데이터 시각화

11. 그래프 디자인의 기본원칙(1)

최대영 교수

고려사이버대학교
THE CYBER UNIVERSITY OF KOREA

Data Visualization

데이터 시각화

11주차

그래프 디자인의 기본 원칙(1)

최대영 교수

1

학습리뷰

1 변수 연관성의 시각화

산점도(Scatter plot)

- 데이터 하나를 x, y좌표 위에 점 하나로 표현

버블 차트(Bubble chart)

- 버블(점)의 크기를 이용해서 속성을 표시

산점도 매트릭스(Scatterplot matrix)

- 모든 변수 간의 관계를 일대일로 산점도로 표현

2

학습리뷰

1 변수 연관성의 시각화

Correlogram

- 상관계수를 색을 이용하여 시각화

경사 차트(Slopegraph)

- 두 변수의 값을 선으로 연결

3

학습리뷰

2 시계열 데이터의 시각화

선 그래프(Line graph)

- 이웃한 두 점(전후 시간의 데이터)을 연결

이동 평균(moving average)을 이용한 평활화(smoothing)

- 한 시점씩 이동하며 특정 구간의 데이터 평균값을 계산하여 표현

국소 추정 산점도 평활(LOESS, locally estimated scatter plot smoothing)

- 저차원의 다항식에 데이터의 부분집합을 적합(fitting)

4

학습리뷰

2 시계열 데이터의 시각화

스플라인(Spline) 모형

- 몇 개의 제어점을 기준으로 구간별로 적합

정의된 함수 형식을 이용한 추세 시각화

- 데이터에 잘 맞는 형식의 함수를 정하여 데이터를 적합

추세 제거(Detrending)

- 장기적으로 지배하는 추세가 있을 때 추세를 제거하여 중요한 편차를 특정하여 강조

5

학습목표

- 잉크 양 비례의 원칙에 대해 설명할 수 있다.
- 오버플로팅과 효과적인 색 사용 방법에 대해 설명할 수 있다.
- 오버플로팅 관련 matplotlib 라이브러리를 이해하고 활용할 수 있다.

6

학습내용

1. 잉크 양 비례의 원칙
2. 오버플로팅과 색
3. 실습

7

01
chapter

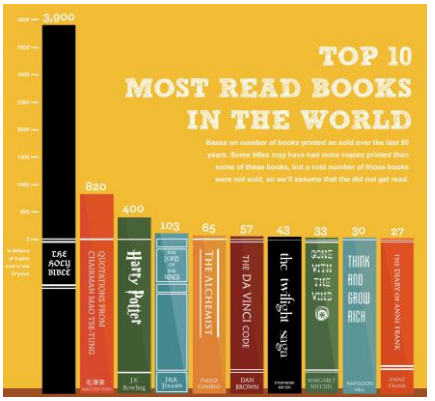
잉크 양 비례의 원칙

8

1. 축이 있는 도표

잉크 양 비례의 원칙(Principle of proportional ink)

시각화에서 음영 영역의 크기는 해당 영역이 나타내는 데이터 값에 비례해야 함
→ 값을 나타내는 데 사용되는 잉크의 양은 값 자체에 비례해야 함



[출처] Calling Bullshit

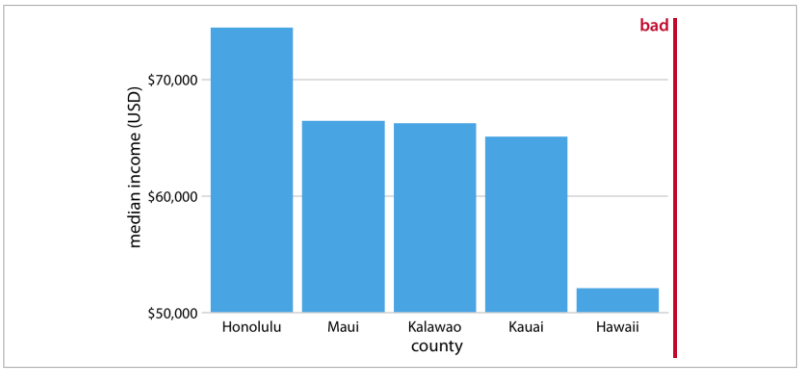
9

1. 축이 있는 도표

선형 축이 있는 막대 도표

모든 막대가 중위 소득 5만 달러에서 시작 → 하와이 카운티와 다른 4개 카운티의 값의 차이가 실제보다 커 보임

하와이 5개 카운티의 중위소득



[출처] Fundamentals of Data Visualization

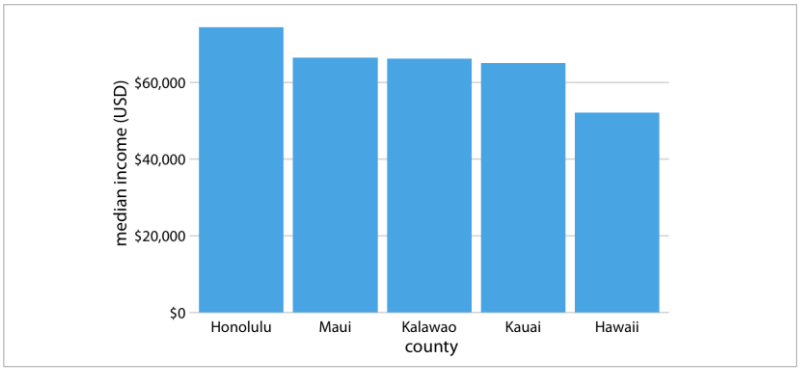
10

1. 축이 있는 도표

선형 축이 있는 막대 도표

모든 막대가 중위 소득 0 달러에서 시작 → 하와이 카운티와 다른 4개 카운티의 값이 차이가 크지 않음을 나타냄

하와이 5개 카운티의 중위소득



[출처] Fundamentals of Data Visualization

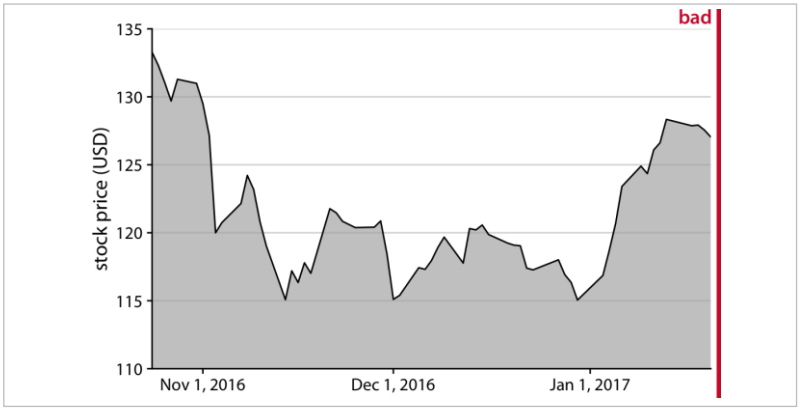
11

1. 축이 있는 도표

선형 축이 있는 선 그래프

주가를 110 달러 기준으로 표현 → 주가가 지나치게 많이 떨어진 것처럼 표현

페이스북 주가



[출처] Fundamentals of Data Visualization

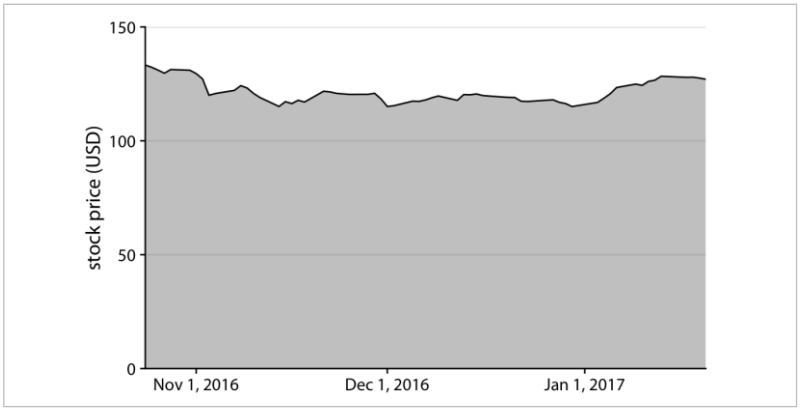
12

1. 축이 있는 도표

선형 축이 있는 선 그래프

주가를 0달러 기준으로 표현 → 주가의 변동폭이 주가에 비해 크지 않음

페이스북 주가



[출처] Fundamentals of Data Visualization

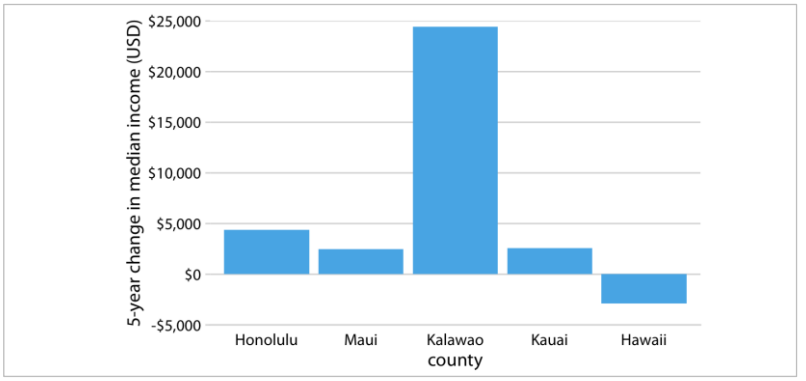
13

1. 축이 있는 도표

막대 도표로 차이를 나타내는 방법

0을 기준으로 중위소득의 변화량을 표현 → 0을 기준으로 하더라도 차이를 분명하게 표현할 수 있음

5년간 중위소득의 변화



[출처] Fundamentals of Data Visualization

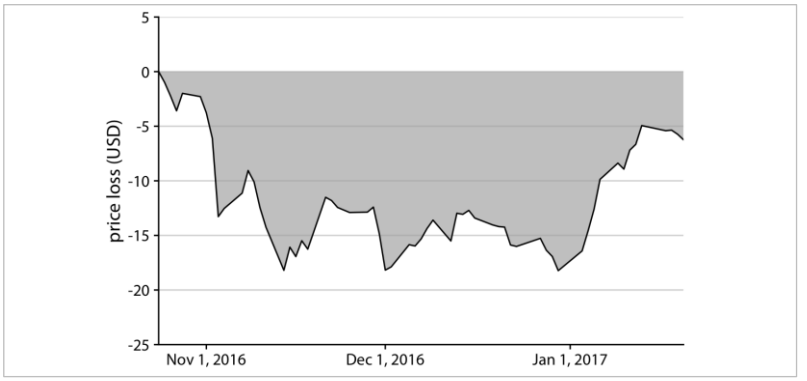
14

1. 축이 있는 도표

선 그래프로 차이를 나타내는 방법

주가 고점으로부터의 차이를 표현 → 0을 기준으로 하더라도 절대적인 주가의 하락을 정확하게 표현할 수 있음

페이스북 주가



[출처] Fundamentals of Data Visualization

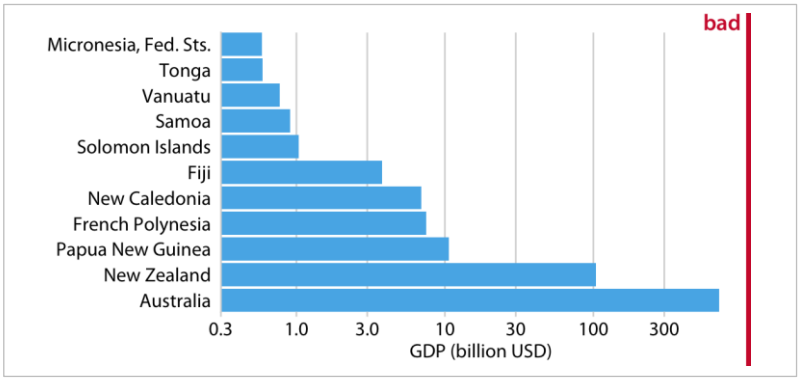
15

1. 축이 있는 도표

로그 축이 있는 막대 도표

막대가 0.3에서 부터 시작 → 막대 길이가 데이터 값을 정확하게 보여주지 않고 0을 왼쪽으로 무한히 멀리에 표현해야 함

오세아니아 대륙 국가의 2007년 GDP



[출처] Fundamentals of Data Visualization

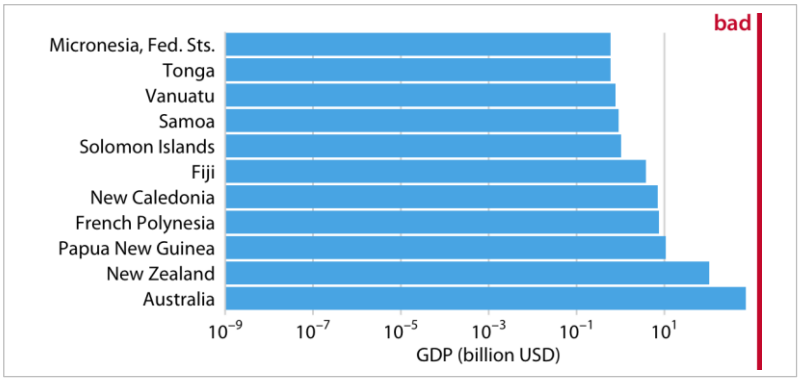
16

1. 축이 있는 도표

로그 축이 있는 막대 도표

막대의 원점을 임의로 왼쪽으로 멀리 보냄 → 여전히 임의의 값에서 시작하는 문제가 발생

오세아니아 대륙 국가의 2007년 GDP



[출처] Fundamentals of Data Visualization

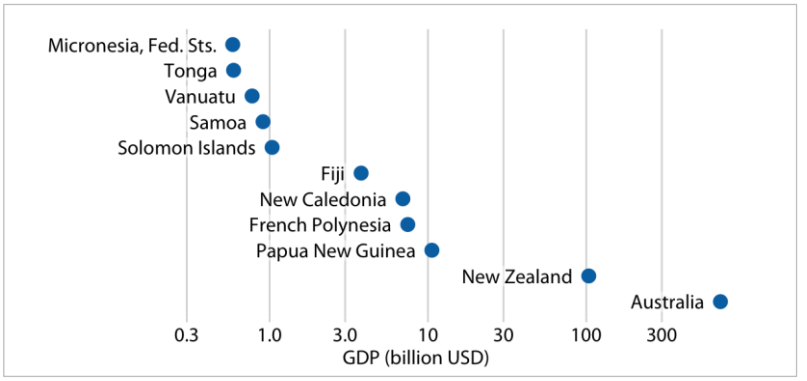
17

1. 축이 있는 도표

점으로 막대를 대체

0을 기준으로 해야 하는 막대 도표의 길이 문제를 해결 → 길이의 개념이 없기 때문에 상대적인 크기를 적절히 표현

오세아니아 대륙 국가의 2007년 GDP



[출처] Fundamentals of Data Visualization

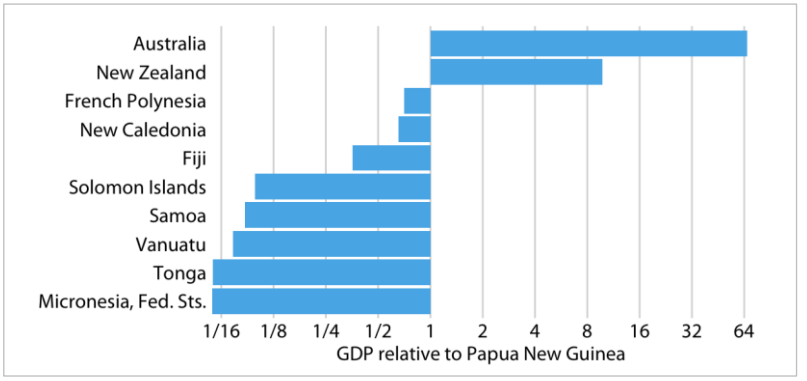
18

1. 축이 있는 도표

비율을 나타내는 로그 스케일 막대 도표

비율을 나타내므로 기준점을 1로 표현

파푸아뉴기니 기준의 오세아니아 대륙 국가의 2007년 GDP



[출처] Fundamentals of Data Visualization

19

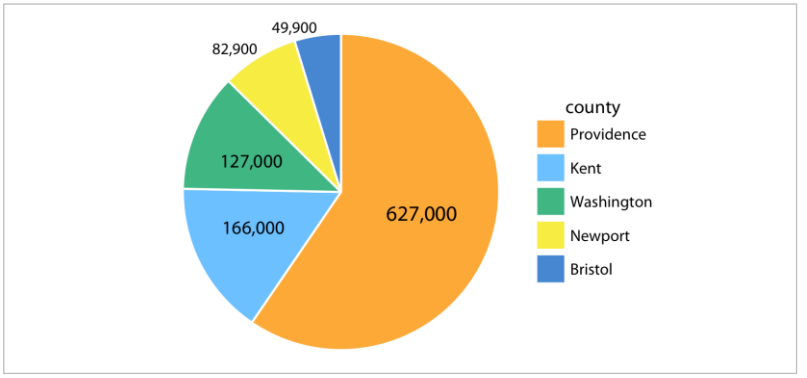
2. 면적을 나타내는 도표

파이 차트(Pie chart)

파이 조각의 면적은 각도에 비례, 각도는 데이터 값에 비례

→ 잉크 양 비례의 원칙에 부합

로드아일랜드주 카운티의 인구



[출처] Fundamentals of Data Visualization

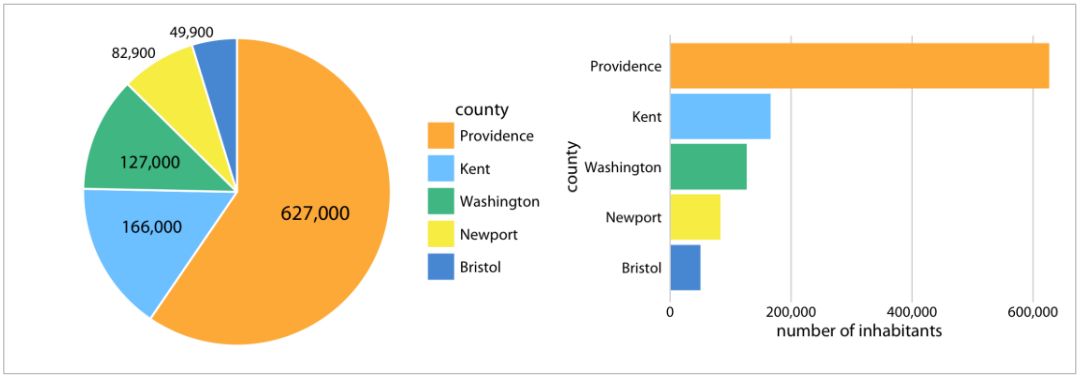
20

2. 면적을 나타내는 도표

파이 차트와 막대 도표의 차이

막대 도표에서 인구 차이가 더 크다고 인식 → 인간의 인지가 면적보다는 길이를 먼저 판단하는 것과 관련

로드아일랜드주 카운티의 인구



[출처] Fundamentals of Data Visualization

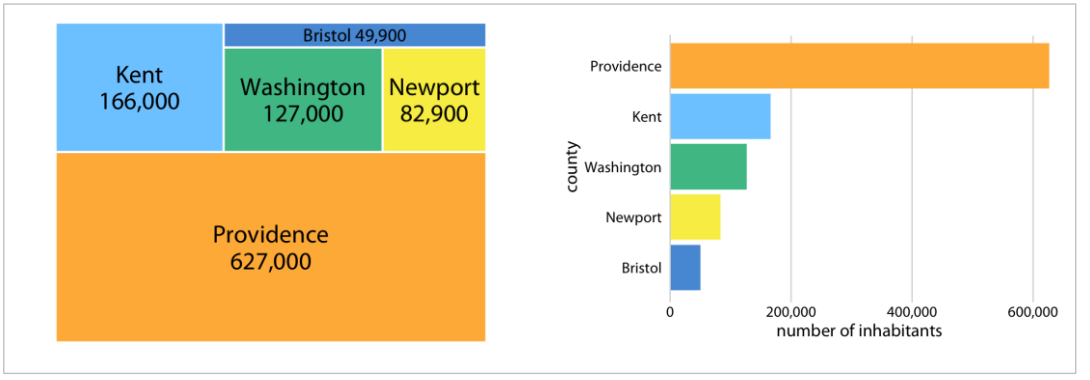
21

2. 면적을 나타내는 도표

트리맵과 막대 도표의 차이

막대 도표에서 인구 차이가 더 크다고 인식 → 인간의 인지가 면적보다는 길이를 먼저 판단하는 것과 관련

로드아일랜드주 카운티의 인구



[출처] Fundamentals of Data Visualization

22

02
chapter

오버플로팅과 색

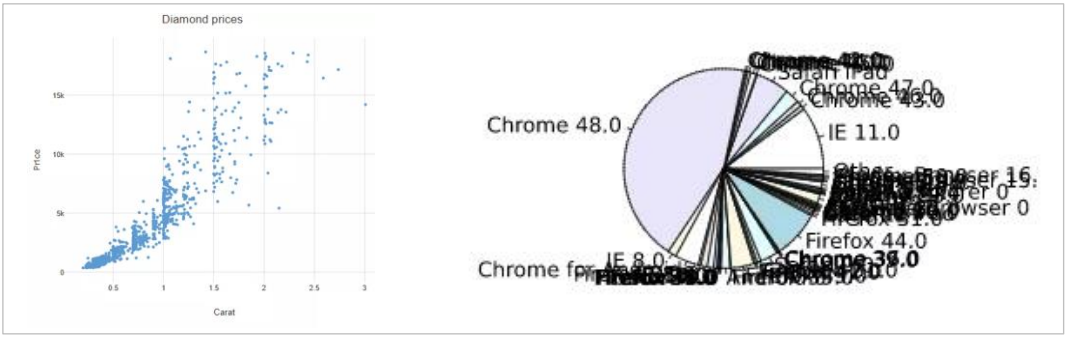
23

1. 오버플로팅

오버플로팅(Overplotting)

여러 개의 점이 같은 좌표에 겹쳐서 표현되어 있는 현상

- 데이터 포인트가 너무 많은 경우에 주로 발생
- 데이터 값이 부정확하거나 반올림 등으로 인해 같은 값이 많은 경우에도 발생



[출처] Fundamentals of Data Visualization

24

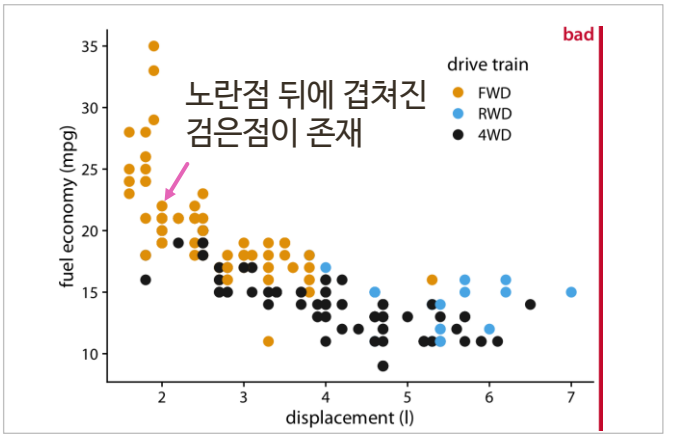
1. 오버플로팅

데이터 값을 반올림하여 겹쳐진 데이터 예시

연비와 엔진 배기량을 반올림

→ 겹쳐진 점으로 인해 데이터의 특성 파악이 어려움
(차종별 데이터 포인트 수, drive train이 다른 차종이 숨겨짐)

인기차종 234종의 도시 주행 연비와 엔진 배기량



[출처] Fundamentals of Data Visualization

25

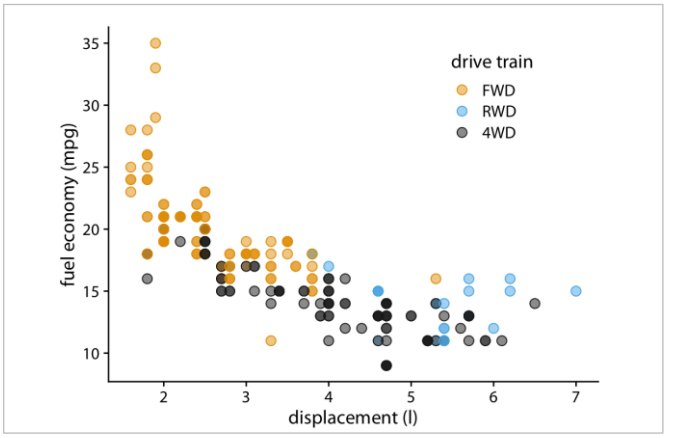
1. 오버플로팅

반투명 색의 사용

점마다 반투명 색을 이용하여 데이터 값의 차이를 표현

→ 몇 개의 점이 겹쳐져 있는지 파악하기가 어려움, 음영의 차이에 대한 설명이 없음

인기차종 234종의 도시 주행 연비와 엔진 배기량



[출처] Fundamentals of Data Visualization

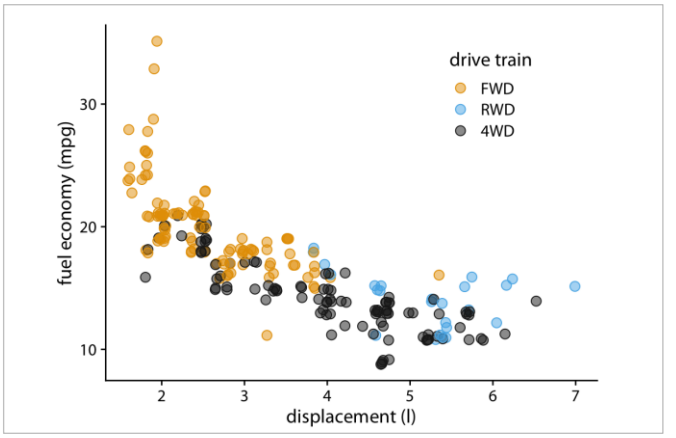
26

1. 오버플로팅

지터링(Jittering)의 사용

점마다 약간씩 위치를 옮겨 표현
→ 부분적으로 겹쳐진 점을 통해 데이터의 양을 파악

인기차종 234종의 도시 주행 연비와 엔진 배기량



[출처] Fundamentals of Data Visualization

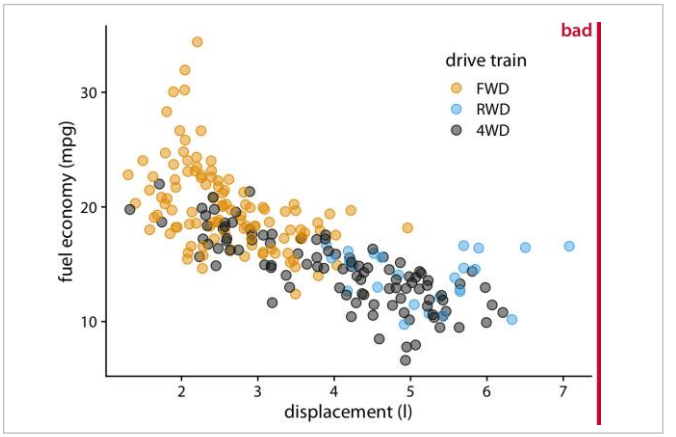
27

1. 오버플로팅

지터링(Jittering)의 사용

점마다 약간씩 위치를 옮겨 표현
→ 위치를 너무 많이 옮기면 데이터가 왜곡됨

인기차종 234종의 도시 주행 연비와 엔진 배기량



[출처] Fundamentals of Data Visualization

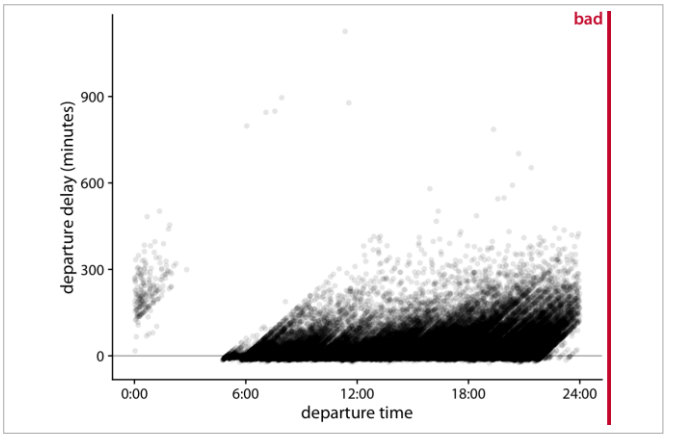
28

1. 오버플로팅

투명도 방법의 한계점

점이 너무 많은 곳은 얼룩처럼 보이고 적은 곳은 흐릿하게 보임

항공편 10만 건의 출발지연 시간



[출처] Fundamentals of Data Visualization

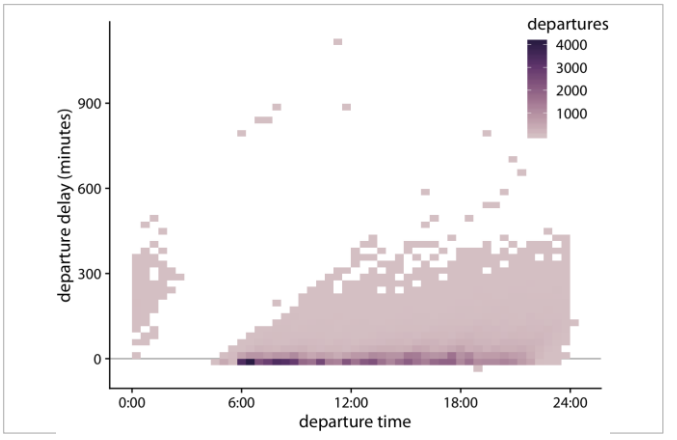
29

1. 오버플로팅

2차원 히스토그램

2차원 평면을 사각형으로 나눠 데이터의 수에 따라 색으로 표현
→ 데이터의 특성을 요약해서 보여주는데 유리

항공편 10만 건의 출발지연 시간



[출처] Fundamentals of Data Visualization

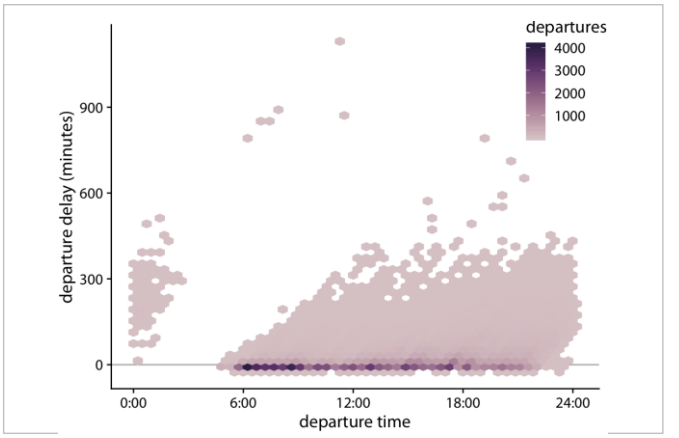
30

1. 오버플로팅

2차원 히스토그램

2차원 평면을 육각형으로 나눠
데이터의 수에 따라 색으로 표현

항공편 10만 건의 출발지연 시간



[출처] Fundamentals of Data Visualization

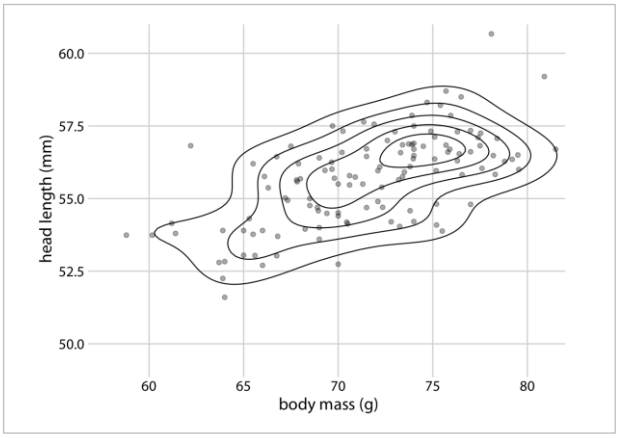
31

1. 오버플로팅

등고선 그래프(Contour lines)

도표 전체의 점 밀도를 추정하고
그에 따라 등고선을 그리는 방법
→ 투명도를 이용하는 방법을
보완하여 분포를 더 잘 표현

큰어치 123마리 데이터



[출처] Fundamentals of Data Visualization

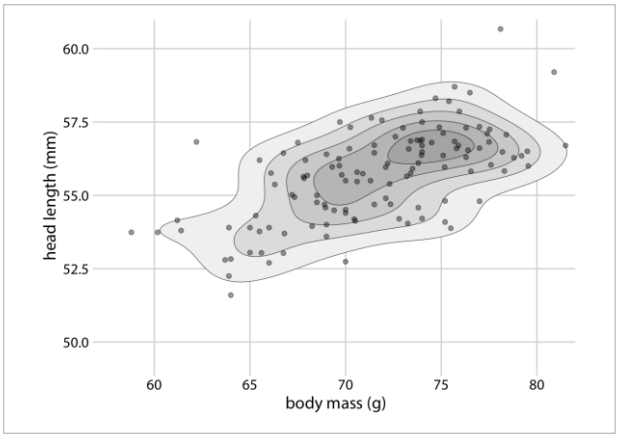
32

1. 오버플로팅

등고선 그래프(Contour lines)

- 도표 전체의 점 밀도를 추정하고 그에 따라 등고선을 그리는 방법
  → 밀도에 따라 음영을 사용하면 분포를 더 두드러지게 나타낼 수 있음

큰어치 123마리 데이터



[출처] Fundamentals of Data Visualization

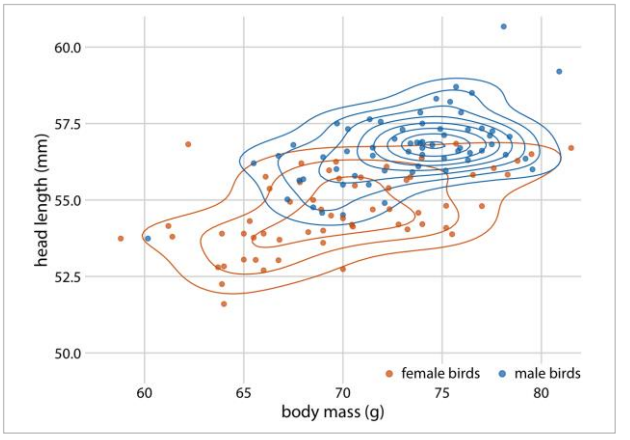
33

1. 오버플로팅

등고선 그래프(Contour lines)

- 등고선의 색을 다르게 하여 여러 범주를 표현 → 범주의 수가 2~3개로 제한되고 분포가 겹쳐져 있지 않은 경우에 효과적

큰어치 123마리 데이터



[출처] Fundamentals of Data Visualization

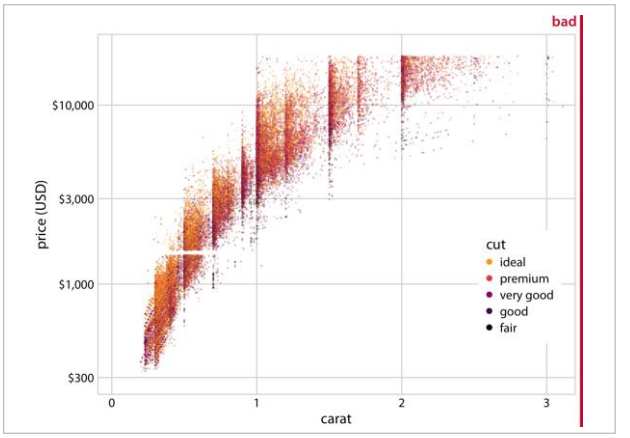
34

1. 오버플로팅

등고선 그래프가 적절하지 않은 경우

범주의 수가 많고 데이터가 많이 겹쳐져 있는 경우 → 전반적인 가격과 캐럿에 대해서는 알 수 있으나 컷(cut) 정보를 알기 어려움

다이아몬드 53940알의 정보



[출처] Fundamentals of Data Visualization

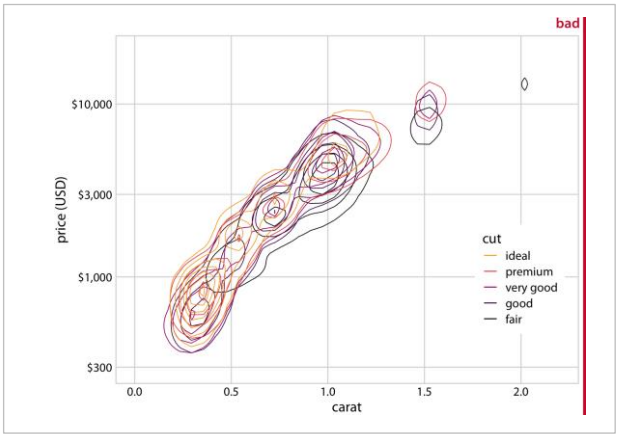
35

1. 오버플로팅

등고선 그래프가 적절하지 않은 경우

범주의 수가 많고 데이터가 많이 겹쳐져 있는 경우 → 등고선으로 표현하여도 컷별 특성을 파악하기 어려움

다이아몬드 53940알의 정보



[출처] Fundamentals of Data Visualization

36

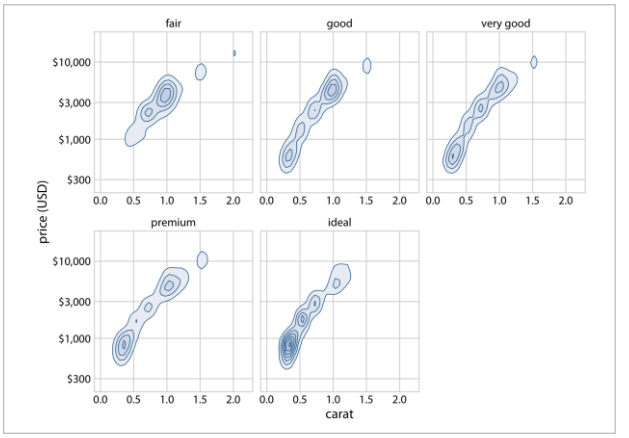
1. 오버플로팅

범주별 등고선 그래프

범주별로 등고선을 그리고 격자를 넣어 범주 간의 비교가 가능하도록 표현

→ 여러 범주가 겹쳐 있을 때 유용

다이아몬드 53940알의 정보



[출처] Fundamentals of Data Visualization

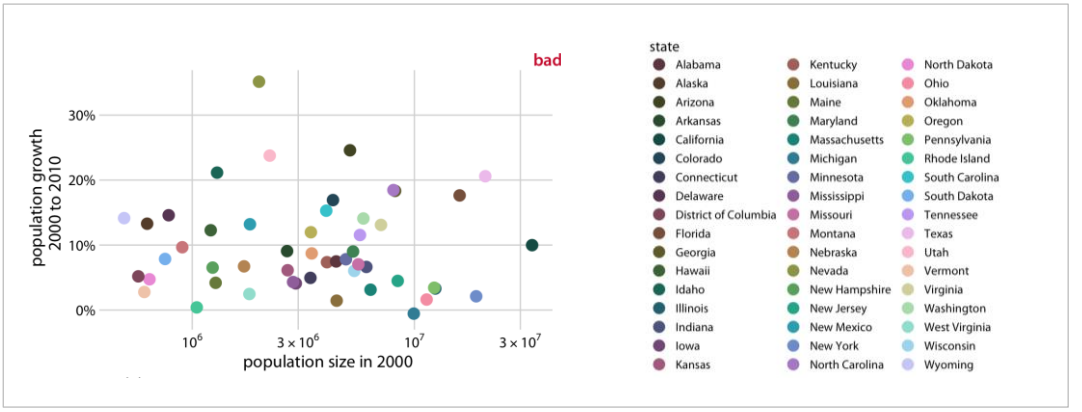
37

2. 효과적인 색 사용

색상 스케일이 효과적인 범주의 수

3~5개의 범주를 색으로 나타낼 때 효과적, 8~10개가 넘어가면 효과가 떨어짐

미국의 주별 인구 성장율과 규모



[출처] Fundamentals of Data Visualization

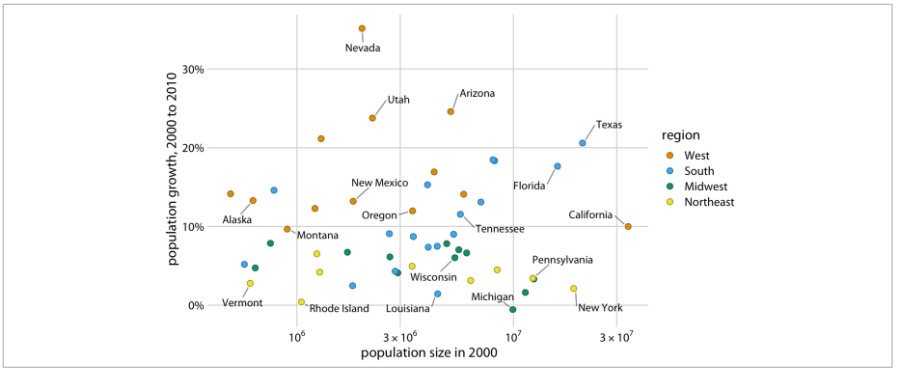
38

2. 효과적인 색 사용

색상 스케일이 효과적인 범주의 수

색상 스케일의 수를 한정하고 대표성을 띄는 중요한 데이터에만 레이블을 표시
→ 도표가 간결하고 메시지 전달력이 좋아짐

미국의 주별 인구 성장율과 규모



[출처] Fundamentals of Data Visualization

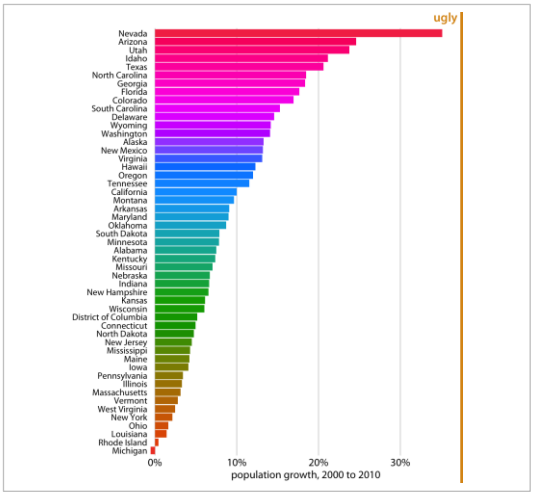
39

2. 효과적인 색 사용

목적이 명확하지 않은 색의 사용

색이 의미하는 바가 없고 채도가 높음
→ 메시지 전달에 도움이 되지 않으며 채도가 높으면 도표를 꼼꼼하게 살피기 어려움(넓은 면적에는 채도가 낮은 색 위주로 사용)

미국의 주별 인구 성장율과 규모



[출처] Fundamentals of Data Visualization

40

2. 효과적인 색 사용

색각 이상자(color-vision deficiency)에 대한 고려

시각화의 색을 고를 때는 색각에 이상이 있는 독자를 고려해야 함

- 녹색약과 적색약: 녹색과 빨간색을 인지하는 능력이 부족
- 청색약: 파란색을 인지하는 능력이 부족

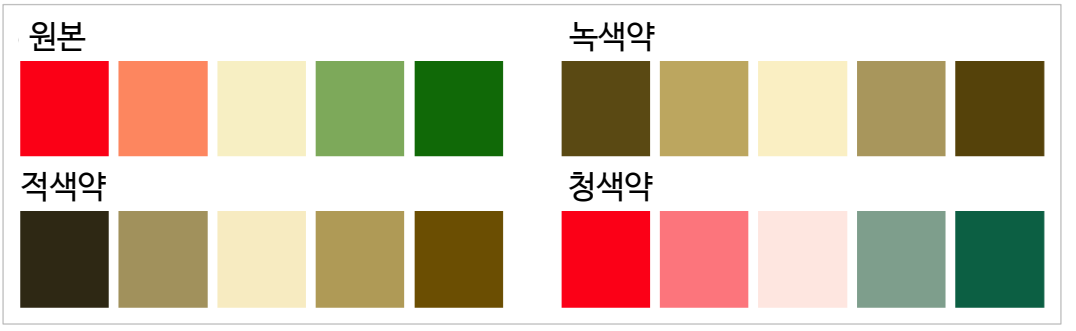
[출처] Fundamentals of Data Visualization

41

2. 효과적인 색 사용

범주형 인코딩과 발산형 인코딩

녹색약과 적색약이 구분하기 어려운 인코딩



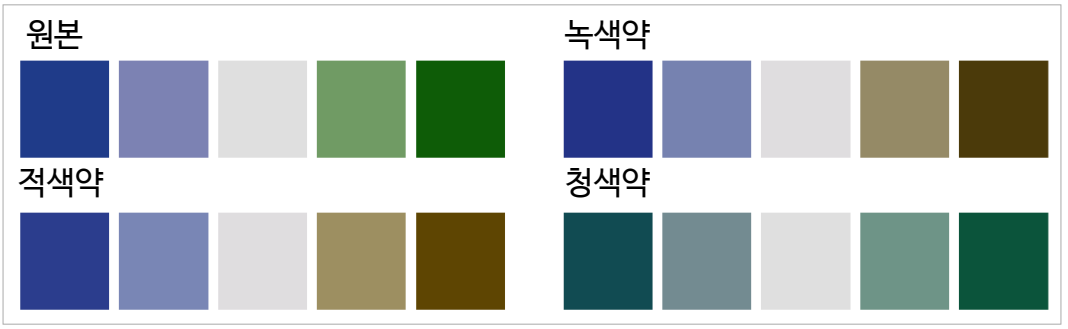
[출처] Fundamentals of Data Visualization

42

2. 효과적인 색 사용

범주형 인코딩과 발산형 인코딩

청색약이 구분하기 어려운 인코딩



[출처] Fundamentals of Data Visualization

43

2. 효과적인 색 사용

색각 이상자를 위한 색상 스케일(Color Brewer의 PiYG)

빨간색과 녹색에 다른 색을 섞어 색각 이상자도 색을 구별할 수 있도록 정함

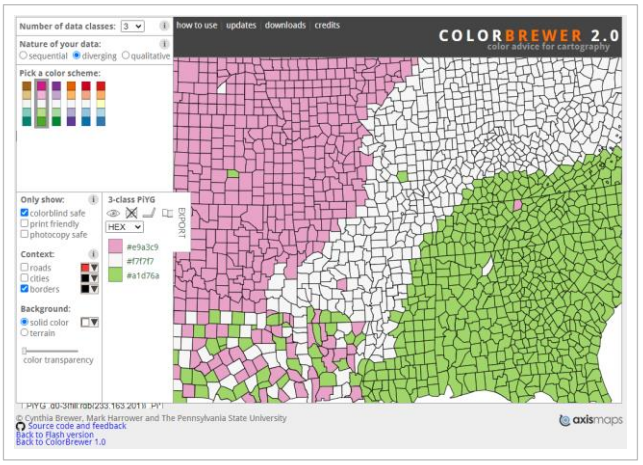
[출처] Fundamentals of Data Visualization

44

2. 효과적인 색 사용

Color Brewer

시각화에 적절한 색상 스케일을 선정하는 도구



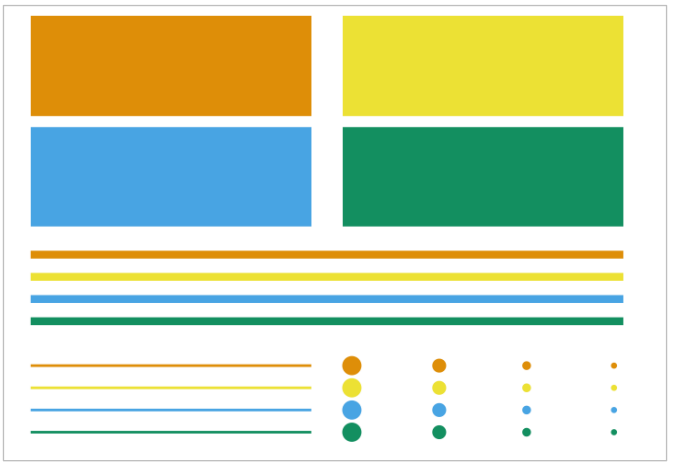
[출처] COLOR BREWER(https://colorbrewer2.org/)

45

2. 효과적인 색 사용

색상 스케일과 크기 속성의 관계

색상 스케일을 사용하는 그래픽 속성의 크기가 작아지면 색을 구별하기 어려워 짐



[출처] Fundamentals of Data Visualization

46

03

chapter

실습

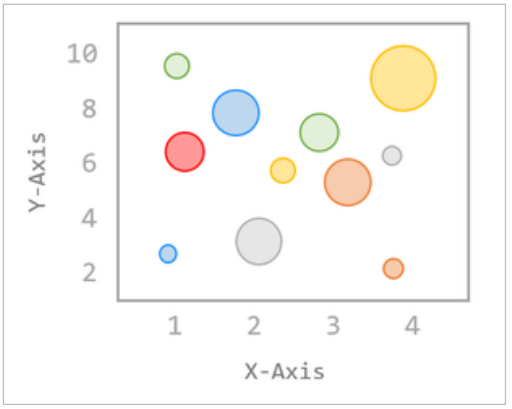
47

1. 오버플로팅

산점도의 속성과 등고선 그래프

산점도의 속성

- scatter(x, y, s=area, c=color, alpha=transparency)



[출처] Matplotlib Tutorial - 파이썬으로 데이터 시각화하기

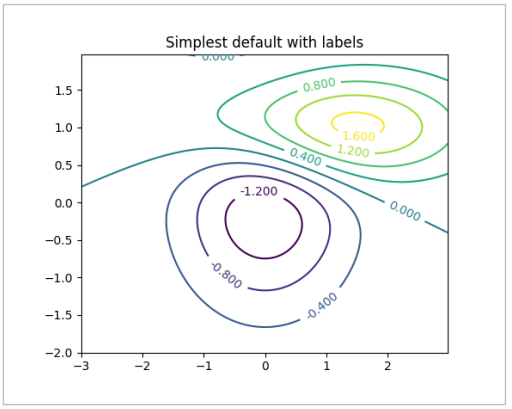
48

1. 오버플로팅

산점도의 속성과 등고선 그래프

등고선 그래프

- contour(X, Y, Z)



[출처] Matplotlib Tutorial – 파이썬으로 데이터 시각화하기

49

학습정리

1 잉크 양 비례의 원칙

잉크 양 비례의 원칙 (Principle of proportional ink)

- 음영 영역의 크기는 해당 영역이 나타내는 데이터 값에 비례해야 함

축이 있는 도표

- 잉크 양 비례의 원칙에 부합하기 위해 막대나 선은 축의 0부터 표현해야 함

면적을 나타내는 도표

- 파이 조각의 면적은 각도에 비례, 각도는 데이터 값에 비례
- 인간의 인지가 면적보다는 길이를 먼저 판단하기 때문에 막대 도표가 파이 차트와 트리맵 보다 데이터 값의 차이를 더 두드러지게 나타냄

50

학습정리

2 오버플로팅과 색

오버플로팅(Overplotting)

- 여러 개의 점이 같은 좌표에 겹쳐서 표현되어 있는 현상

반투명 색 사용

- 점마다 반투명 색을 이용하여 데이터 값의 차이를 표현

지터링(Jittering)

- 데이터를 왜곡하지 않는 한에서 점마다 약간씩 위치를 옮겨 표현

51

학습정리

2 오버플로팅과 색

2차원 히스토그램

- 2차원 평면을 사각형 등으로 나눠 데이터의 수에 따라 색으로 표현

등고선 그래프(Contour lines)

- 도표 전체의 점 밀도를 추정하고 그에 따라 등고선을 그리는 방법

52

학습정리

2 오버플로팅과 색

색상 스케일의 수와 의미

- 3~5개의 범주를 색으로 나타낼 때 효과적
- 색이 의미하는 바를 명확히 하고 채도를 조절하여 피로를 낮춰야 함

색각 이상자(color-vision deficiency)에 대한 고려

- 색각 이상자도 구별할 수 있는 색상 스케일로 인코딩
- Color Brewer 등 도구를 이용해 시각화에 쓰일 색상을 선정

53

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)

54

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