ETL Process

Cities.csv

To fix **encoding errors** in the variable *city_code*:

- Created a list of correct city codes.
- Used fuzzy matching to find the best match for each erroneous city code using the fuzzywuzzy library.
- Created a new DataFrame 'cities_cleaned' with the updated 'city_code' variable.

Products.csv

Products data has been cleaned:

- Created a new DataFrame `product_cleaned` based on the original data.
- Converted hierarchy columns (‘hierarchy1_id','hierarchy2_id', 'hierarchy3_id', 'hierarchy4_id', 'hierarchy5_id’) to categorical dtype.

Sales.csv

Sales data has been cleaned:

- Created a new DataFrame `sales_cleaned` based on the original data.
- Converted 'date' column to datetime format.
- Converted promotional columns ('promo_type_1', 'promo_bin_1', 'promo_type_2', 'promo_bin_2', 'promo_discount_2', 'promo_discount_type_2') to categorical dtype.

Forecast Revenue.csv

- Renamed column 'WEEk' to 'Week' in fore_rev DataFrame
- Converted 'Date' column to datetime dtype and formatted dates as '%d/%m/%Y' in fore_rev_cleaned DataFrame