

Importación de librerías

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

Inicio de la depuración

```
In [ ]: products = pd.read_csv("products_limpieza1.csv")
df_categories = pd.read_csv("categories_limpieza1.csv")
```

```
In [ ]: products.head()
```

```
Out[ ]:
```

	Unnamed: 0	product_id	sku	name	marca_value	short_description	analytic_category
0	0	7349	165774	Pañales CHELINO LOVE Talla3 4-10 KG 36unidades	chelino	Los pañales Chelino, que presentan un novedoso...	Infantil
1	1	7347	165776	Pañales CHELINO LOVE Talla5 13-18 KG 30unidades	chelino	Los pañales Chelino que presentan un novedoso ...	Infantil
2	2	50282	230154	Pañales Pingo Maxi T4 40 uds	pingo	Pingo es el primer pañal ecológico con cuatro ...	Infantil
3	3	7348	165775	Pañales CHELINO LOVE Talla4 9-15 KG 34unidades	chelino	Los pañales Chelino, que presentan un novedoso...	Infantil
4	4	24517	011905	Dodot Pañal T4 9-15Kg 30unds	dodot	¿Qué le pides a un pañal? ¿Absorción máxima, q...	NaN

```
In [ ]: df_categories.head(6)
```

Out []:

	Unnamed: 0	sku	cat1	cat2	cat3
0	0	00.01.10.014	Cosmética y Belleza	Corporal	Hidratación
1	1	00.071697.000.000	Infantil	Juguetes	Mordedores
2	2	000133	Infantil	Higiene infantil	Canastillas y kits bebé
3	3	000147	Higiene y cuidado personal	Facial	Desmaquillantes y limpiadores
4	4	000148	Cosmética y Belleza	Manos	Crema de Manos
5	5	000200	Higiene y cuidado personal	Facial	Mascarillas

Fusión de dataframes

In []:

```
products_full = pd.merge(products, df_categories, on='sku', how='left')
```

In []:

```
products_full.head(6)
```

Out []:

	Unnamed: 0_x	product_id	sku	name	marca_value	short_description	analytic_category
0	0	7349	165774	Pañales CHELINO LOVE Talla3 4-10 KG 36unidades	chelino	Los pañales Chelino, que presentan un novedoso...	Infantil
1	1	7347	165776	Pañales CHELINO LOVE Talla5 13-18 KG 30unidades	chelino	Los pañales Chelino que presentan un novedoso ...	Infantil
2	2	50282	230154	Pañales Pingo Maxi T4 40 uds	pingo	Pingo es el primer pañal ecológico con cuatro ...	Infantil
3	3	7348	165775	Pañales CHELINO LOVE Talla4 9-15 KG 34unidades	chelino	Los pañales Chelino, que presentan un novedoso...	Infantil
4	4	24517	011905	Dodot Pañal T4 9-15Kg 30unds	dodot	¿Qué le pides a un pañal? ¿Absorción máxima, q...	NaN
5	5	11938	011896	Dodot Activity T3 De 4 A 10 Kg 56 Unidades	dodot	Pañales que mantienen la piel del culito seca ...	NaN

```
In [ ]: products_full["cat1"].value_counts()
```

```
Out[ ]: Cosmética y Belleza      3352
        Higiene y cuidado personal 2987
        Salud                  2842
        Infantil              1298
        Nutrición              615
        Veterinaria            20
        salud                  3
        nutrición              1
        Name: cat1, dtype: int64
```

```
In [ ]: products_full["analytic_category"].value_counts()
```

```
Out[ ]: Cosmética y Belleza      6821
        Higiene                  4334
        Infantil                 3857
        Herbolario               2487
        Nutrición                1461
        Ortopedia                496
        Vida Íntima              259
        Óptica                   195
        Perfumeria               145
        Veterinaria              45
        Name: analytic_category, dtype: int64
```

```
In [ ]: #Para unificar categorias pasamos a lower (minúsculas):
        products_full["analytic_category"] = products_full["analytic_category"].str.
```

```
In [ ]: products_full["analytic_category"].head()
```

```
Out[ ]: 0    infantil
        1    infantil
        2    infantil
        3    infantil
        4         NaN
        Name: analytic_category, dtype: object
```

```
In [ ]: products_full[products_full["analytic_category"].isna()].head()
```

Out[]:

	Unnamed: 0_x	product_id	sku	name	marca_value	short_description	analytic_category
4	4	24517	011905	Dodot Pañal T4 9-15Kg 30unds	dodot	¿Qué le pides a un pañal? ¿Absorción máxima, q...	NaN
5	5	11938	011896	Dodot Activity T3 De 4 A 10 Kg 56 Unidades	dodot	Pañales que mantienen la piel del culito seca ...	NaN
6	6	13109	011894	Dodot Pañal Sensitive T1 2-5Kg 30Unds	dodot	Pañales para recién nacido que evitan la irrit...	NaN
14	14	12051	013687	Dodot Sensitive Pañal Primeras Semanas T/0 Has...	dodot	Pañales Dodot para una máxima absorción de pip...	NaN
31	31	13122	011907	Dodot Pañal Pack Semanal T/5 13-18kg 25ud	dodot	Pañales pack semanal\n	NaN

In []: `products_full.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23183 entries, 0 to 23182
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0_x          23183 non-null  int64
1   product_id            23183 non-null  int64
2   sku                   23183 non-null  object
3   name                  23183 non-null  object
4   marca_value           23181 non-null  object
5   short_description     23162 non-null  object
6   analytic_category     20100 non-null  object
7   picture               23177 non-null  object
8   nombre_corto         23183 non-null  object
9   Unnamed: 0_y          11118 non-null  float64
10  cat1                  11118 non-null  object
11  cat2                  11098 non-null  object
12  cat3                  10819 non-null  object
dtypes: float64(1), int64(2), object(10)
memory usage: 2.5+ MB
```

In []: `# Por qué esas variables unnamed:0? Se eliminan...`

```
In [ ]: products_full.drop(columns=["Unnamed: 0_x", "Unnamed: 0_y"]).head()
```

```
Out[ ]:
```

	product_id	sku	name	marca_value	short_description	analytic_category	
0	7349	165774	Pañales CHELINO LOVE Talla3 4-10 KG 36unidades	chelino	Los pañales Chelino, que presentan un novedoso...	infantil	https://www /m /
1	7347	165776	Pañales CHELINO LOVE Talla5 13-18 KG 30unidades	chelino	Los pañales Chelino que presentan un novedoso ...	infantil	https://www /m /
2	50282	230154	Pañales Pingo Maxi T4 40 uds	pingo	Pingo es el primer pañal ecológico con cuatro ...	infantil	https://www /m /
3	7348	165775	Pañales CHELINO LOVE Talla4 9-15 KG 34unidades	chelino	Los pañales Chelino, que presentan un novedoso...	infantil	https://www /m /
4	24517	011905	Dodot Pañal T4 9-15Kg 30unds	dodot	¿Qué le pides a un pañal? ¿Absorción máxima, q...	NaN	https://www /m /

Es necesario que "analytic_category" tome lo que está en el mismo registro en "cat1"

```
In [ ]: products_full["analytic_category"].isna().sum()
```

```
Out[ ]: 3083
```

```
In [ ]: (products_full["analytic_category"] == None).sum()
```

```
Out[ ]: 0
```

```
In [ ]: for pos, na in enumerate(products_full["analytic_category"]): #no funciona...
        if na == None:
            products_full["analytic_category"].iloc[pos] = products_full["cat1"]
```

```
In [ ]: products_full.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23183 entries, 0 to 23182
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0_x           23183 non-null  int64
1   product_id            23183 non-null  int64
2   sku                   23183 non-null  object
3   name                  23183 non-null  object
4   marca_value           23181 non-null  object
5   short_description     23162 non-null  object
6   analytic_category     20100 non-null  object
7   picture               23177 non-null  object
8   nombre_corto          23183 non-null  object
9   Unnamed: 0_y          11118 non-null  float64
10  cat1                  11118 non-null  object
11  cat2                  11098 non-null  object
12  cat3                  10819 non-null  object
dtypes: float64(1), int64(2), object(10)
memory usage: 2.5+ MB
```

```
In [ ]: products_full["analytic_category"].isna().sum()
```

```
Out[ ]: 3083
```

Revisión de tipos

```
In [ ]: products.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23183 entries, 0 to 23182
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0            23183 non-null  int64
1   product_id            23183 non-null  int64
2   sku                   23183 non-null  object
3   name                  23183 non-null  object
4   marca_value           23181 non-null  object
5   short_description     23162 non-null  object
6   analytic_category     20100 non-null  object
7   picture               23177 non-null  object
8   nombre_corto          23183 non-null  object
dtypes: int64(2), object(7)
memory usage: 1.6+ MB
```

```
In [ ]: df_categories.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11159 entries, 0 to 11158
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Unnamed: 0   11159 non-null  int64
1   sku          11159 non-null  object
2   cat1         11159 non-null  object
3   cat2         11139 non-null  object
4   cat3         10859 non-null  object
dtypes: int64(1), object(4)
memory usage: 436.0+ KB
```

```
In [ ]: #test2 = [re.search("([^\;]*)", element).group() for element in categories_se
```

Eliminación de columnas no relevantes

```
In [ ]: products_full.drop(columns=["Unnamed: 0_x", "Unnamed: 0_y"], inplace=True, a
```

Exportación de resultados a CSV

```
In [ ]: products_full.to_csv("df_products_dep.csv")
```