

Tableau를 통한 데이터 시각화와 분석

2019.10.14 (1.5 hours)

소개

서울대학교 언론정보학과 박사과정, 이규호(hci+d lab.)

데이터 사이언스(+저널리즘), 머신러닝(딥러닝), 사회연결망

[Python, R]+Tableau(Visualization), AMOS(SEM), Gephi(Network)

0. 데이터 분석과 TABLEAU

데이터 과학(머신러닝)/저널리즘/사회과학(연구)?

<Google Tensorflow>

<New York Times, The Upshot>

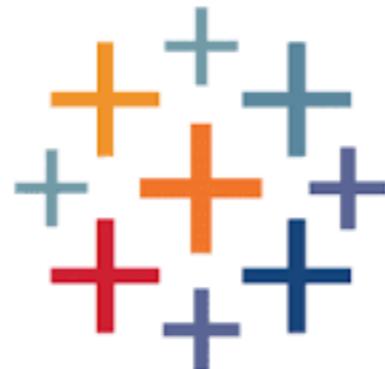
분야는 다르지만 둘 다 시각화를 적절히 사용합니다
1) 데이터 한눈에 파악 2) 중요한 메시지 강조

빅데이터의 발전과 시각화의 시대!

빅데이터를 통한 분석을 어떻게 보여줄 것인가?

“오해에 사로잡힌 사람을 설득할 때는
그의 의견을 데이터와 비교하는 방법이 매우 유용하다” – 팩트폴리스

TABLEAU?



Google Data Studio

Business Intelligence Software

비즈니스 용도로 데이터 분석을 수행하는 소프트웨어
시각화도 가능하지만, 실제로는 더 많은 기능이 존재

왜 TABLEAU인가?



다양한 BI중 자유선택이 가능
Python+Pandas+Matplotlib+Plotly와 같은 대체 조합도 가능

하지만 Tableau는 강력한 인터렉티브 시각화를 손쉽게 구현할 수 있습니다

TABLEAU FOR FREE?

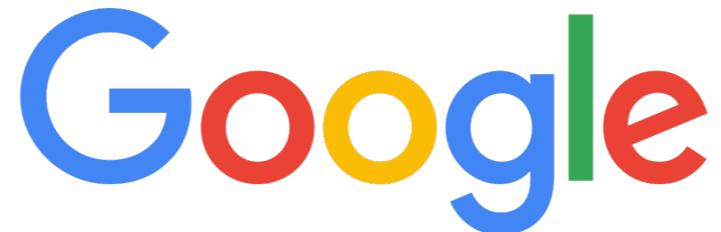


Tableau는 비교적 고가의 소프트웨어
하지만 무료로 Public 버전 사용가능(저장기능제한)
학생/교직원 -> 1년 라이센스 무료(서류 인증 필요, 영어 서류 스캔해서 제출)

TABLEAU학습 준비물



Tableau Desktop(체험판)
Tableau/Public 계정



Google 계정
Google Drive활성화



자신감

1. 데이터 기초 개념

데이터 종류 확인하기

This is what your data should look like

Tidy data = easy analysis

For best success with Tableau, your data should be formatted like a table or spreadsheet as seen here. If your data needs to be prepped before you use it, read on for details on Tableau's built-in tools to help.

A	B	C	D	E	F
Row ID	Order ID	Order Date	Order Priority	Sales	Ship Date
1	CAAB10015140	11/20/16	High	\$221.98	11/22/16
2	CAAB10015140	11/20/16	Critical	\$3,709.40	2/17/16
3	ESKM1637548	2/7/16	Medium	\$5,175.17	10/27/16
4	SGRH9495111	11/14/16	Medium	\$2,892.51	2/9/16
5	INJM156557	7/7/16	Critical		
6	INTS2134092	11/16/14	Critical		
7	INMB1808592	4/24/15	High	\$5,244.84	4/28/15
8	CAAB10015140	11/20/16	High	\$341.96	11/22/16
9	36258	3/16/14	High	\$48.71	3/17/14
10	CAAB10015140	3/16/14	High	\$17.94	3/17/14
11	INPF1912027	6/15/16	High	\$4,626.15	5/2/15
12	MXVFZ171518	11/23/15	Critical	\$2,616.96	1/7/15
13	28879		Critical	\$2,221.80	11/23/15
14	45794		Critical	\$3,701.52	6/17/16
15					
16					

Tableau에서는 주로 Tidy Data를 사용합니다

TIDY DATA?

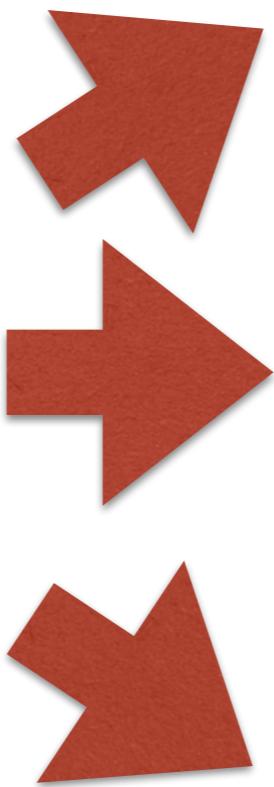


Hadley Wickham
(dplyr, reshape2, ggplot2)

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

VITA.HAD.CO.NZ/PAPERS/TIDY-DATA.PDF

간단히 말해 봅시다



개

흰색

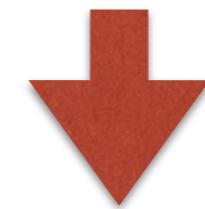
잘생김(?)

...

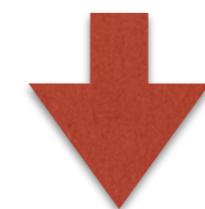
간단히 말해 봅시다

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

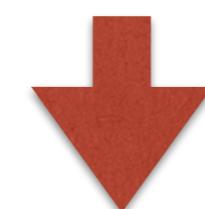
세로에는 서로 다른 특징이 들어갑니다



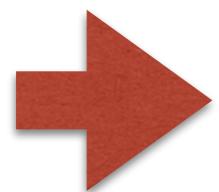
특징 1



특징 2



특징 3



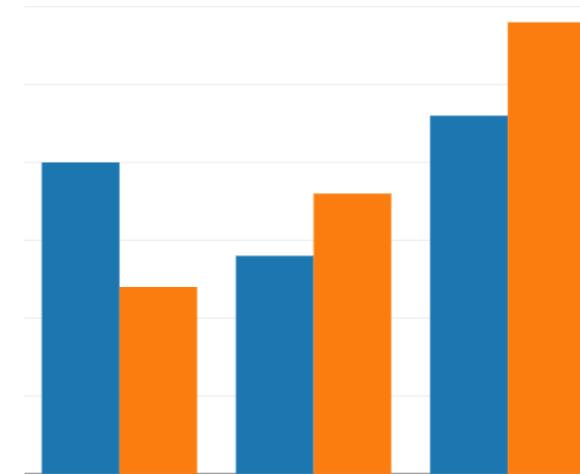
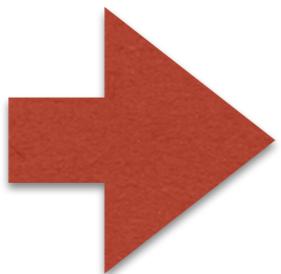
가로에는 각자 다른 조사 대상이 들어갑니다
(e.g. 영희, 민수…/학생1, 학생2/동물1, 동물2…)

TABLEAU와 TIDY DATA의 관계는?

특징

인원	측정값
1	100
2	100
3	100
4	100
5	100
6	100
7	100
8	100
9	100
10	100
11	100
12	100
13	100
14	100
15	100
16	100
17	100
18	100
19	100
20	100
21	100
22	100
23	100
24	100
25	100
26	100
27	100
28	100
29	100
30	100
31	100
32	100
33	100
34	100
35	100
36	100
37	100
38	100
39	100
40	100
41	100
42	100
43	100
44	100
45	100
46	100
47	100
48	100
49	100
50	100
51	100
52	100
53	100
54	100
55	100
56	100
57	100
58	100
59	100
60	100
61	100
62	100
63	100
64	100
65	100
66	100
67	100
68	100
69	100
70	100
71	100
72	100
73	100
74	100
75	100
76	100
77	100
78	100
79	100
80	100
81	100
82	100
83	100
84	100
85	100
86	100
87	100
88	100
89	100
90	100
91	100
92	100
93	100
94	100
95	100
96	100
97	100
98	100
99	100
100	100

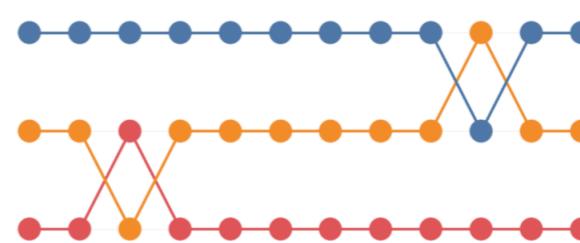
**특징을 섞고, 측정값을 넣으면
시각화가 됩니다**



세로막대차트

특징 : 남녀, 고/저학년

표현가 : 성저



순위차트

특집 : 팀 종류 / 시기(년)

표현값 : 순위

TABLEAU에 적합하지 않은 데이터?

네트워크 데이터(가로/세로 축이 모두 사람일때)
년도별 데이터(년도별로 데이터가 늘어나 있는 형태일때)
텍스트 데이터(가로, 세로축이 아예 존재하지 않음)
....

기타 Tidy Data형식이 아닌 것

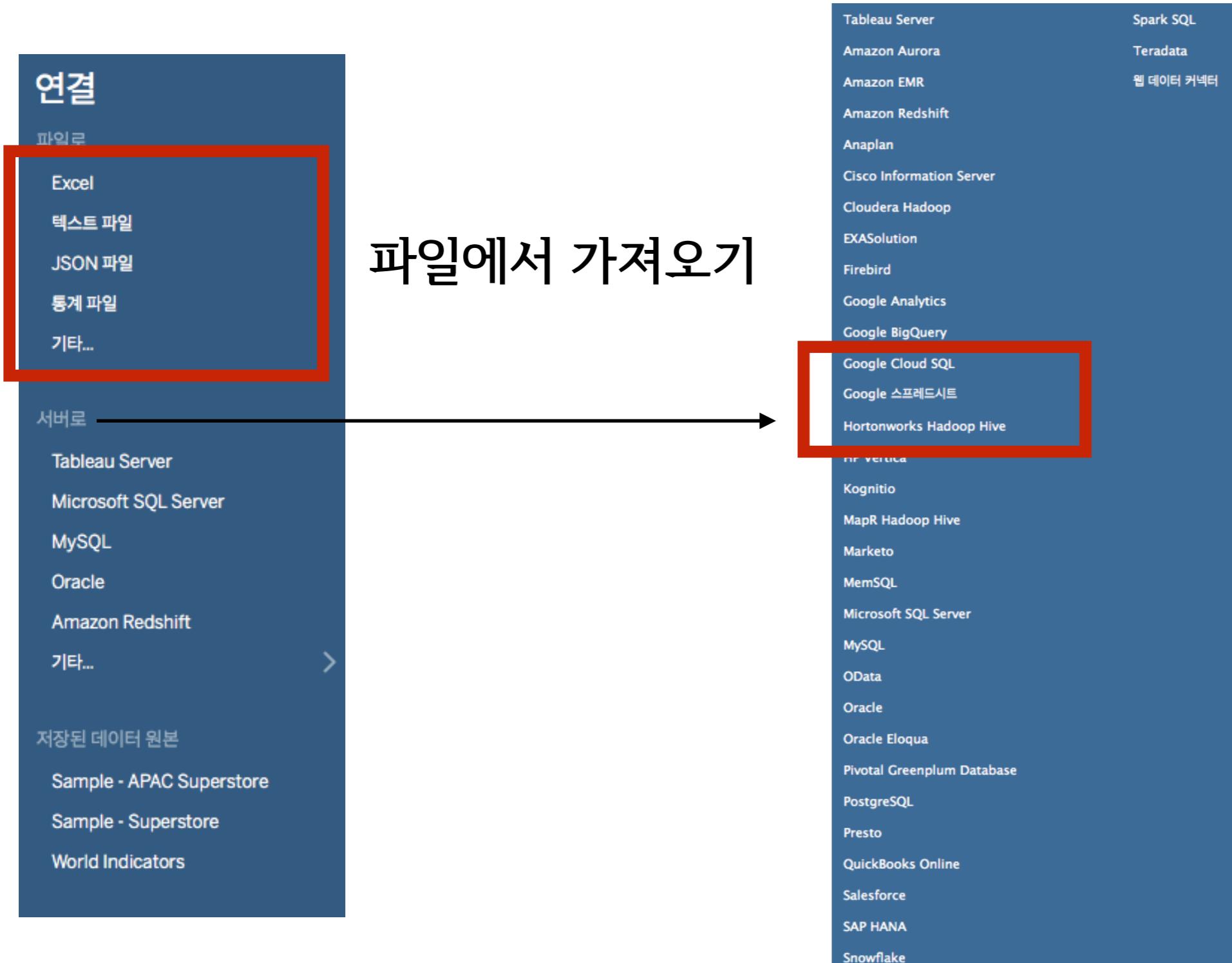
하지만, 이런 데이터도 “전처리”를 잘 하면 쉽게 분석할 수 있습니다
(다음 시간에…)

1. 가볍게 시작하기

DATA!

Tableau에서는 다양한 형태의 데이터 편집 가능
(**CSV**, TSV, **EXCEL**, SPSS, SQL…)

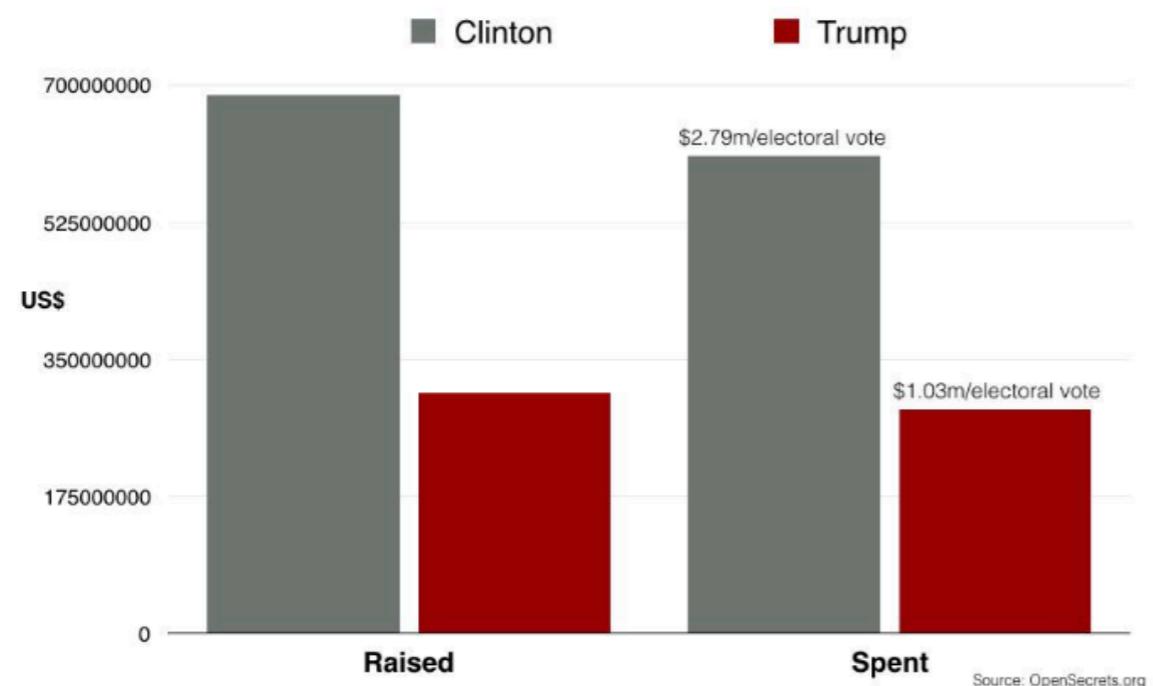
연결(2종류 존재)



추출? 라이브?

Clinton vs Trump: Campaign Funds & Spending 2016

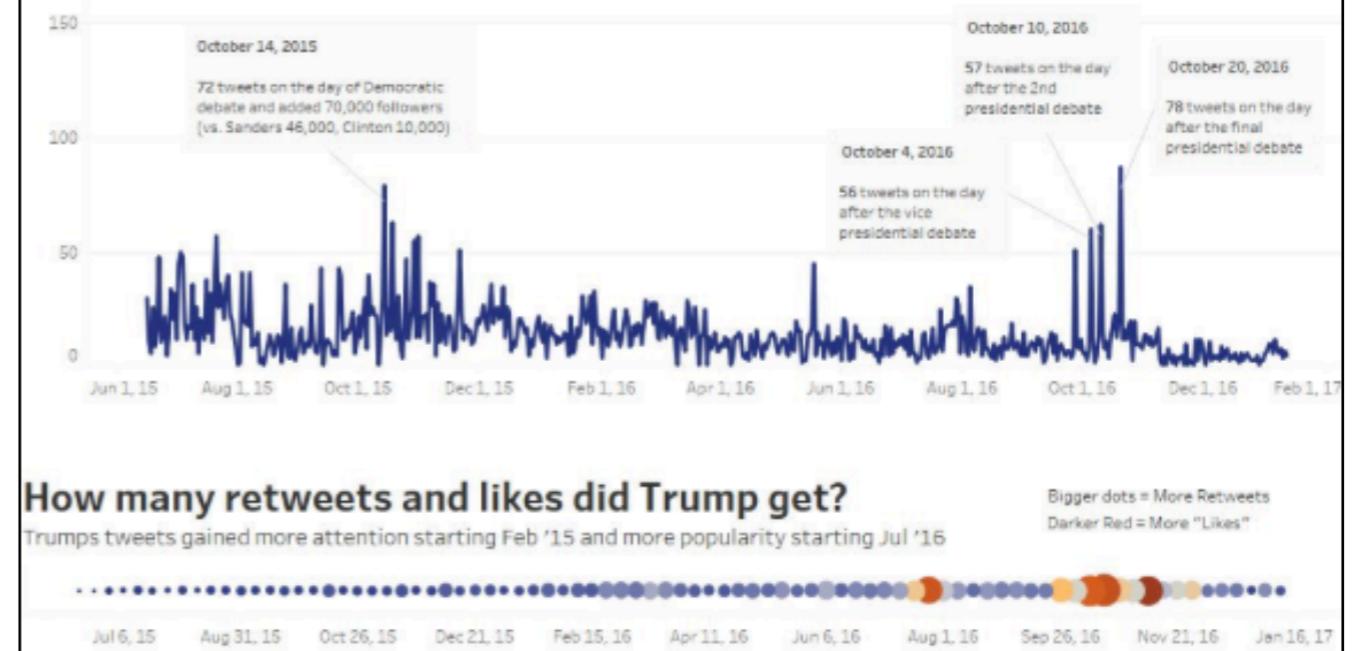
Trump paid 63% less per electoral vote than Clinton



A Glance at Trump's Tweets

How many tweets did Trump post after he ran for President?

Trump tweeted more around presidential debates, vice presidential debate and Democratic debate



How many retweets and likes did Trump get?

Trump's tweets gained more attention starting Feb '15 and more popularity starting Jul '16

과거 데이터, 혼자 작업 - 추출연결(한번 불러오면 바뀌지 않습니다)
협업, 실시간 데이터 - 라이브 연결(원래 파일이 바뀌면 바뀝니다)

DATA!

data.seoul.go.kr
서울 열린 데이터 광장

서울시 지하철호선별 역별 승하차 인원 정보

Sheet	Open API	File			
서울시 지하철호선별 역별 승하차 인원 정보에 대한 파일 명세서를 제공합니다. 명세서를 다운로드 하세요.					
No	파일명	파일크기(KB)	마지막수정일	최초공개일	다운로드
1	CARD_SUBWAY_MONTH_201501.csv	1,002	2016.06.10	2016.06.10	Down
2	CARD_SUBWAY_MONTH_201502.csv	1,030	2016.06.10	2016.06.10	Down
3	CARD_SUBWAY_MONTH_201503.csv	1,018	2016.06.10	2016.06.10	Down
4	CARD_SUBWAY_MONTH_201504.csv	1,007	2016.06.10	2016.06.10	Down
5	CARD_SUBWAY_MONTH_201505.csv	1,043	2016.06.10	2016.06.10	Down

데이터 원본창

subway

subway.csv

연결 추가

subway
텍스트 파일

파일

subway.csv

새 Union

필드 정렬	수정한 날짜
subway.csv	subway.csv
사용일자	노선명
Abc	#
2017. 1. 1.	경원선
1907	역ID
가능	역명
4,194	승차총승객수
3,852	하차총승객수
2017. 1. 1.	경원선
1908	녹양
2,612	
2,468	
2017. 1. 1.	경원선
1909	양주
5,036	
5,004	
2017. 1. 1.	경원선
1910	덕계
1,145	
1,123	
2017. 1. 1.	경원선
1911	덕정
3,823	
4,188	

데이터 조합/편집 가능

차트 작성하기[1]

The screenshot shows a data analysis interface with a blue arrow pointing from the left side to the right side, indicating a transformation or flow from raw data to a summarized view.

Left Side (Raw Data):

- 데이터 분석
- subway
- 차원
- 노선명, 등록일자, 사용일자, 역ID, 역명, 측정값 이름
- 측정값:
 - # 격차별이동인구
 - # 승차총승객수
 - # 하차총승객수
 - # 레코드 수
 - # 측정값

Right Side (Summarized View):

열	노선명	월(사용일자)
행	합계(승차총승객수)	

데이터 편집/속성변경후

조합

문자, 색의 차이는?

앞의 기호(데이터 형태)/색상(데이터 연속성)
서로 연결된 데이터 형식은? 그룹화/계층화

차트 작성하기[2]



요인 선택 후 오른쪽 위 “표현방식” 클릭
Tableau에서는 자동으로 추천 그래프 작성
(직접클릭/더블클릭)

디테일 수정

© MARK ANDERSON

WWW.ANDERTOONS.COM



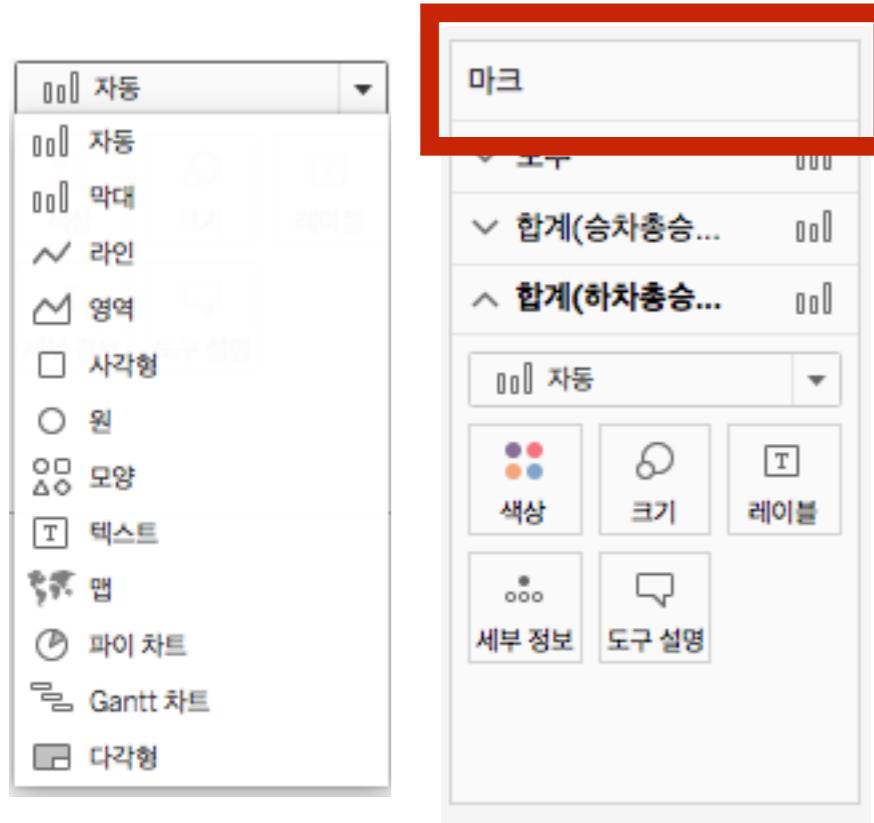
"I am so over the details."

차원

측정값

표현형식

디테일 수정(측정값/표현방식)

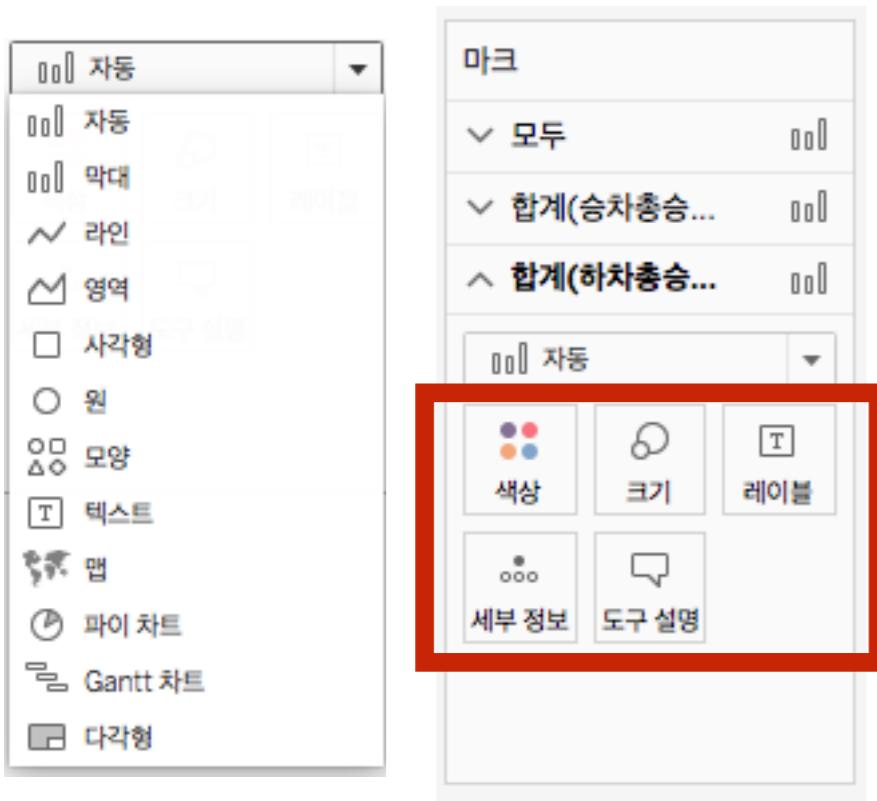


마크(표현 방식 세부 조절 가능)

마크의 조합을 통해 “복잡한”
그래프 생성 가능

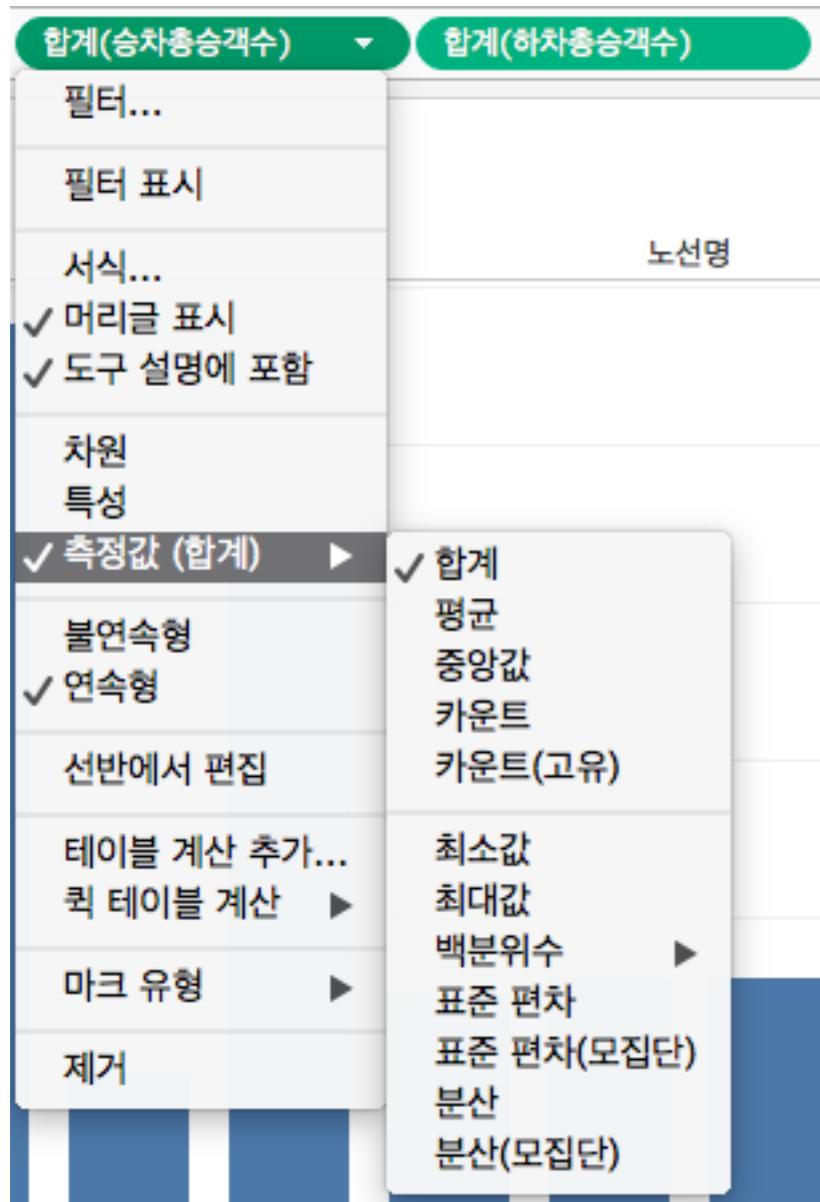
새로운 측정값이 추가될때마다 Stack

디테일 수정(측정값/표현방식)



세부옵션
(세밀한 조절은
아래 Box의 조합으로 가능)

디테일 수정(측정값/표현방식)



내가 원하는 형태의 측정값을 사용 가능
비즈니스 분석에서 빈번하게 쓰는 기능은
(퀵 테이블 계산)에서 사용 가능

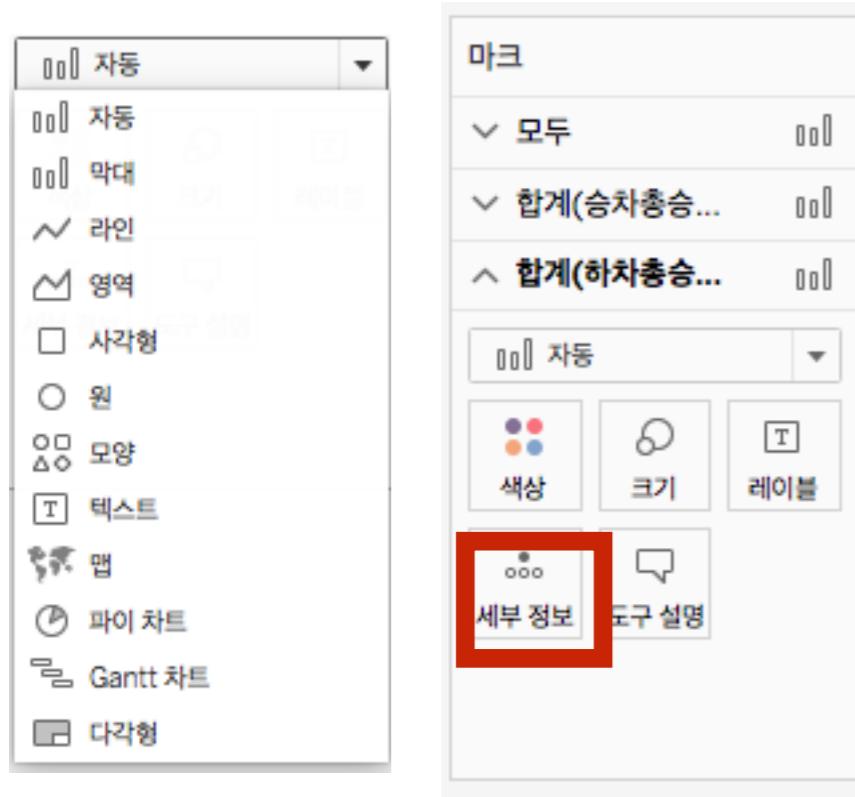
디테일 수정(측정값/표현방식)



카운트 vs 카운트(고유) = “select” vs “select distinct”

“카운트”: 몇개나 있나?, “카운트(고유)”: 몇 종류나 있나?

세부정보 (데이터를 묶고 풀때 사용)



세부정보를 쓰면
“원래는 나눠지지 않는”
정보가 나눠짐

(고급 시각화의 핵심)

2. 데이터 분석

시각화의 2가지 방향[데이터 탐색/분석 정리]

일반 통계
(평균/분포
사분위)

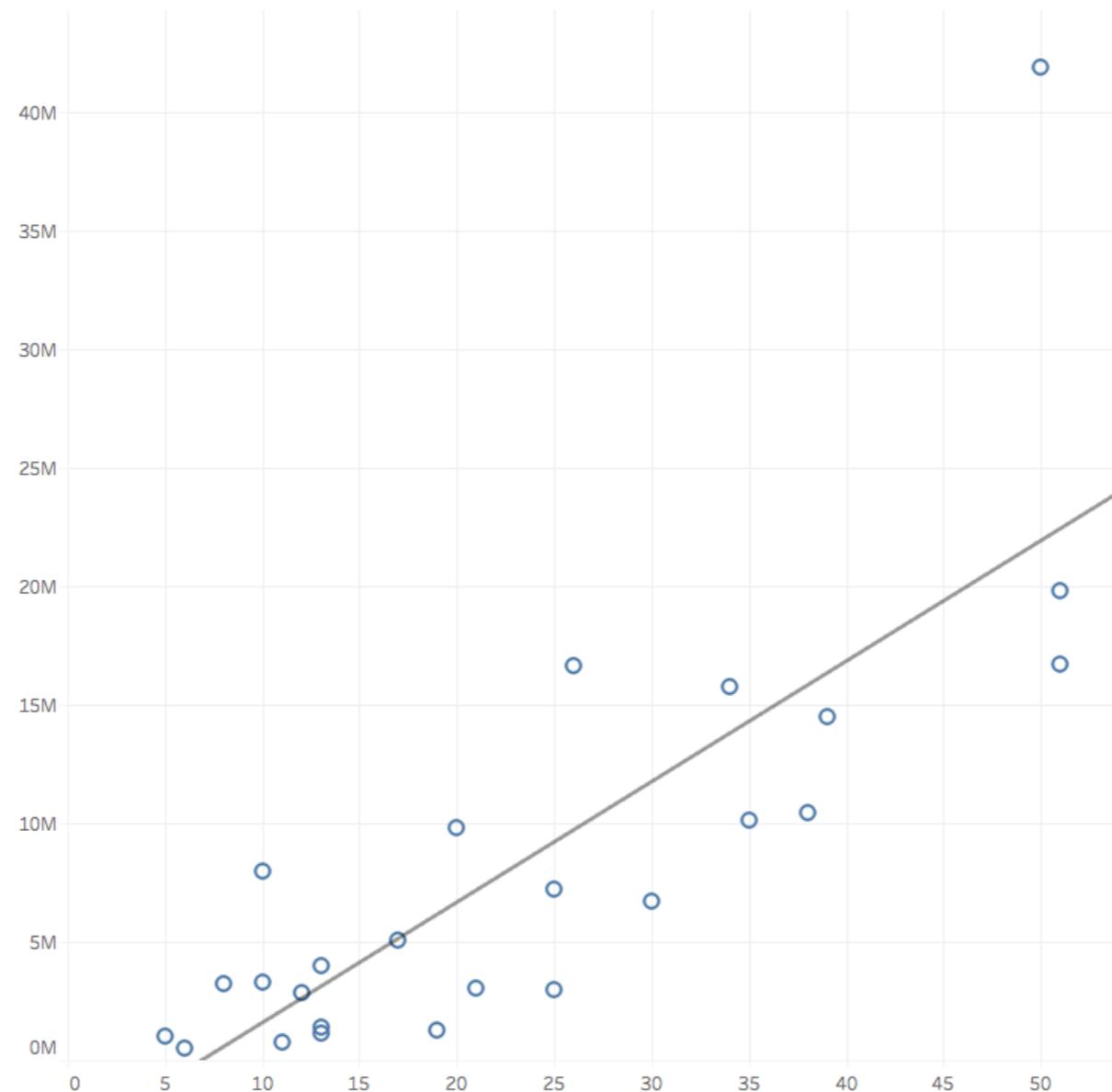
회귀분석
(추세분석)

집단분석
(클러스터링)

시나리오
(What-If?)

- Tableau에서는 분석과 관련해 크게 4가지 기능을 제공
 - + 최근(2019 하반기) : 자연어 검색 기능 추가(Ask Data)
(본 수업에서는 다루지 않음, Desktop에서는 지원불가)

노선별 역수는 이용량과 상관관계가 있을까? [황새의 함정 - 인과와 상관의 위험성]

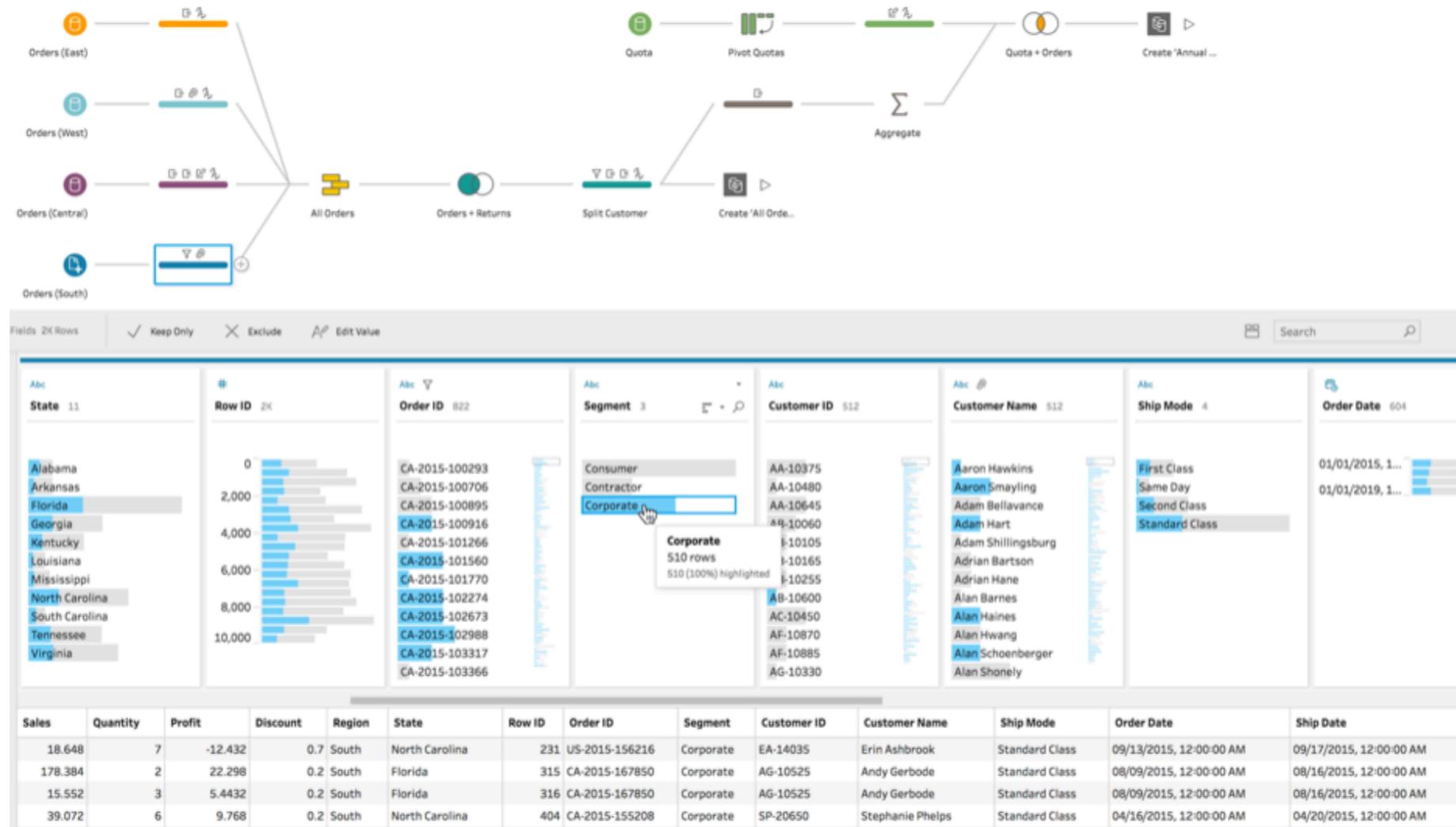


3. 데이터 전처리

왜 중요한가?

데이터 과학(분석) 작업의 80%는 데이터 전처리

TABLEAU PREP?



데이터 전처리 용도로 만들어진 Tableau 연계 프로그램
("대용량", "기업용"에 적합, 본 수업에서는 다루지 않음)

DATA!

weather.go.kr
(기상청, 날씨누리 종합기상정보)

국내 지진 목록



지진 · 화산	관측자료	기후자료	생활과산업
발표정보			
지진	>	국내지진 목록	
지진해일	>	국외지진 목록	

자주 쓰는 전처리 1- 수식 계산

The image shows two screenshots of a data processing tool. On the left, a context menu is open over a column labeled 'Price' with a single data point '0.0'. The menu items include '이름 바꾸기', '값 복사', '숨기기', '계산된 필드 만들기...', '그룹 만들기...', '구간차원 만들기...', '피벗(다중 필드 선택)', and '설명...'. The item '계산된 필드 만들기...' is highlighted. On the right, a formula editor window is displayed with the following code:

```
Name: SortStateField  
Formula:  
CASE [SortBy]  
WHEN 'Profit' THEN SUM([Profit]) * INT([SortOrder])  
WHEN 'Profit Ratio' THEN [Profit Ratio] * INT([SortOrder])  
WHEN 'Sales' THEN SUM([Sales]) * INT([SortOrder])  
END
```

A green checkmark icon at the bottom indicates that the calculation is valid.

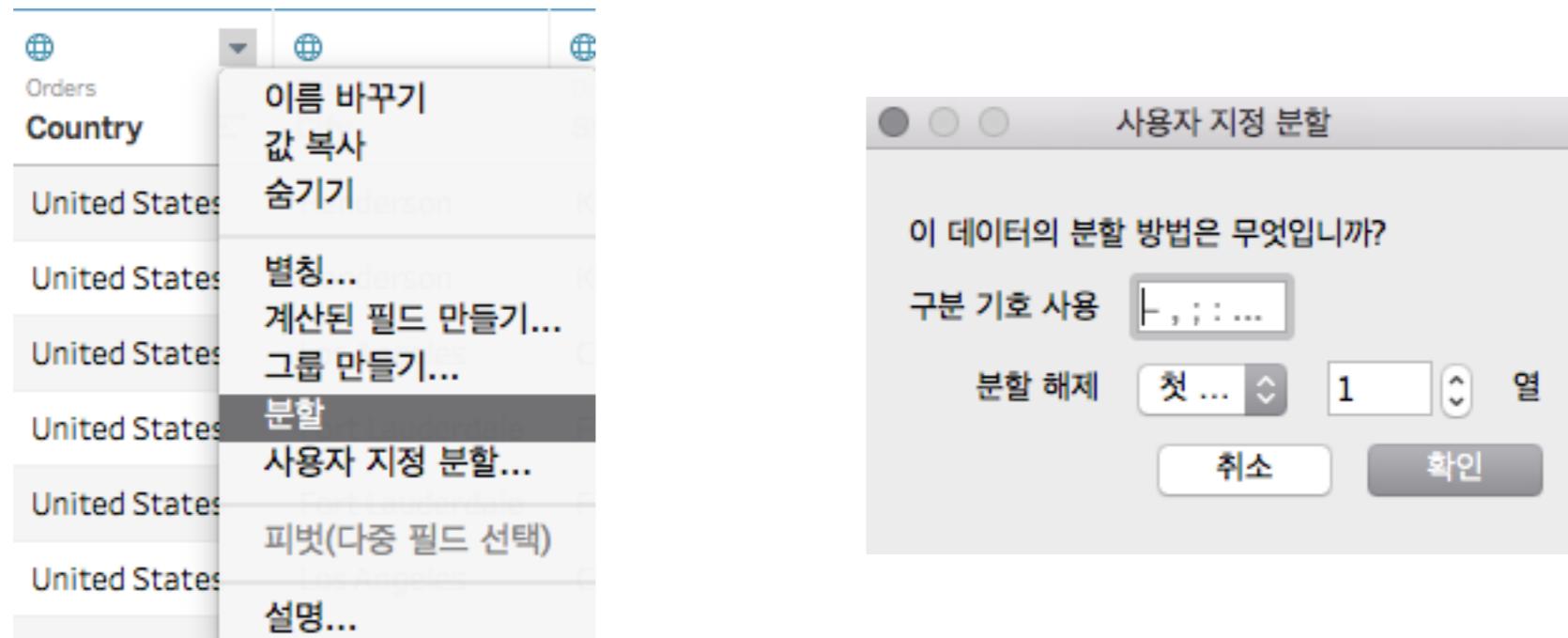
(능력이 된다면) 섬세한 작업도 가능
추천하지는 않음(대부분 Excel, Google Sheets로 처리 가능)
하나라도 “덜” 배우자!

자주 쓰는 전처리 2 - 구간차원 나누기



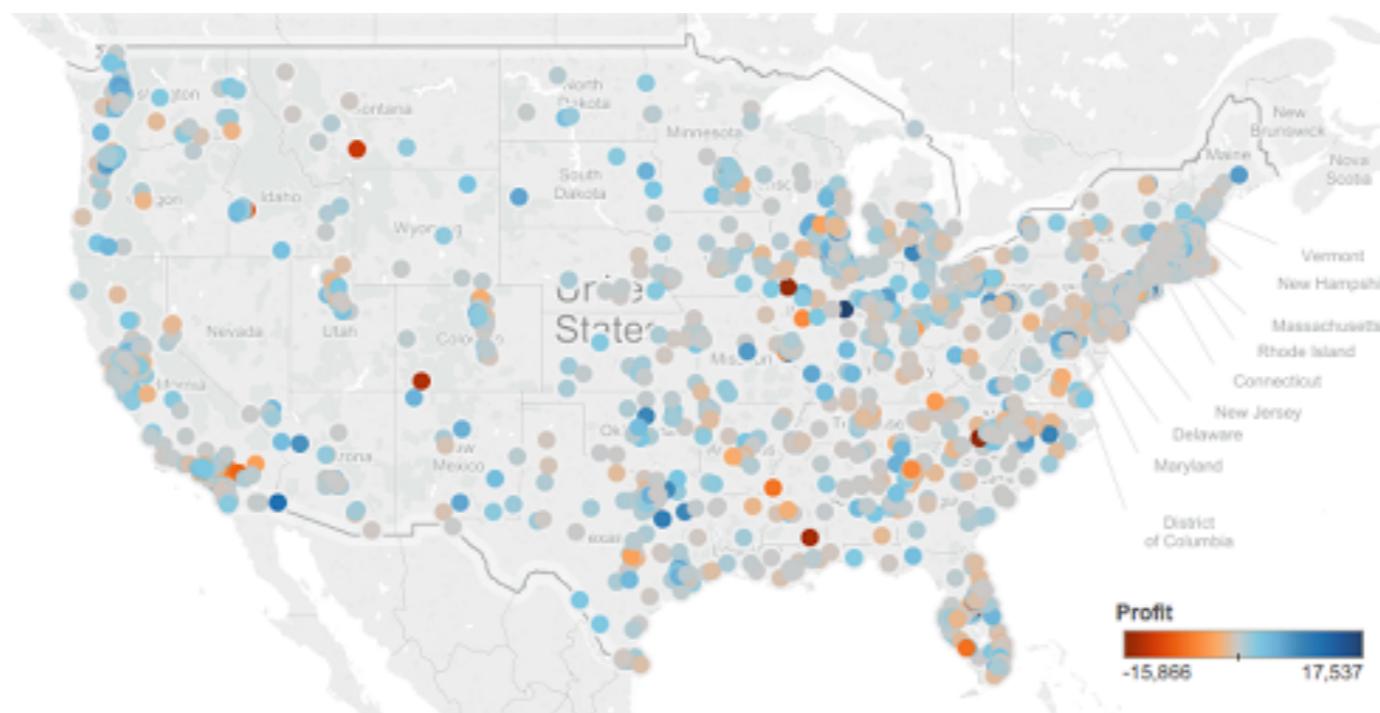
지정된 범위로 데이터 나눠주기
연속형 데이터를 구간별로 나눠서 확인(e.g. 연령대)
“별칭”을 쓰면 그룹 이름을 내가 만들 수 있음

자주 쓰는 전처리 3 - 셀 나누기



특정 문자열 기준으로 나눠주기
년/월/일, 단위, ID 다양한 형태로 사용 가능

MAPPING

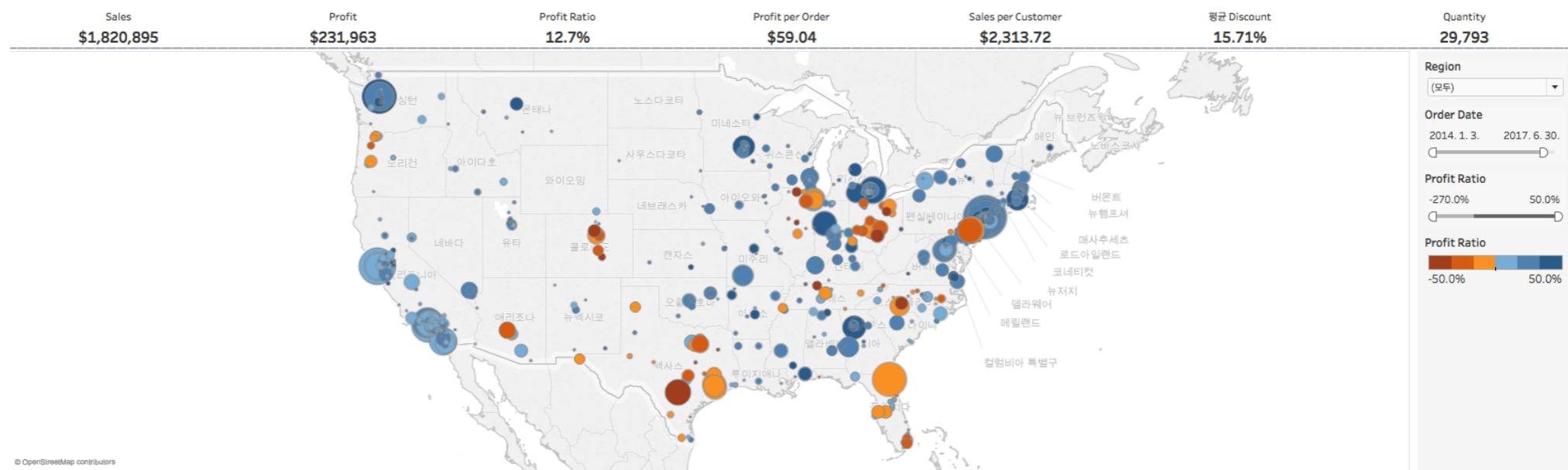


Tableau에는 지리 데이터를 표현하는 2가지 방법 존재

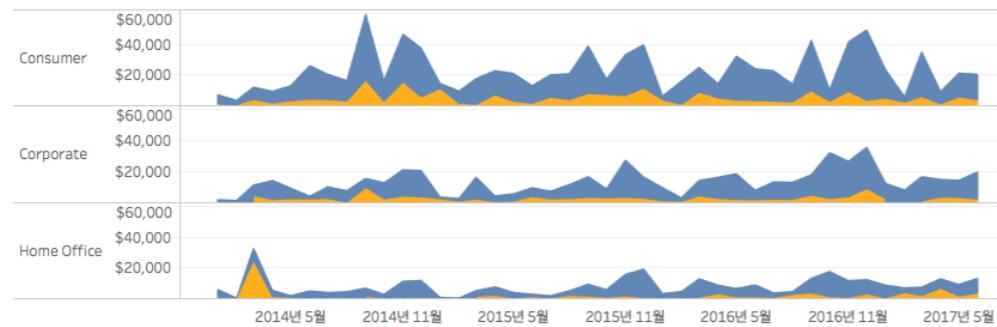
4. 인터랙티브 스토리 텔링(모아보기)

분석 결과 모아보기

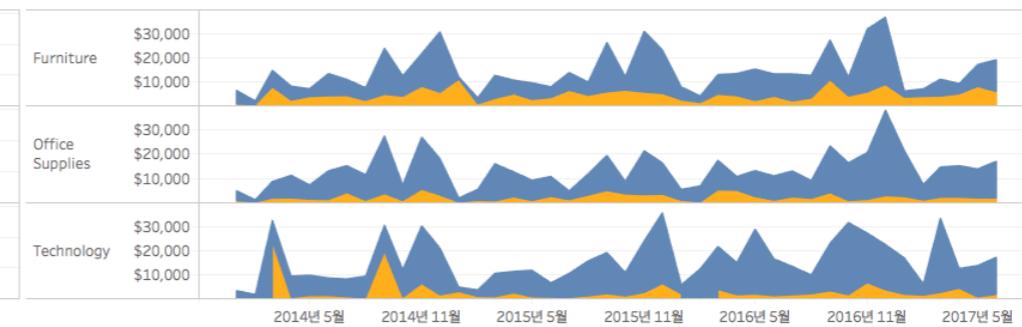
Executive Overview - Profitability (모두)



Monthly Sales by Segment - States: 모두



Monthly Sales by Product Category - States: 모두



각 그래프는 유기적으로 연동 / 필터 설정 가능

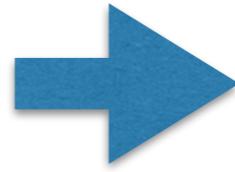
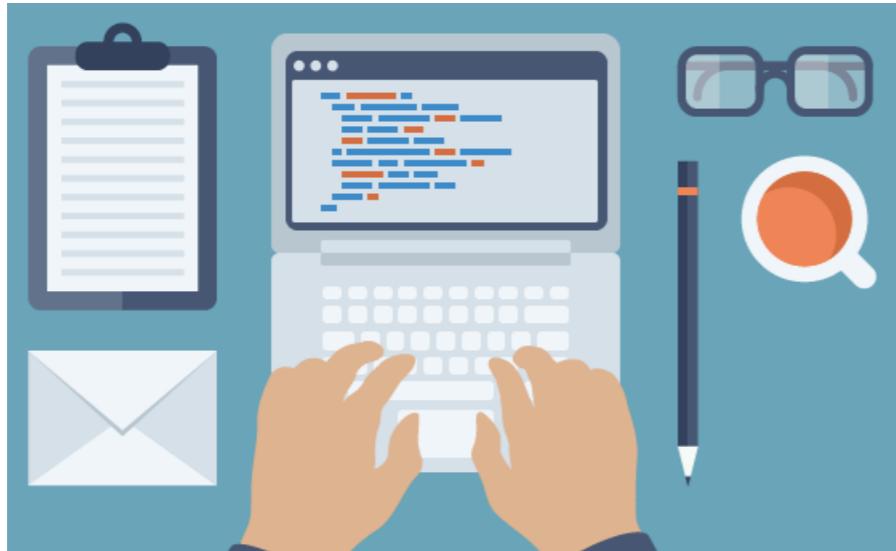
분석 결과 모아보기



워크시트 -> 대시보드 -> 스토리(데이터연동)

5. 온라인에 게시하기

주의!



시각화를 공유하기 전에 자료를 검토해야 합니다
-> 연구목적으로만 사용, 공개 불가 등 제한조건 고려
(Tableau 무료 서버 사용시 자동으로 “자료공유”)

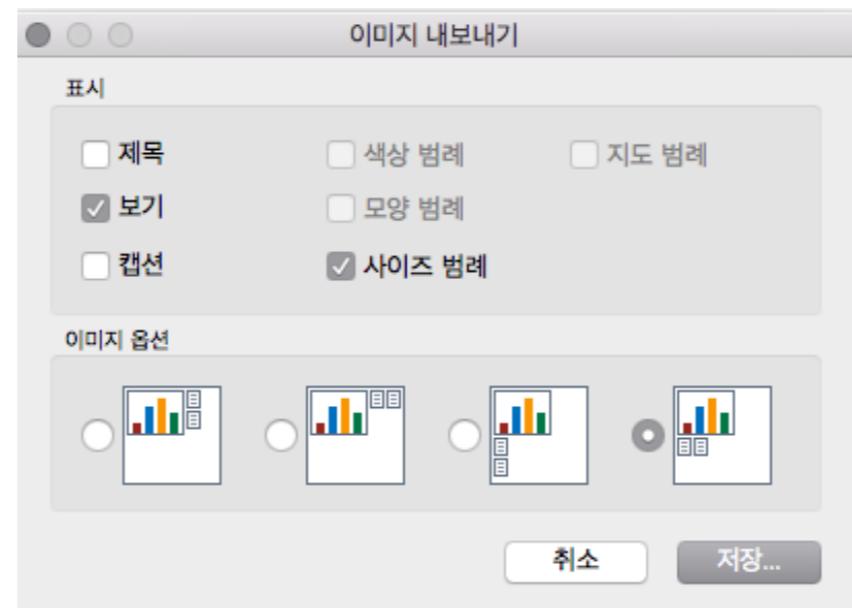
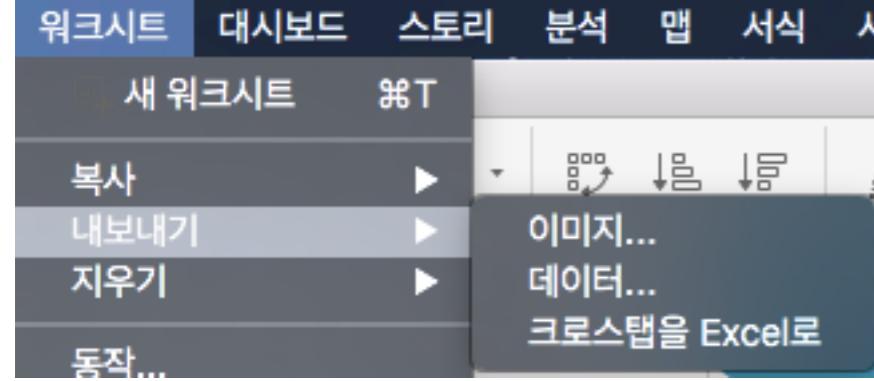
TABLEAU의 공유방법

스냅샷 - 내보내기 시점 그대로(Pdf, Image 형태, 리더 불필요)

Tableau문서(파일) - 데이터 갱신에 따라 변화(**Reader 필요**)
데이터 원본을 포함할지 안 할지 선택 가능(그래도 일부 데이터는 공개)

Tableau문서(온라인) - 데이터 갱신에 따라 변화(웹페이지 형태, **Reader 불필요**)
데이터 원본을 포함할지 안 할지 선택 가능(그래도 일부 데이터는 공개)

스냅샷



워크시트(대시보드) -> 내보내기 -> 옵션 중 선택
파일 -> PDF로 인쇄(통합인쇄) 가능
(+ 2019 PPTX로 내보내기 기능 추가)

TABLEAU문서 공유



twb파일 -> 핵심 요소만 추출(간략하게, 추가 수정 어려움)

twbx -> 모든 요소 포함(데이터셋도 포함, 수정가능, 비공개 데이터일 경우 주의!)

TABLEAU 온라인 공유 (PUBLIC 계정 OR 회사서버 사용)

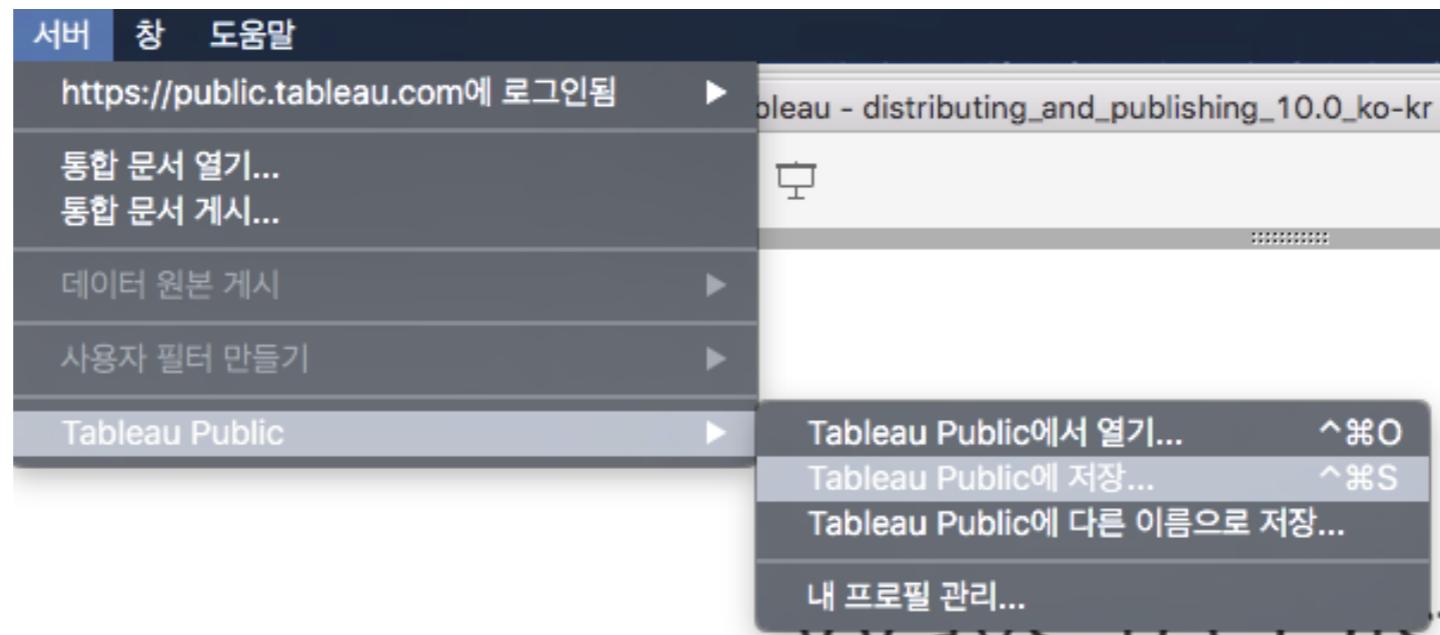


Tableau Public에서는 저장/내보내기가 무조건 서버 상에서만 가능
온라인에 저장하기 전 데이터 추출해야 함(내 컴퓨터에서 업로드가 이뤄지기 때문)
[데이터 추출은 데이터 -> 원하는 데이터 선택 -> 추출로 가능]

TABLEAU문서

The screenshot shows a user profile on Tableau Public. At the top, there's a dark blue header with the Tableau Public logo. Below it is a large circular placeholder for a profile picture. To the right of the placeholder, the name "Lee Gyuho" is displayed, followed by a pencil icon and the word "편집" (Edit). Underneath the name, it says "1 Viz". On the far left, there's a small thumbnail labeled "Viz 1". At the bottom, there are three buttons: "Favorite" with a star icon, "세부 정보 편집" (Edit details) with a pencil icon, and "통합 문서 다운로드" (Download document) with a download icon.

제목

2007-2017데이터 모음

다른 사용자가 Tableau Public에서 Viz를 검색할 때 그냥 지나치지 않도록 눈에 띄는 제목을 주세요.

고정 링크

URL 추가

Viz를 내장해 보십시오. 사이트로 트래픽을 유도하는 탁월한 방법입니다.

설명

testing_02

검색 결과에서 상위에 표시되려면 정확한 설명이 필요합니다.

툴바 설정

- 뷰 컨트롤 표시 실행 취소, 다시 실행, 되돌리기
- 작성자 프로필 링크 표시
- 통합 문서 및 해당 데이터를 다른 사용자가 다운로드할 수 있도록 허용

기타 설정

- 탭으로 통합 문서 시트 표시

public.tableau.com에 로그인 하면 게시한 자료 확인, 공개설정 가능

문서 공유하기



public.tableau.com로그인 한 후, 게시한 자료 선택 -> 오른쪽 아래 공유 URL 얻기