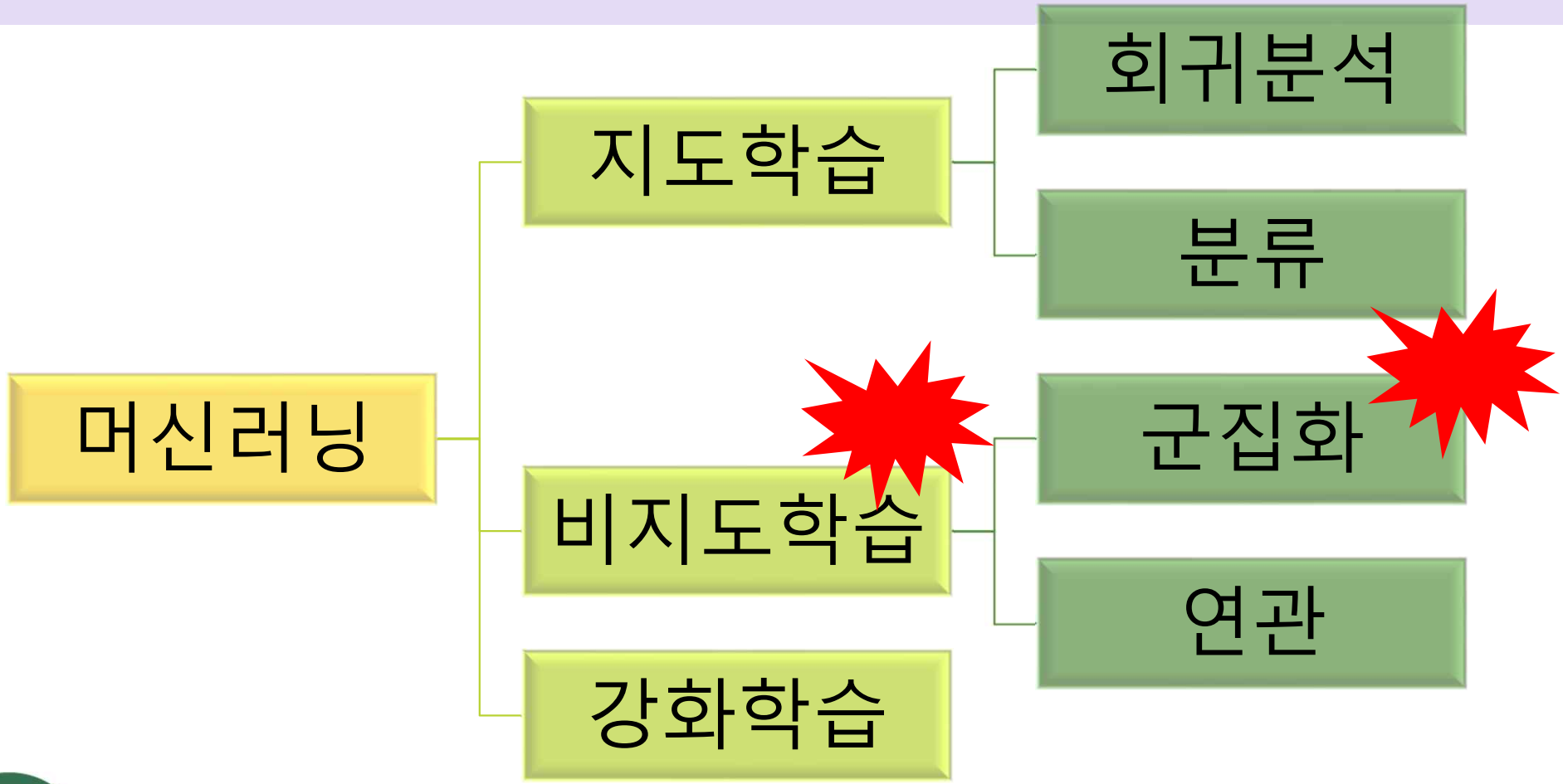
orange 활용 데이터 분석 및 머신 러닝

6차시

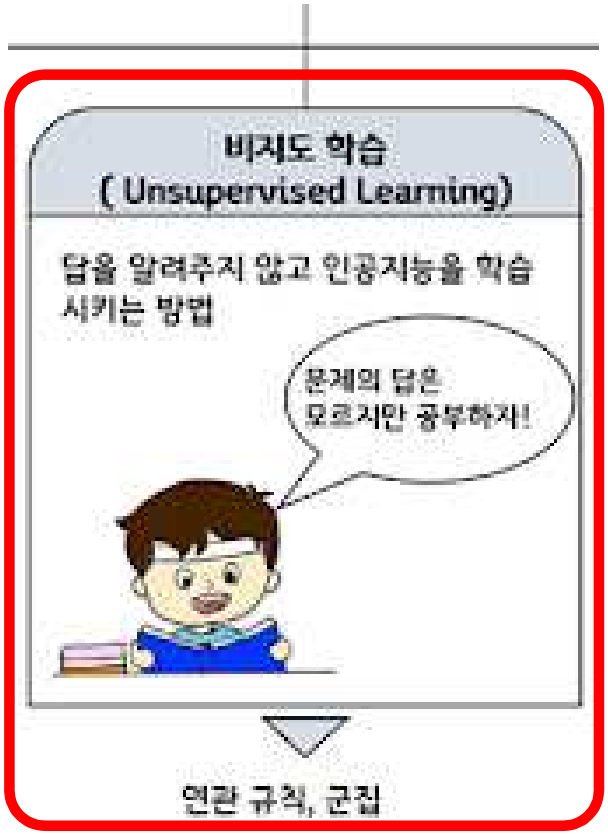
정형데이터를 활용한 군집화 1

K-means와 Hierarchical clustering 모델을 활용한 군집화


머신러닝
지도학습
비지도학습
강화학습
회귀분석
분류
군집화
연관

군집화에 활용하는 데이터의 성질

Feature
Attributes
독립변수
문제
: 정형/비정형데이터

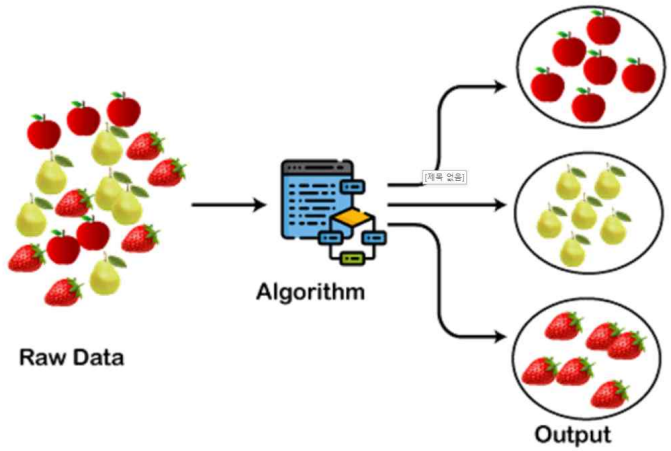


Target : 없음

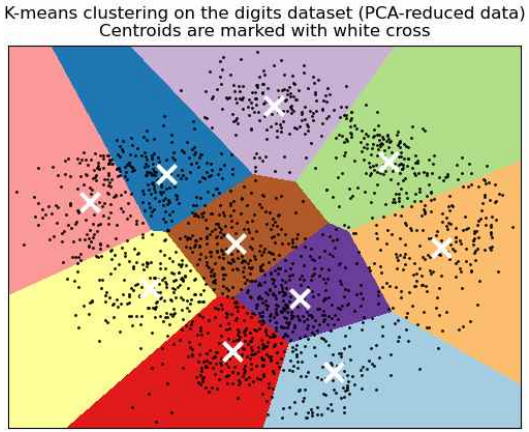
그림 http://ai4school.org

군집화(Clustering) 란?

- **대표적인 비지도학습** 방법으로 문제(데이터)를 주지만 답(레이블)은 주지 않고 데이터의 속성이 유사한 것끼리 모아 분류하는 기법
- **군집화(Clustering)의 주요 모델로는 K-Means (k평균 군집화), Hierarchical Clustering(계층적 군집화)**등이 있음
- 모델이 자기 나름의 계산으로 데이터들을 적절한 개수의 그룹으로 나누어 주지만, 각 그룹이 가지는 의미와 활용도를 찾아내는 것은 분석자의 몫
- **군집화는 데이터 분석 기법 자체가 중요한 게 아니라, 어떻게 변수들을 수집해 그룹을 나누고 맞춤 대응을 할 것인가를 생각해내는 분석가의 통찰이 훨씬 중요하다고 할 수 있음**



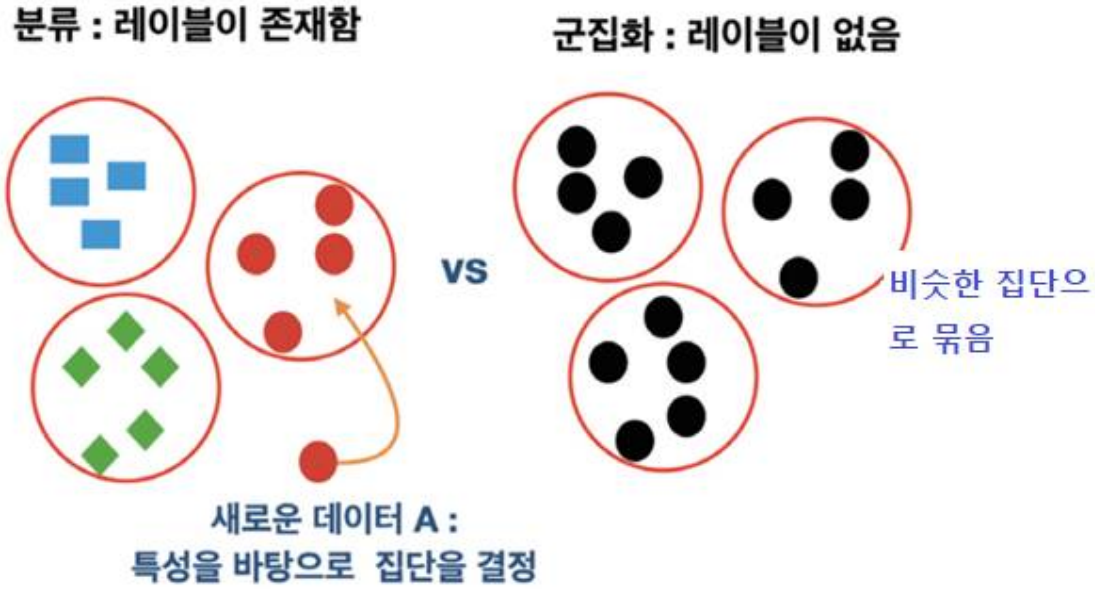
https://www.javatpoint.com/clustering-in-machine-learning



https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html

분류와 군집화의 차이

- 분류 : 서로 다른 색상과 모양(레이블)으로 표시된 소속 집단이 있을 경우 이들의 특성을 이용하여 새로운 데이터 A의 집단을 결정하는 것
- 군집화 : 새로운 데이터의 분류에 집중하기 보다는 기존의 데이터를 가까운 특성 그룹으로 나누어서 특성 그룹의 성격을 파악하는 데 유용

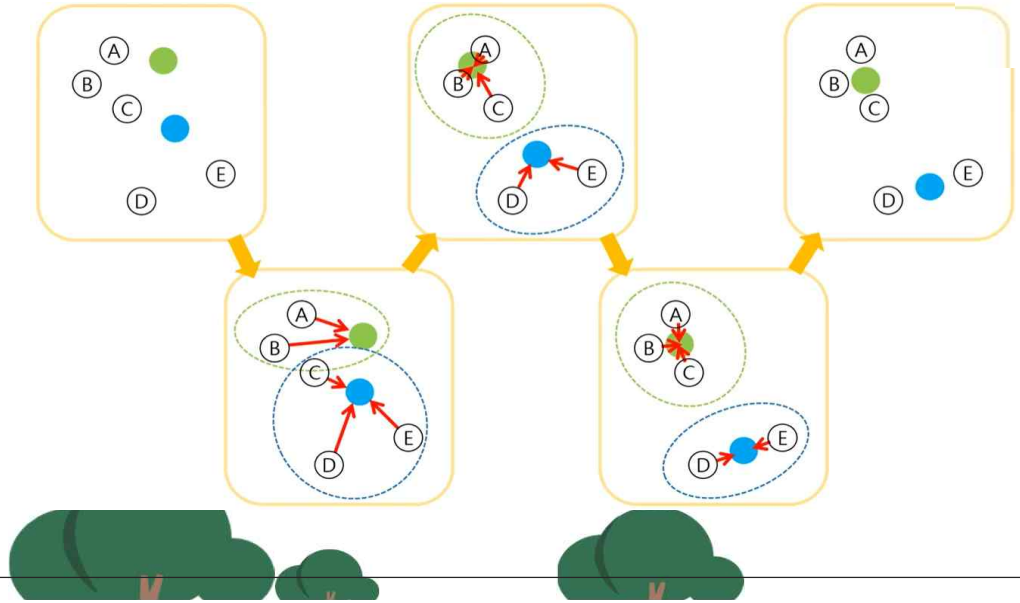


군집화의 대표 알고리즘 : k-Means

- 학습방법

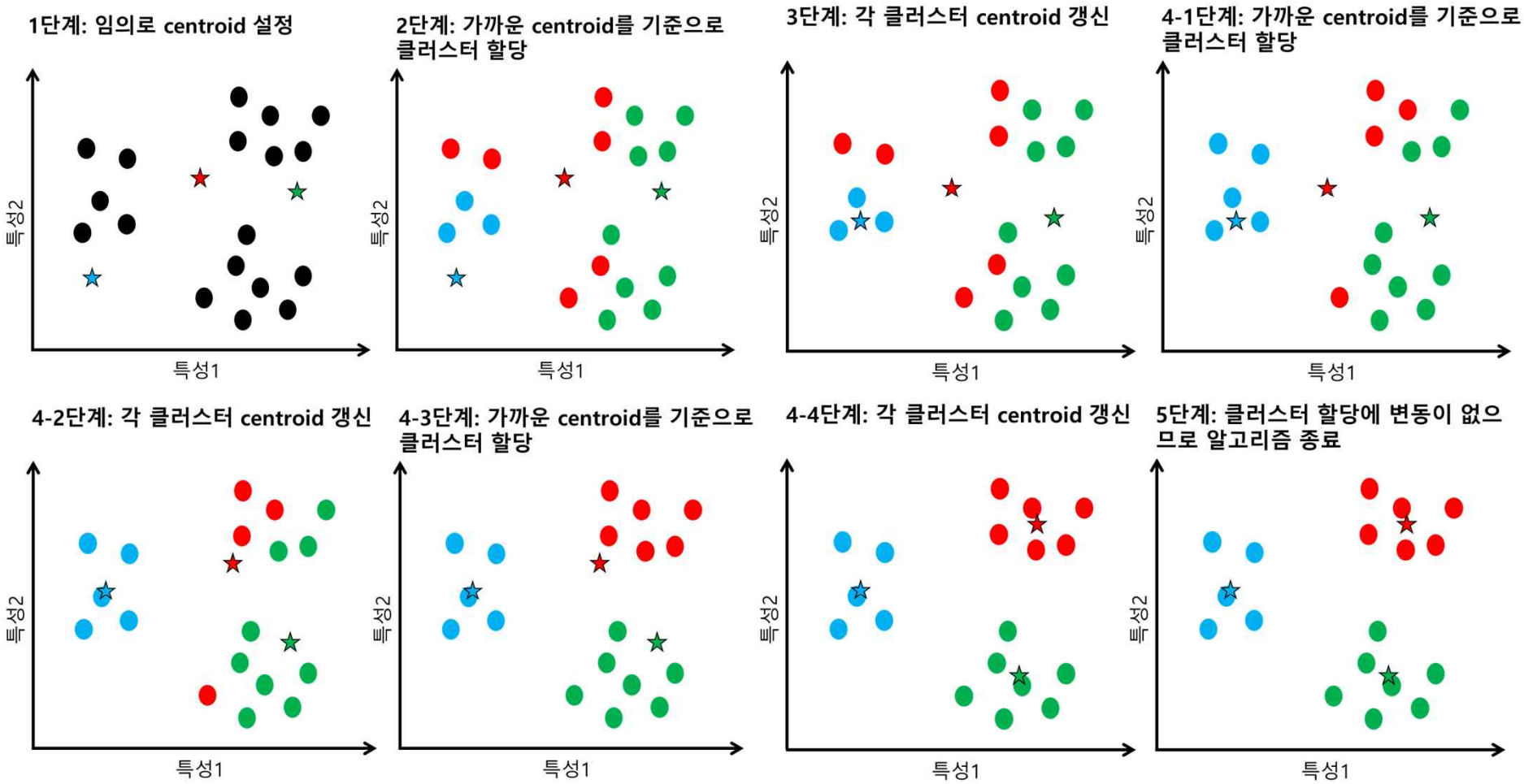
사전에 군집의 개수 k 값을 지정해야 함.

1. 군집 중심점(Centroid)을 지정해 해당 중심에서 가장 가까운 점들을 선택
2. 선택한 점들의 평균 지점으로 군집 중심점을 이동하고 다시 가까운 점들을 선택
3. 1,2 과정을 중심점이 이동하지 않을 때까지 반복



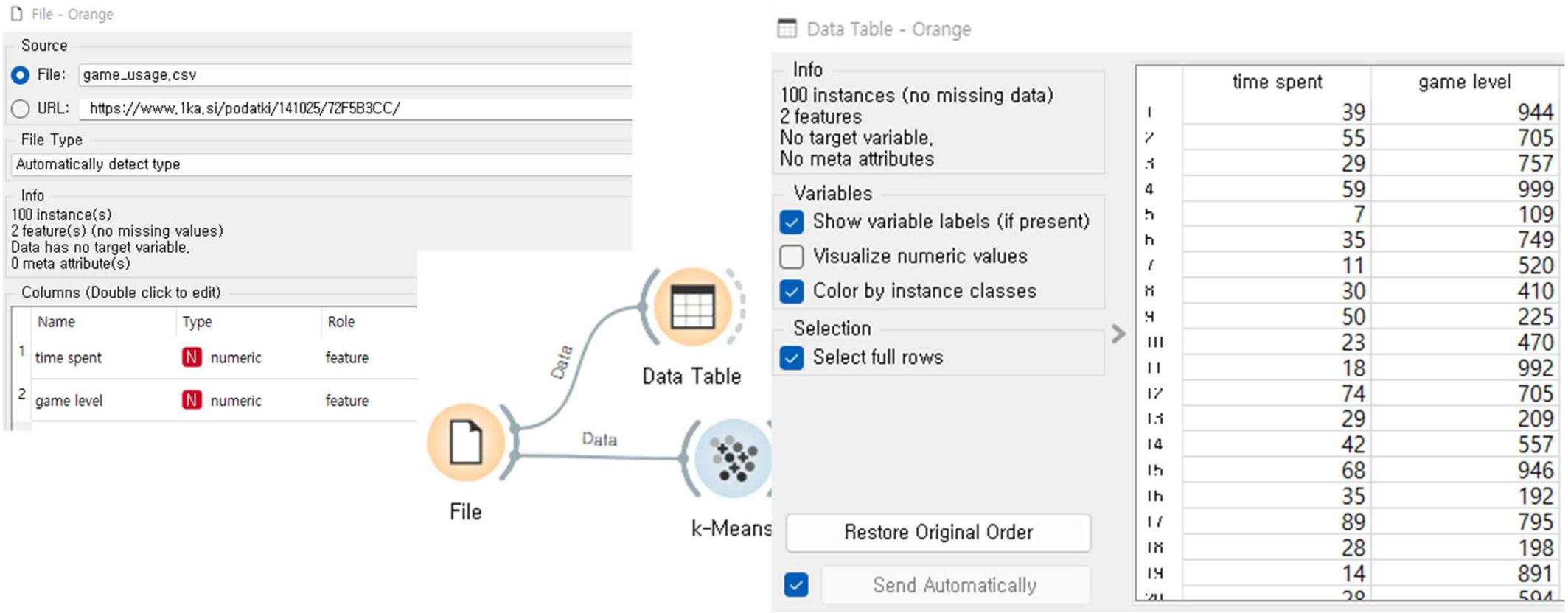
- 장점:
  - 알고리즘이 쉽고 간결
  - 대용량 데이터에도 활용가능
    (시장분석이나 컴퓨터 비전 등
    넓은 분야에서 활용)
- 단점:
  - feature 개수가 많을 경우 정확도 떨어짐,
  - 반복횟수가 많을 경우 시간 오래 걸림,
  - 몇 개의 군집을 선택해야 할지 정하기 어려움
  - 이상치 데이터에 취약

k-means 군집화



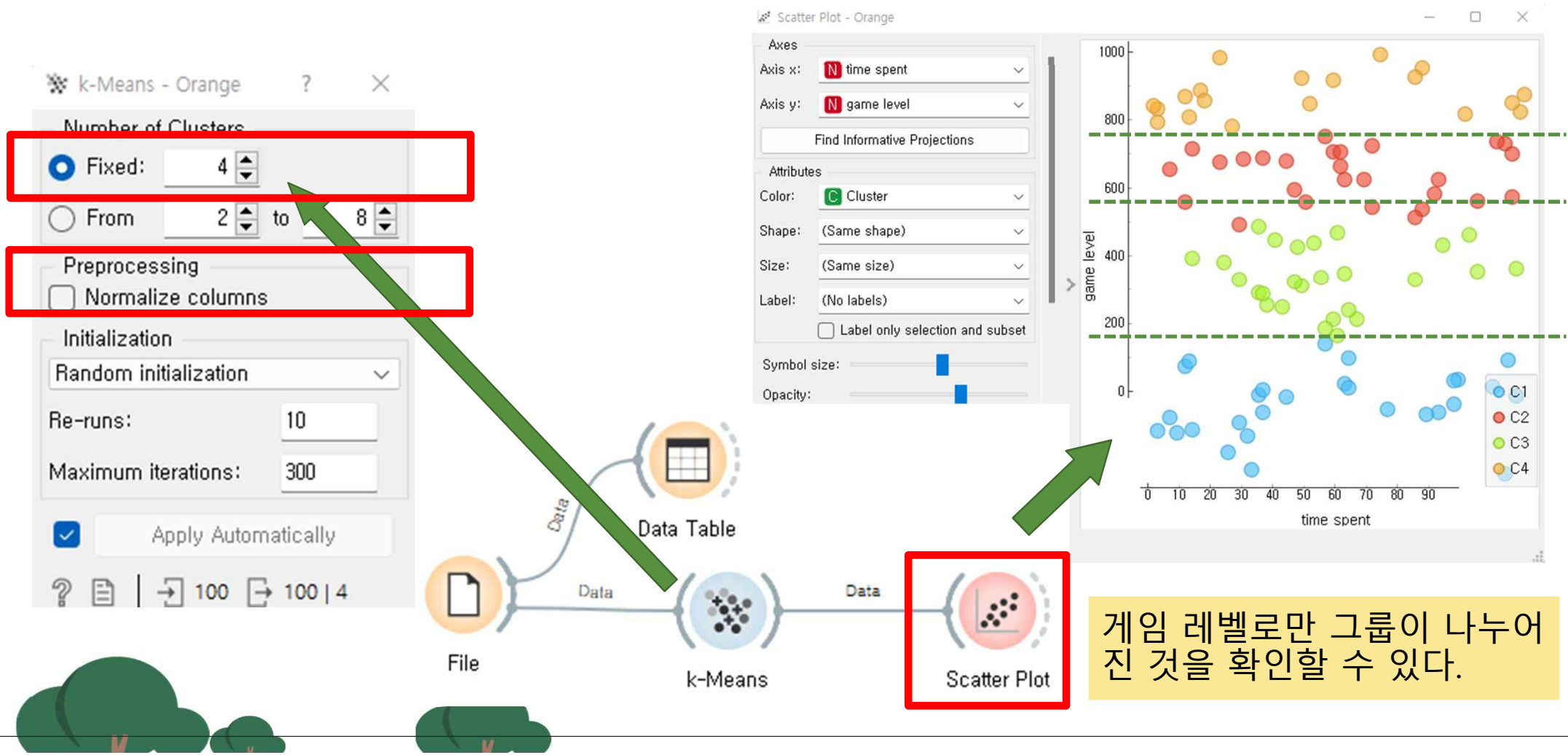
https://bskyvision.com/entry/k-means-클러스터링

게임유저를 사용시간과 레벨을 활용해 그룹지어보자.



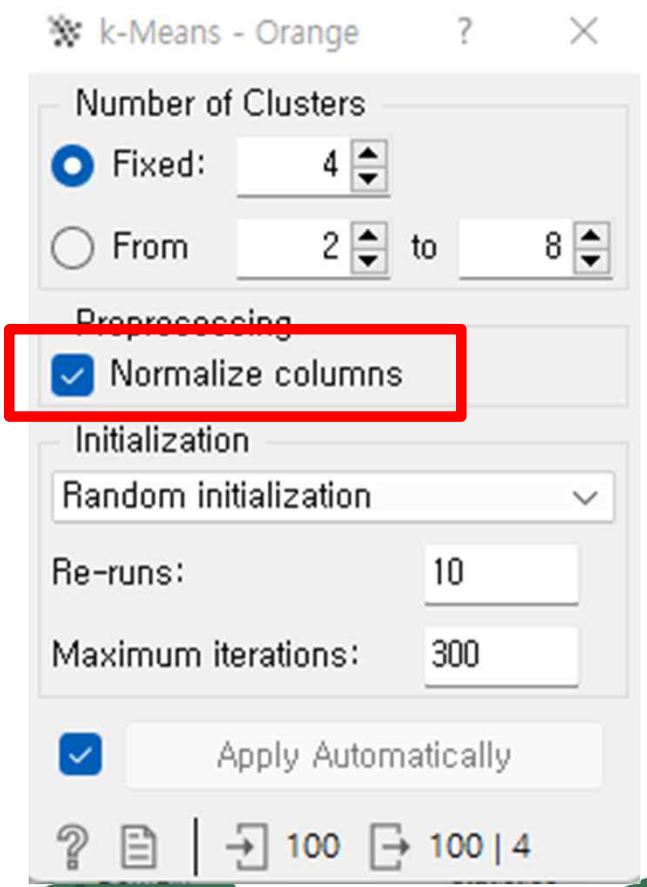
으뜸 데이터 분석과 머신러닝 박동규, 강영민 저 | 생능출판사

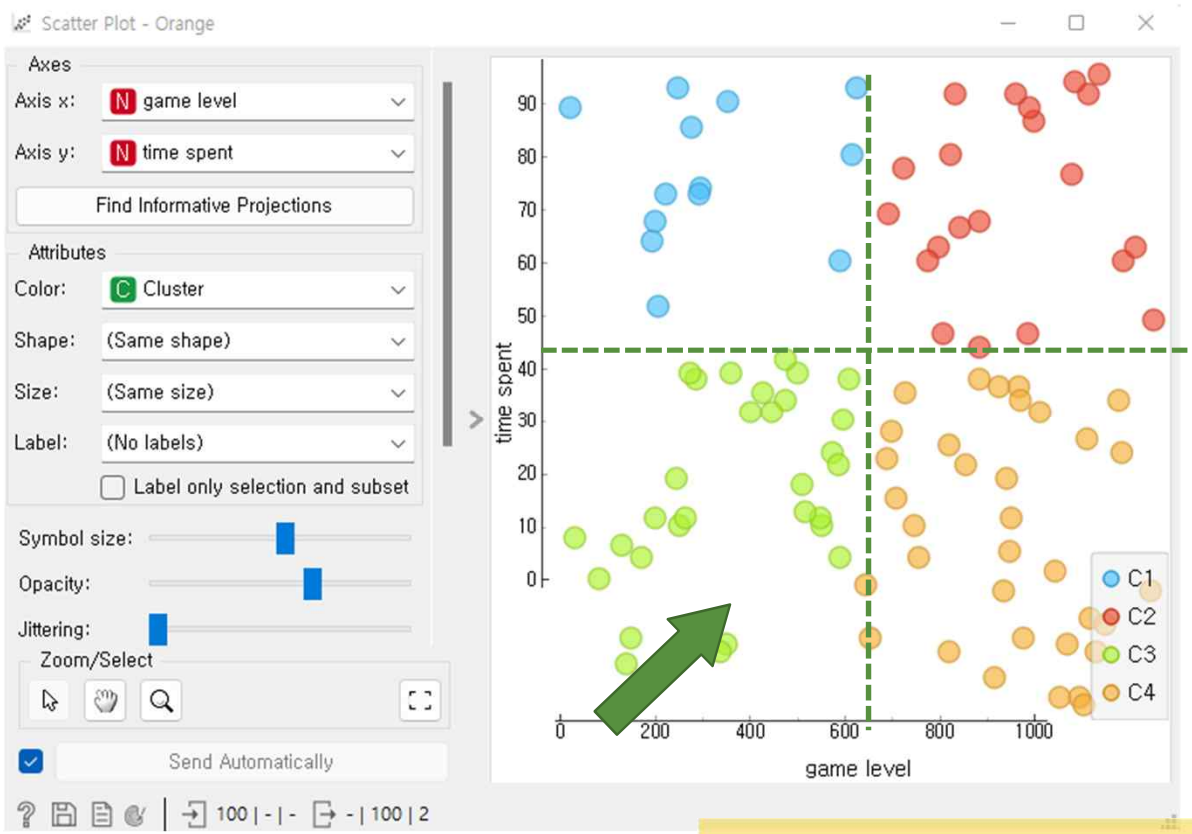
게임유저를 사용시간과 레벨을 활용해 그룹지어보자.



게임 레벨로만 그룹이 나누어진 것을 확인할 수 있다.

데이터 전처리의 필요성





게임레벨과 시간을 기준으로
4개 클러스터로 묶임

정규화 (Normalization)

- 각각의 feature 들이 갖는 값들을 동일한 규모로 변경하는 작업.
- 정규화는 모든 데이터를 0에서 1 사이의 값이 되도록 하는 것이다.
- x 데이터의 최소값이 x_{min}, 최대값이 x_{max}라고 하면 x를 정규화한 $\tilde{x}$는 다음과 같다. (Minmax scaler)

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

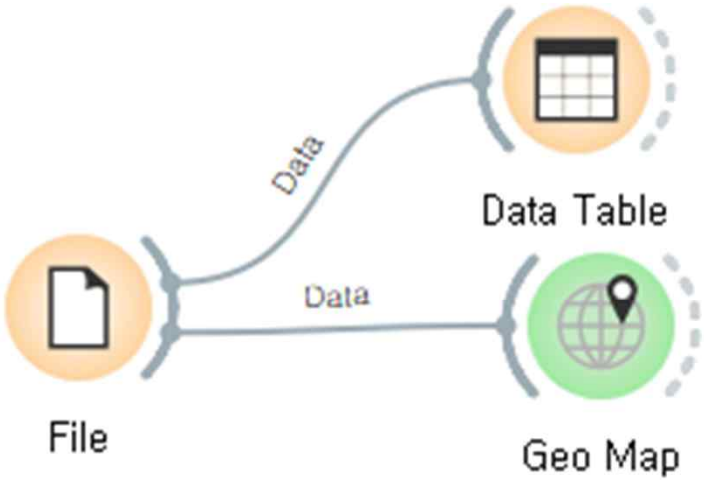
- 입력 데이터를 모두 이러한 방식으로 정규화 하면 **모든 데이터는 0에서 1 사이의 값을 갖게 된다.**

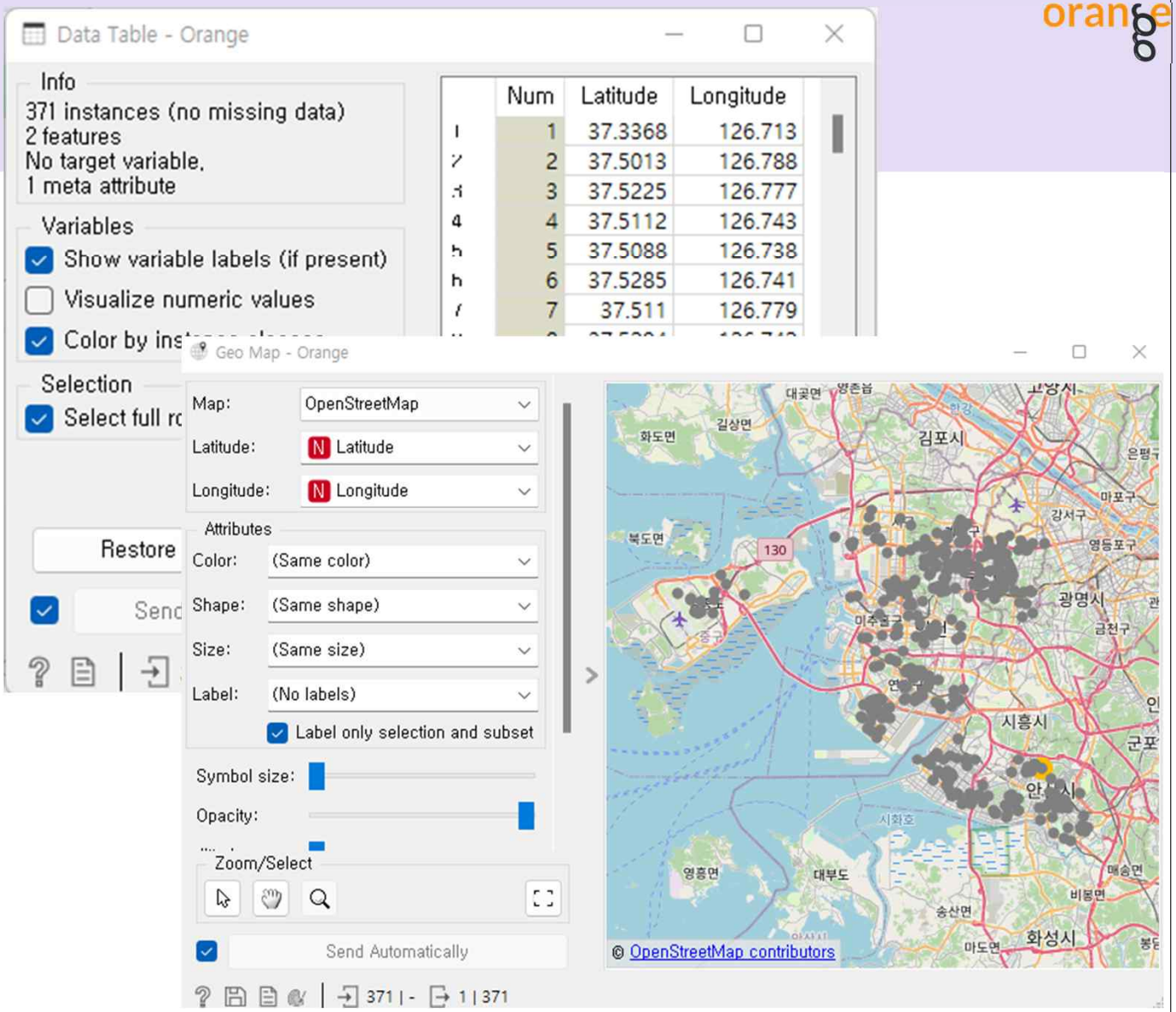
표준화 (Standardization)

- 평균과 분산을 활용하여 데이터를 정제하는 방법
  - **StandardScaler** : 각 특성의 평균을 0, 분산을 1로 변경하여 특성의 스케일을 맞춥니다.
    - 최솟값과 최댓값의 크기를 제한하지 않습니다.
    - $\frac{x-\bar{x}}{\sigma}$
  - **RobustScaler** : 평균과 분산 대신에 중간 값과 사분위 값을 사용합니다.
    - 중간 값은 정렬시 중간에 있는 값을 의미하고, 사분위값은 1/4, 3/4에 위치한 값을 의미합니다.
    - 전체 데이터와 아주 동떨어진 데이터 포인트(이상치)에 영향을 받지 않습니다.
    - $\frac{x-q_2}{q_3-q_1}$ 각각이 사분위값과 중간 값입니다.
  - **MinMaxScaler** : 모든 특성이 0과 1 사이에 위치하도록 데이터를 변경합니다.
    - $\frac{x-x_{min}}{x_{max}-x_{min}}$ 를 이용하여 변경합니다.
  - **Normalizer** : 위와 다른 스케일 조정법으로 특성 벡터의 유클리디안 길이가 1이되도록 조정합니다.
    - 즉 길이가 1인 원 또는 구로 투영하는 것이고, 각도만이 중요할 때 적용합니다.
    - l1, l2, max 옵션을 제공하며 유클리디안 커리인 l2가 기본값입니다.

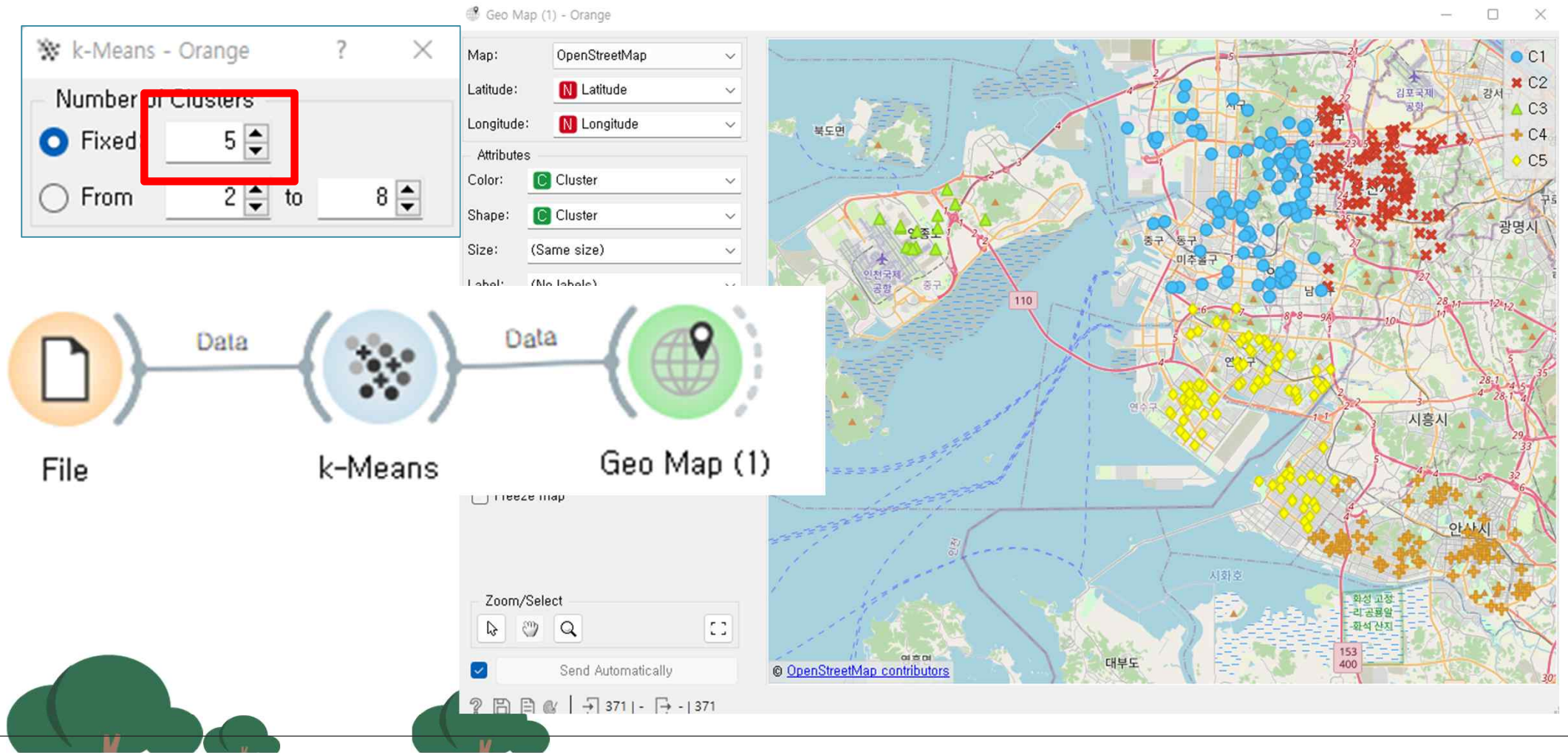
인천지역의 택배정보를 활용하여 군집화 하고 지도에 나타내 보자

데이터 불러오기


Data
Data Table
Data
File
Geo Map


Data Table - Orange
Info
371 instances (no missing data)
2 features
No target variable.
1 meta attribute
Variables
Show variable labels (if present)
Visualize numeric values
Color by ins
Selection
Select full ro
Restore
Num
Latitude
Longitude
1 37.3368 126.713
2 37.5013 126.788
3 37.5225 126.777
4 37.5112 126.743
5 37.5088 126.738
6 37.5285 126.741
7 37.511 126.779
Geo Map - Orange
Map: OpenStreetMap
Latitude: Latitude
Longitude: Longitude
Attributes
Color: (Same color)
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Opacity:
Zoom/Select
Send Automatically
371 | -
1 | 371
© OpenStreetMap contributors

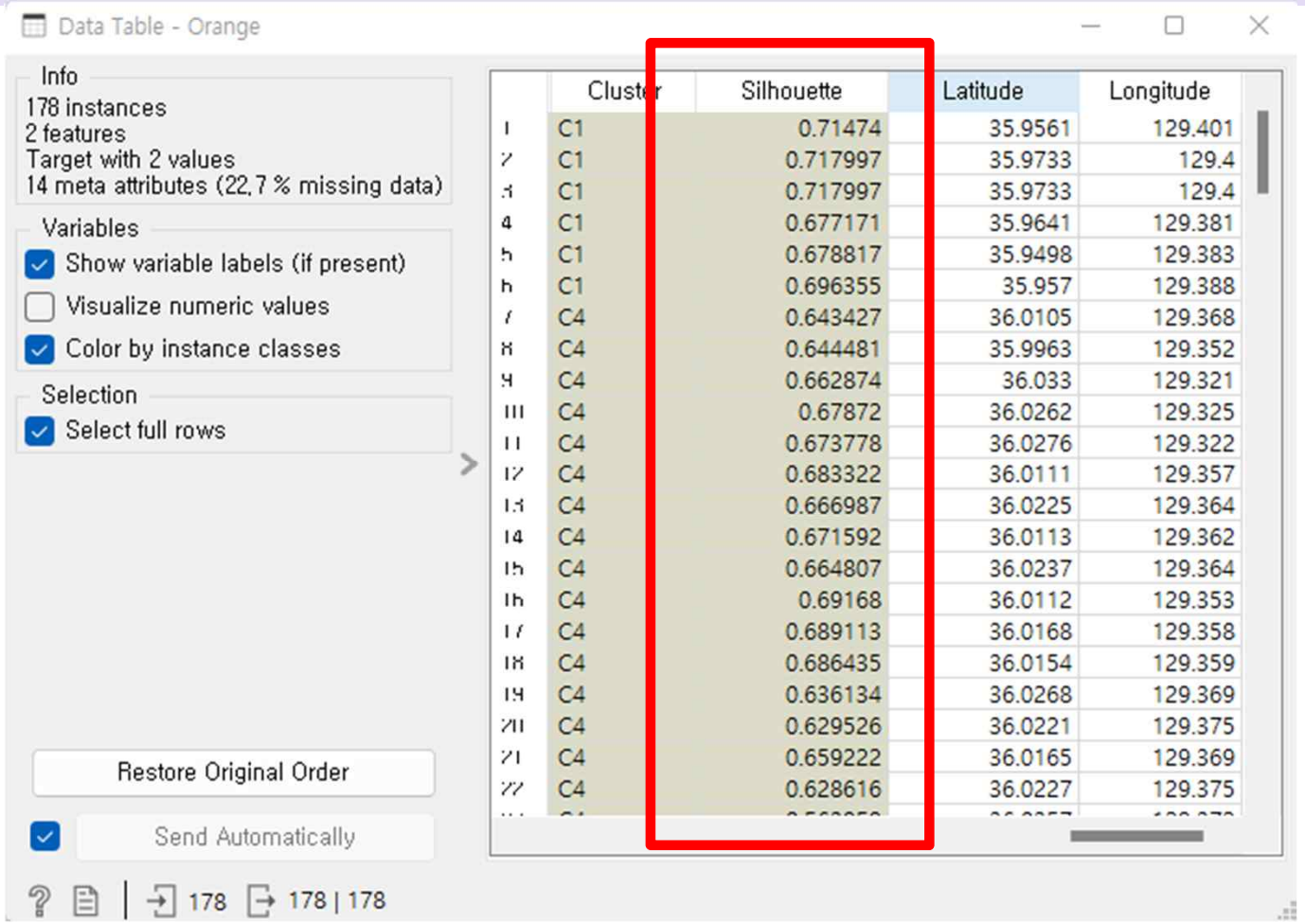
K-Means를 활용하여 군집화 하기



k-Means 군집화의 성능지표 - Silhouette

Silhouette (실루엣계수) :

- 군집 내 거리와 군집 간의 거리를 기준으로 군집 분할의 성과를 측정하는 방식
- 클러스터 안의 테이터들이 다른 클러스터와 비교해 얼마나 비슷한가를 나타내는 군집평가
- 실루엣 지표가 1에 가까울수록 군집화가 잘 되었다고 판단한다.



https://starrykss.tistory.com/1962?category=911653

군집화를 활용하여
강원도의 공원 현황을 알아보고
관리구역을 분할하여 지도에
표시해보자.

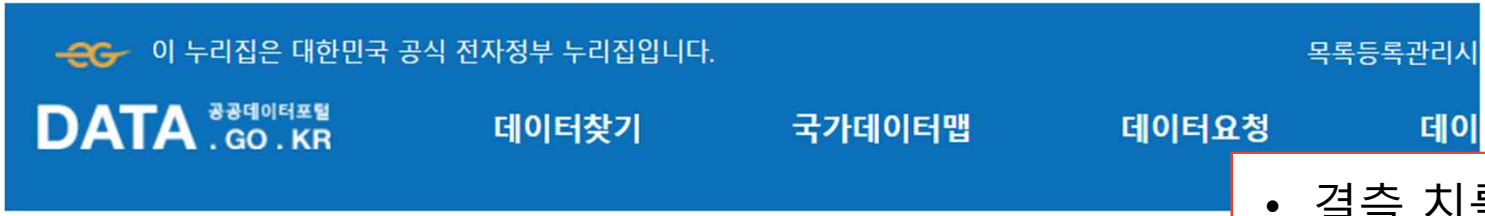
공공 데이터 제공 사이트

공공 데이터 : 나라에서 만들거나 취득하여 관리하는 데이터 (중앙정부, 지방자치단체, 공기업, 공공기관)

사이트명	URL
공공데이터 포털(행정안전부)	http://www.data.go.kr/
국가통계포털(통계청)	http://cosis.kr
고속도로 데이터포털(한국도로공사)	http://data.ex.co.kr
서울 열린 데이터 광장(서울시)	http://data.seoul.go.kr
경기데이터드림(경기도)	http://data.gg.go.kr
부동산 실거래가(국토교통부)	http://rt.molit.go.kr/
마이크로데이터 통합 서비스	https://mdis.kostat.go.kr/index.do
보건복지 데이터포털	https://kdx.kr/main
국민건강보험(NHISS)	https://nhiss.nhis.or.kr/bd/ay/bdaya001iv.do
구글 트랜드	https://trends.google.co.kr/trends/?geo=KR
네이버 데이터랩	https://datalab.naver.com/
한국데이터거래소(민간)	https://data.kihasa.re.kr/
캐글(해외)	https://www.kaggle.com/
Out World in Data(해외)	https://ourworldindata.org

데이터 검색 및 전처리

https://www.data.go.kr/data/15012890/standard.do



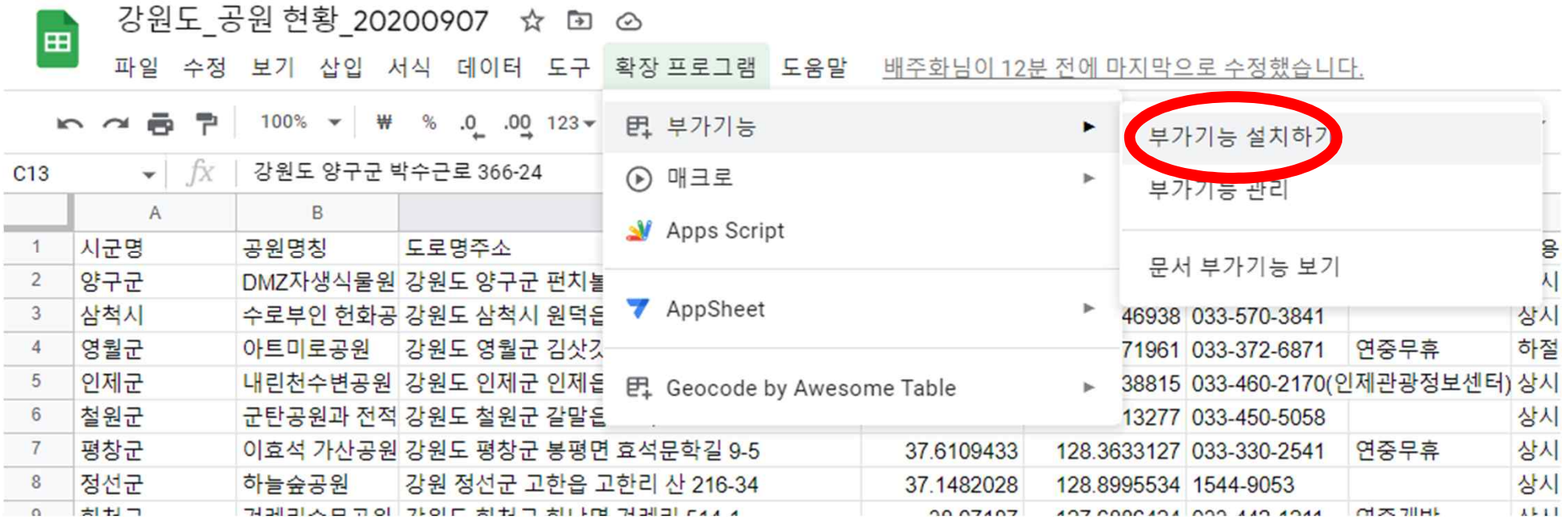
- 결측 치를 확인하여 누락된 정보를 채워 넣을 방법을 생각해보자.
- 필요 없는 속성을 제거하여 데이터의 양을 줄이자.

표준데이터 상세

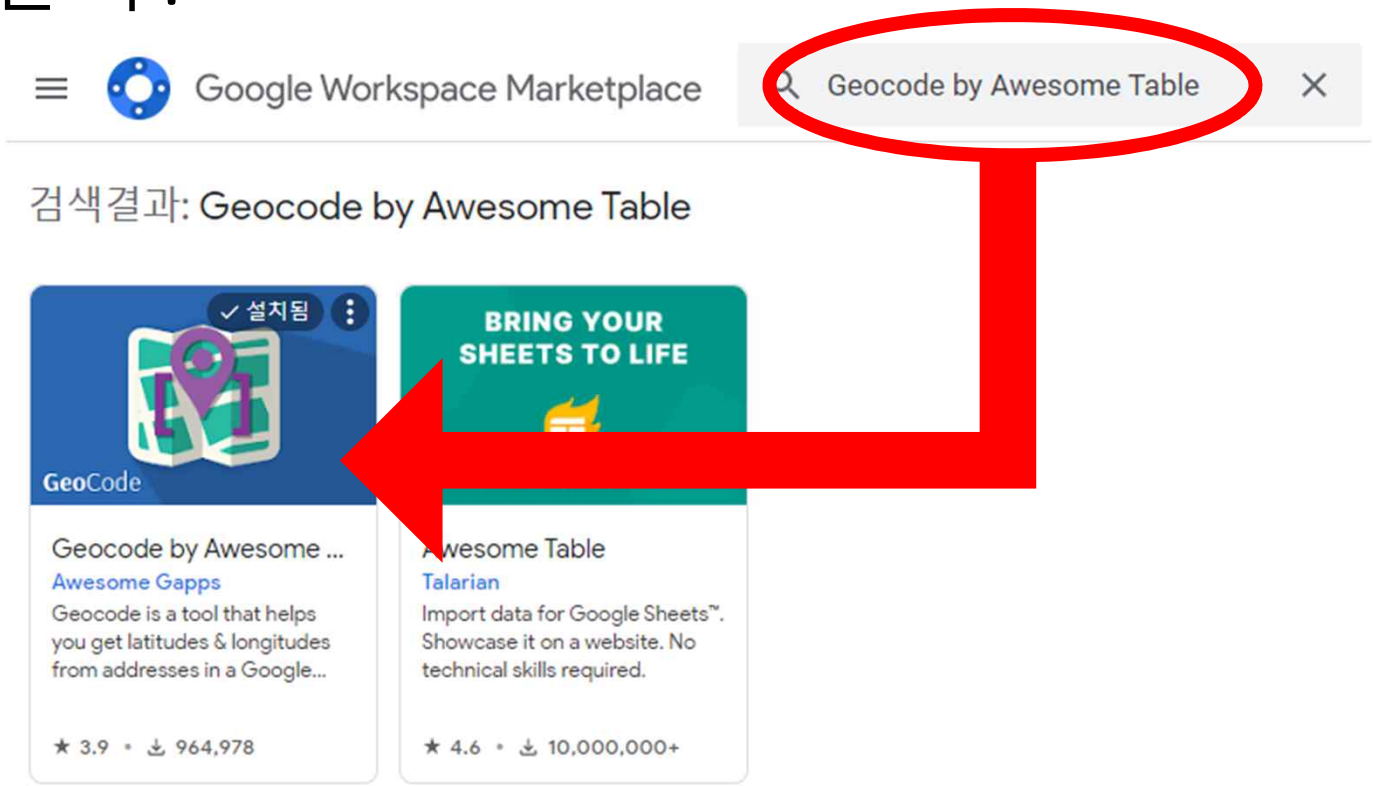
	A	B	C	D	E	F	
1	시군명	공원명칭	도로명주소	전화번호	휴무일	이용시간	주차시설
2	양구군	DMZ자생식물원	강원도 양구군 펀치볼로 916-70	033-480-3034	연중개방	상시이용 가능	주차가능
3	삼척시	수로부인 헌화공원	강원도 삼척시 원덕읍 임원항구로 33-17	033-570-3841		상시이용 가능	주차가능
4	영월군	아트미로공원	강원도 영월군 김삿갓면 영월동로 1121-32	033-372-6871	연중무휴	하절기 09:00~18:00 / 동절기 09:00~17:00	주차가능
5	인제군	내린천수변공원	강원도 인제군 인제읍 내린천로 5693	033-460-2170(인제관광정보센터)		상시이용 가능	
6	철원군	군탄공원과 전적비	강원도 철원군 갈말읍 호국로 4914-27	033-450-5058		상시이용 가능	주차가능
7	평창군	이효석 가산공원	강원도 평창군 봉평면 효석문학길 9-5	033-330-2541	연중무휴	상시이용 가능	
8	정선군	하늘숲공원	강원 정선군 고한읍 고한리 산 216-34	1544-9053		상시이용 가능	주차가능
9	화천군	겨레리수목공원	강원도 화천군 하남면 겨레리 514-1	033-442-1211	연중개방	상시이용 가능	주차가능
10	양구군	양구군민공원	강원도 양구군 양구읍 상리	033-480-2251	연중개방	상시이용 가능	주차가능
11	양구군	양구해시계	강원도 양구군 중앙길 53	033-480-2251	연중개방	상시이용 가능	주차가능
12	양구군	용머리공원	강원도 양구군 파로호로869번길 101	033-480-2251	연중개방	상시이용 가능	주차가능
13	양구군	청춘양구레포츠공원	강원도 양구군 박수근로 366-24	033-480-2251	12월~2월	상시이용 가능	주차가능

주소를 위도 경도 정보로 변환하는 방법

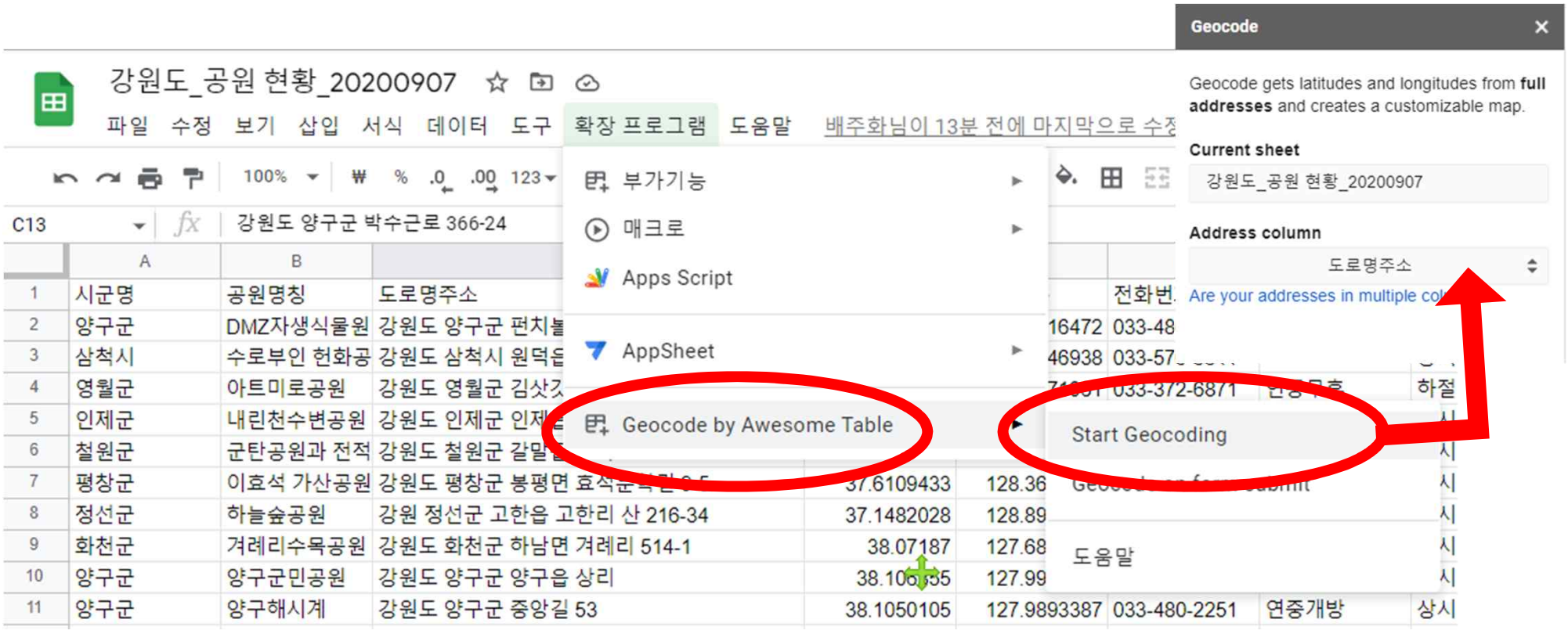
1. 구글스프레드 시트형식으로 파일을 연다.



2. 검색창에서 Geocode by Awesome Table을 선택하고 설치한다.

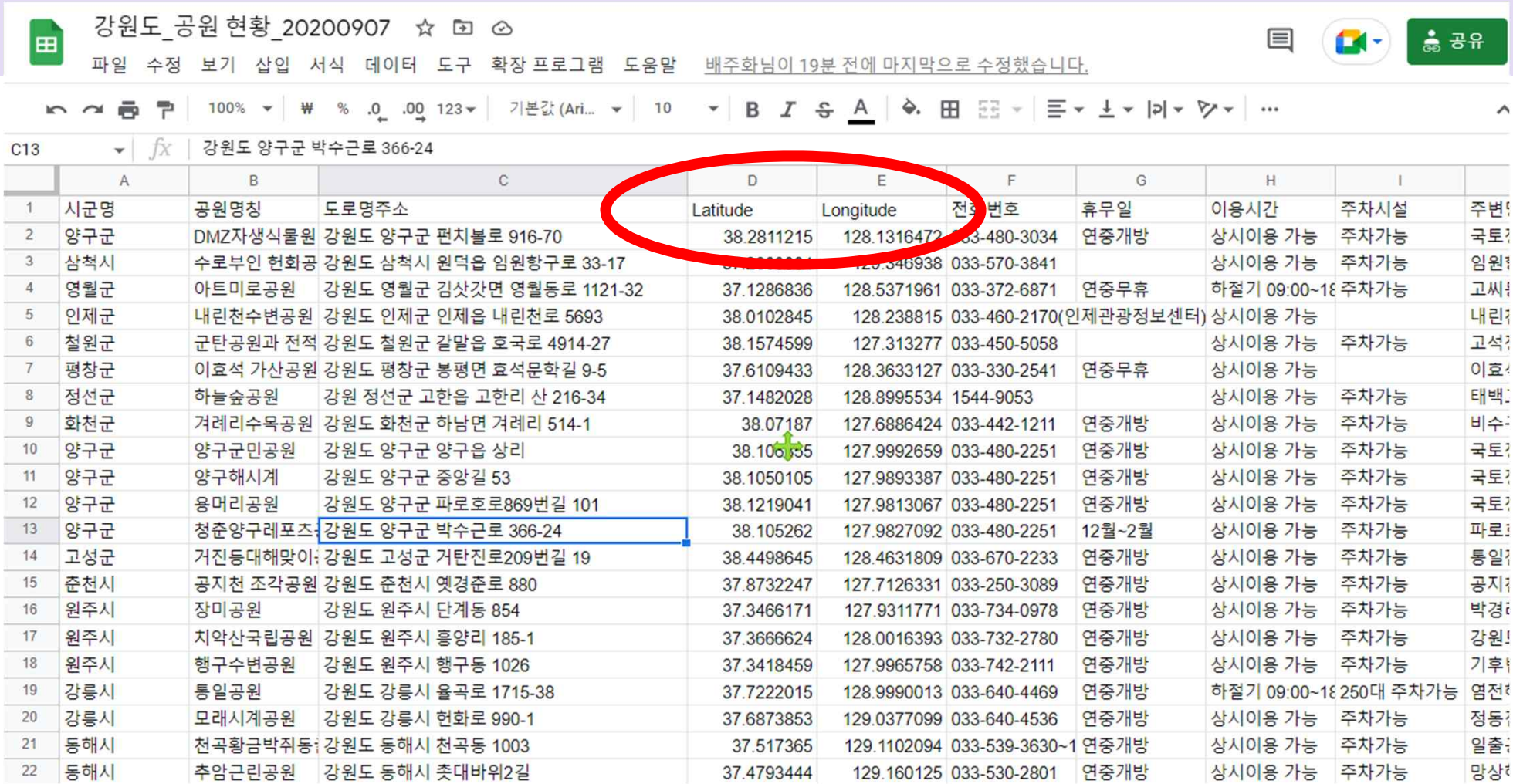


3. 확장 프로그램에서 Geocode를 선택하여 실행한다.



	A	B	C					
1	시군명	공원명칭	도로명주소				전화번	
2	양구군	DMZ자생식물원	강원도 양구군 펀치볼			16472	033-48	
3	삼척시	수로부인 헌화공	강원도 삼척시 원덕읍			46938	033-57	
4	영월군	아트미로공원	강원도 영월군 김삿갓				033-372-6871	하절
5	인제군	내린천수변공원	강원도 인제군 인제					
6	철원군	군탄공원과 전적	강원도 철원군 갈말					
7	평창군	이효석 가산공원	강원도 평창군 봉평면 효석	37.6109433	128.36			
8	정선군	하늘숲공원	강원 정선군 고한읍 고한리 산 216-34	37.1482028	128.89			
9	화천군	겨레리수목공원	강원도 화천군 하남면 겨레리 514-1	38.07187	127.68			
10	양구군	양구군민공원	강원도 양구군 양구읍 상리	38.106	127.99			
11	양구군	양구해시계	강원도 양구군 중앙길 53	38.1050105	127.9893387	033-480-2251	연중개방	상시

- 1일 1000 개 까지 무료 변환 가능



강원도_공원 현황_20200907

파일 수정 보기 삽입 서식 데이터 도구 확장 프로그램 도움말 배주화님이 19분 전에 마지막으로 수정했습니다.

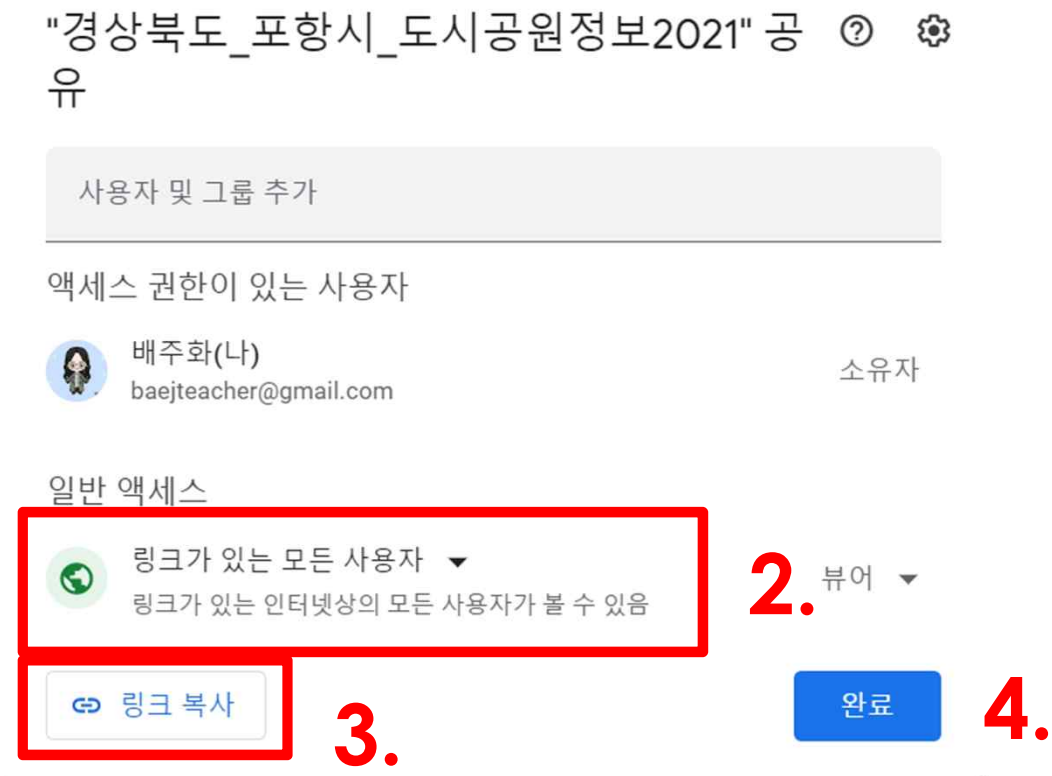
공유

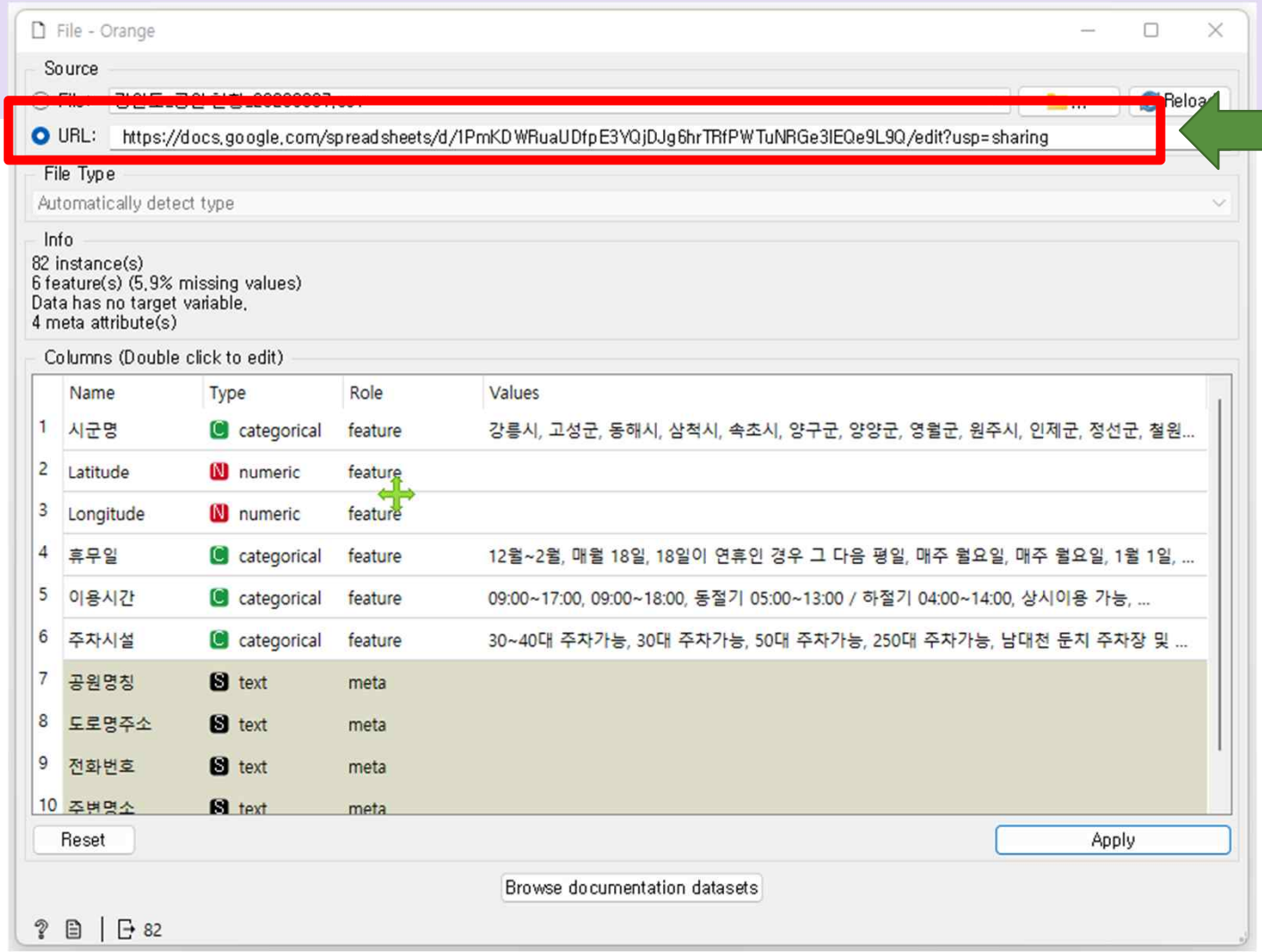
C13 강원도 양구군 박수근로 366-24

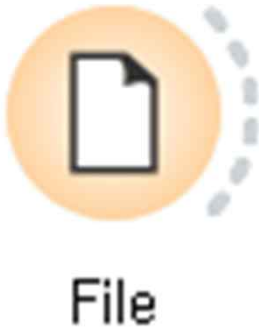
	A	B	C	D	E	F	G	H	I	
1	시군명	공원명칭	도로명주소	Latitude	Longitude	전화번호	휴무일	이용시간	주차시설	주변
2	양구군	DMZ자생식물원	강원도 양구군 펀치볼로 916-70	38.2811215	128.1316472	033-480-3034	연중개방	상시이용 가능	주차가능	국토
3	삼척시	수로부인 헌화공	강원도 삼척시 원덕읍 임원항구로 33-17		129.346938	033-570-3841		상시이용 가능	주차가능	임원
4	영월군	아트미로공원	강원도 영월군 김삿갓면 영월동로 1121-32	37.1286836	128.5371961	033-372-6871	연중무휴	하절기 09:00~18	주차가능	고씨
5	인제군	내린천수변공원	강원도 인제군 인제읍 내린천로 5693	38.0102845	128.238815	033-460-2170(인제관광정보센터)		상시이용 가능		내린
6	철원군	군탄공원과 전적	강원도 철원군 갈말읍 호국로 4914-27	38.1574599	127.313277	033-450-5058		상시이용 가능	주차가능	고석
7	평창군	이효석 가산공원	강원도 평창군 봉평면 효석문학길 9-5	37.6109433	128.3633127	033-330-2541	연중무휴	상시이용 가능		이효
8	정선군	하늘숲공원	강원 정선군 고한읍 고한리 산 216-34	37.1482028	128.8995534	1544-9053		상시이용 가능	주차가능	태백
9	화천군	겨레리수목공원	강원도 화천군 하남면 겨레리 514-1	38.07187	127.6886424	033-442-1211	연중개방	상시이용 가능	주차가능	비수
10	양구군	양구군민공원	강원도 양구군 양구읍 상리	38.106[illegible]55	127.9992659	033-480-2251	연중개방	상시이용 가능	주차가능	국토
11	양구군	양구해시계	강원도 양구군 중앙길 53	38.1050105	127.9893387	033-480-2251	연중개방	상시이용 가능	주차가능	국토
12	양구군	용머리공원	강원도 양구군 파로호로869번길 101	38.1219041	127.9813067	033-480-2251	연중개방	상시이용 가능	주차가능	국토
13	양구군	청춘양구레포츠	강원도 양구군 박수근로 366-24	38.105262	127.9827092	033-480-2251	12월~2월	상시이용 가능	주차가능	파로
14	고성군	거진등대해맞이	강원도 고성군 거탄진로209번길 19	38.4498645	128.4631809	033-670-2233	연중개방	상시이용 가능	주차가능	통일
15	춘천시	공지천 조각공원	강원도 춘천시 옛경춘로 880	37.8732247	127.7126331	033-250-3089	연중개방	상시이용 가능	주차가능	공지
16	원주시	장미공원	강원도 원주시 단계동 854	37.3466171	127.9311771	033-734-0978	연중개방	상시이용 가능	주차가능	박경
17	원주시	치악산국립공원	강원도 원주시 흥양리 185-1	37.3666624	128.0016393	033-732-2780	연중개방	상시이용 가능	주차가능	강원
18	원주시	행구수변공원	강원도 원주시 행구동 1026	37.3418459	127.9965758	033-742-2111	연중개방	상시이용 가능	주차가능	기후
19	강릉시	통일공원	강원도 강릉시 율곡로 1715-38	37.7222015	128.9990013	033-640-4469	연중개방	하절기 09:00~18	250대 주차가능	염전
20	강릉시	모래시계공원	강원도 강릉시 헌화로 990-1	37.6873853	129.0377099	033-640-4536	연중개방	상시이용 가능	주차가능	정동
21	동해시	천곡황금박쥐동	강원도 동해시 천곡동 1003	37.517365	129.1102094	033-539-3630~1	연중개방	상시이용 가능	주차가능	일출
22	동해시	추암근린공원	강원도 동해시 촛대바위2길	37.4793444	129.160125	033-530-2801	연중개방	상시이용 가능	주차가능	망상

구글드라이브 경로로 데이터 불러오기

- 구글 스프레드 시트의 공유 기능을 사용

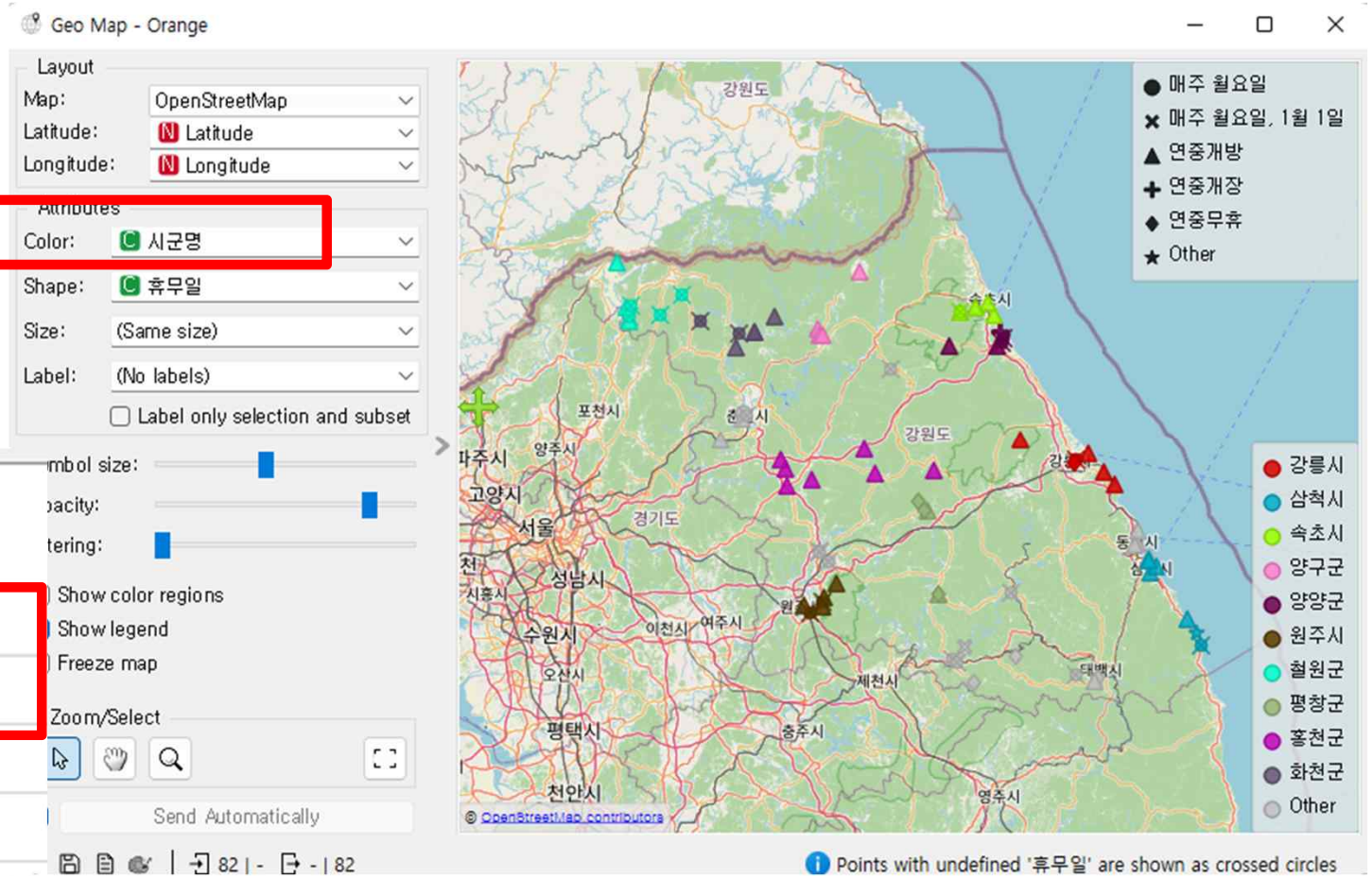



File - Orange
Source
URL:
https://docs.google.com/spreadsheets/d/1PmKDWRuaUDfpE3YQjDJg6hrTRfPWTuNRGe3IEQe9L9Q/edit?usp=sharing
File Type
Automatically detect type
Info
82 instance(s)
6 feature(s) (5.9% missing values)
Data has no target variable.
4 meta attribute(s)
Columns (Double click to edit)
Name
Type
Role
Values
1 시군명 categorical feature 강릉시, 고성군, 동해시, 삼척시, 속초시, 양구군, 양양군, 영월군, 원주시, 인제군, 정선군, 철원...
2 Latitude numeric feature
3 Longitude numeric feature
4 휴무일 categorical feature 12월~2월, 매월 18일, 18일이 연휴인 경우 그 다음 평일, 매주 월요일, 매주 월요일, 1월 1일, ...
5 이용시간 categorical feature 09:00~17:00, 09:00~18:00, 동절기 05:00~13:00 / 하절기 04:00~14:00, 상시이용 가능, ...
6 주차시설 categorical feature 30~40대 주차가능, 30대 주차가능, 50대 주차가능, 250대 주차가능, 남대천 둔치 주차장 및 ...
7 공원명칭 text meta
8 도로명주소 text meta
9 전화번호 text meta
10 text meta
Reset
Apply
Browse documentation datasets
82

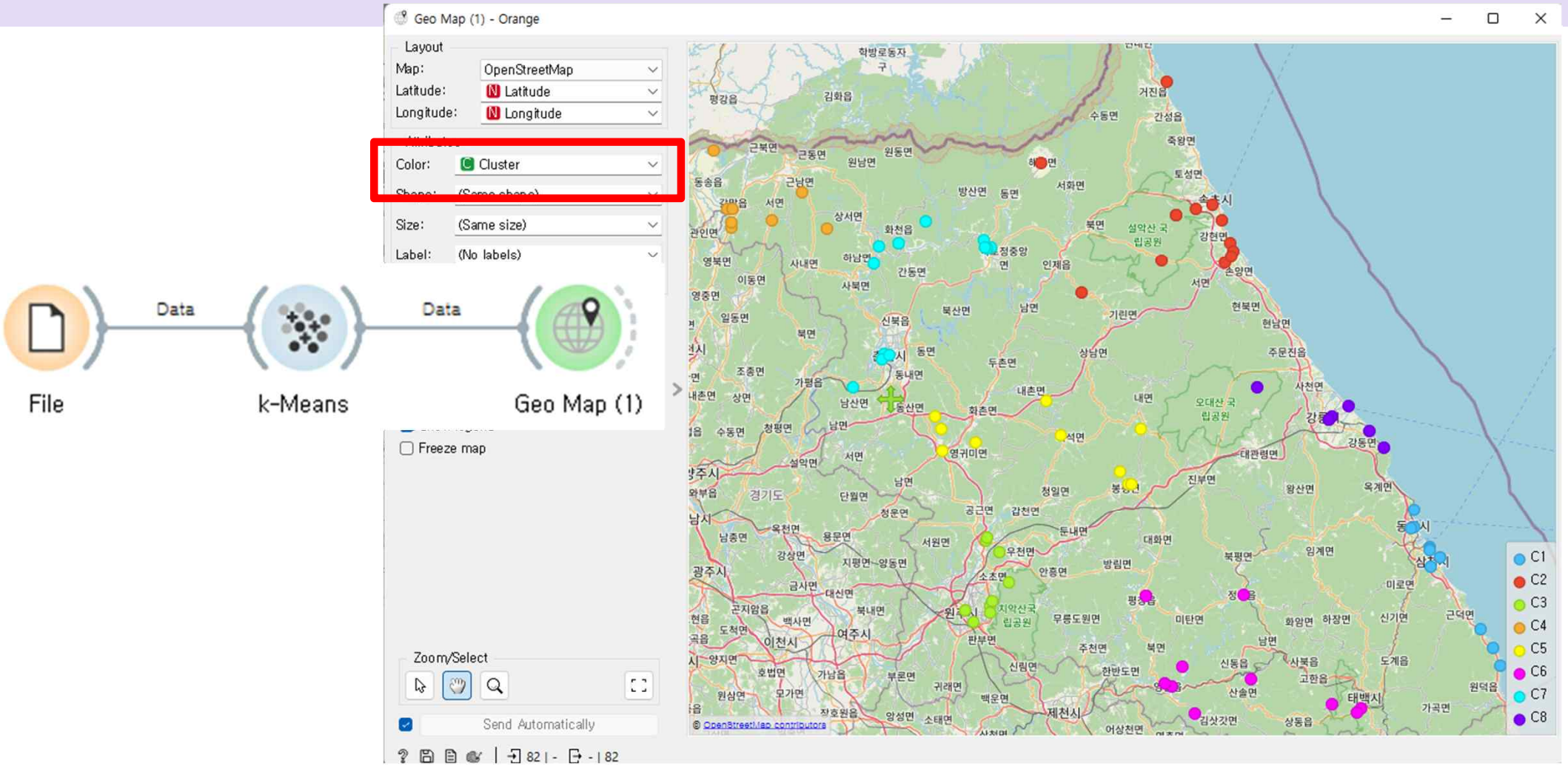

File

지도에 나타내기 - 강원지역 공원

	Name	Type	Role
1	시군명	categorical	feature
2	Latitude	numeric	feature
3	Longitude	numeric	feature
4	휴무일	categorical	feature
5	이용시간	categorical	feature



K-means 군집화


Geo Map (1) - Orange
Layout
Map: OpenStreetMap
Latitude: Latitude
Longitude: Longitude
Color: Cluster
Size: (Same size)
Label: (No labels)
Freeze map
Zoom/Select
Send Automatically
Data
Data
File
k-Means
Geo Map (1)
C1
C2
C3
C4
C5
C6
C7
C8

다음 시간에는 정형데이터를 활용한
비지도학습 - 군집화 두번째 시간으로
계층적 군집화에 대해 알아보고 시각
화 해보도록 하겠습니다.

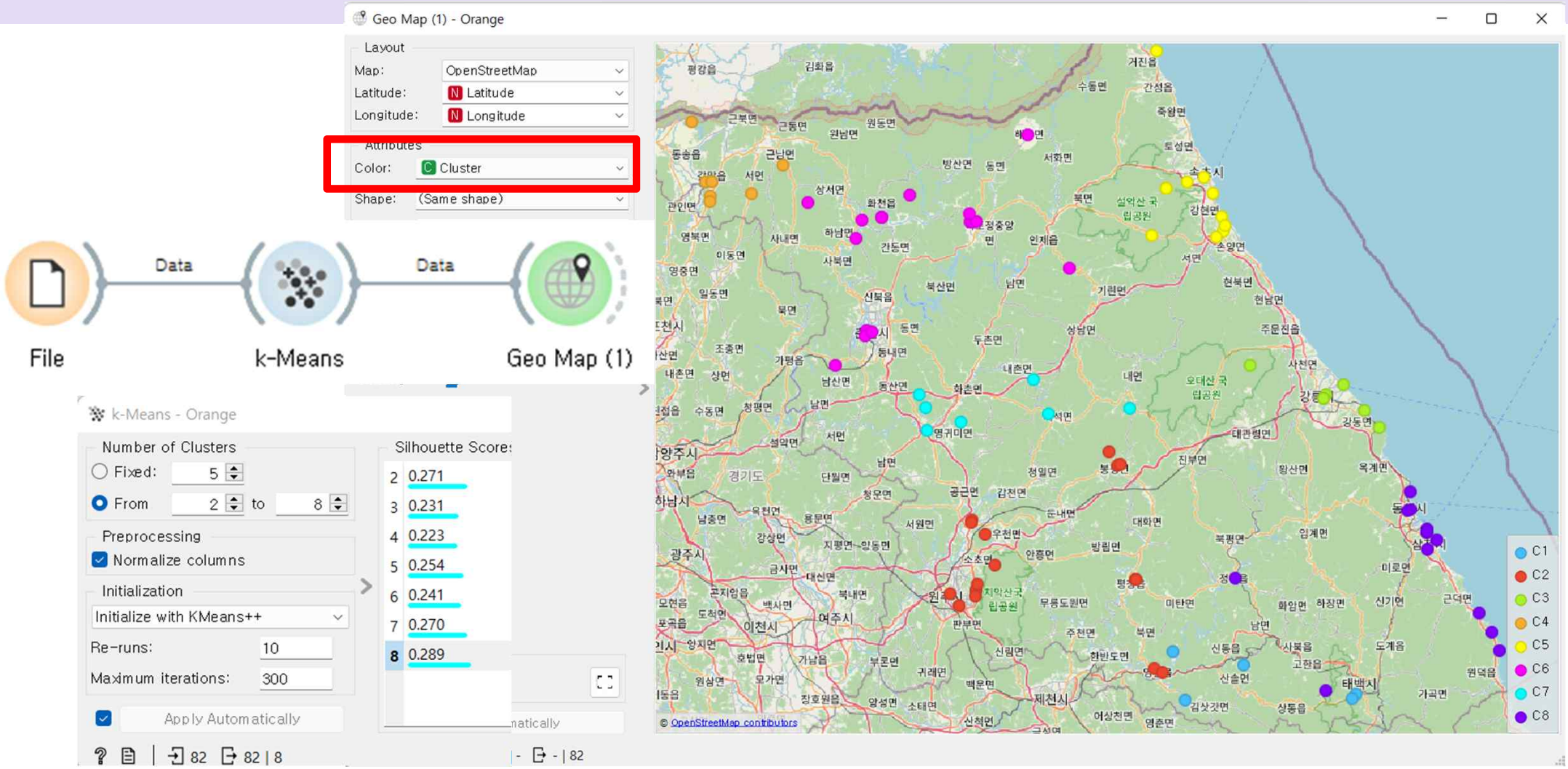
orange 활용 데이터 분석 및 머신 러닝

7차시

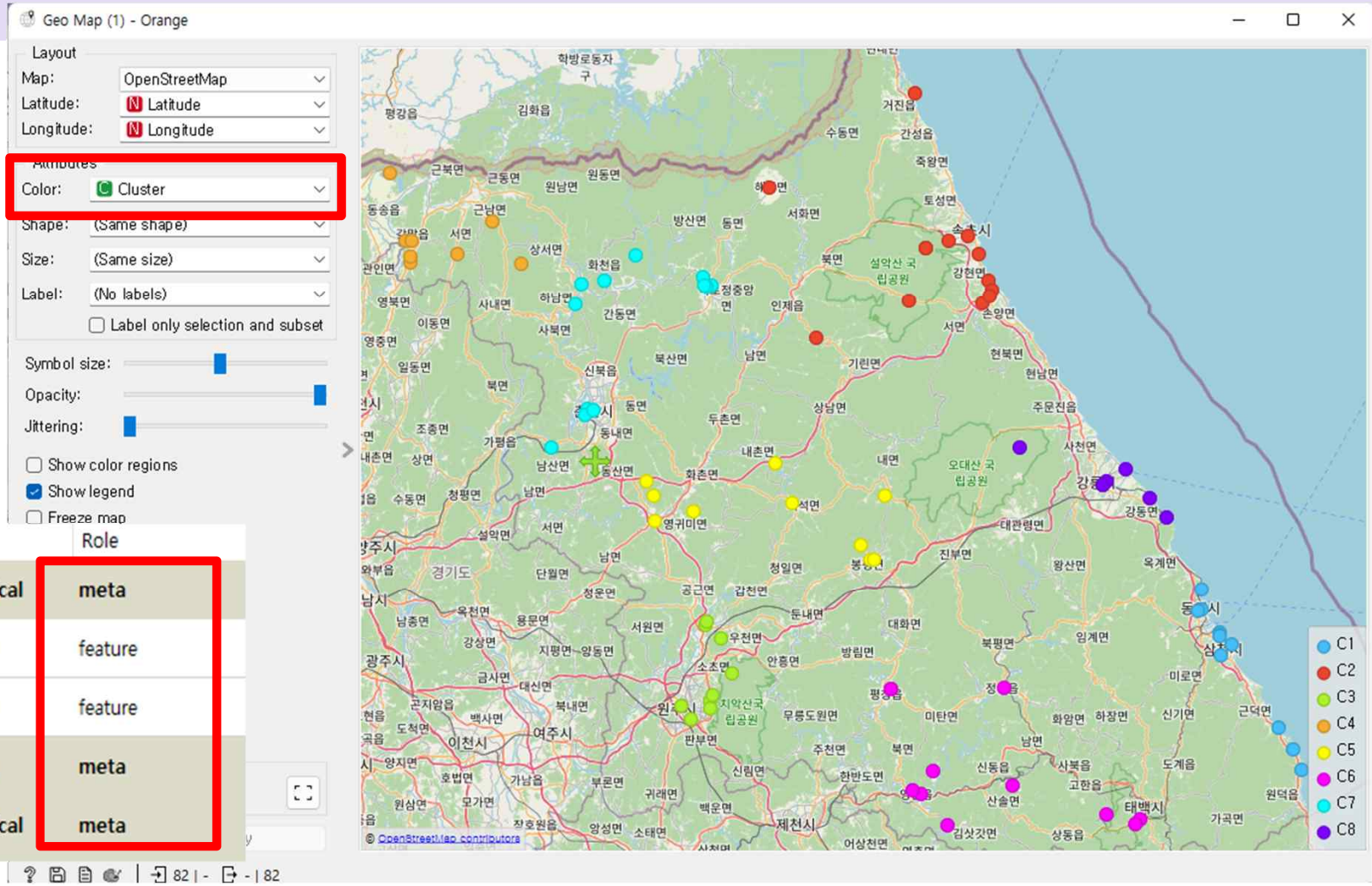
정형데이터를 활용한 군집화 2

Hierarchical clustering 모델을 활용한 군집화

관리구역 나누기 - K-means 군집화

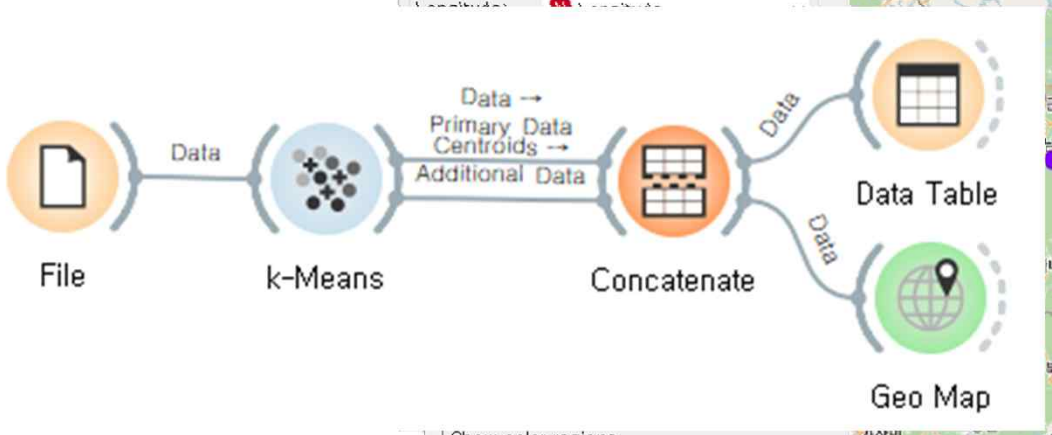


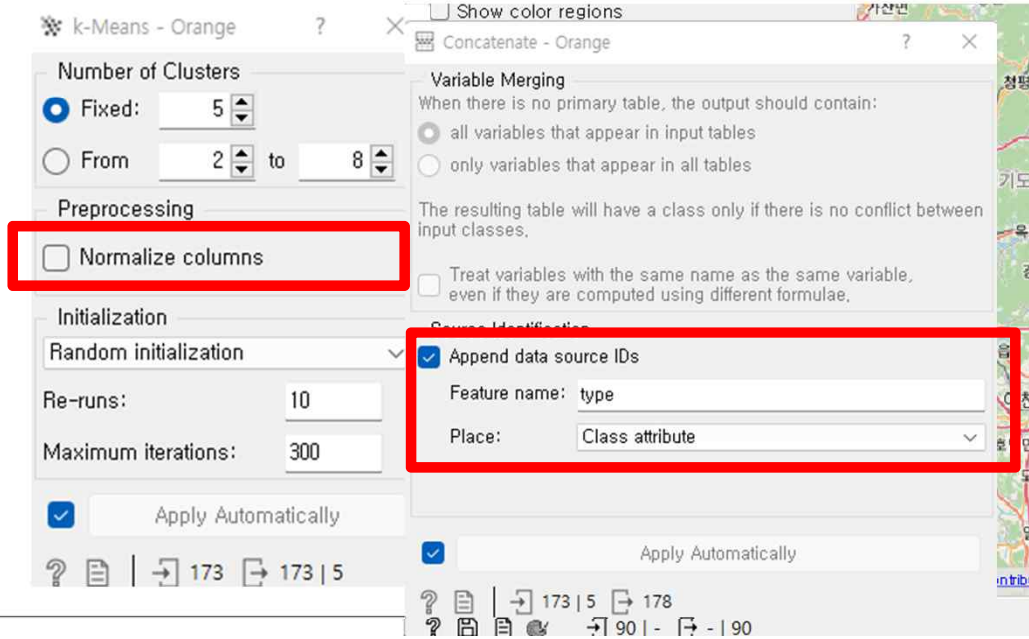
관리구역 나누기 - 위도 경도 정보를 활용한 K-means 군집화

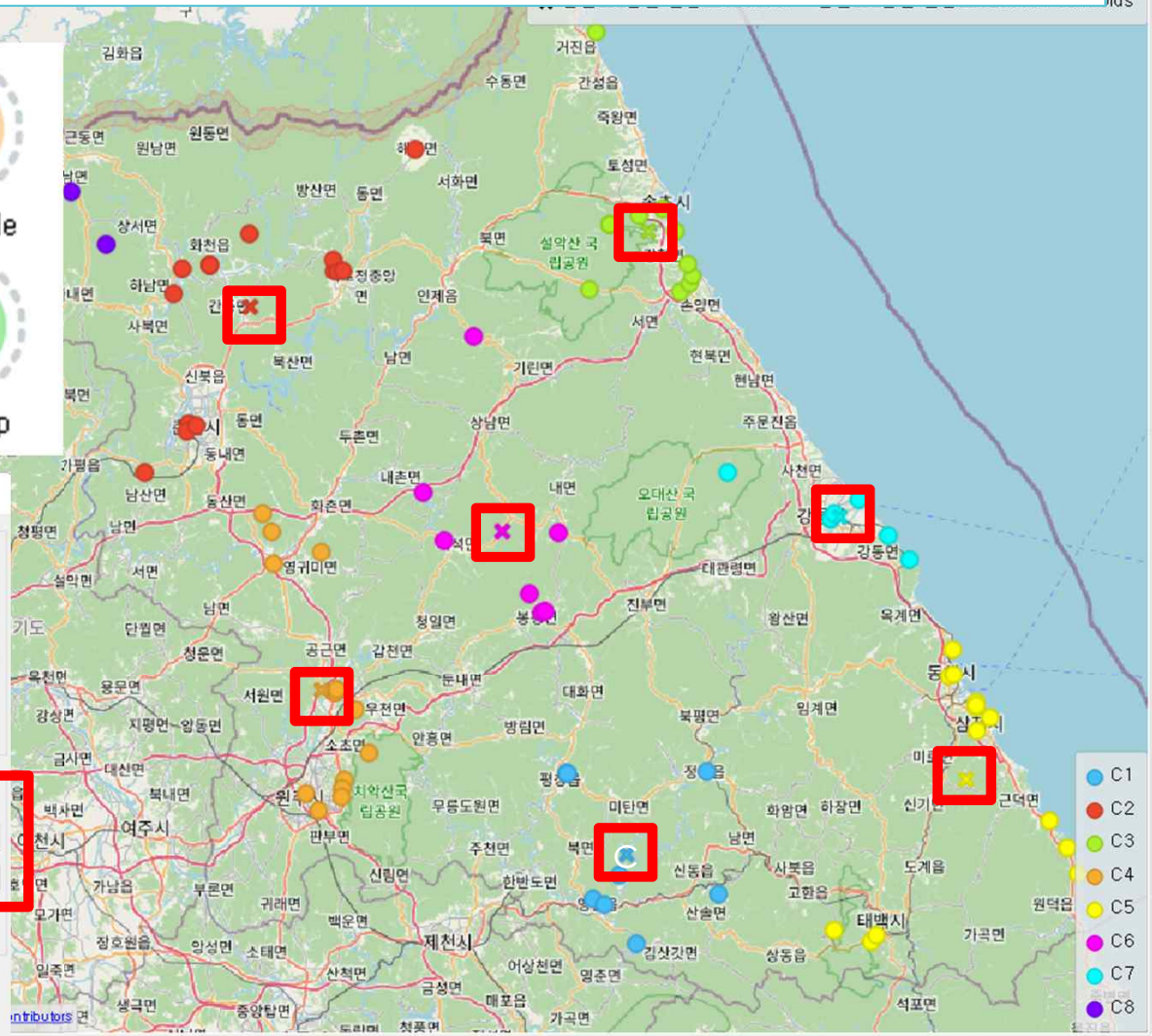


Name	Type	Role
공원구분	**categorical**	**meta**
Latitude	numeric	feature
Longitude	numeric	feature
공원면적	**numeric**	**meta**
공원보유시설(운동시설)	**categorical**	**meta**

k-Means 결과를 지도에 나타내기 - 중심점과 함께 표시



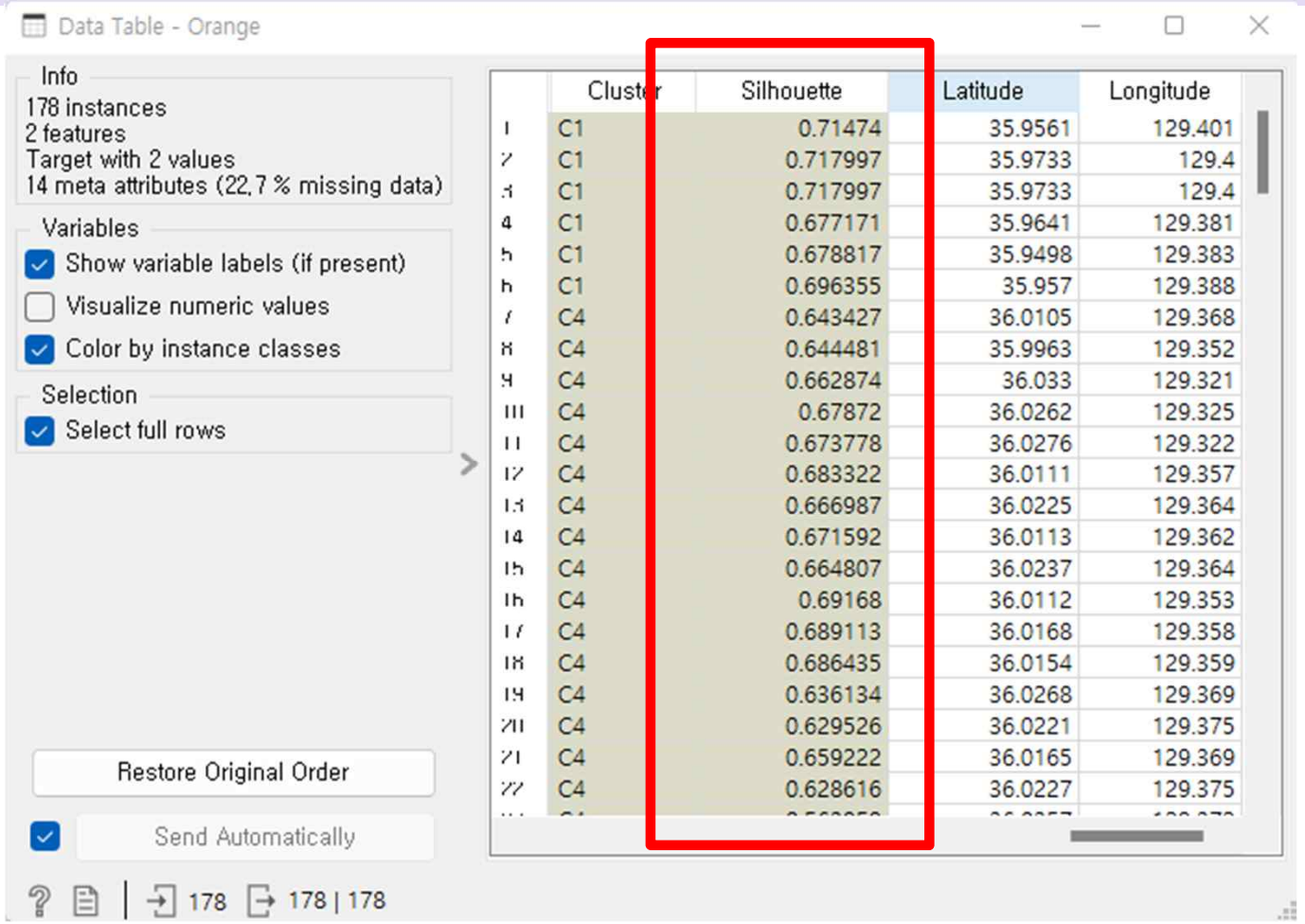




k-Means 군집화의 성능지표 - Silhouette

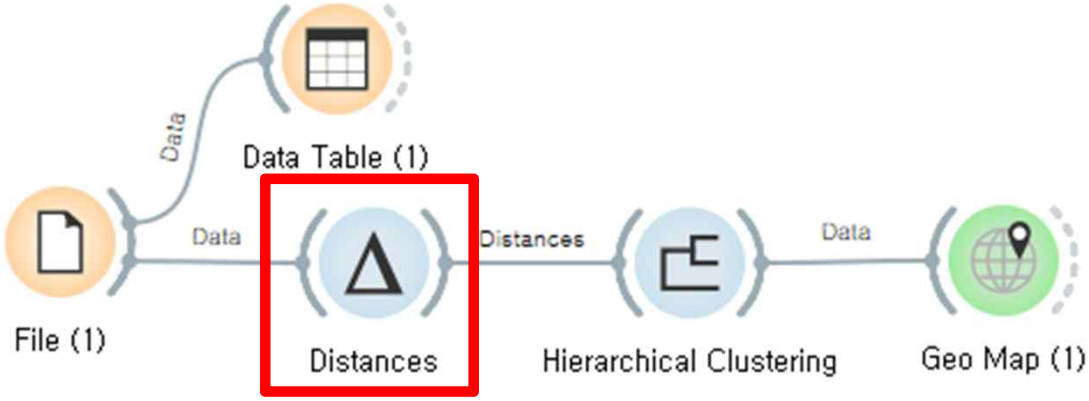
Silhouette (실루엣계수) :

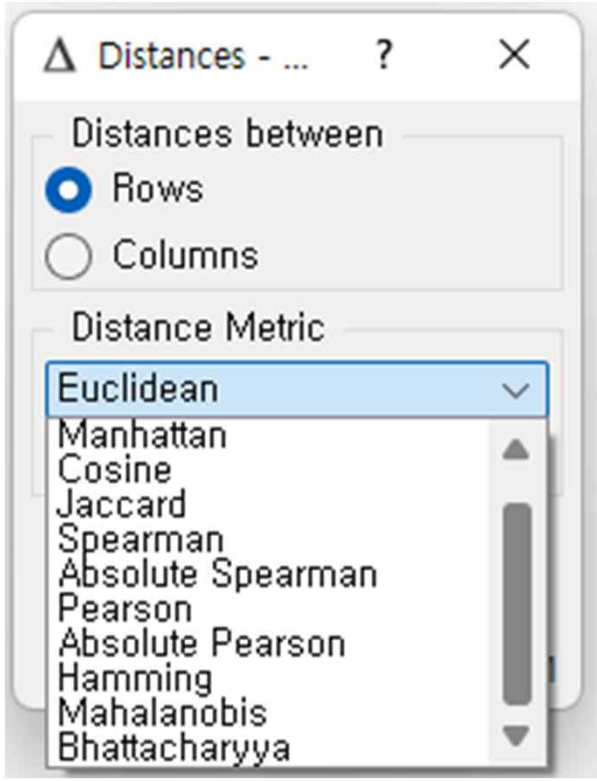
- 군집 내 거리와 군집 간의 거리를 기준으로 군집 분할의 성과를 측정하는 방식
- 클러스터 안의 테이터들이 다른 클러스터와 비교해 얼마나 비슷한가를 나타내는 군집평가
- 실루엣 지표가 1에 가까울수록 군집화가 잘 되었다고 판단한다.



https://starrykss.tistory.com/1962?category=911653

Hierarchical Clustering 모델을 활용하여 군집화 하기



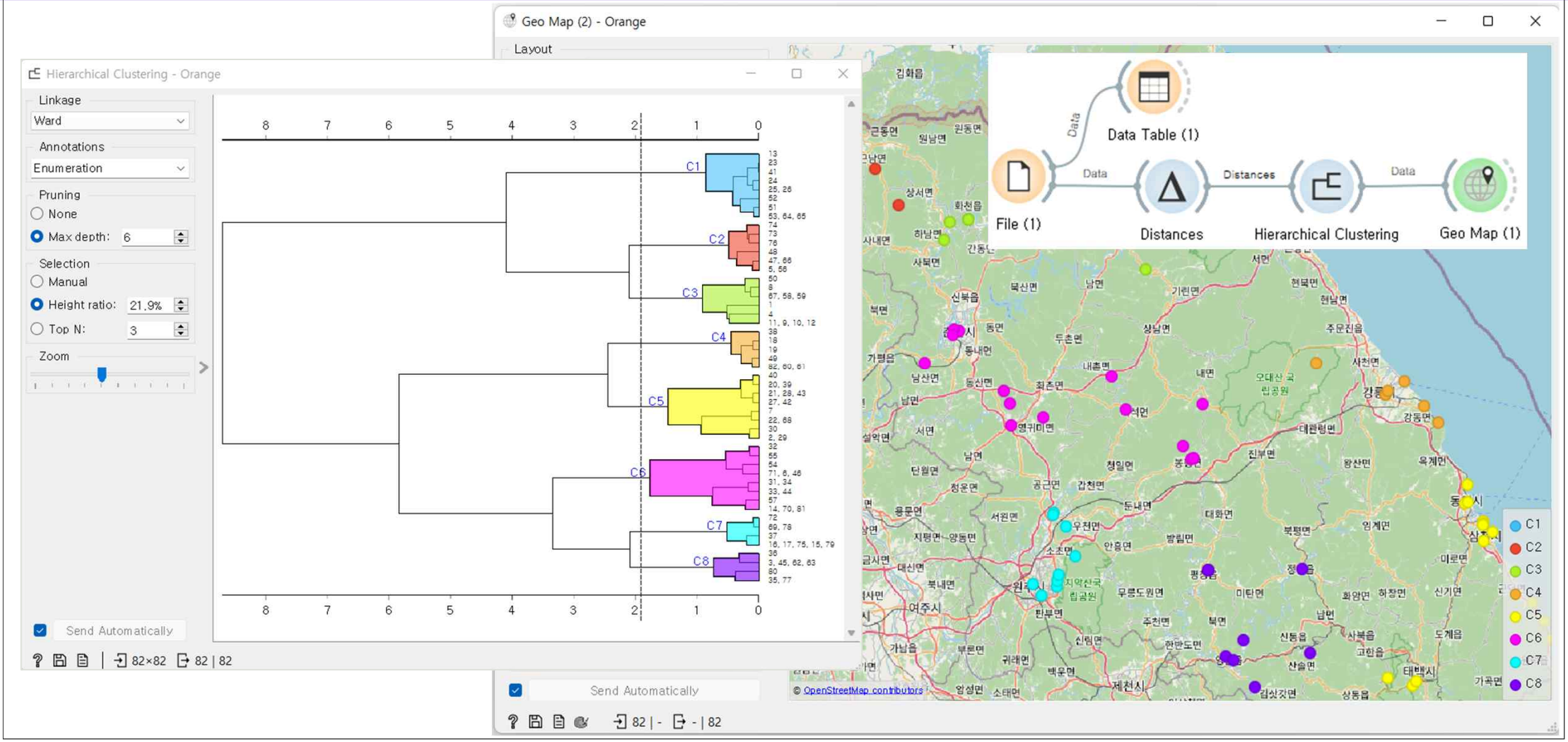


군집 간 거리 측정 방법 (Distance Metrics)

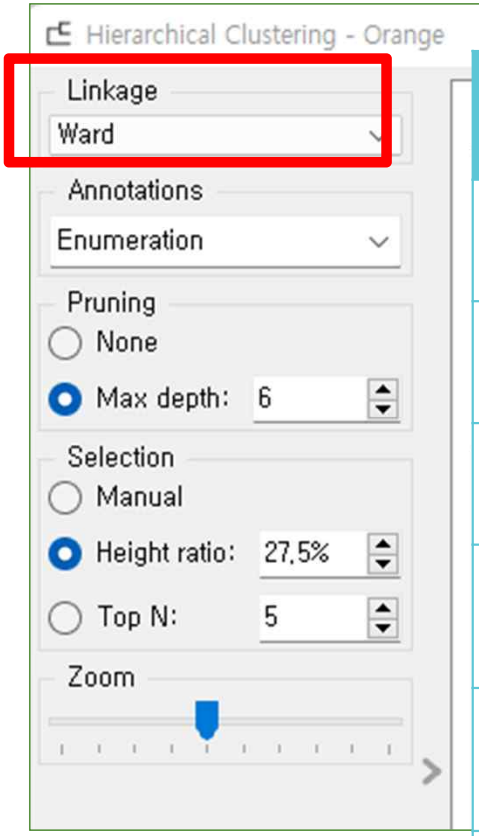
- 유클리드 거리 : 직선거리. 통계적 개념 내포 안됨. 산포정도 반영안됨.
- 맨하튼 거리 : 절댓값들의 합
- 캔버라 거리 : Σ(x-y)/(x+y)
- 표준화 거리 : 통계적 거리. 측정단위 표준화. 표준편차 -> 유클리디안.
- 마할라노비스 거리 : 통계적 거리. 변수의 표준화 + 변수간의 상관성(공분산행렬) 고려
- 자카드계수 : 0과 1사이 값의 유사도 측정. 불린 속성의 개체일 경우, 동일하면 1, 완전 불일치이면 0.
- 코사인거리, 코사인계수

https://starrykss.tistory.com/1962?category=911653

Hierarchical Clustering 모델을 활용하여 군집화 하기

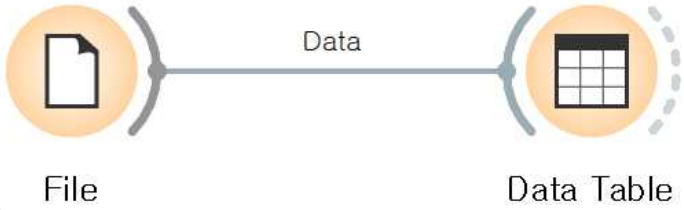


Hierarchical Clustering 모델을 활용하여 군집화 하기



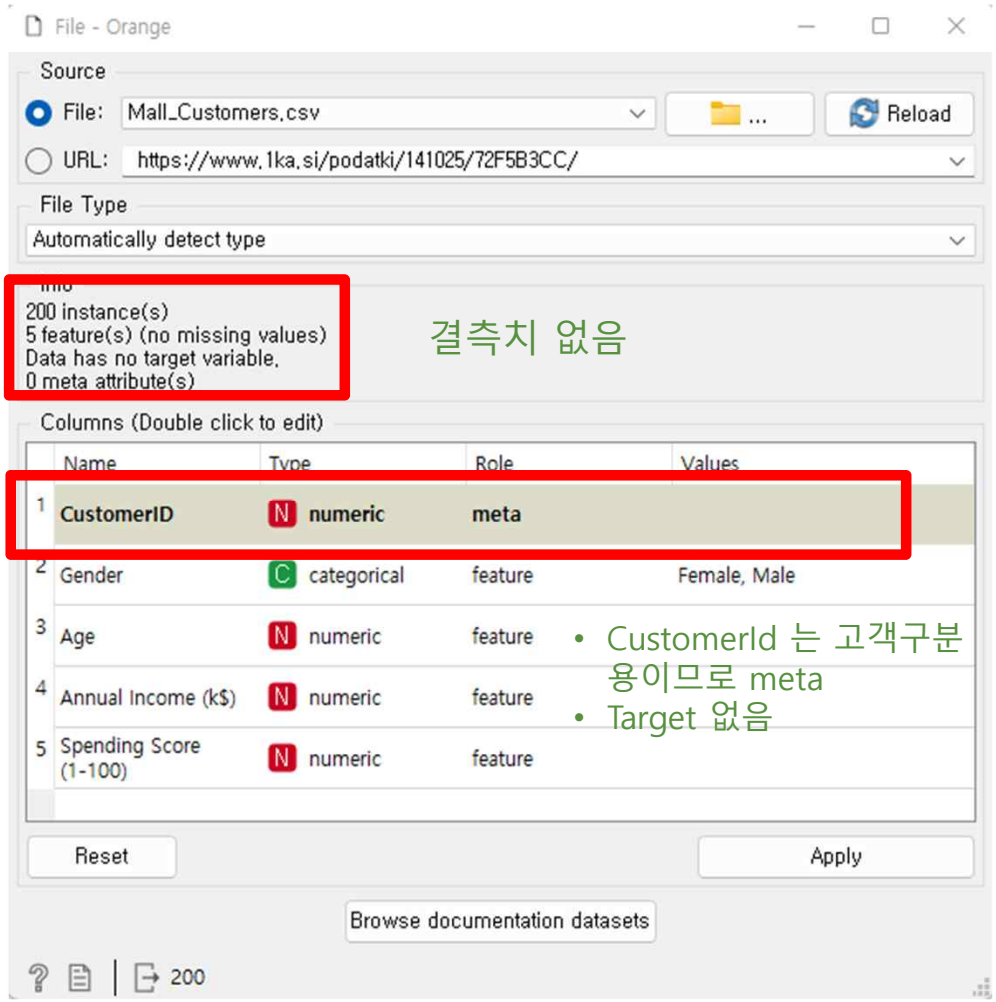
방법	설명
'average' : 평균연결법	비가중 평균 거리(UPGMA)로도 불리우며, 모든 항목에 대한 거리의 평균을 구하면서 군집화 하는 방법. 계산량이 많아질 수 있다.
'centroid' : 중심연결법	중심 거리(UPGMC)로, 두 군집 중심 간의 거리를 측정함, 유클리드 거리에만 적합함
'complete' : 최장 연결법	최장 거리, 완전 연결법이라고도 하며, 두 군집 간 데이터 사이의 거리의 최대값을 측정하는 방식
'single' : 최단연결법	단일 연결법이라고도 하며, 두 군집사이의 거리를 군집 간의 데이터 간의 가장 가까운 거리를 측정하는 방식이다. 고립된 군집을 찾는 데 중점을 둔 방식.
'ward' : 와드 연결법	계층적 군집 내의 오차 제곱합에 기초하여 군집을 수행하는 군집방법으로 크기가 비슷한 군집끼리 병합하는 경향이 있으며 유클리드 거리에만 적합함
'weighted' : 가중평균연결법	가중 평균 거리(WPGMA)

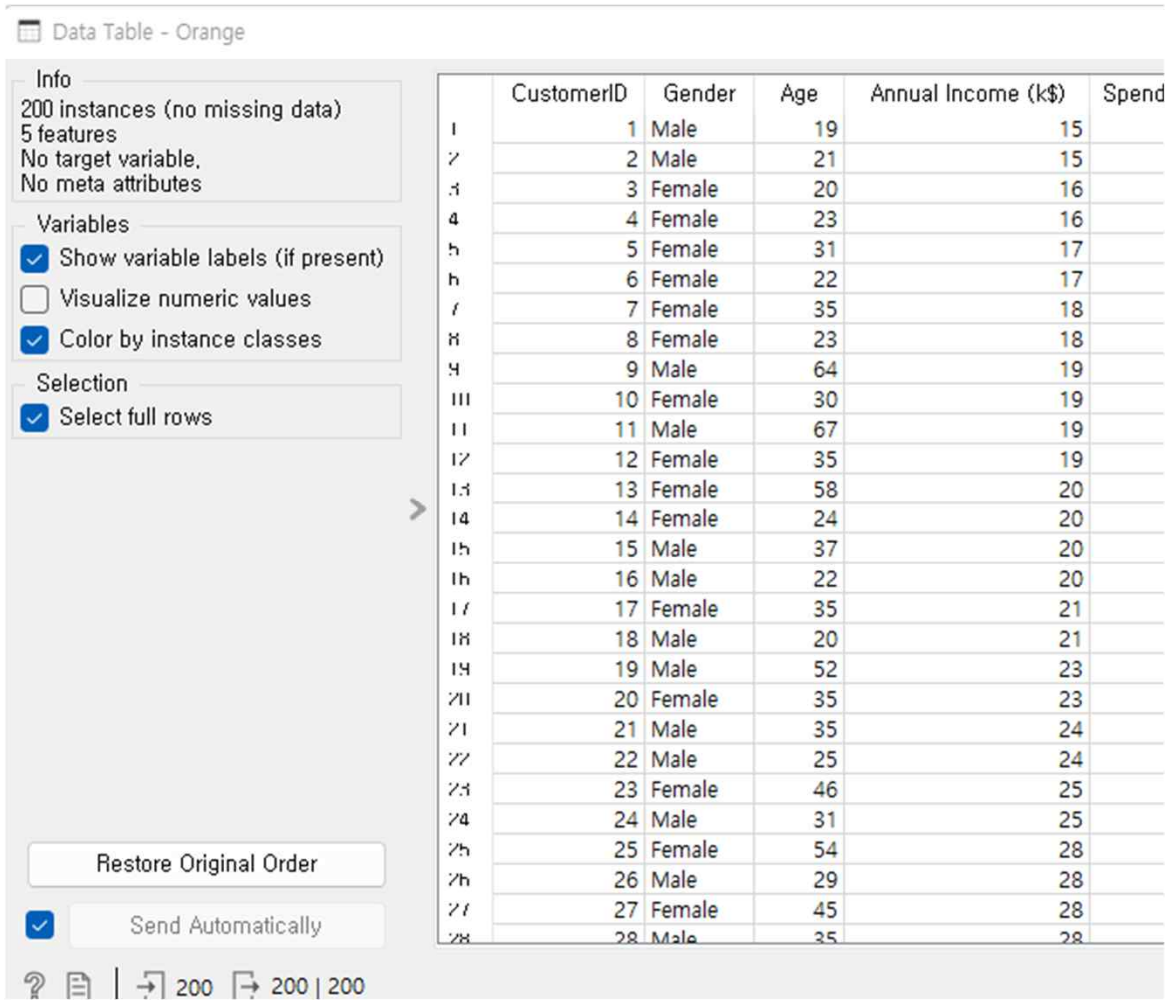
쇼핑몰 고객의 성별, 나이, 연 수입과 소비 점수 데이터를 활용한 고객 클러스터링



	A	B	C	D	E	F
1	Customer	Gender	Age	Annual Income (k$)	Spending Score (1-100)	
2	1	Male	19	15	39	
3	2	Male	21	15	81	
4	3	Female	20	16	6	
5	4	Female	23	16	77	
6	5	Female	31	17	40	
7	6	Female	22	17	76	
8	7	Female	35	18	6	
9	8	Female	23	18	94	
10	9	Male	64	19	3	
11	10	Female	30	19	72	
12	11	Male	67	19	14	
13	12	Female	35	19	99	
14	13	Female	58	20	15	
15	14	Female	24	20	77	
16	15	Male	37	20	13	
17	16	Male	22	20	79	
18	17	Female	35	21	35	
19	18	Male	20	21	66	
20	19	Male	52	23	29	

데이터 준비하기 및 탐색

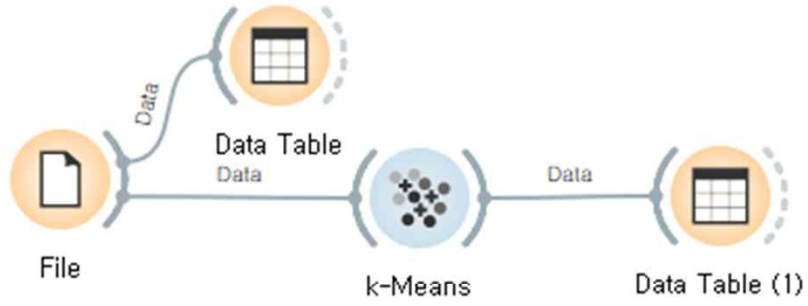

File - Orange
Source
File: Mall_Customers.csv
Reload
URL: https://www.1ka.si/podatki/141025/72F5B3CC/
File Type
Automatically detect type
Info
200 instance(s)
5 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)
결측치 없음
Columns (Double click to edit)
Name Type Role Values
1 CustomerID numeric meta
2 Gender categorical feature Female, Male
3 Age numeric feature
4 Annual Income (k$) numeric feature
5 Spending Score (1-100) numeric feature
• CustomerId 는 고객구분 용이므로 meta
• Target 없음
Reset
Apply
Browse documentation datasets
200

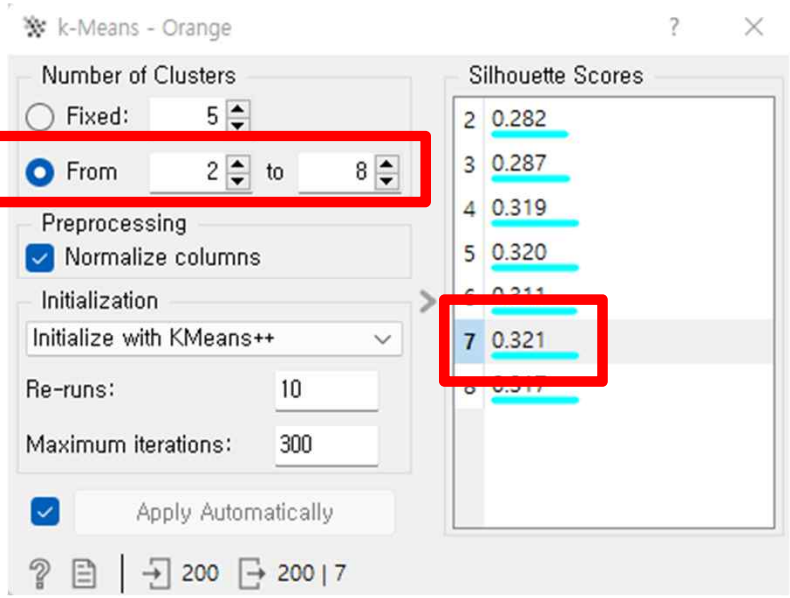

Data Table - Orange
Info
200 instances (no missing data)
5 features
No target variable.
No meta attributes
Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes
Selection
Select full rows
Restore Original Order
Send Automatically
200
200 | 200
CustomerID Gender Age Annual Income (k$) Spend
1 Male 19 15
2 Male 21 15
3 Female 20 16
4 Female 23 16
5 Female 31 17
6 Female 22 17
7 Female 35 18
8 Female 23 18
9 Male 64 19
10 Female 30 19
11 Male 67 19
12 Female 35 19
13 Female 58 20
14 Female 24 20
15 Male 37 20
16 Male 22 20
17 Female 35 21
18 Male 20 21
19 Male 52 23
20 Female 35 23
21 Male 35 24
22 Male 25 24
23 Female 46 25
24 Male 31 25
25 Female 54 28
26 Male 29 28
27 Female 45 28

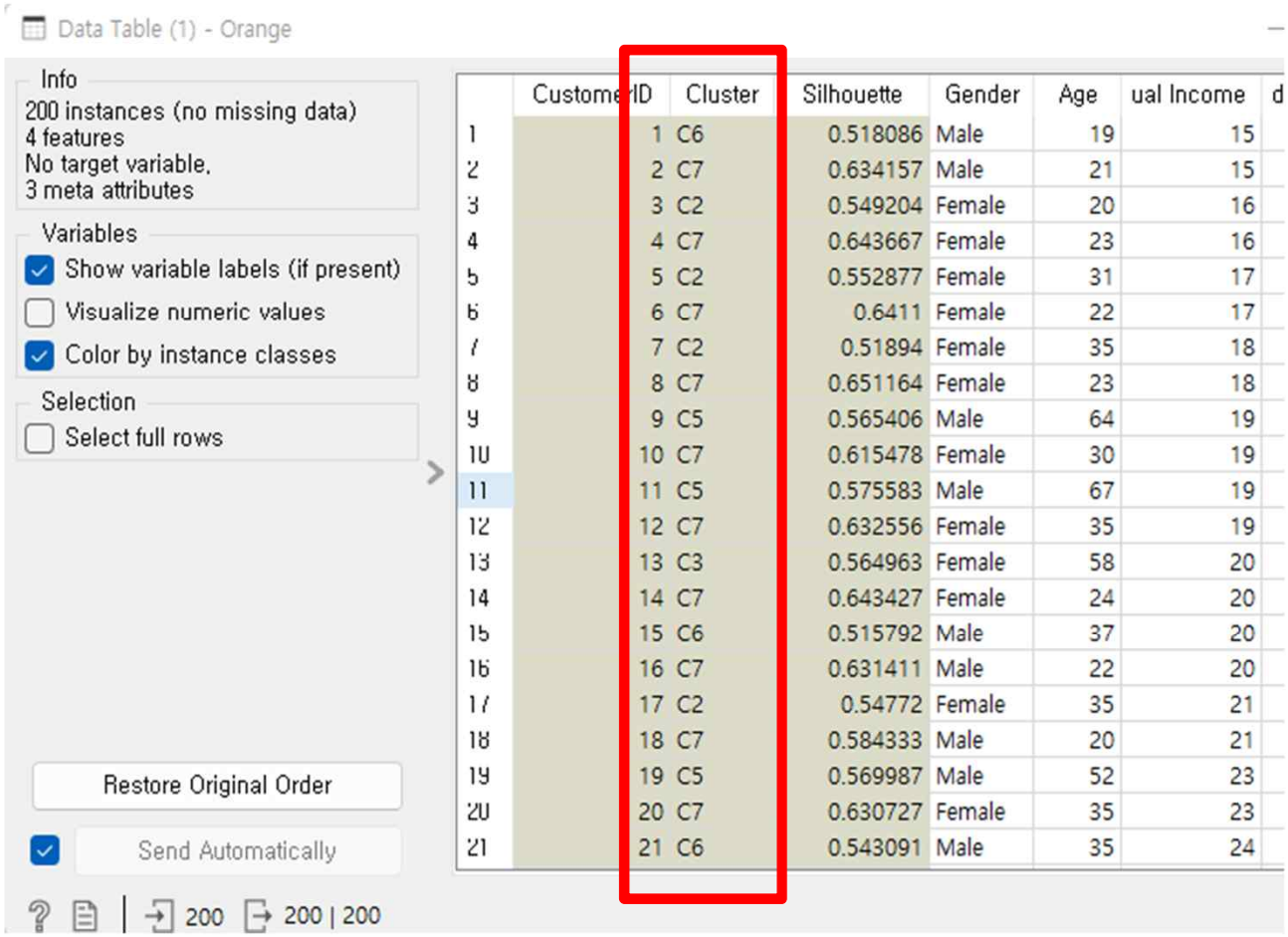
데이터 시각화


Data
Data Table
File
Data
Data
Distributions
Scatter Plot
Distributions - Orange
Variable
Filter...
CustomerID
Gender
Age
Annual Income (k$)
Spending Score (1-100)
Sort categories by frequency
Distribution
Fitted distribution
None
Bin width
10
Smoothing
10
Hide bars
Columns
Split by
Gender
Stack columns
Show probabilities
Show cumulative distribution
Apply Automatically
Female
Male
Probability of 'Gender' at given 'Annual Income (k$)'
Scatter Plot
Axes
Axis x:
Age
Axis y:
Spending Score (1-100)
Find Informative Projections
Attributes
Color:
Shape:
Size:
Label:
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Show all data on mouse hover
Show regression line
Treat variables as independent
Zoom/Select
Send Automatically
Spending Score (1-100)
Age
Probability of 'Gender' at given 'Spending Score (1-100)'

K-means 알고리즘을 활용한 군집화

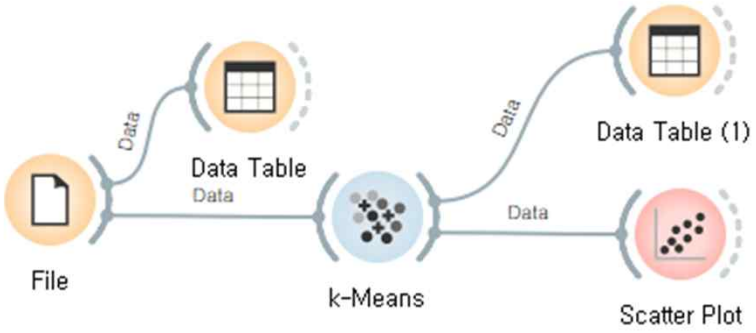


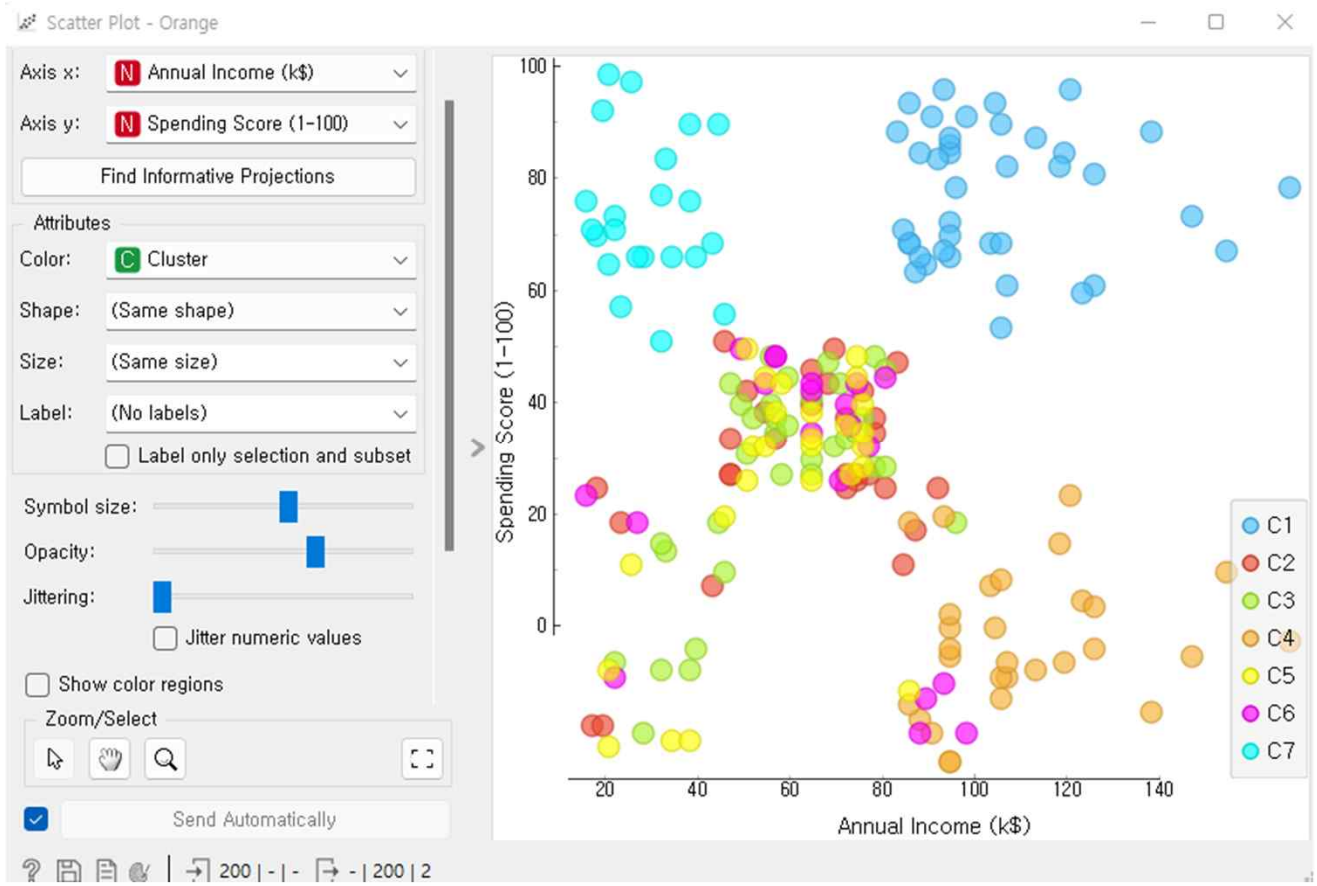




	CustomerID	Cluster	Silhouette	Gender	Age	ual Income	d
1	1	C6	0.518086	Male	19	15	
2	2	C7	0.634157	Male	21	15	
3	3	C2	0.549204	Female	20	16	
4	4	C7	0.643667	Female	23	16	
5	5	C2	0.552877	Female	31	17	
6	6	C7	0.6411	Female	22	17	
7	7	C2	0.51894	Female	35	18	
8	8	C7	0.651164	Female	23	18	
9	9	C5	0.565406	Male	64	19	
10	10	C7	0.615478	Female	30	19	
11	11	C5	0.575583	Male	67	19	
12	12	C7	0.632556	Female	35	19	
13	13	C3	0.564963	Female	58	20	
14	14	C7	0.643427	Female	24	20	
15	15	C6	0.515792	Male	37	20	
16	16	C7	0.631411	Male	22	20	
17	17	C2	0.54772	Female	35	21	
18	18	C7	0.584333	Male	20	21	
19	19	C5	0.569987	Male	52	23	
20	20	C7	0.630727	Female	35	23	
21	21	C6	0.543091	Male	35	24	

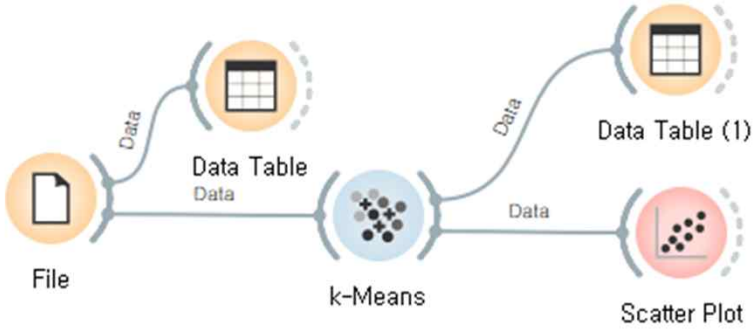
군집화 후 탐색 및 분석



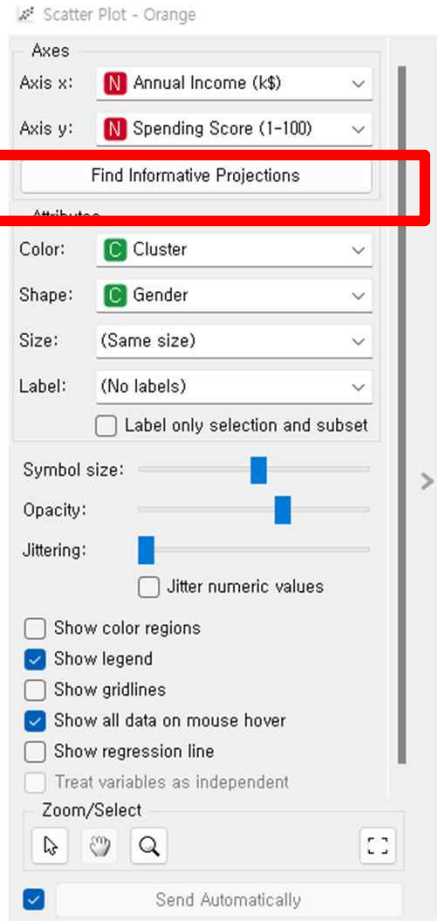


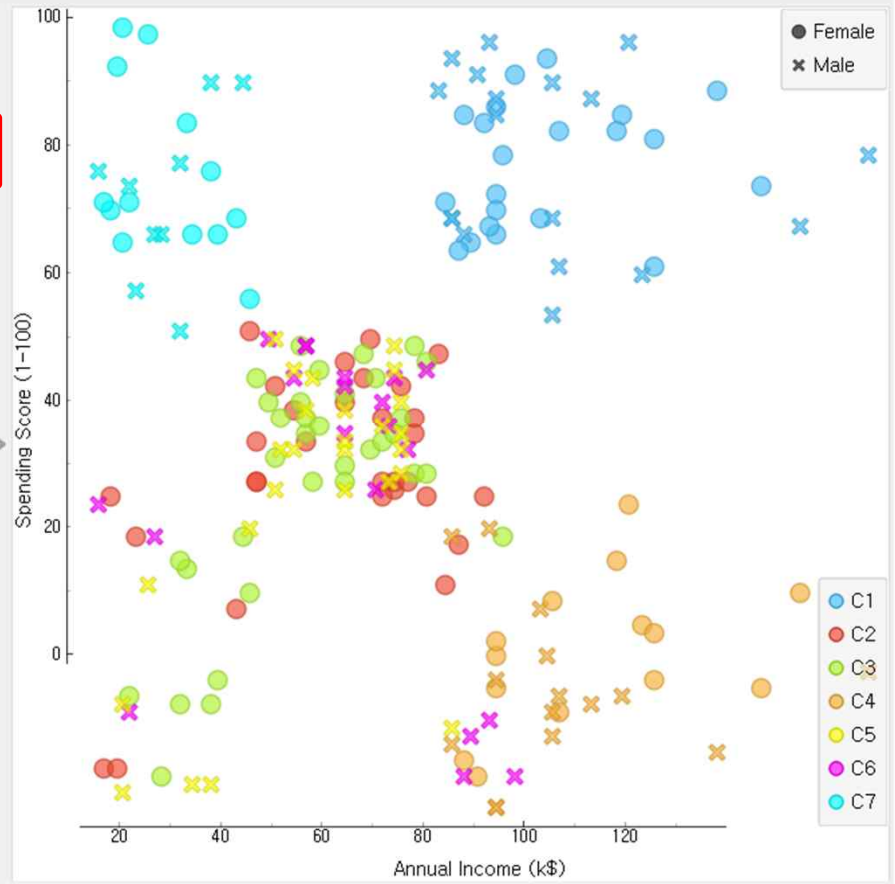
- C1,C7,C4 는 구분의 기준이 명확히 드러나지만 C2, C3, C5, C6 의 구분은 현재 상태에서 드러나지 않음.

군집화 후 탐색 및 분석

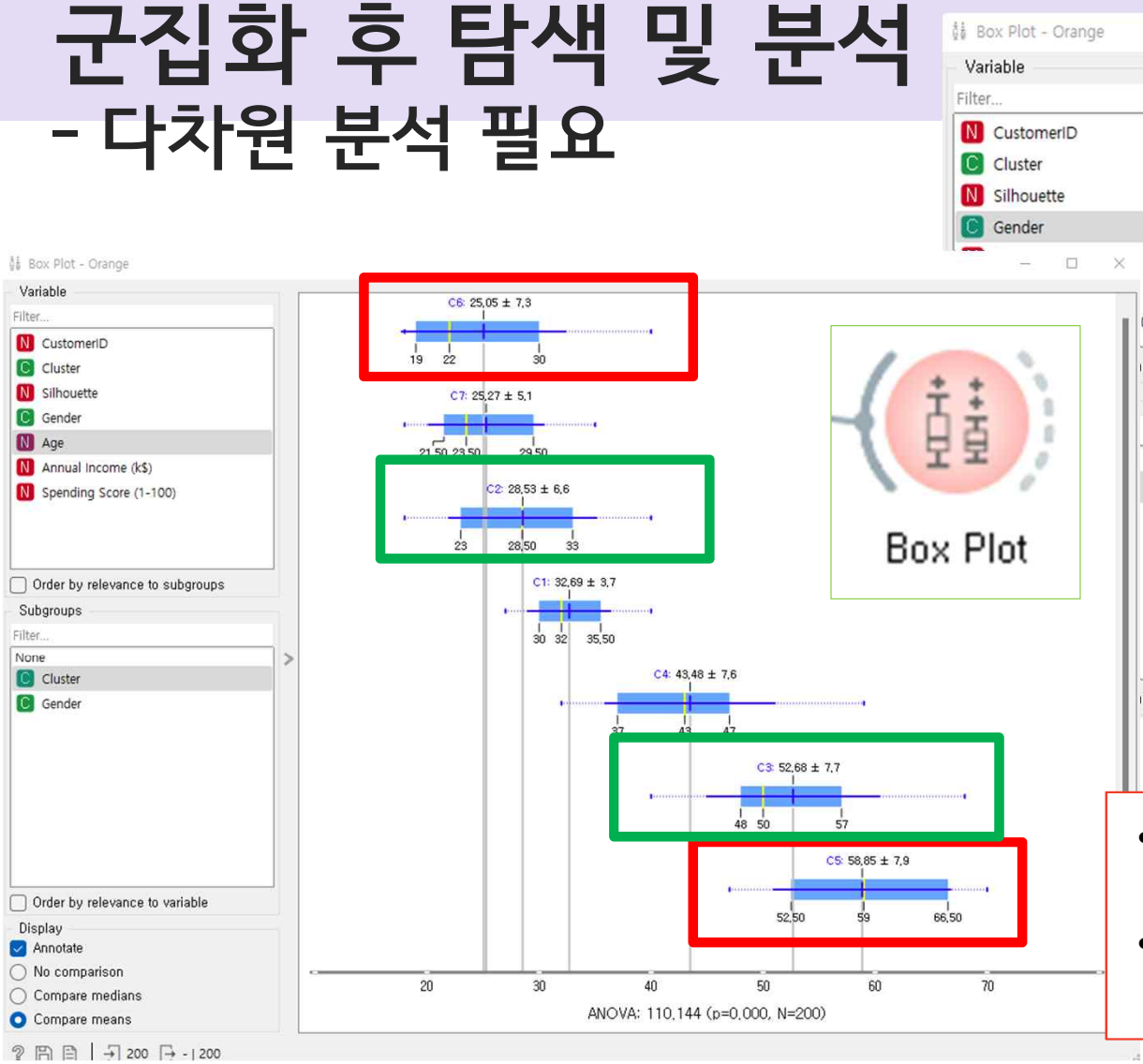

Data
Data Table
Data
Data
Data Table (1)
Data
File
k-Means
Scatter Plot

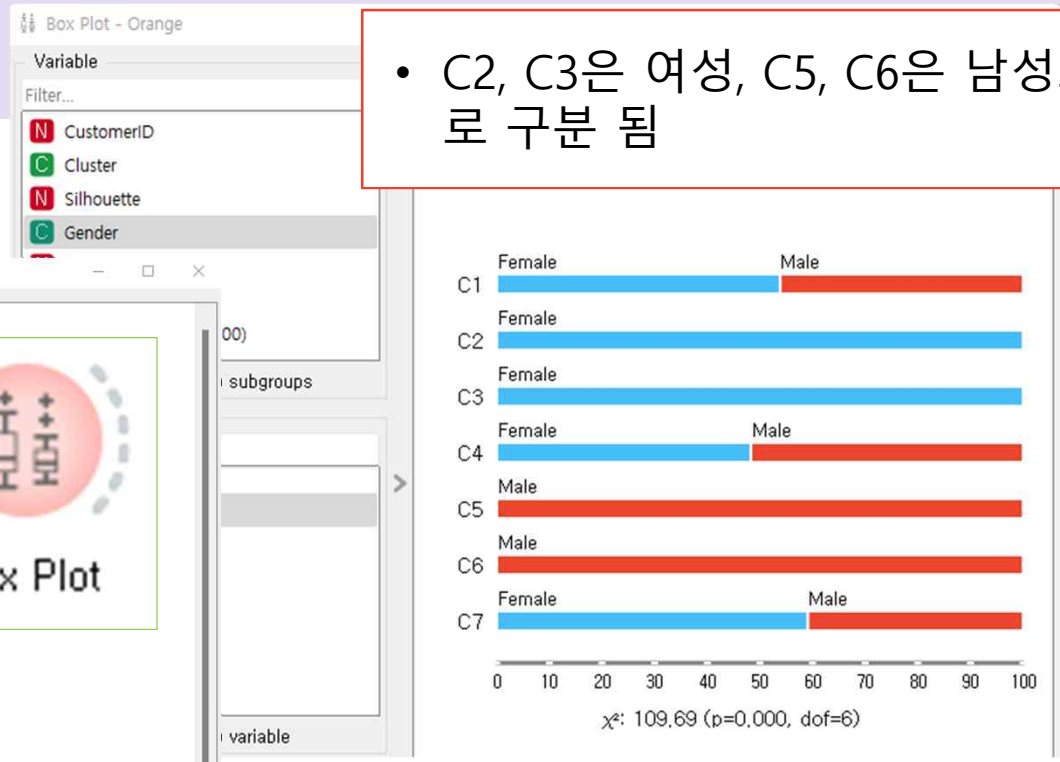

Score Plots - Orange
Filter ...
1 CustomerID, Spending Score (1-100)
2 Annual Income (k$), Spending Score (1-1...
3 Age, Spending Score (1-100)
4 Age, Annual Income (k$)
5 Age, CustomerID
6 Age, Gender
7 Age, Silhouette
8 Gender, Spending Score (1-100)
9 Silhouette, Spending Score (1-100)
10 Annual Income (k$), Silhouette
11 Annual Income (k$), Gender
Finished


Scatter Plot - Orange
Axes
Axis x: Annual Income (k$)
Axis y: Spending Score (1-100)
Find Informative Projections
Color: Cluster
Shape: Gender
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Show all data on mouse hover
Show regression line
Treat variables as independent
Zoom/Select
Send Automatically


Female
Male
C1
C2
C3
C4
C5
C6
C7
Spending Score (1-100)
Annual Income (k$)
0
20
40
60
80
100
120

군집화 후 탐색 및 분석
- 다차원 분석 필요





- C2, C3은 여성, C5, C6은 남성으로 구분 됨

- C2, C3 동일한 여성이나 40대를 기준으로 나이로 구분됨.
- C5, C6은 남성으로 역시 45세 영역에서 나이로 구분 됨

[활용]

- 유통사의 경우 수많은 점포를 군집으로 나누어 군집 별 전략 수립에 활용
- 고객세분화를 통한 시장 세분화된 마케팅 전략 수립

https://analysis-flood.tistory.com/27

다음 시간에는 **비정형데이터 중 이미지 데이터**를 활용한 - **분류와 군집화** 를 다루어 보도록 하겠습니다.