Project 2: Data Cleaning and Storytelling

Time Limit: 5 mins

Main question to be answered:

Should Eniac offer discounts on products?

Choose a few questions from here which help you justify the main question:

- How should products be classified into different categories in order to simplify reports and analysis?
- What is the distribution of product prices across different categories?
    - How many cheap/expensive products do we have?
    - How many sales/revenue do cheap/expensive products generate?
    - How big are the discounts by category?
- How many products are being discounted?
- How big are the offered discounts as a percentage of the product prices?
- How seasonality and special dates (Christmas, Black Friday) affect sales?
    - In which months is Eniac releasing more products?
    - Are sales for a product spiking the month it gets released?
- How could data collection be improved?

Summary of work done so far:

1. **Data cleaning**
    1. Make sure to convert each column from each dataset to the data type it belongs to.
    2. Prices in the products table seem corrupted: some of them have 2 dots or values that are too high —unrealistic.
    3. There's missing data.
2. **Data quality**
    1. Information from different tables should match:
        1. All products being sold must be present in the products table.

2. All orders in the orderlines table should be present in the orders table and vice-versa.

Tips:

Answer business questions

2. Exploring different product categories:
    1. Create product categories by pattern matching names / descriptions.
    2. Analyze categories in terms of revenue, popularity through time...
3. Exploring how sales/revenue evolve through time
4. Detecting different seasonal patterns (holidays, weekends, special days...)

3. **Provide a discount strategy**
    1. Analyze what has happened when discounts have been given.
        1. Discounts are differences between products.price and orderlines.unit_price
    2. Predict what will happen if more discounts are given