Lorenzo Escot
Coordinador del Observatorio de Igualdad de la UCM.
Dpto Economía Aplicada Pública y Política
Facultad de Estudios Estadísticos
Universidad Complutense de Madrid.

Julio E. Sandubete
Facultad de Estudios Estadísticos (Universidad Complutense de Madrid)
Computing and Artificial Intelligence Lab. (Universidad Camilo José Cela)

Morelia, Mich., 6 de octubre de 2022

Análisis de datos aplicado a la Economía

FACULTAD DE ESTUDIOS ESTADÍSTICOS

Grupo de Investigación Complutense num 940051
Análisis de Datos en Estudios Sociales, de Género y Políticas de Igualdad
www.ucm.es/aedipi

Coordinador del ***Observatorio de la Unidad de Igualdad de la UCM***
www.ucm.es/unidaddeigualdad/

unidad de igualdad | UNIVERSIDAD COMPLUTENSE DE MADRID

La Economía es una ciencia empírica

En las **ciencias empíricas**, necesitamos acudir a los "**datos**", a la observación de la realidad para contrastar las hipótesis, para encontrar patrones, para analizar la evolución de los fenómenos observados

De la Econometría tradicional a la Ciencia de Datos

Conjunto de instrumentos, procedimientos, algoritmos, técnicas, que permiten analizar diferentes tipos de datos (no sólo cuantitativos y categóricos) y de diverso tamaño (datos de muestreo pero también datos masivos o Big Data)

¿qué es lo que hace la Ciencia de Datos?, ¿a qué se dedica la Ciencia de Datos? ¿cuál es el flujo de trabajo de la Ciencia de los Datos?

Fuente*: What is Data Science? Understanding Data Science* datacamp

Registros y Estadísticas
Públicos y Open Data

Digitalización de nuestra vida diaria

Muestreos, sondeos,
experimentos

Coockies, Web scraping y APIS de Internet

Fuentes Primarias:
registros internos

Almacenamiento:
Bases de Datos referenciales o SQL
Bases de Datos no tabulares o no SQL

- Almacenamiento y proceso en Local o en la Nube
- Computación en paralelo

Cuantitativas

Categóricas

Textos

Datos Geospaciales

Imágenes

Grafos-Redes Sociales

Sonidos

Recolección y almacenamiento de datos → Depuración de los datos → Exploracion y visualización → Modelización Inferencial y Predictiva

Ingeniería del dato

Analista de datos y científico de datos

¿ CIENCIA de DATOS?

Es algo más que la estadística como Rama de las matemáticas

Ingeniería del Dato (Informática y Computación + Estadística + Áreas de aplicación (Ingenierías, Bioestadística, Marketing, etc)

¿Entidad propia?

Ordenar y comprender los datos:
--¿Los datos miden realmente lo que queremos medir?
-- ¿Cómo se han obtenido esos datos?
-- ¿los datos son representativos de la población que se quiere modelizar o sobre la que se quiere aplicar el modelo?

Depuración
-¿Qué estructura de datos tengo?
-¿Valores perdidos? (imputación de valores perdidos)
-Variables con datos que en realidad no varían
--Valores atípicos, el rango de datos es apropiado

Ordenar
Depurar,
Resumir y
Visualizar
los datos

Una vez elegida la fuente de datos ¿qué hacemos con esos datos?

Ejemplo: Tenemos las respuestas al cuestionario que realiza el INE a 25 mil personas en la encuesta de la EPA (Microdatos de la EPA), ¿Qué hacemos con todos esos datos?

Recolección y almacenamiento de datos → Depuración de los datos → Exploracion y visualización → Modelización Inferencial y Predictiva

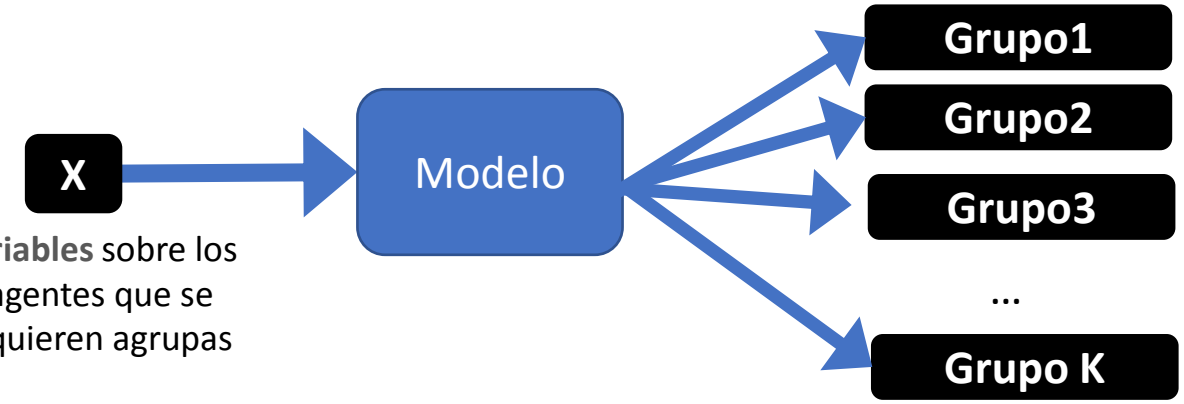
Modelado APLICADO

Modelos Supervisados y Modelos No supervisados

Modelos de agrupación y de reducción de dimensiones

Modelos de Regresión o de Clasificación

El tipo de técnicas a utilizar para el modelado de datos dependerá de los objetivos y de los tipo de datos que tengamos

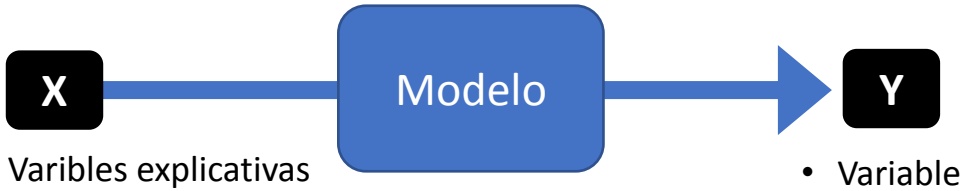


Variables sobre los agentes que se quieren agrupas

Modelos de agrupación y de reducción de dimensiones

Modelos Supervisados y Modelos No supervisados

Modelos de Regresión o de Clasificación



- Varibles explicativas
- Variables independientes
- Tratamieto
- Variable Exógena

- Variable Explicada
- Resultado / Outcome
- Variable Endógena

Leo Breiman (1928- 2005): ***Statistical Modeling: The Two Cultures Statistical Science, Vol. 16, No. 3, 199-231, 2001***

Modelos de agrupación y de reducción de dimensiones

Modelos Supervisados y Modelos No supervisados

Modelos de Regresión o de Clasificación

Inferencia

Objetivo de cuantificar la relación entre variables

Y = F (X, ruido, parámetros)

Supuestos sobre la función de distribución de los estadísticos

Validación con test de bondad de ajuste y análisis de los residuos

Nuevas estructuras de datos, datos masivos Big Data

Predictivos

Y = f (X,)

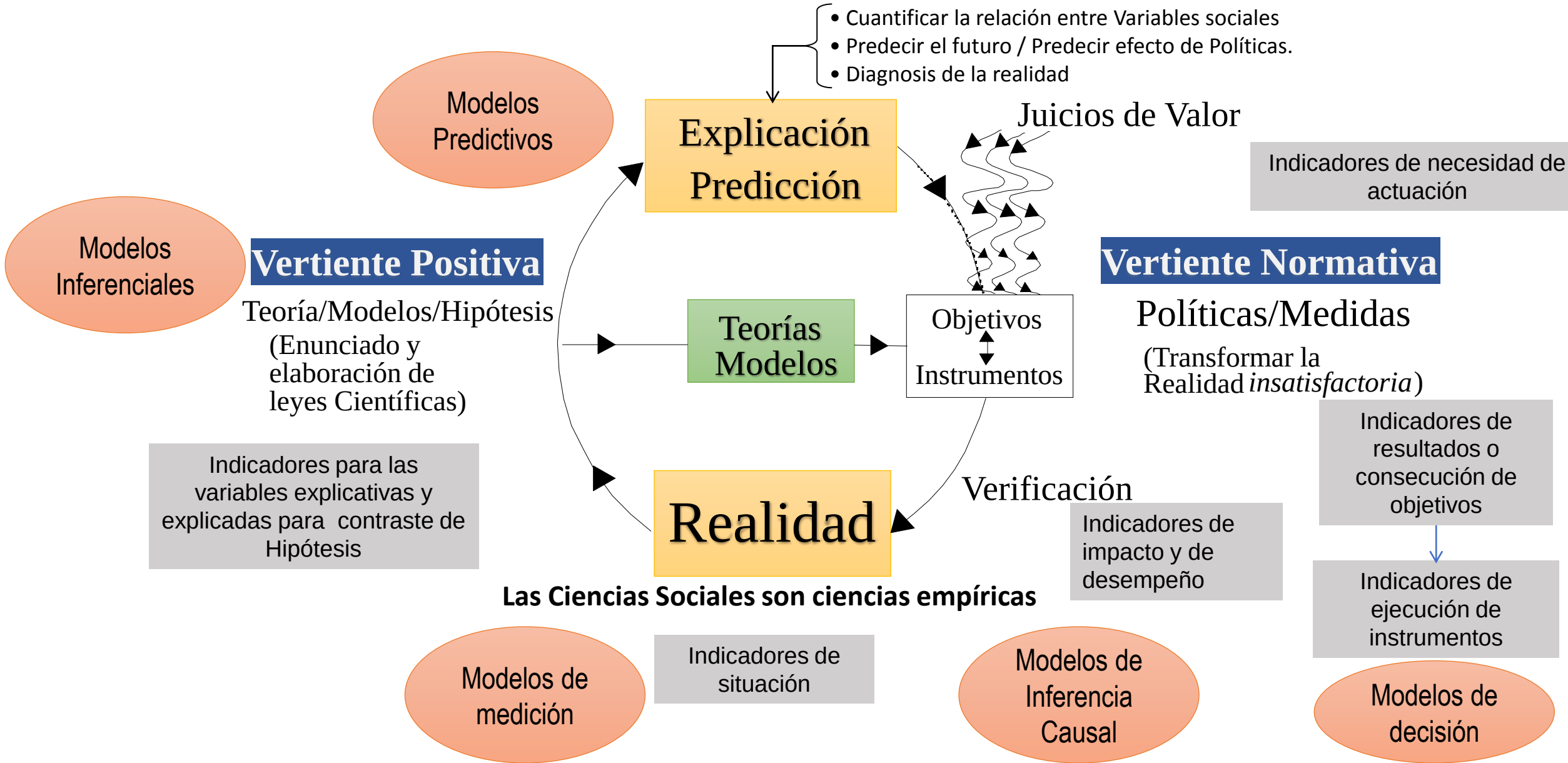
Algoritmos nolineales

f se valida utilizando su exactitud predictiva: Muestra de test y Validación

Modelización algorítmica

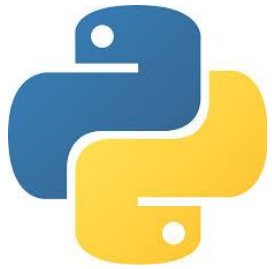
Machine Learning
Deep Leraning
El Internet de las Cosas
Procesamiento en tiempo real

¿Qué papel Juega la Ciencia de Datos en la Economía como Ciencia?



CAMBIO DE PARADIGMA EN LOS RECURSOS Y ENTORNOS DE TRABAJO

EL ECONOMISTA DEL S. XXI TIENE QUE PERDER EL MIEDO A PROGRAMAR



CAMBIO DE PARADIGMA EN LOS RECURSOS Y ENTORNOS DE TRABAJO

TensorFlow

Keras

hadoop

APACHE Spark

python

jupyter

SciPy

SPYDER

NumPy

DASK

bokeh

scikit learn

PC

Numba

pandas

Panel

matplotlib

PYTORCH

CONDA

HoloViews

MySQL

Microsoft SQL Server

PostgreSQL

mongoDB.

+ableau

Qlik

LEAD WITH DATA

Power BI

https://cloud.r-project.org/web/views/

CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Task Views
Other

Documentation
Manuals
FAQs
Contributed

CRAN Task Views

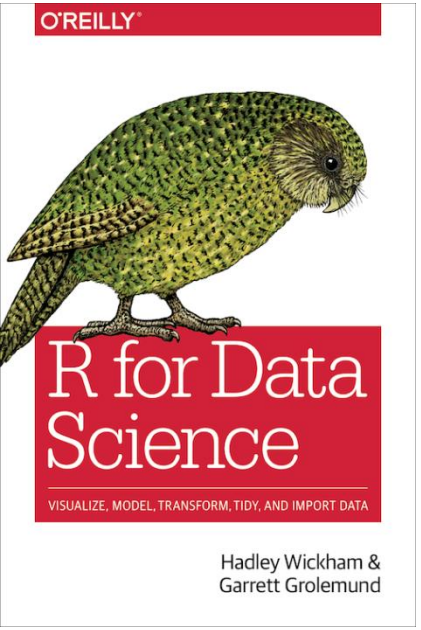
CRAN task views aim to provide some guidance which packages on CRAN are relevant for tasks related to a certain topic. They give a brief overview of the included packages and can be automatically installed using the ctv package. The views are intended to have a sharp focus so that it is sufficiently clear which packages should be included (or excluded) - and they are *not* meant to endorse the "best" packages for a given task.

To automatically install the views, the ctv package needs to be installed, e.g., via
`install.packages("ctv")`
and then the views can be installed via `install.views` or `update.views` (where the latter only installs those packages are not installed and up-to-date), e.g.,
`ctv::install.views("Econometrics")`
`ctv::update.views("Econometrics")`

The resources provided by the CRAN Task View Initiative provide further information on how to contribute to existing task views and how to propose new task views.

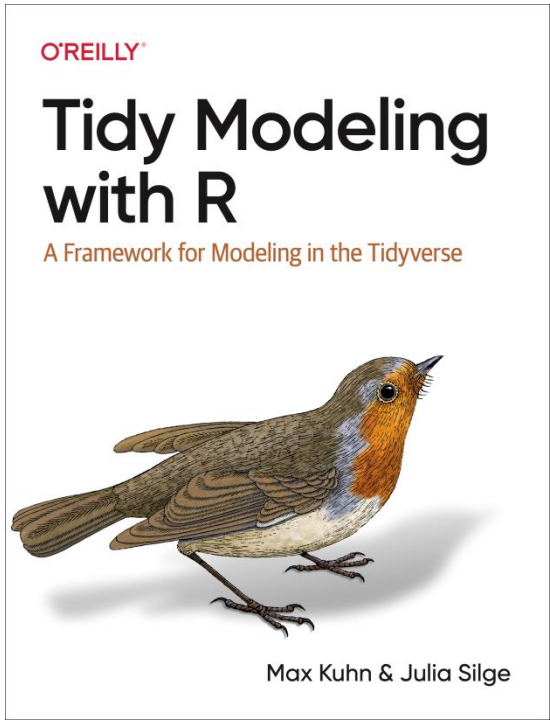
Topics

Agriculture	Agricultural Science
Bayesian	Bayesian Inference
CausalInference	Causal Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
Databases	Databases with R
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
Epidemiology	Epidemiology



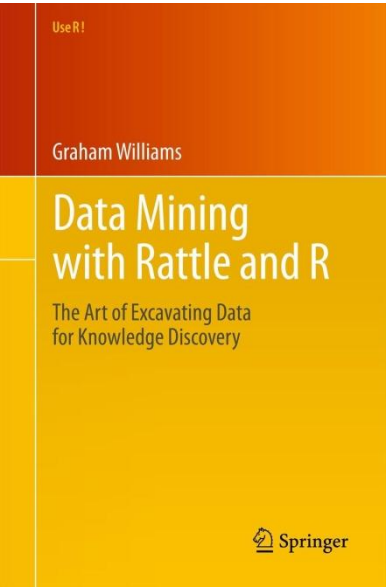
Hadley Wickham and Garrett Grolemund (2016): **R for data science: Import, Tidy, Transform, Visualize, and Model Data.** Editorial O'Reilly Media, Inc. (ISBN-13: 978-1491910399

https://r4ds.had.co.nz/



Max Kuhn and Julia Silge (2022) **Tidy Modeling with R: A Framework for Modeling in the Tidyverse.** Editorial O'Reilly Media, Inc. ISBN: 9781492096481

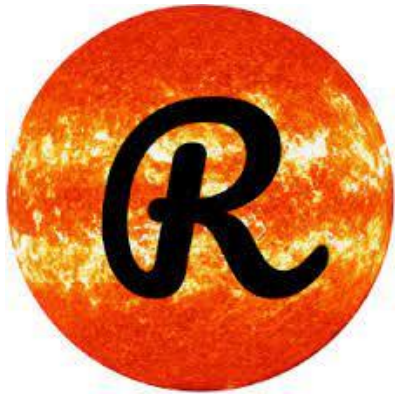
https://www.tmwr.org/



William, G. (2011). **Data Mining with Rattle and R, The art of Excavating Data for Knwledge Discovery**. Springer.

Vincent Nijs (2019) **Radiant – Business analytics using R and Shiny**

https://radiant-rstats.github.io/docs/



“Ponedle ilusión a lo que hagáis, ponedle pasión, sin pasión estaréis perdidos, sin pasión se acabará el amor por lo que hacéis, y sin amor, el día a día se volverá triste y tedioso, sin amor, no habrá esperanza”

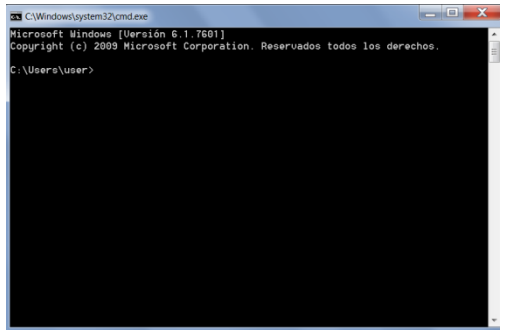
GitHub

github: jsandube/UMSNH2022

https://github.com/jsandube/UMSNH2022

https://www.r-project.org/

Consola o Interface gráfica



RStudio®

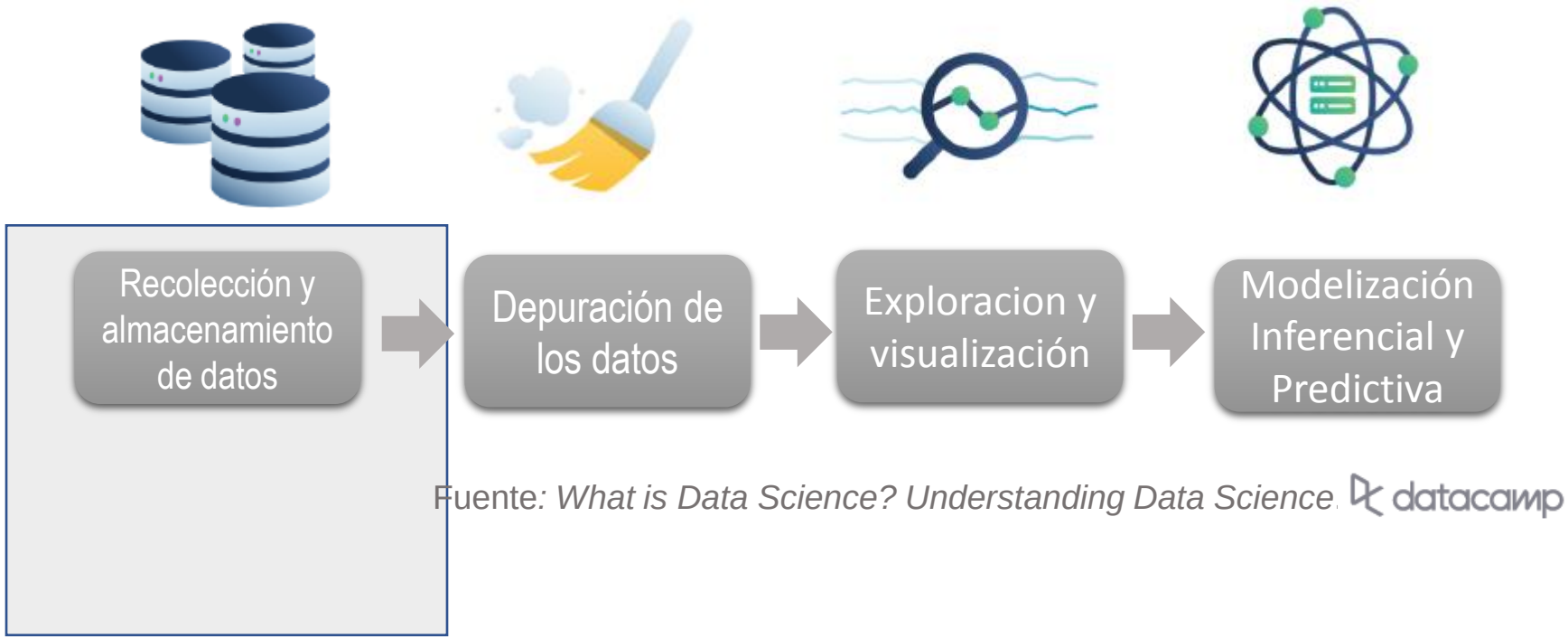
https://www.rstudio.com/

Interface gráfica más agradable de usar (incorpora ayudas a la programación)

RStudio Cloud

https://rstudio.cloud/

VAMOS AL R



Fuente*: What is Data Science? Understanding Data Science*. datacamp

FUENTES DE DATOS PARA SU USO EN NUESTRAS APLICACIONES

Tipología de Fuentes de Datos Empíricos

Elaboración Ajena

- **Fuentes Secundarias** (tablas, gráficos y otros resultados estadísticos publicados por otras investigaciones)

Elaboración Propia

- **Fuentes Primarias**
  (elevado coste)

TIPOLOGIA DE FUENTES
para el análisis confirmatorio

- **Encuestas** (Muestreo)
- **Registros** (Contabilidad)
- **Experimentos**/QuasiExper.
- **Entrevistas** – Cualitativo -

- Entrevistas para **Análisis Cualitativos vs cuantitativos**: ¿permiten hacer inferencia estadística?
- Encuestas vs Registros: **representatividad** (y margen de error de la encuesta)
- Datos Observacionales vs Experimentos: asociación vs **causaliad**

Casos especial: registros de navegación por internet y nuestra huella digital

…….¿cómo elegir fuentes de datos?

Revisión del diseño: **¿sirve realmente la fuente de datos para construir los indicadores?**

Ojo, **cuidado con**

- **las definiciones de las variables/ítems/preguntas incluidos en la fuente de datos ¿representan realmente al fenómeno que quiero estudiar?**,
  Ejemplo: Definición de PARADO (EPA vs. SEPE)

- las poblaciones objeto de análisis ¿es realmente la población que se desea estudiar?, ¿Existen los **sesgos de selección muestral** (endogeneidad de experimentos)?

  Ejemplo: Según las estadísticas aumenta el número de denuncias por violencia de género ¿están aumentando los casos de violencia de género en España? (sesgo, quizás sólo denuncia quien crea que le va a ser útil la denuncia: se estaría sesgando a la baja o infravalorando la importancia de la violencia de género)

  Ejemplo: estudios morosidad nuevos clientes

- Características de las variables utilizadas y las **"trampas" estadísticas**.
  - **ASOCIACION NO ES CAUSALIDAD** número de crímenes y número de agentes de policía
  - **Asociaciones ESPUREAS**: cigüeñas o pastafarismo

¿Cómo elegir fuentes secundarias de elaboración ajena?

- Disponibilidad
  - ¿Están disponibles al público o son accesible a nosotros?
  - ¿Se distribuyen gratuitamente, o tiene algún coste?

- Calidad de los datos
  - Instituciones Públicas (INE, EUROSTAT, Ministerios, Universidades, ONU, UNICEF, Banco Mundial, FMI, OCDE, Organismos internacionales de Turismo, de Trabajo, etc.)
  - Instituciones Privadas (Empresas de investigación o de estudios de mercado, SABI, Bancos y Cajas de Ahorro, etc) **¿son de prestigio?**
  - Investigaciones científicas (tesis, proyectos de investigación subvencionados, o publicadas en revistas científicas de prestigio) **¿son utilizados por otros investigadores?** (*Econlit, Social Science Citation Index*)

- Ejemplos:
  - INE (www.ine.es) Estadística de violencia doméstica y violencia de género
  - Naciones Unidas (www.eclac.cl/mujer)
  - Eurostat (www.eurostat.eu)
  - CGPJ (http://www.poderjudicial.es/cgpj/es/Temas/Estadistica-Judicial/)

Descripción de las fuentes de datos

Antes de comenzar con el análisis de los datos propiamente dicho para contrastar ninguna hipótesis :

- **Describir** muy bien la fuente de datos y la metodología para su elaboración (especialmente cuando no proceden de fuentes estándar o no han sido utilizados previamente por otros investigadores)
    - Identificar la fuente (cualquier otro investigador debería ser capaz de repetir nuestro análisis econométrico)
    - Proporcionar datos técnicos sobre la elaboración de la encuesta (técnica de muestreo, representatividad, y error de muestreo) especialmente en caso de que sea de elaboración propia (incluir cuestionario, datos, etc)
        - Universo (No es lo mismo Encuesta que Registro)
        - Error, nivel de confianza (para proporciones o para medias)
        - Muestreo (aleatorio, estratificado, telefónicas, panel, ¿existen sesgos de selección?)
    - Si se trata de un registro decir también de donde proceden los datos y a qué población representa

Hay que conocer perfectamente ¿qué se está midiendo y cómo se mide?

- Repasar la definición de cada una de las variables (escala de medida, significado, etc.),

  Hay que conocer perfectamente la base de datos que se está utilizando

  Definición de los **ítems, de las preguntas** en los que se concretan los fenómenos que se van a estudiar (ejemplo Violencia de Género medido por el número de denuncias, o medido por el número de sentencias, o medido por el número de muertes)

  ¿qué códigos utiliza?, tiene factores de elevación o de ponderación, ¿se detecta algún dato anómalo?).

- **Revisar el Cuestionario**

… y tras verificar la fuente de datos, se puede pasar a la siguiente fase: ORDENAR Y DEPURAR LOS DATOS

¿Cómo cargo los datos en R?

¿Dónde están los datos?

- OpenData del Banco Mundial y Eurostad, INEGI, Harvard Dataverse
    - Descarga de archivos en excel (mejor siempre en csv)
    - Uso de APIS
    - Webscraping

https://data.worldbank.org/

Fuente*: What is Data Science? Understanding Data Science*. datacamp

DEPURACION DE LOS DATOS

DEPURACION DE LOS DATOS

- **Resumen inicial**: tipo de datos, Max, min, medias, Rangos
- Escalas de Medida: Homogeneizar la forma en que se midem. Ejemplo items en escala de likert ¿van todos ene el mismo sentido
- **Valores Perdidos**:
  - Individuos con demasiados valores perdidos
  - Variables con demasiados valores perdidos
  - **¿Qué se hace con los Valores Perdidos?**:
    - Se eliminan del análisis
    - Se categorizan como una categoría más
    - Se IMPUTAN?

- **Valores Extremos o valores Atípicos**
  - **Qué se hace con los atípicos: ¿son influyentes?**

- **Análisis de la Normalidad, Linealidad o Multicolinealidad exacta, Homocedasticidad**
  - **¿Es necesaria alguna transformación?**
    - Multicolinealidad exacta: Misma información
    - Transformación logarítmica para variables tipo precio

EJEMPLOS DE DEPURACION:

- **Ejemplo de Homogenización e imputación: indicadores mundiales de género**

International based Gender Gap Indexes

World Economic Forum

Global Gender Gap Index

Dimensions	*Indicators*
Economic	Labour-force participation rate
Participation and	Wage equality for similar work
Opportuniy	Estimated earned income
	Legislators, senior officials and managers
	Professional and technical workers
Educational	Literacy rate
Attainment	Enrolment in primary education
	Enrolment in secondary education
	Enrolment in tertiary education
Health and	Sex ratio at birth
Survival	Healthy life expectancy
Political	Women in parliament
Empowerment	Women in ministerial positions
	Years with female/male head of state

4 indicators
146 countries
Year 2022 - World Economic Forum. Global Gender Gap Report
http://reports.weforum.org/globalgender-gap-report-2022

Equal Measures 2030

SDG Gender Index

Dimensions (asociated to gender gap in SDGs)	
SDG1	**Poverty**
SDG2	**Nutrition**
SDG3	**Health**
SDG4	**Education**
SDG5	**Gender Equality**
SDG6	**Water**
SDG7	**Clean Energy**
SDG8	**Work**
SDG9	**Innovation**
SDG10	**Inequalities**
SDG11	**Sustainable cities**
SDG13	**Climates**
SDG16	**Justicie**
SDG17	**Partnerships**

56 key indicators about 14 of the 17 SDG
144 countries coverage
2022 - Equal Measures 2030 (EM2030) SDG Gender Index
https://www.equalmeasures2030.org/who-we-are/

UNDP (United Nations Development Programme)

Gender Inequality Index (GII)

Dimensions	*Indicators*
Reproductive	Maternal motarlity ratio
Health	Adolescent birth rate
Empowerment	Female and male population with at least secondary education
	Female and male shares of Parlamentary seats
Labour Market	Female and male labour force participation ratios

10 key indicators
195 countries coverage
2021 - Human Develpment Reports (UNDP) Gender Composite I.
https://hdr.undp.org/data-center/composite-indices

The OECD Development Centre

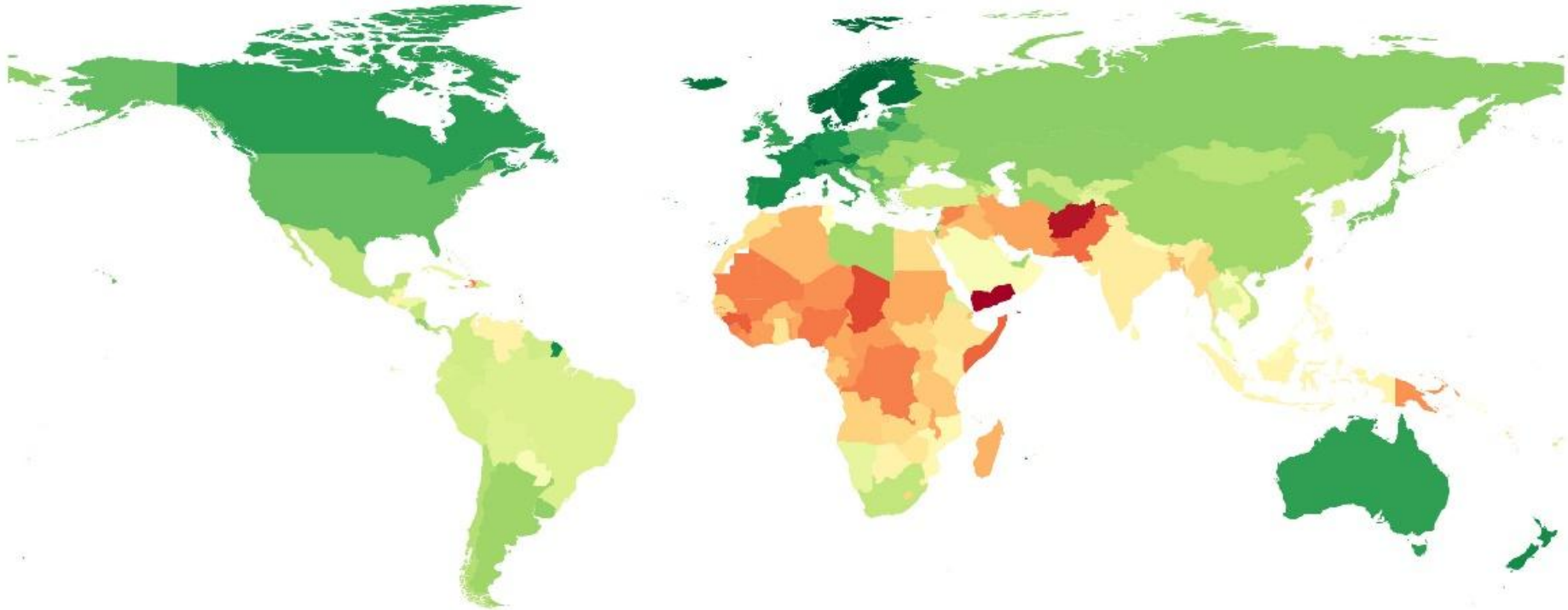
Social Institutions and Gender Index (SIGI)

Dimensions	*Indicators*
Discrimination in	Child marriage
the family	Household responsabilities
	Divorce
	Inheritance
Restricted physical	Violence against women
integrity	Female genital mutilation
	Missing women
	Reproductive autonomy
Restricted access	Secure acces to land assets
to productive and	Secure acces to non-land assets
financial resources	Secure access to formal financial services
	Workplace rights
Restricted civil	Citizenship rights
liberties	Political voice
	Freedon of movement
	Acces to justice

27 variables combined into 16 indicators and 4 dimensions
180 countries coverage
2019 - Global Report for the fourth edition of the SIGI.
https://www.genderindex.org/

Global Gender Gap Index (World Economic Forum)
SDG Gender Index (EquaEqual Measures 2030)
Social Institutions and Gender Index (The OECD Development Centre)
Gender Inequality Index (United Nations Development Programme)

Gender GAP Index



Gender GAP

40 50 60 70 80

Note: Mean Gender GAP Index from World Economic Forum-Equal Measures 2030-OECD-UNPD expresed as Female/Male Ratio, with a range from 0 (for very high discrimiantion) to 100 (for gender equality)

Fuente*: What is Data Science? Understanding Data Science*. datacamp

ANALISIS EXPLORATORIO Y VISUALIZACION DE LOS DATOS

Una vez elegida la fuente de datos ¿qué hacemos con esos datos?

Ordenarlos, visualizarlos y resumirlos

Ejemplo: Tenemos las respuestas al cuestionario que realiza el INE a 25 mil personas en la encuesta de la EPA (Microdatos de la EPA), ¿Qué hacemos con todos esos datos?

Los datos recogen información sobre diferentes fenómenos o agentes que estamos estudiando (individuos, países, empresas, experimentos, accidentes de tráfico…). Utilizamos el concepto de **variable** para representar a dicho fenómeno, esto es para representar a las diferentes características que lo conforman (una variable para cada característica)

Variables : representación de cada una de las características o propiedades del fenómeno que estamos estudiando, (cada variable representa uno de esos aspectos del fenómeno estudiado)
(diferencia entre casos, variable, constante y parámetro)

Partidos jugados
Partidos Ganados
Presupuesto total…

Número de Casos por país
Número Fallecidos
Ocupación camas UCI

Sexo del bebé
Meses de gestación
Peso al nacer
Perímetro craneal

Composición del hogar
Nivel de Estudios del cabeza de familia
Ingresos totales del hogar
Régimen de tenencia de la vivienda

Tipo de Variables

Variables CUANTITATIVAS: Se pueden cuantificar, tienen valor numérico (con una escala de medida).

Ejemplos: edad (años), altura (cm), peso (kilos), ingresos (euros), Incidencia del Covid-19 (casos acumulados en los últimos 14 días por cada 100.000 habitantes), número de hijos (hijos), número de partidos ganados (partidos)

- **Continuas** (valores reales)
- **Discretas** o de recuento (valores enteros)

- Algunas otras tipologías: series temporales, datos sección cruzada o datos de panel

Variables CUALITATIVAS: Atributos, características que indican diferentes categorías que **no se pueden medir o cuantificar numéricamente** (factores)

- **Escala nominal**: a cada atributo se le da un nombre (sexo, nacionalidad, color del pelo)
- **Escala ordinal**: los diferentes atributos guardan una relación de orden (nivel educativo, grado de satisfacción – escala de Likert –)
- De intervalo: Grupos de edad, grupo salarial, ranking Q1, Q2, Q3 … de revistas JCR

Ahora nos centramos en variables unidimensionales. Con variables bidimensionales puede comenzarse a analizar relación o asociación entre variables o entre características de la población (ejemplo, salario con nivel de estudios, peso con altura, nacionalidad con lengua materna)

¿**Cómo podemos resumir los datos** para extraer información relevante?, por ejemplo para hacer una evaluación o seguimiento de la política presupuestaria

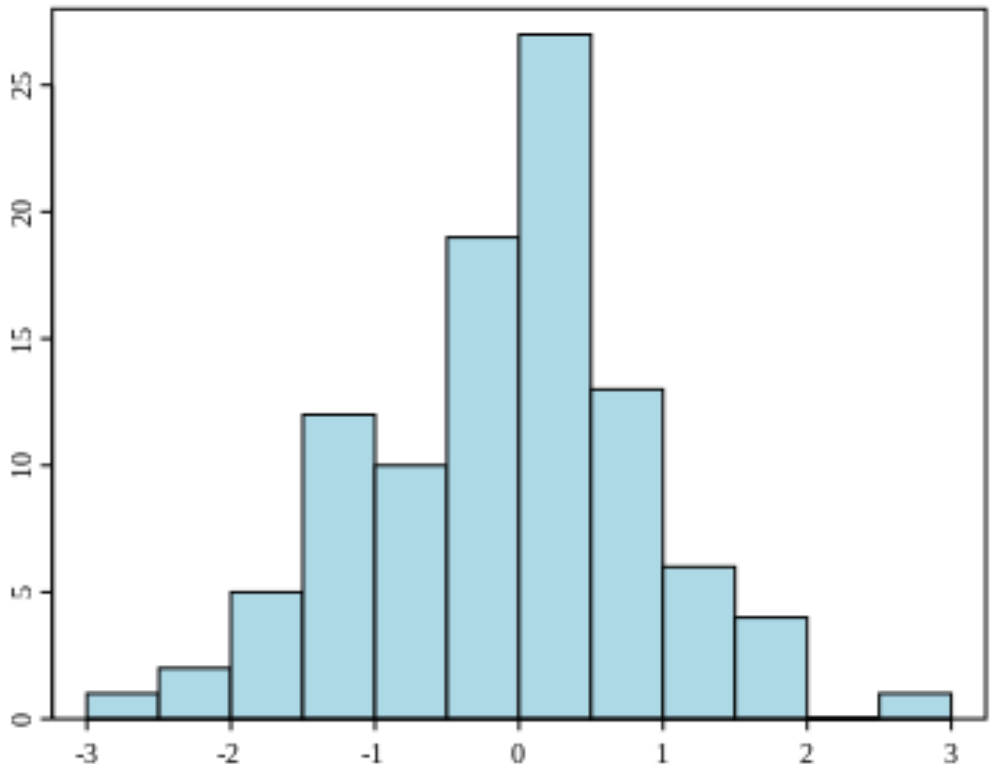
La estadística descriptiva ayuda al resumen y análisis de datos, a obtener indicadores de los diferentes fenómenos que estamos analizando

El tipo de resumen y de análisis estadístico que pueda realizarse dependerá del tipo de variables, diferente para las variables cuantitativas y para las categóricas

Recordamos que nuestra base de datos contiene información de 20 mil individuos que han contestado el cuestionario de la EPA, ¿cómo se resumen esos datos?

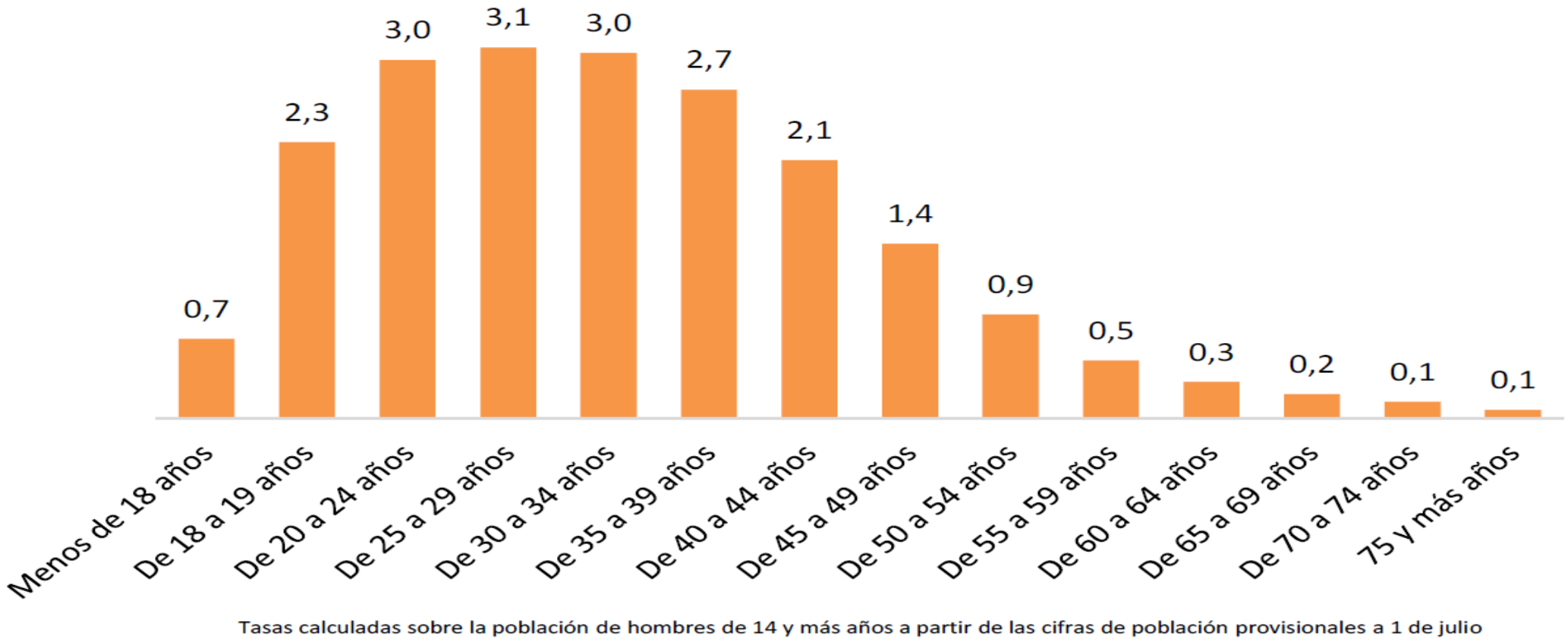
Distribución de frecuencias: **análisis de los posibles valores que puede tomar una variable e importancia de cada uno de ellos**

(Relacionado con fenómenos deterministas y aleatorios)



(Datos anómalos, más probables, asimetrías, etc)

Tasas de víctimas de violencia de género (con orden de protección o medidas cautelares) por edad (tasas por 1.000 mujeres de 14 y más años)

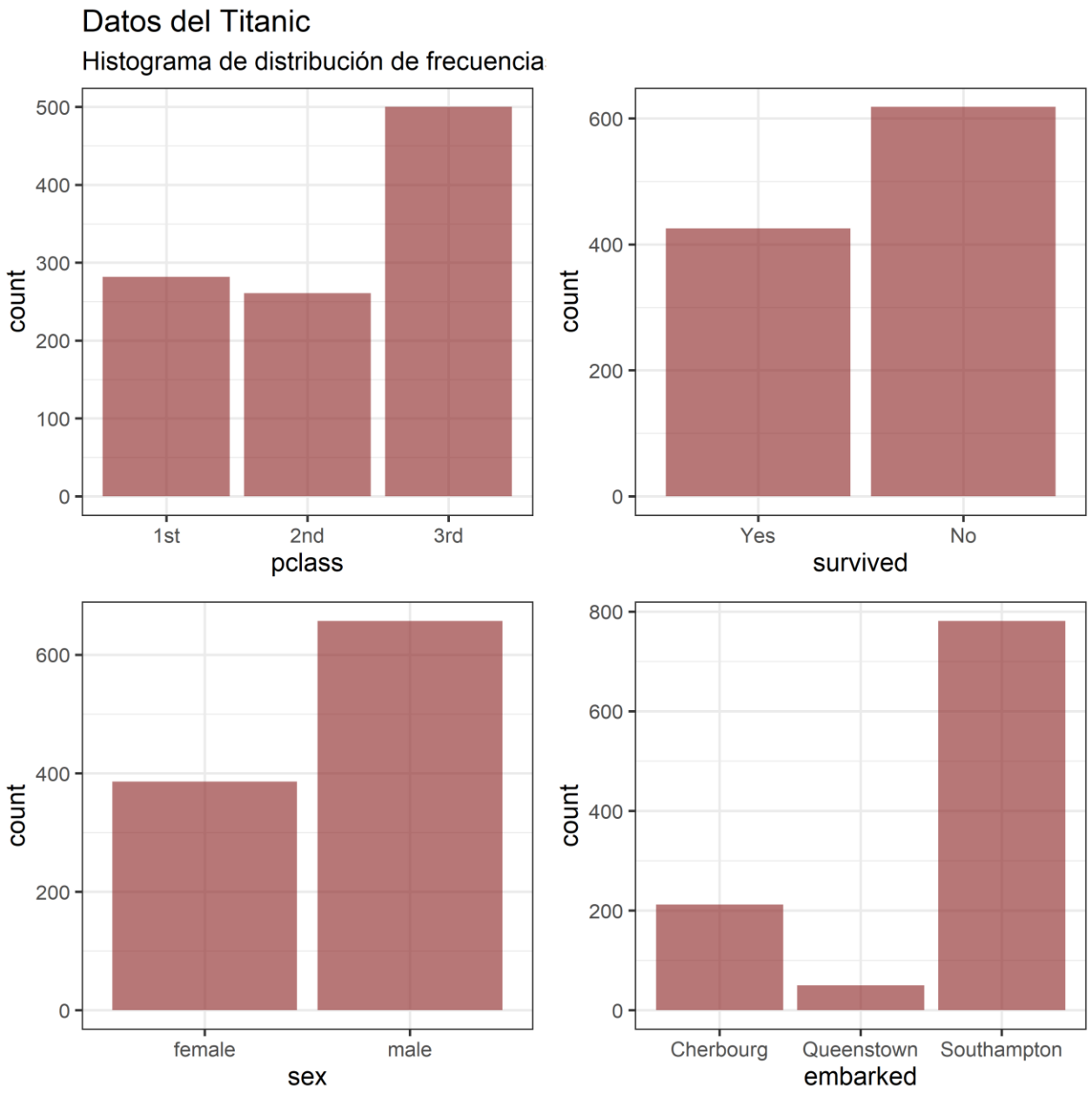


Tasas calculadas sobre la población de hombres de 14 y más años a partir de las cifras de población provisionales a 1 de julio

Estadística de Violencia Doméstica y Violencia de Género. INE, 2016

Caso práctico: Registros de Viajeros del Titanic

https://vnijs.shinyapps.io/radiant



Resumen del Histograma de Variables Continuas

Medidas de Posición Central: Media, mediana, moda

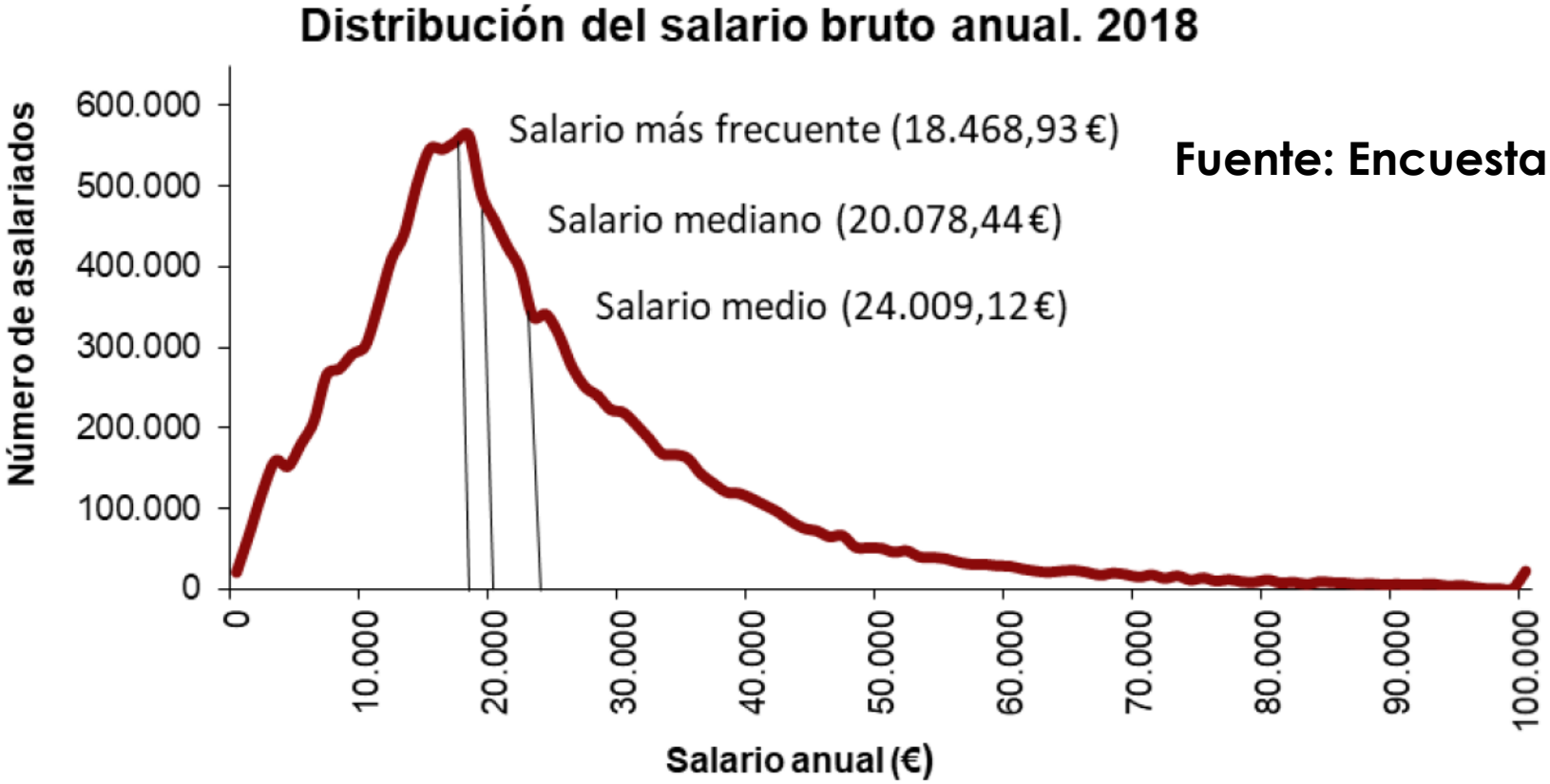
Medidas de posición No central:
Máximo, Mínimo, cuartiles

Medidas de dispersión: varianza (respecto a la Media, mediana, moda)
precisión de la media, mediana, moda

Medidas de forma:
asimetría y apuntamiento/curtosis

Medidas de concentración:
Gini, Índice de disimilitud (Ducan y Ducan)

Ojo, para variables Categóricas sólo MODA y medidas de concentración



Fuente: Encuesta de Estructura Salarial 2018 (INE)

Medida de posición central ¿media o mediana? ¿Segundo municipio más rico de España?

Indicadores de Pobreza

Absolutos: Pobreza cuando la renta media es inferior a un dólar

Relativos (a la renta de los demás ciudadanos de tu entorno)

- Pobreza cuando se gana menos de un 60% de la renta mediana
- Pobreza severa, los ingresos no llegan al 20% de la Renta mediana

- Clase media: entre el 75% y el 200% de la renta mediana
- ¿Ricos: más del 200%?

Cuando la distribución de frecuencias (Histograma), no es simétrica, quizás mejor la moda (pero se descarta porque no hay única forma de calcularla), así que mejor la mediana

Estadística descriptiva bivariante: **variables con dos dimensiones**

Variables bidimensionales: recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase). Podemos utilizar también una variable para cada una de las características del fenómeno que queremos estudiar. En este caso el fenómeno será bivariante, o multivariante

Cuando tenemos variables bidimensionales, o simplemente dos variables (una para cada dimensión), además del resumen de cada una de ellas por separado (descriptivo univariante), se realiza el análisis bivariante para buscar medidas de **ASOCIACION** entre variables o dimensiones

Distribución de frecuencias bivariadas: análisis de los posibles valores que puede tomar cada una de las dos variable y hacer un mapa de las posibles combinaciones de valores dos a dos. La distribución de frecuencias bivariadas recoge el recuento de casos en cada una de esas posibles combinaciones

Ejemplo: Análisis del sexo y Nivel de Estudios

Posibles valores que puede tomar esta variable bivariante (o combinaciones de las dos variables

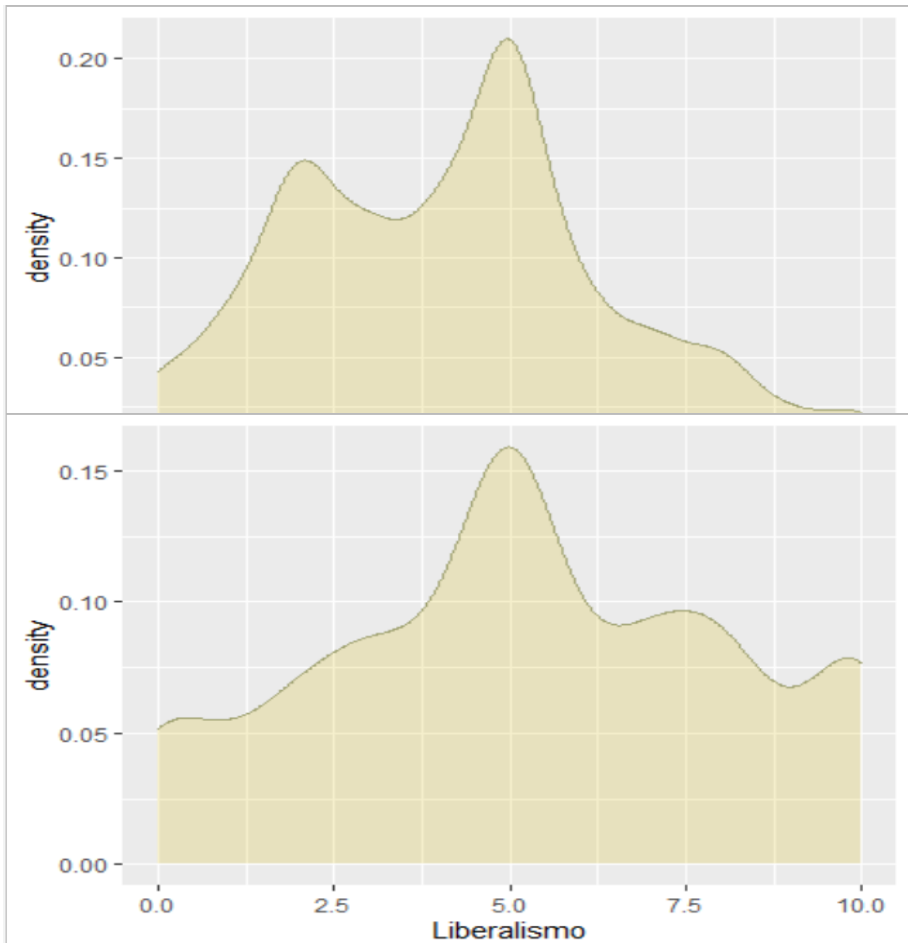
(sexo,estudios)

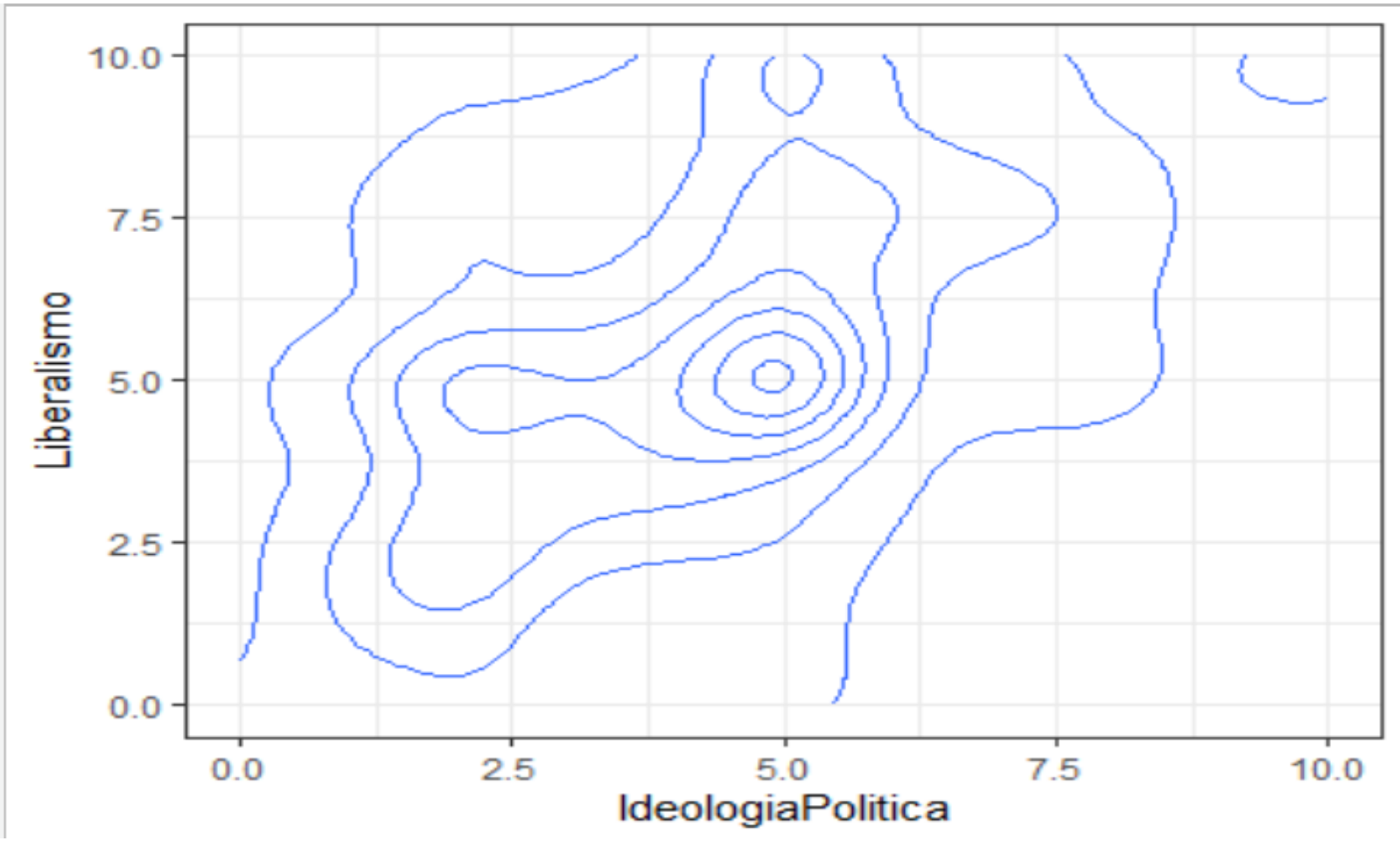
(Varón, Est. Primarios) (Varón, Est. Secundarios) (Varón, Est. Universitarios)

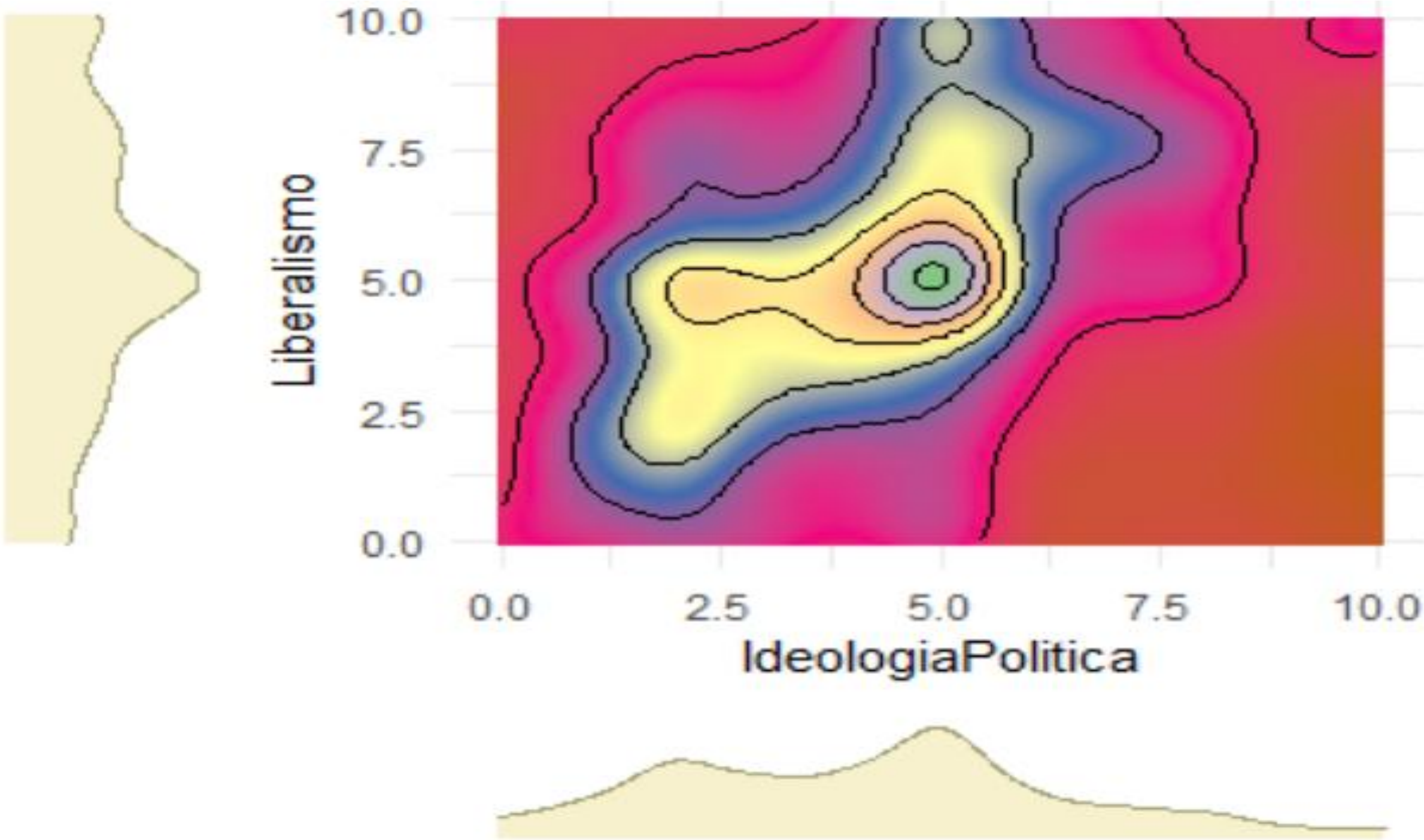
(Mujer, Est. Primarios) (Mujer, Est. Secundarios) (Mujer, Est. Universitarios)

Práctica: Relación entre nivel de estudios y sexo en España

Histogramas de distribución bivariantes para variables continuas






10.0
7.5
5.0
2.5
0.0
Liberalismo
0.0
2.5
5.0
7.5
10.0
IdeologiaPolitica

Medidas de Asociación entre Variables

Una vez analizada la distribución de frecuencias (bi-dimensionales) se analiza **la asociación entre esas dos variables** (el análisis multivariante extiende este análisis bivariante al multivariante). Normalmente se tiene una variable objetivo y se realiza un primer análisis exploratorio descriptivo de la asociación entre esa variable objetivo y cada una de las posibles variables explicativas dos a dos.

Medidas de asociación entre X e Y

X e Y: cuantitativas → Correlación lineal

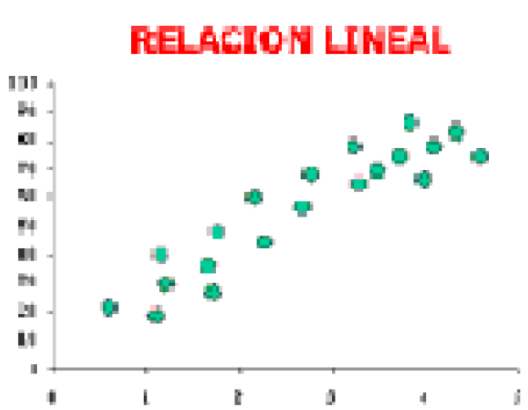
X e Y: cualitativas o categóricas → Independencia – C. Tetracórica

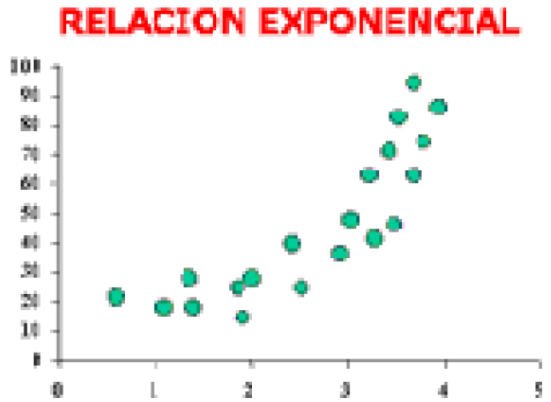
X categóricas e Y cuantitativa (o al revés) → Diferencia de Medias - Corr biserial

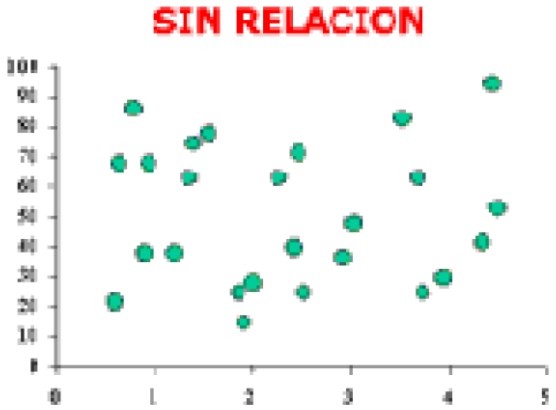
MEDIDAS DE ASOCIACION ENTRE DOS VARIABLES NUMERICAS

Coeficiente de Correlación lineal

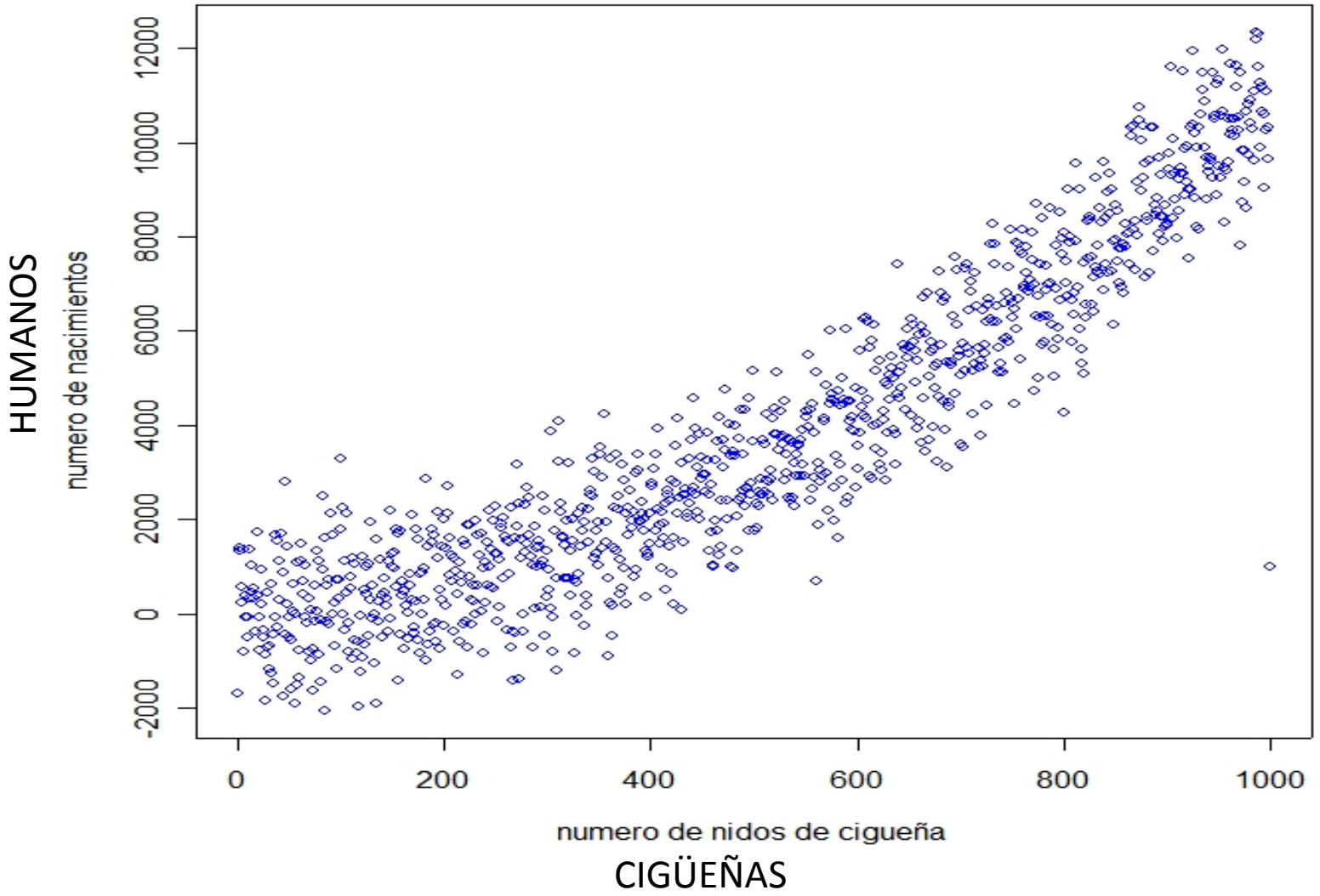
El coeficiente de correlación lineal mide el grado de intensidad en la dependencia lineal entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables la nube de puntos tiene forma longitudinal).



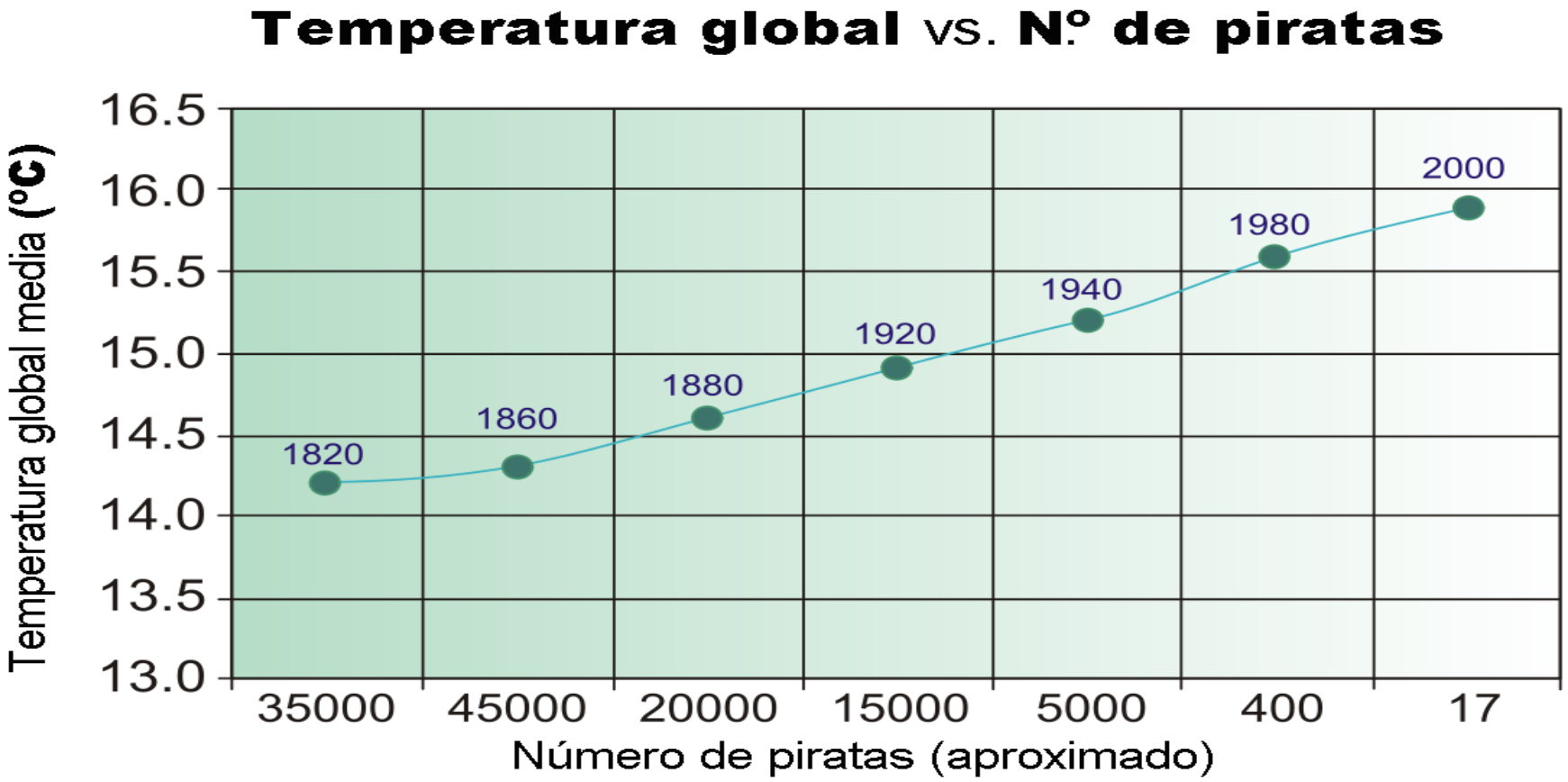




Relación entre el número de niños nacidos en Australia y el número de nidos de cigüeñas



El gráfico parece indicar que existe una clara relación entre los nidos de cigüeña y el número de nacimientos (bebes humanos), lógico si tenemos en cuenta que a los niños los traen las cigüeñas. Véase *Cigüeñas*, Warner Bros Pictures, 2016 (http://www.ciguenaslapelicula.com/)



El gráfico muestra que hay una perfecta correlación entre el aumento de temperaturas del planeta y la disminución de piratas desde el año 1820. En efecto, tal catástrofe climatológica va en aumento como consecuencia de un castigo divino en respuesta a la disminución de los valores religiosos pastafaristas a nivel mundial. La creencia central de esta iglesia pastafarista es que el Monstruo de Espagueti Volador, invisible e indetectable, creó el universo después de beber mucho. La borrachera del monstruo explica las imperfecciones del mundo creado (https://iglesiapastafari.es). Los piratas, corsarios y bucaneros eran en realidad «seres absolutamente divinos». Su imagen de ladrones y forajidos es fruto de la desinformación extendida por los teólogos cristianos en la Edad Media y por los hare krishnas. El pastafarismo afirma que en realidad son «exploradores amantes de la paz y diseminadores de la buena voluntad» que distribuían caramelos entre los niños pequeños, e indican que los piratas modernos no tienen nada que ver con «los bucaneros buscadores históricos de diversión». Aunque, por otra parte, el incremento de actividades relacionadas con la piratería en el golfo de Adén es una prueba adicional de esta teoría, ya que Somalia tiene en la actualidad «el mayor número de piratas y las menores emisiones de carbono del mundo». Véase, 19 de septiembre: Día Internacional de Hablar como los Piratas.

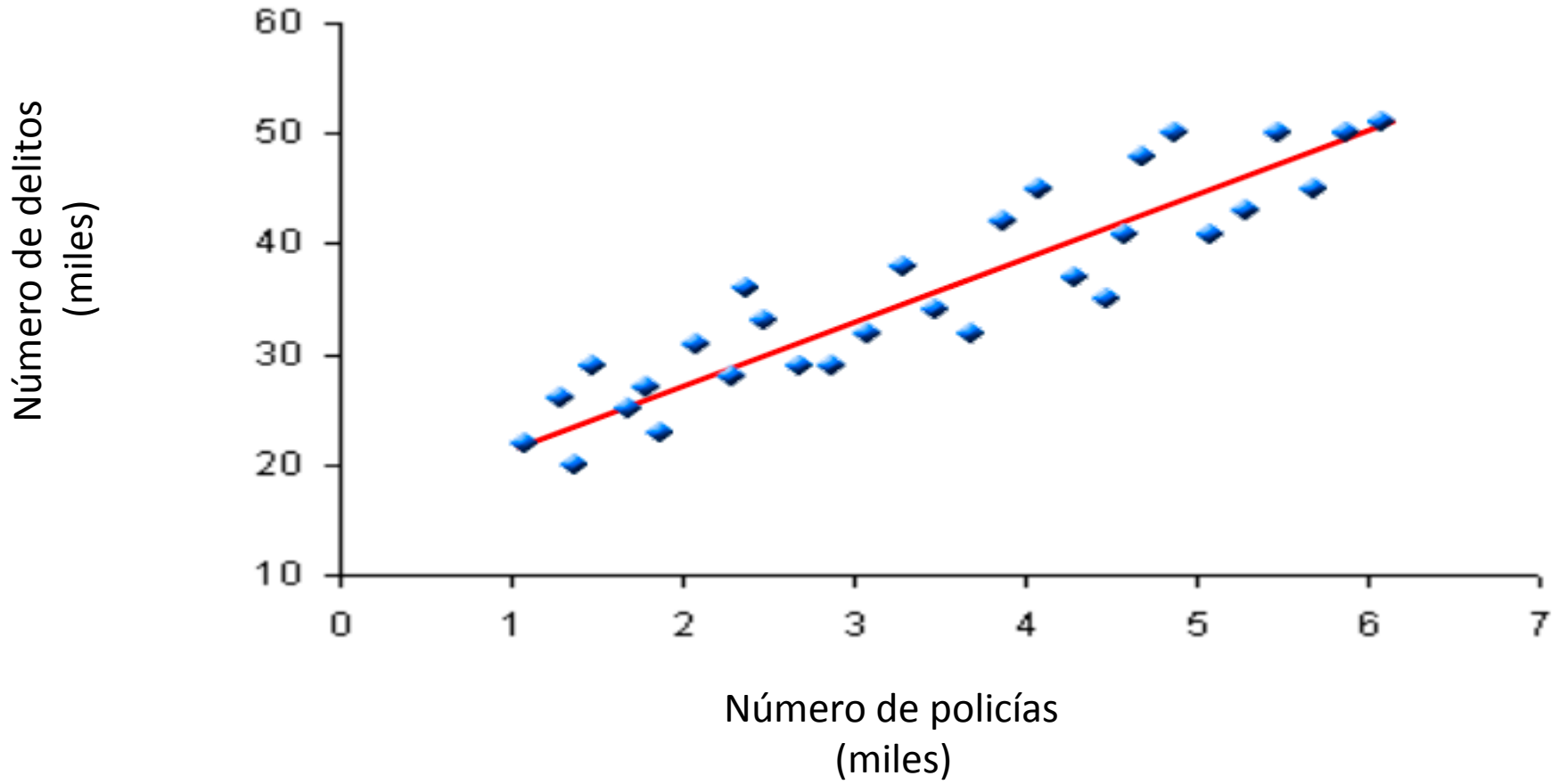
Conclusión: OJO con los datos, a veces las correlaciones son engañosas

Correlación no siempre implica relación o asociación entre variables …

Pero es que a veces, correlación y asociación tampoco implica causalidad …

Asociación entre delincuencia y número de policías

Datos de 30 municipios de Estados Unidos



Medidas de asociación con variables categóricas

Ojo **en variables cualitativas no puede calcularse medias, ni desviaciones típicas, ni correlaciones**

…. Habrá que quedarse con las tablas de contingencia y analizar si existe asociación o no Entre diferentes categorías

Se utilizan tablas de frecuencias y modas

Test Chi2 de independencia

Como la variable sexo es categórica, cuando se quiere analizar las diferencias de género respecto a otras variables categóricas se habla de indicadores de género (Asociación entre el sexo y otras variables categóricas)

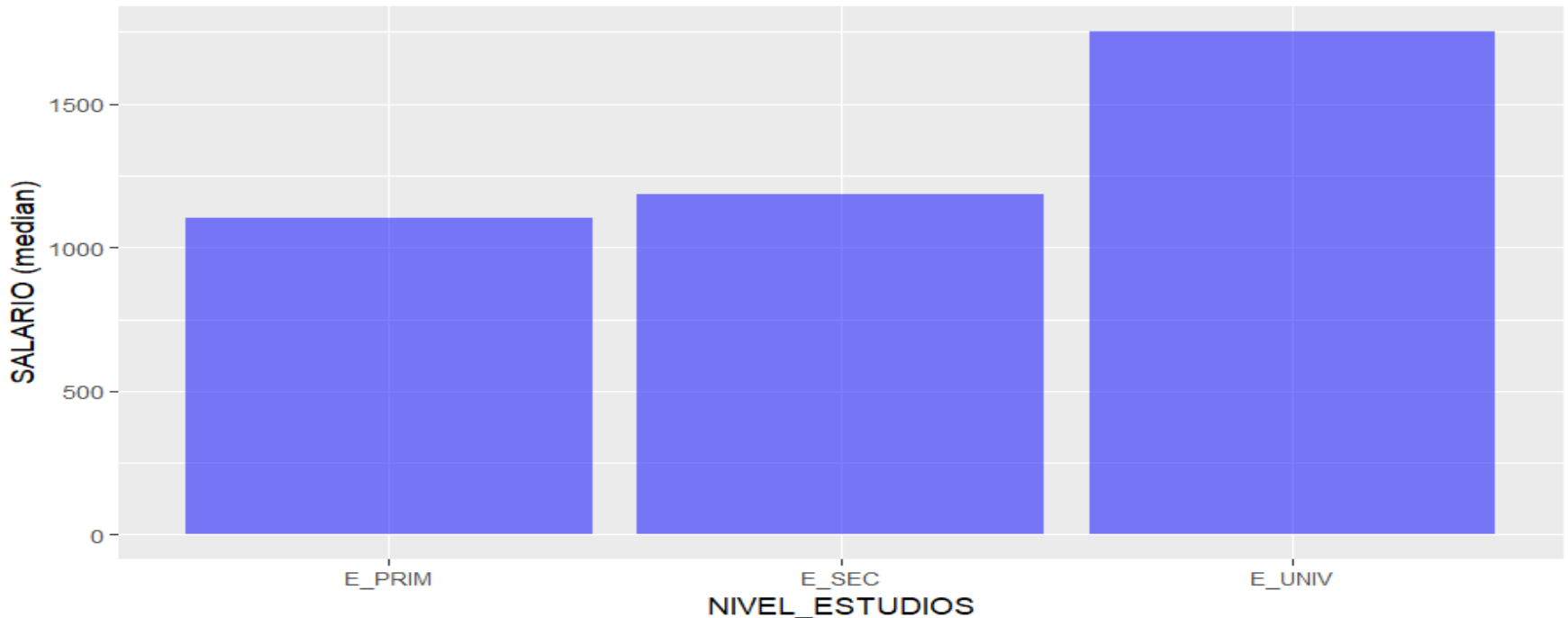
Medidas o indicadores de participación, segregación y concentración (Constituyen una vía indirecta para analizar si hay asociación o no entre categorías

Medidas de Asociación entre X (cuantitativa) e Y (categórica)

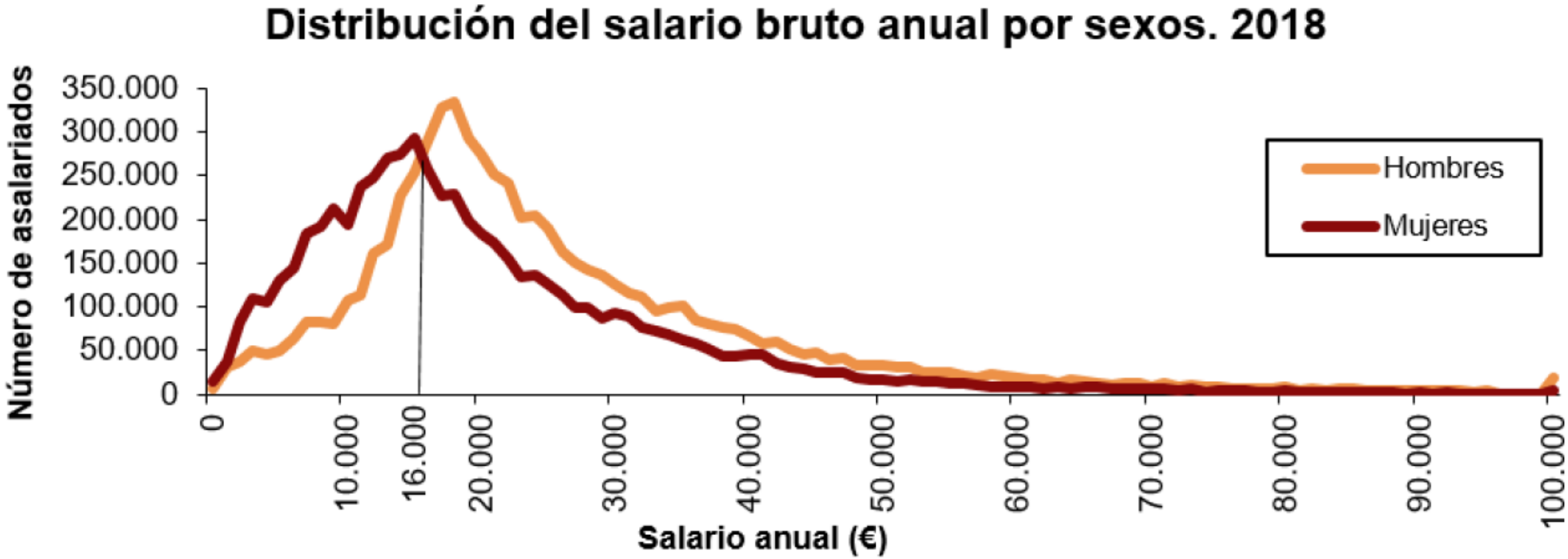
Las variables categóricas no pueden resumirse, pero las cuantitativas sí se pueden resumir en media, mediana, o con medidas de posición no central,....

Analizar la asociación entre una variable CATEGORICA (Nivel de Estudios) y una variable numérica (Salario Medio) es equivalente a analizar si las medias o medianas (o la medida de resumen de la variable numérica) es la misma en todas las categorías o niveles de la variable CATEGORICA

NIVEL_ESTUDIOS	SALARIO
All	All
E_PRIM	1,100.0
E_SEC	1,183.0
E_UNIV	1,750.0
Total	**1,325.7**



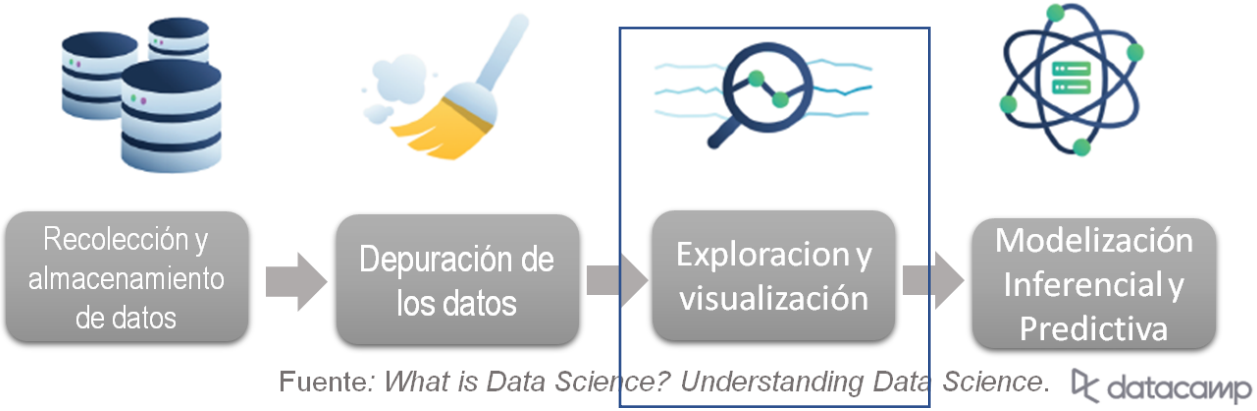
Datos de la *brecha salarial de género* en España



EJEMPLO DE ANALISIS EXPLORATORIO EN R

RADIANT:

- TITANIC
- SALARIOS EN ESPAÑA



Fuente: *What is Data Science? Understanding Data Science.* datacamp

EL ANALISTA DE DATOS Y LOS CUADROS DE MANDO

Dashboard: VISUALIZACION INTERACTIVA

En R: shiny - flexdashborad, (Qlik PowerBI, Tableau, Excel)
https://shiny.rstudio.com/
https://rstudio.github.io/shinydashboard/
https://rstudio.github.io/flexdashboard/articles/examples.html

ATLAS OF ECONOMIC COMPLEXITY, BANCO MUNDIAL
https://atlas.cid.harvard.edu/

tableau

Qlik LEAD WITH DATA

Power BI

ATLAS OF ECONOMIC COMPLEXITY

flexdashboard

Shiny

MODELIZACION