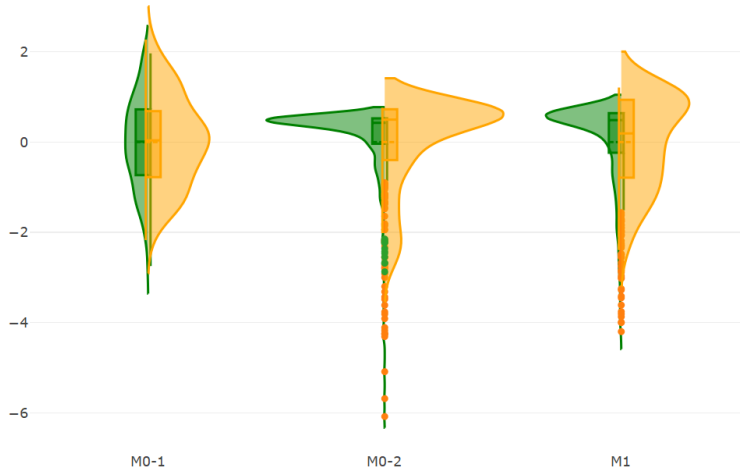


Data Visualization with Python – Day 2 시각화

Dec 2019

Daniel W. Song | song_wooseok@mensakorea.org

Data Visualization with Python – Day 2 시각화

Dec 2019

Agenda

1. Concepts of Visualization
2. EDA
3. Data Visualization(by J. Notebook)

Daniel W. Song | song_wooseok@mensakorea.org

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

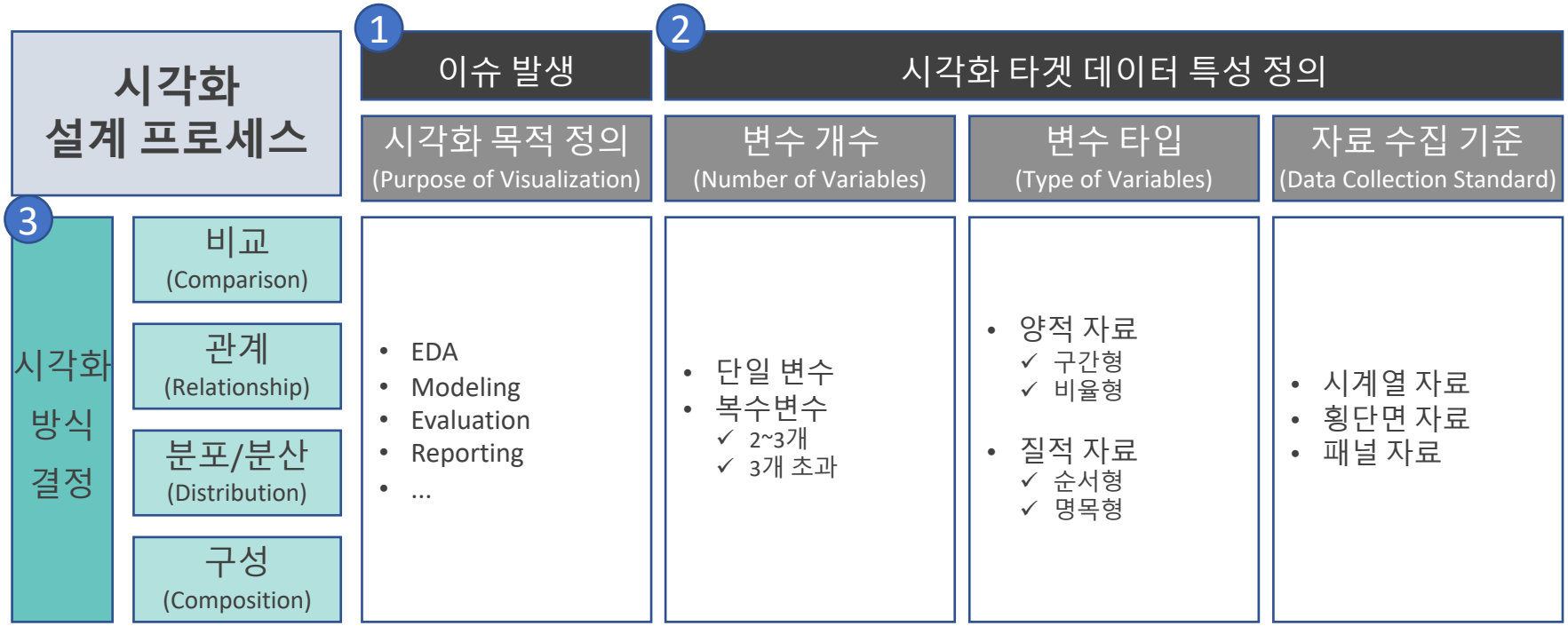
분석 결과가 문제의 답이라면, 시각화는 가장 직관적인 해설

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
- …

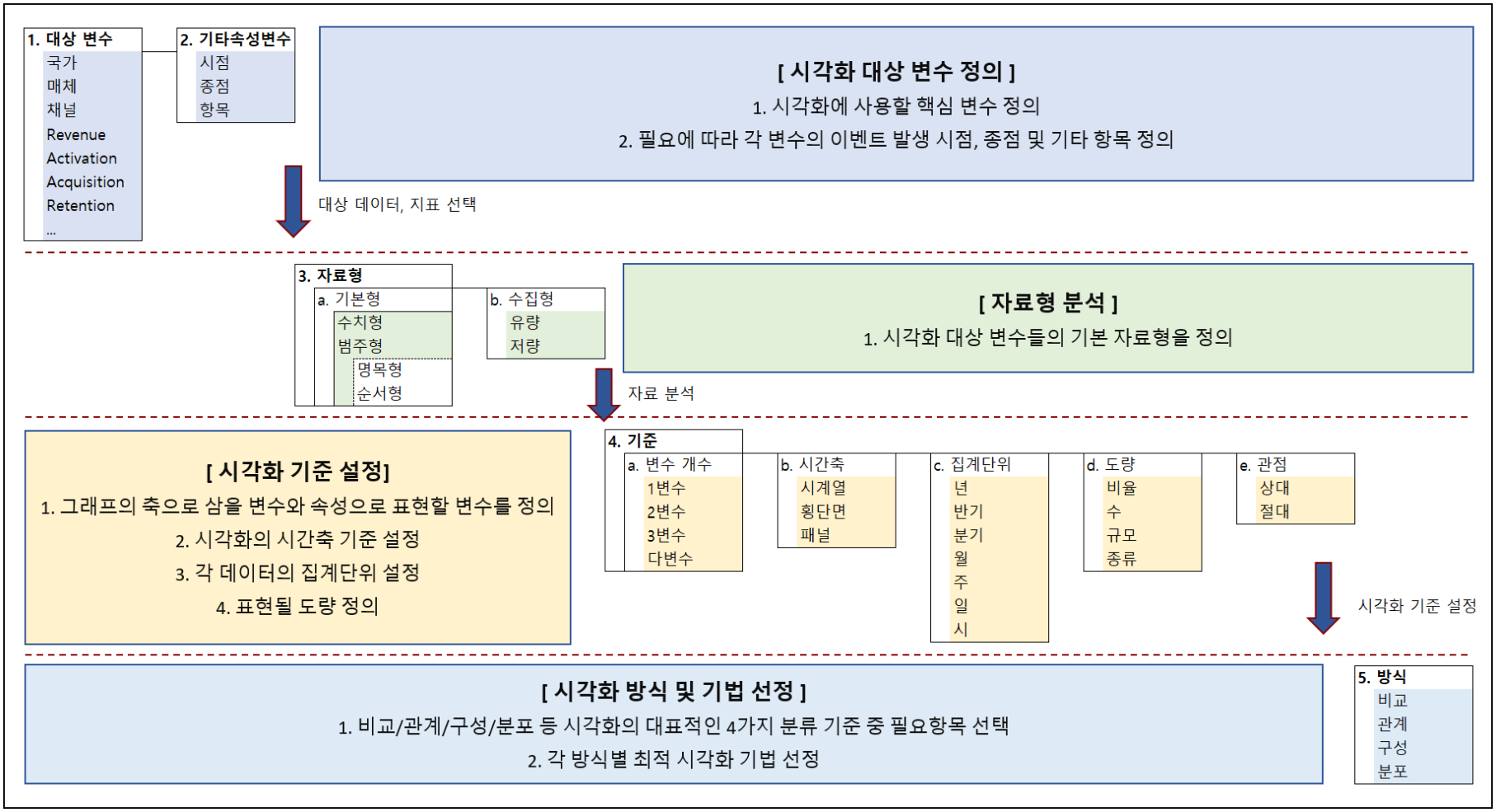
1. Concepts of Visualization | 어떻게 시각화 할 것인가?

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



1. Concepts of Visualization | 어떻게 시각화 할 것인가?

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



1. Concepts of Visualization | 어떻게 시각화 할 것인가?

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

▫ 시각화 체크표

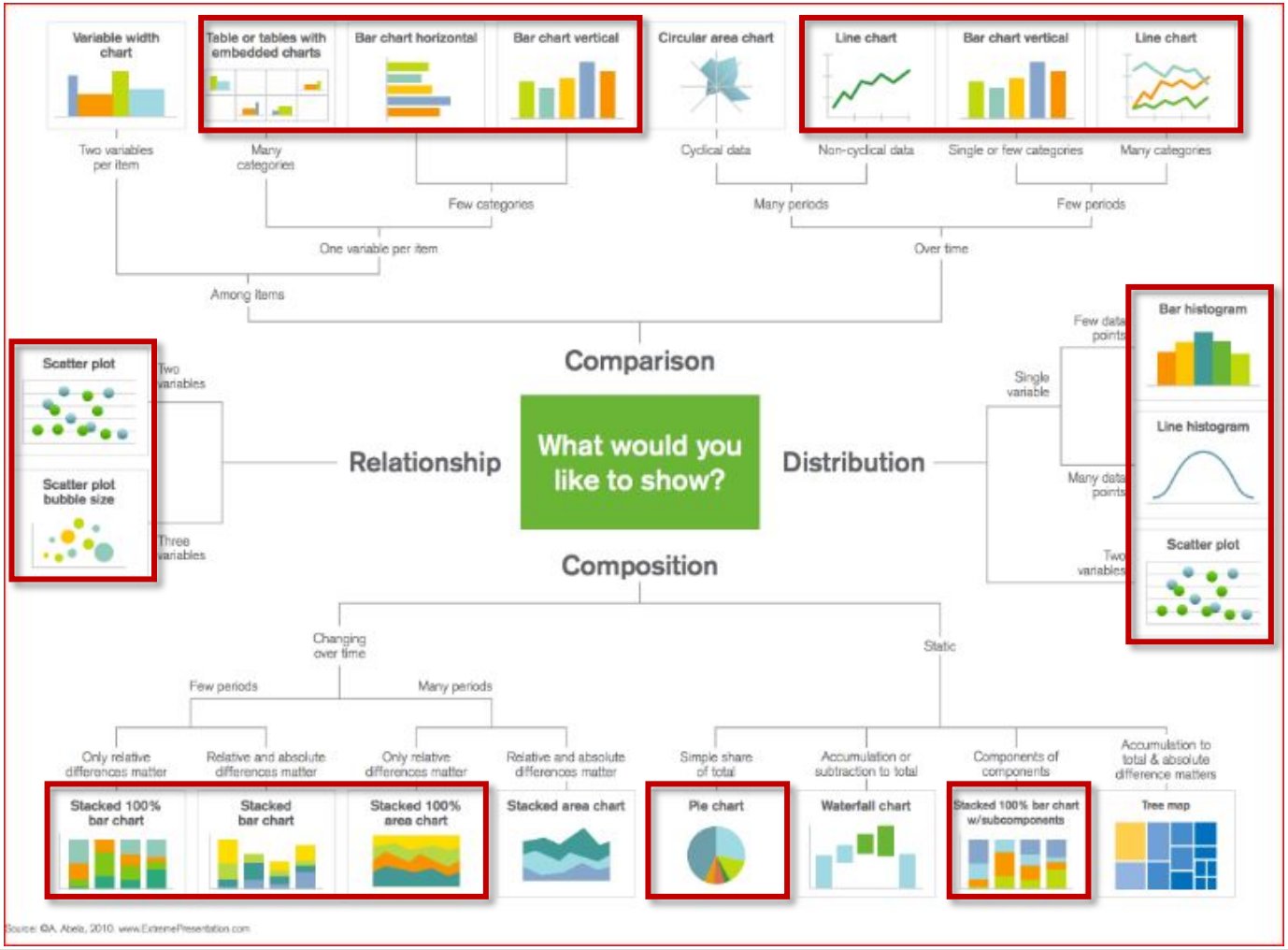
대상변수	자료형		기준					
변수명	기본형	수집형	변수개수	시간축	집계단위	도량	관점	방식

▫ 차트별 시각화 요소 참고표

	표현 가능 변수 개수							데이터 타입별 시각화 목표			
	1변수	2변수	3변수	4변수	5변수	6변수	7변수~	비교	관계	구성	분포
히스토그램_선형	○										○
히스토그램_막대형	○										○
박스플랏	○										○
바이오플랏	○										○
산점도_기본		○							○		○
산점도_버블			○						○		○
산점도_버블컬러				○					○		○
산점도_기본_3차원			○						○		○
산점도_버블_3차원				○					○		○
산점도_버블컬러_3차원					○				○		○
선형도		○	○					○			
레이더차트			○	○	○	○		○			
임베딩차트		○						○			
변수폭차트		○						○			
바차트_기본	○	○						○			
바차트_누적합		○								○	
바차트_누적		○								○	
영역차트_누적합		○								○	
영역차트_누적		○								○	
파이차트	○									○	
폭포차트		○								○	
트리 덴드로그램		○	○	○	○	○	○			○	
트리맵		○	○	○	○	○	○			○	
생키다이어그램		○	○	○	○	○	○			○	

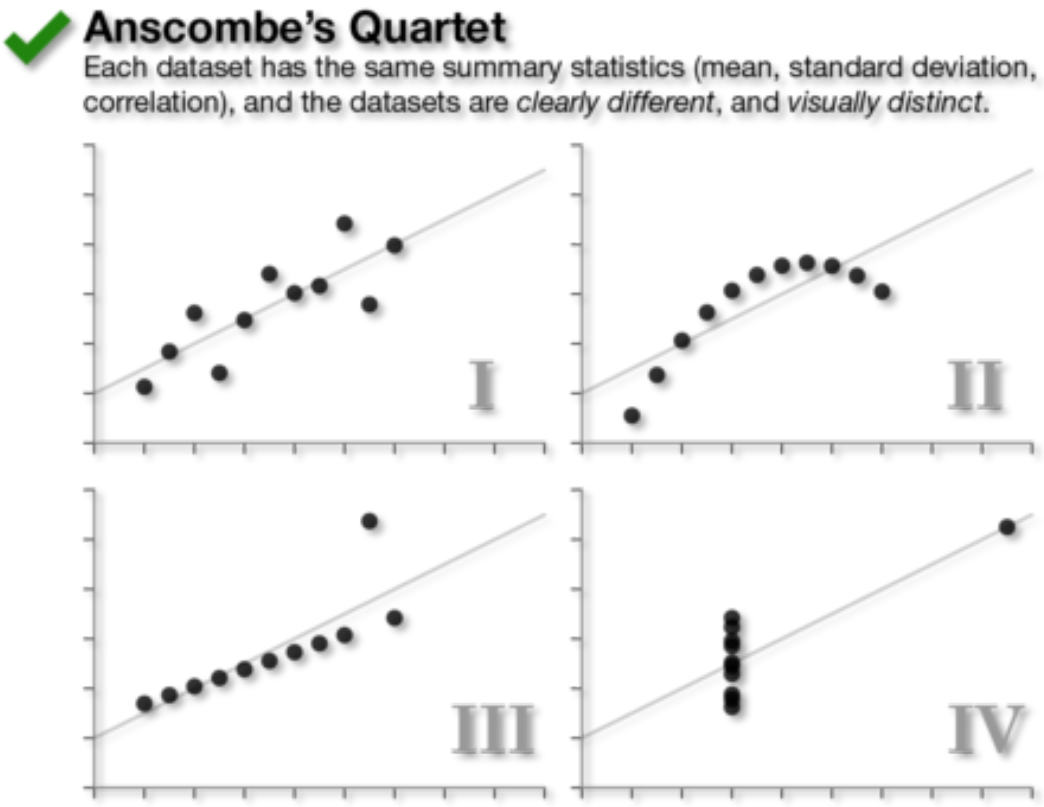
1. Concepts of Visualization | 데이터 유형에 따른 시각화 기법

용도에 맞는 도구 혹은 도구에 맞는 용도

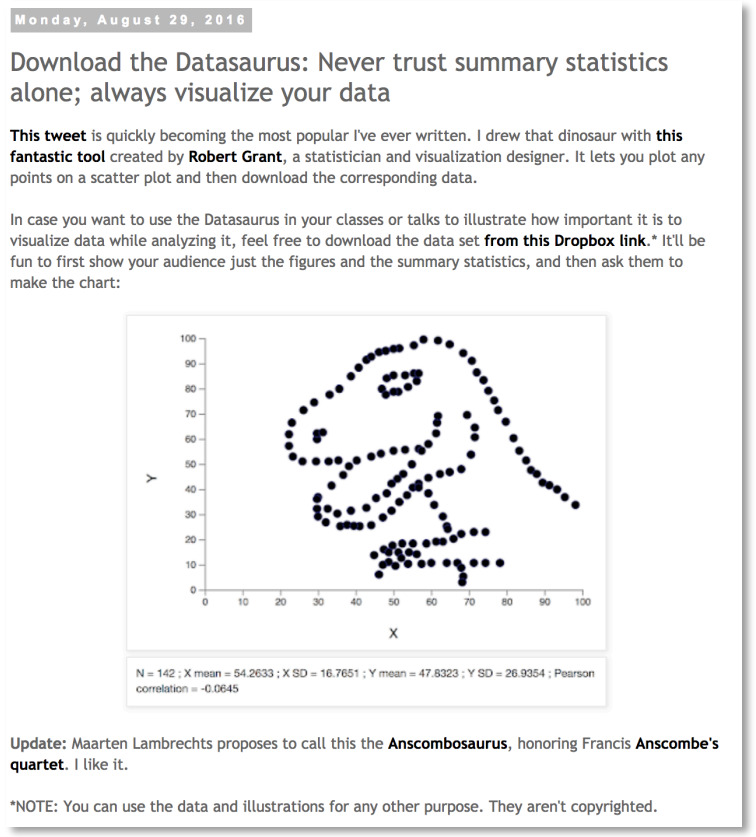


2. 탐색적 자료 분석(EDA) | EDA의 중요성

Anscombe's Quartet과 Datasaurus



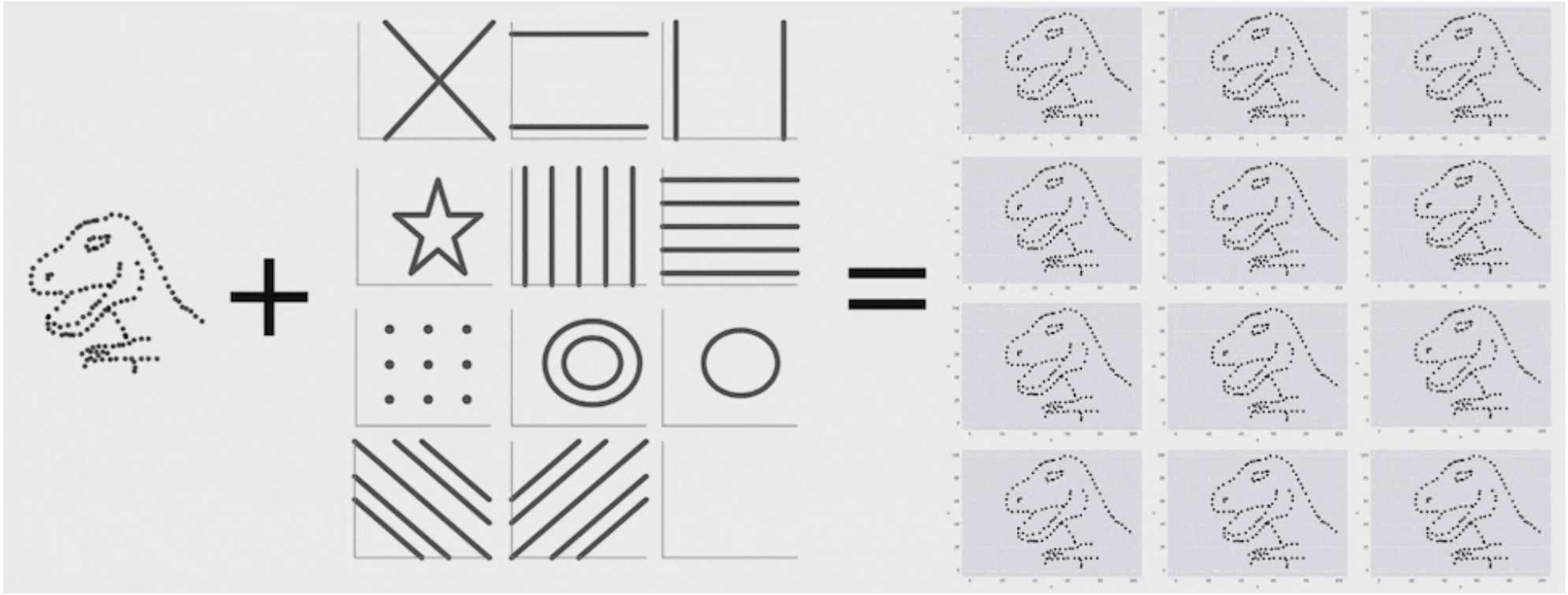
'Anscombe's Quartet', FJ Anscombe, 1973



'Datasaurus', Albert Cairo

2. 탐색적 자료 분석(EDA) | 데이터 유형에 따른 시각화 기법

공룡 12마리



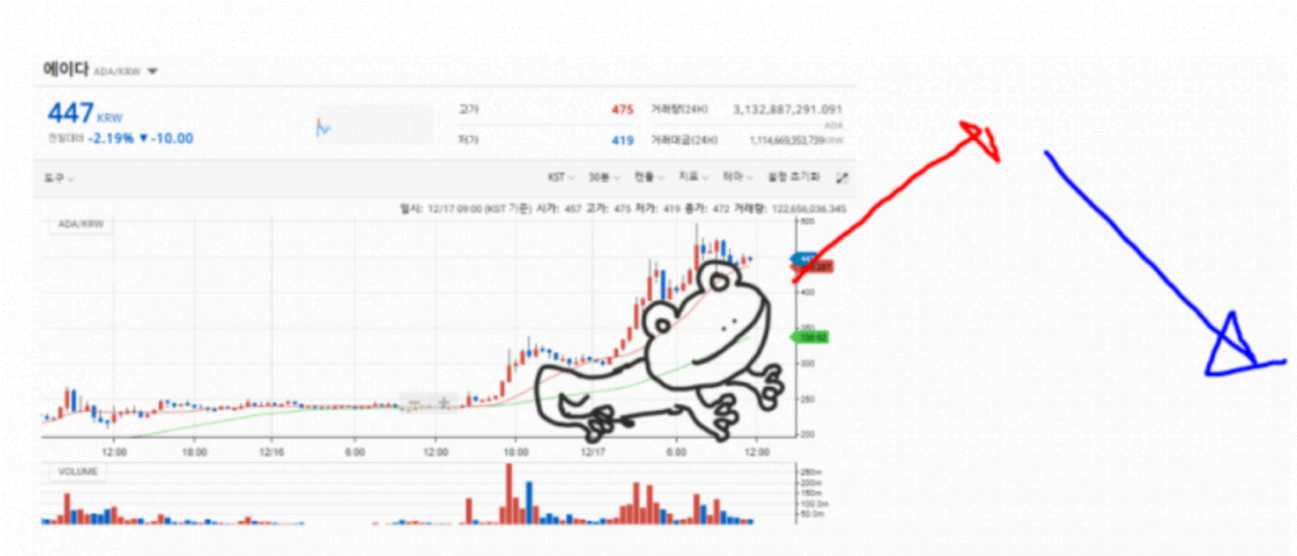
‘The Datasaurus Dozen’

2. 탐색적 자료 분석(EDA) | EDA의 중요성

EDA는 사실 어려운 개념이 아님

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

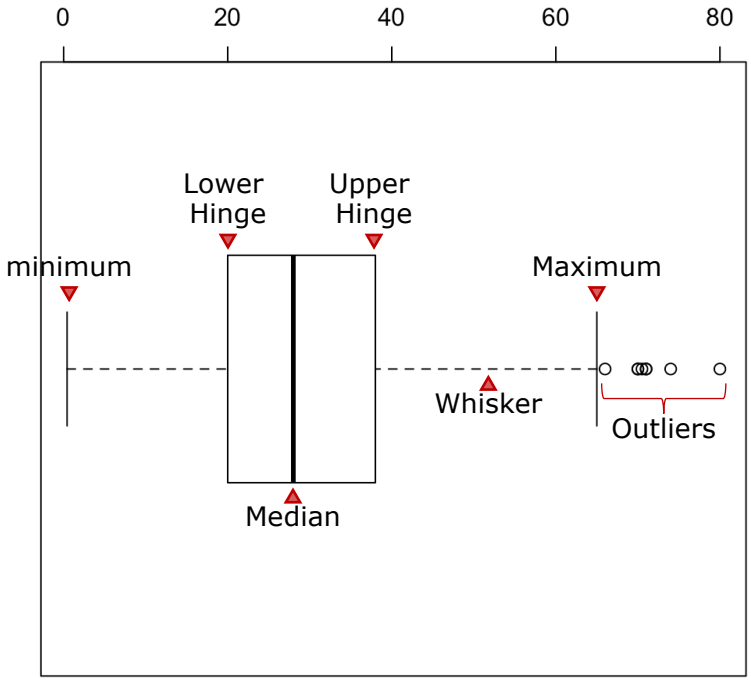
1. 기본 통계 요약
   - ✓ 시각화의 Key
2. 시각화
   - ✓ EDA는 시각화에서 시작하여 시각화로 끝
3. 왜도/첨도
   - ✓ 자료의 분포 패턴을 정량적으로 파악
4. 변수 변환
   - ✓ 왜도/첨도 등을 보정하여 분석 가능한 형태로 변환
5. 평활
   - ✓ 추세선의 과적합 방지
6. 극단값 처리
   - ✓ 절단(trimming) – 특정 극단값 제거
   - ✓ 조정(windsorizing) – 특정 극단값 변환(min or max)

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