Coursera/Johns Hopkins Data Science Specialization

Course 1

The Data Scientist’s Tool Box

Week 1

Video 2: The Data Scientist’s Toolbox

* R is super important🡪Will be set up in the second week of the class
* Most coding will be done in RStudio
* Need to set up a Git/Github account

Video 3: Getting Help

* Use the message boards, either a TA or the prof will respond
* Some important R functions for help: ?rnorm, help.search(“rnorm”), args(“rnorm”), function🡪will produce all available information on the function

Video 4: Finding Answers

* Gotta be willing to go out and find answers on my own. Be clever!
* Start in the forums, then try google, stack overflow [r]

Video 5: R Overview

* Reading Lines of a Text File
* How to figure out when something is wrong with a function, why, and how do we fix it?
* More detailed functions such as lapply

Video 6: Getting and Cleaning Data Overview

* Raw vs clean or processed data, how to connect to rmysql and stuff like that

Video 7: Exploratory Analysis Overview

* Principles of analytic graphics, lots of plots/graphs
* K-means clustering?

Video 8: Reproducible Research Overview

* Create code and docs that will completely reproduce work done in a transparent way so that we can share our work with others easily
* Steps in a data analysis from asking the question🡪creating reproducible code
* The different files required

Video 9: Statistical Inference Overview

* Things I know!!

Video 10: Regression Models Overview

* More things I know!

Video 11: Practical Machine Learning

* Hands on application of machine learning in R

Video 12: Building Data Products

* What do we do with the data or functions once we’ve built them?
* Building an R package is the best way to get my name out there
* Build interactive graphics
* Shiny

Video 13: Installing R on Windows

* Already done but maybe I should update

Week 2

Video 1: Command Line Interface

Video 2: Introduction to Git

Video 3: Introduction to Github

Video 4: Creating a Github Repository

Video 5: Basic Git Commands

Video 6: Basic Markdown

Video 7: Installing R Packages

Video 8: Installing Rtools

Week 3

Video One: Types of Data Science Questions

* Descriptive

Just trying to describe date and interpret what we see

Cannot be generalized beyond the group sampled

Example—US Census

Not trying to make decisions, just descriptions

* Exploratory analysis

Discover new connections

Useful to define future studies

Cannot be used for generalization or prediction

Correlation does not imply causation

Example—brain study

* Inferential analysis

Goal—to use a small sample of data to say something about a larger population

Example—Effect of Air population control on life expectancy in the US

* Predictive Analysis

Goal—To use the data on some objects to predict values for another object

X predicts Y, but does not cause it.

Example—FiveThirtyEight.com

* Causal analysis

Goal---To find out what happens to one variable when you make another variable change.

“gold standard” for data analysis

Example—Fecal treatment helps people recover better

* Mechanistic analysis

To understand the exact changes in variables that lead to the exact changes in other variables for individual objects.

Usually, the only random part is measurement error.

Engineering and physics applications

Video 2: What is data?

* Data are values of qualitative or quantitative variables, belonging to a set of items.
* Variables are measurements or characteristics of an item.
* Qualitative—country of origin, sex… not ordered or numbered
* Quantitative—temperature, height….ordered/numbered
* Data could be a text file, and API, a video file, audio file, spread sheets, excel files
* The most important thing in data science is the question. The second is the data that relates to the question.
* Having data can’t save you if you don’t have a question you want to answer

Video 3: What about big data?

* How much data is there?
* Regardless of the size of the data, you need the right data.

Video 4: Experimental Design

* Why should you care?
* Know and care about the analysis plan
* Formulate your question in advance
* Confounding variables are a problem
* Randomization and blocking
* Sensitivity 🡪Pr(positive | disease)
* Specificity 🡪Pr(negative| no disease)
* Positive Predictive🡪Pr(disease | positive)
* Negative Predictive🡪Pr(no disease | negative)
* Accuracy🡪Pr(correct outcome)

