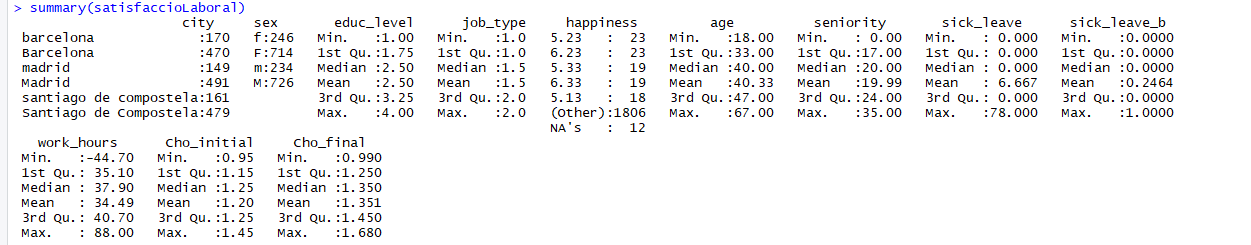
## Carregueu l’arxiu de dades en R i feu una breu descripció del mateix, on s’indiqui el nombre de registres llegits, el nombre de variables i els noms de les variables. Recordeu que abans de carregar l’arxiu, cal inspeccionar quin tipus de format csv es tracta per tal que la seva lectura sigui apropiada.

Càrrega de les dades. El separador de camps és una coma i el símbol decimal és un punt. Depenent de la carpeta:

satisfaccioLaboral <- read.csv("D:/Users/cagarcia/uoc/M2.954 - Estadística avançada/PAC1/rawData.csv", header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE, blank.lines.skip = TRUE)

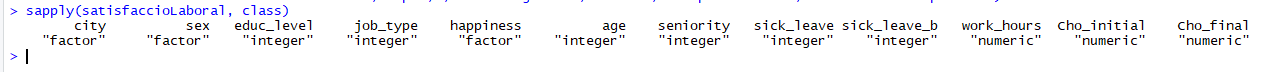
Hi ha 12 variables y el resum (abans de fer cap tractament de dades) és:



Tenim 12 variables de 1920 registres (valor obtingut amb la funció nrow) on:

|  |  |
| --- | --- |
| Nom variable | Tipus |
| city | Factor |
| sex | Factor |
| educ\_level | Integer |
| job\_type | integer |
| happiness | Factor |
| age | Integer |
| Seniority | Integer |
| sick\_leave | Integer |
| sick\_leave\_b | Integer |
| work\_hours | Numeric |
| Cho\_initial | Numeric |
| Cho\_final | Numeric |

El tipus s’ha obtingut amb la funció class (només carregant les dades, sense cap tractament).



2. Indiqueu el tipus de variable estadística de cada variable.

|  |  |
| --- | --- |
| Nom variable | Tipus de variable estadística |
| city | Categòrica nominal |
| sex | Categòrica nominal |
| educ\_level | Categòrica ordinal |
| job\_type | Categòrica ordinal |
| happiness | Quantitativa contínua |
| age | Quantitativa discreta |
| Seniority | Quantitativa discreta |
| sick\_leave | Quantitativa discreta |
| sick\_leave\_b | Categòrica ordinal |
| work\_hours | Quantitativa contínua |
| Cho\_initial | Quantitativa contínua |
| Cho\_final | Quantitativa contínua |

* City. Ciutat de la persona treballadora. És nominal ja que conté informació alfanumèrica.
* Sex. Sexe de la persona treballadora. És nominal ja que conté informació alfanumèrica.
* Educ\_level. Nivell d’educació de la persona treballadora. Conté informació numèrica que amaga informació categòrica (les categories es classifiquen com a 1, 2, 3 o 4).
* Job\_type. Tipus de treball de la persona treballadora. Conté informació numèrica que amaga informació categòrica (les categories es classifiquen com a 1 o 2). L’enunciat fa servir les classificacions C: qualificat i PC: poc qualificat.
* Happiness. Nivell de felicitat de la persona treballadora. Conté numeració contínua.
* Age. Edat de la persona treballadora. Aquesta variable podria ser considerada com a quantitativa contínua (si pensem que podem expressar l’edat amb decimals). Jo considero que la naturalesa de la variable és discreta (es respon 41 anys, no 41.3 ).
* Seniority. Anys d’experiència de la persona treballadora. Mateix cas que la variable Age.
* Sick\_leave. Dies en que la persona treballadora ha estat malalta en el darrer any. La considero quantitativa discreta perquè els dies s’indiquen de forma sencera. El dia o es conta o no es conta.
* Sick\_leave\_b. Variable binària que indica si la persona ha estat malalta. Evidentment, això és una categòrica ordinal.
* Work\_hours. Hores setmanals de feina de promig de la persona. És quantitativa contínua.
* Cho\_initial. Nivell de colesterol de la persona a la penúltima revisió mèdica. És quantitativa contínua.
* Cho\_final. Nivell de colesterol de la persona a la darrera revisió mèdica. És quantitativa contínua.

1. En el cas que R no assigni el tipus apropiat a alguna variable, realitzeu la conversió necessària per tal que el tipus final de cada variable sigui l’apropiat. Per exemple, possibles errors de variables quantitatives amb confusió en el separador decimal.

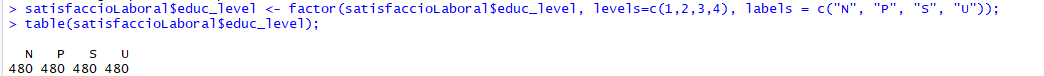
|  |  |  |
| --- | --- | --- |
| Nom variable | Tipus assignat | Tipus objectiu |
| city | Factor | Factor |
| sex | Factor | Factor |
| educ\_level | Integer | Ordered |
| job\_type | integer | Factor |
| happiness | Factor | Numeric |
| age | Integer | Integer |
| Seniority | Integer | Integer |
| sick\_leave | Integer | Integer |
| sick\_leave\_b | Integer | Logical |
| work\_hours | Numeric | Numeric |
| Cho\_initial | Numeric | Numeric |
| Cho\_final | Numeric | Numeric |

Nivell d’educació

La transformació és molt senzilla. Transformem els valors actuals (1, 2, 3 i 4) en els codis indicats a l’enunciat. No tractem els valors buits perquè no n’hi ha.

satisfaccioLaboral$educ\_level <- factor(satisfaccioLaboral$educ\_level, levels=c(1,2,3,4), labels = c("N", "P", "S", "U"));

table(satisfaccioLaboral$educ\_level);



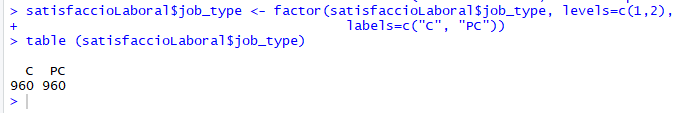
Tipus de feina

La transformació és molt senzilla. Transformem els valors actuals (1 i 2) en codis indicats a l’enunciat (C per a Qualificat i PC per a poc qualificat) . No tractem els valors buits perquè no n’hi ha.

satisfaccioLaboral$job\_type <- factor(satisfaccioLaboral$job\_type, levels=c(1,2),

labels=c("C", "PC"))

table (satisfaccioLaboral$job\_type)



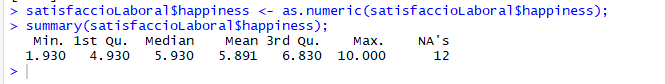
Nivell de felicitat

Les dades tenen números amb “,” en lloc de “.” com a separador de decimals. Substituïm els valors i convertim el tipus. Per ara, mantindrem els valors NA

satisfaccioLaboral$happiness <- gsub(",", ".", satisfaccioLaboral$happiness);

satisfaccioLaboral$happiness <- as.numeric(satisfaccioLaboral$happiness);

summary(satisfaccioLaboral$happiness);

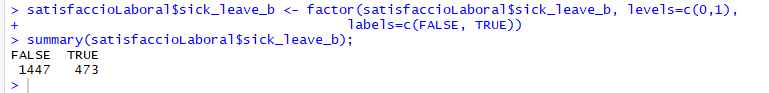


Binari de malaltia

Ara mateix el valor de la columna és 0 (false), 1 (true); convertim a booleà.

satisfaccioLaboral$sick\_leave\_b <- factor(satisfaccioLaboral$sick\_leave\_b, levels=c(0,1), labels=c(FALSE, TRUE));

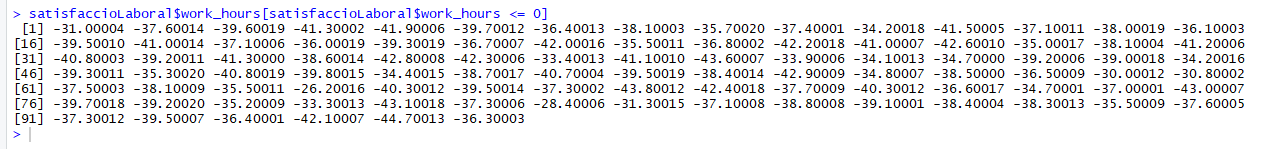
summary(satisfaccioLaboral$sick\_leave\_b);



Hores de feina

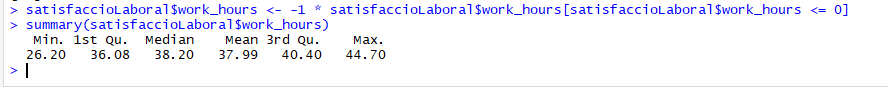
Podem veure que hi ha valors negatius en les hores de feina. Evidentment, això no és possible. No es poden fer un número negatiu de hores.

Els valors semblen molt raonables, amb el signe equivocat.



Podem pensar que es un error d’introducció de dades i que el signe és erroni. Canviem el signe multiplicant per -1.

satisfaccioLaboral$work\_hours <- -1 \* satisfaccioLaboral$work\_hours[satisfaccioLaboral$work\_hours <= 0]



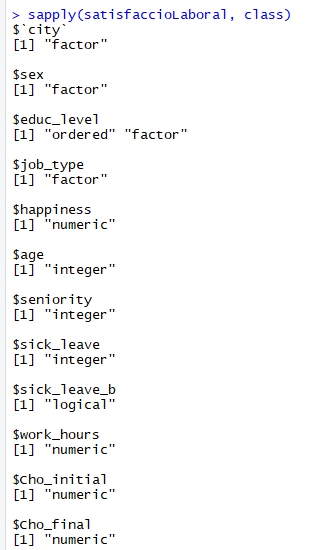
Resum de les variables

Convertim les variables que encara no estan en el tipus correcte.

satisfaccioLaboral$sick\_leave\_b <- as.logical(satisfaccioLaboral$sick\_leave\_b);

satisfaccioLaboral$educ\_level <- as.ordered(satisfaccioLaboral$educ\_level);

Podem veure que ja tenim els tipus de dades objectiu



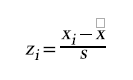
1. Normalitzeu / Estandaritzeu les variables quantitatives.

Nivell de felicitat

El problema que tenim és que hi ha valors NA a la variable. Tenim 12 casos. Tenim vàries estratègies possibles:

* Descartar les mostres
* Assignar un valor en funció d’una variable correlacionada
* Assignar un valor en funció de la distància de les altres mostres
* Assignar el valor de la mitja de les mostres

El rati de valors és molt petit (només 12), així que podem assumir que assignar la mitja és correcte i no afectarà gaire a la mostra. Una vegada fet això, normalitzem la variable:



satisfaccioLaboral$happiness[is.na(satisfaccioLaboral$happiness)] <- mean(satisfaccioLaboral$happiness[!is.na(satisfaccioLaboral$happiness)]);

satisfaccioLaboral$happinessNorm <- (satisfaccioLaboral$happiness - mean(satisfaccioLaboral$happiness)) / sd(satisfaccioLaboral$happiness);

Com a resultat tenim una variable amb mitjana 0 i desviació típica 1.



Altres variables

Primer estandarizem les working hours a un decimal

satisfaccioLaboral$work\_hours <- round(satisfaccioLaboral$work\_hours, digits = 1)

Nomalitzem les altres variables de la mateixa forma, que no tenen valors NA.

satisfaccioLaboral$ageNorm <- (satisfaccioLaboral$age - mean(satisfaccioLaboral$age)) / sd(satisfaccioLaboral$age);

satisfaccioLaboral$seniorityNorm <- (satisfaccioLaboral$seniority - mean(satisfaccioLaboral$seniority)) / sd(satisfaccioLaboral$seniority);

satisfaccioLaboral$sick\_leaveNorm <- (satisfaccioLaboral$sick\_leave - mean(satisfaccioLaboral$sick\_leave)) / sd(satisfaccioLaboral$sick\_leave);

satisfaccioLaboral$work\_hoursNorm <- (satisfaccioLaboral$work\_hours - mean(satisfaccioLaboral$work\_hours)) / sd(satisfaccioLaboral$work\_hours);

satisfaccioLaboral$Cho\_initialNorm <- (satisfaccioLaboral$Cho\_initial - mean(satisfaccioLaboral$Cho\_initial)) / sd(satisfaccioLaboral$Cho\_initial);

satisfaccioLaboral$Cho\_finalNorm <- (satisfaccioLaboral$Cho\_final - mean(satisfaccioLaboral$Cho\_final)) / sd(satisfaccioLaboral$Cho\_final);

S’han creat columnes noves amb els valors normalitzats per tal de no perdre la informació original. Crec que és millor crear columnes noves per tal de fer l’anàlisi i no deixar de tenir les columnes originals.

1. Normalitzeu / Estandaritzeu les variables qualitatives.

Ciutat

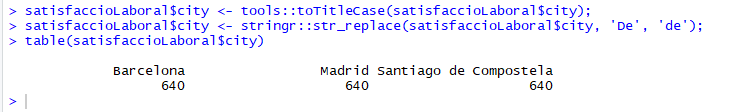
Per a normalitzar, primer llevem els espais en blanc. Després passem totes les paraules a TitleCase (primera lletra en majúscula i les altres en minúscula). Finalment convertim les preposicions a minúscules (en el nostre exemple, només tenim el “de”).

satisfaccioLaboral$city <- trimws(satisfaccioLaboral$city, "both");

satisfaccioLaboral$city <- tools::toTitleCase(satisfaccioLaboral$city);

satisfaccioLaboral$city <- stringr::str\_replace(satisfaccioLaboral$city, 'De', 'de');

table(satisfaccioLaboral$city);

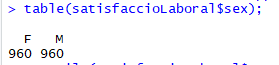


Sexe

Tenim la variable en majúscules i minúscules. La transformem en majúscules, a més de llevar-li els espais en banc.

satisfaccioLaboral$sex <- toupper(trimws(satisfaccioLaboral$sex, "both"));

table(satisfaccioLaboral$sex);



1. Reviseu possibles inconsistències entre variables: age versus seniority i número d’hores setmanals promig amb valor negatiu.

Les hores promig ja han estat modificades al punt 3 per a poder obtenir valors correctes a la normalització. Recordem que hem fet:

satisfaccioLaboral$work\_hours <- -1 \* satisfaccioLaboral$work\_hours[satisfaccioLaboral$work\_hours <= 0]

Per a trobar incoherències entre les variables age i seniority, primer crearem una variable amb l’edat en que la persona va començar a fer feina (age - seniority).

satisfaccioLaboral$startAge <- satisfaccioLaboral$age - satisfaccioLaboral$seniority;

I mirem quants tenien menys de 18 anys quan varen començar a treballar (en principi són dades inconsistents); en tenim 543.

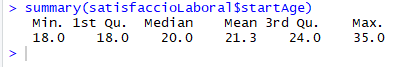
nrow(subset(satisfaccioLaboral, startAge < 18,)))



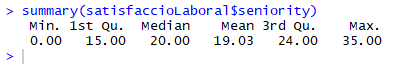
Assignem els valors a seniority tal que si li restem a l’edat mai sigui inferior a 18.

satisfaccioLaboral$seniority[satisfaccioLaboral$startAge < 18] <- satisfaccioLaboral$seniority[satisfaccioLaboral$startAge < 18] - (18 - satisfaccioLaboral$startAge[satisfaccioLaboral$startAge < 18])

Tornem a calcular la variable startAge com age – seniority i validem que ja no hi ha cap valor inconsistent:

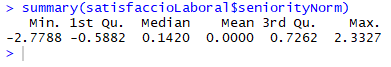


El resum de seniority és:



Recalculem la variable normalitzada de seniority:

satisfaccioLaboral$seniorityNorm <- (satisfaccioLaboral$seniority - mean(satisfaccioLaboral$seniority)) / sd(satisfaccioLaboral$seniority);

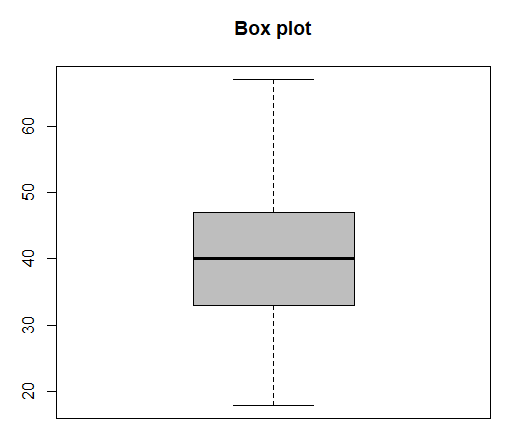


7. Busqueu valors atípics en les variables quantitatives.

i. Presenteu un boxplot per cada variable quantitativa.

Edat

boxplot(satisfaccioLaboral$age,main="Box plot", col="gray")

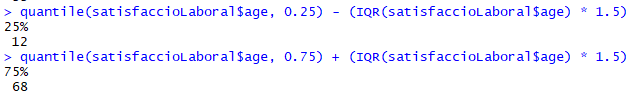


No hi ha valors outliers:

boxplot.stats(satisfaccioLaboral$age)$out



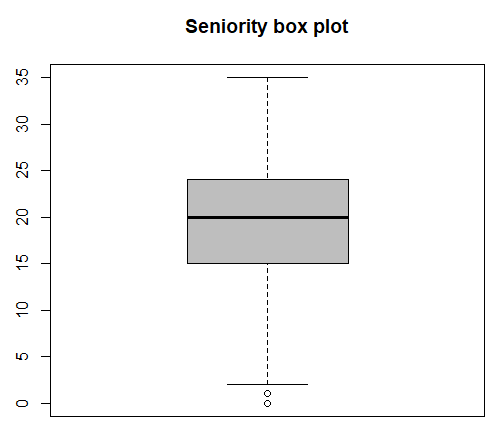
Ens assegurem calculant el rang d’edat en base al rang interquantil:



Efectivament, no hi ha ningú amb menys de 12 anys ni amb més de 68.

Anys d’experiència

boxplot(satisfaccioLaboral$seniority,main="Seniority box plot", col="gray")

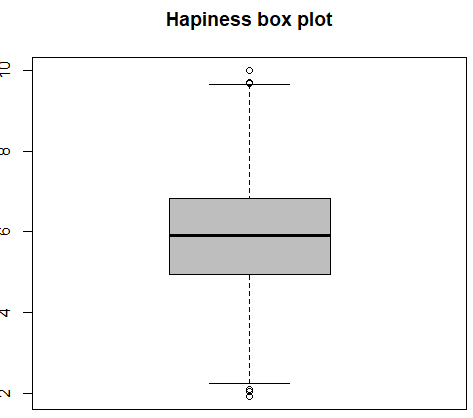


Aquí sí hi ha valors outliers, els que tenen 0 o 1 any d’experiència



Felicitat

boxplot(satisfaccioLaboral$happiness,main="Hapiness box plot", col="gray")



Aquí tenim valors outliers tant per abaix com per adalt:

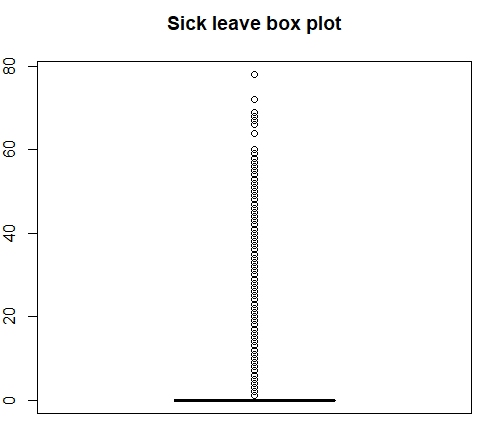
boxplot.stats(satisfaccioLaboral$happiness)$out

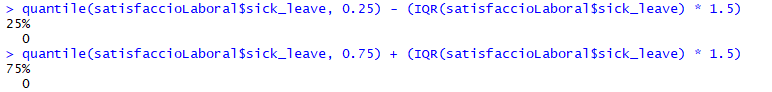


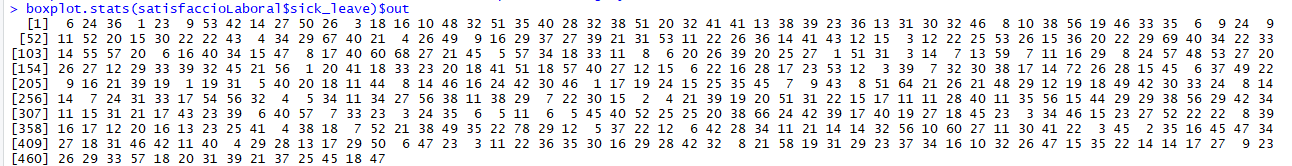
Dies de baixa

Aquí tenim molt de valors outliers; la gran majoria de valors són 0. Això provoca que tant el primer quantil com el tercer siguin 0. Això fa que qualsevol valor diferent de 0 sigui outlier.

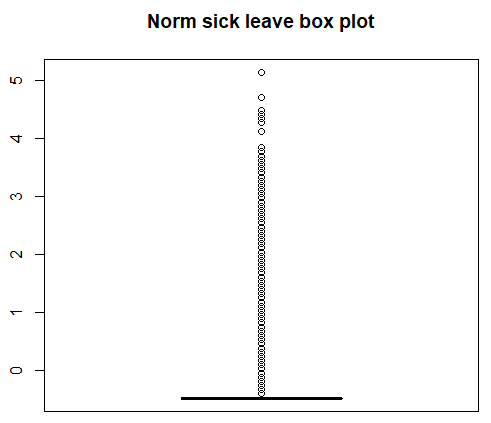
boxplot(satisfaccioLaboral$sick\_leave,main="Sick leave box plot", col="gray")





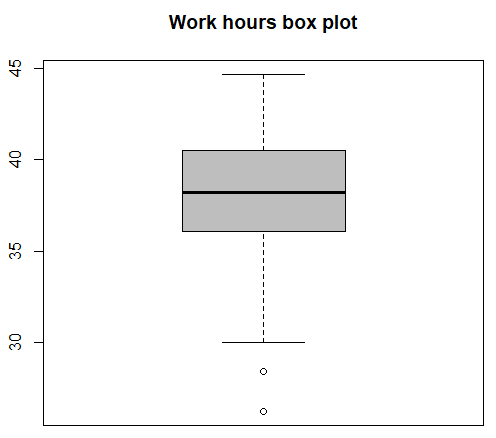


Els resultats són els mateixos si treballem amb la variable normalitzada



Hores de feina

boxplot(satisfaccioLaboral$work\_hours,main="Work hours box plot", col="gray")



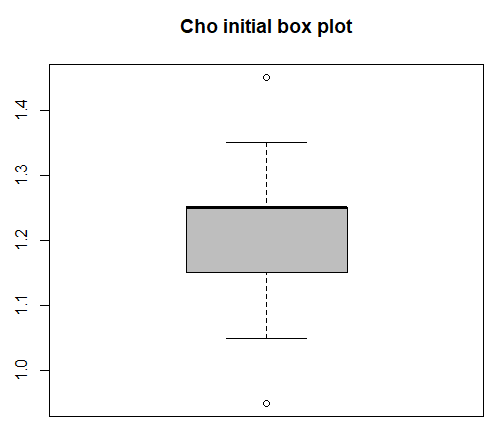
boxplot.stats(satisfaccioLaboral$work\_hours)$out

Tenim 40 valors fora dels límits (tots per abaix)



Colesterol penúltima medició

boxplot(satisfaccioLaboral$Cho\_initial,main="Cho initial box plot", col="gray")



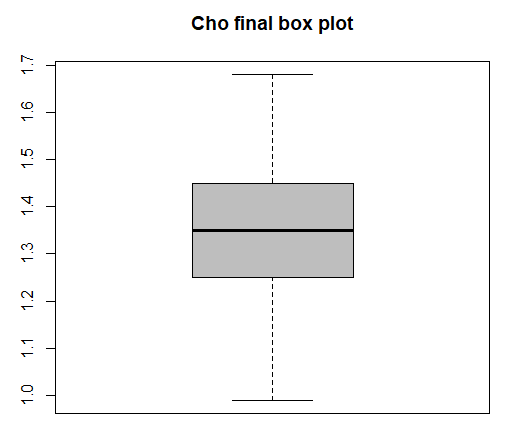
Aquí també tenim valors outliers tant per adalt com per abaix

> boxplot.stats(satisfaccioLaboral$Cho\_initial)$out

[1] 1.45 1.45 0.95 1.45

Colesterol última medició

Aquí no tenim valors outliers, perquè el rang és més ampli. És a dir: hi ha més gent amb el colesterol més baix i amb el colesterol més alt. Això fa que valors que amb la gràfica anterior s’haguessin considerat outliers, aquí són “normals”.



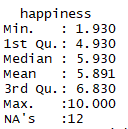
ii. Realizeu un quadre amb les estimacions robustes i no robustes de tendència central i dispersió per a cada variable quantitativa.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Mitjana aritmètica | Mediana | Mitjana retallada (al 5%) | Mitjana winsoritzada (al 5%) | Rang Interquartílic | Desviació estàndard | DAM |
| Happiness | 5.890639 | 5.92 | 5.892238 | 5.888439 | 1.885 | 1.344434 | 1.363992 |
| Age | 40.33125 | 40 | 40.24711 | 40.2749 | 14 | 9.879243 | 10.3782 |
| Seniority | 19.0276 | 20 | 19.30845 | 19.1276 | 9 | 6.847309 | 5.9304 |
| Sick leave | 6.666667 | 0 | 4.640046 | 6.176042 | 0 | 13.89998 | 0 |
| Work hours | 37.98646 | 38.2 | 38.15856 | 38.04271 | 4.325 | 3.443904 | 3.18759 |
| Cho ini | 1.200208 | 1.25 | 1.200116 | 1.200104 | 0.1 | 0.07714255 | 0.14826 |
| Cho fi | 1.351323 | 1.35 | 1.351279 | 1.351651 | 0.2 | 0.1279922 | 0.14826 |

8. Valors perduts

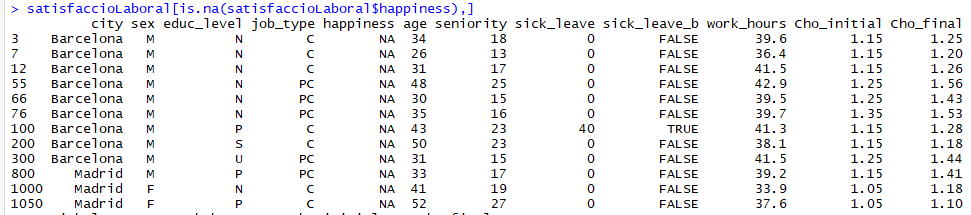
i. Busqueu quines variables i quins registres tenen valors perduts.

L’única variable amb valors perduts és Hapiness (12 NA). Com abans he assignat el valor de la mitja per a fer l’anàlisi, torn enrere el canvi:



Els registres són:

satisfaccioLaboral[is.na(satisfaccioLaboral$happiness),]



ii. Imputeu els valors a partir dels k-veïns més propers usant la distància de Gower amb la informació de totes les variables.

Prendrem com a nombre K = 5, que és el valor per defecte. Considerem totes les variables, no s’exclou cap columna. Per veure el resultat, guardem els valors a la variable happinessComplete. Així podrem comprovar els valors assignats:

satisfaccioLaboral.complet <- VIM::kNN(satisfaccioLaboral)

satisfaccioLaboral.complet$happiness

satisfaccioLaboral$happinessComplete <- satisfaccioLaboral.complet$happiness

Comprovem els valors assignats:

satisfaccioLaboral$happinessComplete[is.na(satisfaccioLaboral$happiness)]



Ara que sabem que està bé, li assignem a la variable:

satisfaccioLaboral$happinessComplete <- satisfaccioLaboral.complet$happiness

Si tornem a calcular els estimadors de la variable:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Mitjana aritmètica | Mediana | Mitjana retallada (al 5%) | Mitjana winsoritzada (al 5%) | Rang Interquartílic | Desviació estàndard | DAM |
| Happiness  Antic | 5.890639 | 5.92 | 5.892238 | 5.888439 | 1.885 | 1.344434 | 1.363992 |
| Happiness Nou | 5.890339 | 5.93 | 5.891904 | 5.888139 | 1.9 | 1.345283 | 1.363992 |

