LNL_HW_week3

Patrick Kelly

Friday, January 31, 2015

Week 3 Homework Assignment: Analysis of Non-Linear Dependencies

1. Description of the data

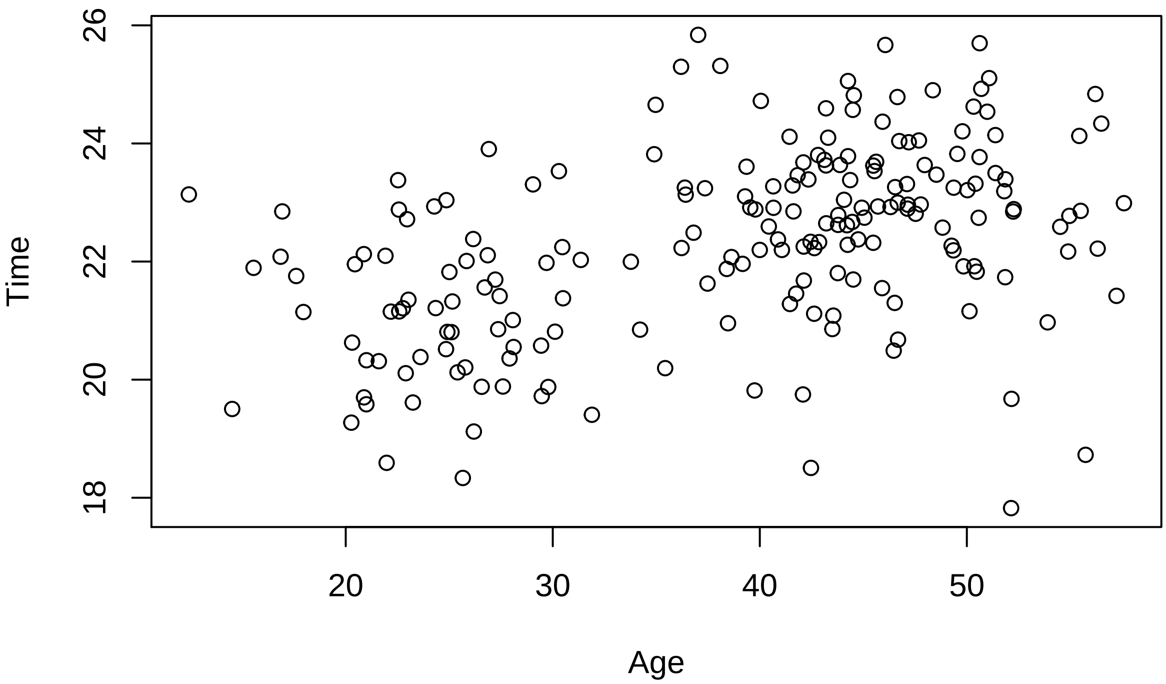
Data in the file Week3_Homework_Project_Data.csv contain observations of time during the day when people watch TV (are logged in to an internet site, active online, etc.) and their age.

Read the data and explore them.

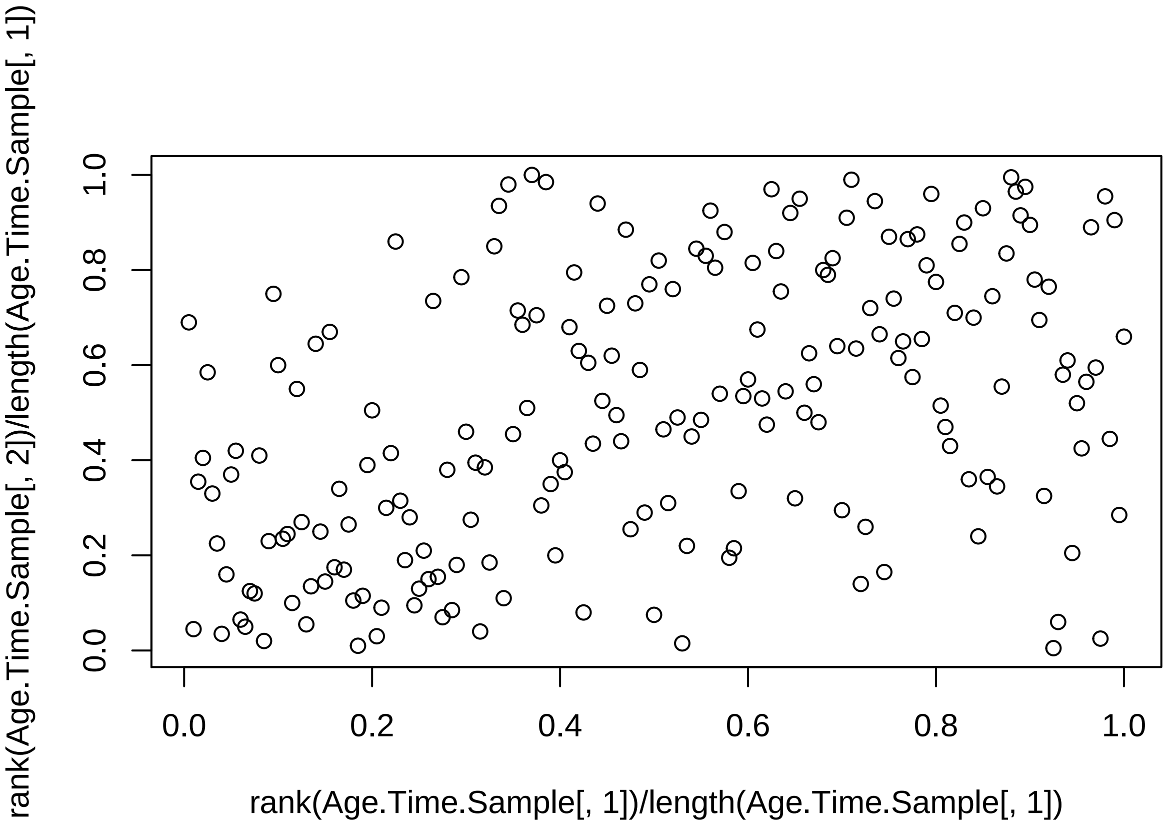
```
Age.Time.Sample<-read.csv(file="C:/Users/Patrick/SkyDrive/Documents/Education/UChicago/Linear_NonLinear/

Age.Time.Sample<-as.matrix(Age.Time.Sample)
Age.Time.Sample[1:10,]
```

```
##             Age     Time
##  [1,] 31.35366 22.02776
##  [2,] 27.43543 21.41511
##  [3,] 22.53053 23.37803
##  [4,] 21.00111 20.32848
##  [5,] 22.58267 21.15540
##  [6,] 29.04533 23.30662
##  [7,] 26.71073 21.56138
##  [8,] 23.61076 20.38302
##  [9,] 28.07124 21.00864
## [10,] 28.10947 20.55258
```

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



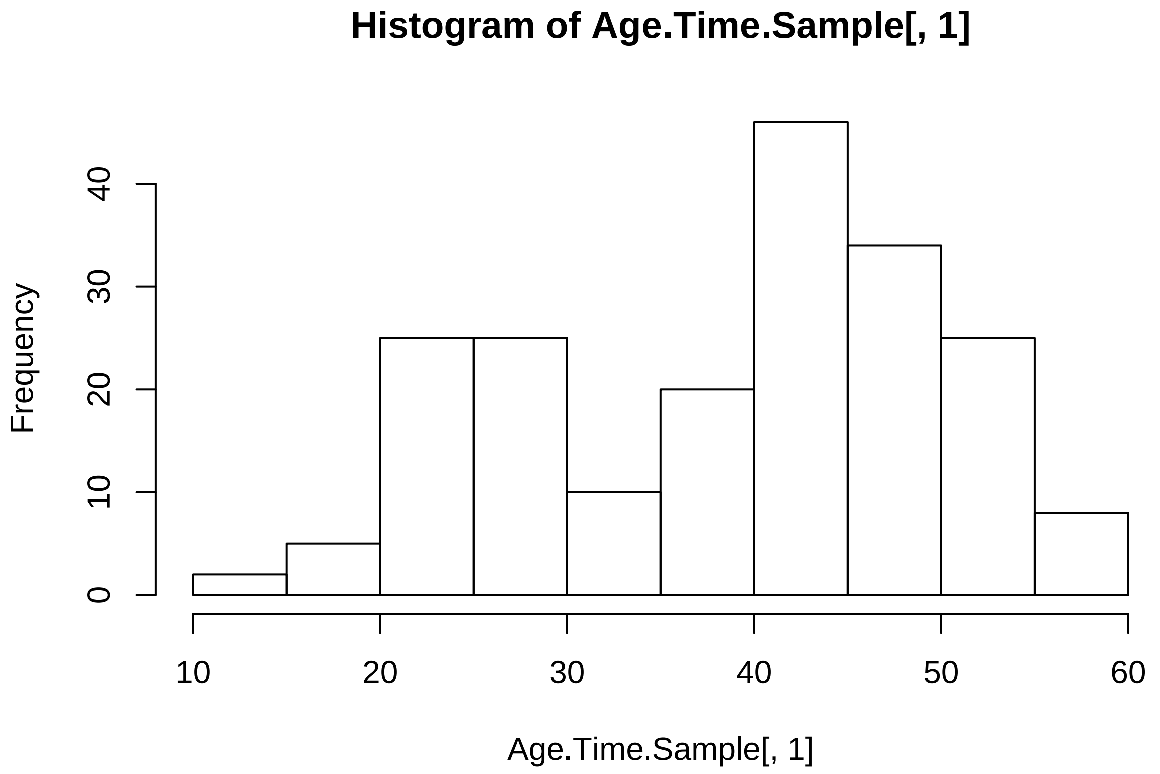
```
plot(rank(Age.Time.Sample[,1])/length(Age.Time.Sample[,1]),rank(Age.Time.Sample[,2])/length(Age.Time.Sa
```



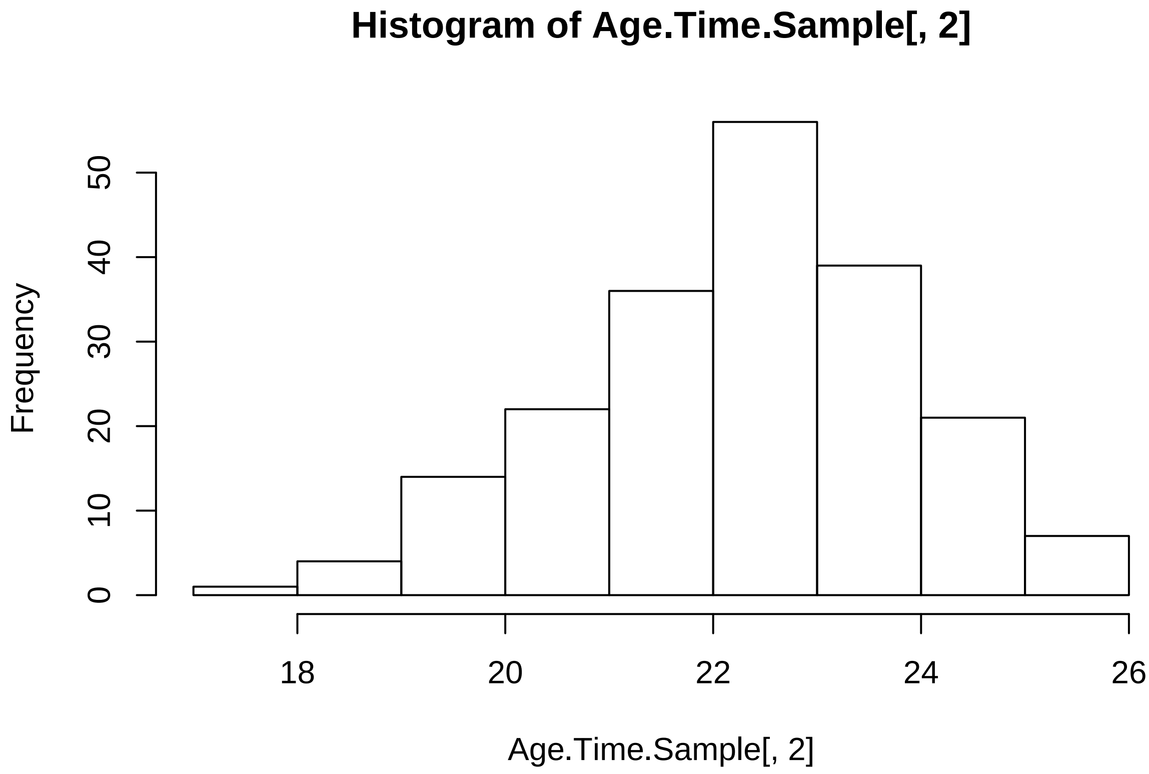
```r
c(cor(Age.Time.Sample)[1,2],cor(Age.Time.Sample)[1,2]^2)
```

```
## [1] 0.4359003 0.1900090
```

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



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



Interpret the initial observations:

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

There appear to be two distinct clusters in the data - One that is towards the upper right section of the plot (higher age and time values) and another that is more in the lower left section of the plot (lower age and time values).

2. What does the empirical copula suggest?

The empircal copula suggests that the clusters observed in the previous plot may not be as definitive as orginally assumed. There does appear to be some loose linear relationship in the data (positive slope).

3. How significant is the amount of correlation?

The correlation moderately exists. The Pearson Correlation Coefficient is .436, definitely suggesting a moderate relationship between the two variables.

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

The first histogram, Age, indicates that there may be two (gaussian) distributions in the data - one centered around ~25 and the other centered around ~43.

The second histogram, Time, shows a left skew. Although this distribution does not directly suggest two distributions (like the previous histogram) it is possible that two seperate distributions also exist and this data with means that are closer together. This is a worthy hypothesis that is worth invetigation given the bimodal distribution of the Age histogram.

Clustering of the data.

Find possible clusters in both variables.

Use `Mclust()` from `Mclust` to find clusters in the age component and time component. Define the `Mclust` object `Age.Clusters` and explore the components of it:

```
library(mclust)
```

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

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

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
```

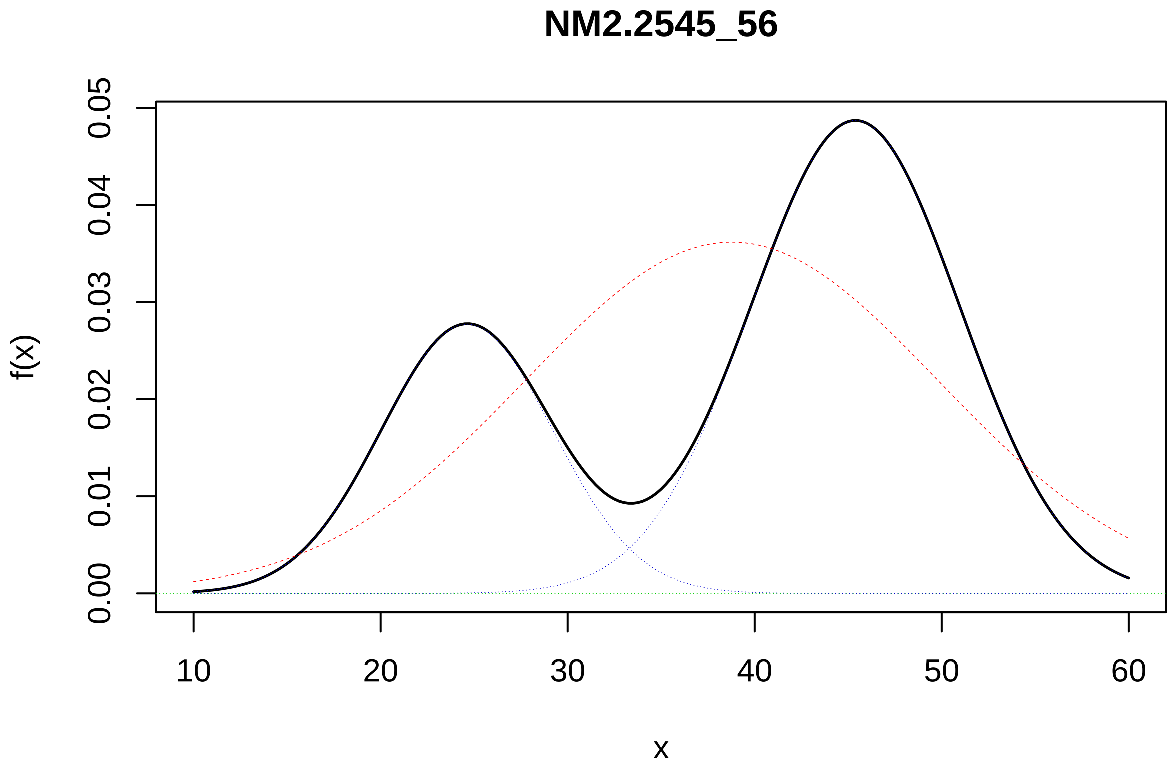
```
Age.Clusters$param
```

```
## $Vinv
## NULL
##
## $pro
## [1] 0.3187374 0.6812626
##
## $mean
##        1        2
## 24.61792 45.38996
##
## $variance
## $variance$modelName
## [1] "V"
##
## $variance$d
## [1] 1
##
## $variance$G
## [1] 2
##
## $variance$sigmasq
## [1] 21.02398 31.12624
##
## $variance$scale
## [1] 21.02398 31.12624
```

```
Age.Clusters.Parameters<-rbind(mu=Age.Clusters$param$mean,sigma=sqrt(Age.Clusters$param$variance$sigmas
#cbind(Mixing.Sequence,Age.Clusters$classification-1,Age.Clusters$uncertainty)
```

Use `norMix()` from `nor1mix` to analyze the mixed Gaussian models classified by `Mclust()`. Define the object `Classified.Mix.Model.Age` using `norMix()` and plot the densities of the normal mix.

```
library(nor1mix)

Classified.Mix.Model.Age <- norMix(mu = Age.Clusters.Parameters[1,], sigma = Age.Clusters.Parameters[2,

plot(Classified.Mix.Model.Age,xout=seq(from=10,to=60,by=.25),p.norm=TRUE,p.comp=TRUE)
```



Define the `Mclust` object `Time.Clusters` and explore the components of it:

```
#force the number of mix components to be 2
Time.Clusters <- Mclust(Age.Time.Sample[,2], G = 2,modelNames = "V")

Time.Clusters$G
```

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

```
Time.Clusters$param
```

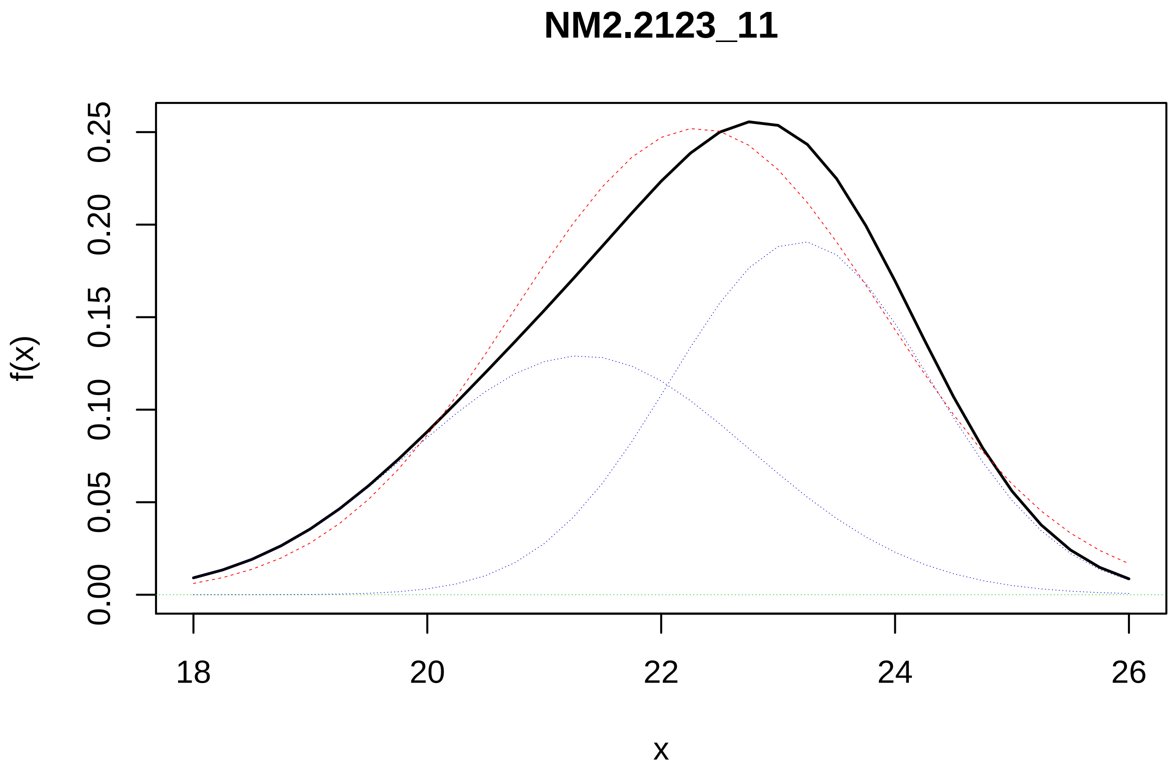
```
## $Vinv
## NULL
##
## $pro
## [1] 0.4668235 0.5331765
##
## $mean
##        1        2
## 21.31969 23.19007
##
## $variance
## $variance$modelName
## [1] "V"
##
## $variance$d
## [1] 1
##
## $variance$G
## [1] 2
##
## $variance$sigmasq
## [1] 2.079294 1.241448
##
## $variance$scale
## [1] 2.079294 1.241448
```

```
Time.Clusters.Parameters<-rbind(mu=Time.Clusters$param$mean,sigma=sqrt(Time.Clusters$param$variance$sig
```

Again, define the object `Classified.Mix.Model.Time` using `norMix` for analyzing the time component.

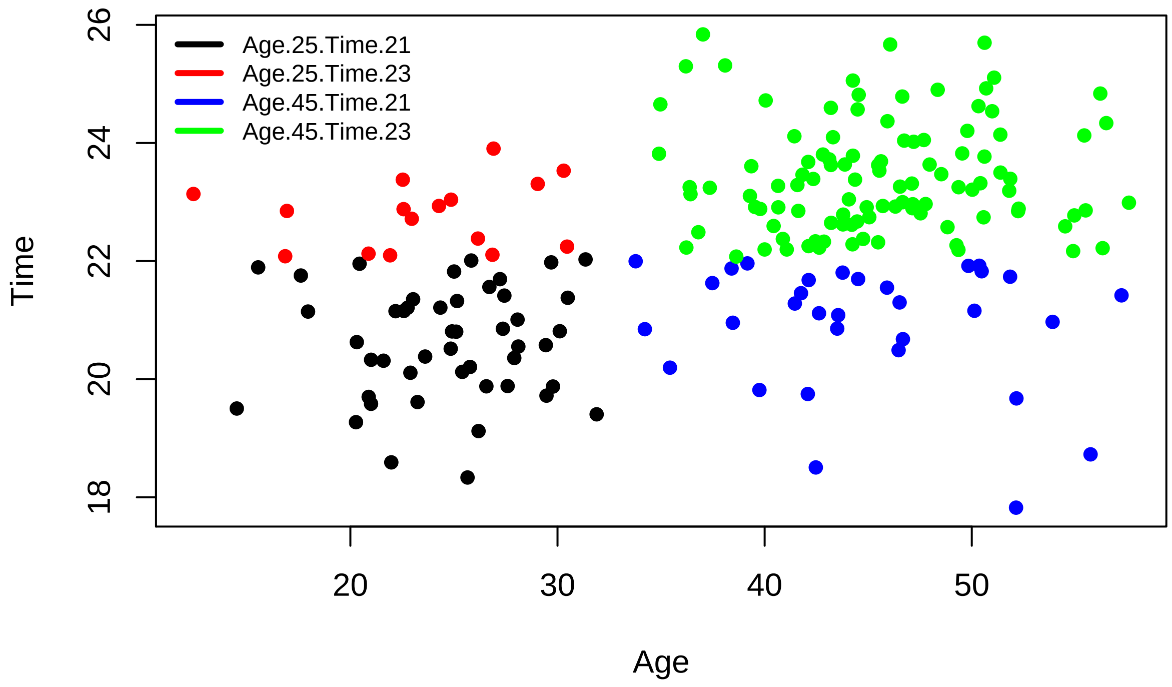
Plot the densities of the mix.

```
Classified.Mix.Model.Time <- norMix(mu = Time.Clusters.Parameters[1,], sigma = Time.Clusters.Parameters

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



Separate the samples into clusters and explore their dependencies

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



Now we clearly see existance of clusters.

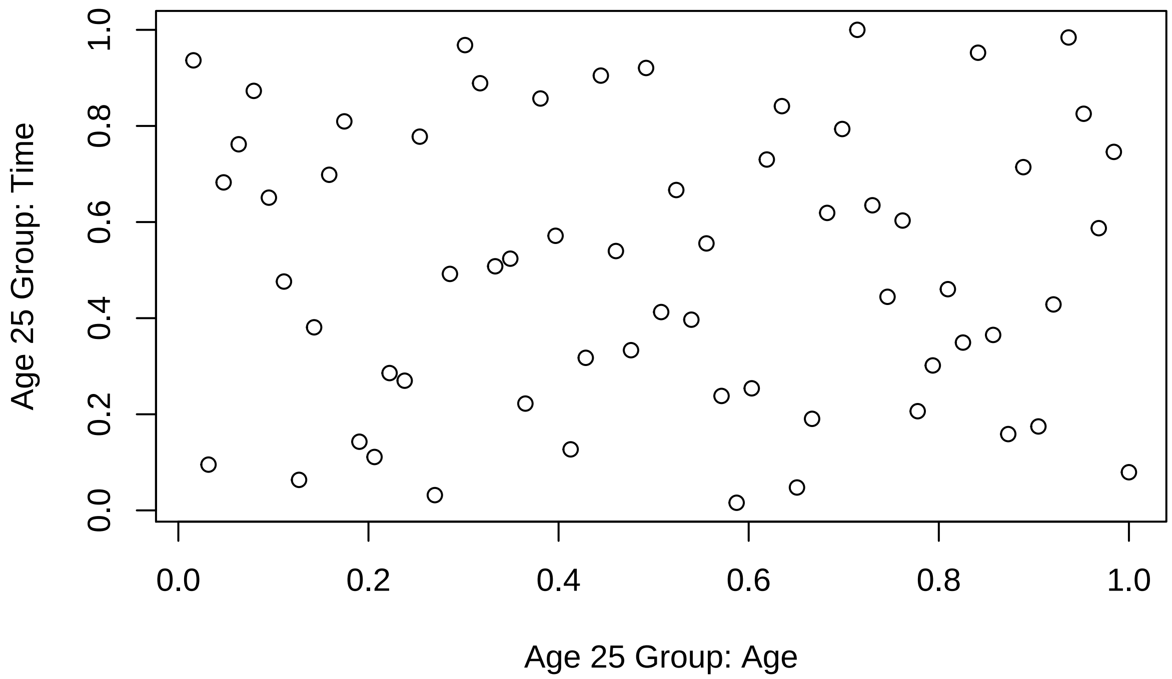
Interpret the results: what dependencies do you see on the chart?

The dependencies observed on the chart (seperated into clusters) above appear to be less than what we previously observed (all observations together). However, I can still see a slight positive relationship in the upper right cluster (green), slight negative in the lower right cluster (blue), and maybe a slight positive relationship in the lower left cluster (black). The only other cluster in the upper left (red) appears to maybe have a slightly positive relationship but there is not at much data as other clusters.

Group the samples by age and by time and explore the dependencies within groups.

Group by age.

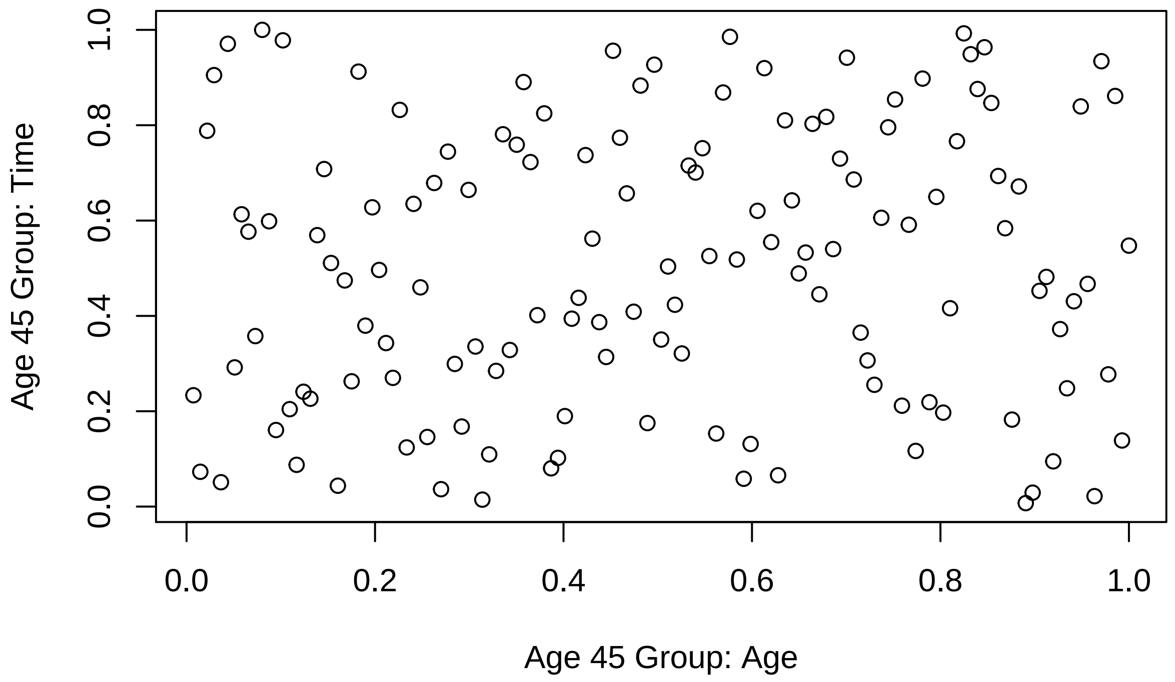
```
#Group by age
Grouped.Data.Age.25<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Age.25[Age.Clusters$classification==1,]<-Age.Time.Sample[Age.Clusters$classification==1,]
Grouped.Data.Age.45<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Age.45[Age.Clusters$classification==2,]<-Age.Time.Sample[Age.Clusters$classification==2,]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



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

```
## [1] -0.01128072
```

```
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")
```

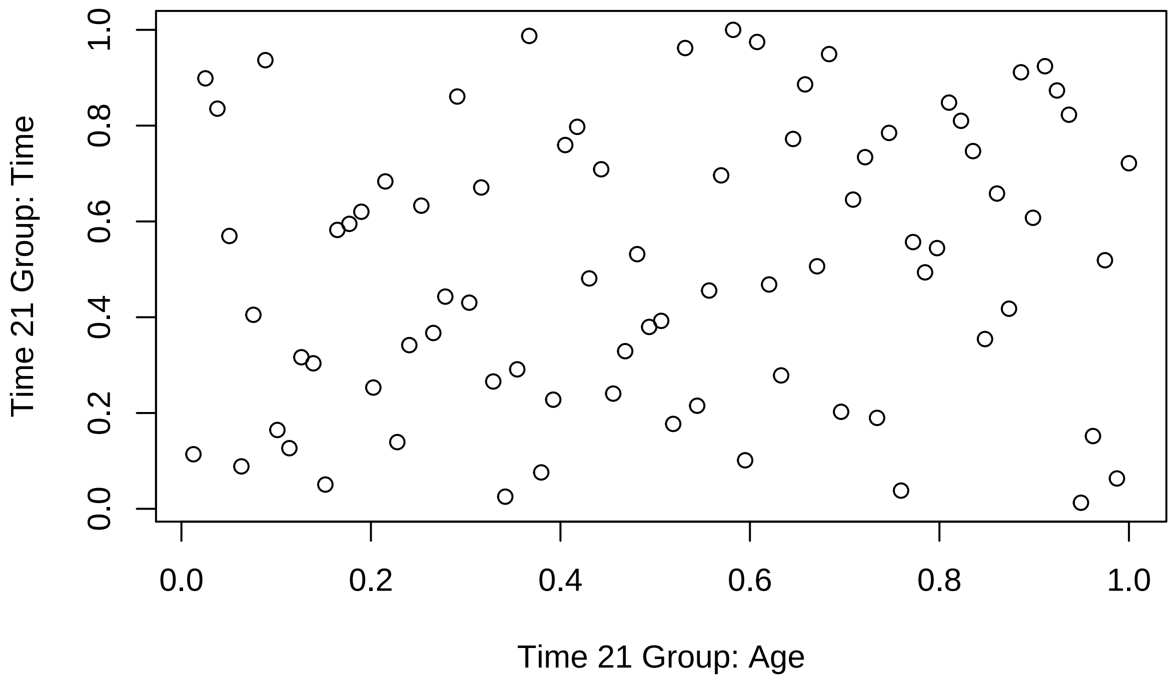


```
cor(na.omit(Grouped.Data.Age.45),method="spearman")[1,2]
```

```
## [1] 0.08611179
```

Group by time

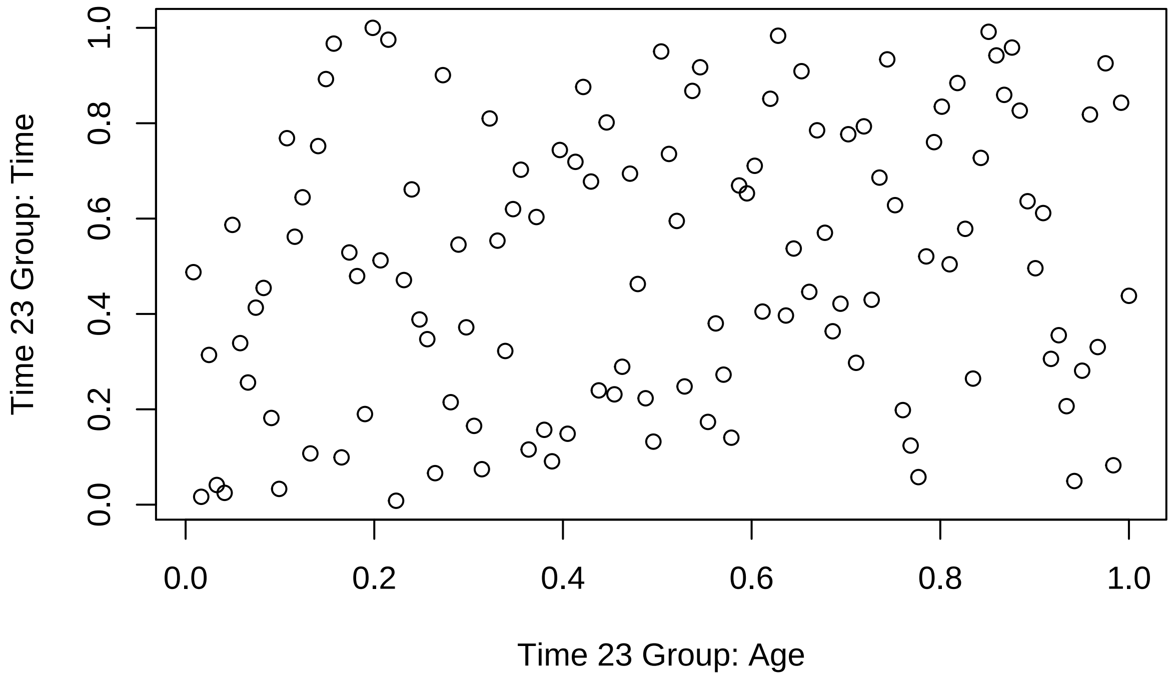
```
#Group by Time
Grouped.Data.Time.21<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Time.21[Time.Clusters$classification==1,]<-
  Age.Time.Sample[Time.Clusters$classification==1,]
Grouped.Data.Time.23<-cbind(Age=rep(NA,200),Time=rep(NA,200))
Grouped.Data.Time.23[Time.Clusters$classification==2,]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



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

```
## [1] 0.1722736
```

```
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")
```



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

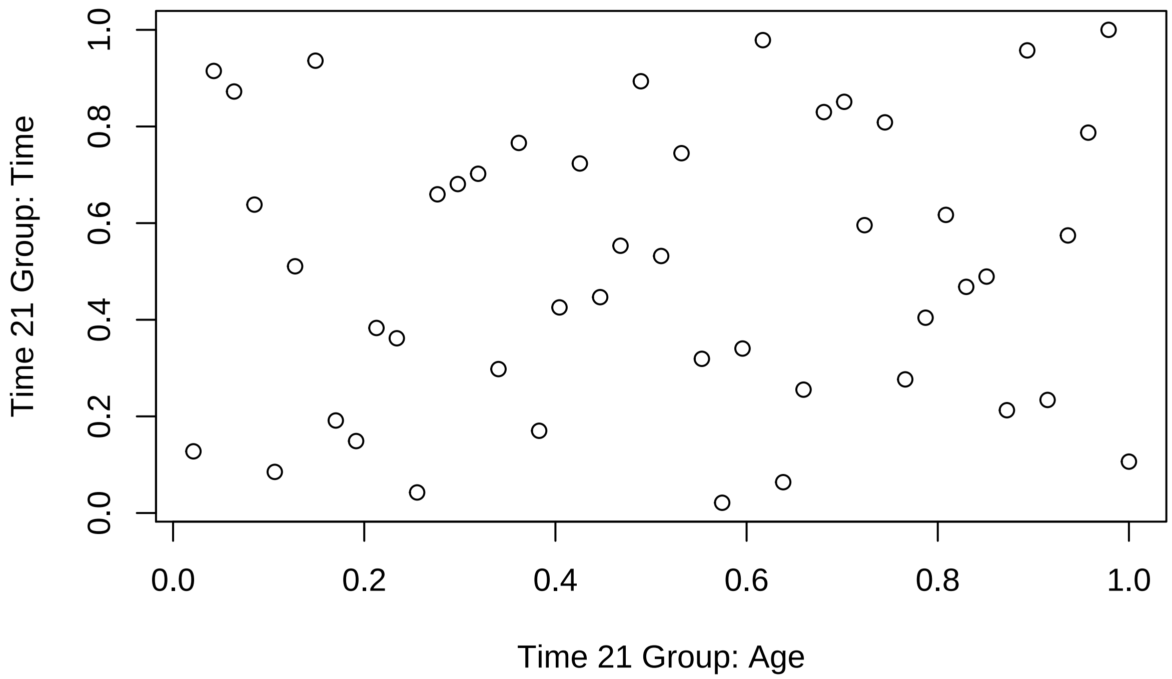
```
## [1] 0.2115364
```

What do you conclude from the results grouped by age and by time?

Through observing the correlations of the different groups by age and by time seperately, it is apparent that the correlations (ie relationships) are considerably lower compared to the unsegmented data. Both of the correlations of the groups split by age are low (< 0.1), while both of the correlations of the groups split by time are positive and moderately low (~ 0.2).

Group by age and by time.

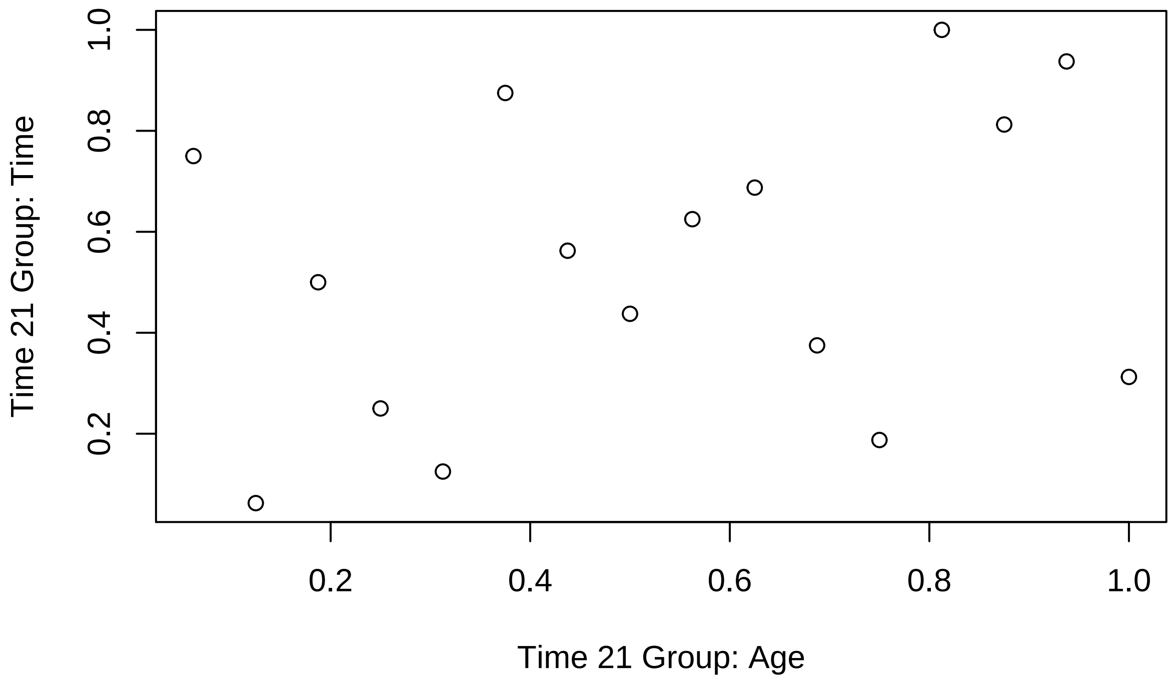
```
#Group by Age and Time
#Grouped.Data.Age.25.Time.21
plot(rank(na.omit(Grouped.Data.Age.25.Time.21[,1]))/length(na.omit(Grouped.Data.Age.25.Time.21[,1])),
     rank(na.omit(Grouped.Data.Age.25.Time.21[,2]))/length(na.omit(Grouped.Data.Age.25.Time.21[,2])),
     xlab="Time 21 Group: Age",ylab="Time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



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

```
## [1] 0.0793247
```

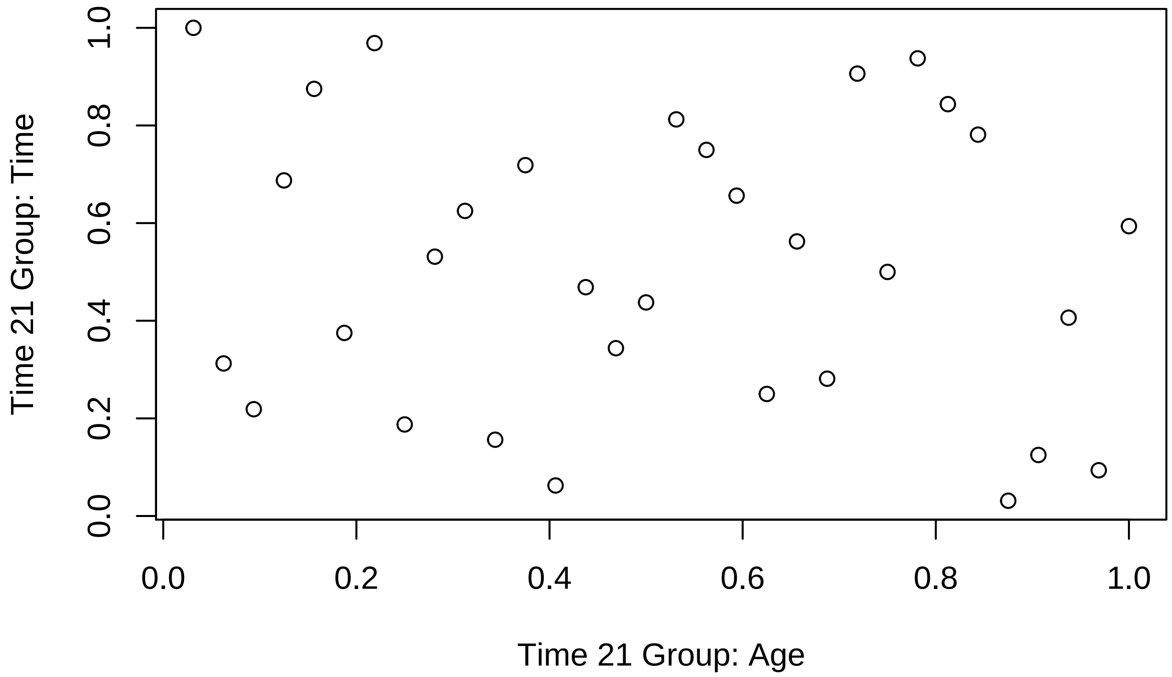
```
#Grouped.Data.Age.25.Time.23
plot(rank(na.omit(Grouped.Data.Age.25.Time.23[,1]))/length(na.omit(Grouped.Data.Age.25.Time.23[,1])),
     rank(na.omit(Grouped.Data.Age.25.Time.23[,2]))/length(na.omit(Grouped.Data.Age.25.Time.23[,2])),
     xlab="Time 21 Group: Age",ylab="Time 21 Group: Time")
```



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

```
## [1] 0.3176471
```

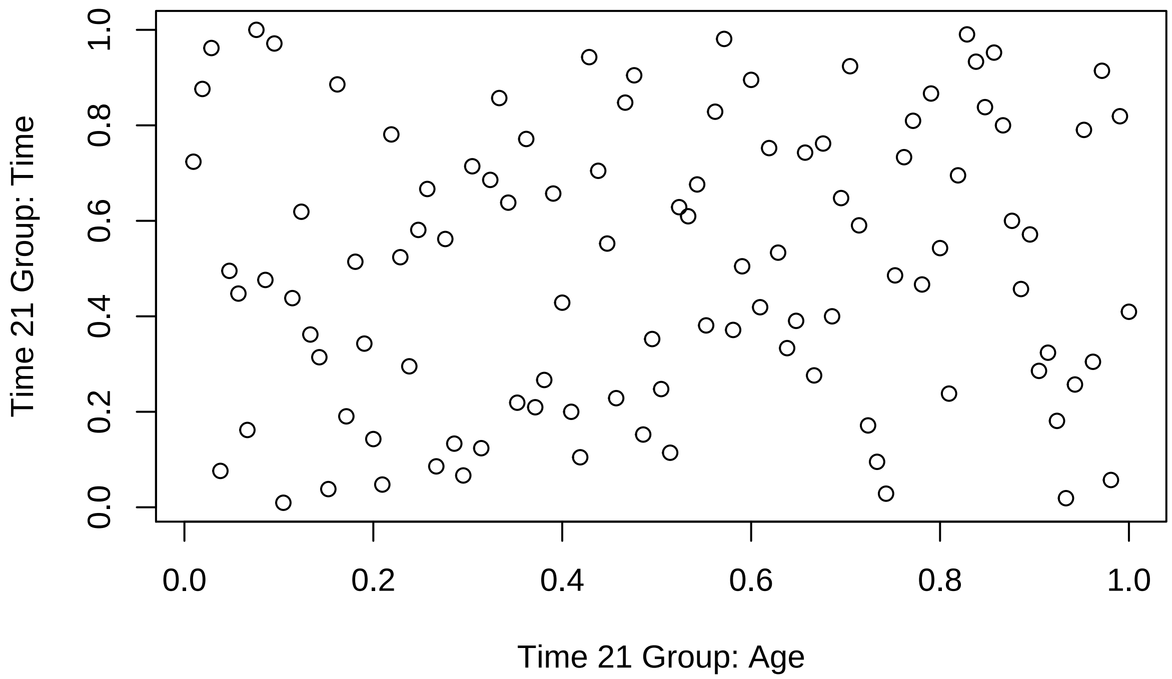
```
#Grouped.Data.Age.45.Time.21
plot(rank(na.omit(Grouped.Data.Age.45.Time.21[,1]))/length(na.omit(Grouped.Data.Age.45.Time.21[,1])),
     rank(na.omit(Grouped.Data.Age.45.Time.21[,2]))/length(na.omit(Grouped.Data.Age.45.Time.21[,2])),
     xlab="Time 21 Group: Age",ylab="Time 21 Group: Time")
```



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

```
## [1] -0.1257331
```

```
#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")
```



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

```
## [1] 0.0871553
```

Interpret the results of dependency analysis by age and time simultaneously

When seperating by age and time the dependencies / relationships vary slightly but are generally moderate to low. All of the clusters other than the upper left cluster (age 25, time 23) have an absolutely value of ~ 0.1. The cluster of age 25, time 23 still only has a moderate dependency of .31.

So overall the relationships and dependencies for the different age and time clusters are all low / moderate - definitely lower than the dependency found in within the unclustered data.

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

Use `normalCopula()` to define the copula objects, then use `fitCopula()` to fit copulas. Use `pobs()` to create pseudo data that `fitCopula()` needs.

Create the object `Gaussian.Copula.Age.25.Time.23.fit` of Gaussian copula fit to the group of age 25 and time 23 and explore it.

```
library(copula)

cop1 <- normalCopula()

p1 <- pobs(na.omit(Grouped.Data.Age.25.Time.23), ties.method = "average")
```

```
Gaussian.Copula.Age.25.Time.23.fit <- fitCopula(cop1,p1)

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

```
## fitCopula() estimation based on 'maximum pseudo-likelihood'
## and a sample of size 16.
##       Estimate Std. Error z value Pr(>|z|)
## rho.1   0.4412     0.1821   2.423   0.0154 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## The maximized loglikelihood is  0.9922
## Optimization converged
## Number of loglikelihood evaluations:
## function gradient
##       22        6
```

Repeat the analysis for the group age 45 and time 21.

```
cop2 <- normalCopula()
p2 <- pobs(na.omit(Grouped.Data.Age.45.Time.21), ties.method = "average")
Gaussian.Copula.Age.45.Time.21.fit <- fitCopula(cop2,p2)
Gaussian.Copula.Age.45.Time.21.fit
```

```
## fitCopula() estimation based on 'maximum pseudo-likelihood'
## and a sample of size 32.
##       Estimate Std. Error z value Pr(>|z|)
## rho.1  -0.2647     0.2031  -1.303    0.192
## The maximized loglikelihood is  0.7887
## Optimization converged
## Number of loglikelihood evaluations:
## function gradient
##       25        5
```

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

In comparing these two clusters' correlations from the parametic models and Spearman correlation coefficients, the corrlations from the parametric models both have the same direction as their corresponding Spearman values with higher absolute values. Therefore correlations of the parametric models suggest a stronger relationship / dependency based on the uniform marginal distributions from the copula.