💼 Case study

Welcome to your second data project! You will keep working for Eniac —the e-commerce tech company— as a Data Analyst. This time you will work with internal data, which is not anonymised… but a bit more chaotic! The complexity of these tasks will require you to use Python, rather than SQL.

The company has high hopes put into the possibilities that come with Data Analysis, and they are especially hopeful that your work can finally settle an ongoing debate: **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

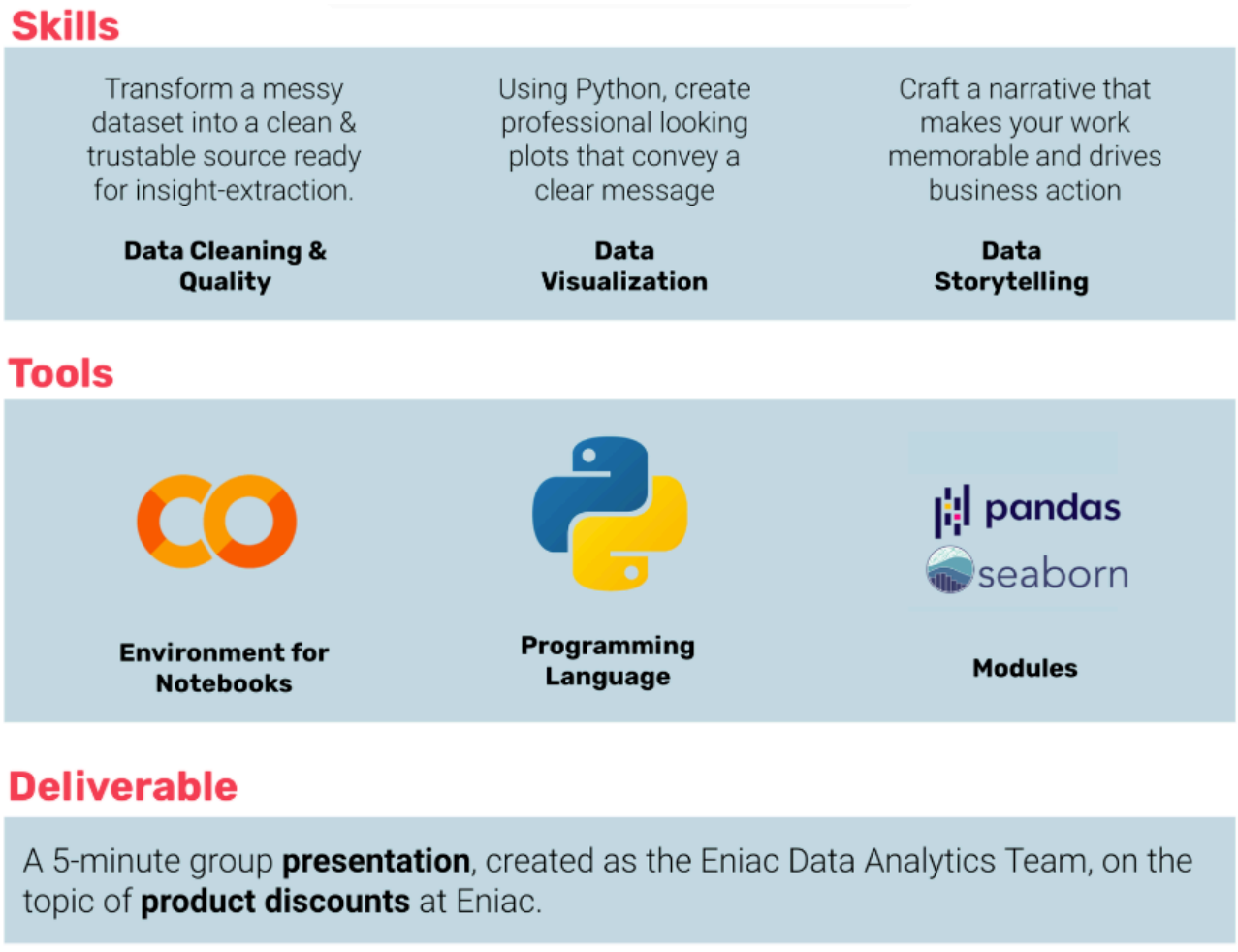
Your analytical and business skills are needed to provide clarity in the following aspects:

- 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?

Project overview

Here's an overview of the project:



Get the data

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
  - **price** – base price of the product, in euros
  - **promo_price** – promotional price, in euros

    - **in_stock** – whether or not the product was in stock at the moment of the data extraction
    - **type** – a numerical code for product type
- `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

🧩 Refine your analysis

1. Re-categorize products

This is a task you could probably spend a hundred hours on and still have some misclassified products. If you have time, try to go for smaller categories (not just “accessories” but “mouse”, “keyboard”, etc.). It’s not a bad idea to have a hierarchy with two or three levels of categories: “accessories” – “storage” – “ssd”.

2. Improve your visualizations

Get fancy with interactive visualizations, make your charts beautiful with Seaborn’s colour palettes or simply make sure that you are using the right plot for whatever you are trying to convey.

Consider especially the following:

- **Your audience will be the board of directors of Eniac.** Adapt to them (e.g. you don’t need to tell them what Eniac sells, but they might want some context on the specific dataset you used to reach your conclusions).

- **Your presentation must not be longer than 5 minutes**. Conciseness is a critical skill we want you to cultivate.
- Make sure you address the main question you were tasked with: **what should Eniac's strategy be concerning discounts?** Tackle other aspects if they're important, and re-formulate the question if you feel it's necessary, but make it the central matter of the presentation.