# R

WEEK 4

KHU **SNS**
HAN JAEYOON

Machine Learning

Data Visualization

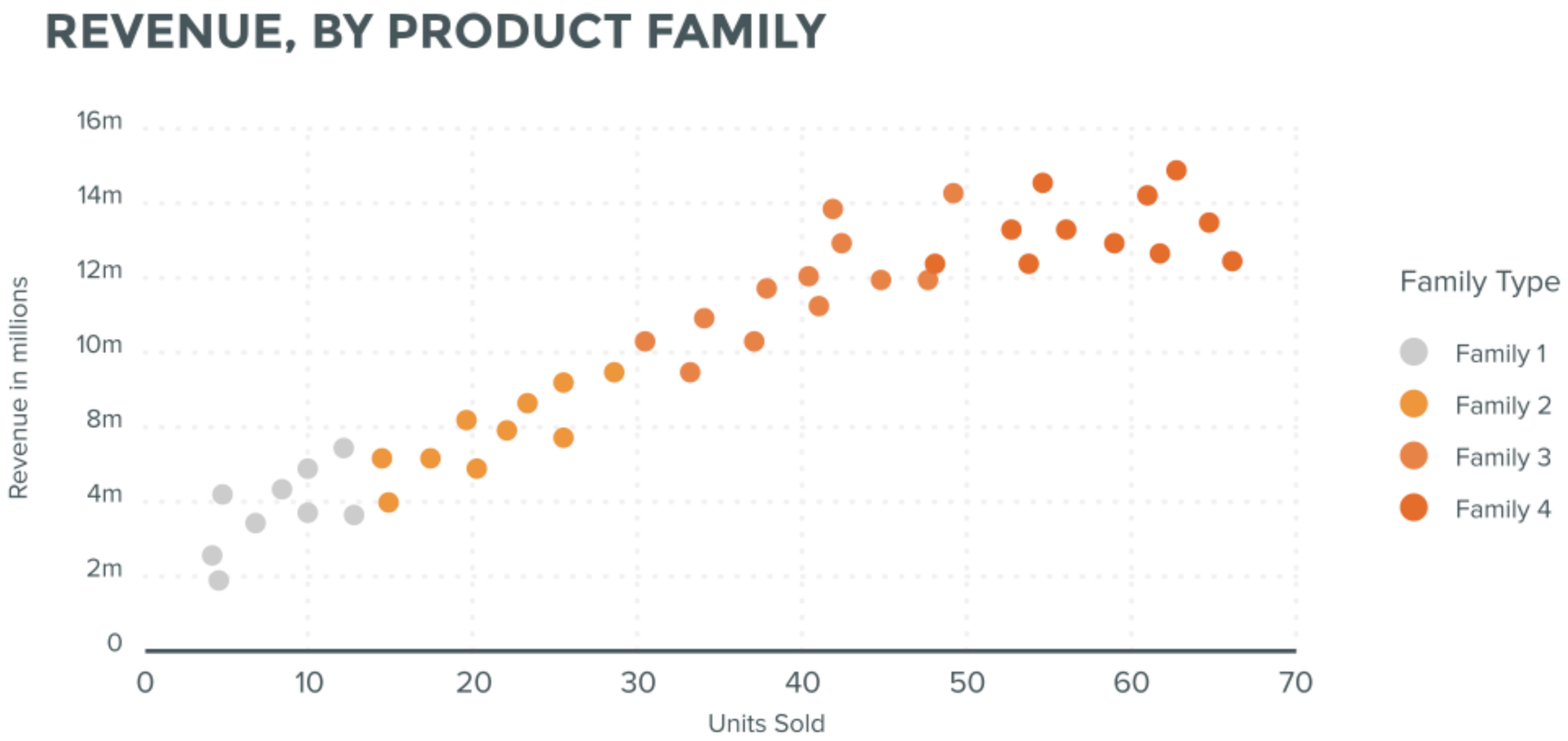
Data Manipulation

Control Statements

Data Types

R Basic Syntax

# DATA VISUALIZATION II




# Scatter Plot

# Scatter Plot

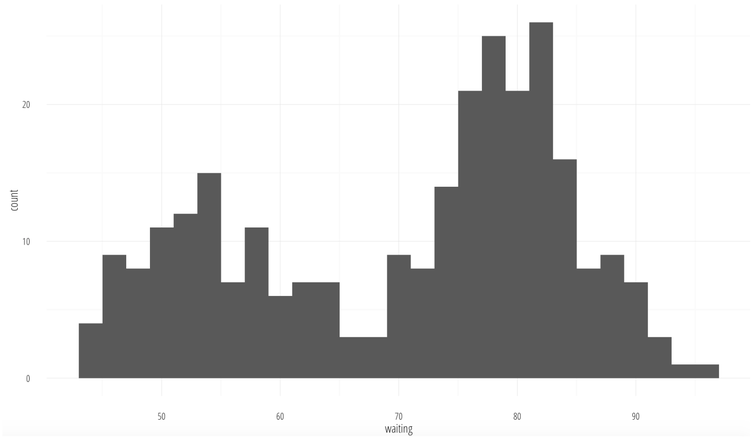
직교 좌표계를 이용해 **두 개 변수 간의 관계**를 나타내는 방법


0
6
12
18
24
0
1
7
13
19
25
+
2
8
14
20
*
*
-
3
9
15
21
.
.
|
|
4
10
16
22
o
O
%
%
5
11
17
23
O
O
#
#

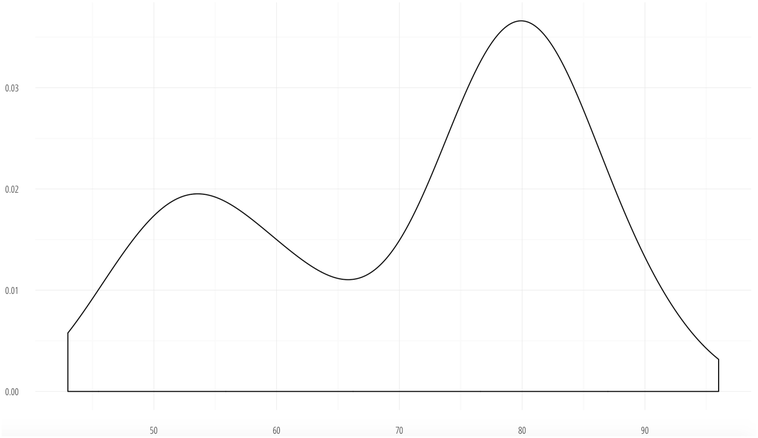
# Data Distribution

데이터의 **분포**를 나타내는 방법은 다양하다.

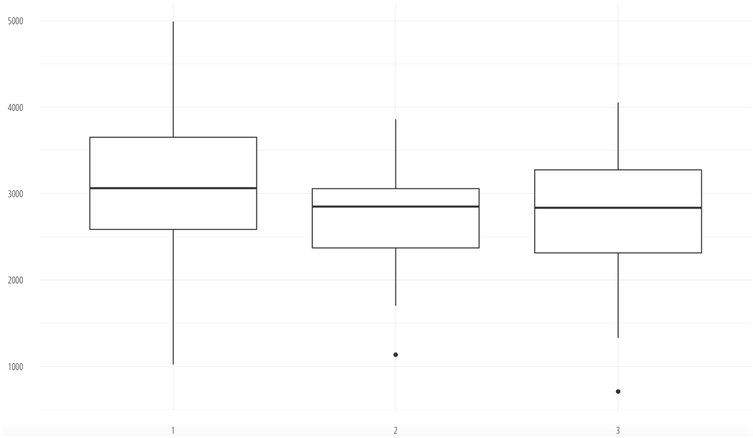
# Data Distribution




Histogram

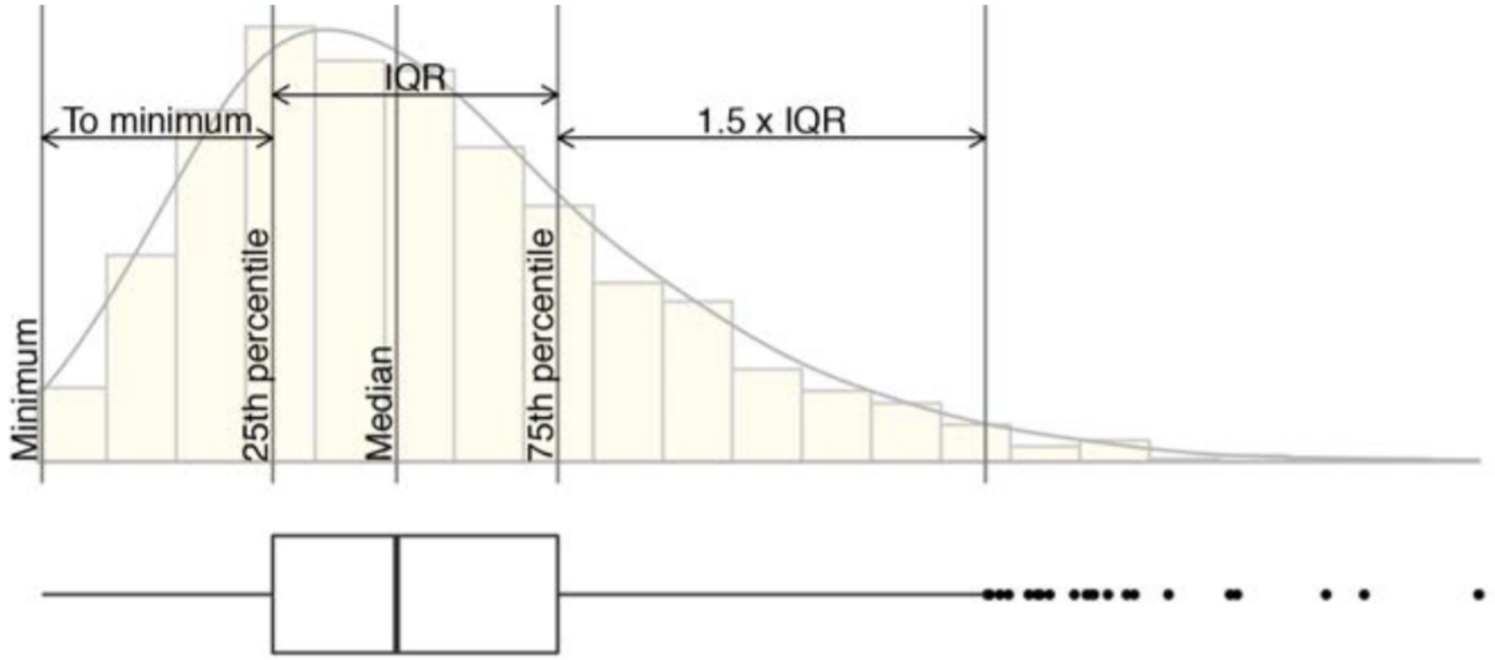



Density curve




Box plot

# Box plot


IQR
To minimum
1.5 x IQR
Minimum
25th percentile
Median
75th percentile

# Draw Maps

지도를 다뤄보자.

# Machine Learning

Machine Learning

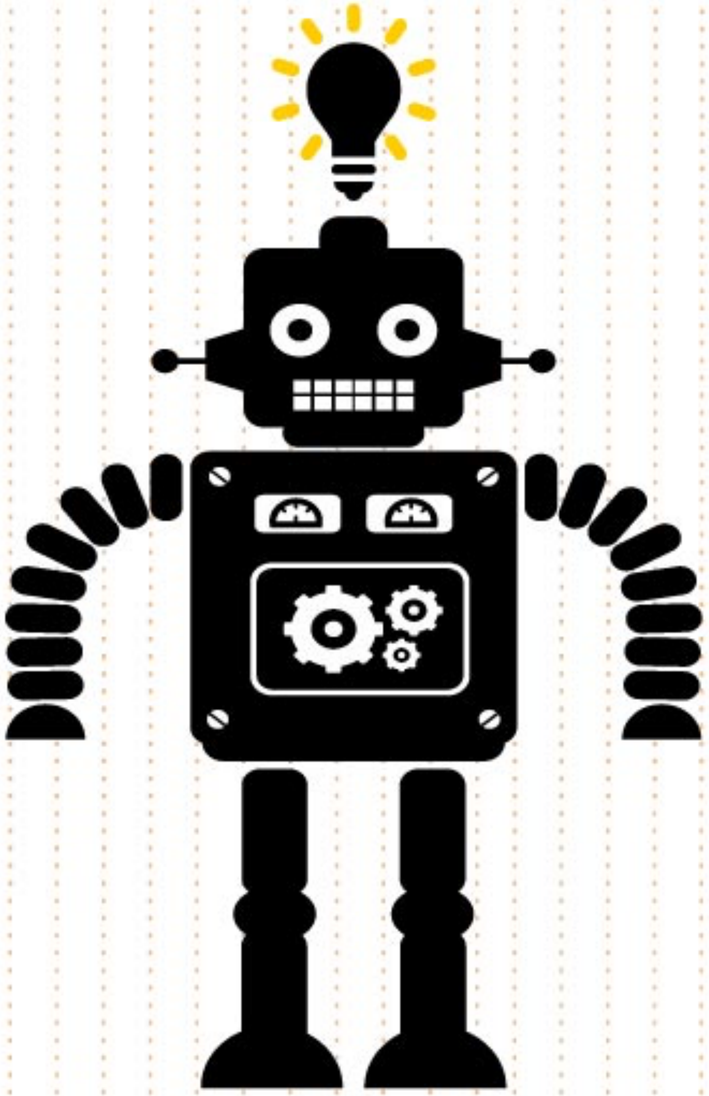
Data Visualization

Data Manipulation

Control Statements

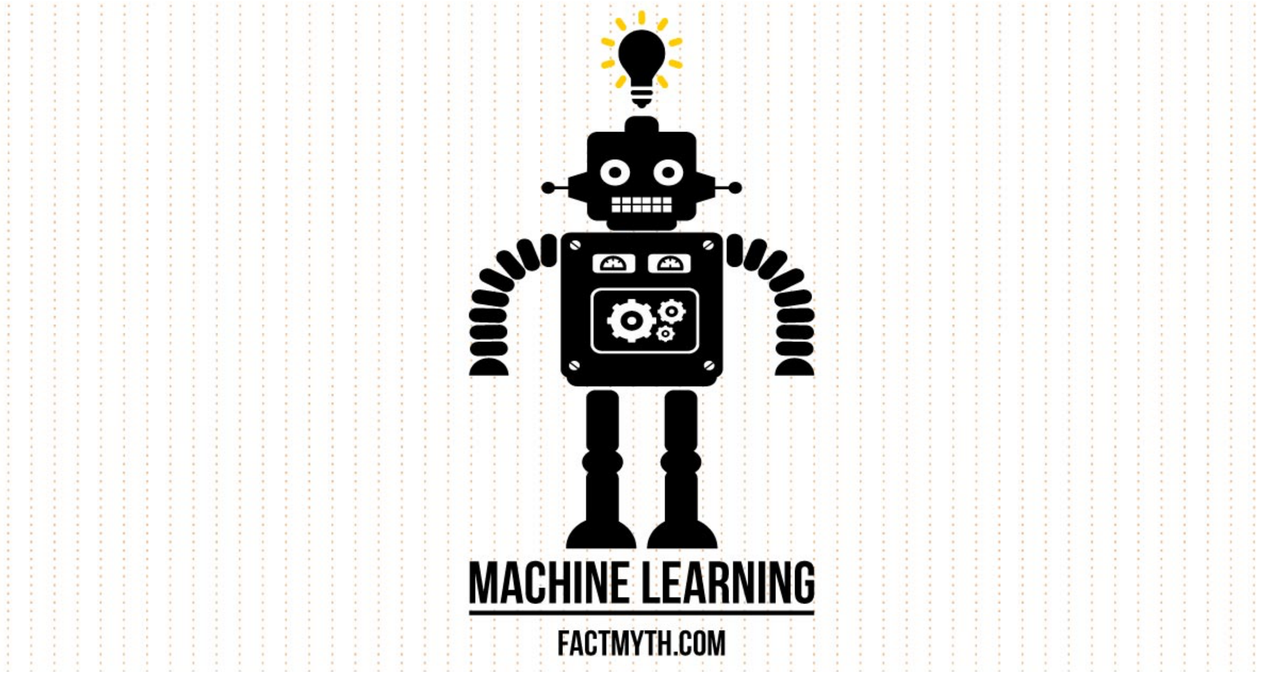
Data Types

R Basic Syntax



MACHINE LEARNING

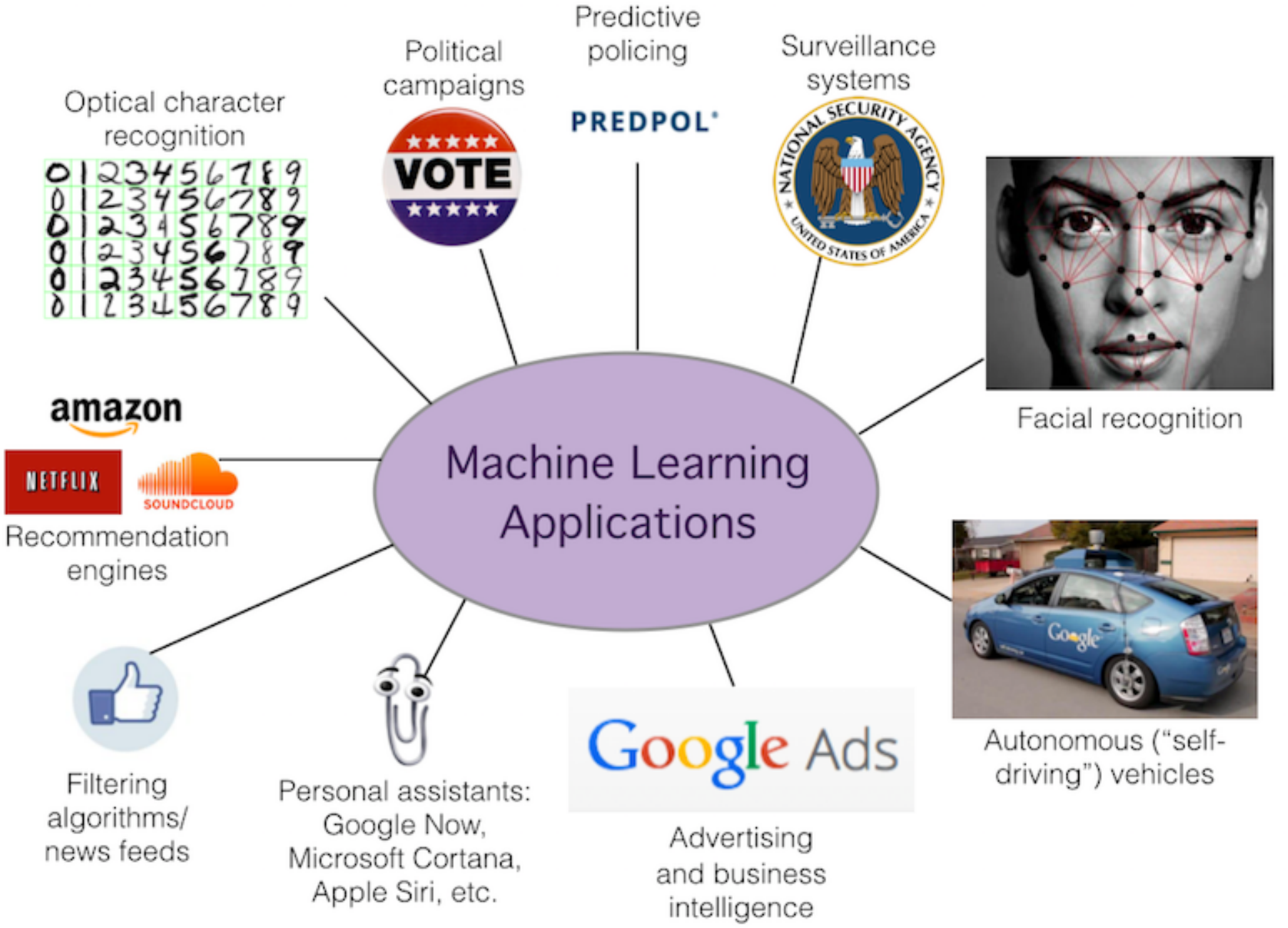
FACTMYTH.COM




**컴퓨터**에게 직접 답을 알려주지 않고,

**데이터**를 통해 컴퓨터가 **학습**을 하여

**문제를 해결**하도록 하는 것


Machine Learning Applications
Optical character recognition
Political campaigns
VOTE
Predictive policing
PREDPOL
Surveillance systems
NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA
Facial recognition
amazon
NETFLIX
SOUNDCLOUD
Recommendation engines
Filtering algorithms/ news feeds
Personal assistants: Google Now, Microsoft Cortana, Apple Siri, etc.
Google Ads
Advertising and business intelligence
Google
Autonomous ("self-driving") vehicles

# 예측?

# 관찰?

# 예측?

## Supervised Learning

가지고 있는 데이터에 **이름**이 붙여져 있다.

위 사진들을 학습하고 새로운 사진들에서 답을 찾는다.

관찰?

# 예측?

## Supervised Learning

학습 데이터 : 테스트 데이터

7 : 3

관찰?

# 관찰?

## Unsupervised Learning

동물들이 무슨 동물인지는 모른다.

단지 비슷해보이는 것들끼리 묶으면 된다.

예측?


INSIGHTS
SPEED
SQL
STORAGE
IT
AZURE
BYOD?
CLOUD
SECURITY
BIG DATA
PATTERNS
DEVELOPERS
SOCIAL
APPS
PLATFORM
WEB
MAGIC
PEOPLE
MEMORY
DEVICES
THINGS PEOPLE CARE ABOUT MOST
INTERNET OF THINGS

# 방법은 다양하다.

2주에 걸쳐서

그나마 간단한 기계학습을

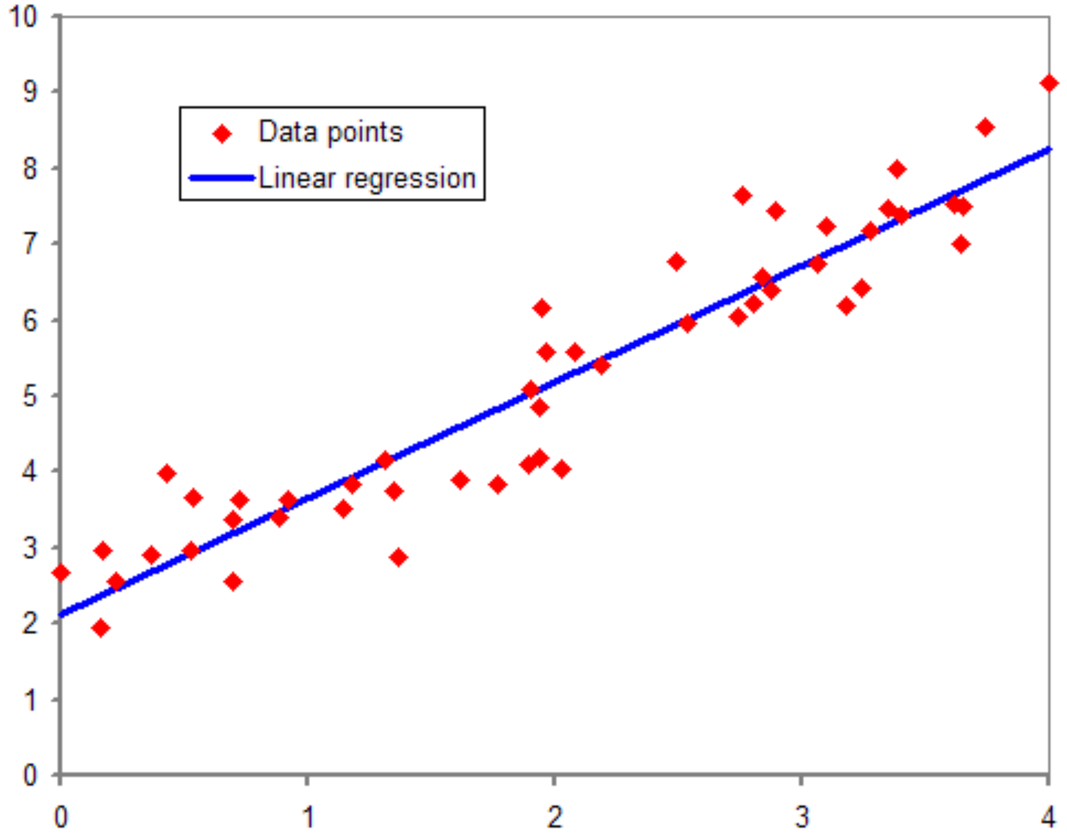
접해보자.

# Linear Regression

선형회귀법

# 선형회귀법
회기 아니다

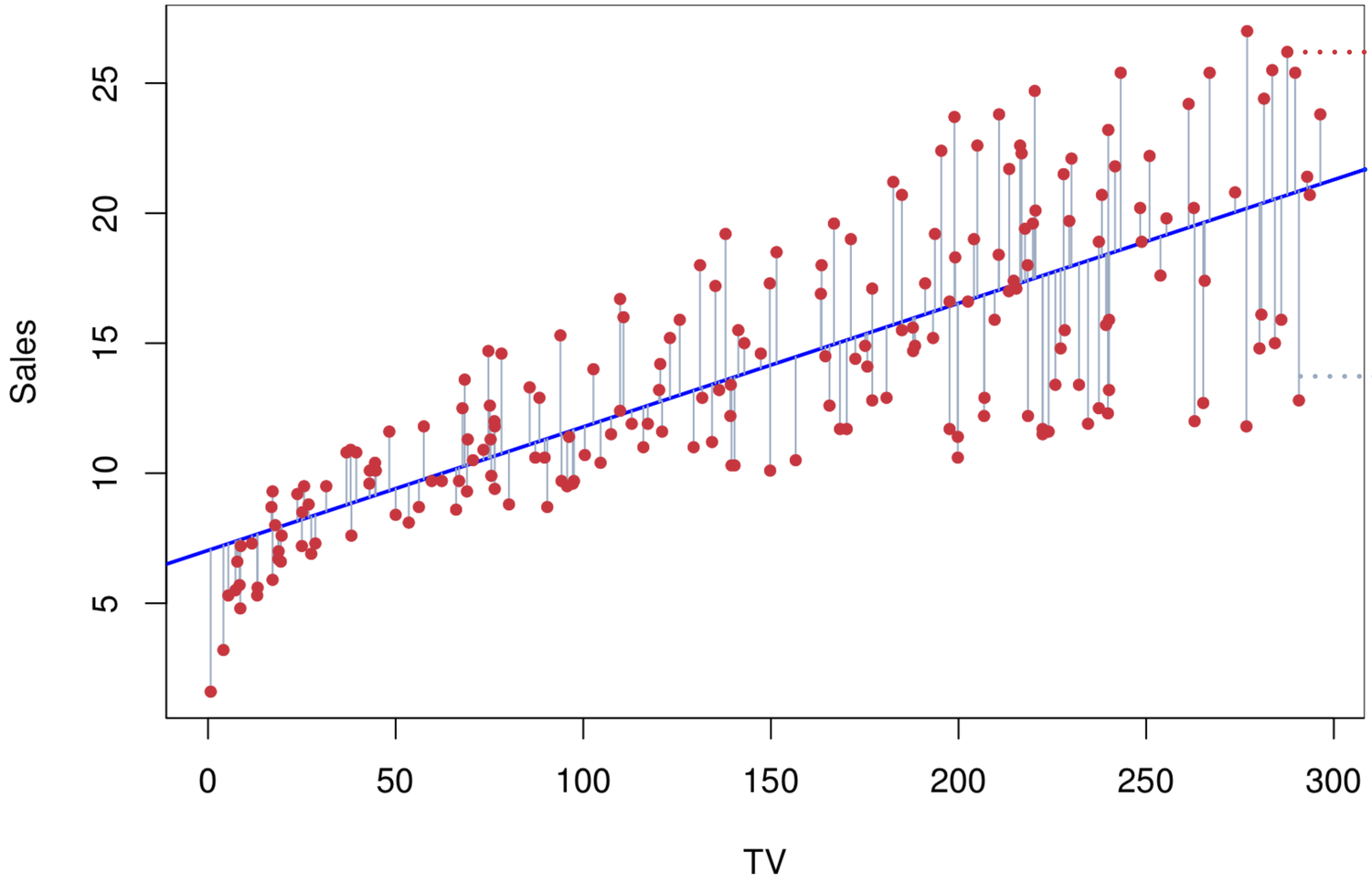
# Linear Regression
## 선형회귀법




종속변수와 한 개 이상의 독립변수 사이의 선형 상관관계를 모델링 하는 기법

# Linear Regression
## 선형회귀법

# Linear Regression
## 선형회귀법




Data Point

Fitted Value /
Regression Line

Residual

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2$$

위의 식이 **최소**가 되는 회귀선을 찾아낸다.

독립변수 : TV 시청량
종속변수 : 광고 판매금액

# Linear Regression
## 선형회귀법


https://ko.wikipedia.org/wiki/선형_회귀

# Linear Regression
## 선형회귀법

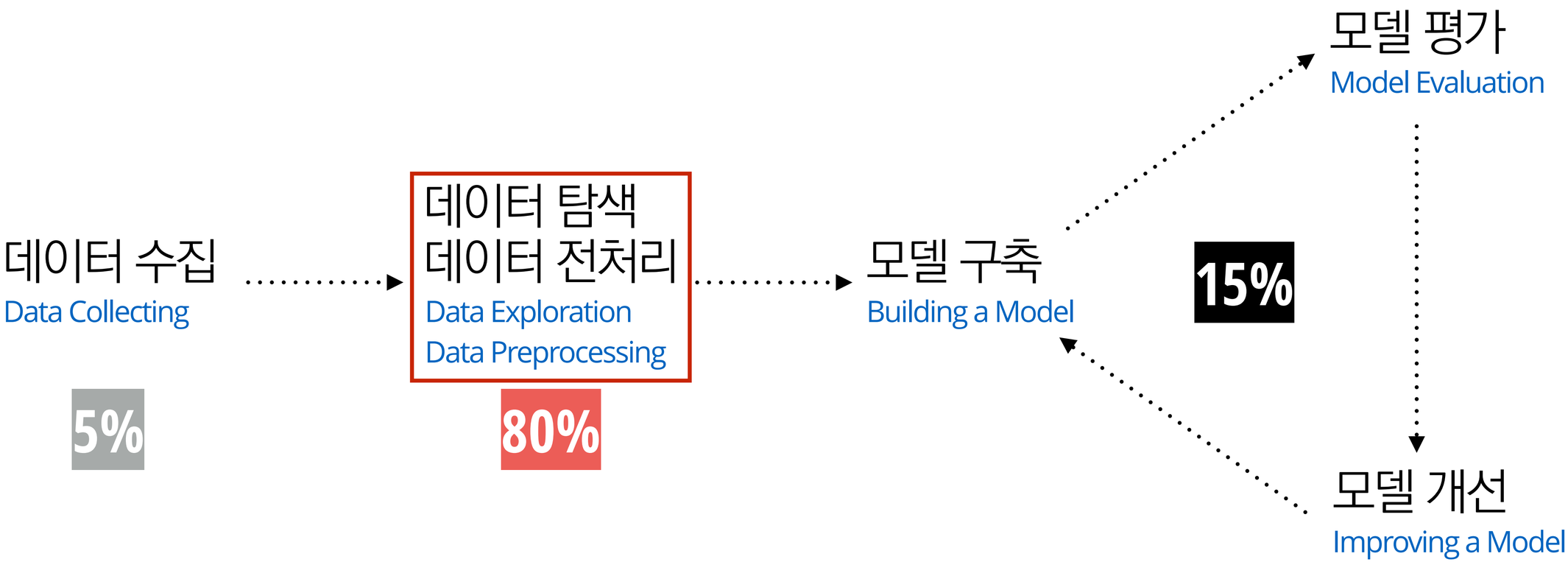
결국, 잔차값을 모두 **계산**해야 된다.

**귀찮다.**
**컴퓨터 시키자.**

# Process


데이터 수집
Data Collecting
데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing
모델 구축
Building a Model
모델 평가
Model Evaluation
모델 개선
Improving a Model

# Process


데이터 수집
Data Collecting
5%
데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing
80%
모델 구축
Building a Model
15%
모델 평가
Model Evaluation
모델 개선
Improving a Model

# Data Collecting

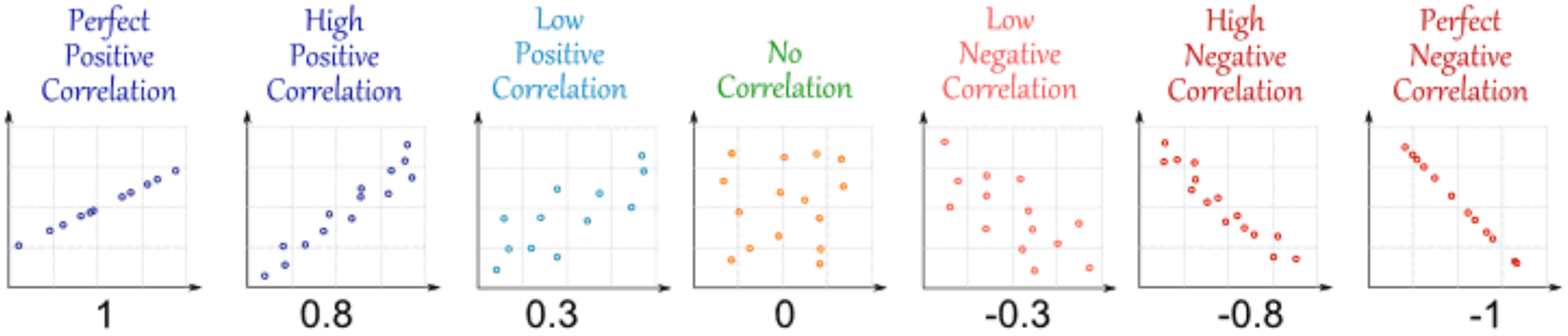

의료비 예측

Predicting Medical Expenses

| | |
|---|---|
| age | 제1순위 보험금 수령인의 나이 |
| sex | 성별 |
| bmi | 신체 용적 지수(BMI, Body Mass Index) |
| children | 보험에서 보장하는 자녀의 수 |
| smoker | 규칙적인 흡연 여부 |
| region | 거주지 (북동, 남동, 남서, 북서) |
| **charges** | **의료비** |

# 상관관계

Correlation



# 다중공선성

## Multicollinearity

회귀분석에서 독립변수들 간에 **강한 상관관계**가 나타나는 문제

회귀계수 추정치의 신뢰성과 안정성에 문제를 발생시킨다.

```
Call:
lm(formula = formula, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11304.9  -2848.1   -982.1   1393.9  29992.8 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348    
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769    
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *  
regionsouthwest   -960.0      477.9  -2.009 0.044765 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7494 
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

**잔차에 대한 정보**

**변수에 대한 회귀계수**
별의 개수에 주목

**모델의 설명력**

```
Call:
lm(formula = formula, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11304.9  -2848.1   -982.1   1393.9  29992.8 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348    
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769    
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *  
regionsouthwest   -960.0      477.9  -2.009 0.044765 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7494 
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

**잔차에 대한 정보**

잔차가 굉장히 넓게 분포되어 있다.

기존 데이터를 보면
당연한 결과임을 알 수 있다.

```
Call:
lm(formula = formula, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

* 이 많을 수록

해당 변수는 **유의한 변수**가 된다.

일반적으로 **0.05 미만**이 좋음

**변수에 대한 회귀계수**
별의 개수에 주목

```
Call:
lm(formula = formula, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

R-squared 값이 0.7509

= 모델이 설명하는 종속변수가 75%

**모델의 설명력**

# 회귀 모형을 개선해보자.

:)