Manchester Metropolitan University

Department of Computing and Mathematics

ASSESSMENT COVER SHEET 2023/24

Unit Code and Title:	6G4Z0025 - Mathematics and Statistics for Data Science
Assessment Set By:	Ismail Adeniran
Assessment ID:	1CWK100 – 1TEST100
Assessment Weighting:	100%
Assessment Title:	1CWK100 – 1TEST100
Type:	Individual
Hand-In Deadline:	9pm on Friday, 12 January 2024
Hand-In Format and Mechanism:	Upload on Moodle Submission Page

Learning outcomes being assessed:

LO1 Choose and apply appropriate basic mathematical and statistical theory and techniques, utilising industry-standard statistical environment, tools, and languages.
LO2 Select the models and methods most relevant to the solution of a given real-world problem.
LO3 Draw and present conclusions by using appropriate mathematical and statistical techniques.

Note: it is your responsibility to make sure that your work is complete and available for marking by the deadline. Make sure that you have followed the submission instructions carefully, and your work is submitted in the correct format, using the correct hand-in mechanism (e.g., Moodle upload). If submitting via Moodle, you are advised to check your work after upload, to make sure it has uploaded properly. If submitting via OneDrive, ensure that your tutors have access to the work. Do not alter your work after the deadline. You should make at least one full backup copy of your work.

Penalties for late submission

The timeliness of submissions is strictly monitored and enforced.

All coursework has a late submission window of 7 calendar days, but any work submitted within the late window will be capped at 40%, unless you have an agreed extension. Work submitted after the 7-day late window will be capped at zero unless you have an agreed extension. See 'Assessment Mitigation' below for further information on extensions.

Please note that individual tutors are unable to grant extensions to assessments.

Assessment Mitigation

If there is a valid reason why you are unable to submit your assessment by the deadline you may apply for assessment mitigation. There are two types of mitigation you can apply for via the unit area on Moodle (in the 'Assessments' block on the right-hand side of the page):

- **Self-certification**: does **not** require you to submit evidence. It allows you to add a short extension (usually, but not always, seven days) to a deadline. This is not available for event-based assessments such as in-class tests, presentations, interviews, etc. You can apply for this extension during the assessment weeks, and the request must be made **before** the submission deadline.

- **Evidenced extensions:** requires you to provide independent evidence of a situation which has impacted you. Allows you to apply for a longer extension and is available for event-based assessment such as in-class test, presentations, interviews, etc. For event-based assessments, the normal outcome is that the assessment will be deferred to the Summer resit period.

Further information about Assessment Mitigation is available on the dedicated Assessments page: https://www.mmu.ac.uk/student-life/course/assessments#ai-69991-0

Plagiarism

Plagiarism is the unacknowledged representation of another person's work, or use of their ideas, as one's own. Manchester Metropolitan University takes care to detect plagiarism, employs plagiarism detection software, and imposes severe penalties, as outlined in the Student Code of Conduct and Regulations for Undergraduate Programmes. Poor referencing or submitting the wrong assignment may still be treated as plagiarism. If in doubt, seek advice from your tutor.

As part of a plagiarism check, you may be asked to attend a meeting with the Unit Leader, or another member of the unit delivery team, where you will be asked to explain your work (e.g. explain the code in a programming assignment). If you are called to one of these meetings, it is very important that you attend.

If you are unable to upload your work to Moodle

If you have problems submitting your work through Moodle, you can email it to the Assessment Team's Contingency Submission Inbox using the email address submit@mmu.ac.uk. You should say in your email which unit the work is for, and provide the name of the Unit Leader. The Assessment team will then forward your work to the appropriate person. If you use this submission method, your work must be emailed **before the published deadline**, or it will be logged as a late submission. Alternatively, you can save your work into a single zip folder then upload the zip folder to your university OneDrive and submit a Word document to Moodle which includes a link to the folder. **It is your responsibility to make sure you share the OneDrive folder with the Unit Leader, or it will not be possible to mark your work.**

Assessment Regulations

For further information see Assessment Regulations for Undergraduate/Postgraduate Programmes of Study on the Student Life web pages.

Formative Feedback:	Model solutions will be provided on Moodle.
Summative Feedback:	Marks on your completed work will be available through Moodle.

Assessment
Mathematics and Statistics for Data Science (6G4Z0025)

Instructions:

- There are 3 questions.
- All answers **MUST** be given within the provided Interactive notebook.
- Answers without code and/or explanations will receive a score of zero.
- You can add as many cells as you require to provide your answers to each question.
- Ensure you save your work and all changes. Then submit the interactive notebook on the Moodle submission page.
- Submit **ONLY ONE** attempt.

Issues and Support: *See the Moodle page for my contact details under the 'Support' section*

QUESTION 1: Does Brain Weight Differ by Age in Healthy Adult Humans? (36 Marks)

The Brainhead.csv dataset provides information on 237 individuals who were subject to postmortem examination at the Middlesex Hospital in London around the turn of the 20th century. Study authors used cadavers to see if a relationship between brain weight and other more easily measured physiological characteristics such as age, sex, and head size could be determined. The end goal was to develop a way to estimate a person's brain size while they were still alive (as the living are not keen on having their brains taken out and weighed). We wish to determine if there is a relationship between age and brain weight in healthy human adults.

RESOURCES: Brainhead.csv dataset and BrainheadDataDictionary.pdf

1. Import the Brainhead.csv dataset. Review the data dictionary to identify each variable in the dataset as categorical or quantitative. If the variable is categorical, further identify it as ordinal, nominal, or an identifier variable. If the variable is quantitative, identify it as discrete or continuous. `(8 marks)`
2. Create a histogram of brain weight and calculate the appropriate summary measures to describe the distribution. `(3 marks)`
3. Display the distribution of age graphically. `(2 marks)`
4. Describe the distribution of age with a numerical summary. `(1 mark)`
5. Draw side-by-side box plots illustrating the distribution of brain weight by age. `(5 marks)`
    a. `Hint: Step 1. Use ***Frame.filterRows*** to create two data frames - a dataframe for the younger than 46 group and another dataframe for the older than 46 group.` For example, for a data frame called 'furniture_df' where 'desks' are labelled 1 and 'chairs' are lableled '2', I can extract a dataframe with just data for desks using: ***let desks_df = furniture_df |> Frame.filterRows (fun key row -> row?desks = 1)***.
    b. `Hint: Step 2. From each dataframe you created, extract just the Brain column`.
    c. `Hint: Step 3. Create two boxplots and combine them.`
6. Calculate and compare the mean and standard deviation of brain weight by age. `(5 marks)`
7. Describe the hypothesis test you would use to test for a statistically significant difference in brain weight by age. `(2 marks)`
8. Identify the appropriate statistical test for your hypotheses in Deliverable 7, and determine if the assumptions for using this test are met. `(3 marks)`
9. Test for a statistically significant difference in brain weight by age at the 0.05 level. `(3 marks)`
10. Calculate a 95% confidence interval for the difference in the mean brain weight for older and younger individuals. `(2 marks)`

11. Summarise your results about the relationship of age and brain weight in healthy adults. `(2 marks)`

QUESTION 2: Preventing Acute Mountain Sickness with Ginkgo Biloba and Acetazolamide (34 Marks)

Acute mountain sickness (AMS) is a common concern for mountain climbers who ascend higher than 2000 m. Characterized by headache, lightheadedness, fatigue, nausea, and insomnia, AMS is caused by a failure to adapt to the acute hypobaric hypoxia experienced at high altitudes. The drug acetazolamide has been used effectively to treat AMS; however, it has a variety of unpleasant side effects that can reduce compliance to taking it. Previous studies suggested that the herbal supplement ginkgo biloba might also be used to prevent AMS without side effects.

To test this hypothesis, healthy western volunteers who were hiking Mt. Everest were randomised to one of four treatments: placebo, ginkgo biloba only, acetazolamide only or ginkgo biloba and acetazolamide. Treatment group as well as incidence of AMS and incidence of headache for the individuals who completed the experiment are presented in Treckers.csv. We wish to determine if ginkgo biloba is as effective in preventing AMS as acetazolamide.

RESOURCES: Treckers.csv dataset and TreckersDataDictionary.pdf

1. Import the Treckers.csv dataset. Open the data dictionary to identify each variable in the dataset as categorical or quantitative. If the variable is categorical, further identify it as ordinal, nominal, or an identifier variable. If the variable is quantitative, identify it as discrete or continuous. `(12 marks)`
2. Create a subset that contains records of participants who were randomised to take ginkgo biloba only and acetazolamide only. `(2 marks)`
    a. Hint: Filter your original frame by **(fun key row -> row?Trt = 2 || row?Trt = 3)** in your F# code. We will use this subset to complete the rest of the deliverables.
3. Create a new column called TrtChar that takes on the value GinkgoBiloba or Acetazolamide for individuals who were assigned to those treatments. Create another new column called AMSChar that takes on the value `Yes` for participants who developed AMS and `No` for participants who did not develop AMS. (You can filter using `AMS_out` if you wish.) `(2 marks)`
4. What number and proportion of hikers developed AMS? `(2 marks)`
5. Calculate the joint and marginal distributions of treatment and AMS. `(3 marks)`
6. Determine the conditional distribution of the incidence of AMS by treatment. `(4 marks)`
7. Display the results of Deliverable 6 in a side-by-side bar chart. `(1 mark)`
8. What is the appropriate test to determine if the proportion of individuals who develop AMS while taking acetazolamide is the same as the proportion who develop AMS while taking ginkgo biloba? Verify that the assumptions for using this test are met. `(5 marks)`
9. Write the hypotheses for the test you identified in Deliverable 8. `(2 marks)`
10. Summarise your conclusions about the effectiveness of ginkgo biloba and acetazolamide as treatments for AMS. `(1 mark)`

QUESTION 3: What Factors Influence Mammal Sleep Patterns? (56 Marks)

All mammals sleep. As any student who has pulled an all-nighter knows, going without sleep or trying to function on too little sleep has a host of deleterious effects. Yet, for something that is so clearly physiologically important, there is a great variety in sleep needs throughout the animal kingdom from animals that seem never to sleep to those that seem never to wake (cats!?!).
Researchers recorded data on sleep duration as well as a set of ecological and constitutional variables for a selection of mammal species. This data appears in the Sleep.csv dataset. We wish to examine the relationship between dreaming and non-dreaming sleep time in this set of mammal species.

RESOURCES: Sleep.csv dataset and SleepDataDictionary.pdf

1. Import the Sleep.csv dataset. Open the data dictionary to identify each variable in the dataset as categorical or quantitative. If the variable is categorical, further identify it as ordinal, nominal, or an identifier variable. If the variable is quantitative, identify it as discrete or continuous. `(22 marks)`
2. Display the distribution of total sleep for the mammal species in the dataset and describe the distribution with some summary statistics: shortest sleep time, longest sleep time, mean sleep time, median sleep time and standard deviation. `(6 marks)`
3. Plot the relationship between nondreaming and dreaming sleep. Do animals who spend more time in dreaming sleep also spend more time in nondreaming sleep or does dreaming sleep decrease as nondreaming sleep increases? Hint: Use Chart.Point. `(2 marks)`
4. What is the appropriate method to model the relationship between time spent in nondreaming sleep and time spent in dreaming sleep? Verify that the assumptions for using this method are met. `(4 marks)`
5. Determine the regression equation that relates time spent in non-dreaming sleep to time spent in dreaming sleep. Interpret the slope. "Hint: The x-axis variable should be non-dreaming sleep". `(3 marks)`
6. Calculate and interpret the correlation and R2 describing the relationship between dreaming and nondreaming sleep time. Interpret both the correlation and R2. `(4 marks)`
7. If a mammal species experiences 5 hours of nondreaming sleep a day, how many hours of dreaming sleep would we expect that animal to get on average? `(2 marks)`
8. Calculate the difference in the number of hours spent in nondreaming and dreaming sleep for each mammal in the dataset. `(1 mark)`
9. What is the appropriate test to determine if mammals spend the same or different numbers of hours in dreaming and nondreaming sleep? Verify that the assumptions for using this test are met. `(3 marks)`
10. Write the hypotheses for the test you identified in Deliverable 9. `(2 marks)`
11. Conduct the hypothesis test and report your conclusion at the 0.05 significance level. `(3 marks)`
12. Create a 95% confidence interval for the mean difference in the number of hours a mammal spends in nondreaming and dreaming sleep. `(2 marks)`
13. Summarise your findings about dreaming and nondreaming sleep in mammals. `(2 marks)`