



유동인구와 카드소비

데이터를 이용한
소비자 성향 분석

Contents



1. 활용데이터 정의



2. 데이터 전처리



3. 탐색적 자료 분석



4. 분석 및 검증



5. 활용방안



활용데이터 정의

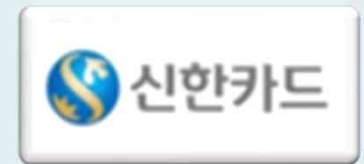
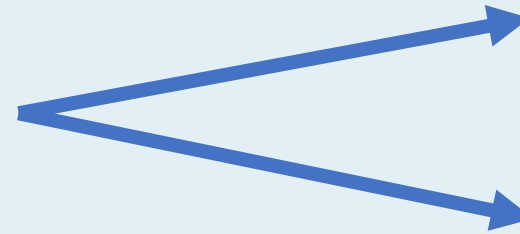


활용 데이터 수집

제7회 데이터 분석 경진대회

2019 빅콘테스트

2019.07.03 (수) ~ 2019.09.10 (화)



- 2019 빅 콘테스트 데이터 분석 경진대회에 많은 기업들이 실 데이터를 제공함.
- 그 중 SK telecom의 유동인구 데이터와 신한카드의 카드 소비 내역 데이터를 활용하였음.

카드 소비 테이블 정의서

No.	컬럼ID	컬럼명	데이터 타입
1	STD_DD	기준일자	VARCHAR(8)
2	GU_CD	구코드	VARCHAR(2)
3	DONG_CD	행정동코드	VARCHAR(3)
4	MCT_CAT_CD	업종코드	VARCHAR(2)
5	SEX_CD	성별코드	VARCHAR(1)
6	AGE_CD	나이코드	VARCHAR(2)
7	USE_CNT	이용건수	NUMBER
8	USE_AMT	이용금액	NUMBER
<ul style="list-style-type: none"> 기간 : 2018년 4월 1일 ~ 2019년 3월 31일 (12개월) 지역 및 카드범위 : 서울시 종로구/노원구, 신용/체크카드 소비 건수/금액 소비 건수/금액은 보정된 금액 온라인거래(전자상거래, 홈쇼핑 등)는 제외 결측치는 존재하지 않는다. 			

카드 소비 테이블 정의서 - 행정동 코드

구코드	행정동코드	구명	행정동명
110	515	종로구	청운요자동
110	530	종로구	사직동
110	540	종로구	삼청동
110	550	종로구	부암동
110	560	종로구	평창동
110	570	종로구	무악동
110	580	종로구	교남동
110	600	종로구	가회동
110	615	종로구	종로 1,2,3,4가동
110	630	종로구	종로 5,6가동
110	640	종로구	이화동
110	650	종로구	혜화동
110	670	종로구	창신1동
110	680	종로구	창신2동
110	690	종로구	창신3동
110	700	종로구	승인1동
110	710	종로구	승인2동

구코드	행정동코드	구명	행정동명
350	560	노원구	월계1동
350	570	노원구	월계2동
350	580	노원구	월계3동
350	595	노원구	공릉1동
350	600	노원구	공릉2동
350	611	노원구	하계1동
350	612	노원구	하계2동
350	619	노원구	중계본동
350	621	노원구	중계1동
350	624	노원구	중계4동
350	625	노원구	중계2,3동
350	630	노원구	상계1동
350	640	노원구	상계2동
350	665	노원구	상계3,4동
350	670	노원구	상계5동
350	695	노원구	상계6,7동
350	700	노원구	상계8동
350	710	노원구	상계9동
350	720	노원구	상계10동

동 개수	
종로구	노원구
17개	19개

카드 소비 테이블 정의서 - 업종, 나이, 성별 코드

한국은행 자체 업종 코드(23개 분류)	
숙박(10)	직물(43)
레저용품(20)	신변잡화(44)
레저업소(21)	서적문구(50)
문화취미(22)	사무통신(52)
가구(30)	자동차판매(60)
전기(31)	자동차정비(62)
주방용구(32)	의료기관(70)
연료판매(33)	보건위생(71)
광학제품(34)	요식업소(80)
가전(35)	음료식품(81)
유통업(40)	수리서비스(92)
의복(42)	

나이코드	코드명
20	25세 미만
25	25-29세
30	30-34세
35	35-39세
40	40-44세
45	45-49세
50	50-54세
55	55-59세
60	60-64세
65	65세이상

성별코드	코드명
M	남자
F	여자

카드 소비 테이블 정의서

Index	STD_DD	GU_CD	DONG_CD	MCT_CAT_CD	SEX_CD	AGE_CD	USE_CNT	USE_AMT
0	20180401	110	515	21	F	30	4	180
1	20180401	110	515	21	F	55	4	22
2	20180401	110	515	21	M	20	35	184
3	20180401	110	515	21	M	25	70	425
4	20180401	110	515	21	M	30	18	82

•
•
•













2152957	20190331	350	720	81	M	35	9	115
2152958	20190331	350	720	81	M	40	5	16
2152959	20190331	350	720	81	M	45	23	259
2152960	20190331	350	720	81	M	50	9	80
2152961	20190331	350	720	81	M	55	9	69
2152962	20190331	350	720	81	M	65	14	129

- 날짜, 구, 동, 업종, 성별, 나이 별로 구분한 약 215만 건의 카드 소비 건수/소비 금액 데이터

유동인구 테이블 정의서

No.	컬럼ID	컬럼명	데이터 타입	No.	컬럼ID	컬럼명	데이터 타입
1	STD_YM	년월	VARCHAR2	18	MAN_FLOW_POP_CNT_6569	남성_6569세 유동인구	NUMBER
2	STD_YMD	년월일	VARCHAR2	19	MAN_FLOW_POP_CNT_70U	남성_70세이상 유동인구	NUMBER
3	BDONG_CD	법정동코드	VARCHAR2	20	WMAN_FLOW_POP_CNT_0004	여성_0004세 유동인구	NUMBER
4	BDONG_NM	법정동명칭	VARCHAR2	21	WMAN_FLOW_POP_CNT_0509	여성_0509세 유동인구	NUMBER
5	MAN_FLOW_POP_CNT_0004	남성_0004세 유동인구	NUMBER	22	WMAN_FLOW_POP_CNT_1014	여성_1014세 유동인구	NUMBER
6	MAN_FLOW_POP_CNT_0509	남성_0509세 유동인구	NUMBER	23	WMAN_FLOW_POP_CNT_1519	여성_1519세 유동인구	NUMBER
7	MAN_FLOW_POP_CNT_1014	남성_1014세 유동인구	NUMBER	24	WMAN_FLOW_POP_CNT_2024	여성_2024세 유동인구	NUMBER
8	MAN_FLOW_POP_CNT_1519	남성_1519세 유동인구	NUMBER	25	WMAN_FLOW_POP_CNT_2529	여성_2529세 유동인구	NUMBER
9	MAN_FLOW_POP_CNT_2024	남성_2024세 유동인구	NUMBER	26	WMAN_FLOW_POP_CNT_3034	여성_3034세 유동인구	NUMBER
10	MAN_FLOW_POP_CNT_2529	남성_2529세 유동인구	NUMBER	27	WMAN_FLOW_POP_CNT_3539	여성_3539세 유동인구	NUMBER
11	MAN_FLOW_POP_CNT_3034	남성_3034세 유동인구	NUMBER	28	WMAN_FLOW_POP_CNT_4044	여성_4044세 유동인구	NUMBER
12	MAN_FLOW_POP_CNT_3539	남성_3539세 유동인구	NUMBER	29	WMAN_FLOW_POP_CNT_4549	여성_4549세 유동인구	NUMBER
13	MAN_FLOW_POP_CNT_4044	남성_4044세 유동인구	NUMBER	30	WMAN_FLOW_POP_CNT_5054	여성_5054세 유동인구	NUMBER
14	MAN_FLOW_POP_CNT_4549	남성_4549세 유동인구	NUMBER	31	WMAN_FLOW_POP_CNT_5559	여성_5559세 유동인구	NUMBER
15	MAN_FLOW_POP_CNT_5054	남성_5054세 유동인구	NUMBER	32	WMAN_FLOW_POP_CNT_6064	여성_6064세 유동인구	NUMBER
16	MAN_FLOW_POP_CNT_5559	남성_5559세 유동인구	NUMBER	33	WMAN_FLOW_POP_CNT_6569	여성_6569세 유동인구	NUMBER
17	MAN_FLOW_POP_CNT_6064	남성_6064세 유동인구	NUMBER	34	WMAN_FLOW_POP_CNT_70U	여성_70세이상 유동인구	NUMBER

유동인구 테이블 정의서

 노원_종로_FLOW_AGE_201804
 노원_종로_FLOW_AGE_201805
 노원_종로_FLOW_AGE_201806
 노원_종로_FLOW_AGE_201807
 노원_종로_FLOW_AGE_201808
 노원_종로_FLOW_AGE_201809
 노원_종로_FLOW_AGE_201810
 노원_종로_FLOW_AGE_201811
 노원_종로_FLOW_AGE_201812
 노원_종로_FLOW_AGE_201901
 노원_종로_FLOW_AGE_201902
 노원_종로_FLOW_AGE_201903

- 기간 : 2018년 4월 1일 ~ 2019년 3월 31일 (12개월)
- 지역 : 서울시 종로구 / 노원구
- 유동인구는 보정된 값

날짜, 구, 동, 업종, 성별, 나이 별로 구분한 한달 유동인구 데이터
2018년 4월 ~ 2019년 3월까지 12개의 CSV 파일

Index	STD_YM	STD_YMD	HDONG_CD	HDONG_NM	MAN_FLOW_POP_CNT_0004	MAN_FLOW_POP_CNT_0509
0	201804	20180401	1111051500	청운효자동	0.05	78.93
1	201804	20180401	1111053000	사직동	1.25	262.54
2	201804	20180401	1111054000	삼청동	0	78.07
3	201804	20180401	1111055000	부암동	0	228.5
4	201804	20180401	1111056000	평창동	0	354.38
5	201804	20180401	1111057000	무악동	0.18	39.69
6	201804	20180401	1111058000	교남동	0.1	71.96
7	201804	20180401	1111060000	가회동	0	37.87
8	201804	20180401	1111061500	종로1.2.3.4가동	0	513.62
9	201804	20180401	1111063000	종로5.6가동	0	110.6

...

WMAN_FLOW_POP_CNT_6064	WMAN_FLOW_POP_CNT_6569	WMAN_FLOW_POP_CNT_70U
1212.56	549.45	811.56
4207.45	2160.84	2972.08
744.6	359.03	421.64
2223.72	1118.48	1498.65
5336.01	2629.75	3820.41
663.07	335.06	505
1005.98	511.36	659.94
496.99	239.37	318.38
11133.7	6688.62	8386.44
3341.49	2045.79	2427.11



데이터 전처리



카드 데이터 한글화

Index	STD_DD	GU_CD	DONG_CD	MCT_CAT_CD	SEX_CD	AGE_CD	USE_CNT	USE_AMT
0	20180401	110	515	21	F	30	4	180
1	20180401	110	515	21	F	55	4	22
2	20180401	110	515	21	M	20	35	184
3	20180401	110	515	21	M	25	70	425
4	20180401	110	515	21	M	30	18	82



Index	STD_DD	GU_CD	DONG_CD	MCT_CAT_CD	SEX_CD	AGE_CD	USE_CNT	USE_AMT
0	20180401	종로구	청운효자동	레저업소 (21)	여자	30세~34세	4	180
1	20180401	종로구	청운효자동	레저업소 (21)	여자	55세~59세	4	22
2	20180401	종로구	청운효자동	레저업소 (21)	남자	25세 미만	35	184
3	20180401	종로구	청운효자동	레저업소 (21)	남자	25세~29세	70	425
4	20180401	종로구	청운효자동	레저업소 (21)	남자	30세~34세	18	82

- 코드로 되어있는 데이터들을 알기 쉽고 그래프로 변형했을 때 보기 편리하도록 한글로 변형

카드 데이터 이상치(outlier) 확인

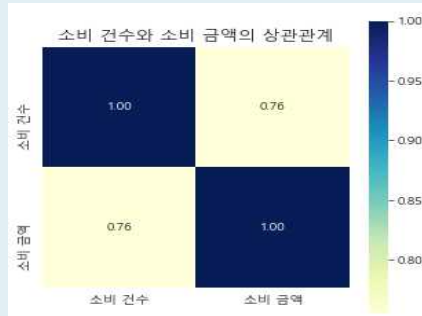
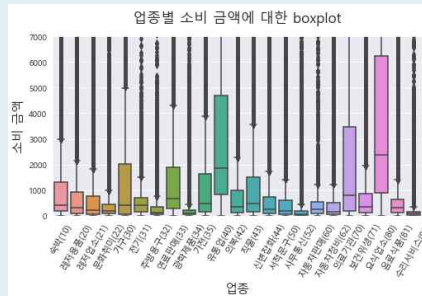
문제

업종별 판매하는 물품이 다르기 때문에 물품에 따른 가격 범위가 다르게 형성될 것으로 생각된다.

소비 건수가 많아지면 소비 금액도 많아질 것으로 생각된다.

소비 건수 or 소비 금액이 크다고 이상치로 판단하는 것을 방지한다.

문제 확인



해결방안

업종별로 나눈 데이터를 기준으로 이상치를 제거한다.

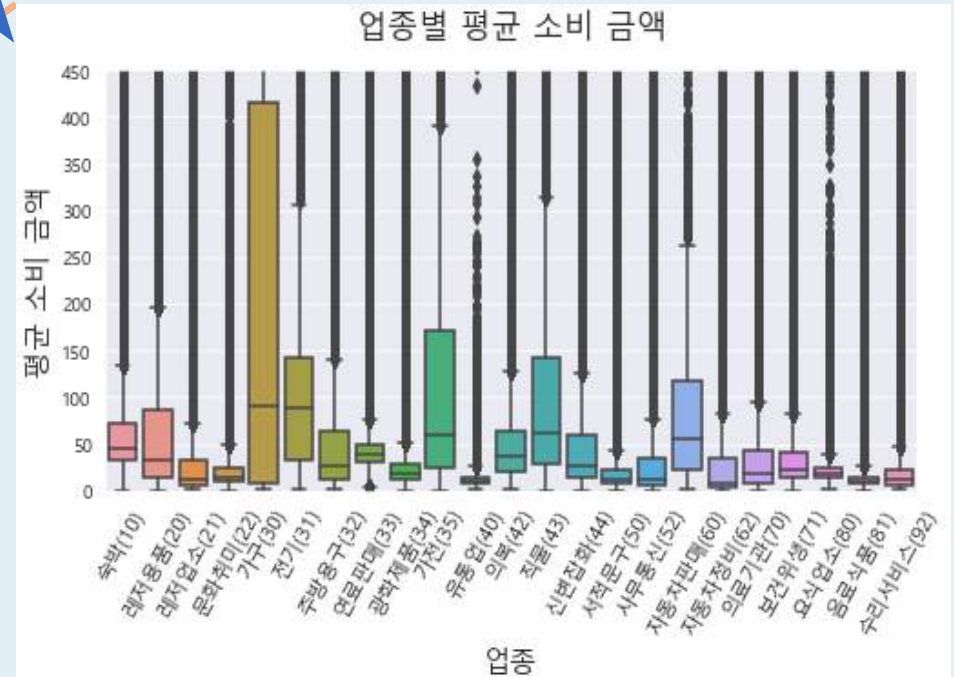
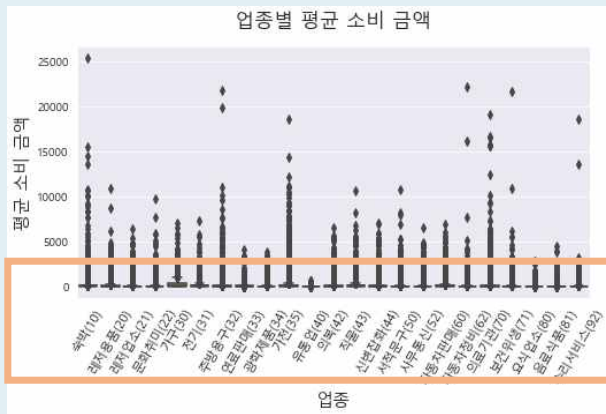
소비 금액을 소비 건수로 나누어 평균 소비 금액을 구한다.

$$\text{평균 소비 금액} = \frac{\text{소비 금액}}{\text{소비 건수}}$$

- 평균 소비 금액 column을 생성하고 데이터를 업종별로 나누어 업종별 평균 소비 금액을 기준으로 이상치를 제거

카드 데이터 이상치(outlier) 확인

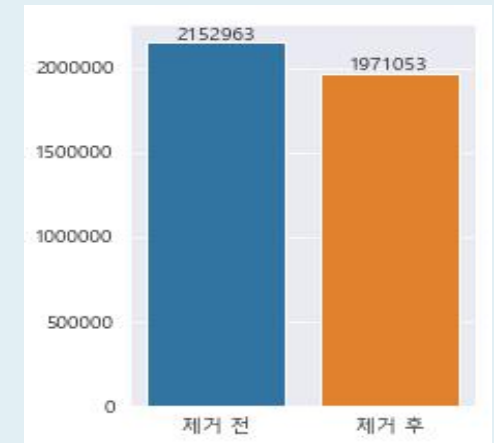
확대



- 업종별 평균 소비 금액의 이상치를 boxplot으로 확인해 본 결과 업종별로 소비 금액이 상이한 것을 확인

카드 데이터 이상치(outlier) 제거

```
card_outlier = pd.DataFrame()
def card_outlier_remove(mct_name):
    card_outlier = card_kr[card_kr["MCT_CAT_CD"] == mct_name]
    Q1 = card_outlier["USE"].quantile(q=0.25)
    Q3 = card_outlier["USE"].quantile(q=0.75)
    IQR = Q3 - Q1
    step = 1.5 * IQR
    outlier_remove = card_outlier[(card_outlier["USE"] > Q1 - step) &
                                   (card_outlier["USE"] < Q3 + step)]
    return outlier_remove
```



- IQR(Inter-Quartile Range) 방법을 사용하여 1.5배 기준으로 산정
- 약 215만 건에서 약 197만 건으로 약 10%정도의 데이터 제거

유동인구 데이터 변형

Index	STD_YM	STD_YMD	HDONG_CD	HDONG_NM	MAN_FLOW_POP_CNT_0004	MAN_FLOW_POP_CNT_0509
0	201804	20180401	1111051500	청운효자동	0.05	78.93
1	201804	20180401	1111053000	사직동	1.25	262.54
2	201804	20180401	1111054000	삼청동	0	78.07
3	201804	20180401	1111055000	부암동	0	228.5
4	201804	20180401	1111056000	평창동	0	354.38
5	201804	20180401	1111057000	무악동	0.18	39.69
6	201804	20180401	1111058000	교남동	0.1	71.96
7	201804	20180401	1111060000	가회동	0	37.87
8	201804	20180401	1111061500	종로1,2,3,4가동	0	513.62
9	201804	20180401	1111063000	종로5,6가동	0	110.6

...

WMAN_FLOW_POP_CNT_6064	WMAN_FLOW_POP_CNT_6569	WMAN_FLOW_POP_CNT_70U
1212.56	549.45	811.56
4207.45	2160.84	2972.08
744.6	359.03	421.64
2223.72	1118.48	1498.65
5336.01	2629.75	3820.41
663.07	335.06	505
1005.98	511.36	659.94
496.99	239.37	318.38
11133.7	6688.62	8386.44
3341.49	2045.79	2427.11



Index	STD_DD	DONG_CD	SEX_CD	AGE_CD	Population
0	20180401	청운효자동	남자	25세 미만	1124.9
1	20180401	청운효자동	남자	25세~29세	1792.3
2	20180401	청운효자동	남자	30세~34세	1936.1
3	20180401	청운효자동	남자	35세~39세	2060.54
4	20180401	청운효자동	남자	40세~44세	1787.98
...					
22315	20190331	상계10동	여자	45세~49세	1921.69
22316	20190331	상계10동	여자	50세~54세	1720.27
22317	20190331	상계10동	여자	55세~59세	1662.68
22318	20190331	상계10동	여자	60세~64세	1388.79
22319	20190331	상계10동	여자	65세 이상	1499.35

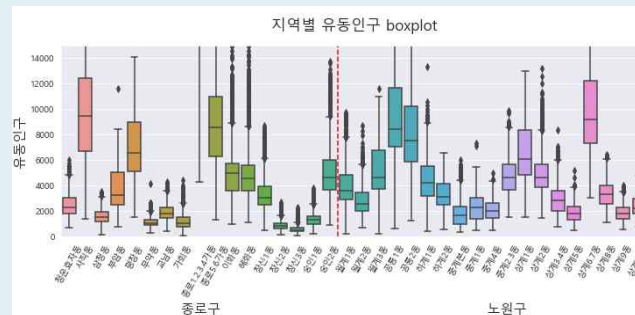
유동인구 원 데이터를 카드 데이터와 결합하기 위해 카드 데이터와 같은 형태로 변환

유동인구 데이터 이상치(outlier) 확인

문제

유동인구는 지역마다 수치가 크게 차이 날 것이라고 예상된다.

문제 확인



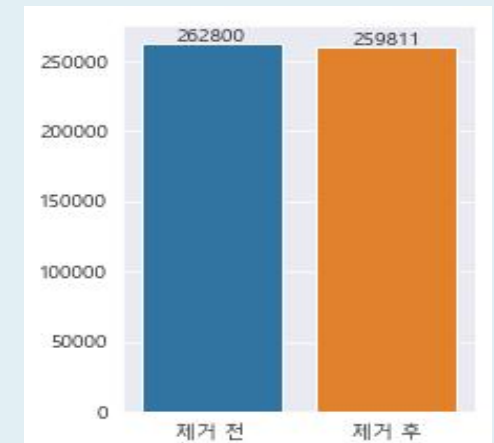
해결방안

지역별로 나눈 데이터를 기준으로 이상치를 제거한다.

- 데이터를 지역별로 나누어 지역별 유동인구를 기준으로 이상치를 제거

유동인구 데이터 이상치(outlier) 제거

```
population_outlier = pd.DataFrame()
def population_outlier_remove(dong_name):
    pop_dong = pop[pop["DONG_CD"] == dong_name]
    Q1 = pop_dong["Population"].quantile(q=0.25)
    Q3 = pop_dong["Population"].quantile(q=0.75)
    IQR = Q3 - Q1
    step = 1.5 * IQR
    outlier_remove = pop_dong[(pop_dong["Population"] > Q1 - step) &
                               (pop_dong["Population"] < Q3 + step)]
    return outlier_remove
```



- 카드 데이터와 마찬가지로 IQR(Inter-Quartile Range) 방법을 사용하여 1.5배를 기준으로 산정
- 약 26만 3천 건의 데이터가 약 26만 건으로 약 2%정도의 데이터를 제거

데이터 결합

카드 데이터

Index	STD_DD	GU_CD	DONG_CD	MCT_CAT_CD	SEX_CD	AGE_CD	USE_CNT	USE_AMT
0	20180401	종로구	청문효자동	연가압소(21)	여자	30세~34세	4	188
1	20180401	종로구	청문효자동	연가압소(21)	여자	55세~59세	4	22
2	20180401	종로구	청문효자동	연가압소(21)	남자	25세 미만	35	184
3	20180401	종로구	청문효자동	연가압소(21)	남자	25세~29세	78	425
4	20180401	종로구	청문효자동	연가압소(21)	남자	30세~34세	18	82

⋮

2152958	20190331	노원구	상계10동	음료식품(81)	남자	40세~44세	5	16
2152959	20190331	노원구	상계10동	음료식품(81)	남자	45세~49세	23	259
2152960	20190331	노원구	상계10동	음료식품(81)	남자	50세~54세	9	88
2152961	20190331	노원구	상계10동	음료식품(81)	남자	55세~59세	9	69
2152962	20190331	노원구	상계10동	음료식품(81)	남자	65세 이상	14	129



유동인구 데이터

Index	STD_DD	DONG_CD	SEX_CD	AGE_CD	Population
0	20180401	청문효자동	남자	25세 미만	1124.9
1	20180401	청문효자동	남자	25세~29세	1792.3
2	20180401	청문효자동	남자	30세~34세	1936.1
3	20180401	청문효자동	남자	35세~39세	2860.54
4	20180401	청문효자동	남자	40세~44세	1787.98

⋮

22315	20190331	상계10동	여자	45세~49세	1921.69
22316	20190331	상계10동	여자	50세~54세	1720.27
22317	20190331	상계10동	여자	55세~59세	1662.68
22318	20190331	상계10동	여자	60세~64세	1388.79
22319	20190331	상계10동	여자	65세 이상	1499.35

=

카드/유동인구 데이터

Index	STD_DD	DONG_CD	SEX_CD	AGE_CD	USE_CNT	USE_AMT	Population
0	20180401	가회동	남자	25세 미만	287	2606	770.11
1	20180401	가회동	남자	25세~29세	565	8394	893.54
2	20180401	가회동	남자	30세~34세	307	6268	937.81
3	20180401	가회동	남자	35세~39세	381	4102	1051.83
4	20180401	가회동	남자	40세~44세	294	3622	856.7

⋮

259743	20190331	혜화동	여자	45세~49세	1600	23345	2926.18
259744	20190331	혜화동	여자	50세~54세	1394	19302	2499.76
259745	20190331	혜화동	여자	55세~59세	708	12146	2413.76
259746	20190331	혜화동	여자	60세~64세	382	5426	2049.7
259747	20190331	혜화동	여자	65세 이상	348	5443	2830.4

- 데이터 전처리를 끝낸 카드 데이터와 유동인구 데이터를 결합
- 유동인구 데이터는 업종별로 분류가 되어있지 않기 때문에 카드 데이터를 업종별로 묶어준 뒤에 merge함수를 이용하여 데이터 셋을 병합

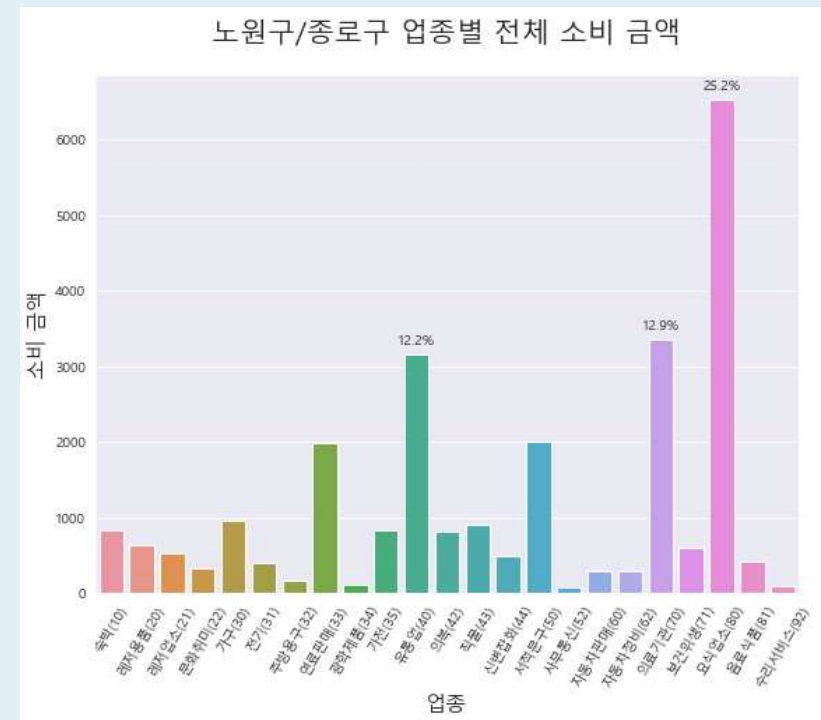
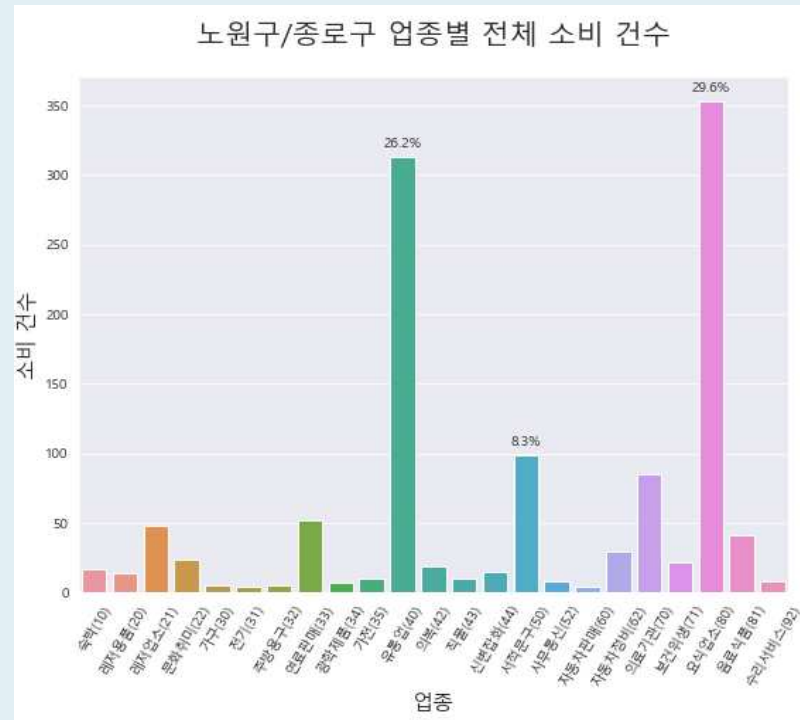


탐색적 자료 분석



업종별 소비건수/소비금액

Index	USE_CNT	USE_AMT
숙박(10)	16.94270629	844.4416216
레저용품(20)	13.60568406	645.1504148
레저업소(21)	47.87705348	523.2548038
문화취미(22)	23.48888831	330.0035936
가구(30)	5.292500747	971.2668061
전기(31)	4.810941571	403.8057313
주방용품(32)	5.414890511	175.1278832
연료판매(33)	51.84511316	1988.615612
광학제품(34)	7.248623811	123.1241072
가전(35)	9.755239755	837.0887072
유통업(40)	312.8029459	3151.004203
의복(42)	18.4658391	826.7078968
직물(43)	10.30917431	905.0197248
신변보호(44)	14.75810701	490.0085151
서적문구(50)	98.79983661	2014.692037
사무통신(52)	7.964847226	87.51772471
자동차판매(60)	4.588740458	294.0041349
자동차정비(62)	29.50752107	291.186386
의료기관(70)	85.39990471	3348.639915
보건위생(71)	22.18024324	610.8187895
요식업소(80)	352.5859992	6525.170349
음료식품(81)	40.99684434	431.247049
수리서비스(92)	8.126481068	102.4853641

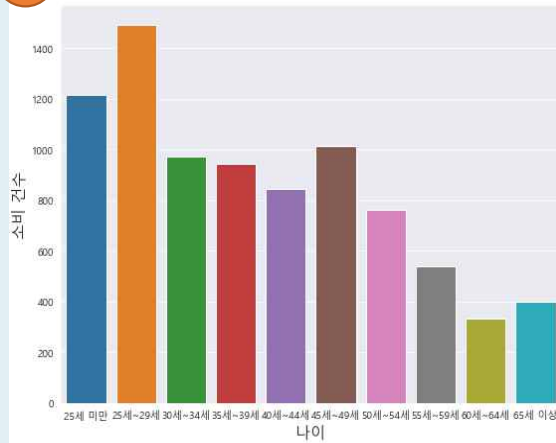


- 카드 소비 건수는 요식업소(80), 유통업(40), 서적문구(50) 순으로 높았다.
- 카드 소비 금액은 요식업소(80), 의료기관(70), 유통업(40) 순으로 높았다.
- 카드 소비의 대부분은 요식업소에 사용되는 것을 알 수 있다.

연령대별 소비 건수/소비 금액/유동인구

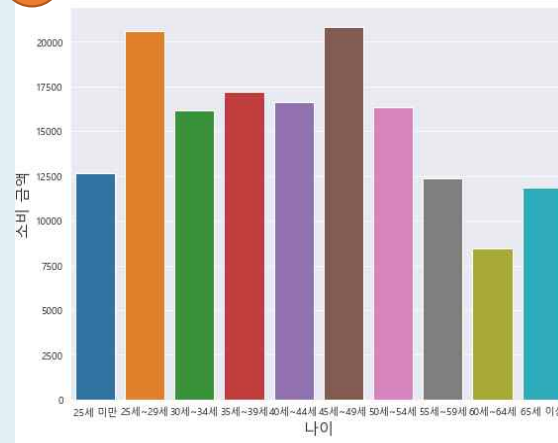
1

연령대별 전체 소비 건수



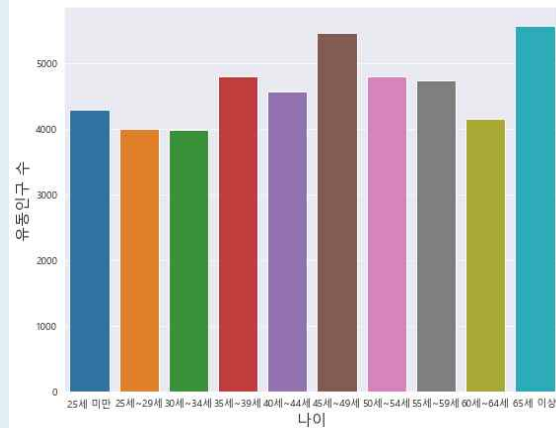
2

연령대별 전체 소비 금액



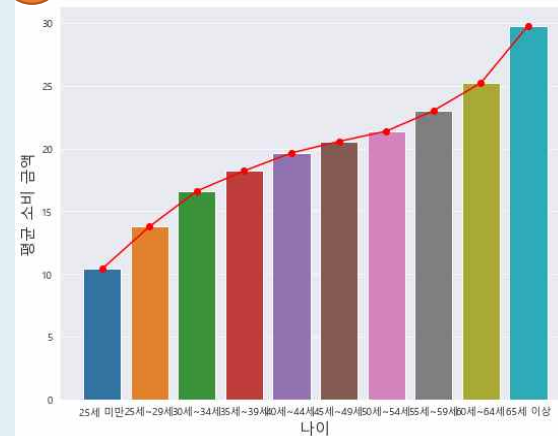
3

연령대별 유동인구



4

연령대별 평균 소비 금액



- 소비 건수가 25~29세에 가장 많았고 연령대가 높아질 수록 점점 줄어드는 것을 볼 수 있다.
- 소비 금액은 45~49세에 가장 많았 25~29세가 뒤를 이었다. 중고생과 대학생 자녀들을 둔 연령대에서 높은 소비가 이루어지는 것으로 생각된다.
- 유동인구는 65세 이상이 제일 많고 45~49세가 그 다음으로 많았다.
- 연령대가 점차 높아질 수록 평균 소비 금액이 점차 늘어나는 것을 확인 할 수 있다.

요일별 소비건수/소비금액/유동인구



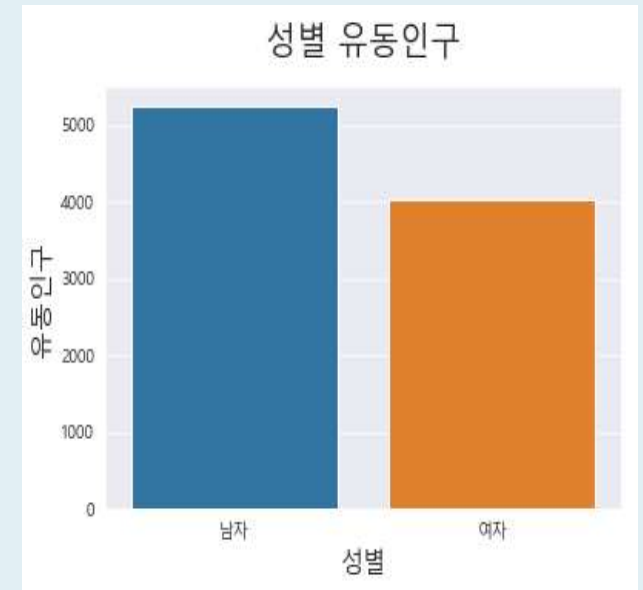
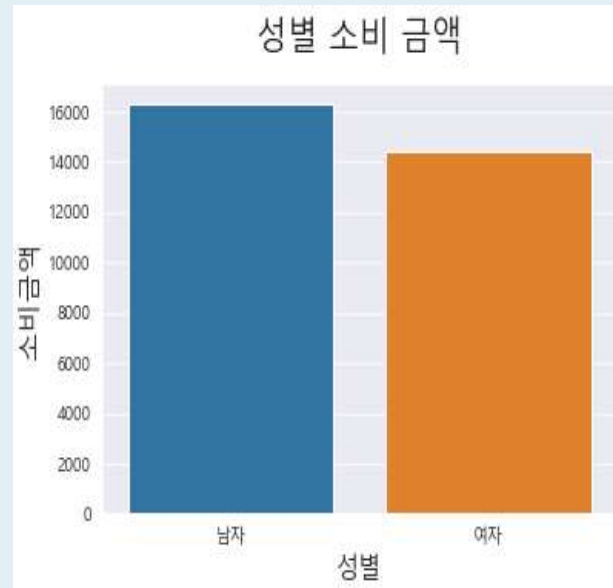
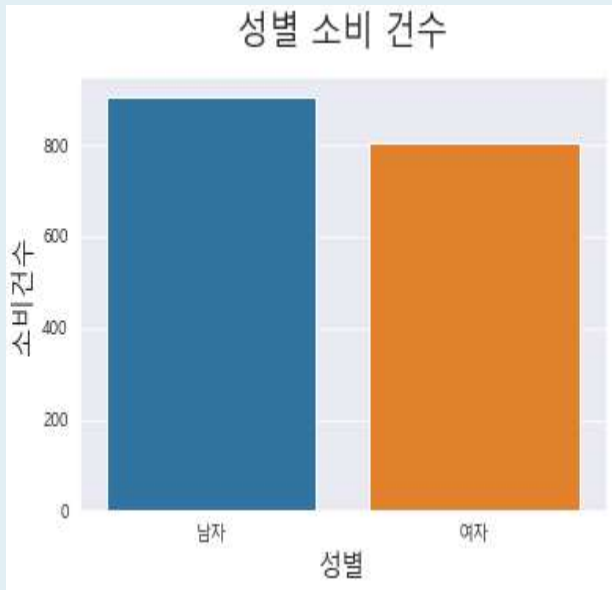
- 2018/4/1 부터 2019/3/31 까지 일요일이 하루 더 많기 때문에 2019/3/31 일요일을 제거한 뒤 그래프를 확인했다.
- 월요일에서 금요일로 갈 수록 소비 건수, 소비 금액, 유동인구가 증가한다. 금요일이 소비 금액이 가장 많고 예상외로 일요일에 소비 건수와 소비 금액, 유동인구가 가장 적은 것을 볼 수 있다.

월별 소비건수/소비금액/유동인구



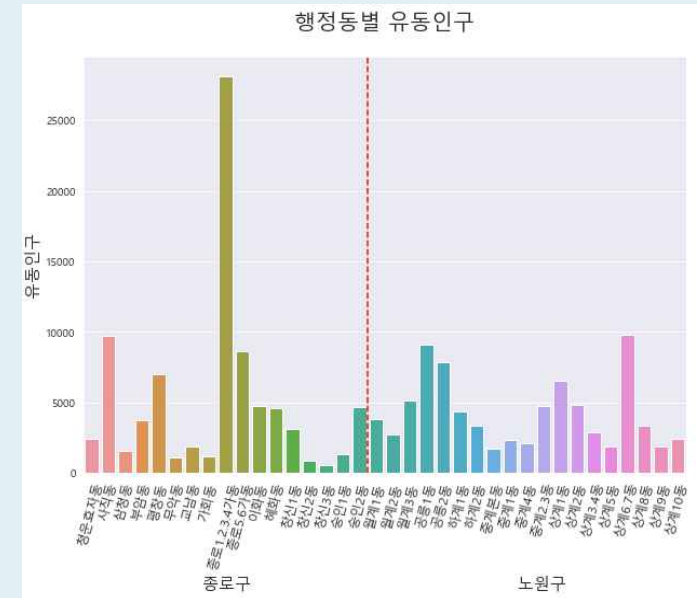
- 소비 건수, 소비 금액은 3월이 가장 많았고 유동인구는 5월이 가장 많았다.
- 날씨가 추운 겨울에는 다른 계절에 비해 유동인구가 적었다. 하지만 12월에는 연말 소비의 영향으로 소비금액이 가장 높은 것으로 생각된다.
- 월별 소비 건수, 소비 금액과 유동인구는 관계가 없어 보이고 12월을 제외한 나머지 월들은 소비 건수와 소비 금액 간의 관계가 있는 것으로 보여진다.

남녀 소비금액/소비건수



- 소비 건수, 소비 금액, 유동인구는 여성보다 남성이 조금 더 많은 것으로 나타났다.

지역별 소비건수/소비금액/유동인구



- 종로구에서는 종로 1,2,3,4동이 압도적으로 소비 건수, 소비 금액, 유동 인구가 가장 많았고 노원구에서는 상계2동이 소비 건수, 소비 금액이 가장 많았고 상계 6, 7동에서 유동인구가 가장 많았다.

종로구/노원구 요식업 시장 니즈 파악

구명	행정동명	성별	주 소비 연령대
종로구	청운효자동	여자	25세 ~ 29세
종로구	사직동	여자	25세 ~ 29세
종로구	삼청동	여자	25세 ~ 29세
종로구	부암동	여자	25세 미만
종로구	평창동	남자	45세 ~ 49세
종로구	무악동	남자	45세 ~ 49세
종로구	교남동	남자	45세 ~ 49세
종로구	가회동	여자	25세 ~ 29세
종로구	종로 1,2,3,4가동	남자	25세 ~ 29세
종로구	종로 5,6가동	남자	45세 ~ 49세
종로구	이화동	여자	25세 ~ 29세
종로구	혜화동	남자	25세 ~ 29세
종로구	창신1동	남자	25세 ~ 29세
종로구	창신2동	남자	45세 ~ 49세
종로구	창신3동	남자	25세 ~ 29세
종로구	승인1동	남자	45세 ~ 49세
종로구	승인2동	남자	25세 ~ 29세

구명	행정동명	성별	주 소비 연령대
노원구	월계1동	남자	25세 미만
노원구	월계2동	남자	25세 미만
노원구	월계3동	남자	35세 ~ 39세
노원구	공릉1동	남자	25세 ~ 29세
노원구	공릉2동	남자	25세 ~ 29세
노원구	하계1동	남자	45세 ~ 49세
노원구	하계2동	남자	45세 ~ 49세
노원구	중계본동	여자	45세 ~ 49세
노원구	중계1동	여자	45세 ~ 49세
노원구	중계4동	남자	45세 ~ 49세
노원구	중계2,3동	여자	45세 ~ 49세
노원구	상계1동	남자	45세 ~ 49세
노원구	상계2동	남자	25세 ~ 29세
노원구	상계3,4동	남자	45세 ~ 49세
노원구	상계5동	여자	45세 ~ 49세
노원구	상계6,7동	남자	25세 ~ 29세
노원구	상계8동	남자	35세 ~ 39세
노원구	상계9동	여자	45세 ~ 49세
노원구	상계10동	여자	25세 미만

- 앞선 업종별 분석에서 가장 많은 소비가 이루어지는 업종은 요식업소에서의 지출이었다.
- 소비와 유동인구가 가장 많았던 종로 1,2,3,4가동의 요식업소에서 남자 25세 ~ 29세 층이 가장 많은 소비를 한 것으로 나타났다.
- 각 동마다 요식업소에서 가장 많은 지출을 하는 성별, 연령을 참고하여 시장 니즈를 파악하고 마케팅 전략을 세운다면 확실한 근거를 가지고 마케팅에 성공할 수 있을 것이다.

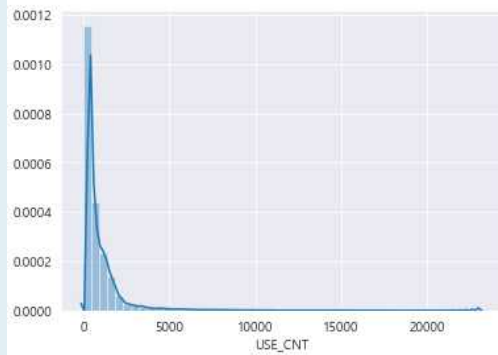


분석 및 검증

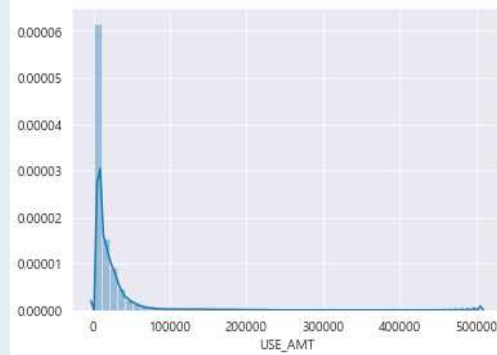


데이터 범주화

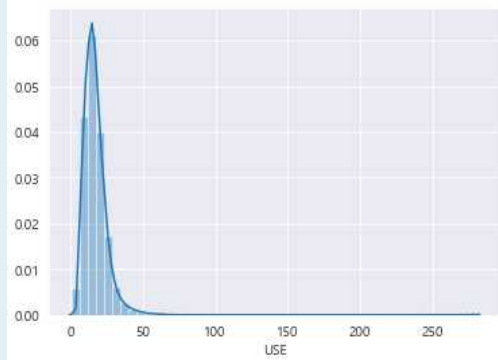
USE_CNT의 데이터 분포도



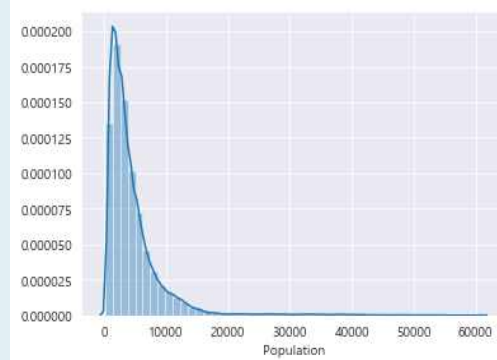
USE_AMT의 데이터 분포도



USE의 데이터 분포도



Population의 데이터 분포도



- 분석에 앞서 다양한 범위에 걸쳐 분포해 있는 연속형 변수들을 범주화하는 작업이 필요하다.
- `display`와 기초 통계량을 확인해 본 결과 `USE_CNT`의 대부분의 데이터는 5000이하에 분포해 있다.
- 그렇기 때문에 5000 미만의 데이터는 분포에 따라 데이터를 잘라내는 `qcut`을 사용하여 1~5까지 범주로 분류하고 5000 이상의 데이터는 6으로 분류한다.
- 나머지 `USE_AMT`, `USE`, `Population`도 마찬가지로 100000, 70, 20000의 기준으로 범주화 적용한다.

더미변수 생성

지역별

가회동	공릉1동
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0

성별

남자	여자
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0

연령별

25세 미만	25세~29세
1	0
0	1
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0

요일별

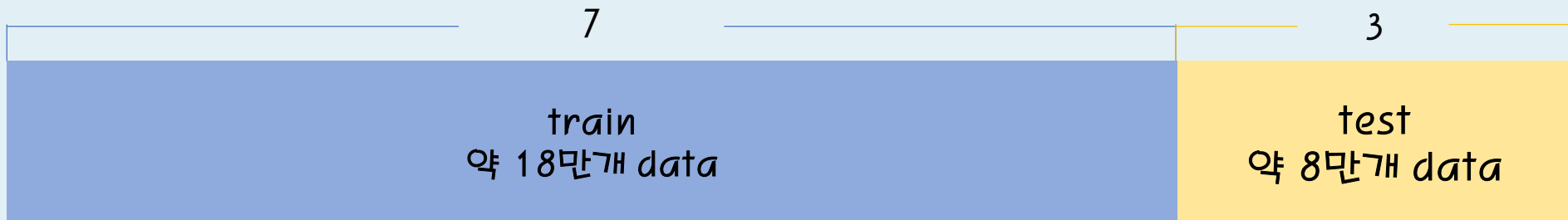
week_일	week_토
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0

월별

month_3월	month_4월
0	1
0	1
0	1
0	1
0	1
0	1
0	1
0	1
0	1
0	1

- 지역별, 성별, 나이별, 요일별, 월별에 따른 소비, 유동인구 성향을 파악하기 위해 더미변수를 생성한다.

데이터 셋 분리



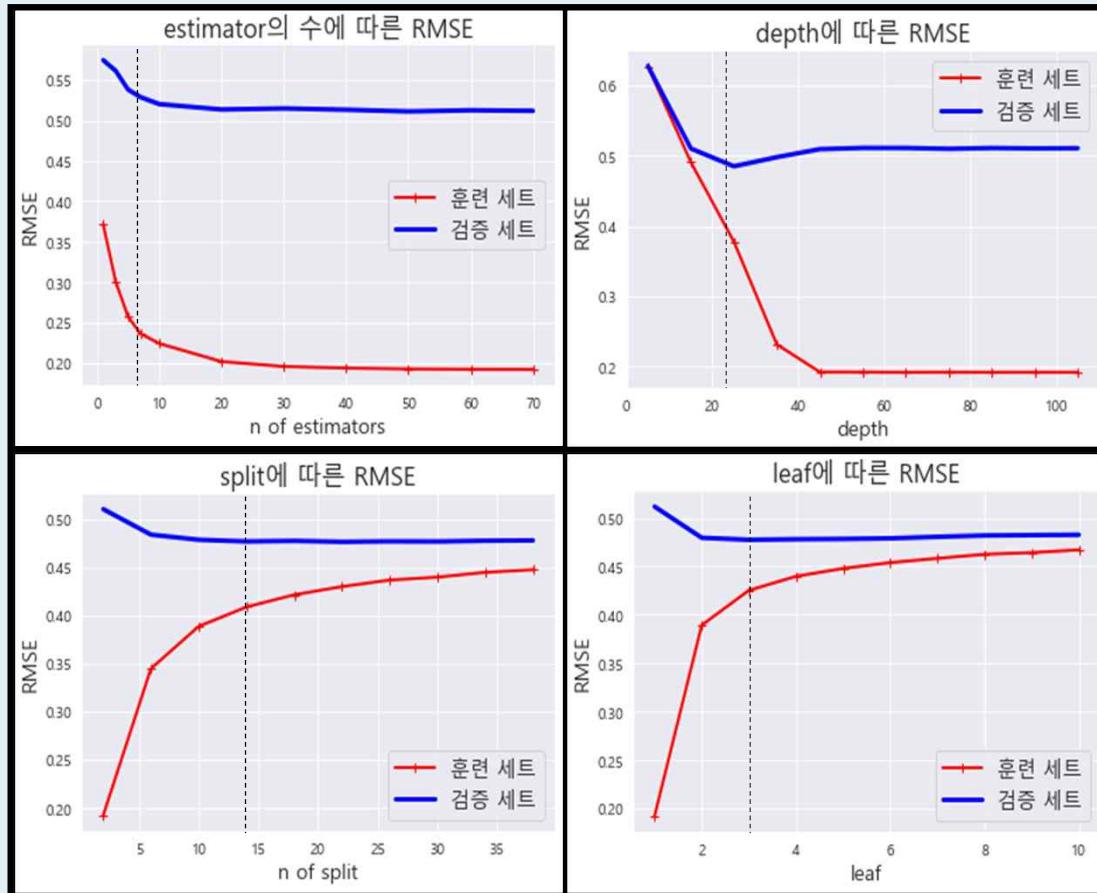
- 지역별, 성별, 나이별, 요일별, 월별에 따른 소비 건수와 유동인구가 소비 금액에 영향을 미치는지 알아보는 것이 분석 목표이므로 독립변수(소비건수, 유동인구)와 종속변수(소비 금액)를 설정한다.
- 전체 data set을 7:3 비율로 분리하여 모델을 생성한다.

분석 기법 선택 과정

분석 기법	정확도
Logistic Regression	0.7396
Linear Discriminant Analysis(LDA)	0.6854
Decision Tree	0.7167
Extra Tree	0.7402
Random Forest	0.7529
Gaussian Naive Bayes	0.4208
KNN Neighbors	0.7389

- 여러 분석 기법들로 모델의 정확도를 확인해 본 결과 tree류의 분석 기법이 높은 정확도를 나타내는 것을 볼 수 있다. 그 중에서도 가장 성능이 좋은 모델은 Random Forest이고 정확도가 약 0.753인 것을 확인 할 수 있다.

하이퍼 파라미터 조정



- 각 하이퍼 파라미터의 값에 따른 train set과 test set의 오차 값을 그래프로 나타내었다. 과적합을 방지하기 위해 적절한 하이퍼 파라미터를 선택하도록 한다.
- 검은 점선을 기준으로 좌측으로 갈수록 과소적합, 우측으로 갈수록 과대적합으로 볼 수 있다.
- `n_estimator`는 5, `max_depth`는 25, `min_samples_split`는 15, `min_samples_leaf`는 3 정도에서 적합하다고 볼 수 있다.

최적의 하이퍼 파라미터 선정

```
n_estimators = [1, 2, 3, 4, 5]
max_depth = [10, 15, 20, 25, 30]
min_samples_split = [5, 10, 15, 20, 25]
min_samples_leaf = [1, 2, 3, 4, 5]

hyperRF = dict(n_estimators = n_estimators, max_depth = max_depth,
               min_samples_split = min_samples_split,
               min_samples_leaf = min_samples_leaf)

gridRF = GridSearchCV(model, hyperRF, error_score='accuracy', cv = 3,
                      verbose = 1, n_jobs = -1)

gridRF.fit(x_train, y_train)
```

```
In [13]: gridRF.best_params_
Out[13]:
{'max_depth': 30,
 'min_samples_leaf': 4,
 'min_samples_split': 25,
 'n_estimators': 5}
```

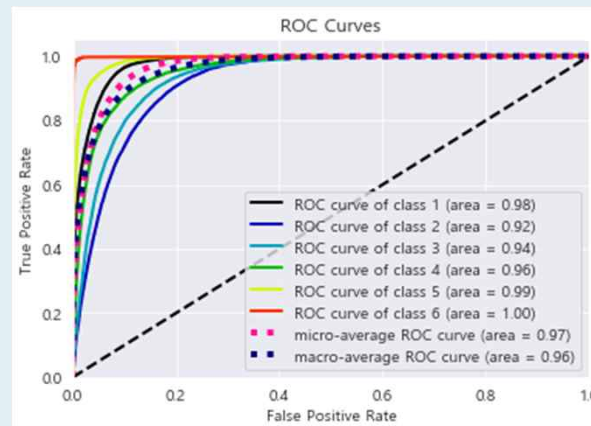
- 이전에 얻은 정보로 하이퍼 파라미터의 주변 값들을 이용해 최적의 하이퍼 파라미터를 찾아내고 모델을 생성한다.
- GridSearchCV으로 각 파라미터들 대입하여 경우의 수를 파악해 점수가 가장 좋은 하이퍼 파라미터를 찾는다.
- n_estimator는 5, max_depth는 30, min_samples_split는 25, min_samples_leaf는 4 일 때 가장 적합한 모델이라고 볼 수 있다.

분석 검증

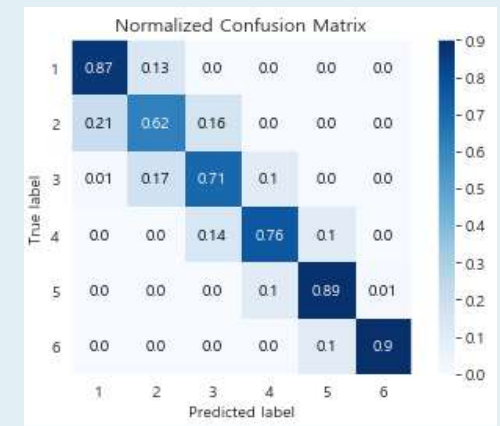
Classification Report

	precision	recall	f1-score	support
1	0.80	0.87	0.83	15268
2	0.67	0.62	0.65	15126
3	0.70	0.71	0.71	15304
4	0.79	0.76	0.77	15313
5	0.89	0.89	0.89	15299
6	0.95	0.90	0.92	1615
accuracy			0.78	77925
macro avg	0.80	0.79	0.80	77925
weighted avg	0.77	0.78	0.77	77925

ROC Curve



Confusion Matrix



Classification Report로 분석 검증을 해 본 결과 5, 6 class에서는 매우 잘 분류하였지만 2, 3 class에서는 다른 class보다 분류 성능이 약간 떨어지는 것을 확인 할 수 있다.

ROC Curve에서는 class 1, 5, 6 98% 이상의 면적을 나타냄으로 높은 분류 성능을 가진다.

Confusion Matrix에서도 class 1, 5, 6에서 87% 이상의 정확도가 나타나는 것을 볼 수 있다.



활용 방안



다양한 카드 상품 개발



건강 카드

의료기관 소비가 많은
65세 이상 남성과 40세 ~
44세 여성을 대상으로 홍보



쇼핑 카드

유통업 소비가 많은
25세 ~ 29세 남성과 45세 ~
49세 여성을 대상으로 홍보



주유 카드

주유 소비가 많은 45세 ~
59세 남성을 대상으로 홍보

요식업 창업 위치/업종 선정

구명	행정동명	주 소비 연령	업종
노원구	중계본동	45 ~ 49세 여성	맛, 가격, 서비스 좋은 카페, 음식점, 주점
	중계 1동		
	중계 2,3동		
	상계 9동		

구명	행정동명	주 소비 연령	업종
종로구	평창동	45 ~ 49세 남성	가성비 좋은 일반 음식점, 주점
	무악동		
	교남동		
	종로 5,6가동		
	창신 2동		
	승인 1동		
노원구	하계 1동		
	하계 2동		
	중계 4동		
	상계 1동		
	상계 3,4동		

구명	행정동명	주 소비 연령	업종
종로구	청운요자동	30세 미만 여성	트렌디한 분위기의 카페, 음식점, 주점
	사직동		
	삼청동		
	부암동		
	가회동		
	이화동		
노원구	상계 10동		

주 소비 연령이 뚜렷하게 구분되는 지역을 대상으로
요식업 창업에 대한 가이드를 마련할 수 있다.

감사합니다.

