Main Examination period 2023 – January – Semester A

MTH765P: Storing, Manipulating and Visualising Data

Duration: 2 hours

The exam is intended to be completed within **2 hours**. However, you will have a period of **4 hours** to complete the exam and submit your solutions.

You should attempt ALL questions. Marks available are shown next to the questions.

All work should be **handwritten** and should **include your student number**. Only one attempt is allowed – **once you have submitted your work, it is final**.

In completing this assessment:

- You may use books and notes.
- You may use calculators and computers, but you must show your working for any calculations you do.
- You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.
- You must not seek or obtain help from anyone else.

When you have finished:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;

Examiners: N. Otter, P. Skraba

All computations should be done by hand where possible, with marks being awarded for showing intermediate steps.

Question 1 [34 marks]. For each of the following, write down a regular expression which best matches the following criteria. For each case provide a short explanation of the regular expression.

(a) For this part you should assume that you are using regular expressions in bash with `grep -P`. Match all numbers which meet the given criteria. You may assume they occur in the middle of a sentence, separated by a space or a comma.

(i) All numbers in the range between -50 (inclusive) and 50 (inclusive). [**7**]

(ii) Let x be the last digit of your student ID. All numbers that do not end in x. [**7**]

(b) For this part you should assume that you are using regular expressions in python with the `RE` package.

You are given a file containing several IBAN numbers, which are of the following form:

```
CC 12 ABCD 0123456789
```

where `CC` is a string of two letters for the country code, `12` is the check number, a string of two numerical digits, `ABCD` is a string of alpha-numeric characters corresponding to the specific bank ID and `0123456789` is the account number, which can be a string of varying length, containing alpha-numeric digits. Note that except for the country code, the remaining letter in the IBAN number may be in lower or uppercase. You may assume that IBAN numbers are stored in a text file with one IBAN number per line. Also, you may assume that the different parts of the IBAN numbers are separated by one blank space, as illustrated in the example above.

(i) Write a regular expression which matches any IBAN number from the countries UK, Germany and France. [**6**]
Note: The country codes are `GB`, `DE` and `FR`, respectively.

(ii) Write a regular expression that matches all IBAN numbers containing a bank account number that has between 7 and 9 characters. [**7**]

(iii) Write a regular expression that matches all IBAN numbers having a bank account number containing only digits. [**7**]

Question 2 [20 marks].
Consider the following data set of daily temperature measurements in Celsius degrees, which were collected once a day at the same time for 10 consecutive days.

day	1	2	3	4	5	6	7	8	9	10
temperature	-7	0	6	7	5	-4	3	16	2	6

(a) In this question you are asked to draw a box-whisker plot. For this, you should compute the following values. Note that you should explain how you perform the computations (e.g., it is not enough to write the value for the first quantile, but you should explain how you compute the value).

- First quantile **[2]**
- Third quantile **[2]**
- Maximum, minimum **[2]**
- Average or median **[2]**

Then draw the plot by hand, specifying how you are using the previously computed values, and whether you are choosing the average or median for your plot. **[4]**

(b) Which samples are the outliers? Choose one of the several definitions given in the lectures, and give your answer specifying which definition you are using. **[4]**

(c) Now assume that the measurements for Day 3 and Day 5 are missing. Interpolate the missing values using (left) zero-order and linear interpolation. **[4]**

Question 3 [24 marks]. You are given the following XML code.

```
<gallery>
    <album name="summer 2022">
             <photo ID="202201">
                    <title>laughing on beach <\title>
             <\photo>
            <photo ID="202202">
                    <title>rock climbing <\title>
            <\photo>
    <\album>
    <album name="selfies">
            <photo ID="2010 123"><\photo>
            <photo ID="2014 01"><\photo>
    <\album>
<\gallery>
```

(a) Draw the tree structure of this code. [**6**]

(b) What are the attributes of each node, if any? [**4**]

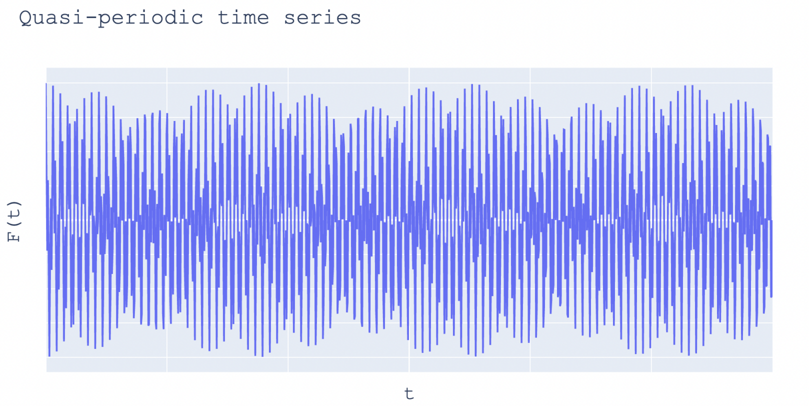
(c) What is the root node of the tree? How many children does it have? [**2**]

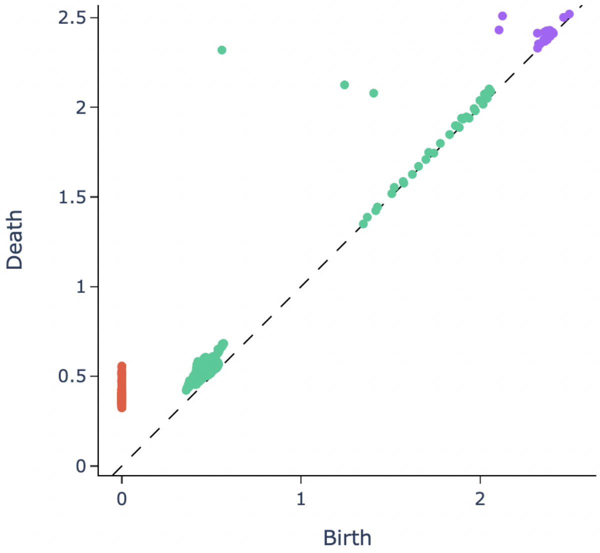
(d) Rewrite the XML code given in the question statement so that all nodes are children of the root. [**6**]

(e) Rewrite the XML code given in the question statement in JSON format. There is no unique way to answer but please ensure you provide a valid JSON encoding the structure and all the information. [**6**]

Question 4 [6 marks].

Consider the two visualisations below. List three mistakes for each of them [**6**]





Question 5 [16 marks].
You are given the following SQL tables.

Genres

id	name
1	Fanstascience
5	Art
7	Music
10	Philosophy
2	Film
11	Mathematics

Books

ISBN	title	author	publ_year	genre
978-1-4000-4437-5	John Cage	K. Silverman	2010	7
978-0-292-77624-1	Sculpting in Time	A. Tarkovsky	1986	2
978-0-8166-6547-1	To Be or not. . . To Bop	D. Gillespie	2009	7
978-0-525-26018-0	Fierce Poise	A. Nemerov	2021	5

(a) Write an SQL query that will return the titles of books whose genre is "Music". Note: you may use that "Music" has id 7. [4]

(b) Write an SQL query that will return title and author names for all books that were published between 2000 and 2011. [4]

(c) What does the following query return?

```
SELECT name FROM Genres INNER JOIN Books ON Books.genre=Genres.id
```

[4]

(d) What does the following query return?

```
SELECT name FROM Genres WHERE id IN (SELECT genre FROM Books)
```

[4]

End of Paper.