SCHOOL OF COMPUTER SCIENCE

TAYLOR'S UNIVERSITY

GROUP ASSIGNMENT (30%)

APRIL 2025 SEMESTER

Module Code : **STATISTICAL INFERENCE AND MODELLING / STATISTICS AND OPERATIONAL RESEARCH**

Module Name : **ITS66804 / MTH60904**

Due Date :

This paper consists of SEVEN (7) pages, inclusive of this page.

Group No : ……..

Tutorial No : ……..

Project Title : ……..

Learning Outcomes:

By the end of this assignment, students will be able to:

1. Choose appropriate statistical tests based on the nature of the data and research questions.
2. Model statistical inferences and predictions using R programming.
3. Visualize outcomes using suitable techniques in R to communicate findings effectively.

Assignment Overview:

In this group assignment, your team will work on a real-world case study dataset sourced from public repositories such as Kaggle. You are required to:

1. Select and download a dataset that is relevant to your interests or aligns with the course objectives.
2. Analyse the dataset to identify the variables and their types (e.g., categorical, continuous).
3. Formulate research questions or hypotheses based on the dataset.
4. Select appropriate statistical tests to answer the research questions.
5. Perform statistical modelling and inference using R.
6. Generate predictions (if applicable) and visualize the results using appropriate visualization techniques in R.
7. Present a comprehensive report detailing your methodology, findings, and interpretations.

Dataset Requirements:

Each group must source its dataset from public repositories such as Kaggle, UCI Machine Learning Repository, or data.gov. The dataset must meet the following criteria:

1. Contains at least 5 variables (a mix of categorical and continuous).
2. Has sufficient observations (minimum 1000 rows).
3. Is relevant to a specific domain (e.g., healthcare, business, social sciences, environmental studies).
4. Includes a clear description of the dataset and its context.

Deliverables:

Each group must submit:

1. A written report (max 10 pages) that includes:
    - Introduction to the case study and dataset.
    - Research questions/hypotheses.
    - Justification for the choice of statistical tests.
    - Description of the modelling approach and its implementation in R.
    - Results, including visualizations and interpretations.
    - Conclusion and recommendations.
2. An R script file (.R) containing all the code used for analysis, modelling, and visualization. The script should be well-commented and organized.
3. A short presentation (10-15 minutes) summarizing the key findings, methodologies, and visualizations.

Rubrics for Evaluation:

Criteria	Excellent (90-100%)	Good (75-89%)	Satisfactory (60-74%)	Needs Improvement (<60%)
Dataset Selection and Understanding	Dataset is highly relevant, well-documented, and thoroughly understood.	Dataset is mostly relevant and well-understood but lacks minor details.	Dataset is somewhat relevant but lacks proper documentation or understanding.	Dataset is irrelevant, poorly documented, or misunderstood.
Research Questions/Hypotheses	Clear, relevant, and well-formulated research questions aligned with the dataset.	Research questions are mostly relevant but lack clarity or alignment in some areas.	Questions are vague or only partially aligned with the dataset.	Research questions are unclear, irrelevant, or absent.
Selection of Statistical Tests	Chooses highly appropriate	Selects mostly appropriate tests	Tests are somewhat	Incorrect or inappropriate test

Criteria	Excellent (90-100%)	Good (75-89%)	Satisfactory (60-74%)	Needs Improvement (<60%)
	statistical tests with clear justification.	with minor gaps in justification.	appropriate but lack proper justification.	selection without justification.
Statistical Modeling & Inference	Accurate and robust modeling with clear interpretation of results.	Modeling is mostly accurate but has minor errors or lacks depth in interpretation.	Modeling contains significant errors or lacks clarity in interpretation.	Poor modeling with major errors and no meaningful interpretation.
Visualization Techniques	Uses highly effective and visually appealing techniques to communicate findings.	Visualizations are mostly effective but could be improved in clarity or design.	Visualizations are basic and fail to fully communicate findings.	Visualizations are poorly designed or fail to communicate findings.
Use of R Programming	Code is clean, efficient, well-documented, and error-free.	Code is mostly clean and functional but has minor issues or lacks documentation.	Code is functional but messy, inefficient, or poorly documented.	Code is disorganized, inefficient, or contains significant errors.
Report Quality	Report is professional, well-structured, and concise with no grammatical errors.	Report is well-structured but has minor issues in clarity or grammar.	Report is poorly structured or contains significant grammatical errors.	Report is unprofessional, disorganized, or difficult to follow.
Presentation	Presentation is engaging, clear, and effectively communicates key findings.	Presentation is mostly clear but lacks engagement or depth in some areas.	Presentation is basic and lacks clarity or engagement.	Presentation is disorganized, unclear, or fails to communicate findings.

Submission Guidelines:

1. Submit the written report in PDF format.
2. Submit the R script as a .R file.
3. Upload all files to MyTimes by the due date.
4. Include a link to the dataset used in your report.
5. Presentations will be conducted during the allocated class session.

Grading Weightage:

- Written Report: 40%
- R Script: 20%
- Presentation: 20%
- Peer Evaluation: 20% (Individual contribution within the group will be assessed via peer feedback forms.)

Important Notes:

1. Collaboration within the group is essential, but each member must contribute meaningfully.
2. Plagiarism will result in a zero grade for the assignment. All sources must be cited appropriately.
3. Late submissions will incur a penalty of 10% per day unless prior approval is granted.
4. Ensure that the dataset you choose is publicly available and properly cited in your report.

Resources:

- Public datasets:
    - Kaggle (https://www.kaggle.com/datasets)
    - UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
    - Data.gov (https://www.data.gov/)
- RStudio (https://www.rstudio.com/)
- Recommended textbooks:

    - "R for Data Science" by Hadley Wickham and Garrett Grolemund
    - "Discovering Statistics Using R" by Andy Field
- Online tutorials and documentation for R packages (e.g., ggplot2, dplyr).

END OF GROUP ASSIGNMENT QUESTIONS