Swiss banknote PCA

Audrey Tedore and Zainab Mahmoud

2022-11-29

STEP 1: Determining whether or not to scale the data

```
setwd("C:/Users/Audrey Tedore/OneDrive - University Of Houston/")
banknotes<-read.csv("C:/Users/Audrey Tedore/OneDrive - University Of Houston/banknotes.csv")
```

```
data=colMeans(banknotes)
colMeans=matrix(data[2:7], nrow=6, ncol=1)
rownames(colMeans)=c("Length","Left","Right","Bottom","Top","Diagonal")
(colMeans)
```

```
##              [,1]
## Length   214.8960
## Left     130.1215
## Right    129.9565
## Bottom     9.4175
## Top       10.6505
## Diagonal 140.4835
```

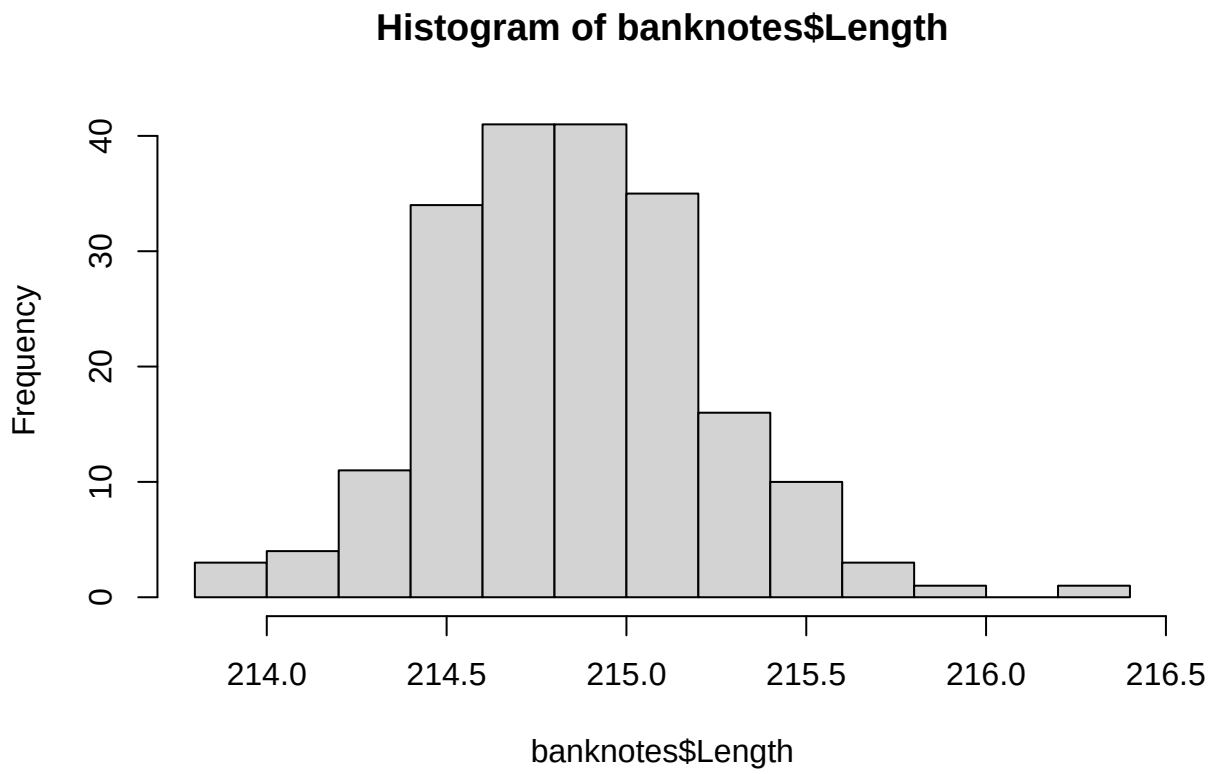
```
(cov.matrix=cov(banknotes[,-1]))
```

```
##               Length        Left       Right     Bottom        Top    Diagonal
## Length     0.14179296  0.03144322  0.02309146 -0.1032462 -0.0185407  0.08430553
## Left       0.03144322  0.13033945  0.10842739  0.2158028  0.1050394 -0.20934196
## Right      0.02309146  0.10842739  0.16327412  0.2841319  0.1299967 -0.24047010
## Bottom    -0.10324623  0.21580276  0.28413191  2.0868781  0.1645389 -1.03699623
## Top       -0.01854070  0.10503945  0.12999673  0.1645389  0.6447234 -0.54961482
## Diagonal   0.08430553 -0.20934196 -0.24047010 -1.0369962 -0.5496148  1.32771633
```

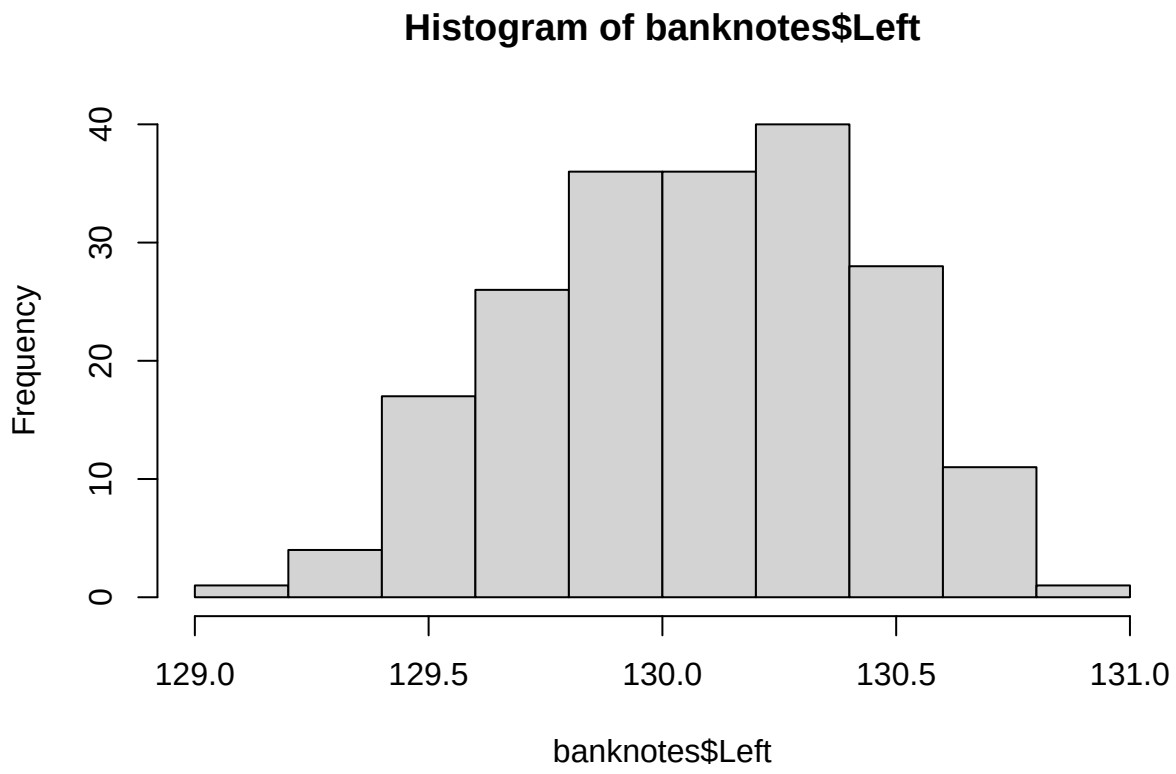
```
(total.variance=sum(diag(cov.matrix)))
```

```
## [1] 4.494724
```

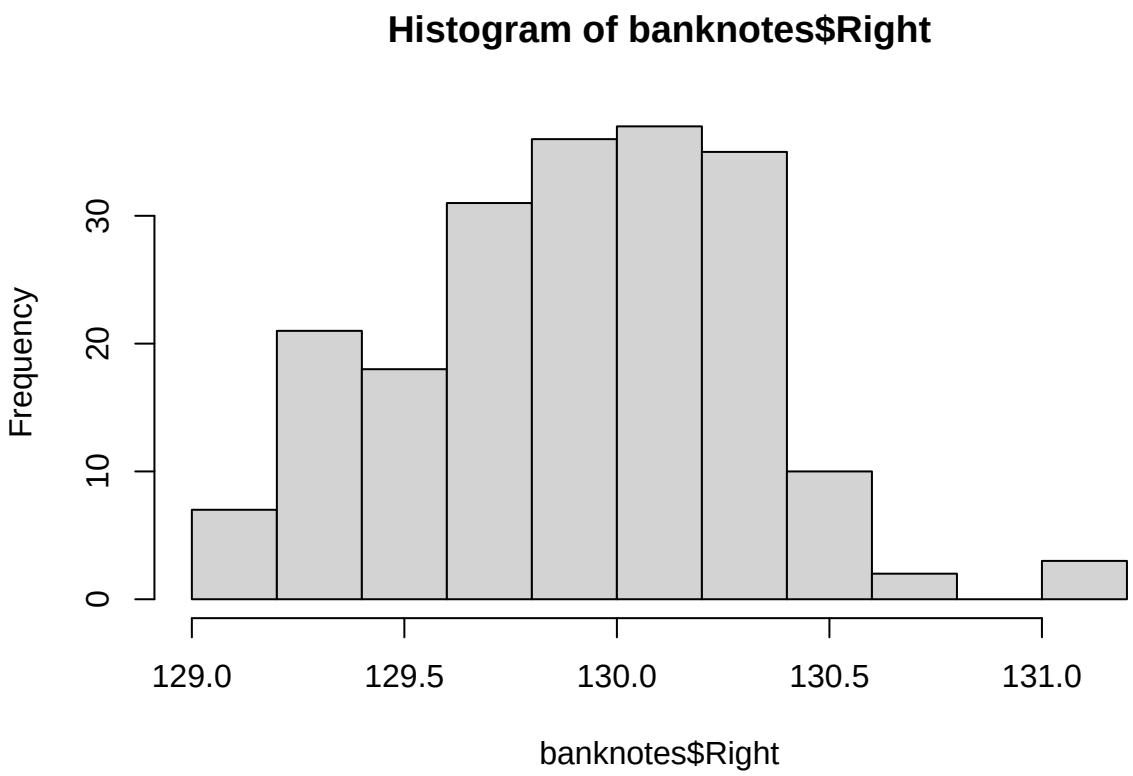
```
hist(banknotes$Length)
```



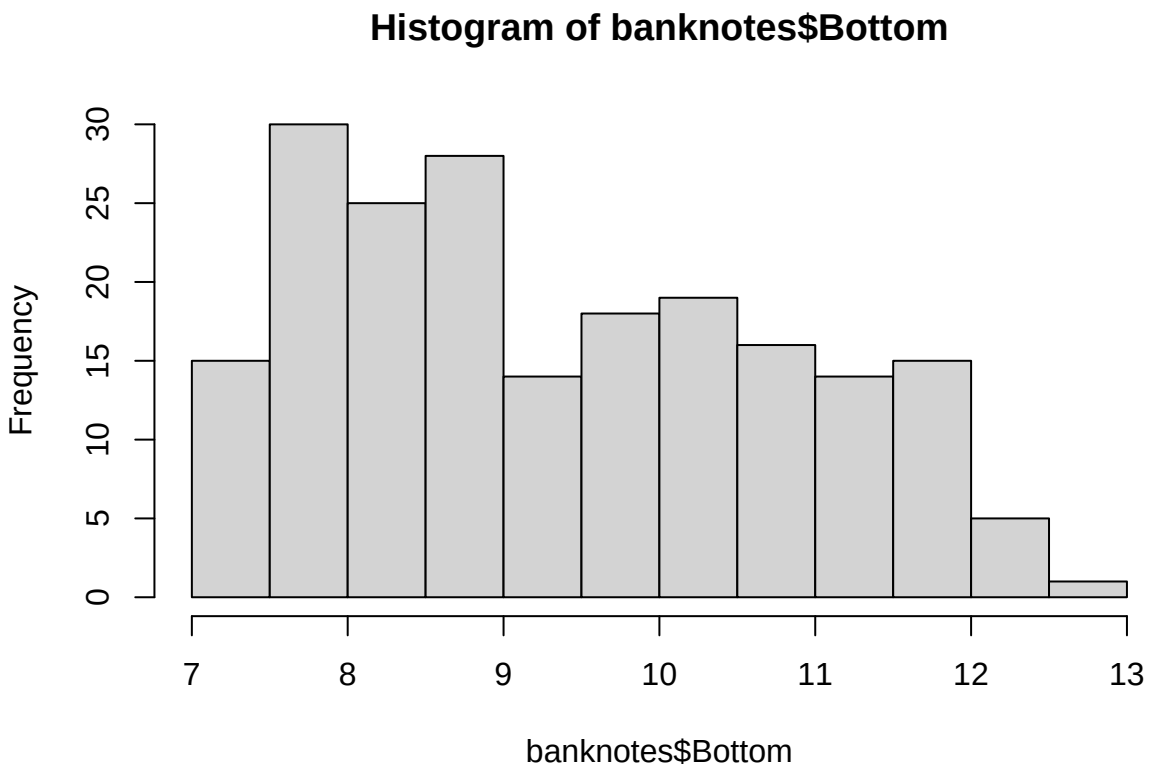

```
hist(banknotes$Left)
```



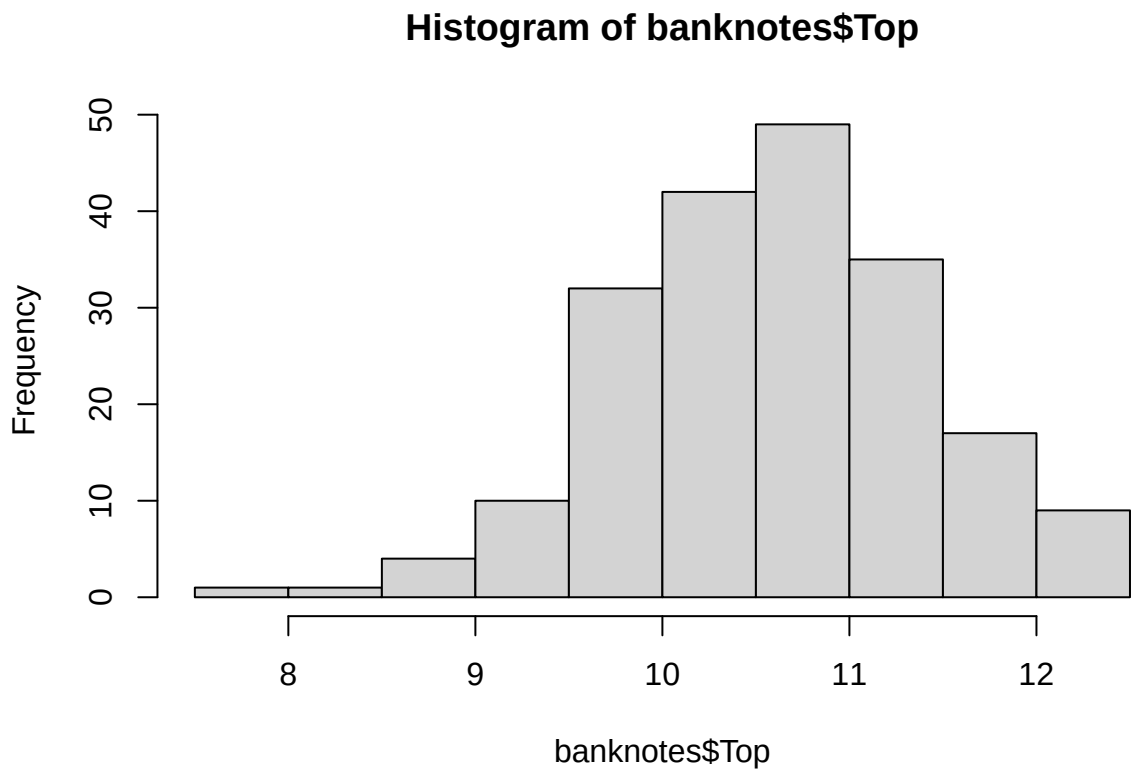

```
hist(banknotes$Right)
```



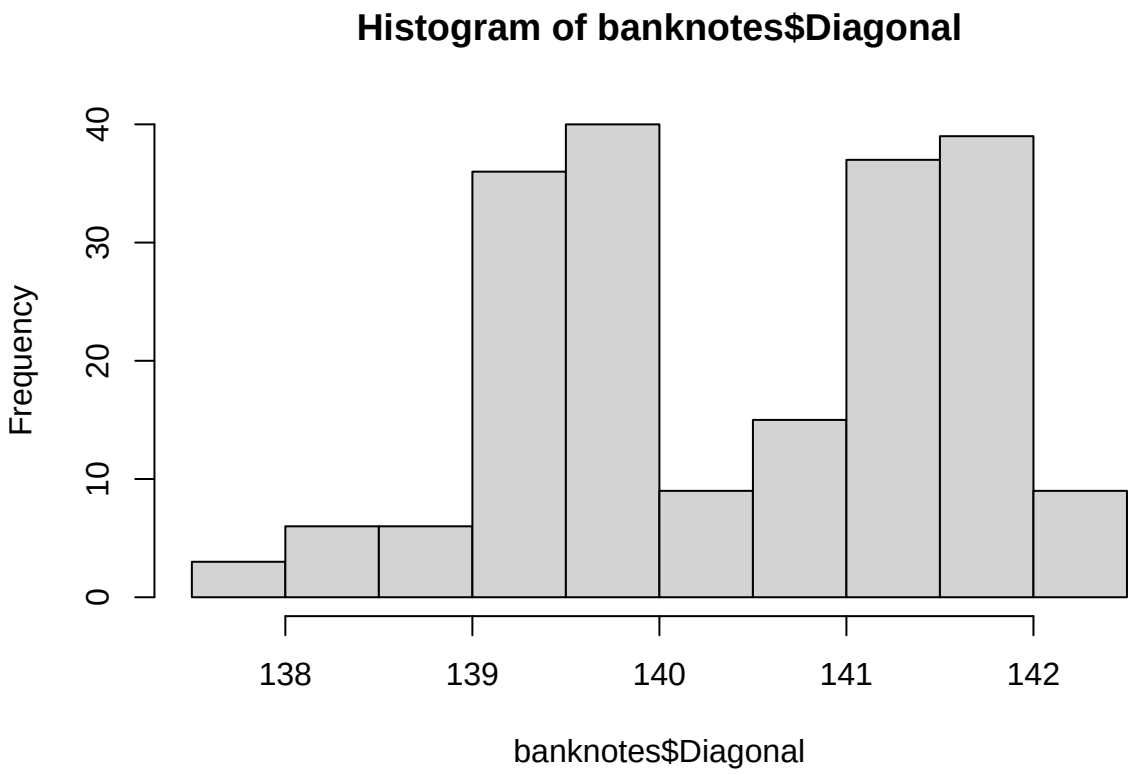

```
hist(banknotes$Bottom)
```




```
hist(banknotes$Top)
```




```
hist(banknotes$Diagonal)
```



STEP 2: Determining how many principal components

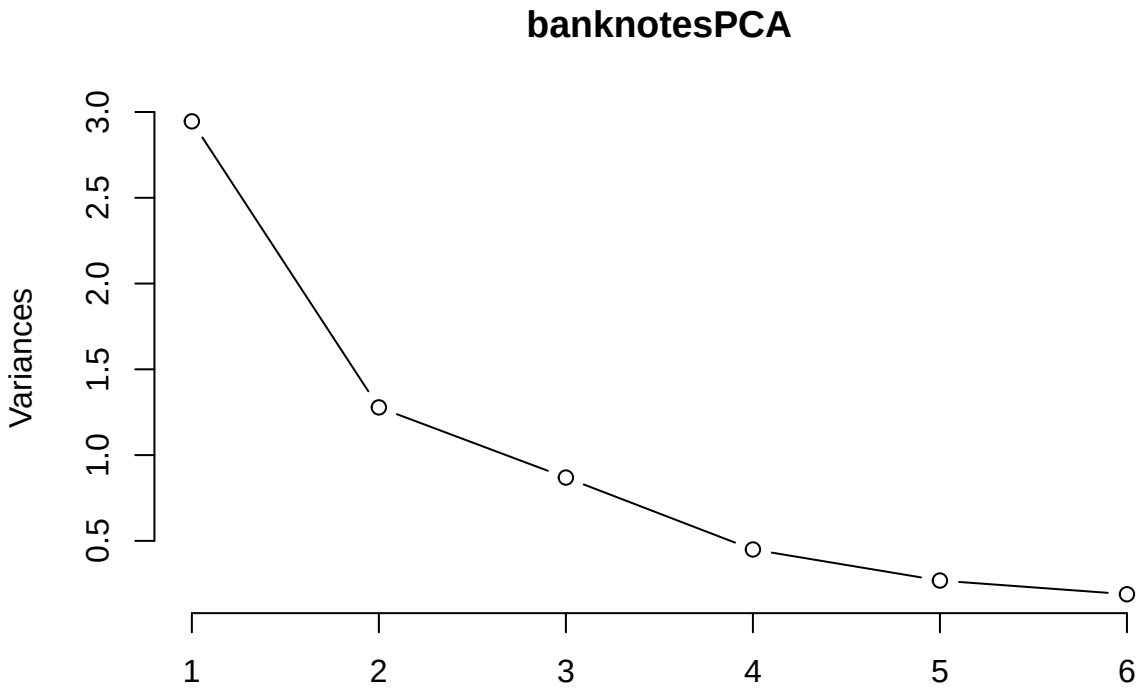
3 principal components can be used instead of the whole data matrix:

The first component PC1 can only account for 49.09% of the variation in bank notes, the first two components can account for 70.39%. The first three components can be used without major loss in accuracy since PC3's cumulative proportion is more than 80% at 84.88%. PC4 has a cumulative proportion of 92.374%, PC5 has a cumulative proportion of 96.852% and PC6 has 100% since all the predictors will be used.

```
banknotesPCA=prcomp(banknotes[,-1],scale.=TRUE)
summary(banknotesPCA)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.7163 1.1305 0.9322 0.67065 0.51834 0.43460
## Proportion of Variance 0.4909 0.2130 0.1448 0.07496 0.04478 0.03148
## Cumulative Proportion  0.4909 0.7039 0.8488 0.92374 0.96852 1.00000
```

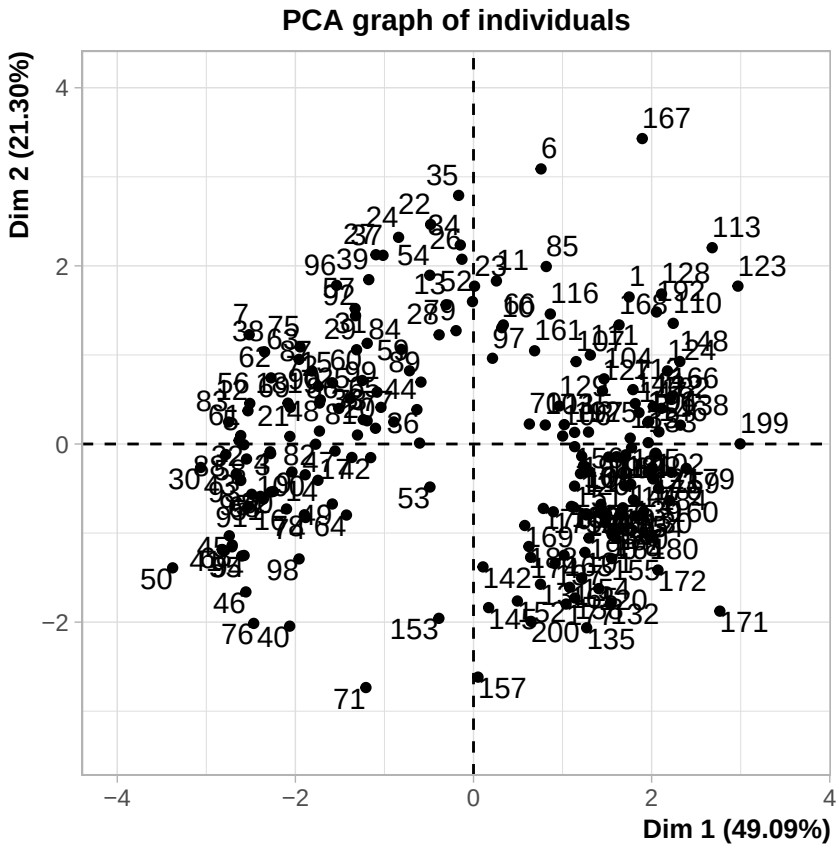
```
screeplot(banknotesPCA,type="lines")
```

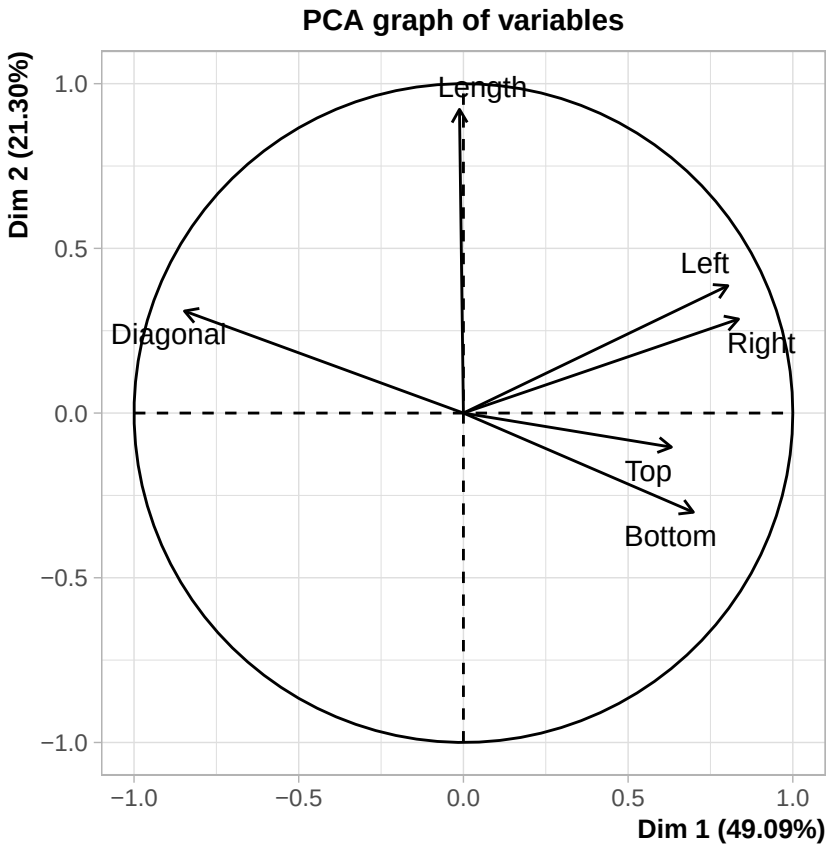


STEP 3: Determining what variables are correlated to the principal components

Left, Right, Bottom,Top,and Diagonal are correlated to PC1. Length is strongly correlated to PC2. Bottom and Top are correlated to PC3.

```
library(FactoMineR)
banknotesPCA2=PCA(banknotes[,-1],scale.unit=TRUE,graph=TRUE)
```


PCA graph of individuals
Dim 2 (21.30%)
Dim 1 (49.09%)
4
2
0
−2
−4
−2
0
2
4




```
banknotesPCA2$var$cor
```

```
##                Dim.1      Dim.2       Dim.3       Dim.4      Dim.5
## Length   -0.01199158  0.9219364 -0.01648225 -0.38536590  0.0304764
## Left      0.80279596  0.3866019  0.09637548  0.26485400 -0.3314768
## Right     0.83526859  0.2854104  0.11510550  0.28856524  0.3183114
## Bottom    0.69810421 -0.3009779  0.54398559 -0.27072283  0.1116898
## Top       0.63139786 -0.1034278 -0.73418921 -0.07392333  0.1139569
## Diagonal -0.84690418  0.3096965  0.10615680  0.26284740  0.1763188
```

STEP 4: Regression using principal components

```
set.seed(101)
sample = sample.int(n = nrow(banknotes),size = round(.80*nrow(banknotes)),
                    replace = FALSE)
train = banknotes[sample,]
test = banknotes[-sample,]
```

```
pca<-banknotesPCA2$ind$coord[sample,1:3]
banknotes.glm = glm(conterfeit ~ pca, family = "binomial", data = train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(banknotes.glm)
```

```
##
## Call:
## glm(formula = conterfeit ~ pca, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9526  -0.0005   0.0000   0.0076   1.4386
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.203      1.932  -1.658   0.0973 .
## pcaDim.1       8.495      4.360   1.948   0.0514 .
## pcaDim.2      -3.108      1.532  -2.029   0.0424 *
## pcaDim.3      -3.346      2.224  -1.505   0.1324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 221.4069  on 159  degrees of freedom
## Residual deviance:   8.2474  on 156  degrees of freedom
## AIC: 16.247
##
## Number of Fisher Scoring iterations: 11
```

```
banknotes.glm.probs=predict(banknotes.glm,type="response")
banknotes.glm.preds=ifelse(banknotes.glm.probs<0.5,"No","Yes")
(conf.mat=table(banknotes.glm.preds,train$conterfeit))
```

```
##
## banknotes.glm.preds  0  1
##                  No  83  1
##                  Yes  1 75
```

```
(train_error_rate=(conf.mat[1,2]+conf.mat[2,1])/sum(conf.mat))
```

```
## [1] 0.0125
```

```
pca<-banknotesPCA2$ind$coord[-sample,1:3]
banknotes.glm.probs.1=predict(banknotes.glm,type="response",newdata=test)
banknotes.glm.preds.1=ifelse(banknotes.glm.probs.1<0.5,"No","Yes")
(conf.mat=table(banknotes.glm.preds.1,test$conterfeit))
```

```
##
## banknotes.glm.preds.1  0  1
##                    No  15  1
##                    Yes  1 23
```

```
(test_error_rate=(conf.mat[1,2]+conf.mat[2,1])/sum(conf.mat))
```

```
## [1] 0.05
```

The test error rate is low and close to the train error rate, indicating that this is a good model.

```
step(banknotes.glm)
```

```
## Start:  AIC=16.25
## conterfeit ~ pca
##
##        Df Deviance     AIC
## <none>       8.247  16.247
## - pca   3  221.407 223.407

##
## Call:  glm(formula = conterfeit ~ pca, family = "binomial", data = train)
##
## Coefficients:
## (Intercept)     pcaDim.1     pcaDim.2     pcaDim.3
##      -3.203        8.495       -3.108       -3.346
##
## Degrees of Freedom: 159 Total (i.e. Null);  156 Residual
## Null Deviance:       221.4
## Residual Deviance: 8.247     AIC: 16.25
```

All 3 principal components are needed in the model.