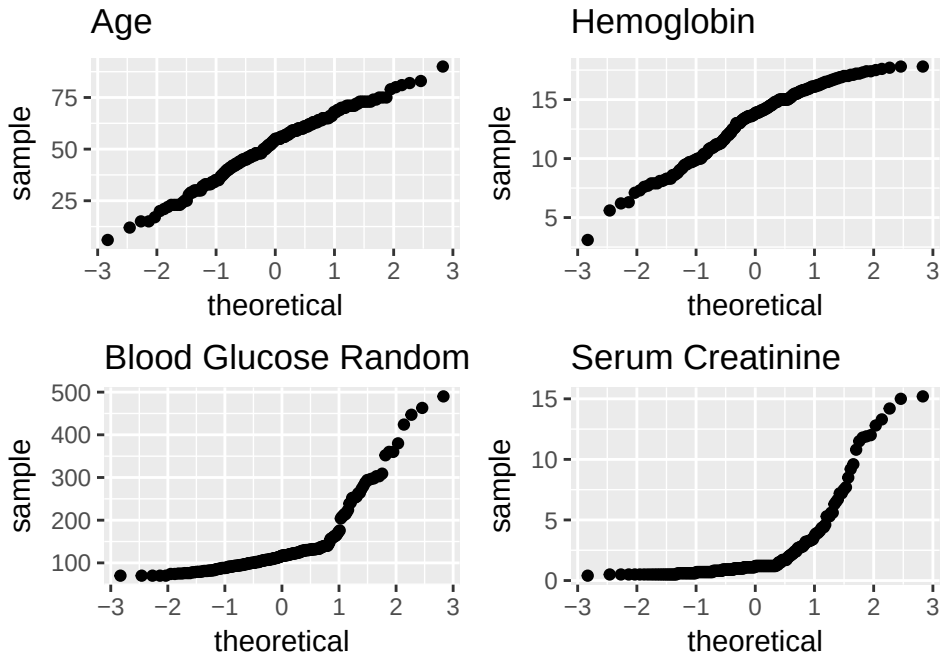
S&DS 563 / F&ES 758b - Multivariate Statistics Homework #2
Principle Components Analysis

Lanxin Jiang (lj345), Grace Sun (ys544), Chenglin Lu (cl939)

2018-02-12

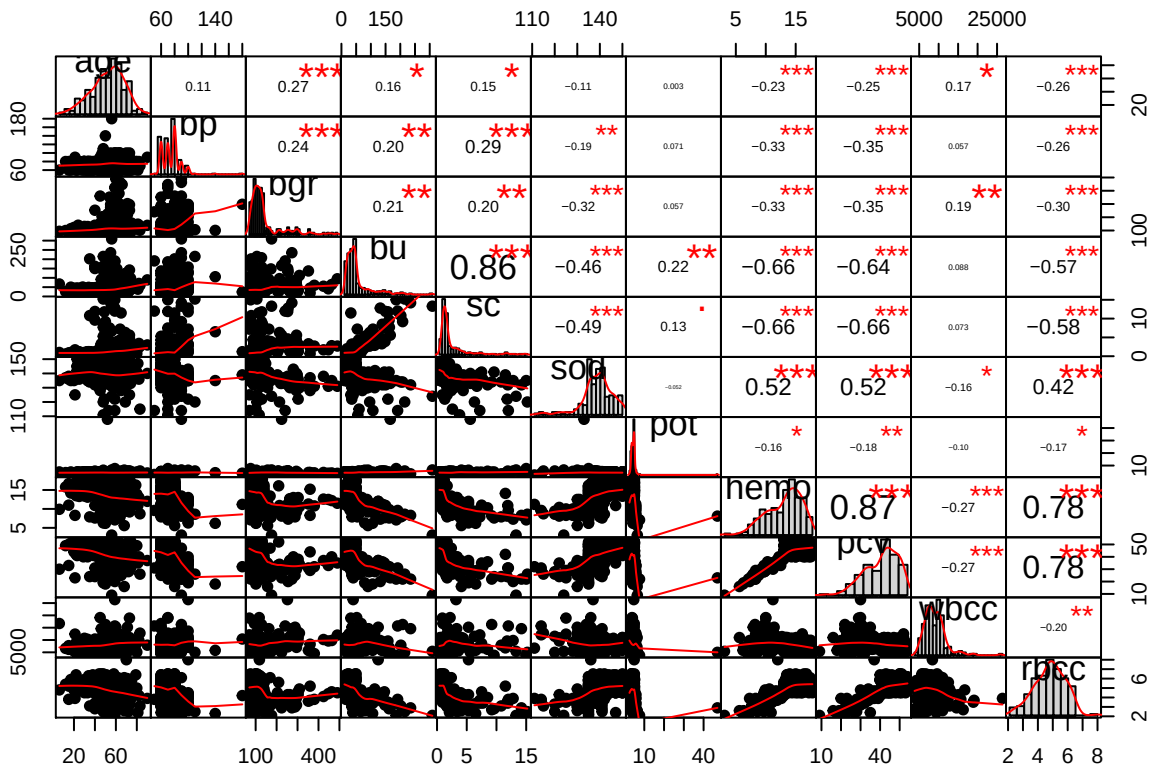
1. First, discuss whether your data seems to have a multivariate normal distribution. Make univariate plots (boxplots, normal quantile plots as appropriate). Then make transformations as appropriate. You do NOT need to turn all this in, but describe what you did.

```
library(RWeka)
library(ggplot2)

CKD <- read.arff("../Chronic_Kidney_Disease/chronic_kidney_disease_full.arff")

# Only Numeric Variables
CKD.numeric <- CKD[,c(1:2,10:18)]
# Remove Missing Observations
CKD.numeric <- CKD.numeric[complete.cases(CKD.numeric),]
#check QQ plot for some numeric variables
gg.qqplot <- function (variable, title) {
    ggplot(CKD.numeric, aes_string(sample=variable)) + ggtitle(title) + stat_qq()
}
p1 <- gg.qqplot("age", "Age")
p2 <- gg.qqplot("bgr", "Blood Glucose Random")
p3 <- gg.qqplot("hemo", "Hemoglobin")
p4 <- gg.qqplot("sc", "Serum Creatinine")
multiplot(p1, p2, p3, p4, cols=2)
```



look for non-linearity, get correlation, make histograms all at once

```
chart.Correlation(CKD.numeric, histogram=TRUE, pch=19)
```

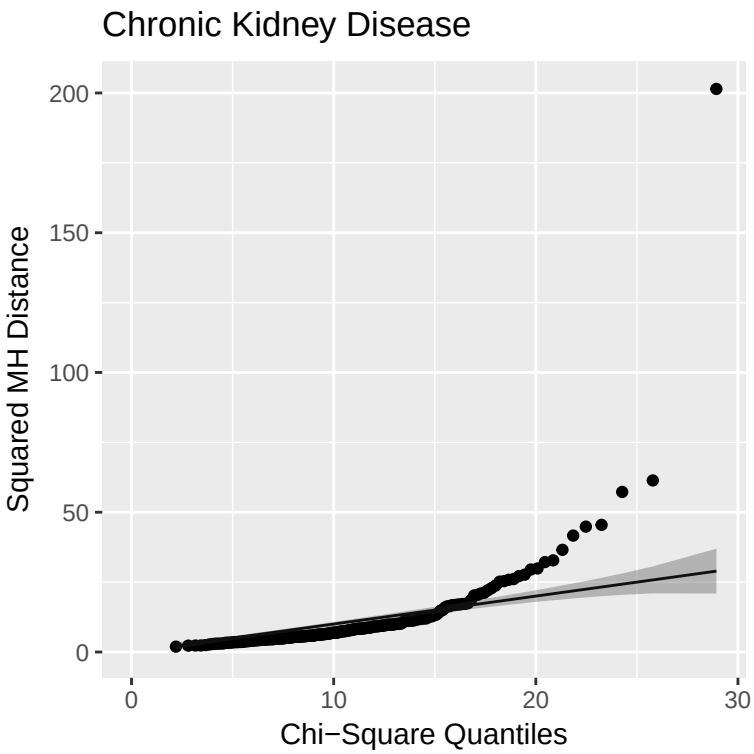


From the above Q-Q plots and histograms, we could observe that most of our variables don't follow the univariate normal distribution. Age and RBCC seem to be normal. BP, BGR, BU, SC, and WBCC are left-skewed, while the others are right-skewed.

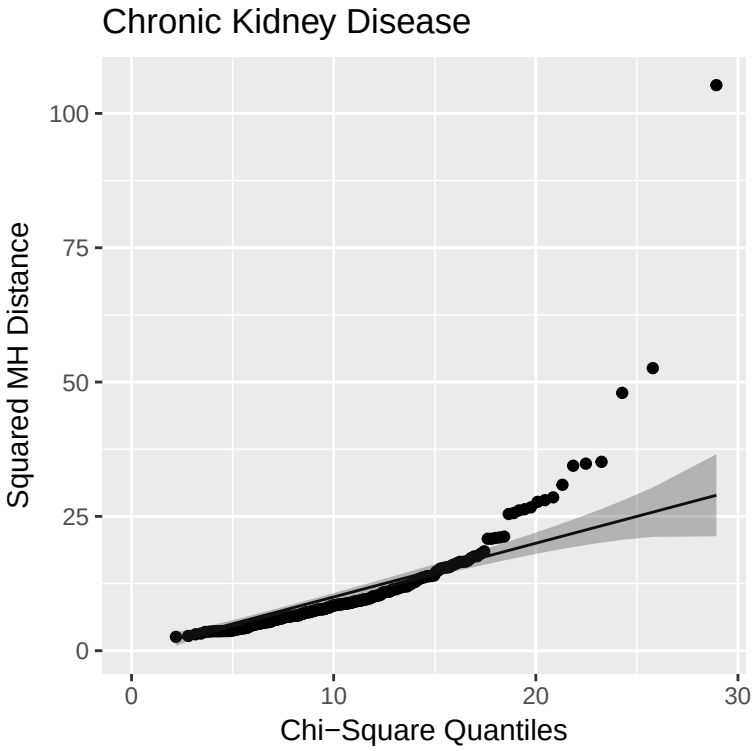
THEN make a chi-square quantile plot of the data. Turn in your chi-square quantile plot as appropriate and comment on what you see.

NOTE that multivariate normality is NOT a requirement for PCA to work!

```
gg.CSQPlot(CKD.numeric,label="Chronic Kidney Disease")
```




```
#log transform our data
gg.CSQPlot(log(CKD.numeric),label="Chronic Kidney Disease")
```

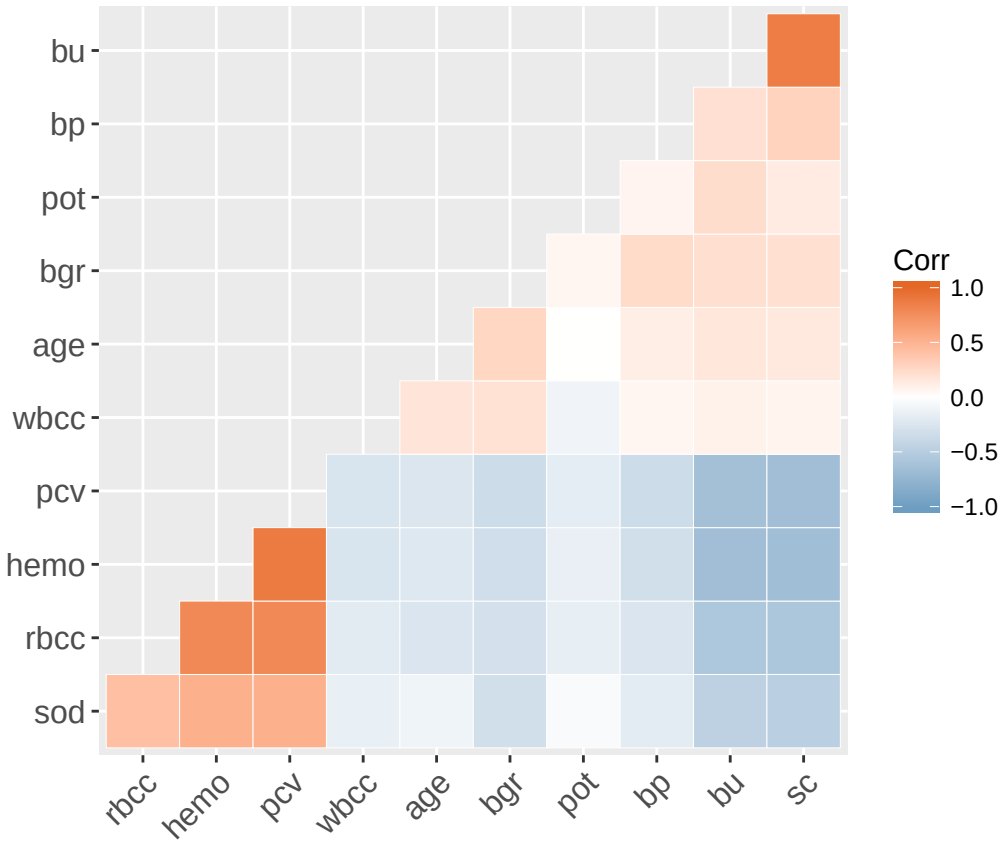


Our data doesn't have multivariat normality. In addition, log transformation doesn't apply to this data. By testing other transformation methods, we conclude that it isn't appropriate to use transformation to achieve normality for our data.

2. Compute the correlation matrix between all variables (SAS and SPSS will provide this for you as part of the PCA procedure in SPSS, click on DESCRIPTIVES. In R use the cor() function.). Comment on relationships you do/do not observe. Do you think PCA will work well?

```
library(ggcorrplot)
CKD.numeric %>%
```

```
  cor(use="pairwise.complete.obs", method = "pearson") %>%
  ggcorrplot(hc.order = TRUE, type = "lower", # lab = TRUE,
    outline.col = "white",
    ggtheme = ggplot2::theme_gray,
    colors = c("#6D9EC1", "white", "#E46726"))
```



From the above correlation matrix plot, the more "orange" indicates that two variables more positively correlated, and the more "blue" indicates more negative correlation. We could observe some negative correlation and some positive ones, but nearly half of them are rarely correlated with one another. PCA could work for this data because of some of the correlation, but it might not work so effectively.

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

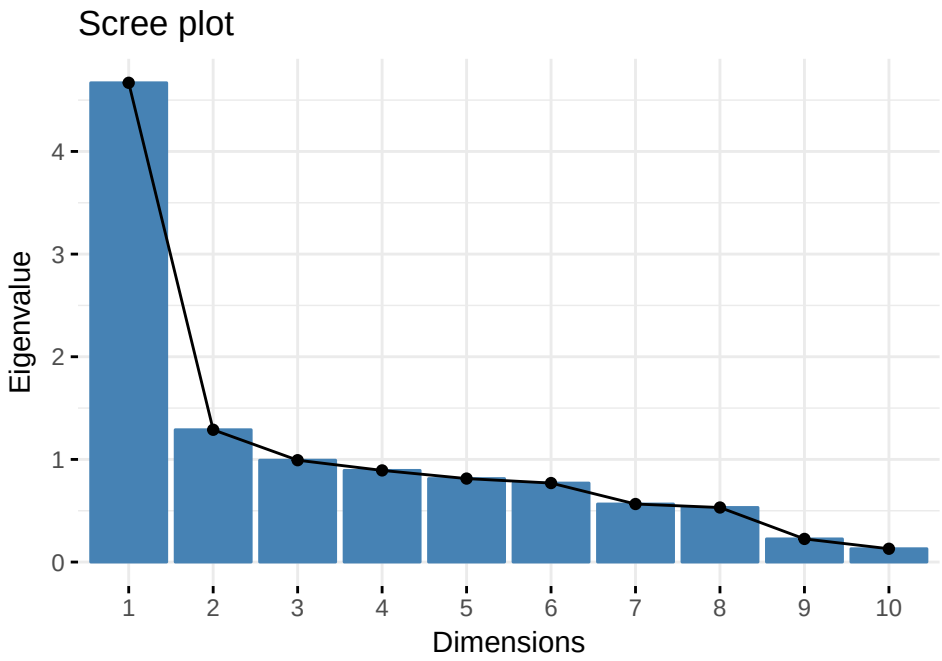
- Total variance explained by a given number of principle components
- The 'eigenvalue > 1' criteria
- The 'scree plot elbow' method (turn in the scree plot)
- Parallel Analysis: think about whether this is appropriate based on what you discover in number 1.

```
library(factoextra)

pc1 <- princomp(CKD.numeric, cor=TRUE)

# Importance | stats:::print.summary.princomp
vars <- pc1$sdev^2
vars <- vars/sum(vars)
pander(rbind(`Standard deviation` = pc1$sdev, `Proportion of Variance` = vars,
    `Cumulative Proportion` = cumsum(vars)),
    split.table = Inf,
    caption = "Importance of components"
)
```

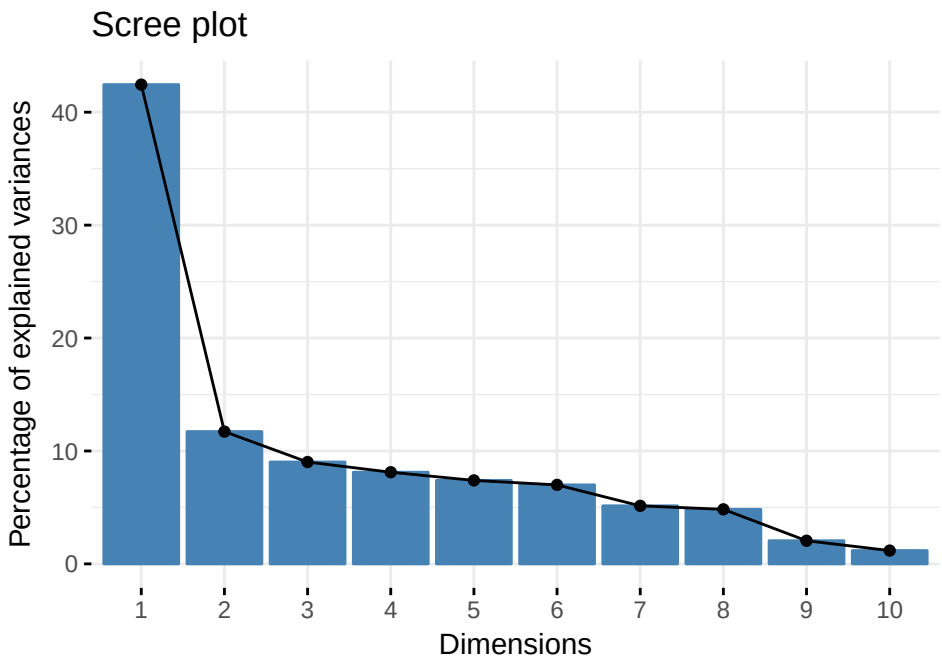
Table 1: Importance of components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
Standard deviation	2.16	1.14	1.00	0.94	0.90	0.88	0.75	0.73	0.47	0.36	0.35
Proportion of Variance	0.42	0.12	0.09	0.08	0.07	0.07	0.05	0.05	0.02	0.01	0.01
Cumulative Proportion	0.42	0.54	0.63	0.71	0.79	0.86	0.91	0.96	0.98	0.99	1.00

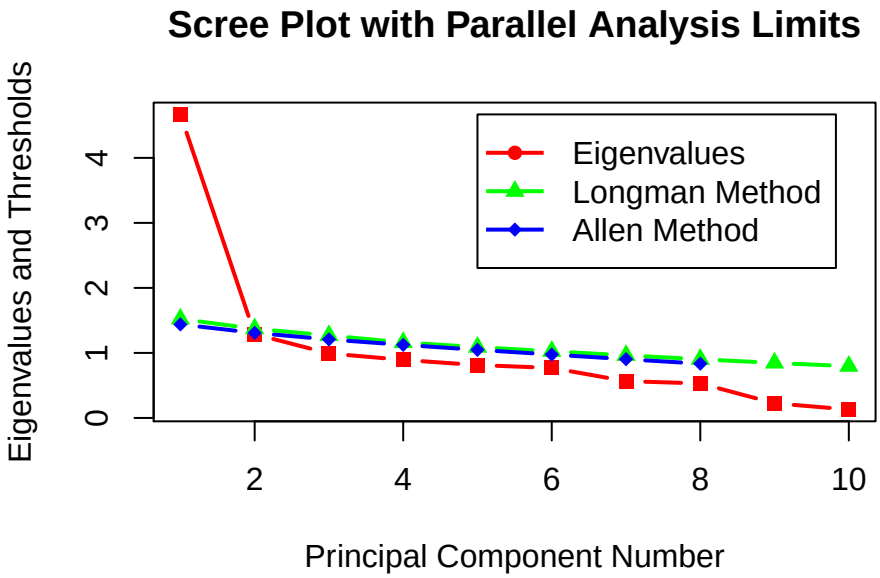
```
# Screeplot
fviz_screeplot(pc1, choice="eigenvalue")
```




```
fviz_screeplot(pc1, choice="variance")
```



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

# Parallel analysis plot
parallelplot(pc1)
```



The first two compenents explain 54% variability. 91% of the overall variability is explained by the first seven principal components.

The first two eigenvalues are larger than 1.

The 'scree plot elbow' method shows that the first two components should retain.

Even though the parallel analysis also shows the first two components should retain, this method isn't appropriate due to nonnormality.

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

```
# Loadings
pander(unclass(pc1$loadings), split.table = Inf)
```

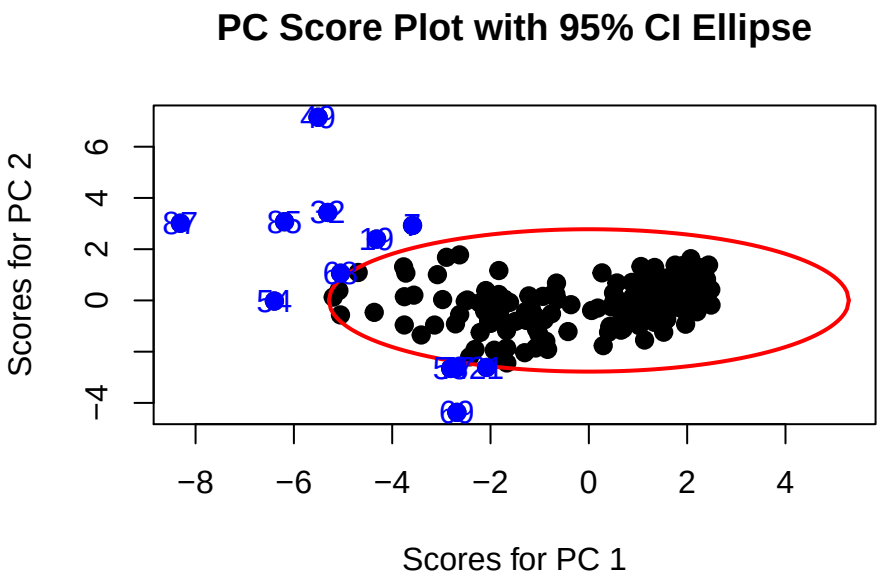
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
age	-0.15	-0.44	0.35	0.39	0.61	0.19	0.21	-0.23	0.03	0.01	-0.02
bp	-0.19	-0.07	0.41	-0.74	-0.09	0.42	0.20	-0.10	-0.08	0.08	0.04
bgr	-0.21	-0.39	0.39	-0.14	-0.05	-0.62	-0.13	0.47	0.00	-0.02	-0.02
bu	-0.37	0.29	-0.15	0.09	0.18	-0.02	0.34	0.35	-0.02	0.53	0.44
sc	-0.38	0.26	-0.18	-0.06	0.23	0.03	0.32	0.29	-0.11	-0.55	-0.44
sod	0.30	0.02	0.19	0.15	0.04	0.55	-0.26	0.68	0.11	-0.03	-0.01
pot	-0.10	0.42	0.61	0.42	-0.45	-0.03	0.20	-0.11	-0.02	-0.05	-0.07
hemo	0.42	0.00	0.08	-0.03	0.09	-0.11	0.29	0.06	-0.47	-0.42	0.55
pcv	0.42	0.01	0.03	-0.03	0.09	-0.12	0.27	0.09	-0.45	0.47	-0.54
wbcc	-0.13	-0.56	-0.29	0.22	-0.57	0.23	0.37	0.12	-0.10	-0.02	-0.02
rbcc	0.38	0.00	0.02	-0.13	0.02	-0.15	0.53	0.06	0.73	-0.04	-0.03

Component one is mostly hemoglobin(hemo) and packed cell volume(pcv), component two is mostly age, potassium(pot), white blood cell count(wbcc). Here our cutoff is 0.4. These two components can explain 54% of the variability. There might be clinical explanation for this result. Kidney experts could dig deeper into this phenomenon.

5. Make a score plot of the scores for at least two pairs of component scores (one and two, one and three, two and three, etc). Discuss any trends/groupings you observe. **As a bonus, try to make a 95% Confidence Ellipse for two of**

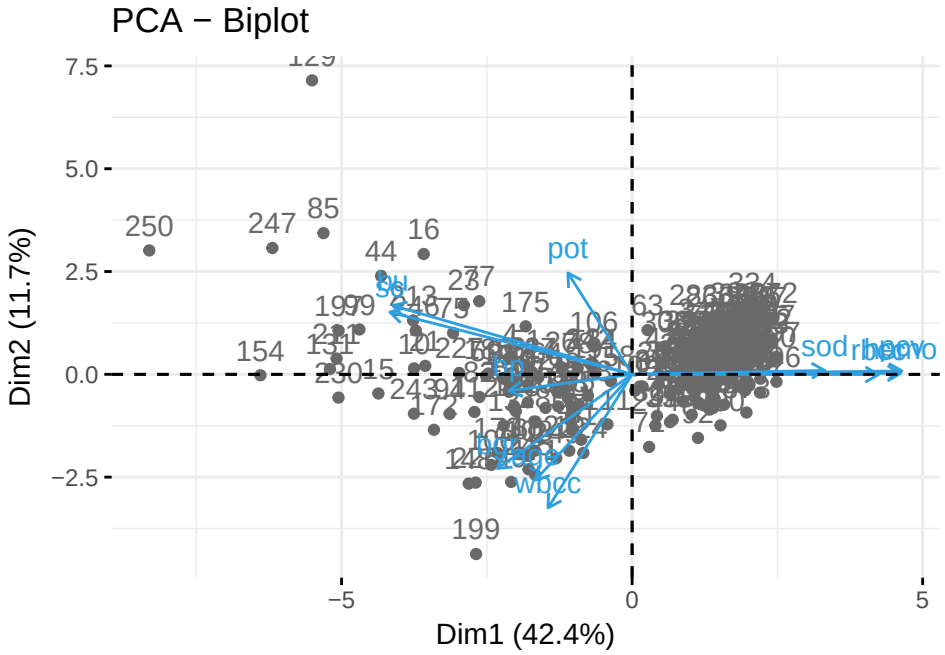
your components. You might want to also try making a bi-plot if you are using R.

```
#make scoreplot with confidence ellipse :
#  c(1,2) specifies to use components 1 and 2
#get function from online

source("http://reuningscherer.net/stat660/r/ciscoreplot.R.txt")

#run the function
ciscoreplot(pc1,c(1,2),c(1:pc1$n.obs))
```



By projecting our data onto the two-dimensional subspace defined by the first and second, or the first and the third principal component axes, we could observe obvious groupings of the black points and some outliers(blue plots), with the red circle as the 95% confidence ellipse. However, the score plot isn't appropraite to our data because our variables don't follow the multivariate normal distribution.

```
# Biplot for first two components
fviz_pca_biplot(pc1, repel = F,
            col.var = "#2E9FDF", # Variables color
            col.ind = "#696969"  # Individuals color
            )
```



The biplot shows the plot of the principal component scores onto which is superimposed a vector representing the coefficients for each of the original variables. Hemoglobin(hemo) and packed cell volume(pcv) are the main vectors in the horizontal (PC1) direction, so we know that the points 254, 247 have unusually small values for these variables.

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.

   The first two components can explain 54% of the variability. Our dataset has obvious correlation between some variables but not among others. Scatterplots of linearity shows variables have linearity but may not be obvious enough. We found some outliers in our score plot and bi-plot. We have 215 observations and 11 variables in the complete case. The ratio of observation to variable is larger than 10, so this sample size works well for PCA. Overall, PCA is effective for this data because of high correlation between some variables, but this method has some limitations in that there is little correlation between certain variables and nonnormality of this dataset.