Lab 02 STATS 769 Due 4:00pm Monday 12 August

The main point of this lab is to **work in Linux**. We will again generate a literate report that describes a simple analysis, but this time we will fit and evaluate a simple logistic regression predictive model and we will use a slightly larger data set (though still in a simple format).

The Data

We will work with eleven CSV files, which are located on the STATS 769 virtual machines: `trips-2018-4.csv` to `trips-2019-2.csv`

Each of these files contains data on 5000 trips on electrics bikes or scooters in Austin, Texas, USA. Each row represents a trip and the file `trip-variables.csv` describes the variables measured for each trip (there are some new ones compared to the last lab).

The Task

1. Import the eleven CSV files into R and combine them into a single data frame.
2. Extract a subset of 1000 rows **from each month** to use as a test set (a total of 11,000 rows); the remaining 44,000 rows are the training set.
3. Transform the training set by removing trips with a distance or duration that is non-positive, then log the duration variable.
4. Create a "long trip" variable that is `TRUE` if the trip distance is greater than 1000 (1km) and `FALSE` otherwise.
5. Using the training set, fit a logistic regression model to predict the proportion of long trips based on trip duration.
6. Transform the test set in the same way as you transformed the training set.
7. Evaluate the model on the test set.

The Report

Your submission should consist of an R Markdown document, submitted via Canvas.

You should write your document so that it can be processed by running it 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`), without any manual intervention.

The CSV files are located in the directory `/course/Labs/Lab02/` on both virtual machines. You should use absolute paths to refer to these files (so your code can be run from anywhere on one of the virtual machines).

Please also submit a processed version of your R Markdown document (PDF or HTML) in case I cannot process your document myself.

Your report should include:

- A description of the data format and how the data were imported to R.
- An explanation of how you created the "long trip" variable.
- A basic exploration of the long trip variable and its relationship to the trip duration variable.
- Model fitting using a training set.
- Model evaluation using a test set.
- **A description of how you performed the analysis and generated the report on Linux** (e.g., any shell commands that you used and any notable differences in how you worked compared to how you would perform tasks on Windows).
- A conclusion summarising the analysis.

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