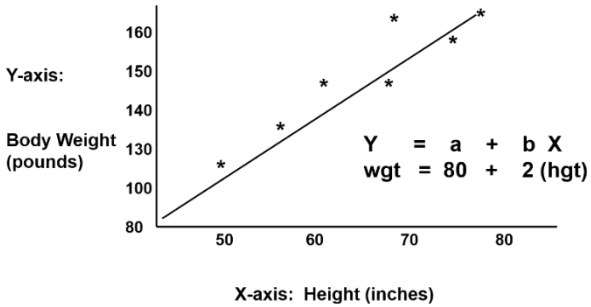
Smart Technology - AIR

Cluster Analysis

방법론적 구분

- 비지도 학습 (Unsupervised Learning)
  - 입력(X) 벡터만 주어짐
  - 군집 (Clustering)
- 지도 학습 (Supervised Learning)
  - 입력(X) 과 출력(Y)이 쌍으로 주어짐
  - 회귀(Regression)와 분류(Classification)



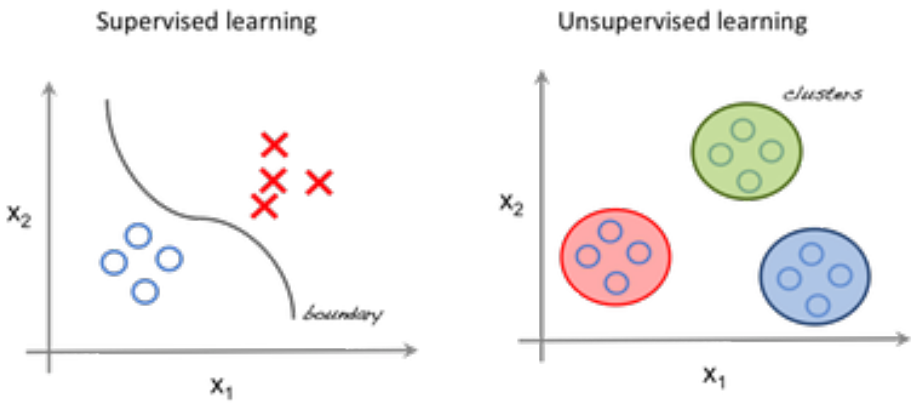
선형회귀 (Linear Regression)

- 강화 학습 (Reinforcement Learning)
  - 주어진 환경에서 행동에 대한 보상을 통해 최적의 행동을 의사결정
- 시계열 분석 (Time Series Analysis)
  - 시간의 흐름에 따라 수집된 자료를 분석
  - 예측 (Forecasting)

비지도학습

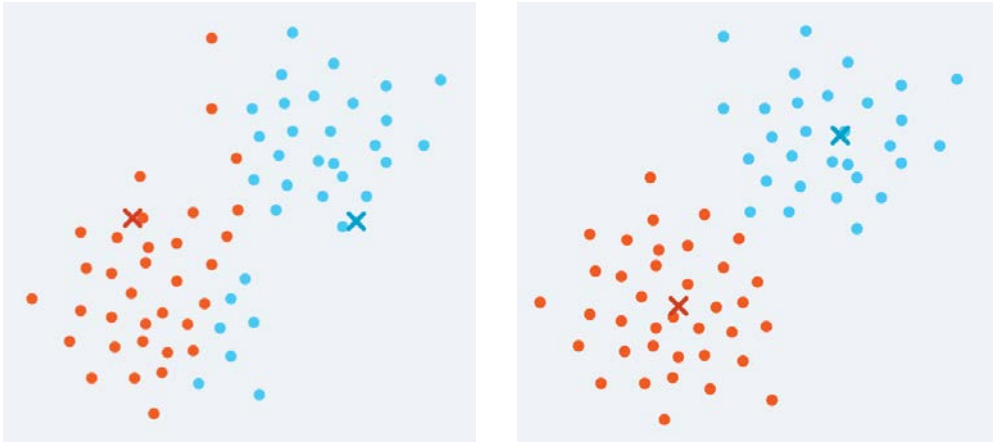
- 비지도학습 (Unsupervised Learning)
  - 학습데이터가 레이블을 갖고 있는 않는 경우, 데이터의 형태가 어떻게 구성되어 있는지를 알아내는 문제
  - 레이블이 없는 데이터의 요약 정보 추출
  - 요약 정보를 통한 전체 데이터의 특징 발견

- 군집 분석 (Cluster Analysis)
  - 주어진 데이터셋 내에 존재하는 몇 개의 군집을 찾아내는 방법

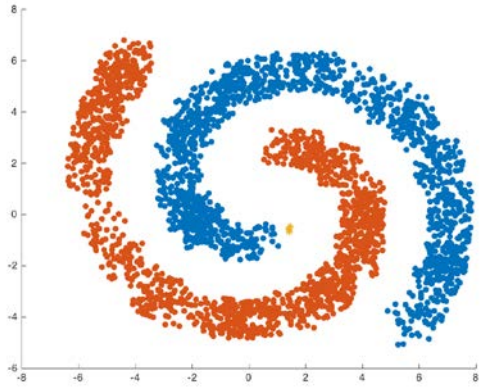


군집 분석

- 분할 기반 군집 (Partition-based clustering)
  - K-means: k 개의 클러스터 중심(좌표 평면의 값)을 찾고, 가장 가까운 중심을 데이터의 클러스터에 포함시킴
  - K-medoids: 데이터 세트의 값의 중심점으로 클러스터를 결정
  - DBSCAN(Density Based Spatial Clustering of Application with Noise): 특정 거리 내의 데이터 밀도를 이용하여 클러스터를 결정



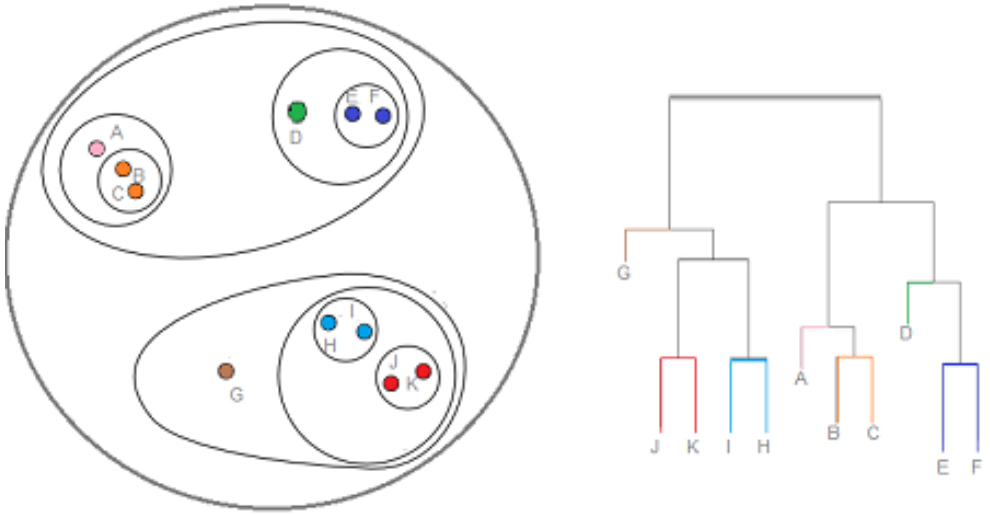
k-means



DBSCAN

군집 분석

- 계층형 군집 (Hierarchical clustering)
  - 병합적 군집: 모든 데이터를 단일 클러스터로 정의하고, 유사성이 높은 두개의 클러스터를 합하는 방식 (Bottom-up 방식)
  - 분할적 군집: 모든 데이터를 포함하는 단일 클러스터로 정의하고, 유사성이 낮은 두개의 클러스터로 분할하는 방식 (Top-down 방식)



군집 분석의 활용

- 고객 세분화: 고객 군집별로 다른 제품과 메시지를 제공
  - 미디어 이용 행태의 차이
  - 개인의 가치관과 라이프 스타일의 차이
- 신용 위험 분류: 과거 신용 기록을 바탕으로 고객 분류
- 도시 계획: 가구의 유형, 가치, 위치 등을 바탕으로 분류
- 생물학: 동식물의 분류

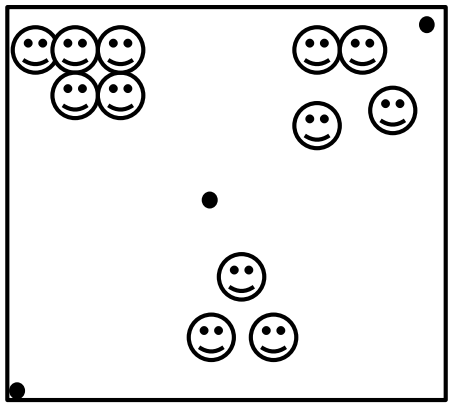
군집 분석의 장단점

- 장점
  - 그룹을 분류할 수 있는 세부 기준이 없는 상황에서, 데이터를 기반으로 그룹을 형성할 수 있음
  - 그룹별로 차별적 속성을 파악하여 효과적으로 대응할 수 있음

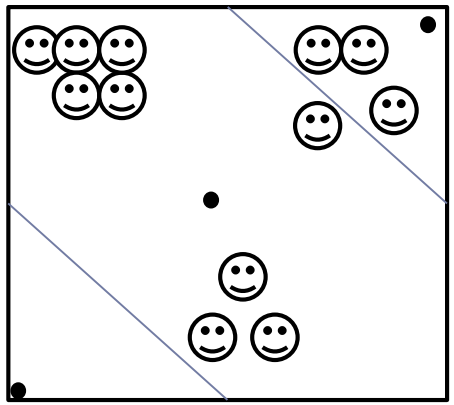
- 단점
  - 비계층적 군집 분석의 경우, 사용자가 그룹의 수를 결정해줘야 하며 결과가 잘 나오지 않을 수 있음
  - 분석 결과에 대한 해석이 어려울 수 있음
  - 타당성 점검이 어려우며, 이상치에 민감하게 반응 함

K-means

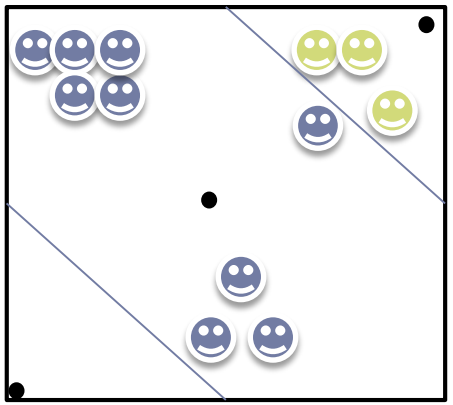
- 공간상에 k개의 중심점(cluster centroid)을 정하고, k개의 클러스터로 구분
  - 의사결정자가 k를 사전에 정해야 함
  - Voronoi diagram: 평면위에 점이 하나 씩 포함되는 다각형으로 분할



초기 중심점



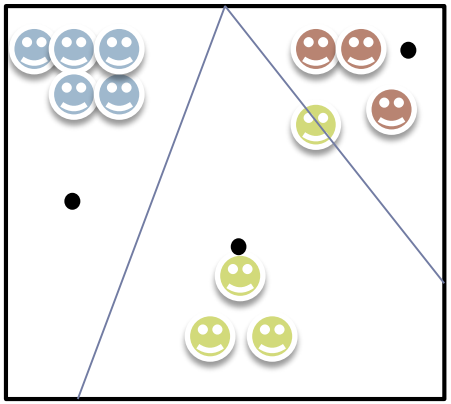
클러스터의 구분



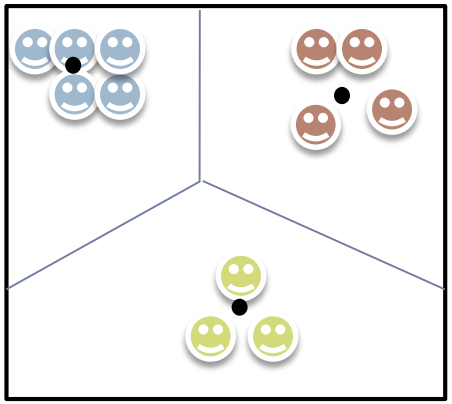
클러스터 할당

K-means

- 공간상에 k개의 중심점(cluster centroid)을 정하고, k개의 클러스터로 구분
  - 의사결정자가 k를 사전에 정해야 함
  - Voronoi diagram: 평면위에 점이 하나 씩 포함되는 다각형으로 분할



중심점 이동



최적의 3-means 클러스터링

E-Mail 마케팅 문제

- Wine 도매점
  - 광고 메일의 홍수 속에서 소비자 타겟 마케팅을 할 수 있는 방법은?
  - 초기 Data (OfferInformation): 작년에 진행된 판매 제안 데이터

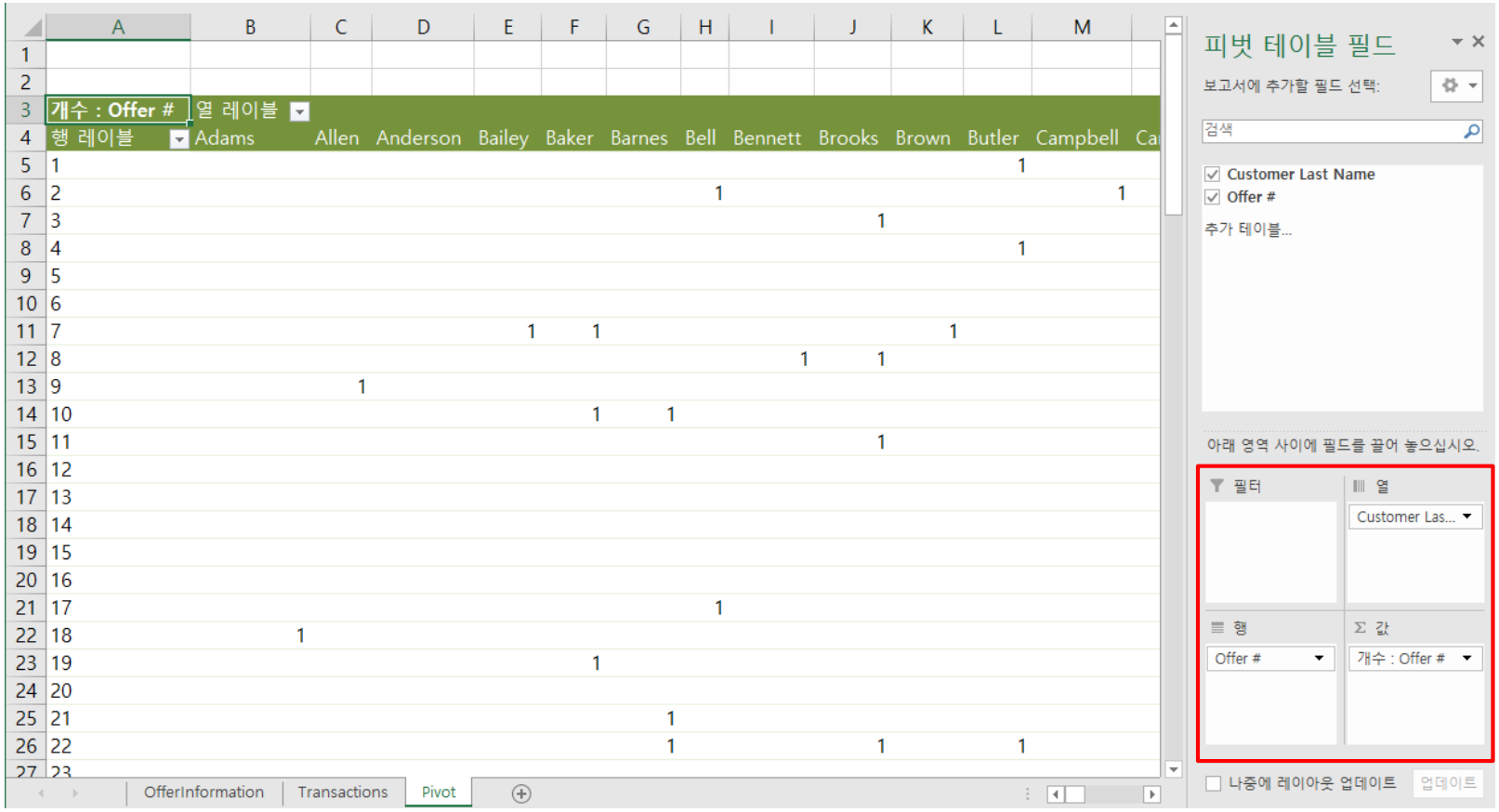
	A	B	C	D	E	F	G
1	Offer #	**Campaign**	**Varietal**	**Minimum Qty (kg)**	**Discount (%)**	**Origin**	**Past Peak**
2	1	January	Malbec	72	56	France	FALSE
3	2	January	Pinot Noir	72	17	France	FALSE
4	3	February	Espumante	144	32	Oregon	TRUE
5	4	February	Champagne	72	48	France	TRUE
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE

  - 초기 Data (Transactions): 작년에 진행된 32번의 판매 제안으로 성사된 거래

	A	B
1	Customer Last Name	Offer #
2	Smith	2
3	Smith	24
4	Johnson	17
5	Johnson	24
6	Johnson	26

거래 데이터의 정렬

- 판매제안별, 소비자별 데이터
  - Transactions를 이용한 Pivot Table 생성



데이터의 병합

- 판매제안과 거래데이터 합하기
  - 새로운 tab 생성 (Matrix)
  - OfferInformation 데이터 Copy & Paste
  - (Offer #제외하고..) Pivot 데이터 Copy & Paste

OfferInformation

	A	B	C	D	E	F	G
1	Offer #	Campaign	Varietal	Minimum	Discount (	Origin	Past Peak
2	1	January	Malbec	72	56	France	FALSE
3	2	January	Pinot Noir	72	17	France	FALSE
4	3	February	Espumante	144	32	Oregon	TRUE
5	4	February	Champagn	72	48	France	TRUE
6	5	February	Cabernet S	144	44	New Zeala	TRUE
7	6	March	Prosecco	144	86	Chile	FALSE
8	7	March	Prosecco	6	40	Australia	TRUE
9	8	March	Espumante	6	45	South Afri	FALSE
10	9	April	Chardonna	144	57	Chile	FALSE
11	10	April	Prosecco	72	52	California	FALSE
12	11	May	Champagn	72	85	France	FALSE
13	12	May	Prosecco	72	83	Australia	FALSE
14	13	May	Merlot	6	43	Chile	FALSE
15	14	June	Merlot	72	64	Chile	FALSE
16	15	June	Cabernet S	144	19	Italy	FALSE

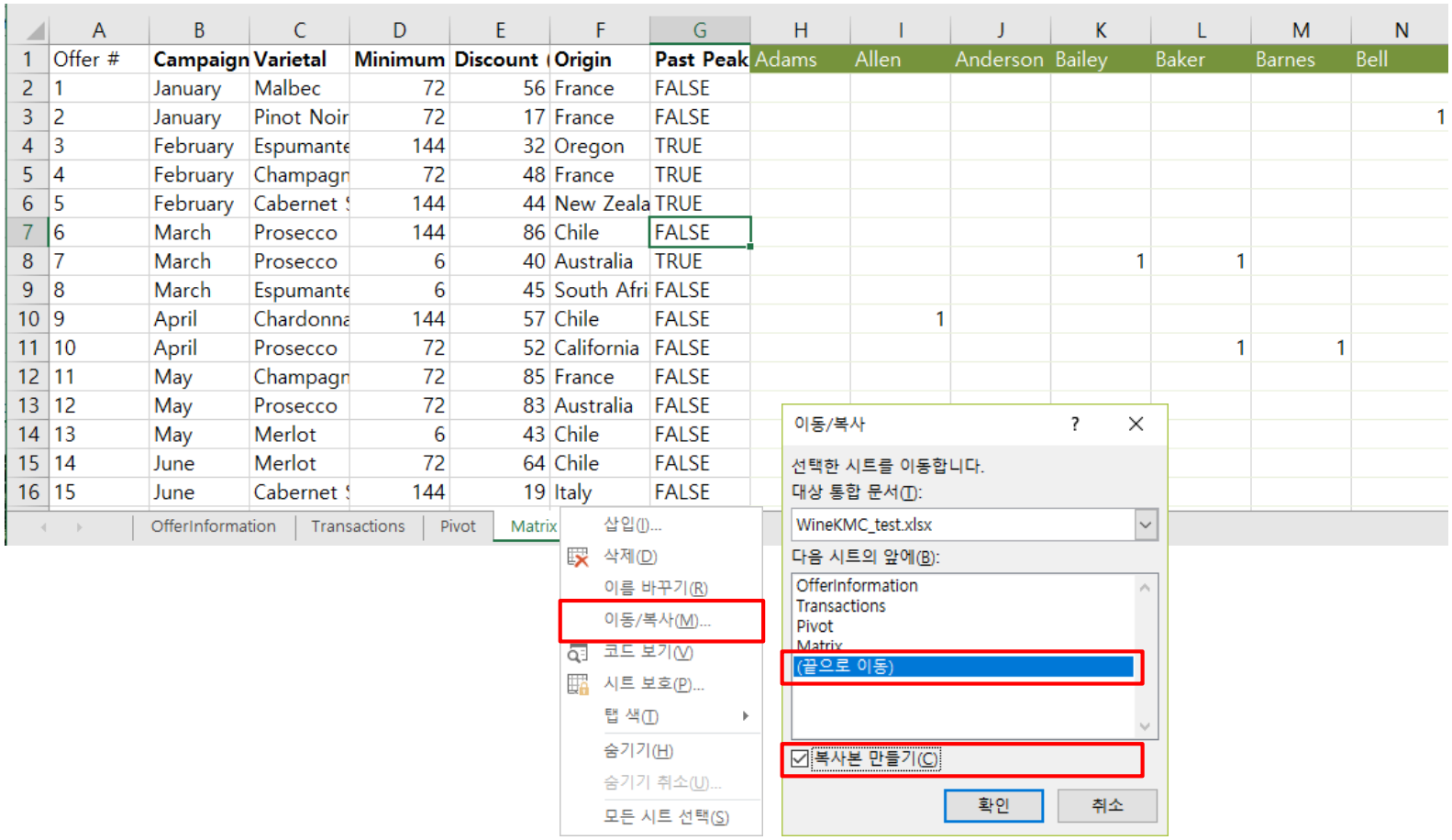
Pivot

	H	I	J	K	L	M	N
1	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell
2							
3							1
4							
5							
6							
7							
8				1	1		
9							
10		1					
11					1	1	
12							
13							
14							
15							
16							

OfferInformation | Transactions | Pivot | Matrix

4개의 클러스터

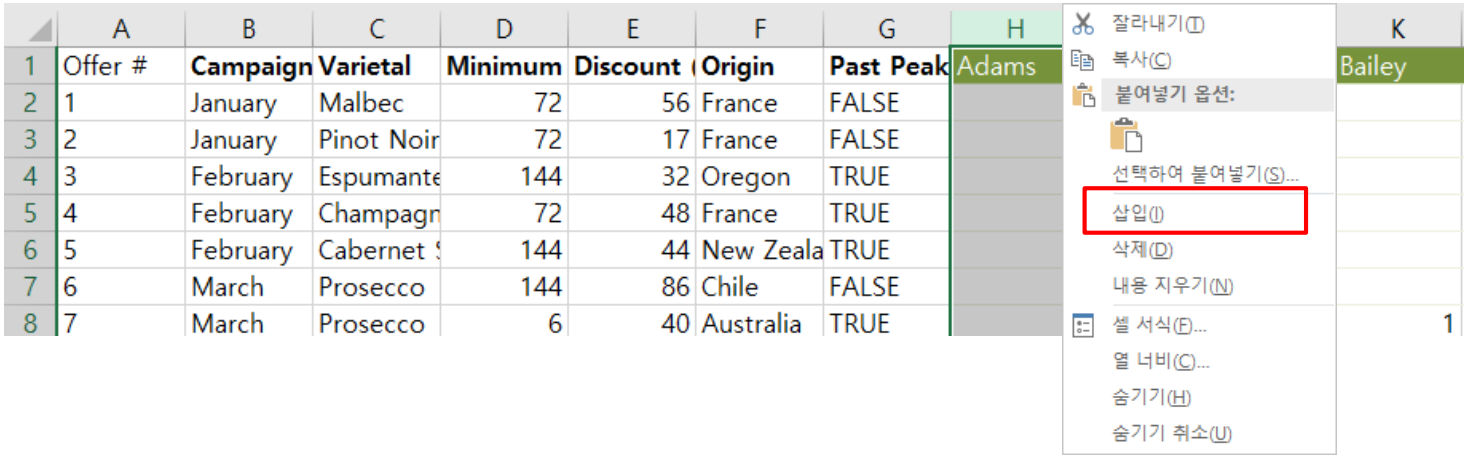
- 4개의 클러스터를 위한 데이터
  - Matrix데이터를 새로운 워크시트로 복사 (이름: 4MC)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Offer #	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell
2	1	January	Malbec	72	56	France	FALSE							
3	2	January	Pinot Noir	72	17	France	FALSE							1
4	3	February	Espumante	144	32	Oregon	TRUE							
5	4	February	Champagn	72	48	France	TRUE							
6	5	February	Cabernet S	144	44	New Zeala	TRUE							
7	6	March	Prosecco	144	86	Chile	FALSE							
8	7	March	Prosecco	6	40	Australia	TRUE				1	1		
9	8	March	Espumante	6	45	South Afri	FALSE							
10	9	April	Chardonna	144	57	Chile	FALSE		1					
11	10	April	Prosecco	72	52	California	FALSE					1	1	
12	11	May	Champagn	72	85	France	FALSE							
13	12	May	Prosecco	72	83	Australia	FALSE							
14	13	May	Merlot	6	43	Chile	FALSE							
15	14	June	Merlot	72	64	Chile	FALSE							
16	15	June	Cabernet S	144	19	Italy	FALSE							

4개의 클러스터

- 4개의 클러스터 Column 생성



	A	B	C	D	E	F	G	H	K
1	Offer #	Campaign	Varietal	Minimum	Discount (	Origin	Past Peak	Adams	Bailey
2	1	January	Malbec	72	56	France	FALSE		
3	2	January	Pinot Noir	72	17	France	FALSE		
4	3	February	Espumante	144	32	Oregon	TRUE		
5	4	February	Champagn	72	48	France	TRUE		
6	5	February	Cabernet S	144	44	New Zeala	TRUE		
7	6	March	Prosecco	144	86	Chile	FALSE		
8	7	March	Prosecco	6	40	Australia	TRUE		1

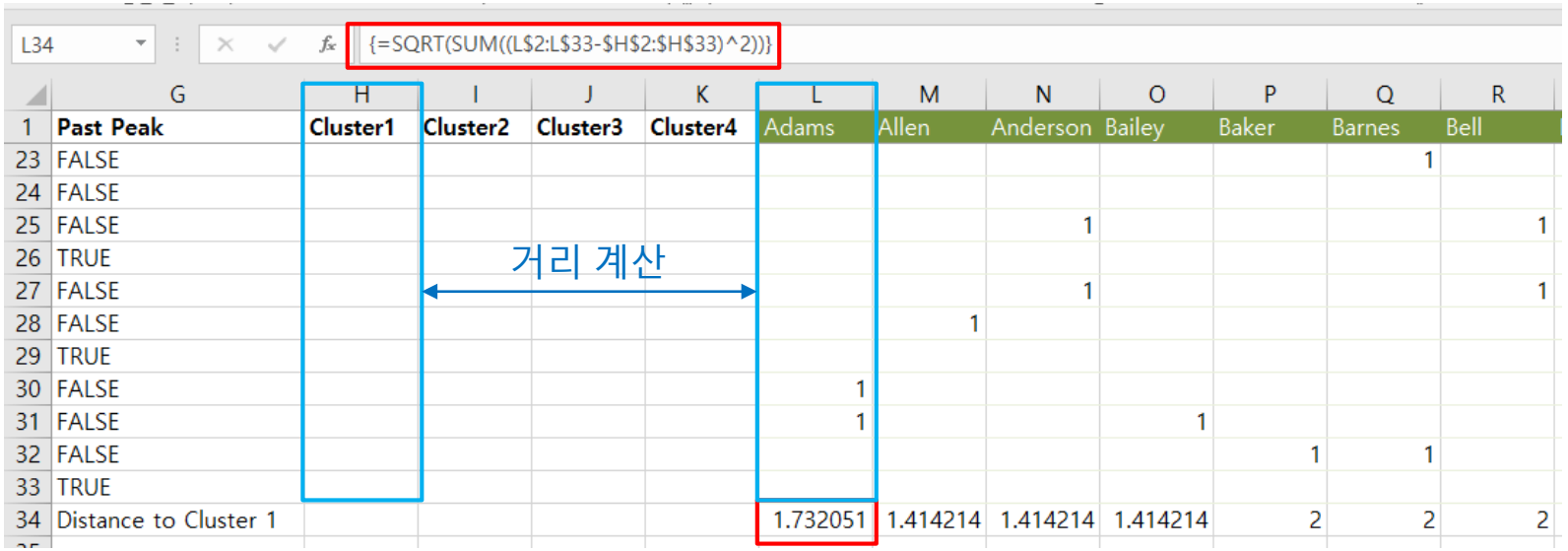
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Offer #	Campaign	Varietal	Minimum	Discount (	Origin	Past Peak	Cluster1	Cluster2	Cluster3	Cluster4	Adams	Allen	Anderson
2	1	January	Malbec	72	56	France	FALSE							
3	2	January	Pinot Noir	72	17	France	FALSE							
4	3	February	Espumante	144	32	Oregon	TRUE							
5	4	February	Champagn	72	48	France	TRUE							
6	5	February	Cabernet S	144	44	New Zeala	TRUE							
7	6	March	Prosecco	144	86	Chile	FALSE							
8	7	March	Prosecco	6	40	Australia	TRUE							
9	8	March	Espumante	6	45	South Afri	FALSE							
10	9	April	Chardonna	144	57	Chile	FALSE						1	
11	10	April	Prosecco	72	52	California	FALSE							

4개의 클러스터

- 소비자와 클러스터 중심점 사이의 거리 계산하기
  - Euclidean Distance: $(x_1, y_1), (x_2, y_2)$

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

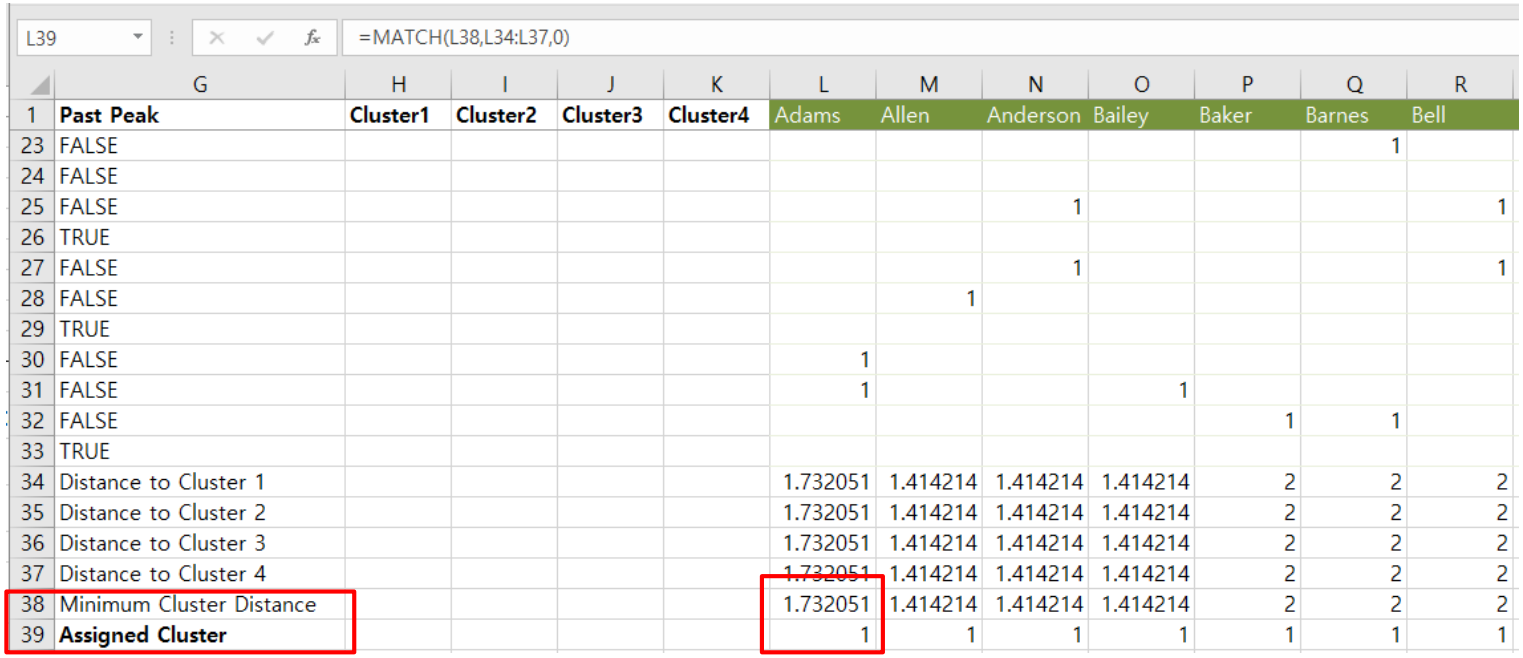
  - 클러스터의 초기 중심점은 (0,0,…,0)으로 설정하고 거리 계산
  - L34 {=SQRT(SUM((L$2:L$33-$H$2:$H$33)^2))} Ctrl + Shift + Enter



	G	H	I	J	K	L	M	N	O	P	Q	R
1	Past Peak	Cluster1	Cluster2	Cluster3	Cluster4	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell
23	FALSE										1	
24	FALSE											
25	FALSE							1				1
26	TRUE											
27	FALSE							1				1
28	FALSE						1					
29	TRUE											
30	FALSE					1						
31	FALSE					1			1			
32	FALSE									1	1	
33	TRUE											
34	Distance to Cluster 1					1.732051	1.414214	1.414214	1.414214	2	2	2

4개의 클러스터

- Cluster 2~4까지의 거리 계산
- 최소 거리 계산하기
  - L38 =MIN(L34:L37)
- 가장 가까운 클러스터 찾기
  - L39 =MATCH(L38,L34:L37,0)



L39 =MATCH(L38,L34:L37,0)

	G	H	I	J	K	L	M	N	O	P	Q	R
1	**Past Peak**	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell
23	FALSE										1	
24	FALSE											
25	FALSE							1				1
26	TRUE											
27	FALSE							1				1
28	FALSE						1					
29	TRUE											
30	FALSE					1						
31	FALSE					1			1			
32	FALSE									1	1	
33	TRUE											
34	Distance to Cluster 1					1.732051	1.414214	1.414214	1.414214	2	2	2
35	Distance to Cluster 2					1.732051	1.414214	1.414214	1.414214	2	2	2
36	Distance to Cluster 3					1.732051	1.414214	1.414214	1.414214	2	2	2
37	Distance to Cluster 4					1.732051	1.414214	1.414214	1.414214	2	2	2
38	Minimum Cluster Distance					1.732051	1.414214	1.414214	1.414214	2	2	2
39	**Assigned Cluster**					1	1	1	1	1	1	1

4개의 클러스터

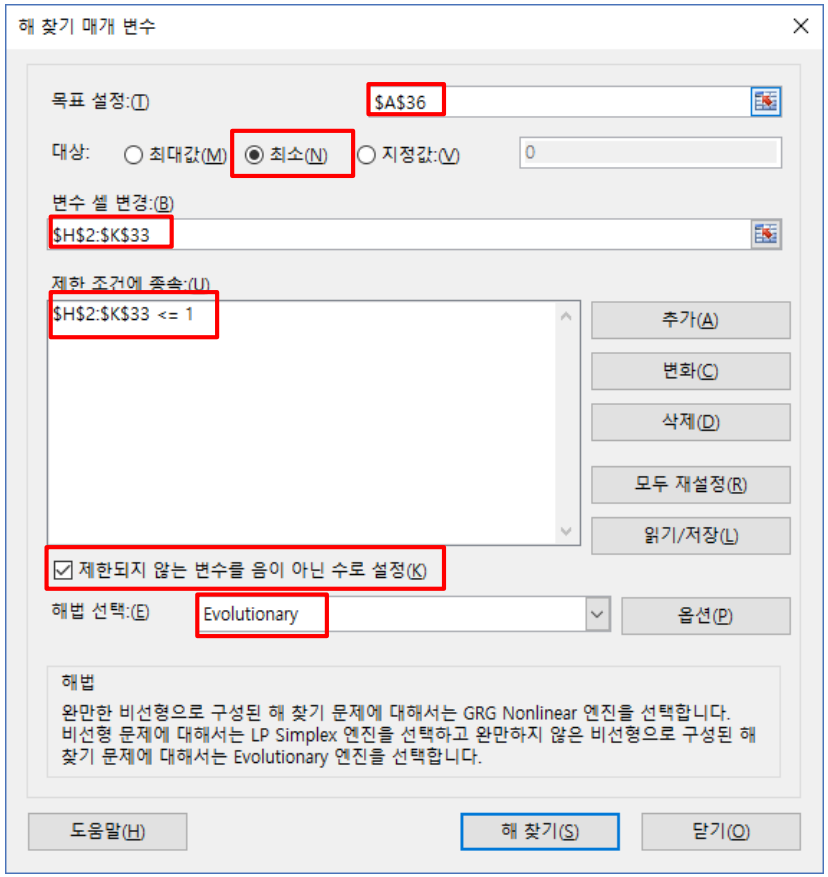
- 거리합 계산하기
- A36 =SUM(L38:DG38)

A36 =SUM(L38:DG38)

	A	B	C	D	E	F	G
1	Offer #	**Campaign**	**Varietal**	**Minimum Qty (kg)**	**Discount (%)**	**Origin**	**Past Peak**
20	19	July	Champagne	12	66	Germany	FALSE
21	20	August	Cabernet Sauvignon	72	82	Italy	FALSE
22	21	August	Champagne	12	50	California	FALSE
23	22	August	Champagne	72	63	France	FALSE
24	23	September	Chardonnay	144	39	South Africa	FALSE
25	24	September	Pinot Noir	6	34	Italy	FALSE
26	25	October	Cabernet Sauvignon	72	59	Oregon	TRUE
27	26	October	Pinot Noir	144	83	Australia	FALSE
28	27	October	Champagne	72	88	New Zealand	FALSE
29	28	November	Cabernet Sauvignon	12	56	France	TRUE
30	29	November	Pinot Grigio	6	87	France	FALSE
31	30	December	Malbec	6	54	France	FALSE
32	31	December	Champagne	72	89	France	FALSE
33	32	December	Cabernet Sauvignon	72	45	Germany	TRUE
34							Distance to Cluster 1
35	Total distance						Distance to Cluster 2
36	174.6366475						Distance to Cluster 3
37							Distance to Cluster 4
38							Min Distance
39							Assigned Cluster

4개의 클러스터

- 최적의 중심점 찾기
  - 목적 함수: 거리합의 최소화 (Non-linear function)
  - Min A36
  - 의사결정변수: 각 Cluster의 중심값
  - ($H$2:$K$33)
  - 제약식: 중심값은 0과 1사이의 값임
  - 0 ≤ $H$2:$K$33 ≤ 1



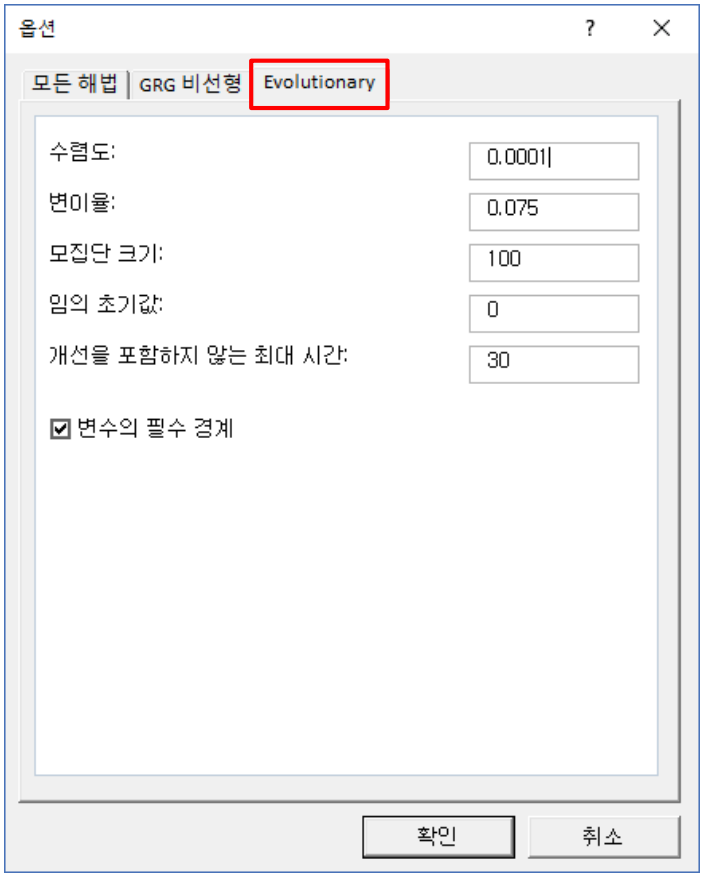
4개의 클러스터

- Evolutionary Algorithm의 옵션 설정
  - 해찾기 창에서 옵션선택

- 옵션 창에서 Evolutionary 선택
  - 수렴도: 종료조건

    (Population의 상위 99%의 목적함수 값의 차이)
  - 변이율: 돌연변이 비율
  - 모집단크기: 해집단의 개수
  - 임의초기값: 난수 생성 Seen number
  - 개선을 포함하지 않는 최대 시간: 종료조건

    (해개선이 이뤄지지 않는 시간 설정)



4개의 클러스터

- 최적의 중심점 찾기
  - 목적함수 값: 140.7
  - (주의) 비선형함수를 다루는 Evolutionary algorithm은 random number를 이용하기 때문에 결과는 매번 달라질 수 있다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Offer #	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	Cluster1	Cluster2	Cluster3	Cluster4	Adams	Allen	Anderson	Bailey
23	22	August	Champagn	72	63	France	FALSE	0.009	0.024	0.023	0.951				
24	23	Septembe	Chardonna	144	39	South Afri	FALSE	0.029	0.023	0.036	0.062				
25	24	Septembe	Pinot Noir	6	34	Italy	FALSE	0.941	0.043	0.017	0.035			1	
26	25	October	Cabernet S	72	59	Oregon	TRUE	0.025	0.034	0.083	0.114				
27	26	October	Pinot Noir	144	83	Australia	FALSE	0.690	0.030	0.090	0.130			1	
28	27	October	Champagn	72	88	New Zeala	FALSE	0.010	0.021	0.087	0.141		1		
29	28	Novembe	Cabernet S	12	56	France	TRUE	0.026	0.017	0.090	0.030				
30	29	Novembe	Pinot Grig	6	87	France	FALSE	0.012	0.619	0.043	0.038	1			
31	30	Decembe	Malbec	6	54	France	FALSE	0.020	0.729	0.079	0.136	1			1
32	31	Decembe	Champagn	72	89	France	FALSE	0.023	0.027	0.211	0.259				
33	32	Decembe	Cabernet S	72	45	Germany	TRUE	0.093	0.013	0.053	0.125				
34							Distance to Cluster 1					2.165886	1.939157	0.739977	1.924189
35	Total Distance						Distance to Cluster 2					1.043837	1.885796	1.865095	1.041527
36	**140.7**						Distance to Cluster 3					1.690595	1.339467	1.42766	1.359019
37							Distance to Cluster 4					2.012004	1.730512	1.781281	1.74921
38							Minimum Cluster Distance					1.043837	1.339467	0.739977	1.041527
39							**Assigned Cluster**					2	3	1	2

4개의 클러스터

- Top Deal 살펴보기 (4MC-TopDealsByCluster 생성)
  - OfferInformation 워크시트 복사, H1~K1 작성

  - 각 클러스터에서 판매제안을 수락한 고객 수 확인하기
  - H2 =SUMIF('4MC'!$L$39:$DG$39,'4MC-TopDealsByCluster'!H$1,'4MC'!$L2:$DG2)

    할당된 클러스터의 결과를 대상으로 → 영역불변 '!$L$39:$DG$39
    할당된 클러스터의 값이 1이면, → Column에 따라 가변 H$1
    해당 Offer를 수락했는지 (0 or 1) 합하기 → Row에 따라 가변 $L2:$DG2

  - I2 =SUMIF('4MC'!$L$39:$DG$39,'4MC-TopDealsByCluster'!I$1,'4MC'!$L2:$DG2)
  - H3=SUMIF('4MC'!$L$39:$DG$39,'4MC-TopDealsByCluster'!H$1,'4MC'!$L3:$DG3)

H2 =SUMIF('4MC'!$L$39:$DG$39,'4MC-TopDealsByCluster'!H$1,'4MC'!$L2:$DG2)

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	1	2	3	4
2	1	January	Malbec	72	56	France	FALSE	0	0	4	6
3	2	January	Pinot Noir	72	17	France	FALSE	4	0	4	2
4	3	February	Espumante	144	32	Oregon	TRUE	0	0	2	4
5	4	February	Champagne	72	48	France	TRUE	0	0	7	5
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0	0	2	2
7	6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
8	7	March	Prosecco	6	40	Australia	TRUE	0	12	4	3
9	8	March	Espumante	6	45	South Africa	FALSE	0	11	6	3
10	9	April	Chardonnay	144	57	Chile	FALSE	0	0	7	3
11	10	April	Prosecco	72	52	California	FALSE	0	0	5	2

4개의 클러스터

- Top Deal 살펴보기 (4MC-TopDealsByCluster)
  - H~K 조건부 서식
  - 필터링
  - H를 내림차순으로 정렬

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4
2	24	September	Pinot Noir	6	34	Italy	FALSE	12	0	0	0
3	26	October	Pinot Noir	144	83	Australia	FALSE	8	0	5	2
4	17	July	Pinot Noir	12	47	Germany	FALSE	7	0	0	0
5	2	January	Pinot Noir	72	17	France	FALSE	4	0	4	2
6	1	January	Malbec	72	56	France	FALSE	0	0	4	6
7	3	February	Espumante	144	32	Oregon	TRUE	0	0	2	4
8	4	February	Champagne	72	48	France	TRUE	0	0	7	5
9	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0	0	2	2
10	6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
11	7	March	Prosecco	6	40	Australia	TRUE	0	12	4	3
12	8	March	Espumante	6	45	South Africa	FALSE	0	11	6	3
13	9	April	Chardonnay	144	57	Chile	FALSE	0	0	7	3
14	10	April	Prosecco	72	52	California	FALSE	0	0	5	2
15	11	May	Champagne	72	85	France	FALSE	0	0	7	6
16	12	May	Prosecco	72	83	Australia	FALSE	0	0	3	2
17	13	May	Merlot	6	43	Chile	FALSE	0	6	0	0

4개의 클러스터

- Top Deal 살펴보기 (4MC-TopDealsByCluster)
  - I를 내림차순으로 정렬

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4
2	30	December	Malbec	6	54	France	FALSE	0	16	2	4
3	29	November	Pinot Grigio	6	87	France	FALSE	0	15	2	0
4	7	March	Prosecco	6	40	Australia	TRUE	0	12	4	3
5	8	March	Espumante	6	45	South Africa	FALSE	0	11	6	3
6	18	July	Espumante	6	50	Oregon	FALSE	0	11	2	1
7	13	May	Merlot	6	43	Chile	FALSE	0	6	0	0
8	24	September	Pinot Noir	6	34	Italy	FALSE	12	0	0	0
9	26	October	Pinot Noir	144	83	Australia	FALSE	8	0	5	2
10	17	July	Pinot Noir	12	47	Germany	FALSE	7	0	0	0
11	2	January	Pinot Noir	72	17	France	FALSE	4	0	4	2
12	1	January	Malbec	72	56	France	FALSE	0	0	4	6
13	3	February	Espumante	144	32	Oregon	TRUE	0	0	2	4
14	4	February	Champagne	72	48	France	TRUE	0	0	7	5
15	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0	0	2	2
16	6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
17	9	April	Chardonnay	144	57	Chile	FALSE	0	0	7	3

4개의 클러스터

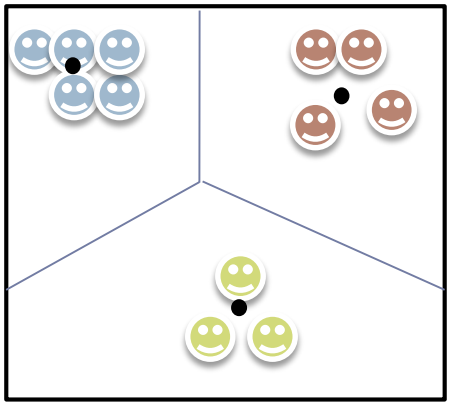
- Top Deal 살펴보기 (4MC-TopDealsByCluster)
  - J,K 를 내림차순으로 정렬

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4
2	31	December	Champagne	72	89	France	FALSE	0	0	10	7
3	4	February	Champagne	72	48	France	TRUE	0	0	7	5
4	9	April	Chardonnay	144	57	Chile	FALSE	0	0	7	3
5	11	May	Champagne	72	85	France	FALSE	0	0	7	6
6	8	March	Espumante	6	45	South Africa	FALSE	0	11	6	3
7	27	October	Champagne	72	88	New Zealand	FALSE	0	0	6	3
8	26	October	Pinot Noir	144	83	Australia	FALSE	8	0	5	2
9	6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
10	10	April	Prosecco	72	52	California	FALSE	0	0	5	2
11	14	June	Merlot	72	64	Chile	FALSE	0	0	5	4
12	16	June	Merlot	72	88	California	FALSE	0	0	5	0

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4
2	22	August	Champagne	72	63	France	FALSE	0	0	0	21
3	31	December	Champagne	72	89	France	FALSE	0	0	10	7
4	6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
5	11	May	Champagne	72	85	France	FALSE	0	0	7	6
6	1	January	Malbec	72	56	France	FALSE	0	0	4	6
7	4	February	Champagne	72	48	France	TRUE	0	0	7	5
8	14	June	Merlot	72	64	Chile	FALSE	0	0	5	4
9	30	December	Malbec	6	54	France	FALSE	0	16	2	4
10	3	February	Espumante	144	32	Oregon	TRUE	0	0	2	4
11	15	June	Cabernet Sauvignon	144	19	Italy	FALSE	0	0	2	4
12	9	April	Chardonnay	144	57	Chile	FALSE	0	0	7	3

평가하기

- 실루엣 (Silhouette) 기법
  - $a(i)$: i 개체가 속한 군집에 있는 요소들과의 평균 거리
  - $b(i)$: i 개체가 속하지 않은 다른 군집과의 평균 거리 중 가장 작은 거리
  - $s(i) = \frac{b(i)-a(i)}{\max\{a(i).b(i)\}}$
  - $-1 \leq s(\mathrm{i}) \leq 1$



평가하기

- 거리 계산하기 (Distances 워크시트: 소비자들 간의 거리)
  - 데이터 개체간의 거리를 계산할 Distance matrix의 생성
  - $\begin{bmatrix} x_1 \\ x_2 \\ \dots \end{bmatrix}$ 와 $\begin{bmatrix} y_1 \\ y_2 \\ \dots \end{bmatrix}$의 거리 $=\sqrt{\sum_i (x_i - y_i)^2}$

  - C3 {=SQRT(SUM((OFFSET(Matrix!$H$2:$H$33,0,Distances!C$1)
    -OFFSET(Matrix!$H$2:$H$33,0,Distances!$A3))^2))}

C$1 (Adams)의 벡터값 - $A3 (Adams)의 벡터값

C3 {=SQRT(SUM((OFFSET(Matrix!$H$2:$H$33,0,Distances!C$1)-OFFSET(Matrix!$H$2:$H$33,0,Distances!$A3))^2))}

	A	B	C	D	E	F	G	H	I	J	K	L
1			0	1	2	3	4	5	6	7	8	9
2			Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Brooks	Brown
3	0	Adams	0.000	2.236	2.236	1.732	2.646	2.646	2.646	1.732	2.646	1.414
4	1	Allen	2.236	0.000	2.000	2.000	2.449	2.449	2.449	2.000	2.449	2.236
5	2	Anderson	2.236	2.000	0.000	2.000	2.449	2.449	1.414	2.000	2.449	2.236
6	3	Bailey	1.732	2.000	2.000	0.000	2.000	2.449	2.449	2.000	2.449	1.000
7	4	Baker	2.646	2.449	2.449	2.000	0.000	2.000	2.828	2.449	2.828	2.236
8	5	Barnes	2.646	2.449	2.449	2.449	2.000	0.000	2.828	2.449	2.449	2.646
9	6	Bell	2.646	2.449	1.414	2.449	2.828	2.828	0.000	2.449	2.828	2.646
10	7	Bennett	1.732	2.000	2.000	2.000	2.449	2.449	2.449	0.000	2.000	1.732

평가하기

- 실루엣 계산하기 (4MC Silhouette 워크시트)
  - 4MC에서 고객이름(L1~DG1), 속한 군집 정보(L39~DG39) 복사
    - 선택하여 붙여넣기, 값 & 행열 바꿈
  - 고객별로 각 군집까지의 평균 거리 계산
  - C2 =AVERAGEIF('4MC'!$L$39:$DG$39,1,Distances!$C3:$CX3)

    4MC의 군집 결과가 1인 경우, Adams로 부터의 거리 데이터의 평균을 구함

C2 =AVERAGEIF('4MC'!$L$39:$DG$39,1,Distances!$C3:$CX3)

	A	B	C	D	E	F
1	Name	Community	Distance from people in 1	Distance from people in 2	Distance from people in 3	Distance from people in 4
2	Adams	2	2.358	1.495	2.318	2.688
3	Allen	3	2.134	2.215	1.980	2.476
4	Anderson	1	0.957	2.215	2.097	2.558
5	Bailey	2	2.134	1.554	2.080	2.462
6	Baker	3	2.562	2.429	2.346	2.703
7	Barnes	4	2.562	2.631	2.423	2.345
8	Bell	1	1.075	2.631	2.495	2.897
9	Bennett	2	2.134	1.575	2.047	2.534
10	Brooks	4	2.562	2.447	2.438	2.297

평가하기

- 실루엣 계산하기 (4MC Silhouette)
  - Closest: 가장 짧은 평균 거리
  - G2 =MIN(C2:F2)
  - Second Closest: 두번째로 짧은 평균 거리
  - H2 =SMALL(C2:F2,2)
  - My Cluster: 본인이 속한 클러스터내의 평균거리
  - I2 =INDEX(C2:F2,B2)
  - Neighboring Cluster: 가장 가까운 이웃 클러스터까지 평균 거리
  - J2 =IF(I2=G2,H2,G2)
  - Silhouette Values: 실루엣 값
  - K2 =(J2-I2)/MAX(J2,I2)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i).b(i)\}}$$

	A	B	C	D	E	F
1	Name	Community	Distance from people in 1	Distance from people in 2	Distance from people in 3	Distance from people in 4
2	Adams	2	2.358	1.495	2.318	2.688
3	Allen	3	2.134	2.215	1.980	2.476
4	Anderson	1	0.957	2.215	2.097	2.558
5	Bailey	2	2.134	1.554	2.080	2.462

G	H	I	J	K	
Closest	Second Closest	My Cluster	Neighboring Cluster	Silhouette Values	
1.495	2.318	1.495	2.318	0.355	
1.980	2.134	1.980	2.134	0.072	
0.957	2.097	0.957	2.097	0.544	
1.554	2.080	1.554	2.080	0.253	
2.346	2.429	2.346	2.429	0.034	
2.345	2.423	2.345	2.423	0.032	
1.075	2.495	1.075	2.495	0.569	Not Bad
1.575	2.047	1.575	2.047	0.231	
2.297	2.438	2.297	2.438	0.058	
1.455	2.294	1.455	2.294	0.365	
2.440	2.565	2.440	2.565	0.049	
1.169	2.279	1.169	2.279	0.487	
1.628	2.506	1.628	2.506	0.351	
2.284	2.562	2.284	2.562	0.109	
1.882	2.038	2.038	1.882	-0.077	Poor

평가하기

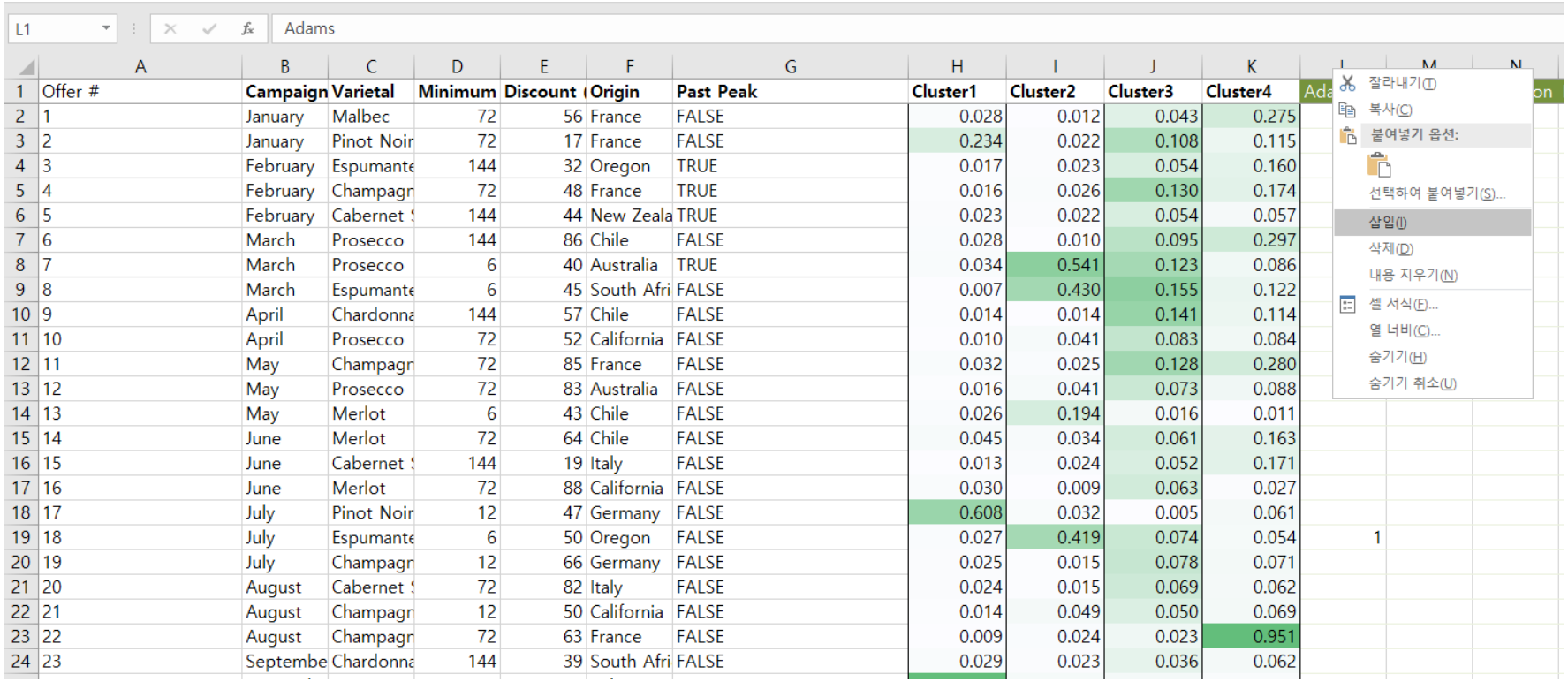
- 실루엣 계산하기 (4MC Silhouette)
  - 최종 실루엣 값 : 각 실루엣 값의 평균
  - M2 =AVERAGE(K2:K101)

M2 =AVERAGE(K2:K101)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Community	Distance from people in 1	Distance from people in 2	Distance from people in 3	Distance from people in 4	Closest	Second Closest	My Cluster	Neighboring Cluster	Silhouette Values		Silhouette
2	**Adams**	2	2.358	1.495	2.318	2.688	1.495	2.318	1.495	2.318	0.355		**0.1492**
3	**Allen**	3	2.134	2.215	1.980	2.476	1.980	2.134	1.980	2.134	0.072		
4	**Anderson**	1	0.957	2.215	2.097	2.558	0.957	2.097	0.957	2.097	0.544		
5	**Bailey**	2	2.134	1.554	2.080	2.462	1.554	2.080	1.554	2.080	0.253		
6	**Baker**	3	2.562	2.429	2.346	2.703	2.346	2.429	2.346	2.429	0.034		
7	**Barnes**	4	2.562	2.631	2.423	2.345	2.345	2.423	2.345	2.423	0.032		
8	**Bell**	1	1.075	2.631	2.495	2.897	1.075	2.495	1.075	2.495	0.569		
9	**Bennett**	2	2.134	1.575	2.047	2.534	1.575	2.047	1.575	2.047	0.231		
10	**Brooks**	4	2.562	2.447	2.438	2.297	2.297	2.438	2.297	2.438	0.058		
11	**Brown**	2	2.358	1.455	2.294	2.660	1.455	2.294	1.455	2.294	0.365		
12	**Butler**	4	2.750	2.565	2.624	2.440	2.440	2.565	2.440	2.565	0.049		
13	**Campbell**	1	1.169	2.432	2.279	2.717	1.169	2.279	1.169	2.279	0.487		
14	**Carter**	2	2.562	1.628	2.506	2.844	1.628	2.506	1.628	2.506	0.351		
15	**Clark**	3	2.562	2.631	2.284	2.627	2.284	2.562	2.284	2.562	0.109		
16	**Collins**	3	2.134	1.882	2.038	2.392	1.882	2.038	2.038	1.882	-0.077		
17	**Cook**	1	0.957	2.215	2.097	2.558	0.957	2.097	0.957	2.097	0.544		

5개의 클러스터

- 5개의 클러스터를 위한 데이터 (5MC 생성)
  - 4MC의 데이터를 새로운 tap으로 복사
  - Cluster5 column 삽입



L1 | Adams

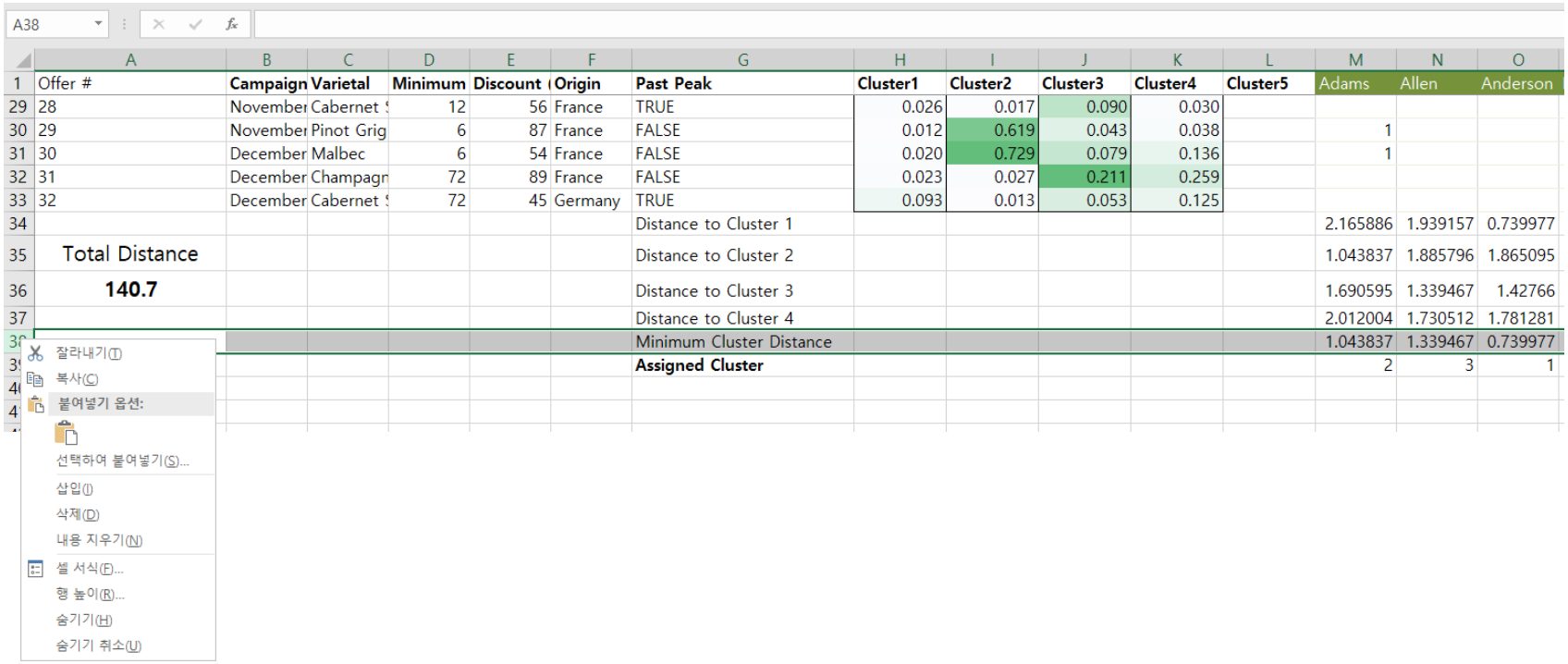
	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum	Discount (	Origin	Past Peak	Cluster1	Cluster2	Cluster3	Cluster4
2	1	January	Malbec	72	56	France	FALSE	0.028	0.012	0.043	0.275
3	2	January	Pinot Noir	72	17	France	FALSE	0.234	0.022	0.108	0.115
4	3	February	Espumante	144	32	Oregon	TRUE	0.017	0.023	0.054	0.160
5	4	February	Champagn	72	48	France	TRUE	0.016	0.026	0.130	0.174
6	5	February	Cabernet S	144	44	New Zeala	TRUE	0.023	0.022	0.054	0.057
7	6	March	Prosecco	144	86	Chile	FALSE	0.028	0.010	0.095	0.297
8	7	March	Prosecco	6	40	Australia	TRUE	0.034	0.541	0.123	0.086
9	8	March	Espumante	6	45	South Afri	FALSE	0.007	0.430	0.155	0.122
10	9	April	Chardonna	144	57	Chile	FALSE	0.014	0.014	0.141	0.114
11	10	April	Prosecco	72	52	California	FALSE	0.010	0.041	0.083	0.084
12	11	May	Champagn	72	85	France	FALSE	0.032	0.025	0.128	0.280
13	12	May	Prosecco	72	83	Australia	FALSE	0.016	0.041	0.073	0.088
14	13	May	Merlot	6	43	Chile	FALSE	0.026	0.194	0.016	0.011
15	14	June	Merlot	72	64	Chile	FALSE	0.045	0.034	0.061	0.163
16	15	June	Cabernet S	144	19	Italy	FALSE	0.013	0.024	0.052	0.171
17	16	June	Merlot	72	88	California	FALSE	0.030	0.009	0.063	0.027
18	17	July	Pinot Noir	12	47	Germany	FALSE	0.608	0.032	0.005	0.061
19	18	July	Espumante	6	50	Oregon	FALSE	0.027	0.419	0.074	0.054
20	19	July	Champagn	12	66	Germany	FALSE	0.025	0.015	0.078	0.071
21	20	August	Cabernet S	72	82	Italy	FALSE	0.024	0.015	0.069	0.062
22	21	August	Champagn	12	50	California	FALSE	0.014	0.049	0.050	0.069
23	22	August	Champagn	72	63	France	FALSE	0.009	0.024	0.023	0.951
24	23	Septembe	Chardonna	144	39	South Afri	FALSE	0.029	0.023	0.036	0.062

잘라내기(T)
복사(C)
붙여넣기 옵션:
선택하여 붙여넣기(S)...
삽입(I)
삭제(D)
내용 지우기(N)
셀 서식(F)...
열 너비(C)...
숨기기(H)
숨기기 취소(U)

1

5개의 클러스터

- 5개의 클러스터를 위한 데이터 (5MC)
  - Cluster5 column 삽입
  - Distance to Cluster 5 row 삽입



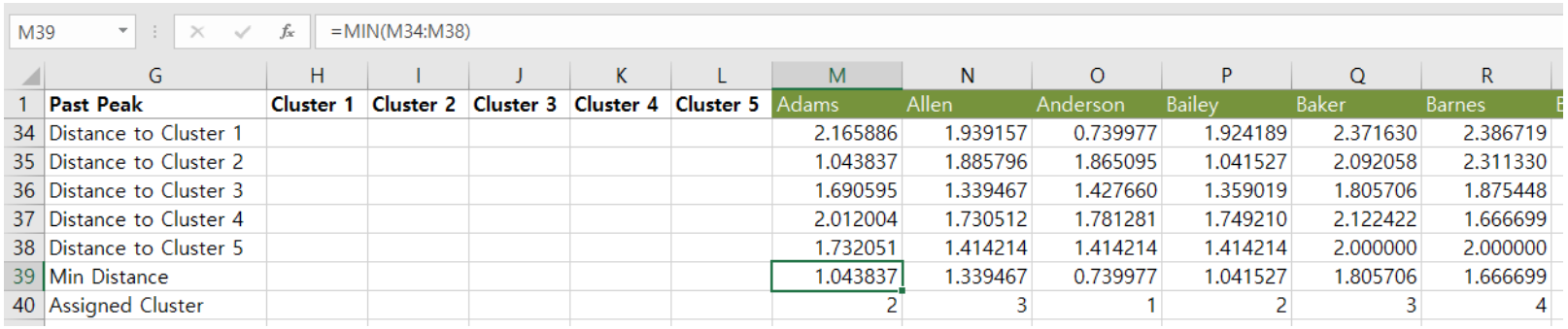
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Offer #	Campaign	Varietal	Minimum	Discount (	Origin	Past Peak	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Adams	Allen	Anderson
29	28	November	Cabernet S	12	56	France	TRUE	0.026	0.017	0.090	0.030				
30	29	November	Pinot Grig	6	87	France	FALSE	0.012	0.619	0.043	0.038		1		
31	30	December	Malbec	6	54	France	FALSE	0.020	0.729	0.079	0.136		1		
32	31	December	Champagn	72	89	France	FALSE	0.023	0.027	0.211	0.259				
33	32	December	Cabernet S	72	45	Germany	TRUE	0.093	0.013	0.053	0.125				
34							Distance to Cluster 1						2.165886	1.939157	0.739977
35	Total Distance						Distance to Cluster 2						1.043837	1.885796	1.865095
36	140.7						Distance to Cluster 3						1.690595	1.339467	1.42766
37							Distance to Cluster 4						2.012004	1.730512	1.781281
38							Minimum Cluster Distance						1.043837	1.339467	0.739977
39							Assigned Cluster						2	3	1

5개의 클러스터

- 5개의 클러스터를 위한 데이터 (5MC)
  - Cluster 5까지의 거리 계산
  - M38 {=SQRT(SUM((M$2:M$33-$L$2:$L$33)^2))}

  - Minimum Cluster Distance에 Cluster5를 포함하여 계산
  - M39 =MIN(M34:M38)

  - 할당되는 Cluster 계산
  - M40 =MATCH(M39,M34:M38,0)



M39 =MIN(M34:M38)

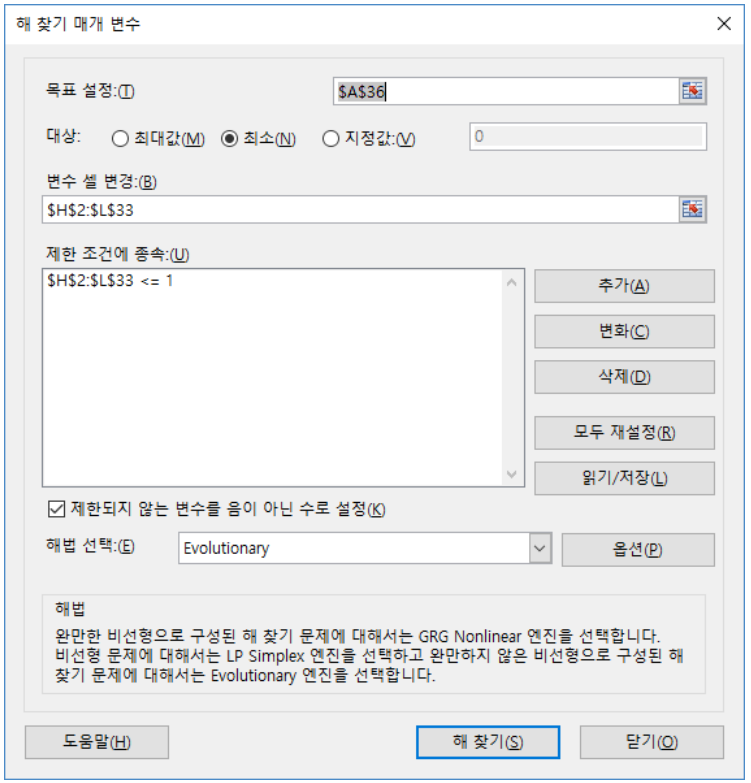
	G	H	I	J	K	L	M	N	O	P	Q	R
1	**Past Peak**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	Adams	Allen	Anderson	Bailey	Baker	Barnes
34	Distance to Cluster 1						2.165886	1.939157	0.739977	1.924189	2.371630	2.386719
35	Distance to Cluster 2						1.043837	1.885796	1.865095	1.041527	2.092058	2.311330
36	Distance to Cluster 3						1.690595	1.339467	1.427660	1.359019	1.805706	1.875448
37	Distance to Cluster 4						2.012004	1.730512	1.781281	1.749210	2.122422	1.666699
38	Distance to Cluster 5						1.732051	1.414214	1.414214	1.414214	2.000000	2.000000
39	Min Distance						1.043837	1.339467	0.739977	1.041527	1.805706	1.666699
40	Assigned Cluster						2	3	1	2	3	4

5개의 클러스터

- 5개의 클러스터를 위한 최적화
  - 목적 함수: 거리합의 최소화 (Non-linear function)
  - =SUM(L38:DG38)
  - Min A36

  - 의사결정변수: 각 Cluster의 중심값
  - ($H$2:$L$33)

  - 제약식: 중심값은 0과 1사이의 값임
  - 0 ≤ $H$2:$L$33 ≤ 1



5개의 클러스터

- 최적의 중심점 찾기
  - 목적함수 값: 135.1
  - (주의) 비선형함수를 다루는 Evolutionary algorithm은 random number를 이용하기 때문에 결과는 매번 달라질 수 있다.

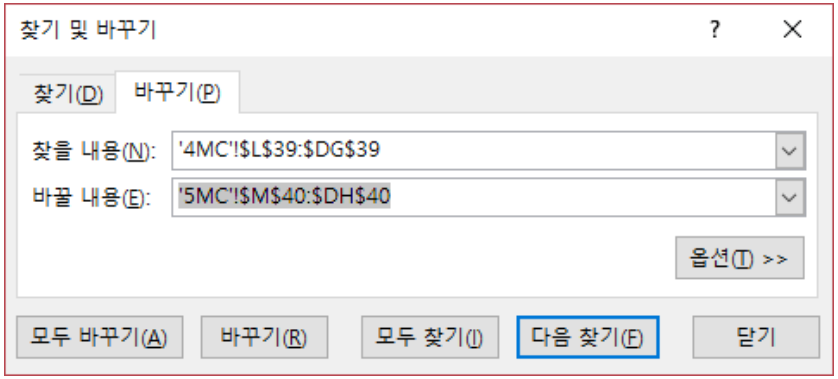
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Offer #	Campaign	Varietal	Minimum	Discount (	Origin	Past Peak	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Adams	Allen	Anderson	Bailey	Baker
23	22	August	Champagn	72	63	France	FALSE	0.007	0.009	0.004	1.000	0.004					
24	23	Septembe	Chardonna	144	39	South Afri	FALSE	0.011	0.007	0.008	0.077	0.072					
25	24	Septembe	Pinot Noir	6	34	Italy	FALSE	1.000	0.011	0.004	0.005	0.009			1		
26	25	October	Cabernet S	72	59	Oregon	TRUE	0.011	0.010	0.008	0.099	0.082					
27	26	October	Pinot Noir	144	83	Australia	FALSE	0.719	0.008	0.000	0.033	0.147			1		
28	27	October	Champagn	72	88	New Zeala	FALSE	0.010	0.011	0.021	0.152	0.112		1			
29	28	Novembe	Cabernet S	12	56	France	TRUE	0.010	0.011	0.000	0.068	0.100					
30	29	Novembe	Pinot Grig	6	87	France	FALSE	0.005	0.679	0.044	0.008	0.048	1				
31	30	Decembe	Malbec	6	54	France	FALSE	0.006	0.769	0.021	0.182	0.051	1			1	
32	31	Decembe	Champagn	72	89	France	FALSE	0.008	0.006	0.013	0.310	0.239					1
33	32	Decembe	Cabernet S	72	45	Germany	TRUE	0.000	0.003	0.004	0.039	0.065					
34							Distance to Cluster 1						2.218456	1.981515	0.724946	1.982409	2.426226
35	Total Distance						Distance to Cluster 2						0.972701	1.937223	1.938234	1.019982	2.130472
36	**135.1**						Distance to Cluster 3						1.954061	1.721054	1.731005	1.721185	2.225835
37							Distance to Cluster 4						2.014427	1.756961	1.863776	1.721341	2.075379
38							Distance to Cluster 5						1.753609	1.297274	1.419349	1.413843	1.829649
39							Minimum Cluster Distance						0.972701	1.297274	0.724946	1.019982	1.829649
40							**Assigned Cluster**						2	5	1	2	5
41																	

5개의 클러스터

- Top Deal 살펴보기 (5MC-TopDealsByCluster 생성)
  - 4MC-TopDealsByCluster 복사
  - 필터해제
  - Offer#의 오름차순 정렬
  - L column을 Cluster5로 지정
  - K2~K33을 L2~L33으로 복사
  - 할당된 클러스터 결과 참조 영역을 변경

    → '4MC'!$L$39:$DG$39를 '5MC'!$M$40:$DH$40 으로 변환

    → '4MC_PPT'!$L2:$DG2를 '5MC_PPT'!$M2:$DH2 으로 변환



5개의 클러스터

- Top Deal 살펴보기 (5MC-TopDealsByCluster)
  - 클러스터 1 내림차순 정렬

	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4	5
2	24	September	Pinot Noir	6	34	Italy	FALSE	12	0	0	0	0
3	26	October	Pinot Noir	144	83	Australia	FALSE	8	0	0	1	6
4	17	July	Pinot Noir	12	47	Germany	FALSE	7	0	0	0	0
5	2	January	Pinot Noir	72	17	France	FALSE	4	0	0	2	4
6	1	January	Malbec	72	56	France	FALSE	0	0	0	5	5
7	3	February	Espumante	144	32	Oregon	TRUE	0	0	1	4	1

  - 클러스터 2 내림차순 정렬

	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4	5
2	30	December	Malbec	6	54	France	FALSE	0	15	1	4	2
3	29	November	Pinot Grigio	6	87	France	FALSE	0	13	2	0	2
4	7	March	Prosecco	6	40	Australia	TRUE	0	12	0	3	4
5	18	July	Espumante	6	50	Oregon	FALSE	0	10	2	1	1
6	8	March	Espumante	6	45	South Africa	FALSE	0	7	10	3	0
7	13	May	Merlot	6	43	Chile	FALSE	0	5	1	0	0
8	24	September	Pinot Noir	6	34	Italy	FALSE	12	0	0	0	0
9	26	October	Pinot Noir	144	83	Australia	FALSE	8	0	0	1	6

5개의 클러스터

- Top Deal 살펴보기 (5MC-TopDealsByCluster)
  - 클러스터 3 내림차순 정렬: South African Espumante?

	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4	5
2	8	March	Espumante	6	45	South Africa	FALSE	0	7	10	3	0
3	29	November	Pinot Grigio	6	87	France	FALSE	0	13	2	0	2
4	18	July	Espumante	6	50	Oregon	FALSE	0	10	2	1	1
5	30	December	Malbec	6	54	France	FALSE	0	15	1	4	2

  - 클러스터 4 내림차순 정렬: High volume? France?

	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4	5
2	22	August	Champagne	72	63	France	FALSE	0	0	0	21	0
3	31	December	Champagne	72	89	France	FALSE	0	0	1	7	9
4	6	March	Prosecco	144	86	Chile	FALSE	0	0	1	6	5
5	1	January	Malbec	72	56	France	FALSE	0	0	0	5	5
6	11	May	Champagne	72	85	France	FALSE	0	0	0	5	8

  - 클러스터 5 내림차순 정렬: High volume? High discounts?

	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Campaign	Varietal	Minimum Qty (k	Discount (%	Origin	Past Peak	1	2	3	4	5
2	31	December	Champagne	72	89	France	FALSE	0	0	1	7	9
3	11	May	Champagne	72	85	France	FALSE	0	0	0	5	8
4	9	April	Chardonnay	144	57	Chile	FALSE	0	0	0	2	8
5	4	February	Champagne	72	48	France	TRUE	0	0	1	4	7
6	26	October	Pinot Noir	144	83	Australia	FALSE	8	0	0	1	6
7	6	March	Prosecco	144	86	Chile	FALSE	0	0	1	6	5
8	1	January	Malbec	72	56	France	FALSE	0	0	0	5	5
9	14	June	Merlot	72	64	Chile	FALSE	0	0	0	4	5
10	27	October	Champagne	72	88	New Zealand	FALSE	0	0	1	3	5

5개의 클러스터

- 실루엣 계산하기 (5MC Silhouette 생성)
  - 4MC Silhouette 복사
  - Column B(Community)값을 5MC에서 복사

34	Distance to Cluster 1	2.218456	1.981515	0.724946	1.982409	2.426226	2.42685	0.878453	1.982454	2.429025
35	Distance to Cluster 2	0.972701	1.937223	1.938234	1.019982	2.130472	2.393979	2.397557	1.34015	2.258149
36	Distance to Cluster 3	1.954061	1.721054	1.731005	1.721185	2.225835	2.221499	2.231645	0.957243	1.723415
37	Distance to Cluster 4	2.014427	1.756961	1.863776	1.721341	2.075379	1.61066	2.303213	1.811456	1.546371
38	Distance to Cluster 5	1.753609	1.297274	1.419349	1.413843	1.829649	1.890532	1.948231	1.492662	1.984186
39	Minimum Cluster Distance	0.972701	1.297274	0.724946	1.019982	1.829649	1.61066	0.878453	0.957243	1.546371
40	**Assigned Cluster**	2	5	1	2	5	4	1	3	4

  - Distance from people in 5 column 삽입 (Column G)
  - F2를 G2에 복사 후, =AVERAGEIF('4MC'!$L$39:$DG$39,5,Distances!$C3:$CX3)로 변경
  - '4MC'!$L$39:$DG$39를 '5MC'!$M$40:$DH$40 으로 변환

  - Closest: H2 =MIN(C2:G2)
  - Second Closest: I2 =SMALL(C2:G2,2)
  - My Cluster: J2 =INDEX(C2:G2,B2)

5개의 클러스터

- 실루엣 계산하기 (5MC Silhouette)
  - 4MC보다 더 잘 구분한 것처럼 보였으나, 0.1492 (4MC) 에서 0.134 (5MC)로 악화됨
  - 노이즈에 의한 현상

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Name	Community	Distance from people in 1	Distance from people in 2	Distance from people in 3	Distance from people in 4	Distance from people in 5	Closest	Second Closest	My Cluster	Neighboring Cluster	Silhouette Values		Silhouette
2	Adams	2	2.358	1.434	2.031	2.651	2.371	1.434	2.031	1.434	2.031	0.294		**0.134**
3	Allen	5	2.134	2.247	1.975	2.451	2.017	1.975	2.017	2.017	1.975	-0.021		
4	Anderson	1	0.957	2.247	2.033	2.537	2.135	0.957	2.033	0.957	2.033	0.529		
5	Bailey	2	2.134	1.483	1.975	2.421	2.124	1.483	1.975	1.483	1.975	0.249		
6	Baker	5	2.562	2.415	2.405	2.660	2.381	2.381	2.405	2.381	2.405	0.010		
7	Barnes	4	2.562	2.657	2.405	2.285	2.468	2.285	2.405	2.285	2.405	0.050		
8	Bell	1	1.075	2.657	2.481	2.877	2.521	1.075	2.481	1.075	2.481	0.567		
9	Bennett	3	2.134	1.732	1.156	2.497	2.186	1.156	1.732	1.156	1.732	0.333		
10	Brooks	4	2.562	2.524	1.988	2.248	2.542	1.988	2.248	2.248	1.988	-0.116		
11	Brown	2	2.358	1.364	2.117	2.622	2.332	1.364	2.117	1.364	2.117	0.356		
12	Butler	4	2.750	2.561	2.600	2.405	2.642	2.405	2.561	2.405	2.561	0.061		
13	Campbell	1	1.169	2.461	2.269	2.696	2.307	1.169	2.269	1.169	2.269	0.485		
14	Carter	2	2.562	1.541	2.305	2.808	2.542	1.541	2.305	1.541	2.305	0.331		
15	Clark	5	2.562	2.657	2.398	2.610	2.295	2.295	2.398	2.295	2.398	0.043		
16	Collins	5	2.134	1.875	1.975	2.363	2.068	1.875	1.975	2.068	1.875	-0.093		
17	Cook	1	0.957	2.247	2.033	2.537	2.135	0.957	2.033	0.957	2.033	0.529		

다른 방법들…

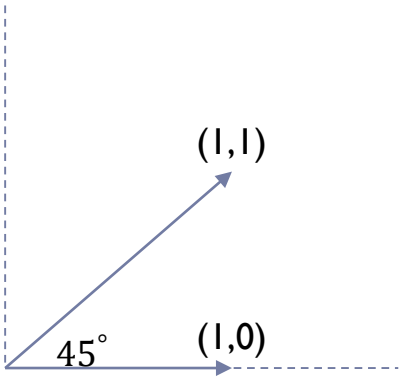
- K-medians clustering
  - 예제의 경우, Vector 값들은 0 또는 1로 구성되어 있음
  - 중심점이 0.113이라는 의미는?
  - 실제 데이터를 구성하는 0 또는 1로만 중심값을 구성할 필요성 존재
  - Outlier에 덜 민감함

- 거리를 계산하는 방식들
  - Euclidean Distance vs. Manhattan distance
  - 0,1구성된 이진데이터의 경우는 Manhattan distance가 더 적절할 수도 있음

  - Symmetric distance vs. Asymmetric distance
  - '구매했다'와 '구매하지 않았다'가 의미의 중요도는 다를 수 있음
    - 재고가 있어서, e-mail을 보지 않아서.. 등 다양한 이유가 존재함
    - 구매하지 않는 것(0) 보다는 구매한 것(1) 을 더 신경써야 할 수도 있음

다른 방법들…

- Asymmetric distance
  - Asymmetric distance인 Cosine distance를 이용
  - Cosine similarity



$$\cos(45^\circ) = \frac{1}{\sqrt{2}\sqrt{1}} = 0.707$$

$$\cos(a,b) = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}}$$

  - 거리 = 1- cosine similarity
  - 분자는 match되는 구매 횟수를 의미, 이를 바탕으로 거리를 계산
  - (1,0) & (1,0) : $\cos(0^\circ) = 1$
  - (1,0) & (0,1) : $\cos(90^\circ) = 0$

다른 방법들…

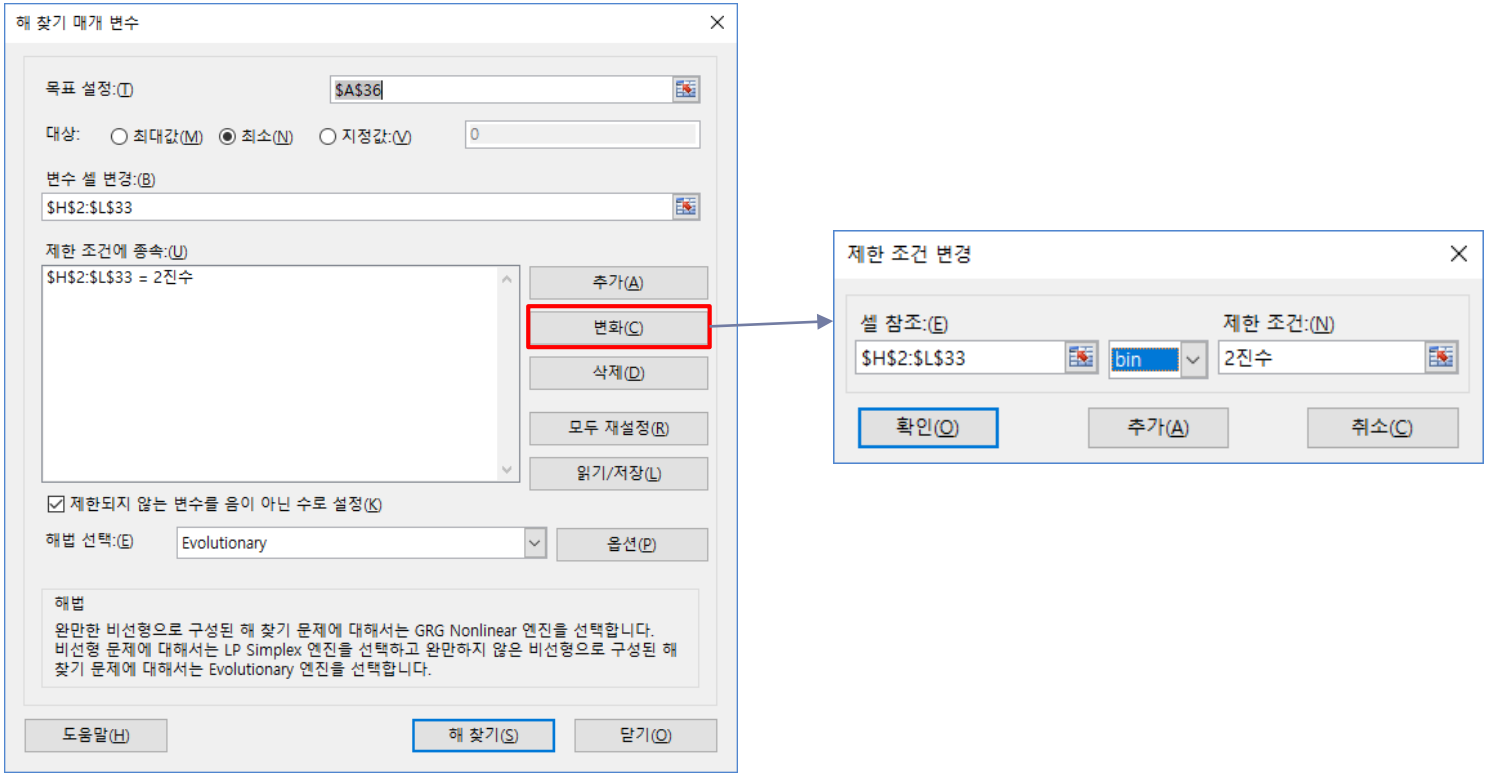
- Clustering에 적용하기 (5MedC 생성)
  - 5MC를 복사하여 5MedC 생성
  - 해찾기 결과 삭제 (H2~L33)
  - 거리 계산 수식 변경 (M34~DH39)
  - M34 =1-SUMPRODUCT(M$2:M$33,$H$2:$H$33)/(SQRT(SUM(M$2:M$33))*SQRT(SUM($H$2:$H$33)))
  - M34 =IFERROR(1-SUMPRODUCT(M$2:M$33,$H$2:$H$33)/(SQRT(SUM(M$2:M$33))*SQRT(SUM($H$2:$H$33))),1)

M34 =IFERROR(1-SUMPRODUCT(M$2:M$33,$H$2:$H$33)/(SQRT(SUM(M$2:M$33))*SQRT(SUM($H$2:$H$33))),1)

	G	H	I	J	K	L	M	N	O	P	Q
1	**Past Peak**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	Adams	Allen	Anderson	Bailey	Baker
29	TRUE										
30	FALSE						1				
31	FALSE						1			1	
32	FALSE										1
33	TRUE										
34	Distance to Cluster 1						1.000000	1.000000	1.000000	1.000000	1.000000
35	Distance to Cluster 2						1.000000	1.000000	1.000000	1.000000	1.000000
36	Distance to Cluster 3						1.000000	1.000000	1.000000	1.000000	1.000000
37	Distance to Cluster 4						1.000000	1.000000	1.000000	1.000000	1.000000
38	Distance to Cluster 5						1.000000	1.000000	1.000000	1.000000	1.000000
39	Min Distance						1.000000	1.000000	1.000000	1.000000	1.000000
40	Assigned Cluster						1	1	1	1	1

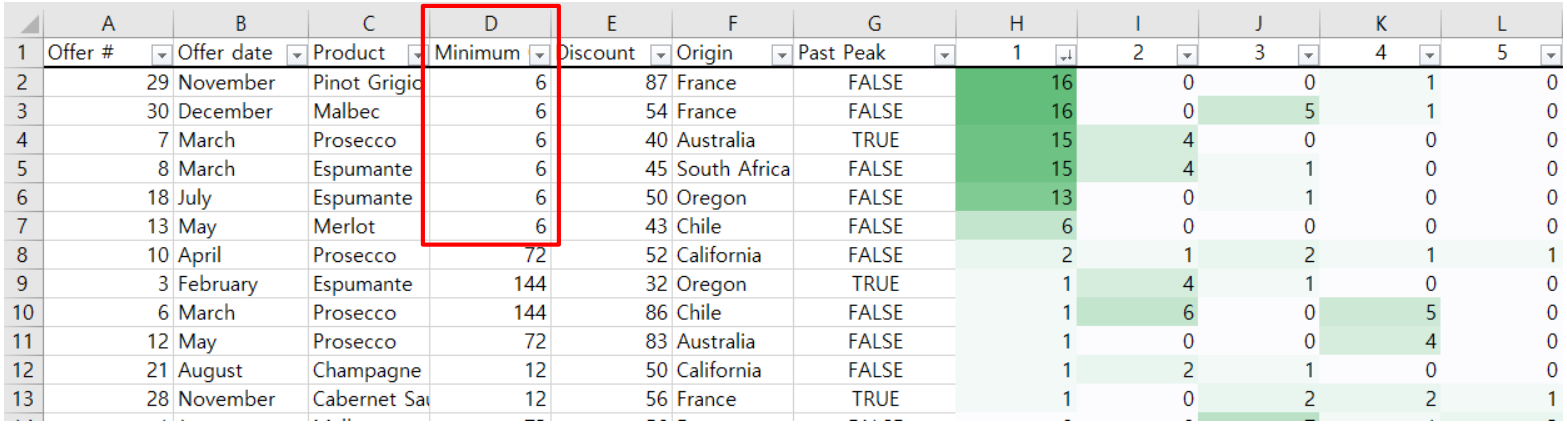
다른 방법들...

- Clustering에 적용하기
  - 의사결정변수를 Binary로 정의



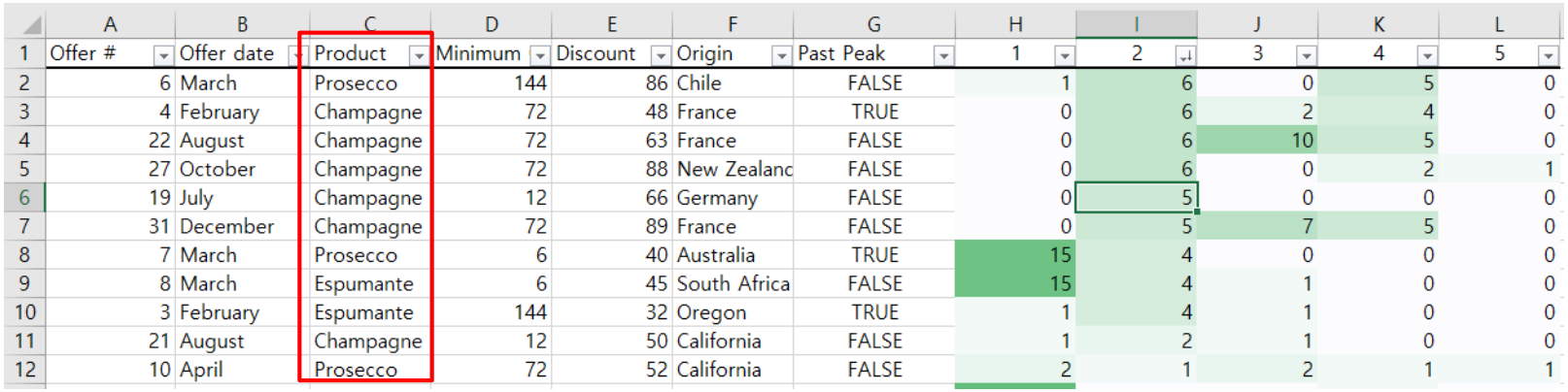
다른 방법들…

- Top Deals 살펴보기
  - Cluster 1 – low volume customers



	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4	5
2	29	November	Pinot Grigio	6	87	France	FALSE	16	0	0	1	0
3	30	December	Malbec	6	54	France	FALSE	16	0	5	1	0
4	7	March	Prosecco	6	40	Australia	TRUE	15	4	0	0	0
5	8	March	Espumante	6	45	South Africa	FALSE	15	4	1	0	0
6	18	July	Espumante	6	50	Oregon	FALSE	13	0	1	0	0
7	13	May	Merlot	6	43	Chile	FALSE	6	0	0	0	0
8	10	April	Prosecco	72	52	California	FALSE	2	1	2	1	1
9	3	February	Espumante	144	32	Oregon	TRUE	1	4	1	0	0
10	6	March	Prosecco	144	86	Chile	FALSE	1	6	0	5	0
11	12	May	Prosecco	72	83	Australia	FALSE	1	0	0	4	0
12	21	August	Champagne	12	50	California	FALSE	1	2	1	0	0
13	28	November	Cabernet Sa	12	56	France	TRUE	1	0	2	2	1

  - Cluster 2 – Sparkling wine



	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4	5
2	6	March	Prosecco	144	86	Chile	FALSE	1	6	0	5	0
3	4	February	Champagne	72	48	France	TRUE	0	6	2	4	0
4	22	August	Champagne	72	63	France	FALSE	0	6	10	5	0
5	27	October	Champagne	72	88	New Zealanc	FALSE	0	6	0	2	1
6	19	July	Champagne	12	66	Germany	FALSE	0	5	0	0	0
7	31	December	Champagne	72	89	France	FALSE	0	5	7	5	0
8	7	March	Prosecco	6	40	Australia	TRUE	15	4	0	0	0
9	8	March	Espumante	6	45	South Africa	FALSE	15	4	1	0	0
10	3	February	Espumante	144	32	Oregon	TRUE	1	4	1	0	0
11	21	August	Champagne	12	50	California	FALSE	1	2	1	0	0
12	10	April	Prosecco	72	52	California	FALSE	2	1	2	1	1

다른 방법들…

- Top Deals 살펴보기
  - Cluster 3 – France

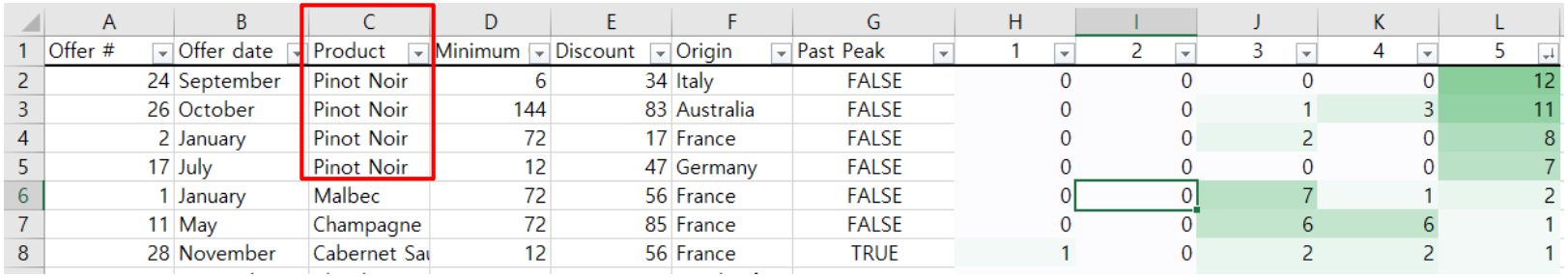
	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4	5
2	22	August	Champagne	72	63	France	FALSE	0	6	10	5	0
3	31	December	Champagne	72	89	France	FALSE	0	5	7	5	0
4	1	January	Malbec	72	56	France	FALSE	0	0	7	1	2
5	11	May	Champagne	72	85	France	FALSE	0	0	6	6	1
6	30	December	Malbec	6	54	France	FALSE	16	0	5	1	0
7	9	April	Chardonnay	144	57	Chile	FALSE	0	0	5	5	0
8	14	June	Merlot	72	64	Chile	FALSE	0	0	4	5	0
9	4	February	Champagne	72	48	France	TRUE	0	6	2	4	0
10	10	April	Prosecco	72	52	California	FALSE	2	1	2	1	1
11	28	November	Cabernet Sa	12	56	France	TRUE	1	0	2	2	1
12	2	January	Pinot Noir	72	17	France	FALSE	0	0	2	0	8

  - Cluster 4 – High volumn

	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4	5
2	11	May	Champagne	72	85	France	FALSE	0	0	6	6	1
3	20	August	Cabernet Sa	72	82	Italy	FALSE	0	0	0	6	0
4	22	August	Champagne	72	63	France	FALSE	0	6	10	5	0
5	31	December	Champagne	72	89	France	FALSE	0	5	7	5	0
6	9	April	Chardonnay	144	57	Chile	FALSE	0	0	5	5	0
7	14	June	Merlot	72	64	Chile	FALSE	0	0	4	5	0
8	15	June	Cabernet Sa	144	19	Italy	FALSE	0	0	1	5	0
9	25	October	Cabernet Sa	72	59	Oregon	TRUE	0	0	1	5	0
10	6	March	Prosecco	144	86	Chile	FALSE	1	6	0	5	0
11	16	June	Merlot	72	88	California	FALSE	0	0	0	5	0
12	4	February	Champagne	72	48	France	TRUE	0	6	2	4	0
13	12	May	Prosecco	72	83	Australia	FALSE	1	0	0	4	0
14	5	February	Cabernet Sa	144	44	New Zealanc	TRUE	0	0	0	4	0
15	32	December	Cabernet Sa	72	45	Germany	TRUE	0	0	0	4	0

다른 방법들…

- Top Deals 살펴보기
  - Cluster 5 – Pinot Noir

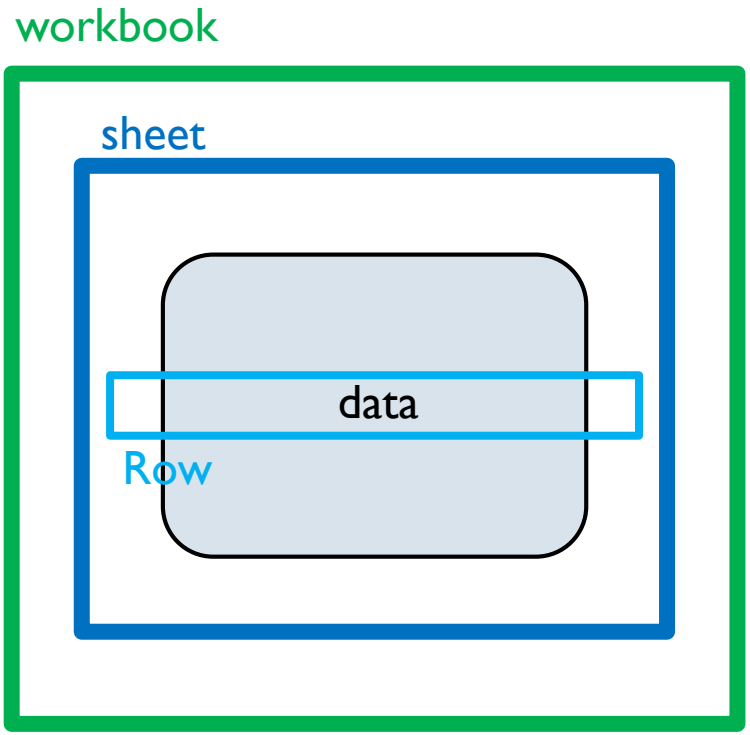


	A	B	C	D	E	F	G	H	I	J	K	L
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4	5
2	24	September	Pinot Noir	6	34	Italy	FALSE	0	0	0	0	12
3	26	October	Pinot Noir	144	83	Australia	FALSE	0	0	1	3	11
4	2	January	Pinot Noir	72	17	France	FALSE	0	0	2	0	8
5	17	July	Pinot Noir	12	47	Germany	FALSE	0	0	0	0	7
6	1	January	Malbec	72	56	France	FALSE	0	0	7	1	2
7	11	May	Champagne	72	85	France	FALSE	0	0	6	6	1
8	28	November	Cabernet Sa	12	56	France	TRUE	1	0	2	2	1

파이썬 설치

- 개발환경 : Jupyter Notebook
- 다운로드 : https://www.anaconda.com/distribution/#download-section
  (window 64bit, 3.7version python)
- 장점
  1. 코드단위로 실행결과를 볼 수 있다.
  2. 아나콘다 설치로 중요한 package가 기본적으로 설치되어 있어서 import만 해서 쓰면 된다.

엑셀에서 데이터 받아오기




```
import pandas as pd
import numpy
import xlrd
book=xlrd.open_workbook('E:/Datasmart/ch02/WineKMC.xlsx')
sheet=book.sheet_by_index(1)
data=[]
for i in range(sheet.nrows) :
    data.append(sheet.row_values(i))
data
```

Pivot table

Customer Last Name / Offer #	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Brooks	Brown	...
1.0	0	0	0	0	0	0	0	0	0	0	...
2.0	0	0	0	0	0	0	1	0	0	0	...
3.0	0	0	0	0	0	0	0	0	1	0	...
4.0	0	0	0	0	0	0	0	0	0	0	...
5.0	0	0	0	0	0	0	0	0	0	0	...
6.0	0	0	0	0	0	0	0	0	0	0	...
7.0	0	0	0	1	1	0	0	0	0	1	...
8.0	0	0	0	0	0	0	0	1	1	0	...

- Import pandas as pd
- pd.pivot(data,
  index=[],
  columns=[]
  values=[]
  fill_value=V
)

K means 와 Silhouette score

K means

- from sklearn.cluster import Kmeans
- kmeans = KMeans(n_clusters=K, random_state=R, max_iter=N).fit(data)

  K = 클러스터링 수
  N = 최대반복수(default=300)
  .fit(data) = 데이터 지정
  kmeans.labels_ = label 결과

  #kmeans.predict() = 예상

Silhouette score

- from sklearn.metrics import silhouette_score
- silhouette_score(data, label, metric='euclidean')

  label = cluster label

- #sli={}

```
sli
```

```
{2: 0.24995851800742472, 3: 0.21154140512205818, 4: 0.1618367827517727}
```

결과

- pt_T[pt_T.clusters==0].sum(axis=0)[:-1]

```
In [50]: wine_df
```

Out[50]:

	Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	clus_0	clus_1	clus_2	clus_3
1	1	January	Malbec	72	56	France	FALSE	0	0	7	3
2	2	January	Pinot Noir	72	17	France	FALSE	0	5	3	2
3	3	February	Espumante	144	32	Oregon	TRUE	0	0	3	3
4	4	February	Champagne	72	48	France	TRUE	0	0	5	7
5	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0	0	2	2
6	6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
7	7	March	Prosecco	6	40	Australia	TRUE	12	1	5	1
8	8	March	Espumante	6	45	South Africa	FALSE	11	0	3	6
9	9	April	Chardonnay	144	57	Chile	FALSE	0	0	3	7
10	10	April	Prosecco	72	52	California	FALSE	0	0	5	2
11	11	May	Champagne	72	85	France	FALSE	0	0	10	3