**Lab 06** STATS 769 Due 4:00pm Tuesday 24 September

The main point of this lab is to **assess large data problems**. We will again generate a literate report that describes a simple linear regression analysis, relating trip duration to trip distance, but we will focus on **measuring the memory requirements** of the analysis.

# The Data

The data set for this lab is a single large CSV file containing electric vehicle trips (a larger version of what we have been using in previous labs). This file is available on the VMs at the following location:

`/course/data.austintexas.gov/Dockless_Vehicle_Trips.csv`

For this lab you should **NOT** load the entire data set into R at any point.

The code snippets shown below are also available on Canvas.

# The Task

1. Determine the size of the file `Dockless_Vehicle_Trips.csv`. How many trips does the file contain?

2. Load 1,000, 10,000, and 100,000 trips from the CSV file into R. Calculate the size of the resulting data frames in R. Explore and explain the contribution of each column to the size of the overall data frame. Estimate the size of the data frame if we loaded the entire CSV file.

   We will work with the data frame containing 100,000 trips for the remainder of the lab.

3. Check for and remove any missing values and/or very large values from the distance and duration variables. The resulting data frame should be named `tripKeep`.

4. Use the following code to remove any non-positive values from distance and duration and log both variables. Measure the **maximum memory used** by this code and explore the **largest individual objects** that are created by this code.

   ```
   trips <- subset(tripsKeep, Trip.Duration > 0 & Trip.Distance > 0)
   trips$logDuration <- log(trips$Trip.Duration)
   trips$logDistance <- log(trips$Trip.Distance)
   ```

   Repeat the process with the following code (and comment on the different memory usage).

   ```
   subset <- tripsKeep$Trip.Duration > 0 & tripsKeep$Trip.Distance > 0
   logDuration <- log(tripsKeep$Trip.Duration[subset])
   logDistance <- log(tripsKeep$Trip.Distance[subset])
   ```

5. Generate test/training group labels by running the following code.

   ```
   labels <- rep(1:10, length.out=length(logDuration))
   groups <- sample(labels)
   ```

   Use the function `mse()` (shown below) to estimate the test error for a simple linear regression model using k-fold cross-validation. What is the **maximum memory** used by R?

```
mse <- function(i, formula) {
    testSet <- groups == i
    trainSet <- groups != i
    fit <- lm(formula,
              data.frame(x=logDistance[trainSet],
              y=logDuration[trainSet]))
    pred <- predict(fit, data.frame(x=logDistance[testSet]))
    mean((pred - logDuration[testSet])^2, na.rm=TRUE)
}
```

6. Plot the simple regression line fit to all of the data (100,000 trips). **Hint:** because there are so many individual points to plot, try the `smoothScatter()` function.

7. Estimate the memory usage if we tried to estimate the test error via k-fold cross-validation on the full data set (the entire CSV file). Does the VM have sufficient RAM to perform this task? What about if an entire lab full of students (40 students) tried to fit the model at the same time on the same VM?

8. Use the shell command `time` to measure the memory required to build your lab report. (Do not run this code within your R markdown file; just show the code and the result from running it in the shell.) Does this correspond to the memory usage measured within R?

# The Report

Your submission should consist of a *tar ball* (as generated by the `tar` shell tool) that contains 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 tar ball can be extracted into a directory 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`) and the markdown document can be processed just by typing `make`.

Your report should include:

- A description of the data format.
- A discussion of the memory usage required to read the data into R.
- A discussion of the memory usage required to transform the data in R.
- A discussion of the memory usage required to estimate test error for a linear regression model.
- A plot of the linear regression model.
- A conclusion summarising the analysis.

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

Marks will be lost for:

- Submission is not a tar ball.
- 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 (or other) code.

---