Introduction

This document helps you, the DSI student, get set up to do DSI Assignment 1.

If you read the assignment, you may think it's a user-friendly way to introduce you to data analysis. "Look!" you might say, "I'll get to analyse my own Facebook posts! That sounds like fun!"

To that I say, you fool! What we do here is not "fun". What we do here involves blood, toil, tears and sweat. (But also stickers. Yay!)

What we do here is to remake you into the optimal data scientist, knowing a whole lot of data science best practices. That way, when you're sitting across the table from a maths PhD and the CEO of a bank, and they're peering down their noses at you wondering why it should be you who'll get the enormous piles of cash that come with being a senior data scientist, and they ask for instance "Have you ever used version control?", you can snort and say "Version control? I learned that for my very first assignment, and have been using it for two years." And you'll show them your Github account and its record of all the work you've put in that they will give you the job, and you'll go on to have many happy years getting requests for donations from the UTS Alumni Society.

Here are the ways this guide will help you become that optimal data scientist:

1) INSTALL - You'll learn to install R, RStudio, and most importantly, Github.

- R is the language we'll be working with to do our data analysis.
- RStudio is the IDE. (That is, the software that allows you to write programs in R.)
- Github is a type of cloud-based hard disk to save code, in a way that remembers all the versions of the code you've saved. It makes it much easier to share your code with your team mates without worrying about what version everyone's up to.
- Installing all of these is the part of the course that involves all that blood, toil, tears and sweat. We'll point you at tutorials that minimises that.

2) START A PROJECT - Although RStudio offers you four different ways to start doing your data analysis, you'll learn which ones to use and which ones to avoid.

- Projects are the best way to write code in R.
- We'll learn how to set up a project that syncs with Github in the cloud, and why this is a good idea.
- We'll learn why it's a bad idea to just start coding on an R or RMarkdown file.

3) WRITE AN RMARKDOWN FILE

- Once we've started a project, we'll learn how to use an RMarkdown file.
- RMarkdown is a file type that works both as a place to code AND a document that ordinary people can read.
- We'll learn how to insert pictures, add formatting like bold and italics, add a footnote, and publish it to Word, PDF or HTML.

4) GET AND CLEAN DATA

- We'll show you little snippets of code to get a data set, clean it and analyse it.
- We'll talk about the joy and wonder of doing reproducible research, and introduce you to several ways you can laugh and point at Microsoft Excel.
- We'll talk about how to think about datasets in principle, so that remembering how to do data analysis becomes like remembering a joke, rather than like memorising a dictionary.

- We'll show you how to put your code into RMarkdown, so you can choose to either show a chart, or show the code that produced the chart, or both.

5) ANALYSE UNSTRUCTURED DATA

- We'll also show you some things you can do with unstructured data, like pictures.
- You'll get an introduction to building dashboards with R, and some examples of why they're so interesting.

Interested? Read on!