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?

Business Intelligence Software

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

왜 TABLEAU인가?

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

하지만 Tableau는 **강력한 인터렉티브 시각화를 손쉽게 구현할 수 있습니다**

TABLEAU FOR FREE?

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

TABLEAU학습 준비물

tableau public

Tableau Desktop(체험판)
Tableau/Public 계정

Google

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.

Data should be in a spreadsheet format with rows and columns

Each column represents one attribute of the data

Data is row level, one row represents one item

	A	B	C	D	E	F
1	Row ID	Order ID	Order Date	Order Priority	Sales	Ship Date
2				High	$221.98	11/22/16
3				Critical	$3,709.40	2/17/16
4				Medium	$5,175.17	10/27/16
5	13524	ESKM1637548	2/7/16	Medium	$2,892.51	2/9/16
6	47221	SGRH9495111	11/14/16	Critical		
7	22732	INJM156557	7/7/16	Critical		
8	30570	INTS2134092	11/16/14	Critical		
9	31192	INMB1808592	4/24/15	High	$5,244.84	4/28/15
10	40099	CAAB10015140	11/20/16	High	$341.96	11/22/16
11	36258	CAAB10015140	3/16/14	High	$48.71	3/17/14
12	36259			High	$17.94	3/17/14
13	28879			High	$4,626.15	5/2/15
14	45794			Critical	$2,616.96	1/7/15
15	4132			Critical	$2,221.80	11/23/15
16	27704	INPF1912027	6/15/16	Critical	$3,701.52	6/17/16

Tableau에서는 주로 Tidy Data를 사용합니다

TIDY DATA?

Hadley Wickham
(d/plyr, 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…)

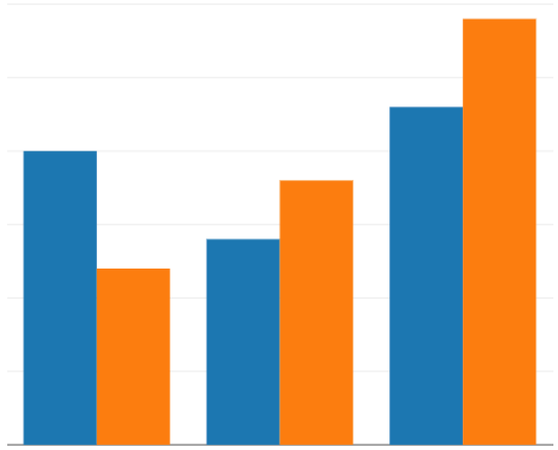
TABLEAU와 TIDY DATA의 관계는?

특징

인원

	측정값		

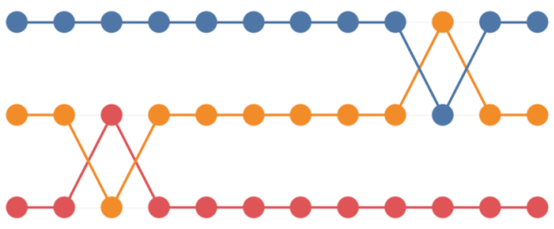
특징을 섞고, 측정값을 넣으면
시각화가 됩니다



세로막대차트

특징 : 남녀, 고/저학년

표현값 : 성적



순위차트

특징 : 팀 종류/시간(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

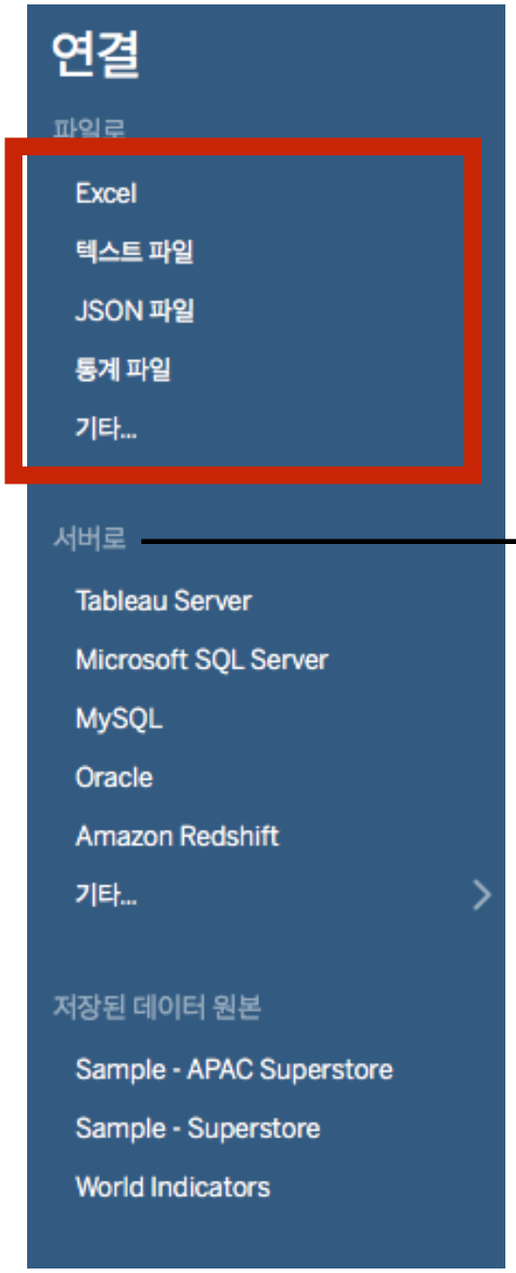
하지만, 이런 데이터도 “전처리”를 잘 하면 쉽게 분석할 수 있습니다
(다음 시간에…)

1. 가볍게 시작하기

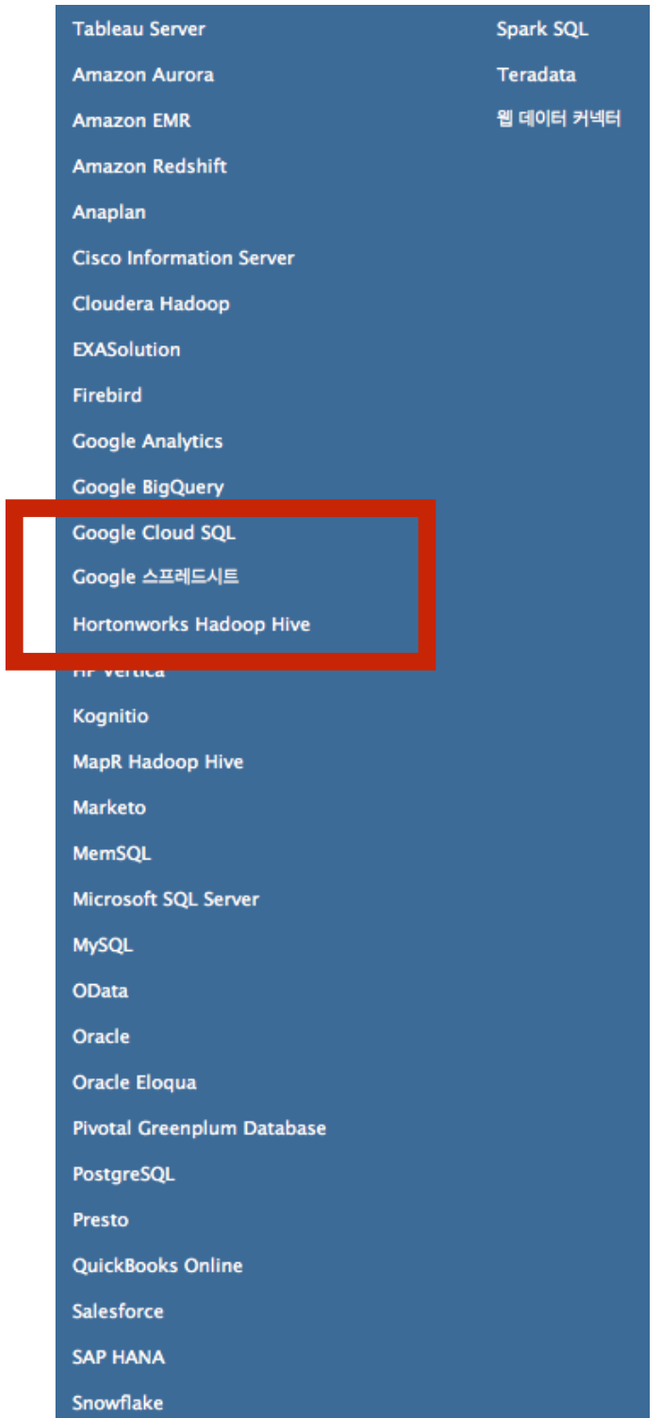
DATA!

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

연결(2종류 존재)

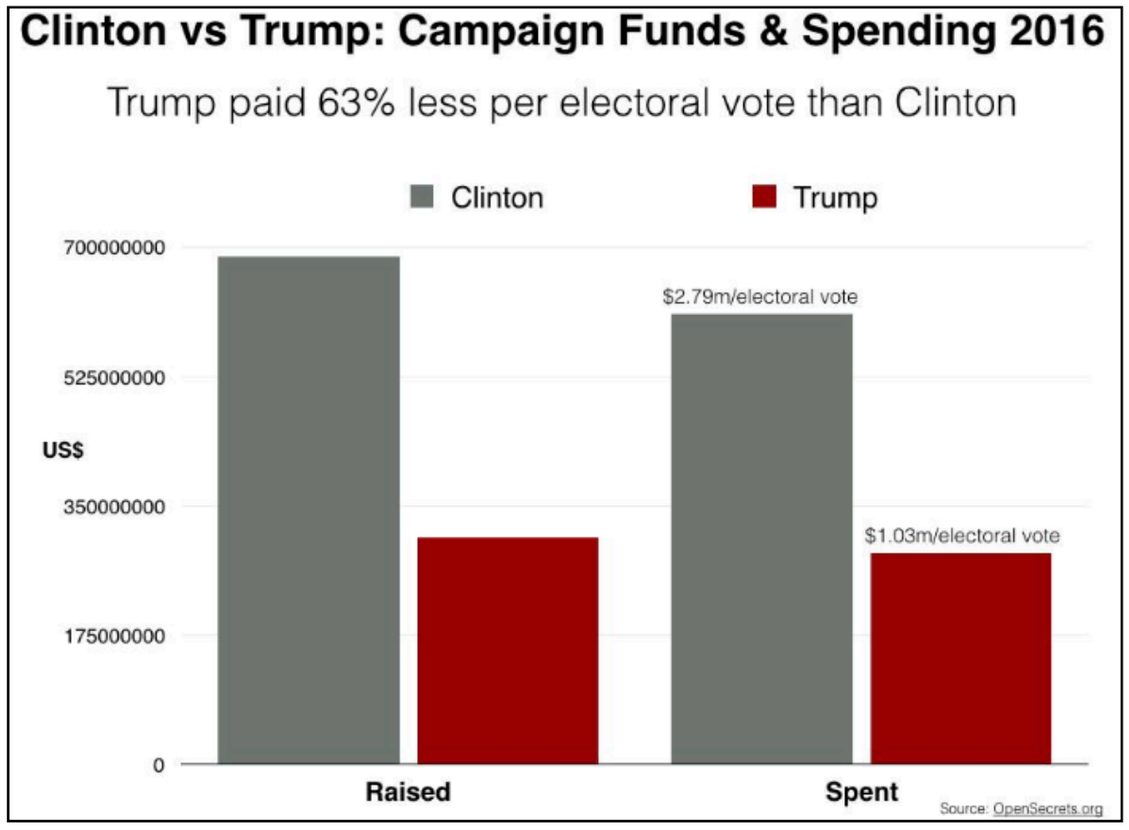


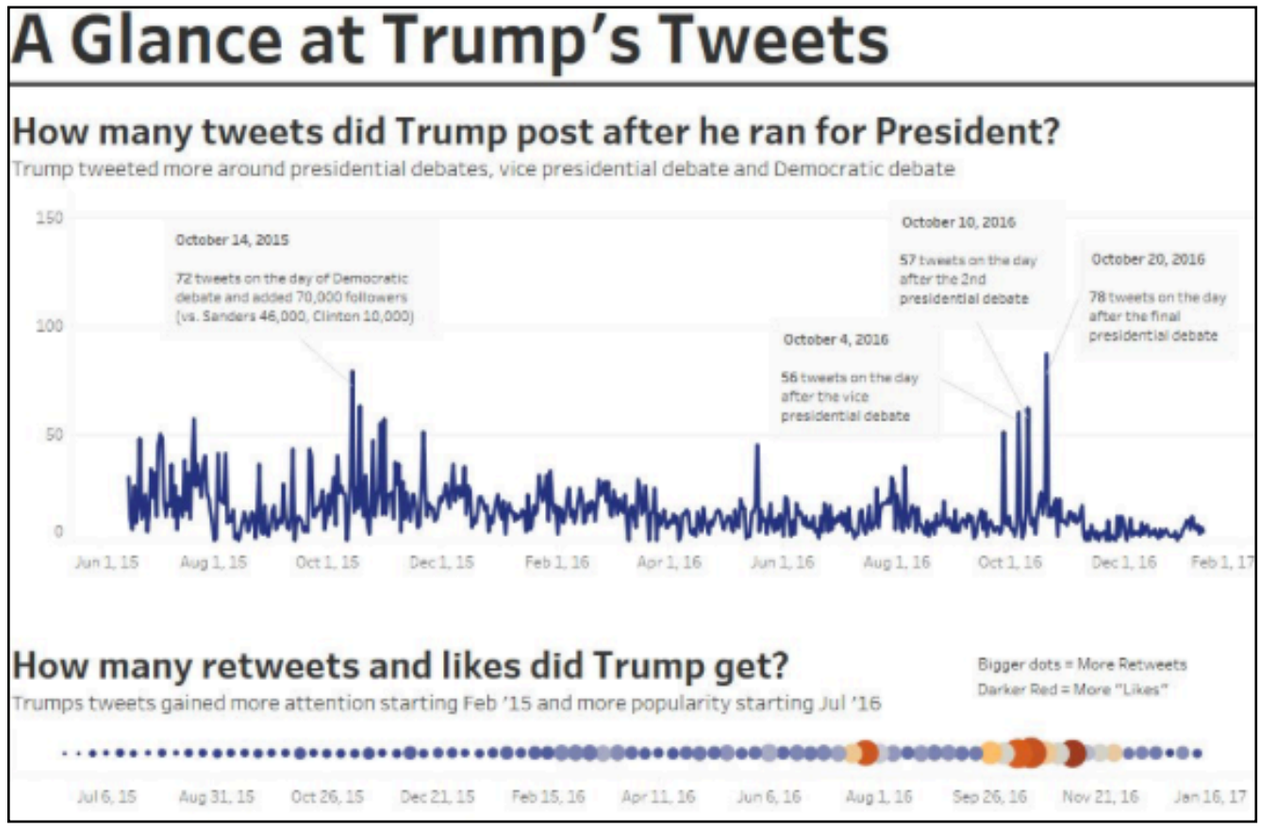
파일에서 가져오기



온라인에서 가져오기

추출? 라이브?





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

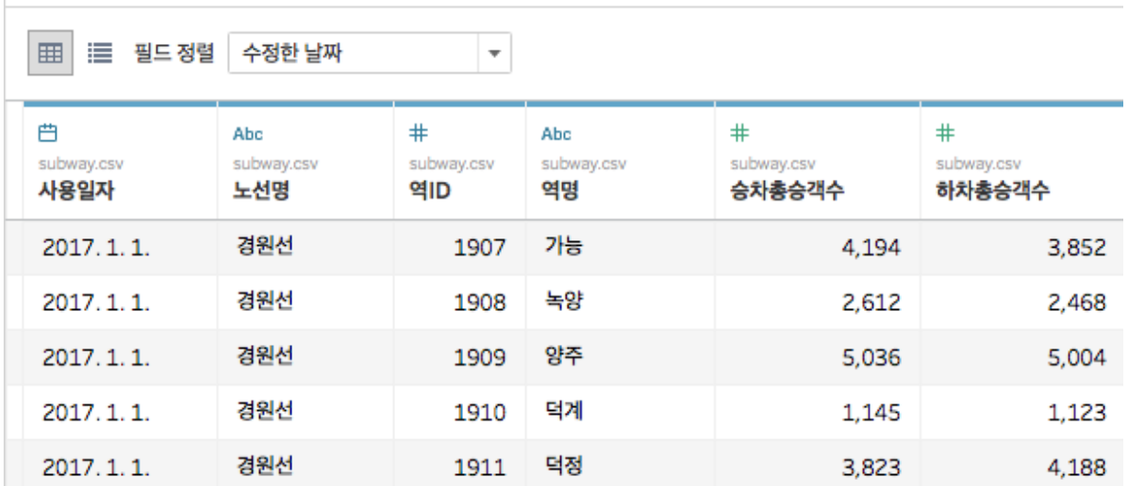
서울시 지하철호선별 역별 승하차 인원 정보

Sheet | Open API | File

너울시 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.csv 사용일자	subway.csv 노선명	subway.csv 역ID	subway.csv 역명	subway.csv 승차총승객수	subway.csv 하차총승객수
2017. 1. 1.	경원선	1907	가능	4,194	3,852
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)

데이터 편집/속성변경후 조합

문자, 색의 차이는?
앞의 기호(데이터 형태)/색상(데이터 연속성)
서로 연결된 데이터 형식은? 그룹화/계층화

차트 작성하기(2)

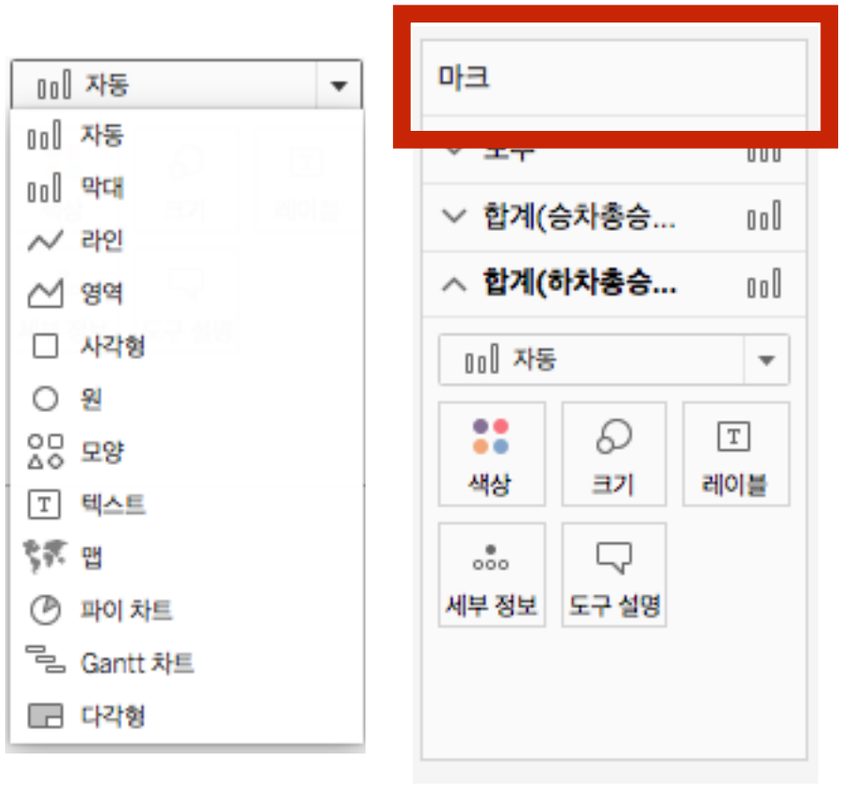
요인 선택 후 오른쪽 위 “표현방식” 클릭
Tableau에서는 자동으로 추천 그래프 작성
(직접클릭/더블클릭)

디테일 수정

"I am so over the details."

차원 측정값 표현형식

디테일 수정(측정값/표현방식)

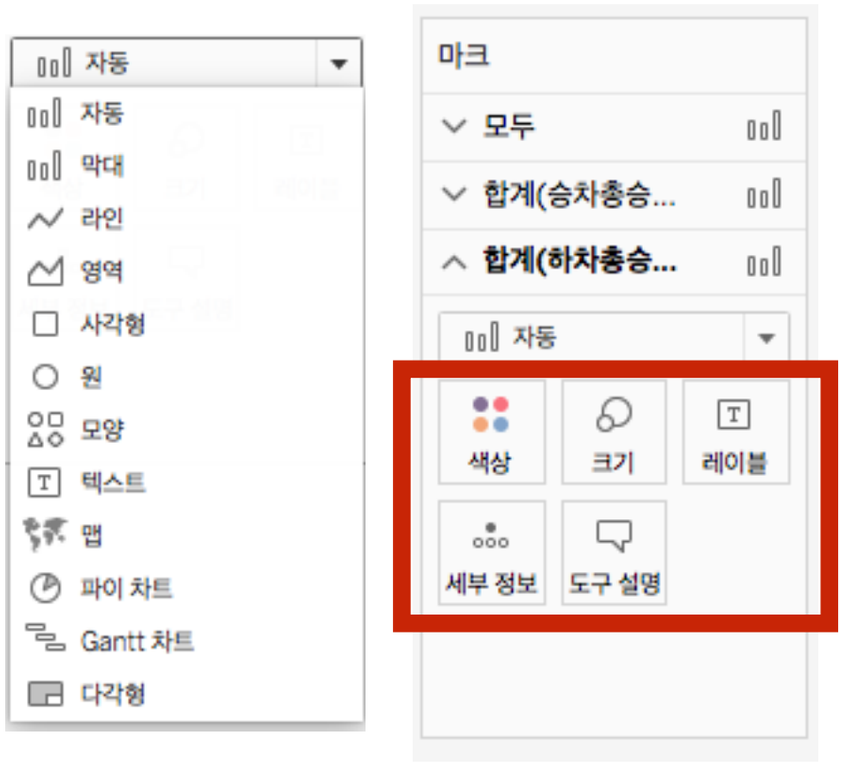


마크(표현 방식 세부 조절 가능)

마크의 조합을 통해 “복잡한” 그래프 생성 가능

새로운 측정값이 추가될때마다 Stack

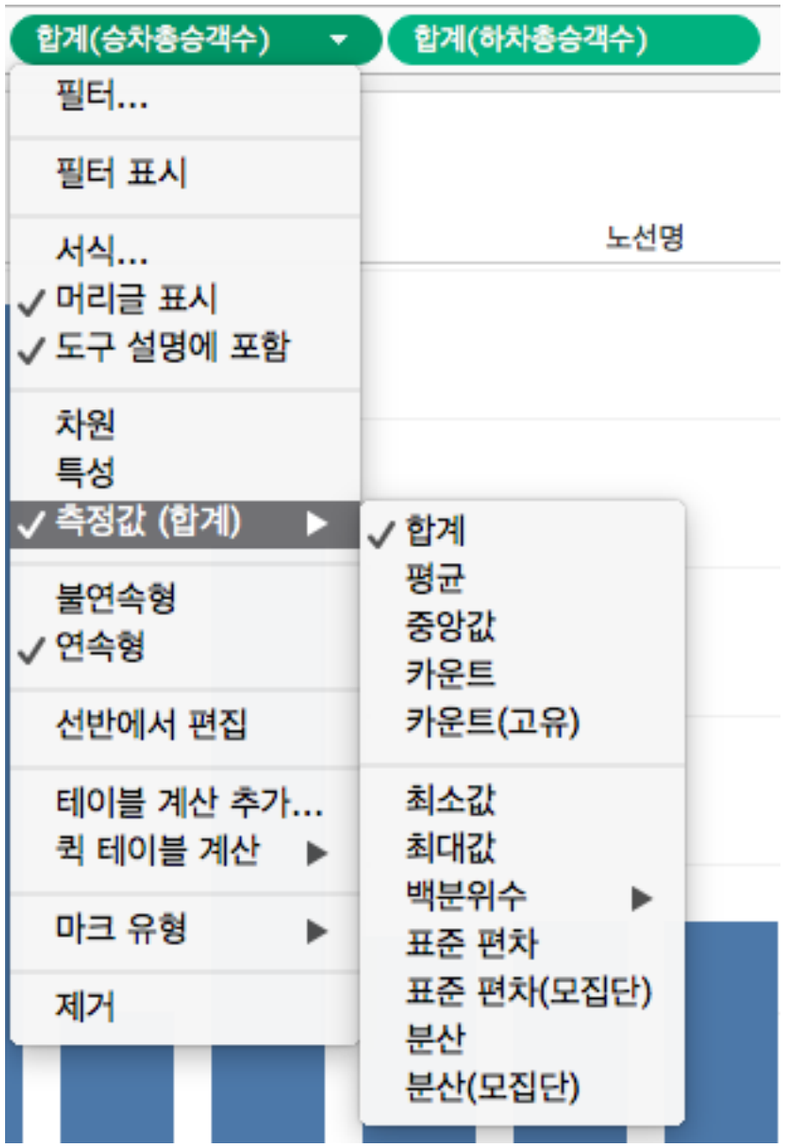
디테일 수정(측정값/표현방식)



세부옵션

(세밀한 조절은
아래 Box의 조합으로 가능)

디테일 수정(측정값/표현방식)



내가 원하는 형태의 측정값을 사용 가능

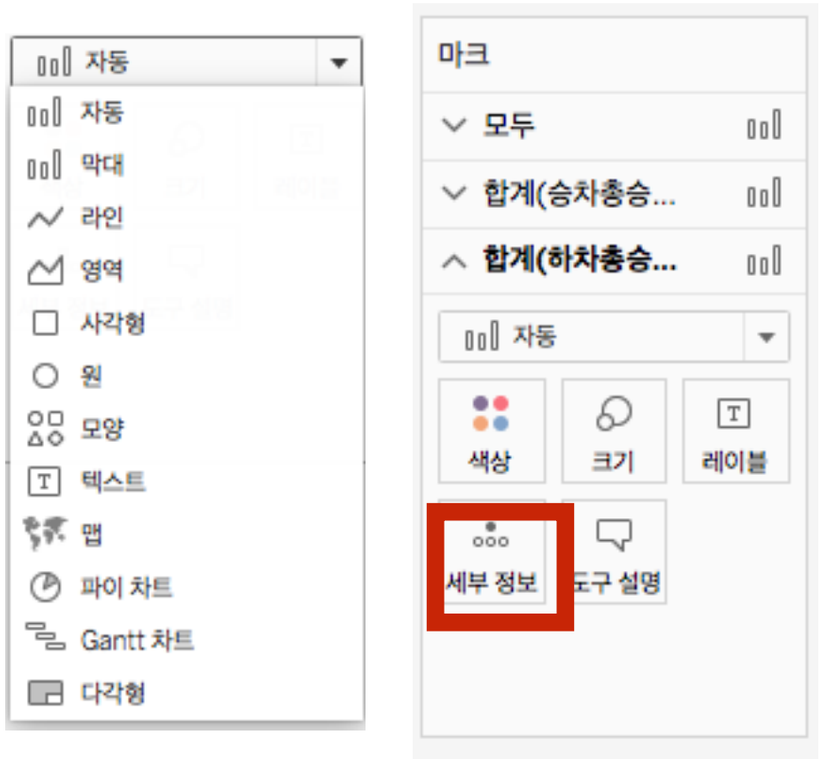
비지니스 분석에서 빈번하게 쓰는 기능은
(퀵 테이블 계산)에서 사용 가능

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)

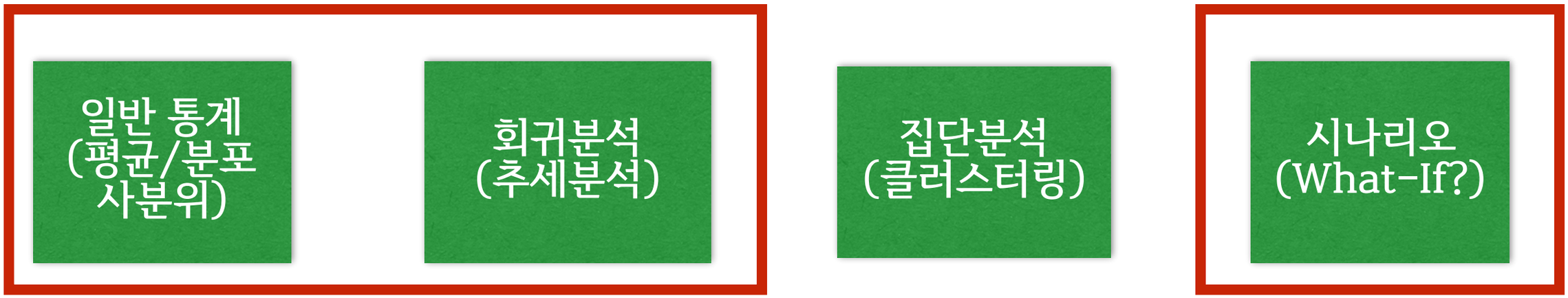


세부정보를 쓰면
“원래는 나눠지지 않는”
정보가 나눠짐

(고급 시각화의 핵심)

2. 데이터 분석

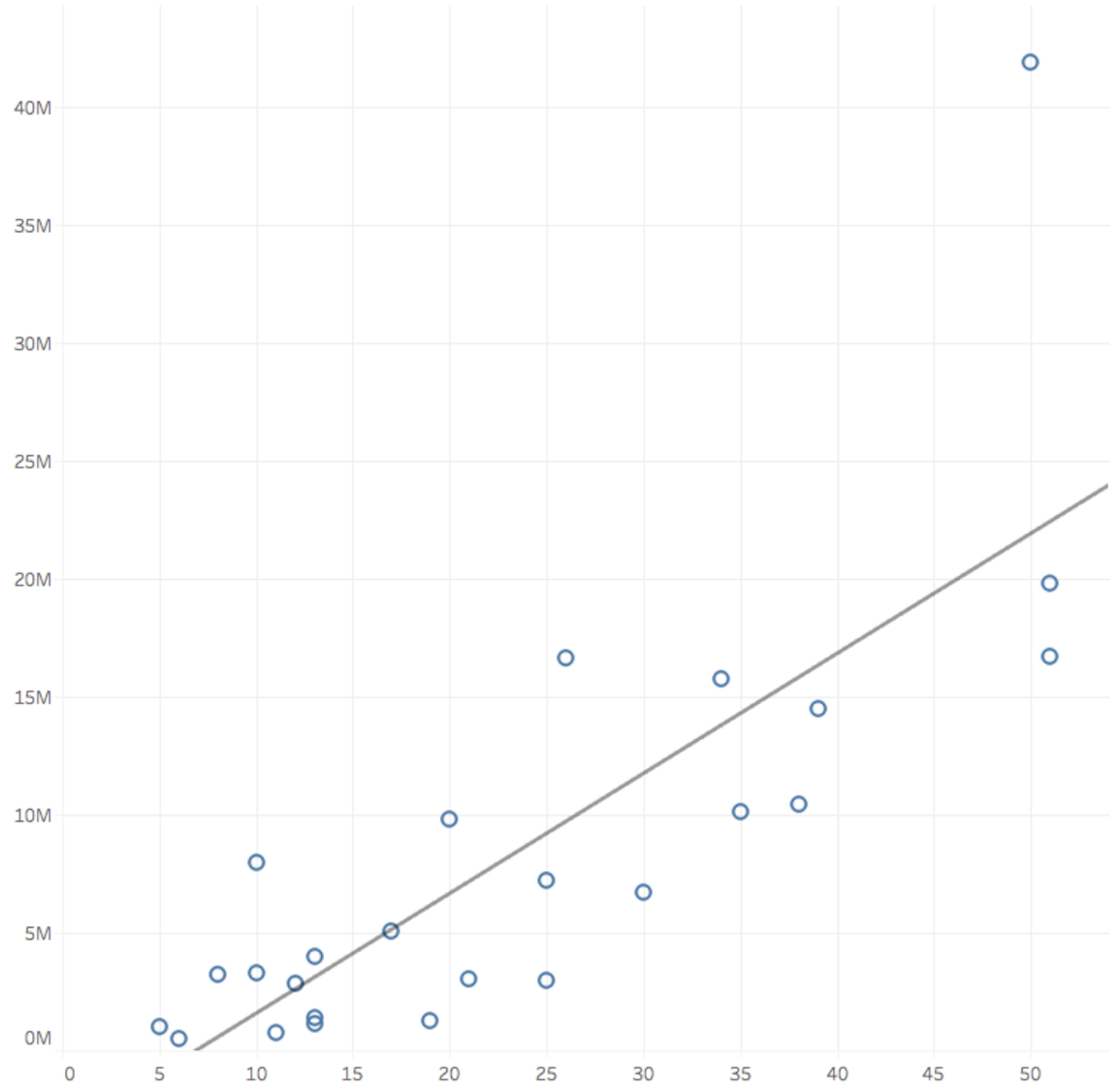
시각화의 2가지 방향(데이터 탐색/분석 정리)



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)

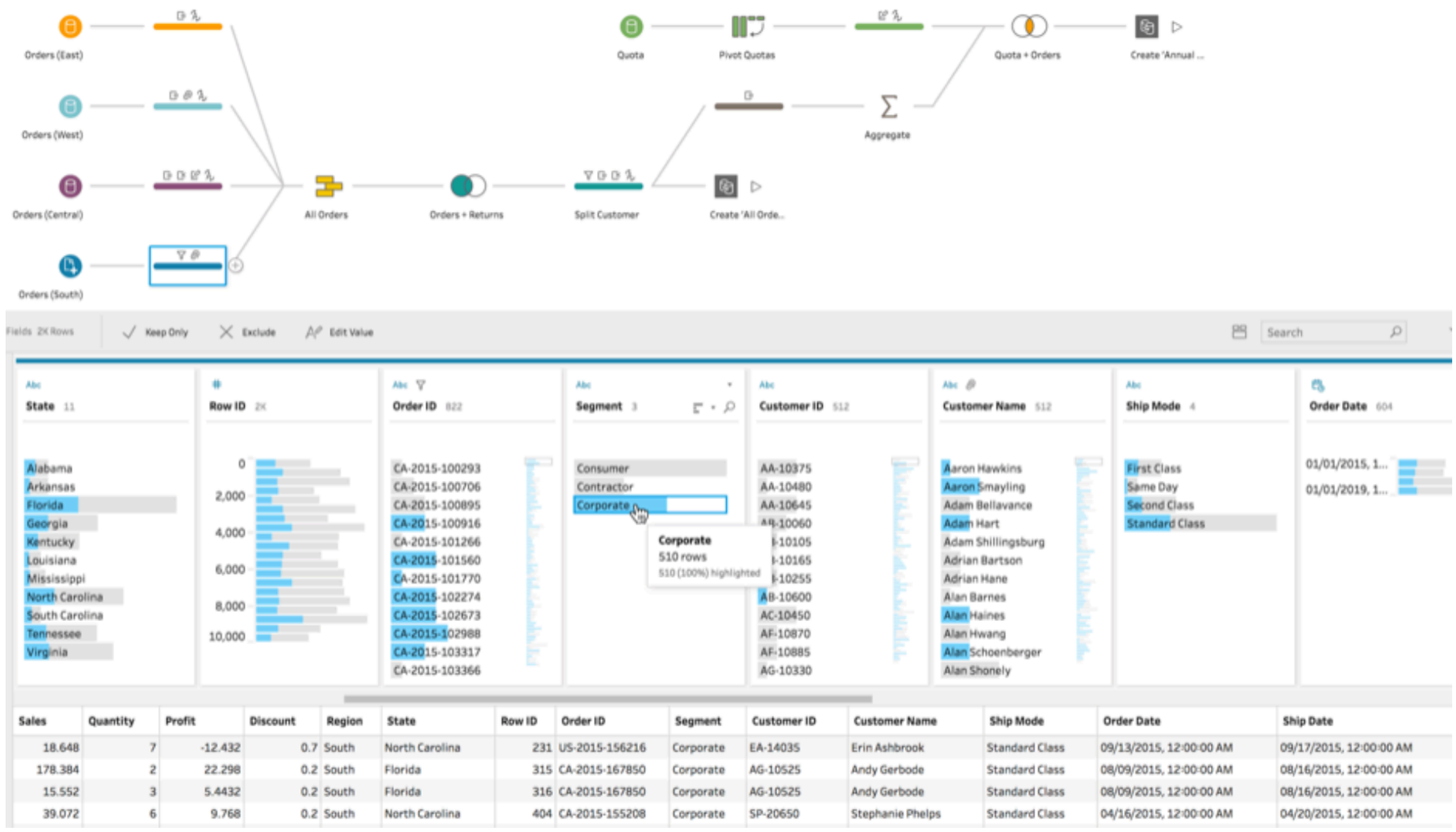


3. 데이터 전처리

왜 중요한가?

데이터 과학(분석) 작업의 80%는 데이터 전처리

TABLEAU PREP?

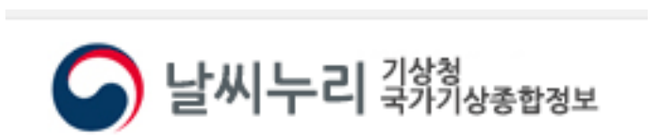


데이터 전처리 용도로 만들어진 Tableau 연계 프로그램
(“대용량”, “기업용”에 적합, 본 수업에서는 다루지 않음)

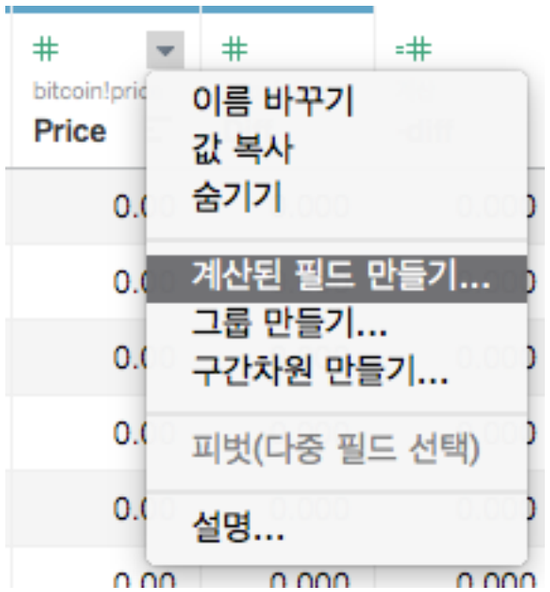
DATA!

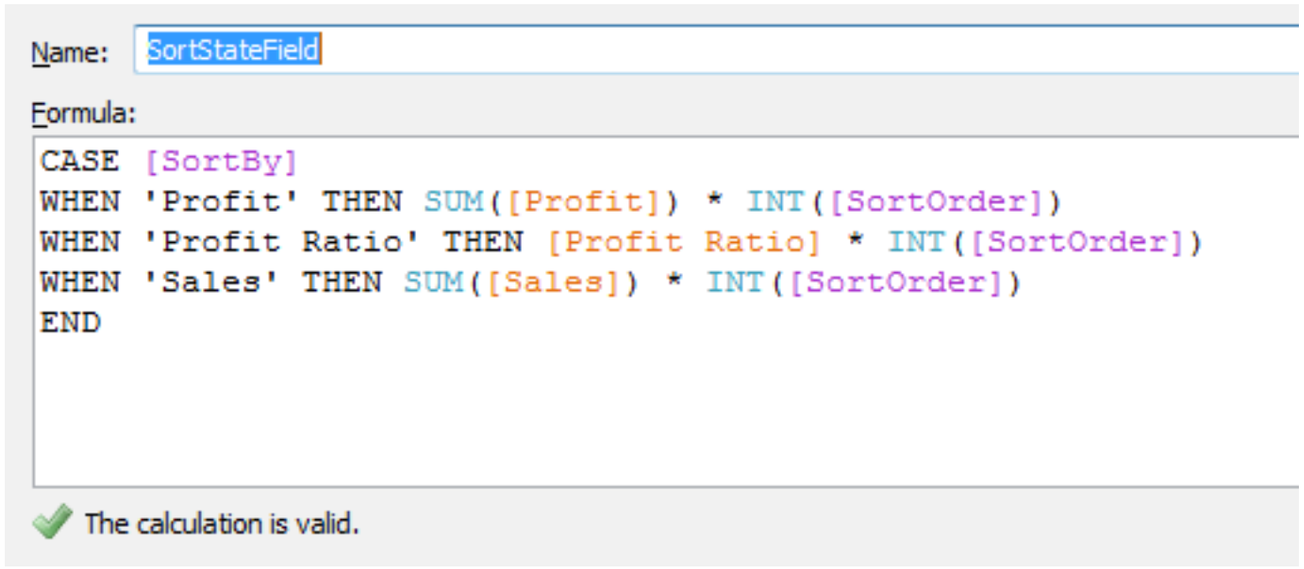
weather.go.kr
(기상청, 날씨누리 종합기상정보)

국내 지진 목록



자주 쓰는 전처리 1 - 수식 계산



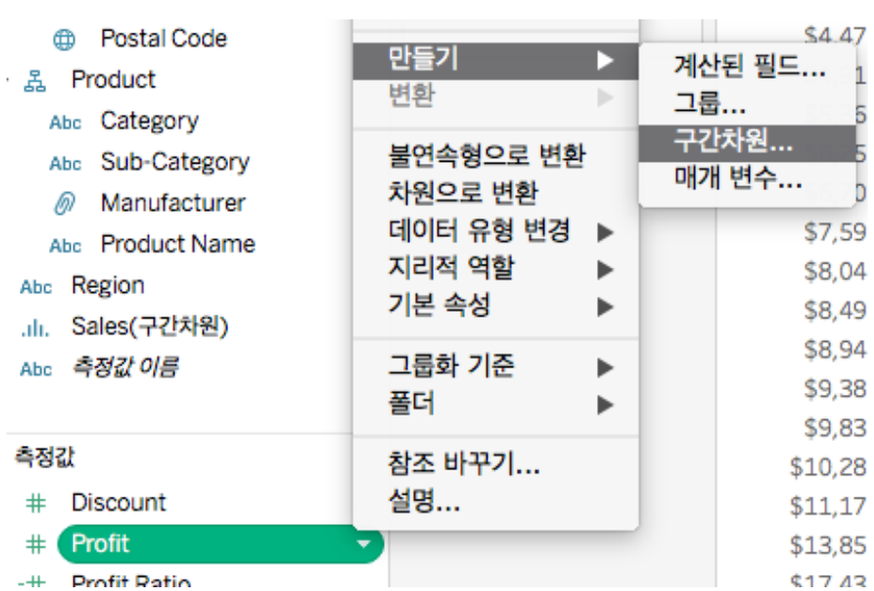


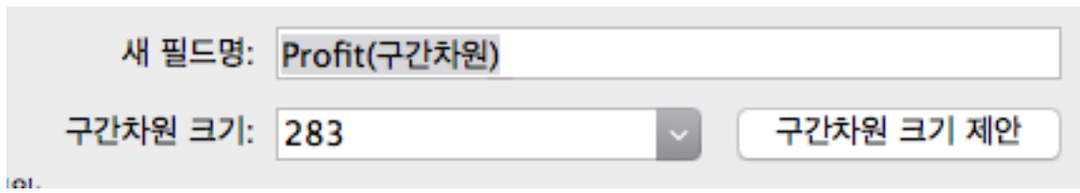
(능력이 된다면) 섬세한 작업도 가능

추천하지는 않음(대부분 Excel, Google Sheets로 처리 가능)

하나라도 “덜” 배우자!

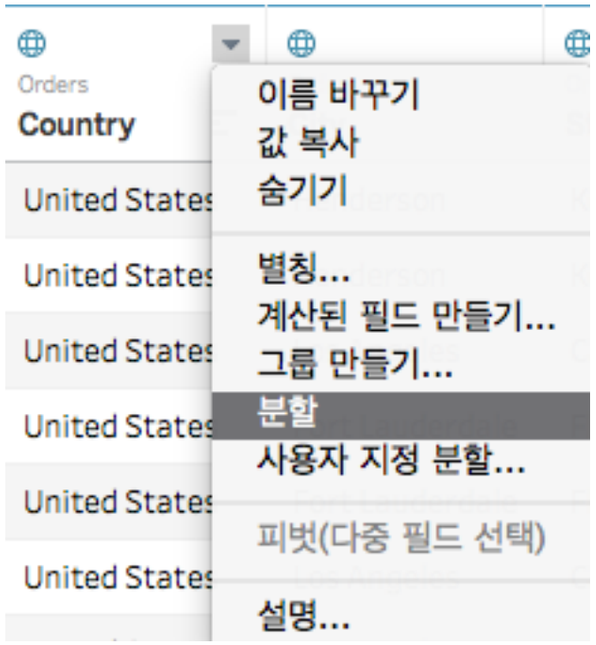
자주 쓰는 전처리 2 - 구간차원 나누기

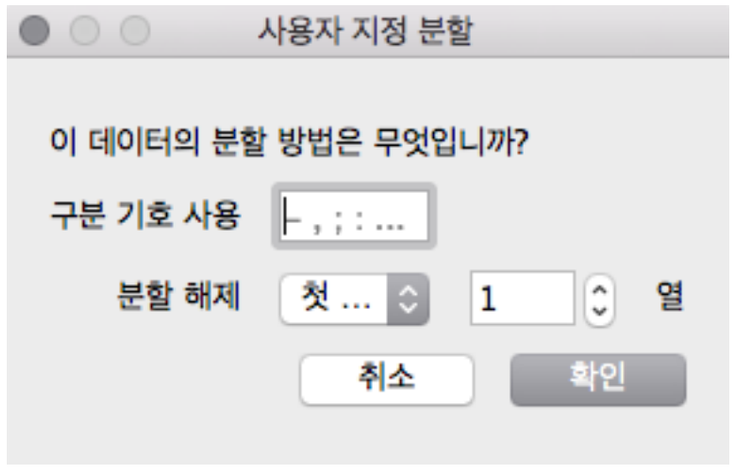




지정된 범위로 데이터 나눠주기
연속형 데이터를 구간별로 나눠서 확인(e.g.연령대)
“별칭”을 쓰면 그룹 이름을 내가 만들 수 있음

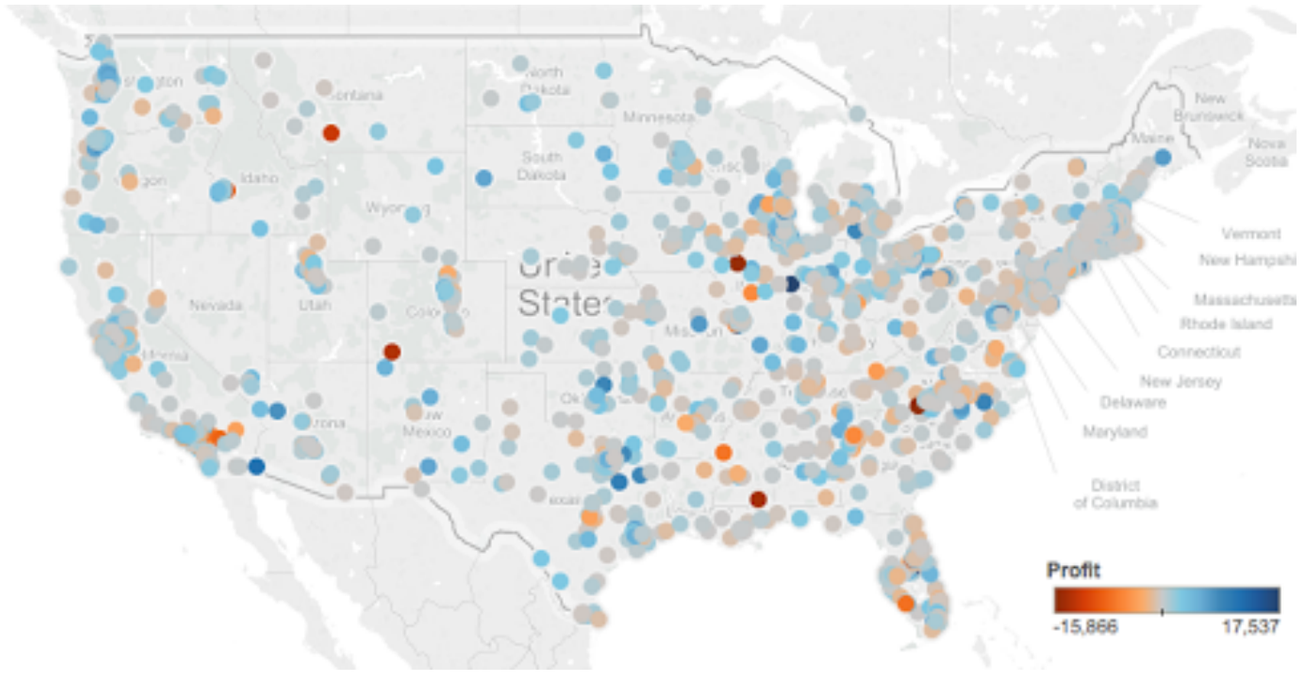
자주 쓰는 전처리 3 - 셀 나누기





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

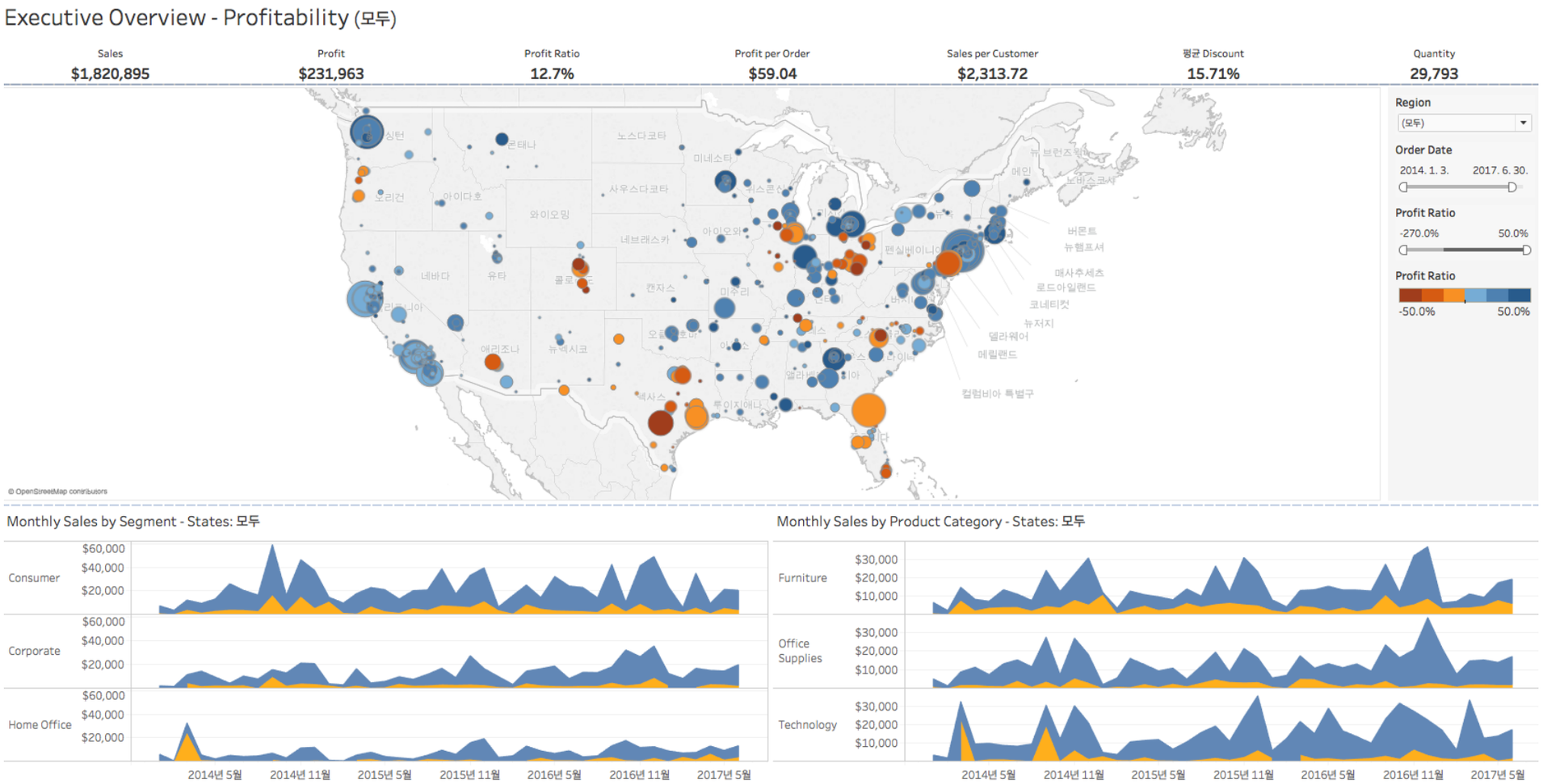
MAPPING



Tableau에는 지리 데이터를 표현하는 2가지 방법 존재

4. 인터렉티브 스토리 텔링(모아보기)

분석 결과 모아보기



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)

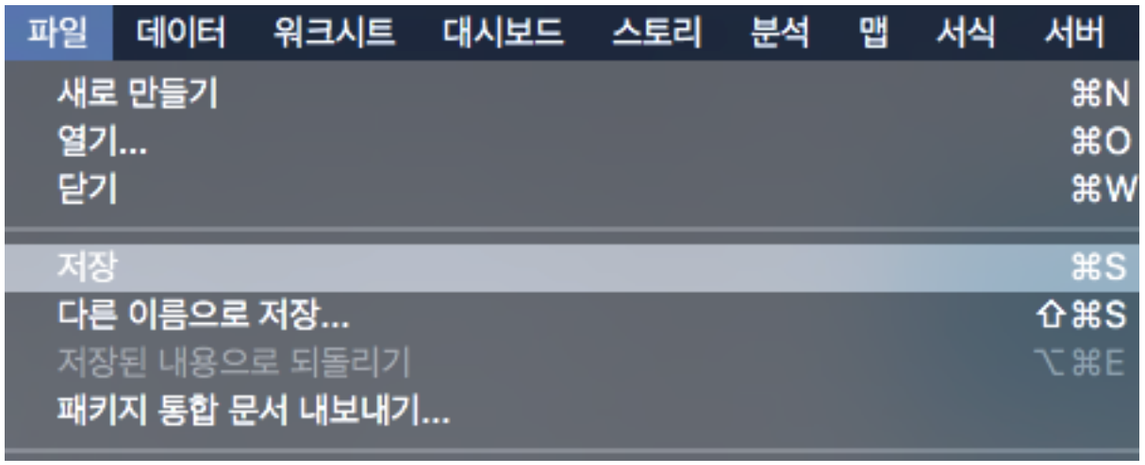
스냅샷

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)

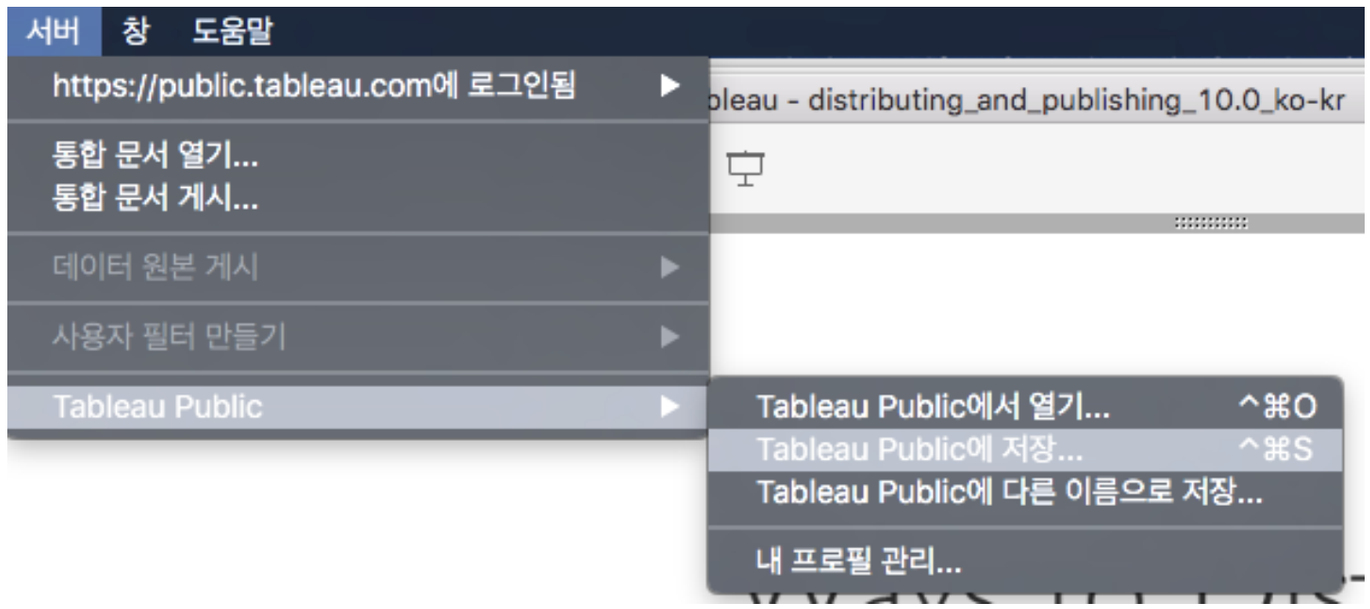
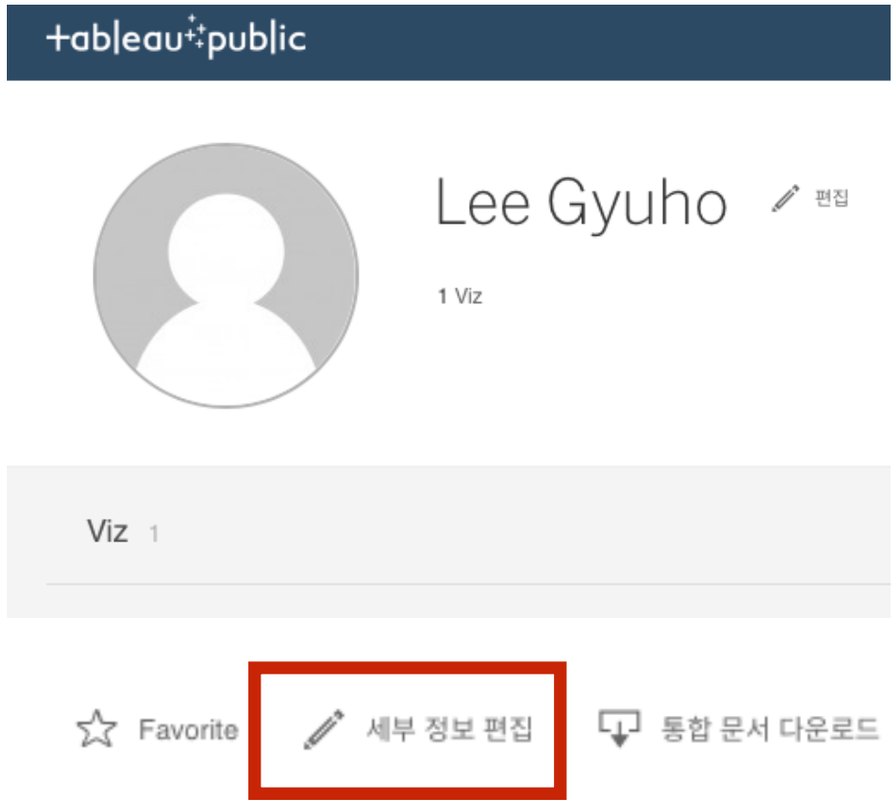


Tableau Public에서는 저장/내보내기가 무조건 서버 상에서만 가능

온라인에 저장하기 전 데이터 추출해야 함(내 컴퓨터에서 업로드가 이뤄지기 때문)

[데이터 추출은 데이터 -> 원하는 데이터 선택 -> 추출로 가능]

TABLEAU문서



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