Big Data - Functional Data Analysis

Final project

Walmart dataset: a functional perspective

Rémi Perrichon

Introduction

You have been hired by Walmart as a junior data scientist. To inform future decision-making on the group's strategy, your manager asks you to analyze weekly sales from 2010-02-05 to 2012-10-26 for multiple stores and departments[1].

Your main objective is to provide a comprehensive analysis of the temporal dynamics of weekly sales. You should present an overview of the sales as well as one or more detailed analyses, focusing on a specific time period of interest, a particular store, or a specific product category. You should justify the relevance of Functional Data Analysis (FDA) for your study and the working assumptions you adopt.

Any method should be motivated, clearly explained and elegantly implemented. Upper management trusts your intellectual honesty: you may fail to master some advanced methods, you can make some simplifying assumptions, yet, there is no way to sweep things under the carpet. Critical thinking is key. Any external reference should be included in a bibliography and accessible upon request.

The use of AI for coding and/or writing the report is allowed, but you must explicitly acknowledge its use. In all cases, your work is your responsibility, and you must be able to answer questions about your approach or any method you use.

Elementary statistical errors will be heavily penalized (e.g., using a correlation coefficient on categorical variables, omitting legends on graphs, etc.).

Brief data description

The original data comes from Kaggle[2]. The dataset provided for this project is a sample of the original dataset, with fewer variables: it is the one recommended for your analysis.

[1] At Walmart, stores are divided into departments to organize products and services efficiently. Departments group similar products together, making it easier for customers to find items and for employees to manage inventory.

[2] https://www.kaggle.com/datasets/ujjwalchowdhury/walmartcleaned

In the `walmart_cleaned_subset.csv` file, weekly sales are recorded:

Name of the variable	Description
Store	The store number
Date	The week of sales
IsHoliday	Binary variable - whether the week is a special holiday week (1) or not (0)
Dept	Department number in each store
Weekly_Sales	Weekly sales in USD

The department numbers' meanings appear to be available online. For example, department number 1 likely refers to 'candy and tobacco.' It is recommended to adopt such nomenclature for the project, especially to ease the interpretation of the results. You must clearly specify the nomenclature you have selected. Of course, you are free to group categories, select only a subset, etc., as long as it is justified.

The holiday events you should consider took place in the following weeks:

Holiday	Dates
Super Bowl	12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labour Day	10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
Thanksgiving	26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
Christmas	31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Mission statement and deadlines

You are expected to provide 3 elements: a R script with your code (i) with a companion Microsoft Word or Latex small report (ii) and some slides for your oral presentation (iii).

The R script and the short companion report (deadline: to be determined by the end of February 2025)

Your R script (no R markdown nor R notebook) should be concise yet very well organized. Don't hesitate to make sections and subsections. Comments are expected.

Sections in the code are fully explained in the companion report (Microsoft Word or Latex). Interpretations are expected in the report. The companion report can only be few pages but special care must be given to spelling and syntax. Make short and insightful sentences.

It is advised to follow the outline that is presented in this document: doing so, you won't forget an important part of the statistical analysis. You are free to organize the sections as you like. The questions are only here to guide you.

Slides for the oral presentation (deadline: a few days before the oral presentation)

You have to send your slides by email (PDF format only !). If you choose to present with animations, the slides must still be submitted as a PDF (check 'include each composition step').

Oral defence (date: to be determined for each group)

Your presentation should last 15 minutes and highlights some part(s) of your analysis. You should focus on the original aspects of your work.

Outline of the analysis

Exploring data

Usual exploration of the raw data.

- Time, place, scope, economic context.
- Assumptions on the data collection, possible biases and limitations.
- Missing values.
- Descriptive statistics.
- Interest of the functional approach.

Data smoothing

You may focus on a single store and/or department to first highlight the method you have chosen (or aggregate by department and/or store of course).

- Discussion of interpolation / smoothing strategies. Interest of the smoothing approach.
- Illustration.
- Generalization.

Registration

- Would registration be helpful here?
- How would you use holiday events to set up an interesting registration problem for a given store and department ?
- Implementation.
- Do descriptive statistics change with registration? Why? Why not?

FPCA

- Would Functional Principal Component Analysis be helpful here?
- Implementation.

Clustering