Minería de datos: PRA1 - Selección y preparación de un juego de datos

Autor: Noelia Pérez Benavent

MAYO 2023

Contents

Introducción

Presentación

En esta práctica abordamos un caso real de minería de datos donde tenemos que poner en juego todos los conceptos trabajados en la asignatura. Hay que trabajar todo el **ciclo de vida del proyecto**, desde el objetivo del proyecto hasta la implementación del conocimiento encontrado, pasando por la preparación, limpieza de los datos, conocimiento de los datos, generación del modelo, interpretación y evaluación. La práctica la dividiremos en dos partes. En esta primera parte (PRA1), abordaremos las primeras fases del proceso, desde los objetivos hasta la preparación de los datos, y en la segunda parte (PRA2) seguiremos con el resto del proceso.

Objetivos

El objetivo global de esta primera parte de la **práctica (PRA1)** consiste en seleccionar uno o varios juegos de datos, y realizar las tareas de **preparación y análisis exploratorio** con el objetivo de disponer de datos listos para después. En la segunda parte (PRA2), **aplicar algoritmos** de clustering, regresión o clasificación, demostrando la correcta asimilación de todos los aspectos trabajados durante el semestre.

Competencias

Las competencias que se trabajan en esta prueba son:

- Uso y aplicación de las TIC en el ámbito académico y profesional.
- Capacidad para innovar y generar nuevas ideas.
- Capacidad para evaluar soluciones tecnológicas y elaborar propuestas de proyectos teniendo en cuenta los recursos, las alternativas disponibles y las condiciones de mercado.
- Conocer las tecnologías de comunicaciones actuales y emergentes así como saberlas aplicar convenientemente para diseñar y desarrollar soluciones basadas en sistemas y tecnologías de la información.
- Aplicación de las técnicas específicas de ingeniería del software en las diferentes etapas del ciclo de vida de un proyecto.
- Capacidad para aplicar las técnicas específicas de tratamiento, almacenamiento y administración de datos.
- Capacidad para proponer y evaluar diferentes alternativas tecnológicas para resolver un problema concreto.

Recursos Básicos

Material docente proporcionado por la UOC.

Formato de entrega PRA_1

El formato de entrega es: - *fichero output*: en formato username_estudiante-PRA1.html/pdf - *fichero ejecutable*: en formato username_estudiante-PR1.Rmd

Se debe entregar la PRA_1 en el buzón de entregas del aula

Nota: Propiedad intelectual

> A menudo es inevitable, al producir una obra multimedia, hacer uso de recursos creados por terceras personas. Es por lo tanto comprensible hacerlo en el marco de una práctica de los estudios de Informática, Multimedia y Telecomunicación de la UOC, siempre y cuando esto se documente claramente y no suponga plagio en la práctica.
>
> Por lo tanto, al presentar una práctica que haga uso de recursos ajenos, se debe presentar junto con ella un documento en que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo y su estatus legal: si la obra esta protegida por el copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia GNU, GPL ...). El estudiante deberá asegurarse de que la licencia no impide específicamente su uso en el marco de la práctica. En caso de no encontrar la información correspondiente tendrá que asumir que la obra esta protegida por copyright.
>
> Deberéis, además, adjuntar los ficheros originales cuando las obras utilizadas sean digitales, y su código fuente si corresponde.

Enunciado

Todo estudio analítico debe nacer de una necesidad por parte del **negocio** o de una voluntad de dotarle de un conocimiento contenido en los datos y que solo podremos obtener a través de una colección de buenas prácticas basadas en la Minería de Datos.

El mundo de la analítica de datos se sustenta en 3 ejes:

A. Uno de ellos es el profundo **conocimiento** que deberíamos tener **del negocio** al que tratamos de dar respuestas mediante los estudios analíticos.

B. El otro gran eje es sin duda las **capacidades analíticas** que seamos capaces de desplegar y en este sentido, las dos prácticas de esta asignatura pretenden que el estudiante realice un recorrido sólido por este segundo eje.

C. El tercer eje son los **Datos**. Las necesidades del Negocio deben concretarse con preguntas analíticas que a su vez sean viables responder a partir de los datos de que disponemos. La tarea de analizar los datos es sin duda importante, pero la tarea de identificarlos y obtenerlos va a ser para un analista un reto permanente.

Como **primera parte** del estudio analítico que nos disponemos a realizar, se pide al estudiante que complete los siguientes pasos:

1. Plantear un problema de analítica de datos detallando los objetivos analíticos y explica una metodología para resolverlos de acuerdo con lo que se ha practicado en las PEC y lo que se ha aprendido en el material didáctico.
2. Seleccionar un juego de datos y justificar su elección. El juego de datos **deberá tener capacidades** para que se le puedan aplicar **algoritmos supervisados**, **algoritmos no supervisados** y **reglas de asociación** y deberá estar alineado con el problema analítico planteado en el paso anterior.

Requisito mínimo: El juego de datos deberá tener como mínimo 500 observaciones con un mínimo de 5 variables numéricas, 2 categóricas y 1 binaria. Además **debe ser distinto**, es importante que no sea un dataset usado en las PEC anteriores.

3. Realizar un análisis exploratorio del juego de datos seleccionado.
4. Realizar tareas de limpieza y acondicionado para poder ser usado en procesos de modelado.
5. Realizar métodos de discretización
6. Aplicar un estudio PCA sobre el juego de datos. A pesar de no estar explicado en el material didáctico, se valorará si en lugar de PCA investigáis por vuestra cuenta y aplicáis SVD (Single Value Decomposition).

- **Algunos recursos**
  - PCA para reducción de dimensiones
  - SVD Singular Value Decomposition

Respuesta a los ejercicios

Ejercicio 1

El objetivo de este análisis de datos es identificar los factores más fuertemente asociados con la mortalidad en pacientes con cáncer de mama y desarrollar un modelo predictivo basado en los datos de diagnóstico disponibles para predecir la mortalidad en estos pacientes.

Para lograr estos objetivos, se seguirá una metodología que comienza por la verificación de datos para asegurarse de que no haya errores o valores faltantes que puedan afectar el análisis. Luego, se llevará a cabo el preprocesamiento de los datos, lo que incluirá la limpieza de los datos, la transformación de variables y la selección de características.

La selección de características implicará la eliminación de las columnas que no son relevantes para el análisis o que contienen demasiados valores faltantes (missing values NA). También se transformarán las variables categóricas en variables numéricas para que puedan ser procesadas por los algoritmos de aprendizaje automático, y se discretizarán las variables continuas si es necesario para mejorar la precisión del análisis.

Posteriormente, se llevará a cabo un análisis exploratorio de datos para comprender mejor las relaciones entre las variables y buscar patrones en los datos. Este análisis incluirá el uso de técnicas de visualización de datos para identificar patrones y tendencias en los datos, así como el análisis de correlación para identificar las variables más fuertemente asociadas con la mortalidad.

Además, se aplicará la técnica de análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos y extraer características significativas que puedan ser útiles para el análisis y la modelización. En lugar de PCA, se podría aplicar la técnica de Single Value Decomposition (SVD) para reducir la dimensionalidad de los datos y extraer características significativas que puedan ser útiles para el análisis y la modelización.

Asimismo, se aplicará la técnica de K-means clustering para agrupar los pacientes en diferentes grupos en función de sus características, lo que permitirá identificar patrones en los datos y explorar las relaciones entre las variables.

En la siguiente etapa, se aplicarán técnicas de aprendizaje automático no supervisado para identificar patrones en los datos y explorar las relaciones entre las variables. Por último, se desarrollará un modelo predictivo que pueda predecir la supervivencia y la mortalidad en pacientes con cáncer de mama.

Este modelo se desarrollará utilizando un conjunto de algoritmos de aprendizaje automático supervisado, como la regresión logística, el árbol de decisión y el random forest, y se ajustarán los modelos utilizando los datos de entrenamiento.

Finalmente, se seleccionará el mejor modelo en función de su rendimiento y se utilizará para predecir la mortalidad y supervivencia en pacientes con cáncer de mama en función de sus características de diagnóstico disponibles.

Ejercicio 2

Se ha seleccionado el dataset “Breast Cancer” de Kaggle (https://www.kaggle.com/datasets/reihanenamdari/breast-cancer). Es un conjunto de datos de pacientes con cáncer de mama obtenido del Programa SEER del NCI, que proporciona información sobre estadísticas de cáncer basadas en la población. El conjunto de datos incluye a pacientes de sexo femenino con cáncer de mama de carcinoma lobulillar y de conducto infiltrante que ha sido diagnosticado entre 2006 y 2010.

Los motivos de elección de este dataset fueron que el cáncer de mama es una de las enfermedades más prevalentes y mortales en mujeres en todo el mundo, y es de gran interés para la investigación médica. La exploración y análisis de los datos relacionados con esta enfermedad pueden proporcionar información valiosa para la mejora de su detección y tratamiento. Además, este conjunto de datos se obtuvo de una fuente confiable y actualizada, el Programa SEER del NCI, lo que garantiza su calidad y su validez científica.

El dataset “Breast cancer” se eligió principalmente por su potencial analítico ya que, contiene información sobre pacientes con cáncer de mama y características asociadas como la edad, la etapa del tumor, el tamaño del tumor, el estado del receptor hormonal, el número de ganglios linfáticos regionales examinados y positivos y la supervivencia (status: dead or alive).

Dado que este dataset tiene más de 4000 registros y 16 variables (múltiples variables numéricas y categóricas), se pueden aplicar varias técnicas de análisis de datos y aprendizaje automático. Para explorar los patrones y

relaciones en los datos, se pueden utilizar técnicas de análisis exploratorio de datos, como la visualización de datos, el análisis de correlación y el análisis de componentes principales (PCA).

Por ejemplo, mediante análisis de correlación se puede analizar la relación entre diferentes variables, como el tamaño del tumor y el estado de los receptores de estrógeno y progesterona, para determinar si existe alguna relación significativa entre ellas.

Para la clasificación de pacientes según su status (dead or alive), se pueden aplicar modelos supervisados como regresión logística, árboles de decisión. Además, se pueden aplicar técnicas de selección de características para determinar qué variables tienen mayor impacto en la supervivencia de los pacientes.

También se pueden aplicar técnicas de agrupamiento, como el algoritmo k-means, para clasificar a los pacientes en grupos similares en función de sus características. Esto puede ayudar a identificar subgrupos de pacientes que pueden tener diferentes pronósticos o resultados (outcomes).

Ejercicio 3

Análisis exploratorio

Primero vamos a instalar y cargas las librerías ggplot2 y dplyr

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
```

```
## Loading required package: ggplot2
```

```
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Ahora cargamos nuestro juego de datos “Breast Cancer”.

```
BreastCancer <- read.csv('Breast_Cancer.CSV', row.names=NULL)
```

Vamos a observar la estructura de nuestro dataset y sus variables.

```
structure = str(BreastCancer)
```

```
## 'data.frame':    4024 obs. of  16 variables:
##  $ Age                   : int  68 50 58 58 47 51 51 40 40 69 ...
##  $ Race                  : chr  "White" "White" "White" "White" ...
##  $ Marital.Status        : chr  "Married" "Married" "Divorced" "Married" ...
##  $ T.Stage               : chr  "T1" "T2" "T3" "T1" ...
##  $ N.Stage               : chr  "N1" "N2" "N3" "N1" ...
##  $ X6th.Stage            : chr  "IIA" "IIIA" "IIIC" "IIA" ...
##  $ differentiate         : chr  "Poorly differentiated" "Moderately differentiated" "Moderately diff
##  $ Grade                 : chr  "3" "2" "2" "3" ...
##  $ A.Stage               : chr  "Regional" "Regional" "Regional" "Regional" ...
##  $ Tumor.Size            : int  4 35 63 18 41 20 8 30 103 32 ...
##  $ Estrogen.Status       : chr  "Positive" "Positive" "Positive" "Positive" ...
##  $ Progesterone.Status   : chr  "Positive" "Positive" "Positive" "Positive" ...
```

```
##  $ Regional.Node.Examined: int  24 14 14 2 3 18 11 9 20 21 ...
##  $ Reginol.Node.Positive : int  1 5 7 1 1 2 1 1 18 12 ...
##  $ Survival.Months       : int  60 62 75 84 50 89 54 14 70 92 ...
##  $ Status                : chr  "Alive" "Alive" "Alive" "Alive" ...
```

Observamos que nuestro dataset BreastCancer contiene *4024 registros* y *16 variables*. Vamos a estructurarlas y describirlas para conocer mejor los datos con los que vamos a trabajar.

- Age: edad de la paciente.
- Race: grupo étnico de las paciente (White, Black, Other).
- Marital Status: estado civil de las pacientes (Single, Married, Divorced, Separated, Widowed).
- T. Stage: representa el tamaño y extensión del tumor principal. Contra más grande es el número de la T, más grande y expandido está el tumor principal (T1, T2, T3 y T4).
- N. Stage: repreenta el número y localización de los ganglios linfáticos que contienen cáncer. Cuanto mayor sea el número después de la N, mayor será el número de ganglios linfáticos que contienen cáncer (N1, N2, N3).
- 6th Stage: determina el grupos de estadio del cáncer de mama (IIA, IIB, IIIA, IIIB, IIIC).
- Differentiate: el grado de diferenciación de un tumor (Well differentiated, Moderately differentiated, Poorly differentiated, Undifferentiated).
- Grade: grado del tumor (1, 2, 3, anaplastic; Grade IV).
- A.Stage: (Regional o Distant) Regional: El cáncer se ha extendido fuera de la mama a estructuras o ganglios linfáticos cercanos. Distante: el cáncer se ha extendido a partes distantes del cuerpo, como los pulmones, el hígado o los huesos.
- Tumor Size: el tamaño del tumor en mm.
- Estrogen Status: (Positive o Negative) Estrógeno positivo: Las células cancerosas que son ER positivas pueden necesitar estrógenos para crecer. Estas células pueden dejar de crecer o morir cuando se tratan con sustancias que bloquean la unión y la acción de los estrógenos. También se denomina receptor de estrógeno positivo. Estrógeno negativo: Los cánceres de mama negativos son un grupo de tumores con mal pronóstico y menos estrategias de prevención y tratamiento del cáncer en comparación con los tumores RE positivos.
- Progesterone Status: (Positive o Negative) Progesterona positiva: Este tipo de cáncer de mama es sensible a la progesterona, y las células tienen receptores que les permiten utilizar esta hormona para crecer. El tratamiento con terapia endocrina bloquea el crecimiento de las células cancerosas. Progesterona negativa: Este tipo de cáncer de mama no tiene receptores de estrógeno ni de progesterona. El tratamiento con fármacos de terapia hormonal no es útil para estos cánceres. Estos cánceres tienden a crecer más rápidamente que los cánceres con receptores hormonales positivos.
- Regional Node Examined: número de nodos examinados.
- Regional Node Positive: número de nodos examinados que han dado positivo.
- Survival months: meses de supervivencia.
- Status: estado de vida (Dead o Alive).

Ejercicio 4

Preprocesamiento y limpieza de datos

Primero vamos a obsrvar las características de nuestro dataset y ver que valores toman. las variables.

```
summary(BreastCancer)
```

```
##       Age            Race           Marital.Status       T.Stage
##  Min.   :30.00   Length:4024        Length:4024        Length:4024
##  1st Qu.:47.00   Class :character   Class :character   Class :character
##  Median :54.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :53.97
##  3rd Qu.:61.00
##  Max.   :69.00
```

```
##    N.Stage            X6th.Stage        differentiate         Grade
##  Length:4024        Length:4024        Length:4024        Length:4024
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    A.Stage             Tumor.Size     Estrogen.Status    Progesterone.Status
##  Length:4024        Min.   :  1.00   Length:4024        Length:4024
##  Class :character   1st Qu.: 16.00   Class :character   Class :character
##  Mode  :character   Median : 25.00   Mode  :character   Mode  :character
##                     Mean   : 30.47
##                     3rd Qu.: 38.00
##                     Max.   :140.00
##  Regional.Node.Examined Reginol.Node.Positive Survival.Months
##  Min.   : 1.00          Min.   : 1.000        Min.   :  1.0
##  1st Qu.: 9.00          1st Qu.: 1.000        1st Qu.: 56.0
##  Median :14.00          Median : 2.000        Median : 73.0
##  Mean   :14.36          Mean   : 4.158        Mean   : 71.3
##  3rd Qu.:19.00          3rd Qu.: 5.000        3rd Qu.: 90.0
##  Max.   :61.00          Max.   :46.000        Max.   :107.0
##     Status
##  Length:4024
##  Class :character
##  Mode  :character
##
##
##
```

Vamos a revisar si existe algun valor nulo. Vemos que no es el caso.

```
colSums(is.na(BreastCancer))
```

```
##                    Age                   Race         Marital.Status
##                      0                      0                      0
##                T.Stage                N.Stage             X6th.Stage
##                      0                      0                      0
##          differentiate                  Grade                A.Stage
##                      0                      0                      0
##             Tumor.Size        Estrogen.Status    Progesterone.Status
##                      0                      0                      0
## Regional.Node.Examined  Reginol.Node.Positive        Survival.Months
##                      0                      0                      0
##                 Status
##                      0
```

```
colSums(BreastCancer=="")
```

```
##                    Age                   Race         Marital.Status
##                      0                      0                      0
##                T.Stage                N.Stage             X6th.Stage
##                      0                      0                      0
##          differentiate                  Grade                A.Stage
##                      0                      0                      0
##             Tumor.Size        Estrogen.Status    Progesterone.Status
##                      0                      0                      0
```

```
## Regional.Node.Examined  Reginol.Node.Positive        Survival.Months
##                      0                      0                      0
##                 Status
##                      0
```

Como podemos ver, no encontramos ningun valor nulo o missing values en nuestro dataset. Ahora vamos a crear histogramas para ver los datos de cada atributo.

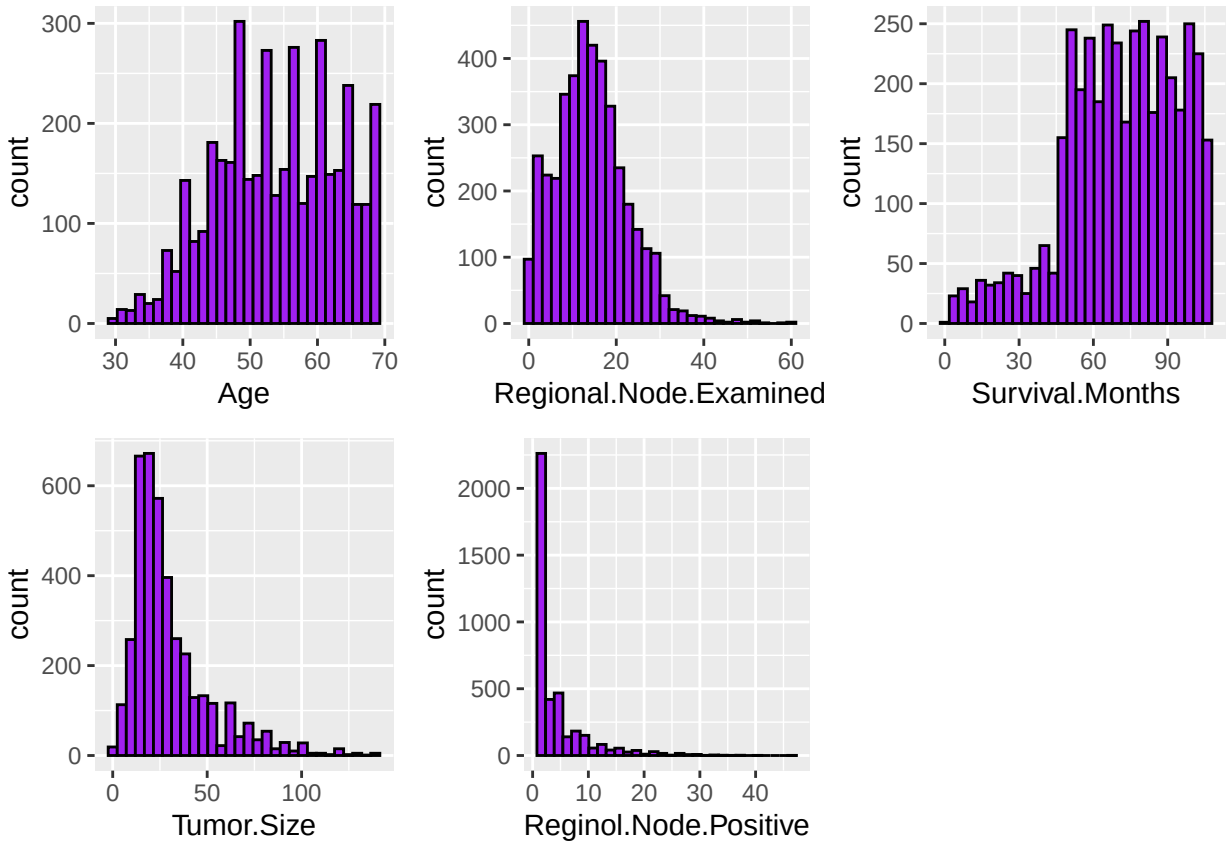
```
if(!require('Rmisc')) install.packages('Rmisc'); library('Rmisc')
```

```
## Loading required package: Rmisc
## Loading required package: lattice
## Loading required package: plyr
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
histList<- list()
n= c(1,10,13,14,15)
BreastCancerAux=BreastCancer %>% select(all_of(n))
for(i in 1:ncol(BreastCancerAux)){
  col <- names(BreastCancerAux)[i]
  ggp <- ggplot(BreastCancerAux, aes_string(x = col)) +
    geom_histogram(bins = 30, fill = "purple", color = "black")
  histList[[i]] <- ggp
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
multiplot(plotlist = histList, cols = 3)
```



Los datos nos muestran que aproximadamente la mayoría de pacientes que componen nuestro dataset se encuentra entre los 45-69 años. Vemos que prácticamente no se contabilizan pacientes menores de 50 años y no hay menores de 30 registradas. La media de edad sería de 54 años (mín 30 - máx 69).

Respecto al número de nódulos examinados vemos como los valores típicos son entre 5-20 nodos (mín 1 - máx 61), siendo la media aproximadamente 15. De ellos, mínimo uno fue positivo y se llegarón a alcanzar 46 nódulos positivos en una paciente (media de nódulos positivos = 4).

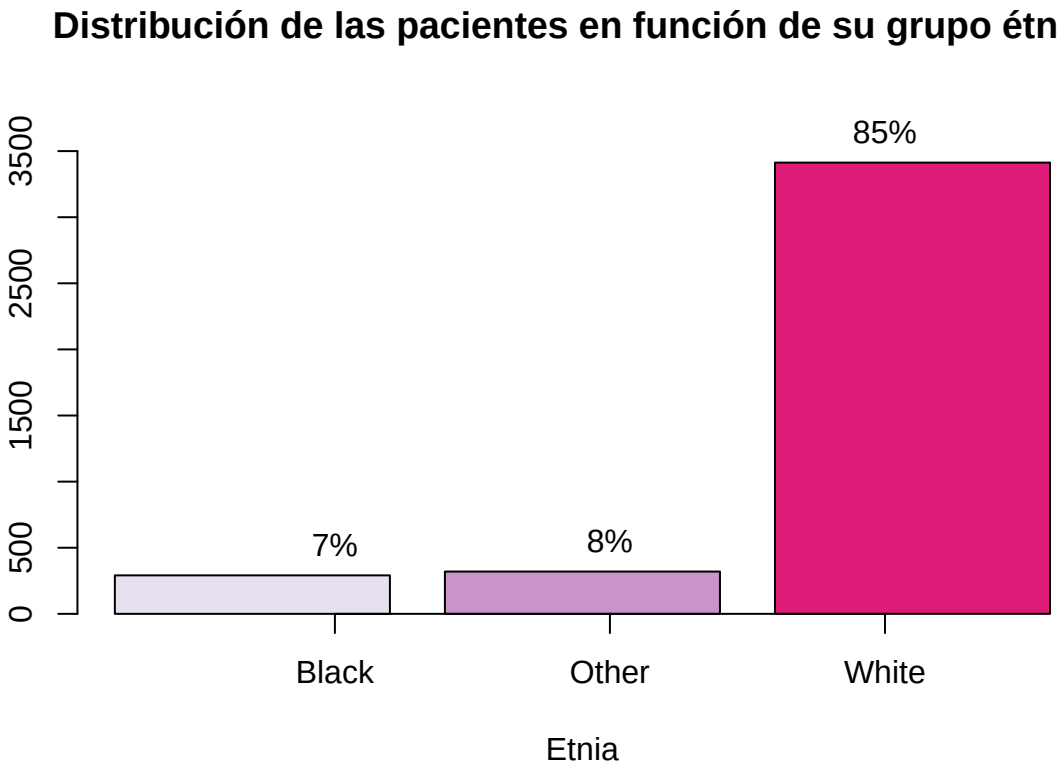
El tamaño medio del tumor fue de 30mmm, registrándose tamaños desde 1mm a 140mm.

Por último, la mínima supervivencia fue de un mes y la máxima 107 meses, estableciéndose la media en 71 meses de supervivencia desde el momento del diagnóstico.

```
if(!require('RColorBrewer')) install.packages('RColorBrewer'); library('RColorBrewer')
```

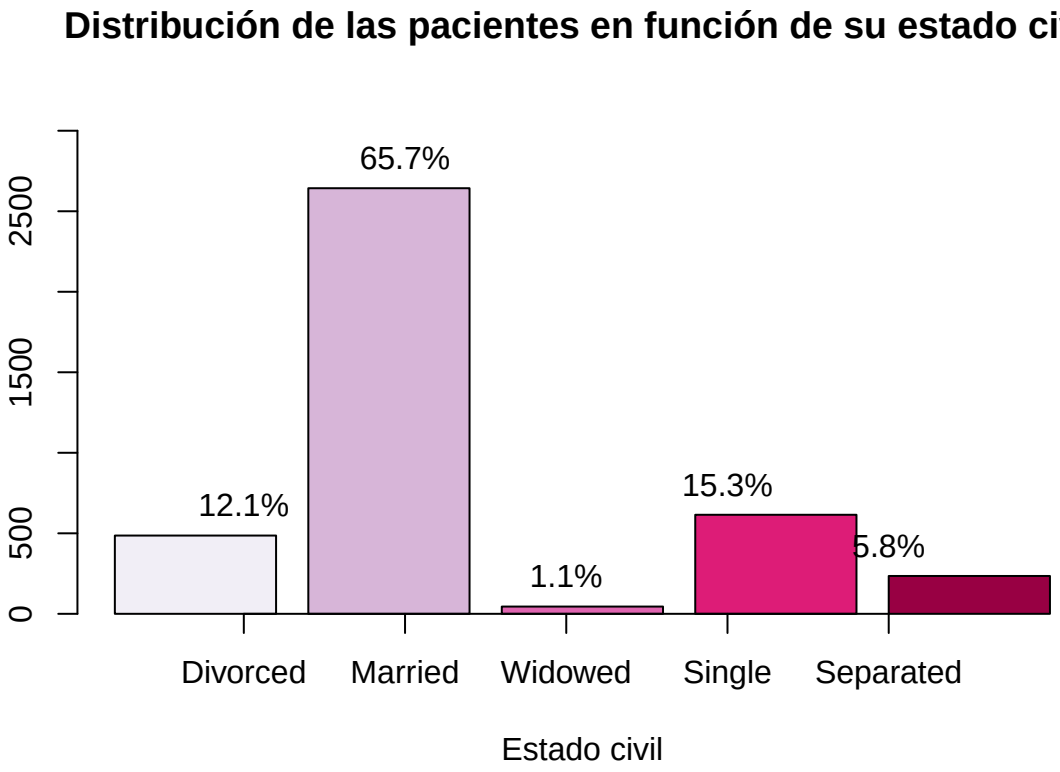
```
## Loading required package: RColorBrewer
```

```
counts <- as.numeric(table(BreastCancer$Race))
porcentajes <- prop.table(counts) * 100
palette <- brewer.pal(3, "PuRd")
barplot(counts, main="Distribución de las pacientes en función de su grupo étnico", xlab="Etnia", col=pa
text(x = 1:3, y = counts, labels = paste0(round(porcentajes), "%"), col = "black", cex = 1.5, pos=3)
names(counts) <- c("Black","Other", "White")
axis(side = 1, at = 1:3, labels = names(counts))
```



Como se aprecia nuestro dataset está desbalanceado en cuánto a la representación etnica de las pacientes, siendo la mayoria pacientes blancas (85%), negras (7%) y el otro 8% representaria a los grupos étnicos restantes.Este desequilibrio en la representación puede tener implicaciones en el análisis y la interpretación de los datos, ya que puede sesgar los resultados en favor de las categorías más representadas.

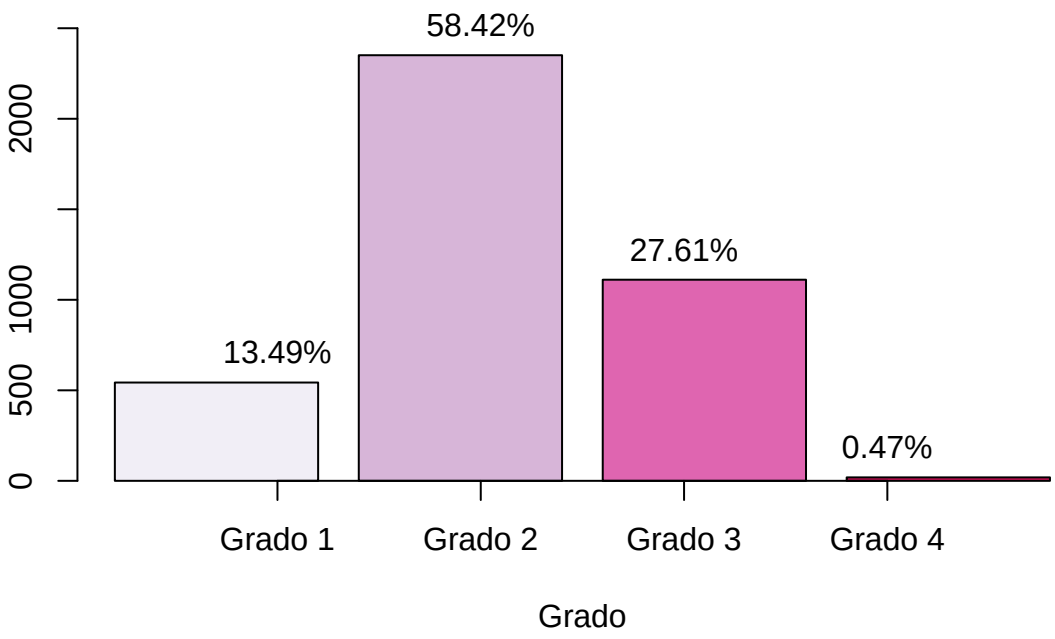
```
counts <- as.numeric(table(BreastCancer$Marital.Status))
porcentajes <- prop.table(counts) * 100
palette <- brewer.pal(5, "PuRd")
barplot(counts, main="Distribución de las pacientes en función de su estado civil", xlab="Estado civil"
text(x = 1:5, y = counts, labels = paste0(round(porcentajes,1), "%"), col = "black", cex = 1, pos = 3)
names(counts) <- c("Divorced","Married", "Widowed", "Single","Separated" )
axis(side = 1, at = 1:5, labels = names(counts))
```



Como se puede observar la mayoría de pacientes de cáncer de mama están casadas (85%), son solteras (15%) o divorciadas (12%).

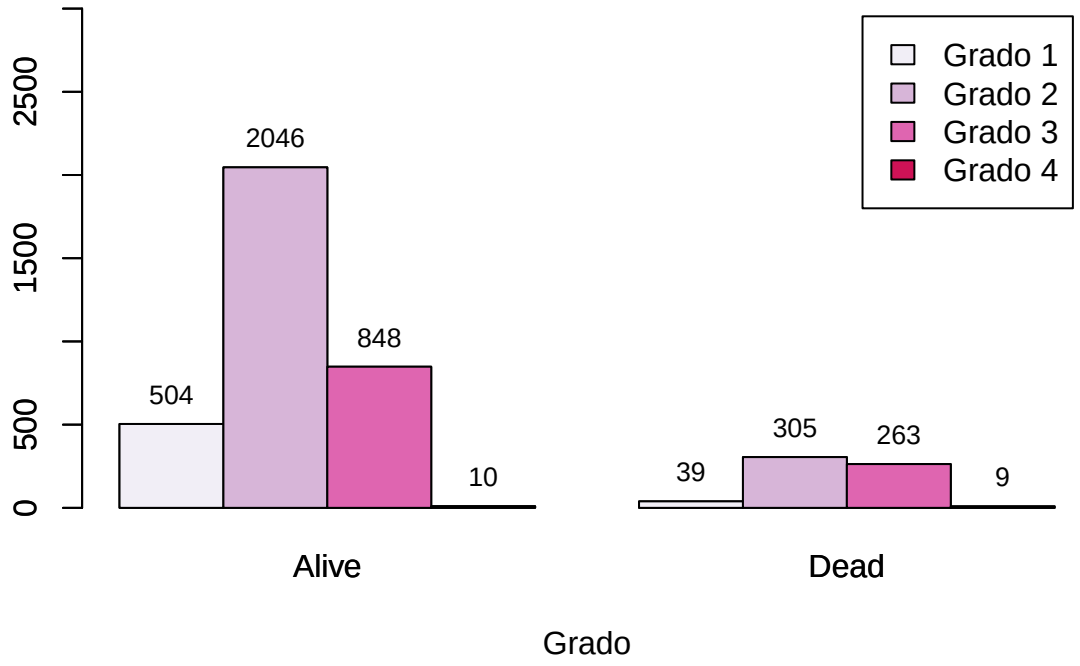
Ahora vamos a empezar a analizar las carácteristicas del cáncer y del propio tumor.

```
counts <- as.numeric(table(BreastCancer$Grade))
counts <- counts[c(2,3,4,1)]
porcentajes <- prop.table(counts) * 100
grado <- factor(c("Grado 1", "Grado 2", "Grado 3", "Grado 4"), levels = c("Grado 1", "Grado 2", "Grado 3
palette <- brewer.pal(4, "PuRd")
barplot(counts[grado], main="Distribución de las pacientes en función del grado del cáncer", xlab="Grado
text(x = 1:4, y = counts, labels = paste0(round(porcentajes,2), "%"), col = "black", cex = 1, pos = 3)
names(counts) <- c("Grado 1", "Grado 2", "Grado 3", "Grado 4")
axis(side = 1, at = 1:4, labels = names(counts)[grado])
```



Como se puede observar la mayoria de pacientes de cáncer de mama de este dataset presentan cáncer en estadio 2 (58%) o estadio 3 (28%). Destacar que los casos de cáncer de mama en estadio 4 no llegan ni al 0.5% de muestras.

```
BreastCancer$Grade <- ifelse(BreastCancer$Grade == "anaplastic; Grade IV", 4, BreastCancer$Grade)
counts <- table(BreastCancer$Grade, BreastCancer$Status)
counts <- counts[c(2,3,4,1),]
rownames(counts) <- c("Grado 1","Grado 2","Grado 3","Grado 4")
palette <- brewer.pal(4, "PuRd")
barplot(counts, main="Distribución de las pacientes en función del grado y del estado del cáncer", xlab=
text(x = barplot(counts, add=TRUE, col=palette, beside=TRUE), y = counts + 5, labels = counts, pos=3, c
```



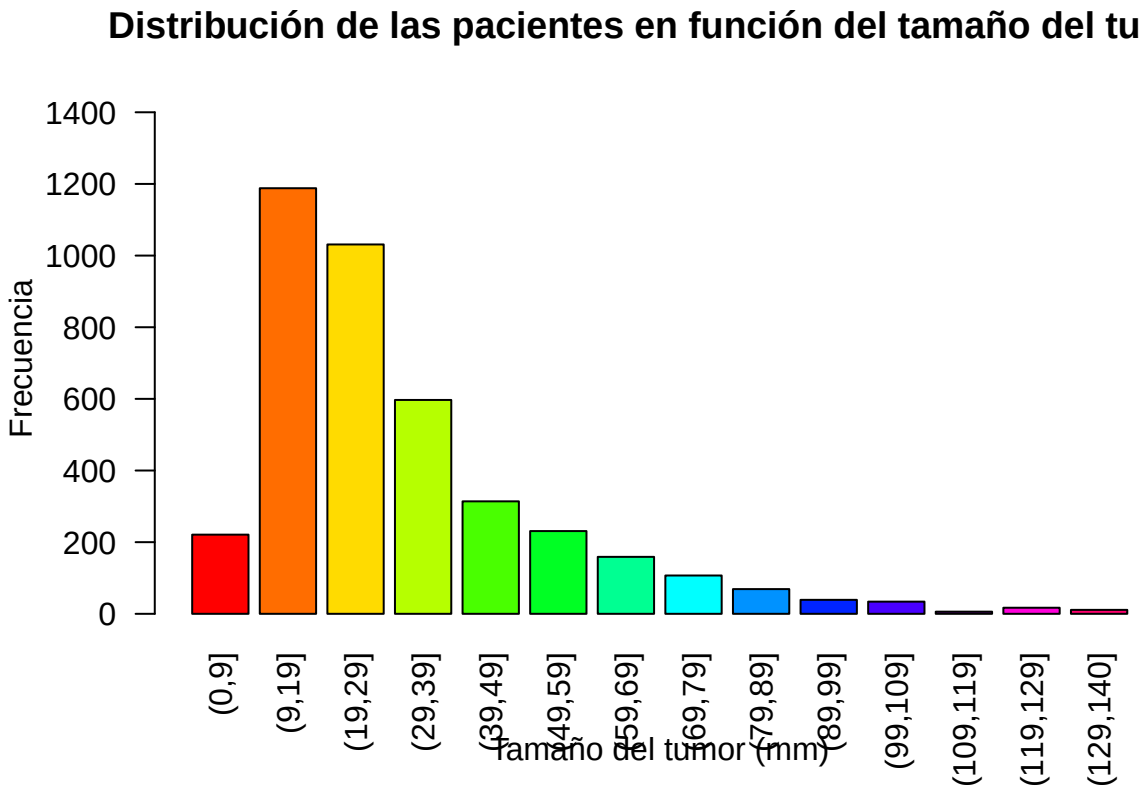

```
freq_pct <- prop.table(counts, 1) * 100
freq_pct_rounded <- round(freq_pct, 1)
table_pct <- cbind(counts, freq_pct_rounded)
colnames(table_pct) <- c("Alive", "Dead", "% Alive", "% Dead")
rownames(table_pct) <- rownames(counts)
table_pct
```

```
##          Alive Dead % Alive % Dead
## Grado 1    504   39    92.8    7.2
## Grado 2   2046  305    87.0   13.0
## Grado 3    848  263    76.3   23.7
## Grado 4     10    9    52.6   47.4
```

Como se puede apreciar las pacientes con cáncer de mama de grado uno en la mayoria de los casos (93,8%) sobreviven a la enfermedad. De forma similar, las pacientes en grado 2, también presentan una tasa de supervivencia alta (87%). Sin embargo, observamos que para cáncers más avanzados (grado 3 y 4) aproximadamente la superviviencia es del 75% y del 52% para el grado 3 y 4, respectivamente. Asimismo, vemos una cierta tendencia a que a mayor grado mayor mortalidad.

Ahora vamos a categorizar la variable Tumor.Size en 14 categorías distintas y vamos a mostrar un barplot con las frecuencias de cada categoría.

```
tumor_size_categories <- cut(BreastCancer$Tumor.Size, breaks = c(0, 9, 19, 29, 39, 49, 59, 69, 79, 89, 9
counts <- table(tumor_size_categories)
barplot(counts, main="Distribución de las pacientes en función del tamaño del tumor", xlab="Tamaño del t
```

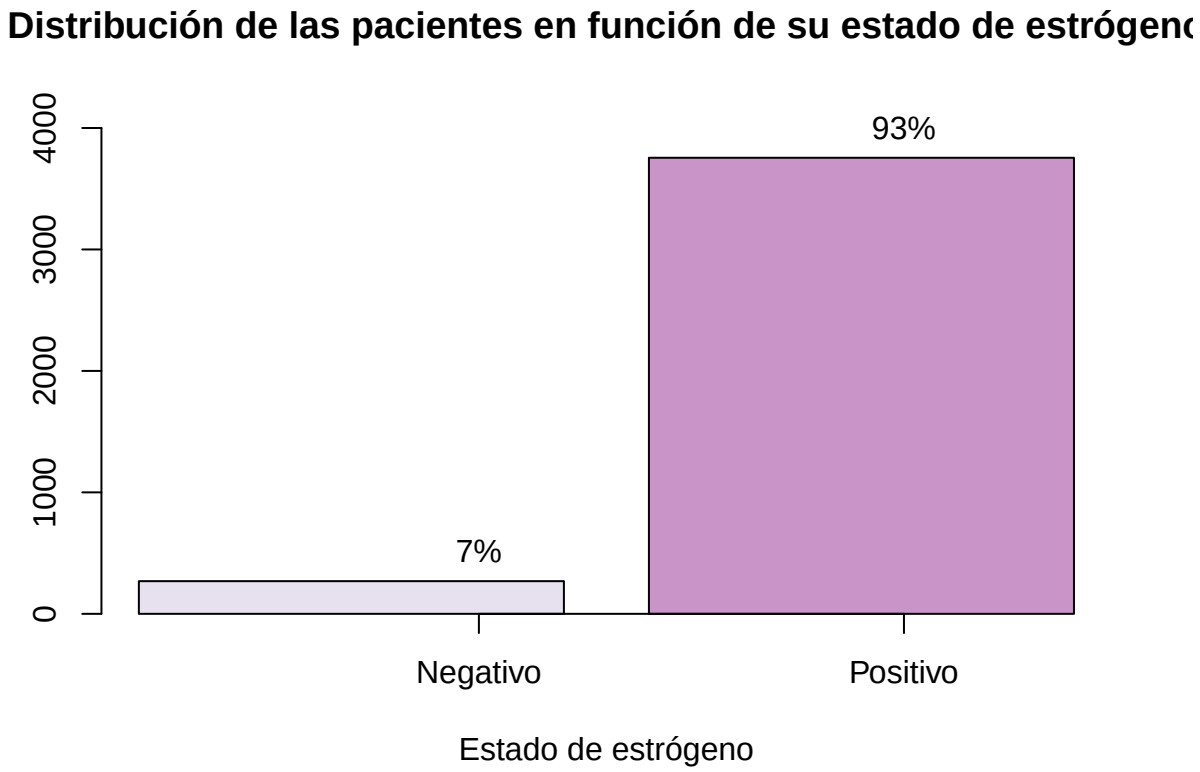


Como se puede observar en la gráfica de barras la mayoría de tumores tienen un tamaño de 9-19mm, 19-29mm y 29-29mm. Ahora vamos a revisar si existe alguna relación entre el tamaño del tumor y el grado del cáncer.

```
counts <- as.numeric(table(BreastCancer$Estrogen.Status))
porcentajes <- prop.table(counts) * 100
palette <- brewer.pal(2, "PuRd")
```

```
## Warning in brewer.pal(2, "PuRd"): minimal value for n is 3, returning requested palette with 3 diffe
```

```
barplot(counts, main="Distribución de las pacientes en función de su estado de estrógeno", xlab="Estado
text(x = 1:2, y = counts, labels = paste0(round(porcentajes), "%"), col = "black", cex = 1.5, pos=3)
names(counts) <- c("Negativo","Positivo")
axis(side = 1, at = 1:2, labels = names(counts))
```



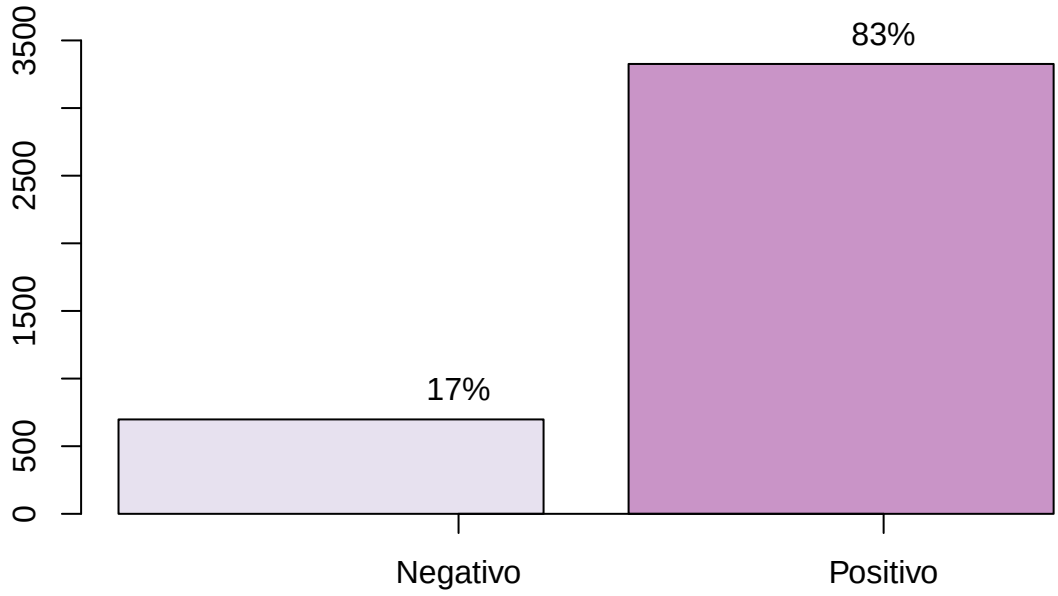
Como podemos observar en la gráfica el 93% de las pacientes que presentan cáncer de mama dan positivo para el receptor de estrógeno. Por consiguiente, la mayoría de pacientes presentan un cáncer que se puede tratar con inhibidores de estrógenos, haciendolo menos agresivo. Ahora vamos a observar el estado del receptor de progesterona.

```
counts <- as.numeric(table(BreastCancer$Progesterone.Status))
porcentajes <- prop.table(counts) * 100
palette <- brewer.pal(2, "PuRd")
```

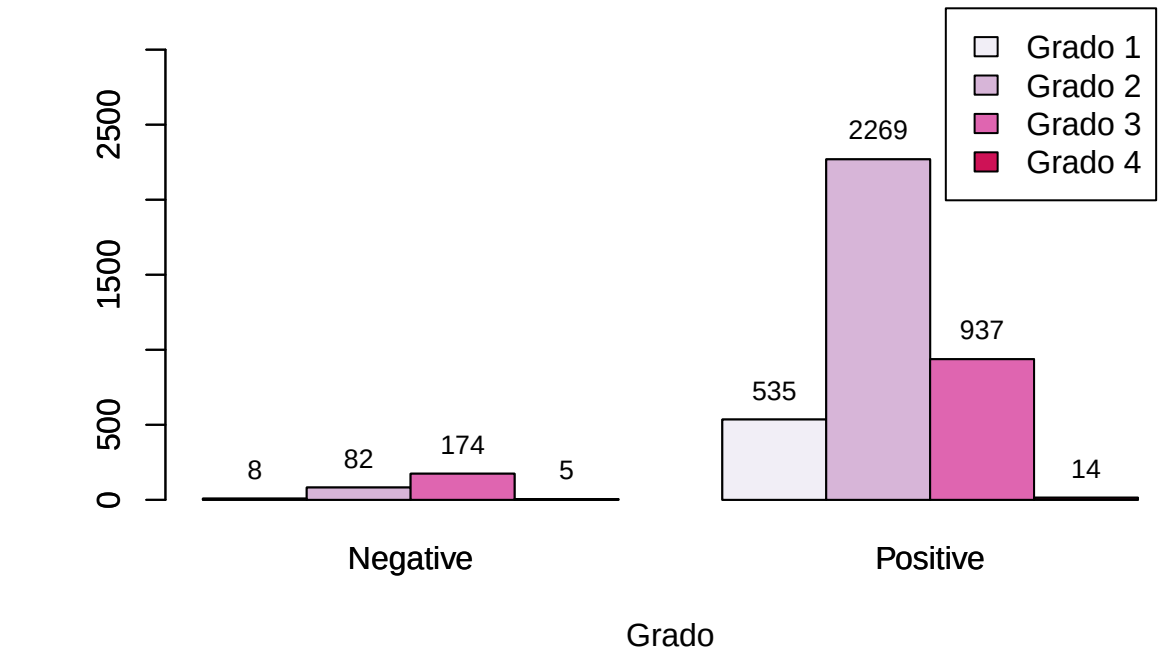
```
## Warning in brewer.pal(2, "PuRd"): minimal value for n is 3, returning requested palette with 3 diffe
```

```
barplot(counts, main="Distribución de las pacientes en función de su estado de progesterona", xlab="Esta
text(x = 1:2, y = counts, labels = paste0(round(porcentajes), "%"), col = "black", cex = 1.5, pos=3)
names(counts) <- c("Negativo","Positivo")
axis(side = 1, at = 1:2, labels = names(counts))
```

Distribución de las pacientes en función de su estado de progestero

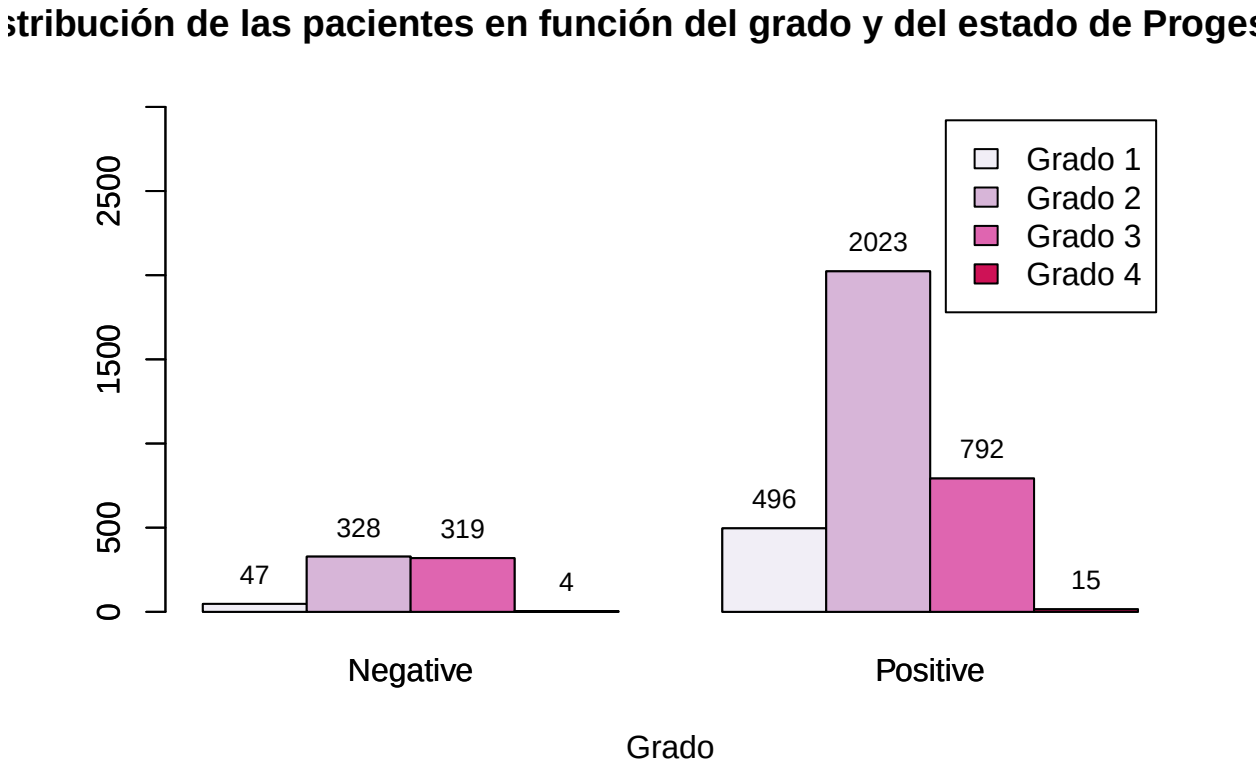



```
counts <- table(BreastCancer$Grade, BreastCancer$Estrogen.Status)
counts <- counts[c(2,3,4,1),]
rownames(counts) <- c("Grado 1","Grado 2","Grado 3","Grado 4")
palette <- brewer.pal(4, "PuRd")
barplot(counts, main="Distribución de las pacientes en función del grado y del estado de estrógeno", xla
text(x = barplot(counts, add=TRUE, col=palette, beside=TRUE), y = counts + 5, labels = counts, pos=3, c
```



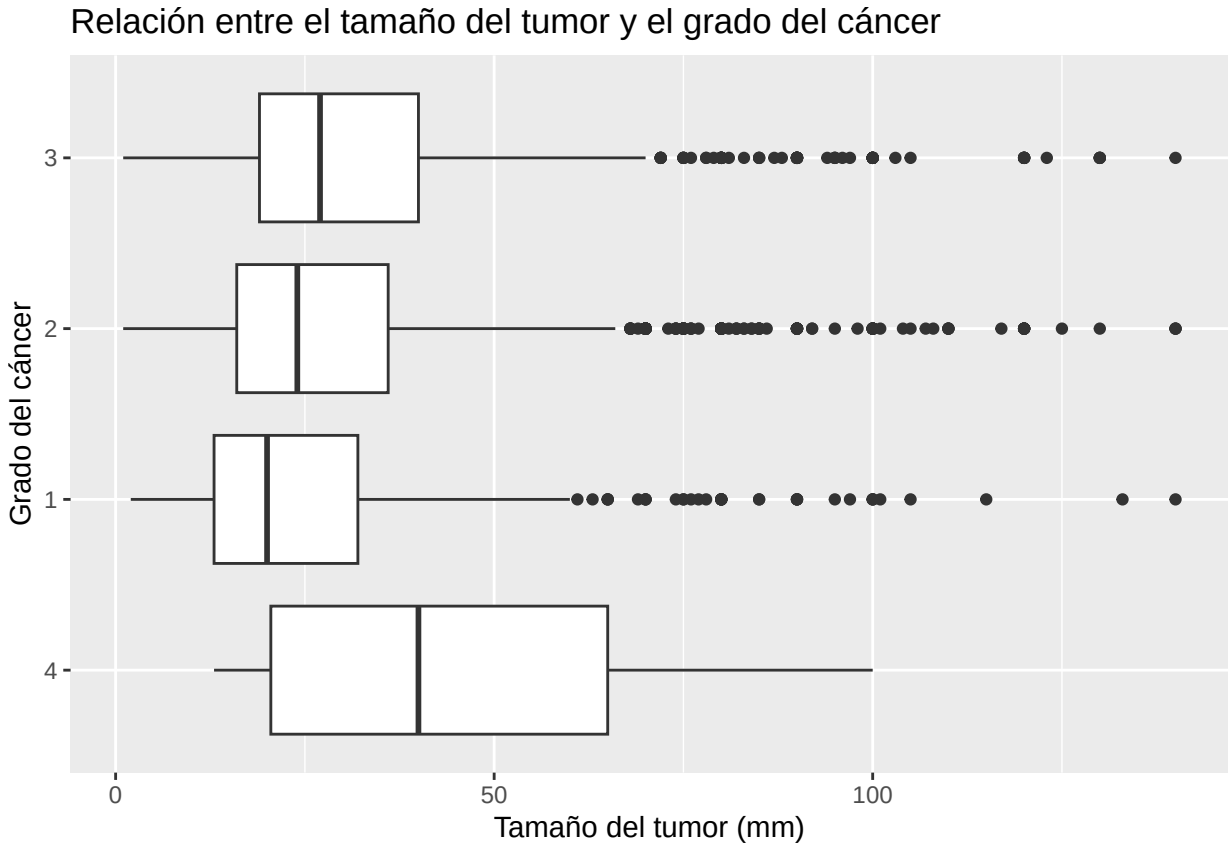
En esta gráfica podemos observar la relación entre el grado de cáncer de mama y el receptor de estrógen

```r
counts <- table(BreastCancer$Grade, BreastCancer$Progesterone.Status)
counts <- counts[c(2,3,4,1),]
rownames(counts) <- c("Grado 1","Grado 2","Grado 3","Grado 4")
palette <- brewer.pal(4, "PuRd")
barplot(counts, main="Distribución de las pacientes en función del grado y del estado de Progesterona",
text(x = barplot(counts, add=TRUE, col=palette, beside=TRUE), y = counts + 5, labels = counts, pos=3, c
```



En esta gráfica podemos observar la relación entre el grado de cáncer de mama y el receptor de progesterona (negativo =0 y positivo=1). Vemos que a medida que aumente el grado del cáncer aumentan los casos de receptor de progesterona negativo, de la misma forma que lo hacían con el receptor de estrógeno.

```
BreastCancer$Grade <- ifelse(BreastCancer$Grade == "anaplastic; Grade IV", "4", BreastCancer$Grade)
library(ggplot2)
ggplot(data = BreastCancer, aes(x = Tumor.Size, y = Grade)) +
  geom_boxplot() +
  scale_x_continuous(name = "Tamaño del tumor (mm)") +
  scale_y_discrete(name = "Grado del cáncer", labels = c("4", "1", "2", "3")) +
  ggtitle("Relación entre el tamaño del tumor y el grado del cáncer")
```



Como podemos apreciar en este diagrama de caja y bigotes si parece existir una relación entre el tamaño del tumor y el grado del cáncer. Se observa que a medida que va aumentando el tamaño del tumor también lo hace el grado, siendo los tumores de cáncer de grado 1 los más pequeños y del grado 4 los más grandes.

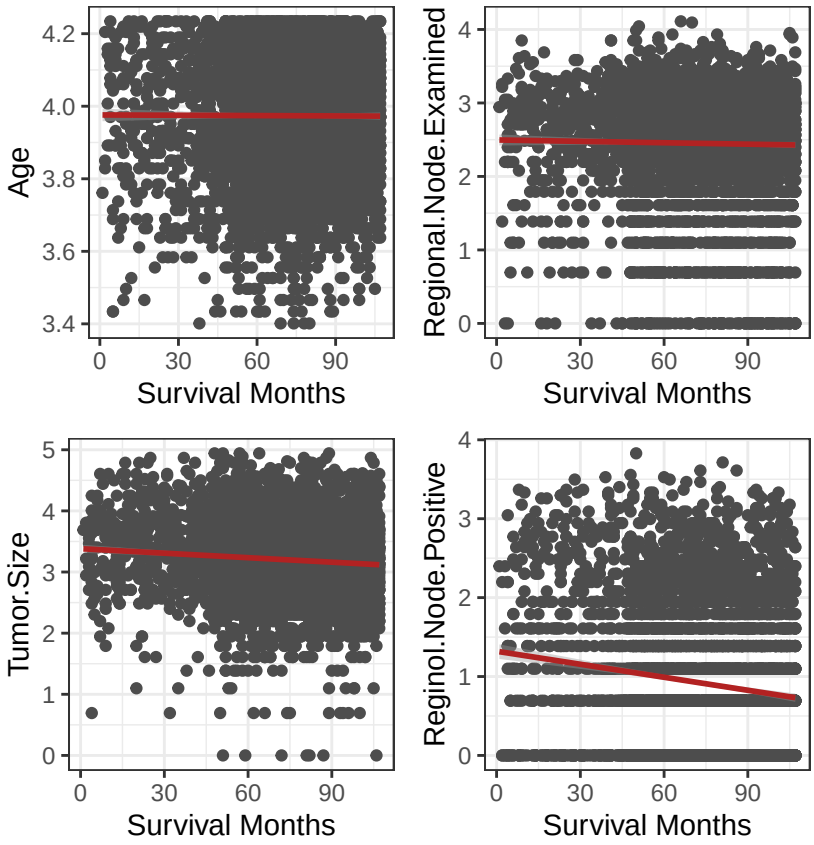
```
n= c(1,10,13,14)
BreastCancerAux=BreastCancer %>% select(all_of(n))
histList2<- vector('list',ncol(BreastCancerAux))
for(i in seq_along(BreastCancerAux)){
  message(i)
histList2[[i]]<-local({
  i<-i
  col <-log(BreastCancerAux[[i]])
  ggp<- ggplot(data = BreastCancerAux, aes(x = BreastCancer$Survival.Months, y=col)) +
    geom_point(color = "gray30")+geom_smooth(method = lm,color = "firebrick")+
    theme_bw() +xlab("Survival Months")+ylab(names(BreastCancerAux)[i])
  })
}
```

```
## 1
## 2
## 3
## 4
```

```
multiplot(plotlist = histList2, cols = 3)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Observamos que los meses de supervivencia aumentan cuando:

- El tamaño del tumor es menor.
- El número de nódulos positivos es menor.

Vamos a utilizar las columnas de interés para realizar la matriz y la vamos a visualizar con la funcion corrplot.

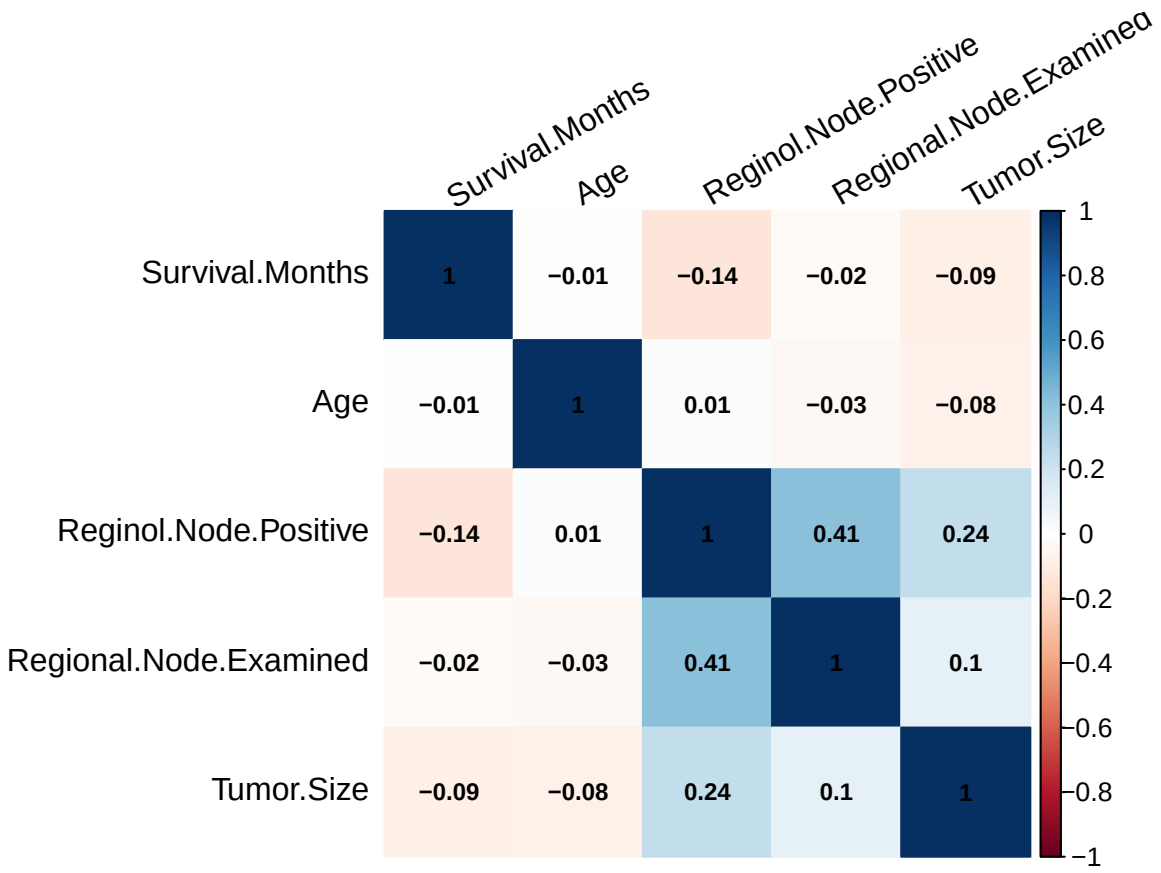
```
if (!require("corrplot")) install.packages("corrplot")
```

```
## Loading required package: corrplot
```

```
## corrplot 0.92 loaded
```

```
library("corrplot")

n <- c(1, 10, 13, 14,15)
factors <- BreastCancer %>% select(all_of(n))
complete_cases <- complete.cases(factors)
factors <- factors[complete_cases, ]
res <- cor(factors)
corrplot(res, method = "color", tl.col = "black", tl.srt = 30, order = "AOE",
         number.cex = 0.75, sig.level = 0.01, addCoef.col = "black")
```



Observamos que no existeapenas correlacion entre las variables excepto para Regional Node Examined y Regional Node Positive, hecho que tiene sentido ya que ambas estan relacionadas.Sin embargo, la correlación no me parece lo suficiente elevada como para eliminar ninguna de las variables en el dataset.

```
age_categories <- cut(BreastCancer$Age, breaks = c(0, 30, 40, 50, 60, Inf), labels = c("0-30", "31-40",
table_grade_age <- table(BreastCancer$Grade, age_categories)
print(table_grade_age)
```

```
##                         age_categories
##                          0-30 31-40 41-50 51-60 61+
##    anaplastic; Grade IV     0     4     5     5   5
##   1                         0    19   162   188 174
##   2                         3   157   663   842 686
##   3                         2   117   366   351 275
```

```
chi_sq <- chisq.test(table_grade_age)
```

```
## Warning in chisq.test(table_grade_age): Chi-squared approximation may be
## incorrect
```

```
print(chi_sq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_grade_age
## X-squared = 51.735, df = 12, p-value = 6.907e-07
```

La prueba de chi-cuadrado muestra que hay evidencia significativa de correlación entre las variables “Grade” y “Age” (p-value = 6.907e-07).

Ejercicio 5

Discretización

Ahora procederemos a discretizar la variable Regional Node Positive ya que era una de las que las distancias entre sus valores era muy grande:

```
if (!require('arules')) install.packages('arules'); library('arules')
```

```
## Loading required package: arules
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

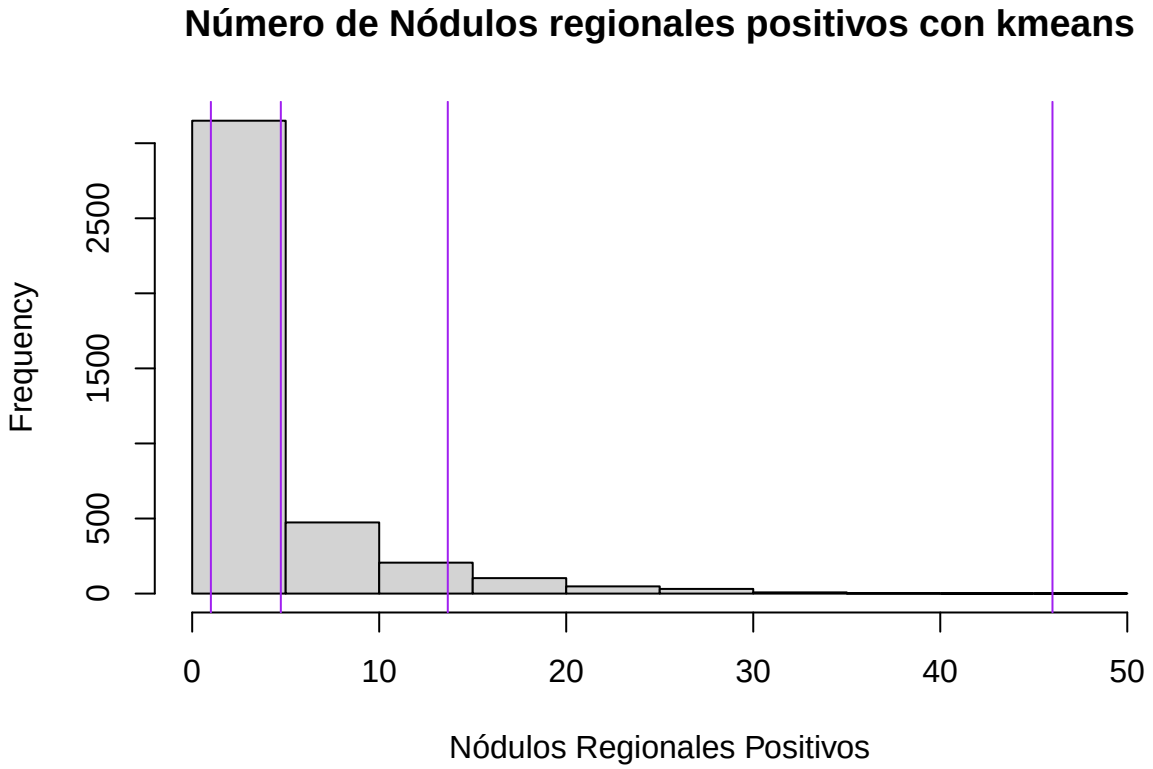
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
set.seed(2)
table(discretize(BreastCancer$Reginol.Node.Positive, "cluster" ))
```

```
##
##    [1,6.23) [6.23,16.4)   [16.4,46]
##        3290         569         165
```

```
hist(BreastCancer$Reginol.Node.Positive, main="Número de Nódulos regionales positivos con kmeans",xlab="
abline(v=discretize(BreastCancer$Reginol.Node.Positive, method="cluster", onlycuts=TRUE),col="purple")
```



Podemos observar que sin pasar ningún parámetro (dejando que el algoritmo escoja las particiones) se obtienen tres

clústers que dividen los nódulos regionales positivos. Asignamos el propio clúster como una variable más al dataset para poder trabajar con ella más tarde.

```
BreastCancer$Reginol.Node.PositiveNM<- (discretize(BreastCancer$Reginol.Node.Positive, "cluster" ))
head(BreastCancer)
```

```
##   Age  Race Marital.Status T.Stage N.Stage X6th.Stage             differentiate
## 1  68 White        Married      T1      N1        IIA     Poorly differentiated
## 2  50 White        Married      T2      N2       IIIA Moderately differentiated
## 3  58 White       Divorced      T3      N3       IIIC Moderately differentiated
## 4  58 White        Married      T1      N1        IIA     Poorly differentiated
## 5  47 White        Married      T2      N1        IIB     Poorly differentiated
## 6  51 White        Single       T1      N1        IIA Moderately differentiated
##   Grade  A.Stage Tumor.Size Estrogen.Status Progesterone.Status
## 1     3 Regional          4        Positive            Positive
## 2     2 Regional         35        Positive            Positive
## 3     2 Regional         63        Positive            Positive
## 4     3 Regional         18        Positive            Positive
## 5     3 Regional         41        Positive            Positive
## 6     2 Regional         20        Positive            Positive
##   Regional.Node.Examined Reginol.Node.Positive Survival.Months Status
## 1                     24                     1              60  Alive
## 2                     14                     5              62  Alive
## 3                     14                     7              75  Alive
## 4                      2                     1              84  Alive
## 5                      3                     1              50  Alive
## 6                     18                     2              89  Alive
##   Reginol.Node.PositiveNM
## 1               [1,4.89)
## 2            [4.89,14.3)
## 3            [4.89,14.3)
## 4               [1,4.89)
## 5               [1,4.89)
## 6               [1,4.89)
```

Normalización

Ahora vamos a normalizar la variable Nódulos Regionales Positivos por el máximo añadiendo un nou valor a los datos que contendrá el valor.

```
BreastCancer$Reginol.Node.PositiveNM<- (BreastCancer$Reginol.Node.Positive/max(BreastCancer[,"Reginol.N
head(BreastCancer$Reginol.Node.PositiveNM)
```

```
## [1] 0.02173913 0.10869565 0.15217391 0.02173913 0.02173913 0.04347826
```

Supongamos que debemos normalizar por la diferencia para ubicar entre 0 y 1 la variable Nódulo Regional Positivo dado que el algoritmo de minería que utilizaremos así lo requiere.

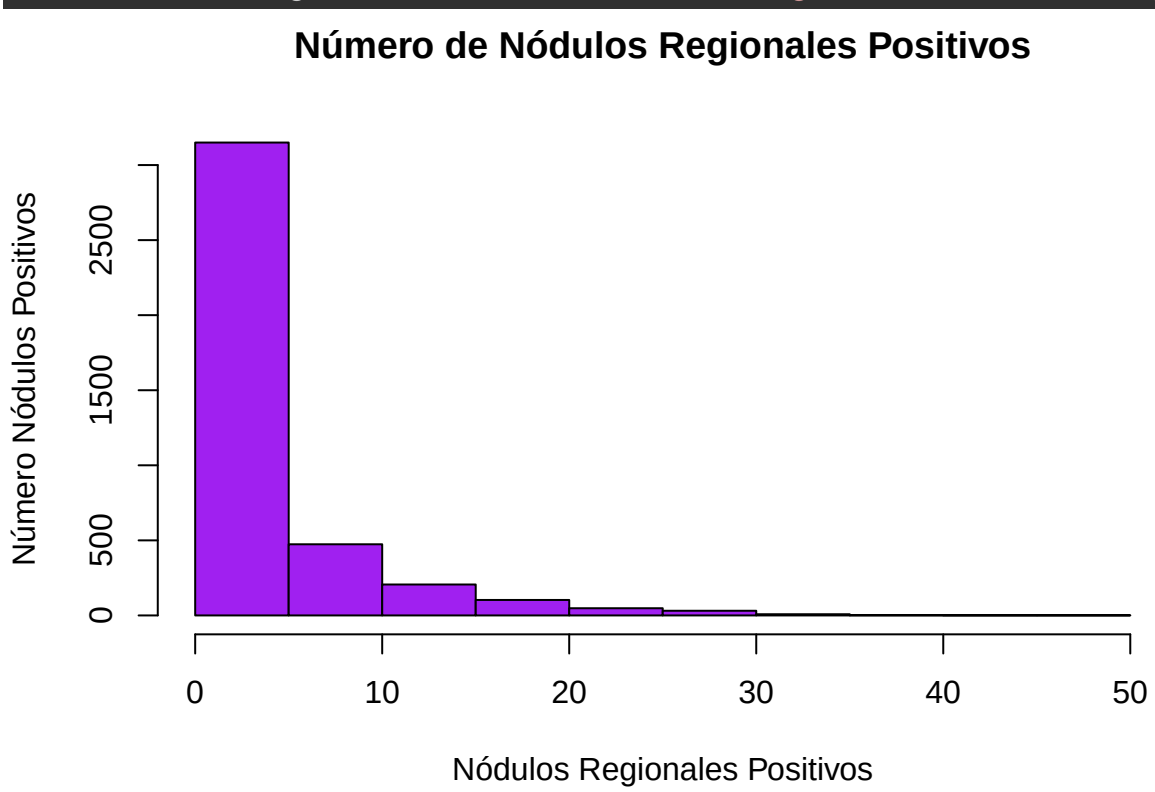
```
BreastCancer$Reginol.Node.PositiveND <- (BreastCancer$Reginol.Node.Positive - min(BreastCancer$Reginol.
max(BreastCancer$Reginol.Node.Positive)
```

```
## [1] 46
```

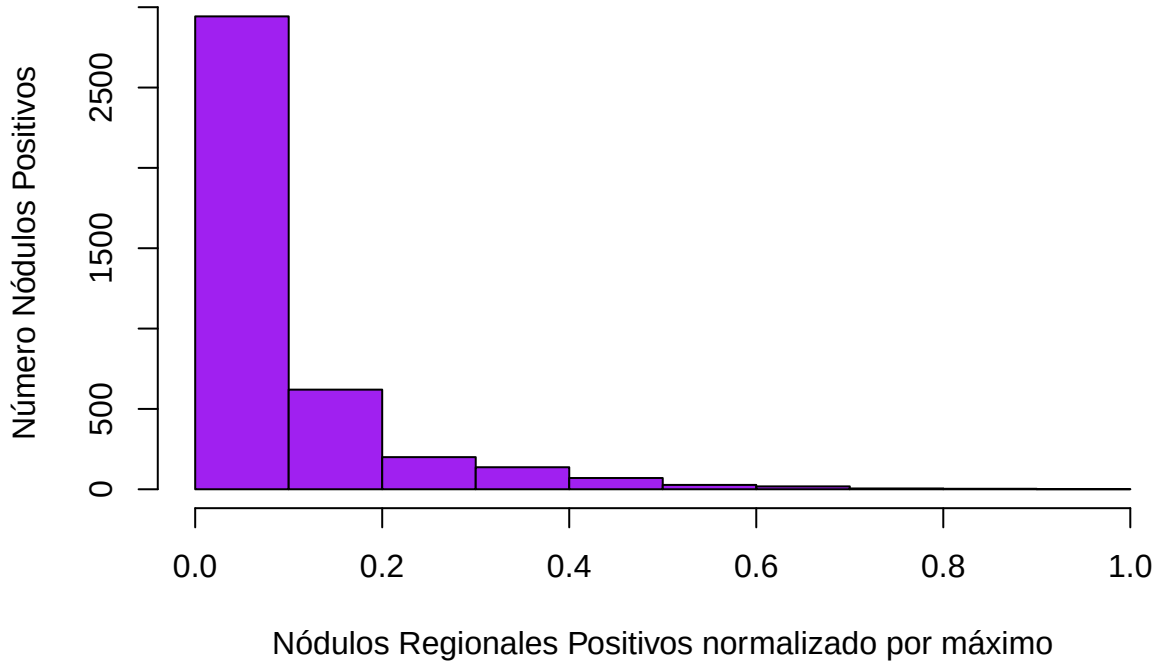
```
min(BreastCancer$Reginol.Node.Positive)
```

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

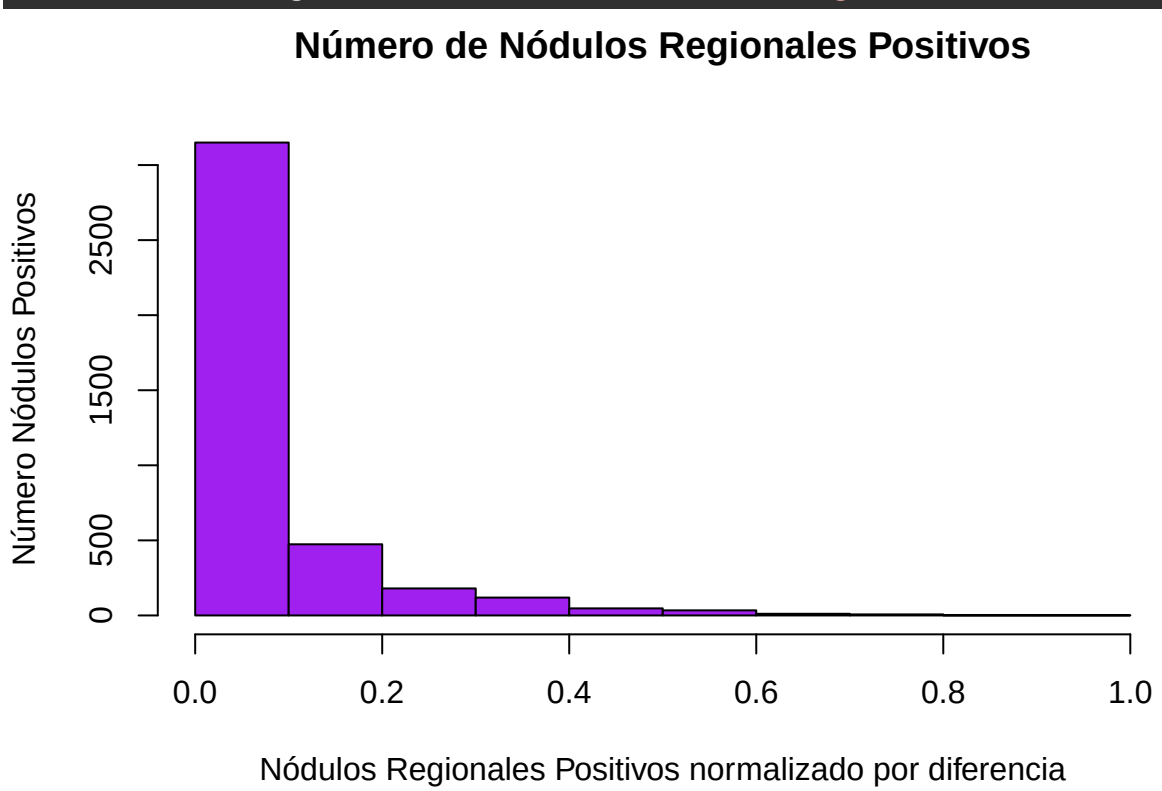
```
hist(BreastCancer$Reginol.Node.Positive,xlab="Nódulos Regionales Positivos", col="purple",ylab="Número
```




```
hist(BreastCancer$Reginol.Node.PositiveNM ,xlab="Nódulos Regionales Positivos normalizado por máximo",
```



```r
hist(BreastCancer$Reginol.Node.PositiveND ,xlab="Nódulos Regionales Positivos normalizado por diferencia
```



A continuación vamos a normalizar las otras columnas, pero primeros vamos a pasar a valores númericos las categorias que sean categóricas.

```r
BreastCancer$Race <- as.numeric(factor(BreastCancer$Race))
BreastCancer$Marital.Status <- as.numeric(factor(BreastCancer$Marital.Status))
BreastCancer$T.Stage <- as.numeric(factor(BreastCancer$T.Stage))
BreastCancer$N.Stage <- as.numeric(factor(BreastCancer$N.Stage))
BreastCancer$X6th.Stage <- as.numeric(factor(BreastCancer$X6th.Stage))
BreastCancer$differentiate <- as.numeric(factor(BreastCancer$differentiate))
BreastCancer$Grade <- as.numeric(factor(BreastCancer$Grade))
BreastCancer$A.Stage <- as.numeric(factor(BreastCancer$A.Stage))
BreastCancer$Estrogen.Status <- as.numeric(factor(BreastCancer$Estrogen.Status))
BreastCancer$Progesterone.Status <- as.numeric(factor(BreastCancer$Progesterone.Status))
BreastCancer$Status <- as.numeric(factor(BreastCancer$Status))

head(BreastCancer)
```

```
##   Age Race Marital.Status T.Stage N.Stage X6th.Stage differentiate Grade
## 1  68    3              2       1       1          1              2     4
## 2  50    3              2       2       2          3              1     3
## 3  58    3              1       3       3          5              1     3
## 4  58    3              2       1       1          1              2     4
## 5  47    3              2       2       1          2              2     4
## 6  51    3              4       1       1          1              1     3
##   A.Stage Tumor.Size Estrogen.Status Progesterone.Status Regional.Node.Examined
## 1       2          4               2                   2                    24
## 2       2         35               2                   2                    14
## 3       2         63               2                   2                    14
```

```
## 4       2         18               2                   2                      2
## 5       2         41               2                   2                      3
## 6       2         20               2                   2                     18
##   Reginol.Node.Positive Survival.Months Status Reginol.Node.PositiveNM
## 1                     1              60      1              0.02173913
## 2                     5              62      1              0.10869565
## 3                     7              75      1              0.15217391
## 4                     1              84      1              0.02173913
## 5                     1              50      1              0.02173913
## 6                     2              89      1              0.04347826
##   Reginol.Node.PositiveND
## 1              0.00000000
## 2              0.08888889
## 3              0.13333333
## 4              0.00000000
## 5              0.00000000
## 6              0.02222222
```

```r
# Obtener el índice de las columnas numéricas
numeric_cols <- sapply(BreastCancer, is.numeric)

# Normalizar las variables numéricas en el rango [0,1]
normalize <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}
BreastCancer[numeric_cols] <- lapply(BreastCancer[numeric_cols], normalize)
```

```r
head(BreastCancer)
```

```
##         Age Race Marital.Status   T.Stage N.Stage X6th.Stage differentiate
## 1 0.9743590    1           0.25 0.0000000     0.0       0.00     0.3333333
## 2 0.5128205    1           0.25 0.3333333     0.5       0.50     0.0000000
## 3 0.7179487    1           0.00 0.6666667     1.0       1.00     0.0000000
## 4 0.7179487    1           0.25 0.0000000     0.0       0.00     0.3333333
## 5 0.4358974    1           0.25 0.3333333     0.0       0.25     0.3333333
## 6 0.5384615    1           0.75 0.0000000     0.0       0.00     0.0000000
##       Grade A.Stage Tumor.Size Estrogen.Status Progesterone.Status
## 1 1.0000000       1 0.02158273               1                   1
## 2 0.6666667       1 0.24460432               1                   1
## 3 0.6666667       1 0.44604317               1                   1
## 4 1.0000000       1 0.12230216               1                   1
## 5 1.0000000       1 0.28776978               1                   1
## 6 0.6666667       1 0.13669065               1                   1
##   Regional.Node.Examined Reginol.Node.Positive Survival.Months Status
## 1             0.38333333            0.00000000       0.5566038      0
## 2             0.21666667            0.08888889       0.5754717      0
## 3             0.21666667            0.13333333       0.6981132      0
## 4             0.01666667            0.00000000       0.7830189      0
## 5             0.03333333            0.00000000       0.4622642      0
## 6             0.28333333            0.02222222       0.8301887      0
##   Reginol.Node.PositiveNM Reginol.Node.PositiveND
## 1              0.00000000              0.00000000
## 2              0.08888889              0.08888889
## 3              0.13333333              0.13333333
## 4              0.00000000              0.00000000
```

```
## 5              0.00000000              0.00000000
## 6              0.02222222              0.02222222
```

Ejercicio 6

El análisis de componentes principales, principal component analysis (PCA) en inglés, es un método que nos permite trabajar con componentes independientres entre si, permitiéndonos representar nuestro dataset en un nuevo sistema de coordenadas (componentes principales) que está mejor adaptado a nuestro juego de datos, entendiendo mejor su variabilidad.

Ahora vamos a aplicar el PCA a nuestro dataset ejecutando la función prcomp().

```
pca.BreastCancer <- prcomp(BreastCancer)
summary(pca.BreastCancer)
```

```
## Importance of components:
##                            PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     0.5414 0.4093 0.3510 0.3396 0.29051 0.26778 0.25972
## Proportion of Variance 0.2585 0.1477 0.1087 0.1017 0.07442 0.06323 0.05948
## Cumulative Proportion  0.2585 0.4062 0.5149 0.6166 0.69103 0.75426 0.81374
##                            PC8     PC9   PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.22648 0.19653 0.1829 0.17713 0.14265 0.12872 0.09187
## Proportion of Variance 0.04523 0.03406 0.0295 0.02767 0.01794 0.01461 0.00744
## Cumulative Proportion  0.85897 0.89303 0.9225 0.95020 0.96815 0.98276 0.99020
##                           PC15    PC16     PC17      PC18
## Standard deviation     0.08547 0.06171 7.42e-17 1.429e-17
## Proportion of Variance 0.00644 0.00336 0.00e+00 0.000e+00
## Cumulative Proportion  0.99664 1.00000 1.00e+00 1.000e+00
```

Summary nos devuelve la proporción de varianza aplicada al conjunto de cada atributo. Observamos como el atributo 1 explica el 0.5414 de variabilidad del total de datos mientras el atributo 16 solo explica el 0,06171.

```
library('factoextra')
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

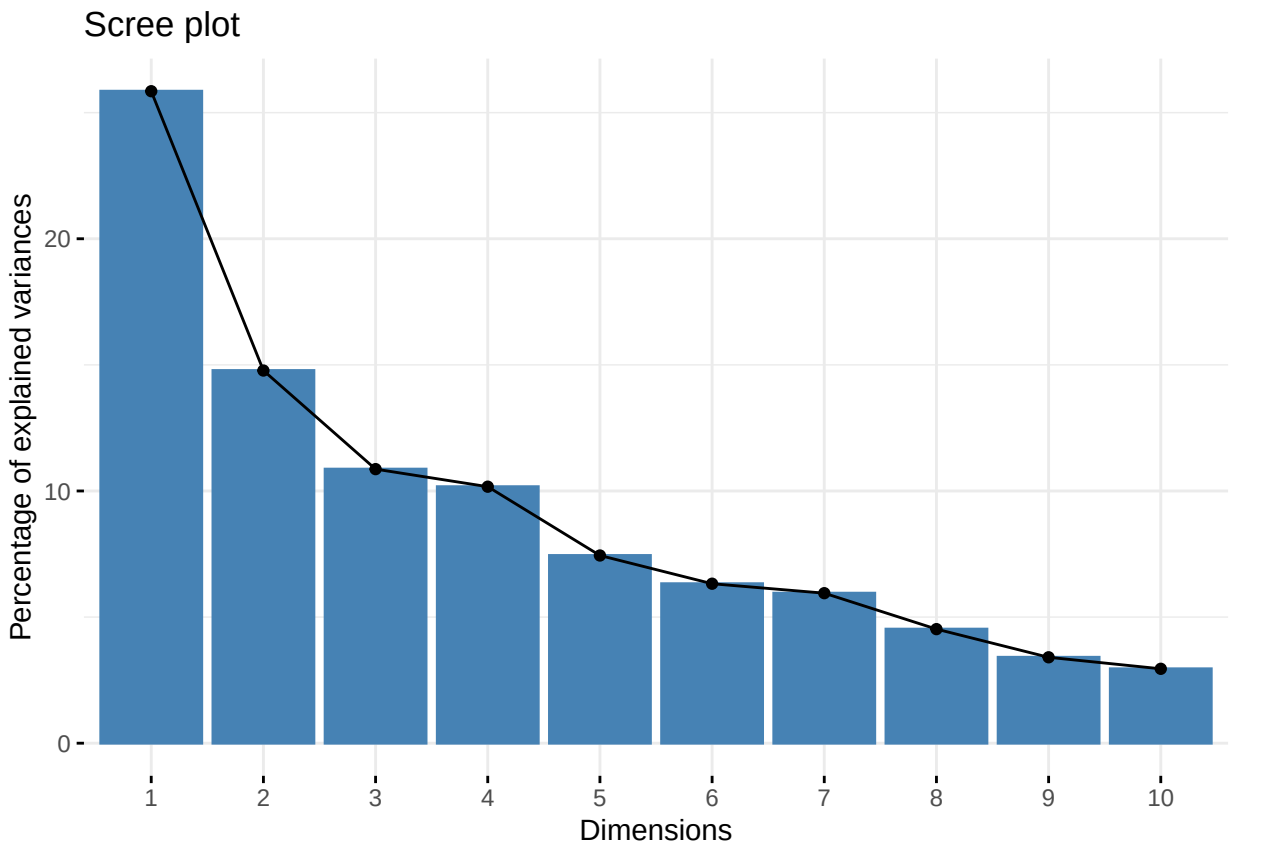
```
ev= get_eig(pca.BreastCancer)
ev
```

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  2.931274e-01     2.584874e+01                     25.84874
## Dim.2  1.675545e-01     1.477540e+01                     40.62413
## Dim.3  1.232263e-01     1.086642e+01                     51.49055
## Dim.4  1.153316e-01     1.017024e+01                     61.66080
## Dim.5  8.439501e-02     7.442171e+00                     69.10297
## Dim.6  7.170449e-02     6.323088e+00                     75.42606
## Dim.7  6.745455e-02     5.948318e+00                     81.37437
## Dim.8  5.129218e-02     4.523078e+00                     85.89745
## Dim.9  3.862443e-02     3.406002e+00                     89.30345
## Dim.10 3.345209e-02     2.949892e+00                     92.25335
## Dim.11 3.137648e-02     2.766860e+00                     95.02020
## Dim.12 2.034888e-02     1.794417e+00                     96.81462
## Dim.13 1.656963e-02     1.461153e+00                     98.27577
## Dim.14 8.439285e-03     7.441981e-01                     99.01997
## Dim.15 7.305593e-03     6.442262e-01                     99.66420
## Dim.16 3.808017e-03     3.358008e-01                    100.00000
## Dim.17 5.505301e-33     4.854718e-31                    100.00000
```

```
## Dim.18 2.042545e-34     1.801170e-32                    100.00000
```

A continuacion se muestra en un histograma el peso de cada atributo sobre el conjunto total de datos:

```
fviz_eig(pca.BreastCancer)
```



Vamos a utilizar el método de Kaiser para decidir cuales de las variables obtenidas será escogida. Este criterio mantendrá todas aquellas variables cuyas varianzas sean superiores a 1.

```
varianza_variables <- pca.BreastCancer$sdev^2
pca.BreastCancer
```

```
## Standard deviations (1, .., p=18):
##  [1] 5.414125e-01 4.093343e-01 3.510361e-01 3.396051e-01 2.905082e-01
##  [6] 2.677769e-01 2.597201e-01 2.264778e-01 1.965310e-01 1.828991e-01
## [11] 1.771341e-01 1.426495e-01 1.287231e-01 9.186558e-02 8.547276e-02
## [16] 6.170913e-02 7.419772e-17 1.429177e-17
##
## Rotation (n x k) = (18 x 18):
##                                  PC1           PC2          PC3         PC4
## Age                      -0.003935856 -0.0150511810  0.046824706 -0.05949927
## Race                     -0.033491842  0.0562563926 -0.048570433  0.04835619
## Marital.Status            0.013226718 -0.0260043841  0.074752898 -0.04563416
## T.Stage                   0.249016385  0.1116306978 -0.014859593  0.07938556
## N.Stage                   0.558416500  0.2555389437  0.024468030  0.15620224
## X6th.Stage                0.532840953  0.2436788077  0.006641180  0.16375635
## differentiate            -0.055396152 -0.0002705696  0.928748063  0.18439864
## Grade                     0.106417593 -0.0875310600 -0.295284910 -0.03109591
## A.Stage                  -0.082690631 -0.0347560941 -0.013817338 -0.01995777
```

```
## Tumor.Size                0.134359454  0.0547857489 -0.008599149  0.04760730
## Estrogen.Status          -0.136036334  0.3672638059 -0.010957132 -0.13576704
## Progesterone.Status      -0.233526839  0.7808564007  0.039780039 -0.38991597
## Regional.Node.Examined    0.079380887  0.0486219191 -0.021530886  0.04413643
## Reginol.Node.Positive     0.165169147  0.0752087792  0.010660584  0.03655195
## Survival.Months          -0.125147087  0.1031087519 -0.082020401  0.31102288
## Status                    0.353270805 -0.2689540735  0.172427897 -0.78976488
## Reginol.Node.PositiveNM   0.165169147  0.0752087792  0.010660584  0.03655195
## Reginol.Node.PositiveND   0.165169147  0.0752087792  0.010660584  0.03655195
##                                  PC5          PC6          PC7           PC8
## Age                      -0.125389715  0.26357927  0.27570540 -0.824357656
## Race                     -0.865102563  0.05012062  0.40143888  0.278829351
## Marital.Status            0.462998977  0.11009149  0.84164119  0.233439243
## T.Stage                  -0.021797992 -0.73862442  0.16848449 -0.176374358
## N.Stage                   0.018708986  0.39749982 -0.06996996  0.070765150
## X6th.Stage               -0.001972055 -0.07670365  0.02265023 -0.049102526
## differentiate            -0.036602829 -0.05148927 -0.05992573  0.071909163
## Grade                     0.072553979 -0.04140779 -0.04497329  0.179269379
## A.Stage                   0.003915498  0.03898096  0.00186459 -0.009428360
## Tumor.Size               -0.008528580 -0.38183128  0.08313977 -0.082103157
## Estrogen.Status          -0.026424458  0.02710127  0.05606804 -0.275376211
## Progesterone.Status       0.058397943 -0.03640167 -0.04579240  0.147729393
## Regional.Node.Examined    0.005982849  0.06250422 -0.01780877  0.040046099
## Reginol.Node.Positive    -0.002669615  0.12570860 -0.02141613  0.021568498
## Survival.Months           0.023444435  0.02887893  0.02290744 -0.069212179
## Status                   -0.096283701 -0.02344097 -0.02626413  0.009175023
## Reginol.Node.PositiveNM  -0.002669615  0.12570860 -0.02141613  0.021568498
## Reginol.Node.PositiveND  -0.002669615  0.12570860 -0.02141613  0.021568498
##                                  PC9          PC10          PC11           PC12
## Age                       3.733483e-01 -0.044084888  0.105029028 -0.0007868454
## Race                      9.606452e-05 -0.022406056  0.023484337 -0.0004443695
## Marital.Status           -4.327245e-02 -0.001250066 -0.008772262  0.0072123700
## T.Stage                   3.546411e-02  0.010000966 -0.022574764 -0.0174024531
## N.Stage                  -3.419320e-02  0.020737326 -0.009661249 -0.0196783808
## X6th.Stage               -1.462791e-04  0.016406238  0.002237891 -0.0319434541
## differentiate             5.874596e-02 -0.216888672  0.191723794 -0.0091378091
## Grade                     2.637628e-01 -0.631704321  0.618308634  0.0170305623
## A.Stage                  -4.812705e-02 -0.043898052  0.009163560 -0.9744433726
## Tumor.Size                4.676763e-04  0.015344409 -0.003506997 -0.1286491968
## Estrogen.Status          -7.896136e-01 -0.198985807  0.284693884  0.0513964046
## Progesterone.Status       3.858984e-01  0.008587728 -0.060699726 -0.0217422620
## Regional.Node.Examined    1.218478e-02  0.011907640  0.001786879 -0.1577403056
## Reginol.Node.Positive    -1.247795e-02  0.005491557 -0.011704119 -0.0319108355
## Survival.Months          -1.822363e-02 -0.669574232 -0.643387664  0.0268517426
## Status                   -8.728377e-02 -0.245497775 -0.263582467  0.0006436894
## Reginol.Node.PositiveNM  -1.247795e-02  0.005491557 -0.011704119 -0.0319108355
## Reginol.Node.PositiveND  -1.247795e-02  0.005491557 -0.011704119 -0.0319108355
##                                 PC13           PC14          PC15          PC16
## Age                       0.021529364 -0.0014426710 -0.020445198 -0.001375121
## Race                     -0.008794653 -0.0077358306 -0.002345647  0.002481384
## Marital.Status            0.002181798 -0.0008714382  0.006497223  0.009253269
## T.Stage                   0.048280221 -0.0160147800  0.197327041 -0.519398524
## N.Stage                  -0.225024517 -0.1858883239 -0.393892809 -0.432656407
## X6th.Stage               -0.111174817 -0.2596547026  0.399643471  0.616635707
```

```
## differentiate            0.015318348 -0.0044616070 -0.003267753 -0.001975332
## Grade                   -0.003298016  0.0197724993 -0.005296695 -0.008247770
## A.Stage                 -0.155020442 -0.0195484283  0.093765458 -0.051626314
## Tumor.Size               0.004997184  0.3021196459 -0.743383554  0.395431029
## Estrogen.Status          0.027819335 -0.0092461558 -0.002748381 -0.002471502
## Progesterone.Status     -0.015532169  0.0023327511 -0.005068386  0.005865695
## Regional.Node.Examined   0.909648678 -0.3403439433 -0.120754114  0.002799150
## Reginol.Node.Positive    0.164239695  0.4791991327  0.152205726 -0.033144953
## Survival.Months          0.002573038 -0.0020434040 -0.009746426  0.010452669
## Status                   0.020473271 -0.0284955366 -0.008397934  0.012002729
## Reginol.Node.PositiveNM  0.164239695  0.4791991327  0.152205726 -0.033144953
## Reginol.Node.PositiveND  0.164239695  0.4791991327  0.152205726 -0.033144953
##                                 PC17          PC18
## Age                     -4.293687e-18  3.488931e-19
## Race                    -1.589775e-17 -2.007606e-17
## Marital.Status           1.328598e-17 -2.904736e-18
## T.Stage                  1.141199e-16  2.673214e-17
## N.Stage                 -1.270470e-16  3.062983e-17
## X6th.Stage              -2.595027e-16 -1.289923e-16
## differentiate            3.209169e-18  1.587923e-17
## Grade                    5.309036e-18 -8.197515e-19
## A.Stage                  2.656791e-17  1.528938e-17
## Tumor.Size              -1.503397e-17  1.163834e-17
## Estrogen.Status         -7.814572e-18 -5.432460e-20
## Progesterone.Status      3.225598e-17  1.003709e-18
## Regional.Node.Examined  -1.343335e-16 -4.179375e-18
## Reginol.Node.Positive   -8.119222e-01 -8.630742e-02
## Survival.Months         -1.700543e-17  9.722065e-18
## Status                  -7.933271e-18 -8.264976e-18
## Reginol.Node.PositiveNM  3.312167e-01  7.462990e-01
## Reginol.Node.PositiveND  4.807055e-01 -6.599916e-01
```

Vemos que no hay ninguna varianza superior a 1. Vamos a probar de escalar los datos para poders seleccionar los datos a escojer.

```
BreastCancer_scale <- scale(BreastCancer)
pca.BreastCancer_scale <- prcomp(BreastCancer)
var_BreastCancer_scale <- pca.BreastCancer_scale$sdev^2
head(var_BreastCancer_scale)
```

```
## [1] 0.29312745 0.16755455 0.12322632 0.11533165 0.08439501 0.07170449
```

Podemos observara como aún así no encontramos valores superiores a uno. Así que vamos a establacer un umbral de un 80-90% de varianza acumulada explicada, que en este caso sería hasta la dimensión 8.

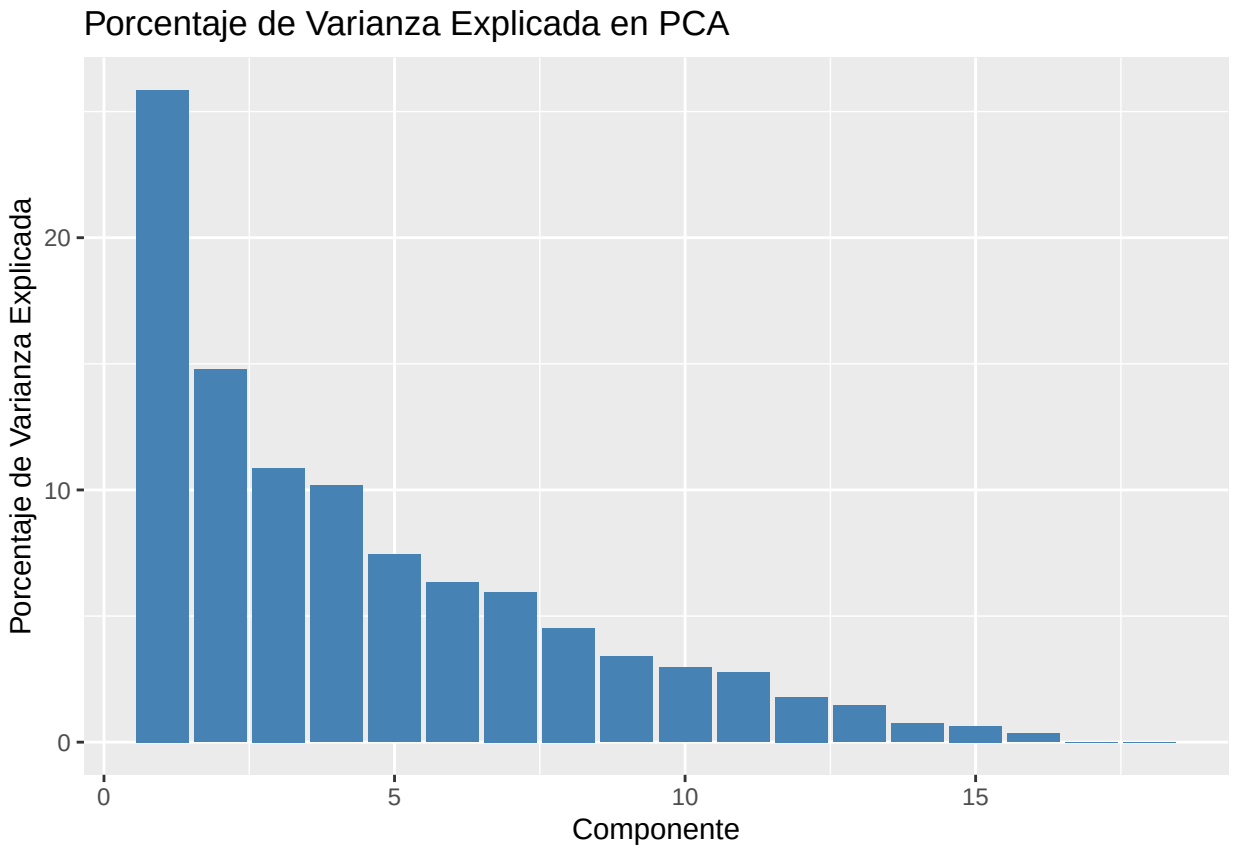
```
library(ggplot2)

valores_propios <- pca.BreastCancer_scale$sdev^2
porcentaje_varianza_explicada <- valores_propios / sum(valores_propios) * 100

# Crear un dataframe con los resultados
df <- data.frame(Componente = 1:length(porcentaje_varianza_explicada),
                 Porcentaje = porcentaje_varianza_explicada)

# Graficar el porcentaje de varianza explicada
ggplot(df, aes(x = Componente, y = Porcentaje)) +
```

```
  geom_bar(stat = "identity", fill = "steelblue") +
  xlab("Componente") +
  ylab("Porcentaje de Varianza Explicada") +
  ggtitle("Porcentaje de Varianza Explicada en PCA")
```



La calidad de representación de las variables en el mapa de factores se llama cos2 (coseno cuadrado, coordenadas cuadradas). Podemos acceder al coz2 de la siguiente manera:

```
var <- get_pca_var(pca.BreastCancer_scale)
var
```

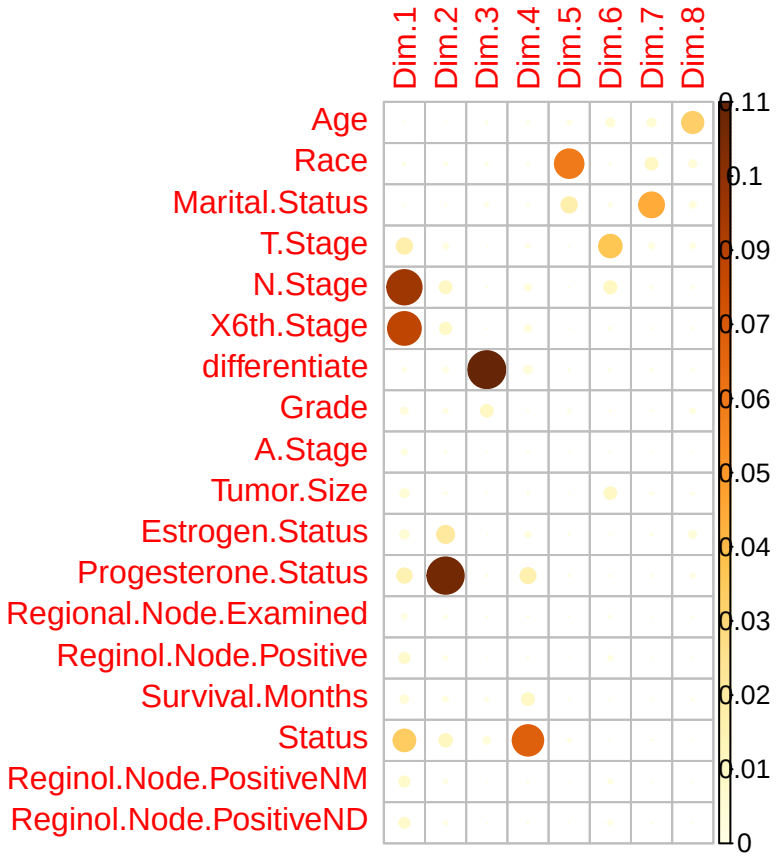
```
## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"
```

```
head(var$coord[,1:10],16)
```

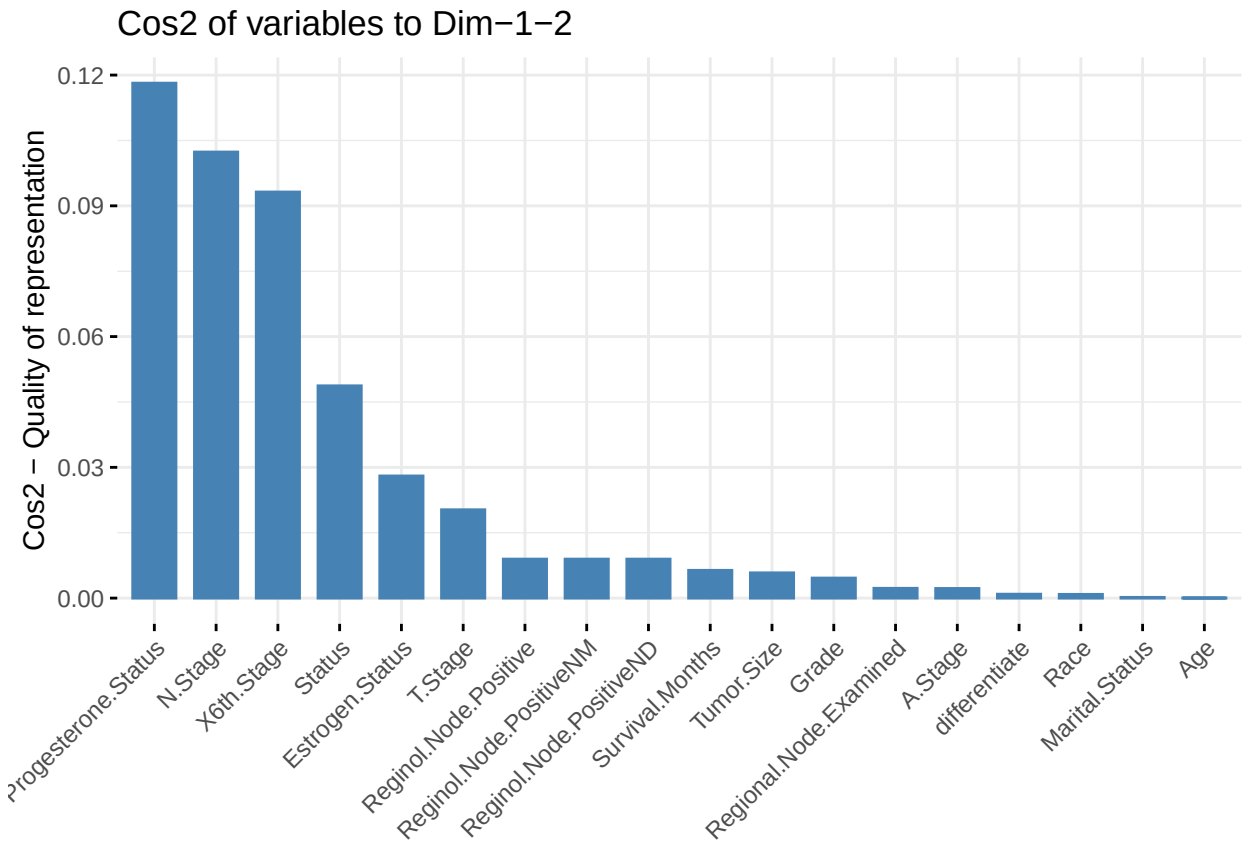
```
##                          Dim.1         Dim.2        Dim.3        Dim.4
## Age                -0.002130922 -0.0061609643  0.016437161 -0.020206258
## Race               -0.018132901  0.0230276696 -0.017049974  0.016422010
## Marital.Status      0.007161110 -0.0106444857  0.026240964 -0.015497594
## T.Stage             0.134820573  0.0456942706 -0.005216253  0.026959742
## N.Stage             0.302333649  0.1046008480  0.008589161  0.053047082
## X6th.Stage          0.288486730  0.0997460878  0.002331294  0.055612496
```

```
## differentiate          -0.029992167 -0.0001107534  0.326024073  0.062622727
## Grade                   0.057615811 -0.0358294629 -0.103655655 -0.010560331
## A.Stage                -0.044769738 -0.0142268605 -0.004850384 -0.006777762
## Tumor.Size              0.072743882  0.0224256848 -0.003018611  0.016167684
## Estrogen.Status        -0.073651766  0.1503336634 -0.003846349 -0.046107184
## Progesterone.Status    -0.126434340  0.3196312879  0.013964229 -0.132417468
## Regional.Node.Examined  0.042977801  0.0199026179 -0.007558118  0.014988957
## Reginol.Node.Positive   0.089424634  0.0307855310  0.003742250  0.012413231
## Survival.Months        -0.067756192  0.0422059461 -0.028792119  0.105624969
## Status                  0.191265215 -0.1100921204  0.060528412 -0.268208211
##                                Dim.5        Dim.6         Dim.7        Dim.8
## Age                    -0.0364267398  0.070580450  0.0716062445 -0.186698685
## Race                   -0.2513193842  0.013421147  0.1042617637  0.063148650
## Marital.Status          0.1345049971  0.029479963  0.2185911683  0.052868800
## T.Stage                -0.0063324954 -0.197786587  0.0437588148 -0.039944871
## N.Stage                 0.0054351138  0.106441285 -0.0181726068  0.016026733
## X6th.Stage             -0.0005728981 -0.020539468  0.0058827215 -0.011120631
## differentiate          -0.0106334219 -0.013787639 -0.0155639193  0.016285827
## Grade                   0.0210775254 -0.011088051 -0.0116804701  0.040600530
## A.Stage                 0.0011374844  0.010438201  0.0004842717 -0.002135314
## Tumor.Size             -0.0024776223 -0.102245612  0.0215930741 -0.018594540
## Estrogen.Status        -0.0076765215  0.007257096  0.0145619984 -0.062366590
## Progesterone.Status     0.0169650811 -0.009747527 -0.0118932092  0.033457424
## Regional.Node.Examined  0.0017380666  0.016737188 -0.0046252961  0.009069551
## Reginol.Node.Positive  -0.0007755452  0.033661865 -0.0055622007  0.004884785
## Survival.Months         0.0068108005  0.007733112  0.0059495236 -0.015675020
## Status                 -0.0279712041 -0.006276950 -0.0068213231  0.002077939
##                                Dim.9       Dim.10
## Age                     7.337451e-02 -0.008063087
## Race                    1.887965e-05 -0.004098048
## Marital.Status         -8.504377e-03 -0.000228636
## T.Stage                 6.969796e-03  0.001829168
## N.Stage                -6.720023e-03  0.003792839
## X6th.Stage             -2.874838e-05  0.003000687
## differentiate           1.154540e-02 -0.039668748
## Grade                   5.183757e-02 -0.115538167
## A.Stage                -9.458457e-03 -0.008028915
## Tumor.Size              9.191290e-05  0.002806479
## Estrogen.Status        -1.551835e-01 -0.036394330
## Progesterone.Status     7.584099e-02  0.001570688
## Regional.Node.Examined  2.394687e-03  0.002177897
## Reginol.Node.Positive  -2.452303e-03  0.001004401
## Survival.Months        -3.581508e-03 -0.122464540
## Status                 -1.715397e-02 -0.044901328
```

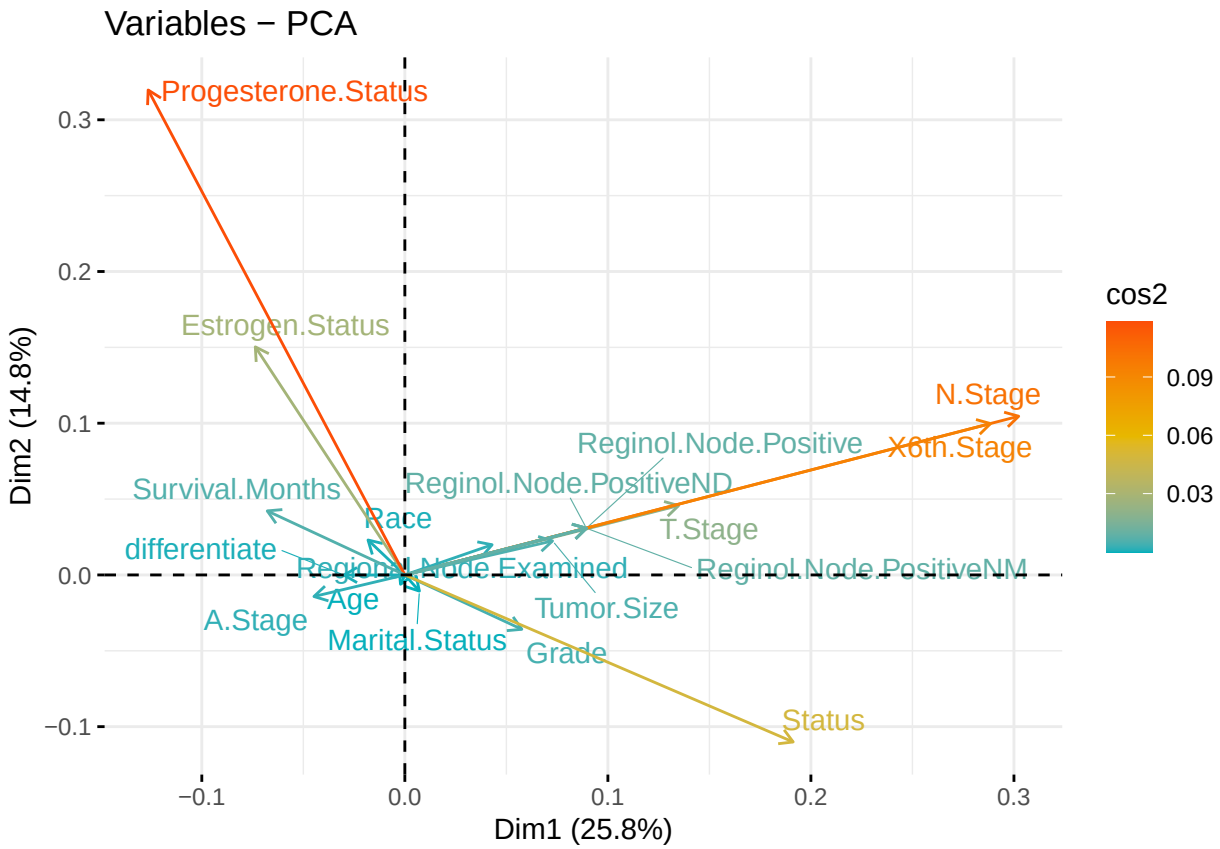
```
corrplot(var$cos2[,1:8], is.corr=FALSE)
```




```
fviz_cos2(pca.BreastCancer_scale, choice = "var", axes = 1:2)
```




```
fviz_pca_var(pca.BreastCancer,
col.var = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE
)
```



Los valores de cos2 se utilizan para estimar la calidad de la representación Cuanto más cercana esté una variable en el círculo de correlaciones, mejor será su representación en mapa de factores (y más importante es interpretar estos componentes) Las variables cercanas al centro de la trama son menos importantes para los primeros componentes.

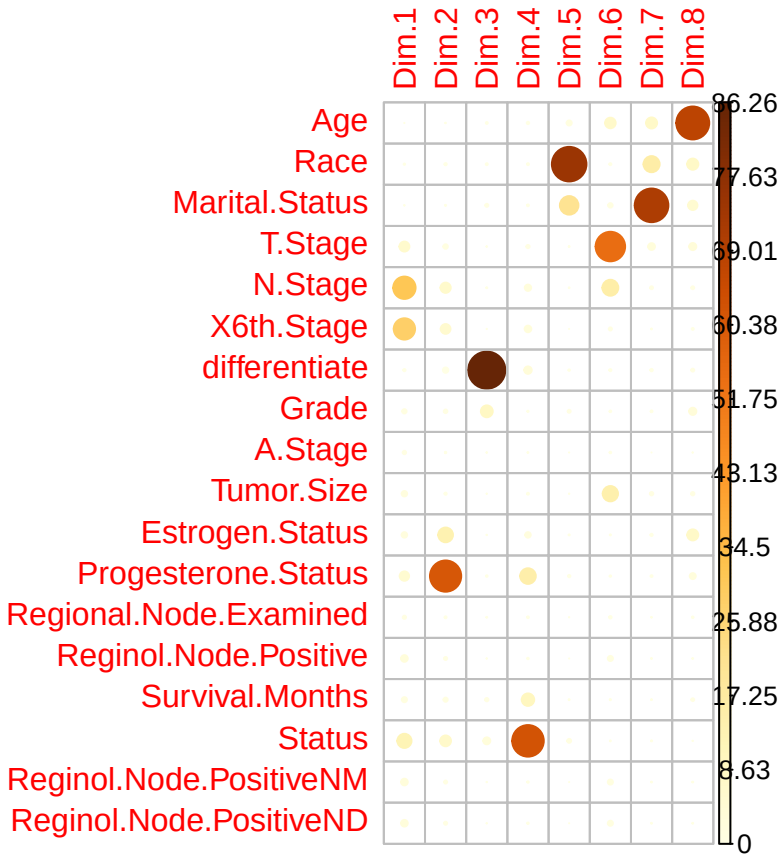
Las contribuciones de las variables en la contabilización de la variabilidad de un determinado componente principal se expresan en porcentaje. Las variables que no están correlacionadas con ningún PC o con las últimas dimensiones son variables con una contribución baja y pueden eliminarse para simplicar el análisis global. La contribución de las variables puede extraerse de la siguiente manera:

```
head(var$contrib[,1:8],16)
```

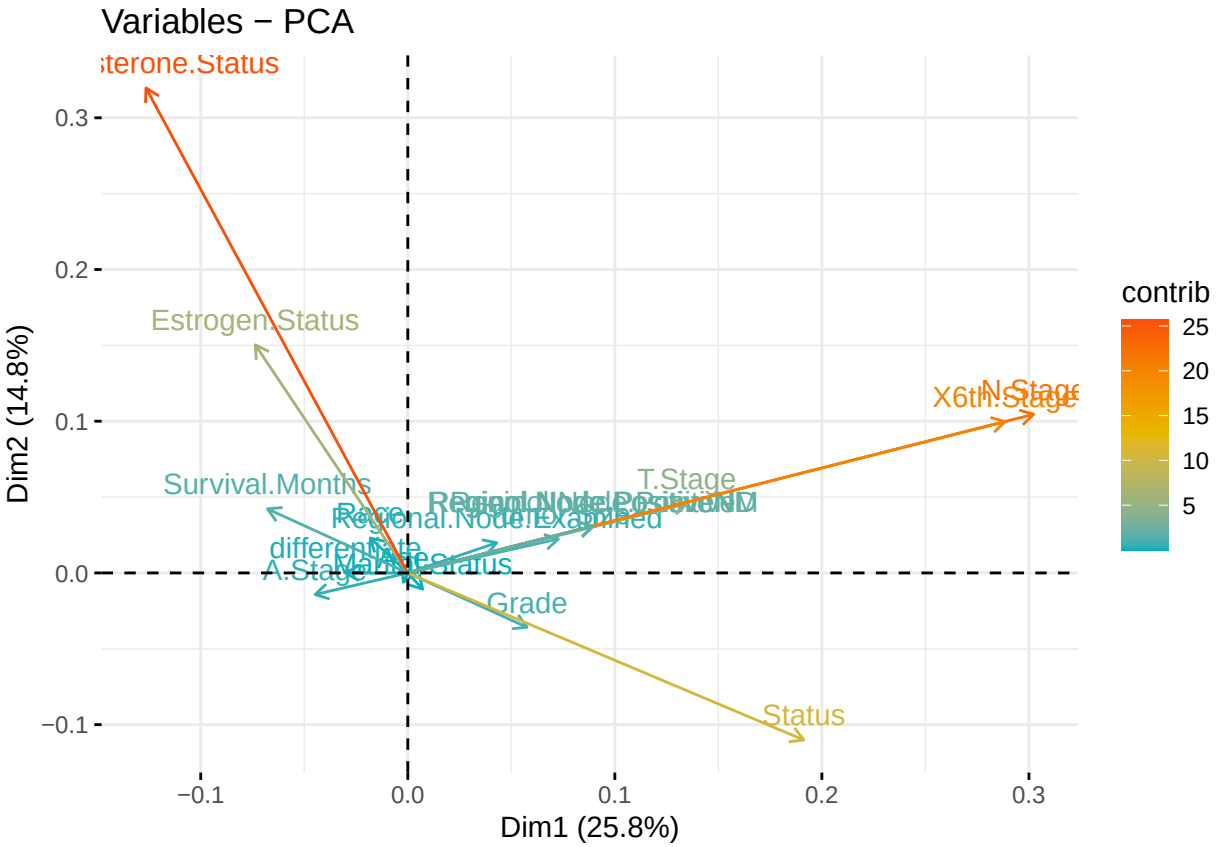
```
##                                Dim.1        Dim.2        Dim.3       Dim.4
## Age                      0.001549096 2.265381e-02  0.219255306  0.35401631
## Race                     0.112170351 3.164782e-01  0.235908697  0.23383210
## Marital.Status           0.017494606 6.762280e-02  0.558799574  0.20824763
## T.Stage                  6.200915978 1.246141e+00  0.022080751  0.63020664
## N.Stage                 31.182898730 6.530015e+00  0.059868449  2.43991385
## X6th.Stage              28.391948072 5.937936e+00  0.004410527  2.68161406
## differentiate            0.306873366 7.320790e-06 86.257296539  3.40028600
## Grade                    1.132470415 7.661686e-01  8.719317810  0.09669556
## A.Stage                  0.683774047 1.207986e-01  0.019091884  0.03983127
## Tumor.Size               1.805246283 3.001478e-01  0.007394536  0.22664550
## Estrogen.Status          1.850588405 1.348827e+01  0.012005875  1.84326885
## Progesterone.Status      5.453478465 6.097367e+01  0.158245149 15.20344658
## Regional.Node.Examined   0.630132526 2.364091e-01  0.046357905  0.19480240
## Reginol.Node.Positive    2.728084707 5.656360e-01  0.011364805  0.13360453
## Survival.Months          1.566179340 1.063141e+00  0.672734616  9.67352336
## Status                  12.480026197 7.233629e+00  2.973137967 62.37285627
```

```
##                                Dim.5       Dim.6        Dim.7        Dim.8
## Age                     1.572258e+00  6.94740314 7.601347e+00 67.956554571
## Race                    7.484024e+01  0.25120768 1.611532e+01  7.774580687
## Marital.Status          2.143681e+01  1.21201369 7.083599e+01  5.449388023
## T.Stage                 4.751525e-02 54.55660364 2.838702e+00  3.110791408
## N.Stage                 3.500262e-02 15.80061069 4.895795e-01  0.500770643
## X6th.Stage              3.889000e-04  0.58834494 5.130330e-02  0.241105808
## differentiate           1.339767e-01  0.26511449 3.591093e-01  0.517092770
## Grade                   5.264080e-01  0.17146051 2.022597e-01  3.213751040
## A.Stage                 1.533113e-03  0.15195149 3.476698e-04  0.008889397
## Tumor.Size              7.273667e-03 14.57951281 6.912222e-01  0.674092840
## Estrogen.Status         6.982520e-02  0.07344791 3.143625e-01  7.583205741
## Progesterone.Status     3.410320e-01  0.13250814 2.096944e-01  2.182397355
## Regional.Node.Examined  3.579448e-03  0.39067771 3.171523e-02  0.160369007
## Reginol.Node.Positive   7.126847e-04  1.58026533 4.586507e-02  0.046520012
## Survival.Months         5.496415e-02  0.08339926 5.247508e-02  0.479032569
## Status                  9.270551e-01  0.05494789 6.898044e-02  0.008418104
```

```
corrplot(var$contrib[,1:8], is.corr=FALSE)
```




```
fviz_pca_var(pca.BreastCancer_scale, col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
)
```



Las variables correlacionadas positivas apuntan al mismo lado de la trama mientras que las negativas al lado opuesto. Se observa que las variables que más aportan a las componentes principaless són N Stage, X6th Stage y Progestorne Status. Observamos como X6th Stage y N Stage estan correlacionadas. De otro lado Progestore Status y Estrogen Status están correlacionadas.

Por lo tanto, se podría concluir que las carácteristicas principales a la hora de determinar un tumor y clasificarlo son aquellas variables que tienen una gran correlación que en este caso serian Progesterone Status, X6th Stage y N Stage.

Rúbrica

Criterios de valoración

Para todas las PRA es **necesario documentar** en cada apartado del ejercicio práctico que se ha hecho, por qué se ha hecho y cómo se ha hecho. Asimismo, todas las decisiones y conclusiones deberán ser presentados de forma razonada y clara, **contextualizando los resultados**, es decir, especificando todos y cada uno de los pasos que se hayan llevado a cabo para su resolución.

- 20% Se plantea un problema propio de minería de datos, se detallan los objetivos analíticos y se explica detalladamente el procedimiento para darles solución.
- 10%. Justificación de la elección del juego de datos donde se detalle el potencial analítico que se intuye. El estudiante deberá visitar los siguientes portales de datos abiertos para seleccionar su juego de datos:
- **Datos abiertos**

  - Google Dataset Search
  - Datos abiertos España
  - Datos abiertos Madrid
  - Datos abiertos Barcelona
  - Datos abiertos Londres
  - Datos abiertos New York

- **Conjuntos de datos para aprendizaje automático e investigación**
  - UCI Machine Learning
  - Datasets for machine-learning research (Wikipedia)
  - Kaggle datasets

Recordad que el mismo dataset **deberá tener capacidades** para que se le puedan aplicar **algoritmos supervisados**, **algoritmos no supervisados** y **reglas de asociación**.

- 20%. Información extraída del análisis exploratorio. Distribuciones, correlaciones, anomalías,...
- 20%. Explicación clara de cualquier tarea de limpieza o acondicionado que se realiza. Justificando el motivo y mencionando las ventajas de la acción tomada.
- 20%. Se realiza un proceso de PCA o SVD donde se aprecia mediante explicaciones y comentarios que el estudiante entiende todos los pasos y se comenta extensamente el resultado final obtenido.
- 10%. Consideración general
  - Se presenta el código y es fácilmente reproducible.
  - Se detalla cada pregunta de manera correcta, mostrando el código, comentando como se ha hecho y porque se ha hecho, comparando los resultados y/o indicando otras alternativas al problema indicado.
  - Se muestran las conclusiones en cada apartado
  - Se indican eventuales citaciones bibliográficas, fuentes internas/externas y materiales de investigación.

Recursos de programación

- Incluimos en este apartado una lista de recursos de programación para minería de datos donde podréis encontrar ejemplos, ideas e inspiración:
  - Material adicional del libro: Minería de datos Modelos y Algoritmos
  - Espacio de recursos UOC para ciencia de datos
  - Buscador de código R

  - Colección de cheatsheets en R
