EJERCICIO DE CLASE

ANÁLISIS DE COMPONENTES PRINCIPALES

El fichero BARRIOS contiene información socio-económica de algunos barrios de Madrid. Para reducir el número de variables e intentar encontrar relaciones, tanto entre variables como entre barrios, realizar los siguientes apartados.

1. Calcular los estadísticos básicos de todas las variables. Comparar sus medias y varianzas.

```
library(pastecs)
#Descriptivos
Est<-stat.desc(datos,basic=FALSE)
```

	P_TOT	P_14	P_65	ANAL	NES	OCU	OCUIN	OCUSER	TEC	PD	TM
median	169.80	36.45	17.60	4.10	21.50	55.35	11.2	36.85	9.20	1.35	15.05
mean	171.67	40.56	19.90	4.47	22.07	59.67	11.7	38.18	9.17	1.36	15.15
SE.mean	10.53	3.64	2.01	0.69	3.04	3.89	1.1	2.35	1.01	0.19	1.96
CI.mean.0.95	22.21	7.68	4.24	1.46	6.40	8.21	2.3	4.96	2.13	0.41	4.13
var	1994.90	238.38	72.62	8.68	165.82	272.41	21.8	99.47	18.36	0.68	68.95
std.dev	44.66	15.44	8.52	2.95	12.88	16.50	4.7	9.97	4.28	0.82	8.30
coef.var	0.26	0.38	0.43	0.66	0.58	0.28	0.4	0.26	0.47	0.61	0.55

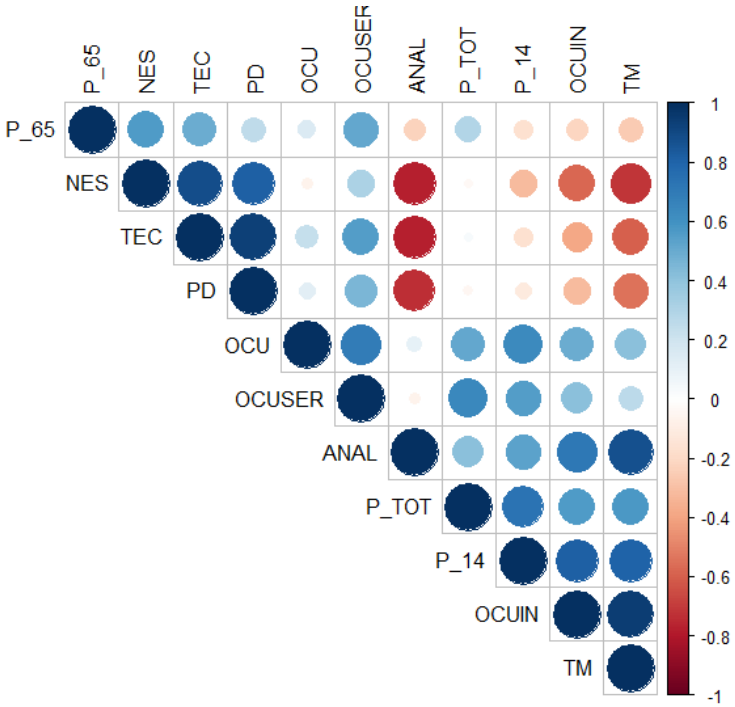
2. **Calcular l**a matriz de correlaciones, y su representación gráfica ¿Cuáles son las variables más correlacionadas? ¿Cómo es el sentido de esa correlacción?

```
datos <- BARRIOS[,-1]
#Matriz correlacioens
R<-cor(datos, method="pearson")
print(R)
```

	P_TOT	P_14	P_65	ANAL	NES	OCU	OCUIN	OCUSER	TEC	PD	TM
P_TOT	1.000	0.74	0.29	0.420	-0.032	0.512	0.57	0.641	0.038	-0.046	0.58
P_14	0.738	1.00	-0.17	0.537	-0.317	0.639	0.82	0.552	-0.161	-0.115	0.80
P_65	0.294	-0.17	1.00	-0.221	0.564	0.152	-0.22	0.515	0.496	0.251	-0.25
ANAL	0.420	0.54	-0.22	1.000	-0.773	0.101	0.71	-0.062	-0.775	-0.738	0.88
NES	-0.032	-0.32	0.56	-0.773	1.000	-0.063	-0.58	0.315	0.890	0.817	-0.72
OCU	0.512	0.64	0.15	0.101	-0.063	1.000	0.49	0.694	0.231	0.127	0.42

OCUIN	0.568	0.82	-0.22	0.713	-0.575	0.491	1.00	0.412	-0.390	-0.310	0.95
OCUSER	0.641	0.55	0.51	-0.062	0.315	0.694	0.41	1.000	0.560	0.458	0.27
TEC	0.038	-0.16	0.50	-0.775	0.890	0.231	-0.39	0.560	1.000	0.938	-0.59
PD	-0.046	-0.11	0.25	-0.738	0.817	0.127	-0.31	0.458	0.938	1.000	-0.54
TM	0.575	0.80	-0.25	0.877	-0.719	0.418	0.95	0.265	-0.593	-0.542	1.00

```
library(corrplot)
corrplot(R, type="upper", order="hclust",tl.col="black", tl.srt=90)
```



3. Realizar un análisis de componentes principales sobre la matriz de correlaciones, **calculando 6 componentes**. Estudiar los valores de los autovalores obtenidos y las gráficas que los resumen. ¿Cuál es el número adecuado de componentes?

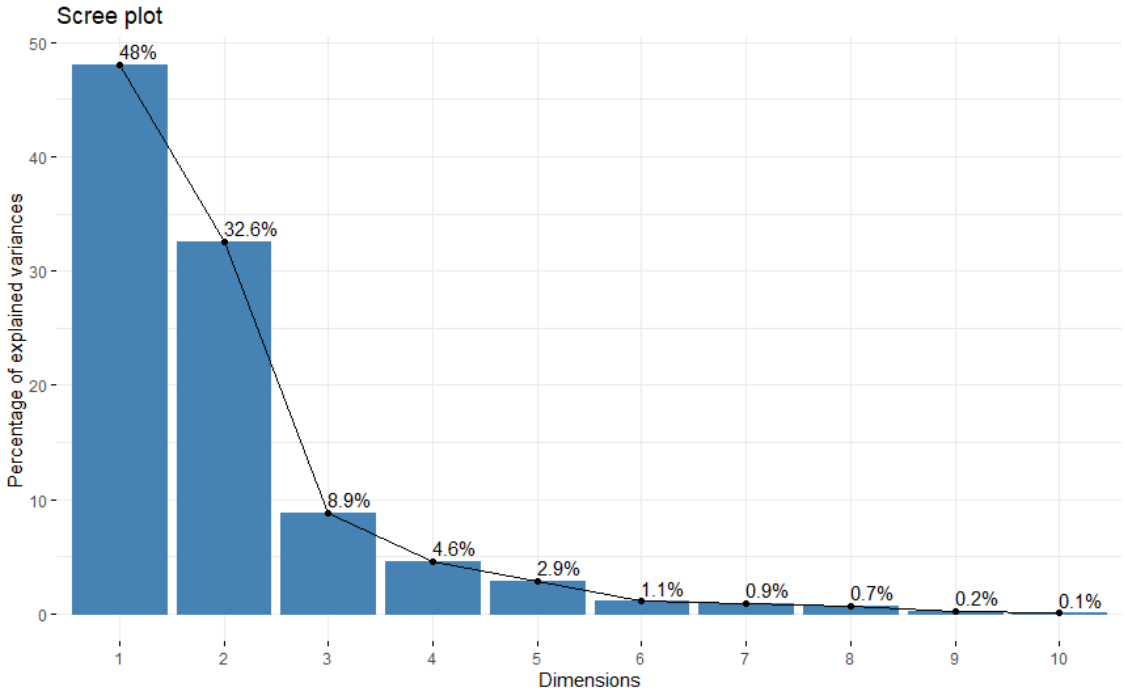
```
library(FactoMineR)
fit<-PCA(datos,scale.unit=TRUE,ncp=6,graph=TRUE)
head(fit)
```

$`eig`			
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	5.279827776	47.99843432	47.99843
comp 2	3.585309098	32.59371908	80.59215
comp 3	0.975584231	8.86894756	89.46110
comp 4	0.505962712	4.59966102	94.06076
comp 5	0.318534034	2.89576395	96.95653
comp 6	0.119957685	1.09052440	98.04705

```
library(factoextra)

# Scree plot

fviz_eig(fit,addlabels=TRUE)
```



4. Hacer de nuevo el análisis sobre la matriz de correlaciones pero ahora **indicando el número de componentes principales que hemos decidido retener.** Sobre este análisis contestar los siguientes apartados.

```
fit<-PCA(datos,scale.unit=TRUE,ncp=3,graph=TRUE)
head(fit)
```

a. ¿Cuál es la expresión para calcular la primera Componente en función de las variables originales?

```
$svd$V
             [,1]        [,2]         [,3]
 [1,]  0.21543198  0.36320961  0.282094194
 [2,]  0.31738326  0.29737923 -0.229411408
 [3,] -0.15614265  0.27143499  0.757757333
 [4,]  0.39755792 -0.07275078  0.249798036
 [5,] -0.36357906  0.23455426  0.088697617
 [6,]  0.14598396  0.38316401 -0.214715744
 [7,]  0.37705097  0.18060272 -0.151272633
 [8,]  0.03368353  0.50511141  0.041616928
 [9,] -0.32242098  0.34136903 -0.135299616
[10,] -0.30158961  0.29181939 -0.366366113
[11,]  0.42274041  0.09694402 -0.009078914
```

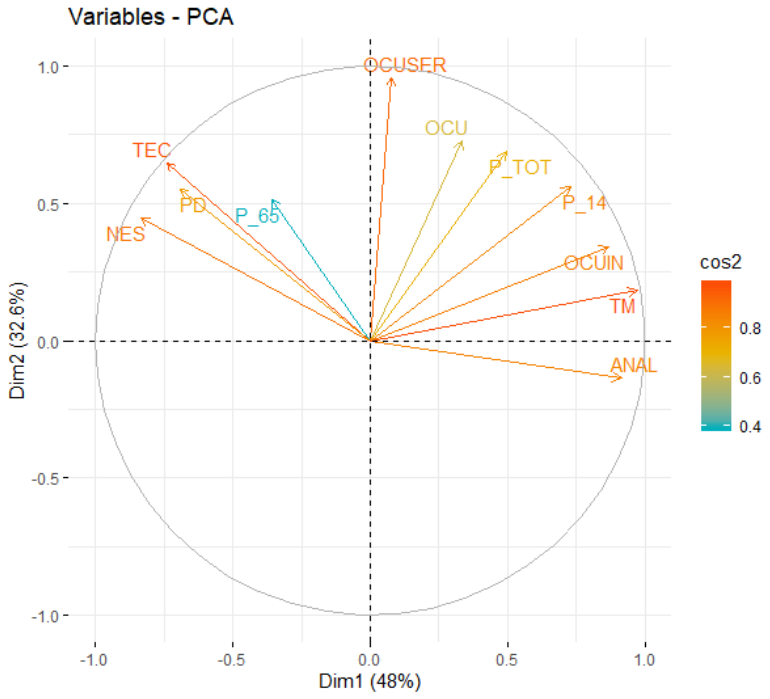
$$CP_1 = 0.21PTOT^* + 0.32P14^* - 0.15P65^* + 0.39ANAL^* - 0.36NES^* + ... + 0.42TM^*$$

b .Mostar una tabla con las correlaciones de las Variables con las Componentes Principales. Para cada Componente indicar las variables con las que está más correlacionada

```
$var$cor
             Dim.1      Dim.2        Dim.3
P_TOT    0.4950169  0.6877342  0.278629140
P_14     0.7292793  0.5630850 -0.226593473
P_65    -0.3587826  0.5139598  0.748449554
ANAL     0.9135036 -0.1377530  0.246729685
NES     -0.8354275  0.4441264  0.087608115
OCU      0.3354401  0.7255177 -0.212078320
OCUIN    0.8663830  0.3419697 -0.149414502
OCUSER   0.0773976  0.9564240  0.041105734
TEC     -0.7408549  0.6463793 -0.133637687
PD      -0.6929889  0.5525575 -0.361865919
TM       0.9713677  0.1835627 -0.008967395
```

c. Comentar los gráficos que representan las variables en los planos formados por las componentes, intentando explicar lo que representa cada componente

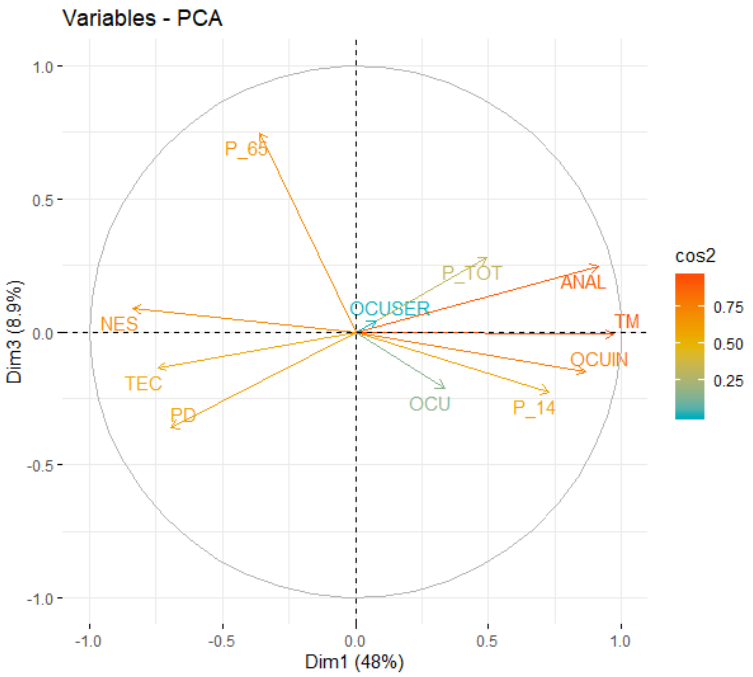
```
fviz_pca_var(fit, axes = c(1, 2), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"),      repel = TRUE )
```



La componente 1 representa el número de trabajadores manuales (TM), el porcentaje de analfabetismo (ANAL), ocupados en industria, Nivel de estudios superiores en negativo y población menor de 14 años.

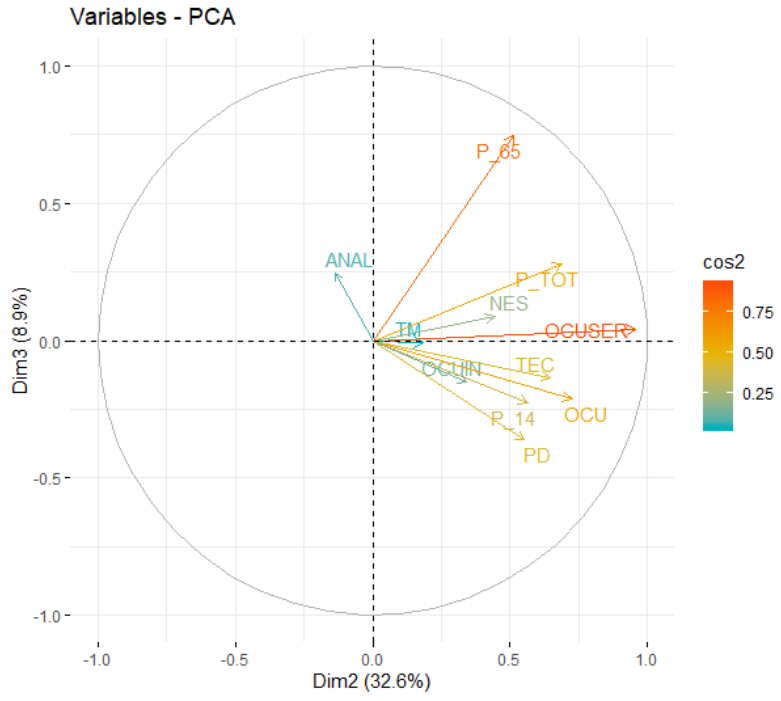
La componente 2 representa a la variable Número de ocupados en servicios, Número de Ocupados y Población Total

```
fviz_pca_var(fit, axes = c(1,3), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE )
```



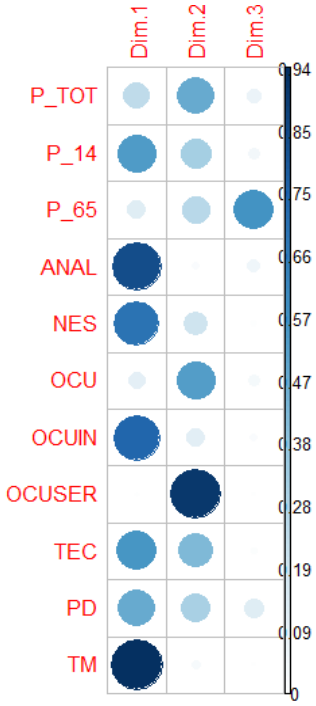
La Componente 3 representa a la población mayor de 65 años.

```
fviz_pca_var(fit, axes = c(2,3), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),  repel = TRUE )
```



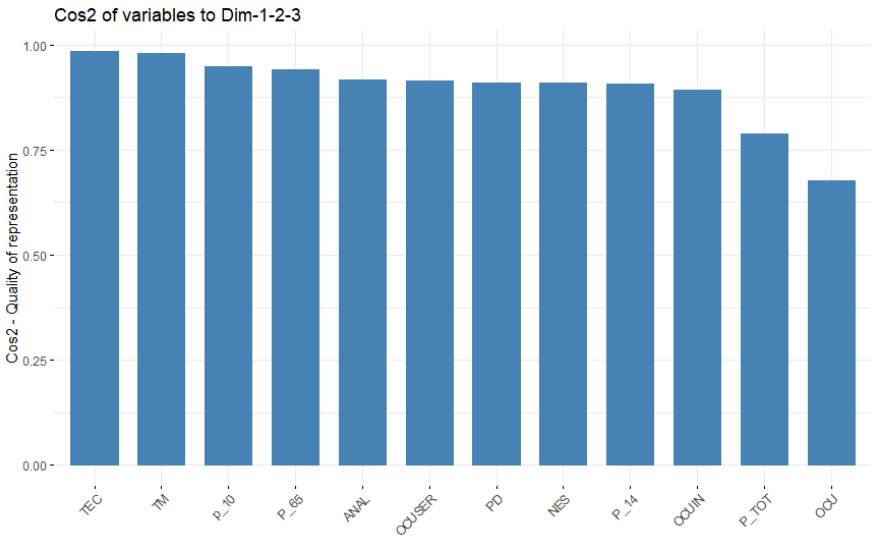
d. Mostrar la tabla y los gráficos que nos muestran la proporción de la varianza de cada variable que es explicado por cada componente. ¿Cuál de las variables es la que está peor explicada?

```
var<-get_pca_var(fit)
print(var$cos2)
corrplot(var$cos2,is.corr=FALSE)
```



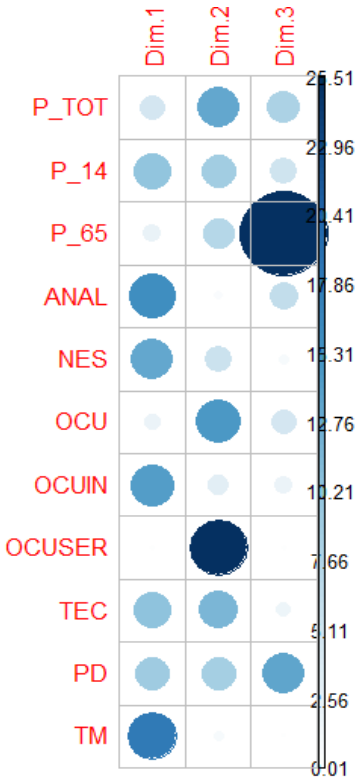
	Dim.1	Dim.2	Dim.3
P_TOT	0.245041752	0.47297834	7.763420e-02
P_14	0.531848328	0.31706467	5.134460e-02
P_65	0.128724987	0.26415465	5.601767e-01
ANAL	0.834488902	0.01897588	6.087554e-02
NES	0.697939037	0.19724828	7.675182e-03
OCU	0.112520085	0.52637592	4.497721e-02
OCUIN	0.750619574	0.11694325	2.232469e-02
OCUSER	0.005990389	0.91474694	1.689681e-03
TEC	0.548866003	0.41780616	1.785903e-02
PD	0.480233567	0.30531975	1.309469e-01
TM	0.943555151	0.03369525	8.041417e-05

```
#Porcentaje de variabilidad explicada por las tres CP
fviz_cos2(fit,choice="var",axes=1:3)
```



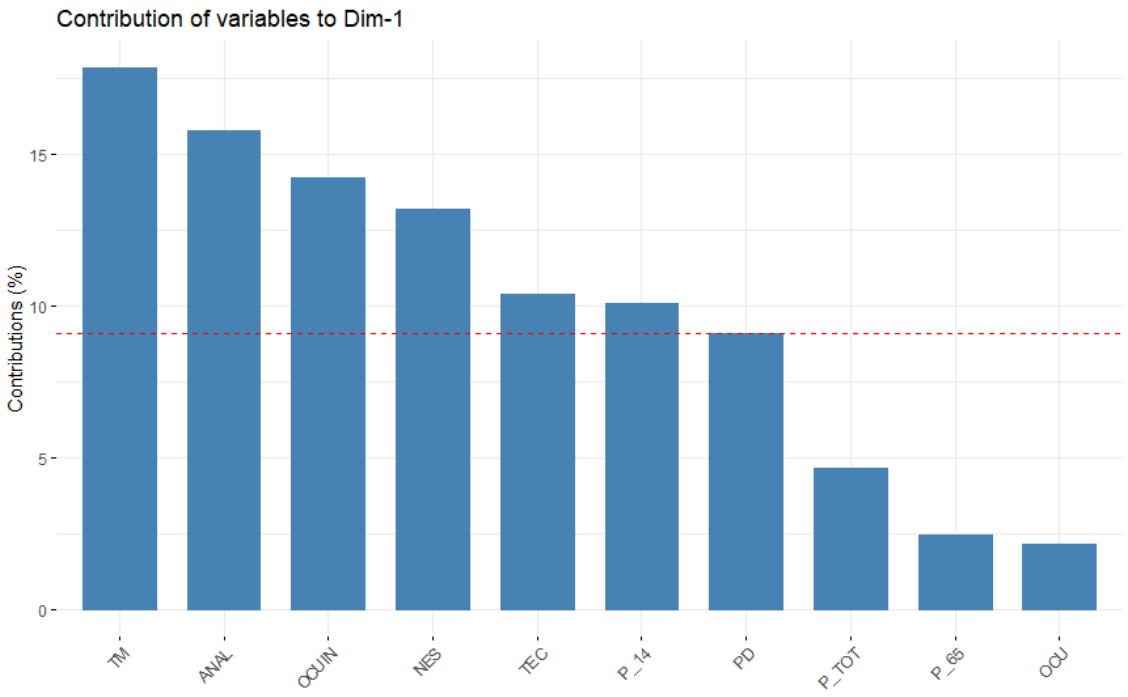
e. Mostrar la tabla y los gráficos que nos muestran el porcentaje de la varianza de cada Componente que es debido a cada variable. ¿Cuál de las variables contribuyen más a cada Componente?

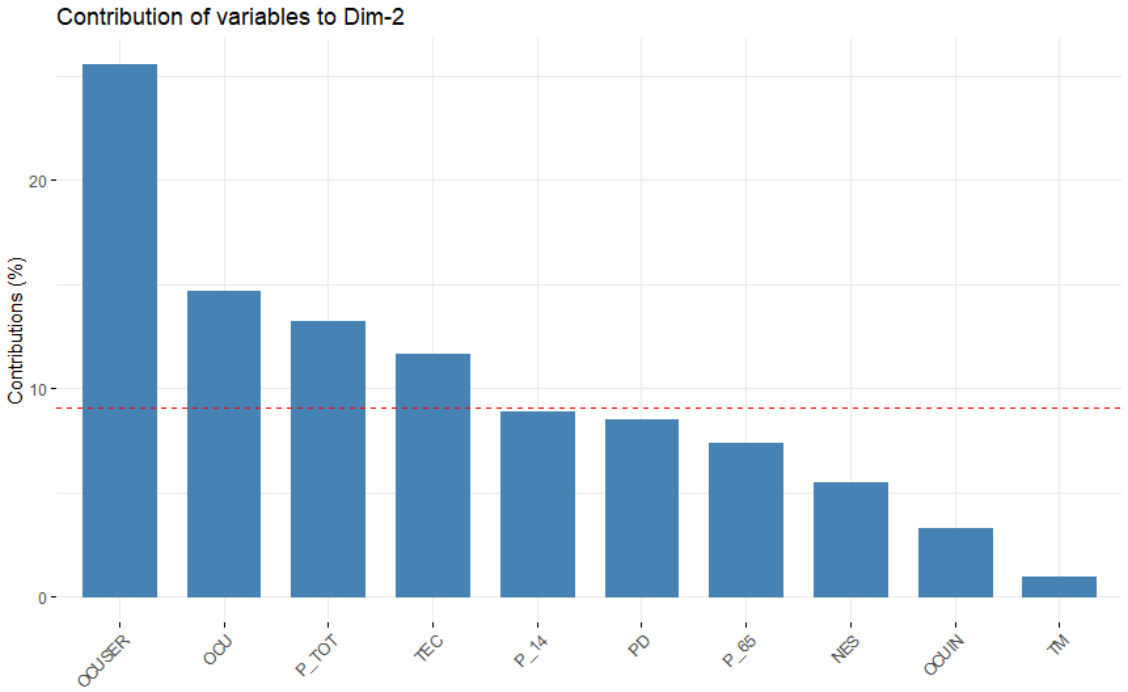
```
corrplot(var$contrib, is.corr=FALSE)
print(var$contrib, digit=2)
```

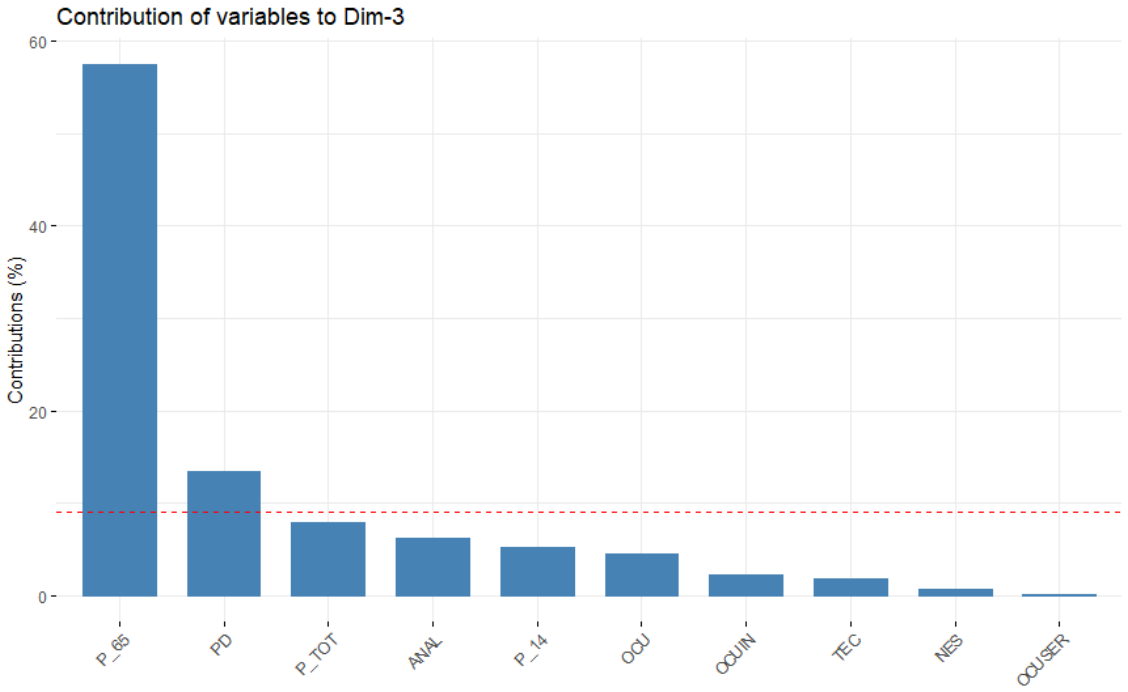



```
  Dim.1 Dim.2   Dim.3
P_TOT   4.64 13.19  7.9577
P_14   10.07  8.84  5.2630
P_65    2.44  7.37 57.4196
ANAL   15.81  0.53  6.2399
NES    13.22  5.50  0.7867
OCU     2.13 14.68  4.6103
OCUIN  14.22  3.26  2.2883
OCUSER  0.11 25.51  0.1732
TEC    10.40 11.65  1.8306
PD      9.10  8.52 13.4224
TM     17.87  0.94  0.0082
```

```
#Contribución de las variables a  cada Componente
fviz_contrib(fit,choice="var",axes=1,top=10)
fviz_contrib(fit,choice="var",axes=2,top=10)
fviz_contrib(fit,choice="var",axes=3,top=10)
```

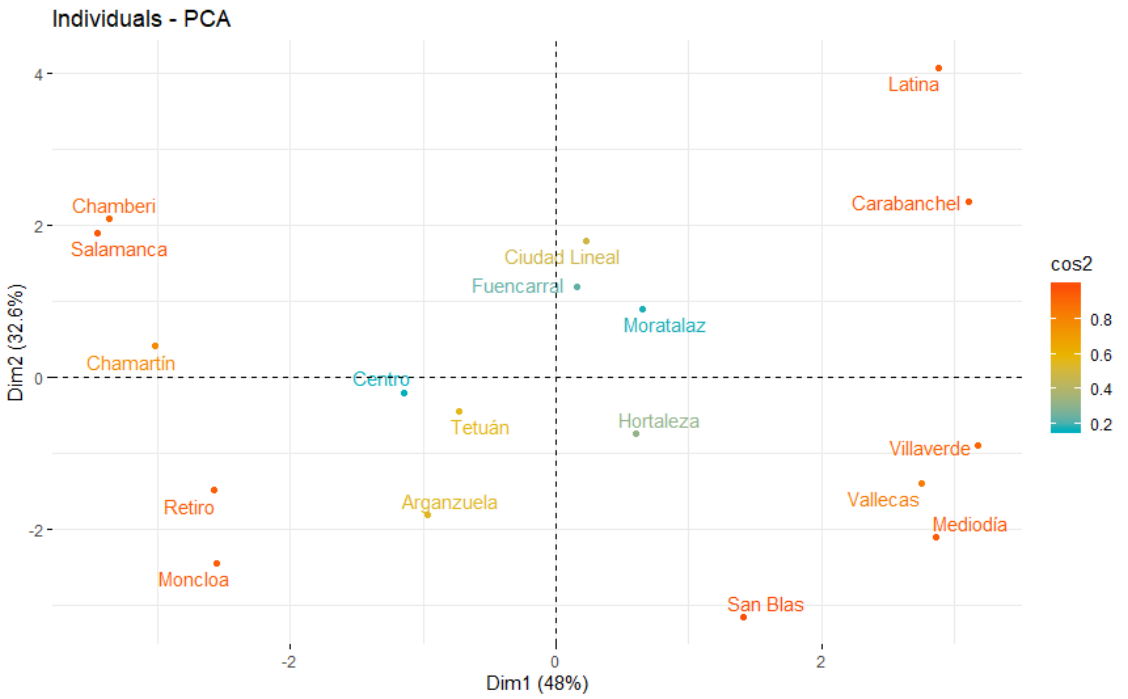




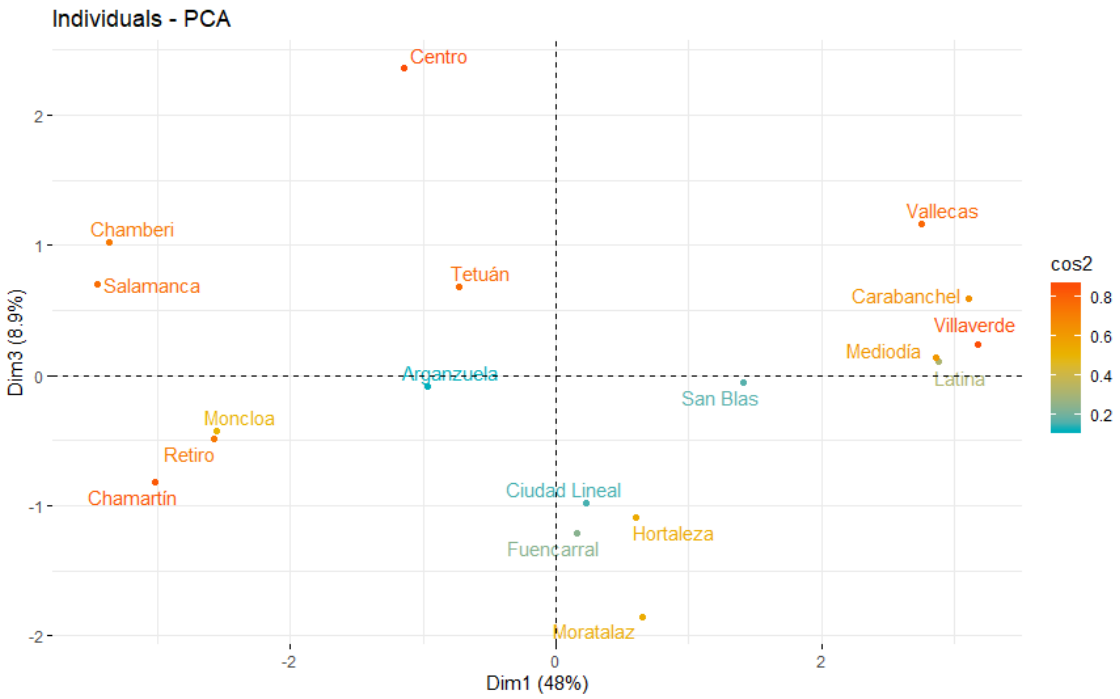


.d. Sobre los gráficos que representan las observaciones en los nuevos ejes Teniendo en cuenta la posición de los barrios en el gráfico ¿Qué barrios tienen una posición más destacada en cada componente?

```
fviz_pca_ind(fit,axes = c(1, 2), col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE) # Avoid text overlapping (slow if many points)
```



```
fviz_pca_ind(fit,axes = c(1, 3), col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```




```
fviz_pca_ind(fit,axes = c(2, 3), col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```