Week 8 Dimension Reduction Analysis

Nitin

October 24, 2020

Perform PCA analysis

Read Data

```
library(tidyverse)
library(readxl)
Data <- read_excel("data.xlsx")
df <- select(Data, -c(3,5,7,9,11,13,15,17,19,21,23,25))
names(df)<-str_replace_all(names(df), c(" " = "." , "," = "" ))
df = na.omit(df)
df$Months <- scale(df$Month)
data_df <- select(df, -c(1))
```

Establish the optimal number of components: visualize the scree plot and explain your decision

Run PCA.

Step 1: Normalize Data

Remove Dependents variable (Y variable) Plywood Price from data sate and scale remaining dataset observation.

```
data_df <- select (data_df, -c(8))
```

```
data_scaled <- as.data.frame(scale(data_df, center = T, scale = T))
```

Step 2 Calculate Covariance Matrix

```
dimn <- dim(data_scaled)
dimn
```

```
## [1] 327  12
```

```
cov_mat = cov(data_scaled)
cov_mat
```

```
##                      Coarse.wool.Price Copra.Price Cotton.Price
## Coarse.wool.Price            1.0000000  0.80064586   0.52832773
## Copra.Price                  0.8006459  1.00000000   0.63699578
## Cotton.Price                 0.5283277  0.63699578   1.00000000
## Fine.wool.Price              0.8921317  0.81696504   0.69700136
## Hard.log.Price               0.5441895  0.52408975   0.47223903
## Hard.sawnwood.Price          0.6638902  0.68409292   0.48576156
## Hide.Price                   0.1710559  0.07429592   0.27863871
## Rubber.Price                 0.7847535  0.76013957   0.74131398
## Softlog.Price               -0.2642042 -0.14439174  -0.07600436
## Soft.sawnwood.Price          0.2058270  0.24368688  -0.11298531
```

```
## Wood.pulp.Price              0.7221412  0.71151770   0.56220087
## Months                       0.8476030  0.72636443   0.22810540
##                    Fine.wool.Price Hard.log.Price Hard.sawnwood.Price
## Coarse.wool.Price       0.89213169     0.54418946           0.6638902
## Copra.Price             0.81696504     0.52408975           0.6840929
## Cotton.Price            0.69700136     0.47223903           0.4857616
## Fine.wool.Price         1.00000000     0.53100171           0.6172997
## Hard.log.Price          0.53100171     1.00000000           0.8140339
## Hard.sawnwood.Price     0.61729969     0.81403388           1.0000000
## Hide.Price              0.22737210     0.09049411           0.1091120
## Rubber.Price            0.82156132     0.61712752           0.7330926
## Softlog.Price          -0.29993968     0.24307907           0.1678083
## Soft.sawnwood.Price     0.06086458     0.18940773           0.3910092
## Wood.pulp.Price         0.76845207     0.35312190           0.5817564
## Months                  0.67845368     0.34147786           0.5665084
##                     Hide.Price Rubber.Price Softlog.Price
## Coarse.wool.Price    0.17105591    0.7847535   -0.26420424
## Copra.Price          0.07429592    0.7601396   -0.14439174
## Cotton.Price         0.27863871    0.7413140   -0.07600436
## Fine.wool.Price      0.22737210    0.8215613   -0.29993968
## Hard.log.Price       0.09049411    0.6171275    0.24307907
## Hard.sawnwood.Price  0.10911199    0.7330926    0.16780832
## Hide.Price           1.00000000    0.0135674    0.24257261
## Rubber.Price         0.01356740    1.0000000   -0.20196995
## Softlog.Price        0.24257261   -0.2019700    1.00000000
## Soft.sawnwood.Price -0.11374738    0.1790261    0.41834661
## Wood.pulp.Price      0.22354226    0.6664118   -0.21193664
## Months              -0.06568999    0.6180625   -0.21725822
##                    Soft.sawnwood.Price Wood.pulp.Price      Months
## Coarse.wool.Price           0.20582696       0.7221412  0.84760298
## Copra.Price                 0.24368688       0.7115177  0.72636443
## Cotton.Price               -0.11298531       0.5622009  0.22810540
## Fine.wool.Price             0.06086458       0.7684521  0.67845368
## Hard.log.Price              0.18940773       0.3531219  0.34147786
## Hard.sawnwood.Price         0.39100915       0.5817564  0.56650840
## Hide.Price                 -0.11374738       0.2235423 -0.06568999
## Rubber.Price                0.17902610       0.6664118  0.61806252
## Softlog.Price               0.41834661      -0.2119366 -0.21725822
## Soft.sawnwood.Price         1.00000000       0.1527109  0.44047048
## Wood.pulp.Price             0.15271093       1.0000000  0.62860719
## Months                      0.44047048       0.6286072  1.00000000
```

Step 3 Calculate Eigenvectors from Covariance Matrix

‘

```
eig_vecs <- eigen(cov_mat)$vectors
eig_vecs
```

```
##             [,1]        [,2]          [,3]        [,4]        [,5]
##  [1,] -0.36633676  0.08564520  0.0998436212  0.16081693 -0.27935782
##  [2,] -0.35738047  0.02416461  0.0643998935  0.02300213  0.18377817
##  [3,] -0.28037734  0.14240052 -0.3974683891 -0.21397787  0.55983439
##  [4,] -0.36652749  0.18022491 -0.0621426204  0.07454600 -0.04059205
##  [5,] -0.26914624 -0.29851517 -0.1882440267 -0.45632935 -0.45731337
```

```
##  [6,] -0.32585381 -0.29893075 -0.0368804622 -0.20022358 -0.21555113
##  [7,] -0.06424554 -0.04224227 -0.6277365456  0.62178995 -0.23219045
##  [8,] -0.35918046  0.05399159 -0.0008806195 -0.24095250  0.19537511
##  [9,]  0.05805273 -0.66109532 -0.2780870254  0.05328196  0.17842564
## [10,] -0.10329472 -0.55142394  0.3821602663  0.26724548  0.33571696
## [11,] -0.32312200  0.12129929 -0.0140708355  0.27645086  0.20933748
## [12,] -0.30808518 -0.01796505  0.4118425669  0.28031088 -0.19255419
##              [,6]         [,7]        [,8]        [,9]       [,10]
##  [1,] -0.25419015  0.008410784 -0.26606603 -0.01544231 -0.13523268
##  [2,] -0.15282628  0.575891373  0.60281484 -0.05142321  0.16342445
##  [3,] -0.21096628 -0.043524129 -0.03685575  0.02844536 -0.51873943
##  [4,] -0.15694766  0.002348110 -0.17699474 -0.45029602  0.52049237
##  [5,]  0.04279097  0.009601146 -0.05124539 -0.42301519 -0.27262796
##  [6,]  0.33254001 -0.173199627  0.37836002  0.42253134  0.08261078
##  [7,] -0.11879979 -0.228845004  0.21377791  0.07291245 -0.02128614
##  [8,] -0.12596098 -0.412867640 -0.20475807  0.36661701  0.40162257
##  [9,] -0.05373448  0.433149069 -0.42376271  0.16267336  0.21305559
## [10,] -0.11057877 -0.424528191  0.17073708 -0.35076609 -0.09592069
## [11,]  0.80811055  0.087486791 -0.21830362 -0.13564557 -0.07922918
## [12,] -0.17983591  0.194839076 -0.20819688  0.35839015 -0.33350653
##              [,11]        [,12]
##  [1,]  0.171342258  0.749153946
##  [2,] -0.265537336  0.122433511
##  [3,]  0.245009794 -0.078990063
##  [4,]  0.413367958 -0.348701810
##  [5,] -0.332829763 -0.130569281
##  [6,]  0.497444851  0.001351631
##  [7,] -0.185377892 -0.084741254
##  [8,] -0.497208359  0.025834004
##  [9,]  0.054250582  0.051932728
## [10,] -0.007636047  0.034502253
## [11,] -0.153339891  0.062371793
## [12,] -0.053909023 -0.513098163
```

Step 4 Select with the largest Eigenvalues

```r
eig_vals <- eigen(cov_mat)$values
eig_vals
```

```
##  [1] 6.31747004 1.73720844 1.42249013 0.91322617 0.50507675 0.31694078
##  [7] 0.26099451 0.18275855 0.14102112 0.08885781 0.06924078 0.04471491
```

Run prcomp function and print useful information.

```r
df.pca <- prcomp(data_scaled)
print(summary(df.pca))
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.5135 1.3180 1.1927 0.9556 0.71069 0.56297 0.51088
## Proportion of Variance 0.5265 0.1448 0.1185 0.0761 0.04209 0.02641 0.02175
## Cumulative Proportion  0.5265 0.6712 0.7898 0.8659 0.90796 0.93437 0.95612
##                            PC8     PC9   PC10    PC11    PC12
## Standard deviation     0.42750 0.37553 0.2981 0.26314 0.21146
```

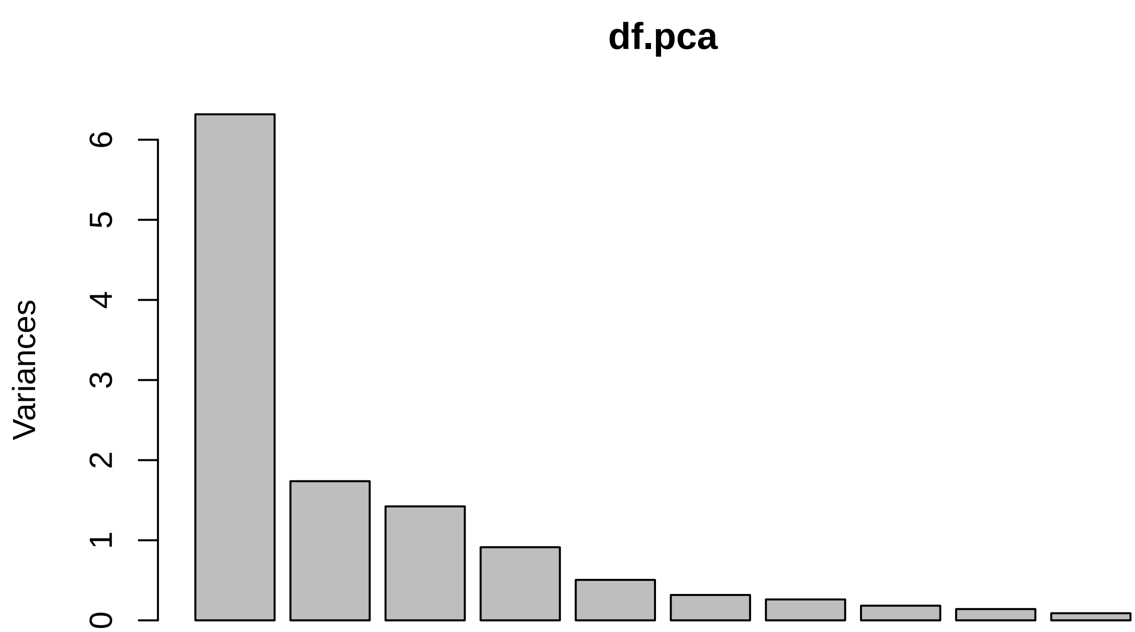
```
## Proportion of Variance 0.01523 0.01175 0.0074 0.00577 0.00373
## Cumulative Proportion  0.97135 0.98310 0.9905 0.99627 1.00000
```

It could be observed that cumulative variance proportion of 86.59 percent is explained by first four principle components PC1, PC2, PC3, PC4.

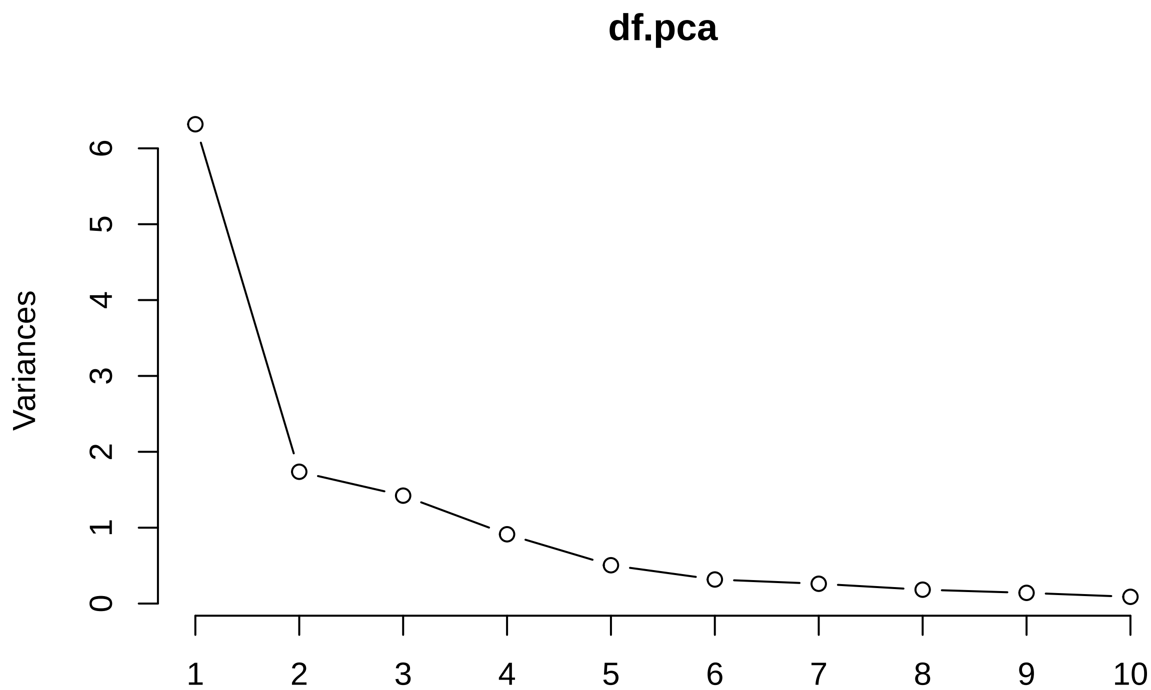
Determine optimum number of PCA.

Optimum number of PCA is four (4) PC1, PC2, PC3, PC4 as cumulative variance proportion of 86.59 percent is explained by first four principle components.

```r
plot(df.pca)
```

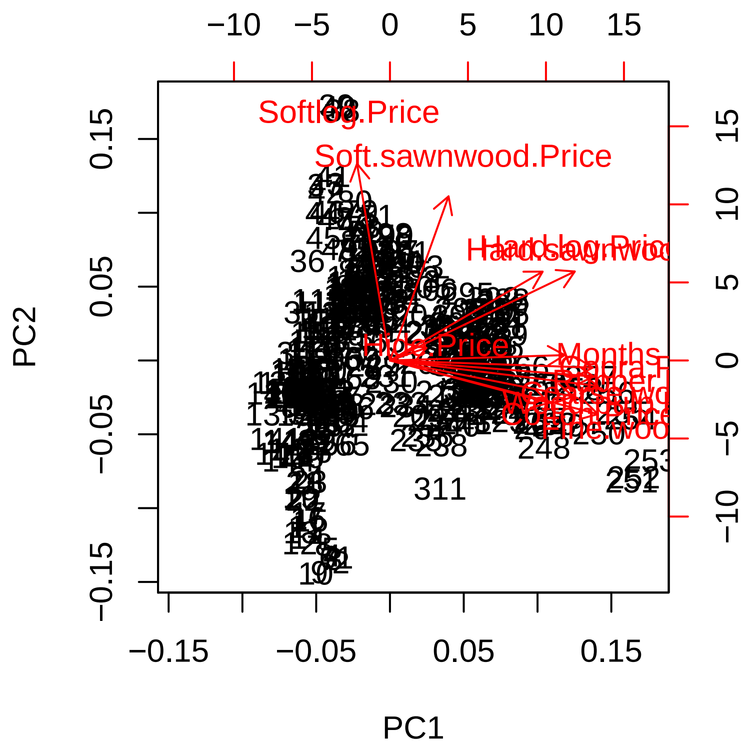



```r
plot(df.pca, type = "l")
```



From two plots above it could be concluded that optimum number of principle components is four PC1, PC2, PC3, PC4.

```
biplot(df.pca)
```



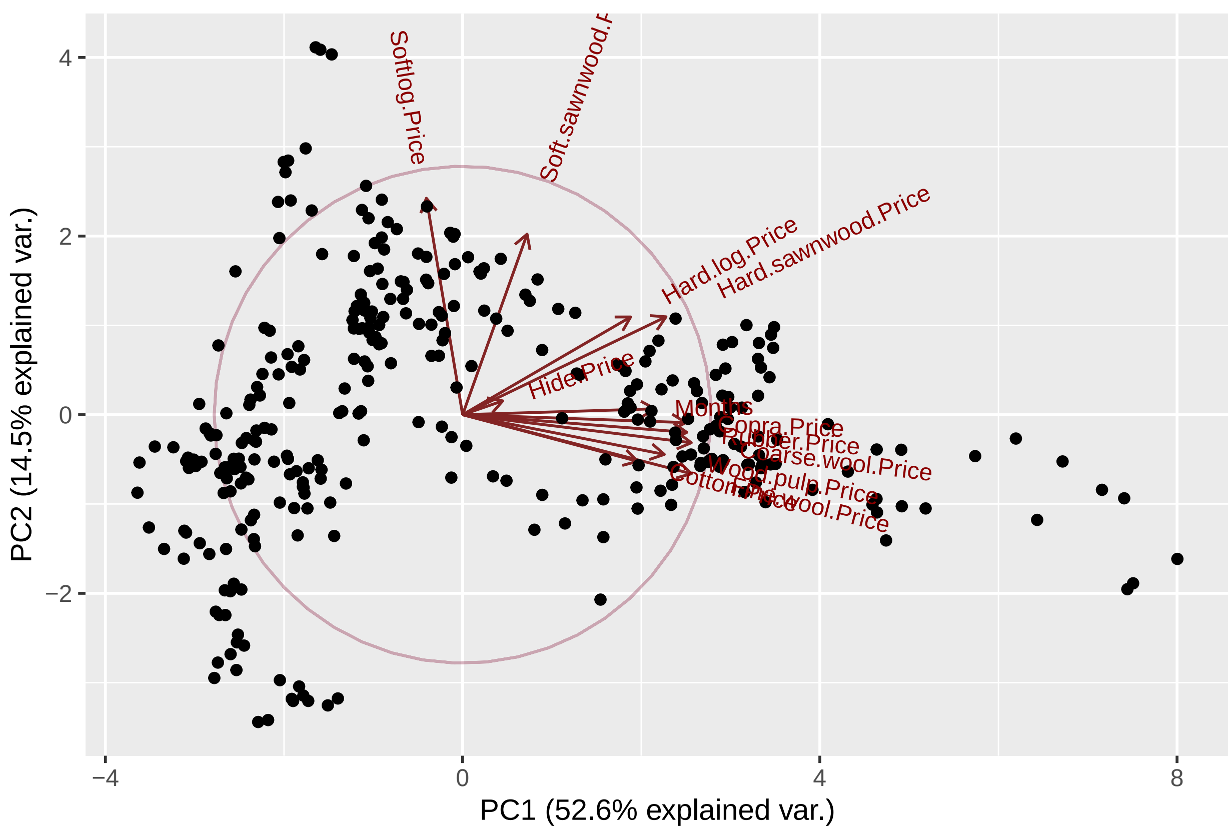

```r
df.pca$rotation
```

```
##                             PC1         PC2           PC3         PC4
## Coarse.wool.Price    0.36633676 -0.08564520  0.0998436212 -0.16081693
## Copra.Price          0.35738047 -0.02416461  0.0643998935 -0.02300213
## Cotton.Price         0.28037734 -0.14240052 -0.3974683891  0.21397787
## Fine.wool.Price      0.36652749 -0.18022491 -0.0621426204 -0.07454600
## Hard.log.Price       0.26914624  0.29851517 -0.1882440267  0.45632935
## Hard.sawnwood.Price  0.32585381  0.29893075 -0.0368804622  0.20022358
## Hide.Price           0.06424554  0.04224227 -0.6277365456 -0.62178995
## Rubber.Price         0.35918046 -0.05399159 -0.0008806195  0.24095250
## Softlog.Price       -0.05805273  0.66109532 -0.2780870254 -0.05328196
## Soft.sawnwood.Price  0.10329472  0.55142394  0.3821602663 -0.26724548
## Wood.pulp.Price      0.32312200 -0.12129929 -0.0140708355 -0.27645086
## Months               0.30808518  0.01796505  0.4118425669 -0.28031088
##                             PC5         PC6          PC7         PC8
## Coarse.wool.Price    0.27935782 -0.25419015  0.008410784 -0.26606603
## Copra.Price         -0.18377817 -0.15282628  0.575891373  0.60281484
## Cotton.Price        -0.55983439 -0.21096628 -0.043524129 -0.03685575
## Fine.wool.Price      0.04059205 -0.15694766  0.002348110 -0.17699474
## Hard.log.Price       0.45731337  0.04279097  0.009601146 -0.05124539
## Hard.sawnwood.Price  0.21555113  0.33254001 -0.173199627  0.37836002
## Hide.Price           0.23219045 -0.11879979 -0.228845004  0.21377791
## Rubber.Price        -0.19537511 -0.12596098 -0.412867640 -0.20475807
## Softlog.Price       -0.17842564 -0.05373448  0.433149069 -0.42376271
## Soft.sawnwood.Price -0.33571696 -0.11057877 -0.424528191  0.17073708
```

```
## Wood.pulp.Price      -0.20933748  0.80811055  0.087486791 -0.21830362
## Months                0.19255419 -0.17983591  0.194839076 -0.20819688
##                             PC9        PC10         PC11         PC12
## Coarse.wool.Price     0.01544231  0.13523268  0.171342258 -0.749153946
## Copra.Price           0.05142321 -0.16342445 -0.265537336 -0.122433511
## Cotton.Price         -0.02844536  0.51873943  0.245009794  0.078990063
## Fine.wool.Price       0.45029602 -0.52049237  0.413367958  0.348701810
## Hard.log.Price        0.42301519  0.27262796 -0.332829763  0.130569281
## Hard.sawnwood.Price -0.42253134 -0.08261078  0.497444851 -0.001351631
## Hide.Price           -0.07291245  0.02128614 -0.185377892  0.084741254
## Rubber.Price         -0.36661701 -0.40162257 -0.497208359 -0.025834004
## Softlog.Price        -0.16267336 -0.21305559  0.054250582 -0.051932728
## Soft.sawnwood.Price   0.35076609  0.09592069 -0.007636047 -0.034502253
## Wood.pulp.Price       0.13564557  0.07922918 -0.153339891 -0.062371793
## Months               -0.35839015  0.33350653 -0.053909023  0.513098163
```

Visualize PCA1 and PCA2 and describe which variables contribute to the PCA

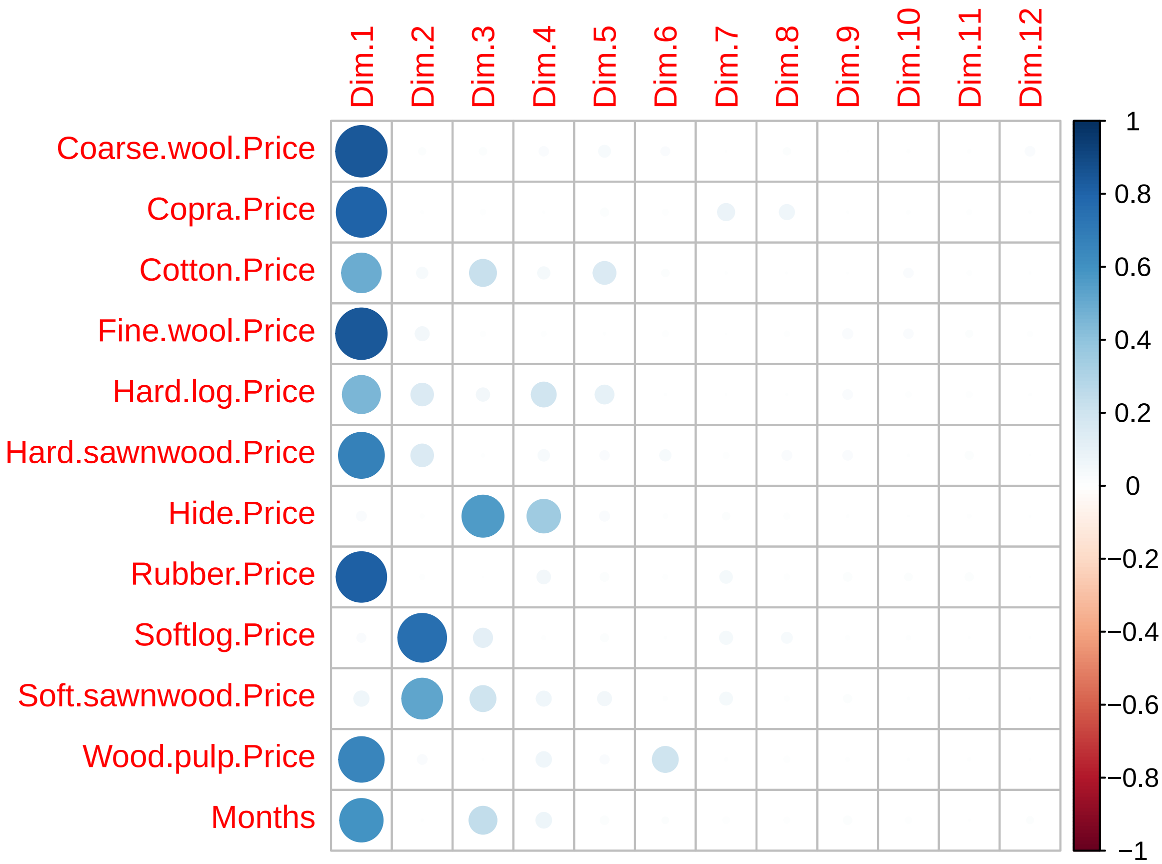
```r
library (ggbiplot)
g <- ggbiplot(df.pca, obs.scale = 1, var.scale = 1, ellipse = TRUE, circle = TRUE) + scale_color_discre
print (g)
```



It could be observed that first two components PC1 and PC2 explains 67.1 percent of variance.

```r
library(factoextra)
library(corrplot)
var <- get_pca_var(df.pca)
```

```r
corrplot(var$cos2, iscorr=FALSE)
```



Reflect how you could use the reduced dimensionality in your final paper

From figure above it could be observed that variable Hide Price does not contribute to principle component 1 (PC1) and principle component 2 (PC2) so it should be eliminated as a variable for analysis in final paper.