Práctica 1

Tipología y ciclo de vida de los datos

Componentes del equipo:

- José Enrique Atiénzar Ibáñez
- Diego Argüelles García

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

La empresa Cárnica SL es una empresa de consultoría española que ofrece sus servicios a clientes de todo el mundo. Sus trabajadores son principalmente consultores *juniors* recientemente titulados con un rango de edad entre los 22 y los 27 años de edad. Estos consultores son enviados rutinariamente a los clientes, muchos de ellos en el extranjero para auditar, aconsejar y mejorar los procesos internos de los mismos. Durante las estancias en clientes, Cárnica SL ofrece a sus trabajadores unas dietas para compensar los gastos extra de manutención durante las estancias fuera del domicilio habitual.

El nuevo responsable del departamento de RRHH ha decidido **optimizar la gestión de dietas de los empleados** para hacer la empresa más *"lean and mean"*. Para ello ha encargado a dos becarios del departamento de informática la creación de un software que actualice periódicamente un listado de precios de consumo para todos los países del mundo. Estos precios podrán ser usados para calcular las dietas apropiadas para los empleados desplazados a clientes dependiendo del país y del perfil del empleado.

El nuevo responsable de RRHH considera que, dado que la gran mayoría de los consultores *juniors* son jóvenes menores de 27 años, y la disposición favorable de los jóvenes al consumo de hamburguesas, la empresa puede usar el precio de un big mac, una bebida y el precio del billete de transporte como base para calcular las dietas de los empleados.

(Fuera del *scope d*e este proyecto quedaría el desarrollar un proyecto de minería de datos para analizar porque el grueso de consultores *juniors* abandona la empresa en un plazo inferior a los 12 meses tras la contratación inicial).

Una vez asignado el proyecto, los dos becarios del departamento de informática identificaron un sitio web https://preciosmundi.com/ que compara precios de productos y servicios en todo el mundo y que actualiza mensualmente un índice de precios de consumo para cada país del mundo. Los becarios decidieron crear un *web scraper* que crease un fichero csv con la información demandada por RRHH. De esta forma, el *web scraper* puede ser ejecutado el primer día laborable de cada mes para obtener un archivo csv con los precios de los productos requeridos para cada país que el departamento de RRHH usará para el cálculo de las dietas de los empleados desplazados.

El archivo cvs puede ser añadido además a los repositorios de información de la empresa para recopilar información temporal sobre los datos usados para calcular las dietas.

2. Título. Definir un título conciso y que sea descriptivo para el dataset.

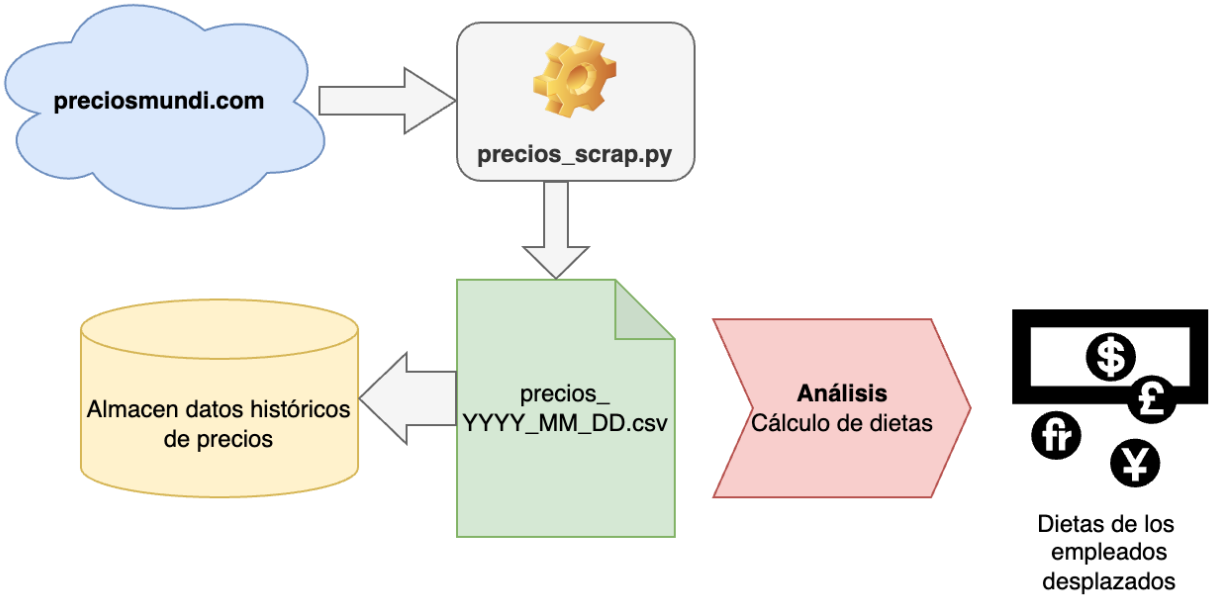
Los becarios decidieron nombrar el fichero *data set* **precios_fecha.cvs,** siendo fecha la fecha en la que se creó el documento. De esta forma se contribuiría a reducir la posibilidad de que por error el departamento de RRHH usase los datos de precios un mes erróneo para el cálculo de unas dietas.

Como título para el data set se decidió: "**Precios relevantes para el cálculo de dietas de manutención para empleados desplazados a clientes**"

3. Descripción del dataset.

El dataset contiene la información de diferentes **precios que son de interés para la empresa a la hora de calcular las dietas de sus trabajadores**. Por cada país del mundo tendrá un registro que contendrá como columna el nombre del país, el continente al que pertenece, el salario medio del país como referencia y una lista de los elementos de consumo que el departamento de RRHH de Cárnica SL está dispuesto a sufragar, básicamente menús económicos, refrescos, café, cerveza (si esta se consume fuera de la jornada laboral y en cantidad moderada), así como el precio del transporte del empleado desde el aeropuerto al hotel y a las instalaciones del cliente. Estos precios serán usados como base por el departamento de RRHH a la hora de determinar las dietas para los empleados desplazados a clientes.

4. Representación gráfica



5. Contenido del dataset

El periodo al que hacen referencia los datos es de abril de 2023. Dado que el sitio web del que hemos obtenido los datos no tiene registro histórico y los precios que hay se actualizan cada mes la idea es **lanzar el script una vez al mes y almacenarlo en un fichero distinto identificado con la fecha**. En una

fase posterior de integración de datos reflejaremos una dimensión temporal mensual para ir agregando los precios de cada mes. **Todos los precios van expresados en euros** y el **salario medio va expresado en Dólar Americano (USD)**.

Nombre del campo	Descripción	Tipo de dato	Ejemplo
País	Nombre del país en castellano	Texto	"Francia"
Continente	Nombre del continente en castellano	Texto	"europa"
Salario medio $	Salario medio del país. Formato con punto como separador de miles y coma para decimales. Expresado en Dólares.	Numérico	2.2443,23
Refresco €	Precio de un refresco expresado en Euros	Numérico	2,30
Capuccino €	Precio de un capuccino expresado en Euros	Numérico	1,52
Cerveza €	Precio de 0,5l de cerveza nacional expresado en Euros	Numérico	3,20
Menú BigMac €	Precio de un menú de Mc Donalds expresado en Euros	Numérico	6,30
Menú del día €	Precio de un menú barato (menú del día)	Numérico	12,53
Billete transporte €	Precio de un billete de ida en transporte público expresado en Euros	Numérico	1,45
Inicio Taxi €	Precio de bajada de bandera del taxi expresado en Euros	Numérico	2,50
Taxi 1Km €	Precio de taxi por 1 km expresado en Euros	Numérico	1,53

6. Propietario

El propietario del sitio web que hemos usado para el scraping de los datos es Víctor Rodríguez Obensa. Este sitio web ofrece una comparativa de precios por países que se actualiza mes a mes así como tablas de precios para cada país.

Tras revisar la sección de aviso legal de la página entendemos que no hay ninguna limitación a la hora de hacer web scraping sobre la misma, si bien para evitar sobrecargar la página, añadimos un sleep de 1

segundo entre llamadas y usamos solo un continente durante la fase de programación para limitar las llamadas al sitio web.

La página específica que “El acceso a EL SITIO WEB y el uso que sus usuarios pueden hacer de la información contenida en la misma, es responsabilidad exclusiva de dichos usuarios.” por lo que entendemos que podemos hacer uso de la información que extraemos de la página.

Otros ejemplos de páginas web similares son:

https://www.expatistan.com/

https://es.numbeo.com/

https://www.priceoftravel.com/

7. Inspiración

La idea de este proyecto surgió inicialmente de nuestro interés en encontrar datos que nos pudieran orientar sobre los precios en posibles destinos de vacaciones. Una vez encontramos el sitio web que usamos como fuente decidimos darle otra orientación al proyecto pensando “**Que haría un responsable de recursos humanos de una consultora sabiendo lo que cuesta un Big Mac en cada país del mundo**”. Pensamos que más interesante que planificar unas vacaciones podríamos usar los datos como fuente de información para asistir a una empresa con clientes en todo el mundo en el cálculo de las dietas apropiadas para sus trabajadores desplazados.

De esta forma decidimos ponernos en el lugar de un nuevo responsable de RRHH en una consultora, decidido a **optimizar el cálculo de las dietas de los trabajadores desplazados a clientes para adaptarlas al país al que el trabajador es desplazado**. La selección de los precios del menú de plato del día o de un Big Mac como indicadores es un guiño al espíritu de algunas consultoras decididas a maximizar sus beneficios reduciendo los que reciben sus empleados más juniors.

Este dataset **es también útil para cualquier persona que quiera hacerse una idea de que presupuesto necesitaría para el día a día en un país extranjero**, por ejemplo durante unas vacaciones o una estancia de estudio. Para el scope de este dataset no hemos tenido en cuenta el precio de los alojamientos ya que estos son generalmente contratados y abonados directamente por las empresas, pero esto podría ser incluido con facilidad si en futuro así lo consideramos necesario

8. Licencia

Una posible opción para este conjunto de datos es la **CC BY-SA 4.0 License**. Esto se debe a que las cláusulas de esta licencia parecen adecuadas para este proyecto.

Permite hacer uso de los datos y su distribución bajo la condición de incluir el nombre del creador del conjunto de datos permitiendo reconocer el trabajo de terceros, las contribuciones respecto al trabajo original y obligando a compartir el trabajo derivado bajo la misma licencia

Permite usar y modificar los datos para usos comerciales y no comerciales en tanto dicha modificación se comparta bajo la misma licencia.

9. Código

El código de nuestra aplicación será accesible en el enlace a Github:
https://github.com/jeatzr/precios-paises

Es un repositorio privado por lo que solo los miembros del equipo de desarrollo y el profesor al que hemos autorizado para corregir la práctica tienen acceso.

Para la realización de la práctica ambos miembros del equipo hemos usado el mismo entorno para evitar problemas:

- **Python 3.8.9**
- Visual Studio Code 1.77.3

Las librerías que hemos importado para el proyecto son las siguientes

```
import pandas as pd
import requests
from bs4 import BeautifulSoup
import time
import re
import datetime
import os
```

Detallaremos para qué usamos cada módulo:

- **pandas (v. 1.5.3)**: Este módulo nos facilita la **creación de un dataframe** con los datos que hemos leído del sitio web y que tenemos guardados en listas. También nos proporciona la valiosa función ***to_csv*** que es la que **finalmente guarda el fichero CSV en disco**.
- **requests (v. 2.28.2)**: Este módulo nos proporciona la función *get* para crear peticiones HTTP tipo GET al sitio web para obtener el código HTML de las distintas páginas que visitará el *scrapper*.
- **beautifulsoup4 (v. 4.12.0):** Este módulo nos proporciona las utilidades para organizar el texto plano de la página codificada en HTML en una estructura de árbol tipo DOM a la que podemos fácilmente acceder para rescatar información.
- **time**: Usaremos su función *sleep* para hacer las pausas de 1 segundo entre las visitas de las páginas para evitar que el sitio web nos bloquee por realizar demasiadas peticiones.
- **re:** Hemos usado esta librería para generar una expresión regular que identifique el salario medio.
- **datetime:** Hemos usado la librería para obtener la fecha actual en modo texto para agregarla al nombre del fichero CSV.
- **os:** Hemos usado esta librería para comprobar la existencia de la carpeta de destino para los datasets y asignarla como destino a la hora de crear el fichero csv.

Ha sido necesario instalar BeautifulSoup y Pandas. Requests viene instalado por defecto en esta versión de Python y el resto de librerías vienen en el núcleo de Python. Por lo tanto si se quiere ejecutar el proyecto con las mismas versiones habría que utilizar el comando *pip* de la siguiente manera:

- pip install beautifulsoup4==4.12.0
- pip install pandas==1.5.3

Un problema encontrado fue que **la columna que alberga el precio en Euros tiene distinta posición** dependiendo si el país tiene una moneda propia diferente al Euro o el Dólar. En dichos países los precios vienen expresados en tres monedas mientras que en países de la Eurozona y en USA solo existen dos columnas para expresar el precio. Al principio no teníamos esta circunstancia en cuenta por lo que obtuvimos datos erróneos así que decidimos **comprobar para cada país cuántas columnas tiene la tabla de precios** para **obtener el índice de columna donde aparece el precio en Euros,** que es la moneda en la que hemos decidido expresar los precios en nuestro dataset.

```
# si solo hay tres columas el precio viene solo en Dólar y Euro
# por lo que el precio en euros esta en la columna 2
if len(rows[0].find_all("th")) == 3:
    i_euro = 2
# en caso contrario suponemos que hay una columna más de precios
# para el precio en moneda loca, por lo que
# el precio en euros está en columna 3
else:
    i_euro = 3
```

Otro de los problemas encontrados es que **no para todos los países existen todos los precios buscados**. No eran muchos los casos pero por ejemplo había algún país como Haití al que en la sección “Servicios y transportes” le faltaba el campo de “Taxi 1Km”. También se detectó que en muchos países islámicos faltaba el precio de la cerveza nacional. Esta circunstancia hace que **no se pueda generalizar y acceder a los precios por su posición en la tabla** ya que en ocasiones se accede al precio erróneo o incluso puede dar un error de ejecución al acceder un índice fuera de rango.

Lo que en principio hacíamos de esta manera, accediendo a los números de línea concretos

```
# En caso de que haya más de una fila leemos los precios
if len(rows) > 1:
    precioBillete = rows[7].find_all("td")[i_col].find(string=True)
    inicioTaxi = rows[5].find_all("td")[i_col].find(string=True)
    taxi1Km = rows[4].find_all("td")[i_col].find(string=True)
# En caso contrario los precios tendrían valor "null"
else:
    precioBillete = "null"
    inicioTaxi = "null"
    taxi1Km = "null"
```

Lo hemos sustituído por una función *getIndexOf* a la cual pasamos la definición textual exacta del precio tal cual viene en la página, la lista de líneas de la tabla y el índice de la columna

```
def getPriceOf(priceDefinition, rows, i_col):
    '''
      Devuelve el valor del precio dado por:
      priceDefinition: Definición del precio tal cual sale en preciosmundi
      rows: lista de elementos <tr> de la tabla
      i_col: índice de la columna para el precio
    '''
    value = "null"
    for row in rows:
        tds = row.find_all("td")
        if tds:
            if tds[0].find(string=True) == priceDefinition:
                value = tds[i_col].find(string=True)
    return value
```

Como se puede observar si no encontrase nada devolvería “null” de manera que tengamos algún valor y sea posteriormente tratable

Dicha función la invocamos de esta manera:

```
precioBillete = getPriceOf(
    'Un billete de ida en transporte público', rows, i_col)
inicioTaxi = getPriceOf(
    'Inicio taxi (tarifa normal)', rows, i_col)
taxi1Km = getPriceOf(
    'Taxi 1km (tarifa normal)', rows, i_col)
```

10. Dataset

DOI del dataset en formato CSV descargable en Zenodo:

https://zenodo.org/record/7859418#.ZEaqY-xBz0p

11. Vídeo

Link del vídeo:

https://drive.google.com/file/d/1_nGlD6HdB9Ti4VV8kelKKNKvAoZQ-Qfb/view?usp=sharing

Contribuciones	Firma
Investigación previa	JEAI, DAG
Redacción de las respuestas	JEAI, DAG
Desarrollo del código	JEAI, DAG
Participación en el video	JEAI, DAG