Drivers of sample number

Engagement team - Freshawater Hackathon

02 August, 2017

# Introduction

We want to know what the drivers of more samples is. We think this might be driven by things like the type of training they got, the amount and types of social engagement, length of involvement, etc.

# Sample effort analysis

The data were provided in .csv format by Ian Thornhill (Earthwatch) on the 19th June 2017 to the engagement team of the hackathon.

I read this in and have a look at the data fixing any of the columns that need to be fixed

rm(list = ls())  
library(lme4)

## Warning: package 'lme4' was built under R version 3.3.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 3.3.3

library(ggplot2)  
library(reshape2)  
library(glmm)

## Warning: package 'glmm' was built under R version 3.3.3

## Loading required package: trust

## Loading required package: mvtnorm

library(effects)

## Warning: package 'effects' was built under R version 3.3.3

## Clean data

raw\_data <- read.csv('data/HWPEngage\_5\_170802.csv', stringsAsFactors = FALSE)  
  
clean\_data <- raw\_data  
  
# First let's fix the date columns  
for(i in c('Date','created', 'Latest')){  
 clean\_data[,i] <- as.Date(as.character(raw\_data[,i]), format = '%d/%m/%Y')  
}   
  
# Make country a factor  
clean\_data$Country <- as.factor(clean\_data$Country)  
  
# Make particip a number (note the #NUM! become NA)  
clean\_data$Particip <- as.numeric(clean\_data$Particip)  
  
# Let's look for NAs  
nrow(clean\_data)

## [1] 7413

summary(clean\_data)

## uid Date Code   
## Min. : 179 Min. :2012-06-11 Length:7413   
## 1st Qu.: 4709 1st Qu.:2014-02-07 Class :character   
## Median : 9316 Median :2014-12-04 Mode :character   
## Mean : 9466 Mean :2014-12-04   
## 3rd Qu.:14229 3rd Qu.:2015-10-05   
## Max. :20313 Max. :2016-12-11   
## NA's :5106   
## Protocol Country TrainTeam Rainfall   
## Length:7413 India :1599 Min. : 3.00 Min. :0.00000   
## Class :character China :1392 1st Qu.:17.00 1st Qu.:0.00000   
## Mode :character UK : 925 Median :21.00 Median :0.00217   
## Brazil : 683 Mean :21.84 Mean :0.14414   
## USA : 523 3rd Qu.:26.00 3rd Qu.:0.18620   
## Mexico : 506 Max. :62.00 Max. :3.16876   
## (Other):1785   
## Temp Paid Team Self   
## Min. :-17.59 Min. :0.0000 Min. :0.000 Min. :0.0000   
## 1st Qu.: 16.16 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:0.0000   
## Median : 21.95 Median :1.0000 Median :1.000 Median :0.0000   
## Mean : 19.99 Mean :0.7798 Mean :0.811 Mean :0.3877   
## 3rd Qu.: 25.10 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000   
## Max. : 34.31 Max. :1.0000 Max. :1.000 Max. :1.0000   
##   
## People Assign TranTrai Coord   
## Min. : 0.000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 2.500 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median : 4.000 Median :1.000 Median :1.0000 Median :0.0000   
## Mean : 4.107 Mean :0.811 Mean :0.5357 Mean :0.2115   
## 3rd Qu.: 5.000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :10.000 Max. :1.000 Max. :1.0000 Max. :1.0000   
##   
## Upload Time Bulk Complex   
## Min. :0.0000 Min. : 1.000 Min. : 1.000 Min. : 1.000   
## 1st Qu.:0.0000 1st Qu.: 3.000 1st Qu.: 2.000 1st Qu.: 2.000   
## Median :1.0000 Median : 4.000 Median : 5.000 Median : 4.000   
## Mean :0.6762 Mean : 5.758 Mean : 5.261 Mean : 6.104   
## 3rd Qu.:1.0000 3rd Qu.: 9.000 3rd Qu.: 7.000 3rd Qu.: 9.000   
## Max. :1.0000 Max. :20.000 Max. :14.000 Max. :18.000   
##   
## Task WQO WQC WQM   
## Min. : 1.00 Min. :1.000 Min. :1.000 Min. :1.00   
## 1st Qu.: 3.00 1st Qu.:2.000 1st Qu.:1.500 1st Qu.:1.75   
## Median : 5.00 Median :3.000 Median :3.000 Median :2.50   
## Mean : 5.67 Mean :2.589 Mean :2.506 Mean :2.50   
## 3rd Qu.: 9.00 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:3.25   
## Max. :14.00 Max. :4.000 Max. :4.000 Max. :4.00   
## NA's :5910 NA's :5906 NA's :5917   
## MedInd MedTrain MedProt Participants2   
## Length:7413 Min. : 0.000 Min. :0.000 Min. : 0.500   
## Class :character 1st Qu.: 1.000 1st Qu.:1.000 1st Qu.: 1.000   
## Mode :character Median : 2.000 Median :1.000 Median : 2.000   
## Mean : 2.435 Mean :2.014 Mean : 3.098   
## 3rd Qu.: 2.000 3rd Qu.:2.000 3rd Qu.: 3.000   
## Max. :32.000 Max. :9.000 Max. :40.000   
## NA's :5903 NA's :5903 NA's :5903   
## created Latest ToNow   
## Min. :2013-02-28 Min. :2013-03-13 Min. : 180.0   
## 1st Qu.:2014-03-21 1st Qu.:2014-06-20 1st Qu.: 614.0   
## Median :2014-12-04 Median :2015-04-08 Median : 898.0   
## Mean :2014-12-03 Mean :2015-03-13 Mean : 898.4   
## 3rd Qu.:2015-09-14 3rd Qu.:2015-12-04 3rd Qu.:1156.0   
## Max. :2016-11-21 Max. :2017-05-09 Max. :1542.0   
##   
## ActivePeriod Years MaxSamp PeriodSamp   
## Min. : 0.00 Min. :0.0000 Min. :41416 Min. : 249.0   
## 1st Qu.: 0.00 1st Qu.:0.0000 1st Qu.:41877 1st Qu.: 736.5   
## Median : 8.00 Median :0.0200 Median :42133 Median : 989.0   
## Mean : 99.74 Mean :0.2734 Mean :42144 Mean : 975.4   
## 3rd Qu.: 111.00 3rd Qu.:0.3000 3rd Qu.:42423 3rd Qu.:1188.0   
## Max. :1522.00 Max. :4.1700 Max. :42903 Max. :1542.0   
## NA's :5903 NA's :5903   
## TeamPoints Blog Comment Invite   
## Min. :0.00000 Min. : 0.00000 Min. : 0.0000 Min. : 0.00000   
## 1st Qu.:0.00000 1st Qu.: 0.00000 1st Qu.: 0.0000 1st Qu.: 0.00000   
## Median :0.03791 Median : 0.00000 Median : 0.0000 Median : 0.00000   
## Mean :0.07712 Mean : 0.04708 Mean : 0.0549 Mean : 0.04087   
## 3rd Qu.:0.08537 3rd Qu.: 0.00000 3rd Qu.: 0.0000 3rd Qu.: 0.00000   
## Max. :1.57072 Max. :40.00000 Max. :48.0000 Max. :55.00000   
##   
## InviteAccepted Pres Quiz Sample   
## Min. : 0.000000 Min. : 0.00000 Min. :0.0000 Min. : 0.000   
## 1st Qu.: 0.000000 1st Qu.: 0.00000 1st Qu.:0.0000 1st Qu.: 0.000   
## Median : 0.000000 Median : 0.00000 Median :1.0000 Median : 0.000   
## Mean : 0.006205 Mean : 0.02091 Mean :0.7035 Mean : 1.211   
## 3rd Qu.: 0.000000 3rd Qu.: 0.00000 3rd Qu.:1.0000 3rd Qu.: 0.000   
## Max. :12.000000 Max. :15.00000 Max. :8.0000 Max. :617.000   
##   
## Share Points BlogTime   
## Min. : 0.000000 Min. : 0.00 Min. :0.0000000   
## 1st Qu.: 0.000000 1st Qu.: 0.00 1st Qu.:0.0000000   
## Median : 0.000000 Median : 5.00 Median :0.0000000   
## Mean : 0.004587 Mean : 35.14 Mean :0.0003065   
## 3rd Qu.: 0.000000 3rd Qu.: 20.00 3rd Qu.:0.0000000   
## Max. :13.000000 Max. :12740.00 Max. :0.2100525   
##   
## CommTime InvTime InviteATime   
## Min. :0.0000000 Min. :0.0000000 Min. :0.000e+00   
## 1st Qu.:0.0000000 1st Qu.:0.0000000 1st Qu.:0.000e+00   
## Median :0.0000000 Median :0.0000000 Median :0.000e+00   
## Mean :0.0003553 Mean :0.0002747 Mean :3.508e-05   
## 3rd Qu.:0.0000000 3rd Qu.:0.0000000 3rd Qu.:0.000e+00   
## Max. :0.2505593 Max. :0.2830882 Max. :6.176e-02   
##   
## PresTime QuizTime SampTime   
## Min. :0.0000000 Min. :0.000000 Min. :0.000000   
## 1st Qu.:0.0000000 1st Qu.:0.000000 1st Qu.:0.000000   
## Median :0.0000000 Median :0.005220 Median :0.000000   
## Mean :0.0001352 Mean :0.006105 Mean :0.008785   
## 3rd Qu.:0.0000000 3rd Qu.:0.010340 3rd Qu.:0.000000   
## Max. :0.0857843 Max. :0.085784 Max. :5.017799   
##   
## ShareTime PointsTime BlogTimeZ   
## Min. :0.000e+00 Min. : 0.00000 Min. :-0.06091   
## 1st Qu.:0.000e+00 1st Qu.: 0.00000 1st Qu.:-0.06091   
## Median :0.000e+00 Median : 0.03952 Median :-0.06091   
## Mean :4.224e-05 Mean : 0.26176 Mean : 0.00000   
## 3rd Qu.:0.000e+00 3rd Qu.: 0.17136 3rd Qu.:-0.06091   
## Max. :1.260e-01 Max. :112.07929 Max. :51.68738   
##   
## CommTimeZ InvTimeZ InviteATimeZ   
## Min. :-0.06577 Min. :-0.04963 Min. :-0.03572   
## 1st Qu.:-0.06577 1st Qu.:-0.04963 1st Qu.:-0.03572   
## Median :-0.06577 Median :-0.04963 Median :-0.03572   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.06577 3rd Qu.:-0.04963 3rd Qu.:-0.03572   
## Max. :57.43694 Max. :66.73746 Max. :69.03625   
##   
## PresTimeZ QuizTimeZ SampTimeZ ShareTimeZ   
## Min. :-0.06035 Min. :-0.8760 Min. :-0.1023 Min. :-0.02518   
## 1st Qu.:-0.06035 1st Qu.:-0.8760 1st Qu.:-0.1023 1st Qu.:-0.02518   
## Median :-0.06035 Median : 0.3692 Median :-0.1023 Median :-0.02518   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.:-0.06035 3rd Qu.: 0.3692 3rd Qu.:-0.1023 3rd Qu.:-0.02518   
## Max. :43.23316 Max. : 9.0856 Max. :51.9826 Max. :71.35558   
##   
## PointsTimeZ BlogTimeZZ Particip   
## Min. :-0.13870 Min. :-0.06833 Min. : 0.500   
## 1st Qu.:-0.13870 1st Qu.:-0.06833 1st Qu.: 1.000   
## Median :-0.11897 Median :-0.06833 Median : 2.000   
## Mean : 0.00000 Mean : 0.00000 Mean : 3.098   
## 3rd Qu.:-0.05976 3rd Qu.:-0.06833 3rd Qu.: 3.000   
## Max. :50.14692 Max. :46.75880 Max. :40.000   
## NA's :5903

# We need to remove people who never record a sample for the  
# sampling effort analysis, how many are there?  
sum(clean\_data$Sample == 0)

## [1] 5903

There are quite a few NAs but these don't appear to be in columns that we really need. They are in the Date column (not sure what this date is) and they are in the observed water quality colunms (WQx), which we dont use currently.

## Metrics

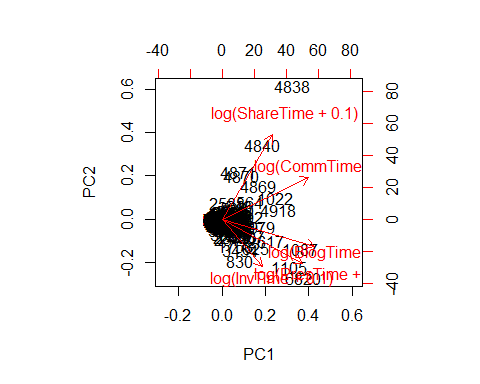
### Create social metrics

We created two different social metrics which quantified the user engagement on the website. The first metric is the communication score which is a combination of Blog, Comment, Invite, Presentation, and Share. To combine these we first devided the values (ie number of blog posts) by the number of weeks the user have been involved in the project, this removes the effect of time which will be in the model already. We then standardised these values so that they could be combined. This was done by Ian and the results are in the XTimeZ columns. I thought it would be a good idea to do a PCA to see if that creates a powerful PCA1.

PCA\_c <- prcomp(~ log(BlogTime+0.1) + log(CommTime+0.1) + log(InvTime+0.1) + log(ShareTime+0.1) + log(PresTime+0.1),  
 data = clean\_data,  
 center = TRUE,  
 scale. = TRUE)  
  
# The first axis explains 44% of the variance  
summary(PCA\_c)

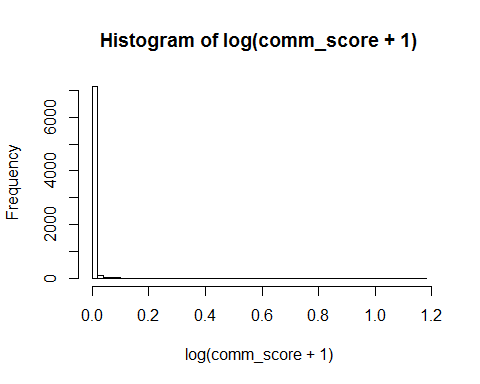
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 1.4861 1.0714 0.9366 0.70473 0.51943  
## Proportion of Variance 0.4417 0.2296 0.1754 0.09933 0.05396  
## Cumulative Proportion 0.4417 0.6713 0.8467 0.94604 1.00000

# This is because the values are highly correlated  
biplot(PCA\_c)



A PCA does not work here as there is not a strong correlation between the social factors. This does not stop us from summing across these values to get the total communication engagement

# These numbers are scaled to have a mean of 0 but actually I want them to have a minimum of 0  
# and max of 1  
  
range01 <- function(x, ...){(x - min(x, ...)) / (max(x, ...) - min(x, ...))}  
  
for(i in c('BlogTimeZ','CommTimeZ','InvTimeZ','ShareTimeZ','PresTimeZ')){  
   
 clean\_data[,i] <- range01(clean\_data[,i])  
 # hist(clean\_data[,i], breaks = 100)  
   
}  
  
comm\_score <- rowSums(clean\_data[,c('BlogTimeZ','CommTimeZ','InvTimeZ','ShareTimeZ','PresTimeZ')])  
  
clean\_data$comm\_score <- comm\_score  
  
# This metric has a big skew  
hist(log(comm\_score + 1), breaks = 50)



# We can turn in into a catagorical  
# Those with 0 (n= 1510), and then the lower and upper 50% for the rest (n=~215 each)  
comm\_cat <- comm\_score  
cat50 <- quantile(x = comm\_score[comm\_score > 0], probs = 0.5)  
comm\_cat[comm\_cat > cat50] <- 2  
comm\_cat[comm\_cat <= cat50 & comm\_cat > 0] <- 1  
table(comm\_cat)

## comm\_cat  
## 0 1 2   
## 7110 153 150

clean\_data$comm\_cat <- comm\_cat

### Sampling period

A note on sampling period. This is the time from the first activity in the project to the last sample collected and can be viewed as the sampling period. This was calculated by Ian.

### Country

We want to account for country in our model as it might be that people in one country sample more than another. This is not something we are interested in but is something we want to account for so we will keep it in the model as a random effect [this is no longer the case as I have to change to a model type that does not allow random effects, it is now included asa fixed effect]. The data was generated by Ian by taking the country that each user was trained in. Almost all users sample in a single country, which is the country that are trained in.

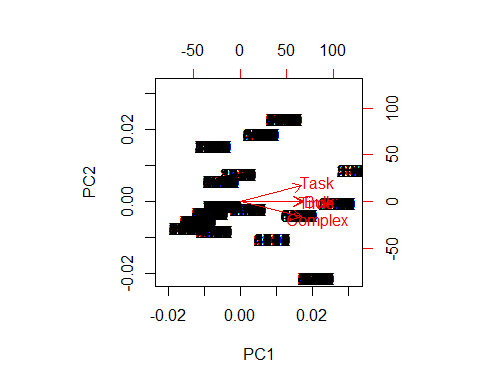
### Difficulty score

The difficulty score is the first axis of an PCA on a number of different metric that quantify the complexity of the tasks users do.

# Let's do a pca of the difficulty metrics   
PCA\_d <- prcomp(~ Time + Bulk + Complex + Task,  
 data = clean\_data,  
 center = TRUE,  
 scale. = TRUE)  
  
# The first axis explains 94% of the variance  
summary(PCA\_d)

## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 1.934 0.34125 0.32547 0.19417  
## Proportion of Variance 0.935 0.02911 0.02648 0.00943  
## Cumulative Proportion 0.935 0.96409 0.99057 1.00000

# This is because the values are highly correlated  
biplot(PCA\_d)



# add to our data  
clean\_data$difficulty\_score <- PCA\_d$x[,1]

It is important to note here that points are clustered together, this is because each protocol has a different difficulty but there are only 28 different protocols, and it look like some must share the same difficulty score. This is important to think about as the location and difficultly are likely to be correlated as a result.

### Group size

We are interested whether being in a larger group when you go out sampling means you are more likely to sample more due to a social factor. This is captured in the Particp variable.

### Looking at data for model

Staff have already been removed from this dataset (that is why there is no staff column this time).

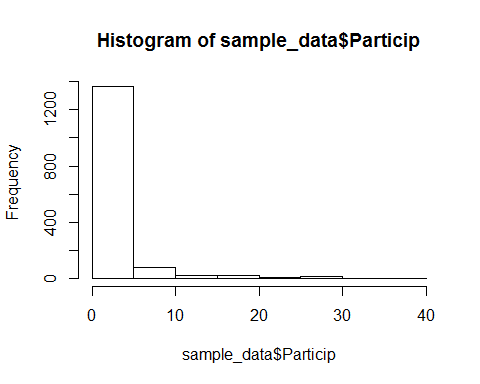
# At this point remove 0 samples  
sample\_data <- clean\_data[clean\_data$Sample > 0, ]  
  
# How much data do we have for the covariates of interest?  
nrow(sample\_data)

## [1] 1510

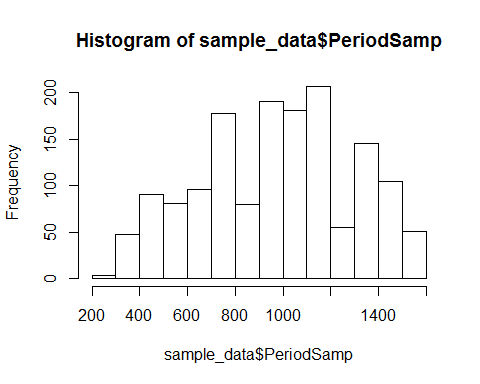
summary(sample\_data[, c('PeriodSamp', 'Paid', 'difficulty\_score', 'comm\_cat', 'Team', 'Rainfall', 'Temp', 'Country', 'Particip')])

## PeriodSamp Paid difficulty\_score comm\_cat   
## Min. : 249.0 Min. :0.0000 Min. :-2.4424 Min. :0.0000   
## 1st Qu.: 736.5 1st Qu.:1.0000 1st Qu.:-1.3928 1st Qu.:0.0000   
## Median : 989.0 Median :1.0000 Median :-0.8681 Median :0.0000   
## Mean : 975.4 Mean :0.8397 Mean :-0.1795 Mean :0.2344   
## 3rd Qu.:1188.0 3rd Qu.:1.0000 3rd Qu.: 0.9943 3rd Qu.:0.0000   
## Max. :1542.0 Max. :1.0000 Max. : 5.3241 Max. :2.0000   
##   
## Team Rainfall Temp Country   
## Min. :0.0000 Min. :0.000000 Min. :-17.59 Brazil :334   
## 1st Qu.:1.0000 1st Qu.:0.000000 1st Qu.: 15.18 India :195   
## Median :1.0000 Median :0.000925 Median : 20.96 UK :183   
## Mean :0.8589 Mean :0.132211 Mean : 19.50 China :177   
## 3rd Qu.:1.0000 3rd Qu.:0.170610 3rd Qu.: 24.98 Mexico :168   
## Max. :1.0000 Max. :3.168760 Max. : 34.31 Canada :153   
## (Other):300   
## Particip   
## Min. : 0.500   
## 1st Qu.: 1.000   
## Median : 2.000   
## Mean : 3.098   
## 3rd Qu.: 3.000   
## Max. :40.000   
##

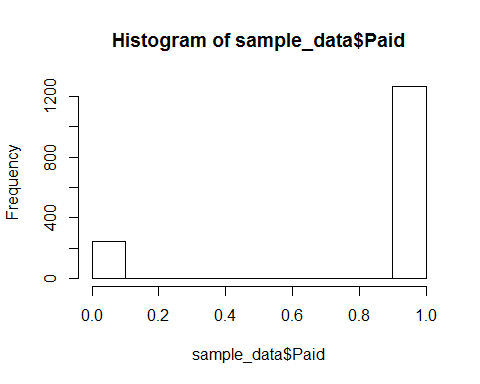
# Could we add in the WQx variables? there are onl a handfull of NAs  
  
# We might be best logging some of our predictor variables  
hist(sample\_data$Particip)



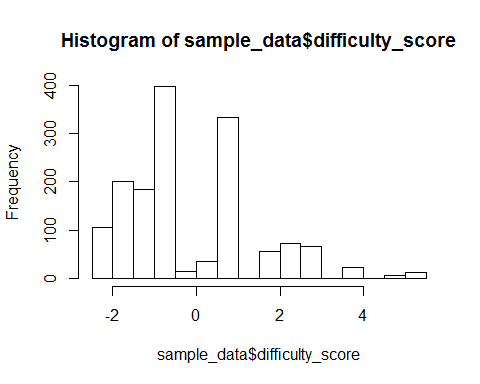
hist(sample\_data$PeriodSamp)



hist(sample\_data$Paid)



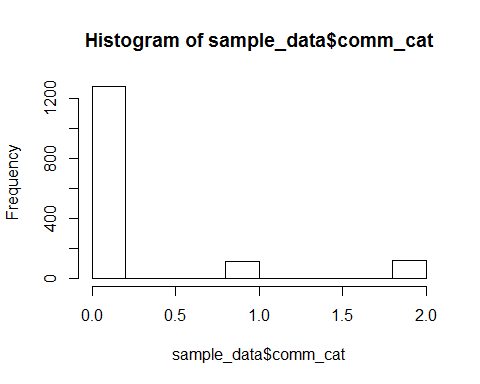
hist(sample\_data$difficulty\_score)



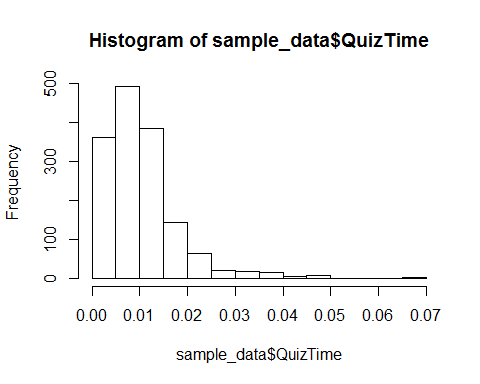
# hist(log(sample\_data$difficulty\_score))  
# hist(log(sample\_data$comm\_score + 1))  
table(sample\_data$comm\_cat)

##   
## 0 1 2   
## 1278 110 122

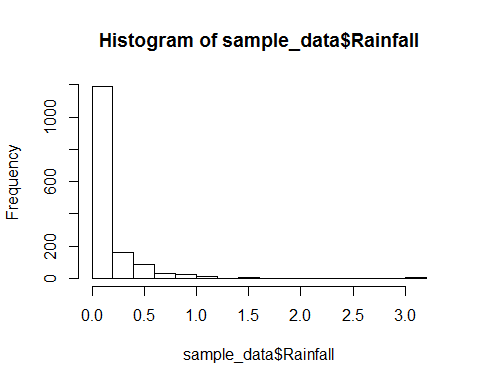
hist(sample\_data$comm\_cat)



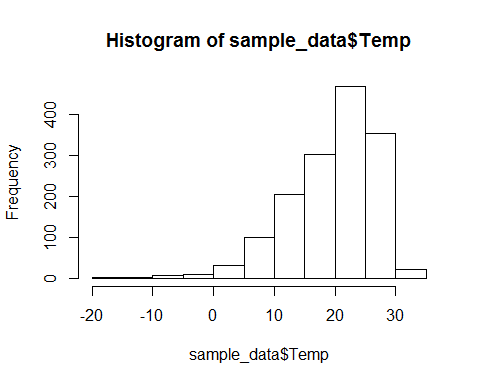
hist(sample\_data$QuizTime)



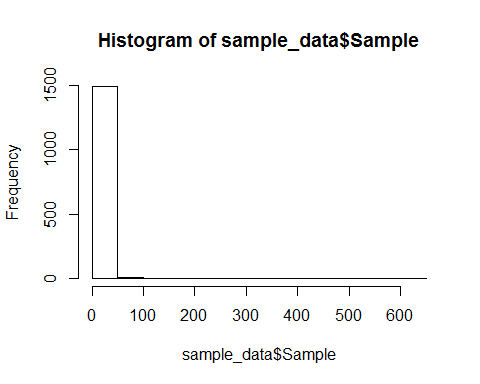
hist(sample\_data$Rainfall)



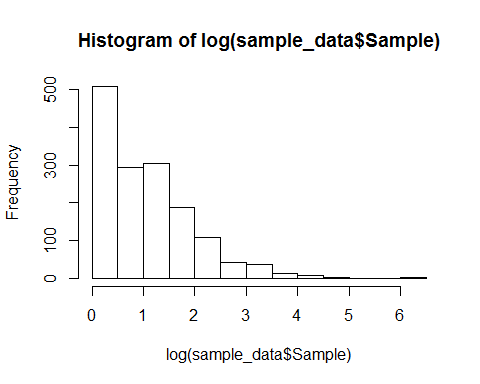
hist(sample\_data$Temp)



hist(sample\_data$Sample)



hist(log(sample\_data$Sample))



summary(sample\_data[,c('PeriodSamp','Paid','Particip','difficulty\_score','comm\_cat','QuizTime','Team','WQC', 'WQO', 'WQM')])

## PeriodSamp Paid Particip difficulty\_score   
## Min. : 249.0 Min. :0.0000 Min. : 0.500 Min. :-2.4424   
## 1st Qu.: 736.5 1st Qu.:1.0000 1st Qu.: 1.000 1st Qu.:-1.3928   
## Median : 989.0 Median :1.0000 Median : 2.000 Median :-0.8681   
## Mean : 975.4 Mean :0.8397 Mean : 3.098 Mean :-0.1795   
## 3rd Qu.:1188.0 3rd Qu.:1.0000 3rd Qu.: 3.000 3rd Qu.: 0.9943   
## Max. :1542.0 Max. :1.0000 Max. :40.000 Max. : 5.3241   
##   
## comm\_cat QuizTime Team WQC   
## Min. :0.0000 Min. :0.000000 Min. :0.0000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:0.005299 1st Qu.:1.0000 1st Qu.:1.250   
## Median :0.0000 Median :0.009315 Median :1.0000 Median :2.500   
## Mean :0.2344 Mean :0.009737 Mean :0.8589 Mean :2.505   
## 3rd Qu.:0.0000 3rd Qu.:0.012534 3rd Qu.:1.0000 3rd Qu.:4.000   
## Max. :2.0000 Max. :0.070000 Max. :1.0000 Max. :4.000   
## NA's :4   
## WQO WQM   
## Min. :1.00 Min. :1.0   
## 1st Qu.:2.00 1st Qu.:1.5   
## Median :3.00 Median :3.0   
## Mean :2.59 Mean :2.5   
## 3rd Qu.:4.00 3rd Qu.:3.5   
## Max. :4.00 Max. :4.0   
## NA's :8 NA's :15

cor(na.omit(sample\_data[,c('PeriodSamp','Paid','Particip','difficulty\_score','comm\_cat','QuizTime','Team','WQC', 'WQO', 'WQM')]))

## PeriodSamp Paid Particip difficulty\_score  
## PeriodSamp 1.000000000 -0.053104769 -0.188274772 0.01478680  
## Paid -0.053104769 1.000000000 -0.007982158 -0.03862182  
## Particip -0.188274772 -0.007982158 1.000000000 0.06128441  
## difficulty\_score 0.014786797 -0.038621819 0.061284412 1.00000000  
## comm\_cat 0.179471554 0.096428501 -0.048792301 -0.02682763  
## QuizTime -0.323329041 0.035822661 0.079071184 -0.04096701  
## Team -0.051823647 -0.173549634 0.104846989 0.44892355  
## WQC -0.001979659 -0.178573889 -0.050011734 -0.05028805  
## WQO -0.035026632 -0.055859539 0.111239201 0.04952178  
## WQM -0.068084386 -0.048692639 0.029747524 0.07157275  
## comm\_cat QuizTime Team WQC  
## PeriodSamp 0.17947155 -0.32332904 -0.05182365 -0.001979659  
## Paid 0.09642850 0.03582266 -0.17354963 -0.178573889  
## Particip -0.04879230 0.07907118 0.10484699 -0.050011734  
## difficulty\_score -0.02682763 -0.04096701 0.44892355 -0.050288052  
## comm\_cat 1.00000000 0.12364715 -0.14336792 -0.045773566  
## QuizTime 0.12364715 1.00000000 -0.09651699 -0.050273379  
## Team -0.14336792 -0.09651699 1.00000000 0.139827160  
## WQC -0.04577357 -0.05027338 0.13982716 1.000000000  
## WQO -0.04241388 -0.02511360 0.23937971 0.120672549  
## WQM -0.01727085 -0.04022076 0.10236787 0.153813770  
## WQO WQM  
## PeriodSamp -0.03502663 -0.06808439  
## Paid -0.05585954 -0.04869264  
## Particip 0.11123920 0.02974752  
## difficulty\_score 0.04952178 0.07157275  
## comm\_cat -0.04241388 -0.01727085  
## QuizTime -0.02511360 -0.04022076  
## Team 0.23937971 0.10236787  
## WQC 0.12067255 0.15381377  
## WQO 1.00000000 0.26853742  
## WQM 0.26853742 1.00000000

We can use a poisson distribution to account for the skew in the samples counts. We need to log the sample counts for this to work but this result in non-integer values which wont work with poission, instead we need to use quasi-poisson. In turn this means we cannot use a mixed effects model so we need to include country as a fixed efect.

m2a <- glm(log(Sample) ~ PeriodSamp +   
 Paid +  
 Particip +  
 difficulty\_score +  
 comm\_cat +  
 QuizTime +  
 Team +  
 WQC +  
 WQO +  
 WQM,  
 data = sample\_data)  
# m2aNB <- MASS::glm.nb(Sample ~ PeriodSamp +  
# Paid +  
# Particip +  
# difficulty\_score +  
# comm\_cat +  
# QuizTime +  
# Team +  
# WQC +  
# WQO +  
# WQM,  
# data = sample\_data[!rownames(sample\_data) %in% c('5480','298','182'),])  
# plot(m2aNB, ask = FALSE)  
# summary(m2aNB)  
# plot(m2a, ask = FALSE)  
# the three outliers without log(samples) are the three people with the most samples  
# negative binomial model does not account for these as well  
summary(m2a)

##   
## Call:  
## glm(formula = log(Sample) ~ PeriodSamp + Paid + Particip + difficulty\_score +   
## comm\_cat + QuizTime + Team + WQC + WQO + WQM, data = sample\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0372 -0.7028 -0.0570 0.5516 4.9389   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.374e-01 1.640e-01 4.495 7.49e-06 \*\*\*  
## PeriodSamp 6.137e-04 8.251e-05 7.438 1.72e-13 \*\*\*  
## Paid -4.217e-01 6.598e-02 -6.391 2.20e-10 \*\*\*  
## Particip -1.432e-02 5.689e-03 -2.518 0.01192 \*   
## difficulty\_score 5.262e-02 1.650e-02 3.190 0.00145 \*\*   
## comm\_cat 3.884e-01 4.193e-02 9.263 < 2e-16 \*\*\*  
## QuizTime 9.249e+00 3.069e+00 3.013 0.00263 \*\*   
## Team -9.882e-02 8.050e-02 -1.228 0.21983   
## WQC 3.458e-04 2.188e-02 0.016 0.98739   
## WQO -3.339e-02 2.264e-02 -1.475 0.14050   
## WQM 2.167e-02 2.209e-02 0.981 0.32676   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.8213251)  
##   
## Null deviance: 1430.0 on 1490 degrees of freedom  
## Residual deviance: 1215.6 on 1480 degrees of freedom  
## (19 observations deleted due to missingness)  
## AIC: 3950.8  
##   
## Number of Fisher Scoring iterations: 2

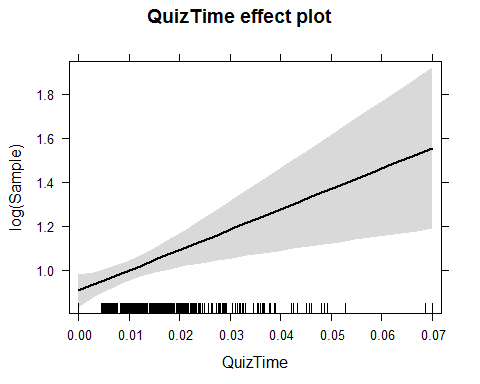
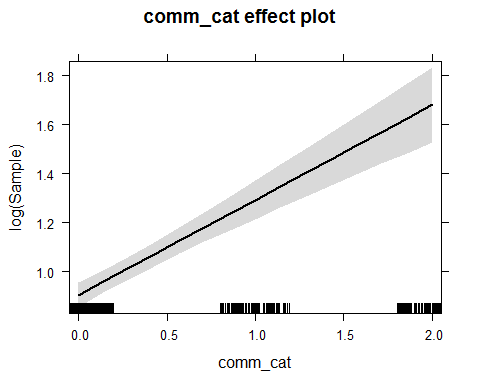
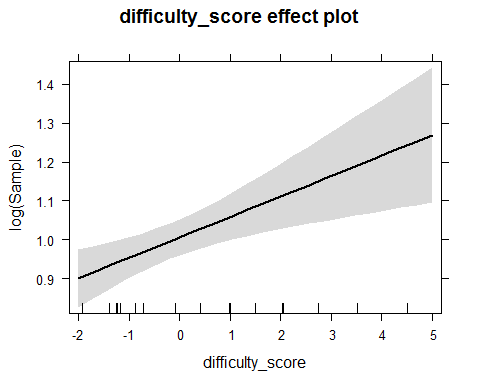
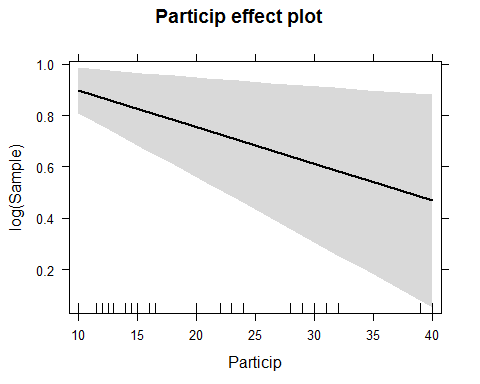
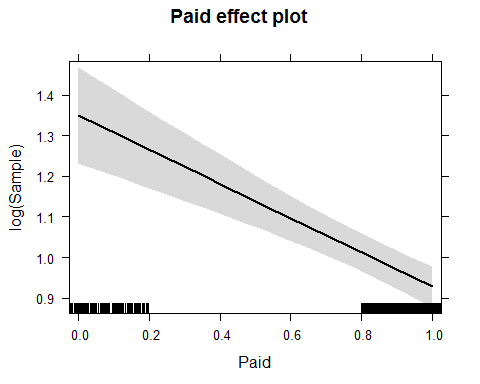
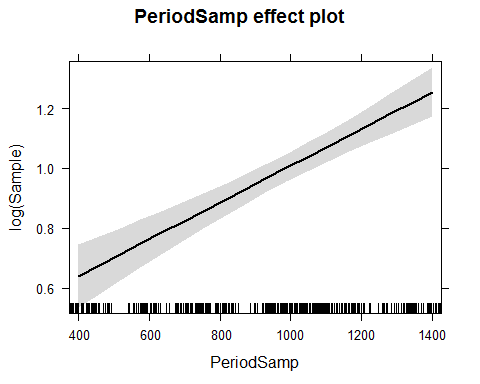
# I think that country correlates with the other variables  
# which might cause issues  
car::vif(m2a)

## PeriodSamp Paid Particip difficulty\_score   
## 1.234161 1.070771 1.061998 1.288022   
## comm\_cat QuizTime Team WQC   
## 1.098870 1.178305 1.434948 1.093217   
## WQO WQM   
## 1.152419 1.108712

# This all looks okay  
  
## I think this is the model we want to go with   
deviance\_explained <- (m2a$null.deviance - m2a$deviance) / m2a$null.deviance   
deviance\_explained

## [1] 0.1499403

# Let's plot the significant results.  
# The y-axis can be viewed as the increase in number of samples  
# get significant factors  
pvals <- summary(m2a)$coefficients[,'Pr(>|t|)']  
to\_plot <- names(tail(pvals, -1)[tail(pvals, -1) <= 0.05])  
for(i in to\_plot){  
 print(plot(Effect(i, m2a)))  
}



# Training retention model

We know that a lot of people who do the training do not go on to submit any records. Why is the case? are there some attibutes of the training that these people undertake, or the people themselves, which predicts retention? Here we use a binomial modelt to see which variables predict retention (0/1 - did they submit 0 records/did they submit >0 records). There is a 6 month period from the last training event to the date of data extraction, we consider this to be a long enough period for our binomial response to be accurate.

## Data

# The data we want is the same as above but we want to keep people with 0 records  
str(clean\_data)

## 'data.frame': 7413 obs. of 67 variables:  
## $ uid : int 3549 3550 3551 3552 3553 3554 3555 3557 444 3558 ...  
## $ Date : Date, format: "2012-09-10" "2012-09-10" ...  
## $ Code : chr "London41191" "London41191" "London41191" "London41191" ...  
## $ Protocol : chr "London" "London" "London" "London" ...  
## $ Country : Factor w/ 13 levels "Argentina","Australia",..: 12 12 12 12 12 12 12 12 12 12 ...  
## $ TrainTeam : int 8 11 8 8 8 8 8 8 11 11 ...  
## $ Rainfall : num 0.055 0.055 0.055 0.055 0.055 ...  
## $ Temp : num 10.6 10.6 10.6 10.6 10.6 ...  
## $ Paid : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Team : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Self : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ People : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Assign : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ TranTrai : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coord : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Upload : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Time : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ Bulk : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ Complex : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ Task : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ WQO : int NA NA NA NA NA NA 2 NA NA NA ...  
## $ WQC : int NA NA NA NA NA NA 4 NA NA NA ...  
## $ WQM : int NA NA NA NA NA NA 3 NA NA NA ...  
## $ MedInd : chr NA NA NA NA ...  
## $ MedTrain : num NA NA NA NA NA NA 2 NA NA NA ...  
## $ MedProt : num NA NA NA NA NA NA 1 NA NA NA ...  
## $ Participants2 : num NA NA NA NA NA NA 2.5 NA NA NA ...  
## $ created : Date, format: "2013-11-26" "2013-11-26" ...  
## $ Latest : Date, format: "2014-07-09" "2014-08-07" ...  
## $ ToNow : int 1271 1271 1271 1271 1271 1271 1271 1271 1524 1271 ...  
## $ ActivePeriod : int 225 254 0 0 0 0 744 0 374 115 ...  
## $ Years : num 0.62 0.7 0 0 0 0 2.04 0 1.02 0.32 ...  
## $ MaxSamp : int NA NA NA NA NA NA 42230 NA NA NA ...  
## $ PeriodSamp : int NA NA NA NA NA NA 1271 NA NA NA ...  
## $ TeamPoints : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Blog : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Comment : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ Invite : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ InviteAccepted : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Pres : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Quiz : int 1 2 0 0 0 0 0 0 0 0 ...  
## $ Sample : int 0 0 0 0 0 0 12 0 0 0 ...  
## $ Share : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Points : int 10 44 0 0 0 0 240 0 0 0 ...  
## $ BlogTime : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CommTime : num 0 0.00551 0 0 0 ...  
## $ InvTime : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ InviteATime : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PresTime : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ QuizTime : num 0.00551 0.01101 0 0 0 ...  
## $ SampTime : num 0 0 0 0 0 ...  
## $ ShareTime : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PointsTime : num 0.0551 0.2423 0 0 0 ...  
## $ BlogTimeZ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CommTimeZ : num 0 0.0208 0 0 0 ...  
## $ InvTimeZ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ InviteATimeZ : num -0.0357 -0.0357 -0.0357 -0.0357 -0.0357 ...  
## $ PresTimeZ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ QuizTimeZ : num 0.369 1.614 -0.876 -0.876 -0.876 ...  
## $ SampTimeZ : num -0.102 -0.102 -0.102 -0.102 -0.102 ...  
## $ ShareTimeZ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PointsTimeZ : num -0.0992 0.035 -0.1387 -0.1387 -0.1387 ...  
## $ BlogTimeZZ : num -0.0683 -0.0683 -0.0683 -0.0683 -0.0683 ...  
## $ Particip : num NA NA NA NA NA NA 2.5 NA NA NA ...  
## $ comm\_score : num 0 0.0208 0 0 0 ...  
## $ comm\_cat : num 0 1 0 0 0 0 0 0 0 0 ...  
## $ difficulty\_score: num -1.92 -1.92 -1.92 -1.92 -1.92 ...

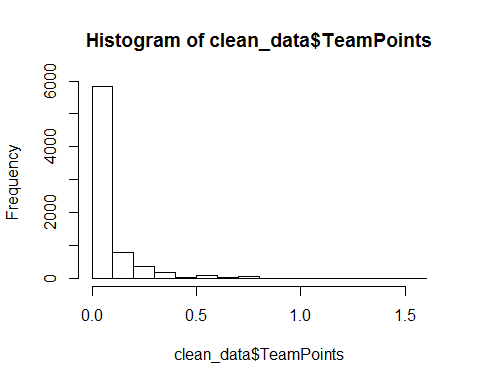
# Let's create our binary column  
clean\_data$Sample01 <- ifelse(clean\_data$Sample > 0, 1, 0)  
  
# Lets veiw the metrics we are going to use  
summary(clean\_data[,c('Paid','TranTrai','Upload','TrainTeam','TeamPoints','Team','Temp','Rainfall')])

## Paid TranTrai Upload TrainTeam   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. : 3.00   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:17.00   
## Median :1.0000 Median :1.0000 Median :1.0000 Median :21.00   
## Mean :0.7798 Mean :0.5357 Mean :0.6762 Mean :21.84   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:26.00   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :62.00   
## TeamPoints Team Temp Rainfall   
## Min. :0.00000 Min. :0.000 Min. :-17.59 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:1.000 1st Qu.: 16.16 1st Qu.:0.00000   
## Median :0.03791 Median :1.000 Median : 21.95 Median :0.00217   
## Mean :0.07712 Mean :0.811 Mean : 19.99 Mean :0.14414   
## 3rd Qu.:0.08537 3rd Qu.:1.000 3rd Qu.: 25.10 3rd Qu.:0.18620   
## Max. :1.57072 Max. :1.000 Max. : 34.31 Max. :3.16876

cor(clean\_data[,c('Paid','TranTrai','Upload','TrainTeam','TeamPoints','Team','Temp','Rainfall')])

## Paid TranTrai Upload TrainTeam TeamPoints  
## Paid 1.00000000 -0.47639003 0.35111697 -0.01044955 0.15001361  
## TranTrai -0.47639003 1.00000000 -0.06611250 0.29712439 -0.01523322  
## Upload 0.35111697 -0.06611250 1.00000000 0.02871120 0.07977444  
## TrainTeam -0.01044955 0.29712439 0.02871120 1.00000000 -0.13469169  
## TeamPoints 0.15001361 -0.01523322 0.07977444 -0.13469169 1.00000000  
## Team -0.25565708 0.51850637 -0.21030089 0.38872124 0.01158423  
## Temp -0.07430298 0.38731757 0.04853446 0.28293604 -0.07537546  
## Rainfall -0.09750582 0.17381448 0.01384540 0.12478629 0.03371263  
## Team Temp Rainfall  
## Paid -0.25565708 -0.07430298 -0.09750582  
## TranTrai 0.51850637 0.38731757 0.17381448  
## Upload -0.21030089 0.04853446 0.01384540  
## TrainTeam 0.38872124 0.28293604 0.12478629  
## TeamPoints 0.01158423 -0.07537546 0.03371263  
## Team 1.00000000 0.44946260 0.04889321  
## Temp 0.44946260 1.00000000 0.13325751  
## Rainfall 0.04889321 0.13325751 1.00000000

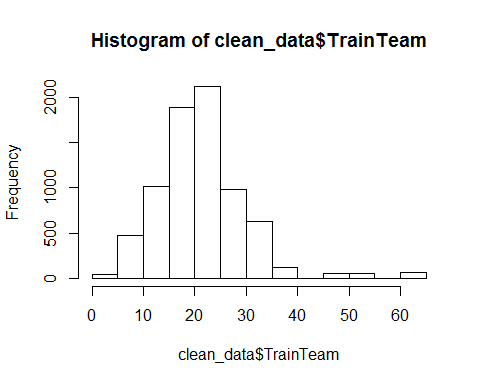
# Take a look at some of these  
hist(clean\_data$TeamPoints)



# A lot of people attend training where nobody has submitted any records  
sum(clean\_data$TeamPoints == 0)

## [1] 2656

hist(clean\_data$TrainTeam)



Now lets do some modelling

b1 <- glm(Sample01 ~ Paid +   
 TranTrai +  
 Upload +  
 TrainTeam +  
 log(TeamPoints+0.001) +  
 Team +  
 Temp +  
 Rainfall,  
 family = binomial(link = "logit"),  
 data = clean\_data)  
# plot(b1, ask = FALSE)  
summary(b1)

##   
## Call:  
## glm(formula = Sample01 ~ Paid + TranTrai + Upload + TrainTeam +   
## log(TeamPoints + 0.001) + Team + Temp + Rainfall, family = binomial(link = "logit"),   
## data = clean\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3535 -0.7244 -0.4929 -0.2964 2.6827   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.564159 0.145413 -3.880 0.000105 \*\*\*  
## Paid 0.085581 0.098844 0.866 0.386589   
## TranTrai -0.089048 0.080970 -1.100 0.271438   
## Upload 0.638058 0.075352 8.468 < 2e-16 \*\*\*  
## TrainTeam -0.031549 0.004594 -6.868 6.53e-12 \*\*\*  
## log(TeamPoints + 0.001) 0.335419 0.016525 20.298 < 2e-16 \*\*\*  
## Team 1.282210 0.108605 11.806 < 2e-16 \*\*\*  
## Temp -0.021100 0.005148 -4.099 4.16e-05 \*\*\*  
## Rainfall -0.140068 0.104687 -1.338 0.180906   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 7494.3 on 7412 degrees of freedom  
## Residual deviance: 6715.2 on 7404 degrees of freedom  
## AIC: 6733.2  
##   
## Number of Fisher Scoring iterations: 5

step(b1)

## Start: AIC=6733.2  
## Sample01 ~ Paid + TranTrai + Upload + TrainTeam + log(TeamPoints +   
## 0.001) + Team + Temp + Rainfall  
##   
## Df Deviance AIC  
## - Paid 1 6715.9 6731.9  
## - TranTrai 1 6716.4 6732.4  
## - Rainfall 1 6717.1 6733.1  
## <none> 6715.2 6733.2  
## - Temp 1 6731.7 6747.7  
## - TrainTeam 1 6765.0 6781.0  
## - Upload 1 6790.0 6806.0  
## - Team 1 6864.9 6880.9  
## - log(TeamPoints + 0.001) 1 7199.0 7215.0  
##   
## Step: AIC=6731.95  
## Sample01 ~ TranTrai + Upload + TrainTeam + log(TeamPoints + 0.001) +   
## Team + Temp + Rainfall  
##   
## Df Deviance AIC  
## - Rainfall 1 6717.9 6731.9  
## <none> 6715.9 6731.9  
## - TranTrai 1 6718.8 6732.8  
## - Temp 1 6731.9 6745.9  
## - TrainTeam 1 6765.3 6779.3  
## - Upload 1 6804.1 6818.1  
## - Team 1 6865.3 6879.3  
## - log(TeamPoints + 0.001) 1 7240.3 7254.3  
##   
## Step: AIC=6731.91  
## Sample01 ~ TranTrai + Upload + TrainTeam + log(TeamPoints + 0.001) +   
## Team + Temp  
##   
## Df Deviance AIC  
## <none> 6717.9 6731.9  
## - TranTrai 1 6721.7 6733.7  
## - Temp 1 6735.0 6747.0  
## - TrainTeam 1 6769.6 6781.6  
## - Upload 1 6807.5 6819.5  
## - Team 1 6872.7 6884.7  
## - log(TeamPoints + 0.001) 1 7240.4 7252.4

##   
## Call: glm(formula = Sample01 ~ TranTrai + Upload + TrainTeam + log(TeamPoints +   
## 0.001) + Team + Temp, family = binomial(link = "logit"),   
## data = clean\_data)  
##   
## Coefficients:  
## (Intercept) TranTrai Upload   
## -0.51119 -0.13874 0.66214   
## TrainTeam log(TeamPoints + 0.001) Team   
## -0.03139 0.33722 1.29655   
## Temp   
## -0.02127   
##   
## Degrees of Freedom: 7412 Total (i.e. Null); 7406 Residual  
## Null Deviance: 7494   
## Residual Deviance: 6718 AIC: 6732

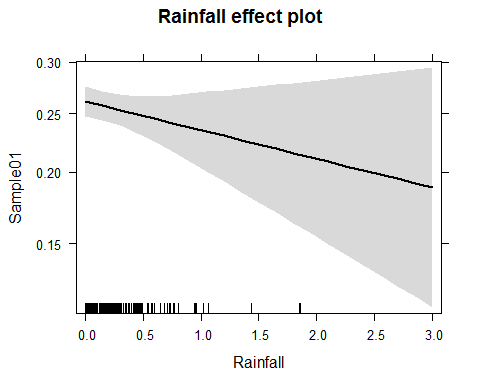
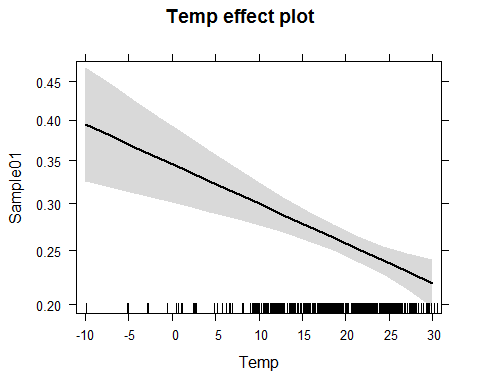
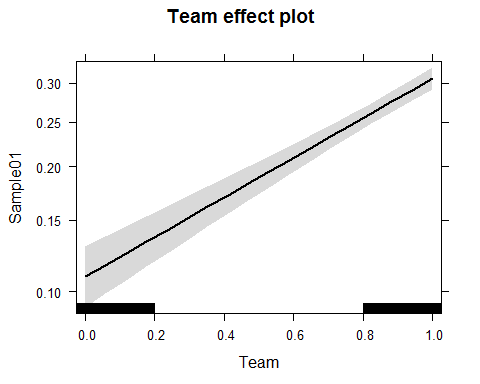
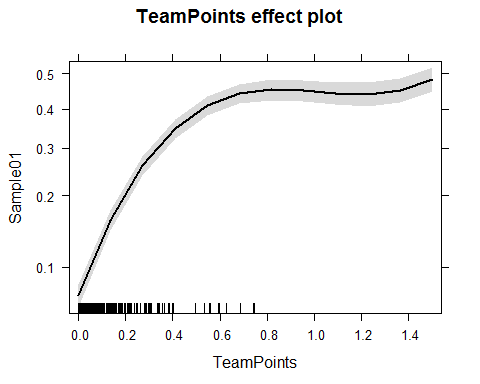
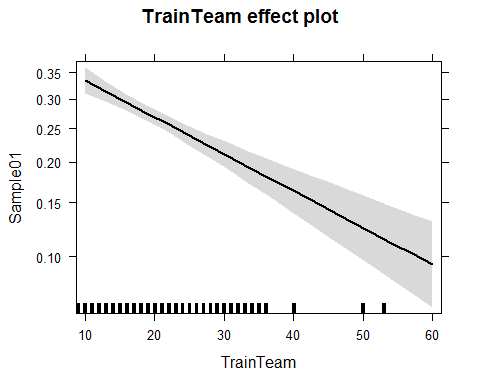
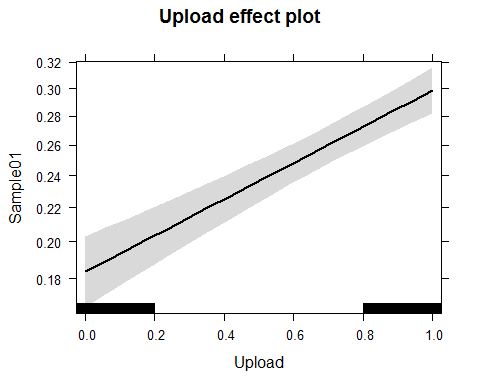
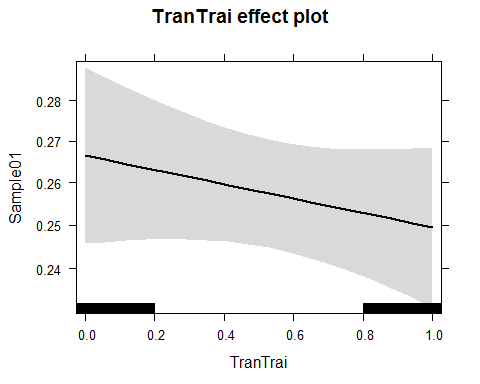
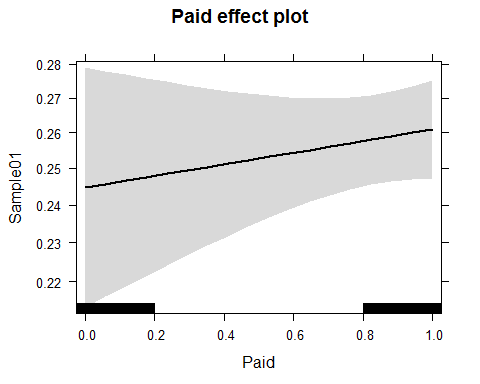
deviance\_explained <- (b1$null.deviance - b1$deviance) / b1$null.deviance   
deviance\_explained

## [1] 0.1039602

# Colinarity does not seem to be a big issue  
car::vif(b1)

## Paid TranTrai Upload   
## 1.497900 1.763880 1.244206   
## TrainTeam log(TeamPoints + 0.001) Team   
## 1.297501 1.087327 1.694864   
## Temp Rainfall   
## 1.409330 1.061657

# Let's plot the significant results.  
# The y-axis can be viewed as probability of retention after training  
for(i in c('Paid','TranTrai','Upload','TrainTeam','TeamPoints','Team','Temp','Rainfall')){  
 print(plot(Effect(i, b1)))  
}



The results here are essentially the same as those we got on the hack day.