Fundamentals of Data Science

Small project - Ref/Def 2024/25 (30 points, 30%)

1 Introduction

You are given a dataset with a representative sample showing the numbers of passengers carried by a small European airline over a three-year period (see Section 2 below). You will need to write a Python code analysing your dataset, producing two plots and calculating two values (see Section 3 below). You are then asked to describe your findings in a short report (see Section 4), and submit this report, along with your Python code, via Studynet/Canvas (see Section 5).

Please, read this brief carefully and in full.

2 Which dataset to use

Please, use dataset "airlinerefdef.csv"

This dataset shows an airline's daily performance during 2021, 2022, and 2023. It is a comma-separated CSV file with four columns: day (or row) ID number (with 0 corresponding to January 1, 2021), date (in yyyy-mm-dd format), number of passengers carried that day (in thousands), and average ticket price on that date (in Euros).

3 What data analysis needs to be done

Write a Python code, which reads the dataset. Then

(A) **All students**: Derive distribution of average **daily** number of passengers flown during a year, and plots it as a bar chart with each column corresponding to **one month**, from January to December (Figure 1) using **whole** 3-year-long dataset. Each bar should represent an average daily passenger number for a specific month of an average year.The plot must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the plot;

(B) **All students**: On the same figure **(Figure 1)** using the same axes plot the Fourier series approximating daily passenger number variation during 2021 for each day of that year. The series should be limited to the first eight terms.

(C) **If the last digit of your ID number is 0, 1, 2 or 3**: create a bar plot showing the average prices of tickets for flights on different days of the week during an average week. Hence, your plot should have 7 columns, representing days from Monday to Sunday. The plot must be included in your report as Figure 2. The plot must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the plot;

(D) **If the last digit of your ID number is 0, 1, 2, or 3**: calculate the total revenues of the airline in 2021, 2022 and 2023, and print these three values in Figure 2;

(E) **If the last digit of your ID number is 4, 5 or 6**: create a scatter plot showing the average daily ticket prices versus daily passenger numbers, and, using linear regression, approximates this data using a linear function. This plot must be included in your report as Figure 2. On the same plot print the formula for the linear regression function. Keep at least two significant digits in the coefficients of the formula. The plot must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the plot;

(F) **If the last digit of your ID number is 4, 5 or 6**: calculate the year-on-year changes of average ticket prices between 2022 and 2021, and 2023 and 2022. Print the two values as percentages in Figure 2;

(G) **If the last digit of your ID number is 7, 8 or 9**: create a bar plot showing the average revenue generated by passengers flying on different days of the week during an average week. Hence, your plot should have 7 columns, representing days from Monday to Sunday. The plot must be included in your report as Figure 2. The plot must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the plot;

(H) **If the last digit of your ID number is 7, 8 or 9**: calculate the fractions of the airline revenue generated by passengers flown by the airline during winter, spring and summer, and print these three values in Figure 2.

4 What to include into your report

Write a short report, which

- includes your name and 8-digit ID number;
- provides a brief description of your dataset (what variables does it contain?, what is the structure of the dataset?);
- includes the figures produced by your code with short, informative captions;
- includes the mathematical formulas you had to use to analyse the data, namely, to calculate the passenger number distribution and the Fourier series, and, when appropriate, to calculate the parameters of linear regression, average prices, revenue distribution, total revenues, fractions of revenues and year-on-year changes;
- includes a discussion of the figures you created and the values you calculated, and any conclusions you can make based on the obtained results. The discussion and conclusion section must be at least one page long (excluding any figures or equations).

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

5 What to submit

- 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.

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.