PYTHON ETL PIPELINE (pandas)

RAW DATA SOURCES

1. products.csv
2. customers.csv
3. sales.csv

EXTRACT DATA

- read_csv()
- Convert date columns
- Initial checks:

.info(),

.shape()

TRANSFORM DATA

CLEANING

- Remove duplicates
- Normalize string values (brand/category)
- Fix types: date → datetime, numeric → float/int
- Handle missing values

ENRICHMENT

- Merge sales + products + customers
- Add calculated fields
- total_revenue
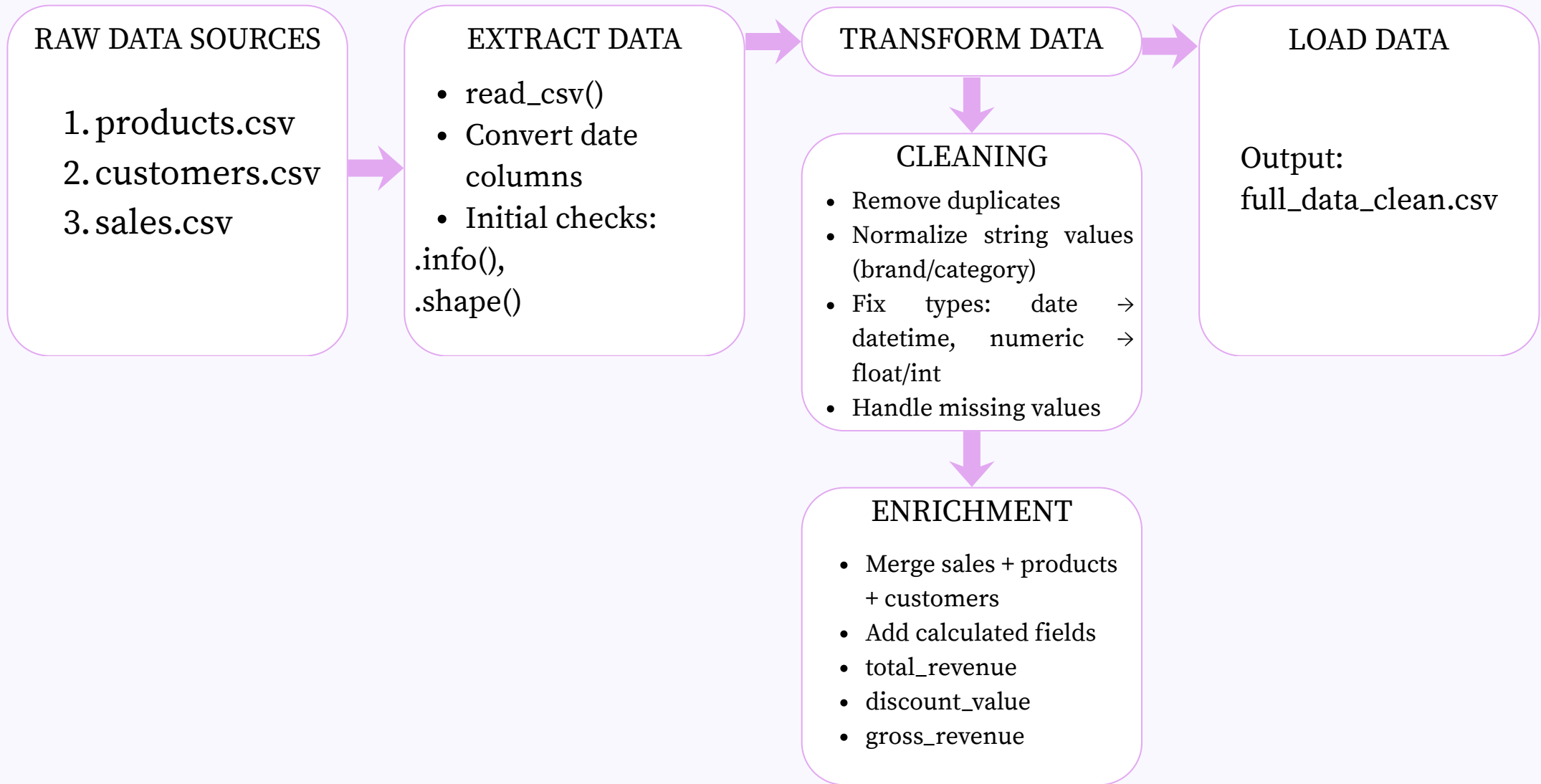- discount_value
- gross_revenue

LOAD DATA

Output:
full_data_clean.csv

PYTHON EDA PIPELINE

IMPORT EDA TOOLS AND DATA

- Pandas
- Matplotlib
- Seaborn
- NumPy
- full_data_clean.csv

DATA OVERVIEW

- .info ()
- .head ()
- .isna().sum()
- .describe()

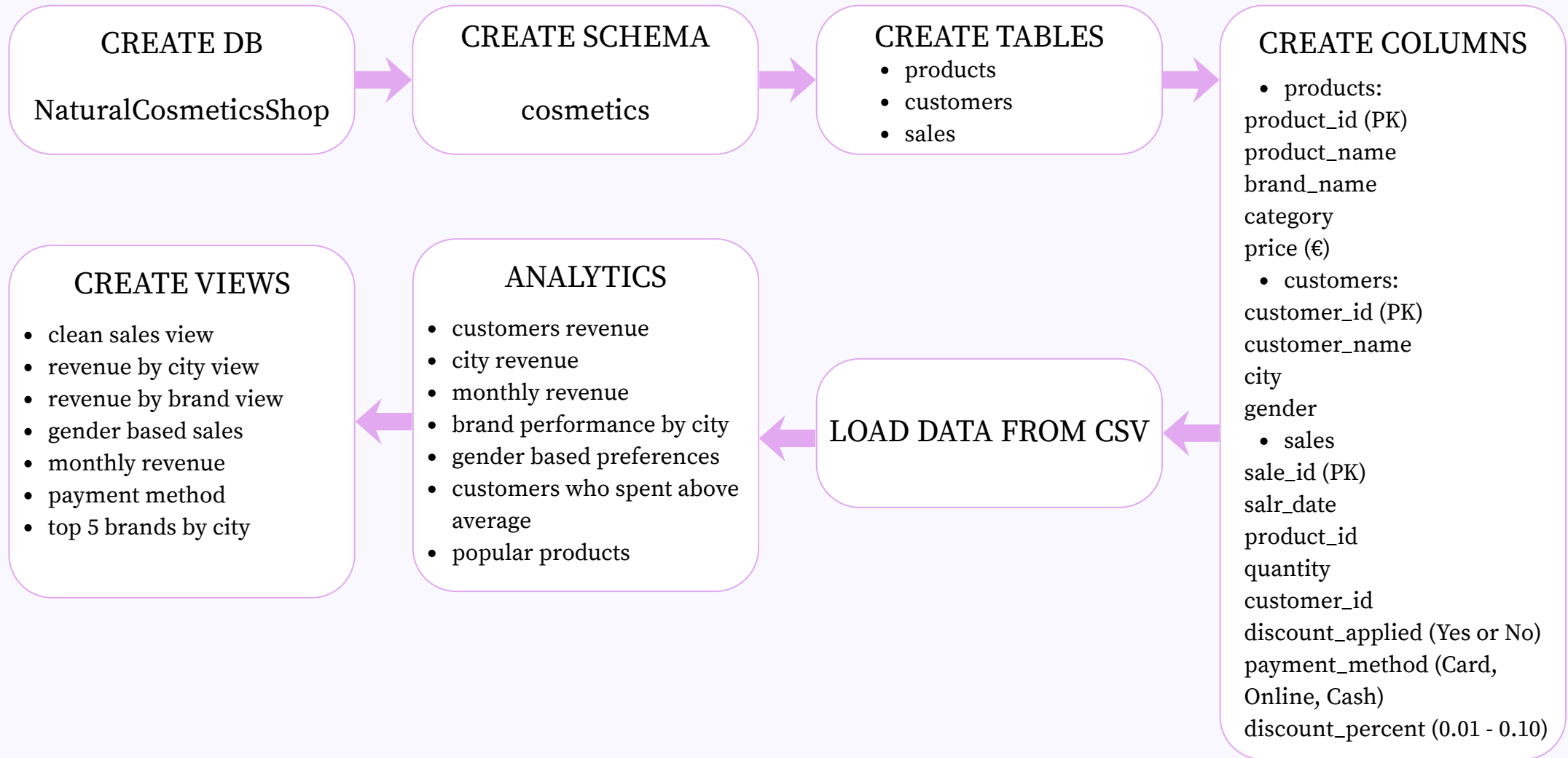**DATA TYPE NORMALIZING AND ADDING 'MONTH' COLUMN (PERIOD)**

VISUALS

- monthly revenue trend
- best selling products
- price vs. quantity sold
- discount rate vs. revenue
- avg revenue by discount
- revenue & discount trends
- correlation heatmap
- customers clustering
- top brands

KEY INSIGHTS

- **Price vs. quantity sold**: The scatterplot shows no clear relationship between product price and quantity sold. Customers consistently buy 1–3 units regardless of price.
- **Discount vs. revenue**: There is no clear relationship between the size of the discount and revenue generated. All discount percent values cluster between 0–10%, and even within these ranges, revenue values are widely scattered.
- **Correlation heatmap**: The correlation heatmap does not reveal any strong relationships between most variables. Only price and revenue-related fields show strong correlation, which makes sense because these values depend on each other. Everything else shows almost no correlation.

SQL (POSTGRESQL) PIPELINE

CREATE DB

NaturalCosmeticsShop

CREATE SCHEMA

cosmetics

CREATE TABLES
- products
- customers
- sales

CREATE COLUMNS
- products:
product_id (PK)
product_name
brand_name
category
price (€)
- customers:
customer_id (PK)
customer_name
city
gender
- sales
sale_id (PK)
salr_date
product_id
quantity
customer_id
discount_applied (Yes or No)
payment_method (Card, Online, Cash)
discount_percent (0.01 - 0.10)

CREATE VIEWS
- clean sales view
- revenue by city view
- revenue by brand view
- gender based sales
- monthly revenue
- payment method
- top 5 brands by city

ANALYTICS
- customers revenue
- city revenue
- monthly revenue
- brand performance by city
- gender based preferences
- customers who spent above average
- popular products

LOAD DATA FROM CSV

POWER BI PIPELINE

CONNECT TO POSTGRESQL & IMPORT VIEWS

↓

CREATE VISUALS →

REVENUE VISUALS

- Gender based revenue
- Revenue by payment method
- Revenue by month
- Revenue by brand
- Revenue by category
- Top 5 brands by city

→

QUANTITY VISUALS

- Customers by gender
- Sales by payment method
- Sales by month
- Sales by brand
- Sales by category

→

DASHBOARD

↓

INSIGHTS

- Gender based revenue

Women spent €180K, men - €152K → €28K (~15.6%) less than women

- Payment method based revenue

Online - €117K (~35.2%), by card - €111K(~33.4%), in cash - €105K (~31.4%)

- Month based revenue

Months with highest revenue: May, October (€30K), and lowest revenue: September (€25K)

- Category based revenue

Category with highest revenue: Serums, Face oils & Targeted treatments(€27K), and lowest revenue: Eye & Lip Care/Specialized areas (€18K)

- Revenue by brand

Brand with highes revenue: Vintner's Daughter (€41K) and lowest revenue: Barulab (€469)

↓

INSIGHTS

- Customers quantity

Women spent 174, men 151

- Payment method based sales quantity

Online - 1792 , by card - 1621, in cash -1587

- Month based sales quantity

Month with highest sales quantity: January (439), and lowest quantity: April (390)

- Category based sales quantity

Category with highest sales quantity:Masks, exofoliants & facal treatments(974), and lowest revenue: Hair Care (415)

- Sales count by brand

Brand with highes sales count: The Ordinary (332) and lowest revenue: Camille Rose (30)