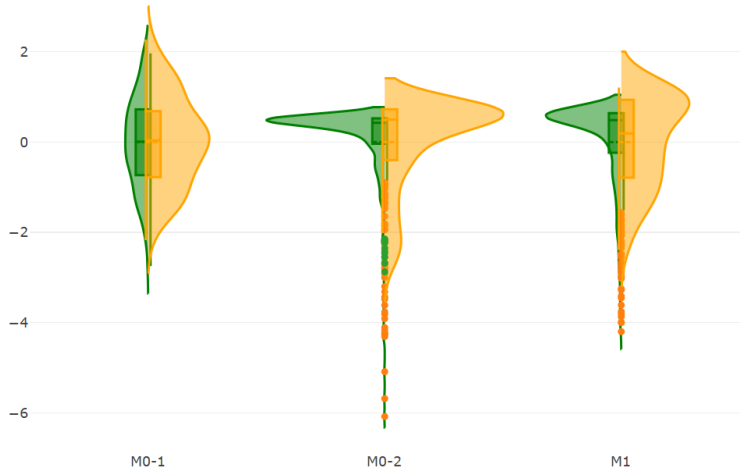
빅데이터 분석 중고급 과정

#7 EDA & Data Visualization

with One Point Tutorial

feat. R

danielwooseok.song@gmail.com



빅데이터 분석 중고급 과정

#6 **Data Manipulation** with One Point Tutorial feat. R

Agenda

1. Concepts of Visualization
2. EDA
3. Data Visualization
4. One Point Tutorial V - matplotlib
5. One Point Tutorial VI - seaborn
6. One Point Tutorial VII - folium
7. One Point Tutorial VIII - pyecharts

danielwooseok.song@gmail.com

1. Concepts of Visualization | 어떤 것을 시각화 할 것인가?

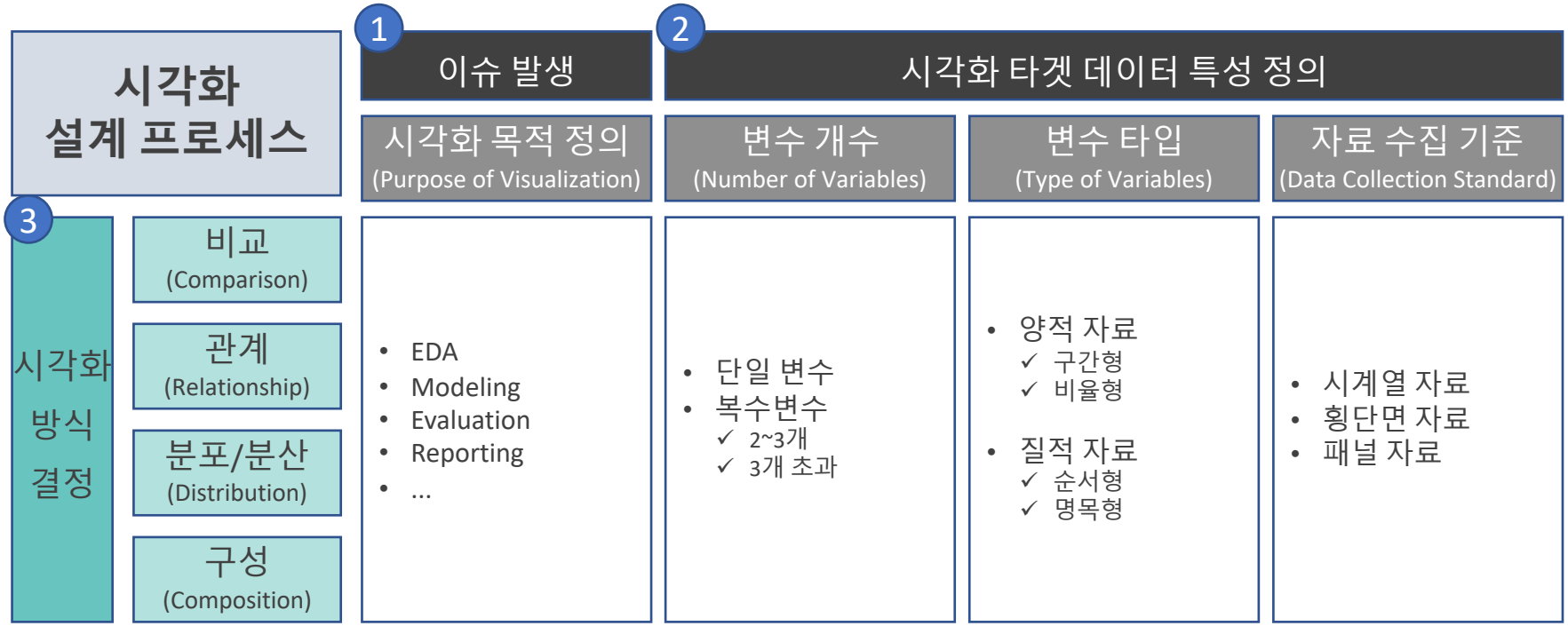
분석 결과가 문제의 답이라면, 시각화는 가장 직관적인 해설

시각화의 대상 :

- 전처리 진행 전의 원천 데이터
- 전처리 진행 후의 가공 데이터
- 개별 변수 내의 항목 구성 및 그 비율
- 개별 변수의 분포(분산)
- 개별 변수간의 관계
- 시간에 따른 데이터의 변동
- 기본 통계치
- 극단값 여부
- 왜도/첨도
- 변수 분류
- 변수 군집
- …

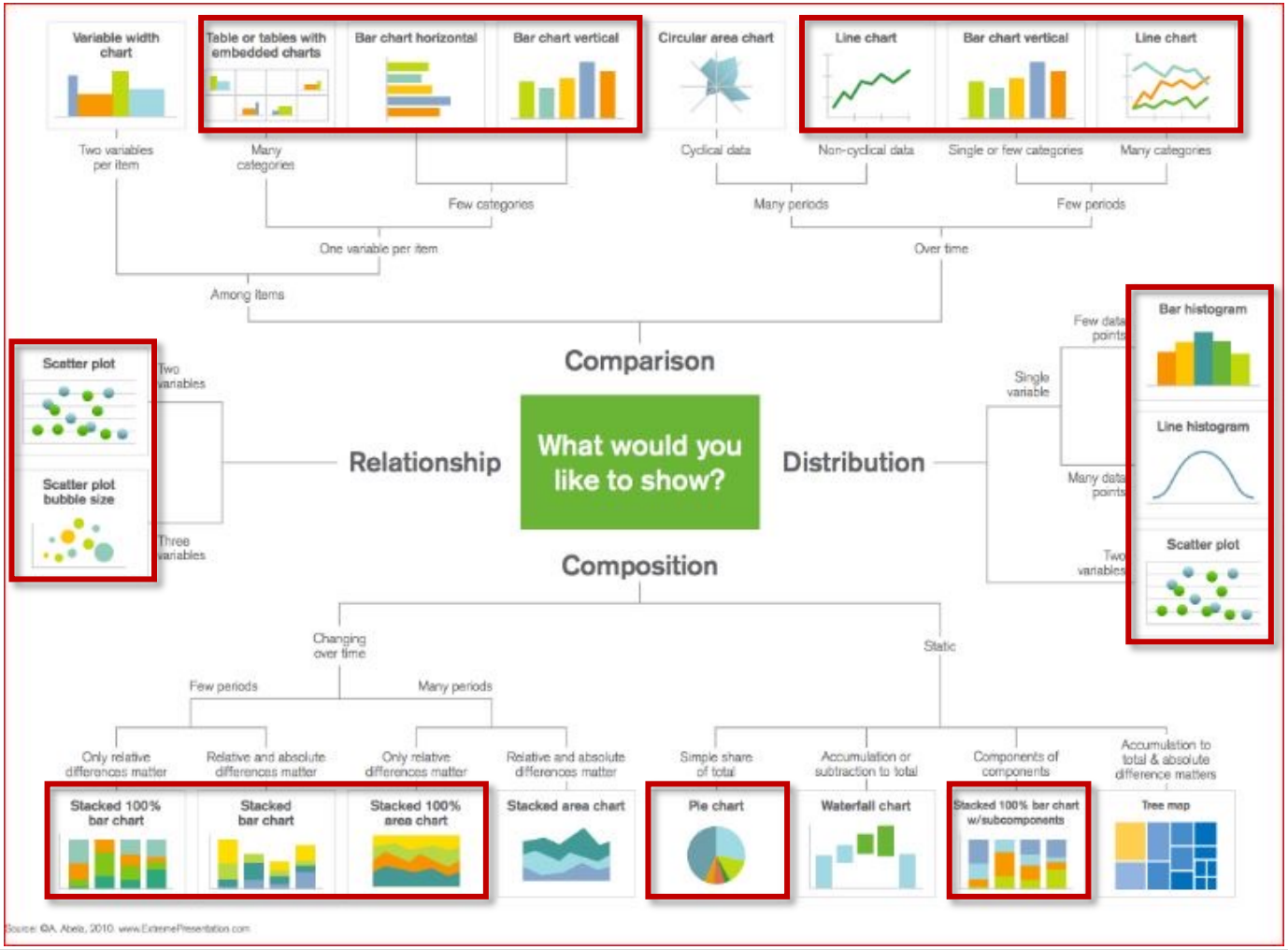
1. Concepts of Visualization | 어떻게 시각화 할 것인가?

실험 설계(Experimental Design)가 필요하듯, 시각화에도 설계가 필요하다.



1. Concepts of Visualization | 데이터 유형에 따른 시각화 기법

용도에 맞는 도구 혹은 도구에 맞는 용도



1. Concepts of Visualization | 왜 데이터 시각화를 해야하는가?

데이터 시각화를 통해 피해갈 수 있는 오류들

데이터와 관련된 일반적 오류 1

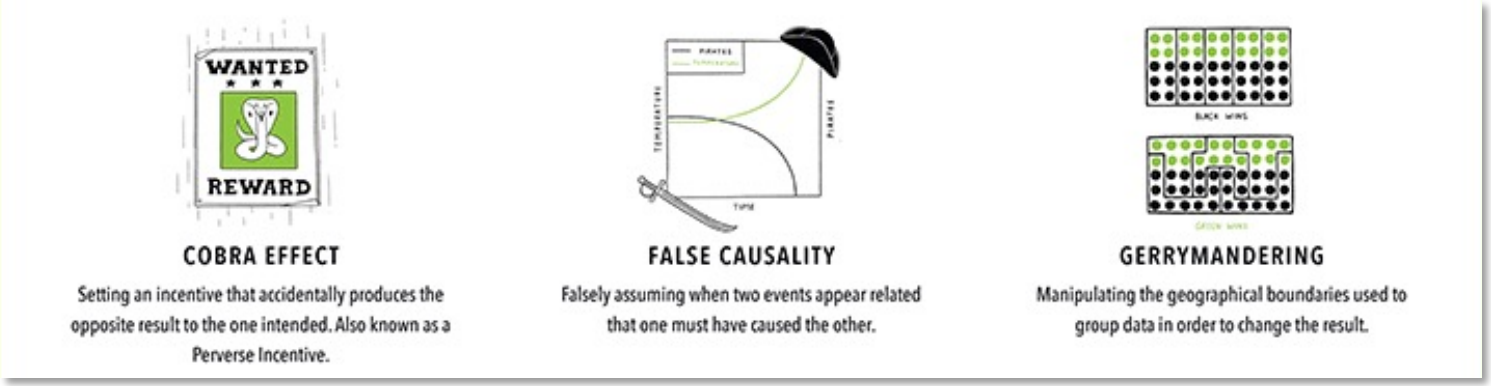
아전인수 *기저귀와 맥주* *죽은 사람은 말이 없다.*

(출처 : GECKOBOARD.COM)

1. Concepts of Visualization 왜 데이터 시각화를 해야하는가?

데이터 시각화를 통해 피해갈 수 있는 오류들

데이터와 관련된 일반적 오류 2



점입가경

까마귀가 날면 배가 떨어진다.

어디서 밑장을?!

(출처 : GECKOBOARD.COM)

1. Concepts of Visualization | 왜 데이터 시각화를 해야하는가?

데이터 시각화를 통해 피해갈 수 있는 오류들

데이터와 관련된 일반적 오류 3

올바른 표본 추출법은?

이 집 음식이 맛있는걸 보니,
이 골목은 맛집촌이군.

다음 동전도 앞면일 확률은?
무려! 1/2 입니다.

기대가 기적을 만든다.

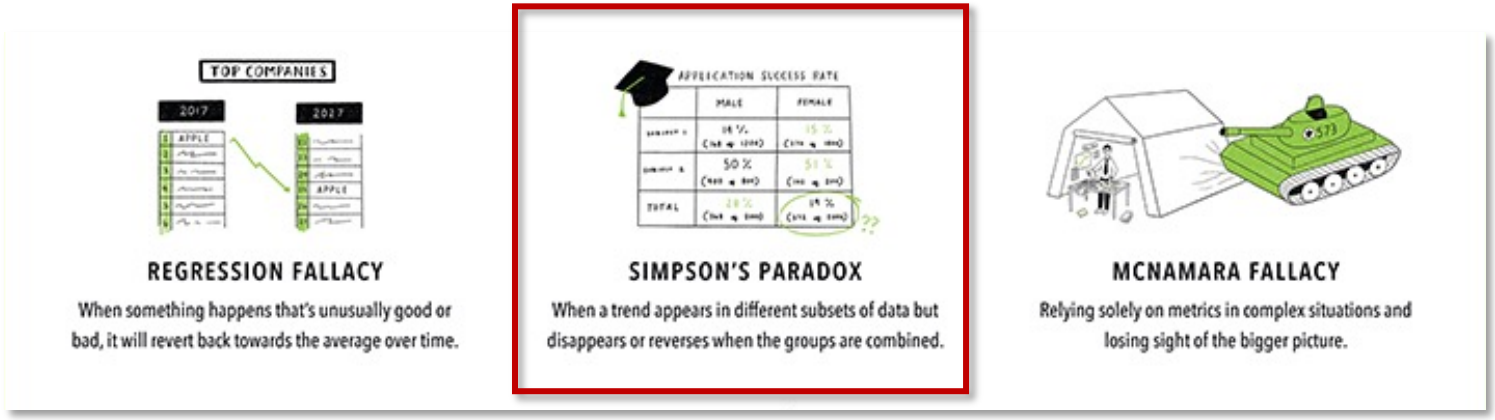
(출처 : GECKOBOARD.COM)

1. Concepts of Visualization | 왜 데이터 시각화를 해야하는가?

데이터 시각화를 통해 피해갈 수 있는 오류들

데이터와 관련된 일반적 오류 4

숫자만으론 와닿지 않는다.



매가 약이다?

반전 블록버스터

별 보고 걷다
돌부리에 걸리다

(출처 : GECKOBOARD.COM)

1. Concepts of Visualization | 왜 데이터 시각화를 해야하는가?

데이터 시각화를 통해 피해갈 수 있는 오류들

데이터와 관련된 일반적 오류 5

모델만으론 알 수 없다.

과유불급

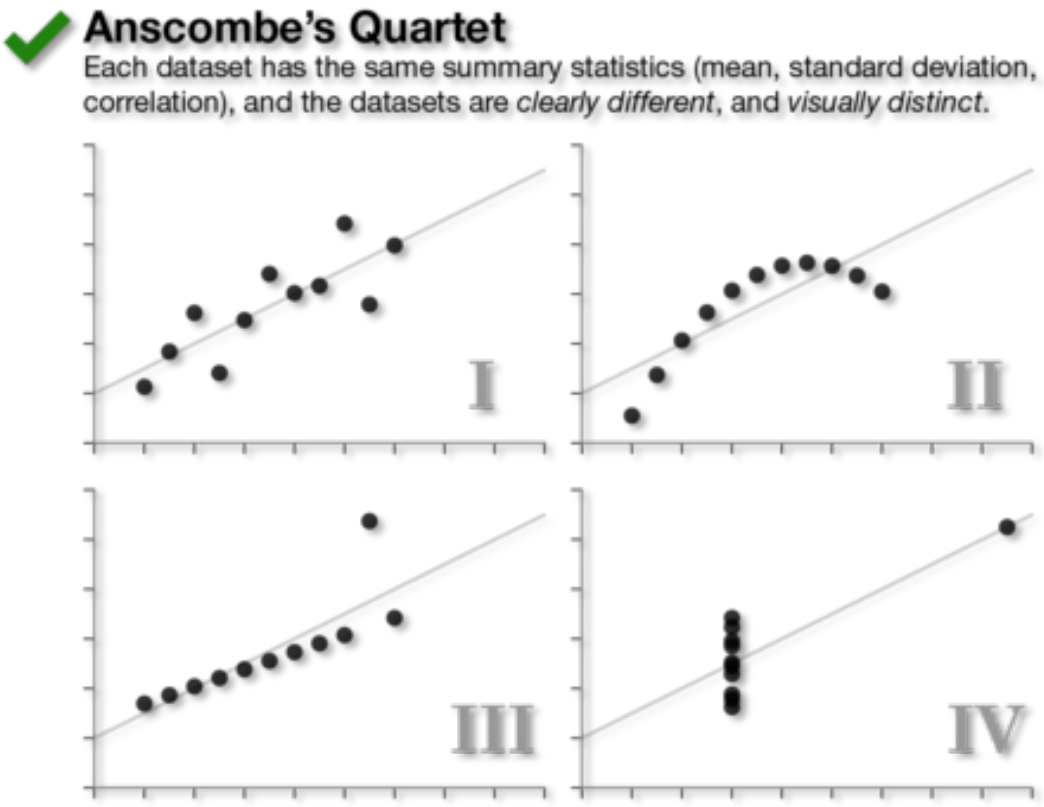
역사는 승자만을 기록한다.

보아야 알 수 있다.

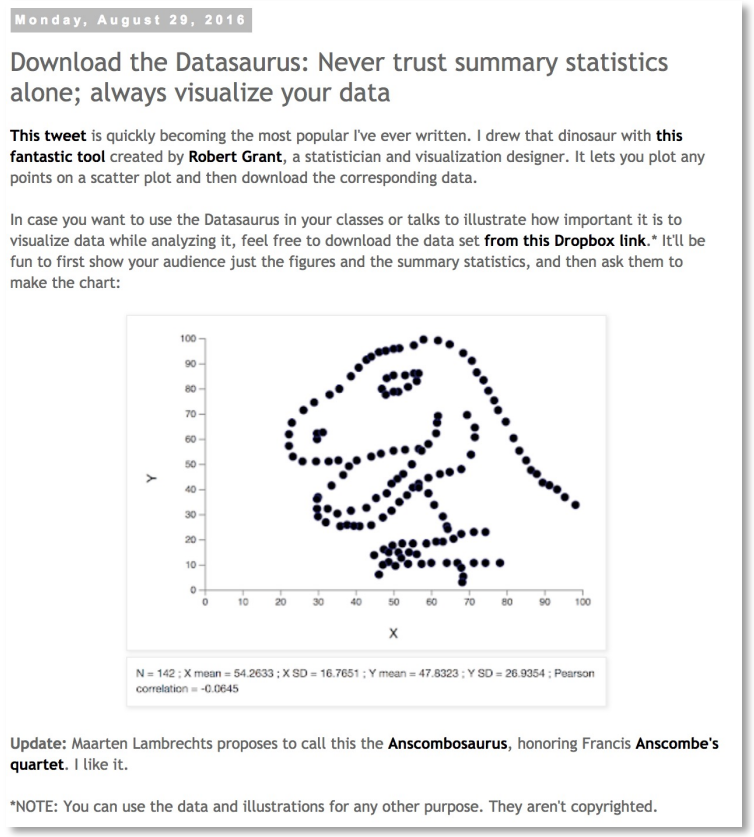
(출처 : GECKOBOARD.COM)

2. 탐색적 자료 분석(EDA) | EDA의 중요성

Anscombe's Quartet과 Datasaurus



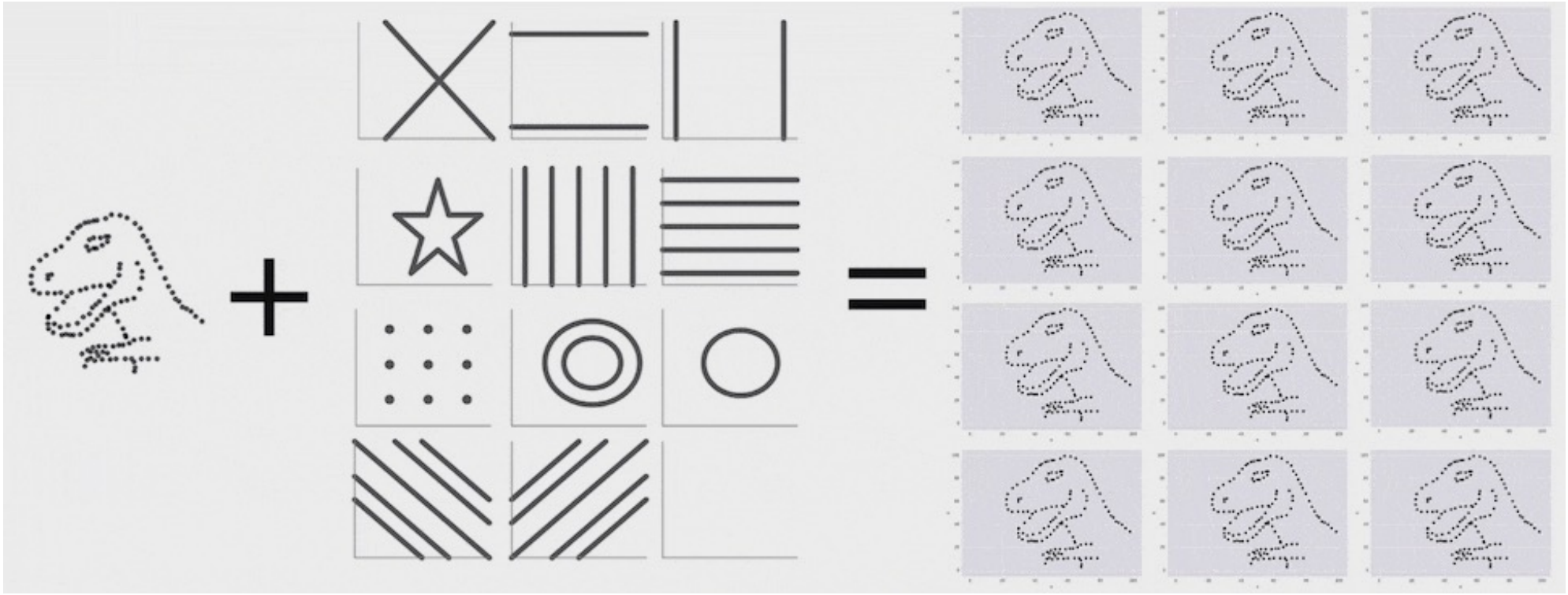
'Anscombe's Quartet', FJ Anscombe, 1973



'Datasaurus', Albert Cairo

2. 탐색적 자료 분석(EDA) | 데이터 유형에 따른 시각화 기법

공룡 12마리



‘The Datasaurus Dozen’

2. 탐색적 자료 분석(EDA) | EDA의 중요성

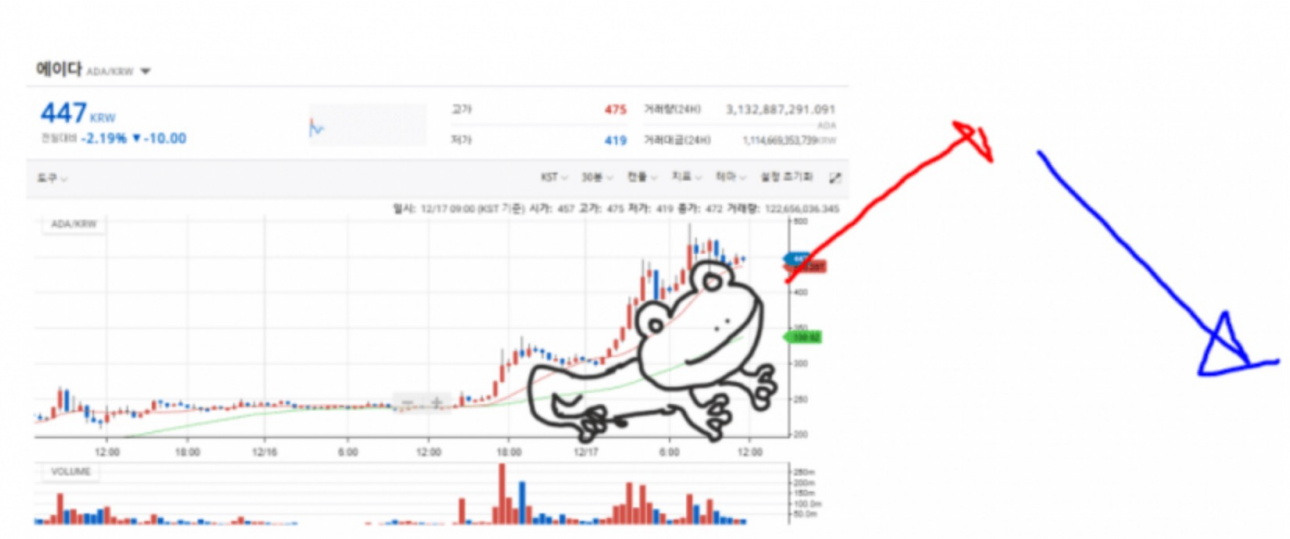
EDA는 사실 어려운 개념이 아님

리플은 오늘 고양이 가족 소풍나왔다.

크게 오르내리락 하지는 않고 이녀석도 아주 서서히 오를 예정이다.

2. 탐색적 자료 분석(EDA) | EDA의 중요성

EDA는 사실 어려운 개념이 아님



코인질은 그냥 처들어가는게 아니라 그래프보고 앞 상황을 예측하고 들어가야 한다.

개구리가 지금이라도 곧 점핑할 것처럼 보인다. 아마 장난질로 급격하게 올랐다가 급격하게 내릴 것 같으니 단타칠 애들만 잘보다 들어와라.

2. 탐색적 자료 분석(EDA) | EDA의 정의와 주제

EDA vs. CDA

EDA, CDA

- EDA
  - ✓ 탐색적 자료 분석(Exploratory Data Analysis), 미지의 특성을 파악하고 구조를 밝히기 위한 다양한 실험 수행
  - ✓ 수치적/계산적/시각적 탐색 작업
- CDA
  - ✓ 확증적 자료 분석(Confirmatory Data Analysis), 수집된 정보 및 자료에 대한 실증적(주로 통계적) 평가에 의한 분석

EDA의 4가지 핵심 주제

- 현시성
  - ✓ 데이터의 구조와 특성을 시각화하여 보여주며, 숨겨진 의미를 찾을 수 있도록 보조
- 잔차
  - ✓ 회귀에 저항하고 있는 특정 잔차들의 의미까지 고려
- 재표현
  - ✓ 간결하고 명료하게 자료를 재구성(log transformation)
- 저항성
  - ✓ 소수의 극단값에 의한 영향 저감(mean vs. median)

2. 탐색적 자료 분석(EDA) | EDA의 기본 기법

기본 기법 이해를 통한 EDA 수행능력 습득을 목표로 함

크게 6가지 기본 기법 필요

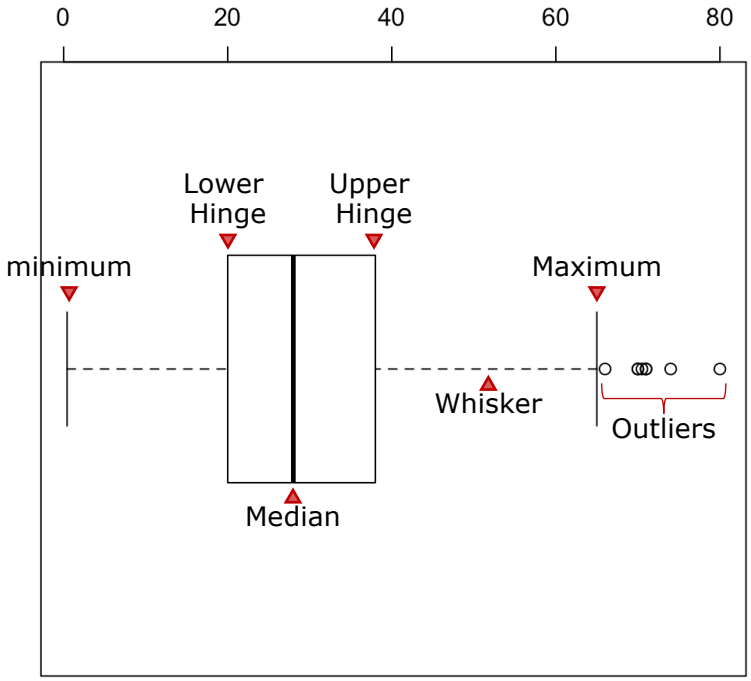
1. 기본 통계 요약
   - ✓ 시각화의 Key
2. 시각화
   - ✓ EDA는 시각화에서 시작하여 시각화로 끝
3. 왜도/첨도
   - ✓ 자료의 분포 패턴을 정량적으로 파악
4. 변수 변환
   - ✓ 왜도/첨도 등을 보정하여 분석 가능한 형태로 변환
5. 평활
   - ✓ 추세선의 과적합 방지
6. 극단값 처리
   - ✓ 절단(trimming) – 특정 극단값 제거
   - ✓ 조정(windsorizing) – 특정 극단값 변환(min or max)

2. 탐색적 자료 분석(EDA) | EDA의 기본 기법

기본 통계수치 요약 개요

5-Number Summary

- mininum < Lower Hinge < Median < Upper Hinge < Max




```
boxplot(titanic_raw$Age, horizontal = TRUE)
```

```
titanic_raw['Age'].plot.(vert=False, kind='box')
```

2. 탐색적 자료 분석(EDA) | EDA의 기본 기법

기본 통계수치 요약

5-Number Summary

- Median(중위수, 중간값)
  - ✓ Sum(개별 자료값)/(N, 자료의 크기)로 표현되는 평균과는 달리, 전체 자료 중 중간에 해당하는 자료의 값을 의미
  - ✓ N이 홀수인 경우 : (N + 1)/2번째 지점의 자료 값
  - ✓ N이 짝수인 경우 : N/2번째와 (N + 1)/2번째 자료 값의 평균
  - ✓ 중위수의 깊이(depth) : 중위수의 순위(rank)로서 d(Median) = (N + 1)/2
  - ✓ 깊이는 기본적으로 min{큰 쪽의 순위, 작은쪽의 순위}를 따름

- Hinge(경첩)
  - ✓ (1 + [d(Median)])/N번째 자료의 값을 의미
  - ✓ Hinge의 깊이 : (1 + [d(Median)])/2

- ✓ fivenum(##)
- ✓ summary(##)

	Depth	Value	
M(Median)	(N + 1)/2	**median**	
H(Hinge)	(1 + [d(M)])/2	**lower hinge**	**upper hinge**
	1	**min**	**Max**

```
fivenum(titanic_raw$Age)
summary(titanic_raw$Age)
```

```
titanic_raw['Age'].plot(vert=False, kind='box')
```

2. 탐색적 자료 분석(EDA) | EDA의 기본 기법

시각화

Stem & Leaf – 줄기 잎 그림

- 숫자형 데이터를 줄기와 잎으로 그려 빈도 및 분포를 시각화
- 직관적으로 데이터의 구조 이해 가능
- stem(데이터, scale = ##, width = ##)
- R Base의 plot들은 그래프 자체의 객체 할당 불가, 그래프 속성 형태로 저장(cf. ggplot2)
- Q) titanic_raw에서 Age 변수를 추출, NA값을 제거 한 후
  1. Argument를 입력하지 않고 데이터만 입력하여 출력하라
  2. Scale 0.5로 출력하라
  3. Width 133으로 출력하라
  4. Scale 0.5, Width 133으로 출력하라
  5. 객체 할당 후 객체를 호출하라

```
> stem(titanic_raw$Age)
```

```
> import stemgraphic
> stemgraphic.stem_graphic(titanic_raw['Age'])
```

2. 탐색적 자료 분석(EDA) | EDA의 기본 기법

시각화

Boxplot

- 숫자형 데이터를 상자형태로 그려 빈도 및 분포를 시각화
- 직관적으로 데이터의 구조 이해 가능
- boxplot(데이터, outline = TRUE/FALSE, na.rm = TRUE/FALSE)
- Q) titanic_raw에서 Age 변수를 추출 후
  1. Boxplot을 출력하라
  2. Outlier를 제거하고 출력하라

```
> boxplot(titanic_age, horizontal = TRUE, outline = FALSE)
```

```
> titanic_raw['Age'].plot(kind='box')
```

2. 탐색적 자료 분석(EDA) | EDA의 기본 기법

왜도/첨도

왜도(Skewness)

- 왜도란, 데이터가 좌/우 방향으로 치우친 정도를 의미
- Skew = $\{(H_U - M) - (M - H_L)\}/\{(H_U - M) + (M - H_L)\}$
- -1 <= Skew <= 1
- H_U = M일 때, skew=−1
- H_L = M일 때, skew=1
- 왜도 보정을 위한 최적 power는?

- Q) 왜도를 자동으로 보정하는 함수를 생성하라

첨도(Kurtosis)

- 첨도란, 데이터가 정규분포를 기준으로 높거나(낮은 분산) 낮은(높은 분산) 정도를 의미
- Kurtosis = E-spread / H-spread – 1.705 **E : 8분위수 , (1 + [d(M)])/2*
- = $(E_U - E_L)/ (H_U - H_L)$ – 1.705
- Kurtosis > 0 : 정규분포보다 뾰족
- Kurtosis < 0 : 정규분포보다 편평

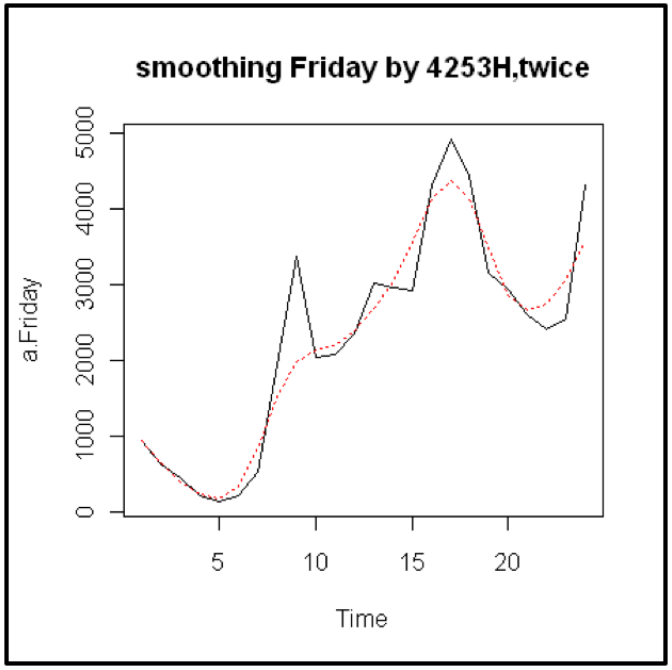
- Q) 숫자형 데이터의 왜도와 첨도를 자동으로 계산하는 함수를 생성하라

2. 탐색적 자료 분석(EDA) | EDA의 기본 기법

평활

평활(Smoothing)

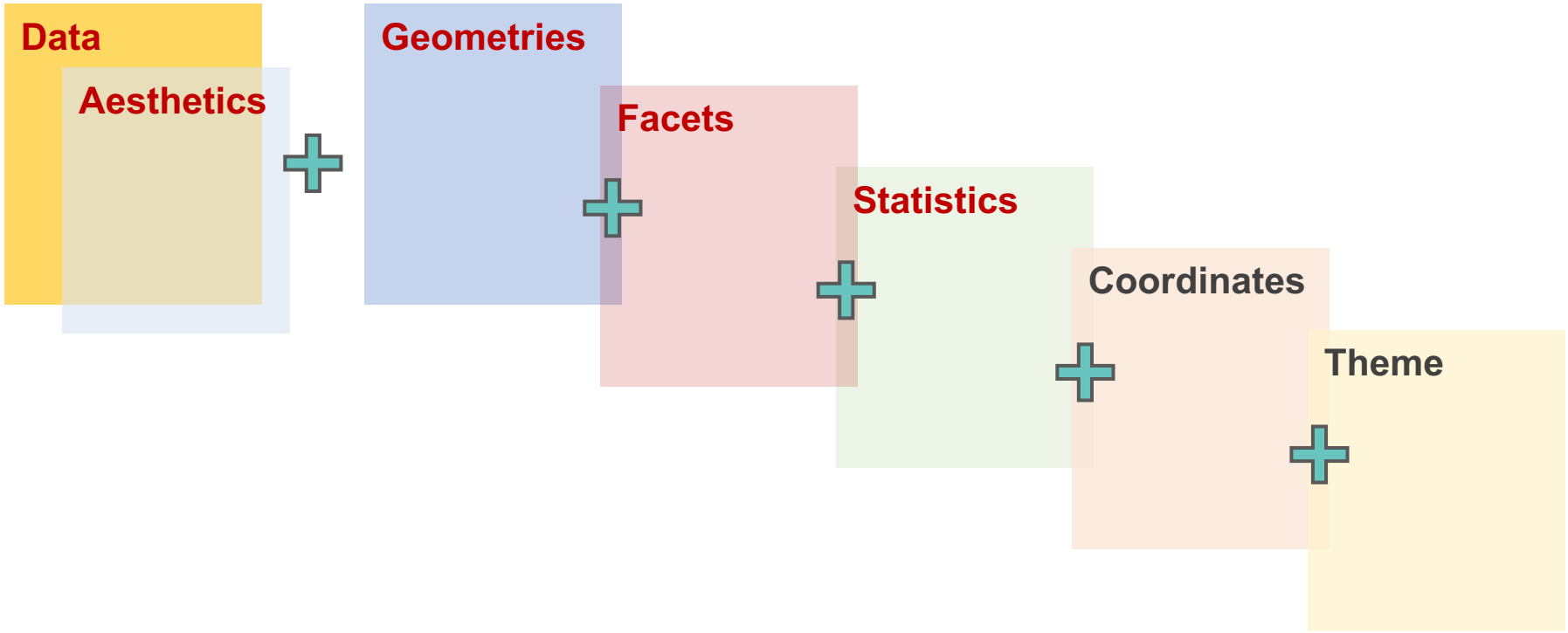
- 평활이란, 자료에 맞는 추세선을 대표성을 강화하여 그리는 기법
- LOESS Smoothing : LOcal regrESSion



3. Data Visualization | ggplot2

ggplot2의 이해

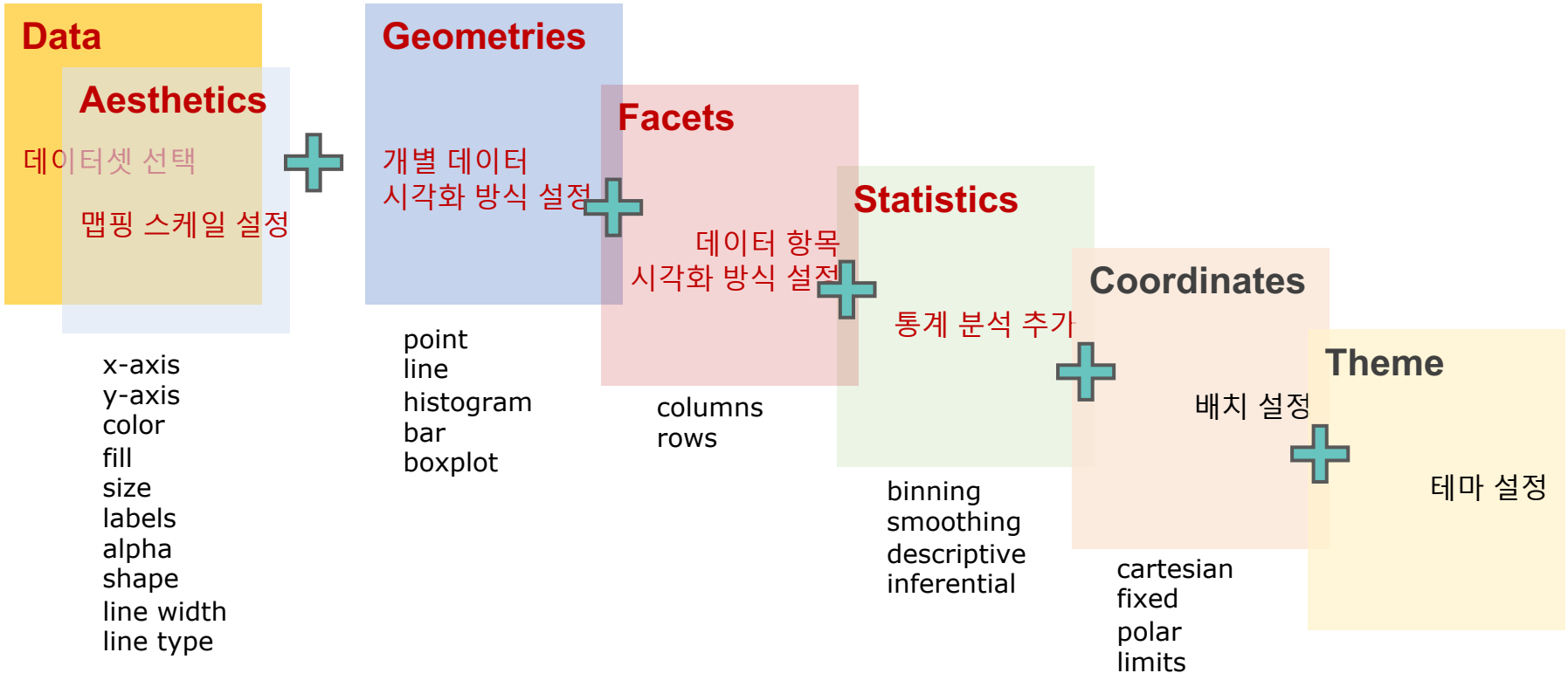
ggplot의 구조



3. Data Visualization | ggplot2

ggplot2의 이해

ggplot의 구조



3. Data Visualization | **ggplot2**

ggplot2의 이해

ggplot의 기본 문법(Data to Statistics)

```
ggplot(데이터, aes(x = 변수1, y = 변수2, color = 색깔, shape = 모양, size = 크기)) +
  geom_표현법(alpha = 투명도)) +
  facet_grid(. ~그리드 분할) +
  stat_smooth(method =  " 평활기준", col =  " 라인 컬러")
```

3. Data Visualization | **ggplot2**

ggplot2의 활용

실습(강의 중 제공)

4. One Point Tutorial Visualization I | matplotlib

matplotlib의 이해와 활용

별첨 튜토리얼 참조

5. One Point Tutorial Visualization II | Seaborn

seaborn의 이해와 활용

별첨 튜토리얼 참조

6. One Point Tutorial Visualization III | folium

folium의 이해와 활용

별첨 튜토리얼 참조

7. One Point Tutorial Visualization IV | pyecharts

pyecharts의 이해와 활용

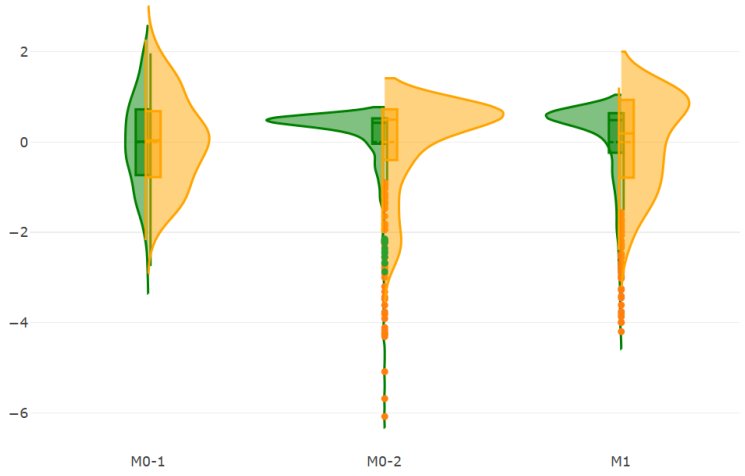
별첨 튜토리얼 참조

빅데이터 분석 중고급 과정

#8 Modeling

with One Point Tutorial
feat. R

danielwooseok.song@gmail.com



빅데이터 분석 중고급 과정

#6 **Data Manipulation** with One Point Tutorial feat. R

Agenda

1. Population & Sample
2. Inference
3. Hypothesis
4. Modeling
5. Evaluation
6. Real World
7. One Point Tutorial IX – 예외 처리

danielwooseok.song@gmail.com

1. Population & Sample | 모집단과 표본

전구의 불량률을 100% 알고 싶다면? 모든 전구를 사용하라

모집단과 표본

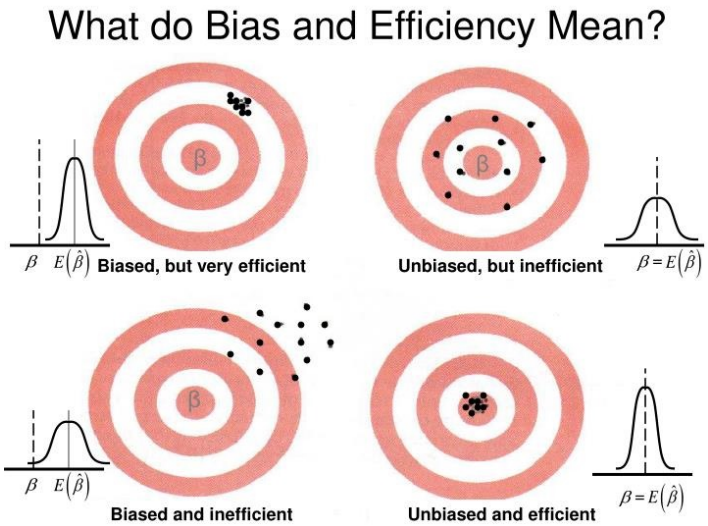
- 데이터 분석의 기본은 데이터 수집
- 데이터 수집은 전통적으로 모집단을 대상으로한 표본조사를 통해 달성
- 빅데이터 시대에는 각종 로그 데이터 등 특정 영역을 대상으로 전수 데이터의 확보와 분석이 점차 현실화
- 분석 대상의 전체 집합이 **모집단(population)**
- 모집단의 특성은 **모수(parameter)**를 통해 표현
- 특정 기준을 가지고 추출된 모집단의 한 부분이 **표본(sample)**
  - ✓ **통계량**(statistic) : 표본에서 계산된 통계
  - ✓ **추정량**(estimator) : 모수 추정에 사용된 통계량으로, 확률변수
  - ✓ **추정값**(estimate) : 추정량이 구현된 값으로, 관찰값

2. Inference | 통계적 추정

일격필살 vs. 난사

통계적 추정

- 통계적 추론(Statistical Inference) : 실험 혹은 관찰을 통해 얻은 데이터를 분석하여 모집단의 정보를 유추하는 것
  - ✓ 추정(Estimation) : 모집단에 대한 유추로서 대상을 하나의 값(point) 혹은 구간(interval)으로 함
  - ✓ 검정(Testing) : 모집단에 대한 예상 혹은 주장을 자료가 뒷받침하는 정도 파악, 모수/분포/모형 등 다양한 검정 존재
- 변동(Variation)
- 편향(Bias)
- 비편향성 vs. 효율성
- 일치성



3. Hypothesis | 가설 검정의 요소

이 산이 맞는가?

가설 검정

- 가설 검정(Hypothesis Testing) : 모집단의 특성에 대한 예상 혹은 주장의 참/거짓을 표본자료를 활용하여 판단하는 행위
  - ✓ **추정(Estimation)** : 모집단에 대한 유추로서 대상을 하나의 값(point) 혹은 구간(interval)으로 함
  - ✓ **검정(Testing)** : 모집단에 대한 예상 혹은 주장을 자료가 뒷받침하는 정도 파악, 모수/분포/모형 등 다양한 검정 존재

- 통계적 가설(Statistical Hypothesis) : 모집단에 대한 예상 혹은 주장
  - ✓ **귀무가설(Null Hypothesis, H_0)** : 기존에 참이라고 믿어지던 가설로서, 검정의 대상
  - ✓ **대립가설(Alternative Hypothesis, H_1 or H_a)** : 귀무가설이 참이 아닌 경우 믿어지는 가설로서 일반적으로 실험의 목적은 귀무가설의 기각

- 검정통계량(Test Statistic) : 검정을 위하여 표존자료에서 구한 통계량

3. Hypothesis | 가설 검정의 요소

뭣이 중헌디?

가설 검정의 오류

- 검정오류(Test Error) : 통계적 가설의 확률을 이야기 할 때, 이것이 맞거나 틀릴 가능성으로서 통계학에서는 틀릴 가능성을 위주로 논의
  - ✓ **제 1종 오류(Type I Error)** : 귀무가설이 옳으나 기각 – 생사람 잡기, 이 확률을 α로 표기하며 검정의 유의수준(significance level)이라 함
  - ✓ **제 2종 오류(Type II Error)** : 대립가설이 옳으나 기각 실패 – 검거 실패, 이 확률을 β로 표기

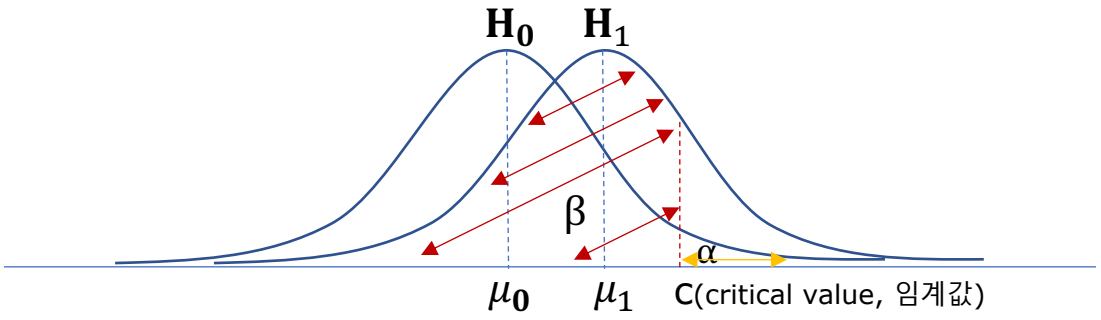
H_0 진위	의사결정	
	기각 불가	기각
사실	$1-\alpha$	α, 1종 오류확률
거짓	β, 2종 오류확률	$1-\beta$, 검정력

3. Hypothesis | 가설 검정의 요소

α vs. $1-\beta$

유의수준(α)과 검정력($1-\beta$)

- **검정력($1-\beta$)** : 대립가설이 사실일 때 귀무가설을 기각할 확률로서, 올바른 결정
- **$1-\alpha$** : 귀무가설이 사실일 때 기각하지 못 할 확률로서, 올바른 결정
- α와 β를 동시에 낮추는 것이 이상적이나, 서로 모순적 관계로서 불가능
- 표본의 크기에 따라 **고정된 α값** 하에서 β감소



4. Modeling | 분산분석 개요

분산분석(ANalysis Of VAriance)이란?

개요

- 분산분석이란, 모집단간 평균의 일치 여부를 분석하기 위해 수행
- 목적은 평균의 일치 여부 확인에 있으나, 이를 수행함에 있어 분산의 역할이 중요하여 분산분석이라 명명
- Q) 어떤 회사의 배터리 유형에 따라 A유형 배터리의 평균 지속시간이 61시간이고, B유형 배터리의 평균 지속시간이 83시간이며, C유형이 73시간인 경우 세가지 유형의 배터리 평균 지속시간은 같은 것인가? 혹은 다른 것인가?
  - ✓ H_0: $\mu_A = \mu_B = \mu_c$
  - ✓ H_1: 모집단의 평균들 중 차이가 존재
- 분산분석의 가정사항
  - ✓ 독립성 가정 : 모집단에서 추출된 표본은 각각 독립
  - ✓ 정규성 가정 : 각 모집단의 반응변수는 정규분포
  - ✓ 등분산 가정 : 반응변수의 분산은 모든 모집단에서 동일

4. Modeling | 일원분산분석(One-way ANalysis Of VAriance)

하나를 보면 열을 알 때

일원분산분석

- 요인(factor) 혹은 독립변수가 하나인 경우로, 요인은 수준(level) 혹은 처리(treatment)로 구성
- 두 개 이상의 모집단에서 추출한 요인의 평균 일치 여부를 분산을 통하여 분석
  - ✓ 표본평균끼리의 변동량 – 처리평균제곱(MSTR)
  - ✓ 표본내부의 변동량 – 오차평균제곱(MSE)
  - ✓ 두 변동량은 F통계량을 통하여 비교

$$\text{F통계량} = \frac{MSTR}{MSE} \sim F(k-1, n-k) \qquad k: \text{비교대상 모집단의 개수}, n: \text{표본의 총 개수 합}$$

- 분산분석표 : 분산분석 수행을 위한 계산식을 정리해놓은 표

	A	B	C
Battery 1	…	…	…
Battery 2	…	…	…
Battery 3	…	…	…
Battery 4	…	…	…
Battery 5	…	…	…
평균	…	…	…
표준편차	…	…	…

▶

변동요인	자유도	제곱합	평균제곱합	F통계량
처리	$k-1$	$SSTR$	$MSTR=\frac{SSTR}{k-1}$	$F=\frac{MSTR}{MSE}$
오차	$n-k$	SSE	$MSE=\frac{SSE}{n-k}$	
총계	$n-1$	SST		

- Q) 실습

4. Modeling | 이원분산분석(Two-way ANalysis Of VAriance)

두 명은 봐야 알 수 있을 때

이원분산분석

- 요인(factor) 혹은 독립변수가 두개인 경우
- 서로 다른 요인의 처리 그룹간의 상호작용(interaction) 존재 가능
- 분산분석표

		A	B	C	$x_{.j}$
A사	Battery 1	...	...	...	
	Battery 2	...	...	...	...
	Battery 3	...	...	...	
B사	Battery 1	...	...	...	
	Battery 2	...	...	...	...
	Battery 3	...	...	...	
	$\overline{x_{i.}}$	...	...	...	

변동요인	자유도	제곱합	평균제곱합	F통계량
요인 1	$a-1$	SS_{TR1}	MS_{TR1}	$F = \frac{MS_{TR1}}{MSE}$
요인 2	$b-1$	SS_{TR2}	MS_{TR2}	$F = \frac{MS_{TR2}}{MSE}$
상호작용	$(a-1)(b-1)$	SS_{INT}	MS_{INT}	$F = \frac{MS_{INT}}{MSE}$
오차	$ab(m-1)$	SSE	MSE	
총계	$abm-1$	SST		

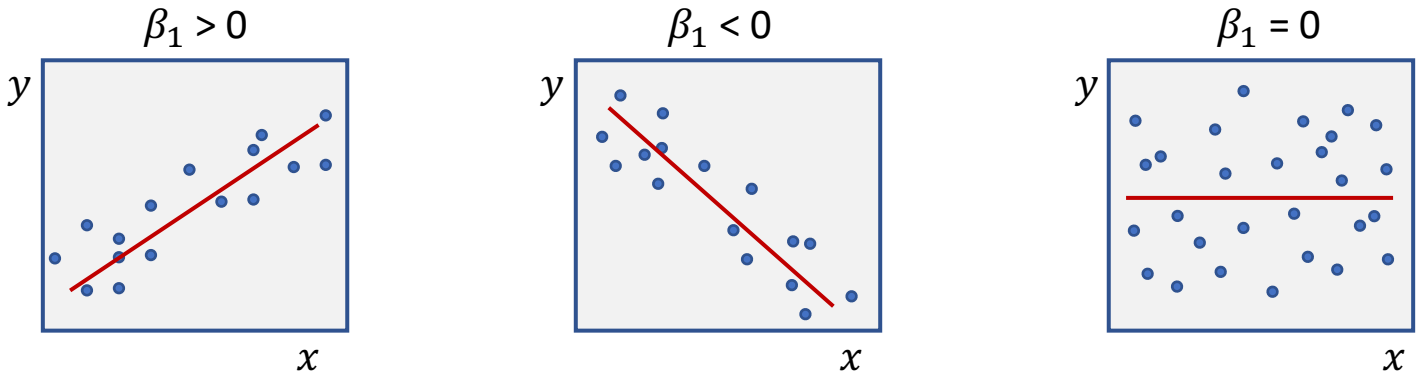
- Q1) 그 이상은?
- Q2) 실습

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 개요

개요(1/2)

- 두 가지 서로 다른 변수의 선형 관계(linear relationship)를 통계적으로 표현하고 분석하는 것
- 모형
  - ✓ $Y = \beta_0 + \beta_1 X + \varepsilon$
  - ✓ Y : 종속변수, β_0 : 절편(intercept), β_1 : 기울기(slope), X : 독립변수, ε : 오차
- 추정된 회귀식
  - ✓ $\hat{y} = b_0 + b_1 x$
  - ✓ b_0, b_1 : 추정치(estimated value)



4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 개요

개요(2/2)

- 최소제곱추정법(Ordinary Least Squares Method)
  - ✓ $\sum_{i=1}^{n} \varepsilon_i^2$가 β_0, β_1에 대하여 최소화 되도록 하여 추정값 b_0, b_1을 찾아내는 방법
  - ✓ $(b_0, b_1) = \arg\min \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min (\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i)^2$ with respect to (β_0, β_1)

- 최소제곱추정법의 추정값 b_0, b_1
  - ✓ $b_1 = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sum_{i=1}^{n}(x_i - x)^2} = r_{xy}\frac{s_y}{s_x}$
  - ✓ $b_0 = \bar{y} - b_1\bar{x}$
  - ✓ x_i : i번째 설명변수 관찰값
  - ✓ y_i : i번째 반응변수 관찰값
  - ✓ $\bar{x}$: x의 평균
  - ✓ $\bar{y}$: y의 평균
  - ✓ $r_{xy} = \frac{s_{xy}}{s_x s_y}$: x와 y의 표본상관계수, s_{xy}는 표본의 공분산(covariance)
  - ✓ s_x : x의 표본 표준편차
  - ✓ s_y : y의 표본 표준편차
  - ✓ n : 관찰값의 수

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 가정

단순선형회귀모형의 가정

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i \qquad i = 1, \cdots, n$$

- 오차항(ε_i)의 평균은 0이고 분산은 σ^2
  즉, $E(\varepsilon_i) = 0,\ \ Var(\varepsilon_i) = \sigma^2$: 등분산 가정

- 오차항(ε_i)들은 서로 독립
- 오차항(ε_i)은 정규분포를 따름 : $\varepsilon_i \sim N(0, \sigma^2)$

$$\varepsilon_i \sim iid\ N(0, \sigma^2)$$

- 반응변수의 확률분포가 정규성까지 만족할 경우, OLS는 모든 비편향 추정량 중 가장 분산이 작은 효율적 추정량

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) SST, MSE

SST = SSR + SSE, MSE

- SST : 총제곱합, Total Sum of Square
  - ✓ SST/DF : 총변동량
- SSR : 회귀제곱합, Regression Sum of Square
  - ✓ SSR/DF : 회귀변동량
- SSE : 오차제곱합, Error Sum of Square
  - ✓ SSE/DF : 오차변동량

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SSR + SSE$$

- MSE : 오차평균제곱, Mean Square Error, 분산 σ^2의 추정값
  - ✓ $s^2 = \hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$, $n\text{-}2 : DF$
  - ✓ 자유도 = 관찰값의 수 – 회귀계수의 수, 현재 β_0, β_1의 두 가지 회귀계수가 존재하므로 자유도는 n - 2

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 분산분석표

단순선형회귀모형의 ANOVA Table

변동요인	제곱합	자유도	평균제곱합	F통계량
회귀	SSR	1	$MSR=\frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
잔차	SSE	$n-2$	$MSE=\frac{SSE}{n-2}$	
총계	SST	$n-1$		

- Q) One-way ANOVA 분산분석표와의 차이는?

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 결정계수

결정계수(Coefficient of Determination) R^2

- Y의 총 변동량 SST에 대해 추정된 회귀식이 설명하는 변동량 SSR의 비율로, R^2라 표기
- 회귀모형의 적합성(Goodness of Fit)에 대한 중요한 판단 기준

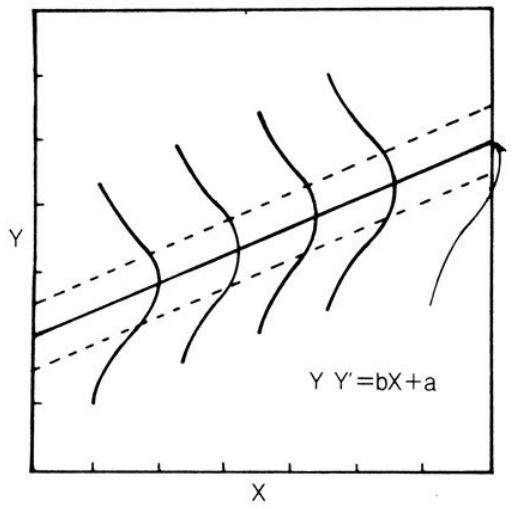
$$R^2 = \frac{\text{회귀변동량}}{\text{총변동량}} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 진단

진단(Diagnostics) 및 처방

- 진단 항목
  - ✓ 모형의 선형성(linearity)
  - ✓ 오차의 정규성(normality), 등분산성(homoscedasticity), 독립성(independency)
  - ✓ 특이값(극단값, outlier) 존재 여부
  - ✓ 영향관찰값(influential observation) 존재 여부



4. Modeling | 선형회귀

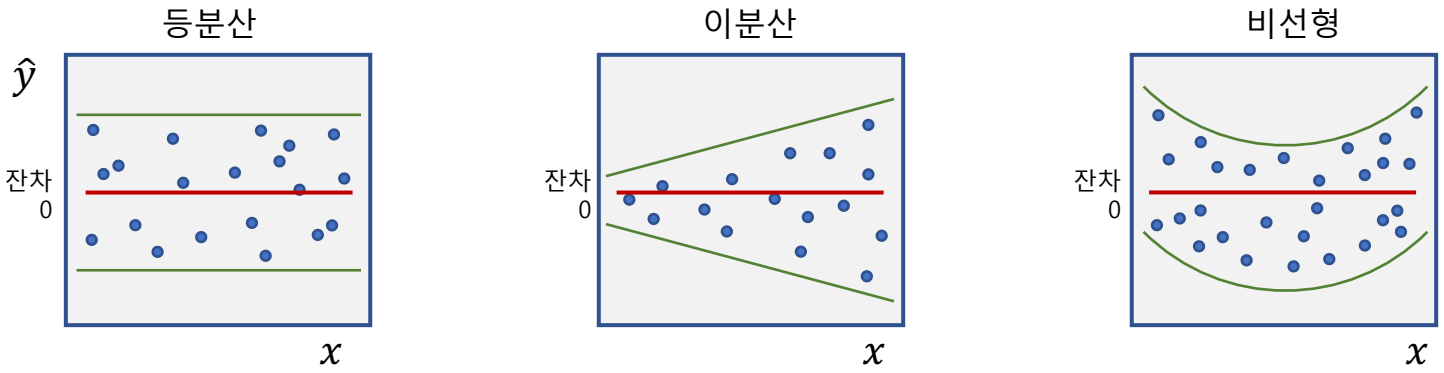
단순선형회귀(Simple Linear Regression) 잔차

잔차, 추정오차(Residual, Estimated Error)

$$e_i = y_i - \hat{y}_i = i - th\ residual$$

- 잔차는 추정된 회귀식이 반응변수 관찰값을 설명하지 못하는 부분
- 회귀분석은 산점도(scatter plot)으로 시작하여 잔차도(residual plot)로 마무리
- 적절한 회귀분석의 경우, 추정된 회귀 값 이외의 잔차에는 어떠한 패턴도 존재하면 안됨
- 잔차에 패턴이 존재하는 경우 회귀모델 재설계 필요

$e = y - \hat{y}$



4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 잔차

잔차에 대한 진단 및 처방

- 등분산 가정 이탈
  - ✓ 가중회귀 수행
  - ✓ 변수변환(제곱근/로그/역수/지수변환 등)
- 독립성 가정 이탈
  - ✓ 자료 추출과정 점검/독립성 검정
- 정규성 가정 이탈
  - ✓ 변수변환
  - ✓ 설명변수 추가,삭제

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 실습

상관계수 구하기

- 상관계수란, 두 확률변수 사이의 선형/비선형 관계를 파악하는 것
  - ✓ 피어슨 상관계수 – 일반적인 선형 상관관계 파악 (-1, 1)
  - ✓ 스피어만 상관계수 – 비선형 상관관계 파악, 이산/순서형 데이터도 가능
  - ✓ cor(변수1, 변수2), rcorr(변수1, 변수2, type =“방식명”)
  - ✓ 상관계수의 검정 – 상관계수의 유의성 파악 – cor.test(변수1, 변수2, method = ”방식명”)

- Q) mtcars 데이터의 wt와 mpg의 상관계수를 구하고 유의성을 파악하라

4. Modeling | 선형회귀

단순선형회귀(Simple Linear Regression) 실습

선형회귀 및 상관계수 구하기

- 상관계수란, 두 확률변수 사이의 선형/비선형 관계를 파악하는 것
  - ✓ 피어슨(Pearson) 상관계수 – 일반적인 선형 상관관계 파악 (-1, 1)
  - ✓ 스피어만(Spearman) 상관계수 – 비선형 상관관계 파악, 이산/순서형 데이터도 가능
  - ✓ cor(변수1, 변수2, method = "방식명")
  - ✓ 상관계수의 검정 – 상관계수의 유의성 파악 – cor.test(변수1, 변수2, method = "방식명")

- Q1) mtcars 데이터의 wt와 mpg의 상관계수를 구하고 유의성을 파악하라
- Q2) 위 데이터를 대상으로 단순선형회귀 수행 후 산점도에 abline을 추가하라
- Q3) 잔차도를 그려 유의성을 판단하라

4. Modeling | 선형회귀

다중선형회귀(Multiple Linear Regression) 개요

개요

- 두 가지 이상의 설명변수와 하나의 반응변수의 선형 관계(linear relationship)를 통계적으로 표현하고 분석하는 것
- 모형
  - ✓ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$
  - ✓ Y : 종속변수, β_0 : 절편(intercept), β_j: j번째 모수, X_j: j번째 독립변수, ε : 오차
- 추정된 회귀식
  - ✓ $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}, \quad i = 1, \cdots, n$
  - ✓ $b_0, b_1, \cdots, b_k$: 모수의 추정치(estimated value)
- *다항회귀 : 2차항 이상의 항이 포함된 회귀 모형으로서, 단순/다중 선형 회귀와 구분

4. Modeling | 선형회귀

다중선형회귀(Multiple Linear Regression) 개요

SST = SSR + SSE, MSE

- SST : 총제곱합, Total Sum of Square
  - ✓ SST/DF : 총변동량
- SSR : 회귀제곱합, Regression Sum of Square
  - ✓ SSR/DF : 회귀변동량
- SSE : 오차제곱합, Error Sum of Square
  - ✓ SSE/DF : 오차변동량

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SSR + SSE$$

- MSE : 오차평균제곱, Mean Square Error, 분산 σ^2의 추정값
  - ✓ $s^2 = \hat{\sigma}^2 = MSE = \frac{SSE}{n-k-1} = \frac{\sum(y_i-\hat{y}_i)^2}{n-k-1}$, n-k-1 : DF
  - ✓ 자유도 = 관찰값의 수 – 회귀계수의 수, 현재 $\beta_0, \beta_1, \cdots, b_k$의 k+1가지 회귀계수가 존재하므로 자유도는 n – (k+1)

다중선형회귀(Multiple Linear Regression) 분산분석표

다중선형회귀모형의 ANOVA Table

변동요인	제곱합	자유도	평균제곱합	F통계량
처리	SSR	k	$MSR=\frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
오차	SSE	$n-k-1$	$MSE=\frac{SSE}{n-k-1}$	
총계	SST	$n-1$		

- Q) One-way ANOVA 분산분석표와의 차이는?

4. Modeling | 선형회귀

다중선형회귀(Multiple Linear Regression) 진단

진단(Diagnostics) 및 처방

- 진단 항목
  - ✓ 모형의 선형성(linearity)
  - ✓ 오차의 정규성(normality), 등분산성(homoscedasticity), 독립성(independency)
  - ✓ 특이값(극단값, outlier) 존재 여부
  - ✓ 영향관찰값(influential observation) 존재 여부
  - ✓ 다중공선성(multicollinearity)

- 다중공선성 : X변수들 사이의 선형독립성이 깨진 상태로서, 하나의 X변수와 다른 X변수들의 선형결합이 높은 상관관계를 가지는 경우
  - ✓ 이 경우, X변수들의 기울기 계산에 오차가 발생하거나 계산 자체가 불가능

- 다중공선성의 진단
  - ✓ 산점도의 선형성
  - ✓ 변수 추가 혹은 제거 시 변수들의 기울기 추정값이 크게 변동
  - ✓ 분산분석시 기울기들이 0이 아님이 유의하나 문제 변수의 기울기는 비 유의한 경우

- Q) 실습

4. Modeling | 변수변환(Variable Transformation)

나의 전장에서 싸우라

변수변환

- 정의
  - ✓ 비선형 관계에 있는 설명변수와 반응변수를 선형 관계로 변화시키기 위해 사용
  - ✓ 로그, 지수, 역수 변환 등등
- 효과
  - ✓ 모형구축 및 결과 해석 용이
  - ✓ 적합성 및 예측력 증가
- Ladder of Power
  - ✓ 함수의 변화 속도별 power로서, power의 크기가 클 수록 변화 증가
  - ✓ J형 자료 : X변수 대상, 1보다 큰 power로 변환(Y변수 기준일 경우, 반대)
  - ✓ L형 자료 : X변수 대상, 1보다 작은 power로 변환(Y변수 기준일 경우, 반대)
  - ✓ $y = x^a = (x^b)^{a/b}$
- Q) 실습

4. Modeling | 로지스틱 회귀분석(Logistic Regression)

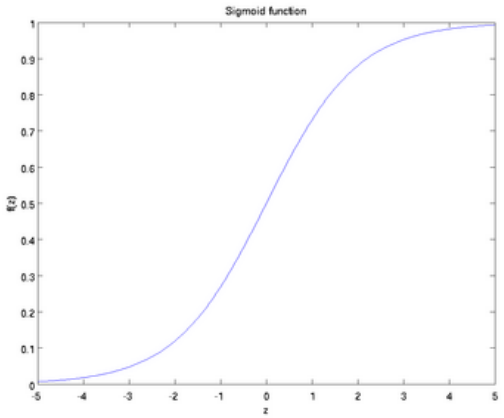
죽느냐 사느냐 그것이 문제일 때

개요

- Y변수(target variable)가 두 가지 원소를 갖는 명목형 자료일 경우 사용하는 비선형 회귀분석
  - ✓ 신용의 우/분량, 부도 판별 등
- 모형

$$E(Y|x_1,\cdots,x_k) = \Pr(Y=1|x_1,\cdots,x_k) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}} = \frac{1}{1+e^{-\beta_0-\beta_1x_1-\beta_2x_2-\cdots-\beta_kx_k}}$$

- 그래프의 형태



- Q) Sigmoid 함수를 작성하고 그래프를 그려보라

4. Modeling | 로지스틱 회귀분석(Logistic Regression)

로짓변환

오즈(Odds), 오즈비(Odds Ratio)

- 오즈 : 특정 사건이 발생할 확률(p)을 발생하지 않을 확률(1-p)로 나눈 값으로서, 사건이 발생할 확률과 발생하지 않을 확률의 상대적 비율

$$\text{오즈}(odds) = \frac{\Pr(\text{사건 발생})}{\Pr(\text{사건 미발생})} = \frac{\Pr(\text{사건 발생})}{1 - \Pr(\text{사건 발생})} = \frac{p}{1-p}$$

- 오즈비 : 두 오즈를 비율로 계산한 것

$$\text{오즈비}(Odds\ Ratio) = \frac{\text{오즈}1(odds1)}{\text{오즈}2(odds2)}$$

- 로지스틱 회귀모델의 오즈(Y=1일 때)

$$\frac{\Pr(Y=1|x_1,\cdots,x_k)}{1-\Pr(Y=1|x_1,\cdots,x_k)} = \frac{\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}}{1-\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}} = \frac{\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}}{\frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}} = e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}$$

- 여기에 로그를 취하여 다중선형회귀모형 형태로 변환한 것이 로그오즈(log-odds) 혹은 로짓(logit)

$$\text{logit}(Y) = \log(odds) = \log(e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}) = \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$$

- Q) 실습 – glm(y~x, family = binomial)

4. Modeling I

Clustering

군집분석

- 객체들의 유사성을 측정하여 유사성이 높은 대상 집단을 분류, 군집 내부 및 군집간의 상이성을 분석
- 특성에 따라 객체들을 복수의 배타 군집으로 분류

- **장점**
  - ✓ 사전 정보의 유무와 관계 없이 유의미한 자료 구조 파악 가능
  - ✓ 다양한 데이터 형태에 적용 가능
- **단점**
  - ✓ 가중치 및 거리의 정의에 따라 결과가 크게 변동
  - ✓ 기준 군집 수 설정 난해
  - ✓ 결과 해석 난해

- **계층적 군집분석**
  - ✓ n개의 군집으로 시작, 거리를 척도로 군집의 개수를 줄여나가는 방식
- **비계층적 군집분석**
  - ✓ n개의 군집을 만들 수 있는 최적 방식 탐색, 대용량 데이터를 처리 할 때 용이(k-means)
  - ✓ elbow, silhouette 등을 활용하여 군집수 최적화

- **Q) 실습**

4. Modeling I

Decision Tree

의사결정 나무

- 의사 결정 규칙과 그 결과들을 트리 구조로 도식화 한 의사결정 지원 도구의 일종
  - ✓ 특정 개체에 특정 값 부여

- **장점**
  - ✓ 해석 및 적용 용이
  - ✓ 비모수적 모델

- **단점**
  - ✓ 안정성 미흡
  - ✓ 과적합 우려

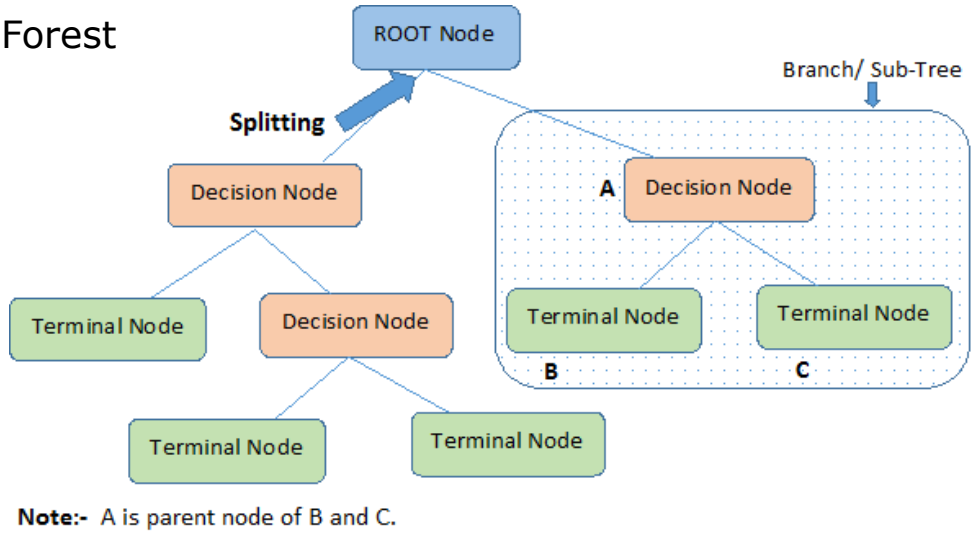
4. Modeling I

Decision Tree

의사결정 나무

- 구성요소
  - ✓ Root Node: 뿌리로서, 전체 분석 대상을 나타냄. 뿌리 노드로부터 다양한 특성을 반영하여 하위 집단을 형성해감
  - ✓ Splitting: 두 가지 이상의 하위 노드를 생성하는 과정
  - ✓ Decision Node: 결정 노드. 또다른 하위 노드를 구성하고 있는 부모(parent) 노드
  - ✓ Leaf Node or Terminal Node: 더이상의 분할(split)이 일어나지 않는 노드
  - ✓ Pruning: 가지치기. 분류 오류를 줄이거나 부적절한 가지를 제거, 분할과 가지치기를 통해 트리가 성장
  - ✓ Branch or Sub-Tree: 의사결정 나무의 한 부분으로서, 부모 노드와 자식 노드들의 집합
  - ✓ Parent Node and Child Node: 상대적인 개념으로서, 한 노드는 상위 노드의 자식 노드이자 하위 노드의 부모 노드

- Q) 실습 – Random Forest



5. Evaluation |

평가

Precision & Recall

- **정의**
  - ✓ 이진 분류(binary classification)을 활용하는 예측/분석에서 모델의 성능을 평가하는데 활용
  - ✓ 정밀도(precision) or 긍정예측치(PPV, Positive Predictive Value) : 예측된 긍정 중 실제 긍정의 비율
  - ✓ 재현율(recall) or 민감도(sensitivity) : 실제 긍정 중 예측된 긍정의 비율
  - ✓ 여기에서의 긍/부정은 정/오 판단이 아닌, 이진 분류상의 상호 배타적 항목을 의미

Precision Recall		실제 정답	
		Positive	Negative
실험 결과	Positive	TP (True Positve)	FP Type I Err. (False Positive)
	Negative	FN Type II Err. (False Negative)	TN (True Negative)

$$\textbf{Precision} = \frac{TP}{TP + FP} \qquad \textbf{Recall} = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1\ Score = \frac{2}{1/recall + 1/precision}$$

- **Q) 실습 - Titanic 데이터에 대한 Random Forest 및 Logistic Regression 모델 예측 결과를 Precision & Recall 표를 만들어 분석하라**

6. Real World |

실전 데이터 사이언스 흐름의 이해

시작부터 활용까지

- 일반적으로 데이터 분석은 목표와 대상을 정의하여 task를 확정하고
- 최종적인 활용 계획까지 수립한 후 진행

- 이러한 단계가 생략될 경우, 무의미하거나 적용 불가능한 작업이 될 가능성 증가
- 한 번 구축된 실험은 내/외부 요인의 변동이 생기기 전까지 재사용 가능
- 분석의 결과 역시 하나의 주요 데이터로 누적하여 추후 분석에 반영

- Q) 실습 - Titanic 데이터를 대상으로 분석 계획 수립부터 회고 및 적용까지의 리포트를 작성하고 토론하라

7. One Point Tutorial IX | 예외 처리

예외 처리의 이해와 활용

별첨 튜토리얼 참조

End of Document