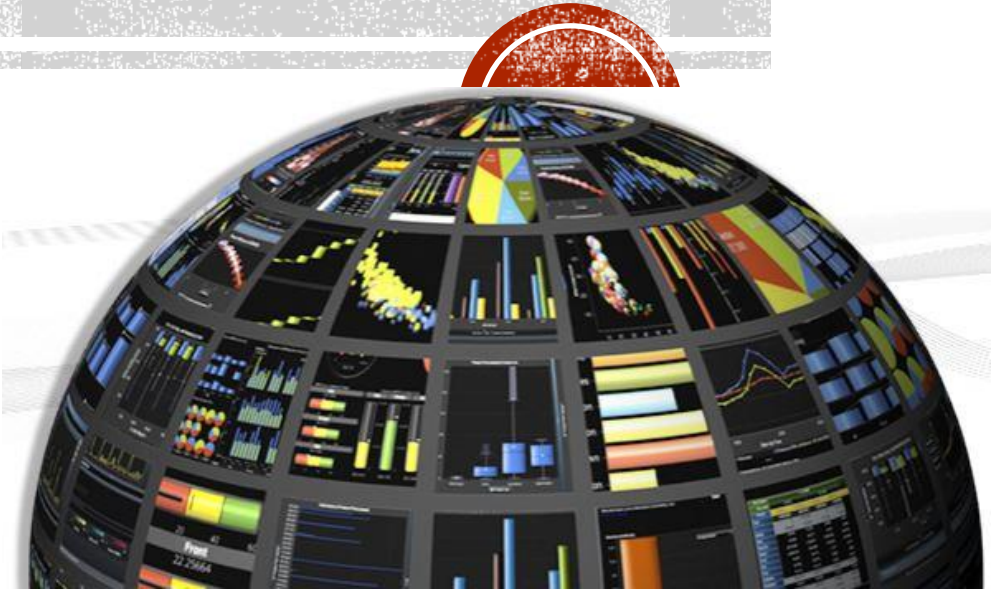
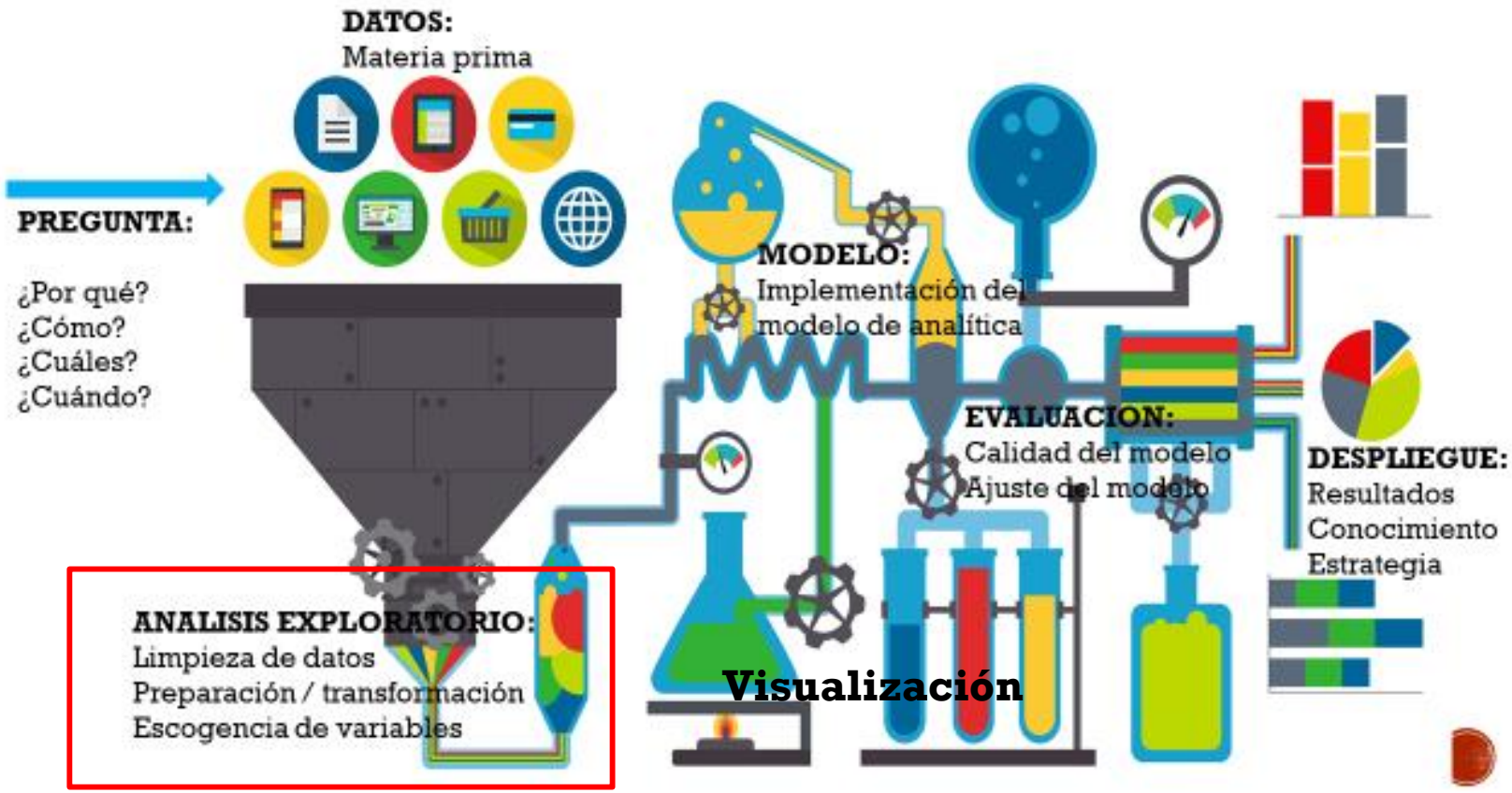
ANÁLISIS EXPLORATORIO

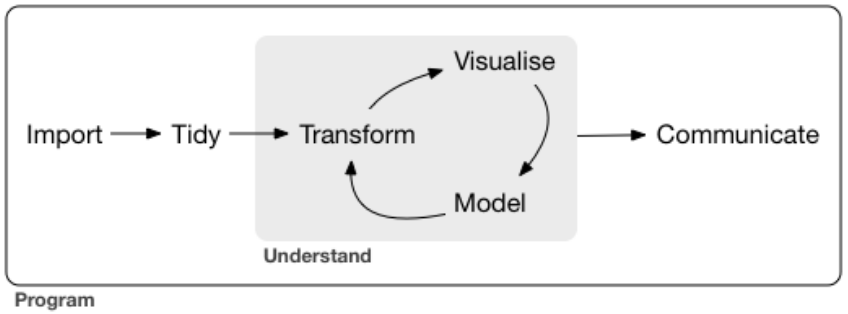
Christian Camilo Urcuqui López, MSc



UNIVERSIDAD ICESI | A OTRO NIVEL

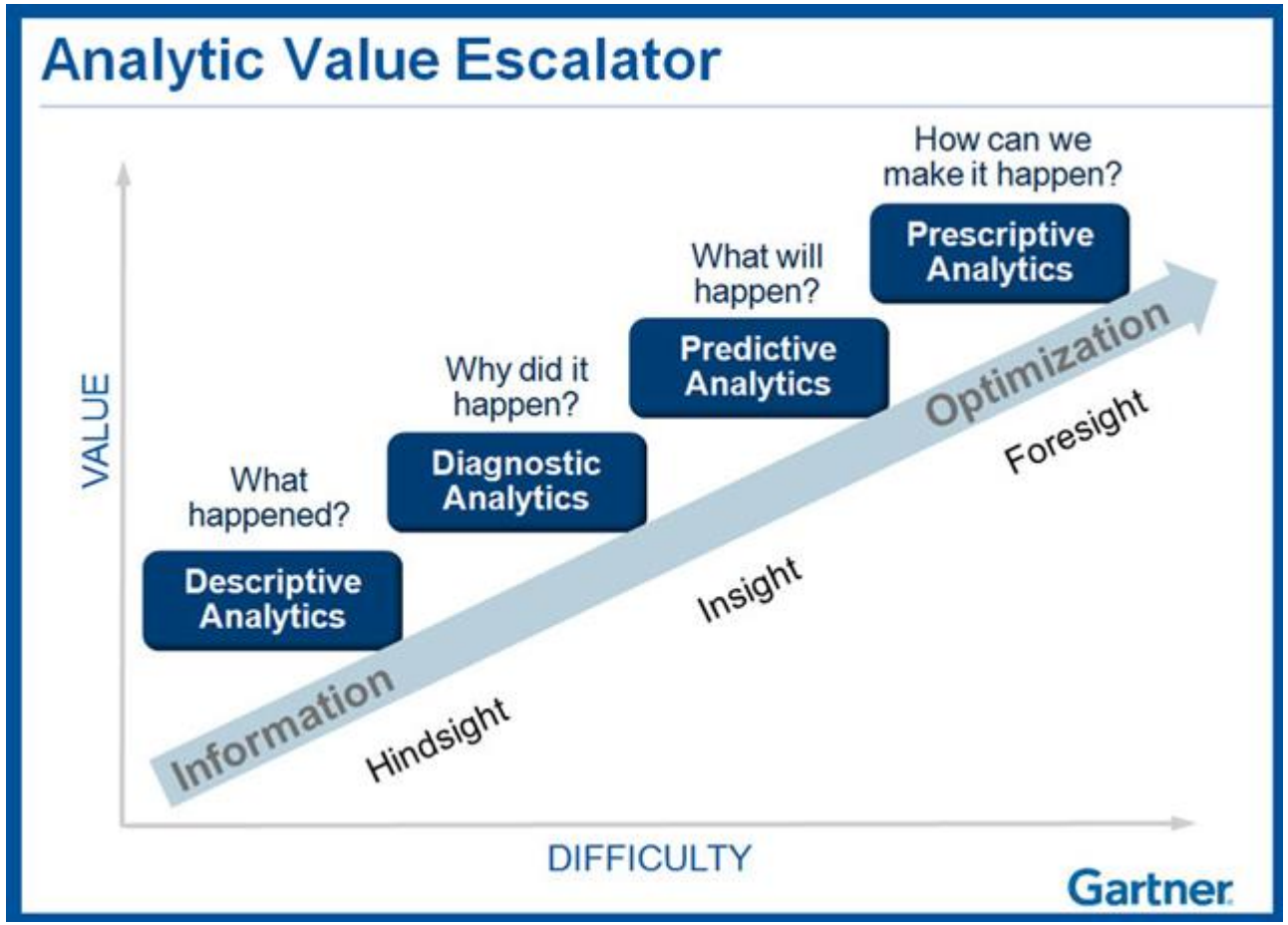
RECORDEMOS





Marco de trabajo típico de un proyecto de ciencia de datos.
R for Data Science

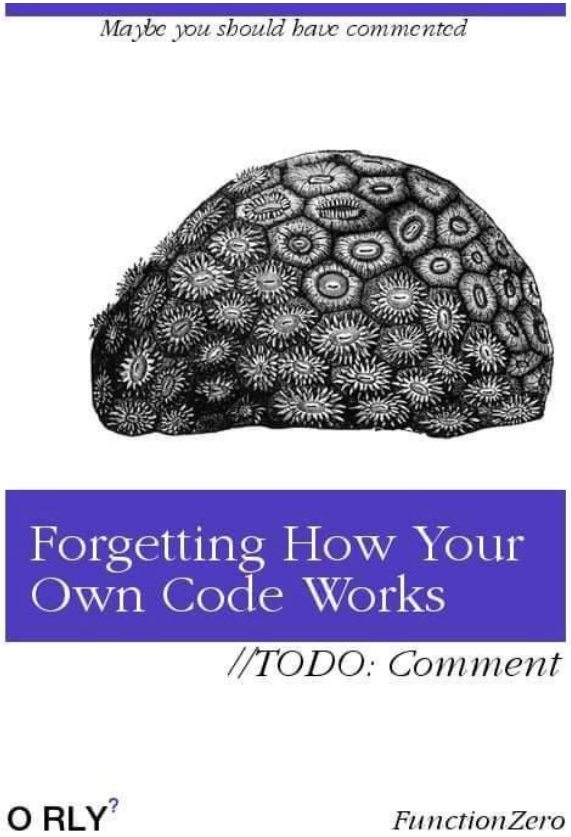
EXPLORATORY DATA ANALYSIS (EDA)



EXPLORATORY DATA ANALYSIS (EDA)

- **Objetivos**
  - Detectar problemas o errores en la información.
  - Entender la estructura de los datos .
  - Encontrar relaciones entre las variables.
  - Identificar los posibles modelos que mejor se adaptarían a los datos.
- Es el primer análisis que se realiza en un proyecto de ciencia de datos.
- Incluye un análisis descriptivo con estadísticos básicos.
- Para este análisis debería existir un elemento crucial en los Proyectos – **El diccionario de datos-**

EXPLORATORY DATA ANALYSIS (EDA)



EXPLORATORY DATA ANALYSIS (EDA)

Diccionario de datos, o un repositorio de metadatos, es un documento que contiene información acerca de la información del proyecto, por ejemplo:

- Nombres
- Tipos de variables
- Descripción
- Relaciones entre variables
- Origen
- Destino
- Formato
- Rango

A partir del diccionario de datos podemos encontrar los distintos problemas en los *datasets*.

EXPLORATORY DATA ANALYSIS (EDA)

TABLE NAME	ATTRIBUTE NAME	CONTENTS	TYPE	FORMAT	RANGE	REQUIRED	PK or FK	FK REFERENCED TABLE
PASSENGER	PASS_NO	Passenger number	CHAR	99999	00001-99999	Y	PK	
	PASS_NAME	Passenger's name	VARCHAR	Xxxxx Xxxxx		Y		
	PASS_TEL	Passenger's telephone number	VARCHAR	999-999-9999		Y		
	PASS_USER	Passenger's username	VARCHAR	Xxxxxxx		Y		
	PASS_PASSWORD	Passenger's password	VARCHAR	Xxxxxxx		Y		
	PASS_ACC	Passenger's account number	CHAR	9999999999	0000000001-9999999999	Y		
DRIVER	DRIVER_NO	Driver number	CHAR	99999	00001-99999	Y	PK	
	DRIVER_NAME	Driver's name	VARCHAR	Xxxxx Xxxxx		Y		
	DRIVER_TEL	Driver's telephone number	VARCHAR	999-999-9999		Y		
	CAR_NO	Car number	CHAR	99999	00001-99999	Y	FK	CAR_NO
CAR	CAR_NO	Car number	CHAR	99999	00001-99999	Y	PK	
	CAR_LICENSE	Number of Car license plate	CHAR	Xxx-xxxx		Y		
	CAR_COLOR	Car color	CHAR	Xxxxx		N		
	CAR_CONDITION	Car condition	TEXT	Xxxxx xxxxx		Y		
	CAR_BRAND	Car brand	CHAR	Xxxxx		N		
INVOICE	INVOICE_NO	Invoice number	CHAR	99999	00001-99999	Y	PK	
	INVOICE_DATE	Invoice date	DATE	dd-mm-yyyy		Y		
	INVOICE_AMOUNT	Invoice total amount	NUMBER	9999.99	35.00-9999.99	Y		
	TRANS_NO	Transportation number	CHAR	99999999	00000001-99999999	Y	FK	TRANS_NO
TRANSPORT	TRANS_NO	Transportation number	CHAR	99999999	00000001-99999999	Y	PK	
	TRANS_SRC	Source of location	TEXT	Xxxxx Xxxxx		Y		
	TRANS_DST	Destination of location	TEXT	Xxxxx Xxxxx		Y		
	TRANS_DISTANCE	Total distance	NUMBER	99.99		Y		
	PASS_NO	Passenger number	CHAR	99999	00001-99999	Y	FK	PASS_NO
	DRIVER_NO	Driver number	CHAR	99999	00001-99999	Y	FK	DRIVER_NO
LOCATION	LOC_NO	Location number	CHAR	99999999	00000001-99999999	Y	PK	
	LOC_NAME	Location name	VARCHAR	Xxxxx Xxxxx		Y		
	PASS_NO	Passenger number	CHAR	99999	00001-99999	Y	FK	PASS_NO
	DRIVER_NO	Driver number	CHAR	99999	00001-99999	Y	FK	DRIVER_NO
REPORT	REPORT_NO	Report number	CHAR	99999	00001-99999	Y	PK	
	REPORT_DETAIL	Report detail	TEXT	Xxxxx Xxxxx		N		
	REPORT_DATE	Report date	DATE	dd-mm-yyyy		Y		
	REPORT_RATE	Report rate	NUMBER	99	01-10	Y		
	PASS_NO	Passenger number	CHAR	99999	00001-99999	Y	FK	PASS_NO

https://sites.google.com/site/happytaxicompany/data-dictionary

EXPLORATORY DATA ANALYSIS (EDA)

- Los principales tipos de problemas que se pueden encontrar en los datos son:
  - El formato de las variables no coincide con el tipo de variable.
  - Observaciones duplicadas.
  - Valores perdidos (NaN).
  - Errores de digitación.
  - Valores fuera de rango o inválidos.

EXPLORATORY DATA ANALYSIS (EDA)

- Defina y determine los tipos de problemas
- Buscar e identificar los problemas
- Corregir los problemas descubiertos
- Documentar el proceso de limpieza

Usualmente, es la actividad que más consume tiempo en un proyecto de ciencia de datos

EXPLORATORY DATA ANALYSIS (EDA)

Datos brutos (raw data)

- Son los datos adquiridos de la fuente original.
- Pueden presentar problemas.
- Usualmente se procesan solo una vez, para prepararlos.

Datos preparados (tidy data)

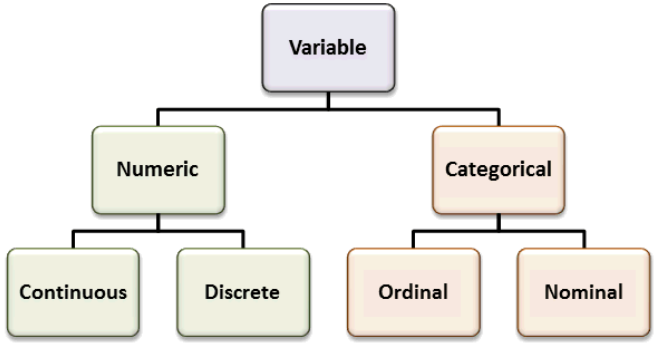
- Los ya se encuentran listos para ser analizados.
- Su preparación incluye: limpieza, transformación, fusiones (merging), extracción de subconjuntos (subsetting).

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23.

UNIVERSIDAD ICESI | A OTRO NIVEL

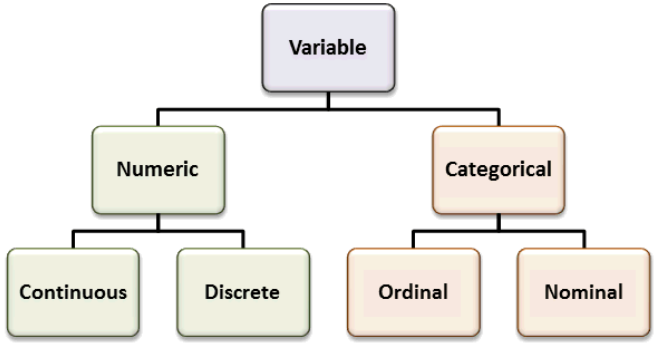
TIPOS DE VARIABLES

- **Variables cuantitativas**, sus valores son numéricos y pueden ser contados o medidos, por ejemplo, ventas netas de una compañía.
  - **Variables discretas**, es una variable numérica que usualmente se obtiene a través del conteo y solamente puede tomar valores específicos de un conjunto, por ejemplo, el número de personas en una ciudad o el número de quejas de los clientes.
  - **Variables continuas**, son variables numéricas que pueden tomar un valor (infinito/decimal) entre dos valores numéricos cualquiera. Usualmente, esta variable se obtiene a partir de mediciones, por ejemplo, la temperatura de un paciente.



TIPOS DE VARIABLES

- **Variables cualitativas**, conocidos también como variables categóricas, sus valores pueden ser contados pero no medidos.
  - **Variables nominales**, son valores que presentan a una categoría y no cuentan con un orden. Estos valores pueden ser contados pero no pueden ser ni medidos y ni ordenados, por ejemplo, género de música y categorías de productos.
  - **Variables ordinales**, son valores numéricos que pueden ser discretos o continuos y que están ya sea ordenadas o jerarquizadas.
  - **Variables binarias**, sus valores hacen parte únicamente a dos categorías que generalmente son opuestos, por ejemplo, 1/0 y verdadero/falso.



TIPOS DE VARIABLES

- **Variables independientes**, sus valores no dependen de otra variable pero posiblemente si puedan influenciar a otras.
- **Variables dependientes**, este tipo de variable si depende de otras variables.
- **Variable aleatoria**, es una variable que puede asumir un valor de un rango de valores basado en la probabilidad.

TIPOS DE VARIABLES

Variables explicativas

Objetivo

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

TIPOS DE VARIABLES

Es importante clasificar las variables ya que de este dependerá el tipo de análisis.

Tipo de dato

- Carácter
  - Gender
- Numérico
  - Prev_Exam_Marks
  - Height
  - Weight
  - Play Cricket

Tipo de variable

- Cualitativa
  - Gender
  - Play Cricket
- Cuantitativa
  - Prev_Exam_Marks
  - Height
  - Weight

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

BIBLIOGRAFÍA

- Lutz, M. (2013). *Learning Python: Powerful Object-Oriented Programming*. " O'Reilly Media, Inc.".
- Villegas, N. & Estrada, D. Introducción al análisis exploratorio. Diplomado en analítica y grandes volúmenes de datos.