Curso de Matemáticas para Data Science: **Estadística Descriptiva**

Francisco Camacho
@el_pachocamacho

[C1] Estadística descriptiva vs. inferencial

Introducción

Estadísticas de un jugador

- **Descriptiva:** resumir historial deportivo.
- **Inferencial:** predecir desempeño futuro del jugador.

Estadísticos descriptivos

Estadísticas

Estadísticas de Jugador

Resumen | Defensivo | Ofensivo | Distribución | Detallado

General | Local | Visitante — Mínimo jgdos | Todos los jugadores

	Jugador	Jgdos	Mins	Goles	Asist	Amar	Roja	TpP	AP%	Aéreos	JdelP	Rating
1	Lionel Messi Barcelona, 33, MP(CD),DL	25(2)	2303	23	8	4	-	5.6	85.4	0.3	17	8.54
2	Robert Lewandowski Bayern, 32, DL	24(1)	2103	35	6	3	-	4.4	76.8	1.7	9	8.05
3	Harry Kane Tottenham, 27, MP(C),DL	28	2457	19	13	1	-	3.9	69.5	2.4	11	7.82
4	Gerard Moreno Villarreal, 28, MP(CD),DL	24	2068	19	5	3	-	3.3	68.7	2.1	11	7.78

¿Puedes mentir con estadística?

- ➔ Dependerá de la definición de quién es el mejor jugador de fútbol.
- ➔ No hay una definición objetiva.
- ➔ Los diferentes estadísticos descriptivos dan nociones diferentes sobre los mismos datos.

“Con frecuencia construimos un caso estadístico con datos imperfectos, como resultado hay numerosas razones por las cuales individuos intelectuales respetables pueden no estar de acuerdo sobre los resultados estadísticos.”

El gran problema de la estadística descriptiva (Naked Statistics, Charles Wheelan).

NEW YORK TIMES BESTSELLER

naked statistics

STRIPPING THE DREAD FROM THE DATA

"Brilliant, funny
. . . the best math teacher
you never had."
—*San Francisco Chronicle*

charles wheelan

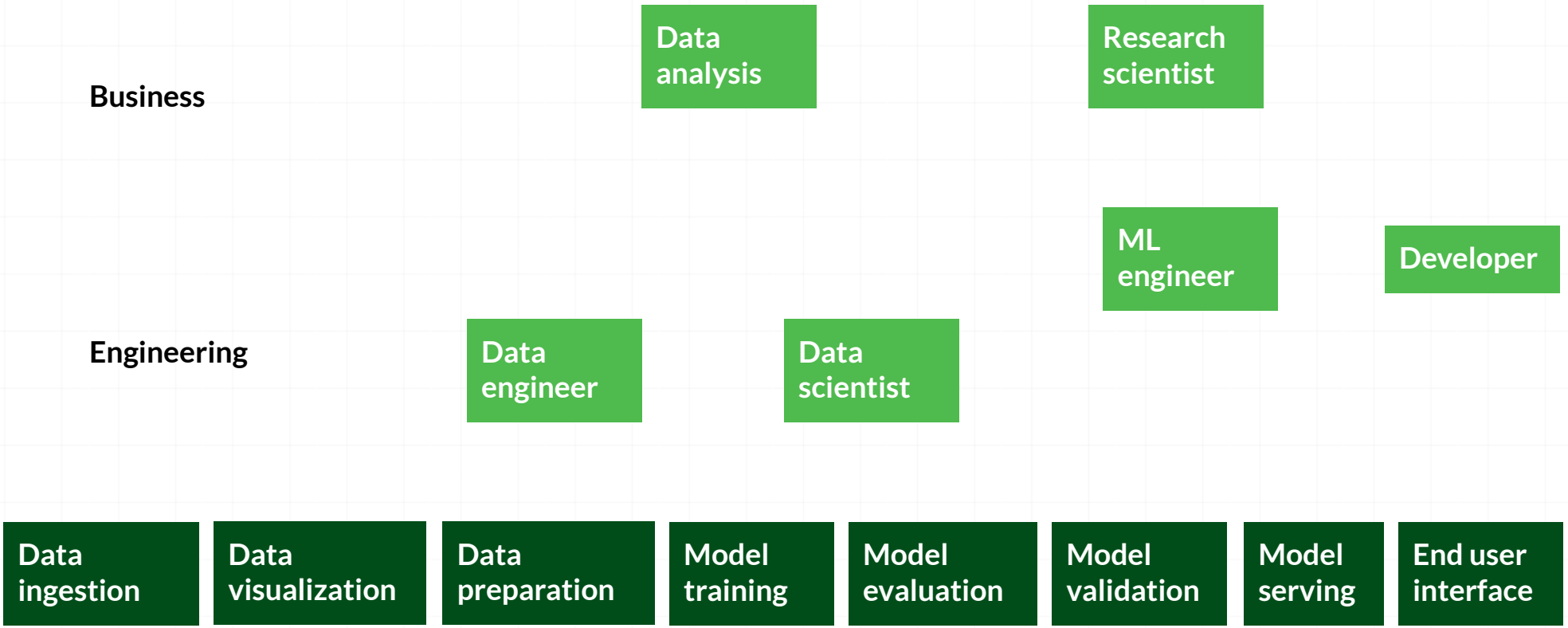
AUTHOR OF *NAKED ECONOMICS*

¿Porque aprender estadística?

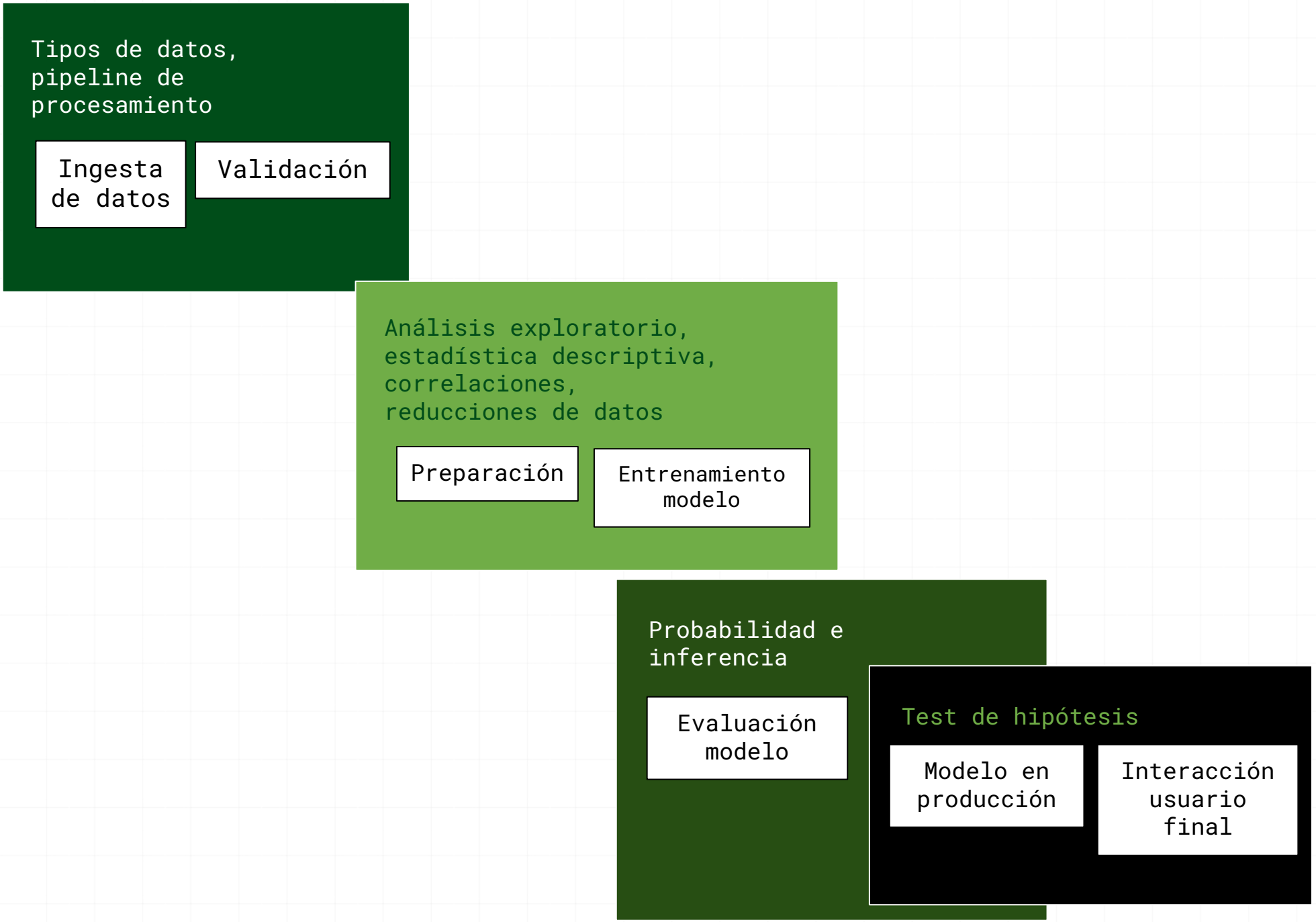
- ➔ Resumir grandes cantidades de información.
- ➔ Tomar mejores decisiones (¿o peores?).
- ➔ Responder preguntas con relevancia social.
- ➔ Reconocer patrones en los datos.
- ➔ Descubrir a quienes usan estas herramientas con fines nefastos.

[C2] Flujo de trabajo en Data Science

Introducción

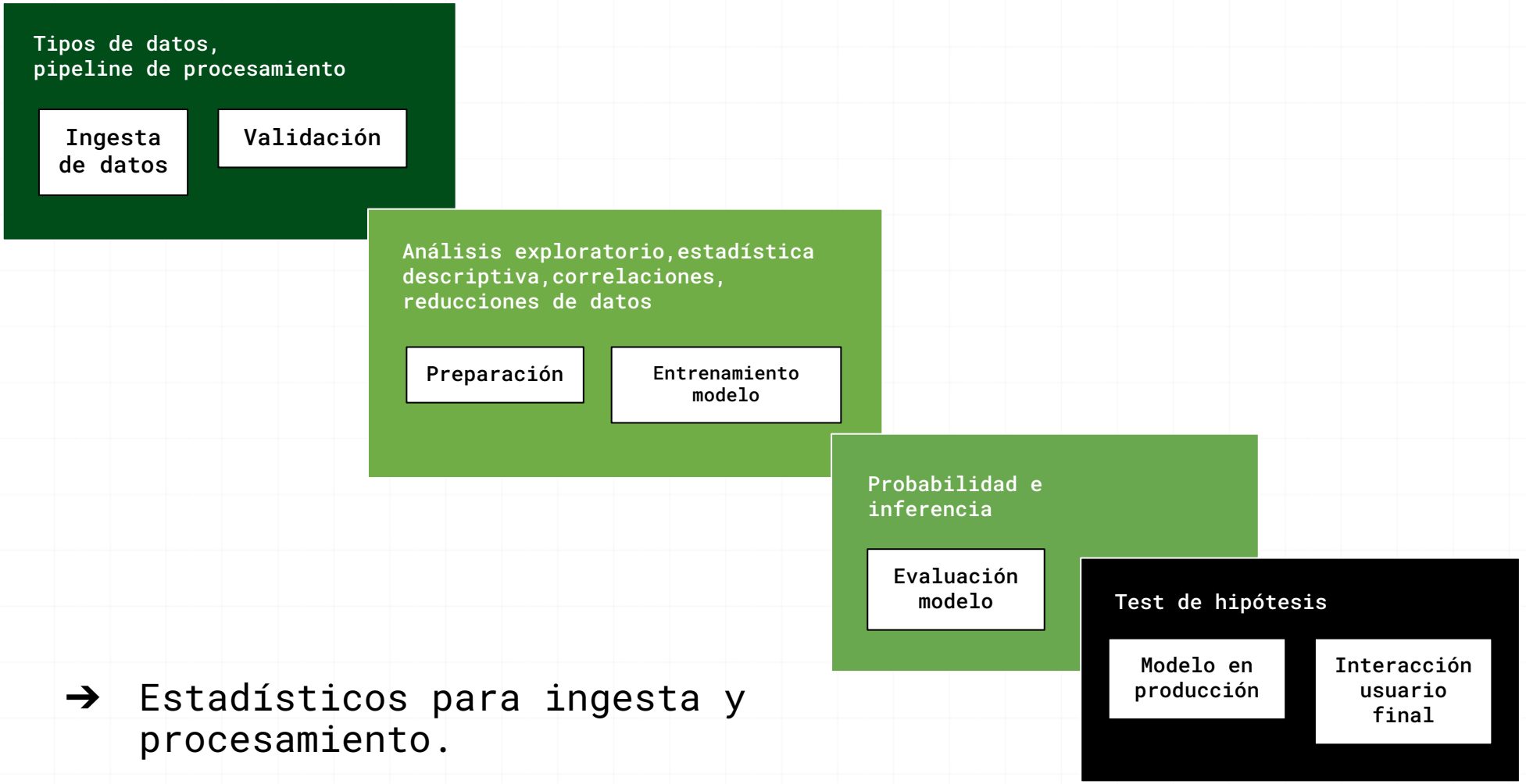


Pasos y roles en el flujo de trabajo de Data Science (Design Patterns in Machine Learning).


Tipos de datos, pipeline de procesamiento
Ingesta de datos
Validación
Análisis exploratorio, estadística descriptiva, correlaciones, reducciones de datos
Preparación
Entrenamiento modelo
Probabilidad e inferencia
Evaluación modelo
Test de hipótesis
Modelo en producción
Interacción usuario final

[C3] Plan del curso

Introducción



- ➔ Estadísticos para ingesta y procesamiento.
- ➔ Estadísticos para analítica y exploración.

[C4] Tipos de datos

Estadística en la ingesta de datos

Tipos de datos

Categóricos
(género, categoría de película, método de pago)

- ordinal
- nominal

Numéricos

(edad, altura, temperatura)

- discretos
- continuos


Ejercicio de identificación
de tipos de datos en python

import pandas as pd

df = pd.read_csv("dataset_sample.csv")

df.describe()

Deepnote

[C5] Medidas de tendencia central

Estadística en la ingesta de datos

¿Tendencia central?

- Media (promedio)
- Mediana (dato central)
- Moda (dato que más se repite)

Diagrama de frecuencias

16
17
16
18
.
.
.
15

20 estudiantes

Edad	Frecuencia
15	3
16	4
17	5
18	6
19	2

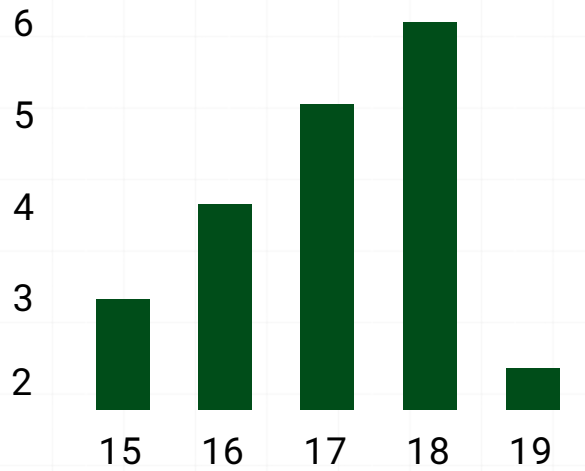


Diagrama de frecuencias

¿Cuándo usar cuál?

- La media es susceptible a valores atípicos.
- La moda no aplica para datos numéricos continuos.

[C6] Metáfora de Bill Gates en un bar

Estadística en la ingesta de datos

[C7] Medidas de tendencia central en Python

Estadística en la ingesta de datos

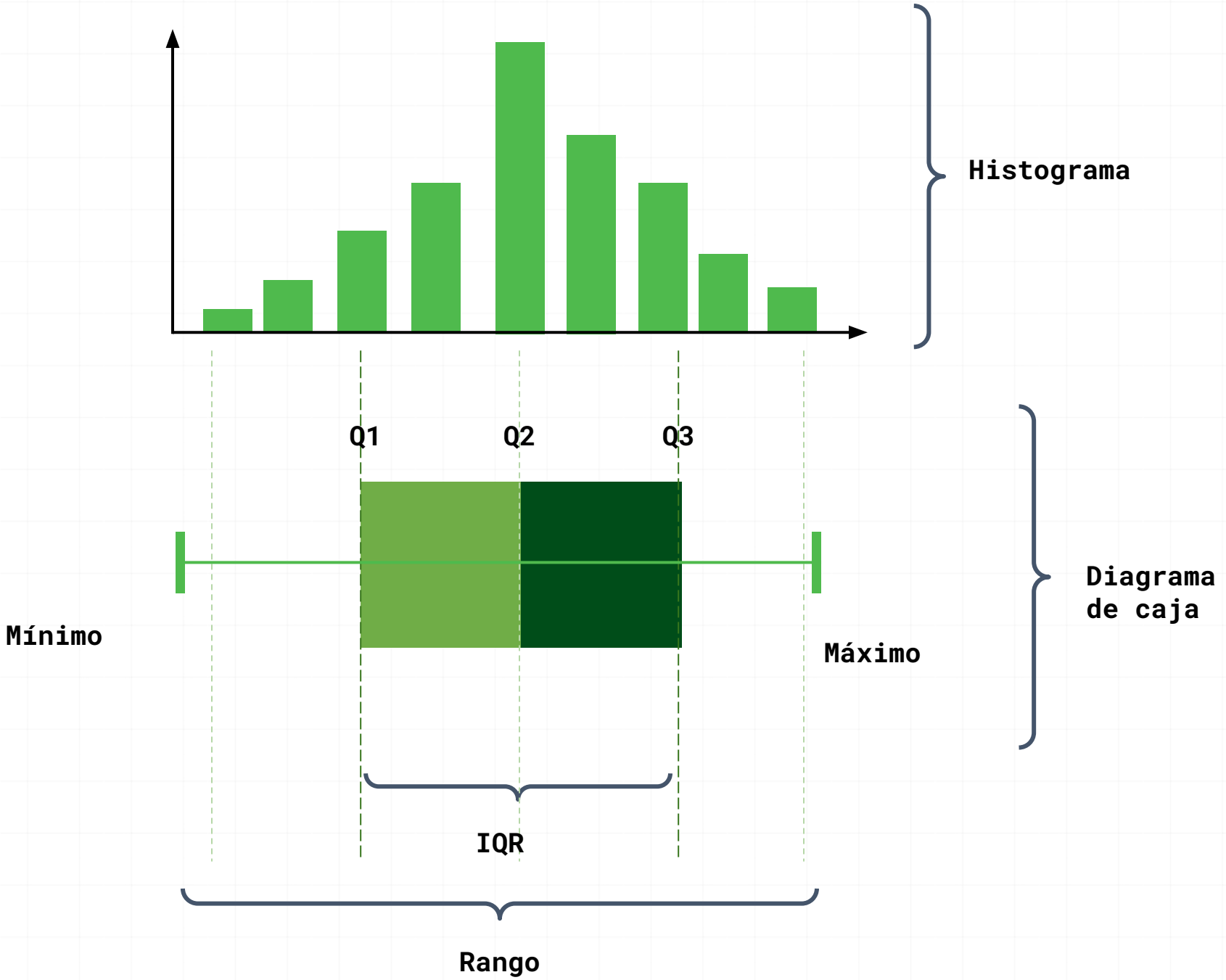
Deepnote

[C8] Medidas de dispersión

Estadística en la ingesta de datos

Dispersión en una distribución

- Rango
- Rango intercuartil
- Desviación estándar


Histograma
Q1
Q2
Q3
Mínimo
Máximo
Diagrama
de caja
IQR
Rango

[C9] Desviación estándar

Estadística en la ingesta de datos

[C10] Medidas de dispersión en Python

Estadística en la ingesta de datos

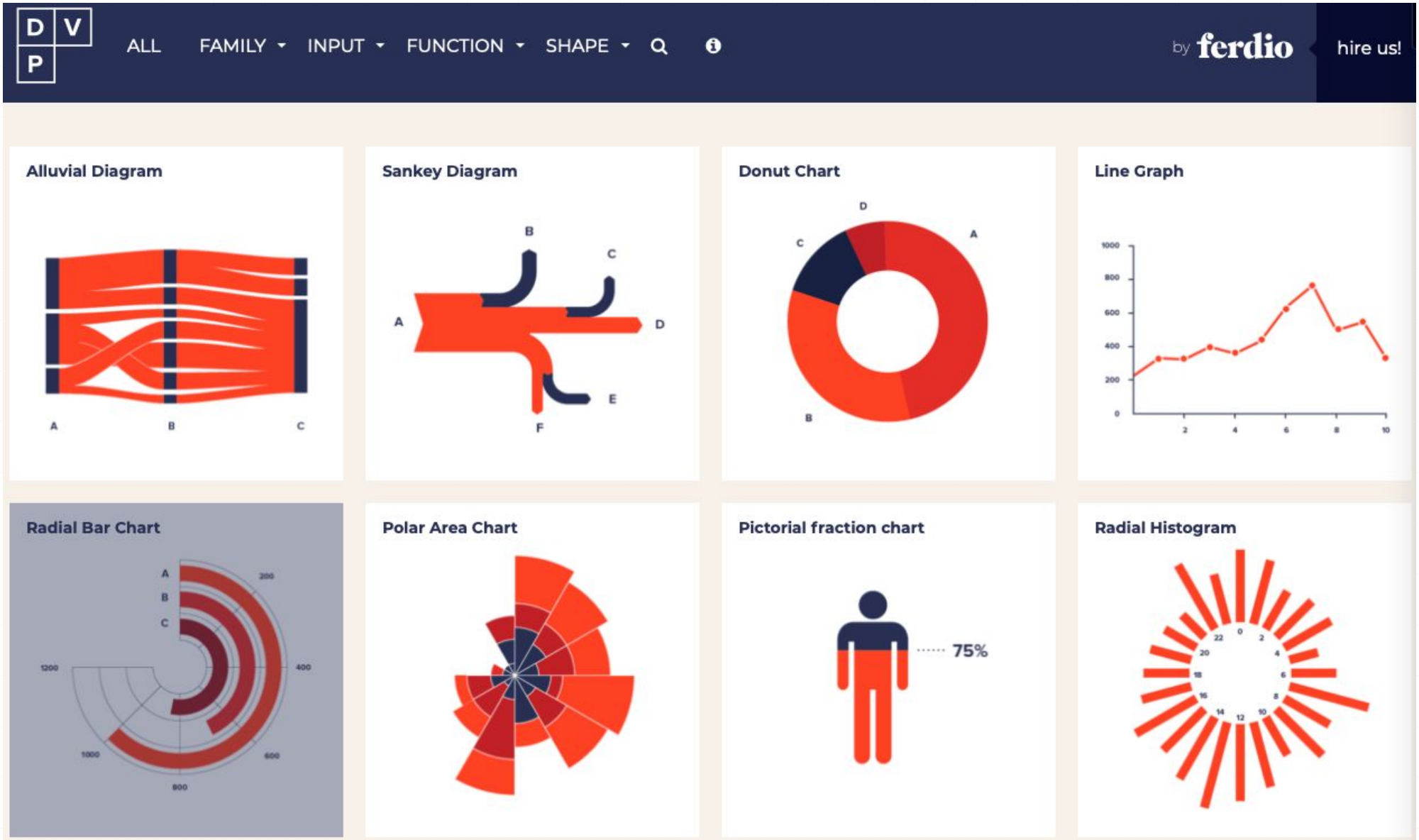
Deepnote

[C11]
Exploración visual de los datos

Estadística en la ingesta de datos

“Una imagen vale más que mil palabras.

Pero una buena imagen...”



[Datavizproject]

Explicación de algunas visualizaciones en el tablero ...

Deepnote

[C12] Diagramas de dispersión en el análisis de datos

Estadística en la ingesta de datos

[C13] Pipelines de procesamiento de datos numéricos

Estadística en la ingesta de datos

Escalamiento lineal

¿Por qué usarlos?

➔ Modelos de machine learning eficientes en el rango [-1, 1].

¿Hay diferentes tipos?

➔ max-min, Clipping, Z-score, Winsorizing, etc.

¿Cuándo usarlos?

➔ Data simétrica o uniformemente distribuida.

[C14] Transformación no lineal

Estadística en la ingesta de datos

Transformación no lineal

¿Por qué usarlos?

➔ Datos fuertemente sesgados, no simétricos.

¿Hay diferentes tipos?

➔ Logaritmos, sigmoides, polinomiales, etc.

¿Cuándo usarlos?

➔ ¡Antes de escalamientos lineales!

Ilustración de métodos de escalamiento en el tablero ...

[C15]
Procesamiento de datos numéricos en Python

Estadística en la ingesta de datos

```
# escalamiento con scikit

from sklearn import datasets, linear_model

X, Y = datasets.load_diabetes(..)
raw = X[:, None, 2]
max_r, min_r = max(raw), min(raw)

scaled = (2*raw - max_r - min_r)/ (max_r - min_r)
```

```
def train():
    linear_model.LinearRegression().fit(raw, Y)

def train_scaled():
    linear_model.LinearRegression().fit(scaled, Y)

raw_time = timeit.timeit(train, number=1000)
scaled_time = timeit.timeit(train_scaled, number=1000)
```

Deepnote

[C16] Pipelines de procesamiento de datos categóricos

Estadística en la ingesta de datos

Mapeos numéricos

Dummy

➔ Representación compacta.

➔ Mejor para inputs linealmente independientes.

One-hot

➔ Permite describir categorías no incluidas inicialmente

[C17]
Procesamiento de datos categóricos en python

Estadística en la ingesta de datos

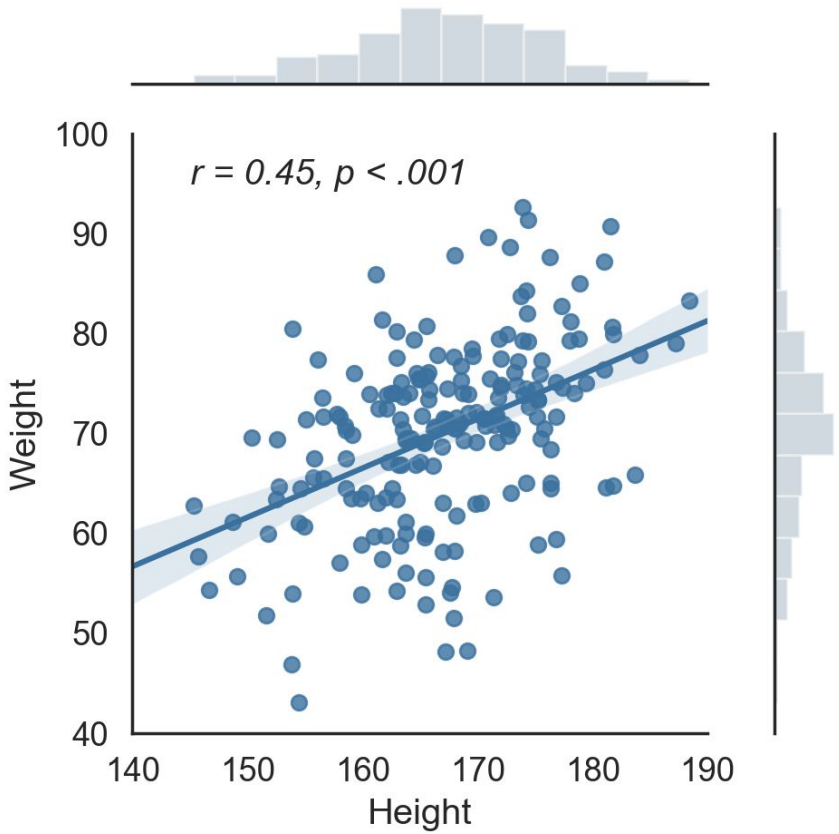
¿Tratar variables numéricas como categóricas?

Deepnote

[C18] Correlaciones

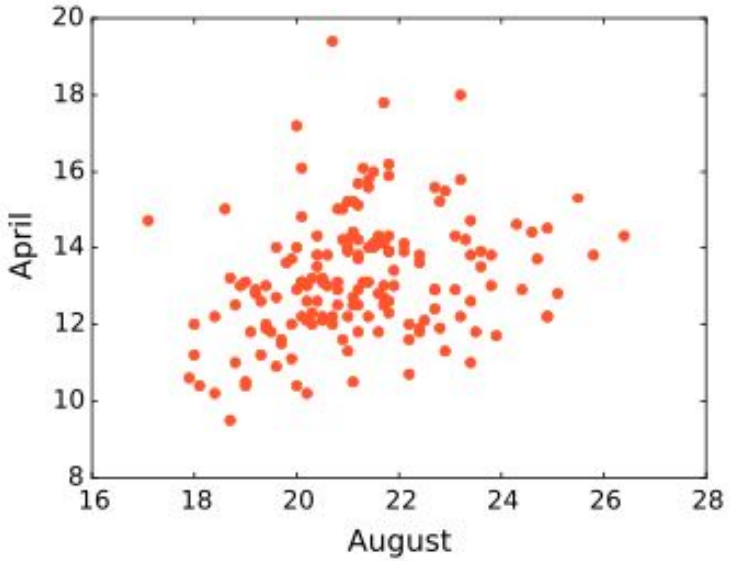
Estadística en la ingesta de datos

Scatterplot o gráfico de dispersión

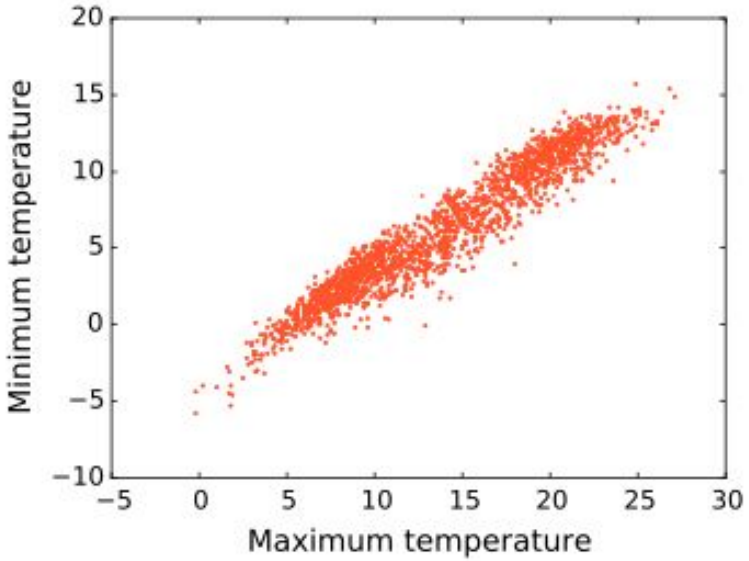


Covarianza y coeficiente de correlación

$\rho = 0.269$



$\rho = 0.962$



Detalle matemático de los conceptos de covarianza y coeficiente de correlación en el tablero ….

[C19] Matriz de covarianza

Estadística en la ingesta de datos

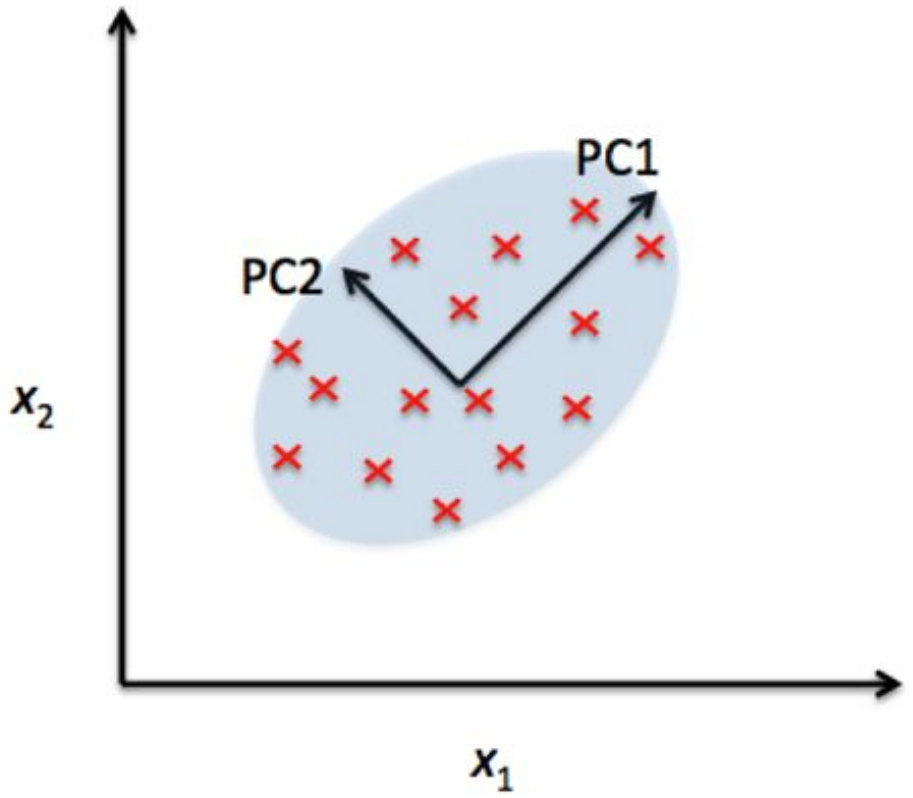
Desarrollo en el tablero ...

Deepnote

[C21] PCA: Análisis de componentes principales

Proyecto de aplicación

Reducción de dimensionalidad



Desarrollo en el tablero ...

[C22] Reducción de dimensionalidad con PCA

Proyecto de aplicación