Web scraping

Para economistas

Alejandro Acosta León

1 de junio de 2024

Economía | udla

Expectativas

¿Qué esperas aprender en este taller?

Escríbanlo en el siguiente Jamboard

Economía | udla

Teoría

Conceptos

- ¿Qué es web scraping?
- Introducción al web scraping y sus aplicaciones.
- Problemas éticos, legales y mal uso del web scraping.
- Alternativas al web scraping.
- Herramientas y librerías necesarias.

Entorno

- Instalación de Python y librerías necesarias (requests, BeautifulSoup, pandas, selenium).
- Configuración del entorno de desarrollo (puede ser Jupyter Notebook, VSCode, etc.).

Economía

Práctica

Nivel 1: data estructurada en HTML

- pandas

Nivel 2: data semi-estructurada

- Inspección de elementos HTML.
- Extracción de datos con BeautifulSoup.
- Descarga de archivos.

Nivel 3: data no estructurada que no usa HTML

- HTML y CSS.
- Inspección de elementos.
- Extracción de datos con BeautifulSoup.

Economía | udla

Algunos recursos importantes

- Material del curso:
  `https://github.com/alejo-acosta/favorita-powerbi/`
- Descarga Power BI:
  `https://powerbi.microsoft.com/en-us/downloads/`
- Dataset original:
  `https://www.kaggle.com/c/favorita-grocery-sales-forecasting/`
- Documentación oficial de Power BI:
  `https://docs.microsoft.com/en-us/power-bi/`
- Lenguaje DAX (Power BI):
  `https://docs.microsoft.com/en-us/dax/`
- Lenguaje M (Power Query):
  `https://docs.microsoft.com/en-us/powerquery-m/`

Economía | udla

¿Qué es Power BI?

- Power BI suele definirse como: Herramienta corporativa de autoservicio para hacer business intelligence.
- BI son herramientas que nos permiten extraer, conectar y visualizar datos.

https://www.gartner.com/doc/reprints?id=1-24ZXJ0MU&ct=210107&st=sb

Economía udla

¿Cuándo no usar Power BI?

Power BI **no** es una herramienta de análisis o modelamiento de datos. Python, R, Stata, julia, etc. tienen una ventaja enorme en este campo. **El secreto es usar la herramienta adecuada para nuestro flujo de trabajo.**

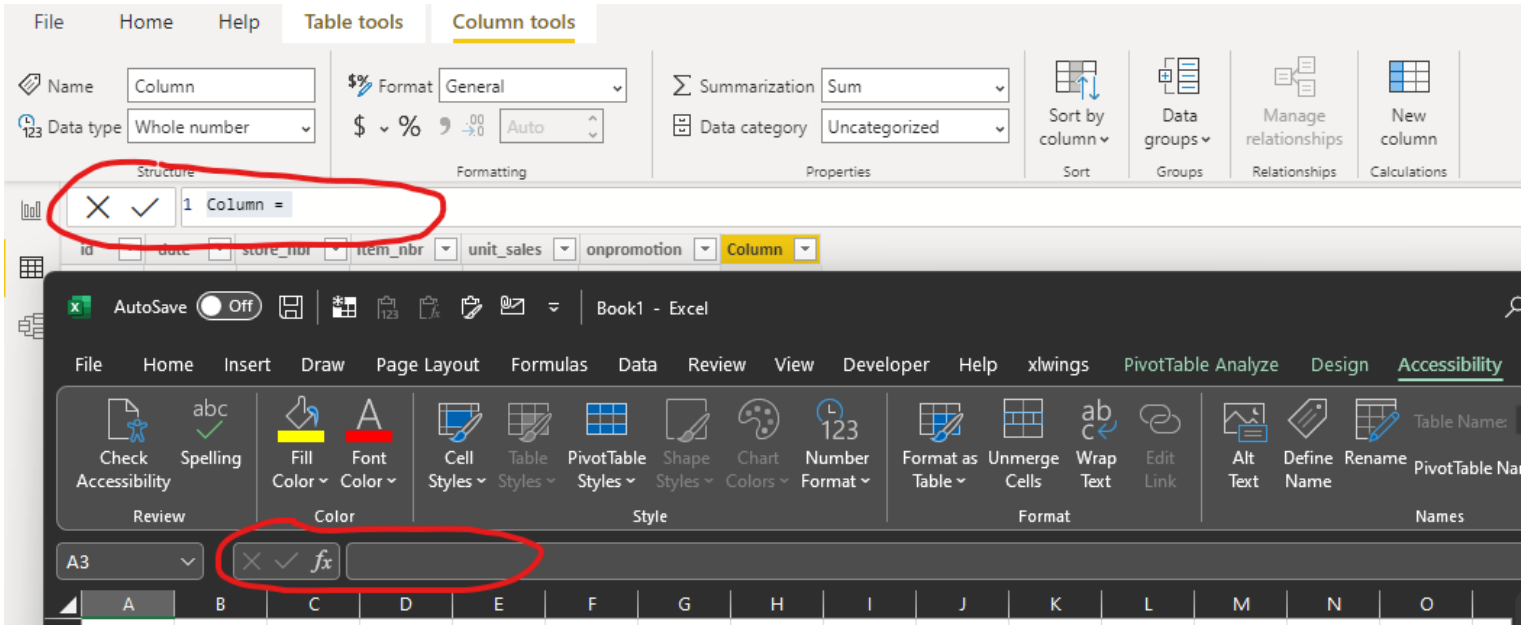
Por ejemplo:

Extracción y manipulación datos	Análisis Exploratorio	Gráficos y reportes	Análisis Estadístico	Modelos
1.Python 2.Excel 3.Power BI 4.Stata	1.Python 2.Power BI 3.Excel 4.Stata	1.Python 2.Power BI 3.Excel	1.Python 2.Stata	1.Stata 2.R 3.Python

Economía | udla

Los lenguajes de Power BI

- Power BI (al igual que Excel) tiene su propio "lenguaje de programación"llamado DAX.



- Sin embargo, también se utiliza otro lenguaje para extraer, transformar y cargar datos (ETL): M que es usado en Power Query.

Economía udla

Dataset

- Utilizaremos las ventas de Corporación la Favorita.
- Solo usaremos 1 % del total de la base.

`https://www.kaggle.com/c/favorita-grocery-sales-forecasting`

Economía | udla

Power BI

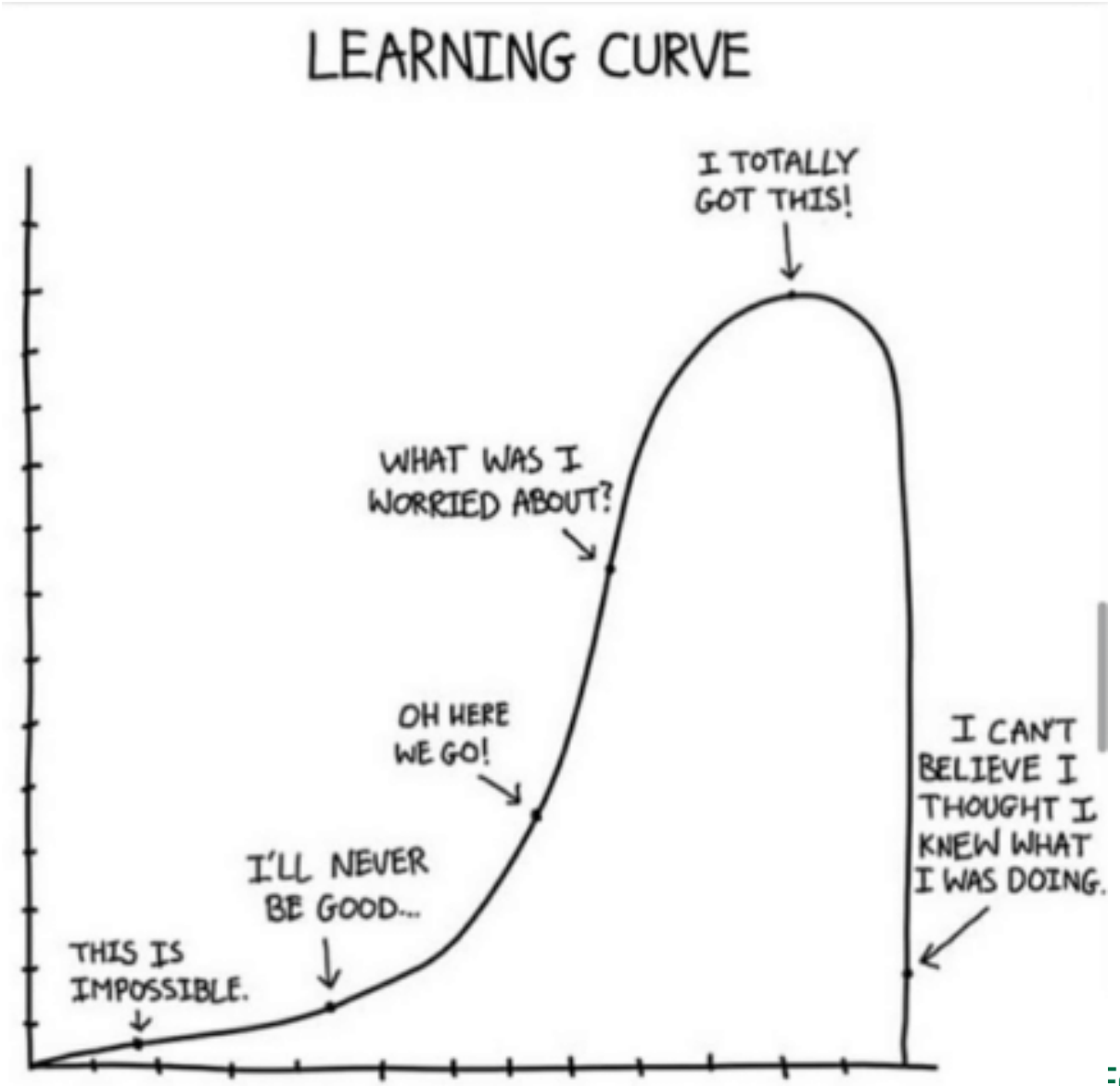
Parte II

En esta sesión revisaremos:

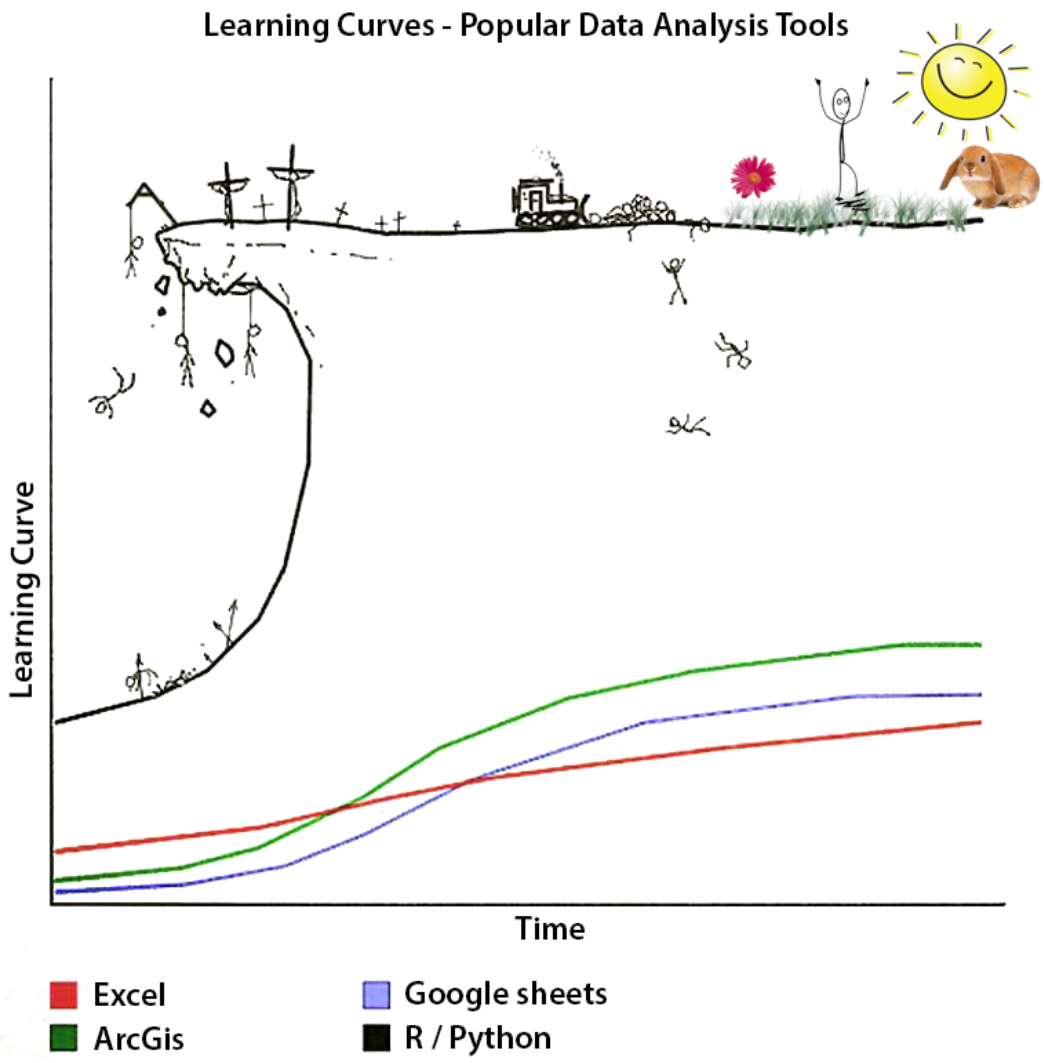
- Refuerzo de tablas 'dimensiones' y modelo relacional de datos.
- Datos long vs wide.
- Métricas.
- Merge y append.
  - Funciones pivot y unpivot.
- Series de tiempo descriptivas (no predictivas).

Economía | udla

Una pequeña motivación



Economía | udb


Learning Curves - Popular Data Analysis Tools
Learning Curve
Time
Excel
ArcGis
Google sheets
R / Python