SCHOOL OF MATHEMATICS AND STATISTICS
Te Kura Mātai Tatauranga

PROJECT INFORMATION 2020

DATA 301 DATA SCIENCE IN PRACTICE 2020 [2/3]

Course Homepage

Information about the DATA 301 project will be posted on Blackboard, under Assessments/Project, and on the SMS course page:
`http://sms.wgtn.ac.nz/Courses/DATA301_2020T2/`

Project Supervisors

Dr Louise McMillan (Course Coordinator)
Room: Cotton CO 429 Tel: – Email: louise.mcmillan@vuw.ac.nz
Professor Richard Arnold
Room: Cotton CO 538 Tel: 463 5668 Email: richard.arnold@vuw.ac.nz
Dr Binh Nguyen
Room: Cotton CO 362 Tel: 463 8896 Email: binh.p.nguyen@vuw.ac.nz
Professor Alejandro Frery
Room: – Tel: – Email: alejandro.frery@vuw.ac.nz

Project Stages

The project is a major component of DATA 301, worth 30% in total of the final course mark.
Objectives

- To experience working as a team towards a shared objective
- To put into practice what you have learned about processing data
- To experience the challenge of exploring a large, and potentially messy, dataset
- To apply your knowledge of modelling and statistical analysis to answer specific questions based on the data
- To demonstrate your ability to communicate your findings to a wider audience

The task will involve processing one or more related datasets, carrying out exploratory analysis, analysing the data in more detail to answer specific questions, and reporting the results. 4 of the lecturers for this course have each proposed a particular dataset, and will each specify what questions the students need to answer with the analysis.

Project Phase 1

Phase 1 of the project will be a group project. Students will be **randomly assigned** into groups of 3. Each group will submit their preferences for which dataset they would like to work on.
We will then assign the groups to the datasets based on those preferences, and each group will be supervised during the semester by the lecturer who proposed their dataset.

Each group will submit a joint report, which is an interim report for the project. The interim report will be submitted as Assignment 3 on 31st August, in Week 6 of the course. This part of the project will be worth 5% of the course grade, and each group will receive a joint mark.
This report will contain Exploratory Data Analysis and also a brief discussion of any Ethics, Privacy and Security concerns relating to the project. The report will also include a short section that describes what each student in the group contributed to the project.

For the interim report, only one student in the group should submit a copy of the group report electronically.

Project Phase 2
Phase 2 of the project is an individual report, which each student must submit in Week 12. Students will continue to work on the dataset they worked on during Phase 1, but they will work individually. The project supervisors will give each student a specific question or two relating to the dataset; these questions will probably be different for each individual student.
Each student will receive an will receive an individual mark for the final project, which will be worth 25% of the course grade (not including the 5% for the EDA assignment).
However, students should include a revised version of the sections from the interim report in the final project submission, and we expect that the content of those sections will be similar to what has been submitted by the other students in their group.
Individual students should submit their final project reports electronically on the due date.

Datasets

There are four datasets to choose from, each provided by one of the project supervisors. Details of each proposal are attached to the end of this document, but in brief, they are:

- **Louise McMillan – Police data**: NZ Police have made a large number of datasets publicly available, and these can be combined with other national datasets from e.g. StatsNZ to answer questions about crime in different regions of New Zealand.
- **Richard Arnold – Time series data**: Many time series datasets relating to environmental are freely available, and can be analysed jointly so that the value of one time series at some point in the future can be used to predict the value of another time series.
- **Binh Nguyen – NHANES data**: NHANES is a large health dataset from the United States, which includes lifestyle information and diagnostic results. The data can be used to determine the prevalence of major diseases and risk factors for diseases.
- **Alejandro Frery – Parkinson's disease tests**: Patients with Parkinson's are asked to draw a spiral to test the deterioriation of their motor function due to the disease. Features of the spiral images drawn by patients with and without Parkinson's can be used to develop an automated diagnostic tool.

If you have a particular dataset you would prefer to work on for your project that is not on this list, you may propose this, but (a) you will need agreement from the other students in your project group; and (b) **you must get approval** from the course coordinator for your chosen dataset. Please submit your request **by the end of Week 3**.

Important Dates

Teaching takes place between Monday 13 July and Friday 16 October 2020, with a 2-week break from 17 August to 30 August. Note that **this year**, the break is after Week 5 and before Week 6.

Week 1: We will randomly assign students to project groups at the end of this week (we will inform students if there are any changes to their groups).
Week 2: Each group should submit their dataset preferences by the end of Wednesday 22nd July.
Week 3: We announce which groups will work on which datasets.
Week 7: Each student will be told what specific analysis questions they should answer in Phase 2, the individual reports
Week 7: The project supervisors will have a short meeting with each student to check on their progress for the project. You are required to arrange a suitable time and to attend the meeting, which can be

via Zoom if preferred.

There will be **no lectures** in Weeks 11 and 12, although there may be optional tutorials on topics relevant to the projects.

Name	Due Week	Due Date	Worth	Content
Assignment 1	W3		10%	Decision Theory, Ethics, Privacy, Security
Assignment 2	W5		10%	Data Wrangling and Integration, Code Proficiency
Assignment 3	start of W6	31 Aug	5%	**Exploratory Data Analysis for Project – End of Phase 1**
Test 1	W7		10%	Decision Theory, Ethics, Privacy, Security, Data Wrangling & Integration, Code Proficiency (i.e. topics from Assign. 1 & 2)
Assignment 4	W8		5%	Code Proficiency, Bash & Git, Modelling (part 1)
Test 2	W9		15%	Code Proficiency, Bash & Git, Modelling (part 1) (i.e. topics from Assignment 4)
Assignment 5	W10		5%	Modelling (part 2), Data Visualization
Project	W12	15 Oct	25%	**Final report, including revised EDA – End of Phase 2**
Test 3	Assessment Week (W13)		15%	Modelling (part 2), Data Visualization, Communication (i.e. topics from Assignment 5)

Software

We expect most students will choose to use R or Python for your report, though you can also use other languages e.g. SAS, SQL, Javascript, but please discuss this with your project supervisor.
You are recommended to use RMarkdown or Jupyter Notebooks to produce your reports, but you can alternatively use LaTeX or Microsoft Word/LibreOffice Writer if you wish.
You can use http://www.overleaf.com as an easy route to learning Latex if you wish.

Analysis and Modelling

During the first part of the project, you will not be given any specific questions to consider when exploring the dataset, because this phase is focused on **exploratory** data analysis. During the second part of the project, you will be given a specific research question or two, which you will answer by analysing the data.
You have been taught a variety of analysis and modelling techniques as part of your data science major. You **do not** need to use all of them in this project! You should choose analysis techniques that are **appropriate** for the dataset and project question(s) you are considering. These **may be basic methods**, such as barplots and scatterplots, or simple linear regression. Use of more sophisticated methods will not, in itself, gain you additional marks unless those techniques are required for the analysis.

Phase 1 Report Marking Scheme

Assignment 3, which is an interim report on the project, will be submitted as a joint report.

The marking scheme outlined below is a guide. Clearly in some proposals more thought and effort will be needed in some aspects of the proposal (to satisfactorily address the issues) than is indicated in this average marking scheme. Also the list here is not meant to be exhaustive, or relevant to every proposal. It is given to jog your memory.

The interim report should have the following content:

Title [1 mark]

- Project title, names and IDs of students in your project group, date of submission

Background and Data [1-3 pages, 6 marks]

- State which dataset(s) your group worked on, and their source
- Explain briefly why the dataset is of interest, or what questions it could be used to answer; assume that the reader has never heard of your dataset
- State the types of data in the dataset(s) and the structure of the dataset(s). Are the data numerical, categorical, or both? Time series? Coordinates? Diagnostic categories? This does **NOT** need to be an exhaustive list of every variable, just a few comments on the overall types.
- State how complete the dataset(s) are (i.e. how many missing, any structure to the missing data, whether there are errors in the data)
- If you used more than one dataset, state what steps you had to take to integrate the datasets

Ethics, Privacy and Security [1-2 pages, 6 marks]

- Brief discussion of any ethical considerations that apply to your project
- Brief discussion of any privacy concerns that might arise connected to your project
- Brief discussion of what steps you could take to keep your project data and results secure (you do **NOT** need to carry this out, you just need to talk about it in the report)

Exploratory Data Analysis [3-8 pages, 15 marks]

For this section, do **NOT** try to summarize everything you can find in the dataset(s). Select a **subset**, highlighting features that you thought were interesting in the data. The plots do not have to be complicated; simple bar charts and scatter plots are fine.

- Several summary tables and/or plots, each describing one, two or three variables in the data that you thought were interesting
- Explain the definitions of the variables in each table/plot
- Comment on the main features of each plot
- Include suitable labels and keys for each plot
- Make sure all plots would be readable if printed in black & white, and adjust the point sizes and/or line thicknesses to improve readability
- Lay out all tables so that they are clearly readable and clearly labelled, and do not use excessive significant figures

Individual Contributions [1 page, 2 marks]

- State what contribution each member of the group made to the data preparation, the analysis and the report

Overall Report [5 marks] These 5 marks will be awarded for overall presentation, clarity and quality of the report. In particular, marks will be awarded for

- A clear logical layout
- Keeping to the page limits for each section
- Key facts being easily located
- Readability of tables and plots

- Clarity of expression [Note: for non-native speakers of English: your English does not need to be perfect, it is the logic and correctness of your presentation that is most important. Nevertheless you are advised to get someone to proof-read your proposal.]
- Clear explanation of how your choice of exploratory plots and tables is relevant to your project, and how the ethical considerations apply to your project (i.e. not just a set of generalities)
- Make sure each time you use/refer to someone else's work you cite the source in the text, and include the reference in the list at the end. It does not need to be a long list; you may only need one or two references.
- *Referencing should be correctly done*: a complete list of references must be included. You can use any referencing style you wish; APA is fine if that's what you like.

Total: 35 marks
You will be expected to include a revised version of the Background, Ethics and EDA sections in the final project report; you do not have to rewrite those sections from scratch. You will be expected to consider any **feedback** you have received for this first report when revising it for the final report, and this will be taken into account when the final report is marked.

Phase 2 Report Marking Scheme

Your mark for Phase 2 of the DATA301 project will be fully based on your individual project report. The report should have the following content:

Title

- State the project title, your name and ID, the names of the other people who were in your project group in Phase 1, the date of submission

Executive Summary [1 page, 5 marks]

- A short summary of your research goal, methods and findings
- Use non-technical language, for a reader with no background in statistics
- 1 page **maximum**

Table of Contents (optional)

Background [1-2 pages, 3 marks]

- Describe the background of the project
- State the aims of the analysis, the specific question(s) you were asked to answer in the Detailed Analysis

Data Description [1-2 pages, 5 marks]

- State the types of data in the dataset(s) and the structure of the dataset(s). Are the data numerical, categorical, or both? Time series? Coordinates? Diagnostic categories? This does **NOT** need to be an exhaustive list of every variable, just a few comments on the overall types.
- Specifically state the attributes of any variables that are used in the Detailed Analysis, in particular for any categorical variables explain the meaning of their categories
- State how complete the dataset(s) are (i.e. how many missing, any structure to the missing data, whether there are errors in the data)

Ethics, Privacy and Security [1-2 pages, 2 marks]

- Include this section from the Phase 1 group report, but **revise it** in response to the feedback that you received about that report
- Brief discussion of any ethical considerations that apply to your project
- Brief discussion of any privacy concerns that might arise connected to your project
- Brief discussion of what steps you could take to keep your project data and results secure (you do **NOT** need to carry this out, you just need to talk about it in the report)

Exploratory Data Analysis [3-4 pages, 5 marks]

- Revise the style and content of the EDA section from the Phase 1 group report in response to the feedback you received about that report
- Select a **subset** of the results from the Phase 1 group report. You do **not** need to include every plot from the EDA assignment. You may want to choose particular plots that turned out to be most relevant for the detailed analysis.
- A few summary tables and/or plots, each describing one, two or three variables in the data that you thought were interesting, or were relevant for the detailed analysis
- Tables can be summary statistics of individual variables, contingency tables for pairs of categorical variables, or results tables for basic analyses
- Explain the definitions of the variables in each table/plot
- Comment on the main features of each plot
- Include suitable labels and keys for each plot
- Make sure all plots would be readable if printed in black & white, and adjust the point sizes and/or line thicknesses to improve readability
- Lay out all tables so that they are clearly readable and clearly labelled, and do not use excessive

significant figures

Detailed Analysis Results [3-4 pages, 15 marks]

- Explain what analysis techniques you used, and why they were suitable for the data and for the question you needed to answer
- Explain any steps you took to manipulate the data, e.g. converting to logarithmic scale, standardizing continuous variables, etc.
- Explain what you did to account for any missing data
- For each part of the analysis, provide plots/tables/diagrams of the results that will help the reader to understand the results of the analysis
- Make sure all plots/tables/diagrams are readable and clearly labelled
- If you have been asked to include predictions, state these clearly
- Include estimates of the **uncertainty** for any results or predictions
- Comment on any biases that you think might exist in the data or the results

Conclusions and Recommendations [1-2 pages, 5 marks]

- State the main conclusions of your work: if you have already presented these in the results, then this is a very short chapter.
- State any recommendations that you would make in light of your findings
- Discuss the limitations of your findings
- (optional) State what future work might be suggested by what you have done

Reference List (optional)

Appendices (optional) Put material here that you want to include in your report, but which is not necessary for every reader to see: only those readers wanting to see a bit more technical detail, or to see tables of data and results that are too complex to include in the main report.

Overall Report [5 marks]

These 5 marks will be awarded for overall presentation, clarity and quality of the report, with the same criteria as for the interim report.

Total: 45 marks

The project report, using font size 12, should be around 10-15 pages including references but excluding any appendices. There should only be about 8 pages of text, i.e. you should not fill 15 pages unless most of them are plots! Excessive length will be penalised. Shorter is OK, as long as the meaning is clear.

The Background and Data Description sections may be expanded versions of the Background and Data section in the Phase 1 group report. The Ethics, Privacy and Security and Exploratory Data Analysis sections are expected to be revised versions of the sections in the Phase 1 report.

The Executive Summary **must not be** longer than one page, and it should be written in a non-technical style, as if intended for readers who do not have a background in statistics or data science.

Proofread your report carefully – and give it to someone else to read. If English is not your first language that won't affect your grade – but what you do write must be clear. (Native English speakers: don't embarrass yourself!)

Avoid plagiarism – if you include any material in your report that appears to be your own because you haven't signalled it comes from somewhere else, then you are committing plagiarism.

The Ethics, Privacy and Security, and Exploratory Data Analysis sections are designed to be worked on as a group task, so these restrictions **do not apply** to those sections. For the Background and Data Description sections, you should revise those sections to be more specific to the individual question you are answering with the analysis, though it is expected that these sections will be similar to the

other students in your group.
Talk to your project supervisor if you're unsure. Plagiarism is the most serious academic offence: you can end up with a zero on your report, a record in the University's Academic Misconduct Register, and maybe worse (suspension or expulsion from the University).

Code Submission

You will need to provide a copy of your code files, or a link to a github/gitlab repository containing your code, when you submit the report. The code will not be marked, but we can use it to check what methods you used to prepare the data and run the analysis.
Note that if you need to perform any data cleaning, this should be **documented**: ideally, any data edits should be scripted in your code, but if you have to manually edit any of the spreadsheets, you should **keep notes** of what you did. This is important for traceability and replicability of your results.

Marking

Each group report, and each individual final report, will be given a provisional grade by the lecturer who has supervised that project, and will also be marked by one of the other lecturers from the course.

Use of Turnitin

Student work provided for assessment in this course may be checked for academic integrity by the electronic search engine `http://www.turnitin.com`. Turnitin is an online plagiarism prevention tool which compares submitted work with a very large database of existing material. At the discretion of the Head of School, handwritten work may be copy-typed by the School and subject to checking by Turnitin. Turnitin will retain a copy of submitted material on behalf of the University for detection of future plagiarism, but access to the full text of submissions is not made available to any other party.