ashishsa_hw3_p3

We perform PCA on 200 bank notes which contains 100 geniuine and 100 counterfeit bank notes.

```r
rm(list = ls())
library(pca3d)
library(purrr)
library(kohonen)
```

```
FALSE
FALSE Attaching package: 'kohonen'

FALSE The following object is masked from 'package:purrr':
FALSE
FALSE     map
```

```r
library(ggplot2)
library(dplyr)
```

```
FALSE
FALSE Attaching package: 'dplyr'

FALSE The following objects are masked from 'package:stats':
FALSE
FALSE     filter, lag

FALSE The following objects are masked from 'package:base':
FALSE
FALSE     intersect, setdiff, setequal, union
```

```r
library(ggfortify)
library(factoextra)
```

```
FALSE Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(gridExtra)
```

```
FALSE
FALSE Attaching package: 'gridExtra'

FALSE The following object is masked from 'package:dplyr':
FALSE
FALSE     combine
```

```r
library(cowplot)
```

```
FALSE
FALSE ********************************************************

FALSE Note: As of version 1.0.0, cowplot does not change the

FALSE   default ggplot2 theme anymore. To recover the previous

FALSE   behavior, execute:
FALSE   theme_set(theme_cowplot())
FALSE ********************************************************
```

```r
library(cluster)
library(lfda)
load("SwissBankNotes.RData")
df1 <- SwissBankNotes
df1 <- as.data.frame(df1)
```

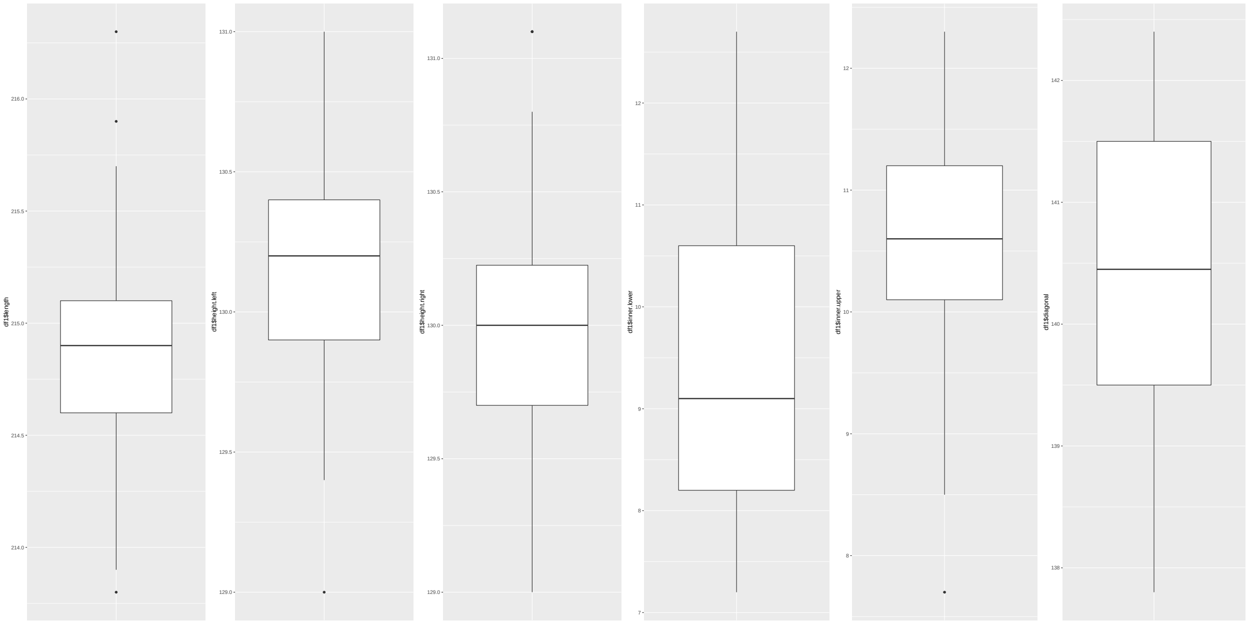
We know that the first 100 Bank Notes are Real and the Second 100 Bank Notes are Counterfeit. We now add one more column to the original dataframe which states that first 100 bank notes are True and second 100 bank notes are False.

```r
status <- c(rep('TRUE',len=100),rep('FALSE',len=100))
df2 <- cbind(df1,status)
```

We now convert the last column into factors.

```r
df2$status <- as.factor(as.character(df2$status))
df2$status <- df2$status %>% map_if(is.factor, as.numeric)
df2$status <- as.numeric(df2$status)
```

```r
plot1 <- ggplot(df1,aes(x=factor(0),df1$length))+geom_boxplot()+
   theme(axis.title.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank())
plot2 <- ggplot(df1,aes(x=factor(0),df1$height.left))+geom_boxplot()+
   theme(axis.title.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank())
plot3 <- ggplot(df1,aes(x=factor(0),df1$height.right))+geom_boxplot()+theme(axis.title.x=element_blank(
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank())
plot4 <- ggplot(df1,aes(x=factor(0),df1$inner.lower))+geom_boxplot()+
   theme(axis.title.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank())
plot5 <- ggplot(df1,aes(x=factor(0),df1$inner.upper))+geom_boxplot()+
   theme(axis.title.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank())
plot6 <- ggplot(df1,aes(x=factor(0),df1$diagonal))+geom_boxplot()+
   theme(axis.title.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank())
plot_grid(plot1, plot2, plot3,plot4,plot5,plot6,ncol=6)
```



We first begin by performing Boxplot Analysis. We first analyze length and observe that it has outliers at 213,215.7 and 217. We analyze height.left and observe that it has outlier at 129. We analyze height.right and observe that it has outliers at 131.5. We analyze inner.lower and observe that it has no outliers. We analyze inner.upper and observe that it has outliers at 8. We analyze diagonal and observe that it has no outliers.

From the Boxplot Analysis we conclude that the Variables need to be scaled as inner upper and inner lower are of different dimension in comparison to the other variables.

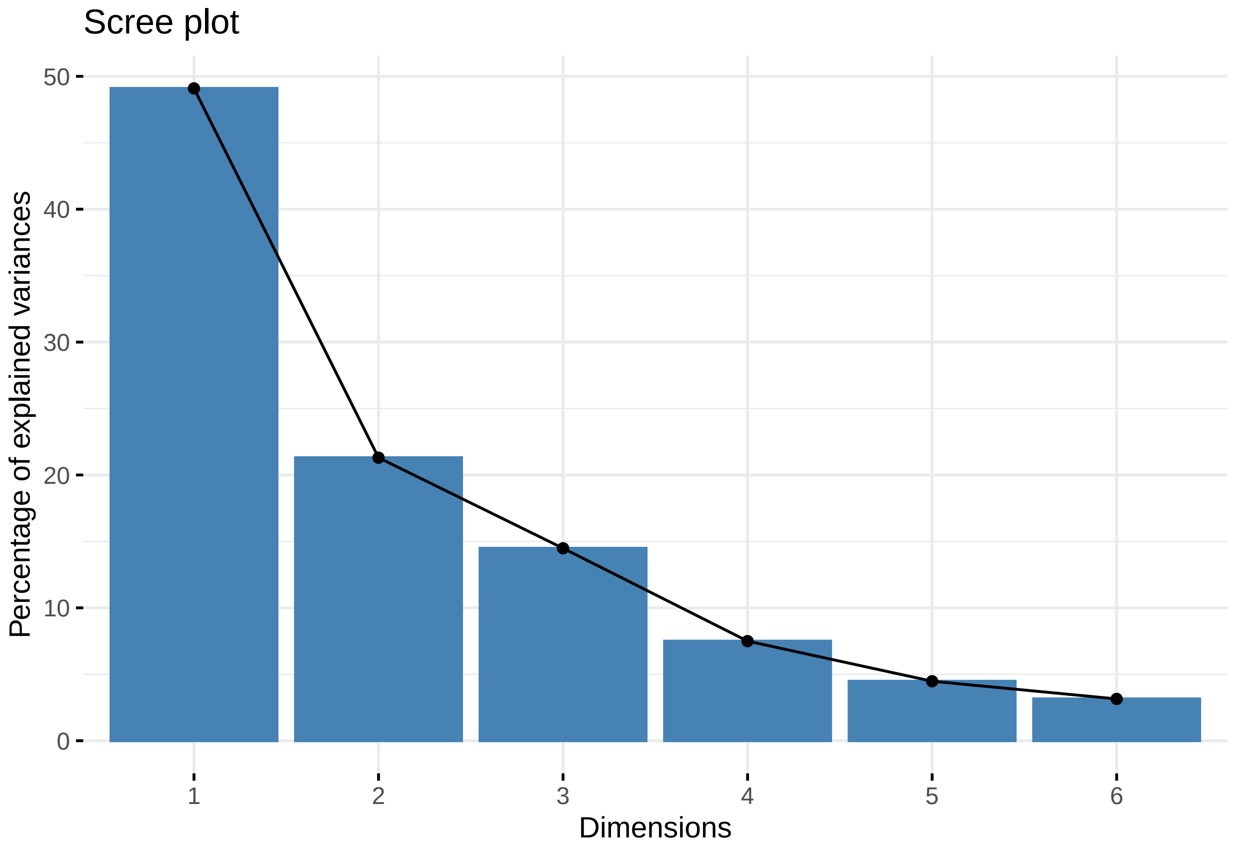
```
m <- apply(df1,2,mean)
s <- apply(df1,2,sd)
z <- scale(df1,m,s)
```

We first Analyze all the variables that is fake and real notes together. We now apply Principal Component Analysis on dataset in R.

```
z_pca <- prcomp(z,center = TRUE,scale. = FALSE)
```

We now plot the Eigen Values for the entire dataset.

```
fviz_eig(z_pca)
```



We summarize the Results for the PCA and observe here that PC1 only accounts for 49Percent of the Variation. While PC2 only accounts for 21Percent of the Variation. Since we have already used scaled data for plotting the dimension of the variables has no significant impact. This indicates that the data that is skewed by is because of false data.

```
summary(z_pca)
```

```
## 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
```

We now check the Eigen Values for each of the 6 variables.

```
EigenValues <- eigen(cov(scale(df1)))
EigenValues$values
```

```
## [1] 2.9455582 1.2780838 0.8690326 0.4497687 0.2686769 0.1888799
```

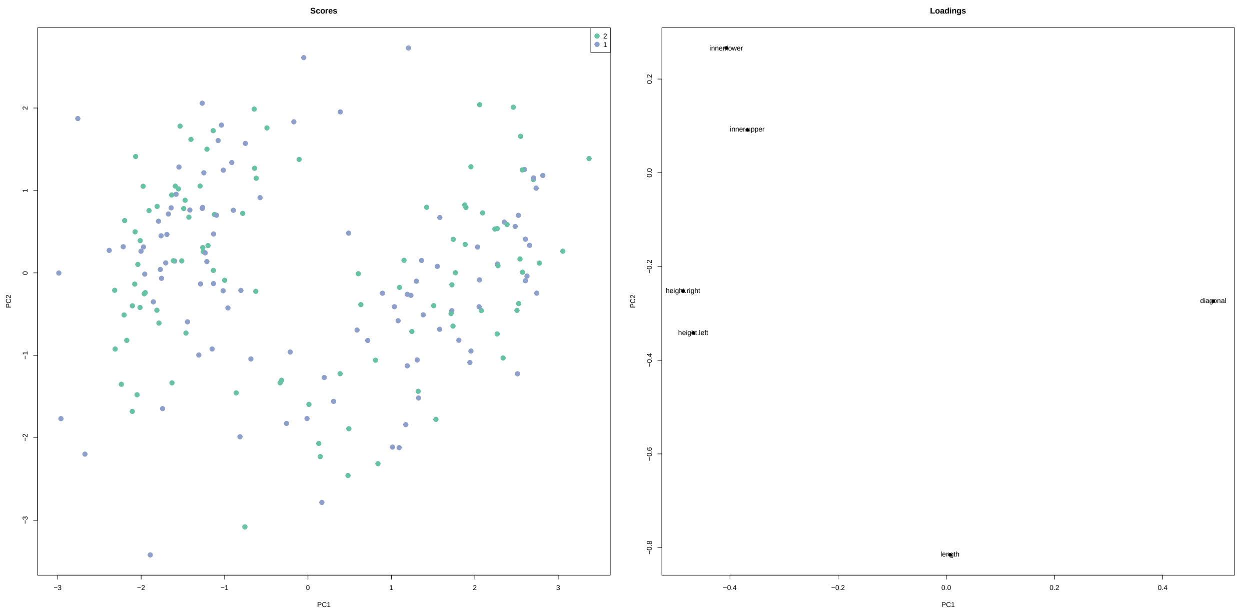
The Eigen Values obtained above are checked using the following method. We calculate the Standard Deviation of the Principle Components and find the square of it. The Values must be same indicating that they are the intended correct value.

```
(z_pca$sdev)^2
```

```
## [1] 2.9455582 1.2780838 0.8690326 0.4497687 0.2686769 0.1888799
```

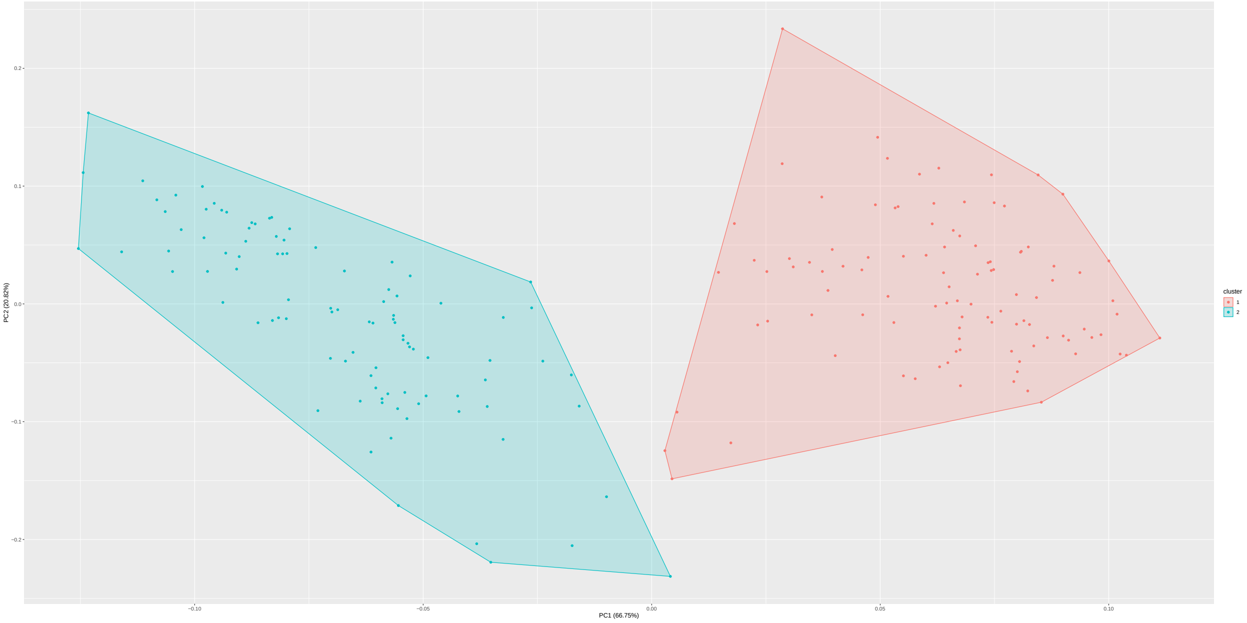
We first plot a score plot. The score plot plots each of the bank notes (Real or Fake) against each of the Principle Components.

```r
PCAcolors <- c("#66c2a5","#8da0cb")[unique(as.integer(df2$status))]


PCAscores <- z_pca$x
PCAloadings <- z_pca$rotation


par(mfrow=c(1,2))
{plot(PCAscores,
     pch=21,
     col=PCAcolors,
     bg=PCAcolors,
     cex=1.5,
     main="Scores"
)
legend("topright",
       legend=unique(df2$status),
       pch=21,
       pt.bg=c("#66c2a5","#8da0cb"),
       pt.cex=1.5,
       col = c("#66c2a5","#8da0cb"))
}

{plot(PCAloadings[,1:2],
     pch=21,
     bg="black",
     cex=1,
     main="Loadings"
)
text(PCAloadings[,1:2],
     labels=rownames(PCAloadings)
)}
```



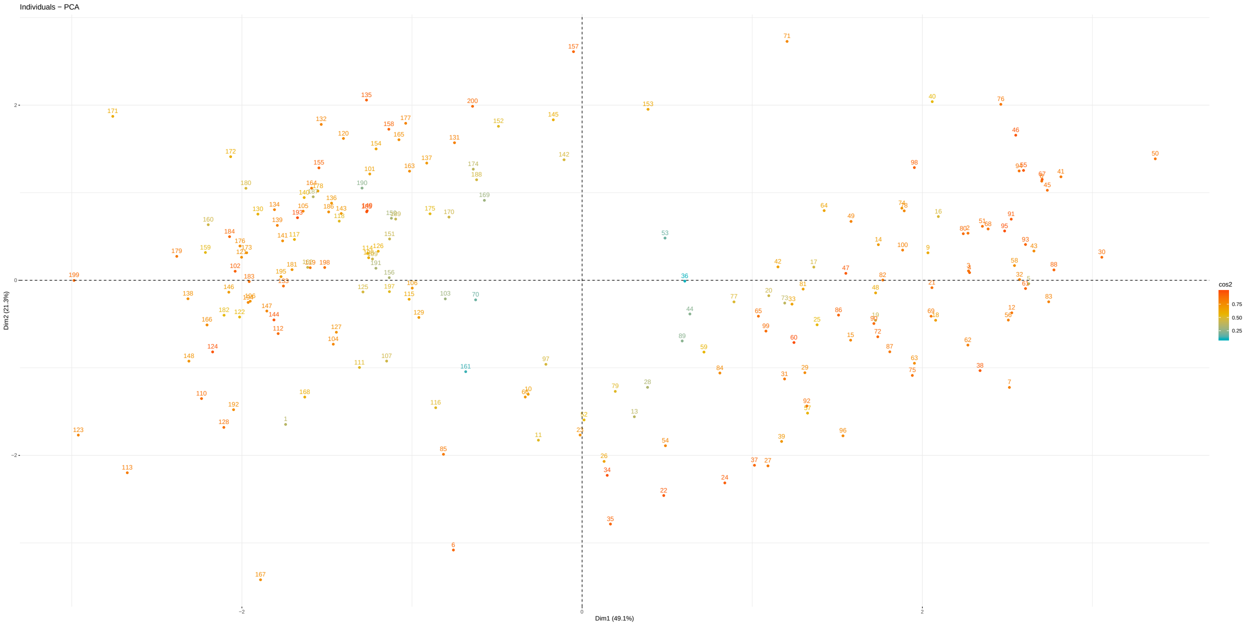
From the score plot we can see a clear distinction between the Real and the Fake Notes. The Real Notes have principle components towards the lower half of the plot.

```
autoplot(fanny(df2[,-7],2),frame=TRUE)
```



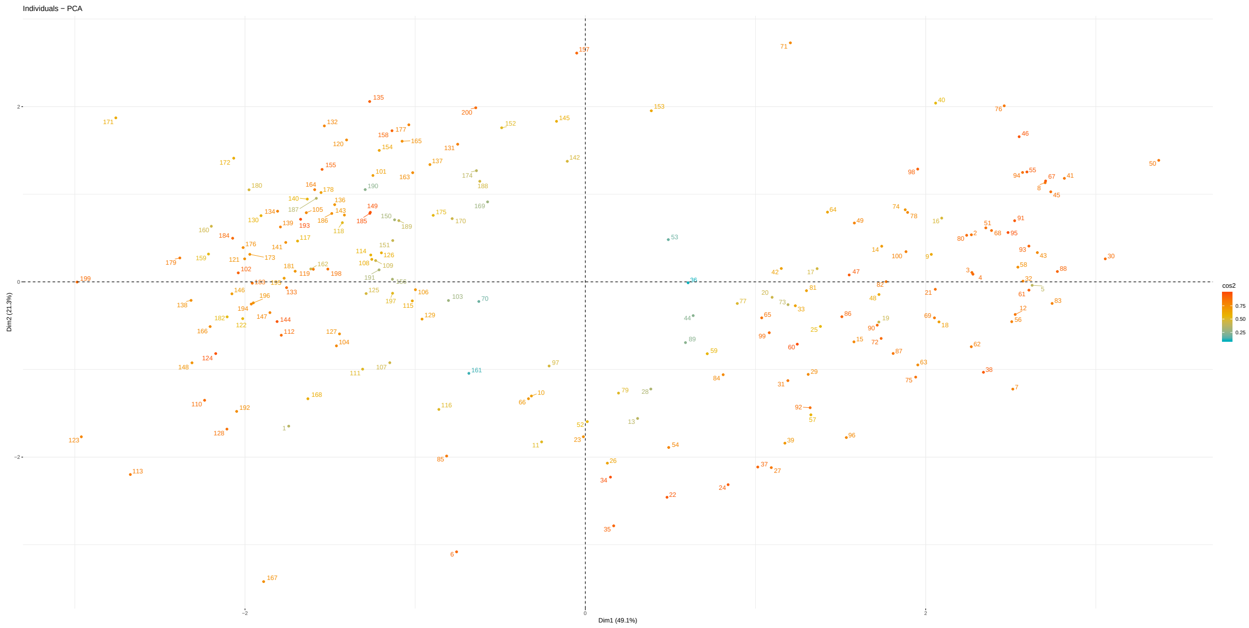
We now map the positively and negatively correlated components for all the True and Fake Notes. We observe the components for each of the fake and real notes and see the 2 clusters (True and Fake Notes)and the components affecting it positively and negatively.

```
fviz_pca_ind(z_pca,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = FALSE)
```



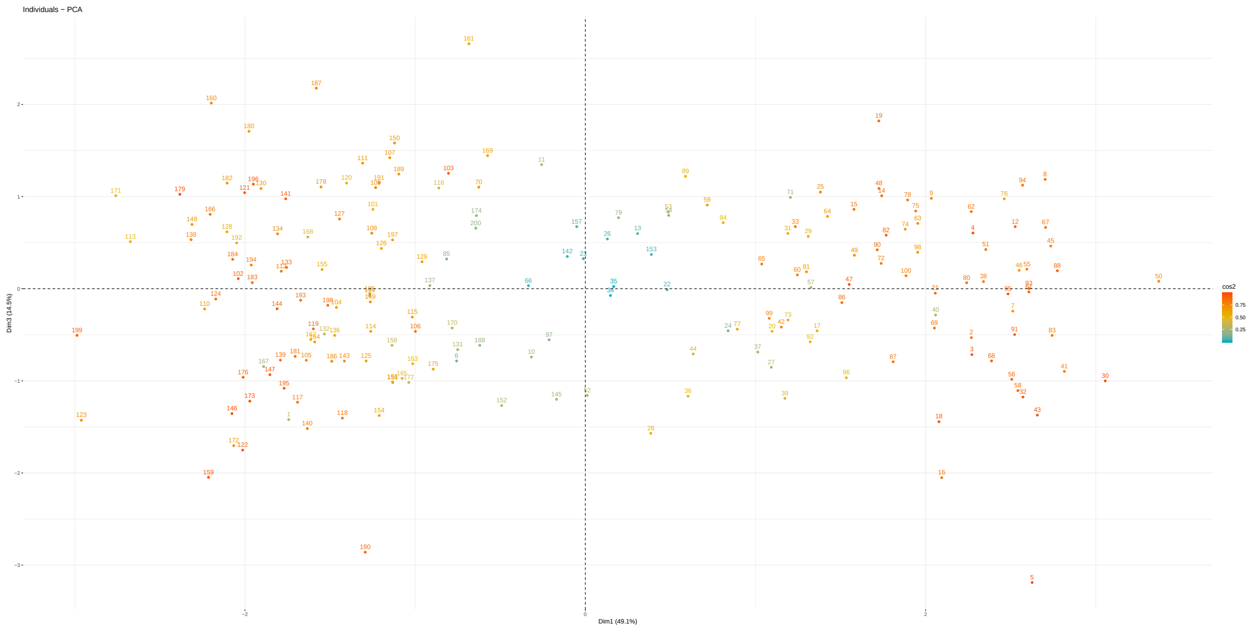
We now map the positively and negatively correlated components for all the True and Fake Notes. We observe the components and the corresponding vectors for each of the fake and real notes and see the 2 clusters (True and Fake Notes)and the components affecting it positively and negatively and its spread across the feature space.

```
fviz_pca_ind(z_pca,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)
```



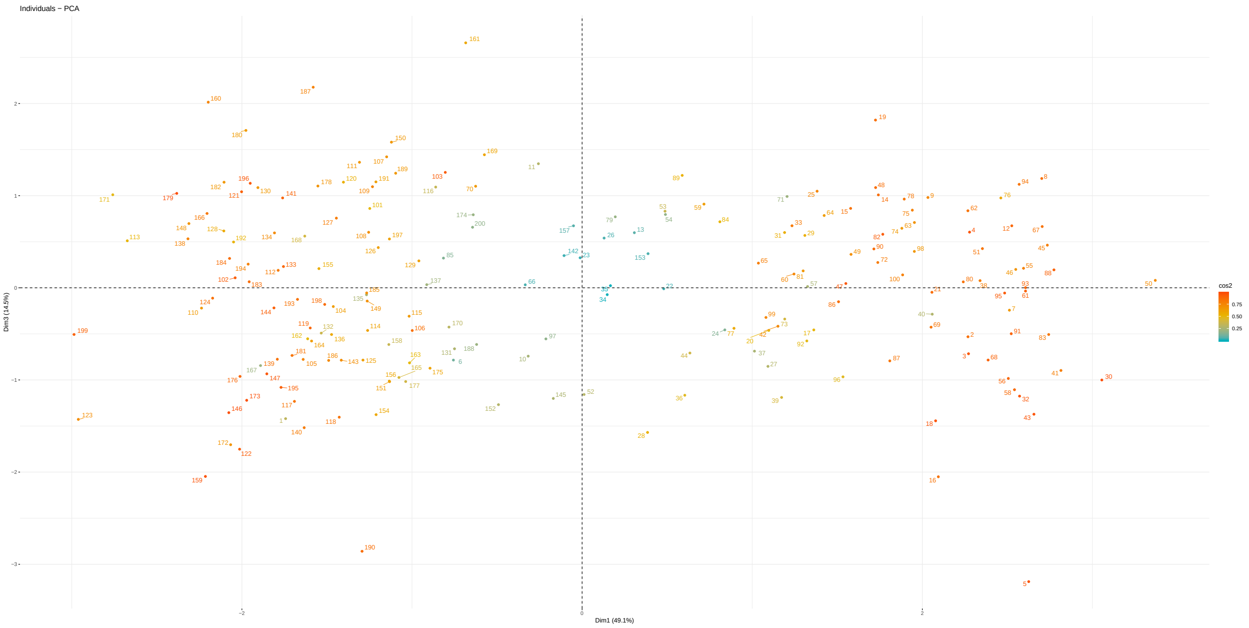
We now map the positively and negatively correlated components for all the True and Fake Notes in the first and third principal components. We observe the components for each of the fake and real notes and see the 2 clusters (True and Fake Notes)and the components affecting it positively and negatively.

```
fviz_pca_ind(z_pca,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = FALSE,axes
```



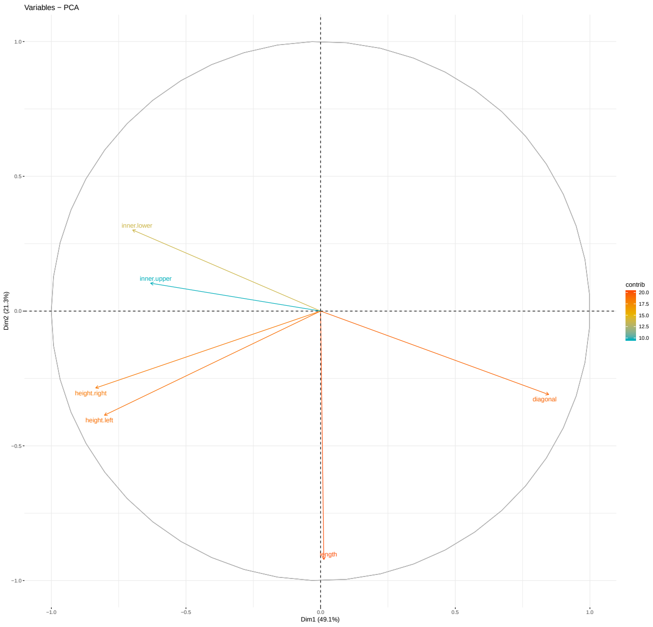
We now map the positively and negatively correlated components for all the True and Fake Notes in the first and third principal components. We observe the components and the corresponding vectors for each of the fake and real notes and see the 2 clusters (True and Fake Notes)and the components affecting it positively and negatively and its spread across the feature space.

```
fviz_pca_ind(z_pca,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE,axes =
```



We now plot the Eigen Vectors across the sample space and observe the distribution of the factors across each of the principal components and study the effect of each of these factors across the sample space.

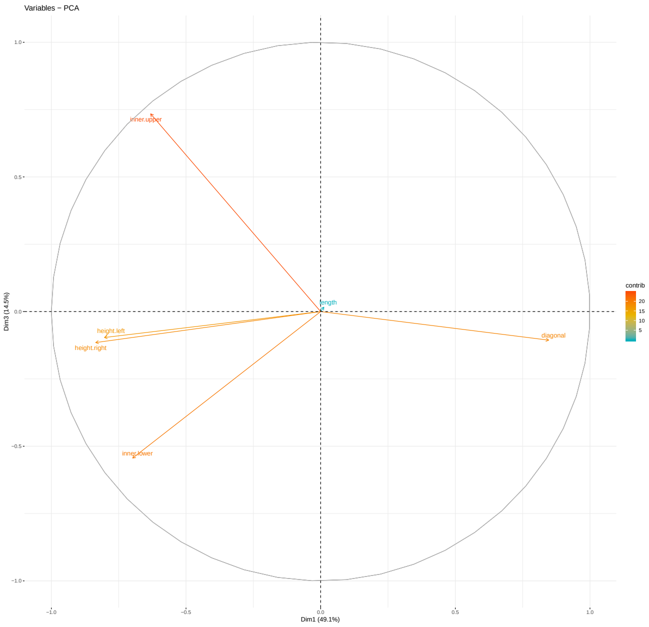
```
fviz_pca_var(z_pca,
col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE
)
```



We can conclude that height(right,left), diagonal, length has the highest contribution across PC1 and PC2 while inner upper has the lowest contribution. inner lower has a moderate contribution to the principle space.

We now plot the Eigen Vectors across the sample space and observe the distribution of the factors across each of the principal components 1 and 3 and study the effect of each of these factors across the sample space.

```
fviz_pca_var(z_pca,
col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE,
axes = c(1, 3)
)
```



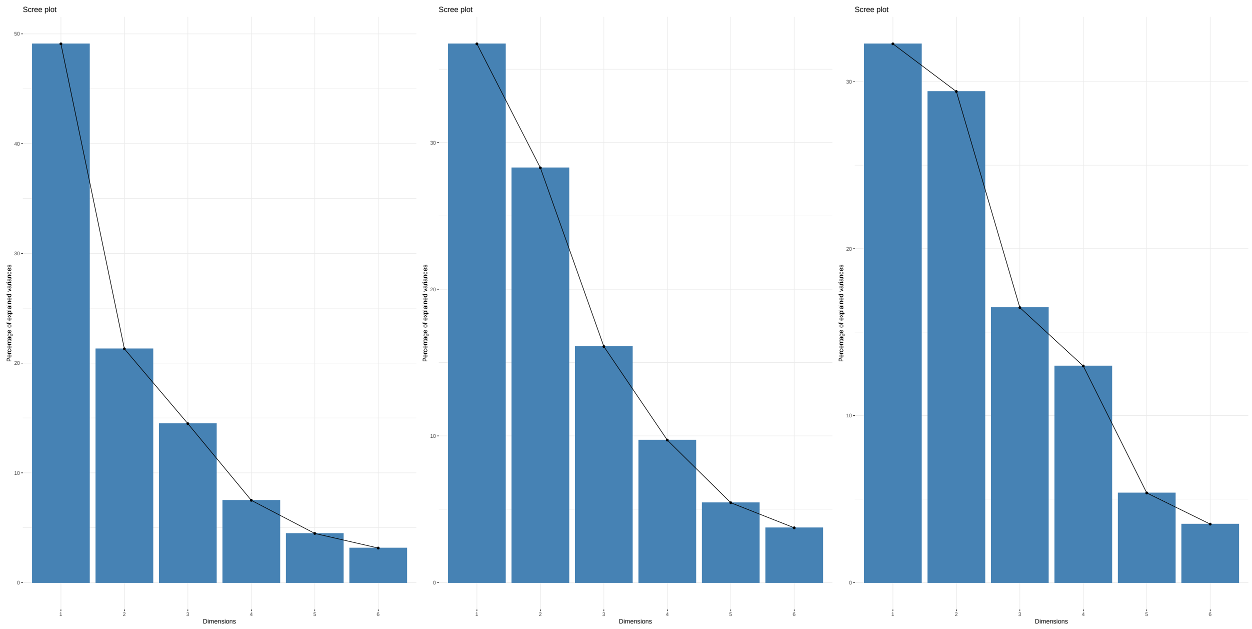
We can conclude that height(right,left), diagonal, inner lower has a moderate contribution to the principle space.length has the lowest contribution to the principle space.inner upper has the highest contribution across PC1 and PC3 while inner upper has the highest contribution.

We now find the principle components for the Real and Fake Notes Separately.

```
d1 <- scale(df2[1:100,-7])
d2 <- scale(df2[101:200,-7])
z_pca1 <- prcomp(d1,center = TRUE,scale. = FALSE)
z_pca2 <- prcomp(d2,center = TRUE,scale. = FALSE)
```

We now plot the Eigen vectors for 1)Both True and False Notes 2)True Notes 3)False Notes

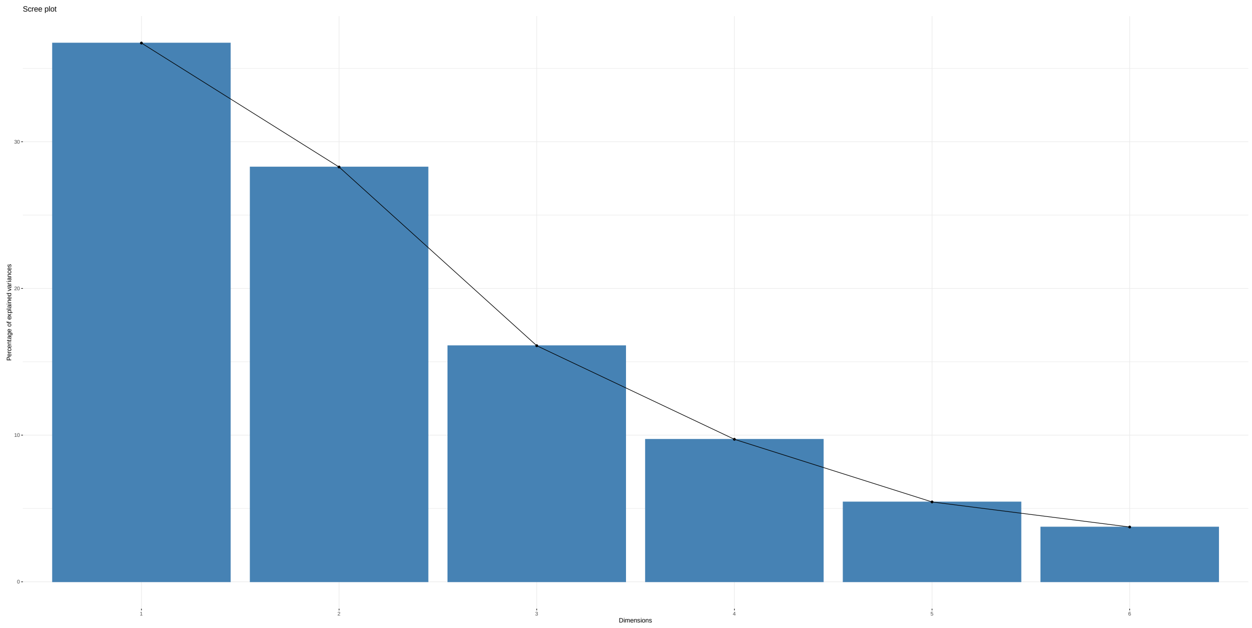
```
plot1 <- fviz_eig(z_pca)
plot2 <- fviz_eig(z_pca1)
plot3 <- fviz_eig(z_pca2)
plot_grid(plot1, plot2, plot3,ncol = 3)
```



The above plot shows comparison of the Percentage Of Variance for 1)Both Real and Fake 2) True 3)Fake notes. We observe that for Real Notes we have highest variances across PC1,PC2. We observe that for Fake Notes we have highest variance across PC1 while all other components are scarcely distributed.

The first plot indicates the Eigen values for True Bank Notes.

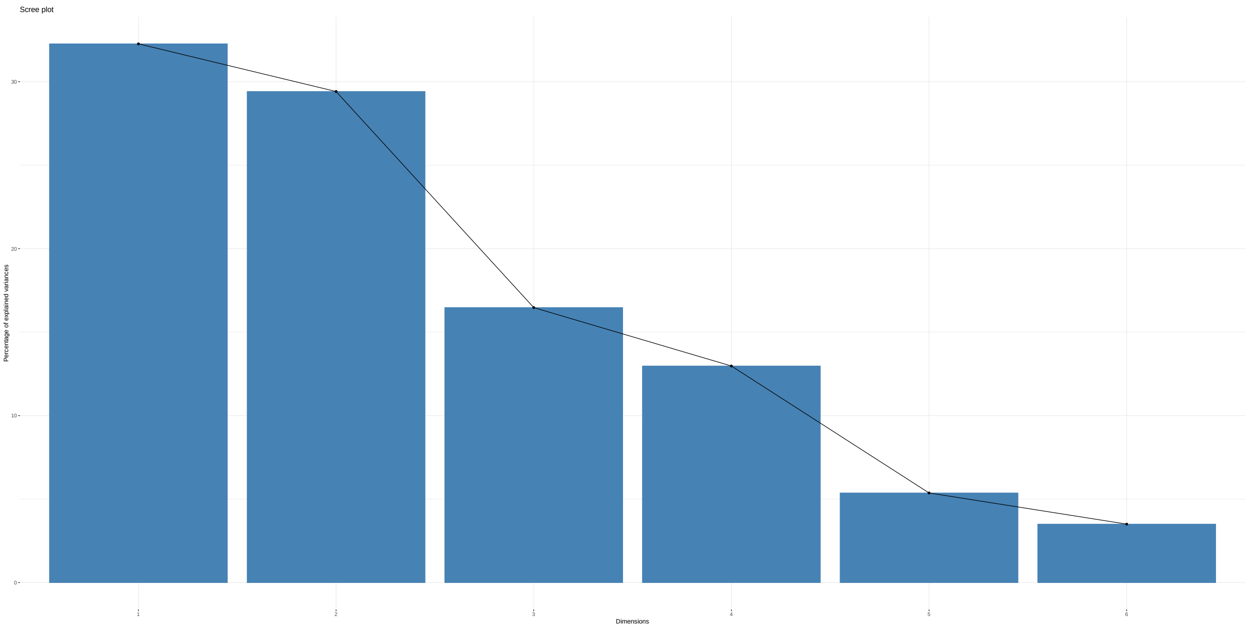
```
plot(plot2)
```



We observe from the first plot that PC1 accounts for 49 Percent variance while PC2 accounts for 28 Percent Variance.

The second plot indicates the Eigen values for Fake Bank Notes.

```
plot(plot3)
```



We observe from the second plot that PC1 accounts for 32 Percent variance while PC2 accounts for 29 Percent Variance.

We now summarize the PCA for both True and Fake Bank Notes.

```
summary(z_pca1)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.4845 1.3026 0.9827 0.76348 0.57156 0.47340
## Proportion of Variance 0.3673 0.2828 0.1610 0.09715 0.05445 0.03735
## Cumulative Proportion  0.3673 0.6501 0.8111 0.90820 0.96265 1.00000
```

```
summary(z_pca2)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6
## Standard deviation     1.3915 1.3285 0.9941 0.8823 0.56755 0.45840
## Proportion of Variance 0.3227 0.2941 0.1647 0.1297 0.05368 0.03502
## Cumulative Proportion  0.3227 0.6169 0.7816 0.9113 0.96498 1.00000
```

We find the Eigen Values for the first 100 values ie., The True Bank Notes.

```
EigenValues1 <- eigen(cov(d1))
EigenValues1$values
```

```
## [1] 2.2038456 1.6967090 0.9657587 0.5828993 0.3266818 0.2241056
```

The Eigen Values obtained above are checked using the following method. We calculate the Standard Deviation of the Principle Components and find the square of it. The Values must be same indicating that they are the intended correct value.

```
(z_pca1$sdev)^2
```

```
## [1] 2.2038456 1.6967090 0.9657587 0.5828993 0.3266818 0.2241056
```

We find the Eigen Values for the second 100 values ie., The Fake Bank Notes.

```
EigenValues2 <- eigen(cov(d2))
EigenValues2$values
```

```
## [1] 1.9362148 1.7648628 0.9883142 0.7783672 0.3221096 0.2101314
```

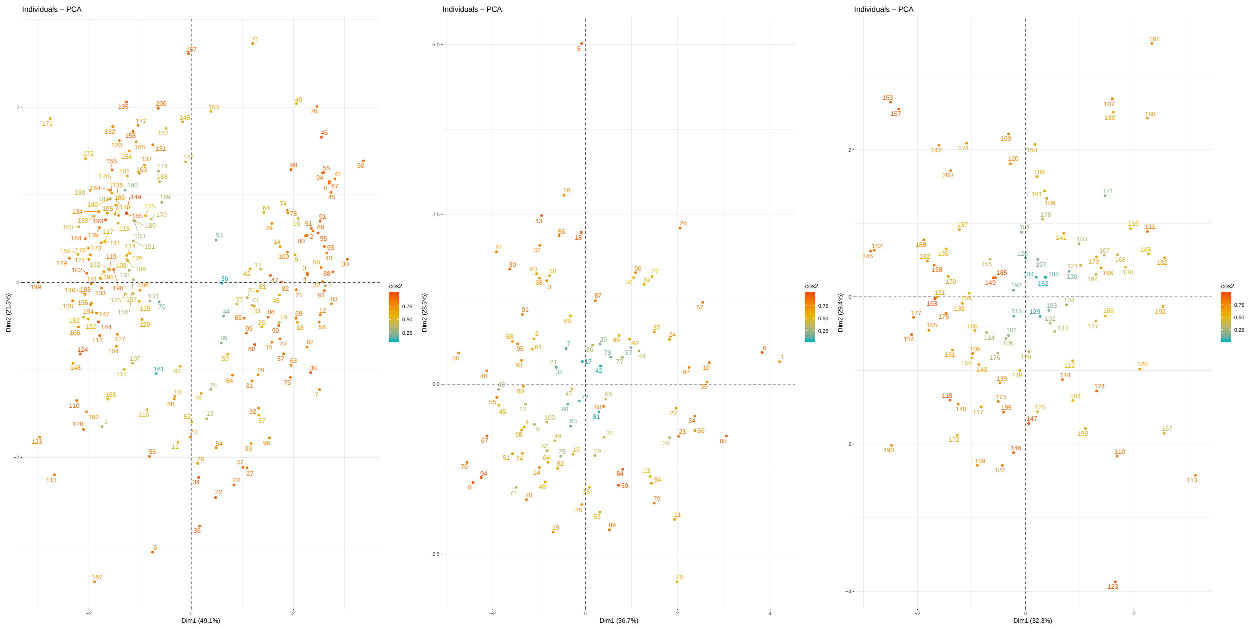
The Eigen Values obtained above are checked using the following method. We calculate the Standard Deviation of the Principle Components and find the square of it. The Values must be same indicating that they are the intended correct value.

```
(z_pca2$sdev)^2
```

```
## [1] 1.9362148 1.7648628 0.9883142 0.7783672 0.3221096 0.2101314
```

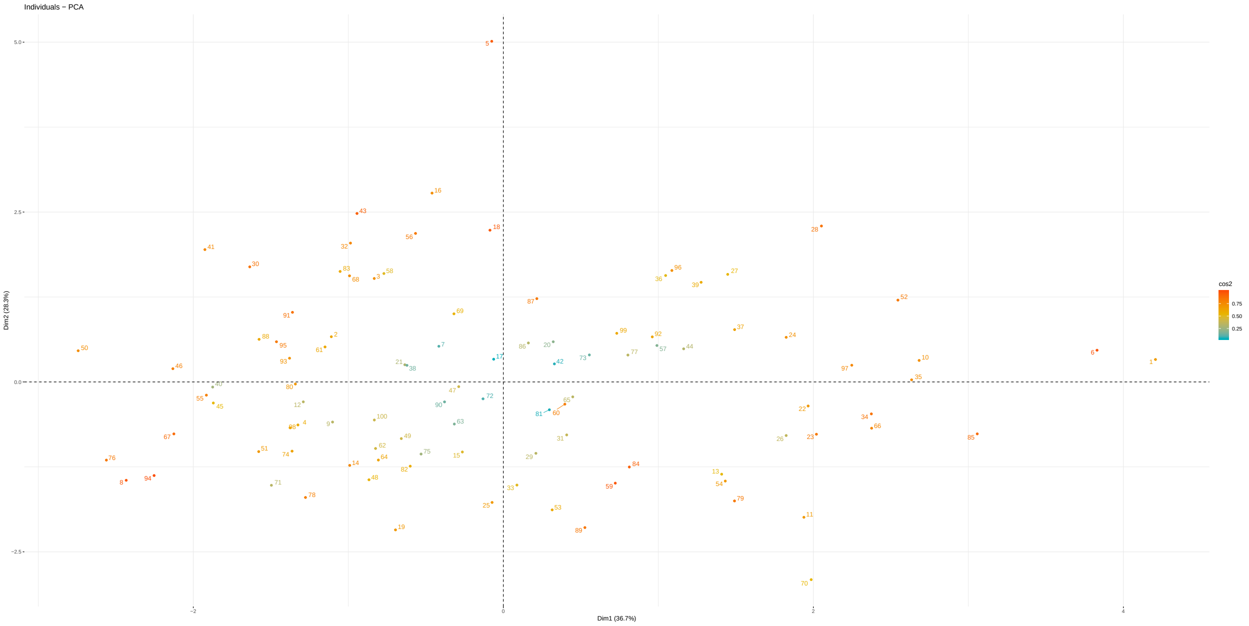
We now plot the principle components and then plot the variables across the principle components. plot 2 indicates the True Bank Notes while the plot3 indicates the False Bank Notes.

```
plot1 <- fviz_pca_ind(z_pca,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = T
plot2 <- fviz_pca_ind(z_pca1,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = T
plot3 <- fviz_pca_ind(z_pca2,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = T
plot_grid(plot1, plot2, plot3,ncol=3)
```

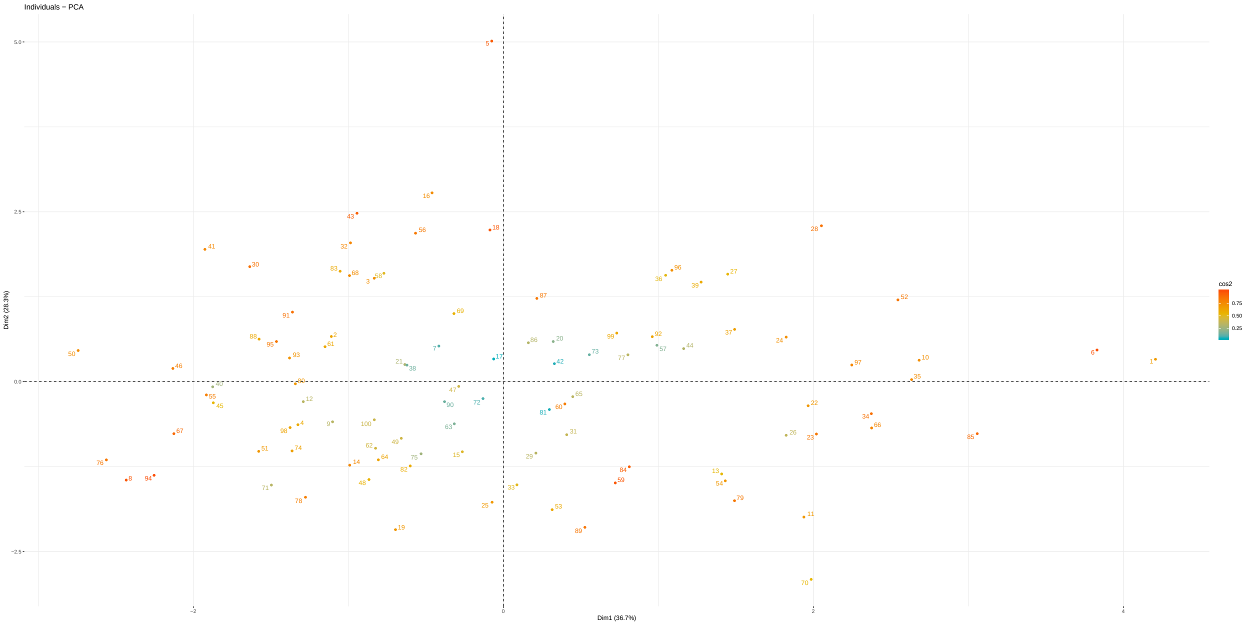


We now plot the Individual PCA components and also the Variables that affect the PCA. We observe that the True Bank Notes have higeher contribution from Inner(lower,upper), height(left,right), medium contribution from length,low contribution from diagonal.

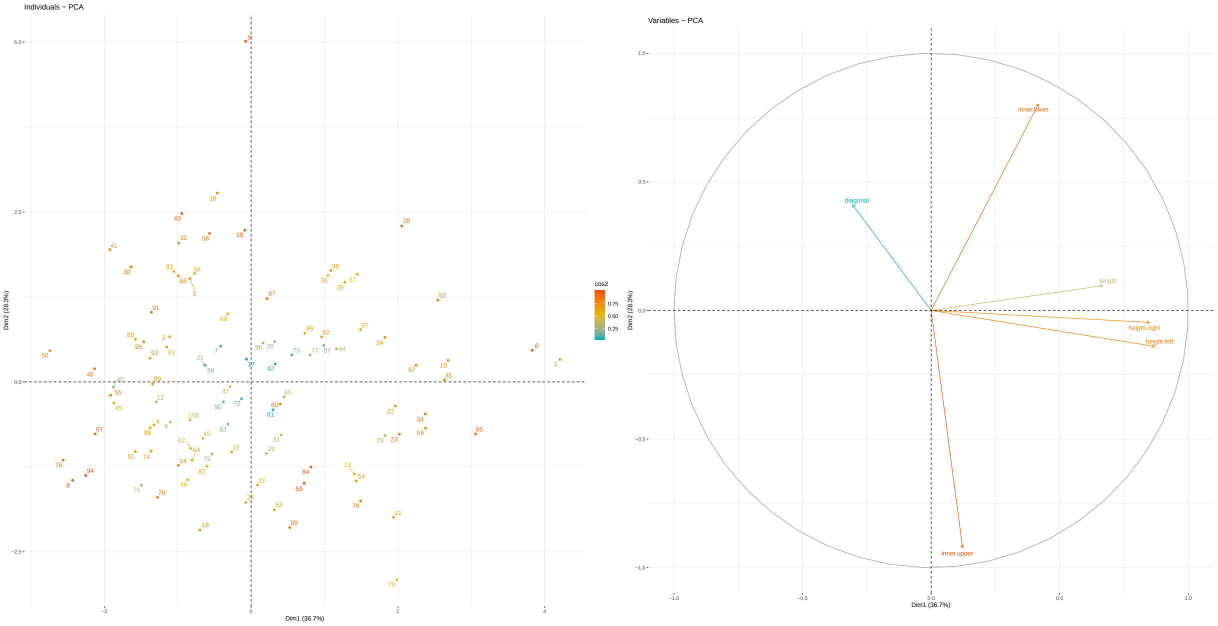
```
plot2
```



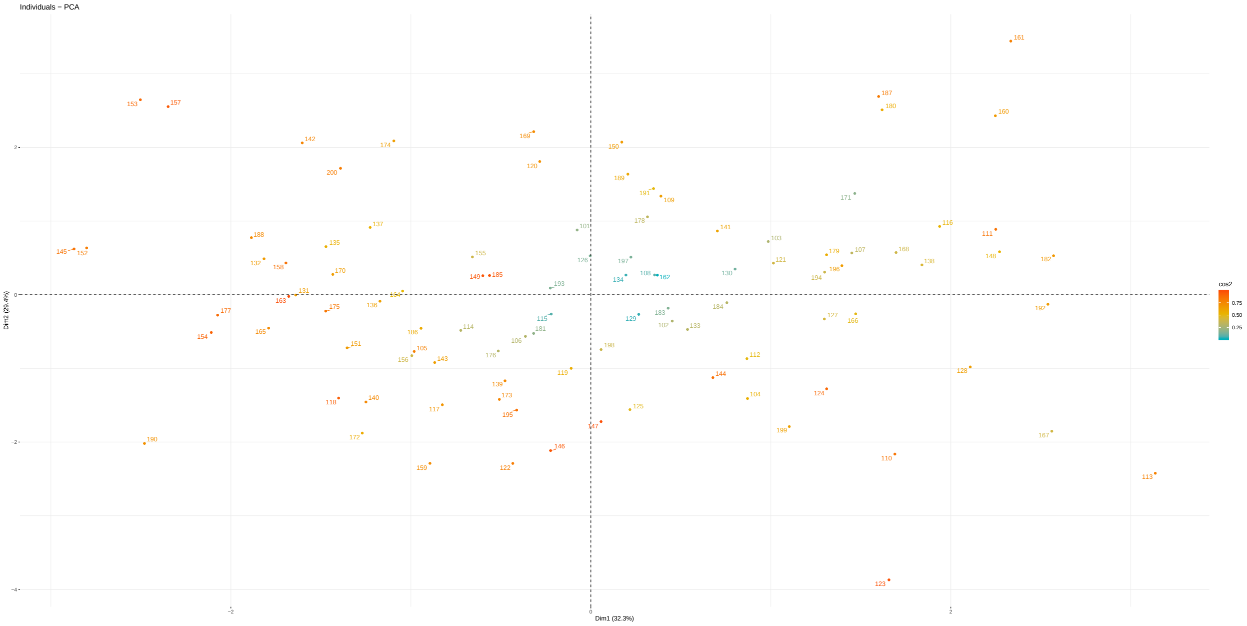
```
plot2
```



```
plot2_1 <- fviz_pca_var(z_pca1,
col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE
)
plot_grid(plot2, plot2_1,ncol = 2)
```

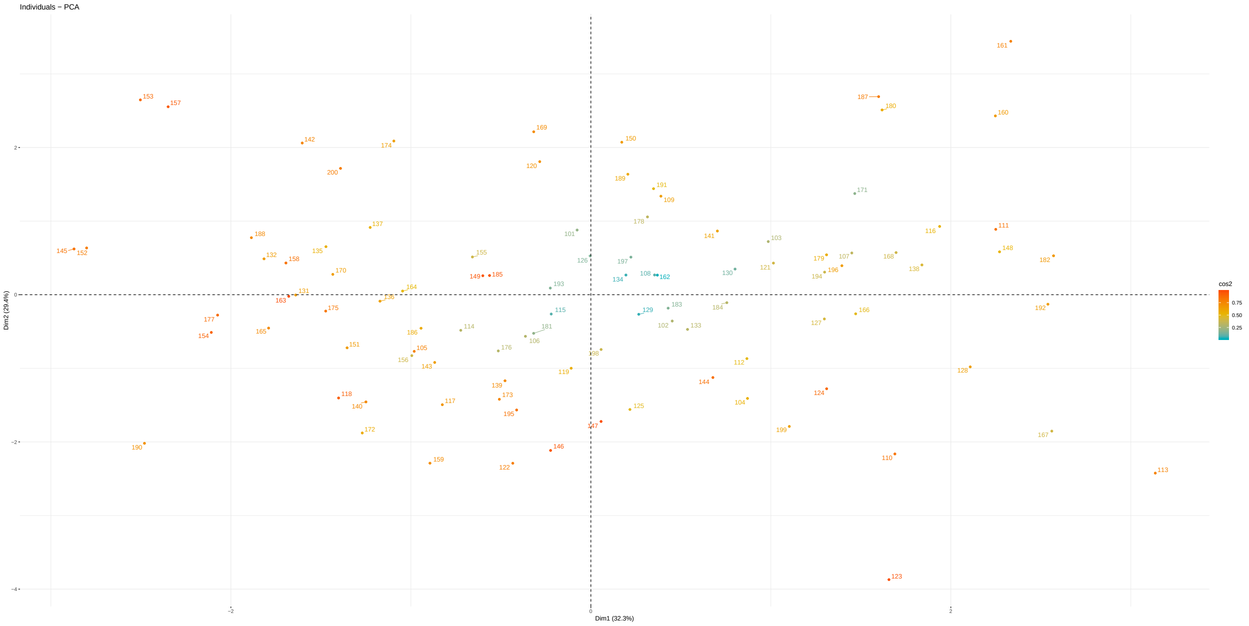


```
plot3
```

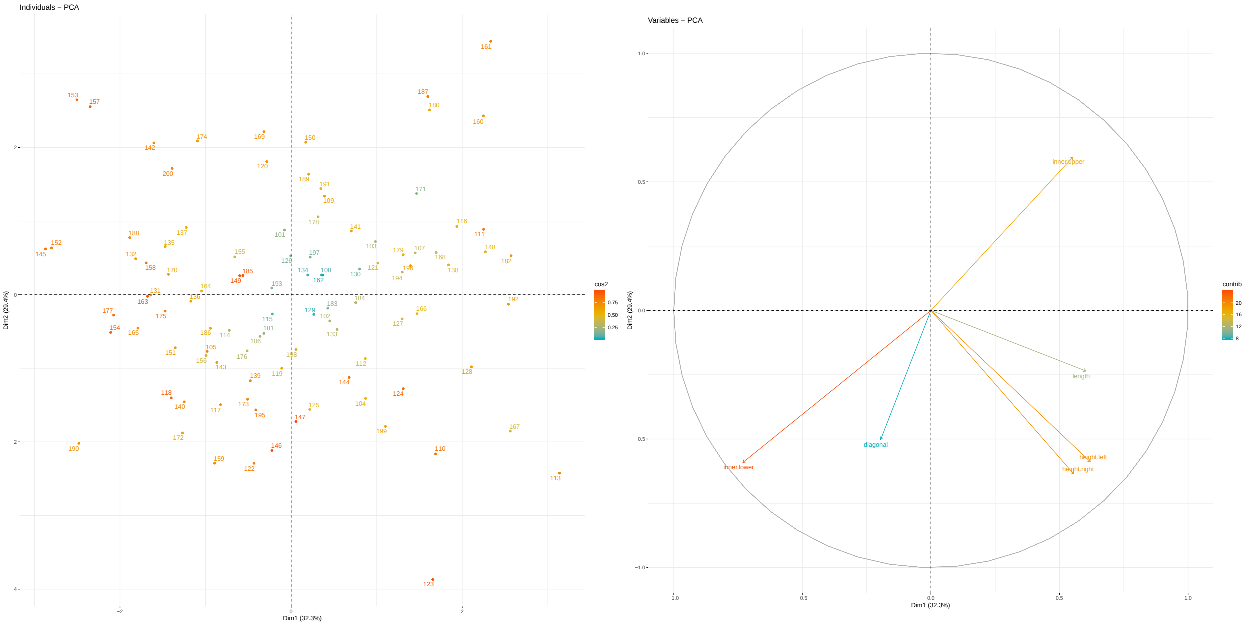


We now plot the Individual PCA components and also the Variables that affect the PCA. We observe that the False Bank Notes have higeher contribution from Inner(lower) and height(left,right), Inner(upper) has high contribution.length has medium contribution.lowest contribution from diagonal.

```
plot3
```

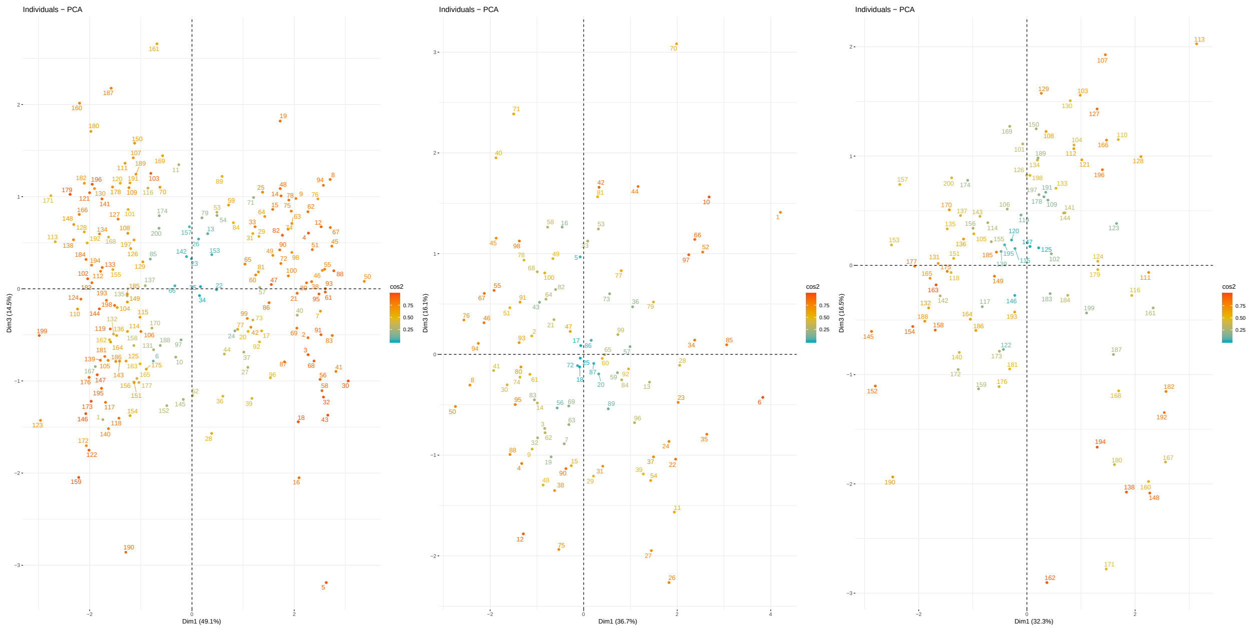


```
plot3_1 <- fviz_pca_var(z_pca2,
col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE
)
plot_grid(plot3,plot3_1,ncol = 2)
```

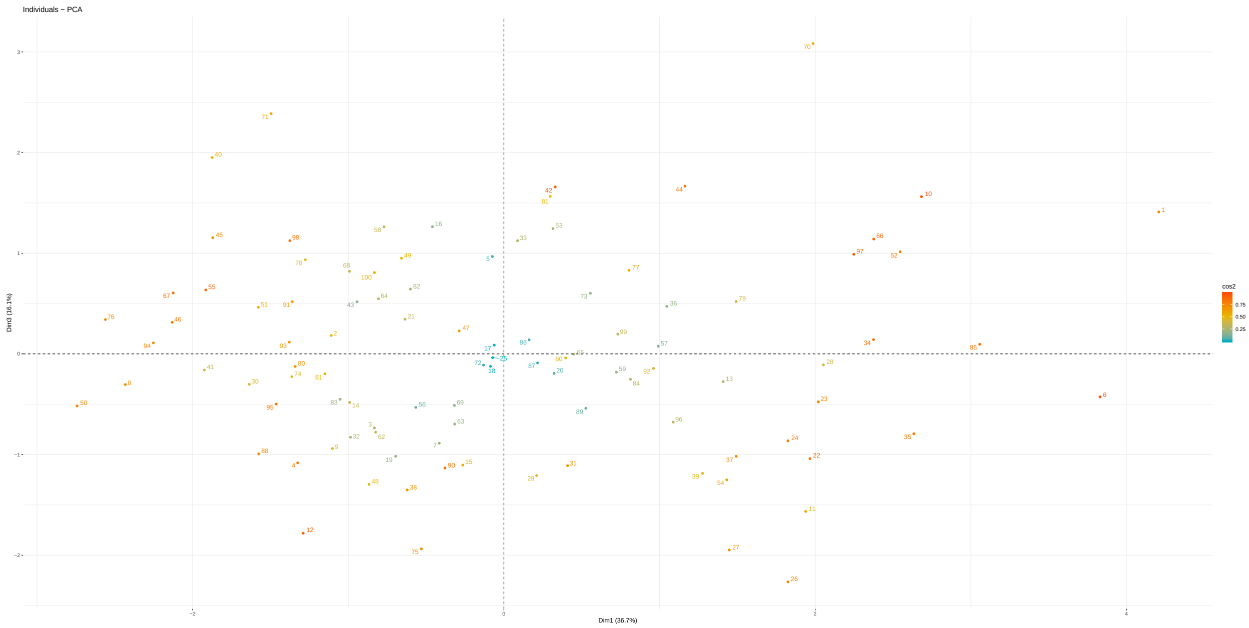


We now plot the Principle Components across First and Third Dimension.

```
plot1 <- fviz_pca_ind(z_pca,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = T
plot2 <- fviz_pca_ind(z_pca1,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = 
plot3 <- fviz_pca_ind(z_pca2,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = 
plot_grid(plot1, plot2, plot3,ncol=3)
```

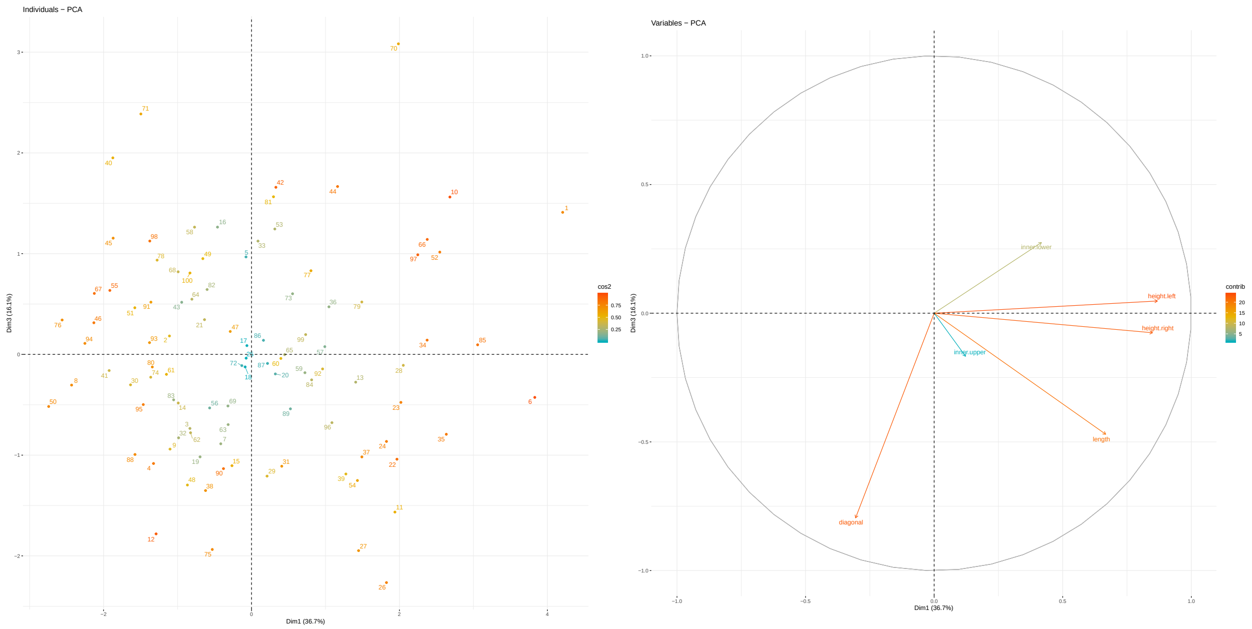


```
plot2
```

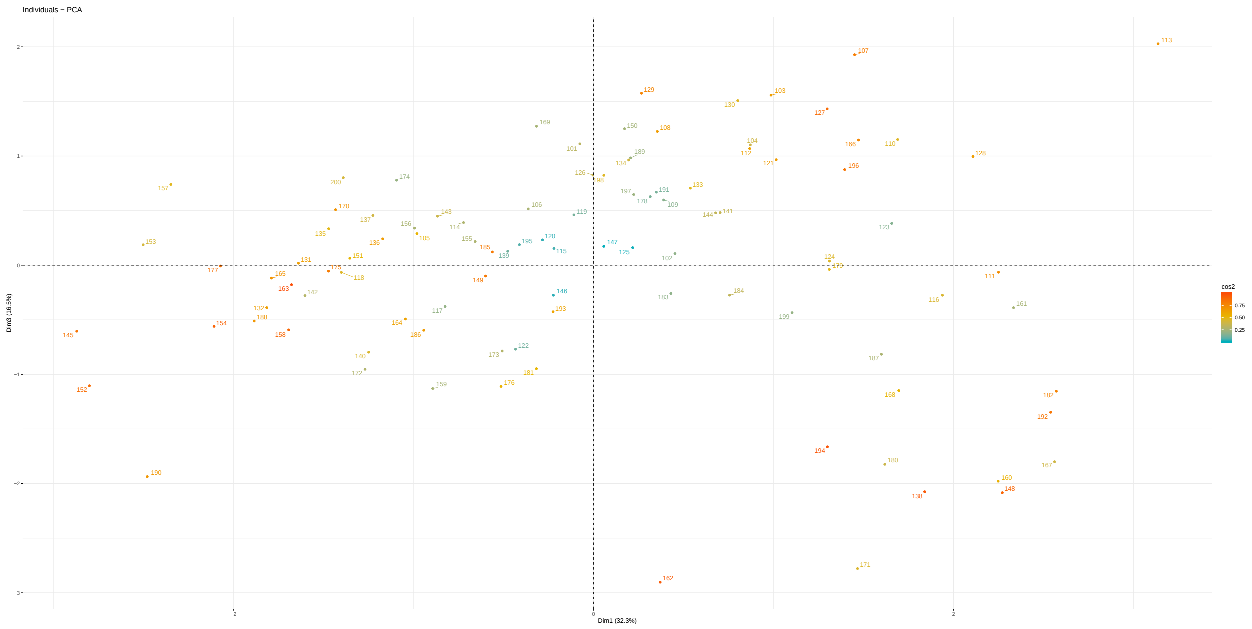


We now plot the Individual PCA components across dimensions 1,3 and also the Variables that affect the PCA. We observe that the True Bank Notes have higher contribution from Inner(lower,upper), height(left,right), medium contribution from length,low contribution from diagonal.

```
plot2 <- fviz_pca_ind(z_pca1,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = '
plot2_1 <- fviz_pca_var(z_pca1,
col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE,
axes = c(1, 3)
)
plot_grid(plot2, plot2_1,ncol = 2)
```

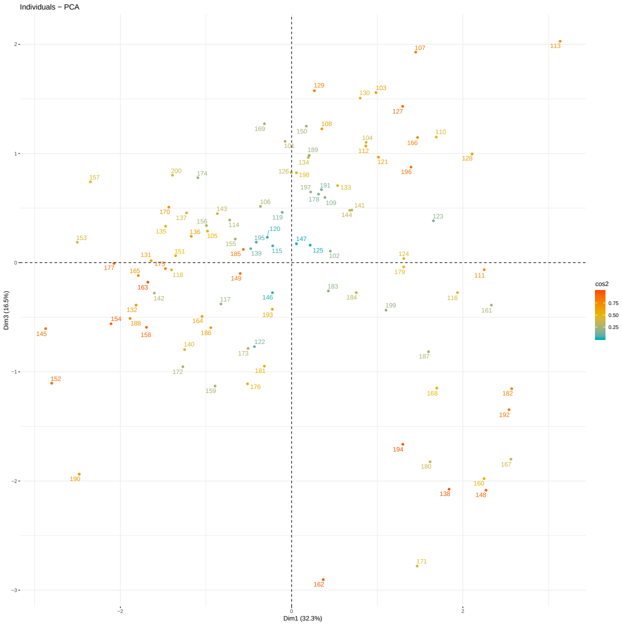


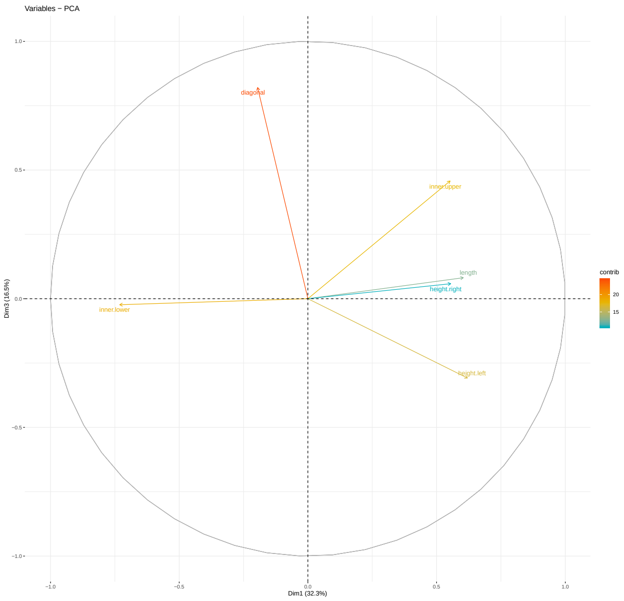
```
plot3
```



We now plot the Individual PCA components across dimensions 1,3 and also the Variables that affect the PCA. We observe that the False Bank Notes have higher contribution from diagonal.Inner(lower,upper), height(left) has medium contribution.length, weight right has low contribution.

```
plot3 <- fviz_pca_ind(z_pca2,col.ind = "cos2",gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),repel = '
plot3_1 <- fviz_pca_var(z_pca2,
col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE,
axes = c(1, 3)
)
plot_grid(plot3, plot3_1,ncol = 2)
```


Individuals − PCA
Dim3 (16.5%)
Dim1 (32.3%)
2
1
0
−1
−2
−3
−2
0
2
cos2
0.75
0.50
0.25
113
107
129
103
130
127
169
150
108
110
104
166
101
189
112
128
121
134
196
157
200
174
126
198
191
133
197
178
109
141
143
106
144
170
137
119
123
156
114
135
120
136
105
195
147
153
155
185
139
115
125
102
124
131
151
177
175
118
165
179
111
149
163
183
142
117
146
184
116
132
199
161
154
193
164
188
145
158
186
140
122
173
187
181
172
152
176
159
168
182
192
194
180
167
160
190
138
148
171
162


Variables − PCA
Dim3 (16.5%)
Dim1 (32.3%)
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
diagonal
inner.upper
length
height.right
inner.lower
height.left
contrib
20
15