Principal Components Analysis on Drug Attitudes

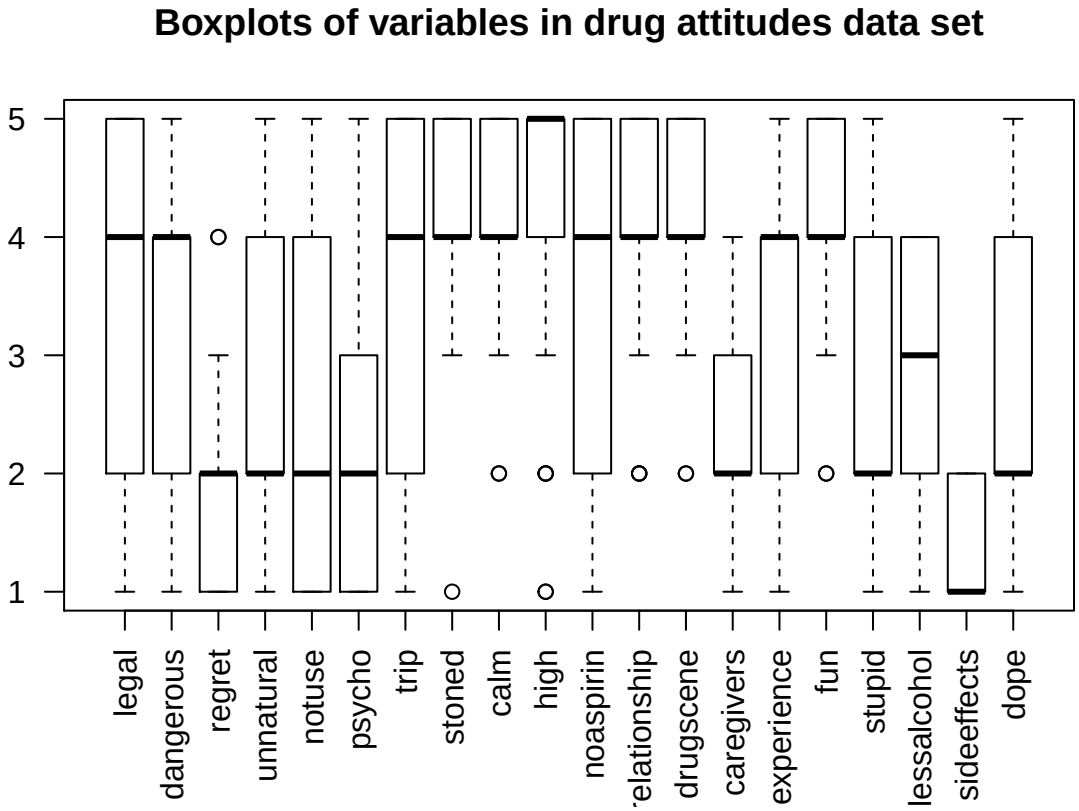
In this project, I apply PCA techniques on a dataset on drug attitudes in the United States to ascertain factors that account for most of the correlation across the data.

```
drug <- read.csv("Data/Drugattitudes.csv")
drug <- drug[complete.cases(drug), ]
```

Multivariate Normal Distribution Although multivariate normality is not necessary to perform PCA, it can affect the interpretation of results, because the relative importance of skewed variables' components can be exaggerated or understated. In addition, it is a required assumption for parallel analysis.

A group of variables that are multivariate normally distributed consist of variables that are themselves normally distributed. Hence, we generate a boxplot of the individual variables, to ensure that their quartiles are consistent with that of a normal distribution, and that there is minimal skewdness.

```
boxplot(drug, las = 2, main = "Boxplots of variables in drug attitudes data set")
```



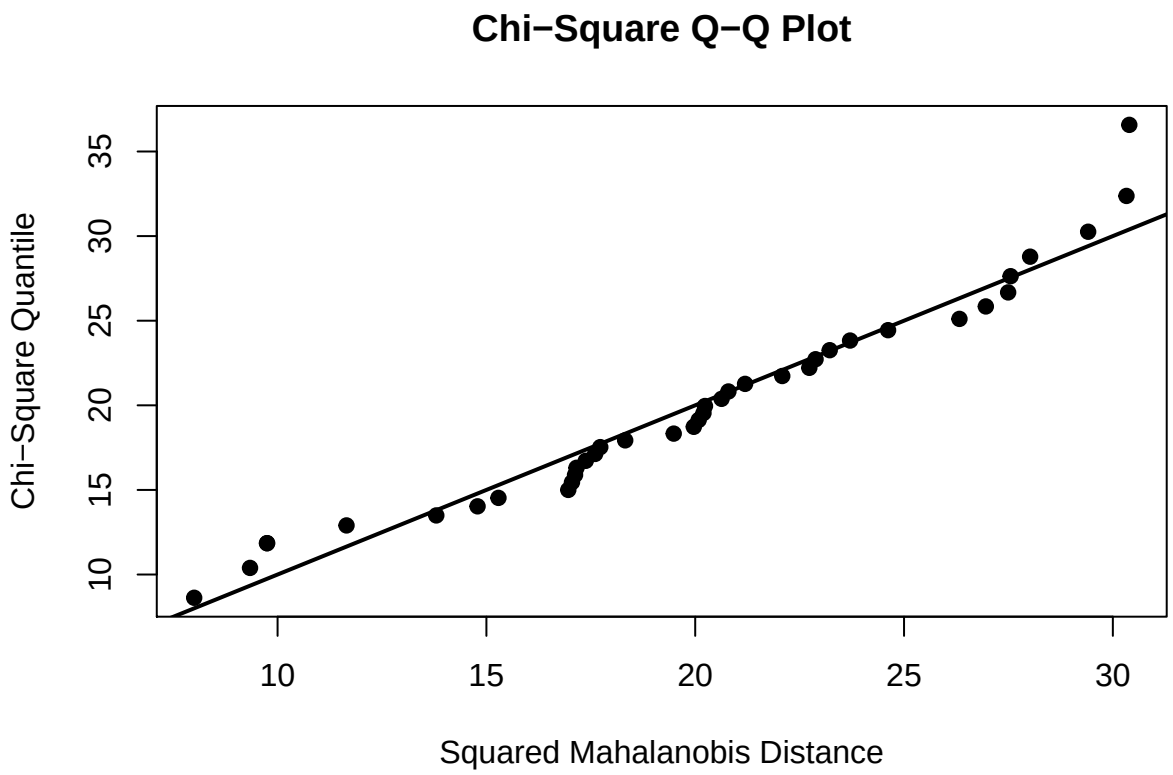
We can see from the boxplot that it is difficult to assume multivariate normality, since there are some variables that are unlikely to have come from a normal distribution. However, it could be argued that this is to be expected, as there are only 5 possible values for each variable. The fact that these variables are discrete over a small range makes it difficult to interpret the boxplots.

Since it is difficult to deduce multivariate normality visually from the above graph, let's apply Mardia's multivariate normality test from the MVN package, and also generate a chi-squared quantile-quantile plot.

```
library(MVN)
```

```
## sROC 0.1-2 loaded
```

```
mardiaTest(drug, qqplot = TRUE)
```




```
##    Mardia's Multivariate Normality Test
## ---------------------------------------
##    data : drug
##
##    g1p            : 253.0384
##    chi.skew       : 1602.576
##    p.value.skew   : 0.1303793
##
##    g2p            : 434.6748
##    z.kurtosis     : -0.5532972
##    p.value.kurt   : 0.5800599
##
##    chi.small.skew : 1741.856
##    p.value.small  : 0.0002314853
##
##    Result         : Data are multivariate normal.
## ---------------------------------------
```

We see that the above R output suggests that our data is in fact consistent with an assumption of multivariate normality. Therefore, we continue to use the variables as they are when performing PCA, but also take some care in interpreting the loadings and when performing parallel analysis, due to the discrete nature of our variables.

Correlation Matrix

```
drug_cor <- round(cor(drug),digits=2)
drug_cor
```

```
##              legal dangerous regret unnatural notuse psycho  trip stoned
```

```
## legal         1.00     -0.31  -0.02     -0.41  -0.23  -0.27  0.78   0.03
## dangerous    -0.31      1.00   0.19      0.26   0.10   0.42 -0.30   0.04
## regret       -0.02      0.19   1.00     -0.04   0.07   0.04  0.01  -0.18
## unnatural    -0.41      0.26  -0.04      1.00   0.34   0.49 -0.43  -0.17
## notuse       -0.23      0.10   0.07      0.34   1.00   0.11 -0.49  -0.60
## psycho       -0.27      0.42   0.04      0.49   0.11   1.00 -0.33   0.05
## trip          0.78     -0.30   0.01     -0.43  -0.49  -0.33  1.00   0.27
## stoned        0.03      0.04  -0.18     -0.17  -0.60   0.05  0.27   1.00
## calm         -0.09      0.09  -0.15      0.01  -0.14  -0.07  0.10   0.46
## high          0.24      0.03  -0.03     -0.26  -0.69  -0.07  0.39   0.60
## noaspirin     0.25     -0.13   0.37      0.03   0.47  -0.04  0.01  -0.39
## relationship  0.36     -0.40  -0.20     -0.30  -0.61  -0.31  0.40   0.39
## drugscene     0.21     -0.09  -0.07     -0.32  -0.37  -0.39  0.20   0.33
## caregivers   -0.04      0.22  -0.01      0.48   0.29   0.42 -0.16  -0.21
## experience    0.36      0.06   0.01     -0.22  -0.44  -0.14  0.46   0.30
## fun           0.24     -0.14  -0.08     -0.07  -0.31  -0.27  0.19   0.20
## stupid       -0.12      0.08   0.20      0.38   0.24   0.33 -0.13  -0.22
## lessalcohol   0.29     -0.25  -0.09     -0.22  -0.30  -0.36  0.31   0.25
## sideeffects   0.05      0.38   0.16      0.24   0.20   0.47 -0.05  -0.17
## dope         -0.06      0.24   0.18      0.57   0.69   0.37 -0.23  -0.41
##                calm  high noaspirin relationship drugscene caregivers
## legal         -0.09  0.24      0.25         0.36      0.21      -0.04
## dangerous      0.09  0.03     -0.13        -0.40     -0.09       0.22
## regret        -0.15 -0.03      0.37        -0.20     -0.07      -0.01
## unnatural      0.01 -0.26      0.03        -0.30     -0.32       0.48
## notuse        -0.14 -0.69      0.47        -0.61     -0.37       0.29
## psycho        -0.07 -0.07     -0.04        -0.31     -0.39       0.42
## trip           0.10  0.39      0.01         0.40      0.20      -0.16
## stoned         0.46  0.60     -0.39         0.39      0.33      -0.21
## calm           1.00  0.32     -0.31         0.29      0.17      -0.25
## high           0.32  1.00     -0.27         0.24      0.37      -0.35
## noaspirin     -0.31 -0.27      1.00        -0.32     -0.10       0.09
## relationship   0.29  0.24     -0.32         1.00      0.46      -0.20
## drugscene      0.17  0.37     -0.10         0.46      1.00      -0.39
## caregivers    -0.25 -0.35      0.09        -0.20     -0.39       1.00
## experience     0.30  0.33     -0.20         0.34      0.03       0.04
## fun            0.03  0.33      0.17         0.24      0.65      -0.22
## stupid        -0.29 -0.22      0.23        -0.37     -0.38       0.25
## lessalcohol   -0.08  0.21     -0.13         0.23      0.10      -0.20
## sideeffects   -0.33  0.05     -0.01        -0.48     -0.40       0.43
## dope          -0.06 -0.32      0.32        -0.54     -0.39       0.53
##               experience   fun stupid lessalcohol sideeffects  dope
## legal               0.36  0.24  -0.12        0.29        0.05 -0.06
## dangerous           0.06 -0.14   0.08       -0.25        0.38  0.24
## regret              0.01 -0.08   0.20       -0.09        0.16  0.18
## unnatural          -0.22 -0.07   0.38       -0.22        0.24  0.57
## notuse             -0.44 -0.31   0.24       -0.30        0.20  0.69
## psycho             -0.14 -0.27   0.33       -0.36        0.47  0.37
## trip                0.46  0.19  -0.13        0.31       -0.05 -0.23
## stoned              0.30  0.20  -0.22        0.25       -0.17 -0.41
## calm                0.30  0.03  -0.29       -0.08       -0.33 -0.06
## high                0.33  0.33  -0.22        0.21        0.05 -0.32
## noaspirin          -0.20  0.17   0.23       -0.13       -0.01  0.32
## relationship        0.34  0.24  -0.37        0.23       -0.48 -0.54
```

```
## drugscene          0.03  0.65  -0.38        0.10       -0.40 -0.39
## caregivers         0.04 -0.22   0.25       -0.20        0.43  0.53
## experience         1.00  0.24  -0.07        0.42       -0.12 -0.21
## fun                0.24  1.00  -0.27        0.14       -0.37 -0.38
## stupid            -0.07 -0.27   1.00        0.04        0.30  0.49
## lessalcohol        0.42  0.14   0.04        1.00        0.00 -0.27
## sideeffects       -0.12 -0.37   0.30        0.00        1.00  0.50
## dope              -0.21 -0.38   0.49       -0.27        0.50  1.00
```

It seems that our data is a good candidate for PCA, as there are a number of variables that are highly correlated with each other. This means that we can reduce the dimensionality of the data set without loosing too much of its information. For example, the pairs of variables high and notuse, dope and notuse, and legal and trip all have correlation coefficients that are greater in magnitude than 0.65. Furthermore, the correlation matrix only tells us about the pairwise correlations - it might be the case that one of the variables can be written as a linear combination of two or more other variables.

On the other hand, there are variables that exhibit almost no correlation with each other. For example, less alcohol and side effects have a corrrelation coefficient of < 0.01. Thus, this suggests that the data set actually contains useful information, and cannot be represented faithfully via only one or two principal components.

Principal Component Analysis

The following R code will carry out principal component analysis on the correlation matrix:

```
pc1 <- princomp(drug, cor=TRUE)

#print(summary(pc1),digits=2,loadings=pc1$loadings,cutoff=0)
```

To decide how many principal components to retain, we first observe the proportion of variance explained:

```
summary(pc1)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.4285513 1.5552369 1.4680532 1.20134817 1.12458209
## Proportion of Variance 0.2948931 0.1209381 0.1077590 0.07216187 0.06323424
## Cumulative Proportion  0.2948931 0.4158312 0.5235902 0.59575206 0.65898630
##                           Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     1.05508151 1.00982885 0.93513772 0.92630490
## Proportion of Variance 0.05565985 0.05098772 0.04372413 0.04290204
## Cumulative Proportion  0.71464615 0.76563387 0.80935800 0.85226004
##                           Comp.10    Comp.11    Comp.12    Comp.13
## Standard deviation     0.77831705 0.70753203 0.68428902 0.60099154
## Proportion of Variance 0.03028887 0.02503008 0.02341257 0.01805954
## Cumulative Proportion  0.88254891 0.90757899 0.93099156 0.94905110
##                           Comp.14    Comp.15     Comp.16     Comp.17
## Standard deviation     0.55807574 0.47608893 0.393979382 0.352546308
## Proportion of Variance 0.01557243 0.01133303 0.007760988 0.006214445
## Cumulative Proportion  0.96462353 0.97595656 0.983717548 0.989931993
##                           Comp.18     Comp.19     Comp.20
## Standard deviation     0.324592996 0.231671848 0.205736912
## Proportion of Variance 0.005268031 0.002683592 0.002116384
## Cumulative Proportion  0.995200024 0.997883616 1.000000000
```

We observe that cumulatively, the first 8 principal components account for 80% of the variance.

Next we can view the eigenvalues of the principal components. Note that in R the value given is the standard deviation rather than the variance, so we square the values as seen below:

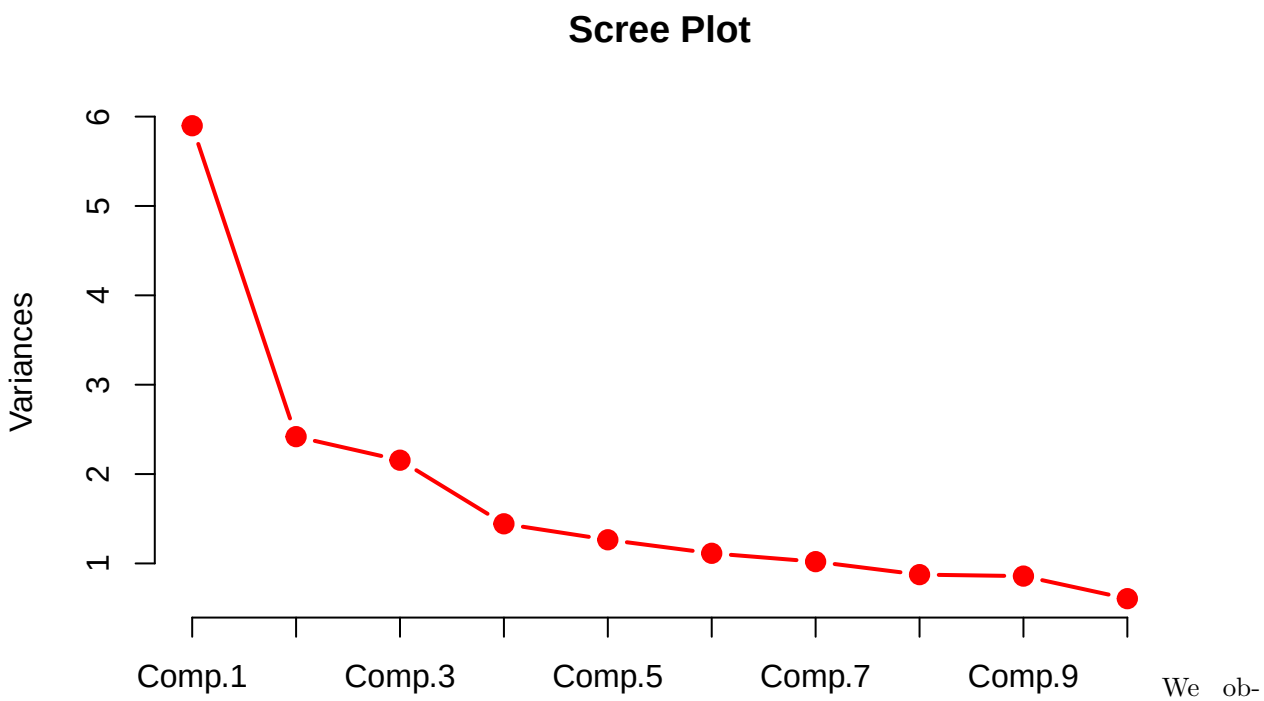
```
(summary(pc1)$sdev)^2
```

```
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
## 5.89786159 2.41876190 2.15518028 1.44323744 1.26468487 1.11319700
##     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
## 1.01975431 0.87448255 0.85804077 0.60577743 0.50060157 0.46825146
##    Comp.13    Comp.14    Comp.15    Comp.16    Comp.17    Comp.18
## 0.36119083 0.31144853 0.22666067 0.15521975 0.12428890 0.10536061
##    Comp.19    Comp.20
## 0.05367185 0.04232768
```

It appears that the first 7 principal components have eigenvalues above 1.

Next, we check the screeplot to look for "elbows":

```
screeplot(pc1,type="lines",col="red",lwd=2,pch=19,cex=1.2,main="Scree Plot")
```



We observe from the graph that two of the significant "elbows" in the plot are at the 2nd and the 4th principal component.

Thus, we will retain the first 3 principal components. This will explain about 52.4% of the variance, and all of the eigenvalues are greater than 1. Furthermore, these 3 principal components occur before the second "elbow" in the screeplot. The eigenvalues of the first 3 principal components are also above thresholds of the Longman and Allen methods.

Examine Loadings

The following code displays the loadings of the first three principal components:

```
print(loadings(pc1,digits=2)[,1:3])
```

```
##                Comp.1      Comp.2      Comp.3
## legal         0.17763135 -0.31917187 -0.38815601
## dangerous    -0.14523821  0.34988847 -0.08297500
## regret       -0.08453309 -0.13865500 -0.14277309
## unnatural    -0.24969426  0.21493513  0.01191595
```

```
## notuse       -0.31614449 -0.21072668  0.17059737
## psycho       -0.22205923  0.33739222 -0.15674177
## trip          0.24448449 -0.17432472 -0.39211123
## stoned        0.23987428  0.36012620 -0.08198553
## calm          0.14077888  0.30862726  0.11714848
## high          0.24860849  0.23904242 -0.20926981
## noaspirin    -0.12996591 -0.43745138 -0.03680139
## relationship  0.30260288  0.01626722  0.02562207
## drugscene     0.25906089 -0.03596770  0.19361236
## caregivers   -0.22718912  0.06830758 -0.24129441
## experience    0.19335131  0.08918780 -0.35397898
## fun           0.20840388 -0.10237294  0.08432925
## stupid       -0.21519668 -0.04083694 -0.24098389
## lessalcohol   0.17298491 -0.09310701 -0.25483367
## sideeffects  -0.21307167  0.11115418 -0.40551110
## dope         -0.31627224 -0.02931426 -0.18076839
```

For the first principal component, the variables "notuse", "relationship", and "dope" had loadings with a magnitude of at least 0.3.

For the second principal component, the variables "legal", "dangerous", "psycho", "stoned", "calm", and "noaspiring" had loadings with magnitudes of above 0.3.

For the third principal component, the variables "legal", "trip", "experience", and "sideffects" had loadings with magnitudes of above 0.3.