Analyse : Merged all.csv

Datum : 19 oktober 2015

Auteur : Rob van Mechelen

Analyse tool: R version 0.98.1103

pdb = read.csv("Merged all.csv", header = TRUE, sep = ";")

> colnames(pdb)

[1] "Patientnummer" "Ziekenhuis"

[3] "Geslacht" "Geboortedatum"

[5] "Telefoon" "Etniciteit"

[7] "Diagnose" "CHADSVASc"

[9] "HASBLED" "Overleden"

[11] "Brief\_verwijzer" "Ontslag"

[13] "Overige\_voorgeschiedenis" "vgblanco"

[15] "vgcoronairlijden" "vgcoronairlijden\_vat"

[17] "vgCABG" "vgpostop\_afib"

[19] "vghartfalen" "vghypertensie"

[21] "vghypercholesterolemie" "vgvaatlijden"

[23] "vgschildklierlijden" "vglonglijden"

[25] "vgdiabetes" "vgCVA"

[27] "vgTIA" "vgembolie"

[29] "vgritmesoornis" "vgritmestoornis\_type"

[31] "verwijzer" "afib\_gezien\_op"

[33] "huidige\_medicatie" "Intakedatum"

[35] "Consulent" "Funktie"

[37] "CVDverhoogd" "Crepiteren"

[39] "Oedeem" "Souffle"

[41] "klgeen" "klpalpitaties"

[43] "kloverslaan" "klduizelingen"

[45] "klcollaps" "kldyspnoe"

[47] "klangina\_pectoris" "klmoe"

[49] "hartslag\_regulair" "hartslag\_irregulair"

[51] "hartslag\_onbekend" "klachten\_sinds\_dag"

[53] "klachten\_sinds\_week" "klachten\_sinds\_maand"

[55] "klachten\_sinds\_jaar" "klachten\_langer\_jaar"

[57] "klachtfrequentie\_dagelijks" "klachtfrequentie\_wekelijks"

[59] "klachtfrequentie\_maandelijks" "klachtfrequentie\_sporadisch"

[61] "klachtduur\_seconden" "klachtduur\_minuten"

[63] "klachtduur\_uren" "klachtduur\_dagen"

[65] "trigger\_geen" "trigger\_inspanning"

[67] "trigger\_emotie" "trigger\_alcohol"

[69] "trigger\_cyclus" "trigger\_houding"

[71] "trigger\_overig" "trigger\_overig\_tekst"

[73] "tijdstip\_overdag" "tijdstip\_nachts"

[75] "tijdstip\_beide" "start\_plotseling"

[77] "start\_geleidelijk" "start\_onbekend"

[79] "einde\_plotseling" "einde\_geleidelijk"

[81] "einde\_onbekend" "EHRA"

[83] "RR\_Systolisch" "RR\_Diastolisch"

[85] "Hartfrequentie" "Regulair\_irregulair"

[87] "Lengte" "Gewicht"

[89] "BMI" "Middelomtrek"

[91] "Heupomtrek" "Roken"

[93] "Jaren\_gerookt" "Alcohol\_hoeveelheid"

[95] "Koffie" "Drugs"

[97] "Drugs\_soort" "Overig"

[99] "Inaktief\_aktief" "Duursporter"

[101] "Datum" "Na"

[103] "K" "Ca"

[105] "Mg" "Kreat"

[107] "GFR" "TSH"

[109] "Trop\_T" "HS\_Trop\_T"

[111] "Trop\_i" "Pro\_BNP"

[113] "Hb" "Cholesterol\_totaal"

[115] "LDL" "HDL"

[117] "Triglyceride" "INR"

[119] "Bilirubine" "ASAT"

[121] "ALAT" "Gamma\_GT"

[123] "Alk\_fosfatase" "Opmerkingen"

[125] "LV\_funktie" "EF"

[127] "LA\_volume" "LA\_diameter"

[129] "RA\_volume" "Aortastenose"

[131] "Aortainsuff" "Mitralisstenose"

[133] "Mitralisinsuff" "Tricuspidaalstenose"

[135] "Tricuspidaalinsuff" "Pulmonaalstenose"

[137] "Pulmonaalinsuff" "Diastolische\_funktie"

[139] "Bijzonderheden" "Antistolling"

[141] "Antistolling\_start" "Beleid"

[143] "ECV" "CCV"

[145] "Vernacalant\_trial" "Ablatie"

[147] "rate\_control\_digitalis" "rate\_control\_betablokker"

[149] "rate\_control\_beta\_1\_blokker" "rate\_control\_calciumantagonist"

[151] "rate\_control\_expectatief" "rhythm\_control\_betablokker"

[153] "rhythm\_control\_sotalol" "rhythm\_control\_disopyramide"

[155] "rhythm\_control\_flecainide" "rhythm\_control\_propafenon"

[157] "rhythm\_control\_amiodarone" "rhythm\_control\_dronedarone"

[159] "overig\_ACE\_remmer" "overig\_diureticum"

[161] "overig\_statine" "overig\_PPI"

[163] "Follow\_up\_afib\_poli" "Follow\_up\_huisarts"

[165] "Follow\_up\_electrofysioloog" "Follow\_up\_algemeen\_cardioloog"

[167] "Doorverwijzing\_ander\_specialisme" "Ander\_specialisme"

>

Dit is iets teveel van het goede. Subset voor initiele analyse.

> library(dplyr)

> subset = select(pdb, Ziekenhuis:HASBLED, Intakedatum)

> str(subset)

'data.frame': 973 obs. of 9 variables:

$ Ziekenhuis : int 1 1 1 1 1 1 1 1 1 1 ...

$ Geslacht : Factor w/ 9 levels "#NULL!","1","2",..: 3 2 3 2 2 2 2 3 2 2 ...

$ Geboortedatum: Factor w/ 911 levels "#NULL!","01-apr-31",..: 103 627 184 287 877 489 320 475 516 94 ...

$ Telefoon : Factor w/ 64 levels "","`","010-4113144",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Etniciteit : Factor w/ 10 levels "#NULL!","1","2",..: 2 6 7 2 2 2 4 2 2 2 ...

$ Diagnose : Factor w/ 5 levels "#NULL!","1","2",..: 2 3 2 5 3 5 3 2 3 2 ...

$ CHADSVASc : Factor w/ 10 levels "#NULL!","0","1",..: 8 3 2 6 6 9 3 7 2 2 ...

$ HASBLED : Factor w/ 8 levels "#NULL!","0","1",..: 3 2 2 3 3 5 2 3 3 2 ...

$ Intakedatum : Factor w/ 432 levels "#NULL!","01-apr-14",..: 307 244 110 187 386 203 365 78 358 411 ...

>

Telefoonnummer niet relevant en #NULL! values oplossen met zo min mogelijk verlies van oorspronkelijke data. Daarom is dit een test fase, waar we samen aan moeten werken om goed te schonen.

table(subset$Geslacht)

#NULL! 1 2 2,1 2,2 2,3 2,4 2,5 3

0 591 382 0 0 0 0 0 0

Wat is in hemelsnaam geslacht 2,3 etc. Voorlopig opgelost door subset te schonen en kom dan op een test bestand.

str(test)

'data.frame': 973 obs. of 7 variables:

$ Ziekenhuis: int 1 1 1 1 1 1 1 1 1 1 ...

$ Etniciteit: int 1 5 6 1 1 1 3 1 1 1 ...

$ Diagnose : int 1 2 1 4 2 4 2 1 2 1 ...

$ CHADSVASc : int 5 1 0 3 3 6 1 4 0 0 ...

$ HASBLED : int 1 0 0 1 1 2 0 1 1 0 ...

$ Sex : Factor w/ 2 levels "1","2": 2 1 2 1 1 1 1 2 1 1 ...

$ age : int 75 62 36 80 68 77 68 61 75 63 ...

test$Sex = as.factor(test$Sex)

levels(test$Sex) = c("1", "2")

boxplot(test$age ~ test$Sex)

tapply(test$age,test$Sex,mean)

tapply(test$age,test$Sex,sd)

group1 = subset(test, test$Sex == "1")

group2 = subset(test, test$Sex == "2")

t.test(group1$age, group2$age, var.equal = TRUE)

Two Sample t-test

data: group1$age and group2$age

t = -5.5663, df = 971, p-value = 3.367e-08

alternative hypothesis: true difference in means is not equal to 0

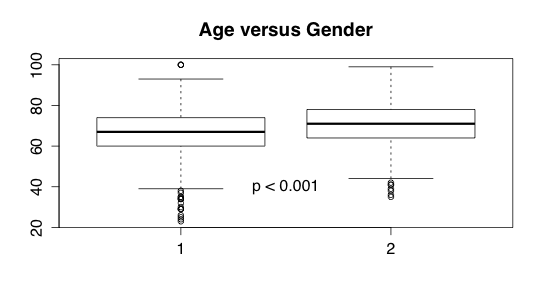
95 percent confidence interval:

-5.826464 -2.789043

sample estimates:

mean of x mean of y

65.93570 70.24346



table(test$Etniciteit)

1 2 3 4 5 6 7 8 9

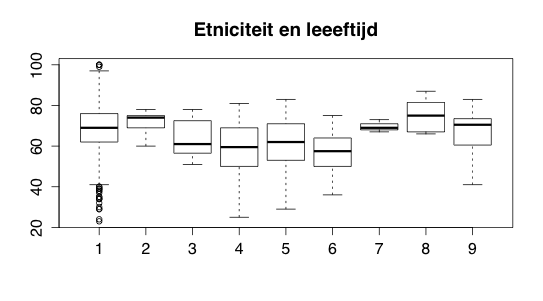
864 9 7 40 19 16 3 7 8

test$Etniciteit = as.factor(test$Etniciteit)

boxplot(test$age ~ test$Etniciteit)

Waarom 9 etniciteitsgroepen?

Moeten we reduceren interessante plot : Niet blanken zijn significant jonger



De leeftijd in alle 6 ziekenhuizen is voor patienten niet echt verschillend.

Leeftijd per ziekenhuis : er zijn 6 ziekenhuizen die mee doen

tapply(test$age,test$Ziekenhuis, mean)

1 2 3 4 5 6

66.22488 69.24074 66.75776 67.32432 68.96000 69.56250

Denk niet veel te beleven hier

Wel interessant is dat patienten met diagnose code 4 het oudst zijn

> tapply(test$age,test$Diagnose, mean)

1 2 3 4

66.85072 66.37396 68.96154 75.84536

>

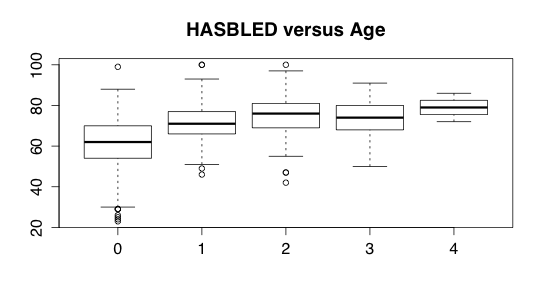
Maar nu komen de interesante zaken :

tapply(test$age, test$HASBLED, mean)

0 1 2 3 4

61.07852 71.22936 75.60795 74.05882 79.00000

>



Wat te denken van de CHADSVASc, weet alleen niet of de score 10 juist is of een verschrijving in de dataset. Moeten we samen naar kijken.

tapply(test$age, test$CHADSVASc, mean)

0 1 2 3 4 5 6 7 10

54.98343 61.94268 68.65873 74.19324 75.84483 76.50000 79.33333 83.83333 49.00000

>

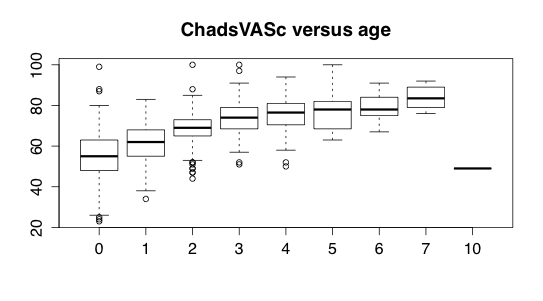
tapply(test$age, test$CHADSVASc, mean)

0 1 2 3 4 5 6 7 10

54.98343 61.94268 68.65873 74.19324 75.84483 76.50000 79.33333 83.83333 49.00000

> boxplot(test$age ~ test$CHADSVASc, main = "ChadsVASc versus age")

>



Nu dit is het voorlopig weer even.

Samenvattend

1. De dataset bestaat uit 973 observaties. Voorlopig gekeken naar 7 variabelen

2. Dataset geschoond voor #NULL! waarden met behoud van zoveel mogelijk observaties. Hier moeten we naar kijken

3. Gender : 591 genoteerd 1, 382 genoteerd 2 (man= 1/vrouw= 2?) Als dit zo is zijn de vrouwen gemiddeld ouder dan de mannen?

4. Statistich significant verschil in Age (p < 0.001)

5. Etniciteit : er zijn opmerkelijk weinig niet blanken : 864 blanken en 109 niet blanken

6. De NIET blanken zijn significant jonger op een enkele uitzondering na in de groepen 7-9 maar deze bevatten < 10 observaties. De reden waarom groep 2 ouder is ?, maar bevat slechts 9 observaties

7. Opmerkelijk verschil in observaties per ziekenhuis range 48-270. De gemiddelde leeftijd per ziekenhuis verschilt niet veel van elkaar, hetgeen suggereert dat de patienten ongeveer van dezelfde leeftijd zijn.

8.Diagnose factor is 1-4 weet niet wat waarvoor staat maar patienten met diagnose 4 zijn gemiddeld het oudst (significant)

9.CHADSVASc zoals te verwachten een hogere score bij een oudere patient (significant)

10. HASBLED score zoals te verwachten een hogere score bij een oudere patient (significant)

Heb p waardes niet voor alles uitgerekend, omdat ik werk met een dirty testset. Maar in de uiteindelijke set zal dit geen echte veranderingen brengen, verwacht ik. Zie dit als een quick and dirty clean (eerste indruk). Het aardige is dat je bijna 1000 observaties hebt.