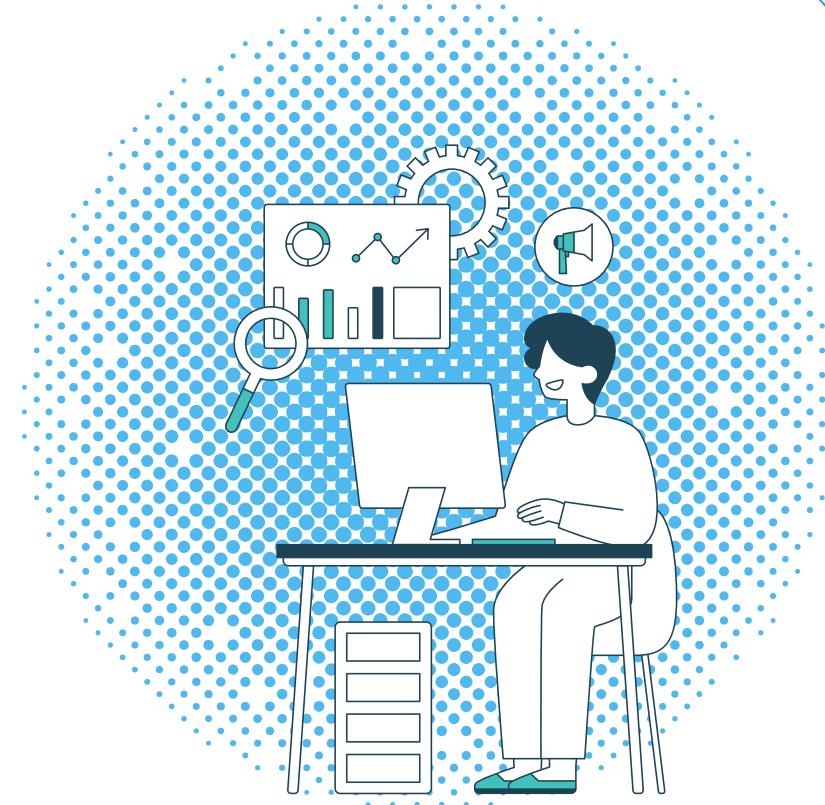


Análisis de datos

Carrera de Especialización en
Inteligencia Artificial

Bimestre 5B2025



Docentes

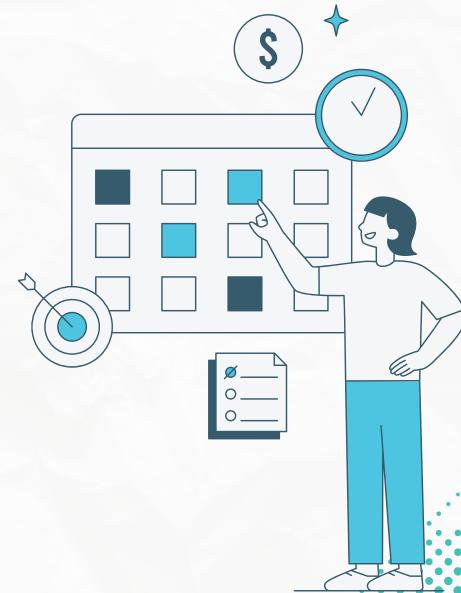


Esp. Lic. María Carina Roldán
macroldan@fi.uba.ar

Esp. Ing. Ariadna Garmendia
arigarmendia@gmail.com

Objetivos de la materia

- 1 Aprender a realizar un análisis exploratorio de datos en detalle.
- 2 Entender qué mecanismos existen para el tratamiento de datos y cuáles son recomendados para el entrenamiento de un modelo de ML.



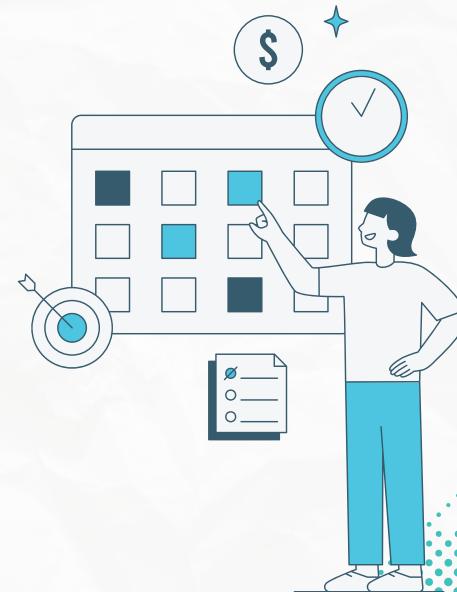
Enfoque

1

Vamos a ver un enfoque práctico y aplicado de los conceptos que se ven en otras materias (matemática, proba, etc.)

2

En esta materia analizamos y preparamos datos pero **no entrenamos modelos**.



Método de evaluación



- Para aprobar la materia deberán realizar **una exposición grupal** en la última clase, que este bimestre se desdoblará en dos sesiones (lo explicaremos al final).
- Se recomienda hacer avances y realizar consultas semanalmente, incorporando los conceptos vistos en cada clase.
- La presentación es obligatoria. No estar presente implica desaprobar la materia.
- La nota final surgirá de la calidad de la exposición.

Programa de la materia

1

Introducción al análisis de datos

23/10

Flujo de trabajo típico. Aplicaciones. Configuración del entorno. Introducción a las principales bibliotecas. Introducción al EDA.

2

Análisis exploratorio de datos (EDA)

30/10

EDA. Visualización de variables numéricas y categóricas. Estadística robusta. Identificación de valores faltantes y outliers. Relación entre variables.

3

Taller práctico - parte 1

6/11

Práctica en salas virtuales: EDA y visualización inicial de un dataset. Buenas prácticas. Puesta en común.

4

Preprocesamiento y limpieza de datos

13/11

Presentación de grupos de trabajo final. Prevención de data leakage. Tratamiento de datos faltantes. Tratamiento de outliers.

5

Feature engineering

20/11

Cardinalidad, codificación, discretización. Desbalance. Creación de nuevos features. Normalización y estandarización.

6

Reducción de dimensionalidad

27/11

Métodos de extracción y técnicas para selección de features. **Bonus:** EDA de datos no estructurados.

7

Taller práctico - parte 2

4/12

Práctica en salas virtuales: preprocesamiento y feature engineering de un dataset. Puesta en común. **Bonus:** Herramientas para automatizar el EDA.



Martes 9/12 - límite para compartir el material para la exposición final.

8

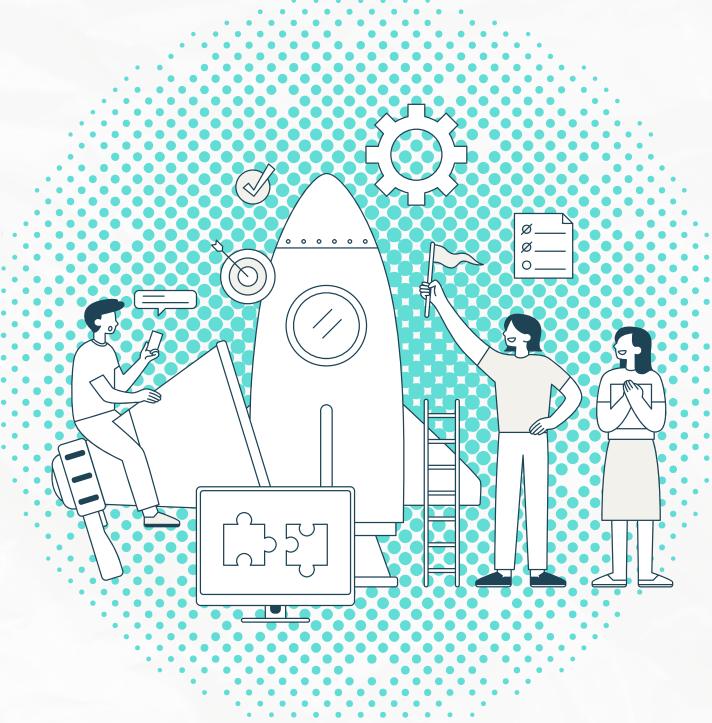
Exposición final

Exposición de trabajos finales

11/12 (desdoblada en 2 sesiones, 11/12 y 15/12).

Información útil

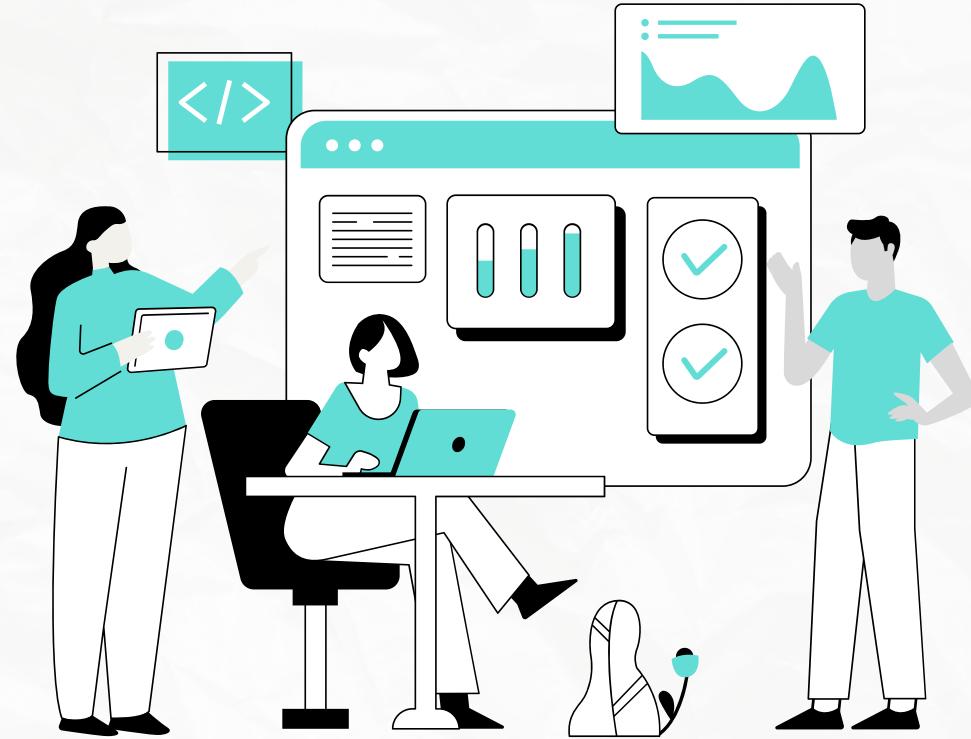
- 1 Material de clase: [Campus posgrado](#)
- 2 Notebooks: [Repositorio de la materia](#)



Repositorio e instalación



Análisis de datos



¿Qué es el análisis de datos?



Es el proceso de explorar, limpiar y modelar datos para extraer información útil y tomar decisiones.



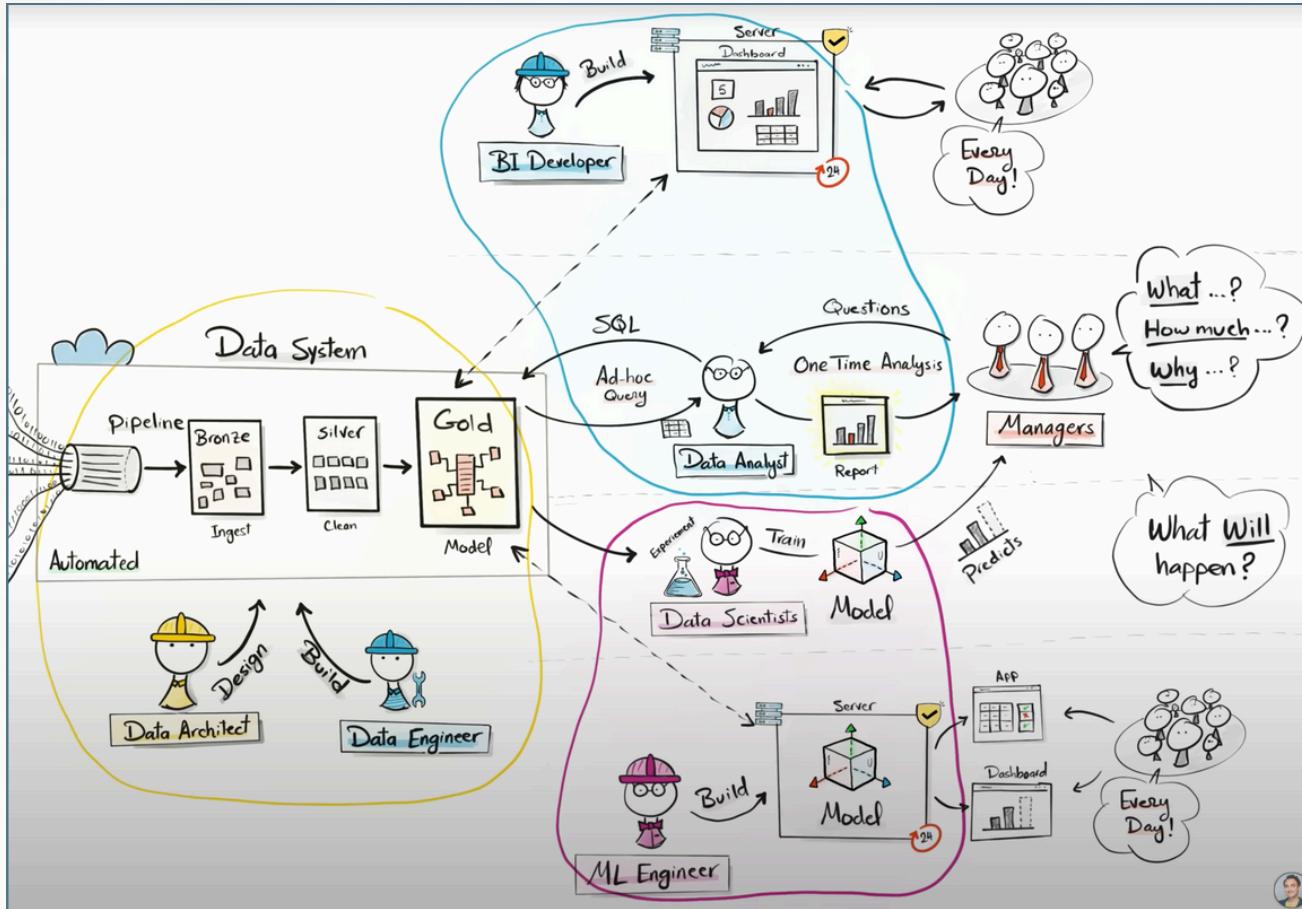
Es clave para garantizar que los modelos de IA trabajen con datos confiables y adecuados para el problema.





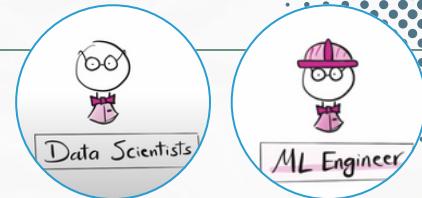
**¿Tienen experiencia
con análisis de datos?
¿Qué tareas incluye?**

Roles en datos y ML



Fuente: <https://www.youtube.com/watch?v=tyJ476aN CYU>

Flujo de trabajo del análisis de datos para IA



- 1 Identificar el problema



- 2 Recopilar los datos.



- 3 EDA y limpieza básica



- 5 Entrenar modelos



- 4 Análisis estadístico y preparación



Bibliotecas para la implementación del flujo de trabajo



Recolección de datos

BeautifulSoup

SQLAlchemy

Otras (otros tipos de datos, automatización)



Keras



YData



PYCARET

EDA, preprocesamiento, limpieza

pandas

NumPy

Análisis estadístico, reducción
de dimensionalidad

scikit
learn

SciPy

Visualización y análisis

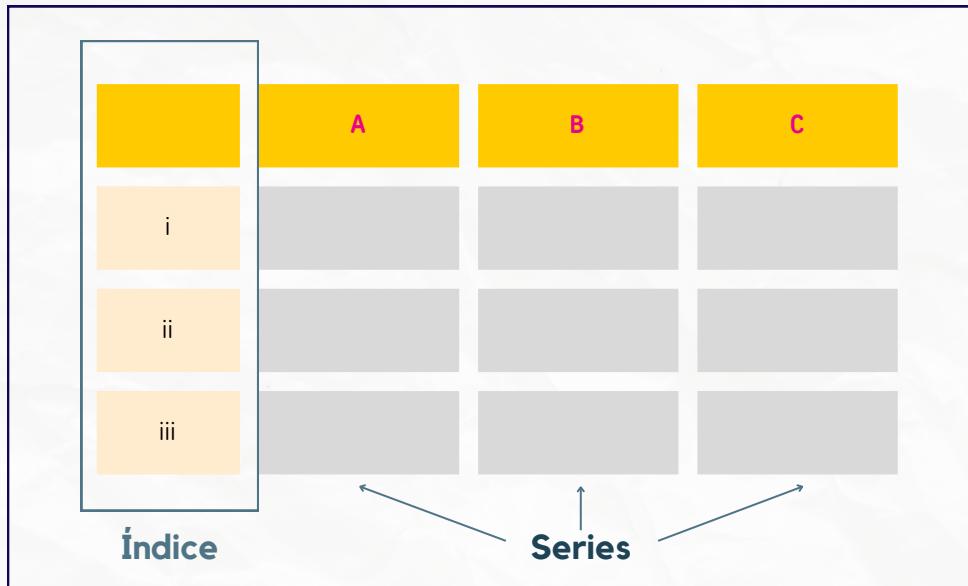
matplotlib

seaborn

statsmodels



Pandas. Terminología básica





Pandas. Documentación

Sitio oficial

<https://pandas.pydata.org/pandas-docs/stable/index.html>

pandas

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

[Install pandas now!](#)

Latest version: 2.2.3

- What's new in 2.2.3
- Release date: Sep 20, 2024
- Documentation (web)
- Download source code

Follow us

Recommended books

Python for Data Analysis by Wes McKinney

Effective Pandas 2 by Wes McKinney

Getting started

- Install pandas
- Getting started

Documentation

- User guide
- API reference
- Contributing to pandas
- Ask a question
- Ecosystem
- Release notes

Community

- About pandas
- Ask a question
- Ecosystem

With the support of:

Cheatsheet

Data Wrangling

with pandas Cheat Sheet
<http://pandas.pydata.org>

Tidy Data – A foundation for wrangling in pandas

In a tidy data set:

 &
 Each variable is saved in its own column
 Each observation is saved in its own row

Creating DataFrames

`df = pd.DataFrame({
 'a': [1, 2, 3],
 'b': [4, 5, 6],
 'c': [7, 8, 9],
}, index=[1, 2, 3])`

Specify values for each column.

Reshaping Data – Change layout, sorting, reindexing, renaming

`pd.melt(df)
Gather columns into rows`

`df.pivot(columns='var', values='val')
Spread values into columns.`

`pd.concat([df1, df2], axis=1)
Append columns of DataFrames`

Specify values for each row.

Subset Observations - rows

`df[df['Length'] > 7]
Extract rows meeting logical condition.`

`df.drop_duplicates()
Remove duplicate rows (only considers columns).`

`df.sample(frac=0.5)
Randomly select 50% of rows.`

`df.sample(n=10)
Randomly select 10 rows.`

`df.nlargest(n, 'value')
Select n largest values.`

`df.nsmallest(n, 'value')
Select n smallest values.`

`df.iat[0, 0]
Select first n entries.`

`df.iat[0, 0]
Select first n entries.`

`df.iat[0, 0]
Select first n entries.`

Subset Variables - columns

`df[['Width', 'Length', 'Species']]
Select multiple columns with specific names.`

`df['Species'].str.contains('green')
Select single column with specific name.`

`df['Species'].str.contains('green').all()
Select if all elements in column matches regular expression regex.`

Subsets - rows and columns

`df.loc[[1, 2], [1, 2, 3]]
Use df.loc[[]] and df.iloc[[]] to select only rows, only columns or both.`

`df.loc[1:2, 1:3]
Use df.loc[1:2] and df.iloc[1:2] to access a single value by row and column.`

`df.iloc[1:2, 1:3]
First index selects rows, second index: columns.`

`df.iat[1:2, 1:3]
Select columns in positions 1, 2 and 3 (first column is 0).`

`df.iat[1, 1:3]
Select rows in positions 1 and 2 (first row is 0).`

`df.iat[1, 1]
Select single value by index.`

Method Chaining

Most pandas methods return a DataFrame so that another pandas method can be applied to the result. This improves readability of code.

`df = (pd.DataFrame({
 'category': ['cat', 'dog', 'cat', 'dog'],
 'variable': 'var1',
 'value': 100
})
 .query("value > 200"))
 .groupby("category")
 .mean()`

Logic in Python (and pandas)

<code>< less than</code>	<code>> greater than</code>	<code>!= not equal</code>
<code>>= greater than or equal</code>	<code>df.column.isin(values)</code>	<code>isin(values)</code>
<code>== equals</code>	<code>df.isnull()</code>	<code>isna()</code>
<code><= less than or equal</code>	<code>df.notnull()</code>	<code>isnotna()</code>
<code>>= greater than or equal</code>	<code>df.iat[0, 0]</code>	<code>iloc[0, 0]</code>

Regular Expression Examples

<code>^/ matches strings containing a pattern /</code>	<code>'Length'</code>	<code>Matches strings ending with word 'Length'</code>
<code>^/s matches strings beginning with the word 'S' or 's'</code>	<code>'Spec1'</code>	<code>Matches strings beginning with the word 'Spec1'</code>
<code>^/s*\$/ matches strings beginning with 'S' and ending with 's'</code>	<code>'*1-3\$'</code>	<code>Matches strings beginning with 'S' and ending with 's'</code>
<code>^/[^s]*\$/ matches strings except the string 'Spec1'</code>	<code>df.Spec1</code>	<code>df.Spec1</code>

Ejemplos prácticos en Jupyter



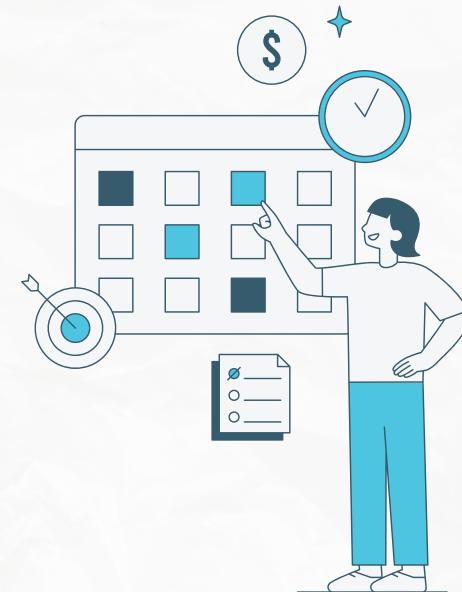


¿Qué es el análisis exploratorio de datos (EDA)?

- El análisis exploratorio sirve para resumir, visualizar y entender los datos.
- Útil para tomar decisiones sobre el manejo de outliers y datos faltantes.
- Ayuda a identificar transformaciones necesarias antes de entrenar modelos de ML: escalamiento, normalización, balance, codificación, discretización, etc.
- Facilita la identificación de patrones, relaciones y la evaluación de la relevancia de los datos.

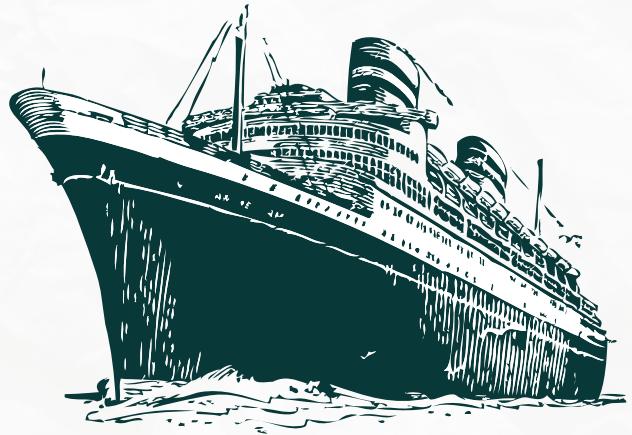
Variables aleatorias y datasets

- Una variable aleatoria (v.a.) es una forma de mapear los resultados de algún fenómeno aleatorio a números.
- Constituyen una herramienta fundamental para modelar y analizar fenómenos que involucran incertidumbre o aleatoriedad.
- Las columnas de un dataset pueden considerarse como un conjunto de observaciones (una muestra) de una variable aleatoria, tomada de una población más grande.



El dataset del Titanic

- Es un dataset muy famoso para aprender IA.
- Lo vamos a estar usando a lo largo del curso para aprender ciertos conceptos.
- Contiene información de pasajeros del barco, sus características (entre otros datos, contiene: edad, género, clase en la que viajaban, costo del pasaje, con cuántos familiares viajaban, etc.), y si sobrevivieron o no.



Ejemplos de v. aleatorias en el dataset del Titanic

Columna	Fenómeno aleatorio
Survived (¿sobrevivió?)	Varía aleatoriamente según factores como la ubicación en el barco, el acceso a botes salvavidas, la clase social y decisiones personales durante el naufragio.
Sex	Varía aleatoriamente según la composición de quienes decidieron o pudieron viajar. Influenciada por factores sociales como roles de género o motivos del viaje.
Age	Variable en función de la población que decidió o pudo viajar en el Titanic. Condicionada por el año de nacimiento de las personas y las circunstancias que los llevaron a ese viaje.
SibSp (cantidad de hermanos o esposo/a)	Influenciada por decisiones familiares (ej., viajar juntos), tamaño de la familia y factores sociales de la época.
Fare (tarifa)	Variable según factores como la clase (primera, segunda, tercera), el puerto de embarque o decisiones individuales de compra.
Embarked (puerto de embarque)	Influenciada por el lugar de origen, itinerario personal y disponibilidad de pasajes desde cada puerto.

Tipos de variables

Cosas que puedo contar o medir

Cuantitativas (numéricas)

Discretas

Toman valores enteros

Cantidad de hijos

Número de visitas al médico

Número de ventas

Continuas

Pueden tomar cualquier valor dentro de un rango

Temperatura

Altura

Edad

Cualidades o información que no se puede medir

Categóricas (cualitativas o atributos)

Nominales

Sin orden

Estado civil

Nombre de ciudad

Día de la semana

Ordinales

Tienen un orden específico (jerarquía o magnitud)

Tallas de ropa (S, M, L)

Opiniones (muy de acuerdo, de acuerdo,...)

Nivel de educación (secundario, terciario, universitario)

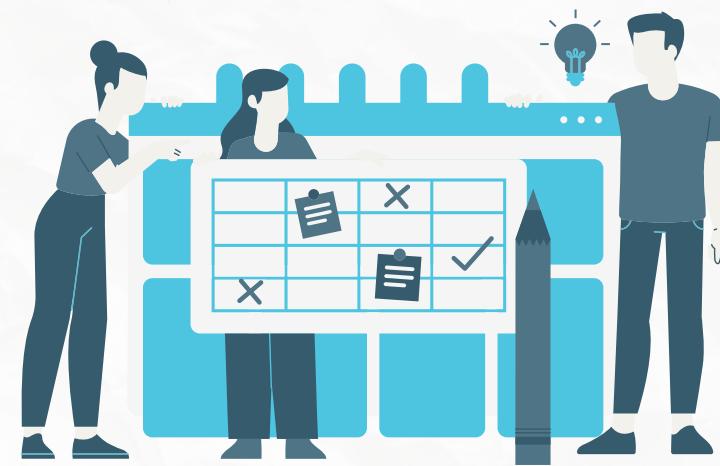
Algunas consideraciones interesantes

- Una variable categórica se puede representar con números (que actúan como un encoding).
- La **edad** en general se considera una variable continua pero generalmente se mide y reporta como discreta (Ej., en años). → Naturaleza de los datos vs. cómo se refleja.
- Las variables **binarias** (Sí/No, 1/0, True/False) son un caso particular de las categóricas nominales.
- Las **fechas** tienen una doble naturaleza:
 - Numérica, cuando interesa medir tiempo transcurrido o tendencias.
 - Categórica, cuando interesa agrupar eventos en una fecha específica (Ej., analizar las ventas en un día particular de la semana).
- ¿Qué pasa con las **coordenadas geográficas**?



Ejercicio opcional

Resolver el ejercicio de la Notebook
clase_01_ejercicio.ipynb (se
encuentra en la carpeta "recursos"
del repositorio).



Descripción del trabajo y exposición final



Armado de grupos para el trabajo final



Planilla con los grupos

Plazo para tener definidos los grupos de trabajo: **jueves 30/10/25**

¿Te quedaron dudas?



- Recordá que podés consultarnos durante la clase prendiendo la cámara y hablando, o simplemente escribiendo en el chat.
- También podés escribirnos por correo las veces que sea necesario (por favor incluir a ambas docentes).



Esp. Lic. María Carina Roldán
macroldan@fi.uba.ar



Esp. Ing. Ariadna Garmendia
arigarmendia@gmail.com

Y si tenés un feedback, un pedido, una sugerencia, etc. y no te animás a decirlo, podés dejarlo por escrito en la encuesta voluntaria y anónima (disponible durante todo el bimestre).