TM351 Data management and analysis

Fall 2019 TMA01 & Project MG

(Cut-off date will be announced)

1. Preamble:

This section contains general rules and guidelines for completing and submitting your TMA.

1.1 General guidelines

This TMA is provided as a pdf file, a lab notebook template and a word template. You should do all the coding work for this TMA (except for the course notebooks of question 2) in the enclosed notebook template and you should write all your answers to the questions, requested summaries and project report (question 5) in the enclosed TMA word template. You will also work through and submit the requested course notebooks of question 2 as a separate zip file.

The TMA requires that you demonstrate an understanding of course concepts and techniques, including your ability to assess the contents and quality of data and the ability to apply those concepts to sample problems. It also tests your ability to formulate your own research question, investigate it, and report your critical findings. Your tutor will be following a detailed marking scheme, but he or she will particularly look for the following:

1. That all work is your own
2. That you have provided references in the proper format whenever required
3. That you have used the course concepts, terminology and prescribed software

1.2 Using the e-library and other external sources.

When asked to do so, you need to search the e-library and the internet to identify relevant material. In particular, you are urged to use the following sources, all of which are freely available to AOU students:

1. AOU's subscribed e-library, accessible through the LMS which includes a number of different resources
2. References provided in your course materials
3. Online manual pages for languages, libraries and tools used
4. Help forums and blogs
5. Other resources

1.3 Submitting your TMA

For this TMA, you will be required to submit a compressed directory containing four different items:

1. The coding work required to support your answers to questions 1, 3 & 4 and 5 all inside the lab notebook template as a sequence of markdown and well commented and **solved** code cells following each question.

2. Your **solved** course notebooks as described in the TMA question 2 in a compressed file.
3. The answers, summaries and conclusions as well as your research report requested in question 5 in the enclosed TMA word template.
4. Any required data sets in a separate directory named: data

Please note that all notebooks you submit must be in a **solved** state and must show all outputs. Your Grader is not obliged to re-run your notebook cells.

Submit your TMA to the LMS system on (or preferably before) the cut-off date. Your tutor will mark your script and post the grades electronically in the approved electronic channel.

1.4 Plagiarism

All work you submit must be yours and in your own words. Your tutor has tools available to him/her to allow the detection of plagiarism from the Internet as well as from other colleagues. Tutors will also manually check your notebooks and reports for similarity. Furthermore, you may be quizzed on the work you submitted and/or asked to demonstrate that it is indeed your own work. The final exam may contain questions that directly relate to the skills you demonstrated in this TMA.

If you copy material that is not your own and submit it as your own you are committing plagiarism. Plagiarism is a serious offence and if a case of plagiarism is detected, the Arab Open University will apply severe penalties and disciplinary procedures.

1.4.1 Quoting and Referencing.

If you wish to quote other materials, including the TM351 learning materials, then you must clearly acknowledge the source according to accepted rules of citation and referencing.

Note that it is not enough to simply post a reference at the end of the document without explicitly stating which parts of your reference are being quoted. Proper citation of external sources must be included. Also, quoting is only used in limited fashion; to stress a certain point using the words of a well-recognized guru, for example. Large amounts of materials copied into your TMA will not be accepted, even if properly quoted. If you need to refer to large amount of external material, you can simply refer to the source.

All references and accompanying citations must be in the Harvard style of referencing.

1.4.2 Getting help and collaborating with colleagues.

You can discuss the TMA with your tutor. Your tutor will help explain unclear points in the TMA and will direct you to useful and appropriate material in the course. However, you should not expect your tutor to supply you with answers to TMA questions. Remember that answering the TMA is ultimately your responsibility, not your tutor's. In addition, working the TMA and overcoming its difficulties by yourself will help you do well in the final examination.

Sharing knowledge and information and holding discussions with your colleagues about the course material is called group learning and is encouraged by the Arab Open University. However, at the end, you should complete the TMA by yourself and answer the TMA, in your own words. Collaborating in answering TMA questions is not allowed, and is not the same as group learning. You are also not allowed to use the course forum to post answers to TMA questions or to collaborate on answering TMA questions.

2. **The questions**

Question 1 (10 marks)

Place all your coding work for this question in the lab notebook template and your final answers in the TMA word template.

The National Oceanic and Atmospheric Administration (NOAA) in the United States publishes U.S. Hourly Precipitation Data, along with related metadata description files in: **Hourly Percipitation Data**. To get the data use the **NCDC ftp server** to access data set 3240 (DSI-3240) for hourly precipitation data and related data documentation. You will need to explore both the metadata and the data itself to be able to answer the following questions:

Visit the website, explore the data and read the documentation associated with the data, then answer the following questions:

1. What is the provenance of the data?
2. When was the data last updated?
3. How complete is the data if the purpose is to investigate the change in precipitation over the past 100 years in the western hemisphere?
4. Why is the reported precipitation amount not always reported on an hourly basis?
5. Is the data mainly quantitative or qualitative?
6. Is the data structured, Semi-structured or unstructured?
7. How do you describe the producer of this data? Primary, Secondary or Tertiary? Explain why?
8. How do you characterize the source of this data: captured? Exhaust, original, derived, or exhaust? Indicate all that apply
9. Which data is contained in the site? Indexical data, attribute data, or metadata? indicate all that apply
10. In case metadata is present indicate which of the following characterizes it: descriptive, structural or administrative? indicate all that apply.

Question 2 (30 marks)

Complete and submit All the following Jupyter notebooks in the form of a "solved" .rar or .zip file: [30 marks]

0.1 Scribble pad
2.2.0 Data file formats, file encodings
2.1 Pandas dataFrames
2.2.1 Data file formats -CSV
2.2.2 Data file formats - JSON
2.2.3 Data file formats - other
3.1 Cleaning data
3.2 Selecting and projecting, sorting and limiting
3.3 Combining data from multiple data sets
3.4 Handling missing data
4.1 Crosstabs and pivot tables
4.2 Descriptive statistics in pandas
4.3 Simple visualisations in pandas
4.4 Activity 4.4 Walkthrough
4.5 Split-apply-combine with SQL and pandas
4.6 Introducing regular expressions
4.7 Reshaping data with pandas
--- show at least three screenshots from OpenRefine
5.1 Anscombe's Quartet - visualising data
5.2 Getting started with maps – folium
8.1 Movies dataset
9.1 SQL DDL
9.2 SQL DML
9.3 SQL views
10.7 Outer join operations
11.1 SQL set operations
11.2 SQL subqueries
14.1 Basic CRUD
14.2 Introduction to accidents
14.3 Using statistical tests

Please note that:
You will receive 1 mark for each completed notebook, including your own scribble pad notebook and screenshots of the OpenRefine tool, for a total of 30 marks.

Please note that:

- Partially completed notebooks will not be counted. All outputs must be shown.
- Please demonstrate your active interaction with each notebook by including your own additions and/or extensions to the code and/or your own additional comments. Use a double hash sign '##' to distinguish your comments from those already provided in the notebook.
- Your tutor may quiz you on the contents of the notebooks you provide.

Question 3 (10 marks)

Place all your coding work for this question in the lab notebook template and your 300-word summary in the TMA word processing template.

In this question, you will download the UK road safety data sets and its supporting documentation, explore it, and write a summary of your understanding of the purpose and contents of these data sets and your assessment of the quality of the data. To do this, you must develop code to explore the data programmatically in a notebook and provide it as part of your answer.

The Department of Transport(DfT) in the UK publishes detailed *road safety data* consisting of three interlinked data sets:

- Accidents data set
- Vehicle data set
- Casualty data set

DfT also publishes metadata for those files, consisting of *Lookup up tables for variables*, which provides the values and labels for each variable. In addition, the Instructions for the *Completion of Road Accident Reports from non-Crash Sources* contains descriptions of each item contained in the three road safety data sets. This part is designed to give you a feel for the data you are investigating.

Download the **Road Safety Data Sets** for 2015 and 2018 in addition to **The Lookup up Tables for Variables** and the **Completion of Road Accident Reports from non-Crash Sources** documents which describes the contents of those data sets.

- Write a 300-word summary in the word processing document including:
    - What you are able to understand from the road safety data sets about:
        - its contents and **(2.5 marks)**
        - the meanings of those contents **(2.5 marks)**
    - Your assessment of the quality of the data provided if the purpose of the data analysis is to determine contributory factors for accidents. For this question, you will need to:
        - Detect and describe the different types of dirty data in the data sets. Use what you studied them in part 3 of TM351: validity, accuracy, completeness, consistency and uniformity as a guideline in your description **(2.5 marks)**.
        - Estimate the amount of dirtiness of the data of each type and discuss its potential impact of the goal of the analysis **(2.5 marks)**.

In order to achieve this, you will need to use tools and/or python code to detect and quantify the dirtiness of the data. It is not enough to report your findings based on visual inspection of the data alone.

The purpose of this summary is to ensure that you had a good look at the data and its metadata descriptions and that you are able to use proper tools and techniques to assess the quality of the data, and that you have a clear idea about its quality.

Question 4 - Project (50 marks)

Place all your coding work for this question in the lab notebook template and your project report in the TMA word processing template.

In answering this question, you will benefit from the experience you gained in the previous question.

In this question, you will formulate your own research question and investigate it and write a report of your findings in your Solution document. The research question should be related to investigating the relationships between a selected independent variable and a selected dependent variable.

The research question must depend on **more than one data set** and therefore must involve combining more than one data set from among the three provided data sets (Accidents, Vehicles and Casualties). Your research question should also combine **more than one year**. Your investigation must include the use of a **proper measure of correlation** to show the relationships between your selected variables and **appropriate visualisations**. Refer to notebooks for part 14 which demonstrate how this can be done. You may use any visualization, for example utilizing folium or matplotlib and tabular reshaping (crosstabs or pivot tables) to investigate the research question and demonstrate your conclusions.

Please note that you may need to reduce the size of the selected sets before combining them by eliminating unwanted columns in order to reduce the space requirements.

For example, you may wish to investigate the relationship between:

- Age of driver and frequency of accidents in general.
- Age band of driver and severity of accidents
- Weather conditions and severity of accident
- ……..

These are of course just examples, there are many other research questions that you can investigate.

You may also limit the scope of your data that is subject to investigation, for example by limiting the time scope or the geographical scope of the data to be investigated.

After you have opened the data sets and explored their contents with the help of the supporting documents, complete your report according to the following report structure:

1. **Executive summary**
   - A brief summary of your project **(2 marks)**
2. **Aims and objectives**
   - A brief description about the general aims of your project **(1 mark)**
   - and more detailed objectives to achieve those aims **(2 marks)**
3. **The Source Data**
   - Describe the data:
     - its sources and its contents **(1 mark)**, and
     - comment on its quality **(1 mark)**
   - Variable classification I:
     - Classify all variables in the "Export Variables" worksheet of **The Lookup up Tables for Variables** workbook into *dependent or independent* variable. **(2 marks)**
     - Organize your answer into a table. **(2 marks)**

- Variable classification II:
    - Further classify all those variables using the *NOIR system of classification* of all attributes (Nominal, Ordinal, Interval and Ratio). **(2 marks)**
    - Organise your answer into a table. **(2 marks)**

4. **The Research Question**
    - State your research question by:
        - identifying the independent variables (**1 mark),** and
        - the dependent variables you wish to investigate **(1 mark)**.
    - Justify why you chose this question by discussing:
        - why you think it is important **(1 mark)**, and
        - why you suspect that there may be a relationship bet. the variables you chose. **(1 mark)**
5. **Analysis and Findings**
    - Produce convincing correlations demonstrating a statistically significant correlation among your chosen independent and dependent variables. You must choose an appropriate statistical method for the types of measure in the variables in your study. **(2 marks)**.
    - Give your critical interpretation and conclusions about those observed correlations. **(2 marks)**
    - Produce tabular summaries of the data in the form of crosstabs or pivot tables **(2 marks)**, along with your critical interpretation of those tables **(2 marks)**.
    - At least two relevant visualisations **(2 marks)** along with your critical interpretations of each visualization **(2 marks)**.
    - Your final answer to the research question you posed **(2 mark)**
    - and critical comment on your conclusions. **(2 mark)**
6. **Project Description**
    - Describe in simple English how you:
        - planned your project **(1 mark)**
        - went about out acquiring your data **(1 mark)**,
        - preparing it **(1 mark)**,
        - analyzing it **(1 mark)**
        - and reporting your findings **(1 mark)**.
7. **Reflection**
    - Reflect on:
        - your experience with the project **(1 mark)**,
        - what you learned **(1 mark)**,
        - what you went well **(1 mark)**,
        - what went wrong **(1 mark)**
        - and how can you benefit from this experience in future projects **(1 mark)**.
8. **References**
    - At least 5 references. All references must be in the Harvard style of referencing and must be accompanied by proper citations in the text. **(5 marks)**

Project Guidelines:

In writing your report, you should follow the guidelines below:

- Your report should be **no more than 3000 words**. This is a strict limit and you will lose marks if you exceed this limit.
- Your report must be accompanied by the related notebook work in the assigned section of the lab notebook template. Refer to Part 5 of your learning materials for additional information about the lab notebook and the project report. Arrange the lab notebook part of this question, into the following sections:
    - Data Acquisition
    - Data Preparation
    - Visualisations and Analysis
- You should include tabular summaries of the data (cross tab or pivot tables) and **at least two visualisations** and appropriate correlations in your report.
- For every visualization (visual of tabular) you include, you must provide a **critical interpretation** of what the visualisation leads you to believe about the data.
- **Document any uncertainties** that you have about the data (if any), for example due to missing data or data that you have reasons to believe is inaccurate or invalid. Explain how those uncertainties will affect your conclusions.
- **Use references** in your report in the Harvard style of referencing. All references must be properly cited in the text in the same Harvard style of referencing. Include a reference to your lab notebook (lab diary) so that your results may be independently verified.