

Importación de librerías

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

Inicio de la depuración

```
In [ ]: # Carga de df
df_ventas = pd.read_csv("df_ventas_2.csv")
df_zipcodes = pd.read_csv("df_products_dep.csv")
```

```
In [ ]: # Se elimina la columna derivada de los indices
df_ventas.drop(columns="Unnamed: 0",inplace=True)
df_zipcodes.drop(columns="Unnamed: 0",inplace=True)
```

```
In [ ]: df_ventas_full = df_ventas.merge(df_zipcodes, how="left", on="product_id")
```

Se hace drop de las categorías, ya que la dos y tres contienen muchos valores perdidos, y la cat1 se encuentra recogida en analytic_category

```
In [ ]: #
df_ventas_full.drop(columns = df_ventas_full.loc[:, ["cat1", "cat2", "cat3"]
```

```
In [ ]: df_ventas_full.isna().sum()
```

```
Out[ ]: item_id          0
        num_order       0
        created_at      0
        product_id      0
        qty_ordered     0
        base_cost       2330
        price           0
        discount_percent 0
        customer_id     0
        zipcode         0
        longitud_zip     0
        country         0
        region          0
        city            0
        date            0
        year            0
        hour            0
        week            0
        day             0
        margin_total    2330
        price_total     0
        sku             27125
        name            27125
        marca_value     27127
        short_description 27286
        analytic_category 90727
        picture         27267
        nombre_corto    27125
        dtype: int64
```

```
In [ ]: df_ventas_full["product_id"] = df_ventas_full["product_id"].astype("float",
```

Borramos columnas con poco valor para visualización y modelización que además contienen valores perdidos y que su imputación estaría expuesta a ser de forma semi aleatoria.

```
In [ ]: df_ventas_full.drop(columns = df_ventas_full.loc[:, ["sku", "short_descripti
```

Exportación final de resultados

```
In [ ]: df_ventas_full.to_csv("df_ventas_3.csv")
```

Imputación por randomforest de los valores perdidos mediante la librería de missingpy