Preprocesamiento

JMason88

7 de octubre de 2018

# Limpio Memoria

rm(list = ls())

# Llamo librerías y llamo script de ‘Mapeo de Hyades’ para traer el dataset final.

library(tidyr)  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v readr 1.1.1 v stringr 1.3.1  
## v ggplot2 3.0.0 v forcats 0.3.0

## -- Conflicts ---------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(readxl)  
library(readr)  
source("mapeo\_hyades.r")  
  
hip <- mapeo\_hyades()

## Parsed with column specification:  
## cols(  
## identifier = col\_character(),  
## cercanaHip = col\_integer(),  
## idcruzHip = col\_integer(),  
## dist = col\_double(),  
## dif\_ra = col\_double(),  
## dif\_de = col\_double()  
## )

## Preprocesamiento del catálogo Hipparcos

Analizo Valores faltantes

sapply(hip, function(x) sum(is.na(x)))

## HIP RA\_J2000 DE\_J2000 Plx pmRA   
## 0 0 0 0 0   
## pmDE Vmag B-V Symbad\_Hyades   
## 0 0 15 0

Existen 15 valores faltantes para la variable ‘B-V’. A continuación se procederá a filtrar el dataframe por esas instancias que presentan valores faltantes.

# Filtro las filas por los valores faltantes

hip[is.na(hip$`B-V`),]

## # A tibble: 15 x 9  
## HIP RA\_J2000 DE\_J2000 Plx pmRA pmDE Vmag `B-V` Symbad\_Hyades  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl>   
## 1 14807 3.19 22.4 19.4 59.2 -130. 10.6 NA FALSE   
## 2 15453 3.32 11.5 16.4 208. 55.3 11.2 NA FALSE   
## 3 15662 3.36 13.4 7.1 71.5 -156. 11.1 NA FALSE   
## 4 15854 3.40 17.5 5.12 14.3 -53.0 8.1 NA FALSE   
## 5 16150 3.47 17.2 13.4 18.7 -7 11.8 NA FALSE   
## 6 17291 3.70 17.0 0.39 9.41 1.01 8.91 NA FALSE   
## 7 18531 3.96 22.9 27.7 157. 11 8.78 NA FALSE   
## 8 19464 4.17 24.1 7.87 -6.84 -15.4 9.55 NA FALSE   
## 9 20472 4.39 11.4 13.2 25.8 -35.1 6.64 NA FALSE   
## 10 21078 4.52 6.79 12.1 -67.2 -90.7 6.78 NA FALSE   
## 11 21334 4.58 15.3 6.01 17.5 1.7 12.1 NA FALSE   
## 12 22266 4.79 13.1 6.12 -3.72 -36.9 12.0 NA FALSE   
## 13 23155 4.98 10.6 18.3 55.2 -153. 11.0 NA FALSE   
## 14 24875 5.33 24.4 2.87 1.24 -7.71 9.64 NA FALSE   
## 15 27465 5.82 19.1 3.32 3.57 1.17 10.3 NA FALSE

En analisis de cluster, un método usualmente aceptado es remover instancias que posean faltantes de la clusterización inicial. Luego con clusters bien definidos, se suele intentar mapear las instancias con faltantes a los clusters más parecidos.

Con esta posibilidad en mente, removeremos 15 estrellas del dataframe para trabajar con un conjunto de datos completos a la hora de clusterizar.

hip\_sin\_faltantes <- hip[!is.na(hip$`B-V`),]  
  
sapply(hip\_sin\_faltantes, function(x) sum(is.na(x)))

## HIP RA\_J2000 DE\_J2000 Plx pmRA   
## 0 0 0 0 0   
## pmDE Vmag B-V Symbad\_Hyades   
## 0 0 0 0

A continuación se procederá a escalar y estandarizar las variables para evitar que las dimensiones de estas incidan o afecten las medidas de distancia dentro del análisis de cluster.

Para las variables RA\_J2000 y DE\_J2000, se realizará un escalado especial que permita conservar las proporciones entre ambas variables, ya que consideramos que una “estandarización” para estas dos podría resultar en una posible distorsión de la medición que se posee de la proyección de la estrella en el cielo.

max\_RA <- max(hip\_sin\_faltantes$RA\_J2000)  
min\_RA <- min(hip\_sin\_faltantes$RA\_J2000)  
max\_DE <- max(hip\_sin\_faltantes$DE\_J2000)  
min\_DE <- min(hip\_sin\_faltantes$DE\_J2000)  
  
max\_RA

## [1] 6.099316

min\_RA

## [1] 2.900032

max\_DE

## [1] 24.49571

min\_DE

## [1] 3.409683

hip\_version1 <- hip\_sin\_faltantes %>%   
 mutate(RA\_J2000 = (max\_RA-RA\_J2000)/(max\_RA-min\_RA)) %>%  
 mutate(DE\_J2000 = (max\_DE-DE\_J2000)/(max\_DE-min\_DE)) %>%  
 mutate(Plx = scale(.$Plx)) %>%  
 mutate(pmRA = scale(.$pmRA)) %>%  
 mutate(pmDE = scale(.$pmDE)) %>%  
 mutate(Vmag = scale(.$Vmag)) %>%  
 mutate('B-V' = scale(.$`B-V`))  
  
summary(hip\_version1)

## HIP RA\_J2000 DE\_J2000 Plx.V1   
## Min. :13510 Min. :0.0000 Min. :0.0000 Min. :-0.872343   
## 1st Qu.:17637 1st Qu.:0.2363 1st Qu.:0.2497 1st Qu.:-0.516191   
## Median :21093 Median :0.4936 Median :0.5035 Median :-0.276607   
## Mean :21220 Mean :0.4877 Mean :0.5032 Mean : 0.000000   
## 3rd Qu.:24943 3rd Qu.:0.7253 3rd Qu.:0.7585 3rd Qu.: 0.116004   
## Max. :28887 Max. :1.0000 Max. :1.0000 Max. :16.265614   
## pmRA.V1 pmDE.V1 Vmag.V1   
## Min. :-4.724499 Min. :-21.281189 Min. :-5.252424   
## 1st Qu.:-0.358126 1st Qu.: -0.025214 1st Qu.:-0.566073   
## Median :-0.236859 Median : 0.223143 Median : 0.053885   
## Mean : 0.000000 Mean : 0.000000 Mean : 0.000000   
## 3rd Qu.: 0.029594 3rd Qu.: 0.343542 3rd Qu.: 0.627180   
## Max. :26.449728 Max. : 3.696902 Max. : 2.887028   
## B-V.V1 Symbad\_Hyades   
## Min. :-1.861028 Mode :logical   
## 1st Qu.:-0.700169 FALSE:2590   
## Median :-0.172014 TRUE :50   
## Mean : 0.000000   
## 3rd Qu.: 0.735872   
## Max. : 5.343554

Se guardan los datos pre procesados dentro de un nuevo CSV llamado ‘hip\_version1.csv’

write\_csv(hip\_version1, path = 'hip\_version1.csv')

## Analisis de Tycho

tyc <- read\_excel("../00.Datos\_y\_TP/hyades\_source.xlsx", sheet = "Tycho")

Remuevo Estrellas del catalogo Hipparcos para no repetir el análisis, excepto las estrellas hyades que servirán como grupo de control para el análisis.

tyc\_version1 <- tyc %>%  
 left\_join(., hip[,c(1,9)]) %>%  
 filter(., !complete.cases(HIP) | Symbad\_Hyades == T)

## Joining, by = "HIP"

summary(tyc\_version1)

## recno TYCID1 TYCID2 TYCID3   
## Min. : 3569 Min. : 51.0 Min. : 1.0 Min. :1   
## 1st Qu.: 53053 1st Qu.: 660.0 1st Qu.: 407.0 1st Qu.:1   
## Median : 57162 Median : 724.0 Median : 789.0 Median :1   
## Mean : 73207 Mean : 899.4 Mean : 895.7 Mean :1   
## 3rd Qu.:105369 3rd Qu.:1298.0 3rd Qu.:1248.0 3rd Qu.:1   
## Max. :156305 Max. :1868.0 Max. :3943.0 Max. :2   
##   
## RA\_J2000\_24 DE\_J2000 pmRA pmDE   
## Min. :2.900 Min. : 3.403 Min. :-149.30 Min. :-187.8   
## 1st Qu.:4.271 1st Qu.: 8.419 1st Qu.: -3.00 1st Qu.: -15.4   
## Median :5.118 Median :14.109 Median : 1.60 Median : -7.1   
## Mean :4.910 Mean :13.967 Mean : 5.04 Mean : -11.4   
## 3rd Qu.:5.664 3rd Qu.:19.465 3rd Qu.: 8.60 3rd Qu.: -2.4   
## Max. :6.100 Max. :24.500 Max. : 179.00 Max. : 116.7   
##   
## BT VT V B-V   
## Min. : 4.347 Min. : 3.647 Min. : 3.540 Min. :-0.4590   
## 1st Qu.:10.572 1st Qu.: 9.813 1st Qu.: 9.726 1st Qu.: 0.3664   
## Median :11.192 Median :10.363 Median :10.285 Median : 0.5712   
## Mean :11.034 Mean :10.220 Mean :10.147 Mean : 0.6915   
## 3rd Qu.:11.647 3rd Qu.:10.763 3rd Qu.:10.696 3rd Qu.: 1.0166   
## Max. :12.853 Max. :11.954 Max. :11.928 Max. : 2.2856   
##   
## HD HIP Plx Symbad\_Hyades   
## Min. : 18053 Min. :15310 Min. :18.74 Mode:logical   
## 1st Qu.: 35207 1st Qu.:20215 1st Qu.:20.66 TRUE:49   
## Median :244138 Median :20661 Median :21.54 NA's:13775   
## Mean :172045 Mean :20710 Mean :21.93   
## 3rd Qu.:284308 3rd Qu.:21008 3rd Qu.:23.13   
## Max. :287468 Max. :28614 Max. :26.96   
## NA's :10106 NA's :13775 NA's :13775

Reviso Faltantes

sapply(tyc\_version1, function(x) sum(is.na(x)))

## recno TYCID1 TYCID2 TYCID3 RA\_J2000\_24   
## 0 0 0 0 0   
## DE\_J2000 pmRA pmDE BT VT   
## 0 0 0 0 0   
## V B-V HD HIP Plx   
## 0 0 10106 13775 13775   
## Symbad\_Hyades   
## 13775

Los únicos faltantes son los IDs cruzados de los otros catalogos.Se procede a remover las columnas de IDs de otros catalogos, incluídos los TYCID1, TYCID2 y TYCID3.

tyc\_version1 <- tyc\_version1 %>% select(., c(1,5:12,16))

Se normaliza las variables.

tyc\_version1 <- tyc\_version1 %>%  
 mutate(RA\_J2000\_24 = scale(.$RA\_J2000\_24)) %>%  
 mutate(DE\_J2000 = scale(.$DE\_J2000)) %>%  
 mutate(pmRA = scale(.$pmRA)) %>%  
 mutate(pmDE = scale(.$pmDE)) %>%  
 mutate(BT = scale(.$BT)) %>%  
 mutate(VT = scale(.$VT)) %>%  
 mutate(V = scale(.$V)) %>%  
 mutate('B-V' = scale(.$`B-V`))

summary(tyc\_version1)

## recno RA\_J2000\_24.V1 DE\_J2000.V1   
## Min. : 3569 Min. :-2.2226528 Min. :-1.6908857   
## 1st Qu.: 53053 1st Qu.:-0.7072803 1st Qu.:-0.8880437   
## Median : 57162 Median : 0.2295475 Median : 0.0228186   
## Mean : 73207 Mean : 0.0000000 Mean : 0.0000000   
## 3rd Qu.:105369 3rd Qu.: 0.8338324 3rd Qu.: 0.8799935   
## Max. :156305 Max. : 1.3152558 Max. : 1.6859550   
## pmRA.V1 pmDE.V1 BT.V1   
## Min. :-7.691565 Min. :-9.254839 Min. :-7.766086   
## 1st Qu.:-0.400655 1st Qu.:-0.209890 1st Qu.:-0.536242   
## Median :-0.171412 Median : 0.225569 Median : 0.183839   
## Mean : 0.000000 Mean : 0.000000 Mean : 0.000000   
## 3rd Qu.: 0.177435 3rd Qu.: 0.472154 3rd Qu.: 0.712286   
## Max. : 8.669377 Max. : 6.720724 Max. : 2.112959   
## VT.V1 V.V1 B-V.V1   
## Min. :-8.632014 Min. :-8.585406 Min. :-2.586995   
## 1st Qu.:-0.534655 1st Qu.:-0.546892 1st Qu.:-0.731198   
## Median : 0.187620 Median : 0.179040 Median :-0.270593   
## Mean : 0.000000 Mean : 0.000000 Mean : 0.000000   
## 3rd Qu.: 0.712911 3rd Qu.: 0.712894 3rd Qu.: 0.730888   
## Max. : 2.276964 Max. : 2.314970 Max. : 3.584344   
## Symbad\_Hyades   
## Mode:logical   
## TRUE:49   
## NA's:13775   
##   
##   
##

Se guarda la primera versión de los datos pre-procesados.

write\_csv(tyc\_version1, path = 'tyc\_version1.csv')