Week 3 Homework

Brian Ritz

Friday, January 30, 2015

Data in the file Week3_Homework_Project_Data.csv contain observations of time during the day when people watch TV (are logged in to an internet site, active online, etc.) and their age.

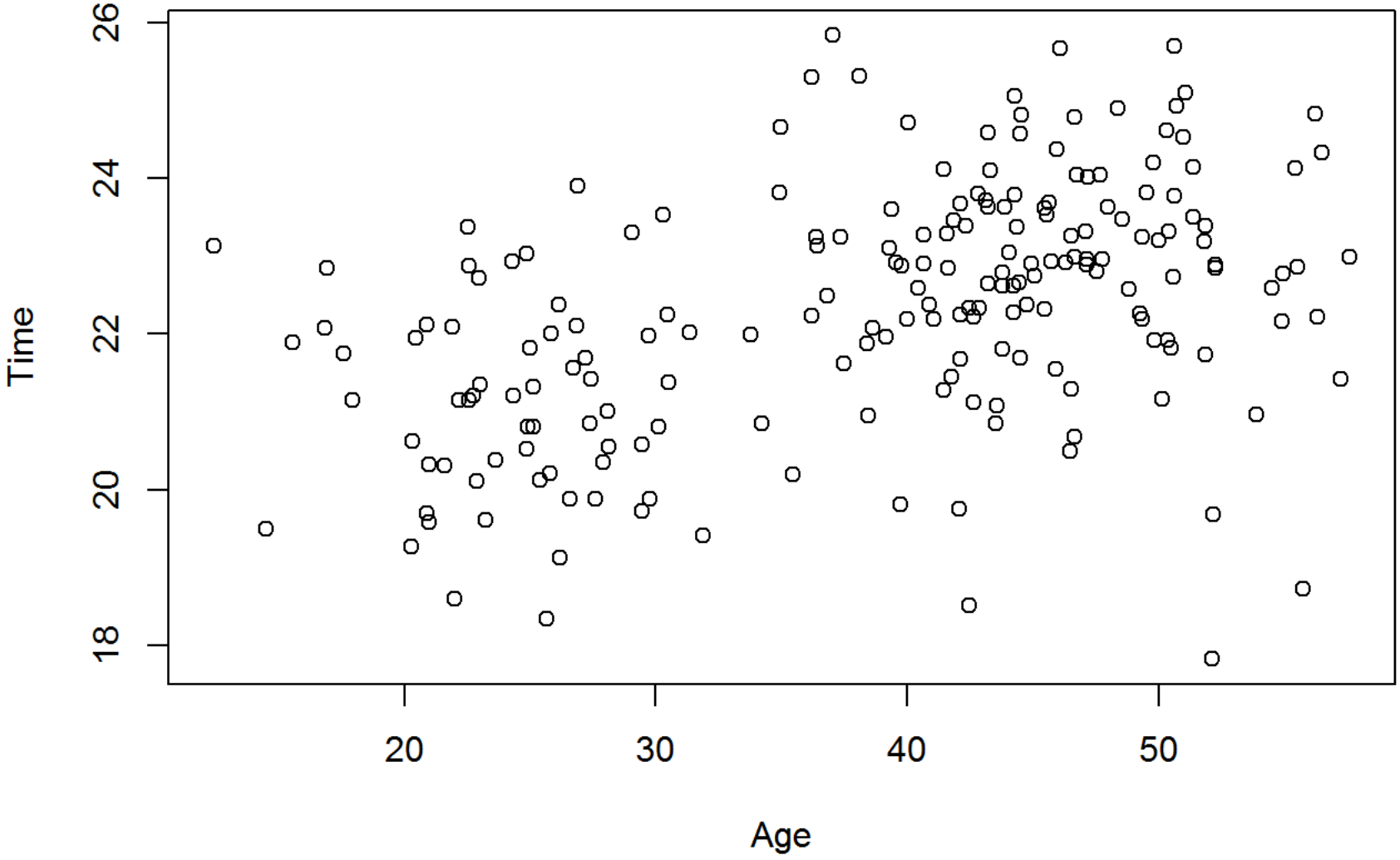
Read in data:

```
Age.Time.Sample<-read.csv(file="Week3_Homework_Project_Data.csv",header=TRUE,sep=",")
Age.Time.Sample<-as.matrix(Age.Time.Sample)
Age.Time.Sample[1:10,]
```

```
##         Age  Time
##  [1,] 31.35 22.03
##  [2,] 27.44 21.42
##  [3,] 22.53 23.38
##  [4,] 21.00 20.33
##  [5,] 22.58 21.16
##  [6,] 29.05 23.31
##  [7,] 26.71 21.56
##  [8,] 23.61 20.38
##  [9,] 28.07 21.01
## [10,] 28.11 20.55
```

```
plot(Age.Time.Sample)
```


PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

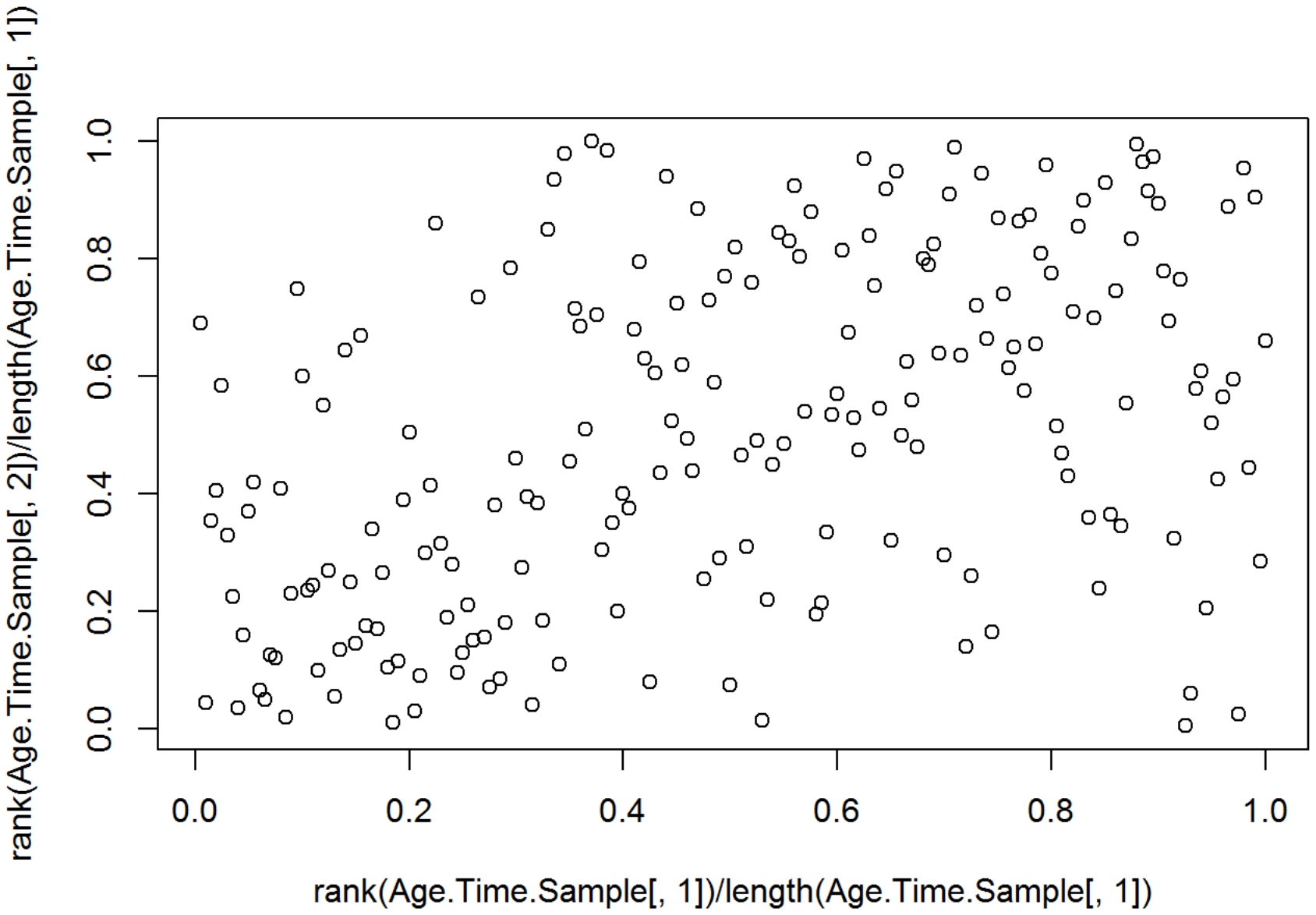


There looks like there might be two clusters here.

```
# Once you rank,the clusters disappear a little bit
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
plot(rank(Age.Time.Sample[,1])/length(Age.Time.Sample[,1]),rank(Age.Time.Sample[,2])/length(Age.Time.Sample
[,1]))
```



PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

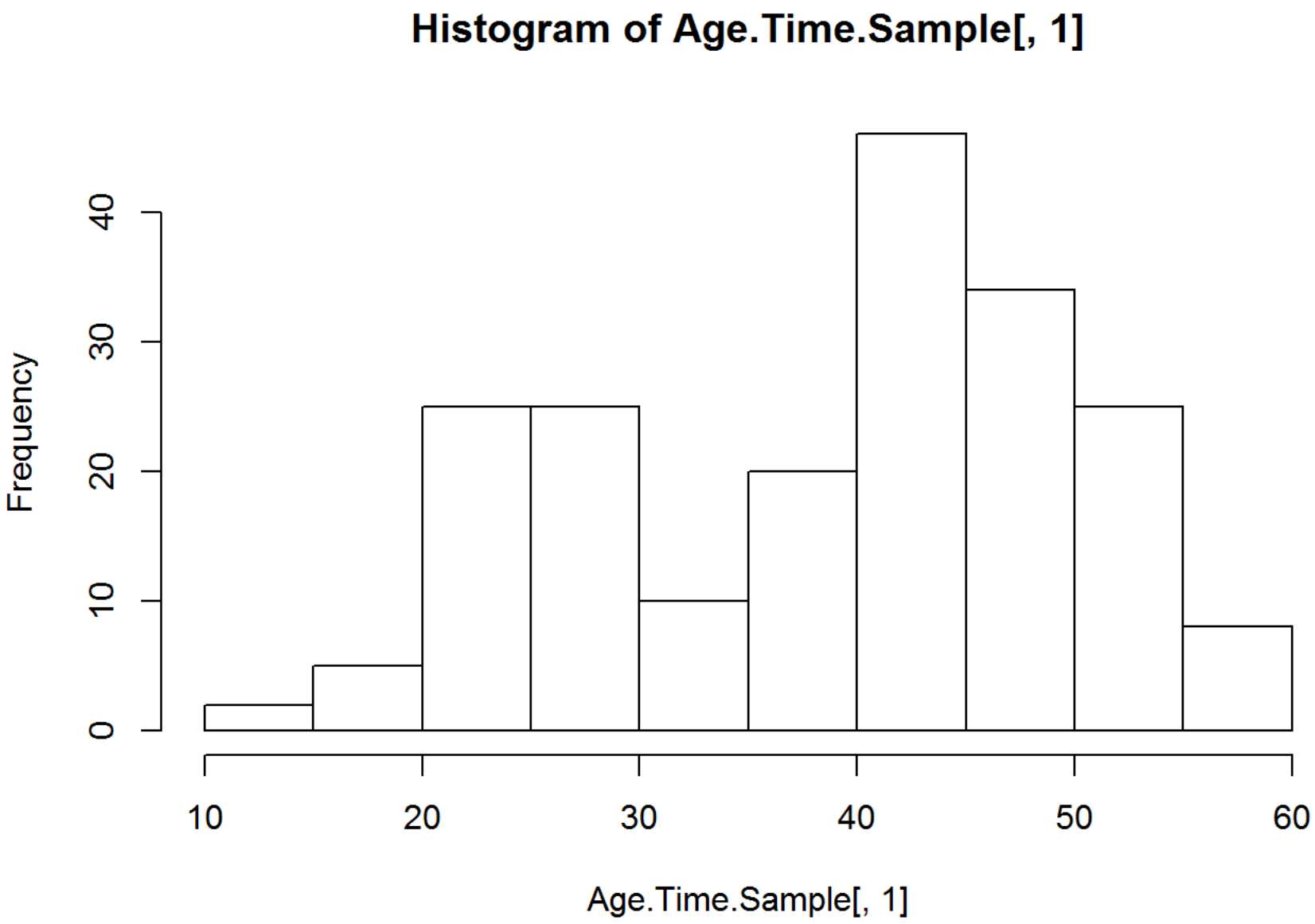
```
# check out the correlation
c(cor(Age.Time.Sample)[1,2],cor(Age.Time.Sample)[1,2]^2)
```

```
## [1] 0.4359 0.1900
```

Now, lets look at the histograms for the samples.

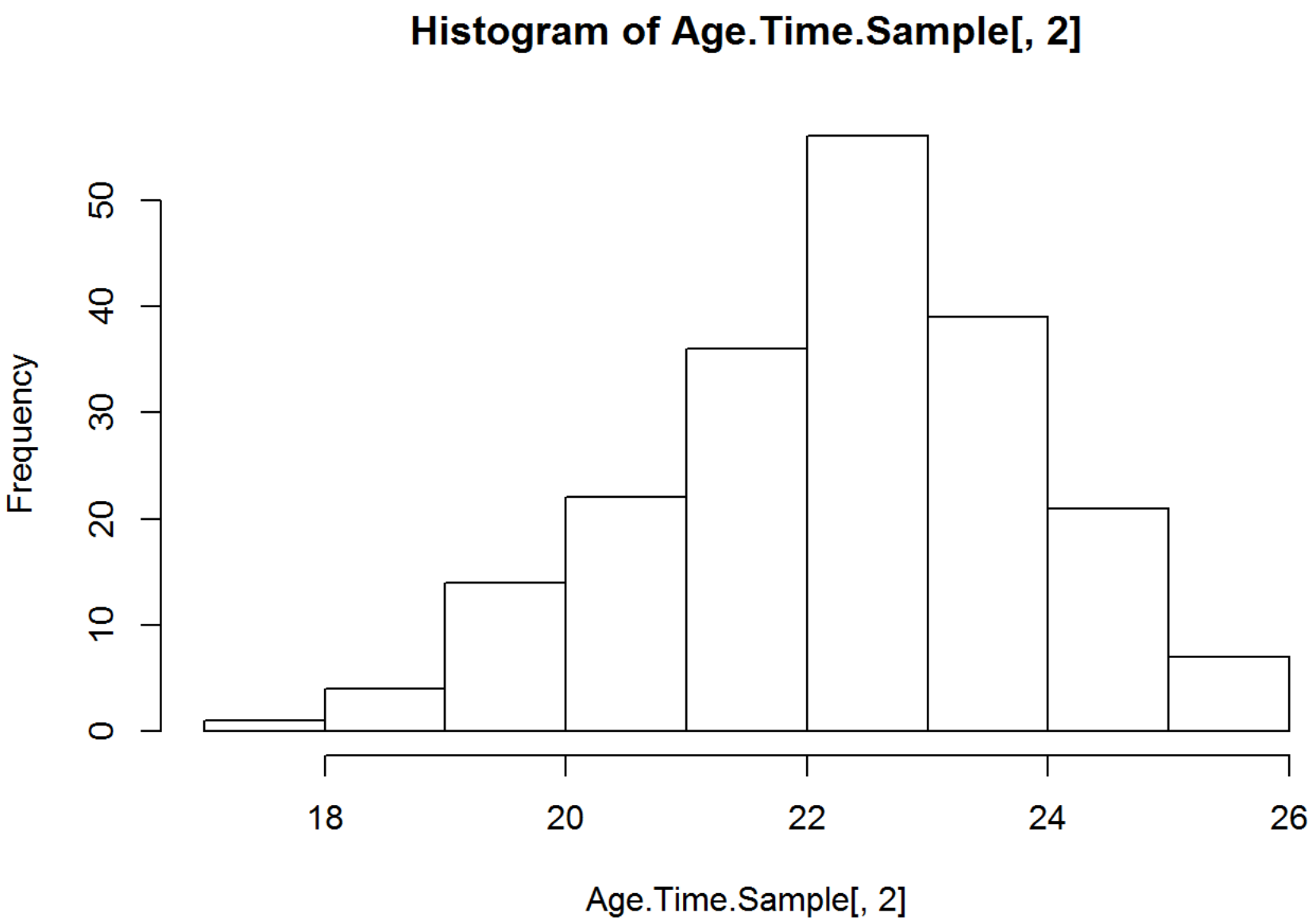
```
hist(Age.Time.Sample[,1])
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
hist(Age.Time.Sample[,2])
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com



Interpret the initial observations

1. What do you see on the scatterplot of Age vs. Time?

I see two clusters one, at the upper right centered at about 40 years old and maybe 23 Time, and the other at about 25

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

years old Age and 21 Time. The first cluster(40, 23) looks like it has more observations in it.

2. What does the empirical copula suggest?

The dependence and the clusters are less evident, but still look like they are present. It lookslike there is positive correlation, because it appears that there are observations missing from the upper left and lower right corners of the copula.

3. How significant is the amount of correlation?

There is a moderate relationship between the data. The relationship decreases when a monotonic function(square function) is applied to the data. The correlation coefficient is .43 for the untransformed variables and .19 for the transformed variables.

4. What do you imply from the shapes of the histograms?

The shapes of the histograms make it look like there is a bimodal distribution of Age, centered around 45 and 25. The histogram for Time looks like it has a fat-tail on the left. It is skewed left. This may be indicitive that there is actually two normal data generating processes with means near to eachother underlying this sample.

Clustering Of the data
Find possible clusters of the values.

```
library(mclust)
```

```
## Package 'mclust' version 4.4
## Type 'citation("mclust")' for citing this R package in publications.
```

```
# use Mclust() to cluster the age component and the time component

Age.Clusters<-Mclust(data=Age.Time.Sample[,1], modelNames="V")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
names(Age.Clusters)
```

```
##  [1] "call"           "data"           "modelName"      "n"
##  [5] "d"              "G"              "BIC"            "bic"
##  [9] "loglik"         "df"             "hypvol"         "parameters"
## [13] "z"              "classification" "uncertainty"
```

```
Age.Clusters$G
```

```
## [1] 2
```

```
Age.Clusters$parameters
```

```
## $Vinv
## NULL
##
## $pro
## [1] 0.3187 0.6813
##
## $mean
##     1     2
## 24.62 45.39
##
## $variance
## $variance$modelName
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
## [1] "V"
##
## $variance$d
## [1] 1
##
## $variance$G
## [1] 2
##
## $variance$sigmasq
## [1] 21.02 31.13
##
## $variance$scale
## [1] 21.02 31.13
```

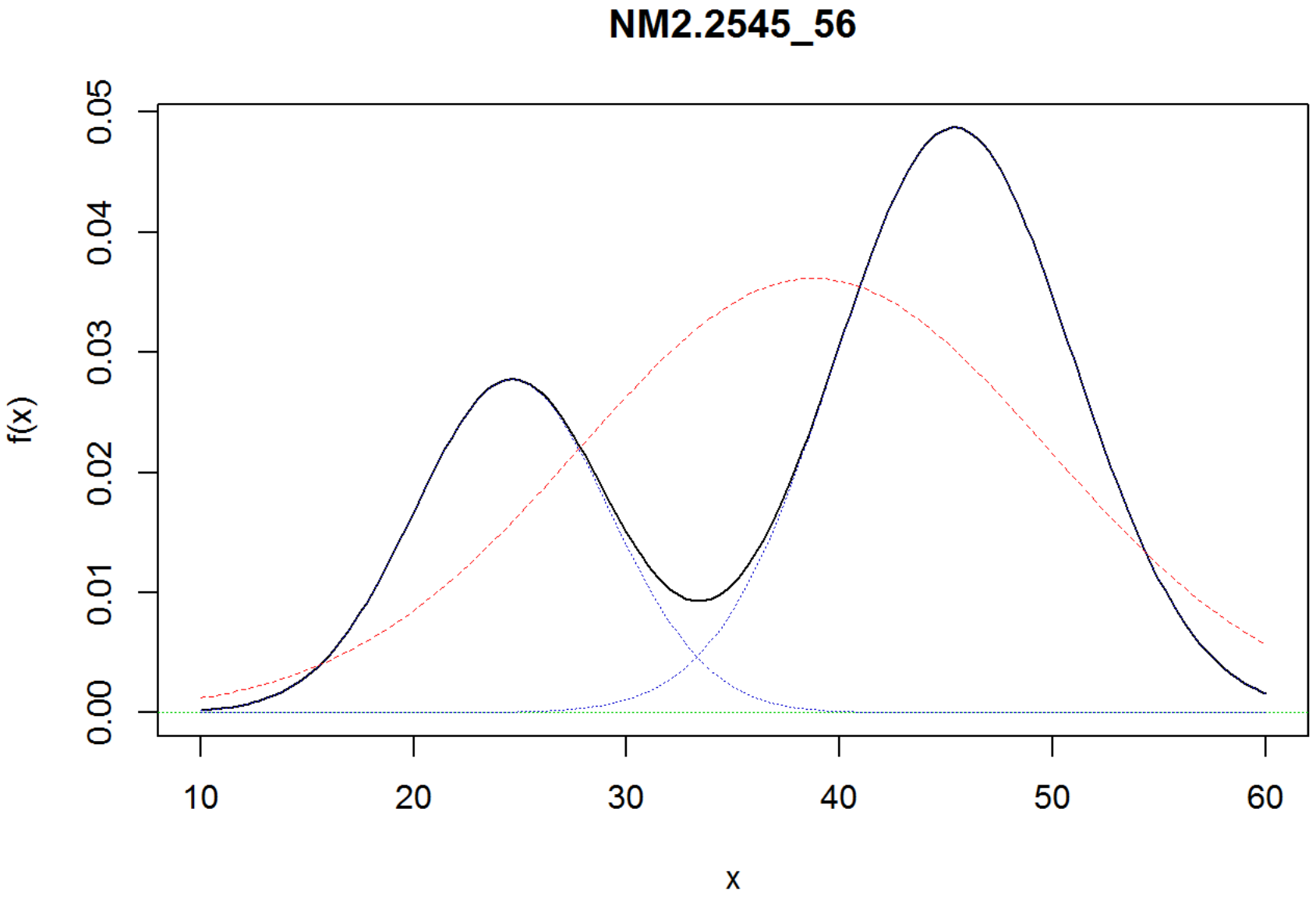
```
Age.Clusters.Parameters<-rbind(mu=Age.Clusters$parameters$mean,
                               sigma=sqrt(Age.Clusters$parameters$variance$sigmasq),
                               pro=Age.Clusters$parameters$pro)
```

The G and parameters match what is given in the homework description.

We now use norMix to analyze the mixed Gaussian models.

```
library(nor1mix)
Classified.Mix.Model.Age <- norMix(Age.Clusters.Parameters["mu",],
                                   sigma=Age.Clusters.Parameters["sigma",],
                                   w=Age.Clusters.Parameters["pro",])
plot(Classified.Mix.Model.Age,xout=seq(from=10,to=60,by=.25),p.norm=TRUE,p.comp=TRUE)
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com



Now let's move on to the time clusters.

```
# Now on to time clusters
Time.Clusters<-Mclust(data=Age.Time.Sample[,2], modelNames="V")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
## Warning: best model occurs at the min or max # of components considered
## Warning: optimal number of clusters occurs at min choice
```

```
names(Time.Clusters)
```

```
##  [1] "call"           "data"           "modelName"      "n"
##  [5] "d"              "G"              "BIC"            "bic"
##  [9] "loglik"         "df"             "hypvol"         "parameters"
## [13] "z"              "classification" "uncertainty"
```

```
Time.Clusters$G
```

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

```
Time.Clusters$parameters
```

```
## $pro
## [1] 1
##
## $mean
## [1] 22.32
##
## $variance
## $variance$modelName
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
## [1] "X"
##
## $variance$d
## [1] 1
##
## $variance$G
## [1] 1
##
## $variance$sigmasq
## [1] 2.503
```

```
# refit and force the number of clusters to be at least 2
Time.Clusters<-Mclust(data=Age.Time.Sample[,2], G=c(2:9), modelNames="V")
```

```
## Warning: best model occurs at the min or max # of components considered
## Warning: optimal number of clusters occurs at min choice
```

```
names(Time.Clusters)
```

```
##  [1] "call"           "data"           "modelName"      "n"
##  [5] "d"              "G"              "BIC"            "bic"
##  [9] "loglik"         "df"             "hypvol"         "parameters"
## [13] "z"              "classification" "uncertainty"
```

```
Time.Clusters$G
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
## [1] 2
```

```
Time.Clusters$parameters
```

```
## $Vinv
## NULL
##
## $pro
## [1] 0.4668 0.5332
##
## $mean
##     1     2
## 21.32 23.19
##
## $variance
## $variance$modelName
## [1] "V"
##
## $variance$d
## [1] 1
##
## $variance$G
## [1] 2
##
## $variance$sigmasq
## [1] 2.079 1.241
##
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
## $variance$scale
## [1] 2.079 1.241
```

```
Time.Clusters.Parameters<-rbind(mu=Time.Clusters$parameters$mean,
                                sigma=sqrt(Time.Clusters$parameters$variance$sigmasq),
                                pro=Time.Clusters$parameters$pro)
```

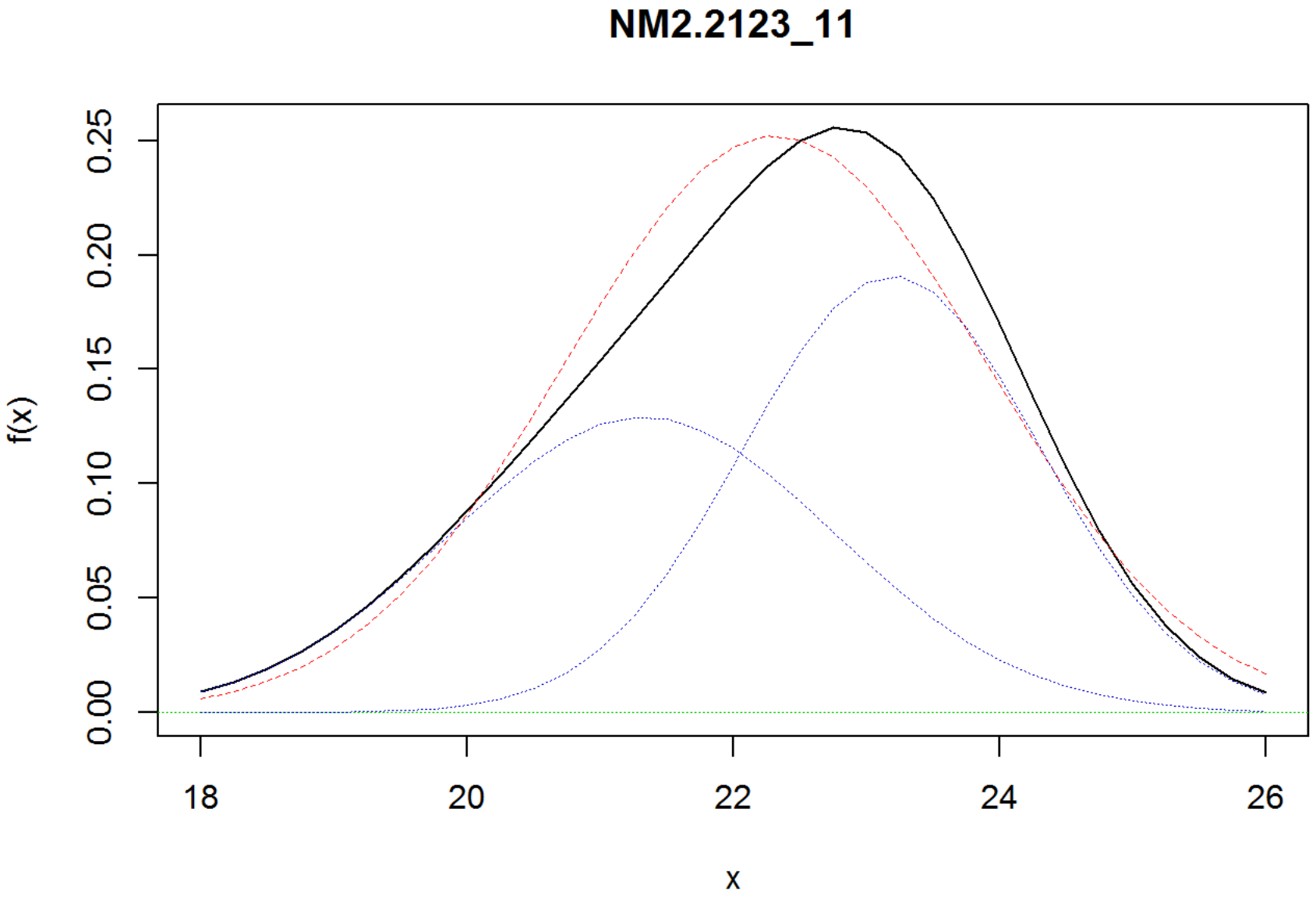
Make a norMix object for time:

```
# make a normix object

Classified.Mix.Model.Time <- norMix(Time.Clusters.Parameters["mu",],
                                    sigma=Time.Clusters.Parameters["sigma",],
                                    w=Time.Clusters.Parameters["pro",])

plot(Classified.Mix.Model.Time,xout=seq(from=18,to=26,by=.25),p.norm=TRUE,p.comp=TRUE)
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com



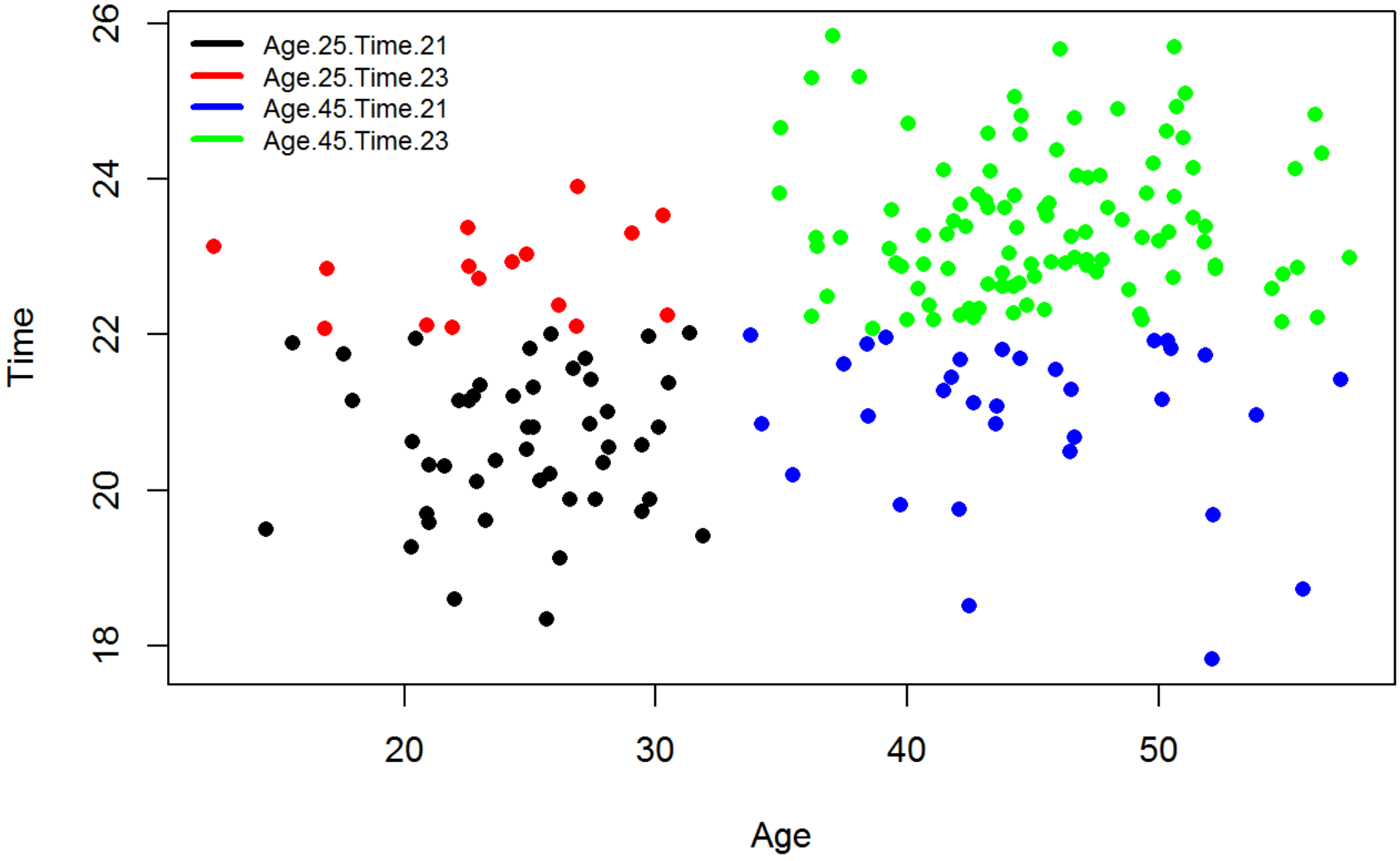
I was right, the means for the two normal distributions are fairly close to eachother; and the distributions overlap for most of their density.

```
#separate samples and explore dependencies
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
Age.Mixing.Sequence<-Age.Clusters$classification
Age.25.Time.21.Mixing.Sequence<-((Age.Clusters$classification==1)&(Time.Clusters$classification==1))
Age.25.Time.23.Mixing.Sequence<-((Age.Clusters$classification==1)&(Time.Clusters$classification==2))
Age.45.Time.21.Mixing.Sequence<-((Age.Clusters$classification==2)&(Time.Clusters$classification==1))
Age.45.Time.23.Mixing.Sequence<-((Age.Clusters$classification==2)&(Time.Clusters$classification==2))
Grouped.Data.Age.25.Time.21<-
  Grouped.Data.Age.25.Time.23<-
  Grouped.Data.Age.45.Time.21<-
  Grouped.Data.Age.45.Time.23<-
  cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Age.25.Time.21[Age.25.Time.21.Mixing.Sequence,]<-
  Age.Time.Sample[Age.25.Time.21.Mixing.Sequence,]
Grouped.Data.Age.25.Time.23[Age.25.Time.23.Mixing.Sequence,]<-
  Age.Time.Sample[Age.25.Time.23.Mixing.Sequence,]
Grouped.Data.Age.45.Time.21[Age.45.Time.21.Mixing.Sequence,]<-
  Age.Time.Sample[Age.45.Time.21.Mixing.Sequence,]
Grouped.Data.Age.45.Time.23[Age.45.Time.23.Mixing.Sequence,]<-
  Age.Time.Sample[Age.45.Time.23.Mixing.Sequence,]
matplot(Age.Time.Sample[,1],cbind(Grouped.Data.Age.25.Time.21[,2],
                                  Grouped.Data.Age.25.Time.23[,2],
                                  Grouped.Data.Age.45.Time.21[,2],
                                  Grouped.Data.Age.45.Time.23[,2]),
        pch=16,xlab="Age",ylab="Time",
        col=c('black','red', 'blue', 'green'))
legend('topleft', c("Age.25.Time.21","Age.25.Time.23","Age.45.Time.21","Age.45.Time.23") ,
       lty=1,lwd=3, col=c('black','red', 'blue', 'green'), bty='n', cex=.75)
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com



What dependencies do you see on the chart? It lookslike within the age45 time 23 cluster, there is a positive dependency. Within age 25s, there does not appear to be much dependency. For Age 45 time 21, there seem to be not as many observations as one would expect at the lower right hand corner, so we may see some dependence there.

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

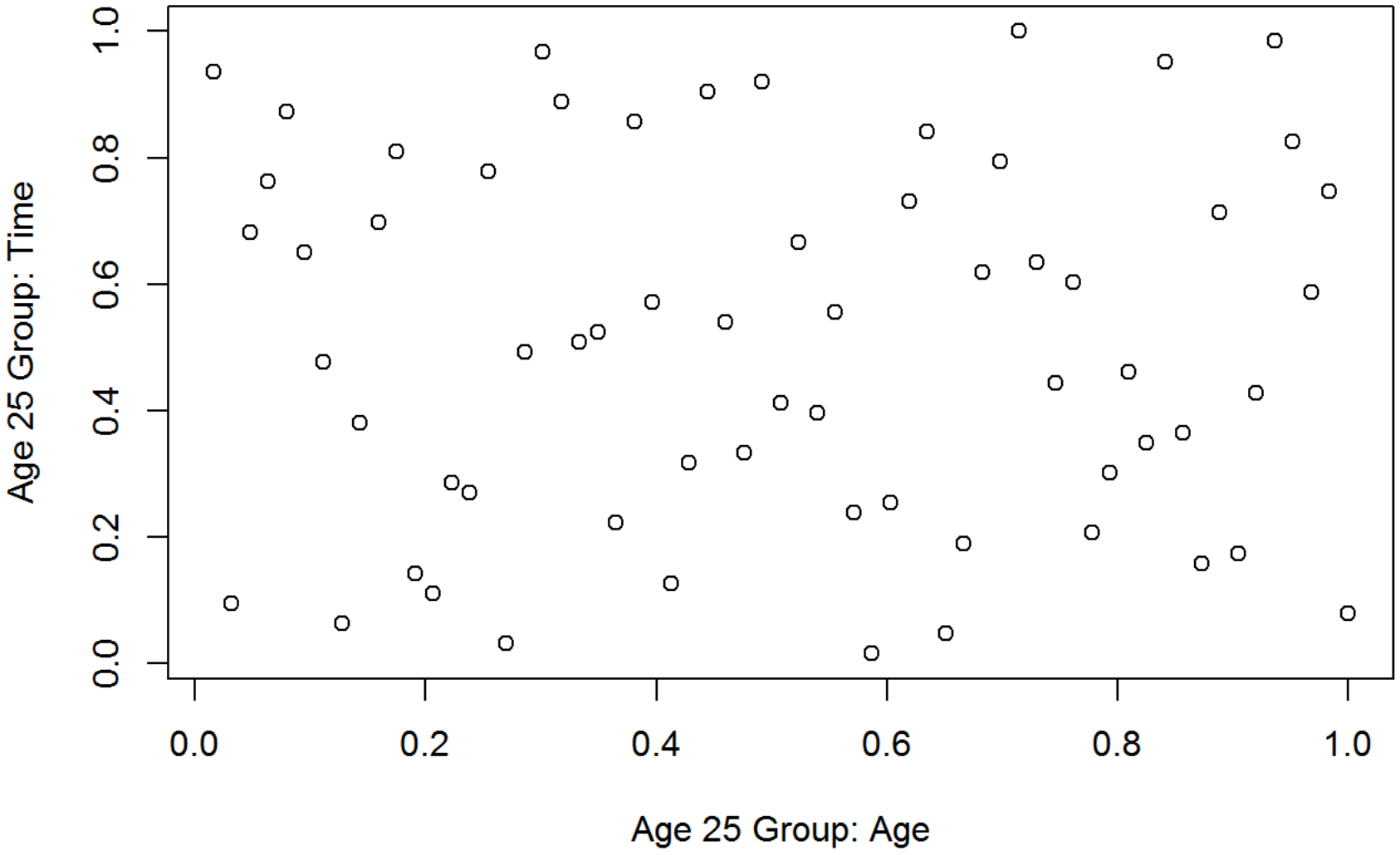
LOOK AT DEPENDENCIES:

Group by age:

```
#Group by age
Grouped.Data.Age.25<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Age.25[Age.Clusters$classification==1,]<-Age.Time.Sample[Age.Clusters$classification==1,]
Grouped.Data.Age.45<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Age.45[Age.Clusters$classification==2,]<-Age.Time.Sample[Age.Clusters$classification==2,]
```

```
plot(rank(na.omit(Grouped.Data.Age.25[,1]))/length(na.omit(Grouped.Data.Age.25[,1])),
     rank(na.omit(Grouped.Data.Age.25[,2]))/length(na.omit(Grouped.Data.Age.25[,2])),
     xlab="Age 25 Group: Age",ylab="Age 25 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com



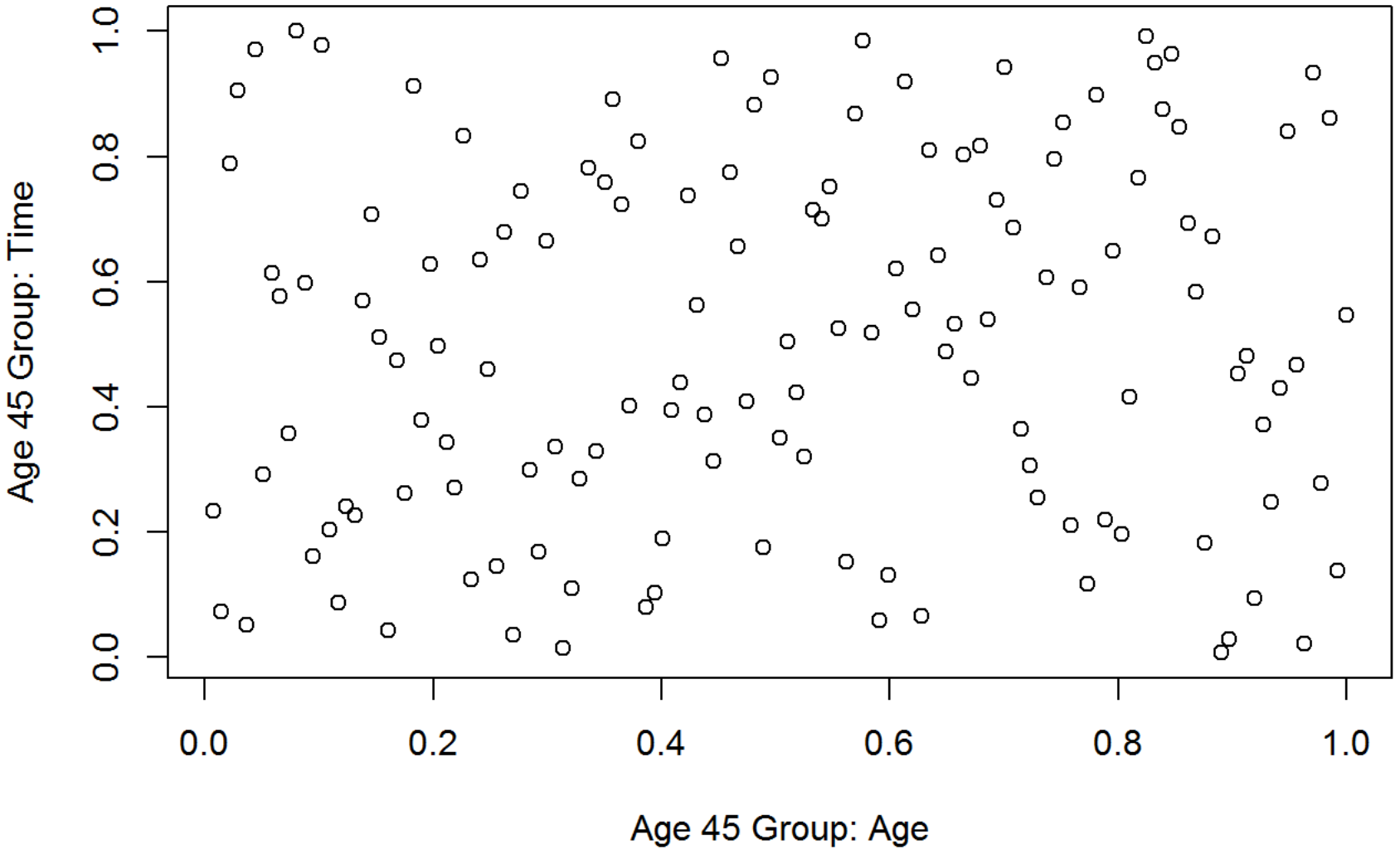

```
cor(na.omit(Grouped.Data.Age.25),method="spearman")[1,2]
```

```
## [1] -0.01128
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
plot(rank(na.omit(Grouped.Data.Age.45[,1]))/length(na.omit(Grouped.Data.Age.45[,1])),
     rank(na.omit(Grouped.Data.Age.45[,2]))/length(na.omit(Grouped.Data.Age.45[,2])),
     xlab="Age 45 Group: Age",ylab="Age 45 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
cor(na.omit(Grouped.Data.Age.45),method="spearman")[1,2]
```

```
## [1] 0.08611
```

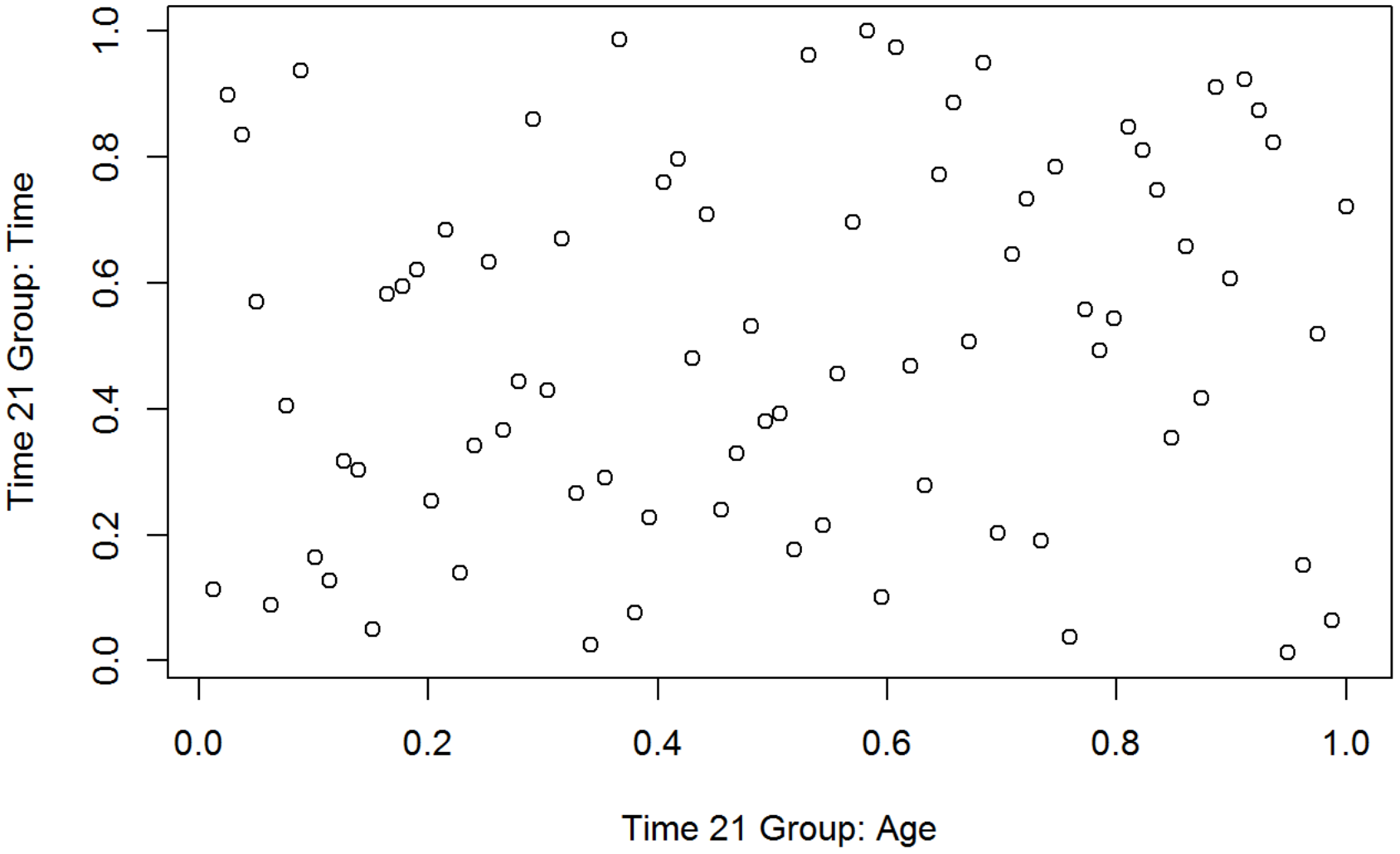
PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

Now group by time:

```
#Group by Time
Grouped.Data.Time.21<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Time.21[Time.Clusters$classification==1,]<-
  Age.Time.Sample[Time.Clusters$classification==1,]
Grouped.Data.Time.23<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Time.23[Time.Clusters$classification==2,]<-
  Age.Time.Sample[Time.Clusters$classification==2,]
```

```
plot(rank(na.omit(Grouped.Data.Time.21[,1]))/length(na.omit(Grouped.Data.Time.21[,1])),
     rank(na.omit(Grouped.Data.Time.21[,2]))/length(na.omit(Grouped.Data.Time.21[,2])),
     xlab="Time 21 Group: Age",ylab="Time 21 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
cor(na.omit(Grouped.Data.Time.21),method="spearman")[1,2]
```

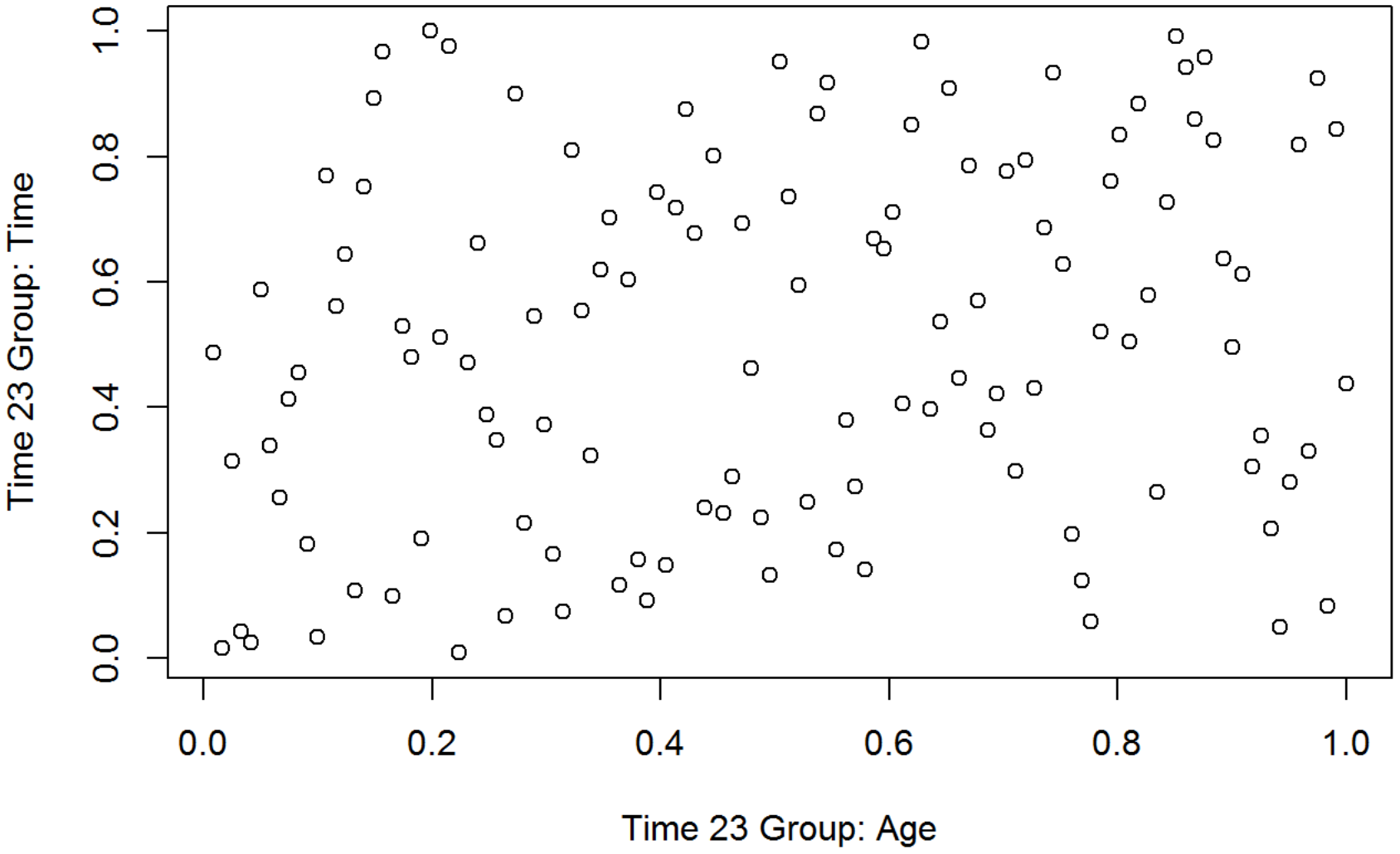
```
## [1] 0.1723
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

The copula looks pretty uniform here, possibly slightly less observations up to the left. Along with the small correlation (.17) the evidence for dependence is not very strong.

```
plot(rank(na.omit(Grouped.Data.Time.23[,1]))/length(na.omit(Grouped.Data.Time.23[,1])),
     rank(na.omit(Grouped.Data.Time.23[,2]))/length(na.omit(Grouped.Data.Time.23[,2])),
     xlab="Time 23 Group: Age",ylab="Time 23 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
cor(na.omit(Grouped.Data.Time.23),method="spearman")[1,2]
```

```
## [1] 0.2115
```

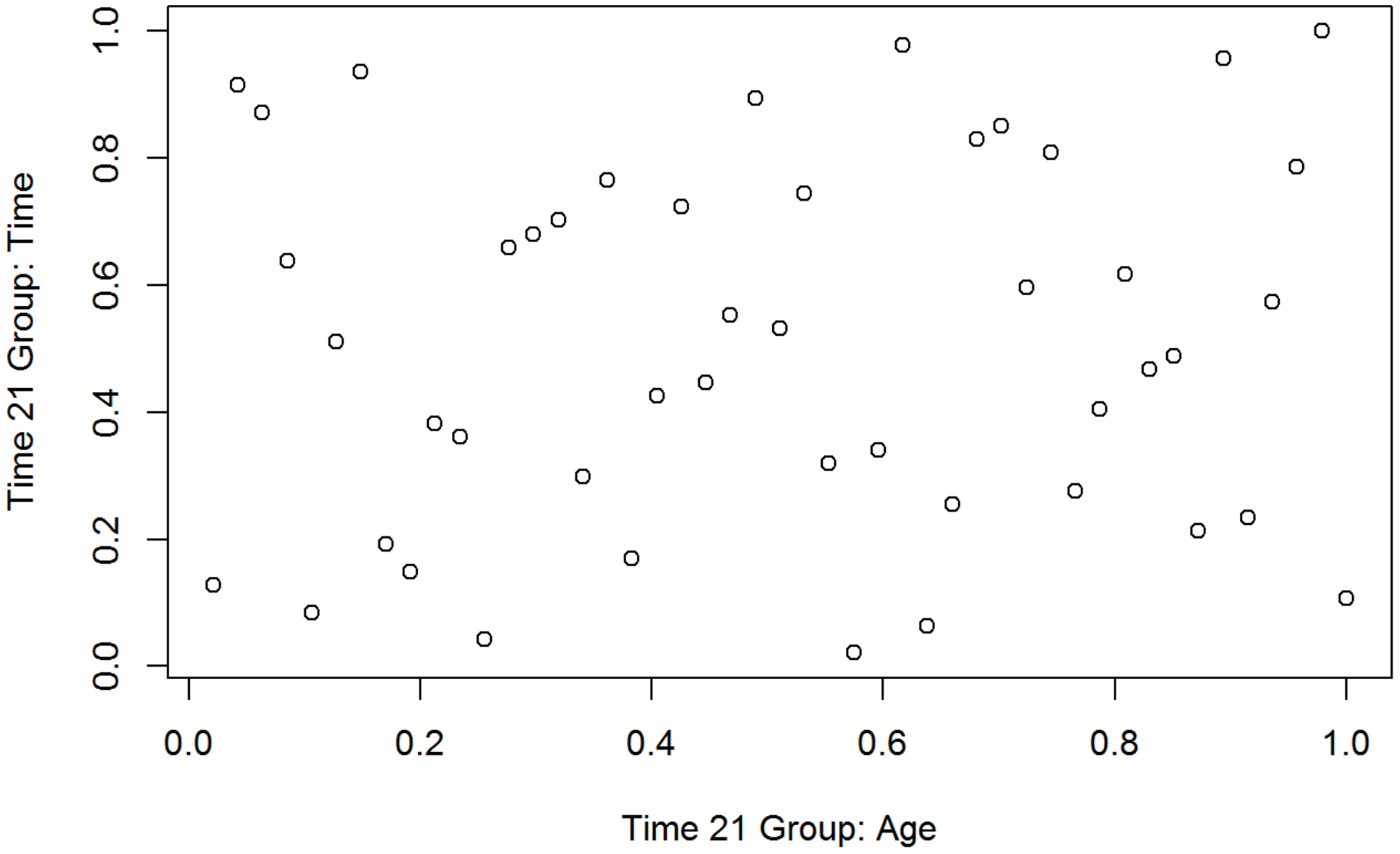
PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

What do you infer from the groupings by age and time? The correlation coefficients are moderate .2 and .17, so there is moderate to low correlation within groups. The empirical copulas appear pretty uniform, so there is not much dependence evident there.

Group by age and time:

```
#Group by Age and Time
#Grouped.Data.Age.25.Time.21
plot(rank(na.omit(Grouped.Data.Age.25.Time.21[,1]))/length(na.omit(Grouped.Data.Age.25.Time.21[,1])),
     rank(na.omit(Grouped.Data.Age.25.Time.21[,2]))/length(na.omit(Grouped.Data.Age.25.Time.21[,2])),
     xlab="Time 21 Group: Age",ylab="Time 21 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
cor(na.omit(Grouped.Data.Age.25.Time.21),method="spearman")[1,2]
```

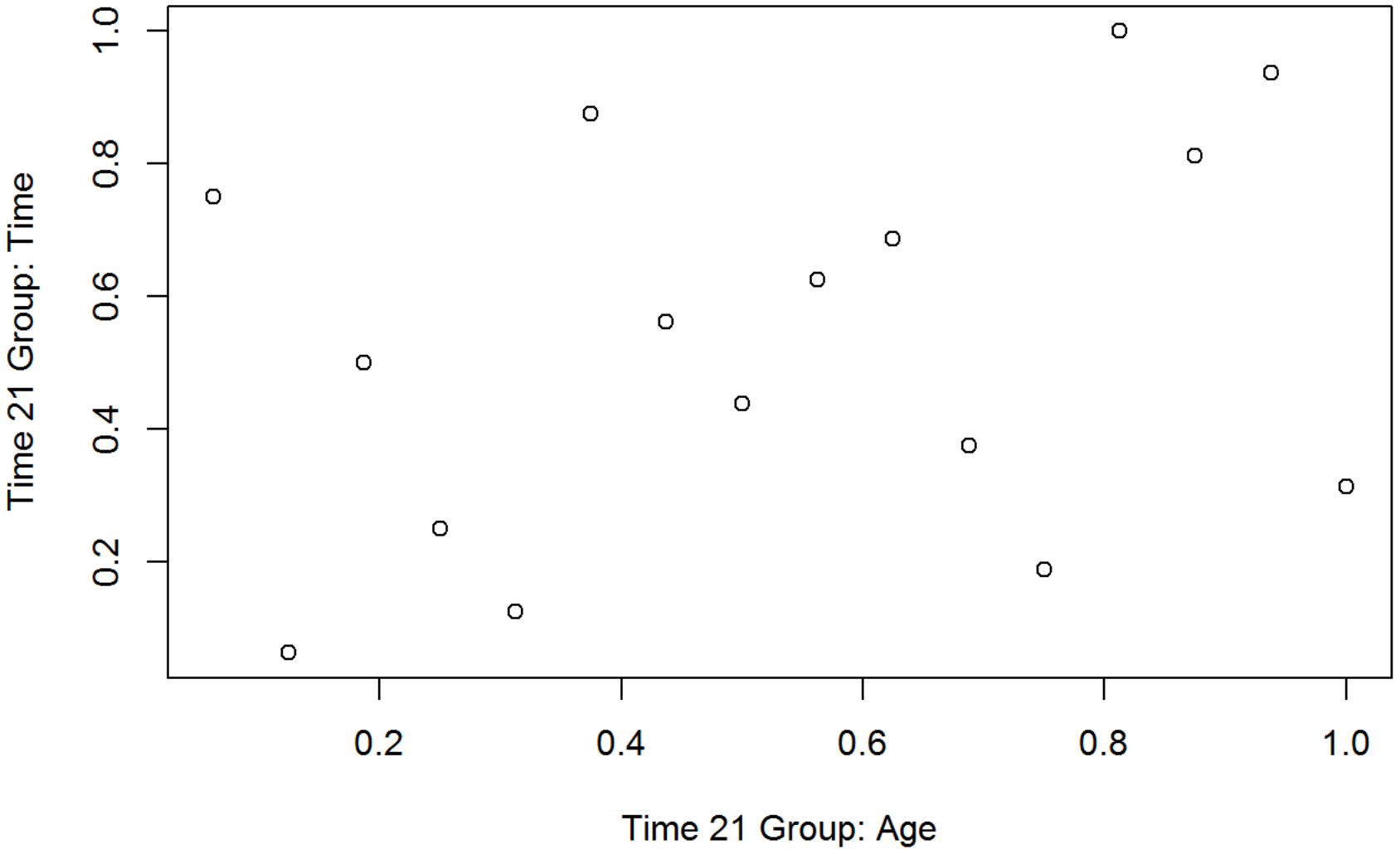
```
## [1] 0.07932
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

Copula is pretty sparse and uniformly distributed. Correlation coefficient is low–evidence for dependence is weak.

```
#Grouped.Data.Age.25.Time.23
plot(rank(na.omit(Grouped.Data.Age.25.Time.23[,1]))/length(na.omit(Grouped.Data.Age.25.Time.23[,1])),
     rank(na.omit(Grouped.Data.Age.25.Time.23[,2]))/length(na.omit(Grouped.Data.Age.25.Time.23[,2])),
     xlab="Time 21 Group: Age",ylab="Time 21 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
cor(na.omit(Grouped.Data.Age.25.Time.23),method="spearman")[1,2]
```

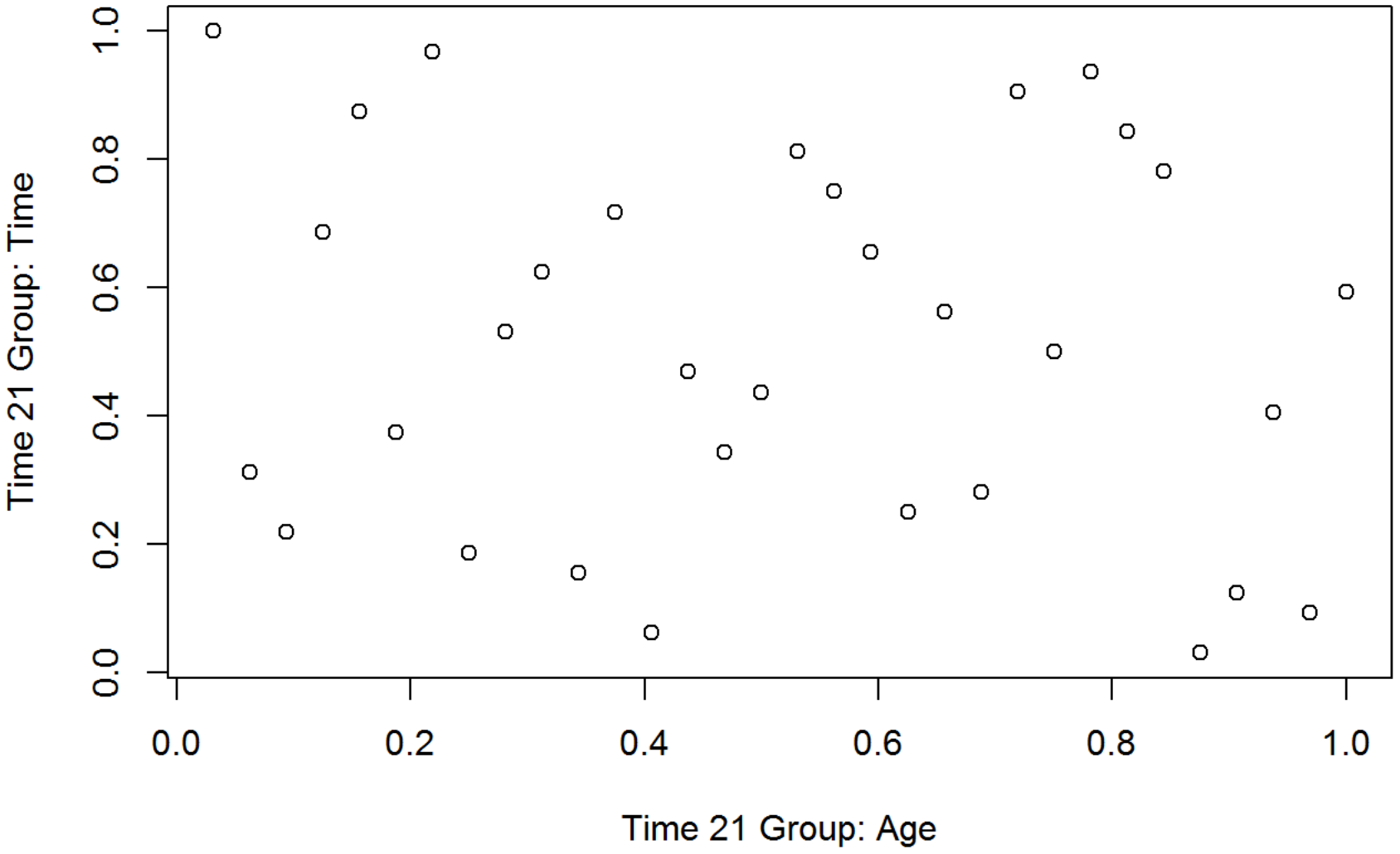
```
## [1] 0.3176
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

There looks to be a modest amount of correlation in the copula, and the correlation coefficient is higher than wev’e seen so far in this analysis. There is moderate evidence for dependence here

```
#Grouped.Data.Age.45.Time.21
plot(rank(na.omit(Grouped.Data.Age.45.Time.21[,1]))/length(na.omit(Grouped.Data.Age.45.Time.21[,1])),
     rank(na.omit(Grouped.Data.Age.45.Time.21[,2]))/length(na.omit(Grouped.Data.Age.45.Time.21[,2])),
     xlab="Time 21 Group: Age",ylab="Time 21 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
cor(na.omit(Grouped.Data.Age.45.Time.21),method="spearman")[1,2]
```

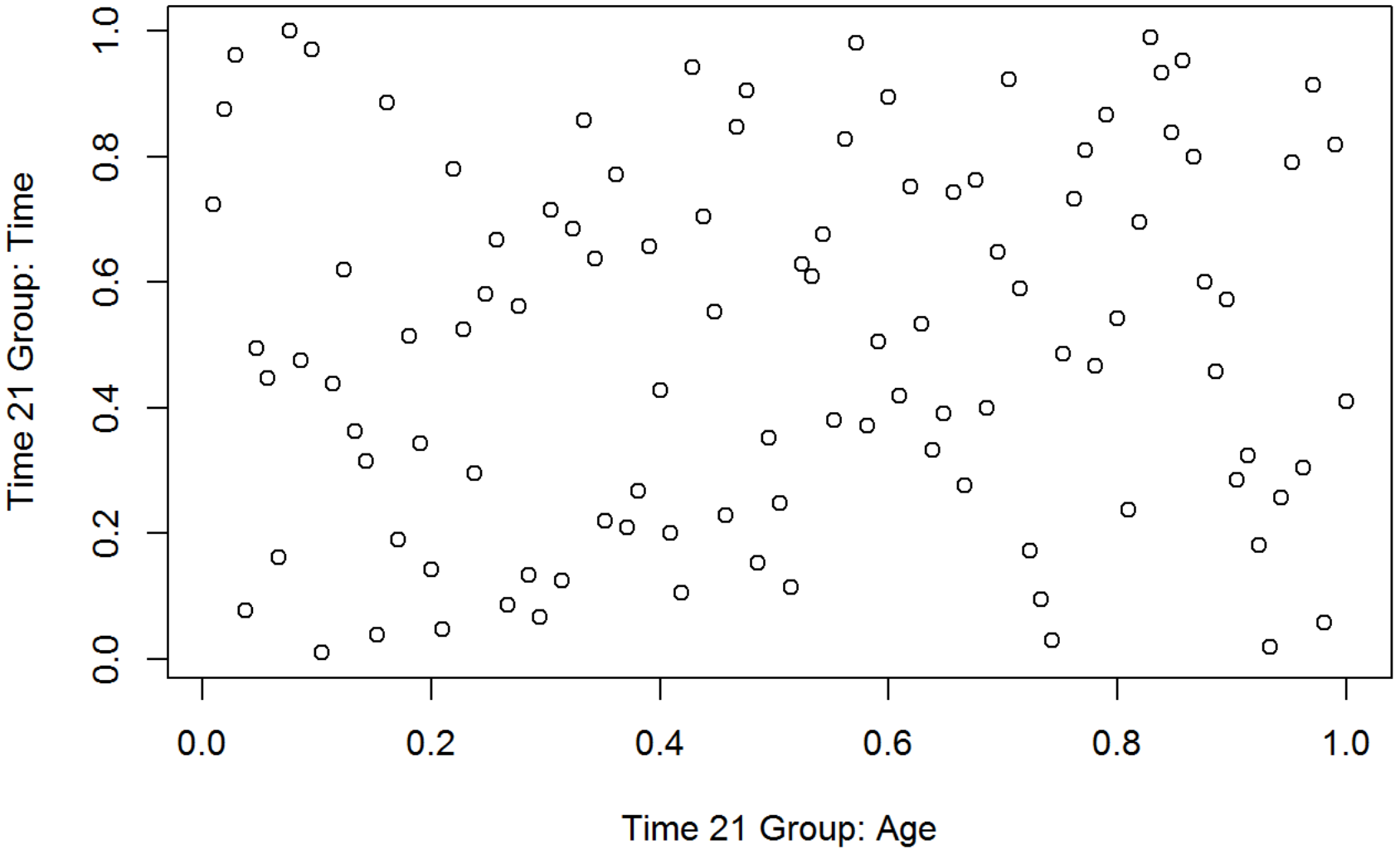
```
## [1] -0.1257
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

Correlation coefficient is slightly negative, which I guessed when first looking at the scatterplot earlier. However, it is still near 0, and the copula looks pretty uniform.

```
#Grouped.Data.Age.45.Time.23
plot(rank(na.omit(Grouped.Data.Age.45.Time.23[,1]))/length(na.omit(Grouped.Data.Age.45.Time.23[,1])),
     rank(na.omit(Grouped.Data.Age.45.Time.23[,2]))/length(na.omit(Grouped.Data.Age.45.Time.23[,2])),
     xlab="Time 21 Group: Age",ylab="Time 21 Group: Time")
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com




```
cor(na.omit(Grouped.Data.Age.45.Time.23),method="spearman")[1,2]
```

```
## [1] 0.08716
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

Copula again looks pretty uniform, so the evidence for dependence is low. Correlation coefficient is 0.08 which is also not indicitive of dependence.

Use copula to fit Gaussian copula to the groups Age.25.Time.23 and Age.45.Time.21.

```
library(copula)

data.to.copula<- pobs(na.omit(Grouped.Data.Age.25.Time.23), ties.method = "average")
Gaussian.Copula.Age.25.Time.23.fit<-fitCopula(normalCopula(), data.to.copula, method="ml")
pobs(na.omit(Grouped.Data.Age.25.Time.23), ties.method = "average")
```

```
##           Age    Time
##  [1,] 0.35294 0.82353
##  [2,] 0.82353 0.76471
##  [3,] 0.11765 0.05882
##  [4,] 0.29412 0.11765
##  [5,] 0.70588 0.17647
##  [6,] 0.94118 0.29412
##  [7,] 0.52941 0.58824
##  [8,] 0.17647 0.47059
##  [9,] 0.47059 0.41176
## [10,] 0.58824 0.64706
## [11,] 0.64706 0.35294
## [12,] 0.23529 0.23529
## [13,] 0.05882 0.70588
## [14,] 0.41176 0.52941
## [15,] 0.76471 0.94118
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
## [16,] 0.88235 0.88235
```

```
data.to.copula<- pobs(na.omit(Grouped.Data.Age.45.Time.21), ties.method = "average")
Gaussian.Copula.Age.45.Time.21.fit<-fitCopula(normalCopula(), data.to.copula, method="ml")
```

Compare the correlations of the parametric models for both groups with Spearman correlations estimated earlier

```
Gaussian.Copula.Age.25.Time.23.fit
```

```
## fitCopula() estimation based on 'maximum likelihood'
## and a sample of size 16.
##       Estimate Std. Error z value Pr(>|z|)
## rho.1    0.441      0.223    1.97    0.048 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## The maximized loglikelihood is  0.992
## Optimization converged
## Number of loglikelihood evaluations:
## function gradient
##       22        6
```

```
cor(na.omit(Grouped.Data.Age.25.Time.23),method="spearman")[1,2]
```

```
## [1] 0.3176
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
Gaussian.Copula.Age.45.Time.21.fit
```

```
## fitCopula() estimation based on 'maximum likelihood'
## and a sample of size 32.
##       Estimate Std. Error z value Pr(>|z|)
## rho.1   -0.265      0.189    -1.4     0.16
## The maximized loglikelihood is  0.789
## Optimization converged
## Number of loglikelihood evaluations:
## function gradient
##       25        5
```

```
cor(na.omit(Grouped.Data.Age.45.Time.21),method="spearman")[1,2]
```

```
## [1] -0.1257
```

The Rho parameter is further away from zero than the spearman correlation coefficients. The parametric models indicate that the data is more correlated than the spearman coefficients would.

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com