1993 - 2018
25
ETSE-UV
25 Anys d'Enginyeria

Campus de Burjassot - Paterna
ETSE-UV
Escola Tècnica Superior d'Enginyeria
Universitat de València

Introducción al Tratamiento de Datos

Tratamiento de Los Datos. Grado en Ciencia de Datos

Marcelino Martínez Sober
Departamento de Ingeniería Electrónica,
Escuela Técnica Superior de Ingeniería
Universidad de Valencia, Avda Universidad S/N
46100, Burjassot (Valencia)
marcelino.martinez@uv.es

IDAL
Intelligent
Data
Analysis
Laboratory

Firmar asistencia en los primeros 15 minutos
Firma AHORA!

CONTENIDO

1. Introducción al Tratamiento de Datos.

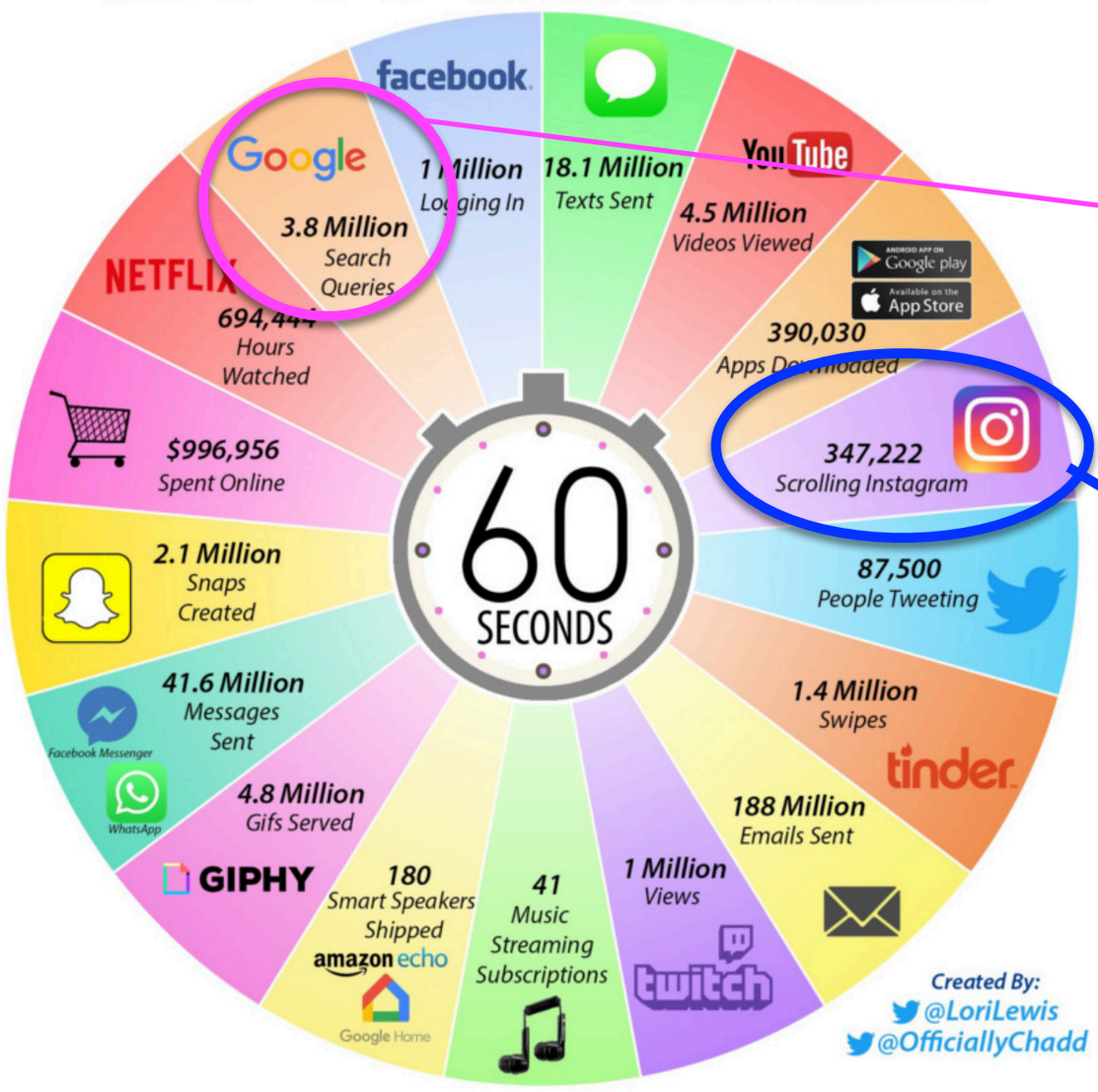
   1.1. ¿Por qué analizar datos?

   1.2. Visión global de un problema de Tratamiento de Datos

Bibliografía:

Ronald K. Pearson - Exploratory Data Analysis Using R (2018, Chapman and Hall_CRC)

¿Por qué analizar Datos?



¿POR QUÉ ANALIZAR DATOS?

Hace una década se indicó que la cantidad total de datos que se estaban produciendo era de, aproximadamente, cinco hexabytes por año ($5 \cdot 10^{18}$ bytes), más que el número total estimado de palabras habladas por los seres humanos en toda nuestra historia. Actualmente según Forbes se genera esa misma cantidad CADA MINUTO. Cada vez más, nos encontramos "ahogándonos en un océano de datos", planteando preguntas como "¿Qué hacemos con todo esto? ¿ Cómo empezamos a darles algún sentido ?

1. **Comprender lo que ha sucedido o lo que está sucediendo;**
2. **Predecir lo** que puede suceder, ya sea en el futuro o en otras circunstancias que aún no hemos visto;
3. Para guiarnos en la toma de decisiones.

¿POR QUÉ ANALIZAR DATOS?

Comprender lo que ha sucedido o lo que está sucediendo;

La asignatura Tratamiento de los Datos se va a centrar en esta primera cuestión. Los apartado 2 y 3 se basan en el desarrollo modelos matemáticos obtenidos a partir de los datos. En la medida en que los datos utilizado sean representativos y fiables las otra etapas permitirán obtener modelos para realizar predicciones con nuevos datos.

La comprensión de los datos, sus características, limitaciones, etc. será imprescindible para la toma de decisiones sobre el problema que se desea analizar.

Existen diferentes tipos de problemas que se pueden plantear: predicción, clasificación, agrupamiento, etc. En todos ellos es IMPRESCINDIBLE un análisis y conocimiento preciso de los DATOS

Brain Storming

¿ Qué consideraciones hemos de realizar para abordar un problema de Tratamiento de datos?

C/ MENORCA (C.C. AQUA) 19
46023 VALENCIA
TELÉFONO: 963319378
09/01/2024 19:30 OP: 67594
FACTURA SIMPLIFICADA: 3075-016-207716

	Descripción	P. Unit	Importe
1	PECHUGA 2 UND		4,95
1	BERENJENA		
	0,690 kg	1,89 €/kg	1,30
1	CALABACIN VERDE		
	0,650 kg	1,69 €/kg	1,10
		TOTAL (€)	**7,35**
		TARJETA BANCARIA	**7,35**

IVA	BASE IMPONIBLE (€)	CUOTA (€)
10%	4,50	0,45
0%	2,40	0,00
TOTAL	**6,90**	**0,45**

TARJ. BANCARIA: **** **** **** 2545
N.C: 077763746 AUT: 455007
AID: A0000000041010 ARC: 3030

MASTERCARD
Importe: 7,35 € DEBIT MASTERCARD

Brain Storming

¿ Qué consideraciones hemos de realizar para abordar un problema de Tratamiento de datos?

1-¿Qué problema o qué preguntas queremos responder ?
2.¿Qué datos podemos /creemos que van a ser necesarios ? ¿Siempre lo sabemos ?
Obtención de los datos
3. De qué tipo son los datos (texto, imágenes, sonidos, numéricos)
4. Disponibilidad de los datos. Acceso (ficheros, BBDD, dispersos en páginas web).
5. Disponibilidad de una clara descripción de los datos (Formato, datos sobre los datos (metadata), cómo importarlos, cómo están organizados,…)
6. El conjunto de datos se va a incrementar paulatinamente o es un conjunto definitivo (automatización de los procesos)

Respuesta DeepSeek(Enero 2025)
¿ Qué consideraciones hemos de realizar para abordar un problema de Tratamiento de datos?

Para abordar un problema de tratamiento de datos de manera efectiva, es importante seguir una serie de consideraciones y pasos estructurados. A continuación, se presentan los aspectos clave a tener en cuenta:

1. Definición del Problema

- **Objetivo claro**: Entender y definir claramente cuál es el problema que se desea resolver o la pregunta que se quiere responder con los datos.
- **Contexto**: Comprender el contexto en el que se enmarca el problema, incluyendo el ámbito de aplicación (negocio, científico, social, etc.).

2. Recopilación de Datos

- **Fuentes de datos**: Identificar y acceder a las fuentes de datos relevantes, ya sean bases de datos, APIs, archivos, sensores, etc.
- **Calidad de los datos**: Evaluar la calidad de los datos, incluyendo su completitud, exactitud, consistencia y relevancia.
- **Volumen**: Considerar el volumen de datos y si es manejable con los recursos disponibles.

3. Limpieza y Preprocesamiento de Datos

- **Limpieza**: Eliminar o corregir datos erróneos, duplicados, o incompletos.
- **Normalización**: Estandarizar los datos para que estén en un formato consistente.
- **Transformación**: Convertir los datos a un formato adecuado para el análisis, lo que puede incluir la codificación de variables categóricas, escalado de características, etc.
- **Manejo de valores faltantes**: Decidir cómo tratar los valores faltantes, ya sea imputándolos, eliminándolos o utilizando técnicas avanzadas como el uso de modelos predictivos.

4. Análisis Exploratorio de Datos (EDA)

- **Visualización**: Utilizar gráficos y visualizaciones para entender la distribución, tendencias, y relaciones entre las variables.
- **Estadísticas descriptivas**: Calcular medidas como medias, medianas, desviaciones estándar, etc., para resumir los datos.
- **Detección de patrones**: Identificar patrones, correlaciones, y posibles anomalías en los datos.

5. Selección de Técnicas y Herramientas

- **Métodos estadísticos**: Decidir si se utilizarán métodos estadísticos tradicionales o técnicas de aprendizaje automático.
- **Algoritmos**: Seleccionar los algoritmos adecuados según el tipo de problema (clasificación, regresión, clustering, etc.).
- **Herramientas**: Elegir las herramientas y lenguajes de programación adecuados (Python, R, SQL, etc.) y las bibliotecas correspondientes (Pandas, NumPy, Scikit-learn, TensorFlow, etc.).

Ver: normalización_datos.Rmd

Respuesta DeepSeek(Enero 2025)

¿ Qué consideraciones hemos de realizar para abordar un problema de Tratamiento de datos?

Para abordar un problema de tratamiento de datos de manera efectiva, es importante seguir una serie de consideraciones y pasos estructurados. A continuación, se presentan los aspectos clave a tener en cuenta:

6. Modelado y Validación

- **Entrenamiento**: Entrenar los modelos con los datos preparados.
- **Validación**: Utilizar técnicas como la validación cruzada para evaluar el rendimiento del modelo.
- **Ajuste de hiperparámetros**: Optimizar los hiperparámetros del modelo para mejorar su rendimiento.
- **Evaluación**: Utilizar métricas adecuadas (precisión, recall, F1-score, RMSE, etc.) para evaluar el modelo.

7. Interpretación de Resultados

- **Análisis de resultados**: Interpretar los resultados del modelo en el contexto del problema original.
- **Visualización**: Presentar los resultados de manera clara y comprensible, utilizando gráficos y resúmenes.
- **Conclusiones**: Extraer conclusiones y recomendaciones basadas en los resultados.

8. Implementación y Despliegue

- **Integración**: Integrar el modelo en los sistemas existentes o en el flujo de trabajo.
- **Monitoreo**: Establecer mecanismos para monitorear el rendimiento del modelo en producción.
- **Mantenimiento**: Planificar el mantenimiento y actualización del modelo según sea necesario.

9. Consideraciones Éticas y Legales

- **Privacidad**: Asegurar que el tratamiento de los datos cumple con las normativas de privacidad (GDPR, CCPA, etc.).
- **Sesgo**: Evaluar y mitigar posibles sesgos en los datos y en los modelos.
- **Transparencia**: Garantizar que los procesos y decisiones basadas en datos sean transparentes y explicables.

10. Documentación y Comunicación

- **Documentación**: Documentar todo el proceso, desde la recopilación de datos hasta la implementación del modelo.
- **Comunicación**: Comunicar los resultados y hallazgos a las partes interesadas de manera clara y efectiva.

Siguiendo estas consideraciones, se puede abordar un problema de tratamiento de datos de manera sistemática y efectiva, maximizando las posibilidades de éxito en la resolución del problema planteado.

Ver: normalización_datos.Rmd

Respuesta Claude(Enero 2026)

¿ Qué consideraciones hemos de realizar para abordar un problema de Tratamiento de datos?

Consideraciones para Abordar un Problema de Tratamiento de Datos

1. **Definición del Problema**
2. **Obtención de Datos**
3. **Exploración Inicial**
4. **Limpieza y Preparación**
5. **Análisis Exploratorio**
6. **Modelado y Resultados**

Principios Transversales

Reproducibilidad • Documentación • Consideraciones Éticas • Iteración Continua • Comunicación Clara

Respuesta Claude(Enero 2026)

¿ Qué consideraciones hemos de realizar para abordar un problema de Tratamiento de datos?

Consideraciones Clave en Cada Fase

1. Definición

- ¿Cuál es el objetivo específico?
- ¿Qué preguntas responder?
- ¿Qué decisiones se tomarán?

2. Obtención

- ¿Datos suficientes y representativos?
- ¿Hay sesgos en la recolección?
- ¿Tenemos permisos para usar datos?

3. Exploración

- Dimensiones y tipos de variables
- Valores faltantes y outliers
- Distribuciones iniciales

4. Limpieza

- Tratamiento de missing values
- Gestión de outliers
- Documentar todas las decisiones

5. EDA

- Aplicar las 4 R's de Tukey
- Análisis univariado y bivariado
- Buscar patrones y relaciones

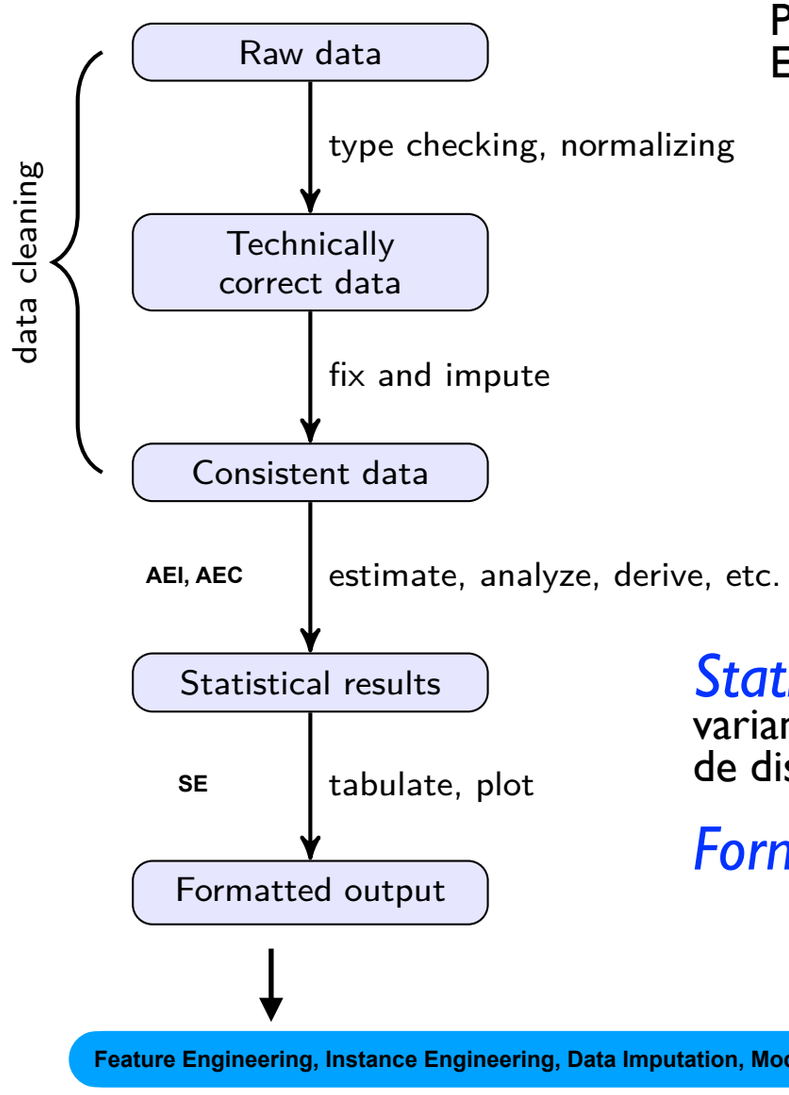
6. Modelado

- ¿Se cumplen los supuestos?
- Validación de resultados
- Interpretación en contexto

Recuerda: El análisis de datos es un proceso iterativo, no lineal

Visión global de un problema de Tratamiento de Datos

El análisis exploratorio de datos es un proceso en el que se inspeccionan, limpian, transforman y modelan datos con el objetivo de extraer información, sugerir conclusiones y servir a otras etapas posteriores.



Raw data: Son los datos originales tal como se obtienen. Esto datos pueden tener elementos faltantes, errores en la obtención, tipos de datos incorrectos (número almacenados como cadenas, etiquetas de categorías erróneas, errores de codificación, etc.) Por lo general importar estos datos como un **data.frame** de R directamente no es trivial. Es necesaria una etapa de preprocesado. (Ej. Ticket compra supermercado)

Technically correct data: Estos datos pueden ser leídos directamente en R y cargados en un data.frame, con valores correctamente etiquetados y almacenados con el tipo correcto (numérico, carácter, factor) etc. Esto no significa que estos estén libres de errores. Por ejemplo, edades negativas, fechas imposibles, datos nos presentes,… Estas inconsistencias van a depender del conjunto de datos a analizar. Una análisis previo es necesario antes de la obtención de modelos con dichos datos.

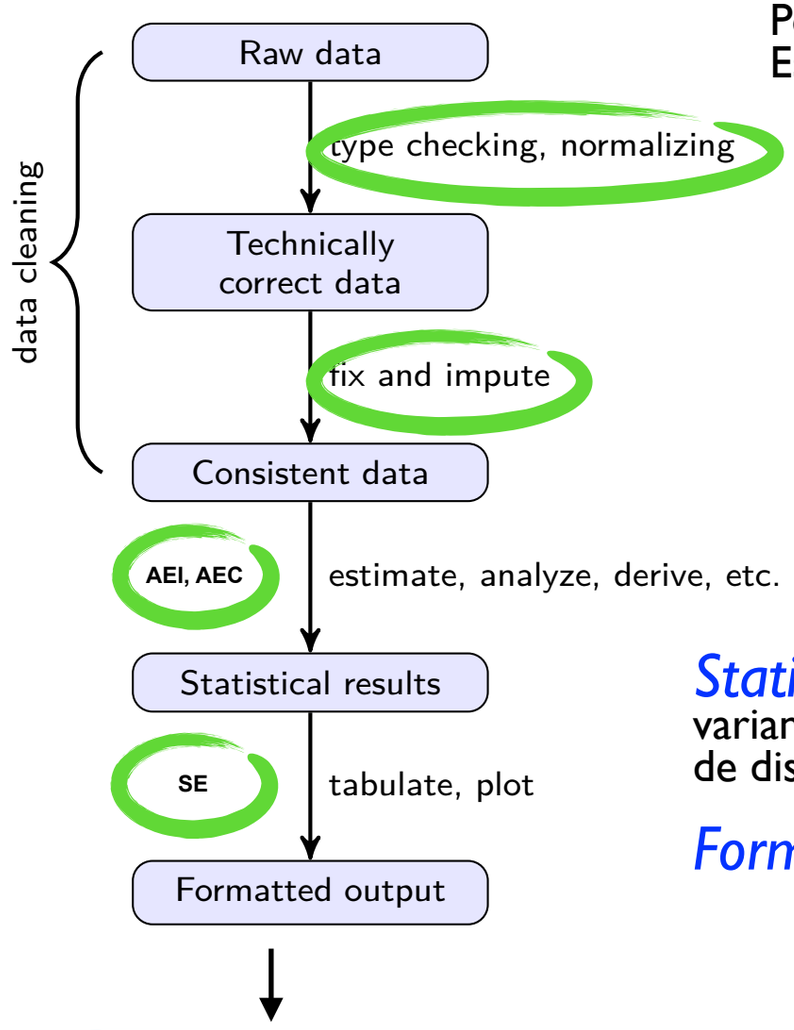
Consistent data: son los tipos de datos con los que normalmente se empieza a realizar un análisis exploratorio de datos. Sin ser una norma rígida, es en este momento cuando **se puede extraer conocimiento sobre posibles patrones en las instancias (AEI)** o **posibles relaciones entre las características (AEC)**.

Statistical result: **Obtenemos información estadística de los datos (SE)**, medias, medianas, varianzas, mínimos, máximos, etc. Este análisis nos permite obtener información sobre el el tipo de distribución, presencia de datos anómalos desde el punto de vista estadístico, etc..

Formated Output: Obtenemos gráficos o tablas para explicar las conclusiones del AED

Visión global de un problema de Tratamiento de Datos

El análisis exploratorio de datos es un proceso en el que se inspeccionan, limpian, transforman y modelan datos con el objetivo de extraer información, sugerir conclusiones y servir a otras etapas posteriores.



Raw data: Son los datos originales tal como se obtienen. Esto datos pueden tener elementos faltantes, errores en la obtención, tipos de datos incorrectos (número almacenados como cadenas, etiquetas de categorías erróneas, errores de codificación, etc.) Por lo general importar estos datos como un **data.frame** de R directamente no es trivial. Es necesaria una etapa de preprocesado.

Technically correct data: Estos datos pueden ser leídos directamente en R y cargados en un data.frame, con valores correctamente etiquetados y almacenados con el tipo correcto (numérico, carácter, factor) etc. Esto no significa que estos estén libres de errores. Por ejemplo, edades negativas, fechas imposibles, datos nos presentes,… Estas inconsistencias van a depender del conjunto de datos a analizar. Una análisis previo es necesario antes de la obtención de modelos con dichos datos.

Consistent data: son los tipos de datos con los que normalmente se empieza a realizar un análisis exploratorio de datos. Sin ser una norma rígida, es en este momento cuando **se puede extraer conocimiento sobre posibles patrones en las instancias (AEI)** o **posibles relaciones entre las características (AEC)**.

Statistical result: **Obtenemos información estadística de los datos (SE)**, medias, medianas, varianzas, mínimos, máximos, etc. Este análisis nos permite obtener información sobre el el tipo de distribución, presencia de datos anómalos desde el punto de vista estadístico, etc..

Formated Output: Obtenemos gráficos o tablas para explicar las conclusiones del AED

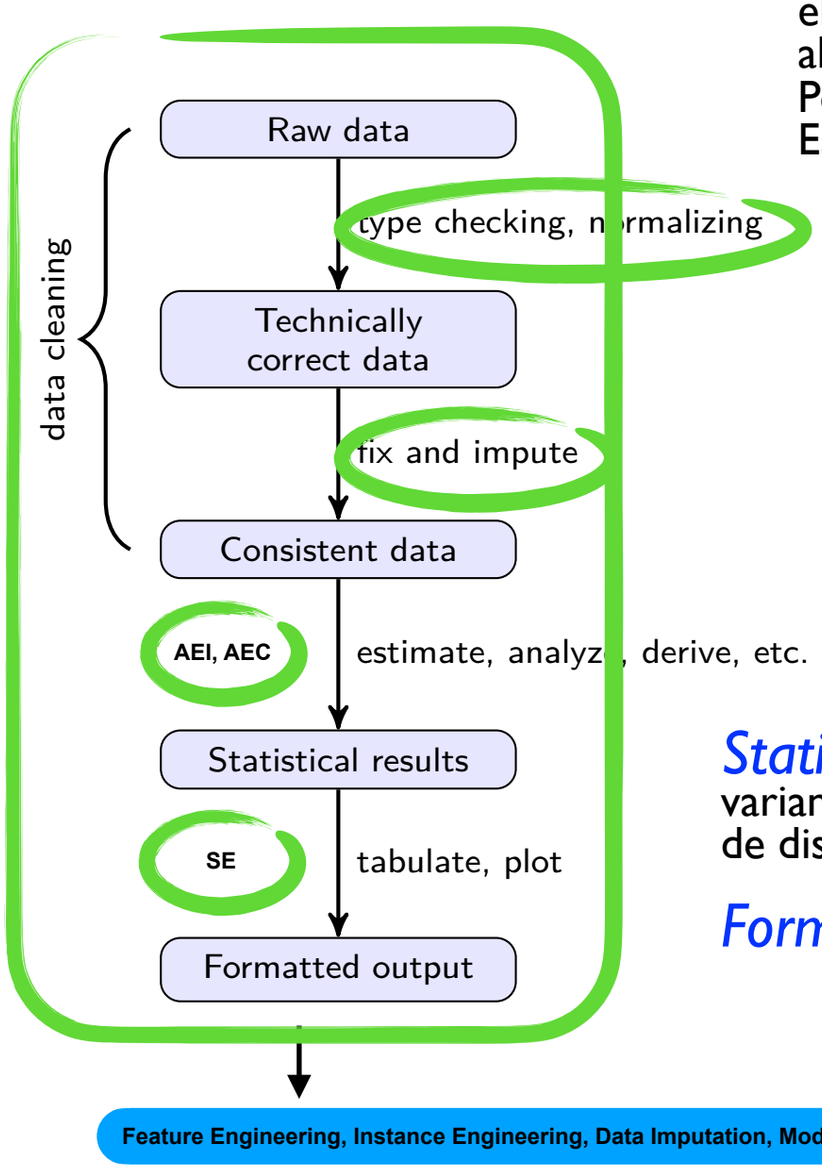
AEC: Análisis Exploratorio de Características. Dimensionality reduction.

AEI: Análisis Exploratorio de Instancias. Búsqueda de prototipos.

SE: Statistical Exploration.

Visión global de un problema de Tratamiento de Datos

El análisis exploratorio de datos es un proceso en el que se inspeccionan, limpian, transforman y modelan datos con el objetivo de extraer información, sugerir conclusiones y servir a otras etapas posteriores.



Raw data: Son los datos originales tal como se obtienen. Esto datos pueden tener elementos faltantes, errores en la obtención, tipos de datos incorrectos (número almacenados como cadenas, etiquetas de categorías erróneas, errores de codificación, etc.) Por lo general importar estos datos como un **data.frame** de R directamente no es trivial. Es necesaria una etapa de preprocesado.

Technically correct data: Estos datos pueden ser leídos directamente en R y cargados en un data.frame, con valores correctamente etiquetados y almacenados con el tipo correcto (numérico, carácter, factor) etc. Esto no significa que estos estén libres de errores. Por ejemplo, edades negativas, fechas imposibles, datos nos presentes,… Estas inconsistencias van a depender del conjunto de datos a analizar. Una análisis previo es necesario antes de la obtención de modelos con dichos datos.

Consistent data: son los tipos de datos con los que normalmente se empieza a realizar un análisis exploratorio de datos. Sin ser una norma rígida, es en este momento cuando **se puede extraer conocimiento sobre posibles patrones en las instancias (AEI)** o **posibles relaciones entre las características (AEC)**.

Statistical result: **Obtenemos información estadística de los datos (SE)**, medias, medianas, varianzas, mínimos, máximos, etc. Este análisis nos permite obtener información sobre el el tipo de distribución, presencia de datos anómalos desde el punto de vista estadístico, etc..

Formated Output: Obtenemos gráficos o tablas para explicar las conclusiones del AED

Es muy importante tener en cuenta el conocimiento adquirido en el AED ya que éste tendrá un efecto directo en etapas posteriores del problema de ciencia de datos (Feature Engineering, Instance Engineering, Imputación de datos, etc..), por lo que es fundamental tener un control sobre la reproducibilidad de TODO el código que implementa el AED.

Ver: Raw2Consistente.Rmd

Ejercicio. Recogida de datos

Rellena la tarea. Nos Autoanalizamos (En Aula Virtual)

Recogida de datos manual:

¿ Crees que el modo de recogida tendrá un efecto sobre el resultado ?

- Tarea Nos Autoanalizamos (hoja de cálculo Google)

Análisis Confirmatorio/Análisis Exploratorio

Análisis
Descriptivo

Análisis
Exploratorio

Análisis
confirmatorio
inferencial

ANÁLISIS CONFIRMATORIO/ANÁLISIS EXPLORATORIO

- Análisis Confirmatorio(ACD): el objetivo es confirmar o refutar una hipótesis. Procedimiento:

  1. El análisis del problema se formula en base a una hipótesis verificable. (El valor medio de la secuencia {x}, recogida bajo condiciones controladas A, es el mismo que el valor medio de la secuencia {y} recogida bajo condiciones controladas B).
  2. Se recoge un conjunto de datos y se computa un estadístico (p.ej. el valor medio de la diferencia entre medias).
  3. Según los resultados del estadístico, la hipótesis es aceptada (la hipótesis de igualdad de medias es consistente con los datos) o rechazada (la hipótesis no es consistente con los datos)
  4. Muy utilizado en ensayos médicos. Requiere de una adecuada preparación del experimento(se verá en otro módulo)

ANÁLISIS CONFIRMATORIO/ANÁLISIS EXPLORATORIO

- **Análisis Exploratorio(AED)**: se utiliza cuando no tenemos información de los datos como para plantear hipótesis. El objetivo es "buscar estructuras dentro de los datos que puedan resultar de interés".

Diferencia:

En AED, no sabemos cómo explicar o describir los datos y vamos analizando posibilidades, mientras que en ACD, las hipótesis están claras de partida y se busca una "prueba" de las mismas.

- Definiciones AED / EDA:
  - "EDA is the art of looking at one or more dataset in an effort to understand the underlying structure of the data contained there"
  - "We look at numbers of graphs and try to find patters. We pursue leads suggested by background information, imagination, patters perceived, and experience with other data analyses"

"It is an empirical fact that using the tools of EDA to root about in large rich collections of data leads to useful results"

Etapas en un problema de análisis de datos.

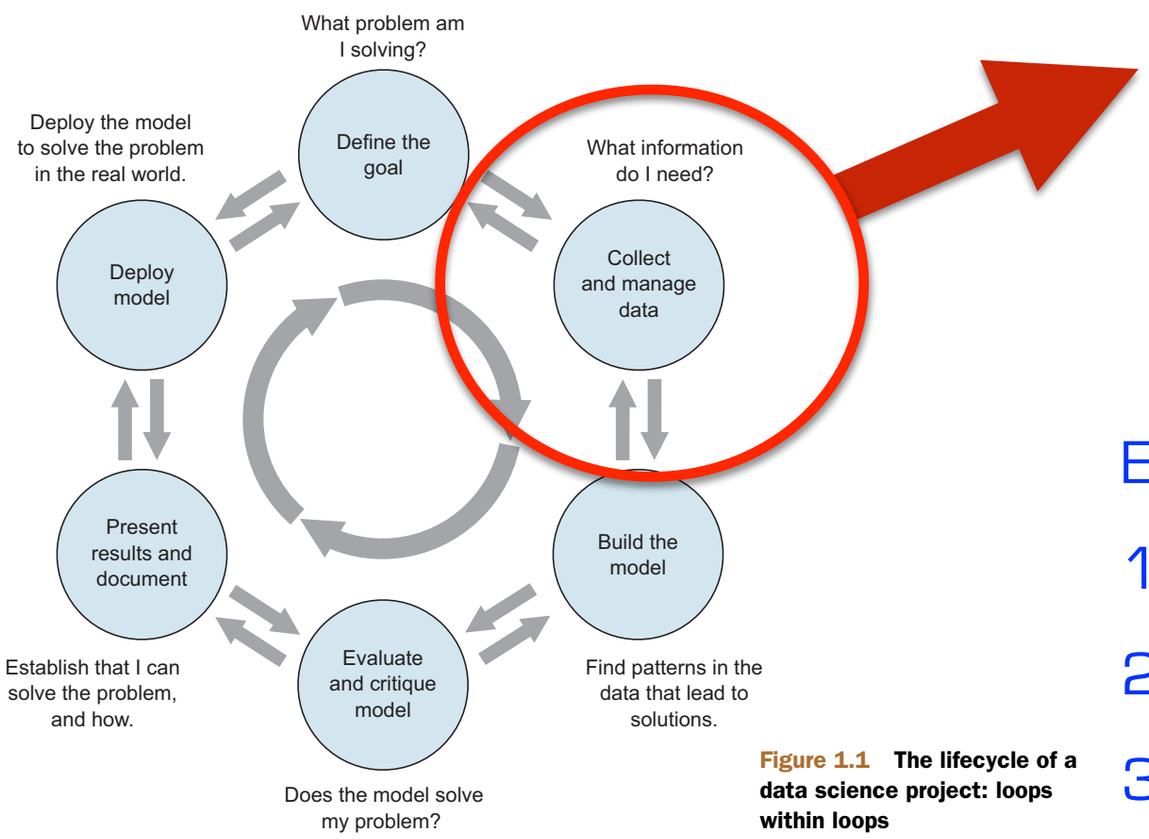
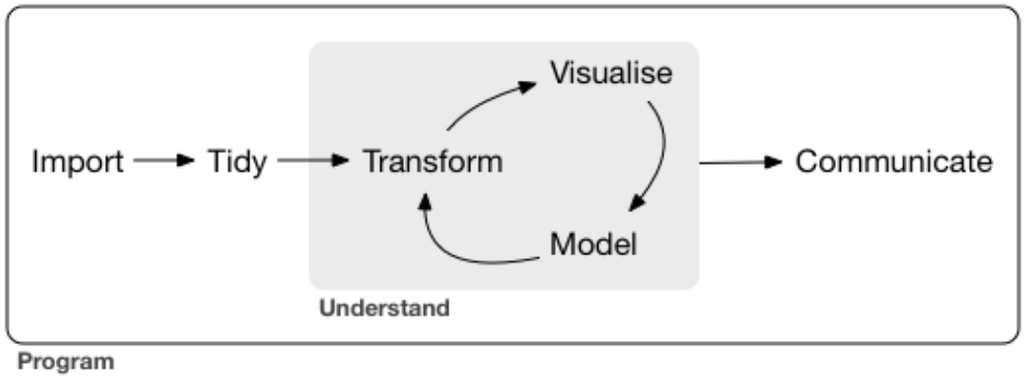


Figure 1.1 **The lifecycle of a data science project: loops within loops**

Fuente: Practical Data Science with R, Nina Zumel,John Mount



Fuente: R for data science http://r4ds.had.co.nz/introduction.html

Etapas Análisis de Datos

1. Importación
2. Preparación de los datos
3. Transformación.
4. Visualización.
5. Modelado.
6. Comunicación de resultados. (Investigación reproducible)

¡¡ Documenta tus Rmarkdowns!!

LAS 4 "R" DEL ANÁLISIS EXPLORATORIO DE DATOS.

Según John Tukey

1. Revelation
Revelación

Revelar la estructura subyacente de los datos mediante visualizaciones y resúmenes estadísticos para hacer visible lo que está oculto en los números brutos.

Ejemplo: Histogramas, boxplots, scatter plots que revelan distribuciones bimodales, outliers o correlaciones no evidentes en tablas.

2. Residuals
Residuos

Examinar lo que queda después de ajustar un modelo o eliminar patrones principales para verificar si el modelo captura adecuadamente la estructura de los datos.

Ejemplo: Tras ajustar una regresión lineal, graficar residuos vs valores ajustados. Un patrón curvo sugiere relación cuadrática.

3. Re-expression
Re-expresión

Transformar variables para simplificar patrones o cumplir supuestos estadísticos, buscando la escala donde los datos son más simples de analizar.

Ejemplo: Aplicar log(x) a datos de ingresos altamente sesgados para obtener una distribución más simétrica y estabilizar varianzas.

4. Resistance
Resistencia

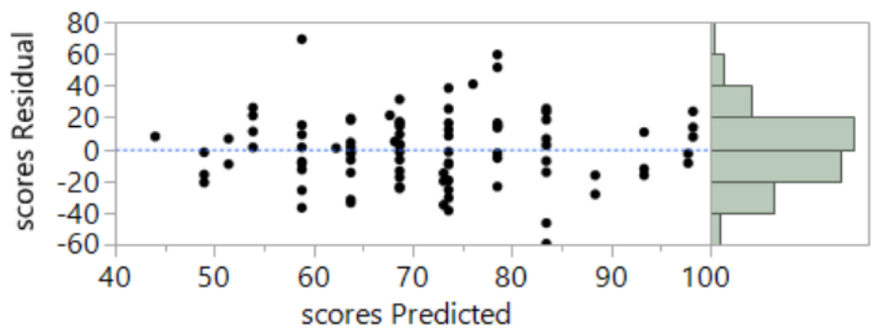
Usar técnicas robustas que resistan la influencia de valores atípicos y datos extremos para obtener conclusiones estables incluso con outliers.

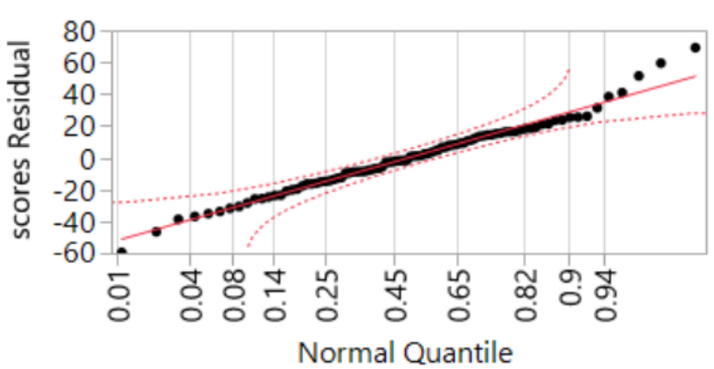
Ejemplo: Usar la mediana en lugar de la media para medir el centro, o el rango intercuartílico en lugar de la desviación estándar.

LAS 4 "R" DEL ANÁLISIS EXPLORATORIO DE DATOS. RESIDUOS

Random residual plot

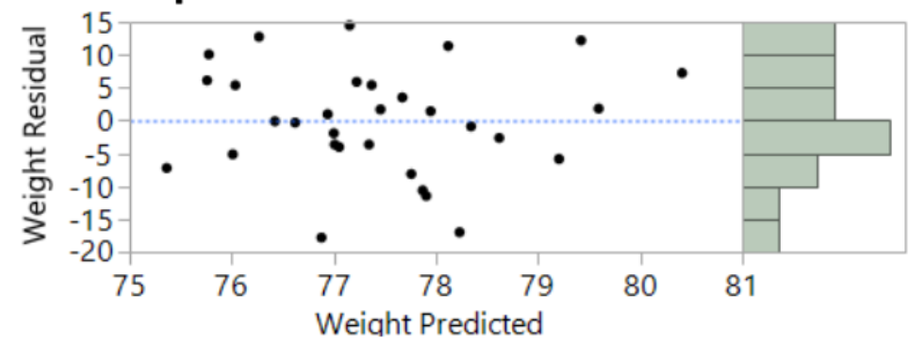
- No systematic pattern
- Normal distribution

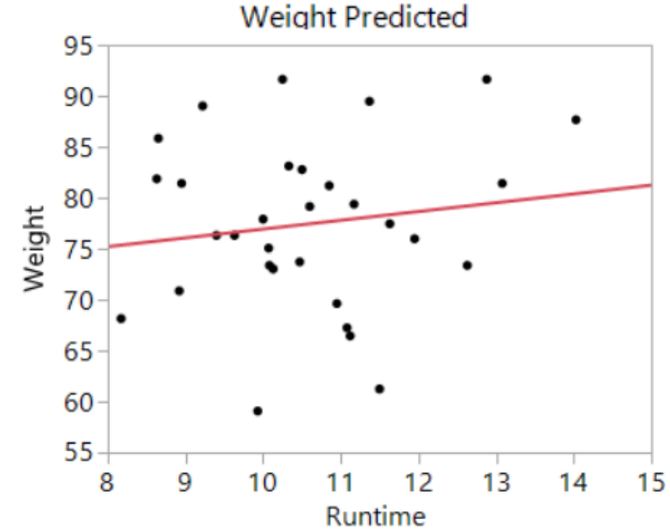




Strange residual patterns

- Fitness data
- Residuals are not normally distributed.
- Explore another model!

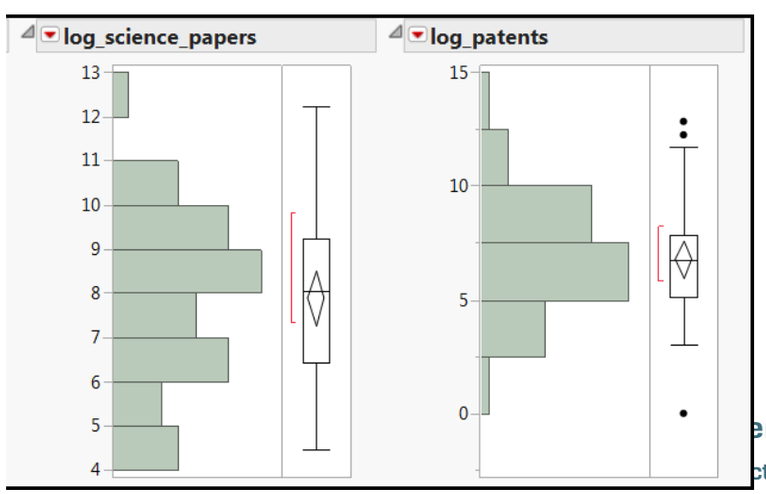




LAS 4 "R" DEL ANÁLISIS EXPLORATORIO DE DATOS. REEXPRESION (TRANSFORMACIÓN)

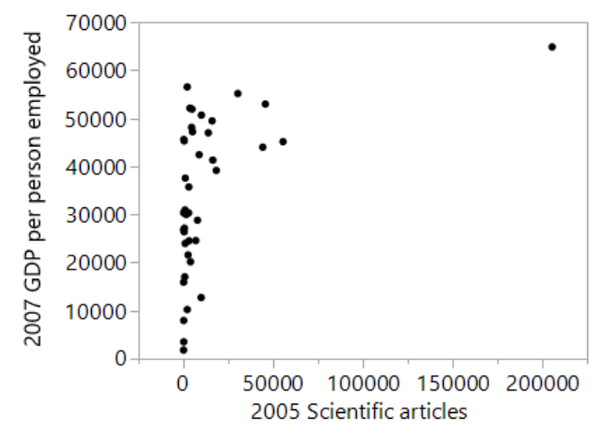
Distribuciones Sesgadas

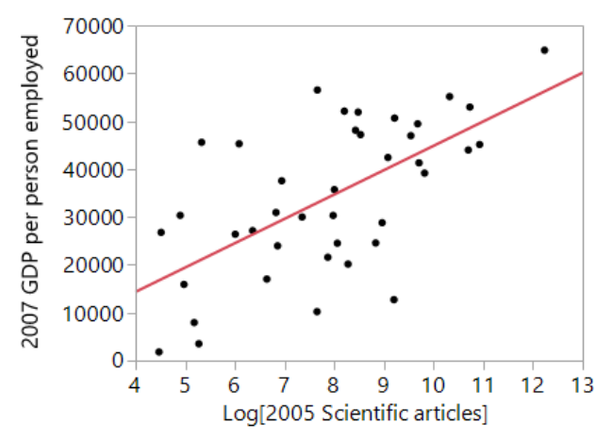
- Distribución de publicaciones científicas y patentes por países.
- **Transformación logarítmica y*=log(y)**



Antes y después

- Un ajuste lineal (regresión) tras la transformación tiene más sentido





Las 4 "R" del Análisis Exploratorio de datos. Resistencia

- La resistencia no es lo mismo que la robustez.
- Resistencia: inmune a los valores atípicos
- Robustez: inmune a violaciones de suposiciones de aplicación
- Usaremos la mediana, la media recortad, etc, para contrarrestar los valores atípicos (más adelante).

LAS 4 "R" DEL ANÁLISIS EXPLORATORIO DE DATOS. VISUALIZACIÓN (REVELATION)

- La visualización de datos es la herramienta principal de EDA. Sin "ver" el patrón de datos,…
- ¿Cómo puedes saber si los residuos son aleatorios o no?
- ¿Cómo puedes detectar si la distribución está sesgada, si la relación es no lineal y decidir si se necesita una transformación?
- ¿Cómo puedes detectar valores atípicos y cómo tratarlos?
- Visualización de datos —> más adelante

FROM RAW TO TECHNICALLY CORRECT DATA

1.Raw Data (datos en bruto)

- No se ha aplicado ningún software sobre los datos.
- No se han modificado valores de los datos.
- No se ha eliminado ningún dato.

2. Tidy Data Set (datos pre-procesados) http://tidyr.tidyverse.org/articles/tidy-data.html (REVISAR)

- Cada **variable (característica)** está representada en una ***columna***.
- Cada ***observación (instancia / registro)*** de una variable está en una ***fila*** distinta.
- Si la información está contenida en diferentes tablas debe haber una variable común que permita enlazar las tablas.
- Incluye una primera fila con los nombres de las variables.
- Los nombres deben ser representativos de las variables que almacenan (FechaNacimiento en lugar de FecNac)

3. Code Book (Metadata) (explicando cada variable y sus valores en el Tidy data Set.)

- Suele ser un fichero de texto que debe incluir: las variables representadas en cada columna, cómo se han medido, sus unidades, etc. Cuánto más completa sea la información más se facilitará el preprocesado (outliers, datos erróneos, etc)

4. Secuencia de instrucciones (programa R, Python) seguida para pasar de Raw Data a Tidy Data(=Consistent Data).

FROM RAW TO TECHNICALLY CORRECT DATA

Gapminder

¿El conjunto de datos CO2 emissions es un conjunto Tidy ?

Consumption CO_2 emissions, tonnes per capita

description: Carbondioxide per Capita (Consumption based) from GM Long Series in Tonnes of CO2 per capita
source_url: http://gapm.io/dco2_consumption_historic

VIEW AS: 🎈 〽️

DOWNLOAD AS: ⬇️CSV ⬇️XLSX

country	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811	1812	1813	1814
Afghanistan	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Angola	0.009	0.009	0.009	0.009	0.009	0.009	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Albania	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Andorra	0.333	0.335	0.337	0.34	0.342	0.345	0.347	0.35	0.352	0.355	0.357	0.36	0.363	0.365	0.368
UAE	0.063	0.063	0.064	0.064	0.064	0.064	0.065	0.065	0.065	0.065	0.066	0.066	0.066	0.066	0.067
Argentina	0.119	0.121	0.122	0.124	0.125	0.127	0.128	0.13	0.132	0.133	0.135	0.137	0.138	0.14	0.142
Armenia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Antigua and Barbuda	0.018	0.018	0.019	0.019	0.019	0.019	0.02	0.02	0.02	0.02	0.02	0.021	0.021	0.021	0.022
Australia	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076
Austria	0.051	0.052	0.052	0.052	0.053	0.053	0.054	0.054	0.07	0.07	0.071	0.071	0.072	0.072	0.073
Azerbaijan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Burundi	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Belgium	1.93	1.95	1.96	1.32	1.33	1.34	1.35	1.36	1.37	1.38	1.39	1.4	1.41	1.42	1.43

Variables: ?????

FROM RAW TO TECHNICALLY CORRECT DATA

Gapminder

¿El conjunto de datos CO2 emissions es un conjunto Tidy ?

Consumption CO_2 emissions, tonnes per capita

description: Carbondioxide per Capita (Consumption based) from GM Long Series in Tonnes of CO2 per capita
source_url: http://gapm.io/dco2_consumption_historic

VIEW AS: 🎈 〽️

DOWNLOAD AS: ⬇️CSV ⬇️XLSX

country	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811	1812	1813	1814
Afghanistan	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Angola	0.009	0.009	0.009	0.009	0.009	0.009	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Albania	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Andorra	0.333	0.335	0.337	0.34	0.342	0.345	0.347	0.35	0.352	0.355	0.357	0.36	0.363	0.365	0.368
UAE	0.063	0.063	0.064	0.064	0.064	0.064	0.065	0.065	0.065	0.065	0.066	0.066	0.066	0.066	0.067
Argentina	0.119	0.121	0.122	0.124	0.125	0.127	0.128	0.13	0.132	0.133	0.135	0.137	0.138	0.14	0.142
Armenia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Antigua and Barbuda	0.018	0.018	0.019	0.019	0.019	0.019	0.02	0.02	0.02	0.02	0.02	0.021	0.021	0.021	0.022
Australia	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076
Austria	0.051	0.052	0.052	0.052	0.053	0.053	0.054	0.054	0.07	0.07	0.071	0.071	0.072	0.072	0.073
Azerbaijan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Burundi	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Belgium	1.93	1.95	1.96	1.32	1.33	1.34	1.35	1.36	1.37	1.38	1.39	1.4	1.41	1.42	1.43

Variables: Country / Year/CO2 ¿*Tidy*?

CodeBook (metadata) Ideal

N		Variables
1	Dataset Name	(DS)
2	Dataset description	(DS)
3	Dataset size (number of included variables)	(DS)
4	Variable name	All
5	Variable Type (nominal, ordinal, real, etc)	All
6	Variable definition, brief	All
7	Variable description, expanded	All
8	Number of observations	All
9	Missing values codes and interpretations	All
10	Number of missing observations	All
11	Defining formulae or procedures	Calc.
12	Information of any cleaning or transformation applied	All
13	Cautions, tips and hints of usage	All
14	List of valid values and their definitions	Cat.
15	Frequency distribution, including missing	Cat.
16	Units of measure	Real
17	Admisible range of values	R/C
18	Descriptive statistics: mean, standard deviation, MAD scale estimate, and 5-numbers summary	R/C
19	Fractions of ESD outliers, specified threshold (criteria)	R/C
20	Fractions of Hampel outliers, specified threshold	Real

DS: Dataset attributes.
All: All variables.
Cat: categorical variables (nominal, ordinal, binary)
Calc: variables calculated from other variables of the dataset.
Real: only for real-valued variables.
R/C: real-valed, integer-valued count variables
ESD: Extreme Studentized Deviate (Test para detección de outliers)

5-numbers summary: minimum, 1st quartile, median, 3rd quartile, maximum
MAD: median absolute difference (MAD=median(|Xi-median(X)|)

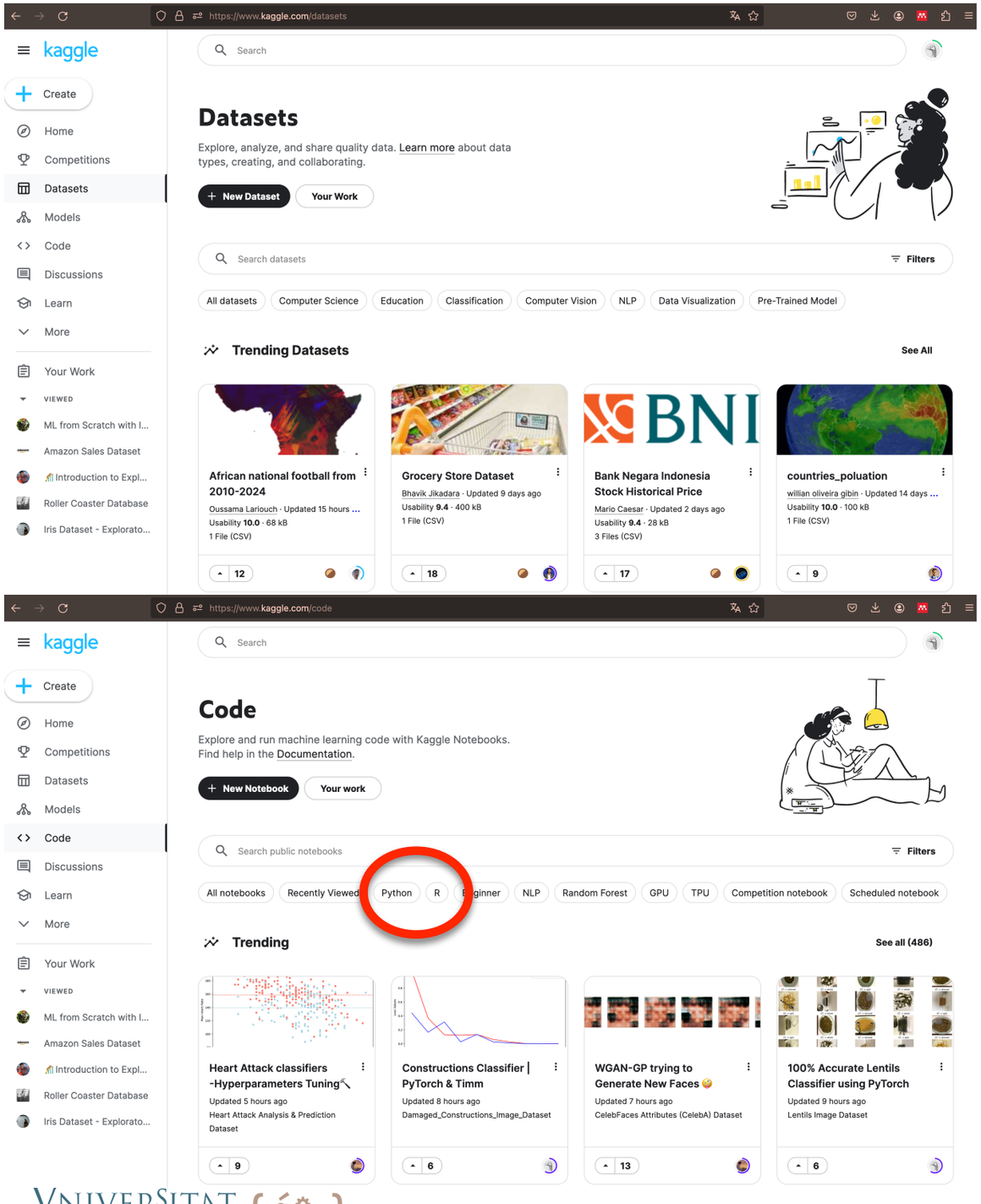
FUENTES DE DATOS

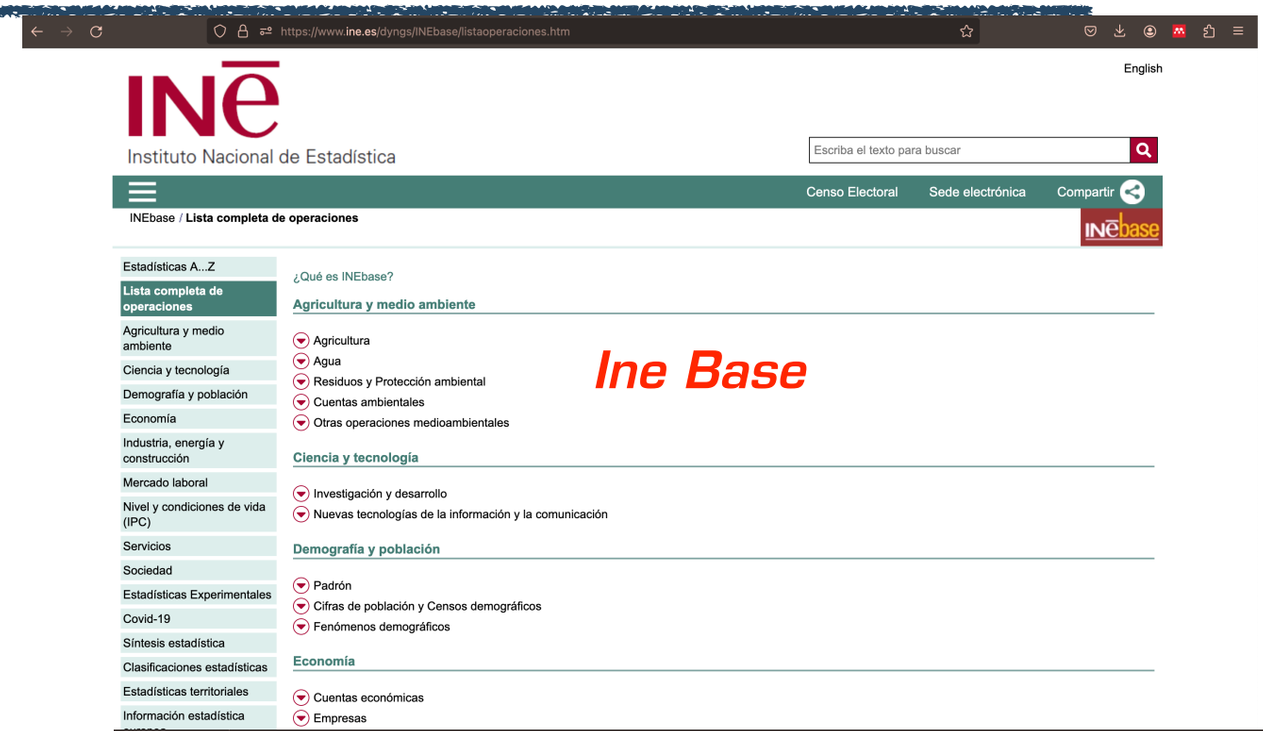
https://datasetsearch.research.google.com/ **https://ropensci.org/**

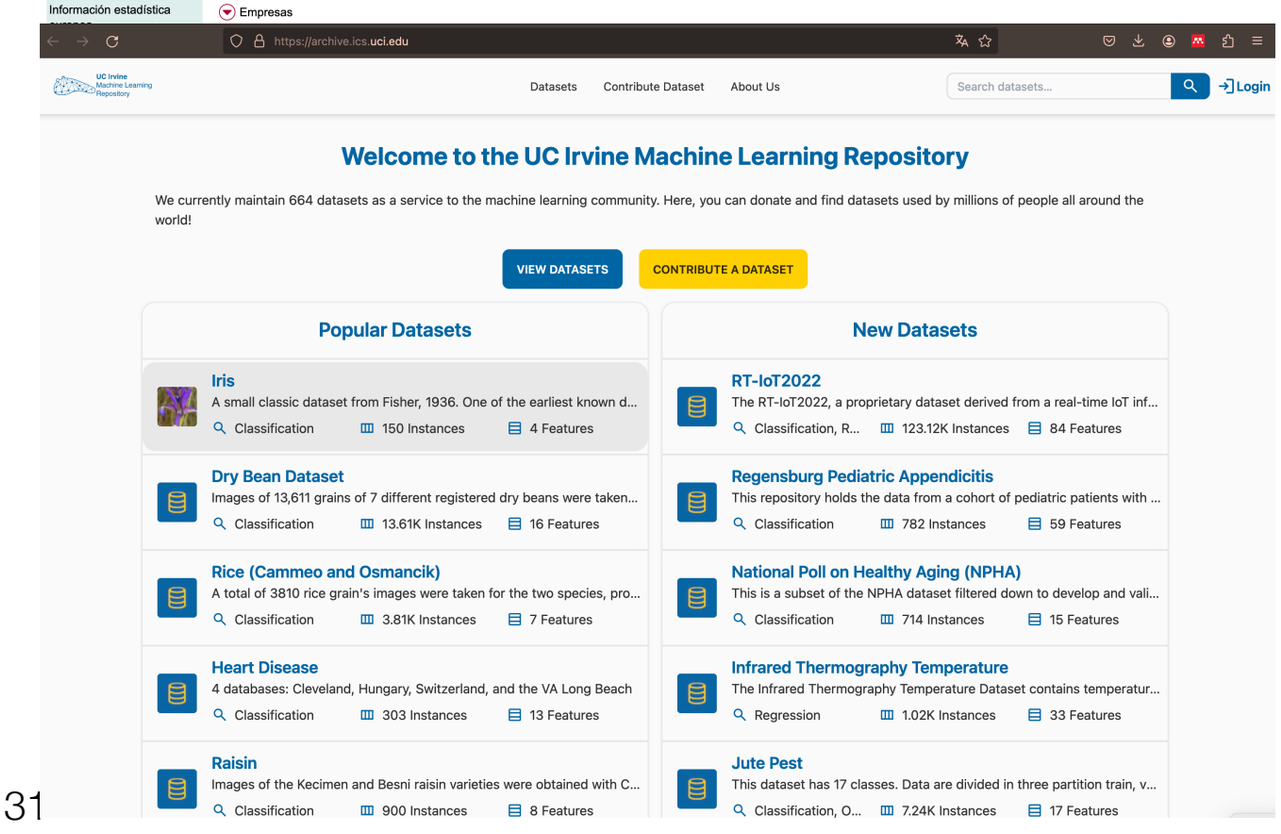
N	Enlace	Variables
1	https://www.r-bloggers.com/when-field-or-lab-work-is-not-an-option-leveraging-open-data-resources-for-remote-research/	When Field or Lab Work is not an Option Leveraging Open Data Resources for Remote Research
	http://e.infogr.am/blog/free-data-sources/	50 Amazing data sources you should know. (Enlaces a todo tipo de datos, gubernamentales, globales, académicos, de salud, etc)
2	http://data.europa.eu/euodp/en/data/	EU Open Data Portal
3	**https://archive.ics.uci.edu/ml/**	**UC Irvine Machine Learning Repository!. Infinidad de conjuntos de datos clasificados por áreas, tipos de datos, etc.**
4	https://www.kaggle.com/datasets	Kaggle, Bases de datos de múltiples problemas propuesto como retos para su resolución, incluye código R y Python
5	**http://www.ine.es/**	Instituto Nacional de Estadística (Ej. Poblac. Municipios)
6	**http://datos.gob.es/es/catalogo**	**Datos abiertos administración pública española.**
7	http://www.dadesobertes.gva.es/es/about	Datos abiertos de la Comunidad Valenciana
8	**https://www.gapminder.org**	**Indicadores diversos a nivel global múltiples campos**
9	**https://opensource.pysport.org/**	**Deportes**

Los portales de transparencia son una fuente de datos gubernamentales (http://transparencia.gob.es/)

FUENTES DE DATOS







SOFTWARE USADO PARA TRATAMIENTO DE LOS DATOS

- En la asignatura Tratamiento de los Datos **se utilizará R** como lenguaje de programación con el entorno RStudio.
- El entorno es multiplataforma y permite:
  - Importar ficheros de datos con múltiples formatos.
  - Realizar el preprocesado y análisis.
  - Generar informes de alta calidad para dar difusión a los resultados usando Rmarkdown.
- Se ha creado un plantilla en formato RMarkdown para la realización de las tareas que se propongan.(Ver Plantilla Rmarkdown en AV)
- Se podría realizar un análisis análogo utilizando Python y sus librerías como pandas, matplotlib, numpy.

Primeros Análisis usando R

Los primeros pasos cuando se analiza un conjunto de datos son los siguientes:

1. Número de registros.
2. Número de variables.
3. Tipos de variables (numéricas, categóricas, fecha, etc)
4. Hay observaciones no registradas (valores perdidos).
5. Las variables del conjunto son las que esperamos.
6. Los valores son consistentes con las variables.
7. Las relaciones entre las variables son las esperadas

CONJUNTOS DE DATOS EN R

- R dispone de múltiples librerías con conjuntos de datos que podemos utilizar directamente sin necesidad de realizar ningún tipo de importación.
- Al arrancar R se cargan, automáticamente los conjuntos de la **librería datasets. (> sessionInfo())**
- **`data(package = "datasets")`** muestra todos los conjuntos presentes en una librería, en este caso, librería datasets.
- No todas las librerías contienen conjuntos de datos.

CONJUNTOS DE DATOS EN R

- Determina los conjuntos de datos que hay en las librerías ***signal*** y ***tidyr***

- Indica de qué tipo son las variables almacenadas en el conjunto '***crabs***' de la librería **MASS**

Kahoot cangrejos !

CONJUNTOS DE DATOS EN R

- Determina los conjuntos de datos que hay en las librerías ***signal*** y ***tidyr***

```
Data sets in package 'signal':

wav                     Example wav file
```

```
Data sets in package 'tidyr':

billboard               Song rankings for Billboard top 100 in the
                        year 2000
cms_patient_care        Data from the Centers for Medicare & Medicaid
                        Services
cms_patient_experience
                        Data from the Centers for Medicare & Medicaid
                        Services
construction            Completed construction in the US in 2018
fish_encounters         Fish encounters
household               Household data
population              World Health Organization TB data
relig_income            Pew religion and income survey
smiths                  Some data about the Smith family
table1                  Example tabular representations
table2                  Example tabular representations
table3                  Example tabular representations
table4a                 Example tabular representations
table4b                 Example tabular representations
table5                  Example tabular representations
us_rent_income          US rent and income data
who                     World Health Organization TB data
who2                    World Health Organization TB data
world_bank_pop          Population data from the world bank
```

- Indica de qué tipo son las variables almacenadas en el conjunto '***crabs***' de la librería **MASS**

```
> library(MASS)
> str(crabs)
'data.frame':   200 obs. of  8 variables:
 $ sp   : Factor w/ 2 levels "B","O": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex  : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ index: int  1 2 3 4 5 6 7 8 9 10 ...
 $ FL   : num  8.1 8.8 9.2 9.6 9.8 10.8 11.1 11.6 11.8 11.8 ...
 $ RW   : num  6.7 7.7 7.8 7.9 8 9 9.9 9.1 9.6 10.5 ...
 $ CL   : num  16.1 18.1 19 20.1 20.3 23 23.8 24.5 24.2 25.2 ...
 $ CW   : num  19 20.8 22.4 23.1 23 26.5 27.1 28.4 27.8 29.3 ...
 $ BD   : num  7 7.4 7.7 8.2 8.2 9.8 9.8 10.4 9.7 10.3 ...
 '
```

CONJUNTO DE DATOS EN R

- > data(package = "MASS")

```
Data sets in package 'MASS':

Aids2                                   Australian AIDS Survival Data
Animals                                 Brain and Body Weights for 28 Species
Boston                                  Housing Values in Suburbs of Boston
Cars93                                  Data from 93 Cars on Sale in the USA in 1993
Cushings                                Diagnostic Tests on Patients with Cushing's
Syndrome
DDT                                     DDT in Kale
GAGurine                                Level of GAG in Urine of Children
Insurance                               Numbers of Car Insurance claims
Melanoma                                Survival from Malignant Melanoma
OME                                     Tests of Auditory Perception in Children with OME
.
.
.
cabbages                                Data from a cabbage field trial
```

INFORMACIÓN DE UN CONJUNTO DE DATOS

> help(package = MASS,topic = Animals)

Brain and Body Weights for 28 Species

Description

Average brain and body weights for 28 species of land animals.

Usage

```
Animals
```

Format

`body`
body weight in kg.

`brain`
brain weight in g.

> library(MASS)
> str(Animals)
'data.frame': 28 obs. of 2 variables:
$ body : num 1.35 465 36.33 27.66 1.04 ...
$ brain: num 8.1 423 119.5 115 5.5 ...

Note

The name `Animals` avoids conflicts with a system dataset `animals` in S-PLUS 4.5 and later.

Source

P. J. Rousseeuw and A. M. Leroy (1987) *Robust Regression and Outlier Detection.* Wiley, p. 57.

References

Venables, W. N. and Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS.* Third Edition. Springer.

INFORMACIÓN DE UN CONJUNTO DE DATOS

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 (numeric) variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	Engine (0 = V-shaped, 1 = straight)
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

Source

Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, **37**, 391–411.

> ??mtcars

```
> str(datasets::mtcars)

'data.frame':    32 obs. of  11 variables:

 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

1993 - 2018
25
ETSE-UV
25 Anys d'Enginyeria

Campus de Burjassot - Paterna
ETSE-UV
Escola Tècnica Superior d'Enginyeria
Universitat de València

Introducción al Tratamiento de Datos

Tratamiento de Los Datos. Grado en Ciencia de Datos

Marcelino Martínez Sober
Departamento de Ingeniería Electrónica,
Escuela Técnica Superior de Ingeniería
Universidad de Valencia, Avda Universidad S/N
46100, Burjassot (Valencia)
marcelino.martinez@uv.es

IDAL
Intelligent
Data
Analysis
Laboratory