

데이터 과학 기초

01

데이터 과학

경북대학교 배준현 교수
(joonion@knu.ac.kr)



01. 데이터 과학

■ 데이터 과학: *data science*

- 수학과 통계학, 컴퓨터과학, 기타 여러 학문 간의 학제 간 융합적 학문 분야
 - an *interdisciplinary field* of study
- 데이터로부터 실행가능한 지식과 통찰을 발견하기 위해
 - to extract actionable *knowledge* and *insights* from data
- 데이터를 과학적 방법으로 연구하는 학문의 한 분야
 - using *scientific methods*



01. 데이터 과학

데이터 과학의 목적

과거를
분석하여

현재
이해하고

미래를
예측한다



- 과학적 방법: *scientific method*
 - 문제 정의: Define a *question*.
 - 가설 수립: Construct an explanatory *hypothesis*.
 - 데이터 수집: Collect **DATA** from *observations*.
 - 데이터 분석: *Analyze* the DATA.
 - 가설 검정: *Test* your hypothesis by doing *experiments*.
 - 결론 도출: Draw a *conclusion*.
 - 결과 전달: Communicate or *publish* your results.
 - 재현 가능성: *Research* and *reproducibility*.



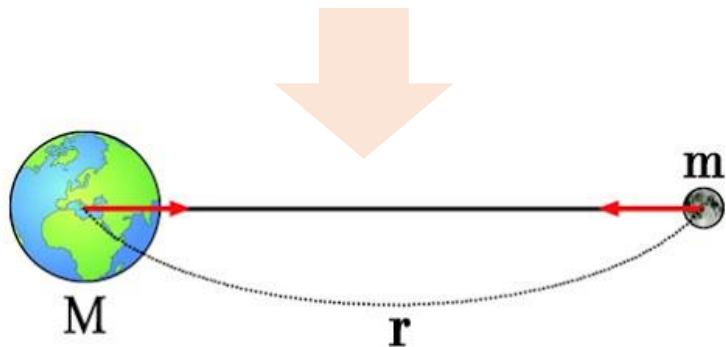
01. 데이터 과학

■ 과학적 방법: *deductive* .vs. *inductive*

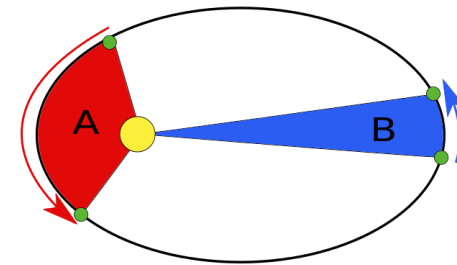
- 연역적 방법: 통찰을 통해 발견한 법칙을 데이터를 통해 검정
- 귀납적 방법: 데이터를 관측하여 유의미한 통찰과 법칙을 발견

뉴턴의 만유인력 법칙

$$F = G \frac{Mm}{r^2}$$



케플러의 행성운동 법칙

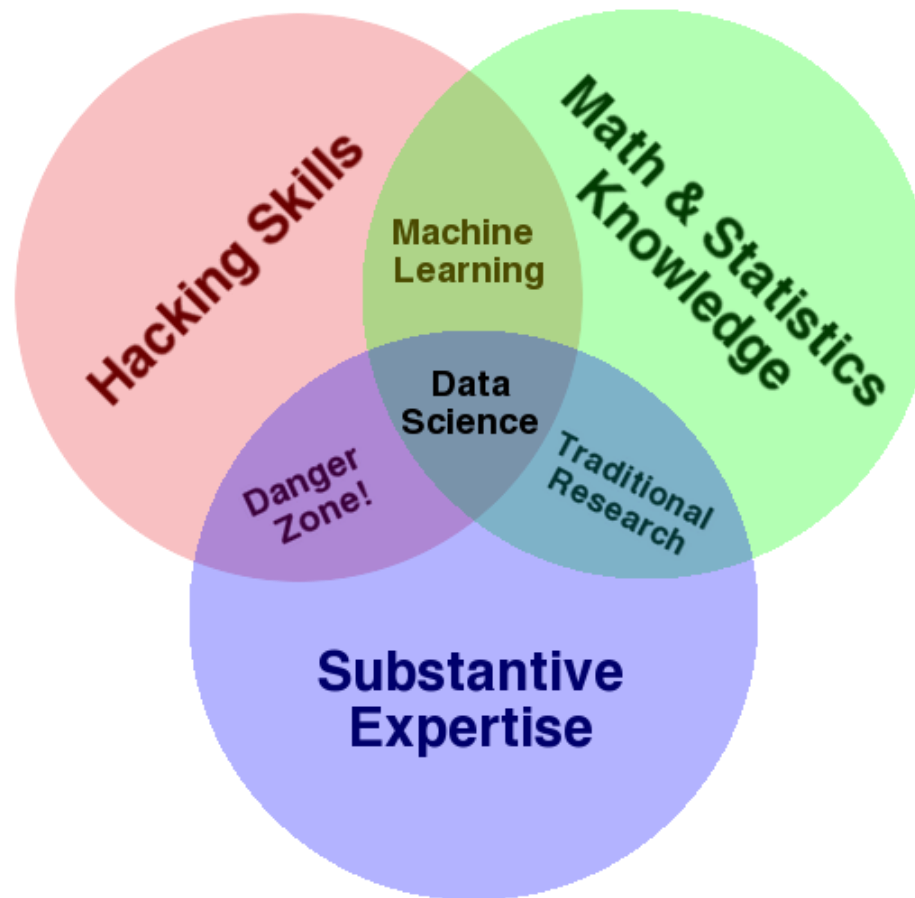


1. 행성은 타원 궤도로 공전한다.
2. 행성이 같은 시간 운동하는 면적이 같다.
3. 행성의 공전주기의 제곱은 긴 궤도 반지름의 세제곱에 비례한다.



■ 학제 간 융합: *interdisciplinary* field

- 수학과 통계학
- 프로그래밍 기술
- 분야별 전문성



The Data Science Venn Diagram
by Drew Conway

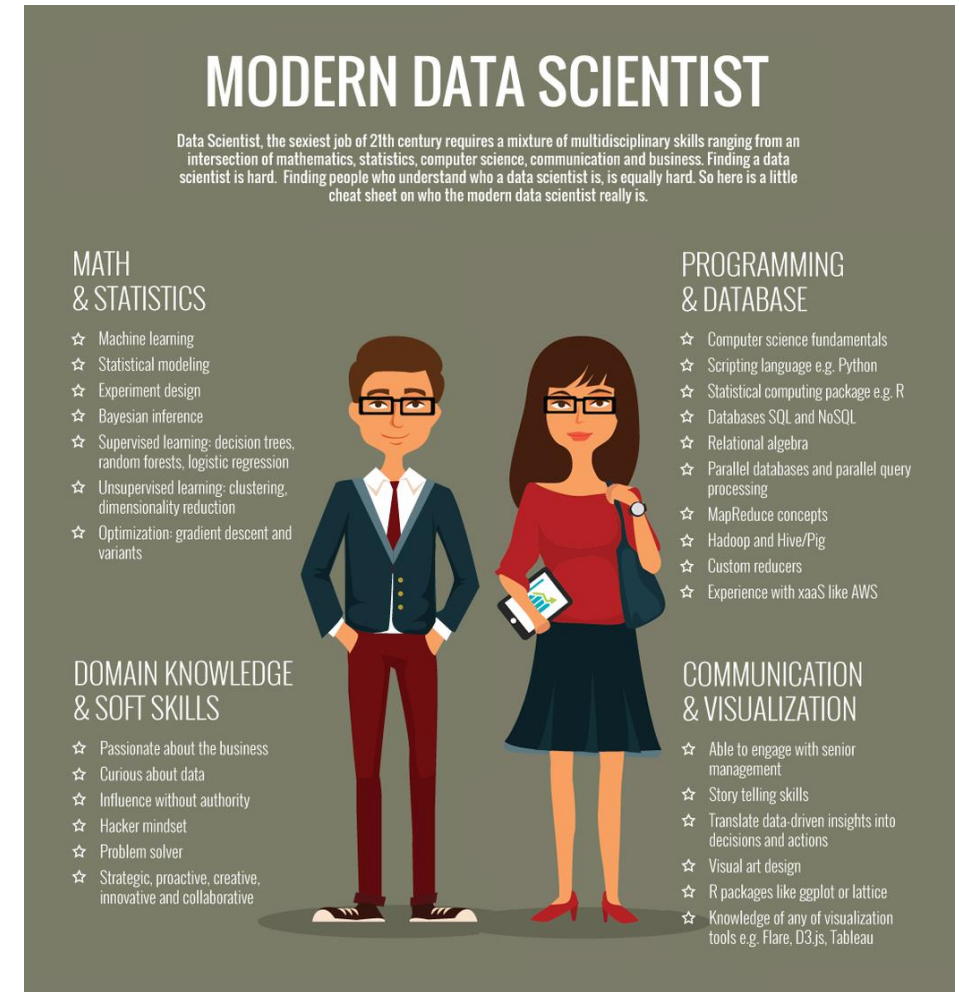
<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>



01. 데이터 과학

■ 데이터 과학자: *data scientist*

- Data Scientist: The *Sexiest Job* of the 21st Century.
 - Harvard Business Review, Oct. 2012.
- Josh Wills says:
 - A data scientist is someone
 - who is *better* at *statistics* than any software engineer
 - and *better* at *software engineering* than any statistician.



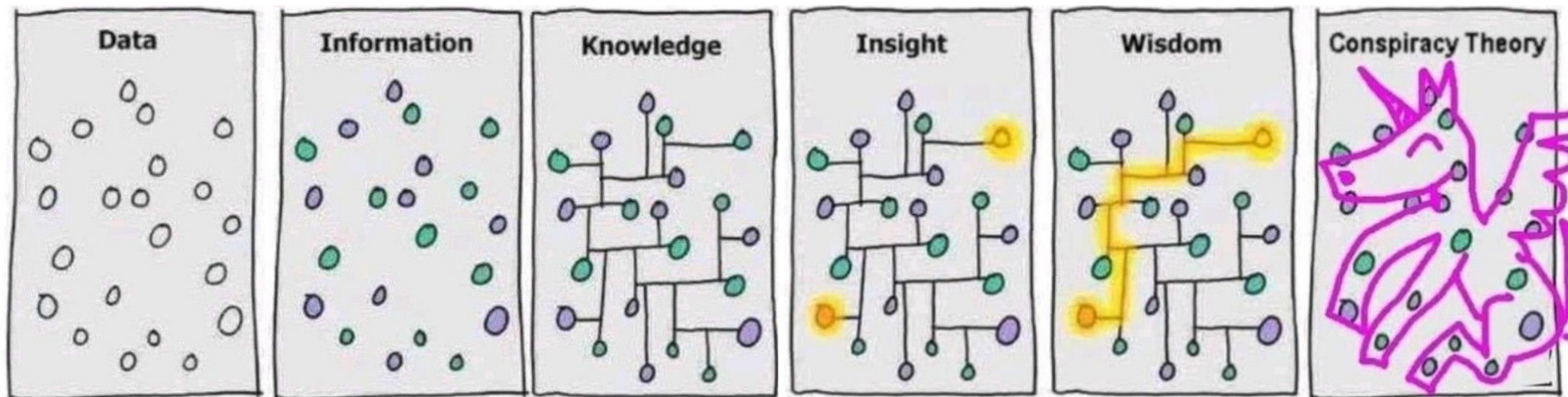
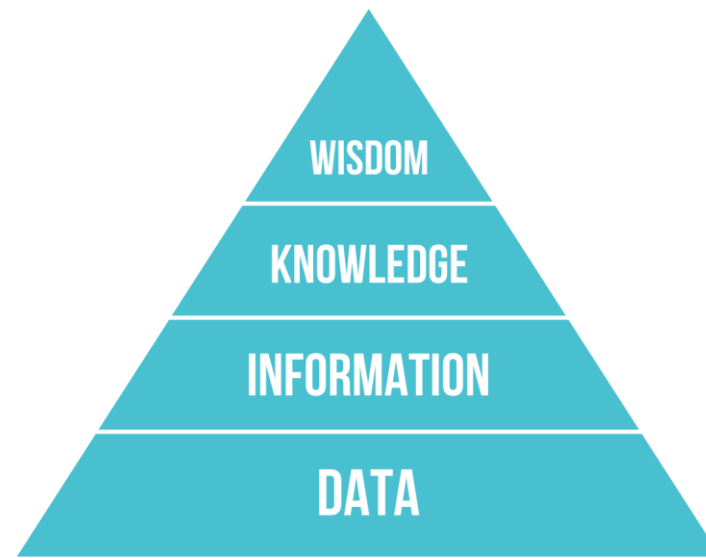
MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing
DISTILLERY



■ DIKW Pyramid:

- 데이터: data
- 정보: information
- 지식: knowledge
- 지혜: wisdom





■ 데이터: from *datum* to *data*

- 데이터: 관찰, 측정, 실험, 또는 조사를 통해 얻는 실체적 사실이나 정보
- 변수(변량): *variable* or *variate*
 - 관찰, 측정, 실험, 또는 조사의 대상이 되는 수량
 - 관측값(*observations*): 변수(변량)에 대한 관측을 통해 얻는 값
- 변수의 유형:
 - 수치형(*numeric*): 수치로 표현할 수 있는 변량. 예) 키, 몸무게
 - 범주형(*categorical*): 범주로 표현할 수 있는 변량. 예) 성별, 혈액형
- 변수의 종류:
 - 독립변수(*feature*): 종속변수에 영향을 주는 변수. 예) 부모의 키
 - 종속변수(*target*): 독립변수로부터 영향을 받는 변수. 예) 자녀의 키



01. 데이터 과학

■ 혼돈의 카오스: 비슷하고 헷갈리는 용어들

수치형 자료 범주형 자료

numeric data *categorical* data

양적 자료 질적 자료

quantitative data *qualitative* data

연속형 자료

continuous data

이산형 자료

discrete data

명목형 자료

nominal data

순서형 자료

ordinal data

독립변수 종속변수

independent variable *dependent* variable

특징변수 목적변수

feature variable *target* variable

설명변수 반응변수

explanatory variable *response* variable

예측변수 결과변수

predictor variable *outcome* variable



01. 데이터 과학

■ 정보: *information*

- 정보: 관찰을 통해 수집한 자료를 실제 문제에 도움되도록 정리한 지식
- 정보의 과학적 정의:
 - 정보란, 불확실성의 해소다!
 - An *information* can be thought of as *the resolution of uncertainty*.

01. 데이터 과학



불확실한 상태
(정보가 없음)



불확실성의 해소
(정보가 있음)



01. 데이터 과학

- 정보량: *quantity* of information
 - 무게의 단위는 kg, 거리의 단위는 km, 정보의 단위는?
 - 정보를 다루기 위해서는 정보를 정량적으로 다룰 수 있어야 함
 - 비트: *bit* = *binary digit*
 - 정보의 최소 단위는 1비트: 0 또는 1 로 표현 가능



01. 데이터 과학

- 정보의 저장과 전송: 정보량의 측정
 - 클로드 섀넌의 공식: $I(x) = -\log_2 p(x)$
 - 세상에서 가장 중요한 공식 중 하나: 디지털 시대를 연 공식
 - 정보의 용량은 어떤 사건 x 가 발생할 확률 $p(x)$ 로 결정할 수 있다.
 - 내일 아침에 해가 동쪽에서 뜰 것이다: 정보량이 매우 적음
 - 내일 아침에 해가 서쪽에서 뜰 것이다: 정보량이 매우 많음





01. 데이터 과학

■ 데이터의 종류:

- **정형 데이터:** *structured data*
 - 숫자형, 범주형 등의 일정한 형식으로 표현할 수 있는 데이터
 - 예) 키, 몸무게, 성별, 혈액형 등
- **비정형 데이터:** *unstructured data*
 - 숫자형, 범주형 등의 일정한 형식이 없는 데이터
 - 예) 텍스트, 이미지, 사운드, 동영상, 또는 이런 데이터들의 혼합
- **반정형 데이터:** *semi-structured data*
 - 일정한 형식이 없지만, 구조적으로 표현할 수 있는 데이터
 - 예) HTML/XML 문서, JSON 포맷 등



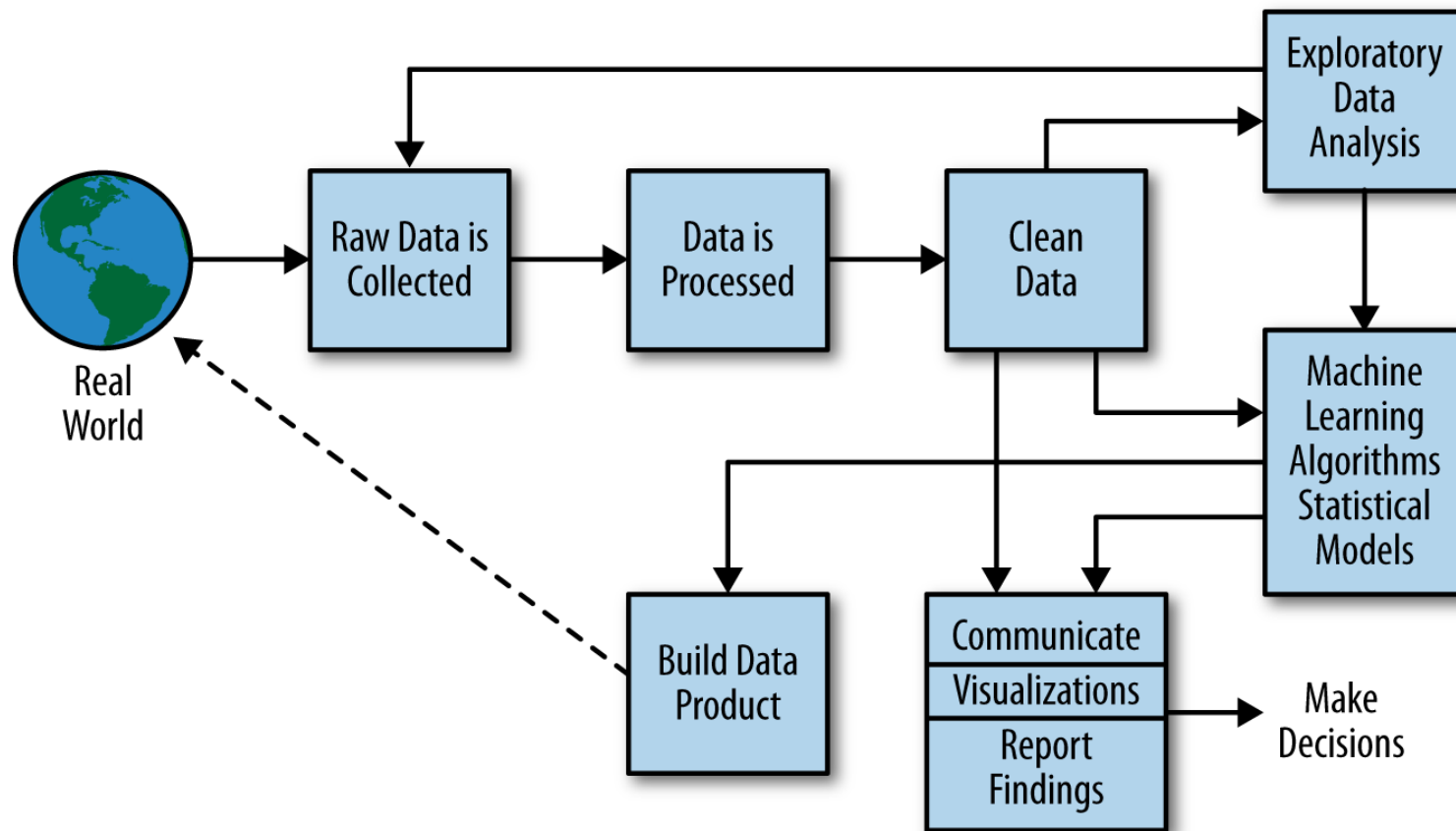
01. 데이터 과학

■ 빅데이터: *BigData*

- 전통적인 데이터 처리 방식으로는 수집, 저장, 분석이 어려운 데이터
- 빅데이터의 세가지 속성(3V):
 - *Volume*: 양적으로 큰 데이터. Tera Bytes, Peta Bytes, etc.
 - *Variety*: 다양한 형태의 데이터. 정형, 비정형, 반정형 데이터
 - *Velocity*: 빠른 속도로 생성되는 데이터. 페이스북, 유튜브, 넷플릭스 등
- 더 중요한 V: (3V + 1V)
 - *Value*: 데이터로부터 얻어낼 수 있는 유의미한 통찰(*insight*)



■ 데이터 과학 프로세스: *Data Science Process*



From: Cathy O'Neil and Rachel Schutt. *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc., 2013.



01. 데이터 과학

■ 문제 정의: *problem definition*

- 데이터 과학의 시작은 문제를 명확히 정의하는 것으로부터 시작
 - 부모의 키가 크면 자녀의 키도 클까?
 - 집값에 영향을 미치는 요인은 무엇일까?
 - 타이타닉호에서 생존한 사람들은 어떤 사람들일까?
 - 고흐가 그린 붓꽃의 품종은 무엇일까?
 - 어떤 종류의 이메일이 스팸 메일일까?
 - 넷플릭스 사용자들에게 어떤 영화를 추천해 주어야 할까?



01. 데이터 과학

■ 데이터 수집: *data collection*

• 데이터화: *datafication*

- a process of taking all aspects of *life* and turning them into *data*.

• 데이터화 의 사례:

- 트위터/페이스북: 생각의 조각을 데이터화
- 구글의 증강현실 안경: 시선의 데이터화
- 링크드인: 직업적 전문가 네트워크의 데이터화

• 데이터화의 중요성:

- Once we *datafy* things,
 - we can *transform* their *purpose*
 - and *turn* the information into *new forms of value*.



01. 데이터 과학

■ 통계학과 데이터 과학:

- 통계학: *Statistics*
 - 데이터를 관찰하고 정리하고 분석하는 방법을 연구하는 전통적인 학문
 - 통계적 추론: *statistical inference*
 - 연구대상에 대한 가설을 세우고 모집단으로부터 표본 추출
 - 추출한 표본에 대해서 가설이 통계적으로 유의미한가를 검정
- 모집단과 표본추출: *population* and *sampling*
 - 빅데이터의 시대: 여전히 표본추출이라는 방식이 필요한가?
 - 예) 2016년 미국 대선: 전통적 여론 조사 .vs. SNS 빅데이터 분석



01. 데이터 과학

■ 데이터 모델링: *data modeling*

- 수학 함수를 통해 데이터의 형태와 구조를 표현하는 것
- 모형 적합: *fitting a model*
 - 관찰된 데이터를 대상으로 데이터의 모형을 추정하는 작업
 - 예) 선형회귀: 데이터를 선형 방정식으로 모델링하는 방법
- 과(대)적합과 과소 적합: *overfitting* and *underfitting*
 - 과적합: 학습 데이터에 과도하게 편향된 모형 적합
 - 과소적합: 학습 데이터를 제대로 설명하지 못하는 모형 적합



01. 데이터 과학

■ 기계학습: *machine learning*

- 데이터를 가장 잘 설명하는 모형을 경험을 통해 스스로 학습하는 알고리즘
- 지도 학습: *supervised* learning
 - 예측하려는 변수의 정답이 있으므로 정답 여부를 확인할 수 있는 경우
 - 예측형 모델(*predictive model*): 회귀(*regression*), 분류(*classification*)
- 비지도 학습: *unsupervised* Learning
 - 해결하려는 문제가 따로 정답이 정해져 있지 않은 경우
 - 설명형 모델(*descriptive model*): 군집화(*clustering*), 연관(*association*)
- 강화 학습: *reinforcement* learning
 - 행동(*action*)에 대한 보상(*reward*)을 통해 학습
 - 딥 러닝: *deep reinforcement learning*



01. 데이터 과학

■ 기계학습 알고리즘의 종류:

- 예측: *prediction*
 - 선형회귀(linear regression): 단순(simple), 다중(multiple), 다항(polynomial)
- 분류: *classification*
 - 로지스틱 회귀(logistic regression), 결정나무(decision tree), k-최근접이웃(kNN)
- 군집화: *clustering*
 - 계층적 군집화(hierarchical clustering), k-평균(k-means)
- 추론: *inference*
 - 나이브 베이지안(naïve Bayesian)
- 인공지능: *artificial intelligence*
 - 다층퍼셉트론(MLP), 합성곱신경망(CNN), 심층신경망(RNN)



01. 데이터 과학

- 데이터 시각화: *data visualization*
 - 데이터 분석의 결과를 시각적으로 표현하고 전달하는 방법





01. 데이터 과학

■ 정보의 시각화

- 문학 작품에서 보는 정보 시각화의 사례: *Going Home*, Pete Hamil, 1971.



Tie a yellow ribbon round the old oak tree, .
노란 리본을 오크나무에 매달아 주세요

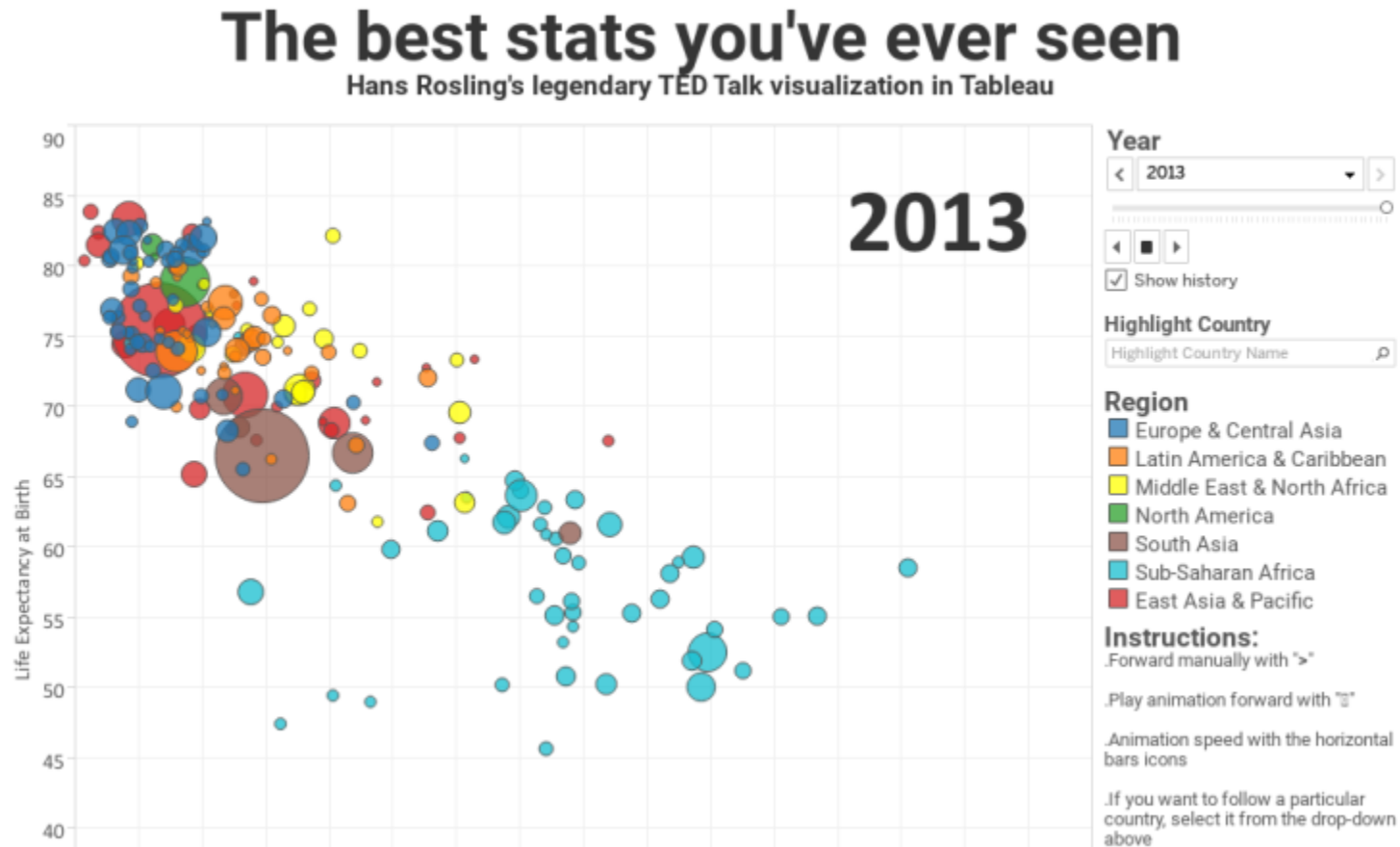
https://youtu.be/ijMgbW8n_HU



01. 데이터 과학

■ 데이터 시각화와 스토리텔링

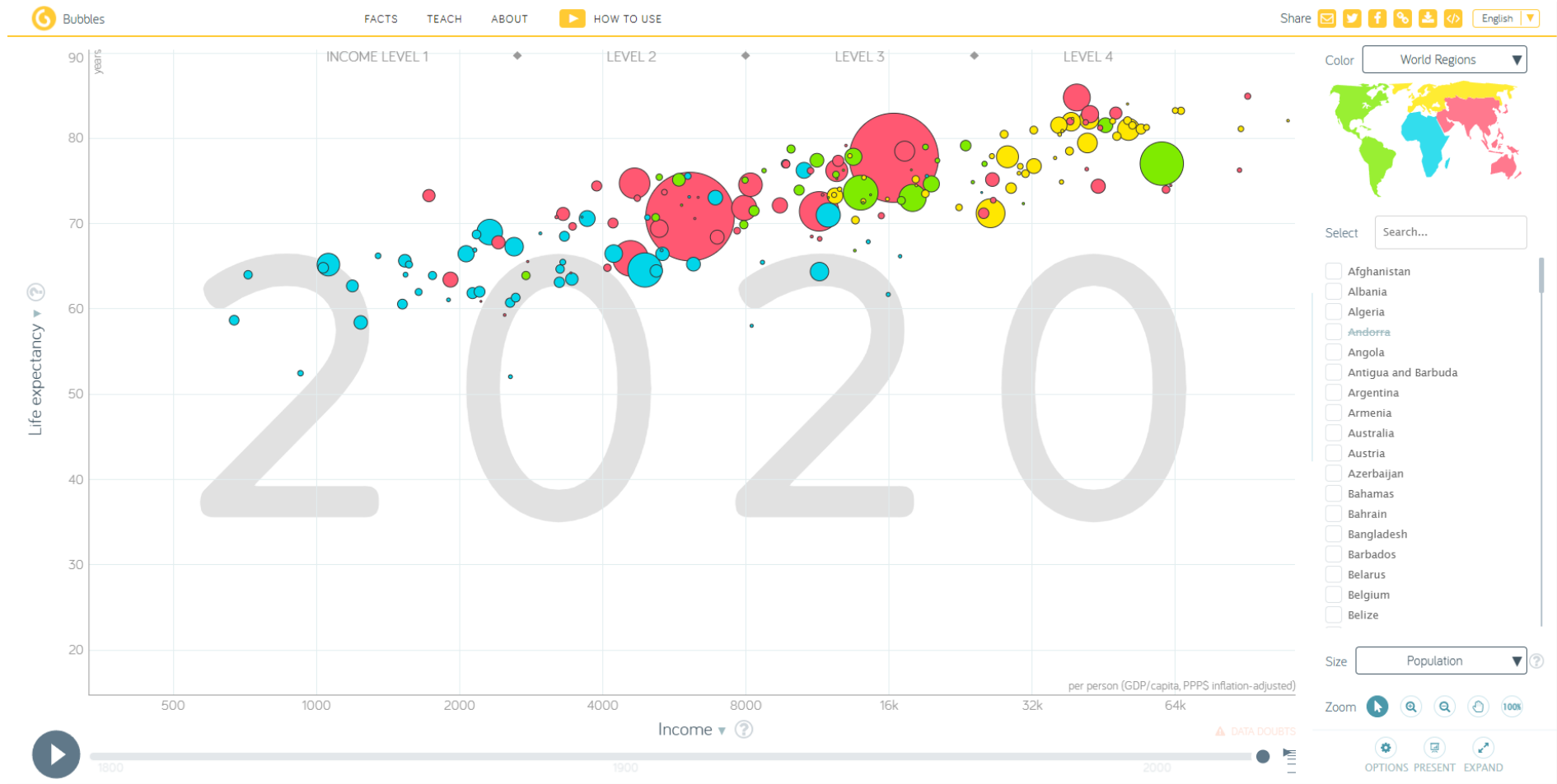
- 데이터 스토리텔링: 데이터 분석을 통해 얻은 통찰을 효과적으로 전달하는 기술



<https://youtu.be/To8QEKmmmog>



01. 데이터 과학



[https://tools-legacy.gapminder.org/tools/#\\$state\\$time\\$value=2020;&chart-type=bubbles](https://tools-legacy.gapminder.org/tools/#$state$time$value=2020;&chart-type=bubbles)



01. 데이터 과학

■ 데이터 과학을 위한 도구:

- *R* / R Studio:
 - 통계학, 데이터 과학을 위한 전통적인 프로그래밍 언어와 도구
 - 데이터 과학을 위한 풍부한 라이브러리, 플랫폼, 커뮤니티가 존재함
- *Python*:
 - 범용 프로그래밍 언어와 플랫폼으로서 데이터 과학을 강력하게 지원
 - Jupyter Notebook, Pandas, Scikit-Learn, TensorFlow, PyTorch, etc.
- *Orange 3*:
 - 코딩 없이 **비주얼 프로그래밍**으로 데이터 과학을 할 수 있는 도구
- *Power BI & Tableau*:
 - Business Intelligence를 지원하는 강력한 데이터 시각화 도구

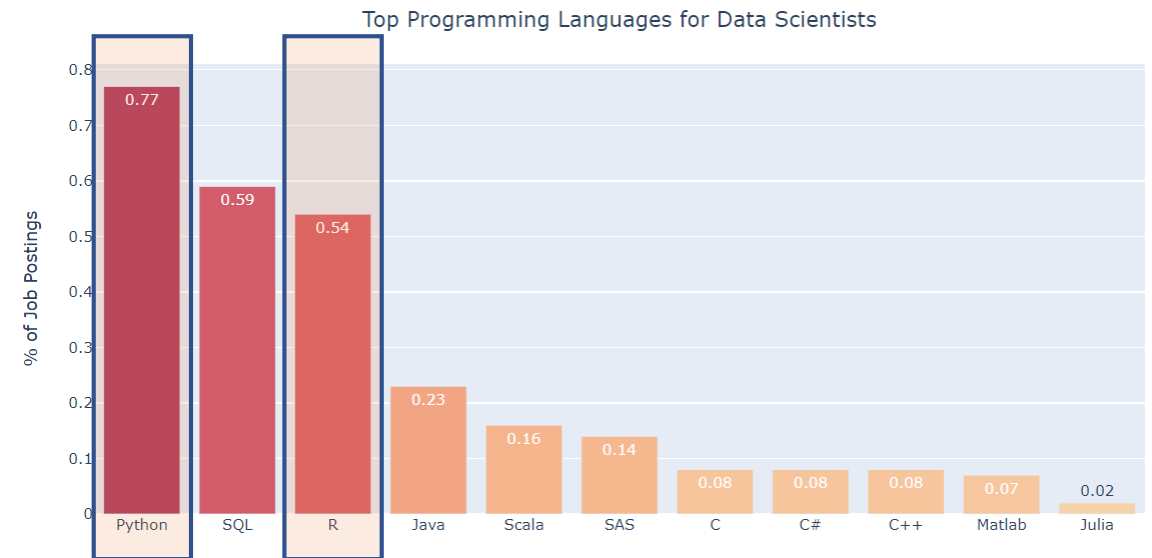


01. 데이터 과학

■ R .vs. Python:

Nov 2021	Nov 2020	Change	Programming Language	Rating	Change
1	2	↑	Python	11.77%	-0.35%
2	1	↓	C	10.72%	-5.49%
3	3		Java	10.72%	-0.96%
4	4		C++	8.28%	+0.69%
5	5		C#	6.06%	+1.39%
6	6		Visual Basic	5.72%	+1.72%
7	7		JavaScript	2.66%	+0.63%
8	16	↑	Assembly language	2.52%	+1.35%
9	10	↑	SQL	2.11%	+0.58%
10	8	↓	PHP	1.81%	+0.02%
11	21	↑	Classic Visual Basic	1.56%	+0.83%
12	11	↓	Groovy	1.51%	-0.00%
13	15	↑	Ruby	1.43%	+0.22%
14	14		Swift	1.43%	+0.08%
15	9	↓	R	1.28%	-0.36%
16	12	↓	Perl	1.22%	-0.29%
17	18	↑	Delphi/Object Pascal	1.22%	+0.36%
18	13	↓	Go	1.21%	-0.16%
19	34	↑	Fortran	1.19%	+0.79%
20	17	↓	MATLAB	1.17%	+0.07%

Source: TIOBE index 2021.

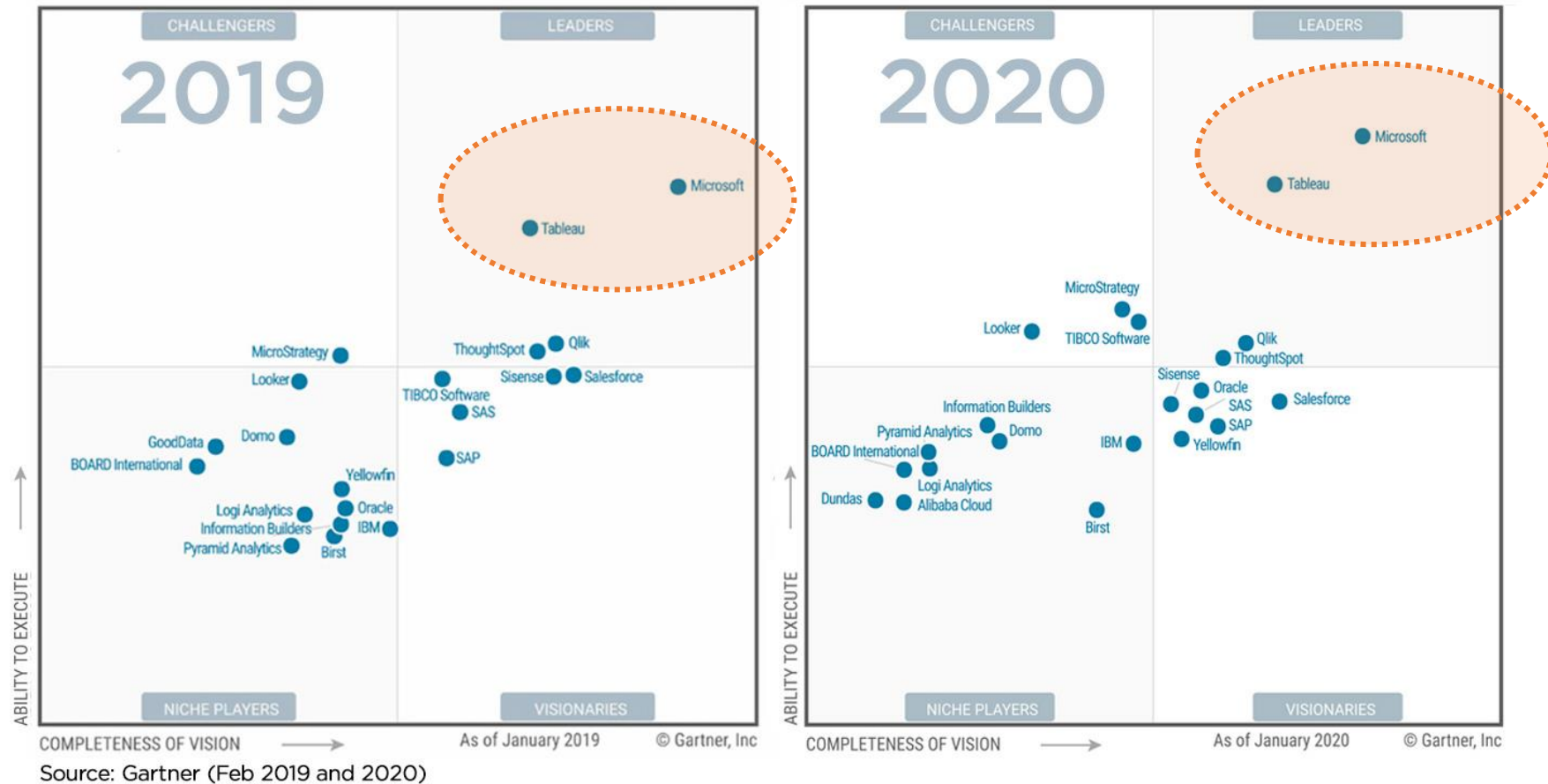


Source: Towards Data Science 2021.



01. 데이터 과학

■ Power BI .vs. Tableau:





01. 데이터 과학

■ Orange 3:

- 프로그래밍 없이 데이터 과학을 제대로 학습할 수 있는 강력한 교육용 도구
- Orange Data Mining: <https://orangedatamining.com/>

[Screenshots](#)[Workflows](#)[Download](#)[Blog](#)[Docs](#)[Workshops](#)

Data Mining Fruitful and Fun

Open source machine learning and data visualization.

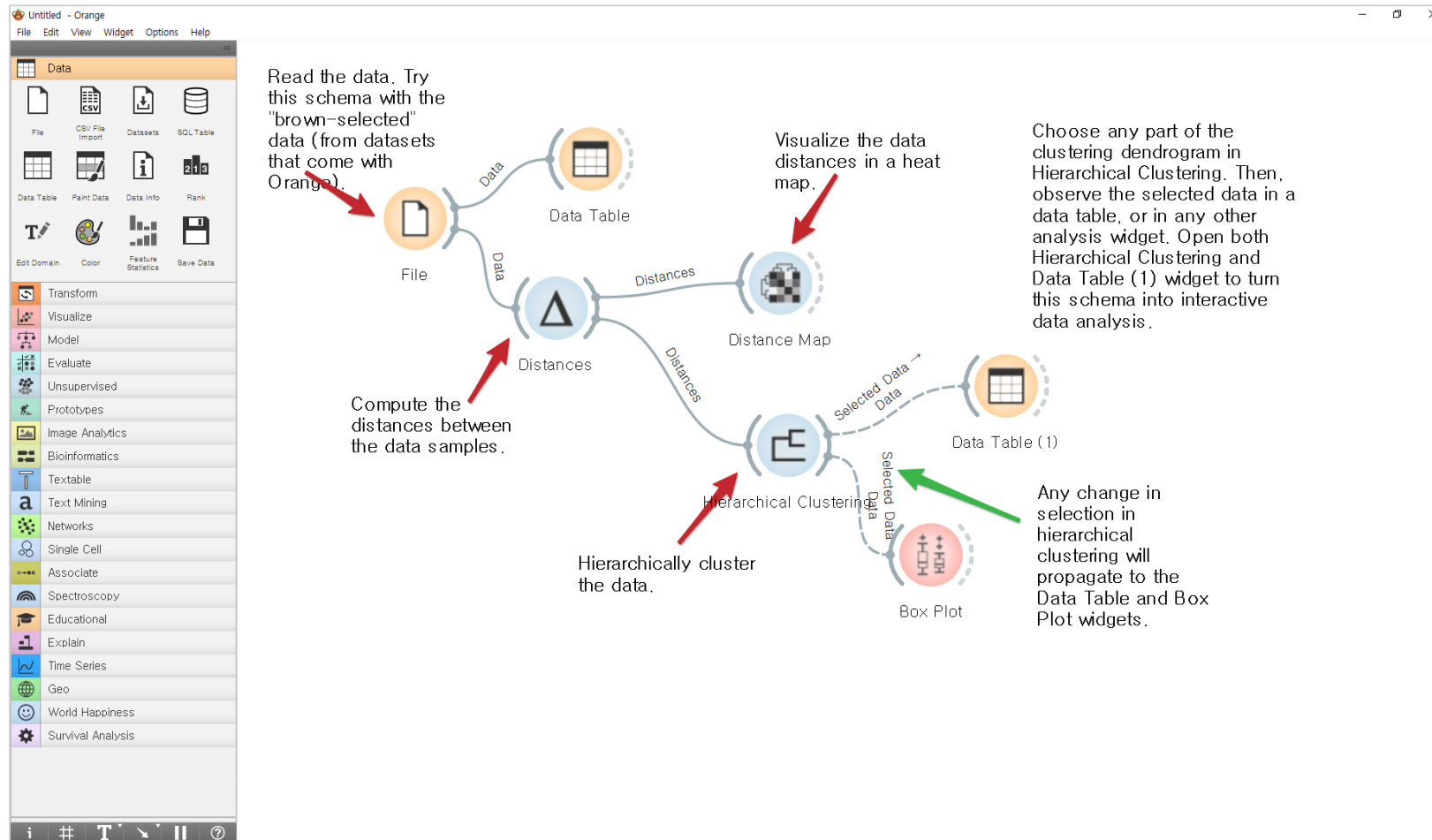
Build data analysis workflows visually, with a large, diverse toolbox.

[Download Orange](#)



01. 데이터 과학

- Orange 설치 및 실행:
 - 워크플로우, 위젯, 애드온



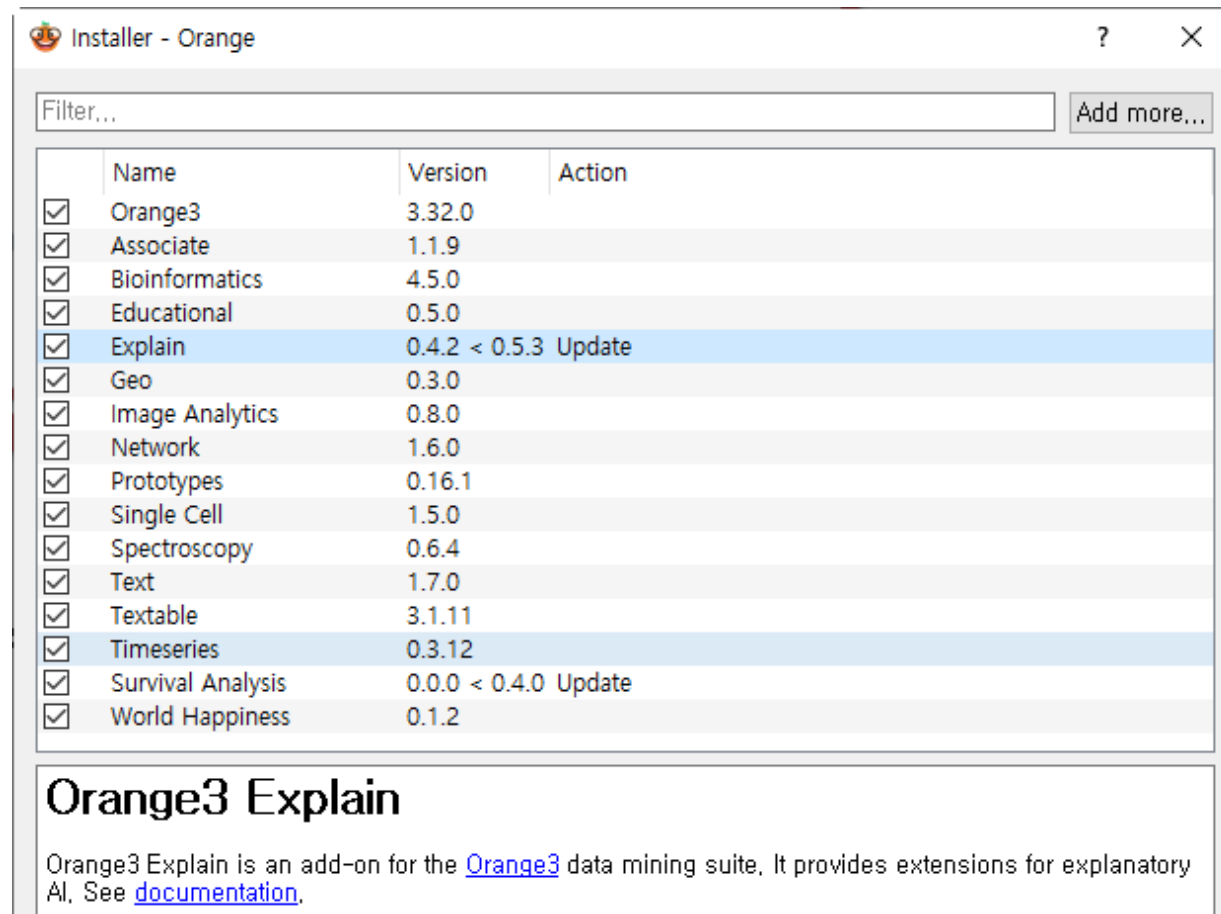
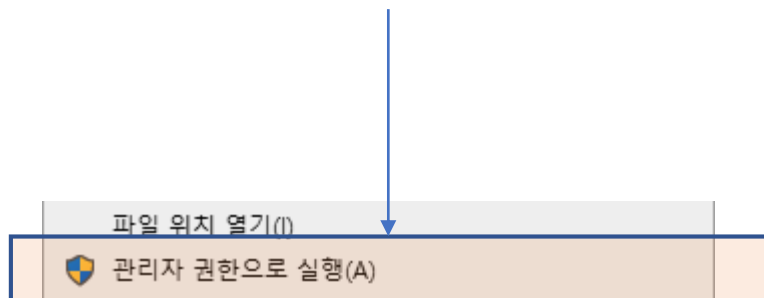


01. 데이터 과학

■ 애드온 설치:

- Orange를 관리자 권한으로 실행
- Options > Add-ons

실행 아이콘을 마우스 우클릭

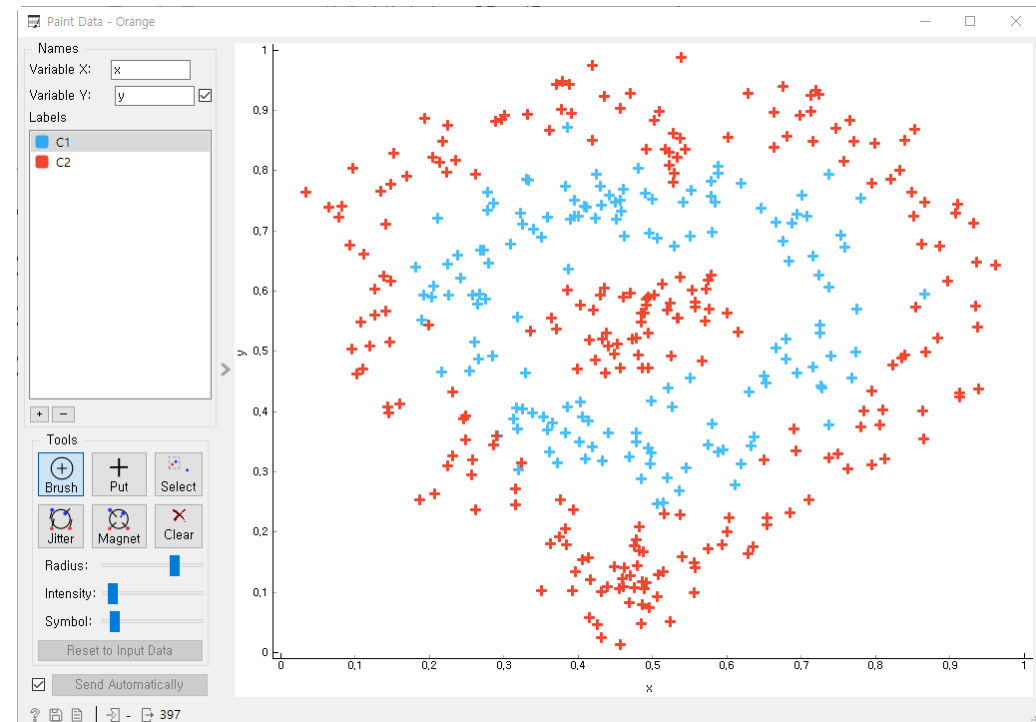
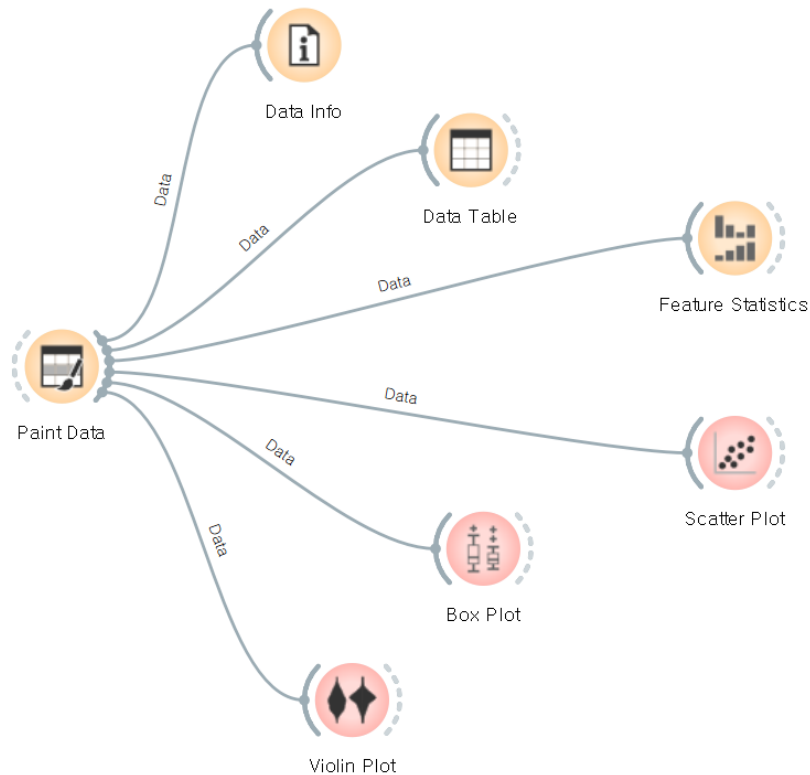




01. 데이터 과학

■ Orange 실습:

- 데이터 만들기: Paint Data
- 데이터 탐색: Data Info, Data Table, Feature Statistics
- 데이터 시각화: Scatter Plot, Box Plot, Violin Plot





01. 데이터 과학

Data Info

Data Info - Orange ? X

Data Set Name
Painted data

Data Set Size
Rows: 397
Columns: 3

Features
Categorical: -
Numeric: 2

Targets
Categorical outcome with 2 values

Meta Attributes
None

Location
Data is stored in memory

Data Attributes

? | 397

Data Table

Data Table - Orange - □ X

Info
397 instances (no missing data)
2 features
Target with 2 values
No meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Class

	Class	x	y
1	C1	0.365922	0.380968
2	C1	0.353427	0.391499
3	C1	0.338226	0.396784
4	C1	0.359091	0.368885
5	C1	0.361726	0.33284
6	C1	0.413931	0.383638
7	C1	0.317106	0.405209
8	C1	0.291482	0.358942
9	C1	0.261058	0.514651
10	C1	0.276253	0.585851
11	C1	0.190374	0.551061
12	C1	0.191897	0.59269
13	C1	0.205415	0.608115
14	C1	0.258064	0.592639
15	C1	0.211255	0.719962
16	C1	0.181743	0.64034
17	C1	0.278064	0.76445
18	C1	0.322329	0.729595
19	C1	0.278773	0.733803
20	C1	0.385964	0.871014
21	C1	0.4241	0.793096
22	C1	0.423747	0.720075

Restore Original Order

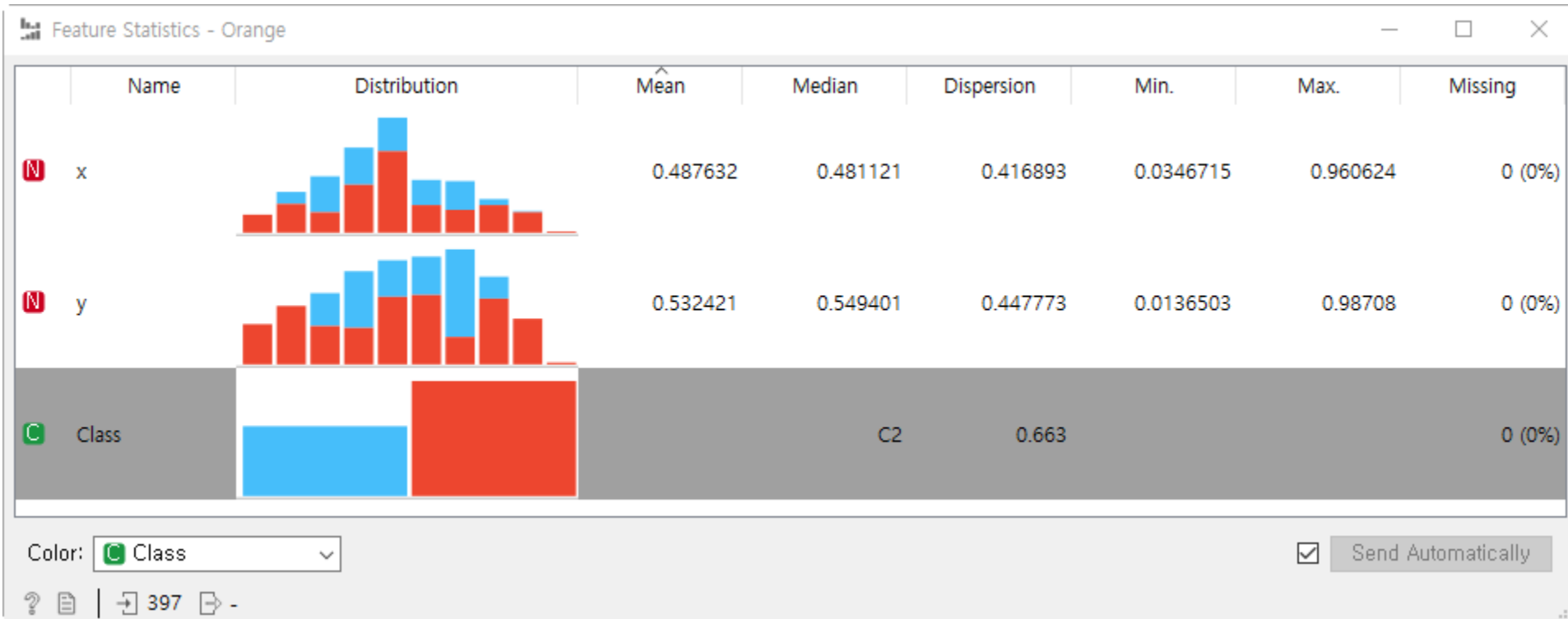
☒ Send Automatically

? | 397 | 397 | 397



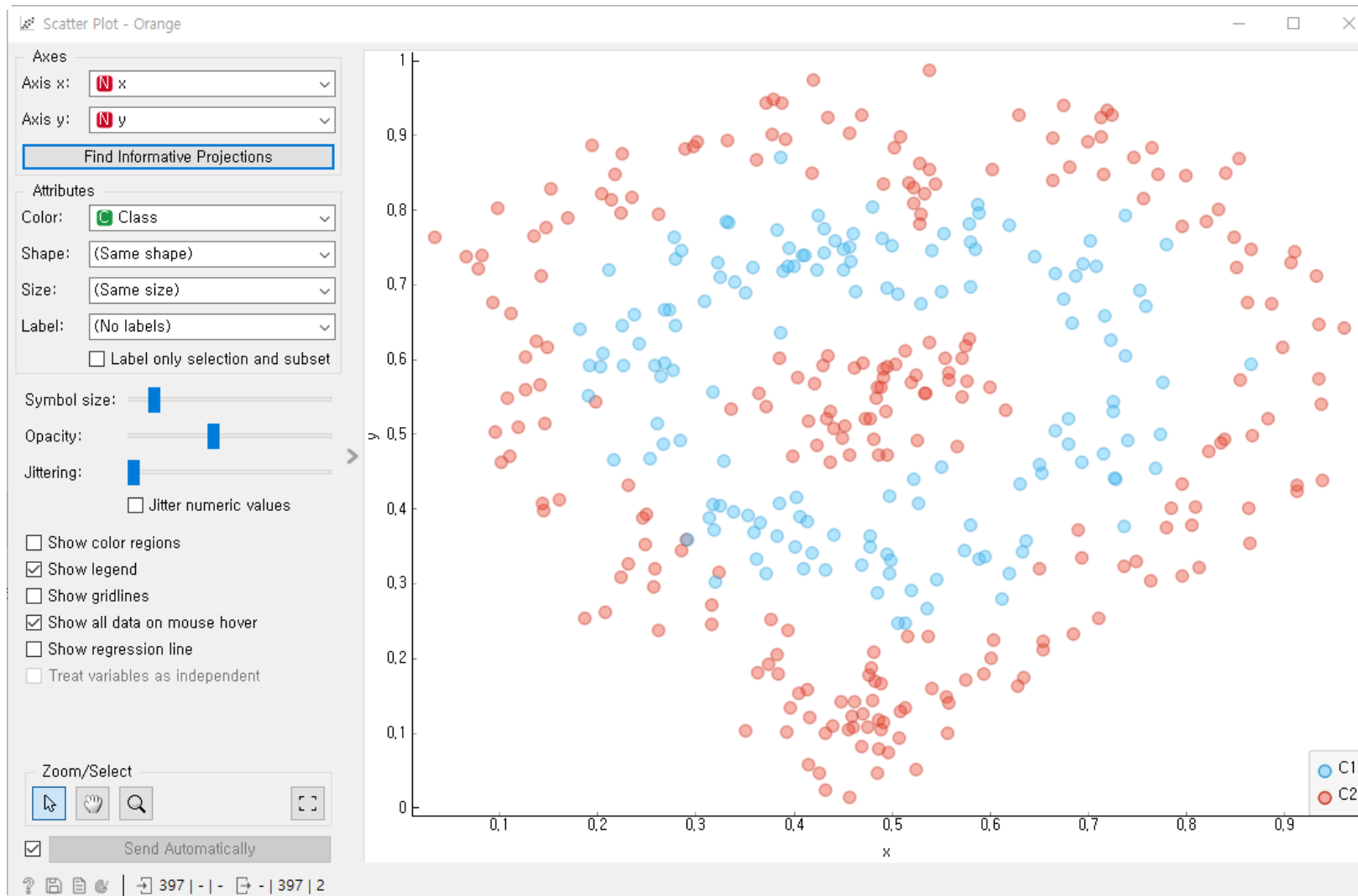
01. 데이터 과학

Feature Statistics



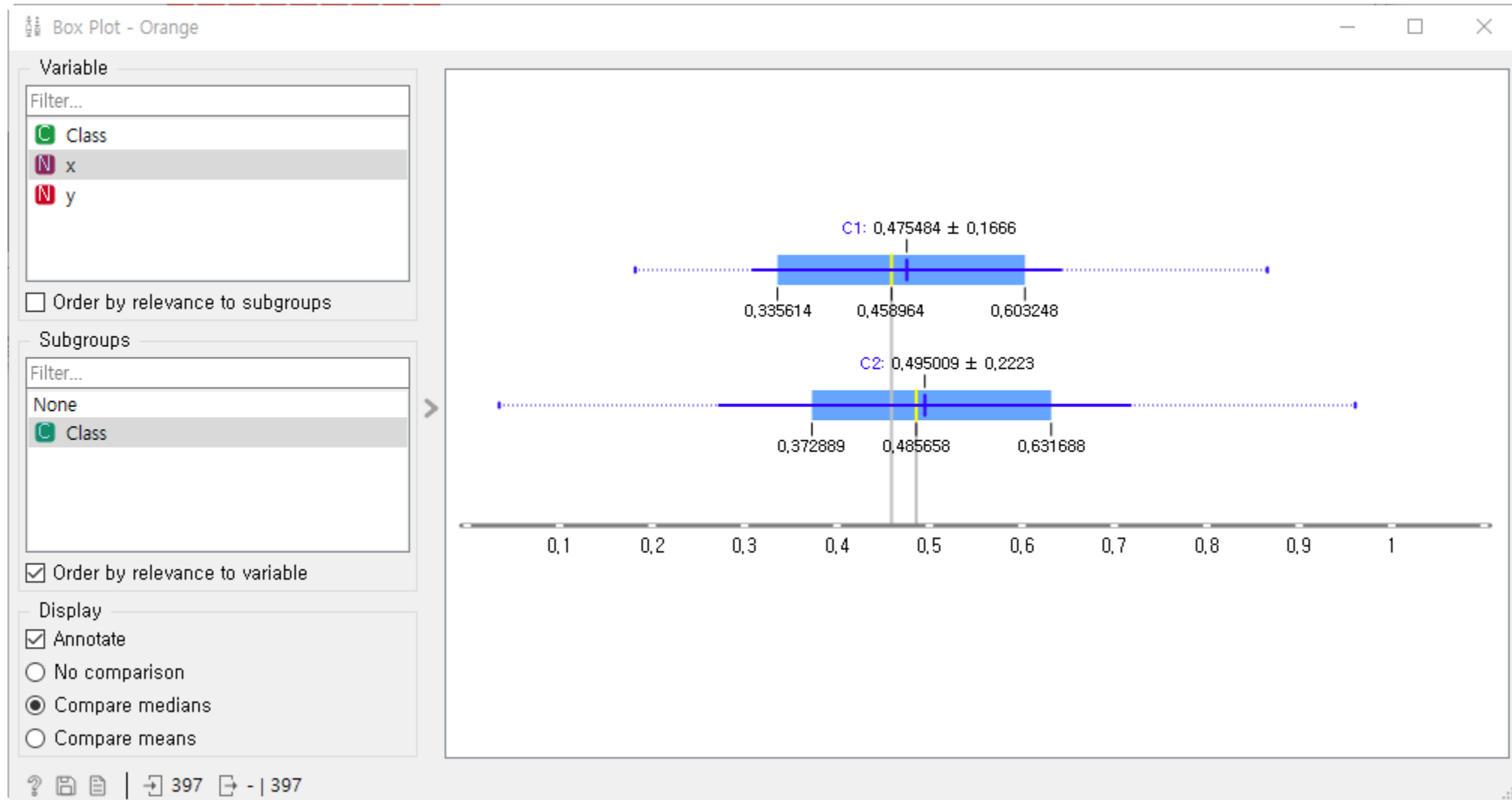


Scatter Plot



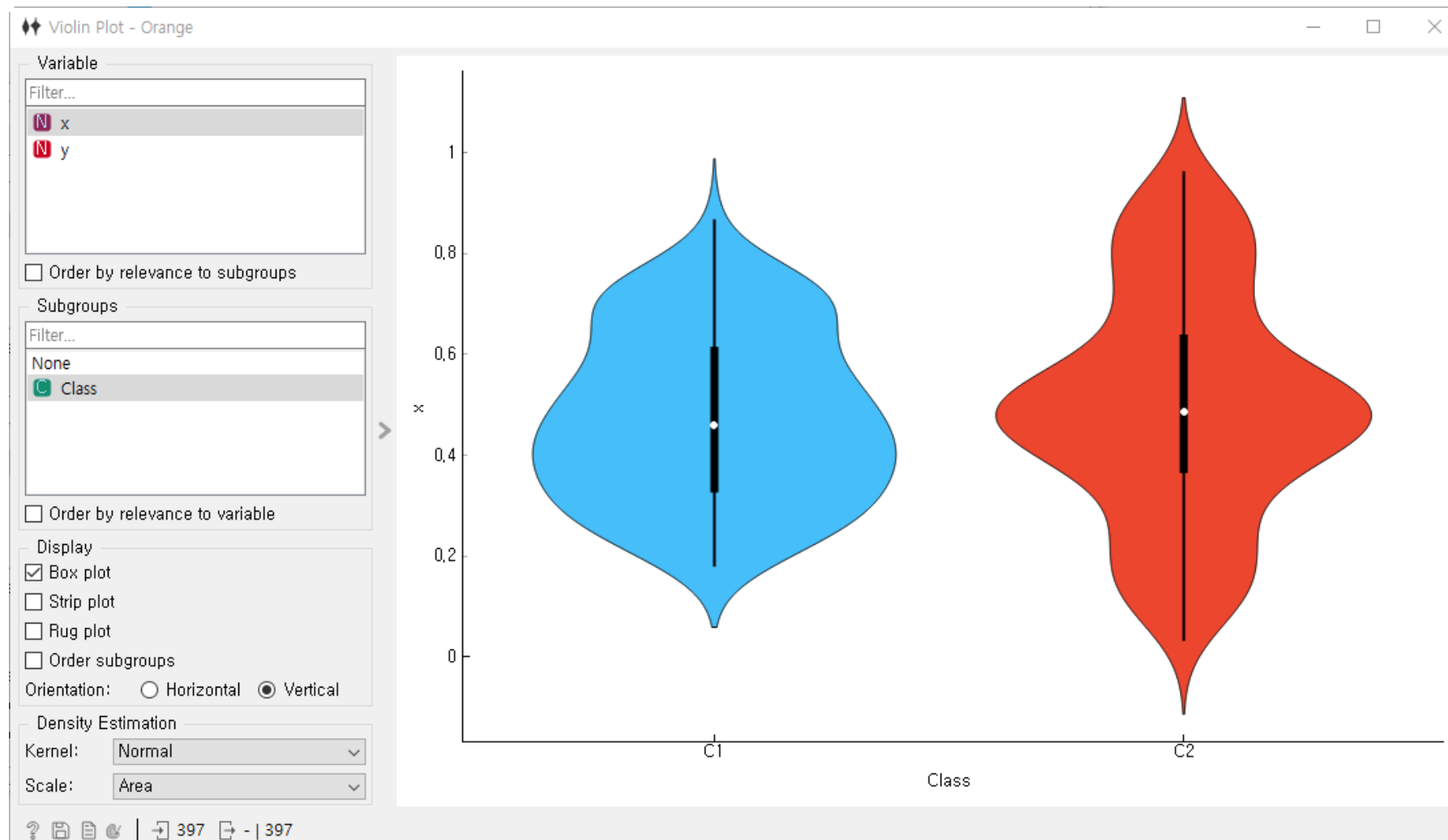


Box Plot





Violin Plot



Tableau란?

Tableau는 데이터를 사용해 문제를 해결하는 방식에 혁신을 가져온 시각적 분석 플랫폼으로, 사람과 조직이 데이터를 최대한 활용하도록 지원합니다.

<https://tableau.com/>



01. 데이터 과학

태블로 주요 제품군

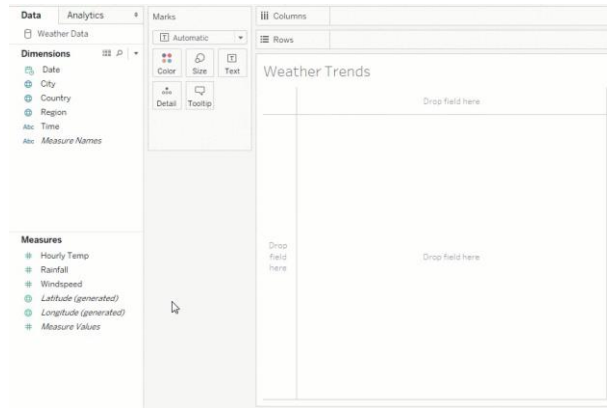


Tableau Desktop

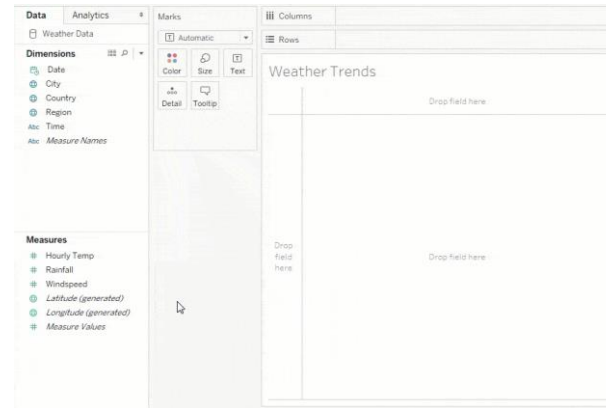


Tableau Public

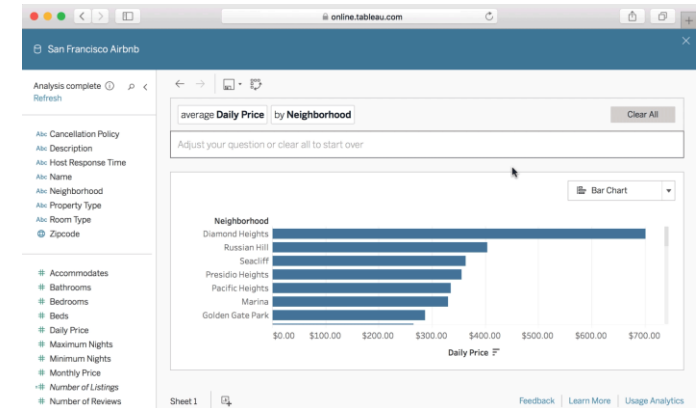


Tableau Online



01. 데이터 과학

■ Tableau 시작하기

- Tableau Public 가입: <https://public.tableau.com>



데이터 스토리텔링

멋진 대화형 비주얼라이제이션을 Tableau의 무료 플랫폼에서 손쉽게 만들어 보십시오. 코딩은 필요 없습니다.



대화의 문을 여세요

전 세계의 비주얼라이제이션 작성자들과 교류합니다. 나의 비주얼라이제이션을 개인 웹 사이트, 블로그 또는 소셜 미디어에 올려 보십시오.



영감 얻기

백만이 넘는 사용자가 만들어 내는 다양한 가능성이 존재하는, 세계에서 가장 광범위한 데이터 시각화 라이브러리를 살펴보고 직접 참여해 보십시오.



01. 데이터 과학

테블로 퍼블릭 계정을 만들기 위해 프로필 만들기에 정보 입력

tableau public

더 알아보기 블로그 리소스 정보 등록 로그인

프로필 만들기

이름
실명을 사용해야 하며, 실명은 커뮤니티의 신뢰성을 높여 줍니다.

이메일
본인의 이메일을 사용하여 Tableau Public에 로그인하십시오. 이 정보는 Tableau만 볼 수 있습니다. Tableau는 사용자의 개인 정보를 절대 제3자에게 양도하거나, 판매하거나, 다른 정보와 교환하지 않습니다.

비밀번호
8자 이상이어야 하며 영문자, 숫자 및 특수 문자가 포함되어야 합니다.

확인

법적 고지 사항 검토
☐ 서비스 약관을 읽고 동의함

☐ 로봇이 아닙니다. reCAPTCHA 개인 정보 보호 · 약관

내 프로필 만들기





01. 데이터 과학

테블로 퍼블릭에 로그인 후 비주얼리제이션 만들기(베타)를 선택

tableau public 더 알아보기 블로그 리소스 정보

Joonion Bae
KNU의 Prof. | Bukgu, Taegu-Kwangyokshi, Korea, Republic of

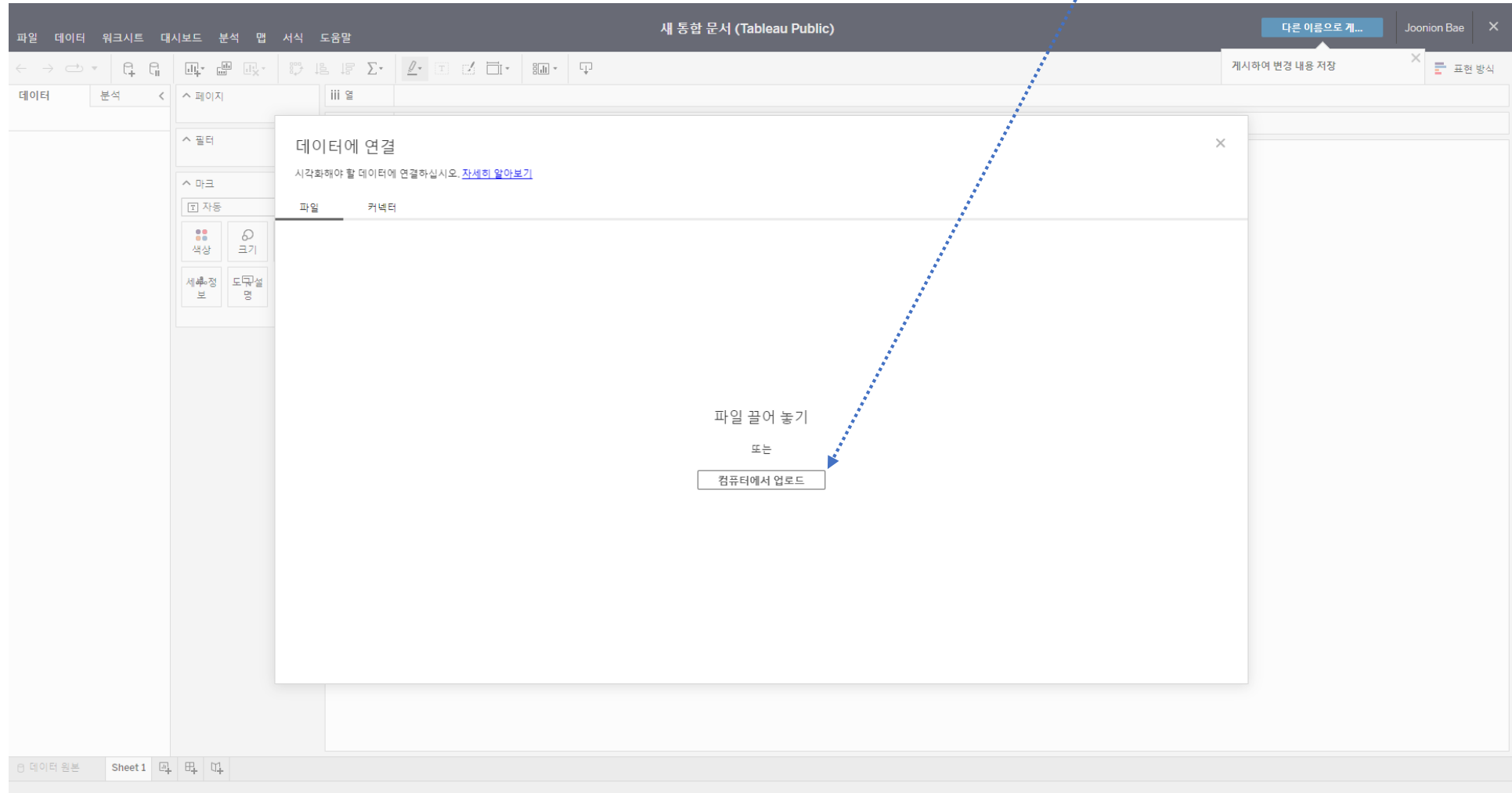
프로필 편집

비주얼리제이션 4 즐거찾기 8 팔로우 8 팔로워 10 비주얼리제이션 만들기(베타)



01. 데이터 과학

[데이터] > [새 데이터 원본] 에서 데이터 원본 선택 (예제 엑셀파일)





01. 데이터 과학

[데이터 원본] 탭에서 데이터 보기

Tableau - CH20_BBOD_Complaints- 13일 후 Tableau 라이선스 만료

파일(F) 데이터(D) 서버(S) 창(N) 도움말(H)

연결: Sample - Superstore (Microsoft Excel)

시트: ☐ 데이터 해석기 사용
데이터 해석기에서 Microsoft Excel 통합 문서를 지우지 못할 수 있습니다.

Orders, People, Returns

연결: ☒ 라이브 ☐ 추출

필터: 0 | 추가

Orders (19개 필드 9994개 행)

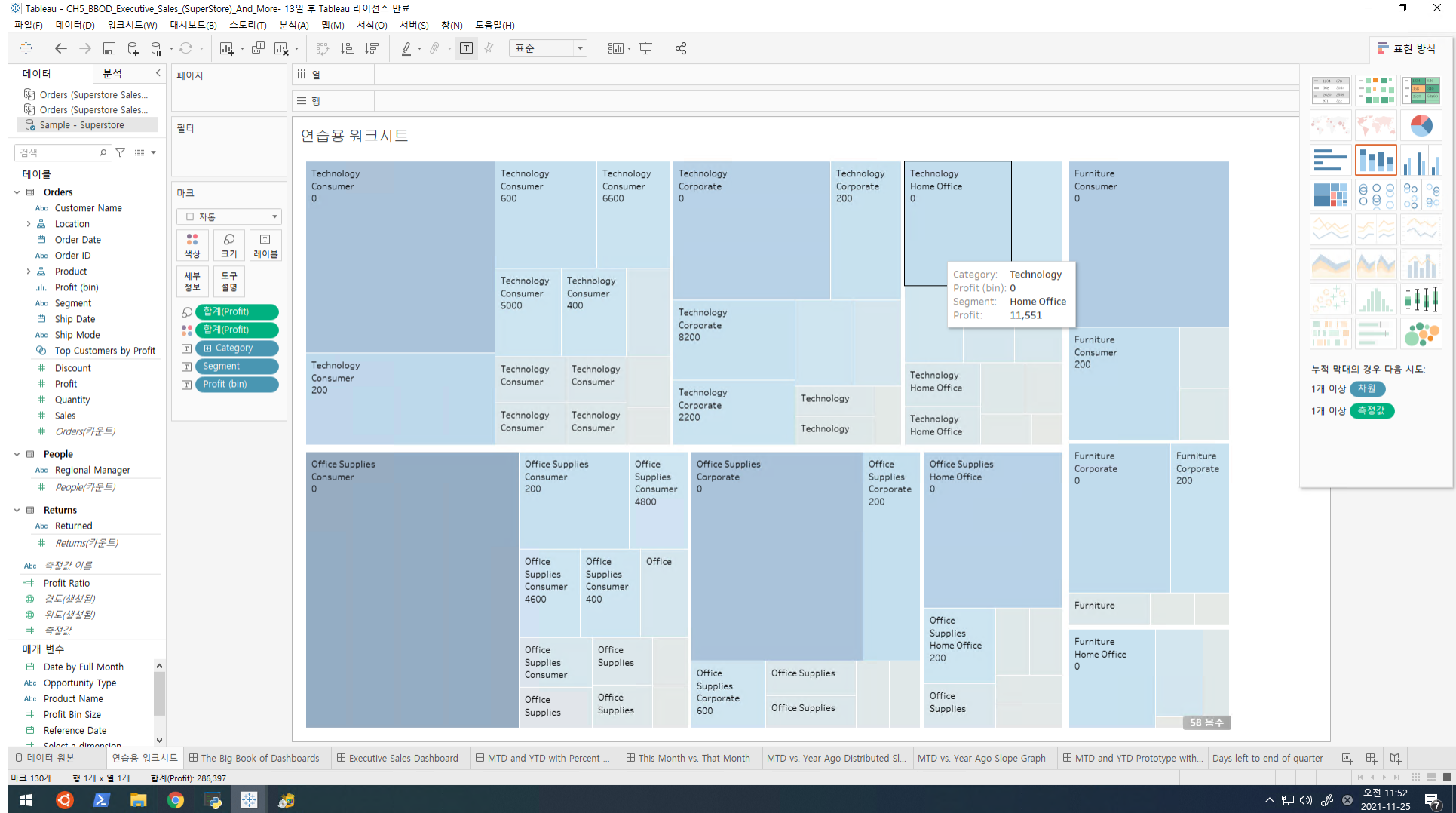
이름	필드명	물리적 테이블	원격 필드명
Order ID	Order ID	Orders	Order ID
Order Date	Order Date	Orders	Order Date
Ship Date	Ship Date	Orders	Ship Date
Ship Mode	Ship Mode	Orders	Ship Mode
Customer Name	Customer Name	Orders	Customer Name
Segment	Segment	Orders	Segment

Order ID	Order Date	Ship Date	Ship Mode	Customer Name	Segment	Country/Region	City	State	Postal Code	Region
CA-2020-152156	2020-11-08	2020-11-11	Second Class	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South
CA-2020-152156	2020-11-08	2020-11-11	Second Class	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South
CA-2020-138688	2020-06-12	2020-06-16	Second Class	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West
US-2019-108966	2019-10-11	2019-10-18	Standard Class	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South
US-2019-108966	2019-10-11	2019-10-18	Standard Class	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South
CA-2018-115812	2018-06-09	2018-06-14	Standard Class	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
CA-2018-115812	2018-06-09	2018-06-14	Standard Class	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
CA-2018-115812	2018-06-09	2018-06-14	Standard Class	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
CA-2018-115812	2018-06-09	2018-06-14	Standard Class	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
CA-2018-115812	2018-06-09	2018-06-14	Standard Class	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West

데이터 원본 | The Big Book of Dashboards | Complaints Dashboard | Image of Dashboard | 시트 6

Joonion Bae

[새 워크시트]를 선택해서 워크시트에 비주얼리제이션 추가하기





01. 데이터 과학

■ 연습문제:

- Orange 3를 다운로드 하고 Add-Ons를 설치해 보시오.
 - Orange 3의 워크플로우 예제들을 실험해 보시오.
- Tableau Public에 가입하여 본인의 계정을 만드시오.
 - 다른 사람들의 비주얼리제이션을 감상해 보시오.

Any Questions?

