Curso: Machine Learning en Agricultura.
Tutor: Procesamiento y limpieza de datos
Taller No 4: Procesamiento de datos con R

Para el desarrollo de este taller debe descargar las bases de datos Soil_Data.csv, Weather_Data.csv y Dataset_Maize_Cordoba_Data_cleanning.csv. Adjuntas por el tutor vía Slack. También descargar el libro de códigos Codebook_Dataset_Maize_Cordoba_Data_ cleaning.xlsx. El ejercicio debe ser desarrollando, utilizando los paquetes dplyr de R y tidyr. El código debe ser generado y descrito en un archivo Rmarkdown.

1. Leer y unificar las bases de datos. Las bases de datos de clima, suelo y manejo están separadas. Utiliza de las funciones join para unificarlas en una sola base de datos.

¿En este caso hay una diferencia entre un inner_join, left_join o right_join?

Después de unificar la base de datos en una sola, utilice los criterios vistos en clase para remover variables que no apartaron al análisis de datos o que son sensibles. Puede apoyarse de la función select.

Comente brevemente las razones por las que decidió removerlas.

2. Ejecutar la limpieza de datos.

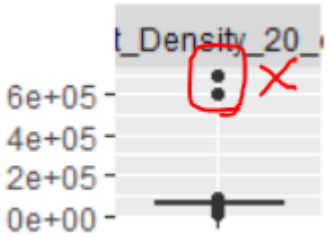
La base de datos tiene varios errores o procedimientos que se deben efectuar antes de proceder con el análisis.

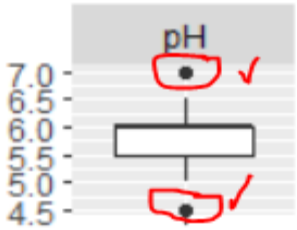
- Coloque las fechas en el formato adecuado
- Calcule y agregue como variable la longitud del ciclo (Fecha de siembra – Fecha de cosecha)
- Convierta las variables las variables que son texto a mayúsculas, puede apoyarse adaptando la siguiente función mutate_if(base_datos_maiz,is.factor,str_to_title).
- La variable de rendimiento parece tener un error de digitación, existe una “,” en vez de “.” , corrija este error adaptando la línea de código mutate(Yield = str_replace(Yield, ",","."), Yield = as.numeric(Yield)). Explique con sus propias palabras esta última línea.

3. Detectar y procesar valores atipicos.

Genere un boxplot para las variables cuantitativas o por lo menos para rendimiento, densidad de planta y la longitud de ciclo calculada en el punto anterior.

Identifique y remplaza por NA, solamente los valores extremamente atípicos, en cada una de las variables por las que hizo los boxplots. A continuación, hay un ejemplo.





los datos de densidad son considerados atípicos y extremos, los de pH son atípicos, pero por no estar tan alejados de la distribución no se consideran extremos.

4. Organizar base de datos final.

Reordene todas las variables de la base de datos, utilizado la función relocate, primero ubique todas las variables de manejo agronómico, en seguida las climáticas, luego las de suelo y finalmente la variable rendimiento.

5. Haga un summary final, y piense en otras modificaciones que se le ocurran que puedan mejorar la base de datos.