```
#############################################################################
# Project:      Credit Scoring Analysis (petit example)
# Description:  Part 3 - Principal Components Analysis
# Data:         CleanCreditScoring.csv
# By:           Gaston Sanchez
# url:          www.gastonsanchez.com
#
# Apply PCA with continuous variables, using Status
# as an illustrative projected variable
#############################################################################


# remember to change your working directory!!! (don't use mine)
# setwd("/Users/gaston/Documents/Gaston/StatsDataMining")

# load package ggplot2
require(ggplot2)

# read cleaned data set
dd = read.csv("CleanCreditScoring.csv", header=TRUE)

# select continuous variables
ddcon = subset(dd, select=c(Seniority, Time, Age, Expenses,
  Income, Assets, Debt, Amount, Price, Finrat, Savings))

# PCA
pca = prcomp(ddcon, scale=TRUE)

# let's check the screeplot
screeplot(pca, type="lines")

# what percentage of total inertia is represented in subspaces?
inertias = 100 * cumsum(pca$sdev^2) / ncol(ddcon)
barplot(inertias, col="gray85", border=NA, names.arg=1:ncol(ddcon),
    main="Percentage of inertia for each dimension")

# how many dimensions do we keep?
nd = 7

# get eigenvalues, eigenvectors (loadings), and projections (ie scores)
egis = pca$sdev[1:nd]^2
U = pca$rotation[,1:nd]
Psi = as.data.frame(pca$x[,1:nd])
Phi = as.data.frame(cor(ddcon, Psi))

# Plot of individuals on first two dimensions
ggplot(data=Psi, aes(x=PC1, y=PC2)) +
geom_hline(yintercept=0, colour="gray65") +
geom_vline(xintercept=0, colour="gray65") +
geom_point(alpha=0.3) +
ggtitle("PCA plot of individuals")

# Plot of variables
# function to create a circle
circle <- function(center=c(0,0), npoints=100)
{
  r = 1
  zz = seq(0, 2*pi, length=npoints)
  xx = center[1] + r * cos(zz)
  yy = center[1] + r * sin(zz)
  return(data.frame(x = xx, y = yy))
}
corcir = circle(c(0,0), npoints = 100)
```

```
# data frame with arrows coordinates
arrows = data.frame(
  x1=rep(0,ncol(ddcon)), y1=rep(0,ncol(ddcon)),
  x2=Phi$PC1, y2=Phi$PC2)

# geom_path will do open circles
ggplot() +
geom_path(data=corcir, aes(x=x, y=y), colour="gray65") +
geom_segment(data=arrows, aes(x=x1, y=y1, xend=x2, yend=y2), colour="gray65") +
geom_text(data=Phi, aes(x=PC1, y=PC2, label=rownames(Phi))) +
geom_hline(yintercept=0, colour="gray65") +
geom_vline(xintercept=0, colour="gray65") +
xlim(-1.1,1.1) + ylim(-1.1,1.1) +
labs(x="pc1 aixs", y="pc2 axis") +
title("Circle of correlations")


# how would you interpret the first two dimensions?
Phi[,1:2]

# Every PC is a linear combination of the standardized original variables
# For instance, PC1 expressed in terms of the observed variables
Z = scale(ddcon)
reg1 = lm(Psi[,1] ~ Z)
print(reg1)
barplot(sort(reg1$coefficients[-1]), border=NA, las=2,
        main="sorted coefficients of linear combination for PC1")

# Let's add information about the credit Status
# and visualize it on the plot formed by PC1-vs-PC2
Psi$Status = dd$Status
# ggplot
ggplot(data=Psi, aes(x=PC1, y=PC2)) +
geom_hline(yintercept=0, colour="gray65") +
geom_vline(xintercept=0, colour="gray65") +
geom_point(aes(colour=Status), alpha=0.3) +
ggtitle=("PCA plot of individuals")

# let's try another plot adding density curves
ggplot(data=Psi, aes(x=PC1, y=PC2, group=Status)) +
geom_hline(yintercept=0, colour="gray65") +
geom_vline(xintercept=0, colour="gray65") +
geom_point(aes(colour=Status), alpha=0.4, size=1.5) +
stat_density2d(aes(colour=Status)) +
ggtitle("PCA plot of individuals\nStatus as illustrative")

# add centers of gravity (cog)
stat1 = tapply(Psi[,1], dd$Status, mean)
stat2 = tapply(Psi[,2], dd$Status, mean)
cog = data.frame(stat1, stat2, Status=c("bad","good"))


ggplot() +
geom_hline(yintercept=0, colour="gray65") +
geom_vline(xintercept=0, colour="gray65") +
geom_point(data=Psi, aes(x=PC1, y=PC2, colour=Status),
    alpha=0.2, size=1.5) +
stat_density2d(data=Psi, aes(x=PC1, y=PC2, colour=Status), alpha=0.4) +
geom_point(data=cog, aes(stat1, stat2, colour=Status), size=5) +
title("PCA plot of individuals\nStatus as illustrative with centers of gravity")
```