Project 3

Deadline:	Submit by midnight Tuesday, 27th of September 2016.
Evaluation:	25% of your final course grade.
Late Submission:	10% deduction from your final mark per day late.
Work	This assignment may be done in groups of up to **two students**.
Purpose:	Learning outcomes 1 - 5 from the course outline.

Project outline:

This project uses real-world GPS tracking data from buses and heavy vehicles on the northern motorway and the northern expressway routes. The data is sensitive and covered by an NDA agreement signed with its owners, Eroad (http://www.eroad.co.nz/). By doing this assignment **you are bound by this agreement** and are expected not to make the data available to anyone, and must destroy the datasets following the completion of this assignment.

This project is split into two parts. This first part will require you to perform data wrangling, integration, analysis and emphasise the predictive modelling. The second part will require you to perform cluster analysis.

The assignment is open-ended. It is up to you to devise what variable(s) you would like to build predictive models on. And it is up to you also to decide what to cluster the data on.

The dataset comprises of 1600 files that have been zipped. Each file represents reading from one vehicle (bus or a truck). The file 'Data Definitions' explains some features of the dataset, 'Northern Motorway Road Segments.csv' links vehicle IDs with definitions of whether a vehicle is a bus or a truck.

Part 1 (supervised learning)

You are asked to consider the Eroad dataset and to creatively formulate a classification problem, for which you can apply prediction modelling based on material covered in class so far.

You are urged to be creative in how you formulate the problems you want to predict. ie. Can you predict what day of the week the buses and trucks are driving on, given their speed, north/south bound direction etc.? Can you predict if the readings are for a bus or a truck? Given speed and time of the day, can you predict the direction? These are some ideas, which you are encouraged to explore and build upon, and feel free to derive and generate new features based on the ones that exist. You could also consider combining this dataset with weather data for those days from an API, which could provide even more questions.

One possible question that Eroad has asked is and which you may consider if it can be answered is: "*The Northern Express route runs from XXX to the Auckland Harbour Bridge. This allows buses to run freely down the motorway and avoid the lengthy delays from the main motorway. Using EROAD data, show how opening up this route to heavy traffic will not severely impact existing flow of public transport, and provide a safer environment for heavy vehicles that are not able to maintain safe following distances on the main motorway.*"

You are also encouraged to use Python's scikit-learn module for machine learning and try using other algorithms. There are many other Python implementations of machine learning algorithms such as Neural Networks (PyBrain) which are not implemented in scikit-learn. You are encouraged to use some of them too.

Part 2 (unsupervised learning)

Formulate a problem using the Eroad data that enables you to perform cluster analysis. ie. Without using bus/truck labels, can you accurately cluster the vehicles into distinct groups that describe trucks and buses? Are traffic patterns for the vehicles distinct enough that they cluster into different days of the week? Use k-means to help you answer some of these and other questions, but also explore using other cluster algorithms from Python's scikit-learn module.

Project Requirements:

The Python code in the submitted 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 notebook by putting them into .py files that you can import and call. This will help the readability of your final notebook by not allowing the python code to distract from your actual findings and discussions. Do not dump dataframe contents in the notebook – show only 5-10 lines at a time – as this severely affects readability. 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.

Structure your notebook appropriately like a report. Run your text through a spell checker extension.

Marking criteria:

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

PART ONE (80%):

Component	Marks	Requirements and expectations
Data Wrangling	20	Quality of data transformation, merging, cleaning, and preparation for modelling.
Data Visualisation	20	Appropriate use of diverse visualisations for communicating results.
Predictive Modelling	40	Diversity of experiments and questions asked of the data. Quality of experimentation, analysis, interpretation and conclusions.

PART TWO (20%):

Component	Marks	Requirements and expectations
Cluster Modelling	20	Diversity of experiments and questions asked of the data. Quality of the evaluation, comparisons and interpretation of results.

Hand-in: **Zip**-up all your **notebooks, python files and derived dataset(s)** you have chosen 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.