

# Principal Component Analysis (PCA)

Pablo E. Gutiérrez-Fonseca

8/26/2021

1. Primer paso: cargar las librerías que necesitas.

```
library(ggplot2)
library(dplyr)
library(missMDA) # Impute
library(ggfortify) # autoplot()
library(cluster) #pam
library(factoextra) #get_pca_var()
library(data.table) # data.table()
library(labdsv) #loadings.pca(pca)

library(devtools)

install_github("vqv/ggbiplot") #ggbiplot
library(ggbiplot)
```

2. Segundo paso: cargar los datos.

```
channel <- read.csv("data/channel_form.csv", header=TRUE)
#head(channel)
```

## 2.1 Vamos a examinar los datos

```
summary(channel)
```

```
##      Forma      NAN_Am      NADBO      NAtemp
## Length:138    Min.   :0.0200    Min.    : 1.310    Min.     :14.67
## Class :character 1st Qu.:0.0400    1st Qu.: 1.930    1st Qu.:24.30
## Mode  :character Median :0.2150    Median : 3.000    Median :26.05
##              Mean  :0.3201    Mean   : 6.164    Mean    :25.84
##              3rd Qu.:0.5000    3rd Qu.: 8.585    3rd Qu.:27.70
##              Max.   :1.5000    Max.    :34.900    Max.     :32.18
##              NA's    :35
##      nit      NASat02      Elevacion      Ancho
## Min.   : 0.00    Min.    : 23.43    Min.     : 3.00    Min.     : 1.000
## 1st Qu.: 0.40    1st Qu.: 86.24    1st Qu.: 25.25    1st Qu.: 2.000
## Median : 0.92    Median : 94.59    Median : 53.00    Median : 3.000
## Mean   :12.00    Mean   : 91.05    Mean    :230.89    Mean    : 3.875
## 3rd Qu.: 1.62    3rd Qu.:100.52    3rd Qu.:269.25    3rd Qu.: 3.000
## Max.   :324.11    Max.    :122.73    Max.     :2370.00    Max.     :16.000
```

```
## NA's :57 NA's :2
## Velocidad Rocas Canto grava
## Min. : 0.000 Min. : 0.00 Min. : 0.00 Min. : 0.00
## 1st Qu.: 3.000 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 3.75
## Median :11.000 Median :10.00 Median :25.00 Median :20.00
## Mean : 9.169 Mean :16.27 Mean :25.46 Mean :17.86
## 3rd Qu.:14.000 3rd Qu.:30.00 3rd Qu.:40.00 3rd Qu.:25.00
## Max. :16.000 Max. :90.00 Max. :80.00 Max. :80.00
## NA's :2 NA's :2 NA's :2 NA's :2
## arena Limo
## Min. : 0.00 Min. : 0.00
## 1st Qu.: 10.00 1st Qu.: 0.00
## Median : 15.00 Median : 7.50
## Mean : 19.83 Mean : 20.51
## 3rd Qu.: 25.00 3rd Qu.: 25.00
## Max. :100.00 Max. :100.00
## NA's :2 NA's :2
```

2.1 Remover la(s) variable(s) que tiene(n) mucho(s) NAs y las Etiquetas (a la funcion lo le gusta), luego las agregamos.

```
channel_1 <- select(channel, -Forma)
summary(channel_1)
```

```
## NAN_Am NADBO NAtemp nit
## Min. :0.0200 Min. : 1.310 Min. :14.67 Min. : 0.00
## 1st Qu.:0.0400 1st Qu.: 1.930 1st Qu.:24.30 1st Qu.: 0.40
## Median :0.2150 Median : 3.000 Median :26.05 Median : 0.92
## Mean :0.3201 Mean : 6.164 Mean :25.84 Mean :12.00
## 3rd Qu.:0.5000 3rd Qu.: 8.585 3rd Qu.:27.70 3rd Qu.: 1.62
## Max. :1.5000 Max. :34.900 Max. :32.18 Max. :324.11
## NA's :35 NA's :57
## NASat02 Elevacion Ancho Velocidad
## Min. : 23.43 Min. : 3.00 Min. : 1.000 Min. : 0.000
## 1st Qu.: 86.24 1st Qu.: 25.25 1st Qu.: 2.000 1st Qu.: 3.000
## Median : 94.59 Median : 53.00 Median : 3.000 Median :11.000
## Mean : 91.05 Mean : 230.89 Mean : 3.875 Mean : 9.169
## 3rd Qu.:100.52 3rd Qu.: 269.25 3rd Qu.: 3.000 3rd Qu.:14.000
## Max. :122.73 Max. :2370.00 Max. :16.000 Max. :16.000
## NA's :2 NA's :2
## Rocas Canto grava arena
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 3.75 1st Qu.: 10.00
## Median :10.00 Median :25.00 Median :20.00 Median : 15.00
## Mean :16.27 Mean :25.46 Mean :17.86 Mean : 19.83
## 3rd Qu.:30.00 3rd Qu.:40.00 3rd Qu.:25.00 3rd Qu.: 25.00
## Max. :90.00 Max. :80.00 Max. :80.00 Max. :100.00
## NA's :2 NA's :2 NA's :2 NA's :2
## Limo
## Min. : 0.00
## 1st Qu.: 0.00
## Median : 7.50
## Mean : 20.51
```

```
## 3rd Qu.: 25.00
## Max.    :100.00
## NA's    :2
```

2.2 Vamos a imputar datos. Esto es comun para set de datos de campo, los cuales tienden a tener ceros (por mal funcionamiento de los equipos, condiciones climáticas adversas que no podemos ir al campo). Se realiza como un paso preliminar para para realizar un PCA en un set de datos completos.

Mas informacion aca: <https://www.rdocumentation.org/packages/missMDA/versions/1.18/topics/imputePCA>

Primero separar e imputar los datos de sustrato y los fisicoquimicos por aparte.

```
# df0 <- channel_1[, 6:13]
# df0

df1 <- select(channel_1, Elevacion, Ancho, Velocidad, Rocas, Canto, grava, arena, Limo)

df1a <- imputePCA(df1, ncp=4, scale = TRUE, method = c("Regularized","EM"),
  row.w = NULL, ind.sup=NULL, quanti.sup=NULL, quali.sup=NULL,
  coeff.ridge = 1, threshold = 1e-06, seed = NULL, nb.init = 1,
  maxiter = 1000)

df2 <- select(channel_1, Elevacion, NAN_Am, NAtemp, NASatO2, nit, NADBO)
df2a <- imputePCA(df2, ncp=4, scale = TRUE, method = c("Regularized","EM"),
  row.w = NULL, ind.sup=NULL, quanti.sup=NULL, quali.sup=NULL,
  coeff.ridge = 1, threshold = 1e-06, seed = NULL, nb.init = 1,
  maxiter = 1000)
```

Unir las dos tablas y seleccionar las columnas para hacer el PCA.

```
df1b <- as.data.frame(df1a) # Sustrata
df2b <- as.data.frame(df2a) # Physicochemical

new_channel <- do.call("merge", c(lapply(list(df1b, df2b), data.frame, row.names=NULL),
  by = 0, all = TRUE, sort = FALSE))[-1]

new_channel2 <- select(new_channel,
  completeObs.Elevacion.x, completeObs.Ancho, completeObs.Velocidad,
  completeObs.Rocas, completeObs.Canto, completeObs.grava, completeObs.arena,
  completeObs.Limo, completeObs.NAtemp, completeObs.NASatO2
)
```

3. Vamos a correr el PCA

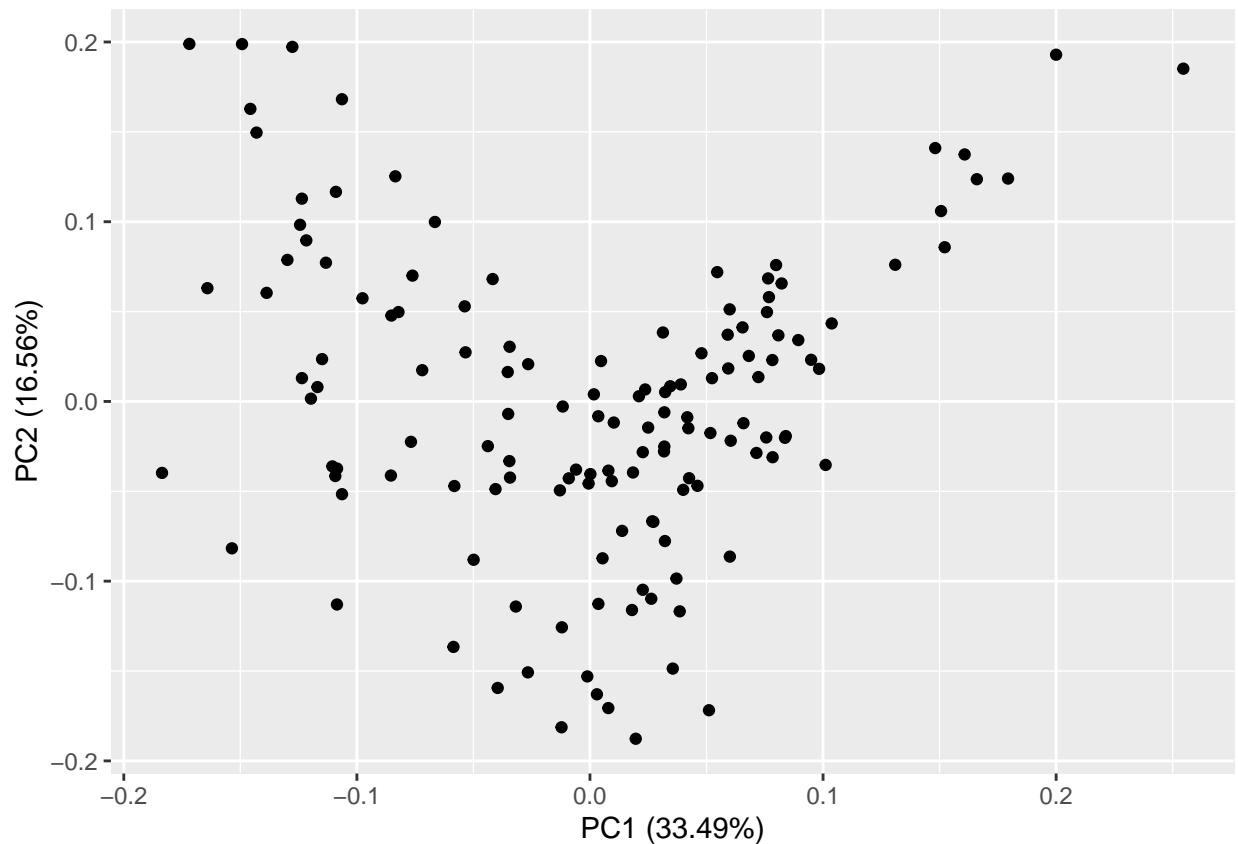
```
channel.pca <- prcomp(new_channel2, center = TRUE, scale =TRUE)
summary(channel.pca)
```

```
## Importance of components:
##              PC1      PC2      PC3      PC4      PC5      PC6      PC7
```

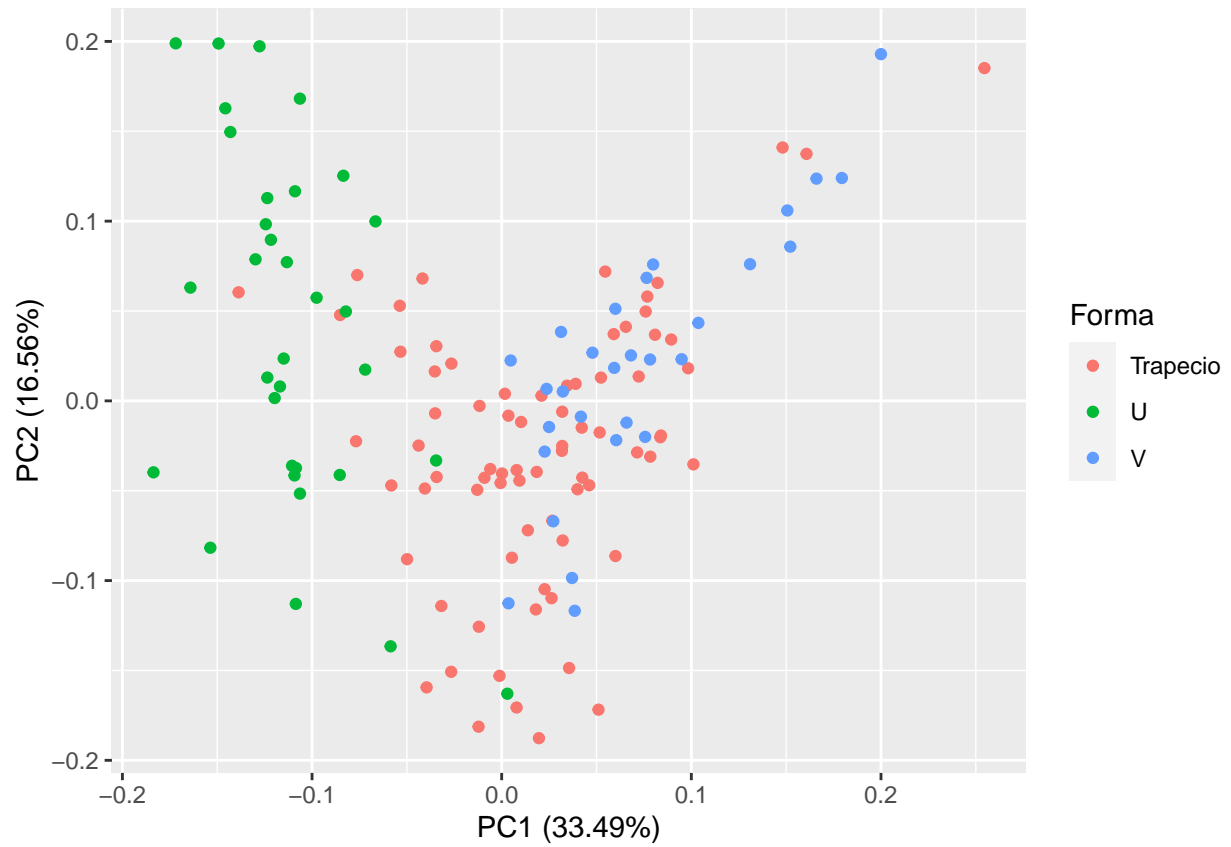
```
## Standard deviation      1.8299 1.2868 1.1895 1.0758 0.89995 0.76399 0.70801
## Proportion of Variance 0.3349 0.1656 0.1415 0.1157 0.08099 0.05837 0.05013
## Cumulative Proportion 0.3349 0.5005 0.6419 0.7577 0.83867 0.89704 0.94717
##                          PC8      PC9      PC10
## Standard deviation      0.62231 0.37269 0.04612
## Proportion of Variance 0.03873 0.01389 0.00021
## Cumulative Proportion 0.98590 0.99979 1.00000
```

3.1 Vamos a ver el grafico.

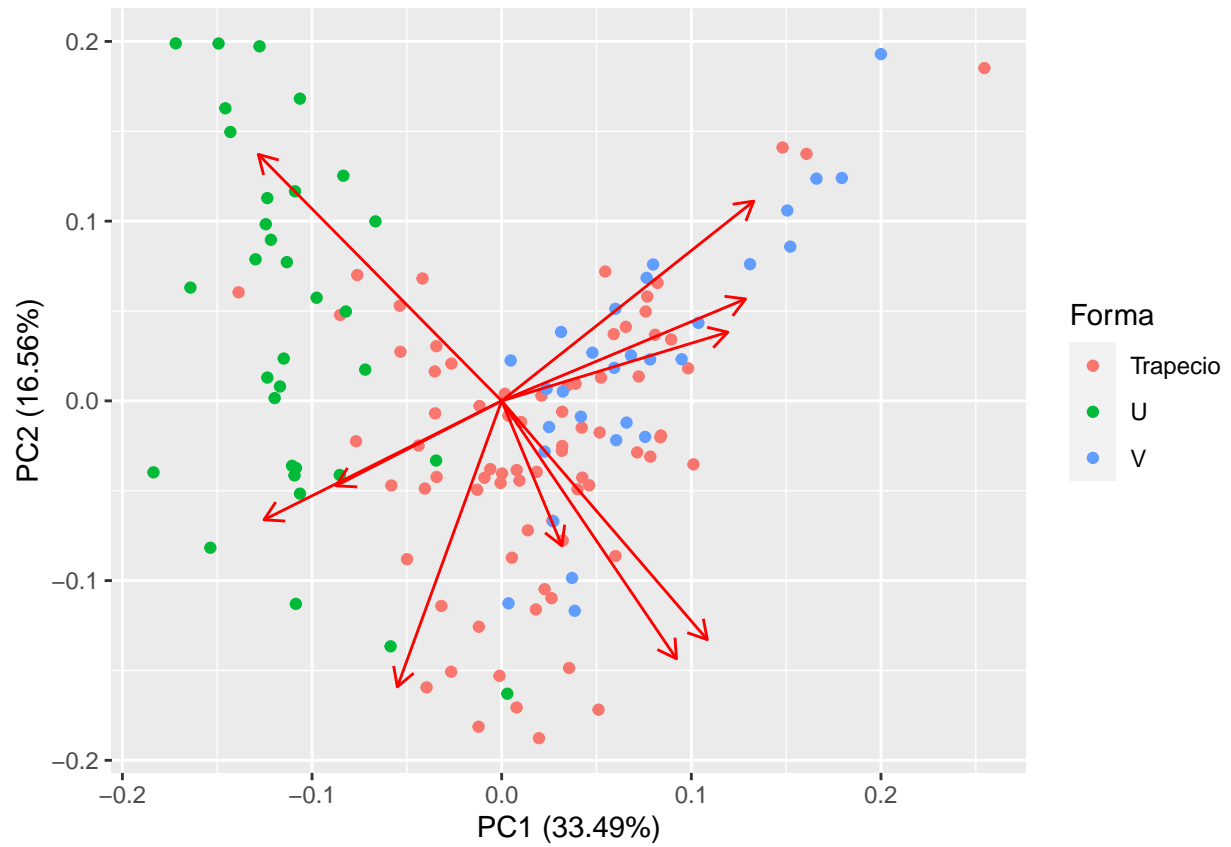
```
autoplot(channel.pca)
```



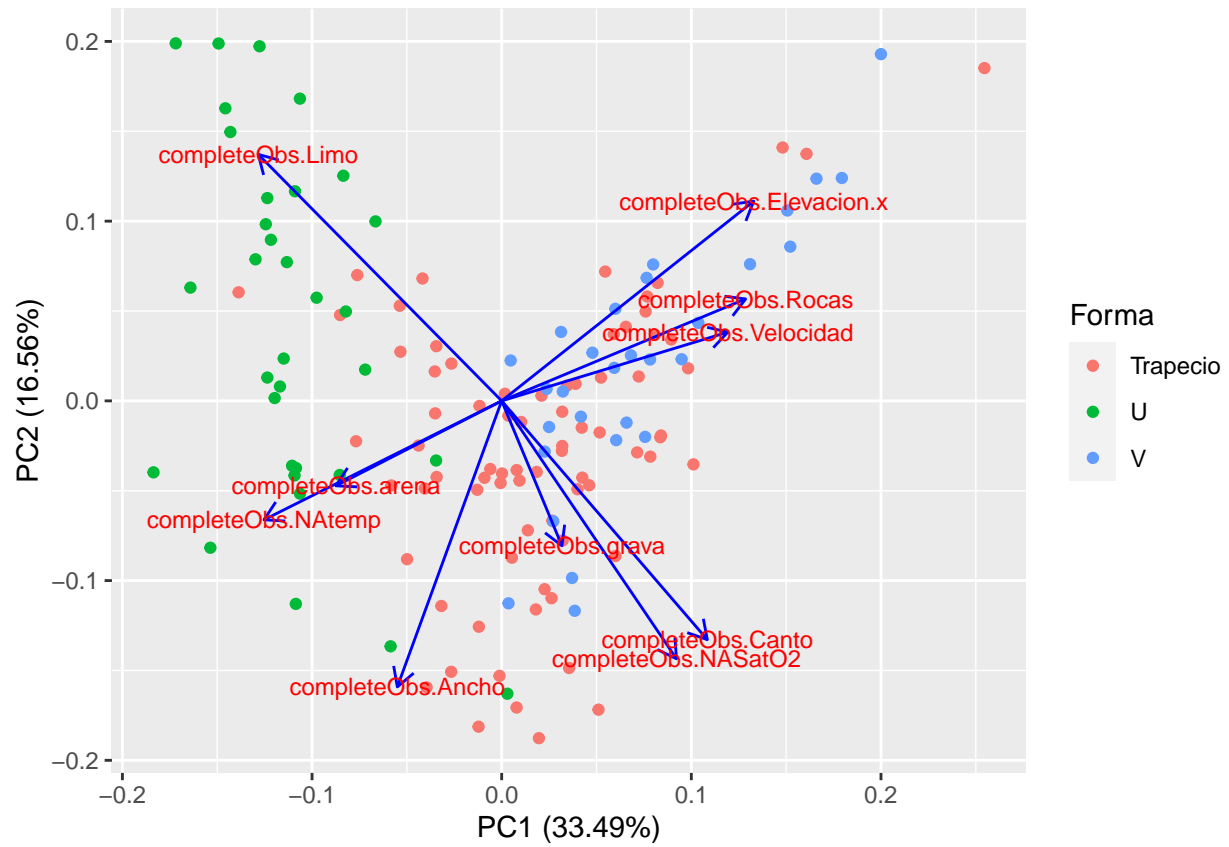
```
autoplot(channel.pca, data = channel, colour = 'Forma')
```



```
autoplot(channel.pca, data = channel, colour = 'Forma', loadings = TRUE)
```

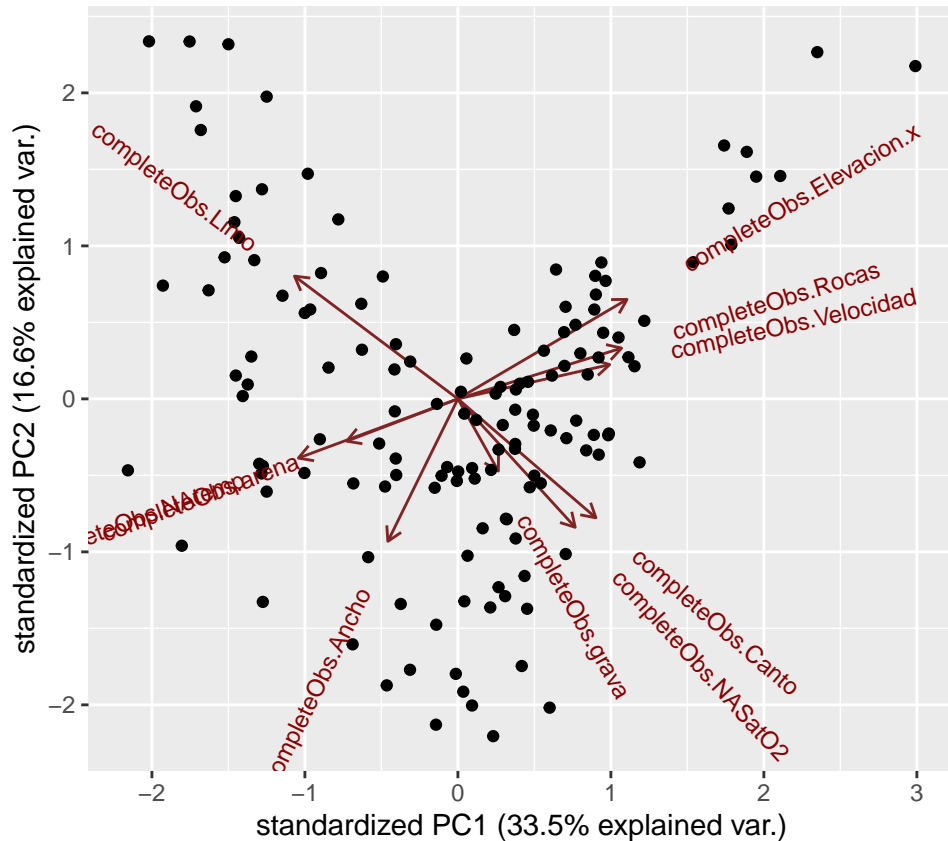


```
autoplot(channel.pca, data = channel, colour = 'Forma', loadings = TRUE,
         loadings.colour = 'blue',
         loadings.label = TRUE, loadings.label.size = 3)
```



Otra manera de ver el grafico

```
ggbiplot(channel.pca, labels=rownames(channel$Forma))
```



3.2 Vamos a ver la contribucion de cada una de las variables.

```
variance <- (channel.pca$sdev)^2
```

```
# Cargar los loadings
loadings <- channel.pca$rotation
round(loadings, 2)[ , 1:3]
```

```
##          PC1   PC2   PC3
## completeObs.Elevacion.x  0.40  0.33 -0.19
## completeObs.Ancho       -0.16 -0.47 -0.44
## completeObs.Velocidad   0.36  0.11  0.41
## completeObs.Rocas       0.38  0.17 -0.18
## completeObs.Canto       0.32 -0.40 -0.10
## completeObs.grava       0.09 -0.24  0.63
## completeObs.arena      -0.26 -0.14 -0.11
## completeObs.Limo       -0.38  0.41 -0.10
## completeObs.NAtemp     -0.37 -0.20  0.36
## completeObs.NASatO2     0.27 -0.43 -0.12
```

```
print(channel.pca)
```

```
## Standard deviations (1, ..., p=10):
## [1] 1.82992315 1.28683255 1.18946223 1.07584685 0.89995265 0.76399277
## [7] 0.70801267 0.62231167 0.37268728 0.04612058
```



```
##
## Rotation (n x k) = (10 x 10):
##
```

	PC1	PC2	PC3	PC4
## completeObs.Elevacion.x	0.39615124	0.3312904	-0.19490883	0.28422267
## completeObs.Ancho	-0.16405478	-0.4744099	-0.43650902	0.10139348
## completeObs.Velocidad	0.35548809	0.1140191	0.41306910	-0.00700442
## completeObs.Rocas	0.38339344	0.1690205	-0.18090196	-0.08759678
## completeObs.Canto	0.32322298	-0.3960706	-0.09665874	-0.34691126
## completeObs.grava	0.09466745	-0.2403600	0.62978697	0.31374507
## completeObs.arena	-0.26046769	-0.1408860	-0.10693686	0.67881136
## completeObs.Limo	-0.38269942	0.4088553	-0.09718531	-0.30305650
## completeObs.NAtemp	-0.37358516	-0.1972495	0.35775624	-0.32980922
## completeObs.NASat02	0.27474489	-0.4278434	-0.11521405	-0.14501367

	PC5	PC6	PC7	PC8
## completeObs.Elevacion.x	-0.24462464	0.0872553041	-0.184651840	0.130557822
## completeObs.Ancho	-0.29139434	0.2302480810	0.239558001	-0.561635255
## completeObs.Velocidad	0.03227524	-0.4860783571	0.124372396	-0.655162051
## completeObs.Rocas	0.52380817	0.4012932650	0.414132060	-0.077544103
## completeObs.Canto	-0.33332751	-0.3829257400	0.167112732	0.351218850
## completeObs.grava	-0.27257872	0.4743647187	-0.114860543	0.004670724
## completeObs.arena	0.33215022	-0.4081015160	0.009855948	0.087775229
## completeObs.Limo	-0.16583016	-0.0003924146	-0.351986218	-0.268395095
## completeObs.NAtemp	0.31146226	0.0456037467	0.178964077	0.098098723
## completeObs.NASat02	0.39949081	0.0422275691	-0.724156487	-0.139749138

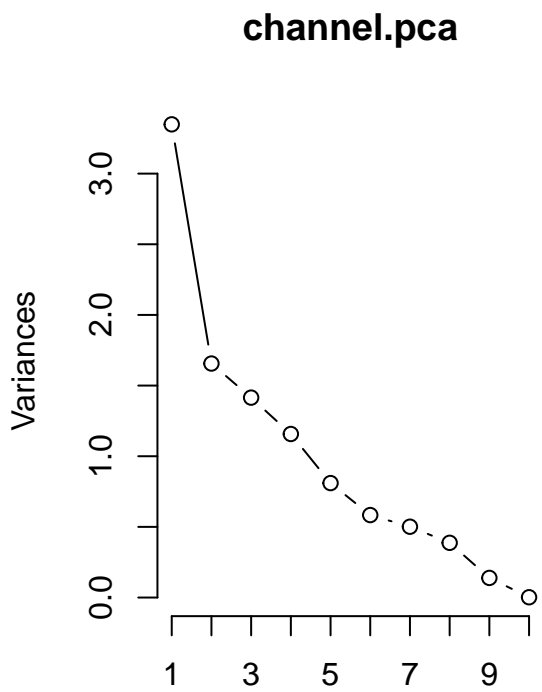
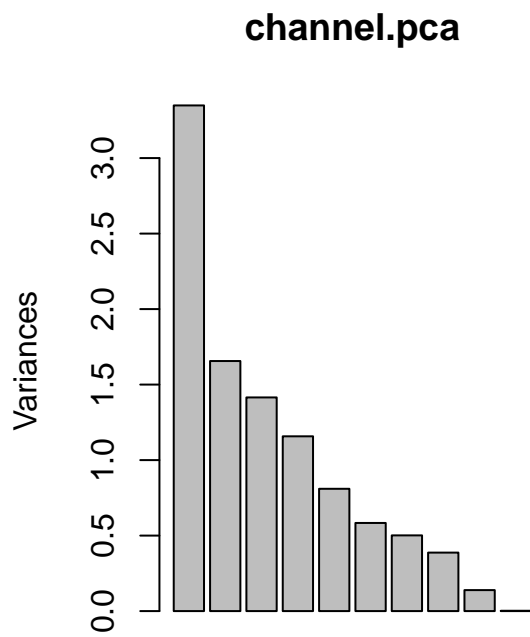
  

	PC9	PC10
## completeObs.Elevacion.x	0.70423149	-0.0007661306
## completeObs.Ancho	0.19091217	-0.0024515868
## completeObs.Velocidad	0.08905142	0.0017012657
## completeObs.Rocas	-0.05514873	-0.4099655074
## completeObs.Canto	0.02763742	-0.4462960258
## completeObs.grava	-0.07731897	-0.3459530412
## completeObs.arena	0.06106687	-0.3894687964
## completeObs.Limo	0.01761358	-0.6011239899
## completeObs.NAtemp	0.66634430	-0.0029056115
## completeObs.NASat02	0.04292839	0.0032717814

```
rownames(loadings) <- colnames(new_channel2)
scores <- channel.pca$x
```

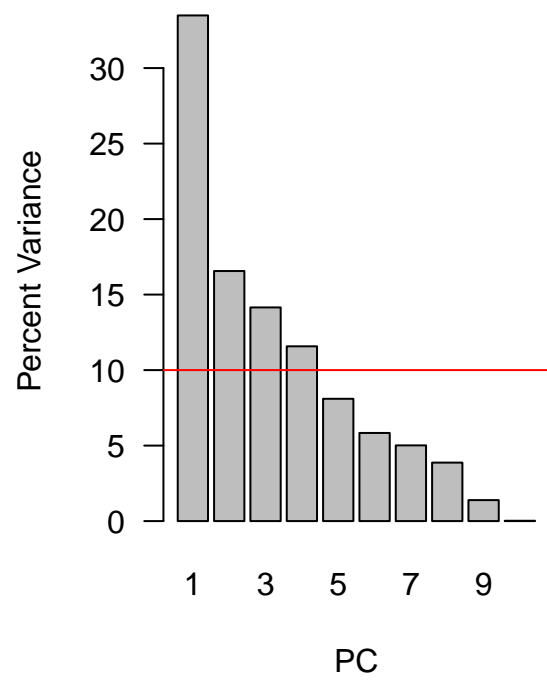
3.3 Ver graficamente lo que explica cada axis.

```
layout(matrix(1:2, ncol=2))
screeplot(channel.pca)
screeplot(channel.pca, type="lines")
```



```
varPercent <- variance/sum(variance) * 100
barplot(varPercent, xlab='PC', ylab='Percent Variance',
names.arg=1:length(varPercent), las=1, col='gray') +
abline(h=1/ncol(new_channel2)*100, col="red")
```

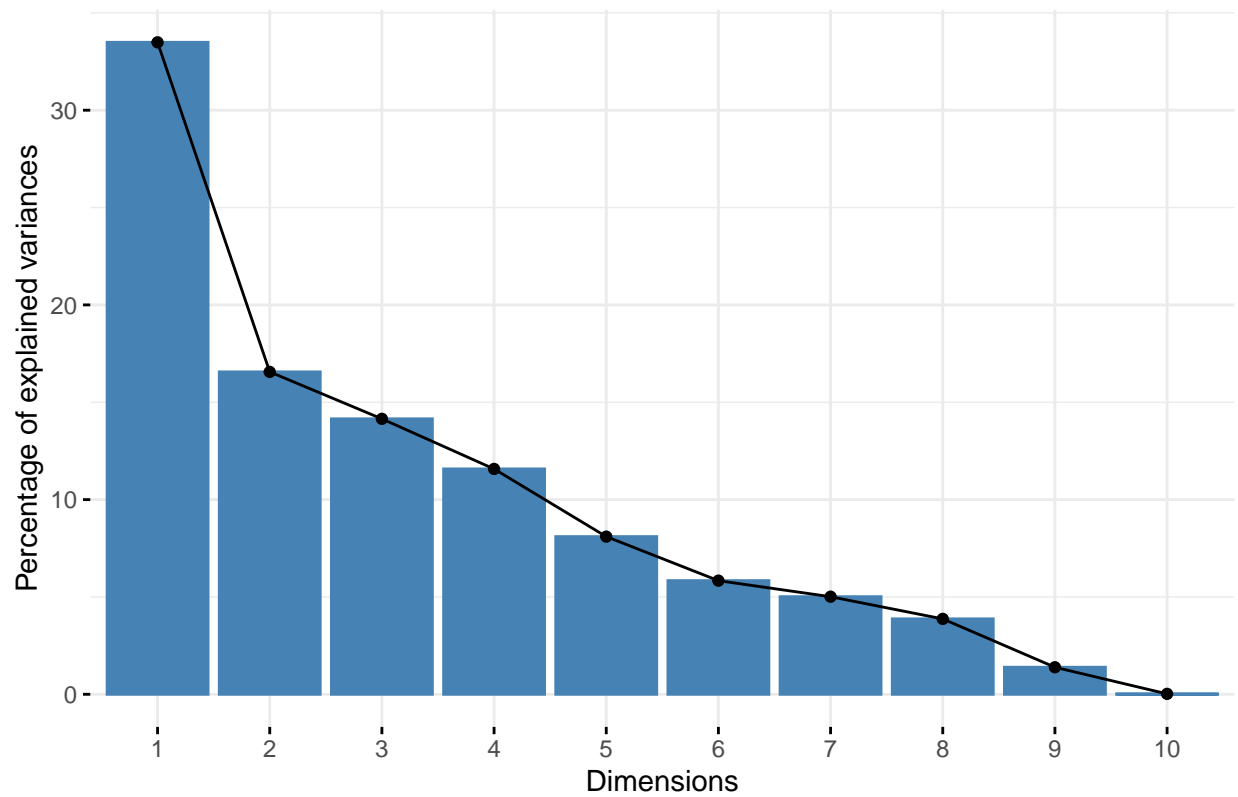
```
## numeric(0)
```



4 Otras formas de visualizar los datos.

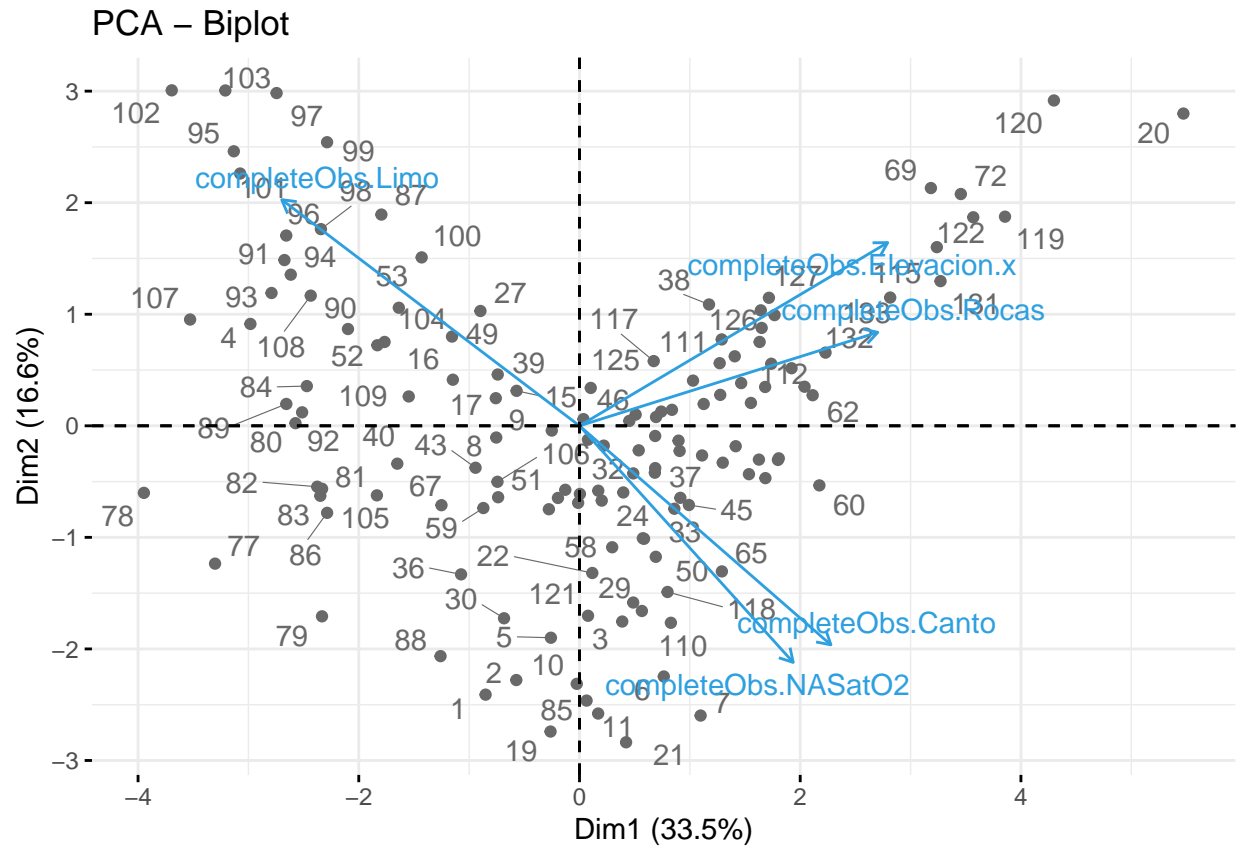
```
fviz_eig(channel.pca)
```

Scree plot



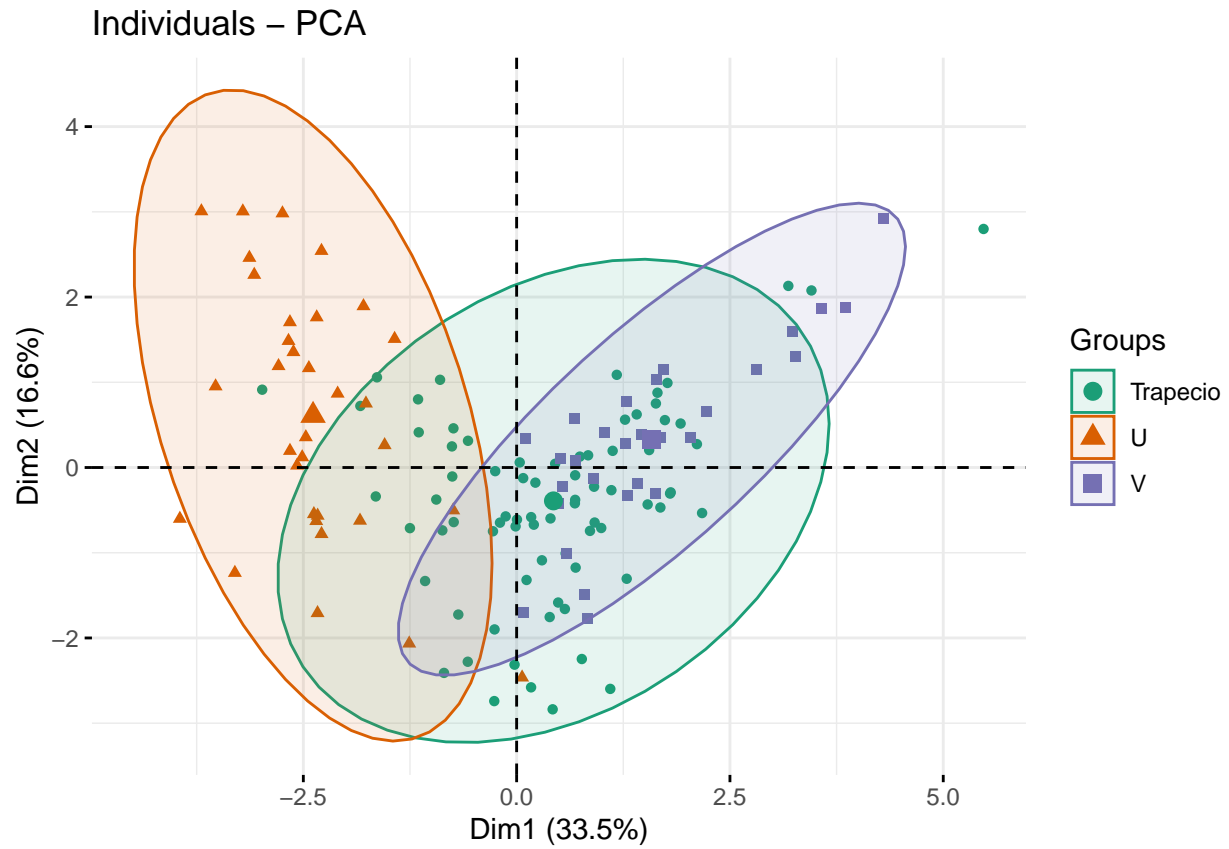
```
fviz_pca_biplot(channel.pca, repel = TRUE,  
  col.var = "#2E9FDF", # Variables color  
  col.ind = "#696969", # Individuals color  
  select.var = list(contrib = 5))
```

```
## Warning: ggrepel: 44 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps
```



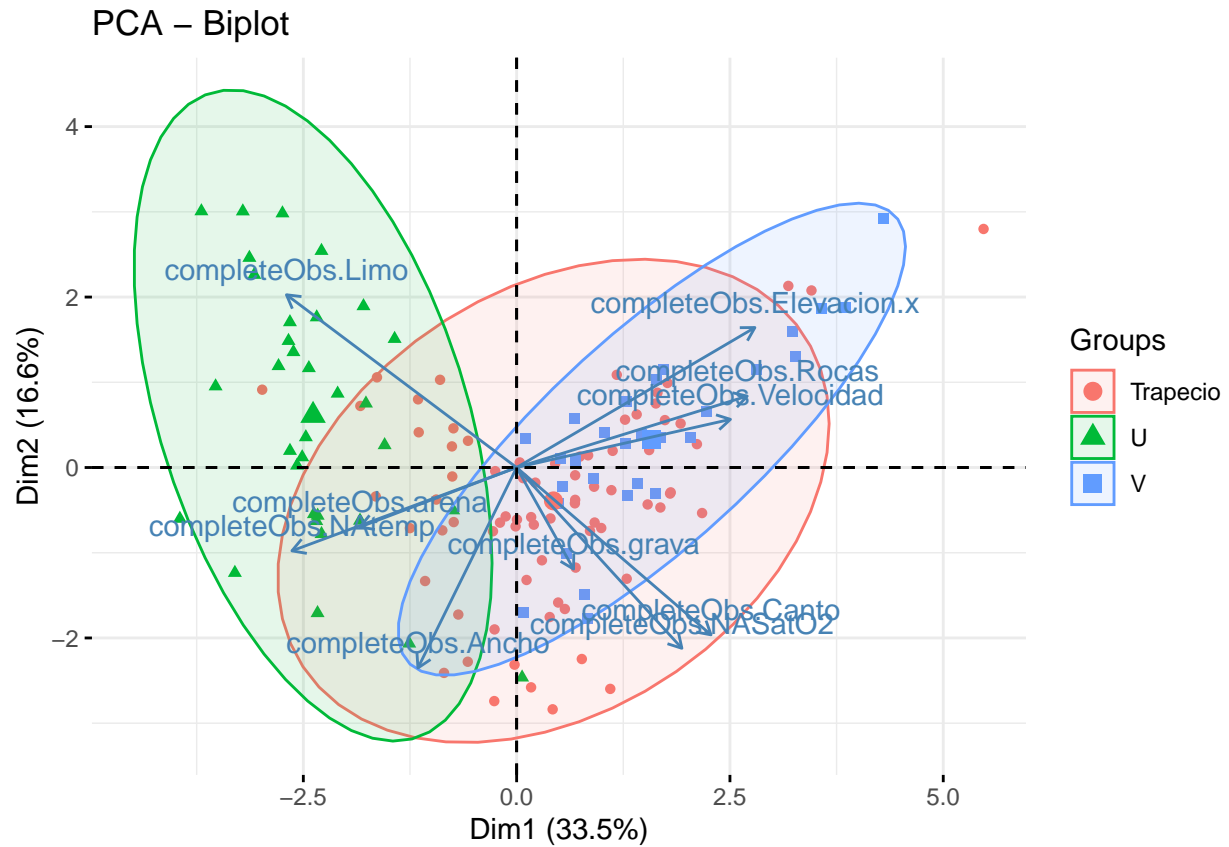
4.1 Con las elipses.

```
fviz_pca_ind(channel.pca, label="none", habillage=channel$Forma,
  addEllipses=TRUE, ellipse.level=0.95, palette = "Dark2")
```



4.2

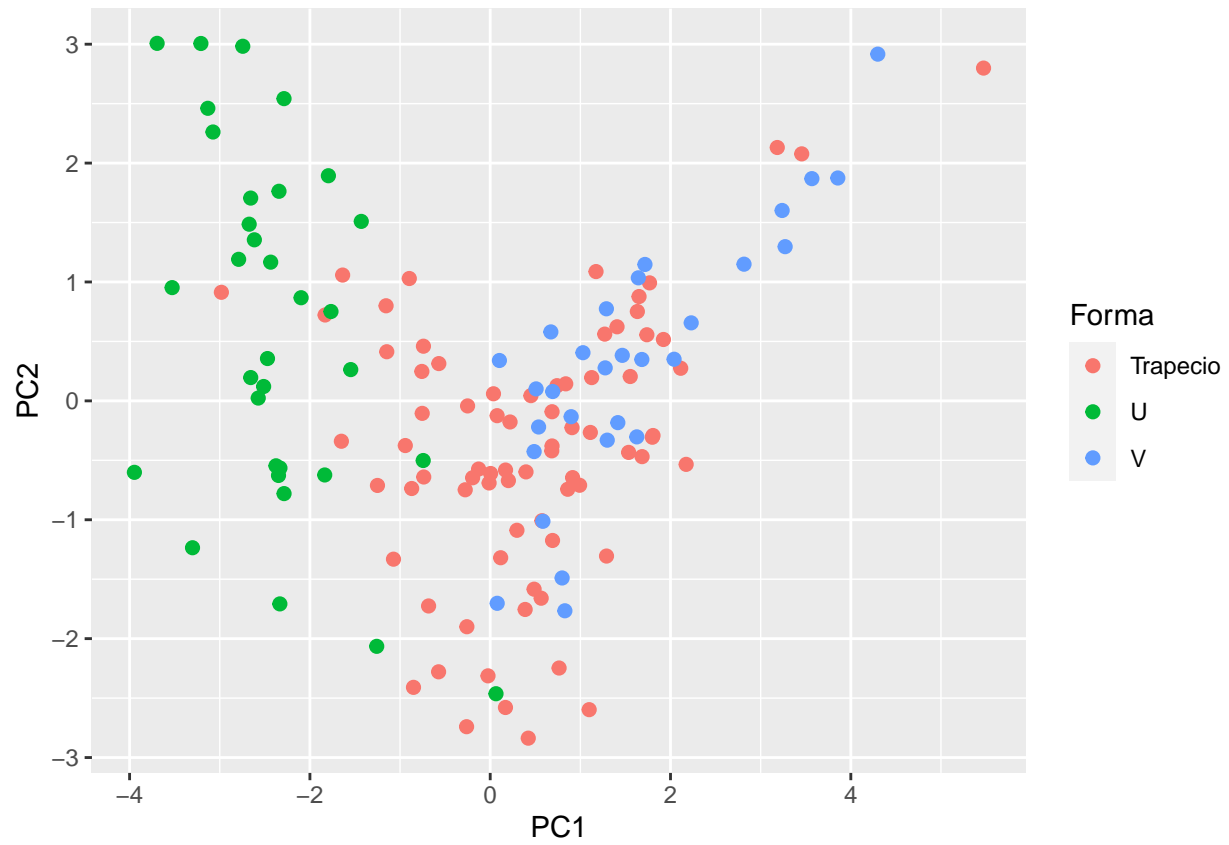
```
fviz_pca_biplot(channel.pca, label = "var", habillage=channel$Forma,
  addEllipses=TRUE, ellipse.level=0.95,
  ggtheme = theme_minimal())
```



5. Convertirlo en una data.frame para trabajarlo en ggplot2

```
data <- data.table(PC1=channel.pca$x[,1], PC2=channel.pca$x[,2], Forma= channel[,1])
data <- data[order(channel$Forma),]

ggplot(data, aes(x=PC1,y=PC2)) +
  geom_point(size = 2, aes(color=Forma))
```



6. Otras enlaces de interes.

# <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-princip>