Data Science Toolbox

Introduction to Git (Version Control)

* Everything is stored in local repositories
* GitBash is a command line environment
* $ is a prompt for a command
* Git – local, on your computer
* Github – remove, on the web
* Create a repo
  + ReadMe – will make repo easy to understand
  + Create a local copy in GitBash
  + Create directory – mkdir/~test-repo
  + Navigate to directory – cd~/test-repo
  + Initialize local Git repository – git init
  + Link remote and local
* Branches
  + To edit a different version from others who are also editing
  + Creates another version of the same directory
* Pull requests
  + Unique feature of github
  + Pick a branch in github
  + Prepare pull request
  + Then the owner decides if they want to merge your changes
* Markdown
  + Text file with a particular format
  + Headings: ## secondary heading ### tertiary heading
  + Lists: \* first item \* second item
* Types of Data science questions
  + Descriptive – describe data, like census data – cannot be generalized without statistical analysis
  + Exploratory – find relationships, but not confirm them – good for future data science projects
  + Inferential analysis – small amount of data – extrapolate – goal of most statistics – estimation and error
  + Predictive analysis – use the data on some objects to predict values
  + Casual analysis – generally use randomized control trials – gold standard
  + Mechanistic analysis – understand exact changes in variables that lead to exact changes in other variables – the only random component is measurement error
* What is data?
  + Qualitative or quantitative variables, belonging to a set of items
  + The most important thing is the question
* Experimental Design
  + Formulate your question in advance
  + Confounding – another variable that may have affected the relationship
  + Prediction quantities

|  |  |  |
| --- | --- | --- |
| Test | Disease | |
| + | - |
| + | True positive | False positive |
| - | False negative | True negative |

* + Sensitivity = P( + test | disease)
  + Specificty= P( - test | no disease)
  + Positive predictive value = P(disease | + test)
  + Negative predictive value = P(no disease | - test)
  + Accuracy = P(correct outcome)

Git Commands

|  |  |
| --- | --- |
| Command | Operation |
| pwd | Print working director |
| ls | Lists files in current directory |
| mkdir | Make directory |
| cp *File to move Directory to move to* | Copy |
| rm | Remove (Like delete, no undo) |
| mv *File to move Renamed file* | Move |
| touch | New file |
| echo | Prints out whatever you provide |
| date | Today’s date |
| clear | Clears command window |
| git add. | Adds new files |
| git add | Updates tracking for updated files |
| git add – A | Does both |
| git commit | Updates local repo, not github |
| git push | Pushes local commits to github |
| git checkout – b *branch name* | Creates another version |
| git branch | Tells you what branch you are on |
| git checkout master | Go back to master branch |
|  |  |

Setting up the github repository and linking the local copy

cd documents/coursera

cd ‘data science’

git init

git remote add origin <https://github.com/tzvia/datasciencecoursera.git>

git pull origin master