Data Fundamentals

Con Python

24 de Febrero, 2023

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Centro de Referencia Nacional en Comercio Electrónico y Marketing CRN Digital
UNIÓN EUROPEA
Barrabés
The Valley
"El FSE invierte en tu futuro"
Fondo Social Europeo

Índice

Data Manipulation

Data Wrangling

Data Wrangling

Hemos tratado datos, visto ejemplos y casuísticas…

Data Wrangling

Hemos tratado datos, visto ejemplos y casuísticas…

Hay más contenido teórico-práctico para manejar mejor los datos

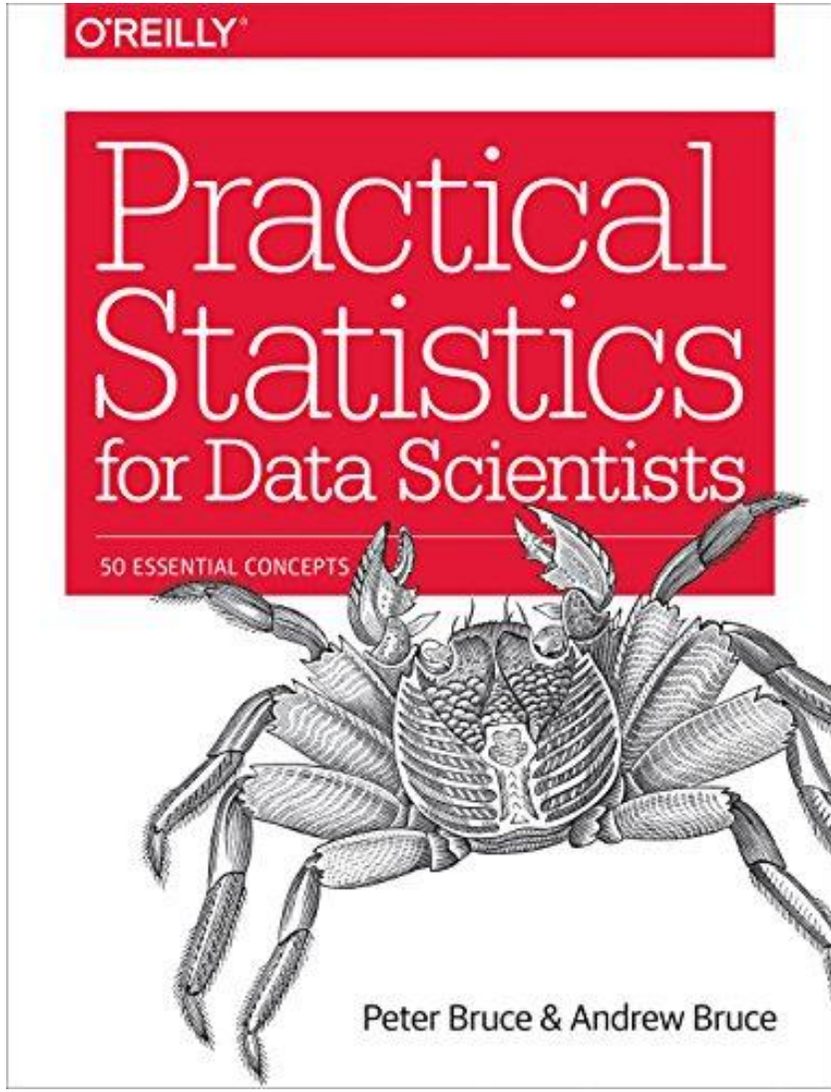
- Data Cleansing
- Data Cleaning
- Data Preparation
- …

Data Wrangling

Hemos tratado datos, visto ejemplos y casuísticas…

Hay más contenido teórico-práctico para manejar mejor los datos

- Data Cleansing
- Data Cleaning
- Data Preparation
- …

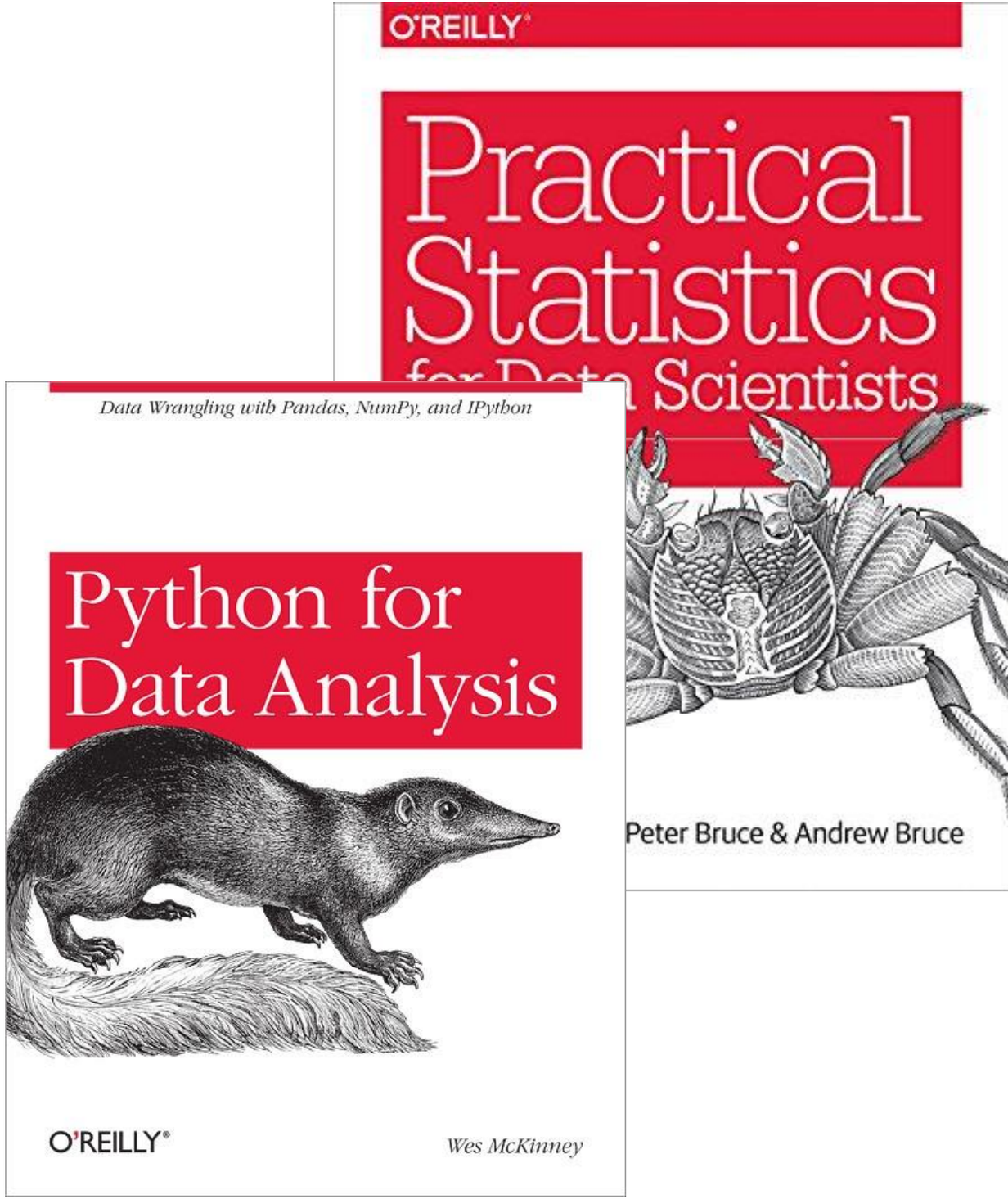


Data Wrangling

Hemos tratado datos, visto ejemplos y casuísticas…

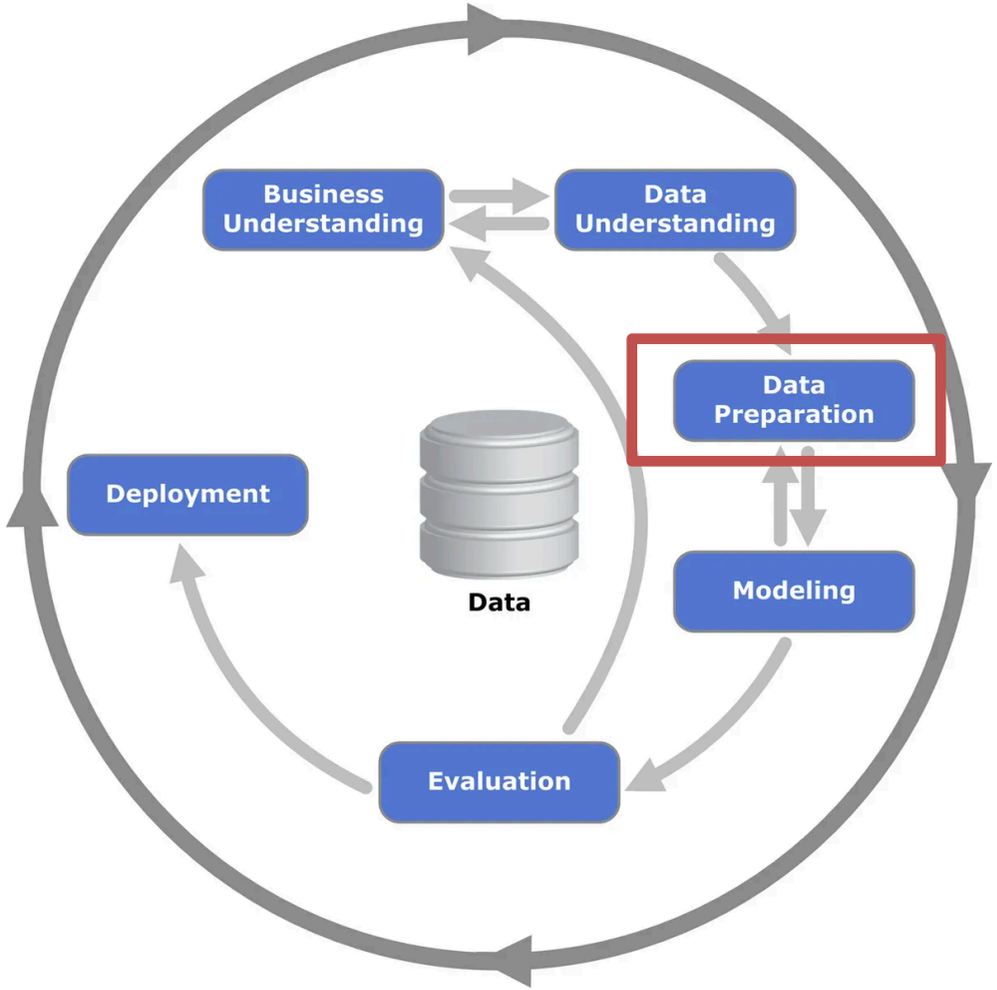
Hay más contenido teórico-práctico para manejar mejor los datos

- Data Cleansing
- Data Cleaning
- Data Preparation
- …



Siguiente Paso

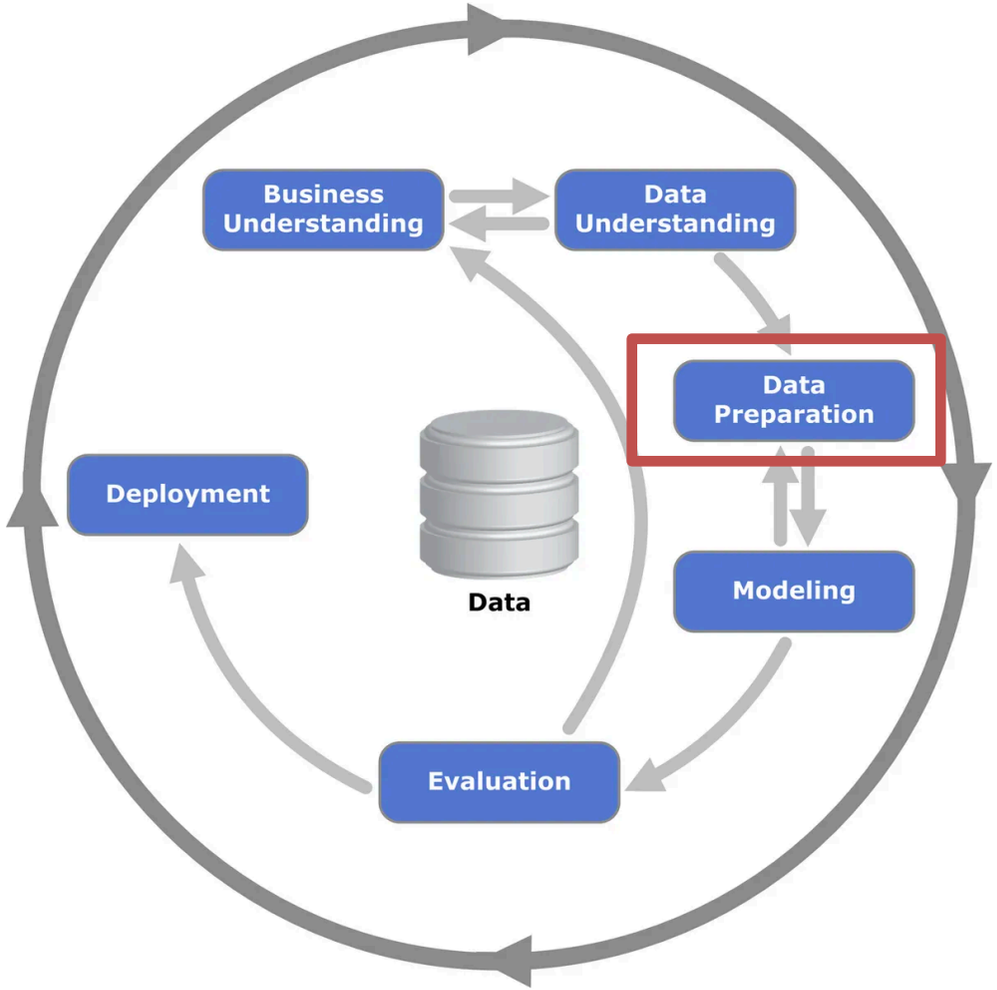
Una vez tenemos bien explorados los datos:



Siguiente Paso

Una vez tenemos bien explorados los datos:

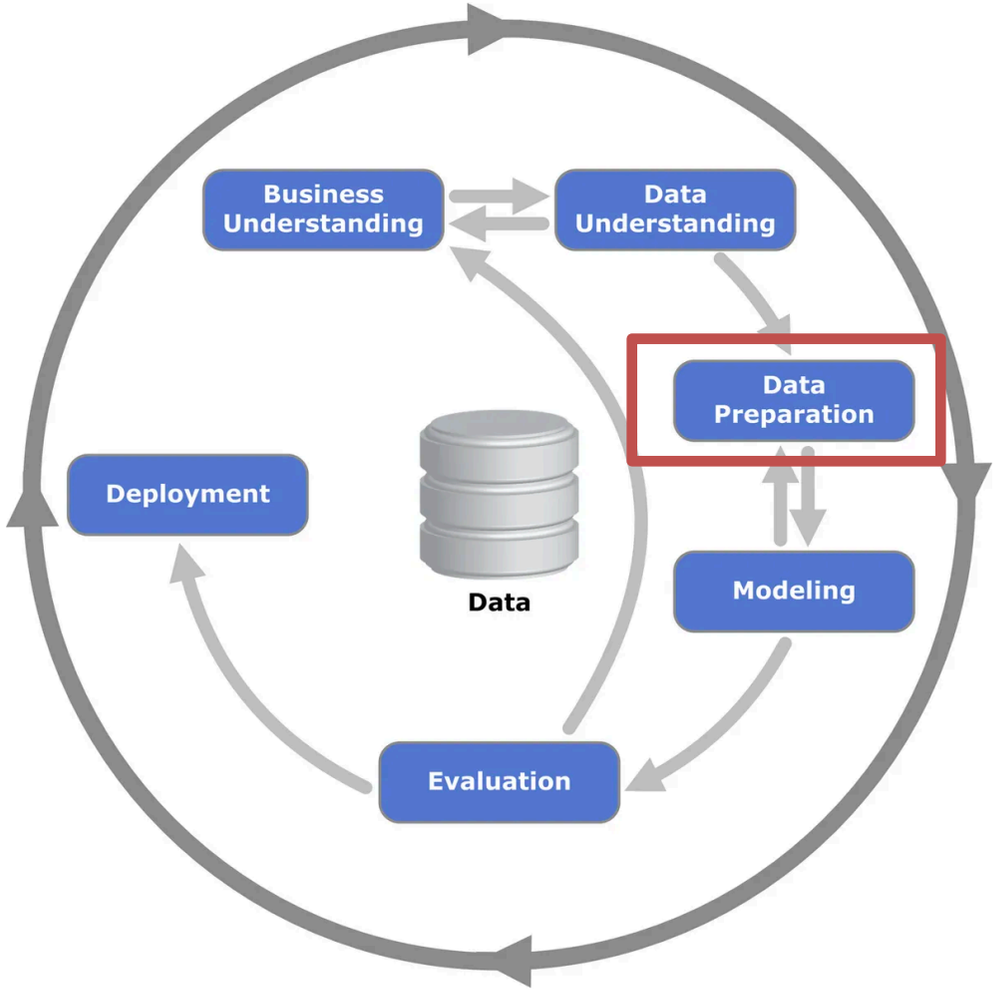
- Incompletitudes



Siguiente Paso

Una vez tenemos bien explorados los datos:

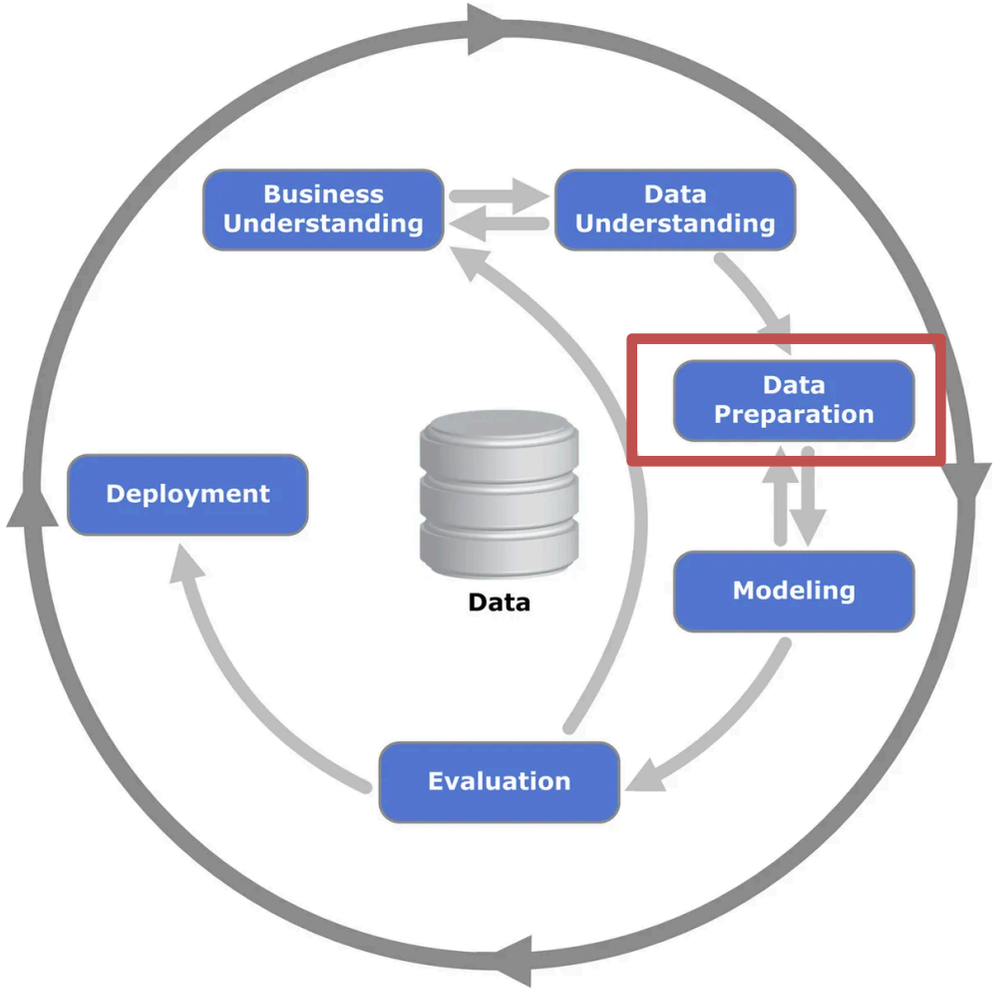
- Incompletitudes
- Nulos



Siguiente Paso

Una vez tenemos bien explorados los datos:

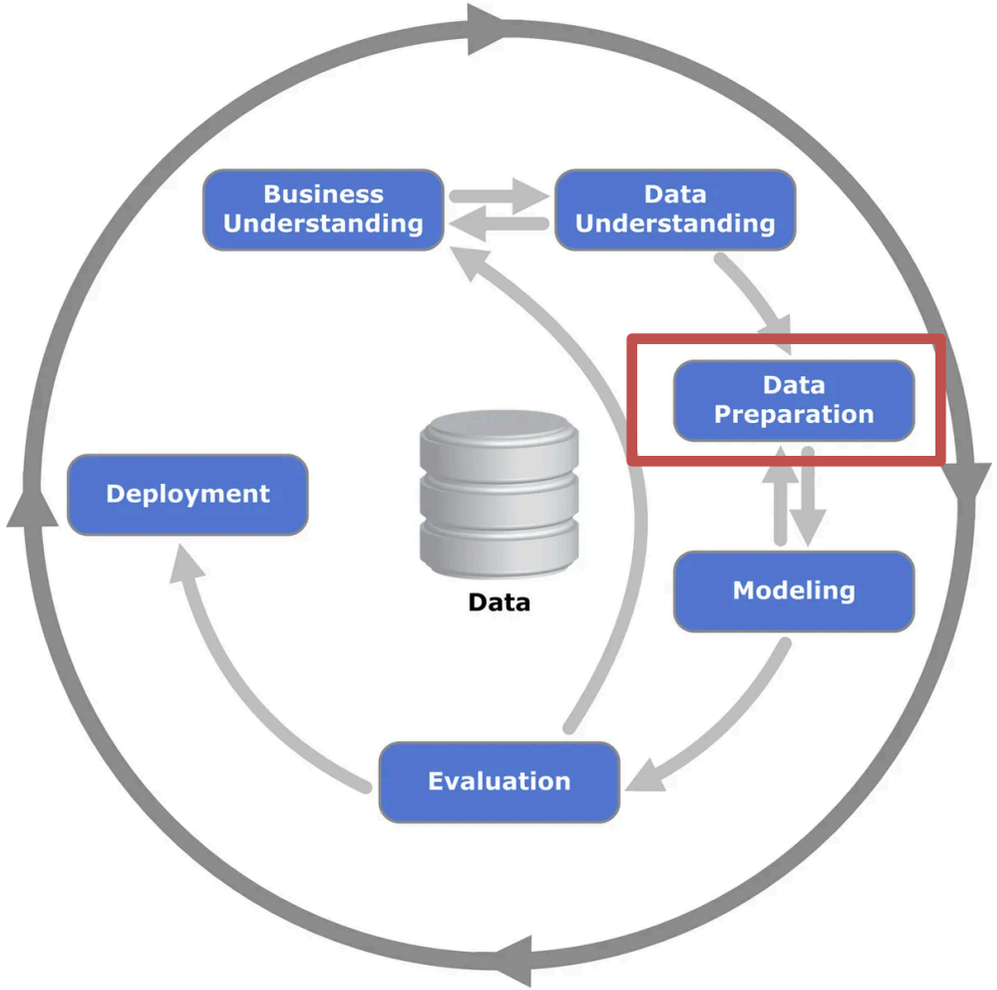
- Incompletitudes
- Nulos
- Incoherentes



Siguiente Paso

Una vez tenemos bien explorados los datos:

- Incompletitudes
- Nulos
- Incoherentes
- Inexistentes



Incompletos

Unidades	Edad	Sexo	PAS	PAD	Peso	Talla
1	×	×	?	?	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	×	×	?	?	×	×
101	×	×	×	×	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	×	×	×	×	?	?
201	×	×	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
300	×	×	?	?	?	?
301	×	×	×	×	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.000	×	×	×	×	×	×

Nota: El símbolo ? representa los valores ausentes y ×, los observados

Incompletos

Tenemos que identificarlos y "marcarlos"

Unidades	Edad	Sexo	PAS	PAD	Peso	Talla
1	×	×	?	?	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	×	×	?	?	×	×
101	×	×	×	×	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	×	×	×	×	?	?
201	×	×	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
300	×	×	?	?	?	?
301	×	×	×	×	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.000	×	×	×	×	×	×

Nota: El símbolo ? representa los valores ausentes y ×, los observados

Incompletos

Tenemos que identificarlos y "marcarlos"

-1, Null, 0, Símbolos… Tenemos que saber dónde están y colocar una señal.

Unidades	Edad	Sexo	PAS	PAD	Peso	Talla
1	×	×	?	?	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	×	×	?	?	×	×
101	×	×	×	×	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	×	×	×	×	?	?
201	×	×	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
300	×	×	?	?	?	?
301	×	×	×	×	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.000	×	×	×	×	×	×

Nota: El símbolo ? representa los valores ausentes y ×, los observados

Incompletos

Tenemos que identificarlos y "marcarlos"

-1, Null, 0, Símbolos… Tenemos que saber dónde están y colocar una señal.

Unidades	Edad	Sexo	PAS	PAD	Peso	Talla
1	×	×	?	?	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	×	×	?	?	×	×
101	×	×	×	×	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	×	×	×	×	?	?
201	×	×	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
300	×	×	?	?	?	?
301	×	×	×	×	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.000	×	×	×	×	×	×

Nota: El símbolo ? representa los valores ausentes y ×, los observados

Incompletos

Incompletos

Habrá veces que no tendremos buena distribución de los datos, cuando esto ocurra diremos que están "imbalanceados" o que no tienen dispersión suficiente.

Incompletos

Habrá veces que no tendremos buena distribución de los datos, cuando esto ocurra diremos que están "imbalanceados" o que no tienen dispersión suficiente.

¿Qué hacemos en este caso?

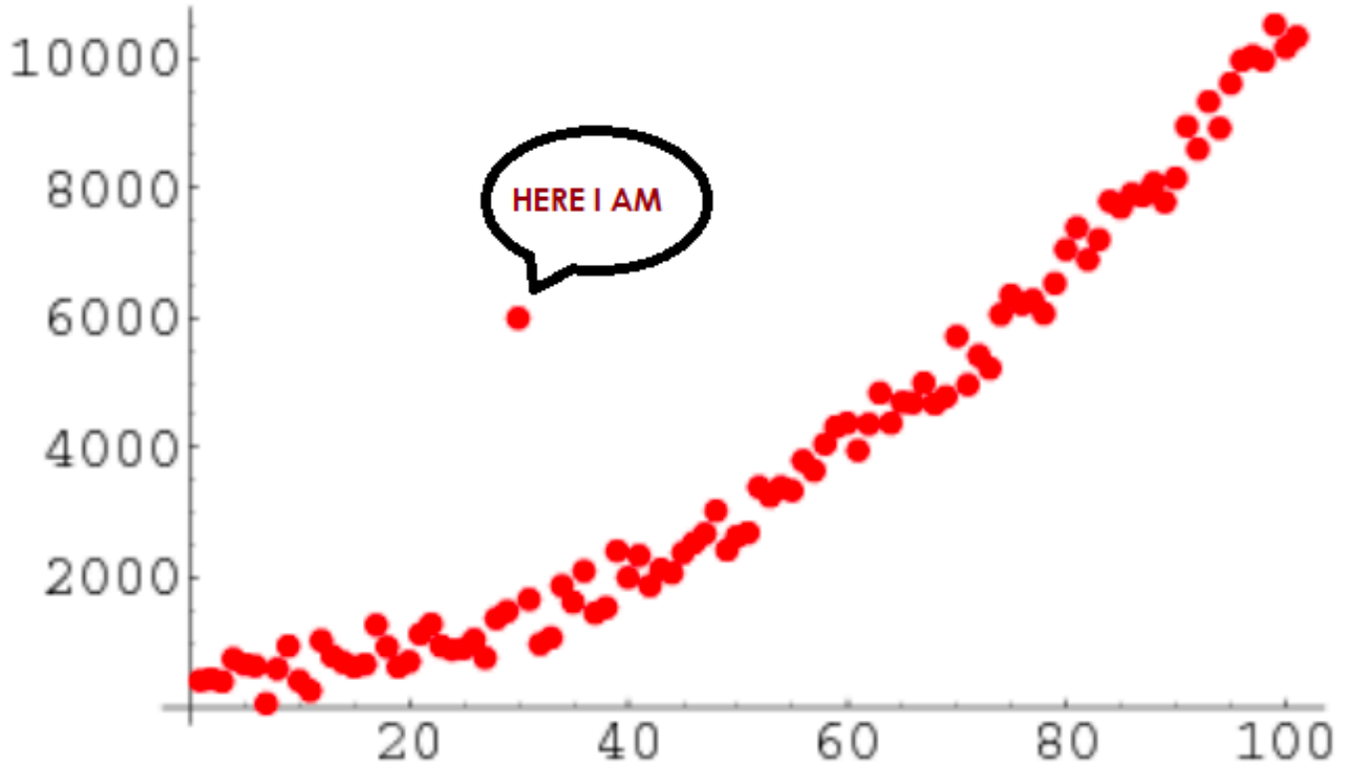
Valores Atípicos

Valores Atípicos

Valores fuera de la normal general...

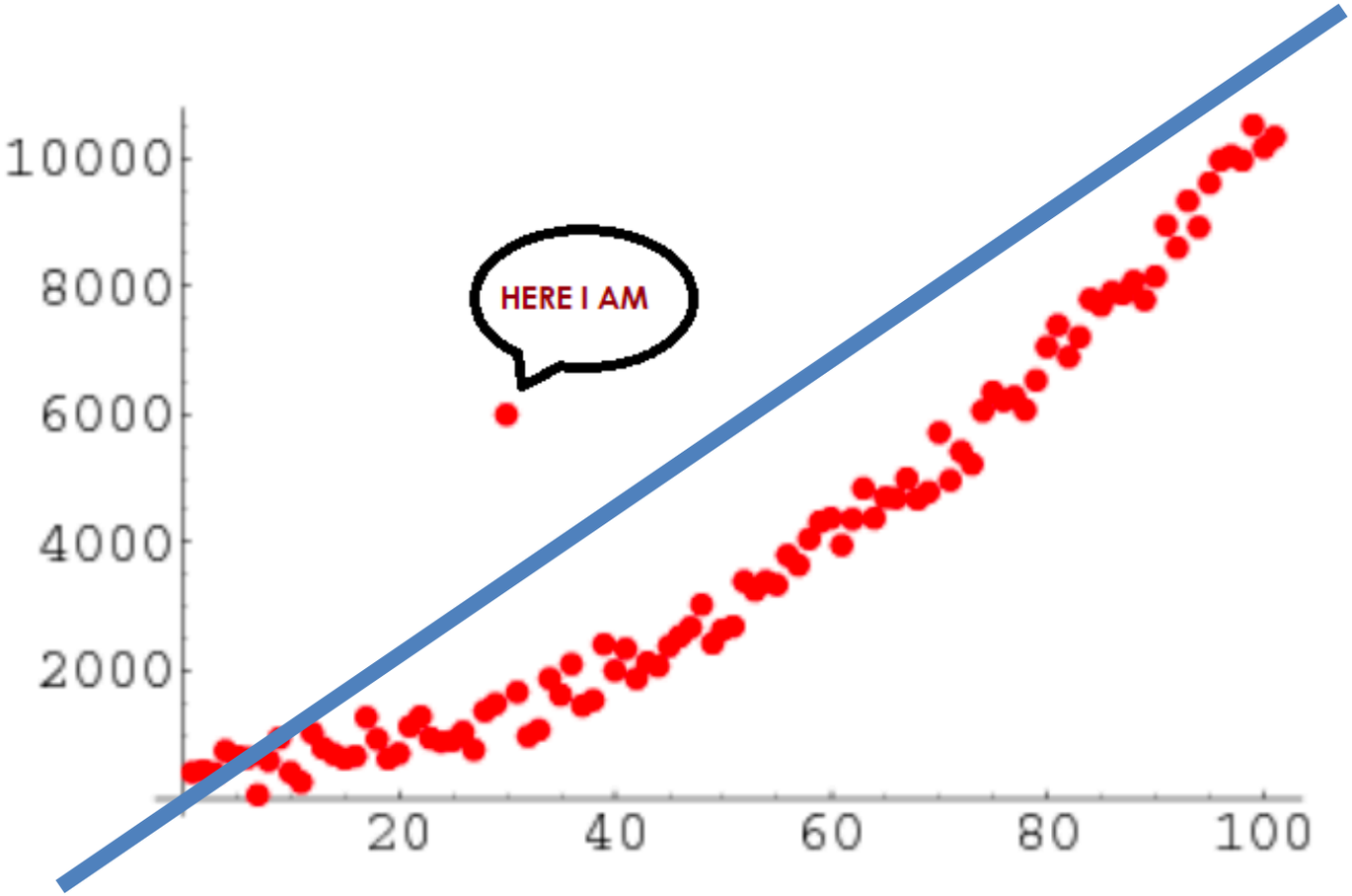
Valores Atípicos

Valores fuera de la normal general…



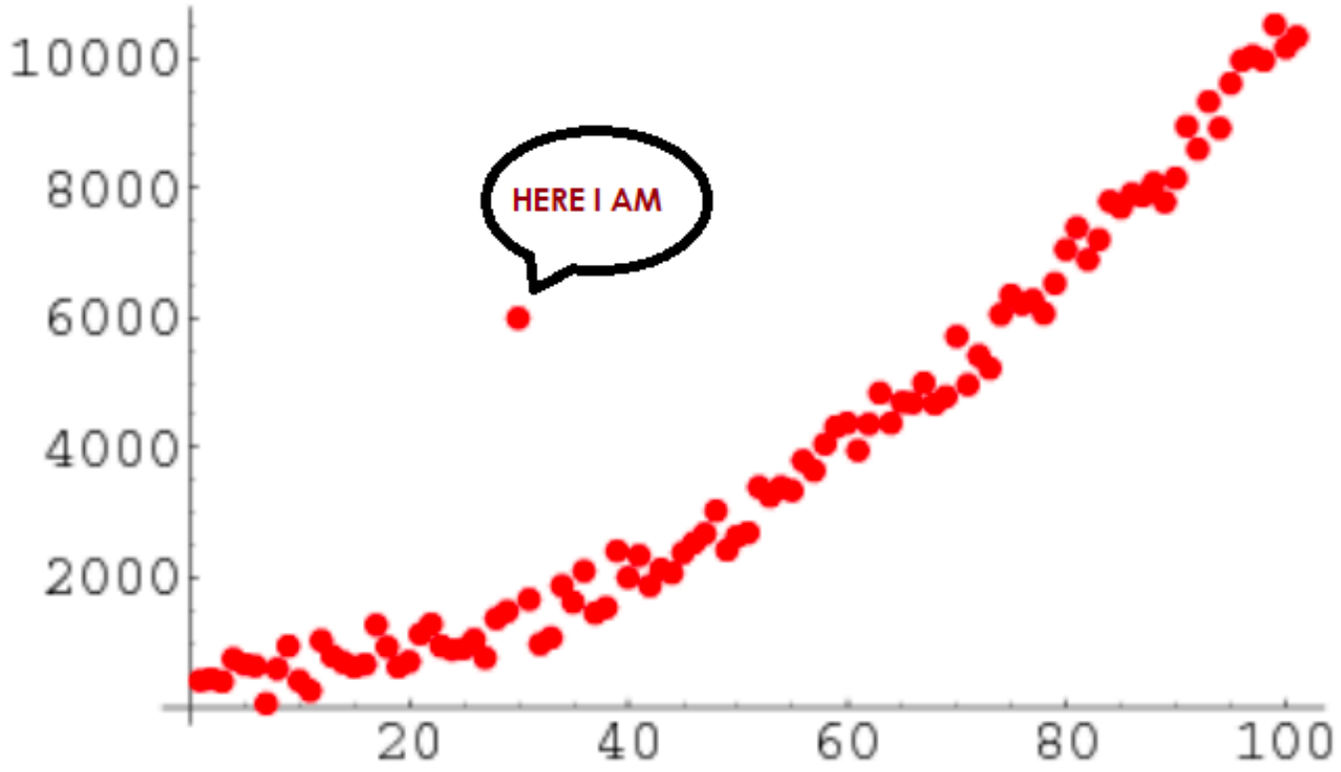
Valores Atípicos

Valores fuera de la normal general…

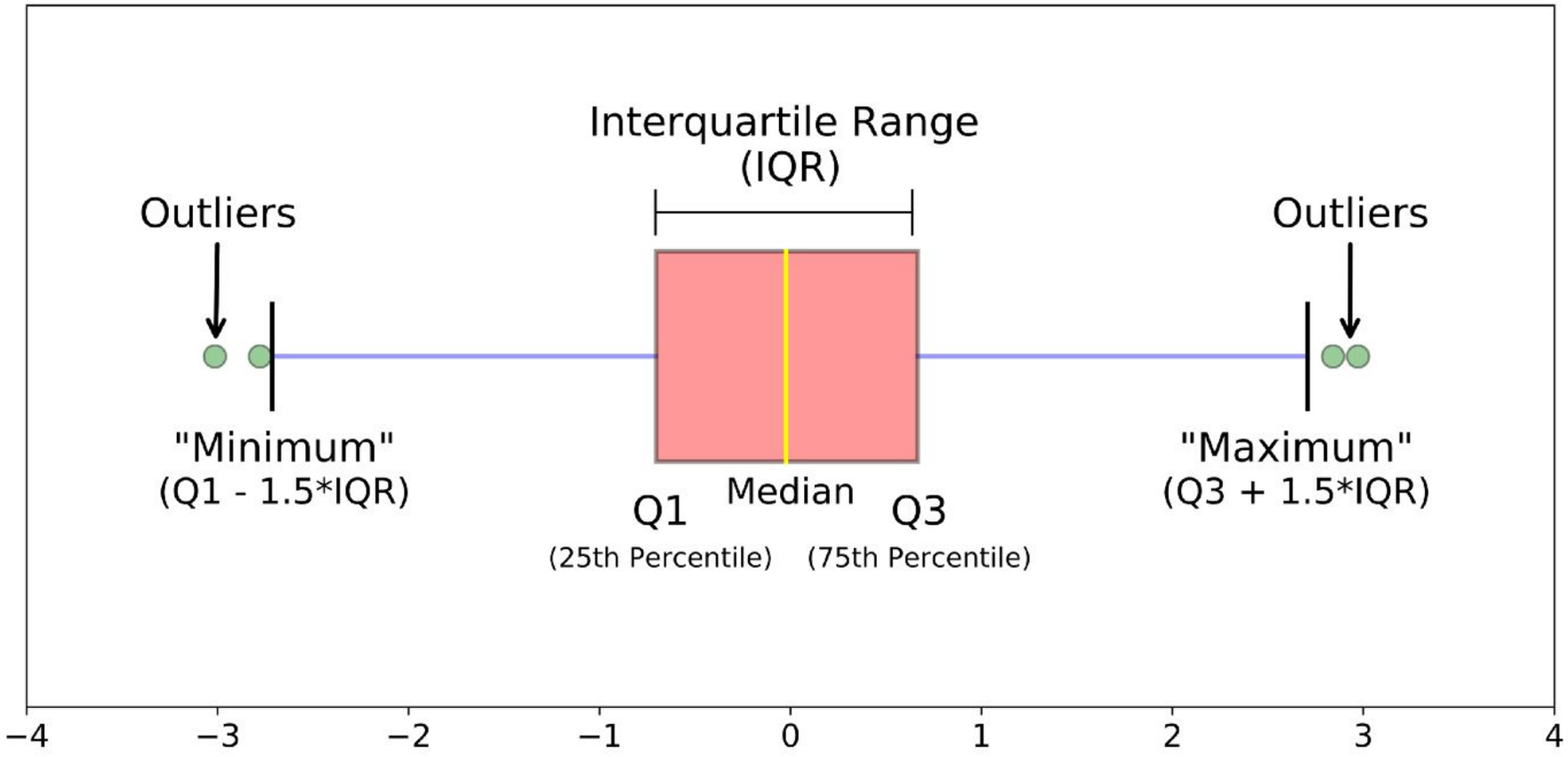


Valores Atípicos

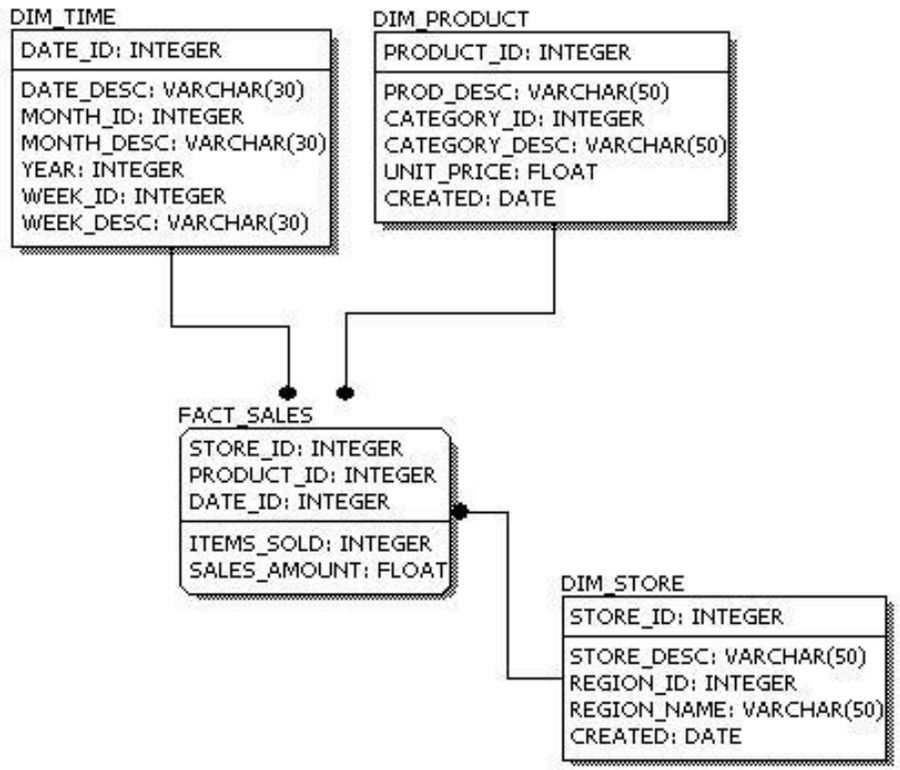
OUTLIERS!!!!



Algunas Métricas

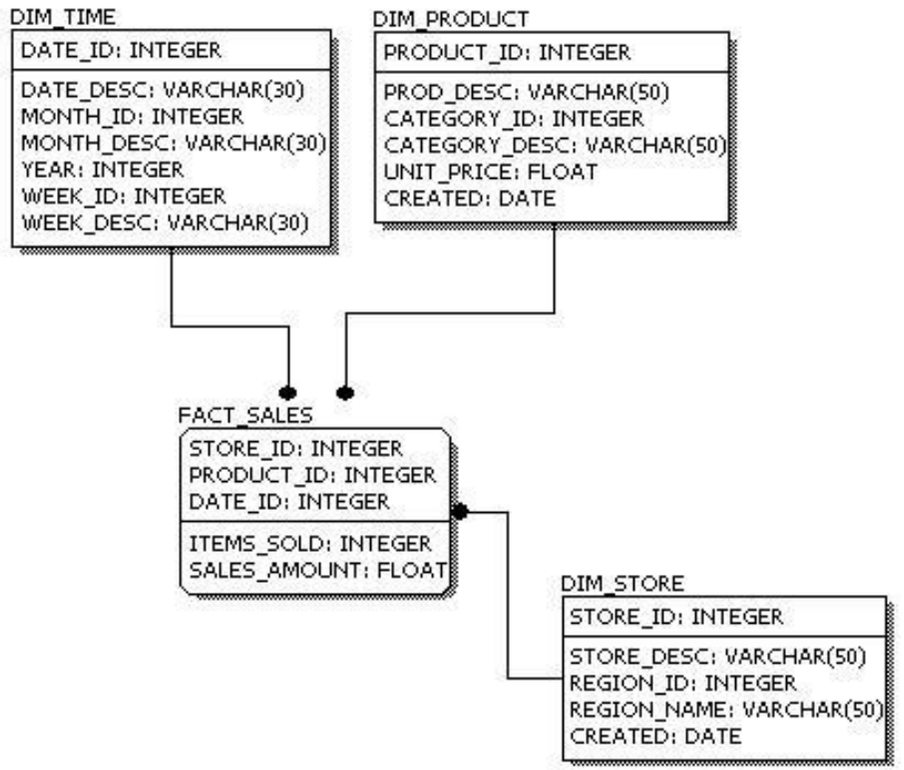


Incoherentes



Incoherentes

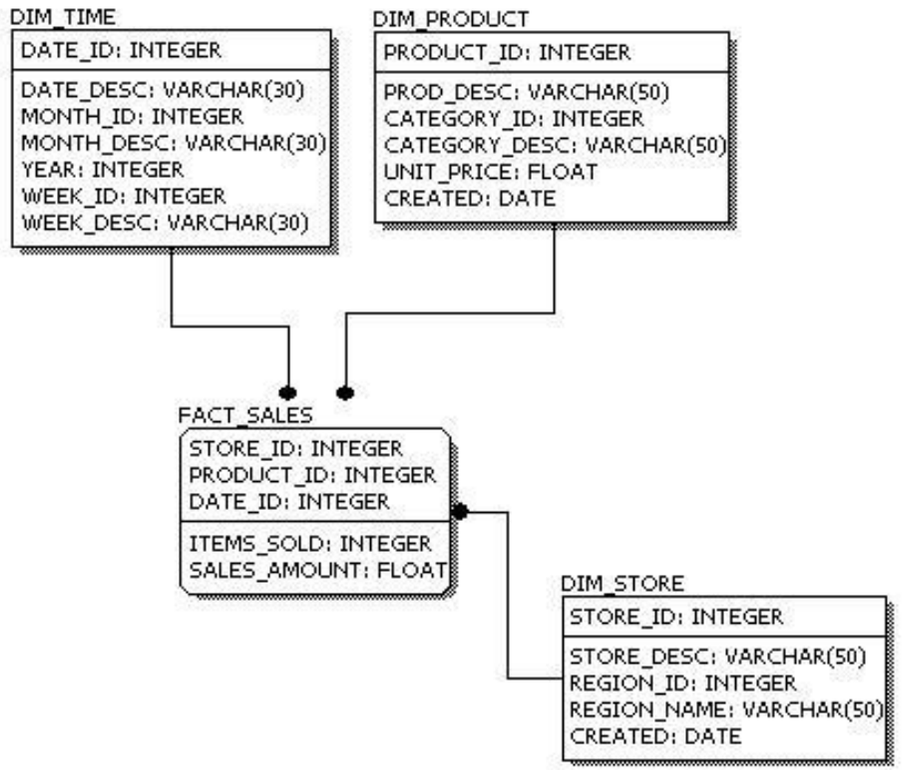
¿Cómo son los datos? **¿Cuál es el rango de valores posible?** **¿Existe un orden?**



Incoherentes

¿Cómo son los datos? **¿Cuál es el rango de valores posible?** **¿Existe un orden?**

¡Un Diccionario de Datos es la respuesta!



Ejercicios

Ejercicio 1

Ejercicio guiado de EDA:

Buscamos Outliers
Identificamos datos incompletos
Creamos un diccionario de Datos
…
Generamos datos Sintéticos para los imbalanceados

"El FSE invierte en tu futuro"

Fondo Social Europeo