**Data Science: Foundations using R**

*Who is a Data Scientist?*

A data scientist is broadly defined as someone who combines the skills of software programmer, statistician, and storyteller/artists to extract the nuggets of gold hidden under mountains of data.

< Data Science simply means using data to answer questions >

*Nuggets:*

1. The data sets we commonly encounter are much messier. It is our job to extract the information we want, corralled into something tidy like a table form, analyse it appropriately and often, visualize our results.
2. We need to reiterate data is important, but it is secondary to your question. A good data scientist asks questions first and seeks out relevant data second.
3. Admittedly, often the data available will limit, or perhaps even enable certain questions you are trying to ask. In these cases, you may have to re-frame your question or answer a related question but the data itself does not drive the question asking.

Some Definitions

Set – In statistics, it’s the population you are trying to discover something about.

Variable – Measurements or characteristics of an item.

Qualitative variable – Measurements or information about qualities

Quantitative variable – Measurements or information about quantities or numerical items

Getting help as a data scientist, common forums: \*\*Stack overflow \*\*Cross Validated

**The Data Science Process:**

1. Formulating the question
2. Finding or generating the data to be used to answer the question
3. Analysing the data: first by exploring and often by modelling the data, which means using some statistical or machine learning techniques
4. Conclusions from the analysis are then drawn and put in a presentable form

* ***Figuring out how to do what you want to do to answer your question of interest is also part of the process.***
* ***An important part of any data science project is effectively communicating the results of the project.***

What is R?

R is both a programming language and environment focused mainly on statistical analysis and graphics.

**R Studio**

* The Source panel is where you store the R commands that you want to save for later, either as a recap of what you did or it’s a way to rerun code.
* A package is a collection of functions, data and code conveniently provided in a nice and complete format for you.
* A library is the place where the packages are located in your computer.
* A repository is a central location where many developed packages are located and available for download.
* Three main repositories are CRAN; Bioconductor and GitHub. [Also: RDocumentation]

***Checking packages already installed:***

Installed.packages( ) or library( )

***Updating packages:***

old.packages( )

update.packages( )

install.packages(“packagename”)

**What is R Project?**

When you create a project, it opens a folder where all files will be kept.

**Version Control:** It is a system that records changes that are made to a file or set of files over time.

***Terminologies:***

Commit: To commit is to save your edits and the changes made.

Push: Updating the repository with your edits.

Pull: Updating your local version of the repository to the current version, since others may have edited in the meanwhile.

Staging: The act of preparing a file for a commit.

Branch: When the same file has two simultaneous copies

Merge: Independent edits of the same file are incorporated into a single, unified file

Conflict: When multiple people make changes to the same file and Git is unable to merge the edits.

Clone: Making a copy of an existing Git repository.

Fork: A personal copy of a repository that you have taken from another person.

**R Markdown**

R Markdown is a way of creating fully reproducible documents in which both text and code can be combined.