R Notebook

Research Question

1. What characteristics of the users is associated with a higher rate of improvement for the health risk factors.
2. Is a higher frequency of usage of the platform (measured through the accumulation of points) associated with a general improvement in health as measured by the risk factors.

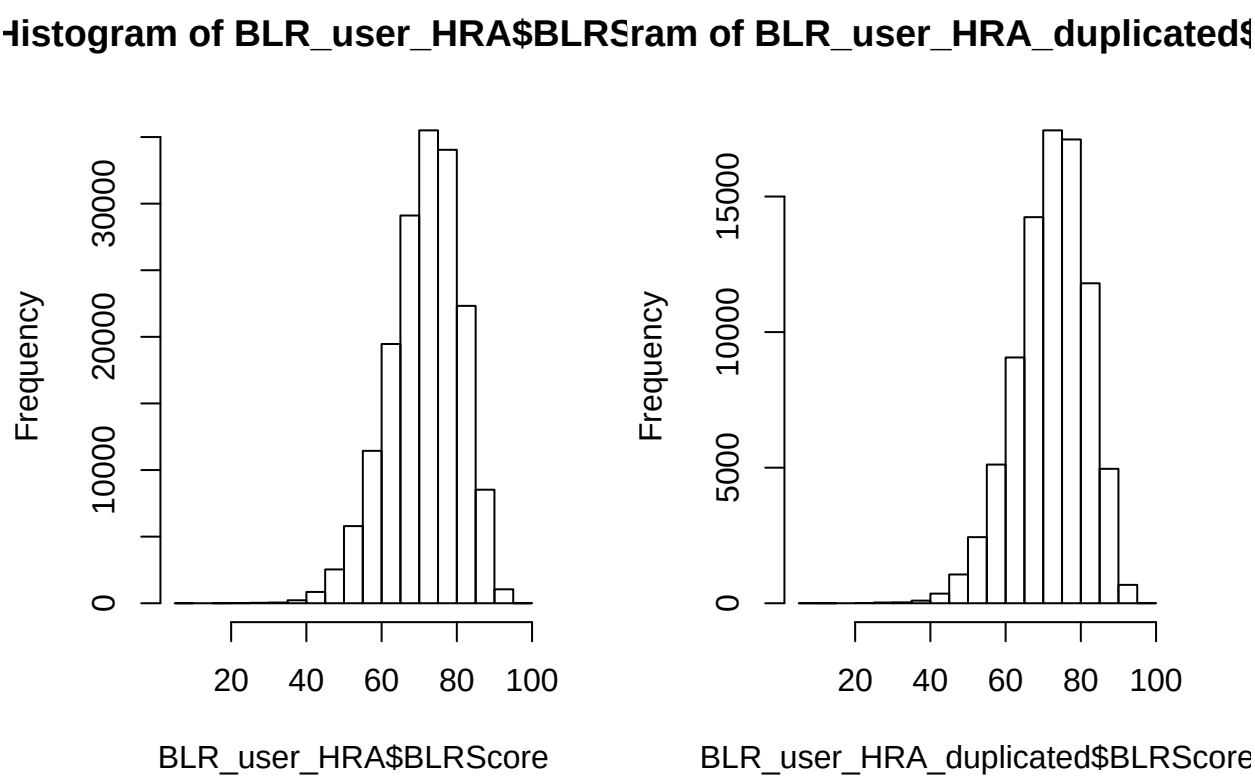
Higervalue of scores means better.

load data

```
BLR_user <- read.csv('BLR_USER.csv')
colnames(BLR_user)[1]<-"UserId"

BLR_user_HRA <- read.csv('BLR_USER_HRA.csv',na.strings = c('NA','NULL'))
colnames(BLR_user_HRA)[1]<-"UserId"
BLR_user_HRA[,1] <- as.character(unlist(BLR_user_HRA[,1]))

p1_points_M <- read.csv("part1- points M.csv")
colnames(p1_points_M)[1]<-"UserId"

p1_points_R <- read.csv("part1- points R.csv")
colnames(p1_points_R)[1]<-"UserId"

p1_points_Z <- read.csv("part1- points Z copy.csv")
colnames(p1_points_Z)[1]<-"UserId"

p2_points_M <- read.csv("part2- points M copy.csv")
colnames(p2_points_M)[1]<-"UserId"

p2_points_R <- read.csv("part2- points R copy.csv")
colnames(p2_points_R)[1]<-"UserId"

p3_points_M <- read.csv("part3- points M.csv")
colnames(p3_points_M)[1]<-"UserId"

p3_points_R <- read.csv("part3- points R copy.csv")
colnames(p3_points_R)[1]<-"UserId"


#concatenate points data with respect to companies
colnames(p1_points_M) <- c('userid','points','CreatedDate')
colnames(p2_points_M) <- c('userid','points','CreatedDate')
colnames(p3_points_M) <- c('userid','points','CreatedDate')
points_M <- rbind(rbind(p1_points_M,p2_points_M),p3_points_M)
rm(p1_points_M,p2_points_M,p3_points_M)
```

```r
colnames(p1_points_R) <- c('userid','points','CreatedDate')
colnames(p2_points_R) <- c('userid','points','CreatedDate')
colnames(p3_points_R) <- c('userid','points','CreatedDate')
points_R <- rbind(rbind(p1_points_R,p2_points_R),p3_points_R)
rm(p1_points_R,p2_points_R,p3_points_R)

colnames(p1_points_Z) <- c('userid','points','CreatedDate')
points_Z <- p1_points_Z
rm(p1_points_Z)
```

```r
#remove status and BLRorginalScore
BLR_user_HRA <- BLR_user_HRA[,-c(5,14)]

#remove rows with N/A BLR column
rmNA<-c()
for(i in 1:nrow(BLR_user_HRA)){
  if (is.na(BLR_user_HRA[i,][4])){
    rmNA <- c(rmNA,i)
  }
}

BLR_user_HRA <- BLR_user_HRA[-rmNA,]
BLR_user_HRA_duplicated <- subset(BLR_user_HRA, duplicated(UserId)|duplicated(UserId,fromLast = TRUE))

rm(rmNA)
rm(i)

#should we keep those user with only one entry?
#though we are interested in improvement, apparently not improvement data can be obtained from user wit
#so I only check the difference between original data from 2 sets.
par(mfrow=c(1,2))
hist(BLR_user_HRA$BLRScore)
hist(BLR_user_HRA_duplicated$BLRScore)
```



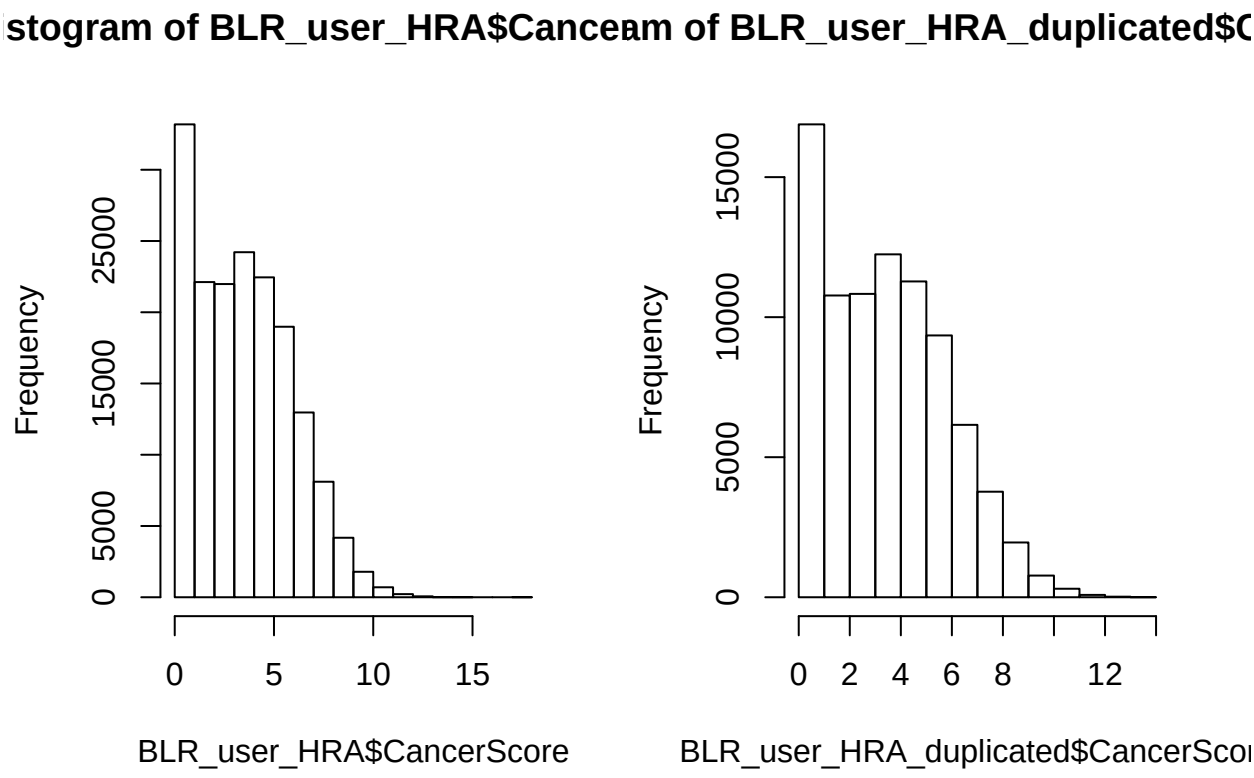

```
summary(BLR_user_HRA$BLRScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   66.00   73.00   71.91   79.00   96.00
```

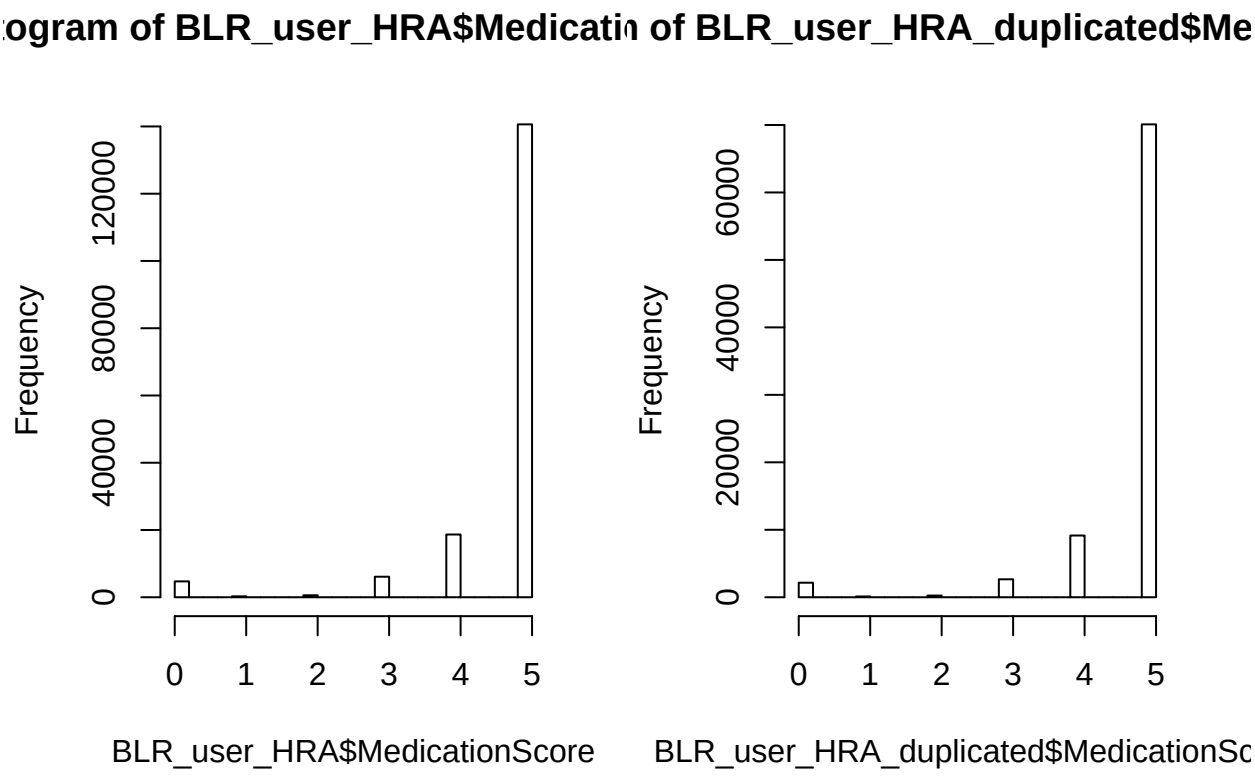
```
summary(BLR_user_HRA_duplicated$BLRScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   67.00   73.00   72.53   79.00   96.00
```

```
#BLR score distributions of data with non-duplicated user and without non-duplicated users are almost t

#randomly picked several other score to see difference between distrition of data with non-duplicated u
par(mfrow=c(1,2))
hist(BLR_user_HRA$CancerScore)
hist(BLR_user_HRA_duplicated$CancerScore)
```




```
par(mfrow=c(1,2))
hist(BLR_user_HRA$MedicationScore)
hist(BLR_user_HRA_duplicated$MedicationScore)
```




```
#distributions are all almost the same
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

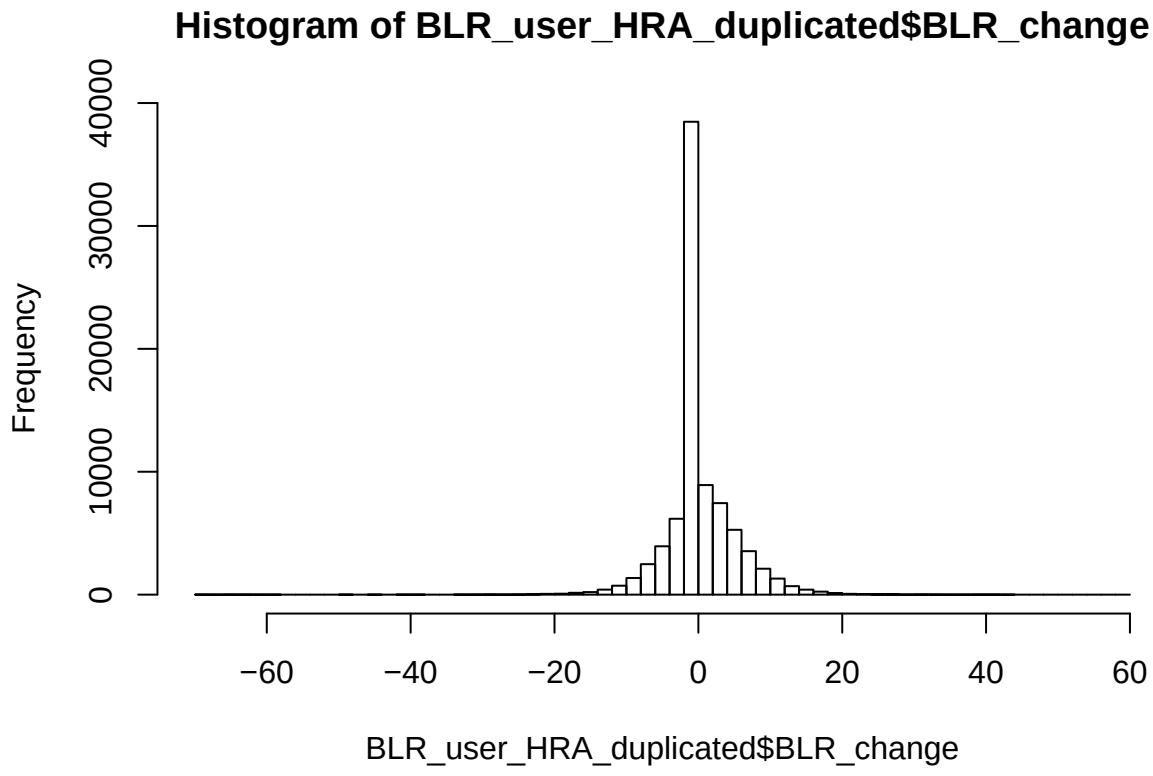
```
BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
   group_by(UserId) %>%
   mutate(BLR_change = BLRScore - lag(BLRScore, default = BLRScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
   group_by(UserId) %>%
   mutate(Heart_change = HeartScore - lag(HeartScore, default = HeartScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
```

```r
  group_by(UserId) %>%
  mutate(Respiratory_change = RespiratoryScore - lag(RespiratoryScore, default = RespiratoryScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(Gastrointestinal_change = GastrointestinalScore - lag(GastrointestinalScore,
                                                               default = GastrointestinalSco

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(Diabetes_change = DiabetesScore - lag(DiabetesScore, default = DiabetesScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(Cancer_change = CancerScore - lag(CancerScore, default = CancerScore[1]))


BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(ArthritisPain_change = ArthritisPainScore - lag(ArthritisPainScore, default = ArthritisPainSc

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(MentalHealth_change = MentalHealthScore - lag(MentalHealthScore, default = MentalHealthScore[

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(SocioFinancial_change = SocialFinancialRelationshipScore -
           lag(SocialFinancialRelationshipScore, default = SocialFinancialRelationshipScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(Diet_change = DietScore - lag(DietScore, default = DietScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(PhysicalActivity_change = PhysicalActivityScore - lag(PhysicalActivityScore,
                                                               default = PhysicalActivitySco

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(FinancialHealth_change = FinancialHealthScore - lag(FinancialHealthScore,
                                                             default = FinancialHealthScore[

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(Medication_change = MedicationScore - lag(MedicationScore, default = MedicationScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
  group_by(UserId) %>%
  mutate(Alcohol_change = AlcoholScore - lag(AlcoholScore, default = AlcoholScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
```

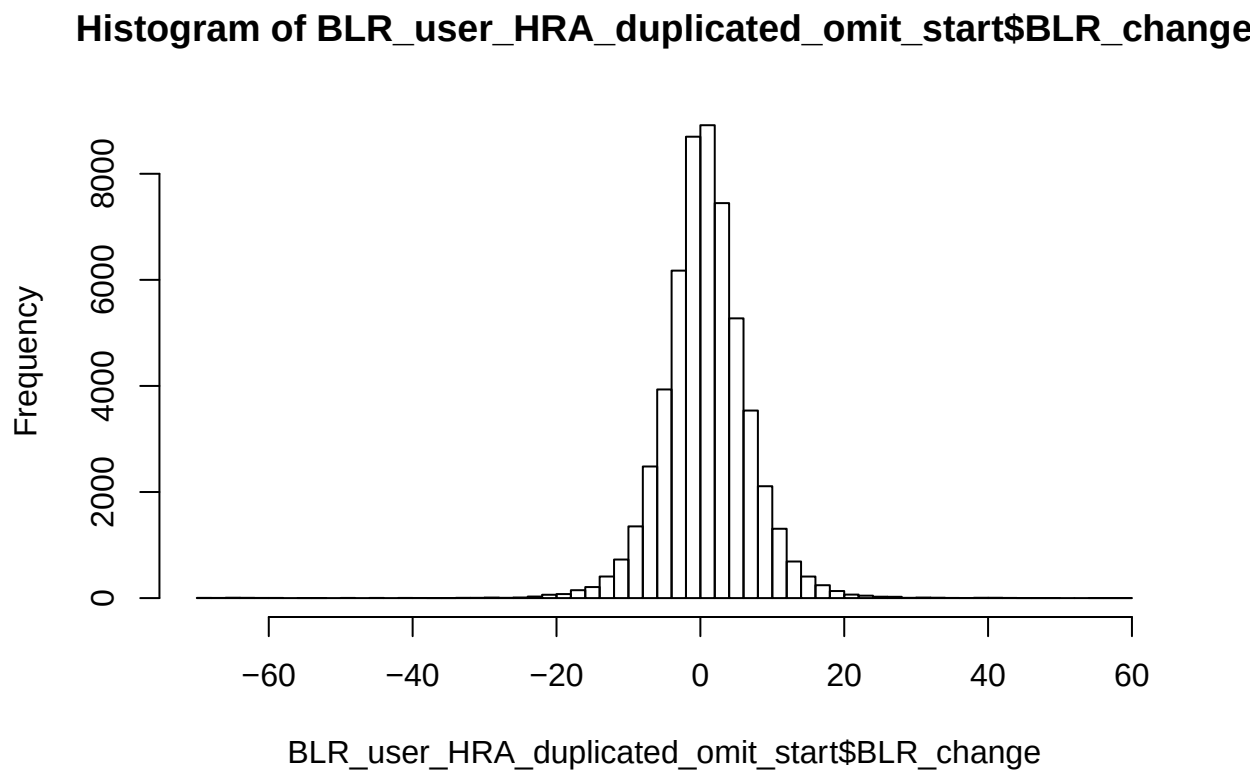
```
   group_by(UserId) %>%
   mutate(Sleep_change = SleepScore - lag(SleepScore, default = SleepScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
   group_by(UserId) %>%
   mutate(Stress_change = StressScore - lag(StressScore, default = StressScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
   group_by(UserId) %>%
   mutate(Smoking_change = SmokingScore - lag(SmokingScore, default = SmokingScore[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated %>%
   group_by(UserId) %>%
   mutate(BMI_change = BMI - lag(BMI, default = BMI[1]))

BLR_user_HRA_duplicated <- BLR_user_HRA_duplicated[order(BLR_user_HRA_duplicated$UserId),]
row.names(BLR_user_HRA_duplicated) <- c(1:nrow(BLR_user_HRA_duplicated))
```

```
## Warning: Setting row names on a tibble is deprecated.
```

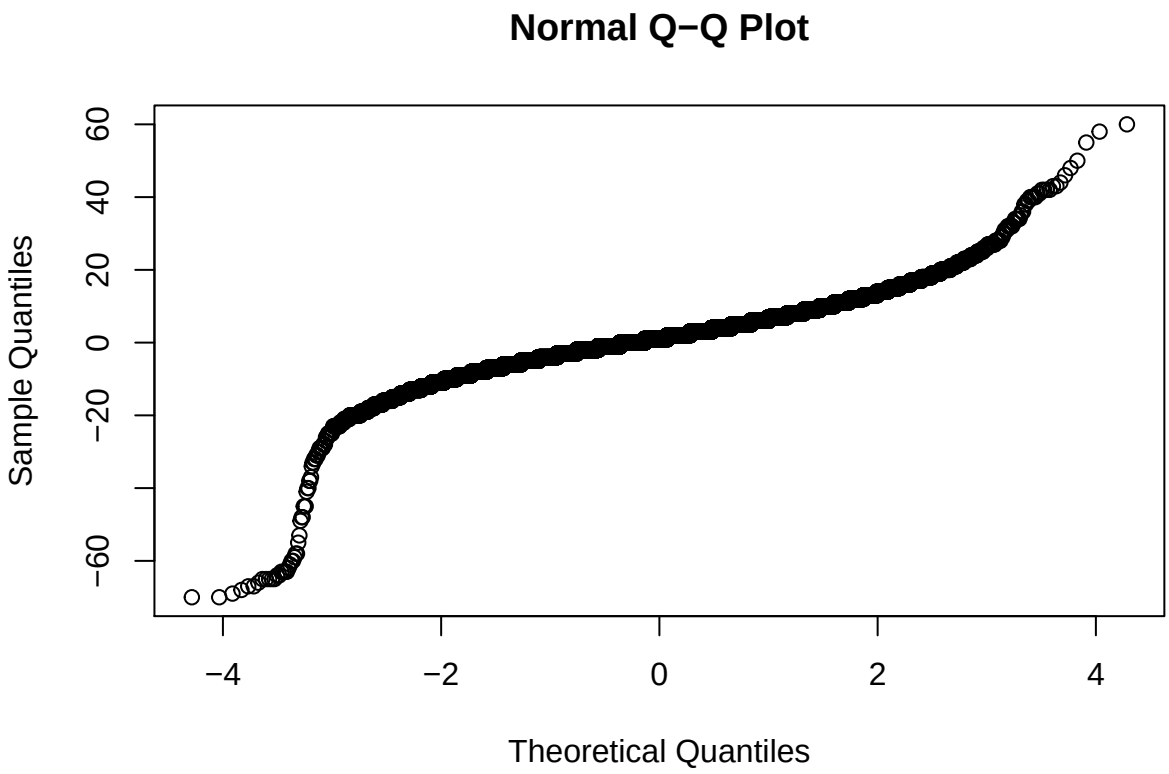
```
#plot a histgram to take a first look at the distribution of change in BLR
hist(BLR_user_HRA_duplicated$BLR_change,breaks = 50)
```



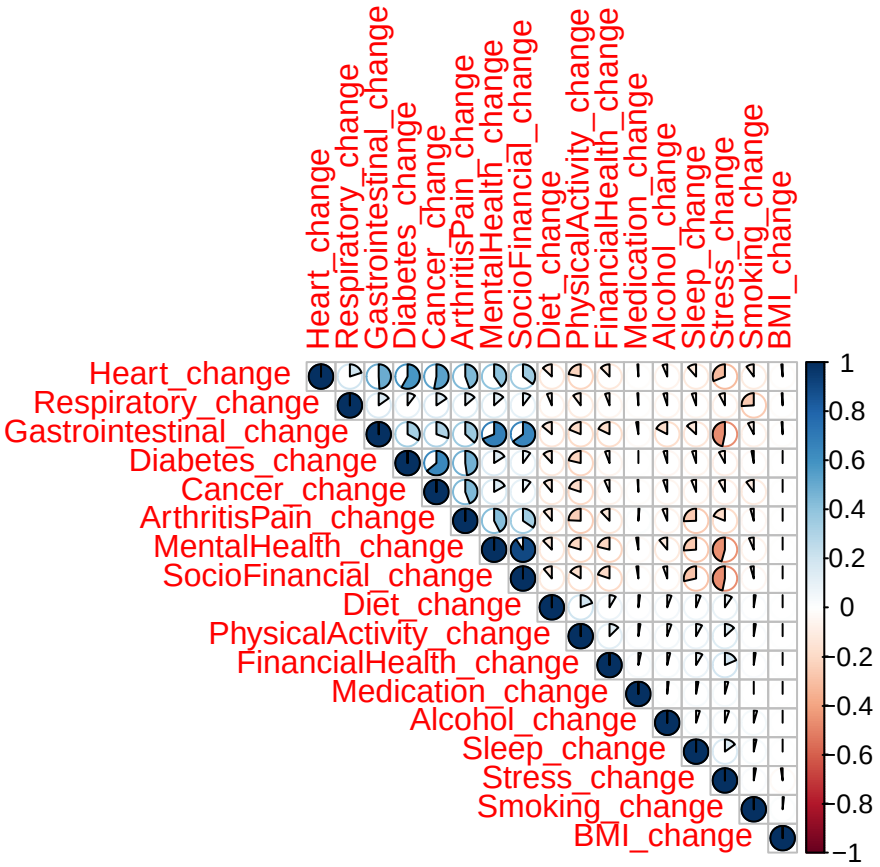
```r
#Looks close to normal but there is a single long bar, which is 0 change
#obviously, it's the starting entry for each user
#the starting entry will always have 0s in all columns represent change value

#create a new dataframe with those so-called starting entry omitted
num_row <- nrow(BLR_user_HRA_duplicated)
zeros <- rep(c(0),times = 17)
start_row <- c(1)
j=2
for(i in 2:num_row){
   #can do so here since I have ordered the dataframe by user id in previous code chunk
   if(all(BLR_user_HRA_duplicated[i,26:42]== zeros,na.rm=TRUE)&
      BLR_user_HRA_duplicated$UserId[i] != BLR_user_HRA_duplicated$UserId[i-1]){
      start_row[j] <- i
      j <- j+1
   }
}
BLR_user_HRA_duplicated_omit_start <- BLR_user_HRA_duplicated[-start_row,]

#plot a new histgram
hist(BLR_user_HRA_duplicated_omit_start$BLR_change,breaks = 50)
```



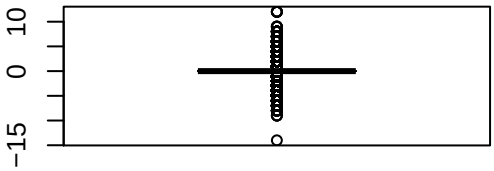

```r
qqnorm(BLR_user_HRA_duplicated_omit_start$BLR_change)
```

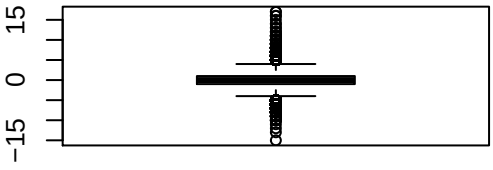


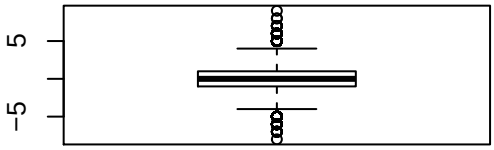

```r
#and correlation plot of predictors
corr <- cor(BLR_user_HRA_duplicated_omit_start[26:42],use = "complete.obs")
corrplot::corrplot(corr,method = 'pie',type='upper')
```

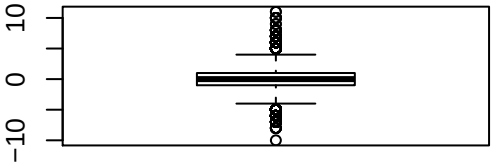


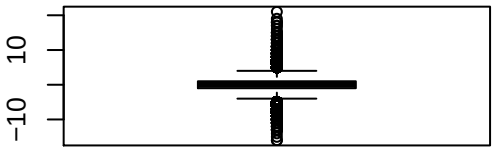

```
par(mfrow = c(2,2))
for(i in 27:42){
   boxplot(BLR_user_HRA_duplicated_omit_start[i])
}
```

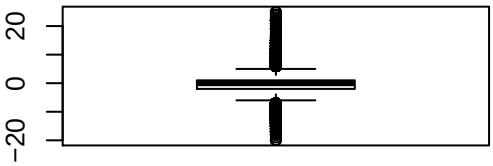

−15
0
10

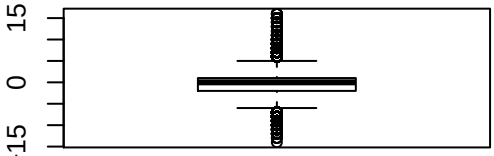

−15
0
15

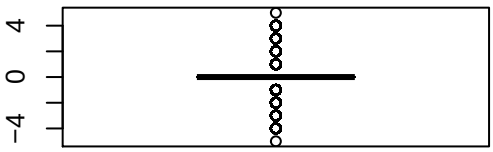

−5
5

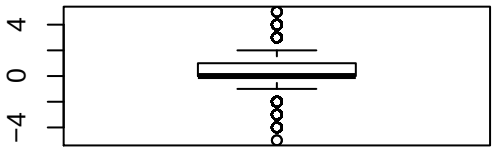

−10
0
10

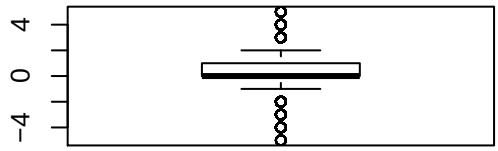

−10
10

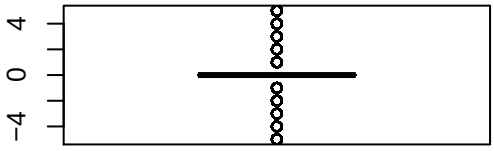

−20
0
20

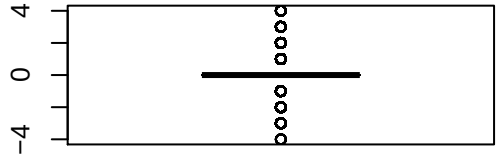

−15
0
15

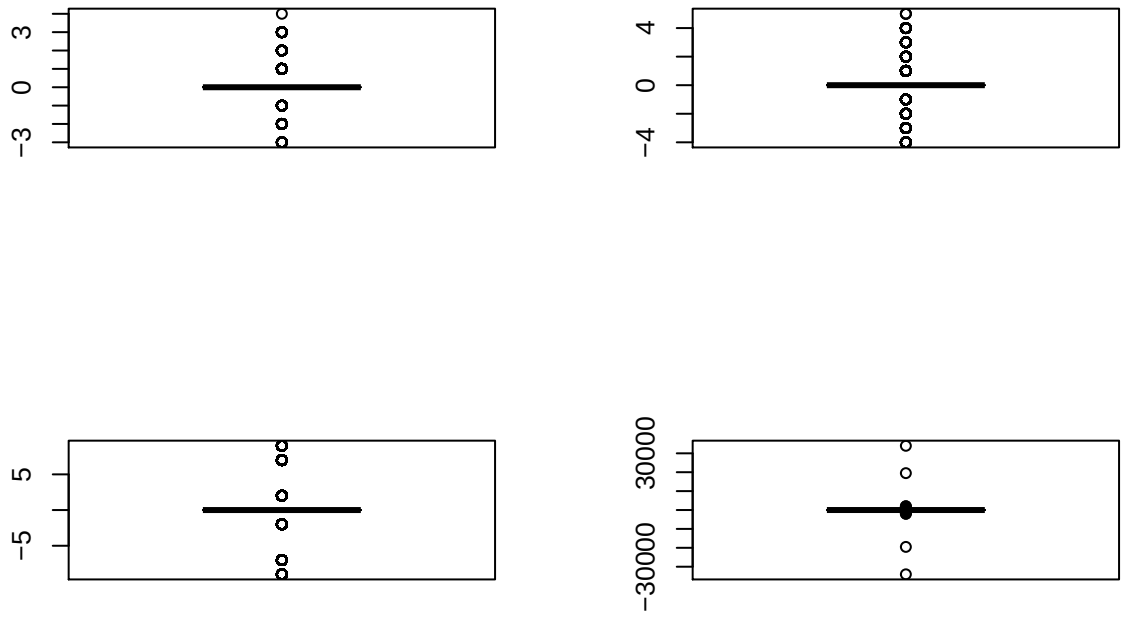

−4
0
4


4
0
−4


4
0
−4


4
0
−4


4
0
−4




```
BLR_user_HRA_duplicated$UserId <- as.character(BLR_user_HRA_duplicated$UserId)
userid <- BLR_user_HRA_duplicated$UserId

BLR.M <- grepl('M', userid)
BLR.R <- grepl('R', userid)
BLR.Z <- grepl('Z', userid)

BLR_user_M <- BLR_user_HRA_duplicated[BLR.M,]
BLR_user_R <- BLR_user_HRA_duplicated[BLR.R,]
BLR_user_Z <- BLR_user_HRA_duplicated[BLR.Z,]

BLR_user_M <- BLR_user_M[order(BLR_user_M$UserId),]
row.names(BLR_user_M) <- c(1:nrow(BLR_user_M))
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
BLR_user_R <- BLR_user_M[order(BLR_user_R$UserId),]
row.names(BLR_user_R) <- c(1:nrow(BLR_user_R))
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
BLR_user_Z <- BLR_user_M[order(BLR_user_Z$UserId),]
row.names(BLR_user_Z) <- c(1:nrow(BLR_user_Z))
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
############################################################

BLR_user_HRA_duplicated_omit_start$UserId <- as.character(BLR_user_HRA_duplicated_omit_start$UserId)
userid <- BLR_user_HRA_duplicated_omit_start$UserId

BLR.M <- grepl('M', userid)
BLR.R <- grepl('R', userid)
BLR.Z <- grepl('Z', userid)

BLR_user_M_omit_start <- BLR_user_HRA_duplicated_omit_start[BLR.M,]
BLR_user_R_omit_start <- BLR_user_HRA_duplicated_omit_start[BLR.R,]
BLR_user_Z_omit_start <- BLR_user_HRA_duplicated_omit_start[BLR.Z,]

BLR_user_M_omit_start <- BLR_user_M_omit_start[order(BLR_user_M_omit_start$UserId),]
row.names(BLR_user_M) <- c(1:nrow(BLR_user_M))
```

```
## Warning: Setting row names on a tibble is deprecated.
```

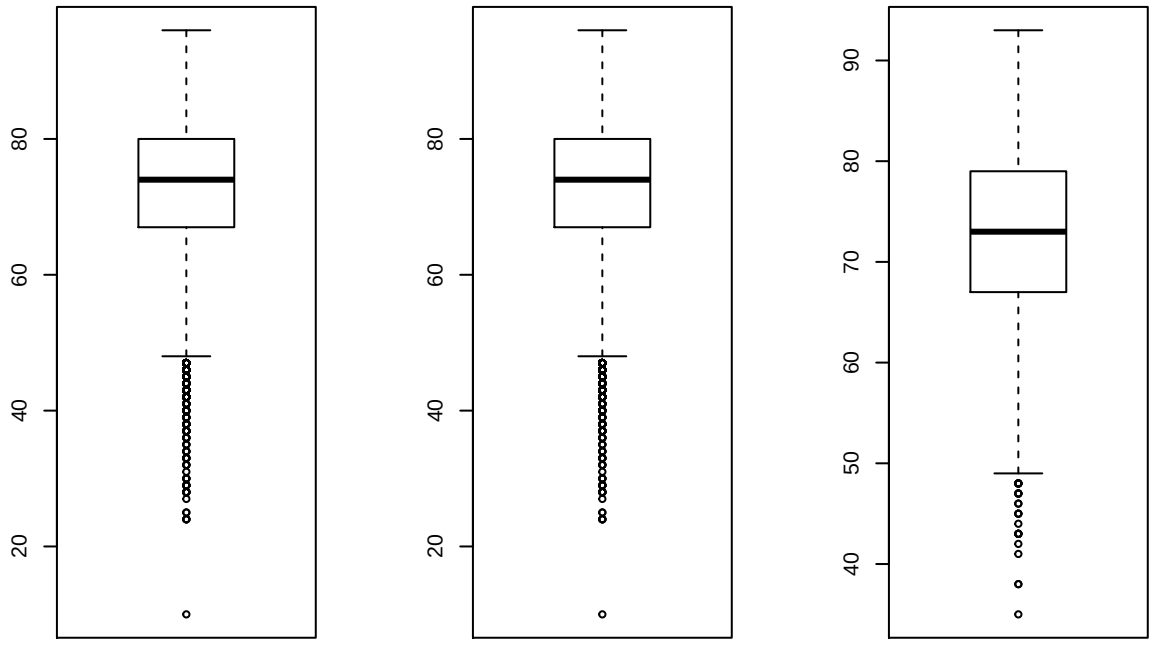
```
BLR_user_R_omit_start <- BLR_user_R_omit_start[order(BLR_user_R_omit_start$UserId),]
row.names(BLR_user_R) <- c(1:nrow(BLR_user_R))
```

```
## Warning: Setting row names on a tibble is deprecated.
```

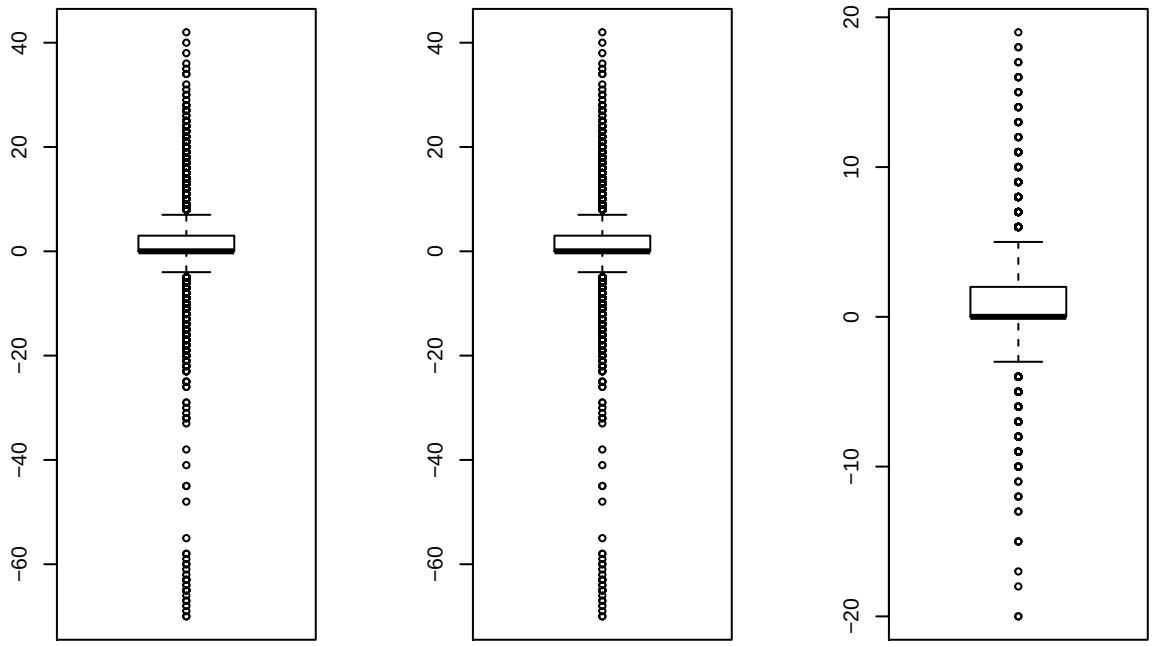
```
BLR_user_Z_omit_start <- BLR_user_Z_omit_start[order(BLR_user_Z_omit_start$UserId),]
row.names(BLR_user_Z) <- c(1:nrow(BLR_user_Z))
```

```
## Warning: Setting row names on a tibble is deprecated.
```

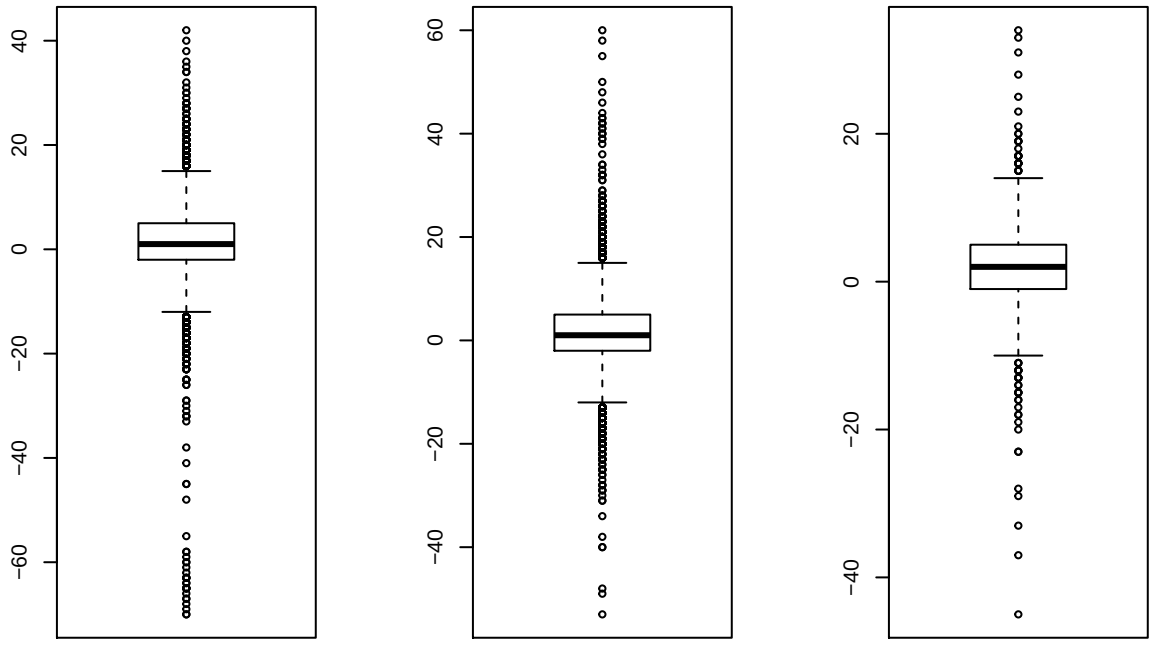
```
par(mfrow = c(1,3))
boxplot(BLR_user_M$BLRScore)
boxplot(BLR_user_R$BLRScore)
boxplot(BLR_user_Z$BLRScore)
```



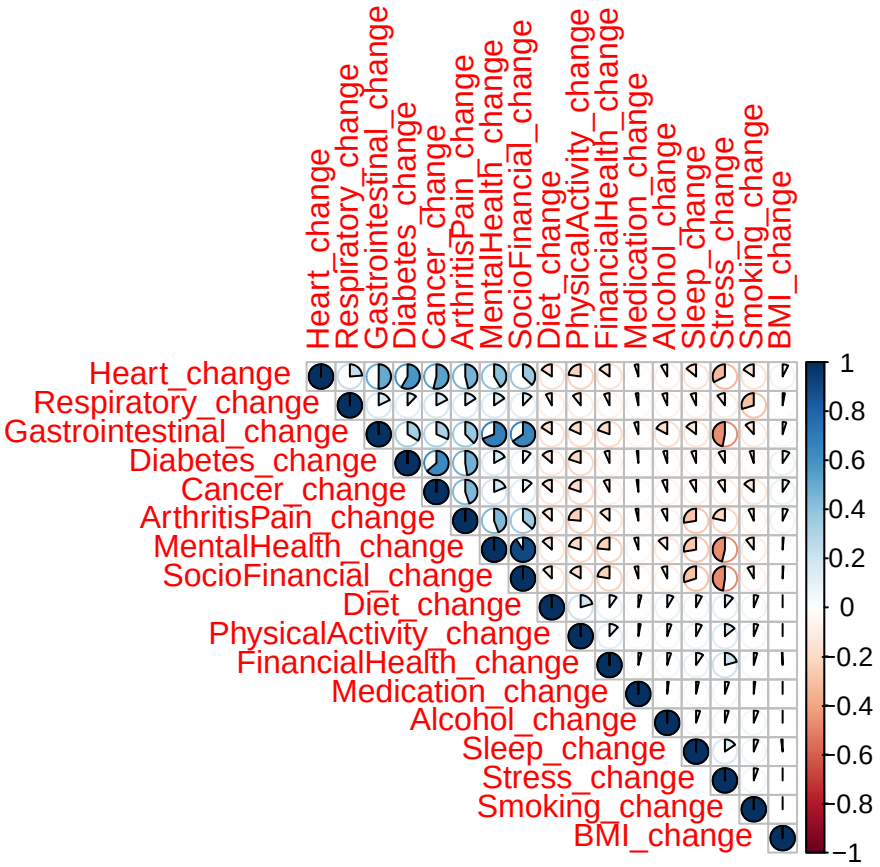

```
par(mfrow = c(1,3))
boxplot(BLR_user_M$BLR_change)
boxplot(BLR_user_R$BLR_change)
boxplot(BLR_user_Z$BLR_change)
```



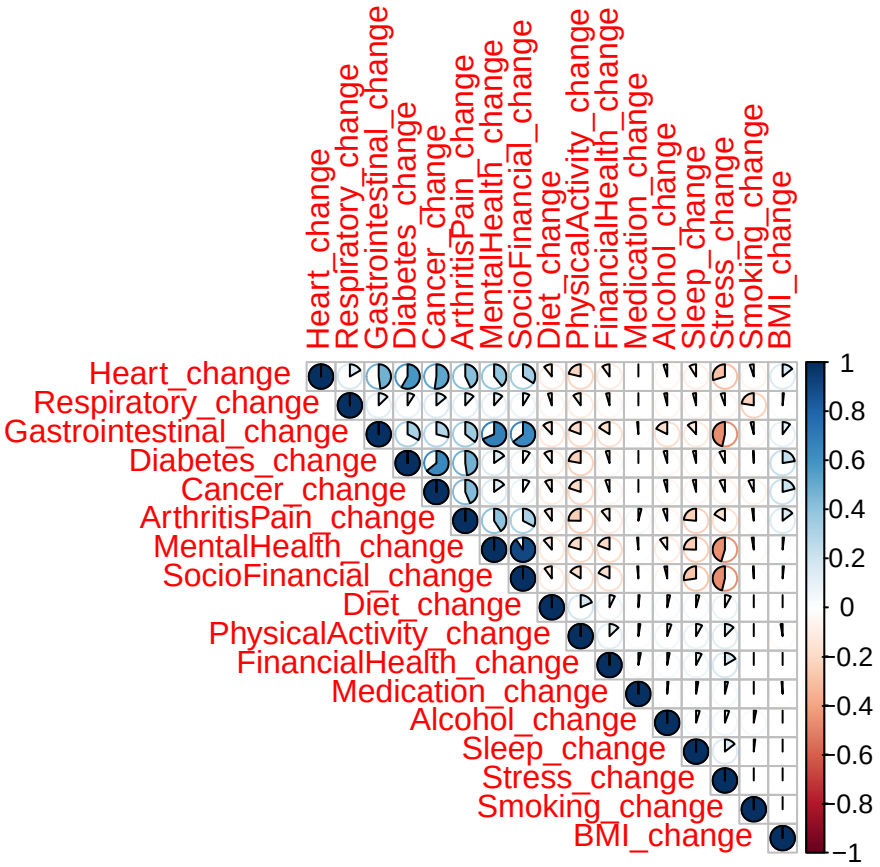

```
par(mfrow = c(1,3))
boxplot(BLR_user_M_omit_start$BLR_change)
boxplot(BLR_user_R_omit_start$BLR_change)
boxplot(BLR_user_Z_omit_start$BLR_change)
```



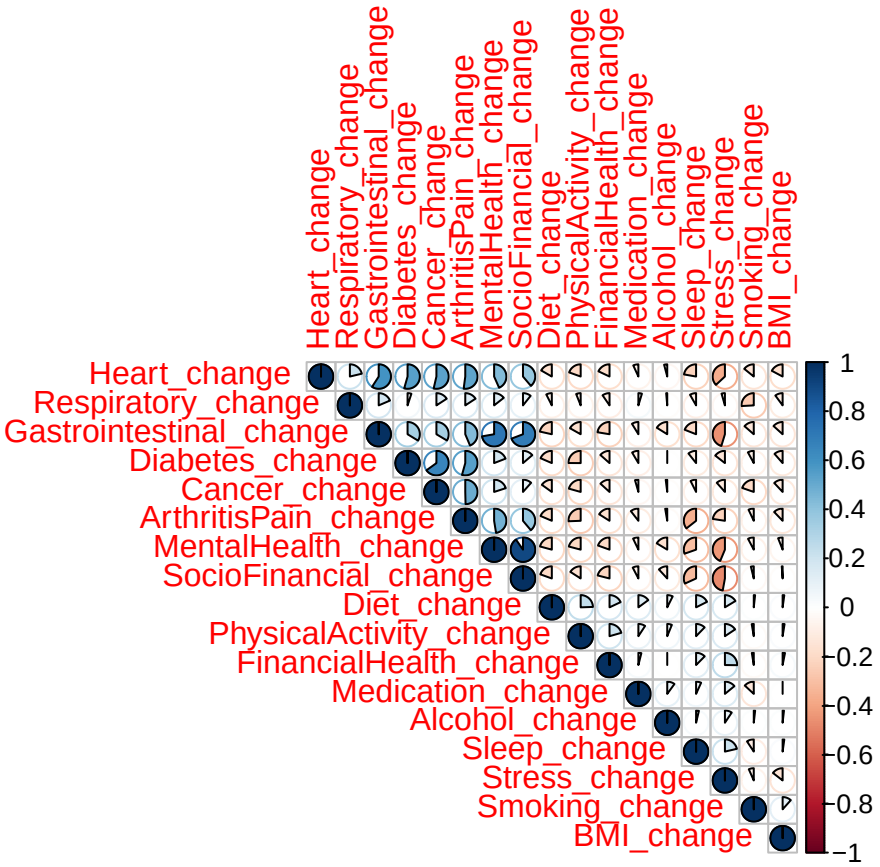

```
par(mfrow=c(1,1))
cor_M <- cor(BLR_user_M_omit_start[26:42],use = "complete.obs")
corrplot::corrplot(cor_M,method = 'pie',type='upper')
```




```
cor_R <- cor(BLR_user_R_omit_start[26:42],use = "complete.obs")
corrplot::corrplot(cor_R,method = 'pie',type='upper')
```




```
cor_Z <- cor(BLR_user_Z_omit_start[26:42],use = "complete.obs")
corrplot::corrplot(cor_Z,method = 'pie',type='upper')
```




```
points_M_positive <- subset(points_M,points > 0)
points_R_positive <- subset(points_R,points > 0)
points_Z_positive <- subset(points_Z,points > 0)

points_M_sum <- aggregate(points~userid,data = points_M_positive,sum)
points_R_sum <- aggregate(points~userid,data = points_R_positive,sum)
points_Z_sum <- aggregate(points~userid,data = points_Z_positive,sum)
rm(points_M,points_M_positive,points_R,points_R_positive,points_Z,points_Z_positive)
sum(nrow(points_M_sum),nrow(points_R_sum),nrow(points_Z_sum))
```

```
## [1] 217729
```

```
nrow(BLR_user_HRA)
```

```
## [1] 170965
```

```
#Pretty confusing
```