

비즈니스 애널리틱스

데이터 시각화(Data Visualization)

SNU Business School

Open Datasets



UCI 데이터셋

· University of California, Irvine에서 제공하는 데이터셋

- 2017년 현재 379개의 자유롭게 활용할 수 있는 토이 데이터셋(toy datasets)을 제공하고 있음

- 경영, 의료, 공학 등 다양한 분야의 데이터를 정형화된 형태로 제공하므로 데이터 시각화와 탐색 등 기초적인 연습을 수행하기에 좋음

- URL: <https://archive.ics.uci.edu/ml/datasets.html>



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)

☐ Repository ☐ Web 

[View ALL Data Sets](#)

Browse Through: 379 Data Sets

Default Task

Classification (277)
Regression (71)
Clustering (60)
Other (53)

Attribute Type

Categorical (37)
Numerical (232)
Mixed (55)

Data Type

Multivariate (296)
Univariate (16)
Sequential (39)
Time-Series (72)
Text (34)
Domain-Theory (22)
Other (21)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998

[Table View](#) [List View](#)

Kaggle 데이터셋

- kaggle.com에서 제공하는 데이터셋

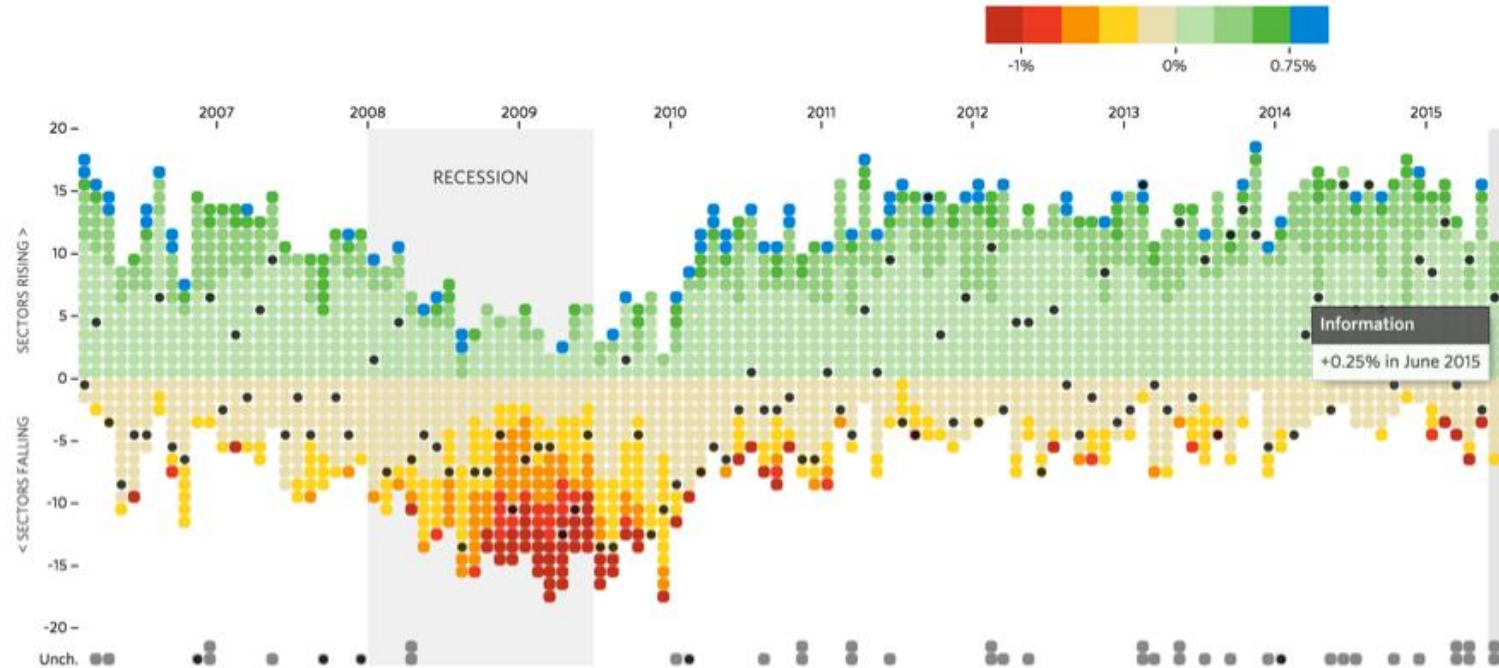
- UCI 데이터셋의 단점이 있다면, 대부분 1900년대의 오래된 데이터가 대부분이므로 시류에 맞지 않고, 현실을 반영하는 데이터가 별로 없다는 점임

- Kaggle에서는 1000개가 넘는 더 많은 종류의 데이터를 제공하며, 최근에 화제가 되고 있는 토픽의 데이터를 다운받을 수 있다

- URL: <https://www.kaggle.com/datasets>



데이터 시각화(Data Visualization)



데이터 시각화

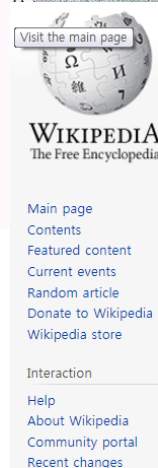
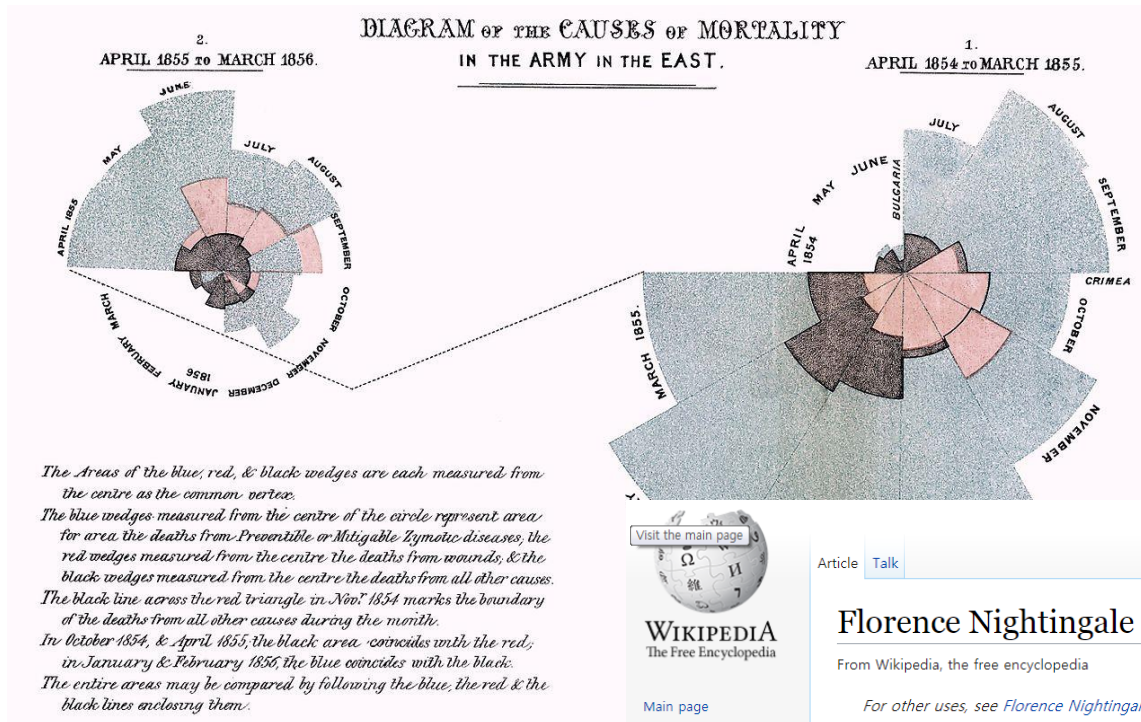
· 데이터 시각화(Data Visualization)[source: https://en.wikipedia.org/wiki/Data_visualization]

- 데이터를 점, 선 등 시각화된 객체로 인코딩해 정보를 효과적으로 전달하고자 하는เทคนิค
- 데이터 시각화의 목적은 복잡한 데이터를 보다 이해와 활용이 가능하도록 만드는 것
- 목적에 따라 바 차트(Bar chart), 히스토그램(Histogram), 산점도(Scatter plot), 네트워크 그래프(Network graph), 히트맵(Heat map) 등 다양한 시각화 도구를 활용할 수 있다



데이터 시각화

- 데이터 시각화(Data Visualization)[source: https://en.wikipedia.org/wiki/Data_visualization]



Article Talk

Read

View source

View history

Search Wikipedia

Florence Nightingale

From Wikipedia, the free encyclopedia

For other uses, see *Florence Nightingale (disambiguation)*.

"*The Lady with the Lamp*" redirects here. For the 1951 film, see *The Lady with a Lamp*.

Florence Nightingale, OM, RRC, DStJ (/ˈflɒrəns ˈnɑːtɪŋɡeɪl/; 12 May 1820 – 13 August 1910) was an English social reformer and statistician, and the founder of modern nursing.

She came to prominence while serving as a manager of nurses trained by her during the **Crimean War**, where she organised the tending to wounded soldiers.^[3] She gave nursing a highly favourable reputation and became an icon of Victorian culture, especially in the persona of "The Lady with the Lamp" making rounds of wounded soldiers at night.^{[4][5]}

Florence Nightingale
OM RRC DStJ



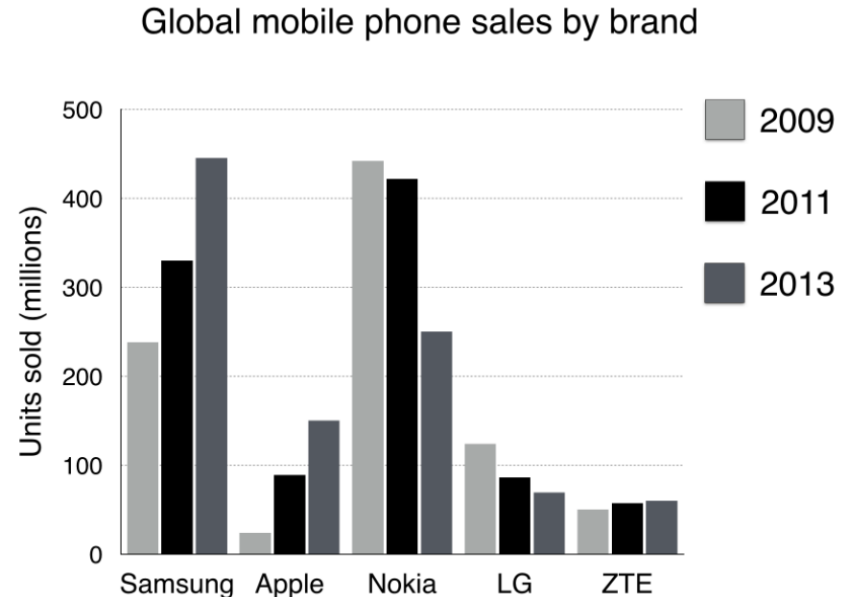
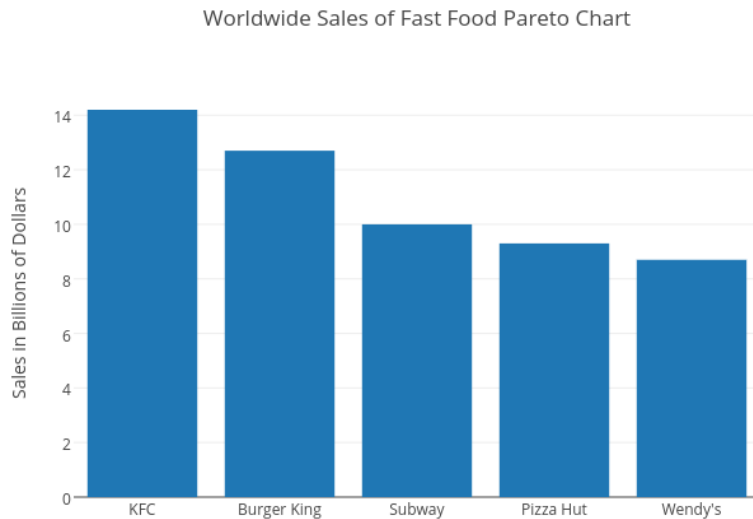
다이어그램(Diagram)

· 다이어그램(Diagram): 시각화 테크닉을 기반으로 정보를 상징적으로 표현한 것. 흔히 그래프(graph)와 동일한 의미로 쓰임[source: <https://en.wikipedia.org/wiki/Diagram>, https://en.wikipedia.org/wiki/Data_visualization]

- 막대 차트(Bar chart): 특정 그룹의 데이터를 직사각형의 막대(bar)의 길이로 표현한 차트

■ 막대는 수직적으로(horizontally) 혹은 수평적으로(vertically) 표현될 수 있다

■ 그룹 간의 값을 비교하는 데 흔히 쓰인다(예시: 기업 간 혹은 부서 간 매출액 비교)



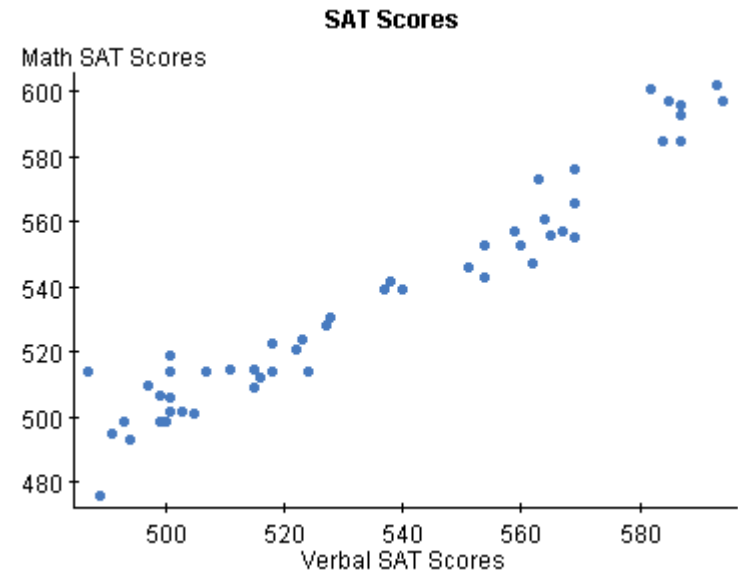
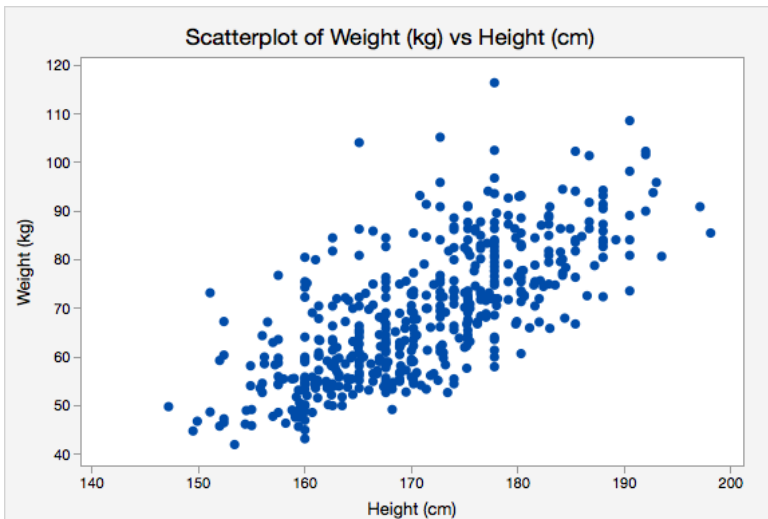
다이어그램(Diagram)

· 다이어그램(Diagram): 시각화 테크닉을 기반으로 정보를 상징적으로 표현한 것. 흔히 그래프(graph)와 동일한 의미로 쓰임[source: <https://en.wikipedia.org/wiki/Diagram>, https://en.wikipedia.org/wiki/Data_visualization]

- 산점도(Scatter plot): 주로 2차원 좌표평면에 두 변수들 간의 관계를 표현하기 위해 산개된 점으로 표현한 다이어그램

■ 일반적으로 x축이 종속 변수, y축이 독립 변수를 상징한다

■ 두 변수들 간의 선형/비선형 관계를 파악하기 위해 흔히 활용됨(예시: 키와 몸무게 간의 상관관계)



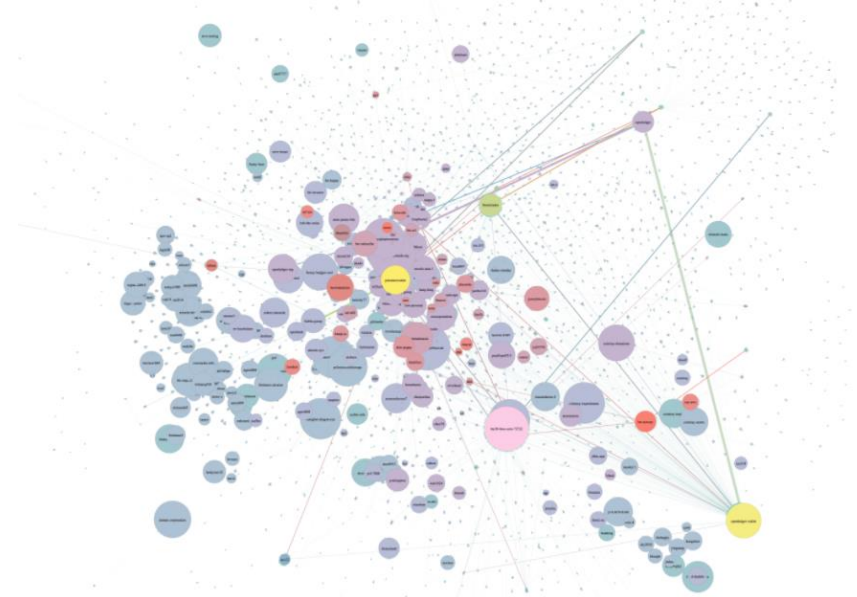
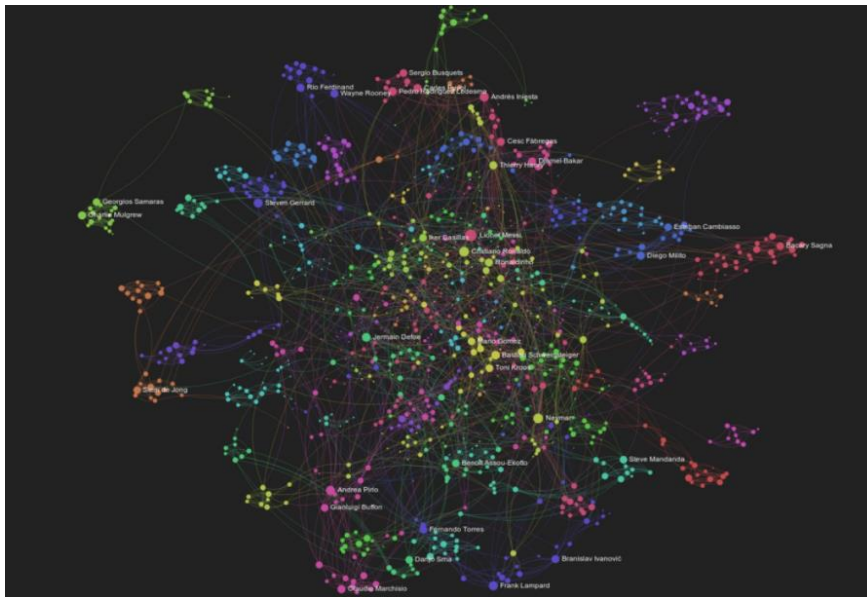
다이어그램(Diagram)

· 다이어그램(Diagram): 시각화 테크닉을 기반으로 정보를 상징적으로 표현한 것. 흔히 그래프(graph)와 동일한 의미로 쓰임[source: <https://en.wikipedia.org/wiki/Diagram>, https://en.wikipedia.org/wiki/Data_visualization]

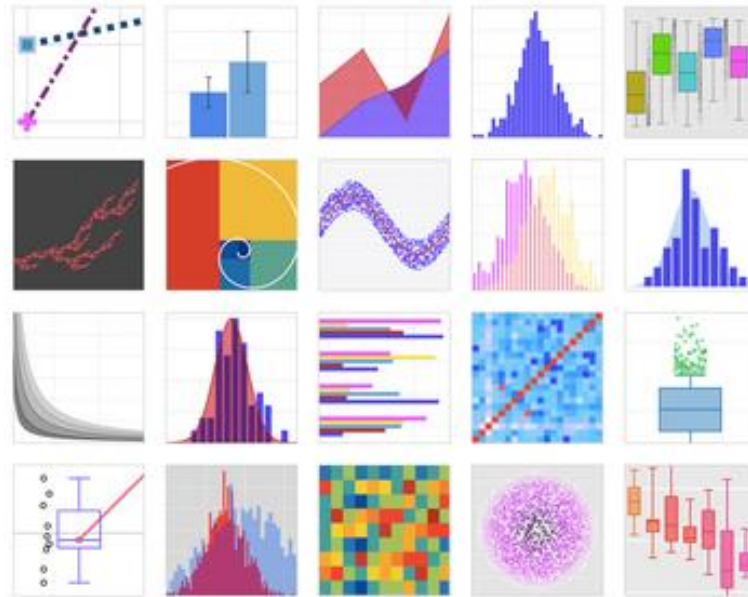
- 네트워크 그래프(network graph): 수많은 노드(node)들 간의 다양한 상호관계를 나타내기 위한 다이어그램

■ 노드의 크기, 색, 연결의 강도, 밀집도 등을 통해 관계를 다채롭게 표현 가능함

■ 몇 개의 변수로 나타내기 어려운 복잡한 사회 현상 혹은 사회연결망(social network)처럼 수많은 객체들 간의 거시적인 경향을 나타내기 위해 활용됨(예시: 온라인 커뮤니티의 연결 구조)



플로틀리(plot.ly)를 활용한 데이터 시각화

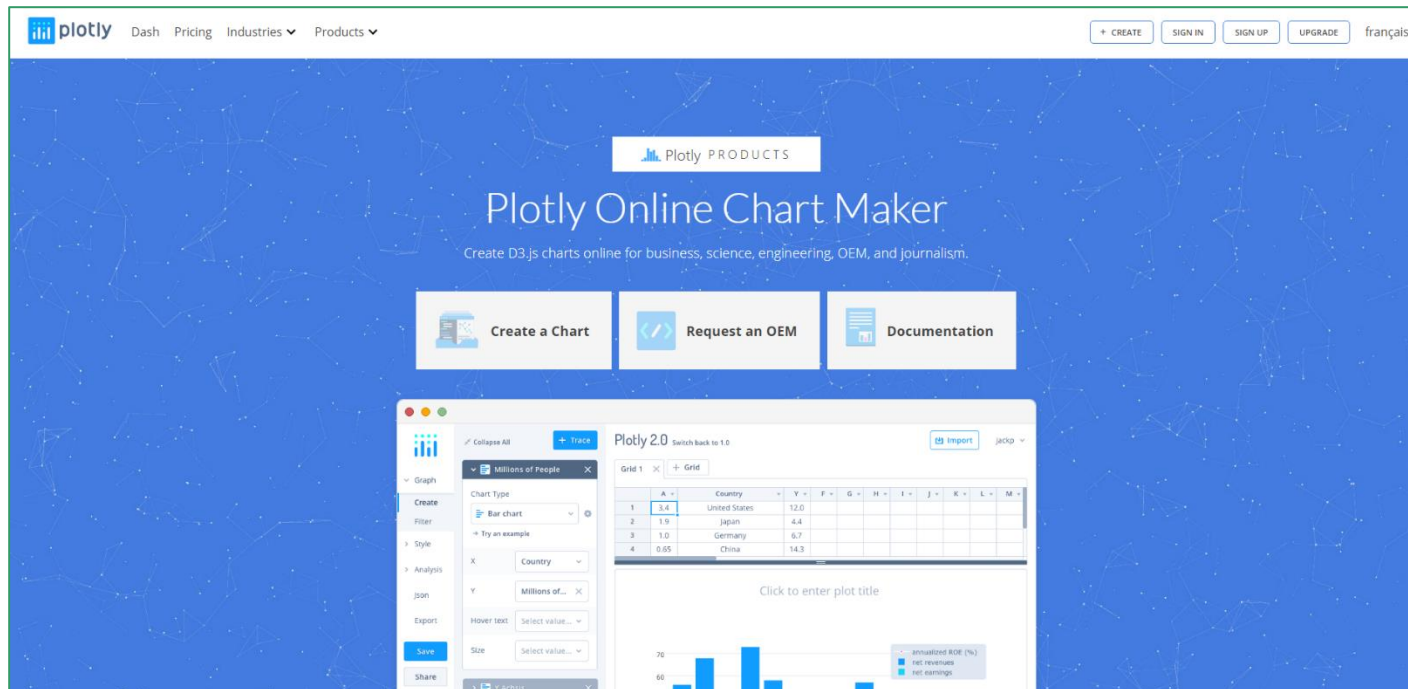


플로틀리(plot.ly)

· 온라인 상으로 차트를 손쉽게 만들고 편집할 있는 서비스를 제공함

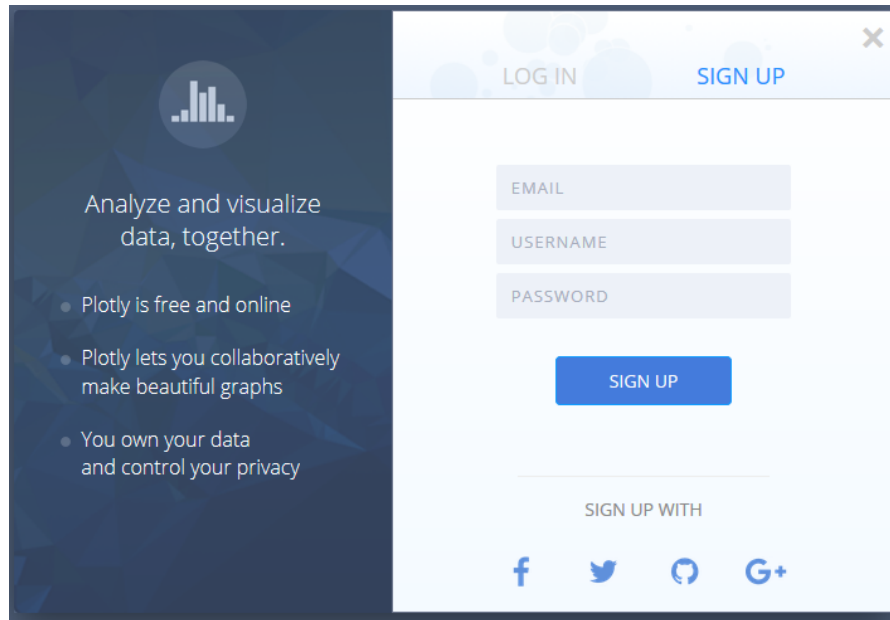
- URL: <https://plot.ly>

- 파이썬, R, 자바스크립 등 다양한 프로그래밍 언어에서도 패키지로 지원됨

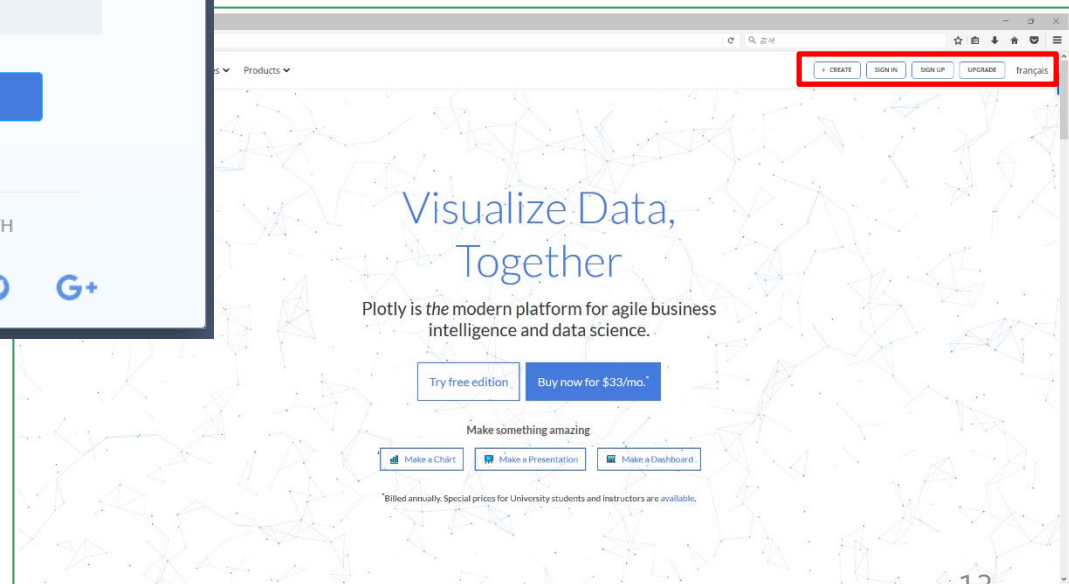


플로틀리 가입하기

- 플로틀리 일반 버전은 가입 후 무료로 사용할 수 있다
 - 메인 화면에서 상단의 'SIGN UP' 클릭 후 e-mail 주소로 가입한다



A modal window for logging in or signing up. It has a dark blue sidebar on the left with a bar chart icon and text: "Analyze and visualize data, together." Below this are three bullet points: "Plotly is free and online", "Plotly lets you collaboratively make beautiful graphs", and "You own your data and control your privacy". The main area is light blue and contains a "LOG IN" link and a "SIGN UP" link. Below these are three input fields for "EMAIL", "USERNAME", and "PASSWORD", followed by a blue "SIGN UP" button. At the bottom, it says "SIGN UP WITH" and shows icons for Facebook, Twitter, GitHub, and Google+.



플로틀리 가입하기

- 로그인 후 다음과 같은 화면이 뜨면 정상적으로 가입이 된 것이다

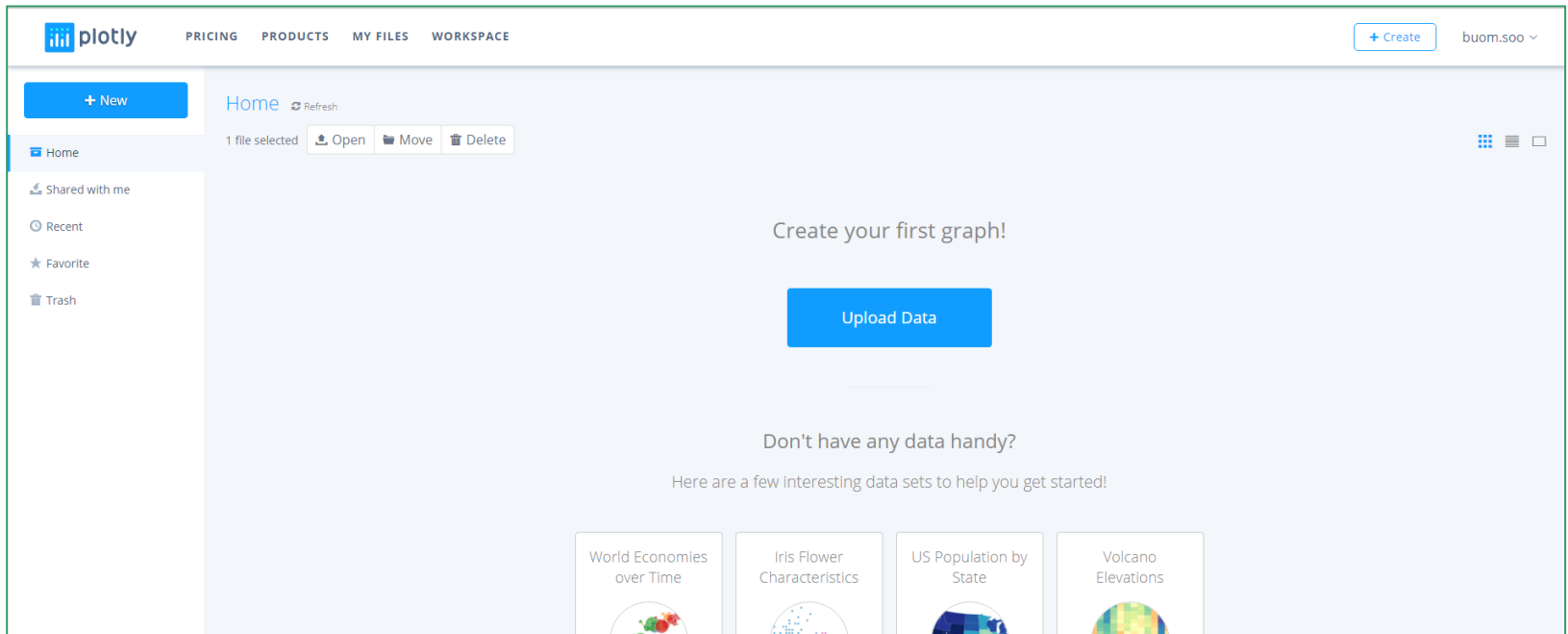
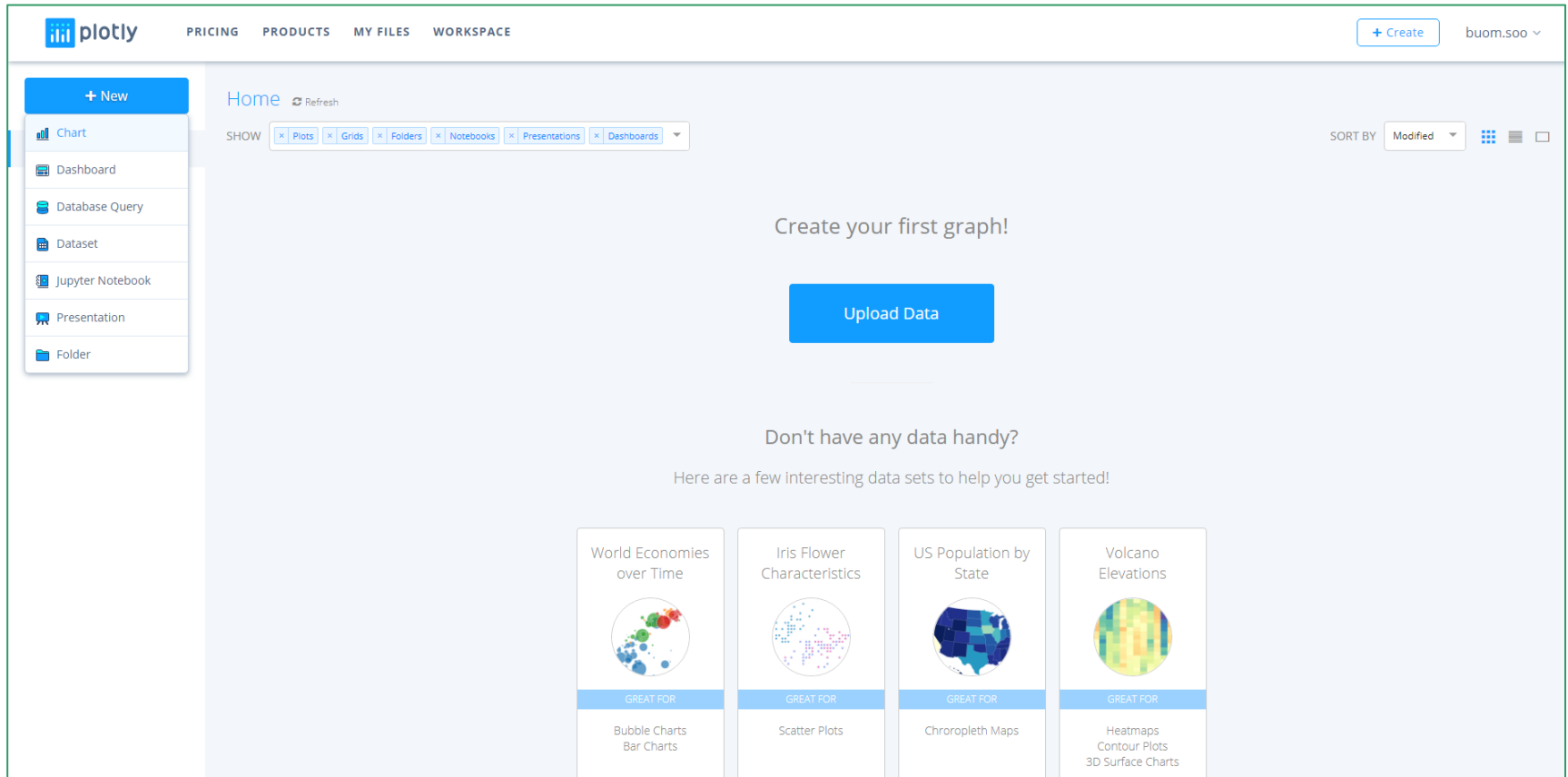


차트 그리기

- 왼쪽 상단의 [+ New] 클릭 후 'Chart'를 선택한다



데이터 셋

· 와인 품질 데이터(data.xlsx)

- 원본 URL: <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

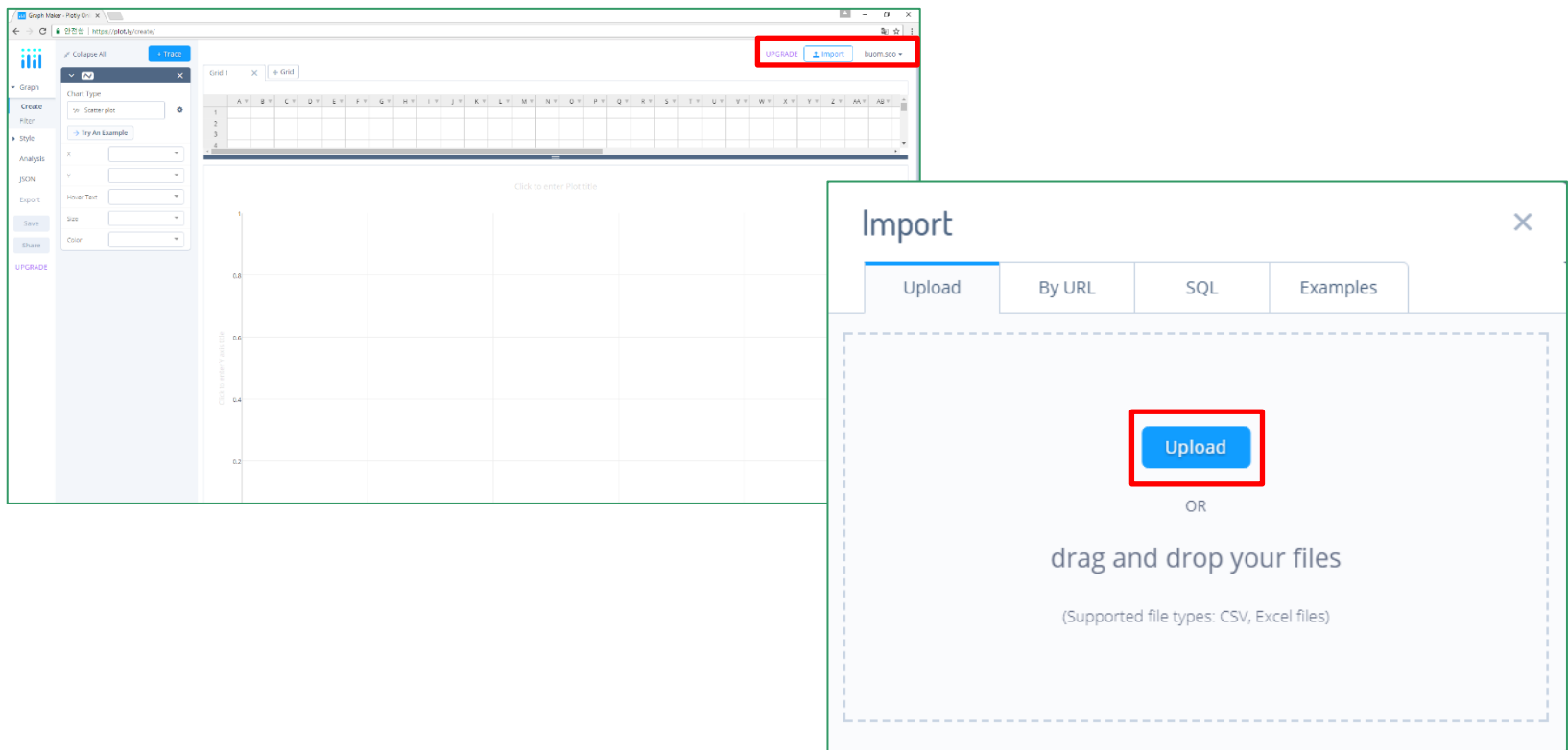
- 와인의 산도, 당도, 밀도 등을 통해 와인의 품질(quality)를 설명하고자 하는 데이터 셋

	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
11	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
12	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
16	8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
17	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
18	8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
19	8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
20	7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4

데이터 불러오기

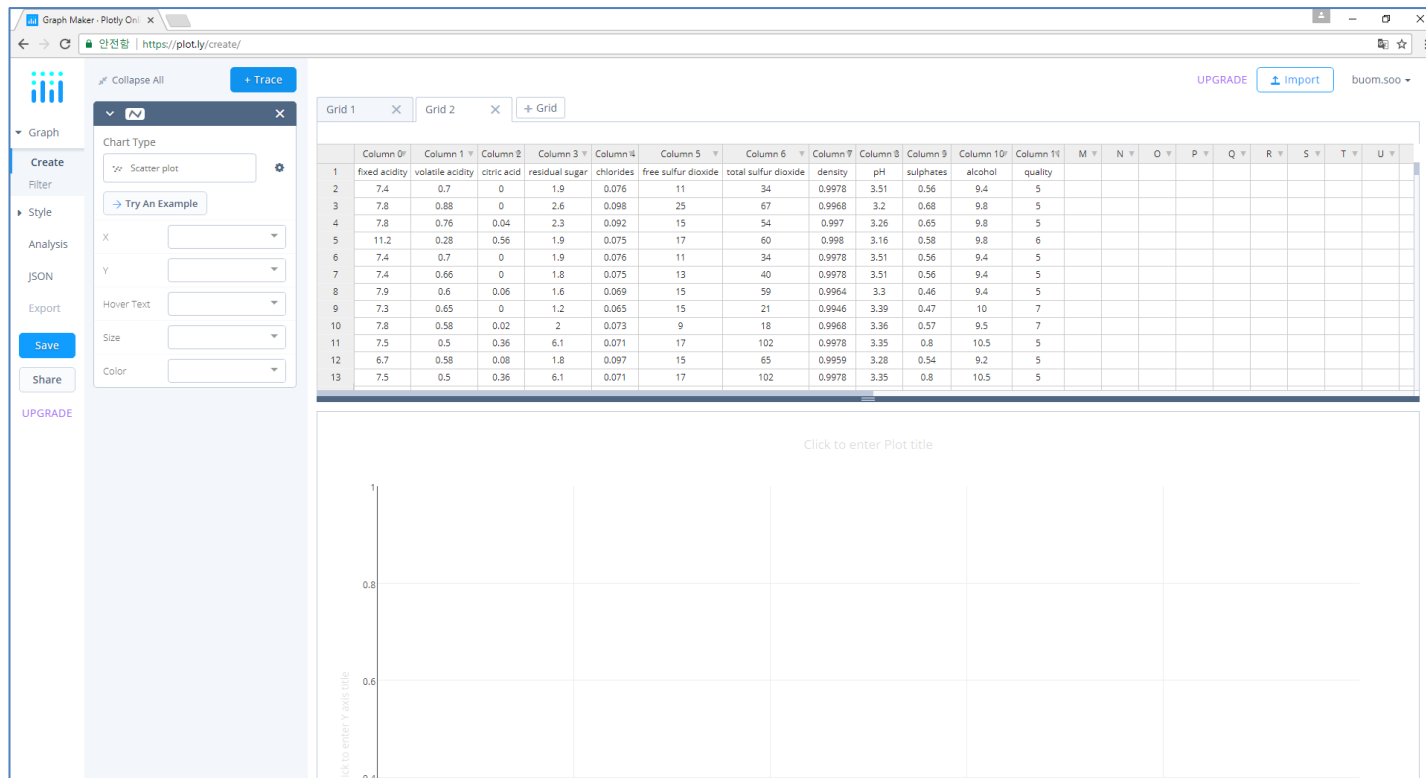
· 오른쪽 상단의 [Import]를 클릭해 데이터를 불러온다

- data.xlsx 파일을 불러온다



데이터 불러오기

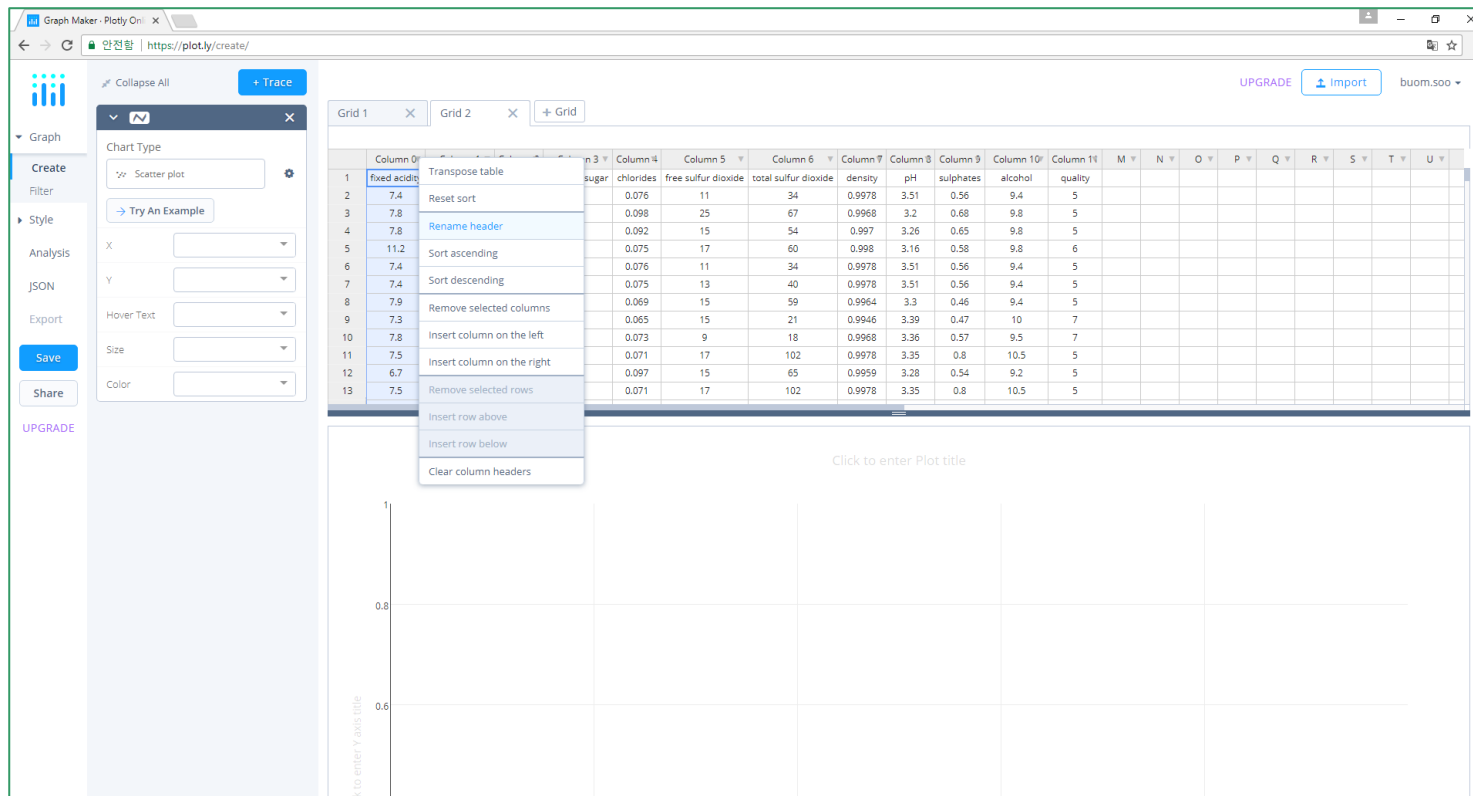
- Grid에 데이터가 추가되면 정상적으로 불러오기가 완료된 것이다
- Grid는 엑셀의 시트(데이터프레임)과 비슷한 개념이라고 생각하면 된다



- Grid의 데이터를 가지고 그래프를 그릴 수 있다

데이터 전처리

- 먼저 column 이름을 재정의하고 header 행을 삭제한다
 - 첫 번째 행을 지우고 거기 있는 이름들을 column 이름으로 옮긴다
 - Column을 누르고 [Rename header]를 클릭하면 이름을 바꿀 수 있다



데이터 전처리

· 먼저 column 이름을 재정의하고 header 행을 삭제한다

- 다음과 같이 column 이름을 바꾼다

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	M	N
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5		
2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5		
3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5		
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6		
5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5		
6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5		
7	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5		
8	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7		
9	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7		
10	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5		
11	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5		
12	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5		
13	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5		

차트 그리기

- 왼쪽 사이드바에서 'Chart type'을 선택할 수 있다
- 각 차트가 다른 용도를 가지므로 알맞게 선택하여 사용한다

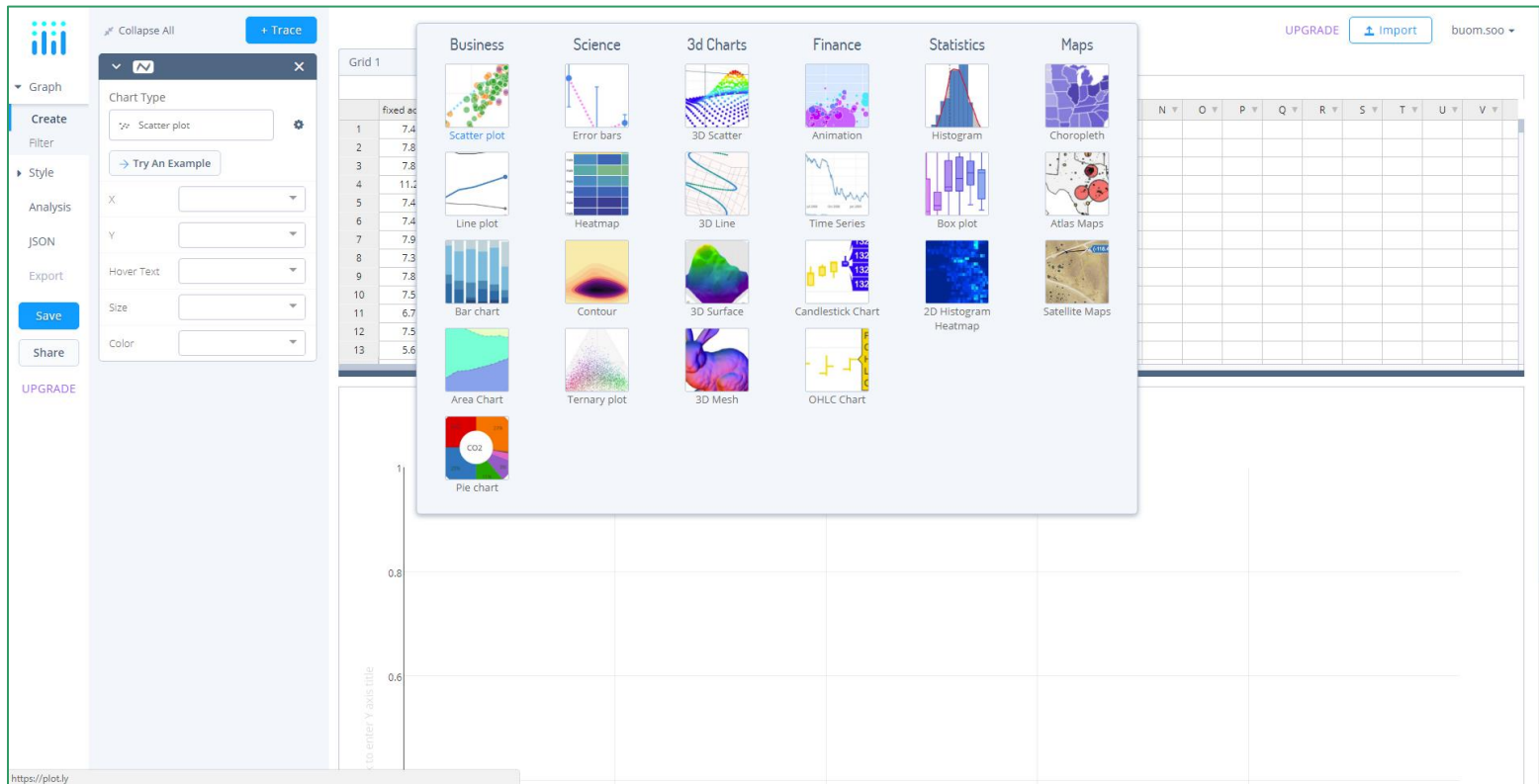


차트 그리기

· 알코올 도수(alcohol)와 와인 품질(quality) 변수 간의 산점도(scatterplot)를 그려보자

- 산점도(scatterplot): 2차원 공간상에 산개된 점(point)들로 두 변수 간의 관계를 나타내는 차트

[source: https://en.wikipedia.org/wiki/Scatter_plot]

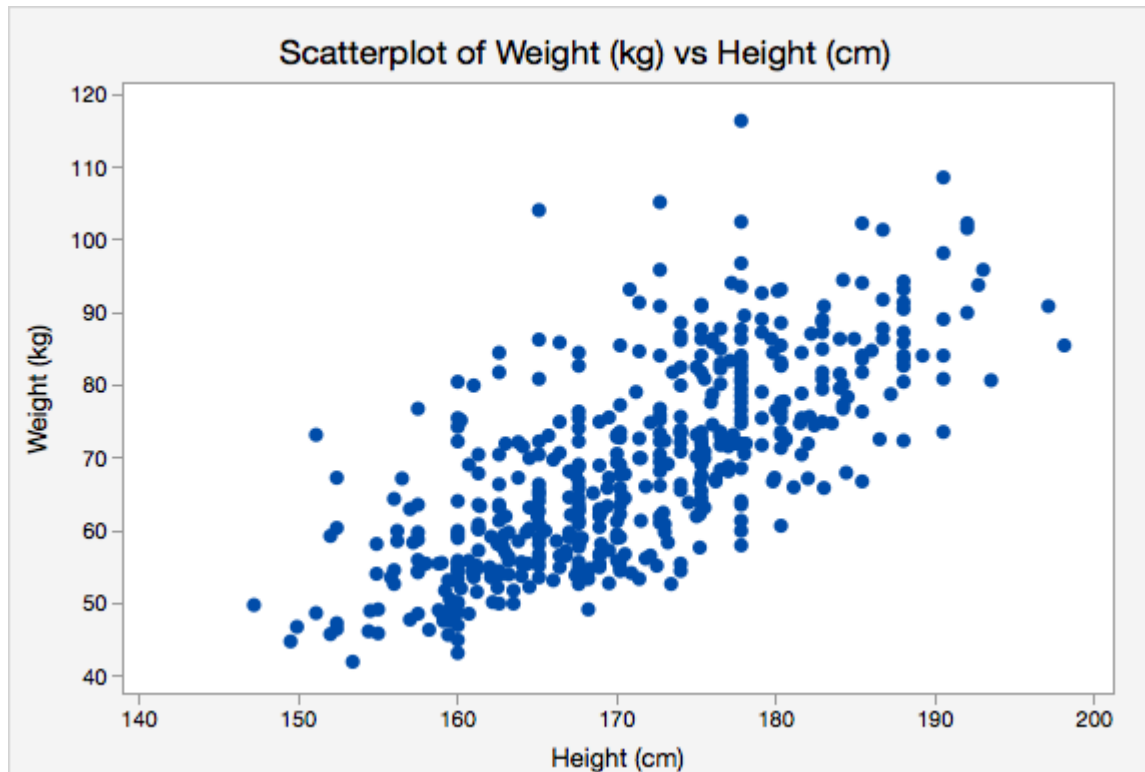


차트 그리기

- 알코올 도수(alcohol)와 와인 품질(quality) 변수 간의 산점도(scatterplot)를 그려보자
 - 좌측 사이드바에서 X는 'alcohol'로 Y는 'quality'를 선택한다

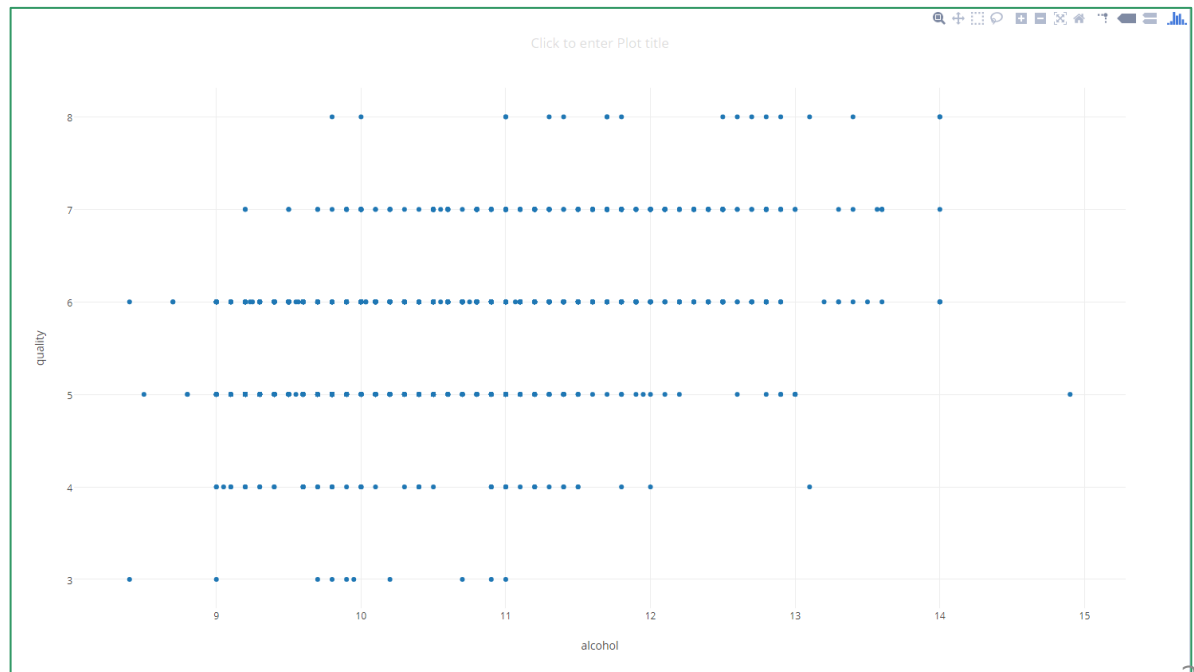
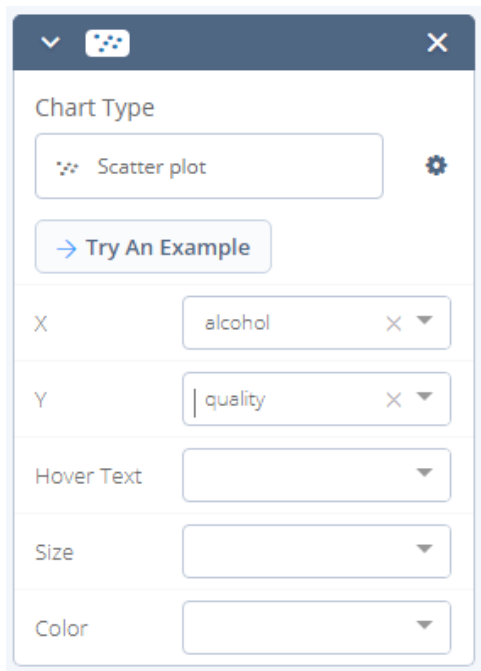


차트 그리기

- 알코올 도수(alcohol)와 와인 품질(quality) 변수 간의 산점도(scatterplot)를 그려보자
 - 각 점의 사이즈(size)와 색(color)를 특정 변수 값에 따라 달라지게 그릴 수도 있다

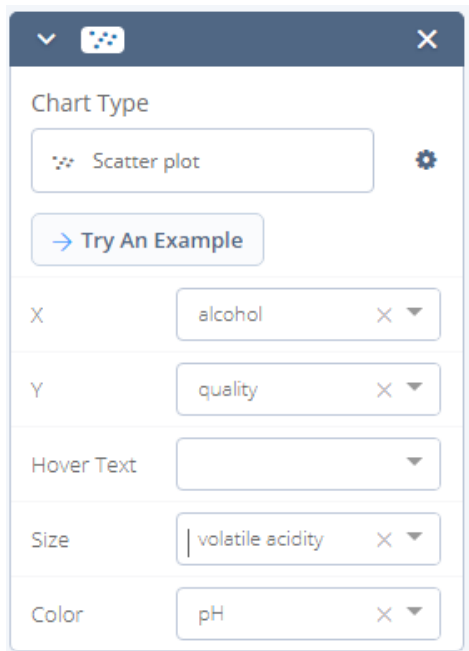


차트 그리기

- 알코올 도수(alcohol)와 와인 품질(quality) 변수 간의 산점도(scatterplot)를 그려보자
 - 산점도의 제목(title)을 설정할 수 있다(클릭 후 입력)

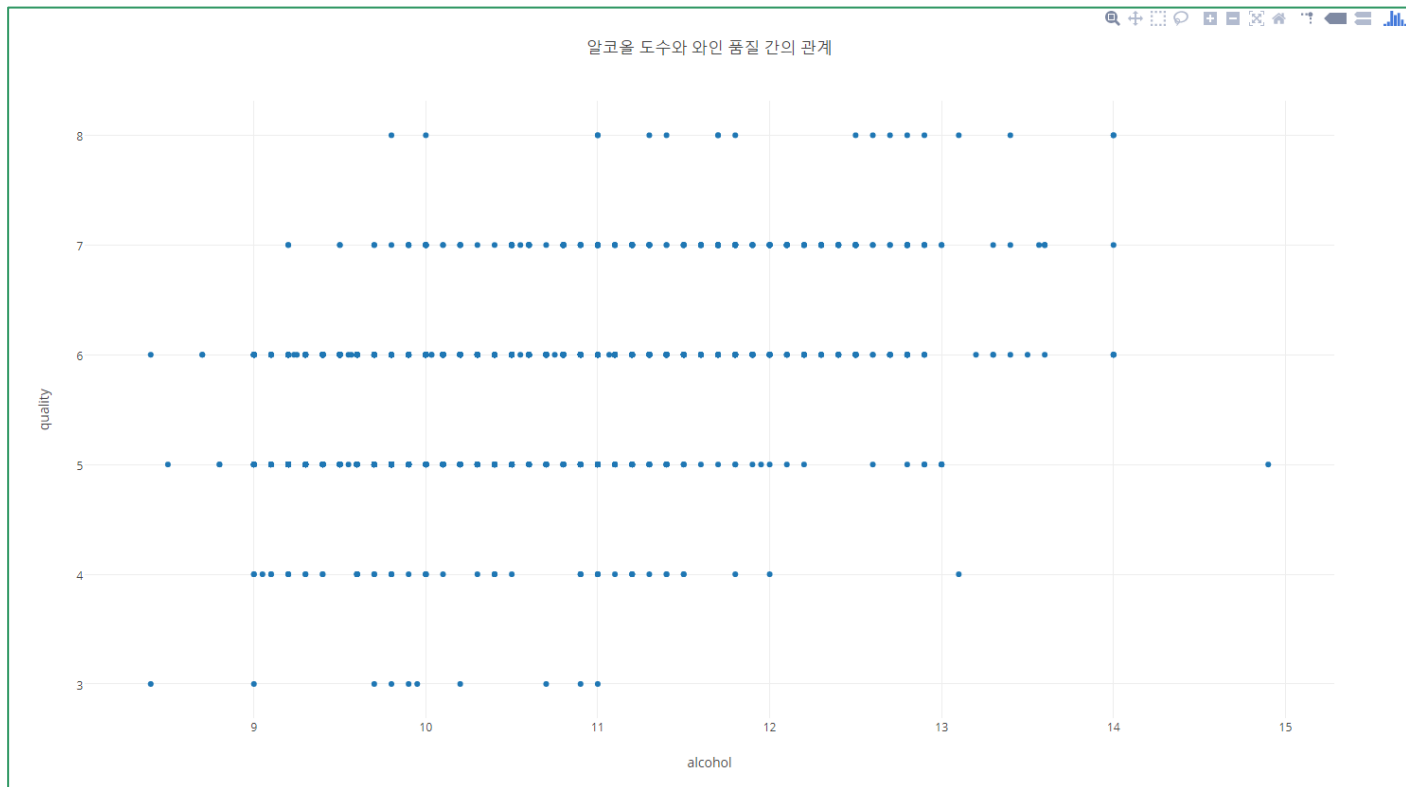


차트 그리기

- 왼쪽의 [Save] 버튼을 통해 내 차트 목록에 저장할 수 있다
 - Public하게 설정할 경우 다른 사람도 나의 차트를 볼 수 있다(공유 기능)

Save

×

Your plot will be private

PLOT

filename

DATA (1)

Grid 2

Public

Private Link

Private

☐

☐

☒

Cancel

Save

UPGRADE

차트 찾기

- 플로틀리에서는 다른 사람들이 공유한 차트도 참고할 수 있다

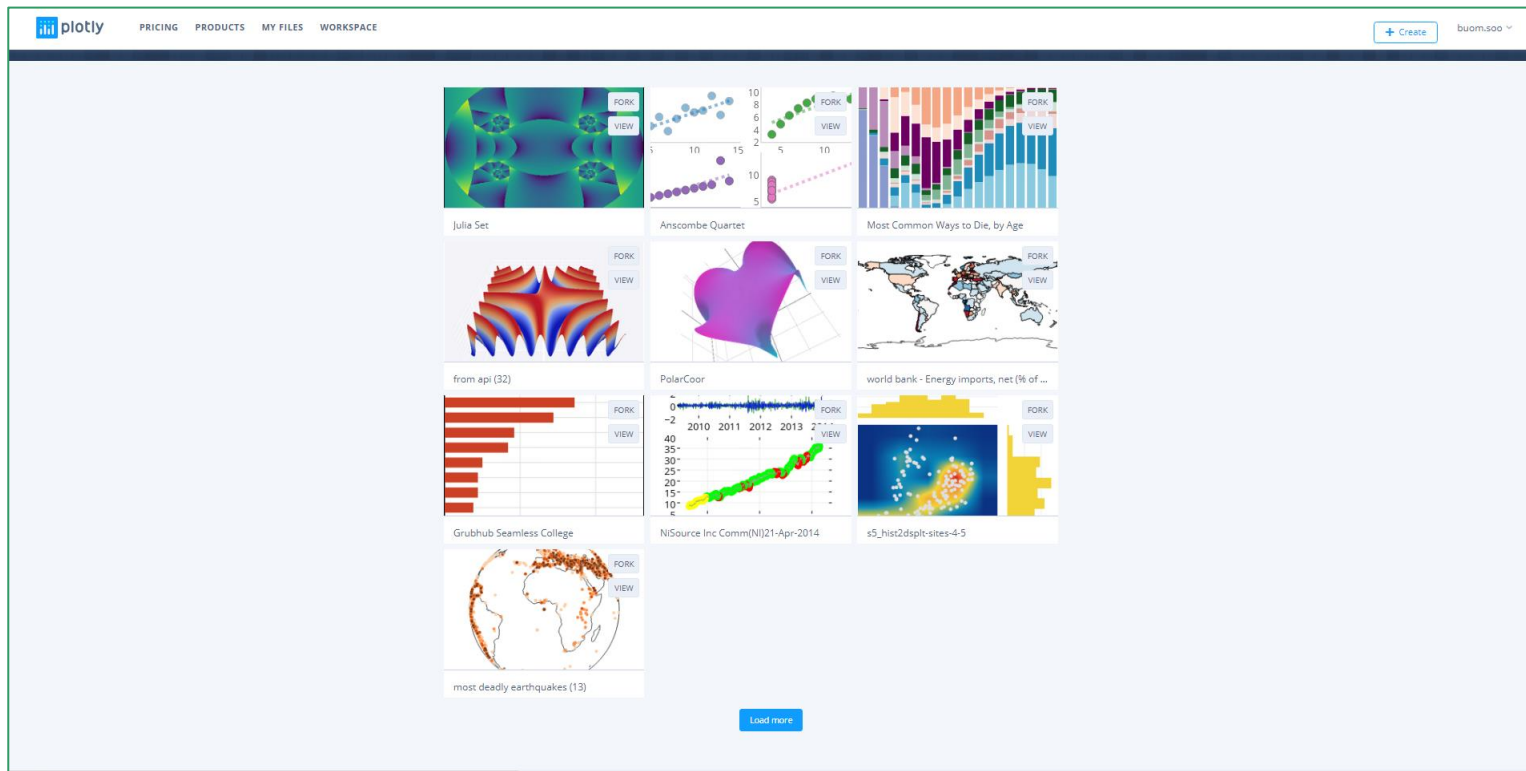


차트 찾기

- 메인 페이지에서 검색어나 조검을 통해 검색하면 참고하고 싶은 차트를 쉽게 찾을 수 있다

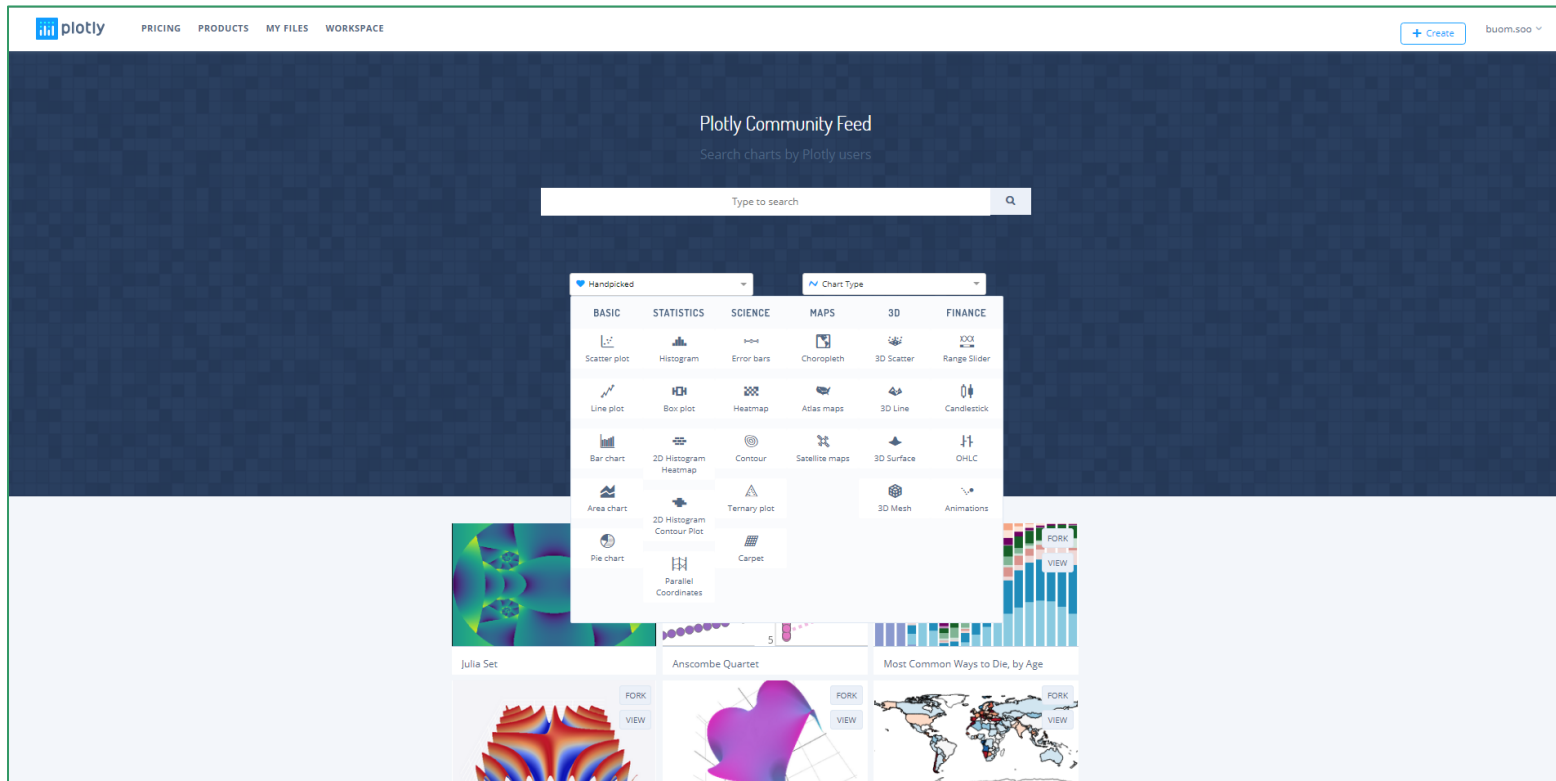
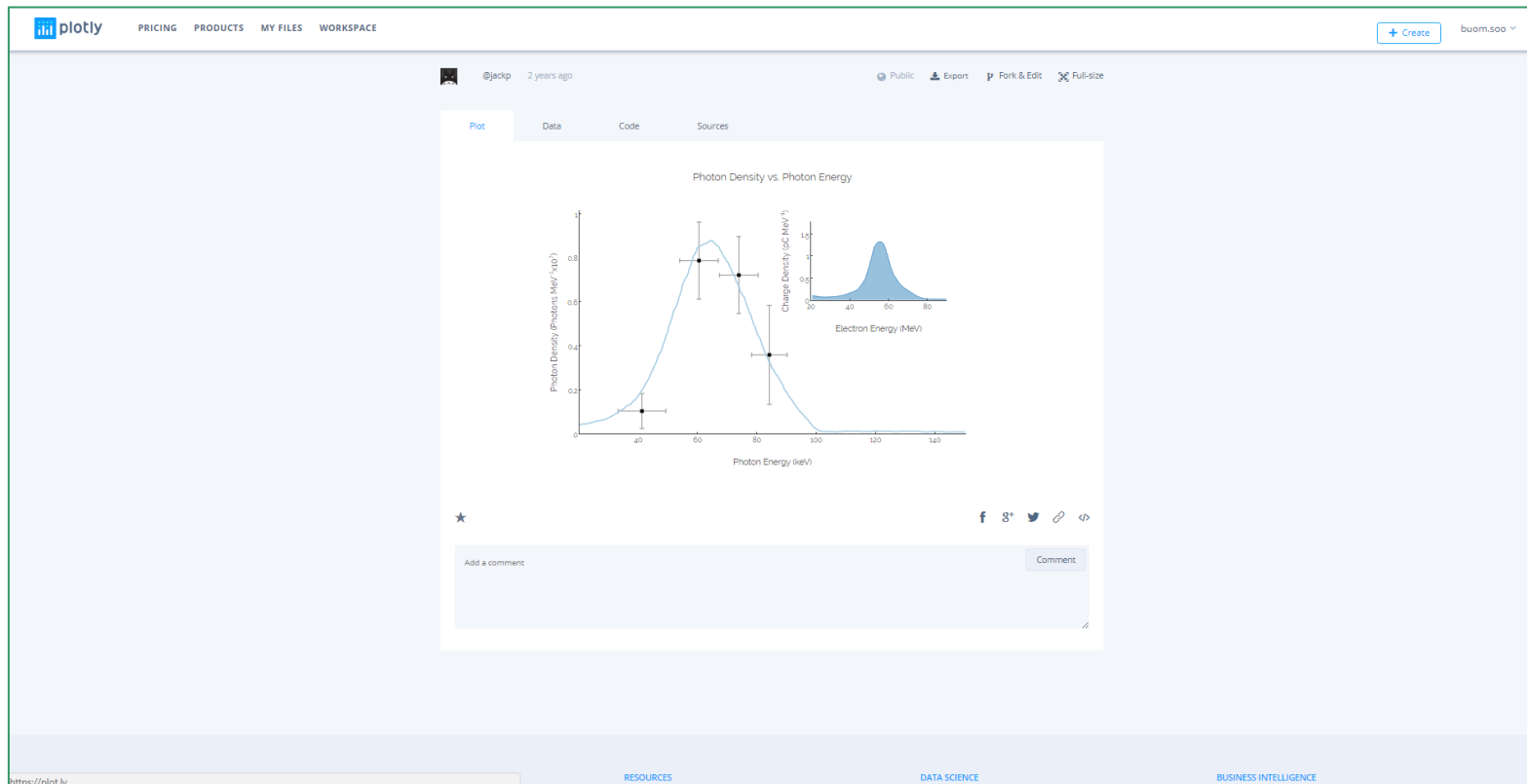


차트 찾기

· 다른 사람이 만든 멋진 그래프를 참고하여 그래프를 쉽고 빠르게 그릴 수 있다

- There is no need to “reinvent the wheel”



실습

- UCI나 Kaggle의 데이터셋, 혹은 본인이 관심 있는 데이터를 가지고 자유롭게 시각화를 실습해 본다

