STUTERN

# The Data Science ToolBox

SGA07_DATASCI

10th December 2019

# Module Overview

- Unix Command Line
- R and RStudio
- Python and Anaconda
- Version Control
- Markdown
- Cloud-based Tools

# Book Keeping

- TA will join the class next week
- Give till the weekend for work reviews
- Can only respond to questions on Tuesday and Thursday

# Outcome

## After this Module, you will;

- Have setup your work tools as a data scientist
- Understand the technical limitation of various tools
- When to use what tool for what job in the data science process
- Learn how to organise your work and keep track of changes
- Structure technical report to accompany your work

# Unix Tools

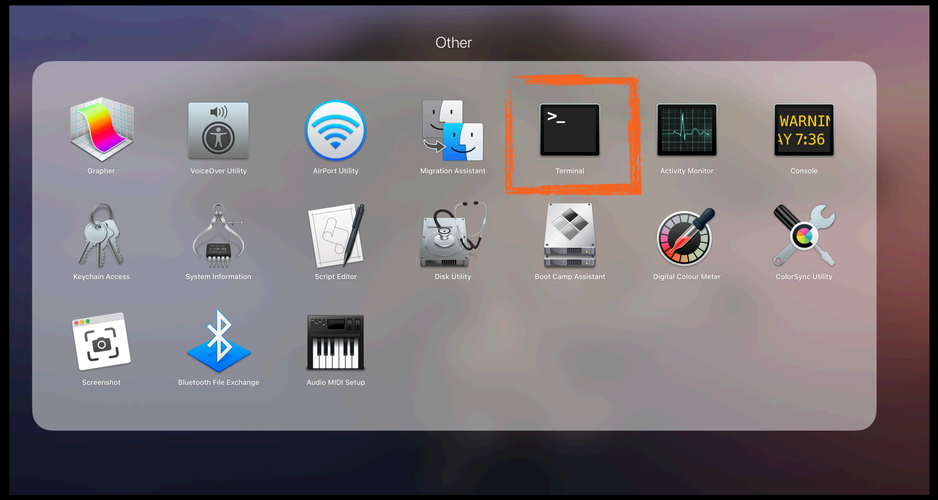
- a multitasking, multi-user computer operating system that exists in many variants
- developed in 1960s in AT&T Bells lab
- standardised through POSIX (Portable Operating System Interface based on Unix) in 1989
- uses command line tools, stdin and stdout operations to perform basic data manipulation of data files represented as lines of text

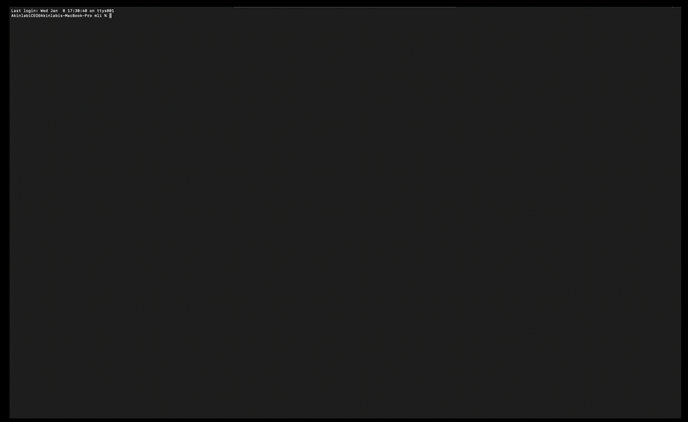
> provides an alternative way to interact with your computer to execute programs for data processing

STUTERN

# Unix Tools

- Available on Unix machines
- Available on Linux machines
- Available on modern Macs as Terminal
- Available on Windows
  - As open source tools for < Windows 10
  - Packaged as powershell in Windows 10



# Basic Unix Commands

- "man" or "-help" : to access information or manual of a command line. *e.g man sort, sort -help*
- "cd" : changes working directory. *e.g cd path/to/directory*
- "mkdir" : make/create directory. *e.g mkdir SGA07_DATASCI*
- "rmdir" : delete/remove directory. *e.g rmdir SGA07_DATASCI*

# Basic Unix Commands

- "cp", "mv", "rm" : copy, move/rename, remove files from a directory respectively. *e.g cp text.file SGA07_DATASCI*
- "ls" : list files in a directory. *e.g ls SGA07_DATASCI*
- "cat" : reads file to stdin and writes the file to stdout. *e.g cat text.file*
- "wc" : write word count of file to stdout. Generates 3 values of number of lines (-l), number of words (-w) and number of characters (-c) in text.file. *e.g wc text.file*

# R

- an open-source programming language that is focused on delivering a better and user-friendly way to do data analysis, statistics and graphical models
- developed in 1995 by Ross Ihaka and Robert Gentleman
- an implementation of S (statistical programming language) an enterprise solution developed at Bell's Laboratories
- CRAN (Comprehensive R Archive Network) is a huge repository of curated R packages - a collection of R functions and data

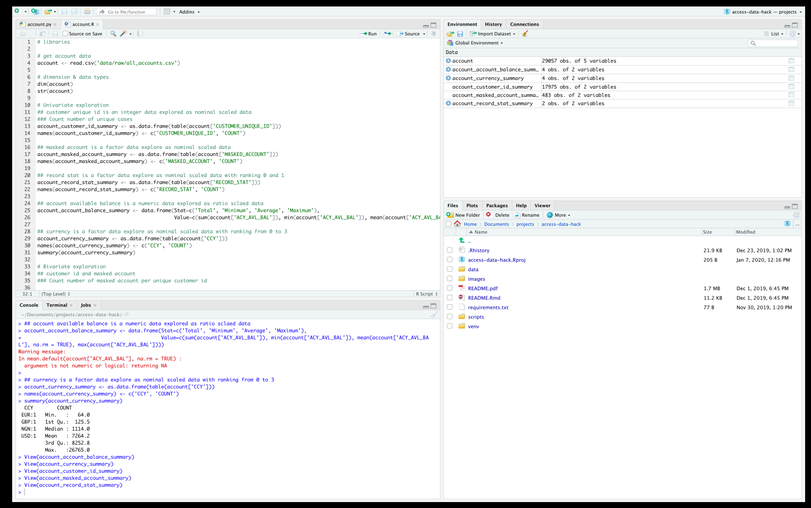
# Some R Packages

- **Dplyr** for easy data manipulation
- **Stringr** for easy text processing
- **Ggplot** for easy data visualisation
- **Caret** for easy data modelling

STUTERN

# RStudio

- an IDE (Integrated Development Environment) for R programming
- available for Linux, Mac & Windows
- provides a great way to manage your workflow as a data scientist with its intuitive view compartments

# Python

- an open-source provides a more general-purpose language with readable syntax unlike R that is built by statisticians
- more effective in deployment and large scale implementation of machine learning models
- created by Guido Van Rossem in 1991 with emphasises on productivity and code readability
- PyPi is the Python Package index and consists of libraries to which users can contribute

# Some R Packages

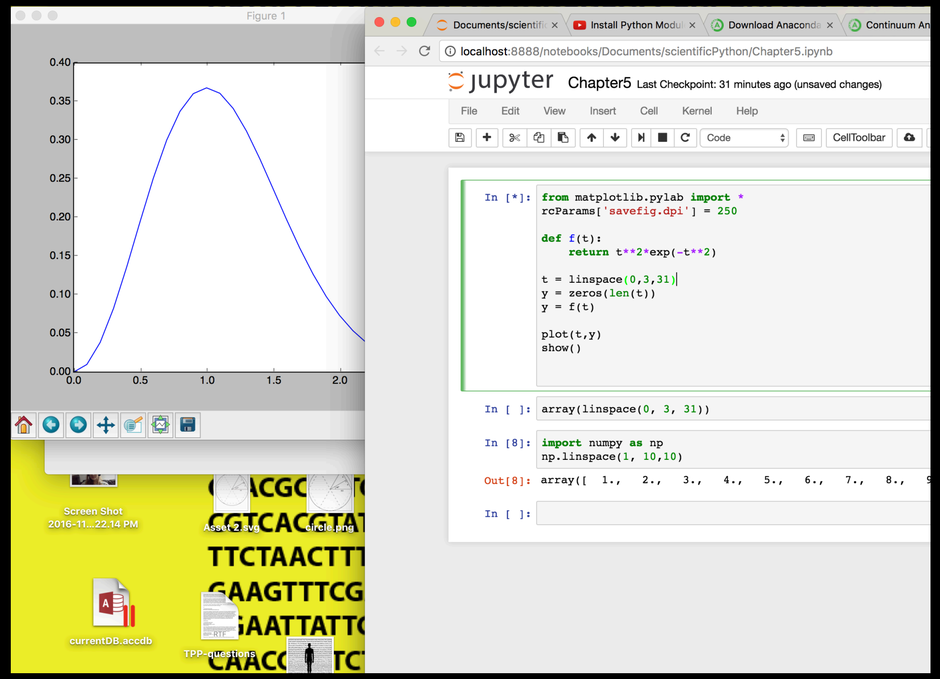
- **Numpy** for effective manipulation of N-dimension arrays
- **Pandas** for effective data structure and analysis
- **Scikit-learn** for effective machine learning
- **Seaborn** for effective data visualisation

STUTERN

# Anaconda

ANACONDA®

- a Python-based data processing and scientific computing platform
- has built in many useful third-party libraries
- serves the same way purpose of RStudio to R and even more



# Practice Lab I

Create your course work directory and project file using Unix & R/Python.

Use the following Instructions:

- Make sure you have unix on your computer
- Create your root directory names *"SGA07_DATASCI"*
- Install R and RStudio Or Python and Anaconda
- Start in a new project in your root directory
- Recreate your data profile and save in your work

# Version Control

- a distributed version control system that allows user to keep entire source code repository and history on their local machine
- was created in 2005 by Linus Torvald to aid the Linux kernel development.

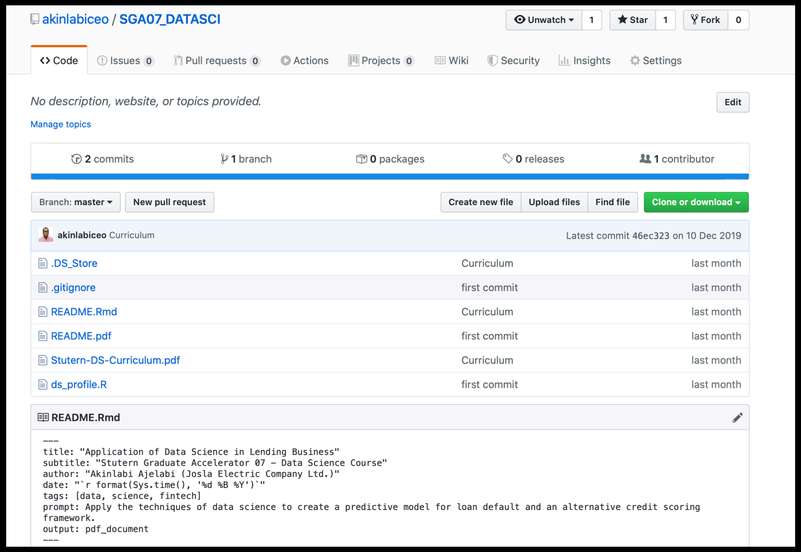
# Key Git Concepts

- **Repository** - this is Git's name for a project
- **Snapshot** - this is the way Git keeps track of your code history
- **Commit** - In Git, history is made up of a series of commits which are stored in the change-log
- **Staging Area** - This is like a shopping basket for version control

STUTERN

# Github

- a web-based git repository hosting service
- allows for code collaboration and storage online as well as adds extra functionality on top of git such as user interface (UI), documentation, bug tracking, feature request, push and pull requests and more!

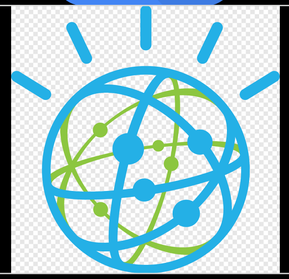


# Markdown

- a lightweight markup language that you can use to add formatting elements to plaintext text documents
- was developed by John Gruber in 2004 to allows web writers to write using an easy-to-read, easy-to-write plain text format, then convert it to structurally valid XHTML
- use it document your methodology, approach and result as well as use it for technical documentation

STUTERN

# Specialised Cloud-based Tools

| | | |
|---|---|---|
|  | Microsoft Power BI | a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. |
|  | Tableau | a powerful and fastest growing data visualisation tool used in the Business Intelligence Industry. It helps in simplifying raw data into the very easily understandable format. |
|  | Google Data Studio | a dashboard and reporting tool that is easy to use, customise, and share. It allows you to transform your data into appealing and informative reports for your audience |
|  | IBM Watson Studio | an integrated environment designed to make it easy to develop, train, manage models, and deploy AI-powered applications and is a SaaS solution delivered on the IBM Cloud |

# Practice Lab 2

Setup your production environment using version control and markdown

Use the following Instructions:

- Create a Github account
- Create your root directory into a version control directory
- Push local directory to Github directory
- Create a ReadMe file using markdown (Add your data science profile)
- Push again to Github directory and share repo link

# Recap/Summary

**At the end of this Module, you should understand;**

- Introduce a broad collection of tools used in data analytics
- Outline the capabilities and uses of each tool
- Provide examples of tool usage
- Allow you to select the appropriate tools to work with
- Based on your preferences, e.g. GUI or command line
- Very quick introduction to each tool
- More information on the web

# Suggested Material

- https://www.cl.cam.ac.uk/teaching/1213/UnixTools/materials.html
- https://www.datascienceatthecommandline.com/index.html
- https://www.howtogeek.com/249966/how-to-install-and-use-the-linux-bash-shell-on-windows-10/
- https://cran.r-project.org
- https://rstudio.com/products/rstudio/
- https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/
- https://r4ds.had.co.nz/introduction.html
- https://www.python.org
- https://www.anaconda.com

# Suggested Material

- https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/
- https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf
- https://www.youtube.com/watch?v=SWYqp7iY_Tc
- https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners
- https://help.github.com/en/github/using-git/getting-started-with-git-and-github
- https://www.markdownguide.org/getting-started/
- https://daringfireball.net/projects/markdown/syntax

# Suggested Material

- https://powerbi.microsoft.com/en-us/
- https://www.tableau.com
- https://datastudio.google.com
- https://www.sas.com/en_us/solutions/business-intelligence.html
- https://www.ibm.com/cloud/watson-studio
- https://aws.amazon.com/machine-learning/