R: manipulación de datos

Un proyecto de datos tiene una gran cantidad de componentes. Sin embargo, en básicamente todos se necesita iterar sobre el ciclo que se muestra en la figura 1.

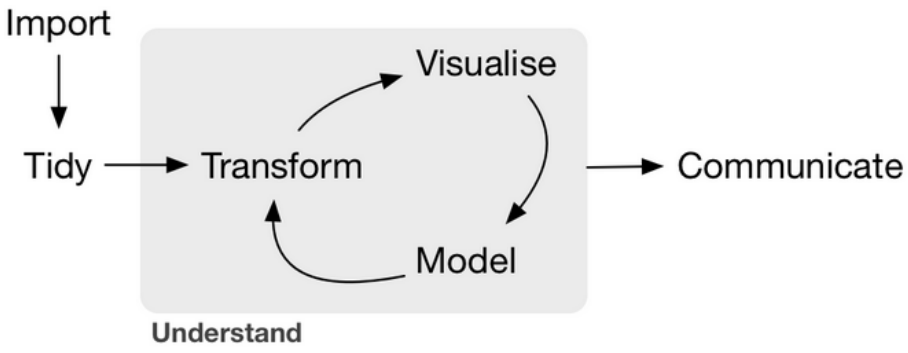

Figura 1: Modelo de las herramientas que se necesitan en un proyecto de datos según **grolemund2016r**

Primero es necesario **importar** nuestros datos a R. Los datos pueden estar en una gran cantidad de formatos o lugares.

Después, normalmente es necesario **arreglar** nuestros datos, es decir, seguir criterios de datos limpios de manera que la manera en la que guardemos los datos equivalga a la semántica de los datos que tenemos. Es muy importante primero limpiar porque esto provee de consistencia a lo largo del análisis.

Posteriormente, en casi todo proyecto, será necesario **transformar** los datos. A veces esto implica enfocarse en un subconjunto de los datos, generar nuevas variables, calcular estadísticos, arreglar los datos de cierta manera, entre muchos otros.

Solamente después de estas etapas podemos empezar a generar conocimiento a partir de los datos. Para esto tenemos dos herramientas fundamentales: la estadística descriptiva (en el diagrama reducido a **visualización**) y la generación de **modelos**. La primera es fundamental pues permite derivar preguntas pertinentes a los datos, encontrar patrones, respuestas, plantear hipótesis. Sin embargo, éstas no escalan de la misma manera que los modelos pues estos, una vez que aceptamos sus supuestos generan los resultados que esperamos o contestan la pregunta planteada.

Por último, necestamos **comunicar** los resultados.

Datos limpios

Mucho del esfuerzo en analítica lidia con la limpieza de datos. Tomar datos de diferentes fuentes y poderlas poner en la forma en la que uno los necesita para realizar analítica toma mucho tiempo y esfuezo. Existen herramientas que permiten que esta parte sea más fácil y eficiente. Entre éstas se encuentran los criterios de datos limpios.

Los conjuntos de datos limpios (*tidy datasets*) permiten manipularlos fácilmente, modelarlos y visualizarlos. Además, tienen una estructura específica: cada variable es una columna, cada observación una fila y cada tipo de unidad observacional es una tabla.

Preparación de datos

Esta actividad incluye una gran cantidad de elementos: desde revisar los outliers, hasta extraer variables de cadenas en datos no estrucutrados, imputación de valores perdidos. Los datos limpios son tan solo un subconjunto de este proceso y lidian como el cómo estructurar los datos de manera que se facilite el análisis.

El estándar de datos limpios está diseñado para facilitar la exploración inicial y el análisis de datos así como simplificar el desarrollo de herramientas para el anásis de datos que trabajen bien con datos limpios.

Los criterios de datos limpios están muy relacionados a los de las bases de datos relacionales y, por ende, al algebra relacional de Codd. Sin embargo, se expresan y enmarcan en lenguaje que le es familiar a estadísticos.

Básicamente, están creados para lidear con conjuntos de datos que se encuentran en el mundo real. Los criterios de datos limpios proporcionan un marco mental a través del cual la intuición es explícita.

Definición de datos limpios

Los datos limpios proporcionan una manera estándar de ligar la estructura de un dataset (es decir su layout físico) con su semántica (su significado).

Estructura de datos

La mayoría de los datos estadísticos están conformados por tablas rectangulares compuestas por filas y columnas. Las columnas casi siempre están etiquetadas *colnames* y las filas a veces lo están.

Tomamos el ejemplo de datos de la figura 2 en donde se presentan datos de un experimento. La tabla contiene dos columas y tres filas, ambas etiquetaadas.

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Figura 2: Típica presentación de datos.

Podemos estructurar los datos de diferentes manera pero la abstracción de filas y columnas solamnete nos permite pensar en la representación transpuesta que se muestra en la figura 3. El layout cambia pero los datos son los mismos. Con columnas y filas, no podemos decir esto de manera apropiada. Además de la simple apariencia, debemos poder describir la semántica -el significado- de los valores que se muestran en una tabla.

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Figura 3: Mismos datos que en 2 pero traspuestos.

Semántica

Un conjunto de datos es una colección de **valores** (normalmente cuantitativos/números o cualitativos/caracteres).

Los valores se organian de dos maneras. Cada valor pertenece a una variable y a una observación. Una variable contiene todos los valores de una medida y del mismo atributo subyacente (por ejemplo, temperatura, duración, altura, latitud) a través de unidades. Una observación, en cambio, contiene todos los valores medidos para la misma unidad (por ejemplo, una persona, un día, un municipio) a través de distintos atributos.

Los mismos datos en las figuras 2 y 3 los pensamos ahora en estos términos. Tenemos 3 variables:

1. *persona* con tres posibles valores (John, Jane, Mary)
2. *tratamiento* con dos posibles valores (a o b)
3. *resultado* con 5 o 6 valores (-, 16, 3, 2, 11, 1)

El diseño del experimento mismo nos habla de la estructura de las observaciones y los posibles valores que pueden tomar. Por ejemplo, en este caso el valor perdido nos dice que, por diseño, se debió de capturar esta variable pero no se hizo (por eso es importante guardarlo como tal). Los valores perdidos estructurales, representan mediciones de valores que no se puede hacer o que no suceden y, por tanto, se pueden eliminar (por ejemplo, hombres embarazados). En la figura 4 se muestran los mismos datos que antes pero pensados tal que las variables son columnas y las observaciones (en este caso, cada punto en el diseño experimental) son filas.

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Figura 4: Observaciones son filas, variables columnas.

Normalmente, es fácil determinar qué son observaciones y qué son variables pero es muy dificil definir en forma precisa variables y observaciones. Por ejemplo, si tienes teléfonos de casa y celulares, se pueden considerar como dos variables distintas en muchos contextos pero en prevención de fraude necesitas una variable que guarde el tipo de teléfono y otra en la que se guarde el número pues el uso regular del mismo número de teléfono por parte de la misma persona puede ayudar a detectarlo.

En general, es más fácil describir las relaciones funcionales entre las variables que entre las filas (el radio, una combinación lineal). También es más fácil hacer comparaciones entre grupos que entre columnas (la suma, el promedio, la varianza, la moda).

Datos limpios

Éstos mapean de forma estándar el significado y la estructura de los datos. Un conjunto de datos se considera sucio o limpio dependendiendo en cómo las filas, columans y tablas mapean a observaciones, variables y tipos. En **datos limpios**:

1. Cada *variable* es una columna.
2. Cada *observación* es una fila.
3. Cada *tipo de unidad observacional* es una tabla.

Esto equivale a la tercera forma normal de Codd enfocado a un solo conjunto de datos y no a datos conectados como en bases relacionales. Los datos sucios son cualquier otro tipo de manera de organizar los datos.

La tabla 4 corresponde a datos limpios: cada fila es una observación, es decir, el resultado de un tratamiento a una persona. Cada columna es una variable. Solo tenemos un tipo de unidad observacional, es decir, cada renglón es una unidad del diseño experimental.

Con los datos así ordenados, suele ser más fácil extraer datos que, por ejemplo, la 2.

Ejercicios

1. Crea un dataframe con los valores de la tabla 2 y otro con los valores de la tabla 4.
2. Extrae el resultado para John Smith, tratamiento a en la primera configuración y en la segunda.
3. Especifica el número de tratamientos con la forma sucia y la forma limpia.
4. Cuál es la media de los resultados: usa la forma 1 y la forma 2.
5. Extrae los tratamientos del tipo a en la forma 2.

Como puedes ver, los datos limpios nos permiten preguntarle cosas a los datos de manera simple y sistemática. En particular, es una estructura muy útil para programación vectorizada como en R (el ejercicio 5) porque la forma se asegura que valores para diferentes variables de la misma observación siempre están apareados.

Por convención, las variables se acomodan de una forma particular. Las variables *fijas*, en este ejemplo, las propias al diseño experimental, van primero y posteriormente las variables *medidas*. Ordenamos éstas de forma que las que están relacionadas sean contiguas.

De sucio a limpio

Los conjuntos de datos normalmente violan estos criterios. Es raro obtener un conjunto de datos con el cuál podemos trabajar de manera inmediata.

Los 5 problemas más comunes para llevar datos sucios a limpios son

1. Los nombres de las columnas son valores, no nombres de variables.
2. Múltiples variables se encuentran en la misma columna.
3. Las variables están guardadas tanto en filas como en columnas.
4. Muchos tipos de unidad observacional se encuentran en la misma tabla.
5. Una sola unidad observarcional se guardó en varias tablas.

Estos problemas pueden ser resueltos con solamente: *melting*, separación de cadenas y *casting*.

Los nombres de las columnas son valores, no nombres de variables

La tabla siguiente muestra datos sucios con este problema. Se muestran distintas religiones con el numero de personas que pertenecen a distintos niveles de ingreso. Dentro de un reporte, este tipo de representación tiene mucho sentido y permite visualizar muchas cosas rápidamente.

El conjunto de datos tiene 3 variables: *religion*, *ingreso* y *frecuencia*. Para arreglarlo, necesitamos *juntar* (melt) las columnas con nombres de niveles de ingreso en una sola columna que contenga esos nombres como valores. En otras palabras, debemos convertir de la columna 2 en adelante en filas.

Con el paquete **tidyr** esto se puede realizar en forma fácil con el comando `gather`.

Religion	X..10k	X.10.20k	X.20.30k	X.30.40k	X.40.50k	X.50.75k	X.75.100k	X.100.150k	X.150k	Don T Know Refused
Agnostic	27	34	60	81	76	137	122	109	84	96
Atheist	12	27	37	52	35	70	73	59	74	76
Buddhist	27	21	30	34	33	58	62	39	53	54
Catholic	418	617	732	670	638	1,116	949	792	633	1,489
Don't know/refused	15	14	15	11	10	35	21	17	18	116
Evangelical Prot	575	869	1,064	982	881	1,486	949	723	414	1,529
Hindu	1	9	7	9	11	34	47	48	54	37
Historically Black Prot	228	244	236	238	197	223	131	81	78	339
Jehovah's Witness	20	27	24	24	21	30	15	11	6	37
Jewish	19	19	25	25	30	95	69	87	151	162
Mainline Prot	289	495	619	655	651	1,107	939	753	634	1,328
Mormon	29	40	48	51	56	112	85	49	42	69
Muslim	6	7	9	10	9	23	16	8	6	22
Orthodox	13	17	23	32	32	47	38	42	46	73
Other Christian	9	7	11	13	13	14	18	14	12	18
Other Faiths	20	33	40	46	49	63	46	40	41	71
Other World Religions	5	2	3	4	2	7	3	4	4	8
Unaffiliated	217	299	374	365	341	528	407	321	258	597

```
limpios <- tidyr::gather(raw, key = income, value = freq, -religion)
```

Religion	Income	Freq
Agnostic	<$10k	27
Atheist	<$10k	12
Buddhist	<$10k	27
Catholic	<$10k	418
Don't know/refused	<$10k	15
Evangelical Prot	<$10k	575

Este tipo de forma de guardar datos es útil también cuando se capturan datos al evitar la repetición de valores.

Múltiples variables se encuentran en la misma columna

Otra forma de datos sucios se encuentra cuando una columna con nombres de variables tiene realmente varias variables dentro del nombre (como en el ejemplo siguiente).

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0		
AE	2000	2	4	4	6	5	12	10		3
AF	2000	52	228	183	149	129	94	80		93
AG	2000	0	0	0	0	0	0	1		1
AL	2000	2	19	21	14	24	19	16		3
AM	2000	2	152	130	131	63	26	21		1
AN	2000	0	0	1	2	0	0	0		0
AO	2000	186	999	1003	912	482	312	194		247
AR	2000	97	278	594	402	419	368	330		121
AS	2000					1	1			

El primer paso es pasar las columnas que son valores de variable a una sola variable

Posteriormente, debemos separar en las columnas apropiadas las variables que estan contenidas en los antiguos nombres de variables.

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	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

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	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3