**Enunciat**

1. **Fes un esquema i descriu com s'agrupen, en funció de l'objectiu del projecte, les diferents tècniques de mineria de dades.**

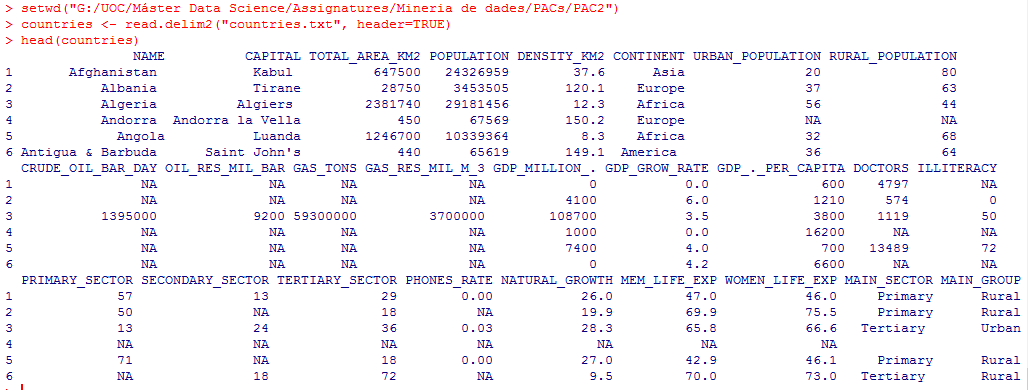
**Per a la resolució d’aquesta pregunta, pren com a punt de partida el material docent i amplia la resposta amb la bibliografia i recursos d'Internet (recorda citar les fonts).**

**Enunciat**

**Contextualitza els exemples de les següents preguntes respecte al projecte que has definit a la PAC1. Si ho desitges, pots redefinir o afinar el projecte.**

1. **Suposem que et trobes a la fase de preparació de dades del projecte que vares proposar a la PAC1. Fes una proposta teòrica de que faries, quins problemes et podries trobar i com els afrontaries. Pels posterior exercicis llegeix com a referència el document del wiki «Anàlisi de dades i estadística descriptiva amb R i R-Commander» i la resta d'enllaços que trobaràs al wiki.**
2. **A partir de les dades del fitxer *country.xlsx* carregades a R analitza i grafica tota la informació que trobis pertinent per conèixer el domini de les dades. Explica el procés seguit, quin objectiu tenies i com ho has plantejat i fet, no només captures de pantalla. Explica els resultats obtinguts.**

Primer de tot especifico la ruta al directori de treball. A continuació carrego el fitxer *country.xlsx* a l’entorn de R, en un dataframe amb nom *countries*. Com que aquesta opció m’ha donat problemes, he transformat el fitxer *countries.xlsx* en un fitxer de text, *countries.txt*, i d’aquesta manera ja s’ha carregat correctament. Comprovo que s’ha carregat correctament amb la funció *head*, que mostra en consola les primeres files.

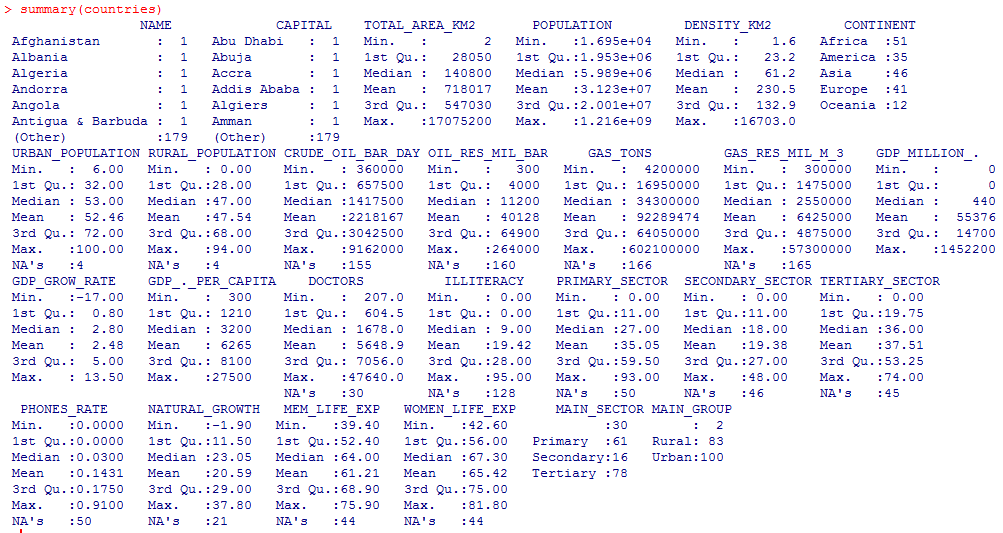


Passo a mirar les dimensions del dataframe countries:



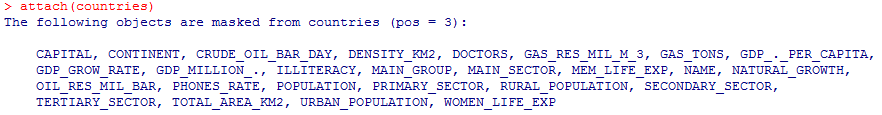
Aquest resultat ens diu que countries té 185 files i 26 columnes.

Sempre és convenient encarar una anàlisi de dades amb una visió general de les variables. Empraré la funció *summary*, que mostra per a cada variable el valor mínim, el máxim, la mitjana, els tres quartils i el número de NA’s, si es tracta de variables numèriques, i la freqüència d’aparició en el cas de les variables de text:



Per exemple, en la variable de text CONTINENT veiem que es llisten els 5 continents i per cada un surt el total de països que hi ha de cada continent. Per la variable numèrica DENSITY\_KM2 es mostren els valors estadístics corresponents.

Per tal de poder realitzar l’anàlisi d’una manera més còmode utilitzaré la funció *attach* per poder accedir a les diferents variables donant només els seus noms, sense haver de fer referència cada vegada al dataframe *countries*:

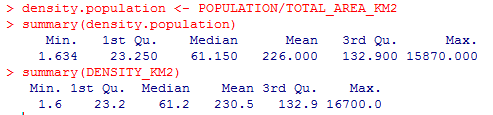


Faig una prova amb una variable, CONTINENT, per exemple:

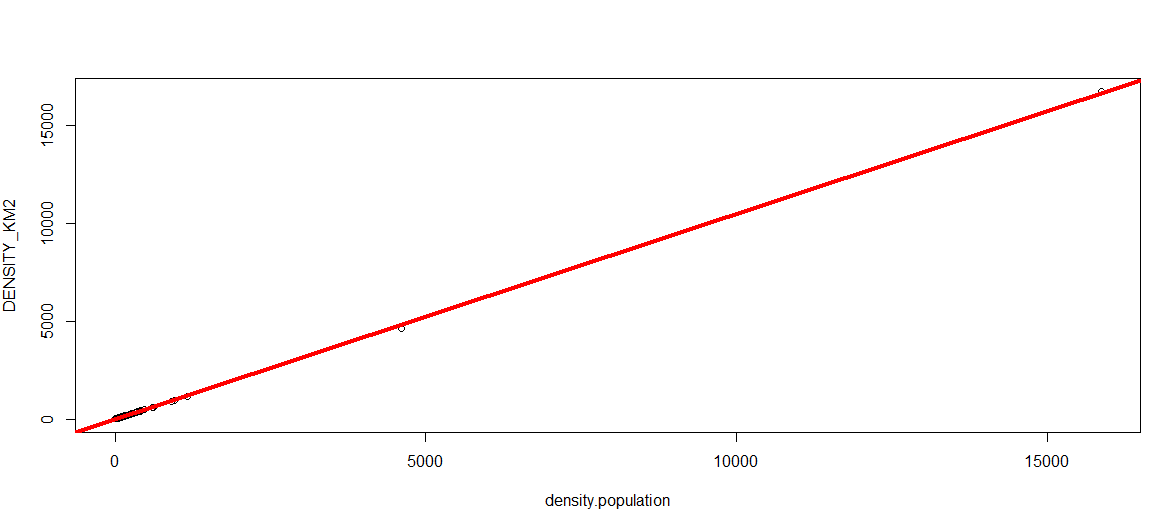


Veiem que ha anat bé.

Hi ha vegades en que unes variables es poden deduir d’unes altres. En el nostre exemple, això hauria de passar, per exemple, amb la variable DENSITY\_KM2, que s’ha de poder deduir de les variables POPULATION i TOTAL\_AREA\_KM2. Anem a veure-ho:



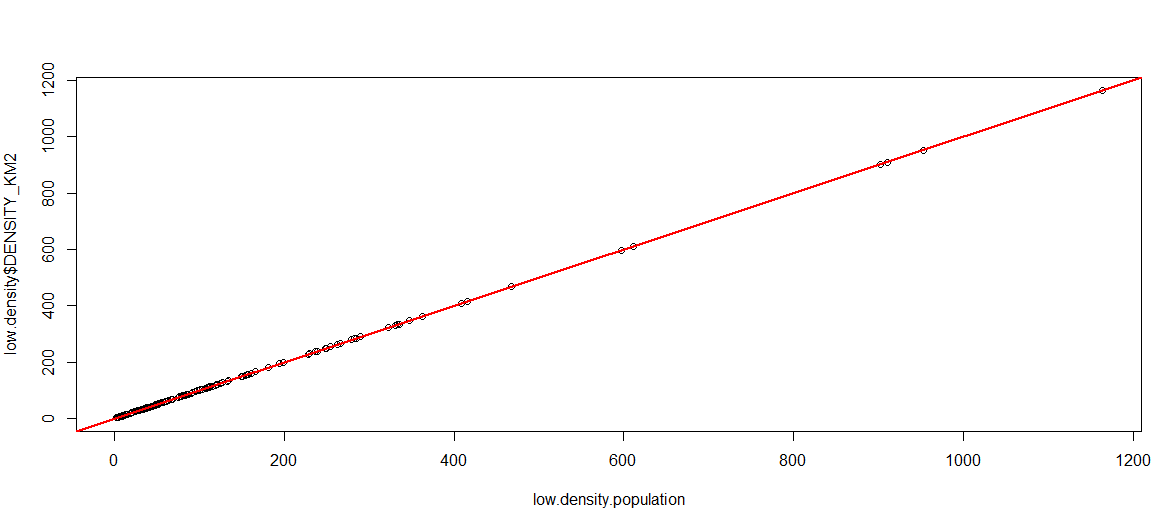
Veiem que, efectivament, els valors que s’obtenen amb la funció summary a la densitat de població donada per la variable DENSITY\_KM2 i a la calculada amb les variables POPULATION i TOTAL\_AREA\_KM2 són molt semblants. Anem a veure-ho fent un diagrama de dispersió:



Veiem que, efectivament, l’ajust és molt bo. Si calculem la correlació entre ambdues variables obtenim:



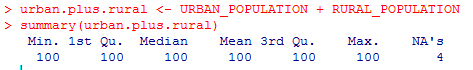
Obtenim una correlació de pràcticament 1. Una altra fet que es pot veure en la representació gràfica és que tenim un parell de valors molt més alts que la resta per la densitat de població. Això ja ho podíem haver deduït amb els resultats de la funció *summary*, en que veiem que el valor màxim està molt per damunt de la mitjana o del tercer quartil. Anem a visualitzar les mateixes dades però obviant aquests dos casos extrems:



Veiem també que hi ha dues variables redundants, URBAN\_POPULATION i RURAL\_POPULATION, ja que, en teoria:

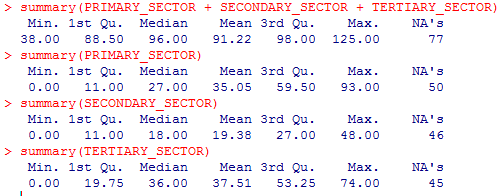
RURAL\_POPULATION + URBAN\_POPULATION = 100

Anem a veure si això es compleix:

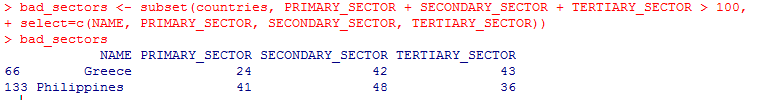


Sí, es compleix exactament. Per tant, podríem eliminar una de les dues variables i no perdríem informació.

Passem a fer una altra comprovació. Sabem que la població es divideix en tres sectors: primari, secondari i terciari, i que la suma dels tres sectors en cada país hauria de ser el 100%. Anem a comprovar-ho:



Veiem que la suma dels tres sectors té el màxim a 125, per sobre del màxim teòric del 100%, i a més ja es veu que es dóna el cas de que sembla bastant habitual tenir valors per sota del 100%, fins i tot tenim el mínim al 30%. Mirant individualment cada sector veiem que hi ha bastants casos en que tenim NA’s en comptes de valors numèrics. Anem a veure quins són els països que ens donen un valor major que el 100%:



Veiem que són Grècia i Filipines. En algun moment deu haver-se produit algun error en el càlcul dels percentatges dels sectors de població.

1. **Discretitza com a mínim un camp del fitxer *country.xlsx*. Redacta un breu escrit detallant què has fet i amb quin propòsit. Quins camps has triat per categoritzar? Quins passos has seguit? Quin resultat has obtingut? Confronta i explica el que has fet amb la lectura del Mòdul 2.**
2. **Proposa i realitza com a mínim una transformació mitjançant expressions de un camp o camps del fitxer *country.xlsx* que vegis interessant. Quins camps has tractat? Amb quin propòsit? Que has obtingut? Quines expressions t'han ajudat? Confronta i explica el que has fet amb la lectura del Mòdul 2.**

**BIBLIOGRAFIA**

* Hernández Orallo, J.; Ramírez Quintana, M. J.; Ferri Ramírez, C.: *Introducción a la Minería de Datos*. Pearson Prentice Hall. Madrid. 2004.
* Mor i Pera, E.; Sangüesa i Solé, R.; Molina Félix, L. C.: *Mineria de dades*. Material docent de la UOC. UOC – Universitat Oberta de Catalunya.
* Grus, N.: Data Science from Scratch. O’Reilly Media. Sebastopol, CA. 2015.
* Outline of machine learning

*https://en.wikipedia.org/wiki/Outline\_of\_machine\_learning*