Drivers of sample number

Engagement team - Freshawater Hackathon

20 June, 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.

# Data

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)

## Loading required package: Matrix

##   
## Attaching package: 'lme4'

## The following object is masked from 'package:stats':  
##   
## sigma

library(ggplot2)  
library(reshape2)  
# library(pscl)  
library(glmm)

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

## Loading required package: trust

## Loading required package: mvtnorm

library(sjPlot)

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

## Warning: Installed Rcpp (0.12.9) different from Rcpp used to build dplyr (0.12.11).  
## Please reinstall dplyr to avoid random crashes or undefined behavior.

# library(sjmisc)

## Clean data

raw\_data <- read.csv('data/HWPEngage\_2\_170619.csv', stringsAsFactors = FALSE)  
  
summary(raw\_data)

## uid TrainingDate Code Protocol   
## Min. : 179 Length:7413 Length:7413 Length:7413   
## 1st Qu.: 4709 Class :character Class :character Class :character   
## Median : 9316 Mode :character Mode :character Mode :character   
## Mean : 9466   
## 3rd Qu.:14229   
## Max. :20313   
##   
## Country TrainTeam Rainfall Temp   
## Length:7413 Min. : 3.00 Min. :0.0000 Min. :-17.59   
## Class :character 1st Qu.:17.00 1st Qu.:0.0000 1st Qu.: 16.16   
## Mode :character Median :21.00 Median :0.0000 Median : 21.95   
## Mean :21.84 Mean :0.1442 Mean : 19.99   
## 3rd Qu.:26.00 3rd Qu.:0.1900 3rd Qu.: 25.10   
## Max. :62.00 Max. :3.1700 Max. : 34.31   
##   
## Paid Team Self People   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:1.0000   
## Median :0.0000 Median :1.000 Median :0.0000 Median :1.0000   
## Mean :0.4106 Mean :0.811 Mean :0.3877 Mean :0.9773   
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :1.0000   
##   
## Assign TranTrai Coord Upload   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :1.000 Median :1.0000 Median :0.0000 Median :1.0000   
## Mean :0.811 Mean :0.5357 Mean :0.2115 Mean :0.6762   
## 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000   
## Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
##   
## Time Bulk Complex Task   
## Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 1.00   
## 1st Qu.: 3.000 1st Qu.: 2.000 1st Qu.: 2.000 1st Qu.: 3.00   
## Median : 4.000 Median : 5.000 Median : 4.000 Median : 5.00   
## Mean : 5.758 Mean : 5.261 Mean : 6.104 Mean : 5.67   
## 3rd Qu.: 9.000 3rd Qu.: 7.000 3rd Qu.: 9.000 3rd Qu.: 9.00   
## Max. :20.000 Max. :14.000 Max. :18.000 Max. :14.00   
##   
## WQO WQC WQM created   
## Min. :1.000 Min. :1.000 Min. :1.000 Length:7413   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000 Class :character   
## Median :3.000 Median :3.000 Median :3.000 Mode :character   
## Mean :2.589 Mean :2.589 Mean :2.589   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :4.000 Max. :4.000 Max. :4.000   
## NA's :5910 NA's :5910 NA's :5910   
## Latest ToNow ActivePeriod Years   
## Length:7413 Min. : 180.0 Min. : 0.00 Min. :0.0000   
## Class :character 1st Qu.: 614.0 1st Qu.: 0.00 1st Qu.:0.0000   
## Mode :character Median : 898.0 Median : 8.00 Median :0.0200   
## Mean : 898.4 Mean : 99.74 Mean :0.2734   
## 3rd Qu.:1156.0 3rd Qu.: 111.00 3rd Qu.:0.3000   
## Max. :1542.0 Max. :1522.00 Max. :4.1700   
##   
## MaxSamp PeriodSamp TeamPoints Blog   
## Length:7413 Min. : 249.0 Min. :0.00000 Min. : 0.00000   
## Class :character 1st Qu.: 736.5 1st Qu.:0.00000 1st Qu.: 0.00000   
## Mode :character Median : 989.0 Median :0.03791 Median : 0.00000   
## Mean : 975.4 Mean :0.07712 Mean : 0.04708   
## 3rd Qu.:1188.0 3rd Qu.:0.08537 3rd Qu.: 0.00000   
## Max. :1542.0 Max. :1.57072 Max. :40.00000   
## NA's :5903   
## Comment Invite InviteAccepted   
## Min. : 0.0000 Min. : 0.00000 Min. : 0.000000   
## 1st Qu.: 0.0000 1st Qu.: 0.00000 1st Qu.: 0.000000   
## Median : 0.0000 Median : 0.00000 Median : 0.000000   
## Mean : 0.0549 Mean : 0.04087 Mean : 0.006205   
## 3rd Qu.: 0.0000 3rd Qu.: 0.00000 3rd Qu.: 0.000000   
## Max. :48.0000 Max. :55.00000 Max. :12.000000   
##   
## Pres Quiz Sample Share   
## Min. : 0.00000 Min. :0.0000 Min. : 0.000 Min. : 0.000000   
## 1st Qu.: 0.00000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.: 0.000000   
## Median : 0.00000 Median :1.0000 Median : 0.000 Median : 0.000000   
## Mean : 0.02091 Mean :0.7035 Mean : 1.211 Mean : 0.004587   
## 3rd Qu.: 0.00000 3rd Qu.:1.0000 3rd Qu.: 0.000 3rd Qu.: 0.000000   
## Max. :15.00000 Max. :8.0000 Max. :617.000 Max. :13.000000   
##   
## Points BlogTime CommTime   
## Min. : 0.00 Min. :0.0000000 Min. :0.0000000   
## 1st Qu.: 0.00 1st Qu.:0.0000000 1st Qu.:0.0000000   
## Median : 5.00 Median :0.0000000 Median :0.0000000   
## Mean : 35.14 Mean :0.0003065 Mean :0.0003553   
## 3rd Qu.: 20.00 3rd Qu.:0.0000000 3rd Qu.:0.0000000   
## Max. :12740.00 Max. :0.2100500 Max. :0.2505600   
##   
## InvTime InviteATime PresTime   
## Min. :0.0000000 Min. :0.000e+00 Min. :0.0000000   
## 1st Qu.:0.0000000 1st Qu.:0.000e+00 1st Qu.:0.0000000   
## Median :0.0000000 Median :0.000e+00 Median :0.0000000   
## Mean :0.0002747 Mean :3.508e-05 Mean :0.0001352   
## 3rd Qu.:0.0000000 3rd Qu.:0.000e+00 3rd Qu.:0.0000000   
## Max. :0.2830900 Max. :6.176e-02 Max. :0.0857800   
##   
## QuizTime SampTime ShareTime   
## Min. :0.000000 Min. :0.000000 Min. :0.000e+00   
## 1st Qu.:0.000000 1st Qu.:0.000000 1st Qu.:0.000e+00   
## Median :0.005220 Median :0.000000 Median :0.000e+00   
## Mean :0.006105 Mean :0.008785 Mean :4.224e-05   
## 3rd Qu.:0.010340 3rd Qu.:0.000000 3rd Qu.:0.000e+00   
## Max. :0.085780 Max. :5.017800 Max. :1.260e-01   
##   
## PointsTime BlogTimeZ CommTimeZ   
## Min. : 0.00000 Min. :-0.06091 Min. :-0.06577   
## 1st Qu.: 0.00000 1st Qu.:-0.06091 1st Qu.:-0.06577   
## Median : 0.03952 Median :-0.06091 Median :-0.06577   
## Mean : 0.26176 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.17136 3rd Qu.:-0.06091 3rd Qu.:-0.06577   
## Max. :112.07929 Max. :51.68738 Max. :57.43694   
##   
## InvTimeZ InviteATimeZ PresTimeZ   
## Min. :-0.04963 Min. :-0.03572 Min. :-0.06035   
## 1st Qu.:-0.04963 1st Qu.:-0.03572 1st Qu.:-0.06035   
## Median :-0.04963 Median :-0.03572 Median :-0.06035   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.04963 3rd Qu.:-0.03572 3rd Qu.:-0.06035   
## Max. :66.73746 Max. :69.03625 Max. :43.23316   
##   
## QuizTimeZ SampTimeZ ShareTimeZ   
## Min. :-0.875990 Min. :-0.1023 Min. :-0.02518   
## 1st Qu.:-0.875990 1st Qu.:-0.1023 1st Qu.:-0.02518   
## Median : 0.369210 Median :-0.1023 Median :-0.02518   
## Mean : 0.000002 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.369210 3rd Qu.:-0.1023 3rd Qu.:-0.02518   
## Max. : 9.085650 Max. :51.9826 Max. :71.35558   
##   
## PointsTimeZ BlogTimeZZ   
## Min. :-0.13870 Min. :-0.06833   
## 1st Qu.:-0.13870 1st Qu.:-0.06833   
## Median :-0.11897 Median :-0.06833   
## Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.05976 3rd Qu.:-0.06833   
## Max. :50.14692 Max. :46.75880   
##

clean\_data <- raw\_data  
  
# First let's fix the date columns  
for(i in c('TrainingDate','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)  
  
# Let's look for NAs  
nrow(clean\_data)

## [1] 7413

summary(clean\_data)

## uid TrainingDate Code   
## Min. : 179 Min. :2012-10-09 Length:7413   
## 1st Qu.: 4709 1st Qu.:2014-04-08 Class :character   
## Median : 9316 Median :2014-12-12 Mode :character   
## Mean : 9466 Mean :2014-12-09   
## 3rd Qu.:14229 3rd Qu.:2015-09-12   
## Max. :20313 Max. :2016-11-29   
##   
## Protocol Country TrainTeam Rainfall   
## Length:7413 India :1599 Min. : 3.00 Min. :0.0000   
## Class :character China :1392 1st Qu.:17.00 1st Qu.:0.0000   
## Mode :character UK : 925 Median :21.00 Median :0.0000   
## Brazil : 683 Mean :21.84 Mean :0.1442   
## USA : 523 3rd Qu.:26.00 3rd Qu.:0.1900   
## Mexico : 506 Max. :62.00 Max. :3.1700   
## (Other):1785   
## Temp Paid Team Self   
## Min. :-17.59 Min. :0.0000 Min. :0.000 Min. :0.0000   
## 1st Qu.: 16.16 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:0.0000   
## Median : 21.95 Median :0.0000 Median :1.000 Median :0.0000   
## Mean : 19.99 Mean :0.4106 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. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :1.0000 Median :1.000 Median :1.0000 Median :0.0000   
## Mean :0.9773 Mean :0.811 Mean :0.5357 Mean :0.2115   
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 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. :1.000   
## 1st Qu.: 3.00 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:2.000   
## Median : 5.00 Median :3.000 Median :3.000 Median :3.000   
## Mean : 5.67 Mean :2.589 Mean :2.589 Mean :2.589   
## 3rd Qu.: 9.00 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :14.00 Max. :4.000 Max. :4.000 Max. :4.000   
## NA's :5910 NA's :5910 NA's :5910   
## 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 Length:7413 Min. : 249.0   
## 1st Qu.: 0.00 1st Qu.:0.0000 Class :character 1st Qu.: 736.5   
## Median : 8.00 Median :0.0200 Mode :character Median : 989.0   
## Mean : 99.74 Mean :0.2734 Mean : 975.4   
## 3rd Qu.: 111.00 3rd Qu.:0.3000 3rd Qu.:1188.0   
## Max. :1522.00 Max. :4.1700 Max. :1542.0   
## 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. :0.2100500   
##   
## 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. :0.2505600 Max. :0.2830900 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. :0.0857800 Max. :0.085780 Max. :5.017800   
##   
## 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   
## Min. :-0.06035 Min. :-0.875990 Min. :-0.1023   
## 1st Qu.:-0.06035 1st Qu.:-0.875990 1st Qu.:-0.1023   
## Median :-0.06035 Median : 0.369210 Median :-0.1023   
## Mean : 0.00000 Mean : 0.000002 Mean : 0.0000   
## 3rd Qu.:-0.06035 3rd Qu.: 0.369210 3rd Qu.:-0.1023   
## Max. :43.23316 Max. : 9.085650 Max. :51.9826   
##   
## ShareTimeZ PointsTimeZ BlogTimeZZ   
## Min. :-0.02518 Min. :-0.13870 Min. :-0.06833   
## 1st Qu.:-0.02518 1st Qu.:-0.13870 1st Qu.:-0.06833   
## Median :-0.02518 Median :-0.11897 Median :-0.06833   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.02518 3rd Qu.:-0.05976 3rd Qu.:-0.06833   
## Max. :71.35558 Max. :50.14692 Max. :46.75880   
##

# We need to remove people who never record a sample  
sum(clean\_data$Sample == 0)

## [1] 5903

clean\_data <- clean\_data[clean\_data$Sample > 0, ]

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

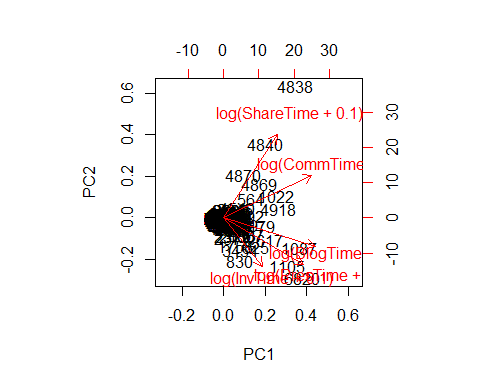
## Create social metrics

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 93% of the variance  
summary(PCA\_c)

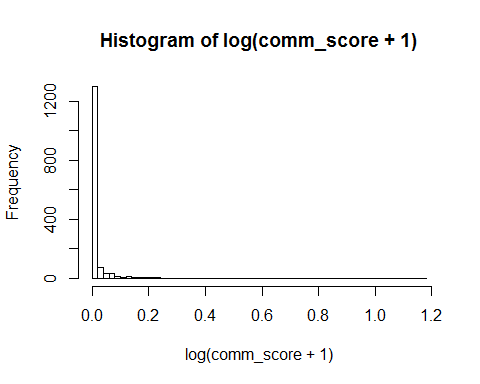
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 1.4972 1.0825 0.9352 0.68343 0.49502  
## Proportion of Variance 0.4483 0.2344 0.1749 0.09342 0.04901  
## Cumulative Proportion 0.4483 0.6827 0.8576 0.95099 1.00000

# 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   
## 1278 120 112

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. 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.

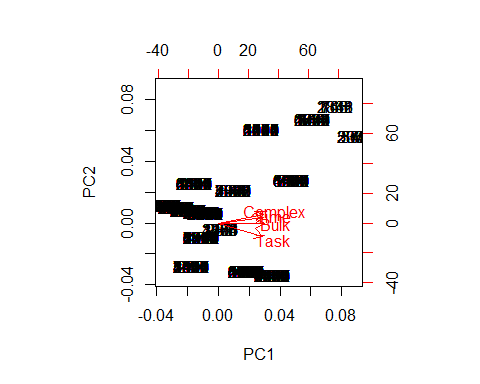
## Difficulty score

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 1.9383 0.37163 0.27037 0.1778  
## Proportion of Variance 0.9393 0.03453 0.01828 0.0079  
## Cumulative Proportion 0.9393 0.97382 0.99210 1.0000

# 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.

## Looking at data for model

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

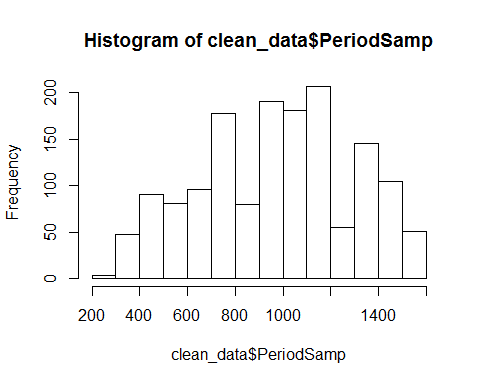
# How much data do we have for the covariates of interest?  
nrow(clean\_data)

## [1] 1510

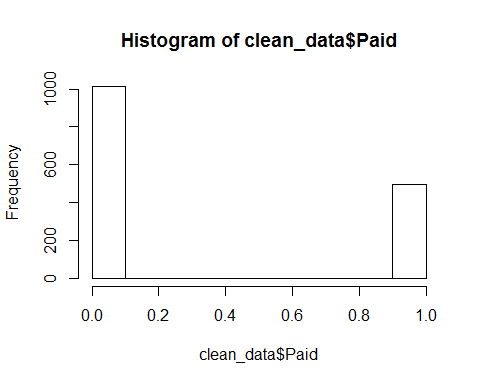
summary(clean\_data[, c('PeriodSamp', 'Paid', 'difficulty\_score', 'comm\_score', 'Team', 'Rainfall', 'Temp', 'Country')])

## PeriodSamp Paid difficulty\_score comm\_score   
## Min. : 249.0 Min. :0.0000 Min. :-2.7046 Min. :0.00000   
## 1st Qu.: 736.5 1st Qu.:0.0000 1st Qu.:-1.4489 1st Qu.:0.00000   
## Median : 989.0 Median :0.0000 Median :-0.8211 Median :0.00000   
## Mean : 975.4 Mean :0.3285 Mean : 0.0000 Mean :0.02068   
## 3rd Qu.:1188.0 3rd Qu.:1.0000 3rd Qu.: 1.3684 3rd Qu.:0.00000   
## Max. :1542.0 Max. :1.0000 Max. : 6.6811 Max. :2.24811   
##   
## Team Rainfall Temp Country   
## Min. :0.0000 Min. :0.0000 Min. :-17.59 Brazil :334   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.: 15.18 India :195   
## Median :1.0000 Median :0.0000 Median : 20.96 UK :183   
## Mean :0.8589 Mean :0.1323 Mean : 19.50 China :177   
## 3rd Qu.:1.0000 3rd Qu.:0.1700 3rd Qu.: 24.98 Mexico :168   
## Max. :1.0000 Max. :3.1700 Max. : 34.31 Canada :153   
## (Other):300

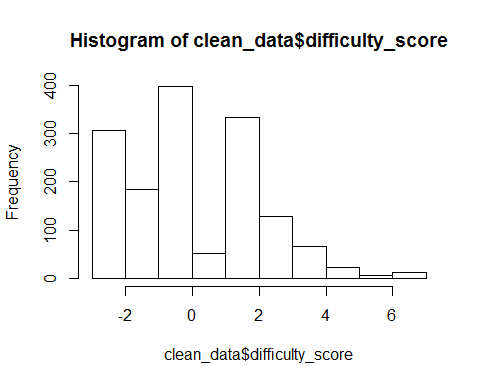
# We are going to lose lots of data through NAs in the Paid column  
# WQS would be great but there are so many NAs we just can't use it  
  
# We might be best logging some of our predictor variables  
hist(clean\_data$PeriodSamp)



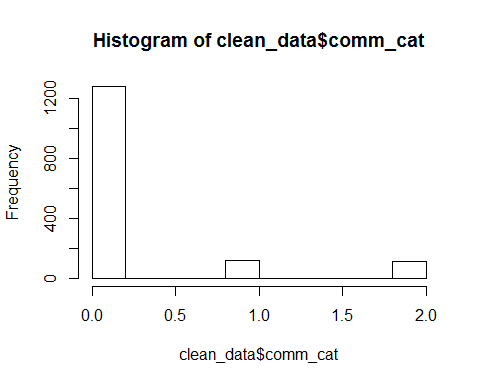
hist(clean\_data$Paid)



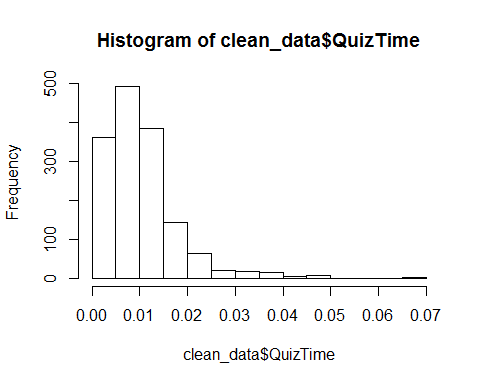
hist(clean\_data$difficulty\_score)



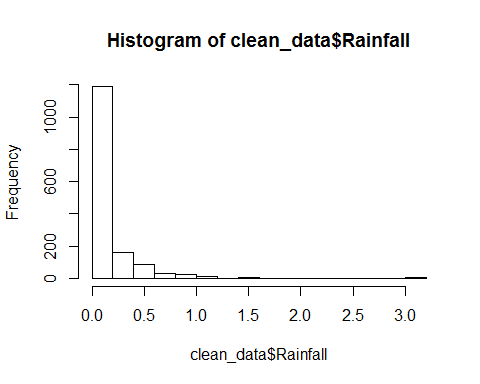
# hist(log(clean\_data$difficulty\_score))  
# hist(log(clean\_data$comm\_score + 1))  
hist(clean\_data$comm\_cat)



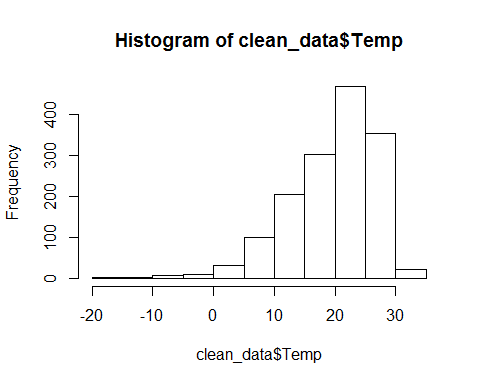
hist(clean\_data$QuizTime)



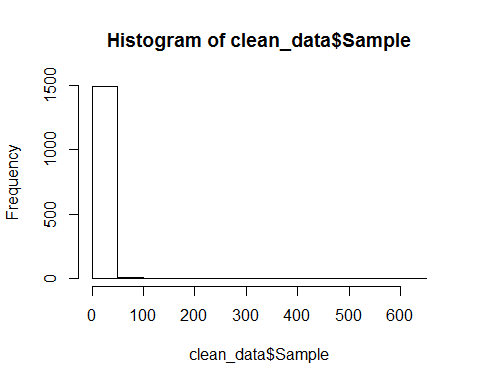
hist(clean\_data$Rainfall)



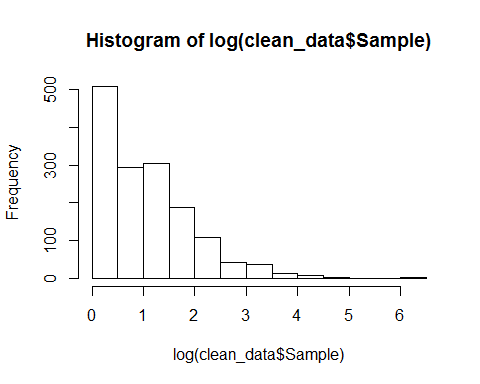
hist(clean\_data$Temp)



hist(clean\_data$Sample)

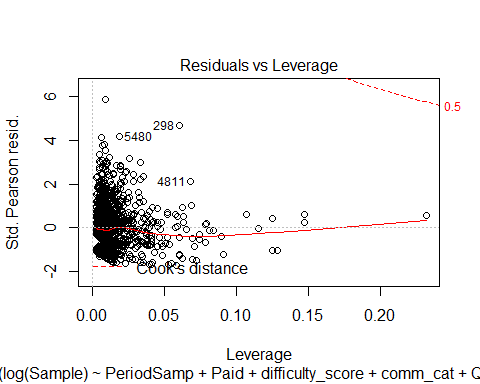
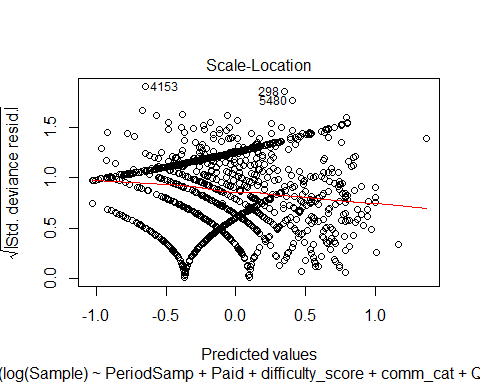
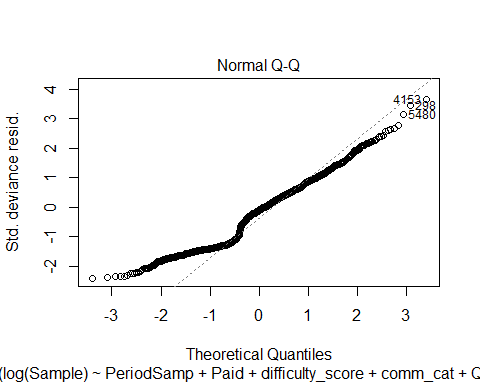
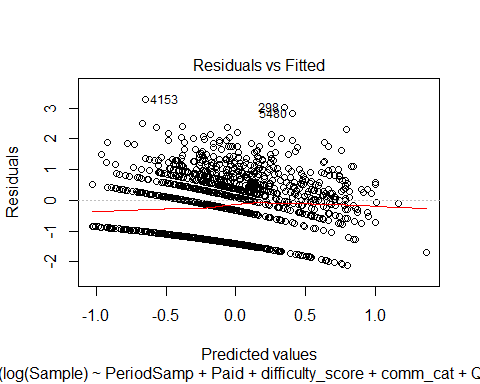


hist(log(clean\_data$Sample))



We can use a poisson distribution to account for the skew in the samples counts. The explanatory variables are in some cases improved by logging.

# # We need to move to a mixed effects model because we want to put country in there  
# m1 <- glmer(Sample ~ scale(PeriodSamp) +   
# Paid +   
# difficulty\_score +   
# comm\_cat +   
# QuizTime +   
# Team +  
# Rainfall +  
# Temp +  
# (1|Country),  
# data = clean\_data,  
# family = 'poisson')  
# plot(m1)  
# # I dont think this model copes with the serious overdispersion of the data  
#   
# # A standard glm with quasipoisson looks okay  
# m2 <- glm(log(Sample) ~ scale(PeriodSamp) +   
# Paid +   
# difficulty\_score +   
# comm\_cat +   
# QuizTime +   
# Team +  
# Rainfall +  
# Temp,   
# # (1|Country),  
# data = clean\_data,  
# family = 'quasipoisson')  
# plot(m2, ask = FALSE)  
# summary(m2)  
  
# lets just add in country as a fixed effect here  
m2a <- glm(log(Sample) ~ PeriodSamp +   
 Paid +  
 difficulty\_score +  
 comm\_cat +  
 QuizTime +  
 # Team + # this has no effct on the model but fixes UK issue  
 Rainfall +  
 Temp +  
 Country,  
 data = clean\_data,  
 family = 'quasipoisson')  
plot(m2a, ask = FALSE)



# the three outliers without log(samples) are the three people with the most samples  
summary(m2a)

##   
## Call:  
## glm(formula = log(Sample) ~ PeriodSamp + Paid + difficulty\_score +   
## comm\_cat + QuizTime + Rainfall + Temp + Country, family = "quasipoisson",   
## data = clean\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1069 -1.1460 -0.1029 0.4783 3.2729   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.594e-01 1.721e-01 -5.575 2.94e-08 \*\*\*  
## PeriodSamp 5.282e-04 8.593e-05 6.147 1.01e-09 \*\*\*  
## Paid -1.230e-01 8.338e-02 -1.475 0.140377   
## difficulty\_score 1.227e-01 2.041e-02 6.009 2.34e-09 \*\*\*  
## comm\_cat 3.320e-01 3.582e-02 9.269 < 2e-16 \*\*\*  
## QuizTime 1.015e+01 3.014e+00 3.368 0.000776 \*\*\*  
## Rainfall -2.444e-01 1.013e-01 -2.414 0.015891 \*   
## Temp -1.496e-02 4.314e-03 -3.468 0.000539 \*\*\*  
## CountryAustralia -3.444e-01 3.842e-01 -0.897 0.370098   
## CountryBrazil 4.541e-01 1.356e-01 3.348 0.000834 \*\*\*  
## CountryCanada 4.054e-01 1.610e-01 2.518 0.011913 \*   
## CountryChina 6.809e-01 1.367e-01 4.981 7.07e-07 \*\*\*  
## CountryFrance 3.340e-01 3.198e-01 1.044 0.296506   
## CountryIndia 6.350e-01 1.711e-01 3.712 0.000213 \*\*\*  
## CountryIndonesia 1.138e+00 2.347e-01 4.847 1.38e-06 \*\*\*  
## CountryMalaysia -2.120e-01 2.156e-01 -0.983 0.325540   
## CountryMexico 7.715e-01 1.627e-01 4.742 2.32e-06 \*\*\*  
## CountrySingapore 1.524e+00 2.161e-01 7.051 2.71e-12 \*\*\*  
## CountryUK 7.895e-01 1.962e-01 4.024 6.00e-05 \*\*\*  
## CountryUSA 3.436e-01 2.072e-01 1.659 0.097406 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasipoisson family taken to be 0.8151555)  
##   
## Null deviance: 1642.0 on 1509 degrees of freedom  
## Residual deviance: 1386.6 on 1490 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 5

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

## [1] 0.155516

# # removing the outliers doesn't really help  
# m2a2 <- glm(Sample ~ PeriodSamp +   
# Paid +  
# difficulty\_score +  
# comm\_cat +  
# QuizTime +  
# Team +  
# Rainfall +  
# Temp +  
# Country,  
# data = clean\_data[!row.names(clean\_data) %in% c('5480', '182', '298'),],  
# family = 'quasipoisson')  
# plot(m2a2, ask = FALSE)  
# # the three outliers without log(samples) are the three people with the most samples  
# summary(m2a2)  
#   
# # Might it be better explained with a negative binomial?  
# m2b <- MASS::glm.nb(Sample ~ PeriodSamp +   
# Paid +   
# difficulty\_score +   
# comm\_cat +   
# QuizTime +   
# Team +  
# Rainfall +  
# Temp +   
# Country,  
# data = clean\_data)  
# plot(m2b, ask = FALSE)  
# summary(m2b)  
#   
# # Might it be better explained with a negative binomial?  
# m3 <- MASS::glm.nb(Sample ~ scale(PeriodSamp) +   
# Paid +   
# difficulty\_score +   
# comm\_cat +   
# QuizTime +   
# Team +  
# Rainfall +  
# Temp,   
# # (1|Country),  
# data = clean\_data)  
# plot(m3, ask = FALSE)  
# summary(m3)  
# # this doesnt look so good  
#   
# # try with an observation level random effect  
# m4 <- glmer(Sample ~ scale(PeriodSamp) +   
# Paid +   
# difficulty\_score +   
# comm\_cat +   
# QuizTime +   
# Team +  
# Rainfall +  
# Temp +  
# (1|Country) +  
# (1|uid),  
# data = clean\_data,  
# family = 'poisson')  
# plot(m4)  
# summary(m4)  
# # this does not seem happy!