데이터 시각화

07. 수량과 분포 데이터의 시각화

최대영 교수

고려사이버대학교
THE CYBER UNIVERSITY OF KOREA

Data Visualization

데이터 시각화

07주차

수량과 분포 데이터의 시각화

최대영 교수

1

학습리뷰

1 시각화 평가

시각화 과정에서 평가해야 할 내용

- 시각화로 해결하려고 하는 문제 정의와 평가
- 문제를 해결하기 위한 데이터와 과업의 결정
- 문제, 데이터, 과업에 맞는 시각화 인코딩 선택
- 시각화 맵핑 알고리즘 구현
- 시각화 상호작용 설계

2

학습리뷰

1 시각화 평가

전체적인 평가의 필요성

- 시각화의 목표를 달성했는지 평가하기 위한 전체적인 관점 필요

전체적인 평가의 방법

- 통찰 기반 평가(Insight-based Evaluation) / 정성적 평가(Qualitative Evaluation)
- 실험 기반 평가(Experimental Evaluation)

3

학습리뷰

2 통찰 기반 평가와 실험 기반 평가

통찰 기반 평가

- 사람들에게 시각화 시스템(또는 도구)을 제공하고 그 시스템이 무엇을 가능하게 하는지를 이해함으로써 시스템의 유용성(utility)을 평가

실험 기반 평가

- 통제 연구(controlled study)를 통해 사람들이 서로 다른 시각화 방법을 사용하여 얼마나 빠르고, 정확하고, 효율적으로 과업을 완성하는지 측정하는 방법

4

학습목표

- 수량 데이터의 특징과 시각화 방법에 대해 설명할 수 있다.
- 분포 데이터의 특징과 시각화 방법에 대해 설명할 수 있다.
- 수량과 분포 데이터 관련 matplotlib 라이브러리를 이해하고 활용할 수 있다.

5

학습내용

1. 수량 데이터의 시각화
2. 분포 데이터의 시각화
3. 실습

6

01
chapter

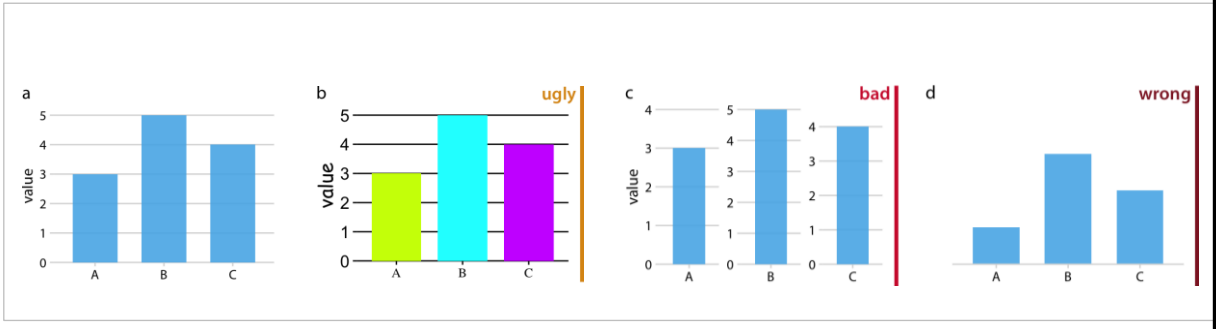
수량 데이터의 시각화

7

1. 막대 도표

문제점이 있는 도표에 대한 표시

- ugly(조악함): 미적으로 빼어나지 않음, 내용은 분명하고 유용
- bad(모호함): 내용을 오도함, 불분명하거나 복잡하여 오해의 소지가 있음
- wrong(틀림): 수학적으로 틀리거나 객관적으로 사실이 아님



[출처] Fundamentals of Data Visualization

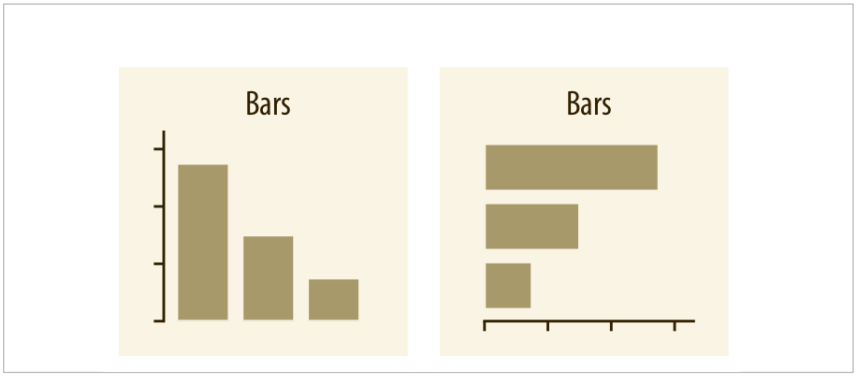
8

1. 막대 도표

막대 도표(Bar plot)

수치 집합의 크기(수량)를 나타내야 하는 경우에 사용

예 자동차 브랜드별 판매량, 도시별 거주 인구 등



[출처] Fundamentals of Data Visualization

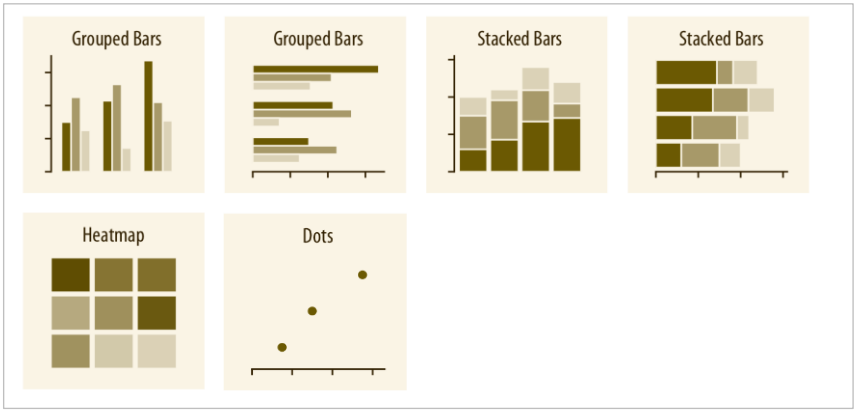
9

1. 막대 도표

막대 도표(Bar plot)

막대 도표의 종류: 기본 막대, 묶은(grouped) 막대, 누적(stacked) 막대

- 막대 도표 대신 점 도표(dot plot)나 히트맵(heatmap) 사용 가능



[출처] Fundamentals of Data Visualization

10

1. 막대 도표

주말 최고 인기 영화 순위 데이터

순위	영화 제목	주말 수익
1	Star Wars: The Last Jedi	$71,565,498
2	Jumanji: Welcome to the Jungle	$36,169,328
3	Pitch Perfect 3	$19,928,525
4	The Greatest Showman	$8,805,843
5	Ferdinand	$7,316,746

[출처] Fundamentals of Data Visualization

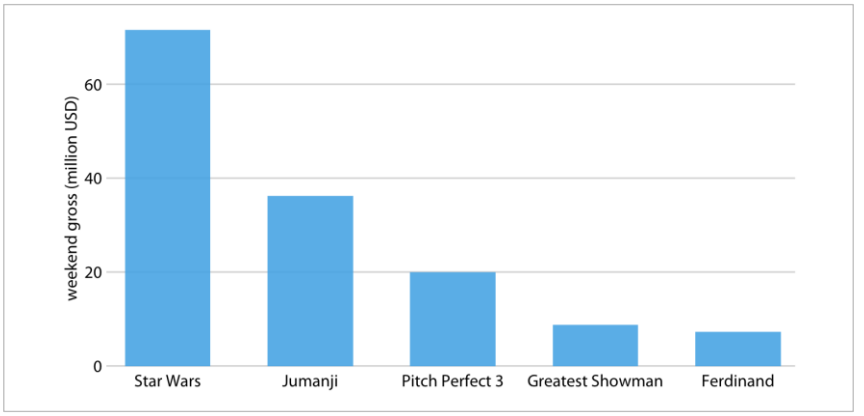
11

1. 막대 도표

기본 막대 도표

레이블(영화 제목)을 가로로 넣는 형태 → 공간을 낭비

주말 최고 인기 영화 순위 예시



[출처] Fundamentals of Data Visualization

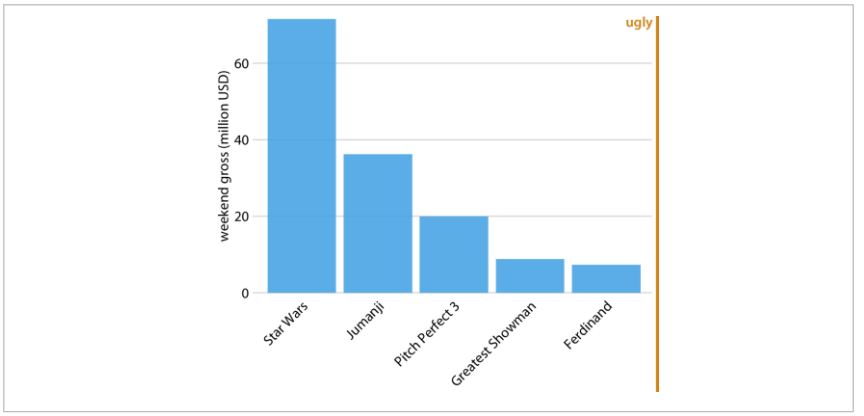
12

1. 막대 도표

기본 막대 도표

레이블을 기울여서 넣은 형태 → 공간을 아낄 수 있으나 글자를 읽기 힘듦

주말 최고 인기 영화 순위 예시



[출처] Fundamentals of Data Visualization

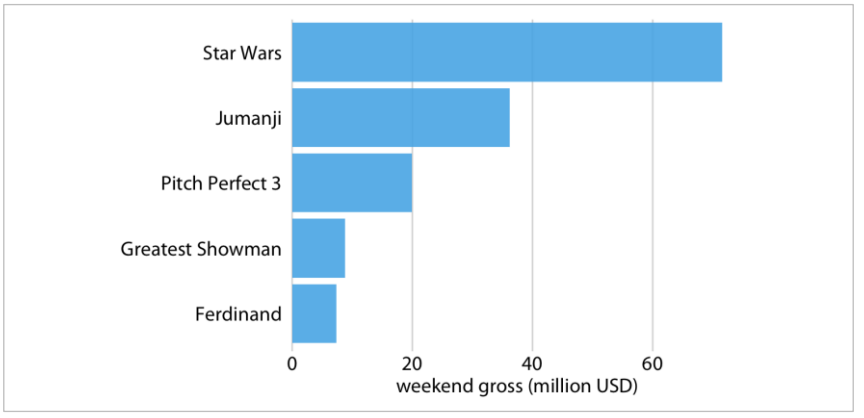
13

1. 막대 도표

기본 막대 도표

가로축으로 막대를 그려 넣은 형태 → 정보를 읽기 쉬움

주말 최고 인기 영화 순위 예시



[출처] Fundamentals of Data Visualization

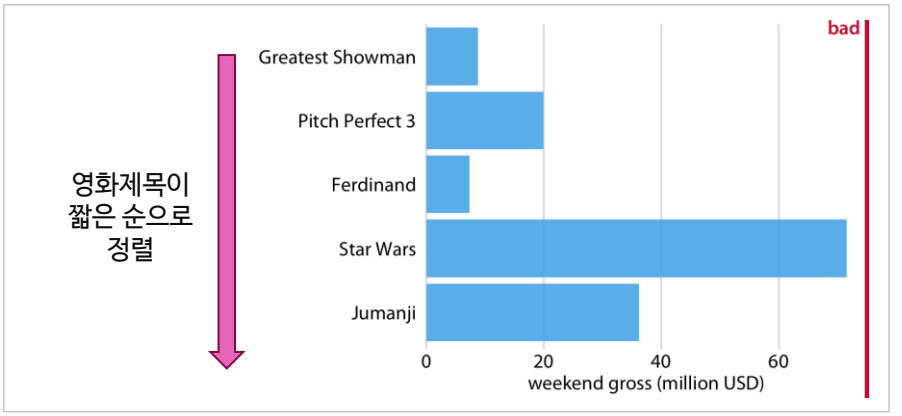
14

1. 막대 도표

기본 막대 도표

막대의 순서를 의미 있게 정하는 것이 매우 중요

주말 최고 인기 영화 순위 예시



[출처] Fundamentals of Data Visualization

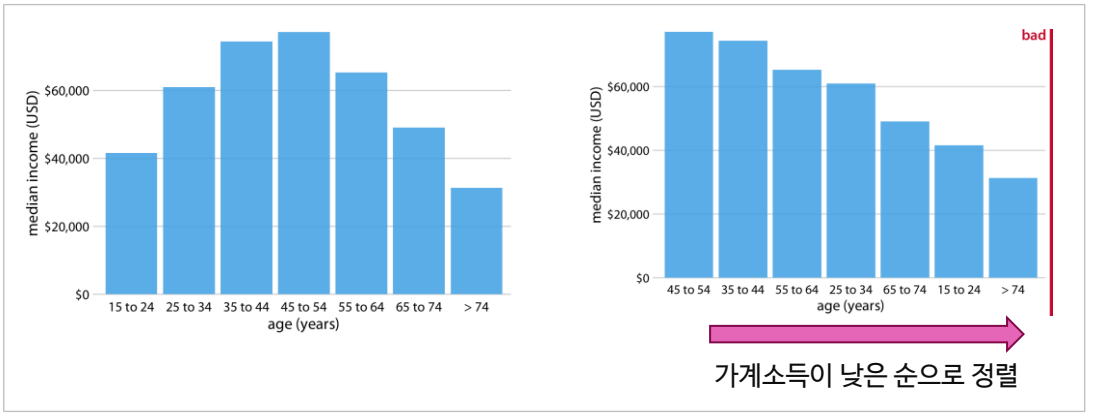
15

1. 막대 도표

기본 막대 도표

막대의 순서를 의미 있게 정하는 것이 매우 중요

미국의 연령별 연간 중위 가계소득 예시



[출처] Fundamentals of Data Visualization

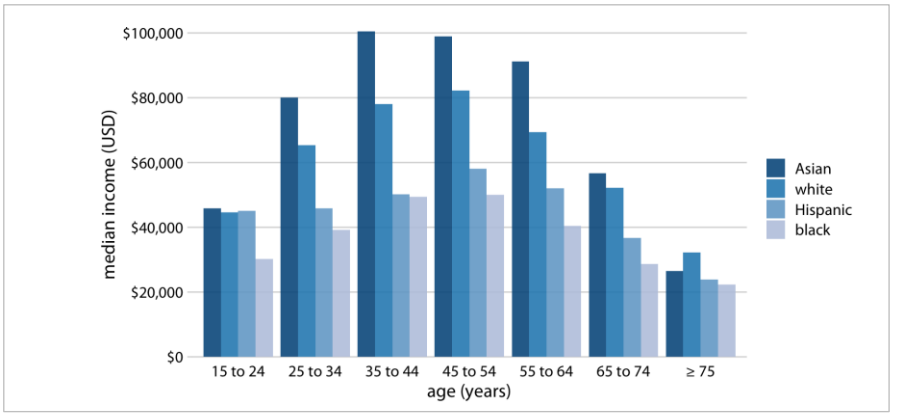
16

1. 막대 도표

묶은 막대 도표(Grouped bar)

두 범주(연령대, 인종)를 동시에 표현 → 한 인종의 연령대별 비교가 어려움

미국의 연령별 연간 중위 가계소득 예시



[출처] Fundamentals of Data Visualization

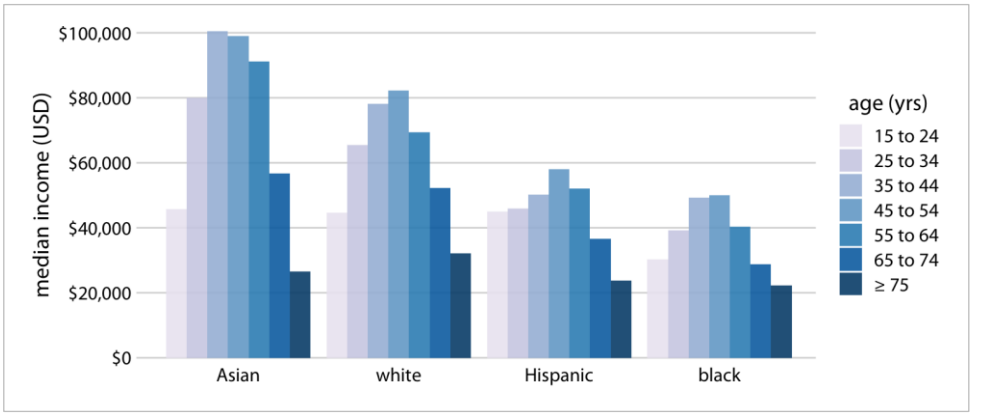
17

1. 막대 도표

묶은 막대 도표(Grouped bar)

인종을 묶고 색으로 연령대를 표현 → 한 연령대의 인종별 비교가 어려움

미국의 연령별 연간 중위 가계소득 예시



[출처] Fundamentals of Data Visualization

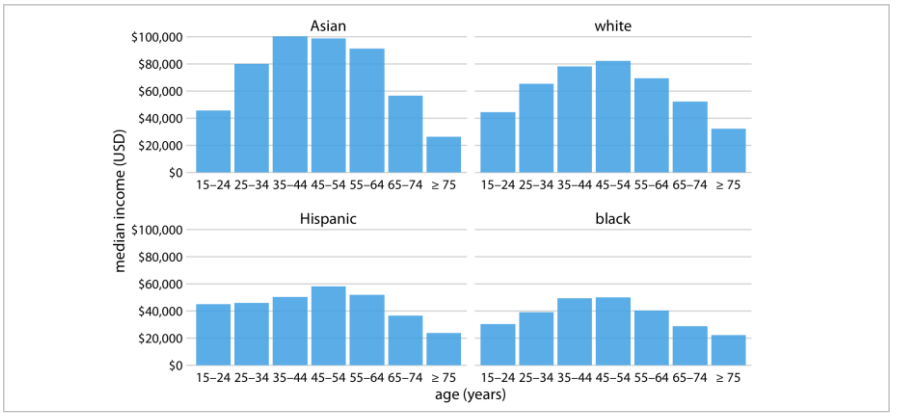
18

1. 막대 도표

묶은 막대 도표(Grouped bar)

범주별로 기본 막대 도표를 그리는 방법 → 인지의 부담이 줄어듦

미국의 연령별 연간 중위 가계소득 예시



[출처] Fundamentals of Data Visualization

19

1. 막대 도표

누적 막대 도표(Stacked bar)

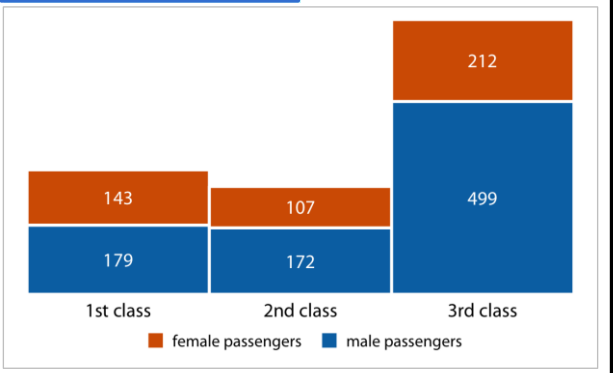
막대들을 쌓아서 합을 도출하는 것이 의미가 있는 경우에 사용

- 타이타닉의 객실 등급별 탑승인원 남녀 합계

막대에 실제 데이터 값을 표현하고 y축을 생략

- 정보를 더욱 간결하게 전달 가능

타이타닉 승객수 예시



[출처] Fundamentals of Data Visualization

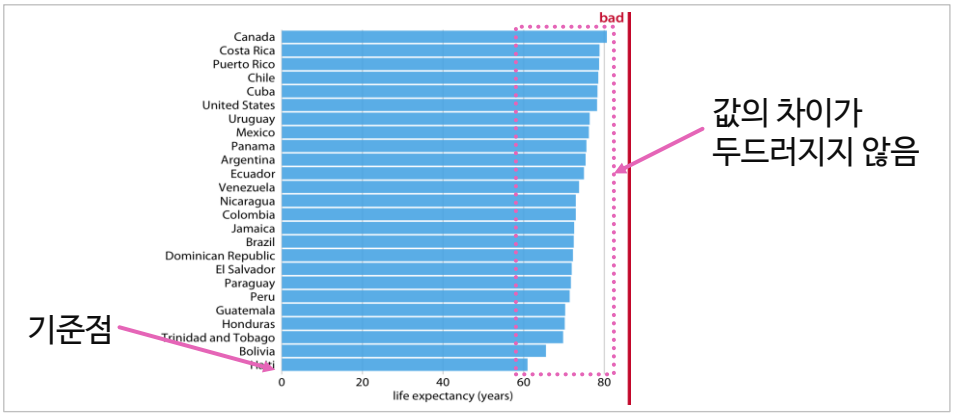
20

2. 점 도표와 히트맵

점 도표(Dot plot)

막대 도표가 기준점 0으로 부터 길이로 정량 값을 표현하는 것의 단점을 보완

아메리카 대륙 25개국의 기대수명 데이터 예시



[출처] Fundamentals of Data Visualization

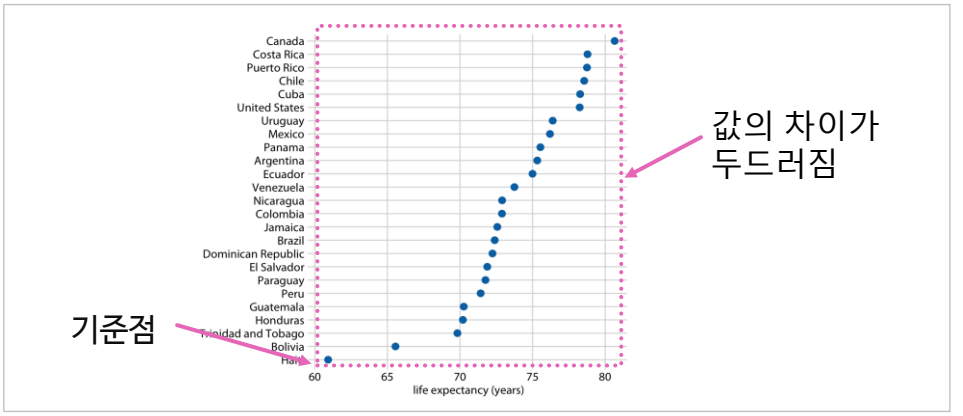
21

2. 점 도표와 히트맵

점 도표(Dot plot)

x축의 범위를 60~81로 제한하여 값의 차이가 두드러지게 표현

아메리카 대륙 25개국의 기대수명 데이터 예시



[출처] Fundamentals of Data Visualization

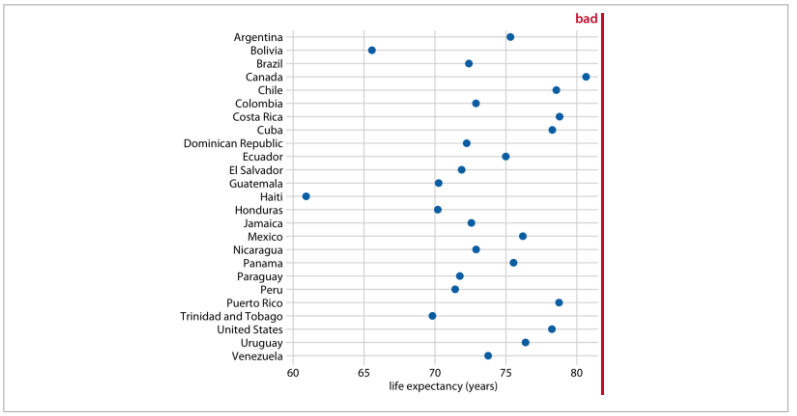
22

2. 점 도표와 히트맵

점 도표(Dot plot)

국가명을 알파벳순으로 정렬 → 요점을 알기 어려움, 국가를 찾기는 쉬움

아메리카 대륙 25개국의 기대수명 데이터 예시



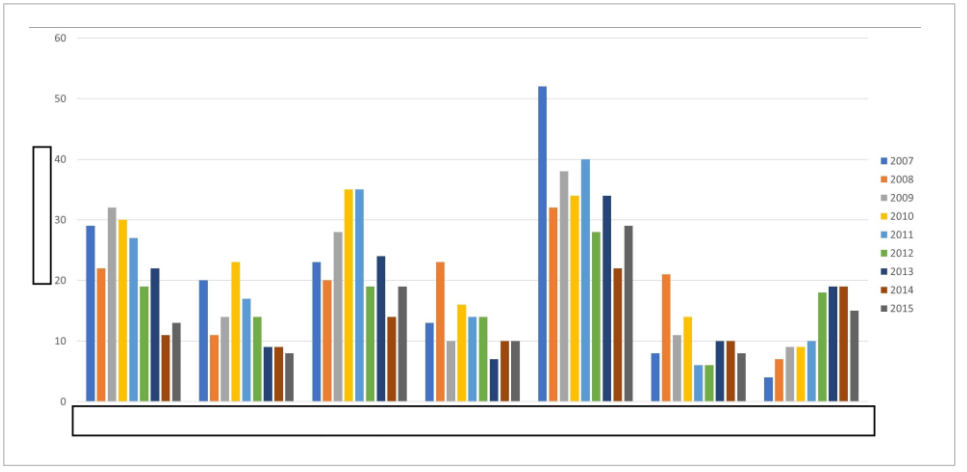
[출처] Fundamentals of Data Visualization

23

2. 점 도표와 히트맵

히트맵(Heatmap)

막대 도표와 점 도표는 데이터 양이 많은 경우 요점을 전달하기 어려움



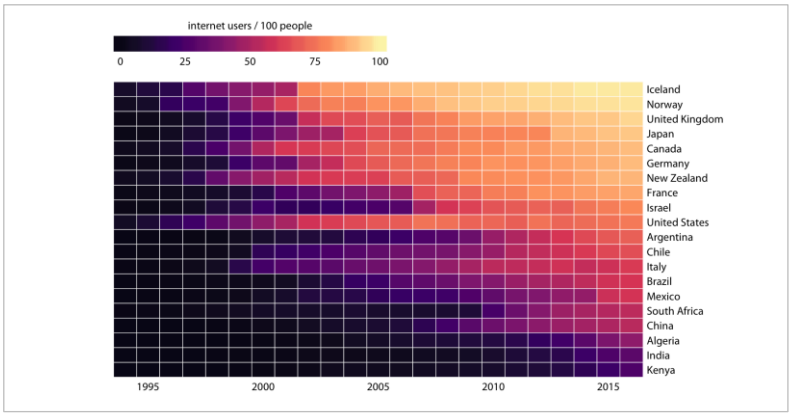
24

2. 점 도표와 히트맵

히트맵(Heatmap)

막대나 점 대신 색으로 데이터 값을 표현 → 전반적인 추세 확인이 쉬움

국가별 인터넷 보급률 예시



[출처] Fundamentals of Data Visualization

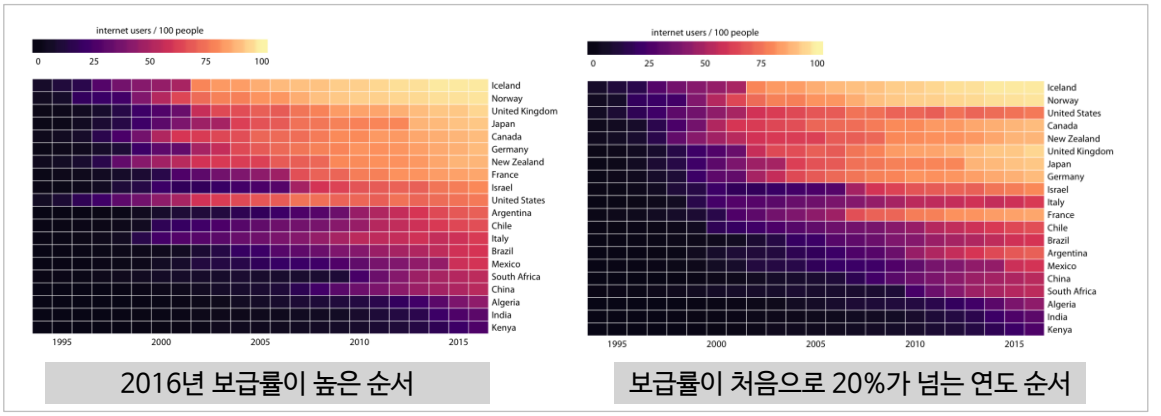
25

2. 점 도표와 히트맵

히트맵(Heatmap)

전하려는 메시지에 따라 국가 정렬 순서를 변경

국가별 인터넷 보급률 예시



2016년 보급률이 높은 순서

보급률이 처음으로 20%가 넘는 연도 순서

[출처] Fundamentals of Data Visualization

26

02

chapter

분포 데이터의 시각화

27

1. 단일 분포의 시각화

타이타닉 승선객의 연령 분포(Age distribution)

연령	인원	연령	인원	연령	인원
0~5	36	26~30	121	51~55	26
6~10	19	31~35	76	56~60	22
11~15	18	36~40	74	61~65	16
16~20	99	41~45	54	66~70	3
21~25	129	46~50	50	71~75	3

[출처] Fundamentals of Data Visualization

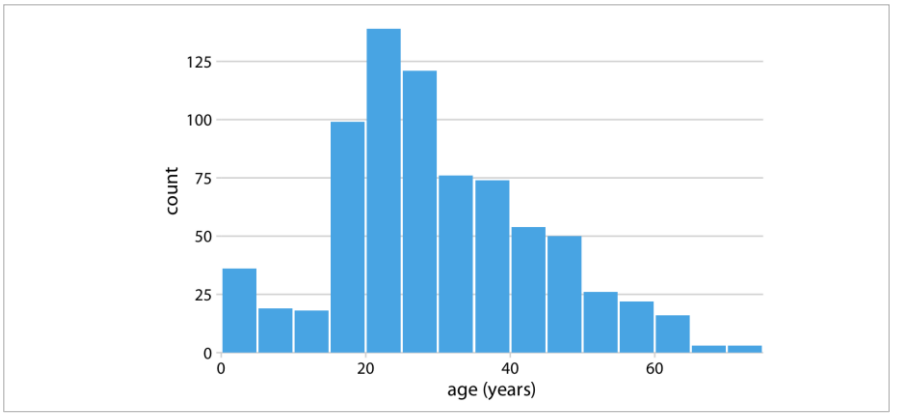
28

1. 단일 분포의 시각화

히스토그램(Histogram)

가로에 범주(연령) 구간, 세로에 구간에 포함된 데이터의 수(인원수)를 표현

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

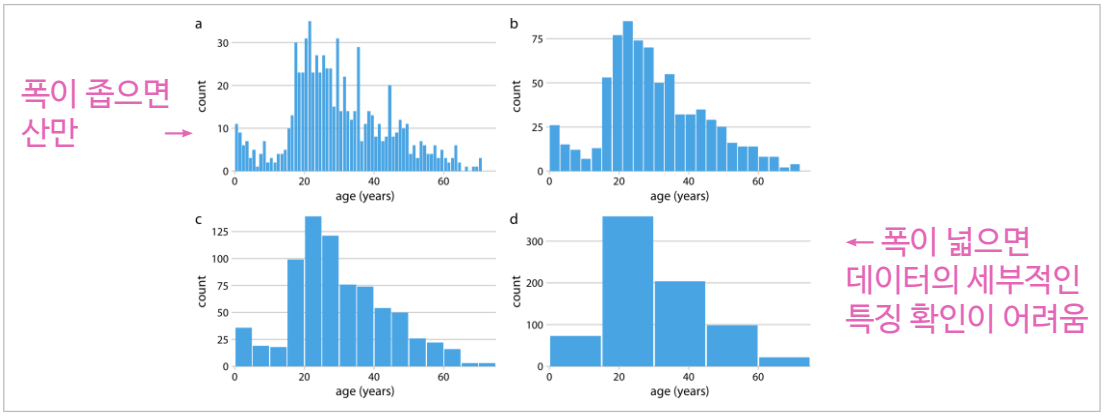
29

1. 단일 분포의 시각화

히스토그램(Histogram)

구간의 폭에 따라 시각화된 분포의 모양이 달라짐

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

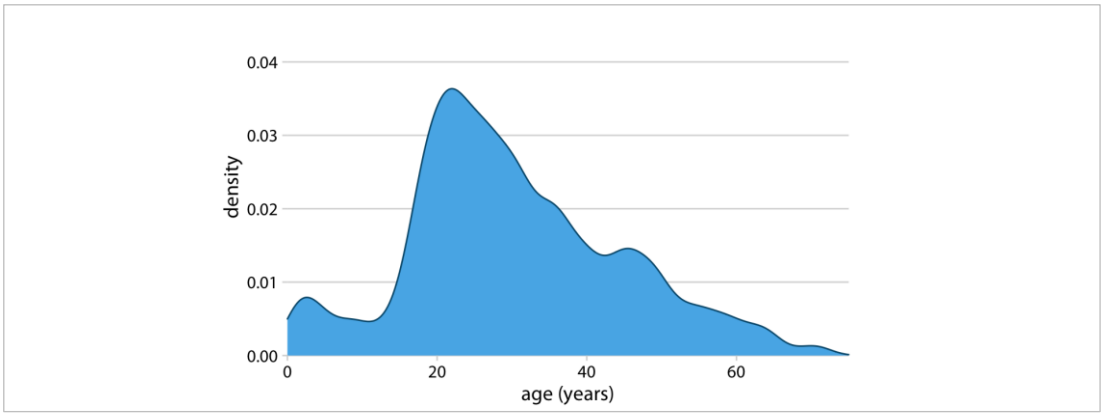
30

1. 단일 분포의 시각화

밀도 도표(Density plot)

데이터의 분포를 연속적인 곡선으로 표현, 곡선 아래의 면적은 1

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

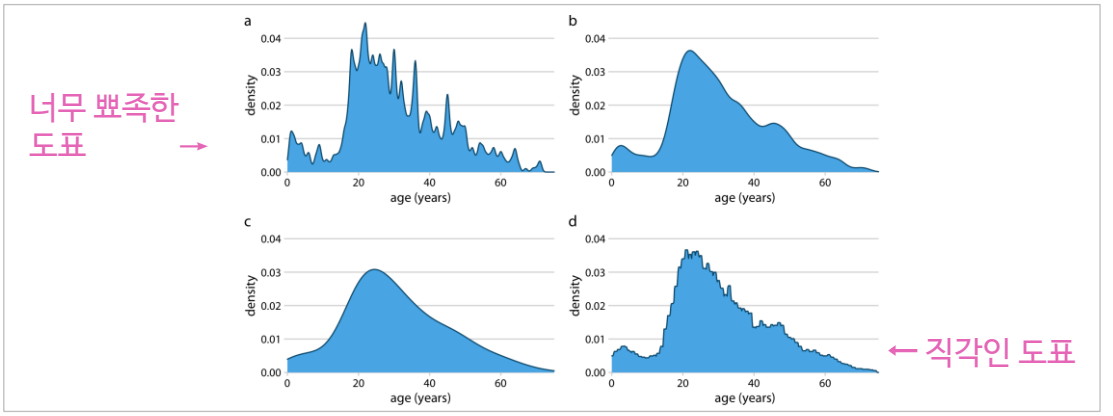
31

1. 단일 분포의 시각화

밀도 도표(Density plot)

밀도 도표를 그리는 방법에 따라 분포가 다르게 표현됨

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

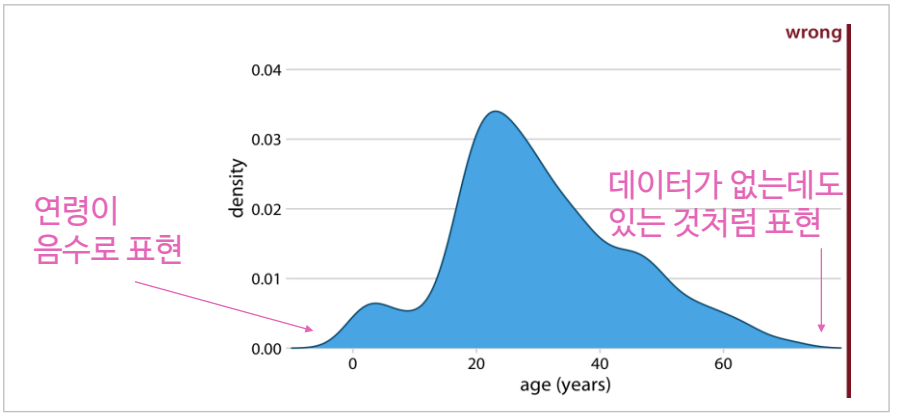
32

1. 단일 분포의 시각화

밀도 도표(Density plot)

데이터를 기반으로 분포를 추정하기 때문에 잘못된 해석이 가능

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

33

1. 단일 분포의 시각화

히스토그램과 밀도 도표의 단점

구간 폭, 모양새 등 표현방법에 따라 결과물이 크게 달라짐

- 원본 데이터를 보여주는 것이 아닌 데이터의 해석에 가까움

대안으로 원본 데이터를 모두 점으로 표시하는 방법이 가능

- 그러나, 데이터의 분포를 나타내는 데는 적절하지 않음

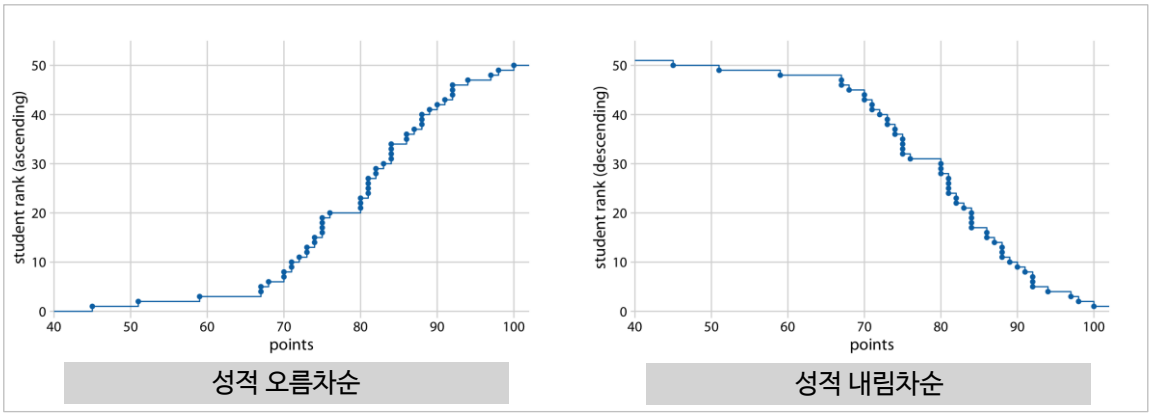
34

1. 단일 분포의 시각화

경험적 누적 분포 함수(Empirical cumulative distribution function)

모든 데이터(성적)를 표현하면서 동시에 분포도 나타내는 방법

50명 학생의 성적분포 예시



성적 오름차순 | 성적 내림차순

[출처] Fundamentals of Data Visualization

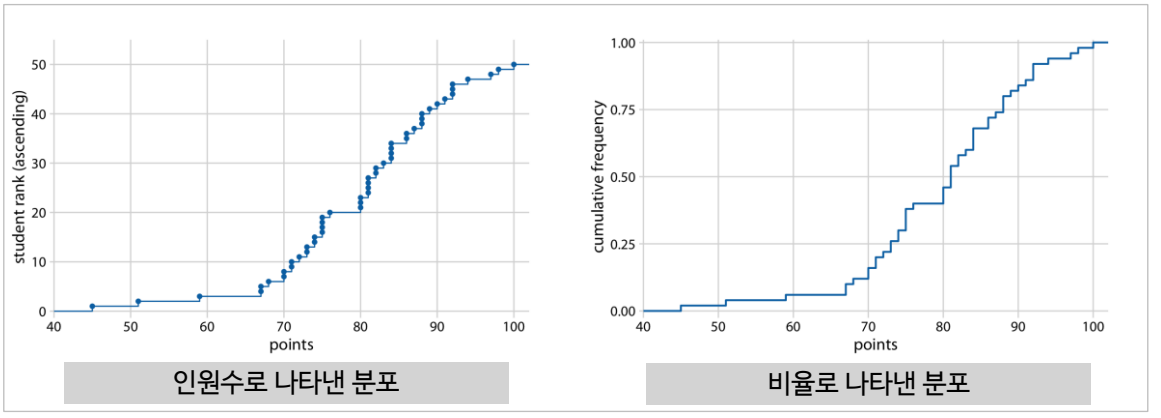
35

1. 단일 분포의 시각화

경험적 누적 분포 함수(Empirical cumulative distribution function)

y축을 비율로 변환하면 분포를 계산하기 쉬워 짐

50명 학생의 성적분포 예시



인원수로 나타낸 분포 | 비율로 나타낸 분포

[출처] Fundamentals of Data Visualization

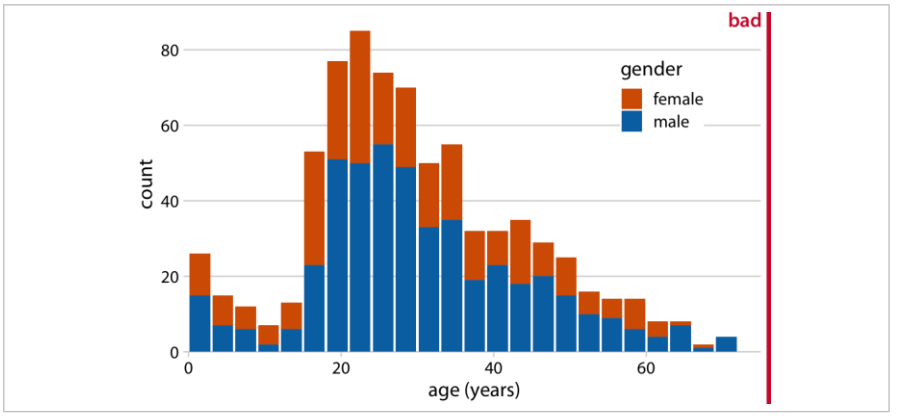
36

2. 여러 분포의 시각화

누적 히스토그램(Stacked histogram)

두 개 이상의 변수의 분포를 하나의 도표에 나타내는 경우에 사용

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

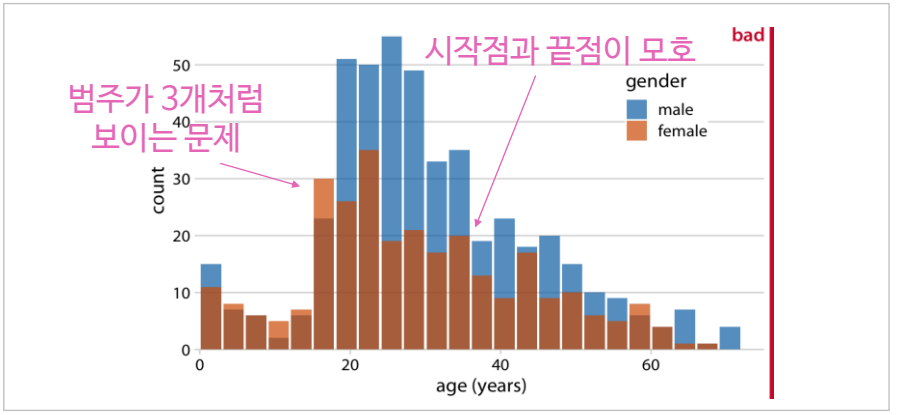
37

2. 여러 분포의 시각화

누적 히스토그램(Stacked histogram)

막대의 투명도를 높여 시작 위치를 표시

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

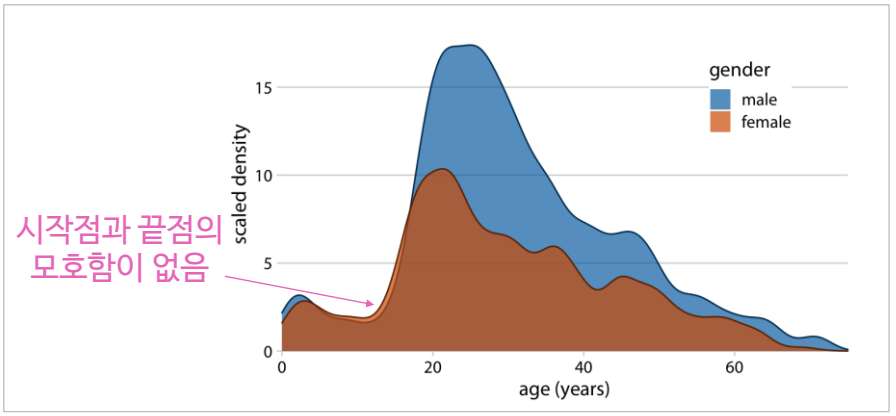
38

2. 여러 분포의 시각화

중첩 밀도 도표(Overlapping density plot)

두 개 이상의 곡선으로 된 밀도 도표를 겹쳐서 표현

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

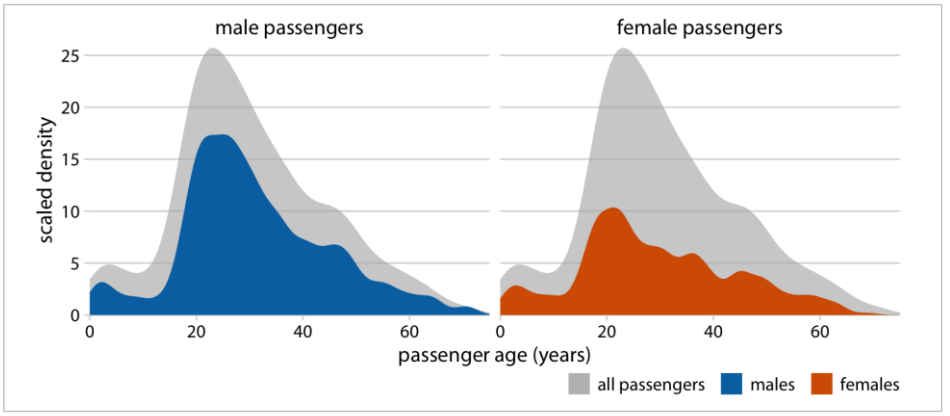
39

2. 여러 분포의 시각화

중첩 밀도 도표(Overlapping density plot)

성별로 나눠서 분포를 표현하면 데이터의 특징을 더 잘 파악할 수 있음

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

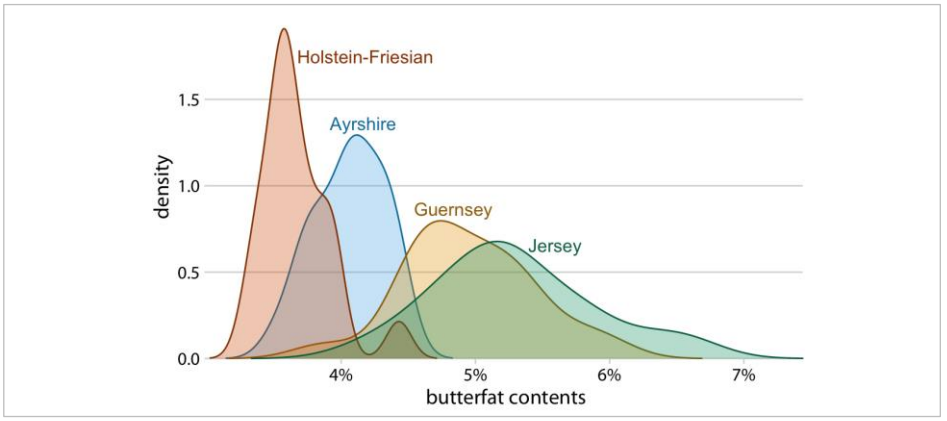
40

2. 여러 분포의 시각화

중첩 밀도 도표(Overlapping density plot)

중첩된 여러 분포를 간명하게 표현 가능

4종류 젖소의 버터지방 함량 예시



[출처] Fundamentals of Data Visualization

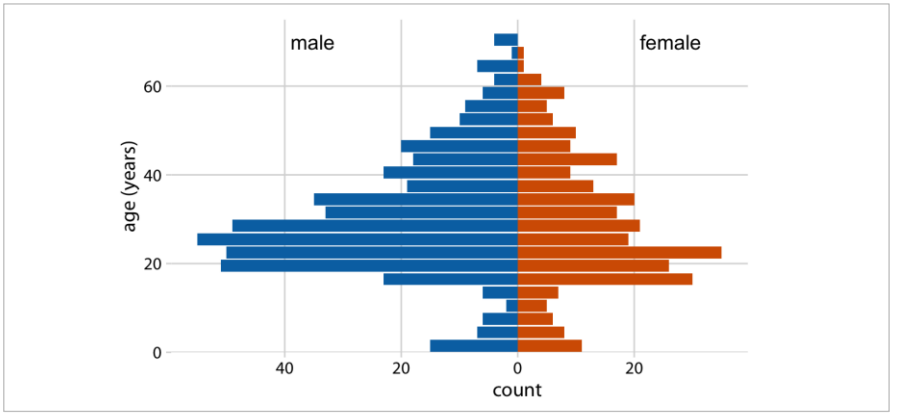
41

2. 여러 분포의 시각화

연령 피라미드(Age pyramid)

히스토그램 2개를 맞대어 표현 → 분포가 3개 이상일 경우에는 그릴 수 없음

타이타닉 승선객 예시



[출처] Fundamentals of Data Visualization

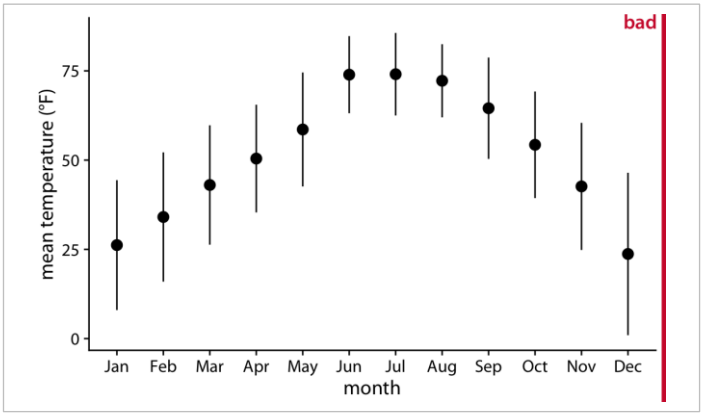
42

2. 여러 분포의 시각화

여러 분포를 표시하는 방법

점으로 평균 또는 중앙값을 표현, 막대(error bar)로 오차를 표현

- 데이터의 분포를 충분히 표현할 수 없음



네브레스카 링컨시의 일별 기온

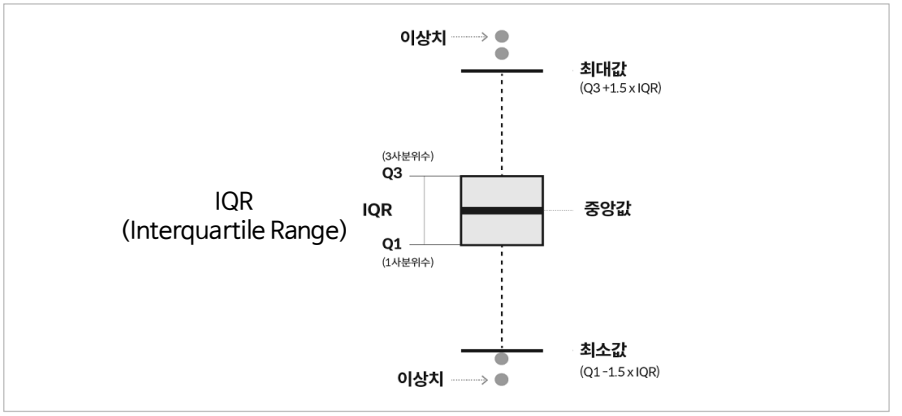
43

2. 여러 분포의 시각화

박스 도표(Box plot)

데이터를 사분위로 나누고 박스안에 50%의 데이터를 표현

- Q1는 25%, Q3는 75% 위치의 데이터를 나타냄



[출처] Fundamentals of Data Visualization

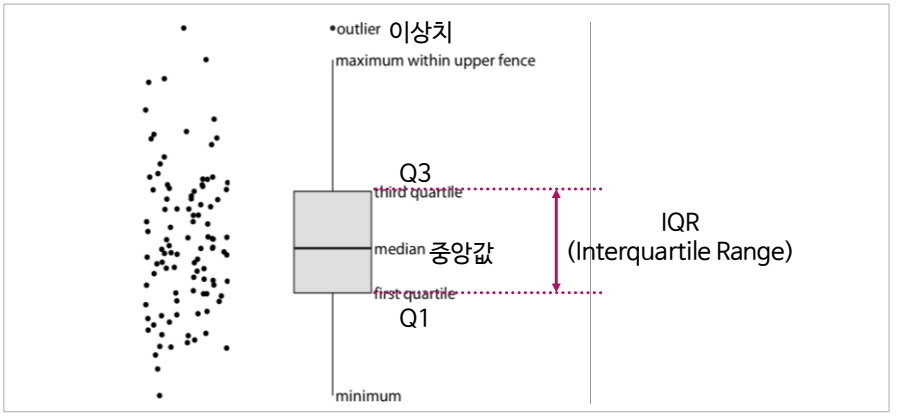
44

2. 여러 분포의 시각화

박스 도표(Box plot)

데이터를 사분위로 나누고 박스안에 50%의 데이터를 표현

- Q1는 25%, Q3는 75% 위치의 데이터를 나타냄



[출처] Fundamentals of Data Visualization

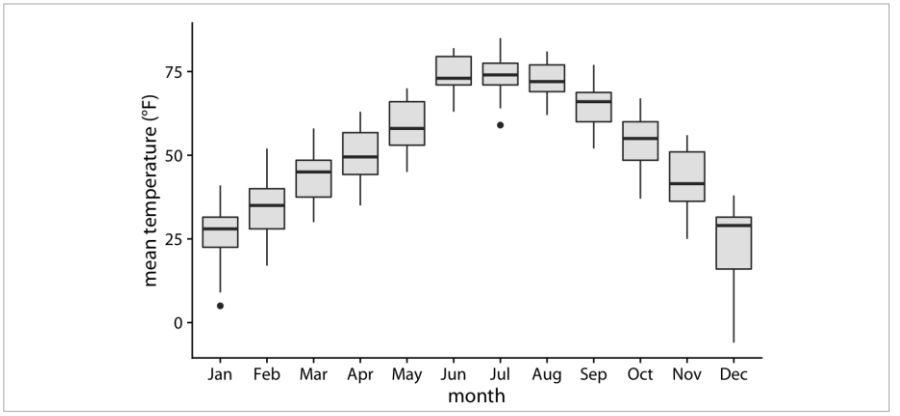
45

2. 여러 분포의 시각화

박스 도표(Box plot)

여러 분포를 간명하게 표현 가능

네브레스카 링컨시의 일별 기온



[출처] Fundamentals of Data Visualization

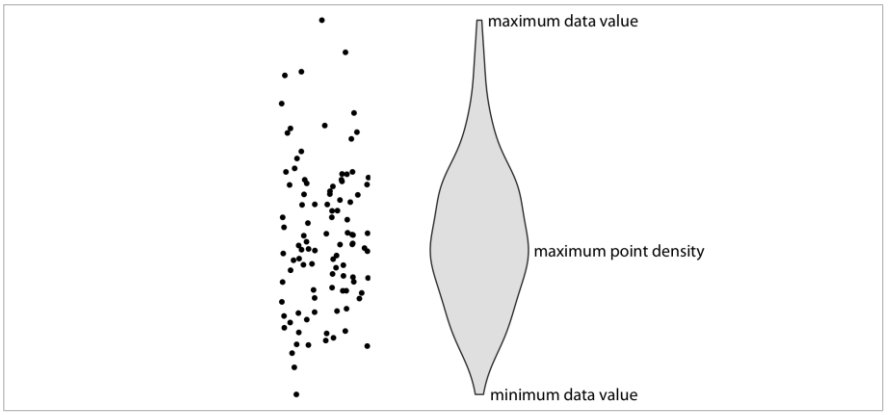
46

2. 여러 분포의 시각화

바이올린 도표(Violin plot)

박스 도표에 비해 데이터의 미묘한 차이를 더 잘 나타낼 수 있음

- 데이터의 양이 충분해야 매끄럽게 표현 가능



[출처] Fundamentals of Data Visualization

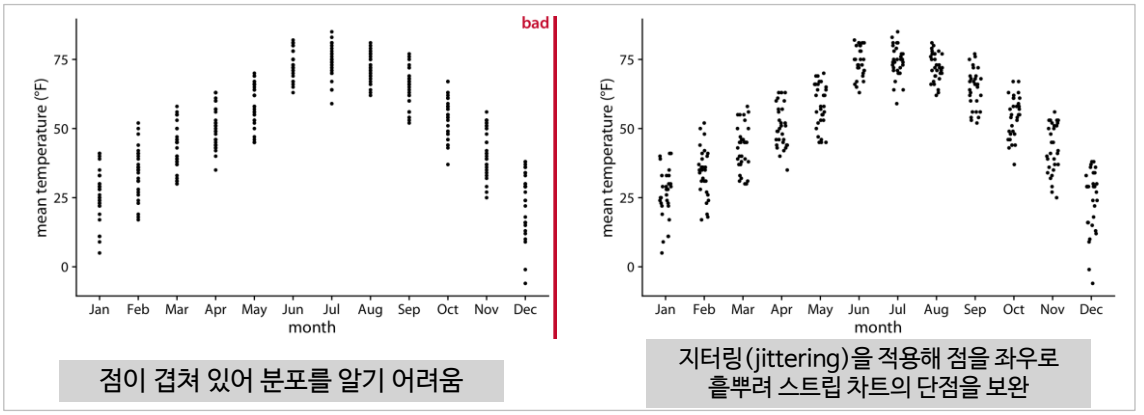
47

2. 여러 분포의 시각화

스트립 차트(Strip chart)

데이터를 점으로 나타내어 분포를 표현

네브레스카 링컨시의 일별 기온



점이 겹쳐 있어 분포를 알기 어려움

지터링(jittering)을 적용해 점을 좌우로 흩뿌려 스트립 차트의 단점을 보완

[출처] Fundamentals of Data Visualization

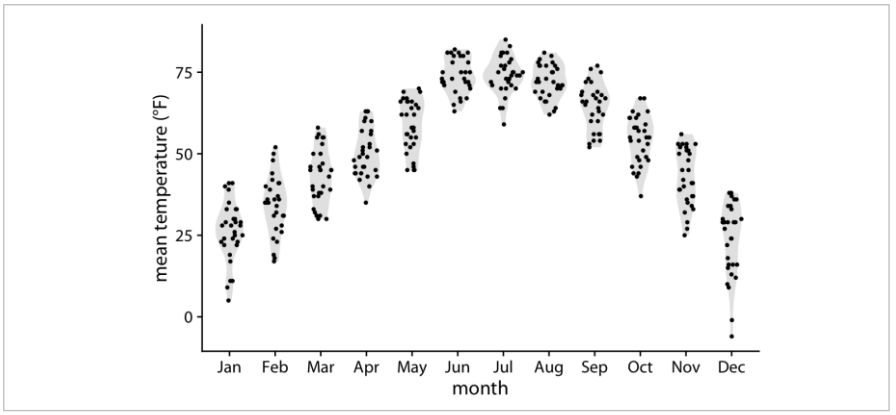
48

2. 여러 분포의 시각화

시나 도표(Sina plot)

바이올린 도표와 지터링한 스트립 도표를 합치는 방법

네브레스카 링컨시의 일별 기온



[출처] Fundamentals of Data Visualization

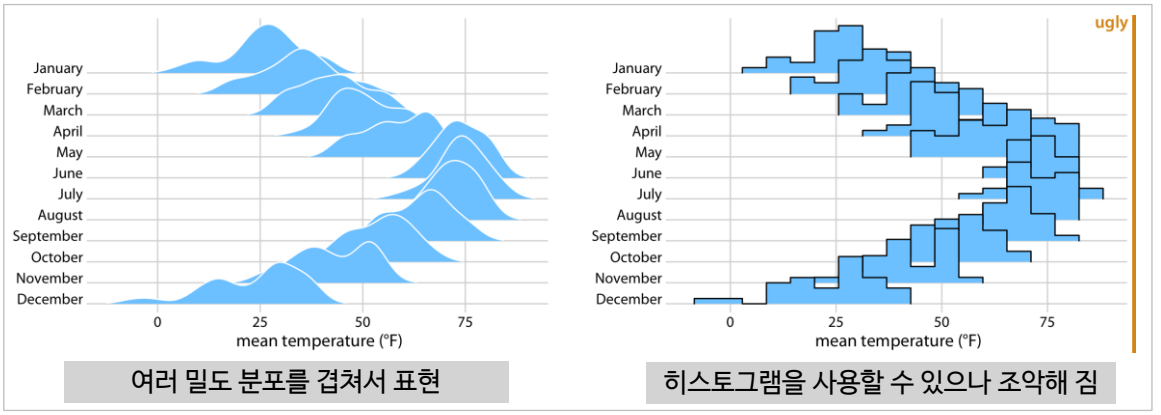
49

2. 여러 분포의 시각화

융기선 도표(Ridgeline plot)

산이 융기한(주변 보다 상승한) 모양을 본뜬 도표

네브레스카 링컨시의 일별 기온



여러 밀도 분포를 겹쳐서 표현

히스토그램을 사용할 수 있으나 조악해 짐

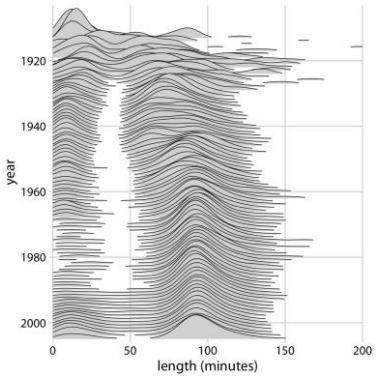
[출처] Fundamentals of Data Visualization

50

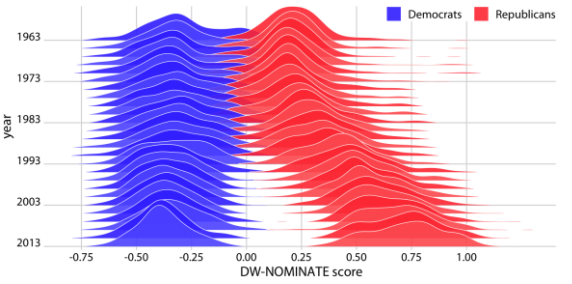
2. 여러 분포의 시각화

융기선 도표(Ridgeline plot)

전체적인 흐름을 파악하는데 용이한 표현



영화 상영시간의 변화



미국 하원의원 투표 양상

[출처] Fundamentals of Data Visualization

51

03

chapter

실습

52

1. 수량 데이터의 시각화

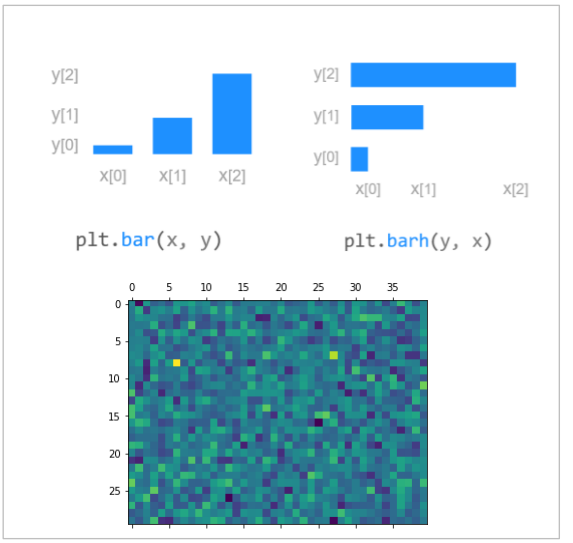
막대 도표와 히트맵

막대 도표

- bar(x, y), barh(y, x)

히트맵

- matshow(arr)



[출처] Matplotlib Tutorial - 파이썬으로 데이터 시각화하기

53

2. 분포 데이터의 시각화

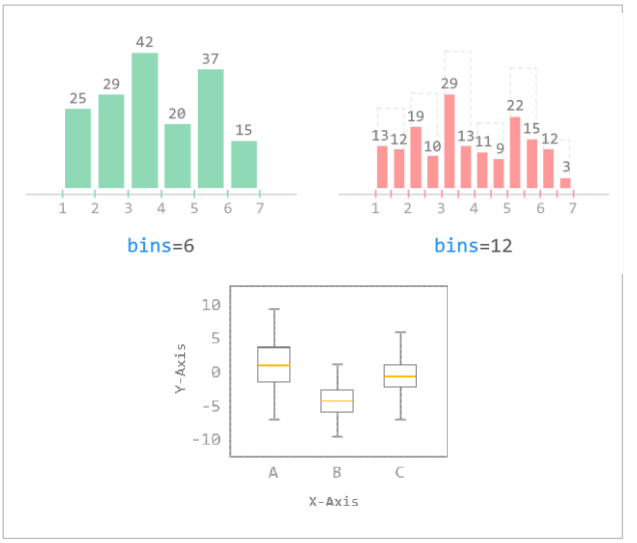
히스토그램과 박스 도표

히스토그램

- hist(x, bins=n_b)

박스 도표

- boxplot(x)



[출처] Matplotlib Tutorial - 파이썬으로 데이터 시각화하기

54

학습정리

1 수량 데이터의 시각화

막대 도표(Bar plot)

- 수치 집합의 크기(수량)를 나타내야 하는 경우에 사용

묶은 막대 도표(Grouped bar)

- 두 가지 이상의 범주를 표현해야 하는 경우에 사용

누적 막대 도표(Stacked bar)

- 막대들을 쌓아서 합을 도출하는 것이 의미가 있는 경우에 사용

55

학습정리

1 수량 데이터의 시각화

점 도표(Dot plot)

- 막대 도표가 기준점 0으로 부터 길이로 정량 값을 표현하는 것의 단점을 보완
- 기준점을 조정하여 데이터를 더 간명하게 표현 가능

히트맵(Heatmap)

- 막대나 점 대신 색으로 데이터 값을 표현하여 전반적인 추세 확인이 쉬운 방법

56

학습정리

2 분포 데이터의 시각화

히스토그램(Histogram)

- 가로에 범주 구간, 세로에 구간에 포함된 데이터의 수를 표현

밀도 도표(Density plot)

- 데이터의 분포를 연속적인 곡선으로 표현, 곡선 아래의 면적은 1

박스 도표(Box plot)

- 데이터를 사분위로 나누고 박스안에 50%의 데이터를 표현

57

참고문헌

「데이터 시각화 교과서」, Claus O. Wilke, 책만, 2020.

「Fundamentals of Data Visualization」, Claus O. Wilke, O'Reilly Media, 2019.

※ 서체 출처 | 넥슨Lv2고딕-(넥슨코리아)www.levelup.nexon.com / 나눔바른고딕(네이버)

58

저작권 안내

계동캠퍼스(03051)서울특별시 종로구 북촌로 106 **안암캠퍼스**(02841)서울특별시 성북구 안암로 145 고려대학교 안암캠퍼스내

www.cuk.edu T.02-6361-2000 F.02-6361-1800