개인 프로젝트

분석 프로그래밍

“L사의 고객 세분화를 통한 맞춤형 상품(서비스) 추천”

20192766 마민정

01 ***Collection, Cleansing*** *Data*

02 ***Generation, Selection*** *Feature*

03 ***Clustering*** *Analysis*

04 ***RFM*** *Analysis*

05 ***Market Basket*** *Anlaysis*

06 ***Recommender*** *System*

Collection
Data

DATA COLLECTION

"**로그 데이터**를 포함하는 온라인 데이터를 나눠서 저장해준다."

Online Data Collection

1. 기존에 만들었던 df 데이터에서 biz_unit이 A01거나 A02거나 A03인 데이터를 뽑아낸다.

2. 로그 데이터와 df 데이터에서 공통되는 열인 clnt_id, trans_id, biz_unit을 기준으로 merge해준다.

3. 온라인 데이터만 다루는 df_on변수가 생성된다.

Offline Data Collection

1. 기존에 만들었던 df 데이터에서 biz_unit이 B01거나 B02거나 B03인 데이터를 뽑아낸다.

2. 오프라인 데이터만 다루는 df_off변수가 생성된다.

Cleansing
Data

DATA CLEANSING

"온라인, 오프라인 별로 **결측치**와 **이상치** 등의 **전처리**를 해준다."

Online Missing Value

1. 모든 값이 널 값인 'sech_kwd' 열과, 전체의 70% 이상이 널 값인 'dvc_ctg_nm' 열 삭제
2. 나머지 결측값이 있는 행 제거

Offline Missing Value

1. 결측값을 조사하였을 때 clac_nm3 열에만 17개의 결측치가 있기 때문에 제거

Online Outlier

1. buy_am이 0인 값은 사지 않았다고 판단하여 제거(277개)
2. 박스플롯을 그렸을 때 0.8*10**7이상인 이상치의 값 2개가 발견되어 이상치 처리를 해줌.

Offline Outlier

1. buy_am가 0이 아닌 경우 물건을 샀다고 판단하여 buy_ct를 1로 치환(6244)
2. 박스플롯을 그렸을 때 그 외 특별한 이상치는 없음.

Feature
Generation

FEATURE GENERATION

"온라인 데이터에서 **주조회시각, 조회빈도, 방문횟수**의
FEATURE를 만들어준다."

	clnt_id	총구매액	평균구매액	최대구매액	구매상품수(중)	내점일수	구매주기	주말방문율	거래당구매건수	구매추세	주구매시간	고가상품구매율	주조회시각
0	34516	808517	8335	223000	45	15	5	0.12	5.4	-1.16	17	0.11	15
1	28454	213640	2513	12800	28	9	8	0.33	8.5	0.25	21	0.00	14
2	25782	1233224	5409	46800	61	52	1	0.25	2.1	1.39	22	0.05	15
3	65774	2499406	5153	49900	77	65	1	0.17	5.4	0.88	19	0.07	13
4	33801	935960	4500	24000	43	51	1	0.23	3.7	-0.41	19	0.07	14
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7389	47847	22900	22900	22900	1	1	0	0.00	1.0	0.00	16	1.00	13
7390	64655	25000	12500	15000	1	1	0	1.00	2.0	0.00	18	0.50	23
7391	41344	99800	99800	99800	1	1	0	0.00	1.0	0.03	18	1.00	10
7392	36971	32000	32000	32000	1	1	0	0.00	1.0	0.01	16	1.00	12
7393	60852	39600	39600	39600	1	1	0	0.00	1.0	-0.10	12	1.00	22

1. 주조회시각

로그데이터의 hit_tm(조회시각) 값에서 시간을 추출하여 clnt_id 별로 가장 많이 조회한 시각을 알려주는 '주조회시각'이라는 피처를 만들어준다.

Feature
Generation

FEATURE GENERATION

"온라인 데이터에서 **주조회시각, 조회빈도, 방문횟수**의
FEATURE를 만들어준다."

	clnt_id	총구매액	평균구매액	최대구매액	구매상품수(중)	내점일수	구매주기	주말방문율	거래당구매건수	구매추세	주구매시간	고가상품구매율	주조회시각	조회빈도
0	34516	808517	8335	223000	45	15	5	0.12	5.4	-1.16	17	0.11	15	10.58
1	28454	213640	2513	12800	28	9	8	0.33	8.5	0.25	21	0.00	14	30.62
2	25782	1233224	5409	46800	61	52	1	0.25	2.1	1.39	22	0.05	15	4.32
3	65774	2499406	5153	49900	77	65	1	0.17	5.4	0.88	19	0.07	13	23.67
4	33801	935960	4500	24000	43	51	1	0.23	3.7	-0.41	19	0.07	14	41.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7389	47847	22900	22900	22900	1	1	0	0.00	1.0	0.00	16	1.00	13	1.00
7390	64655	25000	12500	15000	1	1	0	1.00	2.0	0.00	18	0.50	23	1.48
7391	41344	99800	99800	99800	1	1	0	0.00	1.0	0.03	18	1.00	10	15.27
7392	36971	32000	32000	32000	1	1	0	0.00	1.0	0.01	16	1.00	12	1.00
7393	60852	39600	39600	39600	1	1	0	0.00	1.0	-0.10	12	1.00	22	1.25

2. 조회빈도

로그데이터의 hit_seq(조회일련번호)값을 추출하여 clnt_id와 sess_id별로 얼마나 많은 조회를 했는지 평균값을 알려주는 '조회빈도'라는 피처를 만들어준다.

Feature
Generation

FEATURE GENERATION

"온라인 데이터에서 **주조회시각, 조회빈도, 방문횟수**의
FEATURE를 만들어준다."

	clnt_id	총구매액	평균구매액	최대구매액	구매상품수(중)	내점일수	구매주기	주말방문율	거래당구매건수	구매추세	주구매시간	고가상품구매율	주조회시각	조회빈도	방문횟수
0	34516	808517	8335	223000	45	15	5	0.12	5.4	-1.16	17	0.11	15	10.58	127
1	28454	213640	2513	12800	28	9	8	0.33	8.5	0.25	21	0.00	14	30.62	1378
2	25782	1233224	5409	46800	61	52	1	0.25	2.1	1.39	22	0.05	15	4.32	82
3	65774	2499406	5153	49900	77	65	1	0.17	5.4	0.88	19	0.07	13	23.67	142
4	33801	935960	4500	24000	43	51	1	0.23	3.7	-0.41	19	0.07	14	41.00	123
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7389	47847	22900	22900	22900	1	1	0	0.00	1.0	0.00	16	1.00	13	1.00	1
7390	64655	25000	12500	15000	1	1	0	1.00	2.0	0.00	18	0.50	23	1.48	62
7391	41344	99800	99800	99800	1	1	0	0.00	1.0	0.03	18	1.00	10	15.27	229
7392	36971	32000	32000	32000	1	1	0	0.00	1.0	0.01	16	1.00	12	1.00	1
7393	60852	39600	39600	39600	1	1	0	0.00	1.0	-0.10	12	1.00	22	1.25	5

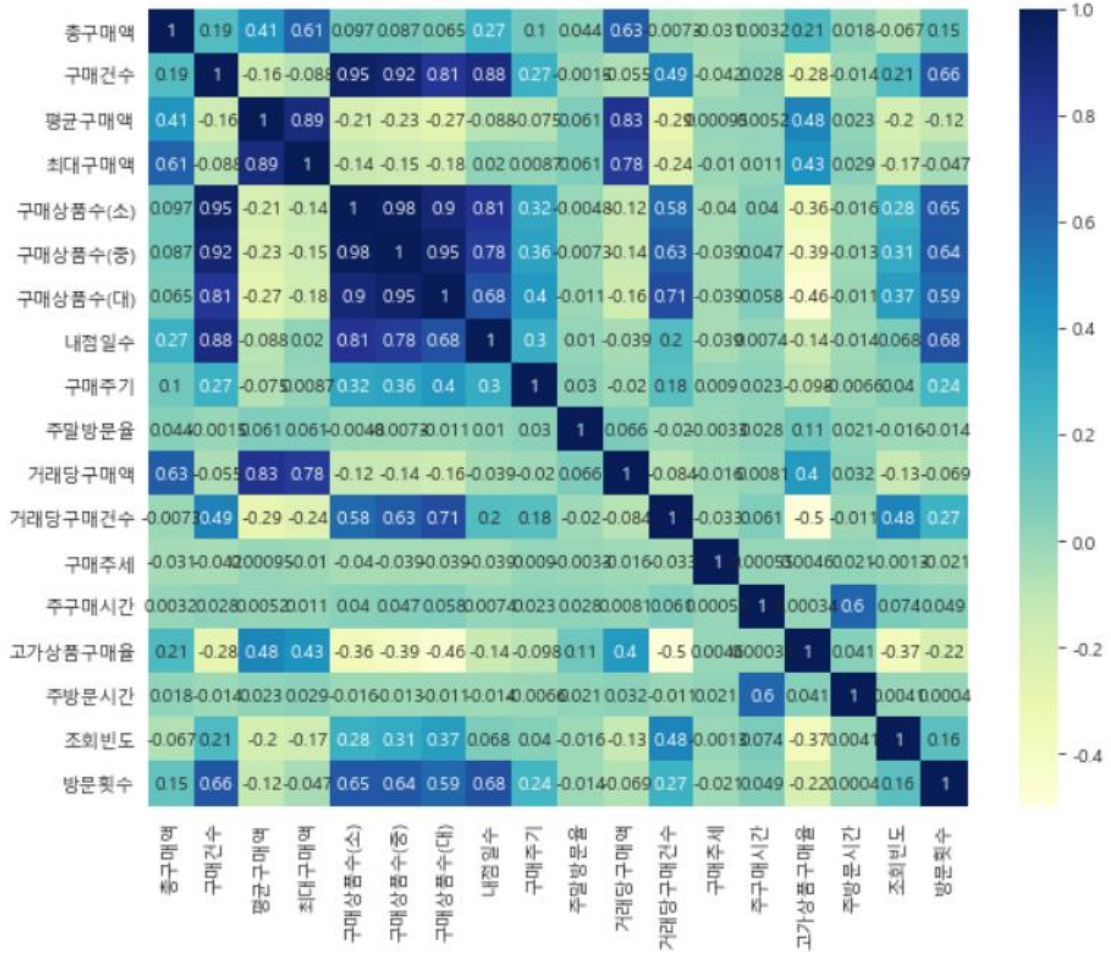
3. 방문횟수

로그데이터의 sess_dt(세션일자)값에서 방문일자를 추출하여 몇 번 방문했는지 알려주는 '방문횟수'라는 새로운 피처를 만들어준다.

Feature
Selection

FEATURE SELECTION

"군집분석이 잘 될 수 있도록 유사한 **변수를 삭제**해준다."



Online Feature 삭제 기준

Heatmap그래프로 상관관계 정도를 그려봄.

상관관계가 0.9가 넘는 열을 살펴보았을 때
구매상품수(소)&(중)&(대), 구매상품수(중)&구매건수,
최대구매액 & 평균구매액 열이 서로 연관이 있음.

구매상품수(중)이 다른 열과 가장 높은 상관관계를 가지고 있기 때문에 남겨놓고 '구매상품수(소)',
'구매상품수(대)'열을 삭제함.

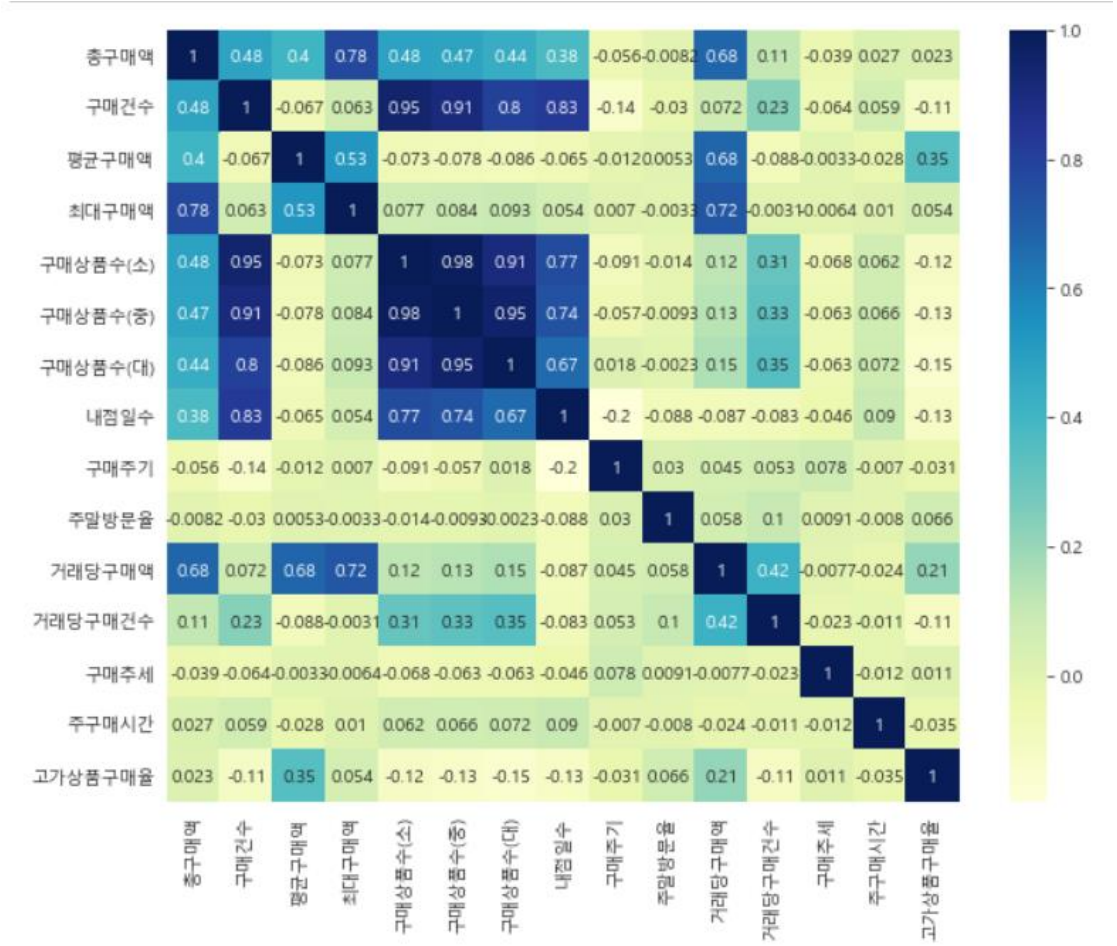
구매상품수(중)을 위에서 남겨놓았기 때문에 '구매건수' 열을 삭제함.

'최대구매액'보다 '평균구매액'이 더 의미 있다고 판단하여 '최대구매액'열을 삭제함.

Feature
Selection

FEATURE SELECTION

"군집분석이 잘 될 수 있도록 유사한 **변수를 삭제**해준다."



Offline Feature 삭제 기준

Heatmap그래프로 상관관계 정도를 그려봄.

상관관계가 0.9가 넘는 열을 살펴보았을 때
구매상품수(소)&(중)&(대), 구매상품수(중)&구매건수,
거래당구매액&최대구매액 열이 서로 연관이 있음.

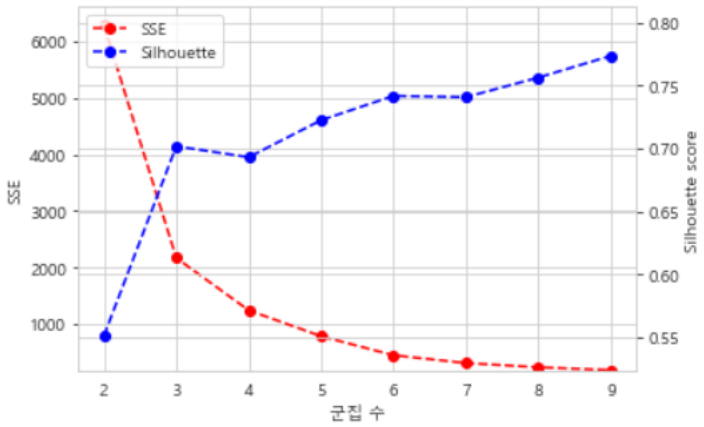
구매상품수(중)이 다른 열과 가장 높은 상관관계를 가지고 있기 때문에 남겨놓고 '구매상품수(소)',
'구매상품수(대)'열을 삭제함.

구매상품수(중)을 위에서 남겨놓았기 때문에
'구매건수',열을 삭제함.

'최대구매액'보다 '평균구매액'이 더 의미 있다고 판단하여
'최대구매액'열을 삭제함.

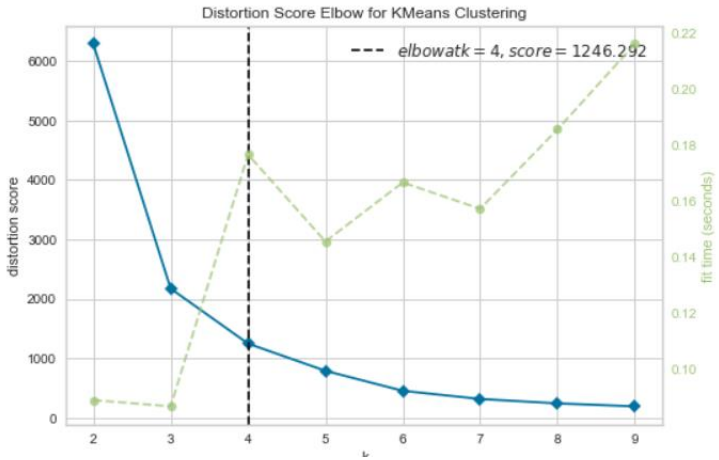
CLUSTER ANALYSIS

"**온라인 데이터**에서 최적의 군집 수를 찾아준다."



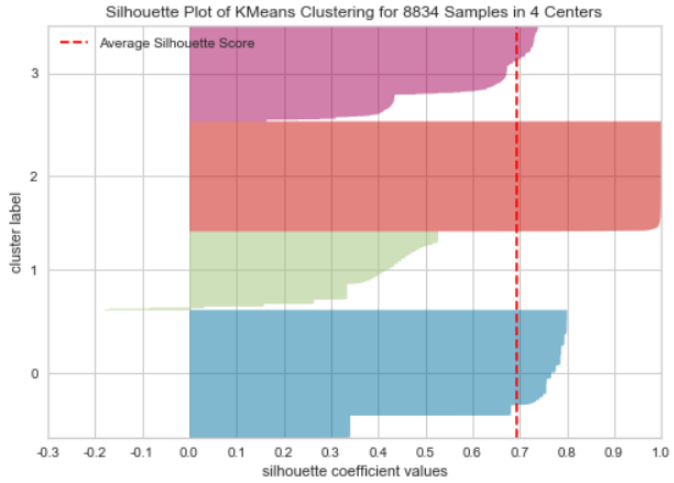
SSE & Silhouette Score

군집 수에 따른 SSE와 Silhouette score 시각화했을 때 3~4부근에서 그래프가 많이 꺾임



Distortion Score

Elbow메소드로 보았을 때 Elbow(팔꿈치)가 4이고 Score가 1246일 때가 최적의 분류

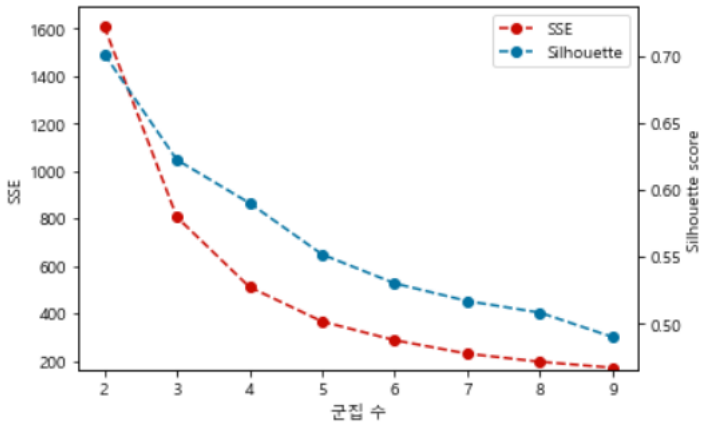


Silhouette Plot

K-Means를 4로 두었을 때 각 군집의 샘플 대부분이 실루엣 점수를 초과해 군집의 수가 적당한 것으로 판단

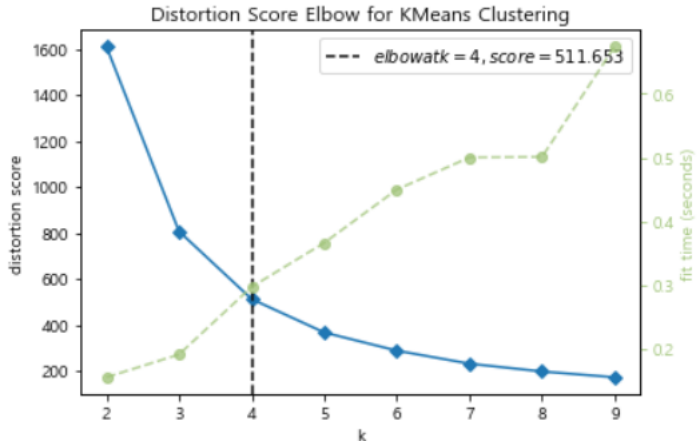
CLUSTER ANALYSIS

"**오프라인 데이터**에서 최적의 군집 수를 찾아준다."



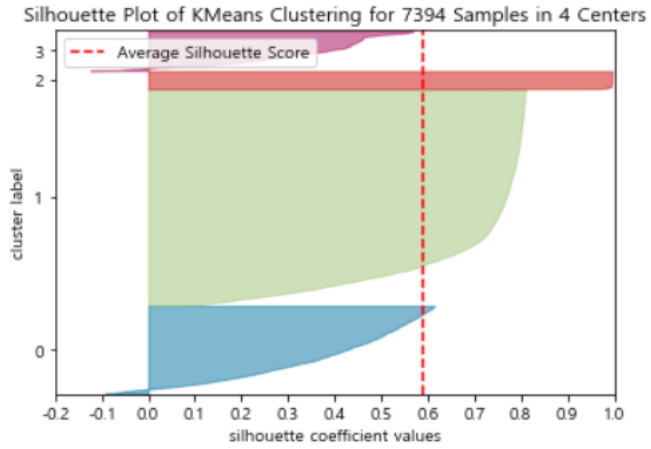
SSE & Silhouette Score

군집 수에 따른 SSE와 Silhouette score 시각화 했을 때 3~5부근에서 그래프가 많이 꺾임



Distortion Score

Elbow메소드로 보았을 때 Elbowatk(팔꿈치)가 4이고 Score가 511일 때가 최적의 분류



Silhouette Plot

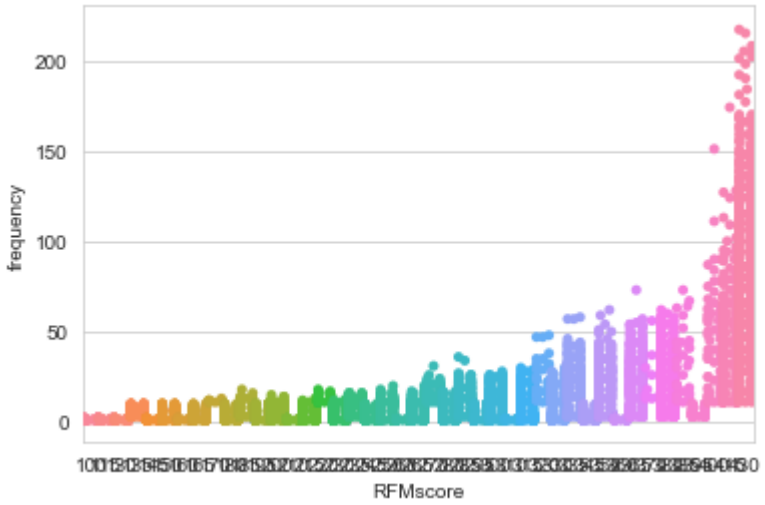
K-Means를 4로 두었을 때 각 군집의 샘플 대부분이 실루엣 점수를 초과해 군집의 수가 적당한 것으로 판단

RFM ANALYSIS

"온라인 데이터에서 RFM을 정의하고 군집별로 **RFM분석**을 해준다. "

온라인 RFM데이터에서 RFMscore별 frequency의 swarmplot을 그렸을 때 점수가 높은 고객의 빈도수가 높음을 알 수 있다.

	세그먼트	recency	frequency	monetary	R_level	F_level	M_level	N	순위	고객분류
0	0	41.032058	13.998179	188180.614208	↑	↑	↓	2745	3	About To Sleep
1	1	41.234597	13.845379	225213.186019	↑	↑	↑	1688	1	Champions
2	2	40.172253	14.378023	207746.267713	↓	↑	↑	2357	2	Potential Loyalists
3	3	40.960861	13.118395	183338.409980	↑	↓	↓	2044	4	Hibernating



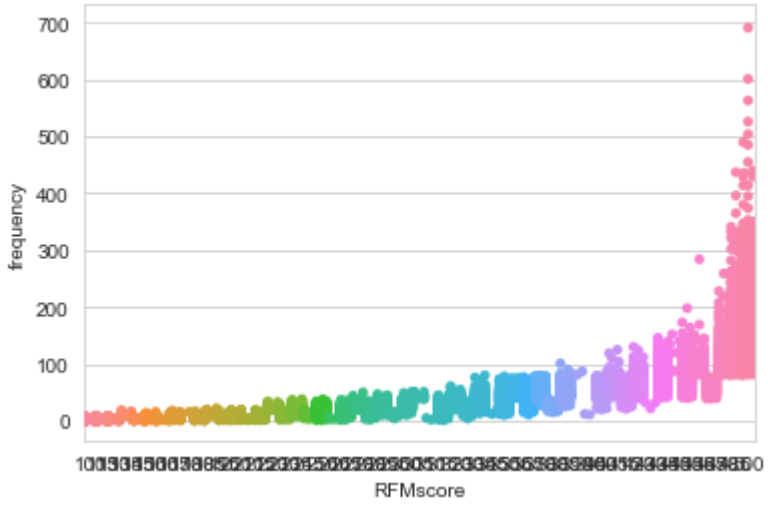
OLINE RFM ANALYSIS

1. 4개의 군집별로 recency, frequency, monetary값의 평균을 구함.
2. 전체 평균과 비교해서 화살표로 표시해준다.
3. 군집의 개수를 구해준다.
4. R이 작을수록, F와 M이 클수록 좋은 군집이므로 순위를 매겨준다.
5. 군집별로 특성을 가장 잘 나타내어 주는 이름을 지정해준다.

RFM ANALYSIS

"오프라인 데이터에서 RFM을 정의하고 군집별로 **RFM분석**을 해준다. "

오프라인 RFM데이터에서 RFMscore별 frequency의 swarmplot을 그렸을 때 점수가 높은 고객의 빈도수가 높음을 알 수 있다.



	세그먼트	recency	frequency	monetary	R_level	F_level	M_level	N	순위	고객분류
0	0	41.032058	13.998179	188180.614208	↑	↑	↓	2745	3	About To Sleep
1	1	41.234597	13.845379	225213.186019	↑	↑	↑	1688	1	Champions
2	2	40.172253	14.378023	207746.267713	↓	↑	↑	2357	2	Potential Loyalists
3	3	40.960861	13.118395	183338.409980	↑	↓	↓	2044	4	Hibernating

OFFLINE RFM ANALYSIS

1. 4개의 군집별로 recency, frequency, monetary값의 평균을 구함.
2. 전체 평균과 비교해서 화살표로 표시해준다.
3. 군집의 개수를 구해준다.
4. R이 작을수록, F와 M이 클수록 좋은 군집이므로 순위를 매겨준다.
5. 군집별로 특성을 가장 잘 나타내어 주는 이름을 지정해준다.

BASKET
Analysis

MARKET BASKET ANALYSIS

"식품이 너무 많은 관계로 **clac_nm1**(상품대분류명)에서 식품이 아닌 것들의 **clac_nm2**(상품중분류명)을 묶어준다."

- **Fashion** <- "Women's Clothing", "Men's Clothing", 'Underwear / Socks and Hosiery / Homewear', "Kids' Clothing", 'Fashion Accessories'
- **Beauty** <- 'Personal Care', 'Home Decor / Lighting', 'Cosmetics / Beauty Care'
- **Health&Sports** <- 'Health Care', 'Outdoor / Leisure Activities', 'Ball Game / Field Sports', 'Seasonal Sports', 'Health / Fitness Training', 'Travel / Leisure Services', 'Sport Fashion'
- **Household_Goods** <- 'Detergents / Hygiene Goods', 'Birth Supplies / Baby Products', 'Kitchenware', 'Tableware / Cooking Utensils', 'Cleaning / Laundry / Bathroom Accessories', 'Home / Kitchen Appliances', 'Refrigerators and Washing Machines'
- **Stationary** <- 'Stationary / Office Supplies', 'Toy', 'Gift Certificates / Cards', 'Books / Records / Instruments'
- **Funiture&Instrument** <- 'Tools / Safety Supplies', 'Heating / Cooling Electronics', 'Automotive Products', 'Cell Phones / Accessories', 'Furniture', 'Computers', 'Video / Audio System Electronics'
- **Others** <- "Others (Non-Products)", "Other Products", "Gardening / Pets", "Bedding / Handicraft"

BASKET
Analysis

MARKET BASKET ANALYSIS

"**지지도**와 **신뢰도**별 규칙을 도출한다."

- 지지도가 0.2이상, 신뢰도가 0.7이상인 **온라인 데이터**에서 규칙 6가지 도출

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Ham and Sausages)	(Frozen Instant Foods)	0.146027	0.296242	0.106520	0.729457	2.462372	0.063261	2.601284
1	(Chilled Instant Foods, Eggs)	(Frozen Instant Foods)	0.131424	0.296242	0.100747	0.766581	2.587685	0.061814	3.014994
2	(Chilled Instant Foods, Instant Noodles)	(Frozen Instant Foods)	0.136971	0.296242	0.106860	0.780165	2.633542	0.066283	3.201306
3	(Instant Noodles, Eggs)	(Frozen Instant Foods)	0.143650	0.296242	0.106973	0.744681	2.513760	0.064418	2.756386
4	(Milk, Eggs)	(Frozen Instant Foods)	0.142404	0.296242	0.100634	0.706677	2.385475	0.058448	2.399262
5	(Instant Noodles, Milk)	(Frozen Instant Foods)	0.140140	0.296242	0.102332	0.730210	2.464912	0.060816	2.608541

- 지지도가 0.2이상, 신뢰도가 0.8이상인 **오프라인 데이터**에서 규칙 20가지 도출

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Biscuits)	(Snacks)	0.380173	0.528131	0.314850	0.828175	1.568124	0.114069	2.746219
1	(Candies)	(Snacks)	0.248850	0.528131	0.205843	0.827174	1.566229	0.074417	2.730312
2	(Beauty, Domestic Fruits)	(Household_Goods)	0.259400	0.537328	0.214633	0.827424	1.539888	0.075251	2.680984
3	(Beauty, Milk)	(Household_Goods)	0.248715	0.537328	0.208006	0.836324	1.556451	0.074365	2.826760
4	(Beauty, Snacks)	(Household_Goods)	0.289965	0.537328	0.235867	0.813433	1.513849	0.080061	2.479924
...	...	...	...	...	...	...	...	...	...
16	(Ham and Sausages, Household_Goods)	(Snacks)	0.249797	0.528131	0.204896	0.820249	1.553117	0.072970	2.625127
17	(Instant Noodles, Household_Goods)	(Snacks)	0.289694	0.528131	0.233973	0.807656	1.529273	0.080977	2.453261
18	(Imported Fruits, Tofu / Bean Sprouts)	(Snacks)	0.250203	0.528131	0.200298	0.800541	1.515799	0.068158	2.365739
19	(Instant Noodles, Milk)	(Snacks)	0.254936	0.528131	0.206384	0.809549	1.532857	0.071744	2.477641
20	(Instant Noodles, Tofu / Bean Sprouts)	(Snacks)	0.245334	0.528131	0.200839	0.818633	1.550057	0.071270	2.601734

Recommeder System

RECOMMENDER SYSTEM

"아이디별로 비슷한 소비를 한 **Neighbors**와 그 이웃이 가장 많이 구매한 10개의 **상품** 추천해준다."

Online Recommender System

clnt_id	0	1	2	3	4	5	6	7	8	9	10
2	2	44452	63702	66283	22833	53058	55422	59881	23328	8183	5153
9	9	68543	14579	7962	67726	66745	8223	35718	44507	14310	48218
23	23	55354	15202	13473	51913	46723	25548	25609	25602	25599	25598
24	24	67553	70174	57427	39801	40335	26337	10126	57210	46922	38885
38	38	42011	32150	17907	5763	9497	14387	13056	70557	70503	56830
...	...	...	...	...	...	...	...	...	...	...	...
72373	72373	45612	52181	54738	42644	47804	28773	8120	65653	34125	42509
72400	72400	61950	50120	14381	69550	57771	37241	4052	44180	36121	23402
72410	72410	34991	5949	18997	24879	41000	53725	53049	31009	42580	12013
72423	72423	37388	71939	11322	25504	25505	25508	25526	25531	25535	25626
72424	72424	27242	70914	4309	54171	57671	2307	15059	13355	9790	21289

clnt_id	recommend_items
2	[Ham, Frozen Korean Pancakes, Bibim Ramens, Wa...
9	[Soy Sauces, Spoon Type Yogurts, Soybean Sprou...
23	[Fresh Milk, Infant / Toddlers' Pants, Women's...
24	[Bibim Ramens, Frozen Korean Pancakes, Frozen ...
38	[Fruit Tea, Frozen Vegetables, Fruit / Vegetab...
...	...
72373	[Makeup Sets, Men's Panties, Women's Running /...
72400	[kelp, Frozen Vegetables, Garlic, Functional M...
72410	[Frozen Tteokbokkis, Grapes, Potato Snacks, Pa...
72423	[Potato Snacks, Chicken Eggs, Crackers, Other ...
72424	[Domestic Porks - Picnics, Bibim Ramens, Ramen...

1. clnt_id 별로 비슷한 소비를 한 K-nearest neighbors 생성
2. 이미 구매한 상품을 제외하고 유사집단에서 가장 많이 구매한 10 개의 상품을 추천

Recommeder System

RECOMMENDER SYSTEM

"아이디별로 비슷한 소비를 한 **Neighbors**와 그 이웃이 가장 많이 구매한 10개의 **상품** 추천해준다."

Offline Recommender System

clnt_id	0	1	2	3	4	5	6	7	8	9	10
9	9	17920	56834	10800	9110	55715	71940	39810	17506	47928	54651
12	12	64406	36089	395	909	49148	67955	53656	51950	58558	25451
20	20	29642	15985	1261	33286	1230	45034	1853	36346	53569	28215
23	23	1316	62234	70144	61990	69697	66940	18829	35999	22401	33794
24	24	63085	28210	26157	64139	15237	25642	41795	37424	22278	6706
...	...	...	...	...	...	...	...	...	...	...	...
72340	72340	39812	4581	44407	25782	5843	30359	52715	64651	52111	29113
72356	72356	48886	48379	13608	50944	24621	30889	19158	12488	56370	58452
72410	72410	69730	36518	45171	39811	39020	3425	53780	58764	38264	58044
72423	72423	40277	14178	21632	185	39254	31178	38411	55408	54627	16203
72424	72424	55592	65946	18943	54272	43143	29263	31054	64383	43791	62878

	clnt_id	recommend_items
0	9	[Trash Bags, Cookie Cakes, Ramens, Imported Be...
1	12	[Dried Anchovies For Parching And Boiling, Fit...
2	20	[Fixed-price Living Products, Chilled Noodles,...
3	23	[Tofu, Sausages, Ramens, Flavored Milk, Genera...
4	24	[Triangle Shaped Gimbabs, Trash Bags, Bananas,...
...	...	...
7389	72340	[Spoon Type Yogurts, Fish Cakes, Chilled Noodl...
7390	72356	[Soft Drink Mixes, Bananas, Ssamjang, Tofu, On...
7391	72410	[Fresh Milk, Trash Bags, Cookies, Cheese, Fixe...
7392	72423	[Trash Bags, General Snacks, Sanitary Pads, Ra...
7393	72424	[Korean Soju, Sausages, Water, Fresh Milk, Sof...

1. clnt_id 별로 비슷한 소비를 한 K-nearest neighbors 생성
2. 이미 구매한 상품을 제외하고 유사집단에서 가장 많이 구매한 10 개의 상품을 추천