Analyzing an Online Retail Dataset

Andrea Passerini

Overview

The goal of this project is to design and implement a small but complete data analysis program in Python, starting from a single raw dataset and ending with a set of summary statistics and plots.

1 Dataset

The dataset to be analized is the **Online Retail II** dataset, made available by the UCI Machine Learning Repository under a CC BY 4.0 licence:

https://archive.ics.uci.edu/dataset/502/online+retail+ii

It contains all the transactions of a UK-based online retailer of gifts and homeware between 01/12/2009 and 09/12/2011: roughly one million rows, distributed as an Excel file with one sheet per year.

Columns

Column	Description
Invoice	Invoice number. Starts with the letter C if the transaction is a cancellation
StockCode	Product code. Most are 5-digit numbers, but a few are special ("POST", "D", "M", ...)
Description	Product name
Quantity	Number of units. Negative for returns/cancellations
InvoiceDate	Date and time of the transaction
Price	Unit price, in GBP
Customer ID	Unique customer identifier. Missing for a significant fraction of rows
Country	Customer's country

The dataset is **not** clean: some of the decisions to be made (what to do with missing customer IDs, how to treat cancellations, which stock codes count as "real" products) do not have a single correct answer. Choices should be documented and justified.

2 Tasks

Task 1 – Load and inspect

Load the dataset and report:

- the number of rows and columns;
- the data type of each column;
- the number of missing values per column;

- the number of distinct customers, products, and countries.

Task 2 – Clean the data

- separate cancelled transactions (Invoice numbers starting with "C") from regular sales;
- identify and set aside non-product stock codes (postage, discounts, manual adjustments, etc.) – decide, and justify, whether to keep or exclude them from the product-level analysis;
- decide what to do with rows that have a missing customer ID (they can still be used for product/country-level statistics, but not for customer-level ones);
- make sure the invoice date is treated as an actual date/time, not plain text.

Task 3 – Engineer new features

Compute at least the following, for later use:

- the revenue of each transaction line (quantity times unit price);
- the year, month, and weekday of each transaction.

Task 4 – Aggregate and summarize

Compute:

- total revenue per country, sorted from highest to lowest;
- the top 10 products by total revenue;
- total revenue per month, over the whole period;
- for each customer: number of orders, number of items purchased, and total amount spent, sorted by total amount spent.

Task 5 – Visualize

Produce the following visualizations:

- a bar chart of total revenue by country (top 10);
- a line chart of total revenue by month over the whole period;
- a histogram of total amount spent per customer.

Task 6 – Optional / bonus

Reorganize the analysis so that it can be reused on a different file without copy-pasting code: someone else should be able to point the program at another, similarly-structured retail dataset and get the same kind of report.