Taller 2 - Componentes

Maria Fernanda Piñeros

23/5/2021

Contents

1 Punto 1

Los siguientes puntos deben realizarse posterior a la limpieza de valores atípicos. Indique los valores omitidos.

La base de datos punto 1 contiene 11 indicadores económicos y sociales de 96 países. Las variables observadas son:

X_1 = Tasa anual de crecimiento de la población,

X_2 = Tasa de mortalidad infantil por cada 1000 nacidos vivos.

X_3= Porcentaje de mujeres en la población activa.

X_4= PNB en 1995 (en millones de dólares).

X_5= Producción de electricidad (en millones kW/h).

X_6= Líneas telefónicas por cada 1000 habitantes,

X_7= Consumo de agua per cápita,

X_8= Proporción de la superficie del país cubierta por bosques,

X_9= Proporción de forestación anual,

X_10= Consumo de energía per cápita,

X_11= Emisión de CO2 per cápita.

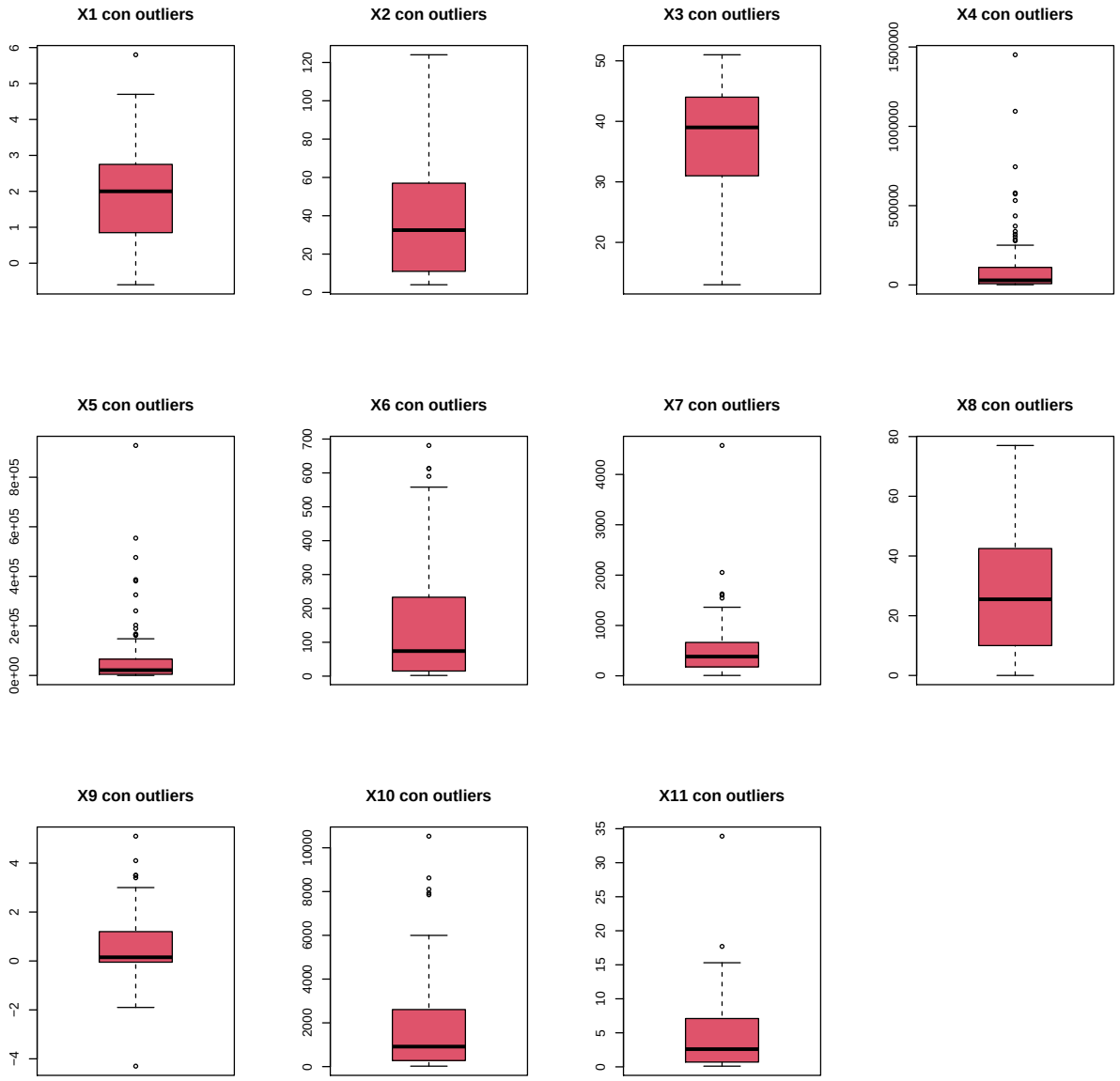
1.1 Descargar datos

```
library(readxl)
setwd("C:/Users/ASUS/Desktop/Universidad/R_LaSalle")
Datos<- read_excel('Base punto 1.xlsx')
```

1.2 Eliminamos los datos atípicos

Se realiza un análisis gráfico por medio de diagramas de caja para detectar datos atípicos o outliers, los cuales serán eliminados .

```
par(mfrow=c(3,4))
boxplot(Datos$`X 1`, main = "X1 con outliers", col=2)
boxplot(Datos$`X 2`, main = "X2 con outliers", col=2)
boxplot(Datos$`X 3`, main = "X3 con outliers", col=2)
boxplot(Datos$`X 4`, main = "X4 con outliers", col=2)
boxplot(Datos$`X 5`, main = "X5 con outliers", col=2)
boxplot(Datos$`X 6`, main = "X6 con outliers", col=2)
boxplot(Datos$`X 7`, main = "X7 con outliers", col=2)
boxplot(Datos$`X 8`, main = "X8 con outliers", col=2)
boxplot(Datos$`X 9`, main = "X9 con outliers", col=2)
boxplot(Datos$`X 10`, main = "X10 con outliers", col=2)
boxplot(Datos$`X 11`, main = "X11 con outliers", col=2)
```



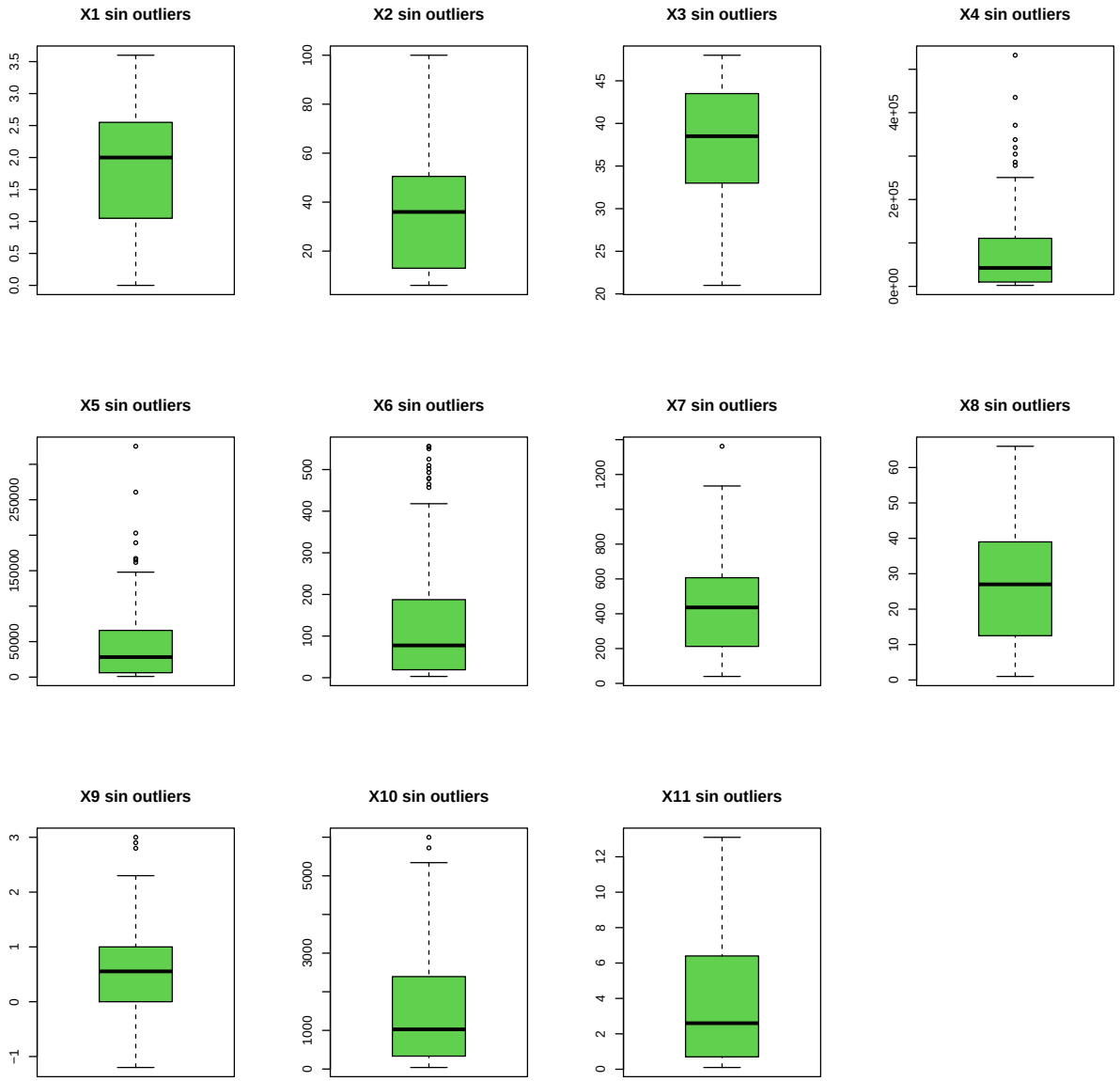
Luego, definimos una función para eliminar todos los datos atíticos que no se encuentren entre el cuantil 0,5 y 0,95

```
impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
```

La cual aplicaremos a cada variable desde x1 hasta x11, obteniendo una base de datos sin valores atípicos.

```
Datos$`X 1` <- impute_outliers(Datos$`X 1`,)
Datos$`X 2` <- impute_outliers(Datos$`X 2`,)
Datos$`X 3` <- impute_outliers(Datos$`X 3`,)
Datos$`X 4` <- impute_outliers(Datos$`X 4`,)
Datos$`X 5` <- impute_outliers(Datos$`X 5`,)
Datos$`X 6` <- impute_outliers(Datos$`X 6`,)
Datos$`X 7` <- impute_outliers(Datos$`X 7`,)
Datos$`X 8` <- impute_outliers(Datos$`X 8`,)
Datos$`X 9` <- impute_outliers(Datos$`X 9`,)
Datos$`X 10` <- impute_outliers(Datos$`X 10`,)
Datos$`X 11` <- impute_outliers(Datos$`X 11`,)
```

De esta forma, al volver a graficar los diagramas de caja después de eliminar los datos atípicos tenemos que:



1.3 Estadística descriptiva

Tabla generada mediante:

library(stargazer) stargazer(Datos, type = "text")

Table 1: Estadistica descriptiva

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
X 1	96	1.871	1.240	−1	0.9	2.7	6
X 2	96	39.062	32.138	4	11	56.5	124
X 3	96	37.281	8.774	13	31	44	51
X 4	96	116,587.400	223,602.000	1,353	8,014.2	109,221.8	1,451,051
X 5	96	69,261.320	134,978.100	6	4,675.8	65,783.5	928,083
X 6	96	165.125	195.991	2	15.5	231.5	681
X 7	96	509.844	574.629	7	178.5	662	4,575
X 8	96	27.333	20.025	0	10	42.2	77
X 9	96	0.553	1.348	−4	−0.03	1.2	5
X 10	96	1,854.427	2,239.285	20	287.8	2,553.5	10,531
X 11	96	4.554	5.222	0.100	0.700	7.050	33.900

.

1.4 Análisis de correlación

MEdiante un cuadro de correlación analizaremos las varibales con mayor coeficiente de correlación

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(dplyr)
```

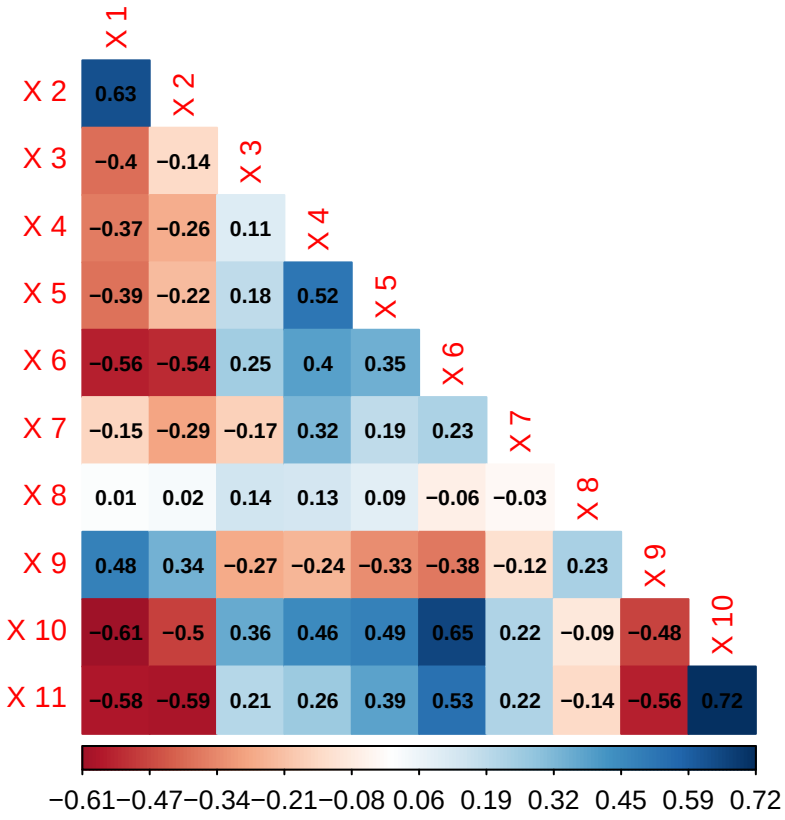
```
##
## Attaching package: 'dplyr'

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

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

```
Datos <- Datos %>%
  select("X 1":"X 11")
corrplot::corrplot(cor(Datos),
                   method="color",
                   type="lower",
                   number.cex=0.7,
```

```
                addCoef.col = "black",
                tl.col="red",
                tl.srt=90,
                tl.cex = 0.9,
                diag=FALSE,
                is.corr = F)
```



Mediante esta matriz, seleccionamos unicamente las varibales que se correlacionen a un grado igual o mayor de 0,5 en valor absoluto con 3 variables o más.

De este modo sólo quedan las variables:

X1, X2, X6, X10, X11

```
Datos<- Datos %>%
  select("X 1", "X 2","X 6", "X 10", "X 11")
```

1.5 Test de barlett

```
library(psych)
cortest.bartlett(Datos,n=96)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 239.2585
##
## $p.value
## [1] 9.804989e-46
##
## $df
## [1] 10
```

Note que el p valor es menor al 0.05. Por tanto se rechaza a favor de la hipotesis alternativa. De este modo al menos dos varianzas son diferentes.

1.6 Gráfico de sedimentaci+on y varianza acumulada

```
cpe<-prcomp(Datos,scale=T)
names(cpe)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

1.6.1 GRáfico de sedimentación

```
library("factoextra")
```

```
## Warning: package 'factoextra' was built under R version 4.0.5
```

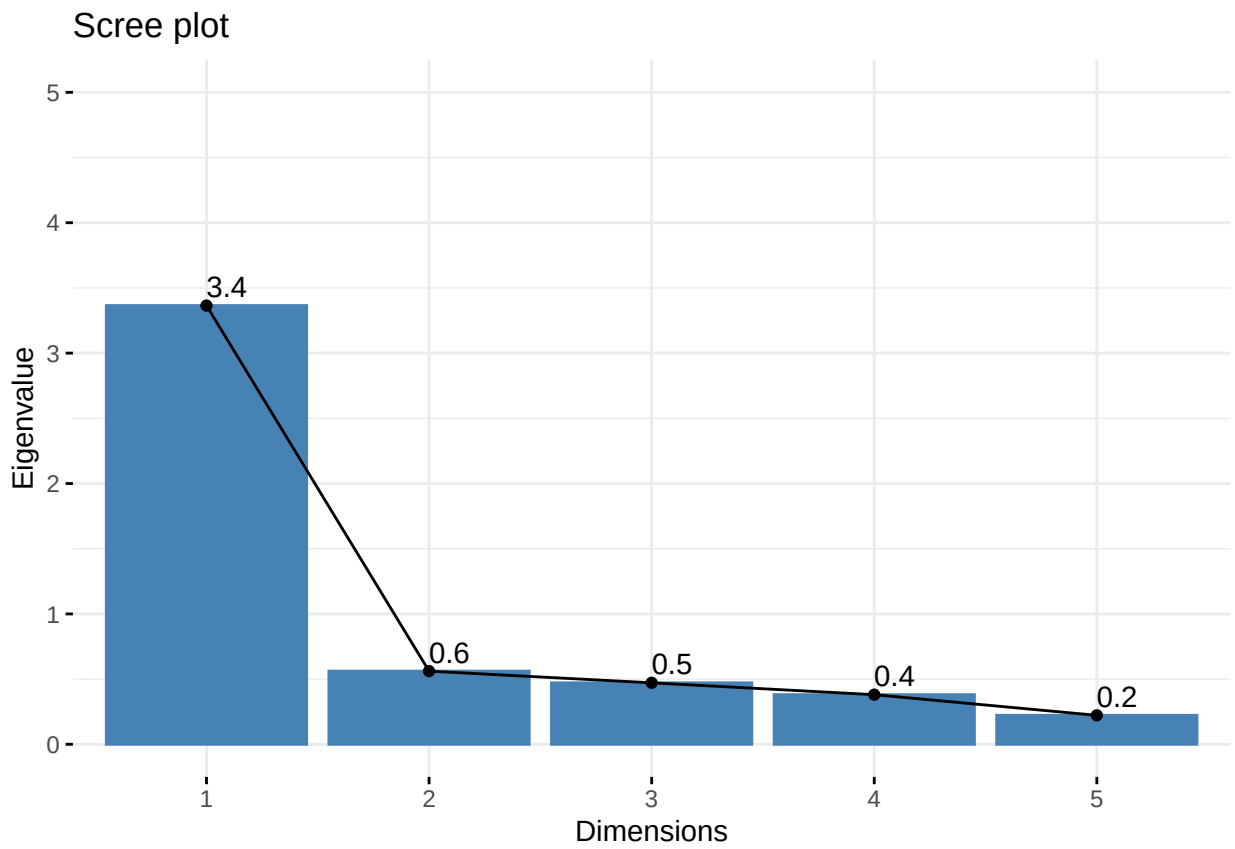
```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

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

```
fviz_eig(cpe,choice = "eigenvalue", addlabels = TRUE, axes = 1,ylim = c(0,5))
```



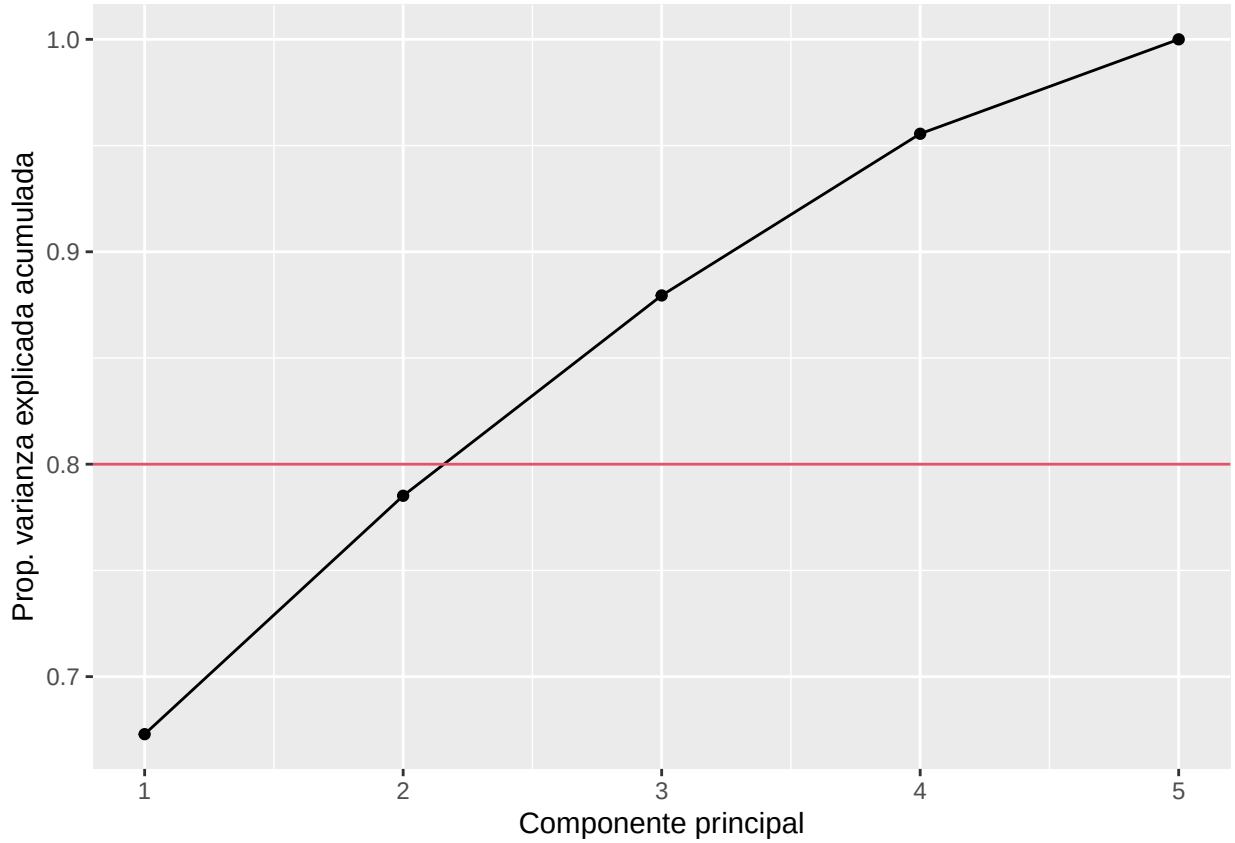
Note que únicamente 1 barra es mayor a 1. Por tanto, me recomienda analizar 1 solo componente. Esto con la consideración que igualmente tendremos el 68% de la información.

1.6.2 Varianza acumulada

```
library(ggplot2)
prop_varianza <- cpe$sdev^2 / sum(cpe$sdev^2);prop_varianza*100
```

```
## [1] 67.292575 11.221694  9.429866  7.611452  4.444413
```

```
prop_varianza_acum <- cumsum(prop_varianza)
ggplot(data = data.frame(prop_varianza_acum, pc=1:5),
       aes(x = pc, y = prop_varianza_acum, group = 1)) +
  geom_point() +
  geom_line() +
  labs(x = "Componente principal",
       y = "Prop. varianza explicada acumulada")+geom_hline(yintercept=0.8,col=2)
```

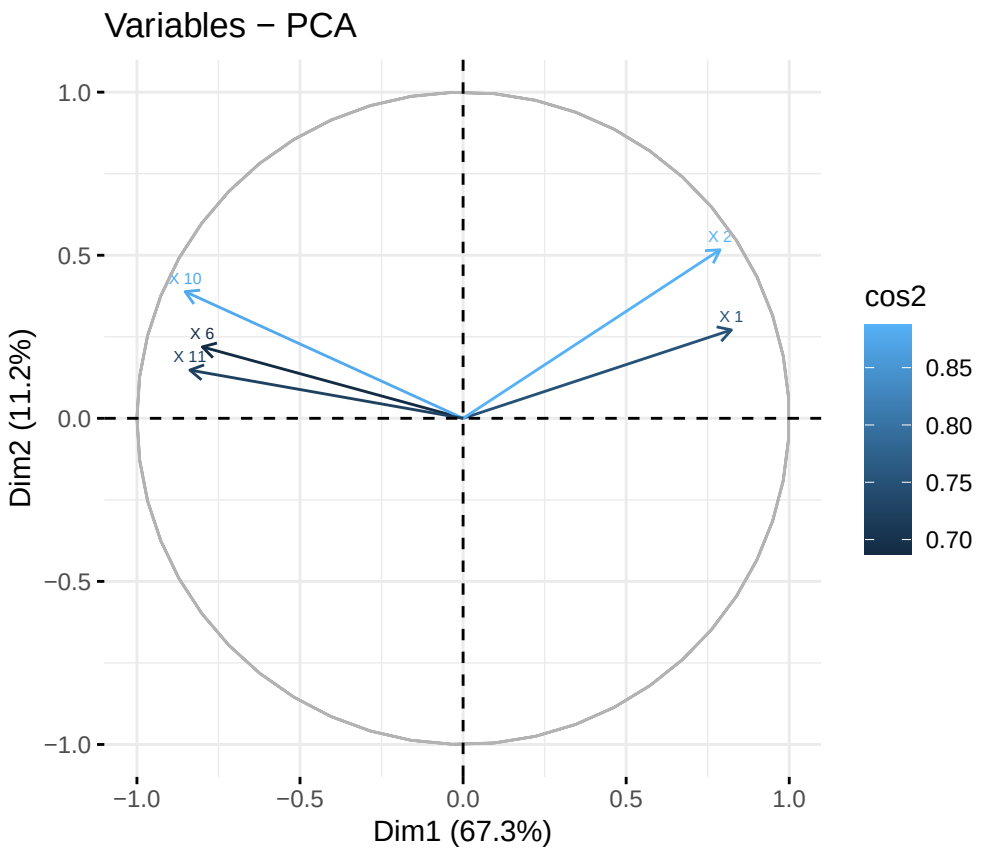


Este criterio me recomienda usar 3 componentes, ya que son los que superan el 0.8 en la gráfica de varianza acumulada.

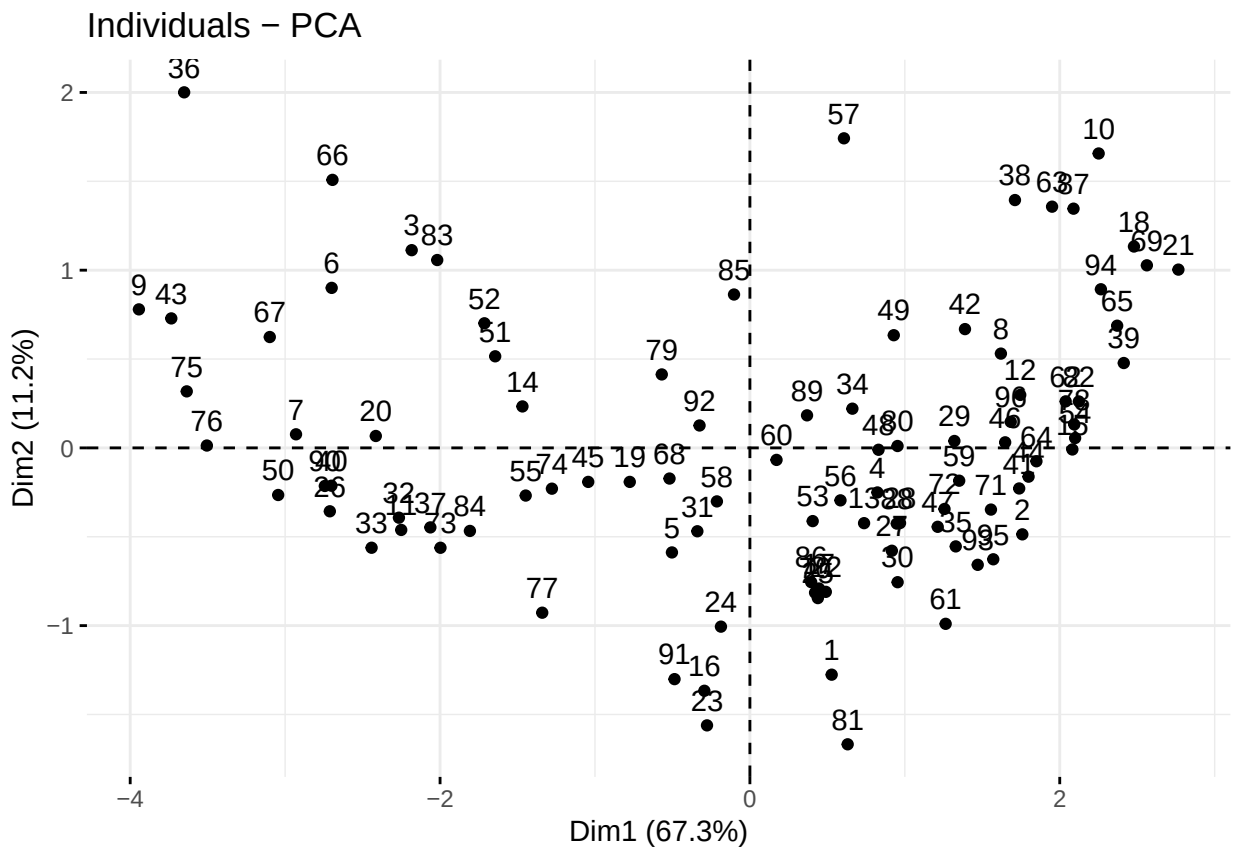
1.7 Identificación de los individuos y creación del valor del indice para cada individuo

Encontramos que X1 y X2 están correlacionadas y X6, X10 y X11 están correlacioandas igualmente en dirección diferente. Graficamente podríamos sugerir que X2 y X10 tienen mayor poder explicativo.

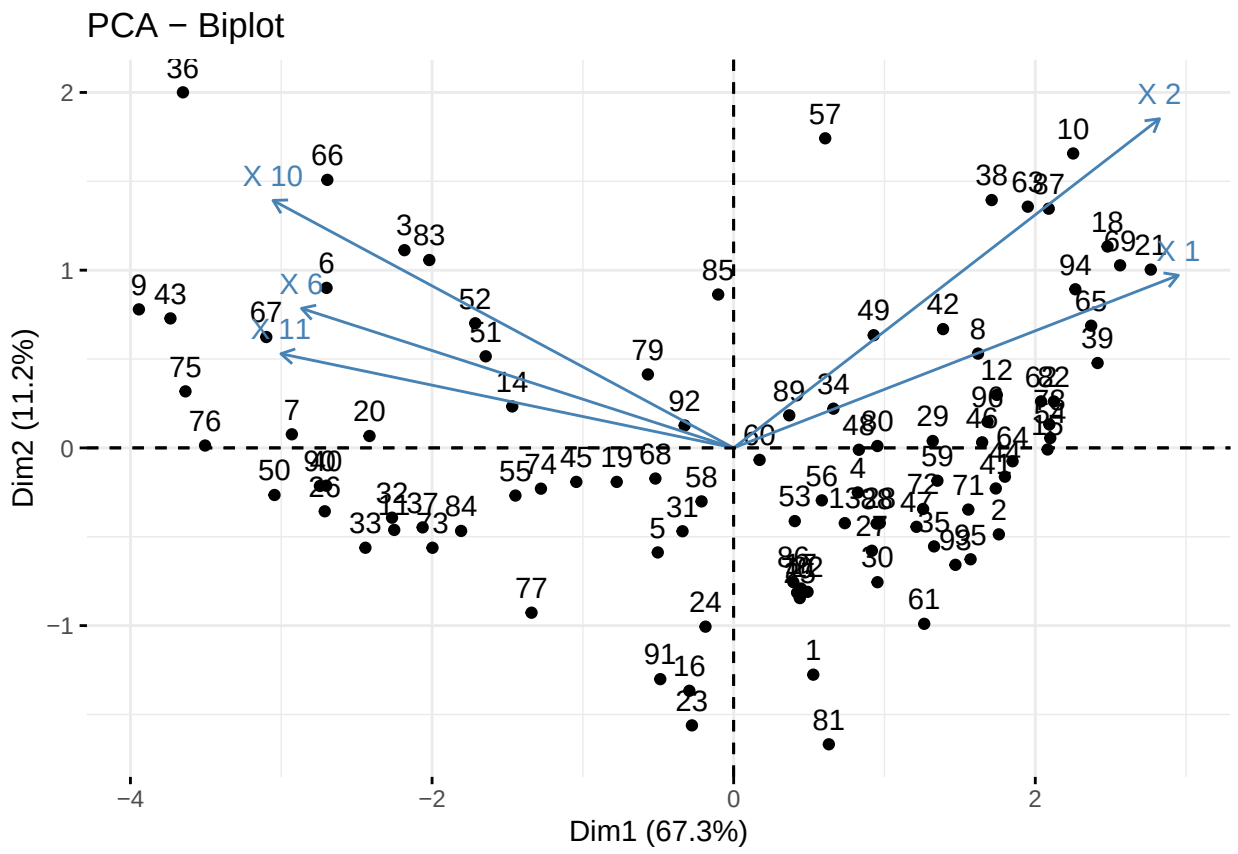
```
fviz_pca_var(cpe, col.var = "cos2",
             geom.var = c("arrow", "text"),
             labelsize = 2,
             repel = FALSE)
```




```
fviz_pca_ind(cpe,addEllipses=F)
```




```
fviz_pca_biplot(cpe)
```



1.8 Especificación de modelos para explicar X1 y X2

Para esto vamos a hacer un modelo lineal e interpretaremos si los coeficientes son significativos al 5% de confianza

1.8.1 X1

Para esto tomaremos las varibales que según la matriz de correlación estuvieron por encima del 50% en valor absoluto en el coeficiente de correlación. Estas son: X2, X6, X10 y X11

```
Reg1 <- lm(`X 1` ~ `X 2`+`X 6`+ `X 10` + `X 11`, data = Datos)
```

1.8.2 X2

Para esto tomaremos las varibales que según la matriz de correlación estuvieron por encima del 50% en valor absoluto en el coeficiente de correlación. Estas son: X2, X6, X10 y X11

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
Reg2 <- lm(`X 2` ~ `X 1`+`X 6`+ `X 10` + `X 11`, data = Datos)
stargazer(Reg1, Reg2, type = "text", digits=3, omit.stat = c("f", 'ser'))
```

```
##
## ==========================================
##                   Dependent variable:
##               ---------------------------
##                   `X 1`          `X 2`
##                    (1)            (2)
## ------------------------------------------
## `X 2`            0.013***
##                  (0.003)
##
## `X 1`                          10.577***
##                                 (2.741)
##
## `X 6`             -0.001       -0.035**
##                  (0.001)        (0.016)
##
## `X 10`          -0.0002**        0.002
##                  (0.0001)       (0.002)
##
## `X 11`            -0.027       -2.478***
##                  (0.030)        (0.808)
##
## Constant         1.796***      28.303***
##                  (0.216)        (7.509)
##
## ------------------------------------------
## Observations       96             96
## R2                0.525          0.499
## Adjusted R2       0.504          0.477
## ==========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Note que para X1, las variables X6 y X11 no son significativas y para X2 la variable X10 no es significativa.