
CSCI E-96

Sept 17, 2018

Intro to R

Data Types & EDA



Agenda

Start	End	Item
		What is R, R-Studio, scripting basics
		Executing a first script
		Visualizing Real Data & EDA
		Working with Geospatial Information
		Break
		R's Data Structures in Detail
		EDA
		Housekeeping, Reading & Homework

Intro to R

Learning Objectives

- What is R?
- What is R Studio?
- Why learn R?
- Scripting Structure



What is R?



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

A screenshot of the RGui (32-bit) window. The window has a menu bar with 'File', 'Edit', 'View', 'Misc', 'Packages', 'Windows', and 'Help'. Below the menu bar is a toolbar with icons for file operations. The main area is the 'R Console', which displays the R startup message: 'R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree" Copyright (C) 2017 The R Foundation for Statistical Computing Platform: i386-w64-mingw32/i386 (32-bit)'. It also includes information about the license, contributors, and how to get help. The prompt '> |' is visible at the bottom of the console.

```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> |
```

R is a language and 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 and statisticians worldwide.

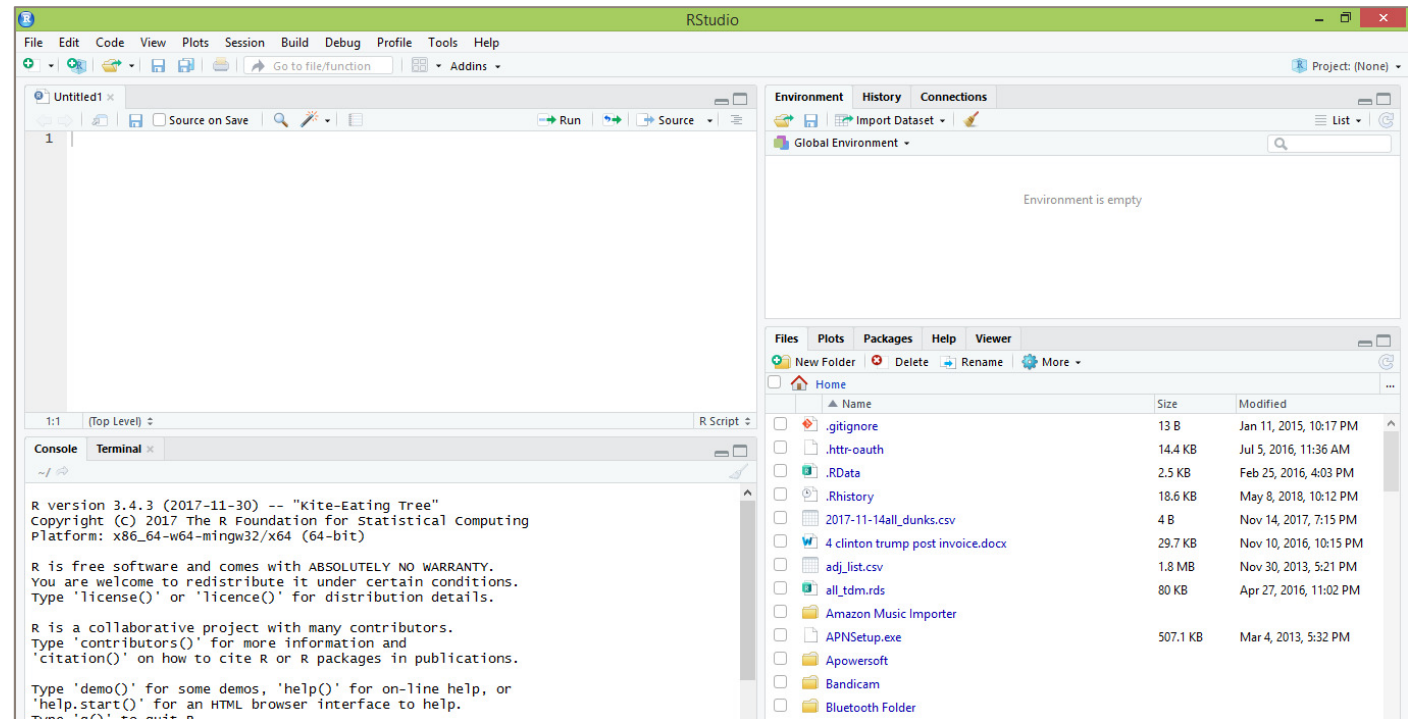
How Companies Use R to Compete in a Data-Driven World, Data-informed.com





What is R Studio?

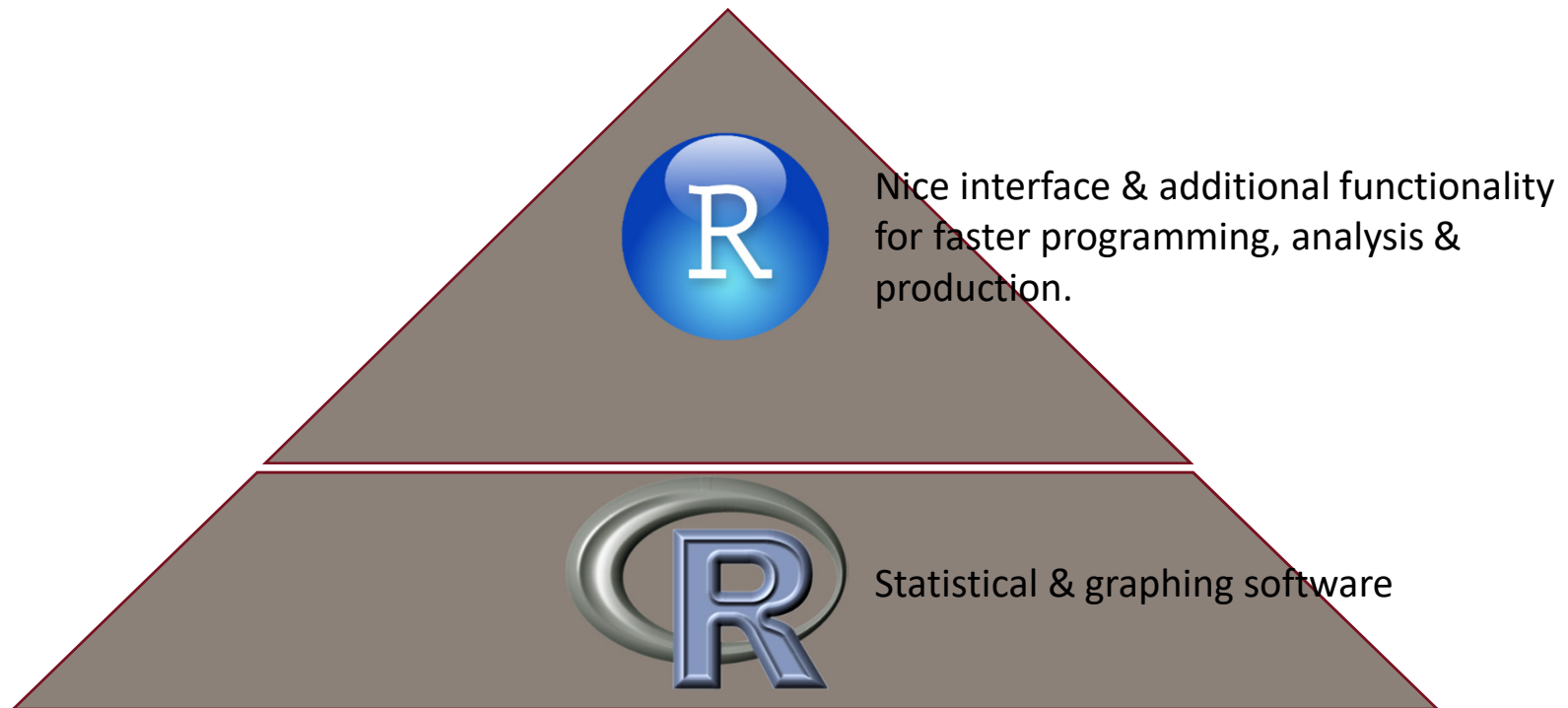
- ✓ IDE – Integrated Development Environment
- ✓ Adds additional functionality e.g. git, shiny projects, markdown templates



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



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

The screenshot shows the RStudio application window with four main panes, each highlighted by a red box and a descriptive text box:

- Scripting:** The top-left pane where R code is written and edited. A text box explains: "Commands are written & adjusted here then executed in the console. Then saved for future quick execution."
- Environment:** The top-right pane showing the current environment. A text box explains: "List of objects & values e.g. loaded 'excel' files".
- Console:** The bottom-left pane where R commands are executed and results are displayed. A text box explains: "Execution of scripts and commands, data exploration and code results".
- Files, Plots, Packages, Help:** The bottom-right pane, which is a file explorer. A text box explains: "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.

R: Do you get rewarded for knowing it?

Top 10 Languages ieee.org

2015	Language Rank	Types	Spectrum Ranking
	1. Java	🌐 📱 🖥️	100.0
	2. C	📱 🖥️ 🧪	99.2
	3. C++	📱 🖥️ 🧪	95.5
	4. Python	🌐 🖥️	93.4
	5. C#	🌐 📱 🖥️	92.2
	6. PHP	🌐 🖥️	84.6
	7. Javascript	🌐 📱	84.3
	8. Ruby	🌐 🖥️	78.8
	9. R	🖥️	74.0

2016	Language Rank	Types	Spectrum Ranking
	1. C	📱 🖥️ 🧪	100.0
	2. Java	🌐 📱 🖥️	98.1
	3. Python	🌐 🖥️	98.0
	4. C++	📱 🖥️ 🧪	95.5
	5. R	🖥️	87.9

2017	Language Rank	Types	Spectrum Ranking
	1. Python	🌐 🖥️	100.0
	2. C	📱 🖥️ 🧪	99.7
	3. Java	🌐 📱 🖥️	99.5
	4. C++	📱 🖥️ 🧪	97.1
	5. C#	🌐 📱 🖥️	93.2
	6. R	🖥️	87.7

Example Salaries



R salaries are very strong and given the language's popularity, the need for more R fluent business leaders and programmers is likely to remain for some time.

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



R Uses Functions, Libraries & Objects

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



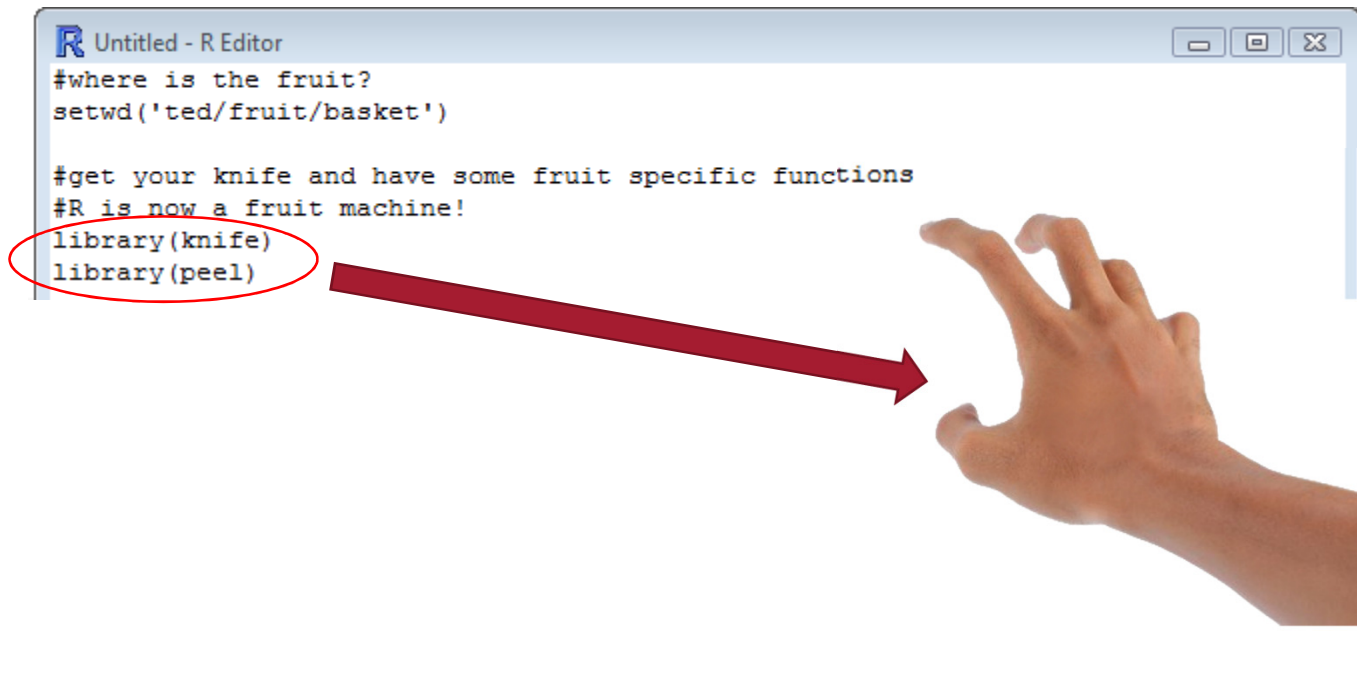
Use the function `setwd` (set working directory) to declare the folder where your data files are within quotes...the fruit basket containing the stuff you want. Ever wonder what the current working directory is? Use `getwd()` with empty parentheses!



The folder path needs to be in quotes and slashes are reversed in Windows not on Apple or Linux.

R Uses Functions, Libraries & Objects

Found the fruit! What tools do I need?



Use `library()` with the name of a specialized package, without quotes, to change R to a specialized piece of software.



Before loading a library use `install.packages("name of package")`. You only need to do this once.

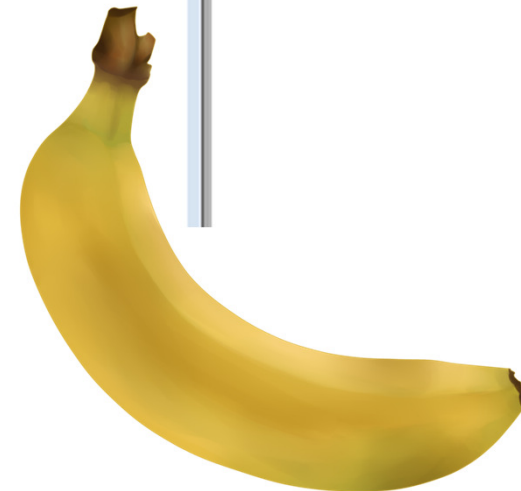
R Uses Functions, Libraries & Objects

Now R is a fruit cutting machine. Let's pick our fruit.

```
R Untitled - R Editor
#where is the fruit?
setwd('ted/fruit/basket')

#get your knife and have some fruit specific functions
#R is now a fruit machine!
library(knife)
library(peel)

#pick up some fruit
banana<-read.fruit('banana.csv')
```



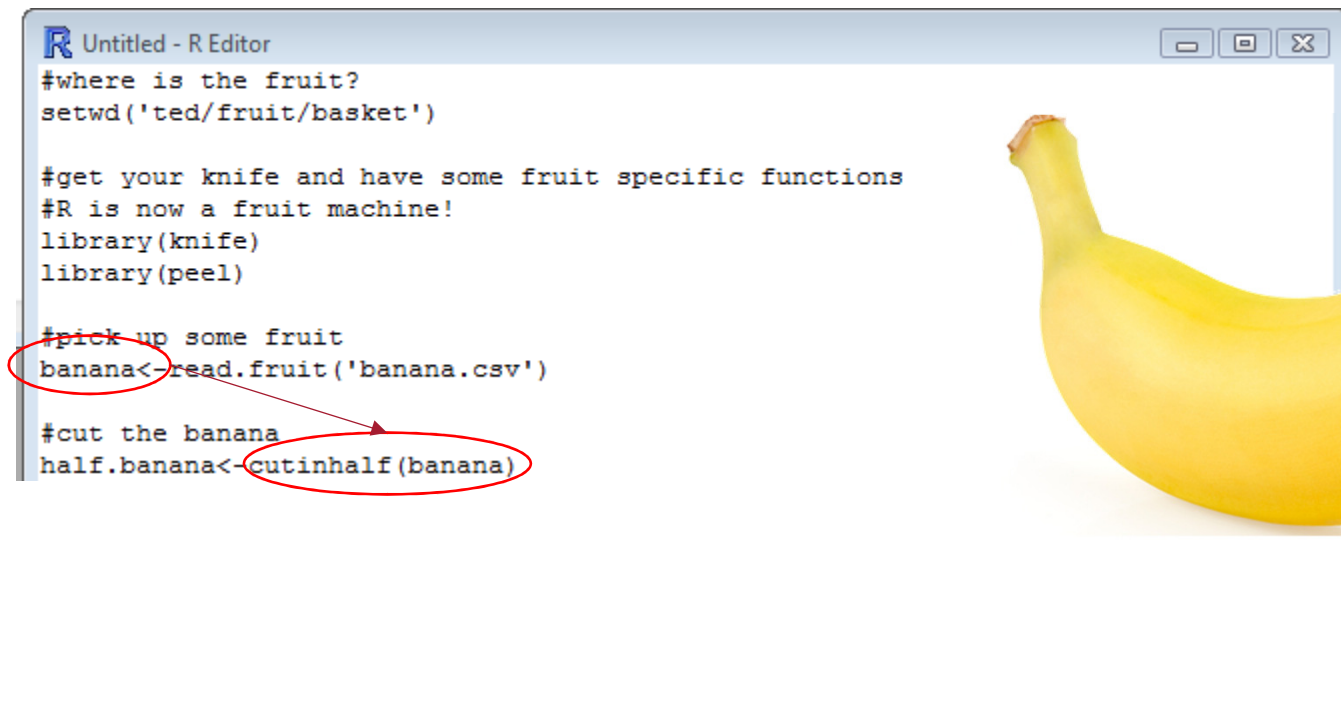
Use a `read.csv()` or similar function to get the object from the working directory. There are multiple read-like functions to get SQL connections, APIs, Excel, and other file types.



The object name must be in quotes and actually be in the directory...watch spelling and capitalization!.

R Uses Functions, Libraries & Objects

Use a function from the **knife** library to create a new object.



The image shows an R Editor window titled 'Untitled - R Editor' with the following code:

```
#where is the fruit?
setwd('ted/fruit/basket')

#get your knife and have some fruit specific functions
#R is now a fruit machine!
library(knife)
library(peel)

#pick up some fruit
banana<-read.fruit('banana.csv')

#cut the banana
half.banana<-cutinhalf(banana)
```

Red circles highlight the function calls `read.fruit('banana.csv')` and `cutinhalf(banana)`. A red arrow points from the first circle to the second. To the right of the code window is a photograph of a yellow banana with a red-handled knife stuck into its side.

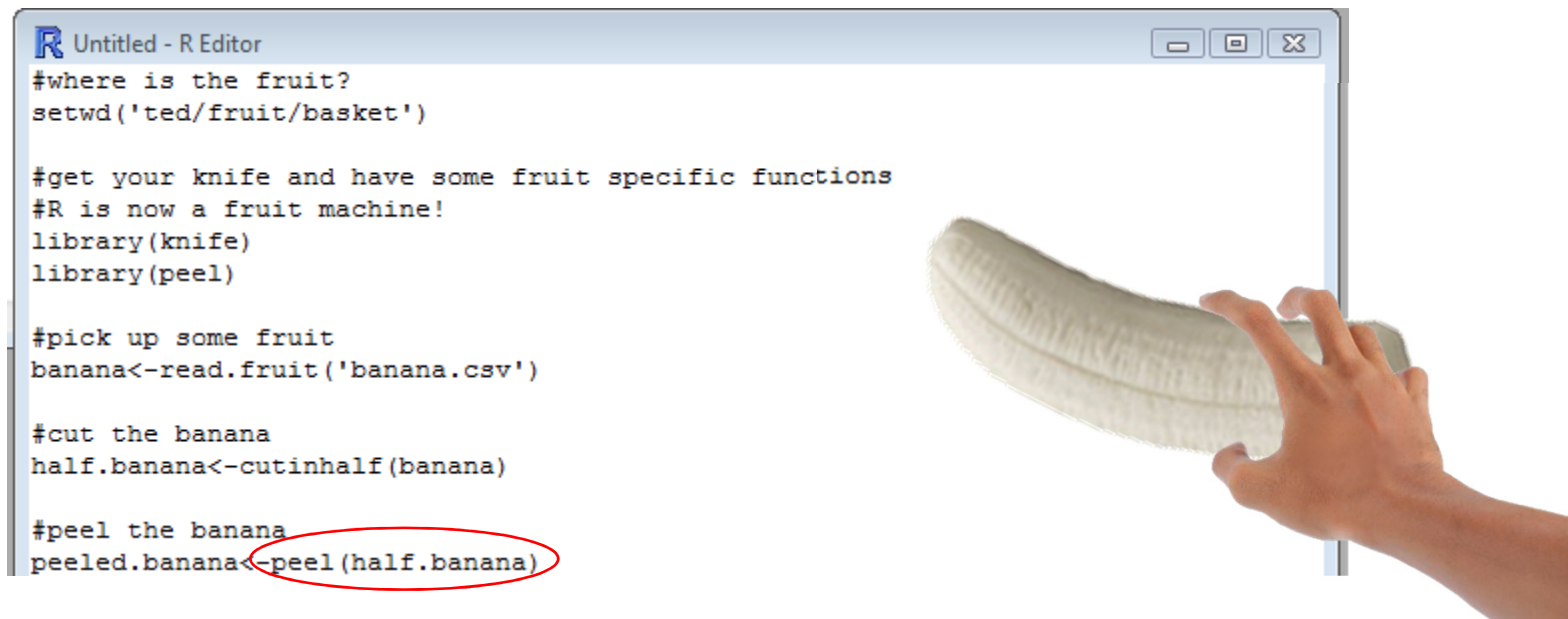
Functions are applied to an object to change it in some way or create a **new object**. Here the R instance has two objects, **banana** & **half.banana**



Be careful, you can overwrite an object pretty easily!

R Uses Functions, Libraries & Objects

Use a function from the **hand** library to create another new object.



```
Untitled - R Editor
#where is the fruit?
setwd('ted/fruit/basket')

#get your knife and have some fruit specific functions
#R is now a fruit machine!
library(knife)
library(peel)

#pick up some fruit
banana<-read.fruit('banana.csv')

#cut the banana
half.banana<-cutinhalf(banana)

#peel the banana
peeled.banana<-peel(half.banana)
```

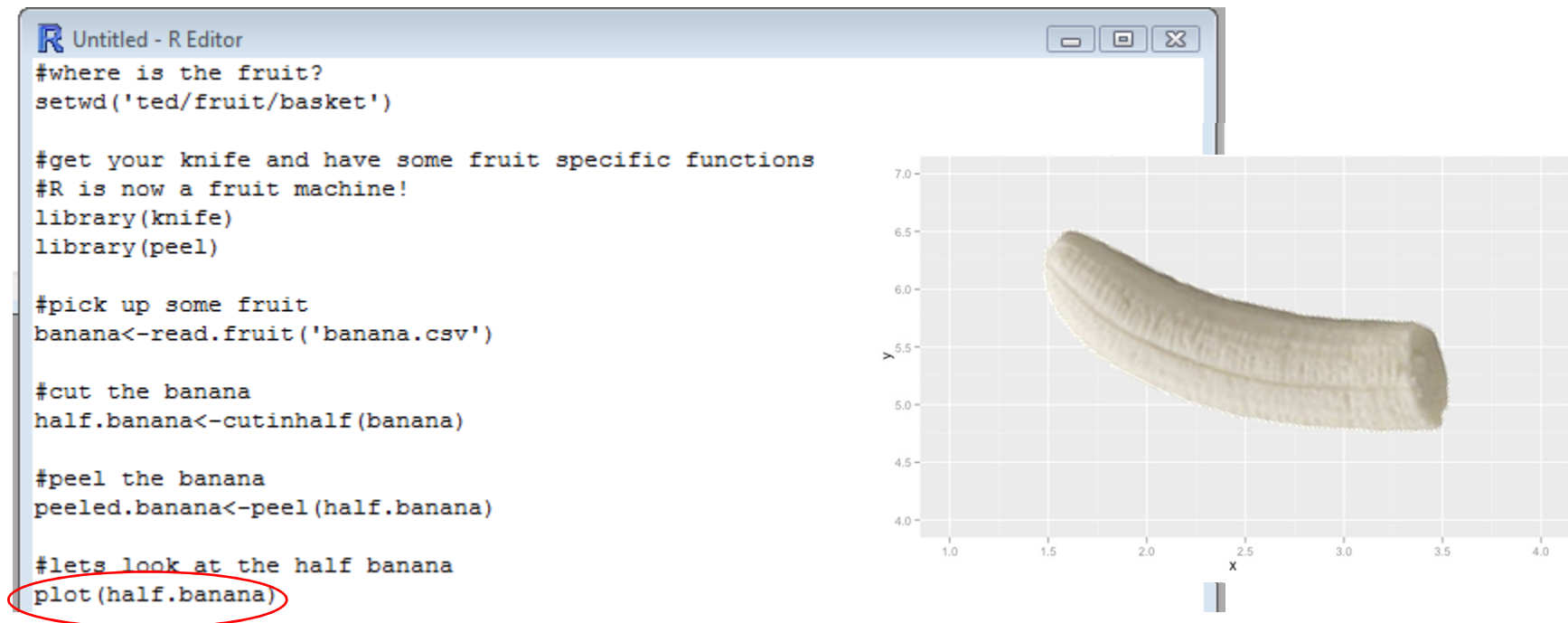
Functions are applied to an object to change it in some way or create a **new object**. Here the R instance has three objects, **banana**, **half.banana** & **peeled.banana**



Be careful, you can overwrite an object pretty easily!

R Uses Functions, Libraries & Objects

Now it's time to consume the outcome.



Here we call **plot** on the object to consume it after our adjustments. You can save files, view graphs, adjust databases, send emails, make reports, update values and more as the result of an R script.



Be careful, you can overwrite objects when SAVING too!

9/19/2018

Kwartler CSCI S-96

16



HARVARD
UNIVERSITY

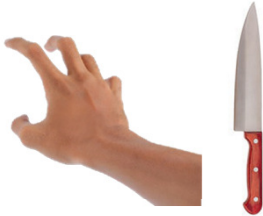
Let's Review.

1



- Use `setwd()` to point to your files.

2



- Load some customized libraries with `library()` for your specific analysis and methodology.

3



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

4



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

5



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

6




- Consume the results by saving, plotting etc.

Open R Studio

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

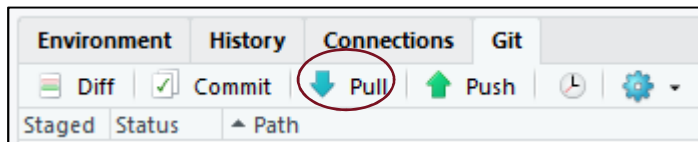
If you don't have it installed yet, please do so now!!



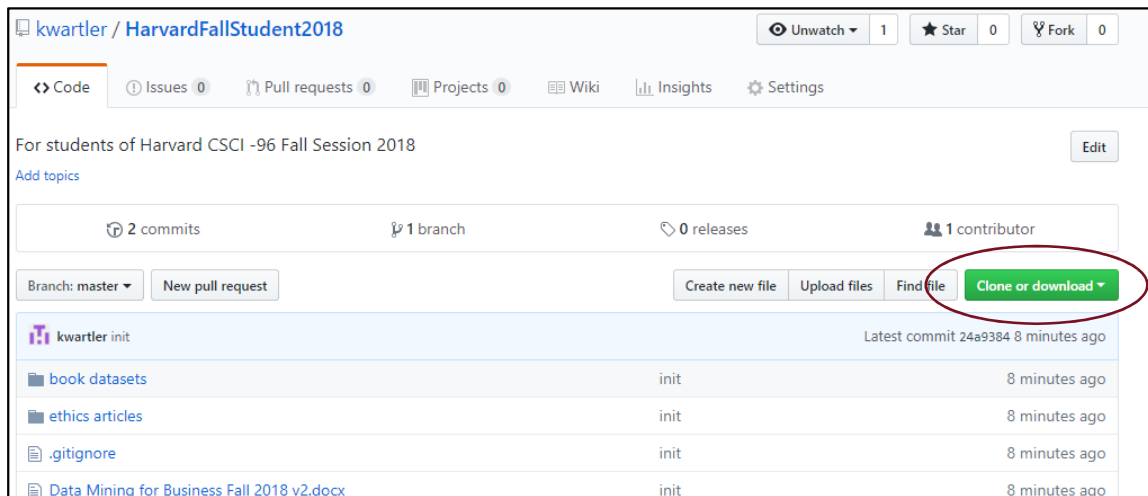
Perform a Git Pull to get the scripts & data

github.com/kwartler/HarvardFallStudent2018

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.



Agenda

Start	End	Item
		What is R, R-Studio, scripting basics
		Executing a first script
		Visualizing Real Data & EDA
		Working with Geospatial Information
		Break
		R's Data Structures in Detail
		EDA
		Housekeeping, Reading & Homework



Let's Practice!

Open:

A_Basic_Test_Drive.R

- Simple Operators ie “+”
- Define Variables using “<-”
- Review objects and types
- Use paste()
- Find help with “?”
- Create a data.frame()
 - Add a column
 - Navigate the DF
 - write.csv()
 - read.csv()
- scatterplot()
- tables()
- barplot()
- Saving a visual as a .jpeg
- “IF” Conditional loops
- “FOR” loops

Open the first script, get familiar with the basic R operations by execution.



Agenda


Start	End	Item
		What is R, R-Studio, scripting basics
		Executing a first script
		Visualizing Real Data & EDA
		Working with Geospatial Information
		Break
		R's Data Structures in Detail
		EDA
		Housekeeping, Reading & Homework




Do or Do Not There is No Try...

X-Ray View All > Star Wars: The Force Awakens (Plus Bo... Options ▾ 🔊 🔍 ✕

Goofs
Continuity: The blood on Finn's helmet changes in shape between shots.

 **Pip Andersen**
Lead Stormtrooper

 **John Boyega**
Finn

General Trivia
Max von Sydow is the second Swedish-born Star Wars cast member, and the fourth of Swedish ancestry to appear in the series. Mark Hamill and Hayden Christensen are both of Swedish descent, while Pernilla August was born in Sweden. Von Sydow and August have also frequently worked with director Ingmar Bergman. Von Sydow previously worked with Carrie Fisher in *Hannah and Her Sisters* (1986).

General Trivia
Gary Oldman auditioned for the role that went to Max von Sydow. This is the second time he was considered for a part in a Star Wars film, as he was approached to voice General Grievous in *Star Wars: Episode III - Revenge of the Sith* (2005).

0:05:53 / 4:11:40 HD

Let's explore Amazon's x-ray feature

Character Background

	A	B	C	D
1	char.name	char.story	char.url	
2	General Hux	Ruthless commander in power struggle with Kylo Ren for the First Order leadership and being exceeded only by Snoke.	http://ia.media-imdb.com/images/M/MV5B1	
3	Poe Dameron	Poe Dameron is portrayed by Oscar Isaac in Star Wars: Episode VII The Force Awakens. Isaac's casting in the film was	http://ia.media-imdb.com/images/M/MV5B1	
4	Maz Kanata	A thousand-year old female pirate and past acquaintance to Han Solo. Around thirty years after the Battle of Endor, Maz l	http://ia.media-imdb.com/images/M/MV5B1	
5	Unkar Plutt	Crolute junk dealer on Jakku. He is very stingy with food ration payments to Rey, until he sees BB-8 with her and offers h	http://ia.media-imdb.com/images/I/81cVbp	
6	Finn	Finn is a former storm trooper (FN-2187) who befriends Poe Dameron, Rey, Han Solo, Chewbacca, and General Organ	http://ia.media-imdb.com/images/M/MV5B1	
7	Snap Wexley	NA	http://ia.media-imdb.com/images/M/MV5B1	
8	Captain Phasma	Legion Commander who reports to General Hux. She wears special armor that can change shape and purpose based u	http://ia.media-imdb.com/images/M/MV5B1	

Official Scenes

	A	B	C
1	defined.scenes	start	end
2	Studio Logo	0	9
3	Star Wars Crawl	9	102
4	The First Order raids a village; Jakku	102	573
5	Poe is held captive by the First Order; Star Destroyer	573	643
6	Rey raids an aging Imperial Star Destroyer; Jakku	643	1004

Character Appearances

1	character	start	end
2	BB-8 Performed By	157	573
3	Lor San Tekka	171	573
4	Poe Dameron	171	573
5	Jakku Villager	236	573
6	Finn	336	573



But first, what is a ggplot?

The first layer is to define a ggplot, with screenTime as the data. The aesthetics (aes) define that information should be colored by each character. However, there is no other information at this point.

```
ggplot(screenTime, aes(colour=screenTime$character))
```



Ggplot is a “grammar of graphics” package. It works by adding layers with an “+” to construct a visual.

Understanding ggplot...

The second layer adds a line segment for each character and defines the size of each.

```
ggplot(screenTime, aes(colour=screenTime$character)) +  
  geom_segment(aes(x=screenTime$start, xend=screenTime$end, y=screenTime$character,  
    yend=screenTime$character), size=3)
```



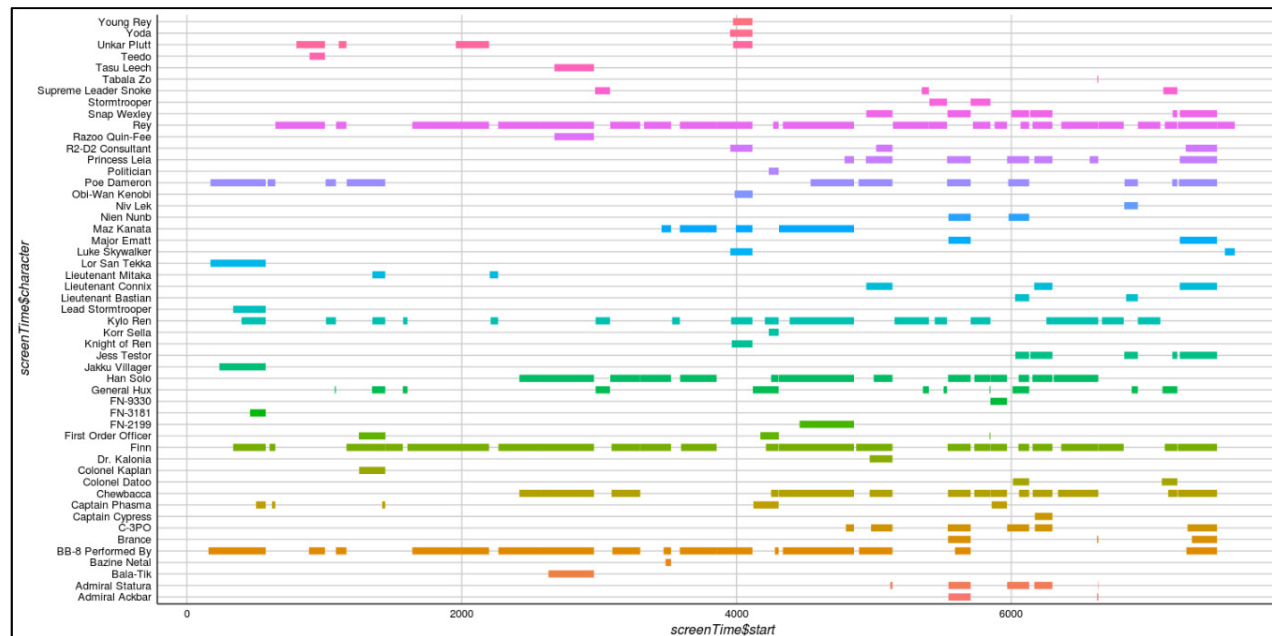
Understanding ggplot...

The third layer changes the background, axis & colors. The fourth layer removes the legend which is redundant in this context.

```
ggplot(screenTime, aes(colour=screenTime$character)) +
```

```
geom_segment(aes(x=screenTime$start, xend=screenTime$end, y=screenTime$character, yend=screenTime$character), size=3) +
```

```
theme_gdocs() + theme(legend.position="none")
```



Let's Practice!

Open:

B_Functions_EDA_Viz.R

- read.csv
- dim()
- table()
- indexing
- subset()
- sample()
- as.matrix()
- barplot()
- ggplot()
- Bokeh::figure()

Open the second script, get familiar with libraries, reading data, functions applied to objects & making visuals.



Agenda

Start	End	Item
		What is R, R-Studio, scripting basics
		Executing a first script
		Visualizing Real Data & EDA
		Working with Geospatial Information
		Break
		R's Data Structures in Detail
		EDA
		Housekeeping, Reading & Homework



Let's Practice on geospatial data!

Open:
C_geospatial.R

- read.csv
- ggplot()
- Google maps with ggmap
- leaflet()

Open the third file, explore geospatial information.



Agenda

Start	End	Item
		What is R, R-Studio, scripting basics
		Executing a first script
		Visualizing Real Data & EDA
		Working with Geospatial Information
		Break
		R's Data Structures in Detail
		EDA
		Housekeeping, Reading & Homework



Data Structures


- R Data Types
- R Data Structures
- Exploratory Data Analysis




Common R Object Types - Vectors

Objects in R can be various forms and even made to be “custom” types.

Numeric/Integer




1
10
12
3.47
82




```
c(1,10,12,3.47)
```

Boolean




TRUE
TRUE
FALSE
TRUE
FALSE




```
c(T, T, F, T, F)  
c(TRUE, TRUE, FALSE, TRUE, FALSE)  
c(T, TRUE, F, TRUE, FALSE)
```

Factors (Distinct Classes)



MALE
FEMALE
FEMALE



```
as.factor(c('MALE','FEMALE','FEMALE'))
```

Unordered

```
[1] MALE FEMALE FEMALE  
Levels: FEMALE MALE
```

High
Med
Low

Ordinal

```
as.factor(c('High','Med','Low'))
```

```
[1] high med low  
Levels: high low med
```

STRING (just text)

```
c('MALE','FEMALE','FEMALE')
```

```
[1] "MALE" "FEMALE" "FEMALE"
```

Cardinality
2
3

In R, a vector can be numeric, Boolean (T/F), factors, or contain strings.

More Complex Common R Object Types - Matrix

Matrices are 2 dimensional data (rows/columns). Each column must be the same type.



RowID	breaks	wool	tension
1	26	A	L
2	30	A	L
3	54	A	L
4	25	A	L
5	70	A	L
6	52	A	L
7	51	A	L
8	26	A	L
9	67	A	L
10	18	A	M



```
> as.matrix(warpbreaks[1:10,])
      breaks wool tension
1    "26"    "A"    "L"
2    "30"    "A"    "L"
3    "54"    "A"    "L"
4    "25"    "A"    "L"
5    "70"    "A"    "L"
6    "52"    "A"    "L"
7    "51"    "A"    "L"
8    "26"    "A"    "L"
9    "67"    "A"    "L"
10   "18"    "A"    "M"
```

All strings

Matrices are organized into rows and columns. In R, the row names are not actually a vector of the matrix but are an attribute of the matrix. In excel you would need a standalone vector to capture that information.



More Complex Common R Object Types – Array

Arrays aren't widely used, except in image analysis (R,G,B matrices)



RowID	breaks	wool	tension		
1	26	A	L	nsion	
2	30	A	L	L	
3	54	A	L	L	nsion
4	25	A	L	L	L
5	70	A	L	L	L
6	52	A	L	L	L
7	51	A	L	L	L
8	26	A	L	L	L
9	67	A	L	L	L
10	18	A	M	L	L
	10	18	A	M	L
		9	67	A	L
		10	18	A	M



```
library(EBImage)

a <- imageData(x)

r <- a[, , 1]
g <- a[, , 2]
b <- a[, , 3]
```

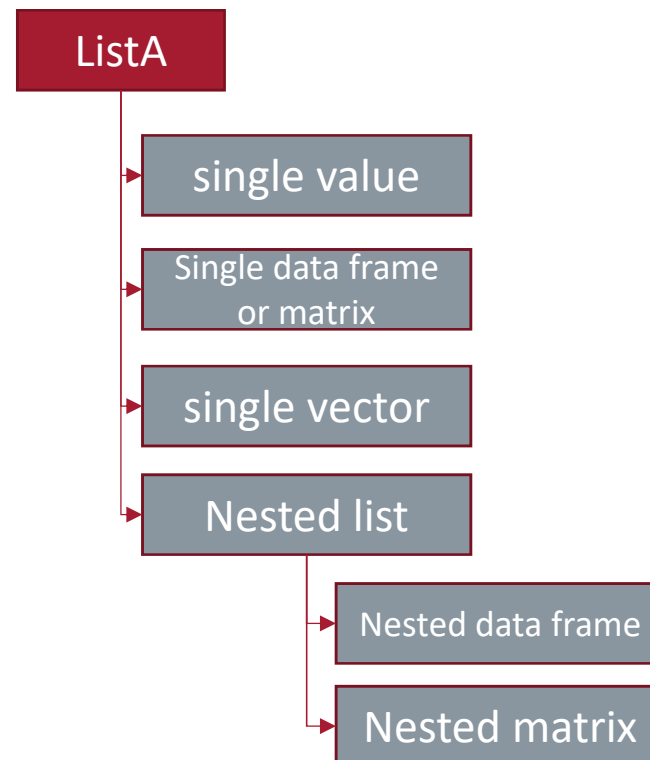
This is an example of extracting RGB data from an image but is not covered in this course since arrays are seldom used.

Arrays can be thought of similar to Excel's workbook which can contain multiple single sheet work books.



More Complex Common R Object Types – List

Lists are multi-dimensional objects that can contain different data types of different lengths.



Lists are useful for data organization but can be complex and difficult to navigate to get specific information.

More Complex Common R Object Types – Data Frame

Data Frames are like 2 dimensional data objects but can have mixed data types.

A data frame is actually a named list but with equal length elements. Being a list lets it contain mixed data types.



RowID	breaks	wool	tension
1	26	A	L
2	30	A	L
3	54	A	L
4	25	A	L
5	70	A	L
6	52	A	L
7	51	A	L
8	26	A	L
9	67	A	L
10	18	A	M



```
> warpbreaks[1:10,]  
  breaks wool tension  
1      26    A      L  
2      30    A      L  
3      54    A      L  
4      25    A      L  
5      70    A      L  
6      52    A      L  
7      51    A      L  
8      26    A      L  
9      67    A      L  
10     18    A      M
```

Integers

Factor

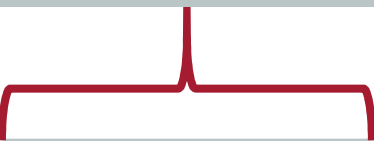
Factor

Data frames are used often because they can hold different types of vectors, but can be switched back and forth with `as.matrix()` and `as.data.frame()`. **Remember that the vector classes could change!!**



Data Structure for Analysis & Modeling

Often the 1st Column is a unique identifier but the identifier could also be a row attribute (not actually a vector)



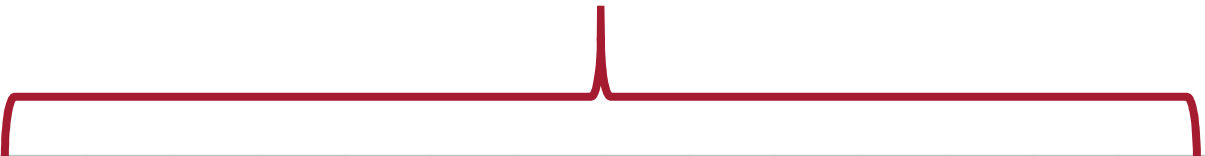
name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100% Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100% Natural Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

Generally we will use data frames to avoid complexity but you will be exposed to other data types.

Data Structure for Analysis & Modeling

Informative features are usually independent & do not lend information to other rows (auto-correlation). Can be called informative columns, independent variables, or features.

Remember in a DF, these can be mixed with decimals, integers, factors, strings, T/F.




name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

Generally we will use data frames to avoid complexity but you will be exposed to other data types.

Data Structure for Analysis & Modeling

If we are doing supervised learning, there is a dependent variable.

This is the outcome and is “dependent” on the informative columns. An analysis with this vector can be binary, classification, or predictive.



name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

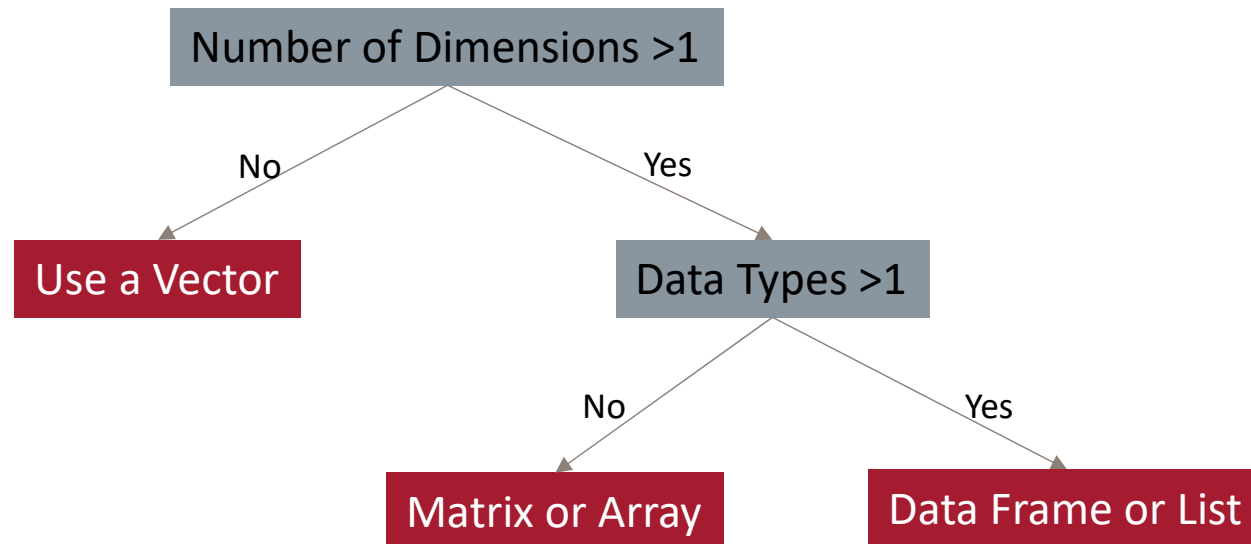
Generally we will use data frames to avoid complexity but you will be exposed to other data types.

Let's Practice!

Open D_R objects.R:

- `c()` to combine values into a vector
- `as.matrix()` to create a matrix object
- `data.frame()`
- `as.list()`
 - List elements by index
 - List elements by name

When should you use a specific data type?



Most analyses start with a data frame, and change classes as needed.

Agenda

Start	End	Item
		What is R, R-Studio, scripting basics
		Executing a first script
		Visualizing Real Data & EDA
		Working with Geospatial Information
		Break
		R's Data Structures in Detail
		EDA
		Housekeeping, Reading & Homework



Data Exploration (EDA)

- Data sets are typically large, complex & messy
- Need to review the data to help refine the task
- Use techniques of Reduction and Visualization



Exploring Data: Sampling to Save Time

- Data mining typically deals with huge databases
- For piloting/prototyping, algorithms and models are typically applied to a sample from a database, to produce statistically-valid results
- Once you develop and select a final model, you use it to “score” (predict values or classes for) the observations in the larger database



Rare Event Over-Sampling

- Often the event of interest is rare
- Examples: response to mailing, fraud in taxes, ...
- Sampling may yield too few “interesting” cases to effectively train a model
- A popular solution: oversample the rare cases to obtain a more balanced training set
- Later, need to adjust results for the oversampling

What are some cases where you think over sampling rare cases makes sense?



Sampling & Oversampling

TABLE 2.4 **SAMPLING IN R**



code for sampling and over/under-sampling

```
# random sample of 5 observations
s <- sample(row.names(housing.df), 5)
housing.df[s,]

# oversample houses with over 10 rooms
s <- sample(row.names(housing.df), 5, prob = ifelse(housing.df$ROOMS>10, 0.9, 0.01))
housing.df[s,]
```

Create an index of random numbers from 1 to the number of rows.

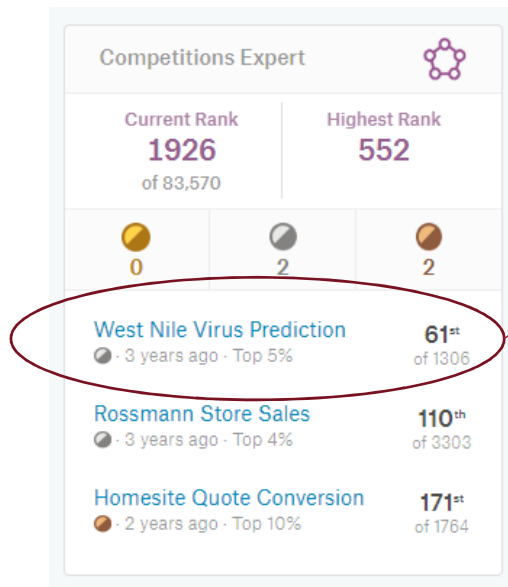
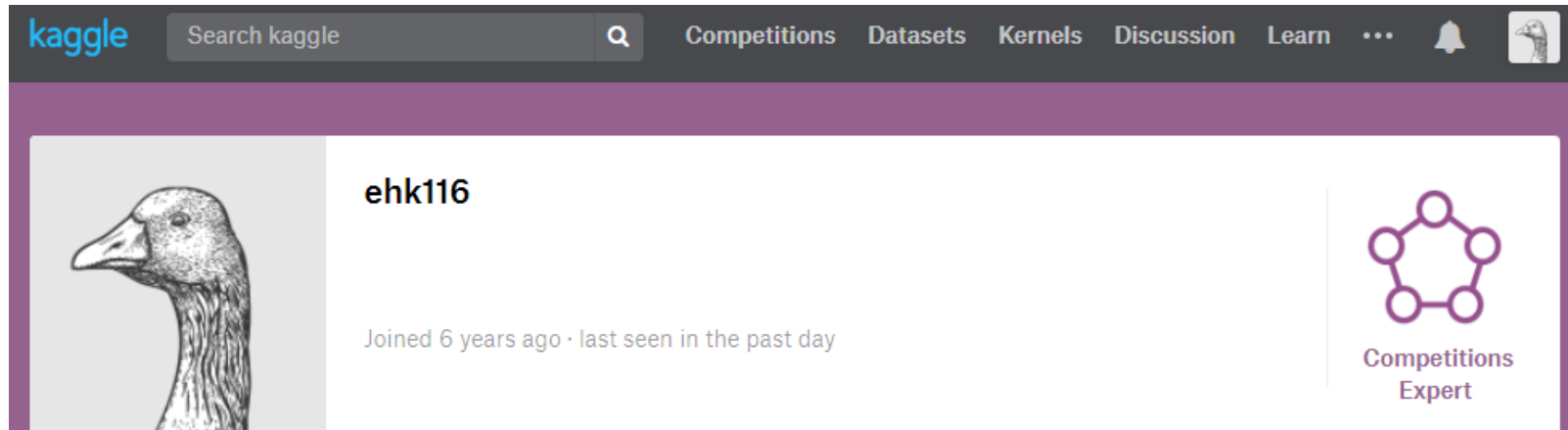
```
idx <- sample(a vector to choose from, the number to choose)
```

Use the index of randomly chosen numbers to select rows

```
dataFrame[ idx, ]
```










What's the value of good EDA?



Asked to predict the presence of West Nile Virus in Chicago mosquitos traps.

EDA let me realize a flaw!

		West Nile Virus Prediction		Predict West Nile virus in mosquitos across the city of Chicago		\$40,000 · 1,306 teams · 3 years ago	
Overview		Data	Kernels	Discussion	Leaderboard	Rules	Team
		My Submissions		Late Submission			
7	▼1	Syowen					
59	▼3	Let's find Mosquito			0.81303		
60	▲35	JustQ			0.81209	56	3y
61	▲19	ehk116			0.81196	25	3y
62	▼8	H2O.ai			0.81171	42	3y
63	▼3	Artem			0.81109	28	3y

Simple EDA by year would show that West Nile was 2x in 2012

After fitting an algorithm, I merely doubled predictions if they were within 2012 for the test set. Not great DS but an easy way to move up the leaderboard.

Let's Practice

Open E_EDA work.R:

- Lots of basic R options
 - `str()`
 - `dim()`
 - `class()`
 - `head()`
 - `nlevels()`
 - `summary()`
 - `cor()`
 - `unique()`
 - `mean()`
 - `colSums()`
 - `is.na()`
- Specific packages make life easier
 - `library(DataExplorer)`
 - `plot_str()`
 - `plot_missing()`
 - `plot_histogram()`
 - `plot_density()`
 - `plot_scatterplot()`
 - `library(radiant.data)`

On this script you will fill in the object, vector and information into the code scaffold. Then spend 5-10min exploring the data with `radiant.data`



Agenda

Start	End	Item
		What is R, R-Studio, scripting basics
		Executing a first script
		Visualizing Real Data & EDA
		Working with Geospatial Information
		Break
		R's Data Structures in Detail
		EDA
		Housekeeping, Reading & Homework

Housekeeping , Reading & Homework

- Prof K is back next week. Email him if you have any questions.
 - Now that the cohort has a level foundation of R knowledge, the real fun begins...applications in a real business scenario!
 - Get your group together, 10-1 OKCupid Case is fast approaching, you can use the EDA functions from today to start working on it.
-

- Chapter 3

6. C2.5 Overfitting
7. C2.6 Data Leakage
8. C2.11 ToyotaCorolla.csv *only “a.”*
9. Day2_Homework_v2.R

