STT481 Capstone in Statistics

Lecture 1: Introduction to R

Chih-Li Sung

09/04/2019 (modified from
http://datascienceandr.org/slide/RBasic-Introduction.html#1)

Agenda

- Basic R
- Application of R
- Install R and Rstudio
- How to learn R - swirl

What is R

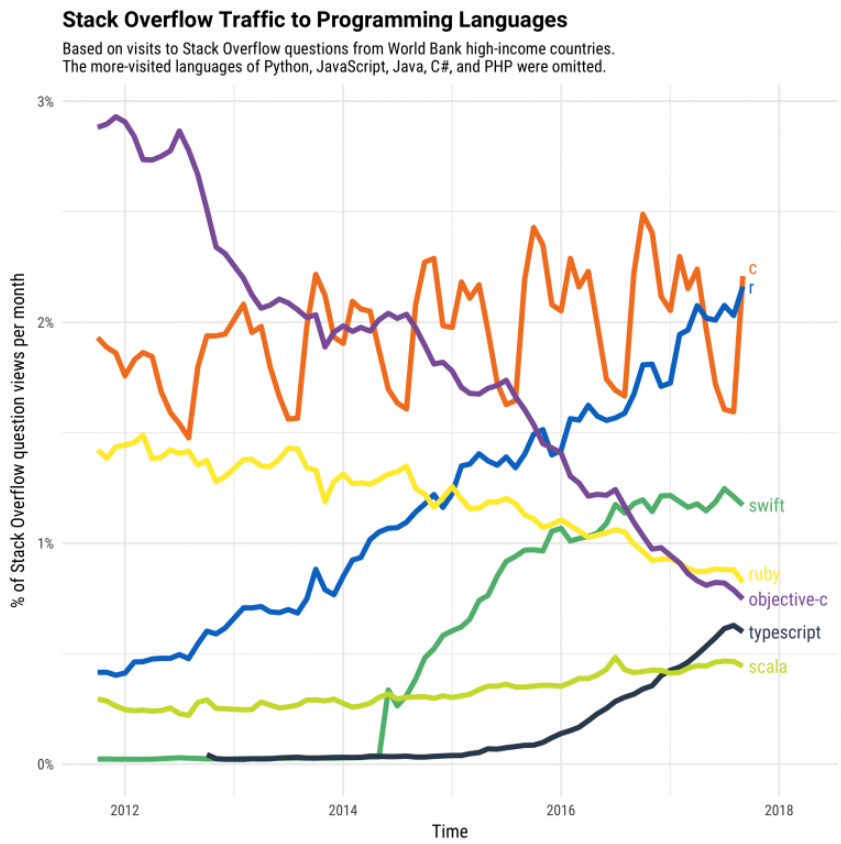
- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.
- R is **free**!
- R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand.
- The current group ("R Core Team") consists of professional statisticians all over the world.

source: link

R is popular

"R and Python are the two most popular programming languages used by data analysts and data scientists" — source: link

- Impressive Growth



R has lots of packages

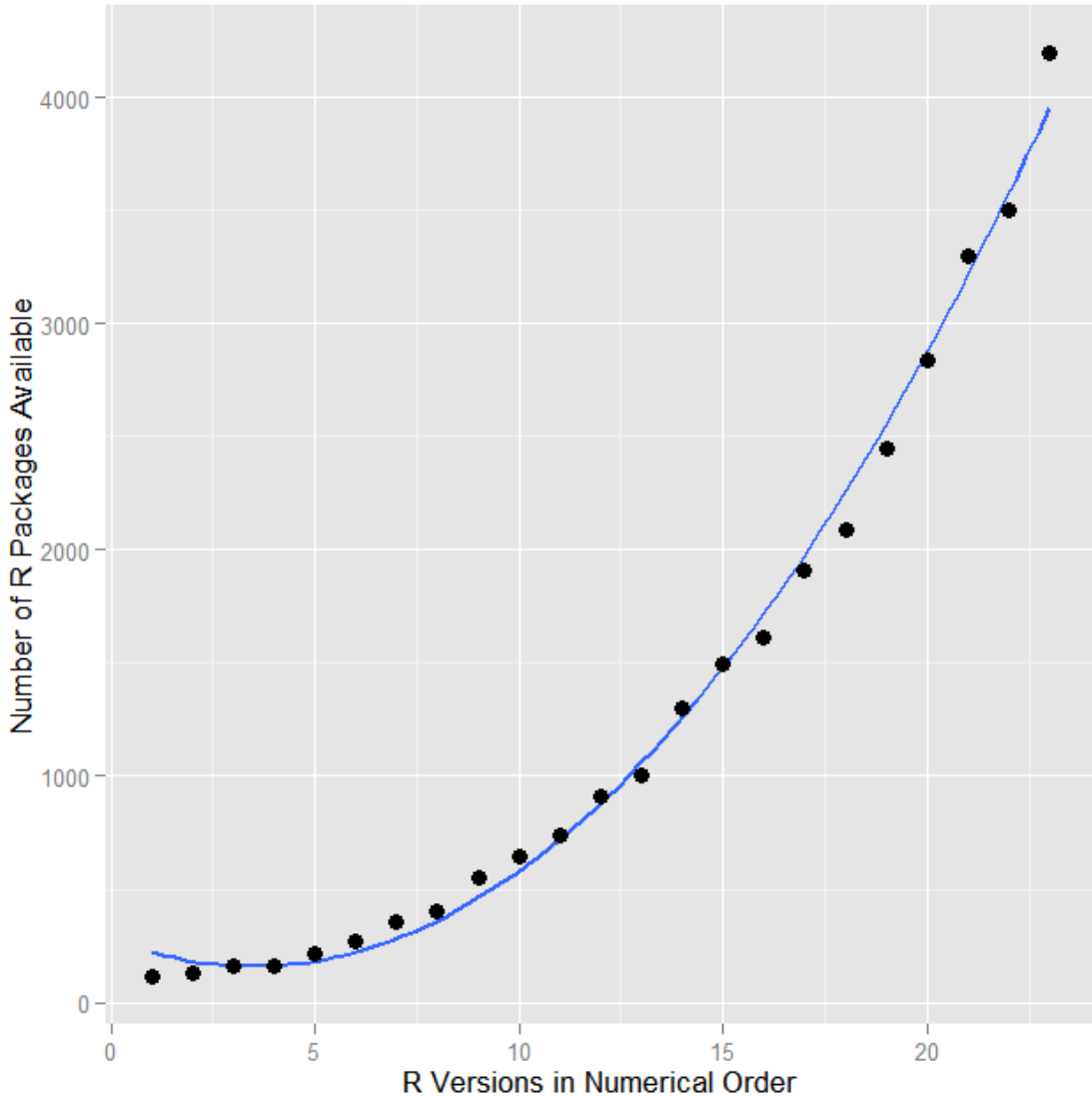


Figure 2:

Easy to integrate R with other languages

Figure 3:

Compared to other languages

- R has very advanced visualization tools
- R has advanced statistical analysis tools

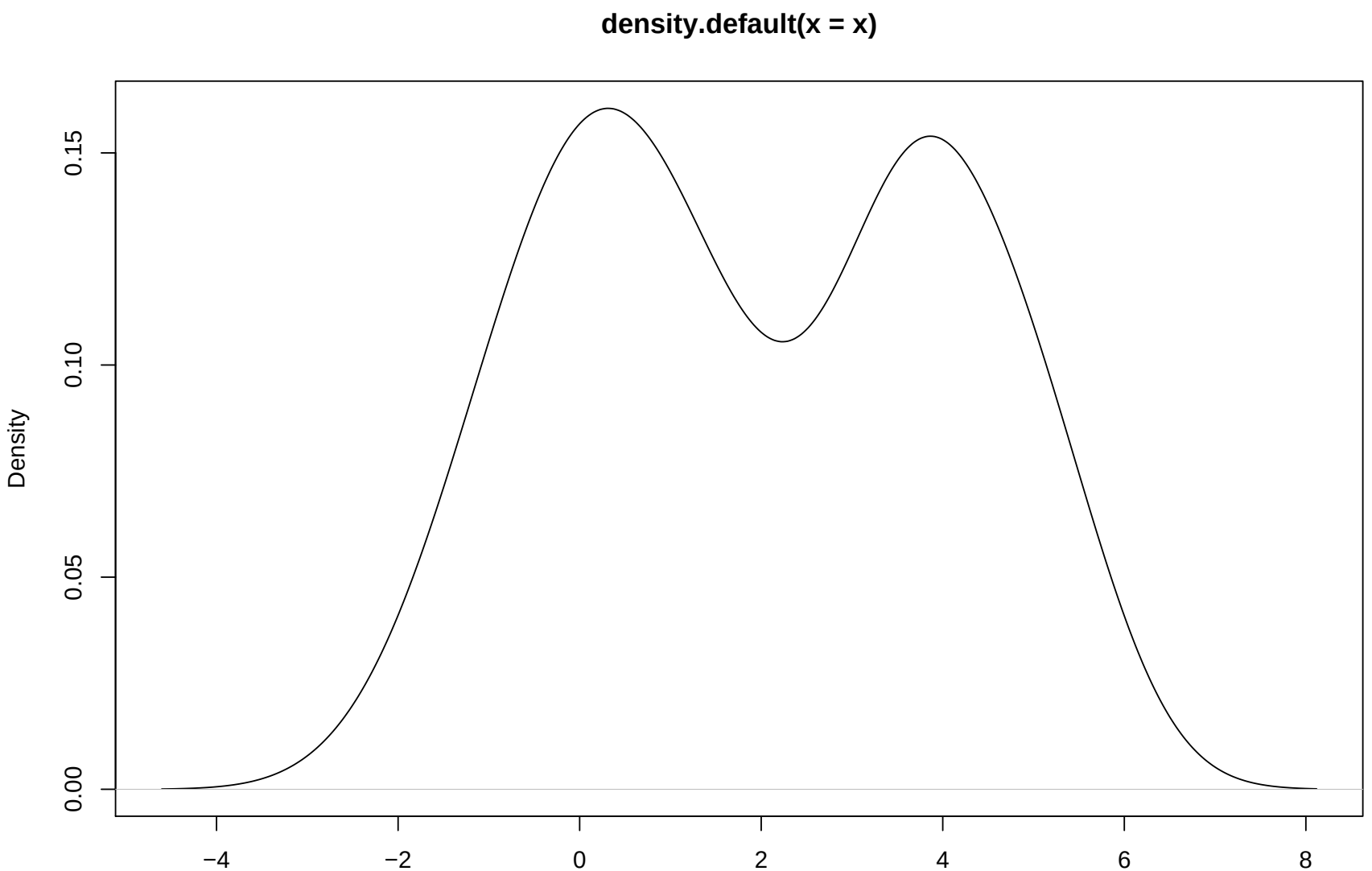
Example (data distribution):

- A lot of statistical theories require normality assumption
- But does a data set really follow a normal distribution?

Example (data distribution):

- One command line:

```
plot(density(x))
```



Example (data distribution):

- Can we test whether it's from a normal distribution? One command line:

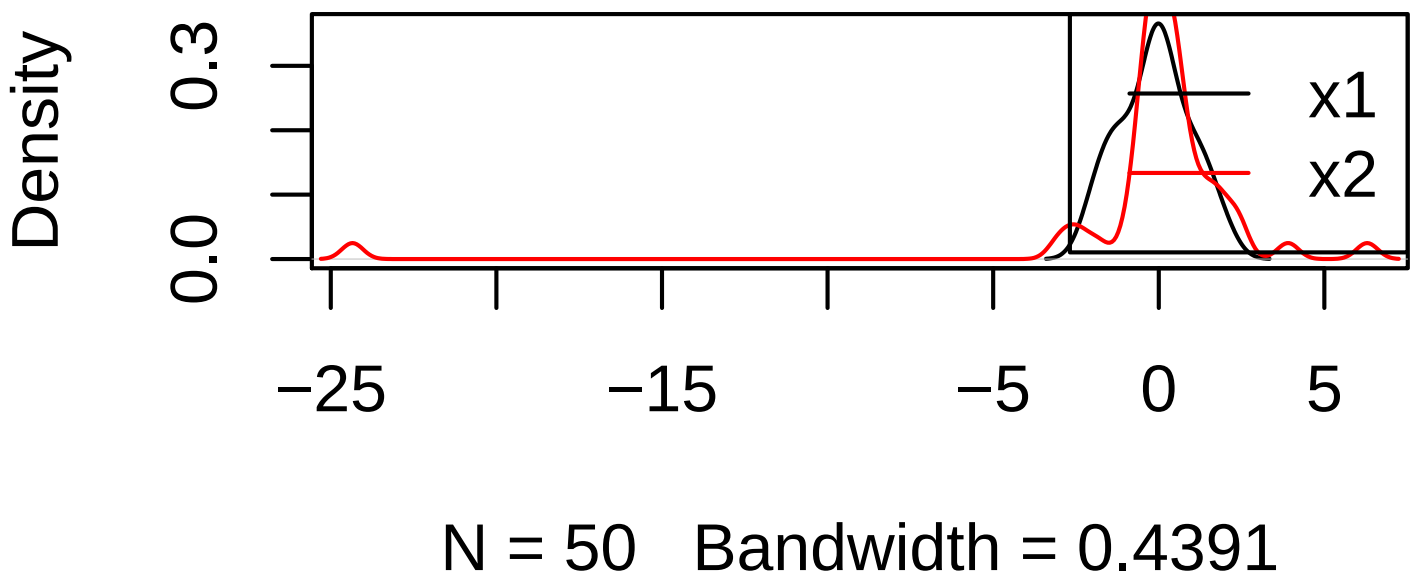
```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95103, p-value = 0.0009704
```

Example (data distribution):

- Compare two data distribution?

```
plot(density(x1), xlim = range(c(x1, x2)),main="")
lines(density(x2), col = 2)
legend("topright", c("x1", "x2"), lty = 1, col = 1:2)
```



Example (data distribution):

- Hypothesis testing?

```
ks.test(x1, x2)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x1 and x2
## D = 0.2, p-value = 0.2719
## alternative hypothesis: two-sided
```

Example (A/B testing):

- I have two advertisements
- Advertisement 1: 10 purchases out of 10000 clicks
- Advertisement 2: 3 purchases out of 5000 clicks
- Are the conversion rates (purchases/clicks) of the two advertisements significantly different?

Example (A/B testing):

- Confidence interval? An R package + one command line:

```
library(binom)
binom.confint(c(10, 3), c(10000, 5000), methods = "exact")
```

```
##   method  x     n  mean        lower       upper
## 1  exact 10 10000 1e-03 0.0004796397 0.001838264
## 2  exact  3  5000 6e-04 0.0001237515 0.001752444
```

```
prop.test(c(10, 3), c(10000, 5000))
```

```
##
##  2-sample test for equality of proportions with continu
##
## data:  c(10, 3) out of c(10000, 5000)
## X-squared = 0.24059, df = 1, p-value = 0.6238
## alternative hypothesis: two.sided
## 95 percent confidence interval:
```

Example (A/B testing):

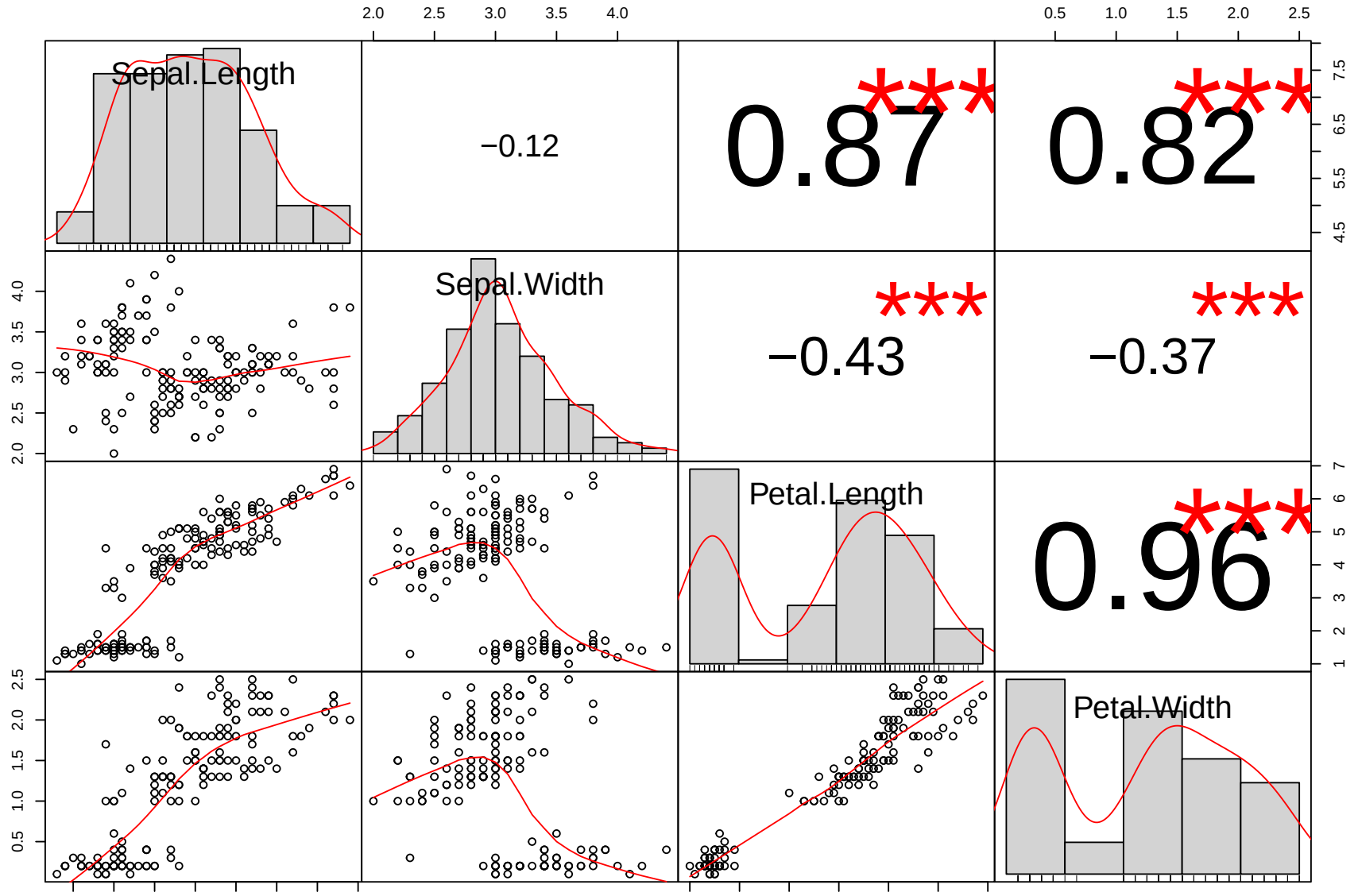
- I want to see if there are other methods

- `exact` - Pearson-Klopper method. See also `binom.test`.
- `asymptotic` - the text-book definition for confidence limits on a single proportion using the Central Limit Theorem.
- `agresti-coull` - Agresti-Coull method. For a 95% confidence interval, this method does not use the concept of "adding 2 successes and 2 failures," but rather uses the formulas explicitly described in the following link: http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval#Agresti-Coull_Interval.
- `wilson` - Wilson method.
- `prop.test` - equivalent to `prop.test(x = x, n = n, conf.level = conf.level)$conf.int`.
- `bayes` - see `binom.bayes`.
- `logit` - see `binom.logit`.
- `cloglog` - see `binom.cloglog`.
- `probit` - see `binom.probit`.
- `profile` - see `binom.profile`.

Figure 4:

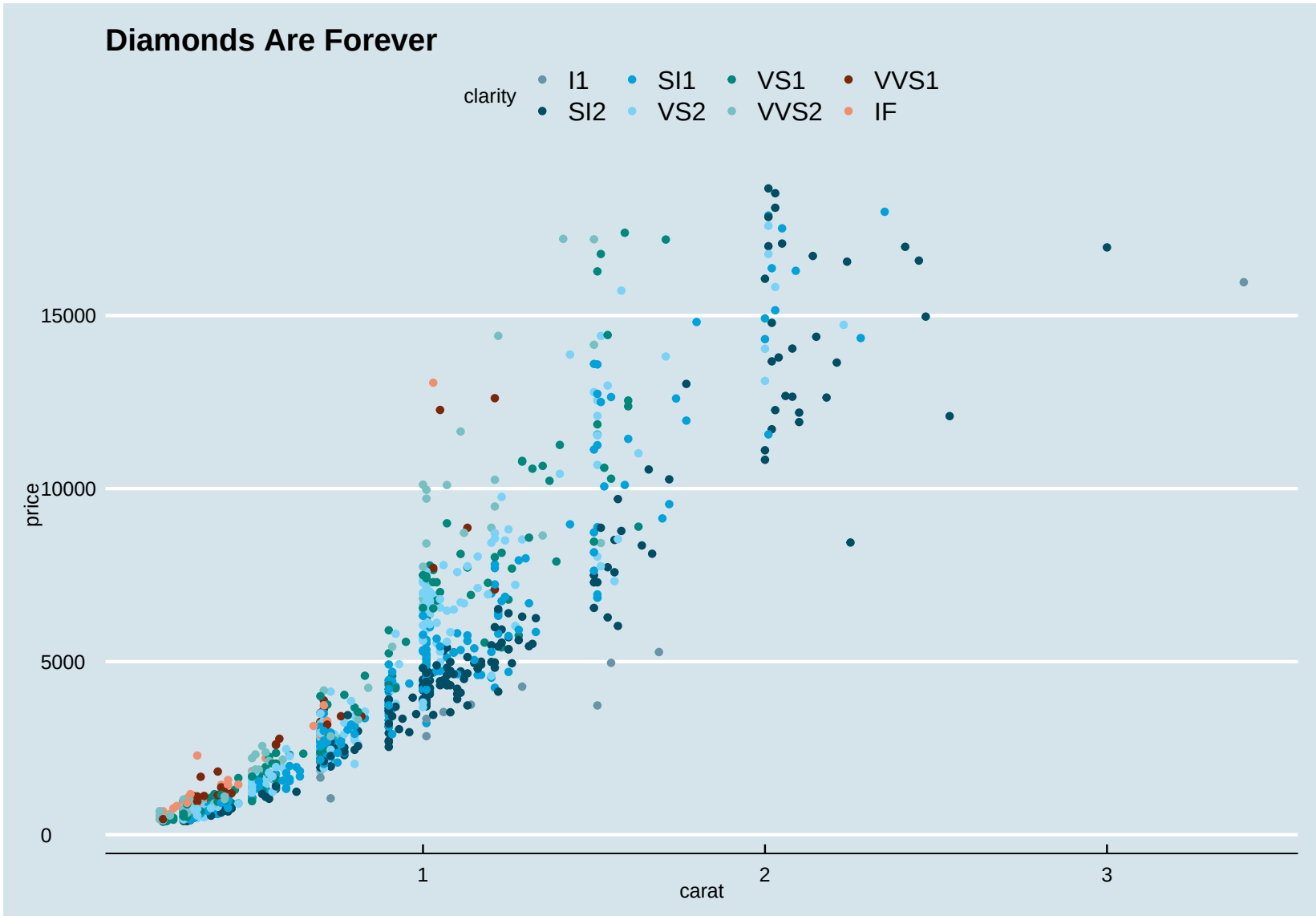
Example (correlation):

```
suppressPackageStartupMessages(library(PerformanceAnalytics
chart.Correlation(iris[-5], bg=iris$Species, pch=21)
```



Example (visualization):

The Economist



Other applications (Stock):

```
library(quantmod)
getSymbols("^TWII")
```

```
## [1] "TWII"
```

```
head(TWII)
```

```
##            TWII.Open TWII.High TWII.Low TWII.Close TWII
## 2007-01-02   7871.41   7937.26  7843.60    7920.80    5
## 2007-01-03   7954.96   7999.42  7917.30    7917.30    5
## 2007-01-04   7929.89   7955.90  7901.24    7934.51    5
## 2007-01-05   7940.20   7942.23  7821.71    7835.57    5
## 2007-01-08   7778.57   7797.57  7736.11    7736.71    4
## 2007-01-09   7778.38   7827.93  7778.38    7790.01    4
```

Other applications (Stock):

```
chartSeries(TWII, subset = "last 4 months", TA = c(addVo()
```

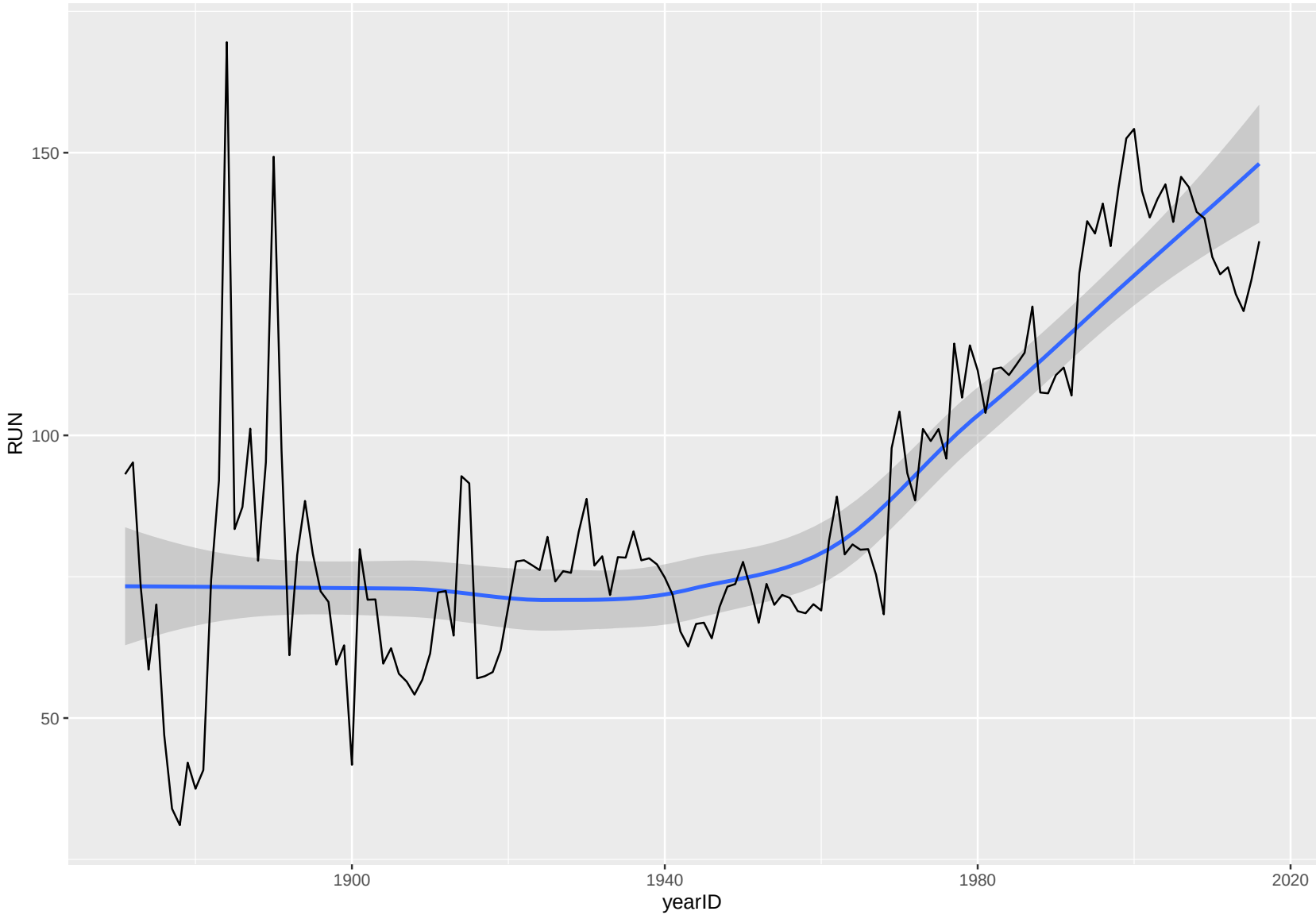
Other applications (Baseball):

```
library(Lahman)
head(Teams[,c("yearID", "name", "Rank", "W", "L", "R", "RA"
```

```
##   yearID                      name Rank  W  L   R  RA
## 1   1871    Boston Red Stockings    3 20 10 401 303
## 2   1871 Chicago White Stockings    2 19  9 302 241
## 3   1871  Cleveland Forest Citys    8 10 19 249 341
## 4   1871    Fort Wayne Kekiongas    7  7 12 137 243
## 5   1871         New York Mutuals    5 16 17 302 313
## 6   1871  Philadelphia Athletics    1 21  7 376 266
```

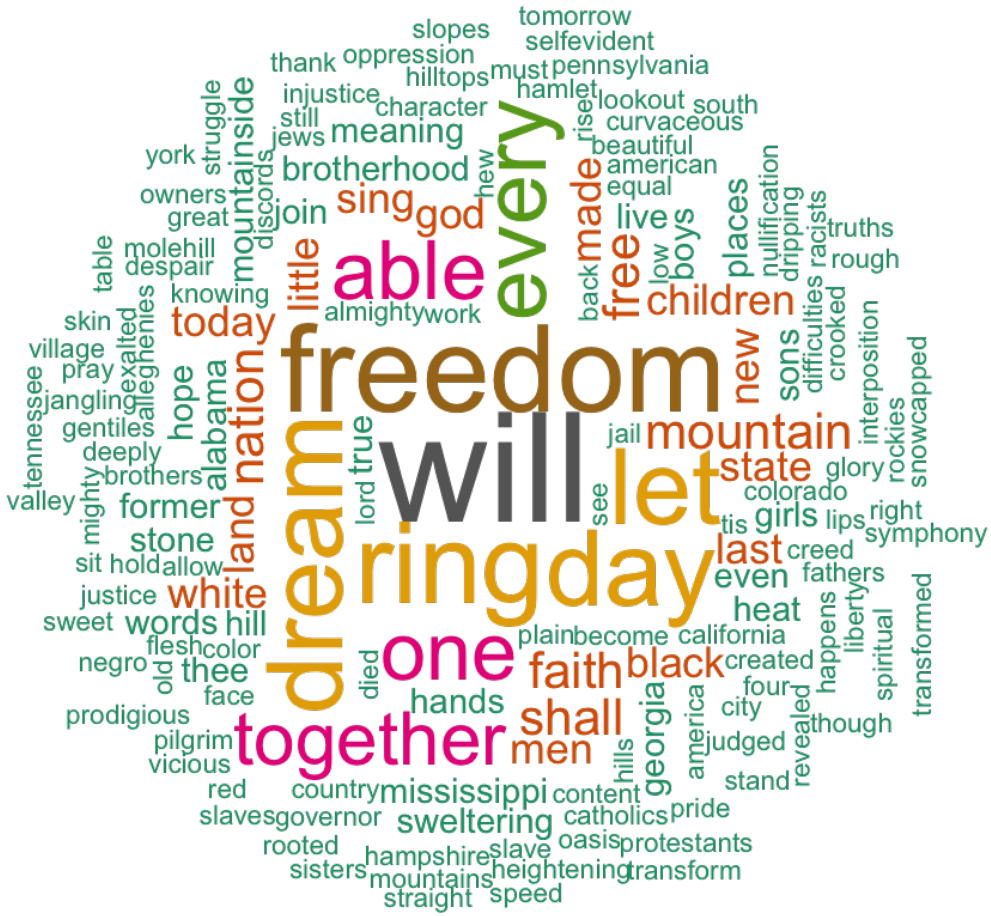
Other applications (Average runs in MLB):

```
## `geom_smooth()` using formula 'y ~ x'
```



Other applications (word cloud):

source: link



Web app:

- Shiny
- Gallery:
  - K-means Example

Big data:

- SparkR
- RHadoop
- MPI:
  - Rmpi
  - pbdMPI

Install R and Rstudio

- live demo

How to learn R - swirl

- live demo

Homework

Homework 1, Q2: Finish the swirl course "R Programming". Finish Section 1-15. Please take a screenshot as below. For each section, in your screenshot, you must have your **last command line** (my_div), the **last message** ("You've reached the end of this lesson! Returning to the main menu..."), and type **your name** after the last message. (Answer No to "Would you like to receive credit for completing this course on Coursera.org?")

```
> my_div
[1] 3.478505 3.181981 2.146460

| You are doing so well!

  |=======================================================================================| 100%
| Would you like to receive credit for completing this course on Coursera.org?

1: No
2: Yes

Selection: 1

| Excellent job!

| You've reached the end of this lesson! Returning to the main menu...

| Would you like to continue with one of these lessons?

1: Exploratory Data Analysis Principles of Analytic Graphs
2: No. Let me start something new.

Selection: Chih-Li Sung
```