ROBERT GORDON UNIVERSITY ABERDEEN

# Assessment Brief - Coursework

| | |
|---|---|
| **Academic Year** | **2023-24** |
| **Semester** | **2** |
| **Module Number** | **CM2606** |
| **Module Title** | **Data Engineering** |
| **Assessment Method** | **Coursework** |
| **Deadline (time and date)** | **21st of April 2024, 12:00 midnight IST** |
| **Submission** | **Assessment Dropbox in the Module Study Area in Campus Moodle.** |
| **Word Limit** | **6,000 words maximum** |
| **Use of Generative Artificial Intelligence (AI) text** | **Is not authorised** |
| **Module Co-ordinator** | **Mohamed Ayoob** |

## What knowledge and/or skills will I develop by undertaking the assessment?

*You will gain skills on data analysis and appraise the process of designing machine learning models on real-world phenomenon while appraising and the comparing these to allow for predictions about the future behaviour of the phenomenon.*

**On successful completion of the assessment students will be able to achieve the following Learning Outcomes:**

1. *Compare and contrast methodologies that are at the forefront in data cleaning, organisation and integration.*
2. *Design and develop an appropriate Extract Transform and Load (ETL/ELT) process, using a state-of-the-art tool for a given data science requirement.*
3. *Develop and integrate a data engineering solution for a machine learning task adopting and extending methods that are informed by current research and industry practices and adopting the Services provided in the AWS Free Tier.*

**Please also refer to the Module Descriptor, available from the module Moodle study area.**

What is expected of me in this assessment?

# Analysing Tropospheric Formaldehyde (HCHO) gas in Sri Lanka

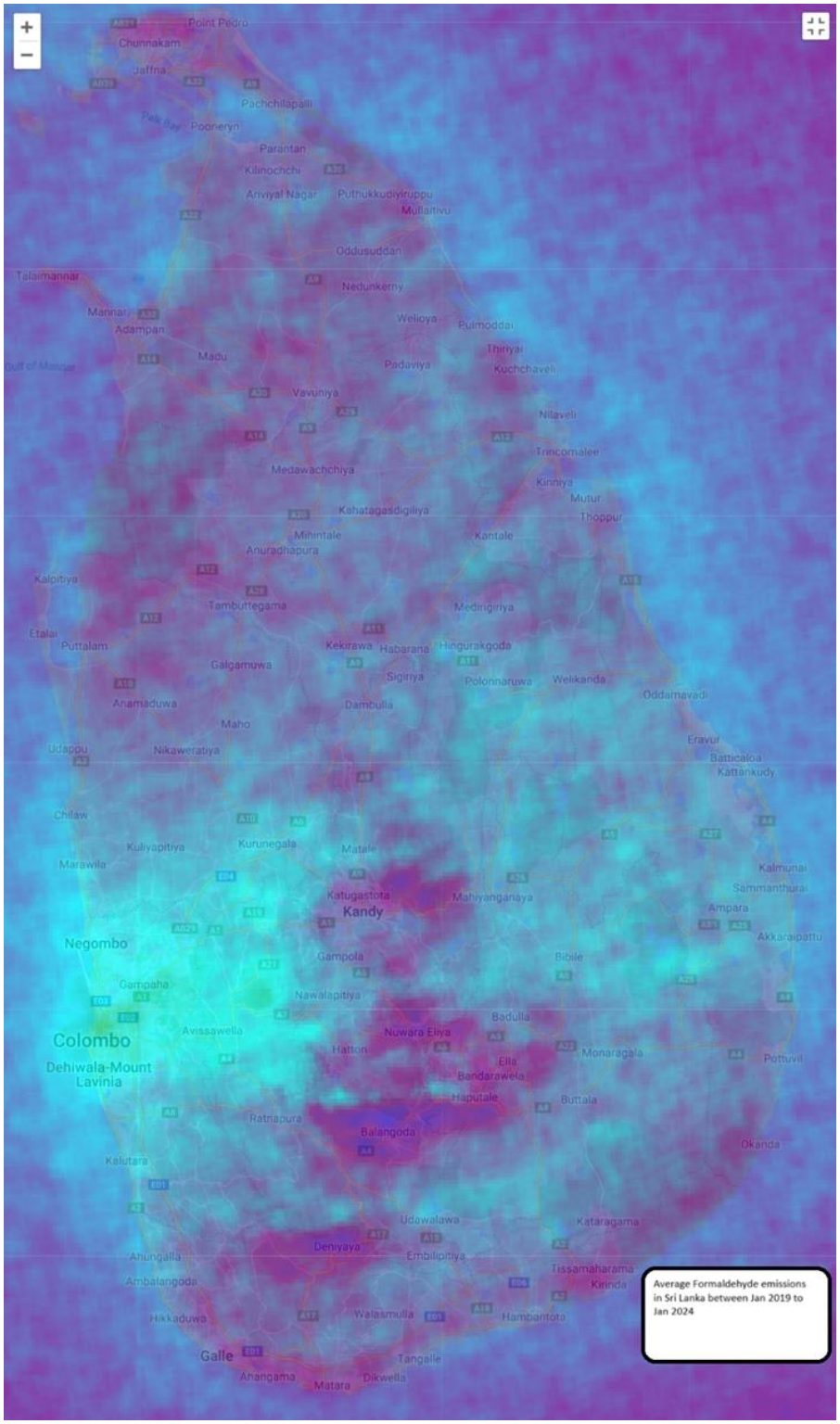


*Figure 1 - mean HCHO concetration map of Sri Lanka between Jan 1 2019 to Dec 31 2023*

## What is expected of me in this assessment?

### Importance of HCHO Monitoring

HCHO is a harmful air pollutant linked to several health problems, including respiratory irritation, eye discomfort, and cancer risks. Understanding spatial and temporal patterns of HCHO is crucial for:

- Assessing air quality and public health risks in different regions.
- Identifying sources of HCHO emissions and developing effective emission control strategies.
- Monitoring the effectiveness of air quality regulations and emission reduction policies.

This coursework provides a valuable opportunity to gain practical data engineering skills while contributing to environmental and public health research. By analysing HCHO data from popular Sri Lankan cities, you can make a meaningful contribution to understanding air quality issues and promoting sustainable development practices.

Satellite observations of formaldehyde (HCHO) are crucial for studying air quality and climate on various scales. This gas plays a key role in breaking down pollutants and forming ozone, a key component of smog. While some HCHO comes from methane breakdown in remote areas, most originates from various human activities like vegetation burning, traffic, and industries. Seasonal changes, fires, and human activities all influence HCHO levels, with its short lifespan making it a good indicator of recent hydrocarbon emissions. Thanks to its advanced technology, TROPOMI instrument in the Sentinel-5P satellite provides detailed images of HCHO distribution, offering valuable insights into air quality and climate processes.

Figure 1. depicts the average heat map of the HCHO concentration in the tropospheric air columns above the airspace of Sri Lanka.

### Objectives of this coursework

- Utilize data engineering skills to analyse HCHO data in Sri Lanka.
- Understand the importance of HCHO monitoring for air quality and climate studies.
- Develop insights into potential HCHO sources and spatial/temporal trends.
- Develop time series prediction algorithms (ML or otherwise)

### Data Description

Link: https://drive.google.com/drive/folders/1xzQ5pIEnaUN2DOyZTqYSJrxFMC8Unx73?usp=sharing

The provided dataset contains daily HCHO measurements (tropospheric HCHO column number density in $mol/m^2$) from the Sentinel-5P satellite from the European Space Agency (ESA) for seven cities in Sri Lanka. The historical data is available from the 2019/1/1 to 2023/12/31 (YYYY-MM-DD). The cities are: 'Colombo Proper', 'Deniyaya, Matara', 'Nuwara Eliya Proper', 'Bibile, Monaragala', 'Kurunegala Proper', 'Jaffna Proper', and 'Kandy Proper'. HCHO is a harmful pollutant linked to respiratory issues and can exacerbate

asthma and other lung diseases. Monitoring helps identify areas with high exposure and assess air quality trends. HCHO plays a role in ozone formation and atmospheric chemistry, impacting climate processes. Understanding its sources and dynamics is crucial for climate models and mitigation strategies. HCHO exposure data informs public health interventions and awareness campaigns to protect vulnerable populations.

Each row of the dataset contains four columns. Table 1. Depicts a small sample of the data from the link. As shown the dataset contains Null values or empty values among other issues. Sometimes certain days would not have readings due to the trajectory of the satellite.

*Table 1 - a sample of the data*

| HCHO reading | Location | Current Date | Next Date |
|---|---|---|---|
| 3.75E-05 | Bibile, Monaragala | 4/1/2019 | 5/1/2019 |
| -1.80E-05 | Bibile, Monaragala | 5/1/2019 | 6/1/2019 |
| 0.000146 | Bibile, Monaragala | 6/1/2019 | 7/1/2019 |
| 2.83E-05 | Bibile, Monaragala | 7/1/2019 | 8/1/2019 |
| | Bibile, Monaragala | 8/1/2019 | 9/1/2019 |

The dataset is also segmented into 3 files. Each file consists of data in the structure discussed above Table 1. However, the file contains the different cities surveyed. The city distribution in each file is as follows:

1. col_mat_nuw_output.csv – contains data about, 'Colombo Proper', 'Deniyaya, Matara', 'Nuwara Eliya Proper'

2. mon_kur_jaf_output.csv – contains data about, 'Bibile, Monaragala', 'Kurunegala Proper', 'Jaffna Proper'

3. kan_output.csv – contains data about, 'Kandy Proper'

## Task(s) – format

Carry out the following tasks and provide the deliverables outlined in the next section below.

1. Data Preprocessing

- Clean and prepare the data: Load the data into a suitable data engineering framework and address missing values, outliers, and format inconsistencies.
- Explore descriptive statistics: Summarize (mean, median, standard deviation) HCHO levels for each city on and across the entire dataset.

## What is expected of me in this assessment?

- Visualize data distribution: Create histograms, boxplots, or other visualizations to understand HCHO distribution.

2. Spatio-Temporal Analysis

- Analyse trends over time: Identify seasonal variations, potential long-term changes, and compare trends across cities, Identify any changes in gas emissions due to the COVID-19 lockdowns.
- Correlate HCHO levels with external factors: Explore potential explanations for observed trends by incorporating data on weather, fire events, or anthropogenic activities. (See if you can join data tables for the locations on rainfall (precipitation or average temperature etc). One source can be (Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC) (noaa.gov)) (**Note**: in this source only Colombo, Kurunegala and Nuwara Eliya are available. It would be sufficient to find any correlations between only these 3 cities.) Feel free to identify other sources.
- Compare spatial patterns: Investigate differences in HCHO distribution between cities and potential contributing factors (e.g., population density). Identify any differences in HCHO emissions due to proximity to the sea, height above sea level etc. of the cities surveyed in question.

3. Machine Learning

- Develop a machine learning model (e.g., ARIMA models, or other time series forecasting) to predict future HCHO levels based on historical data and potentially external factors.
- Evaluate the model's performance using appropriate metrics (e.g., mean squared error, R-squared).

4. Communication and Insights

- Create a comprehensive report summarizing the findings, including visualizations, data tables, and interpretations.
- Acknowledge potential data limitations and uncertainties in your analysis.
- Recommend how your findings could be used for policymaking, or further research.
- Develop a web application (on EC2 or otherwise) or interactive dashboard to visualize and explore the data and the graphs interactively.
- Compare and contrast your findings with similar studies in other regions or on a global scale.

### Deliverables

- Jupyter Notebook or similar script documenting your analysis and code.
- Report (in PDF format) summarizing key findings, visualizations (graphs), limitations and any potential implementation decisions you made using any AWS. (Using AWS is not mandatory for this coursework)
- Viva voce communicating and defending your results and your conclusions and implications.

| How will I be graded? | |
|---|---|
| A number of subgrades will be provided for each criterion on the feedback grid which is specific to the assessment.<br>The overall grade for the assessment will be calculated using the algorithm below*. | |
| **A** | At least 50% of the subgrades to be at Grade A, at least 75% of the subgrades to be at Grade B or better, and normally 100% of the subgrades to be at Grade C or better. |
| **B** | At least 50% of the subgrades to be at Grade B or better, at least 75% of the subgrades to be at Grade C or better, and normally 100% of the subgrades to be at Grade D or better. |
| **C** | At least 50% of the subgrades to be at Grade C or better, and at least 75% of the subgrades to be at Grade D or better. |
| **D** | At least 50% of the subgrades to be at Grade D or better, and at least 75% of the subgrades to be at Grade E or better. |
| **E** | At least 50% of the subgrades to be at Grade E or better. |
| **F** | Failing to achieve at least 50% of the subgrades to be at Grade E or better. |
| **NS** | Non-submission. |

*If the word count is above the specified word limit by more than 10% or the submission contains an excessive use of text within tables, the grade for the submission will be reduced to the next lowest grade.

ROBERT GORDON UNIVERSITY ABERDEEN

# Feedback grid

| GRADE | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| **DEFINITION / CRITERIA (WEIGHTING)** | **EXCELLENT** Outstanding Performance | **COMMENDABLE/VERY GOOD** Meritorious Performance | **GOOD** Highly Competent Performance | **SATISFACTORY** Competent Performance | **BORDERLINE FAIL** | **UNSATISFACTORY** Fail |
| **Question 1 – Data Preprocessing**<br><br>**(Weight 01)** | - Uses appropriate framework for data size and type. - Effectively addresses missing values, outliers, and inconsistencies.<br>- Provides clear explanation of cleaning methods.<br>- Calculates and presents relevant statistics (mean, median, standard deviation) for each city and entire dataset.<br>- Provides clear interpretation of statistics highlighting significant findings.<br>- Creates informative and well-designed visualizations (histograms, boxplots, etc.) that effectively show HCHO distribution.<br>- Provides clear explanation of visuals and identifies key patterns or trends. | - Adequate framework choice considering data size and type.<br>- Addresses missing values, outliers, and inconsistencies but lacks detail in explanation.<br>- Calculates and presents basic statistics but interpretation lacks depth.<br>- Uses appropriate visualizations but explanation lacks specific insights. | - Uses basic framework.<br>- Addresses major data issues but may leave minor inconsistencies.<br>- Calculates some statistics but lacks clarity or completeness.<br>- Visualizations used but lack clarity or completeness in displaying data distribution. | - Framework choice questionable or inefficient.<br>- Limited cleaning applied, leaving substantial issues unaddressed.<br>- Statistics reported but no interpretation or key insights provided.<br>- Visualization choice inappropriate or ineffective in showing HCHO distribution. | - Missing values, outliers, and inconsistencies significantly impact data quality.<br>- Statistics missing or incomplete, hindering understanding of data distribution.<br>- Visualizations poorly designed or missing, hindering understanding of data. | - Data loaded without cleaning, rendering it unusable for analysis.<br>- No attempt to calculate or present descriptive statistics.<br>- No attempt to visualize data distribution. |

| GRADE | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| **DEFINITION / CRITERIA (WEIGHTING)** | **EXCELLENT** Outstanding Performance | **COMMENDABLE/VERY GOOD** Meritorious Performance | **GOOD** Highly Competent Performance | **SATISFACTORY** Competent Performance | **BORDERLINE FAIL** | **UNSATISFACTORY** Fail |
| | | | | | | |
| **Question 2 – Spatio-Temporal Analysis**<br><br>**(Weight 01)** | - Identifies and interprets seasonal and long-term trends in HCHO levels for each city using appropriate time series analysis techniques (e.g., ARIMA, moving averages).<br>- Compares trends across cities and draws meaningful conclusions about potential causes (e.g., urban density, industrial activity).<br>- Specifically analyzes COVID-19 lockdown impact on HCHO emissions and provides evidence-based insights.<br>- Successfully joins data tables with relevant external factors (e.g., weather, fire events, anthropogenic activities) from credible sources, addressing limitations (e.g., data availability).<br>- Conducts rigorous correlation analysis to identify significant relationships between HCHO and external factors, interpreting results with supporting evidence. | - Identifies basic trends but lacks in-depth interpretation or comparison across cities.<br>- COVID-19 impact analysis is present but lacks sufficient detail or evidence.<br>- Attempts to join data tables and performs analysis but may have limitations in addressing data availability or source credibility.<br>- Identifies some correlations but lacks depth in interpretation or exploration of additional factors.<br>- Analysis focuses only on some listed cities or ignores data availability limitations.<br>- Identifies some differences but explanations lack depth or miss key factors.<br>- Conclusions about spatial patterns are basic or unsupported by evidence. | - Identifies some trends but analysis lacks rigor or specific techniques. - Limited comparison across cities and no mention of COVID-19 impact.<br>- Data joining or analysis is limited or unsuccessful.<br>- Correlation analysis is basic or lacks interpretation. - Focuses on only a few external factors or cities.<br>- Limited investigation of spatial differences and lacks consideration of contributing factors. | - Trend analysis is basic or missing entirely. - No comparison across cities or consideration of COVID-19 impact.<br>- No attempt to join data tables or analyze correlations.<br>- No attempt to compare spatial patterns of HCHO distribution. | - Trend analysis is inaccurate or misleading.<br>- COVID-19 impact analysis is irrelevant or harmful.<br>- Correlations are inaccurate or misleading.<br>- Spatial comparison is inaccurate or misleading. | - No attempt to analyze trends in HCHO levels.<br>- No consideration of external factors influencing HCHO levels.<br>- No analysis of spatial distribution of HCHO levels. |

| GRADE | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| **DEFINITION / CRITERIA (WEIGHTING)** | **EXCELLENT** Outstanding Performance | **COMMENDABLE/VERY GOOD** Meritorious Performance | **GOOD** Highly Competent Performance | **SATISFACTORY** Competent Performance | **BORDERLINE FAIL** | **UNSATISFACTORY** Fail |
| | - Includes and analyzes data for all listed cities (Colombo, Kurunegala, Nuwara Eliya) and explores additional sources for other cities.<br>- Investigates and explains differences in HCHO distribution between cities, considering factors like population density, proximity to sea, and height above sea level.<br>- Draws meaningful conclusions about potential causes of spatial variations and supports them with evidence. | | | | | |
| **Question 3 – Machine Learning**<br><br>**(Weight 01)** | - Selects appropriate machine learning model(s) (e.g., ARIMA, LSTM) considering data characteristics and forecasting objectives.<br>- Clearly justifies the chosen model(s) based on theoretical understanding and empirical comparisons.<br>- Effectively preprocesses historical data and incorporates relevant external factors as features.<br>- Applies appropriate feature engineering techniques to enhance model performance.<br>- Trains the model(s) with proper hyperparameter tuning and cross-validation. | - Selects a model but lacks in-depth justification or consideration of alternatives.<br>- Preprocesses data adequately but feature engineering is limited or absent.<br>- Trains the model adequately but hyperparameter tuning and cross-validation are limited.<br>- Uses some evaluation metrics but lacks comprehensive analysis or interpretation.<br>- Offers some interpretations but lacks depth or analysis of key features. | - Uses a basic model without strong justification or understanding of its limitations.<br>- Basic data cleaning performed but no feature engineering used.<br>- Basic model training performed with limited evaluation or discussion of limitations.<br>- Limited interpretation of predictions or feature importance. | - Inappropriately chosen model that does not fit the data or task.<br>- Data preprocessing or feature engineering are insufficient or missing.<br>- Inappropriate or incomplete training and evaluation procedures.<br>- No attempt to interpret model predictions or understand contributing factors. | - No attempt to select or justify a machine learning model.<br>- No data preprocessing or feature engineering are conducted.<br>- No model training or evaluation conducted.<br>- Interpretations are inaccurate or misleading. | - Uses no machine learning approach for forecasting.<br>- Raw data used directly for model training, potentially impacting performance.<br>- Uses a pre-trained model without understanding its limitations and suitability for the task.<br>- No model-based insights provided about HCHO prediction or contributing factors. |

| GRADE | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| **DEFINITION / CRITERIA (WEIGHTING)** | **EXCELLENT** Outstanding Performance | **COMMENDABLE/VERY GOOD** Meritorious Performance | **GOOD** Highly Competent Performance | **SATISFACTORY** Competent Performance | **BORDERLINE FAIL** | **UNSATISFACTORY** Fail |
| | - Evaluates model performance using multiple relevant metrics (e.g., MSE, R-squared, MAE) and interprets results effectively.<br>- Discusses limitations and potential improvements to the model.<br>- Provides clear and meaningful interpretations of the model's predictions and identifies important features influencing HCHO levels.<br>- Explains the relationships between HCHO and external factors based on model insights. | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Question 4 – Communication and Insights**<br><br>**(Weight 01)** | - Presents findings in a well-organized and easy-to-understand report, using logical structure and headings.<br>- Effectively integrates visualizations, data tables, and text to support key points.<br>- Openly discusses potential limitations of data sources, methods, and analysis.<br>- Quantifies and assesses uncertainties in results and their impact on conclusions.<br>- Offers suggestions for future research to address limitations and improve understanding.<br>- Identifies important research questions arising from the work and suggests future studies to address them.<br>- Considers ethical implications of recommendations and potential unintended consequences.<br>- Effectively communicates the importance and impact of findings for air quality, public health, and environment.<br>- Develops a functional and user-friendly web application or interactive dashboard (on EC2 or other platform) to explore and visualize data and graphs.<br>- Compares and contrasts findings with similar studies in other regions. | - Report is organized but may lack clarity or use overly technical language.<br>- Visualizations and data tables are included but not fully integrated with text.<br>- Mentions some limitations but lacks depth in analysis or impact on conclusions.<br>- Uncertainty is acknowledged but not fully explored or quantified.<br>- Offers some recommendations but lacks detail or justification.<br>- Identifies research questions but lacks clear direction or focus.<br>- Communication is clear but may lack impact or engagement.<br>- Creates a basic platform for data exploration but lacks user-friendliness or interactivity. - Design is functional but could be improved for clarity and accessibility.<br>- Mentions other studies but comparison is limited or lacks depth.<br>- Reasons for similarities/differences are not fully explored or supported with evidence.<br>- Acknowledges the importance of HCHO monitoring but discussion lacks depth or specific examples.<br>- Links to impacts are basic or not fully supported by evidence. | - Report structure is basic or inconsistent, making it difficult to follow. - Limited use of visuals or data tables, hindering effective communication.<br>- Limited discussion of limitations, potentially overstating the certainty of findings. - No mention of uncertainties or their implications.<br>- Recommendations are basic or unrealistic, ignoring key findings or context. - No research recommendations provided.<br>- Communication is basic or confusing, hindering understanding of findings. - Tone is unprofessional or biased.<br>- Limited functionality or user interaction in the application/dashboard. - Design may be unclear or hinder user experience.<br>- Basic comparison to other studies provided but lacks insightful discussion or conclusions.<br>- Limited discussion on the significance of HCHO monitoring, lacking in detail or strategic recommendations. | - Report lacks clear organization or direction, making it difficult to understand findings. - Key information is missing or poorly presented.<br>- Ignores potential limitations, implying false confidence in results.<br>- Recommendations are misleading, and not aligned with findings.<br>- Communication is ineffective or inaccurate, potentially misleading the audience.<br>- Application/dashboard is non-functional, unusable, or lacks data visualization.<br>- No comparison with other studies or relevant contextual understanding.<br>- No discussion on the importance of HCHO monitoring for Sri Lanka. | - Report is confusing, misleading, or impossible to follow.<br>- Misrepresents uncertainties or presents misleading conclusions.<br>- No attempt to offer policy or research recommendations.<br>- Findings are not clearly communicated or shared with the intended audience.<br>- No attempt to develop a web application/dashboard for data exploration.<br>- Misrepresents or misinterprets studies, drawing inaccurate conclusions. | - No attempt to create a formal report summarizing the work.<br>- No acknowledgment of limitations or uncertainties.<br><br>- No attempt to communicate findings or their significance.<br>- No attempt to develop a web application/dashboard for data exploration.<br>- Misrepresents or misinterprets studies, drawing inaccurate conclusions. |

***Coursework received late will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost.***

## What else is important to my assessment?

### What is the Assessment Word Limit Statement?

It is important that you adhere to the Word Limit specified above. The Assessment Word Limit Statement can be found in Appendix 2 of the RGU Assessment Policy. It provides detail on the purpose, setting and implementation of wordage limits; lists what is included and excluded from the word count; and the penalty for exceeding the word count.

**What's included in the word count?**

The table below lists the constituent parts which are included and excluded from the word limit of a Coursework; more detail can be found in the full Assessment Word Limit Statement. Images will not be allowed as a mechanism to circumvent the word count.

| Excluded | Included |
|---|---|
| Cover or Title Page | Main Text e.g. Introduction, Methodology, Results, Discussion, Analysis, Conclusions, and Recommendations |
| Executive Summary (Reports) or Abstract | Headings and subheadings |
| Contents Page | In-text citations |
| List of Abbreviations and/or List of Acronyms | Footnotes (relating to in-text footnote numbers) |
| List of Tables and/or List of Figures | Quotes and quotations written within "…" |
| Tables – mainly numeric content | Tables – mainly text content |
| Figures | |
| Reference List and/or Bibliography | |
| Appendices | |
| Glossary | |

**What are the penalties?**

The grade for the submission will be reduced to the next lowest grade if:

- The word count of submitted work is above the specified word limit by more than 10%.
- The submission contains an excessive use of text within Tables or Footnotes.

## What else is important to my assessment?

### What is plagiarism?

Plagiarism is "the practice of presenting the thoughts, writings or other output of another or others as original, without acknowledgement of their source(s) at the point of their use in the student's work. All materials including text, data, diagrams or other illustrations used to support a piece of work, whether from a printed publication or from electronic media, should be appropriately identified and referenced and should not normally be copied directly unless as an acknowledged quotation. Text, opinions or ideas translated into the words of the individual student should in all cases acknowledge the original source" (RGU 2022).

### What is collusion?

"Collusion is defined as two or more people working together with the intention of deceiving another. Within the academic environment this can occur when students work with others on an assignment, or part of an assignment, that is intended to be completed separately" (RGU 2022).

For further information please see Academic Integrity.

### What if I'm unable to submit?

- The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so.
- If you require an extension, you should complete and submit a Coursework Extension Form. This form is available on the RGU Student and Applicant Forms page.
- Further support is available from your Course Leader.

### What additional support is available?

- RGU Study Skills provide advice and guidance on academic writing, study skills, maths and statistics and basic IT.
- RGU Library guidance on referencing and citing.
- The Inclusion Centre: Disability & Dyslexia.
- Your Module Coordinator, Course Leader and designated Personal Tutor can also provide support.

### What are the University rules on assessment?

The University Regulation 'A4: Assessment and Recommendations of Assessment Boards' sets out important information about assessment and how it is conducted across the University.