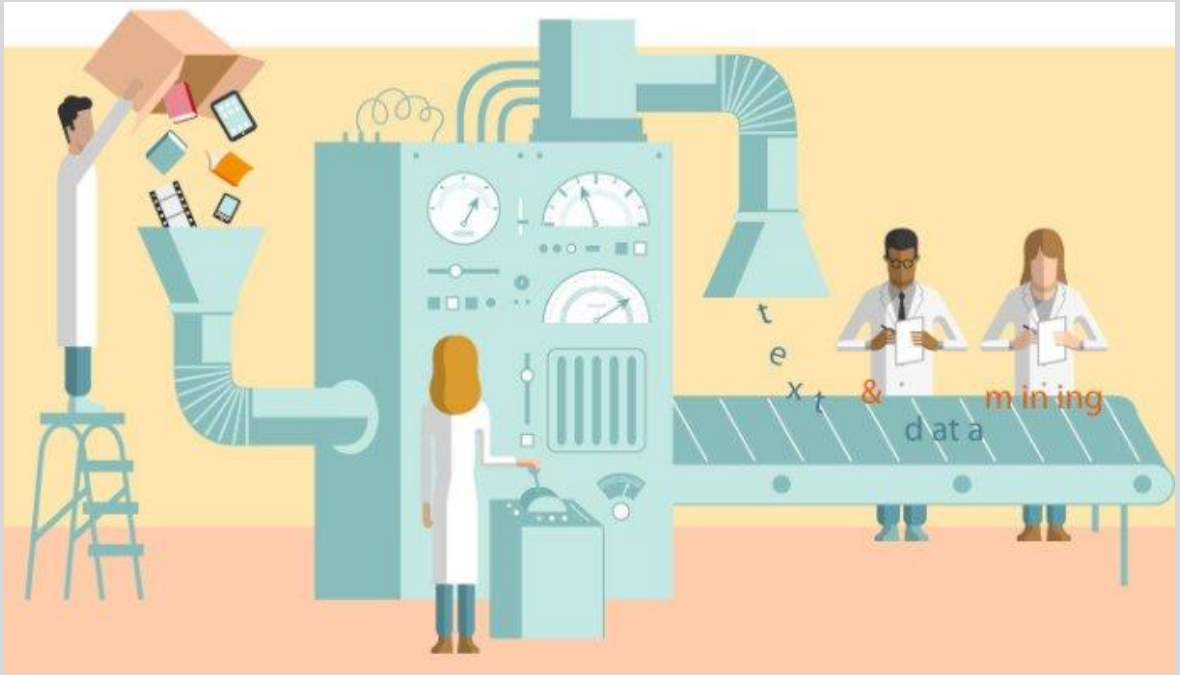


DigitalHouse > Coding School

DATA SCIENCE

UNIDAD 1
MÓDULO 2

Variable Categóricas, Dummies y TimeStamps

Agosto 2017

1. Presentación de conceptos
2. Práctica Guiada (Data Wrangling)
3. Práctica Independiente Dummies

- Una **variable categórica** es una variable que puede tomar un número contable de valores que indican la pertenencia a un grupo (o categoría) con determinadas propiedades cualitativas.
  - Ejemplos: género, estado civil, grupo sanguíneo, color de pelo, etc.
  - Una variable categórica con dos posibles valores se denomina **binaria** o **dicotómica**.

- En estadística se suele asignar etiquetas (o rótulos) numéricos a las variables categóricas
  - Ejemplo: en el caso del estado civil, 0 si soltero y 1 si casado y 2 si divorciado.
  - Los números utilizados para rotular son arbitrarios y en consecuencia las variables categóricas **carecen de ordinalidad**.
  - Luego la principal medida de posición es la moda ya que la mediana y la media no están definidas (y en general cualquier operación numérica tampoco)
  - En general, el software asume que los valores numéricos reflejan cantidades algebraicas.

- Una **variable dummy** es una variable cualitativa que toma valores 0 o 1 para indicar la ausencia o presencia de algún atributo o efecto categórico.
  - Formalmente una variable dummy, Di, puede ser expresada mediante una **función indicadora**:

$$D_i = \mathbb{I}_A(x_i) = \begin{cases} 1 & \text{si } x_i \in A \\ 0 & \text{si } x_i \notin A \end{cases}$$

- ¿Cuál es la relación entre variables categóricas y variables dummies?
  - Una variable categórica con N categorías puede ser expresada en términos de N−1 variables dummies (**one-hot encoding**).
  - Resuelve el problema de interpretar las etiquetas numéricas como un intervalo.
  - Sin embargo si las categorías tienen muchos valores entonces aumenta considerablemente la dimensionalidad de los datos.

- Supongamos que tenemos una variable categórica, C, que registra la ciudad en la que reside una muestra de habitantes de la Argentina.

  - Asumamos que la variable puede tomar 4 posibles valores: Buenos Aires, Rosario, Córdoba y Mar del Plata.
  - Imaginemos que tenemos las siguiente 5 observaciones:

Obs.	Ciudad
1	Rosario
2	Buenos Aires
3	Rosario
4	Mar del Plata
5	Córdoba

- Alternativamente podemos expresar estas observaciones de la variable categórica usando dummies como:

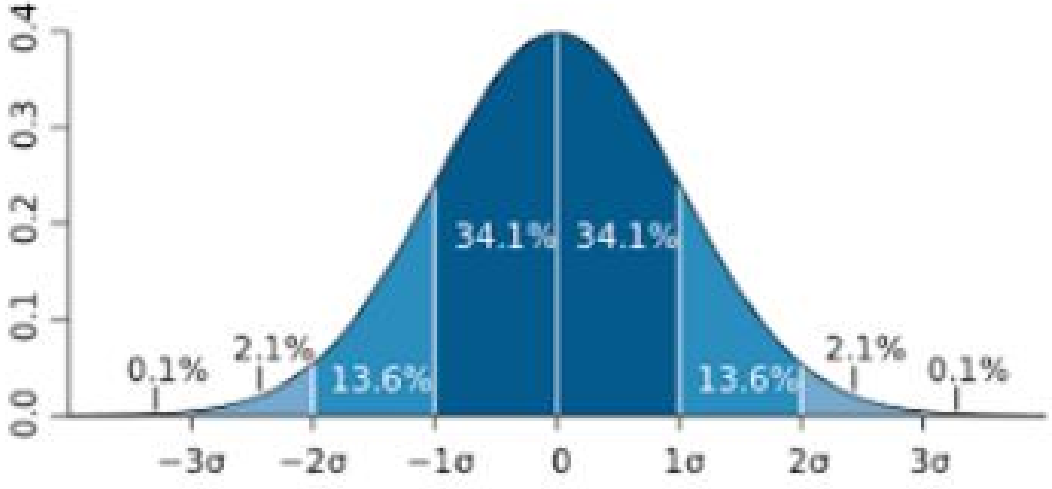
Obs.	Ciudad
1	Rosario
2	Buenos Aires
3	Rosario
4	Mar del Plata
5	Córdoba

Obs.	D_BA	D_C	D_R
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	0
5	0	1	0

- Es importante notar que si existen k categorías, k-1 variables Dummies son suficientes para representarlas.

- Recordemos que un caso muy frecuente en datasets con variables continuas, es que los datos se distribuyan normalmente. Una distribución normal puede caracterizarse por su media y su desvío estándar.
- Esta estimación nos permite esperar determinada cantidad de casos en determinados rangos, basándonos justamente en la media y el desvío.

¿CÓMO SE DISTRIBUYE UNA VARIABLE DUMMY?



Al tomar sólo dos valores, esta variable no se distribuye como una normal, sino como una binomial.

Recordemos que la binomial recibe un único parámetro p…
¿Y cómo calculamos p?

p es la probabilidad de ocurrencia de la categoría. Dentro de la muestra, lo podemos calcular como **la media** de la variable dummy.

¿Y qué significa el desvío estándar?
En una variable binomial, el desvío estándar se puede calcular en función de p.
¡**La media y el desvío ya no son dos parámetros independientes**!

STD(D) = $\sqrt{p(1-p)}$