Assignment 1

Group 11

The Date

0. Load Libraries and Dependencies

```
library(tidyverse)
library(nFactors)
```

1. Problem Statement

The presented data from a conducted study shows the recovery status for 100 participants after taking two different drugs with differing doses. Recovery is measured as the percentage drop in body pathogens before and after taking the respective drug. In order to judge on the effectiveness of the drugs the principal component analysis is applied, to shrink the information to only a subset of the variables' dimensions. The associated identification of useful factor brandings and the reflection of the data on the identified subdimensions should help to evaluate the initial question.
The second section starts with the data preparation and a brief description of the data. Afterwards, in section three, the assumptions for using principal component analysis are described including the way we deal with them. In section four the principal component analysis is applied and the most relevant outputs are described. The last section covers the answering of first, how the factors in the subspace can be labled appropriately and second, which drugs were effective.

2. Descriptive Statistics

Before dealing with the principal component analysis (PCA), it is important to prepare the data. After importing the data and performing the glimpse-function, one can see that the data is stored in a data.frame object with the dimension 100x9.

```
drugdata <- read.table("http://feb.kuleuven.be/martina.vandebroek/public/STATdata/drugsrecovery.txt", h
class(drugdata)
```

```
## [1] "data.frame"
```

```
glimpse(drugdata)
```

```
## Observations: 100
## Variables: 9
## $ ID    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ L500  <int> 15, 10, 10, 10, 10, 20, 15, 5, 15, 10, 5, 20, 15, 20, 20...
## $ L1000 <int> 20, 15, 15, 15, 10, 20, 15, 5, 15, 10, 10, 20, 15, 30, 1...
## $ L2000 <int> 25, 5, 30, 5, 5, 20, 15, 5, 15, 5, 10, 25, 5, 20, 20, 20...
## $ L4000 <int> 30, 15, 30, 5, 25, 5, 35, 10, 55, 35, 20, 40, 30, 75, 30...
## $ R500  <int> 15, 15, 15, 5, 15, 15, 20, 5, 15, 5, 20, 10, 5, 20, 20, ...
## $ R1000 <int> 20, 20, 15, 10, 5, 20, 20, 10, 15, 10, 15, 10, 5, 20, 10...
## $ R2000 <int> 20, 20, 25, 5, 5, 15, 20, 15, 5, 5, 5, 20, 5, 15, 15, 20...
## $ R4000 <int> 30, 30, 40, 25, 65, 35, 25, 20, 25, 30, 20, 30, 25, 65, ...
```

First of all the first column has to be eliminated, as it only stores the ID values. It would be a major mistake taking this column into account when performing PCA, as the results would be heavily biased.

```
drugdata <- drugdata[-1]
glimpse(drugdata)
```

```
## Observations: 100
## Variables: 8
## $ L500  <int> 15, 10, 10, 10, 10, 20, 15, 5, 15, 10, 5, 20, 15, 20, 20...
## $ L1000 <int> 20, 15, 15, 15, 10, 20, 15, 5, 15, 10, 10, 20, 15, 30, 1...
## $ L2000 <int> 25, 5, 30, 5, 5, 20, 15, 5, 15, 5, 10, 25, 5, 20, 20, 20...
## $ L4000 <int> 30, 15, 30, 5, 25, 5, 35, 10, 55, 35, 20, 40, 30, 75, 30...
## $ R500  <int> 15, 15, 15, 5, 15, 15, 20, 5, 15, 5, 20, 10, 5, 20, 20, ...
## $ R1000 <int> 20, 20, 15, 10, 5, 20, 20, 10, 15, 10, 15, 10, 5, 20, 10...
## $ R2000 <int> 20, 20, 25, 5, 5, 15, 20, 15, 5, 5, 5, 20, 5, 15, 15, 20...
## $ R4000 <int> 30, 30, 40, 25, 65, 35, 25, 20, 25, 30, 20, 30, 25, 65, ...
```

One can also see that all the remaining values are of class integer. This is important to know, as correlation matrices can only be calculated using either integers or numerical values. After selecting the relevant variables, it is also useful to check for missing values. This is done by using the apply-function, which outputs the amount of missing values for each column.

```
drugdata %>% apply(2, function(x){
  sum(is.na(x))
})
```

```
##  L500 L1000 L2000 L4000  R500 R1000 R2000 R4000
##     0     0     0     0     0     0     0     0
```

Hence, there arent any missing values presented in the data. Even though the PCA has no distributional assumptions regarding the data it might still be useful to get an idea of the distribution of the variables.

```
summary(drugdata)
```

```
##       L500           L1000           L2000           L4000
##  Min.   : 5.0   Min.   : 5.0   Min.   : 5.00   Min.   : 5.00
##  1st Qu.: 5.0   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:23.75
##  Median :10.0   Median :15.0   Median :15.00   Median :35.00
##  Mean   :12.2   Mean   :14.5   Mean   :17.00   Mean   :36.35
##  3rd Qu.:20.0   3rd Qu.:20.0   3rd Qu.:21.25   3rd Qu.:50.00
##  Max.   :30.0   Max.   :35.0   Max.   :60.00   Max.   :85.00
##       R500           R1000           R2000           R4000
##  Min.   : 5.0   Min.   : 5.0   Min.   : 5.0   Min.   : 5.00
##  1st Qu.: 5.0   1st Qu.:10.0   1st Qu.:10.0   1st Qu.:20.00
##  Median :10.0   Median :15.0   Median :15.0   Median :30.00
##  Mean   :12.4   Mean   :14.3   Mean   :16.6   Mean   :36.35
##  3rd Qu.:15.0   3rd Qu.:20.0   3rd Qu.:20.0   3rd Qu.:45.00
##  Max.   :40.0   Max.   :35.0   Max.   :50.0   Max.   :90.00
```

As one would expect, the range of the variables increase as the dose increases. Taking into consideration that the minimum of each variable is equal to five, this means that due to the higher dose there might be at least some participants whose percentage drop in body pathogens was quite high compared to the participants who got a lower dose. This can also be derived when looking at the mean values or the outlier resistant median values. This might give rise to the assumption that higher dose is more effective. But let's see what is the outcome of the PCA.

3. Assumptions

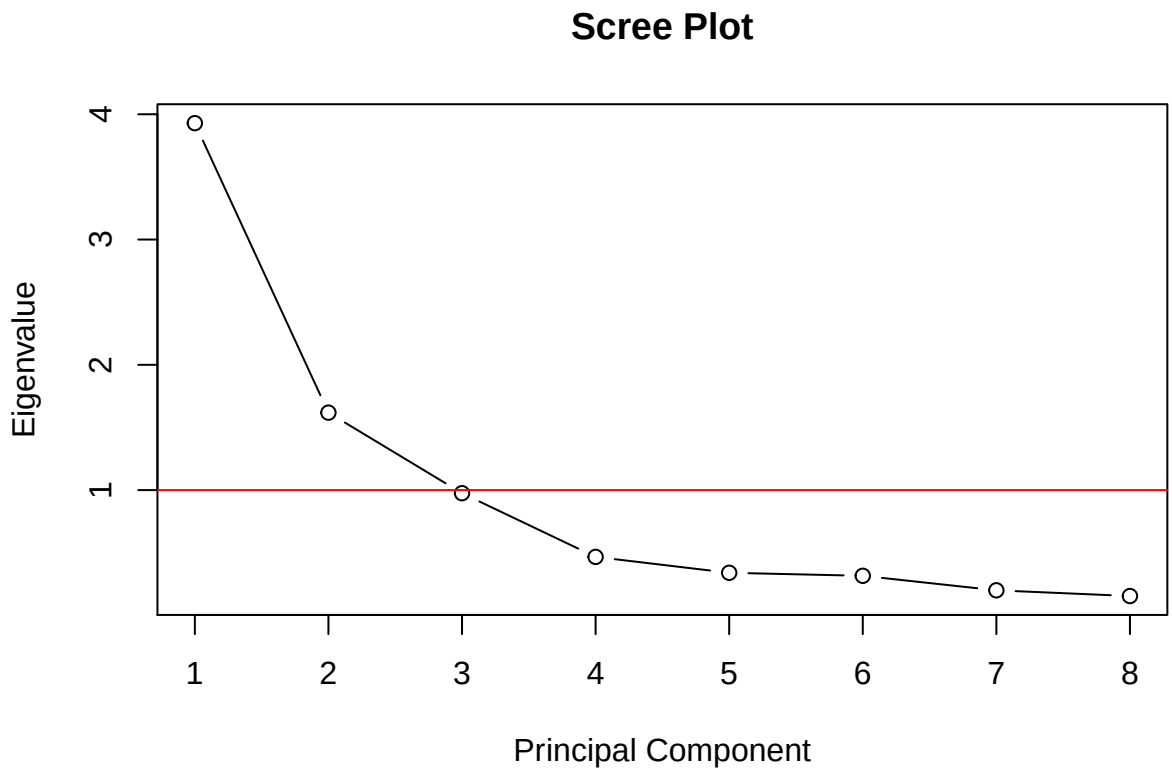
When dealing with PCA there is basically one assumption which has to be met, namely the variables have to be either mean centered or standardized. As all variables are measured in the same scale (from 0% to 100%) the variance observed to not artificial, but important information which has to be considered. For example, if higher drug doses lead to a higher variance then this information has to be included in the model. Therefore mean centering is applied when perfoming PCA. This is done in the upcoming section when using the prcomp-function with the center parameter set to TRUE.

4. Method

Before applying the method the first task is to identify the appropriate number of principal components that effectively summarize the data. Hence, we will decide based on two commonly used procedures. First by visually analyzing the scree plot and second by performing Horn's Parallel procedure.
The scree plot is just the visualization of the eigenvalues in decreasing order. Hence, in the following chunk the eigenvalues are extracted from the correlation matrix of the data and the eigenvalues are ploted. The red line highlights the threshold value of one.

```
# Eigenvalues
eval <- eigen(cor(drugdata))$values

# Scree Plot
plot(eval, xlab = "Principal Component", ylab = "Eigenvalue",
     type = "b", main = "Scree Plot")
abline(h=1,col="red")
```



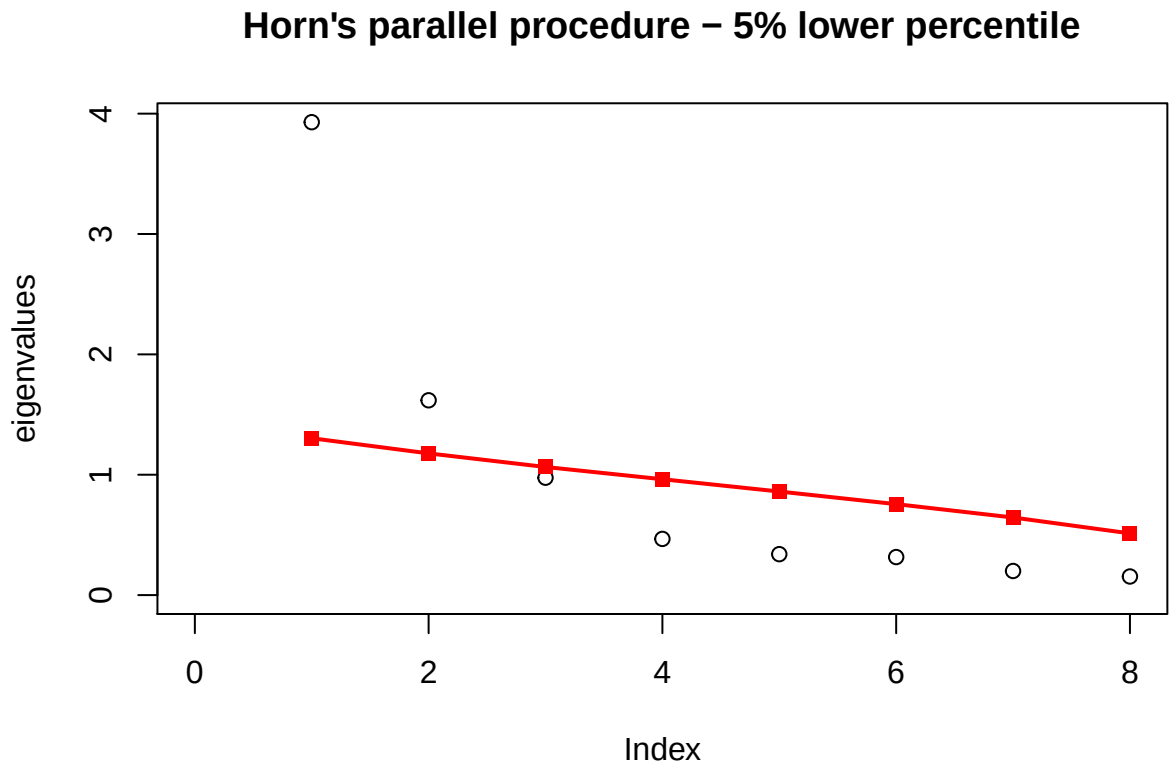
The first characteristic to find would be the so called elbow, which separates the mountain from the debris. As the slope has almost an exponential form, it is difficult to see whether such a distinction can be made. The second criterion to check is to identify which eigenvalues lay above the threshold value of one und which lay below. The first and the second principal component are clearly above the threshold value. The third principal component is a borderliner, as it's value of 0.9753248 lies slightly below the threshold value.

- Horn's Parallel procedure

Compute Eigenvalues associated with many simulated uncorrelated normal variables - retain the ith PC if the corresponding eigenvalue is larger than the 95th percentile of the distribution of the ith largest eigenvalue of the random data (same idea as the previous rule but taking random variation into account)

```
ap <- parallel(subject = nrow(drugdata), var = 8, rep= 1000,cent = 0.05)

plot(eval, xlim=c(0,ncol(drugdata)), ylim = c(0,max(eval)), ylab = "eigenvalues")
lines(ap$eigen$qevpea, type = "o",pch = 15, lwd = 2,col = "red")
title(main= "Horn's parallel procedure - 5% lower percentile")
```



- Perform PCA

```
pr.out <- prcomp(scale(drugdata, center = T))
```

- Variance explained

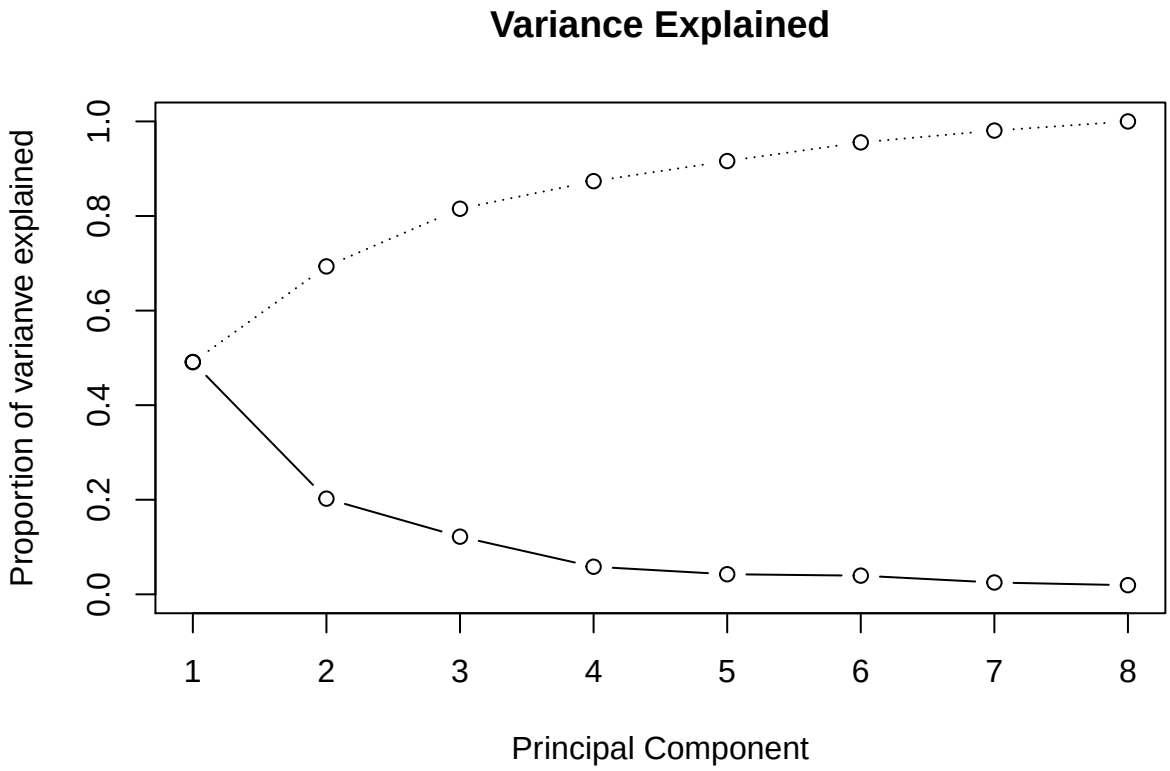
```
summary(pr.out)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.9822 1.2721 0.9876 0.68321 0.58317 0.56204
```

```
## Proportion of Variance 0.4911 0.2023 0.1219 0.05835 0.04251 0.03949
## Cumulative Proportion  0.4911 0.6934 0.8153 0.87368 0.91619 0.95568
##                            PC7     PC8
## Standard deviation     0.44734 0.39303
## Proportion of Variance 0.02501 0.01931
## Cumulative Proportion  0.98069 1.00000
```

```r
# First two principal components explain 69.34% of the variance

# Variance Explained Plot
prop_varex <- eval / sum(eval)
plot(prop_varex, xlab = "Principal Component", ylab = "Proportion of varianve explained",
     type = "b", main = "Variance Explained", ylim = c(0,1))
lines(cumsum(prop_varex), type = "b", lty = 3)
```



5. Interpretation

```r
# Eigenvectors
evec <- pr.out$rotation


# Correlation coefficients between PC's and initial variables
varnames <- names(drugdata)
corrcoef <- matrix(nrow = 2, ncol = ncol(drugdata), byrow = T)
```

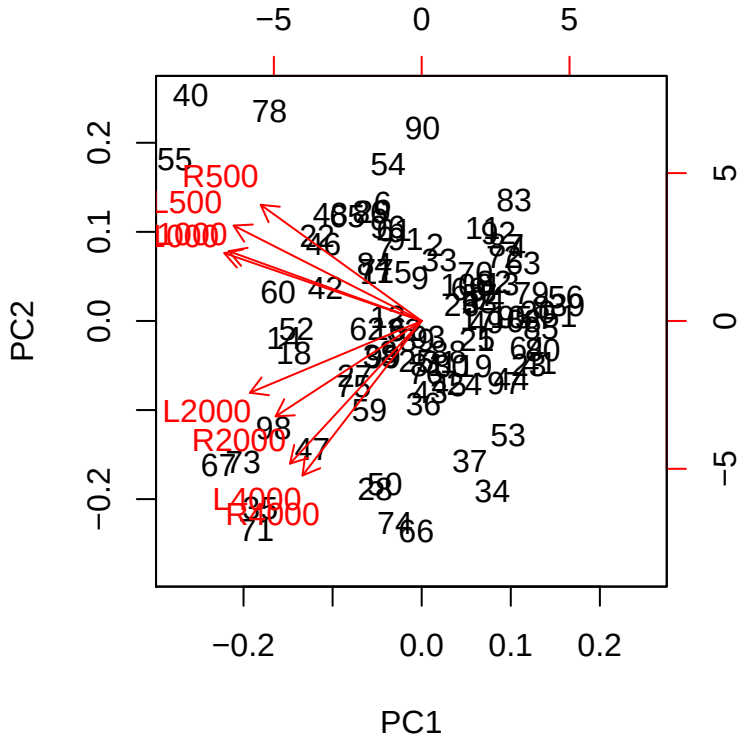
```r
colnames(corrcoef) <- varnames
rownames(corrcoef) <- c("PC1", "PC2")

for( n in 1:length(rownames(corrcoef)) ){
  for( m in 1:nrow(evec) ){
    corrcoef[n,m] <- evec[m,n] * sqrt(eval)[n] / sd(drugdata[,m])
  }
}
corrcoef
```

```
##             L500     L1000       L2000       L4000        R500       R1000
## PC1 -0.12405729 -0.110217 -0.06637818 -0.02838059 -0.09550936 -0.12748647
## PC2  0.06291815  0.037883 -0.02774273 -0.03075025  0.06893419  0.04609256
##            R2000       R4000
## PC1 -0.06647428 -0.02606828
## PC2 -0.04341696 -0.03379422
```

```r
# First principal component is significantly negatively correlated
# with all variables
# Second component discriminates between L2000, L4000, R2000, R4000 on one hand
# (high dose) and L500, L1000, R500 and R1000 on the other hand (low dose)
```

```r
biplot(pr.out, choices = c(1,2))
```



first plot shows two groupings of dosage amount independant of the drug

second plot barely shows any relevant information as there is barely any group seperation -> was to be expected since they explain small portion of the variance only

For the majority of the patients, both drugs had barely any effect on the percentage drop. Lower dose seems to be more effective compared to higher dose, as with a lower dose more patients had a percentage drop.