

수영복 판매 플랫폼 리뷰 데이터 분석 및 사이즈 추천 모델 생성



1조

김나홍 김민수 박영주 박정연 이윤수

목차

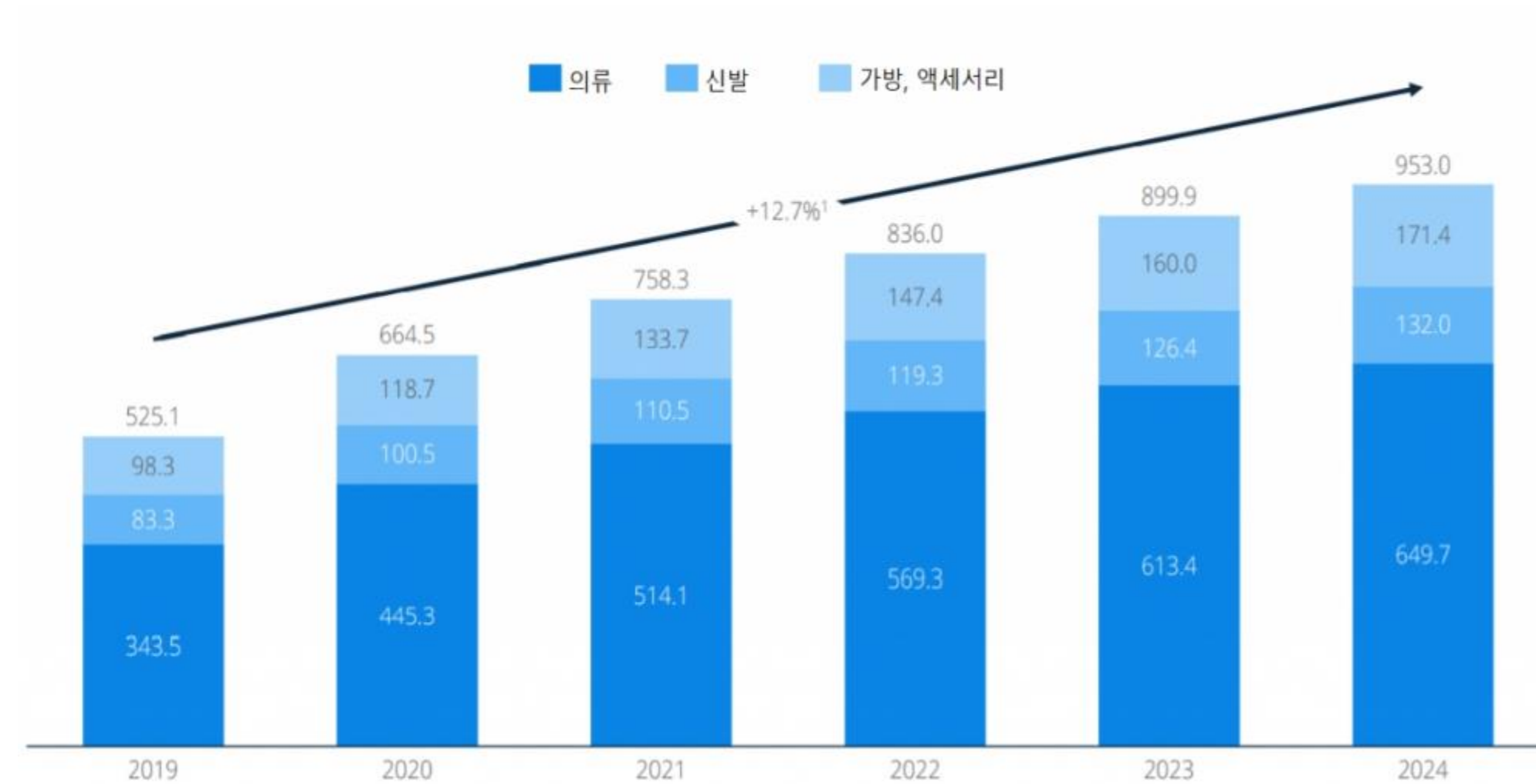
01 주제

02 데이터 소개

03 EDA/전처리

04 모델 학습 및 성능 비교

주제



최근 몇 년간 온라인 패션 시장은 빠르게 성장하고 있으며,
특히 모바일 커머스와 소셜 미디어 기반 소비의 확산은 소비자 행동의 변화를 촉진하고 있습니다.
그중에서도 패션 전자상거래는 기술 기반의 개인화 추천, 실시간 트렌드 반영, 데이터 기반 재고 관리 등을 통해 빠르게 진화하고 있습니다.

구매전환율을 높이는 콘텐츠, 리뷰

구매전환에 있어 리뷰에 대한 영향력은 갈수록 증가하고 있는 추세입니다

이용후기 확인 비율



잠재소비자 중 **97.2%** 가
구매 전 제품에 대한 리뷰를 확인함

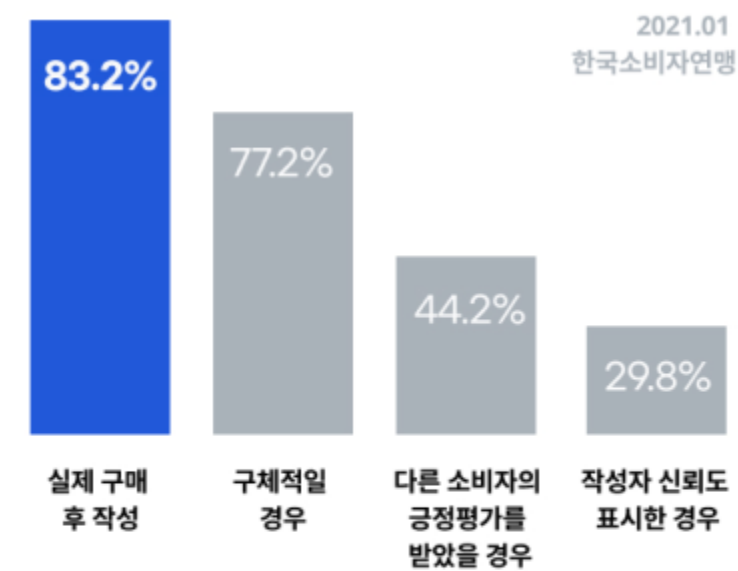
이용후기 없을 때 선택에 미치는 영향



잠재소비자 중 **72.4%** 가
구매 전 제품에 대한 리뷰가 없다면
대체로 구매하지 않음

신뢰하는 이용후기 유형 (중복 포함)

“실제 구매 후 작성된 이용후기”가
가장 신뢰할 만한 리뷰 라는
의견이 **83.2%**로 가장 높음



2021.01
한국소비자연맹

주제

문제 인식

온라인 쇼핑몰에서 사이즈 선택은
여전히 사용자 불만이 많은 영역이다.

특히, **수영복**의 경우
디자인과 소재 및 선호하는 착용감에 있어서
일반 기성복 사이즈와는 달라질 수 있는 특성이 있다.

실제로 많은 수영인들이
여러 번 구매 실패 후 자신에게 맞는 사이즈를 찾는다.

또한 사이즈 교환은
물류비 증가, 고객 만족도 저하의 원인으로
결국 **구매 이탈**로 이어진다.

온라인 의류 쇼핑몰 교환/반품률 비율



주제

목표

작성된 리뷰 데이터를 통해
최적의 사이즈 추천하는
머신러닝 모델 개발

Objective



비전

AI 사이즈 추천을 통해 플랫폼의 반품률 감소
+
소비자가 믿고 참고할 수 있는 추천 정확도 향상

Vision

데이터 소개

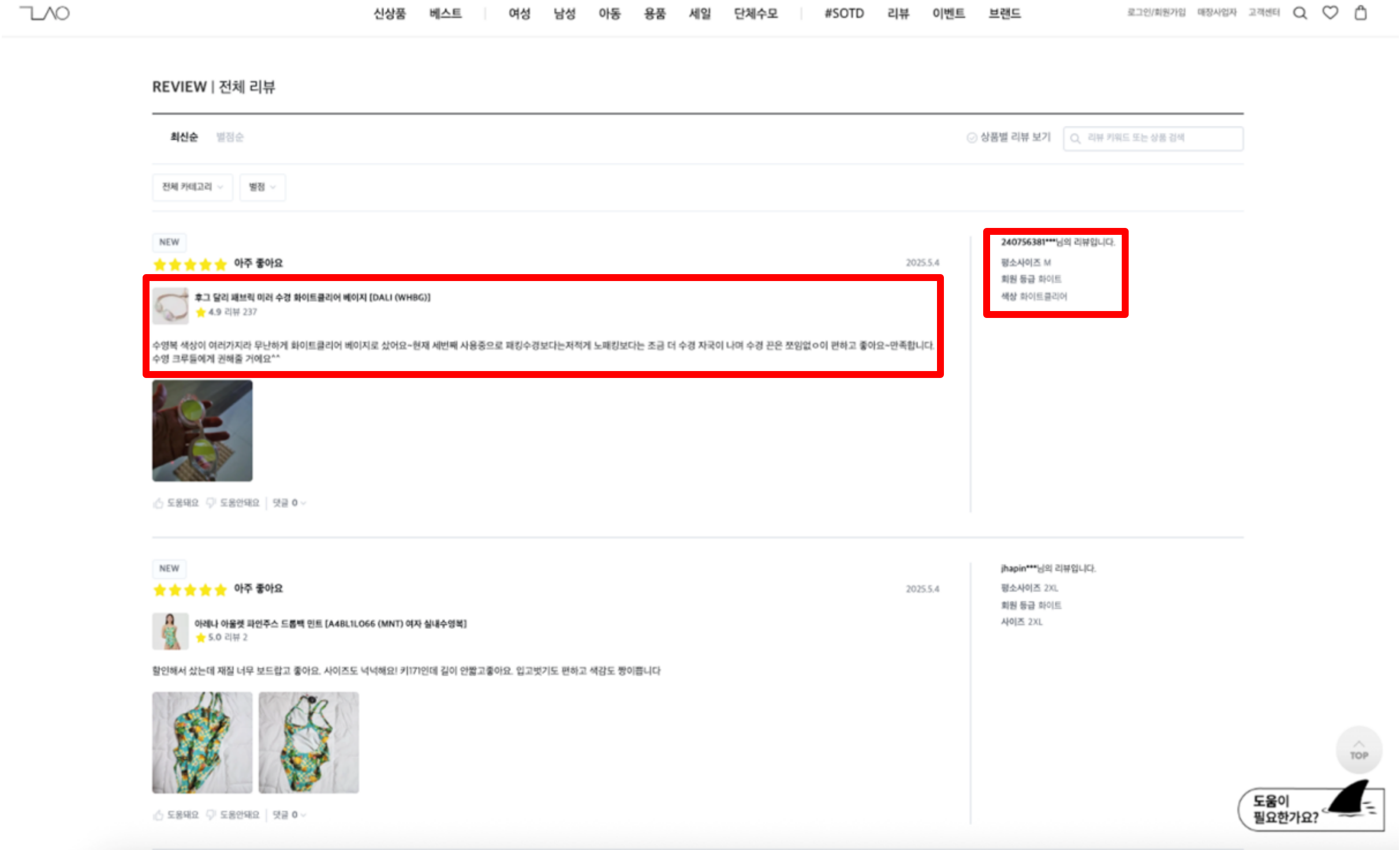
데이터 출처

가나스웜 플랫폼 리뷰 페이지 웹크롤링 실시

데이터 수집

실제 구매자 리뷰 기반 데이터 확보

제품명, 평소 사이즈, 구매 사이즈 등
구체적인 피팅 정보 수집



데이터 소개

1. 초기 데이터셋 (2022.04.01~2025.04.30)

	product_name	score	user_id	review_date	review_text	buy_size	usual_size	height	weight	member_level	has_photo	photo_count	gender
0	스웨이브 트월 포인팅 브이백 탄탄이 멀티 [SWB1WS022 (멀티) 여자 실내수영복]	4	226085929***	2022-05-31	사이즈표에 M 힙둘레 최대길이가 92길래, 2센치정도야~ 하고 시켜서 입었는데 처음...	M	M	157.0	55.0	브론즈	Y	1	Woman
1	나이키 스웜 로고 테이프 패스트백 블랙 [NESSB130-001 여자 실내수영복]	5	225934948***	2022-05-31	좀 작지만 옷에 몸을 맞춰보려고요 ㅎㅎ	M	S	162.0	52.0	화이트	N	0	Woman
2	나이키 스웜 앤서 솔리드 패스트백 탄탄이 미드나잇 네이비 [NESSA040-410A...	4	222749652***	2022-05-31	딱 베이직하고 예뻐요! 승모근부자라서,, 어깨끈이 좀 놀리긴 하지만 입다보면 나아질...	M	S	158.0	46.0	화이트	Y	2	Woman
3	스웨이브 세레니티 크로스에이백 라벤더 [SWB1WS010 (라벤더) 여자 실내수영복]	5	222749652***	2022-05-31	158/ 47인데, s살까하다가 m산건데 s샀으면 큰일날뻔했어요 ㅎㅎ아주 타이트하게...	M	S	158.0	46.0	화이트	Y	1	Woman
4	나이키 스웜 앤서 솔리드 패스트백 탄탄이 제트블랙 [NESSA040-006A 자수 ...	5	ks00***	2022-05-31	어깨가 끈으로 된 나이키 블랙 제품 입었었는데요, 가슴골이 보여서 실내 수영장에서 ...	XL	L	165.0	58.0	화이트	N	0	Woman

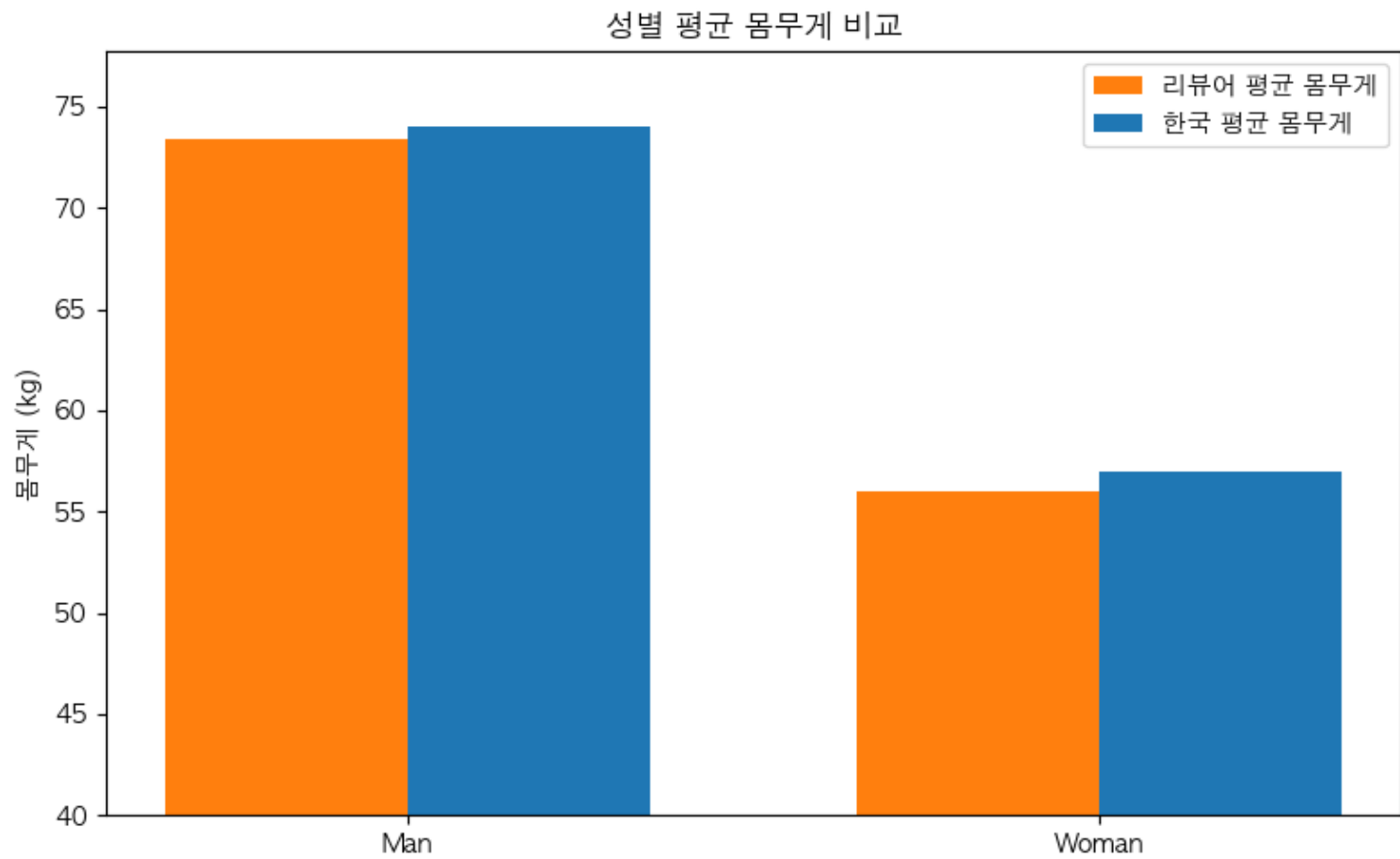
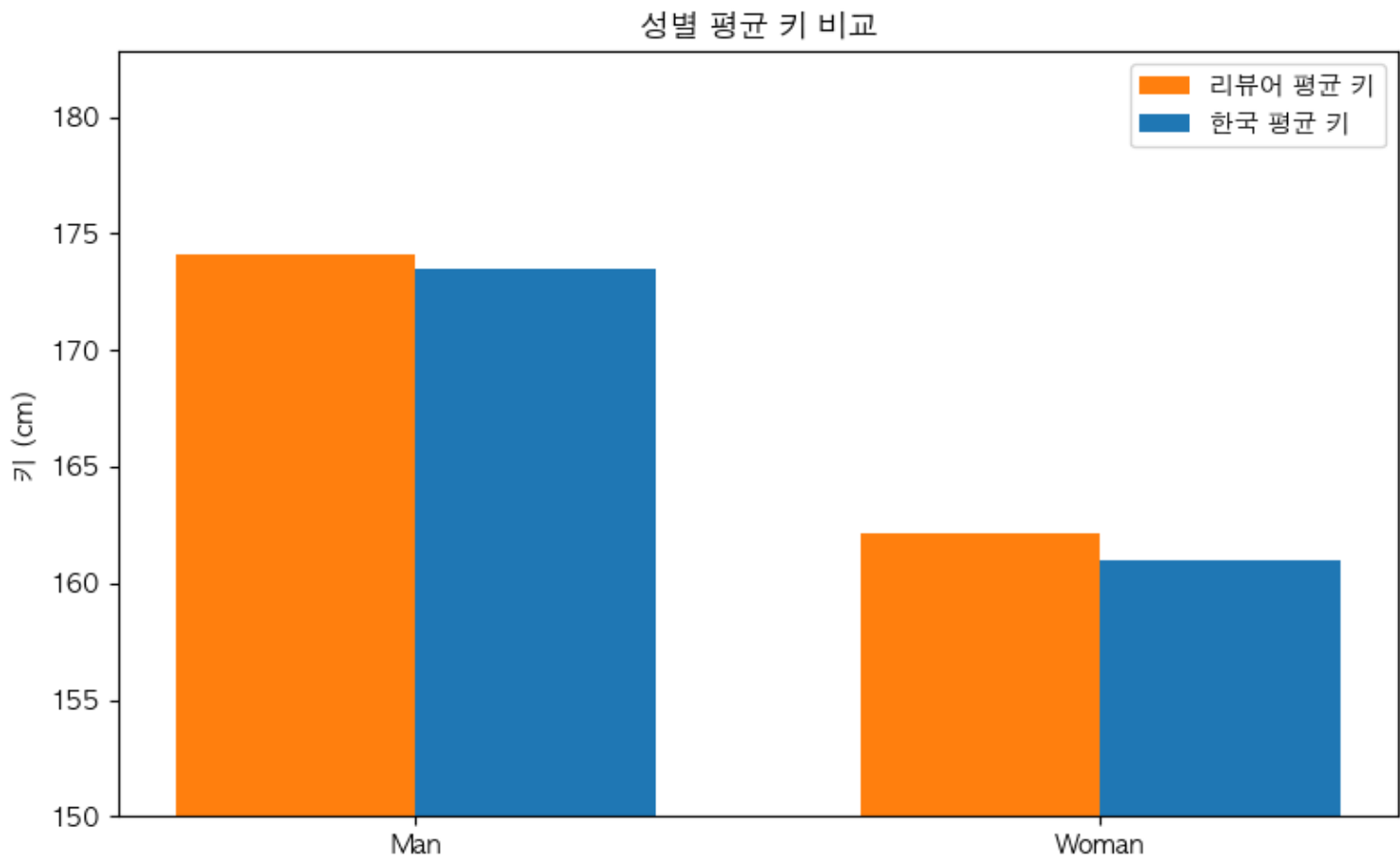
2. 데이터 전처리



usual_size와 buy_size가 두 사이즈 이상 차이가 나면 ‘선물’ 또는 ‘대리 구매’로 의심되어 제거
59,753개의 리뷰가 데이터 분석 및 모델 학습/테스트에 사용

데이터 분석

1.수영복 구매자(리뷰어) 평균 키/몸무게와 대한민국 평균 키/몸무게 비교



	리뷰어 평균 키	리뷰어 평균 몸무게	한국 평균 키	한국 평균 몸무게
Man	174.10	73.38	173.5	74.0
Woman	162.11	55.97	161.0	57.0

- 리뷰어 평균 키는 대한민국 평균보다 남성의 경우 0.6cm, 여성의 경우 1.11cm 더 큼
- 리뷰어 평균 몸무게는 대한민국 평균보다 남성의 경우 0.62kg, 여성의 경우 1.03kg 더 적음

데이터 분석 2. 리뷰 평점 그룹별 주요 키워드 (워드클라우드)

***불용어**: '좀', '조금', '너무', '많이', '잘', '그냥', '다', '더', '같아요', '정말', '그리고', '그래서',
'합니다', '했어요', '보기에', '거의', '이거', '이건', '근데', '해서', '있어요', '없어요', '정도'

☑ 수영복 선택에 있어서 **사이즈**가 가장 중요한 요소임을 확인할 수 있었다.

1~2점 워드클라우드 (불용어 제거)



4~5점 워드클라우드 (불용어 제거)

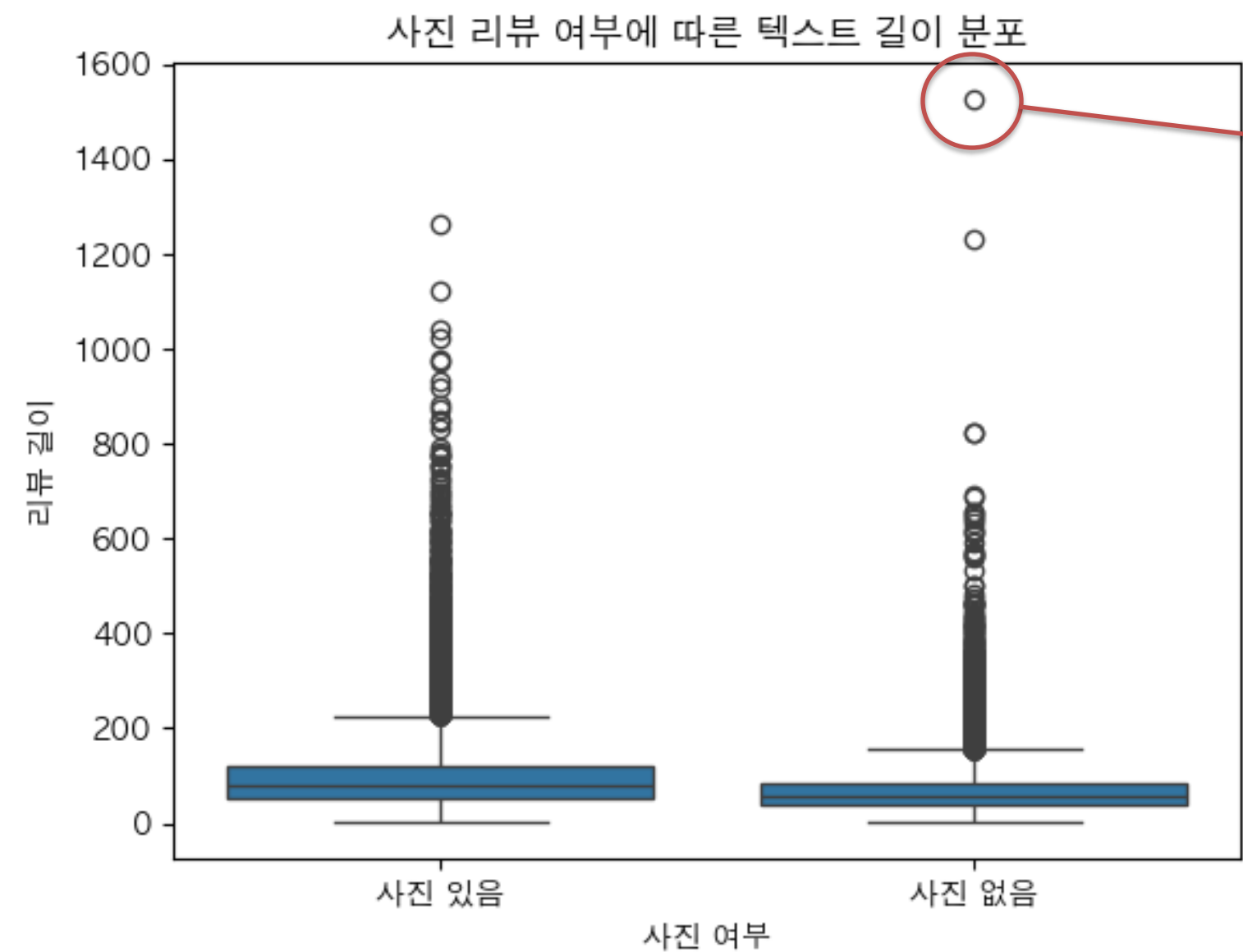


3점 워드클라우드 (불용어 제거)



데이터 분석

3. 리뷰 사진 첨부와 리뷰 텍스트 길이



[아웃라이어 데이터 분석]

53068 160/45-46 토르소 146-147(상체가 긴타입 다른분 보다 3-4cm 길다;) 가슴 80 엉덩이89-90 일반 여성복 44 또는 S, 26 착용 가나스원 상담사 사이즈 추천 26(키,몸무게 기준) 가나스원 사이즈 리뷰 28 추천 (키,몸무게, 평소사이즈 S기준) 검색 수린이가 수영복을 입어보지 않은 상태에서 구매하는게 가장 어려웠어요. 수영복 사이즈 미스로 교환까지해서 원하는 시간안에 입는게 어려울까봐 조마조마 샤크배송있어 든든하네요 수영복이 건조한 상태에서 타이트한게 맞고 물에 들어가면 적당해진다하는데 어느정도가 적당한 타이트함인지 모르겠음 교환해야할 수 있으 니 물은 못 들어감 ^^ 시착경험으로 고를때 기준점 1. 토르소 길이 2. 수영복 재질의 따른 어깨놀림 고려한 사이즈 3. 원하는 수영복 사이즈 후기참고 보통 나이키는 레이스백는 정사이즈를 패스 트백은 한치수 크게 입는게 좋다 추천한다 레이스백도 어깨놀림이 있어 한치수 크게입는걸 추천 그럼 나는 대입해보면 26이 정사이즈고, 한치수 업하면 28 그래서 28을 시켰습니다 가슴, 엉덩이, 토르소 길이가 신체보다 조금씩 수치가 커지고 물속은 다르다해서 걱정했지만 가슴과 엉덩이가 병병하거나 많이 남거나 크지 않았습니다 건조시나 물속이나 그 타이트함이 크게 차이나지 않는다는 걸 느꼈습니다. 물속에서 자리잡고 조금 편해진정도? 타이트함이 크게 다르지 않음 저는 28이 정사이즈였던 것 같으며 30 사이즈를 입으면 편안하겠다 생각했습니다 이유는 1. 어깨놀림이 해소될 것 같다 (수영시 아프거나 하진 않는데 벗으면 빨갛다) 2. 미들컷인데 애매하게 걸려서 가랑이가 찢인다 (크거나 작지는 않아 적당한데 전 좀 더 편한게 좋은타입) 팁 1.니들패치 필요함. BT가 티 나요 2.수린이 1일차라 한달 착용후 내몸에 맞아진다하니 그 생각하면 28을 시키는 것 추천 3. 아예 로우컷 입거나 하이컷이 좋을듯요 (Y존 골반뼈 나온 곳의 걸림이 없이 입는게 활동성 좋은듯 하다는 느낌을 받음) 4. 평소 S,85,90,26 44-55안에 들어가는 사이즈만 입었는데 수영복은 M이 정사이즈로 느껴짐(S시킨거 반품하거나 교환해봄) 수영복은 세부 사이즈는 다르나 탄탄한 재질 들은 큰 오차가 나는 옷이 아닌듯 합니다 26을 입나 28을 입나 엉덩이는 4개입 ㅋㅋ 부유방은 잘 정리해주세요 모델티데일 사진처럼 부유방 나오고 엉덩이 4개로 보임 정도의 차이는 있겠지만 수 영복 입은게 크게 다르지 않은듯요 온라인만 보며 스스로 공부해서 수영복 시켰는데 처음이시면 나이키사세요 수영하시는 분들 후기 보면 나이키가 제일 안늘어나고 오래입는다더라고요. 기본에 충실 함. 이뽀 수영복 잘 산 것 같아요 토르소 길이 꼭 재보시길! 수영복 알아보니 브랜드마다 사이즈 차이는 있지만, 저 보다 키크고 가슴커도 S 입는 분들은 입으시더라고요 타이트함과 편안함 등 취 향차이도 존재하는듯해요 전 몸매가 감당안되서 적당하게 타이트하고 맞는게 좋은것 같아요 부끄럽지만 일부러 신체사이즈 자세히 적었어요 처음이면서 안입어 보고 주문하시는 분들 있을 것 같아서요 도움되었으면 좋겠습니다

Name: review_text, dtype: object

:수영복 사이즈에 관한 상세한 내용이었고,
쇼핑몰의 **사이즈 리뷰 추천에 따라** 사이즈를 골랐으며 결과적으로 **만족**했다
는 내용

➡ 사진을 첨부한 경우 리뷰의 길이가 비교적 길어짐

has_photo	Average of text length
N	68.94
Y	99.47

↓

‘사이즈 추천’의 중요성

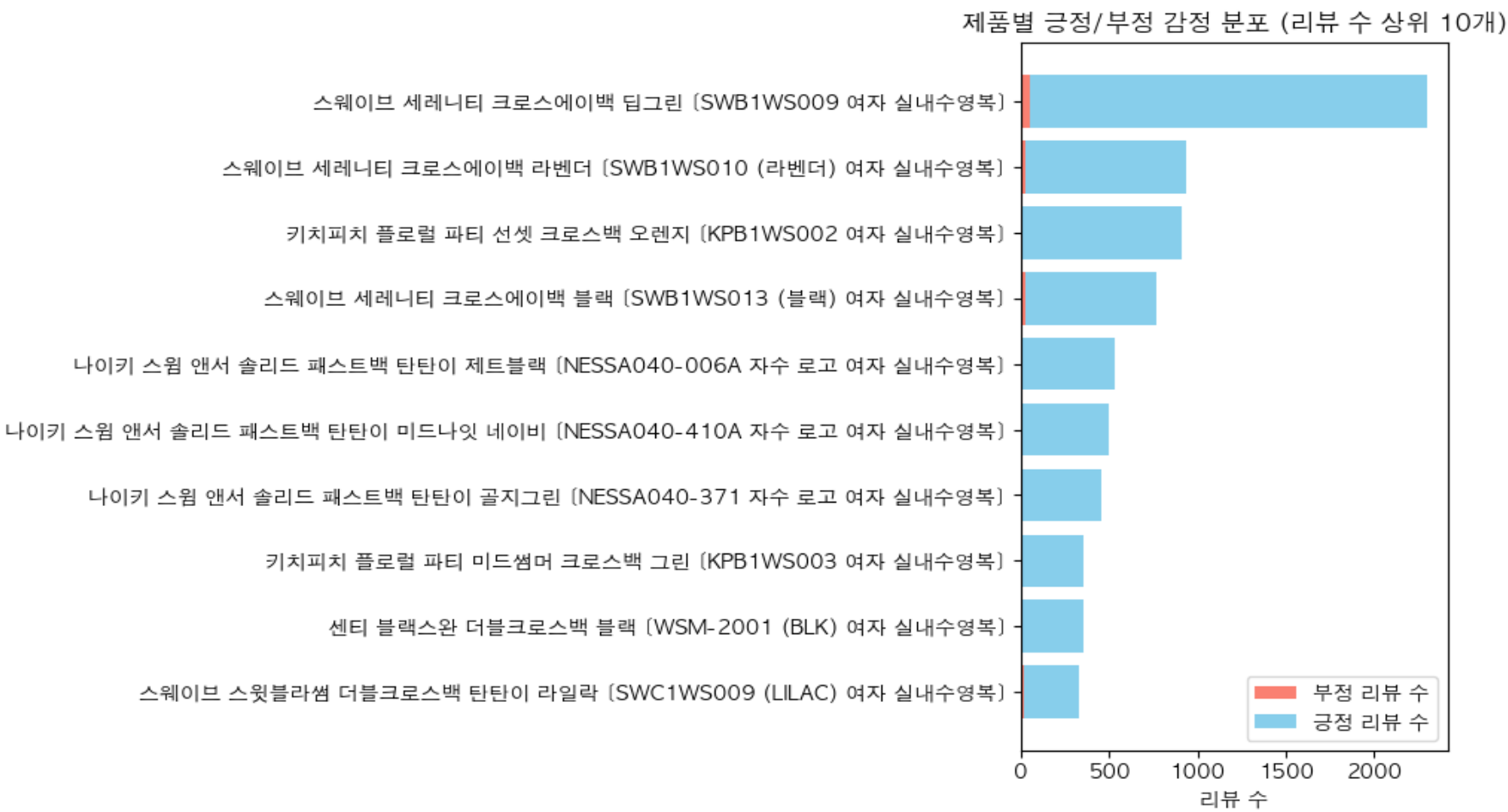
데이터 분석

4. 리뷰 긍정/부정 감정 분석

- 평점 4 이상 → 긍정(1) / 3점 이하 → 부정(0)으로 분류
- 제품별 평균 평점, 긍정 비율, 리뷰 수 집계

<리뷰 수 기준 상위 10개 제품의 긍정/부정 감정 분포>

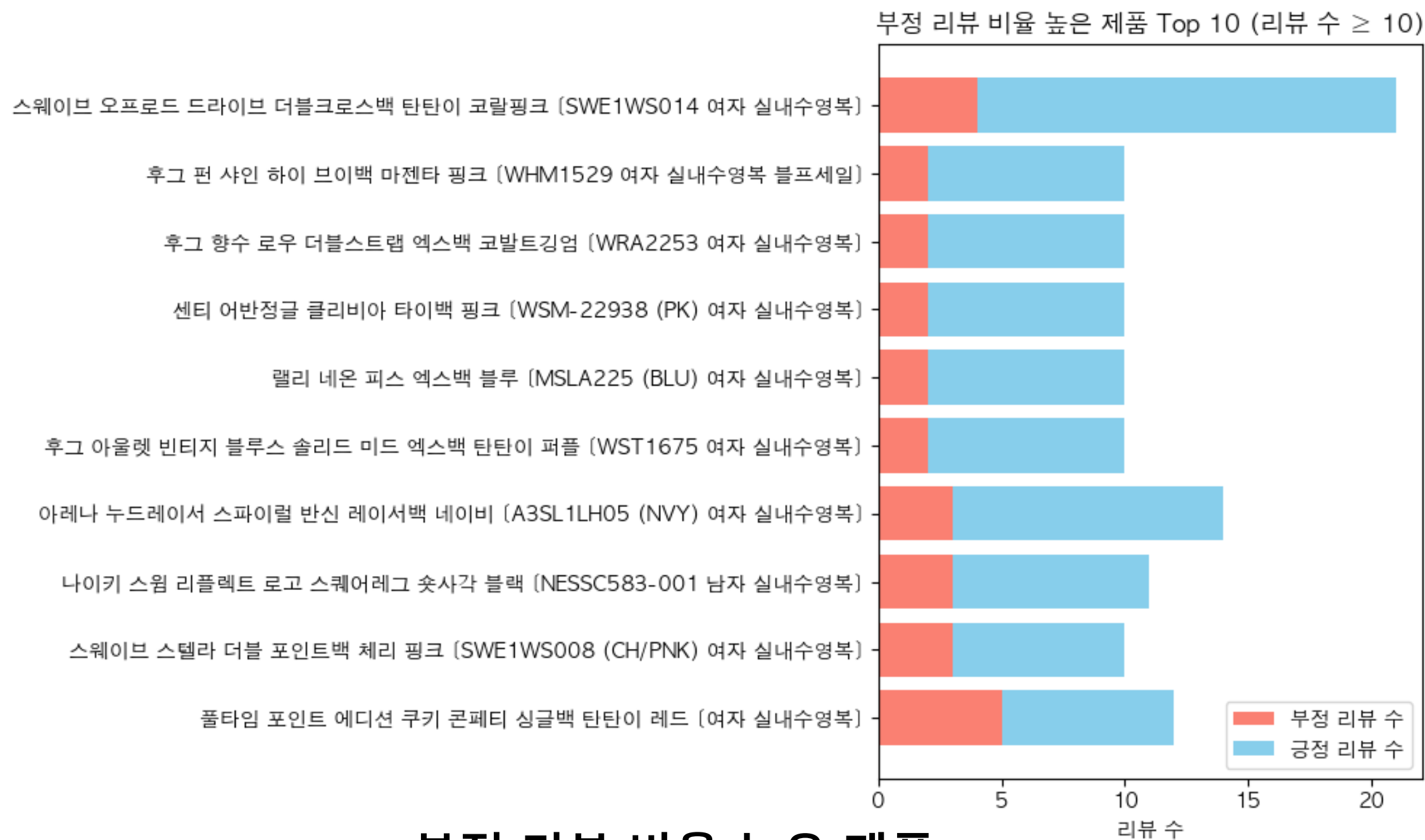
	평균평점	긍정비율	리뷰수
product_name			
스웨이브 세레니티 크로스에이백 딥그린 [SWB1WS009 여자 실내수영복]	4.867	0.980	2303
스웨이브 세레니티 크로스에이백 라벤더 [SWB1WS010 (라벤더) 여자 실내수영복]	4.885	0.977	938
키치피치 플로럴 파티 선셋 크로스백 오렌지 [KPB1WS002 여자 실내수영복]	4.901	0.988	909
스웨이브 세레니티 크로스에이백 블랙 [SWB1WS013 (블랙) 여자 실내수영복]	4.875	0.969	769
나이키 스웜 앤서 솔리드 패스트백 탄탄이 제트블랙 [NESSA040-006A 자수 로고 여자 실내수영복]	4.898	0.979	527
나이키 스웜 앤서 솔리드 패스트백 탄탄이 미드나잇 네이비 [NESSA040-410A 자수 로고 여자 실내수영복]	4.915	0.990	494
나이키 스웜 앤서 솔리드 패스트백 탄탄이 골지그린 [NESSA040-371 자수 로고 여자 실내수영복]	4.945	0.993	451
키치피치 플로럴 파티 미드섬머 크로스백 그린 [KPB1WS003 여자 실내수영복]	4.874	0.986	356
센티 블랙스완 더블크로스백 블랙 [WSM-2001 (BLK) 여자 실내수영복]	4.875	0.980	351
스웨이브 스윗블라썸 더블크로스백 탄탄이 라일락 [SWC1WS009 (LILAC) 여자 실내수영복]	4.795	0.945	327



- 1) 리뷰 건 수 1위는 2위보다 2배 이상의 리뷰가 존재
- 2) 상위 10개 제품 중 특정 제품들이 다수 존재함
- 3) 네번째로 리뷰가 많은 ‘스웨이브’의 경우 리뷰가 많은 다른 제품들에 비해 긍정비율이 다소 낮은 모습을 보임

데이터 분석

4. 리뷰 긍정/부정 감정 분석



<부정 리뷰 비율 높은 제품 Top 10>

(부정리뷰수 / 총리뷰수)

- 구매 대비 만족도가 낮은 제품을 확인하기 위해 부정 비율(부정리뷰수 / 총리뷰수)로 확인한 Top 10
- 마찬가지로, 부정 비율이 높은 제품 또한 여자 실내수영복이 가장 많았음(10개 중 9개)

데이터 분석

4. 리뷰 긍정/부정 감정 분석

	product_name	score	review_text
16160	플타임 포인트 에디션 쿠키 콘페티 싱글백 탄탄이 레드 [여자 실내수영복]	2	예쁘데..조금작아요... <u>조금사이즈</u> 가크면 좋겠네요
20125	플타임 포인트 에디션 쿠키 콘페티 싱글백 탄탄이 레드 [여자 실내수영복]	3	늘어남이란 1도 없는 탄탄이 수영복입니다. 소재는 핑키타는김? 아레나,나이키 처럼 늘어나는 느낌이 한개도없어서 입을때 당황했지만 쫌흔해서 오래입을수 있을것같아요. 색상은 모델착용샷 그대로 입니다. 썩한 느낌보다는 강렬한? 붉은빛이 도는 색입니다. 몸에 딱맞아서 쪼여지는 느낌은 좋은데 가슴선이 짧은 느낌이라 수영복을 자주 위로 올리게 되네요.
20610	플타임 포인트 에디션 쿠키 콘페티 싱글백 탄탄이 레드 [여자 실내수영복]	3	색상 이쁘고 사이즈도 잘 맞아요. 그런데 탄성이 없어 입고 벗을때 불편해요.
20670	플타임 포인트 에디션 쿠키 콘페티 싱글백 탄탄이 레드 [여자 실내수영복]	3	한치수 크게 사야한다해서 L로 샀어요 옆구리는 헐렁하고 엉덩이는 덜 덮히면서 헐렁해요 탄탄이인데 심한 탄탄이라 입을때 약간 불편해요 수영할때는 무리가 없어요 포인트 디자인이 약간 유아틱해서 귀여워 보인다고들 해요 (별로 안귀엽고 싶은데 ㅠㅠ)
22630	플타임 포인트 에디션 쿠키 콘페티 싱글백 탄탄이 레드 [여자 실내수영복]	3	S 했다가 M으로 바꿨는데도 <u>작은느낌</u> 이들어요 ㅠ 엉덩이가 말려올라가는느낌쓰... 수천들은 귀엽다고합니다 ^^

	product_name	score	review_text
22756	스웨이브 스텔라 더블 포인트백 체리 핑크 [SWE1WS008 (CH/PNK) 여자 실내수영복]	3	색상이 차이가 좀 있고 사이즈도 평소 착용하던 브랜드보다 <u>커서</u> 수선했네요
23575	스웨이브 스텔라 더블 포인트백 체리 핑크 [SWE1WS008 (CH/PNK) 여자 실내수영복]	1	물에 젖으면 검은안감 때문에 적색 이중천이라 기포가 많이 심함 입을 첫날....스타트데이인데....기포가 넘 심해 부끄러워 하루를 날려 먹음 입을수 없는 상태라 반품문의 했으나 이중천은 다 그래서 불가하다는 말만 듣고.. 그냥 버릴수 밖에 없었음
28493	스웨이브 스텔라 더블 포인트백 체리 핑크 [SWE1WS008 (CH/PNK) 여자 실내수영복]	3	호연보다 색이. 많이 어두워요 그럭저럭 입을만 뒷태가 예뻐요~

	product_name	score	review_text
2190	나이키 스웜 리플렉트 로고 스퀘어레그 숏사각 블랙 [NESSC583-001 남자 실내수영복]	3	착용감도 좋고 로고나 디자인도 예쁘고 다 좋은데.. 다른후기처럼 앞 로고가 스티커처럼 일어나네요.. 일부러 탈수기도 한번도 안돌리고 구매한지 10일정도 밖에 안돼서 실착한 횟수도 많지 않는데.. 불량인건지 원래 이런제품인건지요..?
3152	나이키 스웜 리플렉트 로고 스퀘어레그 숏사각 블랙 [NESSC583-001 남자 실내수영복]	1	이거 불량인가요\n로고가뜰어져요 ..
39949	나이키 스웜 리플렉트 로고 스퀘어레그 숏사각 블랙 [NESSC583-001 남자 실내수영복]	1	로고 두번입고 바로 떨어진 물속이라 떨어지는게 당연한데..2번입고떨어짐..

	product_name	score	review_text
10400	아레나 누드레이서 스파이럴 반신 레이서백 네이비 [A3SL1LH05 (NVY) 여자 실내수영복]	3	평소 L 입어서..싸이즈 95로 알고있었는데, 90으로 왔더라고요. 구매시 맨아래 참고표를 못봤네요~ 입어보니 원단은 편하긴한데~ 전체적으로 몸을 잡아주는 느낌없고, 다리 아래 시보리 부분이 꼭 쪼이네요~
11306	아레나 누드레이서 스파이럴 반신 레이서백 네이비 [A3SL1LH05 (NVY) 여자 실내수영복]	3	좀얇은감이 있어요 착용했을때 잡아주는게 좀없어요\n한사이즈 <u>작게</u> 입어도 될거같아요
40701	아레나 누드레이서 스파이럴 반신 레이서백 네이비 [A3SL1LH05 (NVY) 여자 실내수영복]	3	배송도 빠르고\n생각보다 디자인도\n괜찮고 착용감도 좋았어요\n항상 아레나 애용하지만\n만족입니다

	product_name	score	review_text
18299	후그 아울렛 빈티지 블루스 솔리드 미드 엑스백 탄탄이 퍼플 [WST1675 여자 실내수영복]	3	가슴과 허벅지 라인 탄탄. 어깨도 끈이 탄탄해서 유연하지 않지만 전반적으로 편한 제품.
21873	후그 아울렛 빈티지 블루스 솔리드 미드 엑스백 탄탄이 퍼플 [WST1675 여자 실내수영복]	2	무난한 기본스탈탄탄이제품 한사이즈크 <u>게</u> 입어도 될듯합니다

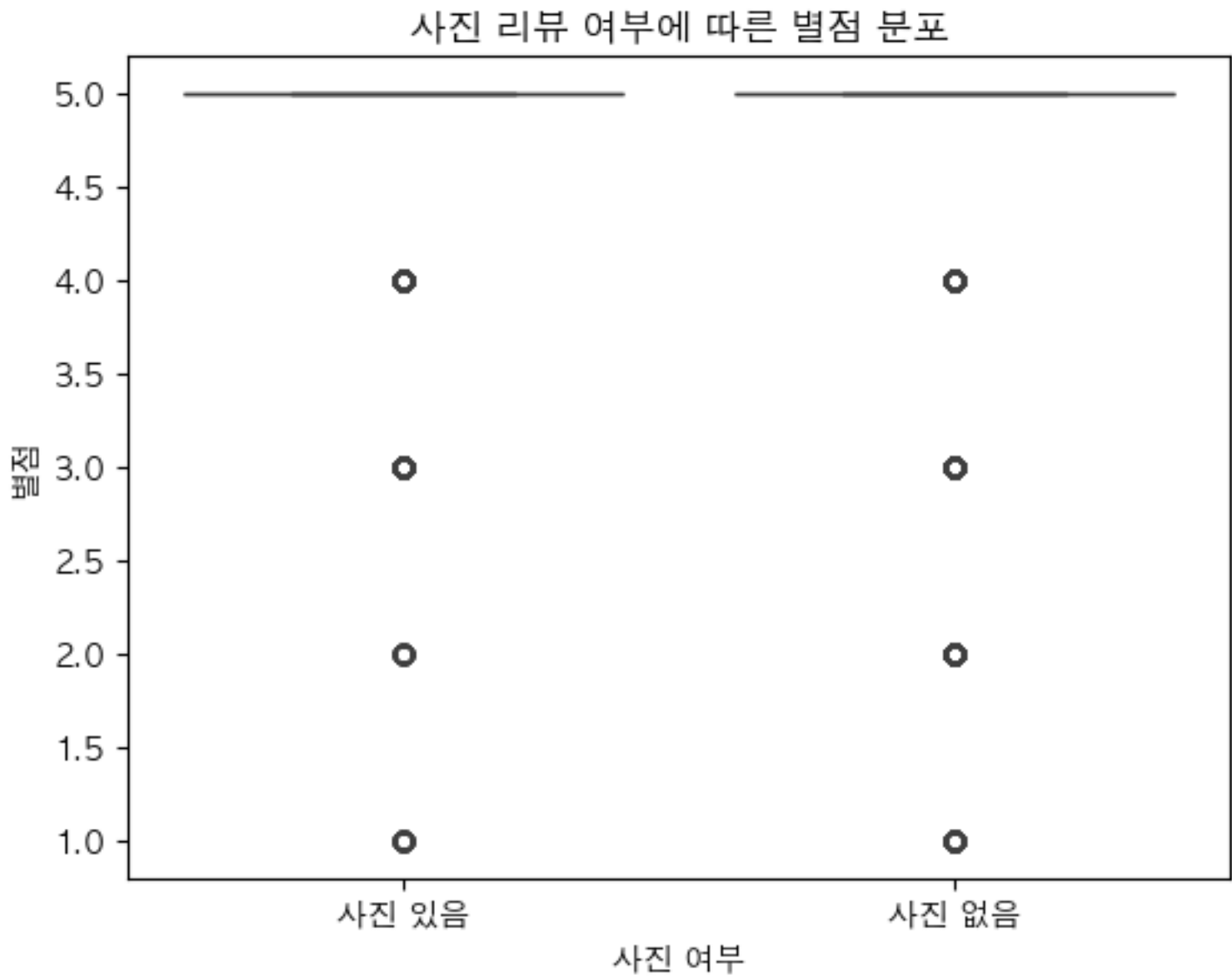
부정 비율이 높은 제품의 리뷰 내용 분석 결과,
사이즈 불만족과 **내구성**에 대한 내용임을 확인



고객 만족도를 향상 시키기 위해서는
정확한 사이즈 추천과 **수영복의 품질 향상**이 중
요

<부정 리뷰 비율 높은 제품 Top 10중 일부 제품 리뷰 랜덤 확인>

데이터 분석 5. 별점에 따른 경향성



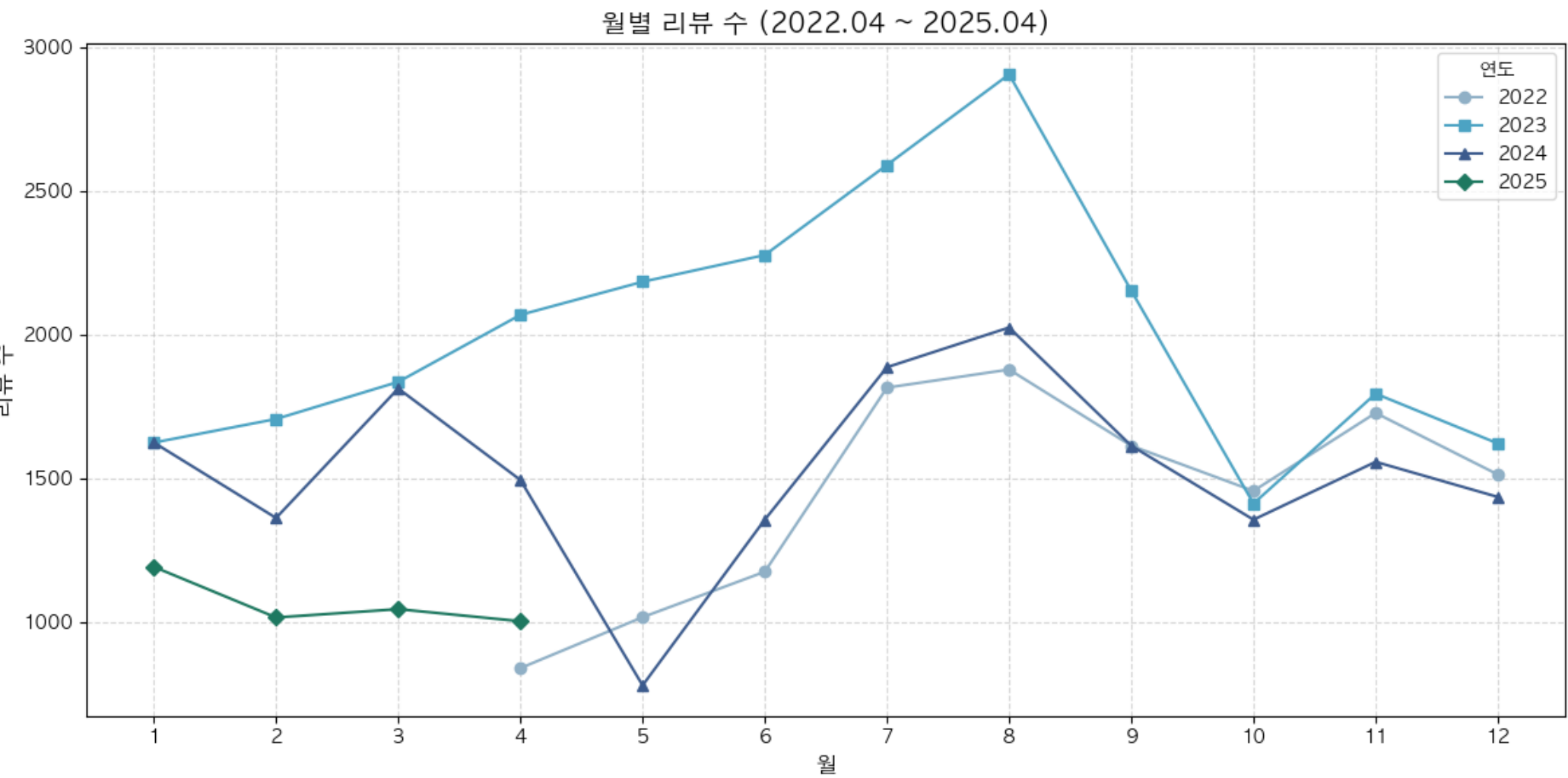
score	count	avg_review_length	avg_height	avg_weight	avg_bmi	사진 첨부 비율
1	208	120.6	163.8	58.8	21.841	0.423
2	233	103.7	162.5	58.1	21.920	0.318
3	986	91.2	162.7	57.4	21.608	0.305
4	4,002	90.1	163.2	57.6	21.592	0.390
5	55,471	81.9	163.6	58.2	21.651	0.462

➡ 다수가 5점을 줌

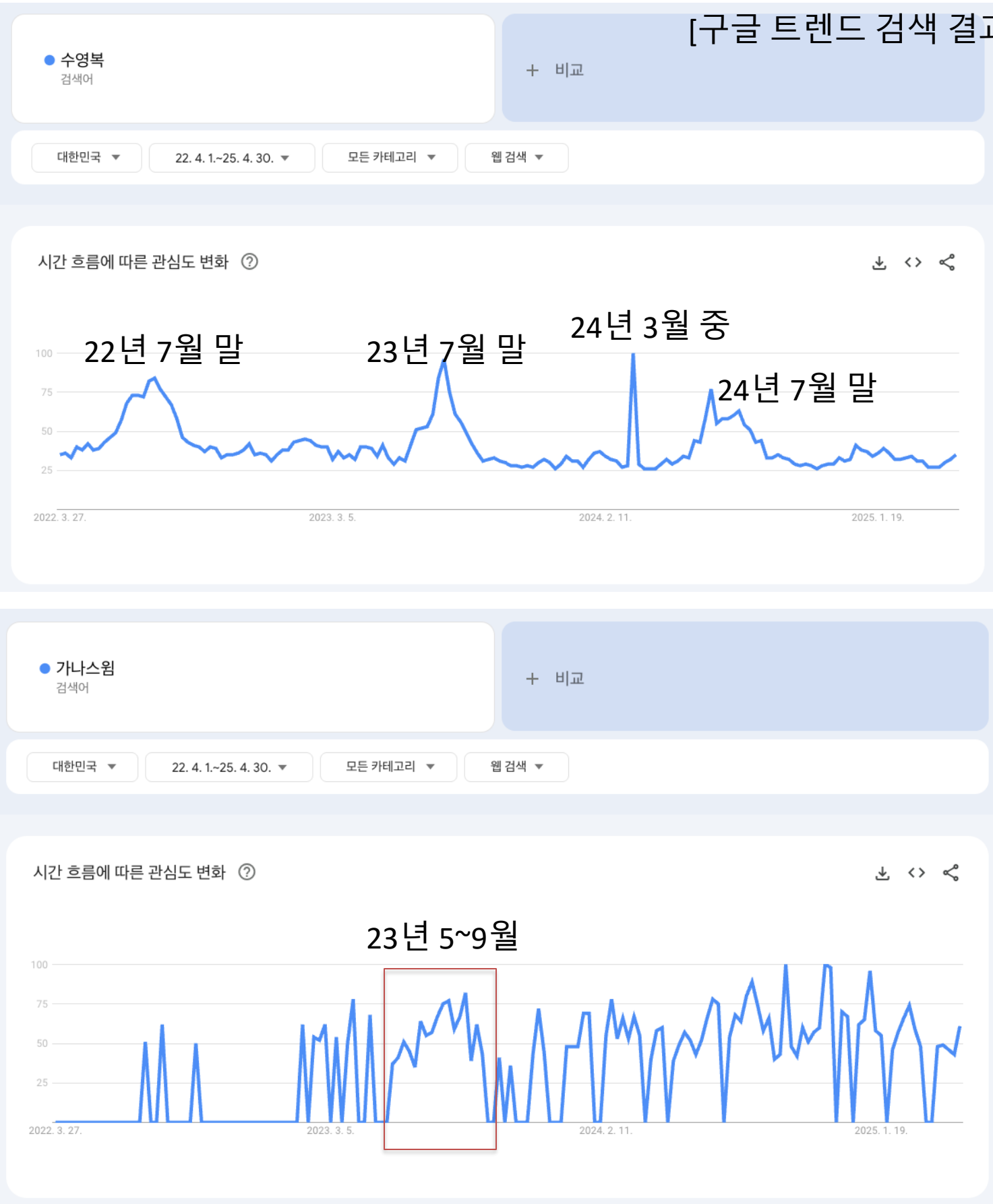
has_photo	Average of score
N	4.86
Y	4.90

데이터 분석

6. 월별 리뷰 분석



[구글 트렌드 검색 결과]



기계학습

1. 데이터

	product_name	height	weight	size_num	usual_size_num	extracted_brand	gender_le	스웨이브	나이키	후그	센티	키치피치	르망고
0	스웨이브 트월 포인팅 브이백 탄탄이 멀티 [SWB1WS022 (멀티) 여자 실내수영복]	157.0	55.0	4	4.0	스웨이브	1	True	False	False	False	False	False
1	나이키 스웜 로고 테이프 패스트백 블랙 [NESSB130-001 여자 실내수영복]	162.0	52.0	4	3.0	나이키	1	False	True	False	False	False	False
2	나이키 스웜 앤서 솔리드 패스트백 탄탄이 미드나잇 네이비 [NESSA040-410A...	158.0	46.0	4	3.0	나이키	1	False	True	False	False	False	False
3	스웨이브 세레니티 크로스에이백 라벤더 [SWB1WS010 (라벤더) 여자 실내수영복]	158.0	46.0	4	3.0	스웨이브	1	True	False	False	False	False	False
4	나이키 스웜 앤서 솔리드 패스트백 탄탄이 제트블랙 [NESSA040-006A 자수 ...	165.0	58.0	6	5.0	나이키	1	False	True	False	False	False	False
...
49547	후그 아울렛 향수 딸기 미드 U 리본크로스 엑스백 코랄 [WSA1728 여자 실내수영복]	165.0	55.0	5	5.0	후그	1	False	False	True	False	False	False
49548	키치피치 플로럴 파티 선셋 크로스백 오렌지 [KPB1WS002 여자 실내수영복]	160.0	58.0	5	5.0	키치피치	1	False	False	False	False	True	False
49549	스웨이브 세레니티 크로스에이백 블랙 [SWB1WS013 (블랙) 여자 실내수영복]	157.0	55.0	4	4.0	스웨이브	1	True	False	False	False	False	False
49550	스웨이브 세레니티 크로스에이백 딥그린 [SWB1WS009 여자 실내수영복]	159.0	55.0	5	5.0	스웨이브	1	True	False	False	False	False	False

- product_name : 상품명 (띄어쓰기로 구분하여 가장 앞 단어가 브랜드명)
- height : 키
- weight : 체중
- size_num : 구매자가 구매한 사이즈의 mapping value
- usual_size_num : 구매자가 평소에 착용하는 사이즈의 mapping value
- extracted_brand : 상품명에서 추출한 수영복의 브랜드명
- gender_le : 성별의 mapping value
- 스웨이브, 나이키, 후그, 센티, 키치피치, 르망고 : 리뷰가 가장 많은 브랜드 TOP6

기계학습 1. 데이터

```
df['extracted_brand'] = df['product_name'].str.split().str[0]
brand_counts = df['extracted_brand'].value_counts()
brand_counts
```

✓ 0.6s

extracted_brand	
스웨이브	11368
나이키	10695
후그	9307
센티	6876
키치피치	5815
르망고	5537
아레나재팬	3440
아레나	2502
애디블	2097
렐리	1318
스피도	597
미즈노	292
루프루프	158
폴타임	157
스포티	134
저스트웨어	124
아이러브비키니	113
이스트본	91
리얼리굿스웜	87
폴핀	84
아만지	33
아디다스	23
움파	19
줄린	17
티어	11
센티아디오스	3
배럴	2

Name: count, dtype: int64

<리뷰 수 TOP 6 브랜드 추출 과정>

```
df_filtered = df_filtered.reset_index(drop=True)
df_filtered
```

✓ 0.0s

	height	weight	size_num	usual_size_num	gender_le	스웨이브	나이키	후그	센티	키치피치	르망고
0	157.0	55.0	4	4.0	1	True	False	False	False	False	False
1	162.0	52.0	4	3.0	1	False	True	False	False	False	False
2	158.0	46.0	4	3.0	1	False	True	False	False	False	False
3	158.0	46.0	4	3.0	1	True	False	False	False	False	False
4	165.0	58.0	6	5.0	1	False	True	False	False	False	False
...
48577	161.0	58.0	4	4.0	1	False	False	False	False	True	False
48578	165.0	55.0	5	5.0	1	False	False	True	False	False	False
48579	160.0	58.0	5	5.0	1	False	False	False	False	True	False
48580	157.0	55.0	4	4.0	1	True	False	False	False	False	False
48581	159.0	55.0	5	5.0	1	True	False	False	False	False	False

48582 rows × 11 columns

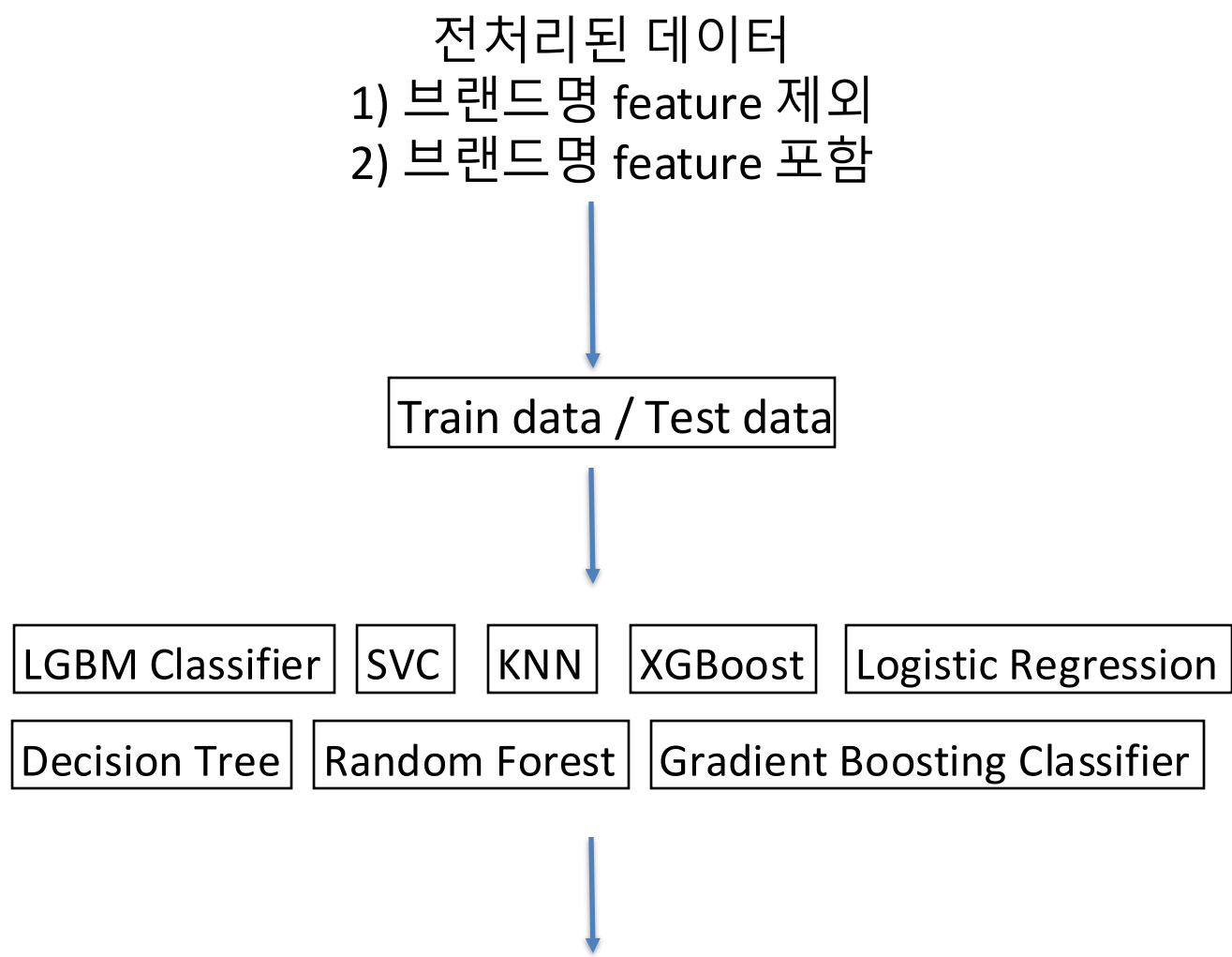
<최종 데이터>

- 총 48,582개의 행과 11개의 columns
- Feature : height, weight, usual_size_num, gender_le, 브랜드 6개 항목
- Target : size_num



브랜드 별 사이즈 가이드라인이 다를 수 있음을 고려하여 feature에 추가하였으나, 확실한 결과 확인을 위해 브랜드를 고려하였을 경우와 고려하지 않았을 경우 비교

기계학습 3. 모델



공통적인 test accuracy 결과:
브랜드명 feature 제외 < 브랜드명 feature 포함

Model	브랜드명 feature 제외	브랜드명 feature 포함
LGBM Classifier	0.624	0.685
SVC	0.583	0.635
KNN	0.610	0.672
XGBoost	0.616	0.662
Logistic Regression	0.596	0.629
Decision Tree	0.627	0.672
Random Forest	0.626	0.679
Gradient Boosting Classifier	0.622	0.677

```
from sklearn.model_selection import train_test_split
from lightgbm import LGBMClassifier
from sklearn.metrics import accuracy_score

X = brand_df[['height_mms','weight_mms','usual_size_num','gender_le']]
y = brand_df['buy_size']

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2, random_state=42)

lgbm = LGBMClassifier(random_state=42)
lgbm.fit(X_train,y_train)

y_pred_tr = lgbm.predict(X_train)
y_pred_ts = lgbm.predict(X_test)

print(f'trian:',accuracy_score(y_train,y_pred_tr))
print(f'test:',accuracy_score(y_test,y_pred_ts))
```

기계학습 3. 모델

1. 불균형 데이터 해결: SMOTE

Size	Before SMOTE	After SMOTE
XS	278	17,317
S	4,774	
M	15,556	
L	17,317	
XL	8,765	
2XL	1,690	
3XL	212	

→ SMOTE 적용 전: 0.68
적용 후 accuracy: 0.78

2. Hyperparameter 조정

```
param_grid = {
    'num_leaves': [31, 50],
    'max_depth': [-1, 10],
    'learning_rate': [0.05, 0.1],
    'n_estimators': [100, 300],
    'min_child_samples': [20, 50],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0]
}
```

```
grid_search = GridSearchCV(
    estimator=lgbm,
    param_grid=param_grid,
    scoring='accuracy',
    cv=3,
    n_jobs=-1,
    verbose=1
)
```

→ Test accuracy: 0.79

프로젝트 한계 및 과제

📌 한계 및 과제

1. 모델 예측 작업에 필요한 더 다양한 feature를 확보하기 힘들었습니다.
허리 둘레, 엉덩이 둘레 등 구체적인 신체 정보를 확보할 수 있었다면 더 정확한 모델을 구현할 수 있었을 것 같다는 아쉬움이 남습니다.
2. 데이터 수집의 한계를 느꼈습니다.
리뷰를 작성한 사람의 연령대, 구매일로부터 얼마나 지나서 리뷰를 작성했는지 등의 정보가 있었다면 더 재밌는 분석도 가능했을 것 같습니다.
3. Gemini api를 통해 리뷰를 분석하고 target값을 업데이트하고자 했습니다.
하지만 사이즈에 대한 만족/불만족을 단순히 구분해내기는 어려웠습니다. 더 고도화된 언어모델을 활용했다면 어땠을까 했습니다.
4. 현재 저희가 구현한 모델은 입력값으로 usual_size가 필요합니다. 완전 처음 수영복을 사는 구매자에게는 여전히 수영복 사이즈 추천이 어렵습니다.
usual_size 피처가 제외된 모델은 성능이 떨어지는 것을 확인했기 때문에 이에 대한 고민이 필요합니다.
5. 브랜드명 feature가 있을 때 성능이 높은 것을 확인했습니다. 이 뜻은 리뷰가 적거나 없는 브랜드에 대해서는 성능이 떨어질 수 있다는 것으로 해석됩니다.
대안으로 다른 브랜드에서 동일한 사이즈의 수영복에 대한 상세 치수가 동일하다면 해당 브랜드 모델에 대한 예측값을 가져와서 사용하는 것은 어떨까