Tipología y ciclo de vida de los datos

# PRA1: Web Scraping – elaboración del dataset

# AUTORES: JOSE LUIS SANTOS DURANGO, MARÍA ISABEL GONZÁLEZ SÁNCHEZ

## **Context****o**

**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 página web que hemos elegido es la página: <https://www.expatistan.com/cost-of-living>

Esta página web ofrece información sobre el costo de vida en distintos países del mundo y distintas ciudades, así como la opción de realizar una comparativa entre dos opciones elegidas. La información viene detallada por distintas categorías: comida, vivienda, ropa, transporte, cuidado personal y entretenimiento, y a su vez se categoriza por un nivel más detallado dentro de cada categoría. El objetivo del proyecto de creación de esta página era ofrecer una visión general del costo de vida en otros países/ciudades para que un expatriado pudiera hacerse una ligera idea del coste de vida en su futuro destino.

La forma en la que se añaden los datos es por crowdsourcing, de forma que los usuarios de la página web añaden los precios de sus propias ciudades y posteriormente mediante un modelo estadístico, se limpian y modelan los datos para eliminar entradas erróneas y poder mostrar el promedio de los precios. A parte, los datos introducidos se combinan con fuentes oficiales de consumo como: [OCDE](https://stats.oecd.org/Index.aspx?DataSetCode=PPP2005), [INE](https://www.ine.es/daco/daco43/notaipca.htm) [Consumer Expenditure Survey](ftp://ftp.bls.gov/pub/special.requests/ce/standard/2007/occup.txt), [Bureau of Labor Statistics, US DOL](http://www.bls.gov/cex/), entre otras. Con un precio fiable de los productos, se convierte a una moneda común mediante el cambio de divisas oficial de la moneda local a la moneda común, y se ponderan las distintas subcategorías, permitiendo así asignar una valoración al país o ciudad en cuestión, para poder generar un ranking.

El nivel del costo de vida en el ranking asociado a cada país y ciudad tiene como base la ciudad de Praga, República Checa, de modo que se le asigna el nivel 100, siendo este el punto de referencia para el resto de comparaciones. El motivo de elegir Praga como punto de referencia es debido a que no es una ciudad muy cara ni muy barata, tiene una gran comunidad de expatriados, y el propietario de la página es expatriado en Praga.

## **Título**

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

En este proyecto hemos extraído dos conjuntos de datos:

* **Cost of living by county:** este título corresponde al conjunto de datos almacenado en el fichero cost\_of\_living\_countries.csv y contiene la información del costo de vida categorizado por país. El scraping ha sido realizado a la página que muestra el ranking por país e individualmente a cada una de las páginas de cada país del que se tiene información.
* **Cost of living by city:** este título responde al conjunto de datos almacenado en el fichero cost\_of\_living\_cities.csv y contiene la información del costo de vida categorizado por ciudad. El scraping ha sido realizado a la página que muestra el ranking por ciudad e individualmente a cada una de las páginas de cada ciudad de la que se tiene información.

## **Descripción del dataset**

**Desarrollar una breve descripción del conjunto de datos que se ha extraído. Es necesario que esta descripción sea coherente con el título elegido.**

Los dos datasets extraídos contienen información de precios de diferentes gastos que podemos tener en un país o ciudad determinada. Tal como se detalla en la pagina, los gastos de un "expatriado típico" (si existe tal cosa) se dividen en 6 grupos:

* Comida
* Alojamiento
* Ropa
* Transporte
* Cuidado Personal
* Entretenimiento

Esta agrupación se basa en informes internacionales sobre variaciones en el costo de vida.

La diferencia entre ambos datasets, es la división geográfica. En el dataset de países, únicamente tenemos la información del país, mientras que en al dataset de ciudades, tenemos la categorización país, ciudad y estado (en caso de que la página tenga esa información). En la siguiente captura podemos apreciar de dónde se extrae la información de forma detallada.

**DATASET DE PAÍSES**

**Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamenteInterfaz de usuario gráfica

Descripción generada automáticamente**

Nombre del país

**Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente**

Ranking

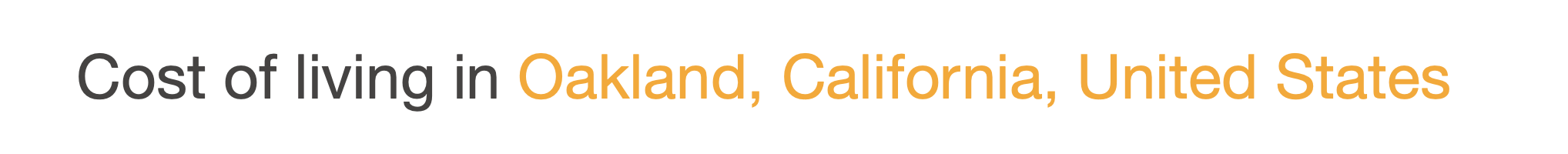
Categoría general

Subcategoría

Precio moneda local

Precio moneda €

**DATASET DE CIUDADES:** la principal diferencia con respecto al scraping de los datos de países, es la extracción del nombre de la tupla ciudad, país, estado:



País

Estado

Ciudad

## **Representación gráfica**

**Dibujar un esquema o diagrama que refleje visualmente el dataset y el proyecto elegido.**

**Logotipo, nombre de la empresa

Descripción generada automáticamente**

**Fuentes de datos:**

Diferentes canales oficiales

Expatriados residentes en varios países actualizan los datos de precios en la propia página.

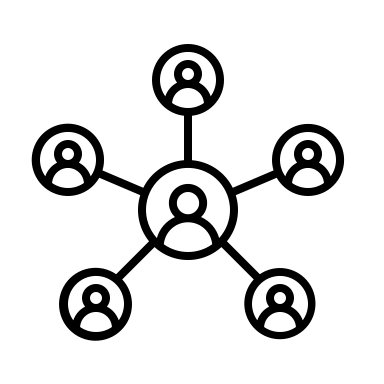
**Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente**

Dibujo de personaje amarillo

Descripción generada automáticamente con confianza baja

Forma, Flecha

Descripción generada automáticamente****

**Logotipo

Descripción generada automáticamente con confianza media**

Patrón de fondo

Descripción generada automáticamente

**Logotipo, nombre de la empresa

Descripción generada automáticamente**Logotipo, Icono

Descripción generada automáticamente

Procesado interno de los datos:

Calidad de datos, verificación, algoritmo de clasificación de precios, software interno para visualización de los datos

Logotipo, nombre de la empresa

Descripción generada automáticamente

Texto

Descripción generada automáticamente

**Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente**

**Diagrama, Forma

Descripción generada automáticamente**

## **Una señal de alto Descripción generada automáticamente con confianza mediaForma Descripción generada automáticamente con confianza baja**Imagen que contiene Icono Descripción generada automáticamenteIcono Descripción generada automáticamente

Análisis de los datos – PRA2

Datasets

Web scraping

## **Contenido**

**Explicar los campos que se incluyen en el dataset y el período de tiempo al que pertenecen los datos.**

Los campos comunes a ambos datasets son los siguientes:

* **Ranking position:** posición que ocupa el país dentro del ranking de costo de vida de los países incluidos en el estudio.
* **Country:** país sobre el que se realiza la valoración de los costos.
* **Category:** categoría general de los costos. Este valor varía entre comida, alojamiento, ropa, transporte, cuidado personal y entretenimiento.
* **Items:** subcategoría del costo.
* **Original Currency:** moneda local del país.
* **Original Currency Value:** costo del ítem en la moneda local.
* **Exchanged Currency:** por defecto EUR.
* **Exchanged Currency Value:** costo del ítem en EUR.

Además, en el dataset de ciudades se añaden los campos adicionales:

* **City:** ciudad sobre la que se realiza la valoración de los costos.
* **State:** estado al que pertenece la ciudad, en caso de que exista.

El periodo de tiempo al que pertenece los datos, tal como se puede observar en la página web: “una vez que tenemos precios fiables para cada producto y ciudad, para comparar dos ciudades primero convertimos los precios de cada producto o servicio a una moneda común utilizando el tipo de cambio actual (actualizado cada 3 días). Luego comparamos cada producto o servicio individual y encontramos la diferencia de precio entre los dos. A continuación, combinamos los resultados de cada producto dentro de su grupo utilizando el peso de cada subgrupo para ajustar la diferencia dentro del grupo. En este momento la diferencia de precio de los grupos se ponderará y combinará para obtener la diferencia de precio final entre ciudades.”

Luego podemos considerar que los datos se actualizan a medida que se van incorporando nuevos valores de las categorías en una ciudad en concreto y cuando ha sido validado por el algoritmo interno de la plataforma, entonces se añaden estos nuevos datos al cálculo de la media ponderada para sacar los rankings. Por lo tanto, podemos considerar que la frecuencia de actualización es alta, y con respecto a los cambios de divisas podemos fijarlo en 3 días.

## **Propietario**

Revisar interesante para prox apartados:

<https://www.expatistan.com/es/como-funciona>

<https://www.expatistan.com/privacy-policy>

<https://www.expatistan.com/faq>

## **Referencias**

* **Expatistan. Cost of living.**

<https://www.expatistan.com>