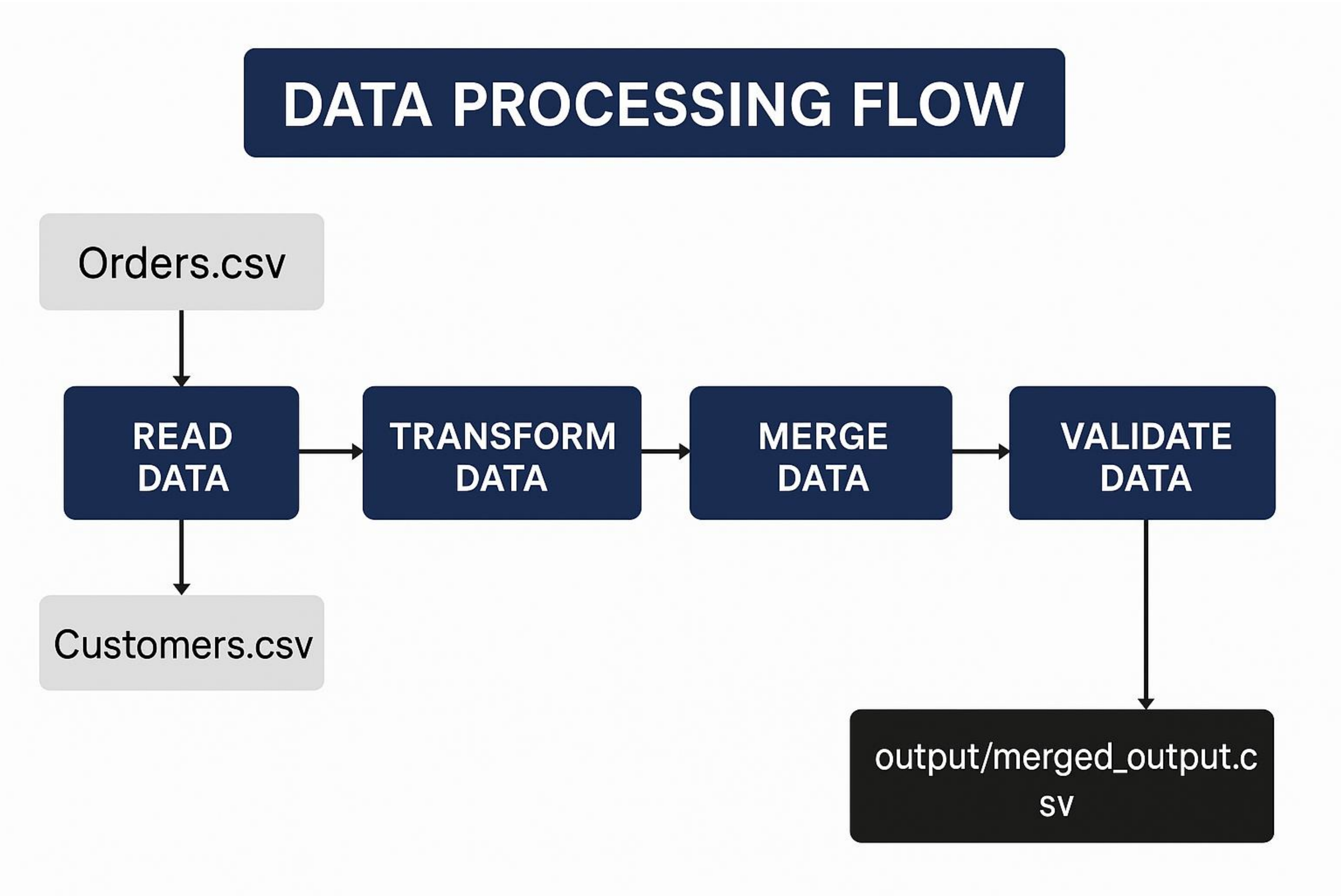
Data Processing Flow



DATA PROCESSING FLOW

Orders.csv

READ DATA

TRANSFORM DATA

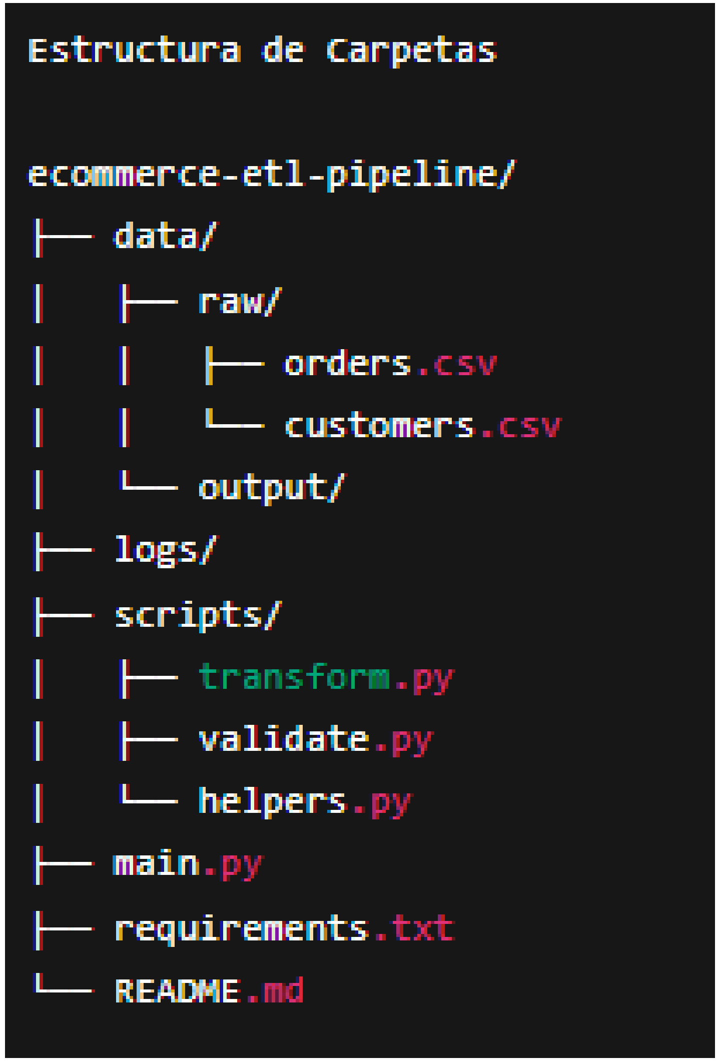
MERGE DATA

VALIDATE DATA

Customers.csv

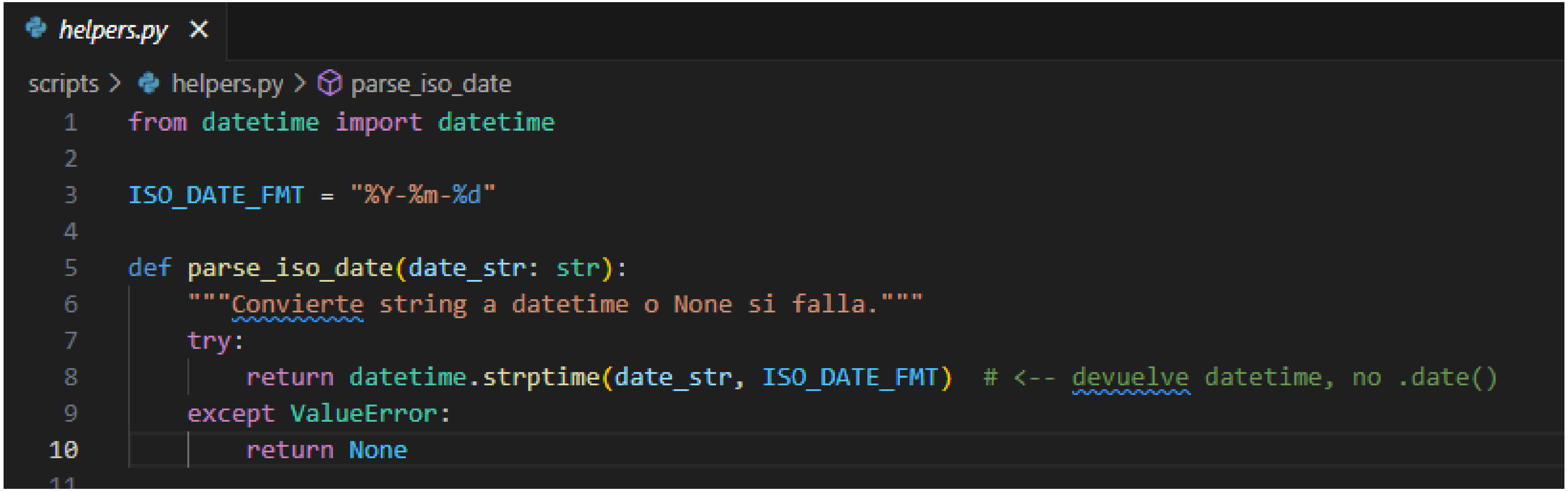
output/merged_output.csv

Folders Structure



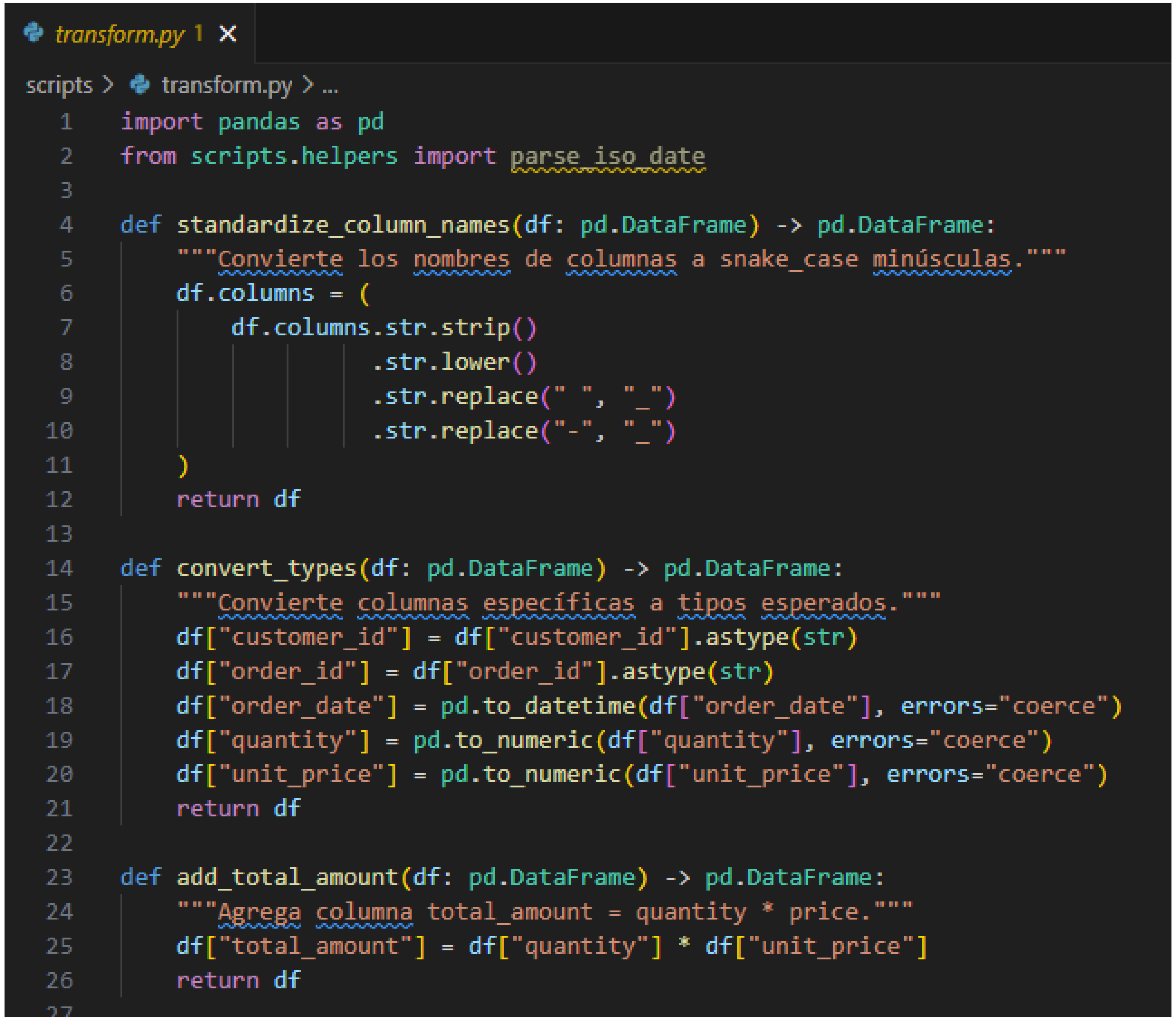
```
Estructura de Carpetas

ecommerce-etl-pipeline/
├── data/
|   ├── raw/
|   |   ├── orders.csv
|   |   └── customers.csv
|   └── output/
├── logs/
├── scripts/
|   ├── transform.py
|   ├── validate.py
|   └── helpers.py
├── main.py
├── requirements.txt
└── README.md
```

Helpers.py



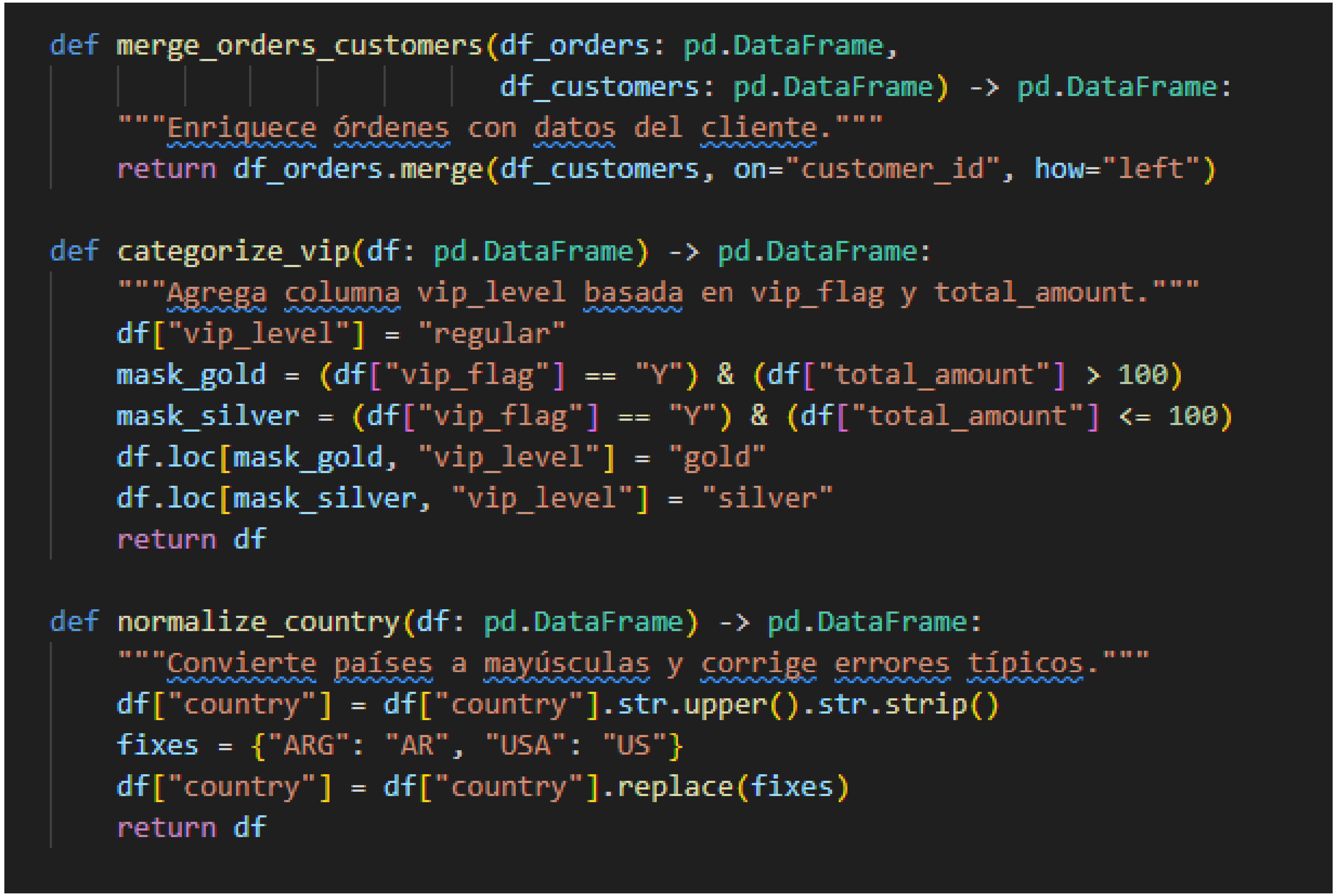
```python
from datetime import datetime

ISO_DATE_FMT = "%Y-%m-%d"

def parse_iso_date(date_str: str):
    """Convierte string a datetime o None si falla."""
    try:
        return datetime.strptime(date_str, ISO_DATE_FMT)  # <-- devuelve datetime, no .date()
    except ValueError:
        return None
```

Transform.py (part one)



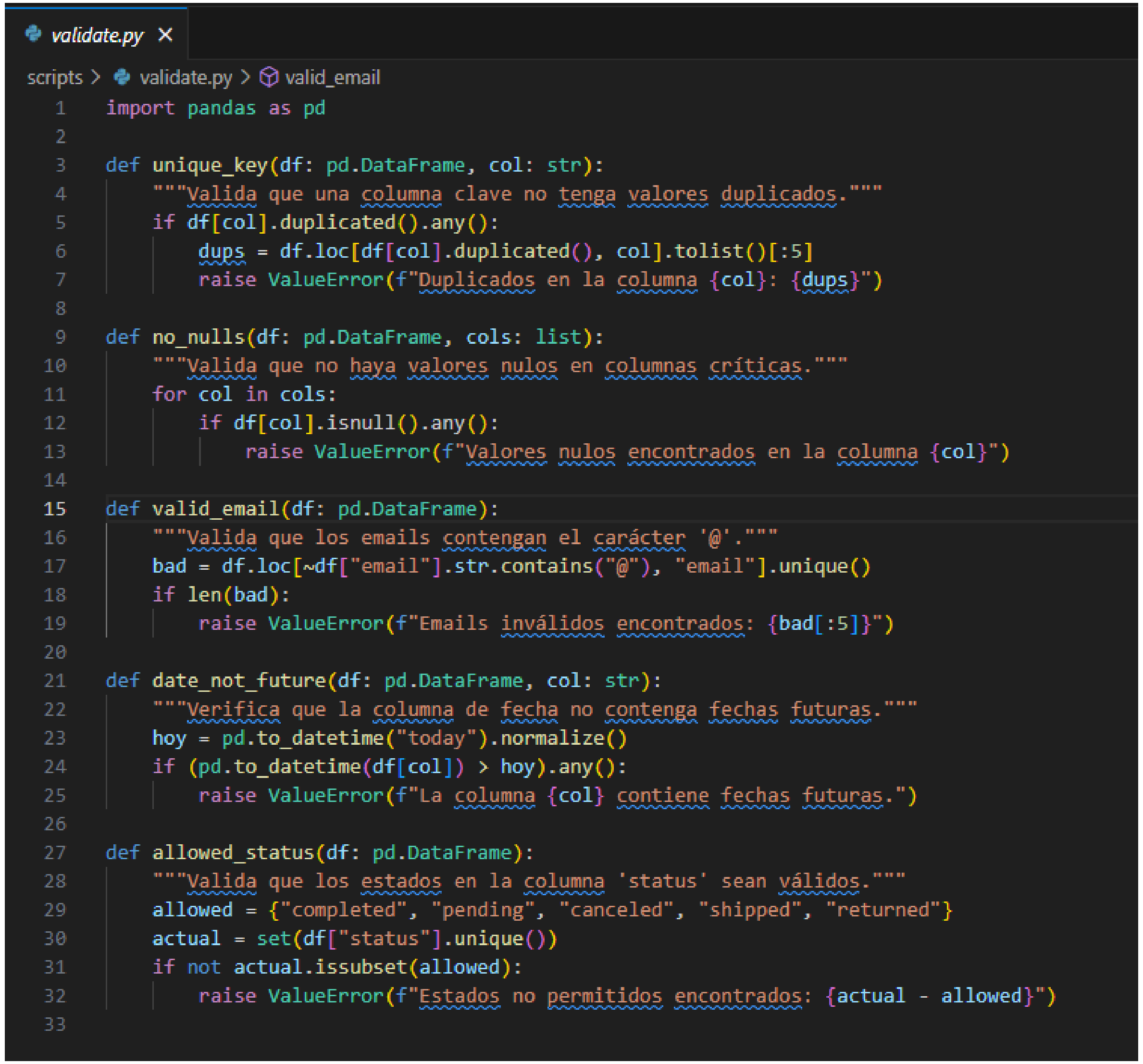
```python
import pandas as pd
from scripts.helpers import parse_iso_date

def standardize_column_names(df: pd.DataFrame) -> pd.DataFrame:
    """Convierte los nombres de columnas a snake_case minúsculas."""
    df.columns = (
        df.columns.str.strip()
                  .str.lower()
                  .str.replace(" ", "_")
                  .str.replace("-", "_")
    )
    return df

def convert_types(df: pd.DataFrame) -> pd.DataFrame:
    """Convierte columnas específicas a tipos esperados."""
    df["customer_id"] = df["customer_id"].astype(str)
    df["order_id"] = df["order_id"].astype(str)
    df["order_date"] = pd.to_datetime(df["order_date"], errors="coerce")
    df["quantity"] = pd.to_numeric(df["quantity"], errors="coerce")
    df["unit_price"] = pd.to_numeric(df["unit_price"], errors="coerce")
    return df

def add_total_amount(df: pd.DataFrame) -> pd.DataFrame:
    """Agrega columna total_amount = quantity * price."""
    df["total_amount"] = df["quantity"] * df["unit_price"]
    return df
```

Transform.py (part two)



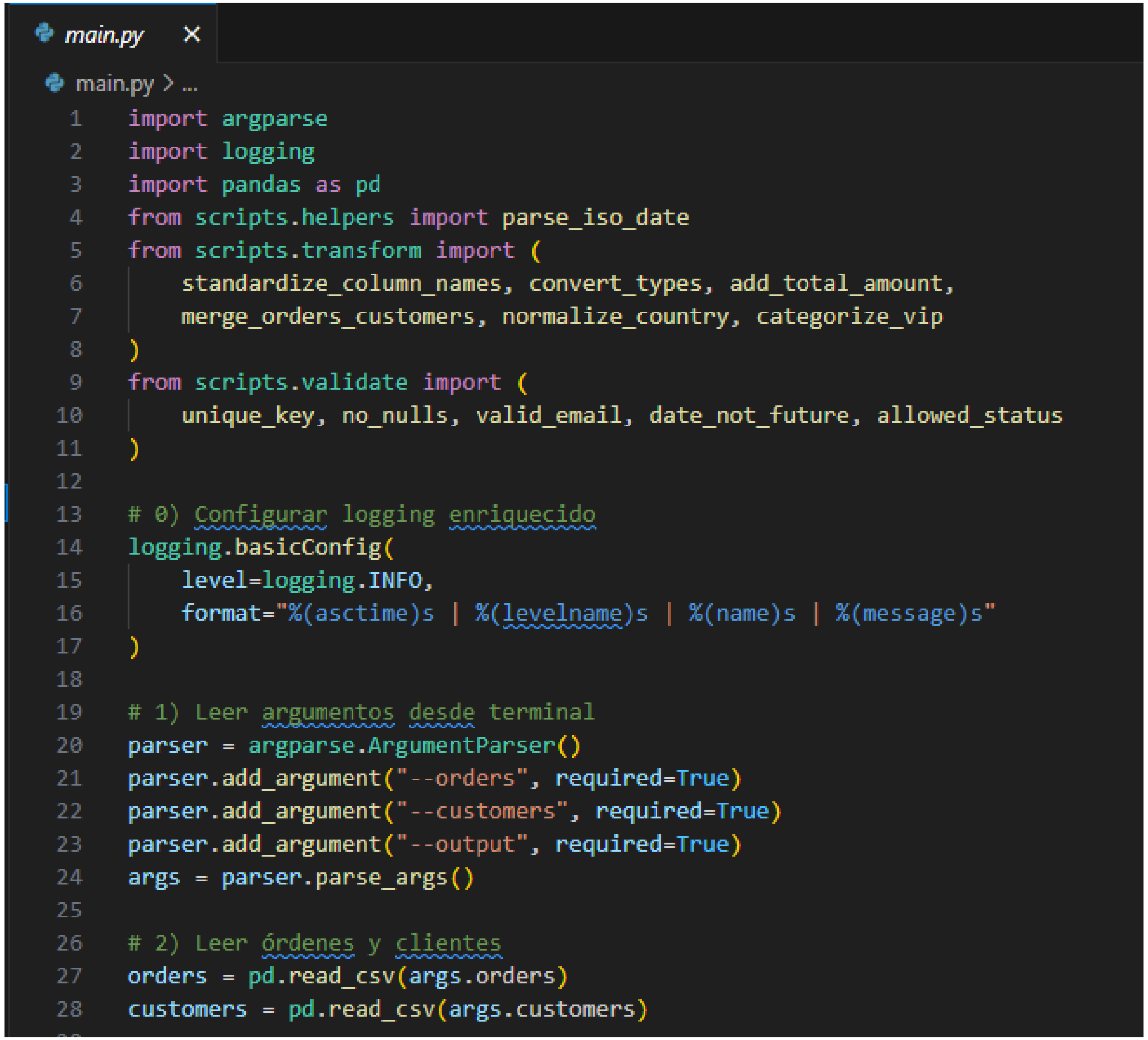
```python
def merge_orders_customers(df_orders: pd.DataFrame,
                           df_customers: pd.DataFrame) -> pd.DataFrame:
    """Enriquece órdenes con datos del cliente."""
    return df_orders.merge(df_customers, on="customer_id", how="left")

def categorize_vip(df: pd.DataFrame) -> pd.DataFrame:
    """Agrega columna vip_level basada en vip_flag y total_amount."""
    df["vip_level"] = "regular"
    mask_gold = (df["vip_flag"] == "Y") & (df["total_amount"] > 100)
    mask_silver = (df["vip_flag"] == "Y") & (df["total_amount"] <= 100)
    df.loc[mask_gold, "vip_level"] = "gold"
    df.loc[mask_silver, "vip_level"] = "silver"
    return df

def normalize_country(df: pd.DataFrame) -> pd.DataFrame:
    """Convierte países a mayúsculas y corrige errores típicos."""
    df["country"] = df["country"].str.upper().str.strip()
    fixes = {"ARG": "AR", "USA": "US"}
    df["country"] = df["country"].replace(fixes)
    return df
```

Validate.py



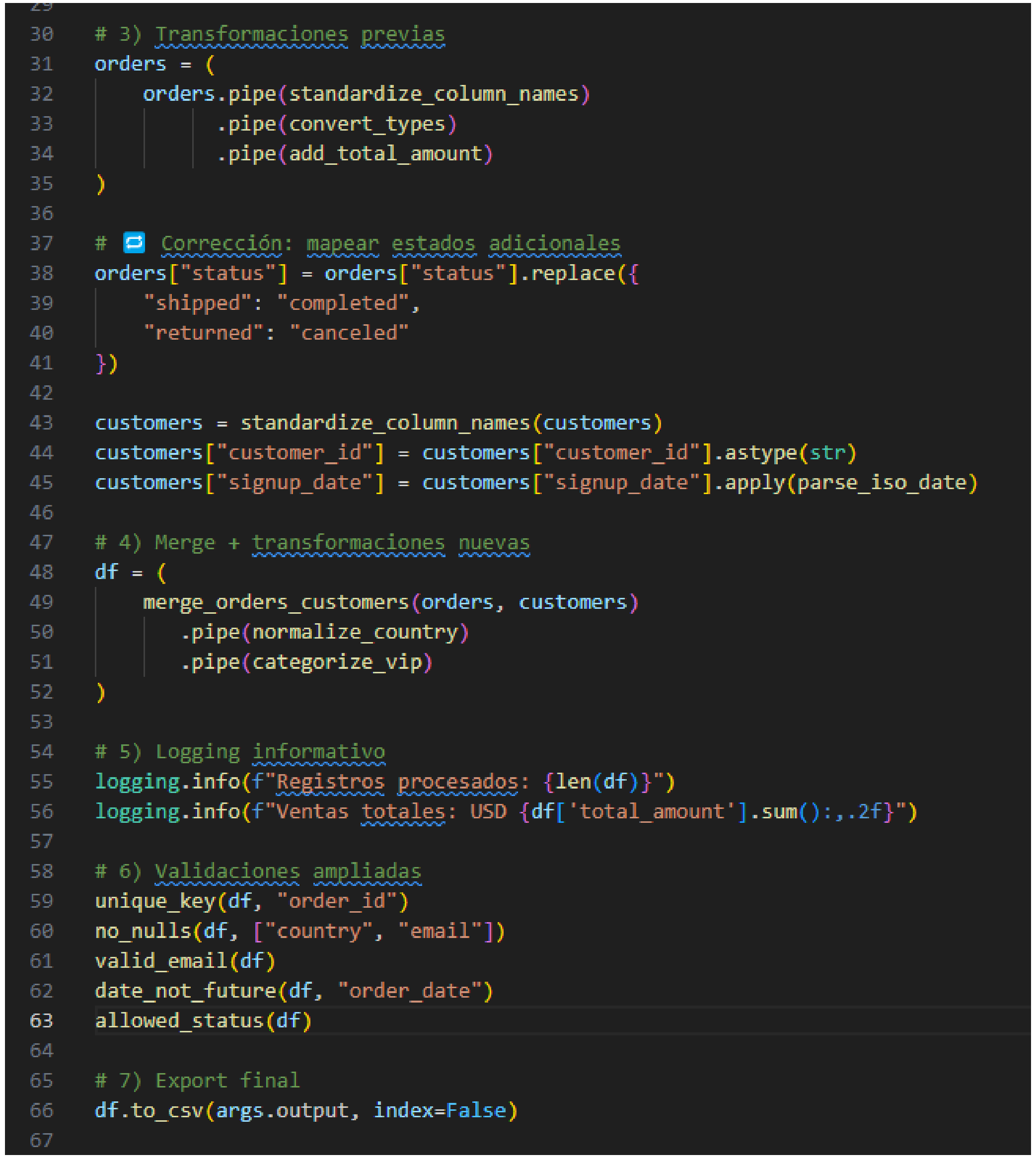
```python
import pandas as pd

def unique_key(df: pd.DataFrame, col: str):
    """Valida que una columna clave no tenga valores duplicados."""
    if df[col].duplicated().any():
        dups = df.loc[df[col].duplicated(), col].tolist()[:5]
        raise ValueError(f"Duplicados en la columna {col}: {dups}")

def no_nulls(df: pd.DataFrame, cols: list):
    """Valida que no haya valores nulos en columnas críticas."""
    for col in cols:
        if df[col].isnull().any():
            raise ValueError(f"Valores nulos encontrados en la columna {col}")

def valid_email(df: pd.DataFrame):
    """Valida que los emails contengan el carácter '@'."""
    bad = df.loc[~df["email"].str.contains("@"), "email"].unique()
    if len(bad):
        raise ValueError(f"Emails inválidos encontrados: {bad[:5]}")

def date_not_future(df: pd.DataFrame, col: str):
    """Verifica que la columna de fecha no contenga fechas futuras."""
    hoy = pd.to_datetime("today").normalize()
    if (pd.to_datetime(df[col]) > hoy).any():
        raise ValueError(f"La columna {col} contiene fechas futuras.")

def allowed_status(df: pd.DataFrame):
    """Valida que los estados en la columna 'status' sean válidos."""
    allowed = {"completed", "pending", "canceled", "shipped", "returned"}
    actual = set(df["status"].unique())
    if not actual.issubset(allowed):
        raise ValueError(f"Estados no permitidos encontrados: {actual - allowed}")

```

Main.py (part one)



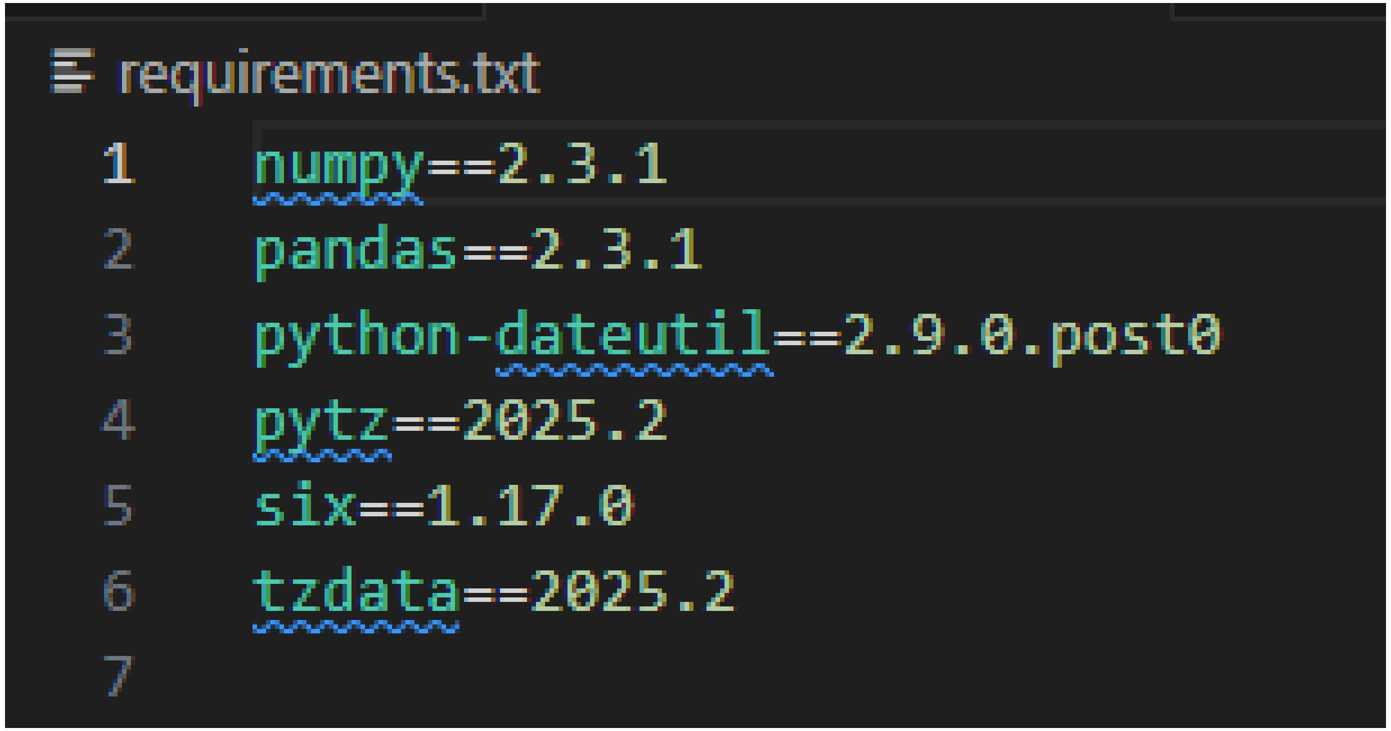
```python
import argparse
import logging
import pandas as pd
from scripts.helpers import parse_iso_date
from scripts.transform import (
    standardize_column_names, convert_types, add_total_amount,
    merge_orders_customers, normalize_country, categorize_vip
)
from scripts.validate import (
    unique_key, no_nulls, valid_email, date_not_future, allowed_status
)

# 0) Configurar logging enriquecido
logging.basicConfig(
    level=logging.INFO,
    format="%(asctime)s | %(levelname)s | %(name)s | %(message)s"
)

# 1) Leer argumentos desde terminal
parser = argparse.ArgumentParser()
parser.add_argument("--orders", required=True)
parser.add_argument("--customers", required=True)
parser.add_argument("--output", required=True)
args = parser.parse_args()

# 2) Leer órdenes y clientes
orders = pd.read_csv(args.orders)
customers = pd.read_csv(args.customers)
```

Main.py (part two)



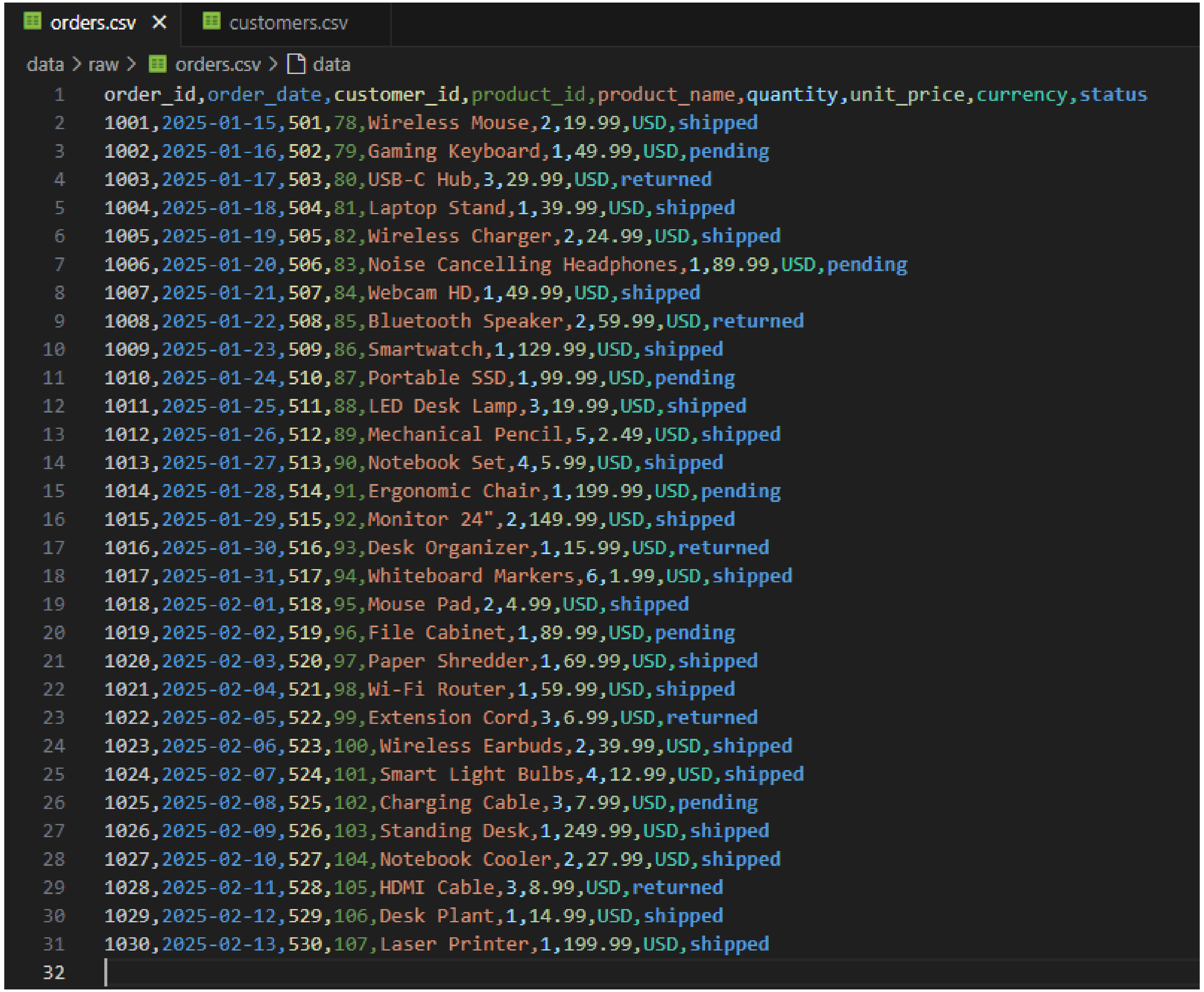
```python
# 3) Transformaciones previas
orders = (
    orders.pipe(standardize_column_names)
          .pipe(convert_types)
          .pipe(add_total_amount)
)

# 🔁 Corrección: mapear estados adicionales
orders["status"] = orders["status"].replace({
    "shipped": "completed",
    "returned": "canceled"
})

customers = standardize_column_names(customers)
customers["customer_id"] = customers["customer_id"].astype(str)
customers["signup_date"] = customers["signup_date"].apply(parse_iso_date)

# 4) Merge + transformaciones nuevas
df = (
    merge_orders_customers(orders, customers)
       .pipe(normalize_country)
       .pipe(categorize_vip)
)

# 5) Logging informativo
logging.info(f"Registros procesados: {len(df)}")
logging.info(f"Ventas totales: USD {df['total_amount'].sum():,.2f}")

# 6) Validaciones ampliadas
unique_key(df, "order_id")
no_nulls(df, ["country", "email"])
valid_email(df)
date_not_future(df, "order_date")
allowed_status(df)

# 7) Export final
df.to_csv(args.output, index=False)
```

Requirements.txt



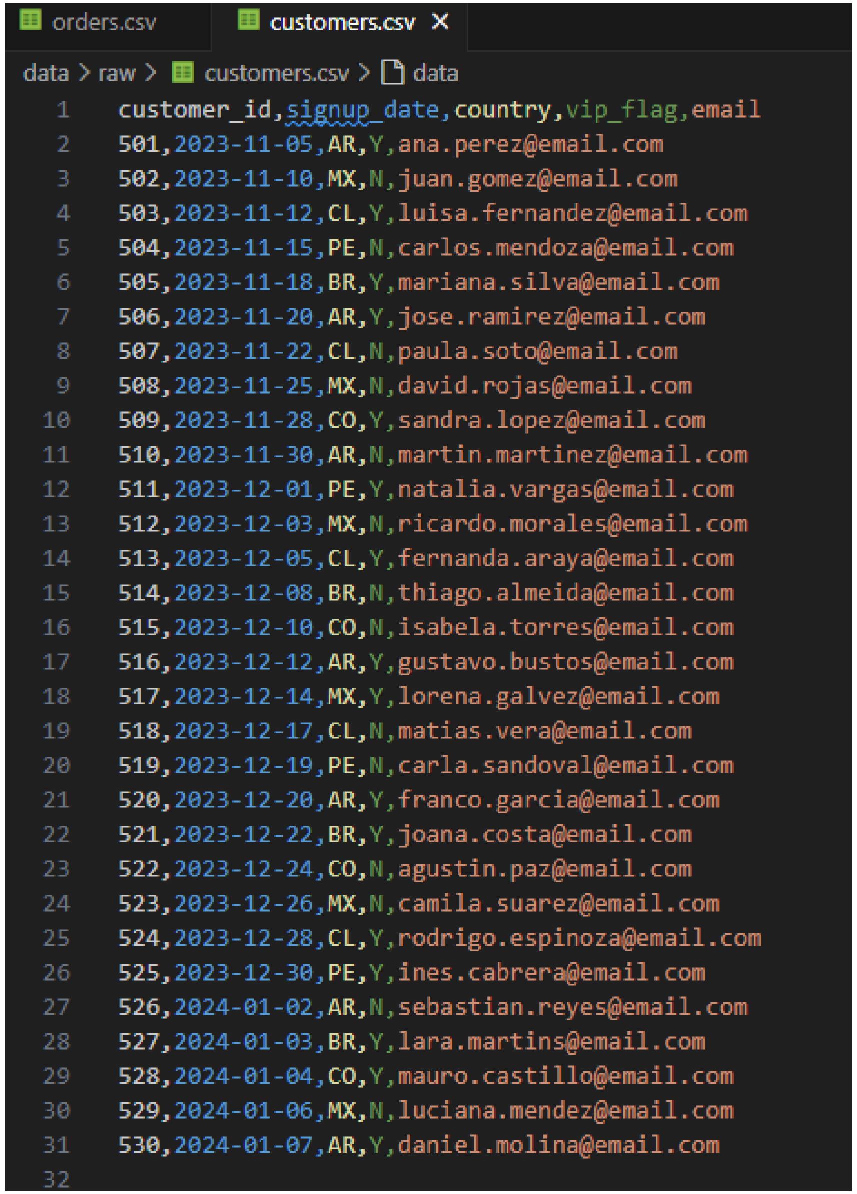
```
numpy==2.3.1
pandas==2.3.1
python-dateutil==2.9.0.post0
pytz==2025.2
six==1.17.0
tzdata==2025.2
```

Orders.csv (raw)



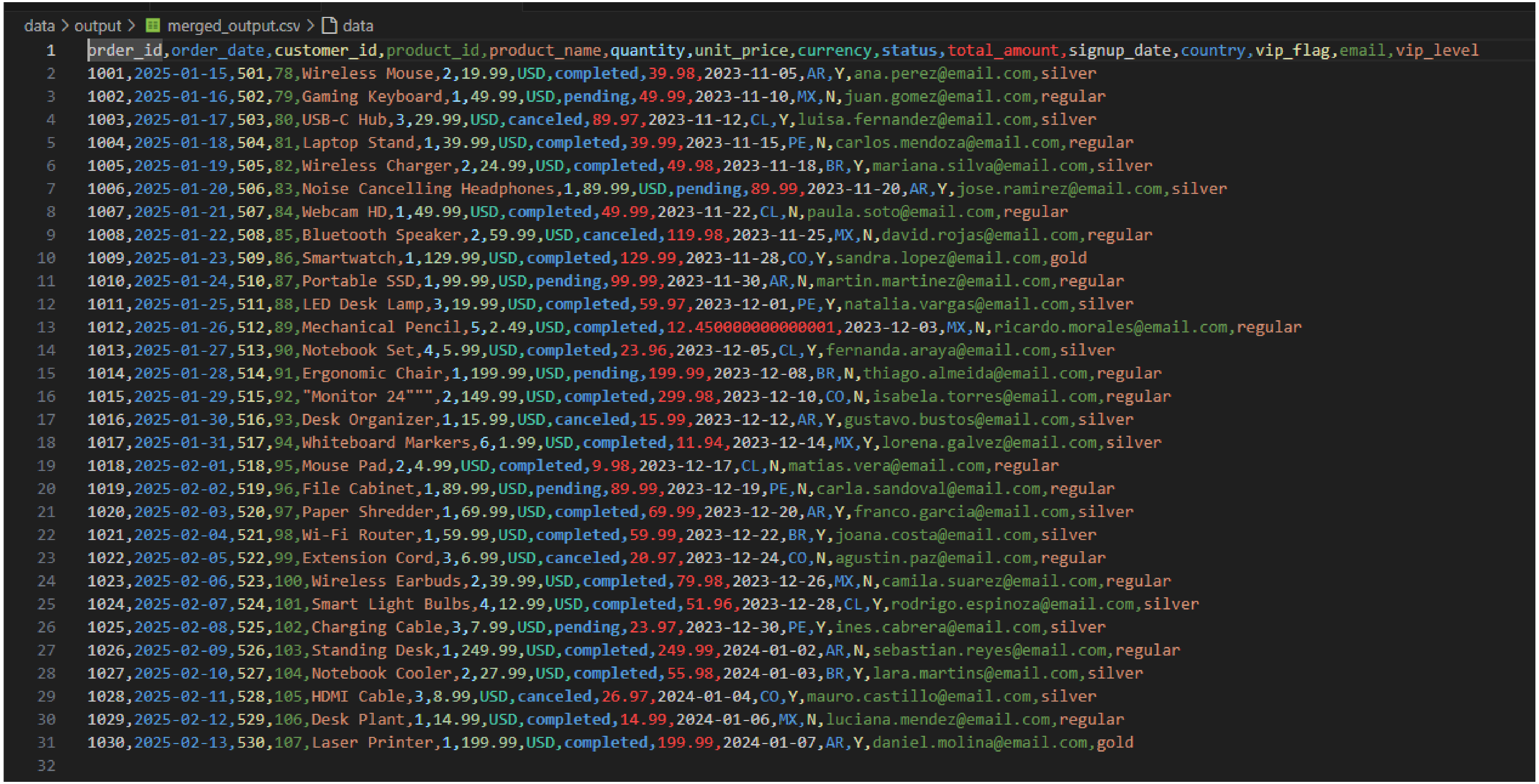
```
order_id,order_date,customer_id,product_id,product_name,quantity,unit_price,currency,status
1001,2025-01-15,501,78,Wireless Mouse,2,19.99,USD,shipped
1002,2025-01-16,502,79,Gaming Keyboard,1,49.99,USD,pending
1003,2025-01-17,503,80,USB-C Hub,3,29.99,USD,returned
1004,2025-01-18,504,81,Laptop Stand,1,39.99,USD,shipped
1005,2025-01-19,505,82,Wireless Charger,2,24.99,USD,shipped
1006,2025-01-20,506,83,Noise Cancelling Headphones,1,89.99,USD,pending
1007,2025-01-21,507,84,Webcam HD,1,49.99,USD,shipped
1008,2025-01-22,508,85,Bluetooth Speaker,2,59.99,USD,returned
1009,2025-01-23,509,86,Smartwatch,1,129.99,USD,shipped
1010,2025-01-24,510,87,Portable SSD,1,99.99,USD,pending
1011,2025-01-25,511,88,LED Desk Lamp,3,19.99,USD,shipped
1012,2025-01-26,512,89,Mechanical Pencil,5,2.49,USD,shipped
1013,2025-01-27,513,90,Notebook Set,4,5.99,USD,shipped
1014,2025-01-28,514,91,Ergonomic Chair,1,199.99,USD,pending
1015,2025-01-29,515,92,Monitor 24",2,149.99,USD,shipped
1016,2025-01-30,516,93,Desk Organizer,1,15.99,USD,returned
1017,2025-01-31,517,94,Whiteboard Markers,6,1.99,USD,shipped
1018,2025-02-01,518,95,Mouse Pad,2,4.99,USD,shipped
1019,2025-02-02,519,96,File Cabinet,1,89.99,USD,pending
1020,2025-02-03,520,97,Paper Shredder,1,69.99,USD,shipped
1021,2025-02-04,521,98,Wi-Fi Router,1,59.99,USD,shipped
1022,2025-02-05,522,99,Extension Cord,3,6.99,USD,returned
1023,2025-02-06,523,100,Wireless Earbuds,2,39.99,USD,shipped
1024,2025-02-07,524,101,Smart Light Bulbs,4,12.99,USD,shipped
1025,2025-02-08,525,102,Charging Cable,3,7.99,USD,pending
1026,2025-02-09,526,103,Standing Desk,1,249.99,USD,shipped
1027,2025-02-10,527,104,Notebook Cooler,2,27.99,USD,shipped
1028,2025-02-11,528,105,HDMI Cable,3,8.99,USD,returned
1029,2025-02-12,529,106,Desk Plant,1,14.99,USD,shipped
1030,2025-02-13,530,107,Laser Printer,1,199.99,USD,shipped

```

Customers.csv (raw)



```
customer_id,signup_date,country,vip_flag,email
501,2023-11-05,AR,Y,ana.perez@email.com
502,2023-11-10,MX,N,juan.gomez@email.com
503,2023-11-12,CL,Y,luisa.fernandez@email.com
504,2023-11-15,PE,N,carlos.mendoza@email.com
505,2023-11-18,BR,Y,mariana.silva@email.com
506,2023-11-20,AR,Y,jose.ramirez@email.com
507,2023-11-22,CL,N,paula.soto@email.com
508,2023-11-25,MX,N,david.rojas@email.com
509,2023-11-28,CO,Y,sandra.lopez@email.com
510,2023-11-30,AR,N,martin.martinez@email.com
511,2023-12-01,PE,Y,natalia.vargas@email.com
512,2023-12-03,MX,N,ricardo.morales@email.com
513,2023-12-05,CL,Y,fernanda.araya@email.com
514,2023-12-08,BR,N,thiago.almeida@email.com
515,2023-12-10,CO,N,isabela.torres@email.com
516,2023-12-12,AR,Y,gustavo.bustos@email.com
517,2023-12-14,MX,Y,lorena.galvez@email.com
518,2023-12-17,CL,N,matias.vera@email.com
519,2023-12-19,PE,N,carla.sandoval@email.com
520,2023-12-20,AR,Y,franco.garcia@email.com
521,2023-12-22,BR,Y,joana.costa@email.com
522,2023-12-24,CO,N,agustin.paz@email.com
523,2023-12-26,MX,N,camila.suarez@email.com
524,2023-12-28,CL,Y,rodrigo.espinoza@email.com
525,2023-12-30,PE,Y,ines.cabrera@email.com
526,2024-01-02,AR,N,sebastian.reyes@email.com
527,2024-01-03,BR,Y,lara.martins@email.com
528,2024-01-04,CO,Y,mauro.castillo@email.com
529,2024-01-06,MX,N,luciana.mendez@email.com
530,2024-01-07,AR,Y,daniel.molina@email.com

```

Merged_output.csv



```
order_id,order_date,customer_id,product_id,product_name,quantity,unit_price,currency,status,total_amount,signup_date,country,vip_flag,email,vip_level
1001,2025-01-15,501,78,Wireless Mouse,2,19.99,USD,completed,39.98,2023-11-05,AR,Y,ana.perez@email.com,silver
1002,2025-01-16,502,79,Gaming Keyboard,1,49.99,USD,pending,49.99,2023-11-10,MX,N,juan.gomez@email.com,regular
1003,2025-01-17,503,80,USB-C Hub,3,29.99,USD,canceled,89.97,2023-11-12,CL,Y,luisa.fernandez@email.com,silver
1004,2025-01-18,504,81,Laptop Stand,1,39.99,USD,completed,39.99,2023-11-15,PE,N,carlos.mendoza@email.com,regular
1005,2025-01-19,505,82,Wireless Charger,2,24.99,USD,completed,49.98,2023-11-18,BR,Y,mariana.silva@email.com,silver
1006,2025-01-20,506,83,Noise Cancelling Headphones,1,89.99,USD,pending,89.99,2023-11-20,AR,Y,jose.ramirez@email.com,silver
1007,2025-01-21,507,84,Webcam HD,1,49.99,USD,completed,49.99,2023-11-22,CL,N,paula.soto@email.com,regular
1008,2025-01-22,508,85,Bluetooth Speaker,2,59.99,USD,canceled,119.98,2023-11-25,MX,N,david.rojas@email.com,regular
1009,2025-01-23,509,86,Smartwatch,1,129.99,USD,completed,129.99,2023-11-28,CO,Y,sandra.lopez@email.com,gold
1010,2025-01-24,510,87,Portable SSD,1,99.99,USD,pending,99.99,2023-11-30,AR,N,martin.martinez@email.com,regular
1011,2025-01-25,511,88,LED Desk Lamp,3,19.99,USD,completed,59.97,2023-12-01,PE,Y,natalia.vargas@email.com,silver
1012,2025-01-26,512,89,Mechanical Pencil,5,2.49,USD,completed,12.450000000000001,2023-12-03,MX,N,ricardo.morales@email.com,regular
1013,2025-01-27,513,90,Notebook Set,4,5.99,USD,completed,23.96,2023-12-05,CL,Y,fernanda.araya@email.com,silver
1014,2025-01-28,514,91,Ergonomic Chair,1,199.99,USD,pending,199.99,2023-12-08,BR,N,thiago.almeida@email.com,regular
1015,2025-01-29,515,92,"Monitor 24""",2,149.99,USD,completed,299.98,2023-12-10,CO,N,isabela.torres@email.com,regular
1016,2025-01-30,516,93,Desk Organizer,1,15.99,USD,canceled,15.99,2023-12-12,AR,Y,gustavo.bustos@email.com,silver
1017,2025-01-31,517,94,Whiteboard Markers,6,1.99,USD,completed,11.94,2023-12-14,MX,Y,lorena.galvez@email.com,silver
1018,2025-02-01,518,95,Mouse Pad,2,4.99,USD,completed,9.98,2023-12-17,CL,N,matias.vera@email.com,regular
1019,2025-02-02,519,96,File Cabinet,1,89.99,USD,pending,89.99,2023-12-19,PE,N,carla.sandoval@email.com,regular
1020,2025-02-03,520,97,Paper Shredder,1,69.99,USD,completed,69.99,2023-12-20,AR,Y,franco.garcia@email.com,silver
1021,2025-02-04,521,98,Wi-Fi Router,1,59.99,USD,completed,59.99,2023-12-22,BR,Y,joana.costa@email.com,silver
1022,2025-02-05,522,99,Extension Cord,3,6.99,USD,canceled,20.97,2023-12-24,CO,N,agustin.paz@email.com,regular
1023,2025-02-06,523,100,Wireless Earbuds,2,39.99,USD,completed,79.98,2023-12-26,MX,N,camila.suarez@email.com,regular
1024,2025-02-07,524,101,Smart Light Bulbs,4,12.99,USD,completed,51.96,2023-12-28,CL,Y,rodrigo.espinoza@email.com,silver
1025,2025-02-08,525,102,Charging Cable,3,7.99,USD,pending,23.97,2023-12-30,PE,Y,ines.cabrera@email.com,silver
1026,2025-02-09,526,103,Standing Desk,1,249.99,USD,completed,249.99,2024-01-02,AR,N,sebastian.reyes@email.com,regular
1027,2025-02-10,527,104,Notebook Cooler,2,27.99,USD,completed,55.98,2024-01-03,BR,Y,lara.martins@email.com,silver
1028,2025-02-11,528,105,HDMI Cable,3,8.99,USD,canceled,26.97,2024-01-04,CO,Y,mauro.castillo@email.com,silver
1029,2025-02-12,529,106,Desk Plant,1,14.99,USD,completed,14.99,2024-01-06,MX,N,luciana.mendez@email.com,regular
1030,2025-02-13,530,107,Laser Printer,1,199.99,USD,completed,199.99,2024-01-07,AR,Y,daniel.molina@email.com,gold
```