All Countries Environmental Data

## Method: PCA

Principal component analysis (PCA), part of descriptive analytics, is used to analyze one table of quantitative data, specifically useful for *high dimensional data* and comparitively lesser data rows. PCA mixes the input variables to give new variables, called principal components. The first principal component is the line of best fit. It is the line that maximizes the inertia (similar to variance) of the cloud of data points. Subsequent components are defined as orthogonal to previous components, and maximize the remaining inertia.

PCA gives one map for the rows (called factor scores), and one map for the columns (called loadings). These 2 maps are related, because they both are described by the same components. However, these 2 maps project different kinds of information onto the components, and so they are *interpreted differently*. Factor scores are the coordinates of the row observations and Loadings describe the column variables. Both can be interpreted through their distance from origin. However, Factor scores are also interpreted by the distances between them and Loadings interpreted by the angle between them.

The distance from the origin is important in both maps, because squared distance from the mean is inertia (variance, information; see sum of squares as in ANOVA/regression). Because of the Pythagorean Theorem, the total information contributed by a data point (its squared distance to the origin) is also equal to the sum of its squared factor scores.

With both Factor and Loadings maps combined we can interpret which grouping criteria of rows of data is most impacted by which columns. This can interpreted visually by observing which a factors and loadings on a particular component and the distance on this component.

PCA also helps in *dimensionality reduction*. Using SVD, we get eigen values arranged in descending order in the diagonal matrix. We can simply ignore the lower eigen values to reduce dimensions. We can also take help of SCREE plot to visually analyze importance of eigen values.

## Dataset

* Data: Measurements of environment conditions in Countries
* Rows: There are 137 observations, 1 for each country.
* Columns: Total 29 variables
* Qualitative: Country (nominal), Happiness (Ordinal).
* Quantitative: Aspect, Slope Crop Land, Tree Canopy Wind Cloud & Multiple variables for Temp & Rain
* Summary of Data

summary(country\_env\_df)

## Country Happiness\_Rank accessibility\_to\_cities  
## Afghanistan: 1 VH:46 Min. : 0.969   
## Albania : 1 H :45 1st Qu.: 55.216   
## Algeria : 1 U :46 Median : 122.848   
## Angola : 1 Mean : 321.889   
## Argentina : 1 3rd Qu.: 338.201   
## Armenia : 1 Max. :3220.715   
## (Other) :131   
## elevation aspect slope cropland\_cover   
## Min. : 8.319 Min. : 94.24 Min. :0.0000 Min. : 0.00553   
## 1st Qu.: 251.926 1st Qu.:168.27 1st Qu.:0.1592 1st Qu.: 6.18871   
## Median : 425.131 Median :176.14 Median :0.5200 Median :19.97215   
## Mean : 603.987 Mean :177.11 Mean :0.7470 Mean :24.28494   
## 3rd Qu.: 814.622 3rd Qu.:186.50 3rd Qu.:1.1985 3rd Qu.:36.14503   
## Max. :2962.817 Max. :245.34 Max. :3.6984 Max. :77.52802   
##   
## tree\_canopy\_cover isothermality rain\_coldestQuart rain\_driestMonth   
## Min. : 0.000 Min. :19.35 Min. : 0.0312 Min. : 0.000   
## 1st Qu.: 4.926 1st Qu.:31.83 1st Qu.: 69.6743 1st Qu.: 2.611   
## Median :16.579 Median :44.35 Median : 144.7110 Median : 16.807   
## Mean :19.109 Mean :48.64 Mean : 217.7677 Mean : 27.408   
## 3rd Qu.:31.197 3rd Qu.:65.48 3rd Qu.: 271.8047 3rd Qu.: 40.790   
## Max. :60.085 Max. :87.00 Max. :1271.9904 Max. :157.769   
##   
## rain\_driestQuart rain\_mean\_annual rain\_seasonailty rain\_warmestQuart  
## Min. : 0.00 Min. : 14.16 Min. : 11.16 Min. : 0.00   
## 1st Qu.: 14.01 1st Qu.: 501.66 1st Qu.: 35.77 1st Qu.: 99.63   
## Median : 67.22 Median : 847.35 Median : 59.71 Median : 217.20   
## Mean : 97.34 Mean :1033.53 Mean : 62.51 Mean : 267.52   
## 3rd Qu.:141.79 3rd Qu.:1486.49 3rd Qu.: 87.52 3rd Qu.: 366.79   
## Max. :537.83 Max. :2972.14 Max. :144.51 Max. :1195.09   
##   
## rain\_wettestMonth rain\_wettestQuart temp\_annual\_range temp\_coldestQuart   
## Min. : 3.757 Min. : 8.929 Min. : 9.275 Min. :-24.7437   
## 1st Qu.: 79.354 1st Qu.: 213.785 1st Qu.:16.848 1st Qu.: 0.4906   
## Median :124.978 Median : 338.736 Median :25.775 Median : 13.0249   
## Mean :173.078 Mean : 454.883 Mean :25.110 Mean : 10.7980   
## 3rd Qu.:270.387 3rd Qu.: 716.121 3rd Qu.:30.863 3rd Qu.: 21.5560   
## Max. :537.011 Max. :1441.200 Max. :50.631 Max. : 26.2576   
##   
## temp\_diurnal\_range temp\_driestQuart temp\_max\_warmestMonth  
## Min. : 6.154 Min. :-19.847 Min. :16.43   
## 1st Qu.: 9.114 1st Qu.: 8.557 1st Qu.:25.50   
## Median :10.622 Median : 19.985 Median :30.72   
## Mean :10.927 Mean : 16.240 Mean :29.98   
## 3rd Qu.:12.612 3rd Qu.: 24.828 3rd Qu.:34.05   
## Max. :16.269 Max. : 35.217 Max. :43.84   
##   
## temp\_mean\_annual temp\_min\_coldestMonth temp\_seasonality   
## Min. :-6.832 Min. :-30.679 Min. : 3.001   
## 1st Qu.: 9.789 1st Qu.: -4.007 1st Qu.: 16.811   
## Median :19.845 Median : 5.713 Median : 48.591   
## Mean :17.215 Mean : 4.872 Mean : 48.573   
## 3rd Qu.:24.395 3rd Qu.: 15.195 3rd Qu.: 71.384   
## Max. :28.250 Max. : 22.202 Max. :145.221   
##   
## temp\_warmestQuart temp\_wettestQuart wind cloudiness   
## Min. : 9.504 Min. : 1.567 Min. :1.203 Min. : 50.36   
## 1st Qu.:18.643 1st Qu.:11.657 1st Qu.:2.202 1st Qu.:154.88   
## Median :24.352 Median :19.084 Median :2.796 Median :199.37   
## Mean :23.185 Mean :18.198 Mean :2.927 Mean :190.10   
## 3rd Qu.:27.270 3rd Qu.:24.748 3rd Qu.:3.586 3rd Qu.:234.12   
## Max. :35.217 Max. :31.853 Max. :4.724 Max. :322.19   
##

* Research Question

How do the 137 countries differ on these variables?

## accessibility\_to\_cities elevation aspect slope  
## Afghanistan(U) 317.71575 1831.7444 201.4298 1.5156001  
## Albania(H) 73.83086 651.8155 192.1303 1.8900753  
## Algeria(H) 1212.79982 556.7583 184.9747 0.1708615  
## Angola(U) 378.20239 1061.4790 174.2569 0.1926286  
## Argentina(VH) 209.21958 682.7993 145.0314 0.6238553  
## Armenia(U) 97.29452 1850.4830 183.5375 2.3188956  
## cropland\_cover tree\_canopy\_cover isothermality  
## Afghanistan(U) 9.511846 0.3746726 35.90442  
## Albania(H) 23.346087 12.8046289 33.16941  
## Algeria(H) 3.690864 0.1766562 40.29895  
## Angola(U) 2.794476 19.8701092 64.33239  
## Argentina(VH) 21.962504 8.8336096 49.85147  
## Armenia(U) 21.338266 6.9929146 32.42359  
## rain\_mean\_annual temp\_mean\_annual wind cloudiness  
## Afghanistan(U) 311.32914 11.53331 3.432747 114.22898  
## Albania(H) 1151.09747 11.46749 2.472694 181.13111  
## Algeria(H) 79.45607 22.96395 4.025770 90.67439  
## Angola(U) 1023.37470 21.61187 2.164405 187.51704  
## Argentina(VH) 539.87247 14.20719 4.270904 159.00653  
## Armenia(U) 501.65877 6.16530 1.968253 191.76167

* PCA

country\_env\_pca <- epPCA(DATA = country\_env\_df\_for\_pca, center = TRUE, scale = TRUE, DESIGN = country\_env\_df$Happiness\_Rank, graphs = FALSE)

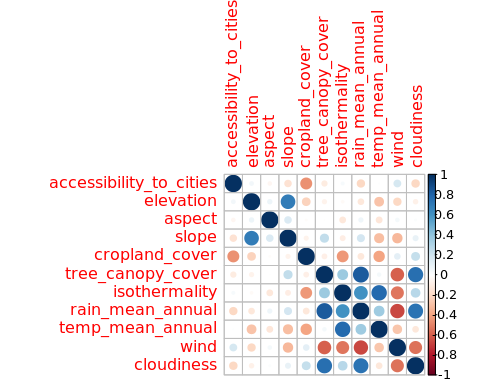
## Results

There are multiple variables representing rain and Temp. Hence, for analysis purposes, lets choose annual mean for Rain and Temp.

### Correlation Plot

Visually analyze multicollinearity in the system

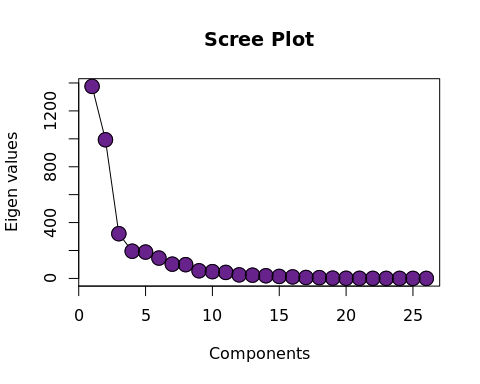
corr\_result = cor(country\_env\_df\_for\_corr)  
corrplot(corr\_result)



### Scree Plot

Gives amount of information explained by corresponding component.

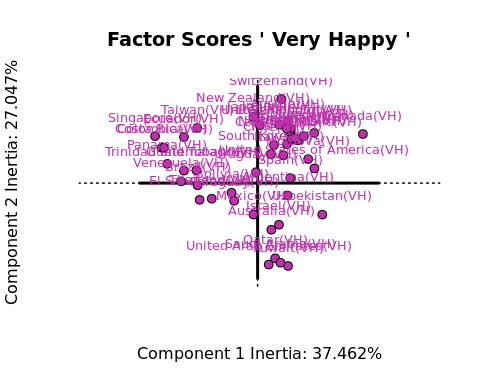
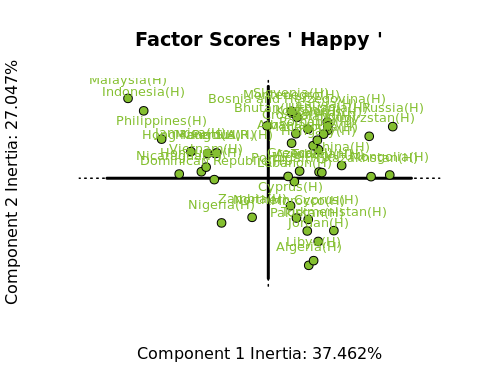
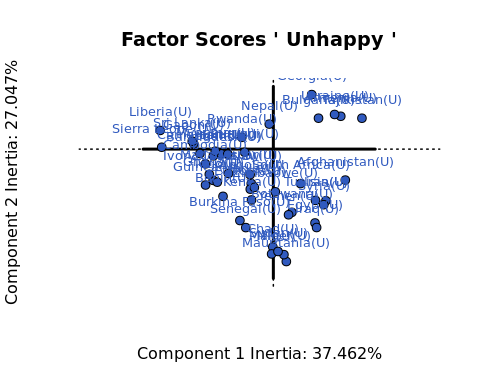
country\_env\_scree <- plot(country\_env\_pca$ExPosition.Data$eigs, ylab = "Eigen values",  
 xlab = "Components",  
 type = "l",  
 main = "Scree Plot",)  
points(country\_env\_pca$ExPosition.Data$eigs, cex = 2, pch = 19, col = "darkorchid4")  
points(country\_env\_pca$ExPosition.Data$eigs, cex = 2, pch = 21, col = "black")



### Factor Scores

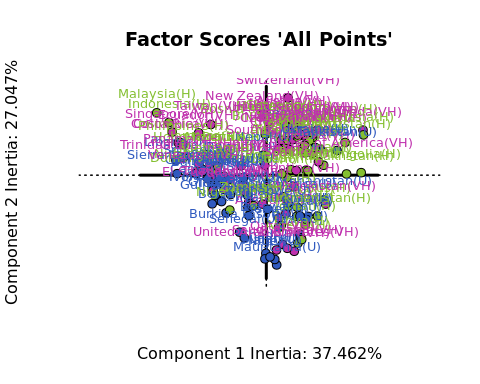
Lets visualize each happiness category “individually” on the Factor Plot to understand the spreed of data over Components 1 and 2.

#par(mfrow=c(2,2))  
  
list\_name <- c('Unhappy', 'Happy', 'Very Happy')  
i = 1  
for (l in groups\_of\_data) {  
 title = list\_name[i]  
 i = i + 1  
 prettyPlot(data\_matrix = country\_env\_pca$ExPosition.Data$fi[l,],   
 dev.new=FALSE,  
 main = paste("Factor Scores '", title, "'"),  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE,  
 display\_points = TRUE, pch = 21, cex = 1.2,   
 col = country\_env\_pca$Plotting.Data$fi.col[l,],   
 display\_names = TRUE,   
 xlab = paste0("Component 1 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[2],3), "%")  
 )  
}



Lets visualize each happiness category “all at once” on the Factor Plot to understand the spreed of data over Components 1 and 2.

prettyPlot(data\_matrix = country\_env\_pca$ExPosition.Data$fi,   
 dev.new=FALSE,  
 main = paste("Factor Scores 'All Points'"),  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE,  
 display\_points = TRUE, pch = 21, cex = 1.2,   
 col = country\_env\_pca$Plotting.Data$fi.col,   
 display\_names = TRUE,   
 xlab = paste0("Component 1 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[2],3), "%")  
 )

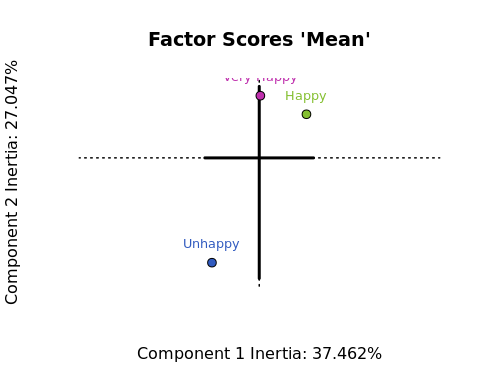


Since the spreed of the data is not clearly differentiated by both components, let’s try to visualize the means of each happiness category.

* Ploting only mean points for each happiness category.

country\_env\_pca\_mean = aggregate(x=country\_env\_pca$ExPosition.Data$fi, by = split(col, col(col)), FUN = mean)  
rownames(country\_env\_pca\_mean) <- c('Unhappy', 'Happy', 'Very Happy')  
  
country\_env\_pca\_mean <- country\_env\_pca\_mean[-1]  
country\_env\_pca\_mean

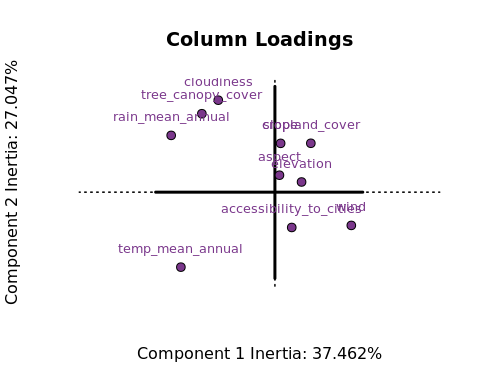
prettyPlot(data\_matrix = country\_env\_pca\_mean,   
 dev.new=FALSE,  
 main = paste("Factor Scores 'Mean'"),  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE,   
 display\_points = TRUE, pch = 21, cex = 1.2,   
 col = unique(col),   
 display\_names = TRUE,   
 xlab = paste0("Component 1 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[2],3), "%")  
 );



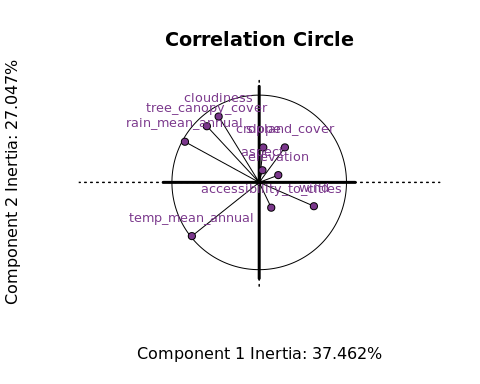
Component 1 and Component 2 seems to clearly differentiate between Very Happy & Happy vs Unhappy countries.

## Loadings

#par(mfrow=c(1,2))  
  
prettyPlot(data\_matrix = country\_env\_pca$ExPosition.Data$fj[c(1,2,3,4,5,6,11, 21, 26,27),],   
 dev.new=FALSE,  
 main = "Column Loadings",  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE,   
 display\_points = TRUE, pch = 21, cex = 1.2,   
 col = country\_env\_pca$Plotting.Data$fj.col[c(1,2,3,4,5,6,11, 21, 26,27),],   
 display\_names = TRUE,   
 xlab = paste0("Component 1 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[2],3), "%")  
 )



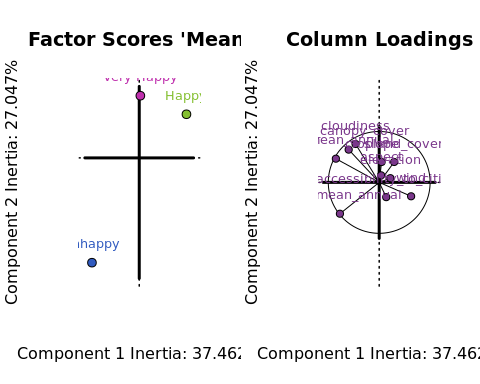
correlationPlotter(data\_matrix = country\_env\_df\_for\_pca[,c(1,2,3,4,5,6,11, 21, 26,27)],  
 factor\_scores = country\_env\_pca$ExPosition.Data$fi[,c(1,2)],   
 dev.new=FALSE,  
 main = "Correlation Circle",  
 x\_axis = 1, y\_axis = 2,   
   
 pch = 21,   
 col = country\_env\_pca$Plotting.Data$fj.col[c(1,2,3,4,5,6,11, 21, 26,27),],   
   
 xlab = paste0("Component 1 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[2],3), "%")  
 )



Rain, compared with Temperature, cloudiness and tree coverage seems to be almost orthogonal to each other, hence are *not* correlated.

## Summary

par(mfrow=c(1,2))  
  
prettyPlot(data\_matrix = country\_env\_pca\_mean,   
 dev.new=FALSE,  
 main = paste("Factor Scores 'Mean'"),  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE,   
 display\_points = TRUE, pch = 21, cex = 1.2,   
 col = unique(col),   
 display\_names = TRUE,   
 xlab = paste0("Component 1 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[2],3), "%")  
 )  
  
correlationPlotter(data\_matrix = country\_env\_df\_for\_pca[,c(1,2,3,4,5,6,11, 21, 26,27)],  
 factor\_scores = country\_env\_pca$ExPosition.Data$fi[,c(1,2)],   
 dev.new=FALSE,  
 main = "Column Loadings",  
 x\_axis = 1, y\_axis = 2,   
   
 pch = 21,   
 col = country\_env\_pca$Plotting.Data$fj.col[c(1,2,3,4,5,6,11, 21, 26,27),],   
   
 xlab = paste0("Component 1 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(country\_env\_pca$ExPosition.Data$t[2],3), "%")  
 )



Viewing the Factor and Loading plots together gives an understanding, based on Component 2, that *Unhappiness* is quite related with Temperature. The more the temperature, the more is unhappiness.