데이터 시각화

10. 변수 연관성과 시계열 데이터의 시각화

최대영 교수

고려사이버대학교
THE CYBER UNIVERSITY OF KOREA

Data Visualization

데이터 시각화

10주차

변수 연관성과 시계열 데이터의 시각화

최대영 교수

1

학습리뷰

1 단일 범주 비율 데이터의 시각화

파이 차트(Pie chart)

- 전체 데이터에 해당하는 원을 부분 데이터가 차지하는 비중에 비례하는 크기의 조각으로 분할

누적 막대 도표(Stacked bar chart)

- 파이 차트의 원 대신 직사각형의 조각으로 표현

병렬 막대

- 누적 막대 도표의 직사각형 조각을 나열하여 비율을 표현

2

학습리뷰

1 단일 범주 비율 데이터의 시각화

누적 밀도

- 연속형 데이터에 대해 밀도를 누적하여 표현

전체 대비 부분 비율의 표현

- 전체 분포를 배경으로 두고 각 범주를 표현

3

학습리뷰

2 내포 비율의 시각화

모자이크 도표(Mosaic plot)

- 가로, 세로의 길이로 범주 내 비율을 표현

트리맵(Treemap)

- 사각형 안에 작은 사각형을 반복적으로 계층을 가지며 쪼개어 표현

4

학습리뷰

2 내포 비율의 시각화

내포 파이 차트(Nested pie)

- 외부원과 내부원으로 구분하여 2개의 범주를 각각 표현

평행 집합 도표(Parallel sets plot)

- 2개 이상의 범주가 있는 데이터의 범주별 분류와 분류간의 관계성을 나타냄

5

학습목표

- 변수 연관성의 시각화 방법에 대해 설명할 수 있다.
- 시계열 데이터의 특징과 시각화 방법에 대해 설명할 수 있다.
- 변수 연관성과 시계열 데이터의 시각화 관련 matplotlib 라이브러리를 이해하고 활용할 수 있다.

6

학습내용

1 변수 연관성의 시각화

2 시계열 데이터의 시각화

3 실습

7

01
chapter

변수 연관성의 시각화

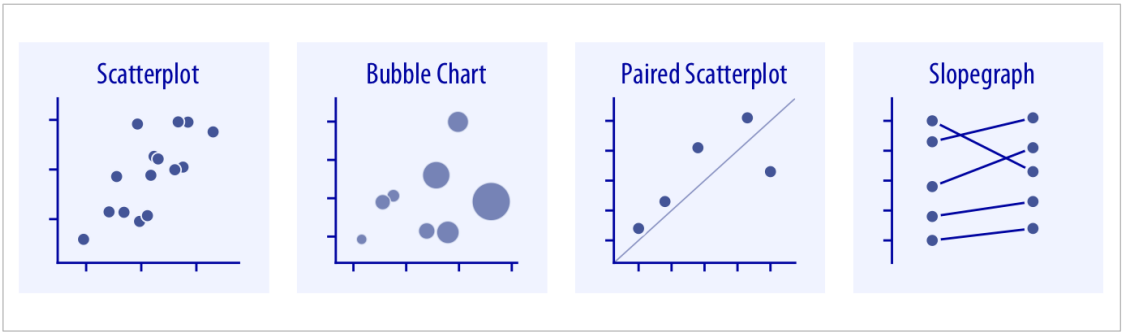
8

1. 산점도와 correlogram

여러 정량 변수의 관계 시각화

데이터셋에 여러 정량 변수가 있을 경우 이 변수 간의 상관관계를 표시

예 동물의 키와 몸무게의 연관성, 학생의 성적과 공부시간의 연관성



[출처] Fundamentals of Data Visualization

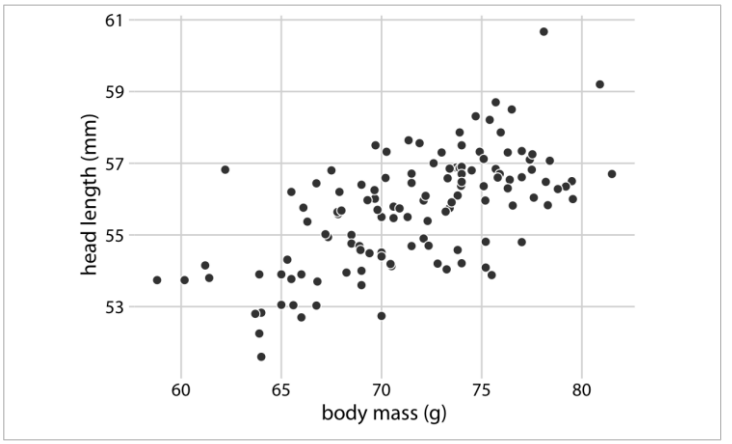
9

1. 산점도와 correlogram

산점도(Scatter plot)

데이터 하나를 x, y좌표 위에 점 하나로 표현 → 점의 분포로 상관관계를 파악

큰어치 123마리 데이터



[출처] Fundamentals of Data Visualization

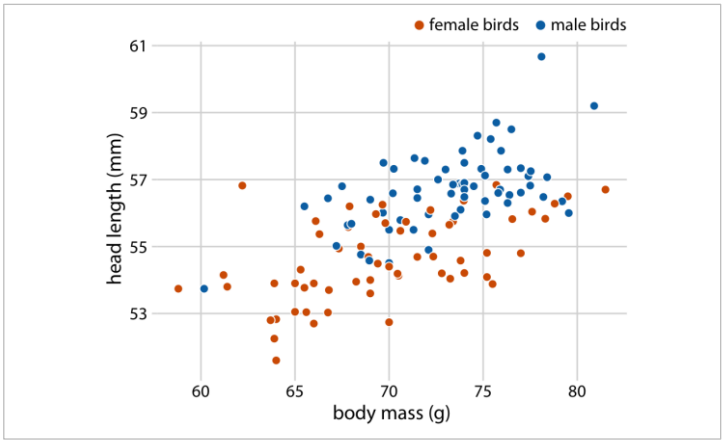
10

1. 산점도와 correlogram

산점도(Scatter plot)

점의 색으로 범주를 구분하여 표현 → 한 도표 내에서 두 범주를 비교 가능

큰어치 123마리 데이터



[출처] Fundamentals of Data Visualization

11

1. 산점도와 correlogram

버블 차트(Bubble chart)

버블(점)의 크기를 이용해서 속성을 표시 → 한 도표 내에서 세 범주를 표현 가능, 두 속성(위치, 크기)을 동시에 사용하여 혼란이 있고 버블의 크기에 제한이 있음

큰어치 123마리 데이터

[출처] Fundamentals of Data Visualization

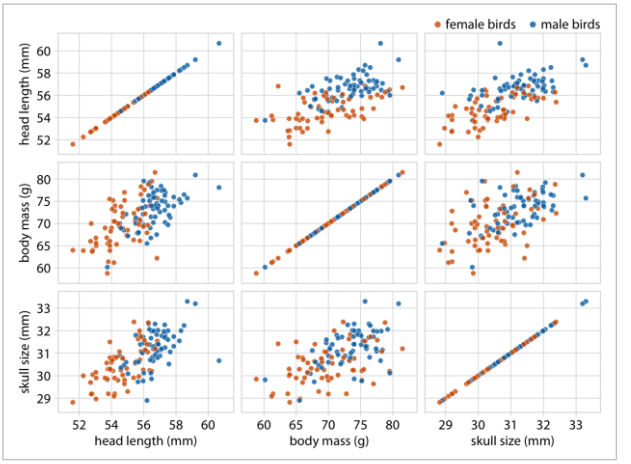
12

1. 산점도와 correlogram

산점도 매트릭스(Scatterplot matrix)

모든 변수 간의 관계를 일대일로 산점도로 표현 → 변수 간의 기본적인 관계를 파악을 할 수 있어 탐색적 데이터 분석을 위해 많이 사용

큰어치 123마리 데이터



[출처] Fundamentals of Data Visualization

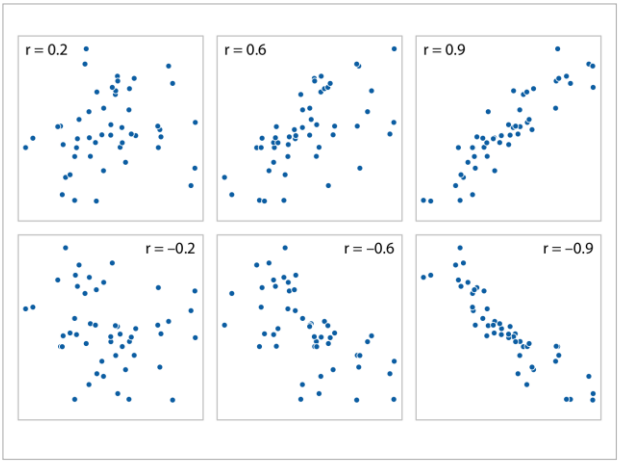
13

1. 산점도와 correlogram

상관계수(Correlation coefficient)

두 변수 사이의 상관관계의 정도를 −1과 1 사이의 수치로 표현
→ 선형관계일수록 1 또는 −1에 가깝고 선형관계가 없으면 0에 가까움

시뮬레이션 데이터



[출처] Fundamentals of Data Visualization

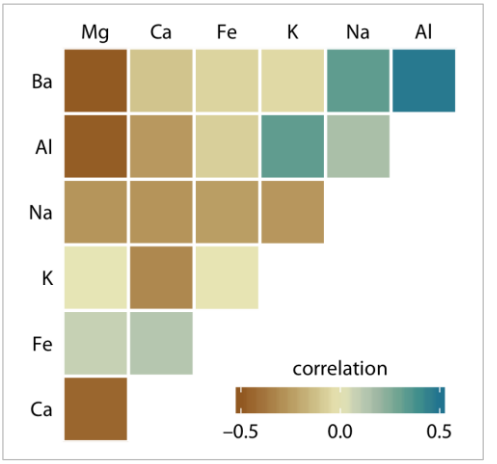
14

1. 산점도와 correlogram

Correlogram

상관계수를 색을 이용하여 시각화
→ 변수가 많을 경우 산점도 매트릭스보다 간명하게 표현 가능, 색의 표현방법 선택이 중요

유리파편의 무기물 함량



[출처] Fundamentals of Data Visualization

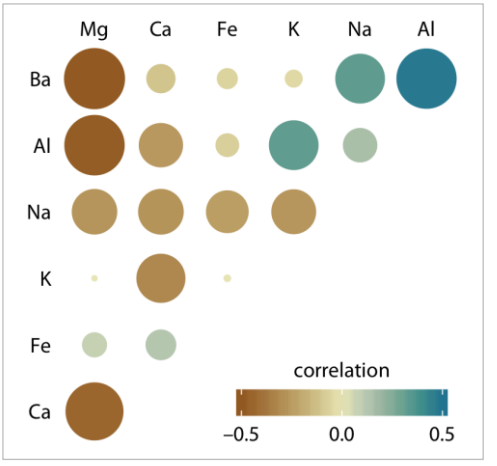
15

1. 산점도와 correlogram

Correlogram

상관계수의 절대값을 원의 크기로 표현 → 상관계수만 가지고는 데이터의 특성을 알기 어려움 (앤스컴 콰르텟의 상관계수)

유리파편의 무기물 함량



[출처] Fundamentals of Data Visualization

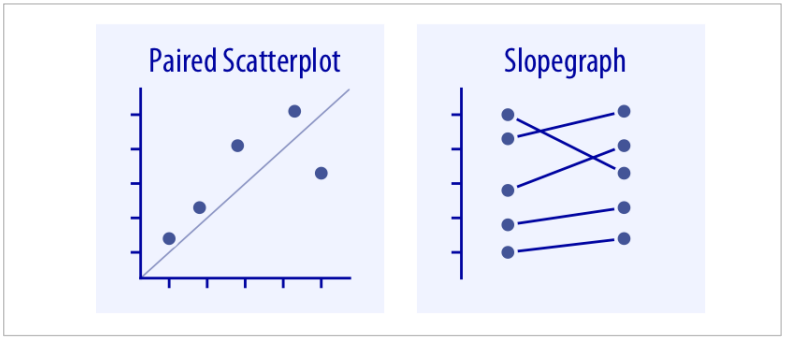
16

2. 쌍 데이터와 경사 차트

쌍 데이터(Paired data)

조금 다른 조건에서 같은 변수에 대한 측정치가 둘 이상 있는 데이터

- 예 피험자 1명에서 비슷한 측정값이 2개 있는 경우(왼팔과 오른팔의 길이)
- 예 피험자 1명의 수치를 두 시점에 반복 측정한 경우(몸무게 2회 측정)



[출처] Fundamentals of Data Visualization

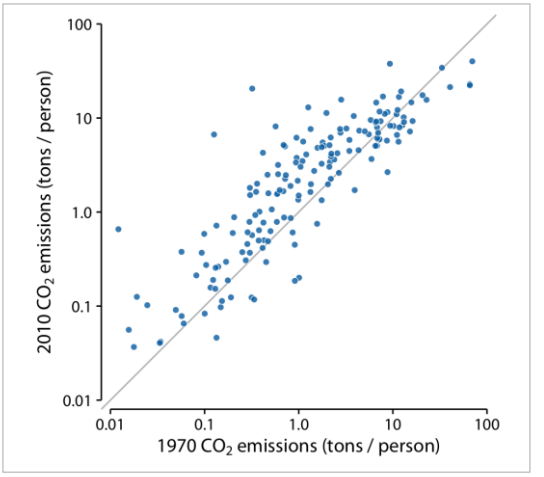
17

2. 쌍 데이터와 경사 차트

사선이 있는 산점도

산점도를 그리고 분포의 참조를 위해 사선을 표현 → 데이터의 전체적인 편차와 관계성 파악에 유용

166개국의 일인당 이산화탄소 배출량



[출처] Fundamentals of Data Visualization

18

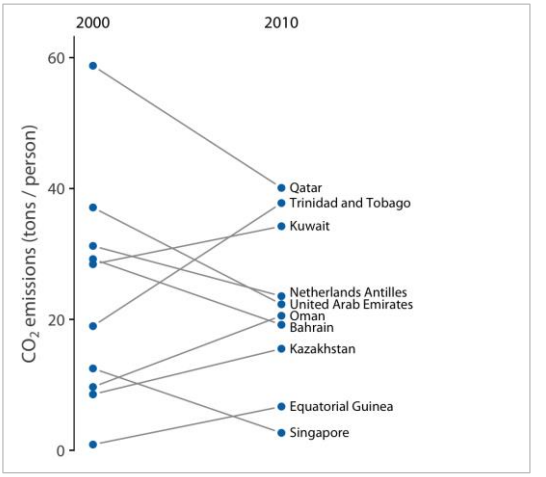
2. 쌍 데이터와 경사 차트

경사 차트(Slopegraph)

두 변수의 값을 선으로 연결

→ 데이터가 적고 개별 데이터의 유사성을 자세히 살피는데 유용

10개국의 일인당 이산화탄소 배출량



[출처] Fundamentals of Data Visualization

19

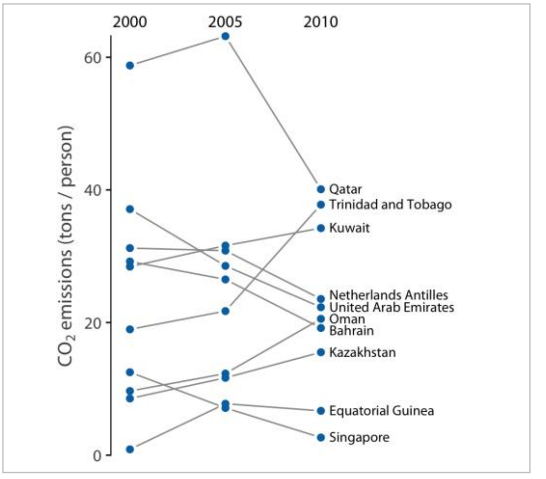
2. 쌍 데이터와 경사 차트

경사 차트(Slopegraph)

측정값이 여러 개인 경우 오른쪽으로 나열

→ 2개 이상의 측정값을 동시에 비교하여 추세를 나타냄

10개국의 일인당 이산화탄소 배출량



[출처] Fundamentals of Data Visualization

20

02

chapter

시계열 데이터의 시각화

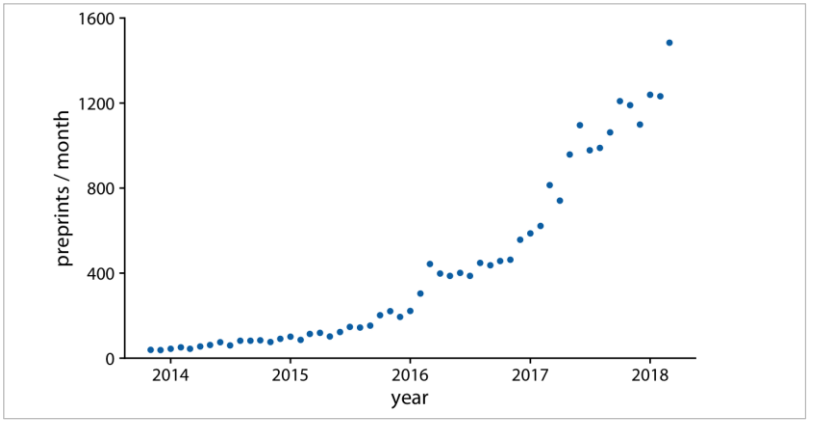
21

1. 선 그래프와 용량-반응 곡선

산점도

월별 bioRxiv에 등록된 논문 건수를 점으로 표현 → 시간 개념이 두드러지지 않음

출간 전 bioRxiv에 등록된 논문 수



[출처] Fundamentals of Data Visualization

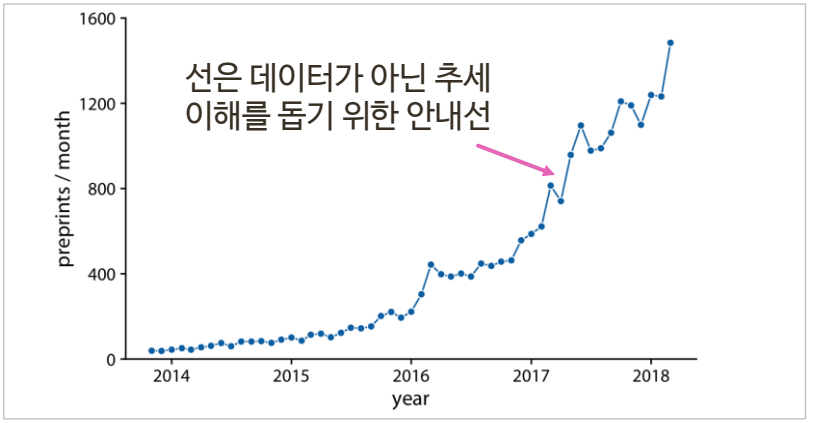
22

1. 선 그래프와 용량-반응 곡선

선 그래프(Line graph)

이웃한 두 점(전후 시간의 데이터)을 연결 → 시간 개념이 두드러지게 나타남

출간 전 bioRxiv에 등록된 논문 수



[출처] Fundamentals of Data Visualization

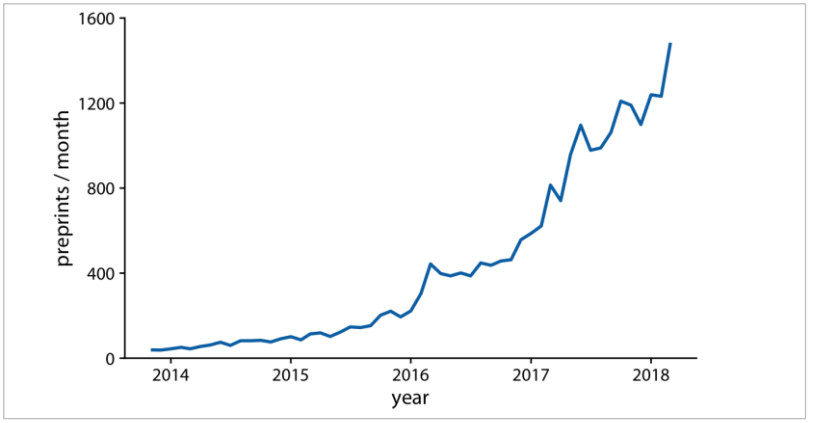
23

1. 선 그래프와 용량-반응 곡선

선 그래프(Line graph)

점을 생략 → 전반적인 추세를 더 잘 보여주며 간명하게 표현

출간 전 bioRxiv에 등록된 논문 수



[출처] Fundamentals of Data Visualization

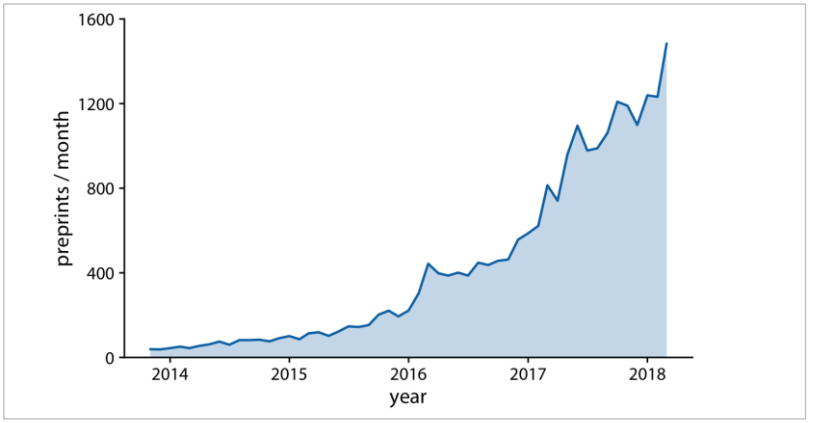
24

1. 선 그래프와 용량-반응 곡선

선 그래프(Line graph)

곡선 아랫부분을 색으로 채움 → 데이터의 흐름이 돋보임

출간 전 bioRxiv에 등록된 논문 수



[출처] Fundamentals of Data Visualization

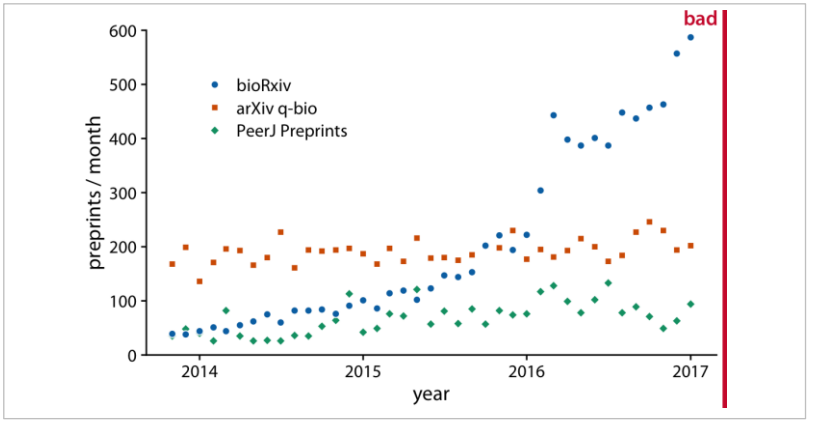
25

1. 선 그래프와 용량-반응 곡선

여러 시계열이 있는 데이터

3개의 논문서버에 등록된 논문의 수를 점으로 표현 → 데이터가 겹쳐 보임

서버 3곳에 등록된 논문 수



[출처] Fundamentals of Data Visualization

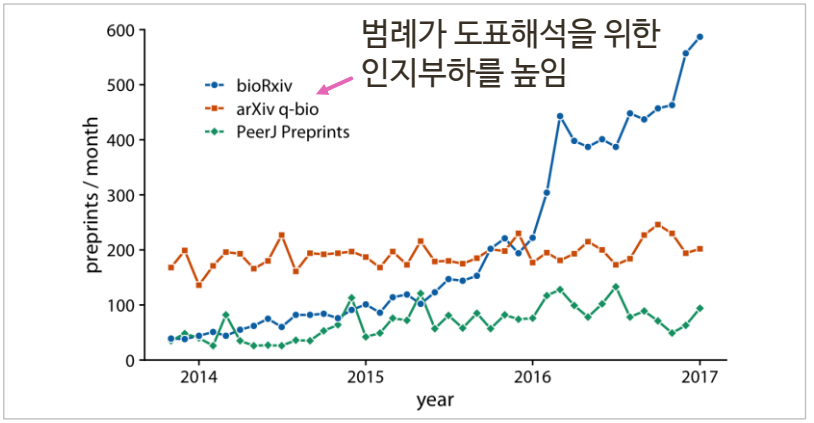
26

1. 선 그래프와 용량-반응 곡선

여러 선이 있는 그래프

범주별로 선으로 연결하여 표현 → 데이터를 구분하여 이해하기 쉬움

서버 3곳에 등록된 논문 수



[출처] Fundamentals of Data Visualization

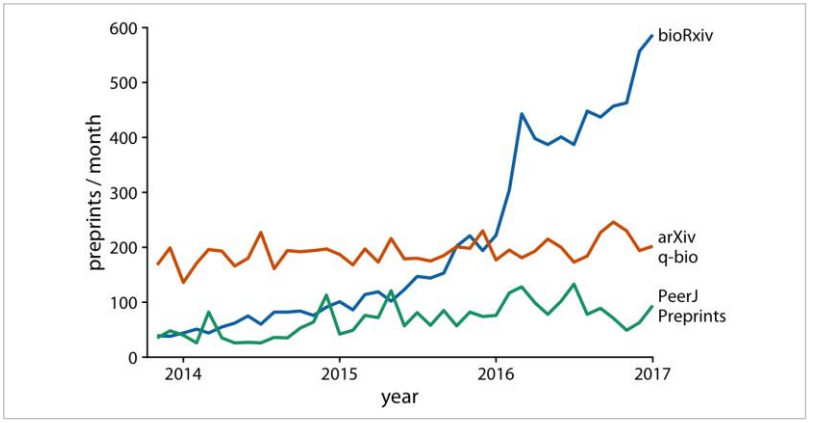
27

1. 선 그래프와 용량-반응 곡선

여러 선이 있는 그래프

범례를 선 옆에 레이블로 표현하고 점을 생략 → 간결하고 이해하기 쉬움

서버 3곳에 등록된 논문 수



[출처] Fundamentals of Data Visualization

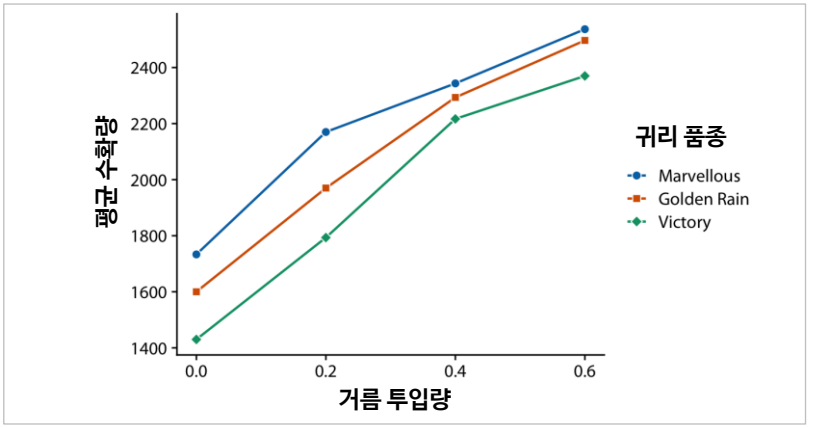
28

1. 선 그래프와 용량-반응 곡선

용량-반응 곡선(Dose-response curve)

시계열이 아니더라도 순서가 있는 데이터에 선 그래프 적용이 가능

귀리 품종별 평균 수확량 데이터



[출처] Fundamentals of Data Visualization

29

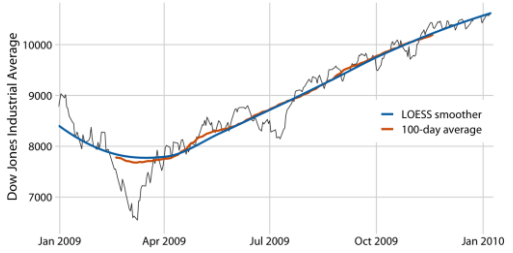
2. 추세의 시각화

추세를 나타내는 방법

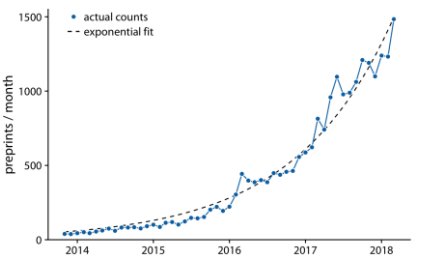
추세 시각화의 중요성

- 데이터의 중요한 특징을 한눈에 파악하기 쉬움
- 추세에서 편차 확인과 노이즈 분리가 쉬움

데이터를 평활화(보정)하는 방법



정의된 함수를 이용하는 방법



[출처] Fundamentals of Data Visualization

30

2. 추세의 시각화

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

한 시점씩 이동하며 특정 구간의 데이터 평균값을 계산하여 표현

→ 데이터의 핵심적인 양상을 보여주고 지엽적 사항이나 노이즈를 없애 줌

다우존스 산업평균지수의 일별 폐장가

[출처] Fundamentals of Data Visualization

31

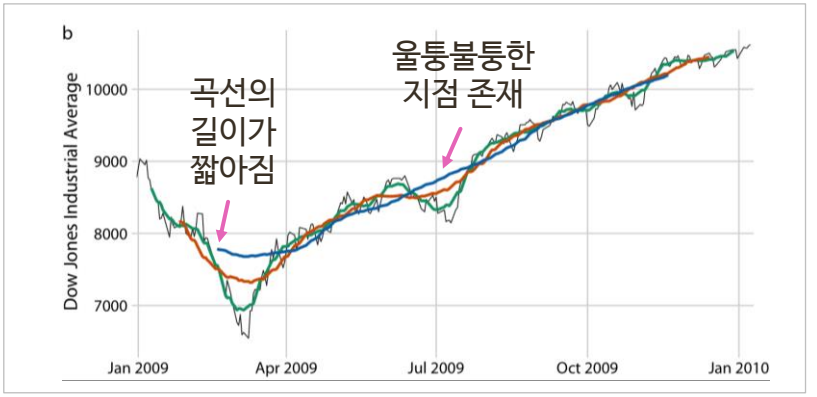
2. 추세의 시각화

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

한 시점씩 이동하며 특정 구간의 데이터 평균값을 계산하여 표현

→ 데이터의 핵심적인 양상을 보여주고 지엽적 사항이나 노이즈를 없애 줌

다우존스 산업평균지수의 일별 폐장가



[출처] Fundamentals of Data Visualization

32

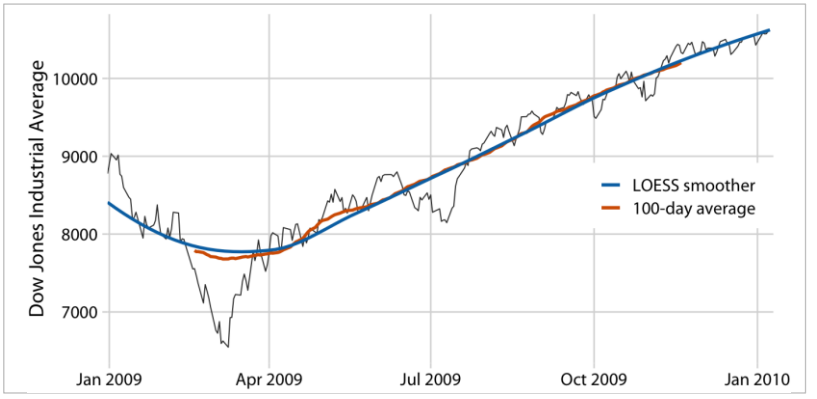
2. 추세의 시각화

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

저차원의 다항식에 데이터의 부분집합을 적합(fitting)

→ 이동 평균보다 평활도가 보통 높으며 계수를 바꿔 평활도를 조정

다우존스 산업평균지수의 일별 폐장가



[출처] Fundamentals of Data Visualization

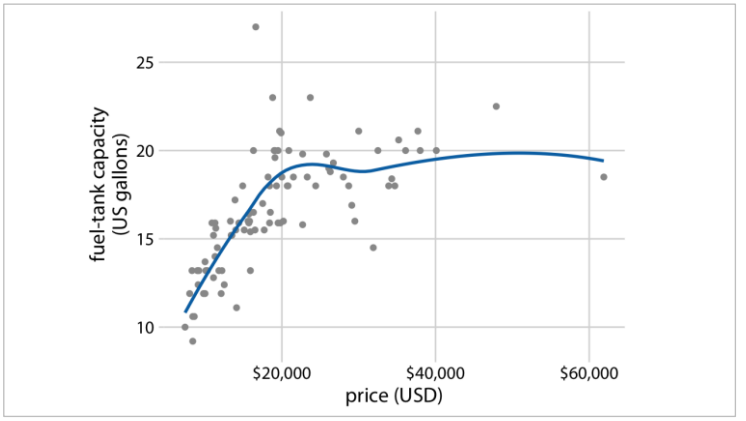
33

2. 추세의 시각화

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

시계열 데이터 뿐만 아니라 어떠한 두 변수의 산점도에도 적용 가능

차량 93대의 연료탱크 용량과 차량 가격



[출처] Fundamentals of Data Visualization

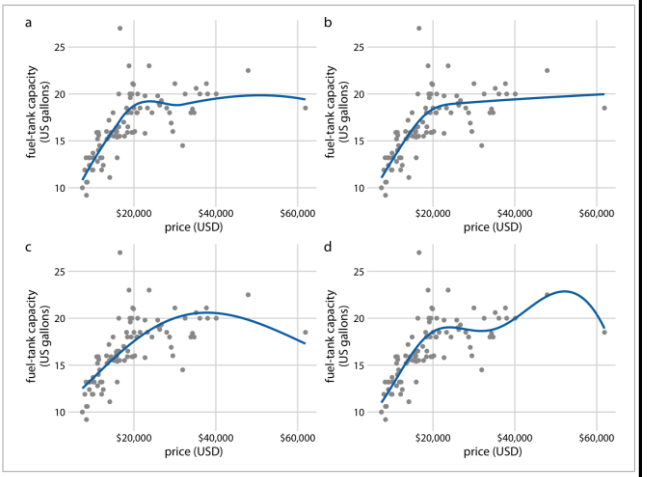
34

2. 추세의 시각화

스플라인(Spline) 모형

- 몇 개의 제어점을 기준으로 구간별로 적합(fitting) → 스플라인의 종류에 따라 다른 평탄화 결과가 나옴

차량 93대의 연료탱크 용량과 차량 가격



[출처] Fundamentals of Data Visualization

35

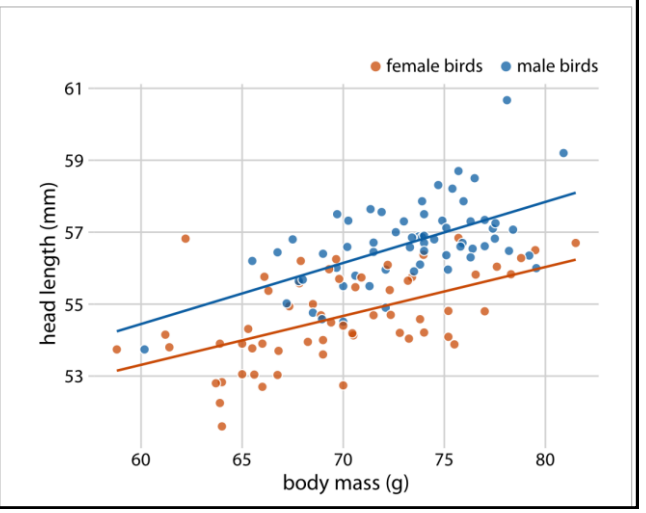
2. 추세의 시각화

정의된 함수 형식을 이용한 추세 시각화

- 데이터에 잘 맞는 형식의 함수를 정하여 데이터를 적합
  → 데이터를 가장 잘 표현하는 계수를 정함

$$y = A + mx$$

큰어치 123마리 데이터



[출처] Fundamentals of Data Visualization

36

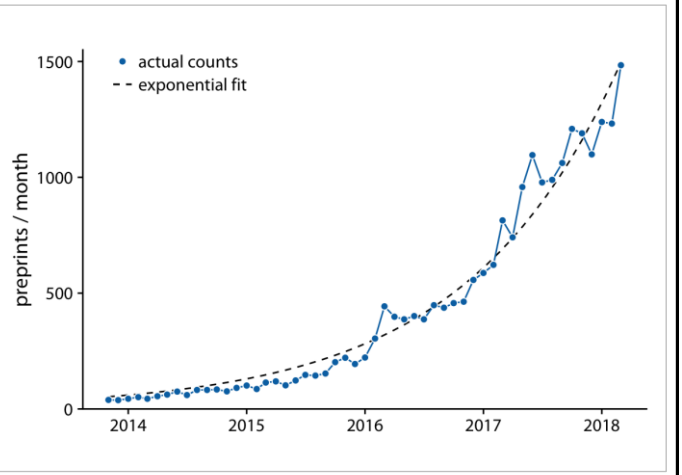
2. 추세의 시각화

정의된 함수 형식을 이용한 추세 시각화

데이터에 잘 맞는 형식의 함수를 정하여 데이터를 적합
→ 데이터를 가장 잘 표현하는 계수를 정함

$$y = A\exp(mx)$$

출간 전 bioRxiv에 등록된 논문 수



[출처] Fundamentals of Data Visualization

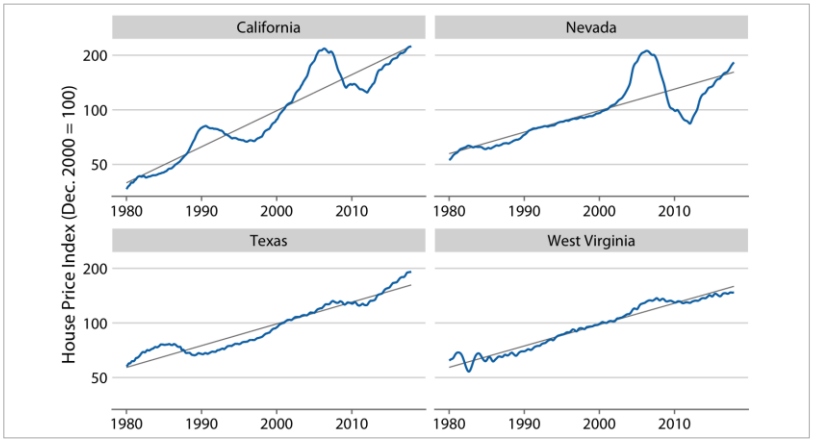
37

2. 추세의 시각화

추세 제거(Detrending)

장기적으로 지배하는 추세가 있을 때 추세를 제거 → 중요한 편차를 특정하여 강조

미국 4개 주의 프레디맥 주택가격 지수



[출처] Fundamentals of Data Visualization

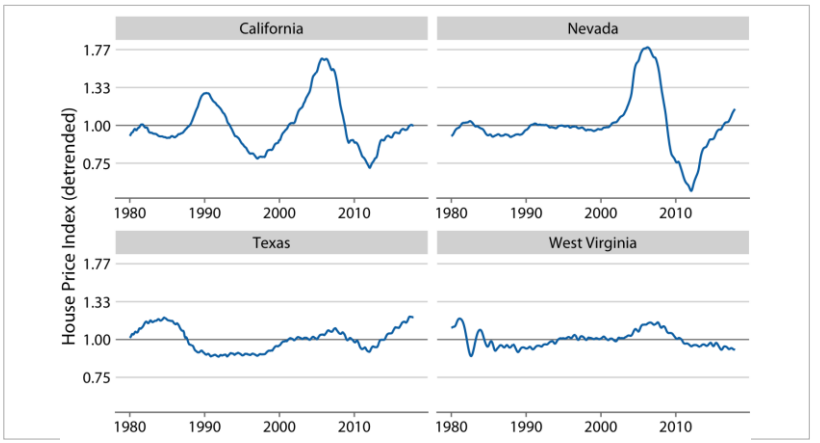
38

2. 추세의 시각화

추세 제거(Detrending)

장기적으로 지배하는 추세가 있을 때 추세를 제거 → 중요한 편차를 특정하여 강조

미국 4개 주의 프레디맥 주택가격 지수



[출처] Fundamentals of Data Visualization

39

03
chapter
실습

40

1. 변수 연관성의 시각화

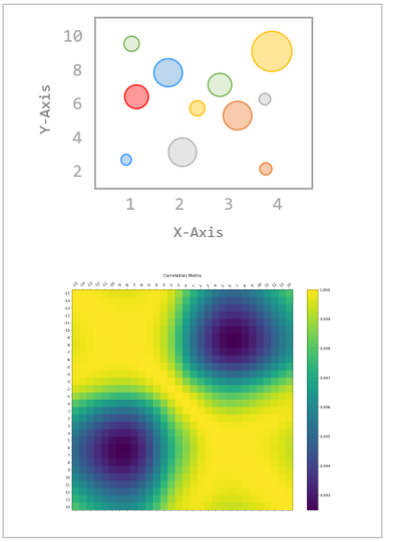
산점도와 correlogram

산점도

- scatter(x, y)

Correlogram

- matshow(df.corr())



[출처] Matplotlib Tutorial – 파이썬으로 데이터 시각화하기

41

2. 시계열 데이터의 시각화

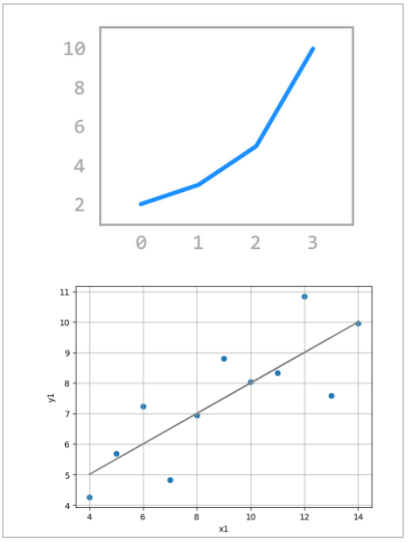
선 그래프와 선형회귀

선 그래프

- plot(x)

선형회귀

- polyfit(x, y, 1)
  * numpy 패키지를 이용



[출처] Matplotlib Tutorial – 파이썬으로 데이터 시각화하기

42

학습정리

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

43

학습정리

1 변수 연관성의 시각화

Correlogram

- 상관계수를 색을 이용하여 시각화

경사 차트(Slopegraph)

- 두 변수의 값을 선으로 연결

44

학습정리

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)

45

학습정리

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

46

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)

47

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