PROJECT 4

Deadline:	Submit by midday Monday, 22 June 2020
Evaluation:	35% of your final course grade.
***Late Submission**:*	**No late submissions accepted since this is the of the semester.**
Work	This assignment is to be done in groups of up **three** students. You will need to fill out and submit a form (to be provided) indicating your contribution to the project. You will be asked to evaluate your group members' as well as your contribution to the project. Identical grades are not guaranteed for each student in a group.
Purpose:	To work in a group setting and to apply machine learning, data mining, visualisation and data sense-making skills learned so far in class, on a chosen real-world problem. Create an artefact/software that demonstrates your work and present this to the class. Learning outcomes 1 - 5 from the course outline.

You are expected to come up with topics for your group at the earliest possible stage so that you can commence work on development. List the chosen topic for your project on the Google Doc. Preferably, discuss your chosen topic and what it is you plan to develop with the teaching staff before commencing work.

Groups must be formed by May 18 and the class Google Doc listing your members and proposed topic must be filled out. Failure to do so will lead to you being assigned randomly to groups.

****

***** 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Plagiarism is the copying or paraphrasing of another person's work, whether published or unpublished, without clearly acknowledging it.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.

PROJECT OUTLINE:

Create a data science-related notebook, a standalone application, web application or any other kind of artefact with which you apply machine learning and data mining/analysis techniques to a chosen real-world problem domain.

Some ideas for possible projects:

1. Current events: COVID19 data analysis; macroeconomic impacts of COVID19.
2. Time-series analysis, time-series forecasting.
3. Recommender engine: create application for making recommendations based on user preferences.
4. Fitness data: analysis of your personal or some group's FitBit data.
5. Twitter: sentiment analysis, text classification, semantic analysis, network visualisation, geospatial visualisation, data storage etc.
6. Facebook: network visualisation, geospatial visualisation, network analysis, natural language processing, data storage etc.
7. Data journalism: data visualisation – implementation of interactive graphs (web enabled), infographics.
8. A live Kaggle competition problem dataset https://www.kaggle.com/competitions (see notes below)
9. Web app that performs some data-related service.
10. Process mining.
11. ...or something entirely different.

Topics **NOT** to cover:

1. Currency markets, BitCoin, share market stock prices
2. Closed Kaggle competition datasets
3. Previously researched topics for which there are existing notebooks
4. Definitely NO to the TITANIC dataset

OTHER NOTES

DATA SOURCES

This is a recommendation, not a requirement: be as original as you can with your data sources. Some datasets are very popular and have come up repeatedly in assignments over the years. Unfortunately, because they are popular there are a lot of online sources that have scripts published for those datasets. In many cases, related assignment submissions involve some form of plagiarism. While the internet is a big place, we have seen a lot of these scripts before and it is easy to catch. Unless you are going to do something genuinely novel with a well-used data source (you will know it is well-used if you can easily find python kernels for it), avoid these data sources. The safest bet is a dataset that is integrated from multiple disparate sources.

WARNING ABOUT CHOOSING A KAGGLE DATASET

Discuss this with the lecturer first. A high standard is set when marking Kaggle-related submissions. If you use a Kaggle dataset, we recommend you do not look at related Kaggle kernels as there can be a temptation to copy what you see. Copying without attribution is plagiarism which could lead to zero marks for this assignment. Be aware that markers are familiar with Kaggle kernels, in part due to marking assignments for other papers and cohorts. We will also be looking through related kernels prior to marking.

TECHNOLOGY

You are encouraged to use Python; however, this is not an absolute pre-requisite for all parts of your project.

If you choose to build a GUI based application, Python does possess libraries that facilitate this; however, you can use Qt or technologies like .NET which allows you to call your Python methods that implement the logic in your application.

In previous years, some students have created web-based applications which have front-end and back-end components that both serve webpages and perform some data science related tasks. If you have web development skills, then you are encouraged to pursue this. It is sufficient that your application run on localhost.

PRESENTATIONS

We will go ahead and conduct presentations despite the current constraints. We will aim for live team presentations over Zoom. Each person in the group will need to present. The presentations will be short and to the point. We would like you to aim for a presentation using only a handful of power point slides, lasting up to 15 minutes, or an application demo lasting up to 20 minutes. Make your presentation **interesting**. Don't focus on technical details. Consider your audience to be tech-savvy executives. Focus instead on the story that you are trying to tell and sell to the audience/decision makers. The presentations will be marked in part by your peers.

PROJECT REQUIREMENTS:

Make sure you do these four things:

1. Submit all your code, experimental code in a mixture of .py and Notebook files as is appropriate for each project. Each project should submit at least one Notebook that contains all the key findings and summaries.

2. Submit a separate document (or include this at the top of a notebook) that details what each team member contributed to the assignment. Not all contributors will be awarded the same mark. Each team member must submit their own version of how each team member contributed.

3. Each member of the class will be marked individually

4. Watch and mark others’ presentations

MARKING CRITERIA:

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

Component	**Marks**
Project presentation not exceeding 15 minutes, or a demonstration not exceeding 20 minutes.	25%
Project python code (or other non-Python code), Notebooks, application of data science, substance and difficulty of the work undertaken.	50%
Originality and creativity	25%

Hand-in: **Zip**-up all your **notebooks, python and other application source files** into a single file. Submit this file via **stream.**

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