SDS363 Problem Set 2

Liana Wang, Yavuz Ramiz Çolak, Ryo Tamaki, David Lieberman

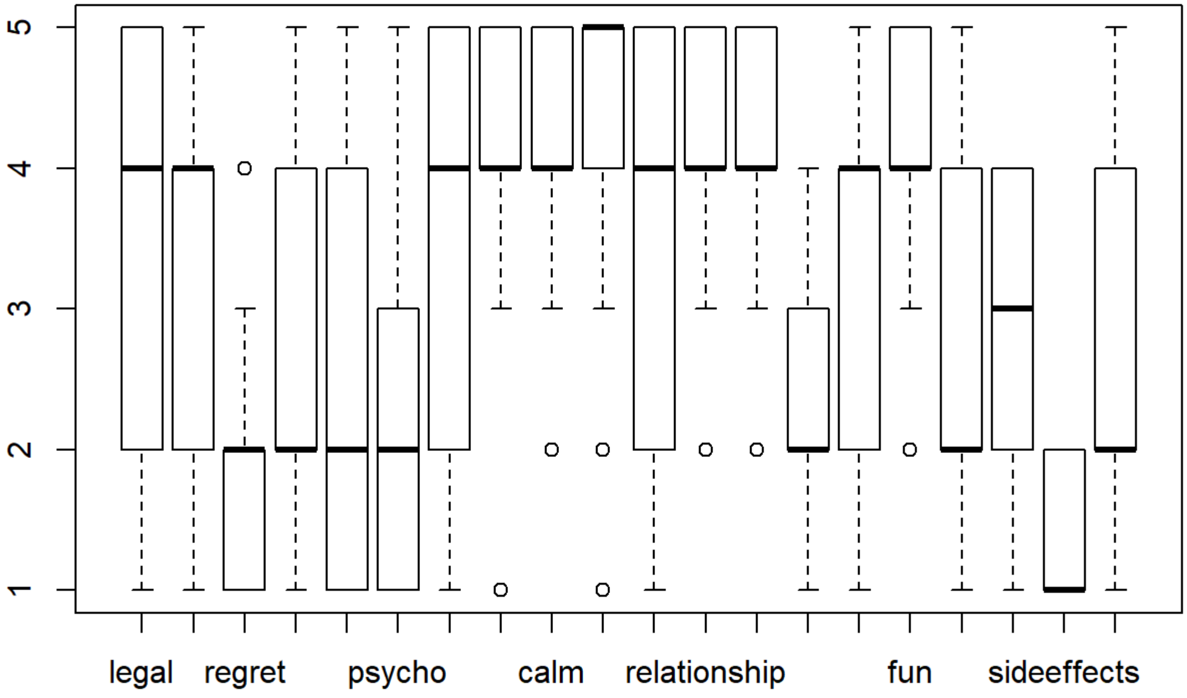
2/11/2019

PROBLEM 1

1). First, discuss whether your data seems to have a multivariate normal distribution. Make univariate plots (boxplots, normal quantile plots as appropriate). Then make transformations as appropriate. You do NOT need to turn all this in, but describe what you did. THEN make a chi-square quantile plot of the data. Turn in your chi-square quantile plot as appropriate and comment on what you see. NOTE that multivariate normality is NOT a requirement for PCA to work!

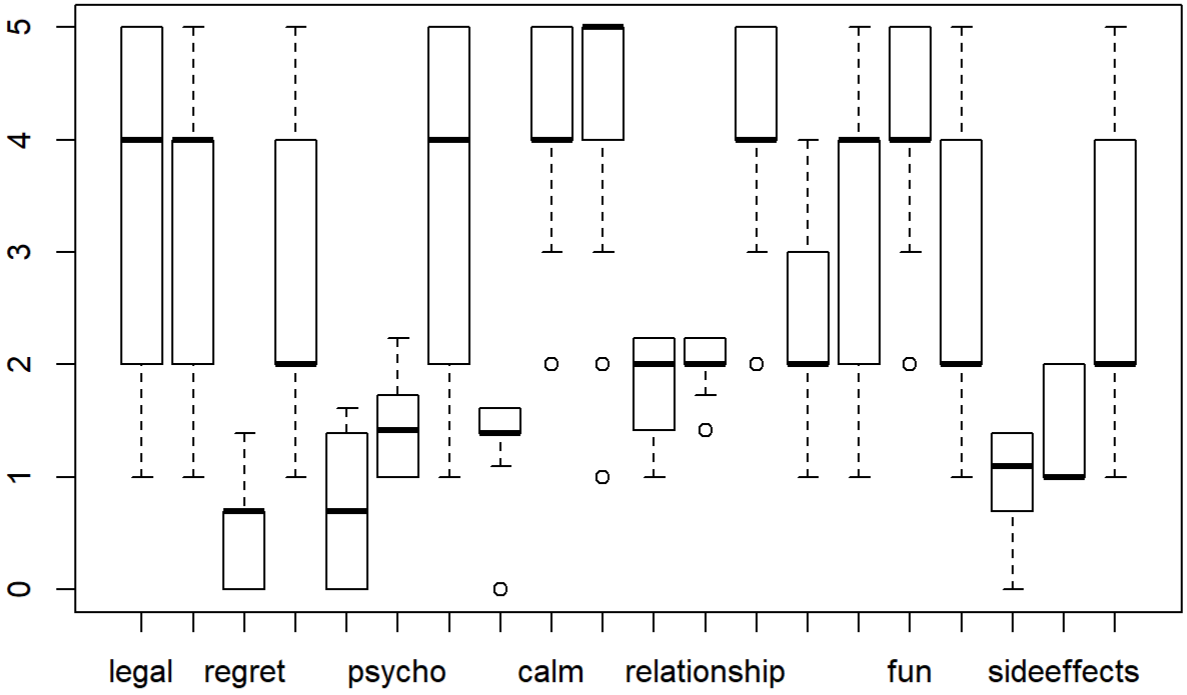
```
data <- read.csv("https://pastebin.com/raw/z6pgskch")
data <- data %>% drop_na()

boxplot(data)
```



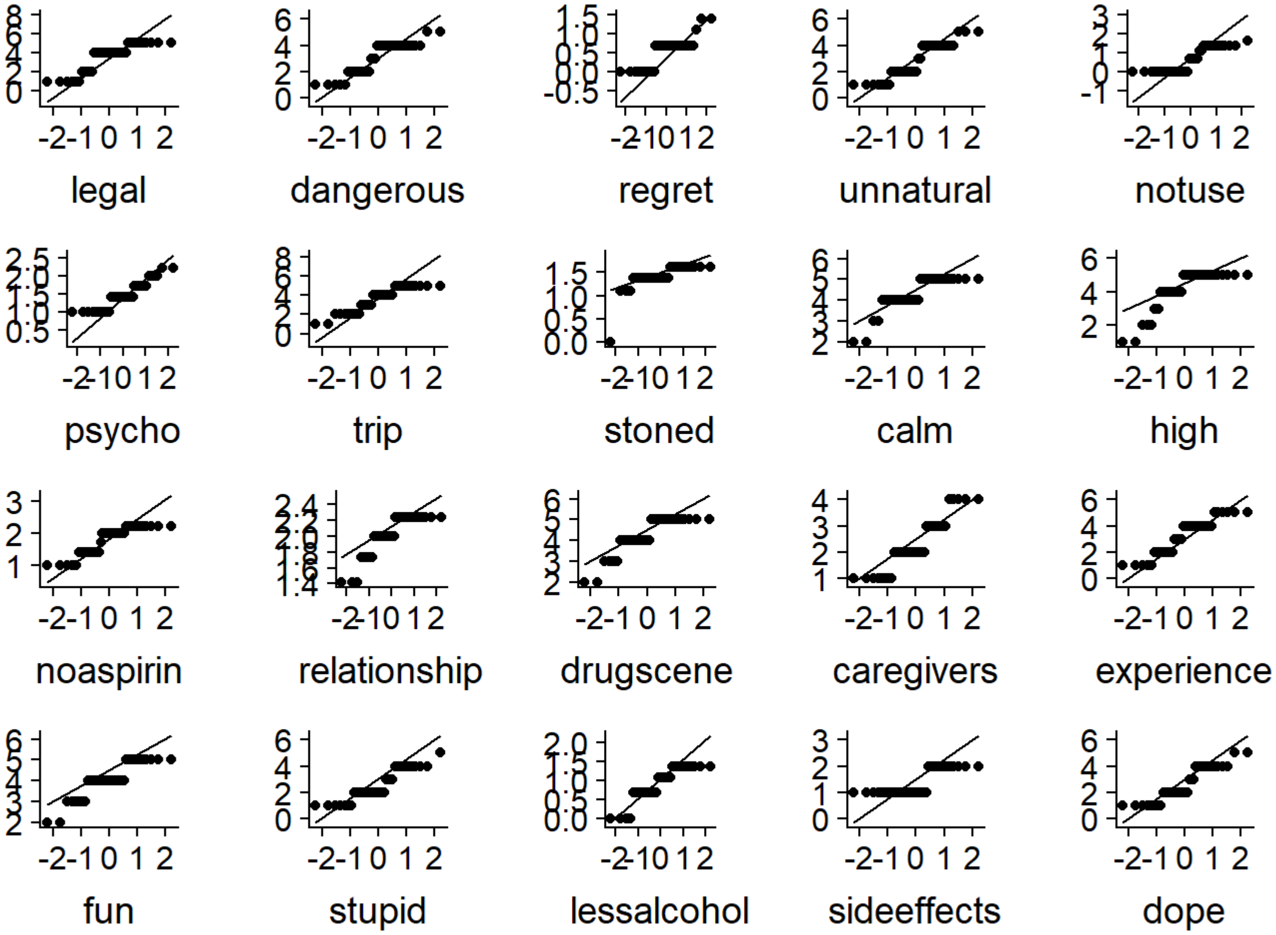
```
data_transformed <- data
data_transformed$regret <- log(data$regret)
data_transformed$notuse <- log(data$notuse)
data_transformed$psycho <- sign(data$psycho)*sqrt(abs(data$psycho))
data_transformed$stoned <- log(data$stoned)
data_transformed$noaspirin <- sqrt(data$noaspirin)
data_transformed$relationship <- sqrt(data$relationship)
data_transformed$lessalcohol <- log(data$lessalcohol)

boxplot(data_transformed)
```



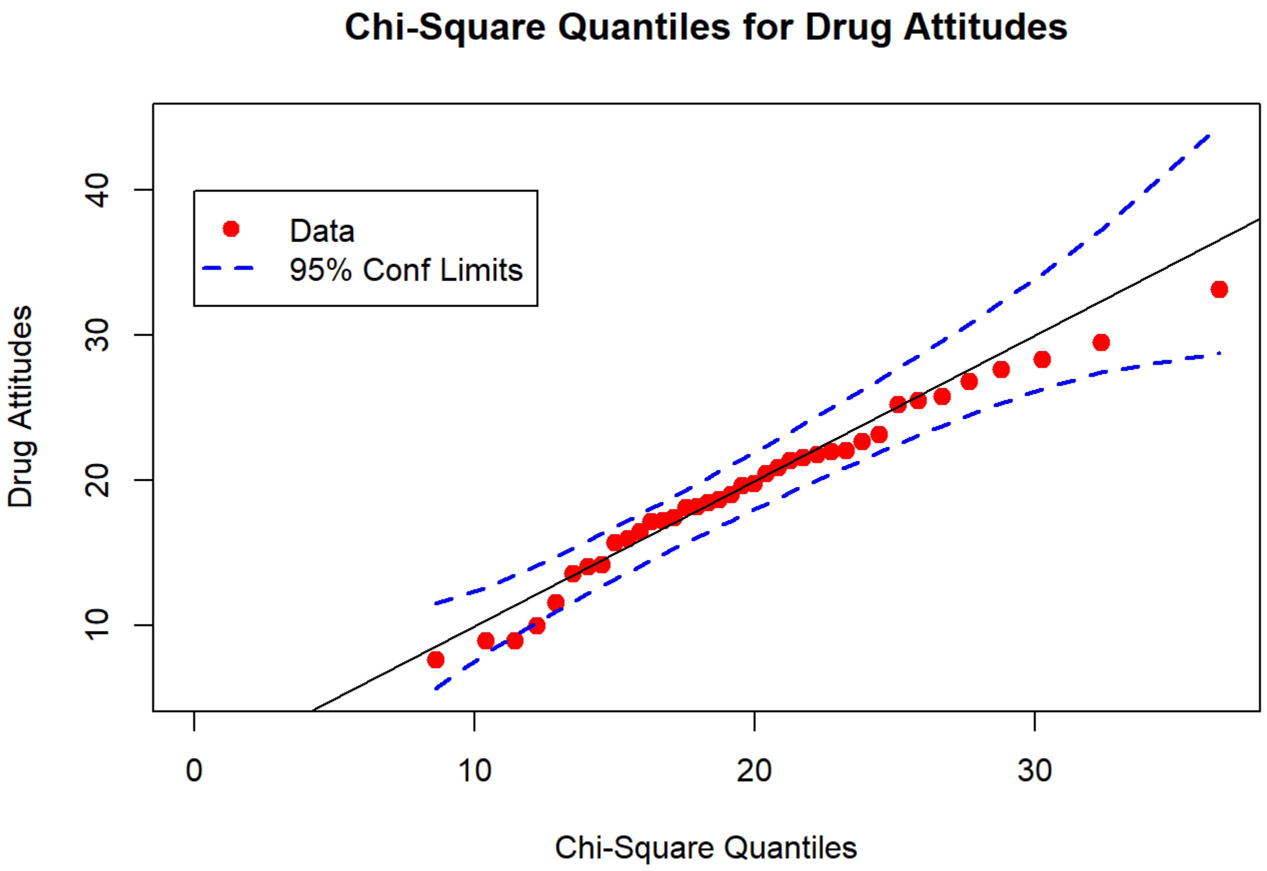
```
QQPlot <- function(x, na.rm = TRUE){
  plots <- list()
  j <- 1
  for (i in names(x)) {
    plots[[i]] <- ggplot(x, aes_string(sample = i)) + stat_qq() + stat_qq_line() + xlab(names(x)
[j]) + ylab("")
    j <- j+1
  }
  plot_grid(plotlist = plots)
}

QQPlot(data_transformed)
```



We first created a boxplot of our data. There is a pretty significant skew on a few of the variables, which we tried to correct through log and square root transformations. After the transformation, we re-created the boxplots. We observed that all variables were distributed more symmetrically and were closer to normal. Furthermore, after transformation, our resulting QQ-plots appeared to fall mostly along the line that would indicate a roughly normal relationship. It is important to note that while some of our QQ plots were partially piece-wise, that was mainly because we included discrete variables. Discrete variables are relevant for our analysis, so we still wanted to include them. Also, our data set is relatively small. With more data points, the QQplots would more closely follow the 45 degree line.

```
CSQPlot<-function(vars,label="Chi-Square Quantile Plot"){
   #usually, vars is xxx$residuals or data from one group and label is for plot
     x<-cov(scale(vars),use="pairwise.complete.obs")
     squares<-sort(diag(as.matrix(scale(vars))%*%solve(x)%*%as.matrix(t(scale(vars)))))
     quantiles<-quantile(squares)
     hspr<-quantiles[4]-quantiles[2]
     cumprob<-c(1:length(vars[,1]))/length(vars[,1])-1/(2*length(vars[,1]))
     degf<-dim(x)[1]
     quants<-qchisq(cumprob,df=degf)
     gval<-(quants**(-1+degf/2))/(exp(quants/2)*gamma(degf/2)*(sqrt(2)**degf))
     scale<-hspr / (qchisq(.75,degf)-qchisq(.25,degf))
     se<-(scale/gval)*sqrt(cumprob*(1-cumprob)/length(squares))
     lower<-quants-2*se
     upper<-quants+2*se
    plot(quants,squares,col='red',pch=19,cex=1.2,xlab="Chi-Square Quantiles",
     ylab=label,main=paste("Chi-Square Quantiles for",label),ylim=range(upper,lower, squares) ,
 xlim=range(c(0,quants)))
    lines(c(0,100),c(0,100),col=1)
    lines(quants,upper,col="blue",lty=2,lwd=2)
    lines(quants,lower,col="blue",lty=2,lwd=2)
    legend(0,range(upper,lower)[2]*.9,c("Data","95% Conf Limits"),lty=c(0,2),col=c("red","blue"
),lwd=c(2,2),
      pch=c(19,NA))
}
CSQPlot(data_transformed,label="Drug Attitudes")
```



Our CSQ plot seems to indicate a roughly normal multivariate distribution, with all data (except for 1 variable, it appears) falling within the 95% confidence interval and a good number of variables falling along the line which indicates normality. The relative skew is, as noted above, likely partially due to the small sample size.

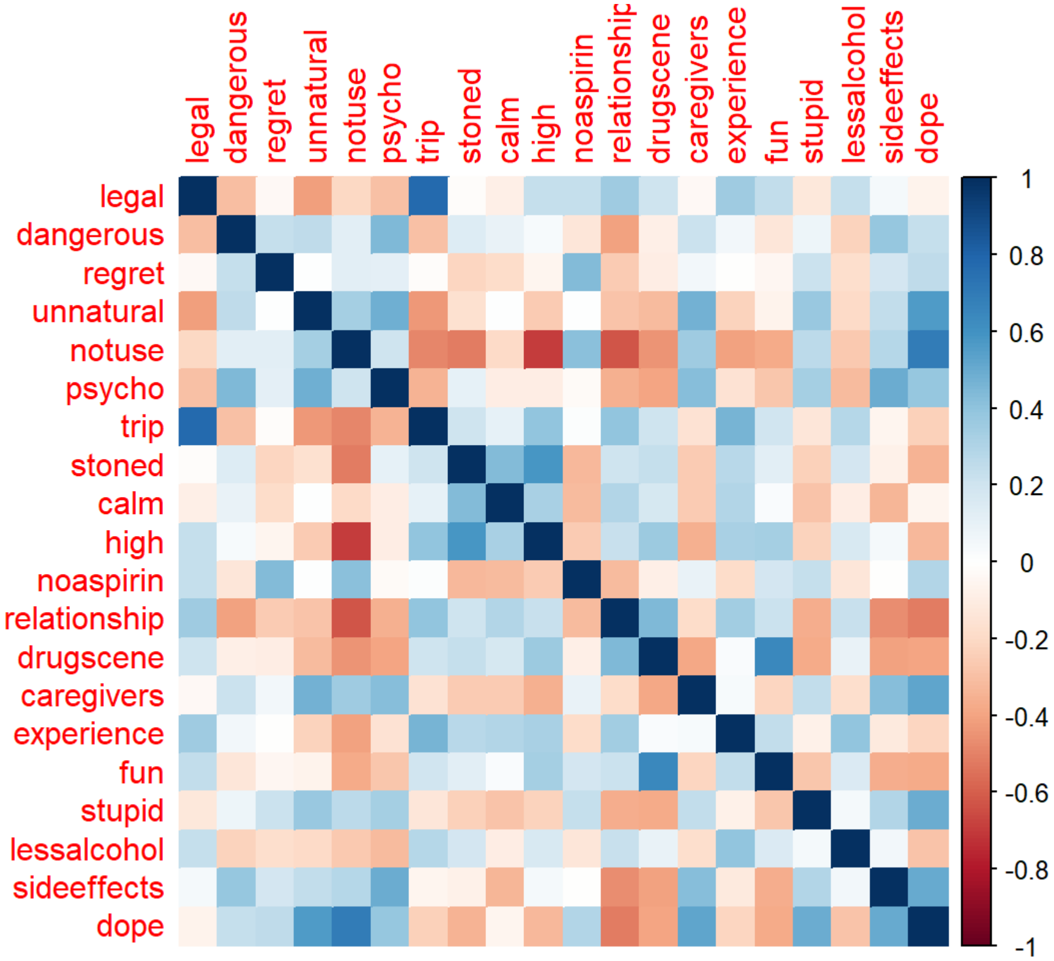
PROBLEM 2

2). Compute the correlation matrix between all variables. Comment on relationships you do/do not observe. Do you think PCA will work well?

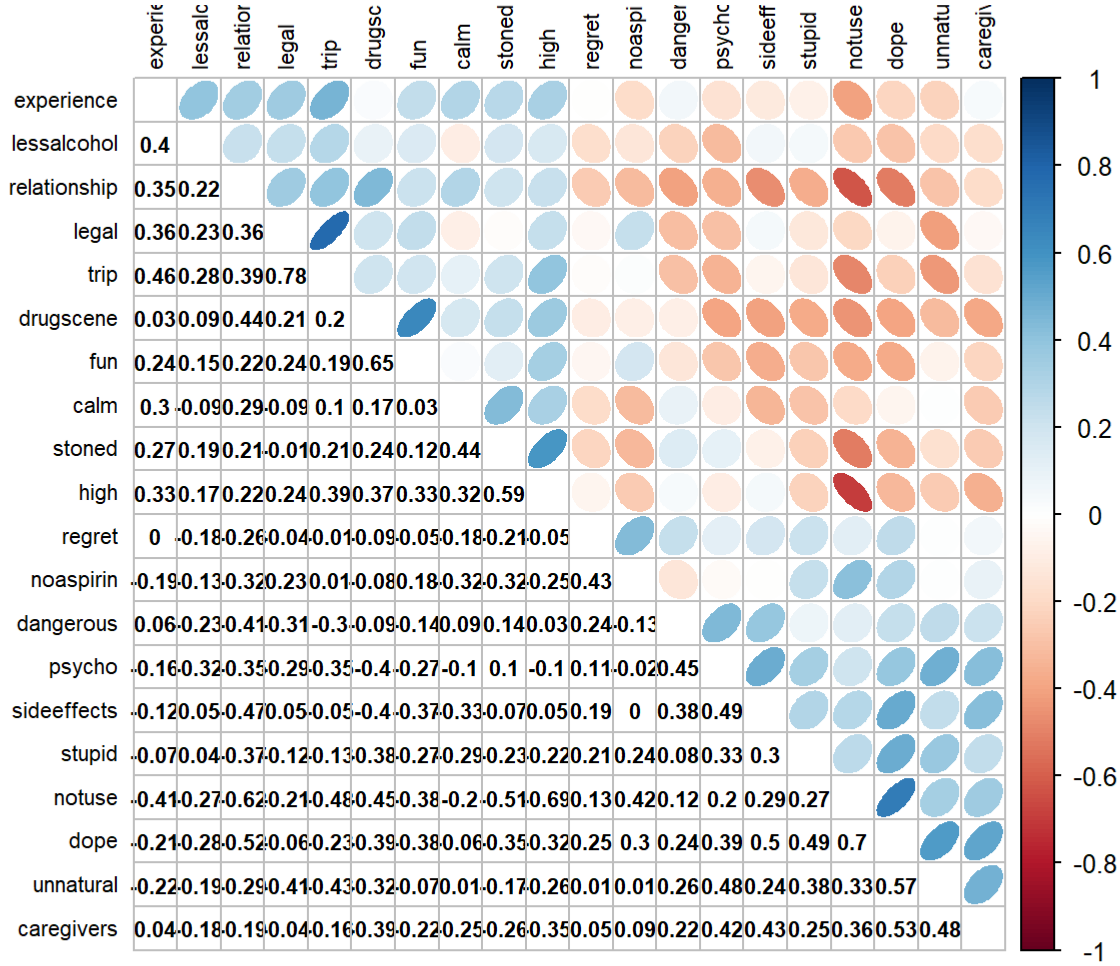
```
round(cor(data_transformed), 2)
```

```
##              legal dangerous regret unnatural notuse psycho  trip stoned
## legal         1.00     -0.31  -0.04     -0.41  -0.21  -0.29  0.78  -0.01
## dangerous    -0.31      1.00   0.24      0.26   0.12   0.45 -0.30   0.14
## regret       -0.04      0.24   1.00      0.01   0.13   0.11 -0.01  -0.21
## unnatural    -0.41      0.26   0.01      1.00   0.33   0.48 -0.43  -0.17
## notuse       -0.21      0.12   0.13      0.33   1.00   0.20 -0.48  -0.51
## psycho       -0.29      0.45   0.11      0.48   0.20   1.00 -0.35   0.10
## trip          0.78     -0.30  -0.01     -0.43  -0.48  -0.35  1.00   0.21
## stoned       -0.01      0.14  -0.21     -0.17  -0.51   0.10  0.21   1.00
## calm         -0.09      0.09  -0.18      0.01  -0.20  -0.10  0.10   0.44
## high          0.24      0.03  -0.05     -0.26  -0.69  -0.10  0.39   0.59
## noaspirin     0.23     -0.13   0.43      0.01   0.42  -0.02  0.01  -0.32
## relationship  0.36     -0.41  -0.26     -0.29  -0.62  -0.35  0.39   0.21
## drugscene     0.21     -0.09  -0.09     -0.32  -0.45  -0.40  0.20   0.24
## caregivers   -0.04      0.22   0.05      0.48   0.36   0.42 -0.16  -0.26
## experience    0.36      0.06   0.00     -0.22  -0.41  -0.16  0.46   0.27
## fun           0.24     -0.14  -0.05     -0.07  -0.38  -0.27  0.19   0.12
## stupid       -0.12      0.08   0.21      0.38   0.27   0.33 -0.13  -0.23
## lessalcohol   0.23     -0.23  -0.18     -0.19  -0.27  -0.32  0.28   0.19
## sideeffects   0.05      0.38   0.19      0.24   0.29   0.49 -0.05  -0.07
## dope         -0.06      0.24   0.25      0.57   0.70   0.39 -0.23  -0.35
##               calm  high noaspirin relationship drugscene caregivers
## legal        -0.09  0.24      0.23         0.36      0.21      -0.04
## dangerous     0.09  0.03     -0.13        -0.41     -0.09       0.22
## regret       -0.18 -0.05      0.43        -0.26     -0.09       0.05
## unnatural     0.01 -0.26      0.01        -0.29     -0.32       0.48
## notuse       -0.20 -0.69      0.42        -0.62     -0.45       0.36
## psycho       -0.10 -0.10     -0.02        -0.35     -0.40       0.42
## trip          0.10  0.39      0.01         0.39      0.20      -0.16
## stoned        0.44  0.59     -0.32         0.21      0.24      -0.26
## calm          1.00  0.32     -0.32         0.29      0.17      -0.25
## high          0.32  1.00     -0.25         0.22      0.37      -0.35
## noaspirin    -0.32 -0.25      1.00        -0.32     -0.08       0.09
## relationship  0.29  0.22     -0.32         1.00      0.44      -0.19
## drugscene     0.17  0.37     -0.08         0.44      1.00      -0.39
## caregivers   -0.25 -0.35      0.09        -0.19     -0.39       1.00
## experience    0.30  0.33     -0.19         0.35      0.03       0.04
## fun           0.03  0.33      0.18         0.22      0.65      -0.22
## stupid       -0.29 -0.22      0.24        -0.37     -0.38       0.25
## lessalcohol  -0.09  0.17     -0.13         0.22      0.09      -0.18
## sideeffects  -0.33  0.05      0.00        -0.47     -0.40       0.43
## dope         -0.06 -0.32      0.30        -0.52     -0.39       0.53
##              experience   fun stupid lessalcohol sideeffects  dope
## legal              0.36  0.24  -0.12        0.23        0.05 -0.06
## dangerous          0.06 -0.14   0.08       -0.23        0.38  0.24
## regret             0.00 -0.05   0.21       -0.18        0.19  0.25
## unnatural         -0.22 -0.07   0.38       -0.19        0.24  0.57
## notuse            -0.41 -0.38   0.27       -0.27        0.29  0.70
## psycho            -0.16 -0.27   0.33       -0.32        0.49  0.39
## trip               0.46  0.19  -0.13        0.28       -0.05 -0.23
## stoned             0.27  0.12  -0.23        0.19       -0.07 -0.35
## calm               0.30  0.03  -0.29       -0.09       -0.33 -0.06
## high               0.33  0.33  -0.22        0.17        0.05 -0.32
## noaspirin        -0.19  0.18   0.24        -0.13         0.00  0.30
## relationship      0.35  0.22  -0.37         0.22        -0.47 -0.52
## drugscene         0.03  0.65  -0.38         0.09        -0.40 -0.39
## caregivers        0.04 -0.22   0.25        -0.18         0.43  0.53
## experience        1.00  0.24  -0.07         0.40        -0.12 -0.21
## fun               0.24  1.00  -0.27         0.15        -0.37 -0.38
## stupid           -0.07 -0.27   1.00         0.04         0.30  0.49
## lessalcohol       0.40  0.15   0.04         1.00         0.05 -0.28
## sideeffects      -0.12 -0.37   0.30         0.05         1.00  0.50
## dope             -0.21 -0.38   0.49        -0.28         0.50  1.00
```

```
#version 1
corrplot(cor(data_transformed), method = "color")
```



```
#version 2
corrplot.mixed(cor(data_transformed), lower.col="black", upper = "ellipse", tl.col = "black", nu
mber.cex=.7, order = "hclust",tl.pos = "lt", tl.cex=.7)
```



The correlation plot shows that while there is low correlation between many of the variables, there are a few spots where variables are moderately or highly correlated. This includes, for example, "notuse" and "relationship"; "trip" and "legal"; "dope" and "notuse," as well as "fun" and "drugscene." Given the noted highly correlated variables, PCA will be helpful in reducing dimensions. It won't necessarily give us two or three highly explanatory variables that sum up the entire dataset, but it will reduce the repetitiveness of current variables and allow us to group them somewhat intelligently.

PROBLEM 3

3). Perform Principle components analysis using the Correlation matrix (standardized variables). Think about how many principle components to retain. To make this decision look at Total variance explained by a given number of principle components The 'eigenvalue > 1' criteria The 'scree plot elbow' method (turn in the scree plot) Parallel Analysis : think about whether this is appropriate based on what you discover in question 1.

```
source("http://www.reuningscherer.net/STAT660/R/parallel.r.txt")
source("http://reuningscherer.net/stat660/r/ciscoreplot.R.txt")
data2 <- data_transformed[,c(names(data_transformed))]
pc1 <- princomp(data2, cor=TRUE)

#To see the cumulative variability accounted for by our components
print(summary(pc1),digits=2,loadings=pc1$loadings,cutoff=0)
```

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
## Standard deviation     2.4218087 1.5557992 1.4517678 1.24438766 1.1081651
## Proportion of Variance 0.2932579 0.1210256 0.1053815 0.07742503 0.0614015
## Cumulative Proportion  0.2932579 0.4142834 0.5196649 0.59708993 0.6584914
##                             Comp.6    Comp.7     Comp.8     Comp.9
## Standard deviation     1.06583377 0.9944123 0.93980558 0.90798998
## Proportion of Variance 0.05680008 0.0494428 0.04416173 0.04122229
## Cumulative Proportion  0.71529151 0.7647343 0.80889603 0.85011832
##                            Comp.10    Comp.11    Comp.12   Comp.13
## Standard deviation     0.79816604 0.69775175 0.68747835 0.5901898
## Proportion of Variance 0.03185345 0.02434287 0.02363132 0.0174162
## Cumulative Proportion  0.88197177 0.90631465 0.92994597 0.9473622
##                            Comp.14    Comp.15     Comp.16     Comp.17
## Standard deviation     0.55886825 0.48973629 0.416658649 0.359795457
## Proportion of Variance 0.01561669 0.01199208 0.008680222 0.006472639
## Cumulative Proportion  0.96297886 0.97497094 0.983651164 0.990123803
##                             Comp.18     Comp.19     Comp.20
## Standard deviation     0.313723489 0.231915301 0.212877457
## Proportion of Variance 0.004921121 0.002689235 0.002265841
## Cumulative Proportion  0.995044924 0.997734159 1.000000000
```

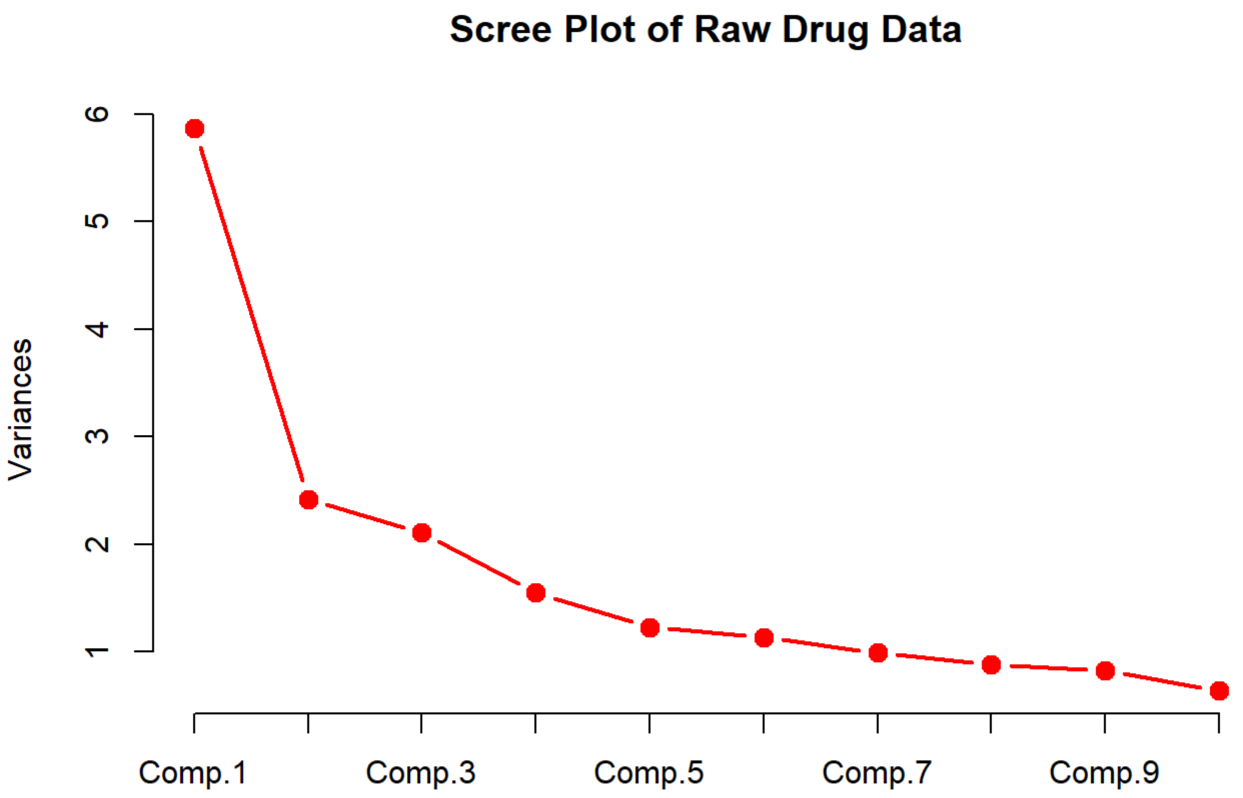
```
## Warning in if (loadings) {: the condition has length > 1 and only the first
## element will be used
```

```
##
## Loadings:
##              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## legal         0.18   0.34   0.38   0.02   0.15   0.15   0.27   0.13
## dangerous    -0.15  -0.36   0.13  -0.32  -0.03   0.05   0.02  -0.38
## regret       -0.12   0.14   0.14  -0.48  -0.11   0.20  -0.26  -0.36
## unnatural    -0.25  -0.19  -0.01   0.02   0.42  -0.32  -0.11   0.18
## notuse       -0.33   0.18  -0.13   0.05   0.06   0.18  -0.06   0.10
## psycho       -0.24  -0.31   0.17  -0.08   0.04  -0.11   0.20   0.03
## trip          0.24   0.20   0.40   0.04   0.07   0.21   0.09   0.12
## stoned        0.20  -0.39   0.14  -0.10  -0.10   0.02  -0.05   0.26
## calm          0.14  -0.33  -0.07   0.01   0.35   0.46  -0.31   0.21
## high          0.24  -0.24   0.25  -0.27  -0.11  -0.01   0.06   0.30
## noaspirin    -0.12   0.42   0.03  -0.37   0.09   0.04  -0.14   0.15
## relationship  0.30   0.01  -0.03   0.25   0.28   0.00   0.11  -0.16
## drugscene     0.26   0.02  -0.17  -0.34   0.16  -0.20   0.18   0.07
## caregivers   -0.24  -0.01   0.23   0.16   0.40  -0.14   0.27  -0.28
## experience    0.19  -0.07   0.38   0.09   0.24   0.04  -0.38  -0.40
## fun           0.21   0.10  -0.06  -0.40   0.32  -0.45  -0.01   0.01
## stupid       -0.22   0.07   0.21   0.04  -0.05  -0.24  -0.42   0.26
## lessalcohol   0.16   0.08   0.23   0.26  -0.23  -0.42  -0.38   0.00
## sideeffects  -0.22  -0.09   0.41  -0.01  -0.25  -0.08   0.30   0.07
## dope         -0.32   0.04   0.18  -0.03   0.27   0.16  -0.05   0.29
##              Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
## legal         0.08   0.02    0.16    0.11    0.26    0.02    0.12
## dangerous     0.29  -0.11    0.39    0.19    0.12   -0.31    0.39
## regret       -0.31  -0.26   -0.42    0.06    0.02   -0.04   -0.24
## unnatural    -0.08  -0.14   -0.23   -0.19    0.25   -0.43    0.10
## notuse        0.41   0.08   -0.01    0.15    0.05    0.16   -0.16
## psycho       -0.31   0.34    0.04    0.22    0.46    0.29   -0.25
## trip         -0.09  -0.04    0.13   -0.07    0.04   -0.53   -0.30
## stoned        0.01   0.39   -0.18    0.40   -0.32   -0.23   -0.16
## calm          0.14  -0.07   -0.09    0.03    0.13    0.10    0.04
## high         -0.05  -0.11   -0.13   -0.40   -0.19    0.28    0.33
## noaspirin    -0.02   0.43   -0.15    0.17   -0.02   -0.01    0.48
## relationship -0.34  -0.19   -0.19    0.35    0.17    0.15    0.33
## drugscene     0.14  -0.40    0.13    0.43   -0.16    0.13   -0.18
## caregivers    0.00   0.11   -0.16    0.04   -0.58   -0.01    0.04
## experience    0.13   0.18    0.15   -0.16   -0.02    0.30   -0.11
## fun           0.18   0.16    0.06   -0.23    0.13    0.07   -0.23
## stupid       -0.38  -0.17    0.50    0.15   -0.19    0.10    0.04
## lessalcohol   0.35  -0.08   -0.31    0.26    0.16   -0.05    0.05
## sideeffects   0.20  -0.21   -0.21   -0.03    0.11    0.16    0.02
## dope          0.15  -0.29   -0.06    0.07   -0.11    0.13   -0.11
##              Comp.16 Comp.17 Comp.18 Comp.19 Comp.20
## legal         0.26    0.02    0.53    0.08    0.31
## dangerous     0.01   -0.05    0.05   -0.09   -0.18
## regret        0.05    0.04    0.24    0.06   -0.01
## unnatural     0.24    0.05   -0.06    0.32    0.16
## notuse        0.12    0.16    0.10    0.48   -0.52
## psycho       -0.16   -0.33   -0.02    0.07   -0.06
## trip         -0.27   -0.10   -0.29    0.08   -0.32
## stoned        0.33    0.24    0.06   -0.05    0.00
## calm         -0.51  0.13  0.17 -0.03  0.16
## high          0.06 -0.25  0.10  0.25 -0.31
## noaspirin    -0.16 -0.03 -0.33  0.01  0.10
## relationship  0.12  0.23 -0.12 -0.10 -0.41
## drugscene    -0.07 -0.14 -0.23  0.32  0.23
## caregivers   -0.30 -0.09  0.23  0.06 -0.01
## experience    0.27  0.04 -0.35  0.17  0.16
## fun          -0.09  0.29  0.17 -0.34 -0.25
## stupid       -0.11  0.24  0.14  0.05 -0.06
## lessalcohol  -0.19 -0.31  0.14 -0.05 -0.06
## sideeffects  -0.19  0.55 -0.27 -0.04  0.15
## dope          0.30 -0.31 -0.16 -0.55 -0.08
```

```
round(pc1$sdev^2,2)
```

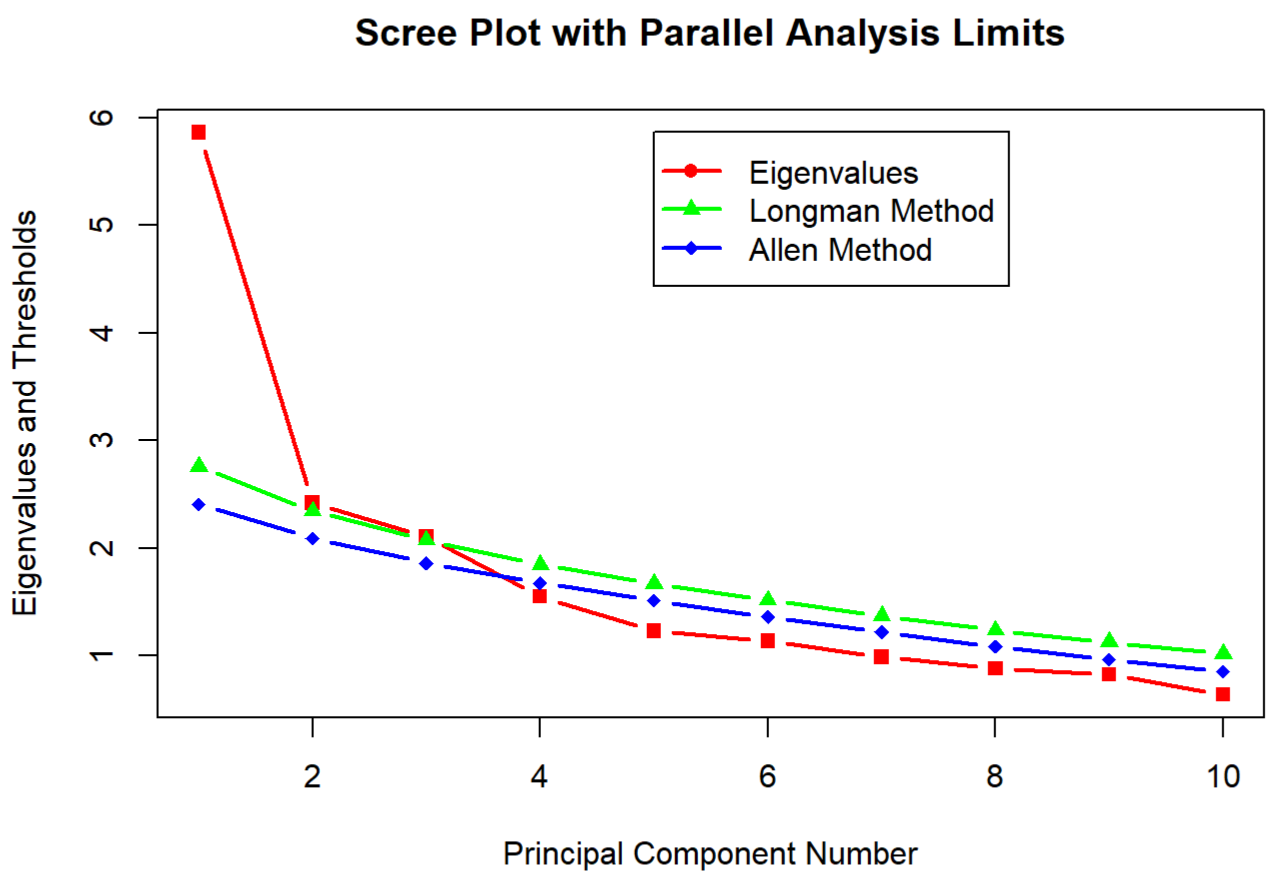
```
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9 
##    5.87    2.42    2.11    1.55    1.23    1.14    0.99    0.88    0.82 
## Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18 
##    0.64    0.49    0.47    0.35    0.31    0.24    0.17    0.13    0.10 
## Comp.19 Comp.20 
##    0.05    0.05
```

```
#We also want to check a screeplot in order to identify any "elbows":
screeplot(pc1,type="lines",col="red",lwd=2,pch=19,cex=1.2,main="Scree Plot of Raw Drug Data")
```



Recall that Parallel Analysis is appropriate if the data has a multivariate normal distribution. Based on our previous results from (1), in particular the Chi-Squared Quantile plot, we note that our data appears to have a multivariate normal distribution — all data points (except for 2) are between the 95% confidence interval bounds with respect to the 45-degree line. Consequently, we would say that our data is appropriate for Parallel Analysis, given their multivariate normality.

```
source("http://www.reuningscherer.net/STAT660/R/parallel.r.txt")
parallelplot(pc1)
```



Following our analysis, we decided to retain 3 principal components. First, when we look at the summary of our PCA, we see that by including components 1,2 and 3, we are able to explain 50% of the variability, which is strong given that we are only including three components. Furthermore, following the eigenvalue rule of thumb, we see that the first three components have eigenvalues that are greater than 1 (in fact, much above 2), and the eigenvalues components beyond 3rd rapidly fall near and below 1. The scree plot shows potential “elbows” at the 2nd and 5th component; while the 3rd component falls beyond the elbow at the 2nd component, we believe the third component’s high eigenvalues and ability to explain more variability merits inclusion (as well as the makeup of its set of contributors, discussed below in problem 4.) As such, including the 3rd component in our judgment seems like the best way of maximizing explanatory power while minimizing the number of components to include.

PROBLEM 4

4). For principle components you decide to retain, examine the loadings (principle components) and think about an interpretation for each retained component if possible.

```
unclass(pc1$loadings)[,1:3]
```

```
##                  Comp.1      Comp.2      Comp.3
## legal         0.1754340  0.33902322  0.37946769
## dangerous    -0.1485403 -0.35885771  0.12764559
## regret       -0.1169510  0.14397665  0.14462069
## unnatural    -0.2505420 -0.19302771 -0.01328092
## notuse       -0.3276529  0.17877540 -0.12829032
## psycho       -0.2400562 -0.30592870  0.16920396
## trip          0.2409972  0.19593640  0.39751253
## stoned        0.1955342 -0.39164280  0.13655290
## calm          0.1440325 -0.32709580 -0.07396334
## high          0.2443258 -0.23715537  0.25034943
## noaspirin    -0.1214688  0.42358343  0.02967649
## relationship  0.2957727  0.01360216 -0.02873280
## drugscene     0.2624187  0.02495031 -0.16658211
## caregivers   -0.2350445 -0.01078438  0.22786639
## experience    0.1881141 -0.06571882  0.37632027
## fun           0.2104343  0.10413867 -0.05566311
## stupid       -0.2206703  0.07199000  0.21382669
## lessalcohol   0.1568316  0.08353340  0.23172442
## sideeffects  -0.2199699 -0.09126361  0.41498822
## dope         -0.3182440  0.04162814  0.17605420
```

Note that following our interpretation of the scree plot with parallel analysis thresholds, we decided to retain the first 3 principal components.

In the first principal component, we see that ‘notuse’ is the primary contributor, with ‘relationship,’ and ‘drugscene’. The questions which produced these variables are here: ‘notuse’: Even if my best friend gave me some hash, I probably wouldn’t use it. ‘relationship’: If people use drugs together, their relationships will be improved. ‘drugscene’: In spite of what the establishment says, the drug scene is really “where it’s at”. *It seems that all of these variables speak to the social aspect of drug usage and people’s attitudes around it as a social activity. People who like the drug scene probably would likely use drugs provided by their best friend would probably also respond that it improves relationships.*

The second principal component’s (Comp.2) primary contributors include ‘noaspirin,’ ‘legal,’ ‘dangerous,’ and ‘calm.’ ‘noaspirin’: I’d have to be pretty sick before I’d take any drug including an aspirin. ‘legal’: All drugs should be made legal and freely available. ‘dangerous’: As a general rule of thumb, most drugs are dangerous and should be used only with medical authorization. ‘calm’: I wish I could get hold of some pills to calm me down whenever I get “up tight”. *The common thread between these variables seems to be that the questions have a semi-medical, individualized nature (with the exception of ‘legal’) to their inquiry that would lead to high correlations between answers (even if that correlation is highly negative). For example, someone who is very suspicious of taking aspirin would probably highly agree that most drugs are dangerous and would highly disagree that taking pills is a legitimate way to calm down. In that case, they might also extrapolate their personal experience with drug usage to others (reflected in ‘legal’) in the belief that they are not medically helpful. Those who who think that drugs are key to feeling better and have individual experience using them in a semi-medical way would likely also want them legal and freely available as a medical recourse for others.*

The third principal component’s (Comp.3) primary contributors are ‘side effects’ and ‘experience.’ ‘sideeffects’: Students should be told about the harmful side effects of certain drugs. ‘experience’: People who make drug legislation should really have personal experience with drugs. *It makes sense that these responses would be* *correlated and form a component since they all reflect something about the respondent's attitudes towards regulations and other people's drug usage–especially young people. Those who think students should know about hamful side effects might want legislators who are drug-free (if they see it as a moral issue), or might really want legislators with personal experience, especially if that experience is negative. Either way, it reflects a concern about drug use in society at-large.*

PROBLEM 5

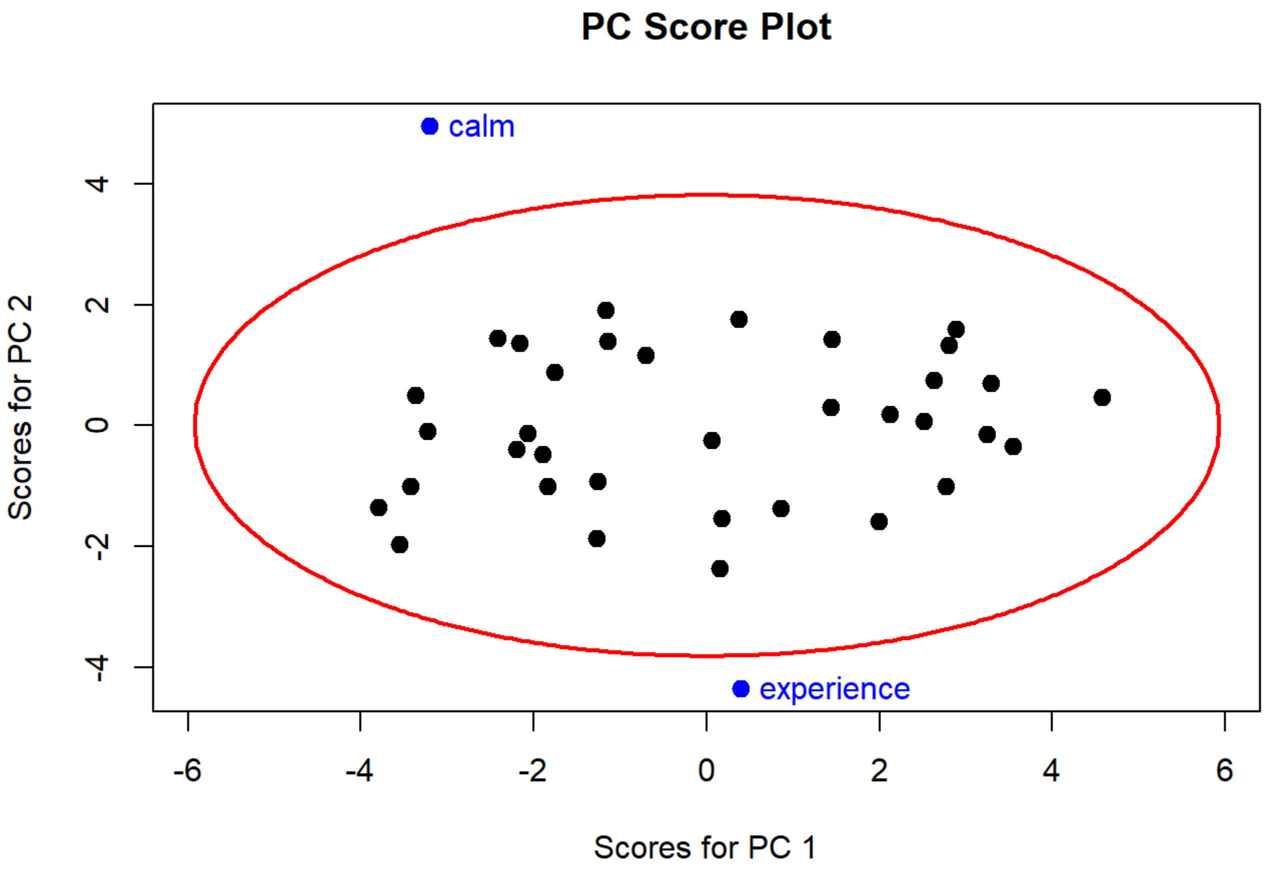
5. Make a score plot of the scores for at least one pair of component scores (one and two, one and three, two and three, etc). Discuss any trends/groupings you observe (probably, this will be 'none'). As a bonus, try to make a 95% Confidence Ellipse for two of your components. You might want to also try making a bi-plot if you're using R.

Define the Score Plot Function

```
ciscoreplot<-function(x,comps,namevec){
  y1<-sqrt(5.99*(x$sdev[comps[1]]^2))
  ymod<-y1-y1%%.05
  y1vec<-c(-y1,seq(-ymod,ymod,by=0.05),y1)
  y2vecpos<-sqrt((5.99-(y1vec^2)/x$sdev[comps[1]]^2)*x$sdev[comps[2]]^2)
  y2vecneg<--sqrt((5.99-(y1vec^2)/x$sdev[comps[1]]^2)*x$sdev[comps[2]]^2)
  y2vecpos[1]<-0
  y2vecneg[1]<-0
  y2vecpos[length(y2vecpos)]<-0
  y2vecneg[length(y2vecneg)]<-0
  plot(x$scores[,comps[1]],x$scores[,comps[2]],pch=19,cex=1.2,ylim=c(min(y2vecneg,x$scores[,comp
s[2]]),max(y2vecpos,x$scores[,comps[2]])),
    main="PC Score Plot", xlab=paste("Scores for PC",comps[1],sep=" "), ylab=paste("Scores for P
C",comps[2],sep=" "),
    xlim=c(min(y1vec,x$scores[,comps[1]]),max(y1vec,x$scores[,comps[1]])))
    lines(y1vec,y2vecpos,col="Red",lwd=2)
    lines(y1vec,y2vecneg,col="Red",lwd=2)
  outliers<-((x$scores[,comps[1]]^2)/(x$sdev[comps[1]]^2)+(x$scores[,comps[2]]^2)/(x$sdev[comps[
2]]^2))>5.99
  points(x$scores[outliers,comps[1]],x$scores[outliers,comps[2]],pch=19,cex=1.2,col="Blue")
  text(x$scores[outliers,comps[1]],x$scores[outliers,comps[2]],col="Blue",lab=namevec[outliers],
 pos=4)
}
```

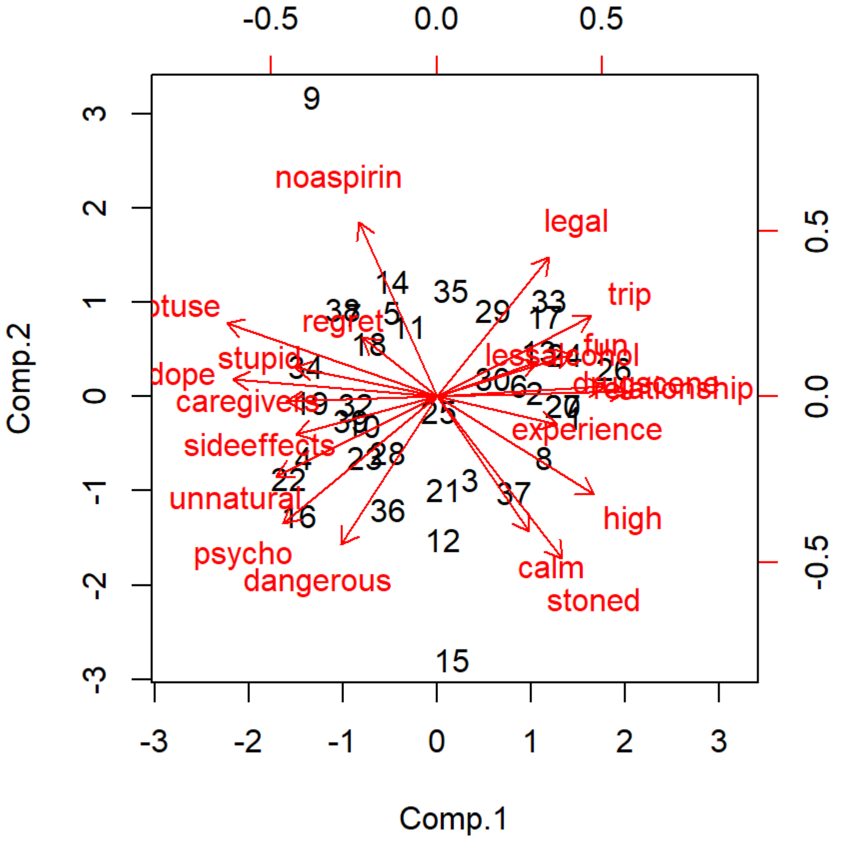
Let's make the scoreplot for the first two components, including a 95% confidence ellipse in the process:

```
ciscoreplot(pc1,c(1,2), names(data))
```



We see a potential grouping: some on the right and some on the left. But the grouping does not look too strong. We also observe 2 outliers: the variables calm and experience.

```
biplot(pc1,choices=c(1,2),pc.biplot=T)
```



PROBLEM 6

6). Write a paragraph summarizing your findings, and your opinions about the effectiveness of using principle components on this data. Include evidence based on scatterplots of linearity in higher dimensional space, note any multivariate outliers in your score plot, comment on sample size relative to number of variables, etc.

From the biplot and scoreplot above, we see that there are not any super apparent groupings of variables that would lead to highly explanatory principal components: in the scoreplot, the dots are fairly scattered, while in the biplot, the vectors point in all directions. This makes sense given our earlier commentary after seeing the correlation plot, where only a few pairs (potentially triplets) of variables seemed to be highly correlated and many others had low to moderate correlation.

Part of this may be due to our sample size - note that we only have 38 (relatively few) observations for 20 (relatively many) variables, which makes it very difficult to even see the distribution of responses per variable, let alone determine very rigorous multivariate relationships. While we have outliers with 'calm' and 'experience' according to the scoreplot alone, it may be due simply to the lack of many responses.

Although we did select three principal components above which reflected the social, medical/individual needs, and society/regulatory-oriented attitudes toward drug usage, and they did explain approximately half the variability of the data, PCA might not be the most effective way to make sense of this data especially since each component included multiple contributors with approximately the same weight of contribution. Perhaps with more observations to the data set, or a more normally distributed set of responses, we could derive more strongly explanatory principal components from the dataset. Given the current state of the data, however, PCA was helpful in reducing the dimensions of the dataset and grouping some variables into these three categories of social, medical/individual *and society/regulatory, which made people’s responses to drug attitudes slightly easier to understand and provided us some insight into the factors that might affect someone’s drug attitudes or different categories in which a respondent might think about drug usage which we did not have before.*