MEMORIA

Estudio sobre los dinosaurios

Análisis Exploratorio de Datos

by Marta Buesa Suárez de Puga

Enero 2022

Objetivo

Desarrollo de un **estudio de dinosaurios basado en un Análisis Exploratorio de Datos** para comprobar si las hipótesis de partida se cumplen o no.

Para ello, se crea un contexto inicial simulado, con un supuesto realista que podría acontecer, donde un Museo de Ciencias Naturales ve la necesidad de desarrollar este estudio, de cara a dar a conocer a los posibles inversores información clave sobre los dinosaurios que se han encontrado por todo el mundo y así descubrir el atractivo de los hallazgos paleontólogos con su consecuente convicción en invertir y apoyar al Museo para seguir desarrollado su labor investigadora también fuera del Museo.

Recursos utilizados

1. Lenguaje de programación → **Python 3.7.4.**

2. Librerias:

- **Numpy:** especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos.
- **Pandas**: especializada en el manejo y análisis de estructuras de datos.
- **Matplotlib**: especializada en la creación de gráficos.
- **Seaborn:** especializada en la creación de gráficos basada en matplotlib pero con una interfaz evolucionada que permite generar fácilmente elegantes gráficos.
- **Selenium**: entorno de pruebas de software para aplicaciones basadas en la web, que permite extraer la información necesaria.

3. Herramienta de **Business Intelligence: Tableau** para la visualización de datos interactivos, creando un discurso/historia con el análisis desarrollado.

4. **Jupiter Notebooks** con **Visual Studio Code.**

5. Power Point programa de presentación para explicar el detalle del estudio desarrollado

BBDD: Datos de referencia

Web 1: Natural History Museum, además de tener una amplia sección de dinosaurios con post y videos con información muy relevante y explicada en muchos casos por su plantilla de paleontólogos, disponen de un directorio de dinosaurios ordenados alfabéticamente encontrados por todo el mundo.

Contenido – archivo CSV
309 dinosaurios en 10 columnas descritas a continuación:
1. Name - nombre del dinosaurio
2. Diet (herbivorous/ carnivorous/ omnivorous) - alimentación
3. Period (nombre del periodo y años comprendidos) - periodo en el que existió el dinosaurio
4. Lived_in - localización donde vivía el dinosaurio
5. Type - tipo de dinosaurio
6. Length (m) - longitud en metros del dinosaurio
7. Taxonomy - taxonomía del dinosaurio
8. Named_by - personas que dieron nombre al dinosaurio
9. Species - la especiae a la que pertenecía el dinosaurio
10. Link - website contenedora de la información del dinosaurio

Web 2: Prehistoria Fandom, con muchas páginas con información relevante de dinosaurios. Disponen de un extenso directorio de dinosaurios con sus principales características.

Extracción a través de web scrapping – archivo CSV
1193 registros en 3 columnas
1. Nombre – nombre del dinosaurio
2. Altura – altura en formato texto
3. Peso – peso en formato texto

Web 3: ABCDino, explorador de dinosaurios con información complementaria.

Extracción a través de web scrapping – archivo CSV
309 registros en 3 columnas
1. Nombre – nombre del dinosaurio
2. Altura – altura en formato texto
3. Peso – peso en formato texto

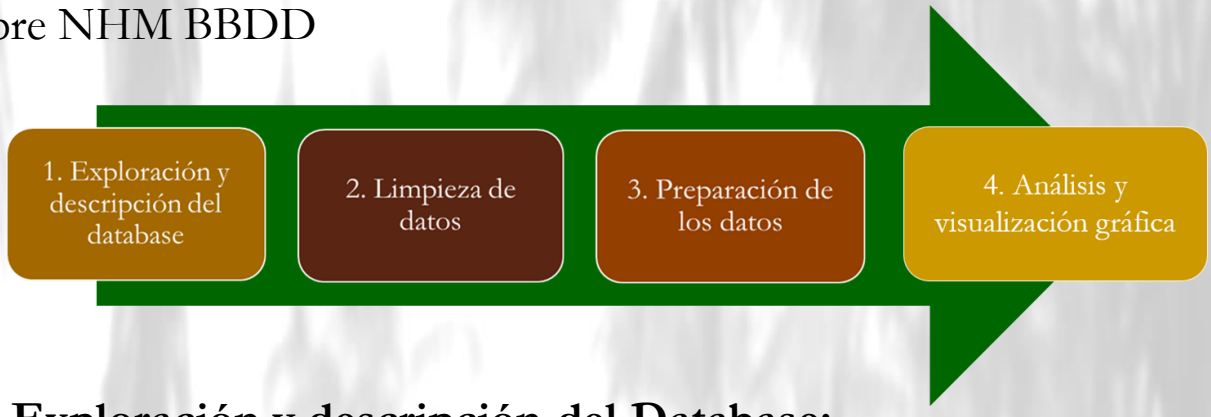
Web 4: ThePaleobiology Database, web desarrollada por paleontólogos, donde existe un exhaustivo detalle de los fósiles encontrados por todo el mundo de distintas criaturas, fauna, flora... Contiene una API para la extracción de BBDD según los filtros que uno desea.

Extracción de 2CSV con información a priori relevante.
Al analizar las extracciones, se detecta la imposibilidad de cruzar estas BBDD con las anteriores, dado que no contienen campo de cruce como podría ser el nombre estandar del dinosaurio.

Status Estudio (día 16 enero de 2022)

ESTUDIO: **100%**

Sobre NHM BBDD



1. **Exploración y descripción del Database:**
   - Observo data obtenido para familiarizarme con su contenido y poder detectar posibles problemas.

2. **Limpieza de datos:**

Tratamiento de columnas:

- Desagregación de contenido en varios columnas.
- Detección de valores nulos y tratamiento de los mismos.
- Eliminación de valores no necesarios de ciertas columnas.
- Transformación de tipo de dato en columna que requiere variable numérica.

3. **Preparación de los datos:**
   - Creación de nueva columna "continente" en función de la información de otra existente "país".
   - Filtrado por columnas que interesa quedarse para el análisis.

4. **Análisis y visualización gráfica**:
   - Desarrollo de diferentes gráficas apoyándome en librerías Matplotlib y Seaborn, y atendiendo a las diferentes variables del database.
   - En paralelo, trabajo con herramienta de Business Intelligence, Tableau, para la creación de tablas con gráficos relevantes, dashboards con visualizaciones clave, y creación de "History".

Status Estudio (día 16 enero de 2022)

COMPLEMENTO 1: 80%

Web Scrapping Preshitoria Fandom. 1193 registros.

PASO 1:

- Localización de elementos que contienen la información buscada en esta web, 'Altura' y 'Peso' para realizar el scrapping de la web a través de librería Selenium.
- Programación para la extracción de data.
- Obtención de los datos y creación de archivo csv.

PASO 2:

- Combinar CSV Fandom con BBDD origen de NHM.
- Columnas que complementan el database: 'Altura' y 'Peso'
- Se detecta que la información extraído no es homogénea, y es difícilmente extraer patrones.
- Se podría trabajar con REGEX.
- Se decide localizar datos a través de nueva fuente.

COMPLEMENTO 2: 70%

Web Scrapping ABCDino. 309 registros.

PASO 1: Igual que en segunda Fase, obtención de data y creación de archivo csv ¡con éxito!

PASO 2:

- Combinar CSV ABCDino con BBDD origen de NHM.
- Columnas que complementan el database: 'Altura' y 'Peso'.
- Paso a analizar y visualizar los datos obtenidos con el cruce.
- Sin embargo , detecto valores faltantes y decido cruzar extracción de dinosaurios con BBDD Fandom.

_PENDIENTE: Terminar el cruce para poder hacer análisis completo.

Agradecimientos

Este dataset en origen está escrapeado de la website "National History Museum" (https://www.nhm.ac.uk) donde llevan a cabo una gran labor, no sólo la física en el propio museo y fuera de el, sino también de desarrollo de contenidos según los hallazgos y características de los dinosaurios descubiertos por todo el mundo. Sin su gran compromiso y aportación a esta gran labor de conocimiento de los dinosaurios no sería posible poder obtener información tan valiosa y educativa para todos los interesados en la materia. Especialmente agradecida al profesor Paul Barrett por compartir su extenso conocimiento y experiencia, que enorme suerte poder trabajar día a día con una colección tan representativa de los dinosaurios y abordando nuevos proyectos por todo el mundo(https://www.nhm.ac.uk/discover/dinosaur-world-tour.html).

También a Alastair Hendry of #NHM_Live, gracias por esas estupendas entrevistas, donde se entran a comentar tantos detalles súper interesantes, nunca perdiendo de vista nuestra referencia de Jurassic Park, están fenomenal.

Asimismo, un agradecimiento especial a "KamranJanjua" por publicar en kaggle (https://www.kaggle.com/kjanjua/jurassic-park-the-exhaustive-dinosaur-dataset) este documento csv, apostando por ofrecer información de los dinosaurios al público en general.

Fuentes de información

Webs:

https://www.nhm.ac.uk/discover/dino-directory/name/name-az-all.html

https://www.kaggle.com/kjanjua/jurassic-park-the-exhaustive-dinosaur-dataset

https://public.tableau.com/app/profile/marta7901/viz/Dino_way/Historia1?publish=yes

https://prehistoria.fandom.com/es/wiki/Categor%C3%ADa:Dinosauria

https://abcdino.com/explorar-dinosaurios-a-al-z/

https://es.wikipedia.org/wiki/E%C3%B3n_fanerozoico

https://dinosaurioss.com

https://es.wikipedia.org/wiki/Dinosauria

https://www.expertoanimal.com/tipos-de-dinosaurios-marinos-nombres-y-fotos-24753.html

https://es.wikipedia.org/wiki/Tax%C3%B3n

https://www.renfe.com/es/ca/grup-renfe/grup-renfe/flota-de-trens/avlo

https://paleobiodb.org/navigator/

Libros:

- ✓ My Encyclopedia of very important dinosaurs DK division of Penguin Random House
- ✓ Dinosaurios Asombrosos Todolibro