Lab 08 STATS 769 Due 4:00pm Tuesday 8 October

The main point of this lab is to gain experience with **measuring execution time** and trying to **write faster code**. We will work with a simple CSV file and calculate test error for a simple linear regression model, exploring the time required for several different approaches.

The Data

The data set for this lab is a single large CSV file containing electric vehicle trips (same as for the previous lab). This file is available on the VMs at the following location:

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

The Task

1. Read the first 100,000 rows from the CSV file using three different approaches:
   - `read.csv()` without specifying `colClasses`.
   - `read.csv()` specifying `colClasses`.
   - `data.table::fread()` specifying `colClasses`.

   In each case, skip the first line of data (because it just contains missing values).

   Compare the execution time for all three approaches. Check that all three approaches produce the same set of data values.

2. Read the entire CSV file into R with `data.table::fread()` and assign the result to a symbol called `tripDT`.

3. Measure the time taken to run the following R code.

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

4. Try to perform the same subsetting and transformation *much faster* by writing R code that uses the special **data.table** syntax. **NOTE:** your code should create new variables in `tripDT` rather than creating a new `tripDTsub` object.

   Check that your new R code produces the same set of data values as the original code (for the subset of rows with positive distance and duration values).

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

```
labels <- rep(1:10, length.out=nrow(tripDT))
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.

```
mse <- function(i, formula) {
    ## cat(i, "\n")
    testSet <- groups == i
    trainSet <- groups != i
```

```
    fit <- lm(formula,
              data.frame(x=tripDT[trainSet, logDistance],
                         y=tripDT[trainSet, logDuration]))
    pred <- predict(fit, data.frame(x=tripDT$logDistance[testSet]))
    mean((pred - tripDT$logDuration[testSet])^2, na.rm=TRUE)
}
```

Measure the time taken to calculate the test error and **profile** the calculation to see where most of the time was spent.

6. Based on your profiling results, try to write new R code to calculate the test error that runs faster. **HINT:** one way to speed up the code involves the `na.action` argument to the `lm()` function.

   Check that your new R code generates approximately the same answer as the original code.

   Please seek assistance from your lecturer if you get stuck with this task - it is not trivial.

7. Use the shell command `time` to measure how long it takes to process your R Markdown document.

   You will **lose marks** if "user" plus "system" is longer than 2 minutes.

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 description of each of "The Task"s.
- 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.
- **R Markdown file takes longer than 2 minutes to run**.
- A "Task" is missing.
- R Markdown file is missing.
- Processed file (pdf or html) is missing.
- Makefile is missing.
- Significantly poor R (or other) code.