Lab 04 STATS 769 Due 4:00pm Monday 26 August

The main point of this lab is to **work with XML and JSON data**. We will again generate a literate report that describes a simple analysis, predicting trip duration from trip distance, but this time in addition to fitting a simple linear regression model, we will also fit a series of polynomial regression models and compare them using k-fold cross-validation.

The Data

The data are again electronic bicycle or scooter trips, but they are provided on each of the VMs in several different formats:

- A set of 10 JSON files, `trips-1.json` to `trips-10.json`, in `/course/Labs/Lab04/JSON/`. Each file contains 10,000 trips.
- A mongoDB database called `"trips"`. The database contains all 100,000 trips.
- A set of 10 XML files, `trips-1.xml` to `trips-10.xml`, in `/course/Labs/Lab04/XML/`. Each file contains 10,000 trips.

The Task

1. Read the data from all 10 JSON files and create an R data frame containing all 100,000 trips, then subset just the *trip_distance* and *trip_duration* for *scooter* trips *from 2018*.
2. Read the data from the mongoDB database, but only request the *scooter* trips *from 2018* and only the variables *trip_duration* and `trip_distance`.
3. Read the data from all 10 XML files, use XPath to extract the `trip_distance` and `trip_duration` elements that have a sibling `year` element with content `"2018"` and a sibling `vehicle_type` element with content `"scooter"`, and create an R data frame containing the distances and durations.
4. Use BaseX to query all 10 XML files at once and extract just the *distances* and *durations* for *scooter* trips *from 2018*.

   NOTE: In the `for` clause, you can specify `collection("file:///absolute/path/to/directory")` instead of `doc("filename")` to access several XML files at once.
5. Check that all four approaches generate a data set that contains the same data values.
6. Fit a series of polynomial regression models that predict the log of trip duration based on the log of trip distance, including progressively higher powers of trip distance (up to 5).
7. Use cross-validation to estimate the test error for each model; after what power does the test error level off ? Produce a plot of test errors against power of the polynomial.

The Report

Your submission should consist of a *tar ball* (as generated by the `tar` shell tool) that includes an R Markdown document *and* a Makefile *and* a processed version of your R Markdown document, submitted via Canvas.

You should write your document and your Makefile so that the markdown document can be processed just by typing `make` anywhere on one of the virtual machines provided for this course (`sc-cer00014-04.its.auckland.ac.nz` or `sc-cer00014-05.its.auckland.ac.nz`).

Your report should include:

- A description of the data format.
- A description of all four methods used to import the data to R.
- The estimates of test error for 5 polynomial models.
- A conclusion summarising the analysis.

Your report should NOT be longer than **10 pages**.

Marks will be lost for:

- More than 10 pages in the report.
- R Markdown file does not run.
- Section of the report is missing.
- R Markdown file is missing.
- Processed file (pdf or docx or html) is missing.
- Makefile is missing.
- Significantly poor R/shell code.
