EDA EXPLANATION AND PROBLEMS

Our goal was to forecast as many ART_COD as possible.

To do so, we merged the following datasets:

anagrafica_prodotti, risultati_annomese, informazioni_business

During this process, we encountered two main issues:

1. Missing product_group values

After the first merge, some ART_CODs had no product_group.

➤ We solved this by creating a mapping script to infer the missing product_group based on known ART_CODs.

2. Incomplete business information

After the second merge, some ART_CODs only had QTA and product_group, with all other columns as NaN.

➤ We decided to keep these entries, creating a dummy variable to flag them.

These may represent discontinued products that still generate sales and can be forecasted, this is the reason why we are trying to keep them in our models.

CODE PART

```
merged_df = pd.merge(annomese,an_prod, on="ART_COD", how="left")
merged_df
#mergiando ho trovato che invece ci sono circa 37k missing value,
#troviamo un modo per mapparli in modo da non perdere info
```

```
df = pd.merge(merged_df,info_bs, on=["ART_COD","ANNOMESE","ITEM_ID"], how="left")
#abbiamo fatto left perchè merge_df non aveva missing values per art_cod
#quindi abbiamo il numero massimo
df['ART_COD'].nunique()
```

```
6968
```

QUESTIONS

- Do you think our approach is appropriate to achieve our main goal?

- Since we observed a strong seasonal trend in our data, we are considering implementing a SARIMA model, do you think it's the right choice?

Annalaura Granata
Daniele Lupico
Gabriele Rizzo