

# 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) # Imputate
library(ggfortify) # autoplot()
library(cluster) #pam
library(factoextra) #get_pca_var()

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

2. Segundo paso: carga los datos.

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

```
##      Forma NAN_Am NADBO NAtemp  nit NASat02 Elevacion Ancho Velocidad Rocas
## 1 Trapecio  0.03  2.38  27.33 0.35   92.04        23   16         5    20
## 2 Trapecio  0.03  2.95  27.81  NA   100.03        31   11         0    20
## 3 Trapecio  0.03  3.13  24.27  NA    96.82        35   14        10    30
## 4 Trapecio  1.15  4.73  27.06 7.54   64.35         9    5         2     0
## 5 Trapecio  0.50  8.16  26.60  NA   110.39        43   11         9    10
## 6 Trapecio  0.53  8.57  23.82  NA   106.09        23   11         5    20
##  Canto grava arena Limo
## 1    25    30    20    0
## 2    45    20    15    0
## 3    30    20    10    0
## 4     0     0    50   50
## 5    40    10    20   20
## 6    60    20     0    0
```

## 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 : 3.822
## 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 :3
## Velocidad Rocas Canto grava
## Min. : 0.000 Min. : 0.00 Min. : 0.00 Min. : 0.0
## 1st Qu.: 3.000 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 2.5
## Median :11.000 Median :10.00 Median :25.00 Median :20.0
## Mean : 9.133 Mean :16.25 Mean :25.65 Mean :17.8
## 3rd Qu.:14.000 3rd Qu.:30.00 3rd Qu.:40.00 3rd Qu.:25.0
## Max. :16.000 Max. :90.00 Max. :80.00 Max. :80.0
## NA's :3 NA's :3 NA's :4 NA's :3
## arena Limo
## Min. : 0.00 Min. : 0.00
## 1st Qu.: 10.00 1st Qu.: 0.00
## Median : 15.00 Median : 10.00
## Mean : 19.79 Mean : 20.62
## 3rd Qu.: 25.00 3rd Qu.: 25.00
## Max. :100.00 Max. :100.00
## NA's :3 NA's :3
```

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, -nit, -NADBO)
summary(channel_1)
```

```
## NAN_Am NAtemp NASat02 Elevacion
## Min. :0.0200 Min. :14.67 Min. : 23.43 Min. : 3.00
## 1st Qu.:0.0400 1st Qu.:24.30 1st Qu.: 86.24 1st Qu.: 25.25
## Median :0.2150 Median :26.05 Median : 94.59 Median : 53.00
## Mean :0.3201 Mean :25.84 Mean : 91.05 Mean : 230.89
## 3rd Qu.:0.5000 3rd Qu.:27.70 3rd Qu.:100.52 3rd Qu.: 269.25
## Max. :1.5000 Max. :32.18 Max. :122.73 Max. :2370.00
##
## Ancho Velocidad Rocas Canto
## Min. : 1.000 Min. : 0.000 Min. : 0.00 Min. : 0.00
## 1st Qu.: 2.000 1st Qu.: 3.000 1st Qu.: 0.00 1st Qu.: 0.00
## Median : 3.000 Median :11.000 Median :10.00 Median :25.00
## Mean : 3.822 Mean : 9.133 Mean :16.25 Mean :25.65
## 3rd Qu.: 3.000 3rd Qu.:14.000 3rd Qu.:30.00 3rd Qu.:40.00
## Max. :16.000 Max. :16.000 Max. :90.00 Max. :80.00
## NA's :3 NA's :3 NA's :3 NA's :4
```

```
##      grava      arena      Limo
## Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
## 1st Qu.: 2.5   1st Qu.: 10.00   1st Qu.: 0.00
## Median :20.0   Median : 15.00   Median : 10.00
## Mean   :17.8   Mean   : 19.79   Mean   : 20.62
## 3rd Qu.:25.0   3rd Qu.: 25.00   3rd Qu.: 25.00
## Max.   :80.0   Max.   :100.00   Max.   :100.00
## NA's   :3      NA's   :3      NA's   :3
```

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>

```
channel_2 <- imputePCA(channel_1, ncp=2, 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)

# class(channel_2)
# as.data.frame(channel_2)
```

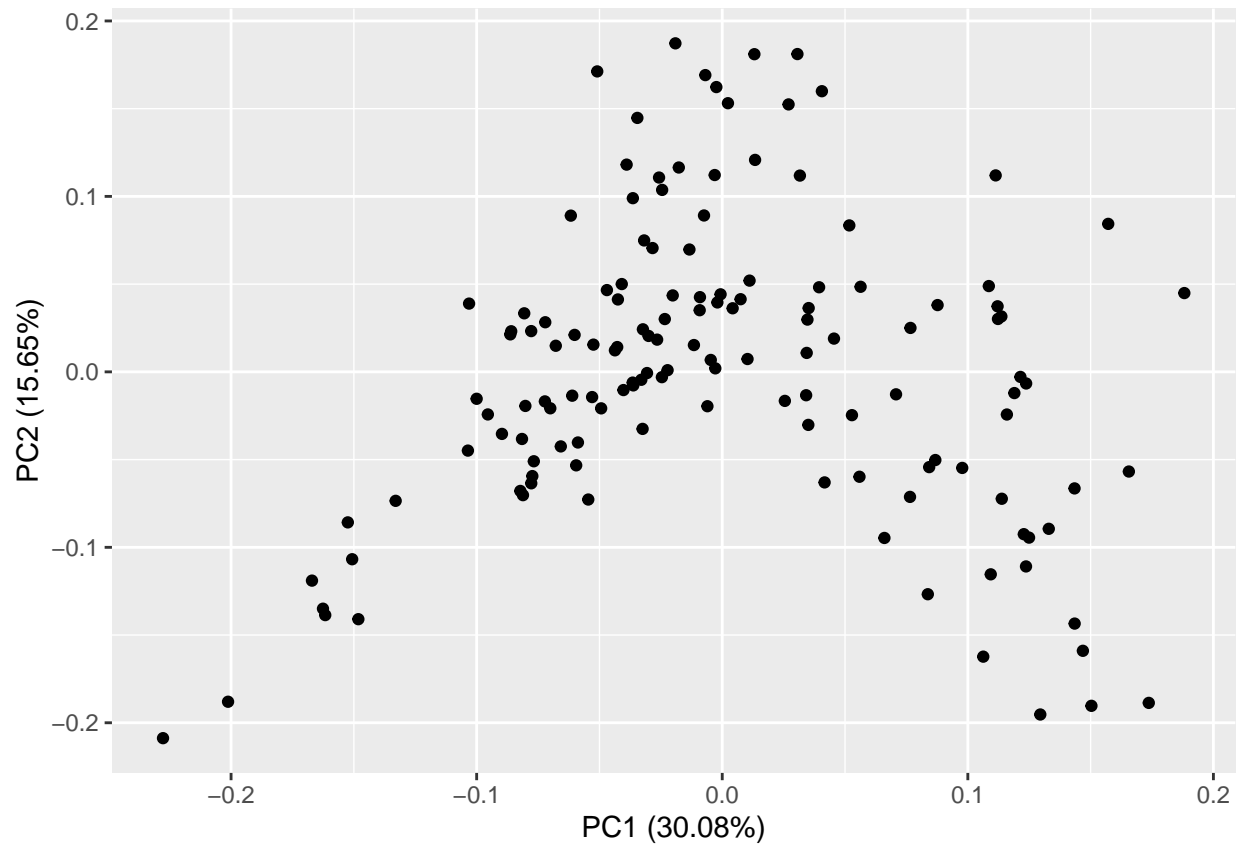
3. Vamos a correr el PCA

```
channel.pca <- prcomp(channel_2$completeObs, center = TRUE, scale = TRUE)
summary(channel.pca)
```

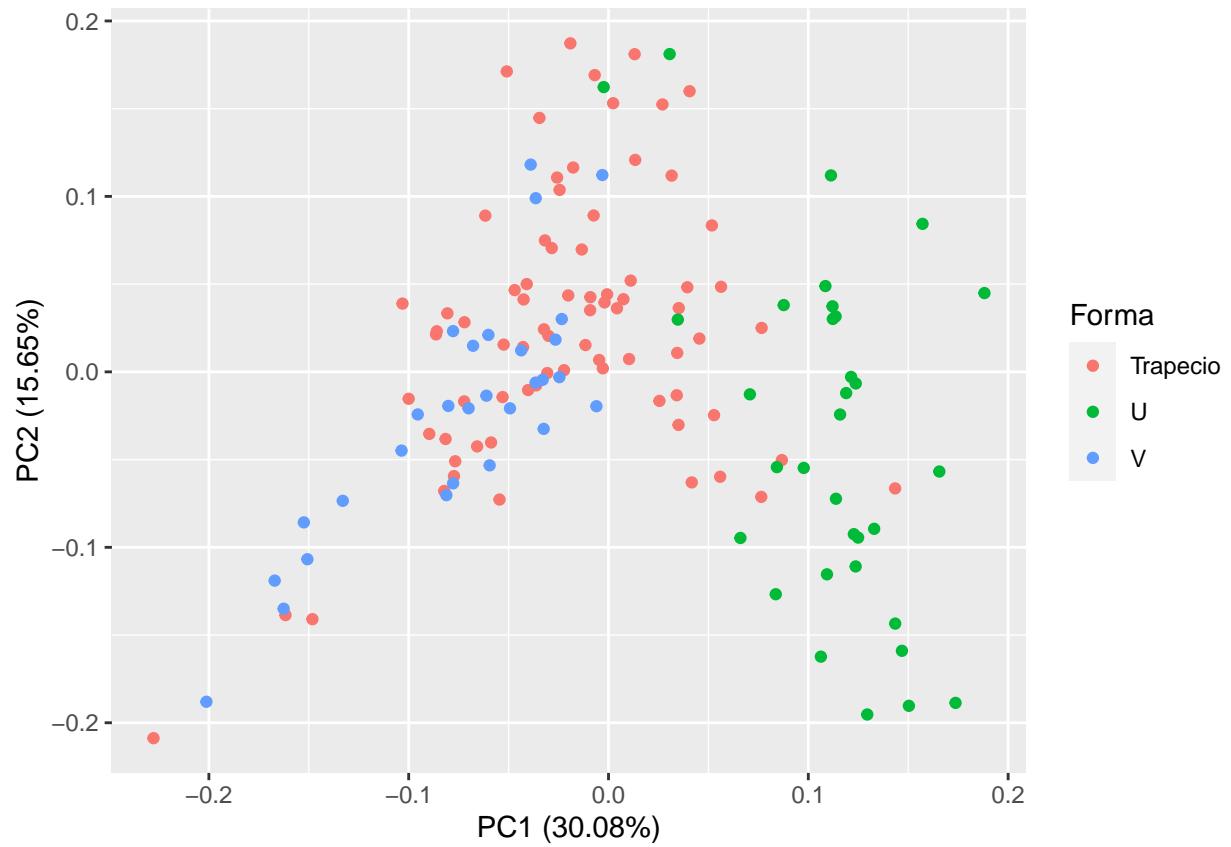
```
## Importance of components:
##      PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation  1.8192 1.3122 1.1975 1.1185 1.00559 0.87738 0.75302
## Proportion of Variance 0.3009 0.1565 0.1304 0.1137 0.09193 0.06998 0.05155
## Cumulative Proportion 0.3009 0.4574 0.5878 0.7015 0.79343 0.86341 0.91496
##      PC8      PC9      PC10      PC11
## Standard deviation  0.6667 0.59031 0.36992 0.07547
## Proportion of Variance 0.0404 0.03168 0.01244 0.00052
## Cumulative Proportion 0.9554 0.98704 0.99948 1.00000
```

3.1 Vamos a ver el grafico.

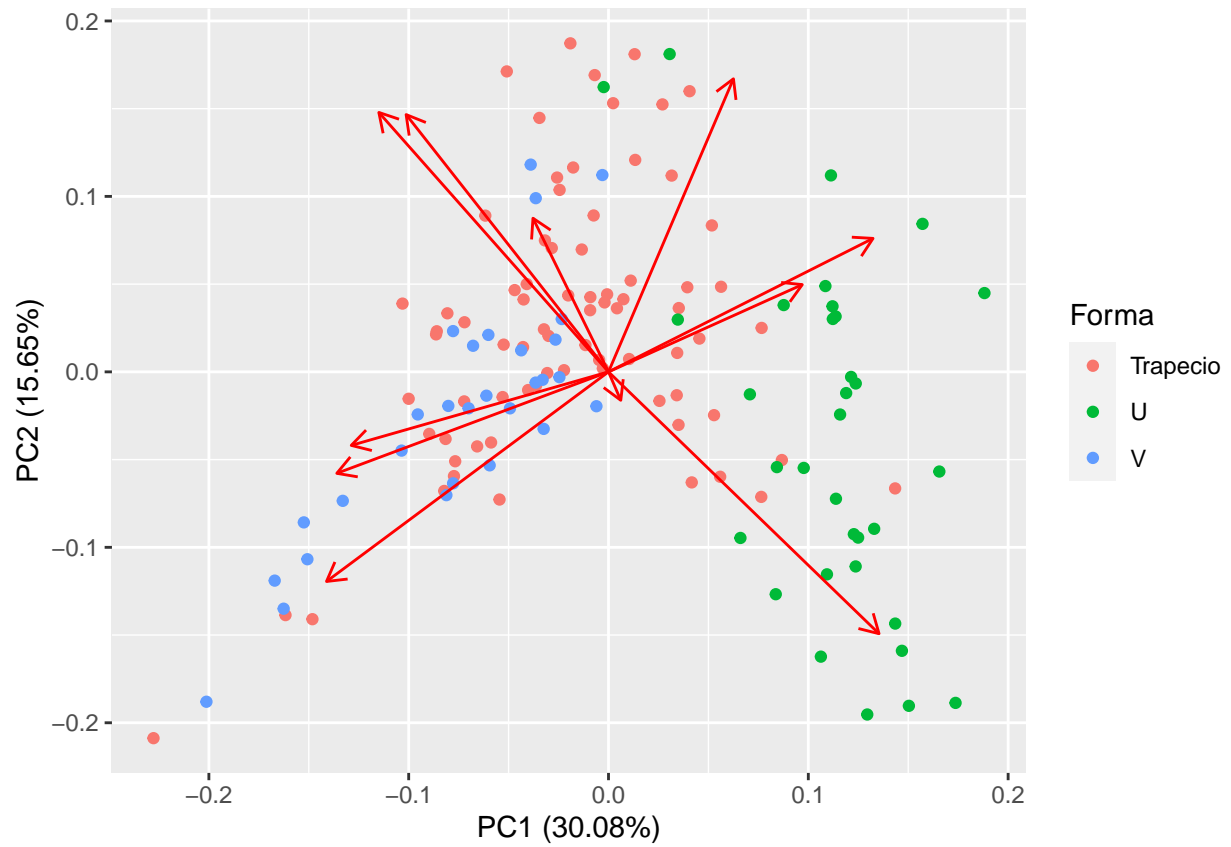
```
autoplot(channel.pca)
```



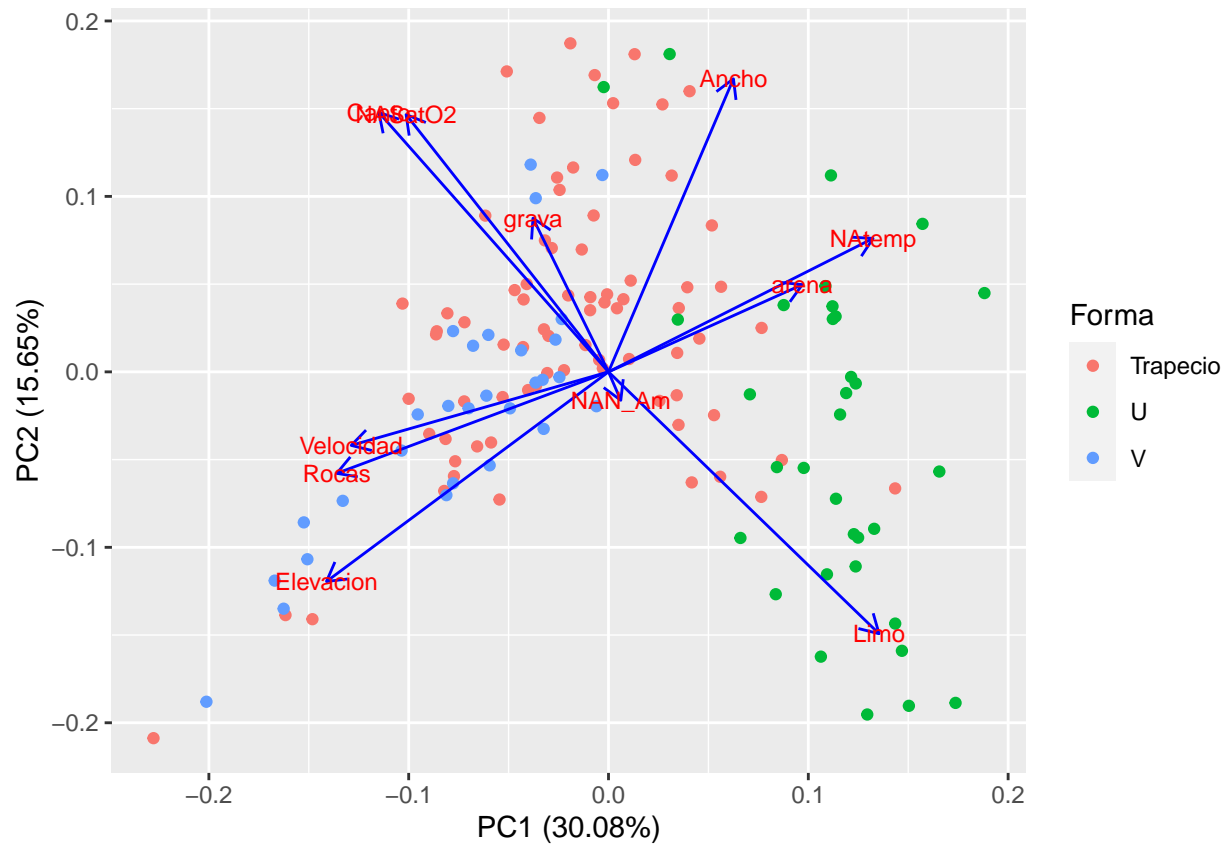
```
autoplot(channel.pca, data = channel, colour = 'Forma')
```



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

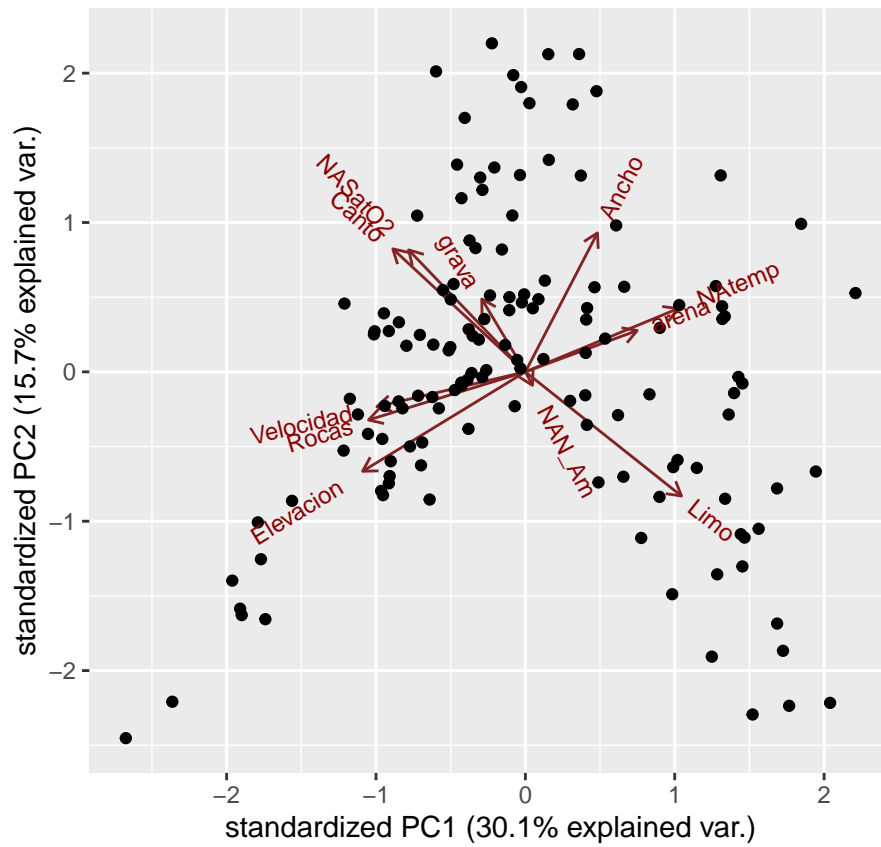


```
autoplot(channel.pca, data = channel, colour = 'Forma', loadings = TRUE,
         loadings.colour = 'blue',
         loadings.label = TRUE, loadings.label.size = 3)
```



Otra manera de ver el grafico

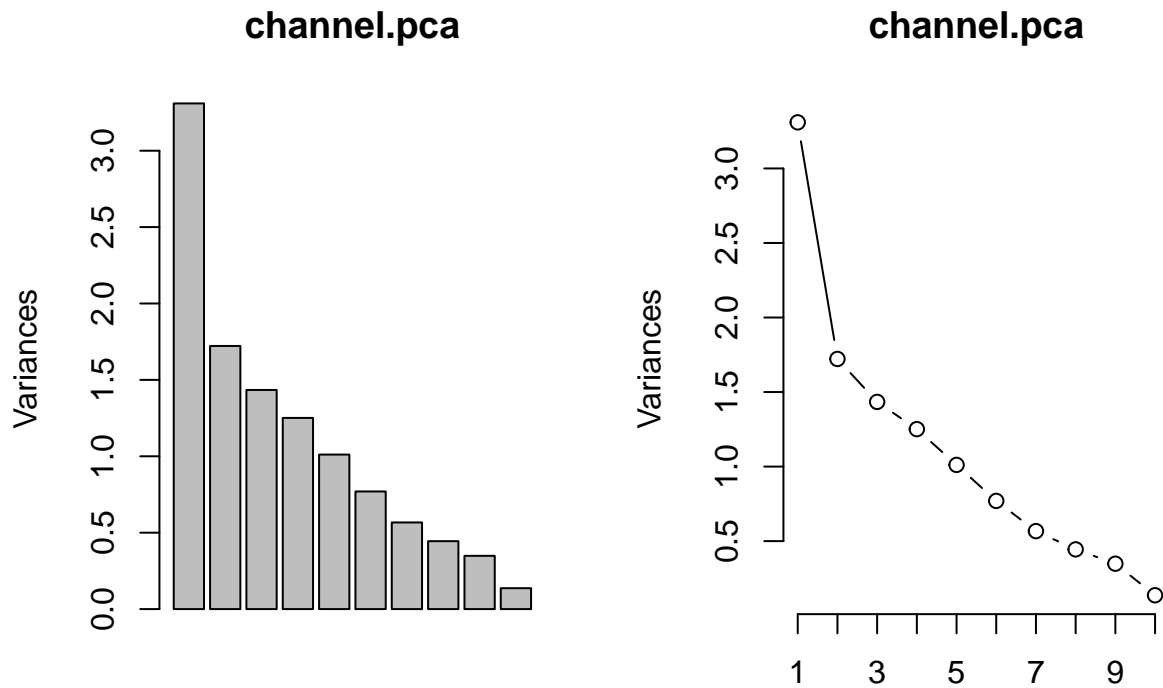
```
ggbiplot(channel.pca, labels=rownames(channel$Forma))
```



3.2 Ver graficamente lo que explica cada axis.

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





3.3 Vamos a ver la contribucion de cada una de las variables. Usamos otra libreria. factoextra

```
get_eigenvalue(channel.pca)
```

| ## |        | eigenvalue  | variance.percent | cumulative.variance.percent |
|----|--------|-------------|------------------|-----------------------------|
| ## | Dim.1  | 3.309339880 | 30.08490800      | 30.08491                    |
| ## | Dim.2  | 1.721987545 | 15.65443223      | 45.73934                    |
| ## | Dim.3  | 1.434048849 | 13.03680772      | 58.77615                    |
| ## | Dim.4  | 1.251144650 | 11.37404227      | 70.15019                    |
| ## | Dim.5  | 1.011205837 | 9.19278033       | 79.34297                    |
| ## | Dim.6  | 0.769787383 | 6.99806712       | 86.34104                    |
| ## | Dim.7  | 0.567041778 | 5.15492525       | 91.49596                    |
| ## | Dim.8  | 0.444432998 | 4.04029998       | 95.53626                    |
| ## | Dim.9  | 0.348471304 | 3.16792094       | 98.70418                    |
| ## | Dim.10 | 0.136844213 | 1.24403830       | 99.94822                    |
| ## | Dim.11 | 0.005695562 | 0.05177784       | 100.00000                   |

```
res.var <- get_pca_var(channel.pca)
res.var$contrib          # Contributions to the PCs
```

| ## |           | Dim.1       | Dim.2      | Dim.3      | Dim.4      | Dim.5      | Dim.6     |
|----|-----------|-------------|------------|------------|------------|------------|-----------|
| ## | NAN_Am    | 0.02920886  | 0.2034616  | 5.8056413  | 31.9945477 | 34.3470814 | 15.246039 |
| ## | NAtemp    | 13.57660503 | 4.4828567  | 15.4148551 | 3.7395130  | 2.1197838  | 14.516177 |
| ## | NASat02   | 7.94579553  | 16.6572006 | 0.4379407  | 8.8574581  | 5.4079970  | 2.598790  |
| ## | Elevacion | 15.42979758 | 11.0563517 | 5.5445224  | 4.1639816  | 0.9325291  | 9.624475  |

```
## Ancho      3.02902190 21.6083873 22.1140351 0.4570409 1.2134531 2.713345
## Velocidad 12.83990263 1.3599088 14.2241968 2.7802734 1.6200790 4.161196
## Rocas      14.32470621 2.5906170 2.5662488 3.2825439 1.7671606 29.965671
## Canto      10.23186117 16.9344566 0.2986754 4.7329896 12.2604278 5.124245
## grava      1.10633855 5.9387801 30.8293418 16.2613832 0.3136215 9.937735
## arena      7.28689335 1.9159508 2.4783815 18.4552620 33.9916827 3.692205
## Limo       14.19986918 17.2520289 0.2861611 5.2750067 6.0261840 2.420122
##           Dim.7      Dim.8      Dim.9      Dim.10      Dim.11
## NAN_Am     0.01132441 5.69938420 6.040932e+00 0.62044106 0.001939282
## NAtemp     0.73662346 0.59088977 2.080669e+00 42.69918930 0.042838114
## NASat02    1.24823008 55.62178693 1.180103e+00 0.02000417 0.024693645
## Elevacion  0.24607958 0.96539511 3.314342e+00 48.68500085 0.037525452
## Ancho      5.68811760 0.01494364 3.793491e+01 5.22569343 0.001047357
## Velocidad  21.91572939 0.18791971 3.942194e+01 1.46336747 0.025485968
## Rocas      23.07674466 4.13345827 1.324401e+00 0.08968473 16.878763447
## Canto      11.02914223 12.77009548 6.399584e+00 0.08647903 20.132043196
## grava      21.59705383 2.73504867 2.422399e-04 0.47831503 10.802140459
## arena      13.91693263 1.00150996 6.374613e-01 0.55983412 16.063886205
## Limo       0.53402214 16.27956826 1.665410e+00 0.07199080 35.989636876
```

```
res.var$coord      # Coordinates
```

```
##           Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
## NAN_Am     0.03109052 -0.05919107 -0.28854069 -0.63269113 0.58933835
## NAtemp     0.67029546 0.27783850 -0.47016652 -0.21630237 -0.14640826
## NASat02    -0.51278980 0.53556971 0.07924824 -0.33289580 0.23385034
## Elevacion  -0.71457991 -0.43633588 0.28197723 0.22824862 0.09710710
## Ancho      0.31660801 0.60999487 0.56313947 0.07561906 -0.11077233
## Velocidad  -0.65185583 -0.15302764 -0.45164359 0.18650802 -0.12799349
## Rocas      -0.68851523 -0.21121104 0.19183655 -0.20265580 0.13367734
## Canto      -0.58189953 0.54000855 0.06544578 -0.24334450 -0.35210533
## grava      -0.19134394 0.31978908 -0.66491189 0.45105812 0.05631482
## arena      0.49106829 0.18163820 0.18852374 0.48052266 0.58628140
## Limo       0.68550852 -0.54504843 0.06406005 -0.25690069 -0.24685446
##           Dim.6      Dim.7      Dim.8      Dim.9      Dim.10
## NAN_Am     0.3425815 -0.008013373 0.159153838 -0.1450893281 -0.029138251
## NAtemp     -0.3342809 0.064629426 0.051245576 0.0851500600 -0.241725815
## NASat02    -0.1414396 -0.084130767 -0.497193700 0.0641273732 -0.005232070
## Elevacion  0.2721911 0.037354705 -0.065502171 0.1074687420 -0.258113553
## Ancho      0.1445233 0.179593995 0.008149508 -0.3635825795 -0.084563935
## Velocidad  -0.1789759 -0.352521406 -0.028899433 -0.3706401926 -0.044749678
## Rocas      -0.4802832 0.361738556 0.135537643 -0.0679349618 0.011078283
## Canto      0.1986097 -0.250079676 0.238232068 0.1493342367 -0.010878490
## grava      0.2765853 0.349949022 -0.110251797 0.0009187691 0.025584105
## arena      -0.1685886 -0.280917821 0.066716120 0.0471314081 -0.027678522
## Limo       0.1364910 -0.055028435 -0.268982849 -0.0761805491 -0.009925485
##           Dim.11
## NAN_Am     0.0003323447
## NAtemp     -0.0015620088
## NASat02    -0.0011859350
## Elevacion  -0.0014619458
## Ancho      0.0002442393
## Velocidad  0.0012048108
## Rocas      0.0310054912
```

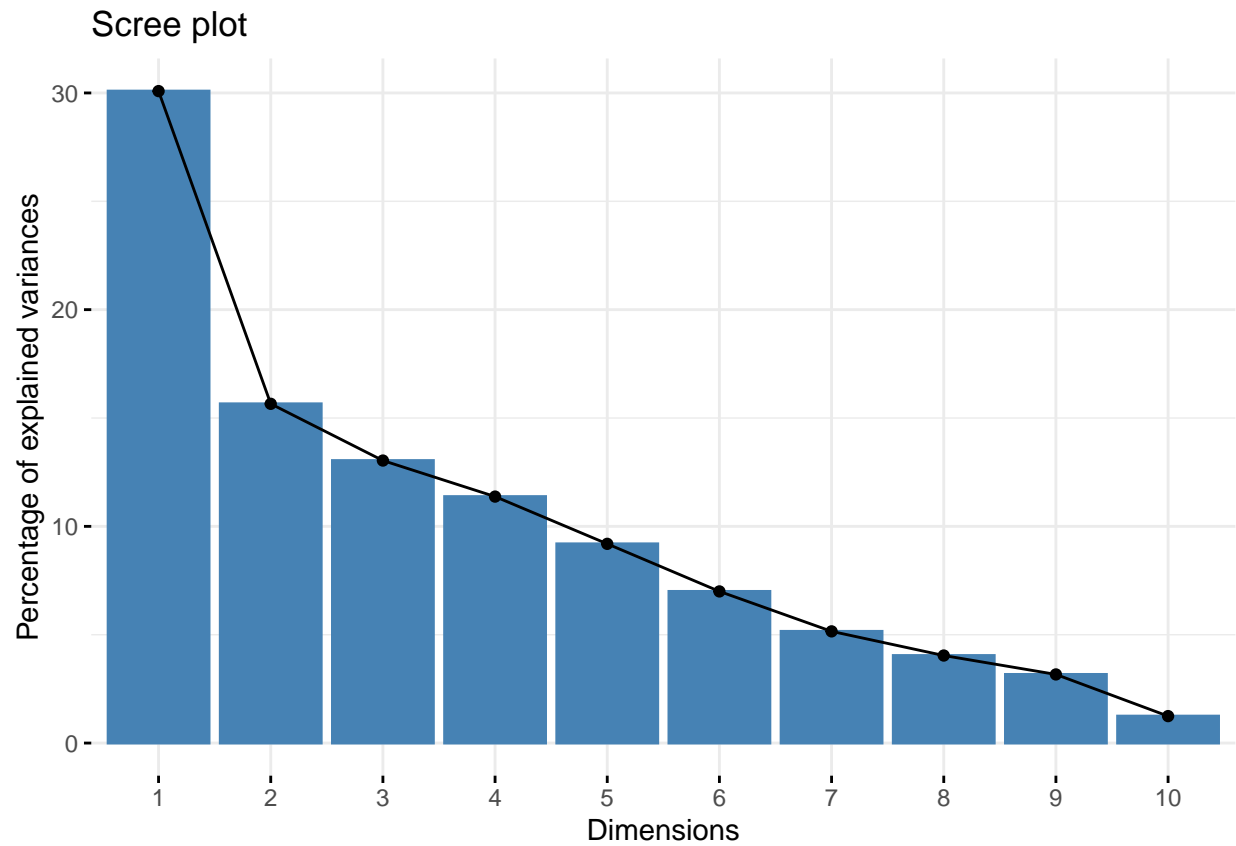
```
## Canto      0.0338619707
## grava      0.0248040851
## arena      0.0302477874
## Limo       0.0452748515
```

```
res.var$cos2      # Quality of representation
```

```
##           Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
## NAN_Am      0.0009666203 0.003503583 0.083255732 0.400298071 0.347319692
## NAtemp      0.4492960046 0.077194234 0.221056552 0.046786717 0.021435378
## NASat02     0.2629533804 0.286834919 0.006280284 0.110819613 0.054685981
## Elevacion   0.5106244449 0.190388999 0.079511160 0.052097433 0.009429788
## Ancho       0.1002406296 0.372093737 0.317126066 0.005718242 0.012270509
## Velocidad   0.4249160183 0.023417459 0.203981931 0.034785242 0.016382333
## Rocas       0.4740532154 0.044610103 0.036801261 0.041069372 0.017869631
## Canto       0.3386070623 0.291609234 0.004283151 0.059216546 0.123978161
## grava       0.0366125028 0.102265054 0.442107821 0.203453425 0.003171359
## arena       0.2411480677 0.032992435 0.035541202 0.230902024 0.343725879
## Limo       0.4699219337 0.297077788 0.004103690 0.065997964 0.060937124
##           Dim.6      Dim.7      Dim.8      Dim.9      Dim.10
## NAN_Am      0.11736208 6.421415e-05 2.532994e-02 2.105091e-02 8.490377e-04
## NAtemp      0.11174370 4.176963e-03 2.626109e-03 7.250533e-03 5.843137e-02
## NASat02     0.02000516 7.077986e-03 2.472016e-01 4.112320e-03 2.737455e-05
## Elevacion   0.07408799 1.395374e-03 4.290534e-03 1.154953e-02 6.662261e-02
## Ancho       0.02088699 3.225400e-02 6.641449e-05 1.321923e-01 7.151059e-03
## Velocidad   0.03203236 1.242713e-01 8.351772e-04 1.373742e-01 2.002534e-03
## Rocas       0.23067195 1.308548e-01 1.837045e-02 4.615159e-03 1.227284e-04
## Canto       0.03944579 6.253984e-02 5.675452e-02 2.230071e-02 1.183415e-04
## grava       0.07649943 1.224643e-01 1.215546e-02 8.441367e-07 6.545464e-04
## arena       0.02842213 7.891482e-02 4.451041e-03 2.221370e-03 7.661006e-04
## Limo       0.01862979 3.028129e-03 7.235177e-02 5.803476e-03 9.851525e-05
##           Dim.11
## NAN_Am      1.104530e-07
## NAtemp      2.439871e-06
## NASat02     1.406442e-06
## Elevacion   2.137285e-06
## Ancho       5.965285e-08
## Velocidad   1.451569e-06
## Rocas       9.613405e-04
## Canto       1.146633e-03
## grava       6.152426e-04
## arena       9.149286e-04
## Limo       2.049812e-03
```

4 Otras formas de visualizar los datos.

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



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

```
## Warning: ggrepel: 41 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps
```

PCA – Biplot

