
Big Data I:

Ingeniería de datos

Felipe Ortega
Dpto. de Estadística e Investigación Operativa
Universidad Rey Juan Carlos

March 24, 2015





(cc)2015 Felipe Ortega.

Algunos derechos reservados.

Este documento se distribuye bajo una licencia Creative Commons
Reconocimiento-CompartirIgual 4.0, disponible en:

<http://creativecommons.org/licenses/by-sa/4.0/es/>

Preparación y transformación de datos

- Limpieza de datos.
- Datos no disponibles.
 - Gestión de valores vacíos.
 - Imputación de datos no disponibles.
- Transformación de datos.
 - *Data munging* o *data wrangling*.
 - Pasar los datos a otro formato o dejarlos preparados para luego poder analizarlos más fácilmente.



Data Wrangler

- Ejemplo de este tipo de herramientas (UW Interactive Data Lab).

The screenshot shows the Data Wrangler interface. On the left, there is a sidebar with a list of transformation scripts under the 'Transform Script' tab. The main area displays a data table with columns for 'Year', 'State', and 'Property_crime_rate'. The table contains data for various states including Alabama, Alaska, Arizona, Arkansas, California, and Colorado, with rows for the years 2004, 2005, 2006, 2007, and 2008.

Transform Script List:

- Split data repeatedly on newline into rows
- Split split repeatedly on ','
- Promote row 0 to header
- Delete empty rows
- Extract from Year between positions 18, 24
- Cut from Year between positions 18, 24
- Cut from Year on 'Alaska'
- Split Year between positions 18, 24
- Split Year on 'Alaska'

Data Table:

	Year	State	Property_crime_rate
0	Reported crime in Alabama	Alabam	
1	2004		4029.3
2	2005		3900
3	2006		3937
4	2007		3974.9
5	2008		4081.9
6	Reported crime in Alaska	Alaska	
7	2004		3370.9
8	2005		3615
9	2006		3582
10	2007		3373.9
11	2008		2928.3
12	Reported crime in Arizona	Arizon	
13	2004		5073.3
14	2005		4827
15	2006		4741.6
16	2007		4502.6
17	2008		4087.3
18	Reported crime in Arkansas	Arkans	
19	2004		4033.1
20	2005		4068
21	2006		4021.6
22	2007		3945.5
23	2008		3843.7
24	Reported crime in California	Califo	
25	2004		3423.9
26	2005		3321
27	2006		3175.2
28	2007		3032.6
29	2008		2940.3
30	Reported crime in Colorado	Colora	

Preparación de datos

- En primer lugar, debemos comprobar que no existen valores extraños ni datos omitidos.
 - Utilizar técnicas básicas de resumen de datos.
 - Técnicas de visualización de datos omitidos.
- Después, tenemos dos opciones:
 - Descartar los casos que contengan variables con datos omitidos.
 - Imputar valores para los datos que faltan, utilizando técnicas avanzadas de imputación de múltiples valores.



Transformación de datos

- Otro paso crucial antes de comenzar nuestro análisis es comprobar la distribución de valores de los parámetros implicados.
 - Muchas técnicas y modelos asumen que los datos siguen una cierta distribución (e.g. Normal), pero puede no ser cierto.
 - De hecho, en la práctica nos encontramos muchas veces con distribuciones sesgadas (*skewed distributions*) o con diferentes apuntamientos (*kurtosis*).
- Posibles objetivos:
 - Reducir la asimetría de la distribución de valores.
 - Transformar una o varias variables de forma que se parezcan más a una distribución Normal (univariante o multivariante).

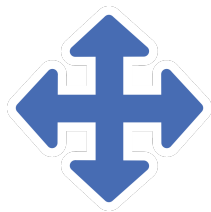


Aspectos adicionales

- Transformación entre diferentes formatos de datos
 - *Wide format vx. long format.*

Wide format

subject	sex	control	cond1	cond2
1	M	7.9	12.3	10.7
2	F	6.3	10.6	11.1
3	F	9.5	13.1	13.8
4	M	11.5	13.4	12.9



Aspectos adicionales

Long format

subject	sex	condition	measurement
1	M	control	7.9
1	M	cond1	12.3
1	M	cond2	10.7
2	F	control	6.3
2	F	cond1	10.6
2	F	cond2	11.1
3	F	control	9.5
3	F	cond1	13.1
3	F	cond2	13.8
4	M	control	11.5
4	M	cond1	13.4
4	M	cond2	12.9

