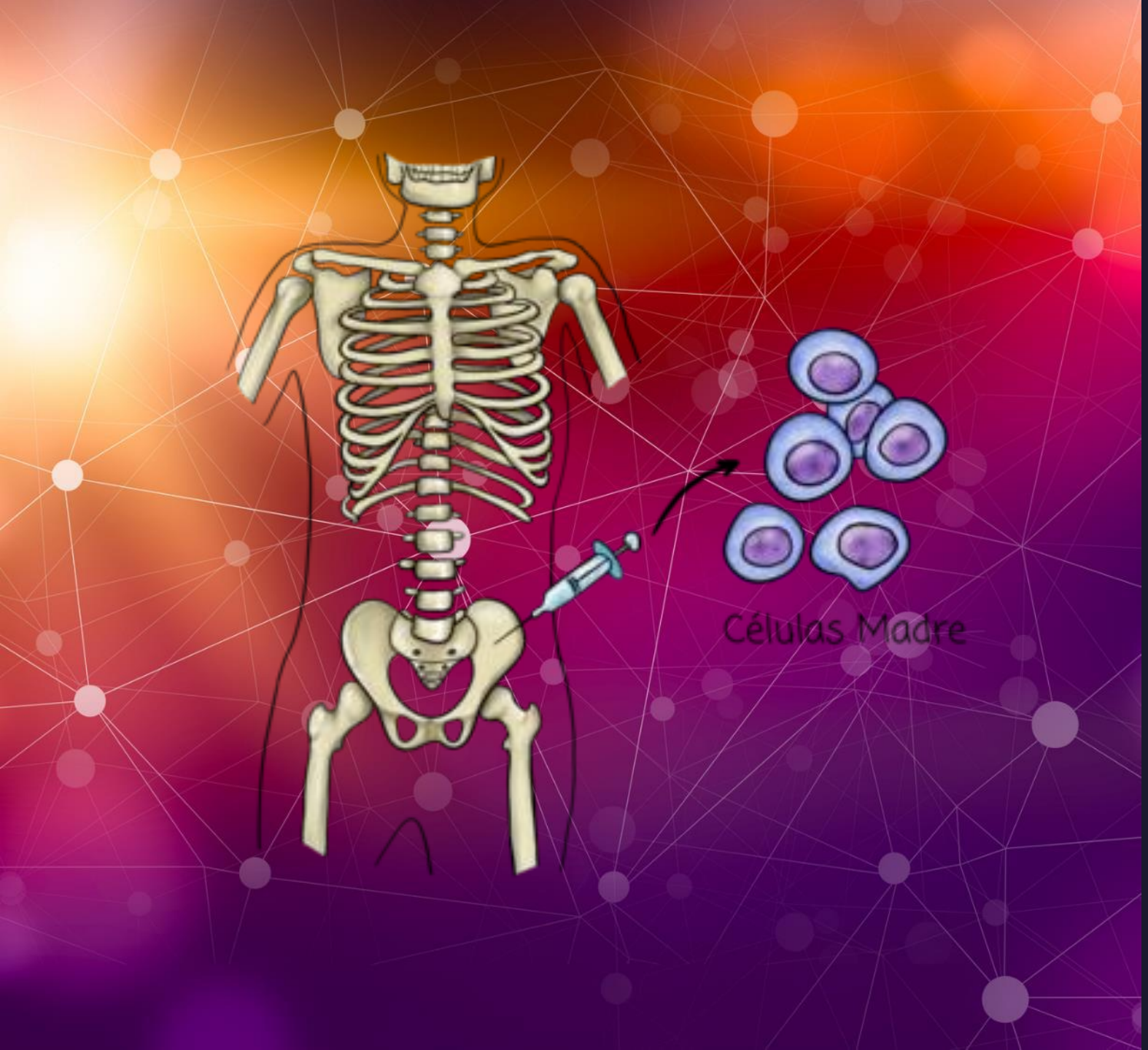


Bone marrow transplant: children Data Set

Paula Poley Ceballos

3ºISA

Análisis Avanzado de Datos Clínicos

ÍNDICE

1. Introducción

El conjunto de datos describe pacientes pediátricos con varias enfermedades hematológicas:

- trastornos malignos

  Es decir, pacientes con
  - leucemia linfoblástica aguda
  - leucemia mielógena aguda
  - leucemia mielógena crónica
  - síndrome mielodisplásico

- Casos no malignos

  Es decir, pacientes con:
  - anemia aplásica grave
  - anemia de Fanconi
  - Adrenoleucodistrofia ligada al cromosoma X

Todos los pacientes → sometieron al trasplante de células madre hematopoyéticas de un donante alogénico no emparentado.

2. GRÁFICOS

Sólo se mostraran algunos gráficos, todos los demás se pueden ver en el .Rmd del proyecto.

Vista de algunos de los datos del conjunto de datos

Lectura del fichero:

```{r}
alldatos<-read.delim("datosMedulaOsea.csv", sep =",", head=TRUE)
alldatos
```

datos | proyectofinal.Rmd*

Filter

	sexo.del.receptor	fuente.de.celulas.madre	donacion	donante.35	IIIV	compatibilidad.de.genero	donante.ABO	receptor.ABO	receptor.Rh	coincidencia.ABO	estado.de.CMV	CMV.del.donante	CMV.del.receptor
1	1	1	22.83014	0	1	0	1	1	1	0	3	1	1
2	1	0	23.34247	0	1	0	-1	-1	1	0	0	0	0
3	1	0	26.39452	0	1	0	-1	-1	1	0	2	0	1
4	0	0	39.68493	1	1	0	1	2	1	1	1	1	0
5	0	1	33.35890	0	0	0	1	2	0	1	0	0	1
6	1	0	27.39178	0	0	0	2	0	1	1	?	?	1
7	0	1	34.52055	0	1	0	0	1	0	1	?	0	?
8	1	0	21.43562	0	1	0	0	1	1	1	1	1	0
9	1	1	32.64110	0	0	0	2	0	1	1	2	0	1
10	1	1	28.78356	0	1	1	1	0	1	1	2	0	1
11	0	1	29.73151	0	0	0	0	-1	1	1	1	1	0
12	0	1	36.80000	1	1	0	1	1	1	0	0	0	0
13	1	1	40.86301	1	0	0	0	1	1	1	1	1	0
14	1	0	30.60274	0	1	0	0	1	1	1	0	0	1
15	1	1	30.67671	0	1	0	1	1	1	0	0	0	0
16	1	0	41.16438	1	0	0	0	-1	1	1	3	1	1
17	1	1	40.23288	1	1	0	2	0	1	1	1	1	0
18	0	1	40.82192	1	1	0	2	0	1	1	1	1	0
19	0	1	43.37534	1	1	0	0	1	1	1	?	0	?
20	1	1	31.74520	0	0	0	1	0	1	1	3	1	1
21	1	1	43.57808	1	0	1	0	1	1	1	3	1	1

Si usamos la función str, podemos ver o explorar la estructura del dataframe que contiene el conjunto de datos. Como podemos ver en la línea superior 'data.frame': , el conjunto de datos tiene 187 observaciones y 36 variables en cada una (ya que anteriormente hemos eliminado una). También podemos ver que tenemos registros int (números enteros), num(numérico), chr (carácter)

```{r}
str(datos)
```

```
'data.frame':   187 obs. of  36 variables:
 $ sexo.del.receptor        : int  1 1 1 0 0 1 0 1 1 1 ...
 $ fuente.de.celulas.madre  : int  1 0 0 0 1 0 1 0 1 1 ...
 $ donacion                 : num  22.8 23.3 26.4 39.7 33.4 ...
 $ donante.35               : int  0 0 0 1 0 0 0 0 0 0 ...
 $ IIIV                     : int  1 1 1 1 0 0 1 1 0 1 ...
 $ compatibilidad.de.genero : int  0 0 0 0 0 0 0 0 0 1 ...
 $ donante.ABO              : int  1 -1 -1 1 1 2 0 0 2 1 ...
 $ receptor.ABO             : chr  "1" "-1" "-1" "2" ...
 $ receptor.Rh              : chr  "1" "1" "1" "1" ...
 $ coincidencia.ABO         : chr  "0" "0" "0" "1" ...
 $ estado.de.CMV            : chr  "3" "0" "2" "1" ...
 $ CMV.del.donante          : chr  "1" "0" "0" "1" ...
 $ CMV.del.receptor         : chr  "1" "0" "1" "0" ...
 $ enfermedad               : chr  "ALL" "ALL" "ALL" "AML" ...
 $ grupo.de.riesgo          : int  1 0 0 0 1 1 0 0 0 0 ...
 $ Txpost.recaida           : int  0 0 0 0 0 1 0 0 0 0 ...
 $ grupo.de.enfermedad      : int  1 1 1 1 1 1 1 0 0 0 ...
 $ coincidencia.HLA         : int  0 0 0 0 1 0 0 3 0 1 ...
 $ discrepancia.de.HLA      : int  0 0 0 0 0 0 0 1 0 0 ...
 $ antigeno                 : chr  "-1" "-1" "-1" "-1" ...
 $ alelo                    : chr  "-1" "-1" "-1" "-1" ...
 $ edad.del.receptor        : num  9.6 4 6.6 18.1 1.3 8.9 14.4 18.2 7.9 4.7 ...
 $ receptor.10              : int  0 0 0 1 0 0 1 1 0 0 ...
 $ agente.del.receptor      : int  1 0 1 2 0 1 2 2 1 0 ...
 $ recaida                  : int  0 1 1 0 0 0 0 0 0 0 ...
 $ aGvHDIIIIV               : int  0 1 1 0 1 1 0 0 1 1 ...
 $ extcGvHD                 : chr  "1" "1" "1" "?" ...
 $ CD34kgx10d6              : num  7.2 4.5 7.94 4.25 51.85 ...
 $ CD3.CD34                 : chr  "1.33876" "11.078295" "19.01323" "29.481647" ...
 $ CD3dkgx10d8              : chr  "5.38" "0.41" "0.42" "0.14" ...
 $ masa.corporal            : chr  "35" "20.6" "23.4" "50" ...
 $ ANC.recuperacion         : int  19 16 23 23 14 16 17 22 15 16 ...
 $ PLT.recuperacion         : int  51 37 20 29 14 70 29 58 14 17 ...
 $ tiempo.hasta.aGvHD.III_IV: int  32 1000000 1000000 19 1000000 1000000 18 22 1000000 1000000 ...
 $ tiempo.de.supervivencia  : int  999 163 435 53 2043 2800 41 45 671 676 ...
 $ estado.de.supervivencia  : int  0 1 1 1 0 0 1 1 0 0 ...
```

Convertimos estas variables con la información dada.

```r
datos$sexo.del.receptor<-factor(datos$sexo.del.receptor, levels=c("0","1"),labels=c("Femenino","Masculino"))
datos$fuente.de.celulas.madre<-factor(datos$fuente.de.celulas.madre, levels=c("0","1"),labels=c("Médula ósea","Sangre periférica"))
datos$donante.35<-factor(datos$donante.35,levels=c("0","1"),labels=c("Menor de 35 años","Mayor o igual a 35 años"))
datos$IIIV<-factor(datos$IIIV, levels=c("0","1"),labels=c("No","Si"))
datos$compatibilidad.de.genero<-factor(datos$compatibilidad.de.genero, levels=c("0","1"),labels=c("Masculino a Femenino","Femenino a Masculino"))
datos$donante.ABO<-factor(datos$donante.ABO, levels=c("-1","0","1","2"),labels=c("B","O","A","AB"))
datos$receptor.ABO<-factor(datos$receptor.ABO, levels=c("-1","0","1","2"),labels=c("B","O","A","AB"))
datos$receptor.Rh<-factor(datos$receptor.Rh, levels=c("0","1"),labels=c("Rh-","Rh+"))
datos$coincidencia.ABO<-factor(datos$coincidencia.ABO, levels=c("0","1"),labels=c("No emparejados","Si emparejado"))
datos$CMV.del.donante<-factor(datos$CMV.del.donante, levels=c("0","1"),labels=c("Ausencia de infección","Presencia de infección"))
datos$CMV.del.receptor<-factor(datos$CMV.del.receptor, levels=c("0","1"),labels=c("Ausencia de infección","Presencia de infección"))
datos$grupo.de.riesgo<-factor(datos$grupo.de.riesgo, levels=c("0","1"),labels=c("Bajo riesgo","Alto riesgo"))
datos$Txpost.recaida<-factor(datos$Txpost.recaida, levels=c("0","1"),labels=c("No","Si"))
datos$grupo.de.enfermedad<-factor(datos$grupo.de.enfermedad, levels=c("0","1"),labels=c("No maligna","Maligna"))
datos$coincidencia.HLA<-factor(datos$coincidencia.HLA, levels=c("0","1","2","3"),labels=c(" 10/10","9/10", "8/10", "7/10"))
datos$discrepancia.de.HLA<-factor(datos$discrepancia.de.HLA, levels=c("0","1"),labels=c(" Coincidencia de HLA","Discrepancia de HLA"))
datos$antigeno<-factor(datos$antigeno, levels=c("-1","0","1"),labels=c("Sin diferencias","Con 1 diferencia", "Con 2 o 3 diferencias"))
datos$alelo<-factor(datos$alelo, levels=c("-1","0","1"),labels=c(" Sin diferencias","Con 1 diferencia", "Con 2,3 o 4 diferencias "))
datos$receptor.10<-factor(datos$receptor.10, levels=c("0","1"),labels=c(" Menos de 10 años","!0 o más años"))
datos$agente.del.receptor<-factor(datos$agente.del.receptor, levels=c("0","1","2"),labels=c(" De 0 a 5 años inclusive","De 5 a 10 años inclusive", "De 10 a 20 años
inclusive"))
datos$recaida<-factor(datos$recaida, levels=c("0","1"),labels=c("No","Si"))
datos$aGvHDIIIIV<-factor(datos$aGvHDIIIIV, levels=c("0","1"),labels=c("Si","No"))
datos$extcGvHD<-factor(datos$extcGvHD, levels=c("0","1"),labels=c("Si","No"))
datos$estado.de.supervivencia<-factor(datos$estado.de.supervivencia, levels=c("0","1"),labels=c("Vivo","Muerto"))

```

Con la función summary podemos tener un resumen estadístico de las variables del dataset, es decir, hacemos un sumario de los datos para verlos facilmente y comprobar errores.

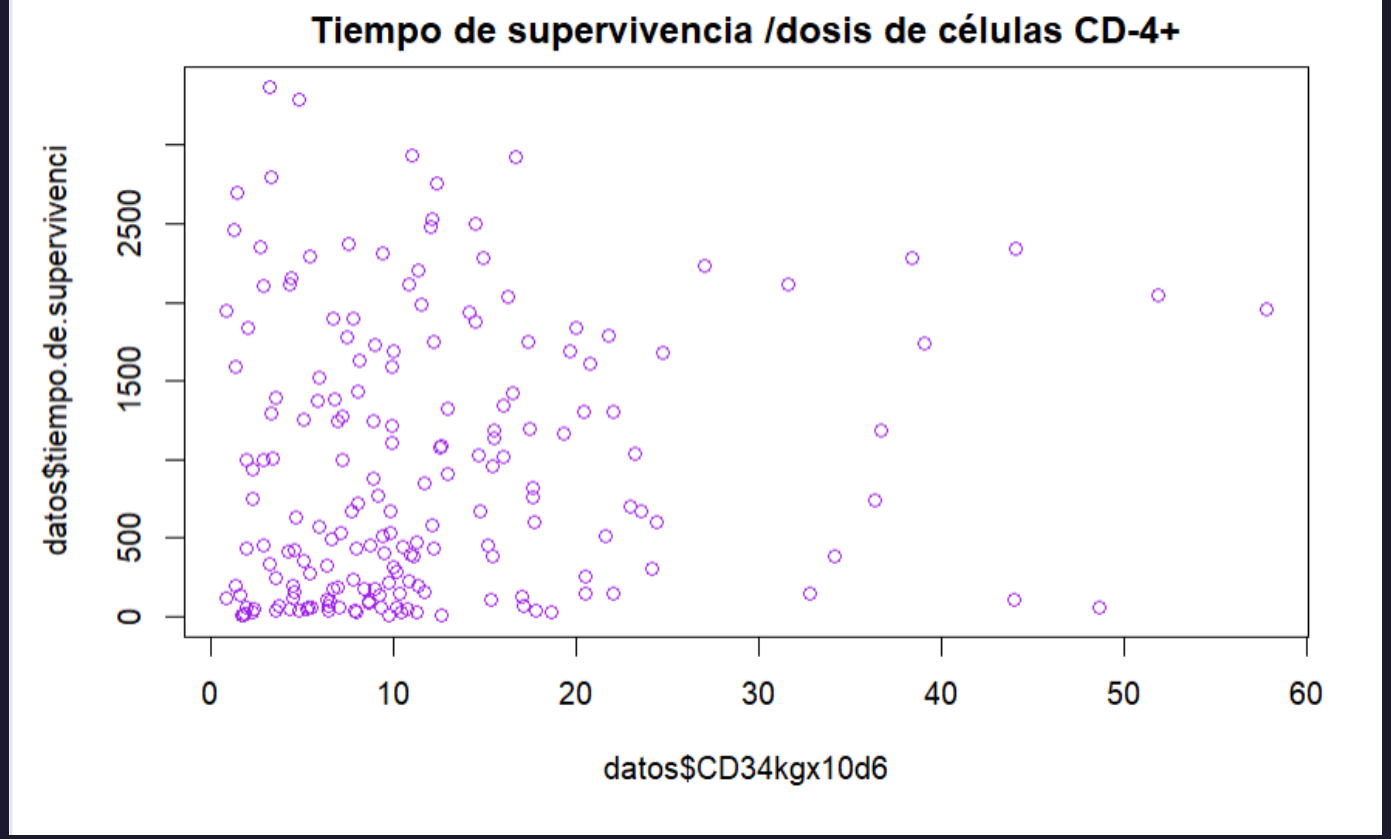
```{r}
summary(datos)
```

```
 sexo.del.receptor      fuente.de.celulas.madre    donacion                    donante.35  IIIV           compatibilidad.de.genero donante.ABO receptor.ABO
 Femenino : 75     Médula ósea      : 42        Min.   :18.65   Menor de 35 años       :104   No: 75   Masculino a Femenino:155        B :28       B   :50
 Masculino:112     Sangre periférica:145        1st Qu.:27.04   Mayor o igual a 35 años: 83   Si:112   Femenino a Masculino: 32        O :73       O   :48
                                                Median :33.55                                                                          A :71       A   :75
                                                Mean   :33.47                                                                          AB:15       AB  :13
                                                3rd Qu.:40.12                                                                                      NA's: 1
                                                Max.   :55.55
 receptor.Rh       coincidencia.ABO estado.de.CMV                CMV.del.donante              CMV.del.receptor  enfermedad          grupo.de.riesgo
 Rh- : 27    No emparejados: 52     Length:187         Ausencia de infección :113    Ausencia de infección : 73     Length:187         Bajo riesgo:118
 Rh+ :158    Si emparejado :134     Class :character   Presencia de infección: 72    Presencia de infección:100     Class :character   Alto riesgo: 69
 NA's:  2    NA's          :  1     Mode  :character   NA's                   :  2    NA's                   : 14     Mode  :character
 Txpost.recaida grupo.de.enfermedad coincidencia.HLA          discrepancia.de.HLA                 antigeno                      alelo    edad.del.receptor
 No:164         No maligna: 32       10/10:94        Coincidencia de HLA:159      Sin diferencias      :93    Sin diferencias        :93   Min.   : 0.600
 Si: 23         Maligna   :155      9/10  :65        Discrepancia de HLA : 28     Con 1 diferencia     :21    Con 1 diferencia       :54   1st Qu.: 5.050
                                    8/10  :23                                     Con 2 o 3 diferencias:65    Con 2,3 o 4 diferencias :32   Median : 9.600
                                    7/10  : 5                                      NA's                 : 8    NA's                   : 8   Mean   : 9.932
                                                                                                                                          3rd Qu.:14.050
                                                                                                                                          Max.   :20.200
          receptor.10               agente.del.receptor recaida  aGvHDIIIIV extcGvHD    CD34kgx10d6      CD3.CD34         CD3dkgx10d8
  Menos de 10 años:99    De 0 a 5 años inclusive :47       No:159   Si: 40     Si  : 28   Min.   : 0.79   Length:187         Length:187
 !0 o más años    :88   De 5 a 10 años inclusive :51       Si: 28   No:147     No  :128   1st Qu.: 5.35   Class :character   Class :character
                        De 10 a 20 años inclusive:89                           NA's: 31   Median : 9.72   Mode  :character   Mode  :character
                                                                                          Mean   :11.89
                                                                                          3rd Qu.:15.41
                                                                                          Max.   :57.78
 masa.corporal      ANC.recuperacion  PLT.recuperacion  tiempo.hasta.aGvHD.III_IV tiempo.de.supervivencia estado.de.supervivencia
 Length:187         Min.   :      9   Min.   :      9   Min.   :     10           Min.   :   6.0          Vivo  :102
 Class :character   1st Qu.:     13   1st Qu.:     16   1st Qu.:1000000           1st Qu.: 168.5          Muerto: 85
 Mode  :character   Median :     15   Median :     21   Median :1000000           Median : 676.0
                    Mean   :  26753   Mean   :  90938   Mean   : 775408           Mean   : 938.7
                    3rd Qu.:     17   3rd Qu.:     37   3rd Qu.:1000000           3rd Qu.:1604.0
                    Max.   :1000000   Max.   :1000000   Max.   :1000000           Max.   :3364.0
```

Podemos representar variables del conjunto de datos gráficamente y mediante la observación de la nube de puntos ver a groso modo si existe correlación.

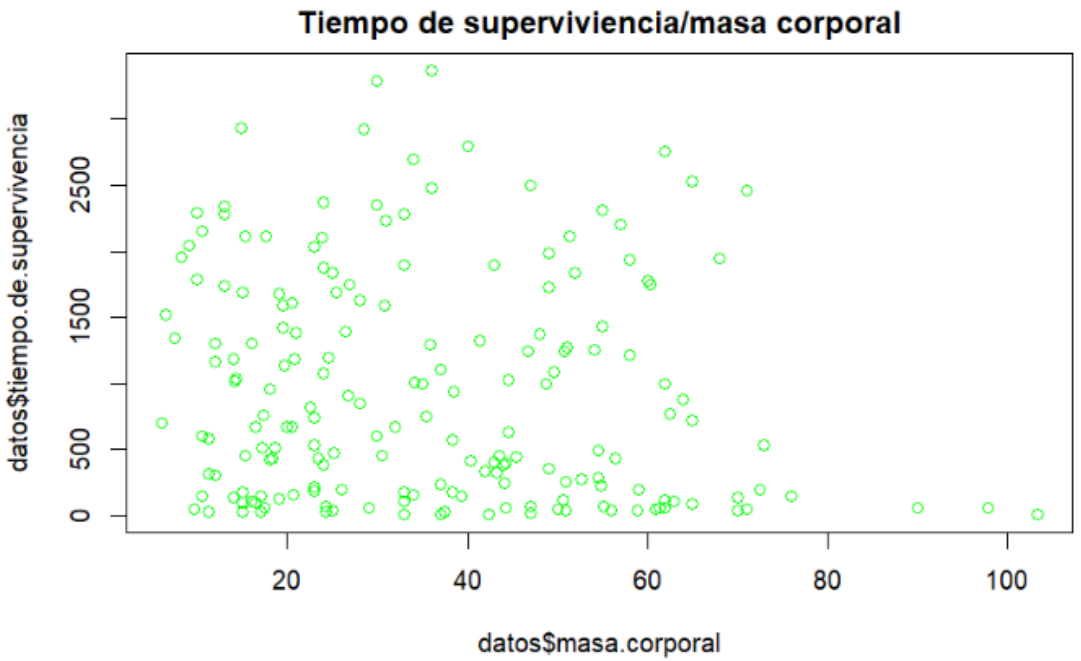
Vamos a ver el tiempo de supervivencia del receptor con la dosis de células CD 4+ por kg de peso corporal del receptor.

```{r}
plot(datos$CD34kgx10d6,datos$tiempo.de.supervivenci, col = "purple",
     main = "Tiempo de supervivencia /dosis de células CD-4+")
```



El tiempo de supervivencia del receptor es decir, el tiempo de observación (si está vivo) o tiempo hasta el evento (si está muerto) en días con la masa corporal del receptor.

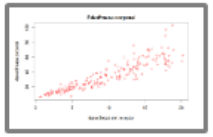
```r
plot(datos$masa.corporal,datos$tiempo.de.supervivencia, col = "green",
     main = "Tiempo de superviviencia/masa corporal")
```



Esta por ejemplo tendría más correlación, la edad del receptor de células madres hematopoyéticas en el momento del trasplante con la masa corporal del receptor de células madre hematopoyéticas en el momento del trasplante.

```{r}
plot(datos$edad.del.receptor,datos$masa.corporal, col = "red",
     main = "Edad/masa corporal")
```

R Console



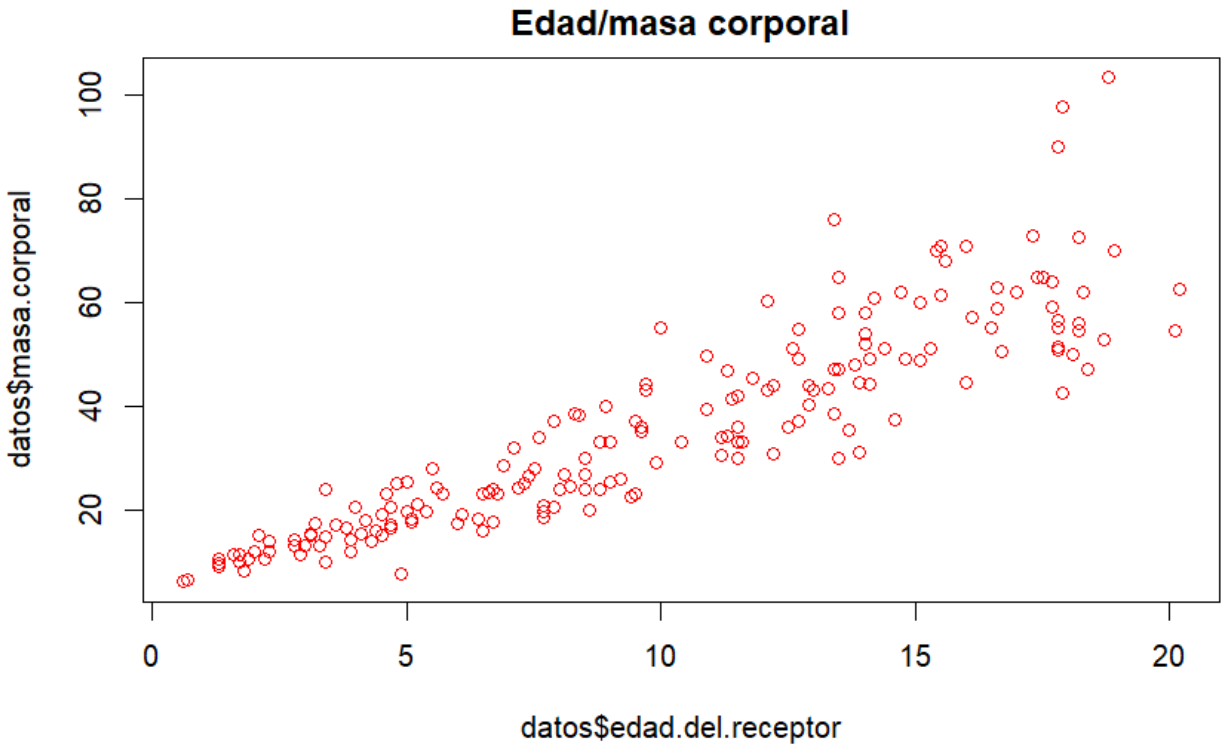


Gráfico de la recaída, es decir de la reaparición de la enfermedad con respecto el sexo del receptor. Podemos ver que tanto del sexo masculino como del sexo femenino hay más pacientes que no han recaído frente a los que si.

```{r}
graf1<-datos%>%
 ggplot(aes(x=sexo.del.receptor, fill=recaida))+
 geom_bar()
graf1
```

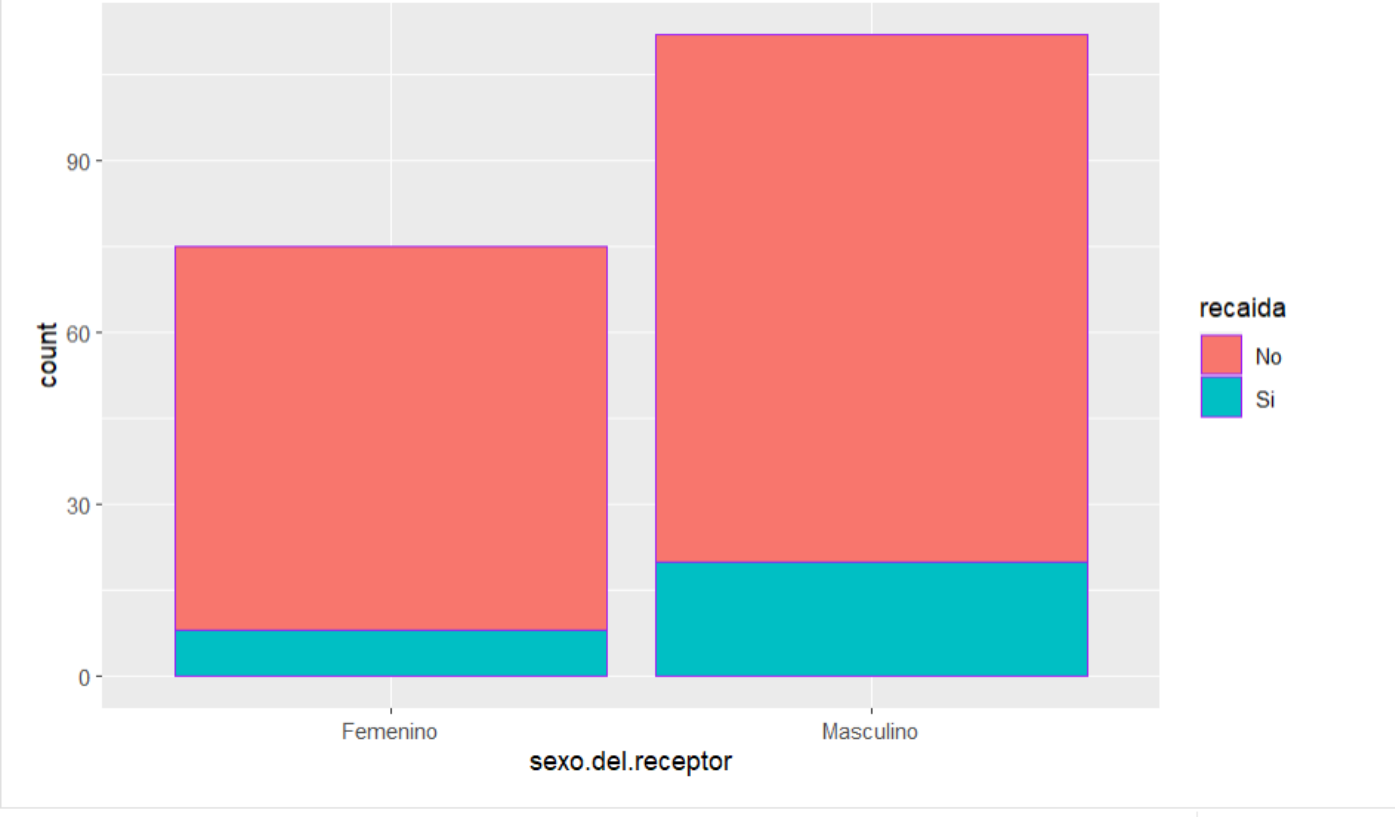


Gráfico de la coincidencia de HLA(antígenos leucocitarios humanos) con respecto el tipo de enfermedad.

```{r}
graf11<-datos%>%
 ggplot(aes(x=discrepancia.de.HLA, fill=grupo.de.enfermedad))+
 geom_bar()
graf11
```

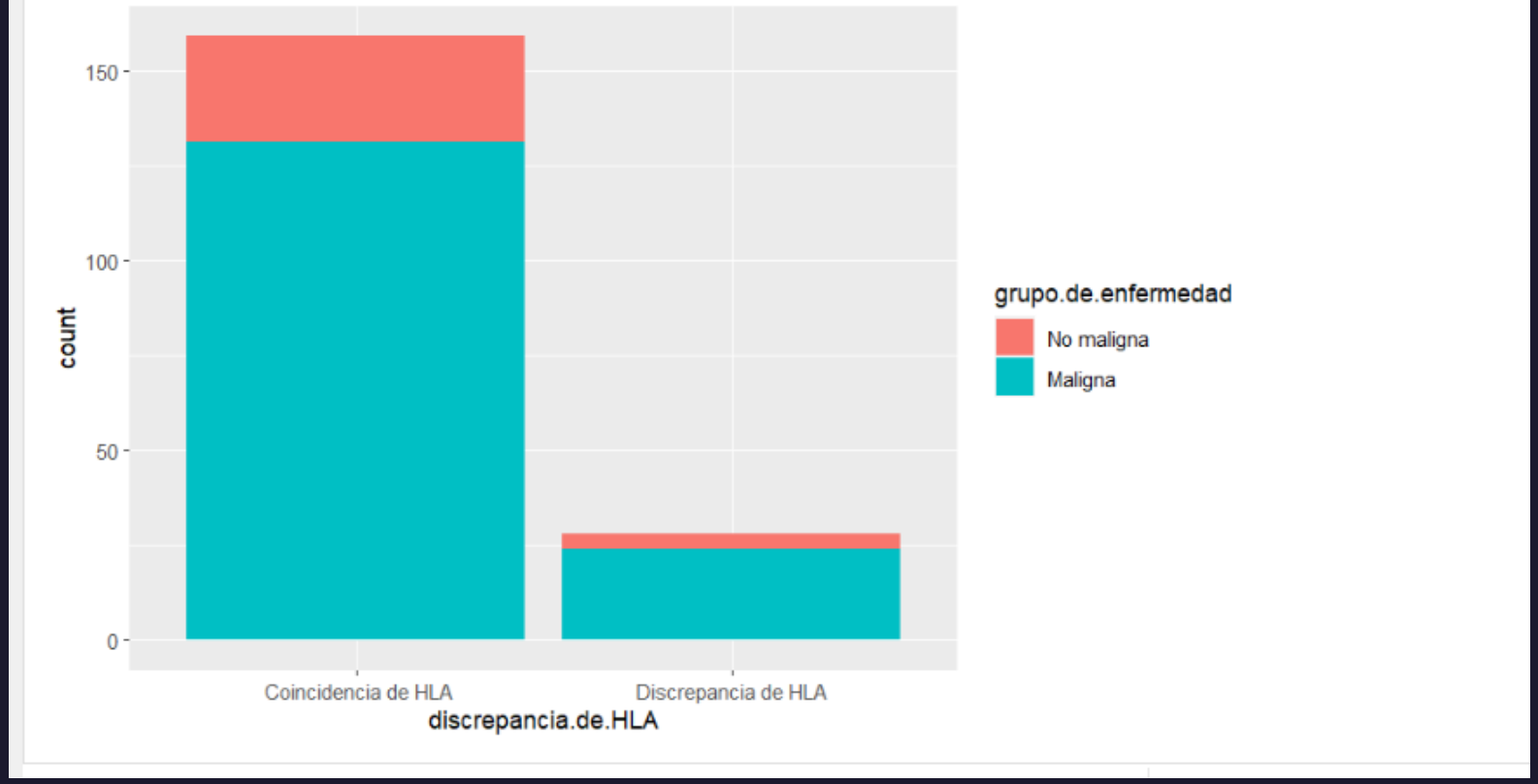


Gráfico de la compatibilidad del donante y receptor según su género con respecto el grupo de riesgo.

```{r}
graf10<-datos%>%
 ggplot(aes(x=compatibilidad.de.genero, fill=grupo.de.riesgo))+
 geom_bar()
graf10
```

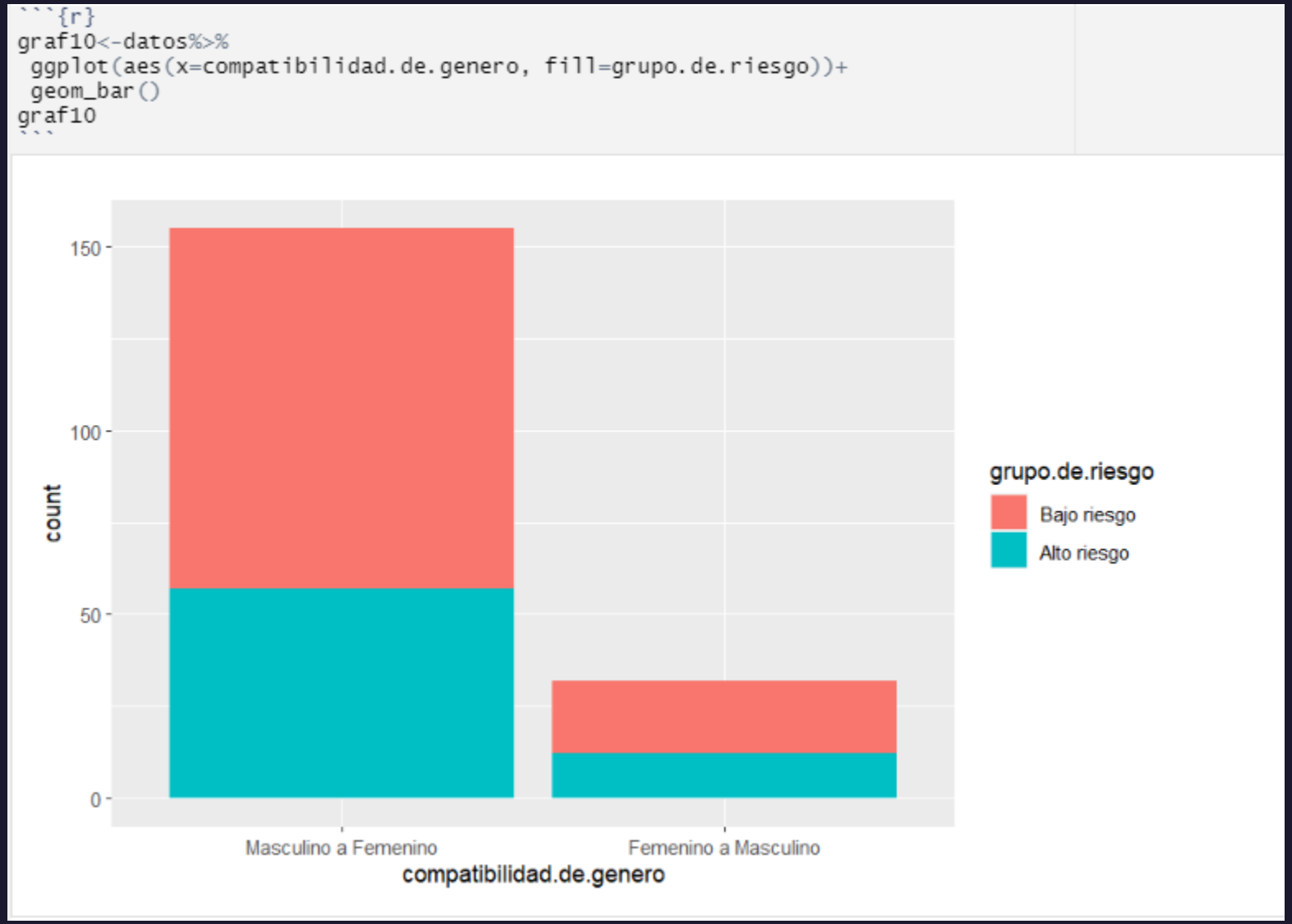
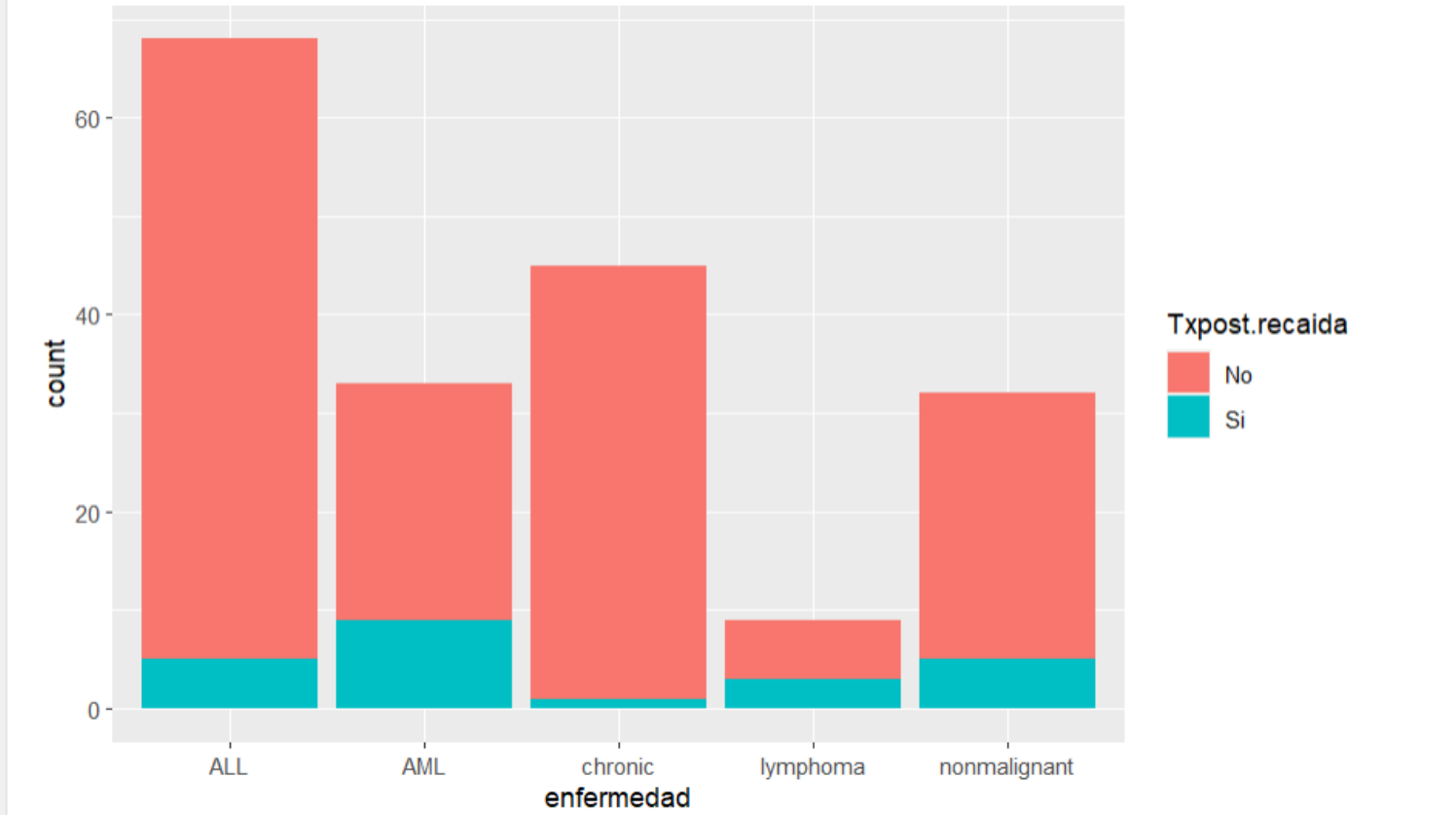


Gráfico del tipo de enfermedad con respecto el segundo trasplante de médula ósea después de la recaída. Como podemos observar que el segundo trasplante de médula ósea después de la recaída afecta mucho a los pacientes que tienen la enfermedad AML. De manera contraria pasa con aquellos pacientes que tengan enfermedad crónica.

```{r}
graf6<-datos%>%
 ggplot(aes(x=enfermedad, fill=Txpost.recaida))+
 geom_bar()
graf6
```



Training and test. Dividimos de forma aleatoriamente nuestro conjunto de datos en un conjunto de entrenamiento (para crear un modelo de predicción), y otro conjunto de test (para ver como de bueno es ese modelo).

```r
mini_datos<-datos[c("sexo.del.receptor", "recaida", "estado.de.supervivencia", "enfermedad","Txpost.recaida","grupo.de.riesgo")]
```

```r
str(mini_datos)
```

```
'data.frame':   187 obs. of  6 variables:
 $ sexo.del.receptor      : Factor w/ 2 levels "Femenino","Masculino": 2 2 2 1 1 2 1 2 2 2 ...
 $ recaida                : Factor w/ 2 levels "No","Si": 1 2 2 1 1 1 1 1 1 1 ...
 $ estado.de.supervivencia: Factor w/ 2 levels "Vivo","Muerto": 1 2 2 2 1 1 2 2 1 1 ...
 $ enfermedad             : chr  "ALL" "ALL" "ALL" "AML" ...
 $ Txpost.recaida         : Factor w/ 2 levels "No","Si": 1 1 1 1 1 2 1 1 1 1 ...
 $ grupo.de.riesgo        : Factor w/ 2 levels "Bajo riesgo",..: 2 1 1 1 2 2 1 1 1 1 ...
```

```r
dimension<-dim(mini_datos[1])
dimension
```

```
[1] 187   1
```

```r
training<- 0.80*dimension
training
```

```
[1] 149.6   0.8
```

```r
ec1 = sort(sample(1:dimension, size=training, replace=FALSE))
```

```r
train<-mini_datos[ec1,]
dim(train)
```

```
[1] 149   6
```

```r
test<-mini_datos[ec1,]
dim(test)
```

```
[1] 149   6
```

Hacemos una regresión logística

```r
Regresion_logistica <- glm(recaida ~ ., family = binomial, data =mini_datos)
summary(Regresion_logistica)
```

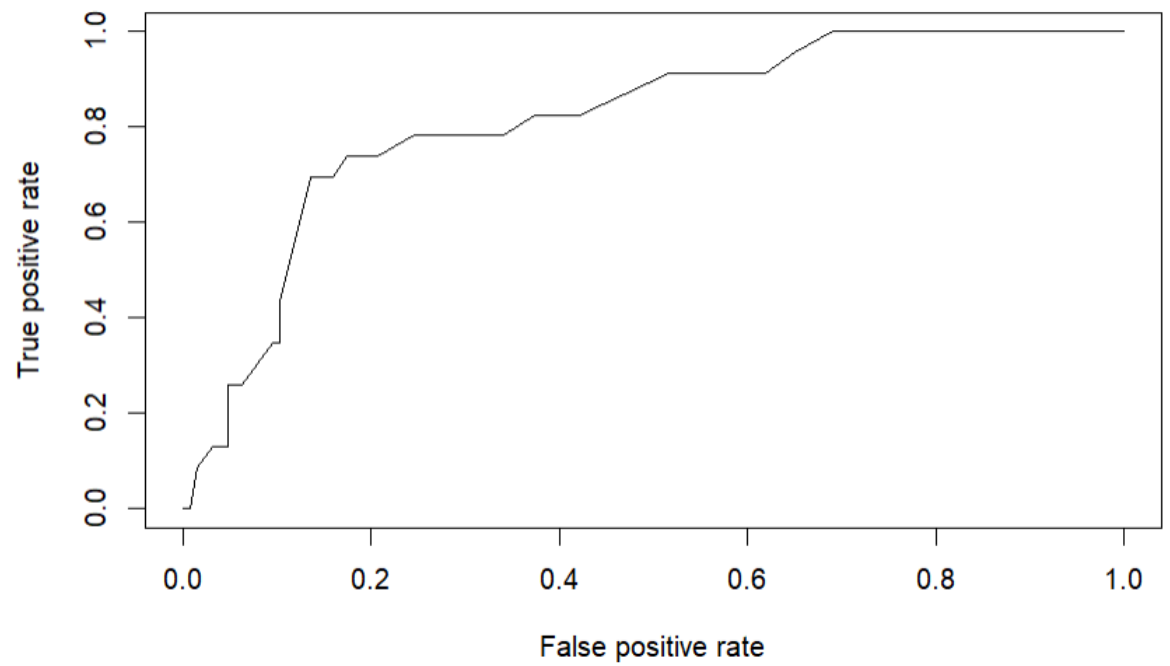
```
Call:
glm(formula = recaida ~ ., family = binomial, data = mini_datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2223  -0.5472  -0.3723  -0.2147   2.5648

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.21352    0.65410  -4.913 8.97e-07 ***
sexo.del.receptorMasculino  0.61677    0.48810   1.264 0.206362
estado.de.supervivenciaMuerto  1.80639    0.54261   3.329 0.000871 ***
enfermedadAML              -0.41629    0.65651  -0.634 0.526021
enfermedadchronic          -0.54522    0.63364  -0.860 0.389536
enfermedadlymphoma          0.12708    0.81011   0.157 0.875354
enfermedadnonmalignant     -1.84410    1.09206  -1.689 0.091289 .
Txpost.recaidaSi            0.06182    0.71348   0.087 0.930958
grupo.de.riesgoAlto riesgo  0.70645    0.55970   1.262 0.206881
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157.92  on 186  degrees of freedom
Residual deviance: 129.12  on 178  degrees of freedom
AIC: 147.12

Number of Fisher Scoring iterations: 6
```

```r
prediccion1<- predict.glm(Regresion_logistica,newdata = test, type="response")
tabla<- table(test$recaida, floor(prediccion1+0.5))
tabla
```

```
     0  1
No 124  2
Si  21  2
```

Curva ROC.
Como podemos ver el modelo que nos sale no es que sea muy bueno pero esta bien.

```{r}
prediccion1 <- ROCR::prediction(prediccion1,test$recaida)
falsos_posi <- performance(prediccion1, "tpr", "fpr")
plot(falsos_posi)
```




```{r}
AUc=performance(prediccion1, measure = "auc")@y.values[[1]]
cat("AUC: ",AUc,"n")
```

```
AUC:  0.8146998 n
```

RANDOM FOREST

```r
randomf <- randomForest(recaida~.,data = test,ntree = 500,mtry = 4)
randomf
```

```
Call:
 randomForest(formula = recaida ~ ., data = test, ntree = 500,      mtry = 4)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 15.44%
Confusion matrix:
    No Si class.error
No 116 10  0.07936508
Si  13 10  0.56521739
```

```r
pred1_rf <- predict(randomf, newdata = test[,-12])
matriz <- confusionMatrix(test$recaida,pred1_rf)
matriz
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No  Si
        No 121   5
        Si  10  13

               Accuracy : 0.8993
                 95% CI : (0.8394, 0.9426)
    No Information Rate : 0.8792
    P-Value [Acc > NIR] : 0.2714

                  Kappa : 0.5768

 Mcnemar's Test P-Value : 0.3017

            Sensitivity : 0.9237
            Specificity : 0.7222
         Pos Pred Value : 0.9603
         Neg Pred Value : 0.5652
             Prevalence : 0.8792
         Detection Rate : 0.8121
   Detection Prevalence : 0.8456
      Balanced Accuracy : 0.8229

       'Positive' Class : No
```

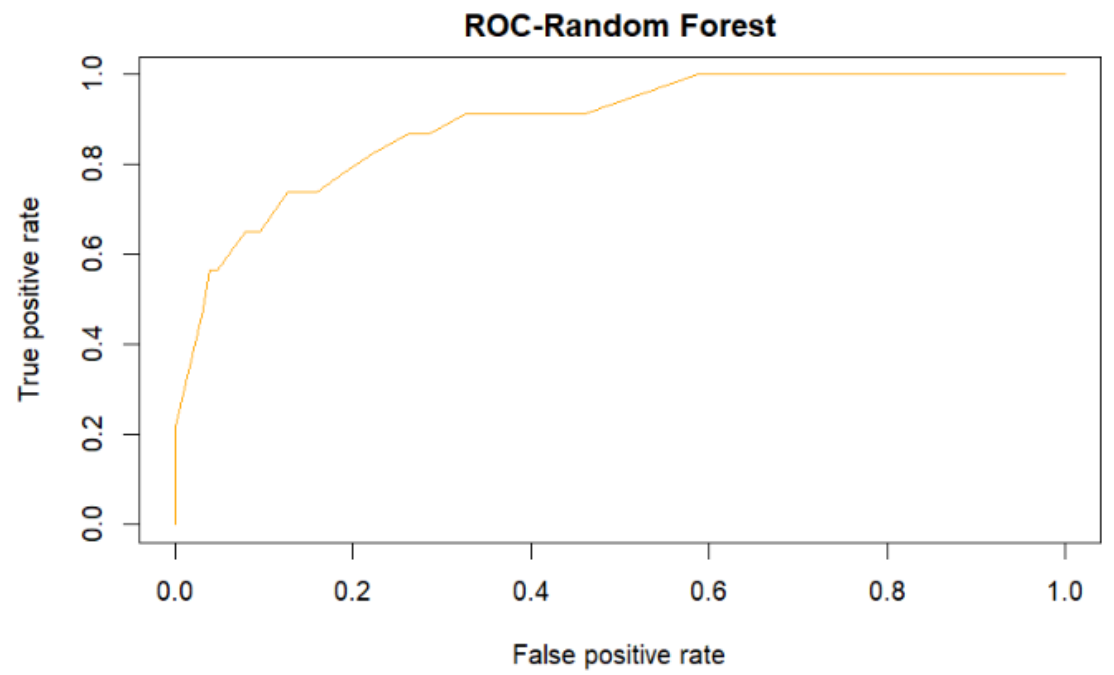
```r
prediccion_rf <-predict(randomf,newdata = test[,-12],type = 'prob')
prediccion_rf <- prediccion_rf[,2]
pred_rf<-prediction(prediccion_rf, test$recaida)
pred_rf
```

```
A prediction instance
  with 149 data points
```

Curva ROC

Aunque ha dado bastante similares, este (Random forest) es mejor puesto que esta con un área bajo la curva de 0.8916... y el otro (Regresión logística) de 0.8146998 aunque de todas maneras este no es que sea totalmente muy bueno.

```r
falsos_randomf <- performance(pred_rf, "tpr", "fpr")
plot(falsos_randomf, main="ROC-Random Forest", col = "orange")
```

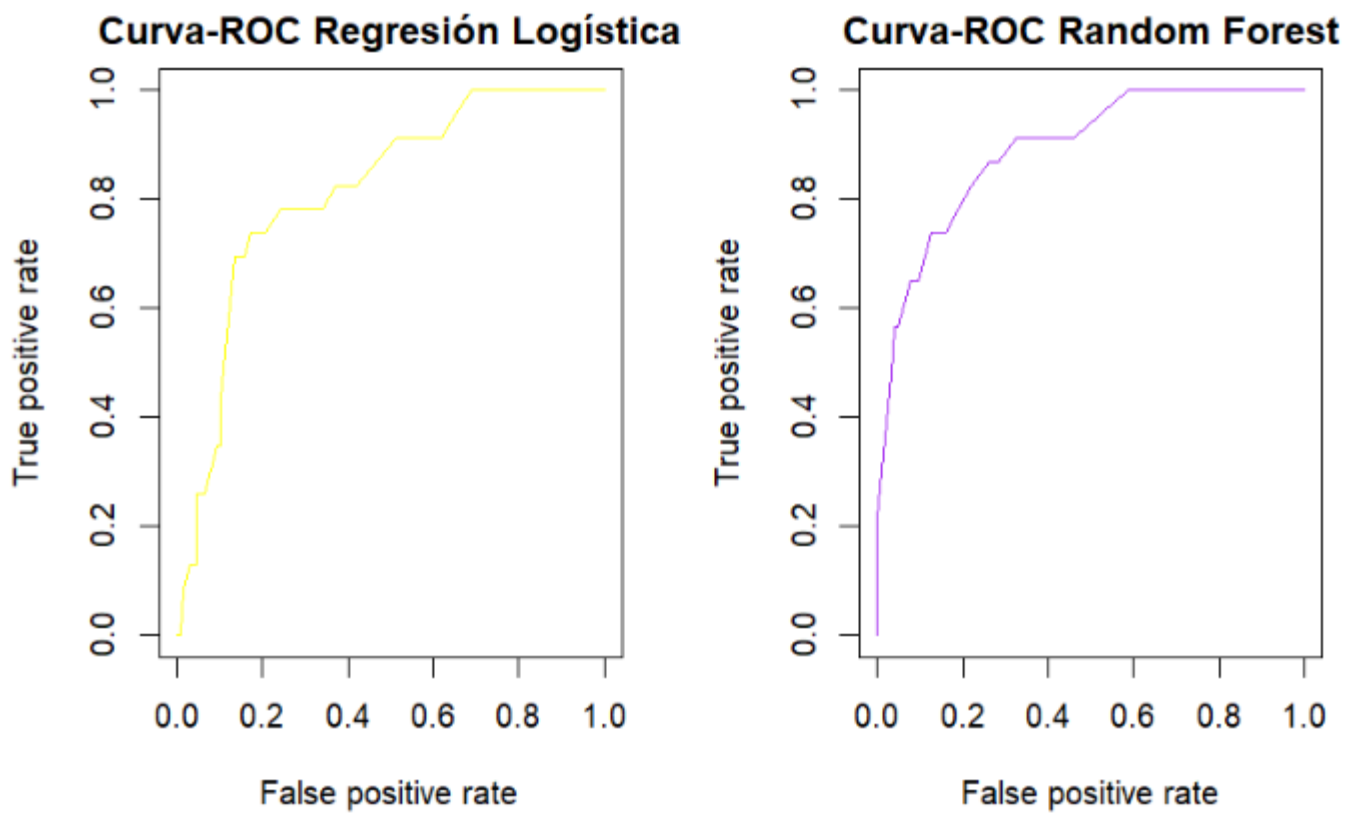



```r
AUC_randomf=performance(pred_rf, measure = "auc")@y.values[[1]]
cat("AUC: ",AUC_randomf,"n")
```

```
AUC:  0.8916494 n
```

Comparación de las curvas ROC

```r
par(mfrow=c(1,2))
plot(falsos_posi, main="Curva-ROC Regresión Logística",col = "yellow")
plot(falsos_randomf, main="Curva-ROC Random Forest",col = "purple")
```



3. Reconomientos:

Fuente: Marek Sikora (marek.sikora '@' polsl.pl), Åukasz WrÃ³bel (lukasz.wrobel '@' polsl.pl), Adam GudyÅ› (adam.gudys '@' polsl.pl)

Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland

URL CONJUNTO DE DATOS UCI: https://archive.ics.uci.edu/ml/datasets/Bone+marrow+transplant%3A+children

¡Gracias!

Paula Poley Ceballos

Análisis avanzado de datos clínicos

3° Ingeniería de la Salud