EVW-AI

GRHECO-Xen

2026-01-15

Abstract

El estudio aborda el problema desde dos ángulos distintos:

1. **Análisis fenotípico**. Se clusterizó a los pacientes basándose en variables clínicas y de laboratorio. Se identificaron 7 clusters, pero la agrupación resultó ser estadísticamente débil (p=0.02), lo que confirma la gran dificultad de clasificar la EVW solo por sus manifestaciones fenotípicas.
2. **Análisis genético**. Se clusterizó a los pacientes basándose en datos genéticos (mutaciones, frecuencia, puntuaciones de patogenicidad). Se identificó un total de 9 clústeres claramente distintos y estadísticamente robustos ($p < 0.001$). La conclusión más relevante es que **estos 9 clusters genéticos muestran una fuerte asociación con la severidad clínica de la enfermedad**, medida a través de la puntuación hemorrágica. Por ejemplo, el *Cluster 8* agrupa a pacientes con los tipos de VWD más severos (como 2N y 2B) y las puntuaciones hemorrágicas más altas.

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

# Load required libraries
library(parallel)
library(FactoMineR)
library(factoextra)
library(missMDA)
library(mclust)
library(tidyverse)
library(ggplot2)
library(ggalluvial)
library(ggridges)
library(reshape2)
library(dunn.test)
library(plotly)
library(flextable)
library(pagedown)

# Parallel setup and seed
options(rf.cores = detectCores() - 1)
set.seed(1923)

# Load custom preprocessing functions
source("./scripts/preprocess_data.R")
```

1. Phenotypic Data Analysis

Las variables fenotípicas disponibles para el estudio son las siguientes:

De las 15 disponibles, se ha decidido prescindir de 2: **Centro de Familia**, al tratarse de un identificar de centro de reclutamiento y familia, y de **FVIII_B_LC1**, ya que se ha comprobado que todos los valores de esta variable son 0.

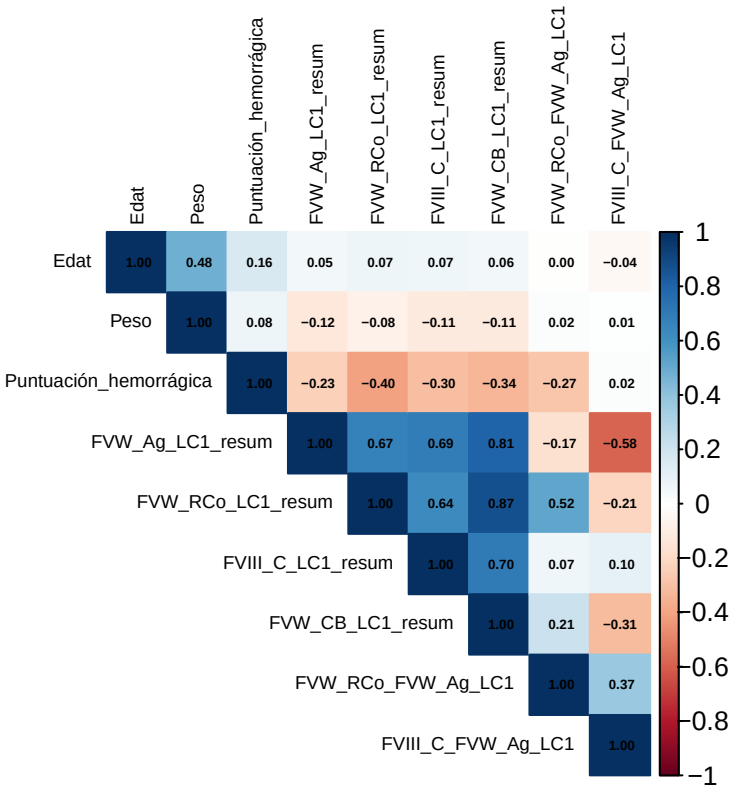
QC Phenotypic Analysis El primer paso en nuestro análisis es realizar un QC de los datos fenotípicos. Con ello podremos comprobar si cada variable está codificada correctamente (numérica, factorial), el rango o niveles de los valores que la componen y el % de datos faltantes.

```
##                                Columna Porcentaje_NA
## 4                                 Peso         39.90
## 5                               Etnia          2.40
## 6                       Grupo_sanguíneo          6.34
## 7                 Puntuación_hemorrágica          0.68
## 8                       FVW_Ag_LC1_resum          5.65
## 9                      FVW_RCo_LC1_resum         12.33
## 10                    FVIII_C_LC1_resum          0.51
## 11                     FVW_CB_LC1_resum          7.53
## 12                  FVW_FVIII_LC1_resum         93.66
## 13 Análisis_multimérico_baja_LC1_resum          4.28
## 14 Análisis_multimérico_alta_LC1_resum         20.72
## 15                   FVW_RCo_FVW_Ag_LC1         13.87
## 16                   FVIII_C_FVW_Ag_LC1          5.99
```

Se observa que **la variable FVW_FVIII_LC1_resum presenta el 94,24% de los datos faltantes**, por lo que se procede a su eliminación ya que la imputación no es la estrategia más adecuada ante tal elevado número de datos faltantes.

Correlation analysis of Phenotypic Data Antes de imputar los datos faltantes y explorar las relaciones entre los individuos y las variables, vamos a proceder a analizar si existe correlación entre las variables.

Se analizó la correlación entre las diferentes variables fenotípicas, mediante la correlación de Spearman, observándose que no existe ninguna variable altamente relacionada (>0.98)

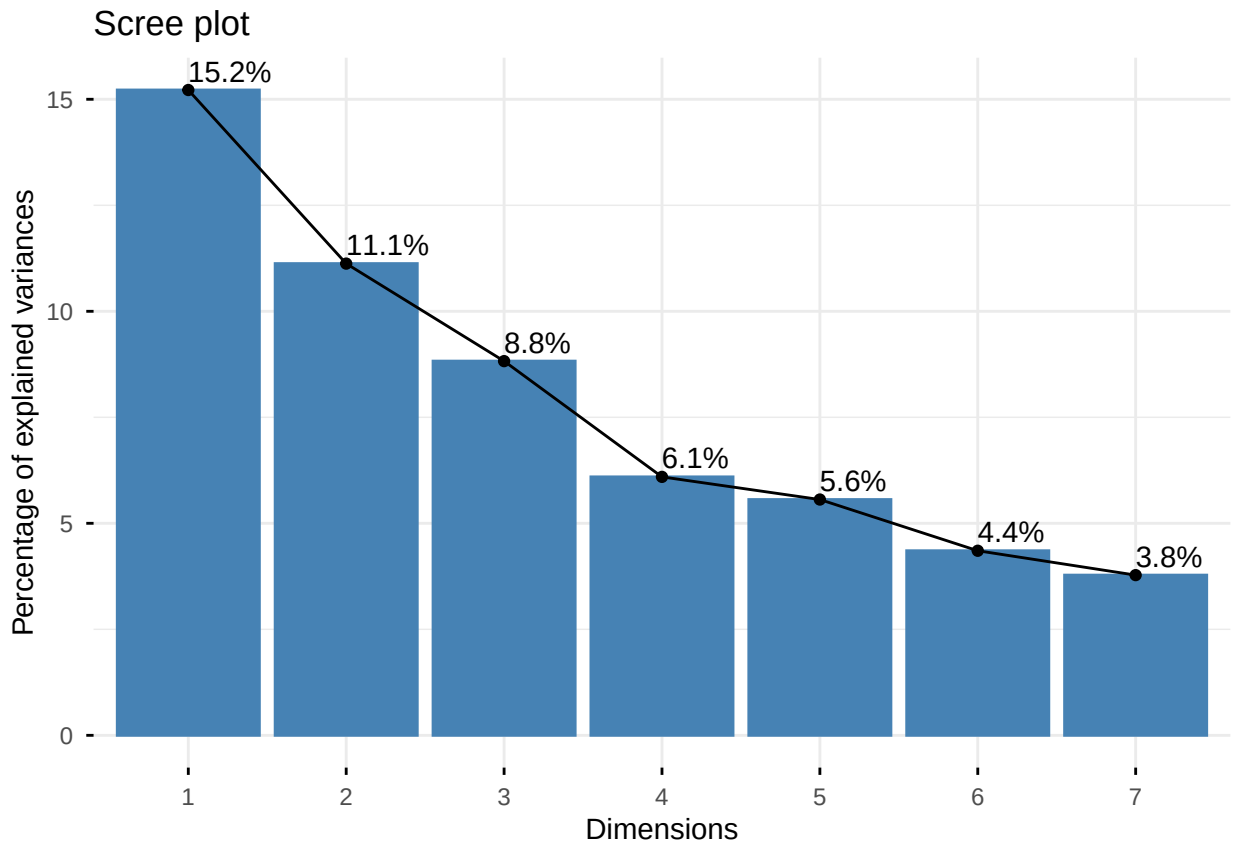


Se puede observar que no hay ninguna variable altamente correlacionada (corr > 0.98), por lo que se emplearán las 15 variables restantes para el análisis factorial de datos mixtos (FAMD)

FAMD + Clusterización Para el análisis de los datos fenotípicos hemos decidido usar FAMD porque es una técnica estadística que permite reducir la dimensionalidad de un conjunto de datos que contiene tanto variables cuantitativas (numéricas) como cualitativas (categóricas); como es el caso de la base de datos fenotípica.

Como comentamos previamente, para hacer el FAMD, primero nos piden imputar los datos faltantes.

Una vez imputado, podemos proceder a hacer el FAMD. Para elegir el número óptimo de NCP, haremos una primera aproximación y, en base al gráfico obtenido con la primera aproximación, elegiremos el valor de NCP donde se aplane la curva. En este caso, NCP = 7

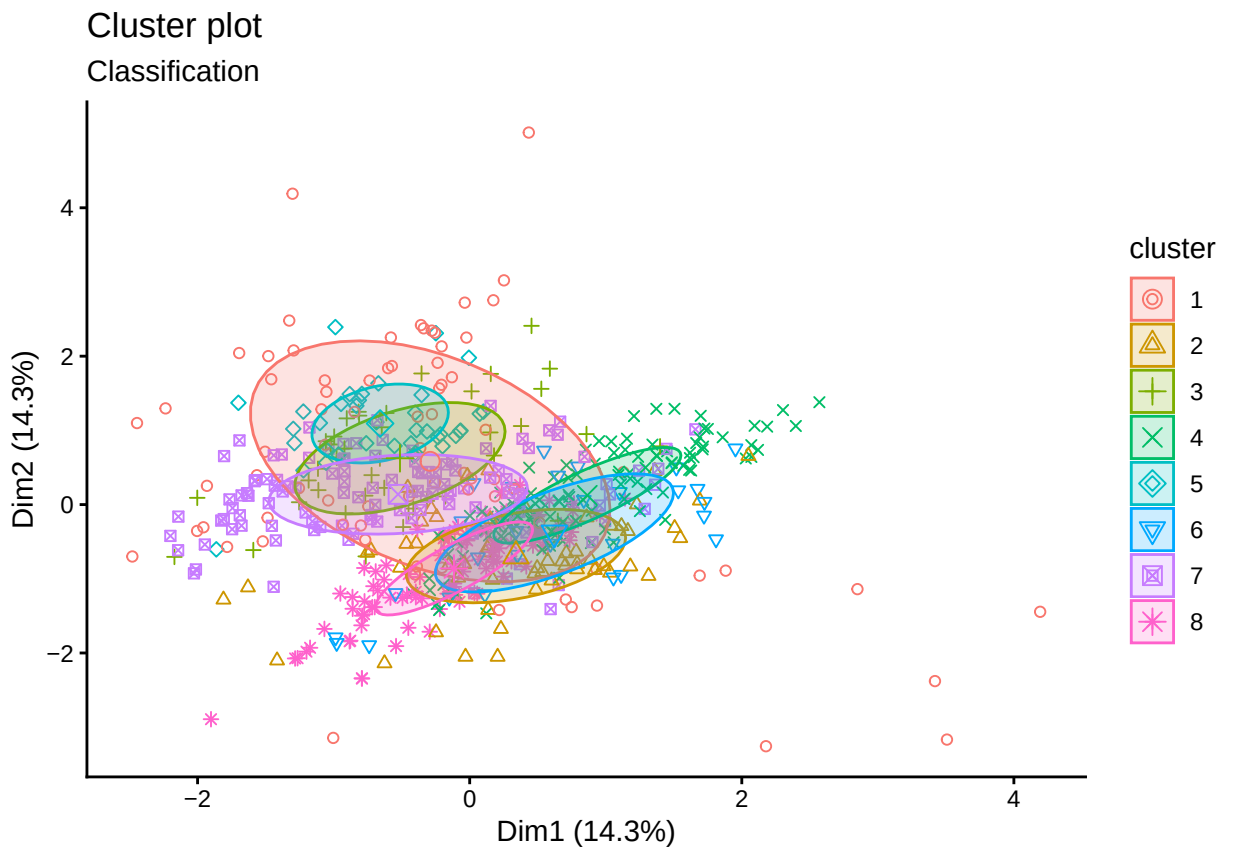


Una vez aplicado el FAMD, hemos reducido y homogeneizado los datos fenotípicos, con lo que ahora podemos clusterizar nuestras muestras.

```
##
## Número óptimo de clusters (fenotípicos): 8
```

```
##
##   1   2   3   4   5   6   7   8
##  76  58  40 121  37  28 119 105
```

Los individuos han sido clusterizados en 7 grupos diferentes, cuya distribución podemos observar en el siguiente gráfico:



Se observa que los 7 clústeres tienen un enorme solapamiento entre ellos, donde los pacientes están *mezclados* en el mismo espacio y apenas se observa una clara diferencia entre ellos. Esto muestra que los datos fenotípicos por sí solos no son un buen método para la diferenciación de pacientes. El fenotipo por sí solo es un mapa "borroso" de la enfermedad.

2. Genotypic Data Analysis

Hemos visto que con los datos fenotípicos no somos capaces de extraer realmente información relevante que nos permita diferenciar grupos de pacientes. Vamos a ver si con los datos genéticos podemos obtener un mapa mucho más nítido e identificar grupos de pacientes más claros.

Las variables genéticas que tenemos disponibles son las siguientes:

```
##   [1] "N_var_total"       "N_var_1"          "N_var_5"
##   [4] "N_alt"             "N_alt_1"          "N_alt_5"
##   [7] "N_exons"           "N_dominis"        "N_aa"
##  [10] "N_C2"              "N_A1"             "N_A2"
##  [13] "N_A3"              "N_D3"             "N_D4"
##  [16] "N_D"               "N_D2"             "N_B"
##  [19] "N_D1"              "N_C1"             "N_CK"
##  [22] "N_SP"              "N_var_impact"     "N_var_frameshift"
##  [25] "N_var_missense"    "N_var_Splice"     "N_var_inframe_indel"
##  [28] "N_var_stop"        "N_var_ACMG"       "N_var_ClinVar"
##  [31] "N_var_SIFT"        "N_var_PolyPhen"   "N_var_AlphaM"
##  [34] "N_var_EVE"         "N_var_SpliceAI"   "N_var_Truncant"
##  [37] "Sum_ClinVar"       "Sum_SIFT"         "Sum_PolyPhen"
```

```
##  [40] "Sum_AlphaM"          "Sum_EVE"             "Sum_Revel"
##  [43] "Sum_ESAD_ESAR"       "Sum_Splice"          "Sum_Truncant"
##  [46] "Sum_Impacte"         "N_aa_TOP_ClinVar"    "Score_TOP_ClinVar"
##  [49] "Freq_TOP_ClinVar"    "N_aa_TOP_SIFT"       "Score_TOP_SIFT"
##  [52] "Freq_TOP_SIFT"       "N_aa_TOP_PolyPhen"   "Score_TOP_PolyPhen"
##  [55] "Freq_TOP_PolyPhen"   "N_aa_TOP_AlphaM"     "Score_TOP_AlphaM"
##  [58] "Freq_TOP_AlphaM"     "N_aa_TOP_EVE"        "Score_TOP_EVE"
##  [61] "Freq_TOP_EVE"        "N_aa_TOP_Revel"      "Score_TOP_Revel"
##  [64] "Freq_TOP_Revel"      "N_aa_TOP_ESAD_ESAR"  "Score_TOP_ESAD_ESAR"
##  [67] "Freq_TOP_ESAD_ESAR"  "N_aa_TOP_Splice"     "Score_TOP_Splice"
##  [70] "Freq_TOP_Splice"     "N_aa_TOP_Truncant"   "Score_TOP_Truncant"
##  [73] "Freq_TOP_Truncant"   "N_aa_TOP_Impacte"    "Score_TOP_Impacte"
##  [76] "Freq_TOP_Impacte"    "Inv_Freq"            "N_var_total_alt"
##  [79] "N_var_5_alt"         "N_var_1_alt"         "N_exons_freq"
##  [82] "N_dominis_freq"      "N_aa_freq"           "N_exons_ClinVar"
##  [85] "N_dominis_ClinVar"   "N_aa_ClinVar"        "N_exons_SIFT"
##  [88] "N_dominis_SIFT"      "N_aa_SIFT"           "N_exons_PolyPhen"
##  [91] "N_dominis_PolyPhen"  "N_aa_PolyPhen"       "N_exons_AlphaM"
##  [94] "N_dominis_AlphaM"    "N_aa_AlphaM"         "N_exons_EVE"
##  [97] "N_dominis_EVE"       "N_aa_EVE"            "N_exons_Revel"
## [100] "N_dominis_Revel"     "N_aa_Revel"          "N_exons_Splice"
## [103] "N_dominis_Splice"    "N_aa_Splice"         "N_exons_ESAD_ESAR"
## [106] "N_dominis_ESAD_ESAR" "N_aa_ESAD_ESAR"      "N_exons_Truncant"
## [109] "N_dominis_Truncant"  "N_aa_Truncant"       "N_exons_Impact"
## [112] "N_dominis_Impact"    "N_aa_Impact"
```

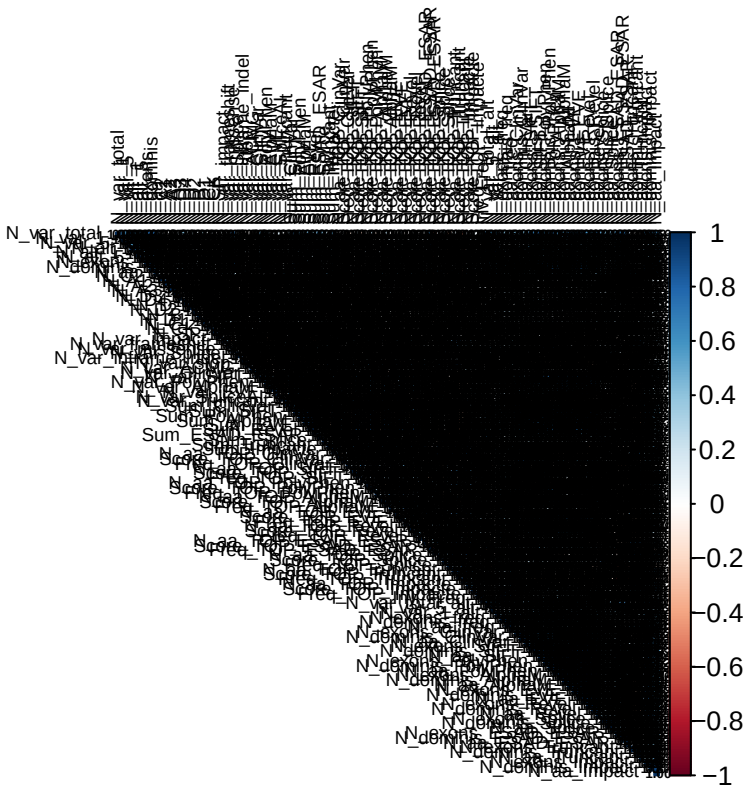
QC Genotypic Analysis Al igual que con los datos fenotípicos, procedemos a realizar el QC de los datos genéticos para ver si está correctamente codificada cada variable, el rango de valores y el % de datos faltantes.

En el QC se observa que tenemos datos completos y no es necesaria la imputación.

```
## [1] Columna       Porcentaje_NA
## <0 rows> (or 0-length row.names)
```

En este caso, se observa que no hay datos faltantes, de manera que nos ahorraremos tener que imputar los datos. Vamos a analizar posibles correlaciones entre variables, para tratar de reducir el número de variables empleadas.

Correlation analysis of Genotypic Data El análisis de correlación, en este caso, sí que muestra una gran cantidad de variables con una alta correlación (> 0.98) que se recogen en la siguiente tabla. Procedemos a la eliminación de una de las 2 variables correlacionadas.




```
## # A tibble: 45 x 3
##    Variable1          Variable2         Correlacion
##    <fct>              <fct>                   <dbl>
##  1 N_aa_TOP_PolyPhen  N_aa_TOP_SIFT           1
##  2 N_aa_TOP_AlphaM    N_aa_TOP_SIFT           1
##  3 N_aa_TOP_Revel     N_aa_TOP_SIFT           1
##  4 N_aa_TOP_AlphaM    N_aa_TOP_PolyPhen       1
##  5 N_aa_TOP_Revel     N_aa_TOP_PolyPhen       1
##  6 N_aa_TOP_Revel     N_aa_TOP_AlphaM         1
##  7 N_aa_TOP_Impacte   N_aa_TOP_Truncant       1.000
##  8 N_var_5_alt        N_alt_5                 1.000
##  9 N_dominis_Truncant N_exons_Truncant        1.000
## 10 N_dominis_Splice   N_exons_Splice          1.000
## # i 35 more rows
```

Aquí también es relevante eliminar cualquier columna donde la varianza sea 0 porque, aunque esto no suceda sobre la base de datos entera, si hacemos algún filtrado de los datos sí que puede ocurrir (típico del perfil genético de subtipos de la enfermedad). Una columna de varianza 0 se carga cualquier resolución en el downstream.

```
geno.mclust <- Mclust(geno_pc_top, G = 1:50)
geno_cl <- genoscp |>
  mutate(clusteres = geno.mclust$classification)
```

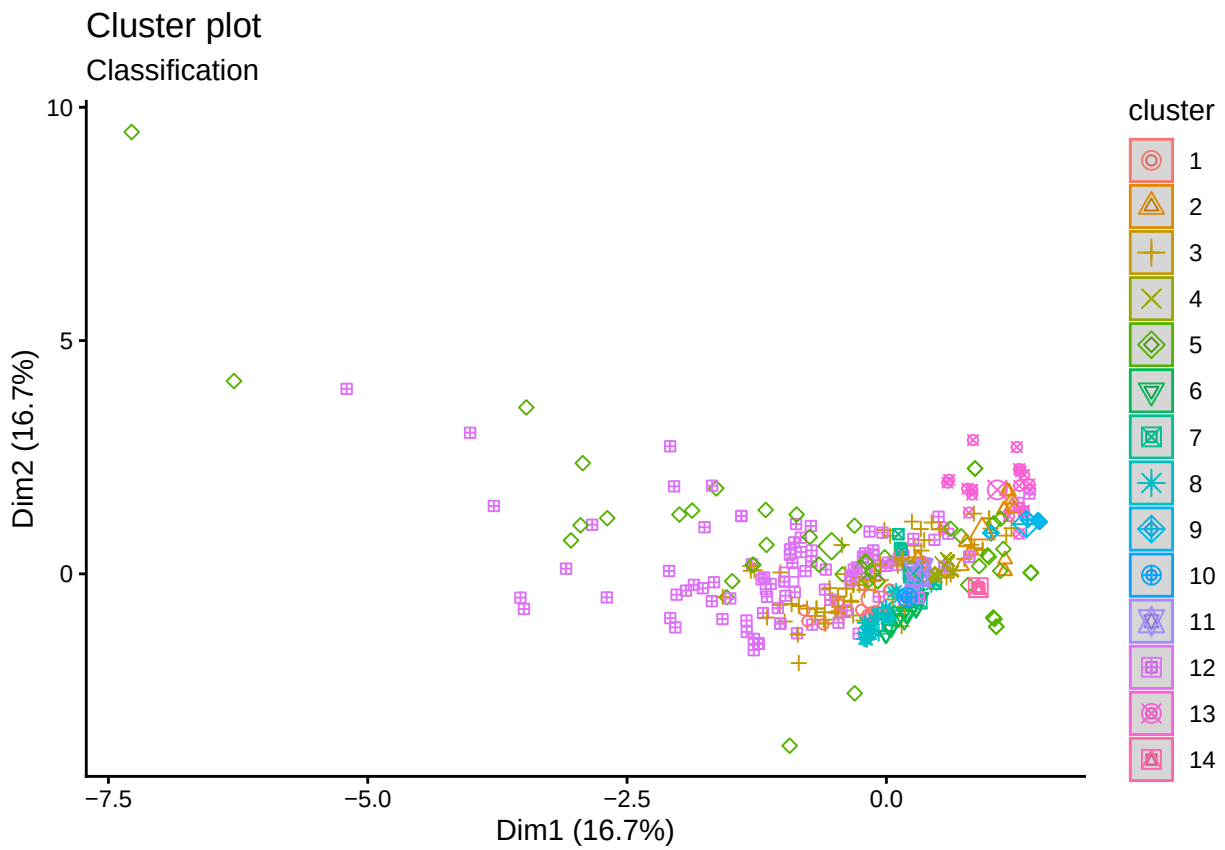
```r
cat("\nNúmero óptimo de clusters (genéticos):", geno.mclust$G, "\n")
```

```
##
## Número óptimo de clusters (genéticos): 14
```

```r
print(table(geno.mclust$classification))
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14
##  23  19  99  46  53  47  24  38  29  24  35 111  22  14
```

Cuya distribución en clústeres es la siguiente:



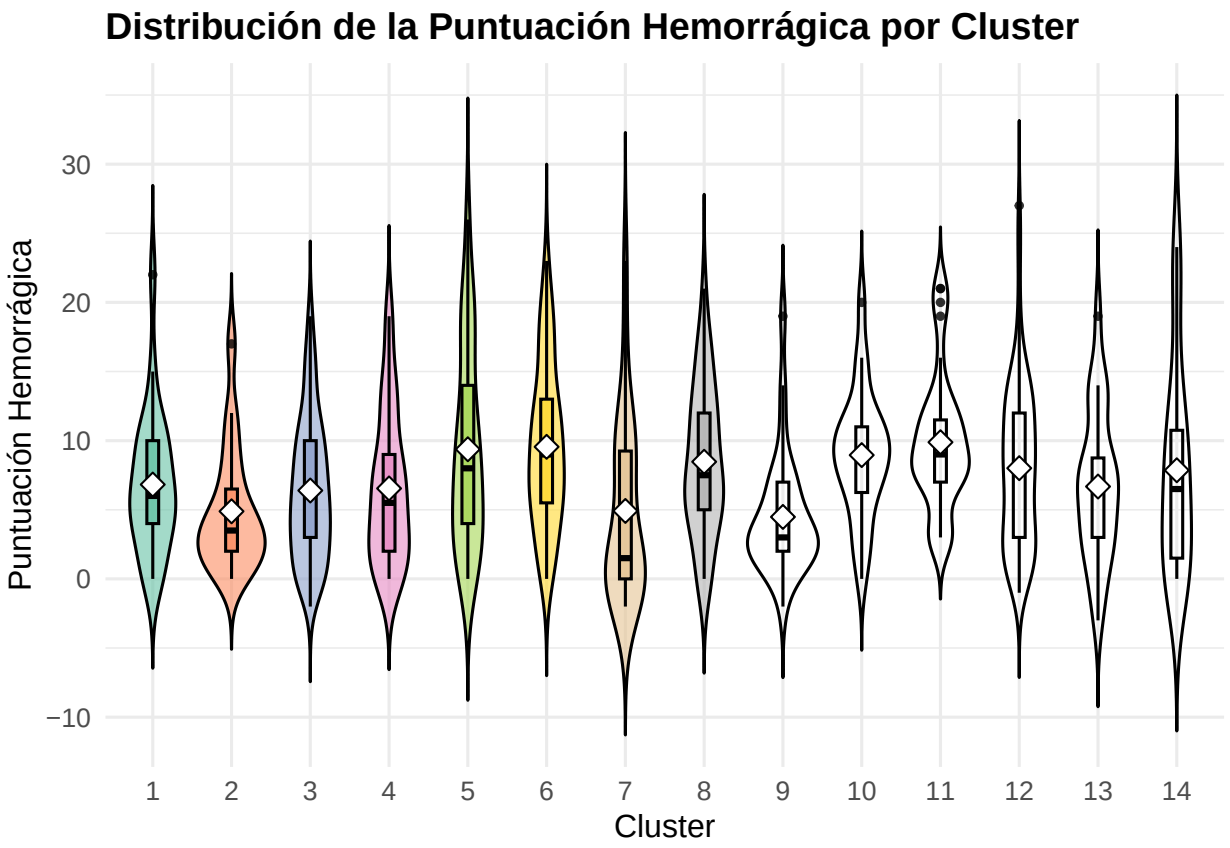
Tras haber realizado la clusterización no supervisada, asociamos el resultado de puntuación hemorrágica a los clústeres de geno para observar la eficacia de nuestra clusterización.

El Kruskal test sí que muestra que hay una distribución diferente en la puntuación hemorrágica:

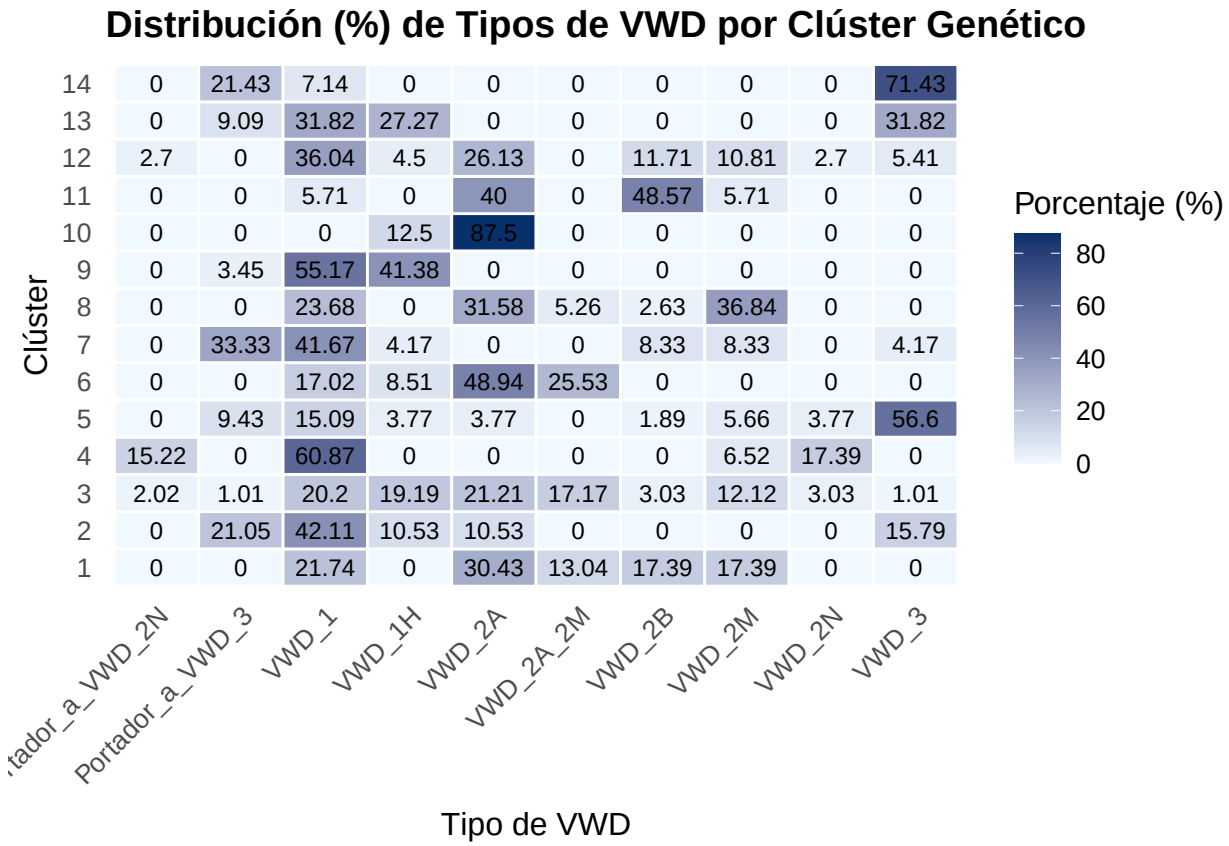
```r
kruskal.test(Puntuacion_Hemorragica ~ clusteres, data = geno_cla)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Puntuacion_Hemorragica by clusteres
## Kruskal-Wallis chi-squared = 46.574, df = 13, p-value = 1.14e-05
```

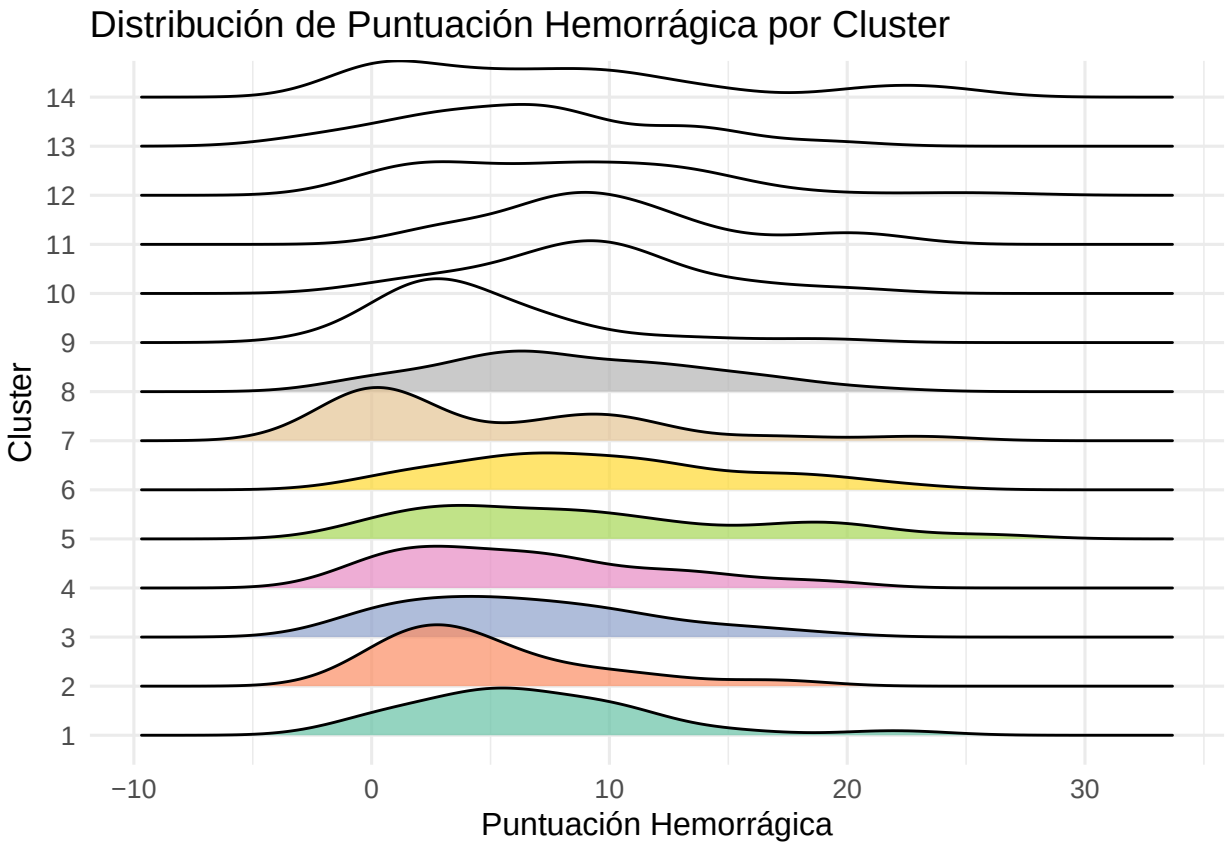
1. Variabilidad de la puntuación hemorrágica entre clústeres Con 7 clústeres no se observa ningún clúster que tenga una puntuación hemorrágica diferenciada.



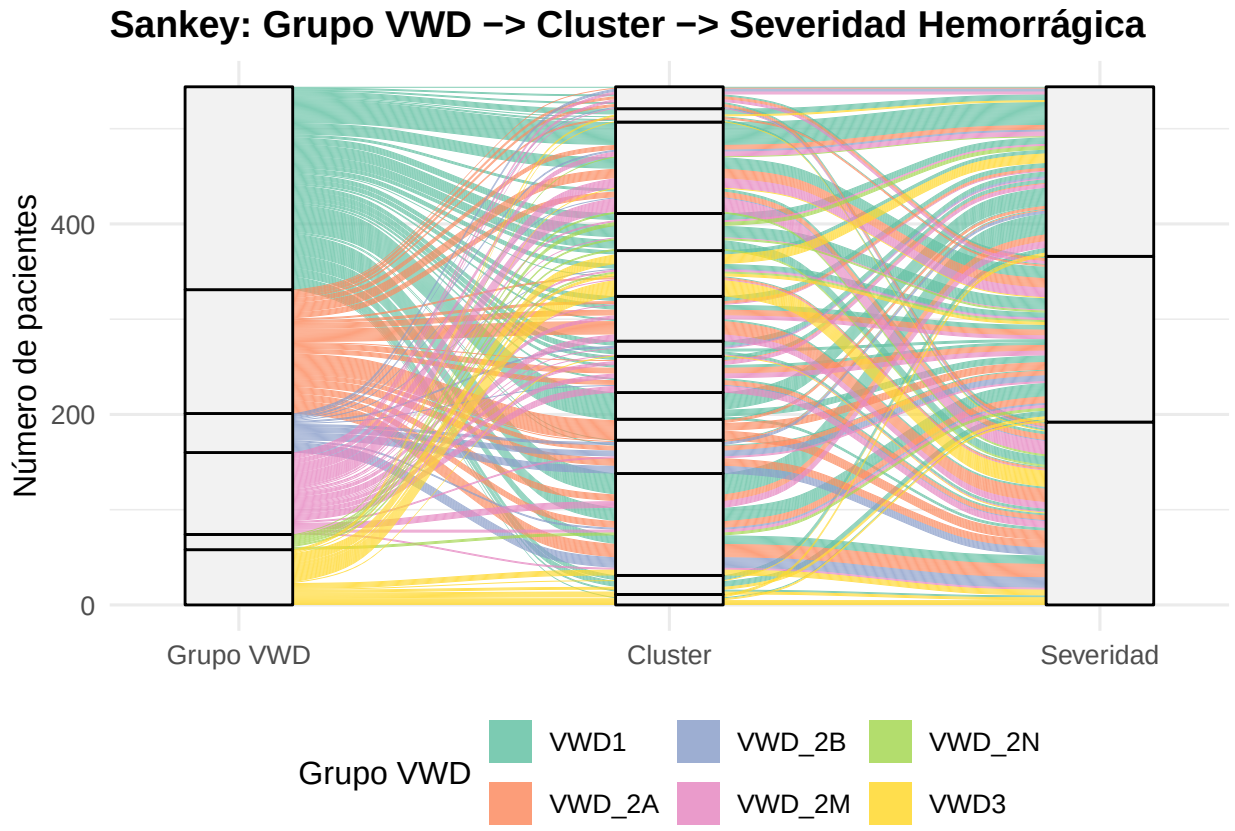
2. Distribución de Tipos de VWD en cada clúster Clúster 1 y Clúster 2 muuuy heterogéneos.

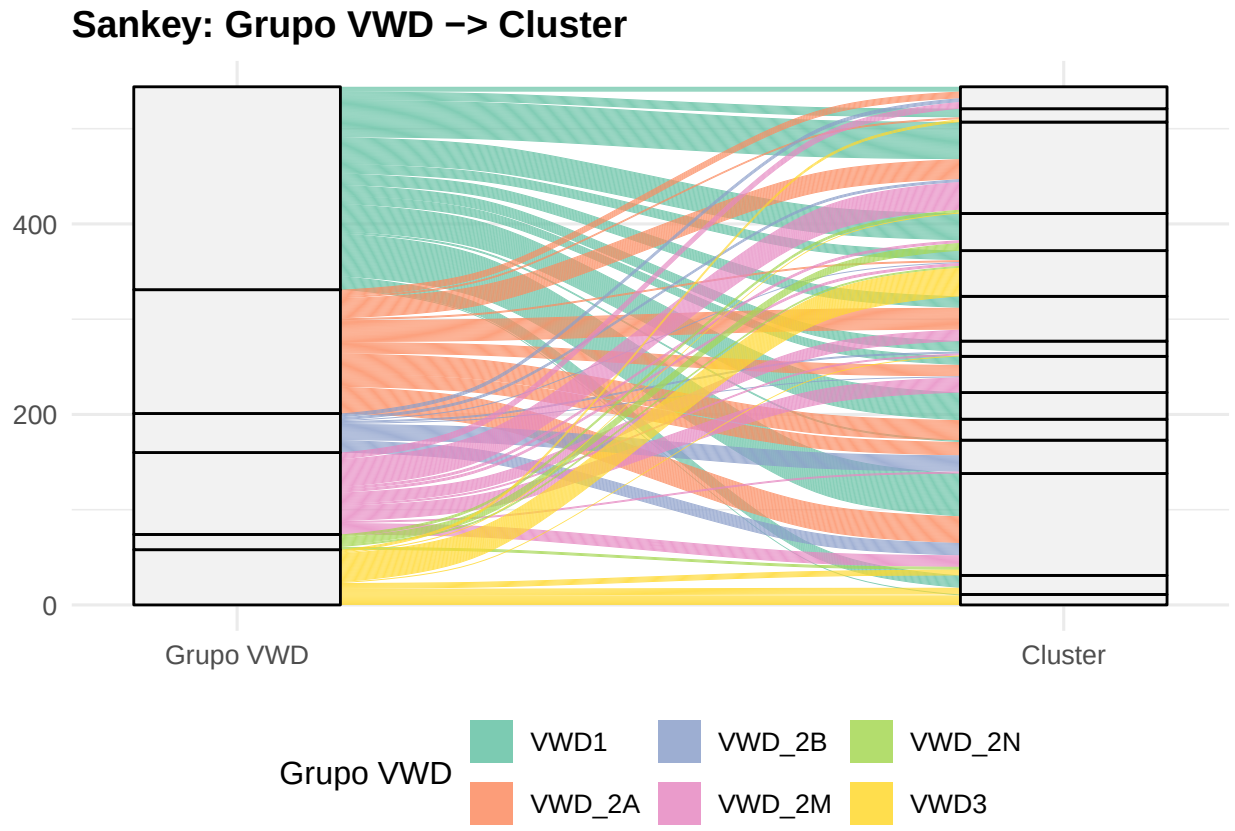


3. Densidad de la Puntuación Hemorrágica Vemos la distribución de la puntuación hemorrágica en los clústeres



4. Integración Genotipo–Clúster–Severidad De nuevo, muy heterogéneo todo.


Sankey: Grupo VWD −> Cluster −> Severidad Hemorrágica
Número de pacientes
400
200
0
Grupo VWD
Cluster
Severidad
Grupo VWD
VWD1
VWD_2A
VWD_2B
VWD_2M
VWD_2N
VWD3



Bonus Track: R2 Se ha estudiado también si existe una asociación lineal del score fenotípico + genético + el tipo de VWD respecto a la Puntuación Hemorrágica. El score fenotípico y score genético se define como el clúster asignado al paciente (en función de las variables fenotípicas y genotípicas respectivamente).

Se comprueba que no existe una relación estadística. De hecho, apenas añade información extra el score genético o fenotípico.

```
##                           R2 R2_Ajustado
## Fenotipico        0.15055989  0.14016464
## Genetico          0.07393685  0.05266685
## VWD               0.18257421  0.16966749
## Feno + Geno       0.17790549  0.14849244
## Feno + Geno + VWD 0.22203546  0.18101551
## Feno + VWD        0.19983051  0.17709034
## Geno + VWD        0.20367223  0.17221943
```

Tampoco si eliminamos a los portadores o a los individuos con un VWD sin clasificar:

```
# Lista de fórmulas que quieres evaluar
modelos <- list(
  "Fenotipico" = Puntuacion_Hemorragica ~ Score_Fenotipico,
  "Genetico"   = Puntuacion_Hemorragica ~ Score_Genetico,
  "VWD"        = Puntuacion_Hemorragica ~ Tipo_VWD,
  "Feno + Geno" = Puntuacion_Hemorragica ~ Score_Fenotipico + Score_Genetico,
  "Feno + Geno + VWD" = Puntuacion_Hemorragica ~ Score_Fenotipico + Score_Genetico + Tipo_VWD,
```

```
  "Feno + VWD" = Puntuacion_Hemorragica ~ Score_Fenotipico + Tipo_VWD,
  "Geno + VWD" = Puntuacion_Hemorragica ~ Score_Genetico + Tipo_VWD
)

# Listas para almacenar resultados
resultados_R2 <- list()

# Loop para imprimir summaries y guardar R2
for (nombre in names(modelos)) {

  #cat("\n----------------------------------------------------------------\n")
  #cat("### Summary del modelo:", nombre, "###\n")
  #cat("----------------------------------------------------------------\n\n")

  modelo <- lm(modelos[[nombre]], data = dataframe_glm_filtrado)

  # Print del summary
  #print(summary(modelo))

  # Guardar R2 en lista
  resultados_R2[[nombre]] <- data.frame(
    Modelo = nombre,
    R2 = summary(modelo)$r.squared,
    R2_Ajustado = summary(modelo)$adj.r.squared
  )
}

# Convertir lista a data.frame final
tabla_R2 <- do.call(rbind, resultados_R2)
tabla_R2[,2:3]
```

```
##                           R2 R2_Ajustado
## Fenotipico        0.15055989  0.14016464
## Genetico          0.07393685  0.05266685
## VWD               0.18257421  0.16966749
## Feno + Geno       0.17790549  0.14849244
## Feno + Geno + VWD 0.22203546  0.18101551
## Feno + VWD        0.19983051  0.17709034
## Geno + VWD        0.20367223  0.17221943
```

Fenogeno data analysis

A partir de aquí, gracias a una falsa alarma decidimos investigar la señal que aportaba la clusterización a través de Geno y Feno.

Esta parte trata con las bases tratadas por QC de Feno y la base completa de Geno (post-análisis) - Con la salvedad de que nos excluimos a los individuos con EVW sin clasificar.

En primer lugar, eliminamos todas las variables relacionadas con futura correlación con el clustering como la variable respuesta, risk groups, VWD (que lo analizamos independientemente)... Preparamos la base para Mclust.

Tenemos un FAMD para el que podemos ya hacer Mclust.

Vemos el número de clústeres:

```
cat("\nNúmero óptimo de clusters (genéticos):", fenogeno_mclust$G, "\n")
```

```
##
## Número óptimo de clusters (genéticos): 11
```

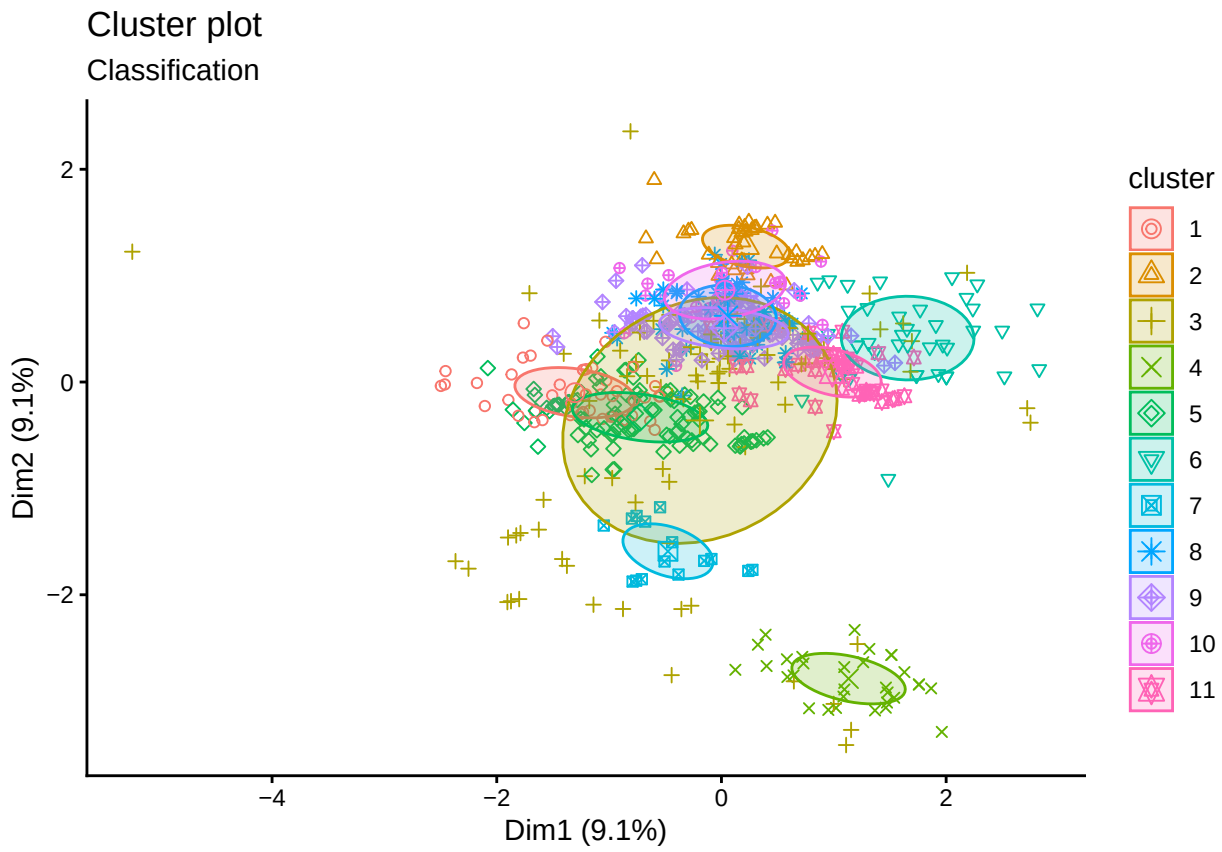
```
print(table(fenogeno_mclust$classification))
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11
##  46  40  89  31  97  38  15  52 110  24  42
```

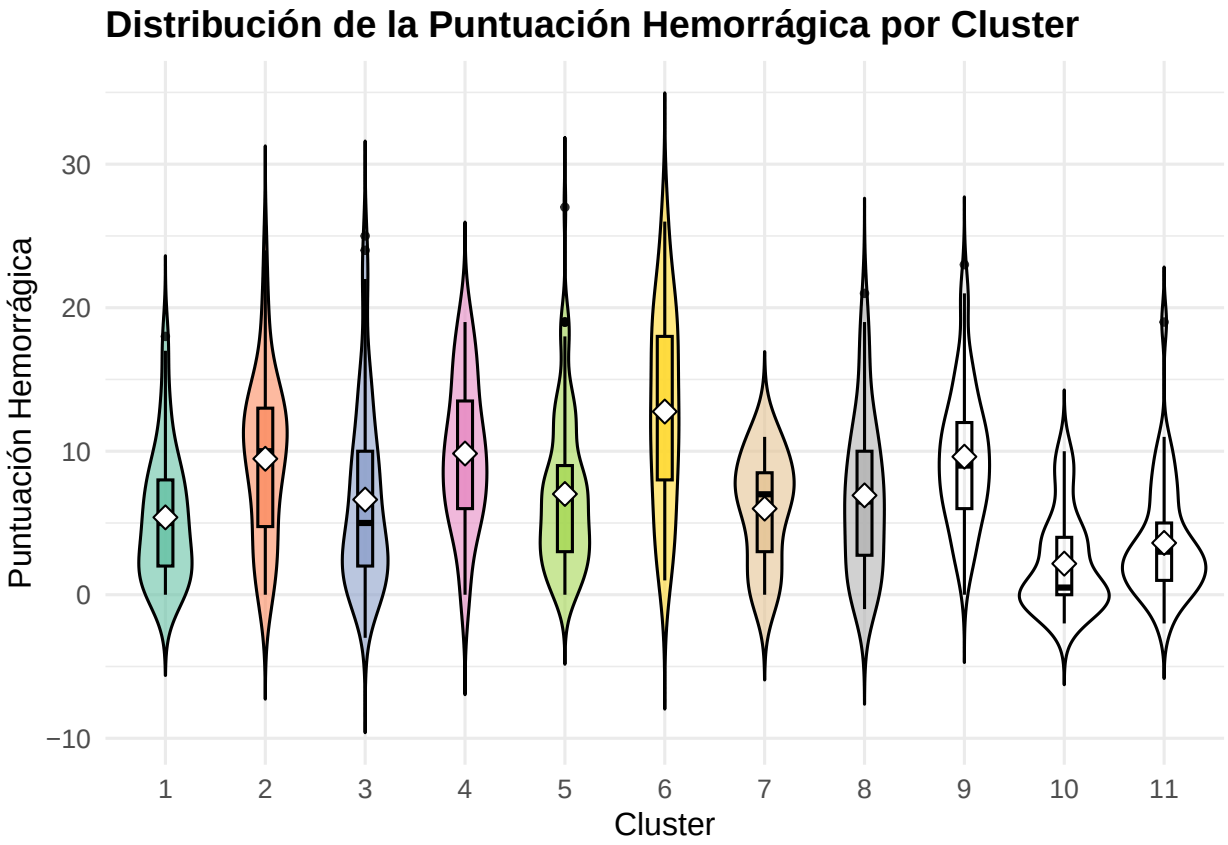
Cuya distribución en clústeres es muy positiva:

- No hay clústeres de uno o dos individuos.
- Número de individuos/clúster < Número de clústeres.
- El tamaño de los clústeres es homogéneo.

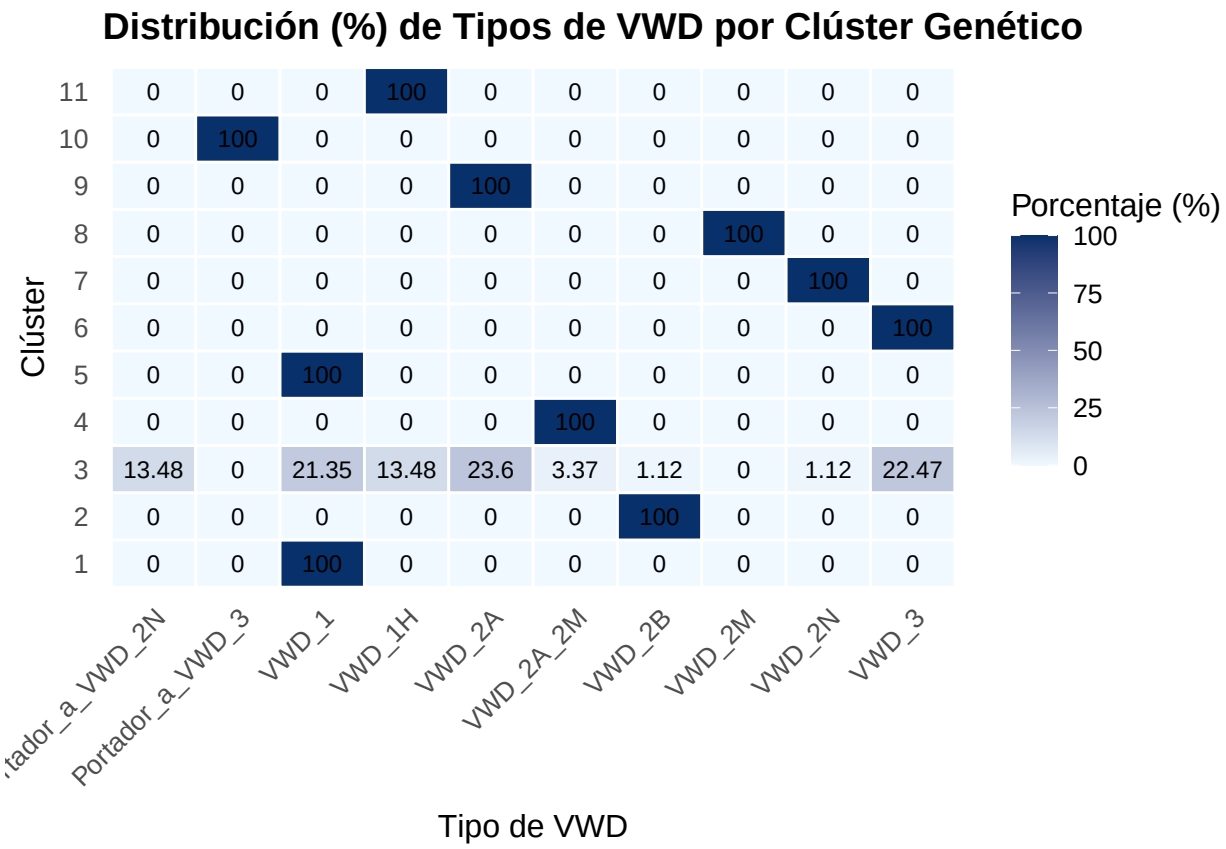
Veamos a continuación el mapa:



En este mapa nos podemos permitir un grado moderado de superposición puesto que solo tenemos las 2 dimensiones principales, y estas cubren un 28,6% de la varianza exclusivamente.

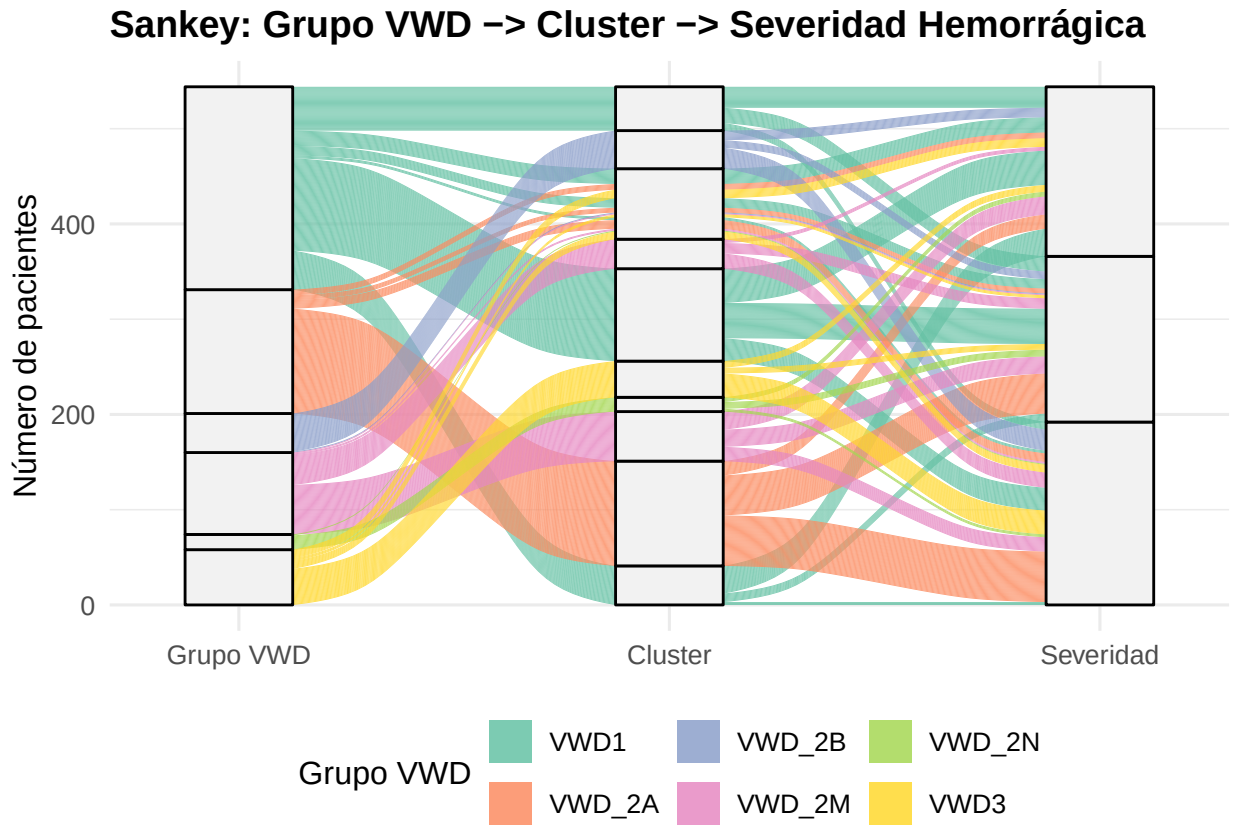


En este gráfico de Puntuación Hemorrágica vemos lo que ya es evidente: Que la señal con la evidencia actual que tenemos -> Puntuación Hemorrágica es muy débil.



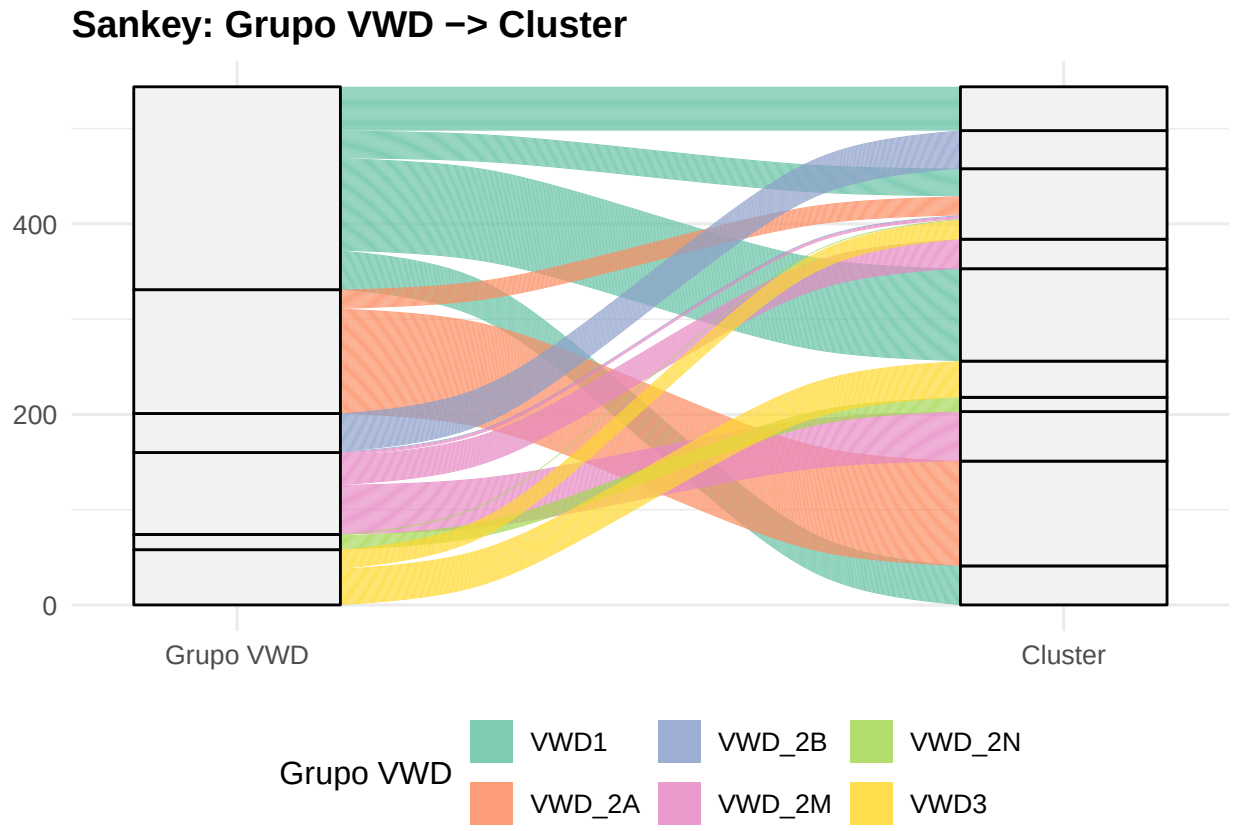
Una de las peores características de esta parte del análisis es que ya no es tan fácil identificar grupos. Podemos ver algunos clústeres que apuntan hacia tipo 3, otras a Portador 2N y 2N, algunas a 2A, otras a 1A, pero no hay mucha claridad entre grupos.

A continuación, vamos a ver cómo mapean los clústeres hacia los grupos y hacia la severidad (en potenciales risk-groups)



Lo que observamos en la mitad izquierda es evidente: no existe una identificación entre clústeres y grupos VWD. Aunque esto parezca negativo, porque nos gustaría encontrar clústeres feno-genotípicos que tuvieran sentido biológico, una interpretación mejor es que esta información será *ortogonal* con respecto de la información VWD, es decir, que toda la varianza que no sea capaz de explicar VWD la podrá complementar fenogeno.

En la mitad derecha tenemos un desafortunado resultado que confirma las sospechas del diagrama de violines: Ni nuestros clústeres ni el grupo VWD son capaces de apuntar ni hacia un score hemorrágico ni hacia unos grupos de riesgo (risk groups).



Este gráfico es tan solo la mitad izquierda del diagrama antes enseñado pero simplificado. Mismas conclusiones.

Por último, vamos a ver el incremento en la capacidad predictora del clústering, medido por el R^2 ajustado y comparado con la práctica clínica actual (a través del tipo VWD)

```
# Lista de fórmulas que quieres evaluar
modelos <- list(
  "FenoGeno" = Puntuación_hemorrágica ~ Score_Fenotipicogenetico,
  "VWD"         = Puntuación_hemorrágica ~ Tipo_VWD,
  "FenoGeno + VWD" = Puntuación_hemorrágica ~ Score_Fenotipicogenetico + Tipo_VWD
)

# Listas para almacenar resultados
resultados_R2 <- list()

# Loop para imprimir summaries y guardar R2
for (nombre in names(modelos)) {

  #cat("\n------------------------------------------------------------\n")
  #cat("### Summary del modelo:", nombre, "###\n")
  #cat("------------------------------------------------------------\n\n")

  modelo <- lm(modelos[[nombre]], data = dataframe_glm)

  # Print del summary
  #print(summary(modelo))
```

```r
  # Guardar R2 en lista
  resultados_R2[[nombre]] <- data.frame(
    Modelo = nombre,
    R2 = summary(modelo)$r.squared,
    R2_Ajustado = summary(modelo)$adj.r.squared
  )
}

# Convertir lista a data.frame final
tabla_R2 <- do.call(rbind, resultados_R2)
tabla_R2[,2:3]
```

```
##                        R2 R2_Ajustado
## FenoGeno       0.1817046   0.1673233
## VWD            0.1825742   0.1696675
## FenoGeno + VWD 0.2173253   0.1936501
```

Según esta información, lo desafortunado es que existe una baja correlación entre el tipo VWD y la puntuación hemorrágica ajustada por un modelo lineal. Lo mismo sucede con la información feno-genotípica y la combinación de Fenogeno y VWD. Sin embargo, una conclusión positiva es que hemos logrado mejorar el ajuste con VWD (la práctica actual) en un 12%.

Conclusiones:

En primer lugar, parece que el análisis tiene luz verde para entrar en el trabajo de clusterización. Hay muchos elementos positivos con respecto a los resultados del algoritmo de Mclust, aunque existen muchos puntos frágiles en mi opinión pertenecientes a los datos.

Algunas consideraciones a este respecto: -El tamaño de la muestra está al borde de ser demasiado bajo (es lo suficientemente grande para la clusterización realizada, pero solo lo bastante grande). Se entiende que el estudio es observacional, lo cual limita nuestras capacidades de pedir más datos. -El mapeo hacia un score hemorrágico con toda la información que tenemos es prácticamente imposible. Es cierto que mejoramos la clasificación al azar (tanto la práctica clínica actual como la información tras el profiling), pero parece completamente impráctica a la hora de tomar decisiones clínicas. Desde mi perspectiva, no sé si se debe a una mala clasificación hemorrágica. Es bastante probable que haya que buscar bibliografía al respecto. Esta bibliografía también ayudará a vender una mejora del 12% como algo clínicamente significativo.

Procedimiento a partir de aquí: Extraeremos la información fenogeno de los pacientes de este informe y elaboraremos un informe nuevo, de validación.

En ese informe, haré una partición de datos aleatorizada por estratificación por tipo de EVW 80% training - 20% test.

Repetiremos el proceso de clusterización con la base training y lo enfrentaremos al test - Aquí forzamos k = 20 clústeres. El objetivo es ver si la nueva asociación en la fracción de testing coincide con la clusterización de este informe.

Elaboraremos una matriz de confusión donde podremos ver la concordancia entre ambos métodos. Dado que la clusterización es una técnica mixta no jerarquizada nuestras métricas de validación tendrán que ser simples, pero bueno, esta debería ser la parte más sencilla. La parte más compleja debería ser el diagnóstico de errores durante la validación.