Data Visualization

서울대학교 데이터 사이언스 부트캠프

Gyuho Lee, hci+d lab., Department of Communication, Seoul National University

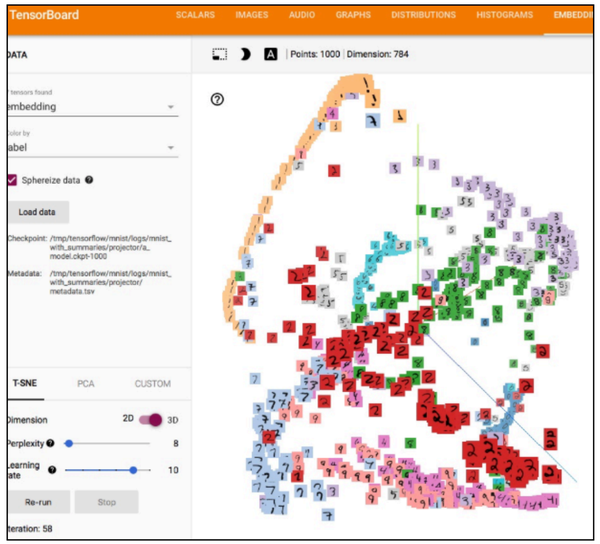
Introduction

Gyuho Lee

- 서울대학교 언론정보학과 hci+d lab. 박사과정
- Computational Social Science (Community / Journalism)
  - 자연어 처리
  - 사회 연결망
  - 데이터 시각화
- Python, R, Tableau, SQL(Database) ... 아마도 재미있어 보이는 모든 것

데이터 시각화?

- 데이터를 그래픽을 통해 전달 하는 과정 (+with 효율, 미학/ 커뮤니케이션)
- 다양한 분야(데이터 과학/저널리즘/경영 등)에서 활용(interdisciplinary)
  - 분야는 다르지만 시각화를 적절히 사용 - 1) 데이터 해석 2) 메시지 전달



https://projector.tensorflow.org

r/dataisbeautiful · Posted by u/chartr OC: 96 10 months ago

Harry Potter Characters: Screen time vs. Mentions In The Books [OC]

OC

r/dataisbeautiful · Posted by u/seracen OC: 1 28 days ago

[OC] Evolution in the rates of common ways people handle and transfer data

https://www.reddit.com/r/dataisbeautiful/comments/kgwl1h/harry_potter_characters_screen_time_vs_mentions
https://www.reddit.com/r/dataisbeautiful/comments/pkosix/oc_evolution_in_the_rates_of_common_ways_people

빅데이터의 발전과 시각화의 시대

데이터는 점점 커지고, 알고리즘은 더욱 복잡해짐

빅데이터/머신러닝을 통한 분석을 어떻게 보여줄 것인가?

“오해에 사로잡힌 사람을 설득할 때는

그의 의견을 데이터와 비교하는 방법이 매우 유용하다”

- 팩트풀니스

(https://youtu.be/Sm5xF-UYgdg)

Tableau?

Business Intelligence Software

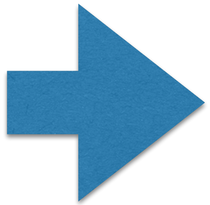
비즈니스 용도로 데이터 분석을 수행하는 소프트웨어

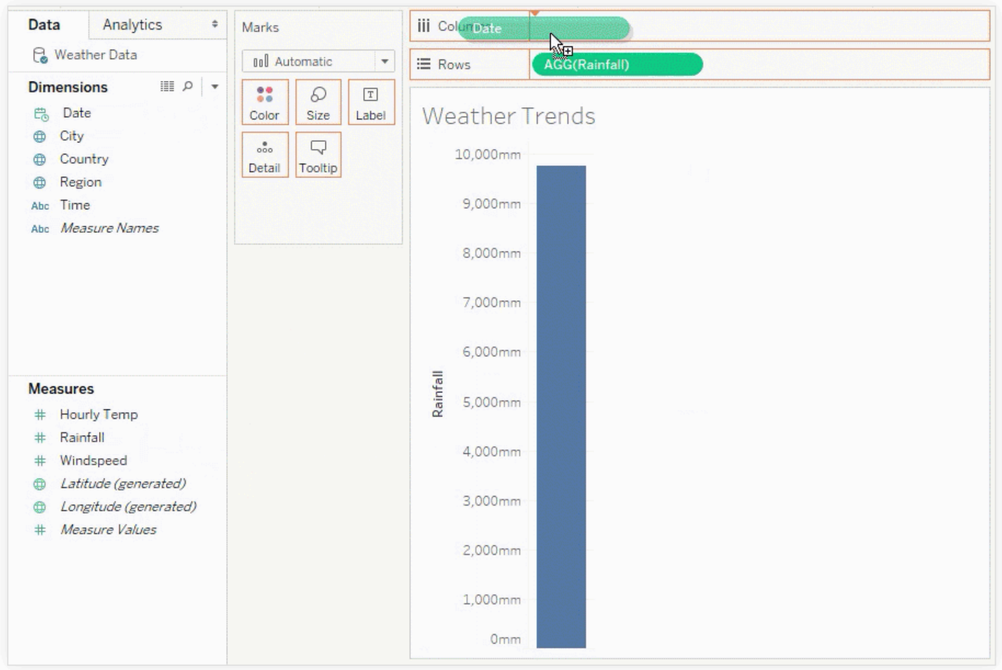
시각화도 가능하지만, 실제로는 더 많은 기능이 존재

Why Tableau?

```
In [9]: df2 = pd.DataFrame({'A': 1.,
   ...:                     'B': pd.Timestamp('20130102'),
   ...:                     'C': pd.Series(1, index=list(range(4)), dtype='float32'),
   ...:                     'D': np.array([3] * 4, dtype='int32'),
   ...:                     'E': pd.Categorical(["test", "train", "test", "train"]),
   ...:                     'F': 'foo'})
   ...:

In [10]: df2
Out[10]:
     A          B    C  D      E
0  1.0 2013-01-02  1.0  3   test  fo
1  1.0 2013-01-02  1.0  3  train  fo
2  1.0 2013-01-02  1.0  3   test  fo
3  1.0 2013-01-02  1.0  3  train  fo
```

```
def create_app(test_config=None):
    # create and configure the app
    app = Flask(__name__, instance_relative_config=True)
    app.config.from_mapping(
        SECRET_KEY='dev',
        DATABASE=os.path.join(app.instance_path, 'flaskr.sqlite'),
    )

    if test_config is None:
        # load the instance config, if it exists, when not testing
        app.config.from_pyfile('config.py', silent=True)
    else:
        # load the test config if passed in
        app.config.from_mapping(test_config)

    # ensure the instance folder exists
    try:
        os.makedirs(app.instance_path)
    except OSError:
        pass

    # a simple page that says hello
    @app.route('/hello')
    def hello():
        return 'Hello, World!'

    return app
```





Python(Pandas+Matplitlib/Plotly)와 같은 대체 조합도 가능 하지만
기술적 요구 사항이 높음 (모르면 사용불가)

Tableau는 강력한 인터렉티브 시각화를 손쉽게 구현/공유 가능
(난이도가 쉽고, 다른 BI에 비해 호환성이 좋은 편)

Why Tableau?

Tableau를 포함한 BI은 대부분 고가의 소프트웨어
하지만, Tableau의 경우 무료버전(Public)과
아카데믹 라이센스(학생/교직원) 1년 무료 제공

Tableau 준비물

Tableau Desktop(체험판/아카데믹)
Tableau/Public 계정

https://www.tableau.com/ko-kr/academic/students
https://public.tableau.com

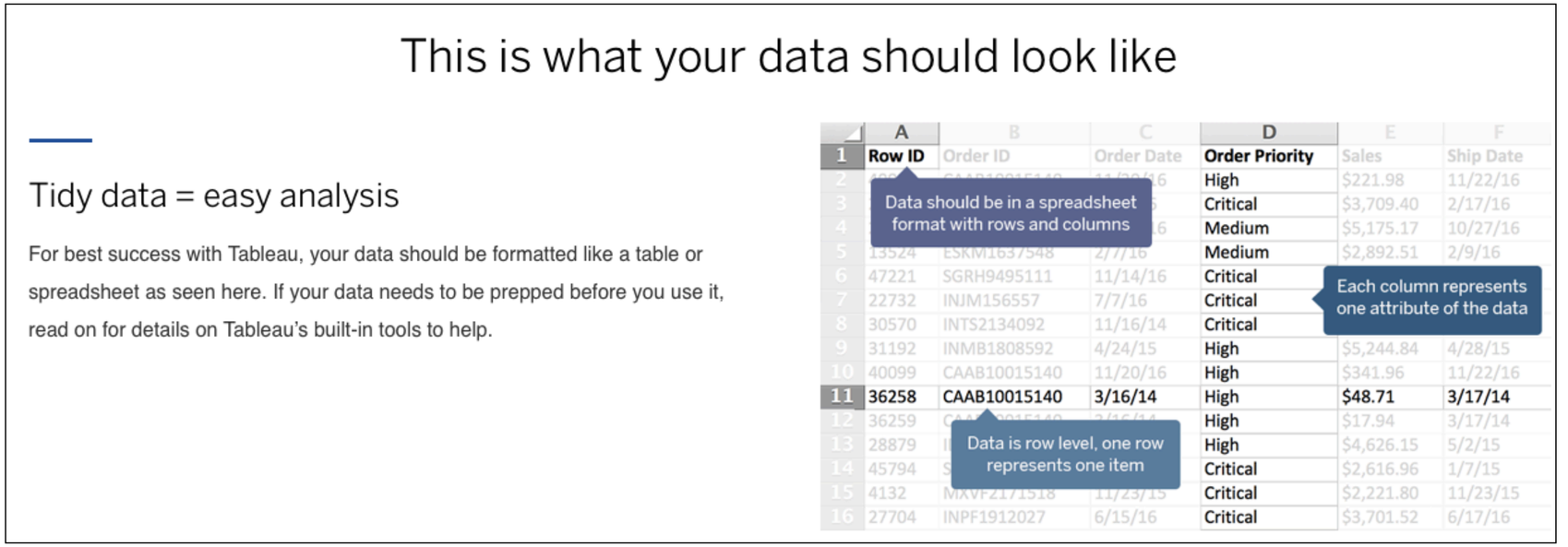
VS_CODE
(시간이 되면 씁니다)

https://code.visualstudio.com

Don’t Panic!
(자신감!)

1. 데이터 기초 개념

데이터 종류 확인



Tableau에서는 주로 Tidy Data를 사용
(데이터 사이언스에서 보편적인 형식)

Tidy Data?

Hadley Wickham
(tidyverse)
(dplyr, reshape2, ggplot2…)

각 변수 = 열(column)
각 관측치 = 행(row)
행과 열은 하나의 기준(실험/관찰)의 결과값

https://commons.wikimedia.org/wiki/File:Hadley-wickham2016-02-04.jpg
vita.had.co.nz/papers/tidy-data.pdf

Tidy Data!

세로(Column) : 서로 다른 특징이 들어갑니다

	종	색	나이
동물1	개	흰색	3
동물2	고양이	흰색	2
동물3	라마	갈색	2
동물4 (…)	개	검은색	3

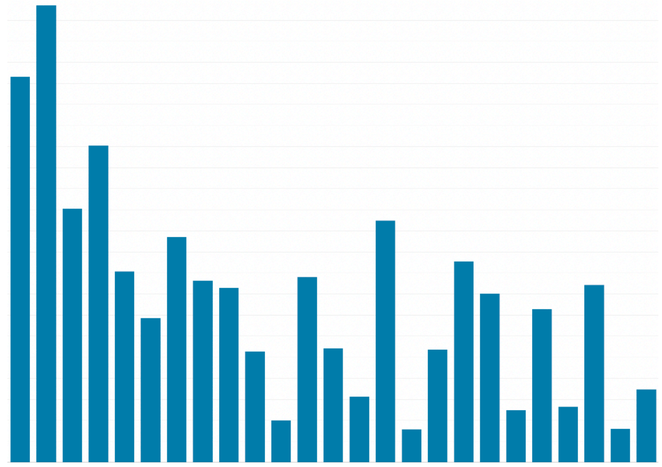
가로(row) : 각자 다른 조사 대상이 들어갑니다
(e.g. 피험자1, 2, 3, 4… / 국가1,2,3,4…)

Visualization과 Tidy Data

특징

가로/세로에 특징(IV)을 섞고
측정값(DV)을 넣으면
시각화 완성

평균 승차객 수
(측정값)



지하철 노선명 (특징)

Tableau에 적합하지 않은 데이터?

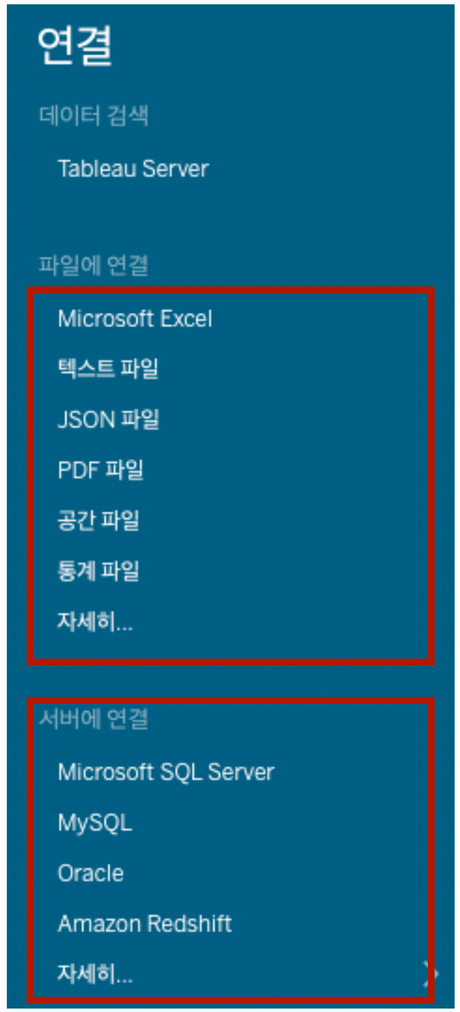
네트워크(연결망) 데이터
텍스트 데이터
년도별 데이터 (Wide Form)
….

기타 Tidy Data형식이 아닌 것

이런 데이터도 “전처리”를 잘 하면 쉽게 분석할 수 있습니다
(다음 시간에…)

2. 가볍게 시작하기

데이터 불러오기



Tableau는 크게 두 종류의 연결 지원 (파일/서버)

Tableau는 현존하는 거의 모든 포맷을 지원
(CSV, TSV, Excel, SPSS...)

서버 차원의 연결 또한 대부분의 유형 지원 (오늘은 배우지 않음)
(SQL, Google Drive/Sheet .... + 추가 커넥터 사용 가능)

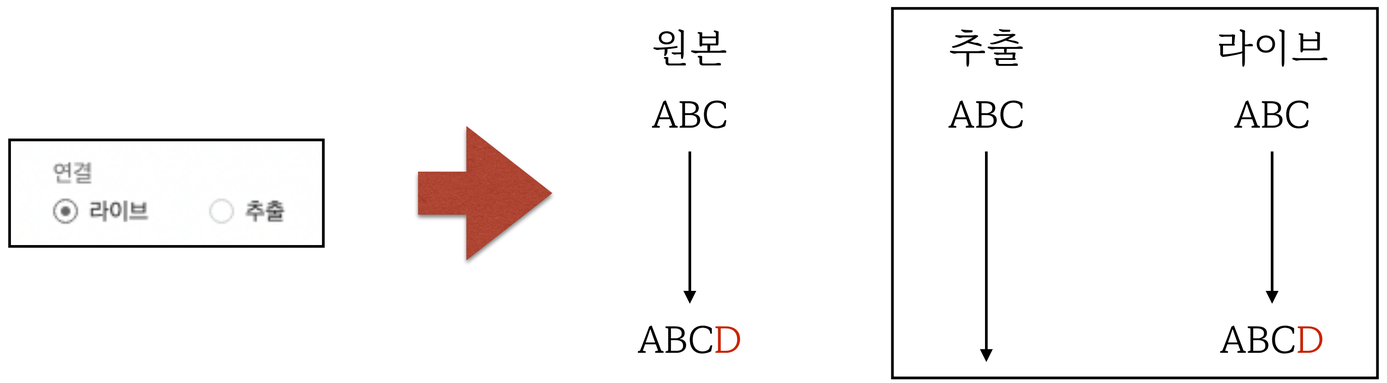
데이터

data.seoul.go.kr
서울 열린 데이터 광장

서울시 지하철호선별 역별 승하차 인원 정보 (2021 상반기)

데이터 원본 - 추출/라이브

데이터 원본창 - 데이터 확인 / 편집/ 조합 가능 (오늘은 조합은 배우지 않음)
그 외로 중요한 개념은 추출/라이브가 존재

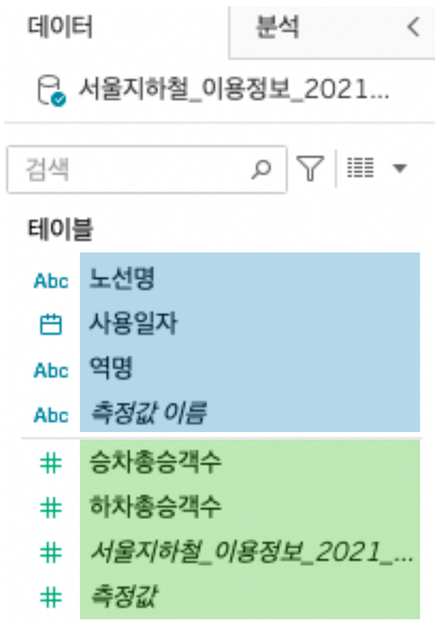


과거 데이터, 혼자 작업 - 추출연결(한번 불러오면 그대로)
협업, 실시간 데이터 - 라이브 연결(원본 파일이 바뀌면 연동)

데이터 선택

Tabelau는 차트 생성을 위해 차원/측정값을 조합

데이터 유형
(숫자, 문자, 날짜, 지리...)



차원 (IV)

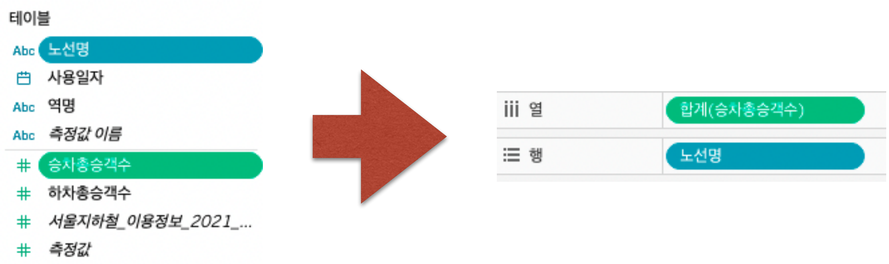
측정값 (DV)

데이터 테이블
데이터 속성(색), 유형/역할(기호) + 계층/그룹화 가능(Drag-Drop)

차트 생성하기

Tabelau는 차트 생성을 위해 2가지 방법 지원

추천 표현방식 적용
(Tableau 추천, 간단하고 효율적)

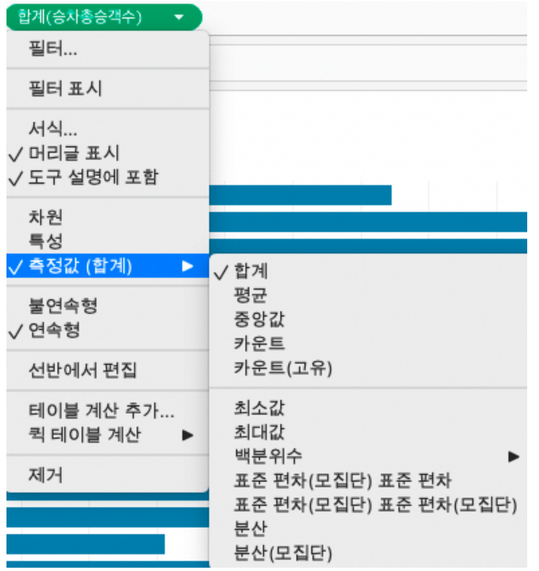


행/열 직접 조합
(세부 사항 조절 가능)

차트 세부사항 조절

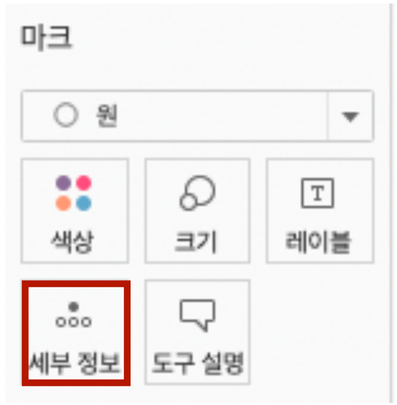
마크 옵션과 컨텍스트 메뉴를 조절하여 차트의 세부사항 조절

마크 옵션을 이용하여
차트의 표현 형식 변경
(클릭/드래그_드롭)

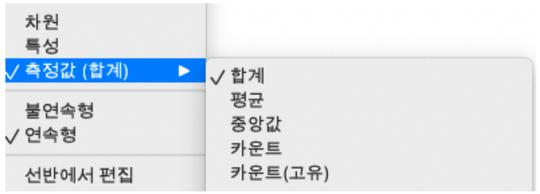


컨텍스트 메뉴 (▽)를 통해
데이터 연산 방식 변경

중요 세부사항



세부정보 : 데이터를 묶어주는 기능
(고급 시각화의 핵심)



카운트/카운트(고유)
데이터를 세는 두가지 방식

2. 데이터 분석

Tableau의 데이터 분석 기능

일반 통계
(평균/분포
사분위)

회귀분석
(추세분석)

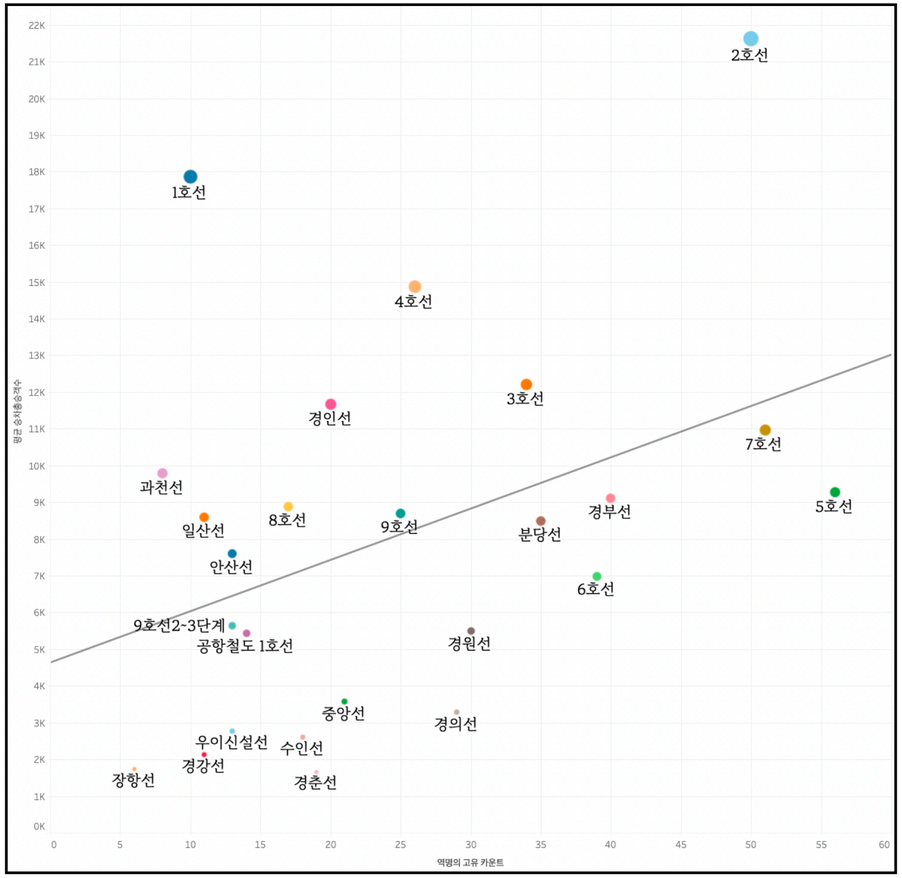
집단분석
(클러스터링)

시나리오
(What-If?)

Tableau에서는 분석과 관련해 기본 4가지 기능 제공

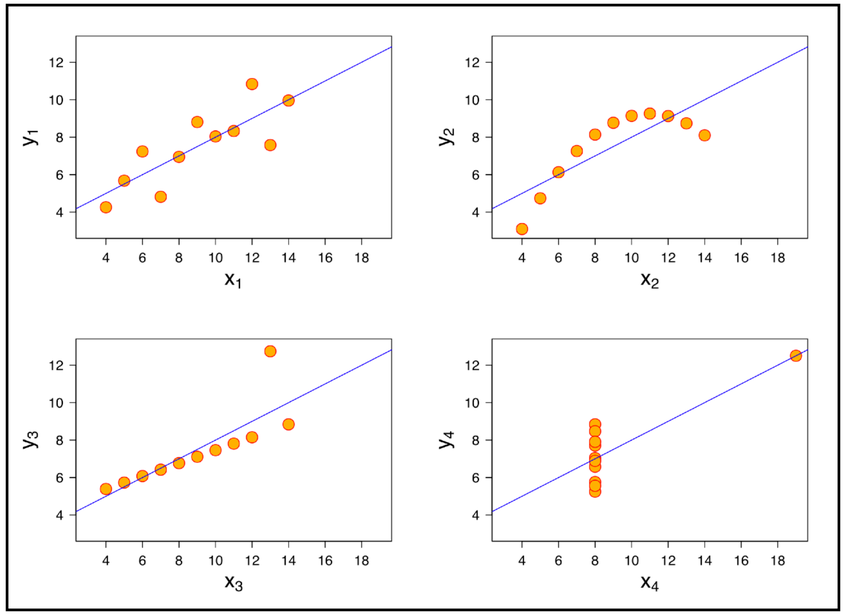
2019 하반기 : 자연어 검색 기능 추가(Ask Data)
(본 수업에서는 다루지 않음, 설정 과정이 복잡함)

통계와 시각화 I



노선별 역수는 이용량과 상관관계가 있을까?
(상어와 아이스크림 문제)

통계와 시각화 II



수치만 본다면 데이터의 문제를 찾기 어렵다
(Anscombe's quartet - 위 데이터의 통계 수치는 모두 동일)

https://en.wikipedia.org/wiki/Anscombe's_quartet
https://commons.wikimedia.org/wiki/File:Anscombe.svg

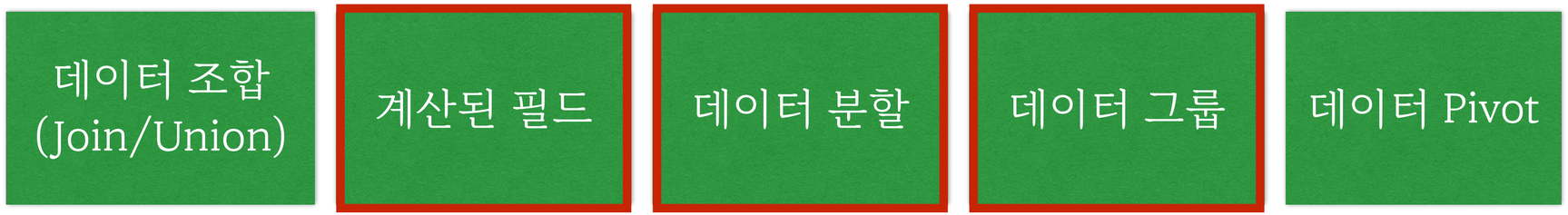
3. 데이터 전처리

데이터

기상청 날씨데이터 서비스
기상자료개방포털

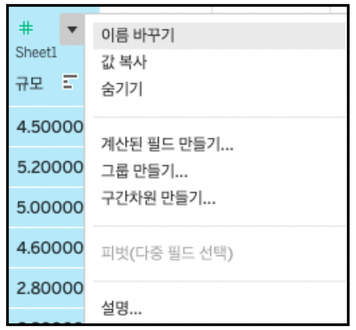
https://data.kma.go.kr/cmmn/main.do
지진 정보

Tableau의 데이터 전처리 기능



데이터 과학의 90% 작업은 전처리 작업
Tableau에서는 데이터 전처리를 위한 여러 기능을 제공
(SQL/Excel과 비슷한 형태)

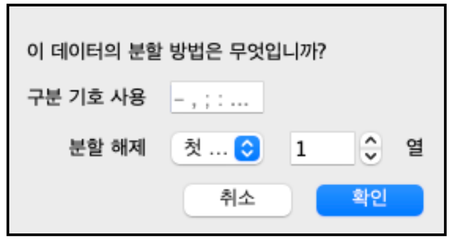
계산된 필드



기존 컬럼을 연산하여 새로운 컬럼 생성

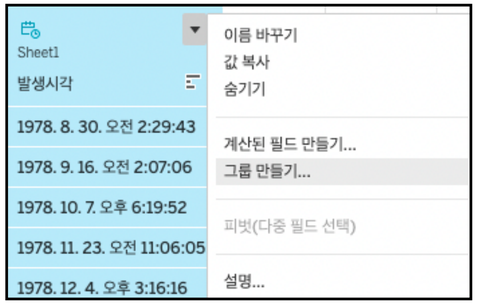
엑셀/SQL과 거의 유사 (굳이 배울 필요는...?)

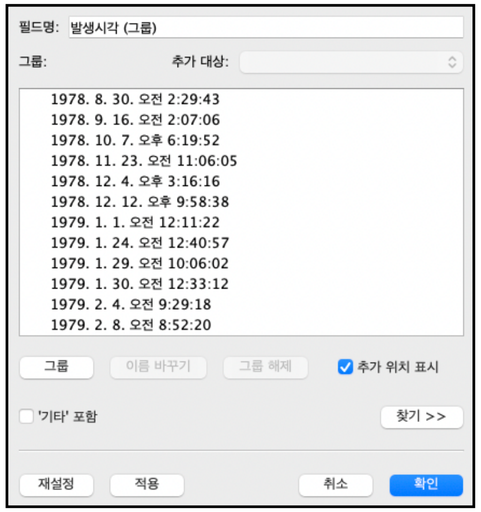
데이터 분할



문자열 데이터를 자동으로 나눠주거나 (분할)
특정한 기준에 따라 나눠줌 (사용자 지정 분할)
(e.g. 주소에서 시/군/구 추출)

데이터 그룹





기존 컬럼의 연속형 데이터를 명목형으로 변환
(e.g. "나이" 변수를 "연령대" 데이터로 변환)

더 복잡한 작업이 필요하다면?

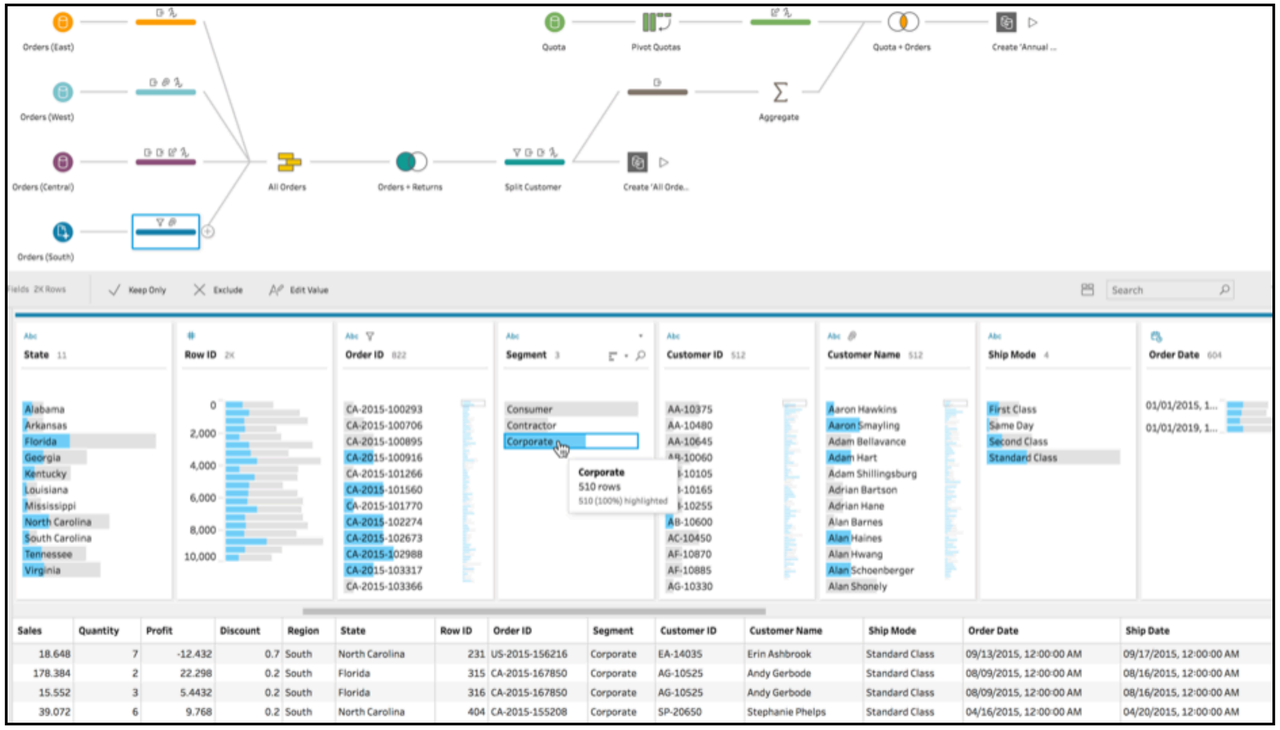


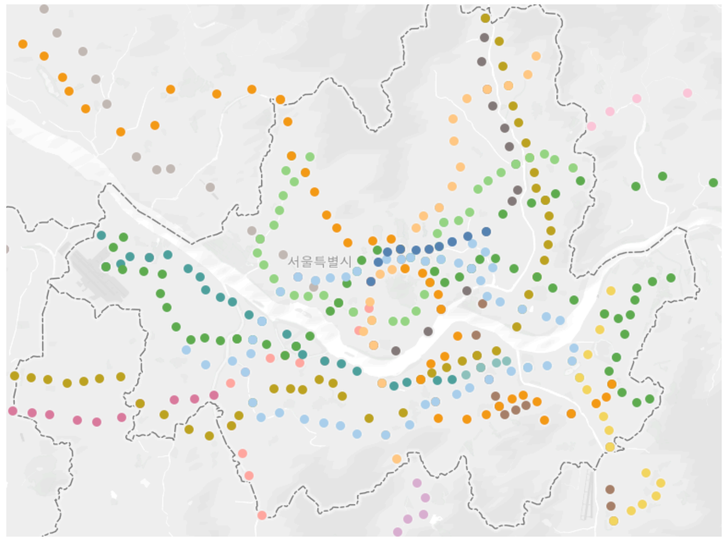
Tableau Prep

전처리 전용으로 만들어진 Tableau 연계 프로그램

(유동적, 대규모, DB 작업에 적합 / 여기서는 다루지 않음)

4. 데이터 시각화 심화

지리 데이터 시각화



지리 데이터의 시각화는 좌표기준(위도/경도), 지역기준 (주소)
두 종류가 존재(오늘은 좌표만 학습)

필터/페이지

	종	색	나이
동물1	개	흰색	3
동물2	고양이	흰색	2
동물3	라마	갈색	2
동물4 (…)	개	검은색	3

	종	색	나이
동물1	개	흰색	3

페이지 

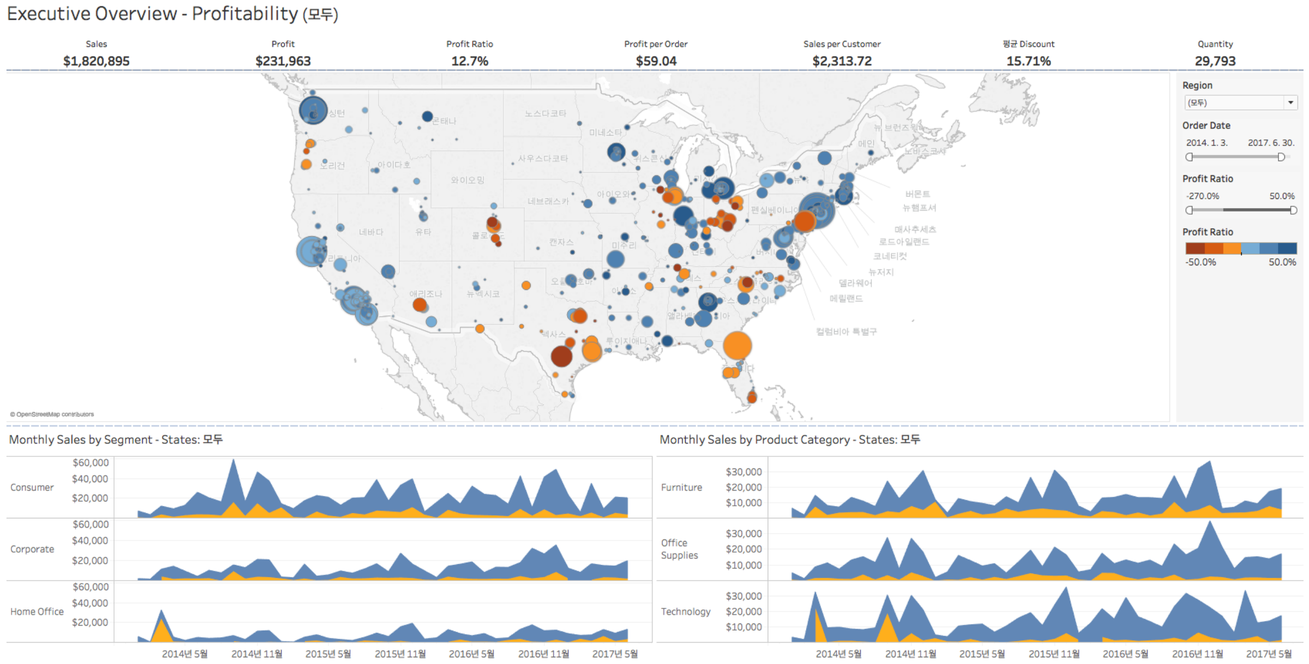
	종	색	나이
동물1	개	흰색	3
동물4 (…)	개	검은색	3

필터 

필터/페이지는 둘 다 조건에 따라 데이터를 추출(+시각화) 할때 쓰지만
다른 활용 방식을 가지고 있음 (오늘은 필터 위주로)

5. 대시보드 구성하기

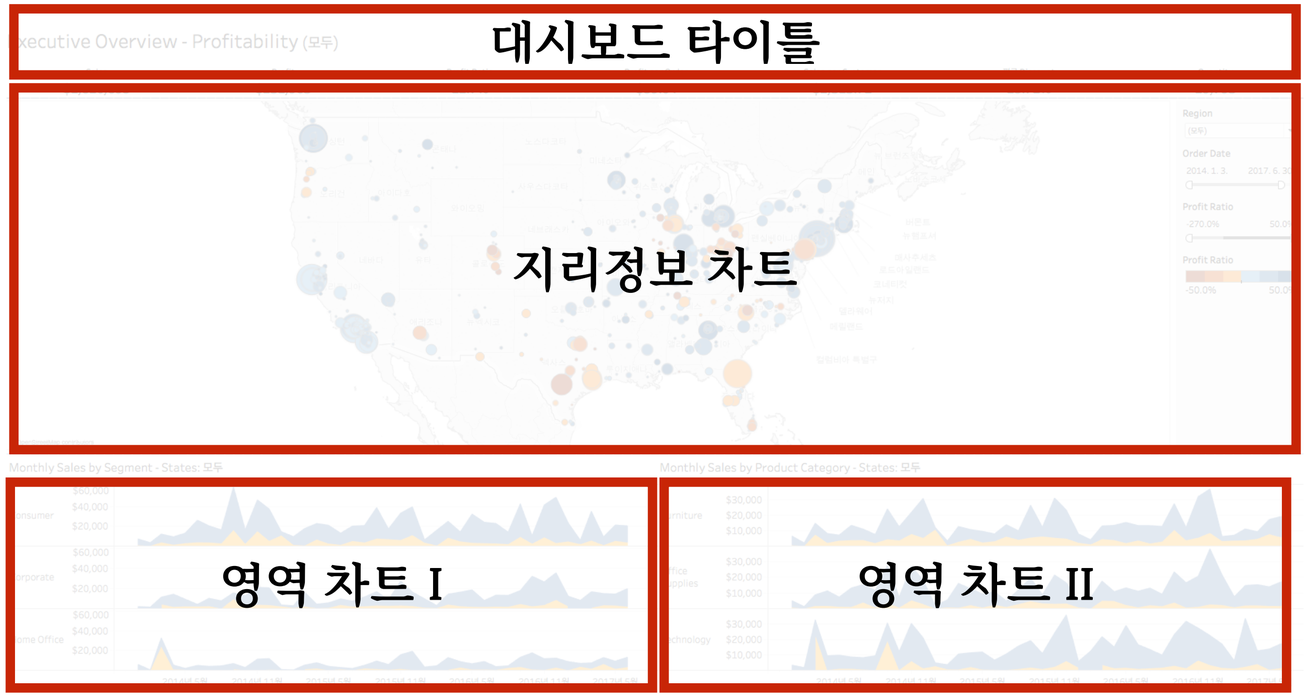
대시보드 조합하기



대시보드 = 개별 시각화의 조합

https://www.tableau.com/ko-kr/solutions/workbook/map-and-track-profitability-with-an-executive-overview

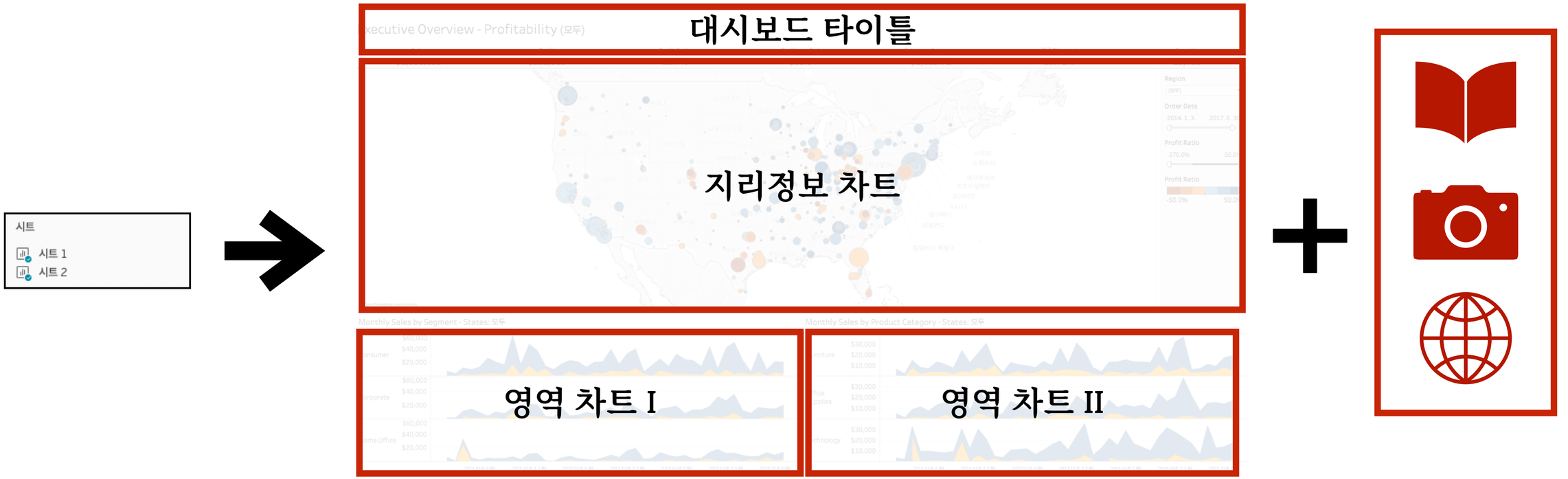
대시보드 조합하기



대시보드 = 개별 시각화의 조합

https://www.tableau.com/ko-kr/solutions/workbook/map-and-track-profitability-with-an-executive-overview

대시보드 조합하기



대시보드 (시트 조합/GRID) + 개체 추가

https://www.tableau.com/ko-kr/solutions/workbook/map-and-track-profitability-with-an-executive-overview

6. 저장/온라인에 게시하기

Tableau의 데이터 저장 (오프라인)

Tableau에서는 Tableau 형식(twb, twbx)으로 데이터를 저장하거나
데이터셋/이미지/파워포인트 형식으로 내보내기가 가능함

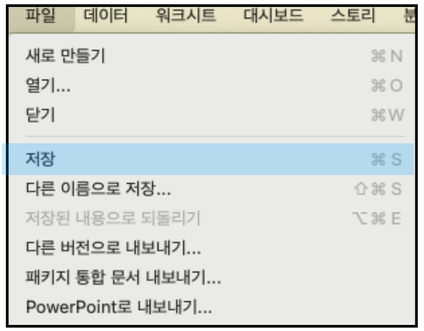
.twb
시각화 구성만

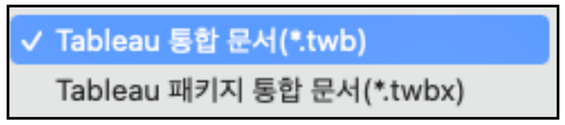
.twbx
시각화+데이터

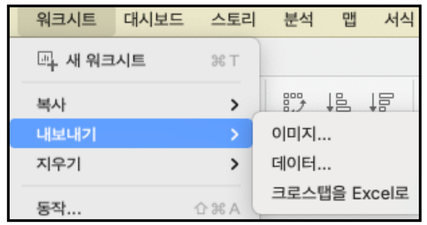
데이터파일

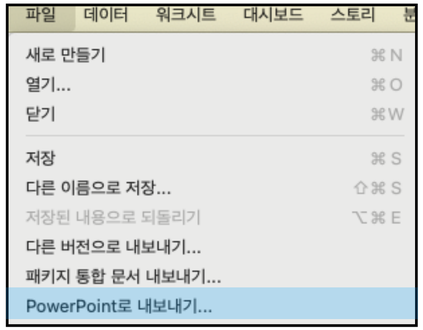
이미지파일

파워포인트









Tableau의 데이터 공유 (온라인 게시)

Tableau에서 제공하는 온라인 공간을 이용하거나(Tableau Public)
회사 등의 공간에서 서버를 구성하여 내부적으로 사용할 수 있음

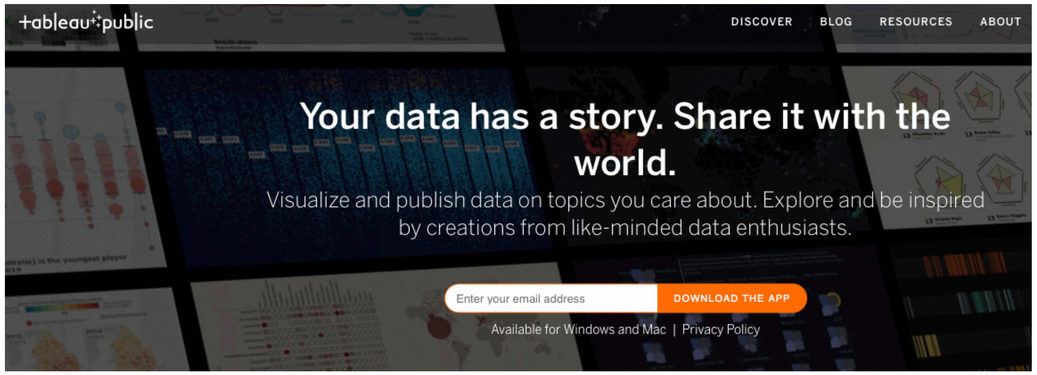
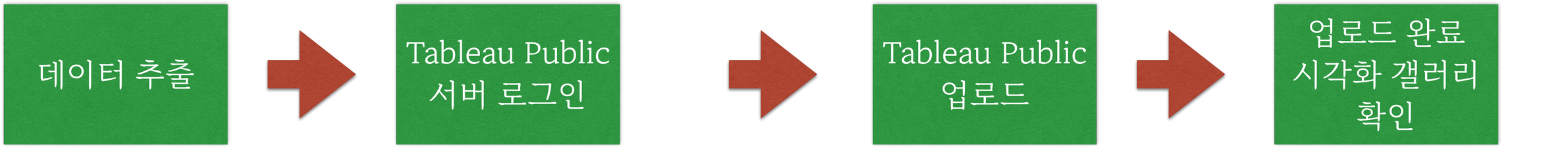


Tableau Public의 경우 무료로 사용 가능하지만 (온라인 공간+프로그램)
Tableau 제공 서버에 데이터가 업로드 되기 때문에 저작권 자료의 사용에 유의할 필요가 있음
(연구목적으로만 사용, 공개 불가 등 제한조건 고려)

Tableau Public을 활용한 온라인 게시



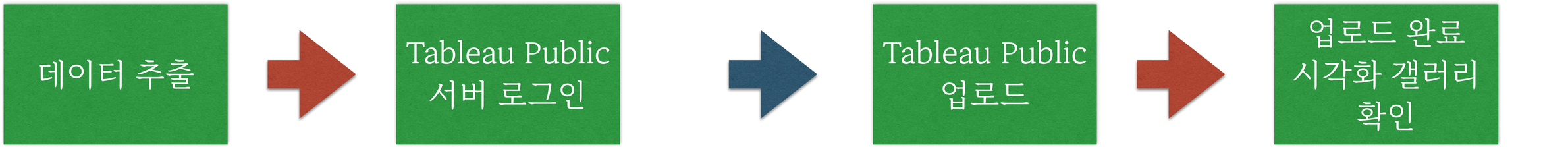
"데이터" 메뉴에서
온라인에 업로드할
자료 선택/추출

"서버"메뉴에서
Tableau Public 선택
Tableau Public에 저장 선택

저장된 시각화/데이터
내 Tableau Public
계정에 업로드 (자동)

내 Public 갤러리에
시각화 자료 업데이트
확인/편집

Tableau Public을 활용한 온라인 게시



"데이터" 메뉴에서
온라인에 업로드할
자료 선택/추출

"서버"메뉴에서
Tableau Public 선택
Tableau Public에 저장 선택

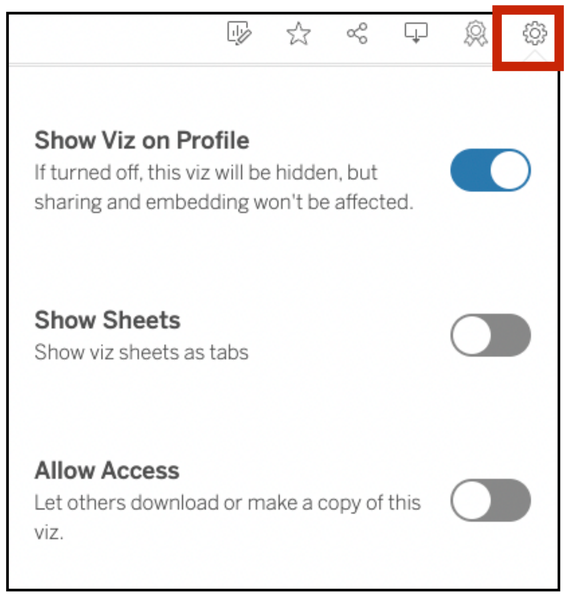
저장된 시각화/데이터
내 Tableau Public
계정에 업로드 (자동)

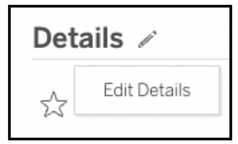
내 Public 갤러리에
시각화 자료 업데이트
확인/편집

Local(PC)

Server(Tableau Public)

Tableau Public 공유 설정





내 온라인 계정 기준으로

- 선택 시각화를 공개/비공개
- 다운로드 허용
- 제목, 설명 등의 내용

등의 디테일 설정 가능

Tableau Public 임베딩/링크공유

Tableau Public의 시각화를 공유하는 방식은 두가지

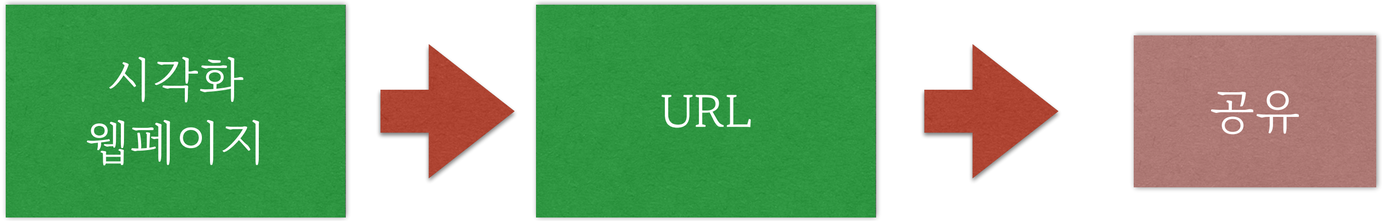
HTML문서에 시각화를 삽입하는 방식

독립된 시각화 페이지에 접속하는 링크 제공 방식

Tableau Public 임베딩/링크공유 예제

시각화 공유 과정 비교

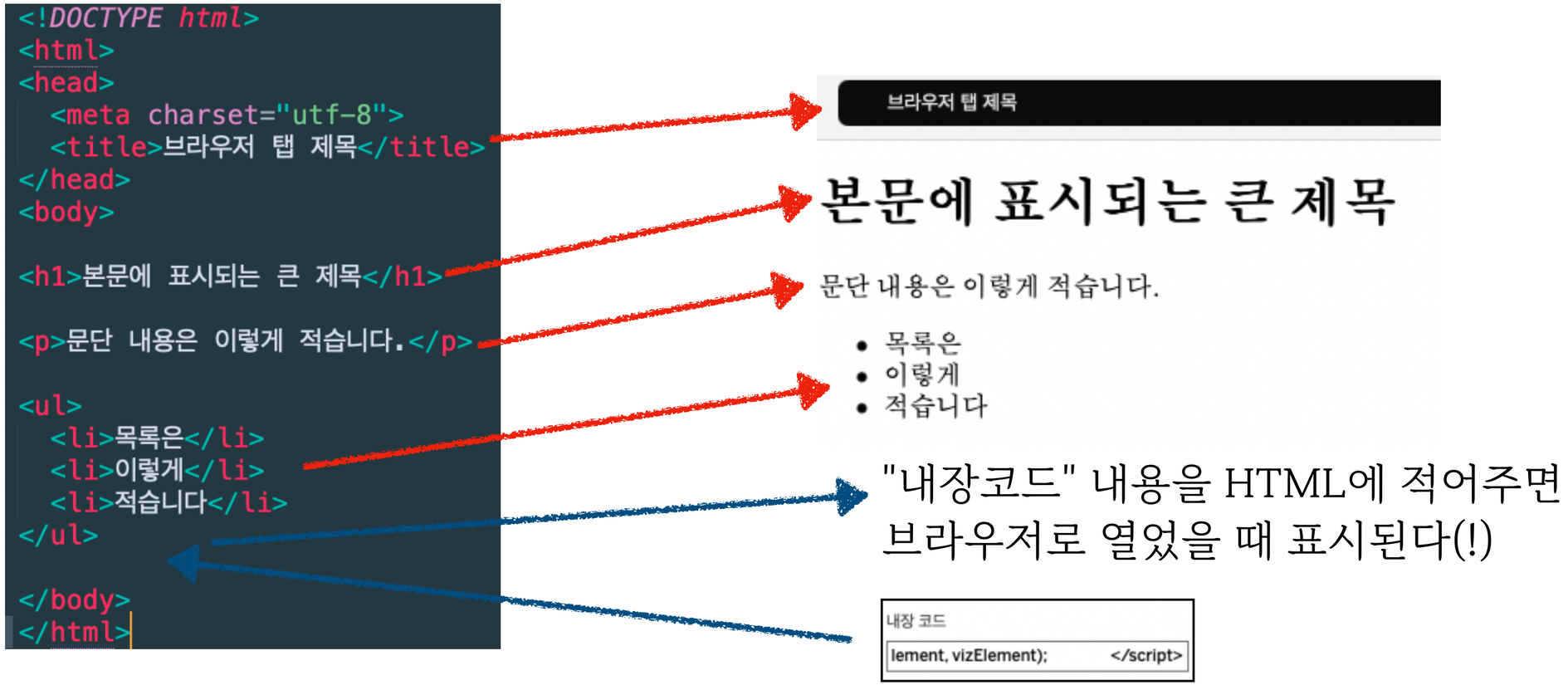
1) Tableau 시각화 페이지 링크 (시각화만 공유 가능, 간편)



2) Tableau 시각화 임베딩 (시각화 추가 내용 공유 가능, 복잡)

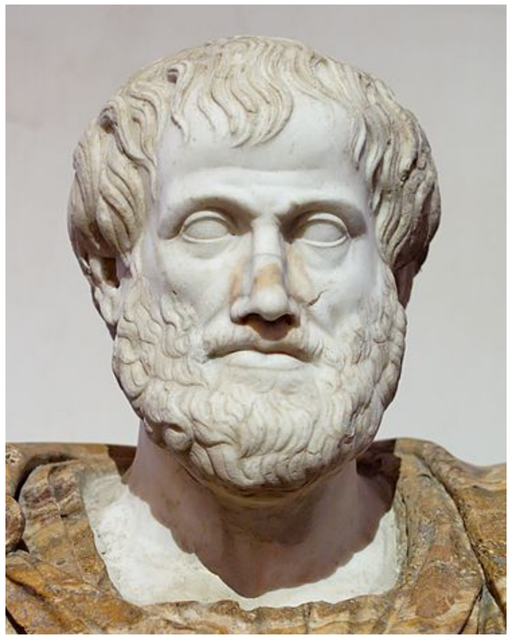
Tableau Public 임베딩

우리가 인터넷에서 보는 페이지는 내부적으로 계층적 구조(HTML)



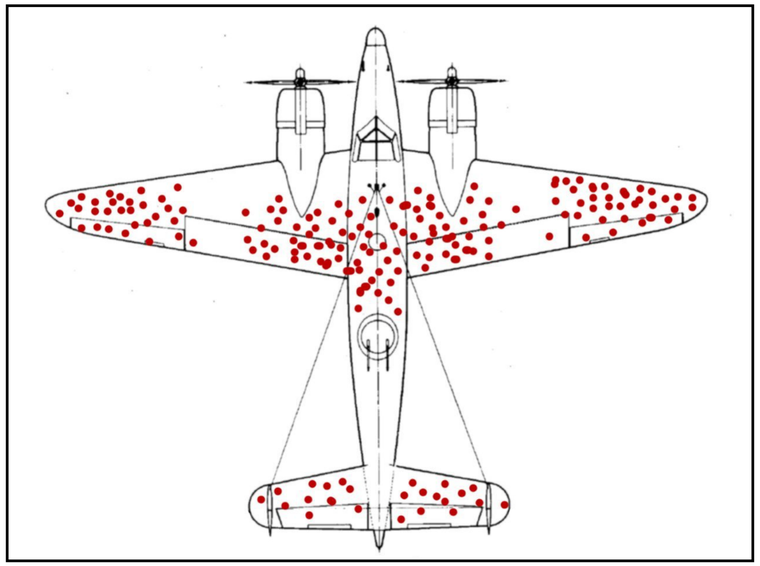
7. 시각화 논의/사례

좋은 데이터 연구의 조건?

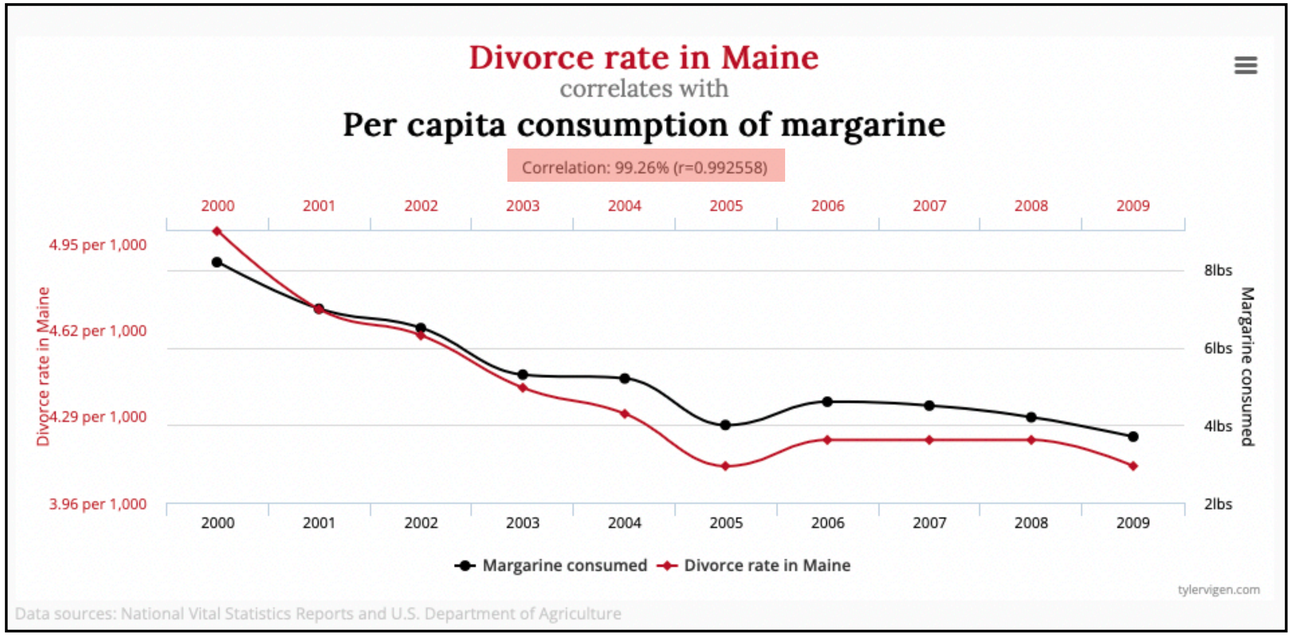


https://en.wikipedia.org/wiki/Aristotle
https://commons.wikimedia.org/wiki/File:Aristotle_Altemps_Inv8575.jpg

데이터 연구의 함정들



https://en.wikipedia.org/wiki/Survivorship_bias
https://commons.wikimedia.org/wiki/File:Survivorship-bias.png



http://tylervigen.com/spurious-correlations

데이터를 통한 사회 분석

Cause of Death - Reality vs. Google vs. Media

(포스팅)
https://www.reddit.com/r/dataisbeautiful/comments/8cwcbu/cause_of_death_reality_vs_google_vs_media_oc/

(시각화)
https://g.redditmedia.com/x1cdV5AkvPOFUAMi1Lar_fQQcpG7VcWhUJvdibJPG0U.gif?fm=mp4&mp4-fragmented=false&s=9627a66446fff1a5dcee66a5cdc17c0b

데이터를 통한 사회 분석

Inside Airbnb
(Murray Cox)

http://insideairbnb.com/

프로젝트 내용
(https://en.wikipedia.org/wiki/Inside_Airbnb)

데이터를 통한 사회 분석

Maps of Heat
(Probable Futures)

https://probablefutures.org/heat/maps-of-heat/

데이터를 통한 사회 분석

The Cheap and Easy Climate Fix That Can Cool the Planet Fast

https://www.bloomberg.com/graphics/2021-methane-impact-on-climate/

데이터를 통한 사회 분석

Poison in the Air

시각화 https://projects.propublica.org/toxmap/

기사 https://www.propublica.org/article/toxmap-poison-in-the-air

데이터를 통한 사회 분석

Tracking Federal Purchases to Fight the Coronavirus

시각화
https://projects.propublica.org/coronavirus-contracts/

+Federal Procurement Data System
https://www.fpds.gov/fpdsng_cms/index.php/en/

마무리 / 오늘 다루지 않은 내용

데이터 연결/조작

온라인 데이터 연결(+크롤링), 데이터셋 조합(Join, Union 등..), DB다루기

시각화 응용/심화

지리 데이터, 시계열 데이터 심화(+Time Series 조합)

페이지, 필터 기능 심화

파라메터 응용(인터렉티브 핵심!)

머신러닝/통계

Python, R + 통계기법 시각화(+상관관계 분석)

기타

온라인 공유 심화 + 팁 + 모범사례(?)

특이한 시각화 기법들(워드 크라우드, 순위 차트, 간트 차트 등…)

…