Fundamentals of Data Science

Small project - Semester B 2024/25 (30 points, 30%)

1 Introduction

You are given a dataset containing sales data for a superstore somewhere in Europe (see Section 2 below). You will need to write a Python code analysing your dataset, producing two specific plots and calculating two specific values (see Section 3 below). Then you will need to describe your results in a short report (see Section 4). You will need to submit your report, your code, and the obtained values via Studynet/Canvas (see Section 5).

Please, read this brief in full. The dataset you have to use and the analysis you have to perform will be different for different students.

2 Which dataset to use

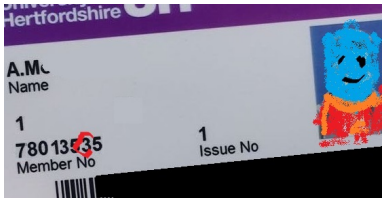

Figure 1: Second to last digit of your student ID card number

The dataset you will need to use depends on the **SECOND TO LAST digit** of your student ID card number, namely:

- **If this digit is "0"**, please, use dataset "sales0.csv"
- **If this digit is "1"**, please, use dataset "sales1.csv"
- **If this digit is "2"**, please, use dataset "sales2.csv"
- **If this digit is "3"**, please, use dataset "sales3.csv"
- **If this digit is "4"**, please, use dataset "sales4.csv"
- **If this digit is "5"**, please, use dataset "sales5.csv"
- **If this digit is "6"**, please, use dataset "sales6.csv"
- **If this digit is "7"**, please, use dataset "sales7.csv"
- **If this digit is "8"**, please, use dataset "sales8.csv"
- **If this digit is "9"**, please, use dataset "sales9.csv"

The download links for your dataset and its brief description can be found on the description of this assignment on Canvas.

3 What data analysis needs to be done

All students need to write Python code, which

(A) Reads the data;

(B) Creates a **monthly** plot showing average **daily** number of items sold by the superstore during a **typical year**, and plots it as a bar chart (**Figure 1** in your report). In other words, you need to plot a bar chart with twelve bars corresponding to **one month** from January to December, with the height of each bar showing the average number of items sold daily during that month. The plot must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the plot;

(C) **On the same figure** (Figure 1 in your report) plots the Fourier series (as a curve) approximating daily number of items sold during the first year in your dataset (2022). The plot should have 365 values (one for every day). The series should be limited to the first eight terms;

(D) Creates a scatter plot of average daily prices plotted against numbers of items sold by the superstore in that day (**Figure 2** in your report). The plot must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the plot;

(E) Uses linear regression to find a linear approximation of the "price vs number" (the scatter plot in Figure 2) and plots the obtained linear function **in Figure 2**, over the scatter plot;

(F) Calculates values X and Y (see below), and prints them **in Figure 2**.

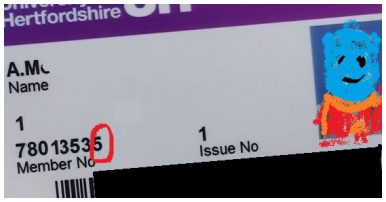

Figure 2: Last digit of your student ID card number

The values X and Y you will need to calculate depend on the **LAST DIGIT** of your student ID card number:

- **If this digit is "0"**, please, evaluate the fraction (in percents) of the superstore's revenue generated during the summer months (value X), and the fraction (in percents) of the superstore's revenue generated during the autumn months (value Y). (By summer months here we mean June, July and August. By autumn months here we mean September, October and November.) Print values X and Y in Figure 2;
- **If this digit is "1"**, please, evaluate the average price of items sold by the superstore during the summer months (value X), and the average price of items sold by the superstore during the summer months (value Y). (By summer months here we mean June, July and August. By autumn months here we mean September, October and November.) Print values X and Y in Figure 2;
- **If this digit is "2"**, please, evaluate the total revenue of the superstore in 2021 (value X), and the total revenue of the superstore in 2022 (value Y). Print values X and Y in Figure 2;
- **If this digit is "3"**, please, evaluate the total revenue of the superstore generated by sales of grocery items in 2021 (value X), and the total revenue of the superstore generated by sales of non-grocery items during the same year (value Y). Print values X and Y in Figure 2;
- **If this digit is "4"**, please, evaluate the fraction of the revenue of the superstore generated by sales of grocery items (value X), and the fraction of the revenue of the superstore generated by sales of non-grocery items (value Y). Print values X and Y in Figure 2;

- **If this digit is "5"**, please, find the day of a typical year when the average prices are usually highest (value X), and the day of a typical year when the average prices are usually lowest (value Y). Both days of the year should be given in form of the day number, i.e. an integer between 1 (corresponding to January 1) and 365 (corresponding to December 31). Print values X and Y in Figure 2;
- **If this digit is "6"**, please, find the day of (a typical) year which generates highest revenue for the superstore (value X), and the day of a typical year when the average prices are usually highest (value Y). The day of the year should be given in form of the day number, i.e. an integer between 1 (corresponding to January 1) and 365 (corresponding to December 31). Print values X and Y in Figure 2;
- **If this digit is "7"**, please, evaluate the total revenue of the superstore generated by sales of grocery items in 2022 (value X), and the total revenue of the superstore generated by sales of grocery items in 2022 (value Y). Print values X and Y in Figure 2;
- **If this digit is "8"**, please, evaluate the fraction of the revenue of the superstore generated by sales of non-grocery items in 2021 (value X), the fraction of the revenue of the superstore generated by sales of non-grocery items in 2022 (value Y). Print values X and Y in Figure 2;
- **If this digit is "9"**, please, evaluate the year-on-year change of average price of grocery items between 2021 and 2022 (value X), and year-on-year change of average price of grocery items between 2022 and 2023 (value Y). Print values X and Y in Figure 2. By year-on-year change of of average price here we mean the difference of the average price in the current year and in the previous year normalised (i.e. divided by) the average price in the previous year;

4 What to include into your report

Write a short report, which

- includes your 8-digit ID number;
- provides a brief description of your dataset (what variables does it contain?, what is the structure of the dataset?) [no more than 1/3 of A4 page];
- includes Figures 1 and 2 produced by your code with short, informative captions [no more than one page, figures should be decipherable, font sizes should be similar to the font sizes used in the report];
- provides the formulas you used to derive the Fourier approximation, values X and Y [no more than 1/2 of A4 page];
- ends with a discussion of the results any conclusions that can be made based on Figures 1 and 2, and values X and Y [minimum half a page, but no more than 1 page].

Your report should be no longer than three A4 pages with 2cm margins; the font should be Arial 11 or similar, with single line spacing. The text and the equations in your report must be machine-readable, i.e. the **text and the formulas cannot be included as images**.

5 What to submit

- Your ID number, **X and Y values**;
- **Your Python code** as a $[IDnumber].py$ file, where $[IDnumber]$ is your 8-digit student ID number;
- **Your report** in PDF format as a $[IDnumber].pdf$ file, where $[IDnumber]$ is your 8-digit student ID number.
- **You discussion** (from your report) as a plain text.

Important:
(a) do not submit any other files;
(b) Colab/Jupyter/etc notebooks are not accepted;
(c) Only files submitted via Canvas are considered;
(d) When rounding numbers, keep at least two significant digits.