
R Recipes 데이터 분석

강사 : 문성민

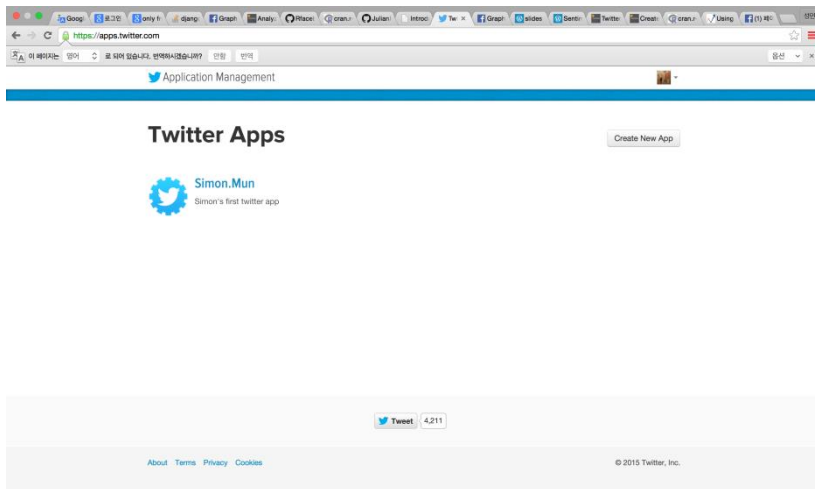
Text Analysis utilizing twitter

Text Analysis utilizing twitter

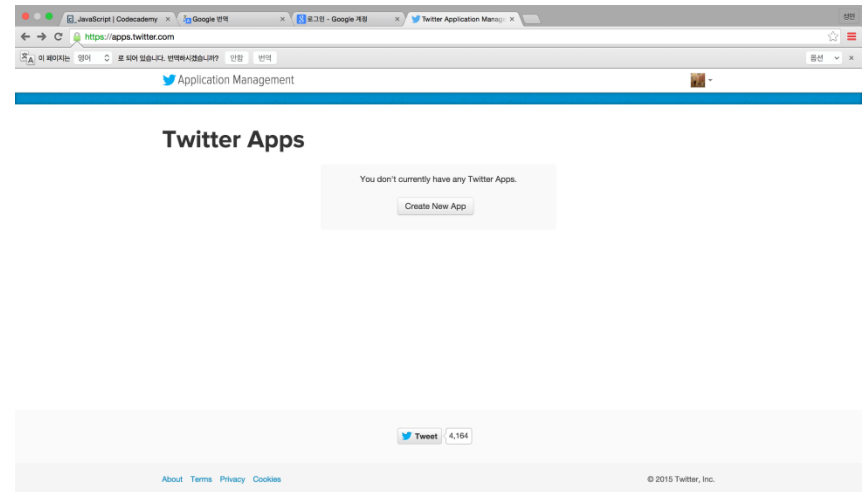
● 트위터 계정 생성

- <https://apps.twitter.com>

1) Process



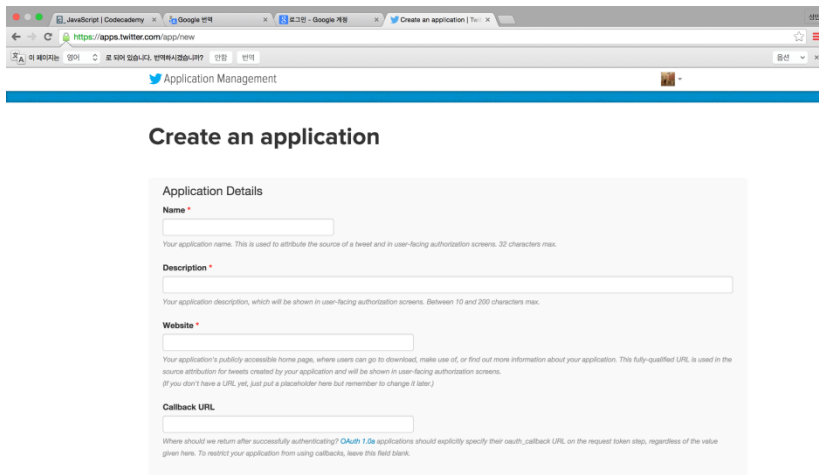
2) Process



Text Analysis utilizing twitter

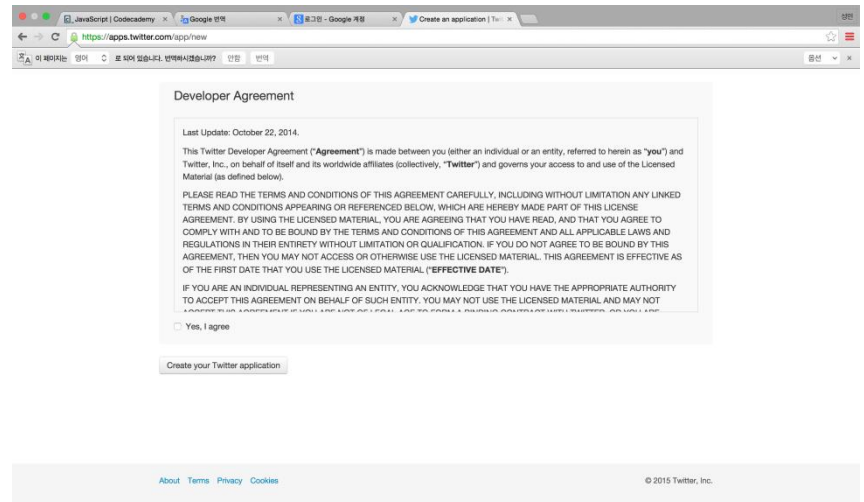
- 트위터 계정 생성
 - <https://apps.twitter.com>

3) Process



The screenshot shows the 'Create an application' page on the Twitter Developer portal. The page has a blue header with the Twitter logo and 'Application Management'. The main heading is 'Create an application'. Below it, the 'Application Details' section contains four input fields: 'Name', 'Description', 'Website', and 'Callback URL'. Each field has a small red asterisk indicating it is required. Below each field is a small line of explanatory text. The 'Name' field is labeled 'Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.' The 'Description' field is labeled 'Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.' The 'Website' field is labeled 'Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL, yet, just put a placeholder here but remember to change it later.)' The 'Callback URL' field is labeled 'Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.'

4) Process



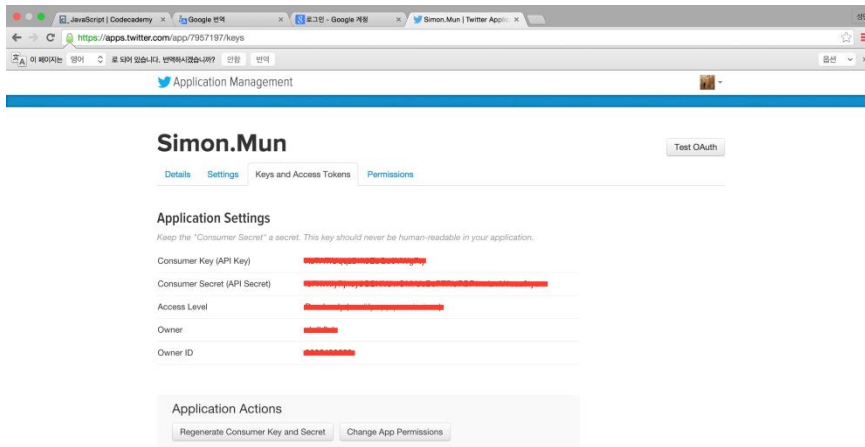
The screenshot shows the 'Developer Agreement' page on the Twitter Developer portal. The page has a blue header with the Twitter logo and 'Create an application'. The main heading is 'Developer Agreement'. Below it, the 'Developer Agreement' section contains a text area with the following text: 'Last Update: October 22, 2014. This Twitter Developer Agreement ("Agreement") is made between you (either an individual or an entity, referred to herein as "you") and Twitter, Inc., on behalf of itself and its worldwide affiliates (collectively, "Twitter") and governs your access to and use of the Licensed Material (as defined below). PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("EFFECTIVE DATE"). IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL FOR ANY PURPOSE OTHER THAN THAT FOR WHICH IT WAS PROVIDED TO YOU.' Below the text area is a checkbox labeled 'Yes, I agree'. At the bottom of the page, there is a button labeled 'Create your Twitter application' and a footer with links for 'About', 'Terms', 'Privacy', and 'Cookies', and a copyright notice '© 2015 Twitter, Inc.'

Text Analysis utilizing twitter

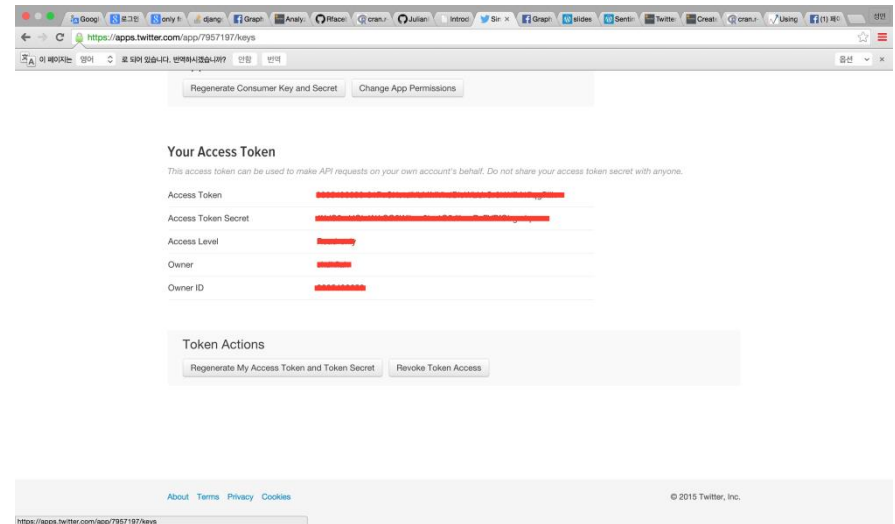
● 트위터 계정 생성

- <https://apps.twitter.com>

5) Process



6) Process



Text Analysis utilizing twitter

● 라이브러리 설치

```
> library(bitops)
> library(RCurl)
> library(RJSONIO)
> library(twitter)
> library(ROAuth)
> library(RColorBrewer)
> library(devtools)
> install_github("twitter", username="geoffjentry")
Downloading github repo geoffjentry/twitteR@master
Installing twitteR
'/Library/Frameworks/R.framework/Resources/bin/R' --vanilla CMD INSTALL \

'/private/var/folders/28/g8cf_pvx46s5phqwr6qq7jw0000gn/T/Rtmp8qGmY/devtoolscb924cc3a7ae/geoffj
entry-twitteR-563a23c' \
  --library='/Library/Frameworks/R.framework/Versions/3.1/Resources/library' \
  --install-tests

* installing *source* package 'twitteR' ...
** R
** inst
** preparing package for lazy loading
Creating a generic function for 'as.data.frame' from package 'base' in package 'twitteR'
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (twitteR)
Reloading installed twitteR

Attaching package: 'twitteR'

The following object is masked from 'package:plyr':

    id

The following objects are masked from 'package:dplyr':

    id, location

경고메시지:
Username parameter is deprecated. Please use geoffjentry/twitteR
```

GitHub에서 twitteR패키지의 최신버전을 다운로드한다.

Text Analysis utilizing twitter

- 유저 정보 입력

```
> api_key <- "[REDACTED]"
>
> api_secret <- "[REDACTED]"
>
> access_token <- "[REDACTED]"
>
> access_token_secret <- "[REDACTED]"
>
> setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)
[1] "Using direct authentication"
```

- <https://apps.twitter.com>에서 로그인 후 제공받은 api_key, api_secret, access_token, access_token_secret을 입력한다.

Text Analysis utilizing twitter

- Greece에 관련된 텍스트 1000개 크롤링

```
> Greece.tweets = searchTwitter("Greece" , n = 1000)
```

- Greece에 관련된 텍스트만 추출

```
> library(plyr)
>
> Greece.text = laply(Greece.tweets,function(t)t$getText())
```

- 모든 문자 소문자로 변환

```
> Greece.text <- tolower(Greece.text)
>
```

- Rt를 빈공간으로 바꾸기(삭제)

```
> Greece.text <- gsub("rt", "", Greece.text)
>
```

- 유저이름 삭제(@||w+)

```
> Greece.text <- gsub("@\\w+", "", Greece.text)
>
```


Text Analysis utilizing twitter

- 문장 부호 제거

```
> Greece.text <- gsub("[[:punct:]]", "", Greece.text)  
>
```

- 링크 제거

```
> Greece.text <- gsub("http\\w+", "", Greece.text)  
>
```

- 탭 제거

```
> Greece.text <- gsub("[ \\t]{2,}", "", Greece.text)  
>
```

- 시작 부분의 문자 제거

```
> Greece.text <- gsub("^ ", "", Greece.text)  
>
```

- 끝 부분의 문자 제거

```
> Greece.text <- gsub(" $", "", Greece.text)
```

Text Analysis utilizing twitter

- Corpus생성

```
>  
> Greece.text.corpus <- Corpus(VectorSource(Greece.text))
```

- Tm_map을 활용하여 Stop words 삭제

```
> Greece.text.corpus <- tm_map(Greece.text.corpus, function(x)removeWords(x,stopwords()))
```

- TermDocumentMatrix를 활용하여 수치형 데이터로 형변환

```
> myTdm <- TermDocumentMatrix(Greece.text.corpus, control = list(wordLengths = c(2, Inf)))  
>
```

Text Analysis utilizing twitter

- 10번 이상 출현한 명사 나타내기

```
> findFreqTerms(myTdm, lowfreq = 10)
```

[1] "1greek"	"2015"	"230815"	"24"	"akan"
[6] "allowed"	"anda"	"atau"	"athens"	"bailout"
[11] "bid"	"border"	"chicago"	"conservative"	"copyright"
[16] "coreoo"	"crisis"	"cross"	"dan"	"direction"
[21] "ditangan"	"form"	"gives"	"government"	"greece"
[26] "greek"	"gt"	"holidays"	"just"	"kekalkan"
[31] "last"	"lesvos"	"like"	"macedonia"	"macedonias"
[36] "malaysia"	"migrants"	"mtvhottest"	"najib"	"nasib"
[41] "new"	"night"	"now"	"npr"	"nyc"
[46] "one"	"opposition"	"otrachicago"	"pay"	"refugees"
[51] "refugeesgr"	"ringgit"	"sama"	"seaside"	"see"
[56] "segalanya"	"selamatkan"	"sepei"	"stage"	"syrian"
[61] "tsipras"	"undur"	"update"	"visit"	

- Bailout과 관련된 명사 찾기

```
> findAssocs(myTdm, "bailout", 0.25)
```

```
$bailout
```

poll	dutch	weakens	allnight	backs	deal
0.73	0.61	0.61	0.59	0.59	0.59
debate	shows	suppo	wilnews	truegreece	costs
0.59	0.56	0.56	0.55	0.48	0.39
seats	three	vvd	240815120213	coalition	grip
0.39	0.39	0.37	0.27	0.27	0.27
maintain	oligarchs	paid	pnews	yanis	
0.27	0.27	0.27	0.27	0.27	

Text Analysis utilizing twitter

- ggplot2를 이용하여 막대 그래프 그리기
- 라이브러리 불러오기, 단어에 따른 빈도수 합하기, 10이상 단어 추출

```
> library(ggplot2)
>
> termFrequency <- rowSums(as.matrix(myTdm))
>
> termFrequency <- subset(termFrequency, termFrequency >= 10)
```

- 데이터 프레임 형태로 형변환, X와 Y생성

```
> termFrequency <- as.data.frame(termFrequency)
>
> X = row.names(termFrequency)
>
> Y = termFrequency$termFrequency
```

☐ ☐ ☐ ☐

Text Analysis utilizing twitter

- 단어 빈도수에 따른 워드 클라우드 생성

- 라이브러리

```
> library(wordcloud)
```

```
>
```

- 행렬로 형변환

```
> m <- as.matrix(myTdm)
```

```
>
```

- 빈도 값 합산

```
> wordFreq <- sort(rowSums(m), decreasing = TRUE)
```

```
>
```

- 팔레트 생성

```
> pal <- brewer.pal(8, "Dark2")
```

```
>
```

☐ ☐ ☐ ☐

- ```
> wordcloud(words = names(wordFreq), freq = wordFreq, min.freq = 10, random.order = F, rot.p
r = 0.1, colors = pal)
```



---

# **Text Analysis utilizing twitter**

- Clustering Analysis**



# What is Cluster Analysis?

## ● 개념(Concept)

- N개의 관찰치를 대상으로 p개의 변수를 측정했을 때, 관측한 p개의 변수 값을 이용하여 N개의 관찰치 사이의 유사성(similarity)의 정도를 측정하여 관찰치들을 가까운 순서대로 군집화하는 통계적 분석 방법이다.
- 두 관찰치 사이의 유사성을 측정하는 여러 방법(유클리디안, 유클리디안 제곱거리, 코사인값, 상관계수, 체비셰프, 블록, 민코브스키, 커스텀거리,...) 중 가장 일반적으로 이용되는 방법은 유클리디안 거리 또는 상관계수 공식을 이용한 변수가 유사성 측정이다.
- 데이터의 특징을 나타내는 변수간에 값의 차이 혹은 단위의 차이가 클 때는 데이터를 표준화시켜 사용하는 것이 일반적이다.

## ● 유사성 관련 수식(Formula)

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

유클리디안 거리 측정 공식

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

상관계수 거리 측정 공식

# What is Cluster Analysis?

---

- 표준화 관련 수식(Formula)

$$Z_i = \frac{W_i - \bar{W}}{S_W} \quad \bar{W} = \frac{\sum_{i=1}^n W_i}{n} \quad S_W = \sqrt{\frac{\sum (W_i - \bar{W})^2}{n}}$$

- 군집화 방법의 개념(Concept)

- 계층적 군집 분석 : 각 관찰치들 사이의 유사성, 거리 행렬을 구한 뒤에 관찰치들을 가까운 순서대로 연결해 가는 방법(최단연결법, 최장연결법, 중심연결법, 평균연결법, 워드의 방법 등이 있다.)
- 비계층적 군집 분석 : 비 계층 적 군집 방법은 K-means clustering으로 불리며 관찰치들이 속할 군집의 수(K)를 미리 정한 뒤 정해진 군집으로 관찰치들을 포함시키는 방법이다.

# Text Analysis utilizing twitter

---

- 단어 빈도수에 따른 군집 분석
- removeSparseTerms를 활용하여 Corpus상의 0값 제거

```
> myTdm2 <- removeSparseTerms(myTdm, sparse = 0.95)
>
```

- 행렬로 형변환

```
> m2 <- as.matrix(myTdm2)
>
```

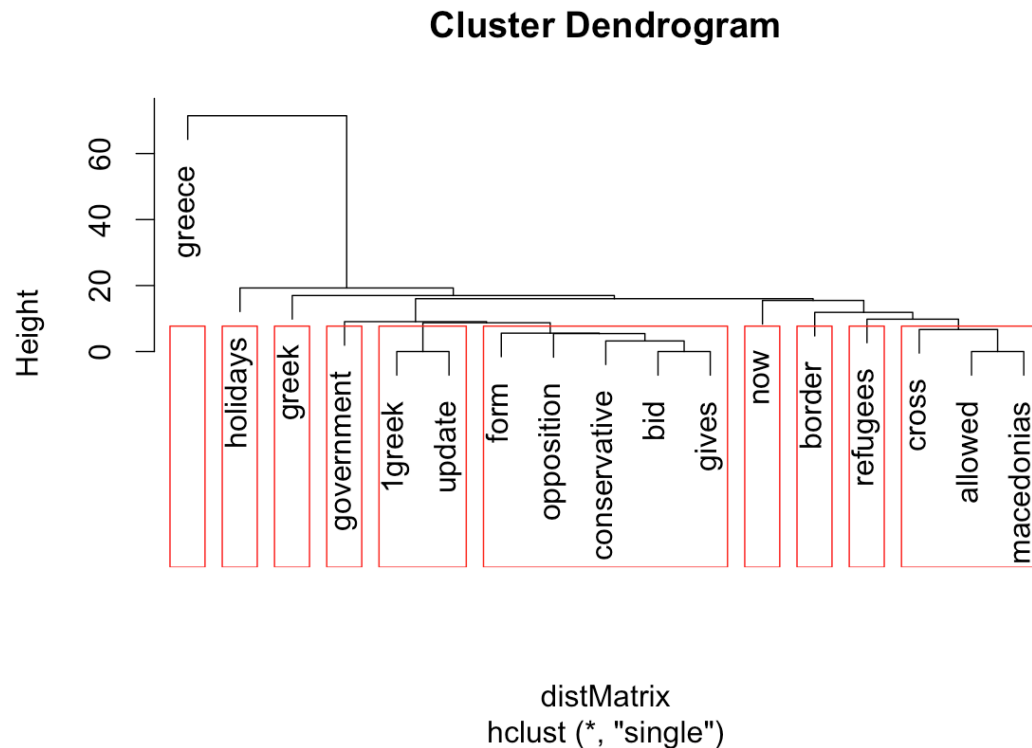
- 유클리디안 매트릭스 생성

```
> distMatrix <- dist(scale(m2),method="euclidean")
```

# Text Analysis utilizing twitter

- 유클리디안, 최단거리 연결법을 활용한 군집 수형도(덴드로그램)

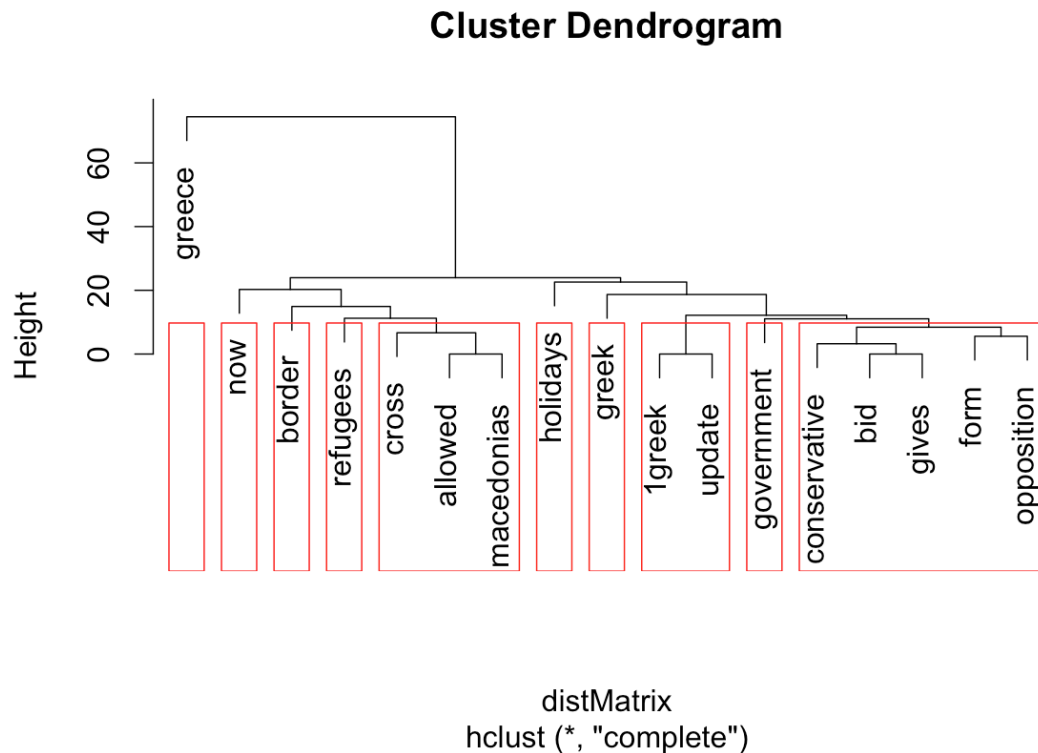
```
> single <- hclust(distMatrix, method = "single")
>
> plot(single)
>
> rect.hclust(single, k = 10)
```



# Text Analysis utilizing twitter

- 유클리디안, 최장거리 연결법을 활용한 군집 수형도(덴드로그램)

```
> complete <- hclust(distMatrix, method = "complete")
>
> plot(complete)
>
> rect.hclust(complete, k = 10)
```



---

# Regression Analysis with R

# What is Regression Analysis?

---

## ● 개념(Concept)

- 독립변수와 종속변수 간에 존재하는 연관성을 분석하기 위하여 관측된 자료에서 이들간의 함수적 관계식을 통계적으로 추정하는 방법이다.
- 독립변수의 수에 따라 단순회귀분석과 다중회귀분석으로 나뉜다.
- 회귀분석을 시행하기에 앞서 기본 가정(정규성, 등분산성, 독립성)들이 모두 충족 되어야 한다.
- 회귀분석은 잔차(측정값-실제값)의 제곱의 합을 최소로 하는 최소 제곱법을 사용한다.
- 독립변수에 의해 설명되는 종속변수의 비율 값으로 결정계수를 사용한다.
- 회귀모형의 유의성을 검증하기 위해 F값을 사용하고 회귀계수의 유의성을 검증하기 위해 T값을 활용한다.

## ● 회귀 분석 용어 정리

- 종속변수 : 분석의 대상이 되는 변수
- 독립변수 : 종속변수에 영향을 미치는 변수
- 잔차 : 추정 값과 실제 값의 차이값
- 정규성 : Q-Q plot에서 두 변수가 유사한 정도
- 등분산성 : 잔차들을 사용한 산점도에서 잔차들이 고루 퍼져 있는 정도
- 독립성 : 더빈 왓슨 값이 0~4이내이고 2일 때는 가장 독립성을 만족한다.

# What is Regression Analysis?

- 회귀모형식

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i + e_i \quad b_0 = \bar{Y} - b_1 \bar{X}, b_1 = \frac{S_{xy}}{S_{xx}}, e_i = \hat{Y}_i - Y_i$$

- 분산 분석표

| Source(요인) | Df(자유도)   | SS(제곱합) | Ms(평균제곱)                  | F                     |
|------------|-----------|---------|---------------------------|-----------------------|
| Reg(회귀)    | K-1       | SSR     | $MSR = \frac{SSR}{k-1}$   | $F = \frac{MSR}{MSE}$ |
| Error(오차)  | N-(k-1)-1 | SSE     | $MSE = \frac{SSE}{n-k-1}$ |                       |
| Total      | N-1       | SST     |                           |                       |

K=변수의 수  
N=총 데이터의 수

$$S_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i^2 - n\bar{X}^2$$

$$S_{xx} = \sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$$

$$S_{yy} = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

$$SST = S_{yy}$$

$$SSR = b_1^2 S_{xx} = b_1 S_{xy}$$

$$SSE = SST - SSR$$



# What is Regression Analysis?

- T값, F값,  $R^2$ (결정계수)값 관련 공식

$$SE(b_1), (\text{표준오차}) = \sqrt{\frac{MSE}{S_{xx}}}$$

$$T_{\text{값}} = \frac{\widehat{b_1} - 0}{SE(\widehat{b_1})} = \frac{b_1}{SE(b_1)}$$

$$F_{\text{값}} = (t_{\text{값}})^2$$

$$R^2(\text{결정계수}) = \frac{SSR}{SST}$$

- 잔차에 대한 가정사항

- 잔차들의 합은 0이다.
- 잔차들의  $X_i$ 에 의한 가중합은 0이다.
- 잔차들의  $\widehat{X}_i$ 에 의한 가중합은 0이다.
- 잔차에 의해 생성된 회귀식은 항상 평균점( $\bar{X}, \bar{Y}$ )을 지난다.

# Regression Analysis with R

## ● 예제(Example)

- 9575명의 몸무게와 체질량 지수에 대하여 기록된 데이터

## ● 변수 설명

- Weight = 몸무게에 대한 수치형 데이터(kg)
- BMI = 체질량 지수( $\frac{Weight}{Height^2}$ )

```
> getwd()
[1] "/Users/Seongmin_M/Downloads"
>
> setwd("/Users/Seongmin_M/Downloads")
>
> NHANES<-read.csv("NHANES_1.csv",head=T)
>
> head(NHANES)
 X Weight BMI
1 1 82.7 29.4
2 2 85.6 29.6
3 3 71.5 23.1
4 4 93.8 30.0
5 5 81.6 29.7
6 6 68.3 21.9
>
> str(NHANES)
'data.frame': 9575 obs. of 3 variables:
 $ X : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Weight: num 82.7 85.6 71.5 93.8 81.6 68.3 67.7 86.5 61.7 85 ...
 $ BMI : num 29.4 29.6 23.1 30 29.7 21.9 26.3 31.9 20.9 26.7 ...
```

# Regression Analysis with R

- 데이터 확인

```
> getwd()
[1] "/Users/Seongmin_M/Downloads"
>
> setwd("/Users/Seongmin_M/Downloads")
>
> NHANES<-read.csv("NHANES_1.csv",head=T)
>
> head(NHANES)
 X Weight BMI
1 1 82.7 29.4
2 2 85.6 29.6
3 3 71.5 23.1
4 4 93.8 30.0
5 5 81.6 29.7
6 6 68.3 21.9
>
> str(NHANES)
'data.frame': 9575 obs. of 3 variables:
 $ X : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Weight: num 82.7 85.6 71.5 93.8 81.6 68.3 67.7 86.5 61.7 85 ...
 $ BMI : num 29.4 29.6 23.1 30 29.7 21.9 26.3 31.9 20.9 26.7 ...
```

- Head함수와 str함수를 사용하여 데이터의 형태를 확인하여 준다.

# Regression Analysis with R

## ● 회귀식 생성

```
> lm.NHANES=lm(NHANES$BMI~NHANES$Weight)
>
> lm.NHANES
```

Call:

```
lm(formula = NHANES$BMI ~ NHANES$Weight)
```

Coefficients:

|             |                |
|-------------|----------------|
| (Intercept) | NHANES\$Weight |
| 5.3798      | 0.2618         |

- 회귀식은  $BMI = 5.3798 + 0.2618 \times \text{Weight}$ 이다.

## ● 회귀모형 검증

```
> anova(lm.NHANES)
```

Analysis of Variance Table

Response: NHANES\$BMI

|                | Df   | Sum Sq | Mean Sq | F value | Pr(>F)        |
|----------------|------|--------|---------|---------|---------------|
| NHANES\$Weight | 1    | 48813  | 48813   | 13305   | < 2.2e-16 *** |
| Residuals      | 3705 | 13593  | 4       |         |               |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

- p값이 0.05이하 이므로 회귀모형은 유의 하다.

# Regression Analysis with R

## ● 회귀계수 검증

```
> summary(lm.NHANES)
```

Call:

```
lm(formula = NHANES$BMI ~ NHANES$Weight)
```

Residuals:

| Min     | 1Q      | Median  | 3Q     | Max     |
|---------|---------|---------|--------|---------|
| -7.1021 | -1.3043 | -0.0928 | 1.2004 | 12.6110 |

Coefficients:

|                | Estimate | Std. Error | t value | Pr(> t )   |
|----------------|----------|------------|---------|------------|
| (Intercept)    | 5.37983  | 0.17969    | 29.94   | <2e-16 *** |
| NHANES\$Weight | 0.26178  | 0.00227    | 115.35  | <2e-16 *** |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.915 on 3705 degrees of freedom  
(5868 observations deleted due to missingness)

Multiple R-squared: 0.7822, Adjusted R-squared: 0.7821

F-statistic: 1.33e+04 on 1 and 3705 DF, p-value: < 2.2e-16

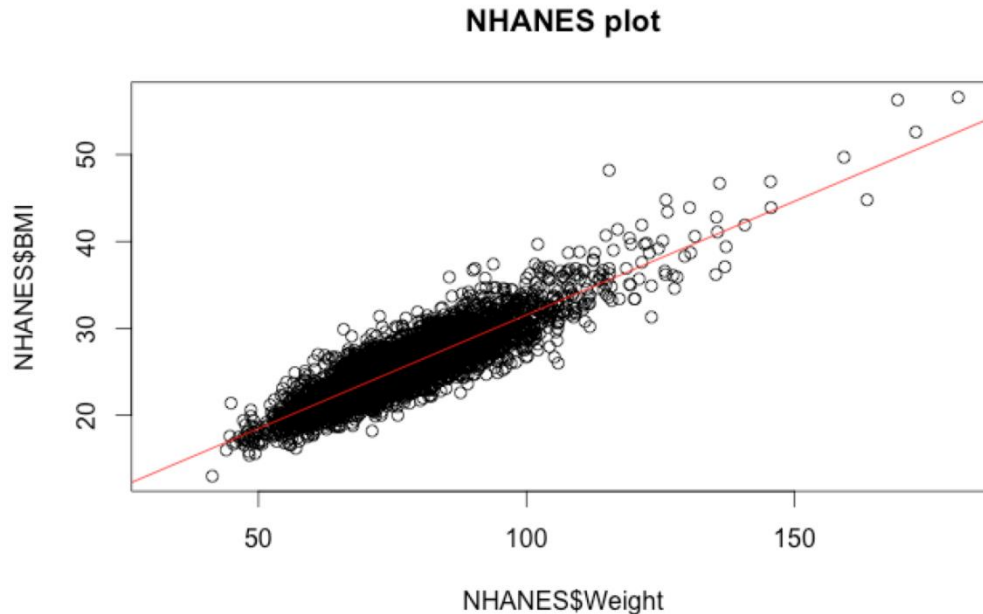
- 상수항과 독립변수의 p값이 모두 0.05이하이므로 회귀계수는 유의하다.
- 설명력은 78.21%의 설명력을 지니고 있다.

# Regression Analysis with R

- 공분산성 검증

```
> vcov(lm.NHANES)
```

|                | (Intercept)   | NHANES\$Weight |
|----------------|---------------|----------------|
| (Intercept)    | 0.0322871504  | -4.015012e-04  |
| NHANES\$Weight | -0.0004015012 | 5.150687e-06   |



- 공분산 값(절대값)이 10 이하이고 점들이 한 직선상에 일치하지 않으므로 다중 공선성은 없다고 볼 수 있다.

# Regression Analysis with R

## ● 독립성 검증

```
> install.packages("lmtest")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/lmtest_0.9-33.tgz'을 시도합니다
Content type 'application/x-gzip' length 266752 bytes (260 Kb)
URL을 열었습니다
=====
downloaded 260 Kb

The downloaded binary packages are in
/var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpXqG9Ju/downloaded_packages
> library(lmtest)
필요한 패키지를 로딩중입니다: zoo

다음의 패키지를 부착합니다: 'zoo'

The following objects are masked from 'package:base':

 as.Date, as.Date.numeric

다음의 패키지를 부착합니다: 'lmtest'

The following object is masked from 'package:RCurl':

 reset

>
> dwtest(lm.NHANES)

Durbin-Watson test

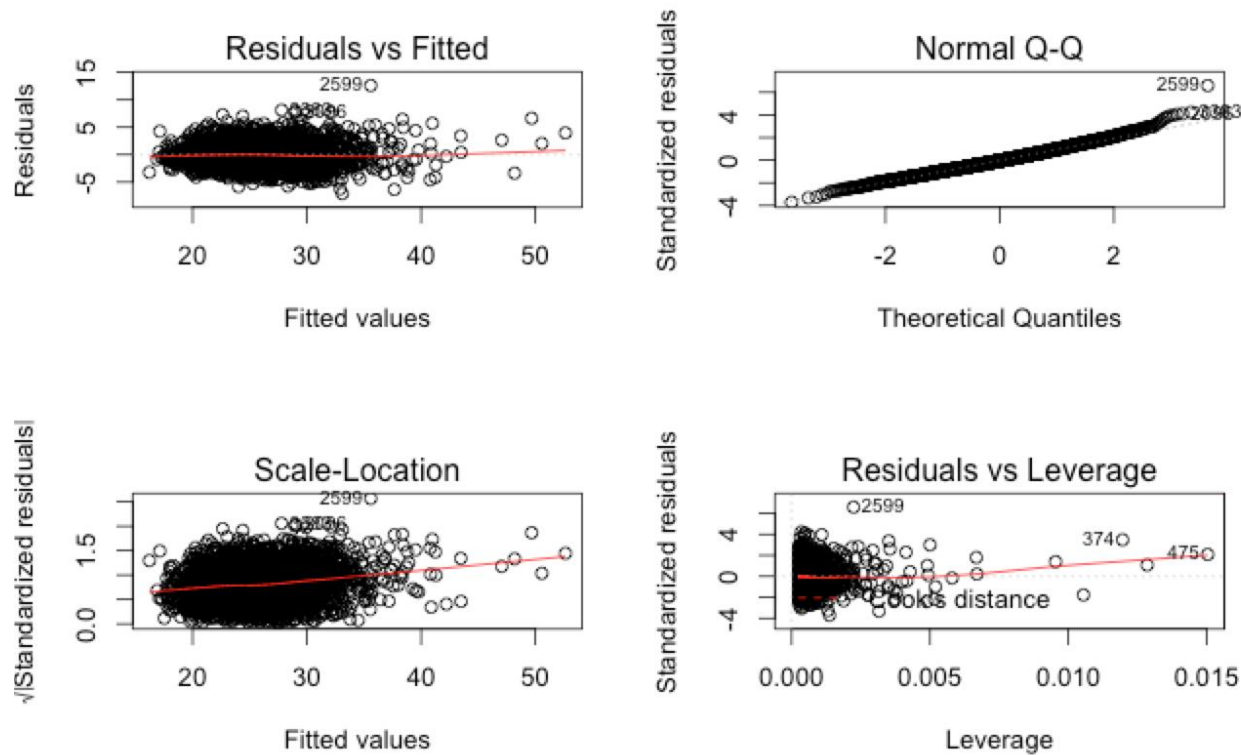
data: lm.NHANES
DW = 1.9466, p-value = 0.05209
alternative hypothesis: true autocorrelation is greater than 0
```

- Durbin-watson값이 1.9466이므로 독립성을 충족한다.

# Regression Analysis with R

## ● 등분산성, 정규성 검증

```
> par(mfrow=c(2,2))
>
> plot(lm.NHANES)
```



- 잔차가 상하에 고루 분포하므로 등분산성이 만족한다고 할 수 있다.
- QQ도표는 직선의 형태를 띠므로 정규성은 만족한다.



# Regression Analysis with R

## ● 이상치 확인

```
> install.packages("car")
```

URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/car\_2.0-25.tgz'을 시도합니다

Content type 'application/x-gzip' length 1386779 bytes (1.3 Mb)

URL을 열었습니다

=====

downloaded 1.3 Mb

The downloaded binary packages are in

/var/folders/28/g8cf\_pvx46s5phqgwr6qq7jw0000gn/T//RtmpzB2XaL/downloaded\_packages

```
> library(car)
```

경고메시지:

패키지 'car'는 R 버전 3.1.3에서 작성되었습니다

```
>
```

```
> outlier.test(lm.NHANES)
```

|      | rstudent | unadjusted | p-value   | Bonferonni | p |
|------|----------|------------|-----------|------------|---|
| 2599 | 6.629329 | 3.8603e-11 | 1.431e-07 |            |   |

- 2599번이 이상치 임을 확인 할 수 있으며 표에서도 나타내 주고 있다.

# Regression Analysis with R

## ● 이상치 확인

```
> NHANES[2599,]
 X Weight BMI
2599 2599 115.4 48.2
>
> lm.NHANES$fitted[2599]
 2599
35.58903
>
> lm.NHANES$residuals[2599]
 2599
12.61097
```

- 이상치의 BMI는 48.2이고 Weight는 115.4이다.
- 회귀식을 대입하여 BMI를 구한 결과 35.58903이 나올 것으로 예상 되었고 이는 실제값과 많은 차이가 난다.
- 2599번의 잔차는 12.61097 로 매우 높다.
- 그러므로 이상치를 빼고 다시 분석 할 필요가 있다.

## ● 결론

- 계수와 모형, 독립성, 등분산성, 정규성을 만족하므로 회귀 모형을 사용할 수 있다.

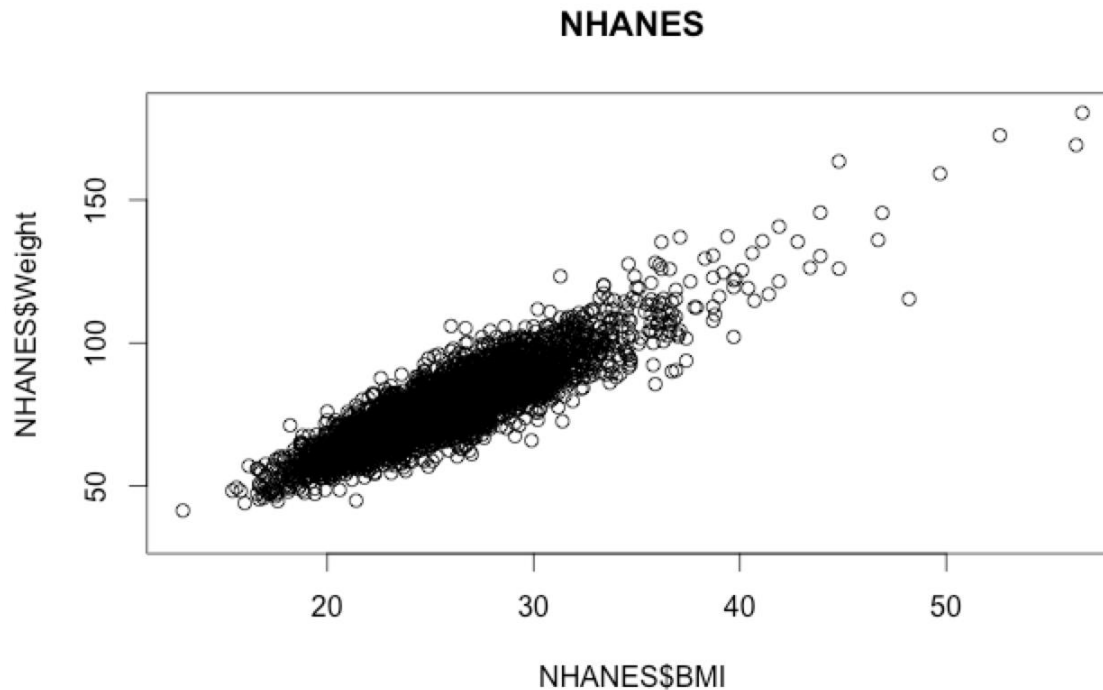
---

# Hebin visualization with R

# Hebin visualization with R

- 일반 시각화(산점도)

```
> par(mfrow=c(1,1))
>
> plot(NHANES$BMI,NHANES$Weight, main="NHANES")
```

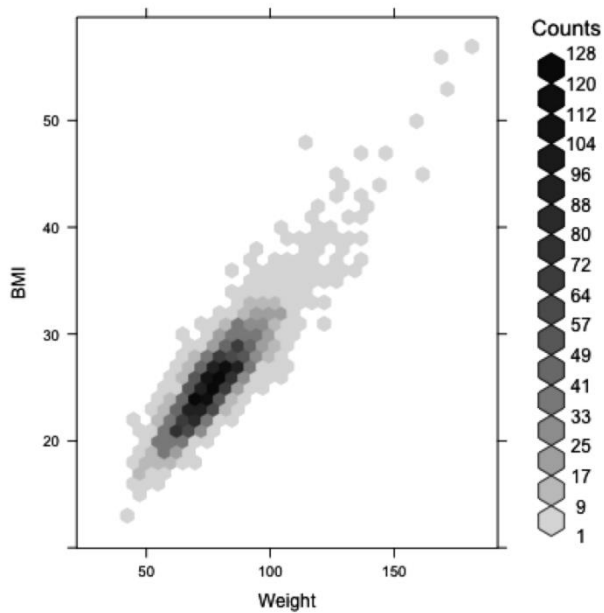


- 데이터의 수가 많고 겹치는 부분이 많아 데이터의 특성을 파악하기 힘들다.

# Hebin visualization with R

- Hebin패키지를 활용한 시각화

```
> install.packages("hebin")
Warning in install.packages :
 package 'hebin' is not available (for R version 3.1.2)
>
> library(hexbin)
>
> hexbinplot(BMI ~ Weight, data=NHANES)
```

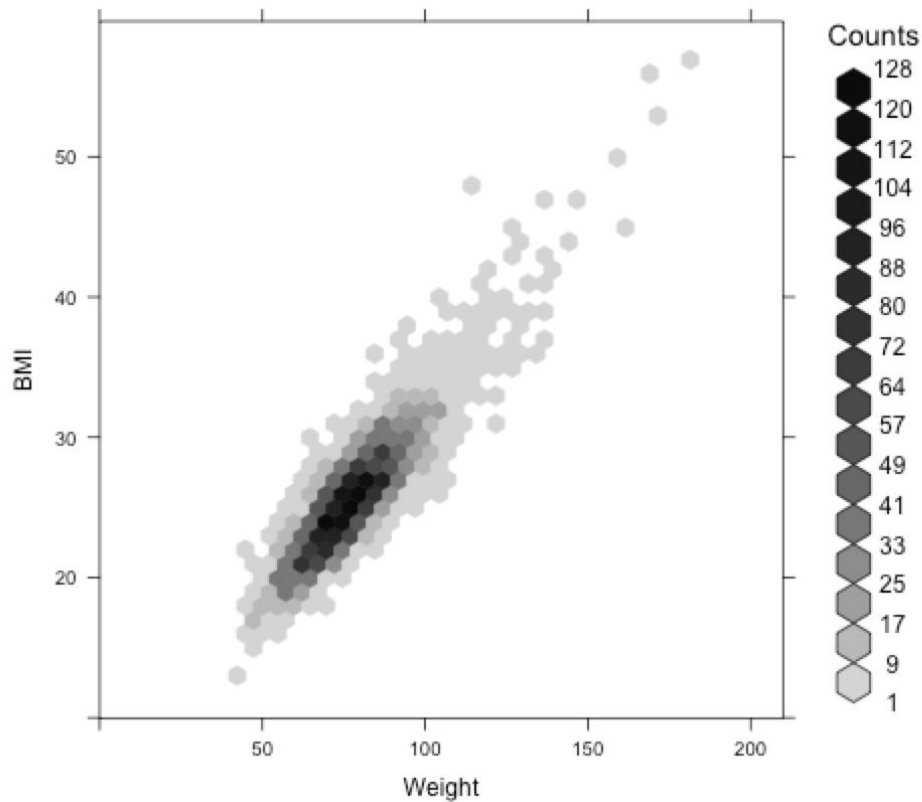


- 데이터가 겹치는 부분에 명도를 달리하여 표시하였다.

# Hebin visualization with R

- 창크기 확대

> hexbinplot(BMI ~ Weight, data=NHANES, xlim=c(0,210))

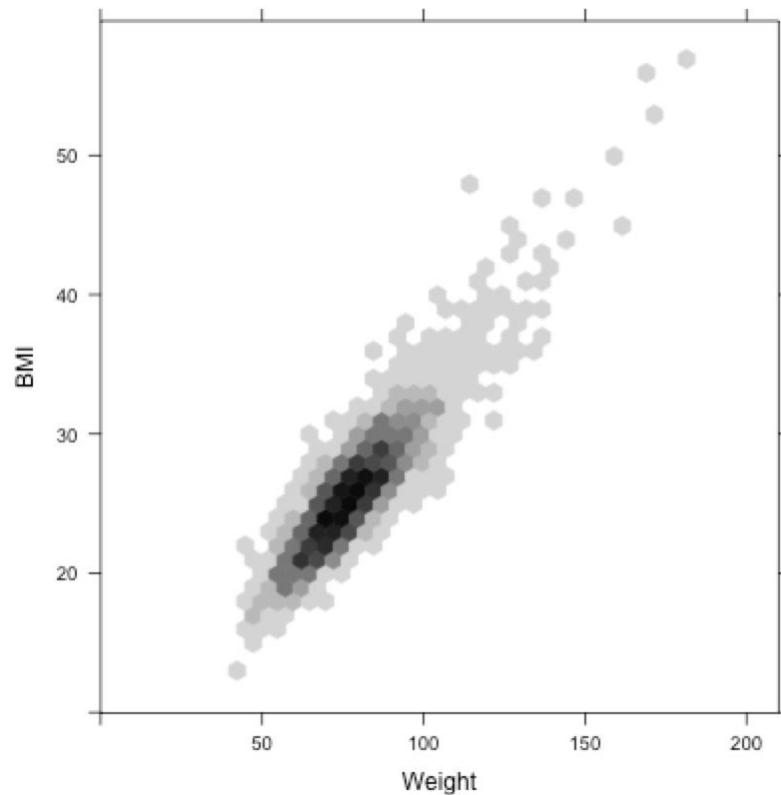


- 창의 크기를 확대하였다.

## Hebin visualization with R

- 범례 삭제

```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, data=NHANES)
```

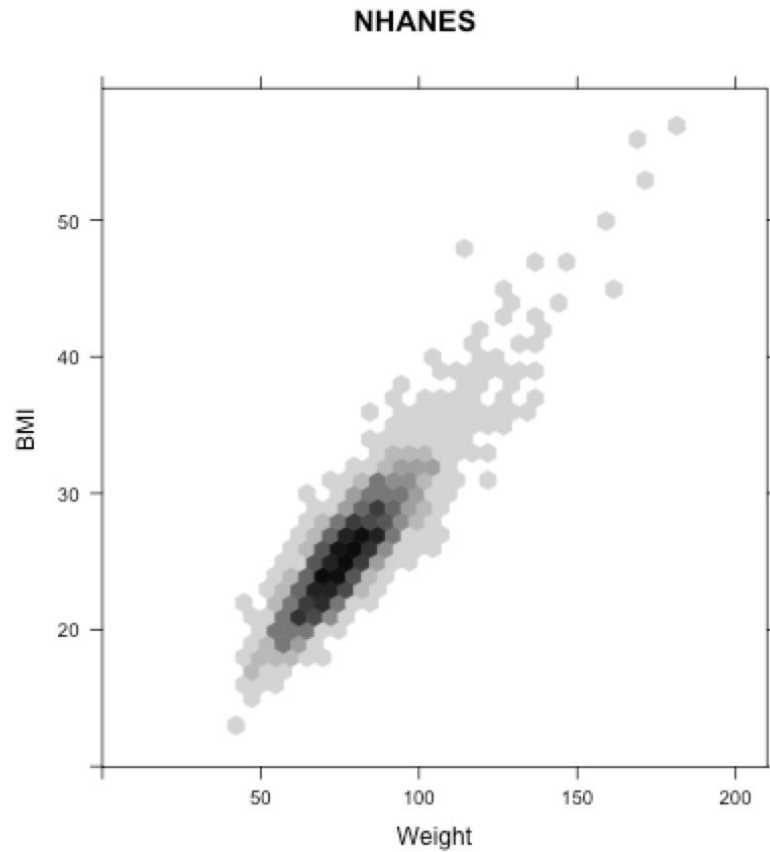


- 범례를 삭제하였다.

# Hebin visualization with R

- 제목 삽입

```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, data=NHANES, main="NHANES")
```



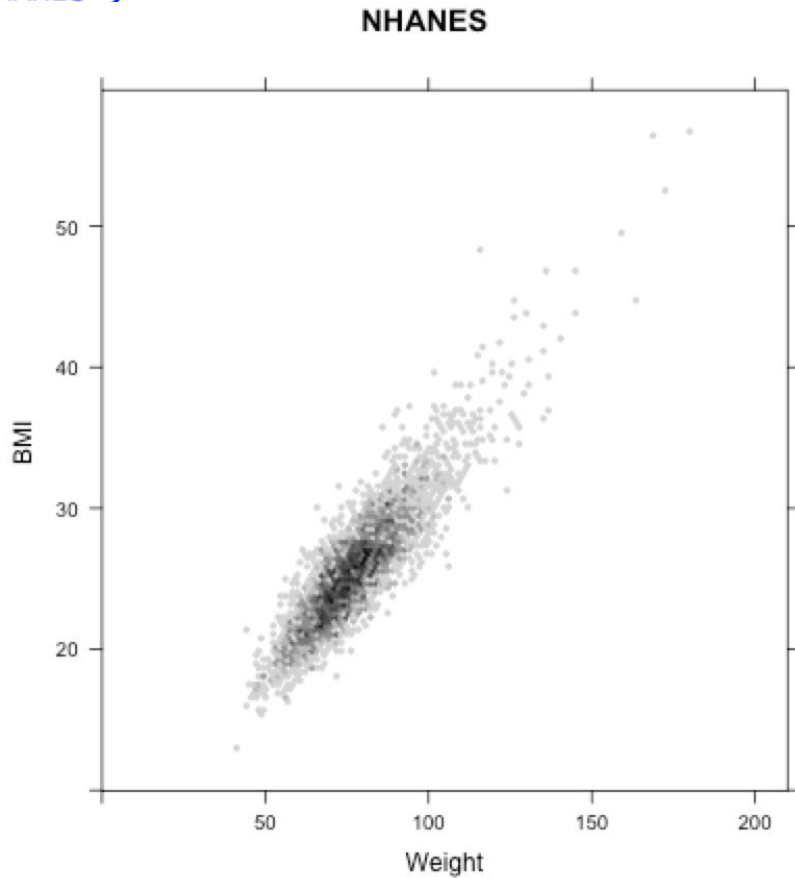
- 제목을 삽입하였다.



# Hebin visualization with R

- 점크기 조정

```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, xbins=100, data=NHANES, main="NHANES")
```

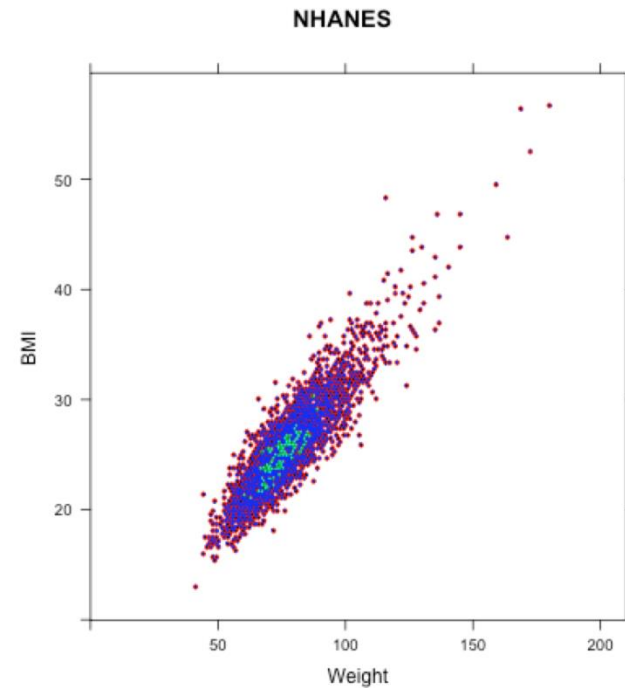
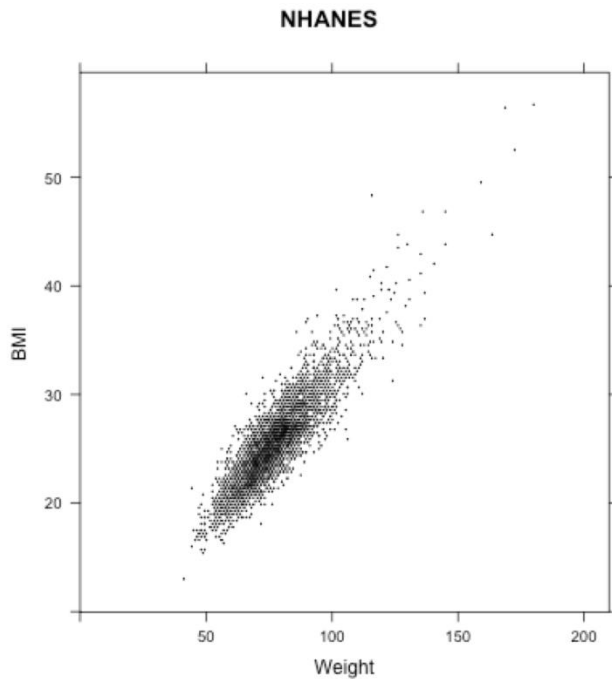


- 점의 크기를 축소하였다.

# Hebin visualization with R

## ● 스타일 조정

```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, xbins=100, data=NHANES, style
= "lattice",main="NHANES")
>
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, xbins=100, data=NHANES, style
= "nested.centroids",main="NHANES")
```



- 스타일 변경에 따라 표현법이 상이해 지는 것을 확인 할 수 있다.