TMA 02

This tutor-marked assignment (**TM351 TMA 02**) must be submitted by 12 noon (UK local time) on **11 March 2021**.

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. The TMA will take about 15 hours to complete all activities.

Preparation - before you start work

For Question 2 of this TMA you will be directed to work in an IPython Notebook. This is available, together with the data files for TMA 02, in the TMA 02 section of the Assessment page of the TM351 website, in a zip archive titled `2020J_TMA02.zip` . Now carry out the following steps:

1. Download this archive, and either place it in your VCE's shared directory (if you are using a local installation), or upload it to the VCE (if you are using the remote installation). Unzipping it will create four directories:

    - a directory `2020j_green_collection`, which contains data for question 1,
    - a directory `2020J_TMA02/`, which contains a readme and sample notebooks for question 2,
    - a directory `2020J_TMA02_data`, which contains the data files for questions 2 and 3, and
    - a directory `Reports`, which contains some background reports on the provided datasets.

2. Rename the directory `2020J_TMA02/` so that the new name has your OU student PI (personal identifier) at the beginning, i.e. *yourPI*`_2020J_TMA02/`.

3. Inside the renamed directory, you will find a number of notebooks. In Question 2(b) you may use these as the basis of your answer (one notebook outlines some initial explorations of the data, the second illustrates how to generate graphs from a subset of the data). If you do not modify these notebooks, there is no need to return them with your TMA.

By including your PI in the directory name and all filenames you will allow your tutor to identify your work.

You should ensure that you always work in the *yourPI*`_2020J_TMA02/` directory when working on your TMA. 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 directory named *yourPI*_2020J_TMA02/. These files should be:

- a *Solution Document* called *yourPI*_TMA02_solution.docx (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 created in answering Question 2(b) and a reference to each file you created in answering Question 3(c)
- a lab Notebook file containing your answer to Question 2(b), called *yourPI*_TMA02_Question2b.ipynb
- the files you are submitting in answer to Question 3(c)
- any additional data files that you have added, created or updated, in the data/ subdirectory of the *yourPI*_2020J_TMA02/ directory.

Note that you should not return any files (e.g. sample Notebooks, data files) that you have not changed.

Zip this entire directory, then check that all your files are present in the resulting archive. Next, 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 the TMA/EMA service refuses to accept your file, check that it does not contain any unchanged data files and that there are no files with very long names (especially in the .ipynb_checkpoints directory, which may be hidden by your operating system and, if necessary, can be deleted).

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

Question 1 (40 marks)

> You should be able to answer parts a) to e) of this question when you have completed Parts 8–14 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.

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

There are still a number of organisations who have a web presence but that use forms that people can complete as part of a document and send via email as an attachment.

A small retail company called Baby Green Collection has been designing and creating baby clothes from material that can be recycled. Their material sourcing and production is international and their main customer base is in London.

They currently enable orders to take place via phone, email and web-based forms. All the data is collected and stored on spreadsheets.

You have been contacted by the company to help design and develop a new database solution. They have sent you a sample of their 2019 data to illustrate the kind of data collection they currently use to keep track of the company's production, sales and clients (see the spreadsheet files in the `2020j_green_collection` directory. Note that some of the spreadsheet files contain more than one worksheet).

The data stored by the service needs to be accessed by several different groups of people with different requirements and responsibilities, including the office manager, the finance manager, the production manager, sales and client support, and the person tasked with maintaining the data storage services.

One of your tasks is to create a data storage system to store the relevant data about the clients and the products. Baby Green has stated that they need to store data about the staff of the organisation using the Baby Green system. In addition they need to record all contacts between staff in sales & customer support and customers.

a. Illustrate to Baby Green the potential benefits and disadvantages of using spreadsheet and database storage solutions. Provide examples to illustrate the key benefits of the database solution. What might change in their data management process when moving from a spreadsheet to a database system?

   Highlight any legal and security considerations that the organisation should consider.

   (10 marks)

b. You will now need to consider how you might implement a database to store Baby Green data.
   Review the data files that Baby Green have shared with you.

   Outline a possible structure for the data if you were to store it in a MongoDB database.

   (6 marks)

c. Now, imagine that you decide to store the data from Baby Green in a relational database instead. Using example data they have shared, examine the data and provide a set of relational tables.

   List the columns you would include in each table. What would you choose for the primary key of each table? Define the relationships between the 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.

   (14 marks)

d. Data quality is important when tracking product production and sales. Some of the communications between the suppliers and Baby Green, and Baby Green and their clients, will be via documents and images, e.g. images illustrating design flaws, quality of the materials used and letters/details about returns. Hence, there will be important data that may not fit the structured data format.

   Compare and contrast the suitability of relational and document databases to support storage and analysis of this dataset and the additional materials. Your answer should address four specific areas, which may include:

   i. the ability to store data which does not conform to a structured format

   ii. the ability of database-using programs to access and process the data

   iii. the ease of changing database definitions to encompass changes in the data structures being held

   iv. the role of constraints to enforce data integrity.

   (8 marks)

e. Analyse the data that Baby Green has shared with you. Identify two areas where you feel the data is incomplete. Explain the advantages and challenges for Baby Green to collect and store this data.

   (2 marks)

Question 2 (35 marks)

> You should be able to answer this question when you have completed Parts 8–17 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.

Write your answers to Questions 2(a) and 2(c) directly into your *Solution Document*. Question 2(b) requires you to create and 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 the data found in the `2020J_TMA02_data/Deprivation_Index` folder. The files include data and metadata about indices of deprivation.

To understand and explore this data, it is recommended that you review the English indices of deprivation 2015 research report (Department for communities and local government, 2015), which we have included in the `Reports` folder.

When you discuss this investigation with your tutor and other students, please limit your discussions to the data.

a. This part is designed to give you a feel for the data you are investigating.

   - Read the English indices of deprivation 2015 research report (Department for communities and local government, 2015), Chapters 1 and 2 and sections 1-3 of Chapter 3 (pp.7-27). This gives some background about English indices of deprivation.

   - 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 description about English indices of deprivation.

   - Ensure you comment on both data collection and integrity.

   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, and how the background may limit your conclusions. We do not want you to write about the results presented in the report; you will be carrying out your own data analysis in the next part of the question.

   (5 marks)

b. In this part you create and work in a notebook called *yourPI*_TMA02_Question2b.ipynb to investigate English indices of deprivation.

   As mentioned in the 'Preparation' section, we have prepared some notebooks which may help you get started and can be used to explore the datasets. These notebooks are explained in `tm351_tma02_readme.ipynb`. You may wish to use these notebooks to get started, but you are under no obligation to do so. Feel free to use *other methods* to store and manipulate the data. The following step by step approach can be used as a starting point but other approaches can be considered. However, explain your choices in the notebook.

   Create a new notebook called *yourPI*_TMA02_Question2b.ipynb. Treat it 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.

   In addition, each discrete manipulation of the data should be presented in its own code cell (or cells, if it is clearer to break the code up a bit) and be preceded by at least one markdown cell explaining what the code is intended to do.

   - Open *yourPI*_TMA02_Question2b.ipynb and start to explore the dataset. How you explore the data depends on how it is stored. For example, import one or two csv files from the `2020J_TMA02_data/Deprivation_Index/` directory into a DataFrame. Use some simple ***pandas*** commands such as `head()` and `describe()` to explore the content of the dataset. Remember to add explanatory comments to your code.

- The data records the deprivation indices for different geographical areas in England. You should investigate patterns in the data. You may decide to restrict your analysis to specific geographical areas or deprivation indices. However, before you can do this you need to briefly examine all the data to make sure that you select a suitable subset for analysis. If appropriate, use a suitable statistical test to quantify the differences.

- Use the Notebook to create and label at least three plots to visualise some aspects of the data. For example, you may create plots for different geographical areas or different deprivation indices.

  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 investigations and visualisations tell you.

- If you modify any of the supplied data files, or create new data files of your own, place them in a new `data` subdirectory of your *`yourPI`*`_2020J_TMA02` directory.

(20 marks)

c. 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. You should present your results in a form that highlights the relevant results; you must include at least two visualisations. You should critically evaluate your results and their presentation, including mentioning any confounding factors that may weaken your conclusions. These could include concerns about the reliability or coverage of the data, or other influences which are not included.

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.

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. 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 an additional dataset and answer two questions: one on an additional dataset and one that requires a combination of the additional dataset and the deprivation index data. You will need to use a data mining technique, either classification or clustering, at some point in the EMA data investigation. You will learn about these techniques in Parts 20 and 21.

   We have identified two further additional datasets you could use. They are:

   - UK census report DC6206EW (nomis, n.d.a), of population by socio-economic group, ethnic group, sex, age and area, found in the directory `census` (Terms and conditions, nomis, n.d.b.). This directory contains four csv files covering different age groups. The directory also contains the Output Area Lookup table (Office for National Statistics, n.d.a.) as output-area classification.csv which will allow you to connect the LSOA codes in the deprivation indices to the MSOA codes in the census report. (Terms and conditions (Office for National Statistics, n.d.) for the lookup table).

   - GCSE and equivalent results in England: 2014 to 2015: Information about the achievements of young people at the end of key stage 4 in England. See `key stage 4` directory for the data. There are two supporting reports about the data collection, *Department for Education, Main Text, SFR 01/2016* and *Quality and methodology information: SFR 01/2016*.

   Open the datasets and explore what is in them, and see if there is any metadata. Consider what questions you could ask of the datasets and how you would go about answering them.

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

   i. Which additional dataset you have chosen.

   ii. A question you can investigate using this additional dataset and why this dataset can answer it.

iii. Another question, which you can answer by *combining* the deprivation index dataset and the additional dataset you have chosen, and why this combination of datasets can answer it.

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 questions. Engaging with the EMA at this early stage will give you the best chance of successfully completing the EMA and the module.

(10 marks)

b. 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 could you address the investigation question you proposed in Part (a)?

- Describe how you intend to analyse the data and visualise your results. Explain the tools and techniques you could use. This is an opportunity for your tutor to give feedback on your approach before you do too much work on it.

Note: You do not need to discuss data mining in this assignment, but you must consider how data mining can be used in your EMA when you study Parts 20 and 21.

(5 marks)

c. In this section you start of plan your EMA.

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 four weeks at the end of the module for working on the EMA. 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.doc` or `docx`.

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_2020J_TMA02/` directory.

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)

References

Department for Communities and Local Government (2015) *The English indices of deprivation 2015*. Available at: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 (Accessed: 21 August 2020).

Department for Communities and Local Government (2016) *The English indices of deprivation 2015 – Frequently Asked Questions (FAQs).* Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/579151/English_Indices_of_Deprivation_2015_-_Frequently_Asked_Questions_Dec_2016.pdf (Accessed: 21 August 2020).

Department for Communities and Local Government (2015) *The English Index of Multiple Deprivation (IMD) 2015 – Guidance.* Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/464430/English_Index_of_Multiple_Deprivation_2015_-_Guidance.pdf (Accessed: 21 August 2020).

Department for Education (2016) *Revised GCSE and equivalent results in England: 2014 to 2015: Main Text, SFR 01/2016*. Available at: https://www.gov.uk/government/statistics/revised-gcse-and-equivalent-results-in-england-2014-to-2015 (Accessed: 8 September 2020).

Department for Education (2016) *Revised GCSE and equivalent results in England, 2014 to 2015: Quality and methodology information: SFR 01/2016.* Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/493295/SFR01_2016_QualityandMethodology.pdf (Accessed: 8 September 2020).

nomis (n.d.a.) *UK census report DC6206EW*. Available at https://www.nomisweb.co.uk/query/construct/summary.asp?mode=construct&version=0&dataset=682 (Accessed: 31 August 2020).

nomis (n.d.b.) *Terms and Conditions*. Available at: https://www.nomisweb.co.uk/home/terms.asp (Accessed: 31 August 2020).

Office for National Statistics (n.d.a.) *Output Area to LSOA to MSOA to Local Authority District (December 2017) Lookup with Area Classifications in Great Britain*. Available at: https://geoportal.statistics.gov.uk/datasets/output-area-to-lsoa-to-msoa-to-local-authority-district-december-2017-lookup-with-area-classifications-in-great-britain (Accessed: 31 August 2020).

Office for National Statistics (n.d.b) *Terms and Conditions*. Available at: www.ons.gov.uk/help/termsandconditions (Accessed: 31 August 2020).

UK Government (n.d.) *Open Government Licence (OGL), Version 3.0.* Available at: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ (Accessed: 21 August 2020).

Acknowledgements

Grateful acknowledgement is made to the following sources:

The indices of deprivation 2015 and GCSE results 2014–2015 have been published using the Open Government Licence (OGL) version 3.0, see: www.nationalarchives.gov.uk/doc/open-government-licence/version/3/