R 프로그래밍

(16주차)

2016. 06. 18(토)

장운호

(ADP 002-0004)

목차

Ⅰ. 정규 표현식 소개

1. 기본 문자열 처리

동일한 기능을 하는 함수가 다수 존재하는 바,
각각의 장단점을 파악해 두면, 상황에 따라 유용하게 사용 가능

문자열 관련 함수	함수 설명
grep(패턴, 문자열) str_detect(문자열, 패턴)	패턴이 들어가 있는 문자열벡터의 원소번호를 찾아줌
paste("문자열",...,sep=" ") str_c("문자열", sep=" ")	여러 개의 문자열을 하나로 합쳐줌
nchar(문자열) str_length(문자열)	문자열의 글자수를 세어줌
substr(문자열, start, stop) str_sub(문자열, start, stop)	문자열의 일부를 지정하여 추출해냄
gsub(패턴, 바꿀문자열, 문자열) str_replace_all(문자열, 패턴)	특정문자열을 별도로 지정한 문자열로 교체(substitute)함
문자열 %in% 특정 문자열 벡터	문자열이 특정 문자열 벡터내에 존재할 경우 TRUE를 반환

2. 정규 표현식 (Regular Expression)

문자열 내에 특정한 패턴을 찾아내기 위한 목적으로 개발된 언어(문법체계)임.

- 정규 표현식의 효과적인 활용을 위해서는 메타문자를 포함하는 패턴지정 방법에 대한 이해가 필요함.

구분	메타문자	메타문자의 의미
문자 지정	**. (점)**	임의의 문자 한 개를 의미합니다.
반복 지정	?	선행문자패턴이 0개 혹은 1개 나타납니다.
	+	선행문자패턴이 1개 이상 반복됩니다.
	*	선행문자패턴이 0개 이상 반복됩니다.
	{...}	(interval) 반복수를 직접 지정할 수 있습니다. 예를 들어 {3} : 3번 반복 {,7} : 7번 이하 {2,5} : 2~5번 반복
위치지정	^	라인의 제일 앞부분을 의미합니다.
	$	라인의 제일 끝부분을 의미합니다.
그룹 지정	[...]	안에 지정된 문자들 그룹 중에 한 문자를 지정합니다.
	[^...]	안에 지정된 그룹의 문자를 제외한 나머지(여집합)를 지정합니다.
기타	\	(escape) 메타의 의미를 없애줍니다.
	|	(alternation) OR연산을 합니다.
	( )	괄호는 패턴을 그룹화하는 역할을 합니다.

자료) 손에 잡히는 vim, advanced 리눅스 시스템 네트워크 프로그래밍 (http://sunyzero.tistory.com/215)

2. 정규 표현식 (Regular Expression)

정규 표현식을 익히는 가장 빠른 방법은 연습을 통해서 메타문자의 활용을 익히는 것임.

```
load("myDirData.RData")
head(myDirData)
str(myDirData)
myDirData
```

```
# 메타문자 : ^ (문장의 첫글자를 의미함)
grepIndex <- grep("^jan", x = myDirData)
myDirData[grepIndex]
length(myDirData[grepIndex])

# 메타문자 : $ (문장의 마지막을 의미함)
grepIndex <- grep("pdf$", x = myDirData)
myDirData[grepIndex]
length(myDirData[grepIndex])

# 메타문자 : ? (선행문자 패턴이 0개 혹은 1개 나타납니다)
grepIndex <- grep("_?.png$", x = myDirData)
grepIndex
myDirData[grepIndex]
length(myDirData[grepIndex])
```

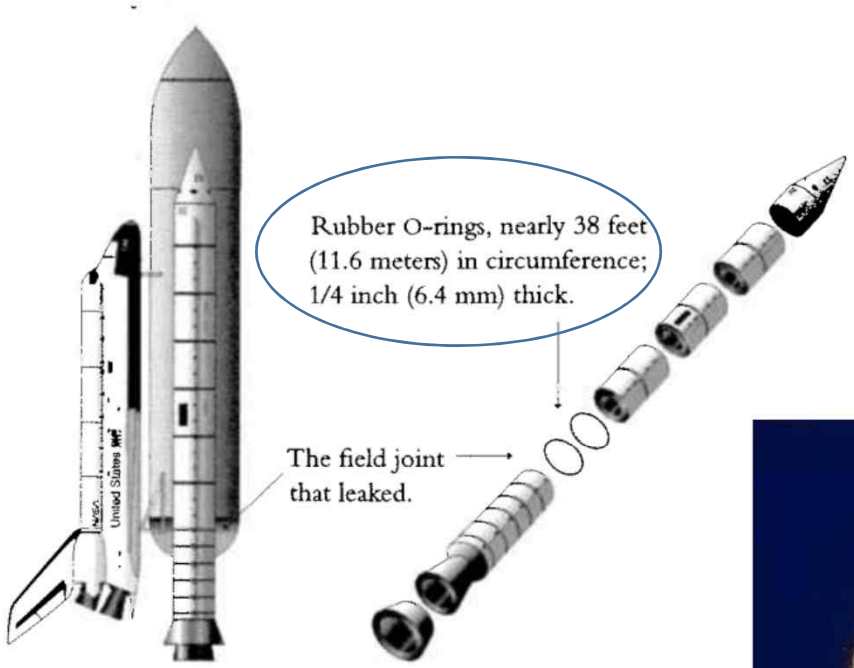
2. 정규 표현식 (Regular Expression)

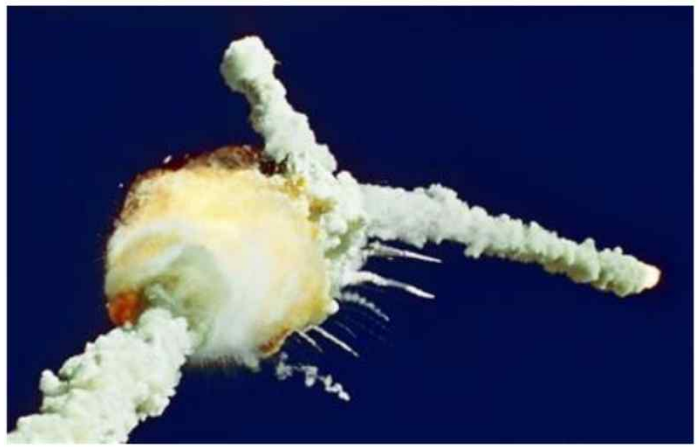
```
# 메타문자 ? (선행문자패턴이 0개나 1개)

grepIndex <- grep("nd?.png$", x = myDirData)
grepIndex
myDirData[grepIndex]

# 메타문자 : + (선행문자 패턴이 1개이상 나타납니다)
grepIndex <- grep("^jun[0-9a-zA-Z_]+.png$", x = myDirData)
grepIndex
myDirData[grepIndex]

# 메타문자 : + (선행문자 패턴이 1개이상 나타납니다)
grepIndex <- grep("[가-힣]+₩₩.RData$", x = myDirData)
grepIndex
myDirData[grepIndex]

# 메타문자 + 와 * 의 차이를 느껴보세요.
grepIndex <- grep("[가-힣]*₩₩.RData$", x = myDirData)
grepIndex
myDirData[grepIndex]
```

Ⅱ. 데이터 시각화

1. Space Shuttle "Challenger" (Tufte, "Visual Explanation")

두께 6.4 mm, 둘레길이 11.6 미터인 패킹 파열로 인한 충격으로 폭파됨. (1986년)

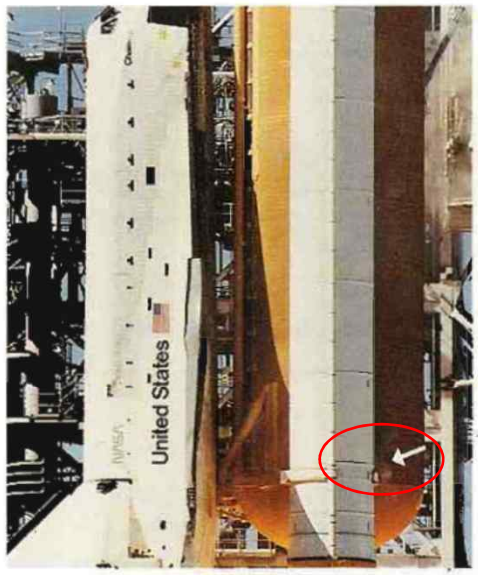




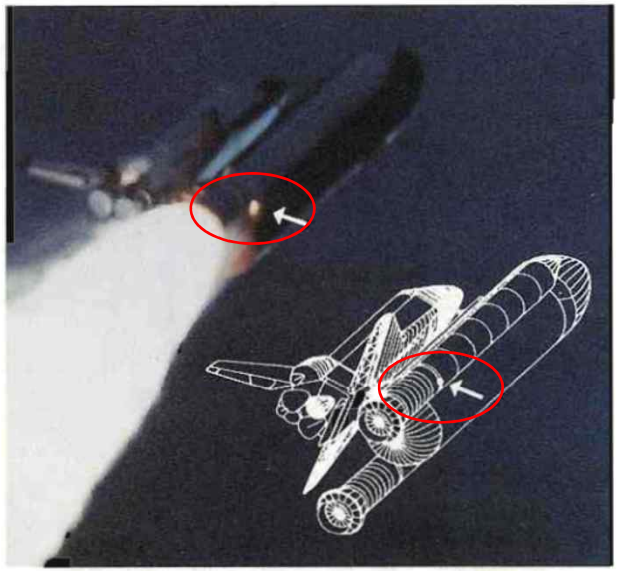
http://msnbcmedia1.msn.com/j/msnbc/Components/Photos/050709/050609_columbia_hmed_6p.hmedium.jpg

1. Space Shuttle "Challenger" (Tufte, "Visual Explanation")

발사 예정일(D-day) 당시 매우 추운 날씨가 예보 되었으나, 발사를 강행하였음.



Less than 1 second after ignition, a puff of smoke appeared at the aft joint of the right booster, indicating that the O-rings burned through and failed to seal. At this point, all was lost.

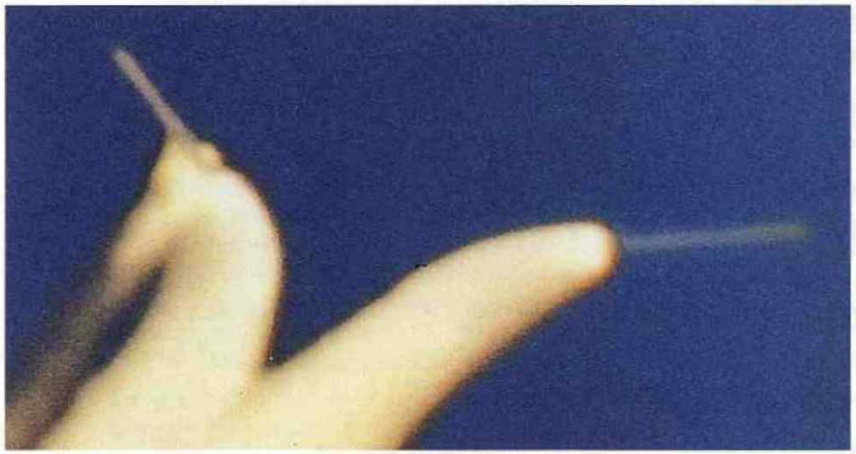


On the launch pad, the leak lasted only about 2 seconds and then apparently was plugged by putty and insulation as the shuttle rose, flying through rather strong cross-winds. Then 58.788 seconds after ignition, when the Challenger was 6 miles up, a flicker of flame emerged from the leaky joint. Within seconds, the flame grew and engulfed the fuel tank (containing liquid hydrogen and liquid oxygen). That tank ruptured and exploded, destroying the shuttle.

1. Space Shuttle "Challenger" (Tufte, "Visual Explanation")

안타까운 사실은 소중한 인명이 희생된, 이 사건을 사전에 막을 수 있었다는 것임.

- D-day 전날까지도 전문가들이 모여 안전성 문제에 대한 검토가 있었음.



As the shuttle exploded and broke up at approximately 73 seconds after launch, the two booster rockets crisscrossed and continued flying wildly. The right booster, identifiable by its failure plume, is now to the left of its non-defective counterpart.

The flight crew of Challenger 51-L. Front row, left to right: Michael J. Smith, pilot; Francis R. (Dick) Scobee, commander; Ronald E. McNair. Back row: Ellison S. Onizuka, S. Christa McAuliffe, Gregory B. Jarvis, Judith A. Resnik.

1. Space Shuttle "Challenger" (Tufte, "Visual Explanation")

위험을 인지할 만한 충분한 量의 정보를 가지고 있었음에도 불구하고,
올바른 의사결정을 하지 못한 이유는, 핵심을 전달할 수 없었던 자료에 기인함.

[D-day 전날 회의자료]

TEMPERATURE CONCERN ON

SRM JOINTS

27 JAN 1986

Solid Rocket Motor (고체연료 엔진)

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
22A	None	None	0.280	None	None	36°--66°
22A	NONE	NONE	0.280	NONE	NONE	338°-18°
15A	0.010	154.0	0.280	4.25	5.25	163
15B	0.038	130.0	0.280	12.50	58.75	354
15B	None	45.0	0.280	None	29.50	354
13B	0.028	110.0	0.280	3.00	None	275
11A	None	None	0.280	None	None	--
10A	0.040	217.0	0.280	3.00	14.50	351
2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

BLOW BY HISTORY
SRM-15 WORST BLOW-BY
- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

MOTOR	M&T	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

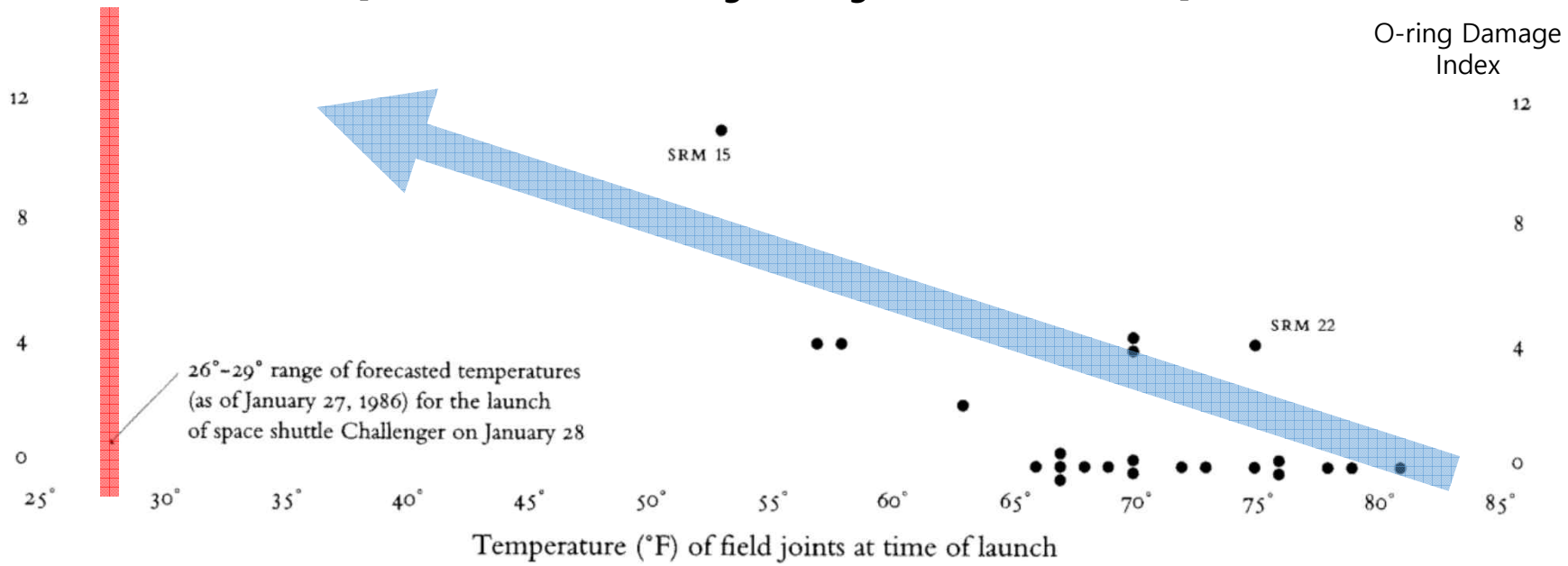
MOTOR	O-RING
DM-1	47
DM-2	52
QM-3	48
QM-4	51
SRM-15	53
SRM-22	75
SRM-25	29 27

1. Space Shuttle "Challenger" (Tufte, "Visual Explanation")

X좌표와 Y좌표를 일정한 간격으로 두고 點만 찍었서도,
돌이킬 수 없는 손실을 막을 수 있었음.

- 이를 통계적인 기법을 적용하여 분석한 논문도 존재함.

[발사당일 온도와 O-ring Damage Index간의 산점도]



주 : 화살표 등은 이해를 돕기 위한 목적으로 추가 반영

2. Anscombe's Quartet

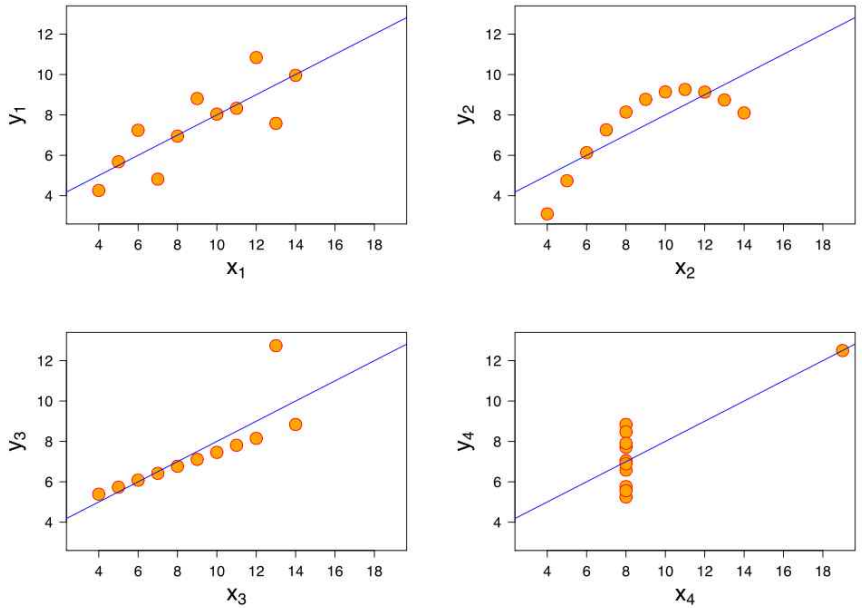
데이터 셋간의 평균, 분산, 선형 회귀직선의 계수들이 정확히 같거나, 거의 유사함.

- 기본 통계량 수치는 유사하나, 산점도로 표현 시, 그 차이가 매우 명확하게 드러남.
  ※ 데이터 Review時에는 반드시 Outlier 존재 여부나 데이터의 "분포(모양)"를 확인해야 함.

[데이터 Set]

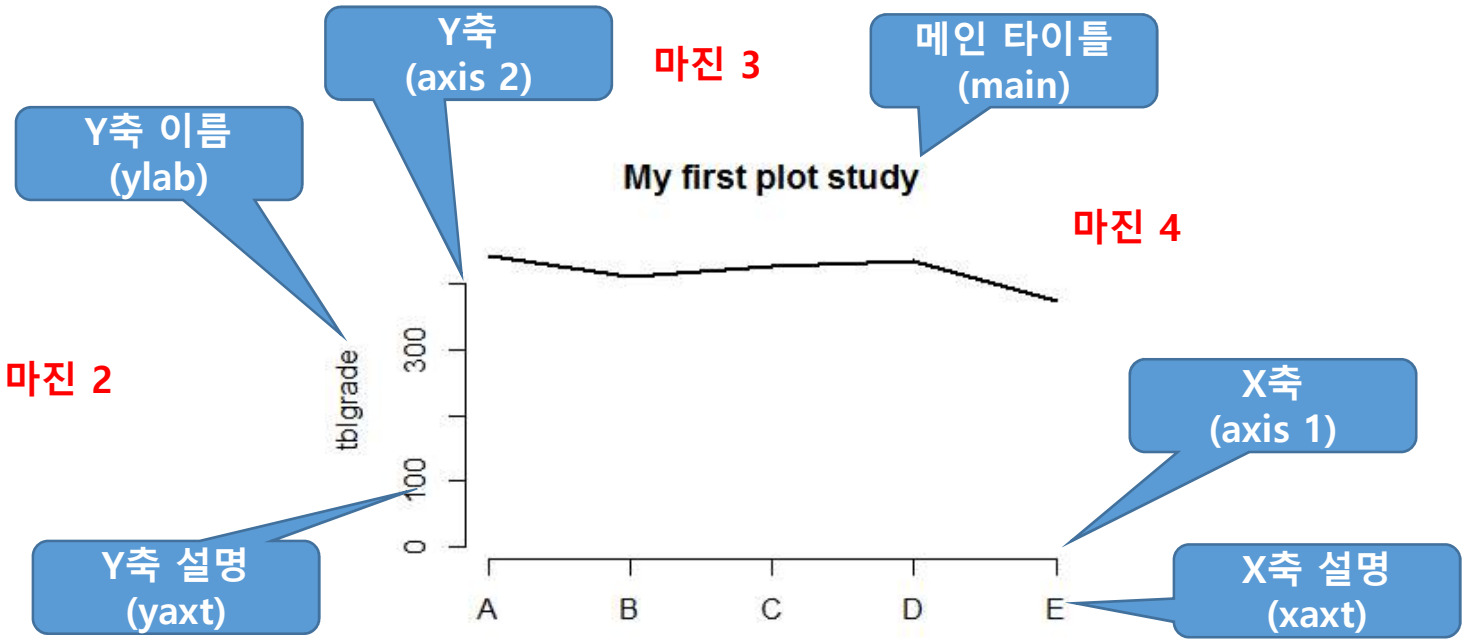
1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

[시각화 결과]



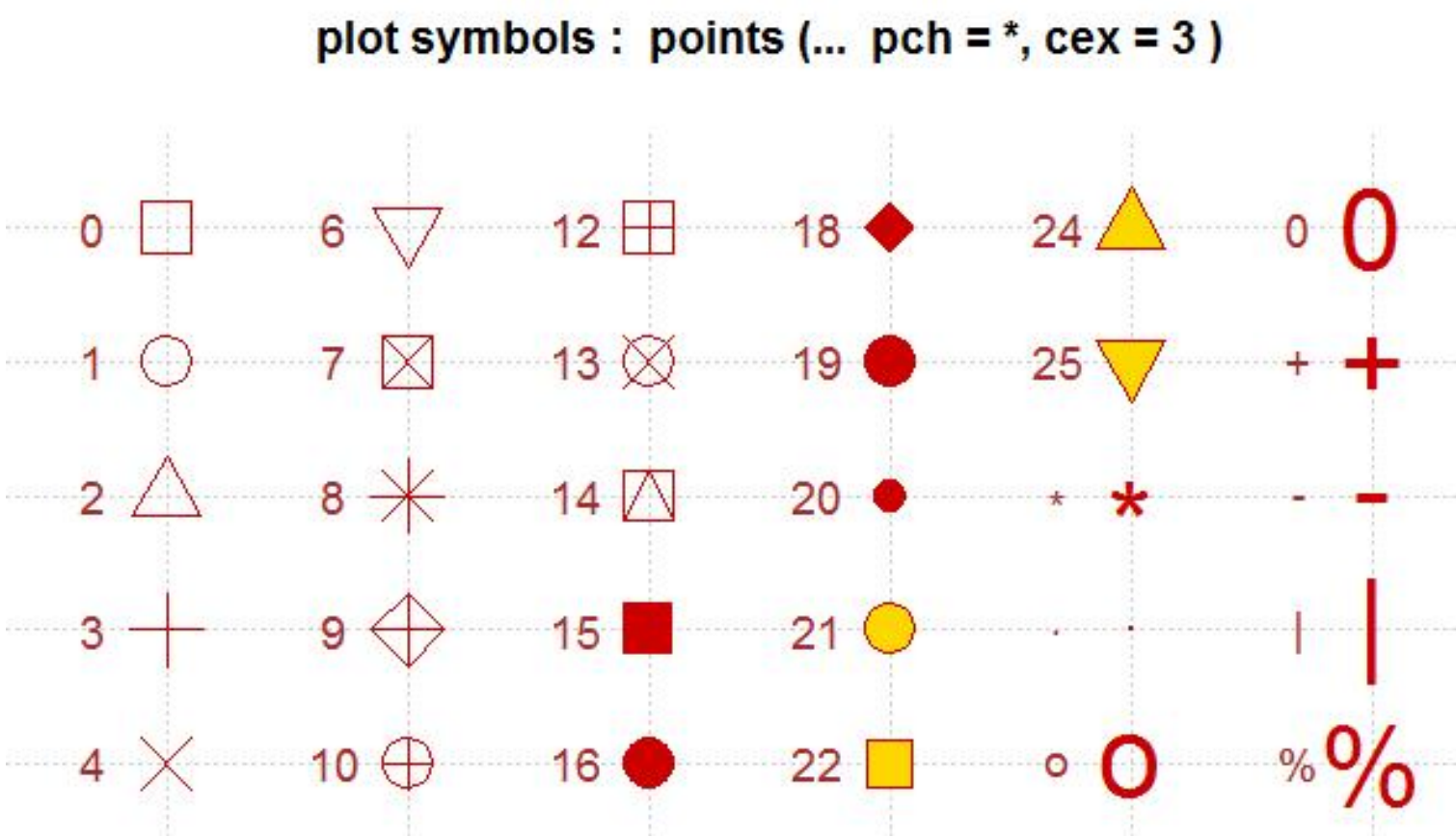
자료: Wikipedia (일부 수정)

※ 도표의 각 영역을 지칭하는 명칭



※ plot 인자 지정(1)

번호 지정으로 그래프상의 점들의 모양을 지정할 수 있음.



3. plot 주요 키워드 소개

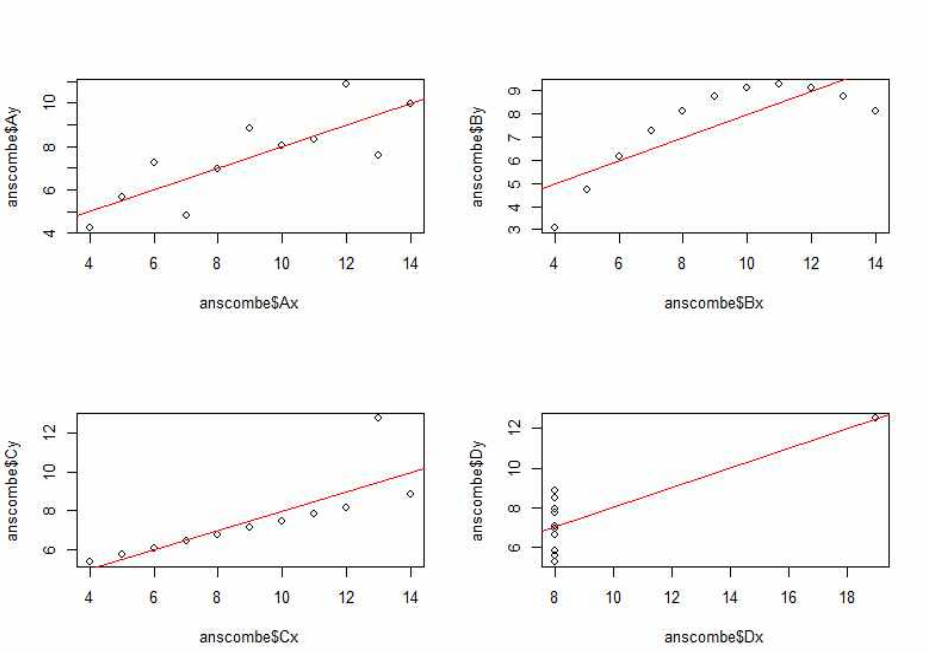
기타 자주 사용되는 plot함수의 argument 또는 외부 출력 함수는 아래와 같음.

구분	주요 내용	비고
lty, lwd	Line type과 line width (숫자 혹은 문자열로 지정가능)	(0=blank, 1=solid (default), 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash) 또는 "blank", "solid", "dashed", "dotted", "dotdash", "longdash", or "twodash",
type	선그래프, 점그래프 등 지정	type=l (선그래프), type=d (선/점그래프 혼합)
cex	글자크기 조정	**정수의 숫자로 지정**
col	선이나 문자열의 색상지정	**숫자 혹은 문자열로 지정**
las	축레이블의 수직/수평여부 지정	0: always parallel to the axis [default], 1: always horizontal, 2: always perpendicular to the axis, 3: always vertical.
pdf, png	pdf 혹은 png 그림파일로 출력	File 이름 및 그림의 크기 별도 지정 필요 ※ 사용후 dev.off() 적용 필수

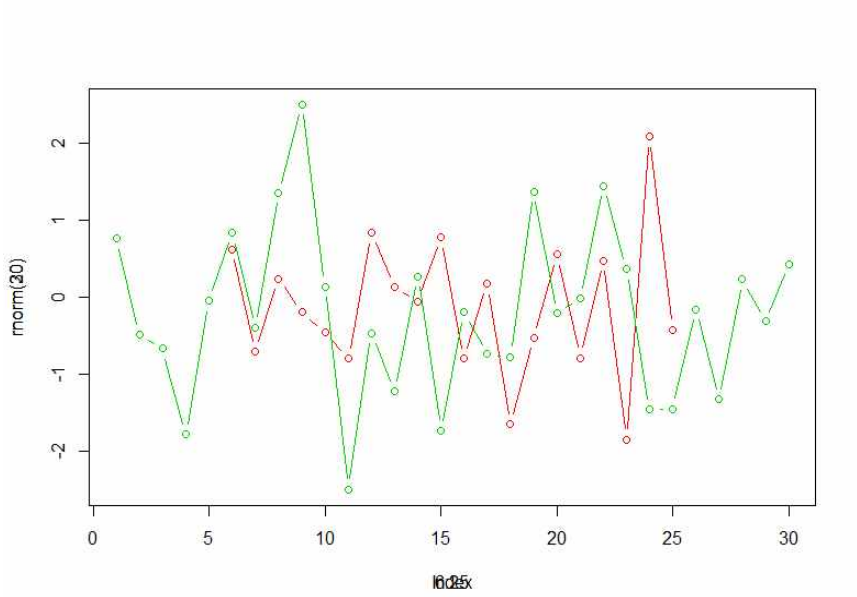
※ plot 인자 지정(2)

par 함수를 통한 파라미터 지정으로 다양한 효과를 줄 수 있음.

par(mfrow=c(2,2))



par(new=T)

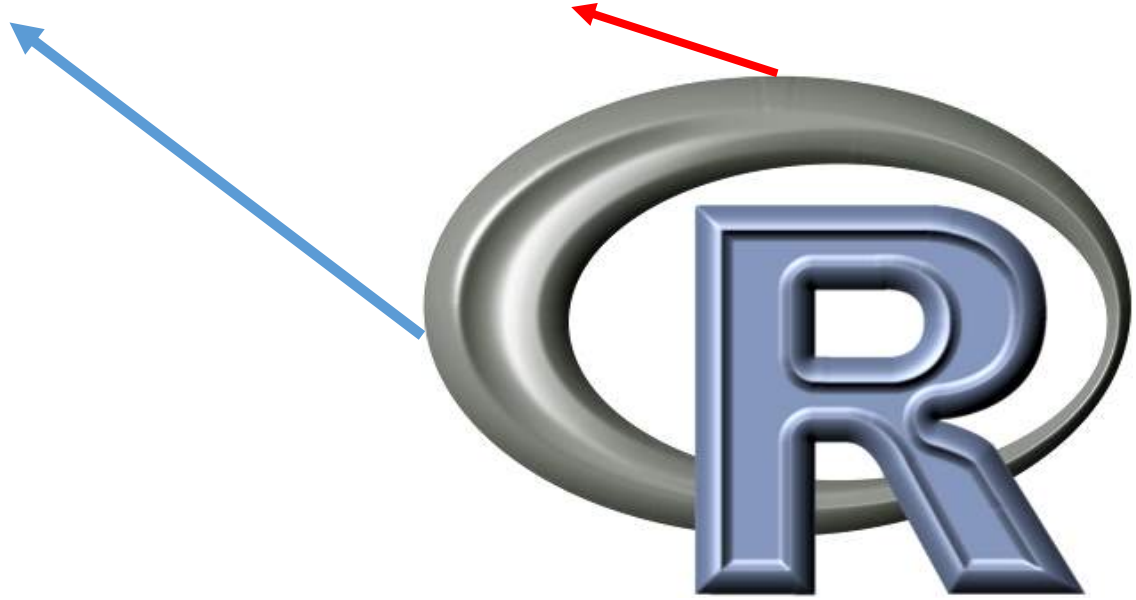


Ⅲ. Wrap-Up

기획(Planning) (우메다 사토시, “최고의 기획자는 세 번 계략을 짠다”)

이상과 현실의 갭을 메꾸는 방법과 프로세스를 찾는 것

기획력 = 발상력 + 실행력



End of Document.

감사합니다.