

# Women's E-Commerce Clothing Reviews

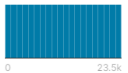
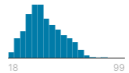



- 23,000 Customer Reviews and Ratings -

## 1. 데이터 이해

1) **Dataset:** 고객이 작성한 리뷰를 중심으로 진행되는 여성 의류 전자 상거래 데이터 세트

## 2) Content

: 이 데이터세트에는 23,486개의 행과 10개의 특징 변수가 포함됨. 각 행은 고객 리뷰에 해당하며 변수를 포함함.

About this file									
This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers									
#	Clothing ID	# Age	Title	Review Text	# Rating	# Recommended L...	# Positive Feedba...	Division Name	
id	Unique ID of the product	Age of the reviewer	Title of the review	review	Product rating by reviewer	Whether the product is recommended or not by the reviewer	Number of positive feedback on the review	Name of the division product is in	
			<div><div>[null]</div><div>Love it!</div><div>Other (19540)</div></div>	<div><div>[null]</div><div>Perfect fit and I've g...</div><div>Other (22638)</div></div>				General General Petite Other (1516)	
0	767	33		Absolutely wonderful - silky and sexy and comfortable	4	1	0	Intimates	
1	1090	34		Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never...	5	1	4	General	
2	1077	60	Some major design flaws	I had such high hopes for this dress and really wanted it to work for me. i initially ordered the pe...	3	0	0	General	
3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothin...	5	1	0	General Petite	

- Clothing ID: 검토 중인 특정 부분을 나타내는 정수 범주형 변수

- Age: 검토자 연령의 양의 정수 변수

- Title: 검토 제목 문자열 변수

- Review Text: 검토 본문에 대한 문자열 변수

- Rating: 고객이 1 (Worst)에서 5 (Best)까지 부여한 제품 점수에 대한 양의 순서 정수 변수

- Recommended IND: 고객이 제품을 추천하는 위치를 나타내는 이진 변수

(고객이 제품을 권장하는 경우: 1, 권장하지 않는 경우: 0)

- Positive Feedback Count: 이 리뷰가 긍정적이라고 생각하는 다른 고객의 수를 문서화하는 양의 정수

- Division Name: 제품 상위 레벨의 카테고리 이름

- Department Name: 제품 부서 이름의 카테고리 이름

- Class Name: 제품 클래스 이름의 카테고리 이름

## 2. 데이터 활용 사례

### 1) kernel 소개

- 주제: Predicting Sentiment from Clothing Reviews (의류 리뷰에서 감정 예측)

- 다양한 방법을 사용하여 고객의 리뷰로부터 고객의 정서를 예측.

### 2) 전체 과정(자료 중심)

#### ① Importing Modules and Reading the Dataset

: 모듈을 가져와서 자료를 읽고, 일부 열을 제거하여 새로운 dataset 정의

Out[1]:

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Intimates	Intimate	Intimates
1	1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

#### ② Data frame에 단어 개수 추가 및 일부 단어 사용 횟수 확인

: 유용한 정보를 도출해내기 위해 wordcounts 함수 정의

Out[2]:

	Review Text	Rating	Class Name	Age	Word Counts
0	Absolutely wonderful - silky and sexy and comf...	4	Intimates	33	{'absolutely': 1, 'and': 2, 'comfortable': 1, ...}
1	Love this dress! it's sooo pretty. i happene...	5	Dresses	34	{'am': 1, 'and': 2, 'bc': 2, 'be': 1, 'below': ...}
2	I had such high hopes for this dress and reall...	3	Dresses	60	{'and': 3, 'be': 1, 'bottom': 1, 'but': 2, 'ch': ...}
3	I love, love, love this jumpsuit. it's fun, fl...	5	Pants	50	{'and': 1, 'but': 1, 'compliments': 1, 'every': ...}
4	This shirt is very flattering to all due to th...	5	Blouses	47	{'adjustable': 1, 'all': 1, 'and': 1, 'any': 1, ...}

#### ③ WordCloud 사용하여 리뷰에서 클래스 이름, 일부 선택된 단어 및 모든 단어의 밀도 표시

: '단어 밀도' 시연

(‘사랑, 증오, 환상, 후회’와 같은 고객 정서를 보여주는 단어 선택함 → 제품 클래스 이름을 확인하여 고객들이 가장 선호하는 클래스를 알아냄 → wordcloud 모듈을 사용하여 선택한 단어의 단어 수와 클래스 이름을 나타내는 표의 처음 다섯 줄을 프린트함.)

```
Selected Words
love      8951
great     6117
super     1726
happy     785
glad      614
dtype: int64

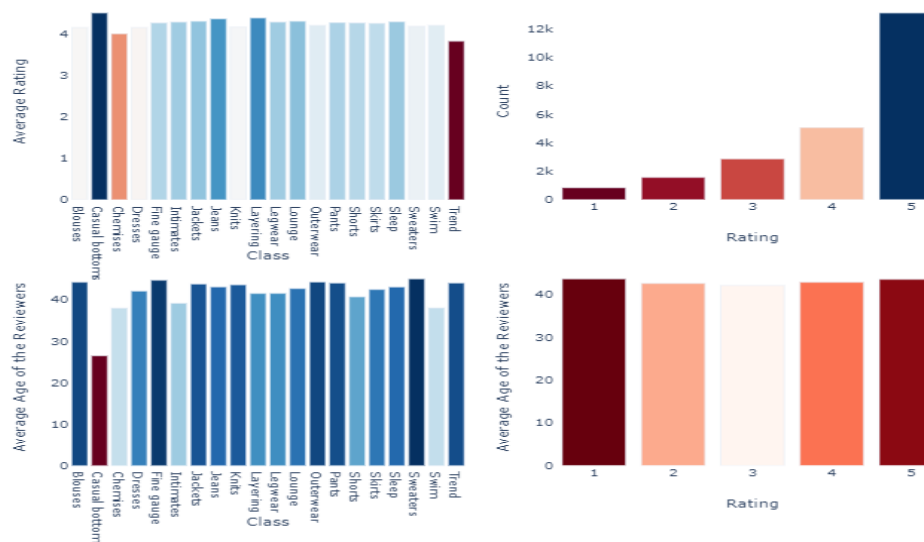
Class Names
Dresses   6319
Knits     4843
Blouses   3097
Sweaters  1428
Pants     1388
Name: Class Name, dtype: int64
```



- Love, great, super 등 긍정적인 단어가 더 많이 사용됨.
- 고객들은 대부분 드레스, 니트, 블라우스를 선호함
- Dress, love가 모든 리뷰에서 빈번하게 사용됨.

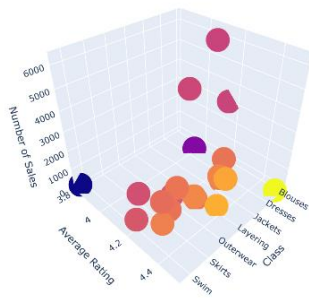
#### ④ 등급, 클래스 이름 및 연령 사이의 관계 알아보기

: 관계성을 알아보기 위해 아래의 동적 차트 사용



- 대부분의 평가가 긍정적이고, 클래스 간 평균 등급이 비슷해 보임.
- 반면에, 나이를 보면 평균 연령은 등급에 따라 크게 달라지지 않음.
- 평균 나이는 Casual bottom을 제외한 클래스 이름 사이에서 약간씩 변화함.

Average Rating & Class & Number of Reviewers



## ⑤ 감정 분류기 만들기

: dataset에 긍정적/부정적을 나타내는 칼럼이 없기 때문에 새로운 감성 칼럼을 정의함.

- 4등급 이상의 리뷰: 긍정 (새로운 프레임에서 true로)
- 2등급 이하의 리뷰: 부정 (새로운 프레임에서 false로)
- 3과 같은 중립 등급을 가진 선은 포함하지 않음.

→ training과 test sets로 자료를 쪼갬.

→ 모델들을 하나씩 맞춤.

### Logistic Regression

```
In [7]: start=dt.datetime.now()
lr = LogisticRegression()
lr.fit(X_train,y_train)
print('Elapsed time: ',str(dt.datetime.now()-start))
```

Elapsed time: 0:00:01.056909

### Naive Bayes

```
In [8]: start=dt.datetime.now()
nb = MultinomialNB()
nb.fit(X_train,y_train)
print('Elapsed time: ',str(dt.datetime.now()-start))
```

Elapsed time: 0:00:00.011419

### Support Vector Machine (SVM)

```
In [9]: start=dt.datetime.now()
svm = SVC()
svm.fit(X_train,y_train)
print('Elapsed time: ',str(dt.datetime.now()-start))
```

Elapsed time: 0:00:48.484381

### Neural Network

```
In [10]: start=dt.datetime.now()
nn = MLPClassifier()
nn.fit(X_train,y_train)
print('Elapsed time: ',str(dt.datetime.now()-start))
```

Elapsed time: 0:01:41.359712

## ⑥ Evaluating Models

: Data frame에 결과 추가

Out[11]:

	Review Text	Rating	Class Name	Age	Word Counts	Sentiment	Logistic Regression	Naive Bayes	SVM	Neural Network
19218	I love this dress's gentle blue lace. the silh...	5	Dresses	35	{'and': 1, 'as': 1, 'blue': 1, 'chest': 1, 'dr...	True	True	True	True	True
3530	Beautiful choice...beautiful fit for my daught...	5	Knits	51	{'beautiful': 2, 'body': 1, 'choice': 1, 'daug...	True	True	True	True	True
15663	If you are shaped anything like me, you will h...	4	Dresses	25	{'am': 1, 'and': 2, 'anything': 1, 'are': 1, '...	True	True	True	True	True
21310	This top is so cute and of spectacular quality...	5	Blouses	33	{'10': 1, '34c': 1, 'all': 1, 'almost': 1, 'an...	True	True	True	True	True
15154	First saw this poncho on a petite blog and aft...	5	Sweaters	56	{'after': 1, 'and': 5, 'below': 1, 'blog': 1, ...	True	True	True	True	True

## ⑦ 정밀도 - Recall -F1-Score

- 앞서, ROC 곡선과 혼동 매트릭스를 이용했지만 최종 평가를 내리기에는 부족함.

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FN + FP}$$

- 평가 모델 부분에서 모든 평가 지표의 결과를 보면, Naïve Bayes와 Logistic Regression에서 최상의 결과를 얻을 수 있음. 따라서 두 가지 모두 정서를 예측하는데 매우 효과적임. 반면에, Naïve Bayes는 시간이 적게 걸리므로 더 큰 dataset를 가질 때 중요한 장점이 될 수 있음.

Logistic Regression				
	precision	recall	f1-score	support
False	0.78	0.64	0.70	458
True	0.96	0.98	0.97	3665
accuracy			0.94	4123
macro avg	0.87	0.81	0.84	4123
weighted avg	0.94	0.94	0.94	4123
Naive Bayes				
	precision	recall	f1-score	support
False	0.79	0.65	0.71	458
True	0.96	0.98	0.97	3665
accuracy			0.94	4123
macro avg	0.87	0.82	0.84	4123
weighted avg	0.94	0.94	0.94	4123
Support Vector Machine (SVM)				
	precision	recall	f1-score	support
False	0.00	0.00	0.00	458
True	0.89	1.00	0.94	3665
accuracy			0.89	4123
macro avg	0.44	0.50	0.47	4123
weighted avg	0.79	0.89	0.84	4123
Neural Network				
	precision	recall	f1-score	support
False	0.73	0.63	0.68	458
True	0.95	0.97	0.96	3665
accuracy			0.93	4123
macro avg	0.84	0.80	0.82	4123
weighted avg	0.93	0.93	0.93	4123

### 3) 결론

- 고객의 리뷰에는 Love, great, super 등 긍정적인 단어가 부정적인 단어보다 더 많이 사용됨.
- 고객들은 대부분 드레스, 니트, 블라우스를 선호함
- Dress, love가 모든 리뷰에서 빈번하게 사용됨.
- 대부분의 평가가 긍정적이고, 클래스 간 평균 등급이 비슷해 보이는 반면에, 나이를 보면 평균 연령은 등급에 따라 크게 달라지지 않음.
- 평균 나이는 Casual bottom을 제외한 클래스 이름 사이에서 약간씩 변화함.

### 3. 데이터를 프로젝트에 적용한다면?

여러 브랜드를 한눈에 볼 수 있는 쇼핑몰 APP을 만들어서

- 각 쇼핑몰마다 리뷰에 빈번하게 등장하고 긍정적인 표현이 많이 사용된 제품 또는 키워드 순서대로 제품 정렬
- 각 쇼핑몰마다 제품마다 리뷰에 높은 빈도수 키워드를 4개 정도 보여줌
- 각 쇼핑몰마다 가장 높은 빈도수 키워드를 보여줌.

이런 식으로 쇼핑몰과 각 제품의 특징을 한눈에 볼 수 있도록 하면 좋을 것 같습니다.