

Importación de librerías

```
In [ ]: import pandas as pd
import re
```

Variables destacadas

```
In [ ]: pathToDF = "../../../Inputs/Creados - Proyecto/"
fileToDF = "dfVentasDefinitivo.csv"

pathToExport = "../Outputs/"
fileToExport = "ProdsToScrap.csv"
```

Importación de datos

```
In [ ]: df = pd.read_csv(f"{pathToDF}{fileToDF}")
```

```
In [ ]: #Sólo ejecutar una vez, elimina primera columna si la exportación de dicho C
df.drop(columns=df.columns[0], axis=1, inplace=True)
```

Comprobación preliminar de los datos (tamaño y composición del dataframe)

```
In [ ]: print(df.shape)
df.loc[df["product_id"]==12708]

(810167, 26)
```

```
Out[ ]:
```

	item_id	num_order	created_at	product
809559	ffcca82d7cf99085286c8b609b5986bb	d95c0ceef2ad6902e57f2b8a00c3f57c	2018-09-30 06:31:41	12

1 rows × 26 columns

Depuración de datos

Sólo se escogen las columnas referidas a los productos en sí.

```
In [ ]: df = df[["product_id", "name"]]

print(df.shape)
df.head(2)

(810167, 2)
```

```
Out[ ]:
```

	product_id	name
0	5645	Weleda Hombre Crema Hidratante 30 ml
1	28743	Gynea Gestagyn Men 60 Capsulas

Se comprueba que los tamaños de ambas columnas en sus valores únicos no coincide.

```
In [ ]: print(len(df["product_id"].unique()))
        len(df["name"].unique())
```

```
19787
```

```
Out[ ]: 19778
```

Por tanto, se de toma como referencia la variable con menor número de únicos para generar el dataframe preliminar para tratar.

Las comprobaciones de que los valores no se hubieran descuadrado, se realizó en el documento de testeo.

```
In [ ]: df = df.drop_duplicates(subset=['name'])
        print(df.shape)
        df.iloc[[1]]
```

```
(19778, 2)
```

```
Out[ ]:
```

	product_id	name
1	28743	Gynea Gestagyn Men 60 Capsulas

Eliminación de cantidades en Productos

Cada columna del dataframe resultante se transforma a lista, lo más reseñable es la aplicación de REGEX para limpiar los nombres de los productos únicos y así obtener un resultado más consistente.

La aplicación de dicha expresión regular también se aplicó para intentar reducir el número de nombres de productos únicos y así mejorar la eficiencia del WebScrapping posterior.

```
In [ ]: idModded = df.index.tolist()
        productModded = df["product_id"].unique().tolist()
        nameModded =[ele.rstrip() for ele in list(map(lambda x: re.sub("\d+\s*\S*\w+
```

Creación de CSV con los datos tratados

Con las listas ya depuradas en el apartado anterior, se genera y comprueba el dataframe final.

```
In [ ]: dfDef = pd.DataFrame({"id":idModded,"product_id":productModded,"name":nameMc  
dfDef["name"].drop_duplicates()  
print(dfDef.shape)  
dfDef.head(2)
```

(19778, 3)

```
Out[ ]:   id  product_id      name  
0  0      5645  Weleda Hombre Crema Hidratante  
1  1      28743      Gynea Gestagyn Men
```

Exportación de resultados

```
In [ ]: # Importante conservar el "index=False" para evitar inconsistencias en impor  
dfDef.to_csv(f"{pathToExport}{fileToExport}", index=False)
```