Tutor Marked Assignment (TMA) - 2020-2021/Spring

TM351: Data Management and Analysis

Cut-Off Date : TBA **Total Marks : 100**

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 in the enclosed notebook template and you should write all your answers to the questions, requested summaries and project report in the enclosed TMA word template. You will also work through and submit the requested course notebooks 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 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 in a compressed file.
3. The answers, summaries and conclusions as well as your research report requested 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.

The questions

Question 1 (30 marks- 1.5 Marks each)

Complete and submit all the following Jupyter notebooks in the form of a "solved" .rar or .zip file:

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 visualizations 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

Please note that:

You will receive 1.5 mark for each completed notebook., including your own scribble pad notebook 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 2 (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 use Pima Indians Diabetes Dataset. Download the data set from https://www.kaggle.com/uciml/pima-indians-diabetes-database.

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Write a summary of your understanding of the purpose and contents of this dataset 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.

Write a summary (~ 300 words) in word processing document which includes the following:

- The contents of the above dataset with detailed description **(4 marks)**
- The quality of the data with respect to validity, accuracy, completeness, consistency, and uniformity **(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 **(3.5 marks)**.

Question 3 – Project (45 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. Your investigation must include the use of a proper measure of correlation to show the relationships between your selected variables and appropriate visualizations. Visualizations can be provided either by utilizing folium or matplotlib.

For example, you may wish to investigate the relationship between

- Blood pressure and diabetes
- Number of pregnancies and diabetes
- Age and diabetes
- ……..

The above given relationships are just samples, there are many other research questions that you can investigate.

Study the data sets, analyze it and then prepare a research document 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 Research Question
    - State your research question by:
        i. identifying the independent variables **(2 marks)**, and
        ii. the dependent variables you wish to investigate **(2 marks)**.

4. 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. **(4 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 **(4 marks)**, along with your critical interpretation of those tables **(2 marks).**
- At least two relevant visualizations **(6 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)**

5. Reflection
   - Reflect on:
     i. your experience with the project **(2 mark),**
     ii. what you learned **(1 mark)**,
     iii. what you went well **(1 mark)**,
     iv. what went wrong **(1 mark)**
     v. and how can you benefit from this experience in future projects **(1 mark)**

6. References
   - At least 6 references. All references must be in the Harvard style of referencing and must be accompanied by proper citations in the text. **(6 marks)**

Question 4 – (15 marks)

(a) Compare *any three* recent trending data visualization tools.
(b) What is data pre-processing? Explain the role of pre-processing techniques in data analysis.
(c) Briefly explain Google Colab. Use Python and any dataset of your choice and explain one example (analysis/visualization) done using Colab.

End of TMA Questions
