You Ready?

Applied Statistics

Dr Tao Zou

ANU – RSFAS

Last Updated: 26/07/2017

What is Statistics

- Statistics is the science of learning from data … (ASA: http://www.amstat.org/careers/whatisstatistics.cfm.)
- Statisticians collect and **analyse data**, then calculate results using **a specific design** (model). They **draw conclusions and make decisions** in the face of **uncertainty**. (ASA: http://www.amstat.org/careers/careersinstatisticspresentation.cfm.)
- This course focuses on: data type ⟷ statistical modelling.

What are the data and modelling I used?

- PM2.5 concentration and weather data in Beijing ⟷ nonparametric regression model (Environmental Science).

S ⟷

⟷

Course Description

Data Type	Model	Period
Continuous Data	Simple Linear Regression; Multiple Linear Regression.	Week 1-6
Discrete Data 1. Categorical Data	Nominal: Logistic Regression.	Week 7-12
	Ordinal: Logistic Regression.	
2. Count Data	Binomial: Logistic Regression.	
	Poisson: Log-Linear Regression.	
Multivariate Data	Multivariate Analysis.	

Bootstrap: one powerful tool to provide the inferences for all the above models.

Questions & Answers

- Due to the large number of the enrollment, students are **strongly encouraged** to **ask questions during the lectures, the tutorials and the consultation hours**, instead of in emails.
- The lecturer will leave several minutes at the start and the end of each lecture for Q&A.
- Questions & Answers forum has been opened on Wattle. Before you post your questions at the forum, please check the past Q&A first.
- The tutors and the lecturer will monitor the forum on Wattle.
- For those students who send emails for Q&A, their questions may be copied to the forum for everyone if the questions are not private. And answers will be given in the forum.

Recommended Textbook

- "The Statistical Sleuth" by F. L. Ramsey and D. W. Schafer. This book is available in the University bookstore.

Assessments

Assessment Task	Value	Due Date	Date for Return of Assessment	Linked Learning Outcomes
1. Quiz (online)	10% (compulsory but redeemable)	12:00 pm, Wed, Week 5	The week after submission.	1 through 3.
2. Assignment 1	10% (optional and redeemable)	12:00 pm, Wed, Week 8	The week after submission.	1 through 5.
3. Assignment 2	10% (optional and redeemable)	12:00 pm, Wed, Week 11	The week after submission.	1 through 7.
4. Final exam	70%	TBA		1 through 7.

- Due to the redeemable nature of the quiz and assignments, late submission will not be accepted without appropriate documentation.
- The meaning of "redeemable" will be explained in the following R project.

Quiz

- This quiz is compulsory, and is to be attempted online on Wattle individually.
- Announcements will be made during lectures and on Wattle site regarding the availability of the quiz.
- This quiz will require the use of R to analyse real data and there will be a mix of multiple choice questions and numerical evaluation questions.

Assignments

- 2 take-home problem sets.
- Students should individually attempt all the questions, show appropriate interpretation and computation details, summarise and report on the findings of the analysis, as well as discuss results.
- Assignments are required to be typed and contain relevant R code and graphics.
- Hard Copy Submission: Assignments are submitted via the physical assignment box in front of the admin office at level 4, CBE Building (26C). The cover sheet must use the assignment cover sheet template. Assignments must include the cover sheet available on Wattle site.

Final Exam

- Final: 15 minutes reading time and 3 hours writing time.

The permitted material for the final exam will be:

- Calculator (non-programmable).
- Unannotated paper-based dictionary (no approval required).
- Two A4 pages with notes on both sides.

Computing

- R: http://www.r-project.org.
- No prior knowledge of R is assumed in this course.
- "The Undergraduate Guide to R" by T. Martin.

https://sites.google.com/site/undergraduateguidetor/manual-files.

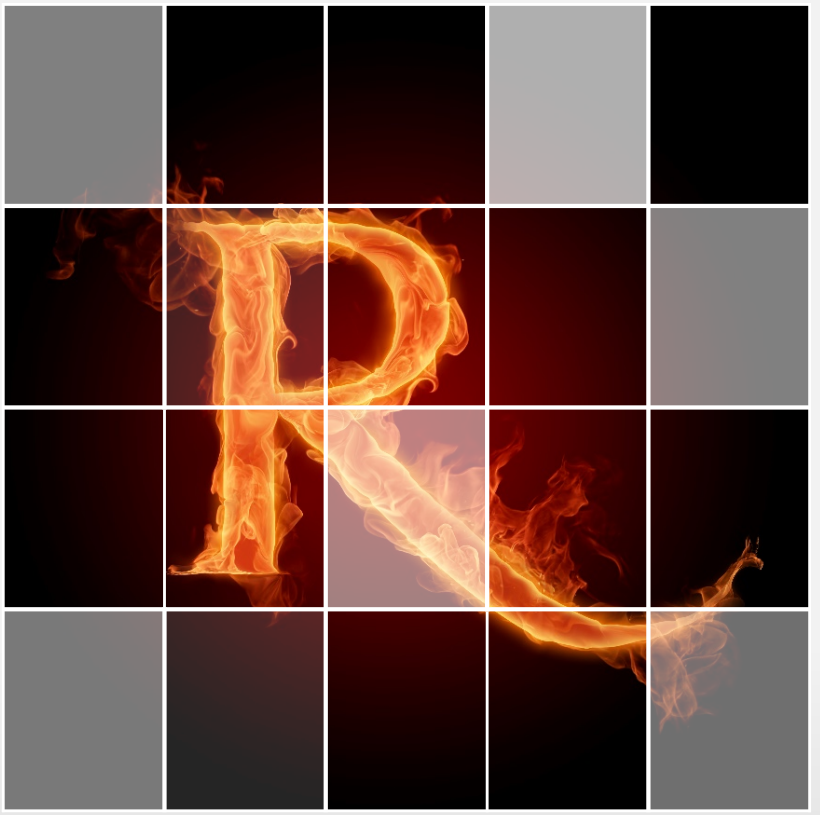
- RStudio (http://www.rstudio.com) is also an excellent user interface for R. You can download and install RStudio but it will not be introduced during class.

This Course ≠ R

- The ultimate goal of this course is to implement appropriate statistical modelling given different types of real data.
- R is just a tool to realise the statistical computing when statistical models have been built.
- The more important thing in this course is to understand statistical concepts. That is the only way to perform statistical analysis properly.

You R Ready!

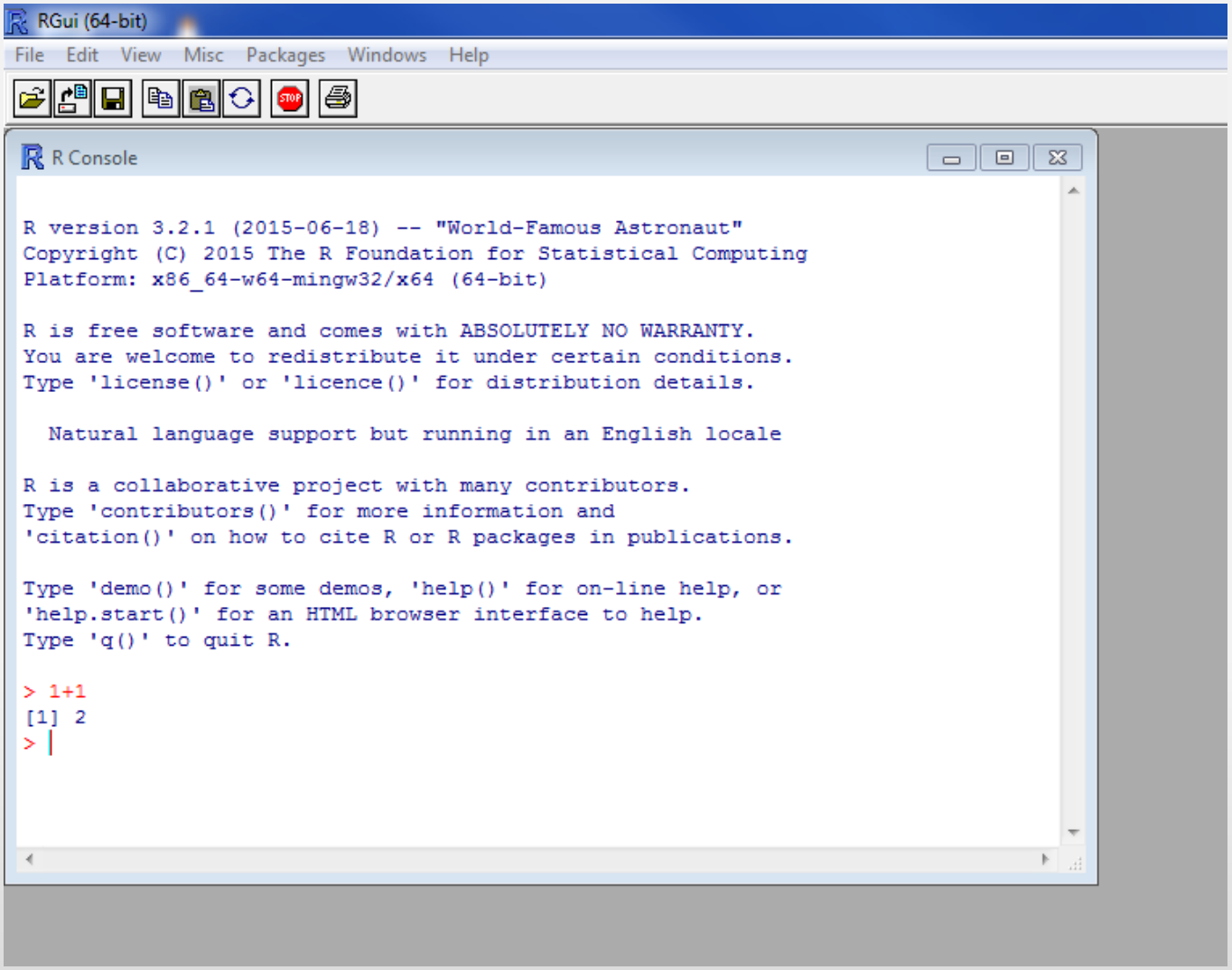
Why R?



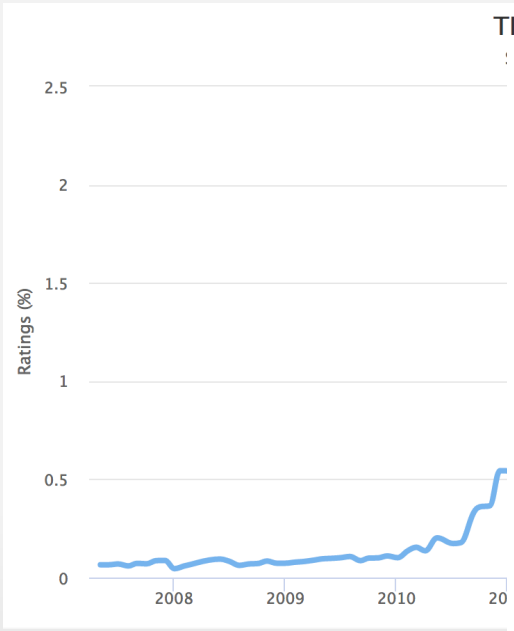
Creation of R

Ross Ihaka

Robert Gentleman


RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
R version 3.2.1 (2015-06-18) -- "World-Famous Astronaut"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-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.
> 1+1
[1] 2
>

Why R ?



Nov 2014	Nov 2013	Change	Programming Language	Ratings	Change
1	1		C	17.469%	-0.69%
2	2		Java	14.391%	-2.13%
3	3		Objective-C	9.063%	-0.34%
4	4		C++	6.098%	-2.27%
5	5		C#	4.985%	-1.04%
6	6		PHP	3.043%	-2.34%
7	8	^	Python	2.589%	-0.52%
8	10	^	JavaScript	2.088%	+0.04%
9	12	^	Perl	2.073%	+0.55%
10	11	^	Visual Basic .NET	2.061%	+0.09%
11	-	^^	Visual Basic	1.657%	+1.66%
12	31	^^	R	1.548%	+1.14%
13	9	vv	Transact-SQL	1.408%	-1.11%
14	13	v	Ruby	1.211%	-0.09%
15	17	^	Delphi/Object Pascal	0.957%	+0.31%

http://www.tiobe.com/tiobe_index

R in Industry

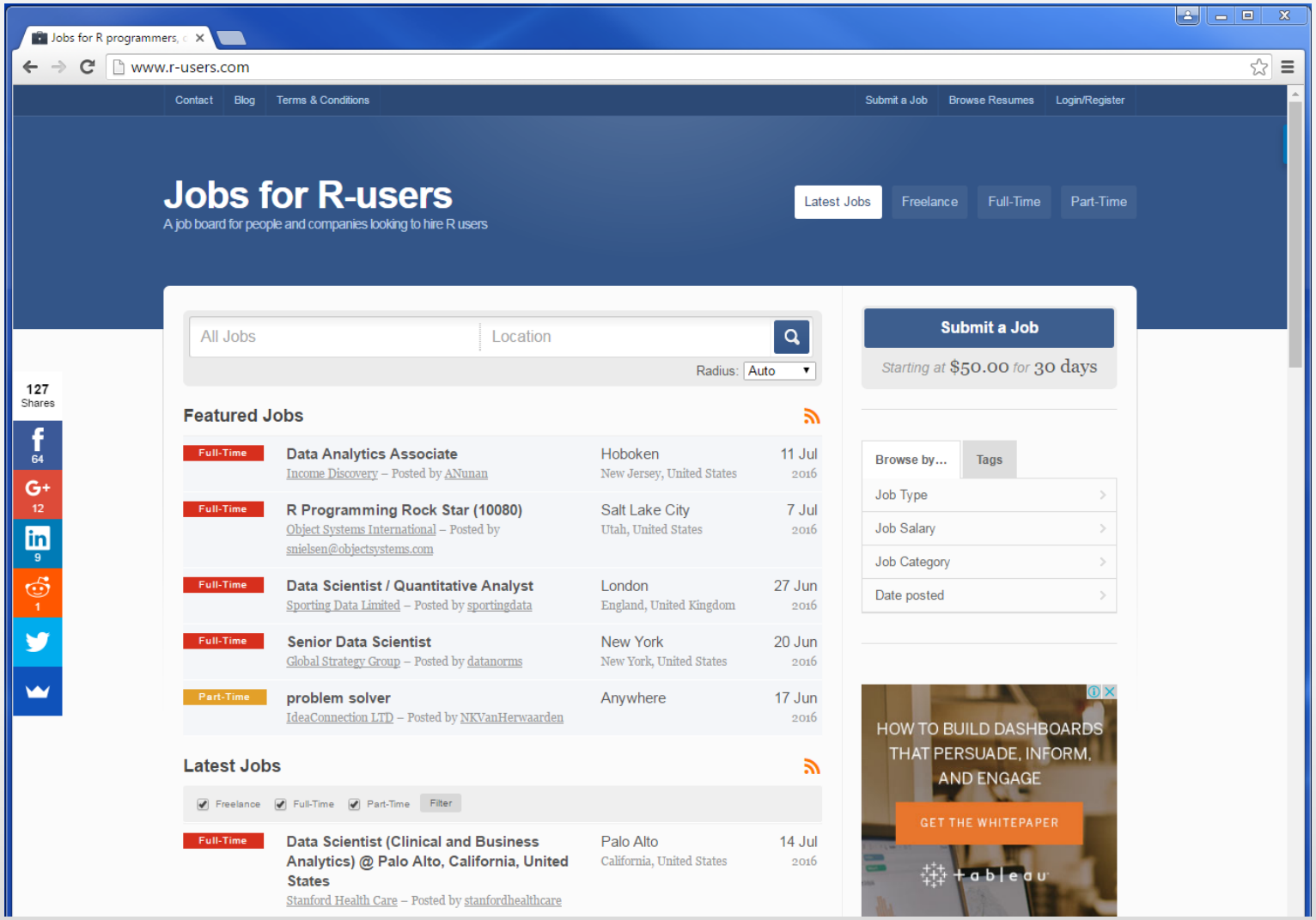
REVOLUTION ANALYTICS

Type	Subsidiary
Industry	Statistical software
Predecessor	Revolution Computing
Founded	2007
Headquarters	Mountain View, CA, United States
Key people	David Rich, CEO
Products	Revolution R
Revenue	8-11 Million in 2009
Owner	Microsoft[1]
Parent	Microsoft
Website	revolutionanalytics.com

REVOLUTION ANALYTICS

+

Microsoft


Jobs for R programmers,
www.r-users.com
Contact
Blog
Terms & Conditions
Submit a Job
Browse Resumes
Login/Register
Jobs for R-users
A job board for people and companies looking to hire R users
Latest Jobs
Freelance
Full-Time
Part-Time
All Jobs
Location
Radius: Auto
Submit a Job
Starting at $50.00 for 30 days
127 Shares
64
12
9
1
Featured Jobs
Full-Time
Data Analytics Associate
Income Discovery – Posted by ANunan
Hoboken
New Jersey, United States
11 Jul 2016
Full-Time
R Programming Rock Star (10080)
Object Systems International – Posted by snielsen@objectsystems.com
Salt Lake City
Utah, United States
7 Jul 2016
Full-Time
Data Scientist / Quantitative Analyst
Sporting Data Limited – Posted by sportingdata
London
England, United Kingdom
27 Jun 2016
Full-Time
Senior Data Scientist
Global Strategy Group – Posted by datanorms
New York
New York, United States
20 Jun 2016
Part-Time
problem solver
IdeaConnection LTD – Posted by NKVanHerwaarden
Anywhere
17 Jun 2016
Latest Jobs
Freelance
Full-Time
Part-Time
Filter
Full-Time
Data Scientist (Clinical and Business Analytics) @ Palo Alto, California, United States
Stanford Health Care – Posted by stanfordhealthcare
Palo Alto
California, United States
14 Jul 2016
Browse by...
Tags
Job Type
Job Salary
Job Category
Date posted
HOW TO BUILD DASHBOARDS THAT PERSUADE, INFORM, AND ENGAGE
GET THE WHITEPAPER
tableau

R in China

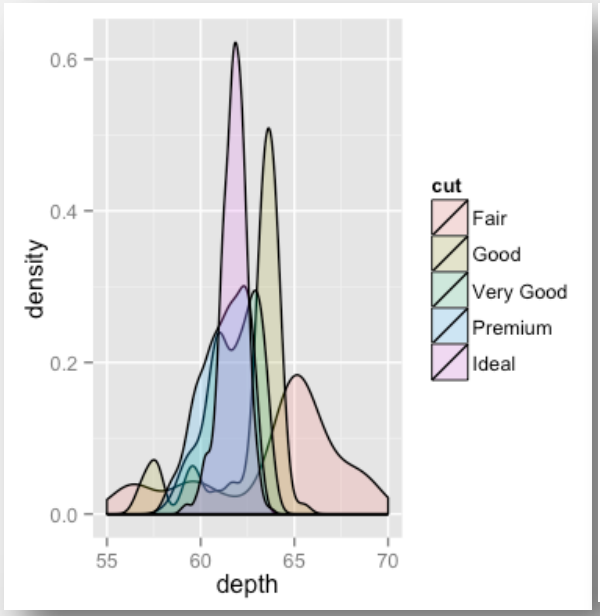
What can R be used for?

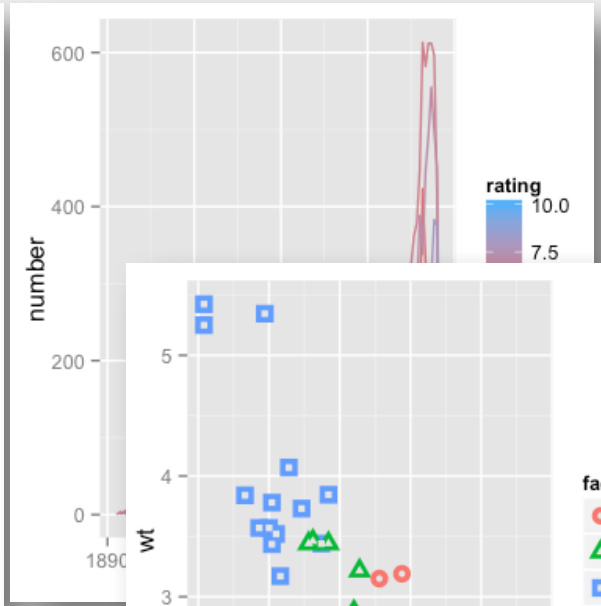
- Covering a wide variety of statistical and graphical techniques (more than 1,1054 additional packages by now).

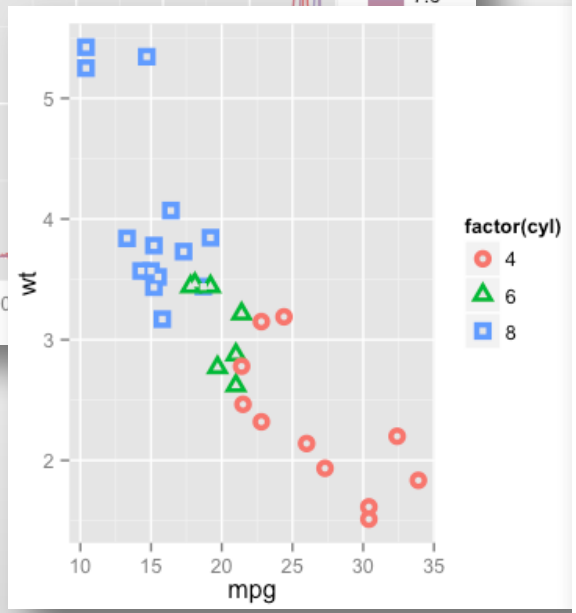
- Open source, but strong community.

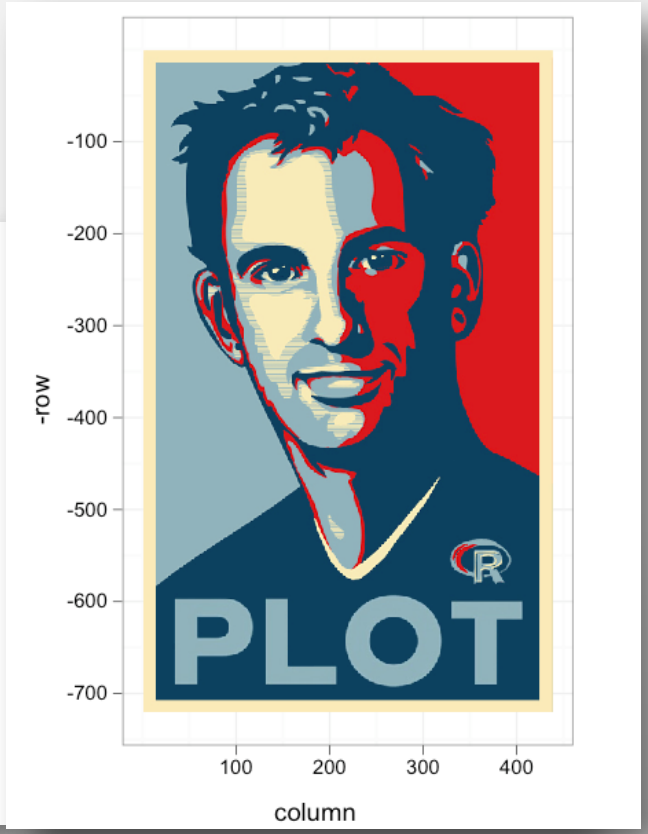
- Extensible:
  - R is supported by MySQL, Hadoop …
  - Python, C … can be linked and called in R.

What can R be used for?

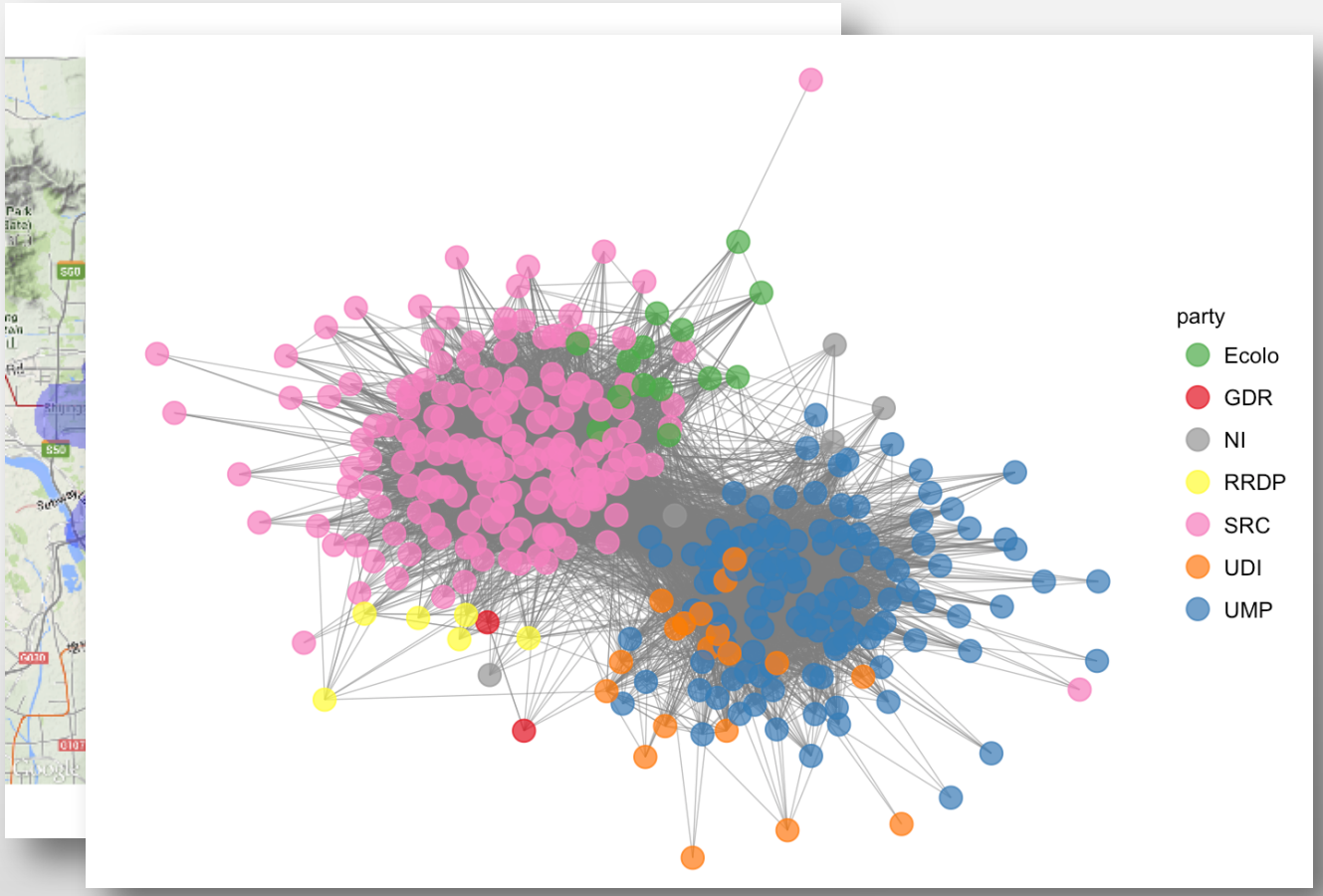








What can R be used for?

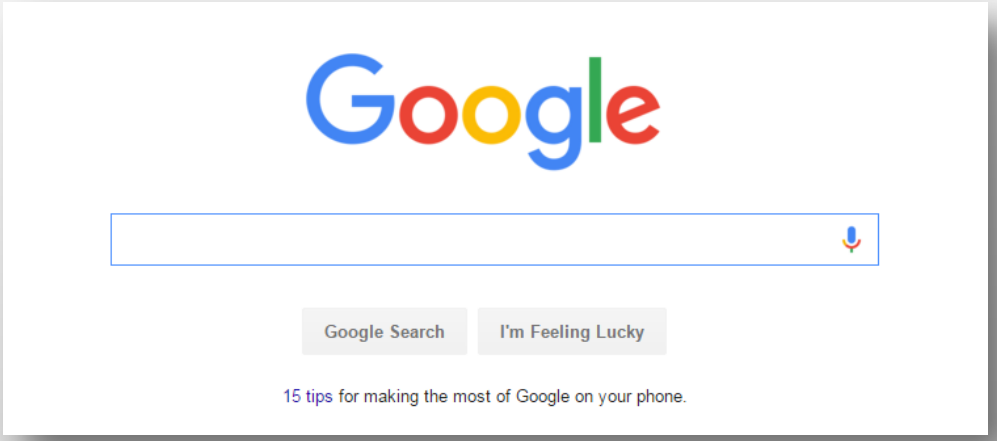


What can R be used for?

- Play 2048 using R!

Thanks to References

- CluBear (WeChat Group)
- Capital of Statistics



Data Types and Languages in R

Data Types in R

- Three basic types:
  - numeric
  - character
  - logical (TRUE / FALSE)

```
> y1<-1
> y1
[1] 1
> class(y1)
[1] "numeric"
>
> y2='1'
> y2
[1] "1"
> class(y2)
[1] "character"
>
> y3='one'
> y3
[1] "one"
> class(y3)
[1] "character"
```

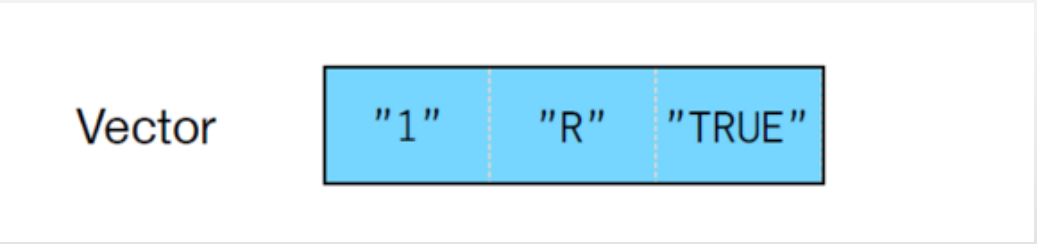
```
> y4 <- 3 == 4
> y4
[1] FALSE
> class(y4)
[1] "logical"
>
> y5 <- "a"=="a"
> y5
[1] TRUE
> class(y5)
[1] "logical"
```

Data Types in R

- Vector

```
> vector1=c(1,2,3,4)
> vector2<-c('a','b','c','d')
> length(vector1)
[1] 4
>
> vector3=1:4
> vector2[3]
[1] "c"
>
> vector2[c(2,3)]
[1] "b" "c"
> vector2[c(2,3,1,1,3)]
[1] "b" "c" "a" "a" "c"
>
> which(vector2=='d')
[1] 4
> vector2[c(1,3)]
[1] "a" "c"
>
> seq(1,10,by=2)
[1] 1 3 5 7 9
```

```
> vector4=c(1,'R',TRUE)
> class(vector4)
[1] "character"
```



Data Types in R

- Some functions for vectors: min max mean quantile sort

```
> aa=c(3,5,10,4.5,-1,-3)
> bb=c(1,1,1,2,3,3,1,2,4,1,2,4,4,2,3,4,1,2,3,4)
> cc=letters[bb]
> letters
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q" "r" "s"
[20] "t" "u" "v" "w" "x" "y" "z"
> cc
 [1] "a" "a" "a" "b" "c" "c" "a" "b" "d" "a" "b" "d" "d" "b" "c" "d" "a" "b" "c"
[20] "d"
>
> min(aa)
[1] -3
> max(aa)
[1] 10
> median(aa)
[1] 3.75
> quantile(aa)
    0%    25%    50%    75%   100%
-3.000  0.000  3.750  4.875 10.000
>
> sort(aa)
[1] -3.0 -1.0  3.0  4.5  5.0 10.0
> sort(aa,decreasing=TRUE)
[1] 10.0  5.0  4.5  3.0 -1.0 -3.0
```

Data Types in R

- Matrix: cbind() rbind()

```
> cbind(1:3,2:4)
     [,1] [,2]
[1,]    1    2
[2,]    2    3
[3,]    3    4
> rbind(1:3,2:4)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    3    4
>
> M=matrix(1:9,nrow=3,ncol=3)
>
> M[2,3]
[1] 8
> M[c(1,2),c(2,3)]
     [,1] [,2]
[1,]    4    7
[2,]    5    8
> M[,3]
[1] 7 8 9
```

Data Types in R

Data frame: different data types can be in different columns.

data={ “x” : c(1, … ,10);

“y” : c(‘A’ , … , ‘J’)}

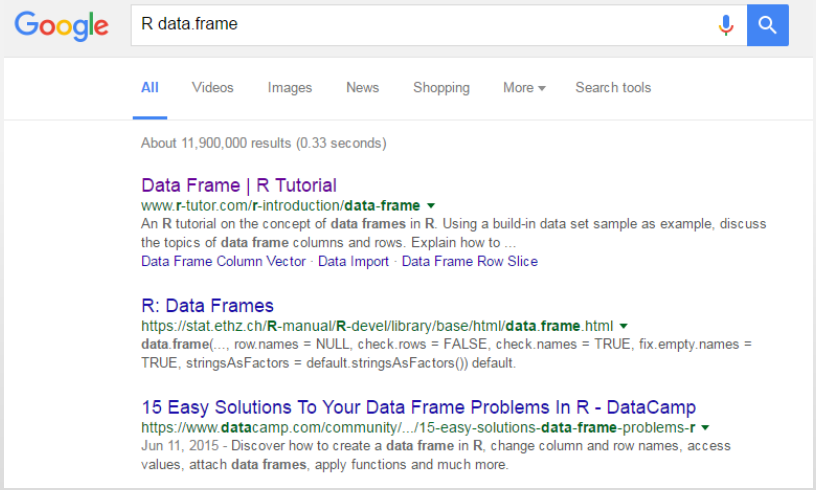
```
> LETTERS
 [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N"
[15] "O" "P" "Q" "R" "S" "T" "U" "V" "W" "X" "Y" "Z"
> data=data.frame(x=1:10,y=LETTERS[1:10])
> data[,1]
 [1]  1  2  3  4  5  6  7  8  9 10
> data$x
 [1]  1  2  3  4  5  6  7  8  9 10
> class(LETTERS[1:10])
[1] "character"
> data$y
 [1] A B C D E F G H I J
Levels: A B C D E F G H I J
> class(data$y)
[1] "factor"
>
> summary(data)
       x               y
 Min.   : 1.00   A      :1
 1st Qu.: 3.25   B      :1
 Median : 5.50   C      :1
 Mean   : 5.50   D      :1
 3rd Qu.: 7.75   E      :1
 Max.   :10.00   F      :1
                 (Other):4
```

How to get help in R

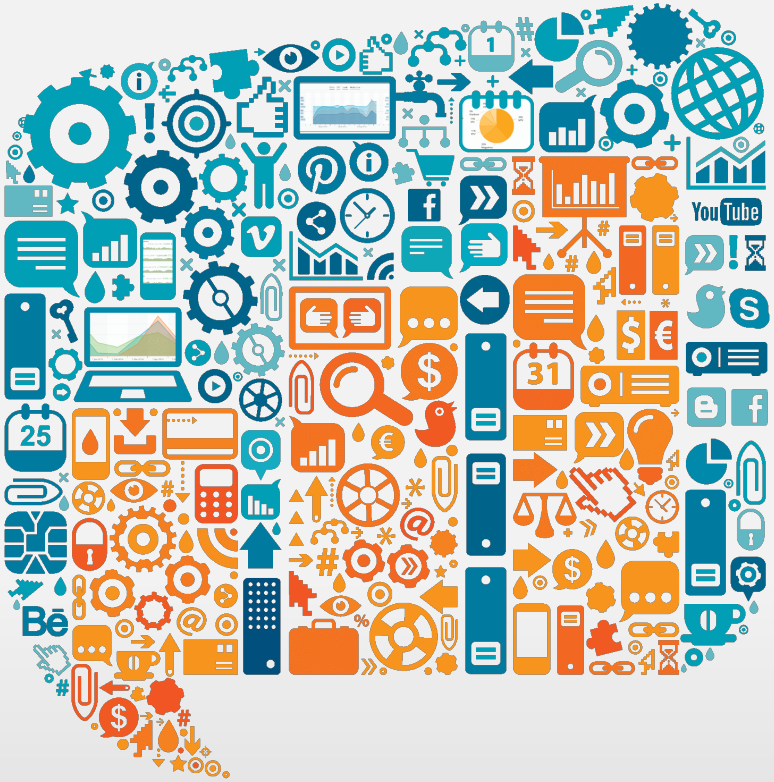
- ? or ??

```
> ?data.frame
> ??data.frame
```

- Google



R Project: Assessment Data



"The best way to learn how to program is by doing it." – Garry Tan

Read CSV

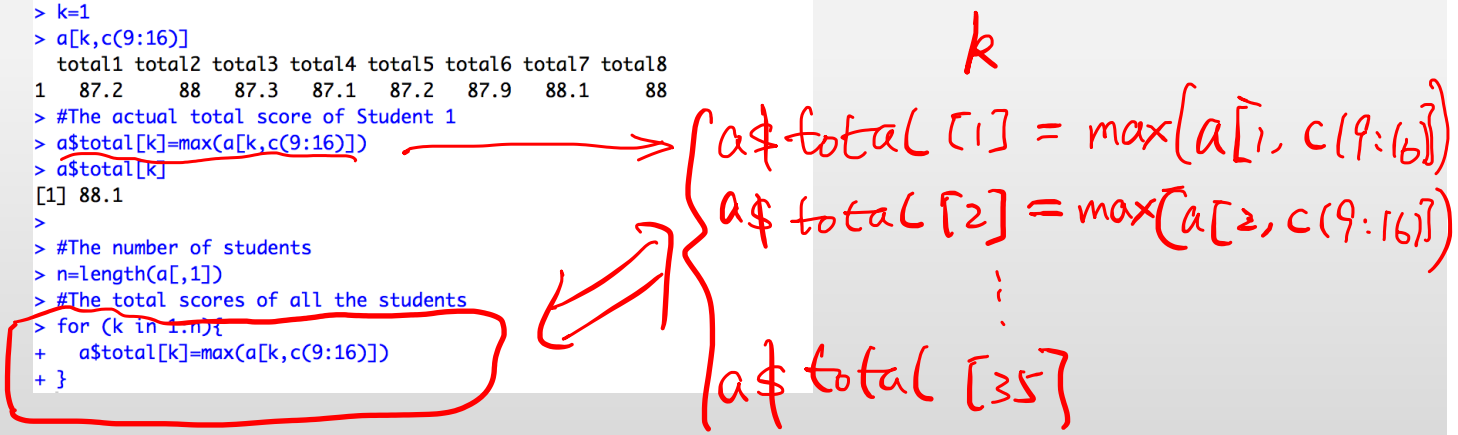
```
> a=read.csv(file='assessment.csv',header=T)
>
> head(a)
  X UniqueNo gender residence Quiz Assignment1 Assignment2 Final
1 1        1   Male     India   80          87          89    88
2 2        2   Male Indonesia   85          81          87    88
3 3        3   Male     Korea   90          99         100    94
4 4        4   Male  Mongolia   85          93          97    83
5 5        5 Female  Thailand   95          91          90    86
6 6        6 Female Hong Kong  100          93          82    95
> summary(a)
       X           UniqueNo       gender       residence       Quiz
 Min.   : 1.0   Min.   : 1.0   Female:11   Australia: 5   Min.   : 80.00
 1st Qu.: 9.5   1st Qu.: 9.5   Male  :24   Korea     : 5   1st Qu.: 82.50
 Median :18.0   Median :18.0               USA       : 5   Median : 90.00
 Mean   :18.0   Mean   :18.0               Canada    : 4   Mean   : 88.71
 3rd Qu.:26.5   3rd Qu.:26.5               China     : 4   3rd Qu.: 95.00
 Max.   :35.0   Max.   :35.0               India     : 2   Max.   :100.00
                                          (Other)   :10
  Assignment1     Assignment2        Final
 Min.   :80.00   Min.   : 82.0   Min.   : 79.00
 1st Qu.:87.00   1st Qu.: 88.0   1st Qu.: 86.00
 Median :90.00   Median : 90.0   Median : 88.00
 Mean   :89.43   Mean   : 90.6   Mean   : 89.86
 3rd Qu.:93.00   3rd Qu.: 94.0   3rd Qu.: 94.50
 Max.   :99.00   Max.   :100.0   Max.   :100.00
```

Redeemable for Three Assessments

- Normal case:
  - Total = Assessment 1 × 10% + Assessment 2 × 10% + Assessment 3 × 10% + Final × 70%.
- If you think the grade of Assessment 1 is too low:
  - Total = Assessment 2 × 10% + Assessment 3 × 10% + Final × 80%. (Drop Assessment 1.)
- If you think the grades of Assessments 1&2 are too low:
  - Total = Assessment 3 × 10% + Final × 90%. (Drop Assessments 1&2.)
- If you think the grades of Assessments 1&2&3 are all low:
  - Total = Final × 100%. (Drop Assessments 1&2&3.)

How to realise the above in R?

```
> #Drop nothing
> a$total1=a$Quiz*0.1+a$Assignment1*0.1+a$Assignment2*0.1+a$Final*0.7
>
> #Drop 1 assessment
> a$total2=a$Assignment1*0.1+a$Assignment2*0.1+a$Final*0.8
> a$total3=a$Quiz*0.1+a$Assignment2*0.1+a$Final*0.8
> a$total4=a$Quiz*0.1+a$Assignment1*0.1+a$Final*0.8
>
> #Drop 2 assessments
> a$total5=a$Quiz*0.1+a$Final*0.9
> a$total6=a$Assignment1*0.1+a$Final*0.9
> a$total7=a$Assignment2*0.1+a$Final*0.9
>
> #Drop 3 assessments
> a$total8=a$Final*1.0
>
> a$total=0
> #The 8 possible total scores of Student 1
> k=1
> a[k,c(9:16)]
  total1 total2 total3 total4 total5 total6 total7 total8
1   87.2     88   87.3   87.1   87.2   87.9   88.1     88
> #The actual total score of Student 1
> a$total[k]=max(a[k,c(9:16)])
> a$total[k]
[1] 88.1
>
> #The number of students
> n=length(a[,1])
> #The total scores of all the students
> for (k in 1:n){
+   a$total[k]=max(a[k,c(9:16)])
+ }
```



Write CSV

```
> head(a)
  X UniqueNo gender residence Quiz Assignment1 Assignment2 Final total1 total2
1 1        1   Male     India   80          87          89    88   87.2   88.0
2 2        2   Male Indonesia   85          81          87    88   86.9   87.2
3 3        3   Male     Korea   90          99         100    94   94.7   95.1
4 4        4   Male  Mongolia   85          93          97    83   85.6   85.4
5 5        5 Female  Thailand   95          91          90    86   87.8   86.9
6 6        6 Female Hong Kong  100          93          82    95   94.0   93.5
  total3 total4 total5 total6 total7 total8 total
1   87.3   87.1   87.2   87.9   88.1     88  88.1
2   87.6   87.0   87.7   87.3   87.9     88  88.0
3   94.2   94.1   93.6   94.5   94.6     94  95.1
4   84.6   84.2   83.2   84.0   84.4     83  85.6
5   87.3   87.4   86.9   86.5   86.4     86  87.8
6   94.2   95.3   95.5   94.8   93.7     95  95.5
> b=summary(a)
>
> write.csv(a,file='assessment_result.csv')
> write.csv(b,file='summary.csv')
```