Intro to Data Analytics and Visualizations

Simple Problem

John is having some problems in school and his parents are wondering what is going on and how to fix it? When asking John, he says he is doing everything right. How can we investigate this problem and come up with solutions using data?

What is Data Science?

1. Statistics
2. Machine Learning
3. Software and data storage and access
4. Project Management
5. Field knowledge
6. Terminology and relationship with Data Analytics and Visualizations

Examples

1. Netflix
2. Amazon

Roles in the Data Science Process

1. Project Sponsor
2. Client
3. **Data Scientist** (other names for this roles are possible)
4. Data Architect
5. Operations

TOOLS

What are the tools we will use?

1. Methods
2. Datasets
3. Software:
   - 3.1 R and R Studio
   - 3.2 GitHub for version control
   - 3.3 Python (limited)

Introduction to R

Outline

1. What is R
2. Why use R
3. Installing R in your own computer
4. R studio
5. Your first script

What is R?

R is a powerful, versatile, and free statistical programming language. Scientists, statisticians, analysts, students and others who are interested in statistical analysis, data visualization, etc. are using R to do so.

Data analysis is done in R by writing or using built in scripts and functions in the R language. The R environment is not only equipped with all the standard methods, but also some of the most recent cutting-edge techniques.

R is open source. This means that you can download and use R for free, and additionally the source code is open and available for inspection and modification.

Why use R?

* R is free and open.

* R is a language. You learn much more than just point and click.

* R has excellent tools for graphics and data visualization.

* R is flexible. You are not restricted to the built in set of functions, you can use them and extend them with your own.

You can make your analysis your own!

How to Obtain R for your own computer?

Windows:
http://cran.r-project.org/bin/windows/base/

MacOs X:
http://cran.r-project.org/bin/macosx/

Version Control with GitHub

- Why use Git?
- How to install Git?
- Go to https://github.com/ and create an account.
- Go to http://git-scm.com/downloads and click the link specific for your system.
- Download a free GUI specific for your system http://git-scm.com/downloads/guis .
- Follow through with installation and setup instructions.
- Note: Book with instructions: http://git-scm.com/book

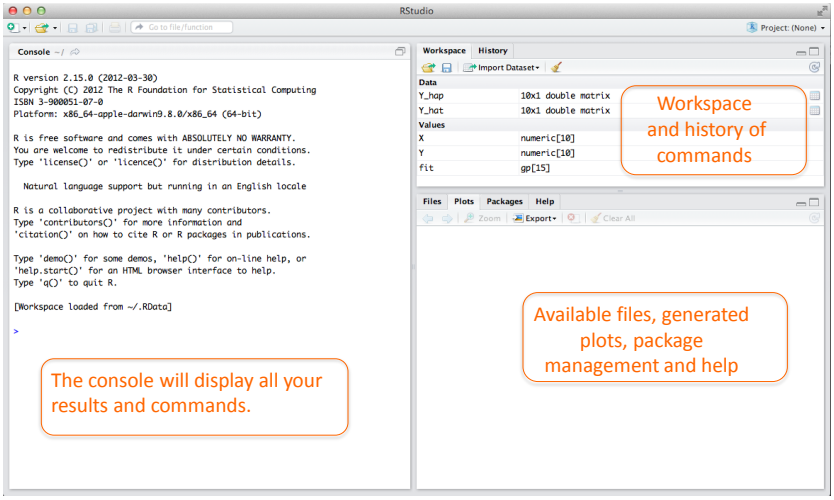
How to Obtain R Studio for your own computer? (after you got R and Git)

http://www.rstudio.com/products/rstudio/download/

We are installing RStudio Desktop.

Then choose the installer for your own system (Windows/Mac) and click on it. Follow through with the installation instructions.

R Studio



R Studio menu

- Open files.
- New files.
- Save files.
- Run scripts (Ctrl+R in Windows).
- ”New Folder”. Create a folder “Rfolder” on your computer, where you will save all your R files.

First script

- Set working directory to the R folder you created
- Create “Hello World” object
- Create comments with “#”
- Get to help from menu or type help.start()
- Save R script and workspace to your R folder

Version Control with GitHub

- Why use Git?
- How to install Git?