TM351 Data management and analysis

TMA 02

Contents

WEB 05030 2
2.1

TMA 02

This tutor-marked assignment (**TM351 TMA 02**) must be submitted by 12 noon (UK local time) on **02 March 2017**.

This module requires all assignments to be submitted electronically. To submit an assignment, please follow the link(s) from your StudentHome page to the online TMA/EMA service.

If you foresee any difficulty with submitting your assignment on time, you should contact your tutor well in advance of the cut-off date.

For further information about policy, procedure and general submission of assignments please refer to the Assessment Handbook, which can also be accessed via your StudentHome page.

TMA 02 assesses your work on Parts 8–17 of TM351.

Preparation - before you start work

For Question 2 of this TMA you will be directed to work in an IPython Notebook. This will be available, together with the PDF and data files for TMA 02, in the TMA 02 section of the `Assessment resources` page of the TM351 website, in a zip archive titled `tm351_tma02_16j.zip`. Now carry out the following steps:

1 Download this archive and unzip it to extract the files it contains into your virtual machine's shared folder. Unzipping it will create a folder `TMA02_2016J/` which will hold the PDF for Question 2(a)(i), the Notebook for Question 2(b), an additional (optional) Notebook for Question 3, and a `data/` folder containing additional data files.

2 Rename this folder in such a way that the new name has your OU student PI (personal identifier) at the beginning, i.e. *yourPI*`_TMA02_2016J/`.

3 Inside the renamed folder, make a copy of the supplied Question 2 Notebook.

4 Rename the copy, again putting your PI at the start of the Notebook filename: `TMA02_Question2b.ipynb` should be renamed `yourPI_TMA02_Question2b.ipynb`.

Your PI in the folder name and Notebook filename will allow your tutor to identify your work.

You should ensure that you always work in this folder when working on your TMA. When working through the Notebook-based question you should work on the renamed copy: the original Notebook will then be available in case you require it as a backup. Do remember to back up your files regularly.

You should now be ready to start the TMA questions.

Submitting your completed TMA

When you have completed the TMA, all your files should be in the folder named *yourPI*`_TMA02_2016J/`. These files should be:

- a *Solution Document* (in .doc or .docx format) containing your answers to Questions 1, 2(a), 2(c) and 3(a), along with a reference to the Notebook you generated in answering Question 2(b) and a reference to each file you created in answering Question 3(b)
- the original PDF for Question 2
- the original Question 2 Notebook file
- your updated Question 2 Notebook file containing your answer to Question 2(b) (with *yourPI* added to the front of the Notebook filename)
- the original additional Notebook for Question 3(a)
- the files you are submitting in answer to Question 3(b)
- the original data files in the `data/` folder
- any additional data files that you have added, created or updated, in the `data/` folder.

Zip this entire folder, then check that all your files are present in the resulting archive. Then submit your zip file to the online TMA/EMA service. After the cut-off date, your tutor can then download, mark and return your work.

If any of the above process is unclear, contact your tutor or post a help request on the TM351 Technical help forum as soon as possible.

Question 1 (40 marks)

You should be able to answer this question when you have completed Parts 8 to 14 of the module.

This question is intended to test your understanding of different database structures.

This question tests the following learning outcomes:

- Understand the similarities and differences between at least two different database models, and how they are used to manage data collections.
- Select an appropriate database model for a data collection.

You should spend no more than four hours on this question.

This question does not require you to work through a Notebook. Write your answers to all parts of Question 1 directly into your *Solution Document*.

Scenario for Question 1

A large service sector is that of commercial babysitting services. A typical example is www.sitters.co.uk. Upon registering, a family can book a babysitter for a given timeslot. The service selects a babysitter, handles the booking details and makes appropriate arrangements with the babysitter. The service also handles payments by the client families, and ensures that all its registered babysitters have been cleared by the Disclosure and Barring Service (DBS)[1].

The data stored by the service needs to accessed by several different groups of people with different requirements and responsibilities, including the office manager, the finance manager, the line managers for the babysitters, and the person tasked with maintaining the data storage services.

In this question, you should imagine that you have been recruited to set up the data management system for a local babysitting service, called Open Sitters. One of your tasks is to create a data storage system to store the relevant data about the clients and the babysitters. Open Sitters has stated that they need to store data about the various families who use the service, the babysitters managed by the service, and details of each booking made by each family.

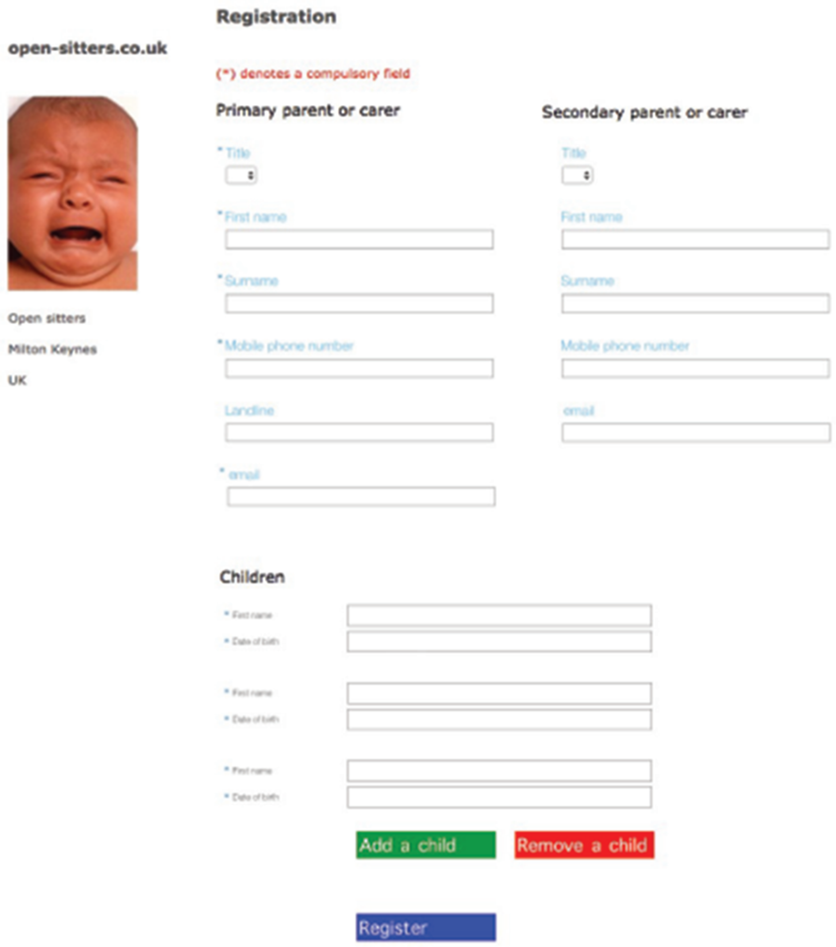
(a) The TM351 module materials and Ponniah (2003) identify several potential difficulties of using file-oriented systems to store data for such scenarios.

   ◦ Give two potential difficulties from a practical point of view in using a file-oriented system for managing the Open Sitters data.
   ◦ Give one potential difficulty from a legal point of view in using a file-oriented system for managing the Open Sitters data.

(b) You now need to consider how you might implement the database to store the service's data.

   A mock-up of the family registration page is shown in Figure 1. The site allows arbitrary numbers of children to be registered with each family; the "Add a child" and "Remove a child" buttons are used to set the total number of children to be registered with the family.

[1] A DBS check is required under UK law for employees whose job involves working with children.



open-sitters.co.uk

Open sitters

Milton Keynes

UK

Registration

(*) denotes a compulsory field

Primary parent or carer

*Title

*First name

*Surname

*Mobile phone number

Landline

*email

Secondary parent or carer

Title

First name

Surname

Mobile phone number

email

Children

*First name

*Date of birth

*First name

*Date of birth

*First name

*Date of birth

Figure 1 Example of Open Sitters family registration site (adapted from a similar page on sitters.co.uk)

Suppose that you decided that you decided to store the data from this web form in a document database, and chose MongoDB. Outline a possible structure for the data if you were to store it in a MongoDB database. You may use a diagram.

(4 marks)

(c) Now, imagine that you decide to store the data from this web form in a relational database instead. You decide to use two relational tables: a “family” table, which stores information about the carers, and a “child” table to store the information about the children.

List the columns you would include in each table. What would you choose for the primary key of each table? How would you declare the relationship between the two tables? Describe all the constraints on the data that are implied by the above scenario. State any assumptions you make. If you make no assumptions, state this.

(12 marks)

(d) Highlight the key differences between your document-based model of the data, and your relational model of the data.

(4 marks)

(e) Suppose that, in the relational model, there is another relation “booking”, in which each row represents a distinct booking made. You decide to add an extra column to the “family” table which stores the total number of bookings made by each family. If a registered family has not yet made a booking, the stored value of their total number of bookings is zero. When a registered family makes a new booking, the booking is recorded in the “booking” table, and the value in the extra column in the “family” table is increased by 1.

What DBMS function ensures that this two-step process maintains data integrity, i.e. that the total number of bookings recorded in the family table is equal to the total number of bookings recorded in the “booking” table? Briefly explain how data integrity can be maintained.

(4 marks)

(f) Relational databases can enforce more regularity on the data they contain than document databases can. Give three examples of relational properties that help constrain the data in a relational database.

In contrast, document databases are generally ‘schema free’. Give one advantage and one disadvantage of using this approach.

Draw on examples from the babysitting service scenario to illustrate your answers.

(10 marks)

Question 2 (35 marks)

You should be able to answer this question when you have completed Parts 8 to 17 of the module.

This question is designed to get you started on a data investigation that will be developed into a larger investigation for the end-of-module assessment (EMA).

This question tests the following learning outcomes:

- Use data to answer a practical question.
- Use appropriate software packages to explore a dataset.
- Write a report detailing a systematic approach to analysing a dataset.

Ensure that you have made a copy of the `TMA02_Question2b.ipynb` Notebook and renamed it so that it has your personal identifier (PI) at the front of the Notebook filename (i.e. `yourPI_TMA02_Question2b.ipynb`).

You should spend no more than three and a half hours on this question.

Write your answers to Questions 2(a) and 2(c) directly into your *Solution Document*. Question 2(b) requires you to work in the `yourPI_TMA02_Question2b.ipynb` Notebook. Write the filename of your Question 2(b) Notebook in your *Solution Document* under the heading ‘Question 2(b)’.

Scenario for Question 2

You are to investigate patterns in votes cast in the EU (Brexit) referendum.

For this question you are provided with an Electoral Commission dataset: http://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information.

The dataset you will be using is the full results data. It is contained in the file `EU-referendum-result-data.csv` in the `TMA02_2016J/data/` folder.

Note that you will be investigating the referendum results in this TMA and the EMA. When you discuss this work with your tutor and other students, please limit your discussions to the data and not observations about the correctness or otherwise of anyone's vote in the referendum or opinion on EU membership.

(a) You should spend no more than half an hour on this part. This part is designed to give you a feel for the data you are investigating.

- Read the introduction to ‘The 2016 EU Referendum Report’ `2016-EU-referendum-report.pdf` (in the `TMA02_2016J/` folder) downloaded from http://www.electoralcommission.org.uk/__data/assets/pdf_file/0008/215279/2016-EU-referendum-report.pdf (accessed 13 October 2016). This gives some background to the referendum and results.
- Visit the BBC’s summary of the referendum results at http://www.bbc.co.uk/news/politics/eu_referendum/results. Browse the website. Use the ‘interactive map’ and list of local results to view selected parts of the data.
- In your *Solution Document*, use no more than 100 words to write bullet points to briefly observe what you have been able to find out from the PDF and website about how the referendum results were reported.

The purpose of writing this short summary is to demonstrate that you have explored the background and got some feel for the data you will be working with. We do not want you to write about the pattern of voting: you will be carrying out your own data analysis in the next part of the question.

(5 marks)

(b) You should spend no more than two hours on this part. In this part you will work in the `yourPI_TMA02_Question2b.ipynb` Notebook to investigate the EU referendum data.

Treat this Notebook as a lab notebook: keep all the work you do and don't tidy it up or delete work that turns out to be a dead end. Use level 1 headings in Markdown cells in the Notebook to help your tutor identify regions in the Notebook that demonstrate you have performed the required steps.

- Open your copy of the Notebook. Import the dataset `EU-referendum-result-data.csv` from the `TMA02_2016J/data/` folder into a DataFrame. Use some simple *pandas* commands such as `head()`, `describe()` and `unique()` to explore the content of the dataset and decide which columns are of interest to you. Add explanatory comments to your code.
- The data is organised by region and district. Investigate patterns in the votes between regions.
- Use the matplotlib facility provided in the Notebook to create and label at least three plots to visualise some aspects of the data. You may choose to create plots for different regions and/or for different turnout rates.

  Make sure you label any plots you decide to use in your report for part (c).
- Include notes in your lab Notebook critically evaluating what you think your visualisations tell you.

(20 marks)

(c) You should spend no more than one hour on this part. In this part you will use your findings from parts (a) and (b) to write a report using the following outline structure:

Aims and objectives
Background
Sources of data
Analysis pipeline
Findings
Conclusions
References

Present your report in your *Solution document*. Your report should be no more than 600 words. Some sections may be very short. Include evidence in the form of screenshots and plots, as appropriate. You should include at least two visualisations. For every visualisation you include, critically evaluate what it tells you about the data.

Document any uncertainties that you have about the data. If you have no uncertainties, then state this. Comment on whether there are risks to the report linked to the uncertainties you have documented. If there are risks, state what those risks are.

You should use references in your report, as appropriate, to support your conclusions and give a context for your investigation. Include a reference to the Notebook you used in your investigation so that your results may be independently verified.

(10 marks)

Question 3 (25 marks)

This question is designed to help you in the planning stage of the EMA. This is your opportunity to develop a work plan so your tutor can give you helpful, timely feedback.

A good answer to this question will mean you have mapped out your EMA work and got a good start at understanding what the EMA requires.

This question tests the following learning outcomes:

- Use data to answer a practical question.
- Present an analysis of a dataset to a variety of audiences.

You should spend no more than two and a half hours on this question

Write your answer to Question 3(a) directly into your *Solution Document*. Question 3(b) requires you to create two files. Write the filenames of these files in your *Solution Document* under the heading 'Question 3(b)'.

(a) You should spend no more than one and a half hours on this part. This part is designed to prompt you to think about the question you will explore for your EMA. It provides an opportunity for you to get feedback from your tutor to inform your EMA work.The EMA requires you to investigate a new dataset and answer two questions: one on the new dataset and one that requires a combination of the two datasets. You will need to use a data mining technique, either classification or clustering, at some point in the EMA data investigation.

We have provided three further datasets in the `tm351_tma02_16j.zip` file.

The three further datasets are:

- Population Estimates for UK, England and Wales, Scotland and Northern Ireland, obtained from the Office of National Statistics and cached as `ukmye.zip`. The MYBE1 table within that file gives breakdowns by age; the MYBE3 table gives population change and immigration numbers. (Terms and conditions.)
- Voting records from the 2015 UK general election, provided by the Electoral Commission, cached as . `RESULTS FOR ANALYSIS.csv` is likely to be the most useful, but `RESULTS.csv` could also be helpful to understand the

data. You may want to use the Ward to Westminster Parliamentary Constituency to Local Authority District Lookup datasheet to relate parliamentary constituencies to local authority districts (cached as `wards.csv`; Terms and conditions).

- UK census report DC6206EW, of population by socio-economic group, ethnic group, sex, age, and area, cached as `DC6206EW.csv`.(Terms and conditions.)

Open the datasets and explore what is in them. Consider what questions you could ask of the datasets and how you would go about answering them. You should pay attention to the metadata about these datasets; some of it is given in additional files, some is included in the dataset files themselves.

Write a sentence or two in your *Solution document* for each item below. Justify why you have made your choices.

1 Which additional dataset you have chosen.

2 A question you can investigate using this additional dataset any this dataset can answer it.

3 Another question, which you can answer by combining the Brexit vote dataset and the additional dataset you have chosen, and why this combination of datasets can answer it.

4 A proposal for how data mining will be useful in your investigation.

At this stage, you are not committing to the exact questions or techniques that you will explore in the EMA. The feedback from your tutor will help you to refine your question. Beginning to engage with the EMA question at this early stage will give you the best chance of successfully completing the EMA and the module.

(15 marks)

(b) You should spend no more than one hour on this part.

If you have not yet done so, read the EMA. Think about what you are being asked to do and how this builds on the analysis you did for Question 2. How are you going to address the investigation question you proposed in part (a)?

- Set out a work plan of activities and milestones for completing the EMA. Use the sample data investigation and report as a guide to what you should aim to achieve. The study planner allocates two clear weeks at the end of the module for working on the EMA in addition to a further week during which the EMA cut-off date occurs. However, it is best to start work on your EMA as soon as possible to allow for discussion and time to develop your thoughts.

  Your work plan can take any form that you find helpful, for instance a list or a diagram. It should be realistic and it should be possible to modify the plan and add detail as you progress.

  Submit your work plan in a document named `yourPI_workplan`.

- Start an IPython Notebook that you can use as a project 'diary' in which to record your explorations, results, notes and what you need to do. It should contain at least one initial brief entry, perhaps importing and briefly examining the dataset you chose in part (a). Name the Notebook `yourPI_project_diary.ipynb`.

This is your personal lab Notebook and it will not be assessed beyond the initial Notebook you submit for TMA 02. However, you may want to include it as supporting evidence for your EMA.

Make sure that your work plan and project diary are in the *yourPI*`_TMA02_2016J/` folder.

There does not need to be much detail at this stage, but you need to have a clear vision as to how you are going to proceed.

Your tutor will provide some feedback based on what you submit. Complete this part to the best of your ability to give a good basis for informing discussion with your tutor as you work on the remainder of the module.

(10 marks)