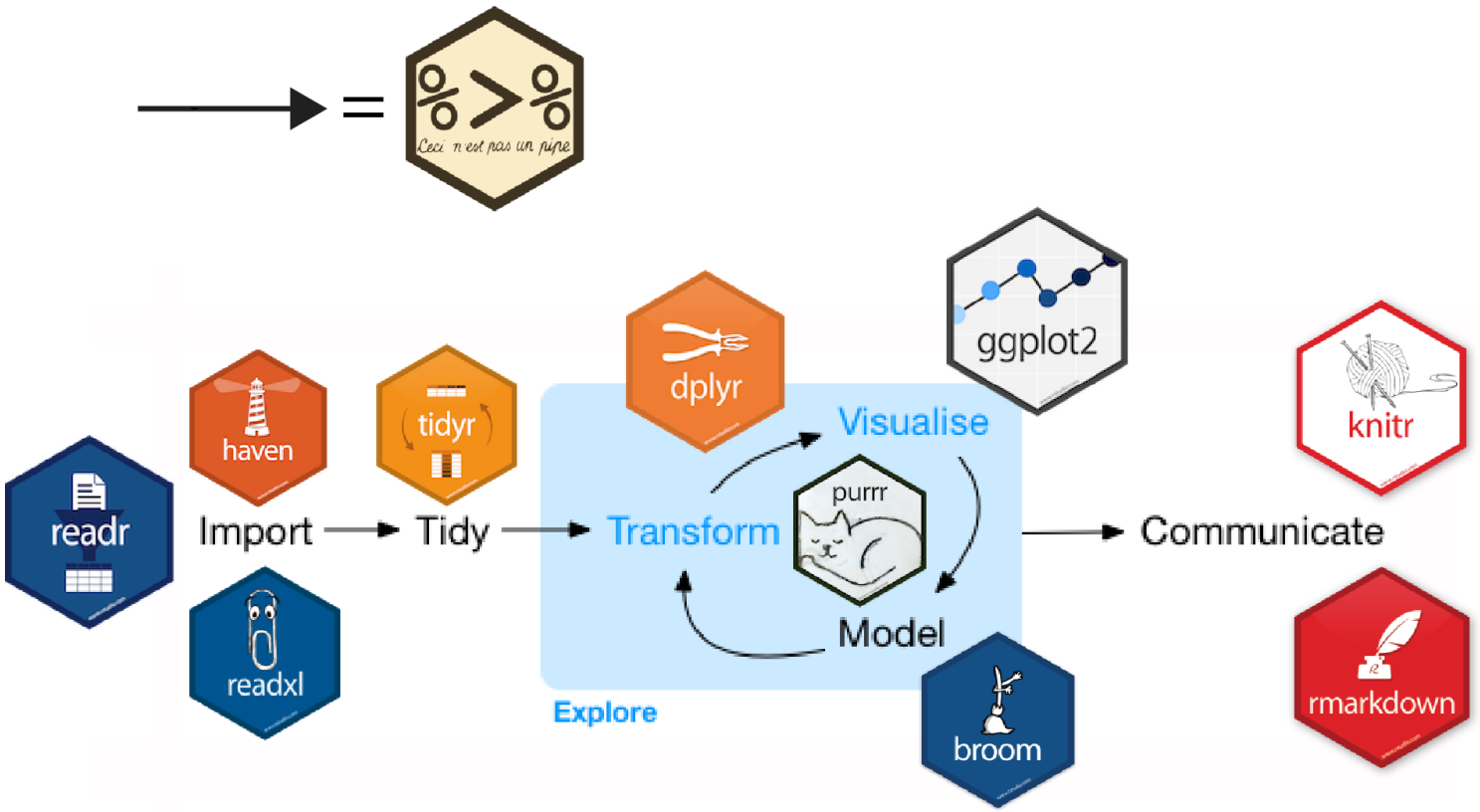
Ordenando datos con R

Usos y ejemplos

Mariel Lovatto
ma.lvtto@gmail.com

2018-11-26

Contenido



data.frames factors strings

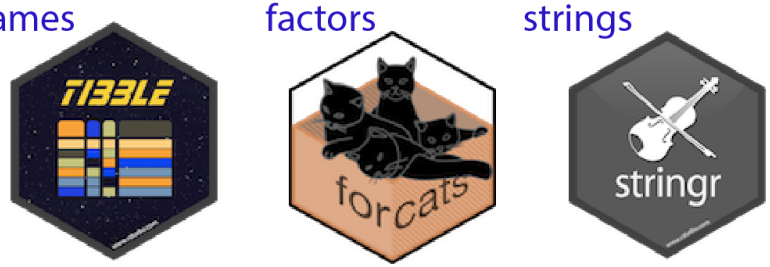


Figure 1: Tidyverse

Tidy Data

1. Cada columna representa una variable
2. Cada fila representa una observación.
3. En cada celda hay un valor de la variable para una observación.

Messy Data

Los encabezados de columna son valores, no nombres de variables

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
Agnostic	84	34	109	60	81	76	137
Atheist	74	27	59	37	52	35	70
Buddhist	53	21	39	30	34	33	58
Catholic	633	617	792	732	670	638	1116
Don't know/refused	18	14	17	15	11	10	35
Evangelical Prot	414	869	723	1064	982	881	1486
Hindu	54	9	48	7	9	11	34
Historically Black Prot	78	244	81	236	238	197	223
Jehovah's Witness	6	27	11	24	24	21	30
Jewish	151	19	87	25	25	30	95

Messy Data

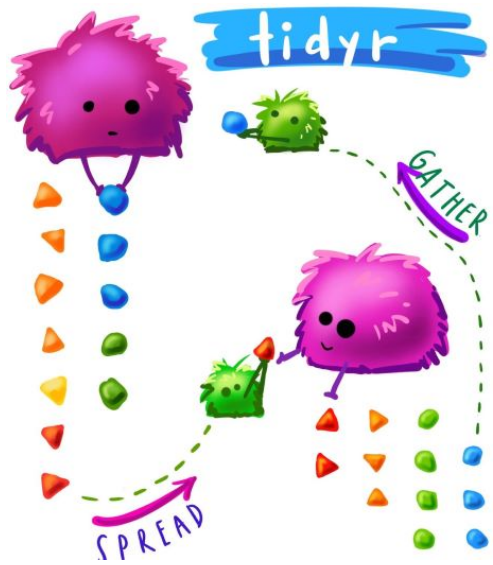
Los encabezados de columna son valores, no nombres de variables

```
nombres <- names(raw)[-1]

raw2 <- raw %>%
  gather(nombres
         , key = "income"
         , value = "counts") %>%
  arrange(religion)

raw22 <- raw2[rep(1:nrow(raw2), raw2[,3]),1:2]
```

religion	income
Mainline Prot	$30-40k
Mainline Prot	$20-30k
Mainline Prot	<$10k
Mainline Prot	$40-50k
Unaffiliated	$20-30k
Unaffiliated	<$10k
Catholic	>150k
Unaffiliated	$30-40k

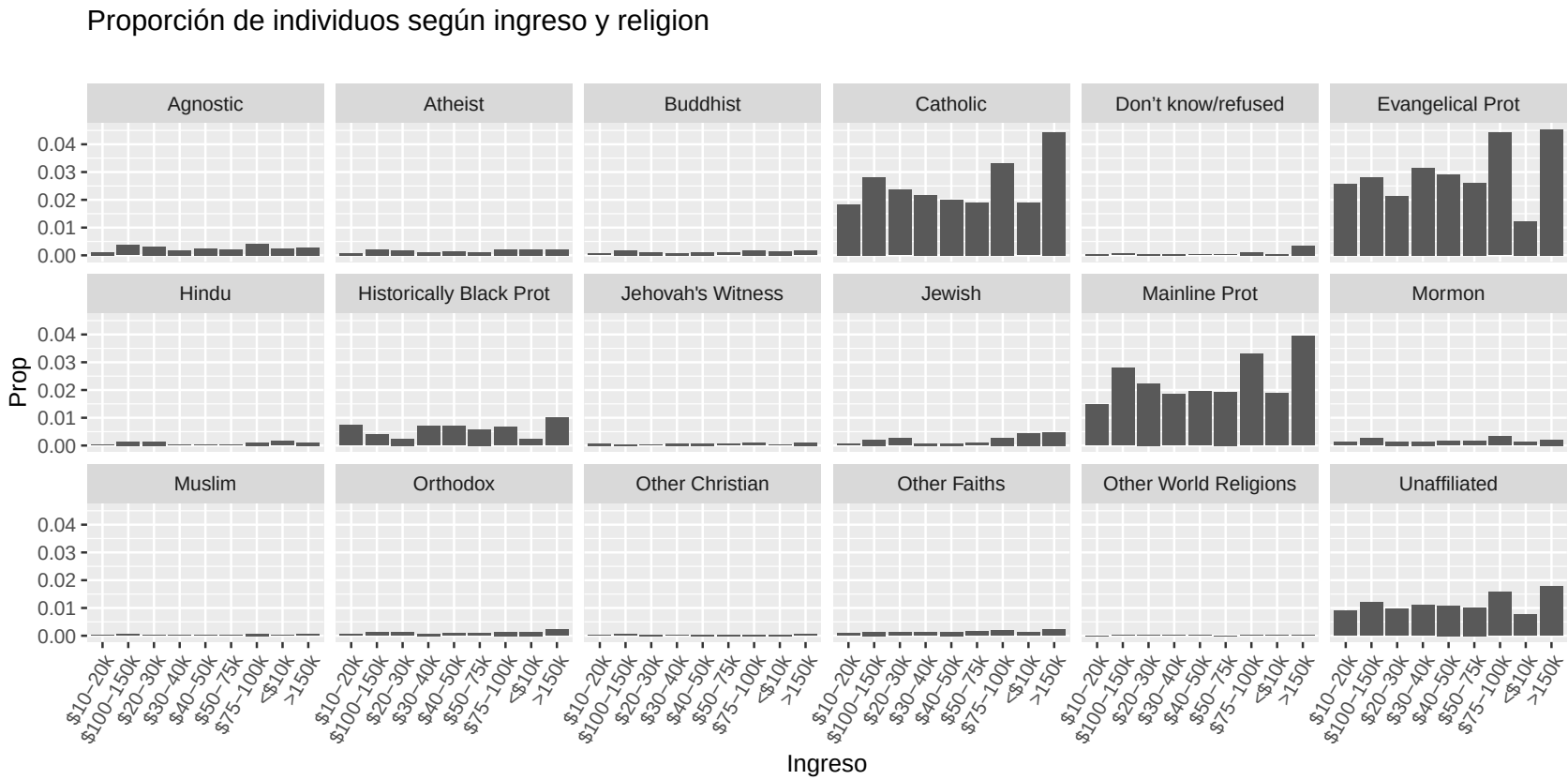


twitter.com/allison_horst

Visualización

```
p <- ggplot(raw22, aes(x = income))+
  geom_bar(aes(y = ..count../sum(..count..))) +
  facet_wrap(~religion, nrow = 3) +
  theme(axis.text.x=element_text(angle=60, hjust=1)) +
    labs(title = "Proporción de individuos según ingreso y religion", x = "Ingreso", y = "Prop",
         subtitle = "")


p
```



Messy Data

Múltiples variables se almacenan en una columna

country	year	m014	m1524	m2534	f014	f1524	f2534	f3544	f4554
AD	2000	0	0	1	NA	NA	NA	NA	NA
AE	2000	2	4	4	3	16	1	3	0
AF	2000	52	228	183	93	414	565	339	205
AG	2000	0	0	0	1	1	1	0	0
AL	2000	2	19	21	3	11	10	8	8
AM	2000	2	152	130	1	24	27	24	8
AN	2000	0	0	1	0	0	1	0	0
AO	2000	186	999	1003	247	1142	1091	844	417
AR	2000	97	278	594	121	544	479	262	230
AS	2000	NA	NA	NA	NA	NA	NA	NA	1

Messy Data

Múltiples variables se almacenan en una columna

```
nombres <- names(raw)
tb <- raw %>%
  gather(nombres[-c(1,2)]
         , key = "column"
         , value = "cases"
         , na.rm = TRUE) %>%
  arrange(country, column, year)

tb$sex <- str_sub(tb$column, 1, 1)

ages <- c("04" = "0-4", "514" = "5-14", "014" = "0-14",
          "1524" = "15-24", "2534" = "25-34", "3544" = "35-44",
          "4554" = "45-54", "5564" = "55-64", "65"= "65+", "u" = NA)

tb$age <- factor(ages[str_sub(tb$column, 2)], levels = ages)

tb <- tb[c("country", "year", "sex", "age", "cases")]
```

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	f	0-14	3

Messy Data

Las variables se almacenan en filas y columnas

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	1	tmin	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	2	tmax	NA	27.3	24.1	NA	NA	NA	NA	NA
MX17004	2010	2	tmin	NA	14.4	14.4	NA	NA	NA	NA	NA
MX17004	2010	3	tmax	NA	NA	NA	NA	32.1	NA	NA	NA
MX17004	2010	3	tmin	NA	NA	NA	NA	14.2	NA	NA	NA
MX17004	2010	4	tmax	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	4	tmin	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	5	tmax	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	5	tmin	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	6	tmax	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	6	tmin	NA	NA	NA	NA	NA	NA	NA	NA
MX17004	2010	7	tmax	NA	NA	28.6	NA	NA	NA	NA	NA
MX17004	2010	7	tmin	NA	NA	17.5	NA	NA	NA	NA	NA
MX17004	2010	8	tmax	NA	NA	NA	NA	29.6	NA	NA	29.0
MX17004	2010	8	tmin	NA	NA	NA	NA	15.8	NA	NA	17.3
MX17004	2010	10	tmax	NA	NA	NA	NA	27.0	NA	28.1	NA
MX17004	2010	10	tmin	NA	NA	NA	NA	14.0	NA	12.9	NA
MX17004	2010	11	tmax	NA	31.3	NA	27.2	26.3	NA	NA	NA
MX17004	2010	11	tmin	NA	16.3	NA	12.0	7.9	NA	NA	NA
MX17004	2010	12	tmax	29.9	NA	NA	NA	NA	27.8	NA	NA
MX17004	2010	12	tmin	13.8	NA	NA	NA	NA	10.5	NA	NA

Messy Data

Las variables se almacenan en filas y columnas

```
nombres <- names(raw)

tb <- raw %>%
  gather(nombres[-c(1,2,3,4)]
         , key = "variable"
         , value = "value"
         , na.rm = TRUE)

tb$day <- as.integer(str_replace(tb$variable, "d", ""))

tb$date <- as.Date(ISOdate(tb$year, tb$month, tb$day))

tb <- tb[c("id", "date", "element", "value")]

tb <- arrange(tb, date, element)
```

Messy Data

Las variables se almacenan en filas y columnas

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7
MX17004	2010-03-05	tmax	32.1
MX17004	2010-03-05	tmin	14.2
MX17004	2010-03-10	tmax	34.5
MX17004	2010-03-10	tmin	16.8
MX17004	2010-03-16	tmax	31.1
MX17004	2010-03-16	tmin	17.6
MX17004	2010-04-27	tmax	36.3
MX17004	2010-04-27	tmin	16.7
MX17004	2010-05-27	tmax	33.2
MX17004	2010-05-27	tmin	18.2
MX17004	2010-06-17	tmax	28.0
MX17004	2010-06-17	tmin	17.5
MX17004	2010-06-29	tmax	30.1
MX17004	2010-06-29	tmin	18.0
MX17004	2010-07-03	tmax	28.6

twitter.com/allison_horst

```
tb1 <- tb %>%
  spread(tb[,c(3,4)]
         , key = "element"
         , value = "value"
         , convert = TRUE)

tb2 <-  mutate(tb1,date= as.character(date))
```

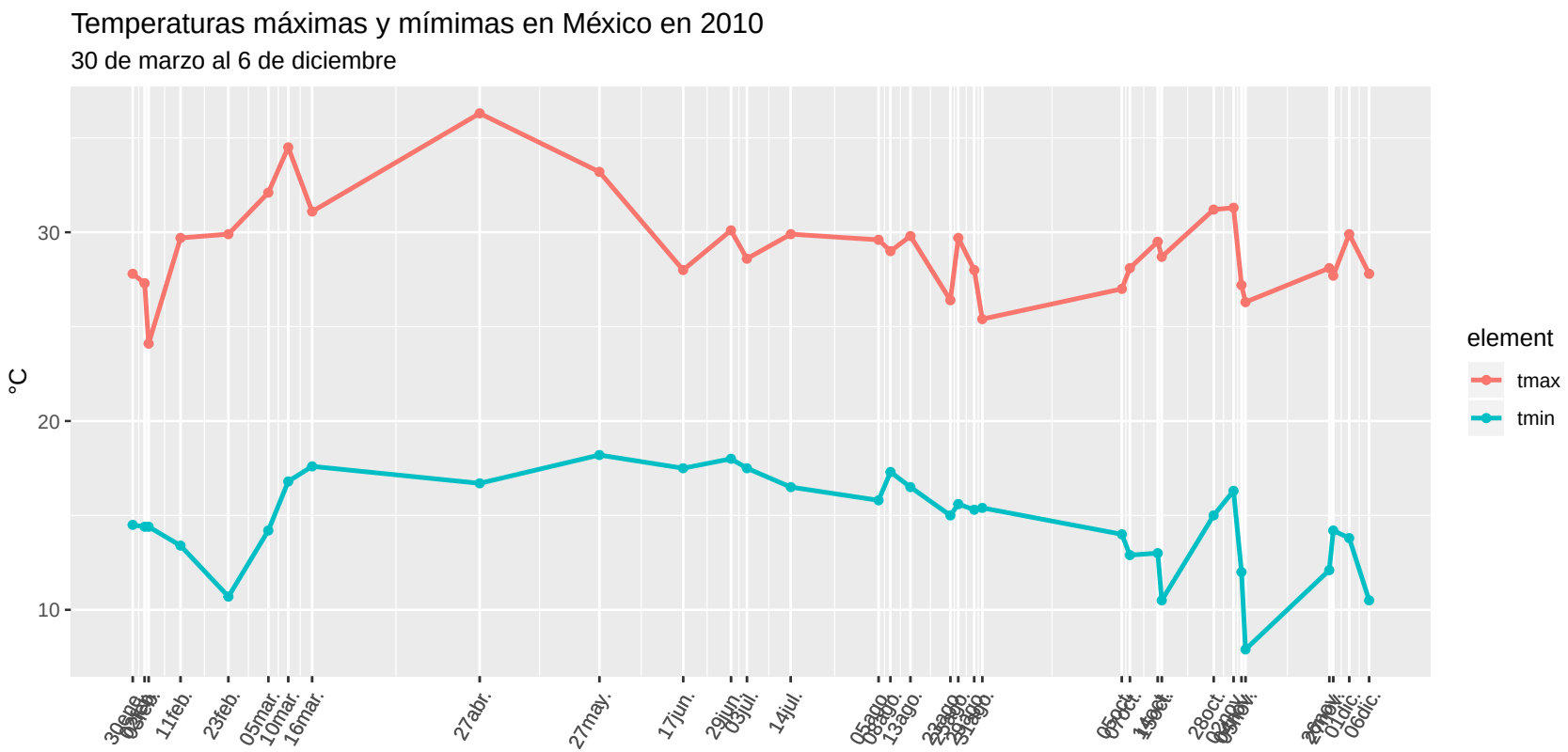
id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2
MX17004	2010-06-17	28.0	17.5
MX17004	2010-06-29	30.1	18.0
MX17004	2010-07-03	28.6	17.5
MX17004	2010-07-14	29.9	16.5
MX17004	2010-08-05	29.6	15.8

twitter.com/allison_horst

Visualización

```
P <- tb %>%
    ggplot(aes(x = date, y = value, color = element)) +
    geom_line(size = 1) + geom_point()+
    labs(title = "Temperaturas máximas y mímimas en México en 2010", x = "", y = "°C",
         subtitle = "30 de marzo al 6 de diciembre")


P + scale_x_date(date_labels = "%d%b", breaks = tb$date)+
  theme(axis.text.x=element_text(angle=60, hjust=1))
```



Messy Data

Una unidad de análisis se almacena en varias tablas

id	nroCuestionario	Domicilio	Fracción	Radio	a71Trabajo	estrato	clase
3912	19	Castelli 871	7	8	1	5	3
3908	36	Av. Gral. Paz 6620	7	9	1	3	2
4790	41	Pavón 697	7	2	1	2	2
1358	43	Piedras 7149	7	5	1	7	3
3164	60	Pje. Cervantes 4075	26	4	1	8	3
4050	79	Mendoza 4552	33	5	1	7	3
3596	89	Estrada 2400	33	5	1	7	3
4585	97	Roque Sáenz Peña 3060	20	8	1	4	2
1613	150	Marcial Candioti 6932	6	14	1	7	3
1368	151	Sarmiento 7435	6	9	1	9	4
4165	176	Ayacucho 1969	6	4	1	4	2
718	182	Javier de la Rosa 3198	6	12	1	4	2
1388	184	Javier de la Rosa 3120	6	12	1	2	2
4155	189	San Jerónimo 7952	31	14	1	5	3
3202	201	Santiago de Chile 1051	26	17	1	4	2
3315	205	Pje. Mitre 1755	26	1	1	5	3
3760	209	Pje. Mitre 1759	26	1	1	5	3
4064	217	Juan Díaz de Solís 1297	26	4	2	11	4
2611	221	San Jerónimo 7953	31	14	1	1	1
2669	243	Mendoza 3565	21	11	1	4	2

Frac2010	Rad2010	geometry
05	01	list(c(5428461, 5428462.5, 5428462.5, 5428703.5, 5428805.5, 5428871.5, 5429340, 5429447.5, 5429473, 5429529, 5429659, 5429659, 5429643.5, 5429630, 5429603.5, 5429574.5, 5429560.5, 5429549.5, 5429549.5, 5429504, 5429476.5, 5429476.5, 5429476.5, 5429476, 5429476.5, 5429477, 5429478, 5429478, 5429461.5, 5429461.5, 5429387.5, 5429333.5, 5429302.5, 5429180, 5429155, 5429099, 5429046, 5428993, 5428993, 5428915, 5428901.5, 5428828.5, 5428788, 5428740.5, 5428712, 5428680, 5428641.5, 5428641.5, 5428598.5, 5428550, 5428500.5, 5428459.5, 5428287, 5428251.5, 5428172, 5428131.5, 5427991.5, 5427991.5, 5427989.5, 5427989.5, 5427991, 5428000.5, 5428011.5, 5428025.5, 5428038.5, 5428051.5, 5428063.5, 5428075.5, 5428085, 5428095, 5428107.5, 5428102, 5428102, 5428101, 5428101, 5427663, 5427596, 5427509.5, 5427465, 5427292.5, 5427052.5, 5426813.5, 5426746.5, 5426677.5, 5426522, 5426249, 5426249, 5426256.5, 5426261.5, 5426274.5, 5426279.5, 5426279.5, 5426284.5, 5426291, 5426321, 5426382.5, 5426432, 5426458.5, 5426491.5, 5426529.5, 5426562.5, 5426584, 5426594, 5426607.5, 5426629, 5426665, 5426704, 5426727, 5426757.5, 5426779.5, 5426783.5, 5426785, 5426782, 5426770.5, 5426763.5, 5426744.5, 5426729.5, 5426713, 5426694.5, 5426666, 5426647.5, 5426640.5, 5426640.5, 5426650.5, 5426678, 5426720, 5426768, 5426812.5, 5426862, 5426933.5, 5426979.5, 5427024.5, 5427064, 5427107.5, 5427170, 5427214.5, 5427225.5, 5427225.5, 5427837, 5428012.5, 5428165, 5428461, 6509093.5, 6509119.5, 6509119.5, 6509058, 6509029, 6509010, 6508878, 6508844.5, 6508833, 6508816, 6508781, 6508781, 6508733.5, 6508686, 6508588.5, 6508492.5, 6508441.5, 6508393.5, 6508393.5, 6508406, 6508413, 6508413, 6508341, 6508286.5, 6508229.5, 6508173.5, 6508109, 6508109, 6508113.5, 6508113.5, 6508132, 6508145, 6508153, 6508184, 6508191, 6508205, 6508217.5, 6508231, 6508231, 6508252.5, 6508254.5, 6508275, 6508281, 6508278.5, 6508276.5, 6508271.5, 6508262, 6508262, 6508251.5, 6508249.5, 6508252.5, 6508261, 6508304.5, 6508314.5, 6508339.5, 6508348.5, 6508388, 6508388, 6508363.5, 6508345, 6508330, 6508315, 6508302, 6508282, 6508265, 6508245.5, 6508227, 6508205.5, 6508182.5, 6508152.5, 6508086.5, 6508022, 6508022, 6508004, 6508004, 6508314.5, 6508356, 6508401, 6508418.5, 6508469, 6508490, 6508496.5, 6508507, 6508521, 6508591.5, 6508776, 6508776, 6508799.5, 6508818, 6508880.5, 6508938.5, 6509008, 6509046, 6509080.5, 6509104, 6509115.5, 6509079, 6509033, 6508980, 6508897, 6508806, 6508690.5, 6508644, 6508617.5, 6508599.5, 6508589.5, 6508593, 6508626.5, 6508678.5, 6508738, 6508784.5, 6508839.5, 6508895, 6508964.5, 6509006.5, 6509081.5, 6509115, 6509143, 6509159, 6509181, 6509205, 6509225, 6509262.5, 6509281.5, 6509303, 6509306, 6509296, 6509283, 6509253, 6509211.5, 6509163.5, 6509152, 6509153.5, 6509179, 6509286, 6509390.5, 6509433, 6509433, 6509274.5, 6509219.5, 6509177.5, 6509093.5))

Messy Data

Una unidad de análisis se almacena en varias tablas

```
names(dat15)[4] <- "Frac2010"
radios$Frac2010 <- as.numeric(radios$Frac2010)
datos <- left_join(radios, dat15)
```

Frac2010	Domicilio	a71Trabajo	clase	geometry
5	Formosa 6421	1	3	list(c(5428461, 5428462.5, 5428462.5, 5428703.5, 5428805.5, 5428871.5, 5429340, 5429447.5, 5429473, 5429529, 5429659, 5429659, 5429643.5, 5429630, 5429603.5, 5429574.5, 5429560.5, 5429549.5, 5429549.5, 5429504, 5429476.5, 5429476.5, 5429476.5, 5429476, 5429476.5, 5429477, 5429478, 5429478, 5429461.5, 5429461.5, 5429387.5, 5429333.5, 5429302.5, 5429180, 5429155, 5429099, 5429046, 5428993, 5428993, 5428915, 5428901.5, 5428828.5, 5428788, 5428740.5, 5428712, 5428680, 5428641.5, 5428641.5, 5428598.5, 5428550, 5428500.5, 5428459.5, 5428287, 5428251.5, 5428172, 5428131.5, 5427991.5, 5427991.5, 5427989.5, 5427989.5, 5427991, 5428000.5, 5428011.5, 5428025.5, 5428038.5, 5428051.5, 5428063.5, 5428075.5, 5428085, 5428095, 5428107.5, 5428102, 5428102, 5428101, 5428101, 5427663, 5427596, 5427509.5, 5427465, 5427292.5, 5427052.5, 5426813.5, 5426746.5, 5426677.5, 5426522, 5426249, 5426249, 5426256.5, 5426261.5, 5426274.5, 5426279.5, 5426279.5, 5426284.5, 5426291, 5426321, 5426382.5, 5426432, 5426458.5, 5426491.5, 5426529.5, 5426562.5, 5426584, 5426594, 5426607.5, 5426629, 5426665, 5426704, 5426727, 5426757.5, 5426779.5, 5426783.5, 5426785, 5426782, 5426770.5, 5426763.5, 5426744.5, 5426729.5, 5426713, 5426694.5, 5426666, 5426647.5, 5426640.5, 5426640.5, 5426650.5, 5426678, 5426720, 5426768, 5426812.5, 5426862, 5426933.5, 5426979.5, 5427024.5, 5427064, 5427107.5, 5427170, 5427214.5, 5427225.5, 5427225.5, 5427837, 5428012.5, 5428165, 5428461, 6509093.5, 6509119.5, 6509119.5, 6509058, 6509029, 6509010, 6508878, 6508844.5, 6508833, 6508816, 6508781, 6508781, 6508733.5, 6508686, 6508588.5, 6508492.5, 6508441.5, 6508393.5, 6508393.5, 6508406, 6508413, 6508413, 6508341, 6508286.5, 6508229.5, 6508173.5, 6508109, 6508109, 6508113.5, 6508113.5, 6508132, 6508145, 6508153, 6508184, 6508191, 6508205, 6508217.5, 6508231, 6508231, 6508252.5, 6508254.5, 6508275, 6508281, 6508278.5, 6508276.5, 6508271.5, 6508262, 6508262, 6508251.5, 6508249.5, 6508252.5, 6508261, 6508304.5, 6508314.5, 6508339.5, 6508348.5, 6508388, 6508388, 6508363.5, 6508345, 6508330, 6508315, 6508302, 6508282, 6508265, 6508245.5, 6508227, 6508205.5, 6508182.5, 6508152.5, 6508086.5, 6508022, 6508022, 6508004, 6508004, 6508314.5, 6508356, 6508401, 6508418.5, 6508469, 6508490, 6508496.5, 6508507, 6508521, 6508591.5, 6508776, 6508776, 6508799.5, 6508818, 6508880.5, 6508938.5, 6509008, 6509046, 6509080.5, 6509104, 6509115.5, 6509079, 6509033, 6508980, 6508897, 6508806, 6508690.5, 6508644, 6508617.5, 6508599.5, 6508589.5, 6508593, 6508626.5, 6508678.5, 6508738, 6508784.5, 6508839.5, 6508895, 6508964.5, 6509006.5, 6509081.5, 6509115, 6509143, 6509159, 6509181, 6509205, 6509225, 6509262.5, 6509281.5, 6509303, 6509306, 6509296, 6509283, 6509253, 6509211.5, 6509163.5, 6509152, 6509153.5, 6509179, 6509286, 6509390.5, 6509433, 6509433, 6509274.5, 6509219.5, 6509177.5, 6509093.5))

Visualización

```
da1 <-  datos %>%
  group_by(Rad2010, Frac2010) %>%
  summarise(clase1 = sum(clase==1)/length(clase),
            clase2 = sum(clase==2)/length(clase),
            clase3 = sum(clase==3)/length(clase),
            clase4 = sum(clase==4)/length(clase))


p1 <- ggplot()+ geom_sf(data=da1, aes(fill=clase1))+
  scale_fill_viridis_c()

p2 <- ggplot()+ geom_sf(data=da1, aes(fill=clase2))+
  scale_fill_viridis_c()

p3 <- ggplot()+ geom_sf(data=da1, aes(fill=clase3))+
  scale_fill_viridis_c()

p4 <- ggplot()+ geom_sf(data=da1, aes(fill=clase4))+
  scale_fill_viridis_c()

ggarrange(p1,p2,p3,p4)
```

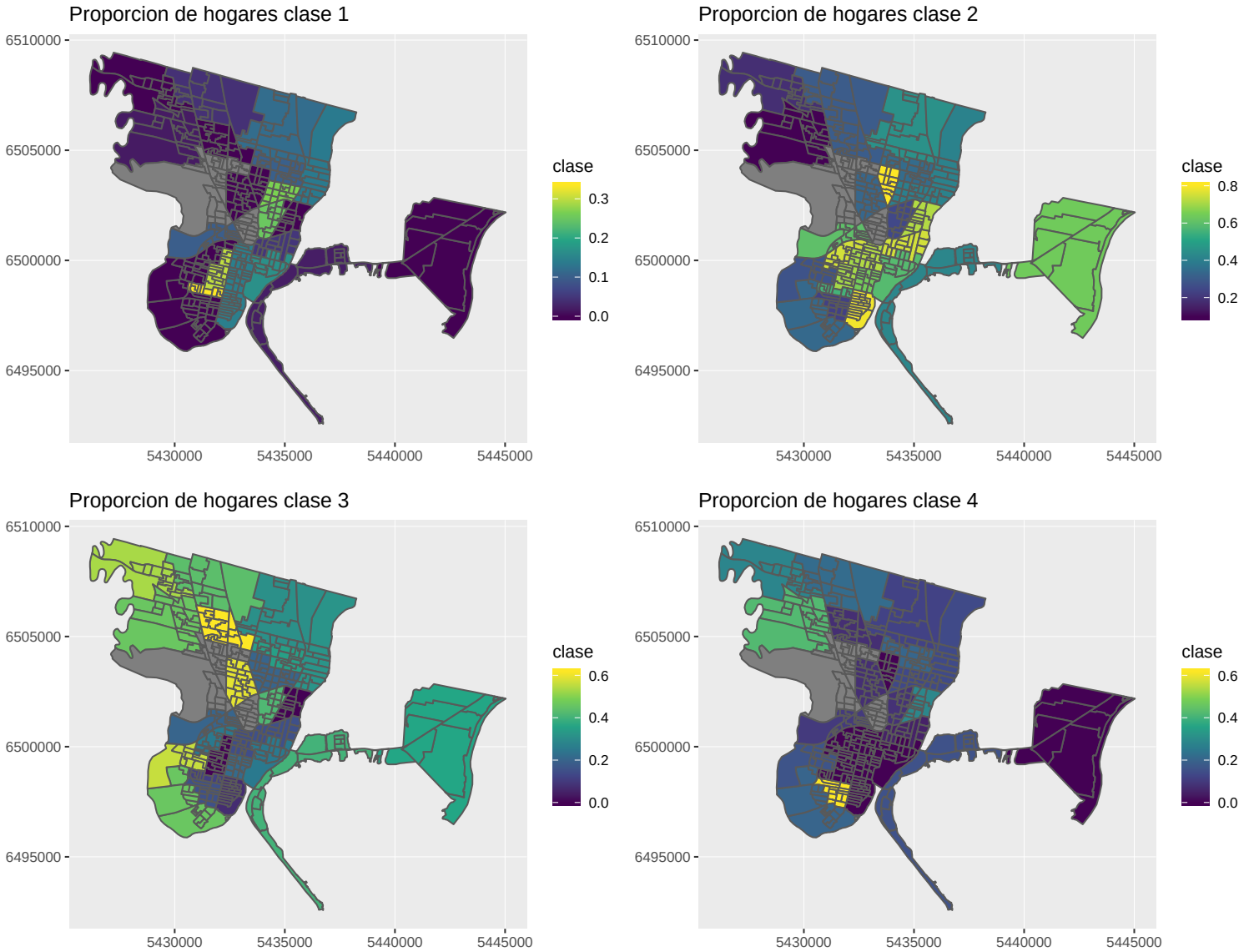


Figure 2: Proporción de hogares en función de la clase

Conclusiones

Funciones y paquetes utilizados

base	dplyr	stringr	tidyr	ggplot2
as.integer	filter	str_sub	gather	ggsf
as.Date	arrange	str_replace	spread	ggplot
as.numeric	%>%			geom_bar
as.character	mutate			geom_path
ISOdate	summarize			geom_point
	group_by			

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

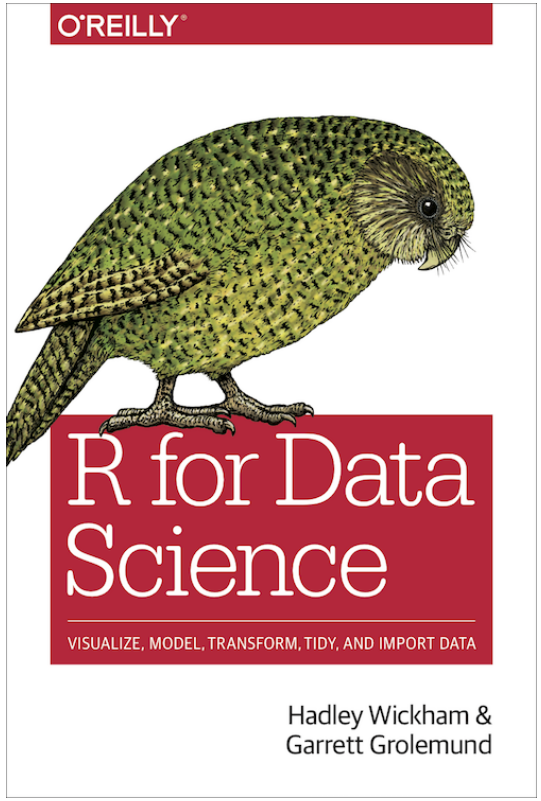
Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.



GRACIAS

twitter.com/allison_horst