     "AddressParticipantStateProvince"
[23] "AddressParticipantCounty"      "AddressParticipantCity"
[25] "EventID"                       "ParticipantGender"
> colnames(biketeam)
 [1] "InternalEventName"              "EventID"
 [3] "TeamID"                         "TeamName"
 [5] "TeamCreationDate"               "TeamDivision"
 [7] "Company"                        "NumberofParticipants"
 [9] "TotalFeesPaid"                  "TeamTotalConfirmed"
[11] "TotalOnlineGifts"               "TotalOfflineConfirmedGifts"
[13] "TotalOfflineUnconfirmedGifts"   "TeamGoal"
[15] "TotalConfirmedGiftsinTeamHistory"
```

Fig 2. Displaying the names of various columns in both the datasets

```
> nrow(participants)
[1] 13212
> nrow(biketeam)
[1] 29937
```

Fig 3. Displaying the number of rows for both the datasets

From the above output, we can see that figure 2 shows the column names in both the datasets to be merged and the command “nrow” gives the number of rows present in both the datasets as shown in the figure 3.

Data Cleaning

The data cleaning has been performed for both the chosen datasets on a separate basis on the start. Both the datasets, i.e., Participants and Bike Teams were chosen separately and cleaned partially using Microsoft excel and partially using the analytical language R.

For the Participants dataset:

We removed all the blank team names, team creation date, team division, member ID, Participant Email Status, Participant Employer (We removed the employer name which was a number), Participant Occupation, Participant Connection to MS, Address – Participant State/Province, Address - Participant County. We also changed the "blank" to "other" for participant gender.

For the Bike Team dataset:

We removed all blank captain email donation, Team Division, Previous Event Team Members. We also removed "friends & family" from team captain accept email column as it was irrelevant.

Below mentioned is the code used for cleaning the Null values from the bike teams dataset:

```
#Setting the working directory
workingdirectory= "C:\\Users\\imalpot\\Desktop\\"
setwd(workingdirectory)
#read the target file
datafile_biketeams=read.csv(file.choose(),header = TRUE)
```

```
#Omit the null values
refined_datafile <- na.omit(datafile_biketeams)
#Check the column names of the datafile
str(refined_datafile)
#Delete the unnecessary coloumns not required for analysis
refined_datafile2 = subset(refined_datafile, select = -
c(CaptainEmailDomain,TeamCaptainAcceptEmail) )
#Write the datafile into a new csv file
write.table(refined_datafile2,"New_Bike_Teams.csv",sep=",", row.names =FALSE)
```

Data Consolidation

We consolidated the Participants and the Bike Team dataset and merged it into a single dataset

"total", using the common column, "Team ID". Below is the code for the same:

```
#merging the datasets

total <- merge(participants,biketeam ,by="TeamID")

#checking the column names for the merged dataset

colnames(total)

nrow(total)
```

Figure 4 given below depicts the screenshot for the same:

```
> #merging the datasets
> total <- merge(participants,biketeam ,by="TeamID")
>
> #checking the column names for the merged dataset
> colnames(total)
 [1] "TeamID"                          "SecurityCategoryName"
 [3] "Fiscal Year"                     "InternalEventName.x"
 [5] "EventDate"                       "ParticipationTypeName"
 [7] "TeamName.x"                      "TeamCreationDate.x"
 [9] "TeamDivision.x"                  "ContactID"
[11] "MemberID"                        "RegistrationDate"
[13] "IsPriorParticipant"              "TotalofAllConfirmedGifts"
[15] "TotalFromParticipant"            "TotalNotFromParticipant"
[17] "NumberFromParticipant"           "NumberNotFromParticipant"
[19] "ParticipantEmployer"             "ParticipantOccupation"
[21] "ParticipantConnectiontoMS"       "AddressParticipantStateProvince"
[23] "AddressParticipantCounty"        "AddressParticipantCity"
[25] "EventID.x"                       "ParticipantGender"
[27] "InternalEventName.y"             "EventID.y"
[29] "TeamName.y"                      "TeamCreationDate.y"
[31] "TeamDivision.y"                  "Company"
[33] "NumberofParticipants"            "TotalFeesPaid"
[35] "TeamTotalConfirmed"              "TotalOnlineGifts"
[37] "TotalOfflineConfirmedGifts"      "TotalOfflineUnconfirmedGifts"
[39] "TeamGoal"                        "TotalConfirmedGiftsinTeamHistory"
>
```

Fig 4. Data Consolidation shown by merging the different datasets

Data Transformation

We created the dummy variables for the relevant categorical variables to use the same in our model. We created dummy variables for the variables TeamDivision, IsPriorParticipant, ParticipantConnectiontoMS, ParticipantGender. Below is the code for the same:

```
#Creating dummy variables (data transformation)
library(dummies)
bike_TeamDivision=dummy(total[,c('TeamDivision.x')], sep='_')
colnames(bike_TeamDivision)
colnames(bike_TeamDivision)=c('TeamDivision_1','TeamDivision_2','TeamDivision_3','TeamDivi
sion_4','TeamDivision_5','TeamDivison_6', 'TeamDivison_7')
bike_TeamDivision=as.data.frame(bike_TeamDivision)
total=data.frame(total,bike_TeamDivision)



bike_ispriorparticipant=dummy(total[,c('IsPriorParticipant')], sep ='_')
```

```
colnames(bike_ispriorparticipant)
colnames(bike_ispriorparticipant)=c('bike_ispriorparticipant_1','bike_ispriorparticipant_2')
bike_ispriorparticipant=as.data.frame(bike_ispriorparticipant)
total=data.frame(total,bike_ispriorparticipant)



bike_participantconnectiontoms=dummy(total[,c('ParticipantConnectiontoMS')], sep='_')

colnames(bike_participantconnectiontoms)
colnames(bike_participantconnectiontoms)=c('bike_participantconnectiontoms_1','bike_partici
pantconnectiontoms_2','bike_participantconnectiontoms_3','bike_participantconnectiontoms_4
','bike_participantconnectiontoms_5','bike_participantconnectiontoms_6','bike_participantconn
ectiontoms_7','bike_participantconnectiontoms_8','bike_participantconnectiontoms_9',

'bike_participantconnectiontoms_10','bike_participantconnectiontoms_11','bike_participantcon
nectiontoms_12', 'bike_participantconnectiontoms_13', 'bike_participantconnectiontoms_14')
bike_participantconnectiontoms=as.data.frame(bike_participantconnectiontoms)
total=data.frame(total,bike_participantconnectiontoms)



bike_gender=dummy(total[,c('ParticipantGender')], sep='_')

colnames(bike_gender)
colnames(bike_gender)=c('bike_gender_1','bike_gender_2','bike_gender_3')
bike_gender=as.data.frame(bike_gender)
total=data.frame(total,bike_gender)
```

Given below are some of the screenshots displaying the transformation of the variables

```
> head(bike_TeamDivision)
  TeamDivision_1 TeamDivision_2 TeamDivision_3 TeamDivision_4 TeamDivision_5
1              0              1              0              0              0
2              0              1              0              0              0
3              0              1              0              0              0
4              0              1              0              0              0
5              0              1              0              0              0
6              0              0              0              1              0
  TeamDivison_6 TeamDivison_7
1             0             0
2             0             0
3             0             0
4             0             0
5             0             0
6             0             0
> |
```

Fig 5. Transformed variable "bike_TeamDivision"

```
> head(bike_ispriorparticipant)
  bike_ispriorparticipant_1 bike_ispriorparticipant_2
1                         0                         1
2                         0                         1
3                         0                         1
4                         0                         1
5                         0                         1
6                         0                         1
>
```

Fig 6. Transformed variable "bike_ispriorparticipant"

```
> head(bike_participantconnectiontoms)
  bike_participantconnectiontoms_1 bike_participantconnectiontoms_2 bike_participantconnectiontoms_3 bike_participantconnectiontoms_4 bike_participantconnectiontoms_5
1                                0                                0                                0                                0                                0
2                                0                                0                                0                                0                                0
3                                0                                0                                0                                0                                0
4                                0                                0                                0                                0                                0
5                                0                                0                                0                                1                                0
6                                0                                0                                0                                0                                0
  bike_participantconnectiontoms_6 bike_participantconnectiontoms_7 bike_participantconnectiontoms_8 bike_participantconnectiontoms_9
1                                0                                0                                0                                0
2                                0                                0                                0                                1
3                                0                                0                                1                                0
4                                0                                0                                0                                0
5                                0                                0                                0                                0
6                                0                                0                                0                                0
  bike_participantconnectiontoms_10 bike_participantconnectiontoms_11 bike_participantconnectiontoms_12 bike_participantconnectiontoms_13
1                                 0                                 0                                 0                                 0
2                                 0                                 0                                 0                                 0
3                                 0                                 0                                 0                                 0
4                                 0                                 0                                 0                                 0
5                                 0                                 0                                 0                                 0
6                                 0                                 0                                 0                                 0
  bike_participantconnectiontoms_14
1                                 1
2                                 0
3                                 0
4                                 1
5                                 0
6                                 1
>
```

Fig 7. Transformed variable "bike_participantconnectiontoms"

```
> head(bike_gender)
  bike_gender_1 bike_gender_2 bike_gender_3
1             0             1             0
2             0             1             0
3             1             0             0
4             0             1             0
5             1             0             0
6             1             0             0
>
```

Fig 8. Transformed variable "bike_gender"

Data Reduction using PCA and FA

"Principal component analysis (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components". PCA or Principal Component Analysis helps in analyzing the columns which are relatively less important using the Eigen values.

We have done PCA on the variables "TotalofAllConfirmedGifts","TotalFromParticipant", "TotalNotFromParticipant","NumberFromParticipant","NumberNotFromParticipant","TotalFees Paid","TeamTotalConfirmed","TotalOnlineGifts","TotalOfflineConfirmedGifts","TotalOfflineUnc onfirmedGifts","TeamDivision_1","TeamDivision_2","TeamDivision_3","TeamDivision_4","Tea mDivision_5","TeamDivison_6","TeamDivison_7","bike_ispriorparticipant_1","bike_ispriorparti cipant_2", "bike_gender_1", "bike_gender_2", and "bike_gender_3".

Below is the code for the same:

```
#Data reduction using PCA

reduction_data.pca2 = total[c("TotalofAllConfirmedGifts"  ,"TotalFromParticipant"
,"TotalNotFromParticipant"  ,"NumberFromParticipant"   ,"NumberNotFromParticipant"
,"TotalFeesPaid","TeamTotalConfirmed",  "TotalOnlineGifts",  "TotalOfflineConfirmedGifts",
"TotalOfflineUnconfirmedGifts","TeamDivision_1","TeamDivision_2","TeamDivision_3","TeamDi
vision_4","TeamDivision_5","TeamDivison_6","TeamDivison_7","bike_ispriorparticipant_1","bik
e_ispriorparticipant_2","bike_gender_1","bike_gender_2","bike_gender_3" )]

pcamodel_reduc2 = princomp(reduction_data.pca2,cor=TRUE)


#checking Eigen Values

pcamodel_reduc2$sdev^2
```

```
> #checking Eigen Values
> pcamodel_reduc2$sdev^2
      Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6
4.295429e+00 2.549684e+00 2.090732e+00 1.898736e+00 1.603326e+00 1.226555e+00
      Comp.7       Comp.8       Comp.9      Comp.10      Comp.11      Comp.12
1.128056e+00 1.017534e+00 1.009885e+00 1.004904e+00 1.000080e+00 9.976643e-01
     Comp.13      Comp.14      Comp.15      Comp.16      Comp.17      Comp.18
7.451926e-01 6.956573e-01 4.799393e-01 1.704621e-01 8.357836e-02 2.583737e-03
     Comp.19      Comp.20      Comp.21      Comp.22
2.661447e-13 1.311166e-13 5.587852e-15 0.000000e+00
```

Fig 9. PCA modelling showing the standard deviation of the component variables

```
#plotting the graph
plot(pcamodel_reduc2,main="Screeplot")
```

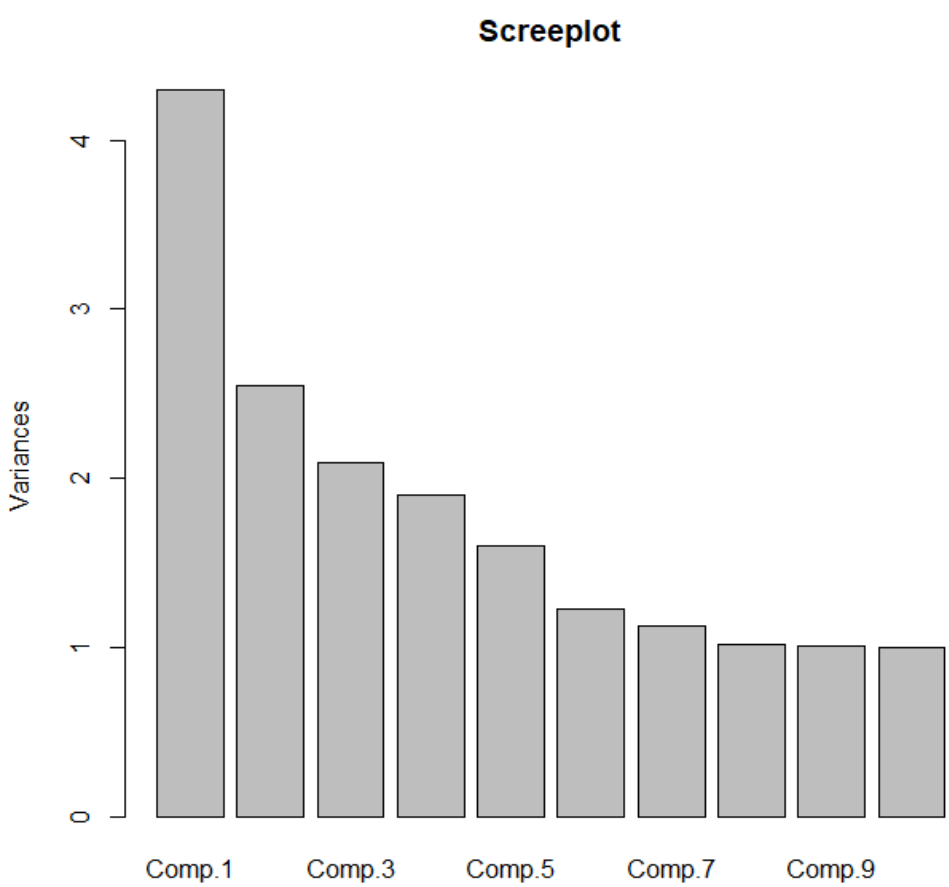

Fig 10. Screeplot for the variance values of the different components

We can see from the figure 6 that the total number of variable are not shown in this graph. It is only showing 10 components out of 22. Hence, we shall use “screeplot” function which will do the work.

Below is the code for the same:

screeplot(pcamodel_reduc2, npcs = 22, type = "lines")

The output for the same is shown below in figure 11:

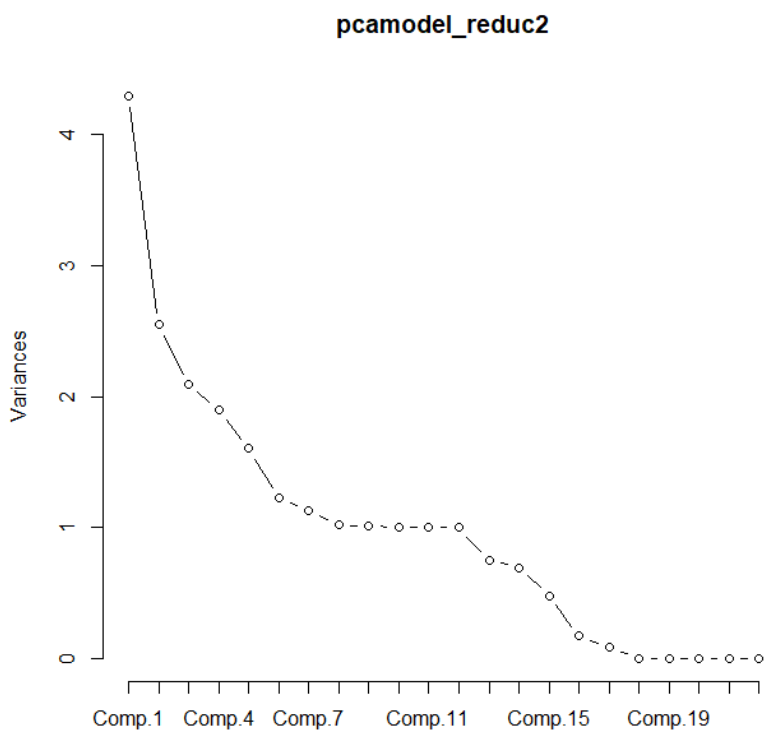

Fig 11. PCA modelling for all the components

We can see from the figure 11 that the graph is falling continuously and abruptly after the 12th component. So we shall now perform the Factor Analysis in order to identify which are the

important components out of the 22 variables. We have used the “psych” and “GPArotation” library for the same.

Below is the code for factor analysis with factors = 12:

```
##confirm results of PCA
##FA
temp <- total[,c(14,15,16,17,18,36,37,38,41,42,43,44,45,46,47,48,49,64,65,66)]
colnames(temp)
library(psych)
library(GPArotation)
fa(r=cor(temp), nfactors=12, rotate="varimax", SMC=FALSE, fm="minres")
```

Below is the output for the same:

```
Factor Analysis using method =  minres
Call: fa(r = cor(temp), nfactors = 12, rotate = "varimax", SMC = FALSE,
    fm = "minres")
Standardized loadings (pattern matrix) based upon correlation matrix
                              MR1   MR5   MR4   MR3   MR2   MR6   MR7   MR8
TotalofAllConfirmedGifts      0.97  0.06  0.03  0.01  0.01 -0.01  0.00 -0.01
TotalFromParticipant          0.14  0.07  0.04  0.02 -0.02  0.00  0.00 -0.01
TotalNotFromParticipant       1.00  0.06  0.03  0.01  0.01 -0.01  0.00 -0.01
NumberFromParticipant         0.07  0.03  0.00  0.01  0.01 -0.01  0.00  0.00
NumberNotFromParticipant      0.56  0.12  0.08  0.01 -0.03  0.02  0.00  0.00
TotalOnlineGifts              0.10  1.00  0.03  0.00  0.03 -0.03 -0.01 -0.01
TotalOfflineConfirmedGifts    0.13  0.89  0.02  0.01  0.05 -0.03 -0.01 -0.01
TotalOfflineUnconfirmedGifts  0.03  0.45  0.01 -0.01  0.04  0.02 -0.01 -0.01
TeamDivision_1               -0.02  0.10 -0.04  0.05  0.94 -0.26 -0.10 -0.09
TeamDivision_2                0.00 -0.03  0.00  0.01  0.00 -0.01  1.00 -0.01
TeamDivision_3                0.01 -0.01  0.03 -0.02 -0.01  1.00 -0.02 -0.01
TeamDivision_4                0.01 -0.08  0.03 -0.04 -0.95 -0.24 -0.10 -0.09
TeamDivision_5               -0.02 -0.02  0.01  0.00  0.00 -0.01  0.00  0.00
TeamDivison_6                -0.01 -0.04 -0.01 -0.01  0.00 -0.01 -0.01  1.00
TeamDivison_7                 0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.00
bike_ispriorparticipant_1    -0.08 -0.03 -0.99 -0.02  0.03 -0.02  0.00  0.01
bike_ispriorparticipant_2     0.08  0.03  0.99  0.02 -0.03  0.02  0.00 -0.01
bike_gender_1                -0.01  0.01 -0.02 -1.00 -0.04  0.01 -0.01  0.00
bike_gender_2                 0.01 -0.01  0.02  1.00  0.04 -0.01  0.01  0.00
bike_gender_3                -0.01  0.02 -0.04 -0.01  0.01 -0.01  0.00  0.00
                              MR12   MR9  MR11  MR10    h2      u2 com
TotalofAllConfirmedGifts      0.20  0.00  0.00  0.00 1.00  0.0050 1.1
TotalFromParticipant          0.90  0.00 -0.01  0.00 0.84  0.1612 1.1
TotalNotFromParticipant       0.04  0.00  0.00  0.00 1.00  0.0011 1.0
NumberFromParticipant         0.38 -0.01  0.01  0.00 0.15  0.8515 1.1
NumberNotFromParticipant      0.08 -0.01  0.00  0.00 0.35  0.6525 1.2
TotalOnlineGifts              0.04 -0.01 -0.02  0.00 1.01 -0.0052 1.0
TotalOfflineConfirmedGifts    0.05  0.00 -0.02  0.00 0.82  0.1831 1.1
TotalOfflineUnconfirmedGifts  0.04 -0.01  0.03  0.00 0.21  0.7879 1.1
TeamDivision_1                0.00 -0.07  0.01 -0.01 1.00  0.0028 1.2
TeamDivision_2                0.00  0.00  0.00  0.00 1.00  0.0049 1.0
TeamDivision_3               -0.01 -0.01 -0.01  0.00 1.00  0.0044 1.0
TeamDivision_4                0.01 -0.06 -0.01 -0.01 1.00  0.0029 1.2
TeamDivision_5               -0.01  1.00  0.00  0.00 1.00  0.0050 1.0
TeamDivison_6                 0.00  0.00  0.00  0.00 1.00  0.0049 1.0
TeamDivison_7                -0.01  0.00  0.00  1.00 1.00  0.0050 1.0
bike_ispriorparticipant_1    -0.02 -0.01  0.02  0.00 1.00  0.0025 1.0
bike_ispriorparticipant_2     0.02  0.01 -0.02  0.00 1.00  0.0025 1.0
bike_gender_1                -0.02  0.00 -0.05  0.00 1.00  0.0025 1.0
bike_gender_2                 0.02  0.00 -0.06  0.00 1.00  0.0025 1.0
bike_gender_3                 0.01  0.00  1.00  0.00 1.00  0.0050 1.0
```

Fig 12a. Factor Analysis Results 1

```
                       MR1  MR5  MR4  MR3  MR2  MR6  MR7  MR8 MR12  MR9 MR11 MR10
SS loadings           2.32 2.04 1.99 1.99 1.81 1.12 1.02 1.01 1.01 1.00 1.00 1.00
Proportion Var        0.12 0.10 0.10 0.10 0.09 0.06 0.05 0.05 0.05 0.05 0.05 0.05
Cumulative Var        0.12 0.22 0.32 0.42 0.51 0.56 0.61 0.67 0.72 0.77 0.82 0.87
Proportion Explained  0.13 0.12 0.12 0.11 0.10 0.06 0.06 0.06 0.06 0.06 0.06 0.06
Cumulative Proportion 0.13 0.25 0.37 0.48 0.59 0.65 0.71 0.77 0.83 0.88 0.94 1.00

Mean item complexity =  1.1
Test of the hypothesis that 12 factors are sufficient.

The degrees of freedom for the null model are  190  and the objective function was  74.99
The degrees of freedom for the model are 16  and the objective function was  48.6

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.02

Fit based upon off diagonal values = 1Warning messages:
1: In cor.smooth(R) : Matrix was not positive definite, smoothing was done
2: In cor.smooth(R) : Matrix was not positive definite, smoothing was done
3: In cor.smooth(r) : Matrix was not positive definite, smoothing was done
4: In fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate,  :
  An ultra-Heywood case was detected.  Examine the results carefully
5: In cor.smooth(r) : Matrix was not positive definite, smoothing was done
```

Fig 12b. Factor Analysis Results 2

As seen in the above figure 12a and 12b of the FA analysis, only 12 out of 22 variables are required. The 12 significant variables are mentioned below:

- Factor1: TotalofAllConfirmedGifts
- Factor2: TotalOnlineGifts
- Factor3: bike_ispriorparticipant_1
- Factor4: bike_gender_2
- Factor5: TeamDivision_1
- Factor6: TeamDivision_3
- Factor7: TeamDivision_2
- Factor8: TeamDivision_6
- Factor9: TotalFromParticipant

- Factor10: TeamDivision_5
- Factor11: bike_gender_3
- Factor12: TeamDivison_7

Hence, We shall remove the rest columns from the dataset. The columns to be removed from the dataset are as given below:

- TotalNotFromParticipant
- NumberFromParticipant
- NumberNotFromParticipant
- TotalOfflineConfirmedGifts
- TotalOfflineUnconfirmedGifts
- TeamDivision_4
- bike_ispriorparticipant_2
- bike_gender_1

The code for reducing the non-necessary columns from the dataset is:

```
reduced_total <- total[,c(-16,-17,-18,-37,-38, -44, -49, -64)]
```

Data Dictionary

Below table shows all the variables with their description and data type:

Variable Name	Description	DataType
Participation Type Name	3 registration options: cyclist, virtual cyclist, or volunteer	Factor
Team Division	Team type: corporate, friends & family, other	Factor
Contact ID	Unique participant contact ID	Integer
Member ID	Unique participant data warehouse ID	Integer
Participant Accept Email	Acceptance of email from the participants	String
Registration Date	Date of the registration	DateTime
Registration Active Status	The current status of the registration	Factor
Is Team Captain	If the participant is the captain of the team	Factor
Is Secondary Registration	TRUE = Someone else registered this person after their own registration; FALSE = this is a primary registration	Factor
Is Prior Participant	YES = participated in previous iteration of this event; N/A = did not	Factor
Emails Sent	Number of emails sent via online participant center	Integer
Total of All Confirmed Gifts($)	All donations for this participant	Integer
Total From Participant($)	Total donations received BY this participant (the participant is same as the donor)	Integer
Total Not From Participant($)	Total donations ON BEHALF OF this participant (donations made by people other than the donor)	Integer
Number From Participant	Number of donations BY this participant	Integer
Number Not From Participant	Number of donations ON BEHALF OF this participant	Integer
Participant Email Status	Email status of the participant in a team	Factor
Participant Employer	Employer of the participant	String
Participant Occupation	Occupation of the participant	String
Participant Connection to MS	How is the participant related to the disease	Factor
Address - Participant State/Province	State from which the participant belong to	String
Address - Participant County	Country from which the participant belong to	String
Address - Participant City	City from which the participant belong to	String
Address - Participant ZIP/Postal Code	Zip code from where the participant belong to	String

Registration Type	Typically 3 registration options: cyclist, virtual cyclist, or volunteer	Factor
Participant Gender	Gender of the participant	Factor
Participant Goal($)	Fundraising goal for the participant	Integer
Suggested Participant Goal($)	The system-generated suggested goal (can be modified by the user)	Integer
Event Type	Campaign (Bike, Walk, MuckFest, etc.)	Factor
Internal Event Name	Name of the event	String
Event ID	Unique event identifier	Integer
Team ID	Unique team identifier	Integer
Team Creation Date	Date on which the team was created	DateTime
Team Captain Contact ID	Unique contact ID of the team captain	Integer
Captain Email Domain	Domain of the captain's email	Factor
Team Captain Accept Email	Whether the captain has accepted the invitation for the event	Factor
Team Division	Whether the team is corporate, friends & family or other	Factor
Company	Company affiliated with the team	String
Number of Participants	Total number of participants in a team	Integer
Total Fees Paid	Registration fees paid by all the participants of a team	Integer
Team Total Confirmed ($)	Total amount of donations for all team members and team gifts	Integer
Total Online Gifts($)	Total amount of team's donations received online	Integer
Total Offline Confirmed Gifts($)	Total amount of team's donations received offline (cash/check)	Integer
Total Offline Unconfirmed Gifts($)	Total amount of team's donations received offline but never received by NMSS	Integer
Team Goal($)	Goal for the team for fundraising	Integer
Total Confirmed Gifts in Team History($)	Total donations received from this team	Integer
Previous Event Fiscal Year	The most recent year that the team participated	Factor
Previous Event Internal Name	Name of the event	String
Previous Event Team Name	Name of the team	String
Previous Event Confirmed Gifts($)	Team's fundraising	Integer
Previous Event Team Members	Total number of members in a team	Integer
Event Date	Event date of the current team	DateTime

Descriptive Statistics

The descriptive statistics to study a few significant variables in the dataset are shown below:

1. “TeamTotalConfirmed” variable

```
> describe(data3$TeamTotalConfirmed)
data3$TeamTotalConfirmed
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
   12376        0     2253        1    51556    70929     1085     2200     6036
     .50      .75      .90      .95
   17009    48507   132356   285148

lowest :      0.0      5.0     25.0     30.0     35.0
highest: 345520.3 377688.5 383232.7 390189.5 396236.2
```

Fig 13. Description of the variable “TeamTotalConfirmed”

As seen in the figure 13, there are no missing values in this data. Mean of the total donation earned by the teams is $ 51,556.

2. “NumberofParticipants” variable

```
> describe(data3$NumberofParticipants)
data3$NumberofParticipants
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
   12376        0      113        1    42.76    49.45        3        4        9
     .50      .75      .90      .95
      22       49      128      171

lowest :   1   2   3   4   5, highest: 193 205 206 215 235
```

Fig 14. Description of the variable “NumberofParticipants”

As seen in the figure 14, there are no missing values in this data. Mean of the total number of participants is 42.76.

3. Distribution of Gender:

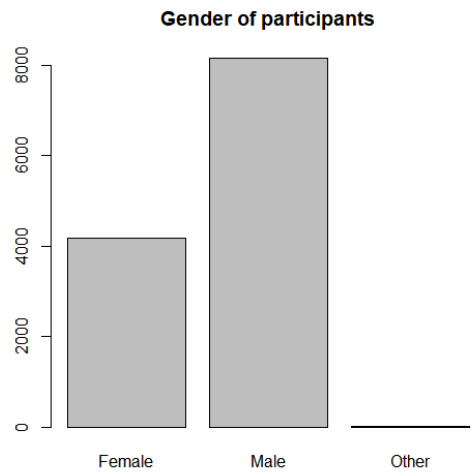

Fig. 15. Gender Distribution

We can see from the above figure 15 that number of males are approx. double than the number of females participating in the competition. There was few unmentioned gender, which is denoted by "other".

4. Distribution of Prior Participant:

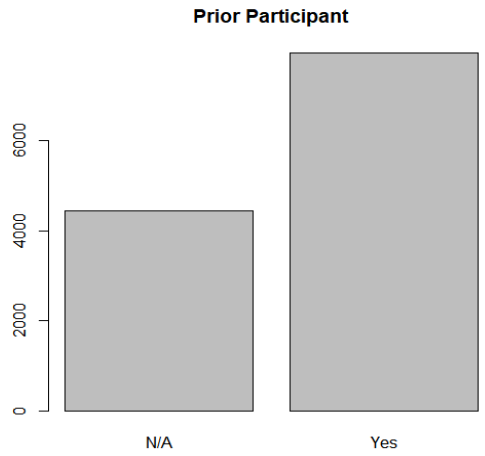

Fig. 16. Distribution of the variable Prior Participant

As seen in the figure 16, the number of participants who already participated in the event is far more than the new participants.

5. Distribution of Participant connection to MS:

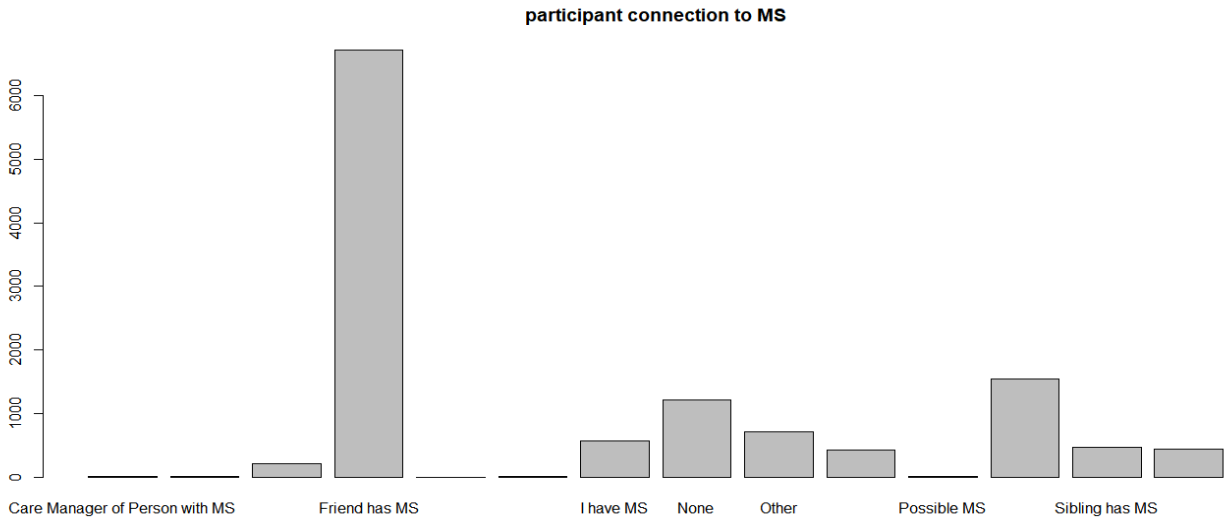

Fig 17. Distribution of the variable "ParticipantconnectiontoMS"

As shown in the figure 17, most of the participants has a friend who is suffering from this disease.

6. Distribution of Participant occupation:

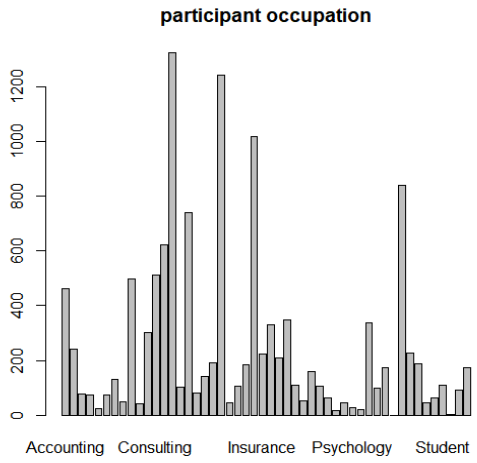

Fig 18. Distribution of the variable "ParticipantOccupation"

As seen in the figure 18, most of the participants are from consulting and insurance background.

7. Distribution of participants per year:

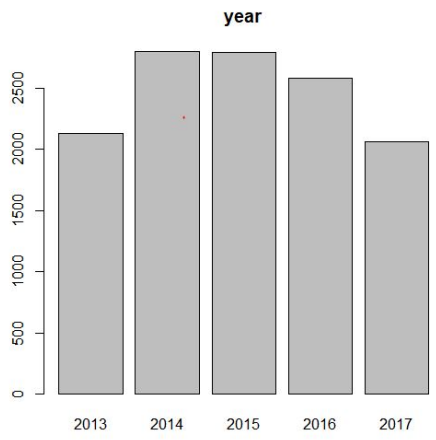

Fig 19. Distribution of the participants with year

We can see from the figure 19 that participation was on peak for 2014 and 2015 and it has a decreasing trend after that.

Data Modeling

We are analyzing the industries which have the strongest involvement in the Bike MS in the last five years, and the occupations which were responsible for most of the fundraising. In the second business question, we are analyzing about the occupation which were responsible for most of the fundraising.

Two modelling techniques used are:

- Linear Regression
- Decision Tree

Linear Regression

For the first business question, the strength of the involvement of industries can be measured by the number of participants participated from each industry. In this case, number of participants is a continuous variable, so we shall use the linear regression model in order to predict strongest involvement of the industries. The target variable in this case would be “NumberofParticipants” and one of the predictor variables will be “TeamDivision”, as this variable contains the information about the various industries such as Corporate.

For the second business question, the target variable will be “TeamTotalConfirmed”, which is again a continuous variable. It has the information about the total donations gathered by teams from 2013 to 2017. One of the predictor variables will be occupation, which contains the information about the occupation of all the participants. We have done linear regression, as it helps us to determine the relation between the predictor and the target variable. We shall know about the various occupations which have generated the maximum number of fundraising using this model.

Decision Tree

Decision tree is another modelling technique which we have used in our analysis. We have used both the regression and classification trees in our analysis since in the first business question , we have a continuous target variable for which the regression tree would fit the best.

For the second business question, we have used classification tree since the target variable is a categorical variable.

In the decision tree, the top nodes are the best predictor as they are highly correlated with the target variable. Hence, we would take into account all the nodes of our decision tree but our main focus would be on the top most node.

Assumptions

For both the modelling techniques we have used, there are several assumptions which are to be considered before proceeding with the results of the same. The assumptions of both the linear regression and the decision tree are mentioned below.

Linear Regression

Below are the assumptions that we have taken for linear regression modeling:

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

Below is the model for which we have checked the assumptions:

Fit 3 Assumptions:
fit3 = lm(log(NumberofParticipants) ~ ParticipantOccupation + TeamDivision_1 +TeamDivision_2 + TeamDivision_3 + TeamDivision_5 + sqrt(TeamTotalConfirmed), data = data3.train)

We have taken log of NumberofParticipants and square root of TeamTotalConfirmed in order to fulfill all the assumptions:

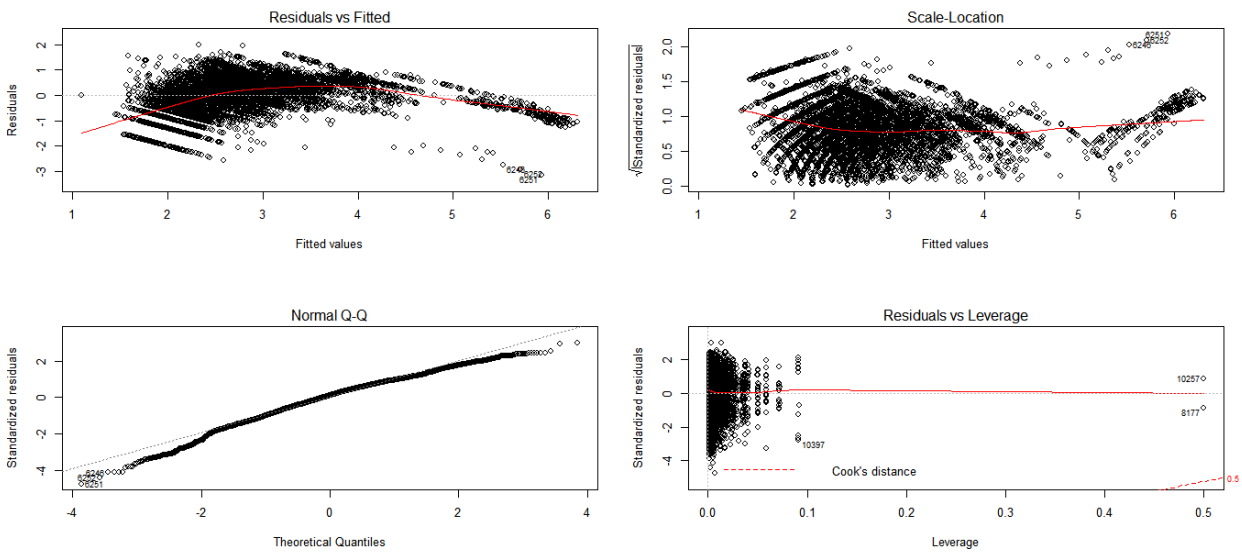

Fig 20. Conditions of Linearity, Normality and Homoscedasticity for Fit3 model

From the above figure 20, we can see that our model is fulfilling the linearity, normality and homoscedasticity. For checking auto correlation, we have done Durbin Watson test, and for checking multi collinearity, we have performed the VIF test.

```
> durbinWatsonTest(fit3)
 lag Autocorrelation D-W Statistic p-value
   1       0.6836516     0.6320152       0
 Alternative hypothesis: rho != 0
```

Fig 21. Durbin Watson Statistic for Fit3 model

Since the statistic value is greater than 0.05 as shown in the figure 21, we fail to reject null hypothesis, i.e. no auto correlation.

```
> vif(fit3)
                              GVIF Df GVIF^(1/(2*Df))
ParticipantOccupation     1.139259 49        1.001331
TeamDivision_1            1.226720  1        1.107574
TeamDivision_2            1.023664  1        1.011763
TeamDivision_3            1.085641  1        1.041941
TeamDivision_5            1.011330  1        1.005649
sqrt(TeamTotalConfirmed)  1.076499  1        1.037545
```

Fig 22. VIF values for Fit3 model

As seen in the figure 22, VIF for our final model is less than 3, so we can say that there is no multi-collinearity among the variables.

Now, considering the assumptions for the Fit6 model, we get

Fit6 model assumptions:
fit6 = lm(sqrt(sqrt(TeamTotalConfirmed)) ~ ParticipantOccupation + log(NumberofParticipants), data = data3.train)

In this model as well, we have taken log and square root in order to make the distribution normal.

Below are the graphs in the figure 23 through which we can confirm that there is linearity, normality and homoscedasticity among the variables.

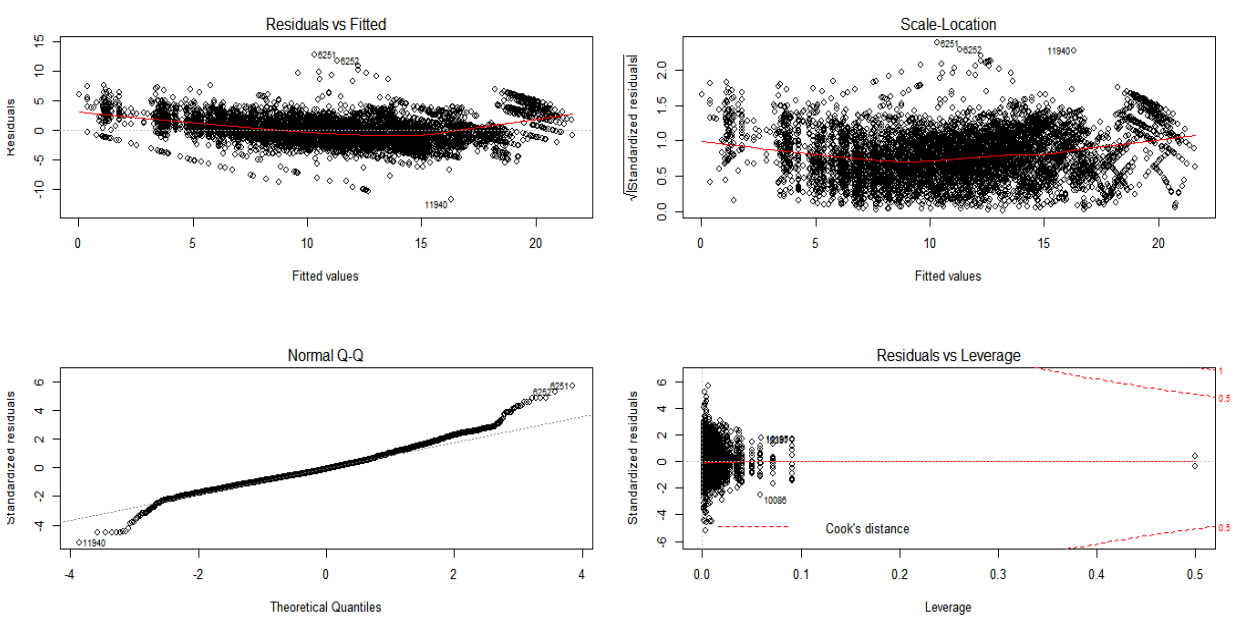

Fig 23. Conditions of Linearity, Normality and Homoscedasticity for Fit6 model

As seen below from the above Durbin Watson test, the statistic value is greater than 0.05, that means we fail to reject null hypothesis, i.e. there are no auto correlation among the variables.

```
> durbinWatsonTest(fit6)
 lag Autocorrelation D-W Statistic p-value
   1       0.7183493     0.5627186       0
 Alternative hypothesis: rho != 0
```

Fig 24. Durbin Watson Test for Fit6 model

```
> vif(fit6)
                              GVIF Df GVIF^(1/(2*Df))
ParticipantOccupation     1.028201 49        1.000284
log(NumberofParticipants) 1.028201  1        1.014003
```

Fig 25. VIF values for Fit6 model

The VIF test again confirms that we don't have any multi collinearity among the variables, as the value of VIF is less than 3.

Decision Tree

Some of the assumptions for the decision tree are that independent variables have non-overlying and appropriate levels to be used as a standard for splitting the decision tree. The splitting data should give the significant variables for the target variable.

Model Goals

Linear Regression

For the business questions we are doing the analysis on, linear regression gives us the best results. For the first business question, number of participants is a continuous variable, so we shall use the linear regression model in order to predict strongest involvement of the industries.

For the second business question, we are analyzing about the occupation which were responsible for most of the fundraising. Here, the target variable will be "TeamTotalConfirmed", which is again a continuous variable. We have done linear regression, as it helps us to determine the relation between the predictor and the target variable.

Decision Tree

Decision tree helps us to figure out the best significant variable for the target variable which we might not get in the linear regression. Since we can make use of both the classification as well as the regression tree as we have continuous target variable for one question and categorical target variable for another one, decision tree would fit the best.

Data Splitting and Sub-Sampling

Data splitting is a technique of partitioning the dataset into two different portions, Training and testing subsets. We do it mainly for cross-validation purposes. This technique is a traditional

approach in data science where we train the training dataset by using different model so that we can realize a better model to use on whole dataset.

We have chosen 70-30 splitting in our project, where 70% is for the training dataset and 30% is for the testing dataset. we chose 70-30 because,

1. As we have a huge dataset, we have decided to split the data in 70-30 manner instead of 50-50.
2. We took higher percentage in training dataset as we want to assure that we have enough data so that we can properly identify the perfect model for this dataset

Below is the code that we have used for data splitting:

```
#### Set the percentages of your subsets
train.size = 0.7
test.size = 0.3

#### Calculate the sample sizes
train2 = floor(train.size * nrow(data3))
test2 = floor(test.size * nrow(data3))

#### Determine the indices each subset will have
#### 1) randomly select the indices for the training set
#### 2) determine the remaining indices not in the training set
#### 3) from the list of indices in Step 2, randomly select
#### indices for the validation set
#### 4) determine the testing-subset indices by selecting those
```

```
#### not in the validation-subset
indices.train = sort(sample(seq_len(nrow(data3)), size=train2))
indices.valid_test = setdiff(seq_len(nrow(data3)), indices.train)
indices.test = sort(sample(indices.valid_test, size=test2))
#### Use the indices to select the data from the dataframe
data3.train = data3[indices.train,]
data3.test = data3[indices.test,]
nrow(data3.train)
nrow(data3.test)
```

Figure 26 shows the screenshot of number of rows for the training and testing data:

```
> nrow(data3.train)
[1] 8663
> nrow(data3.test)
[1] 3712
>
```

Fig 26. Count of rows in the training and testing datasets

For the variables "TeamTotalConfirmed", it can be seen that there is not much difference among the mean, standard deviation and median for the training, testing and the main dataset.

```
> describe(data3$TeamTotalConfirmed)
   vars     n     mean       sd median  trimmed      mad min      max    range
X1    1 12376 51555.62 86279.78  17009 28282.46 19895.01   0 396236.2 396236.2
   skew kurtosis     se
X1 2.54     5.72 775.57
> describe(data3.train$TeamTotalConfirmed)
   vars    n     mean       sd median  trimmed      mad min      max    range
X1    1 8663 51462.66 85869.83  16933 28336.68 19885.24   0 396236.2 396236.2
   skew kurtosis     se
X1 2.53     5.66 922.59
> describe(data3.test$TeamTotalConfirmed)
   vars    n    mean       sd median  trimmed     mad min      max    range skew
X1    1 3712 51779.1 87250.85  17170 28155.76 20015.1   0 396236.2 396236.2 2.57
   kurtosis      se
X1     5.82 1432.07
```

Fig 27. Mean, Standard Deviation and Median of the training, testing and the actual dataset for "TeamTotalConfirmed" variable

For the variable, "NumberofParticipants", we can see as given in the figure 28 that mean, median and standard deviation for all the training, testing and main dataset is almost the same.

```
> describe(data3$NumberofParticipants)
   vars     n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 12376 42.76 52.67     22   30.79 23.72   1 235   234 1.93     2.99 0.47
> describe(data3.train$NumberofParticipants)
   vars    n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 8663 43.02 52.87     22   31.06 23.72   1 235   234 1.92     2.95 0.57
> describe(data3.test$NumberofParticipants)
   vars    n  mean   sd median trimmed   mad min max range skew kurtosis   se
X1    1 3712 42.17 52.2     21   30.17 22.24   1 235   234 1.96     3.07 0.86
```

Fig 28. Mean, Standard Deviation and Median of the training, testing and the actual dataset for "NumberofParticipants" variable

The same facts as stated above can be reemphasized through the t-test which we have performed.

Figure 29 provides the screenshot for the same.

```
> t.test(data3$TeamTotalConfirmed,data3.train$TeamTotalConfirmed)

        Welch Two Sample t-test

data:  data3$TeamTotalConfirmed and data3.train$TeamTotalConfirmed
t = -0.29694, df = 18671, p-value = 0.7665
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2723.704  2007.040
sample estimates:
mean of x mean of y
 51555.62  51913.96
```

Fig 29. T-test results for the training dataset of "TeamTotalConfirmed" variable

From the above t-test we can see that the p-value is 0.7665 which is greater than 0.05, so we fail to reject the null hypothesis, i.e. the difference between the mean of the two sample is zero.

Building the Models

For measuring the strength of the industries participating in the Bike MS competition from 2013 to 2017, we have used linear regression model. We have made 3 models in order to check which one is the best one, so we could test the efficiency of that model using the test data. We have used the training data to create the three models.

Linear Regression Model:

For the 1st model, below is the code:

```
> fit1 = lm(NumberofParticipants ~ ParticipantOccupation + bike_gender_2 +
+ TeamDivision_1 + TeamDivision_2 + TeamDivision_3 + TeamDivision_5,  data = data3.train)
> summary(fit1)
```

Fig 30. Programming Code for the 1st regression model

Figure 31 given below provides the output of the 1st regression model

```
Call:
lm(formula = NumberofParticipants ~ ParticipantOccupation + bike_gender_2 +
    TeamDivision_1 + TeamDivision_2 + TeamDivision_3 + TeamDivision_5,
    data = data3.train)

Residuals:
   Min     1Q Median     3Q    Max
-84.12 -30.39 -14.32  13.13 193.78

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                              22.5553     2.9067   7.760 9.49e-15 ***
ParticipantOccupationAdministrative, Support, and Clerical -1.3298     4.7043  -0.283  0.77743
ParticipantOccupationAdvertising                         -9.4102     7.2079  -1.306  0.19174
ParticipantOccupationAerospace and Defense               -9.2967     7.3281  -1.269  0.20460
ParticipantOccupationAgriculture, Forestry, and Fishing  -5.8109    11.4816  -0.506  0.61280
ParticipantOccupationArchitecture                         0.4672     7.1575   0.065  0.94796
ParticipantOccupationArts and Entertainment               2.2636     5.7626   0.393  0.69447
ParticipantOccupationAviation and Airlines              -23.3170     9.3800  -2.486  0.01294 *
ParticipantOccupationBanking and Financial Services      -2.9994     3.8677  -0.776  0.43806
ParticipantOccupationClergy                               3.8048     9.3881   0.405  0.68528
ParticipantOccupationConstruction and Landscaping        -9.1029     4.4348  -2.053  0.04014 *
ParticipantOccupationConsulting                          14.9410     3.8311   3.900 9.69e-05 ***
ParticipantOccupationEducation and Training              11.3063     3.6808   3.072  0.00214 **
ParticipantOccupationEngineering                          6.5160     3.2384   2.012  0.04424 *
ParticipantOccupationEnvironment                         -4.0651     6.2518  -0.650  0.51556
ParticipantOccupationExecutive/Management                 5.8015     3.5539   1.632  0.10262
ParticipantOccupationFacilities, Maintenance, and Repair -7.1063     7.1174  -0.998  0.31809
ParticipantOccupationFire, Law Enforcement, and Security  8.9692     5.8304   1.538  0.12400
ParticipantOccupationGovernment                          10.8622     5.0886   2.135  0.03282 *
ParticipantOccupationHealthcare                           3.6699     3.2468   1.130  0.25838
ParticipantOccupationHomemaking                          -0.7073     9.3835  -0.075  0.93992
ParticipantOccupationHotel, Gaming, Leisure, and Travel  17.2267     6.4538   2.669  0.00762 **
ParticipantOccupationHuman Resources                      5.6602     5.1605   1.097  0.27274
ParticipantOccupationInformation Technology (IT)         15.2256     3.3612   4.530 5.98e-06 ***
ParticipantOccupationInsurance                           -4.4722     4.8241  -0.927  0.35392
ParticipantOccupationLegal and Paralegal                  5.8906     4.2792   1.377  0.16868
ParticipantOccupationManufacturing                        5.5057     5.0897   1.082  0.27940
ParticipantOccupationMarketing                            6.8650     4.1685   1.647  0.09962 .
ParticipantOccupationMedia                               11.0465     6.4228   1.720  0.08549 .
ParticipantOccupationMilitary                             8.4691     8.2804   1.023  0.30644
ParticipantOccupationNonprofit                            4.1038     5.5026   0.746  0.45581
ParticipantOccupationOil and Gas                         -4.0343     5.9973  -0.673  0.50116
ParticipantOccupationPersonal Care and Service           12.5610     8.0062   1.569  0.11671
ParticipantOccupationPhotography                          5.1736    15.2846   0.338  0.73501
ParticipantOccupationProperty Management                 20.2226     9.2350   2.190  0.02857 *
ParticipantOccupationPsychology                          -7.4183    11.2198  -0.661  0.50851
ParticipantOccupationPublishing                          20.7970    13.6088   1.528  0.12650
ParticipantOccupationReal Estate, Rental, and Leasing     6.3251     4.2221   1.498  0.13414
ParticipantOccupationRestaurant and Food Services        12.2401     6.4592   1.895  0.05813 .
ParticipantOccupationRetail/Wholesale                    -4.9582     5.2802  -0.939  0.34774
ParticipantOccupationRetired                            -19.5553    49.9017  -0.392  0.69516
ParticipantOccupationSales                               34.7975     3.4540  10.074  < 2e-16 ***
ParticipantOccupationScience and Biotechnology           -0.9085     4.7813  -0.190  0.84931
ParticipantOccupationSkilled Work and Trades              4.8000     5.3255   0.901  0.36744
ParticipantOccupationSocial Work                         -8.2000     8.9868  -0.912  0.36156
ParticipantOccupationStock Broker/Investment Advisor     -0.4574     8.0235  -0.057  0.95454
ParticipantOccupationStudent                             12.9722     6.3154   2.054  0.04000 *
ParticipantOccupationTechnical Account Manager          -44.6855    35.3423  -1.264  0.20613
ParticipantOccupationTelecommunications                   8.5881     6.7322   1.276  0.20210
ParticipantOccupationTransportation and Warehousing       9.5858     5.3406   1.795  0.07271 .
bike_gender_2                                            -2.8652     1.2451  -2.301  0.02140 *
TeamDivision_1                                           30.6302     1.1650  26.293  < 2e-16 ***
TeamDivision_2                                           -3.5387     5.1469  -0.688  0.49176
TeamDivision_3                                            3.6366     2.1977   1.655  0.09802 .
TeamDivision_5                                            6.5599     7.8418   0.837  0.40288
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.82 on 8608 degrees of freedom
Multiple R-squared:  0.1178,    Adjusted R-squared:  0.1122
F-statistic: 21.28 on 54 and 8608 DF,  p-value: < 2.2e-16
```

Fig 31. Output of the 1st Regression Model

For the 2nd model, below is the code:

```
> fit2 = lm(NumberofParticipants ~ ParticipantOccupation + TeamDivision_1 +
+ TeamDivision_2 + TeamDivision_3 + TeamDivision_5,  data = data3.train)
> summary(fit2)
```

Fig 32. Working Code for the 2nd regression model

Figure 33 split up in 3 different parts shows the output of the 2nd regression model

```
Call:
lm(formula = NumberofParticipants ~ ParticipantOccupation + TeamDivision_1 +
    TeamDivision_2 + TeamDivision_3 + TeamDivision_5, data = data3.train)

Residuals:
   Min     1Q Median     3Q    Max
-84.76 -30.04 -14.38  13.15 193.51

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                                21.2005     2.8472
ParticipantOccupationAdministrative, Support, and Clerical -0.1449     4.6772
ParticipantOccupationAdvertising                           -9.7577     7.2081
ParticipantOccupationAerospace and Defense                -10.0654     7.3223
ParticipantOccupationAgriculture, Forestry, and Fishing    -6.5470    11.4800
ParticipantOccupationArchitecture                          -0.2008     7.1534
ParticipantOccupationArts and Entertainment                 2.1761     5.7639
ParticipantOccupationAviation and Airlines                -24.2880     9.3729
ParticipantOccupationBanking and Financial Services        -3.5846     3.8603
ParticipantOccupationClergy                                 2.5955     9.3757
ParticipantOccupationConstruction and Landscaping         -10.2675     4.4069
ParticipantOccupationConsulting                            14.3117     3.8223
ParticipantOccupationEducation and Training                11.4860     3.6809
ParticipantOccupationEngineering                            5.4824     3.2079
ParticipantOccupationEnvironment                           -4.6891     6.2474
ParticipantOccupationExecutive/Management                   4.8724     3.5318
ParticipantOccupationFacilities, Maintenance, and Repair   -8.1889     7.1036
ParticipantOccupationFire, Law Enforcement, and Security    8.0278     5.8175
ParticipantOccupationGovernment                            10.8761     5.0898
ParticipantOccupationHealthcare                             3.9276     3.2457
ParticipantOccupationHomemaking                             0.5803     9.3691
ParticipantOccupationHotel, Gaming, Leisure, and Travel    16.7534     6.4521
ParticipantOccupationHuman Resources                        6.2204     5.1560
ParticipantOccupationInformation Technology (IT)           14.1840     3.3314
```

Fig 33a. Output of the 2nd regression model

```
ParticipantOccupationPhotography                          4.4527    15.2852
ParticipantOccupationProperty Management                 20.1645     9.2373
ParticipantOccupationPsychology                          -6.8523    11.2199
ParticipantOccupationPublishing                          19.5361    13.6012
ParticipantOccupationReal Estate, Rental, and Leasing     5.8380     4.2178
ParticipantOccupationRestaurant and Food Services        11.7074     6.4567
ParticipantOccupationRetail/Wholesale                    -5.5484     5.2753
ParticipantOccupationRetired                            -18.2005    49.9107
ParticipantOccupationSales                               34.0068     3.4377
ParticipantOccupationScience and Biotechnology           -1.3855     4.7780
ParticipantOccupationSkilled Work and Trades              3.6814     5.3046
ParticipantOccupationSocial Work                         -7.4940     8.9838
ParticipantOccupationStock Broker/Investment Advisor     -1.7188     8.0068
ParticipantOccupationStudent                             13.0552     6.3169
ParticipantOccupationTechnical Account Manager          -43.2538    35.3457
ParticipantOccupationTelecommunications                   7.8526     6.7263
ParticipantOccupationTransportation and Warehousing       8.4718     5.3199

TeamDivision_1                                           30.5533     1.1648
TeamDivision_2                                           -3.7797     5.1471
TeamDivision_3                                            3.6308     2.1983
TeamDivision_5                                            6.5862     7.8437
                                                        t value Pr(>|t|)
(Intercept)                                               7.446 1.05e-13 ***
ParticipantOccupationAdministrative, Support, and Clerical -0.031 0.975290
ParticipantOccupationAdvertising                         -1.354 0.175863
ParticipantOccupationAerospace and Defense               -1.375 0.169282
ParticipantOccupationAgriculture, Forestry, and Fishing  -0.570 0.568491
ParticipantOccupationArchitecture                        -0.028 0.977609
ParticipantOccupationArts and Entertainment               0.378 0.705780
ParticipantOccupationAviation and Airlines               -2.591 0.009577 **
ParticipantOccupationBanking and Financial Services      -0.929 0.353127
ParticipantOccupationClergy                               0.277 0.781918
ParticipantOccupationConstruction and Landscaping        -2.330 0.019836 *
ParticipantOccupationConsulting                           3.744 0.000182 ***
ParticipantOccupationEducation and Training               3.120 0.001812 **
ParticipantOccupationEngineering                          1.709 0.087477 .
ParticipantOccupationEnvironment                         -0.751 0.452934
ParticipantOccupationExecutive/Management                 1.380 0.167745
ParticipantOccupationFacilities, Maintenance, and Repair -1.153 0.249031
ParticipantOccupationFire, Law Enforcement, and Security  1.380 0.167641
ParticipantOccupationGovernment                           2.137 0.032640 *
ParticipantOccupationHealthcare                           1.210 0.226276
ParticipantOccupationHomemaking                           0.062 0.950612
ParticipantOccupationHotel, Gaming, Leisure, and Travel   2.597 0.009432 **
ParticipantOccupationHuman Resources                      1.206 0.227684
ParticipantOccupationInformation Technology (IT)          4.258 2.09e-05 ***
ParticipantOccupationInsurance                           -0.991 0.321509
ParticipantOccupationLegal and Paralegal                  1.302 0.193113
ParticipantOccupationManufacturing                        0.887 0.374945
ParticipantOccupationMarketing                            1.583 0.113451
```

Fig 33b. Output of the 2nd regression model

```
ParticipantOccupationMedia                                  1.623 0.104589
ParticipantOccupationMilitary                               0.886 0.375768
ParticipantOccupationNonprofit                              0.837 0.402790
ParticipantOccupationOil and Gas                           -0.824 0.409913
ParticipantOccupationPersonal Care and Service              1.628 0.103577
ParticipantOccupationPhotography                            0.291 0.770823
ParticipantOccupationProperty Management                    2.183 0.029066 *
ParticipantOccupationPsychology                            -0.611 0.541395
ParticipantOccupationPublishing                             1.436 0.150939
ParticipantOccupationReal Estate, Rental, and Leasing       1.384 0.166351
ParticipantOccupationRestaurant and Food Services           1.813 0.069834 .
ParticipantOccupationRetail/Wholesale                      -1.052 0.292935
ParticipantOccupationRetired                               -0.365 0.715373
ParticipantOccupationSales                                  9.892  < 2e-16 ***
ParticipantOccupationScience and Biotechnology             -0.290 0.771842
ParticipantOccupationSkilled Work and Trades                0.694 0.487702
ParticipantOccupationSocial Work                           -0.834 0.404211
ParticipantOccupationStock Broker/Investment Advisor       -0.215 0.830034
ParticipantOccupationStudent                                2.067 0.038791 *
ParticipantOccupationTechnical Account Manager             -1.224 0.221085
ParticipantOccupationTelecommunications                     1.167 0.243062
ParticipantOccupationTransportation and Warehousing         1.592 0.111317
TeamDivision_1                                             26.231  < 2e-16 ***
TeamDivision_2                                             -0.734 0.462771
TeamDivision_3                                              1.652 0.098636 .
TeamDivision_5                                              0.840 0.401113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.83 on 8609 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.1118
F-statistic: 21.57 on 53 and 8609 DF,  p-value: < 2.2e-16
```

Fig 33c. Output of the 2nd regression model

Working Code for the 3rd regression model is shown in the figure 34 below:

```
> fit3 = lm(log(NumberofParticipants) ~ ParticipantOccupation + TeamDivision_1 +
+ TeamDivision_2 + TeamDivision_3 + TeamDivision_5 + sqrt(TeamTotalConfirmed),  data = data3.train)
> summary(fit3)
```

Fig 34. Working code for the 3rd regression model

Figure 35 shows the output of the 3rd regression model below.

```
Call:
lm(formula = log(NumberofParticipants) ~ ParticipantOccupation +
    TeamDivision_1 + TeamDivision_2 + TeamDivision_3 + TeamDivision_5 +
    sqrt(TeamTotalConfirmed), data = data3.train)

Residuals:
     Min       1Q   Median       3Q      Max
-3.11382 -0.39961  0.06775  0.46629  1.99173

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                            1.687e+00  3.838e-02  43.953  < 2e-16 ***
ParticipantOccupationAdministrative, Support, and Clerical  1.241e-01  6.209e-02   1.998 0.045715 *
ParticipantOccupationAdvertising                       1.291e-01  9.570e-02   1.349 0.177296
ParticipantOccupationAerospace and Defense             2.694e-02  9.723e-02   0.277 0.781763
ParticipantOccupationAgriculture, Forestry, and Fishing  -4.730e-02  1.524e-01  -0.310 0.756301
ParticipantOccupationArchitecture                      -1.998e-01  9.497e-02  -2.103 0.035458 *
ParticipantOccupationArts and Entertainment            -2.279e-02  7.652e-02  -0.298 0.765846
ParticipantOccupationAviation and Airlines             9.600e-02  1.245e-01   0.771 0.440664
ParticipantOccupationBanking and Financial Services    1.573e-01  5.125e-02   3.069 0.002157 **
ParticipantOccupationClergy                            1.541e-01  1.245e-01   1.238 0.215767
ParticipantOccupationConstruction and Landscaping      -3.668e-02  5.852e-02  -0.627 0.530754
ParticipantOccupationConsulting                        -1.101e-01  5.080e-02  -2.168 0.030179 *
ParticipantOccupationEducation and Training            -2.123e-02  4.889e-02  -0.434 0.664127
ParticipantOccupationEngineering                       -1.697e-02  4.259e-02  -0.399 0.690241
ParticipantOccupationEnvironment                       -2.220e-01  8.294e-02  -2.676 0.007455 **
ParticipantOccupationExecutive/Management              -2.318e-02  4.689e-02  -0.494 0.621123
ParticipantOccupationFacilities, Maintenance, and Repair  -1.695e-02  9.431e-02  -0.180 0.857410
ParticipantOccupationFire, Law Enforcement, and Security  3.210e-02  7.724e-02   0.416 0.677746
ParticipantOccupationGovernment                        1.596e-02  6.759e-02   0.236 0.813322
ParticipantOccupationHealthcare                        3.632e-02  4.309e-02   0.843 0.399325
ParticipantOccupationHomemaking                        -5.956e-02  1.244e-01  -0.479 0.632070
ParticipantOccupationHotel, Gaming, Leisure, and Travel  2.605e-01  8.567e-02   3.041 0.002369 **
ParticipantOccupationHuman Resources                   2.353e-01  6.845e-02   3.437 0.000590 ***
ParticipantOccupationInformation Technology (IT)       4.987e-02  4.425e-02   1.127 0.259739
ParticipantOccupationInsurance                         1.463e-01  6.405e-02   2.284 0.022406 *
ParticipantOccupationLegal and Paralegal               -1.174e-01  5.680e-02  -2.066 0.038814 *
ParticipantOccupationManufacturing                     7.244e-03  6.733e-02   0.108 0.914331
ParticipantOccupationMarketing                         -1.283e-01  5.534e-02  -2.318 0.020472 *
ParticipantOccupationMedia                             -1.838e-02  8.525e-02  -0.216 0.829253
ParticipantOccupationInformation Technology (IT)       4.987e-02  4.425e-02   1.127 0.259739
ParticipantOccupationInsurance                         1.463e-01  6.405e-02   2.284 0.022406 *
ParticipantOccupationLegal and Paralegal               -1.174e-01  5.680e-02  -2.066 0.038814 *
ParticipantOccupationManufacturing                     7.244e-03  6.733e-02   0.108 0.914331
ParticipantOccupationMarketing                         -1.283e-01  5.534e-02  -2.318 0.020472 *
ParticipantOccupationMedia                             -1.838e-02  8.525e-02  -0.216 0.829253
ParticipantOccupationMilitary                          -7.919e-02  1.098e-01  -0.722 0.470613
ParticipantOccupationNonprofit                         -1.482e-01  7.301e-02  -2.030 0.042425 *
ParticipantOccupationOil and Gas                       -1.037e-01  7.947e-02  -1.305 0.191898
ParticipantOccupationPersonal Care and Service         2.785e-01  1.063e-01   2.621 0.008792 **
ParticipantOccupationPhotography                       2.223e-01  2.029e-01   1.095 0.273390
ParticipantOccupationProperty Management               1.052e-01  1.227e-01   0.858 0.391073
ParticipantOccupationPsychology                        -3.871e-01  1.489e-01  -2.599 0.009372 **
ParticipantOccupationPublishing                        -1.760e-01  1.806e-01  -0.975 0.329646
ParticipantOccupationReal Estate, Rental, and Leasing  -1.644e-01  5.605e-02  -2.933 0.003364 **
ParticipantOccupationRestaurant and Food Services      -2.526e-01  8.574e-02  -2.946 0.003229 **
ParticipantOccupationRetail/Wholesale                  -1.615e-01  7.004e-02  -2.306 0.021151 *
ParticipantOccupationRetired                           -9.432e-01  6.626e-01  -1.424 0.154620
ParticipantOccupationSales                             -6.017e-02  4.580e-02  -1.314 0.188950
ParticipantOccupationScience and Biotechnology         1.511e-01  6.343e-02   2.382 0.017258 *
ParticipantOccupationSkilled Work and Trades           -1.174e-01  7.042e-02  -1.668 0.095395 .
ParticipantOccupationSocial Work                       -1.492e-01  1.193e-01  -1.251 0.211009
ParticipantOccupationStock Broker/Investment Advisor   -1.007e-01  1.063e-01  -0.948 0.343303
ParticipantOccupationStudent                           -1.461e-02  8.389e-02  -0.174 0.861724
ParticipantOccupationTechnical Account Manager         -5.959e-01  4.693e-01  -1.270 0.204141
ParticipantOccupationTelecommunications                -2.863e-02  8.930e-02  -0.321 0.748524
ParticipantOccupationTransportation and Warehousing    2.679e-02  7.063e-02   0.379 0.704449
TeamDivision_1                                         3.361e-01  1.575e-02  21.337  < 2e-16 ***
TeamDivision_2                                         2.493e-01  6.834e-02   3.649 0.000265 ***
TeamDivision_3                                         5.136e-02  2.919e-02   1.759 0.078536 .
TeamDivision_5                                         6.287e-01  1.041e-01   6.037 1.64e-09 ***
sqrt(TeamTotalConfirmed)                               6.972e-03  5.142e-05 135.573  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6615 on 8608 degrees of freedom
Multiple R-squared:  0.7151,    Adjusted R-squared:  0.7133
F-statistic: 400.1 on 54 and 8608 DF,  p-value: < 2.2e-16
```

Fig 35. Output of the 3rd regression model

We have made 3 models for the analysis of the second question which asks for the occupations which have the highest contribution for the National Bike MS.

As mentioned above, we are again using the linear regression for all of the 3 models, and for the best model we shall use the test data in order to check for the efficiency of the model.

Below is the code for the 1st regression model for the second business question:

```
> fit4 = lm(TeamTotalConfirmed ~ ParticipantOccupation +  bike_participantconnectiontoms_2 + bike_participantconnectiontoms_3 +
+  bike_participantconnectiontoms_4 + bike_participantconnectiontoms_5 +
+  bike_participantconnectiontoms_6 + bike_participantconnectiontoms_7 +
+  bike_participantconnectiontoms_8 + bike_participantconnectiontoms_9 +
+  bike_participantconnectiontoms_10 + bike_participantconnectiontoms_11 +
+  bike_participantconnectiontoms_12+ bike_participantconnectiontoms_13+ bike_gender_2 + bike_gender_3, data = data3.train)
>
```

Fig 36. Working Code of 1st regression model of the 2nd business question

Figure 37 shows the output of the same as given below

```
Residuals:
    Min      1Q  Median      3Q     Max 
-103333  -43306  -28242     788  366812 

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                              38613.2     6862.8   5.626 1.90e-08 ***
ParticipantOccupationAdministrative, Support, and Clerical  -7551.2     7989.9  -0.945 0.344636    
ParticipantOccupationAdvertising                        -13536.2    12245.1  -1.105 0.269003    
ParticipantOccupationAerospace and Defense              -10876.1    12454.2  -0.873 0.382527    
ParticipantOccupationAgriculture, Forestry, and Fishing -13120.8    19506.6  -0.673 0.501201    
ParticipantOccupationArchitecture                        10700.4    12159.1   0.880 0.378867    
ParticipantOccupationArts and Entertainment               1837.9     9801.6   0.188 0.851261    
ParticipantOccupationAviation and Airlines              -27929.5    15920.4  -1.754 0.079411 .  
ParticipantOccupationBanking and Financial Services      -9834.5     6570.1  -1.497 0.134469    
ParticipantOccupationClergy                             -13878.6    15943.9  -0.870 0.384068    
ParticipantOccupationConstruction and Landscaping       -15684.5     7540.5  -2.080 0.037553 *  
ParticipantOccupationConsulting                          27928.0     6516.5   4.286 1.84e-05 ***
ParticipantOccupationEducation and Training               6419.5     6214.7   1.033 0.301655    
ParticipantOccupationEngineering                          9920.9     5510.7   1.800 0.071848 .  
ParticipantOccupationEnvironment                          1743.7    10634.3   0.164 0.869762    
ParticipantOccupationExecutive/Management                 5717.1     6044.7   0.946 0.344277    
ParticipantOccupationFacilities, Maintenance, and Repair -20612.0    12095.4  -1.704 0.088395 .  
ParticipantOccupationFire, Law Enforcement, and Security  2770.0     9901.2   0.280 0.779668    
ParticipantOccupationGovernment                          15302.1     8636.9   1.772 0.076478 .  
ParticipantOccupationHealthcare                           -885.2     5508.6  -0.161 0.872331    
ParticipantOccupationHomemaking                         -11025.0    15941.2  -0.692 0.489204    
ParticipantOccupationHotel, Gaming, Leisure, and Travel  15259.4    10965.5   1.392 0.164083    
ParticipantOccupationHuman Resources                       628.6     8766.3   0.072 0.942836    
ParticipantOccupationInformation Technology (IT)         15913.5     5715.9   2.784 0.005380 ** 
ParticipantOccupationInsurance                          -19465.7     8191.9  -2.376 0.017513 *  
ParticipantOccupationLegal and Paralegal                  8834.4     7269.4   1.215 0.224292    
ParticipantOccupationManufacturing                        2000.5     8649.1   0.231 0.817095    
ParticipantOccupationMarketing                           11821.8     7082.2   1.669 0.095108 .  
ParticipantOccupationMedia                               16224.2    10903.0   1.488 0.136774    
ParticipantOccupationMilitary                             2245.3    14092.5   0.159 0.873418    
ParticipantOccupationNonprofit                           -6337.9     9345.3  -0.678 0.497668    
ParticipantOccupationOil and Gas                          6607.8    10189.2   0.649 0.516673    
ParticipantOccupationPersonal Care and Service            -231.4    13602.1  -0.017 0.986427    
ParticipantOccupationPhotography                         -6449.1    25948.2  -0.249 0.803723    
ParticipantOccupationProperty Management                 24255.2    15675.9   1.547 0.121830    
ParticipantOccupationPsychology                          -10349.9    19094.0  -0.542 0.587799    
ParticipantOccupationPublishing                          26708.9    23114.3   1.156 0.247912    
ParticipantOccupationReal Estate, Rental, and Leasing    25616.2     7168.4   3.573 0.000354 ***
ParticipantOccupationRestaurant and Food Services        21653.0    10970.6   1.974 0.048444 *  
ParticipantOccupationRetail/Wholesale                   -12168.4     8979.7  -1.355 0.175422    
ParticipantOccupationRetired                            -45750.3    84821.4  -0.539 0.589644    
ParticipantOccupationSales                               49184.1     5871.9   8.376  < 2e-16 ***
ParticipantOccupationScience and Biotechnology           -7698.1     8121.4  -0.948 0.343218    
ParticipantOccupationSkilled Work and Trades              5156.0     9052.5   0.570 0.568987    
ParticipantOccupationSocial Work                        -12323.3    15271.0  -0.807 0.419704    
ParticipantOccupationStock Broker/Investment Advisor       261.5    13627.8   0.019 0.984690    
ParticipantOccupationStudent                             13255.0    10753.7   1.233 0.217760    
ParticipantOccupationTechnical Account Manager          -41430.3    60144.6  -0.689 0.490939    
ParticipantOccupationTelecommunications                  10045.0    11435.3   0.878 0.379739    
ParticipantOccupationTransportation and Warehousing       3806.5     9083.1   0.419 0.675168    
bike_participantconnectiontoms_2                         47875.6    27342.5   1.751 0.079989 .  
bike_participantconnectiontoms_3                         -7490.1     8471.2  -0.884 0.376619    
bike_participantconnectiontoms_4                          6630.2     5048.0   1.313 0.189067    
bike_participantconnectiontoms_5                        -45449.6    49166.2  -0.924 0.355300    
bike_participantconnectiontoms_6                         -4412.3    26073.0  -0.169 0.865620    
bike_participantconnectiontoms_7                          9727.1     6521.7   1.492 0.135865    
bike_participantconnectiontoms_8                          2043.0     5683.4   0.359 0.719257    
bike_participantconnectiontoms_9                          4311.4     6189.9   0.697 0.486123    
bike_participantconnectiontoms_10                         8735.3     6912.4   1.264 0.206369    
bike_participantconnectiontoms_11                           35.8    32433.4   0.001 0.999119    
bike_participantconnectiontoms_12                        15535.9     5527.1   2.811 0.004952 ** 
bike_participantconnectiontoms_13                        -3849.5     6747.9  -0.570 0.568372    
bike_gender_2                                            -2282.8     2140.2  -1.067 0.286185    
bike_gender_3                                             3846.1    16800.6   0.229 0.818932    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84580 on 8599 degrees of freedom
Multiple R-squared:  0.03676,	Adjusted R-squared:  0.0297 
F-statistic: 5.209 on 63 and 8599 DF,  p-value: < 2.2e-16
```

Fig 37. Output of the 1st regression model for the 2nd business question

Working code for the 2nd regression model for the 2nd business question is given below.

```
> fit5 = lm(TeamTotalConfirmed ~ ParticipantOccupation + bike_gender_2 +
+ bike_gender_3, data = data3.train)
>
> summary(fit5)
```

Fig 38. Working Code for the 2nd regression model for the 2nd business question

Figure 39 depicts the output of the above model

```
Call:
lm(formula = TeamTotalConfirmed ~ ParticipantOccupation + bike_gender_2 +
    bike_gender_3, data = data3.train)

Residuals:
   Min     1Q Median     3Q    Max
-94326 -43343 -28433    313 367583

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                             45585.55    4796.29   9.504  < 2e-16 ***
ParticipantOccupationAdministrative, Support, and Clerical  -7599.85    7994.54  -0.951 0.341819
ParticipantOccupationAdvertising                        -14312.91   12248.14  -1.169 0.242606
ParticipantOccupationAerospace and Defense              -11901.55   12450.89  -0.956 0.339160
ParticipantOccupationAgriculture, Forestry, and Fishing -14917.41   19511.86  -0.765 0.444572
ParticipantOccupationArchitecture                         9708.23   12163.33   0.798 0.424802
ParticipantOccupationArts and Entertainment               1600.23    9789.90   0.163 0.870162
ParticipantOccupationAviation and Airlines              -28687.83   15929.65  -1.801 0.071752 .
ParticipantOccupationBanking and Financial Services     -10267.96    6572.79  -1.562 0.118280
ParticipantOccupationClergy                             -15657.00   15936.61  -0.982 0.325904
ParticipantOccupationConstruction and Landscaping       -15840.11    7538.39  -2.101 0.035647 *
ParticipantOccupationConsulting                          28427.58    6516.99   4.362  1.3e-05 ***
ParticipantOccupationEducation and Training               5919.03    6211.68   0.953 0.340674
ParticipantOccupationEngineering                         10037.10    5508.66   1.822 0.068481 .
ParticipantOccupationEnvironment                          3506.60   10625.32   0.330 0.741390
ParticipantOccupationExecutive/Management                 5151.82    6040.65   0.853 0.393761
ParticipantOccupationFacilities, Maintenance, and Repair -21902.90   12089.91  -1.812 0.070072 .
ParticipantOccupationFire, Law Enforcement, and Security  3382.70    9892.43   0.342 0.732398
ParticipantOccupationGovernment                          15263.99    8617.39   1.771 0.076546 .
ParticipantOccupationHealthcare                           -559.47    5502.26  -0.102 0.919013
ParticipantOccupationHomemaking                         -12688.86   15941.92  -0.796 0.426087
ParticipantOccupationHotel, Gaming, Leisure, and Travel  14045.70   10966.63   1.281 0.200310
ParticipantOccupationHuman Resources                      -311.74    8765.17  -0.036 0.971630
ParticipantOccupationInformation Technology (IT)         15128.25    5713.02   2.648 0.008111 **
ParticipantOccupationInsurance                          -21317.43    8189.27  -2.603 0.009255 **
ParticipantOccupationLegal and Paralegal                  7661.21    7264.61   1.055 0.291641
ParticipantOccupationManufacturing                         869.16    8649.33   0.100 0.919959
ParticipantOccupationMarketing                           12245.80    7084.60   1.729 0.083933 .
ParticipantOccupationMedia                               17166.71   10909.51   1.574 0.115627
ParticipantOccupationMilitary                             -685.18   14052.59  -0.049 0.961113
ParticipantOccupationNonprofit                           -6434.20    9340.78  -0.689 0.490949
ParticipantOccupationOil and Gas                          6291.40   10190.82   0.617 0.537014
ParticipantOccupationPersonal Care and Service             525.96   13601.59   0.039 0.969155
ParticipantOccupationPhotography                         -6924.17   25960.82  -0.267 0.789694
ParticipantOccupationProperty Management                 25094.73   15683.64   1.600 0.109623
ParticipantOccupationPsychology                         -14115.77   19066.27  -0.740 0.459106
ParticipantOccupationPublishing                          27951.56   23128.14   1.209 0.226868
ParticipantOccupationReal Estate, Rental, and Leasing    25224.76    7163.10   3.521 0.000431 ***
ParticipantOccupationRestaurant and Food Services        22745.67   10967.62   2.074 0.038119 *
ParticipantOccupationRetail/Wholesale                   -11411.75    8966.52  -1.273 0.203156
ParticipantOccupationRetired                            -42995.55   84809.34  -0.507 0.612191
ParticipantOccupationSales                               48740.03    5870.75   8.302  < 2e-16 ***
ParticipantOccupationScience and Biotechnology           -7194.11    8126.30  -0.885 0.376027
ParticipantOccupationSkilled Work and Trades              4216.99    9047.24   0.466 0.641150
ParticipantOccupationSocial Work                        -13164.80   15267.20  -0.862 0.388550
ParticipantOccupationStock Broker/Investment Advisor       -59.79   13629.16  -0.004 0.996500
ParticipantOccupationStudent                             15414.23   10734.37   1.436 0.151047
ParticipantOccupationTechnical Account Manager          -38675.55   60065.08  -0.644 0.519661
ParticipantOccupationTelecommunications                  11139.90   11438.74   0.974 0.330146
ParticipantOccupationTransportation and Warehousing       2027.37    9074.29   0.223 0.823215
bike_gender_2                                            -2619.39    2122.91  -1.234 0.217285
bike_gender_3                                              737.54   16754.62   0.044 0.964889
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84670 on 8611 degrees of freedom
Multiple R-squared:  0.03339,	Adjusted R-squared:  0.02767
F-statistic: 5.833 on 51 and 8611 DF,  p-value: < 2.2e-16
```

Fig 39. Output of the 2nd regression model for the 2nd business question

Working code for the final model is mentioned below:

```
> fit6 = lm(sqrt(sqrt(TeamTotalConfirmed)) ~ ParticipantOccupation + log(NumberofParticipants), data = data3.train)
>
> summary(fit6)
```

Fig 40. Working code of the 3rd regression model for the 2nd business question

Figure 40 depicts the screenshot of the above code.

```
Call:
lm(formula = sqrt(sqrt(TeamTotalConfirmed)) ~ ParticipantOccupation +
    log(NumberofParticipants), data = data3.train)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6396  -1.4635  -0.1999   1.2864  12.9005

Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                               0.997655   0.137222   7.270 3.90e-13 ***
ParticipantOccupationAdministrative, Support, and Clerical -0.334380   0.209759  -1.594 0.110948
ParticipantOccupationAdvertising                         -0.646149   0.323293  -1.999 0.045678 *
ParticipantOccupationAerospace and Defense               -0.713089   0.328372  -2.172 0.029914 *
ParticipantOccupationAgriculture, Forestry, and Fishing  -0.528489   0.514921  -1.026 0.304755
ParticipantOccupationArchitecture                         0.930601   0.320861   2.900 0.003737 **
ParticipantOccupationArts and Entertainment               0.251796   0.258449   0.974 0.329956
ParticipantOccupationAviation and Airlines               -1.360668   0.420102  -3.239 0.001204 **
ParticipantOccupationBanking and Financial Services      -0.307030   0.173162  -1.773 0.076251 .
ParticipantOccupationClergy                              -0.111919   0.420088  -0.266 0.789924
ParticipantOccupationConstruction and Landscaping        -0.040158   0.197735  -0.203 0.839071
ParticipantOccupationConsulting                           0.774581   0.171517   4.516 6.38e-06 ***
ParticipantOccupationEducation and Training               0.501829   0.163943   3.061 0.002213 **
ParticipantOccupationEngineering                          0.162304   0.143907   1.128 0.259418
ParticipantOccupationEnvironment                          0.898334   0.280256   3.205 0.001354 **
ParticipantOccupationExecutive/Management                 0.319669   0.158430   2.018 0.043651 *
ParticipantOccupationFacilities, Maintenance, and Repair -0.023422   0.318553  -0.074 0.941388
ParticipantOccupationFire, Law Enforcement, and Security  0.153947   0.260537   0.591 0.554614
ParticipantOccupationGovernment                           0.338618   0.227495   1.488 0.136665
ParticipantOccupationHealthcare                           0.157992   0.145159   1.088 0.276445
ParticipantOccupationHomemaking                           0.576740   0.420104   1.373 0.169835
ParticipantOccupationHotel, Gaming, Leisure, and Travel  -0.496983   0.289565  -1.716 0.086142 .
ParticipantOccupationHuman Resources                     -0.427974   0.231281  -1.850 0.064283 .
ParticipantOccupationInformation Technology (IT)          0.090772   0.149509   0.607 0.543777
ParticipantOccupationInsurance                           -0.771501   0.216151  -3.569 0.000360 ***
ParticipantOccupationLegal and Paralegal                  0.771519   0.191651   4.026 5.73e-05 ***
ParticipantOccupationManufacturing                       -0.062236   0.227496  -0.274 0.784420
ParticipantOccupationMarketing                            0.397536   0.186959   2.126 0.033504 *
ParticipantOccupationMedia                                0.854919   0.287786   2.971 0.002980 **
ParticipantOccupationHealthcare                           0.157992   0.145159   1.088 0.276445
ParticipantOccupationHomemaking                           0.576740   0.420104   1.373 0.169835
ParticipantOccupationHotel, Gaming, Leisure, and Travel  -0.496983   0.289565  -1.716 0.086142 .
ParticipantOccupationHuman Resources                     -0.427974   0.231281  -1.850 0.064283 .
ParticipantOccupationInformation Technology (IT)          0.090772   0.149509   0.607 0.543777
ParticipantOccupationInsurance                           -0.771501   0.216151  -3.569 0.000360 ***
ParticipantOccupationLegal and Paralegal                  0.771519   0.191651   4.026 5.73e-05 ***
ParticipantOccupationManufacturing                       -0.062236   0.227496  -0.274 0.784420
ParticipantOccupationMarketing                            0.397536   0.186959   2.126 0.033504 *
ParticipantOccupationMedia                                0.854919   0.287786   2.971 0.002980 **
ParticipantOccupationMilitary                            -0.008867   0.370406  -0.024 0.980901
ParticipantOccupationNonprofit                           -0.090848   0.246502  -0.369 0.712475
ParticipantOccupationOil and Gas                          0.651462   0.268446   2.427 0.015254 *
ParticipantOccupationPersonal Care and Service           -0.642507   0.358975  -1.790 0.073515 .
ParticipantOccupationPhotography                         -0.113491   0.685229  -0.166 0.868456
ParticipantOccupationProperty Management                  0.604243   0.414100   1.459 0.144554
ParticipantOccupationPsychology                           1.121675   0.503377   2.228 0.025886 *
ParticipantOccupationPublishing                           0.311344   0.610082   0.510 0.609833
ParticipantOccupationReal Estate, Rental, and Leasing     1.115346   0.188882   5.905 3.66e-09 ***
ParticipantOccupationRestaurant and Food Services         1.061905   0.289377   3.670 0.000244 ***
ParticipantOccupationRetail/Wholesale                     0.269508   0.236605   1.139 0.254710
ParticipantOccupationRetired                              2.188103   2.239097   0.977 0.328485
ParticipantOccupationSales                                0.630071   0.154494   4.078 4.58e-05 ***
ParticipantOccupationScience and Biotechnology           -0.291234   0.214347  -1.359 0.174277
ParticipantOccupationSkilled Work and Trades              0.333459   0.237881   1.402 0.161015
ParticipantOccupationSocial Work                          0.881557   0.402860   2.188 0.028678 *
ParticipantOccupationStock Broker/Investment Advisor      0.371648   0.358963   1.035 0.300539
ParticipantOccupationStudent                              0.725390   0.283167   2.562 0.010433 *
ParticipantOccupationTechnical Account Manager            0.637916   1.585599   0.402 0.687459
ParticipantOccupationTelecommunications                   0.340998   0.301646   1.130 0.258315
ParticipantOccupationTransportation and Warehousing       0.226721   0.238584   0.950 0.341998
log(NumberofParticipants)                                 3.593723   0.019714 182.291  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.235 on 8612 degrees of freedom
Multiple R-squared:  0.8005,    Adjusted R-squared:  0.7993
F-statistic: 691.1 on 50 and 8612 DF,  p-value: < 2.2e-16
```

Fig 41. Output of the 3rd regression model for the 2nd business question

Decision Tree:

The working code for the 2 decision trees created is given below. We created 2 decision trees.

- A regression tree is being created when the target variable is "NumberofParticipants" which is a continuous variable.
- A classification tree has been created when the target variable is "ParticipationOccupation" which is a categorical variable.

```
#decision tree1 – Regression Tree
tree2 <- rpart(NumberofParticipants ~ TeamDivision_1 + TeamDivision_2 + TeamDivision_3 +
TeamDivision_5 + TeamTotalConfirmed, data = data3.train, method='class')

#decision tree2– Classification Tree
tree1 <- rpart(ParticipantOccupation ~ NumberofParticipants + TeamDivision_1 +
TeamDivision_2 + TeamDivision_3 + TeamDivision_5 + TeamTotalConfirmed, data = data3.train,
method='class')
```

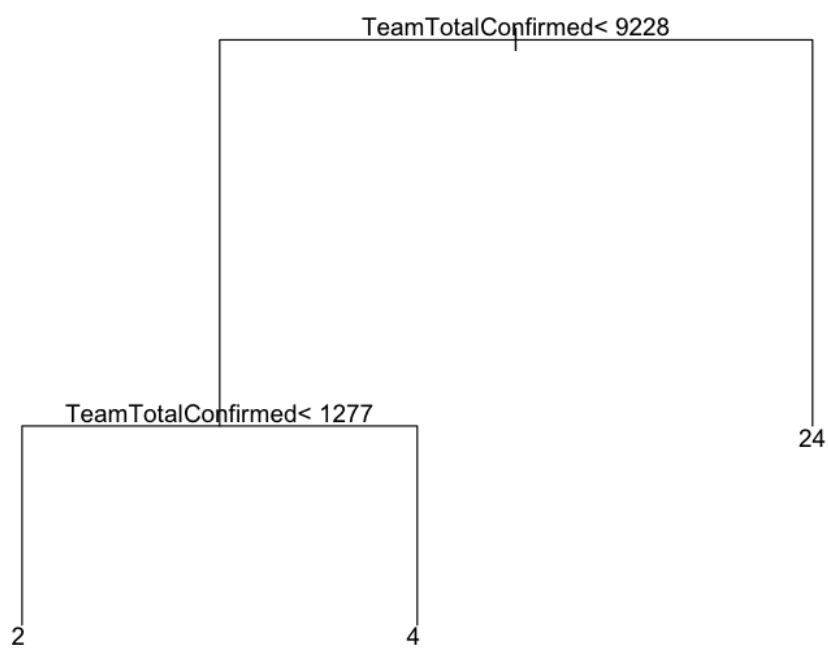

Fig 42. Regression Tree Output

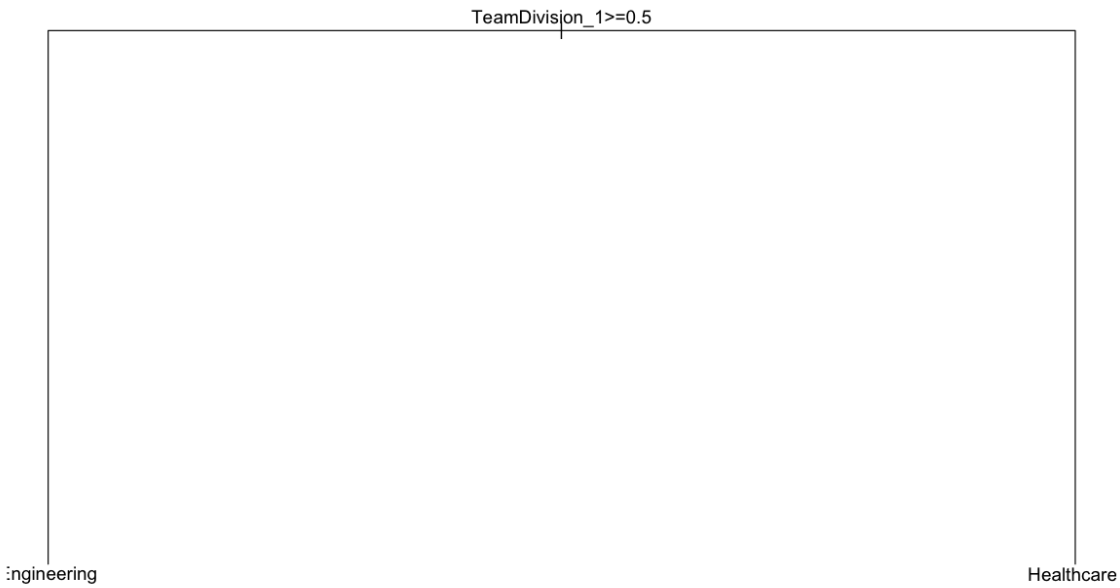

Fig 43. Classification Tree Output

Interpretation of Results

The interpretation of both the regression models and the decision trees are mentioned below.

Interpretation of Regression Models

According to our evaluation for the first question, i.e. which industries have the strongest involvement in the Bike MS competition, below are significant variables that we came across through our analyzation:

1) Administrative, support and clerical
2) Advertising

3) Architecture
4) banking and Financial Services
5) Consulting
6) Environment
7) Hotel, gaming, leisure and travel
8) Human Resources
9) Insurance
10) Legal and Paralegal
11) Insurance
12) Marketing
13) Military
14) Personal Care and Service
15) Psychology
16) Real Estate, Rental and Leasing
17) Restaurant and Food Services
18) Science and Biotechnology
19) TeamDivision_1 - Corporate
20) TeamDivision_2 - Corporation
21) TeamDivision_5 – Organization (Clubs, civic groups, etc)

These have the strongest involvement in the Bike MS competition as per our deduction.

The question that we are analyzing for our second model is, the occupations which are responsible for most of the fund raising. Below are the significant occupations that was the output from our model:

1) Administrative, Support and Clerical
2) Aerospace and Defense
3) Agriculture, Forestry, and Fishing
4) Architecture
5) Aviation and Airlines
6) Consulting
7) Education and Training
8) Environment
9) Insurance
10) Legal and Paralegal
11) Manufacturing
12) Marketing
13) Real Estate, Rental and Leasing
14) Restaurant and Food Services
15) Sales
16) Skilled Work and Trades
17) Stock Broker/Investment Advisor

These are the occupations which are most responsible for the highest fund raising.

Interpretation of Decision Trees:

In the regression decision tree as given in the figure 42, we can see that if "team_total_confirmed" is less than 9228 then it will go to the left part of the tree or else there will be 24 participants in a team. Now in the left side, if "team_total_confirmed" is less than 1277, then there will be two participants in a team.

If "team_total_confirmed" is not less than 1277, then there will be four participants in a team. Hence, according to the decision tree, the number of participants in a team is related to the total number of teams confirmed for the campaign.

In the classification decision tree, we can see if "team_division_1" >= 0.5, i.e. for a team from a corporate company, will be either part of engineering or healthcare domain. Hence, according to the classification decision tree, if the team division 1 (i.e., corporates) is greater than 0.5, i.e., if the corporate team division is present then the participant occupation would be Engineering or else it would be Healthcare domain as shown in the figure 43.

Assessing the Models

For our first final model, i.e. Fit3, figure 44 depicts the screenshot of the R square, adjusted R square and F statistic values:

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

Residual standard error: 0.6615 on 8608 degrees of freedom
Multiple R-squared:  0.7151,    Adjusted R-squared:  0.7133
F-statistic: 400.1 on 54 and 8608 DF,  p-value: < 2.2e-16
```

Fig 44. R-square, Adjusted R-square and F-statistic for Fit3 model

As we can see, both of the R-squared and adjusted R-squared value is approx. 71%, which is a very good value. Also, the p-value of the F-statistic for this model is less than 0.05, which means it is a significant model and we can proceed with this.

For our second final model, i.e. Fit3, figure 45 depicts the screenshot of the R square, adjusted R square and F statistic values:

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

Residual standard error: 2.235 on 8612 degrees of freedom
Multiple R-squared:  0.8005,    Adjusted R-squared:  0.7993
F-statistic: 691.1 on 50 and 8612 DF,  p-value: < 2.2e-16
```

Fig 45. R-square, Adjusted R-square and F-statistic for Fit3 model

Again, both the values of R-squared and adjusted R-squared values are approx. 80%. This means the data are very closely fitted to the regression line. Also, the p-value for the F-statistic is also less than 0.05, which means that this model is significant, and we can proceed with the model and use the testing data to check the efficiency of the same.

Strength and Weakness of the models

Below are the strengths of the first final model, i.e. Fit3:

- This model is clearly showing those industries which has the strongest involvement based upon the number of participants who are participating in the competition from 2013 to 2017.

- We can further cleanse the output by choosing the variables having the highest co-efficient value.

Below is the weakness of the first final model, i.e. Fit3:

- This model has made many of the variables insignificant, which might have strong involvement in the Bike MS competition logically.

Below are the strengths of the second final model, i.e. Fit6:

- This model is also showing all those variables along with their co-efficient values which has contributed towards the highest fund raising, taking the total funds as the target variable.
- The R-square value is also quite high, which further strengthens our model.

Below is the weakness of the second final model, i.e. Fit6:

- The only weakness that this model has, it has only taken all those variables which are statistically significant. There might have also contributed towards the fund raising, but due to the low amount, it has discarded them.

Justification of the choice of the models

For the chosen final models, we have performed the multiple regression using the test dataset.

Below is the screenshot of the multiple regression for fit3, using the test dataset:

```
> #Testing the best model using the test data:
> fitTest1 = lm(sqrt(sqrt(TeamTotalConfirmed)) ~ ParticipantOccupation + log(NumberofParticipants),  data = data3.test)
> summary(fitTest1)

Call:
lm(formula = sqrt(sqrt(TeamTotalConfirmed)) ~ ParticipantOccupation +
    log(NumberofParticipants), data = data3.test)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2454  -1.4700  -0.2186   1.2999  11.4544

Coefficients:
                                                        Estimate Std. Error
(Intercept)                                              0.83532    0.21741
ParticipantOccupationAdministrative, Support, and Clerical -0.16984    0.33090
ParticipantOccupationAdvertising                        -0.10189    0.46865
ParticipantOccupationAerospace and Defense              -1.08877    0.51841
ParticipantOccupationAgriculture, Forestry, and Fishing  0.08506    0.77461
ParticipantOccupationArchitecture                        1.40929    0.54013
ParticipantOccupationArts and Entertainment              1.04800    0.44407
ParticipantOccupationAviation and Airlines              -0.17747    0.63263
ParticipantOccupationBanking and Financial Services      0.19320    0.26166
ParticipantOccupationClergy                              0.73493    0.77481
ParticipantOccupationConstruction and Landscaping       -0.13684    0.31772
ParticipantOccupationConsulting                          1.17182    0.26315
ParticipantOccupationEducation and Training              0.80373    0.25800
ParticipantOccupationEngineering                         0.68646    0.22940
ParticipantOccupationEnvironment                         1.02570    0.49971
ParticipantOccupationExecutive/Management                0.76618    0.24754
ParticipantOccupationFacilities, Maintenance, and Repair 0.16338    0.50870
ParticipantOccupationFire, Law Enforcement, and Security 0.35755    0.39591
ParticipantOccupationGovernment                          1.40696    0.38931
ParticipantOccupationHealthcare                          0.51481    0.22978
ParticipantOccupationHomemaking                          0.75576    0.59560
ParticipantOccupationHotel, Gaming, Leisure, and Travel  0.33904    0.45576
ParticipantOccupationHuman Resources                    -0.22685    0.36263
ParticipantOccupationInformation Technology (IT)         0.34506    0.23276
ParticipantOccupationInsurance                          -0.26188    0.35145
ParticipantOccupationLegal and Paralegal                 0.96333    0.30702
ParticipantOccupationManufacturing                       0.45547    0.35355
ParticipantOccupationMarketing                           0.75711    0.29215
ParticipantOccupationMedia                               1.17304    0.47571
ParticipantOccupationMilitary                            0.80486    0.56554
ParticipantOccupationNonprofit                           0.43749    0.38949
ParticipantOccupationOil and Gas                         1.19611    0.48325
ParticipantOccupationPersonal Care and Service          -0.65512    0.51864
ParticipantOccupationPhotography                        -0.27705    1.02395
ParticipantOccupationProperty Management                 0.78879    0.61307
ParticipantOccupationPsychology                          1.81086    0.87198
ParticipantOccupationPublishing                          0.63958    0.77460
ParticipantOccupationReal Estate, Rental, and Leasing    1.59585    0.30239
ParticipantOccupationRestaurant and Food Services        1.36264    0.42845
ParticipantOccupationRetail/Wholesale                    0.93555    0.38325
ParticipantOccupationSales                               1.00664    0.24353
ParticipantOccupationScience and Biotechnology          -0.15850    0.34767
ParticipantOccupationSkilled Work and Trades             0.96480    0.35988
ParticipantOccupationSocial Work                         0.56611    0.70585
ParticipantOccupationStock Broker/Investment Advisor     0.63056    0.50872
ParticipantOccupationStudent                             0.95451    0.43852
ParticipantOccupationTechnical Account Manager           0.92117    1.60088
ParticipantOccupationTelecommunications                  1.10487    0.43853
ParticipantOccupationTransportation and Warehousing      0.71034    0.38936
log(NumberofParticipants)                                3.53251    0.03046
```

Fig 46a. Multiple Regression for Fit3 model

```
                                                    t value Pr(>|t|)
(Intercept)                                           3.842 0.000124 ***
ParticipantOccupationAdministrative, Support, and Clerical  -0.513 0.607795
ParticipantOccupationAdvertising                     -0.217 0.827905
ParticipantOccupationAerospace and Defense           -2.100 0.035779 *
ParticipantOccupationAgriculture, Forestry, and Fishing  0.110 0.912563
ParticipantOccupationArchitecture                     2.609 0.009113 **
ParticipantOccupationArts and Entertainment           2.360 0.018328 *
ParticipantOccupationAviation and Airlines           -0.281 0.779084
ParticipantOccupationBanking and Financial Services   0.738 0.460347
ParticipantOccupationClergy                           0.949 0.342921
ParticipantOccupationConstruction and Landscaping    -0.431 0.666716
ParticipantOccupationConsulting                       4.453 8.72e-06 ***
ParticipantOccupationEducation and Training           3.115 0.001852 **
ParticipantOccupationEngineering                      2.992 0.002786 **
ParticipantOccupationEnvironment                      2.053 0.040184 *
ParticipantOccupationExecutive/Management             3.095 0.001982 **
ParticipantOccupationFacilities, Maintenance, and Repair  0.321 0.748097
ParticipantOccupationFire, Law Enforcement, and Security  0.903 0.366525
ParticipantOccupationGovernment                       3.614 0.000306 ***
ParticipantOccupationHealthcare                       2.240 0.025121 *
ParticipantOccupationHomemaking                       1.269 0.204558
ParticipantOccupationHotel, Gaming, Leisure, and Travel  0.744 0.456987
ParticipantOccupationHuman Resources                 -0.626 0.531632
ParticipantOccupationInformation Technology (IT)      1.482 0.138308
ParticipantOccupationInsurance                       -0.745 0.456225
ParticipantOccupationLegal and Paralegal              3.138 0.001716 **
ParticipantOccupationManufacturing                    1.288 0.197740
ParticipantOccupationMarketing                        2.591 0.009594 **
ParticipantOccupationMedia                            2.466 0.013714 *
ParticipantOccupationMilitary                         1.423 0.154773
ParticipantOccupationNonprofit                        1.123 0.261412
ParticipantOccupationOil and Gas                      2.475 0.013362 *
ParticipantOccupationPersonal Care and Service       -1.263 0.206617
ParticipantOccupationPhotography                     -0.271 0.786733
ParticipantOccupationProperty Management                  1.287 0.198304
ParticipantOccupationPsychology                           2.077 0.037897 *
ParticipantOccupationPublishing                           0.826 0.409038
ParticipantOccupationReal Estate, Rental, and Leasing     5.277 1.39e-07 ***
ParticipantOccupationRestaurant and Food Services         3.180 0.001483 **
ParticipantOccupationRetail/Wholesale                     2.441 0.014689 *
ParticipantOccupationSales                                4.134 3.65e-05 ***
ParticipantOccupationScience and Biotechnology           -0.456 0.648497
ParticipantOccupationSkilled Work and Trades              2.681 0.007375 **
ParticipantOccupationSocial Work                          0.802 0.422595
ParticipantOccupationStock Broker/Investment Advisor      1.240 0.215239
ParticipantOccupationStudent                              2.177 0.029570 *
ParticipantOccupationTechnical Account Manager            0.575 0.565044
ParticipantOccupationTelecommunications                   2.520 0.011794 *
ParticipantOccupationTransportation and Warehousing       1.824 0.068175 .
log(NumberofParticipants)                               115.959  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.246 on 3662 degrees of freedom
Multiple R-squared:  0.7937,    Adjusted R-squared:  0.7909
F-statistic: 287.5 on 49 and 3662 DF,  p-value: < 2.2e-16
```

Fig 47b. Multiple regression for Fit 3 model

The R-squared and adjusted R-square values increased! It is now approx. 80% which is same as the previous model! Also, this model is significant too, which can be confirmed through the p-value of the f-statistics.

Now, we would perform the same for the second final model and check its efficiency using the test data. Figure 48 shows the screenshot of the output of the multiple regression model that we ran using the test dataset:

```
> fitTest2 = lm(log(NumberofParticipants) ~ ParticipantOccupation + TeamDivision_1 +
+ TeamDivision_2 + TeamDivision_3 + TeamDivision_5 + sqrt(TeamTotalConfirmed),  data = data3.test)
> summary(fitTest2)

Call:
lm(formula = log(NumberofParticipants) ~ ParticipantOccupation +
    TeamDivision_1 + TeamDivision_2 + TeamDivision_3 + TeamDivision_5 +
    sqrt(TeamTotalConfirmed), data = data3.test)

Residuals:
     Min       1Q   Median       3Q      Max
-2.93607 -0.39395  0.06592  0.47727  1.89799

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                               1.708975   0.061740
ParticipantOccupationAdministrative, Support, and Clerical  0.078534   0.099022
ParticipantOccupationAdvertising                         -0.009027   0.140001
ParticipantOccupationAerospace and Defense                0.138211   0.154910
ParticipantOccupationAgriculture, Forestry, and Fishing   0.068763   0.231398
ParticipantOccupationArchitecture                        -0.288507   0.161328
ParticipantOccupationArts and Entertainment              -0.055308   0.132776
ParticipantOccupationAviation and Airlines               -0.234719   0.188970
ParticipantOccupationBanking and Financial Services       0.086283   0.078241
ParticipantOccupationClergy                              -0.258654   0.231749
ParticipantOccupationConstruction and Landscaping         0.020352   0.094931
ParticipantOccupationConsulting                          -0.140033   0.078775
ParticipantOccupationEducation and Training              -0.141462   0.077450
ParticipantOccupationEngineering                         -0.096224   0.068598
ParticipantOccupationEnvironment                         -0.303991   0.149300
ParticipantOccupationExecutive/Management                -0.077905   0.074029
ParticipantOccupationFacilities, Maintenance, and Repair  0.066972   0.151963
ParticipantOccupationFire, Law Enforcement, and Security -0.076298   0.118635
ParticipantOccupationGovernment                          -0.152740   0.116765
ParticipantOccupationHealthcare                          -0.005703   0.068888
ParticipantOccupationHomemaking                          -0.101664   0.178034
ParticipantOccupationHotel, Gaming, Leisure, and Travel   0.021121   0.136280
ParticipantOccupationHuman Resources                      0.229668   0.108308
ParticipantOccupationInformation Technology (IT)          0.030527   0.069554
```

Fig 48a. Multiple regression for Fit 6 model

ParticipantOccupationInsurance	0.202909	0.105079
ParticipantOccupationLegal and Paralegal	-0.120604	0.091773
ParticipantOccupationManufacturing	-0.033819	0.105674
ParticipantOccupationMarketing	-0.093383	0.087345
ParticipantOccupationMedia	-0.170169	0.142206
ParticipantOccupationMilitary	-0.367362	0.169057
ParticipantOccupationNonprofit	-0.244982	0.116384
ParticipantOccupationOil and Gas	-0.160830	0.144433
ParticipantOccupationPersonal Care and Service	0.348340	0.154998
ParticipantOccupationPhotography	0.352041	0.306131
ParticipantOccupationProperty Management	0.134688	0.183354
ParticipantOccupationPsychology	-0.314113	0.260507
ParticipantOccupationPublishing	-0.363661	0.231502
ParticipantOccupationReal Estate, Rental, and Leasing	-0.310362	0.090564
ParticipantOccupationRestaurant and Food Services	-0.284017	0.128199
ParticipantOccupationRetail/Wholesale	-0.254554	0.114520
ParticipantOccupationSales	-0.098387	0.072954
ParticipantOccupationScience and Biotechnology	0.169414	0.103825
ParticipantOccupationSkilled Work and Trades	-0.157731	0.107628
ParticipantOccupationSocial Work	-0.161058	0.210833
ParticipantOccupationStock Broker/Investment Advisor	-0.109302	0.152198
ParticipantOccupationStudent	-0.077436	0.131161
ParticipantOccupationTechnical Account Manager	-0.625727	0.478273
ParticipantOccupationTelecommunications	-0.198568	0.131149
ParticipantOccupationTransportation and Warehousing	0.055075	0.116388
TeamDivision_1	0.337504	0.024287
TeamDivision_2	0.104352	0.103517
TeamDivision_3	0.083009	0.045474
TeamDivision_5	0.622011	0.155859
sqrt(TeamTotalConfirmed)	0.007053	0.000082

	t value	Pr(>|t|)	
(Intercept)	27.680	< 2e-16	***
ParticipantOccupationAdministrative, Support, and Clerical	0.793	0.427777	
ParticipantOccupationAdvertising	-0.064	0.948595	
ParticipantOccupationAerospace and Defense	0.892	0.372343	
ParticipantOccupationAgriculture, Forestry, and Fishing	0.297	0.766360	
ParticipantOccupationArchitecture	-1.788	0.073807	.
ParticipantOccupationArts and Entertainment	-0.417	0.677030	
ParticipantOccupationAviation and Airlines	-1.242	0.214280	
ParticipantOccupationBanking and Financial Services	1.103	0.270194	
ParticipantOccupationClergy	-1.116	0.264454	
ParticipantOccupationConstruction and Landscaping	0.214	0.830254	
ParticipantOccupationConsulting	-1.778	0.075547	.
ParticipantOccupationEducation and Training	-1.826	0.067857	.
ParticipantOccupationEngineering	-1.403	0.160783	
ParticipantOccupationEnvironment	-2.036	0.041811	*
ParticipantOccupationExecutive/Management	-1.052	0.292704	
ParticipantOccupationFacilities, Maintenance, and Repair	0.441	0.659449	
ParticipantOccupationFire, Law Enforcement, and Security	-0.643	0.520177	
ParticipantOccupationGovernment	-1.308	0.190921	
ParticipantOccupationHealthcare	-0.083	0.934023	
ParticipantOccupationHomemaking	-0.571	0.568008	
ParticipantOccupationHotel, Gaming, Leisure, and Travel	0.155	0.876845	
ParticipantOccupationHuman Resources	2.121	0.034031	*
ParticipantOccupationInformation Technology (IT)	0.439	0.660764	
ParticipantOccupationInsurance	1.931	0.053558	.
ParticipantOccupationLegal and Paralegal	-1.314	0.188876	
ParticipantOccupationManufacturing	-0.320	0.748965	
ParticipantOccupationMarketing	-1.069	0.285083	
ParticipantOccupationMedia	-1.197	0.231523	
ParticipantOccupationMilitary	-2.173	0.029843	*
ParticipantOccupationNonprofit	-2.105	0.035364	*
ParticipantOccupationOil and Gas	-1.114	0.265556	
ParticipantOccupationPersonal Care and Service	2.247	0.024675	*
ParticipantOccupationPhotography	1.150	0.250231	
ParticipantOccupationProperty Management	0.735	0.462645	
ParticipantOccupationPsychology	-1.206	0.227982	
ParticipantOccupationPublishing	-1.571	0.116298	

Fig 48b. Multiple regression for Fit 6 model

```
ParticipantOccupationReal Estate, Rental, and Leasing       -3.427 0.000617 ***
ParticipantOccupationRestaurant and Food Services           -2.215 0.026791 *
ParticipantOccupationRetail/Wholesale                       -2.223 0.026291 *
ParticipantOccupationSales                                  -1.349 0.177546
ParticipantOccupationScience and Biotechnology               1.632 0.102825
ParticipantOccupationSkilled Work and Trades                -1.466 0.142865
ParticipantOccupationSocial Work                            -0.764 0.444970
ParticipantOccupationStock Broker/Investment Advisor        -0.718 0.472706
ParticipantOccupationStudent                                -0.590 0.554965
ParticipantOccupationTechnical Account Manager              -1.308 0.190852
ParticipantOccupationTelecommunications                     -1.514 0.130096
ParticipantOccupationTransportation and Warehousing          0.473 0.636096
TeamDivision_1                                              13.896  < 2e-16 ***
TeamDivision_2                                               1.008 0.313488
TeamDivision_3                                               1.825 0.068018 .
TeamDivision_5                                               3.991 6.71e-05 ***
sqrt(TeamTotalConfirmed)                                    86.010  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6709 on 3658 degrees of freedom
Multiple R-squared:  0.7075,    Adjusted R-squared:  0.7033
F-statistic: 166.9 on 53 and 3658 DF,  p-value: < 2.2e-16
```

Fig 48c. Multiple regression for Fit 6 model

From the above screenshot, the R-squared and the adjusted R-squared value is again 70%. This is inline with our previous findings. Also, this model is significant which can be confirmed by the p-value of the F-statistics.

Conclusion

We have used two different modelling techniques, i.e., Linear Regression and Decision tree. We built 3 different models each for both the business questions and 1 decision tree for each of the business question.

We chose the Fit3 and Fit6 as our final models which gave us the best efficiency of 80% and 70% respectively for both the business questions as shown from the R-square values.

On the other hand, the 2 decision tree models, i.e., regression and classification tree gave us the most significant component for the target variables respectively. These details in-turn helped us to predict strongest involvement of the industries and to determine the relation between the predictor and the target variables as well.

References

Some of the references from where this project has been studied and datasets and extracts have been obtained are given below:

http://www.teradatauniversitynetwork.com/Community/Student-Competitions/2018/Data-Challenge/Datasets/

http://www.teradatauniversitynetwork.com/Community/Student-Competitions/2018/Data-Challenge/Business-Questions/

https://en.wikipedia.org/wiki/Principal_component_analysis