Práctica 2: Limpieza y análisis de datos

Brais Rodriguez Martinez

12 de diciembre de 2021

Contents

Descripción del dataset

El juego de datos seleccionado es el dataset de "Absenteeism at work" obtenido de la página web https://www.kaggle.com/kewagbln/absenteeism-at-work-uci-ml-repositiory. Éste incluye 21 atributos físicos, psicológicos y ciscunstanciales de 740 trabajadores que permiten realizar los estudios señalados en el apartado anterior.

Se realiza la carga del fichero. Éste, está separado por ";", por lo que se debe especificar este separador. Por otra parte, se definen unas cabeceras más descriptivas:

```
cabeceras <- c("ID", "Razon.ausencia", "Mes", "Dia.semana", "Estacion", "Coste.transporte",
               "Distancia.hogar.trabajo", "Tiempo.servicio", "Edad", "Carga.trabajo.por.dia",
               "Objetivo", "Fallo.disciplinario", "Educacion", "Hijos", "Bebedor.social",
               "Fumador.social", "Mascotas", "Peso", "Altura", "IMC", "Horas.ausente")
absentismo <- read.csv("./Absenteeism_at_work.csv", col.names = cabeceras, sep = ";",
                       encoding = "UTF-8")

str(absentismo)
```

```
## 'data.frame':    740 obs. of  21 variables:
##  $ ID                     : int  11 36 3 7 11 3 10 20 14 1 ...
```

```
##  $ Razon.ausencia         : int  26 0 23 7 23 23 22 23 19 22 ...
##  $ Mes                    : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ Dia.semana             : int  3 3 4 5 5 6 6 6 2 2 ...
##  $ Estacion               : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Coste.transporte       : int  289 118 179 279 289 179 361 260 155 235 ...
##  $ Distancia.hogar.trabajo: int  36 13 51 5 36 51 52 50 12 11 ...
##  $ Tiempo.servicio        : int  13 18 18 14 13 18 3 11 14 14 ...
##  $ Edad                   : int  33 50 38 39 33 38 28 36 34 37 ...
##  $ Carga.trabajo.por.dia  : num  240 240 240 240 240 ...
##  $ Objetivo               : int  97 97 97 97 97 97 97 97 97 97 ...
##  $ Fallo.disciplinario    : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Educacion              : int  1 1 1 1 1 1 1 1 1 3 ...
##  $ Hijos                  : int  2 1 0 2 2 0 1 4 2 1 ...
##  $ Bebedor.social         : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ Fumador.social         : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ Mascotas               : int  1 0 0 0 1 0 4 0 0 1 ...
##  $ Peso                   : int  90 98 89 68 90 89 80 65 95 88 ...
##  $ Altura                 : int  172 178 170 168 172 170 172 168 196 172 ...
##  $ IMC                    : int  30 31 31 24 30 31 27 23 25 29 ...
##  $ Horas.ausente          : int  4 0 2 4 2 2 8 4 40 8 ...
```

Una vez cargado, se verifica su estructura:

1. ID: identificador único que recibe el trabajador.

2. Razon.ausencia: identificador de la razón de la abstinencia. Sus posibles valores van de 1 a 28. Su significado se verá más adelante.

3. Mes: número que identifica el mes del año en el que se produjo la ausencia.

4. Dia.semana: número del día de la semana en el que se produjo la ausencia.

5. Estacion: estación del año en la que se produjo la ausencia.

6. Coste.transporte: cantidad de dinero que cuesta el transporte del hogar del trabajador al trabajo.

7. Distancia.hogar.trabajo: distancia en km del hogar del trabajador al trabajo.

8. Tiempo.servicio: tiempo que el trabajador dedica al servicio semanalmente.

9. Edad: años que tiene el trabajador.

10. Carga.trabajo.por.dia: minutos trabajador por día.

11. Objetivo: objetivo del trabajador, sobre 100.

12. Fallo.disciplinario: indicador booleano que informa si se ha producido un fallo disciplinario por la falta.

13. Educacion: nivel de educación recibida por el trabajador.

14. Hijos: número de hijos.

15. Bebedor.social: indicador de si el trabajador es bebedor social o no.

16. Fumador.social: indicador de si el trabajador es fumador social o no.

17. Mascotas: número de mascotas.

18. Peso: peso en kg del trabajador.

19. Altura: altura en cm del trabajador.

20. IMC: índice de masa corporal del trabajador.

21. Horas.ausente: horas totales en las que el trabajador estuvo ausente.

Importancia y objetivos de los análisis.

El problema a tratar es el estudio de las condiciones, tanto físicas como psicológicas o circunstanciales que provocan el absentismo laboral. De esta manera, se quiere observar si existen relaciones entre estas características y la ausencia mayor o menor del trabajador.

Así, se quieren generar indicadores de relación entre los diferentes atributos para observar si existe dicha correlación. Por otra parte, también se quiere observar si hay ciertos patrones en las características de los trabajadores que provocan las ausencias, formando grupos de ausencias e identificándolos.

Finalmente, sería interesante saber si un trabajador va a ausentarse o cuántas horas podría estar ausente a partir de todos sus datos.

Limpieza de los datos

Preprocesado de los datos

Primeramente se adapta la columna de razón de ausencia, creando una nueva que defina con palabras la razón de la ausencia:

```
absentismo.mod <- absentismo

absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 0] <-
  "Desconocido"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 1] <-
  "Enfermedad infecciosa y parasitaria"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 2] <-
  "Neoplasias"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 3] <-
  "Enfermedades de la sangre"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 4] <-
  "Enfermedades endocrinas, nutricionales y metabólicas"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 5] <-
  "Enfermedades y desórdenes mentales"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 6] <-
  "Enfermedades del sistema nervioso"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 7] <-
  "Enfermedades del ojo y anexos"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 8] <-
  "Enfermedades del oído y apófisis mastoides"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 9] <-
  "Enfermedades del sistema circulatorio"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 10] <-
  "Enfermedades del sistema respiratorio"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 11] <-
  "Enfermedades del sistema digestivo"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 12] <-
  "Enfermedades de la piel y tejido subcutáneo"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 13] <-
  "Enfermedades del sistema musculoesquelético y del tejido conectivo"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 14] <-
  "Enfermedades del sistema genitourinario"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 15] <-
```

```
  "Embarazo, parto y puerperio"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 16] <-
  "Ciertas condiciones que se originan en el período perinatal"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 17] <-
  "Malformaciones congénitas, deformaciones y anomalías cromosómicas"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 18] <-
  "Anomalías no identificadas"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 19] <-
  "Lesiones, intoxicaciones y otras por causas externas"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 20] <-
  "Causas externas de morbilidad y mortalidad"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 21] <-
  "Factores que influyen en la salud"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 22] <-
  "Seguimiento del paciente"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 23] <-
  "Consulta médica"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 24] <-
  "Donación de sangre"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 25] <-
  "Examen en laboratorio"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 26] <-
  "Ausencia no justificada"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 27] <-
  "Fisioterapia"
absentismo.mod$Razon.ausencia.def[absentismo.mod$Razon.ausencia == 28] <-
  "Dentista"

absentismo.mod$Razon.ausencia.def[is.na(absentismo.mod$Razon.ausencia.def)]
```

```
## character(0)
```

```
head(absentismo.mod)
```

```
##   ID Razon.ausencia Mes Dia.semana Estacion Coste.transporte
## 1 11             26   7          3        1              289
## 2 36              0   7          3        1              118
## 3  3             23   7          4        1              179
## 4  7              7   7          5        1              279
## 5 11             23   7          5        1              289
## 6  3             23   7          6        1              179
##   Distancia.hogar.trabajo Tiempo.servicio Edad Carga.trabajo.por.dia Objetivo
## 1                      36              13   33               239.554       97
## 2                      13              18   50               239.554       97
## 3                      51              18   38               239.554       97
## 4                       5              14   39               239.554       97
## 5                      36              13   33               239.554       97
## 6                      51              18   38               239.554       97
##   Fallo.disciplinario Educacion Hijos Bebedor.social Fumador.social Mascotas
## 1                   0         1     2              1              0        1
## 2                   1         1     1              1              0        0
## 3                   0         1     0              1              0        0
## 4                   0         1     2              1              1        0
## 5                   0         1     2              1              0        1
## 6                   0         1     0              1              0        0
```

```
##   Peso Altura IMC Horas.ausente            Razon.ausencia.def
## 1   90    172  30              4       Ausencia no justificada
## 2   98    178  31              0                   Desconocido
## 3   89    170  31              2               Consulta médica
## 4   68    168  24              4 Enfermedades del ojo y anexos
## 5   90    172  30              2               Consulta médica
## 6   89    170  31              2               Consulta médica
```

De la misma manera, se hace el mismo proceso para el mes, el día de la semana y la estación del año:

```
absentismo.mod$Mes.def[absentismo.mod$Mes == 0] <- "Desconocido"
absentismo.mod$Mes.def[absentismo.mod$Mes == 1] <- "Enero"
absentismo.mod$Mes.def[absentismo.mod$Mes == 2] <- "Febrero"
absentismo.mod$Mes.def[absentismo.mod$Mes == 3] <- "Marzo"
absentismo.mod$Mes.def[absentismo.mod$Mes == 4] <- "Abril"
absentismo.mod$Mes.def[absentismo.mod$Mes == 5] <- "Mayo"
absentismo.mod$Mes.def[absentismo.mod$Mes == 6] <- "Junio"
absentismo.mod$Mes.def[absentismo.mod$Mes == 7] <- "Julio"
absentismo.mod$Mes.def[absentismo.mod$Mes == 8] <- "Agosto"
absentismo.mod$Mes.def[absentismo.mod$Mes == 9] <- "Septiembre"
absentismo.mod$Mes.def[absentismo.mod$Mes == 10] <- "Octubre"
absentismo.mod$Mes.def[absentismo.mod$Mes == 11] <- "Noviembre"
absentismo.mod$Mes.def[absentismo.mod$Mes == 12] <- "Diciembre"

absentismo.mod$Mes.def[is.na(absentismo.mod$Mes.def)]
```

```
## character(0)
```

```
absentismo.mod$Dia.semana.def[absentismo.mod$Dia.semana == 2] <- "Lunes"
absentismo.mod$Dia.semana.def[absentismo.mod$Dia.semana == 3] <- "Martes"
absentismo.mod$Dia.semana.def[absentismo.mod$Dia.semana == 4] <- "Miércoles"
absentismo.mod$Dia.semana.def[absentismo.mod$Dia.semana == 5] <- "Jueves"
absentismo.mod$Dia.semana.def[absentismo.mod$Dia.semana == 6] <- "Viernes"

absentismo.mod$Dia.semana.def[is.na(absentismo.mod$Dia.semana.def)]
```

```
## character(0)
```

```
absentismo.mod$Estacion.def[absentismo.mod$Estacion == 1] <- "Primavera"
absentismo.mod$Estacion.def[absentismo.mod$Estacion == 2] <- "Verano"
absentismo.mod$Estacion.def[absentismo.mod$Estacion == 3] <- "Otoño"
absentismo.mod$Estacion.def[absentismo.mod$Estacion == 4] <- "Invierno"

absentismo.mod$Estacion.def[is.na(absentismo.mod$Estacion.def)]
```

```
## character(0)
```

```
head(absentismo.mod)
```

```
##   ID Razon.ausencia Mes Dia.semana Estacion Coste.transporte
## 1 11             26   7          3        1              289
## 2 36              0   7          3        1              118
## 3  3             23   7          4        1              179
## 4  7              7   7          5        1              279
## 5 11             23   7          5        1              289
## 6  3             23   7          6        1              179
##   Distancia.hogar.trabajo Tiempo.servicio Edad Carga.trabajo.por.dia Objetivo
## 1                      36              13   33               239.554       97
```

```
## 2                      13             18   50              239.554       97
## 3                      51             18   38              239.554       97
## 4                       5             14   39              239.554       97
## 5                      36             13   33              239.554       97
## 6                      51             18   38              239.554       97
##   Fallo.disciplinario Educacion Hijos Bebedor.social Fumador.social Mascotas
## 1                   0         1     2              1              0        1
## 2                   1         1     1              1              0        0
## 3                   0         1     0              1              0        0
## 4                   0         1     2              1              1        0
## 5                   0         1     2              1              0        1
## 6                   0         1     0              1              0        0
##   Peso Altura IMC Horas.ausente            Razon.ausencia.def Mes.def
## 1   90    172  30             4       Ausencia no justificada   Julio
## 2   98    178  31             0                   Desconocido   Julio
## 3   89    170  31             2               Consulta médica   Julio
## 4   68    168  24             4 Enfermedades del ojo y anexos   Julio
## 5   90    172  30             2               Consulta médica   Julio
## 6   89    170  31             2               Consulta médica   Julio
##   Dia.semana.def Estacion.def
## 1         Martes    Primavera
## 2         Martes    Primavera
## 3      Miércoles    Primavera
## 4         Jueves    Primavera
## 5         Jueves    Primavera
## 6        Viernes    Primavera
```

Se aplica el mismo proceso para el nivel de educación:

```
absentismo.mod$Educacion.def[absentismo.mod$Educacion == 1] <- "Instituto"
absentismo.mod$Educacion.def[absentismo.mod$Educacion == 2] <- "Graduado"
absentismo.mod$Educacion.def[absentismo.mod$Educacion == 3] <- "Post-graduado"
absentismo.mod$Educacion.def[absentismo.mod$Educacion == 4] <- "Máster y doctor"

absentismo.mod$Educacion.def[is.na(absentismo.mod$Educacion.def)]
```

```
## character(0)
```

```
head(absentismo.mod)
```

```
##   ID Razon.ausencia Mes Dia.semana Estacion Coste.transporte
## 1 11             26   7          3        1              289
## 2 36              0   7          3        1              118
## 3  3             23   7          4        1              179
## 4  7              7   7          5        1              279
## 5 11             23   7          5        1              289
## 6  3             23   7          6        1              179
##   Distancia.hogar.trabajo Tiempo.servicio Edad Carga.trabajo.por.dia Objetivo
## 1                      36             13   33              239.554       97
## 2                      13             18   50              239.554       97
## 3                      51             18   38              239.554       97
## 4                       5             14   39              239.554       97
## 5                      36             13   33              239.554       97
## 6                      51             18   38              239.554       97
##   Fallo.disciplinario Educacion Hijos Bebedor.social Fumador.social Mascotas
## 1                   0         1     2              1              0        1
```

```
## 2                  1         1     1              1              0        0
## 3                  0         1     0              1              0        0
## 4                  0         1     2              1              1        0
## 5                  0         1     2              1              0        1
## 6                  0         1     0              1              0        0
##   Peso Altura IMC Horas.ausente            Razon.ausencia.def Mes.def
## 1   90    172  30             4       Ausencia no justificada   Julio
## 2   98    178  31             0                   Desconocido   Julio
## 3   89    170  31             2               Consulta médica   Julio
## 4   68    168  24             4 Enfermedades del ojo y anexos   Julio
## 5   90    172  30             2               Consulta médica   Julio
## 6   89    170  31             2               Consulta médica   Julio
##   Dia.semana.def Estacion.def Educacion.def
## 1         Martes    Primavera     Instituto
## 2         Martes    Primavera     Instituto
## 3      Miércoles    Primavera     Instituto
## 4         Jueves    Primavera     Instituto
## 5         Jueves    Primavera     Instituto
## 6        Viernes    Primavera     Instituto
```

Para futuros estudios, también sería interesante crear dos nuevas columnas que indican si el trabajador tiene hijos y otra que indica si el trabajador tiene mascotas:

```
absentismo.mod$Tiene.hijos[absentismo.mod$Hijos > 0] <- 1
absentismo.mod$Tiene.hijos[absentismo.mod$Hijos == 0] <- 0

absentismo.mod$Tiene.hijos[is.na(absentismo.mod$Tiene.hijos)]
```

```
## numeric(0)
```

```
absentismo.mod$Tiene.mascotas[absentismo.mod$Mascotas > 0] <- 1
absentismo.mod$Tiene.mascotas[absentismo.mod$Mascotas == 0] <- 0

absentismo.mod$Tiene.mascotas[is.na(absentismo.mod$Tiene.mascotas)]
```

```
## numeric(0)
```

```
head(absentismo.mod)
```

```
##   ID Razon.ausencia Mes Dia.semana Estacion Coste.transporte
## 1 11             26   7          3        1              289
## 2 36              0   7          3        1              118
## 3  3             23   7          4        1              179
## 4  7              7   7          5        1              279
## 5 11             23   7          5        1              289
## 6  3             23   7          6        1              179
##   Distancia.hogar.trabajo Tiempo.servicio Edad Carga.trabajo.por.dia Objetivo
## 1                      36              13   33               239.554       97
## 2                      13              18   50               239.554       97
## 3                      51              18   38               239.554       97
## 4                       5              14   39               239.554       97
## 5                      36              13   33               239.554       97
## 6                      51              18   38               239.554       97
##   Fallo.disciplinario Educacion Hijos Bebedor.social Fumador.social Mascotas
## 1                   0         1     2              1              0        1
## 2                   1         1     1              1              0        0
## 3                   0         1     0              1              0        0
```

```
## 4                   0         1     2              1              1        0
## 5                   0         1     2              1              0        1
## 6                   0         1     0              1              0        0
##   Peso Altura IMC Horas.ausente            Razon.ausencia.def Mes.def
## 1   90    172  30             4       Ausencia no justificada   Julio
## 2   98    178  31             0                   Desconocido   Julio
## 3   89    170  31             2               Consulta médica   Julio
## 4   68    168  24             4 Enfermedades del ojo y anexos   Julio
## 5   90    172  30             2               Consulta médica   Julio
## 6   89    170  31             2               Consulta médica   Julio
##   Dia.semana.def Estacion.def Educacion.def Tiene.hijos Tiene.mascotas
## 1         Martes    Primavera     Instituto           1              1
## 2         Martes    Primavera     Instituto           1              0
## 3      Miércoles    Primavera     Instituto           0              0
## 4         Jueves    Primavera     Instituto           1              0
## 5         Jueves    Primavera     Instituto           1              1
## 6        Viernes    Primavera     Instituto           0              0
```

Con el objetivo de realizar las representaciones más claras, se crea una nueva columna que indica SÍ o NO dependiendo del indicador de las columnas booleanas:

```r
absentismo.mod$Fallo.disciplinario.def[absentismo.mod$Fallo.disciplinario == 1] <- "SÍ"
absentismo.mod$Fallo.disciplinario.def[absentismo.mod$Fallo.disciplinario == 0] <- "NO"

absentismo.mod$Bebedor.social.def[absentismo.mod$Bebedor.social == 1] <- "SÍ"
absentismo.mod$Bebedor.social.def[absentismo.mod$Bebedor.social == 0] <- "NO"

absentismo.mod$Fumador.social.def[absentismo.mod$Fumador.social == 1] <- "SÍ"
absentismo.mod$Fumador.social.def[absentismo.mod$Fumador.social == 0] <- "NO"

absentismo.mod$Tiene.hijos.def[absentismo.mod$Tiene.hijos == 1] <- "SÍ"
absentismo.mod$Tiene.hijos.def[absentismo.mod$Tiene.hijos == 0] <- "NO"

absentismo.mod$Tiene.mascotas.def[absentismo.mod$Tiene.mascotas == 1] <- "SÍ"
absentismo.mod$Tiene.mascotas.def[absentismo.mod$Tiene.mascotas == 0] <- "NO"

head(absentismo.mod)
```

```
##   ID Razon.ausencia Mes Dia.semana Estacion Coste.transporte
## 1 11             26   7          3        1              289
## 2 36              0   7          3        1              118
## 3  3             23   7          4        1              179
## 4  7              7   7          5        1              279
## 5 11             23   7          5        1              289
## 6  3             23   7          6        1              179
##   Distancia.hogar.trabajo Tiempo.servicio Edad Carga.trabajo.por.dia Objetivo
## 1                      36              13   33               239.554       97
## 2                      13              18   50               239.554       97
## 3                      51              18   38               239.554       97
## 4                       5              14   39               239.554       97
## 5                      36              13   33               239.554       97
## 6                      51              18   38               239.554       97
##   Fallo.disciplinario Educacion Hijos Bebedor.social Fumador.social Mascotas
## 1                   0         1     2              1              0        1
## 2                   1         1     1              1              0        0
```

```
## 3                  0         1     0              1              0        0
## 4                  0         1     2              1              1        0
## 5                  0         1     2              1              0        1
## 6                  0         1     0              1              0        0
##   Peso Altura IMC Horas.ausente           Razon.ausencia.def Mes.def
## 1   90    172  30             4       Ausencia no justificada   Julio
## 2   98    178  31             0                   Desconocido   Julio
## 3   89    170  31             2               Consulta médica   Julio
## 4   68    168  24             4 Enfermedades del ojo y anexos   Julio
## 5   90    172  30             2               Consulta médica   Julio
## 6   89    170  31             2               Consulta médica   Julio
##   Dia.semana.def Estacion.def Educacion.def Tiene.hijos Tiene.mascotas
## 1         Martes    Primavera     Instituto           1              1
## 2         Martes    Primavera     Instituto           1              0
## 3      Miércoles    Primavera     Instituto           0              0
## 4         Jueves    Primavera     Instituto           1              0
## 5         Jueves    Primavera     Instituto           1              1
## 6        Viernes    Primavera     Instituto           0              0
##   Fallo.disciplinario.def Bebedor.social.def Fumador.social.def Tiene.hijos.def
## 1                      NO                SÍ                 NO              SÍ
## 2                      SÍ                SÍ                 NO              SÍ
## 3                      NO                SÍ                 NO              NO
## 4                      NO                SÍ                 SÍ              SÍ
## 5                      NO                SÍ                 NO              SÍ
## 6                      NO                SÍ                 NO              NO
##   Tiene.mascotas.def
## 1                 SÍ
## 2                 NO
## 3                 NO
## 4                 NO
## 5                 SÍ
## 6                 NO
```

Por otra parte, con el objetivo de mantener el sentido de los datos, se pasa el atributo de carga de trabajo por día de minutos a horas, igual que otros atributos como Tiempo.servicio u Horas ausente:

```
absentismo.mod$Carga.trabajo.por.dia <- absentismo.mod$Carga.trabajo.por.dia/60

summary(absentismo.mod$Carga.trabajo.por.dia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.432   4.073   4.404   4.525   4.904   6.315
```

Discretización

Para análisis próximos, se procede a discretizar la variable IMC para obtener los diferentes estados IMC de una persona:

```
absentismo.mod["Rango.IMC"] <- cut(absentismo.mod$IMC, breaks = c(0,18.5,25.0,30,40),
                                   labels = c("Inferior", "Normal","Sobrepeso","Obeso"))

head(absentismo.mod$Rango.IMC)
```

```
## [1] Sobrepeso Obeso     Obeso     Normal    Sobrepeso Obeso
## Levels: Inferior Normal Sobrepeso Obeso
```

Comprobación de elementos a cero o vacíos

Se realiza un resumen de los datos para ver si, a priori, parecen correctos o no.

```
summary(absentismo.mod)
```

```
##        ID         Razon.ausencia       Mes           Dia.semana
##  Min.   : 1.00   Min.   : 0.00   Min.   : 0.000   Min.   :2.000
##  1st Qu.: 9.00   1st Qu.:13.00   1st Qu.: 3.000   1st Qu.:3.000
##  Median :18.00   Median :23.00   Median : 6.000   Median :4.000
##  Mean   :18.02   Mean   :19.22   Mean   : 6.324   Mean   :3.915
##  3rd Qu.:28.00   3rd Qu.:26.00   3rd Qu.: 9.000   3rd Qu.:5.000
##  Max.   :36.00   Max.   :28.00   Max.   :12.000   Max.   :6.000
##     Estacion     Coste.transporte Distancia.hogar.trabajo Tiempo.servicio
##  Min.   :1.000   Min.   :118.0    Min.   : 5.00           Min.   : 1.00
##  1st Qu.:2.000   1st Qu.:179.0    1st Qu.:16.00           1st Qu.: 9.00
##  Median :3.000   Median :225.0    Median :26.00           Median :13.00
##  Mean   :2.545   Mean   :221.3    Mean   :29.63           Mean   :12.55
##  3rd Qu.:4.000   3rd Qu.:260.0    3rd Qu.:50.00           3rd Qu.:16.00
##  Max.   :4.000   Max.   :388.0    Max.   :52.00           Max.   :29.00
##       Edad       Carga.trabajo.por.dia    Objetivo      Fallo.disciplinario
##  Min.   :27.00   Min.   :3.432         Min.   : 81.00   Min.   :0.00000
##  1st Qu.:31.00   1st Qu.:4.073         1st Qu.: 93.00   1st Qu.:0.00000
##  Median :37.00   Median :4.404         Median : 95.00   Median :0.00000
##  Mean   :36.45   Mean   :4.525         Mean   : 94.59   Mean   :0.05405
##  3rd Qu.:40.00   3rd Qu.:4.904         3rd Qu.: 97.00   3rd Qu.:0.00000
##  Max.   :58.00   Max.   :6.315         Max.   :100.00   Max.   :1.00000
##    Educacion         Hijos       Bebedor.social   Fumador.social
##  Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :1.000   Median :1.000   Median :1.0000   Median :0.00000
##  Mean   :1.292   Mean   :1.019   Mean   :0.5676   Mean   :0.07297
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.00000
##     Mascotas           Peso             Altura           IMC
##  Min.   :0.0000   Min.   : 56.00   Min.   :163.0   Min.   :19.00
##  1st Qu.:0.0000   1st Qu.: 69.00   1st Qu.:169.0   1st Qu.:24.00
##  Median :0.0000   Median : 83.00   Median :170.0   Median :25.00
##  Mean   :0.7459   Mean   : 79.04   Mean   :172.1   Mean   :26.68
##  3rd Qu.:1.0000   3rd Qu.: 89.00   3rd Qu.:172.0   3rd Qu.:31.00
##  Max.   :8.0000   Max.   :108.00   Max.   :196.0   Max.   :38.00
##  Horas.ausente     Razon.ausencia.def   Mes.def          Dia.semana.def
##  Min.   :  0.000   Length:740         Length:740         Length:740
##  1st Qu.:  2.000   Class :character   Class :character   Class :character
##  Median :  3.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :  6.924
##  3rd Qu.:  8.000
##  Max.   :120.000
##  Estacion.def       Educacion.def       Tiene.hijos     Tiene.mascotas
##  Length:740         Length:740         Min.   :0.0000   Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :1.0000   Median :0.0000
##                                        Mean   :0.5973   Mean   :0.3784
##                                        3rd Qu.:1.0000   3rd Qu.:1.0000
##                                        Max.   :1.0000   Max.   :1.0000
```

```
##  Fallo.disciplinario.def Bebedor.social.def Fumador.social.def
##  Length:740              Length:740         Length:740
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##  Tiene.hijos.def    Tiene.mascotas.def     Rango.IMC
##  Length:740         Length:740         Inferior :  0
##  Class :character   Class :character   Normal   :390
##  Mode  :character   Mode  :character   Sobrepeso:146
##                                        Obeso    :204
##
##
```

```
absentismo.mod[is.na(absentismo.mod)]
```

```
## character(0)
```

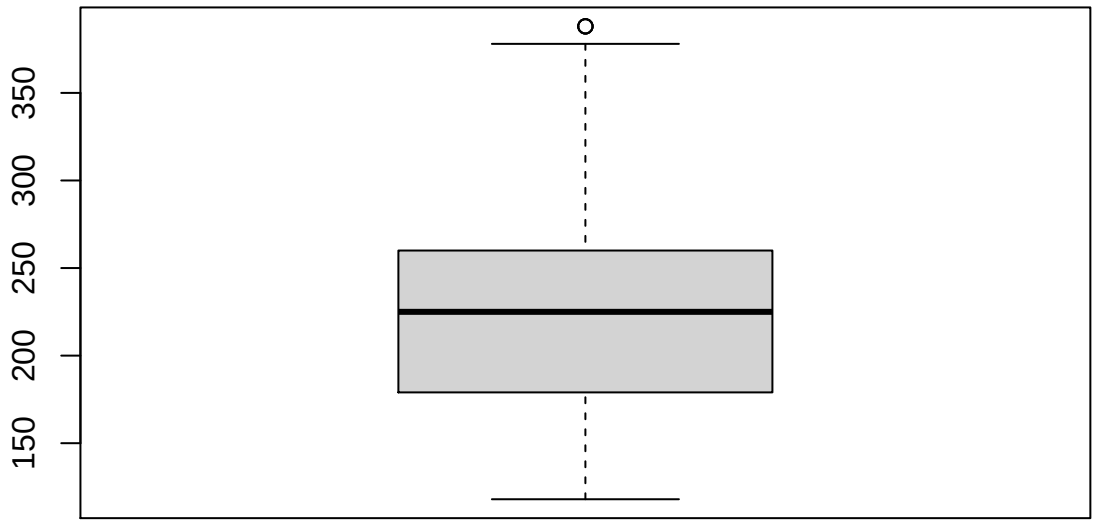
A primera vista, parece que todos los datos son correctos. Además, tampoco hay ningún NA en éstos, por lo que no será necesario realizar correcciones sobre los mismos.

Comprobación de valores extremos

Una vez analizados los atributos, interesaría saber si existen valores extremos en: Coste.transporte, Distancia.hogar.trabajo, Tiempo.servicio, Edad, Carga.trabajo.por.dia, IMC y Horas.ausente:

Primeramente, Coste.trasporte:

```
boxplot <- boxplot(absentismo.mod$Coste.transporte)
```



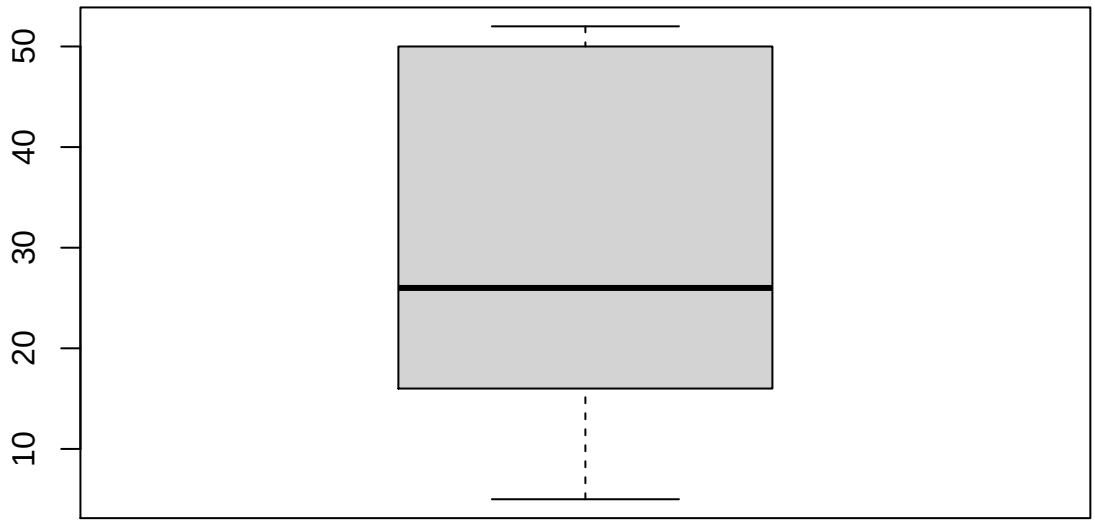

```
boxplot$out
```

```
## [1] 388 388 388
```

Se puede observar, en el caso del coste de transporte, hay 3 registros que muestran un valor extremo por arriba, concretamente 388. En este caso, no se considera un error, sino que es un valor real desviado de la media.

Segundo, Distancia.hogar.trabajo:

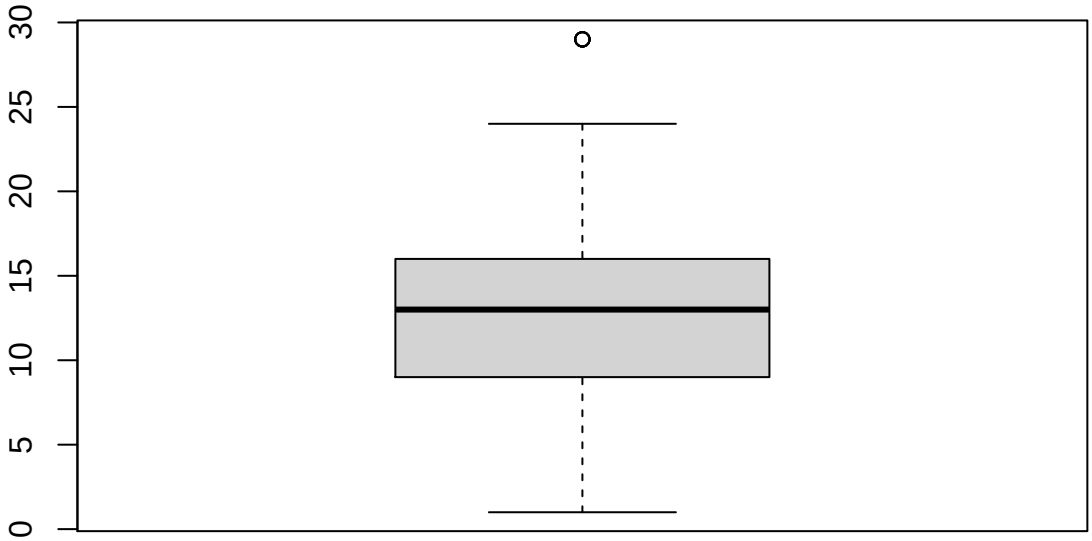
```
boxplot(absentismo.mod$Distancia.hogar.trabajo)
```



Para este caso se puede ver que no hay valores extremos, por lo que parece que todos los trabajadores viven en las cercanías de la oficina, ya que no hay ninguna distancia fuera de lo común.

Por otra parte, Tiempo.servicio:

```
boxplot <- boxplot(absentismo.mod$Tiempo.servicio)
```



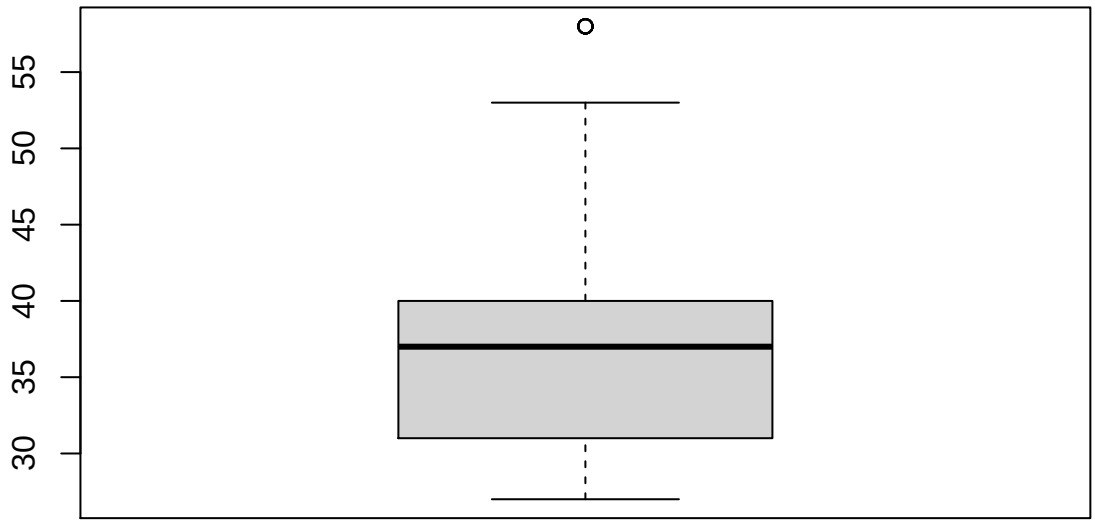

```
boxplot$out
```

```
## [1] 29 29 29 29 29
```

Para Tiempo.servicio, el outlier es 29. Esto muestra que hay algunas personas que tienen mucha experiencia ya en la compañía, como podría ser algún jefe o director, por lo que no se considera un valor erróneo.

Para el atributo Edad:

```
boxplot <- boxplot(absentismo.mod$Edad)
```



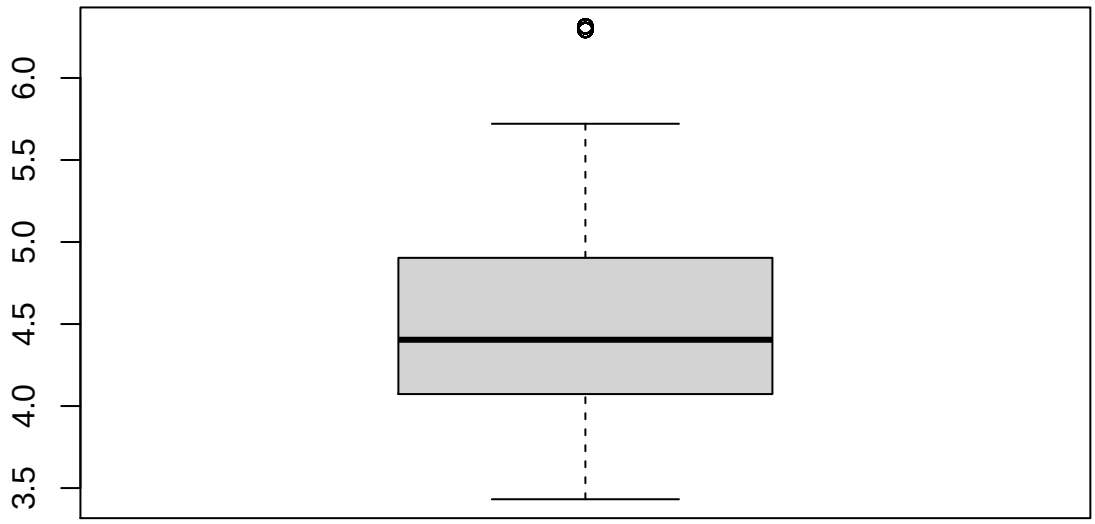

```
boxplot$out
```

```
## [1] 58 58 58 58 58 58 58 58
```

Se observa el número 58 como un outlier por la parte superior, mostrando que hay una o varias personas de avanzada edad. Igual que los casos anteriores, se considera un valor dentro de lo normal, ya que una persona de esa edad puede seguir trabajando sin problema.

Ahora, para el atributo Carga.trabajo.por.dia:

```
boxplot <- boxplot(absentismo.mod$Carga.trabajo.por.dia)
```



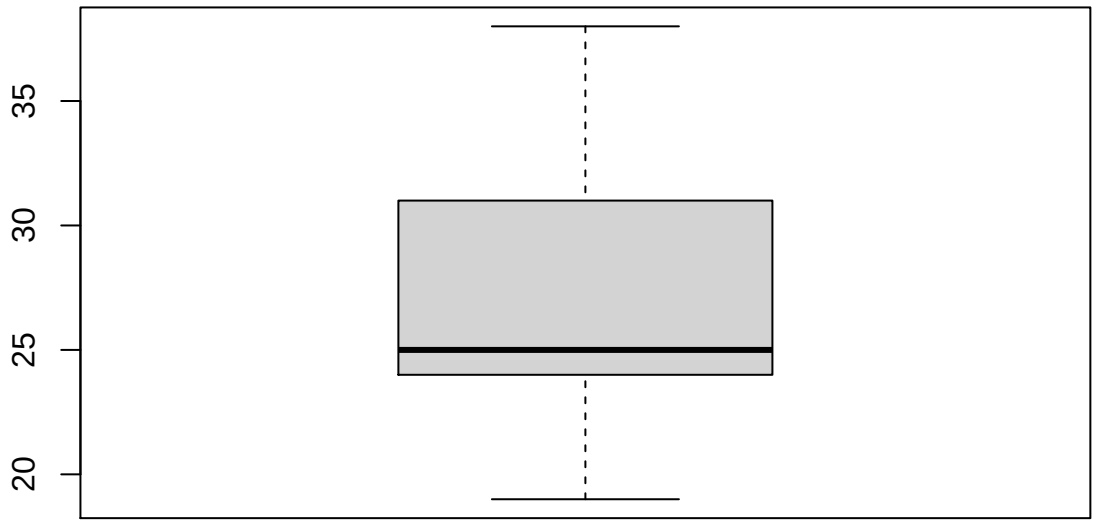

```
boxplot$out
```

```
##  [1] 6.314733 6.314733 6.314733 6.314733 6.314733 6.314733 6.314733 6.314733
##  [9] 6.314733 6.314733 6.314733 6.314733 6.314733 6.314733 6.314733 6.314733
## [17] 6.292500 6.292500 6.292500 6.292500 6.292500 6.292500 6.292500 6.292500
## [25] 6.292500 6.292500 6.292500 6.292500 6.292500 6.292500 6.292500 6.292500
```

En este caso ya hay más outliers por la parte superior. Se podrían llegar a considerar datos correctos. Aunque sea una cantidad de horas diarias dentro de lo normal, sería necesaria una investigación acerca del porqué están tan alejadas de la media; por ejemplo, es posible que se trabaje los 7 días de la semana y 6.3 horas al día se puede considerar una cifra inadecuada al superarse las 40 horas semanales.

Pasando al atributo IMC:

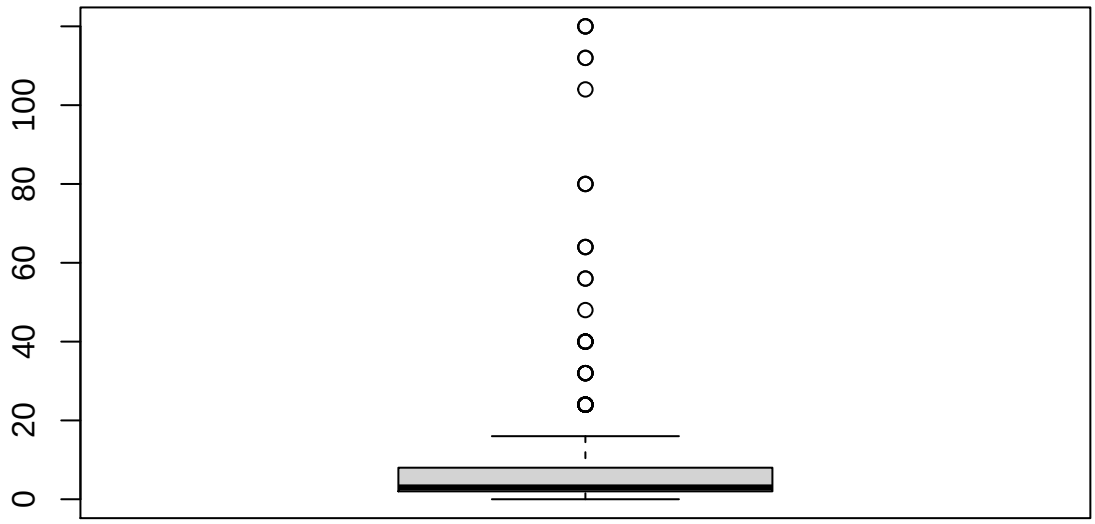
```
boxplot(absentismo.mod$IMC)
```



No se ha obtenido ningún outlier para IMC, lo que lleva a pensar que la forma física de los trabajadores es similar.

Por último, el caso de Horas.ausente:

```
boxplot <- boxplot(absentismo.mod$Horas.ausente)
```




```
boxplot$out
```

```
##  [1]  40  40  32  32  40  24  64  56  40  40  24  24  24  56  24  24  24  24  80
## [20]  32  24  32  40  64 120  32  24 120  40  24 112  24  32  80  24 112  24 104
## [39]  24  64  48  24 120  80
```

En este último caso hay bastantes outliers y todos muy variados. Todos ellos son valores correctos ya que no se salen de la realidad, pero en el entorno de la compañía sería necesario ver las razones de las ausencias.

Análisis de los datos

Selección de los grupos de datos a analizar

A primera vista, los factores de la estación del año y el rango IMC en el que se encuentra el trabajador pueden ser indicadores críticos que causan ausencias de mayor o menor medida. Por otra parte, por si fuese necesario en futuros análisis, se quiere agrupar a los trabajadores por nivel de educación.

Agrupación por nivel de educación

Los niveles de educación del dataset son "Graduado", "Instituto", "Máster y doctor" y "Post-graduado":

```
summary(factor(absentismo.mod$Educacion.def))
```

```
##        Graduado       Instituto Máster y doctor   Post-graduado
##              46             611               4              79
```

```
absentismo.mod.graduado <-
absentismo.mod[absentismo.mod$Educacion.def == "Graduado",]
```

```
absentismo.mod.instituto <-
absentismo.mod[absentismo.mod$Educacion.def == "Instituto",]
absentismo.mod.masterdoctor <-
absentismo.mod[absentismo.mod$Educacion.def == "Máster y doctor",]
absentismo.mod.postgraduado <-
absentismo.mod[absentismo.mod$Educacion.def == "Post-graduado",]
```

Agrupación por estación del año

Por otra parte, las estaciones son "Primavera", "Verano", "Otoño" e "Invierno":

```
summary(factor(absentismo.mod$Estacion.def))
```

```
##  Invierno     Otoño Primavera    Verano
##       195       183       170       192
```

```
absentismo.mod.primavera <-
absentismo.mod[absentismo.mod$Estacion.def == "Primavera",]
absentismo.mod.verano <-
absentismo.mod[absentismo.mod$Estacion.def == "Verano",]
absentismo.mod.otonho <-
absentismo.mod[absentismo.mod$Estacion.def == "Otoño",]
absentismo.mod.invierno <-
absentismo.mod[absentismo.mod$Estacion.def == "Invierno",]
```

Agrupación por rango de IMC

Finalmente, entre los rangos de IMC se distinguen "Inferior", "Normal", "Sobrepeso" y "Obesidad":

```
summary(factor(absentismo.mod$Rango.IMC))
```

```
##    Normal Sobrepeso     Obeso
##       390       146       204
```

```
absentismo.mod.inferior <-
absentismo.mod[absentismo.mod$Rango.IMC == "Inferior",]
absentismo.mod.normal <-
absentismo.mod[absentismo.mod$Rango.IMC == "Normal",]
absentismo.mod.sobrepeso <-
absentismo.mod[absentismo.mod$Rango.IMC == "Sobrepeso",]
absentismo.mod.obesidad <-
absentismo.mod[absentismo.mod$Rango.IMC == "Obeso",]
```

Comprobación de la normalidad y homogeneidad de la varianza

A continuación, se va a comprobar la normalidad de los distintos atributos numéricos que se han destacado hasta ahora:

Primeramente, Coste.trasporte:

```
shapiro.test(absentismo$Coste.transporte)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  absentismo$Coste.transporte
## W = 0.94566, p-value = 7.717e-16
```

Segundo, Distancia.hogar.trabajo:

```
shapiro.test(absentismo$Distancia.hogar.trabajo)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  absentismo$Distancia.hogar.trabajo
## W = 0.87832, p-value < 2.2e-16
```

Por otra parte, Tiempo.servicio:

```
shapiro.test(absentismo$Tiempo.servicio)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  absentismo$Tiempo.servicio
## W = 0.94317, p-value = 3.161e-16
```

Para el atributo Edad:

```
shapiro.test(absentismo$Edad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  absentismo$Edad
## W = 0.92845, p-value < 2.2e-16
```

Ahora, para el atributo Carga.trabajo.por.dia:

```
shapiro.test(absentismo$Carga.trabajo.por.dia)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  absentismo$Carga.trabajo.por.dia
## W = 0.9232, p-value < 2.2e-16
```

Pasando al atributo IMC:

```
shapiro.test(absentismo$IMC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  absentismo$IMC
## W = 0.94565, p-value = 7.694e-16
```

Por último, Horas.ausente:

```
shapiro.test(absentismo$Horas.ausente)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  absentismo$Horas.ausente
## W = 0.40081, p-value < 2.2e-16
```

Ninguno de los atributos estudiados sigue una distribución normal ya que el test de shapiro ha obtenido unos p-value inferiores al nivel de significancia, es decir, de 0.05.

Una vez comprobada la normalidad de los atributos, toca comprobar si existe homocedasticidad:

```
fligner.test(Distancia.hogar.trabajo ~ Coste.transporte, data = absentismo.mod)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Distancia.hogar.trabajo by Coste.transporte
## Fligner-Killeen:med chi-squared = 213.74, df = 23, p-value < 2.2e-16
```

Se comprueba que no se verifica la hipótesis nula, es decir, la varianza no es la misma ya que el p-value es inferior al nivel de significancia 0.05.

Pruebas estadísticas

Matriz de correlaciones

Es importante ver las correlaciones entre la diversas variables vistas:

```
absentismo.cor <- cor(absentismo.mod[,c(c(2:21))])
round(absentismo.cor, 2)
```

```
##                          Razon.ausencia   Mes Dia.semana Estacion
## Razon.ausencia                     1.00 -0.08       0.12    -0.12
## Mes                               -0.08  1.00      -0.01     0.41
## Dia.semana                         0.12 -0.01       1.00     0.05
## Estacion                          -0.12  0.41       0.05     1.00
## Coste.transporte                  -0.12  0.14       0.03     0.04
## Distancia.hogar.trabajo            0.16  0.00       0.12    -0.06
## Tiempo.servicio                    0.05 -0.06       0.02    -0.01
## Edad                              -0.08  0.00       0.00    -0.01
## Carga.trabajo.por.dia             -0.12 -0.17       0.02     0.15
## Objetivo                           0.09 -0.46       0.03    -0.06
## Fallo.disciplinario               -0.55  0.11      -0.02     0.15
## Educacion                         -0.05 -0.07       0.06     0.00
## Hijos                             -0.06  0.08       0.10     0.05
## Bebedor.social                     0.07  0.06       0.04    -0.05
## Fumador.social                    -0.12 -0.04       0.01    -0.05
## Mascotas                          -0.06  0.05      -0.03     0.01
## Peso                               0.00  0.02      -0.13    -0.03
## Altura                            -0.08 -0.07      -0.08    -0.03
## IMC                                0.04  0.05      -0.10    -0.01
## Horas.ausente                     -0.17  0.02      -0.12    -0.01
##                          Coste.transporte Distancia.hogar.trabajo
## Razon.ausencia                      -0.12                    0.16
## Mes                                  0.14                    0.00
## Dia.semana                           0.03                    0.12
## Estacion                             0.04                   -0.06
## Coste.transporte                     1.00                    0.26
## Distancia.hogar.trabajo              0.26                    1.00
## Tiempo.servicio                     -0.35                    0.13
## Edad                                -0.23                   -0.15
## Carga.trabajo.por.dia                0.01                   -0.07
```

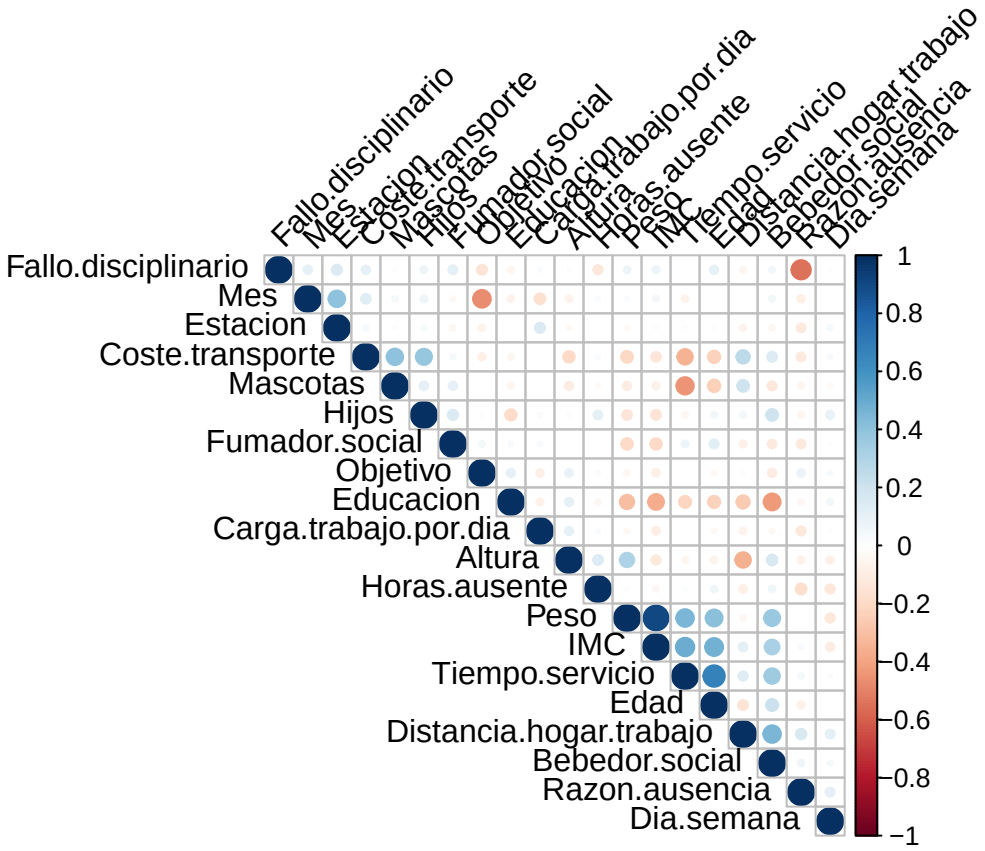
```
## Objetivo                          -0.08                  -0.01
## Fallo.disciplinario                0.11                  -0.06
## Educacion                         -0.06                  -0.26
## Hijos                              0.38                   0.05
## Bebedor.social                     0.15                   0.45
## Fumador.social                     0.04                  -0.08
## Mascotas                           0.40                   0.21
## Peso                              -0.21                  -0.05
## Altura                            -0.19                  -0.35
## IMC                               -0.14                   0.11
## Horas.ausente                      0.03                  -0.09
##                          Tiempo.servicio  Edad Carga.trabajo.por.dia Objetivo
## Razon.ausencia                      0.05 -0.08                -0.12     0.09
## Mes                                -0.06  0.00                -0.17    -0.46
## Dia.semana                          0.02  0.00                 0.02     0.03
## Estacion                           -0.01 -0.01                 0.15    -0.06
## Coste.transporte                   -0.35 -0.23                 0.01    -0.08
## Distancia.hogar.trabajo             0.13 -0.15                -0.07    -0.01
## Tiempo.servicio                     1.00  0.67                 0.00    -0.01
## Edad                                0.67  1.00                -0.04    -0.04
## Carga.trabajo.por.dia               0.00 -0.04                 1.00    -0.09
## Objetivo                           -0.01 -0.04                -0.09     1.00
## Fallo.disciplinario                 0.00  0.10                 0.03    -0.15
## Educacion                          -0.21 -0.22                -0.07     0.10
## Hijos                              -0.05  0.06                 0.03    -0.01
## Bebedor.social                      0.35  0.21                -0.03    -0.10
## Fumador.social                      0.07  0.12                 0.03     0.05
## Mascotas                           -0.44 -0.23                 0.01     0.01
## Peso                                0.46  0.42                -0.04    -0.04
## Altura                             -0.05 -0.06                 0.10     0.09
## IMC                                 0.50  0.47                -0.09    -0.09
## Horas.ausente                       0.02  0.07                 0.02     0.03
##                          Fallo.disciplinario Educacion Hijos Bebedor.social
## Razon.ausencia                         -0.55     -0.05 -0.06           0.07
## Mes                                     0.11     -0.07  0.08           0.06
## Dia.semana                             -0.02      0.06  0.10           0.04
## Estacion                                0.15      0.00  0.05          -0.05
## Coste.transporte                        0.11     -0.06  0.38           0.15
## Distancia.hogar.trabajo                -0.06     -0.26  0.05           0.45
## Tiempo.servicio                         0.00     -0.21 -0.05           0.35
## Edad                                    0.10     -0.22  0.06           0.21
## Carga.trabajo.por.dia                   0.03     -0.07  0.03          -0.03
## Objetivo                               -0.15      0.10 -0.01          -0.10
## Fallo.disciplinario                     1.00     -0.06  0.07           0.05
## Educacion                              -0.06      1.00 -0.19          -0.42
## Hijos                                   0.07     -0.19  1.00           0.21
## Bebedor.social                          0.05     -0.42  0.21           1.00
## Fumador.social                          0.12      0.03  0.16          -0.11
## Mascotas                                0.02     -0.05  0.11          -0.12
## Peso                                    0.07     -0.30 -0.14           0.38
## Altura                                 -0.01      0.10 -0.01           0.17
## IMC                                     0.08     -0.37 -0.14           0.32
## Horas.ausente                          -0.12     -0.05  0.11           0.07
##                          Fumador.social Mascotas  Peso Altura   IMC
```

```
## Razon.ausencia                   -0.12    -0.06  0.00  -0.08  0.04
## Mes                              -0.04     0.05  0.02  -0.07  0.05
## Dia.semana                        0.01    -0.03 -0.13  -0.08 -0.10
## Estacion                         -0.05     0.01 -0.03  -0.03 -0.01
## Coste.transporte                  0.04     0.40 -0.21  -0.19 -0.14
## Distancia.hogar.trabajo          -0.08     0.21 -0.05  -0.35  0.11
## Tiempo.servicio                   0.07    -0.44  0.46  -0.05  0.50
## Edad                              0.12    -0.23  0.42  -0.06  0.47
## Carga.trabajo.por.dia             0.03     0.01 -0.04   0.10 -0.09
## Objetivo                          0.05     0.01 -0.04   0.09 -0.09
## Fallo.disciplinario               0.12     0.02  0.07  -0.01  0.08
## Educacion                         0.03    -0.05 -0.30   0.10 -0.37
## Hijos                             0.16     0.11 -0.14  -0.01 -0.14
## Bebedor.social                   -0.11    -0.12  0.38   0.17  0.32
## Fumador.social                    1.00     0.11 -0.20   0.00 -0.20
## Mascotas                          0.11     1.00 -0.10  -0.10 -0.08
## Peso                             -0.20    -0.10  1.00   0.31  0.90
## Altura                            0.00    -0.10  0.31   1.00 -0.12
## IMC                              -0.20    -0.08  0.90  -0.12  1.00
## Horas.ausente                    -0.01    -0.03  0.02   0.14 -0.05
##                          Horas.ausente
## Razon.ausencia                   -0.17
## Mes                               0.02
## Dia.semana                       -0.12
## Estacion                         -0.01
## Coste.transporte                  0.03
## Distancia.hogar.trabajo          -0.09
## Tiempo.servicio                   0.02
## Edad                              0.07
## Carga.trabajo.por.dia             0.02
## Objetivo                          0.03
## Fallo.disciplinario              -0.12
## Educacion                        -0.05
## Hijos                             0.11
## Bebedor.social                    0.07
## Fumador.social                   -0.01
## Mascotas                         -0.03
## Peso                              0.02
## Altura                            0.14
## IMC                              -0.05
## Horas.ausente                     1.00
```

Se instala el paquete corrplot para representar de forma gráfica la correlación entre los atributos del dataset:

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

corrplot(absentismo.cor, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Con esta representación se contesta a una de las preguntas planteadas para el análisis: comprobar si existe relación entre las características físicas y circunstanciales de los empleados y las ausencias. Las razones y las horas de la ausencia, en este contexto, no parecen ser factores que se vean afectados en gran medida por ninguna característica de las recopiladas en el dataset. De todas maneras sí puede verse relación entre el coste del transporte y el número de hijos y mascotas, entre el IMC del trabajador con su edad y el indicador de bebedor, y éste a su vez con el tiempo de servicio y la distancia del hogar al trabajo.

Así, no se ha podido averiguar si algunas de las características mostradas afecta más o menos a las ausencias, pero sí se han descubierto otras relaciones a las que habría que prestar atención si se muestra interés por la salud de los trabajadores, como la relación entre el IMC con la edad y el alcoholismo.

Modelo supervisado: árbol de decisión

A través de la matriz de correlación no ha sido posible ver las relaciones entre las ausencias y las diferentes características, por lo que se va a crear un árbol de decisión para intentar observar qué atributos son los que más se tienen en cuenta. Para ello, se va a crear una nueva variable de clasificación de horas de ausencia "Nivel.ausencia", que etiqueta las ausencias como 1, considerándose como "Corta/Nula" y que abarca entre 0 y 2 horas de ausencia; como 2, considerándose como "Media" y que abarca entre 2 horas y una jornada de 7 horas, y finalmente más de 7 horas se considera como una ausencia de larga duración, casi al nivel de una baja laboral:

```
absentismo.mod$Nivel.ausencia <- cut(absentismo.mod$Horas.ausente,
                                     breaks = c(-1,2,7,120), labels = c(1, 2, 3))

absentismo.mod$Nivel.ausencia <- as.factor(absentismo.mod$Nivel.ausencia)

summary(absentismo.mod$Nivel.ausencia)
```

```
##   1   2   3
## 289 180 271
```

Para crear el árbol de decisión, primero se divide el dataset en las columnas independientes y el objetivo, en este caso el atributo Nivel.ausencia. Se han escogido la razón de la ausencia, el mes, el día de la semana, el coste de transporte, la distancia del hogar al trabajo, el tiempo de servicio, la carga de trabajo por día, si es bebedor social, el IMC y la edad como los atributos que más podrían afectar a la duración de las ausencias:

```
set.seed(666)
y <- absentismo.mod[,c("Nivel.ausencia")]
X <- absentismo.mod[,c("Razon.ausencia", "Mes", "Dia.semana",
                       "Coste.transporte", "Distancia.hogar.trabajo",
                       "Tiempo.servicio", "Carga.trabajo.por.dia",
                       "Bebedor.social", "IMC", "Edad")]
```

Con el dataset dividido entre las variables que van a realizar la predicción y la variable objetivo, se dividen ambas muestras en training y test con una proporción del 75% para training y un 25% para test. De esta manera, el árbol será creado con el 75% de los datos y se probará con el 25% restante.

```
split_prop <- 3
max_split<-floor(nrow(X)/split_prop)
tr_limit <- nrow(X)-max_split
ts_limit <- nrow(X)-max_split+1

trainX <- X[1:tr_limit,]
trainy <- y[1:tr_limit]
testX <- X[ts_limit+1:nrow(X),]
testy <- y[ts_limit+1:nrow(X)]
```

Y se ejecuta un summary de las muestras para comprobar si son más o menos parecidas:

```
summary(trainX);
```

```
##  Razon.ausencia       Mes           Dia.semana    Coste.transporte
##  Min.   : 0.00   Min.   : 1.000   Min.   :2.000   Min.   :118.0
##  1st Qu.:13.00   1st Qu.: 3.000   1st Qu.:3.000   1st Qu.:179.0
##  Median :23.00   Median : 7.000   Median :4.000   Median :225.0
##  Mean   :19.19   Mean   : 6.587   Mean   :3.889   Mean   :223.7
##  3rd Qu.:26.00   3rd Qu.: 9.000   3rd Qu.:5.000   3rd Qu.:260.0
##  Max.   :28.00   Max.   :12.000   Max.   :6.000   Max.   :388.0
##  Distancia.hogar.trabajo Tiempo.servicio Carga.trabajo.por.dia Bebedor.social
##  Min.   : 5.00           Min.   : 3.00   Min.   :3.432         Min.   :0.0000
##  1st Qu.:16.00           1st Qu.:10.00   1st Qu.:4.073         1st Qu.:0.0000
##  Median :26.00           Median :13.00   Median :4.417         Median :1.0000
##  Mean   :30.03           Mean   :12.64   Mean   :4.617         Mean   :0.6174
##  3rd Qu.:50.00           3rd Qu.:16.00   3rd Qu.:5.106         3rd Qu.:1.0000
##  Max.   :52.00           Max.   :29.00   Max.   :6.315         Max.   :1.0000
##       IMC            Edad
##  Min.   :19.0   Min.   :27.00
##  1st Qu.:24.0   1st Qu.:33.00
##  Median :25.0   Median :37.00
##  Mean   :26.9   Mean   :36.64
##  3rd Qu.:31.0   3rd Qu.:40.00
##  Max.   :38.0   Max.   :58.00
```

```
summary(trainy)
```

```
##   1   2   3
## 183 118 193
```

```
summary(testX)
```

```
##  Razon.ausencia       Mes           Dia.semana    Coste.transporte
##  Min.   : 0.00   Min.   : 0.000   Min.   :2.000   Min.   :118.0
##  1st Qu.:13.00   1st Qu.: 3.000   1st Qu.:3.000   1st Qu.:179.0
##  Median :23.00   Median : 5.000   Median :4.000   Median :225.0
##  Mean   :19.23   Mean   : 5.788   Mean   :3.975   Mean   :216.4
##  3rd Qu.:27.00   3rd Qu.:10.000   3rd Qu.:5.000   3rd Qu.:235.0
##  Max.   :28.00   Max.   :12.000   Max.   :6.000   Max.   :378.0
##  NA's   :495     NA's   :495      NA's   :495     NA's   :495
##  Distancia.hogar.trabajo Tiempo.servicio Carga.trabajo.por.dia Bebedor.social
##  Min.   :10.00           Min.   : 1.00   Min.   :3.703         Min.   :0.0000
##  1st Qu.:16.00           1st Qu.: 9.00   1st Qu.:3.961         1st Qu.:0.0000
##  Median :26.00           Median :12.00   Median :4.404         Median :0.0000
##  Mean   :28.74           Mean   :12.38   Mean   :4.341         Mean   :0.4653
##  3rd Qu.:45.00           3rd Qu.:17.00   3rd Qu.:4.585         3rd Qu.:1.0000
##  Max.   :52.00           Max.   :29.00   Max.   :5.226         Max.   :1.0000
##  NA's   :495             NA's   :495     NA's   :495           NA's   :495
##       IMC             Edad
##  Min.   :19.00   Min.   :28.00
##  1st Qu.:24.00   1st Qu.:30.00
##  Median :25.00   Median :37.00
##  Mean   :26.24   Mean   :36.07
##  3rd Qu.:31.00   3rd Qu.:40.00
##  Max.   :38.00   Max.   :58.00
##  NA's   :495     NA's   :495
```

```
summary(testy)
```

```
##    1    2    3 NA's
##  106   61   78  495
```

Como se puede ver, los datos son muy similares entre las muestras, por lo que podemos asumir que están bien distribuídos.

Para poder crear el árbol de decisión, es necesario importar el paquete C50:

```
if(!require(C50)){
    install.packages('C50', repos='http://cran.us.r-project.org')
    library(C50)
}
```

Con las muestras de train y test listas, se procede a la creación del modelo de árbol de decisión:

```
model <- C50::C5.0(trainX, trainy, rules=TRUE )
summary(model)
```

```
##
## Call:
## C5.0.default(x = trainX, y = trainy, rules = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Mon Dec 27 18:31:06 2021
## -------------------------------
##
```

```
## Class specified by attribute `outcome'
##
## Read 494 cases (11 attributes) from undefined.data
##
## Rules:
##
## Rule 1: (32, lift 2.6)
##  Razon.ausencia <= 0
##  ->  class 1  [0.971]
##
## Rule 2: (49/15, lift 1.9)
##  Razon.ausencia > 22
##  Razon.ausencia <= 23
##  Edad <= 34
##  ->  class 1  [0.686]
##
## Rule 3: (267/132, lift 1.4)
##  Razon.ausencia > 22
##  ->  class 1  [0.506]
##
## Rule 4: (19/2, lift 3.6)
##  Razon.ausencia > 22
##  Razon.ausencia <= 27
##  Coste.transporte <= 189
##  Tiempo.servicio > 12
##  Carga.trabajo.por.dia > 3.992567
##  Carga.trabajo.por.dia <= 4.3551
##  IMC > 24
##  Edad > 34
##  Edad <= 43
##  ->  class 2  [0.857]
##
## Rule 5: (25/4, lift 3.4)
##  Razon.ausencia > 22
##  Edad > 34
##  Edad <= 36
##  ->  class 2  [0.815]
##
## Rule 6: (34/6, lift 3.4)
##  Razon.ausencia > 22
##  Mes > 5
##  Tiempo.servicio <= 12
##  Carga.trabajo.por.dia > 3.943817
##  Edad > 34
##  ->  class 2  [0.806]
##
## Rule 7: (3, lift 2.0)
##  Coste.transporte <= 118
##  Carga.trabajo.por.dia <= 3.943817
##  Edad <= 43
##  ->  class 3  [0.800]
##
## Rule 8: (8/1, lift 2.0)
##  Razon.ausencia > 24
```

```
##  IMC > 32
##  ->  class 3  [0.800]
##
## Rule 9: (195/40, lift 2.0)
##  Razon.ausencia > 0
##  Razon.ausencia <= 22
##  ->  class 3  [0.792]
##
## Rule 10: (55/16, lift 1.8)
##  Coste.transporte > 248
##  Distancia.hogar.trabajo <= 42
##  Edad <= 34
##  ->  class 3  [0.702]
##
## Default class: 3
##
##
## Evaluation on training data (494 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##      10  105(21.3%)   <<
##
##
##     (a)   (b)   (c)    <-classified as
##    ----  ----  ----
##     159     7    17    (a): class 1
##      36    56    26    (b): class 2
##      15     4   174    (c): class 3
##
##
##  Attribute usage:
##
##  100.00% Razon.ausencia
##   33.81% Edad
##   15.59% Coste.transporte
##   11.34% Carga.trabajo.por.dia
##   11.13% Distancia.hogar.trabajo
##   10.73% Tiempo.servicio
##    6.88% Mes
##    5.47% IMC
##
##
## Time: 0.0 secs
```

Una vez obtenidas las reglas, se comentan en detalle:

- Regla 1: las ausencias por razones desconocidas serán cortas o sin ausencia con un 97.1%.
- Regla 2: las ausencias por asistencias a consulta (23) en trabajadores con una edad de 34 años o inferior serán cortas o sin ausencia conun 68.6%.
- Regla 3: aquellas ausencias por asistencia a consulta, al dentista, al fisioterapeuta, al laboratorio, por donación de sangre o que no estén justificadas, serán cortas con una confianza del 50.6%.

- Regla 4: con un 85.7% de confianza, las ausencias por asistencias a consulta, fisioterapia, laboratorio, donación de sangre o que no están justificadas, de trabajadores cuyo coste de transporte es igual o superior a 189, un tiempo de servicio mayor que 12, una carga de trabajo por día entre 3.99 y 4.35 horas, un IMC considerado obesidad y una edad entre 34 y 43 años, serán de duración próxima a una jornada laboral.
- Regla 5: con un 81.5% de confianza, las ausencias por asistencias a consulta, dentista, fisioterapia, donaciones de sangre, a laboratorio o no justificadas y una edad de 35 o 36 años serán de duración próxima a una jornada laboral.
- Regla 6: las ausencias por las razones nombradas en la anterior regla, en los meses de junio a diciembre, con un tiempo de servicio menor o igual a 12, una carga de trabajo por dia superior a 3.94 horas y una edad superior a 34 años serán de duración media con un 80.6% de seguridad.
- Regla 7: si el coste de transporte de un trabajador que se ausenta es igual o menor que 118, la carga de trabajo es inferior o igual a 3.94 y su edad es igual o inferior a 43 años, su ausencia será de larga duración con un 80% de confianza.
- Regla 8: si la razón de ausencia es por asistencia al dentista, fisioterapeuta, laboratorio o no está justificada y el IMC es superior a 32, la duración será de larga duración con un 80% de confianza.
- Regla 9: si la razón de la ausencia no es una de las nombradas en la regla 6, la ausencia será de larga duración con un 79.2% de seguridad.
- Regla 10: si el coste de transporte es superior a 248, la distancia del hogar al trabajo es inferior o igual a 42 y la edad es inferior o igual a 34, la duración será de larga duración con un 70.2% de confianza.

Tras echar un vistazo a las reglas queda claro que la razón de la ausencia es un factor determinante para saber si la duración de la ausencia va a ser corta, media o larga. Por lo general, las razones que más se tienen en cuenta son las asistencias a la consulta, al dentista, al laboratorio, a sesiones de fisioterapia, a donaciones de sangre o simplemente no está justificada. Por otra parte, la edad también influye la cantidad de horas de ausencia, cuanto menor es la edad, mayor es la duración de la ausencia. El IMC también ejerce cierta influencia, corroborando que a IMCs altos mayor es el número de horas ausente.

Una vez se ha creado el modelo del árbol de decisión, es buen momento para medir el porcentaje de aciertos de éste:

```
predicted_model <- predict( model, testX, type="class" )
mat_conf<-table(testy,Predicted=predicted_model)

porcentaje_correct<-100 * sum(diag(mat_conf)) / sum(mat_conf)
print(sprintf("El %% de registros correctamente clasificados es: %.4f %%",
              porcentaje_correct))
```

```
## [1] "El % de registros correctamente clasificados es: 60.4082 %"
```

Se obtiene un % de aciertos del 60.40%, al límite de lo que se podría considerar aceptable (>60%).

Adicionalmente, se calcula la matriz de confusión para observar de manera más detallada el rendimiento del modelo. Para ello se usa el método CrossTable del paquete gmodels:

```
if(!require(gmodels)){
    install.packages('gmodels', repos='http://cran.us.r-project.org')
    library(gmodels)
}
```

Ahora, se crea la matriz de confusión:

```
CrossTable(testy, predicted_model,prop.chisq  = FALSE, prop.c = FALSE,
           prop.r =FALSE,dnn = c('Reality', 'Prediction'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  245
##
##
##               | Prediction
##      Reality |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##            1 |        76 |         9 |        21 |       106 |
##              |     0.310 |     0.037 |     0.086 |           |
## -------------|-----------|-----------|-----------|-----------|
##            2 |        35 |         4 |        22 |        61 |
##              |     0.143 |     0.016 |     0.090 |           |
## -------------|-----------|-----------|-----------|-----------|
##            3 |         8 |         2 |        68 |        78 |
##              |     0.033 |     0.008 |     0.278 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |       119 |        15 |       111 |       245 |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

Se puede observar que el 31% de las muestras han sido clasificadas con éxito como duración normal (1), el 1.6% como duración media (2) y el 27.8% como de larga duración. Por otra parte, se observa que la clasificación que mejor ha predecido ha sido la correspondiente a 3, habiendo predecido solamente un 3% como 1 y un 0.8% como 2. El árbol también ha tenido un rendimiento bueno clasificando la etiqueta 1, con un 3% de los casos clasificados como 2 y un 8% como 3. En cambio, para predecir los casos de duración media (2) ha tenido más dificultades, con un 14.3% de los casos clasificados como 1 y un 9% como 3.

Las columnas que se han escogido para formar parte del modelo han sido cuidadosamente seleccionadas, combinadas y probadas para obtener el mayor porcentaje de aciertos, por lo que una solución general para mejorar el rendimiento del árbol sería aumentar el número de muestras totales, sobre todo aquellas con duraciones de ausencia medias, ya que se observó al principio de este apartado que ésta era la que menos filas tenía.

Creación de un modelo no supervisado

Como modelo no supervisado se decide usar el algoritmo de kmeans para agrupar los registros de absentismo para formar poblaciones que permitan identificar a sus miembros por la similitud de sus características.

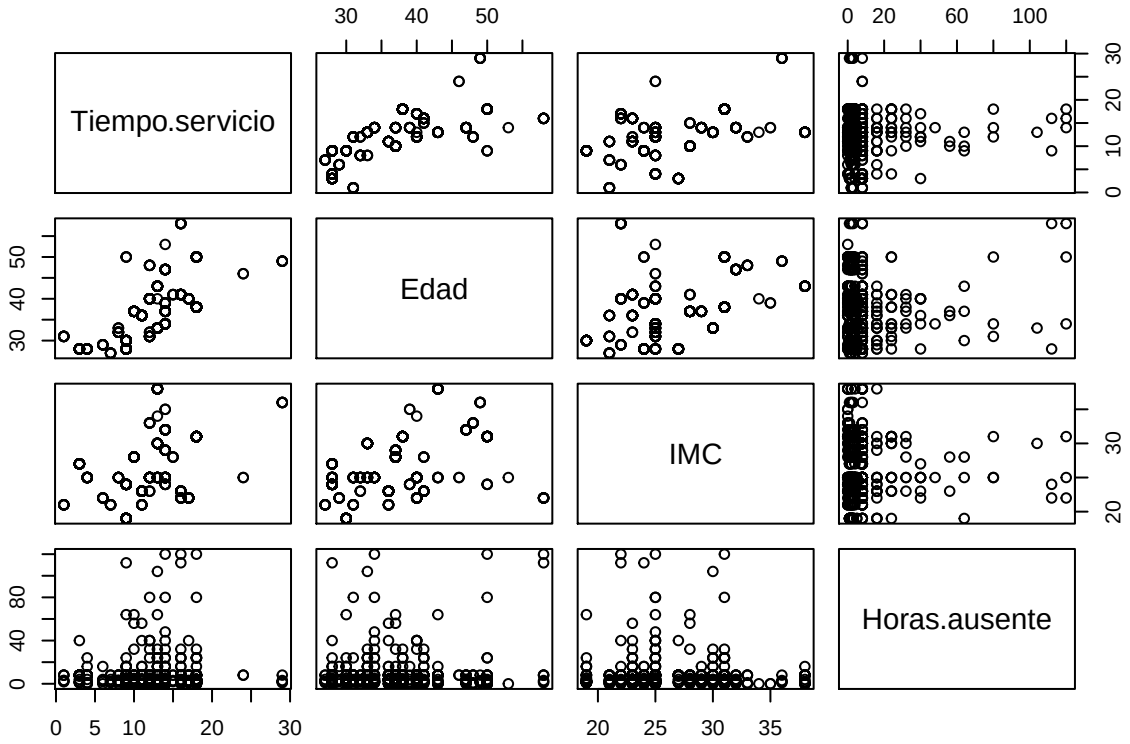
Primeramente, se importa la librería de "cluster":

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

Antes de aplicar el algoritmo de kmeans sería interesante observar cómo se agrupan los datos de clientes:

```
absentismo.clustering <- absentismo.mod[,c("Tiempo.servicio", "Edad", "IMC",
                                           "Horas.ausente")]

plot(absentismo.clustering)
```



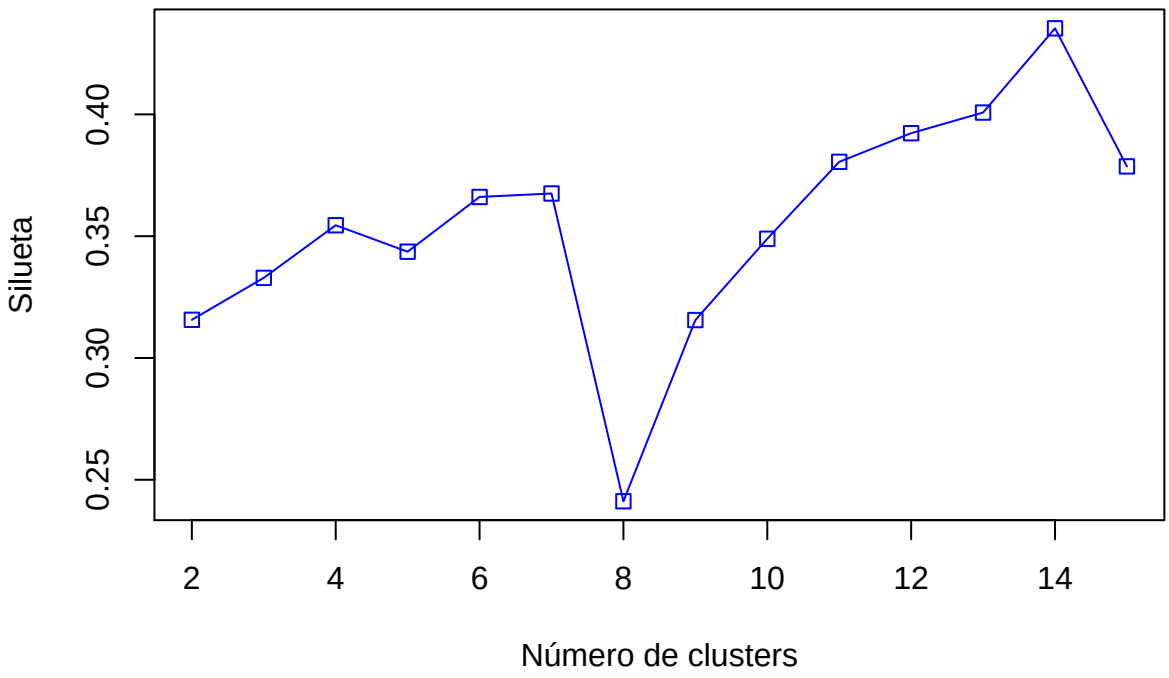
A primera vista, no parece que haya muchos grupos ya que los datos están bastante agrupados. Se podrían vislumbrar dos grupos echando un vistazo a la relación entre los diferentes atributos y Horas.ausente, correspondiéndose el de la izquierda a los trabajadores que tienen pocas horas de ausencia y el otro a los que tienen bastantes más.

Para corroborar esta hipótesis, se ejecuta kmeans con diferentes números de agrupaciones para ver cuál puede ser la mejor:

```
set.seed(1234)
d <- daisy(absentismo.clustering)
resultados <- rep(0, 15)
for (i in c(2,3,4,5,6,7,8,9,10,11,12,13,14,15))
{
  fit           <- kmeans(absentismo.clustering, i)
  y_cluster     <- fit$cluster
  sk            <- silhouette(y_cluster, d)
  resultados[i] <- mean(sk[,3])
}
```

A continuación, se representa el array de resultados, que contiene las medidas para cada implementación de kmeans. Cuanto mayor sea la medida quiere decir que el número de clústers seleccionados para esa implementación es mejor:

```
plot(2:15,resultados[2:15],type="o",col="blue",pch=0,xlab="Número de clusters",
     ylab="Silueta")
```




```
which(resultados[2:15]==max(resultados[2:15]))+1
```

```
## [1] 14
```

```
resultados[which(resultados[2:15]==max(resultados[2:15]))+1]
```

```
## [1] 0.4353156
```

En este caso, se obtiene que el número óptimo de clústers es 14, muy por encima de lo que se había visto en la primera representación.

Para obtener unos resultados más fiables, se usa la función kmeansruns que ejecuta kmeans, para ello se instala el paquete "fpc".

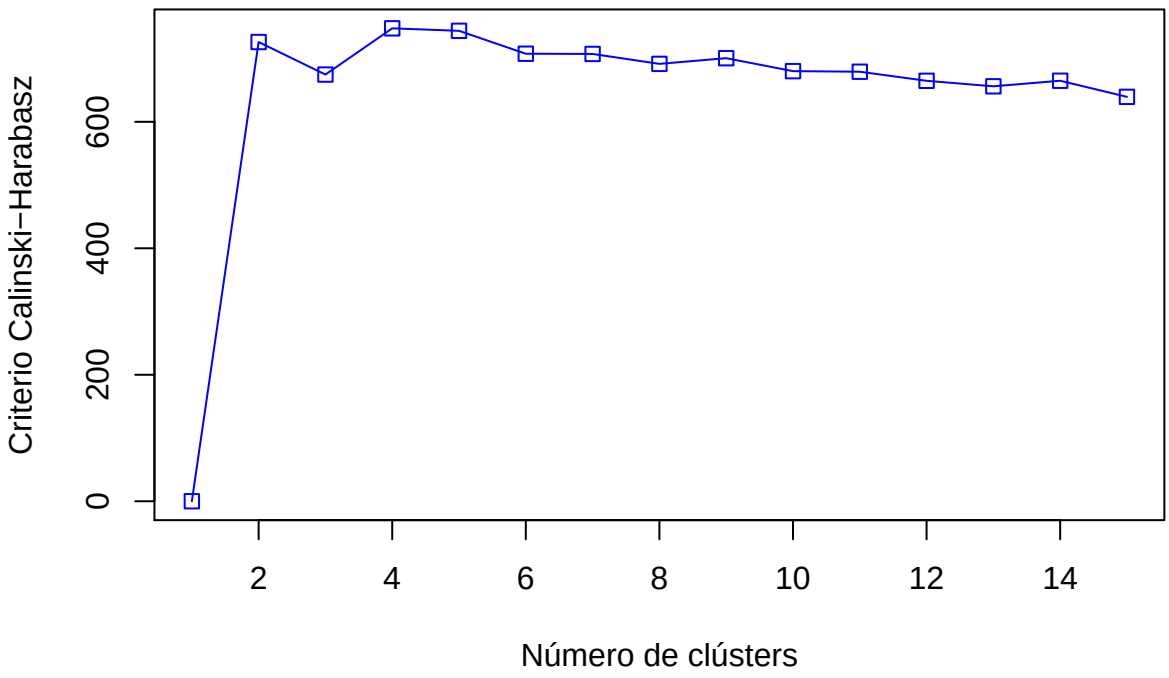
```
if (!require('fpc')) install.packages('fpc'); library('fpc')
```

Usando esta función, se calcula el número de clústers óptimos atendiendo a dos criterios: la silueta media ("asw") y Calinski-Harabasz ("ch").

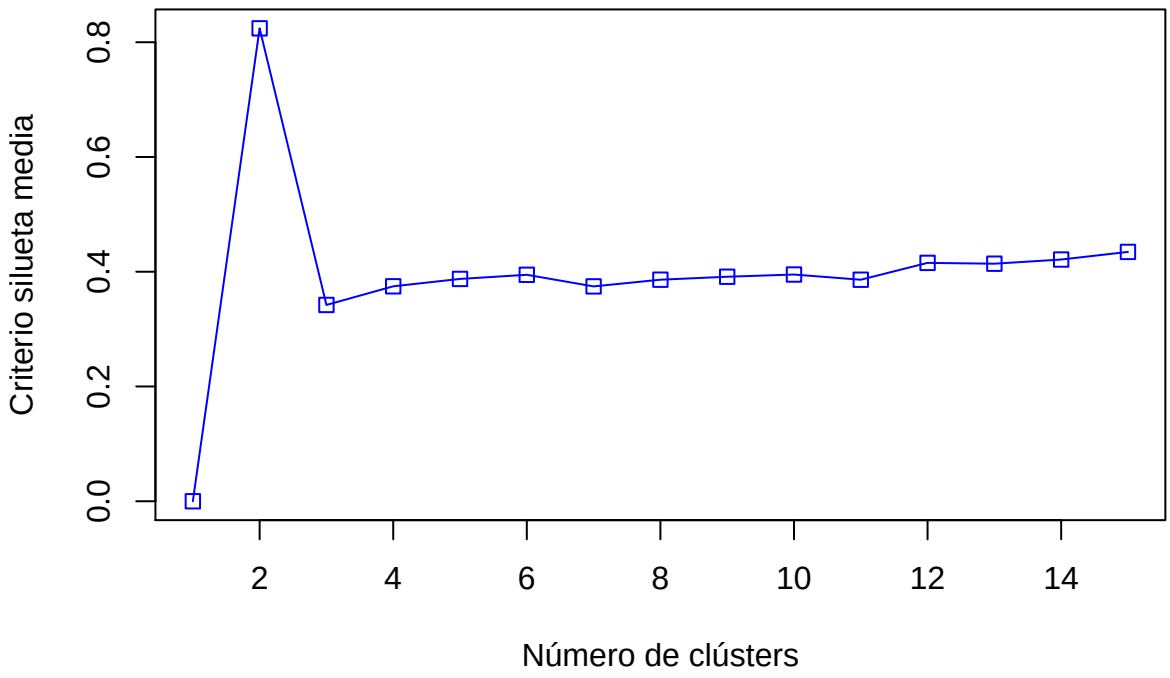
```
fit_ch  <- kmeansruns(absentismo.clustering, krange = 1:15, criterion = "ch")
fit_asw <- kmeansruns(absentismo.clustering, krange = 1:15, criterion = "asw")
```

Y una vez se obtienen los resultados, se procede a comprobar el mejor resultado:

```
plot(1:15,fit_ch$crit,type="o", col="blue", pch=0, xlab="Número de clústers",
     ylab="Criterio Calinski-Harabasz")
```




```
plot(1:15,fit_asw$crit,type="o", col="blue", pch=0, xlab="Número de clústers",
     ylab="Criterio silueta media")
```




```
which(fit_ch$crit==max(fit_ch$crit))
```

```
## [1] 4
```

```
which(fit_asw$crit==max(fit_asw$crit))
```

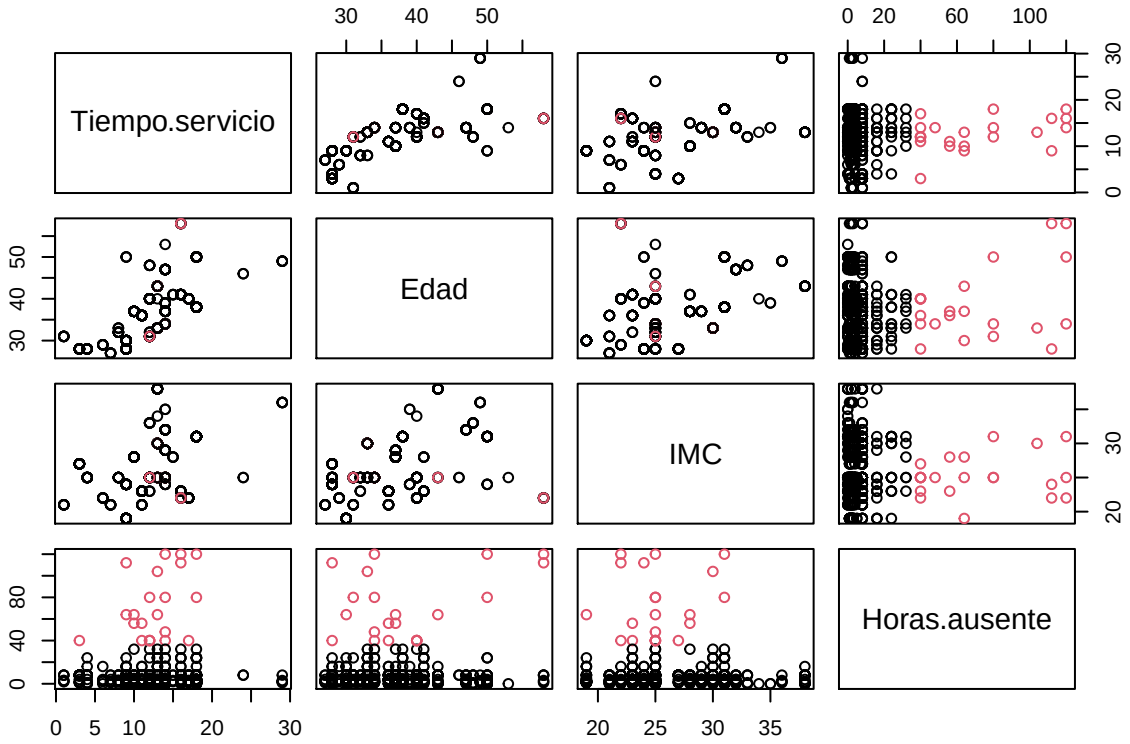
```
## [1] 2
```

Parece que estos dos criterios tienen un pensamiento similar al propuesto en la primera representación, que además parece algo más lógico.

Con estos datos acerca del número de clústers óptimos, se decide aplicar kmeans con 2 y 4 agrupaciones.

Primeramente, se aplica kmeans con 2 clústers:

```
cl2 <- kmeans(absentismo.clustering, 2)
with(absentismo.clustering, pairs(absentismo.clustering, col=c(1:2)[cl2$cluster]))
```

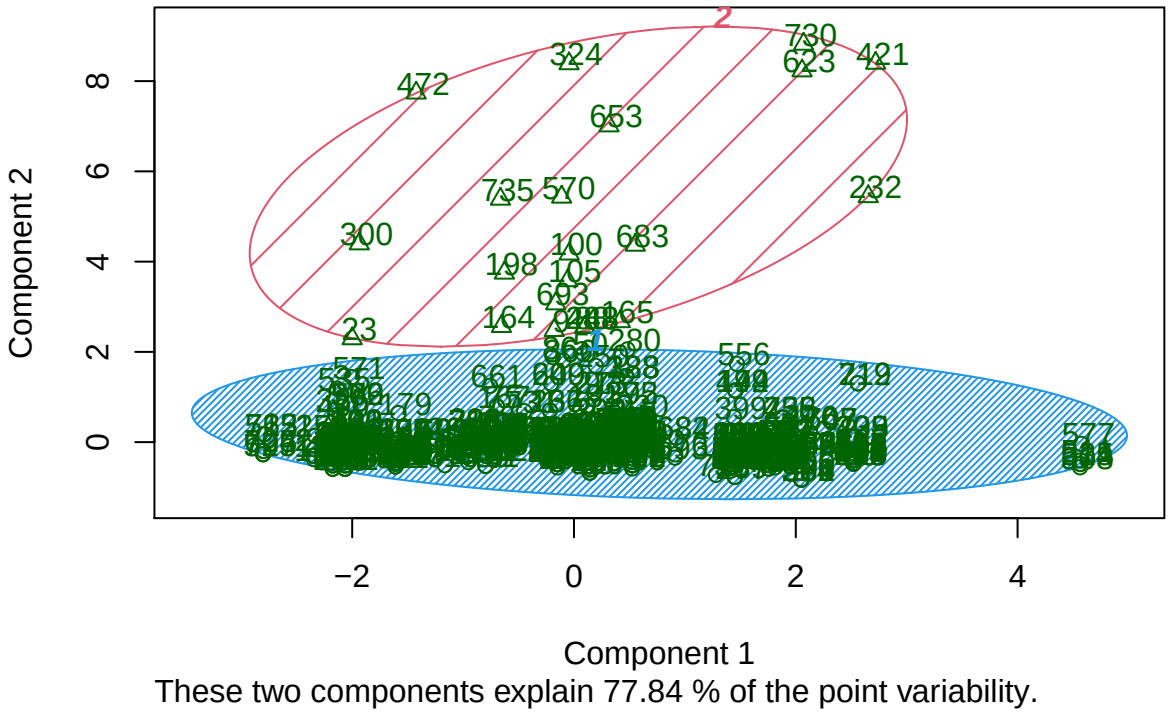


Con la representación del número de grupos, se realizan las siguientes observaciones:

- Grupo 1: a primera vista, el grupo más poblado. Las edades, los tiempos de servicio y el IMC de los miembros de este grupo están bastante dispersos, pero se destaca que, con bastante diferencia, se agrupan en cifras muy bajas de horas de ausencia.
- Grupo 2: es un grupo minoritario cuyos tiempos de servicio son medios, sus IMCs son medios-altos, sus edades dispersas y sus horas de ausencia son medias y altas.

En el siguiente gráfico se pueden ver los clústers formados de una manera bastante clara:

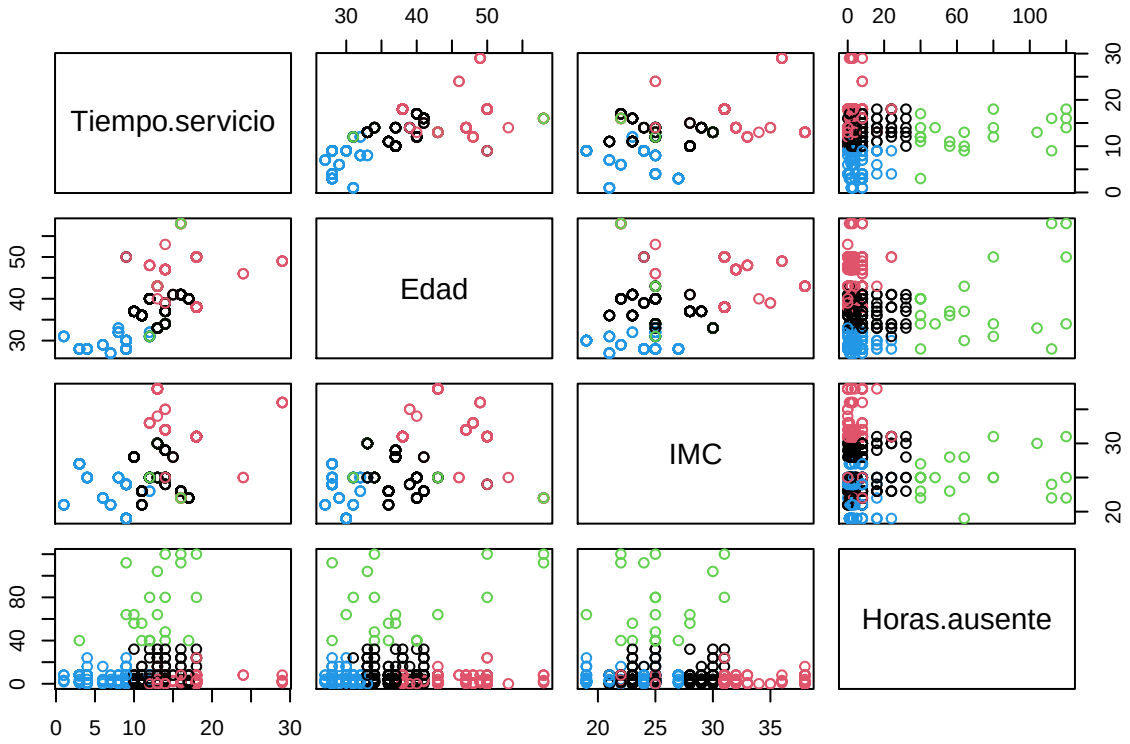
```
clusplot(absentismo.clustering, cl2$cluster, color=TRUE, shade=TRUE, labels=2,
         lines=0, main = "Clústers de trabajadores")
```



En este diagrama, se pueden observar a los dos grupos bien diferenciados. Como se comentaba, uno de ellos contiene más miembros y se extiende a lo largo de la componente 1, pero se mantiene en valores bajos de la componente 2. El otro, en cambio, tiene menos miembros y se mantiene en valores medios de cada componente.

Una vez vista la aplicación de kmeans con 2 clústers, se procede a realizar la implementación de éste con 4:

```
cl4 <- kmeans(absentismo.clustering, 4)
with(absentismo.clustering, pairs(absentismo.clustering, col=c(1:4)[cl4$cluster]))
```

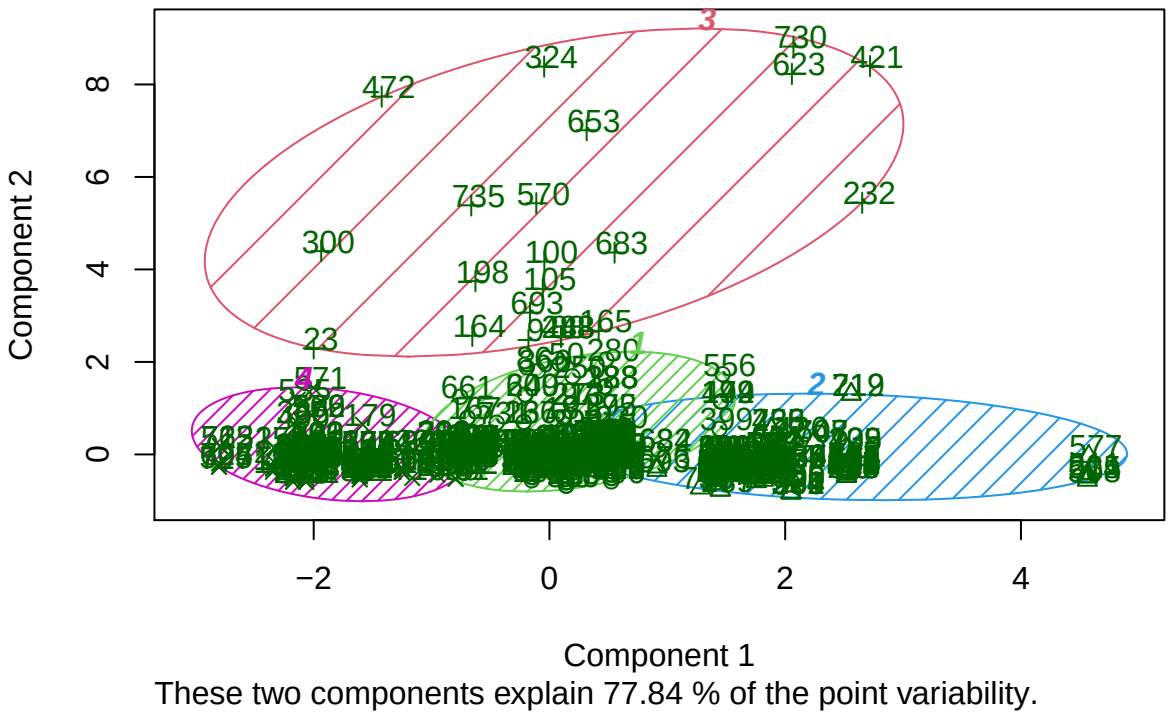


A continuación, se describen las observaciones de los 4 grupos obtenidos en esta última representación:

- Grupo 1: se corresponde a trabajadores jóvenes con un tiempo de servicio bajo y con IMC dentro de lo normal. En cuanto a las horas de ausencia, es un grupo cuyo tiempo ausente es muy bajo. Posiblemente sean trabajadores jóvenes que llevan poco tiempo en la empresa, pudiéndose ver que no es un incentivo para que falten más horas.
- Grupo 2: son trabajadores que tienen un tiempo de servicio y una edad medias, cuyos IMCs están dentro de la normalidad y las horas de ausencia son muy bajas también, aunque muy ligeramente por encima del anterior grupo. Estos trabajadores se corresponden a los que ya tienen experiencia y llevan un tiempo en la empresa.
- Grupo 3: el tiempo de servicio de los trabajadores de este grupo es similar al anterior y, en algunos casos, muy alto. Las edades, de la misma manera, también son más altas, y pocos IMC están dentro de los valores óptimos ya que la mayoría se sitúan en valores altos. Por el contrario, las horas de ausencia de este grupo son igualmente bajas. Probablemente sean los trabajadores más experimentados y que más tiempo llevan en la empresa, pero que no llevan vidas muy saludables.
- Grupo 4: se corresponde al grupo con menos individuos. El tiempo de servicio de éstos es medio, las edades variadas y tienen IMC dentro de la normalidad, pero las horas de ausencia son medias y altas, por encima de los valores de otros trabajadores.

Como se hizo anteriormente, se representan los clústers de manera bastante clara:

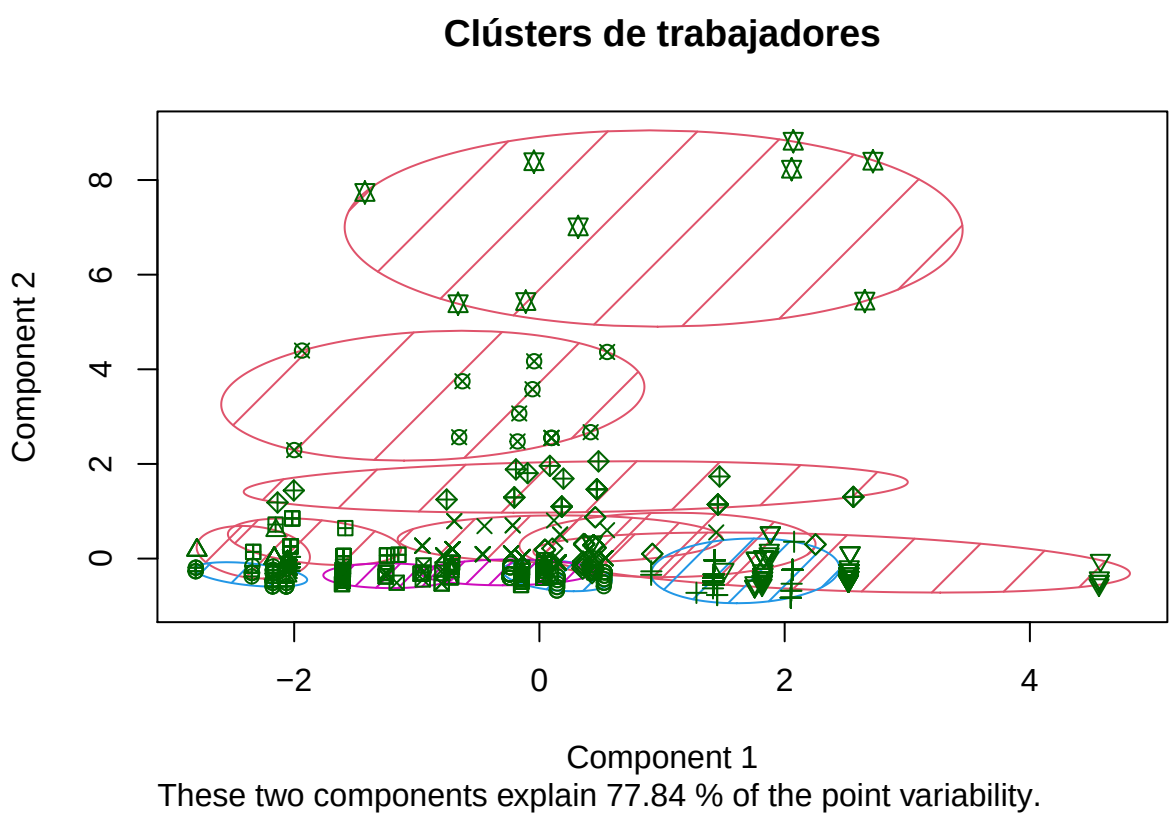
```
clusplot(absentismo.clustering, cl4$cluster, color=TRUE, shade=TRUE, labels=2,
         lines=0, main = "Clústers de trabajadores")
```



Para este caso, se pueden observar los 4 grupos representados. De éstos, 3 de ellos se distribuyen a lo largo de la componente 1, mientras que el cuarto grupo está más centrado y se extiende casi uniformemente.

Por último, simplemente se muestra el clusplot para los 14 clústers que se habían

```
cl14 <- kmeans(absentismo.clustering, 14)
clusplot(absentismo.clustering, cl14$cluster, color=TRUE, shade=TRUE,
         lines=0, main = "Clústers de trabajadores")
```



Para este último caso, se observa que ha dividido incluso más el grupo inferior e incluso ha dividido el grupo superior en dos grupos.

Se puede observar que las dos primeras representaciones de clustering son similares: la primera tiene 2 agrupaciones y, aunque la segunda tenga 4, se vislumbra que es el grupo poblado del primer clustering el que se ha dividido para formar 3 grupos. Las diferencias principales son que el primer clustering hacía más diferencia entre los trabajadores que se ausentaban más horas y los que menos, mientras que el segundo se ha visto que tiene más en cuenta el tiempo de servicio, las edades y los IMC, además de las horas ausentes.

Finalmente, se crea el dataset final. Se incluyen en él algunas de las nuevas columnas creadas, como Razon.ausencia.def y se quitan otras por no considerarse útiles para las siguientes fases, como Objetivo o Fallo.disciplinario, que no se han usado. A continuación, se muestra el dataset final con las columnas ordenadas:

```
absentismo.final <- absentismo.mod[,c("ID", "Razon.ausencia", "Razon.ausencia.def",
                                      "Mes.def", "Mes", "Dia.semana.def", "Dia.semana",
                                      "Estacion.def", "Coste.transporte", "Distancia.hogar.trabajo",
                                      "Tiempo.servicio", "Carga.trabajo.por.dia", "Educacion",
                                      "Educacion.def", "Hijos", "Tiene.hijos", "Bebedor.social",
                                      "Fumador.social", "Mascotas", "Tiene.mascotas", "IMC",
                                      "Horas.ausente", "Edad", "Rango.IMC")]

write.csv(absentismo.final, "Absenteeism_at_work_final.csv")
```

Conclusiones

A través de todo el informe visto, se han podido ver y contestar ciertos aspectos planteados al principio del mismo. Primeramente, se quería averiguar si había algún tipo de correlación entre las características físicas, circunstanciales y laborales del trabajador, entre ellas mismas y con el factor de las horas de ausencia y/o la razón de la misma. A primera vista, con este análisis no se ha visto una correlación fuerte entre dichas características, las horas de ausencia y la razón, pero sí se han descubierto otras relaciones que pueden ser muy importantes para garantizar la buena salud y la estabilidad del trabajador. Por una parte se ha visto que hay relación entre el coste del trasporte del trabajador y el número de hijos y mascotas, por lo que la empresa podría ofrecer bonos de transporte según estos factores. Por otra parte, existe una relación entre el IMC del trabajador con su edad y el indicador de si es bebedor o no, relacionado a su vez este último con el tiempo de servicio y la distancia del hogar al trabajo. Tal vez, es posible que en la empresa haya un ambiente que queme mentalmente a la gente poco a poco, descuidando su salud (valores de IMC altos) o recurriendo a la bebida (no necesariamente considerándolo alcoholismo). Por ello, sería interesante que la empresa estudiase otros factores que ahora mismo no están en este dataset y ponga atención a estos factores para evitar problemas de salud.

Con un estudio más a fondo, que ha incluido un método supervisado de clasificación como es un árbol de decisión, se ha podido ver que la razón de la ausencia es un factor que afecta en gran medida a la duración de la ausencia, es decir, si va a ser una ausencia corta (menos de dos horas), media (entre 2 y 7 horas) o larga (más de 7 horas). También se ha podido ver que las razones más comunes son las visitas a la consulta, al dentista, por una cita para una muestra biológica, fisioterapia o análisis de sangre. En menor medida, la edad y el IMC son otros factores que también influyen en la magnitud de la ausencia: a mayor edad e IMC, mayor es la ausencia. Además, aparte de esta información que el árbol ha proporcionado, también se ha obtenido la capacidad de poder ver si un trabajador se va a ausentar poco o mucho según sus características. Globalmente, la eficacia del árbol ha alcanzado el 60.4%, una cifra aceptable.

La necesidad de agrupar las ausencias por características similares se ha logrado tras aplicar el modelo no supervisado de clustering, usando kmeans. En el estudio se han obtenido 2 y 4 grupos, pero preferiblemente se pone el foco en los 4 grupos, ya que ofrecen más variedad. Estos grupos han atendido a la edad, el tiempo de servicio y el IMC por considerarse características físicas diferenciadoras. Esto ha permitido formar un grupo de trabajadores jóvenes y con tiempos de servicio bajos, aparentemente motivados, ya que han empezado hace poco y sus horas de ausencia son bajas; el segundo grupo recoge a trabajadores con edades y antigüedades medias, que también han visto aumentadas sus horas de ausencia; el tercero es similar al segundo, pero los IMCs de los trabajadores son altos, pudiendo ver que no llevan vidas muy saludables; y el cuarto, el menos poblado, el más variado y con horas de ausencia altas. Así, se tiene una visión de los distintos perfiles de trabajadores y ausencias hay, siendo útil para predecir futuras magnitudes de ausencia.

```
table_list<-data.frame(Contribuciones = character(0), Firma = character(0))

table_list[nrow(table_list) + 1,] = c("Investigación previa", "Brais Rodríguez Martínez")

table_list[nrow(table_list) + 1,] = c("Redacción de las respuestas", "Brais Rodríguez Martínez")

table_list[nrow(table_list) + 1,] = c("Creación del código", "Brais Rodríguez Martínez")

knitr::kable(table_list, format = "simple")
```

Contribuciones	Firma
Investigación previa	Brais Rodríguez Martínez
Redacción de las respuestas	Brais Rodríguez Martínez
Creación del código	Brais Rodríguez Martínez