

**36103, Statistical Thinking for Data Science**

**Assessment Brief**

**Autumn/2018**

**Assessment Summary:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Deliverable** | **Description** | **Type** | **Weight** | **Due** |
| **Exploration of data skills and issues** | R “Vignette” + 3 comments | Individual | 10% | 11.59pm Thursday March 29th  (i.e. before Easter Friday!) |
| **Data analysis project** | **Part A:**  Project plan | Group | 10% | 11:59pm Monday 29th April |
| **Part B:**  Presentation &  Report | Group | 20% | **Presentation:** Monday 14th May during class  **Report:** 11.59pm Sunday2 27th May |
| **Part C:**  Project review | Individual | 20% | 11:59pm Monday 4th June |
| **Individual project exploration** | Blog post | Individual | 40% | 11.59pm Friday 22rd June |
| **Total** | | | **100%** |  |

All feedback comments and marks will be made available to students in [REVIEW](https://uts.review-edu.com/uts)

**Citation:** Proper referencing is mandatory ([Harvard UTS style](http://www.lib.uts.edu.au/help/referencing/harvard-uts-referencing-guide) 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 named, 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:** 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.

**Extensions and Late Penalties:** If unavoidable circumstances arise students can apply before the assignment due date for an extension of up to one week 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).

**All assessments in a subject must be submitted at a Pass level.** 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 1 week require a formal application for [Special Consideration](https://www.uts.edu.au/current-students/managing-your-course/classes-and-assessment/special-circumstances/special) 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.

Here is an explanation of this rational, and an explanation of the different audiences that you should consider when preparing your deliverables:

**Assessment Task 1:** You are learning how to write help pages for peers.   
**Advice:** write for practicing data scientists who want to learn how to use R packages.

**Assessment Task 2:** You are learning about the process involved in tendering for data analysis projects. Often non-expert decision makers will be involved in this process, as well as teams of expert data science professionals so you must learn how to communicate at multiple levels.  
**Advice:** Pay close attention to the task description for more details as to how the various subprojects should be targeted.

**Assessment Task 3:** You are writing a technical data analysis report – this is the most scientific form of writing that you will submit throughout the semester.   
**Advice:** follow a format similar to a scientific paper.

**Assessment Task 1: Exploration of data skills and issues**

**Weight: 10%**

**Length: 500-700 words (not including code)**

**Due: 11.59pm Thursday March 29th**

**Individual assignment**

**Note: This is a pass/fail assessment. If you satisfactorily complete it then you will receive a full 10% (with the grade you would have gotten recorded in REVIEW so that you can calibrate it).**

**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 (<https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>) 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. It is a good one to emulate in this task!

**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 your classmates, by explaining how to make use of a core statistical modelling technique using an existing R package.

You will do this using the following sequence of steps:

1. Think of a **statistical question** and find a way to answer it using R. (NB. If you are unsure whether your question is appropriate then be sure to ask Kirsty on the relevant [Help me!](https://canvas.uts.edu.au/courses/371/discussion_topics/5114) Forum topic in canvas).
2. Write up your problem and solution in a RMarkdown blog that follows a structure similar to a R Vignette. Make sure that you include examples of the code you generate being used on a real dataset in your document.
3. Post your blog on your CIC Around blog.   
   NB: **Mark your blog post with the STDS-Vignette tag**.
4. Go and read the other Vignettes that your classmates have posted! (Make sure they are from Autumn 2018 – i.e. were posted recently, you should be meeting each other during this exercise!)
5. Leave at least three comments on those blogs that provide constructive feedback and/or suggest other tools for solving the same problem. NB. 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?

**Places to get ideas and help**

1. Look for the **STDS-Vignette** tag on CIC-Around. This will give you some examples of the types of documentation produced in previous runs of this assessment item. What types of posts do you find most helpful? (NB. The instructions for what type of Vignette to produce have changed, so take care that you are producing a document that meets the guidelines above!)
2. Read a number of R Vignettes to get a feel for the format. This will also give you a chance to explore existing R documentation and so learn about a number of R packages and their functionality. Here are some good ones, but feel free to go and find some more:
   1. httr is an excellent tool for collecting data on the web <https://cran.r-project.org/web/packages/httr/vignettes/quickstart.html>
   2. Tidyr vignette: ftp://[cran.r-project.org/pub/R/web/packages/tidyr/vignettes/tidy-data.html](http://cran.r-project.org/pub/R/web/packages/tidyr/vignettes/tidy-data.html)
   3. Dplyr vignettes: <https://cran.r-project.org/web/packages/dplyr/> (In particular see the Introduction to dplyr one! <https://cran.r-project.org/web/packages/dplyr/vignettes/introduction.html> )
   4. curl is another R package for getting data from the web. It is a good one to use if you already use curl but httr is probably a bit more user friendly: <https://cran.r-project.org/web/packages/curl/vignettes/intro.html>
   5. Data.table vignettes: <https://cran.r-project.org/web/packages/data.table/index.html>
   6. Best practices for writing a API package: <https://cran.r-project.org/web/packages/httr/vignettes/api-packages.html>
   7. Getting started with rsdmx, which helps with getting data from the web in XML format: <https://cran.r-project.org/web/packages/rsdmx/vignettes/quickstart.html>
3. Read the R Vignette on Vignettes: <http://r-pkgs.had.co.nz/vignettes.html>   
   This will give you information about how to make actual Vignettes (which could be important if you ever end up writing a package in R!)

**Assessment Criteria:**

|  |  |  |  |
| --- | --- | --- | --- |
| **SLO** | **CILO** | **Assessment Criteria** | **Weight** |
| 3 | 3.1 | Clear and concise explanation of a realistic statistical problem, and a technique that provides a solution for it using a collection of appropriate R packages and functions. | 50% |
| 7 | 3.3 | At least three reflective comments on other blogs that provide constructive feedback on how they could be improved and/or suggest other tools for solving the same problem. | 50% |
| **Sub Total** | | | **100** |
| **Total (10%)** | | | **0 or 10** |

**Assessment Task 2: Data analysis project**

**Weight: 50% (see below for breakdown to different parts and due dates)**

**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.

You **must keep a timestamped reflective journal** throughout your groupwork project. You should do this using the GoingOK tool which is available at: <http://goingok.org/>

**Part A (Proposal)**

**Weight:** 10%

**Length: 750-1000 words (not including code samples which should be included in an appendix)**

**Group Assessment**

**Due date: 11:59pm Monday 29th April**

**Submitted: On Canvas in PDF format by *one group member*. Be sure to follow the naming convention defined at the front of this brief.**

**Task**

In this Assessment item, 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.
2. Work to define a set of well specified research questions for your broad area of interest. (Note: at least some of these should be actionable – see Part B).
3. Look for a range of datasets that might help to answer their questions.
4. Refine at least one research questions so that it can be answered by a **Predictive Regression Model**. (Note: you can ask other questions too, but you must produce at least one regression model in Assessment 2 Part B.)
5. Write a proposal that summarises the following:
   1. the rationale for the project,
   2. the research questions,
   3. the range of datasets examined as well as those chosen for the analysis (include details about how you might be merging different datasets here),
   4. the regression modelling techniques to be employed and,
   5. any issues that you anticipate might arise in carrying out the project.
   6. Include an Appendix that contains code demonstrating the data acquisition and merger processes that you have used to date.

**Assessment Criteria: Part A**

|  |  |  |  |
| --- | --- | --- | --- |
| **SLO** | **CILO** | **Assessment Criteria** | **Weight** |
| 2 | 1.4 | The clarity of the research questions articulated, with a clear intention to make the invisible visible for a well defined set of stakeholders. | 30% |
| 4 | 1.2 | The process of data acquisition demonstrates mastery of key R functionality (e.g. interfacing with APIs) and a creative attempt to solve the problem of finding data that can answer the research questions. | 50% |
| 3 | 4.2 | An eloquent and well justified proposal. | 20% |
| **Sub Total** | | | **100** |
| **Total (10%)** | | | **/10** |

**Part B (Presentation and Report)**

**Weight:** 20%

**Length:**

* **Presentation: 10-15mins (to be negotiated)**
* **Report: 5000 words (not including code which can be included in an appendix)**

**Group Assessment**

**Due date:**

* **Presentation:** **Monday 14th May** during class
* **Report: 11.59pm Monday 27th May** (Submitted using Canvas in PDF format by *one group member*. See naming convention defined at the front of this brief.)

**Task**

Groups will now work in teams to implement the project proposed in Part A.

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. Your challenge in this part of the project will be to communicate with both sets of stakeholders:

1. Presentations will be made during an on campus session (on Monday 14th May). 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).
2. You will have a short time, and one slide, to define your target audience before your presentation begins, and identify the problem that your analysis seeks to address.
3. Feedback will be given from your cohort, members of the teaching team, and from a panel of experts who will take on the role of high end decision makers.
4. Reports should then attempt to incorporate ideas provided by the panel of expert decision makers, and feedback received from the rest of the class.

**Assessment Criteria: Part B (Presentation)**

|  |  |  |  |
| --- | --- | --- | --- |
| **SLO** | **CILO** | **Assessment Criteria** | **Weight** |
| 3 | 4.2 | Quality of presentation, and responses to questions, that are both adapted for a non-technical audience. | 30% |
| 2 | 1.4 | Well formulated research questions, preliminary results that are easy to read and appropriate (including graphical and numerical summaries), and clearly delineated limitations of the study. | 30% |
| 6 | 4.3 | An exemplary exploratory data analysis is presented which provides new insights presented in a format that enables rapid sensemaking. | 40% |
| **Sub Total** | | | **100** |
| **Total (10%)** | | | **/10** |

**Assessment Criteria: Part B (Report)**

|  |  |  |  |
| --- | --- | --- | --- |
| **SLO** | **CILO** | **Assessment Criteria** | **Weight** |
| 5 | 2.2 | Exemplary use of predictive statistical models with assumptions and mitigation strategies clearly justified. | 30% |
| 1 | 2.4 | Evidence of a well articulated alignment between the exploratory analysis, the models generated, features chosen, and their limitations. | 40% |
| 3 | 3.1 | The meaning is communicated to a technical target audience with clarity and fluency. An excellent evaluation of model implications is included along with an executive summary that provides quick recommendations to C-suite decision makers. | 30% |
| **Sub Total** | | | **100** |
| **Total (10%)** | | | **/10** |

**Part C (Project Review and Contribution to Community)**

**Weight:** 20%

**Length: 700-1000 words**

**Individual Assessment**

**Due date: 11:59pm Monday 4th June**

**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.

In Part C of Assessment 2, you will present evidence of your contribution to the Statistical Thinking community, along with a reflective review of your professional practice in the project.

You should follow the recommendations of this guide in crafting your reflective review: <https://intranet.birmingham.ac.uk/as/libraryservices/library/skills/asc/documents/public/Short-Guide-Reflective-Writing.pdf>. This review will need to make use of evidence generated during your groupwork 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 printout of your GoingOK journal 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, a data analysis of your contributions etc.)

You 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.**

**Be careful:** some of the fora you use may have time/space limits! For example, slack does not store messages beyond a limit. You should be storing evidence as you go!

**Task**

In 700-1000 words you will construct a reflective review thatconsiders:

1. **The group dynamics of your team.** 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 dynamics “healthy”?
2. **Your contribution to the broader Statistical Thinking community.** How have you helped out people beyond your group? 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 fora? **(It should be a lot for top marks!)**

NB. If the reflections of your group suggest to 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: Part C**

|  |  |  |  |
| --- | --- | --- | --- |
| **SLO** | **CILO** | **Assessment Criteria** | **Weight** |
| 7 | 3.3 | An exemplary contribution to your group and to the broader Statistical Thinking community, as demonstrated by evidence in the appendix. | 50% |
| 1 | 2.4 | A thoughtful critique of your contributions is provided, along with reflections about how you plan to change your processes in future collaborative work to achieve better outcomes. | 50% |
| **Sub Total** | | | **100** |
| **Total (20%)** | | | **/20** |

**Assessment Task 3: Individual project exploration**

**Weight: 40%**

**Length: 2000 word blog post to CIC Around**

**Due: 11.59pm Friday 22nd June**

**Individual Assessment**

**Rationale**

While completing Assessment Task 2, you have probably found new research questions and 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, so building upon your statistical knowledge with Advanced Topics of personal interest.

**Task**

Working individually, you will extend your explorations in the data analysis that was completed for Assessment Task 2. You must make use of at least one new statistical method introduced in Module 3 to extend ideas for further exploration that you have discussed in the logbook that you created during Assessment Task 2. 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 Assessment Task 2, and defines a new research question. This post should resemble a short academic paper.

Your report should take the form of a blog post on CIC Around. It should include:

1. The background that led to your new analysis and its aims
2. A justification of the methodologies that you are using to answer your question
3. The results that you have obtained and what they imply for your research question
4. Conclusions that you can draw from your analysis
5. A reflection upon how your new analysis enhances the insights gained in Assessment Task 2.
6. **Please mark your blog post with the STDS-Further-Explorations tag**.

**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% |
| 5 | 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% |
| **Sub Total** | | | **100** |
| **Total (%)** | | | **/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 |