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.

```
BARRIOS<- read_excel("C:/Users/reven/OneDrive/Desktop/Master Big data/
Clases/ACP y Factorial/BARRIOS.xlsx")
datos<- as.data.frame(BARRIOS)
rownames(datos)<-datos[,1]
datos<-datos[,-1]
```

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

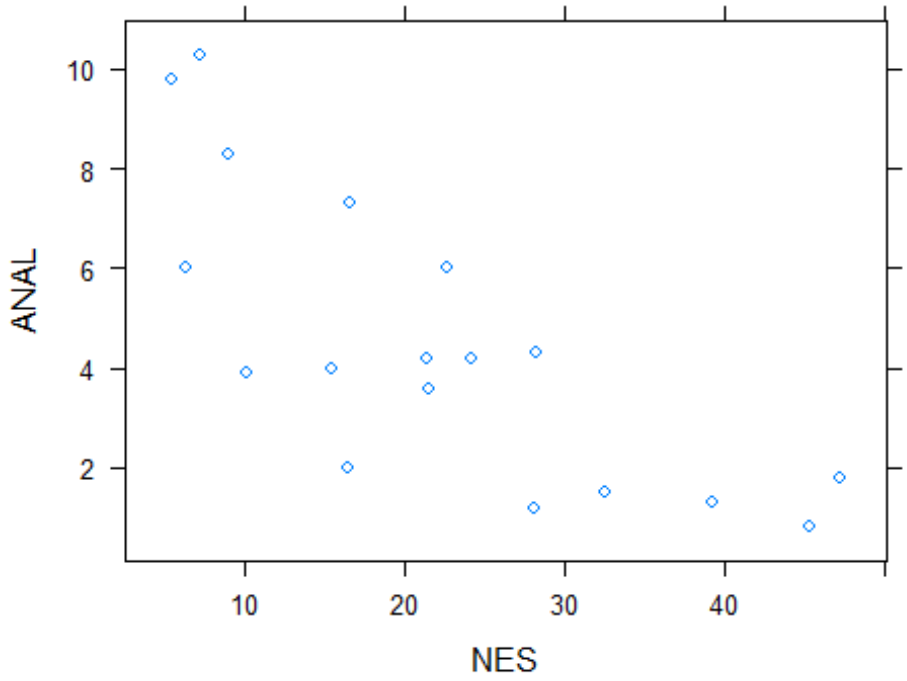
```
#Descriptivos
Est<-stat.desc(datos,basic=FALSE)
knitr::kable(Est,  digits =2,caption = "Estadísticos descriptivos")
```

Estadísticos descriptivos

	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	11.25	36.85	9.20	1.35	15.05
mean	171.67	40.56	19.90	4.47	22.07	59.67	11.66	38.18	9.17	1.36	15.15
SE.mean	10.53	3.64	2.01	0.69	3.04	3.89	1.10	2.35	1.01	0.19	1.96
CI.mean.0.95	22.21	7.68	4.24	1.46	6.40	8.21	2.32	4.96	2.13	0.41	4.13
var	1994.90	238.38	72.62	8.68	165.82	272.41	21.80	99.47	18.36	0.68	68.95
std.dev	44.66	15.44	8.52	2.95	12.88	16.50	4.67	9.97	4.28	0.82	8.30
coef.var	0.26	0.38	0.43	0.66	0.58	0.28	0.40	0.26	0.47	0.61	0.55

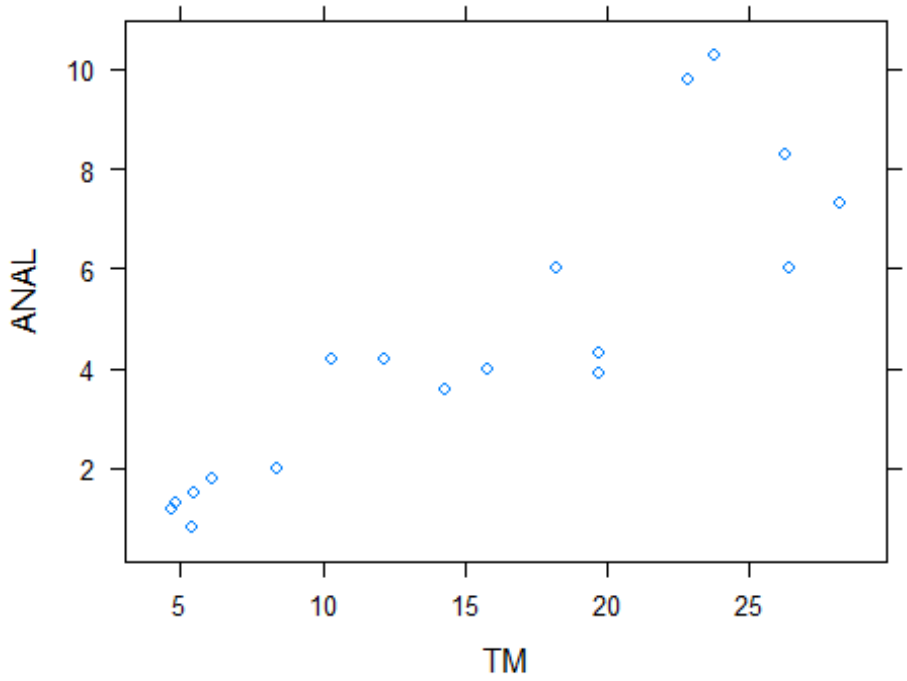
2. Representar el gráfico de dispersión de las variables NES, ANAL y de las varianles ANAL,TM.

```
xyplot(ANAL ~ NES, data =datos)
```



Podemos ver una relación lineal de tipo inverso, es decir, si en un barrio hay mayor porcentaje de nivel de estudios superiores entonces hay un menor porcentaje de analfabetismo.

```
xyplot(ANAL ~ TM, data =datos)
```



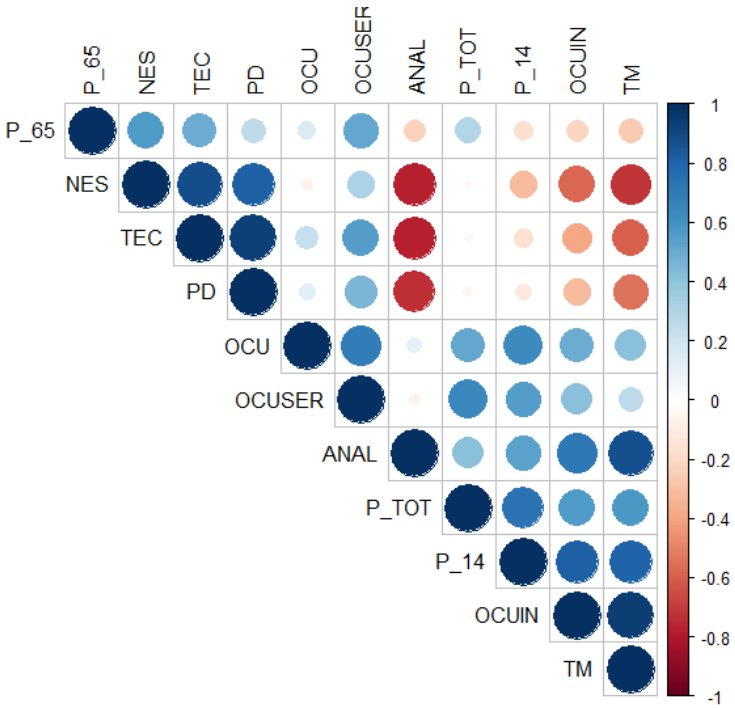
Aquí la relación es directa, a medida que aumenta el porcentaje de trabajadores manuales aumenta también el nivel de analfabetismo.

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

```
 R<-cor(datos, method="pearson")
knitr::kable(R,  digits =2,caption = "Correlaciones")
```

	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

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



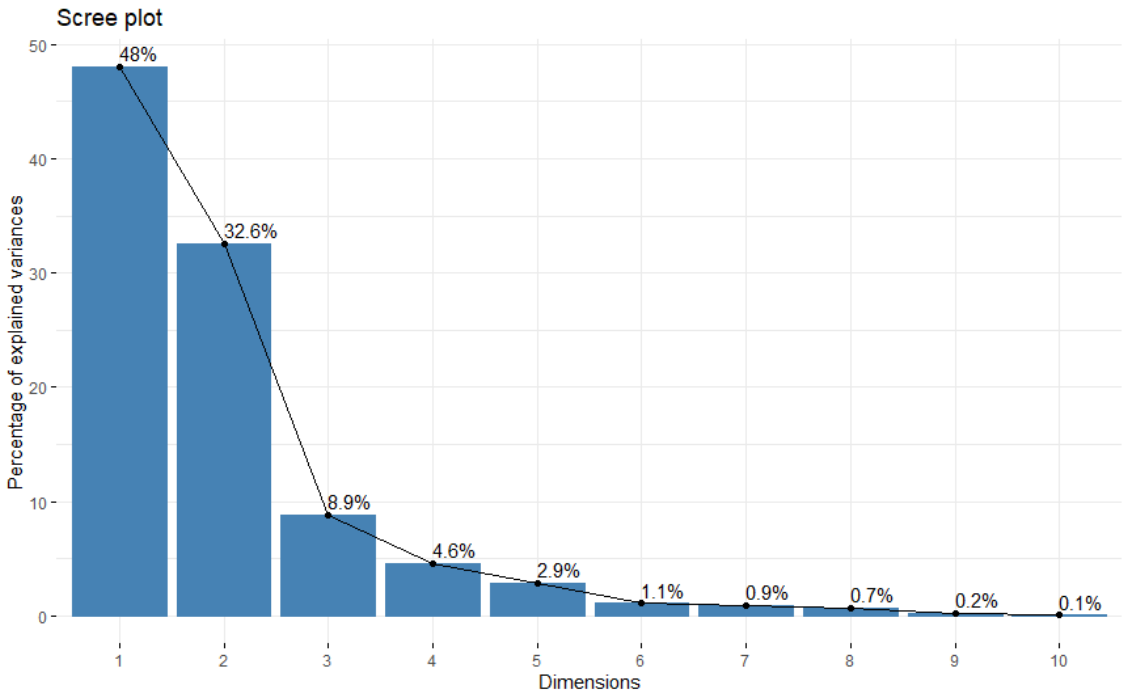
4. Realizar un análisis de componentes principales sobre la matriz de correlaciones, **calculando 7 componentes**. Estudiar los valores de los autovalores obtenidos y las gráficas que los resumen. ¿Cuál es el número adecuado de componentes?

```
fit<-PCA(datos,scale.unit=TRUE,ncp=7,graph=TRUE)

eig<-get_eigenvalue(fit)
knitr::kable(eig,  digits =2,caption = "Autovalores")
```

Autovalores

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	5.28	48.00	48.00
Dim.2	3.59	32.59	80.59
Dim.3	0.98	8.87	89.46
Dim.4	0.51	4.60	94.06
Dim.5	0.32	2.90	96.96
Dim.6	0.12	1.09	98.05
Dim.7	0.10	0.93	98.98
Dim.8	0.08	0.71	99.70
Dim.9	0.02	0.22	99.92
Dim.10	0.01	0.07	99.99
Dim.11	0.00	0.01	100.00

```
fviz_eig(fit,addlabels=TRUE)
```



5. 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)
```

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

```
knitr::kable(fit$svd$V,  digits =3,caption = "Autovectores")
```

Autovectores

0.215	0.363	0.282
0.317	0.297	-0.229
-0.156	0.271	0.758
0.398	-0.073	0.250
-0.364	0.235	0.089
0.146	0.383	-0.215
0.377	0.181	-0.151
0.034	0.505	0.042
-0.322	0.341	-0.135
-0.302	0.292	-0.366
0.423	0.097	-0.009

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

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

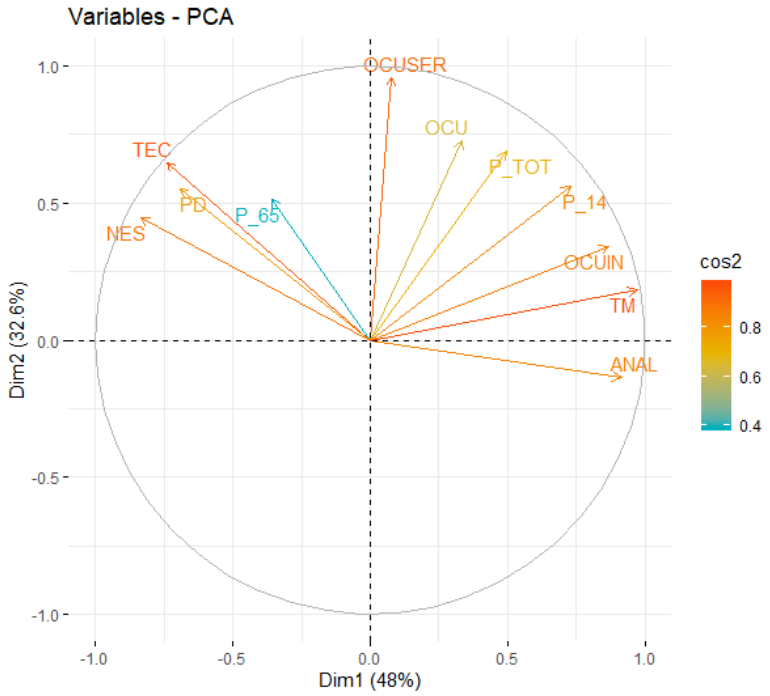
```
#Guardamos los estadísticos asociados a las variables en el objeto var

var<-get_pca_var(fit)

knitr::kable(var$cor,  digits =2,caption = "Correlaciones de la CP con
las variables")
```

Correlaciones de la CP con las variables

	Dim.1	Dim.2	Dim.3
P_TOT	0.50	0.69	0.28
P_14	0.73	0.56	-0.23
P_65	-0.36	0.51	0.75
ANAL	0.91	-0.14	0.25
NES	-0.84	0.44	0.09
OCU	0.34	0.73	-0.21
OCUIN	0.87	0.34	-0.15
OCUSER	0.08	0.96	0.04
TEC	-0.74	0.65	-0.13
PD	-0.69	0.55	-0.36
TM	0.97	0.18	-0.01

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

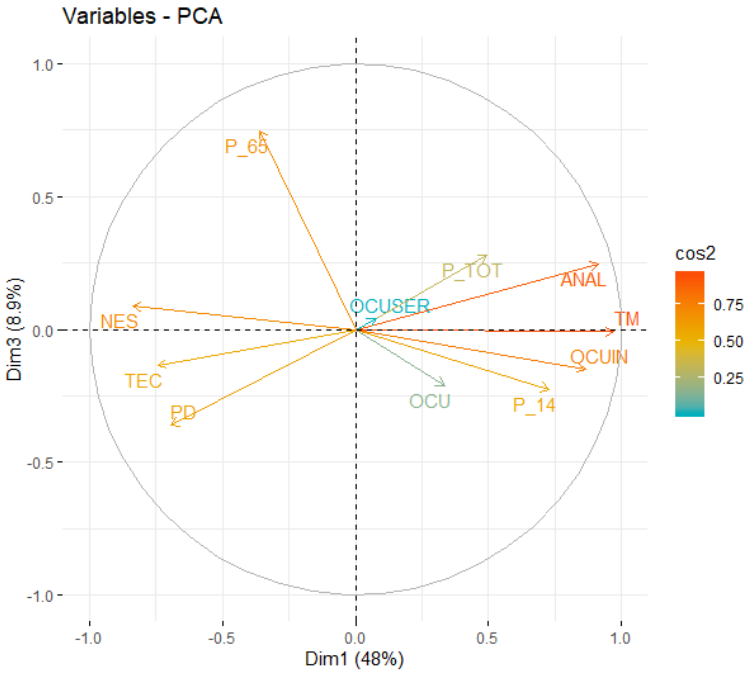
```
# Representación gráfica variables
fviz_pca_var(fit, axes = c(1, 2), col.var="cos2", gradient.cols = c("#
00AFBB", "#E7B800", "#FC4E07"),repel = TRUE )
```



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.

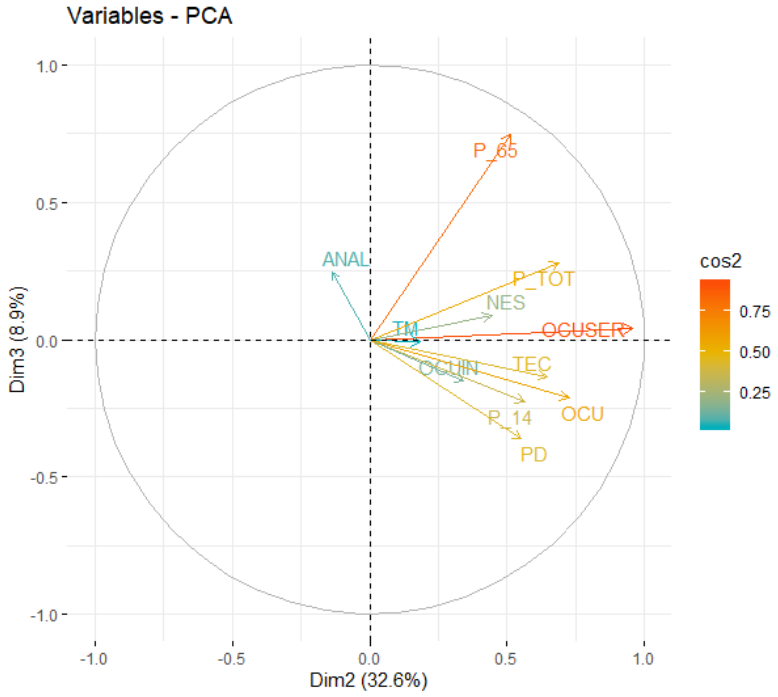
La componente 2 representa a la variable Número de ocupados en servicios, Número de Ocupados y Población Total

```
fviz_pca_var(fit, axes = c(1,3), col.var="cos2", gradient.cols = c("#0
0AFBB", "#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("#0
0AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE )
```



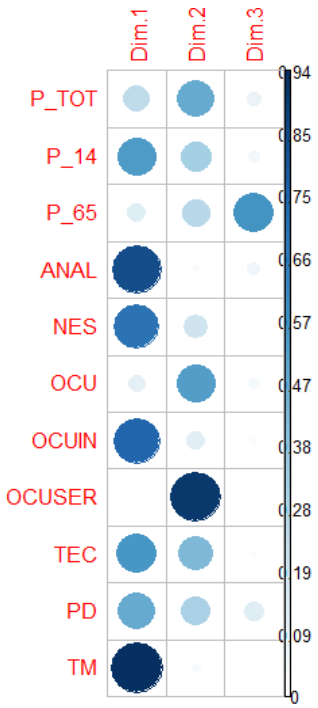
d. Mostrar la tabla y los gráficos que nos muestran la proporción de la varianza de cada variable que es explicado por cada componente. ¿Cuál de las variables es la que está peor explicada?

```
knitr::kable(var$cos2,  digits =2,caption = "Cosenos al cuadrado")
```

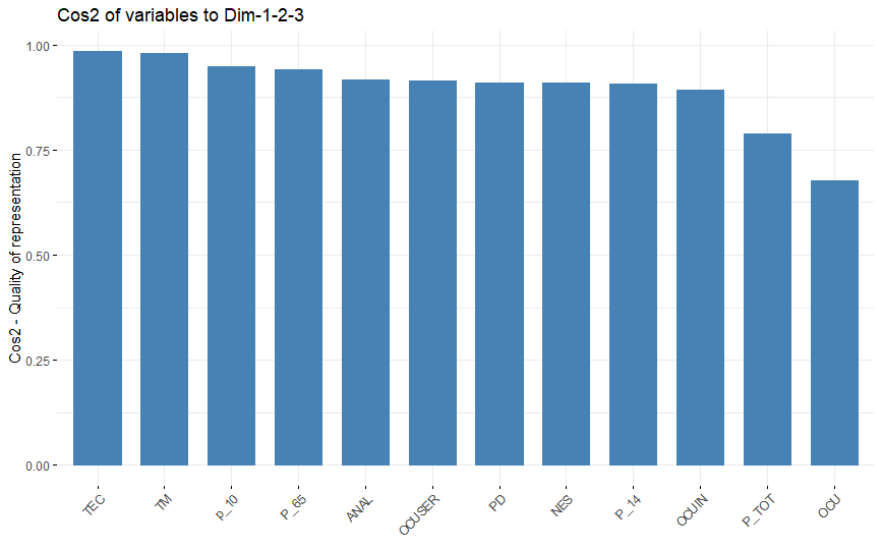
Cosenos al cuadrado

	Dim.1	Dim.2	Dim.3
P_TOT	0.25	0.47	0.08
P_14	0.53	0.32	0.05
P_65	0.13	0.26	0.56
ANAL	0.83	0.02	0.06
NES	0.70	0.20	0.01
OCU	0.11	0.53	0.04
OCUIN	0.75	0.12	0.02
OCUSER	0.01	0.91	0.00
TEC	0.55	0.42	0.02
PD	0.48	0.31	0.13
TM	0.94	0.03	0.00

```
 # Representación gráfica de los cosenos
corrplot(var$cos2,is.corr=FALSE)
```




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



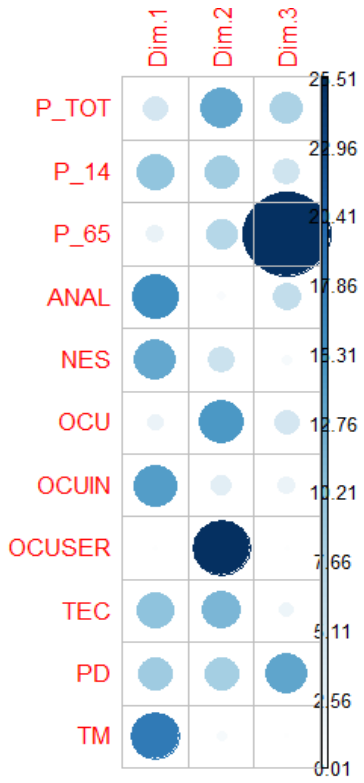
e. Mostrar la tabla y los gráficos que nos muestran el porcentaje de la varianza de cada Componente que es debido a cada variable. ¿Qué variables contribuyen más a cada Componente?

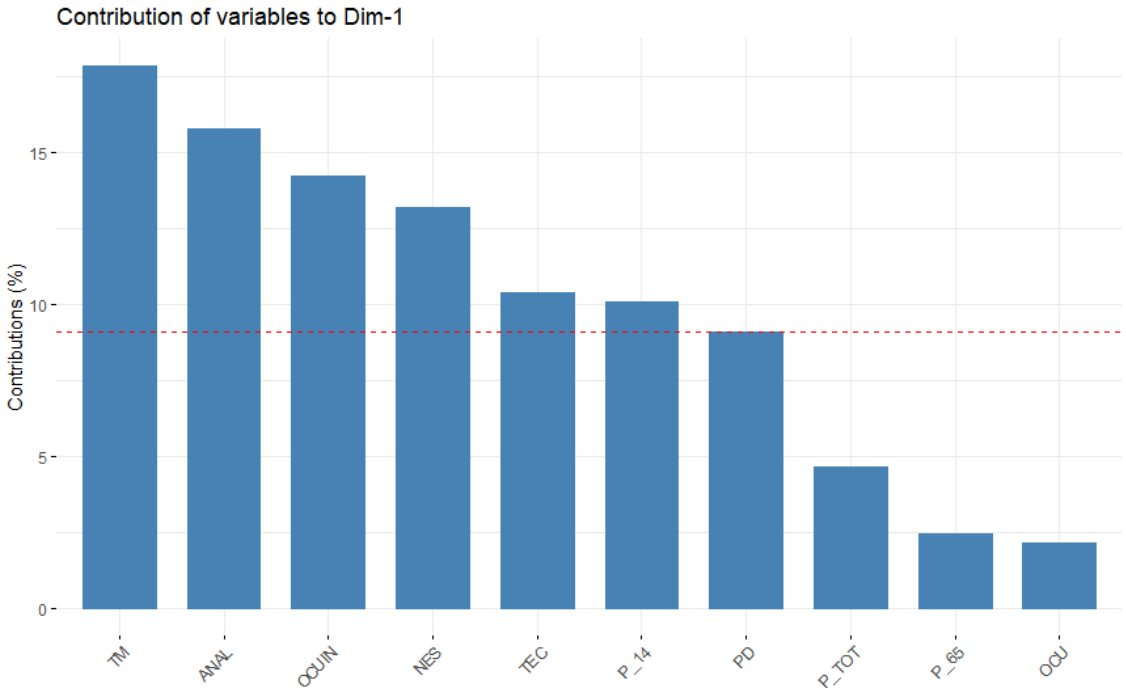
```
knitr::kable(var$contrib,  digits =2,caption = "Contribuciones")
```

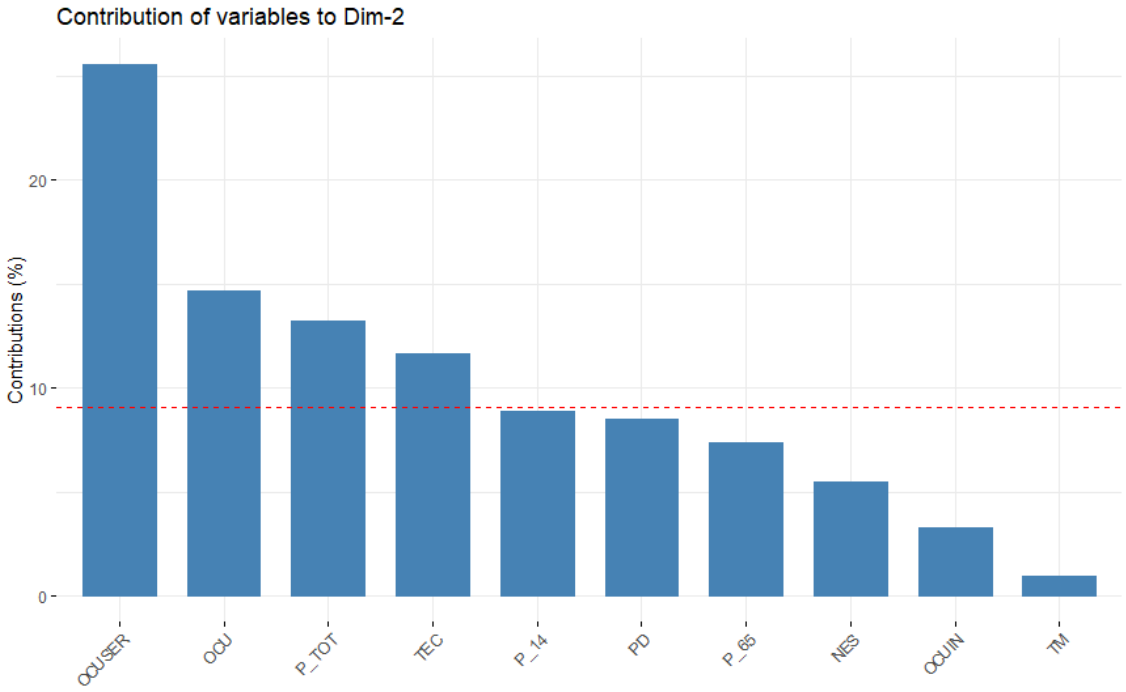
Contribuciones

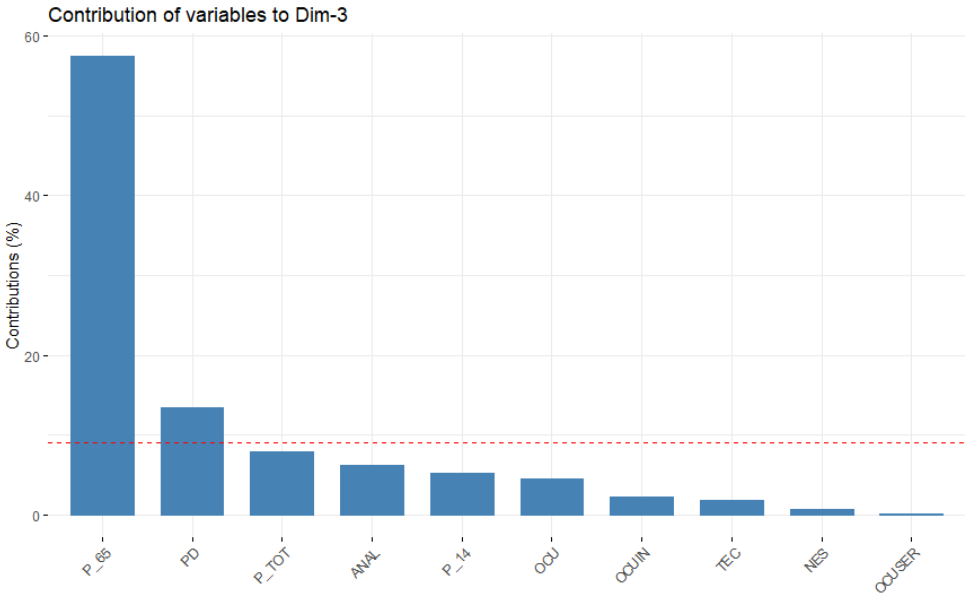
	Dim.1	Dim.2	Dim.3
P_TOT	4.64	13.19	7.96
P_14	10.07	8.84	5.26
P_65	2.44	7.37	57.42
ANAL	15.81	0.53	6.24
NES	13.22	5.50	0.79
OCU	2.13	14.68	4.61
OCUIN	14.22	3.26	2.29
OCUSER	0.11	25.51	0.17
TEC	10.40	11.65	1.83
PD	9.10	8.52	13.42
TM	17.87	0.94	0.01

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


Dim.1
Dim.2
Dim.3
P_TOT
P_14
P_65
ANAL
NES
OCU
OCUIN
OCUSER
TEC
PD
TM
25.51
22.96
20.41
17.86
15.31
12.76
10.21
7.66
5.11
2.56
0.01

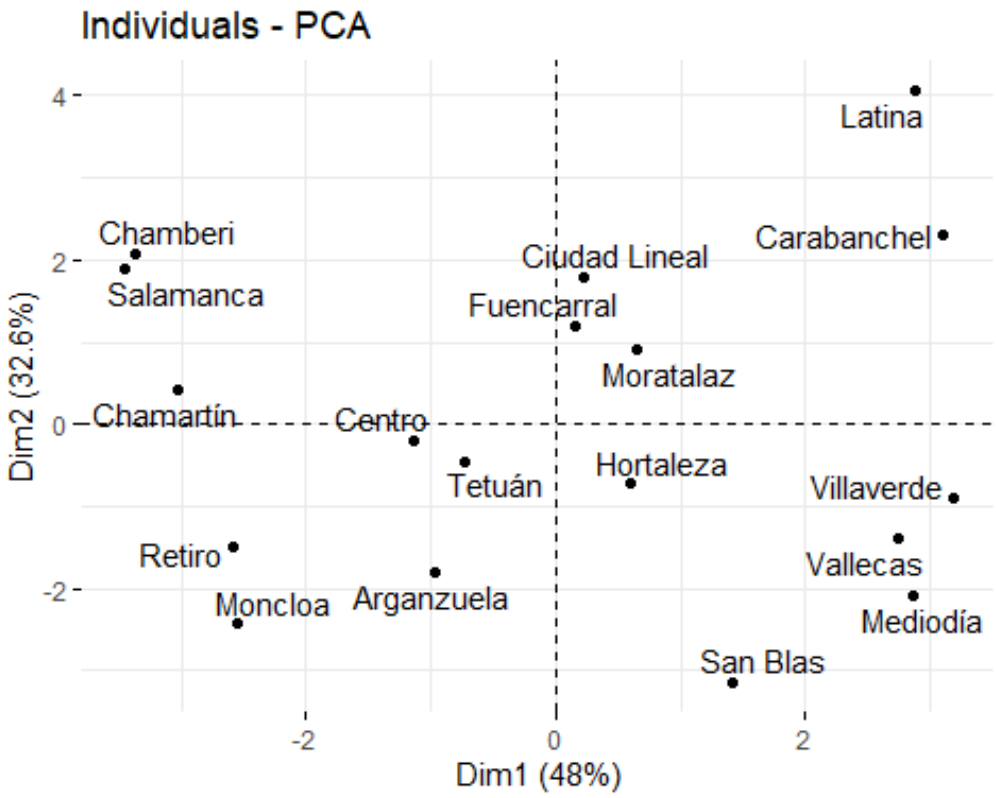

Contribution of variables to Dim-1
Contributions (%)
0
5
10
15
TM
ANAL
OCUIN
NES
TEC
P_14
PD
P_TOT
P_65
OCU



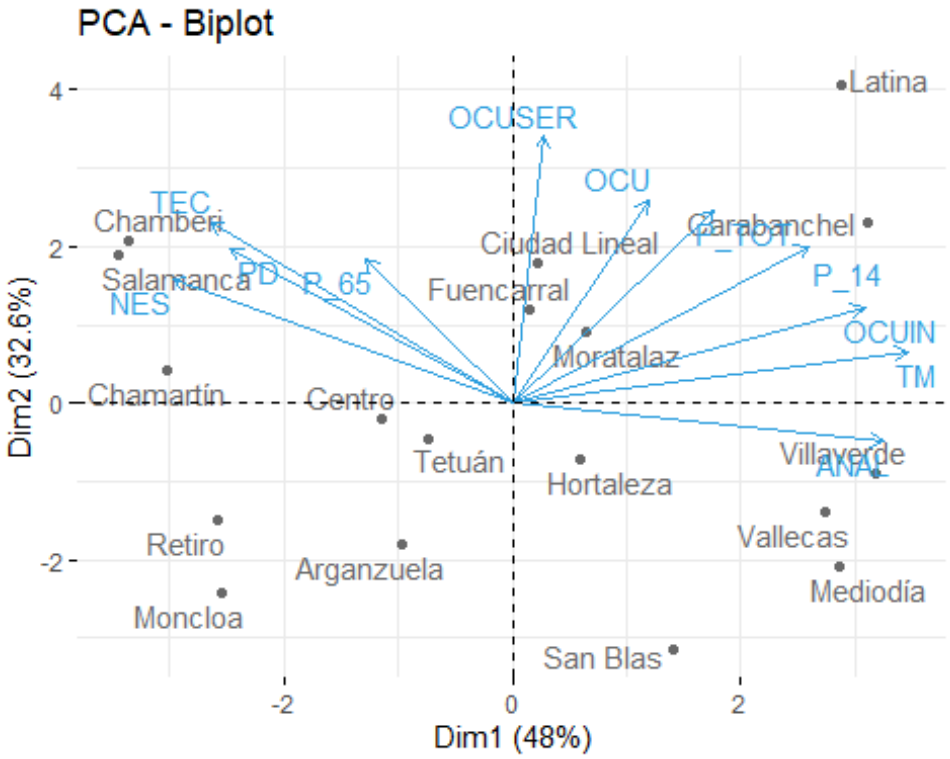


f. Sobre los gráficos que representan las observaciones en los nuevos ejes, y ayudándonos del biplot, teniendo en cuenta la posición de los barrios en el gráfico. Comentar las características socioeconómicas de algunos grupos de barrios

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



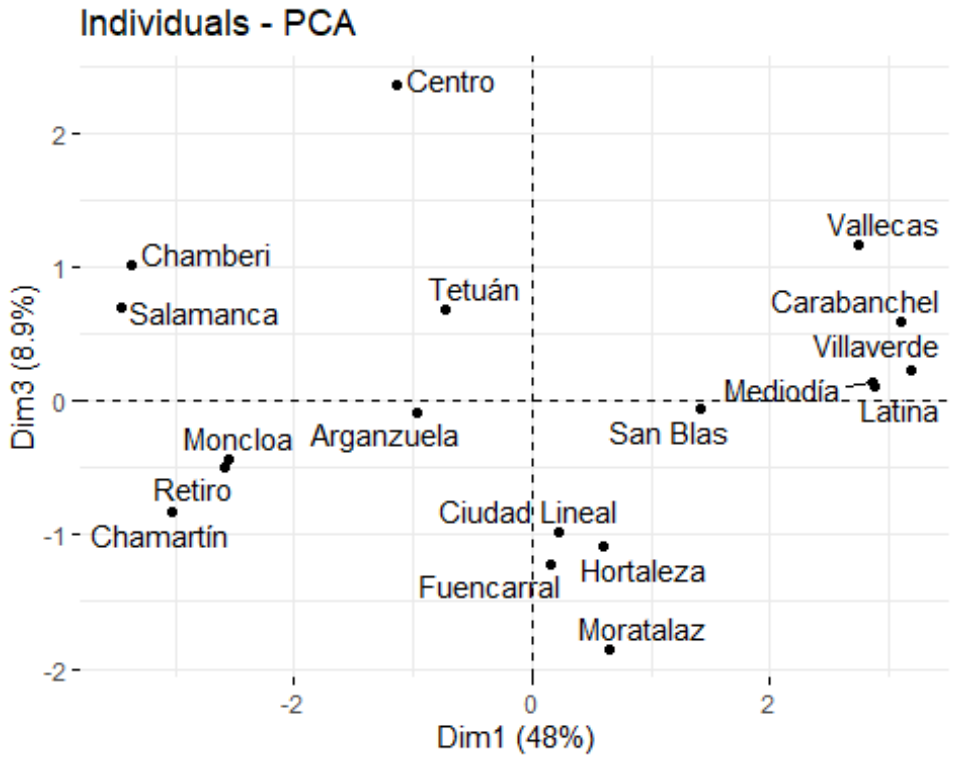

```r
#Representación conjunta de los individuos y las variables en los plan
os de las CP

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



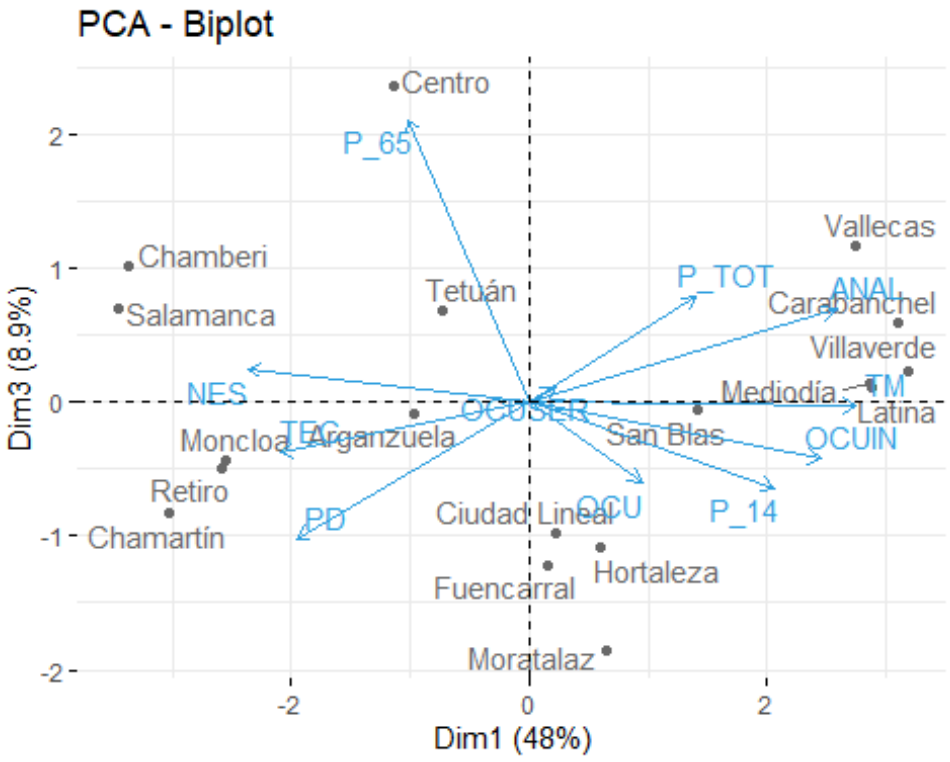
Vemos que Villaverde, Vallecas y Mediodia tienen un comportamiento similar. Tienen un valor alto de la CP1 lo que significa alto porcentaje de personas qe trabajan en trabajos manuales y alto porcentaje de analfabetismo. Mientras que Chamberi y Salamanca también son similares pero en este caso tienen un alto porcentaje de población con nivel de estudios superiores y trabajadores en puestos directivos.

Para ver el comportamiento de los barrios en la Componente 3 representamos los siguienets gráficos.

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




```
fviz_pca_biplot(fit, repel = TRUE,axes = c(1, 3), col.var = "#2E9FDF",
col.ind = "#696969")
```



g) Qué valor tiene Salamanca en la Componente 2?, ¿Y Villaverde?, ¿Qué barrio tiene un valor más alto de la Componente 3?

```
ind<-get_pca_ind(fit)
knitr::kable(ind$coord,  digits =3,caption = "Valores de los individuos en las Cp")
```

Valores de los individuos en las Cp

	Dim.1	Dim.2	Dim.3
Centro	-1.144	-0.197	2.363
Arganzuela	-0.963	-1.798	-0.079
Retiro	-2.581	-1.486	-0.488
Salamanca	-3.459	1.893	0.698
Chamartín	-3.023	0.410	-0.823
Tetuán	-0.731	-0.447	0.686
Chamberi	-3.367	2.083	1.021
Fuencarral	0.154	1.191	-1.215
Moncloa	-2.555	-2.437	-0.428
Latina	2.880	4.062	0.105
Carabanchel	3.105	2.306	0.593
Villaverde	3.178	-0.899	0.234
Mediodía	2.861	-2.093	0.138

Vallecas	2.749	-1.399	1.170
Moratalaz	0.653	0.901	-1.857
Ciudad Lineal	0.229	1.787	-0.979
San Blas	1.414	-3.148	-0.054
Hortaleza	0.601	-0.730	-1.087