PEC 1 (20% nota final)

Orientaciones en la solución de la PEC 1

Ejercicio 1 [70%]

Después de leer el recurso “Calvo, M., Pérez, D., Subirats, L. (2019). Introducción al ciclo de vida de los datos.” contesta las siguientes preguntas con tus propias palabras:

1 ¿Qué perfil profesional relacionado con la ciencia de datos te gustaría ser? ¿Y cuál te gustaría menos ser? Razona la respuesta (máximo 200 palabras). [5%]

Me gustaría ser científico/a de datos puesto que me gusta limpiar y organizar los datos, soy una persona curiosa, y conozco los lenguajes R, Python, Matlab y Spark. El que menos me gustaría ser es analista de negocio porque no me parecen tan interesantes los procesos de negocio.

2 Lista los diferentes factores que influyen en la calidad de los datos y pon un ejemplo diferente al que se explica en los materiales (máximo 300 palabras). [30%]

- Exactitud: Un economista introduce los datos con puntos indicando miles, pero va a Estados Unidos y se interpreta como un decimal.
- Completitud: Un vendedor de casas introduce los datos que describen la casa que quiere vender, sin dejar ningún valor en blanco.
- Consistencia: Si se mide el índex de demasiado corporal (IMC), tiene que estar en consonancia con pes/levantada^2.
- Puntualidad: Diferencia entre la fecha de un examen, el resultado de la nota, y la introducción al expediente académico.
- Unicidad: En un examen de una asignatura, se comprueba que no hay una misma persona con dos notas.
- Validez: En una base de datos hay un valor de una persona que pesa 1000 kg. Este dato no se válida porque se encuentra fuera de rango, una persona no puede pesar 1000 kg.

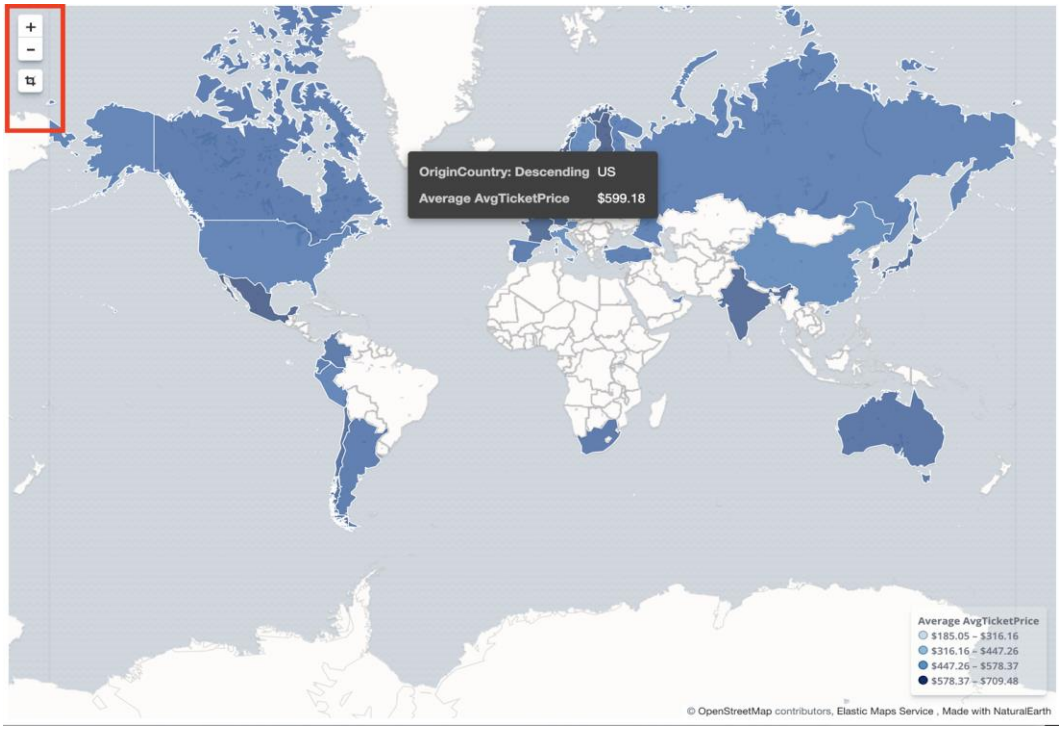
3 ¿En qué tipos de base de datos no es necesario conocer a priori los datos que se quieren almacenar? Pon tres ejemplos de tecnologías que utilicen estas bases de datos (máximo 100 palabras). [10%]

Son las bases de datos no relacionales. Tres ejemplos de estas bases de datos son: MongoDB, Redis y Cassandra.

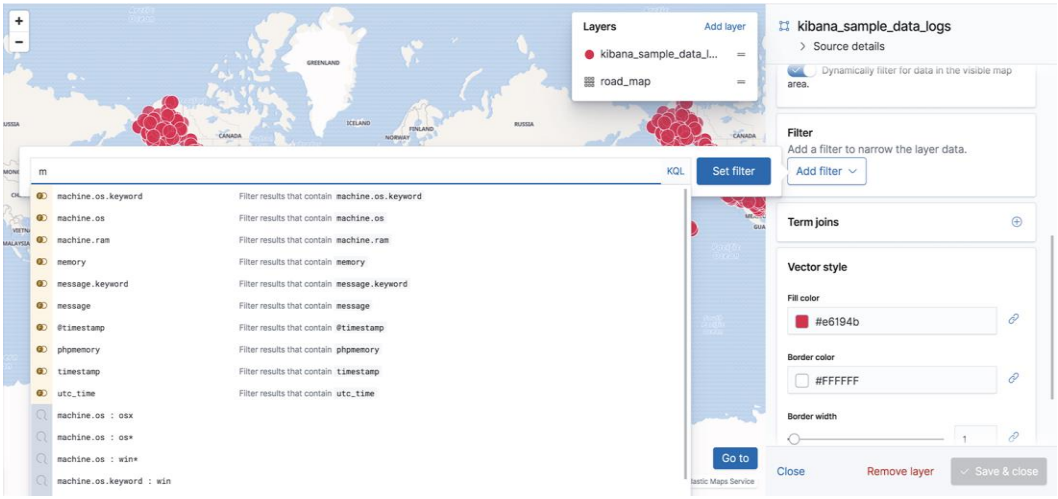
4 Pon ejemplos visuales con imágenes de las 7 visualizaciones que permiten un nivel más alto de abstracción (adjuntar las 7 imágenes). [25%]

Panorama general:

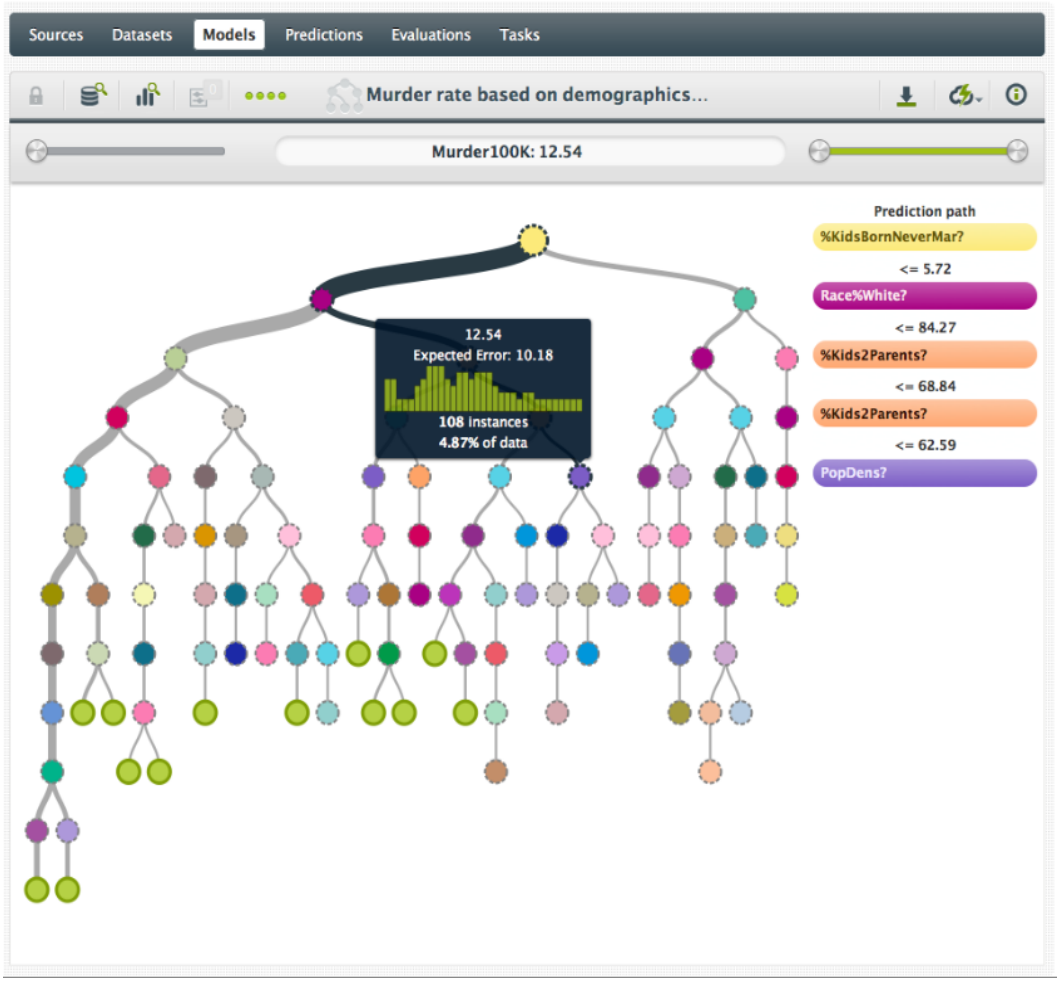
Acercamiento:



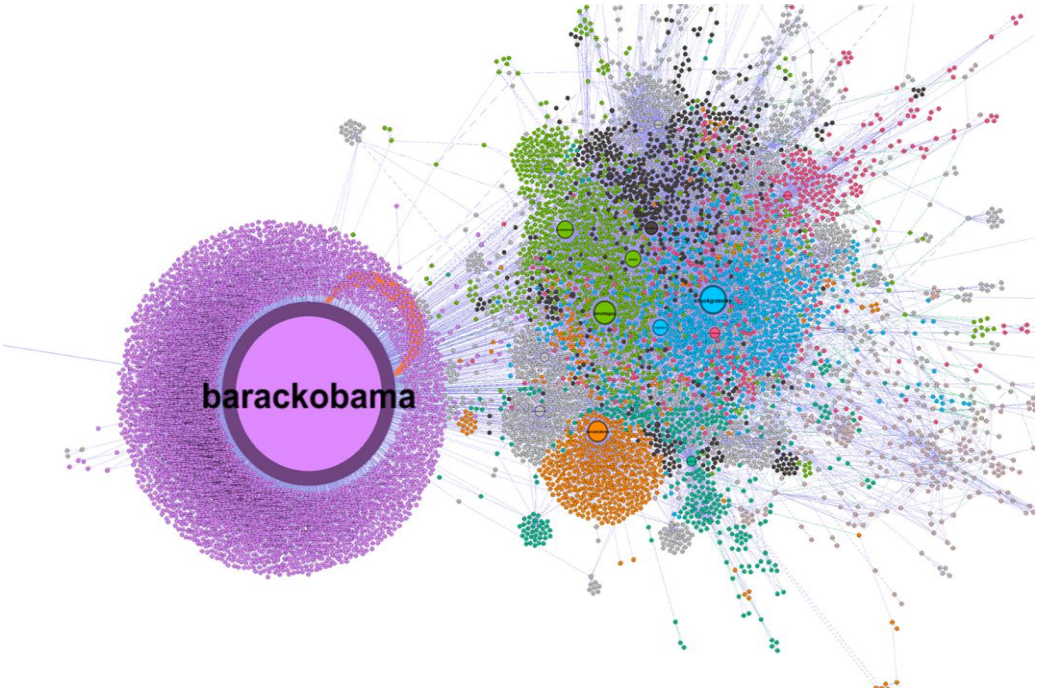
Filtraje:



Detalles a petición:

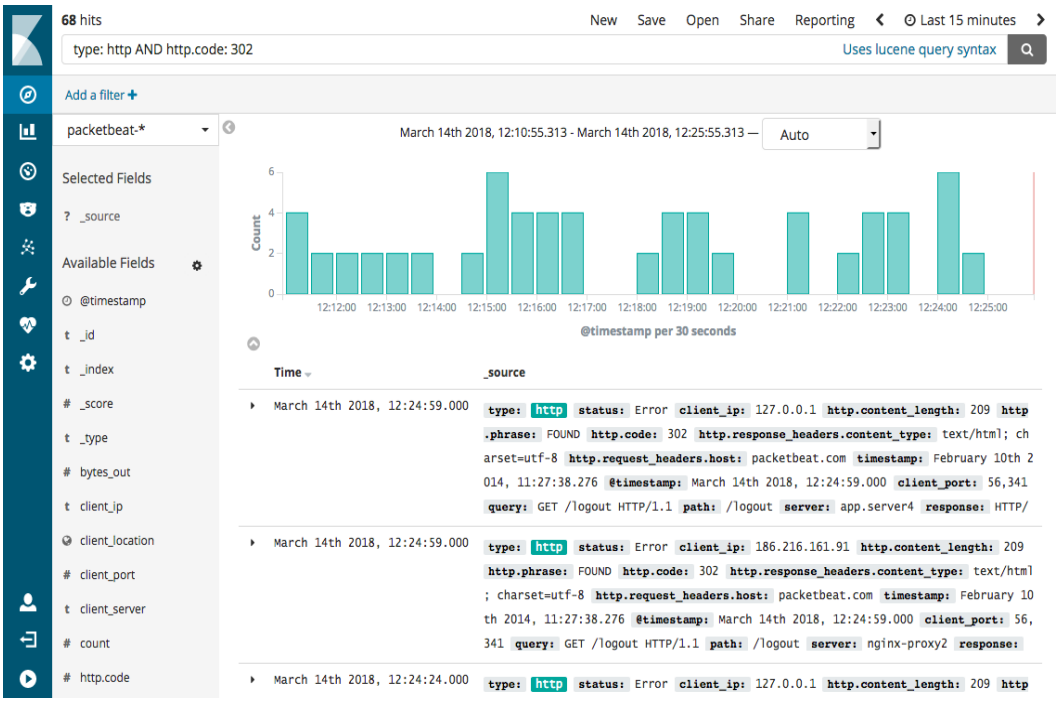


Relaciones:



Historial:

Extracción:



Ejercicio 2 [30%]

Después de leer el recurso “Subirats, L., Calvo, M. (2019). Web Scraping.”, capítulos 1 y 7. Contesta las siguientes preguntas con tus propias palabras:

1 ¿Por qué crees que es necesario hacer web scraping)? (máximo 100 palabras). [10%]

El web scraping es necesario cuando no se dispone de una API o de un repositorio de datos, o cuando estos: no son gratuitos, limitan el número de acceso por segundo/día, etc., y/o el API no permite extraer toda la información de interés.

2 ¿Por qué es importante analizar el contenido del archivo *robots.txt*? ¿Qué riesgo corremos si no lo hacemos? (máximo 100 palabras). [10%]

Porque es en este archivo donde la mayor parte de páginas web indican las restricciones a tener en cuenta cuando se pretende rastrearlas.
Aunque estas restricciones son solo una sugerencia y nunca una obligación, es recomendable tenerlas en cuenta para reducir las posibilidades de ser bloqueados y/o evitar problemas legales futuros.

3 Explica como evitarías saturar el servidor con peticiones web (máximo 100 palabras). [10%]

Se puede evitar colocando retrasos definidos entre las peticiones. De este modo, no se sobrecarga el servidor y no se provocan alarmas que resulten en un bloqueo.

Criterios de valoración

La ponderación de los ejercicios es la siguiente:

- Ejercicio 1.1: 5%
- Ejercicio 1.2: 30%
- Ejercicio 1.3: 10%
- Ejercicio 1.4: 25%
- Ejercicio 2.1: 10%
- Ejercicio 2.2: 10%
- Ejercicio 2.3: 10%

Se evaluará la precisión de los ejemplos, así como el respeto al número de palabras máximo establecido para cada pregunta. La idoneidad y claridad de las respuestas también será evaluada.