Lab 01 STATS 769 Due 4:00pm Monday 5 August

The point of this lab is to make sure that everyone can generate a literate report (using R Markdown) and that everyone can fit and evaluate a simple linear regression predictive model using a small data set in a simple format.

The Data

We will work with three CSV files (available on Canvas):
`trips-2018-7.csv`
`trips-2018-8.csv`
`trips-2018-9.csv`

Each of these files contains data on 5000 trips on electrics bikes or scooters in Austin, Texas, USA. Each row represents a trip, with the following variables measured:

`type`
Vehicle type (bicycle or scooter).

`duration`
Trip duration, in seconds.

`distance`
Trip distance, in meters.

`hour`
The hour of the day during which trip occurred, in local time (US/Central).

`day`
The day of the week on which the trip occurred, in local time (US/Central), where Sunday = 0, and so on.

`month`
The month # the trip occurred, in local time (US/Central), where 1 = January, etc.

`year`
The year the trip occurred.

The Task

1. Import the three 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 3000 rows); the remaining 12000 rows are the training set.
3. Using the training set, fit a linear regression model to predict trip duration based on trip distance.
4. Evaluate the model on the test set.

The Report

Your submission should consist of an R Markdown document (or similar), submitted via Canvas.

You should write your document so that I can process it on my computer without any manual intervention. For example, do not include any calls to `setwd()` or `file.choose()`.

For this lab, you should write code that assumes that the CSV files are in the current working directory.

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

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 training and test sets.
- A basic exploration of variables to be used in the data analysis (e.g., plots of distributions and plots of relationships).
- Model fitting using a training set.
- Model evaluation using a test set.
- A conclusion summarising the analysis.

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