Sessió 11

Exemple: Anàlisi exploratòria d'un conjunt de dades

En aquesta sessió farem un repàs del que hem après en les pràctiques anteriors fer una anàlisis exploratòria d'unes dades.

La pràctica/lliurament consistirà en carregar el fitxer, observar les seves característiques, preparar-lo per treballar-hi, analitzar algunes de les variables que conté i detectar les relacions que hi ha entre elles.

El conjunt de dades en el que treballarem conté mesures d'altures màximes d'espècies de plantes que creixen en diferents llocs i característiques bioclimàtiques i geogràfiques del lloc. La taula següent té la informació d'algunes de les variables del fitxer:

Taula de variables

height	altura màxima en metres de la planta
logheight	logaritme de l'altura
lat	latitud (en valor absolut)
long	longitud
alt	altura sobre el nivell del mar
temp	temperatura mitjana anual
diurn.temp	temperatura mitjana diurna
isotherm	mitjana del rang de temperatures mensual (màxima del mes - mínima del mes) dividit pel rang anual (màxima anual - mínima anual)
temp.seas	("temperature seasonality"). És una mesura de la variació de temperatura al llarg de l'any (desviació o CV de la variable temperatura mensual)
temp.max.warm	temperatura màxima del mes més càlid
temp.min.cold	mínima del mes més fred
temp.ann.range	temperatura màxima anual menys temperatura mínima
temp.mean.wetqr	temperatura mitjana de l'estació humida
temp.mean.dryqr	temperatura mitjana de l'estació seca
temp.mean.warmqr	temperatura mitjana de l'estació càlida
temp.mean.coldqr	temperatura mitjana de l'estació freda
rain	precipitació anual
rain.wetm	precipitació en el mes humit
rain.drym	precipitació en el mes sec
rain.seas	mesura de la variació de precipitació mensual
rain.wetqr	precipitació en l'estació humida
rain.warmqr	precipitació en l'estació càlida
rain.coldqr	precipitació en l'estació freda
LAI	("leaf area index"). Es defineix com la superfície que ocupen les fulles per unitat de superfície de la terra. Mesura la densitat de vegetació.
NPP	("net primary productivity"). Indica el grau d'acumulació de CO2 a l'atmosfera en ecosistemes terrestres.
hemisphere	hemisferi (-1 és sud i 1 és nord)

11.1 Guió d'anàlisi d'un conjunt de dades

1. **Càrrega del fitxer:**
   - Importa el fitxer **Plant_height.xlsx** que trobaràs al campus Virtual i assigna'l a un objecte anomenat `df.plantes`.
2. **Observació de les característiques:**
   - De quin tipus és l'objecte `df.plantes`? Si no és un data.frame, fes la transformació per a que ho sigui.
   - Quants casos té l'objecte? Quantes variables?
   - De quin tipus són les variables?
3. **Preparació de l'objecte:**
   (a) Visualitza els 3 primers casos i determina si caldria canviar el tipus d'alguna variable. Si és així, fes-ho (potser hauràs de tabular alguna variable per comprovar si és un factor).
   (b) Crea una variable única ajuntant les variables "site" i "Family" amb una barra baixa i assigna-la com a nom de les files.
   (c) Compta quants valors perduts hi ha a cada un dels casos i elimina aquells que tinguin més de 6 casos perduts.
   (d) Compta quants valors perduts hi ha a cada variable i elimina aquelles que tinguin més de 15 casos perduts.
   (e) La variable `growthform` té 10 valors perduts. En aquests casos coneixem almenys l'altura de la planta (`height`), que ens pot ajudar a predir quin és el seu valor de `growthform`. Estudia l'altura en els diferents grups definits per la variable `growthform` i, segons això, assigna un valor de la variable `growthform` en cada un dels 10 casos on té valor perdut. Exclou els tipus de `growthform` que tenen un sol cas o dos.
   (f) Les variable `LAI` té 6 valors perduts. En cada un d'aquests casos coneixem el valor de la variable `rain`. Calcula la recta de regressió de `LAI` sobre la variable `rain`. Substitueix els valors perduts de la variable `LAI` per la predicció que dona la recta de regressió en funció de l'altura.
   (g) Fes el mateix amb la variable `NPP`
   (h) Re-codifica la variable `hemisphere` en la mateixa variable per a que sigui un factor que indiqui si la planta està a l'hemisferi "North" o "South".
   (i) Crea una nova variable al data.frame anomenada `tropic` que divideixi la terra pels tròpics de Càncer i Capricorni (+23º27'). Per fer-ho has de canviar el signe de la latitud per a l'hemisferi sud. Anomena els factors "Cancer", "Tropical" i "Capricorn". Utilitza la funció cut().
4. **Anàlisi:** En aquest apartat estudiarem les variables `height`, `LAI`, `rain`,`temp`, `hemisphere`, `growthform`, `tropic` i dues variables numèriques més de la teva elecció.
   (a) Especifica quines són les teves variables, de quin tipus i fes-ne un petit estudi univariant.
   (b) Utilitzant les dues variables `hemisphere` i `tropic` que acabes de crear, quina zona té unes temperatures més altes (variable `temp`)?
   (c) Té sentit el que hem fet en l'apartat (e) de preparació de les variables per omplir els valors perduts de la variable `growthform`? Justifica-ho amb una taula.
   (d) Té sentit el que hem fet en l'apartat (f) i (g) de preparació de les variables per omplir els valors perduts de les variable `LAI` i `NTT`? Explica en què et bases.
   (e) Estudia la variable `LAI` conjuntament amb la resta de variables numèriques que has triat. Pren mesures i presenta gràfics. Pots posar-los en la mateixa finestra. Recorda que l'escala de `LAI` ha de ser la mateixa per poder-los comparar.
   (f) Estudia la variable `LAI` en els grups definits pels diferents factors. Presenta en una mateia finestra gràfics de la variable `LAI` respecte diferents factors. Recorda que perquè els gràfics es puguin comparar bé cal que l'escala sigui la mateixa. Si cal també pots agrupar una de les teves variables numèriques.
   (g) Repeteix el procediment dels apartats (e) i (f) anteriors amb la variable `height`. Potser en comptes d'utilitzar `height` convé més utilitzar `logheight`. Tria una de les dues segons el que observis.

5. **Explica els resultats:**
   - Fes un informe final en Word del que has descobert. Utilitza la funcionalitat de generació d'informes amb RStudio, tal com s'explica a l'apartat 5 de la pràctica 7.
   - L'informe ha d'incloure els resultats i les instruccions en R que has fet servir per obtenir-los.
   - En l'arxiu de word també hi has d'incloure comentaris sobre les conclusions a les que has arribat. Per generar l'arxiu de word pots utilitzar la funcionalitat de RStudio "generar informe" (mira't la secció final d'aquesta pràctica).
   - Les teves explicacions les podràs editar i pulir un cop hagis generat l'informe, obrint-lo amb word.

Preguntes

L'objectiu és descobrir quines variables geogràfiques i climàtiques influeixen més en el creixement (altura) de les plantes.

1. Influeix la latitud en el creiement de la planta?
2. Hi ha diferències de creixement entre els dos hemisferis (comparant regions amb latituds semblants)?
3. Què és més determinant per al creixement? la pluja? la temperatura?
4. Amb quina o quines variables està més relacionada la variable `LAI`?