USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

2nd R-Programming Bootcamp

Vivek Singh

Information Systems Decision Sciences (ISDS)

MUMA College Of Business

vivek4@mail.usf.edu

Web: http://vivek4.myweb.usf.edu

About Me

Professional Experience

Alcatel·Lucent

SingTel

NUS National University of Singapore

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

Technologies

C

Linux

amazon web services™

python

R

Teaching @ USF

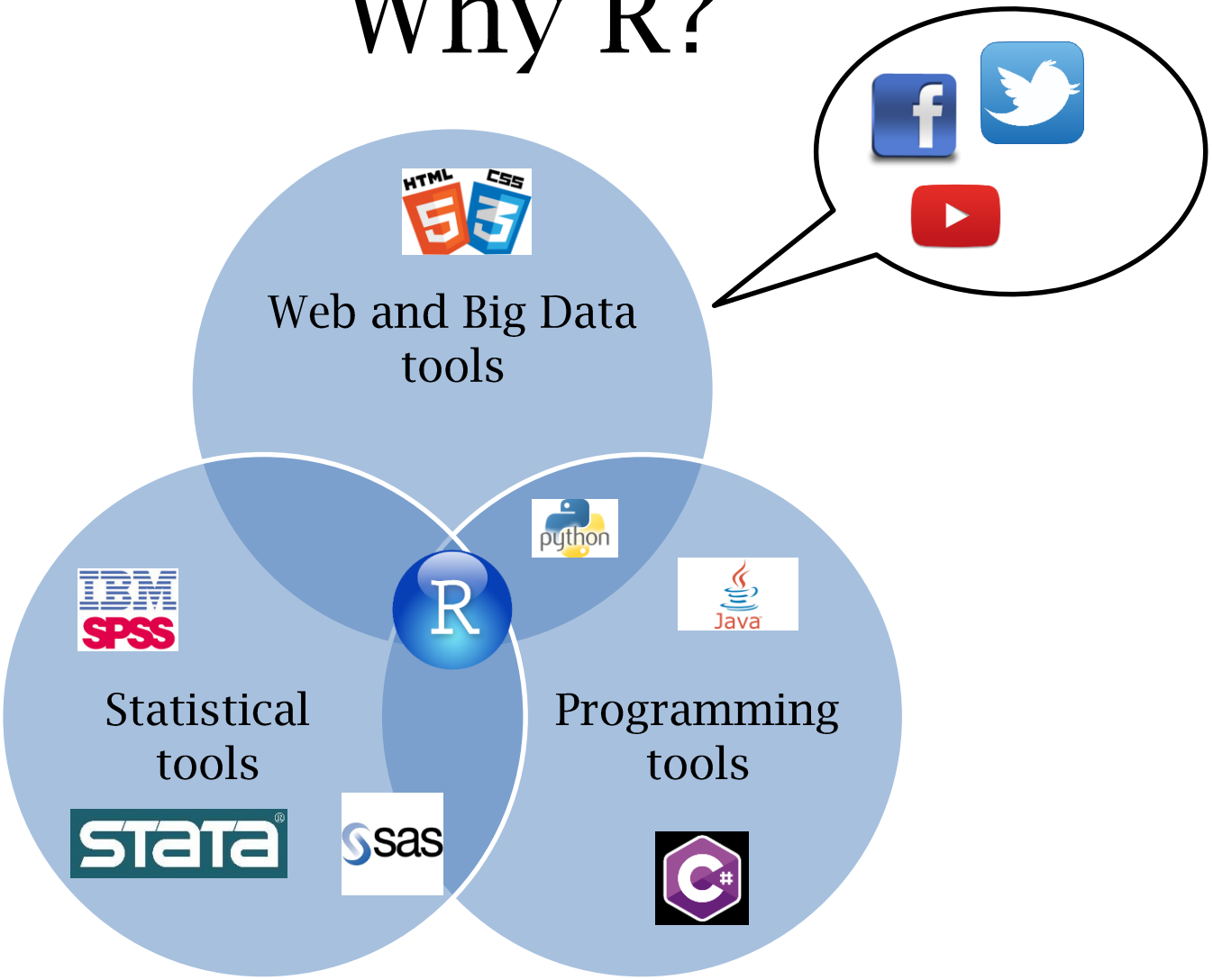
Sprint 2017

ISM 4930: Applied Data Science (Cloud computing and Real-time Business)

Outline

- Motivation – Why R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, YouTube
- CIRCE @USF (Cluster Computing)

Why R?


Web and Big Data tools
Statistical tools
Programming tools
python
Java
IBM SPSS
STATA
sas

R Programming

- Programming environment
  - Data manipulation
  - Computation
- Statistical analysis
- Visualization
- Built from S language

Ross Ihaka

Programming Language Designer

George Ross Ihaka is an Associate Professor of Statistics at the University of Auckland who is recognized, along with Robert Gentleman, as one of the originators of the R programming language. Wikipedia

Born: 1954, Waiuku, New Zealand

Residence: Auckland, New Zealand

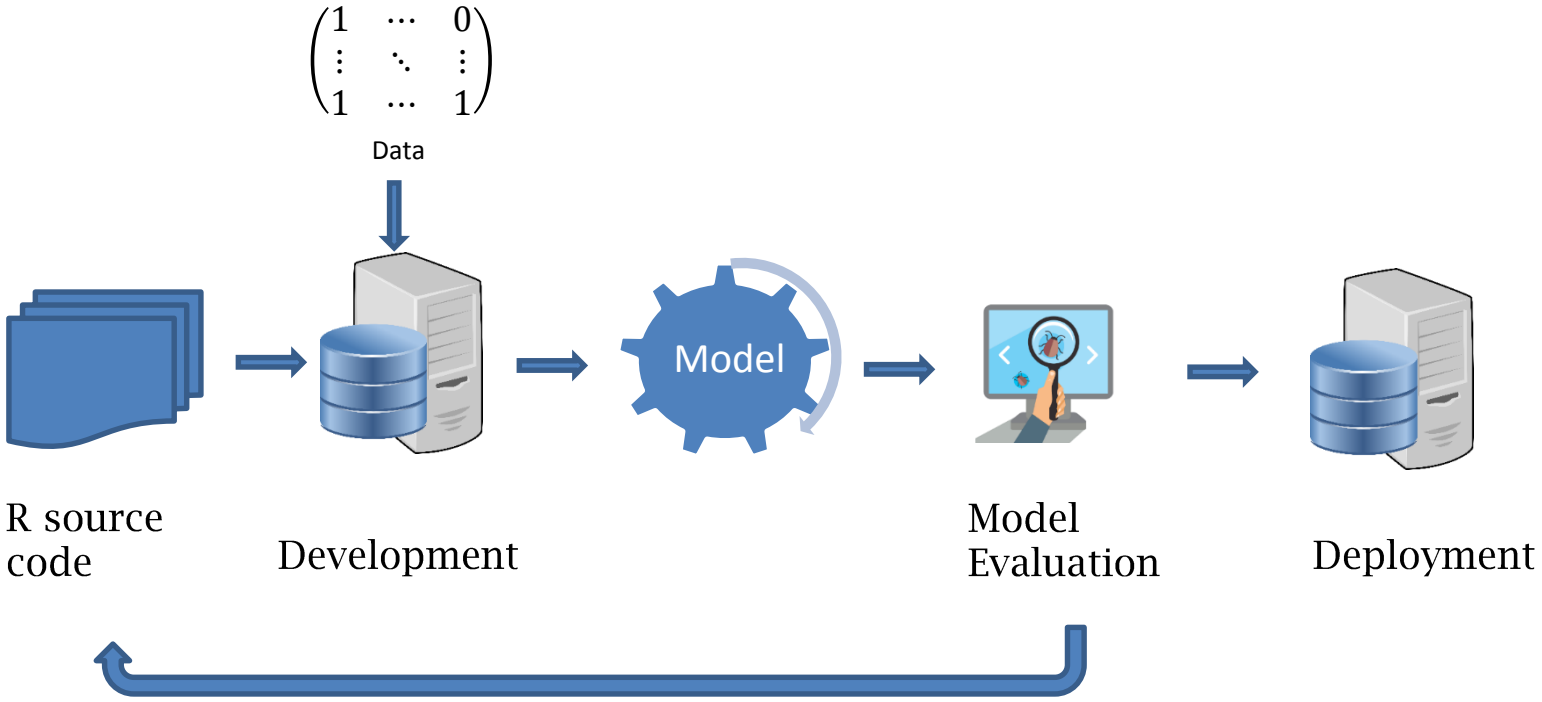
Known for: R

Alma maters: University of Auckland, University of California, Berkeley

Robert Gentleman

Programming Language Designer

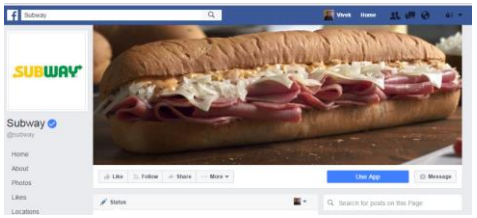
Data Analytics Life Cycle

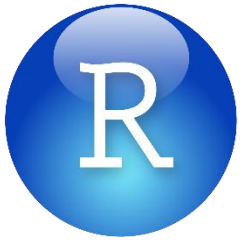


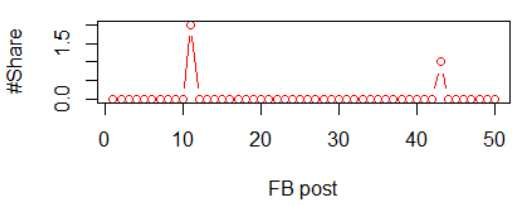
$ 1,000,000

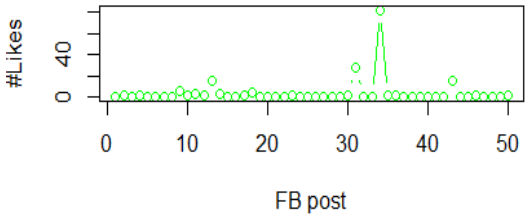
https://www.kaggle.com/c/data-science-bowl-2017

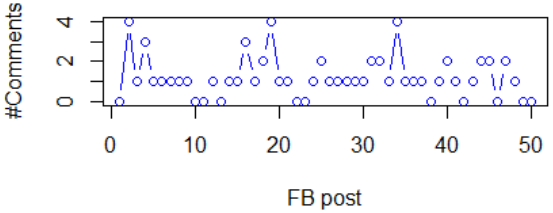
Social Media Analytics Demo with R

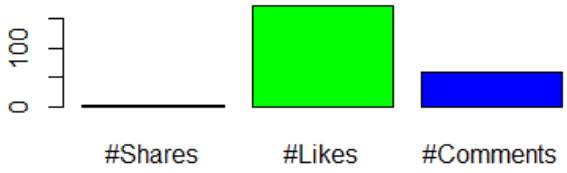












DV: #Share	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	−0.42	0.15	−2.77	0.01	**
#Likes	0.12	0.01	18.88	< 2e−16	***
#Comments	0.24	0.09	2.59	0.01	*

(December 16, 2016)

Run R code on CIRCE @USF

GPU Hardware

Card Model	Quantity	Memory	Additional Info
NVIDIA Kepler K20	40	6GB	2013 Expansion
NVIDIA Fermi	8	2GB	

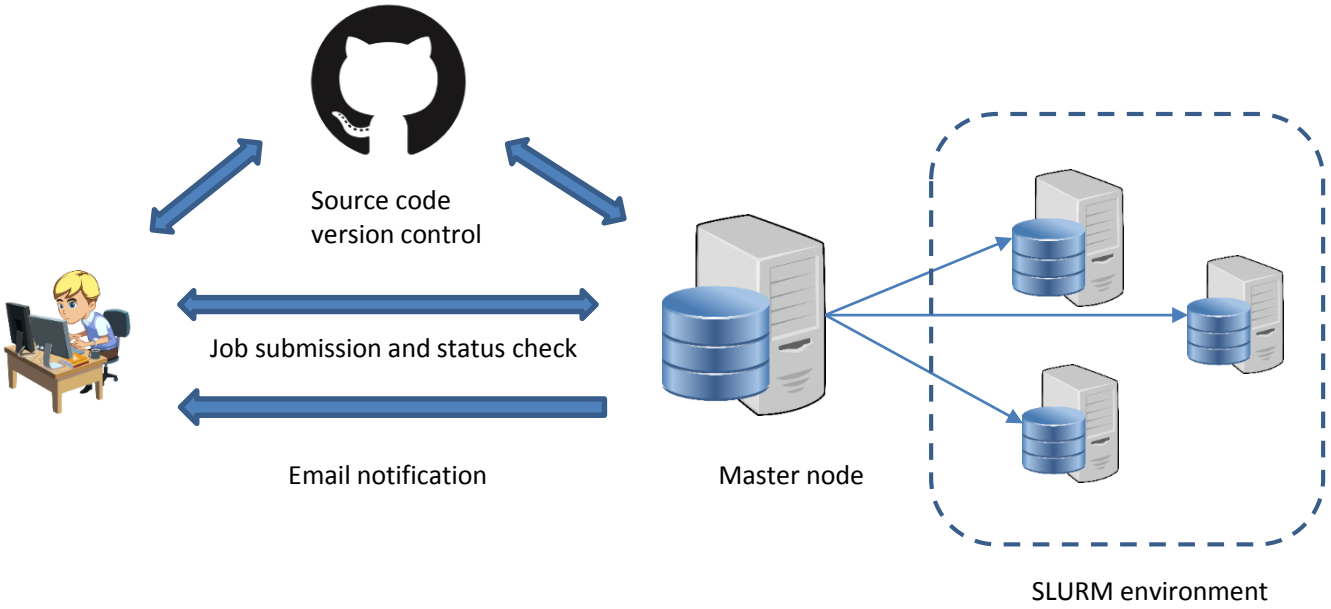
Server Hardware

Nodes	Core Count	Processors	Memory per node	Interconnect	Additional Info	
	138	1656	2 x Intel Xeon E5649 (Six Core)	24GB	QDR InfiniBand	
	128	2048	2 x Intel Xeon E5-2670 (Eight Core)	32GB	QDR InfiniBand	2013 Expansion
	68	816	2 x Intel Xeon E5-2630 (Six Core)	24GB	QDR InfiniBand	
	40	800	2 x Intel Xeon E5-2650 v3 (10-core)	128GB	QDR InfiniBand	hii02 partition
	36	288	2 x AMD Opteron 2384 (Quad Core)	16GB	DDR InfiniBand	
	34	408	2 x AMD Opteron 2427 (Six Core)	24GB	DDR InfiniBand	
	20	320	2 x Intel Xeon E5-2650 v2 (Eight Core)	192GB	QDR InfiniBand	hii01 partition
	20	320	2 x Intel Xeon E5-2650 v2 (Eight Core)	64GB	QDR InfiniBand	hii01 partition
	16	192	2 x Intel Xeon E5-2620 (Six Core)	64GB	QDR InfiniBand	hii01 partition
	4	48	2 x Intel Xeon E5649 (Six Core)	24GB	QDR InfiniBand	Login nodes
	4	80	2 x Intel Xeon E5-2650 v3 (10-core)	512GB	QDR InfiniBand	2015 Large-memory nodes
	2	32	2 x AMD Opteron 6128 (Eight Core)	192GB	DDR InfiniBand	Large-memory nodes
	2	32	2 x AMD Opteron 6128 (Eight Core)	18GB	DDR InfiniBand	
	1	16	4 x Intel Xeon E7330 (Quad Core)	132GB	SDR InfiniBand	Large-memory node
	1	16	2 x Intel Xeon E5-2650 (Eight Core)	32GB	QDR Infiniband	Chemistry GPU node
Totals						
	520	7168		24.6TB		

File System Hardware

File System Path	File System Type	Interconnect	Available Size	Backed Up?	Long-Term Storage	Additional Info
/home	GPFS	QDR Infiniband	2.4PB	Daily	Yes	home directory space for secure file storage

Run R code on CIRCE @USF



Reference

- (Petra Kuhnert and Bill Venables) *An Introduction to R Software for Statistical Modelling & Computing*, CSIRO Mathematical and Information Sciences Cleveland, Australia

- Other books and materials
  - CRAN project: https://cran.r-project.org/other-docs.html

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, Workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

R studio

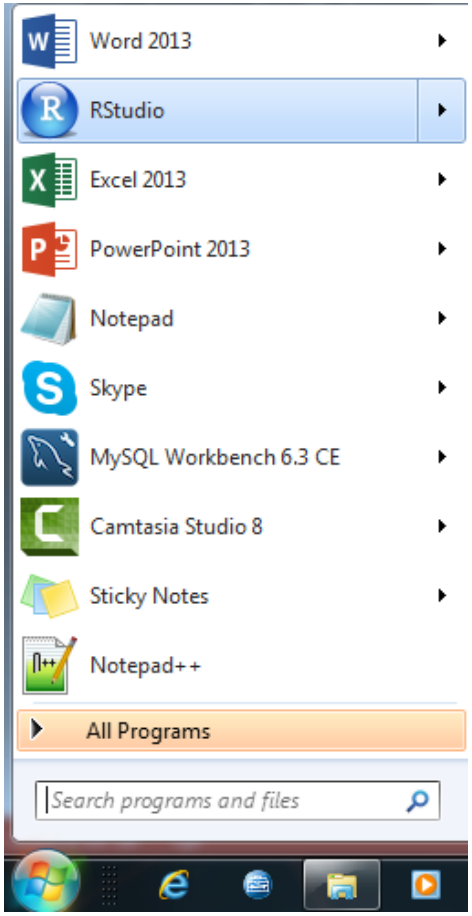
- Download R–studio desktop
  - https://www.rstudio.com/products/rstudio/download3/
  - AGPL license
  - Windows, Mac OS X, Linux (Ubuntu)

Programming IDE– R Studio

- Coding
- Execution
- Debugging

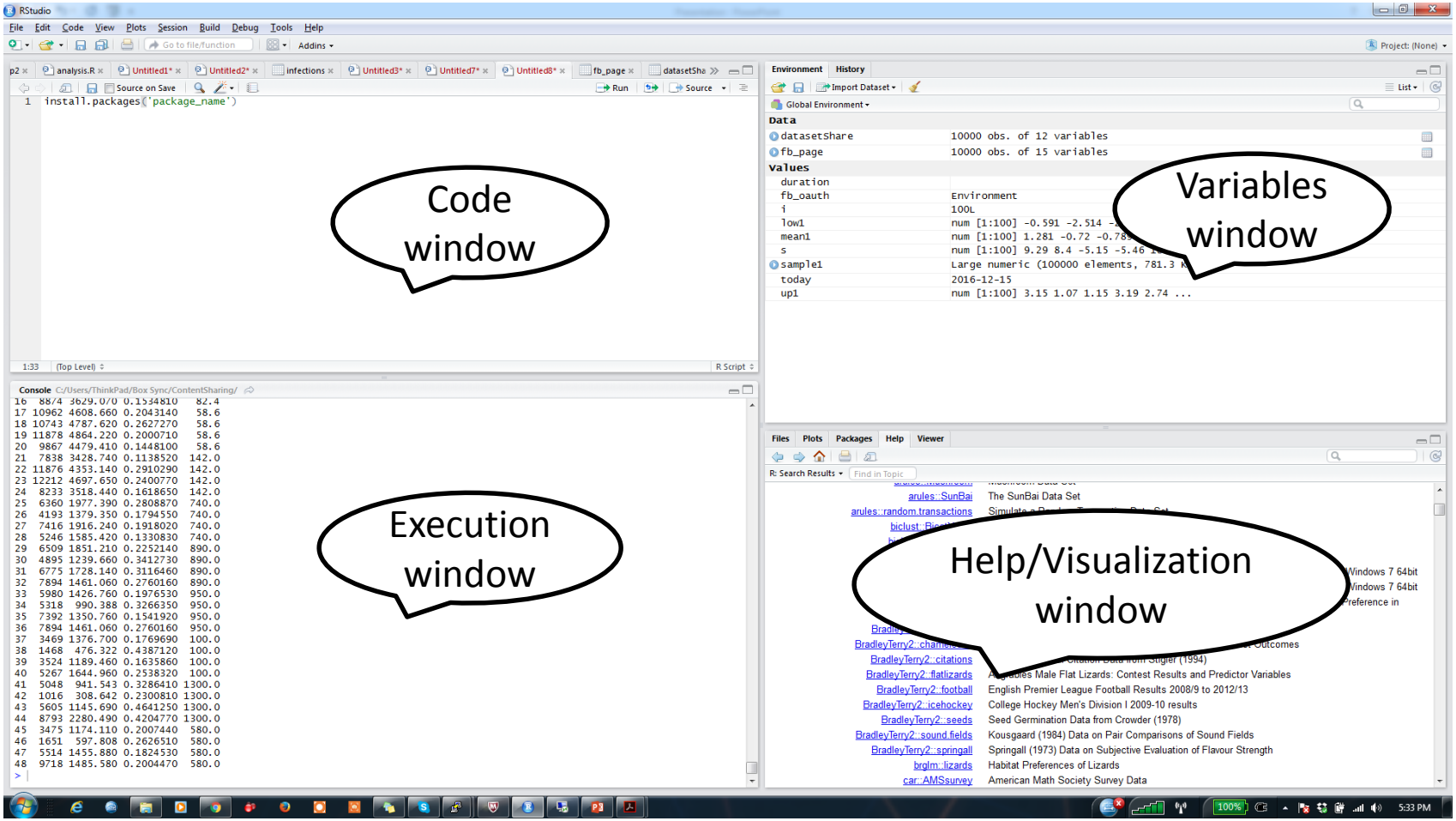
- Batch mode execution
- R–workspace

R–studio interface



To quit R, close the R studio or use q() function

R-studio interface



Hello world!!

Method 1:

Write following code in the execution window

>print('Hello World')

Clear console: CTRL+l

Method 2:

Write the code in the code window

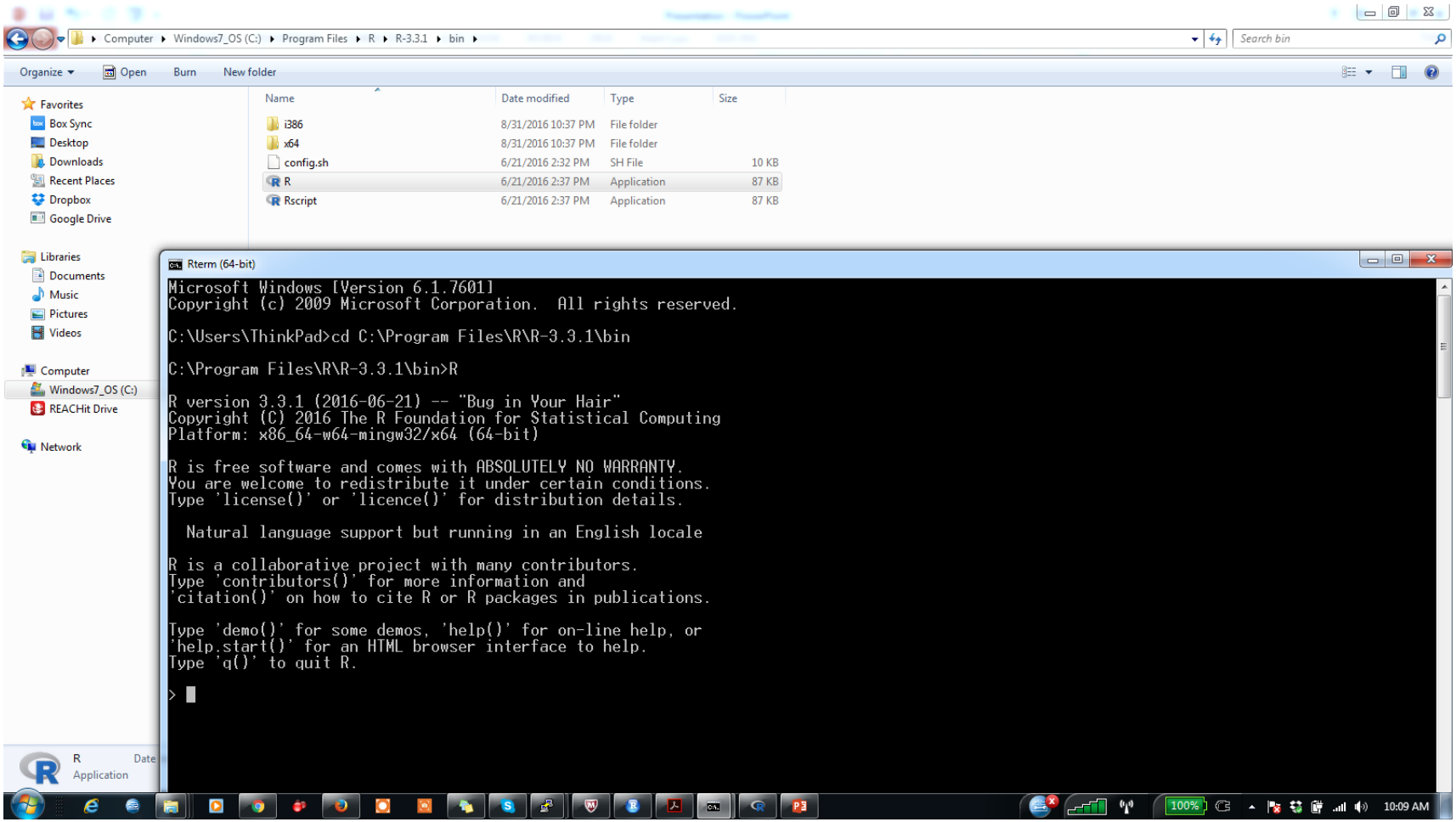
Save as 'script.R' (optional)

Select the code and click Run button OR (CTRL+ Enter)

Save workspace – save.image(file='helloWorld.RData')

Running from Command prompt: R CMD BATCH script.R

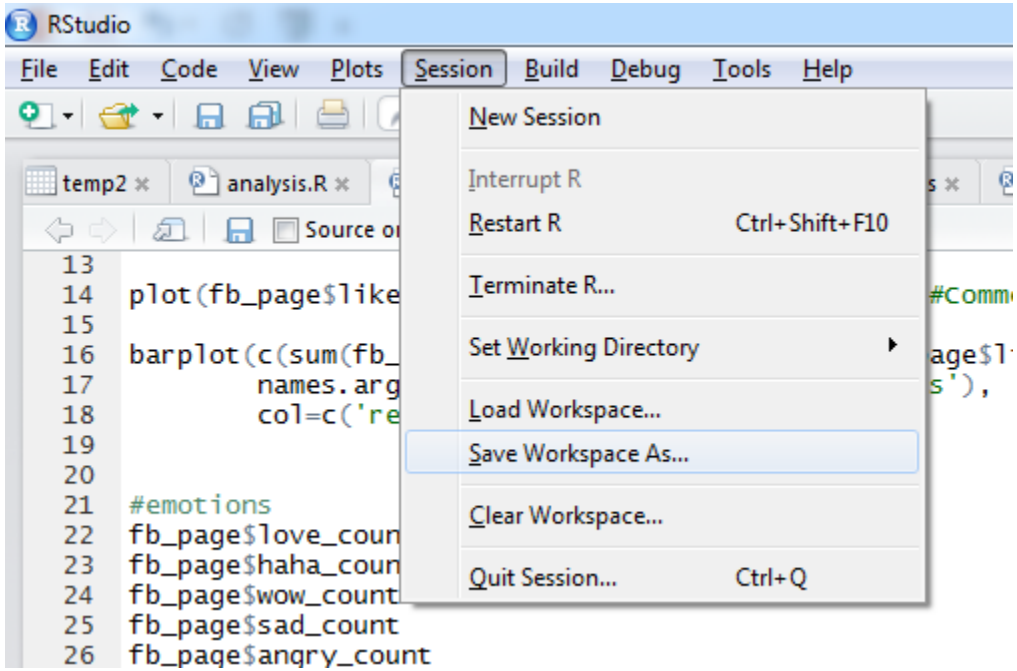
Running R via command Line



Running R via command Line

- Why?
  - Good for automation and running on Clusters such as USF– CIRCE (http://www.usf.edu/it/research-computing/)
- Command
  - "C:\Program Files\R\R-3.3.1\bin\R.exe" CMD BATCH script.R
  - Creates an output file: script.Rout

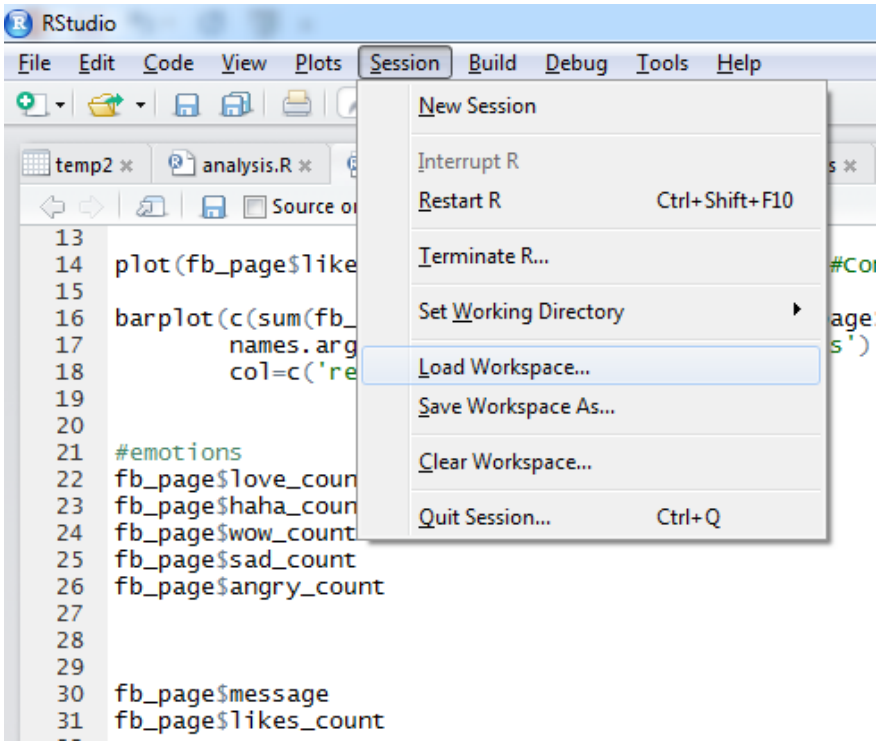
Save R- workspace



save.image("C:/Users/ThinkPad/Box Sync/R workshop/helloworld.RData")

*Keep one workspace for each project

Load R– workspace

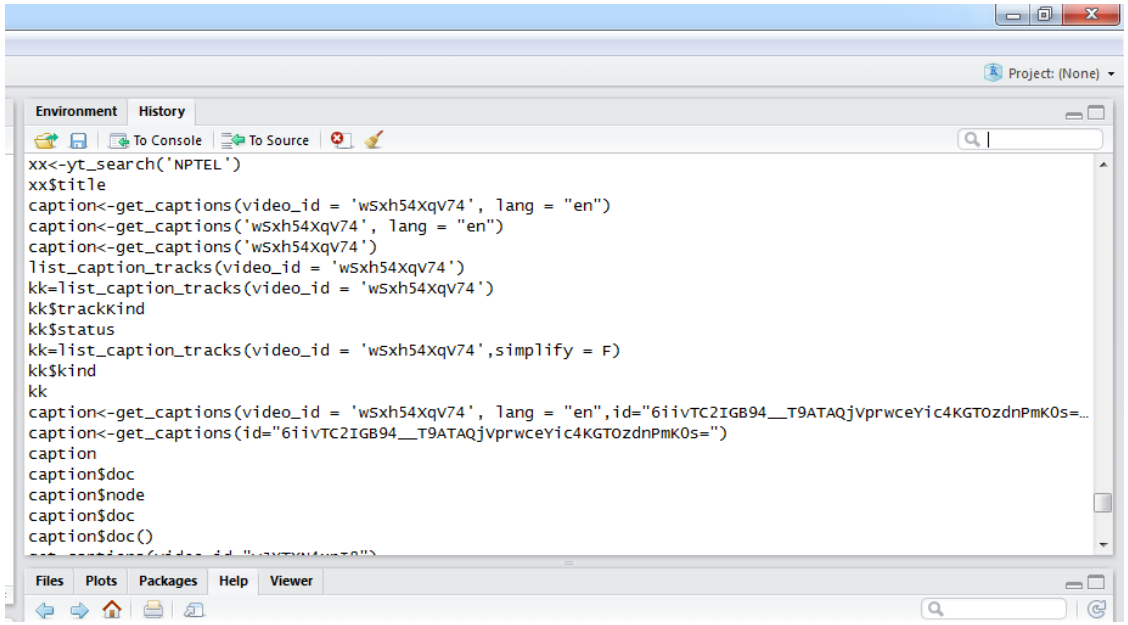


load("C:/Users/ThinkPad/Box Sync/R workshop/helloworld.RData")

https://github.com/vivek14632/R-Programming-workshop/tree/master/workspace

History of commands

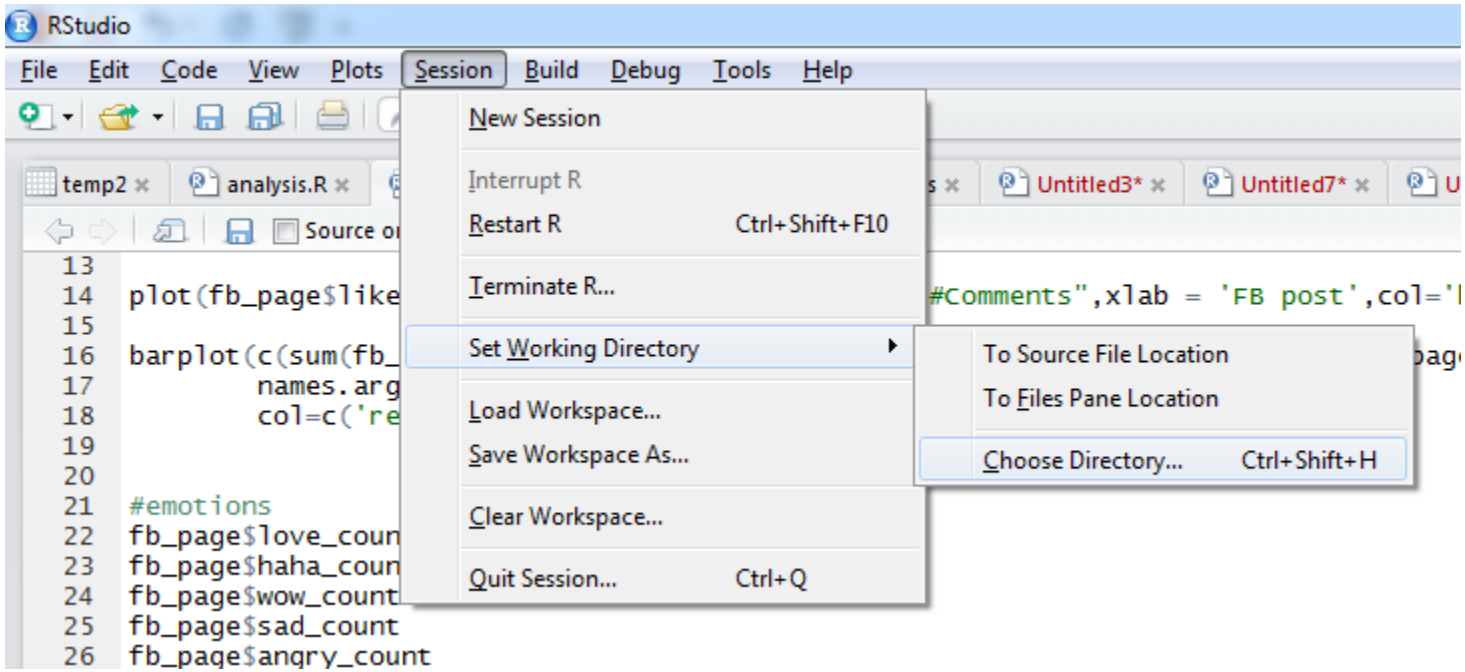
- history()



R–working directory

- Check working directory
  - getwd()
- Set working directory
  - setwd('path_to_directory')

R–working directory



setwd("C:/Users/ThinkPad/Box Sync/R workshop/code")

Debug R program

- print () function
- Line by line execution
- Breakpoints
  - Using GUI
  - browser() function

5 minutes break!!

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

R Objects

- Data types: Numeric, Integer, Logical, Complex, Character, Raw
- Different types of Objects
  - Vector
    - Set of elements of same mode: logical, numeric (integer, double), complex, character, list
  - Matrix
    - Rows and columns of same mode: logical, numeric (integer or double), complex or character.
  - Data frame: Similar to matrix but the columns can be of different modes
  - List: generalization of vector with a collection of data objects
- Class of an Object
- Example code: https://github.com/vivek14632/R-Programming-workshop/tree/master/ObjectsInR

Vectors

- Different types of vectors
  - Numeric vectors
  - Character vectors
  - Logical vector
  - Complex vector

https://github.com/vivek14632/R-Programming-workshop/blob/master/ObjectsInR/vectors.R

Vectors

- Creating a vector
  - c() function
  - seq() function
  - rep() function
  - : operator
  - Creating vector at run-time
- Even a single value is a vector
- Vector repetition: benefits and challenges

https://github.com/vivek14632/R-Programming-workshop/blob/master/ObjectsInR/vectors.R

Matrix

- Converting vector to matrix
  - dim() function
    - Creates matrix by column
    - Can also convert matrix to vector
  - matrix() function
    - matrix (vectorName, #rows,#columns)
    - By row: matrix (vectorName, #rows,#columns, byRow=T)
- rbind() function
- cbind() function

https://github.com/vivek14632/R-Programming-workshop/blob/master/ObjectsInR/matrix.R

Data Frame

- Similar to matrix
- Contains data columns with different modes – character, numeric, logical,etc.
- Convert matrix to data frame
  - data.frame()
- Columns names: names()

https://github.com/vivek14632/R-Programming-workshop/blob/master/ObjectsInR/dataFrame.R

List

- Combination of vector, matrix, data frames with different data types
- Used for storing different forms of output and return it from a function
- Display the output

https://github.com/vivek14632/R-Programming-workshop/blob/master/ObjectsInR/list.R

Functions

- Using library functions
- User defined functions
- Checking function definitions
- Modifying library functions (optional)

Example code: https://github.com/vivek14632/R-Programming-workshop/tree/master/functions

Loops

- Different types of loops
  - For loop
  - Repeat loop
  - While loop
- 'For' and 'While' loop is most widely used.

https://github.com/vivek14632/R-Programming-workshop/tree/master/loops

Conditionals

- If condition
- Else condition
- Else if condition
- Ifelse condition
- https://github.com/vivek14632/R-Programming-workshop/blob/master/conditionals/conditionals.R

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

Different file formats

- Clipboard
- CSV – read and write
- JSON format
- XLSX format
- User inputs via command line
- Text file
- System directory

Example code: https://github.com/vivek14632/R-Programming-workshop/tree/master/fileIO

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

Packages in R

- Installation
- Loading
- Updating packages
- Uninstall packages
- Example code:
  - https://github.com/vivek14632/R-Programming-workshop/blob/master/package/packages.R

5 minutes break!!

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

Data manipulation functions

- Table
- Subset
- Split
- Sort
- cbind()
- rbind()
- date and time
- apply

Example code: https://github.com/vivek14632/R-Programming-workshop/tree/master/dataManipulation

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

DATABASE CONNECTION

MySQL database and R

- Package: RMySQL
- Steps
  - Install package
  - Load package
  - create database connection
  - execute SQL query
- https://github.com/vivek14632/R-Programming-workshop/blob/master/databaseConnection/mysql.R

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

VISUALIZATION

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

Visualization

- Useful functions for graphs
- plot() – frequently used function for plotting
- xyplot()
- legend() – adding legend
- points() – addling points to an existing plot
- lines() – adding lines to an existing plot

Plot() basic parameters

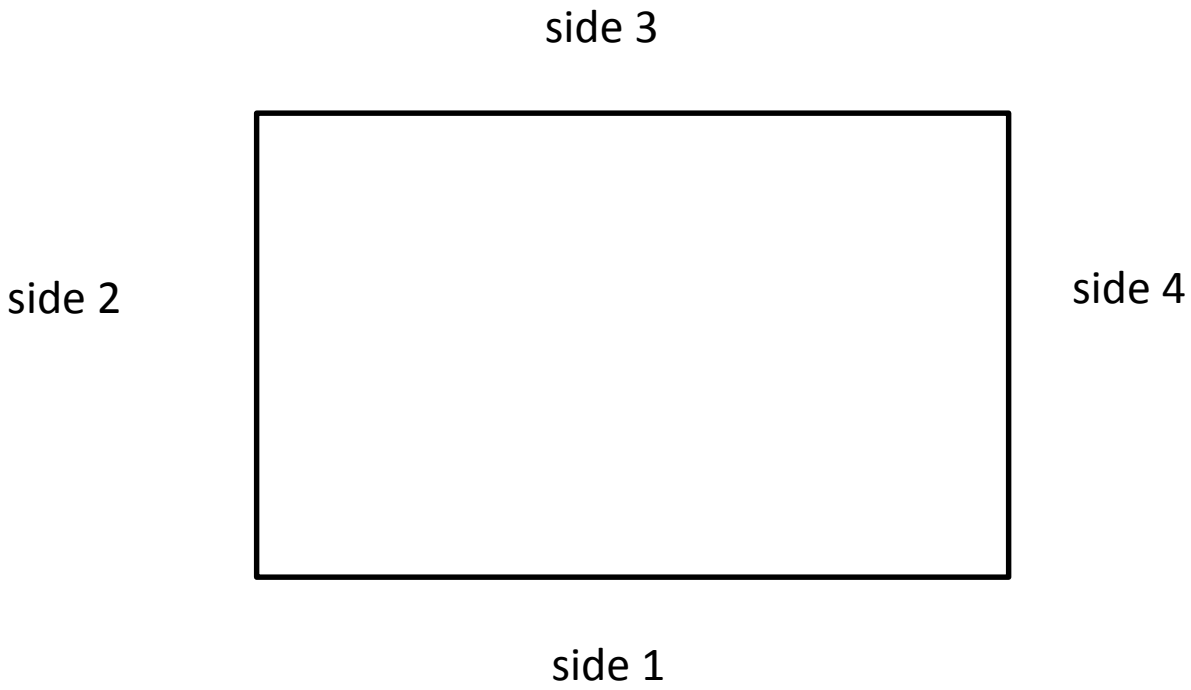
- type
  - 'l' → line
  - 'p' → point
  - 'b' → both line and point
- ylab →"label of Y-axis"
- xlab → "label of X-axis"
- xlim → range of X-axis
  - c(lower Value, Upper value)
- ylim → "range of Y-axis"
  - c(lower Value, Upper value)
- main → "title of the plot"

Plot() parameters

- pch → type of character used in the point plots
- lty→ line types
- col→ color of lines and points
- bty→ type of the box to enclose the graph
- lab→ change axis scale
- Please check the URL for different forms of points, lines, and colors
  - http://www.statmethods.net/advgraphs/parameters.html

axis() function

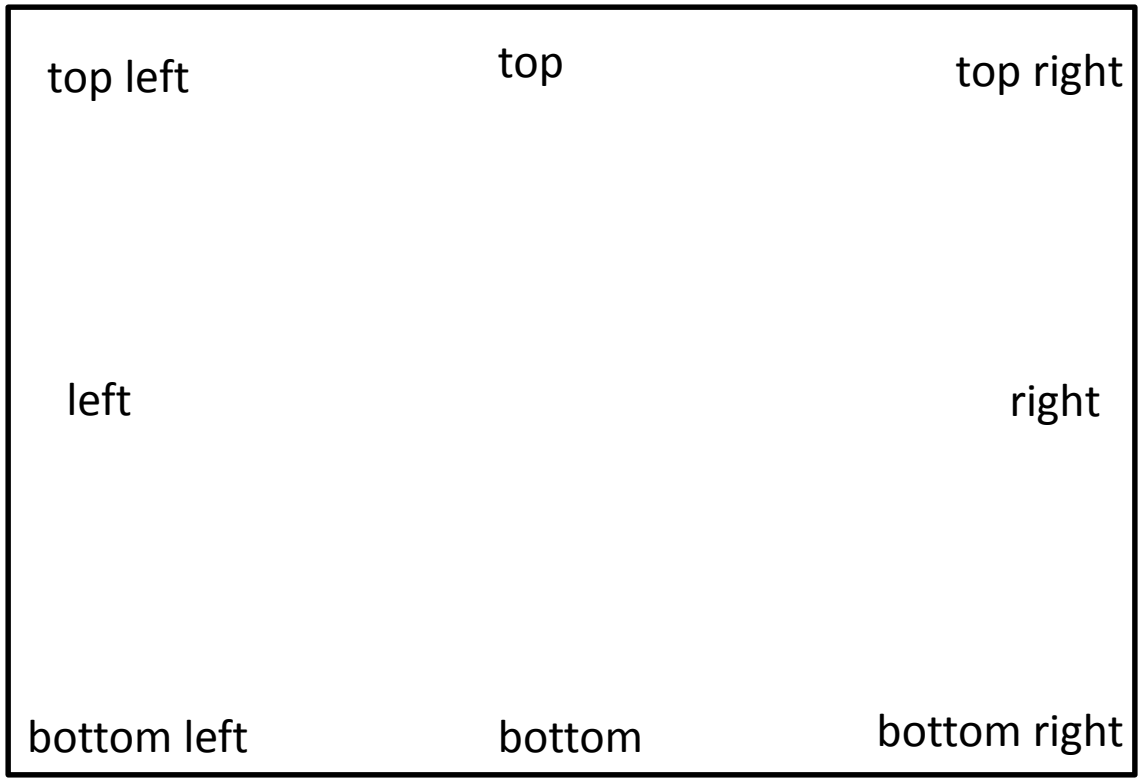
- Add axis to an existing plot



axis() parameters

- side → side number to add axis
  - select one of the values {1,2,3,4}
- labels → whether to show label on the axis or not
  - select one of the values {T,F}
- tick → whether to add tick to the axis or not
  - select one of the values {T,F}
- line → distance between label and graph
  - any real number preferably between [0,1]
- pos→ shift position of the axis

legend () function



legend() function

- legend('topright',**c('Minimum price','Maximum Price')**,pch = c(1,0),col=c('black','red'),lty = c(2,3))
- c('Minimum price','Maximum Price') → variable names

Multiple plots

- par(mfrow=c(number of rows,number of columns))
  - Example: par(mfrow=c(2,3))
  - Appears by row
- par(mfcol=c(number of rows,number of columns))
  - Appears by columns

Types of plots

- Generic Plot or Plot
- Density plot
- Histogram
- Geographical Map
- QQ plot
- Time series plots

https://github.com/vivek14632/R-Programming-workshop/tree/master/Visualization

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

DATASETS

Accessing datasets in R

>install.packages('MASS')

>library('MASS')

#dataset

>quine

Other R dataset packages

(1) datasets

http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

https://github.com/vivek14632/R-Programming-workshop/tree/master/datasets

USF MUMA COLLEGE OF BUSINESS UNIVERSITY OF SOUTH FLORIDA

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

USF MUMA
COLLEGE OF BUSINESS
UNIVERSITY OF SOUTH FLORIDA

R AND SOCIAL MEDIA

Social Media APIs

- R Packages
  - Facebook API – Rfacebook
  - Twitter API – twitterR
  - Youtube API – tuber
  - Google Trends API – gtrendsR
- Steps
  - Authentication using Oauth
  - Use API functions
- https://github.com/vivek14632/R-Programming-workshop/tree/master/socialMedia

Outline

- Motivation – Why learn R?
- Programming IDE – R studio, workspace, Console
- Programming – objects, loops, conditionals, function
- File Input/output
- R– Packages
- Data manipulation
- Database connector – MySQL
- Visualization
- Datasets
- Social Media APIs – Twitter, Facebook, Google Trends, Youtube
- CIRCE @USF (Cluster Computing)

Working with CIRCE

- SLURM Job scheduler
- R script
- Submission script
- Important command

Submission script

```
#!/bin/bash
#
#SBATCH --comment=r-test
#SBATCH --ntasks=4
#SBATCH --job-name=r-test
#SBATCH --output=output.%j.r-test
#SBATCH --time=01:00:00

#### SLURM 4 processor R test to run for 1 hour.

module purge
module add apps/R/3.1.2

mpirun Rmpi test.R
```

Important commands

- sbatch: Submit jobs to SLURM
- squeue: Check your job status
- scancel: cancel your job

https://github.com/vivek14632/R-Programming-workshop/tree/master/CIRCE

Optional

- Distribution
  - Working with standard distribution such as Normal, Poisson, and Uniform.

- Financial data
  - Quantmod library
  - https://github.com/vivek14632/R-Programming-workshop/blob/master/qunatitativeTrading/gettingData.R