MA12005: COURSEWORK

ANNA DAUPARE (AD2865@BATH.AC.UK).

Set date: Wednesday 22nd Nov 2023
Due date: Wednesday 13nd Dec 2023

Plagiarism warning

This is an individual assignment. You should not discuss your work with anyone else. The work which you hand in must be your own. Under no circumstances should you exchange solutions with other students. After you hand your work in, you could be asked by an examiner to explain it verbally as part of the marking process. Any evidence of cheating will entail disciplinary procedures. See http://www.bath.ac.uk/quality/documents/QA53.pdf.

Provided Files

As part of this assessment you are provided with the following files:

- This document, which describes the problem and the assessment procedure.
- The zip file which contains the files you may use as part of the assessment.

Coursework Guidelines

Remember all the way back to our very first lecture? I told you my definition of a data scientist, a person that extracts knowledge from data. This coursework is set to assess your ability to do this *and* to communicate that knowledge to a data layman.

The idea is that you've been given a large set of data, in this case a database of earthquake information. You've been asked to explore, clean and analyse this dataset, draw conclusions and, perhaps most importantly, disseminate your conclusions to a policy maker who will use the information to decide upon certain actions.

The coursework has been split into 5 sections and the majority of these has been further subdivided into set guided tasks and open ended components. It carries 100 marks in total.

You have at your disposal a 'helper notebook'. This is contained in the zip file I provide you with on Moodle. Within the helper notebook is a function that will be helpful in section 4.

Submission Instructions

You should submit a zip file on Moodle that is named *****CW1.zip where ***** is your Bath username. For instance, my submission would be named ad2865.zip. Within this zip file, ensure you have included:

(1) **A single jupyter notebook file**. This should be clearly formatted and well commented to enable the marker to easily understand which question you are answering and the logic behind the commands you invoke. It should also clearly display the results. This single jupyter notebook will cover sections 1-4 of this coursework.

(2) **A single pdf file of no more than 2 pages**. This will detail the report that describes your findings to the policy maker. This can be written in word or any other equivalent word processing document or in LaTeX but *must* be submitted in pdf format. Please do not submit hand written documents. This report should be in an Arial or Helvetica font with size 11 and margins of no more than 2cm.

Questions

In the interest of fairness to all students **I cannot reply to individual questions on this coursework**. I will collect any questions which I receive and answer them (if appropriate) when all students are present in the lectures. I will not answer any questions after 10th Dec, so please have a careful look at the assignment as soon as possible and direct any questions on clarification to `ad2865@bath.ac.uk`. Put "MA12005 coursework query" in the subject line so I can clearly filter.

Marking:

The coursework is 100% of the final grade for this module! Marking will be done in two stages: first, by a number of automated tests that also checks for plagiarism; second, by hand. If your code works as required and looks reasonable to the eye you will receive full marks for implementation. Marks will also be assigned for comments and will be based on how understandable the fully commented code is. For automated tests that fail, partial marks may still be given, but only if style and comments make it possible to examine the code properly.[1] In principle, and within reason, your code will not be marked for efficiency.

Notes on Tutor guidance:

You can work on the coursework during the computer lab sessions (but should of course not work together). This section describes what tutors can, and cannot do, to help you with your coursework.

- **What does xyz mean?** Tutors *can* explain mathematically and algorithmically concepts to you and *cannot* tell you whether your code is correct.
- **Error messages.** Tutors *can* explain what a Python error message means and *cannot* tell you what to do to your program to make the message go away. They *may* (depending on context) be able to ask you questions like "how do you know ..." to help *you* work out what to do to your program.
- **How do I program this?** Tutors *cannot* write your algorithms for you. They *may* (depending on context) be able to say "have you looked at Anna's model solution to Tickable blah blah blah".
- **Tutors can** explain what a piece of Python means in general (but not what it is doing in *your* code).
- **Tutors can** say "that looks like an extremely complicated way of doing xyz: why don't you start again and look at Anna's solution to blah blah blah".
- **Tutors can** say "why don't you insert some print statement so you can see what the values actually are?"

The main idea is that tutors can facilitate your ability to undertake the coursework but not do it for you.

[1] This is meant to reflect the real world: perfectly working code can live undocumented and unmaintained for long periods of time; broken but well-documented code may be fixed; broken, undocumented code is a source of utter despair and will be discarded where possible...

1. Data Exploration (10 marks total)

This question is testing your ability to use the knowledge gained within the lectures for some rudimentary data exploration. You should create your own Jupyter notebook for solutions and comments. This should be in the same style as done weekly in the labs.

Task 1.1. Within the zip file I provided you there is a comma separated variable file. Import the `earthquakes.csv` file as a dataframe. Use the `read_csv` method, which you have seen in your lectures.

[2 marks]

Task 1.2. Calculate the average magnitude of all the earthquakes in the dataset. Find location name and time were the earthquake is the closest to the average magnitude?

[3 marks]

Task 1.3. Identify the five deepest earthquakes in the dataset. Create a DataFrame containing information about these earthquakes, including time, place, latitude, longitude, and depth. Sort the DataFrame by depth in descending order and print the result.

[5 marks]

2. Data Cleaning (30 marks total)

In this section you are required to clean polluted data from the dataframes you obtained in the previous section. You will clean the data in preparation for analysis and interpretation in the following sections.

Task 2.1. Check if there are any `NaN` entries in the dataset.Using the rows where `NaN` are present create a new data frame with the following columns: impact.significance,location.depth, location.distance location.full. Then, remove these entries from the dataset using methods you have learnt in the lectures.

[2 marks]

Task 2.2. Check if there are any duplicates, if there are duplicates, count duplicates and identify duplicate rows. After identifying and displaying the duplicate records, remove all but one record from each set of duplicates in the original dataset.

[3 marks]

Task 2.3. Change 15 integer entries to string entries in 'impact magnitude column' randomly. Identify these misinformed entries and substitute misinformed entries with the median impact.magnitude value.

[5 marks]

Task 2.4. Entries with impact.significance > 1000 or < 0 are considered to be outliers. Find these outliers, and output them in a dataframe, containing time, place, magnitude and significance. If an outlier is negative change it to 0, if the outlier value is larger than 1000 change it to 1000.

[5 marks]

Task 2.5. Create separate dataframes containing the subset of the earthquakes that occurred within a specified magnitude range. Specifically, create one dataframe for earthquakes with magnitudes between 5.0 and 5.9, and another dataframe for earthquakes with magnitudes between 6.0 and 6.9. Each dataframe should contain the following columns: Time, Magnitude, Place, Latitude, and Longitude.

[10 marks]

Task 2.6. Given the nature of the dataset, describe the methodology that you would employ to fill the missing values. Explain the reasoning behind your chosen methodology.

[5 marks]

3. Data Manipulation (20 marks total)

In this section you will conduct some basic data manipulation and summary statistics using the data frames you cleaned in the previous sections.

Task 3.1. Create separate data frames that displays the following descriptive statistics for earthquake data within specific states (5 countries of your choice from csv file):

- Mean earthquake depth,
- Standard deviation of earthquake depths,
- Median earthquake depth,
- Mean earthquake magnitude,
- Standard deviation of earthquake magnitudes,
- Median earthquake magnitude.

[10 marks]

Task 3.2. For each of the countries in Task 3.1, determine the mean earthquake frequency for magnitude ≥ 2 and the average period (T) that elapses between successive earthquakes. Calculate the mean frequency in earthquakes/day. Calculate the mean period by calculating all the inter-earthquake intervals in days, and divide by the number of intervals. Present your results in a data frame.

[5 marks]

Task 3.3. Combine the data frame from Task 3.1 into single dataset and produce new overall statistics described in the previous two tasks for the combined data. Produce new single data frame in the combined statistics for each of the individual locations and the combined location (i.e the total of 6 rows).

[5 marks]

4. Data visualisation (20 marks total)

In this section, you will visualise the data.

Task 4.1. Create a box plot to visualize the distribution of earthquake magnitudes for 3 different states of your choice, over a specific time period of your choice.

[3 marks]

Task 4.2. Create a bar chart to compare the distribution of earthquake magnitudes across multiple regions during a specific time period of your choice. Compare the distribution of earthquake magnitudes by region.

[3 marks]

Task 4.3. Create a scatter plot of earthquake locations, where marker size is independent of magnitude, for all earthquakes. Save the plot as an HTML version to the same directory as the Jupyter notebook you are using.

[3 marks]

Task 4.4. Create a heatmap of earthquake activity over time using your dataset. The heatmap should visualize the frequency of earthquakes in each state over specific time intervals of your choice.

[4 marks]

Task 4.5. Create three different world maps that correspond to 3 time periods that visualize earthquake activity (number of earthquakes) for all states, for the following time periods:

- $t \leq t_1$
- $t_1 < t < t_2$
- $t \geq t_2$

where t_1 = 8th August 2016 at 00:00 , t_2 = 16th August 2016 at 00:00

5. Report

Task 5. Compose an executive summary, less than two pages, detailing the main observations a policy maker should draw from your results. This should be written in plain English, should avoid jargon and be easy for a lay-person to read.

Below are some suggestions that should enable you to have a coherent structure:

- Provide a concise overview of the conclusions derived from the data.
- Explain the key assumptions you have made in the analysis.
- Interpret and draw conclusions from the data.
- Detail any reservations you have in your findings.
- Suggest next steps for a more detailed analysis.

This will be assessed on the following grounds:

- Clarity of language.
- Presentation.
- Whether the arguments are supported by the data.
- Added value.

On that note, should you wish to conduct a more in depth analysis, this is allowed. Any additional code should be included within the Jupyter submission under the title "Section 5". This will contribute to the added value aspect of the marking scheme.

[20 marks]