GENERALITAT VALENCIANA
CONSELLERIA DE TRANSPARÈNCIA, RESPONSABILITAT SOCIAL, PARTICIPACIÓ I COOPERACIÓ

CÀTEDRA TRANSPARÈNCIA I PARTICIPACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

MUGI

TALLER

LIMPIEZA DE DATOS

TECHFEST
2 MAYO 2017

David Pardo – dapargi@upv.es

Taller de Limpieza de datos

Objetivo del taller

Reflexionar y entender la importancia de la Limpieza de datos antes de la explotación y visualización de los mismos

Aprender a utilizar la herramienta Open Refine (interfaz web amigable y sencilla para el usuario) para la limpieza de datos, su explotación y visualización con Google Fusion Tables

Taller de Limpieza de datos

Dirigido a

Cualquier ciudadano interesado en iniciarse en la explotación y visualizaciones con datos que necesite conocimientos sobre limpieza de sus datos

- Estudiantes cualquier edad 14 a 99 años
- Académicos / Científicos
- Periodistas
- Activistas / lobbys
- Profesionales de cualquier ámbito (empresarial, gubernamental)
- Cualquier ciudadano

Sin competencias en programación, análisis de datos (*business intelligence*, minería de datos, etc.)
Con mínimas competencias digitales (navegar, guardar archivos, … o preguntar)

Para nivel avanzado: son útiles conocimientos mínimos en hojas de cálculo y estadística (fórmulas, funciones, tablas dinámicas, tipos de gráficos) y … programación

Taller de Limpieza de datos

Índice del taller

- Introducción a la limpieza de datos: la problemática de datos erróneos y su importancia

- Utilización de la herramienta Open Refine (interfaz web amigable y sencilla para el usuario) para la limpieza de datos y, su posterior explotación y visualización.
  - Instalación
  - Importación de datos
  - LIMPIEZA y explotación de datos
  - Técnicas avanzadas
  - Exportación de datos

- Introducción a la Visualización de datos con Google Fusion Tables)
  - Importación e introducción de datos
  - Creación de hojas de gráficos
  - Creación de mapas
  - Exportación de datos y de gráficos o mapas para su publicación

Taller de Limpieza de datos

Introducción a la limpieza de datos

Introducción a la limpieza de datos: la problemática de datos erróneos y su importancia

Taller de Limpieza de datos

Introducción a la limpieza de datos > **Revolución digital: datos**

Taller de Limpieza de datos

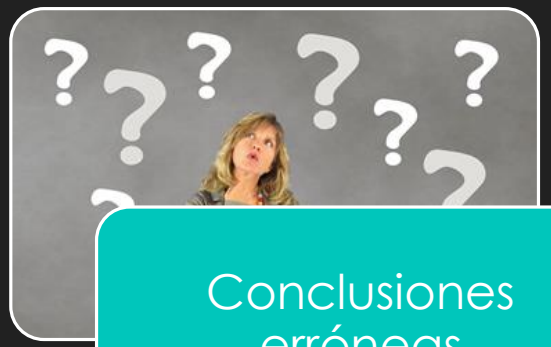
Introducción a la limpieza de datos > **Razones para limpiar datos**

Información incorrecta o inconsistente

Datos estadísticos oficiales

Datos clientes, productos, de producción en empresas

Datos académicos, científicos, …



Conclusiones erróneas ➔Decisiones incorrectas

investigaciones académicas o científicas, informes empresariales o gubernamentales

Trabajos de clase o personales, ciudadanos Trabajos periodísticos, …

Malas inversiones (a escala pública o privada)

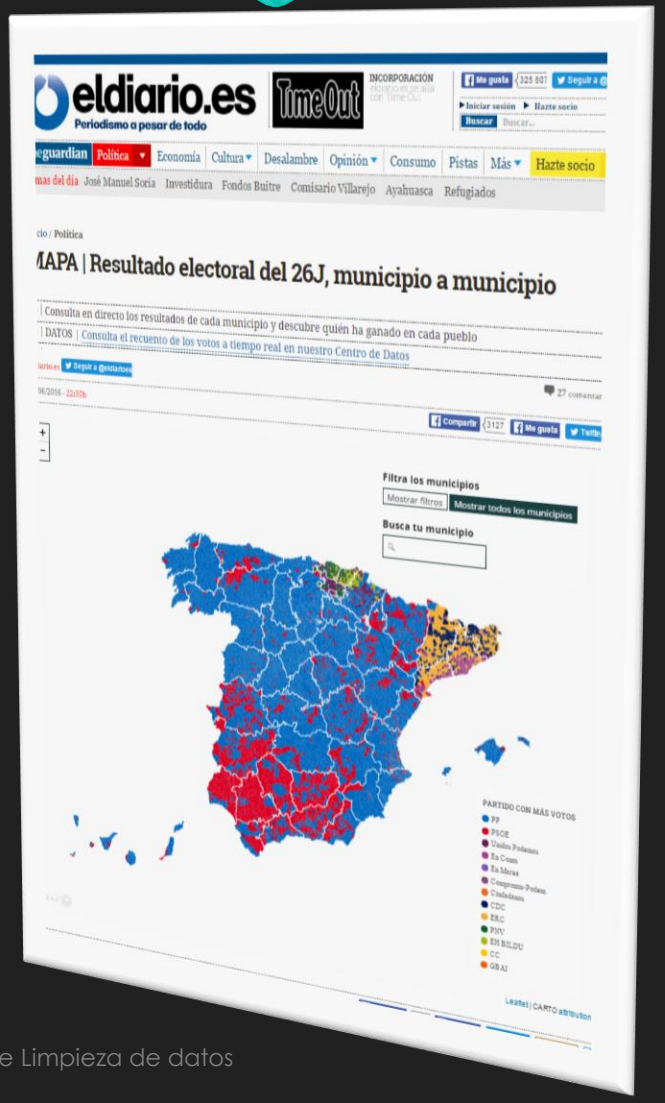
Empeoramiento de servicios públicos: sanidad, escuelas, pérdida de bienestar,

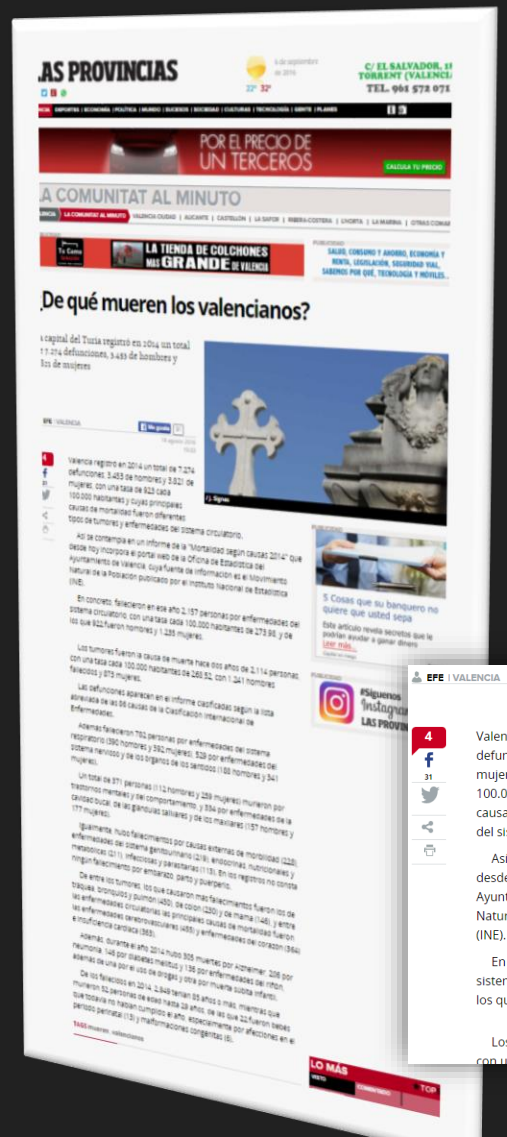
Pérdida de dinero, bancarrota

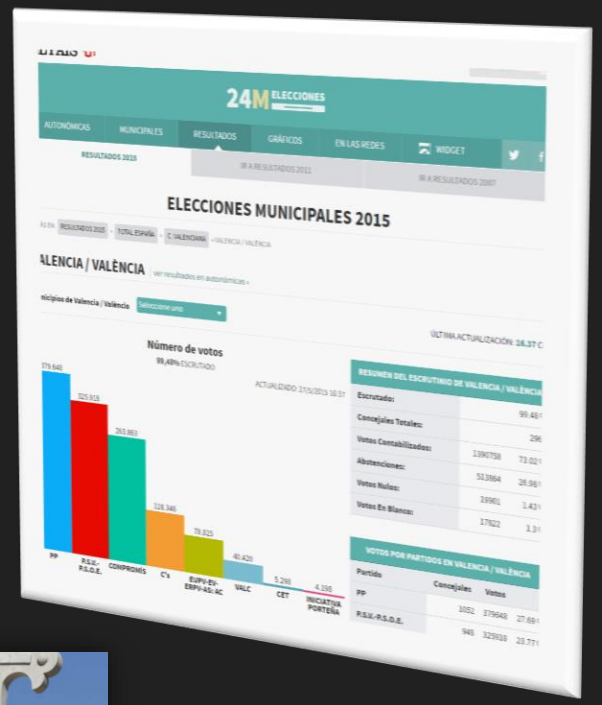
Pérdida de credibilidad

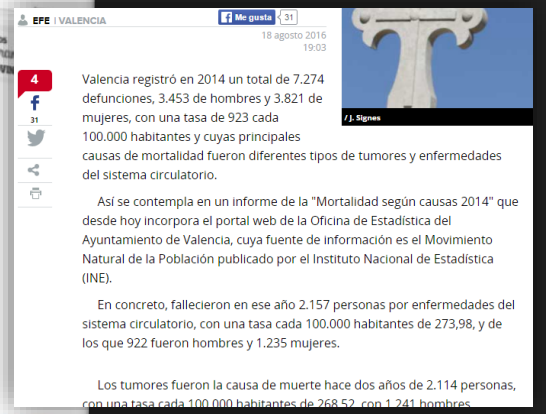
Taller de Limpieza de datos

Introducción a la limpieza de datos > **La visualización de datos**









Taller de Limpieza de datos

Introducción a la limpieza de datos > **Etapa inicial de la visualiz. datos**

- Limpieza de datos: **fase INICIAL** de un proyecto de visualización de datos tras la recogida (sensores, encuestas, entrevistas, …) y procesado de datos (en tablas)

- **VISUALIZACIÓN DE DATOS** (*data visualization*) es el proceso de búsqueda, interpretación, contrastación y comparación de datos que permite un conocimiento en profundidad y detalle de los mismos de tal forma que se transformen en información comprensible para el usuario (*Wikipedia*)
  - ➔ Comunicar información (normalmente de forma visual) vía gráficos estadísticos, tablas, gráficos, mapas o infografías para ayudar la usuarios a analizar y tomar decisiones

Taller de Limpieza de datos

Introducción a la limpieza de datos > **Etapas de la Visualización de datos**

Recoger, recopilar y procesar → **Verificación / Interpretación** → **LIMPIEZA de datos** → **Explotación y Análisis** → **Comunicación (Visualización y narración)**

Taller de Limpieza de datos

Introducción a la limpieza de datos > **Etapas de la Visualización de datos**

1. **Recoger, recopilar y procesar**: Búsqueda y recogida datos (encuestas, entrevistas, sensores, descargas web, APIs, …) y convertirlos por ejemplo en tablas

2. **Verificación / Interpretación**: Contextualizar datos / investigación (de dónde, son fiables, verificar descripción datos y categorías)

3. **LIMPIEZA de datos,** errores y **adecuación** archivos para su explotación y elaboración de tablas / gráficos / mapas finales:

4. **Explotación y Análisis** de datos

5. **Comunicación (Visualización y narración)**: Explotación datos, conversión a gráficos o mapas, infografías (visualización) y narración de historias / personalizar / humanizar ...

Taller de Limpieza de datos

Introducción a la limpieza de datos > **La limpieza de datos**

Los científicos de datos dedican al menos el 60% de su tiempo a la limpieza y organización de datos.

La recogida de conjuntos de datos viene en segundo lugar en el 19% de su tiempo, lo que significa que los científicos de datos pasan alrededor del 80% de su tiempo en la preparación y gestión de datos para el análisis.

El 57% de los científicos de datos consideran que la limpieza y organización de datos son la parte menos agradable de su trabajo

Fuente: What's The Big Data?, Data Scientists Spend Most of Their Time Cleaning Data

Taller de Limpieza de datos

Introducción a la limpieza de datos > **Errores**

Imaginarse un estudio (se han ido recogiendo datos de diferentes personas y mediciones) incluso es la unión de varios estudios (en varios departamentos de una empresa, páises o en años diferentes) donde los datos son almacenados finalmente en una gran tabla de datos (muchas filas con muchas columnas y datos entre ellas).

Taller de Limpieza de datos

Introducción a la limpieza de datos > **Errores y tareas de limpieza de datos**

- **VALIDACIÓN de los datos:**
  - Ej. Validación de fechas (y si mezclamos formatos de distintos idiomas)
    - 2/5/2017 OK, 2/15/2017 ERROR ¿ y 2/5/17 ? ¿inglés: 5/21/2017 y 5/2/2017 ?
  - Ej. Validación de datos cuantitativos (años, cantidades de dinero, ….):
    - Separadores de miles y decimales diferentes en función del idioma (ES: 149.597.870,7 / EN: 149,597,870.7)
    - Cumplan un rango: ¿ Edad 213 ? ¿Horas al día que ves la TV: 28 ?
    - Cumplan unos determinados valores: "Lunes", "Martes", "Miércoles", "Jueves", "Viernes", Sábado", "Domingo"
- **DETECCIÓN de columnas, filas y celdas VACÍAS (sin valor, null, "")**

  A la hora de resumir (totales) es importante conocer si es un 0 o es una ausencia de información (no se ha recogido o se desconoce)
  - Ej, En una columna recogemos las temperaturas de los últimos 7 años y podemos encontrarnos que en una año, una población su valor está vacío.

    ¿Era 0º ? ¿Cuando haga la media de esa población o de ese año lo tengo en cuenta?
  - Ej. En una fila, al preguntar por servicios extra del hotel está vacía ¿no tiene servicios extra o simplemente no se introdujo la información?

Taller de Limpieza de datos

Introducción a la limpieza de datos > **Errores y tareas de limpieza de datos**

- **Normalización de valores**
  - Inconsistencias: Representación mismo valor de diferentes formas (o idiomas):
    - ¿Qué país visitaste este verano? "USA", "EE.UU.", "Estados Unidos", "United States", "Estados Unidos de América"
    - Comunidad Autónoma: "Rioja, La" "La Rioja" "Rioja (La)"
    - Sexo: / Hombre, H, Varón, V, Masculino o Mujer, M, Hembra, H, Femenino ¡¡¡ ¿2 H ? !!!
    - Última serie de TV que has visto? "Por 13 razones", "13 razones", "Por trece razones", "Trece razones", "13 reasons why", "Thirteen reasons why"
    - ¿Equipo de fútbol preferido? "Valencia", "VLC", "Valencia CF". "Valencia S.A.D."
  - Errores ortográficos / tipográficos
    - Tildes,: València, Valencia, VALÈNCIA, VALENCIA, València/Valencia,
    - V por b y viceversa, Álaba, Alaba, Álava, Álaba o Schwarzenegger …..
    - espacios en blanco (antes, después y varios entre palabras): "El Corte Inglés" " El Corte Inglés" "El Corte Inglés "
    - letras intercambiadas "El Corte Ignlés"
    - mayúsculas/minúsculas: "El corte inglés" "El Corte Inglés" "EL CORTE INGLÉS"
  - Mezclas de escalas (uniformar)
    - Presupuesto: 3,500,000 o 3,5 M
    - Distancias (al sol): 149 597 870 700 metros o 149.597.870,7 km o 1 ua
    - Peso: 730 grs o 0,73 kg

Taller de Limpieza de datos

Introducción a la limpieza de datos > **Errores y tareas de limpieza de datos**

- **ADECUACIÓN del formato humano al del ordenador** :
  - Ej. Datos en filas cuando se necesitan en columnas

    Para permitir luego explotar la información de forma fácil:
    ordenar, filtrar, pivotar o crear gráficos o totales de forma sencilla

Día	Encargados
Lunes	Jose, Carlos, Antonio
Martes	David, Jose, Pablo
Miércoles	Carlos, Javier, Pablo

Día	Encargado 1	Encargado 2	Encargado 3
Lunes	Jose	Carlos	Antonio
Martes	David	Jose	Pablo
Miércoles	Carlos	Javier	Pablo

Día	Encargado
Lunes	Jose
Lunes	Carlos
Lunes	Antonio
Martes	David
Martes	Jose
Martes	Pablo
Miércoles	Carlos
Miércoles	Javier
Miércoles	Pablo

Comunidad Autónoma	Población
Andalucía	
2016	8.388.107
2015	8.399.043
2014	8.402.305
Aragón	
2016	1.308.563
2015	1.317.847
2014	1.325.385
Asturias, Principado de	
2016	1.042.608
2015	1.051.229
2014	1.061.756
Balears, Illes	
2016	1.107.220
2015	1.104.479
2014	1.103.442
Canarias	
2016	2.101.924
2015	2.100.306
2014	2.104.815

Comunidad Autónoma	2016	2015	2014
Andalucía	8.388.107	8.399.043	8.402.305
Aragón	1.308.563	1.317.847	1.325.385
Asturias, Principado de	1.042.608	1.051.229	1.061.756
Balears, Illes	1.107.220	1.104.479	1.103.442
Canarias	2.101.924	2.100.306	2.104.815

Comunidad Autónoma	Año	Población
Andalucía	2016	8.388.107
Andalucía	2015	8.399.043
Andalucía	2014	8.402.305
Aragón	2016	1.308.563
Aragón	2015	1.317.847
Aragón	2014	1.325.385
Asturias, Principado de	2016	1.042.608
Asturias, Principado de	2015	1.051.229
Asturias, Principado de	2014	1.061.756
Balears, Illes	2016	1.107.220
Balears, Illes	2015	1.104.479
Balears, Illes	2014	1.103.442
Canarias	2016	2.101.924
Canarias	2015	2.100.306
Canarias	2014	2.104.815

  - Varios tipos de datos en una columna ➔ Dividirlos en varias columnas

Equipo	Títulos
Valencia	6 Ligas, 7 Copas, 1 Supercopa de España, 3 Copas de la Uefa, 2 Supercopas de Europa, 1 Recopa de Europa

Equipo	Ligas	Copas	Supercopas Espña	Copas UEFA	Supercopas Europa	Recopas Europa
Valencia	6	7	1	3	2	1

  - Eliminación de información redundante que no aporta información
    - Eliminación de columnas que no se utilicen o de información derivada
    - Eliminación de registros (filas) duplicadas (recogidos 2 veces o al juntar información de varias fuentes) o registros sumatorios incluidos previamente

Taller de Limpieza de datos

Herramienta OpenRefine

Una herramienta gratuita, de código abierto para trabajar con datos desarreglados/desorganizados

LIMPIEZA DE DATOS

Taller de Limpieza de datos

Herramienta OpenRefine

Web; http://openrefine.org/

- Anteriormente denominado Google Refine (oct, 2012)
- Actual versión *(3 marzo 2017): OpenRefine 2.7-rc2 Release Candidate 2*
- Programa en Java que se ejecuta en local en un navegador (no hace falta conexión).
  Una vez ejecutado, sino se abre automáticamente en tu navegador preferido poner la dirección http://127.0.0.1:3333/
- Tipos de archivos: TSV, CSV, *SV, Excel (.xls, xlsx), JSON, XML, RDF –XML, Google documentos.

Taller de Limpieza de datos

Herramienta OpenRefine > Operaciones básicas

Herramienta *open source* para limpieza de información y su exploración mediante facetas:

- **Importar** información en varios formatos
- **Explorar** información fácilmente mediante **facetas** para filtrar de forma simple o múltiple y ordenar columnas
- **LIMPIAR DATOS**
  - **Eliminar / Añadir / Ordenar / Renombrar columnas**
  - **Vaiidar datos y Normalizar datos**
    - **Convertir** tipos de datos a Fecha / Numérico / Texto
    - Aplicar **transformaciones** a celdas: quitar espacios al principio o al final, dos espacios consecutivos, cambiar min/mayúsculas, convertir a número o texto o fecha, ….
    - **Dividir** columnas en varias o **transponer** (poner columnas en filas)
    - **Clustering o agrupaciones**: Encontrar automáticamente agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa y unirlas en un único valor o comprobar duplicados (normalizar).
  - **Exportar** archivos final o el proyecto entero o solo los cambios realizados en un proyecto y exportarlos en formato JSON en archivo txt para luego aplicarlos a otros archivos iguales.

Limpieza de datos
Herramienta OpenRefine > Avanzado

AVANZADO

- Posee un lenguaje para la elaboración de facetas o transformación de columnas:

  GREL: General Refine Expression Language (también se puede utilizar Python o Jython)

  https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language

  Valores: value / cells["Nombre Provincia"].value

  Control: If(value>10, "Mayor de 10", "Menor o igual a 10") > < == !=

  Condiciones: and(a,b, ...) / or(a,b,...) / not(a)

  Cadenas:

  length(s) / startsWith(s, b) / toUppercase(s) / toLowercase(s) / trim(s) /

  substring(s, n1, n2) / indexOf(s, s2) / split(s, separ)

- Permite el uso de expresiones regulares (utilizando el lenguaje "Java's regular expresión" por ejemplo, para filtrar en el texto

Limpieza de datos

Herramienta OpenRefine > Clustering (avanzado)

Fingerprint: métodos por defecto de clustering en Open refine

- Método rápid y simple. Funciona relativamente bien en variedad de contextos y es el que produce menos falsos positivos (es el método por defecto
- El proceso que sigue (el orden de las operacioneses relevante):
  - 1- Elimina espacios en blanco en cabecera y cola
  - 2- Cambia todos los caracteres a minúsculas
  - 3- Elimina todos los caracteres de puntuación y de control (puntos, comas)
  - 4- Divide la cadena en piezas o elementos separadas por espacios en blanco
  - 5- Ordena las piezas y elimina duplicados
  - 6- Une las piezas de nuevo
  - 7- Normaliza caracteres a su representación ASCII (por ejemplo "gödel" → "godel")

    Cuidado que "godél" también acabaría en godel y podría ser un falso positivo. Este método puede ser menos efectivo cuando se tienen datos con caracteres extendidos

Taller de Limpieza de datos

Herramienta OpenRefine > EJERCICIOS

EJERCICIO 1

En el Portal de datos abiertos de la GVA buscar, visualizar y descargar la distribución CSV del conjunto de datos:
“Datos mortalidad - Municipios - Causas CV- 2013”

Importarlos a Open Refine y limpiar y explorar los datos para contestar a las preguntas que se indiquen.

Taller de Limpieza de datos

Herramienta OpenRefine > EJERCICIOS

EJERCICIO 2

Importar el archivo "datos-mortalidad-provincias-causas-cie10-2007-SUCIO" que se puede descargar del siguiente enlace

https://goo.gl/RpUMiV

Realizar limpieza de archivo

- Importar el archivo "datos-mortalidad-provincias-causas-cie10-2007-SUCIO"
- Quitar espacios de CAUSA
- Mediante Cluster normalizar VALÈNCIA y CASTELLÓN ir cambiando sus parámetros
- Igual con CAUSA (probar con distintas funciones y tener cuidado
- Mediante Edit en Faceta normalizar el GENERO y Edit en columnas

Taller de Limpieza de datos

Herramienta OpenRefine > EJERCICIOS

EJERCICIO 3: Avanzado: adecuación de las filas y columnas (transponer)

Tenemos una base de datos con la población de cada provincial desde 2001 a 2009. Cada registro contiene:

- el codigo de la provincial,
- nombre provincial,
- año
- población (de dicha provincia en ese año)

Queremos ampliarlo con los datos de 2010 a 2016 aportados por el INE en su portal de datos

DATOS: Buscar en Demografía y población > Padrón. Población por municipios > Cifras oficiales de población de los municipios españoles > Resumen por provincias > Población por provincias y sexo (elegir todas las provincias sin el Total Nacional, Sexo Total, 2016 a 2010)

ACCIONES:

- Importarlos a Open Refine:
- Separar la columna de las Provincias en Código y NombreProvincia con función substring(value,inicio,fin)
- Transponer en 2 columnas (Key:Año, Value:Población)
- Sino se ha rellenado automáticamente "Fill Down" la columna NombreProvincia
- Seleccionar mediante Faceta los datos de 2016 solo y **exportar a Excel** o csv

Taller de Limpieza de datos

Herramienta OpenRefine > EJERCICIOS

EJERCICIO 4

Importar el archivo “Estadísticas alumnos” que se puede descargar del siguiente enlace

https://goo.gl/fGMGRq

Realizar limpieza de archivo en especial normalizar los valores

Intentar encontrar valores anómalos (altos en las 2 últimas columnas) con técnicas como la ordenación, faceta numérica o faceta custom con el logaritmo del valor

Google Fusion Tables
(tablas dinámicas de Google)

Una herramienta de Google para
obtener, visualizar y compartir tablas de datos

Limpieza de datos
Google Fusion Tables

Aplicación experimental de Google para obtener, visualizar y compartir tablas de datos. Permite:

- Crear o importar tablas.
- Crear (pestañas)
  - Tablas/vistas: Editar, explorar, filtrar
  - Tarjetas (cards): Visualizar información en tarjetas (cards):
  - Mapas (es possible geolocalizar automáticamente o indicar la columna).
    - Permite configurar estilos de puntos (iconos único, basado en un columna, por intervalos), polígonos, líneas
    - Configurar leyenda o ventana de información .
  - Resúmenes
  - Gráficos
    - De dispersion (scatter)
    - Líneas
    - Áreas
    - Columnas y barras (histogramas)
    - De sectores (tarta)
    - Red

Limpieza de datos
Google Fusion Tables

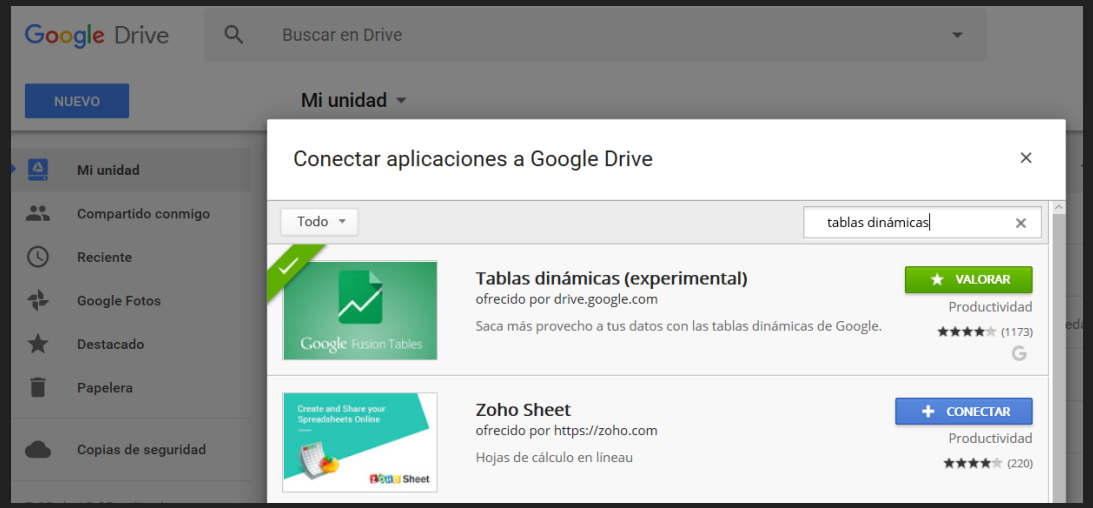
Para ejecutar (abrir una tabla dinámica de Google)

- Desde la URL: https://support.google.com/fusiontables/answer/2571232?hl=en

  Buscar en Google "Fusion tables"

- En Google Drive > Nuevo > Más > TABLAS DINÁMICAS DE GOOGLE

  Nota: Puede ser necesario la primera vez hacer clic en [conectar aplicaciones] y buscar la aplicación "tablas dinámicas"



Taller de Limpieza de datos

Google Fusion Tables > EJERCICIOS

EJERCICIOS TABLE FUSION TABLES

Representa los siguientes conjunto de datos en gráficos y mapas

Aparcabicis de Valencia (datos abiertos Ayto Valencia)

Acoso escolar (descargar https://goo.gl/DQfCvY) y haz una 2ª representación con la tabla de comunidades https://goo.gl/CiTtDX (hacer una copia iniciado sesión)

Población por provincias en el 2016 (https://goo.gl/wR9R87)

(Nota KML de provincias (https://goo.gl/frdiwl)

Taller de Limpieza de datos

FIN