Introduction to R and RStudio

Kwang-Yeol Park

2018 3 8

Course overview - 1Q

- 1st week: Introduction to R and RStudio
- 2nd week: Data structure
- 3rd week: Graph I
- 4th week: Data manipulation I
- 5th week: Data manipulation II
- 6th week: Graph II
- 7th week: RMarkdown
- 8th week: Midterm exam

Course overview - 2Q

- 9th week: Basic statistics I
- 10th week: Basic statistics II
- 11th week: Graph III
- 12th week: Artificial Intelligence
- 13th week: Presentation of Project
- 14th week: Git
- 15th week: Github
- 16th week: Final exam

진료: 4차 산업혁명

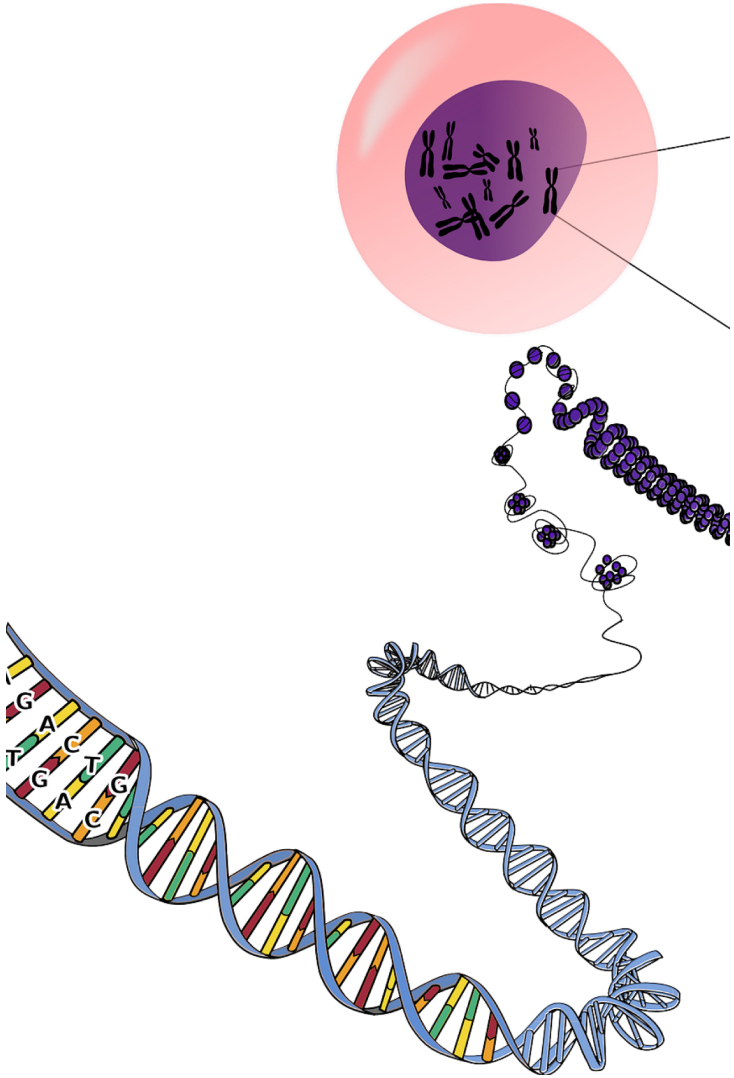
제 1차 산업혁명	제 2차 산업혁명	제 3차 산업혁명	제 4차 산업혁명
18세기	19세기~20세기 초	20세기 후반	2015년~
증기기관 기반의 기계화 혁명	전기 에너지 기반의 대량생산 혁명	컴퓨터와 인터넷 기반의 지식정보 혁명	IOT/CPS/인공지능 기반의 만물 초 지능 혁명
증기기관을 활용하여 영국의 섬유공업이 거대산업화	공장에 전력이 보급 되어 벨트 컨베이어를 사용한 대량 생산보급	인터넷과 스마트 혁명으로 미국주도의 글로벌 IT기업 부상	사람,사물,공간을 초 연결, 초 지능화 하여 산업구조 사회 시스템 혁신

원격진료

AI for practice

IoT + Big Data

Genomic data into clinic



연구: Big data

- 보건의료 빅데이터: 100만명 cohort
- 유전체 데이터

- MS-EXCEL
- SPSS
- SAS

- R
- Python

Prologue

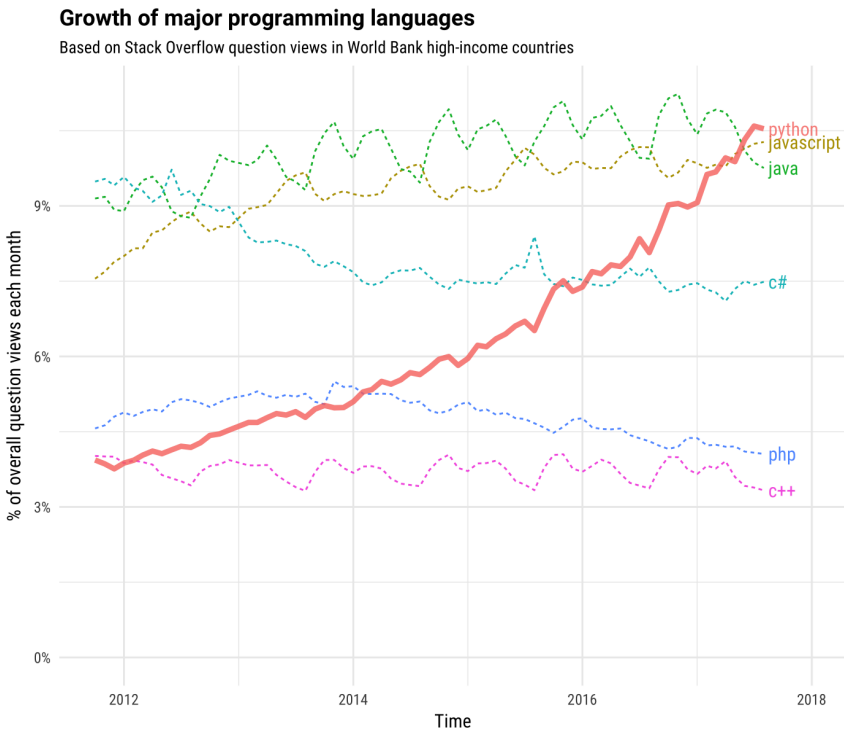
Why R ?

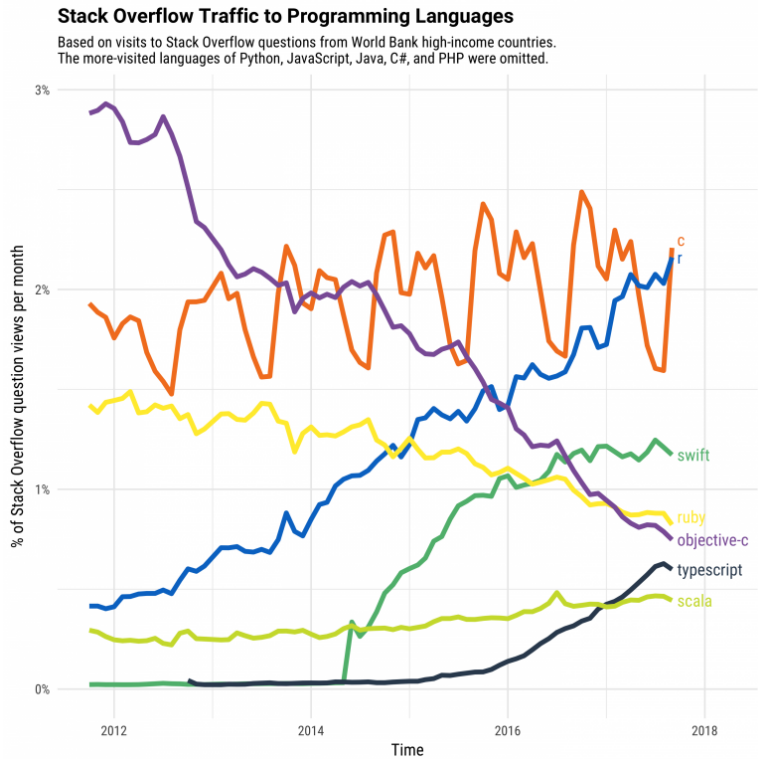
- A language and environment for statistical computing and graphics
- Allow the user to program algorithms and use libraries programmed by others
- Cool Graphics!
- Easy access to uptodate statistical methods
- Reproducible research
- **Life will be easier**

The 2017 Top Programming Languages - IEEE Spectrum

Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C		99.7
3. Java		99.5
4. C++		97.1
5. C#		87.7
6. R		87.7
7. JavaScript		85.6
8. PHP		81.2
9. Go		75.1
10. Swift		73.7

https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

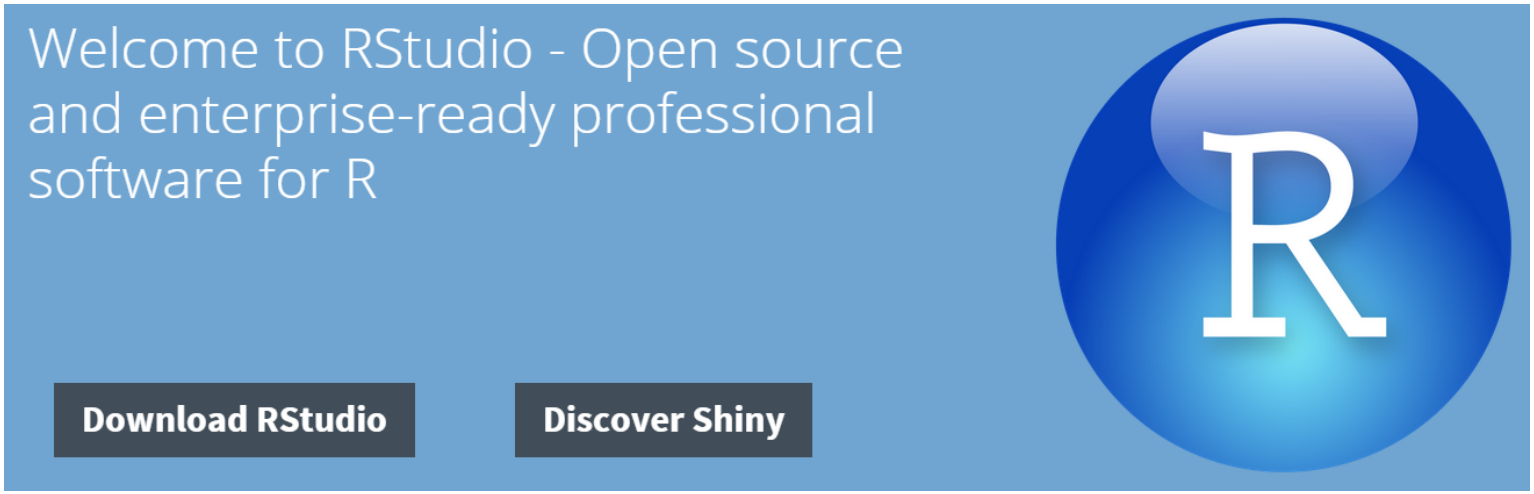




https://stackoverflow.blog/2017/10/10/impressive-growth-r/

RStudio

- GUI for R
- Free!
- Easy to use



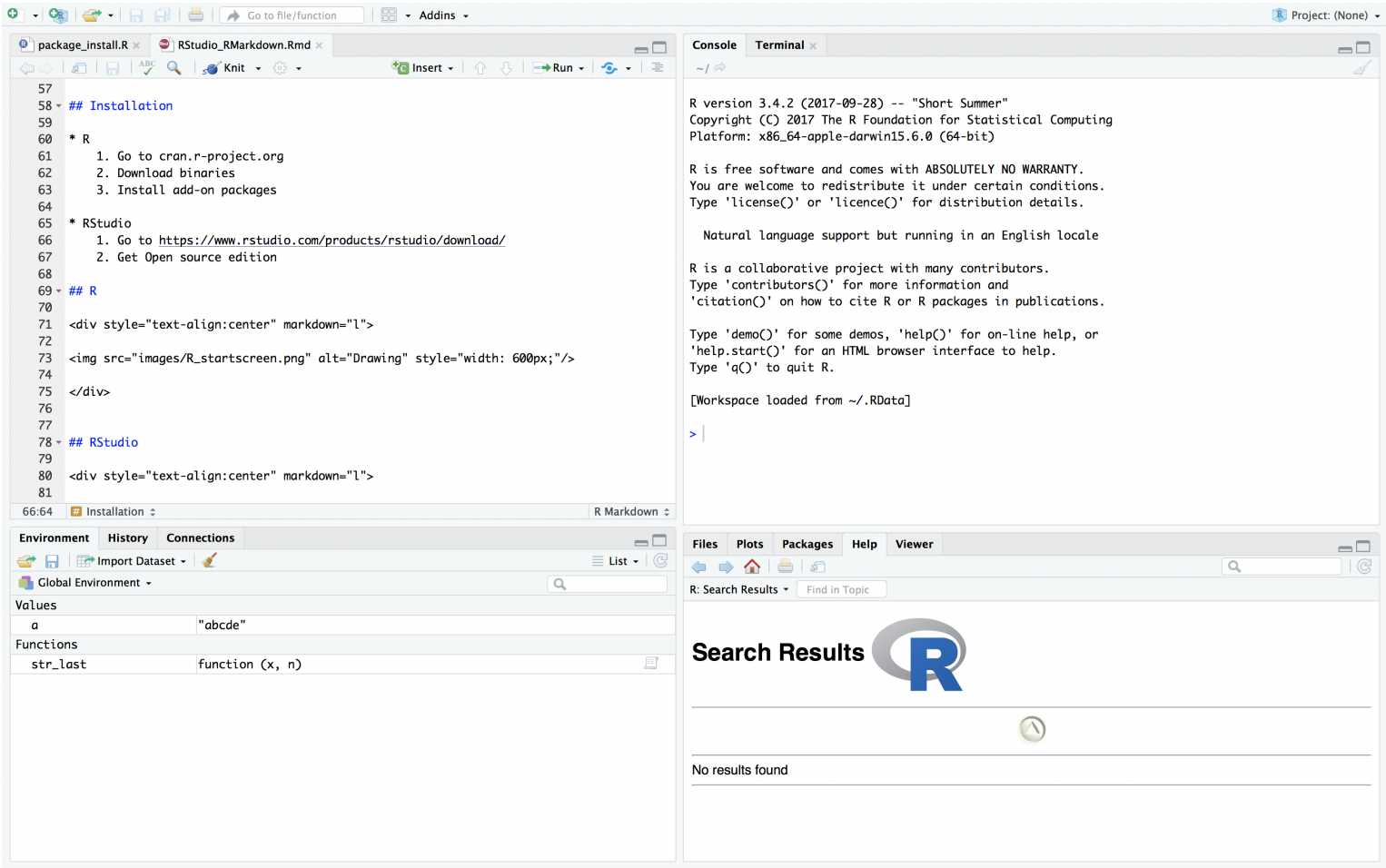
Installation

- R
  1. Go to https://cran.r-project.org
  2. Download binaries
  3. Install add-on packages
- RStudio
  1. Go to https://www.rstudio.com/products/rstudio/download/
  2. Get Open source edition

R

```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

[R.app GUI 1.70 (7463) x86_64-apple-darwin15.6.0]

[Workspace restored from /Users/kwangyeolpark/.RData]
[History restored from /Users/kwangyeolpark/.Rapp.history]

>
```

RStudio


Go to file/function
Addins
Project: (None)
package_install.R
RStudio_RMarkdown.Rmd
Knit
Insert
Run
57
58 ## Installation
59
60 * R
61 1. Go to cran.r-project.org
62 2. Download binaries
63 3. Install add-on packages
64
65 * RStudio
66 1. Go to https://www.rstudio.com/products/rstudio/download/
67 2. Get Open source edition
68
69 ## R
70
71
72
73
74
75
76
77
78 ## RStudio
79
80
81
66:64
Installation
R Markdown
Console
Terminal
~/
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.
[Workspace loaded from ~/.RData]
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
a "abcde"
Functions
str_last function (x, n)
Files
Plots
Packages
Help
Viewer
R: Search Results
Find in Topic
Search Results
No results found

Rstudio

- Make new file> file
- Comment/Uncomment> code
- Set working directory> session
- packages> tools
- Make RMarkdown file: later

You can use python or bash

```python
for i in [1, 2, 3, 4, 5]:
  print(i)
```

```
## 1
## 2
## 3
## 4
## 5
```

```bash
pwd
python --version
```

```
## /Users/kwangyeolpark/Dropbox/WorkingWithMyself/강의/SW중심대학/2018_1Q2Q_lecture
## Python 2.7.10
```

Simple calculation in R

```
1 + 3
```

```
## [1] 4
```

```
a <- c(100, 234, 356, 477, 888)
mean(a)
```

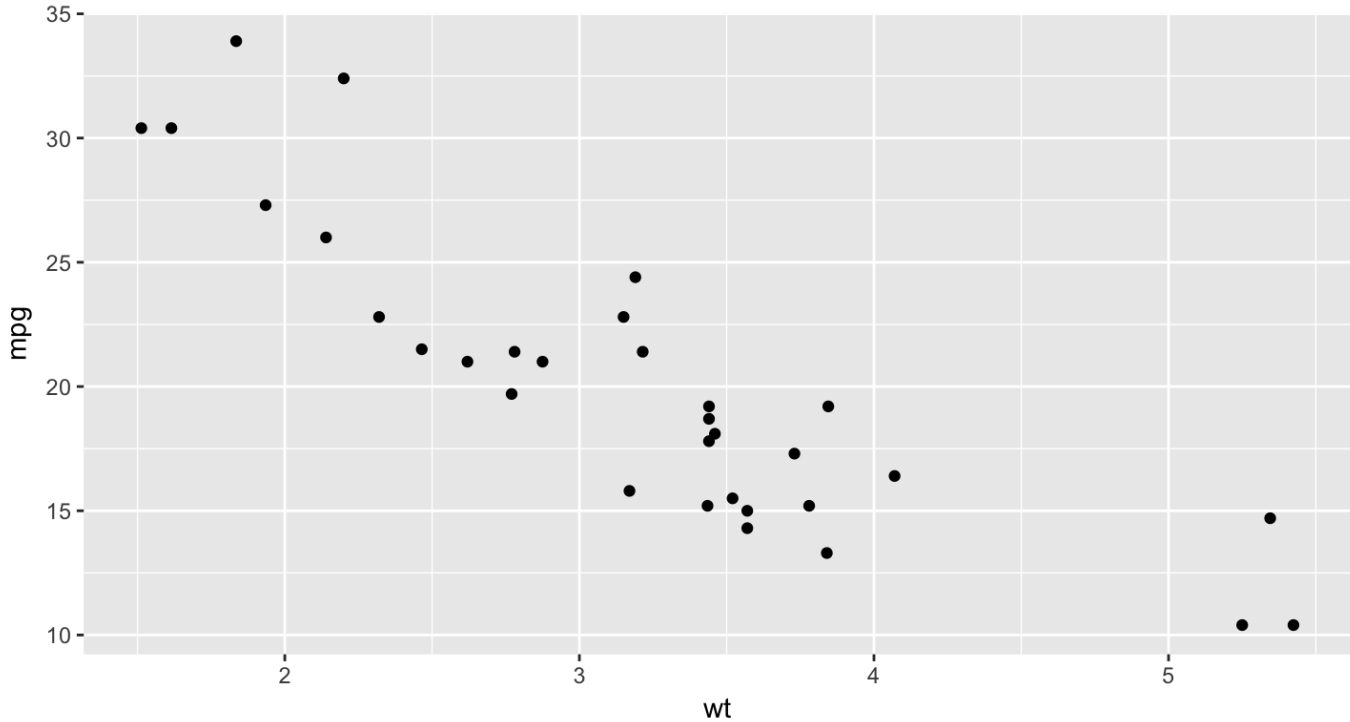
```
## [1] 411
```

```
sd(a)
```

```
## [1] 301.2308
```

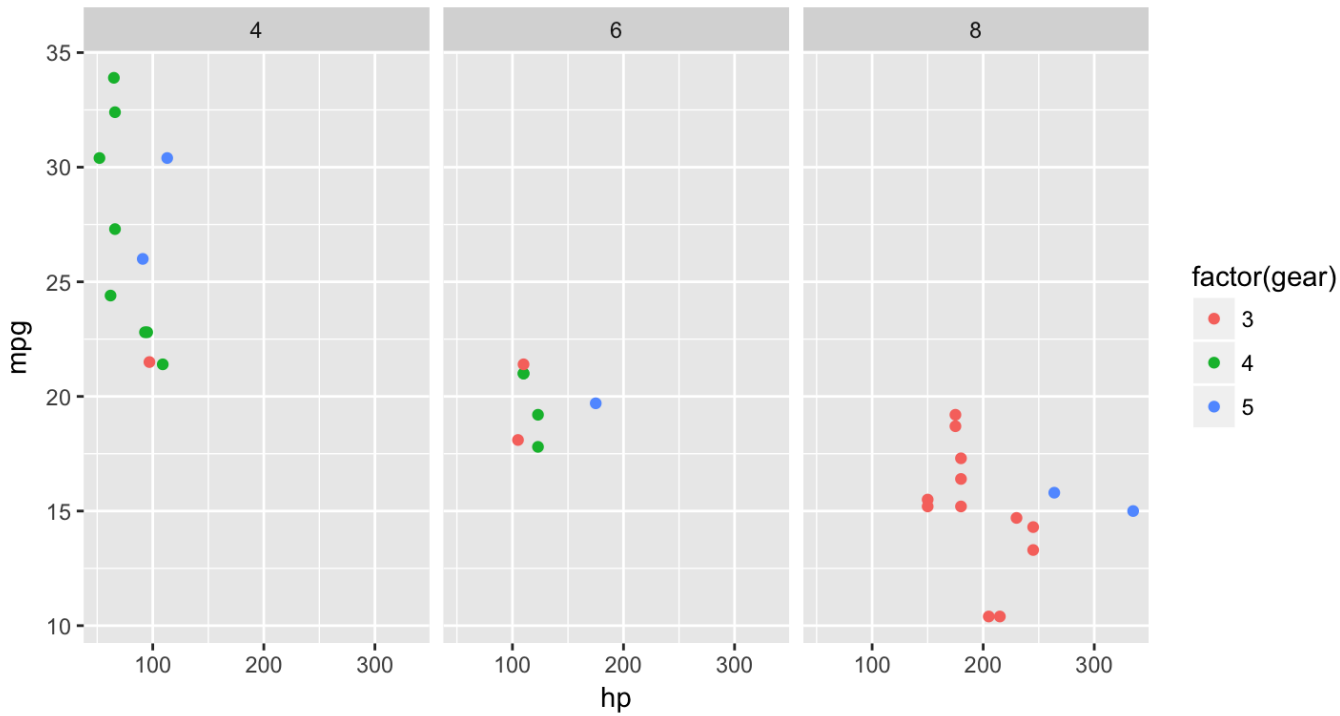
Simple plot in R

```
qplot(wt, mpg, data = mtcars)
```

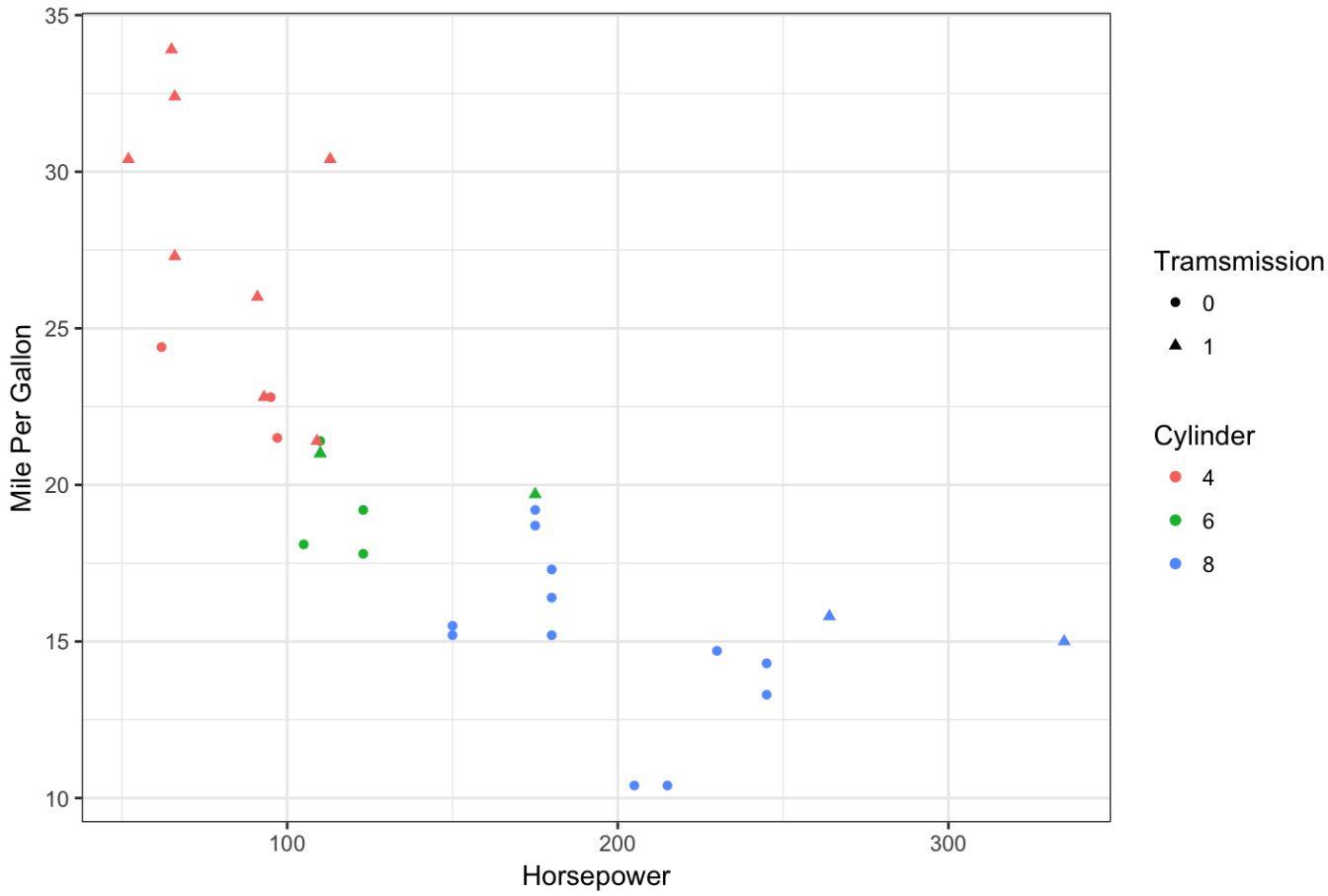


Another plot in R

```
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point(aes(color=factor(gear))) + facet_wrap( ~ cyl)
```



Another plot in R



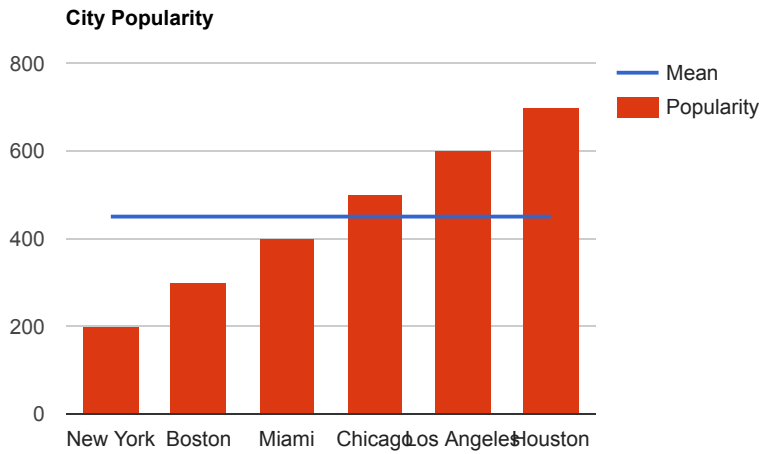
Demo

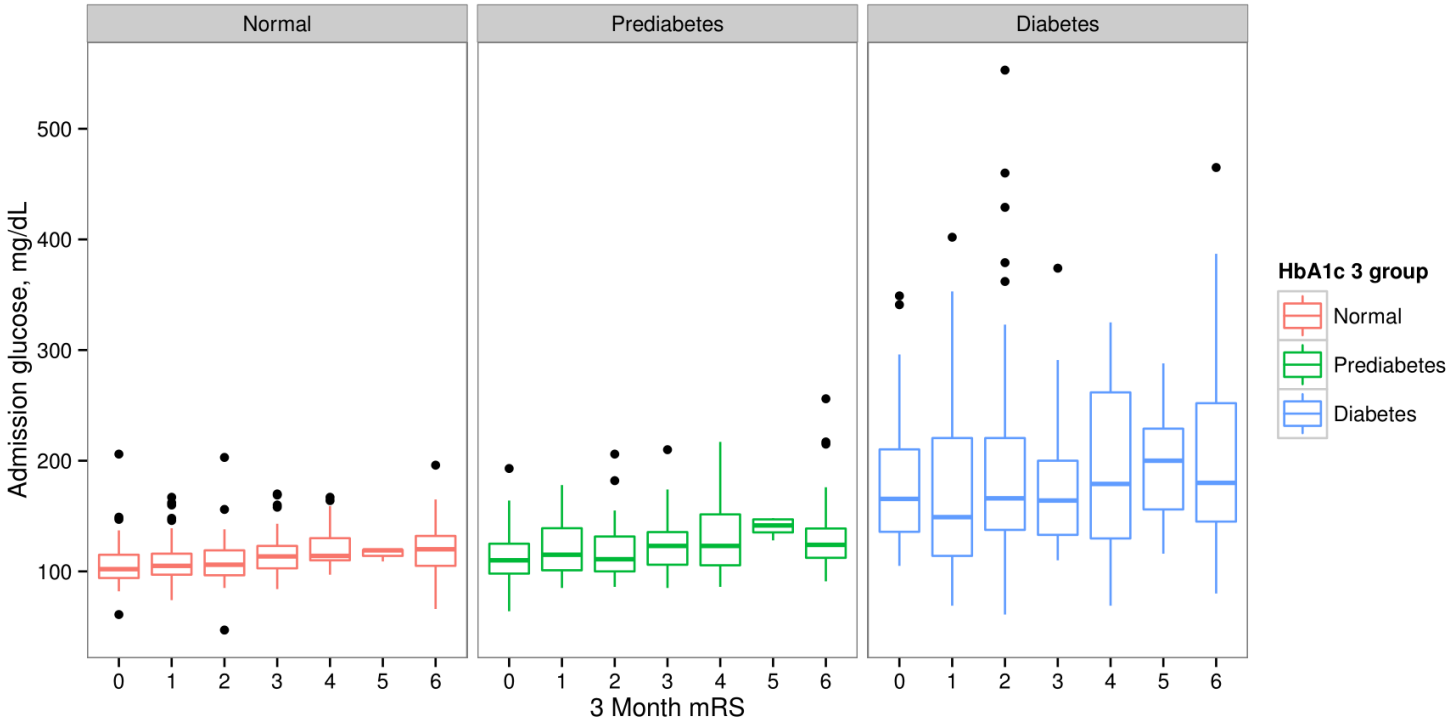
```
demo(graphics)
```

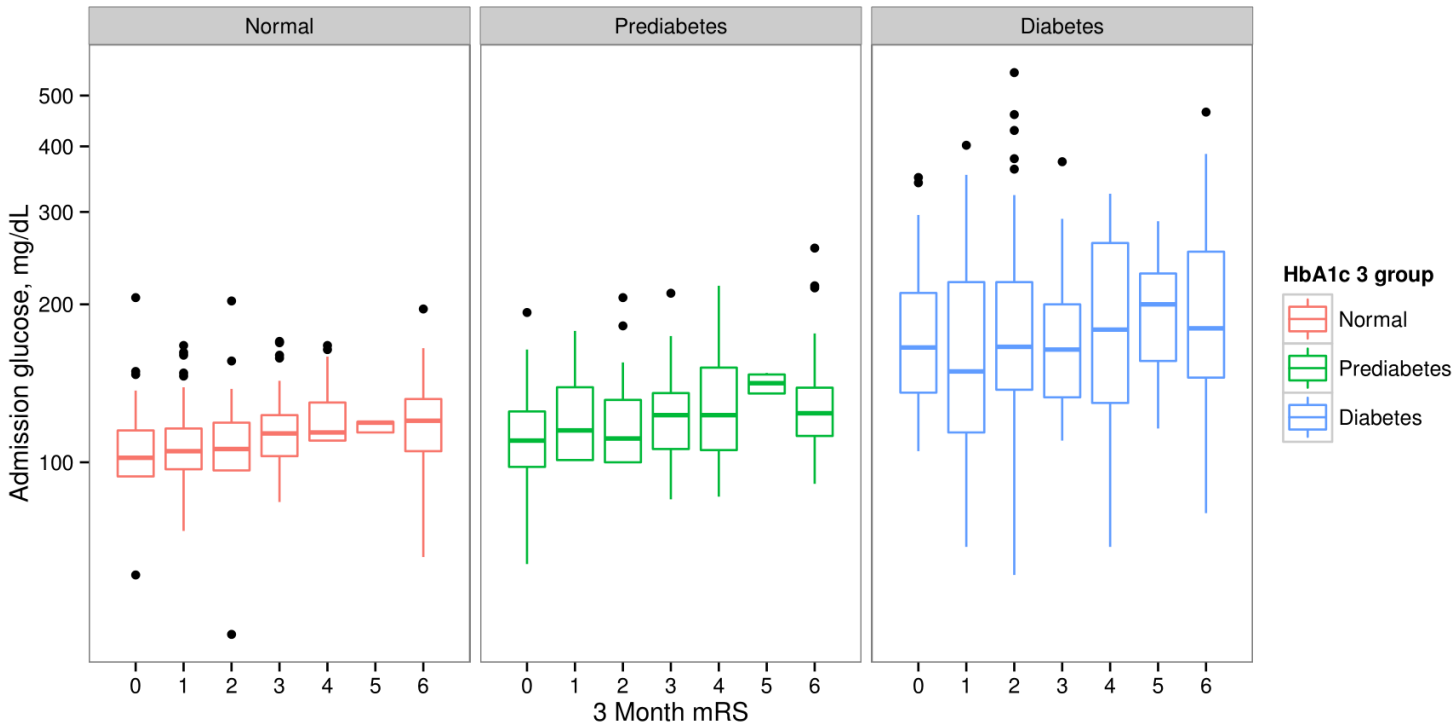
Combo chart in R

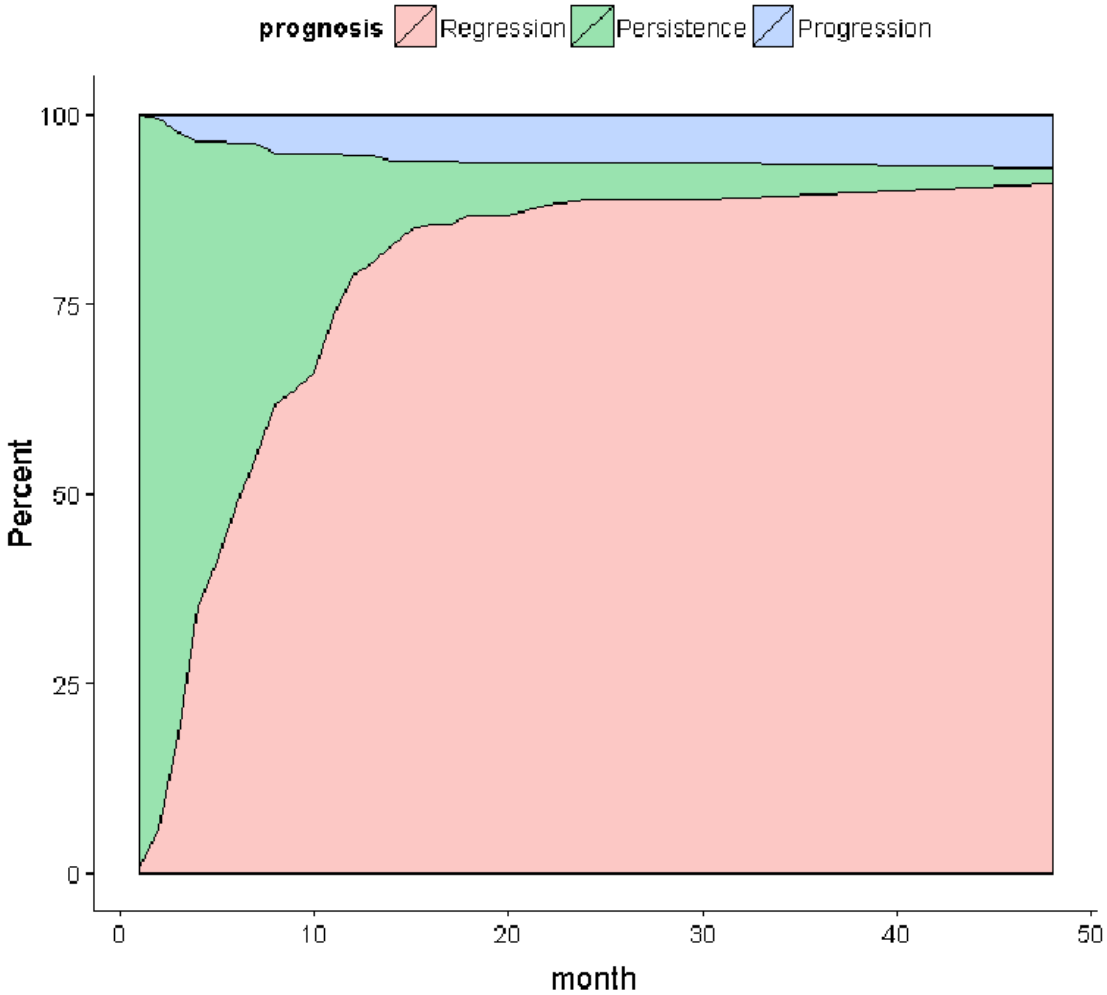
```
CityPopularity$Mean=mean(CityPopularity$Popularity)
CC <- gvisComboChart(CityPopularity, xvar='City',
                     yvar=c('Mean', 'Popularity'),
                     options=list(seriesType='bars',
                                  width=450, height=300,
                                  title='City Popularity',
                                  series='{0: {type:"line"}}'))
plot(CC)
```

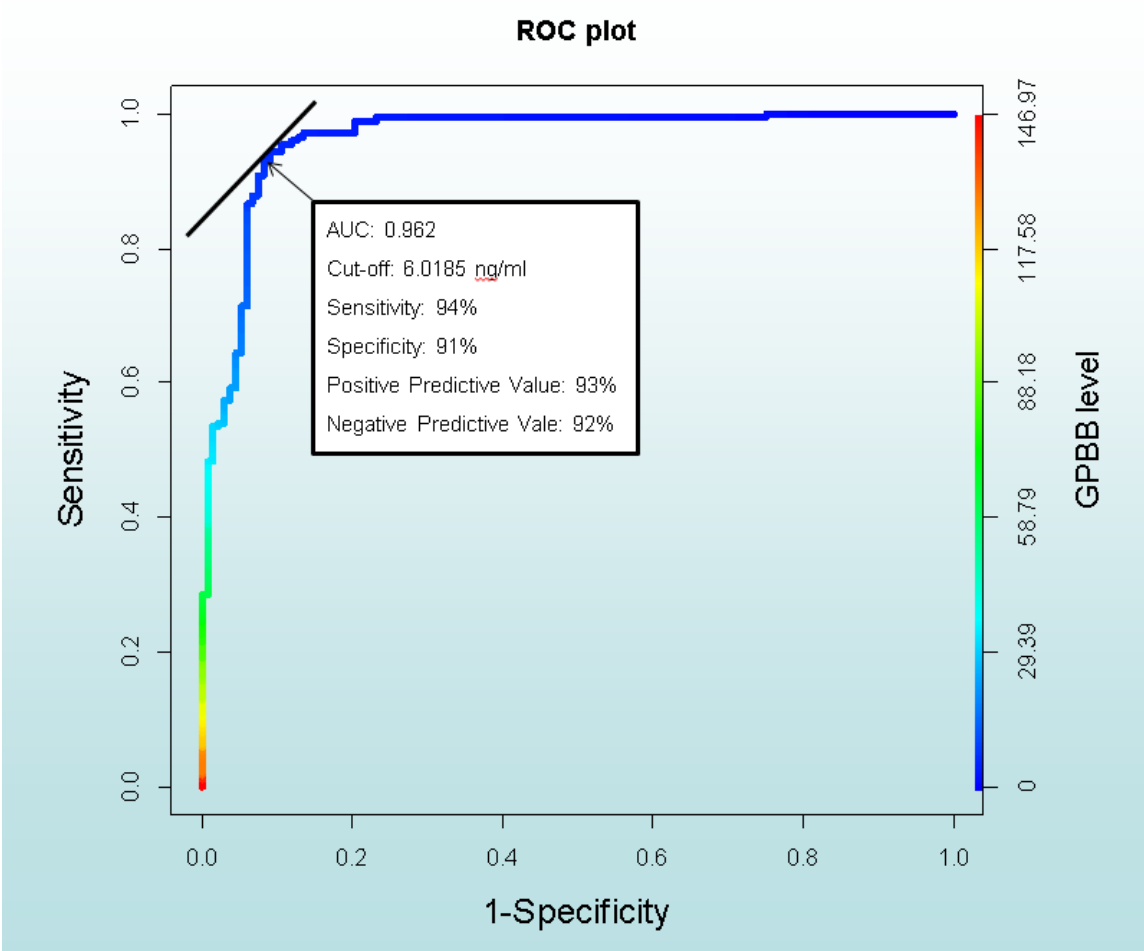
Combo chart in R



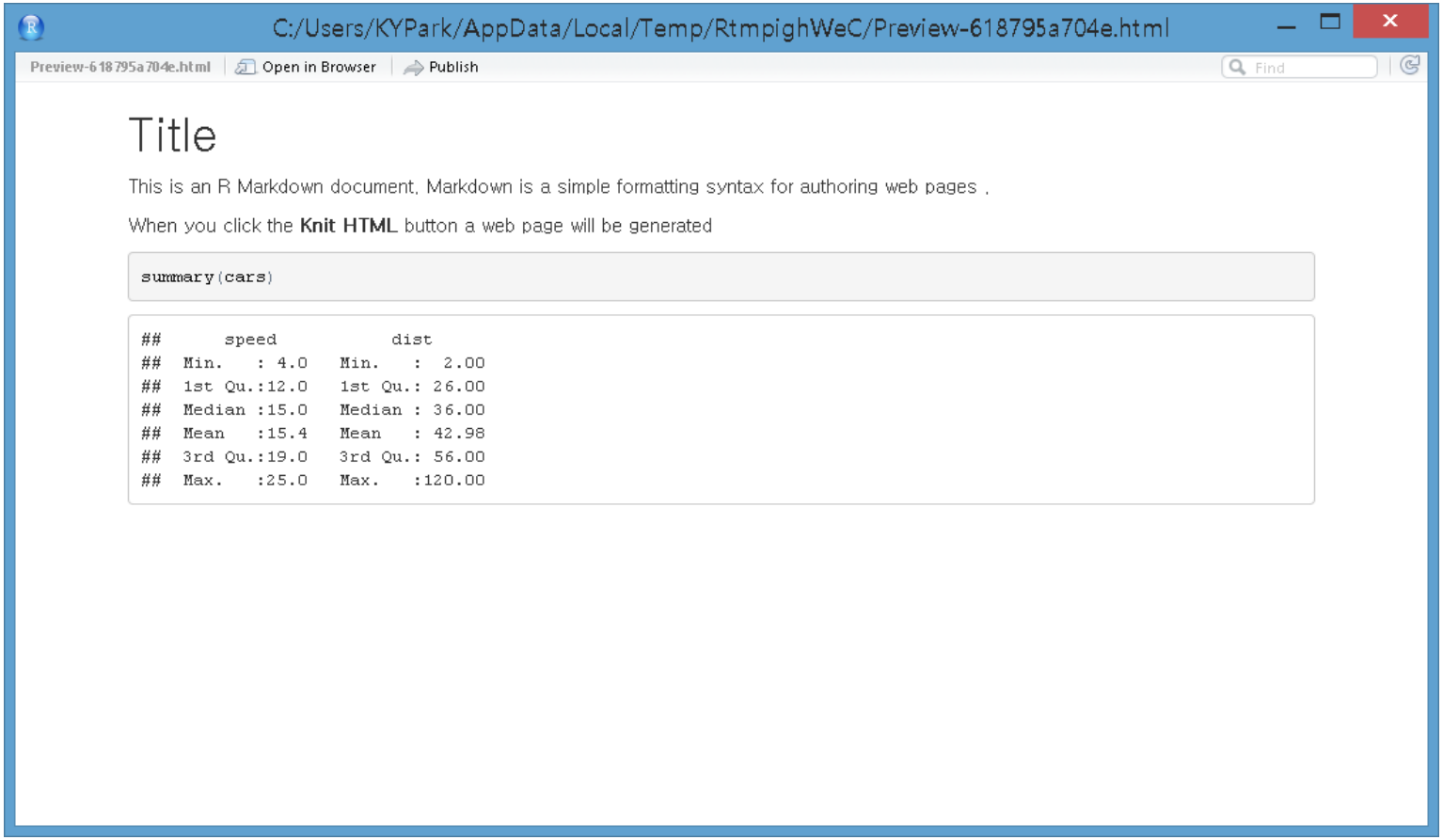

Normal
Prediabetes
Diabetes
Admission glucose, mg/dL
500
400
300
200
100
0
1
2
3
4
5
6
3 Month mRS
HbA1c 3 group
Normal
Prediabetes
Diabetes


Normal
Prediabetes
Diabetes
Admission glucose, mg/dL
500
400
300
200
100
0 1 2 3 4 5 6
3 Month mRS
HbA1c 3 group
Normal
Prediabetes
Diabetes

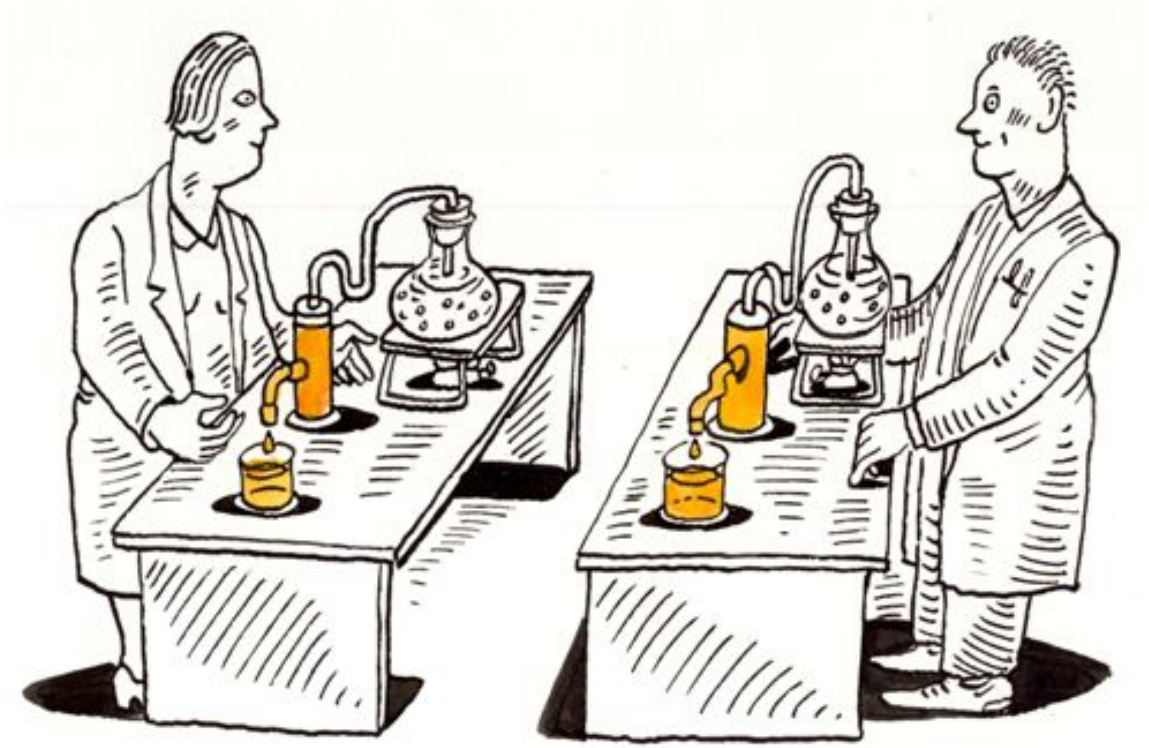

prognosis
Regression
Persistence
Progression
100
75
50
25
0
Percent
0
10
20
30
40
50
month


ROC plot
AUC: 0.962
Cut-off: 6.0185 ng/ml
Sensitivity: 94%
Specificity: 91%
Positive Predictive Value: 93%
Negative Predictive Vale: 92%
Sensitivity
1-Specificity
GPBB level
0.0
0.2
0.4
0.6
0.8
1.0
0
29.39
58.79
88.18
117.58
146.97

R Markdown


C:/Users/KYPark/AppData/Local/Temp/RtmpighWeC/Preview-618795a704e.html
Preview-618795a704e.html
Open in Browser
Publish
Find
Title
This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages .
When you click the Knit HTML button a web page will be generated
summary(cars)
speed   dist
Min.   : 4.0   Min.   : 2.00
1st Qu.:12.0   1st Qu.: 26.00
Median :15.0   Median : 36.00
Mean   :15.4   Mean   : 42.98
3rd Qu.:19.0   3rd Qu.: 56.00
Max.   :25.0   Max.   :120.00

Reproducible Research



Doing Research

- Collecting and cleaning data
  - MS-EXCEL or MS-ACCESS
  - SPSS
  - txt file (CSV, comma-separated values)
- Analysis
  - SPSS or SAS
  - R
- Writing
  - MS-WORD
  - LATEX
  - HTML

Problems

- Modification of data
  - Addition of new data
  - Error correction in dataset after analysis
- The connection between dataset and tables/graph might be broken easily.
- Writing methods section based on the analysis you did 3 months ago.
- Repetitive analyses are boring!

Case

Hi Dr. Park,

I have starting working on GPBB manuscript (ASH as well).

I need a paragraph from you describing the statistical methodology you used when you analyzed the data for the ISC abstract **a few years ago**.

Can you also send me the list of final study cohort (300 patients) to me?

Thanks,

GPBB study

Kwang-Yeol Park

Monday, August 17, 2015

DB history

August 6, 2012

1. Ross gave me 193 cases and 100 controls.

- 4 cases: not stroke (excluded by Svetlana)
- 9 cases: infarction volume is missing (2) or 0(7)

2. So, the remaining should be 180 cases and 100 controls.(By Svetlana)

- 8 cases and 3 controls: Baseline GPBB value is missing. (Ross said there are some samples lost.)

3. Therefore, the initial cohort consisted of 172 cases and 97 controls.
4. ISC abstract was based on this initial cohort of (172 + 97 = 269) cases and controls.

January 21, 2013

1. 36 controls were added to the cohort. Therefore, the study population was expanded to 172 cases and 133 controls (305 subjects in total). GPBB presentation at the ISC was based on this cohort.

Reproducible research matters

- Cleaning data + Analysis + Writing
- Combining tool
  - R (and Rstudio) + LaTeX or Markdown
- Output
  - PDF, HTML, MS-WORD