Small project - Semester A 2025/26

Fundamentals of Data Science module / 7PAM2020-0901

1 Introduction

You are given two datasets containing public transport usage information in an European city (see Section 2 for a description of the data). The first dataset contains information on the actual number of passengers using different modes of transport, as well as average ticket prices or total revenues generated by the ticket sales in 2019. The second dataset is a statistically representative sample of individual journeys in 2022, with each element containing the date and time of the trip, information on the mode of transport, ticket price, distance travelled, and the duration of the trip.

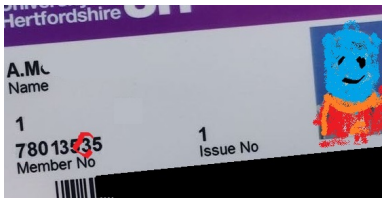
Both datasets can include up to three different modes of transport: buses, trams, and metro (including all types of suburban and commuter trains, underground, overground, metro etc). However, not every city has all three modes of transport.

For simplicity, New Year's, Easter, and Christmas days are excluded from the dataset, because public transport usage on these days can vary significantly from other days. Furthermore, information on some other days may be missing from the first dataset (2019) for technical reasons.

You will need to write Python code to analyse your datasets, plot three graphs, and calculate three values (see Section 3). You will then need to describe your results in a short report (see Section 4). You will need to submit the report, code, and calculated values via Studynet/Canvas (see Section 5).

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

2 Which datasets to use



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:

	Dataset for 2019	All passengers in 2019	Representative sample for 2022	Bus passengers in 2022	Tram passengers in 2022	Metro passengers in 2022
0	2019data0.csv	358 066 262	2022data0.csv	95 341 772	62 002 755	212 037 551
1	2019data1.csv	228 689 197	2022data1.csv	97 440 752	0	119 844 941
2	2019data2.csv	3 023 290 173	2022data2.csv	703 444 216	802 612 335	1 025 314 200
3	2019data3.csv	2 650 924 571	2022data3.csv	638 225 089	852 502 319	1 119 097 445
4	2019data4.csv	198 486 658	2022data4.csv	95341772	0	122 034 174
5	2019data5.csv	365 516 088	2022data4.csv	95 341 772	0	212 037 551
6	2019data6.csv	45 491 315	2022data6.csv	39 537 756	0	7 064 380
7	2019data7.csv	2 576 009 432	2022data7.csv	683 089 225	952 605 310	1 210 897 821
8	2019data8.csv	1 067 652 913	2022data8.csv	266 107 756	344 481 229	458 859 118
9	2019data9.csv	1 829 875 937	2022data9.csv	433 020 544	594 306 621	742 099 762

The download links for the datasets and their brief descriptions can be found in the description of this assignment on Canvas.

3 What data analysis needs to be done

3.1 All students...

All students need to write Python code, which

(A) Reads the data from the two datasets provided to you. You must not change the files (including their filenames) in any way. Your code should be able to read the datasets provided they are in the same folder as your code, irrespective of the folder name;

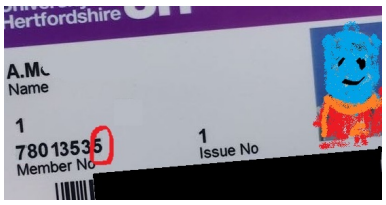
(B) Creates a figure showing four plots: scatter plots of total daily passenger number using public transport in 2019 and 2022, and smoothed line plots total daily passenger number using public transport in 2019 and 2022. The smoothed line plots have to be created using the first 8 terms of the Fourier series representing the original daily passenger data. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 1** in your report;

(C) Creates a figure with two bar plots showing average numbers of passenger using public transport on different days of the week, from Monday to Sunday, in 2019 and in 2022. Therefore, each bar plot should have seven bars. The figure must show the X, Y and Z values you have to calculate (see Section 3.2 for your X, Y and Z values). The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 2** in your report;

(D) Creates a figure which show a scatter plot of the price (in Euros, vertical axis) plotted versus the length of trip (in km, horizontal axis) for all metro journeys made in 2022, and a linear fit (of the scatter plot data) obtained using linear regression. The formula for the linear function representing the fit must be shown in the figure. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 3** in your report.

3.2 Depending on your card number...

In addition to the above analysis, you have to produce a fourth figure and calculate three values (we will call them X, Y and Z). The fourth figure you have to produce and the values you have to calculate depend on the **LAST DIGIT** of your student ID card number:



Last digit of your student ID card number.

If this DIGIT IS "0", please

- Create a figure which shows two bar plots: a bar plot showing total number of journeys made by bus, tram and train in 2019 (three bars), and the same for 2022. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Calculate the total number of weekend (Saturday-Sunday) journeys made by bus (value X),
- the total number of weekend (Saturday-Sunday) journeys made by tram (value Y),
- and the total number of weekend (Saturday-Sunday) journeys made by train (value Z).

If this DIGIT IS "1", please

- Create a figure which shows two bar plots: a bar plot showing the total number of journeys made during peak hours, the total number of journeys made during off-peak hours on working days (Monday-Friday), and the total number of journeys made on weekends in 2019 (three bars), and the same for 2022. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Calculate the fraction of the revenue (in %) generated by journeys made during peak hours in 2019 (value X),
- the fraction of the revenue (in %) generated by the journeys made during off-peak hours on working days (Monday-Friday) in 2019 (value Y),
- and the fraction of the revenue (in %) generated by weekend (Saturday-Sunday) journeys in 2019 (value Z).

If this DIGIT IS "2", please

- Create a bar plot showing annual variation of the monthly-averaged ticket price in 2019. Each bar should represent one month from January to December. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Calculate the average peak ticket price in 2019 (value X),
- Calculate the average price of off-peak tickets for journeys between Monday and Friday in 2019 (value Y),
- Calculate the average price of tickets for weekend trips in 2019 (value Z).

If this DIGIT IS "3", please

- Create a figure which shows two bar plots: a bar plot showing total number of journeys made by bus, tram and train in 2019 (three bars) during weekends (Saturday-Sunday), and the same for 2022. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Calculate the fraction of weekend (Saturday-Sunday) journeys made by bus (value X),
- the fraction of weekend (Saturday-Sunday) journeys made by tram (value Y),
- and the fraction of weekend (Saturday-Sunday) journeys made by train (value Z).

If this DIGIT IS "4", please

- Create a bar plot showing distribution of all journeys made in 2022 in respect of their travel time (duration). The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Calculate the number of all journeys made by bus in 2019 (value X),
- the number of all journeys made by tram in 2019 (value Y),
- and the number of all journeys made by metro in 2019 (value Z).

If this DIGIT IS "5", please

- Create a bar plot showing distribution of all journeys in made 2022 in respect of their average velocity. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Evaluate the average speed of bus journeys in 2022 (value X),
- the average speed of train journeys in 2022 (value Y),
- and the average speed of all journeys in 2022 (value Z).

If this DIGIT IS "6", please

- Create a figure which shows two bar plots: a bar plot showing a fraction (in percent) of journeys made by bus, tram and train in 2019 (three bars), and the same for 2022. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.

- Calculate the fraction (percentage) of all journeys made during spring months (value X),
- the fraction (percentage) of all journeys made during summer months (value Y),
- and the fraction (percentage) of all journeys made during autumn months (value Z).

If this DIGIT IS "7", please

- Create a bar plot showing average hourly distribution of journeys made on Friday in 2022. The plot should have 24 bars representing hours from 0 to 23. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Calculate the fraction (percentage) of the annual revenue made by the journeys during spring months (value X),
- the the fraction (percentage) of the annual revenue made by the journeys during summer months (value Y),
- and the fraction (percentage) of the annual revenue made by the journeys during autumn months (value Z).

If this DIGIT IS "8", please

- Create a figure which shows two bar plots: a bar plot showing monthly public transport revenue in 2019 and monthly public transport revenue in 2022. Hence, each bar plot should have 12 bars, each representing the total revenue during a month, from January to December. The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Calculate the fraction of the revenue generated by buses in 2022 (out of the total 2022 public transport revenue) (value X),
- the fraction of the revenue generated by trams in 2022 (out of the total 2022 public transport revenue) (value Y),
- and the fraction of the revenue generated by trains in 2022 (out of the total 2022 public transport revenue) (value Z).

If this DIGIT IS "9", please

- Create a scatter plot of the length of all journeys made in 2022 (in km) versus their durations (in minutes). The figure must have appropriate axes names, ticks, labels and legend. Your student ID number must be clearly shown on the figure. This figure must have only one panel. This figure has to be included as **Figure 4** in your report.
- Evaluate the total revenue generated by bus journeys in 2019 (value X),
- the total revenue generated by tram journeys in 2019 (value Y),
- and the total revenue generated by train journeys in 2019 (value Z).

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?) [250–1000 symbols];
- includes Figures 1-4 produced by your code with short, informative captions (no more than two pages in total, figures should be decipherable, font sizes should be similar to the font sizes used in the report);
- provides all mathematical formulas you used to produce Figures 1-4 and the X, Y and Z values (including, but not limited to, Fourier series, linear regression etc)
- ends with a discussion of the results any conclusions that can be made based on Figures 1-4 and the X, Y and Z values (for instance, what is the difference between the public transport use in 2019 and 2022 in terms of the annual, monthly, weekly and daily variation of passenger numbers, travel time, prices and revenues? What could be the reason for this difference?) [1000–3000 symbols]

Your report should be no longer than FIVE 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. Text and formulae included as images will be considered a substantial mistake, resulting in a significantly lower grade.**.

5 What to submit

- **Your report** in PDF format as a [$IDnumber$].*pdf* file, where [$IDnumber$] is your 8-digit student ID number;
- **Your Python code** as a [$IDnumber$].*py* 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 instead of the Python codes;
(c) Only files submitted via Canvas are considered;
(d) When rounding numbers, keep at least two significant digits.