

제 5회 L.POINT Big Data Competition 분석 보고서

최서현

INDEX

I. 제안하는 서비스

- ① 서비스의 자세한 과정
- ② 서비스가 실시되는 예시
- ③ 서비스의 기대 효과

II. 데이터 전처리

- ① 데이터 탐색
- ② 데이터 전처리

III. 예측 모형

- ① 모델링
- ② 결과

IV. 선호지수 개발

- ① “브랜드 선호도“
- ② 모든 상품군의 브랜드 선호도 구해보기
- ③ 브랜드 선호도의 유용성

V. 인사이트

- ① 모델의 변수 중요도를 통해 봤을 때
- ② 브랜드 선호도를 통해 봤을 때

A dimly lit clothing store interior. In the foreground, there's a long wooden display case on the left and a round wooden table in the center. To the right, there are racks of clothes and a display featuring a model boat. The background shows more clothing racks and a wall with a grid of small items. The lighting is soft, with several pendant lights hanging from the ceiling.

I . 제안하는 서비스

“고객이 다음 달에 구매할 상품을 예측하여
고객별 자동 맞춤형 마케팅을 실시”

① 서비스의 자세한 과정

단계 1. 각 고객들이 다음 달에 어떤 상품군을 구매할 지 예측

단계 2. 해당 상품군이 브랜드 선호도가 높은 상품군인지 평가

단계 3. 브랜드 선호도를 고려하여 맞춤형 마케팅 진행

※ 브랜드 선호도란 제가 이번 프로젝트에서 개발한 선호 지수로써, 이 수치가 높을수록 고객들은 여러 브랜드 제품들을 찾아보기 이전에 이미 뚜렷히 선호하는 브랜드가 있고 이를 주로 구매하는 경향이 강한 상품군임을 의미합니다. 고객의 마음 속에 이미 사고자 하는 브랜드가 분명하게 자리잡고 있으면 다른 브랜드의 상품을 추천해주는 방법은 비교적 효과가 떨어집니다. 이런 경우에는 해당 브랜드 제품에 2+1 프로모션을 제안하여 더욱 큰 소비를 이끌어내는 등 다른 방식의 마케팅을 적용하는 것이 더욱 효과적입니다. 이처럼 브랜드 선호도는 마케팅 방법을 구상하는 데에 중요한 역할을 차지합니다. 특히 이런 브랜드 선호도는 상품군 별로 유의미한 차이를 보이고 있습니다. 자세한 소개와 개발 과정은 IV. 선호 지수 개발 파트에서 설명하고 있습니다.

I. 제안하는 서비스

② 서비스가 실시되는 예시

	고객 A	고객 B
단계 1	해당 고객이 다음 달에 스포츠 패션을 구매할 것으로 예측	해당 고객이 다음 달에 화장품/뷰티케어를 구매할 것으로 예측
단계 2	스포츠 패션은 비교적 낮은 브랜드 선호도를 보임 즉, 나이키만을 선호하거나 아디다스만을 선호하는 식의 고객들이 적고, 다양한 브랜드의 상품을 둘러보고 소비를 결정하는 고객들이 많음	화장품/뷰티케어는 비교적 높은 브랜드 선호도를 보임 즉, 설화수 상품만을 선호하거나 로레알 상품만을 선호하는 식의 고객들이 많음
단계 3	고객이 구매할 예정이었던 스포츠 패션 상품군에 대해 다양한 브랜드 제품들을 일목요연하게 정리하여 고객이 무엇을 구매할 지 찾아보기 전에 이메일, 팝업 메시지 등을 통해 미리 추천해줌	고객이 구매할 예정이었던 화장품/뷰티케어 상품군에 대해 동일 브랜 드 제품을 2 개 이상 구매 시 할인 혜택, 혹은 2+1 등의 혜택 등을 제공

I. 제안하는 서비스

③ 서비스의 기대효과

- 1) 구매할 예정이었던 상품군의 여러 제품들을 먼저 추천해줌으로써 고객의 소비 만족도를 높임
- 2) 구매할 예정이었던 상품군에 2+1 등의 각종 맞춤형 프로모션을 제공함으로써 고객의 소비 만족도도 높이고 더 높은 수요도 유도할 수 있음
- 3) 별도의 장비나 다수의 인력이 필요 없는 자동 서비스로써 비용이 적고 데이터가 쌓일 수록 더욱 성능이 정교해짐



회사의 매출을 극대화

A dimly lit clothing store interior. In the center, a round wooden table holds several folded items. To the left, a long display case with glass tops contains various accessories. The background features racks of clothes and a wall with a grid of small, colorful items. On the right, a large display features a blue boat model and more clothing. The floor is made of light-colored wood, and the ceiling has several pendant lights hanging from it.

II. 데이터 전처리

II. 데이터 전처리

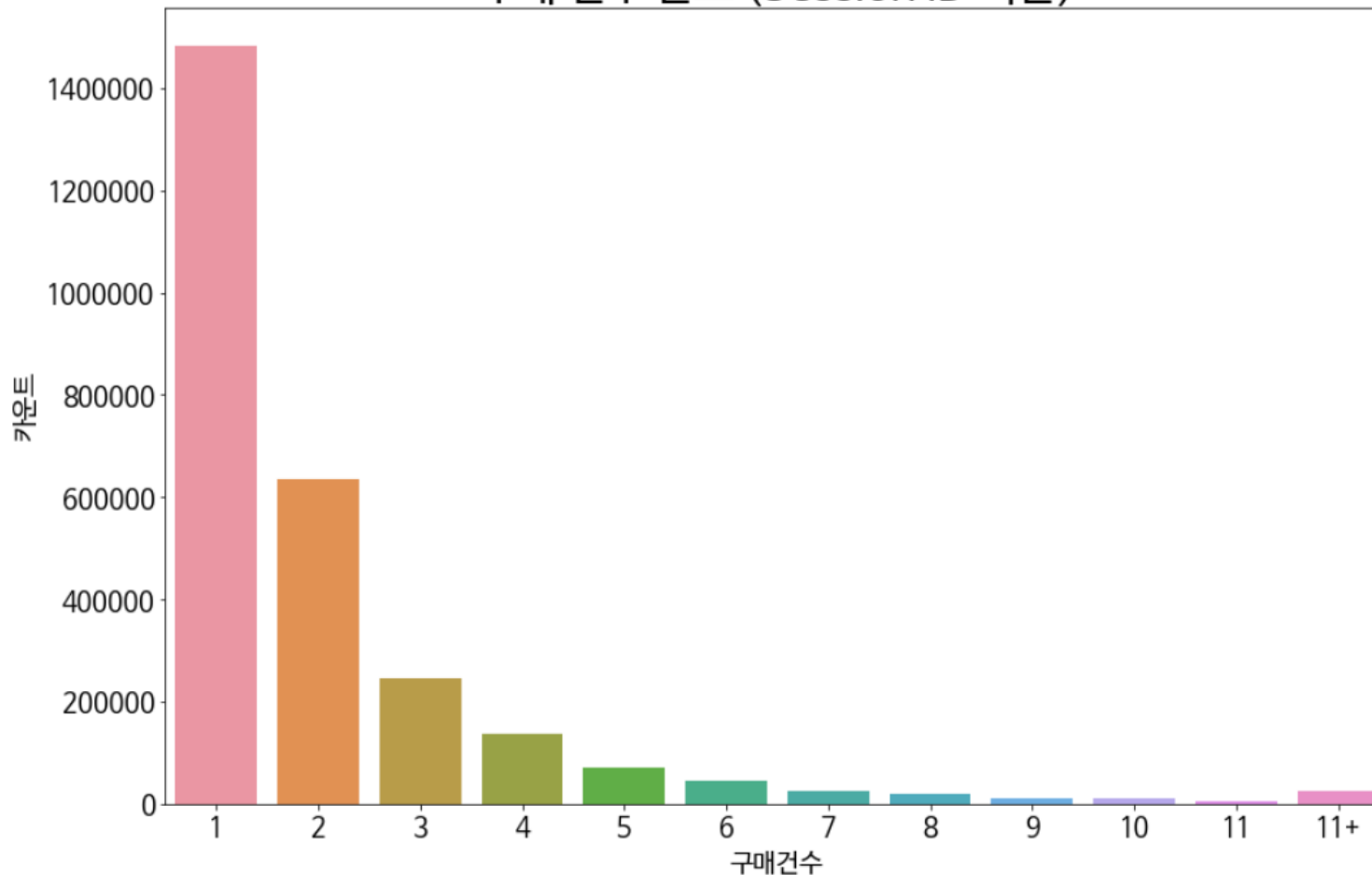
① 데이터 탐색 - 1) 변수 별 기본 정보 확인

Product	Search1	Search2	Session	Master	Custom
<ul style="list-style-type: none"> - 테이블 크기 5,024,906 * 8 - 자료 형태 <ol style="list-style-type: none"> 1) CLNT_ID : int64 2) SESS_ID : int64 3) HITS_SEQ : int64 4) PD_C int64 5) PD_ADD_NM : object 6) PD_BRA_NM : object 7) PD_BUY_AM : object 8) PD_BUY_CT : object - 기타 정보 <ol style="list-style-type: none"> 1) CLNT ID 고유 갯수 : 922,737 개 2) SESS ID 고유 개수 : 2,425,886 개 3) 브랜드 고유 개수 : 22,716 개 4) 히트 넘버 고유 개수 500 개 5) 상품 코드 고유 개수 : 847,652 개 	<ul style="list-style-type: none"> - 테이블 크기 2,884,943 * 4 - 자료 형태 <ol style="list-style-type: none"> 1) CLNT_ID : int64 2) SESS_ID : int64 3) KWD_NM : object 4) SEARCH_CNT : int64 - 기타 정보 <ol style="list-style-type: none"> 1) SEARCH_CNT의 통계량 평균 : 1.91 표준편차 : 2.34 최소값 : 1 중간값 : 1 최대값 : 98 2) 키워드 고유값 개수 : 81,539개 	<ul style="list-style-type: none"> - 테이블 크기 8,051,172 * 3 - 자료 형태 <ol style="list-style-type: none"> 1) SESS_DT : int64 2) KWD_NM : object 3) SEARCH_CNT : object - 기타 정보 <ol style="list-style-type: none"> 1) 키워드 고유값 개수 : 81,539 개 	<ul style="list-style-type: none"> - 테이블 크기 2,712,907 * 9 - 자료 형태 <ol style="list-style-type: none"> 1) CLNT_ID : int64 2) SESS_ID : int64 3) SESS_SEQ : int64 4) SESS_DT : int64 5) TOT_PAG_VIEW_CT : float64 6) TOT_SESS_HR_V : object 7) DVC_CTG_NM : object 8) ZON_NM : object 9) CITY_NM : object - 기타 정보 <ol style="list-style-type: none"> 1) DVC_CTG_NM 고유 갯수 : 3 개 2) ZON_NM 고유 갯수 : 16 개 3) CITY_NM 고유 갯수 : 163 개 4) TOT_PAG_VIEW_CT 통계량 평균 : 85.16 표준편차 : 87.58 최소값 : 1 중간값 : 55 최대값 : 499 	<ul style="list-style-type: none"> - 테이블 크기 847,652 * 5 - 자료 형태 <ol style="list-style-type: none"> 1) PD_C : int64 2) PD_NM : object 3) CLAC1_NM : object 4) CLAC2_NM : object 5) CLAC3_NM : object - 기타 정보 <ol style="list-style-type: none"> 1) CLAC1_NM 고유값 개수 : 37 개 2) CLAC2_NM 고유값 개수 : 128 개 3) CLAC3_NM 고유값 개수 : 898 개 4) 상품명 고유값 개수 : 817,431개 	<ul style="list-style-type: none"> - 테이블 크기 671,679 - 자료 형태 <ol style="list-style-type: none"> 1) CLNT_ID : int64 2) CLNT_GENDER : object 3) CLNT_AGE : int64 - 기타 정보 <ol style="list-style-type: none"> 1) CLNT_GENDER 분포 F : 570,616 개 M : 101,063 개 2) 30,40대가 가장 많은 산 모양을 띠

II. 데이터 전처리

① 데이터 탐색 - 2) 구매 건수

구매 건수 분포 (Session ID 기준)

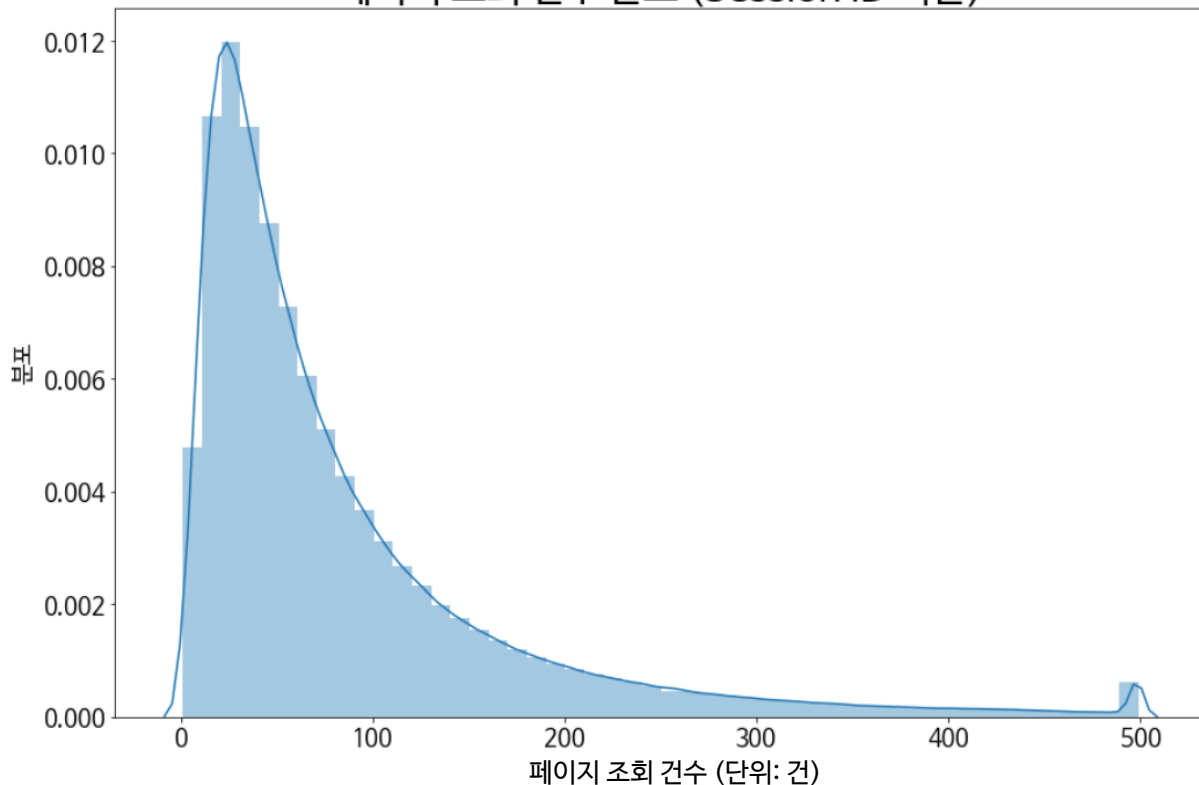


구매 건수의 분포는 왼쪽과 같다.
제품 1개를 구매하는 경우가 가장
많고 그 이후로는 빈도수가 급격히
줄어들고 있다.

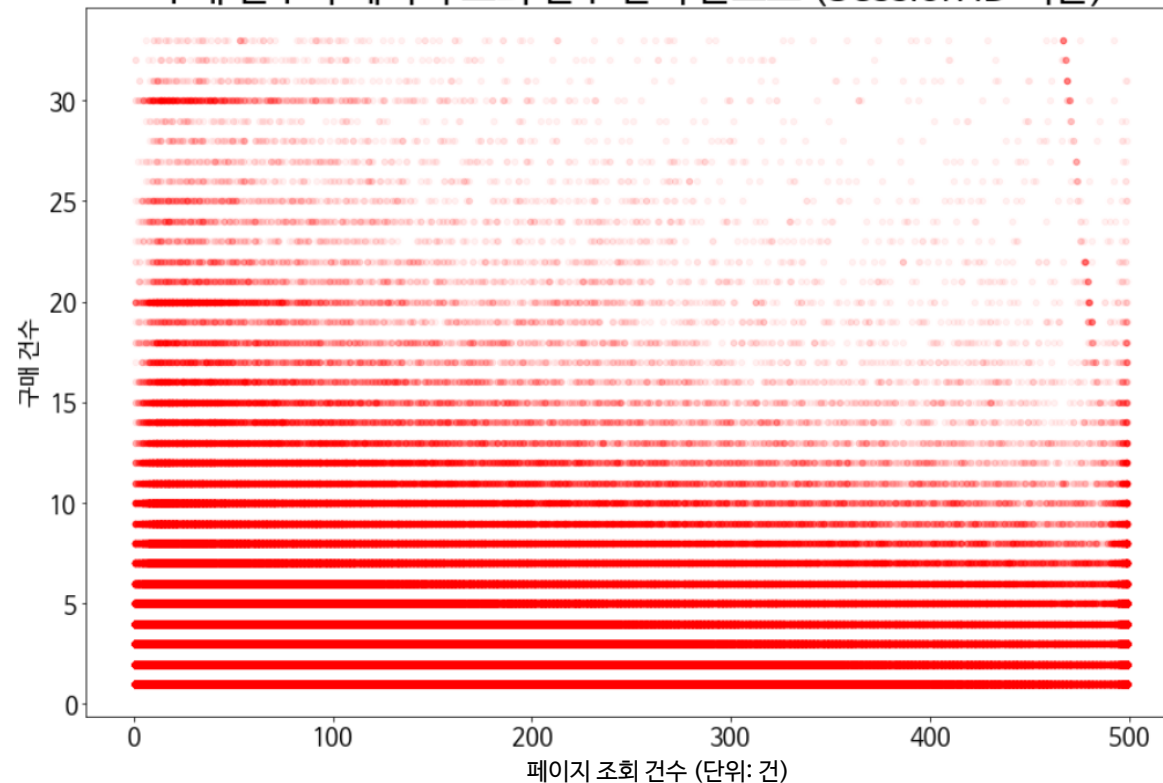
II. 데이터 전처리

① 데이터 탐색 - 3) 페이지 조회 건수

페이지 조회 건수 분포 (Session ID 기준)



구매 건수와 페이지 조회 건수 간의 분포도 (Session ID 기준)

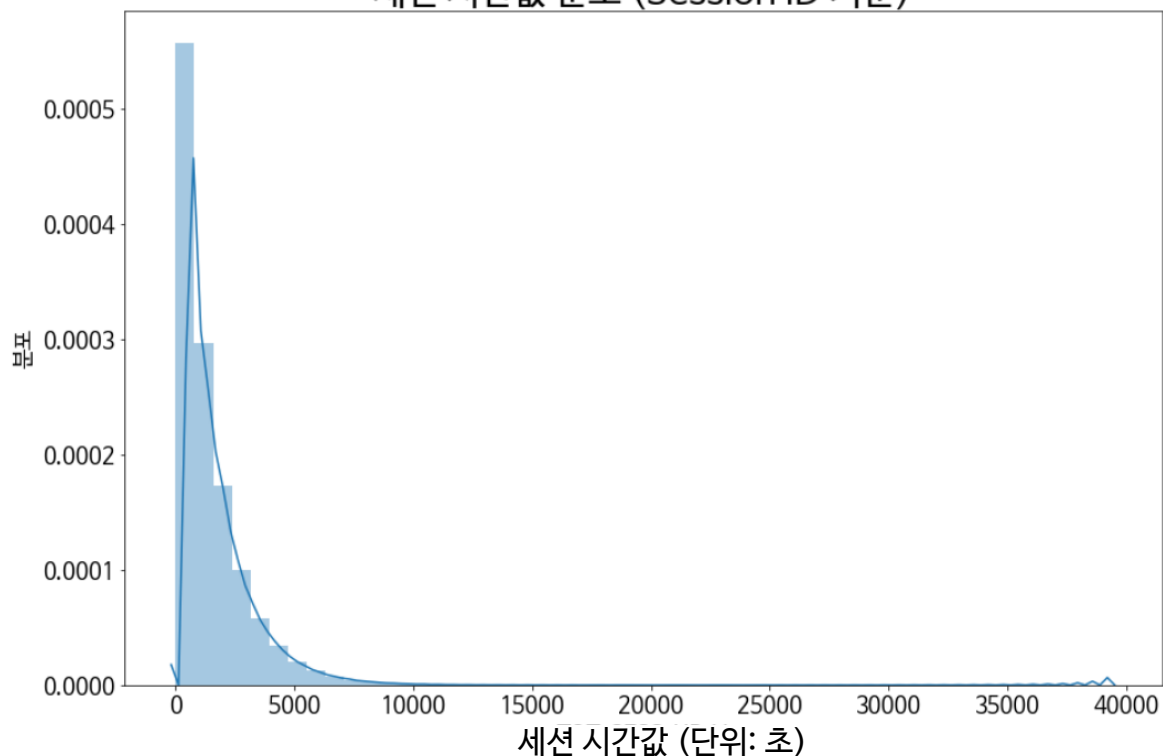


평균 페이지 조회 건수는 85 건, 상위 99%의 페이지 조회 건수는 448 건이며 최고 페이지 조회 건수는 499 건이다.
구매 건수가 높을수록 오히려 페이지 조회 건수가 낮아지는 모양새를 보이고 있다.

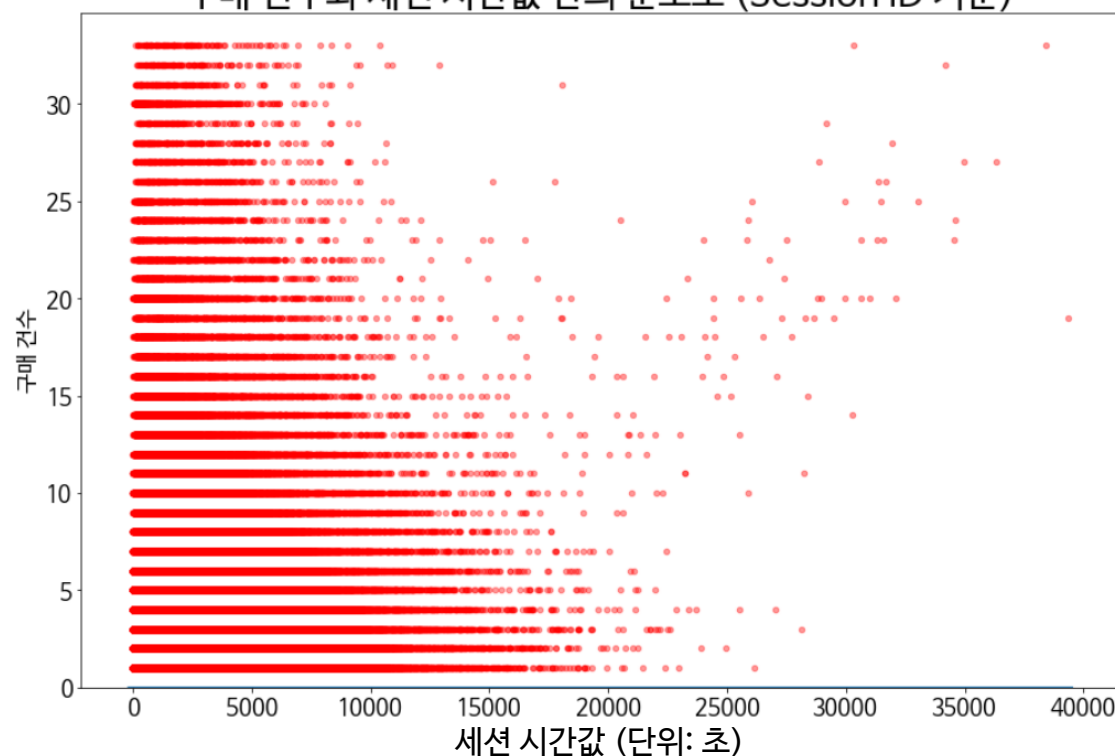
II. 데이터 전처리

① 데이터 탐색 - 4) 세션 시간 값

세션 시간값 분포 (Session ID 기준)



구매 건수와 세션 시간값 간의 분포도 (Session ID 기준)

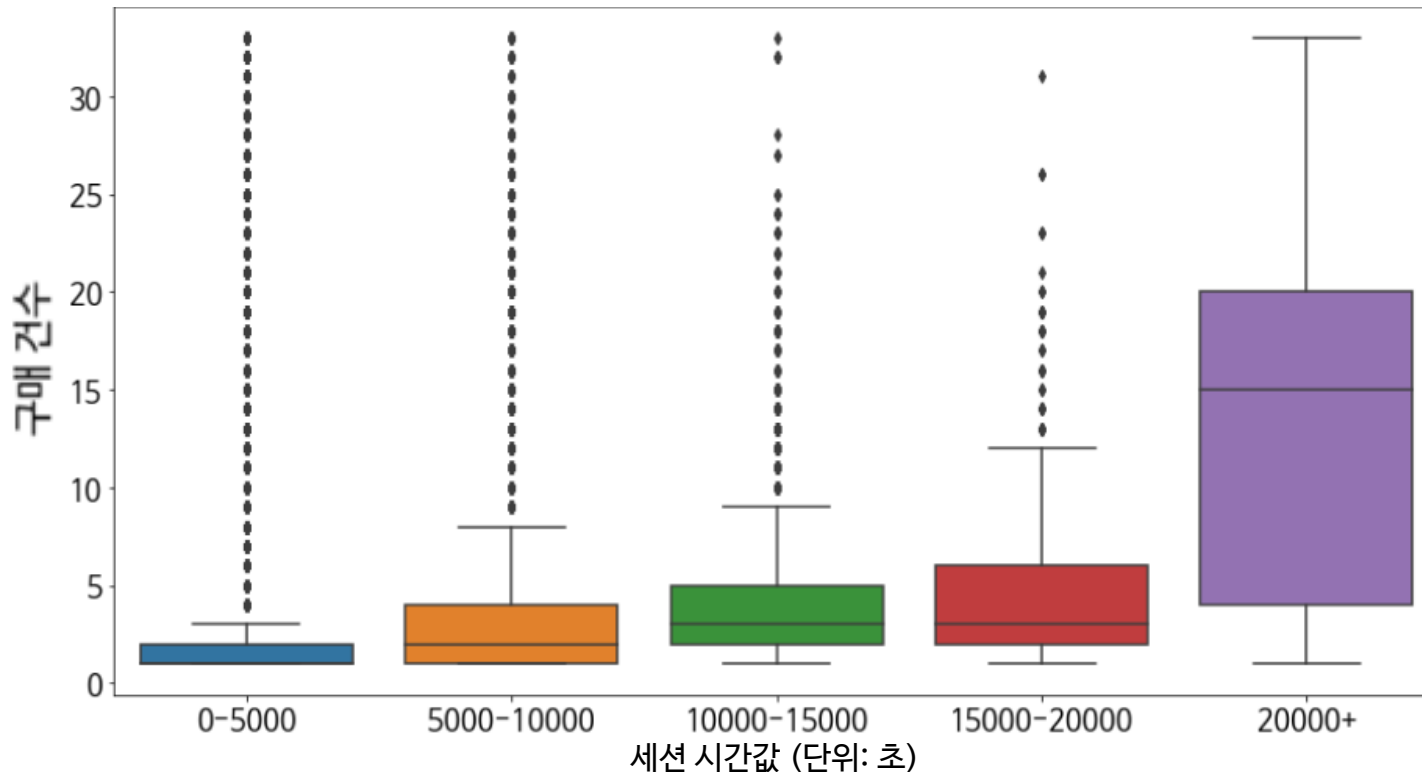


평균 세션 시간 값은 1,452 초, 상위 99%의 세션 시간 값은 6,989 초이며 최고 세션 시간 값은 39,329 초이다.
세션 시간값 또한 페이지 조회 건수와 마찬가지로 높을수록 높은 구매 건수를 보이는 고객들의 수가 줄어들고 있음을 볼 수 있다.
한편 약 20,000초 이후부터는 약한 양의 상관관계도 동시에 보이고 있다.

II. 데이터 전처리

① 데이터 탐색 - 4) 세션 시간 값

구매 건수가 50 미만인 경우, 세션 시간값 간에 따른 구매 건수 박스플롯 (Session ID 기준)

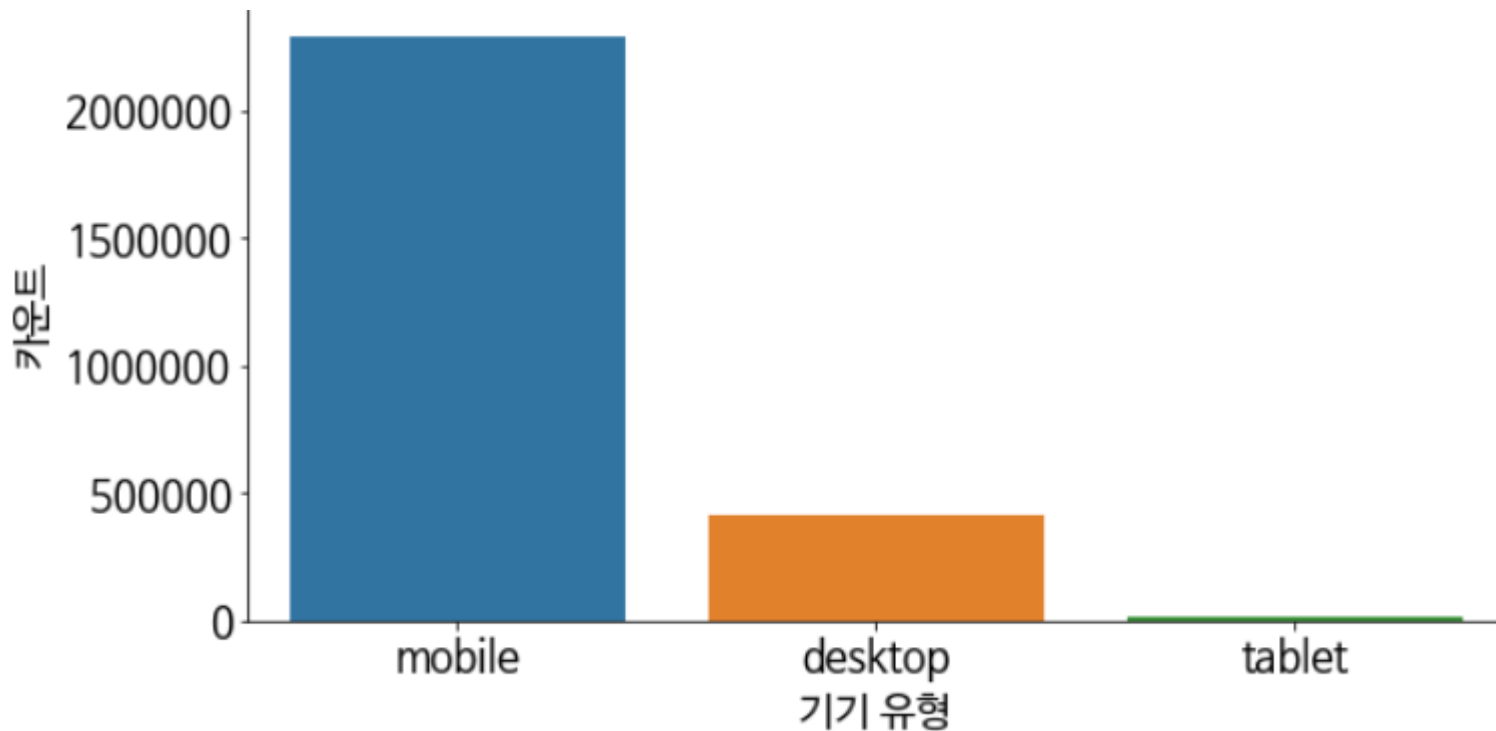


박스 플롯으로 살펴보면 살짝 양의 상관관계를 보이고 있다. 특히 20,000초 이후부터는 세션 시간값이 클수록 구매 건수도 확연히 증가하고 있다. 하지만 20,000초 이후부터는 관측치 수가 워낙 적어서 의미있는 패턴이라고는 하기 힘들 것이라고 판단된다.

II. 데이터 전처리

① 데이터 탐색 - 5) 기기 유형

기기 유형 분포(Session ID 기준)

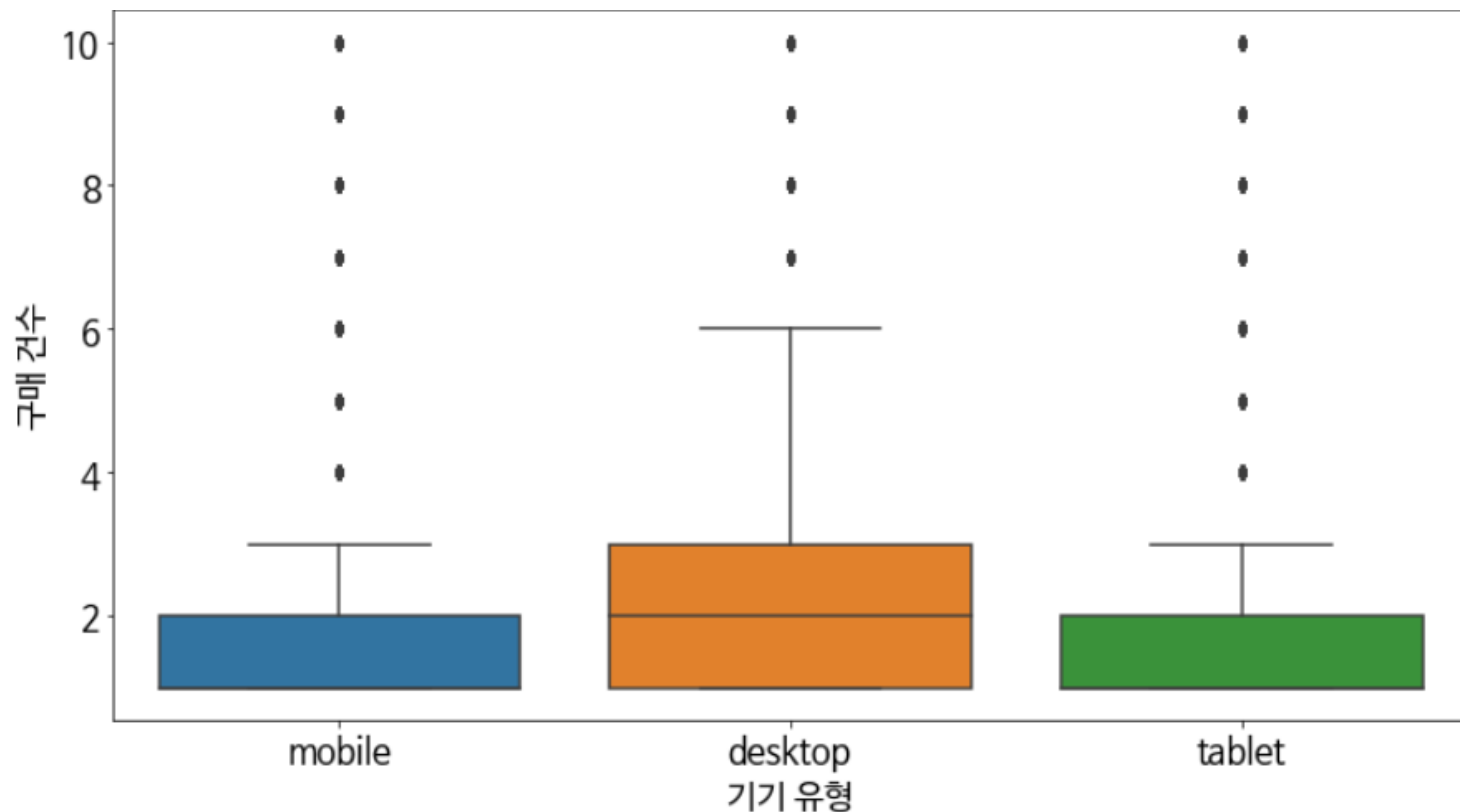


모바일 세션 로그인 수가 가장 많고 데스크탑 세션 로그인 수는 모바일의 약 1/5 정도이다. 태블릿의 수치는 거의 미미하다. 자세한 수치를 살펴보자면 모바일은 2,289,681번, 데스크탑은 411,168번, 태블릿은 12,057번의 세션 로그인이 있었다.

II. 데이터 전처리

① 데이터 탐색 - 5) 기기 유형

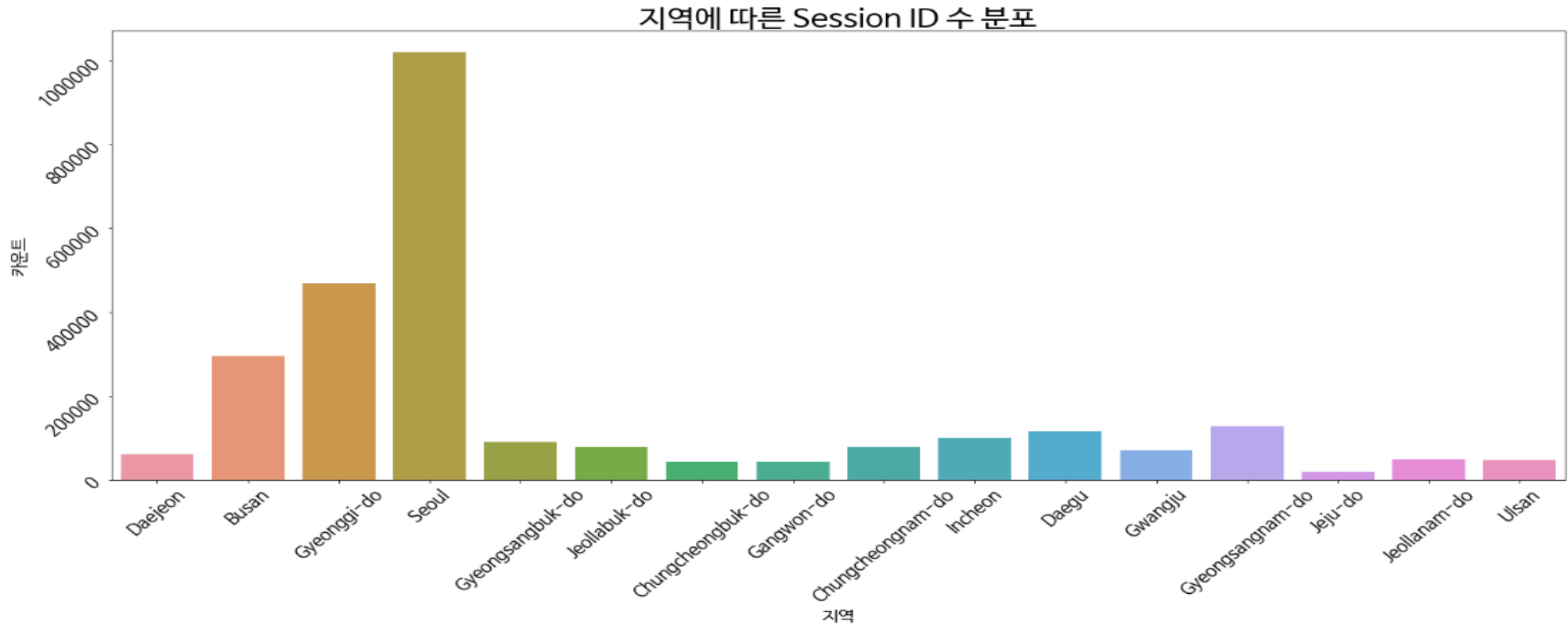
기기 유형에 따른 구매 건수 박스플롯 (Session ID 기준)



데스크탑의 경우 더 높은 구매 건수를 보이고 있다. 기기 유형이 구매 건수를 예측하는 데에 어느 정도 유의미한 변수임을 알 수 있다. 모바일과 태블릿의 경우 각각 평균 1.86과 2.00의 구매 건수를 보이는 반면 데스크탑은 2.34의 구매 건수를 보이고 있다. 흥미로운 결과다. 데스크탑 고객들의 특성을 조금 더 자세히 살펴볼 필요가 있어 보인다.

II. 데이터 전처리

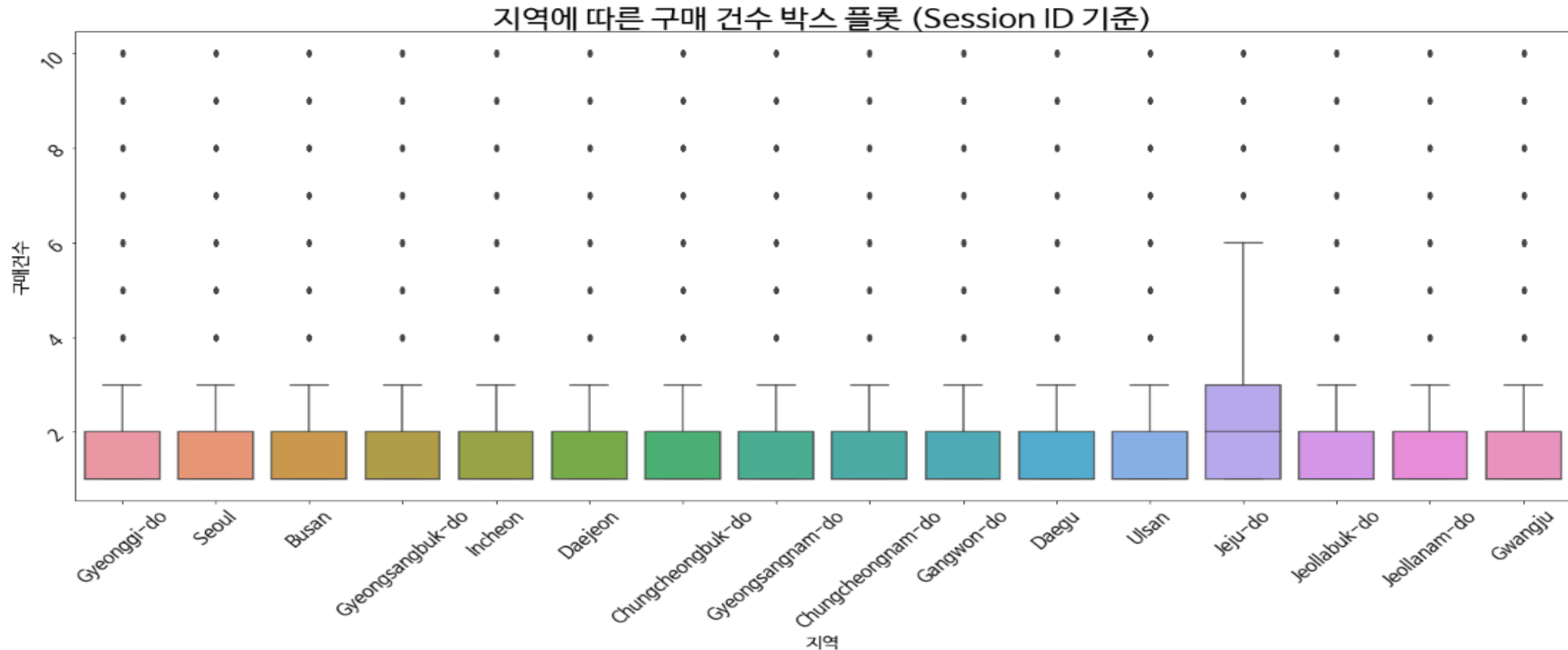
① 데이터 탐색 - 6) 지역



서울, 경기도, 부산이 타 지역들에 비해 눈에 띄게 높은 수치를 보이고 있다.

II. 데이터 전처리

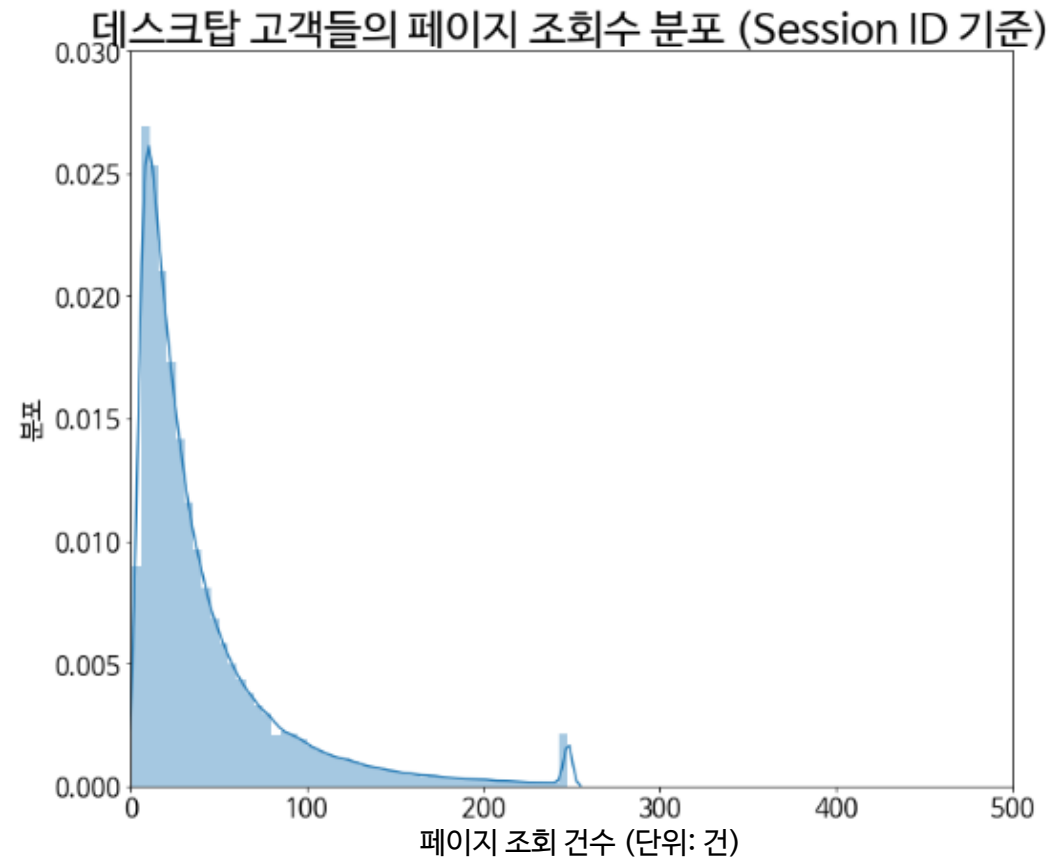
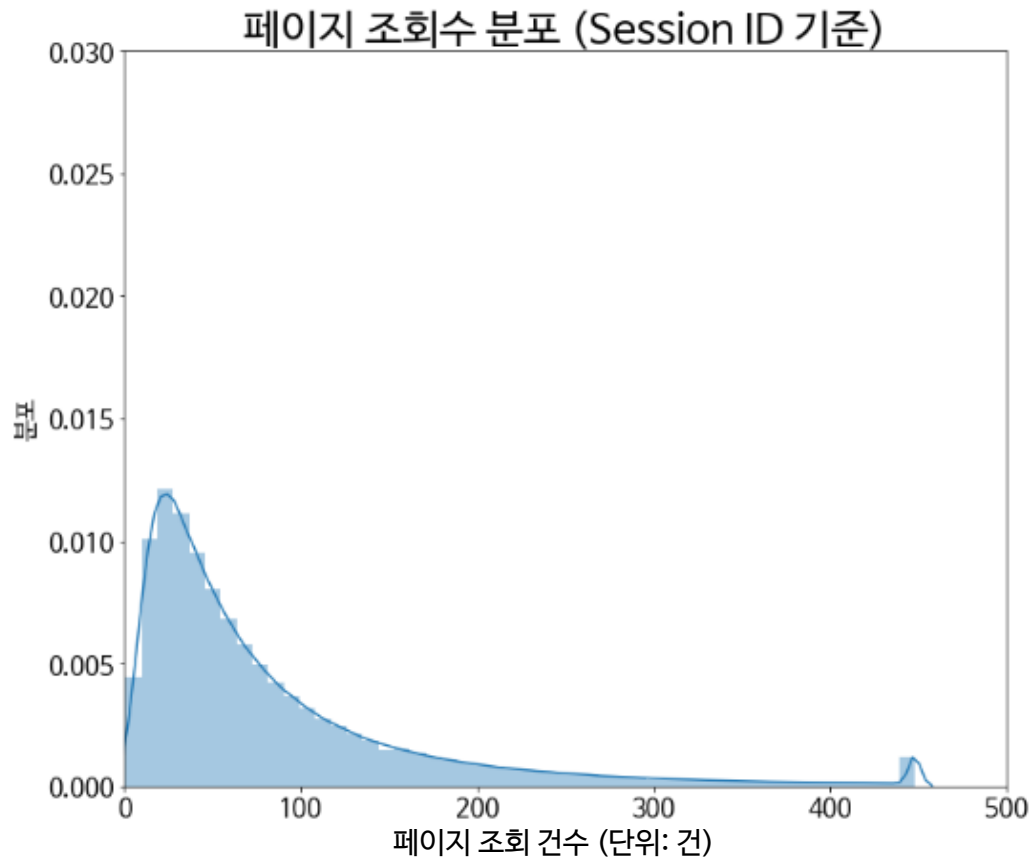
① 데이터 탐색 - 6) 지역



지역별로 제주도를 제외하면 거의 평균 구매 분포가 차이가 없다. 이전 자료를 살펴보면 제주도는 애초에 관측치 수가 너무 적기에 이 결과가 유의미한 정보라고 보긴 힘들 것이다.

II. 데이터 전처리

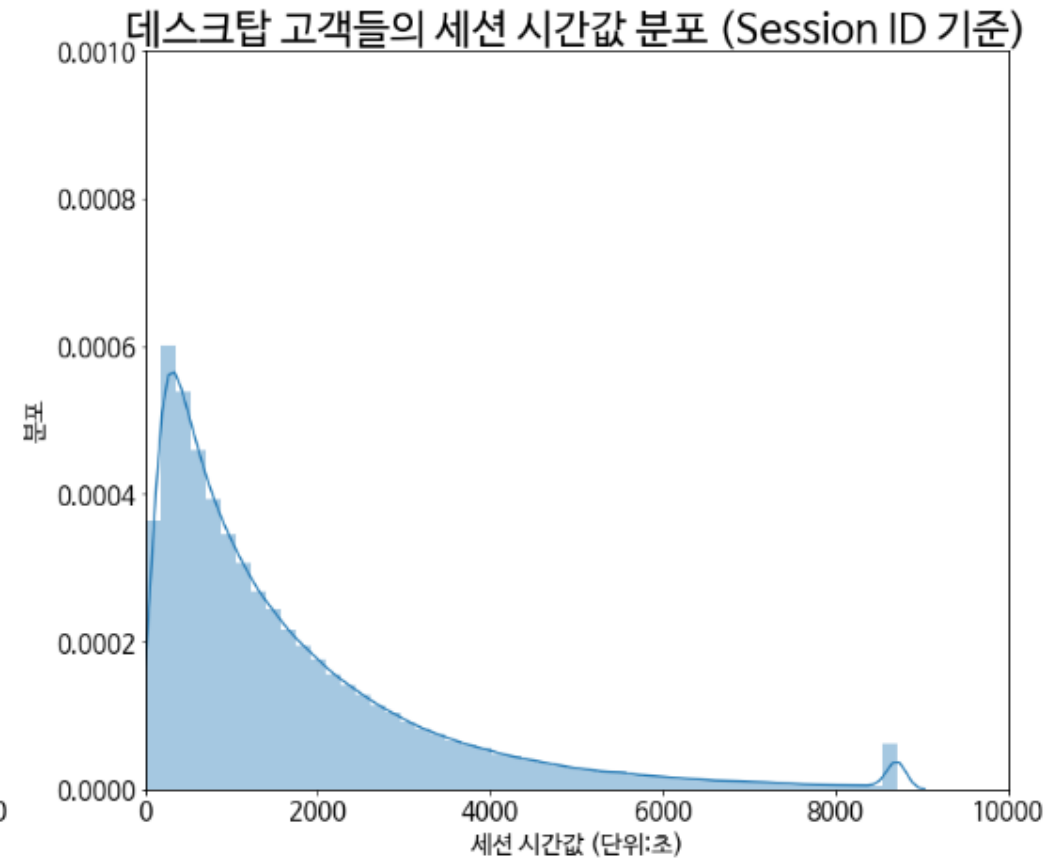
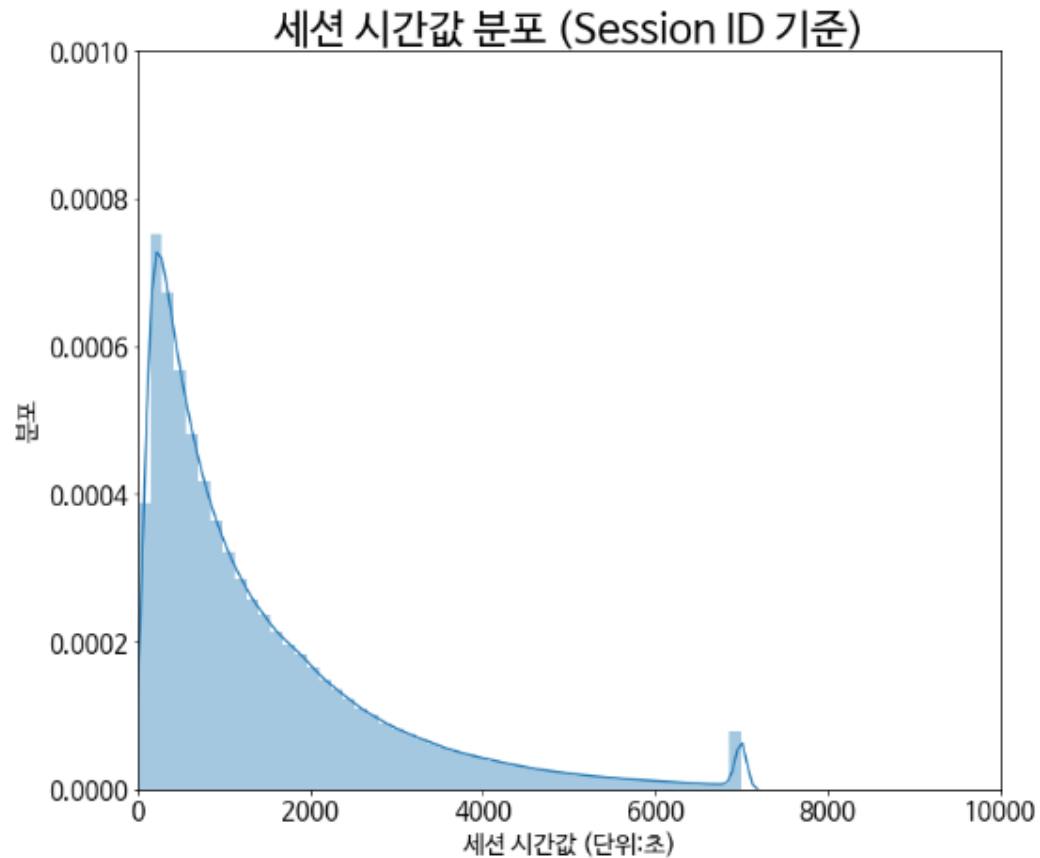
① 데이터 탐색 - 7) 데스크탑 고객들의 특징 (페이지 조회 건수)



앞서 데스크탑 고객들이 모바일, 태블릿 유저에 비해 평균 구매 건수가 확연히 높았기에 이들의 특징을 좀 더 자세히 살펴보고자 한다. 위 그래프를 보면, 데스크탑 고객들의 경우 세션의 페이지 조회 건수 분포가 확연히 적음을 볼 수 있다.

II. 데이터 전처리

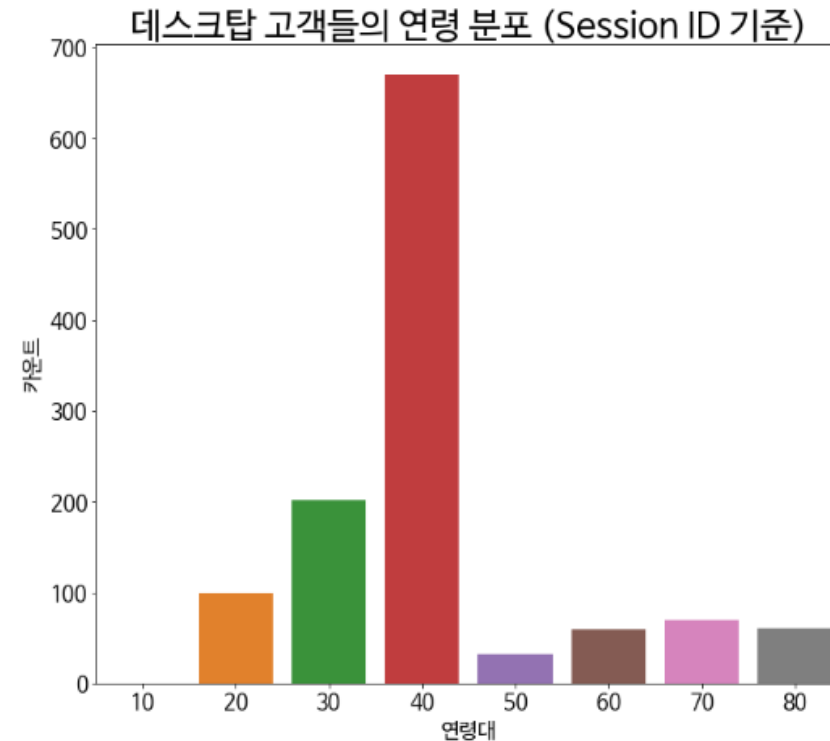
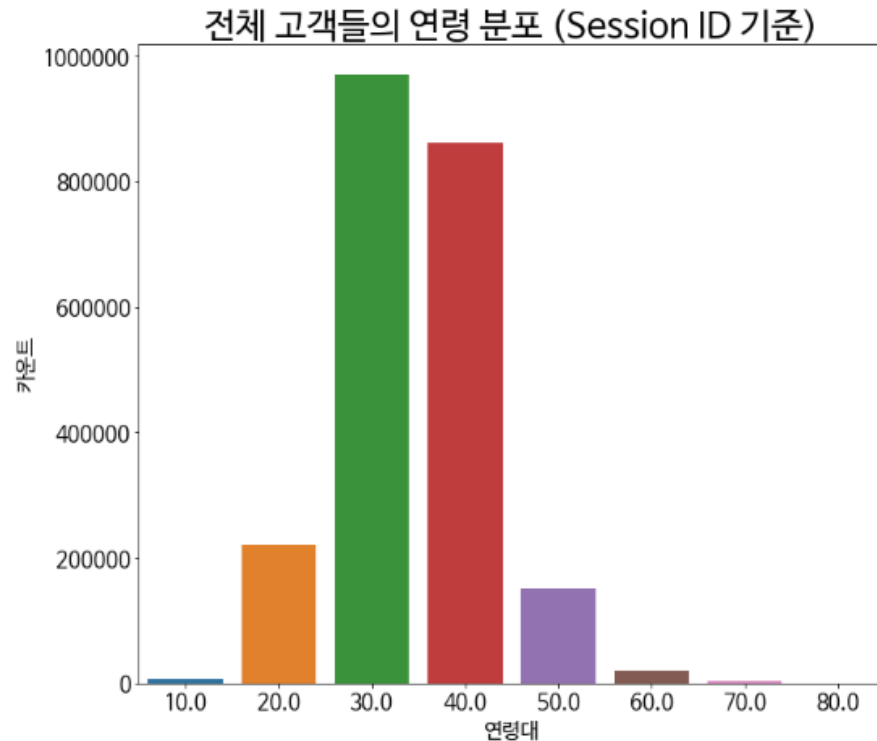
① 데이터 탐색 - 8) 데스크탑 고객들의 특징 (세션 시간값)



데스크탑 고객들은 세션 시간값의 분포가 꼬리가 좀더 길다. 즉 전반적으로 전체 고객들의 경우보다 세션 시간값이 높음을 알 수 있다.

II. 데이터 전처리

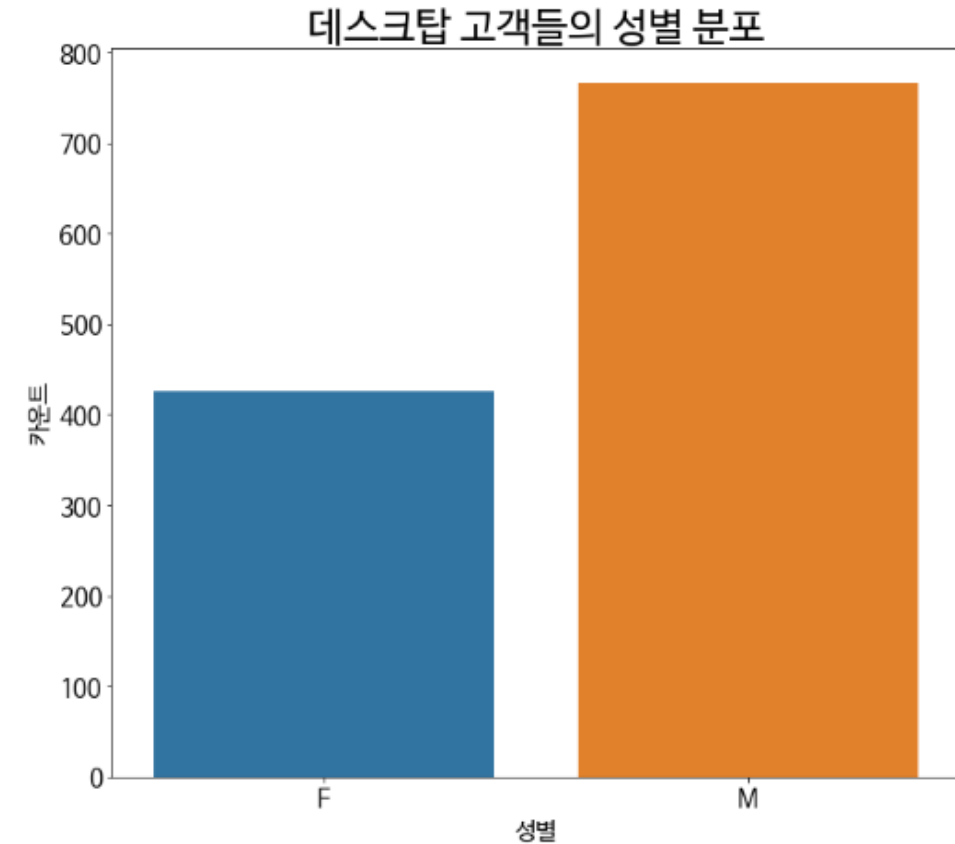
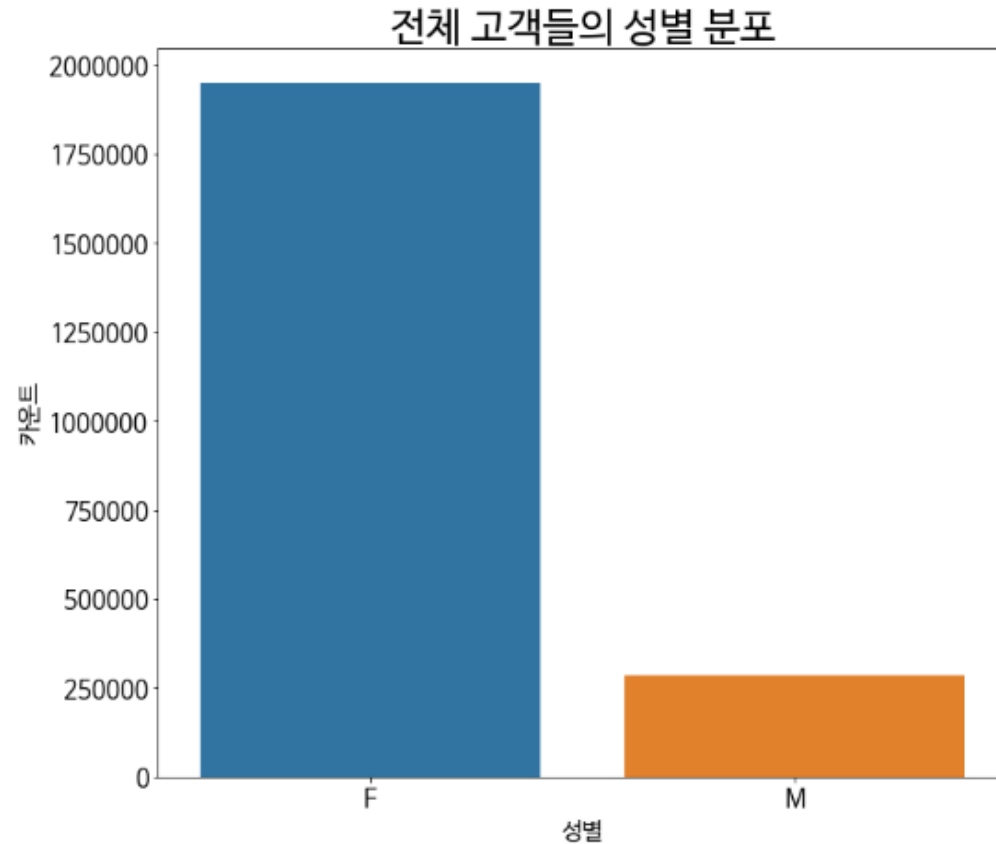
① 데이터 탐색 - 9) 데스크탑 고객들의 특징(연령대)



데스크탑 고객들의 연령대 분포는 전체 고객들의 연령대 분포에 비해 40대, 60~80대의 비율이 확연히 높음을 볼 수 있다. 하지만 데스크탑의 경우 연령대와 성별 정보가 수집이 어려워 결측치 상당히 많았다. 때문에 관측되는 데이터 수가 절대적으로 적었기에 큰 의미를 가지는 결과라고는 보기 힘들다.

II. 데이터 전처리

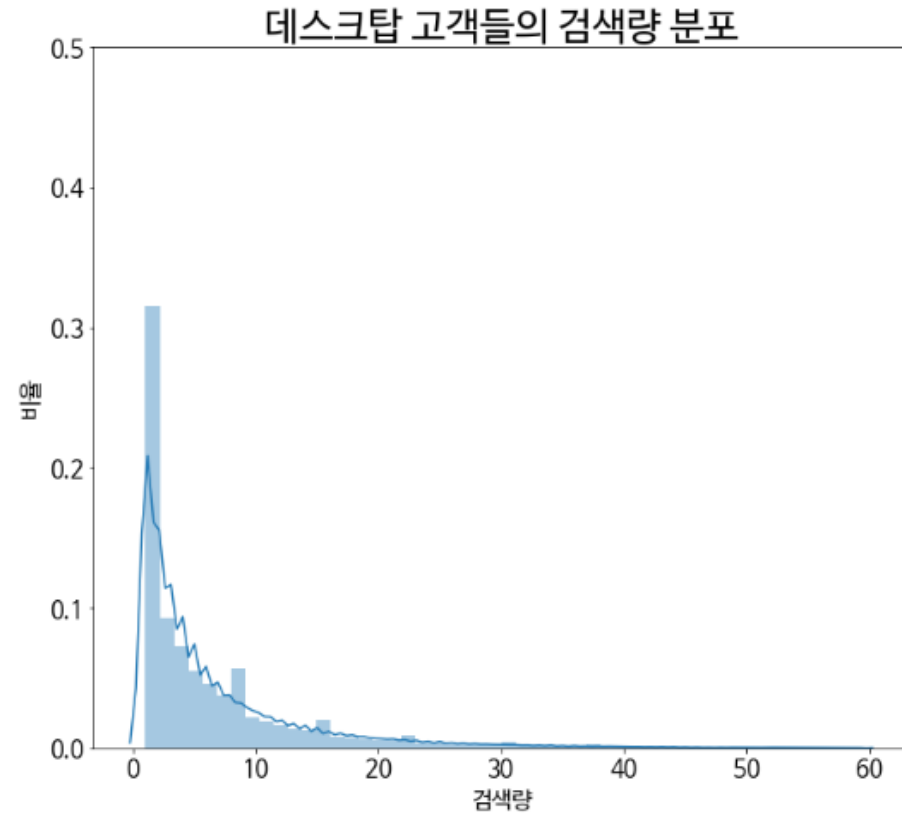
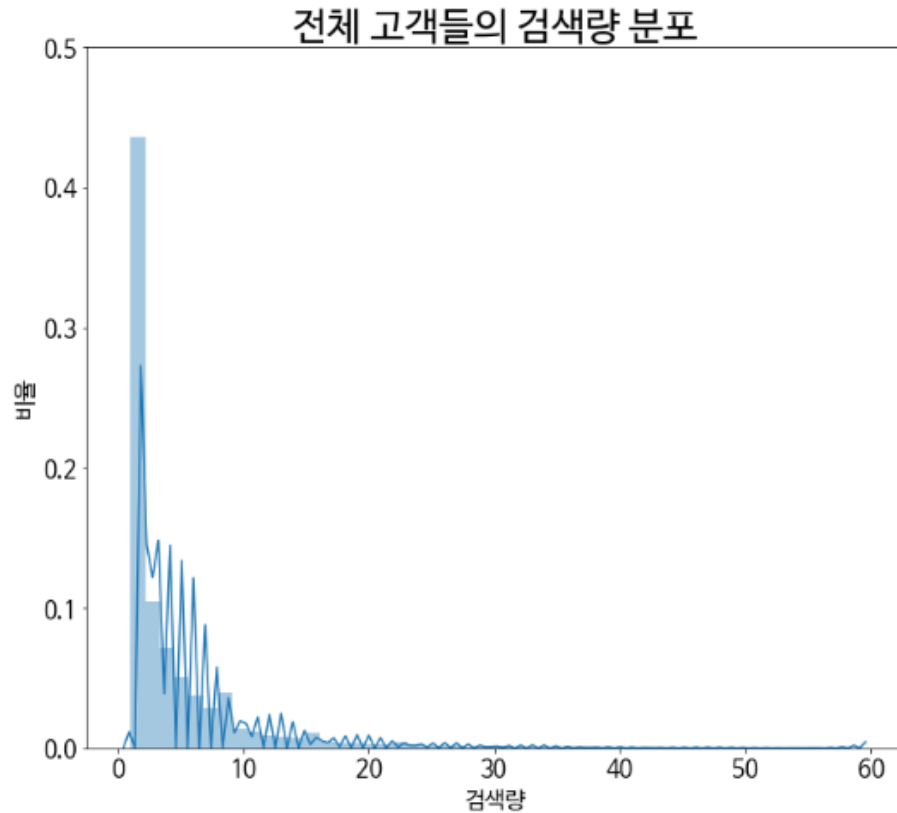
① 데이터 탐색 - 10) 데스크탑 고객들의 특징 (성별)



데스크탑 고객들의 경우 전체 고객들보다 남성 고객들의 분포가 훨씬 많음을 볼 수 있다. 하지만 앞서 연령대 정보와 마찬가지로 관측되는 수가 너무 적기에 유의미한 결과라고 보기는 힘들다.

II. 데이터 전처리

① 데이터 탐색 - 11) 데스크탑 고객들의 특징 (검색어 통계량)



데스크탑 고객들의 검색량 분포가 좀더 꼬리가 길다. 하지만 전반적으로 큰 차이는 없다.

II. 데이터 전처리

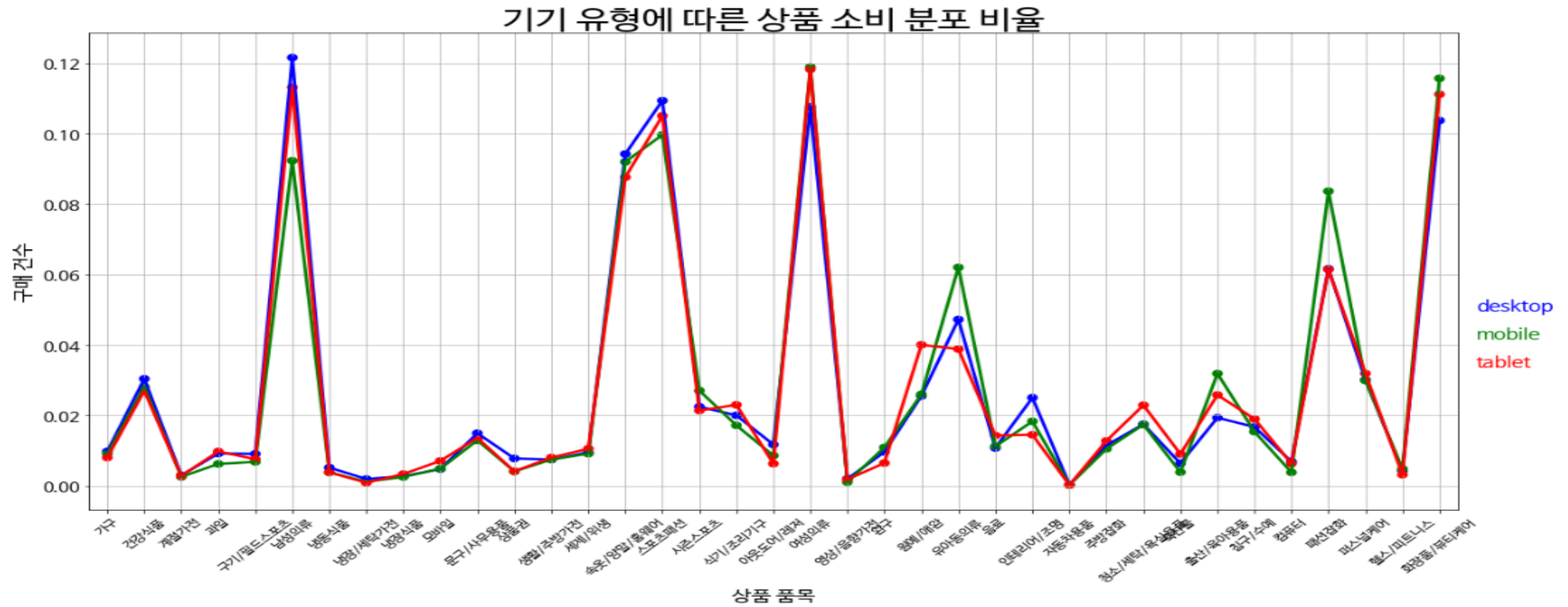
① 데이터 탐색 - 12) 데스크탑 고객들의 특징 (검색어 분포)



데스크탑 고객들의 검색어는 주로 티셔츠, 남성, 팬츠, 셔츠, 여성, 원피스, 블라우스 등이 많은 것을 알 수 있다.

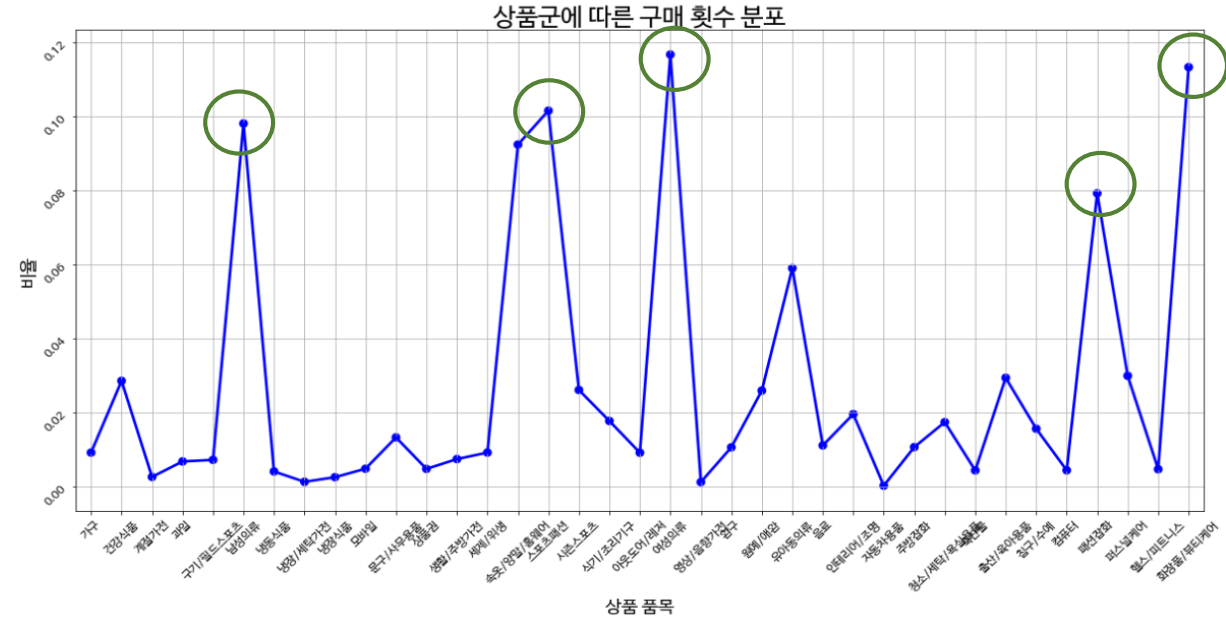
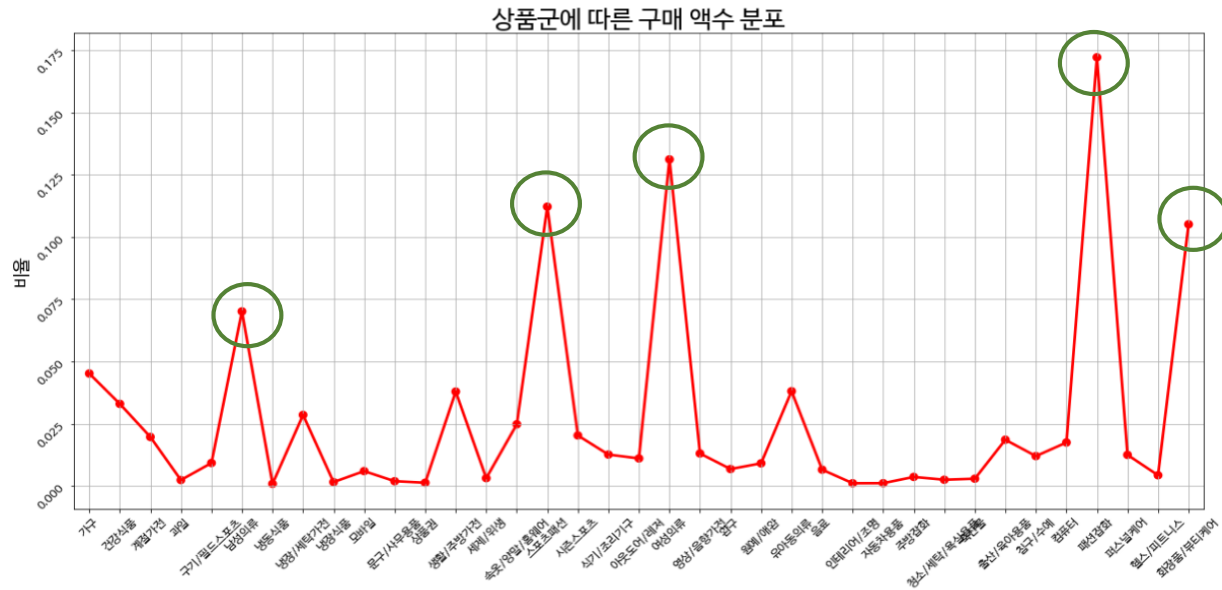
II. 데이터 전처리

① 데이터 탐색 - 13) 데스크탑 고객들의 특징(구매 품목)



데스크탑 고객들은 주로 남성외류, 스포츠패션에서 비교적 높은 수요를 보이고 있다.

① 데이터 탐색 - 14) 주요 품목 설정

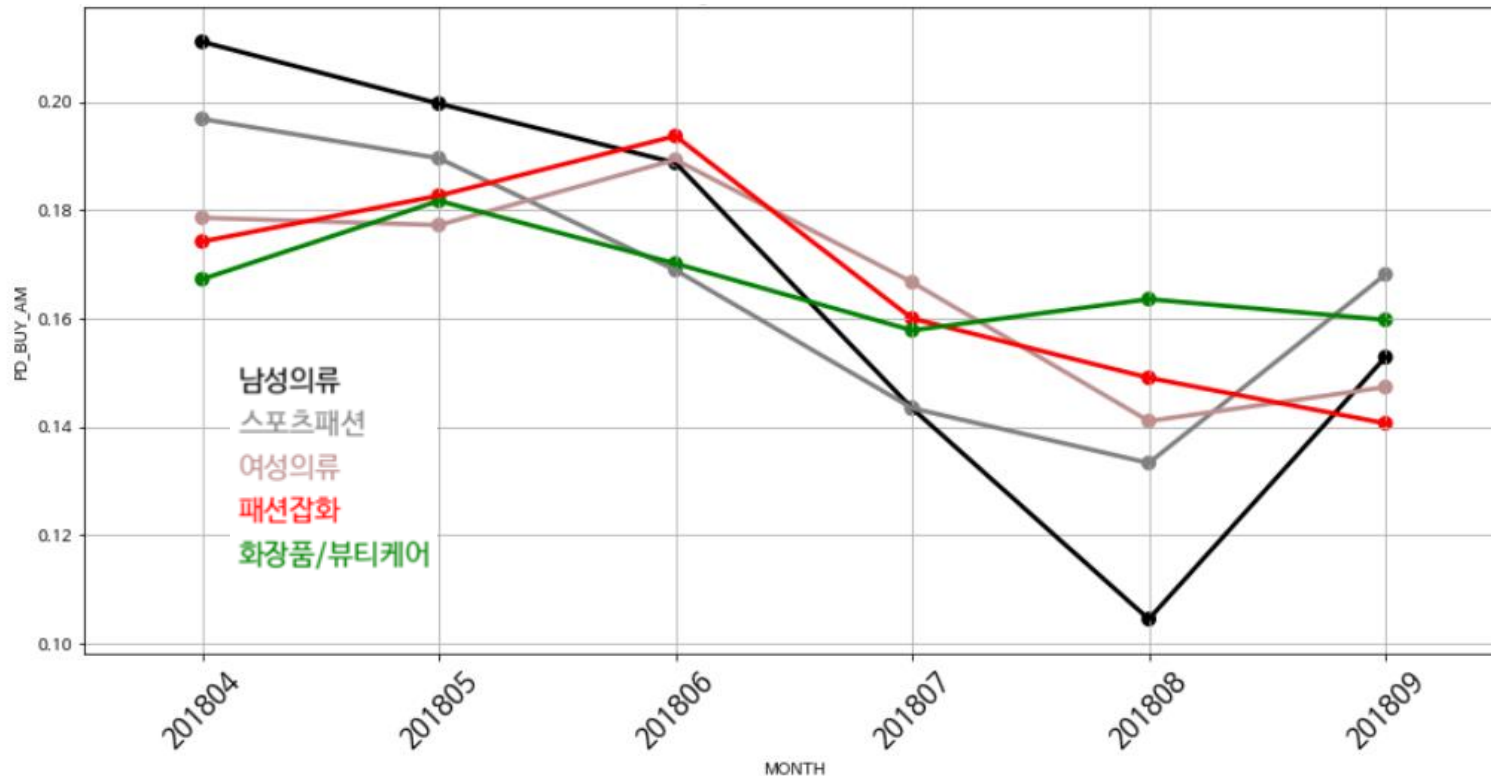


남성의류, 여성의류, 스포츠패션, 패션잡화, 화장품/뷰티케어는 구매 액수와 구매 횟수 면에서 모두 최상위권을 위치하고 있다. 즉 이 다섯 가지의 상품군은 회사의 매출에 중대한 비중을 차지하는 주요품목이다. 이 다섯 가지 제품군을 주요 품목으로 설정하고 이를 중점적으로 분석해보겠다.

II. 데이터 전처리

① 데이터 탐색 - 15) 주요 품목 특징

월 별 주요 상품 매출액 분포 비율



앞에서 설정한 5가지 주요 품목군의 월별 매출 기록이다. 남성의류와 스포츠패션이 8월에 특히 매출이 낮은 모습을 볼 수 있다. 8월을 나타내는 파생 변수를 추가하는 것이 좋은 아이디어로 보인다.

II. 데이터 전처리

① 데이터 탐색 - 살펴본 내용 정리

변수 별 기본 정보 외에 알아낸 내용들

- 페이지 조회수와 세션 시간값은 모두 구매 건수와 반비례하는 경향이 있다. 다만 세션 시간값은 20,000초 이후로는 구매건수와 약하게 정비례한다.
- 데스크탑 고객, 제주도 고객은 구매 건수가 높은 경향이 있음
- 데스크탑 고객들은 페이지 조회건수는 낮은 반면 세션 시간값은 높은 경향이 뚜렷함
- 데스크탑 고객들은 주로 40대 이상의 고객들이 많고 남성이 많음 (하지만 결속치가 매우 많아서 관측된 데스크탑 고객들의 연령, 성별 정보는 약 10,000개도 안 되는 수준임)
- 데스크탑 고객들의 검색량은 주로 많은 편임
- 데스크탑 고객들의 검색 단어는 주로 티셔츠, 남성, 팬츠, 셔츠, 여성, 원피스, 블라우스 등이 많음
- 데스크탑 고객들은 스포츠 패션, 남성 의류에 대해 비교적 높은 수요를 보임

II. 데이터 전처리

① 데이터 탐색 - 살펴본 내용 정리

변수 별 기본 정보 외에 알아낸 내용들

- 구매 건수와 구매 액수를 기준으로 봤을 때 가장 규모가 큰 상위 5가지 상품군은 남성의류, 여성의류, 스포츠패션, 패션잡화, 화장품/뷰티케어임
(이 다섯 상품군을 주요 품목군으로 설정할 것임)
- 위 주요 상품군들은 계절에 따라서 수요 폭의 차이가 있음
- 특히 8월에는 스포츠 패션과 남성 의류가 눈에 띄는 감소를 보이고 있음

II. 데이터 전처리

② 데이터 전처리 - 1) 원하는 데이터 형태

- 월 기준의 시계열 데이터
- (고객 ID, 월)을 복합키로 갖는 형태의 데이터
- 주요 품목군의 다음 달 구매 여부를 나타나는 종속 변수까지 달려 있는 형태 (주요 품목군의 다음 달 구매 여부를 예측하는 것이 모델링의 목적임)

② 데이터 전처리 - 2) 클렌징 (변수 제거, 데이터 형태 변경, 결측치 처리)

Product 변수 정리

1. 변수제거 : 'HITS_SEQ', 'PD_ADD_NM', 'PD_BRA_NM'
2. 데이터 형태 Character 형으로 바꾸기 : 'CLNT_ID', 'SESS_ID'
3. 데이터 형태 Int형으로 바꾸기 : 'PD_BUY_AM', 'PD_BUY_CT'

Search1 변수 정리

1. 데이터 형태 Character 형으로 바꾸기 : 'CLNT_ID', 'SESS_ID'

Search2 변수 정리

1. 데이터 형태 Int형으로 바꾸기 : 'SEARCH_CNT'
2. 데이터 형태 Date형으로 바꾸기 : 'SESS_DT'

Custom 변수 정리

1. 데이터 형태 Character 형으로 바꾸기 : 'CLNT_ID', 'CLNT_AGE'

Session 변수 정리

1. 변수제거 : 'SESS_SEQ'
2. 데이터 형태 Character 형으로 바꾸기 : 'CLNT_ID', 'SESS_ID'
3. 데이터 형태 Int형으로 바꾸기 : 'TOT_SESS_HR_V', 'TOT_PAG_VIEW_CT'
4. 데이터 형태 Date형으로 바꾸기 : 'SESS_DT'
5. 결측치 값 평균치로 채워 넣기 : 'TOT_SESS_HR_V', 'TOT_PAG_VIEW_CT'

Master 변수 정리

1. 변수제거 : 'PD_NM', 'CLAC2_NM', 'CLAC3_NM'

② 데이터 전처리 - 3) 클렌징 (스케일링, 범주형 변수 처리, 통합)

1. 고빈도 팩터라이징 인코딩

- 고유 값 중에서 빈도가 높은 순으로 100개 까지는 0부터 99까지 숫자로 바꿨다.
- 빈도 순위가 100위 밖인 데이터 값들과 결측치는 모두 -99로 바뀌었다.
- 적용 변수 : 'DVC_CTG_NM', 'ZON_NM', 'CITY_NM', 'KWD_CNT', 'CLNT_AGE'

2. 원핫 인코딩

- 변수의 특성 차이를 더욱 확실히 나타내고 싶으면서 고유 값의 개수가 적은 변수는 고빈도 팩터라이징 인코딩이 아닌 원핫 인코딩을 적용했다.
- 적용 변수 : 'CLNT_GENDER'

3. 스케일링

- 변수 별로 데이터 분포를 비슷하게 만들어주기 위해 표준화 스케일링을 적용하였다.
- 적용 변수 : 'TOT_SESS_HR_V', 'TOT_PAG_VIEW_CT'

4. 하나의 테이블로 통합

- 위 6개의 데이터 테이블 중 Search2를 제외한 5개의 테이블을 하나로 통합했다.
- 'CLNT_ID', 'SESS_MONTH'를 복합키로 하고 이를 기준으로 통합했다. ('SESS_MONTH'는 새로 만든 변수로써, 뒤에 피처 엔지니어링에서 소개하고 있음)
- 'CLNT_ID', 'SESS_MONTH'를 기준으로 합칠 때 "KWD_NM", "DVC_CTG_NM", "ZON_NM", "CITY_NM"의 데이터 값은 가장 빈도수가 높은 값으로 설정했다. (본래 "SESS_ID" 기준으로 돼있던 변수들임)

② 데이터 전처리 - 4) 피처 엔지니어링

1. 월 변수 추가

- 앞서 데이터 탐색 과정에서 주요 품목군들이 월마다 수요 양상이 다름을 고려하여 월 변수를 추가하였다.
- 변수명 : 'SESS_MONTH'

2. 휴가철의 여부를 나타내는 변수 추가

- 월 중에서도 8월의 경우 다른 때와 비교하여 매우 상이한 수요 양상을 보였기에 이를 판별하고자 8월 인지를 나타내는 변수를 추가하였다.
- 변수명 : 'IS_VACATION'

3. 직전 월(月)에 5가지의 주요 품목군을 통틀어 구매 금액이 얼마나 되는지 나타내는 변수 추가

- 이전에 구매를 얼마나 많이 했는 지의 여부가 다음달 구매 여부를 예측하는 데 중요한 정보라고 생각했다. 하지만 4번과 같이 5가지의 주요 상품군들 각각을 모두 하나의 열로 나타내면 열의 수가 너무 많아져 모델링에 좋지 않은 영향을 끼치기에 주요 상품군을 통틀어서 구매한 액수를 나타내는 한 개의 변수를 추가했다.
- 4월 고객 데이터는 전월의 구매 여부에 대한 데이터가 없으므로 삭제하였다.
- 변수명 : "PREV_MONTH_PD_BUY_AM"

4. 직전 월(月)에 5가지의 주요 품목군을 각각 몇 개 구매 했는지 나타내는 변수 추가

- 이전에도 구매를 했던 고객인지 여부가 다음 달 구매 여부를 예측하는 데에 중요한 정보라고 생각했고 직전 달에 주요 품목 구매 횟수가 어떻게 되는지 나타내는 변수를 추가했다. (LAG -1 변수의 성격)
- 4월 고객 데이터는 전월의 구매 여부에 대한 데이터가 없으므로 삭제하였다.
- 변수명 : "PREV_MONTH_남성의류", "PREV_MONTH_스포츠패션", "PREV_MONTH_여성의류", "PREV_MONTH_패션잡화", "PREV_MONTH_화장품/뷰티케어"

5. 다음 달에 주요 품목 5가지를 각각 구매했는지 여부를 나타내는 변수 추가 (타겟변수)

- 이는 예측할 타겟 변수이다.
- 9월 데이터는 다음 달의 구매 여부에 대한 데이터가 없으므로 삭제하였다.
- 변수명 : "NEXT_MONTH_남성의류", "NEXT_MONTH_스포츠패션", "NEXT_MONTH_여성의류", "NEXT_MONTH_패션잡화", "NEXT_MONTH_화장품/뷰티케어"

II. 데이터 전처리

② 데이터 전처리 - 5) 결과물 형태

전처리를 마친 데이터는 702,594 * 28 형태를 가진 월 단위 시계열 데이터이다.

Ex) 상위 6개의 관측치 확인

CLNT_ID	SESS_MON TH	PD_BUY_A M	CLAC1_NM _남성의류	CLAC1_NM _스포츠패 션	CLAC1_NM _여성의류	CLAC1_NM _패션잡화	CLAC1_NM _화장품/뷰 티케어	TOT_PAG_ VIEW_CT	TOT_SESS_ HR_V	DVC_CTG_ NM	ZON_NM	CITY_NM	KWD_CNT	CLNT_AGE	CLNT_GEN DER_F	is_vacation	PREV_MO NTH_PD_B UY_AM	PREV_MO NTH_남성 의류	PREV_MO NTH_스포 츠패션	PREV_MO NTH_여성 의류	PREV_MO NTH_패션 잡화	PREV_MO NTH_화장 품/뷰티케 어	NEXT_MONTH_ 남성의류	NEXT_MONTH_ 스포츠패션	NEXT_MONTH_ 여성의류	NEXT_MONTH_ 패션잡화	NEXT_MONTH_ 화장품/뷰티케 어
10000	5	77800	0	0	0	2	0	-0.39519	-0.43275	0	0	0	-99	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1000021	7	315600	0	0	0	0	6	0.241013	-0.25293	0	1	1	-99	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1000028	5	158500	0	0	5	0	0	-0.49522	-0.41401	1	2	2	-99	-99	0	0	0	0	0	0	0	0	0	0	0	0	0
1000033	6	124000	0	0	0	0	1	-0.4512	-0.56622	0	1	1	-99	0	1	0	0	0	0	0	0	0	0	0	0	0	0
100004	7	20800	0	1	0	0	1	-0.23514	-0.29472	0	3	3	-99	2	1	0	0	0	0	0	0	0	1	0	0	0	0
100004	8	40000	0	2	0	0	0	-0.33917	-0.35196	0	3	3	-99	2	1	1	0	0	1	0	0	2	0	0	0	0	1

22개의 설명 변수

5개의 타겟 변수

III. 예측 모형



III. 예측 모형

① 모델링 - 1) 목적

“고객이 다음 달에 구매할 상품을 예측하자”

① 모델링 - 2) 원리 소개

1. XGBoost 모델 100개를 묶어서 앙상블을 꾸림

- 타겟 변수의 클래스 비중이 약 10:1 정도로 불균형했기에 이를 해결하고자 앙상블을 적용함. 클래스가 1인 데이터를 (전체사이즈/2)로 다운 샘플링, 클래스가 9인 데이터를 (전체사이즈/20)으로 다운 샘플링하여 1:1 비율을 맞춘 트레인셋을 100개 구성
- 각 100개의 트레인셋에서 모델을 하나씩 적합을 하여 100개의 모델을 만듦
- 들어오는 테스트 셋마다 100개의 모델 모두가 예측을 하고 100개 모두 1이라고 예측한 경우를 1, 그 외는 0으로 최종 결과를 반환 (기준점을 100개로 잡은 것은 1~100까지 모두 실험해본 결과 100이 F1 Score가 가장 높았기 때문)

2. 트레인, 테스트셋 비율은 9:1로 구성

- 트레인셋은 약 63만개, 테스트셋은 약 7만 개

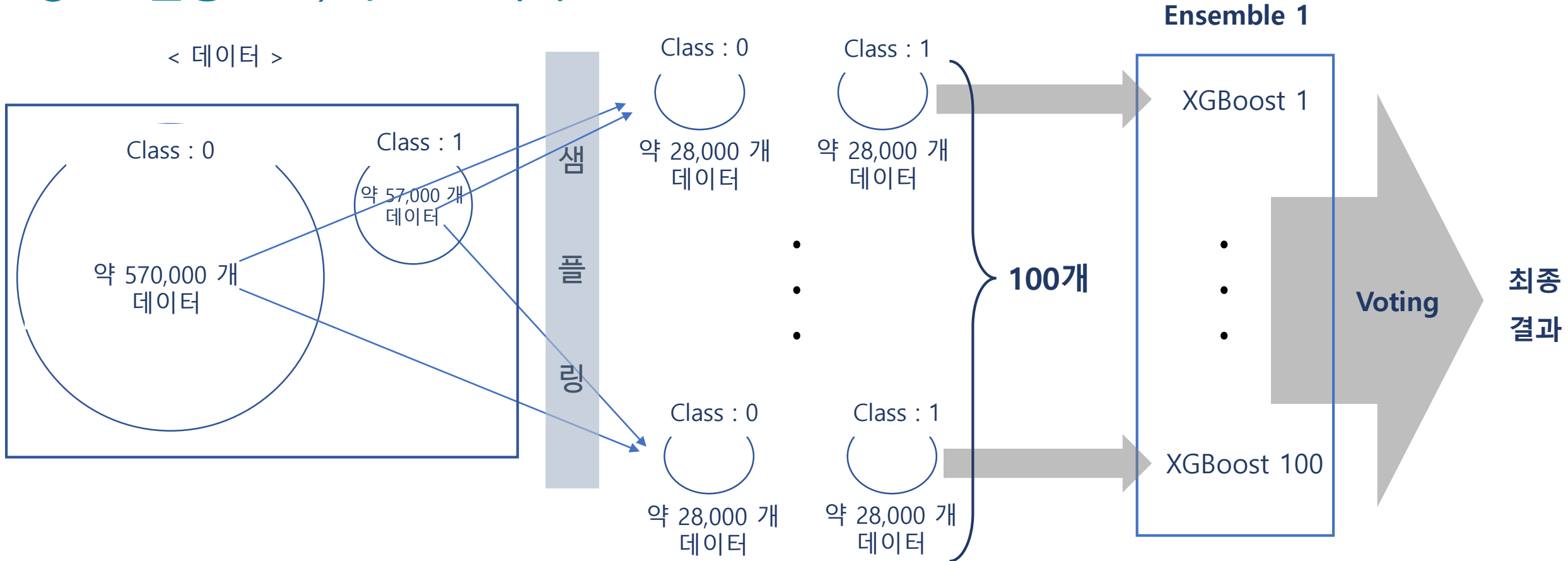
3. 각 모델의 파라미터는 다음과 같음

- max_depth=4
- nthread=4
- colsample_bytree=0.8
- colsample_bylevel=0.9
- min_child_weight=10
- n_jobs=4
- scale_pos_weight=10

4. 5개의 주요 품목별로 각각 이런 식의 모델을 하나씩 만들어서 총 5개의 앙상블 모델을 만듦

III. 예측 모형

① 모델링 - 3) 구조 도식화



위와 같은 하나의 앙상블 모델을 주요 품목마다 하나씩 만들어서 총 5개의 모델로 예측을 진행했다.

III. 예측 모형

② 결과

남성의류

실제 예측 \	0	1
0	60,428	2,299
1	5,692	1,846



0.316

스포츠패션

실제 예측 \	0	1
0	56,781	2,912
1	8,919	2,648



0.309

여성의류

실제 예측 \	0	1
0	53,357	2,294
1	10,398	4,211



0.399

패션잡화

실제 예측 \	0	1
0	55,303	2,958
1	9,089	2,910



0.326

화장품/뷰티케어

실제 예측 \	0	1
0	52,382	2,561
1	11,540	3,777



0.349

F1 Score

※ F1 Score = $\frac{2(PREC \cdot REC)}{PREC + REC}$

A dimly lit clothing store interior. In the center, a round wooden table holds several folded items. To the left, a long display case is filled with various accessories. The background features racks of clothing and a wall with a grid of small, colorful items. The floor is made of light-colored wood. The overall atmosphere is modern and minimalist.

IV. 선호지수 개발

IV. 선호지수 개발

“브랜드 선호도란 해당 상품군의 고객들이 자신만의 선호하는 브랜드가
뚜렷한 경향이 얼마나 강한 지를 나타내는 지수이다.”

IV. 선호지수 개발

① “브랜드 선호도” - 2) 개념 소개

Ex)

ID	구매한 상품군	구매한 브랜드	구매시 검색어
1	퍼스널케어	아베다	아베다 스무드
2	퍼스널케어	아베다	아베다
3	스포츠패션	아디다스(의류)	나이키레깅스

두 고객 모두 구매한 브랜드와 검색한 브랜드가 일치.
“아베다”를 검색하고 아베다 상품을 구매했음.
아베다에 대한 선호도가 뚜렷함

검색한 브랜드와 구매 브랜드가 불일치.
“나이키레깅스”를 검색하고 아디다스 상품을 구매함.
처음부터 아디다스를 사고자 했던 생각은 없었으나
아디다스를 구매함. 특정 브랜드에 대한 선호도가 뚜렷
하지 않은 채로 상품 구매가 이루어짐

퍼스널 케어는 스포츠 패션에 비해 브랜드 선호도가 뚜렷한 경향을 보인다.

IV. 선호지수 개발

① “브랜드 선호도” - 3) 개발 과정

※ 앞선 예시 계속 사용하겠습니다.

과정 1. 브랜드 열과 검색어 열에 있는 데이터 값들을 띄어쓰기, 괄호 기준으로 쪼갬 (리스트 형태가 됨)

구매한 상품군	구매한 브랜드	구매시 검색어
퍼스널케어	[아베다]	[아베다, 스무드]
퍼스널케어	[아베다]	[아베다]
스포츠패션	[아디다스, 의류]	[나이키, 레깅스]

과정 3. ‘해당 상품군이 획득한 점수 합/ 상품군 수’의 수식으로 모든 상품군마다 결과값을 구함

상품군	브랜드 선호 지수
퍼스널케어	$(2 + 2) / 2 = 2$
스포츠 패션	$0 / 1 = 0$

과정 2. 각 행마다 브랜드 리스트의 원소들이 검색어 리스트의 원소들과 같거나 포함될 때마다 +1점, 반대로도 똑같이 진행

구매한 상품군	구매한 브랜드	구매시 검색어
퍼스널케어	[아베다]	[아베다, 스무드]
퍼스널케어	[아베다]	[아베다]
스포츠패션	[아디다스, 의류]	[나이키, 레깅스]

= 2



“스무드”는 브랜드 리스트의 원소와 같거나 포함되지 않으므로 영향 無

= 2



“아디다스”, “의류”와 “나이키“, “레깅스” 모두 서로 같거나 포함되지 않으므로 영향 無

= 0



“아디다스”, “의류”와 “나이키“, “레깅스” 모두 서로 같거나 포함되지 않으므로 영향 無

IV. 선호지수 개발

② 모든 상품군의 브랜드 선호도 구해보기

〈상품군 상품권의 데이터 예시〉

상품군	브랜드 선호도
상품권	0.646516
화장품/뷰티케어	0.554699
유아동의류	0.459436
스포츠패션	0.440630
냉장식품	0.421572
퍼스널케어	0.405034
축산물	0.402070
시즌스포츠	0.400563
남성의류	0.396624
패션잡화	0.388964
출산/육아용품	0.373582

ID	구매시 검색어	구매한 상품군	구매한 브랜드
2912141	[씨유]	상품권	씨유
1487448	[세븐일레븐]	상품권	세븐일레븐
2114024	[cu]	상품권	세븐일레븐
2114024	[세븐일레븐]	상품권	세븐일레븐
1227213	[지에스25]	상품권	지에스25
2154076	[gs25]	상품권	지에스25
1094189	[파리바게뜨]	상품권	파리바게뜨

〈상품군 냉장/세탁가전의 데이터 예시〉

음료	0.151943
과일	0.130857
청소/세탁/육실용품	0.129972
컴퓨터	0.120496
문구/사무용품	0.098888
냉장/세탁가전	0.053380

ID	구매시 검색어	구매한 상품군	구매한 브랜드
6414777	[이어캡]	냉장/세탁가전	[LG전자]
1604266	[B267S]	냉장/세탁가전	[LG전자]
1572024	[TV]	냉장/세탁가전	[LG전자]
3840666	[냉장고]	냉장/세탁가전	[삼성전자]
3840666	[세탁기]	냉장/세탁가전	[삼성전자]
3840666	[꽃배달]	냉장/세탁가전	[삼성전자]
5634928	[삼성드럼세탁기]	냉장/세탁가전	[삼성전자]

브랜드 선호도가 가장 높은 상품권의 경우와 가장 낮은 냉장/세탁가전의 경우를 몇 개의 관측치를 통해서 비교해보니 검색어와 구매한 브랜드 간의 유사성 차이가 뚜렷함을 알 수 있다.

③ 브랜드 선호도의 유용성

- 고객들이 각각 선호하는 자신만의 브랜드가 뚜렷하고 이 브랜드 제품을 주로 구매하는지의 여부는 회사 입장에서 어떤 판매 전략을 세울지 결정하는 데 중요한 정보이다.
- 직전 슬라이드에서 보았듯이 실제로 상품군마다 이런 차이가 유의미하게 드러나고 있는 것을 보아, 이를 토대로 다양하고 차별화된 마케팅 방법을 구상할 수 있는 여지가 굉장히 크다.

A dimly lit retail store interior, likely a clothing boutique. The space features wooden flooring, a wall of colorful patterned tiles on the left, and various clothing displays. In the center, a round table holds folded garments under several red pendant lights. To the right, a long wooden display case contains a model of a blue boat. Clothing racks with shirts and jackets are visible in the background.

V. 인사이트

① 모델의 변수 중요도를 통해 봤을 때

Features	남성 의류 예측 모델	스포츠패션 예측 모델	여성의류 예측 모델	패션잡화 예측 모델	화장품/뷰티케어 예측 모델
CITY_NM	0.020408	0.025606	0.013387	0.014895	0.019836
CLAC1_NM_남성의류	0.07415	0.043801	0.029451	0.040623	0.04104
CLAC1_NM_스포츠패션	0.046259	0.105795	0.033467	0.041977	0.038304
CLAC1_NM_여성의류	0.043537	0.031671	0.10174	0.05281	0.042408
CLAC1_NM_패션잡화	0.036054	0.044474	0.046854	0.099526	0.045144
CLAC1_NM_화장품/뷰티케어	0.038776	0.037736	0.056894	0.060257	0.090287
CLNT_AGE	0.068027	0.057278	0.083668	0.055518	0.03762
CLNT_GENDER_F	0.036735	0.01752	0.038153	0.016926	0.023256
DVC_CTG_NM	0.035374	0.039757	0.03079	0.040623	0.038988
KWD_CNT	0.012245	0.012803	0.006024	0.00677	0.018468
PD_BUY_AM	0.087075	0.08221	0.078983	0.078538	0.081395
PREV_MONTH_PD_BUY_AM	0.071429	0.083558	0.086345	0.087339	0.097127
PREV_MONTH_남성의류	0.042857	0.021563	0.014726	0.014895	0.01026
PREV_MONTH_스포츠패션	0.021769	0.046496	0.006693	0.019634	0.019836
PREV_MONTH_여성의류	0.026531	0.018194	0.048862	0.021666	0.022572
PREV_MONTH_패션잡화	0.023129	0.015499	0.023427	0.038592	0.0171
PREV_MONTH_화장품/뷰티케어	0.015646	0.018194	0.021419	0.025728	0.062927
SESS_MONTH	0.101361	0.104447	0.104418	0.09411	0.084131
TOT_PAG_VIEW_CT	0.1	0.088275	0.089023	0.085308	0.098495
TOT_SESS_HR_V	0.070068	0.068733	0.056894	0.077183	0.080027
ZON_NM	0.011565	0.018194	0.006693	0.007448	0.014364
is_cation	0.017007	0.018194	0.022088	0.019634	0.016416

※ 변수 중요도가 0.07이 넘는 경우는 노란색으로 칠했습니다.

- 구매 월, 페이지 조회 건수, 세션 값, 구매 총액, 이전달 구매 총액이 대체적으로 유의미한 변수인 것으로 드러남. XGBoost가 앙상블 모델이라서 회귀분석만큼 변수의 영향력을 정교하게 알 수는 없지만 EDA의 결과를 참조하여 변수 중요도 결과값들을 해석하자면,

- 1) 이전 월, 당 월에 지출한 액수와 구매 건수가 많을수록 다음 달에 어떤 상품을 구매할 가능성이 높음
- 2) 페이지 조회건수, 세션 시간값이 적을수록 다음 달에 어떤 상품을 구매할 가능성이 높음
- 3) 남성의류, 스포츠패션, 여성의류, 패션잡화는 7,8월 여름이 되면서 수요가 감소하는 반면 화장품/뷰티케어는 수요가 꾸준한 경향이 강하고 이는 수요를 예측하는 데에 큰 작용을 함

- 동일 품목의 구매 건수가 다음달 구매 여부를 예측하는 데에 중요한 변수인 것으로 나타남. 예를 들면 남성의류를 예측할 땐 남성의류의 구매건수, 스포츠패션을 예측할 땐 스포츠 패션의 구매 건수가 중요한 식임


② 브랜드 선호도를 통해 봤을 때

상품군	브랜드 선호도
상품권	0.646516
화장품/뷰티케어	0.554699
유아동의류	0.459436
스포츠패션	0.440630
냉장식품	0.421572
퍼스널케어	0.405034
축산물	0.402070
시즌스포츠	0.400563
남성의류	0.396624
패션잡화	0.388964
출산/육아용품	0.373582
건강식품	0.365876
식기/조리기구	0.355631
여성의류	0.350874
자동차용품	0.343832
생활/주방가전	0.342150
헬스/피트니스	0.329146
숙옷/양말/홈웨어	0.321594
가구	0.275234
침구/수예	0.271798
아웃도어/레저	0.269594
구기/필드스포츠	0.268674
영상/음향가전	0.256015
원예/애완	0.241316
모바일	0.220215
세제/위생	0.196841
완구	0.196440
인테리어/조명	0.194500
냉동식품	0.187416
계절가전	0.181105
주방잡화	0.179101
음료	0.151943
과일	0.130857
청소/세탁/육실용품	0.129972
컴퓨터	0.120496
문구/사무용품	0.098888
냉장/세탁가전	0.053380

선호도
감소

- 브랜드 선호도가 높은 상품군들에는 화장품/뷰티케어, 스포츠패션, 남성의류, 퍼스널케어 등으로 주로 자신만의 개성을 나타낼 만한 패션, 뷰티 상품이나 기호 식품 등이 주로 많은 것을 알 수 있음. 즉 자신만의 개성을 드러낼 만한 상품군일수록 본인만의 선호하는 브랜드가 있는 경우가 많음
- 반면 브랜드 선호도가 낮은 상품군들에는 청소/세탁/용실용품, 냉장/세탁가전제품, 문구/사무용품, 주방잡화, 컴퓨터 등 자신만의 개성을 드러내는 상품이라기보다는 생활과 관련된 생필품 등이 많은 것을 알 수 있음. 즉 생활에 필요한, 누구나 필요한 상품군일수록 본인만의 선호하는 브랜드가 딱히 없고 여러 브랜드를 둘러보면서 제품을 구매하는 경향이 강함

※ 변수 중요도가 가장 높은 상품군부터 내림차순으로 정렬된 표입니다.



끝까지 봐주셔서 진심으로 감사합니다