Practical Data Science/Analytics
(Feature Engineering)

Write R Scripts or use R to perform any mathematical operations while solving the following problems.

Problem 1: Covariance matrix, Standardized Covariance matrix & Correlation matrix

The dataset Olympics.dat contains the result of the decathlon at the olympic games in Atlanta at 1996 and is available at algorithmica repository. Answer the following questions:

a) Find the covariance matrix for given dataset.
b) Apply z-score transformation to covariance matrix and find the resultant matrix.
c) Find the correlation matrix for given dataset.
d) What are your observations? Can you prove your observation mathematically?
e) Find PCA using covariance matrix and make a biplot for first two principal components.
f) Find PCA using correlation matrix and make a biplot for first two principal components.
g) Which of the two plots in e) and f) seems more advisable? And why?
h) Answer the following questions from the the plot you selected in last part.
   - Which discipline has high correlation with the total number of points (i.e. punkte)?
   - Which variable is displayed badly by the projection?
   - State two disciplines with high positive correlation.
   - State two disciplines with high negative correlation.
   - State two disciplines which are uncorrelated.

Problem 2: Separating US states according to their violence

We want to generate an index that separates US states best according to their violence. USArrests dataset is directly available in R. You can load it using data(USArrests) function. This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Do the following tasks on USArrests dataset:

a) Make a new data set containing only info on murder, assault and rape for each state. Compare the ranges and spread of the three variables.

b) Draw the star plot using the following R function:

```
stars(data, draw.segments = TRUE, key.loc = c(21,1))
```

Practical Data Science/Analytics (Feature Engineering)

Answer the following questions by performing PCA using princomp with cor=F.

c) Compute the total variance of data using trace of covariance matrix.

d) Remember that PC1 tries to explain most of the variance in the data. Which variable do you expect to get the largest loadings for PC1? Look at the loadings of computed PC1. Do they match your expectation?

e) Compute the total variance of data across PC1, PC2 and PC3. Do you see any difference between the original data and projected data variances? If yes, why?

Answer the following questions by performing PCA using princomp with cor=T.

f) Look at the loadings for PC1. Check that PC1 is a (weighted) average of the three violence measures. By looking at the loadings of PC2, describe the violence profile of states that have a high PC2 score.

g) Verify that the PC1 and PC2 scores are uncorrelated. Make a plot of PC2 versus PC1.

h) Answer the following questions by looking at the plot:

i. What is the most violent state according to PC1? And the least violent?
ii. What is the state with the highest PC2 score? Look back at the star plot to see if this state fulfills your answer to f).
iii. Look at some states that are close together in the plot, for example Alaska, Nevada and California. Are their star plots similar as well?
iv. Look at some states that are far apart in the plot made, for example Nevada and main. How do their star plots look?

Problem 3: Understanding relationship among performances in tests

R library provides covariance matrix ability.cov and you can access with data(ability.cov). It contains the covariance matrix of the results of six ability and intelligence tests which were given to 112 individuals. We are interested in the question of whether the performance in and the correlation between the six tests can be explained by two or three variables describing some general concept of intelligence. Do the following tasks:

a) Convert the covariance matrix to correlation matrix using cov2cor() function.

b) Apply PCA and find the principal components. How many principal components are required to cover 95% of variance?

c) Interpret the meaning of first 4 principal components.