실습 4

k-Means를 사용한 택배 배송 위치 군집화

[Lecture] Dr. HeeSuk Kim

4

택배 배송 위치 군집화로 세상을 바꿔 봐!

k-Means를 사용하여 택배 배송 위치를 군집화해 보자.

데이터 종류:

정형 데이터

사용하는 모델:

K-Means

1 해결해야 할 문제는 무엇일까?

문제 상황

해를 거듭할수록 택배 수요는 폭발적으로 늘어나고 있으며, 택배 비용 문제도 이슈화되고 있다. 택배 수요는 늘고 있지만, 택배원 수급 및 택배 비용은 제자리걸음을 하고 있다. 이러한 상황을 분석하기 위해 특정 지역의 택배 위치 데이터로 군집화하고, 우리가 미처 보지 못했던 경향, 패턴 등을 파악하여 택배 배송 문제를 해결할 방법을 찾아보면 어떨까?

인천 연안 지역 택배 위치 데이터로 **군집화하여 택배 배송 문제를 해결**할 인공지능 모델을 만들어 보자.

2 데이터를 준비하자!

데이터 다운로드 링크

(https://bit.ly/3AFvAJo)

1 외부 데이터 다운로드

- Delivery.csv 파일 다운로드

- 371개의 인천 연안 지역 택배 배송 위치의 위도(Latitude)와 경도(Longitude) 확인 가능

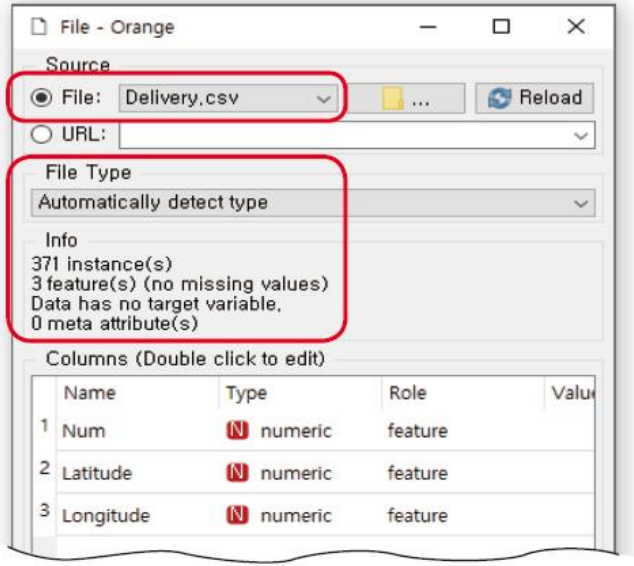
	A	B	C	D
1	Num	Latitude	Longitude	
2	1	37.3368	126.7128	
3	2	37.5013	126.7878	
4	3	37.5225	126.7774	
5	4	37.51118	126.7432	
6	5	37.50878	126.7385	
7	6	37.52849	126.7415	
8	7	37.511	126.779	
368	367	37.29901	126.833	
369	368	37.52454	126.6223	
370	369	37.49137	126.6781	
371	370	37.52737	126.6235	
372	371	37.45626	126.7052	

그림 4-1 인천 연안 지역 택배 배송 위치 데이터

2 데이터 불러오기

① 택배 배송 위치 데이터 불러오기

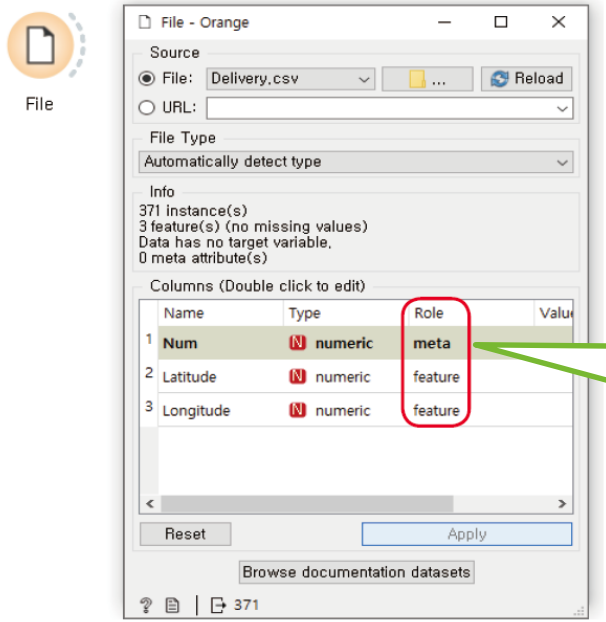
- Data 카테고리 – [File]위젯 가져온 후, 더블 클릭하여 Delivery.csv 파일 불러오기



데이터 정보(Info)로
3개의 속성으로 구성된
371개의 데이터가
담겨 있음을 확인

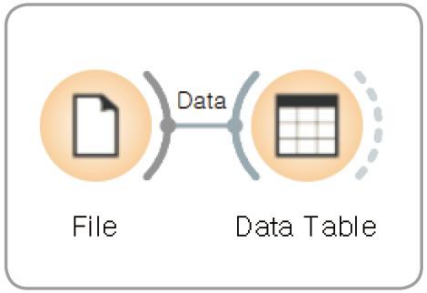
② 속성 역할(Role) 변경하기

- 3개의 속성 중 일련번호(Num)는 참고만 하기 위해 meta로 설정하고, 나머지 속성은 feature 로 설정



Num: meta
Latitude: feature
Longitude: feature

- Data 카테고리 – [Data Table] 위젯을 가져와서 [File]위젯과 연결
- [Data Table] 위젯을 더블 클릭하여 데이터 정보(Info) 살펴보기



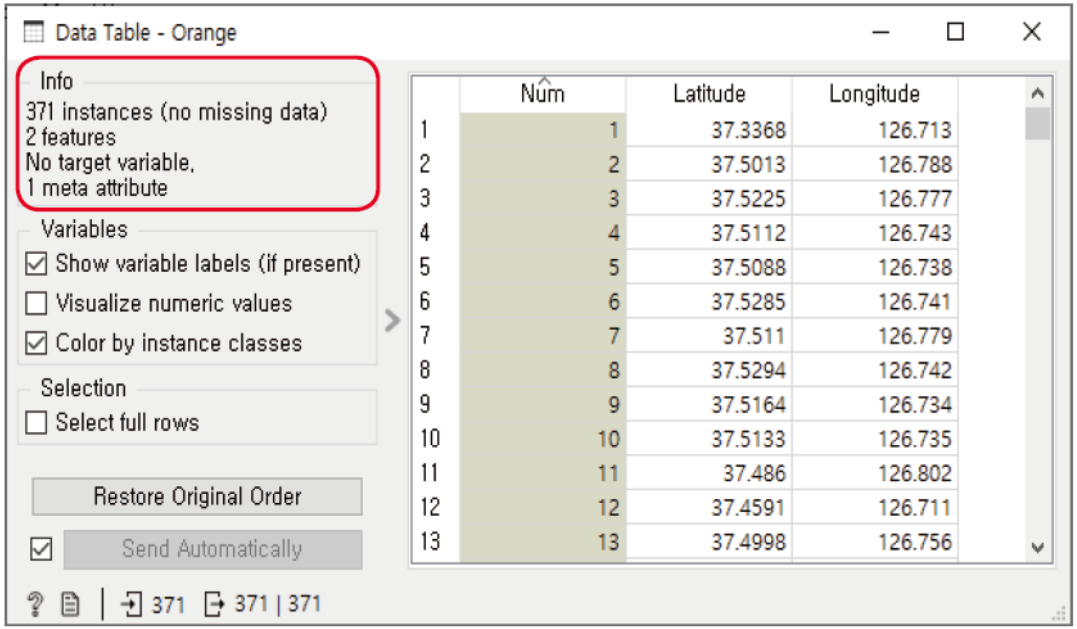


그림 4-2 속성 역할(Role)을 변경한 택배 배송 위치 데이터

3 데이터 속성 정보 확인하기

◆ 택배 배송 위치 데이터

속성명	속성 정보
Num	일련번호
Latitude	위도
Longitude	경도

Latitude와 Longitude는 기계학습에 영향을 미치는 feature이다.

AI랑 친해지기

군집화

(1) 군집화의 개념 및 종류

- 정답(레이블)이 **없는** 데이터로 학습을 하는 **비지도 학습**에서 가장 대표적인 것이 **군집화**이다.
  군집화는 결과에 대한 사전 지식은 없지만, 해당 데이터를 통해 의미가 있는 결과를 얻고자 할 때 사용하는 기계학습이다.

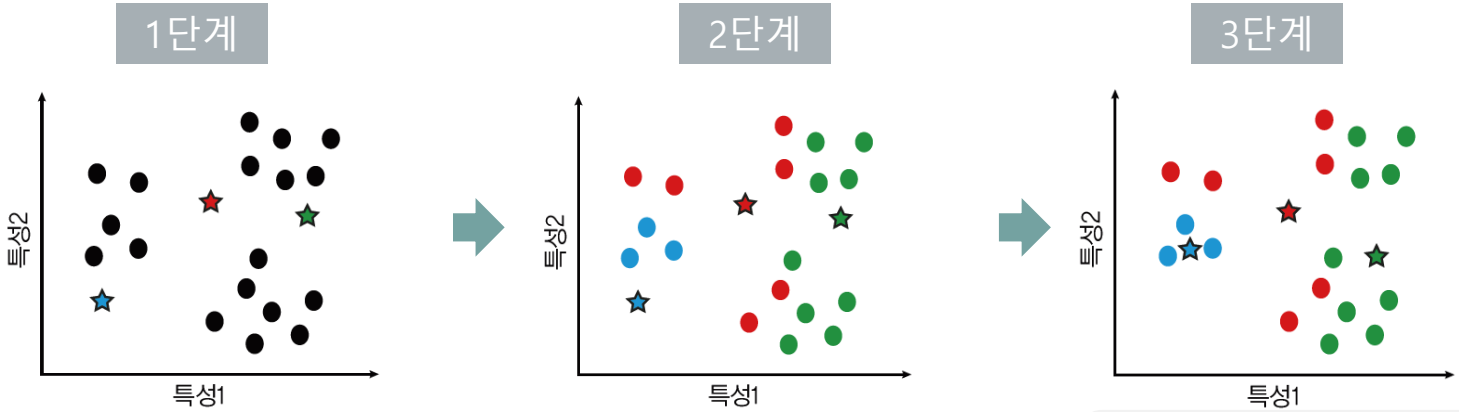
- 군집화 종류로는 k-Means, DBSCAN, Hierarchical clustering 등이 있다.
  데이터 특성에 따라 속도나 군집성능에 차이가 있으므로, 데이터 특성에 따라 적절한 군집화를 선택한다. 이 활동에서는 k-Means를 사용한다.

(2) k-Means(k-평균)의 군집화 과정

요약 정리

• k-Means는 **k개의 중심점**을 찍은 후, 이 **중심점**에서 **각 점 간 거리**의 **합**이 **최소화**가 되는 **중심점 k개**의 위치를 찾고, **다시** 이 **중심점**에서 **가장 가까운 점들을 기준으로 묶는 작업**을 통해 **군집화**한다.

(2) k-Means(k-평균)의 군집화 과정

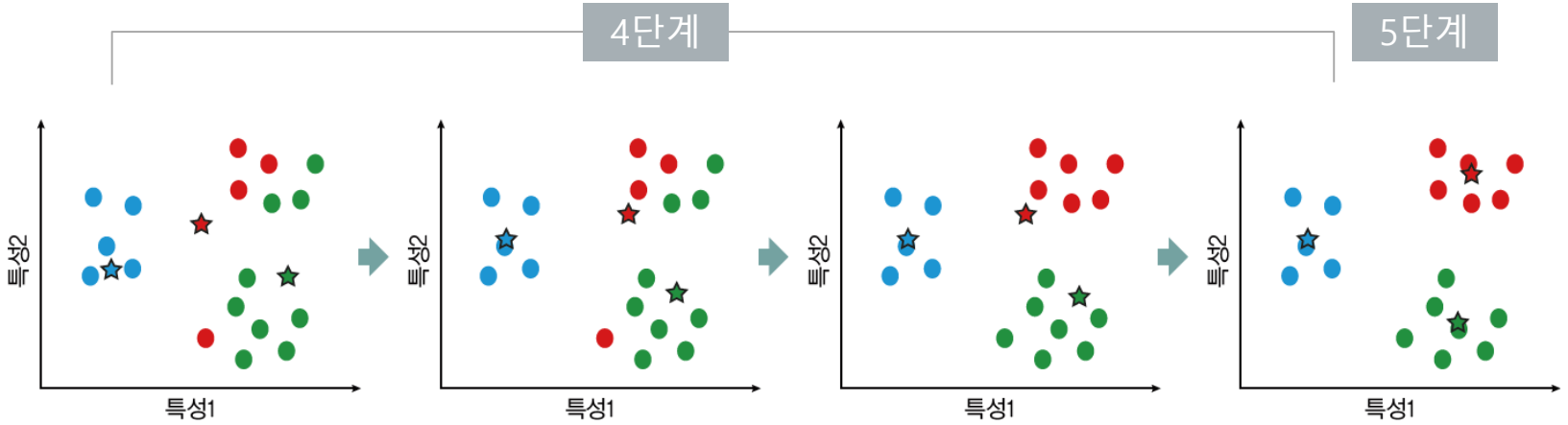


데이터 셋에서 k개의 **중심(centroid)을 임의로 지정**

중심으로부터 모든 데이터가 얼마나 떨어져 있는지 계산한 후, **각 데이터에서 가장 가까운 중심 지정**

[2단계] 과정에서 할당된 결과를 바탕으로 **중심을 새롭게 지정**

(2) k-Means(k-평균)의 군집화 과정



[2단계]~[3단계] 과정을 반복한다.

중심이 더는 변하지 않으면 멈춘다.

3개의 별표가 중심(centroid)이 되고,이 과정에서는 k는 3이다.

그림 4-3 k-Means의 군집화 과정

3 어떤 모델을 선택하고 학습시킬까?

1 학습 모델 선택하기

① 모델 선택하기

- Unsupervised(비지도 학습) 카테고리의 여러 비지도 학습 모델 중 [k-Means] 위젯을 캔버스로 가져와서 [File] 위젯과 연결

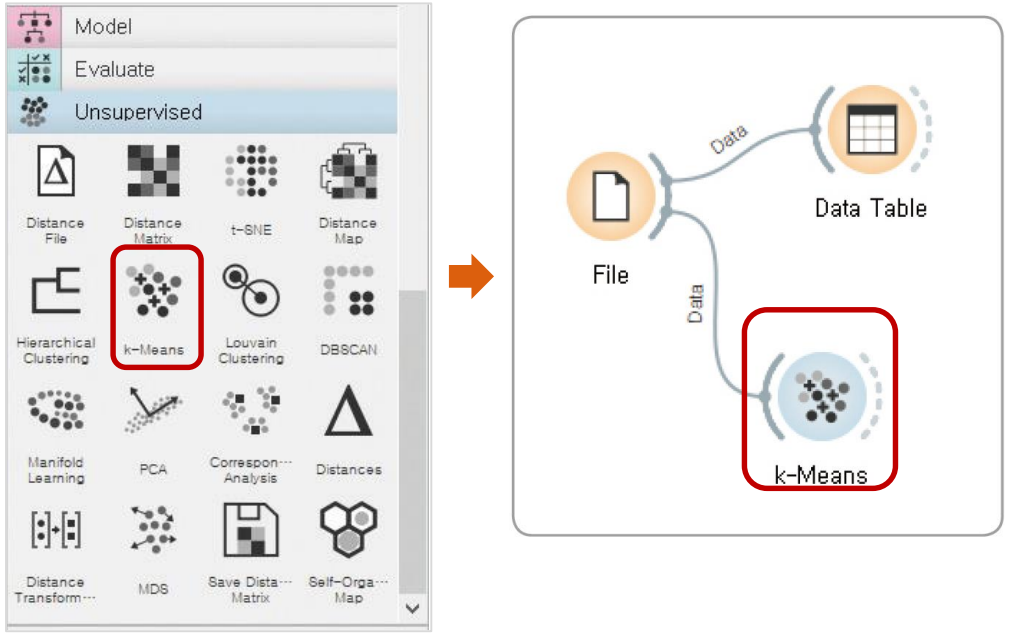
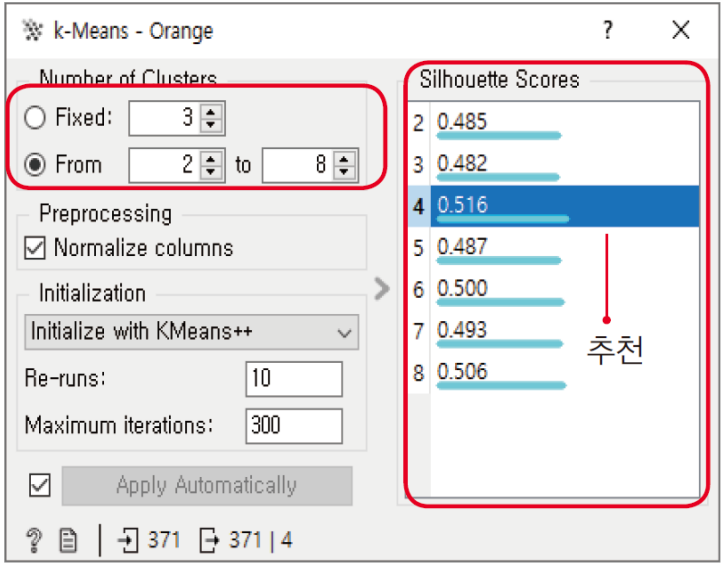


그림 4-4 [k-Means] 위젯을 [File] 위젯에 연결

② [k-Means] 위젯 살펴보기

- [k-Means] 위젯을 더블 클릭한 후 아래와 같은 대화 창이 나타나면 설정 변경을 통해 가장 적합한 군집 개수를 파악하여 군집화함.



Number of Clusters(군집 개수)
- **Fixed**: 내가 원하는 군집의 개수를 설정한다.
- **From**: 실루엣 점수(Silhouette Scores)를 보여 주는 **범위**를 설정한다.

Silhouette Scores: 해당 범위 내에서 **가장 높은 점수의 군집 개수를 추천**한다.

그림 4-5 [k-Means] 위젯 창

2 모델 학습시키기

① 군집 설정하기

- [그림 4-5]에서 확인한 결과, 가장 높은 실루엣 스코어(Silhouette Scores)가 0.516이므로 [그림 4-6]에서 k(Fixed)를 4로 설정

- Fixed값으로 군집 개수를 설정하면 출력으로 설정한 개수만큼 군집 생성

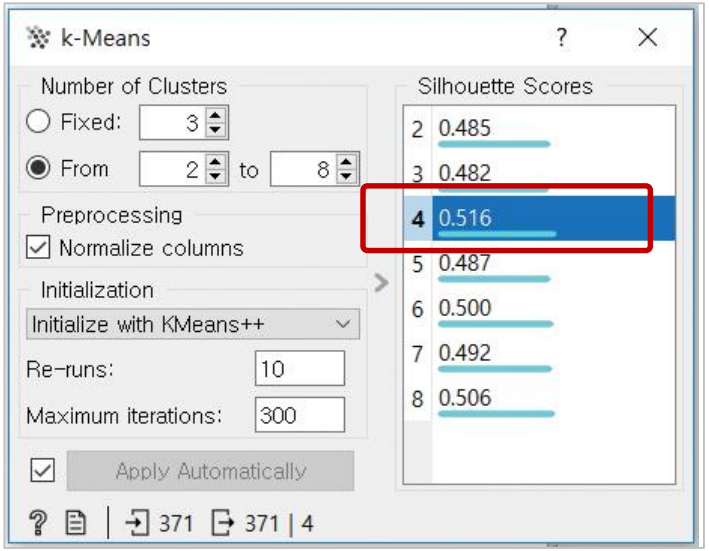
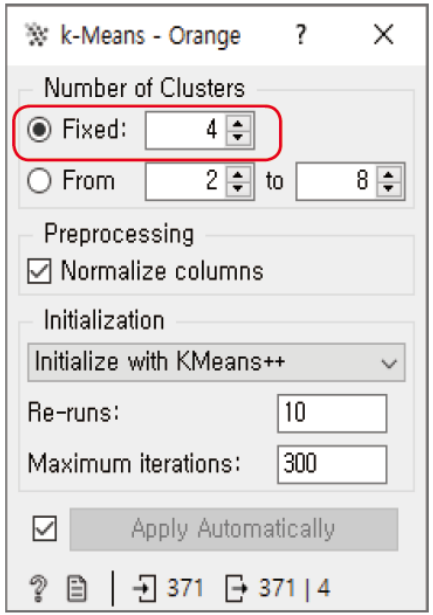


그림 4-5 [k-Means] 위젯 창

② 출력 확인하기

- [k-Means] 위젯을 더블 클릭한 후 아래와 같은 대화 창이 나타나면 설정 변경을 통해 가장 적합한 군집 개수를 파악하여 군집화함.



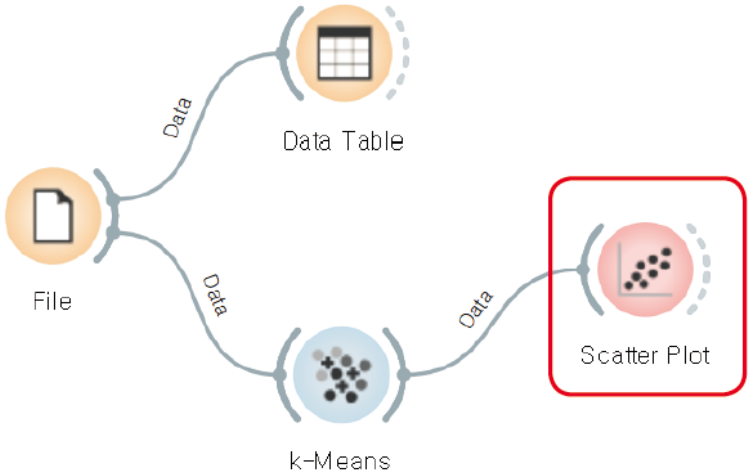


그림 4-6 군집 개수를 4로 설정한 [k-Means] 위젯에 [Scatter Plot] 위젯 연결

4 모델의 성능을 확인해 보자!

1 학습 결과 확인하기

① 군집화 그래프 확인하기

- 연결한 [Scatter Plot] 위젯 더블 클릭
  → [그림 4-7]과 같이 산점도 형태의 군집화한 그래프 출력

- 군집별로 색을 다르게 설정하면 군집이 어떻게 형성되었는지 쉽게 파악할 수 있음.

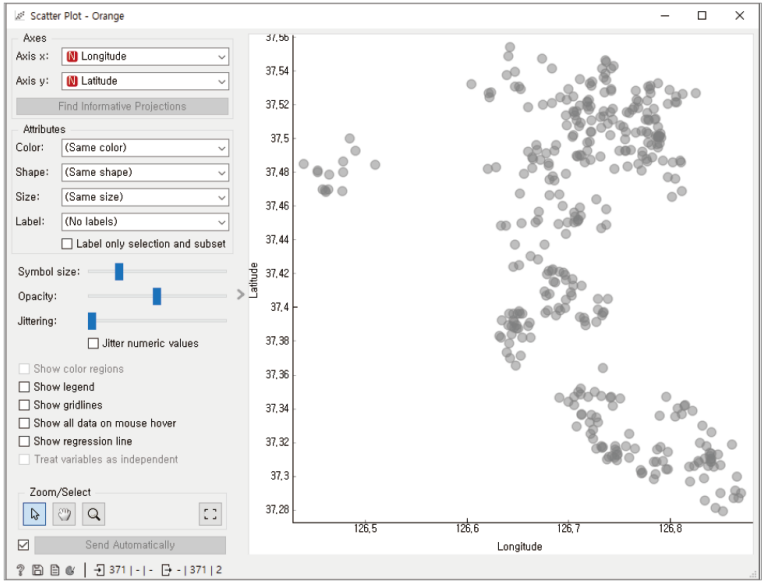


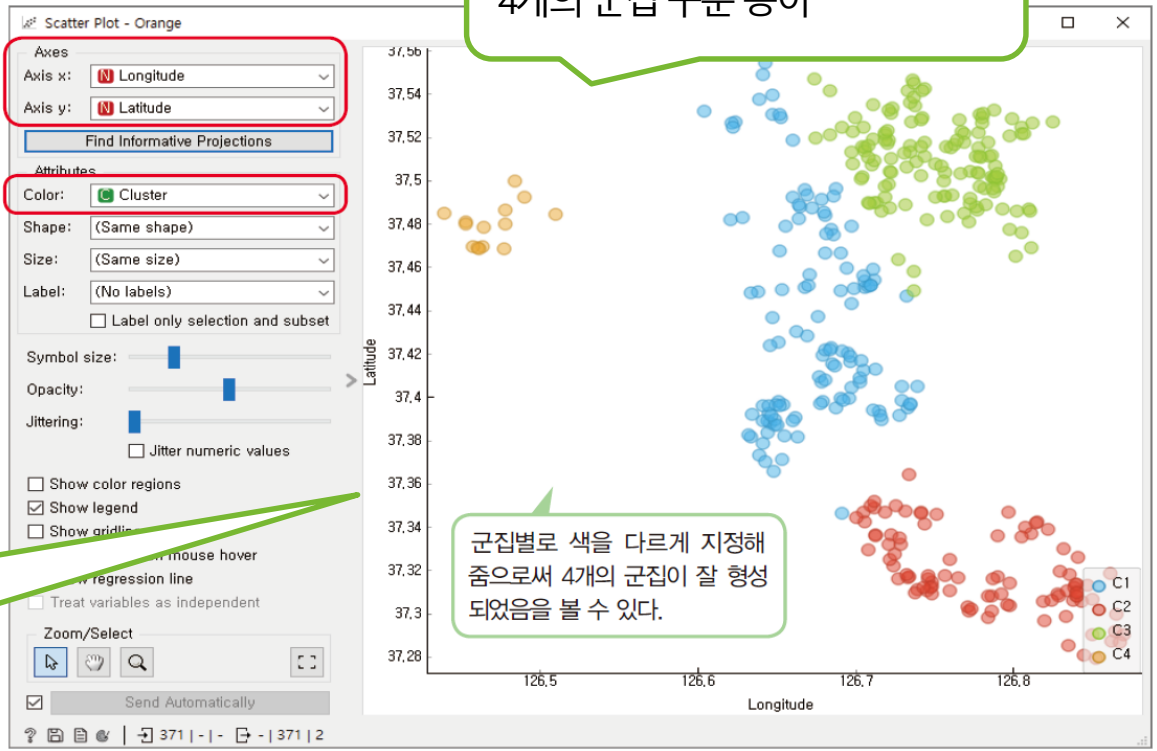
그림 4-7 k-Means 모델로 형성된 군집화의 시각화

② x, y축 설정과 색 지정하기

x축: 경도(Longitude)
y축: 위도(Latitude)

color: 군집(Cluster)

군집별로 색을 다르게 지정하면 4개의 군집 구분 용이



군집별로 색을 다르게 지정해 줌으로써 4개의 군집이 잘 형성되었음을 볼 수 있다.

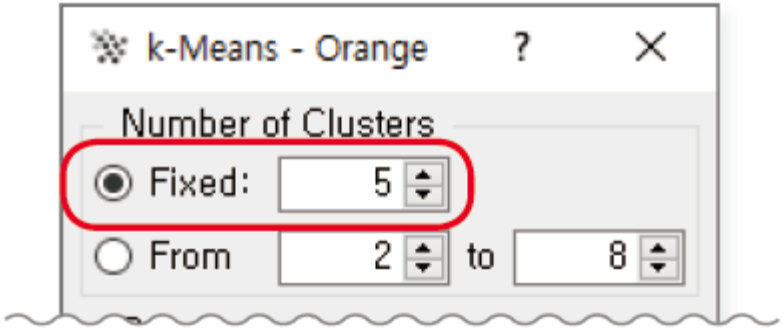
군집 색상은 Orange 3 버전에 따라 다르게 나타날 수 있다.

그림 4-8 색으로 구분한 4개의 군집

2 성능 결과 확인하기

① 군집 점 늘리기

- 군집 점의 수를 다양하게 늘렸을 때의 군집 상황을 살펴보기 위해 [k-Means] 위젯을 더블 클릭하여 Fixed의 값을 5, 6, 7, 8로 바꾸어 가면서 택배 거점에 따른 군집 점 형성



군집 색상은 Orange 3 버전에 따라 다르게 나타날 수 있다.

군집 점이 5일 때(k=5) **군집 점이 6일 때(k=6)** **군집 점이 7일 때(k=7)** **군집 점이 8일 때(k=8)**

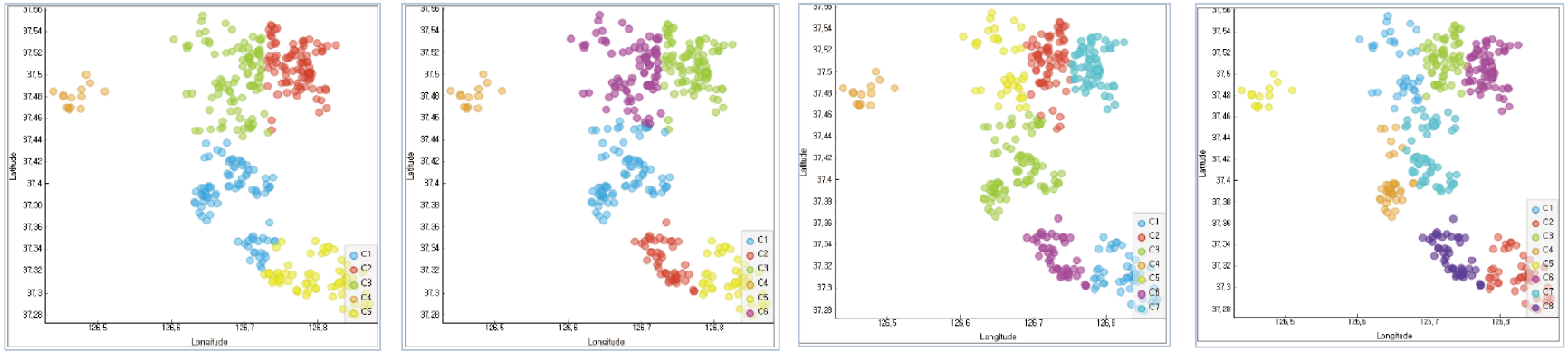
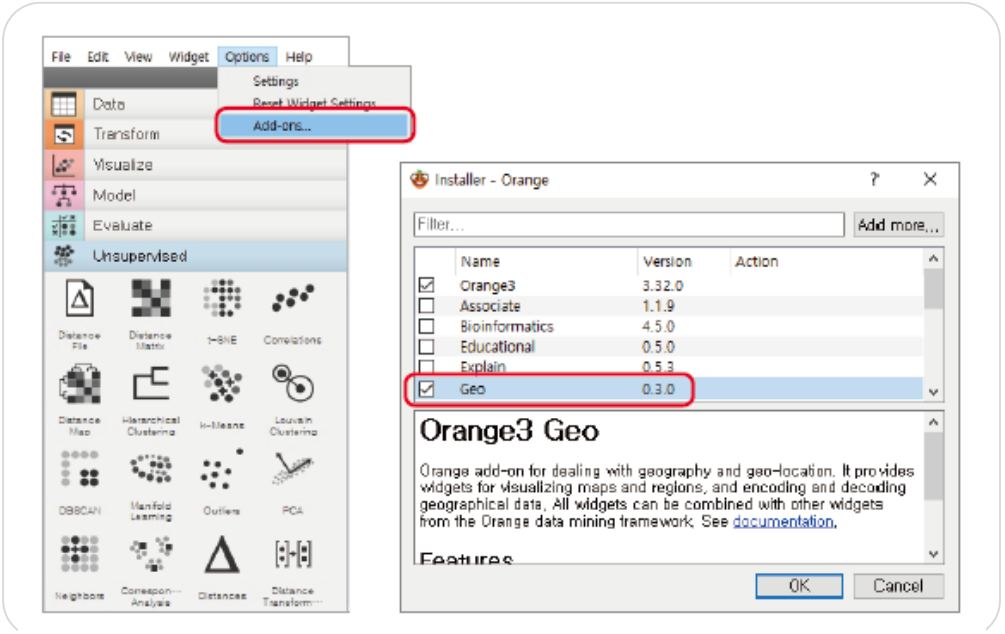


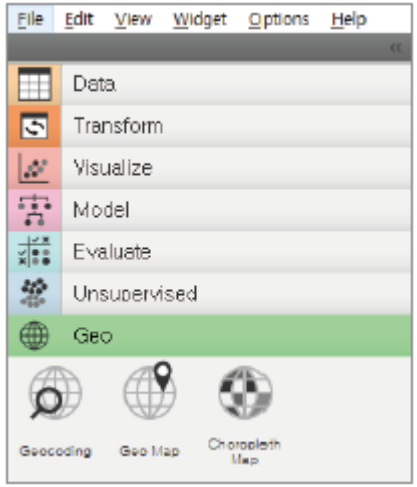
그림 4-9 군집 점 수에 따라 형성된 군집

일반적으로 군집 점은 보통 4개 이상 존재하지만 [그림 4-9]처럼 군집 점이 5에서 8로 늘어날수록 더 작은 군집이 형성되면서 세분화되는 것을 알 수 있다.

② 인천 연안 지도로 보기

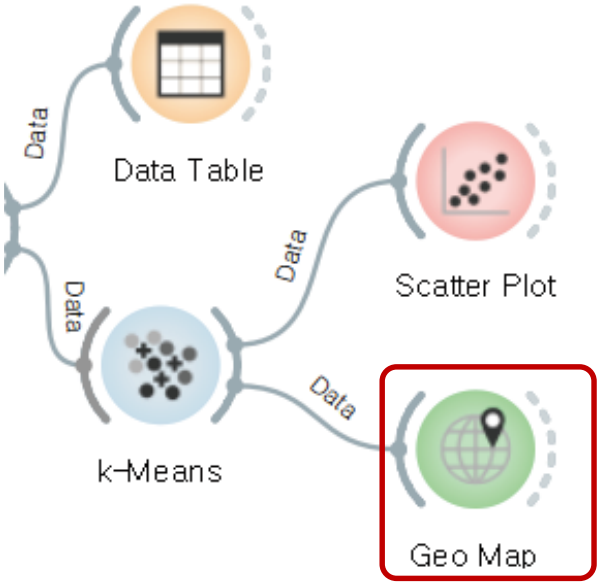
- Options 메뉴에서 [Add-ons]를 클릭하면 [Installer] 창이 나타난다.
- Geo에 체크(☑) 하고 OK를 클릭하면 Geo 카테고리가 추가로 설치된다.



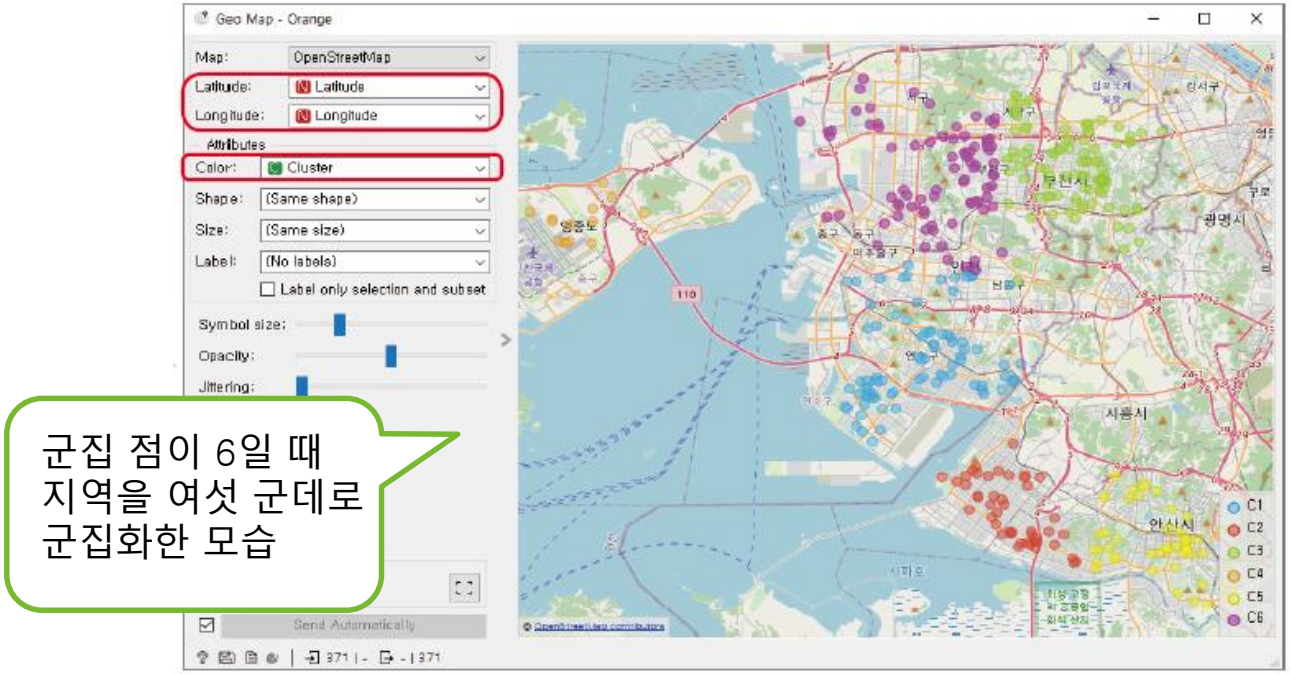


Geo를 [Add-ons]를 한 후 Orange3를 재실행해야 Geo카테고리가 나타난다.

- [k-Means] 위젯에 Geo 카테고리의 [Geo Map] 위젯을 새롭게 연결



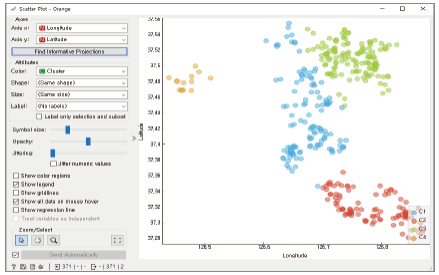
- 위도와 경도를 설정
- Orange3 내에서도 군집화한 곳에 지도를 동시에 표시할 수 있음.
- 색상은 군집(Cluster)으로 설정하여 군집화된 곳이 어느 지역인지 쉽게 가능



군집 점이 6일 때 지역을 여섯 군데로 군집화한 모습

- [Geo Map] 위젯을 이용하면
- k-Means 군집화와 인천 연안 지역 지도를 합한
- 지역별 군집화한 모습 확인 가능

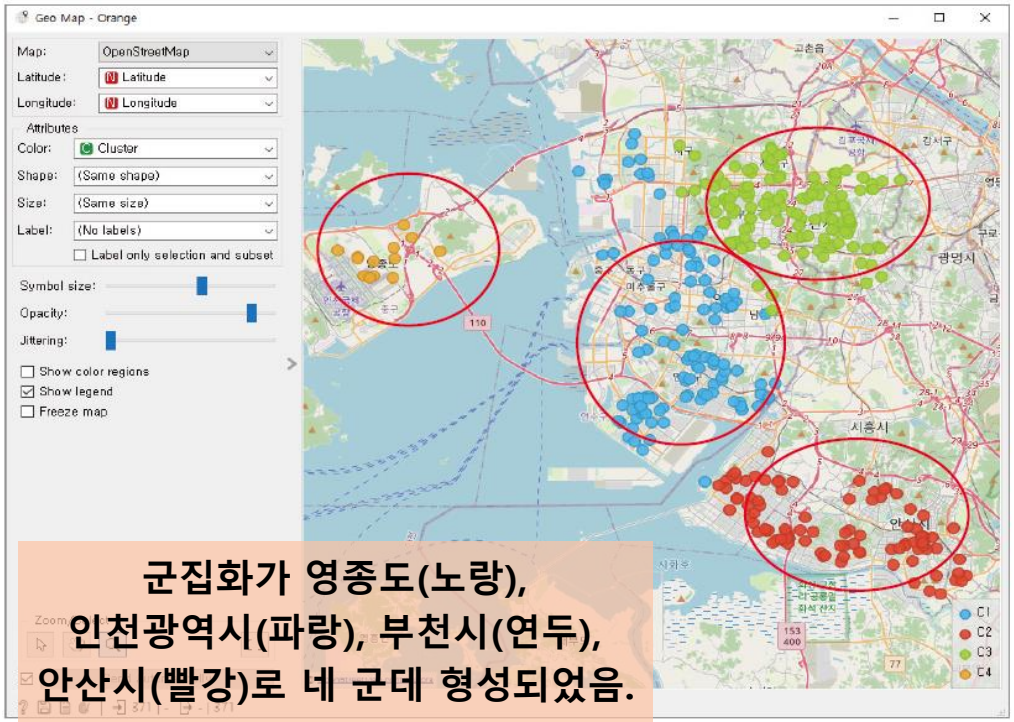
[준비 데이터]



k-Means로 군집화

인천 연안 지도

[실행 결과]

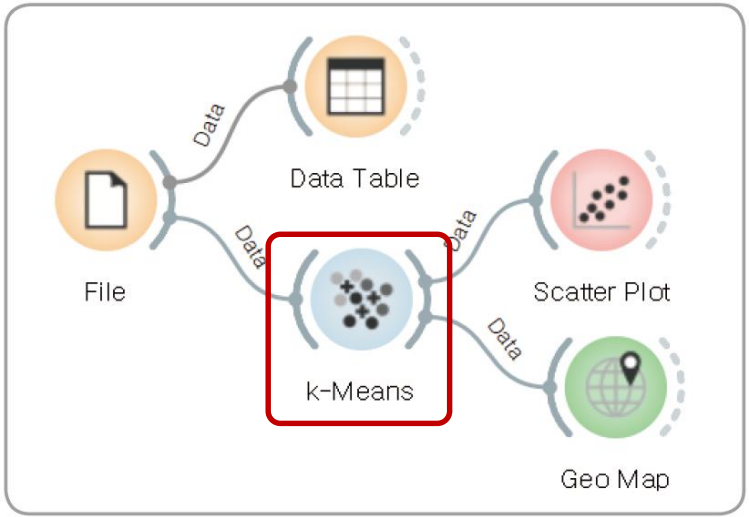


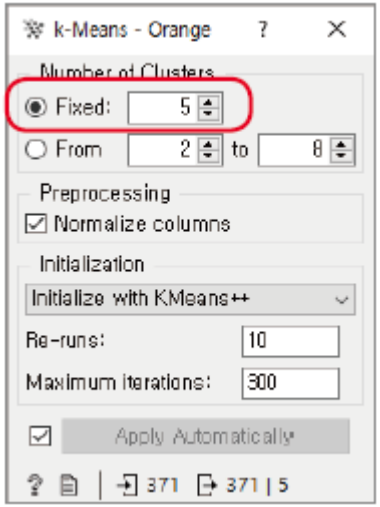
군집화가 영종도(노랑), 인천광역시(파랑), 부천시(연두), 안산시(빨강)로 네 군데 형성되었음.

인천 연안 지도에 지역별로 군집화한 모습

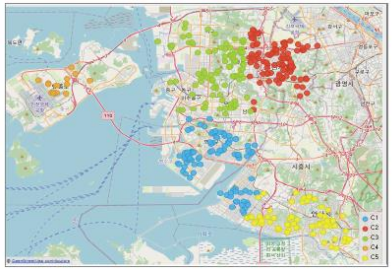
② 군집 살펴보기

- [k-Means] 위젯을 더블 클릭하여 Fixed의 값을 5, 6, 7, 8로 바꾸어 가면서 인천 연안 지역 지도와 택배 위치 데이터를 활용하여 군집 상황을 살펴보도록 한다.

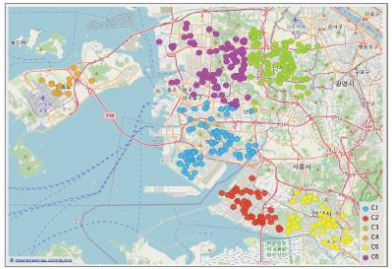




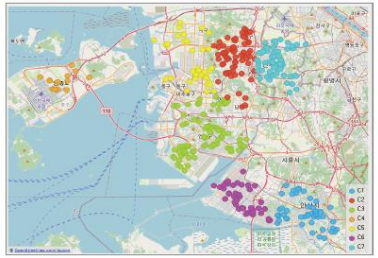
군집 점이 5일 때(k=5)



군집 점이 6일 때(k=6)



군집 점이 7일 때(k=7)



군집 점이 8일 때(k=8)

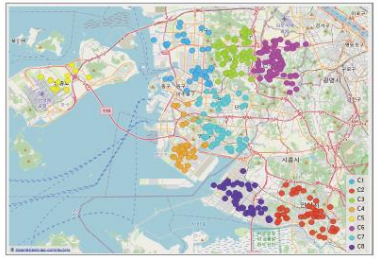


그림 4-10 군집 점을 기반으로 한 모델 평가

위와 같은 군집 상황을 분석하여 택배 집하장 선정 등에 적용할 수 있다.
최근에는 코로나의 영향으로 택배 물류 양이 계속 늘어나고 있다.
이에 따라 유류비, 인건비, 택배 거점 등 고려해야 할 것 또한 한둘이 아니다.

따라서 인천광역시, 안산시 등의 **행정 구역 기반의 택배 거점**을 **군집 점을 기반으로 변경**한다면 **비용을 절약**할 뿐만 아니라 택배 집하장 선정 등을 최적화할 수 있다.

Orange3 장점

Q Geo 카테고리의 역할은 무엇이고 어떤 위젯이 있나요?

A 데이터에 포함된 위도와 경도 정보를 활용하여 데이터를 지도에 표시하여 시각화 기능을 제공하는 카테고리이다.

[Geocoding(지오코딩)] 위젯: 위도와 경도 정보를 이용하여 지역 이름을 알려 주거나 그 반대 작업을 하는 위젯

[Geo Map(지리 지도)] 위젯: 지도에 데이터를 원으로 표시하는 위젯

[Choropleth Map(등치 지도)] 위젯: 선택한 통계 변수에 따라 영역의 색을 달리하여 지도에 데이터를 시각화하는 위젯

정리하기

k-Means 모델을 이용하여 서해안 지역의 택배 위치 데이터로 군집화를 해 보았다.

현재 택배 집하장 선정 등은 행정 구역 위주로 운용된다.
이 군집화를 통해 행정 구역 바탕이 아닌 군집화된 지역을 중심으로 운용하면
인건비, 유류비 등의 비용을 절감할 수 있다.

이처럼 군집화는 데이터에서 내가 알지 못했던 것을 발견하는 방법으로
사람이 파악하기 힘든 본질적인 문제나 숨겨진 특징 및 구조를 연구할 때 주로 사용
군집화와 같은 비지도 학습을 한 후 그 결과를 토대로 지도 학습을 다시 하는 경우
도 있다.

Q & A