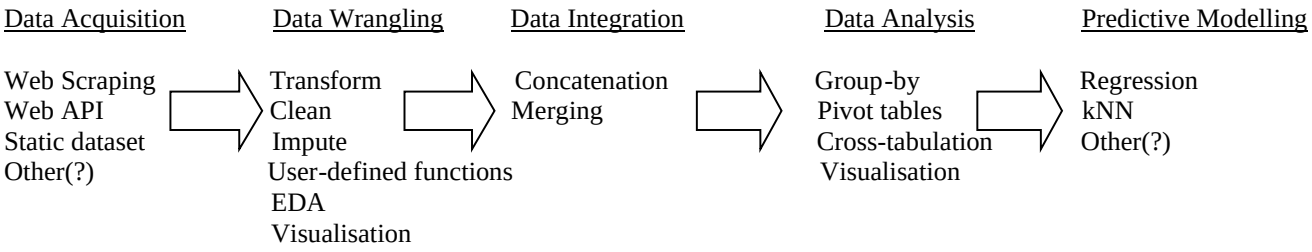
Project 2

Deadline:	Hand in by midnight Wednesday, 15 April 2020.
Evaluation:	25% of your final course grade.
Late Submission*:*	See Course Guide
Work	This assignment is to be done **individually**.
Purpose:	Implement the entire data science/analytics workflow. Learn to correctly apply and reason about using different machine learning techniques to solve real-world problems. Gain skills in extracting data from the web using APIs and web scraping. Build on the data wrangling, data visualization and introductory data analysis skills gained up to this point as well as problem formulation and presentation of findings. Learning outcomes 1 - 5 from the course outline.

Project outline:

This project requires that you apply machine learning techniques taught so far to build predictive regression models on current, topical and original data from your chosen domain. You are expected to carry out an entire data science/analytics workflow by: (1) acquiring data from multiple sources, (2) performing data wrangling, (3) integrating the data, (4) conducting analysis to answer some key research questions and finally (5) perform predictive modelling.



The data should primarily come from sources such as web APIs or scraped web pages. This data can also be combined with a static datasets found in various repositories if needed, or datasets you used from your project 1. The important point is that you are predicting continuous valued outputs and are entirely free to choose a domain or a combination of domains that interest you.

Project Requirements:

Project details:

- Each student must work on prediction analysis from different domains/data sources. Therefore, once you have chosen your domain, you must register it on a Google Docs document linked on Stream. Do this as soon as possible in case someone else wants to do the same domain/data source (first-in-first-served).

Suggestions and questions to consider in you experiments:

- Build regression and kNN models and compare their outputs.
- Experiment with models using different feature types. Which features are most effective? Why?
- Experiment with kNN using different distance metrics and different values of *k*, and compare. Which values of *k* are most robust for the size of your dataset and your problem domain? Are variables in your data having different scales affecting the algorithm's accuracy? How have you tried to overcome this?
- Experiment with linear, multiple linear and polynomial regression models and compare. At what point does a regression model become too complex and no longer captures the true relationships in the data?
- How reliable are your prediction models? What do the confidence intervals and prediction bands tell you? Could you recommend this predictive model to a client? Would you expect this model to preserve its accuracy on data beyond the range it was built on?

Submit a Jupyter Notebook that contains your most integral parts of analysis, together with a thorough description of findings. This notebook will be the one that is marked, but you can submit other notebooks as an appendix. **The Python code in the notebook must be entirely self-contained and all the experiments and the graphs must be replicable**.

Do not use absolute paths, but instead use relative paths if you need to. Consider hiding away some of your Python code in your 'final notebook' by putting them into .py files that you can import and call. This will help the readability of your final

notebook by removing unnecessary python code that can clutter and distract from your actual findings and discussions.

You may install and use any additional Python packages you wish that will help you with this project. When submitting your project, include a README file that specifies what additional python packages you have installed in order to make your project repeatable on my computer, should I need to install extra modules.

Your notebook must have a heading, abstract (a brief summary of your project together with key findings), an introduction to your research context and research questions, data sources, then the body of your experimental findings, explanations and discussions of the findings and a conclusion. Run your text through an IPython Notebook spell-checker extension.

NOTE: Topics of web scraping, using web APIs and kNN algorithms will be covered in weeks 5 and 6. Therefore, begin your assignment as soon as you can using concepts covered thus far. Once material in weeks 5 and 6 is covered, you will be able to complete all remaining components of this assignment.

Marking criteria:

Marks will be awarded for different components of the project using the following rubric:

Component	Marks	Requirements and expectations
Data Acquisition	15	Diversity of sources: data from a web API and data scraped from a web site should be included to get maximum marks; appropriate use of merging and concatenation.
Data Wrangling	10	Thoroughness in data cleaning, visualisations, handling of missing values, outliers.
Data Analysis	20	Quality of your exploratory data analysis and the presentation of the characteristics of the data, discussion of assumptions being made if any. Formulation of the problem as a machine learning problem and the diversity of techniques used to achieve this. Presentation of findings.
Predictive Modelling	40	Diversity of experiments. Quality of the evaluation, comparisons and interpretation of results.
Originality	15	Originality of the datasets, research questions and the code. The degree to which the problem domain is topical and presented in an interesting way.

Hand-in: Make sure that the notebook you submit has all the outputs embedded. Also, export your notebook into HTML. Z**ip**-up all your **notebooks (.ipynb and .html)** and **dataset(s)** you have chosen, as well as any other **.py files** you might have written, into a single file and submit through Stream. Do not email your submission to the lecturer unless there are problems with the submission site.

If you have any questions or concerns about this assignment, please ask the lecturer sooner rather than closer to the submission deadline.

****

*** Plagiarism ***

****

It is mandatory that any assessment items that you submit during your University study are your own work. Massey University takes a firm stance on academic misconduct, such as plagiarism and any form of cheating.

Plagiarism is the copying or paraphrasing of another person's work, whether published or unpublished, without clearly acknowledging it. It includes copying the work of other students and reusing work previously submitted by yourself for another course. **It also includes the copying of code from unacknowledged sources**.

Academic integrity breaches impact on students as it disadvantages honest students and undermines the credibility of your qualification. Plagiarism, and cheating in tests and exams will be penalised; it is likely to lead to loss of marks for that item of assessment

and may lead to an automatic failing grade for the course and/or exclusion from reenrolment at the University.

Please see the Academic Integrity Guide for Students on the University website for more information. The Guide steps you through the University Academic Integrity Policy and Procedures. For example, you will find definitions of academic integrity misconduct, such as plagiarism; how misconduct is determined and managed; and where to find resources and assistance to help develop the skills of academic writing, exam preparation and time management. These skills will help you approach university study with academic integrity.