Eniac

With Pandas

# Whether or not it’s beneficial **to discount products**?

The Marketing Team Lead is convinced that offering discounts is beneficial in the long run. She believes discounts improve customer acquisition, satisfaction, and retention, and allow the company to grow.

The main investors in the Board are worried about offering aggressive discounts. They have pointed out how the company’s recent quarterly results showed an increase in orders placed, but a decrease in the total revenue. They prefer that the company positions itself in the quality segment, rather than competing to offer the lowest prices in the market.

## Business questions

Your analytical and business skills are needed to provide clarity in the following aspects:

* How should products be classified into different categories to simplify reports and analysis?
* What is the distribution of product prices across different categories?
* How many products are being discounted?
* How big are the offered discounts as a percentage of the product prices?
* How do seasonality and special dates (Christmas, Black Friday) affect sales?
* How could data collection be improved?

At the end of this project (your instructor will set the exact date), you are expected to give a presentation to the company board and present the results of your analysis. The presentation needs to be concise, but also engaging and convincing.

# A problem with the data

Before you get too deep into the business questions, there is some bad news we need to share with you: the **data appears to have some inconsistencies** —it must have been corrupted.

The Database Administrator, Khader, blames the buggy pipeline that connects the online store with the database, while the Software Engineer, Lina, claims the issues have to do with wrong encodings and bad maintenance practices in our database. As Khader and Lina argue with one another and try to fix the data source, the company board does not have time to lose, they want your business insights. In short**, you will have to assess the data quality of the dataset and take on many data cleaning tasks before you can even start to think about the business questions.**

How mad should you be about that? A popular figure shared all the time in Data Science communities is that Data Analysts and Scientists spend 80% of their time cleaning and preparing data. Ironically, there is no reliable data source for that number, but if you just talk with people in the field, they will all tell you it’s a major part of the job. There is little merit in applying advanced methods and fancy solutions to perfectly clean and formatted datasets. The real value good Data Scientists can bring is usually to make dirty data usable and trustable —with the necessary disclaimers.

# Project Overview

Here’s an overview of the project:

A group of logos and text

Description automatically generated

# Get the data.

# [**Get the data from this Drive folder**](https://drive.google.com/drive/folders/1mHBDnFvMOgxnZIVSAw1LyT8b2-qbzB8T?usp=sharing)**. (4 datasets)**

Here’s a description of each table and its columns:

**orders.csv** – Every row in this file represents an order.

* **order\_id**– a unique identifier for each order.
* **created\_date**– a timestamp for when the order was created.
* **total\_paid**– the total amount paid by the customer for this order, in euros.
* **state**
  + “Shopping basket” – products have been placed in the shopping basket.
  + “Place Order” – the order has been placed but is awaiting shipment details.
  + “Pending” – the order is awaiting payment confirmation.
  + “Completed” – the order has been placed and paid, and the transaction is completed.
  + “Cancelled” – the order has been cancelled and the payment returned to the customer.

**orderlines.csv** – Every row represents each one of the different products involved in an order.

* **id**– a unique identifier for each row in this file.
* **id\_order**– corresponds to orders.order\_id.
* **product\_id**– an old identifier for each product, nowadays not in use.
* **product\_quantity**– how many units of that product were purchased on that order.
* **sku** – stock keeping unit: a unique identifier for each product.
* **unit\_price**– the unitary price (in euros) of each product at the moment of placing that order.
* **date**– timestamp for the processing of that product.

**products.csv**

* **sku**– stock keeping unit: a unique identifier for each product.
* **name**– product name.
* **desc**– product description.
* **in\_stock**– whether the product was in stock at the moment of the data extraction.
* **type**– a numerical code for product type.
* **promo\_price**– promotional Price, in euros.

**brands.csv**

* + **short**– the 3-character code by which the brand can be identified in the first 3 characters of products.sku
  + **long**– brand name

🡪First findings:

* 99% of the products in orders with a total paid below 20k are still in the shopping cart.
* Apple, Bose, Jaybird, Mophie, Startech, Unknown are repeated – so they have different short names. 🡪 clean table
* <https://calmcode.io/pandas-pipe/introduction.html>
* <https://tomaugspurger.net/posts/modern-1-intro/>

## 🧩 Keep asking questions

As you know already, exploring the data is critical. We performed an initial exploration using basic pandas methods at the beginning of the project and detected data cleaning & data quality issues. As these issues get solved and the analysis progresses, more concrete questions should arise, for example:

* What is the time period that the dataset covers?
* What is the overall revenue for that time?
* Are there seasonal patterns in the evolution of sales?
* What are the most sold products?
* What are the products that generate the most revenue?

Keep the project goals in mind and be as curious, inquisitive, and creative as possible, asking questions to the data.