Limpieza y procesamiento de bases de datos

Hugo Andrés Dorado

Científico de datos

hugo.doradob@gmail.com

Motivación.

- Muchas de las bases de datos en agricultura se generan sin ninguna estructura que permita fácilmente ejecutar análisis de datos.
- Existen variables que no tiene sentido mantenerlas porque son ajenas a nuestros intereses dentro del análisis de datos.
- Algunas variables no pueden o no deben ser utilizadas en la forma que vienen originalmente.
- Existen variables que por cuestiones de privacidad deben ser removidas de la base de datos.

Aspectos a preguntarnos sobre las variables en una primera revisión la base de datos

- ¿Esta variable contribuye a mi objetivo de investigación? (predicción, agrupamiento, …)

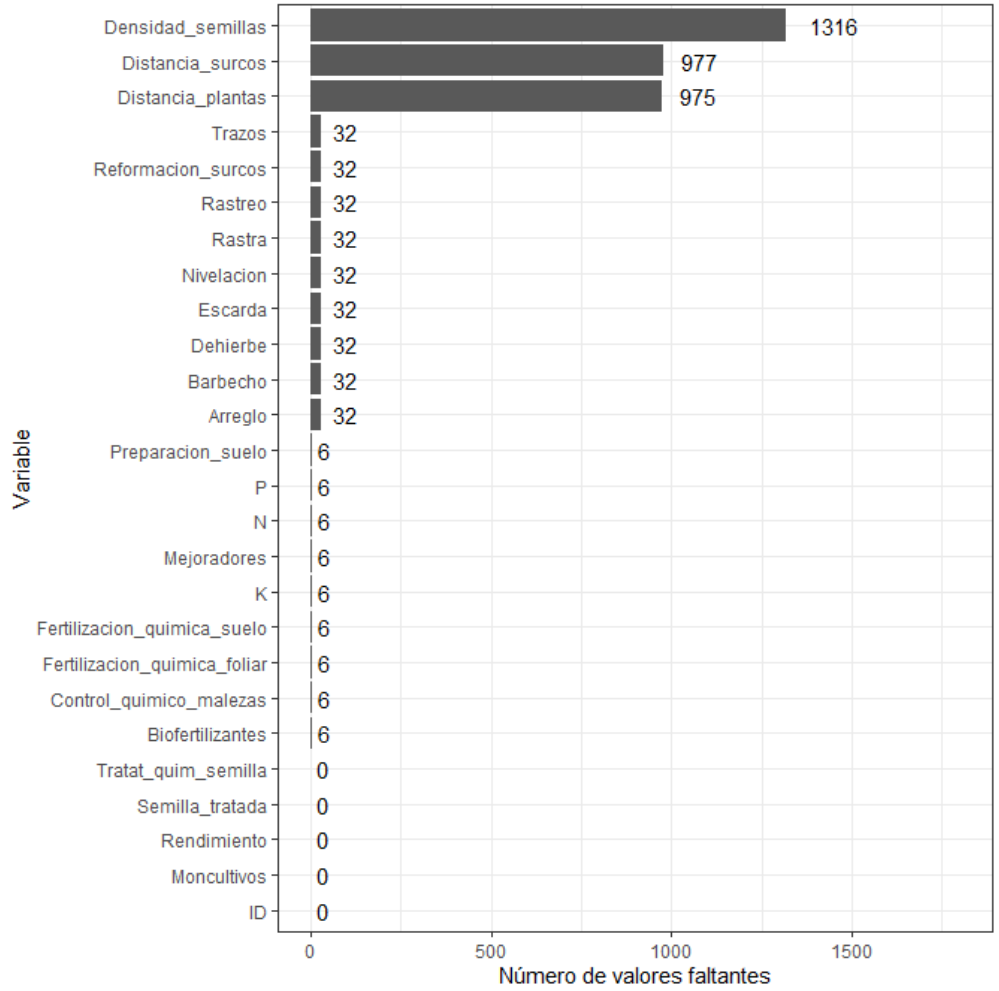
- ¿Esta variable tiene una escala uniforme? (kilogramos, hectárea, categorías,…)

- ¿La escala actual de la variable permite vincularla al análisis o necesita una transformación? (Fechas, muchas categorías)

Remover datos personales

ID	FECHA_ENCUESTA	NOM_PROD	CEDULA	LUGAR_EXPE	TEL_MOVIL	TEL_FIJO	CORREO ELECTRONICO	NOM_FINCA	DIR_RESIDENCIA
212		ANDRES EDUARDO MEJIA HERNANDEZ		NA	NA			EL GUARANGO	NA
213		BAIRO DE JESUS CIRO RESTREPO		NA	NA			BAIRO CIRO	NA
214		BERNARDO DE JESUS HENAO ESCOBAR		NA	NA			EL MIRADOR	NA
215		BERTHA INES CASTAÑO MUÑOZ		NA	NA			VILLA ELISA	NA
216		BERTULFO ROMAN FLOREZ		NA	NA			PARAJE PANTALIO	NA
217		C.I. FRUTY GREEN S.A.		NA	NA			EL CEBADERO	NA
218		CARLOS ALBERTO BEDOYA BEDOYA		NA	NA	NA		LA FURA1	NA
219		CARLOS ANDRES ROMAN		NA	NA	NA		EL MORRITO	NA
220		CARLOS MARIO SALAZAR BERMUDEZ		NA	NA	NA		NA	NA
221		CONSTRUCCIONES Y FINCAS (RICARDO ECHA	80004	NA	NA			LA PRADERA	NA
222		DIANA CAROLINA RAMIREZ GALVIS	1	NA	NA			LA SAMARIA	NA
223		DIEGO FERNANDO SALAS 3393200 E1 132		NA	NA			LOS FAROLES	NA
224		DIEGO LEON TOBON BEDOYA		NA	NA			EL SALADERO 2	NA
225		EDGAR DARIO TOBON TOBON		NA	NA			EL SALADERO	NA
226		EDUARDO ANTONIO ECHEVERRI ANGEL		NA	NA			PLAYITA LINDA	NA
227		EFRAIN DE JESUS BEDOYA BEDOYA		NA	NA			LA FORTUNA	NA
228		EFRAIN DE JESUS HENAO ESCOBAR		NA	NA			LA GEMELA	NA
229		EFRAIN DE JESUS VILLEGAS BEDOYA		NA	NA			NA	NA
230		ELENA GONZALEZ AYALA		NA	NA			NA	NA
231		ELKIN DE JESUS OSPINA MONTOYA		NA	NA			SAN MIGUEL	NA
232		ENRIQUE MONTOYA PELAEZ		NA	NA			NA	NA
233		ERIKA MARIA CASTAÑO GOMEZ		NA	NA			LEJANIAS	NA
234		FRANCISCO ALBERTO VARGAS MORALES		NA	NA			LA MERCED	NA
235		FABIAN CADAVID ORTIZ	NA	NA	NA			LAS COLINAS	NA
237		GABRIEL EDUARDO BOTERO MUÑOZ		NA	NA			LA LUCIA	NA
238		GLADIS DE JESUS ZAPATA DE CASTAÑO		NA	NA			PROVIDENCIA	NA
239		GLORIA ISAZA DE RESTREPO		NA	NA			EL MANANTIAL	NA
241		GUSTAVO ALONSO BEDOYA BOTERO		NA	NA			NA	NA
242		GUSTAVO RESTREPO ACEVEDO		NA	NA			ALTOS DE MAZARELO	NA
243		HECTOR DANIEL HENAO BOTERO		NA	NA	5		LOS SOLECITOS	NA

Remover variables con muchos datos faltantes



Remover variables que sean redundantes

Fechas

F.Siembra	F.Cosecha	Dias Ciclo
6/12/2018	8/18/2018	67
6/7/2018	8/13/2018	67
6/13/2018	8/22/2018	70
6/12/2018	8/25/2018	74
6/16/2018	8/25/2018	70
6/6/2018	8/17/2018	72
6/12/2018	8/19/2018	68
6/13/2018	8/22/2018	70
6/14/2018	8/22/2018	69
6/10/2018	8/23/2018	74
6/13/2018	2018/08/24	72
6/18/2018	8/26/2018	69
6/14/2018	8/20/2018	67
6/8/2018	8/15/2018	68
6/4/2018	8/13/2018	70

Coordenadas

DEPARTAMENTO	MUNICIPIO	lat	lon
Estelí	Pueblo Nuevo	13.272903	-86.536228
Estelí	Pueblo Nuevo	13.402430	-86.497179
Estelí	Pueblo Nuevo	13.370056	-86.514091
Estelí	Pueblo Nuevo	13.330616	-86.483238
Estelí	Pueblo Nuevo	13.316430	-86.492824
Estelí	Pueblo Nuevo	13.324065	-86.498764
Estelí	Pueblo Nuevo	13.315068	-86.507950
Estelí	Pueblo Nuevo	13.327409	-86.507067
Estelí	Condega	13.344149	-86.334541
Estelí	Condega	13.290238	-86.467527
Estelí	Condega	13.275412	-86.456048
Estelí	Condega	13.404080	-86.301450
Estelí	Condega	13.412028	-86.291793
Estelí	Esteli	13.265412	-86.334682
Estelí	Esteli	12.986344	-86.312953
Estelí	Esteli	13.120721	-86.416268

Rendimiento

Area	Rendimiento	Producción	Rendimiento/mz
1	6.12	6.12	6.12
1	3.82	3.82	3.82
1	0.50	0.50	0.50
1	10.42	10.42	10.42
1	23.50	23.50	23.50
1	6.50	6.50	6.50
1	8.00	8.00	8.00
1	8.00	8.00	8.00
1	19.00	19.00	19.00
1	26.00	26.00	26.00
1	2.48	2.48	2.48
1	8.00	8.00	8.00
1	21.25	21.25	21.25
1	20.00	20.00	20.00
1	14.00	14.00	14.00
0.5	2.00	2.00	4.00
0.5	2.00	2.00	4.00
1	2.00	2.00	2.00

Variables que necesitan transformación (Sumarización)

Filas repetidas, (generalmente varias aplicaciones de algún fertilizante o control)

ID_EVENTO	ID_PROD	FECHA_FERTI	TIPO_PROD_FERTI	CANTIDAD_PROD_FERTI
43	52	4/13/2013	Quimica	300
43	52	5/15/2013	Quimica	225
44	54	4/25/2013	Quimica	300
44	54	5/25/2013	Quimica	250
44	54	5/25/2013	Quimica	100
46	55	3/27/2013	Quimica	300
46	55	4/26/2013	Quimica	234
46	55	4/26/2013	Quimica	550

Variables convertidas a frecuencias y totales

ID_EVENTO	FrecFerQu	TotFerQui
43	2	525
44	3	650
46	3	1084
53	1	100

Variables

Fertilizaciones
Monitoreos
Variables climáticas
Riego
Labores
…

	A	B	C	D	E	F
1	DATE	ESOL	RAIN	RHUM	TMAX	TMIM
.557	4/5/2009	412.8747	0	70.99139	36	24.3016
.558	4/6/2009	513.9043	0	75.20833	34.8	24.9
.559	4/7/2009	396.5338	0	73.85714	34.1	25.6
.560	4/8/2009	397.8491	0	74.09524	33.9	25.4
.561	4/9/2009	448.4498	0	76.82609	34.6	24.9
.562	4/10/2009	481.8188	0	66.20671	39	24.8
.563	4/11/2009	448.1053	0	73.66386	35.9	25.4

Indicadores climáticos transformados (Acumulados, promedios, frecuencia, máximos o mínimos)

ID	FECHA_SIEMBRA	FECHA_COSECHA	ANO_COS	RENDIMIENTO_HA	TMAXavg	TMINavg	TEMPavg	GDaccu11	RANGO_Diurno_avg	Eneraccu
RC38_2009_5	4/5/2009	8/3/2009	2009	5600	33.11977441	23.67722572	28.39850006	1957.651791	9.442548692	43981.57
RC38_2009_6	4/5/2009	8/3/2009	2009	5775	33.11977441	23.67722572	28.39850006	1957.651791	9.442548692	43981.57
RC38_2009_7	4/5/2009	8/3/2009	2009	4200	33.11977441	23.67722572	28.39850006	1957.651791	9.442548692	43981.57
RC27_2013_3037	10/22/2012	2/19/2013	2013	5262	34.06942149	24.20578512	29.13760331	1880.2564	9.863636364	43883.31
RC38_2013_129	10/22/2012	2/19/2013	2013	5265	34.06942149	24.20578512	29.13760331	1880.2564	9.863636364	43883.31
RC38_2013_130	10/24/2012	2/21/2013	2013	5284	34.16363636	24.2107438	29.18719008	1873.7553	9.952892562	43962.81
RC38_2013_134	11/2/2012	3/2/2013	2013	6720	34.30661157	24.30743802	29.30702479	1862.2728	9.999173554	44100.78

Discusión con expertos

Variables comunes que se mantienen

Clima	Suelo	Manejo agronómico	Rendimiento/Rentabilidad
• Temperatura mínima media • Rango diurno medio • Energía solar acumulada • Frecuencia de días con temperatura máxima superior a 34 ° C • Precipitación acumulada • Frecuencia de días con temperatura mínima inferior a 8 ° C • Humedad relativa media	• Contenido de arcilla • Contenido de limo • Contenido orgánico del suelo • Capacidad de intercambio catiónico • Saturación básica • Drenaje	• Infiltración • Cultivar • Tratamiento de semillas • Tipo de labranza • Número de escarda mecánica • Número de aplicaciones de (fertilizaciones, fertilizantes foliares, bio fertilizantes, herbicidas post-siembra, insecticidas) • Cantidad total de nitrógeno aplicada • Cantidad total de fósforo aplicada • Cantidad total de fósforo aplicada • Cantidad total de potasio aplicada	

Estructura y limpieza de una base de datos

Estructura ideal de una base de datos

Cada fila representa una unidad de análisis y cada columna representa una variable.

	A	B	C	D	E
1	ID	Sowing_Date	Harvest_Date	Variety	Yield
2	RC61_2008_989	2008-03-07	2008-07-05	ACARIGUA	6700
3	RC62_2010_207	2010-07-22	2010-11-25	ACD 2526	9125
4	RC62_2011_275	2011-03-11	2011-07-15	ACD 2526	6375
5	RC62_2012_361	2011-09-08	2012-01-12	ACD 2526	6875
6	RC62_2011_303	2011-04-25	2011-08-29	ACD 2528	7500
7	RC62_2011_213	2010-08-30	2011-01-03	ACD 2540	6563
8	RC62_2011_274	2011-03-09	2011-07-13	caracoli	6250
9	RC62_2010_76	2009-12-19	2010-04-24	CHICALA	5600
10	RC62_2011_336	2011-08-06	2011-12-10	CHICALA	4625
11	RC62_2011_345	2011-08-22	2011-12-26	CHICALA	4687
12	RC62_2011_348	2011-08-23	2011-12-27	CHICALA	5163
13	RC62_2012_372	2011-09-14	2012-01-18	CHICALA	6875
14	ENA_2007a_106386	2007-02-21	2007-07-01	CIMARRON BARINAS	6937.5
15	ENA_2007a_100234	2007-03-21	2007-07-25	CIMARRON BARINAS	7500
16	ENA_2007a_102633	2007-04-14	2007-09-25	CIMARRON BARINAS	8187.5
17	ENA_2007a_101504	2007-05-14	2007-10-09	CIMARRON BARINAS	8000
18	ENA_2007a_100400	2007-05-26	2007-10-06	CIMARRON BARINAS	5187.5
19	ENA_2007a_100150	2007-05-26	2007-10-13	CIMARRON BARINAS	7812.5
20	ENA_2008a_101504	2008-03-01	2008-07-02	CIMARRON BARINAS	6562.5
21	ENA_2008a_100234	2008-04-29	2008-09-09	CIMARRON BARINAS	7000

Asegurarse de crear o identificar un ID que le permita conectar las bases de datos

Agregar un diccionario de datos

Variable	Descripción	Unidad	Rango
ID	Identificador	-	AVT_1,AVT_10,AVT_101,AVT_102,AVT_103,AVT_104,AVT_105,AVT_106,AVT_1
Region	Region	-	I,IV,RACCN,V,VI,RACCS,II,III
Departamento	Departamento	-	Nueva Segovia ,Granada,Caribe Norte,Carazo,Masaya,Rio San Juan,Jinotega,M
Municipio	Municipio	-	Jalapa,Nandaime,Siuna,Diriomo,Diria,Jinotepe,Santa Teresa,Masatepe,Morri
Latitude	Latitud	Decimales	[14.81,11.07]
Longitude	Longitud	Decimales	[-83.71,-87.38]
FechaGerminacion	Fecha de germinación	dd/mm/anio	[2018-11-13,2015-06-17]
FechaCosecha	Fecha de cosecha	dd/mm/anio	[2019-02-13,2015-10-19]
CrecimientoTotal	Número de dias de crecimiento	dias	[86.0,143.0]
Area	Area	mts2 ó manzanas	mts2 ó manzanas
Unidad	Unidad de area	Unidad de area	mts2, manzanas

Como mínimo se sugiere, por cada variable:

- Nombre corto
- Nombre completo (descripción)
- Unidad de medida
- Rango [Max - Min], o posibles categorías

Errores comunes - Estandarizar formato

Fechas

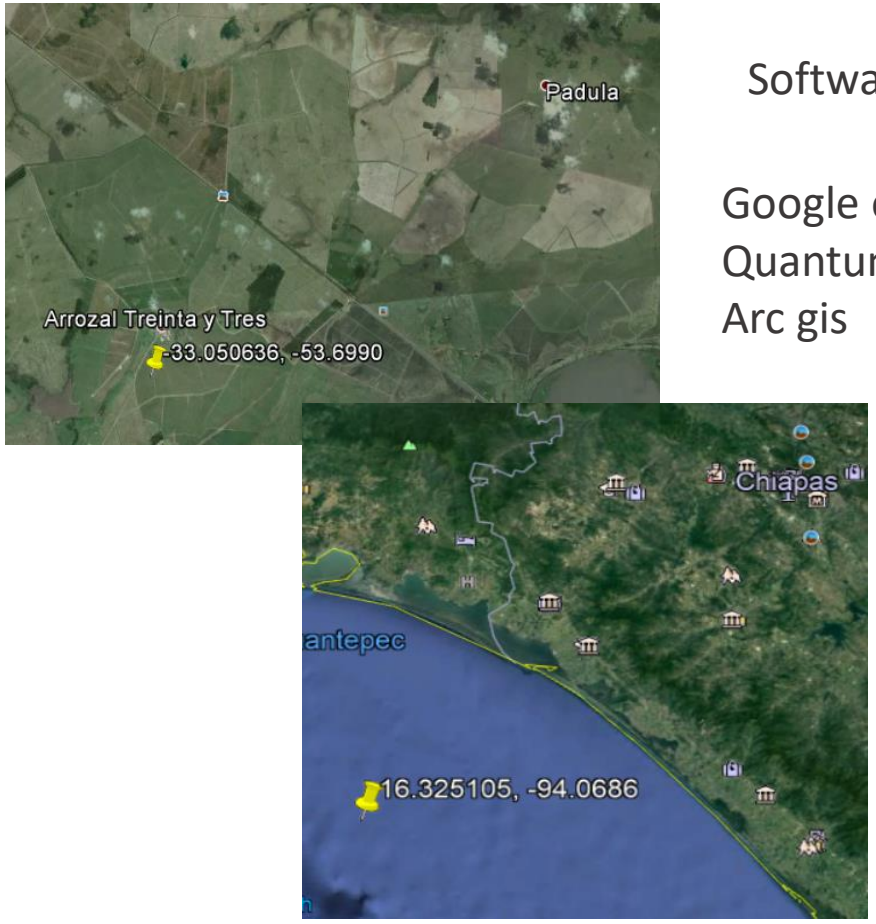
FECHA DE Germinacion	FECHA DE COSECHA	Rendimiento unitario qq/mz
6/2/2016	5-Oct-16	62.7
6/2/2016	25-Oct-16	39.9
6/10/2016	8-Oct-16	128
6/14/2016	10-Oct-16	80.49
6/6/2018	6-Oct-16	132.99
6/10/2016	8-Oct-16	74.91
6/8/2016		99
Ago.04/2016	8-Dec-16	123.75
2016		116.05
2016		118.184
Jul.16/2016	Nov.23/2016	137.5
Jul.06/2016	Nov.15/2016	53.999
Jun.27/2018	Nov.15/2016	41.316
Jun.27/2016	Nov.16/2016	109.197
Jun.24/2017	Nov.10/2016	34.9965
Jul.07/2016	Dic.01/2016	150.0015
Jun.10/2016	Oct.12/2016	139.348
Jun.17/2016	Oct.20/2016	158.4

Coordenadas mal registradas

37	5°08'27.5"	-75°54'31.3"	RISARALDA	APIA
38	5°08'42.3"	-75°55'02.2"	RISARALDA	APIA
39	5°78'41.0"	-75°05'02.8"	RISARALDA	APIA
40	5°67'16.8"	-75°04'17.0"	RISARALDA	APIA
41	5°08'17.8"	-75°54'18.4"	RISARALDA	APIA
42	5°09'41.8"	-75°55'26.4"	RISARALDA	APIA
43	5°09'41.6"	-75°55'26.1"	RISARALDA	APIA
44	5°09'35.3"	-75°55'10.0"	RISARALDA	APIA

Problemas de coordenadas y fechas

Revisión de coordenadas



Software de apoyo

Google earth
Quantum gis
Arc gis

Revisión de fechas registradas

O	P	Q	R	S
	ManeraSiemb	FechaGerminacion	FechaCosecha	CrecimientoTotal
711082	Chorrillo	6/5/2016	10/5/2016	122
646509	Chorrillo	7/10/2016	11/9/2016	122
500618	Espeque	8/10/2018	12/6/2019	483
085106	Espeque	8/29/2018	12/19/2018	112
496744	Espeque	9/14/2018	1/8/2019	116
021449	Espeque	8/15/2018	12/12/2018	119
534115	Espeque	8/17/2018	12/15/2018	120
756709	Espeque	8/31/2018	9/27/2018	27
235164	Chorrillo	7/6/2016	11/5/2016	122
569002	Chorrillo	6/27/2016	10/22/2016	117
905146	Chorrillo	6/27/2016	10/23/2016	118
401172	Chorrillo	6/24/2016	10/17/2016	115
086117	Chorrillo	6/5/2016	10/5/2016	122
436118	Chorrillo	6/7/2016	10/1/2016	116
038814	Espeque	6/10/2016	10/12/2016	124
544835	Espeque	6/17/2016	10/20/2016	125

Unidades de medida

Producción

ID	PN_ANO_P	UNIDAD_PN_PLÁTANO
130	15.00	CAJAS X SEMANA
138	85.00	CAJAS X SEMANA
142	40.00	CAJAS X SEMANA
145	25.00	CAJAS X SEMANA
147	22.00	CAJAS X SEMANA
148	7.00	CAJAS X SEMANA
152	16.00	CAJAS X SEMANA
153	40.00	CAJAS X SEMANA
155	180.00	CAJAS X AÑO
156	15.00	CAJAS X SEMANA
157	4800.00	CAJAS X AÑO
159	50.00	CAJAS X SEMANA
161	950.00	CAJAS X AÑO
166	4576.00	CAJAS X AÑO
167	390.00	CAJAS X AÑO
724	300.00	CARTONES
1017	55647	KILOGRAMO
1017	54000	KILOGRAMO
1017	43200	KILOGRAMO
1017	64818	KILOGRAMO

Coordenadas geográficas

ID	ALT_LOTE_	LAT_LOTE_	LONG_LOT	DIST_SIEM
735	1886	2.21603	-75.57572	6X6
744	932	2.41042	-75.52153	7X7
750	900	2.16405	-75.64209	8X8
926	1700	2.44	76,-73	7x5
2054	1772	2	76.67375	7X7
2149	1776msnm	2.57164°N	76.63855°O	7x5
2150	1762msnm	2.57193°N	7664061°O	7x5
2151	1766msnm	2.57062°N	76.64054°O	7X5
2165	1781msnm	2.58491°N	76.63085°O	7x5

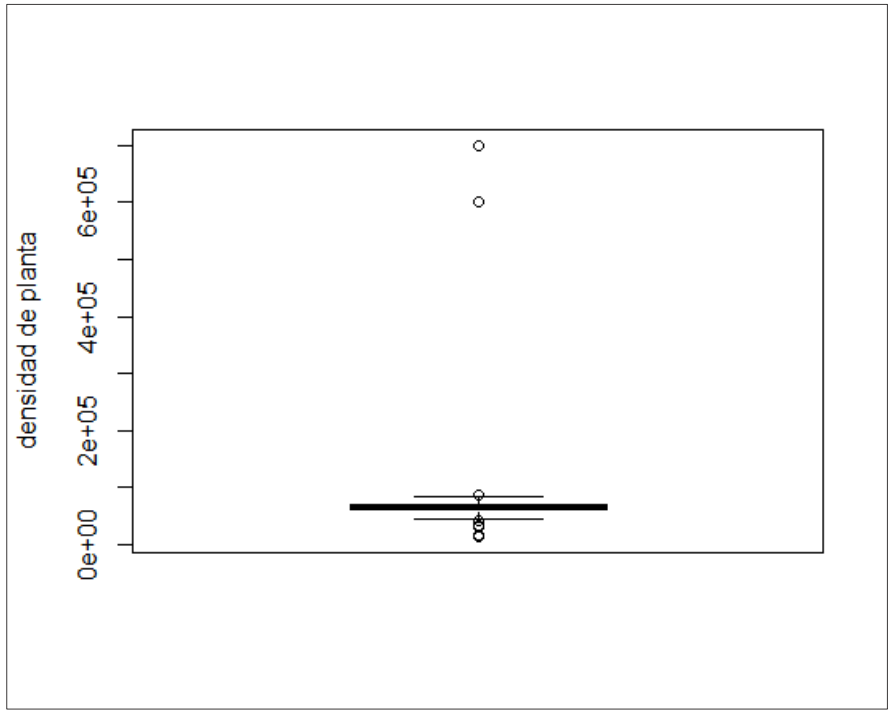
Área

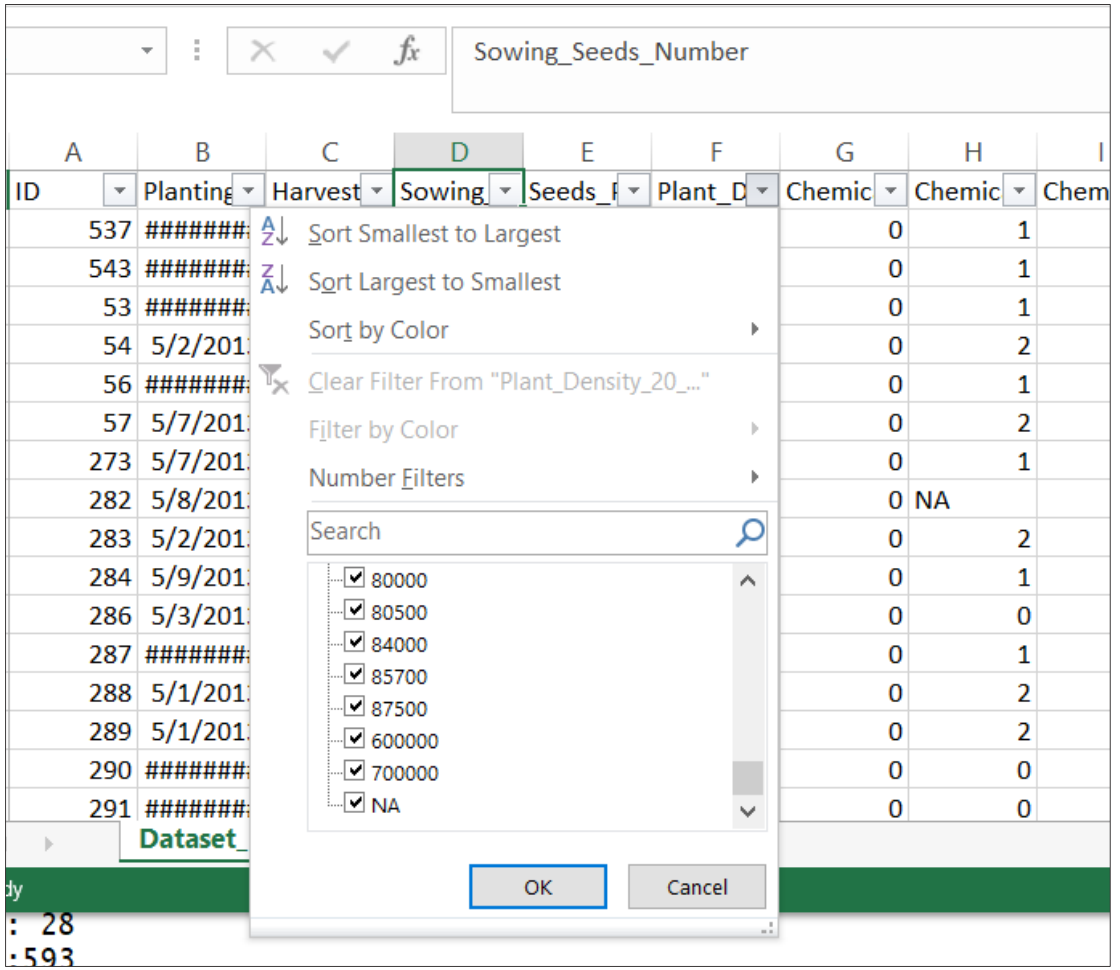
NO_LOTES_MANGO	NO_ARBO	AREA_MA	UNIDAD_MEDID
4	25	0.001	HECTAREAS
3.5	2	0.25	HECTAREAS
1	5	0.5	HECTAREAS
1	50	0.5	HECTAREAS
1	25	0.5	HECTAREAS
2	400	4	HECTAREAS
1	50	1.5	HECTAREAS
1	35	0.6	HECTAREAS
1	105	0.5	HECTAREAS
1	200	2	HECTAREAS
1	40	0.5	HECTAREAS
1	1300	1	HECTAREAS
1	2	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
2	800	5.76	HECTAREAS
220	220	2	HECTAREAS
NA	700	NA	HECTAREAS
NA	NA	8	FANEGADAS
2	160	1.5	HECTAREAS
NA	NA	NA	HECTAREAS
3	450	5	HECTAREAS
5	NA	6	HECTAREAS
2	600	7	HECTAREAS
NA	NA	NA	HECTAREAS

Errores en los caracteres

VARIEDAD	Manera DE SIEMBR	FECHA DE Germinacio
INTA Fortaleza Secano	Chorrillo	junio
INTA L9	Chorrillo	
INTA Fortaleza Secano	Chorrillo	
INTA L9	Chorrillo	junio
Inta L9	Maquinaria	7/22/2016
Inta L9	Maquinaria	7/18/2017
INTA Dorado	Bueyes	Julio
Inta F. Secano	Bueyes	8/17/2016
INTA L9	Espeque	INTA L9
INTA Fortaleza Secano	Espeque	INTA Fortaleza Secano
Inta San Juan	Se Perdio	
INTAL8,	Espeque	10/10/2016
Inta L9	Espeque	10/12/2016
INTA Chinandega	Chorrillo	10/11/2016
INTA Chinandega	Espeque	10/8/2016
INTA L8, INTA L9 INTA Fortaleza Secano	Chorrillo	No Siguieron para 2017
INTA L9	Chorrillo	6/10/2017
Inta L9	Chorrillo	6/15/2017

Valores atípicos





Algunos criterios para estructura y limpieza de bases de datos

- Colocar en cada fila nuestra unidad de observación y en cada columna la variable.
- Crear un diccionario de datos.
- Verificar que todas los registros de una misma variables tengan la misma unidad de medida.
- Resolver problemas de coordenadas y fechas.
- Resolver problemas de mayúsculas y minúsculas.
- Identificar valores atípicos.

Prácticas en R

Funciones en R

- Funciones **filter, arrange, select** y **mutate** para limpiar bases de datos (**dplyr**)
- Funciones **group_by** y **summarise** para resumir bases de datos (**dplyr**)
- Funciones **pivot** para re-estructurar bases de datos (**tidyr**)
- Funciones **join** para unir base de datos (**dplyr**)

Limpieza de datos

Subconjuntos de Observaciones

dplyr::**filter(iris, Sepal.Length > 7)**
Extrae filas que cumplen criterios lógicos.

dplyr::**distinct(iris)**
Remueve filas duplicadas.

dplyr::**sample_frac(iris, 0.5, replace = TRUE)**
Selecciona una fracción de filas al azar.

dplyr::**sample_n(iris, 10, replace = TRUE)**
Selecciona n filas al azar.

dplyr::**slice(iris, 10:15)**
Selecciona filas por posición.

dplyr::**top_n(storms, 2, date)**
Selecciona y ordena las n entradas mas altas (por grupo si los datos estan agrupados).

Lógica in R - ?Comparison, ?base::Logic

<	Menor de	!=	No equivale a
>	Mayor a	%in%	Membrecia de grupo
==	equivale a	is.na	Es **NA**
<=	Menos o equivalente a	!is.na	No es **NA**
>=	Mayor o equivalente a	&,|,!,xor,any,all	Operadores Booleanos

Subconjuntos de Variables

dplyr::**select(iris, Sepal.Width, Petal.Length, Species)**
Selecciona columnas por nombre o funciones de ayuda.

Funciones de ayuda para for select - ?select

select(iris, **contains(".")**)
Selecciona columnas cuyos nombres contienen una cadena de caracteres.

select(iris, **ends_with("Length")**)
Selecciona columnas cuyos nombres terminan con una cadena de caracteres.

select(iris, **everything()**)
Selecciona todas las columnas.

select(iris, **matches(".t.")**)
Selecciona columnas cuyo nombre cumple con una expresión regular.

select(iris, **num_range("x", 1:5)**)
Selecciona columna con nombres x1, x2, x3, x4, x5.

select(iris, **one_of(c("Species", "Genus"))**)
Selecciona columnas cuyos nombres están en un grupo de nombres.

select(iris, **starts_with("Sepal")**)
Selecciona columnas cuyos nombres comienzan con una cadena de caracteres.

select(iris, **Sepal.Length:Petal.Width**)
Selecciona todas las columnas entre Sepal.Length and Petal.Width (incluyente).

select(iris, **-Species**)
Selecciona todas las columnas excepto Species.

Resumir Datos

dplyr::**summarise(iris, avg = mean(Sepal.Length))**
Resume datos a una sola fila de valores.

dplyr::**summarise_each(iris, funs(mean))**
Aplica la función *summary* a cada columna.

dplyr::**count(iris, Species, wt = Sepal.Length)**
Cuenta el número de valores únicos para cada variable (con o sin ponderación).

Group Data

dplyr::**group_by(iris, Species)**
Agrupo datos en filas por los valores en Species.

dplyr::**ungroup(iris)**
Remueve la agrupación del data frame.

iris %>% group_by(Species) %>% summarise(…)
Calcula una fila separada con el resumen para cada grupo.

https://github.com/rstudio/cheatsheets/raw/master/translations/spanish/data-wrangling-cheatsheet_Spanish.pdf

Pivot

Pivot a lo largo

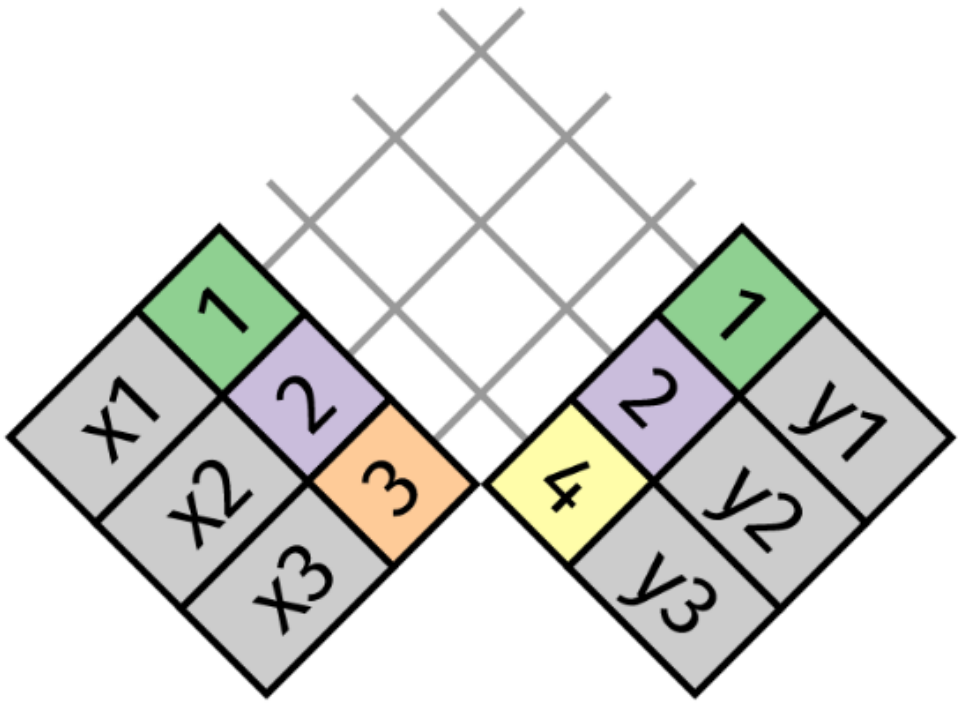
```
# A tibble: 286 x 4
   Nommunicipio Nommodalidad Variable       value
   <chr>        <chr>        <chr>          <dbl>
 1 HopelchÃ©n  Temporal     Area         25044.
 2 HopelchÃ©n  Temporal     Rendimiento      2.08
 3 Sacramento   Riego        Area           27.5
 4 Sacramento   Riego        Rendimiento     3.1
 5 Teopisca     Riego        Area          501
 6 Teopisca     Riego        Rendimiento     3.25
 7 Teopisca     Temporal     Area         3329
 8 Teopisca     Temporal     Rendimiento     1.5
 9 PantelhÃ³    Temporal     Area         2432
10 PantelhÃ³    Temporal     Rendimiento     1.27
# ... with 276 more rows
```

Pivot a lo ancho

```
# A tibble: 143 x 4
   Nommunicipio              Nommodalidad    Area Rendimiento
   <chr>                     <chr>          <dbl>       <dbl>
 1 HopelchÃ©n               Temporal     25044.          2.08
 2 Sacramento              Riego           27.5         3.1
 3 Teopisca                Riego          501           3.25
 4 Teopisca                Temporal      3329           1.5
 5 PantelhÃ³               Temporal      2432           1.27
 6 Reforma                 Temporal      1755           1.37
 7 Pantepec                Temporal       955           1.08
 8 Amatenango de La Frontera Temporal     3940           1.77
 9 Hidalgo del Parral      Temporal       110.          1.64
10 San Dimas               Temporal      5000           0.8
# ... with 133 more rows
```

Uniones

- Una unión es una forma de conectar cada fila en x con cero, una o más filas en y

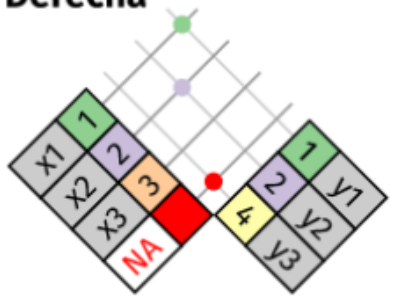


Uniones de transformación

Izquierda

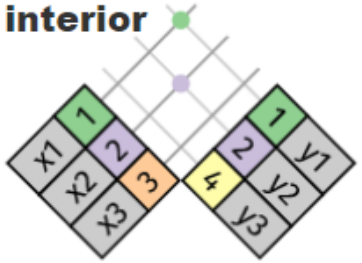
llave	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

Derecha



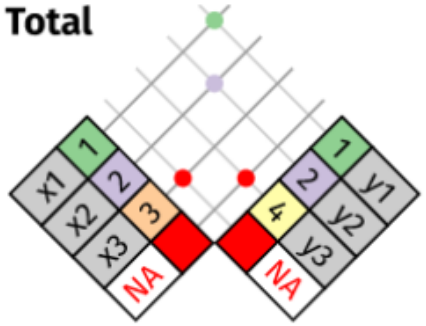
llave	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Unión interior



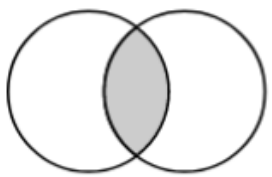
llave	val_x	val_y
1	x1	y1
2	x2	y2

Total

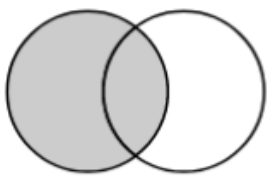


llave	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

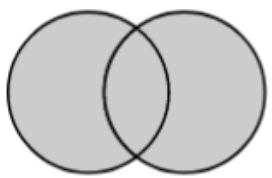
Uniones de transformación



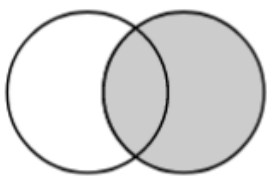
Unión interior
`inner_join(x,y)`



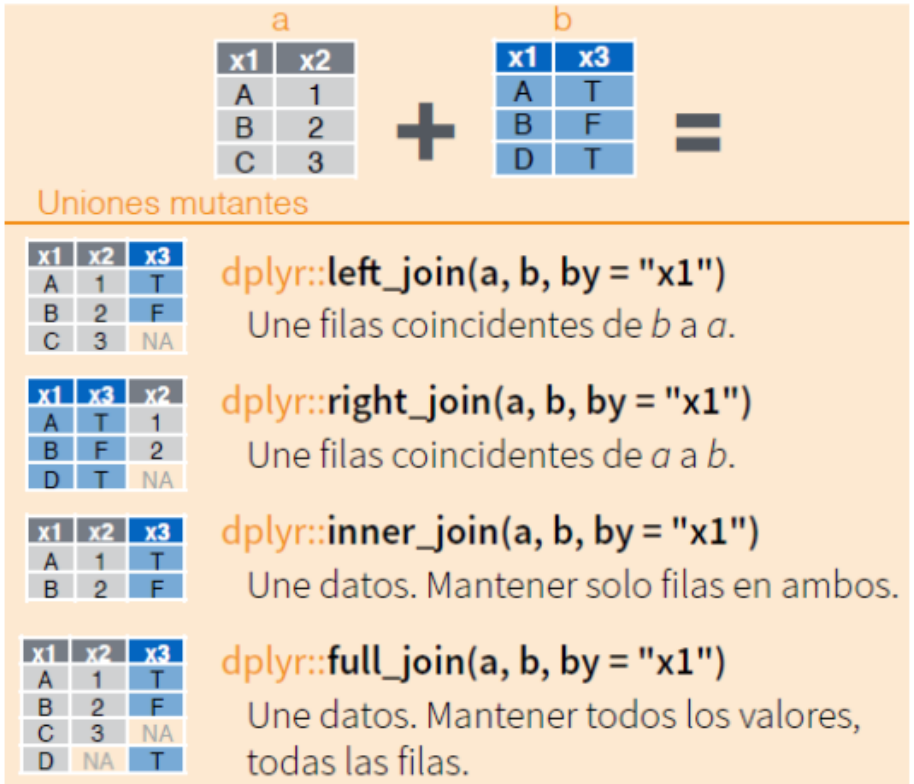
Unión por la izquierda
`left_join(x,y)`



Unión total
`full_join(x,y)`

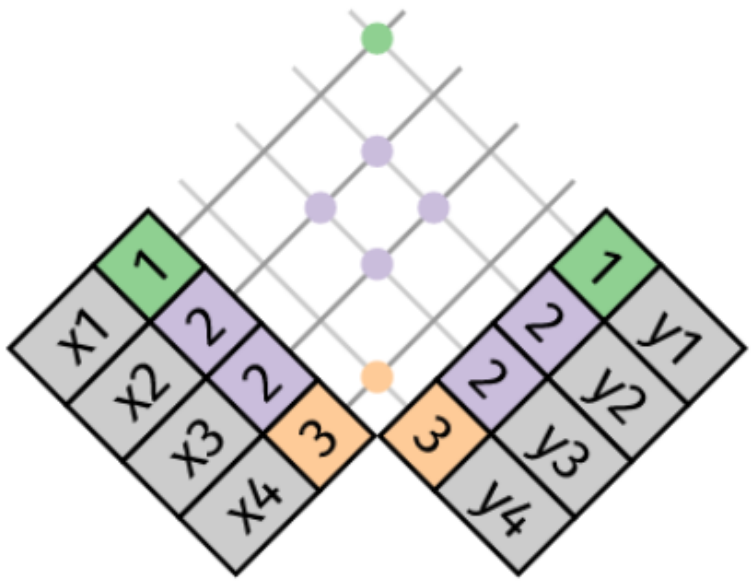


Unión por la derecha
`right_join(x,y)`



Claves duplicadas

- Cuando unes claves duplicadas, se obtienen todas las posibles combinaciones, es decir, el producto cartesiano



llave	val_x	val_y
1	x1	y1
2	x2	y2
2	x2	y3
2	x3	y2
2	x3	y3
3	x4	y4

Preguntas

- ¿De las bases de datos que ha procesado, utiliza criterios para descartar variables antes de analizar SI/NO?, ¿cuales?
- ¿Aplica alguna transformación o sumarización a alguna de las variables que analiza SI/NO?, ¿cuales?

- ¿Siempre que organiza una base de datos crea un diccionario de datos SI/NO?
- ¿Cuales son los errores más frecuente que detecta en las bases de datos que analiza?

- ¿Ejecutó los scripts de las prácticas con dplyr y tidyverse SI/NO?, coméntenos las experiencias. (dudas, dificultades, descubrimientos)

Gracias!

Hugo Andrés Dorado.

Científico de datos

hugo.doradob@gmail.com

Conocimiento generado a partir de proyectos de:

Alianza

Bioversity International

CIAT
Centro Internacional de Agricultura Tropical
Desde 1967 Ciencia para cultivar el cambio

CLIMATE CHANGE AGRICULTURE AND FOOD SECURITY