

# Laboratorio #1

Paul Belches, José Cifuentes, Oscar Juárez

7/23/2020

## Análisis de los componentes principales

Primero, debemos definir solo las variables cuantitativas útiles de nuestro set de datos

```
datosNumericos <- dataSet[,c("MSSubClass", "LotArea", "OverallQual", "OverallCond", "YearBuilt", "YearRemodAd", "TotalBsmtSF", "TotalFtSq", "LowQualFinSF", "HighQualFinSF", "WoodDeckSF", "OpenPorchSF", "EnclosedPorch", "3SeasonPorch", "ScreenedPorch", "YrSold", "MOSold", "YrMoSold", "SaleType", "SaleCondition", "SalePrice")]
datosNumericos[is.na(datosNumericos)] <- 0
```

Ahora analizamos si es posible utilizar el análisis factorial para formar las combinaciones lineales de las variables.

```
pafDatos<-paf(as.matrix(datosNumericos))
pafDatos$KMO
```

```
## [1] 0.7318
```

```
pafDatos$Bartlett
```

```
## [1] 21612
```

```
#summary(pafDatos)
```

Obtenemos un **KMO** de *0.73*, lo cual significa una aceptable adecuación muestral. Por otra parte, el **Bartlett** es de *21612* lo cual es un valor alto indicando menos probabilidad que la matriz sea una matriz identidad.

También debemos validar el **nivel de significación** de la prueba

```
cortest.bartlett(datosNumericos[, -1])
```

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

```
## $chisq
## [1] 21045
##
## $p.value
## [1] 0
##
## $df
## [1] 435
```

El valor p es de 0, dado que es mayor a 0.05, el análisis factorial podría no funcionar. Vale la pena mostrar la matriz de correlación.

```
cor(datosNumericos[, -1], use = "pairwise.complete.obs")
```

Del último resultado, cabe destacar los **15 valores con mayor correlación** entre ellos.

VARIABLE 1	VARIABLE 2	CORRELACIÓN
GarageCars	GarageArea	0.882
TotRmsAbvGrd	GrLivArea	0.825
X2ndFlrSF	GrLivArea	0.687
BedroomAbvGr	TotRmsAbvGrd	0.676
BsmtFullBath	BsmtFinSF1	0.649
FullBath	GrLivArea	0.63
TotRmsAbvGrd	X2ndFlrSF	0.616
HalfBath	X2ndFlrSF	0.611
GarageCars	OverallQual	0.6
YearRemodAdd	YearBuilt	0.593
YearBuilt	OverallQual	0.572
GarageArea	OverallQual	0.562
YearRemodAdd	OverallQual	0.551
GarageCars	YearBuilt	0.538
BedroomAbvGr	GrLivArea	0.521

Podemos notar una correlación significativa entre la cantidad de carros que caben en un garage, junto con el área de esta. Además, parece haber correlación entre el total de habitaciones en un segundo nivel junto con el GrLivArea.

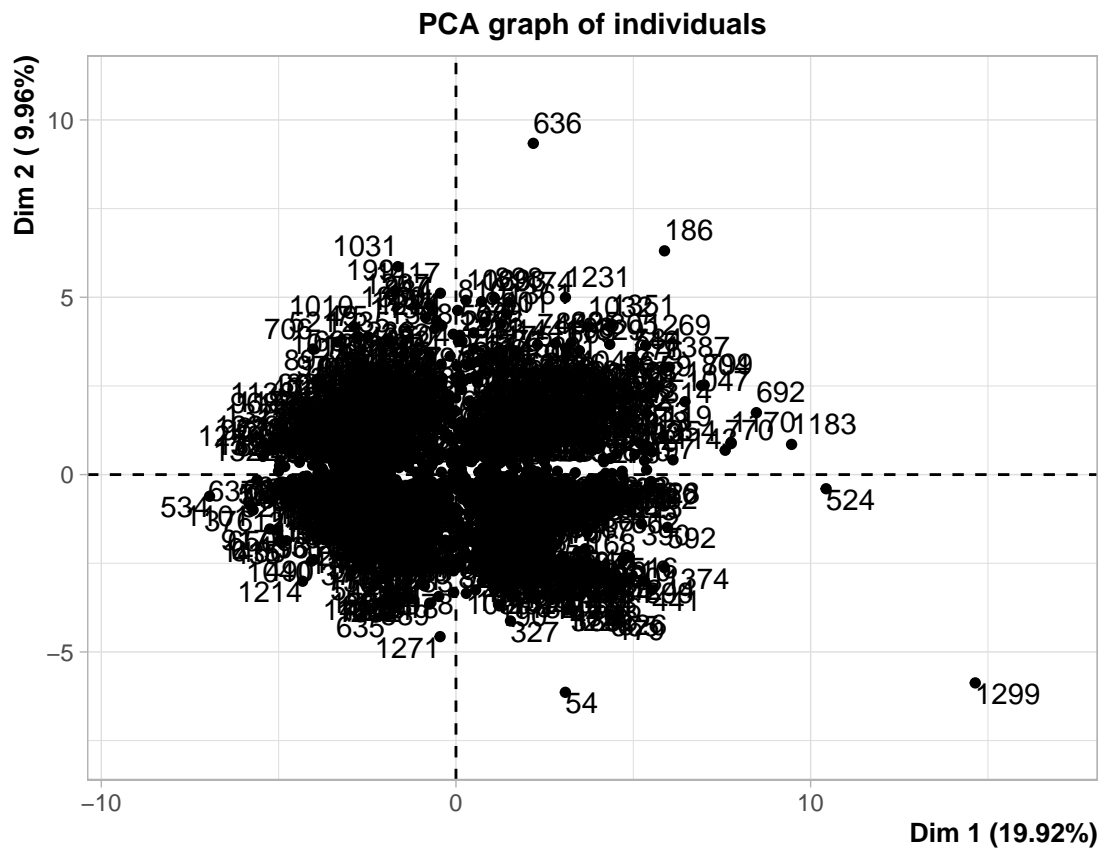
Normalizamos los datos y obtenemos un resumen de los componentes.

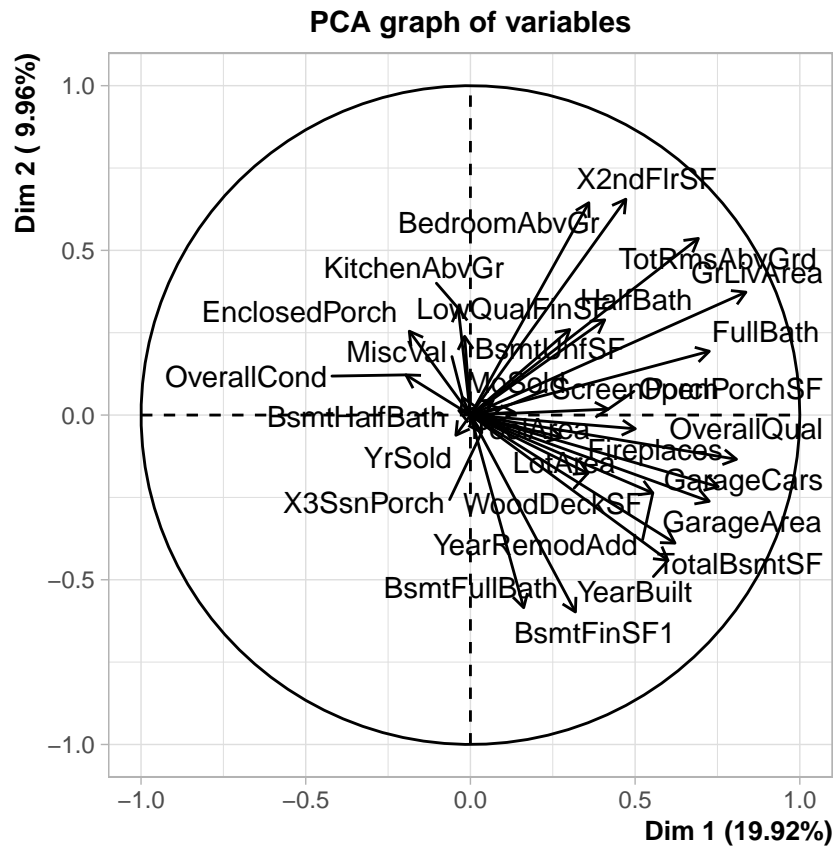
```
compPrinc<-prcomp(datosNumericos, scale = T)
summary(compPrinc)
```

```
## Importance of components:
##
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7      PC8
## Standard deviation  2.445  1.7423  1.4373  1.3191  1.1974  1.0875  1.071  1.0465
## Proportion of Variance 0.193  0.0979  0.0666  0.0561  0.0462  0.0382  0.037  0.0353
## Cumulative Proportion 0.193  0.2907  0.3574  0.4135  0.4597  0.4979  0.535  0.5702
##
##          PC9      PC10     PC11     PC12     PC13     PC14     PC15     PC16
## Standard deviation  1.0291  1.0096  1.0060  0.9931  0.9846  0.9423  0.9332  0.9136
## Proportion of Variance 0.0342  0.0329  0.0326  0.0318  0.0313  0.0286  0.0281  0.0269
## Cumulative Proportion 0.6044  0.6373  0.6699  0.7017  0.7330  0.7616  0.7897  0.8167
##
##          PC17     PC18     PC19     PC20     PC21     PC22     PC23     PC24
## Standard deviation  0.8949  0.8724  0.8388  0.7887  0.7660  0.6888  0.6287  0.55076
## Proportion of Variance 0.0258  0.0245  0.0227  0.0201  0.0189  0.0153  0.0127  0.00979
## Cumulative Proportion 0.8425  0.8670  0.8898  0.9098  0.9287  0.9440  0.9568  0.96658
##
##          PC25     PC26     PC27     PC28     PC29     PC30     PC31
## Standard deviation  0.51748  0.50757  0.40124  0.3817  0.31936  0.25552  0.19134
## Proportion of Variance 0.00864  0.00831  0.00519  0.0047  0.00329  0.00211  0.00118
## Cumulative Proportion 0.97522  0.98353  0.98872  0.9934  0.99671  0.99882  1.00000
```

Ahora podemos darnos una idea del comportamiento con los siguientes dos gráficos

```
compPrincPCA<-PCA(datosNumericos[,-1],ncp=ncol(datosNumericos[,-1]), scale.unit = T)
```

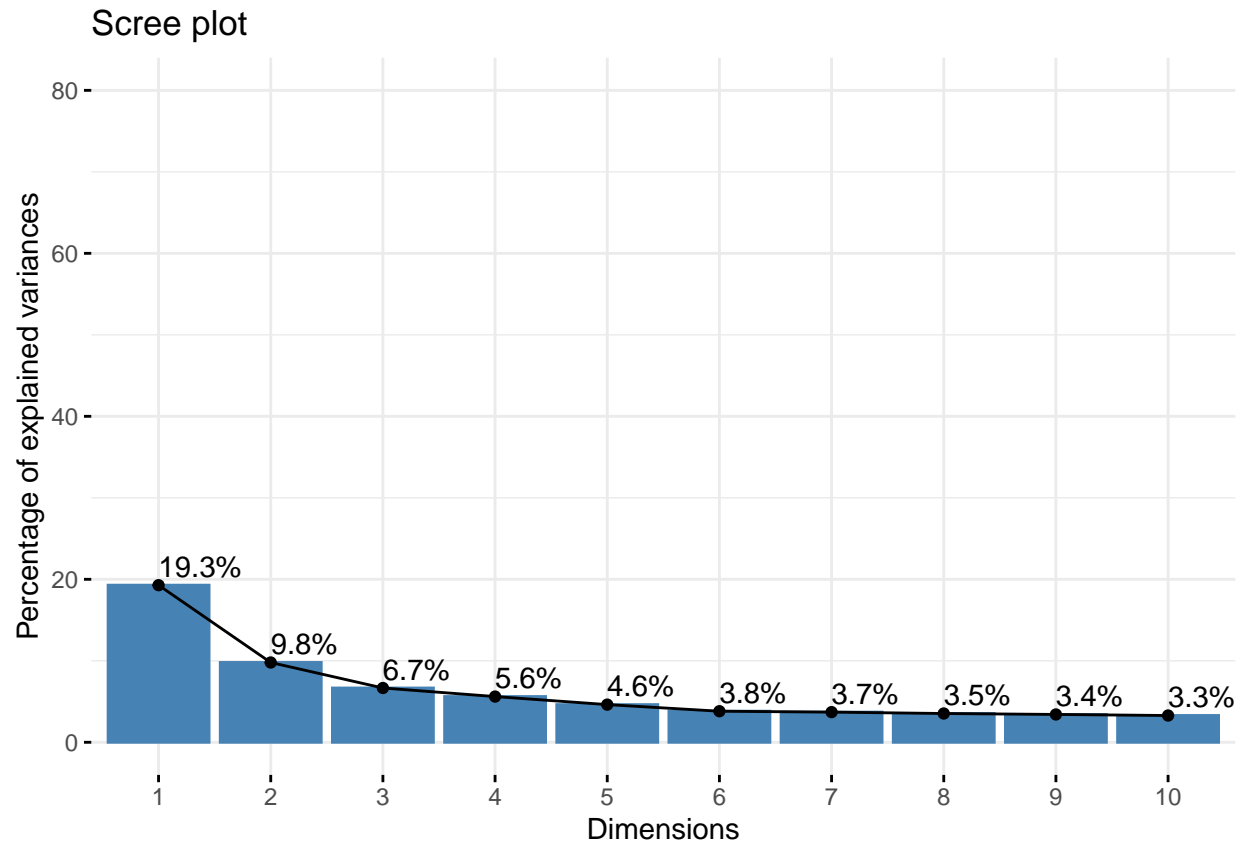




```
# summary(compPrincPCA)
```

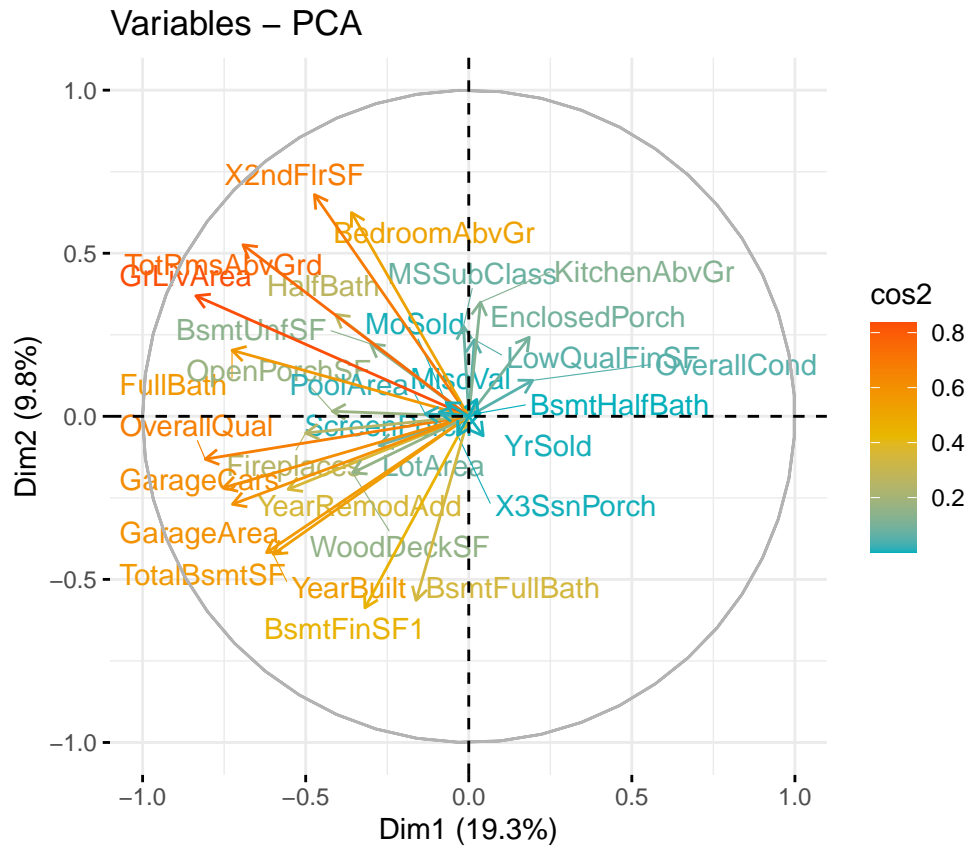
Hacemos plot de los componentes principales.

```
fviz_eig(compPrinc, addlabels = TRUE, ylim = c(0, 80))
```



Ahora, se muestra la calidad de la representación de los componentes respecto a las dos primeras dimensiones.

```
# En la siguiente gráfica se ilustra la calidad de la representación de los componentes en las dos pr
fviz_pca_var(compPrinc, col.var = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE # Avoid text overlapping
)
```



Hacemos matriz de correlación con los datos.

```
var<-get_pca_var(compPrinc)
corrplot(var$cos2, is.corr = F)
```

