Assignment 2

```
library(tidyverse)
```

1. For the data matrix below:

```
x <- matrix(c(4 ,1,-1,-3,1, 2,0,-1,5,-1), nrow=5)
x
```

```
##      [,1] [,2]
## [1,]    4    2
## [2,]    1    0
## [3,]   -1   -1
## [4,]   -3    5
## [5,]    1   -1
```

(a) Calculate the sample variance-covariance matrix.

```
var(x)
```

```
##       [,1]  [,2]
## [1,]  6.80 -2.25
## [2,] -2.25  6.50
```

(b) Calculate the correlation matrix.

```
cor(x)
```

```
##           [,1]      [,2]
## [1,]  1.000000 -0.338432
## [2,] -0.338432  1.000000
```

(c) Standardize the variables to have mean 0 and standard deviation 1.

```
xs <- scale(x)
xs
```

```
##            [,1]       [,2]
## [1,]  1.3805370  0.3922323
## [2,]  0.2300895 -0.3922323
## [3,] -0.5368755 -0.7844645
## [4,] -1.3038405  1.5689291
## [5,]  0.2300895 -0.7844645
## attr(,"scaled:center")
## [1] 0.4 1.0
## attr(,"scaled:scale")
## [1] 2.607681 2.549510
```

(d) In R find the eigenvectors of the correlation matrix of x.

```r
t(xs)%*%(xs)/(nrow(xs) - 1)
```

```
##           [,1]      [,2]
## [1,]  1.000000 -0.338432
## [2,] -0.338432  1.000000
```

```r
var(xs)
```

```
##           [,1]      [,2]
## [1,]  1.000000 -0.338432
## [2,] -0.338432  1.000000
```

```r
eigen(cor(x))
```

```
## eigen() decomposition
## $values
## [1] 1.338432 0.661568
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7071068 -0.7071068
## [2,]  0.7071068 -0.7071068
```

(e) Using prcomp() function, find the loadings for the principal components of x.

```r
prcomp(x, scale = TRUE)
```

```
## Standard deviations (1, .., p=2):
## [1] 1.1569062 0.8133683
##
## Rotation (n x k) = (2 x 2):
##             PC1       PC2
## [1,] -0.7071068 0.7071068
## [2,]  0.7071068 0.7071068
```

2. Body fat data. The data consists of observations taken on a sample of 88 males. In this question you will look at PCA of the variables variables were measured:

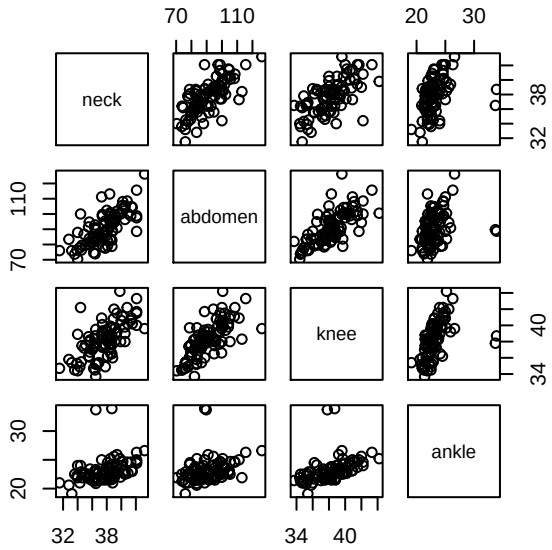
- Neck circumference (cm)
- Abdomen circumference (cm)
- Knee circumference (cm)
- Ankle circumference (cm)

```r
bodyfat <- read.table("data/bodyfat.txt", header = TRUE) %>%
  select(neck, abdomen, knee, ankle)
head(bodyfat)
```

```
##   neck abdomen knee ankle
## 1 36.2    85.2 37.3  21.9
## 2 38.5    83.0 37.3  23.4
## 3 34.0    87.9 38.9  24.0
## 4 37.4    86.4 37.3  22.8
## 5 34.4   100.0 42.2  24.0
## 6 39.0    94.4 42.0  25.6
```

Use pairs to construct a scatterplot matrix. Are there any outliers? If so, which cases are they?

```
pairs(bodyfat)
```




```
bodyfat %>% filter(ankle > 30)
```

```
##   neck abdomen knee ankle
## 1 38.7    88.7 38.7  33.9
## 2 36.5    89.7 37.8  33.7
```

There are two outliers with extreme ankle values, but non extreme values on other variables. They are observations 31 and 84.

(b) Carry out a principal components analysis of the data. What percentage of the variability in the dataset is accounted for by the first component? What percentage of the variability in the dataset is accounted for by the first two components? Examine the scree diagram and comment. (You will find the code for the screeplot in h1code.R).

```
scree_ggplot <- function(p) {
  e <- p$sdev ^ 2
  df <- data.frame(e = e / sum(e), ind = 1:length(e))
  df %>%
    ggplot(aes(ind, e)) +
    geom_line(linetype = "dotted", size = 0.7) +
    geom_point(colour = "orange", size = 2.5) +
```

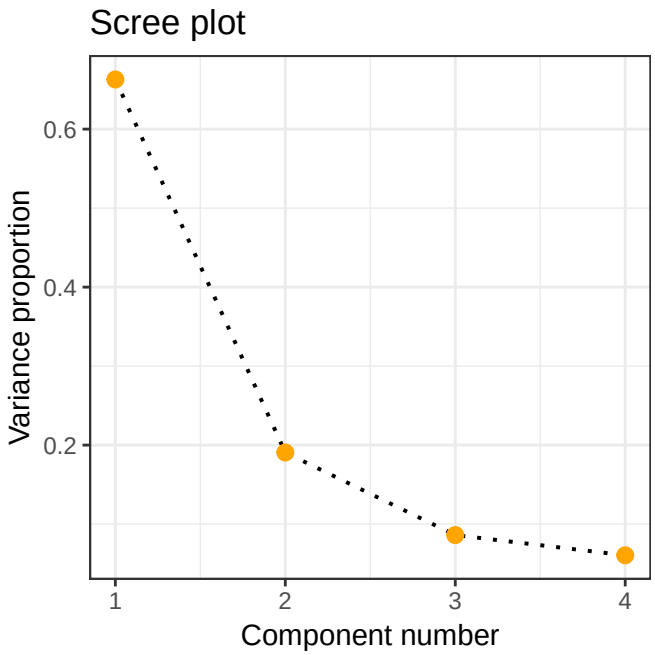
```
    labs(x = "Component number", y = "Variance proportion",
         title = "Scree plot") +
    theme_bw()
  }


p <- prcomp(bodyfat, scale = TRUE)
p$rotation[,1:2]
```

```
##                PC1         PC2
## neck    0.5283837  0.21938447
## abdomen 0.5351101  0.35203936
## knee    0.5460990  0.05660109
## ankle   0.3691120 -0.90814925
```

```
summary(p)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.6283 0.8731 0.58678 0.49183
## Proportion of Variance 0.6629 0.1906 0.08608 0.06047
## Cumulative Proportion  0.6629 0.8535 0.93953 1.00000
```

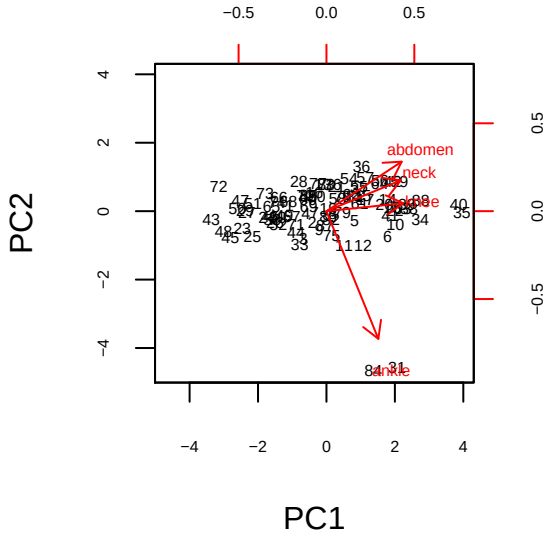
```
scree_ggplot(p)
```



$\approx$ 66% of variability explained by the 1st PC, $\approx$ 85% by the first 2 PCs and $\approx$ 94% by the first 3 PCs.

(c) What does the first component measure? the second component? Make a biplot to assist your interpretations. Are there any outliers? What can you say about the outliers from the plot?

```
biplot(p, scale = 0, cex=c(0.5, 0.5), cex.axis = 0.5)
```



The first component is a weighted average of the variables. It is an overall measure of size. The second component is a contrast of neck and abdomen with ankle. It is a measure of the difference between top size and ankle. The visible outliers are 84 and 31, with the big ankle values.

(d) Omiting any outliers identified, repeat parts (b) and (c).

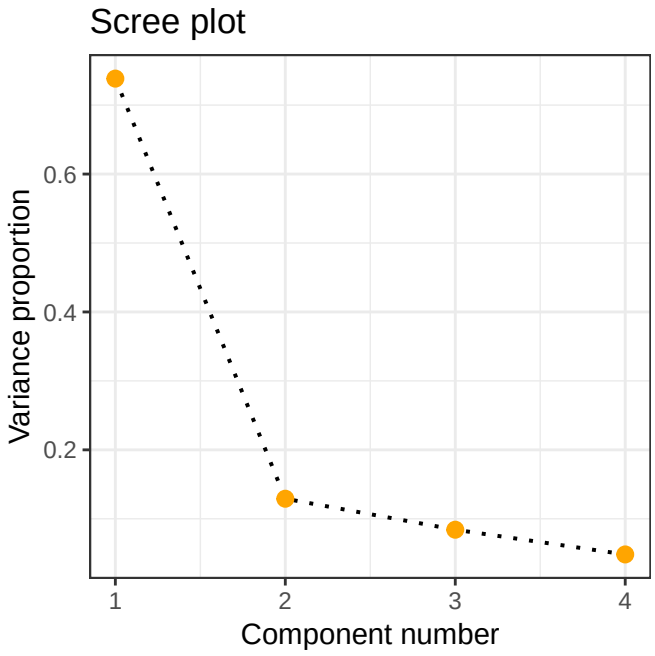
```
p <- prcomp(bodyfat %>% filter(ankle < 30), scale = TRUE)
p$rotation[,1:2]
```

```
##                PC1        PC2
## neck    0.5002906  0.3230518
## abdomen 0.5005250  0.5460175
## knee    0.5251301 -0.1447403
## ankle   0.4726758 -0.7593106
```

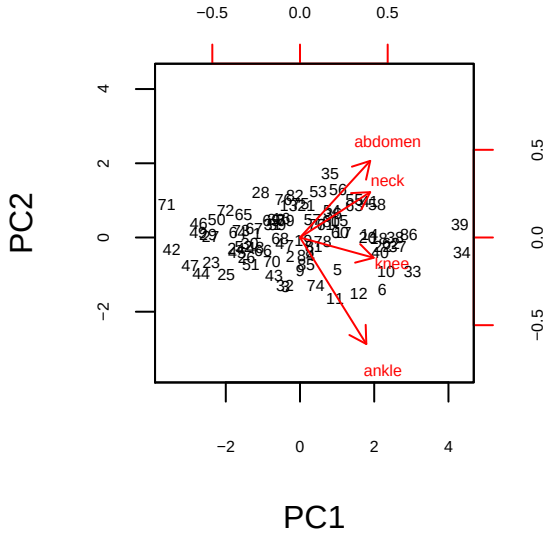
```
summary(p)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7186 0.7186 0.58020 0.43962
## Proportion of Variance 0.7384 0.1291 0.08416 0.04832
## Cumulative Proportion  0.7384 0.8675 0.95168 1.00000
```

```
scree_ggplot(p)
```




```
biplot(p, scale=0, cex=c(.5,.5), cex.axis=.5)
```



The first component is a weighted average of the variables. It is an overall measure of size. The second component is a contrast of neck and abdomen with knee and ankle. It is a measure of the difference between top size and lower size. The high weight people stick out on the first component, but are not that extreme.

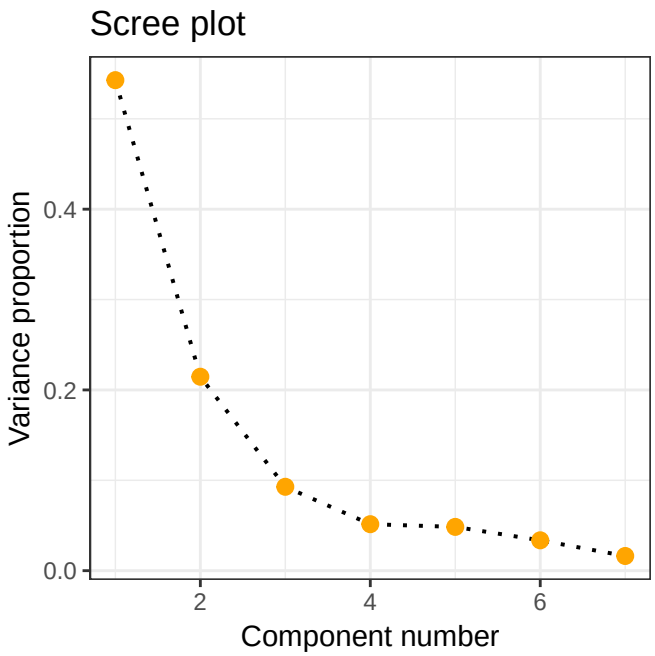
3. A 1902 study obtained measurements on seven physical characteristics for each of 3000 criminals. The seven variables measured were (1) head length (2) head breadth (3) face breadth (4) left finger length (5) left forearm length (6) left foot length (7) height. Using the correlation matrix given below, find the principal components of the data and interpret the results. What percentage of the variability in the dataset is accounted for by the first component? What percentage of the variability in the dataset is accounted for by the first two components? Examine the scree diagram and comment.

```
crimcorr <- matrix(c(
  1.000, 0.402, 0.396, 0.301, 0.305, 0.339, 0.340,
  0.402, 1.000, 0.618, 0.150, 0.135, 0.206, 0.183,
  0.396, 0.618, 1.000, 0.321, 0.289, 0.363, 0.345,
  0.301, 0.150, 0.321, 1.000, 0.846, 0.759, 0.661,
  0.305, 0.135, 0.289, 0.846, 1.000, 0.797, 0.800,
  0.339, 0.206, 0.363, 0.759, 0.797, 1.000, 0.736,
  0.340, 0.183, 0.345, 0.661, 0.800, 0.736, 1.000), nrow = 7, byrow = TRUE)
colnames(crimcorr)<- c("Head-L","Head-B","Face-B",
                       "L-Fing","L-Fore","L-Foot", "Height")
```

```
V <- eigen(crimcorr)

V$values/sum(V$values)
```

```
## [1] 0.54278208 0.21461832 0.09282963 0.05143670 0.04845178 0.03360759
## [7] 0.01627391
```

```
V$sdev <- sqrt(V$values)
scree_ggplot(V)
```



Proportion variance explained by the 1st PC is 0.54, first two is 0.76 etc.First PC is a measure of overall size of the person. Second PC contrasts head measurements with the rest. Third PC is the head length.

4. For each of the following situations, answer, if possible: (i) Is it a classification or regression problem? (ii) Are we most intererest in inference or prediction? (iii) Provide n and p. For each predictor described state whether it is categorical or quantitative. (iv) Indicate whether we would expect the performance of a flexible learning method to be better or worse than an inflexible method.

(a) We have a set of data on 500 worldwide tech firms. For each firm, information on profit, CEO salary, number of employees, average employee salary, and home country is recorded. We are interested in the relationship between CEO salary and other measurements.

Regression, inference, n = 500, 1 response, 4 predictors. All predictors are quantitative, except country which is categorical. Inflexible better for inference.

(b) A company wishes to launch a new product. They want to know in advance whether it will be a success or failure. They collect data on 20 similar products, and record whether they succeeded or not, price charged, marketing budget, and 10 other variables.

Classification, prediction, n = 20, 1 response (binary), 12 predictors. All described predictors are quantitative. Inflexible
better because there are so many predictors relative to n.

(c) A dataset was collected to related the birthweight of babies to the days of gestation and gender.

Regression, inference, n = unknown, 1 response (quantitative), 2 predictors, birthweight quantitative and gender categorical. Inflexible better to understand predictors response association.

(d) Observations were collected on 56 attributes from 32 lung cancer patients belonging to one of 3 classes.

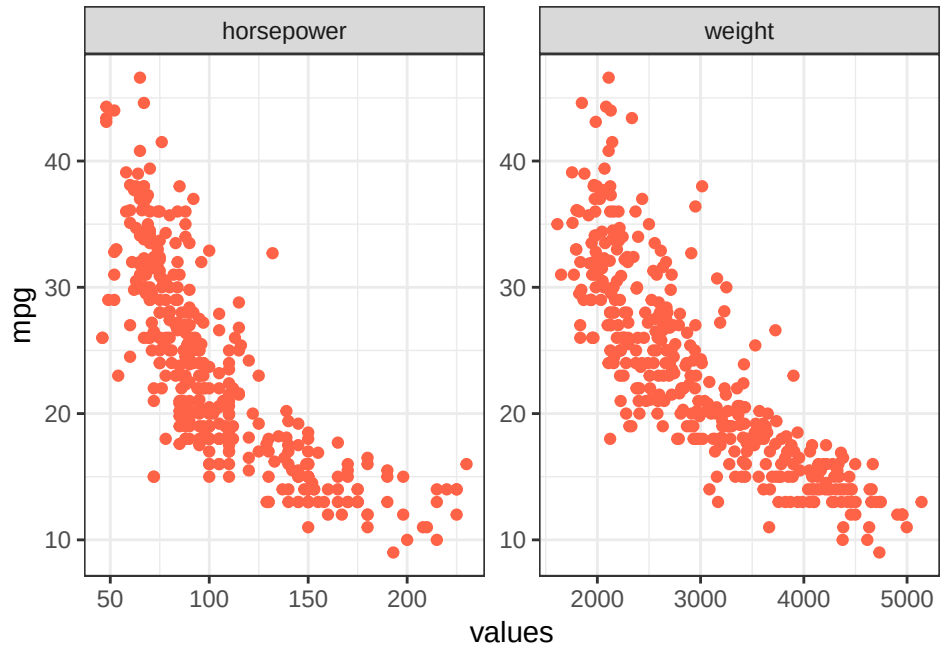
Classification, prediction, n = 32, 1 response (categorical, 3 classes), 56 predictors, structure unknown. Inflexible since n is so large relative to p.

5. In this exercise you will conduct an experiment to compare the fits on a linear and flexible model fit. You will use the Auto data from the package ISLR and explore the relationship between the response mpg with weight and horsepower.

```
# install.packages("ISLR") #home computer, first time only
library(ISLR)
auto <- Auto[complete.cases(Auto[,c(1,4,5)]), ] # to remove NAs
```

(a) Plot the response (miles per gallon) vs weight and horsepower. What do they tell you about the relationship between mpg and the predictors?

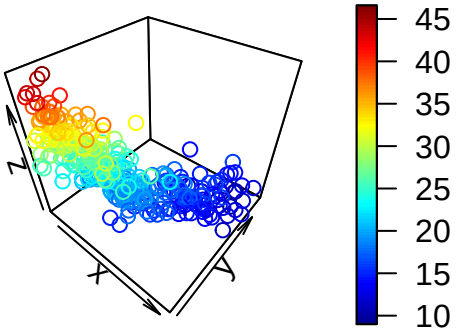
```
auto_ggplot <- auto %>%
  select(mpg, weight, horsepower) %>%
  gather(key = "key", value = "value", -mpg)

auto_ggplot %>%
  ggplot(aes(x = value, y = mpg)) +
  facet_wrap(~key, scales = 'free') +
  geom_point(colour = "tomato") +
  labs(x = "values") +
  theme_bw()
```



mpg goes down as weight or horsepower goes up, and both plots show curvature.

(b) Make a 3d plot of weight, horsepower and mpg (see commands above). What do they tell you about the relationship between mpg and the predictors?

```
library(plot3D)
library(plot3Drgl)

scatter3D(auto$weight, auto$horsepower, auto$mpg)
```




```
# scatter3Drgl(auto$weight, auto$horsepower, auto$mpg)
```

The plot shows that points lie on a surface, not a plane, so a linear fit is not appropriate.

(c) Next, divide the data into a training set and a test set as follows:

```
set.seed(123)
train <- sample(nrow(auto), round(0.8 * nrow(auto)))
auto_train <- auto[train,]
auto_test <- auto[-train,]
```

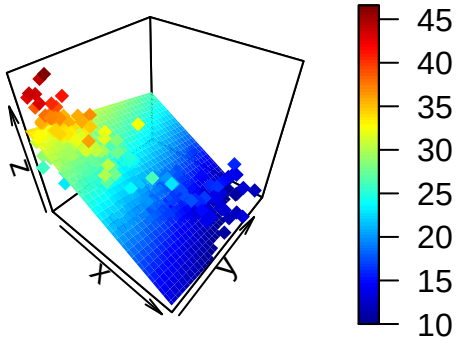
Fit a linear regression model to mpg versus weight and horsepower on AutoTrain. Call the fit f1. Examine summary(f1) and comment on the significance of the predictors.

```
f1 <- lm(mpg ~weight + horsepower, data = auto_train)
summary(f1)
```

```
##
## Call:
## lm(formula = mpg ~ weight + horsepower, data = auto_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4101  -2.7431  -0.4644   2.5079  16.0258
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.1626326  0.8950964   51.57  < 2e-16 ***
## weight      -0.0060579  0.0005553  -10.91  < 2e-16 ***
## horsepower  -0.0431712  0.0120932   -3.57 0.000414 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.339 on 311 degrees of freedom
## Multiple R-squared:  0.7087, Adjusted R-squared:  0.7068
## F-statistic: 378.3 on 2 and 311 DF,  p-value: < 2.2e-16
```

(d) Plot the fitted surface and the data. (See lecture notes for code). Does the linear surface look like a good fit?

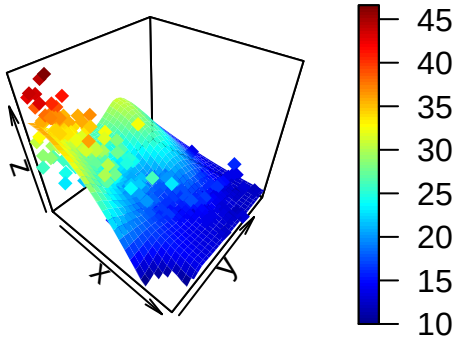
```
wt1 <- seq(1610, 5140, length.out = 30)
hp1 <- seq(45, 230, length.out = 30)
pred <- predict(f1, expand.grid(weight = wt1, horsepower = hp1))
pred <- matrix(pred, 30, 30)

scatter3D(auto_train$weight, auto_train$horsepower, auto_train$mpg,
          pch = 18, surf = list(x = wt1, y = hp1, z = pred))
```



We can see that for high values of z = mpg, points lie far away and mostly above from the fitted planes so the linear fit does not look appropriate.

(e) Use loess to fit a surface to the same data. Call the fit f2. Plot the fitted surface and the data. Does the loess surface look like a good fit?

```
f2 <- loess(mpg ~ weight + horsepower, data = auto_train)
pred <- predict(f2, expand.grid(weight = wt1, horsepower = hp1))
pred <- matrix(pred,30,30)

scatter3D(auto_train$weight,auto_train$horsepower,auto_train$mpg,
          pch = 18, surf = list(x = wt1, y = hp1, z = pred))
```



Now it looks like we captured the pattern of the association, but a smoother surface might be better.

(f) Calculate the MSE for both fits on the training data. What do these numbers tell you?

```
mean(residuals(f1)^2)
```

```
## [1] 18.64557
```

```
mean(residuals(f2)^2)
```

```
## [1] 16.12921
```

The train MSE for the flexible model is smaller.

(g) Calculate the MSE for both fits on the test data. What do these numbers tell you?

```
pred1 <- predict(f1, auto_test)
mean((pred1 - auto_test$mpg)^2)
```

```
## [1] 14.81678
```

```
pred2 <- predict(f2, auto_test)
mean((pred2 - auto_test$mpg)^2)
```

```
## [1] 14.70665
```

The test MSE is about the same for both fits: choose the simpler model (OLS).