

SSG SHOP 상품 분류 자동화 모델 구축 및 사이즈 적정성 분석 보고서

작성자: 이선우

작성일: 2025년 12월 5일

1. 서론

1.1 연구 배경 및 목적

온라인 편집숍 SSG SHOP은 하반기 대규모 입고 예정인 Black/Blue 계열 상품의 효율적 운영과 사이즈 관련 고객 불만(VOC) 해소를 위해 본 프로젝트를 수행한다. 첫째, 수작업 분류 방식의 한계를 극복하기 위한 머신러닝 기반 자동 분류 모델 구축, 둘째, 브랜드별 표기 사이즈와 실측 치수 불일치로 인한 반품·클레임 감소를 위한 사이즈 특성 분석을 목표로 한다.

1.2 데이터 및 분석 과제

데이터 구성

- 이미지 데이터: Black/Blue 계열 5개 품목(dress, pants, shirt, shoes, shorts) 이미지
- shirts.csv: 브랜드별 표기 사이즈 및 실측 치수(Brand Name, Type, Size, Chest(cm), Front Length(cm), Across Shoulder(cm))

수행 과제

- **Task 1:** 5개 카테고리 자동 분류 모델 구축 (전처리-알고리즘 비교-최적 모델 제안)
- **Task 2:** 셔츠 사이즈 적정성 분석
 - Q1: 동일 Brand Size 내 브랜드 간 실측 치수 차이 검증
 - Q2: 소재(Type)가 실제 치수에 미치는 영향 분석
 - Q3: 특정 두 브랜드 간 동일 사이즈 기준 실측 차이 비교

2. Task 1: 이미지 기반 의류 분류 모델

2.1 데이터 개요 및 전처리

데이터 구성

- 총 5,993개 이미지, 10개 클래스(black/blue × dress/pants/shirt/shoes/shorts)
- 클래스별 불균형 존재: 최소 299개(blue_shorts) ~ 최대 871개(black_pants)
- Train:Val:Test = 6:2:2 비율로 stratified split

전처리 전략

- Resize: 224×224, ImageNet 정규화 적용
- Train 증강: RandomHorizontalFlip, RandomRotation, ColorJitter
- WeightedRandomSampler로 클래스 불균형 보정

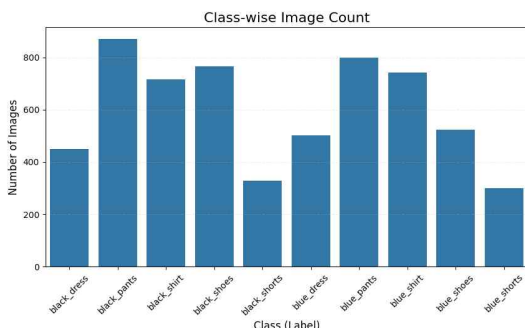


그림 1 클래스별 이미지 수 분포

2.2 Baseline 모델: 색 히스토그램 + HOG 특징 기반 분류

특징 추출

- 색 히스토그램(RGB 각 채널 16 bins) + HOG(9 orientations, 8×8 pixels/cell)
- 총 8,148차원 특징 벡터, StandardScaler로 정규화

모델 성능 (Test Set)

모델	Accuracy	Macro F1	특징
Logistic Regression	0.671	0.644	계산 효율적, 해석용이
Linear SVM	0.575	0.540	수렴 불안정

2.3 CNN 모델: ResNet-18 전이학습

모델 구성

- ImageNet 사전학습 ResNet-18 사용
- 최종 FC layer만 10-class 분류용으로 교체
- Adam optimizer (lr=1e-4), CrossEntropyLoss

학습 전략

- Max 8 epochs, Early Stopping (patience=2)
- Best model: Epoch 5 (Val Loss: 0.0712)

최종 성능 (Test Set)

모델	Accuracy	Macro F1	향상률
Baseline (LogReg)	0.671	0.644	-
CNN (ResNet-18)	0.969	0.968	50.2%

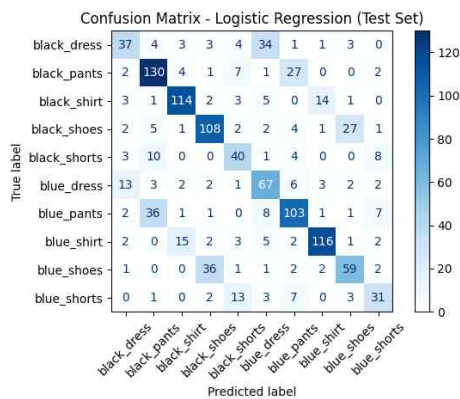


그림 2 Baseline(LogReg)

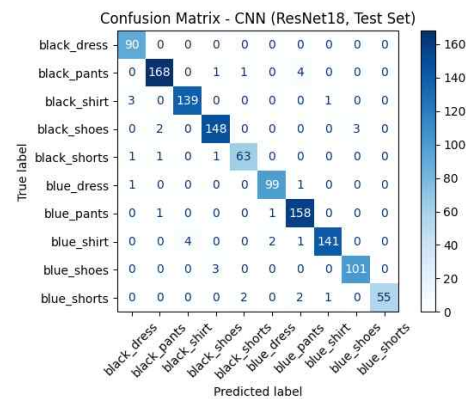


그림 3 CNN

2.4 결론

전통적 특징 기반 모델은 색상·텍스처 정보만으로 67% 정확도 달성했으나, CNN은 심층 특징 학습을 통해 97% 정확도로 약 50% 성능 향상을 보였다. 의류 형태·패턴·구조적 정보를 효과적으로 학습한 CNN이 실용 적용에 적합함을 확인했다.

3. TASK 2: shirts 데이터 통계 분석

3.1 데이터 개요 및 전처리

원본 데이터

- 834개 관측치, 7개 변수 (Brand Name, Type, Size, Brand Size, 실제 치수 3개)

- Across Shoulder 결측 5개, Type 불일치(Cotton/cotton)

전처리

- Type 통합, 결측 제거, IQR 기반 이상치 제거
- 최종 771개 샘플 사용

상관관계 분석

- Size ↔ Across Shoulder: 0.80 (강한 상관)
- Size ↔ Front Length: 0.50 (중간 상관)
- Size ↔ Chest: 0.27 (약한 상관)

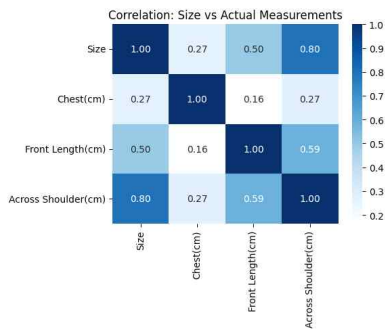


그림 4 치수 간 상관계수

3.2 Q1: 동일 Brand Size라도 브랜드별 실제 치수 차이가 존재하는가?

분석 방법

- One-way ANOVA: Chest(cm)/ Across Shoulder(cm) ~ Brand Name (Brand Size별)
- 대상: L, XL 사이즈 (표본수 충분)

ANOVA 결과

Brand Size	변수	F-statistic	p-value	해석
L	Chest(cm)	22.59	0.043	차이 있음
XL	Chest(cm)	205.63	0.005	차이 있음
L	Across Shoulder(cm)	59.24	0.017	차이 있음
XL	Across Shoulder(cm)	28.97	0.034	차이 있음

시각화: 빈도 상위 10개 브랜드 Stripplot

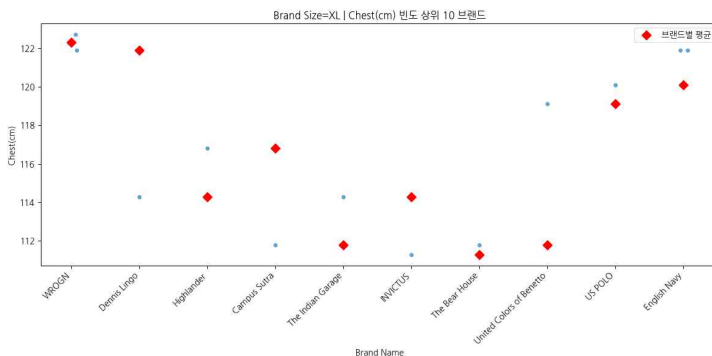


그림 5 Brand Size:XL/Chest(cm)의 브랜드별 치수 분포

결론

동일 사이즈 표기(L, XL)에도 브랜드 간 실제 치수는 통계적으로 유의한 차이 존재. 소비자는 사이즈 표기만으로 핏을 판단하기 어렵다.

3.3 Q2: 소재(Type)가 실제 치수에 영향을 주는가?

단순 ANOVA 결과 (Type_clean: Cotton vs Polyester)

종속변수	F-statistic	p-value	해석
Chest(cm)	1.52	0.217	차이 없음
Front Length(cm)	0.20	0.657	차이 없음
Across Shoulder(cm)	0.75	0.388	차이 없음

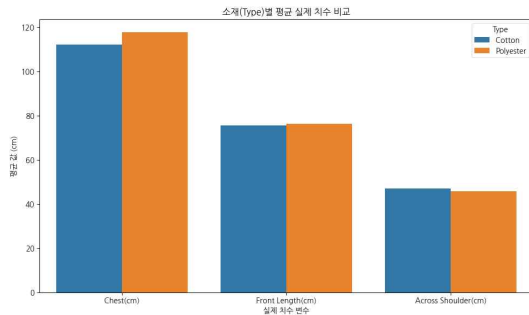


그림 6 소재별 평균 치수 Barplot

-> 소재만으로는 유의한 차이 없음

Brand Size 통제 후 다중회귀

『Chest(cm) ~ Type_clean + Brand Size(더미)』

- Type_clean[Polyester] 계수: 5.199 (p = 0.022)
- 해석: Brand Size를 통제하면 소재 효과가 유의하게 나타남

결론

Brand Size가 치수 변동의 주요 원인이므로, 이를 통제해야 소재의 순수 효과(약 5cm 차이)가 드러난다. 사이즈를 고려하지 않으면 소재 효과가 과소평가된다.

3.4 Q3: 서로 다른 두 브랜드 간 동일 Size 기준 실제 치수 차이가 있는가?

분석 대상

- 상위 2개 브랜드: WROGN vs English Navy
- Brand Size: L, XL, M

t-test 결과

Brand Size	변수	mean_WROGN	mean_English Navy	diff	p-value	해석
L	Chest(cm)	116.6	116.8	-0.2	0.860	차이 없음
L	Across Shoulder(cm)	47.7	51.5	-4.1	0.004	차이 있음
XL	Across Shoulder(cm)	49.2	53.9	-4.7	0.007	차이 있음
XL	Across Shoulder(cm)	44.1	47.5	-3.5	0.010	차이 있음

결론

두 브랜드는 Chest(가슴둘레)에서는 유사하나, Across Shoulder(어깨너비)에서 약 3-5cm의 유의한 차이 존재. 동일 사이즈라도 브랜드별 어깨 핏이 다를 수 있어 사이즈 선택 시 주의 필요.

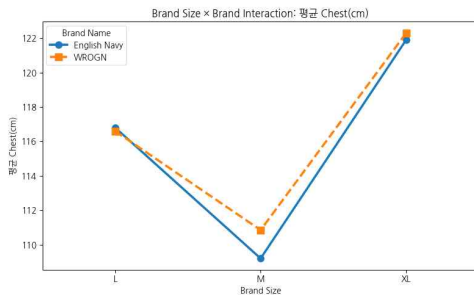


그림 7 Brand Size × Chest(cm) 라인 그래프

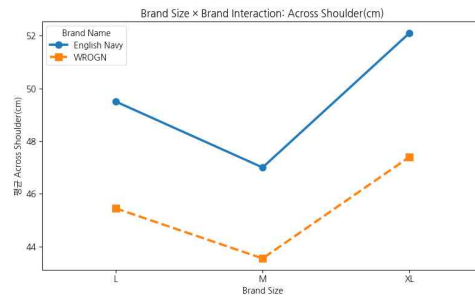


그림 8 Brand Size × Across Shoulder(cm) 라인 그래프

4. 종합 결론

TASK 1

- CNN 전이학습(ResNet-18)이 전통적 ML 대비 50% 성능 향상(97% 정확도) 달성
- 의류 분류에서 심층 특징 학습의 우수성 검증

TASK 2

- Q1: 동일 Brand Size 내 브랜드 간 치수 차이 유의함 (ANOVA $p < 0.05$)
- Q2: 소재 효과는 Brand Size 통제 후 유의하게 나타남 (다중회귀)
- Q3: 특정 브랜드 쌍에서 어깨너비 차이 존재 (t-test $p < 0.01$)

실무 시사점: 온라인 의류 쇼핑에서 사이즈 표기만으로는 실제 핏 예측 어려움. 브랜드별·소재별 상세 치수 정보 제공 필요.

5. 연구의 한계 및 향후 과제

5.1 데이터 표본수 제약

본 연구에서 사용한 shirts 데이터는 전체 165개 브랜드 중 대부분이 5개 미만의 샘플을 보유하고 있어, ANOVA 및 t-test 수행 시 다음과 같은 문제가 발생했다:

- 통계적 검정력 부족: 브랜드별 $n < 10$ 인 경우가 다수로, 유의미한 차이를 탐지하기 어려움
- 시각화 한계: Boxplot 대신 Stripplot을 사용해야 했으며, 빈도 상위 10개 브랜드로 제한
- 일반화 가능성 저하: 소수 브랜드 기반 결론을 전체 시장으로 확장하기 어려움

5.2 대표 치수 변수 선정의 논리적 근거 부족

본 분석에서는 Chest(가슴둘레)를 주요 분석 대상으로 삼았으나, 이를 선택한 명확한 이론적·실무적 근거가 부족했다:

- 상관관계 분석 결과: Size와의 상관계수가 0.27로 가장 낮았음에도 Chest를 대표 변수로 사용
- 소비자 관점 미반영: 실제 의류 핏에서 어깨너비(Across Shoulder)나 기장(Front Length)이 더 중요할 수 있으나 검증하지 못함

향후 연구에서는 이러한 한계점을 보완하여 보다 견고하고 실무적으로 유용한 분석 결과를 도출할 필요가 있다.

[부록]

부록 A. TASK 1 추가 분석 자료

A-1. 이미지 속성 상세 분석

- 이미지 해상도 분포 (샘플 30개 기준)

해상도 (width × height)	빈도
474 × 474	8
474 × 711	3
474 × 592	2
474 × 632	2
427 × 538	1

- 색공간: 전체 샘플 RGB 통일

- 파일 포맷: 100% JPEG

해석: 해상도 불균일로 인해 Resize 전처리 필수

A-2. 전처리된 Train 데이터 샘플 시각화



증강 효과 확인

- RandomHorizontalFlip, RandomRotation으로 데이터 다양성 확보

- ColorJitter로 조명 변화에 강건한 모델 학습

A-3. Baseline 모델 상세 성능 지표

Validation Set 성능

모델	Accuracy	Macro F1
Logistic Regression	0.676	0.646
Linear SVM	0.586	0.557

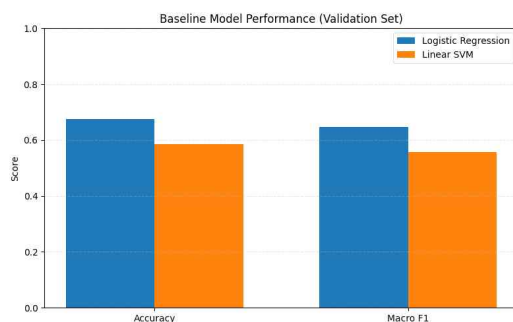
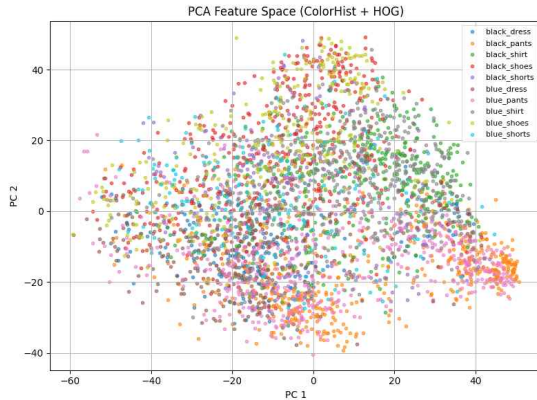


그림 10 Baseline 모델 성능 비교 그래프
해석: 로지스틱 회귀가 선형 SVM보다 모든 지표에서 우수하나, 둘 다 70% 미만의 한계

A-4 PCA: 특징 분포 시각화



해석

- 색상(ColorHist)과 형태(HOG) 특징을 PCA로 2차원 축소했으나, 대부분의 의류 카테고리끼리 서로 겹쳐 있어 클래스 간 경계가 모호함
- 일부 데이터(예: 우측 하단의 black_pants)가 뭉쳐 있는 경향은 보이지만, 전반적으로 이 두 주성분(PC1, PC2)만으로는 명확한 분류가 어려움

A-5. PCA 2D 기반 결정 경계 비교

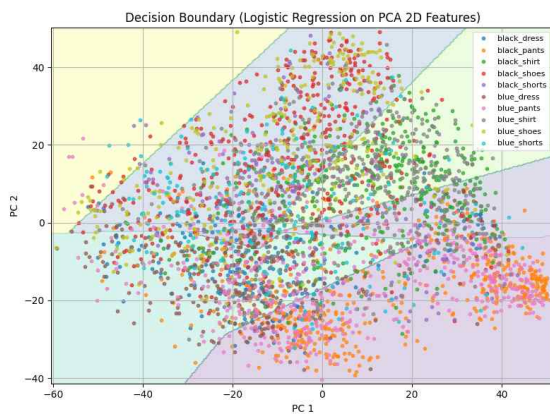


그림 12 로지스틱 회귀 결정 경계

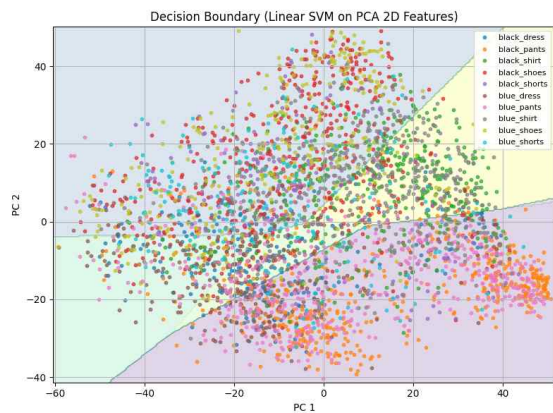


그림 13 선형 SVM 결정 경계

해석

- Logistic Regression: 데이터 공간을 직선(초평면)들을 교차시켜 다각형 형태로 나누었으나, 복잡하게 혼재된 클래스 영역을 직선으로 무리하게 나누려는 경향을 보임
- Linear SVM: Logistic Regression과 마찬가지로 선형 경계를 형성하고 있으며, 경계가 너무 단순하여 섞여 있는 데이터들을 정교하게 갈라내지 못함
- 두 모델 모두 PC1, PC2 두 개의 축만으로는 데이터의 비선형적 특성을 반영하지 못해 분류 성능이 낮을 것임을 시각적으로 확인시켜줌

A-6. Baseline 오분류 사례 분석

공통 오분류 패턴

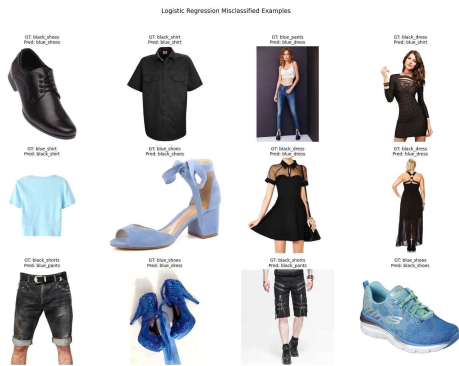


그림 14 로지스틱 회귀 오분류 샘플 (12개)



그림 15 선형 SVM 오분류 샘플 (12개)

- shorts ↔ pants 혼동: 기장 정보 부족
- 유사 색상 shoes 간 혼동: 형태 특징 학습 부족
- dress의 다양한 스타일 구분 실패

A-7. CNN 학습 과정 상세

Epoch별 성능 변화

Epoch	Train Loss	Train Acc	Val Loss	Val Acc	Val F1
1	0.4870	0.8684	0.1409	0.9583	0.9505
2	0.1338	0.9688	0.1011	0.9683	0.9625
3	0.0815	0.9800	0.0930	0.9741	0.9713
4	0.0656	0.9808	0.1162	0.9566	0.9489
5	0.0508	0.9869	0.0712	0.9783	0.9764
6	0.0330	0.9917	0.0769	0.9791	0.9743
7	0.0308	0.9903	0.0796	0.9716	0.9640

- Best Model: Epoch 5 (Val Loss 최저)
- Early Stopping: Epoch 7에서 발동 (patience=2)

A-8. CNN 오분류 사례 심층 분석

CNN 오분류 샘플 (총 37개 중 12개)

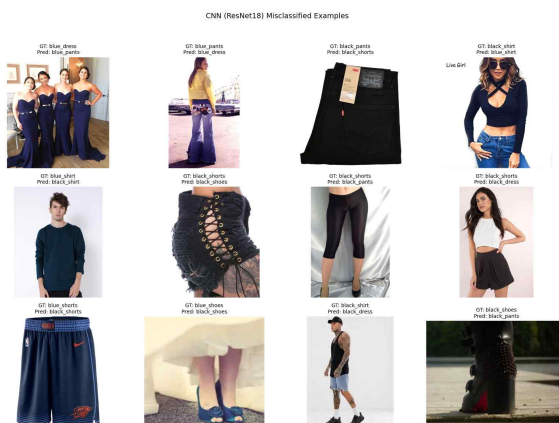


그림 16 CNN 오분류 샘플
오분류 원인 분석

- 극단적 촬영 각도 (측면, 부분 클로즈업)

- 액세서리/장식이 과도한 복잡한 디자인

오분류율: 3.1% (37/1199) - Baseline 대비 1/10 수준

부록 B. TASK 2 추가 분석 자료

B-1. 수치형 변수 분포

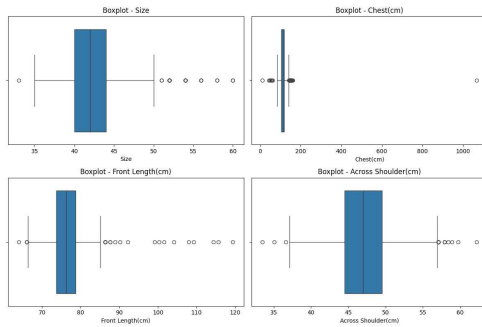


그림 17 수치형 변수 박스플롯

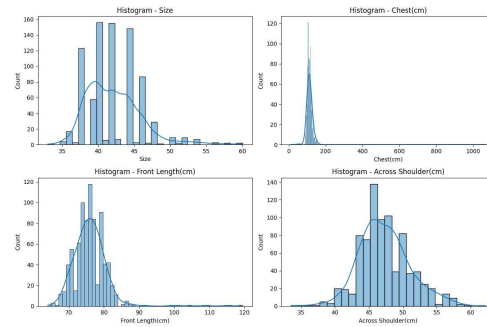


그림 18 수치형 변수 히스토그램

정규성 검토

- Size: 약 42 중심의 정규분포
- Chest: 우측 꼬리 긴 분포 (대형 사이즈 일부 존재)
- Front Length: 정규분포에 근사
- Across Shoulder: 정규분포에 근사

B-2. 이상치 탐지 상세 결과

IQR 기반 이상치 통계

변수	하한	상한	이상치 개수	이상치율
Size	34.0	50.0	26	3.1%
Chest(cm)	83.2	141.2	33	4.0%
Front Length(cm)	66.2	86.2	23	2.8%
Across Shoulder(cm)	37.0	57.0	12	1.8%

이상치 예시

- Size 52-60 범위: FXXL, TXL 등 비표준 사이즈
- Chest 146+ cm: 측정 오류 또는 특수 체형용
- Front Length 100+ cm: 롱 셔츠/원피스형

B-3. Q1 보충: 다섯 가지 Brand Size 전체 ANOVA 결과

S, M, L, XL, XXL 사이즈 종합 분석

Brand Size	변수	F-stat	p-value	해석
S	Chest(cm)	6.91	0.135	차이 없음
S	Front Length(cm)	2.25	0.358	차이 없음
S	Across Shoulder(cm)	9.59	0.099	차이 없음
M	Chest(cm)	6.58	0.141	차이 없음
M	Front Length(cm)	3.62	0.241	차이 없음
M	Across Shoulder(cm)	28.77	0.034	차이 있음

L	Chest(cm)	22.59	0.043	차이 있음
L	Front Length(cm)	2.98	0.285	차이 없음
L	Across Shoulder(cm)	59.24	0.017	차이 있음
XL	Chest(cm)	205.63	0.005	차이 있음
XL	Front Length(cm)	4.21	0.211	차이 없음
XL	Across Shoulder(cm)	28.97	0.034	차이 있음
XXL	Chest(cm)	26.96	0.152	차이 없음
XXL	Front Length(cm)	3.76e+25	1.3e-13	차이 있음
XXL	Across Shoulder(cm)	31.06	0.142	차이 없음

종합 해석

- M~XL 사이즈: Chest와 Across Shoulder에서 브랜드 간 차이 유의
- Front Length: 대부분 사이즈에서 브랜드 간 차이 미미 (XXL 제외)
- S 사이즈: 모든 치수에서 브랜드 간 차이 비유의 (표본 부족 영향)

B-4. Q2 보충: 소재별 상세 비교

Type_clean 분포

Type	빈도	비율
Cotton	766	99.4%
Polyester	5	0.6%

소재별 평균 치수

Type	Chest(cm)	Front Length(cm)	Across Shoulder(cm)
Cotton	112.1	75.5	47.0
Polyester	117.6	76.2	45.8

해석: Polyester 표본 극소량(n=5)으로 통계적 신뢰도 제한적, 참고용으로만 해석

B-5. Q3 보충: Chest(cm), Front Length(cm) t-test 결과

두 브랜드 간 비교

변수	Brand Size	mean_WROGN	mean_English Navy	p-value	해석
Chest	L	116.60	116.8	0.859972	차이없음
Chest	XL	122.30	121.9	0.422650	차이없음
Chest	M	110.85	109.2	0.422650	차이없음
Front	L	75.30	77.45	0.297532	차이없음
Front	XL	77.50	80.00	0.194386	차이없음
Front	M	74.30	74.95	0.685351	차이없음

해석: 기장과 가슴둘레는 두 브랜드 간 유의한 차이 없음. 어깨너비와 달리 표준화된 경향

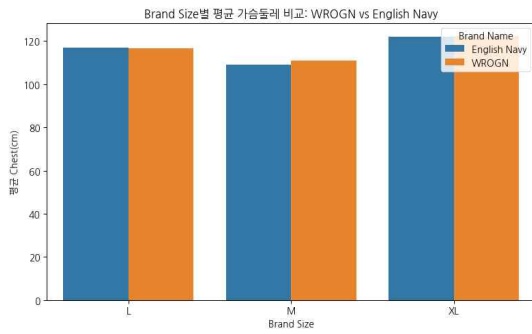


그림 19 Brand Size × Brand Chest(cm) 막대 그래프

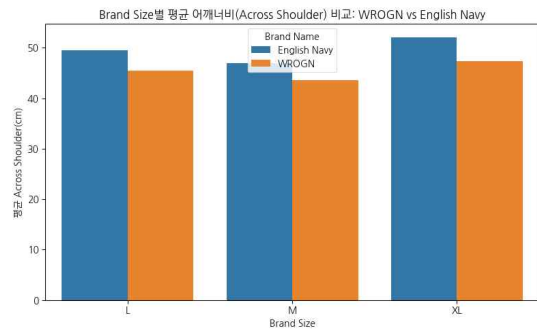


그림 20 Brand Size × Brand Across Shoulder(cm) 막대 그래프

B-6. 브랜드별 분포 (상위10개)

Brand Name별 샘플 수 (데이터 크기 (L, XL, M만 사용))

Brand Name	count
WROGN	6
English Navy	6
Roadster	3
Highlander	3
Campus Sutra	3
The Indian Garage	3
INVICTUS	3
The Bear House	3
United Colors of Benetto	3
Mast & Harbour	3

해석: 주요 브랜드도 사이즈별 표본이 매우 부족