

36103 Statistical Thinking for Data Science

Assessment Brief Spring 2022

This document outlines each assessment task for STDS, including; key dates, submission formats, weightings, assessment briefs, and associated Subject Learning Objectives (SLO) and Course Intended Learning Outcomes (CILO).

Students should familiarise themselves with this document and refer to it throughout each assessment task.

For help, tips, and more information, head to Canvas and look for each assessment task in the [‘Modules’](#) tab.

36103 Assessment Overview

Assessment summary table with key dates

Deliverable	Description	Type	Weight	Due
Writing documentation	R “Vignette” + 3 comments	Individual	10%	11.59pm Sunday 14 th August
Data analysis project	Part A: Project plan	Group	10%	11:59pm Sunday 4 th September
	Part B: Presentation & Report	Group	20%	Presentation: Saturday 10 th September on-campus Report: 11.59pm Sunday 2 nd October
	Part C: Project review	Individual	20%	11:59pm Sunday 9 th October
Individual project exploration	Blog Post in Canvas	Individual	40%	11.59pm Sunday 6 th November
Total			100%	

All feedback comments and marks will be made available to students in [REVIEW](#)

Additional information

Citations: Proper referencing is mandatory ([APA style](#) preferred) for all externally sourced material.

Submission Requirements:

- All assignments need to be submitted via Canvas unless otherwise instructed
- Submissions must include a title page with subject, assessment task, student name(s) and IDs, date of submission, and the title of the assessment.
- Please use the following file naming format for each submission:
 - For individual assignments: StudentName_AssignmentName_Date.
 - For group assignments: TeamName_AssignmentName_Date
- Please use embedded objects instead of linked objects for content sourced externally.

Length penalties: Submissions exceeding task Length by more than 20% will be penalised (10% of the overall assessment mark). Tables, figures, references, and appendices are not included in word limits.

Plagiarism: Assessments will be checked against both other students' work and external sources for evidence of plagiarism. Any confirmed cases of plagiarism may result in reduced marks for an assessment.

Extensions and late penalties: If unavoidable circumstances arise, students can apply before the assignment due date for an extension of up to five days by sending an email to the subject coordinator. This needs to outline the reason they are unable to submit on time and include an outline of how far they have progressed with the assignment.

- Extension requests submitted after the due date will only be considered in exceptional circumstances; note that work, travel or hardware issues are not valid reasons.
- Unless formal extension dates are agreed upon in writing each late submission will be penalised 10% per day after the due date (to a Pass grade).

Minimum grade requirements: All assessments in a subject must be submitted at a Pass level at a minimum. If a submitted assignment does not achieve a Pass grade on submission, students will be given an opportunity to resubmit for a maximum of a Pass grade at the subject coordinator's discretion.

Special Consideration: Extensions of more than five days require a formal application for [Special Consideration](#) in accordance with university policy.

36103 Assessment Structure Rationale

The assessment structure for this course has been designed to expose you to many of the different communication forms that you will need to use as practicing data scientists. Thus, each assessment task will require you to communicate in a markedly different format.

Assessment Task 1:

AT1 focusses on developing students' ability to write good documentation for peers. Students will identify a statistical, data or coding problem and solve it using data and R. Students will write their chosen problem as an R vignette, aimed at helping other data scientists.

Advice: Take the opportunity to learn something new and that interests you, and is likely to help other students.

Assessment Task 2:

AT2 focusses on exposing students to each key step in a data science project cycle using real-world data. Students will work in teams to propose, execute and critically reflect upon a full data science research project. Students will learn how to develop and pitch a research idea, execute their chosen project, communicate findings both verbally and in report format, and self-evaluate project progress and areas for improvement. AT2 is split into three subtasks.

Advice: Pay close attention to the task description for more details as to how the various subprojects should be targeted.

Assessment Task 3:

AT3 focusses on developing individual students' capacity to perform technical analyses using advanced methods. Students will work individually to write a technical report, deep-diving into a research topic and communicating outcomes to a technical audience.

Advice: follow a format similar to a scientific paper. Read real-world examples get a feel for the format.

Assessment Task 1: Writing documentation

Key information

Weight: 10%

Length: Up to 700 words (not including code)

Type: Individual assignment

Submission format: AT1 Canvas Discussion Forum

Note: This is a pass/fail assessment. If you satisfactorily complete it then you will receive a full 10%. However, actual % marks will be available in REVIEW for reference

Rationale

The ability to create useful documentation is a vital skill for data scientists. In addition to preparing reports and presentations that decision makers and data custodians can use, you may well find yourself creating new software products, often for other people or for teams of people. Being able to document these products is an essential capability.

Form

The R Vignette is a type of long form documentation increasingly being used to document new software packages built using R. They provide more details than simple help pages, and include examples of how specific functionality can be used in a non-trivial scenario. For example, the [tidy data vignette](#) does an excellent job of both defining the problem of messy data, and describing the capabilities of a package that has been designed to solve this problem.

The best examples/vignettes are self-contained. Other users should be able to run all code in the Vignette without having to alter it. Students should try to ensure that datasets in their examples are easily retrievable, either by loading the data directly using R code, or by providing hyperlinks to relevant datasets in-text

Task

In this Assessment item, you will create a blog post that takes inspiration from the R Vignette form. This should be designed to help other STDS students by explaining how to solve a problem you might encounter in this subject--and possibly need to master in your group project--using an existing R package.

You will do this using the following sequence of steps:

1. Consider your strengths and weaknesses in the data science cycle according to self-assessment survey and ongoing experience. Work out an area where you would like to strengthen your skills. CILO 3.1
2. Within that area, think of problem you might have and find a way to resolve it using R. (If you are unsure whether your question is appropriate then be sure to ask Kyle on the relevant 'Help me!' Forum topic in Canvas). Check Canvas for some ideas. CILO 3.1
3. Write up your problem and solution in RMarkdown that follows a structure similar to an R Vignette. Make sure that you include examples of the code you generate being used on a real dataset in your document. CILO 3.1
 - a. Check out rmarkdown.rstudio.com to get started.
4. Submit your Vignette on the Canvas Discussion Page: [Assignment One - R Vignette - Submission and Comments](#). Either:

- a) Attach your R Vignette as a .pdf. To do this, knit your RMarkdown file into pdf format, or if you find it easier, knit into word format and then convert to .pdf.
 - b) Publish your R Vignette to Rpubs and provide the link (URL) here. Note that if you do this, your assignment is available for the general internet public to see. ^{CILO 3.1}
5. Leave at least three comments on other students' vignettes, providing constructive feedback and/or suggest other tools for solving the same problem. Constructive feedback involves more than simply saying: "great post". How could the document have been improved? Is there any advice you can give? Maybe you have spotted a minor mistake in the code? Maybe there is another package that you have come across that solves the same problem in a different way? ^{CILO 3.3}
 6. Feel free to incorporate feedback and submit a second updated version for final assessment. Especially if feedback highlights a key error.

Note: To facilitate feedback and comments for CILO 3.3, students should submit an initial R Vignette prior to the cut-off date, and then submit a final version should any changes be required.

Assessment Criteria

SLO	CILO	Assessment Criteria	Weight
3	3.1	Clarity and brevity in explaining a realistic problem, and the appropriateness of the technique devised to solve it using relevant R packages and functions.	50%
7	3.3	Depth of insight and constructiveness of at least three comments on other blogs suggesting improvements and/or use of different tools.	50%
Contribution to final grade (%)			up to 10

Assessment Task 2: Data analysis project

Key information

This Assessment Task is split into three parts; A) a project proposal, B) execution of the proposed project, communicated in both presentation and technical report format, C) a self-assessed project review and evidence of active participation in the STDS community. See below for further details for each task.

Marks for each part total to 50% of your final mark for this subject.

You must keep a timestamped reflective journal throughout your group project.

You need to do this using the private reflective journal discussion forum that has been set up for you in Canvas under the Discussions tab. The reflective journal provides evidence for marks towards Assessment Task 2C.

Overarching rationale

As a data scientist one of your core duties will be to work with a team of people to analyse complex datasets, and to report back the results of that analysis to stakeholders from a variety of backgrounds, who often have different needs and capabilities. This assessment task will give you a chance to experience the complexity that can often arise in this situation, and gain experience executing a full data science project cycle.

AT2A: Project proposal

Key information

Weight: 10%

Length: Up to 1000 words (not including code samples in appendix)

Type: Group assignment

Submission format: Project proposal submitted in Canvas in PDF format, single submission per group.

Note: Be sure to follow the naming convention defined at the front of this brief.

Rationale

The ability to formulate and propose realistic data science research ideas is a key skill. This involves identifying a problem area, defining suitable research questions, and identifying datasets and techniques that can answer these questions.

Form

Teams will submit a project proposal defining the research problem and questions, with data and exploratory analyses to support the aims and objectives of the project. Teams will also include a self-contained code appendix to showcase code techniques used throughout.

Task

In this Assessment task, you will work in a team to produce a project proposal for your data analysis project. You will do this by following this sequence of steps:

1. Define a broad research area of interest (e.g. public health, climate change, demographic change, finance etc.) and form groups accordingly. ^{CILO 1.4}
2. Work to define a set of well specified research questions for your broad area of interest. Questions should be formulated so they can be tackled by the analysis in AT2B. ^{CILO 1.4 + 4.2}
3. Look for a range of datasets that might help to answer these questions. ^{CILO 1.2}
4. Describe and summarise identified datasets via exploratory data analyses. This includes identifying problematic data (i.e. missing values, outliers) and the use of data visualisations to support proposed questions. ^{CILO 1.2}
5. Refine at least one research question so that it can be answered by a Regression Model. (Note: you can ask other questions too, but you must produce at least one regression model in Assessment 2B.). The proposal doesn't need to include any regression results, only the proposed model structure. ^{CILO 1.4}
6. Write a proposal that summarises the following:
 - a. The rationale and stakeholders for the project. ^{CILO 1.4 + 4.2}
 - b. The research questions. ^{CILO 1.4 + 4.2}
 - c. The range of datasets examined as well as those chosen for the analysis (include details about how you merged the different datasets and an assessment on whether the granularity of the data sources is sufficient to answer your research questions). ^{CILO 1.4, 4.2 + 1.2}
 - d. Exploratory analyses and visualisations to support the proposal. This could also include supporting information from external sources.
 - e. The proposed regression modelling techniques. ^{CILO 1.4}
 - f. Any issues that you anticipate might arise in carrying out the project. ^{CILO 1.4 + 4.2}
 - g. Include an Appendix that contains code used for data acquisition and merger processes, and exploratory analyses. As with AT1, code should be self-contained and able to be executed by teaching staff. Code should be clearly written and well formatted. ^{CILO 1.2}

Assessment Criteria

SLO	CILO	Assessment Criteria	Weight
2	1.4	Clarity in articulating the research questions along with a well-defined proposal for making the invisible visible for a specified set of stakeholders.	30%
4	1.2	Level of expertise using key R functionality demonstrated in the process of data acquisition, and creativity in solving the problem of finding and merging datasets that can answer the research questions.	50%
3	4.2	Eloquence and robustness of the argument used to justify the proposal.	20%
Contribution to final grade (%)			up to 10

AT2B: Stakeholder presentation and technical report

Rationale

Different stakeholders often have different expectations as to how statistical information and models will be communicated to them. Thus, senior managers (who are frequently the final decision makers) often expect a brief presentation, but will rely upon a separate set of recommendations from an in house team of people who are more expert in a domain. These recommendations are often derived from a combination of reports and presentations. The aim of assessment task is to communicate with both types of stakeholders. To get to the communication stage teams will execute their proposed project from AT2A.

Form

Part B is split into two subtasks; B1 is a group presentation aimed at presenting findings to a broader stakeholder audience, B2 is a technical report aimed towards other data science professionals.

Part B1 (Stakeholder presentation)

Key information

Weight: 10%

Length: 6 minutes

Type: Group assignment

Submission format: In-person group presentation during on-campus workshop

Form

Teams will work on their proposed project, resulting in an in-class team presentation aimed at higher level decision makers.

Task

Teams will now work in teams to implement the project proposed in AT2A (taking on board feedback received from the teaching team).

1. Execute the proposed project from AT2A
 - a. Be sure to include at least one regression model.
 - b. Keep track of your code, this will be included as an appendix for the report for AT2B.
2. Prepare your group presentation CILO 4.2, 1.4 + 4.3
 - a. Presentations should briefly introduce the problem and relevance, methods employed, results, and conclusions.
 - b. Remember your target audience – non-technical decision makers.
 - c. Work together to develop a strong slide deck.
 - d. Think carefully about any data visualisations you include. There is a difference between technical visualisations to help you understand things vs visualisations to convey a message to an audience.
 - e. Practice your presentation as a group to ensure you stay on time. Practice runs will help things flow much better when presenting live.

3. Present your results. Presentations will be made during an on campus session. They should be pitched for an audience of senior decision makers who are not necessarily expert in the domain (e.g. a CEO or CIO). CILO 4.2, 1.4 + 4.3
4. Feedback will be given from your cohort and members of the teaching team.

Assessment Criteria

SLO	CILO	Assessment Criteria	Weight
3	4.2	Quality and professionalism of presentation, and responses to questions, that are well tailored for your specified non-technical audience.	30%
2	1.4	Strength of rationale and framing for research questions and preliminary results (including graphical and numerical summaries), along with clearly delineated limitations of the proposed study.	30%
6	4.3	Effectiveness in presenting the data analysis to the defined stakeholder that enables rapid sensemaking and prompts new insights.	40%
Contribution to final grade (%)			up to 10

Part B2 (Technical report)

Key information

Weight: 10%

Length: Up to 5000 words (not including code which should be included in an appendix)

Type: Group assignment

Submission format: Technical report submitted in canvas in PDF format, single submission per group

Form

Teams will write up their executed research project as a technical report. This report should include sufficient detail for the target audience to understand the project aim, assess the suitability of chosen datasets and methods, and critically evaluate the conclusions drawn. This includes model assumptions and evidence that they have been tested. Teams should draw upon feedback from AT2B1. Appropriate use of data visualisation and summary tables should be included to communicate key outcomes. Any externally source data or information should be correctly cited. An appendix containing the code used to undertake the research and produce outcomes should be included, and should be self-contained and well formatted.

Task

Teams will continue working on their project, and create a report aimed at a technical audience.

1. Continue to execute the proposed project from AT2A
 - a. Incorporate any suitable feedback from AT2B1 presentations. This may include changes to the analysis, additional analyses, or including new datasets.
2. Work together to write up the technical report
 - a. Ensure the report follows a logical structure. A useful format to follow would be; executive summary, introduction, methods, results, discussion, and conclusion. CILO 3.1
 - b. Define your research topic, aims and objectives. CILO 2.4 + 3.1
 - c. Describe and summarise the datasets used. CILO 2.4
 - d. Use correct terminology and report appropriate statistics throughout. CILO 3.1
 - e. Ensure you include at least one regression model. CILO 2.2
 - f. Include visualisations and tables. Include figure/table legends and refer to them in the main text. CILO 3.1
 - g. Include an executive summary at the start of the report, this should be a self-contained summary of the project and outcomes, aimed at non-technical readers, this forms a key part of CILO 3.1 CILO 3.1
 - h. Be sure to cite any externally sourced information, including datasets. CILO 3.1
 - i. Include a code appendix, following previous formats where code is self-contained and well formatted. CILO 2.4 + 3.1
3. Finalise and submit the report
 - a. Before submission, check for spelling and grammatical errors. CILO 3.1
 - b. Format the report to enhance readability, use headings and choose appropriate fonts, etc. CILO 3.1
 - c. Submit the report in Canvas in pdf format.

Assessment Criteria

SLO	CILO	Assessment Criteria	Weight
5	2.2	Appropriateness of statistical models with assumptions and mitigation strategies clearly justified.	30%
1	2.4	Clarity and strength of alignment between the exploratory analyses, models generated, features chosen, and their limitations.	40%
3	3.1	Clarity and fluency in communicating your findings to a technical target audience. Soundness of the model interpretation and implications, and professionalism of the executive summary for C-suite decision makers.	30%
Contribution to final grade (%)			up to 10

AT2C: Project Review and Contribution to Community

Key information

Weight: 20%

Length: Up to 1000 words

Type: Individual assignment

Submission format: Evidence and critical self-evaluation submitted via Canvas in PDF format

Rationale

As a data science professional, you will often have to undergo performance reviews, or argue the case for a promotion. It is important that you learn how to make use of evidence to justify claims that you make about your contribution to projects, and of your broader influence in the wider data science community. This will be assisted by an ability to reflect upon events that occur during projects, and the identification of strategies for improvement.

Form

In AT2C, students will present evidence of their contribution to the STDS community, along with a reflective review of their professional practice in the project. Students should follow the recommendations of [this guide](#) in crafting your reflective review. This review will need to make use of evidence generated throughout the project, which should be curated in an appendix. The appendix must contain:

1. At least one R script that you generated during the group project (whether or not it was used in the final project deliverables).
2. A copy of your reflective journal (generated on Canvas) which should include at least 5 reflective entries over the life of your group project.
3. A curated collection of other artefacts that help you to support the claims that you are making in your reflective review (e.g. slack posts, emails, forum posts, minutes and agenda documents, github reports, a data analysis of your contributions etc.)

Students should reference the items from your Appendix in your reflection, explaining how they provide evidence of your claims. That is, you should use this evidence to make your case.

Task

In 1000 words students will construct a reflective review that considers:

1. The group dynamics of your team. CILO 2.4
 - a. You might want to consider some of the following questions: What went well? What did not work so well? What would you try next time to try and generate a better team dynamics? How did your team dynamics affect the statistical modelling process? Was your group dynamic “healthy”?
 - b. Working in teams has its challenges: if you identify issues try to bring them up early and work together as a team to fix them.
2. Your contribution to the broader Statistical Thinking community. CILO 3.3
 - a. How have you helped out people beyond your team? What responses have you made to people’s questions in the forums and slack? Have you asked any questions that provoked an interesting discussion? How have you contributed to the class discussion fora, blogs, or to other open fora such as StackExchange?

Note: If the reflections of your group suggest to the subject coordinator that you have not contributed to the earlier deliverables for Assessment Task 2 then your marks for those items may be adjusted as appropriate. This may include a 0% mark.

Assessment Criteria

SLO	CILO	Assessment Criteria	Weight
7	3.3	Depth of evidence demonstrating your contribution to your group and to the broader Statistical Thinking community.	50%
1	2.4	Insightfulness and criticality in reviewing your contributions and identifying strategies for improvement in future collaborative work to achieve better outcomes.	50%
Contribution to final grade (%)			up to 20

Assessment Task 3: Individual project exploration

Key information

Weight: 40%

Length: Up to 2000 words

Type: Individual assignment

Submission format: Blog post submitted in Canvas

Rationale

While completing Assessment Task 2, you have probably found new research questions and data analysis avenues that were beyond the scope of your group project. In this task you will perform a detailed statistical investigation that extends your group project analysis. This assessment give students an opportunity to work individually to explore more advanced statistical methods.

Form

Working individually, you will extend your explorations in the data analysis that was completed for AT2. You must make use of at least one new statistical method introduced in Module 3, or another method discussed with teaching staff to extend ideas for further exploration that you have discussed in the reflective journal that you created during AT2. You may also want to make use of other methods that you consider relevant, or to use new data sets that extend your group dataset.

To complete this task you will need to write an extended blog post that introduces a new research problem that builds upon the knowledge about the problem domain that you explored in AT2, and defines a new research question. This post should resemble a short academic paper. Students are encouraged to read other academic publications to familiarize themselves with the format.

Task

Your report should take the form of a blog post submitted in Canvas. It should include:

1. The background that led to the new analysis. Formulate or update research questions to match the new analysis. CILO 1.4
2. A justification of the methodologies that used to answer research questions. Including any new datasets. CILO 1.4 + 2.2
3. The results and what they imply for the research questions. Include relevant tables and figures. CILO 1.4 + 4.3
4. Conclusions drawn from the final results, including a critical evaluation of outcomes and limitations. Suggest next steps for further work. CILO 4.3
5. A comparison of how the new analyses enhance the insights gained in AT2. CILO 1.4 + 4.3
6. Reports should include a self-contained and well formatted code appendix. CILO 2.2
7. Any externally source material should be cited throughout.

Assessment Criteria

SLO	CILO	Assessment Criteria	Weight
2	1.4	Novelty and coherence of the new analysis (including aims, new research questions and argumentation).	40%
3	2.2	Soundness of the statistical methodology that also shows evidence of having applied relevant new analytical methods to address the new research question.	30%
6	4.3	Appropriateness of the interpretation applied to the results, where the resulting conclusions are well justified and answer the stated research question.	30%
Contribution to final grade (%)			up to 40

Appendix: CILOs

CIC Gradate Attributes (GAs) & Course Intended Learning Outcomes (CILOs)

Graduate Attributes	Course Intended Learning Outcomes
1. Sociotechnical systems thinking	1.1 Understanding relationships & processes within systems
	1.2 Exploring and testing models and describing behaviours of complex systems
	1.3 Making predictions and informing data discovery
	1.4 Making the invisible visible
2. Creative, analytical and rigorous sense making	2.1 Critiquing trends and theoretical frameworks
	2.2 Exploring, interpreting and visualising data
	2.3 Understanding uncertainty, ambiguity and complexity
	2.4 Designing & managing data investigations
3. Create value in problem solving and inquiry	3.1 Developing strategies for innovation
	3.2 Examining and articulating data value
	3.3 Working together
4. Persuasive and robust communication	4.1 Developing communication skills
	4.2 Engaging audiences
	4.3 Informing decision making
5. Ethical citizenship and leadership	5.1 Becoming a reflective data practitioner
	5.2 Embracing ethical responsibilities
	5.3 Leading data science