Data Mining for Business Harvard CSCI E-96

Ted Kwartler

https://www.linkedin.com/in/edwardkwartler/

ehk116@gmail.com

Agenda

Start	End	Item
		Introductions
		Syllabus Review
		Break
		Intro to R
		R Installation, environment & git
		Simple R Scripting

Introduction

Edward (Ted) Kwartler

https://www.linkedin.com/in/edwardkwartler/

Education

Work

Side Hustles

HARVARD UNIVERSITY

Introductions

HARVARD UNIVERSITY

Name? Where are you located?

On Piazza…

- Where are you from?
- Profession
- Why did you take this course?
- What do you hope to learn by the end of the course?
- Tell us something interesting about yourself. Hobby? Family? Etc.

This helps Instructor and TA get to know you along with what may be of interest for your learning style.

Agenda

Start	End	Item
		Introductions
		Syllabus Review
		Break
		Intro to R
		R Installation, environment & git
		Simple R Scripting

Syllabus Review

Course Syllabus: Data Mining for Business

CSCI E-96 – *although this course includes business topics at its heart it is a computer science course. Please be mindful of this.*

Harvard Extension Spring 2020
Dates: January 27, 2020 – May 16, 2020
Time: Monday 8-10pm
Building: Harvard Hall 101
Instructor: Ted Kwartler, MBA
Email:
ehk116@gmail.com
edwardkwartler@fas.harvard.edu
Office Hrs: Available upon request
Optional Lab: Fridays 10AM EST (most weeks); https://zoom.us/j/8591855557

Important URLs:

Piazza (class discussion board, post LOTS of questions!)
https://piazza.com/class/k5cooiqzd9h53p

Max	Min	Grade
100	90	A
89.9	87	B+
86.9	80	B
79.9	77	C+
76.9	70	C
69.9	67	D+
66.9	60	D
59.9	0	F

Community Learning Expectations

I expect:

- Kindness > Cleverness
- Active learning & participation
- To make adjustments along the way so learning will be engaging & fun
- Academic rigor

I do **not** expect:

- Expert programmers or production quality coding skills – *not really a programming course*
- Significant math skills – *not a heavy math course*

My goal for your learning experience is that the topics are:

- **Interesting**
- **Applicable** as a professional or future business person
- **Perspective building** as a consumer affected by these methods

Community Learning Expectations

The class expects (of the instructor & cohort):

- …
- …
- …

The class does not expect:

- …
- …
- …

The class hopes learning this material will be:

- …
- …
- …

Although this is rhetorical, reflect on what you can contribute to make this the BEST course you have ever taken.

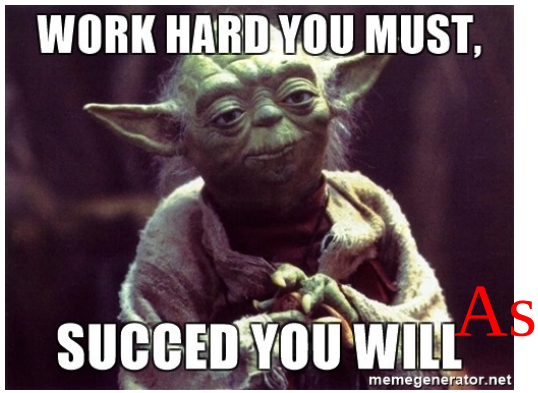
Some motivation

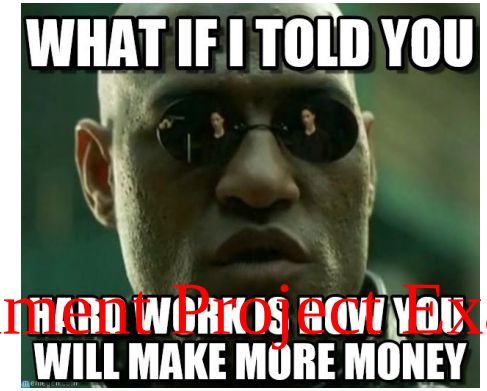
- Focus on your learning outcomes
- Don't focus on your grade
- You own the level of effort & engagement in your learning

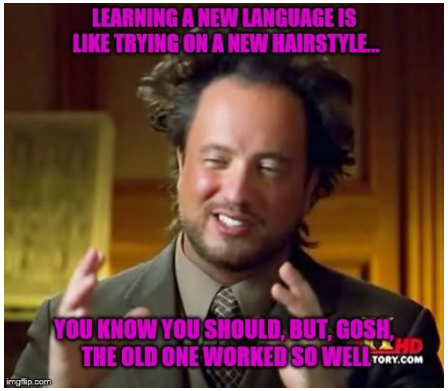
- I am focused on your learning outcomes
- I don't care if you need a specific grade to graduate
- I take my role seriously mirroring your effort & engagement to support learning

If you are struggling, ask Prof help, attend labs, request office hours, email or even smoke signals, whatever it takes to succeed.

Your professor loves memes.







Agenda

Start	End	Item
		Introductions
		Syllabus Review
		Break
		Intro to R
		R Installation, environment & git
		Simple R Scripting

Intro to R

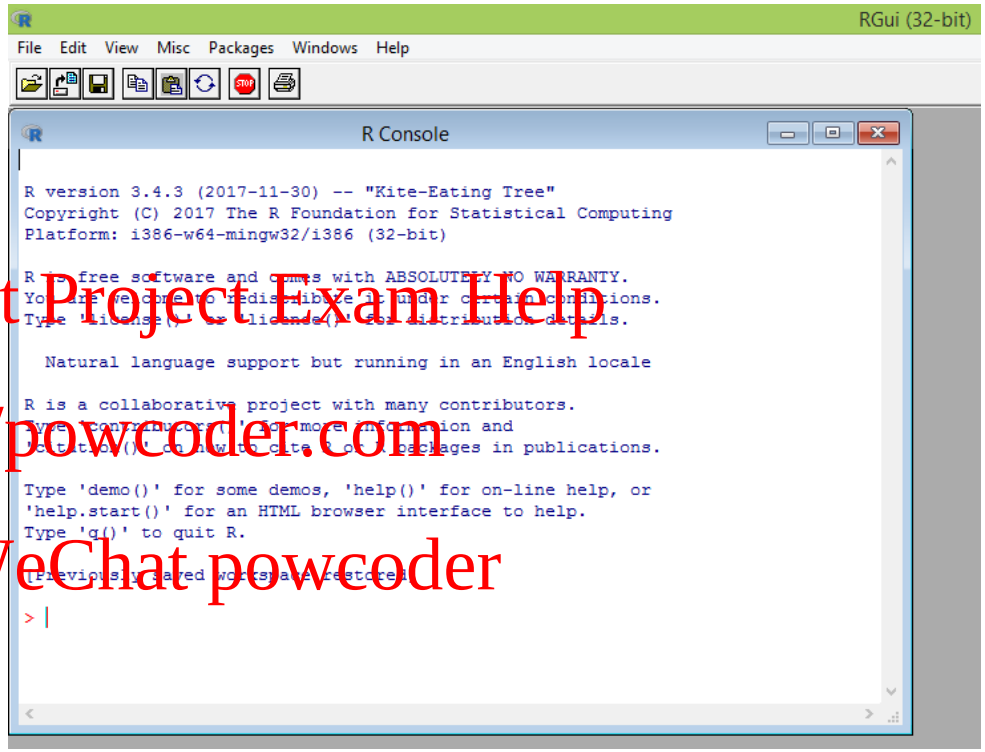
Learning Objectives

- What is R?
- What is R Studio?
- Scripting Structure in this course

Why? This ensures a basic fluency before the course really starts.

What is R?

- ✓ Flexible
- ✓ Open source
- ✓ Academic & growing in industry
- ✓ Language agnostic, SQL, Weka, C, Fortran, Java etc



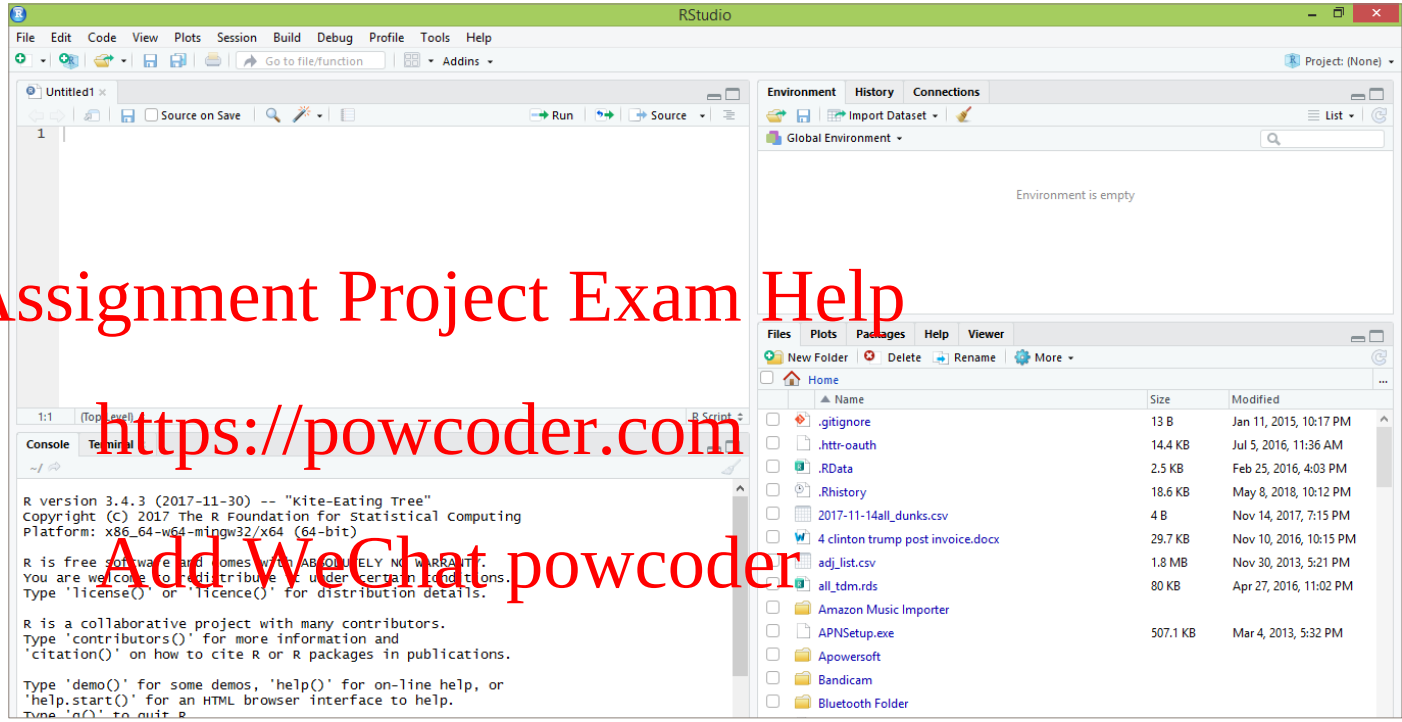
R is a language & environment for statistical computing and graphics. It is the **most popular statistical software** in circulation today and is used by more than **2 million** data scientists & statisticians worldwide.

How Companies Use R to Compete in a Data-Driven World, Data-informed.com

HARVARD UNIVERSITY

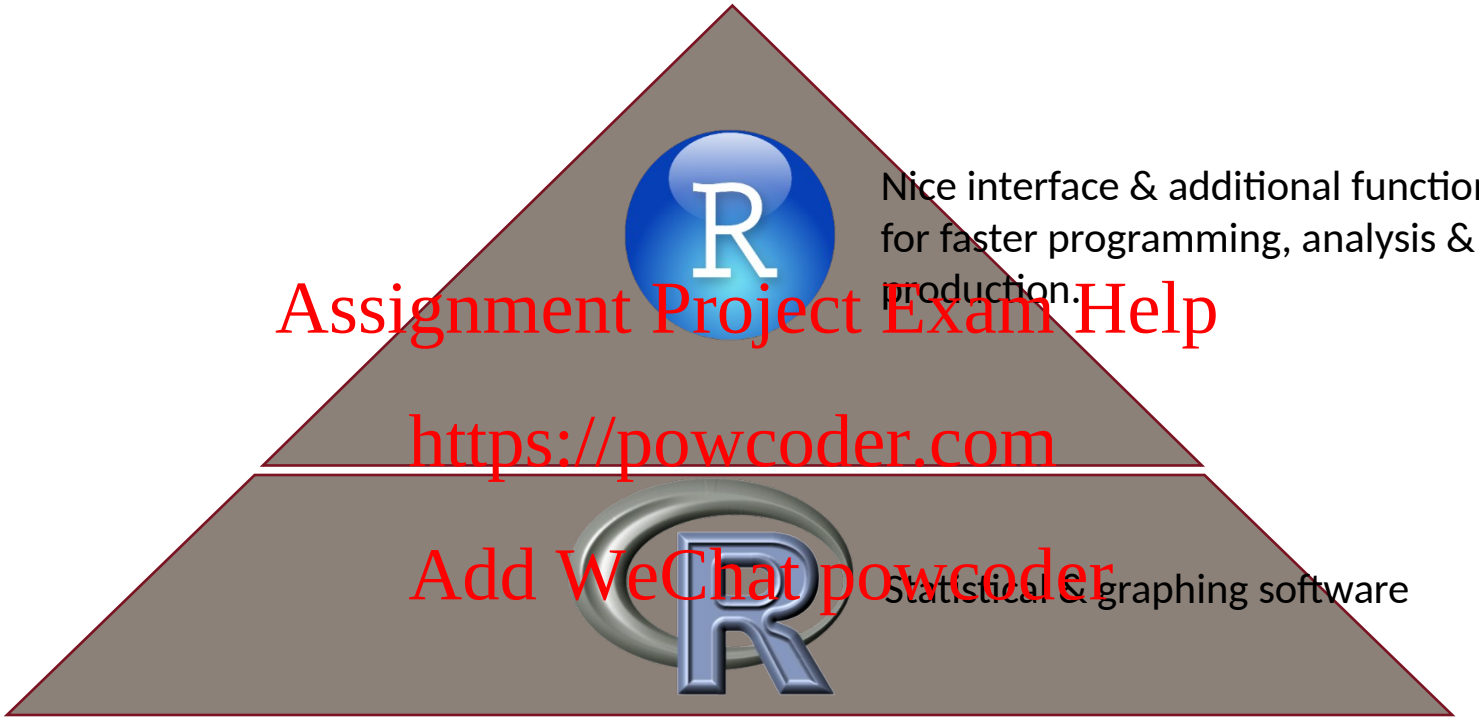
What is R Studio?

- ✓ IDE – Integrated Development Environment
- ✓ Adds additional functionality e.g. git, shiny projects, markdown templates



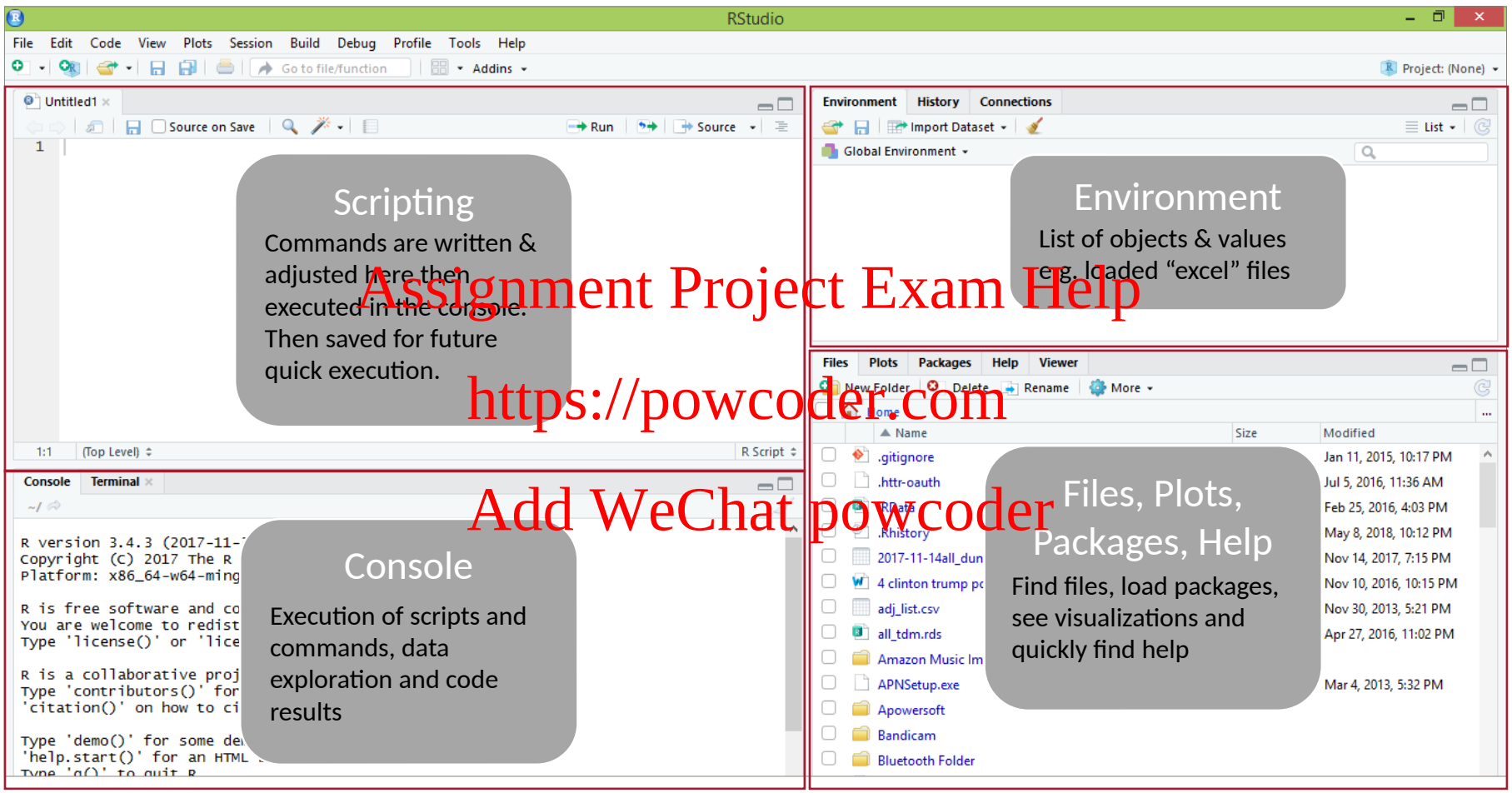
R studio is the most popular IDE for R although there are others & you don't actually need it to execute R code.

What is the relationship between R & R studio?



R Studio sits atop of the installed R version. Without base R, R studio cannot function. By programmatically accessing base R, R Studio improves the interface and functionality.

R Studio has four main panes



Scripting

Commands are written & adjusted here then executed in the console. Then saved for future quick execution.

Environment

List of objects & values
eg. loaded "excel" files

Console

Execution of scripts and commands, data exploration and code results

Files, Plots, Packages, Help

Find files, load packages, see visualizations and quickly find help

R Studio works on Linux, Windows and iOS.

R: Where is it being used?

Media
Google
Facebook
Twitter
Foursquare
Kickstarter
New York Times
Economist

Services
Zillow
Trulia
eHarmony
DataSong
PredictWise
Nationwide

Finance
Lloyd's Bank
Credit Suisse
American Century
Australia and New Zealand Banking Group

Technology
SAS
Oracle
IBM
Teradata
Coursera
SAP
DataCamp
Alteryx
TIBCO
OneTick
Amazon
Google
Microsoft

U.S Government
Food & Drug Administration

National Weather Service

National Institute of Standards in Technology

R is ubiquitous in various industries used for analysis, prototyping and visualization. However, more often python is used in *production* environments.

Ok, it's awesome but…

Pros

- VERY extensible
  - 10K+ packages
- Built by stats, made for stats
- Free, open source
  - Cutting edge
  - Diverse applications
- Outstanding graphical capabilities
- Large community-based help

Cons

- Tough to do "big data"
  - In memory data constraints w/o extra effort
- Official documentation is terse
- Not as polished as a commercial application
- Slow compared to lower level languages
- Production worthy apps can be *difficult* to create
- Large community-based help

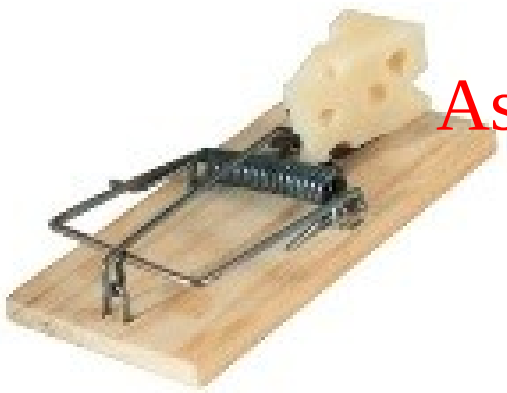
A basic R workflow

Let's eat a banana for breakfast. **Where is the fruit?**

```
setwd('ted/fruit/basket')
```


The item of interest needs to be in the "working directory"

R Trap!



WINDOWS (local laptop):

- In `setwd()` the slashes have to be switched (or "escaped")!


```
setwd("~/desktop")
```

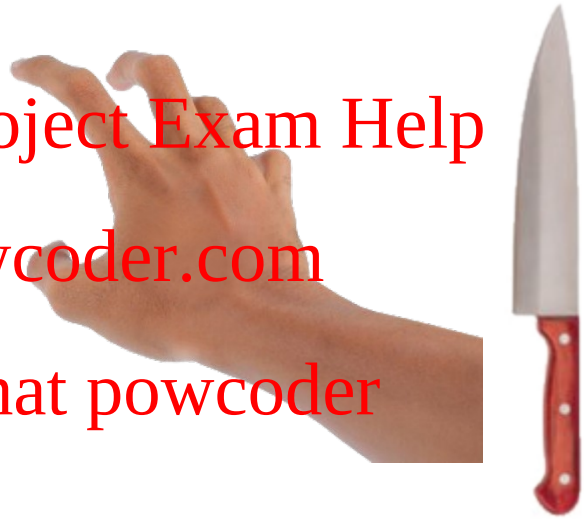


```
> setwd("~\desktop")
Error: '\d' is an unrecognized escape in character string starting ""~\d
```

R uses functions, libraries & objects

Found the fruit basket! **What tools do I need?**

```
setwd("ted/fruit/basket")
library(knife)
library(peel)
```

Change R into a breakfast preparing machine with specialized libraries.

Another trap!

Before loading a library use

```
install.packages("name of package").
```

You only need to do this once per environment

But, will need to repeat it w/each new environment!

The course readme has the packages to install.

```
# Individually you can use
# install.packages('packageName') such as below:
install.packages('ggplot2')

# or
install.packages('pacman')
pacman::p_load(ggplot2, ggthemes, rbokeh, maps,
               ggmap, leaflet, radiant.data, DataExplorer,
               vtreat, dplyr, MLmetrics, pROC,
               MLmetrics, caret, e1071, plyr,
               rpart.plot, randomForest, forecast, dygraphs,
               lubridate, jsonlite, tseries, ggcorrplot,
               arules,fst, recommenderlab,reshape2,
               TTR,quantmod, htmltools,
               PerformanceAnalytics,rpart, data.table,
               pbapply, rbokeh, stringi, tm, qdap,
               dendextend, wordcloud, RColorBrewer,
               tidytext, radarchart, openNLP, xml2, stringr,
               devtools, flexdashboard, rmarkdown, httr)

# Additionally we will need this package from a different repo
install.packages('openNLPmodels.en',
                 repos= 'http://datacube.wu.ac.at/')
```


Don’t install them now, it takes awhile.

R uses functions, libraries & objects

Now R is a cutting & peeling machine, let's pick our fruit.

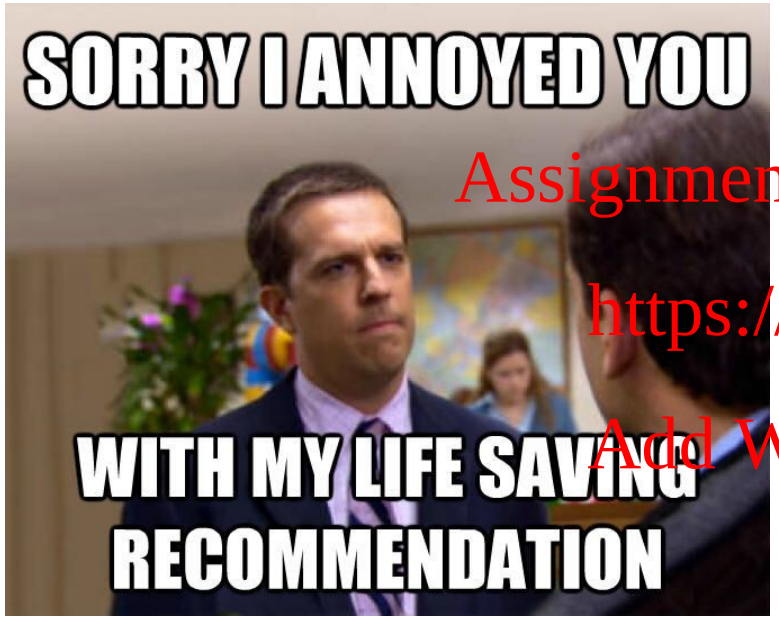
```
setwd("ted/fruit/basket")

library(knife)
library(peel)

banana <- read.fruit("Banana.csv")
```

Must

- Be In the working directory or full path declared
- Be inside quotes
- Have correct capitalization, spacing & spelling matter.

R now has the object called **banana** in memory ie "opening a spreadsheet"

What if my file is Excel or ???

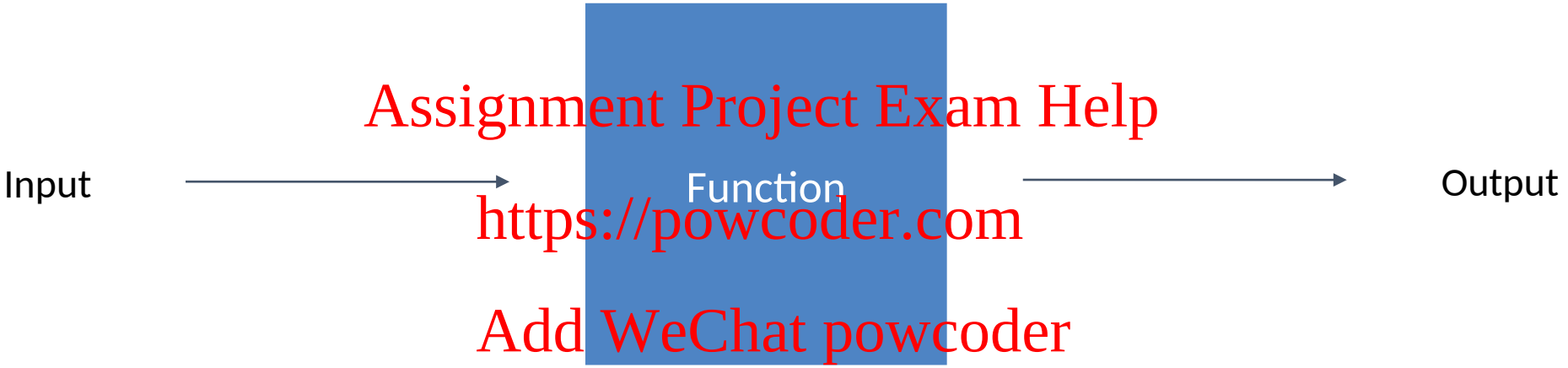


R can open many files types with different functions

- CSV
- Excel
- SaS
- SPSS
- Rds, rda, fst etc
- Connections
  - Databases (SQL, mongoDB)
  - S3 Buckets
  - APIs

Recommendation: to get started stick with CSV. Most software can export CSV. Most class files are CSV.

R uses functions, libraries & objects

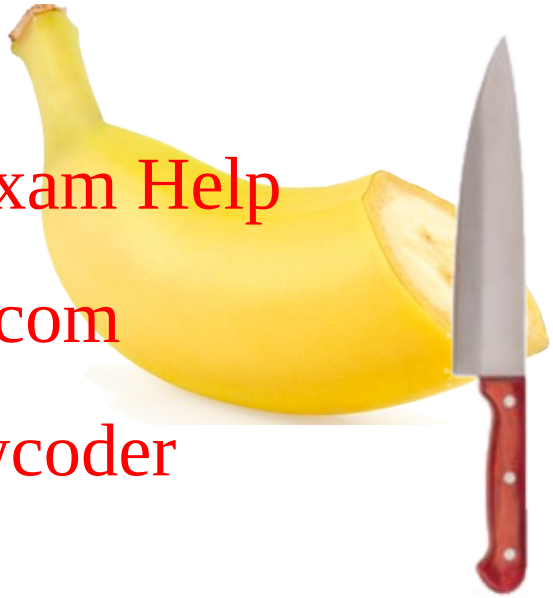


R uses predefined functions to accomplish things.

R uses functions, libraries & objects


```
halfBanana <- cutinhalf(banana )
```


Output | Function (Input)



“ <- ” is the assignment operator

R uses functions, libraries & objects

```
setwd("ted/fruit/basket")
library(knife)
library(peel)
banana <- read.fruit("Banana.csv")
halfBanana <- cutInHalf(banana)
halfBanana <- peel(halfBanana)
```

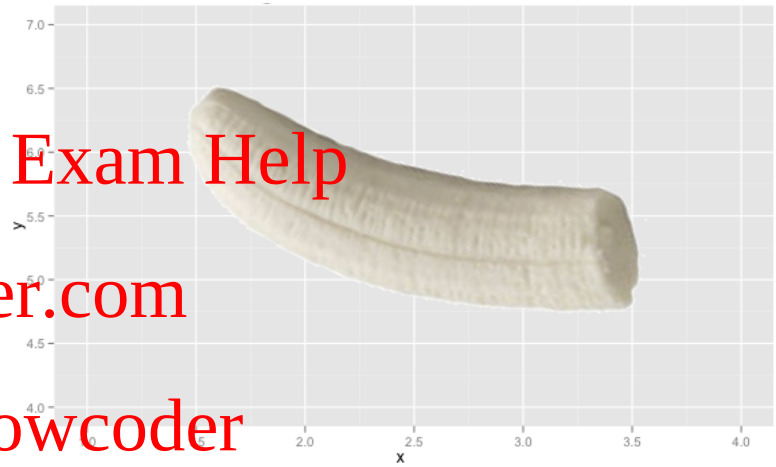
Output | Function (Input)

Results of an object can be used in the next function.

R uses functions, libraries & objects

```
setwd("ted/fruit/basket")
library(knife)
library(peel)
banana     <- read.fruit("Banana.csv
halfBanana <- cutinhalf(banana)
halfBanana <- peel(halfBanana)

plot(halfBanana)
```

Function (Input)



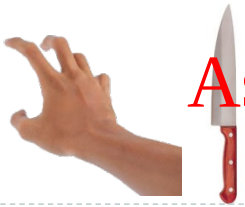
You don't always have to declare an output object & you can also save items to disk.

Workflow review

1

- Use `setwd()` to point to your files & where to save outputs.

2



- Load some customized libraries with `library()` for your specific analysis and methodology. *Sometimes we will also create customized functions not in a library to aid our analysis.*

3

- Read in the file so the object is "in-memory" with `read.csv()` or similar.

4

- Apply a function(s) from a library to adjust or create new objects in memory. The pseudo code for this is:

```
object<-function(applied to object)
```

5

- Consume the results by saving, plotting etc.

Agenda

Start	End	Item
		Introductions
		Syllabus Review
		Break
		Intro to R
		R Installation, environment & git
		Simple R Scripting

Install R Studio & Git

Prerequisites:

- Textbook: Data Mining for Business Analytics: Concepts, Techniques, and Applications in R

ISBN-10: 1118879368

Harvard Coop Bookstore link for the book: https://tinyurl.com/300-CSCI-E-96-F18-1

- Software: R & R-Studio
- Access to git software to download data sets and class material or ability to download directly from the Internet
- A webcam or other method to record case presentations & upload to the University's approved site
- Be prepared to obtain a free zoom account as each group will need a single zoom participant to record case presentations

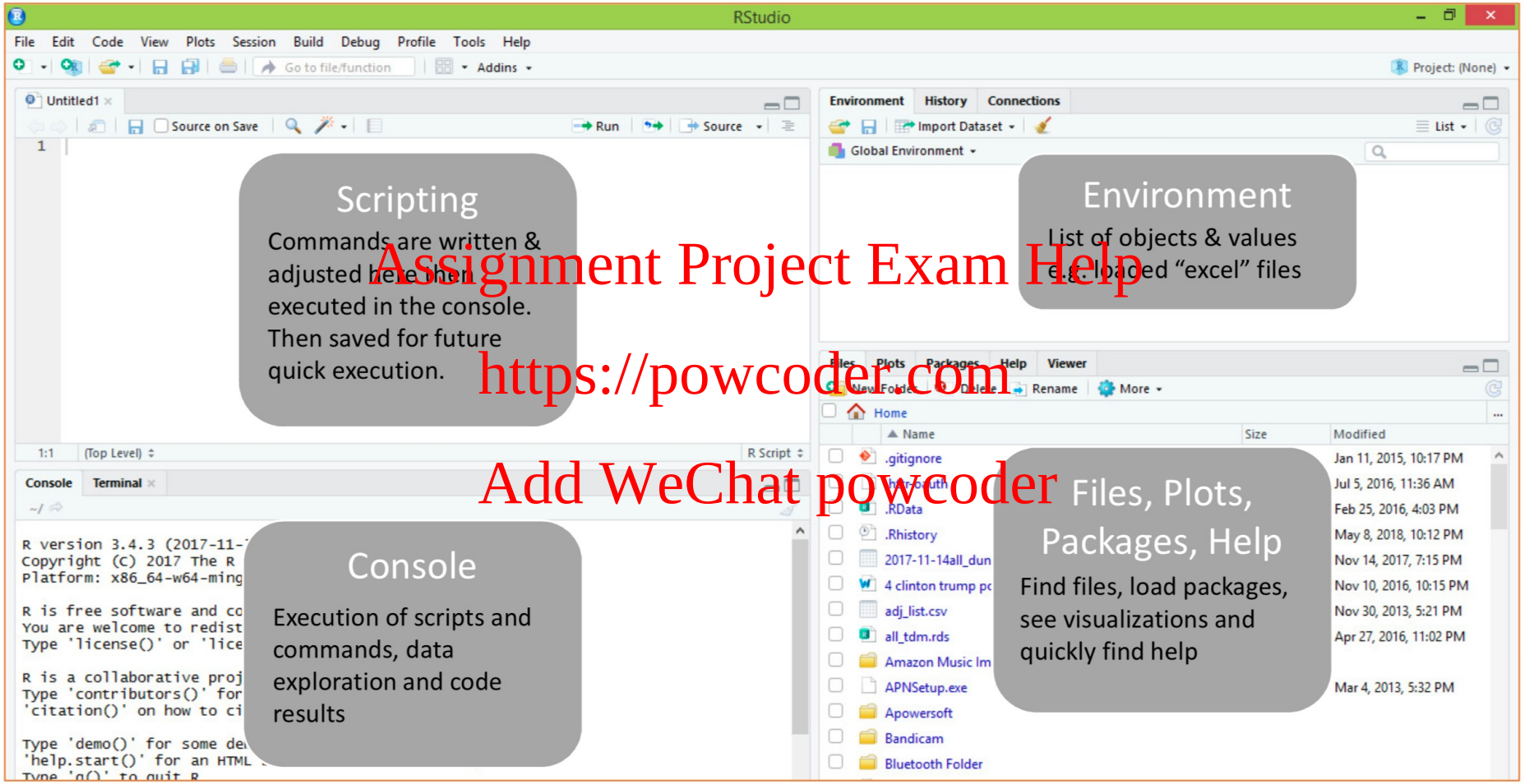
If you don't have it installed yet, please do so now!! The rest of the course and your learning utilizes R Studio.

R Studio Cloud

- Also, create an account for RStudio Cloud: https://rstudio.cloud .
  - Once you're logged in, use (click) "Your Workspace" or create a new space.
  - Click the "New Project" button.
  - Change the Title of the Project as you wish.
  - In Console, type *install.packages("rmarkdown")* and press the Enter button

- Uploading and Downloading Files in RStudio Cloud
  - See https://support.rstudio.com/hc/en-us/articles/200713893-Uploading-and-Downloading-Files

R Studio



Libraries to install…we cover a lot in this course!

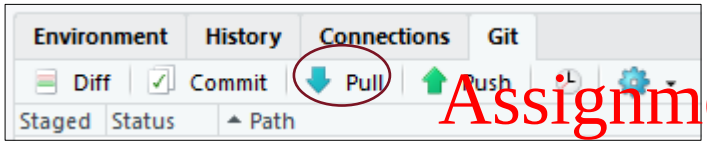
Please review the readme file of the Student repo!!

```
# Individually you can use
# install.packages('packageName') such as below:
install.packages('ggplot2')

# Or
install.packages('pacman')
pacman::p_load(ggplot2, ggthemes, rbokeh, maps,
               ggmap, leaflet, radiant.data, DataExplorer,
               dygraphs, plotly, modelmetrics, pROC,
               MLmetrics, caret, e1071, plyr,
               rpart.plot, randomForest, forecast, dygraphs,
               lubridate, jsonlite, tseries, ggseas,
               arules,fst, recommenderlab,reshape2,
               TTR,quantmod, htmltools,
               PerformanceAnalytics,rpart, data.table,
               pbapply, rbokeh, stringi, tm, qdap,
               dendextend, wordcloud, RColorBrewer,
               tidytext, radarchart, openNLP, xml2, stringr,
               devtools, flexdashboard, rmarkdown, httr)

# Additionally we will need this package from a different repo
install.packages('openNLPmodels.en',
                 repo= 'http://datacube.wu.ac.at/')
```

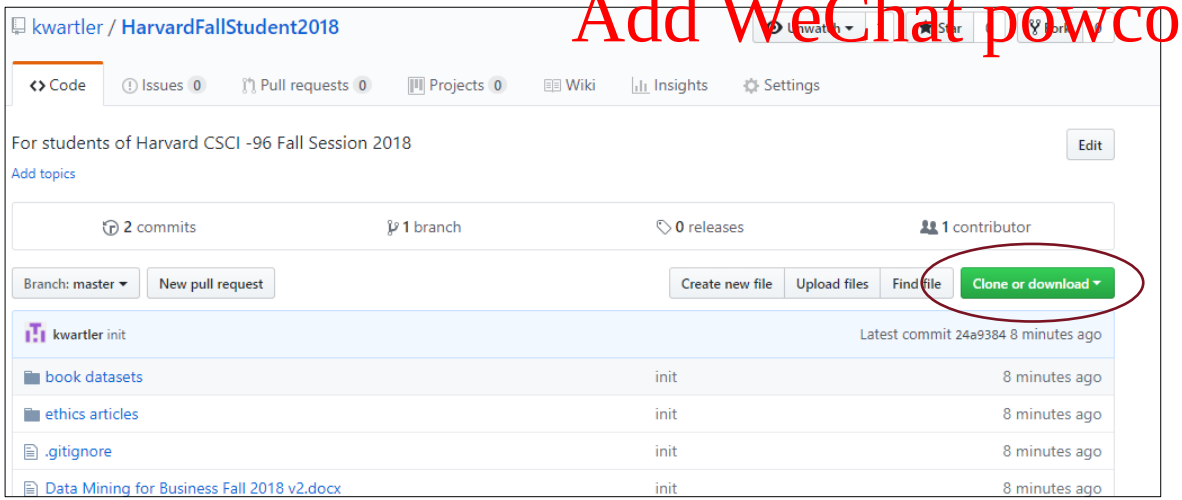
To get scripts, ppts & data use the class repo

https://github.com/kwartler/Harvard_DataMining_Business_Student

If you have git software, when do a "git pull" in Rstudio.



Alternatively you can download a zip of the repo on github.com but this can be cumbersome with file updates.



Rmd (Markdown) & HTML

- For nearly all assignments, you will be required to submit your solutions in the forms of Rmd and HTML files. This will ensure to show not only your R code but also the outputs (numbers, tables, and graphs) generated by the code.
  - R files will NOT be accepted.
- YouTube demo:
  - https://youtu.be/80y2HbyLUJ0

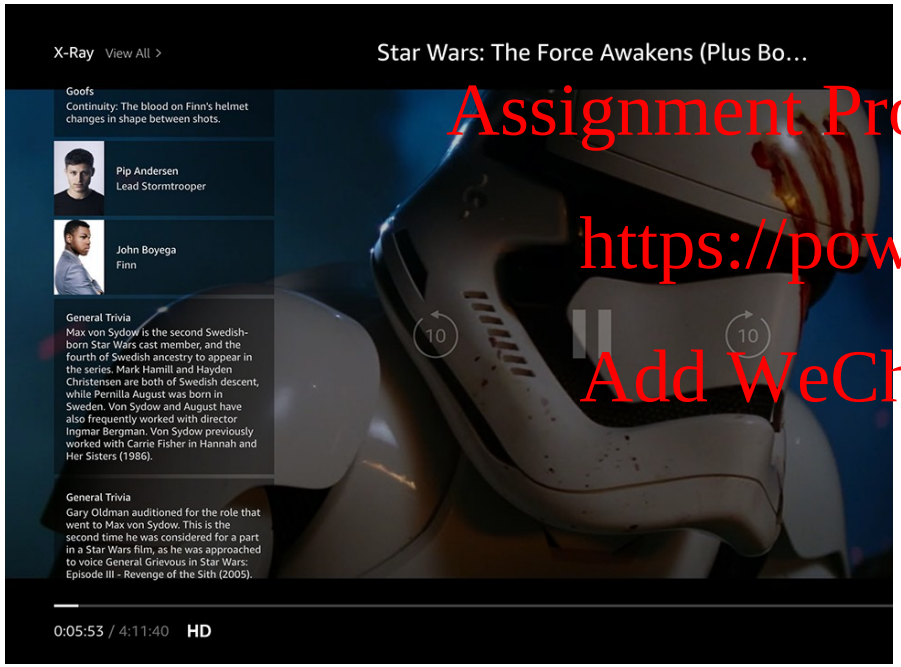
Agenda

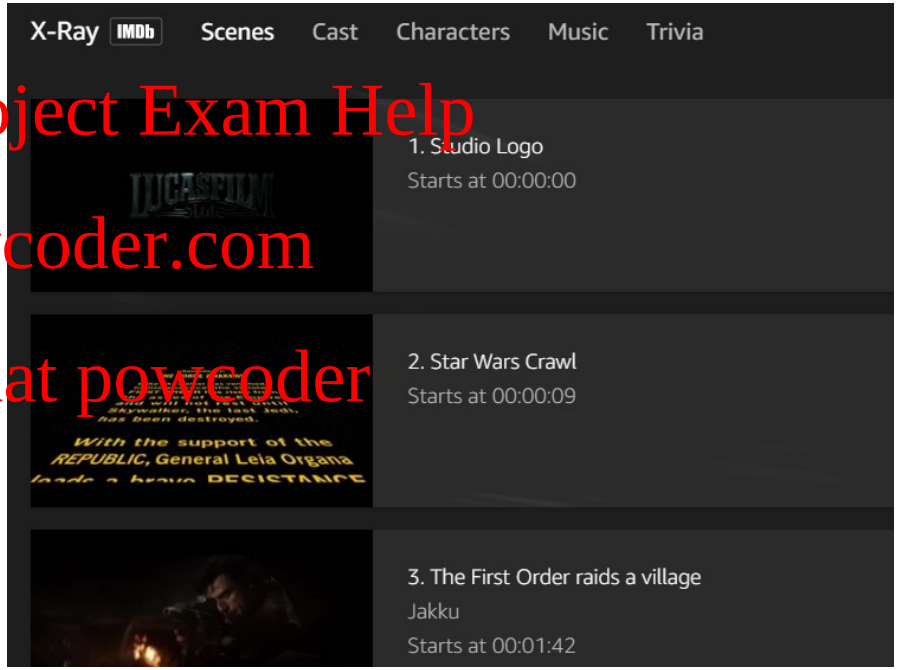
Start	End	Item
		Introductions
		Syllabus Review
		Break
		Intro to R
		R Installation, environment & git
		Simple R Scripting

Amazon X-Ray: A basic R-Script

Left Hand Screen:
Actor Information & Trivia

Center “Pop-over” Screen:
Scenes, Cast, Characters, Music & Trivia





HARVARD UNIVERSITY

Let's practice!

Open A_demo_script.R

id	name	fictionalLocation	start	end
/xray/scene/1	1. Studio Logo	NA	0	9000
/xray/scene/2	2. Star Wars Crawl	NA	9000	102000
/xray/scene/3	3. The First Order raids a village	Jakku	102000	573000
/xray/scene/4	4. Poe is held captive by the First Order	Star Destroyer	573000	643000
/xray/scene/5	5. Rey raids an aging Imperial Star Destroyer	Jakku	643000	1004000

Learning Outcome: Following a consistent workflow, load, clean, & analyze real data with multiple data types to establish basic familiarity.

Housekeeping , Reading & Homework

- Make sure you have R Studio Installed
  - Sign up at RStudio Cloud
- Recommended to install Git locally for "*git pull*"
  - Connect your local to the class repo
- Practice knitting a Rmd file to generate its HTML file
- Piazza case group organization: Facilitated by Jay.

- Chapter 1 & Chapter 2

1. Initial Reflection essay (12-15 sentences)
2. Piazza introduction post

3. C2.1 Data Mining Techniques

Feb 4th

4. C2.2 Data Partition
5. C2.3 Data Sample
6. C2.4 Modeling Steps