고객 해지 모델 군집분석

데이터 지식서비스 공학과
이원욱

Since 2015
Dankook University Machine Learning Lab

ML Lab
DANKOOK UNIVERSITY

1. 데이터

*사용 데이터

1) 4월 8일에 해당하는 고객 데이터
-VOC(고객 문의 데이터)
-contact(고객의 계약 정보 데이터)
-customer(고객 정보 데이터)

2) Feature Importance 데이터
-해당 데이터로 학습한 RF 모델의 해지 예측 확률
-각 변수별로 Column의 Feature Importance

3)데이터의 columns 유형
-continuous column: 70
-category, binary columns:39

4)전체건수: 36733*109

	iptv_comb_yn	pstn_comb_yn	mphon_comb_yn	mphon_sbsc_yn	smph_use_yn
0	1	0	0	1	1
1	1	0	0	0	0
2	0	1	1	1	1
3	1	0	0	1	0
4	0	1	0	0	0
...	...	...	...	...	...
36728	1	1	0	0	0
36729	1	1	0	0	0
36730	0	0	0	1	1
36731	1	0	1	1	1
36732	1	0	1	1	1

36733 rows × 109 columns

2. PCA

*FAMD PCA

-Mixed Data(continuous, category 등, 여러 유형의 변수가 섞인 데이터) 사용하는 PCA 방법의 한 종류 *일반 PCA는 연속형 데이터일 때만 가능함

-사용 목적

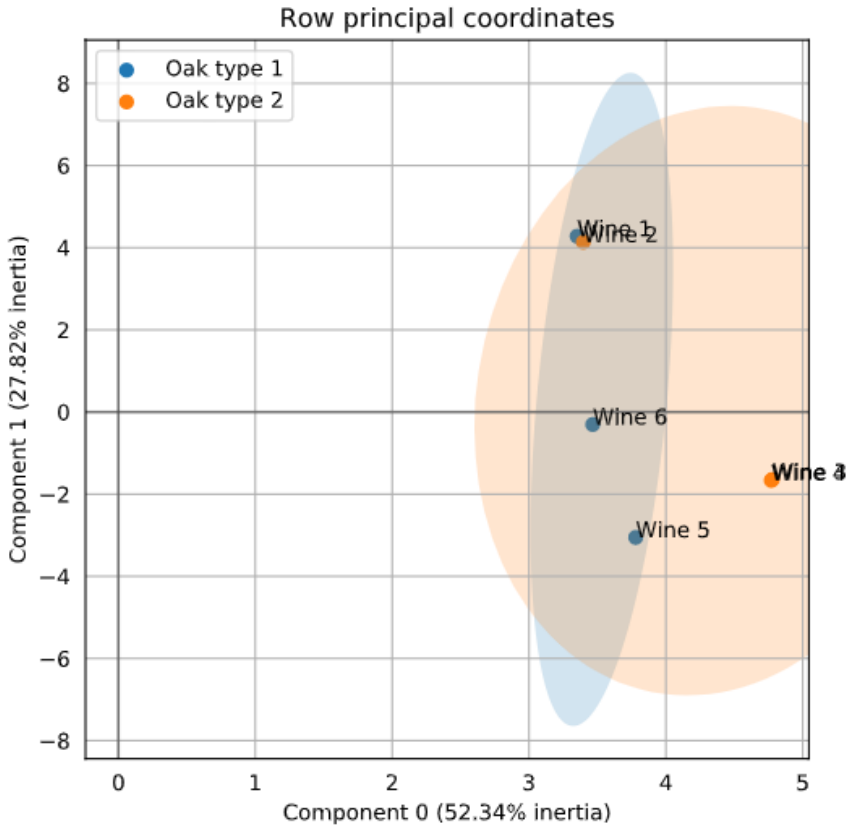
1)군집분석은 거리를 계산하여 군집을 묶어 주는데, Mixed Data에는 적합한 거리 계산식이 없음.

-Euclidean은 연속형 데이터일 때만 가능하고, Jaccard등의 다른 거리 계산 식은 연속형 데이터를 반영하기 힘들기 때문

2)데이터를 2차원으로 축소하여 표에 나타내는 T-SNE역시, binary data에는 사용할 수 없음

2)Feature의 개수가 많음(109개)

-따라서 FAMD PCA를 통해 Continuous한 형태의 Feature를 Extraction하고 K-Means clustering을 진행함



2. PCA

*FAMD PCA

성분 개수	comp0	comp1	comp2	comp3	comp4	comp5	comp6	comp7	comp8	comp9	합계
설명력	0.6925	0.0497	0.0282	0.0220	0.1708	0.0155	0.0124	0.1112	0.0098	0.0085	0.8672

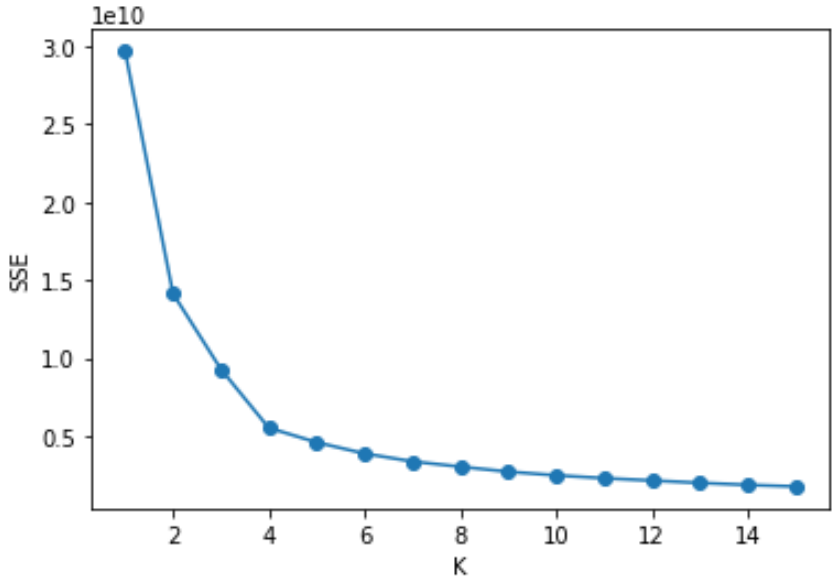
- Comp0 일 때 69%의 설명력을 가지며, featur의 개수 대비 상당히 높은 설명력을 나타냄.
- Comp1부턴 값이 현저히 낮아지며, 크게 의미가 없음을 나타냄
- K-means clustering으로 군집이 형성된 분포를 확인하면, 주성분이 3개 일때 군집이 가장 좋게 형성됨

3. K-means Clustering

*주어진 데이터를 k개의 군집으로 묶은 알고리즘

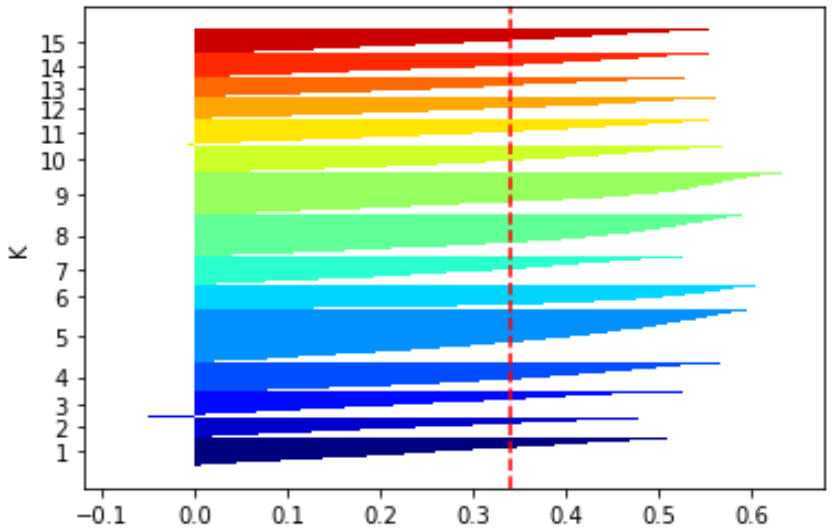
*군집의 개수 설정 방법

1)elbow



- 군집의 개수에 따라 SSE값의 격차가 줄어드는 지점을 찾아 군집의 개수를 설정하는 방법. (k=4)

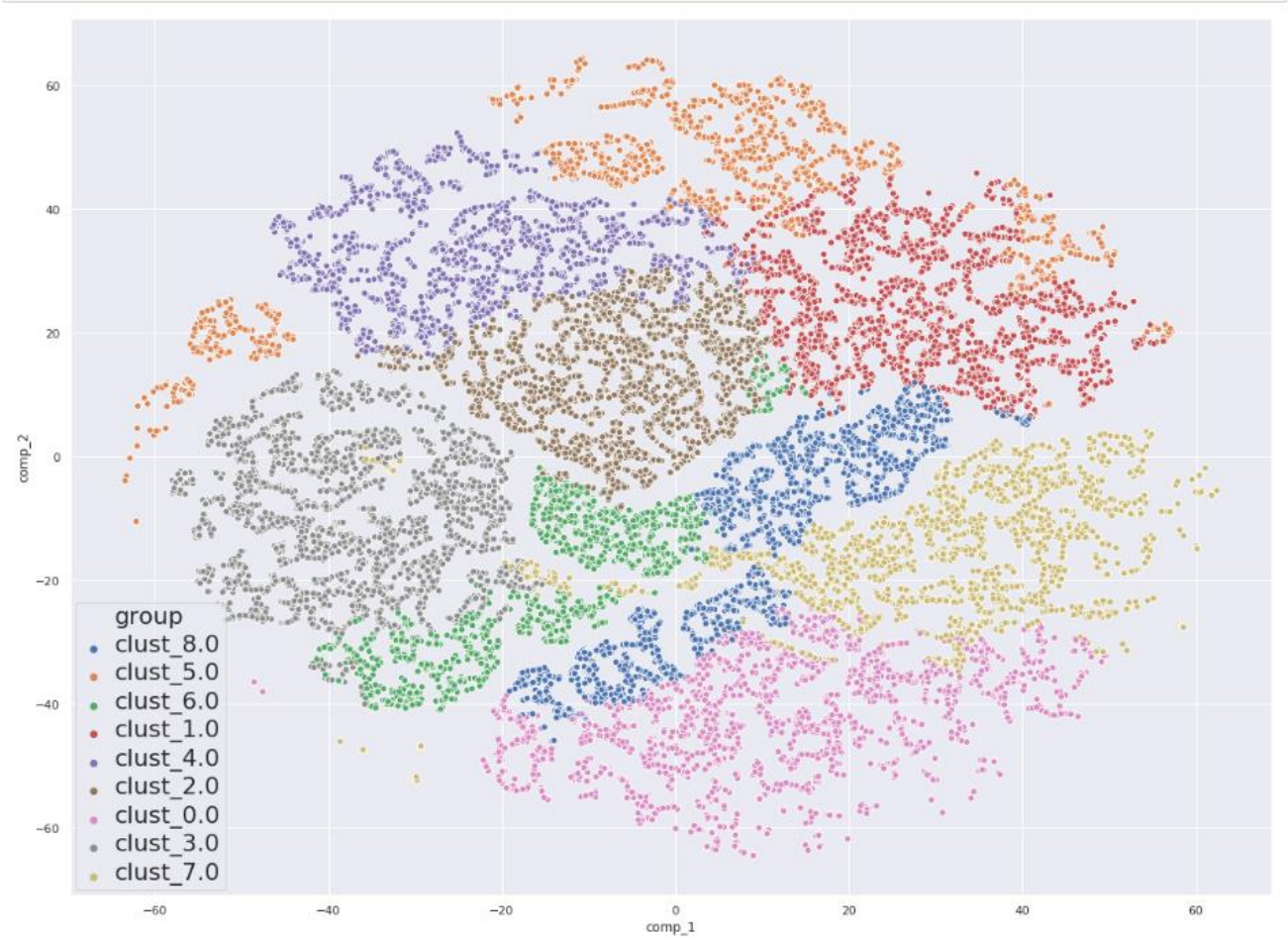
2)Silhouette



- 군집의 개수에 따라 -1에서 1사이의 값을 가지며, 가장 높은 값을 가진 군집의 개수가 적절함(k=9)

3. K-means Clustering

*고객을 최대한 세분화 하는 것을 목표로, 앞선 2가지 방법을 적용하였을 때, 최적의 9개의 군집이 가장 적합하다고 판정



*T-SNE로 군집의 분포를 확인하였을 때, 군집이 적절하게 형성되는 것을 확인할 수 있음

4. Clustering Analysis

*형성된 9개의 군집을 바탕으로 군집들에 유의미한 결과가 있는지 확인하기 위해 통계 분석 진행

1)Feature Importance 데이터에서 Importance가 높은 상위 20개의 변수 추출

2)상위 20개의 데이터가 모두 연속형 변수였음.

	cancel_count	new_date_delta	now_chage_prod_sbsc_date_delta	svc_rl_use_day_num	opn_cont_rl_use_mons_num	r6m_inet_avg_arpu_amt
0	0	281.0	281.0	243.0	8.0	21999.500
1	0	102.0	102.0	64.0	3.0	43601.144
2	0	5518.0	5518.0	5376.0	177.0	19066.667
3	0	1636.0	405.0	858.0	53.0	21778.000
4	0	4795.0	4795.0	4670.0	154.0	45786.666
...	...	...	...	...	...	...
36728	0	3509.0	398.0	3469.0	114.0	21614.833
36729	0	3509.0	398.0	3469.0	114.0	21614.833
36730	0	691.0	691.0	653.0	22.0	21780.000
36731	0	419.0	142.0	381.0	13.0	24585.000
36732	0	896.0	896.0	858.0	29.0	27494.333

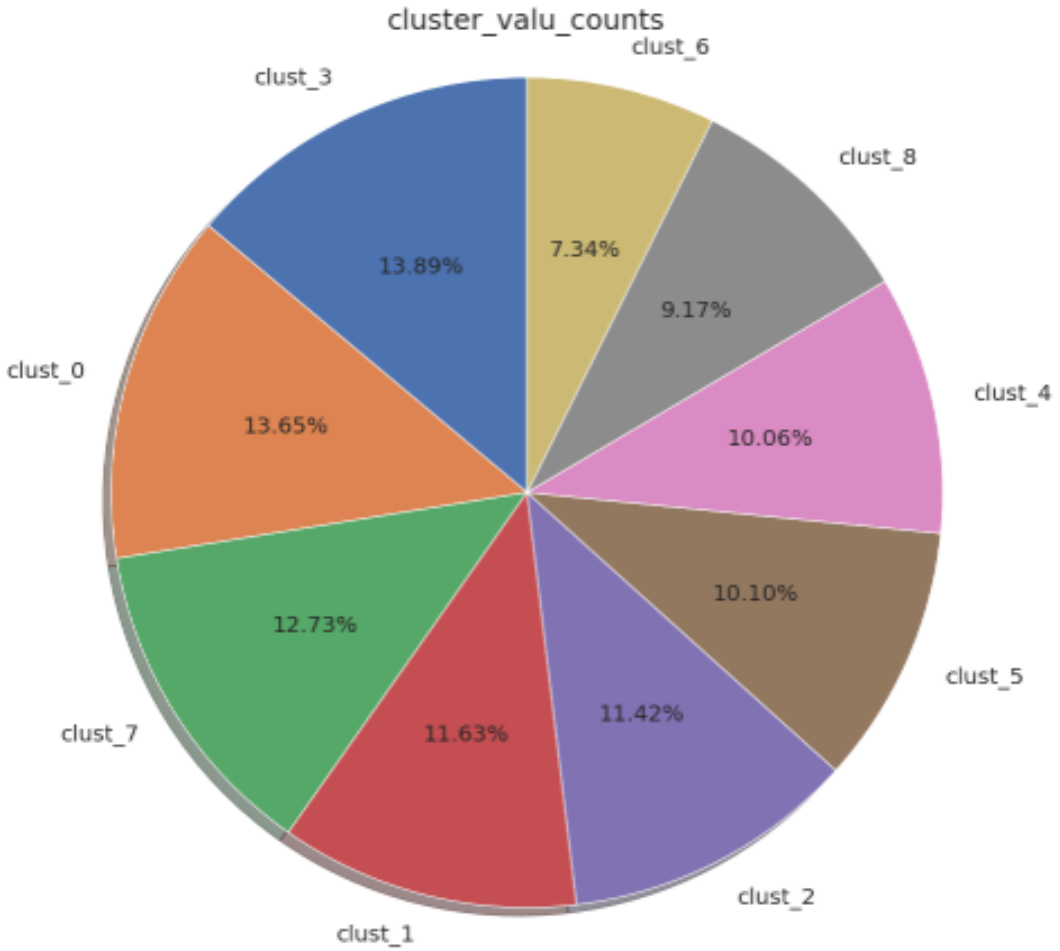
상위 20개의 변수 명

'해지_cnt',
'new_date_delta',
'now_chage_prod_sbsc_date_delta',
'svc_rl_use_day_num',
'opn_cont_rl_use_mons_num',
'r6m_inet_avg_arpu_amt',
'r6m_avg_arpu_amt',
'r3m_avg_arpu_amt',
'cust_age',
'r3m_inet_avg_arpu_amt',
'rmonth_tot_bill_amt',
'svc_use_mons_num',
'inet_engt_exp_rmnd_mons_num',
'engt_rmnd_mons_num',
'r3m_iptv_avg_arpu_amt',
'iptv_engt_exp_rmnd_mons_num',
'comb_engt_exp_rmnd_mons_num',
'mship_rmnd_score',
'r6m_mphon_avg_arpu_amt',
'mphon_comb_circuit_num'

4. Clustering Analysis

*군집별 데이터 개수

	clust_number	value
0	clust_3	5103
1	clust_0	5014
2	clust_7	4675
3	clust_1	4273
4	clust_2	4196
5	clust_5	3711
6	clust_4	3697
7	clust_8	3367
8	clust_6	2697



4. Clustering Analysis

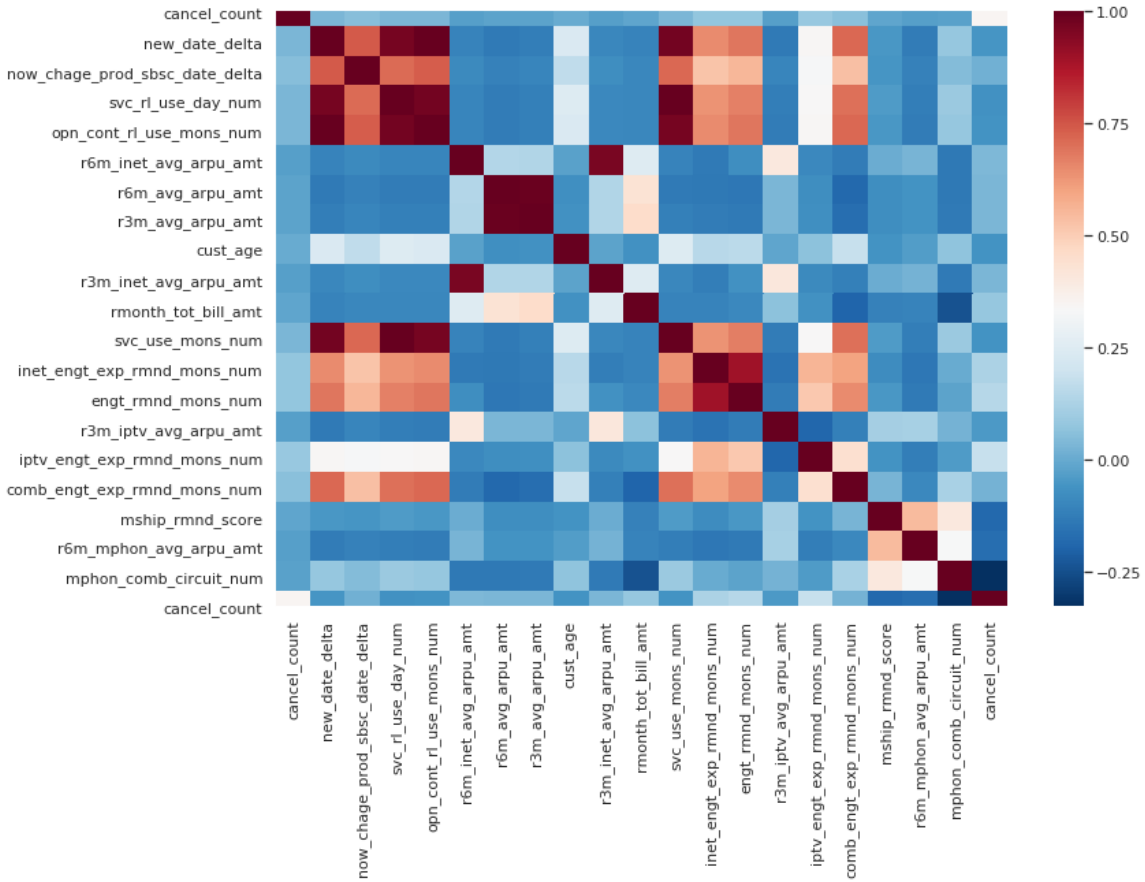
*상위 20개의 변수의 pearson correlation matrix 계산

-빨간색으로 갈수록 두 변수 사이에 상관관계가 높음.

-자기 자신을 제외하고 svc_use_mons_num, svc_rl_use_nm 등의 변수들의 상관관계가 높음

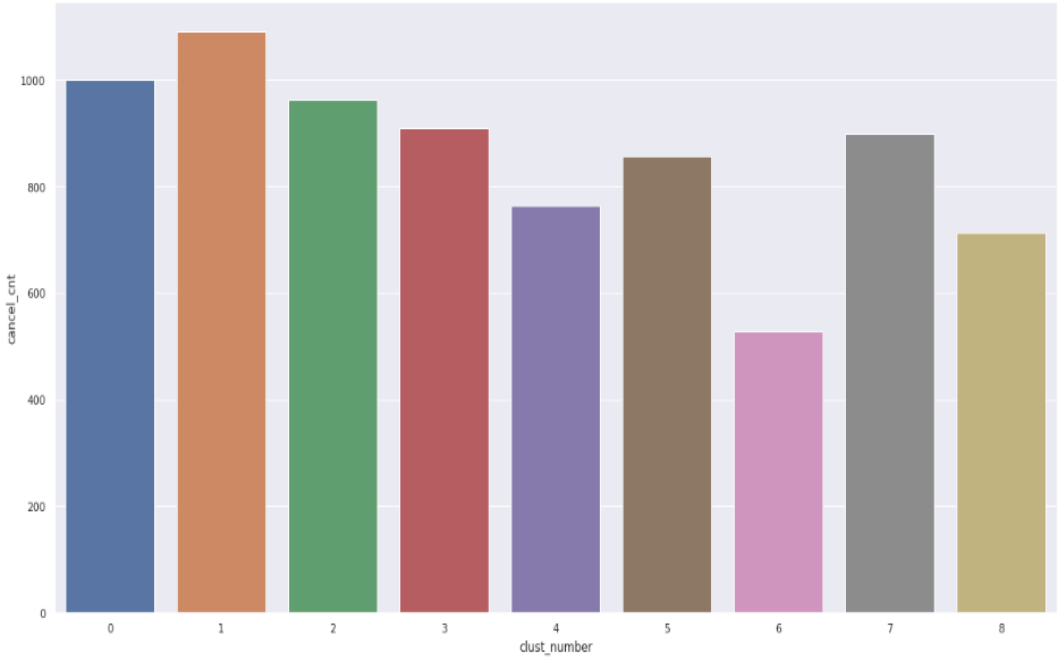
-가장 importance가 높은 해지_cnt와 해지 확률 등은 중요한 변수지만 높은 상관관계를 가진 변수가 없었으나, 두 변수끼리 양의 상관관계가 0.349로 가장 높았음.

cancel_count	0.348998
new_date_delta	0.996629
now_chage_prod_sbsc_date_delta	0.741779
svc_rl_use_day_num	0.996755
opn_cont_rl_use_mons_num	0.996629
r6m_inet_avg_arpu_amt	0.966260
r6m_avg_arpu_amt	0.994618
r3m_avg_arpu_amt	0.994618
cust_age	0.247834
r3m_inet_avg_arpu_amt	0.966260
rmonth_tot_bill_amt	0.456297
svc_use_mons_num	0.996755
inet_engt_exp_rmnd_mons_num	0.896741
engt_rmnd_mons_num	0.896741
r3m_iptv_avg_arpu_amt	0.411862
iptv_engt_exp_rmnd_mons_num	0.560846
comb_engt_exp_rmnd_mons_num	0.718354
mship_rmnd_score	0.548195
r6m_mphon_avg_arpu_amt	0.548195
mphon_comb_circuit_num	0.408488
cancel_prob	0.348998

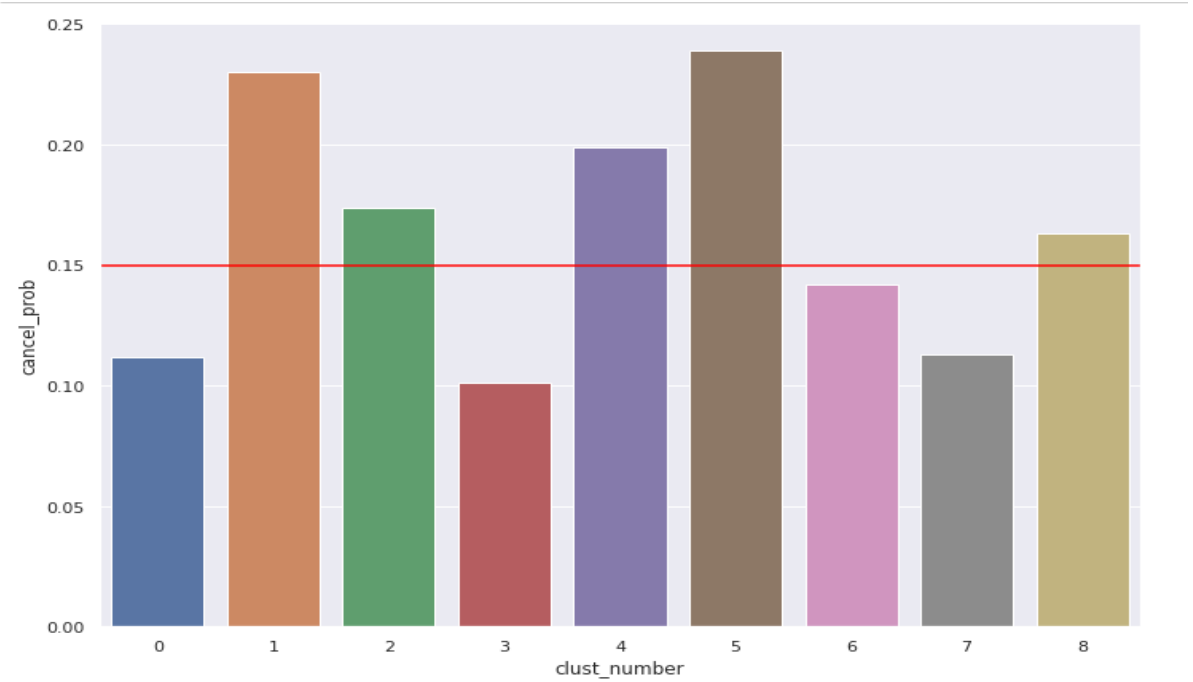


4. Clustering Analysis

*군집별 해지 문의 건수



*군집별 해지 확률



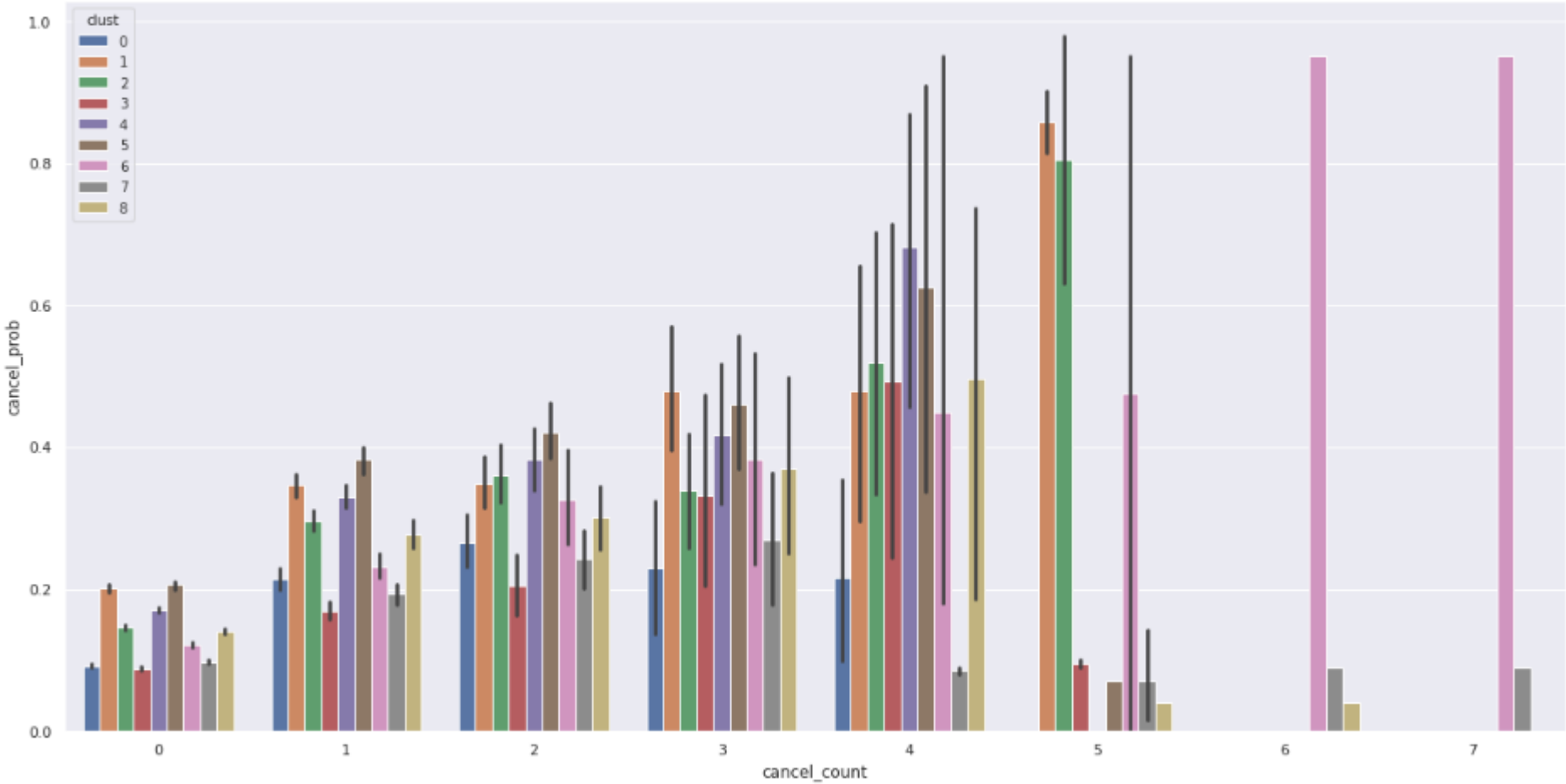
*빨간색선: 전체 해지 확률 평균

-2번 군집은 해지건수와 해지 확률이 모두 높음
-3번 7번 군집은 해지건수는 많은 편이나 해지 확률은 낮은 편임

4. Clustering Analysis

*군집 별 해지 문의 건수와 해지확률 관계표

*X축은 해지 건수이며, y축은 해지 확률



*해지건수가 4번일 경우 4번 군집은 해지할 확률이 0.62인 반면, 7번 군집은 해지건수가 4번일 경우에도 해지확률이 0.1임. 즉, 4번 군집의 고객들이 해지 위험도가 높음

*6번 군집일 경우 해지 문의가 가장 많은 편이며, 해지 확률 또한 가장 높음.

*7번 군집은 해지 문의는 많이 하지만, 해지는 하지 않는 고객 집단으로 확인할 수 있음(black consumer 같은...? 이건 그냥 제 생각입니다.)