

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

```
In [ ]: df_categories = pd.read_csv("products_categories.csv", sep=";", encoding='la
```

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

```
Out[ ]:
```

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

```
In [ ]: # df_categories[df_categories["cat1"].isna()]
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11159 entries, 0 to 11158
Data columns (total 4 columns):
#   Column  Non-Null Count  Dtype
---  -
0    sku      11159 non-null    object
1    cat1     10741 non-null    object
2    cat2     10721 non-null    object
3    cat3     10441 non-null    object
dtypes: object(4)
memory usage: 348.8+ KB
```

```
In [ ]: categories_na = df_categories[df_categories["cat1"].isna()]
```

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

```
Out[ ]:
```

	sku	cat1	cat2	cat3
30	001268,Salud,"Huesos, articulaciones y músculo...	NaN	NaN	NaN
44	002655,Salud,Botiquín,"Apósitos, tiritas y gasas"	NaN	NaN	NaN
49	002870,Salud,"Huesos, articulaciones y músculo...	NaN	NaN	NaN
50	002872,Salud,"Huesos, articulaciones y músculo...	NaN	NaN	NaN
51	002873,Salud,"Huesos, articulaciones y músculo...	NaN	NaN	NaN

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 418 entries, 30 to 11082
Data columns (total 4 columns):
#   Column  Non-Null Count  Dtype
---  -
0    sku      418 non-null     object
1   cat1       0 non-null       object
2   cat2       0 non-null       object
3   cat3       0 non-null       object
dtypes: object(4)
memory usage: 16.3+ KB
```

```
In [ ]: categories_sep = categories_na["sku"].str.split(",", expand=True)
```

```
In [ ]: categories_sep = categories_sep.rename(columns={0: "sku", 1: "cat1"})
```

```
In [ ]: categories_sep["cat2,3"] = categories_sep[[2,3,4]].apply(" ".join, axis=1)
```

```
In [ ]: categories_sep = categories_sep.drop(labels=[2,3,4], axis=1)
```

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

```
Out[ ]:
```

	sku	cat1	cat2,3
30	001268	Salud	"Huesos, articulaciones y músculos", Colágeno
44	002655	Salud	Botiquín, "Apósitos, tiritas y gasas"
49	002870	Salud	"Huesos, articulaciones y músculos", Colágeno
50	002872	Salud	"Huesos, articulaciones y músculos", Colágeno
51	002873	Salud	"Huesos, articulaciones y músculos", Colágeno

```
In [ ]: categories_sep_2_3 = categories_sep["cat2,3"].str.split(' ', expand=True)
```

```
In [ ]: categories_sep_2_3 = categories_sep_2_3.rename(columns={0: "cat2", 1: "cat3", 2: "2"})
```

```
In [ ]: categories_sep_2_3.head()
```

```
Out[ ]:
```

	cat2	cat3	2
30	Huesos, articulaciones y músculos		, Colágeno
44	Botiquín,	Apósitos, tiritas y gasas	
49	Huesos, articulaciones y músculos		, Colágeno
50	Huesos, articulaciones y músculos		, Colágeno
51	Huesos, articulaciones y músculos		, Colágeno

```
In [ ]: categories_sep_2_3.head()
```

```
Out[ ]:
```

	cat2	cat3	2
30	Huesos, articulaciones y músculos	, Colágeno	
44	Botiquín,	Apósitos, tiritas y gasas	
49	Huesos, articulaciones y músculos	, Colágeno	
50	Huesos, articulaciones y músculos	, Colágeno	
51	Huesos, articulaciones y músculos	, Colágeno	

```
In [ ]: for pos, na in enumerate(categories_sep_2_3["cat2"]):
        if na == "":
            categories_sep_2_3["cat2"].iloc[pos] = categories_sep_2_3["cat3"].il
            categories_sep_2_3["cat3"].iloc[pos] = categories_sep_2_3[2].iloc[pc

categories_sep_2_3.drop(columns=[2], inplace=True)
```

```
In [ ]: categories_sep_2_3["cat3"]=categories_sep_2_3["cat3"].str.replace("^[,]\s",
categories_sep_2_3["cat2"]=categories_sep_2_3["cat2"].str.replace("[,]\s$",
```

```
In [ ]: categories_sep_2_3.head()
```

```
Out[ ]:
```

	cat2	cat3
30	Huesos, articulaciones y músculos	Colágeno
44	Botiquín Apósitos, tiritas y gasas	
49	Huesos, articulaciones y músculos	Colágeno
50	Huesos, articulaciones y músculos	Colágeno
51	Huesos, articulaciones y músculos	Colágeno

```
In [ ]: categories_sep_clean = pd.merge(categories_sep, categories_sep_2_3, left_incl
categories_sep_clean.drop(columns="cat2,3", inplace=True)
```

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

```
Out[ ]:
```

	sku	cat1	cat2	cat3
30	001268	Salud	Huesos, articulaciones y músculos	Colágeno
44	002655	Salud	Botiquín Apósitos, tiritas y gasas	
49	002870	Salud	Huesos, articulaciones y músculos	Colágeno
50	002872	Salud	Huesos, articulaciones y músculos	Colágeno
51	002873	Salud	Huesos, articulaciones y músculos	Colágeno

```
In [ ]: df_categories.drop(df_categories[df_categories["cat1"].isna()].index, inplace=
```

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10741 entries, 0 to 11158
Data columns (total 4 columns):
#   Column  Non-Null Count  Dtype
---  -
0    sku      10741 non-null    object
1   cat1      10741 non-null    object
2   cat2      10721 non-null    object
3   cat3      10441 non-null    object
dtypes: object(4)
memory usage: 419.6+ KB
```

```
In [ ]: # for i in df_categories["cat1"]:
#       if i is na:
#           df_categories["cat1"][i].drop(axis=0, inplace=True)
```

```
In [ ]: df_categories["cat1"].isna().sum()
```

```
Out[ ]: 0
```

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10741 entries, 0 to 11158
Data columns (total 4 columns):
#   Column  Non-Null Count  Dtype
---  -
0    sku      10741 non-null    object
1   cat1      10741 non-null    object
2   cat2      10721 non-null    object
3   cat3      10441 non-null    object
dtypes: object(4)
memory usage: 419.6+ KB
```

```
In [ ]: df_categories_limpieza1 = pd.concat([df_categories, categories_sep_clean])
```

```
In [ ]: # df_categories_limpieza1 = df_categories.append(categories_sep_clean, ignore_index=True)
```

```
In [ ]: df_categories_limpieza1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11159 entries, 0 to 11082
Data columns (total 4 columns):
#   Column  Non-Null Count  Dtype
---  -
0    sku      11159 non-null    object
1   cat1      11159 non-null    object
2   cat2      11139 non-null    object
3   cat3      10859 non-null    object
dtypes: object(4)
memory usage: 435.9+ KB
```

```
In [ ]: df_categories_limpieza1[df_categories_limpieza1["cat2"].isna()].head()
```

```
Out[ ]:
```

	sku	cat1	cat2	cat3
<b>31</b>	001469	Veterinaria	NaN	NaN
<b>743</b>	113159=575349	Veterinaria	NaN	NaN
<b>8172</b>	570052	Veterinaria	NaN	NaN
<b>8173</b>	570235	Veterinaria	NaN	NaN
<b>8174</b>	570237	Veterinaria	NaN	NaN

```
In [ ]: df_categories["cat1"] = df_categories["cat1"].str.lower()  
df_categories["cat2"] = df_categories["cat2"].str.lower()  
df_categories["cat3"] = df_categories["cat3"].str.lower()
```

```
In [ ]: df_categories_limpieza1.to_csv("categories_limpieza1.csv")
```

```
In [ ]: # categories_na.to_csv("categories_na.csv")
```

```
In [ ]: # categories_na_df = pd.read_csv("categories_na.csv", sep=',')
```

```
In [ ]: # categories_na_df
```