

Análisis de Datos

Tema 0 - Introducción

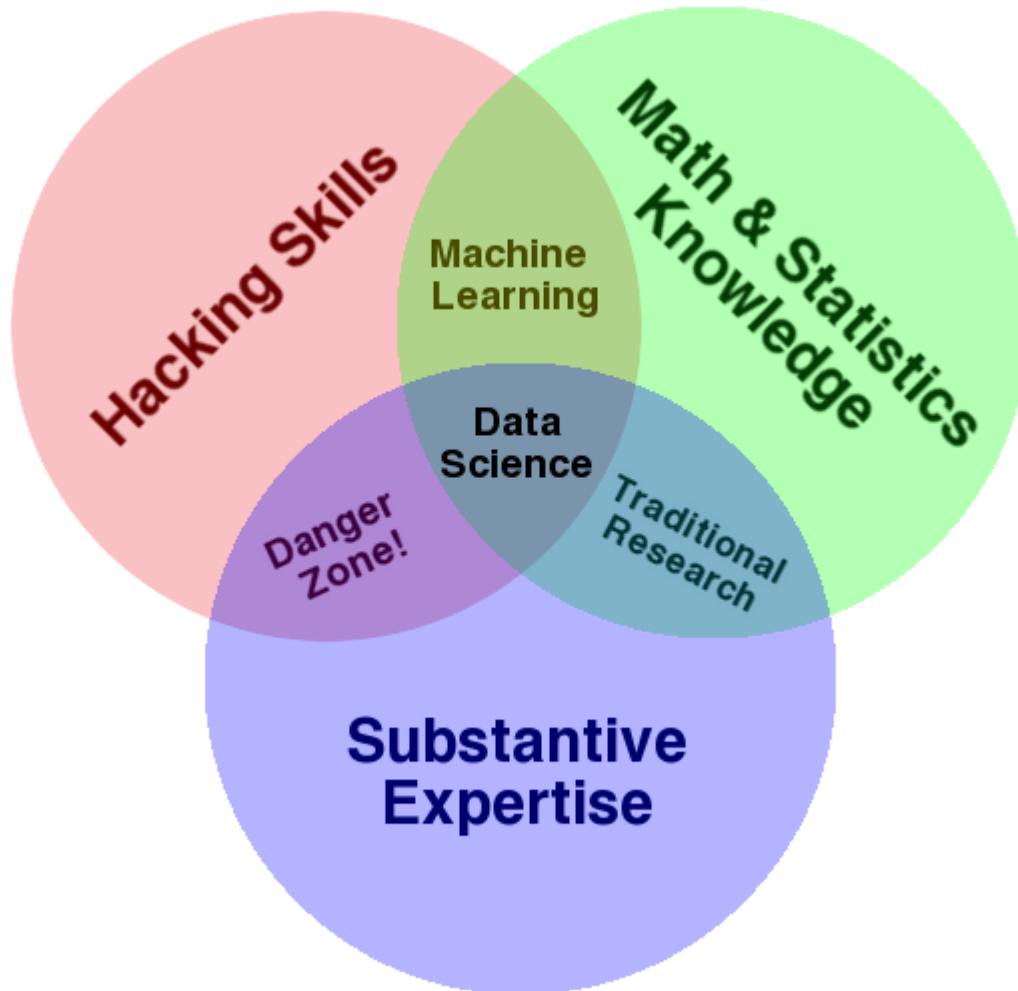
Roi Naveiro

¿Qué es la Ciencia de Datos?

*"Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to **understand and analyze** actual phenomena with **data**. It employs techniques and theories drawn from many fields within the context of **mathematics, statistics, information science, and computer science**."*

-Wikipedia

¿Qué es la Ciencia de Datos?



¿Qué son los Datos?

"A collection of discrete units of information that in their most basic forms convey quantity, quality, fact, statistics, or other basic units of meaning."

[-Wikipedia](#)

¡Definición bastante vaga!

Tipos de Datos

- Cualquier unidad de información es un dato
- Una distinción importante
 - Datos estructurados
 - Datos no estructurados

Datos Estructurados

Datos tabulares

solutions-jun3.csv

New Open Save Print Import Copy Paste Format Undo Redo AutoSum Sort A-Z Sort Z-A Gallery Toolbox Zoom Help

Verdana 10 B I U

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|----|----|------------|------------|------------|------------|-----------|--------|---|---|---|---|---|---|---|---|---|
| | id | problem_id | subject_id | start | stop | time_left | answer | | | | | | | | | |
| 2 | 1 | 498 | 17 | 1307119989 | 1307120016 | 2369 | A | | | | | | | | | |
| 3 | 2 | 150 | 15 | 1307119991 | 1307120009 | 2376 | D | | | | | | | | | |
| 4 | 3 | 313 | 16 | 1307119994 | 1307120009 | 2376 | E | | | | | | | | | |
| 5 | 4 | 12 | 13 | 1307119995 | 1307120019 | 2366 | B | | | | | | | | | |
| 6 | 5 | 273 | 14 | 1307119996 | 1307120028 | 2357 | A | | | | | | | | | |
| 7 | 6 | 101 | 19 | 1307119996 | 1307120021 | 2364 | B | | | | | | | | | |
| 8 | 7 | 105 | 18 | 1307119998 | 1307120048 | 2337 | B | | | | | | | | | |
| 9 | 8 | 162 | 12 | 1307120004 | 1307120042 | 2343 | C | | | | | | | | | |
| 10 | 9 | 70 | 15 | 1307120011 | 1307120038 | 2347 | C | | | | | | | | | |
| 11 | 10 | 300 | 16 | 1307120012 | 1307120092 | 2293 | B | | | | | | | | | |
| 12 | 11 | 494 | 17 | 1307120017 | 1307120075 | 2310 | D | | | | | | | | | |
| 13 | 12 | 357 | 13 | 1307120021 | 1307120118 | 2267 | A | | | | | | | | | |
| 14 | 13 | 522 | 19 | 1307120025 | 1307120152 | 2233 | D | | | | | | | | | |
| 15 | 14 | 232 | 14 | 1307120030 | 1307120158 | 2227 | C | | | | | | | | | |
| 16 | 15 | 344 | 15 | 1307120041 | 1307120117 | 2268 | B | | | | | | | | | |
| 17 | 16 | 160 | 17 | 1307120079 | 1307120249 | 2136 | D | | | | | | | | | |
| 18 | 17 | 516 | 16 | 1307120094 | 1307120159 | 2226 | B | | | | | | | | | |
| 19 | 18 | 472 | 12 | 1307120119 | 1307120170 | 2215 | A | | | | | | | | | |
| 20 | 19 | 43 | 15 | 1307120122 | 1307120140 | 2245 | C | | | | | | | | | |
| 21 | 20 | 353 | 13 | 1307120144 | 1307120199 | 2186 | C | | | | | | | | | |
| 22 | 21 | 218 | 15 | 1307120152 | 1307120272 | 2113 | E | | | | | | | | | |
| 23 | 22 | 69 | 16 | 1307120163 | 1307120188 | 2197 | D | | | | | | | | | |
| 24 | 23 | 562 | 16 | 1307120190 | 1307120301 | 2084 | D | | | | | | | | | |
| 25 | 24 | 121 | 19 | 1307120253 | 1307120294 | 2091 | E | | | | | | | | | |
| 26 | 25 | 297 | 15 | 1307120277 | 1307120342 | 2043 | B | | | | | | | | | |
| 27 | 26 | 495 | 13 | 1307120281 | 1307120353 | 2032 | E | | | | | | | | | |
| 28 | 27 | 94 | 14 | 1307120288 | 1307120343 | 2042 | E | | | | | | | | | |
| 29 | 28 | 22 | 18 | 1307120310 | 1307120365 | 2020 | C | | | | | | | | | |
| 30 | 29 | 64 | 19 | 1307120310 | 1307120385 | 2000 | B | | | | | | | | | |
| 31 | 30 | 502 | 16 | 1307120323 | 1307120336 | 2049 | B | | | | | | | | | |
| 32 | 31 | 44 | 16 | 1307120339 | 1307120352 | 2033 | A | | | | | | | | | |
| 33 | 32 | 315 | 14 | 1307120348 | 1307120362 | 2023 | B | | | | | | | | | |
| 34 | 33 | 385 | 15 | 1307120352 | 1307120553 | 1832 | E | | | | | | | | | |
| 35 | 34 | 550 | 13 | 1307120356 | 1307120444 | 1941 | B | | | | | | | | | |
| 36 | 35 | 92 | 14 | 1307120368 | 1307120397 | 1988 | B | | | | | | | | | |
| 37 | 36 | 395 | 16 | 1307120377 | 1307120426 | 1959 | D | | | | | | | | | |
| 38 | 37 | 267 | 17 | 1307120382 | 1307120515 | 1870 | E | | | | | | | | | |
| 39 | 38 | 257 | 14 | 1307120401 | 1307120427 | 1958 | C | | | | | | | | | |
| 40 | 39 | 312 | 19 | 1307120407 | 1307120548 | 1837 | D | | | | | | | | | |
| 41 | 40 | 321 | 18 | 1307120431 | 1307120449 | 1936 | A | | | | | | | | | |
| 42 | 41 | 220 | 16 | 1307120437 | 1307120510 | 1875 | A | | | | | | | | | |

Datos No Estructurados

Todo lo demás

- Imágenes
- Audio
- Vídeo
- Texto

Tipos de Análisis de Datos

- Descriptivo
- Exploratorio
- Inferencial
- Predictivo
- Causal

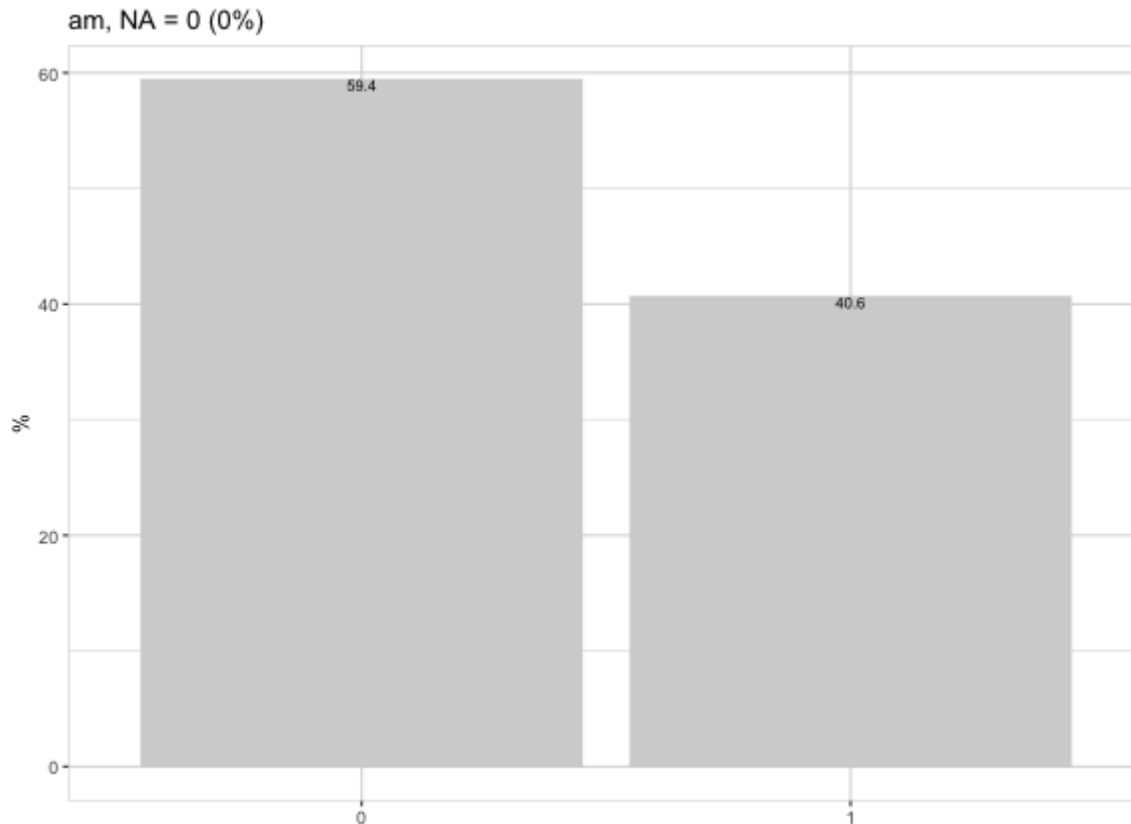
Análisis Descriptivo

Objetivo: resumir la información presente en un conjunto de datos

- Primer tipo de análisis de datos a realizar
- En general, las descripciones no se pueden generalizar sin la ayuda de modelos estadísticos

Análisis Descriptivo

La base de datos **mtcars** de R contiene información extraída de la *1974 Motor Trend US magazine* acerca de 10 aspectos de diseño de rendimiento de 32 vehículos. La variable **am** se refiere a la transmisión (0 = automática, 1 = manual)



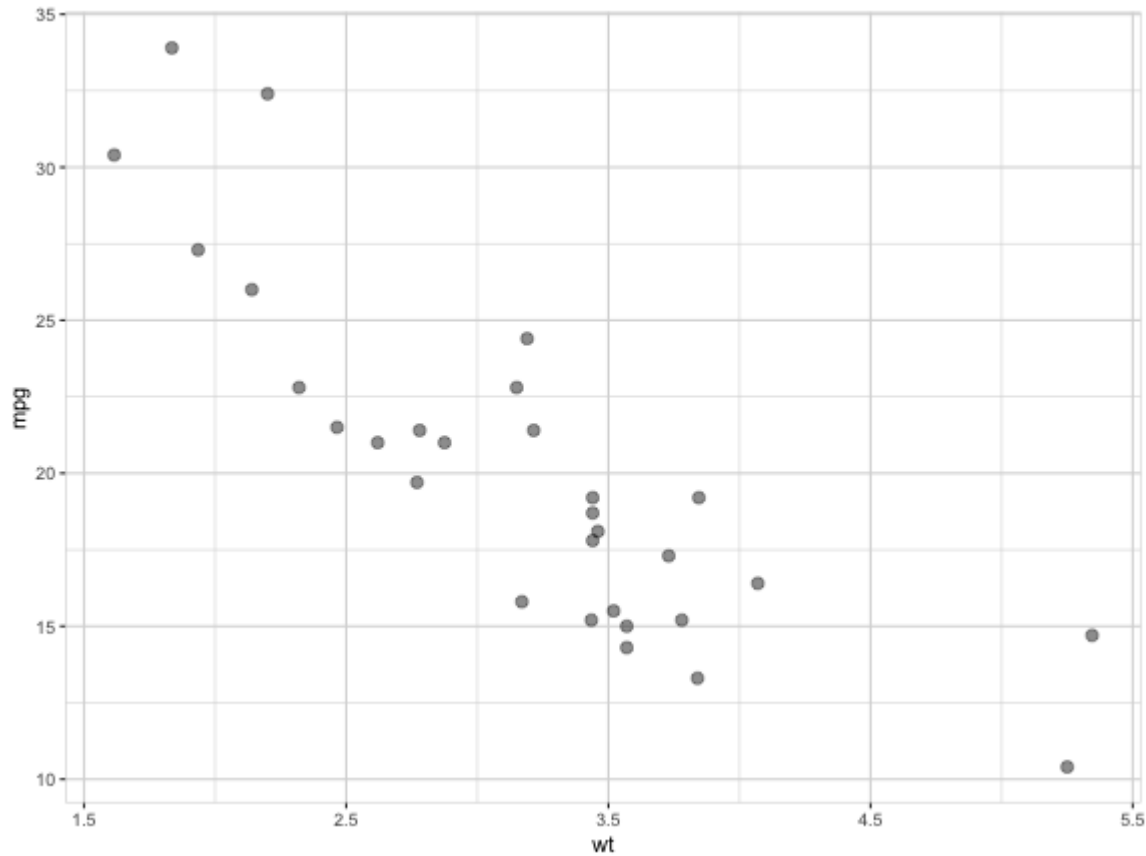
Análisis Exploratorio

Objetivo: descubrir relaciones entre características de los datos

- Motiva preguntas de investigación

Análisis Exploratorio

La variable **wt** se refiere al peso del vehículo y **mpg** a las millas por galón



Análisis Inferencial

Objetivo: usar una muestra pequeña de datos acerca de una población para extraer alguna información acerca de la misma.

- Aquí entra en juego la **estadística**

Análisis Inferencial

[< Previous Article](#) | [Next Article >](#)

Epidemiology:

January 2013 - Volume 24 - Issue 1 - p 23–31

doi: 10.1097/EDE.0b013e3182770237

Air Pollution

Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 U.S. Counties for the Period from 2000 to 2007

Correia, Andrew W.^a; Pope, C. Arden III^b; Dockery, Douglas W.^c; Wang, Yun^a; Ezzati, Majid^d; Dominici, Francesca^a

FREE SDC

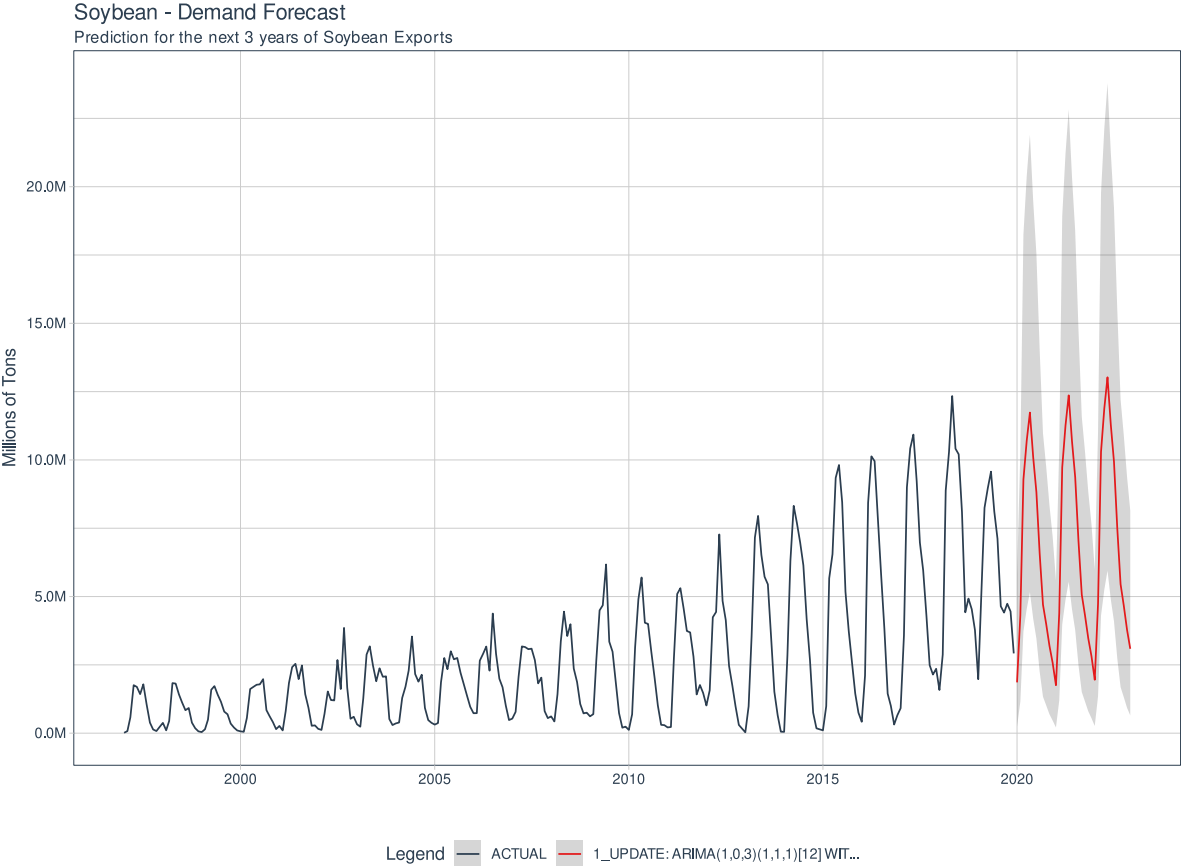
[Article Outline](#)

Análisis Predictivo

Objetivo: utilizar datos sobre un conjunto de objetos para predecir el valor de una variable en un objeto nunca antes visto

- X predice Y no implica que X sea causa de Y

Análisis Predictivo



[linkedin.com/in/lucianobatistads/](https://www.linkedin.com/in/lucianobatistads/)

Fuente

Análisis Causal

Objetivo: encontrar qué le sucede a una variable cuando se modifica el valor de otra

- Las relaciones causales usualmente identifican efectos medios, no efectos individuales

Análisis Causal



The NEW ENGLAND
JOURNAL of MEDICINE

SUBSCRIBE
OR RENEW

→



Q

≡



Concise summaries of
clinical study results

INTERACTIVE MEDICAL CASE

More Than a Little Unsteady



IMAGE CHALLENGE

What is the diagnosis?



ORIGINAL ARTICLE

Once-Weekly Dulaglutide for the
Treatment of Youths with Type 2
Diabetes



PERSPECTIVE

Bringing Sickle-Cell Treatments
to Children in Sub-Saharan
Africa

Editor's Note: This article was published on December 10, 2020, at NEJM.org.

ORIGINAL ARTICLE

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D., Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D., Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D., [et al.](#), for the C4591001 Clinical Trial Group*

≡

🔖

Article

Figures/Media

Metrics

13 References

5659 Citing Articles

Letters

December 31, 2020

N Engl J Med 2020; 383:2603-2615

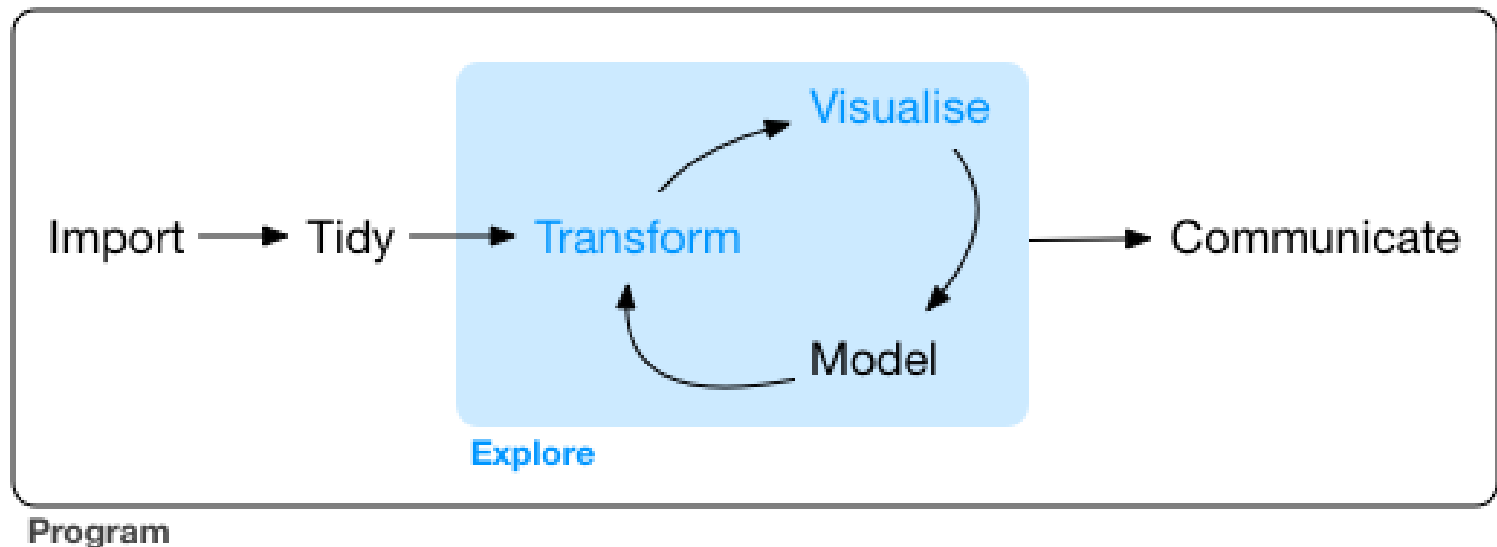
DOI: 10.1056/NEJMoa2034577

Chinese Translation [中文翻译](#)

Este curso

Adentrarnos en la ciencia de datos a través de R

- Datos Tabulares (observaciones x variables)
- Análisis Descriptivo, Exploratorio, Inferencial



Aprenderemos las herramientas fundamentales de R para las distintas fases de este esquema.

Las fases de un proyecto de análisis de datos

1. **Importación:** cargar datos en R procedentes de: base de datos, fichero, aplicación web (API), etc.
2. **Organización:** almacenar datos de manera consistente con análisis.
3. **Transformación:** filtrado, creación de variables derivadas, etc.
4. **Visualización:** generación de preguntas, descubrimiento de tendencias.
5. **Análisis Exploratorio:** Visualización + transformación aplicados de forma sistemática.
6. **Modelización:** confirmar hipótesis, responder preguntas.
7. **Comunicación:** comunicar resultados

Las fases de un proyecto de análisis de datos

- 1,2,3 → Data Wrangling
- 4,5 → Data exploration
- 5 → Modelización
- 6 → Comunicación

Todo esto usando el lenguaje de programación R.

Programa

- Tema 1 - Programación en R
- Tema 2 - Análisis Exploratorio de los Datos
- Tema 3 - Data Wrangling
- Tema 4 - Modelización
- Tema 5 - Comunicación de resultados

R y RStudio

¿Qué es R/RStudio?

- R es un lenguaje de programación especializado en estadística
- RStudio es una interfaz para programar en R



[Instalación de R y RStudio](#)

Presentación

- Análisis de Datos ECO-A
- Profesor: Roi Naveiro
- Email: roi.naveiro@cunef.edu
- Web del curso: <https://roinaveiro.github.io/CU-analisis-datos/>

Horario

- *Horario clases:*
 - Lunes 15.30 - 16.30
 - Jueves 16.30 - 18.00
 - Viernes 16.30 - 18.00
- *Asistencia:* Obligatoria, al menos 80%

Evaluación

- **Convocatoria Ordinaria**
 - Evaluación continua: Examen 1: 20%
 - Evaluación continua: Examen 2: 20%
 - Examen Ordinario Final: 60% (toda la materia)
- **Convocatoria Extraordinaria:** Examen Extraordinario Final (60%) + Evaluación continua

Recursos Interesantes

- [Análisis de datos acerca de la evolución de las tendencias musicales](#)
- [A year as told by FitBit](#)
- [Charla TED](#)
- [RMarkdown](#)

Bibliografía

- [Hands-On Programming with R](#), Grolemund (2014)
- [R for Data Science](#), Wickham and Grolemund (2016)
- [Data Visualization, A practical introduction](#), Healy (2018)
- [Data Science Specialization](#), Johns Hopkins University, Coursera