

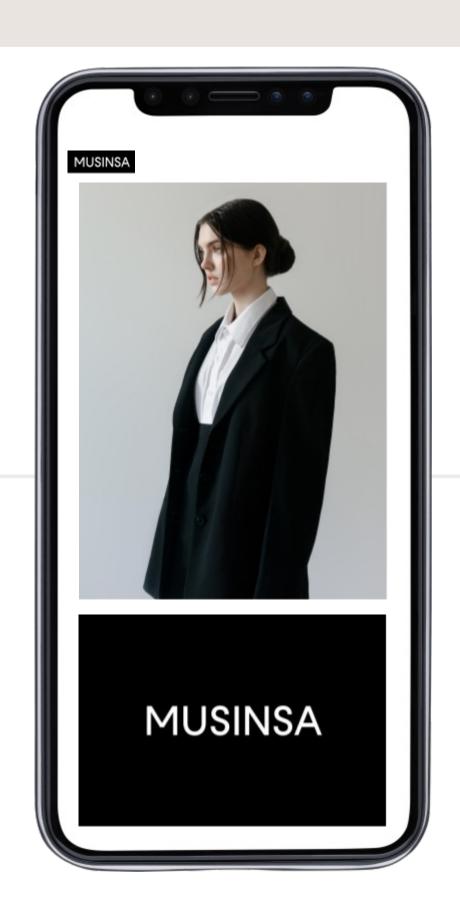
목차

01. 연구 배경

- · 주제 선정 배경 · 가설 설정

03. 데이터 분석 / 시각화

- · 다중회귀분석 · 토픽모델링



02. 데이터 탐색

- · 월간 랭킹 크롤링 데이터 · 전처리 전 데이터 · 전처리 후 데이터

04. 결론

• 활용방안 및 기대효과

1. 연구배경

- 주제 선정 및 가설 설정



소비자들은 사이트에서 제공하는 랭킹 시스템을 참고하여 구매를 결정



랭킹은 매출, 판매 수량, 구매후기 수, 조회수를 반영한 공식에 의해 선정



가설 설정



평점, 조회수, 좋아요는 실제 구매로 이어지지 않을 수 있음 → 순위와 상관관계가 높지 않을 것



2.

구매후기 개수는 실구매와 직접적인 연관이 있음 → 순위와 상관관계가 높을 것

각 요인들이 순위에 어느 정도의 영향을 미치는지 분석





- 무신사 카테고리별 월간 랭킹 크롤링

크롤링 준비

```
import pandas as pd
import requests
from bs4 import BeautifulSoup as bs
from tadm import tadm
import time
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.options import Options
chrome_options = Options()
chrome_options.headless = True
def musinsa_rank(category_num,page_num):
    url = f"https://www.musinsa.com/ranking/best?period=month&age=ALL&mainCategory=O{category_num}
    &subCategory=&leafCategory=&price=&golf=false&kids=false&newProduct=false&exclusive=false&discount=false&soldOut=false&page={page_num}
    &viewType=small&priceMin=&priceMax="
    response = requests.get(url)
    html = bs(response.text, 'lxml')
    musinsa_rank_df = rbnl(html)
    return musinsa_rank_df
```

- 무신사 카테고리별 월간 랭킹 크롤링

순위/상품명/URL/브랜드명

```
| def rbn|(html):
    musinsa_rank_df = pd.DataFrame()

#순위 볼기
    rank_html = html.select('#goodsRankList > li > p')
    rank_no_list = []

for i in rank_html:
        rank_no_list.append(i.string.strip())

musinsa_rank_df['순위'] = rank_no_list

#링크와 의류명 볼기
    link_name_html = html.select('#goodsRankList > li > div.li_inner > div.article_info > p.list_info > a')
    link_list = []

for i in link_name_html:
    link_list.append(i['href'])
    name_list.append(i['title'])

musinsa_rank_df['상품명']=name_list
    musinsa_rank_df['VRL']=link_list
```

```
#브랜드 이름 뽑기
brand_html = html.select('#goodsRankList > li > div.li_inner > div.article_info > p.item_title > a')
brand_list = []

for i in brand_html:
    brand_list.append(i.string)

musinsa_rank_df['브랜드']=brand_list

#상세 페이지 크롤링
musinsa_rank_df2 = specific_info(link_list)

#데이터 프레임 옆으로 함치기
musinsa_rank_df = pd.concat([musinsa_rank_df, musinsa_rank_df2], axis=1)

return musinsa_rank_df
```

- 무신사 카테고리별 월간 랭킹 크롤링

조회수/누적 판매량/좋아요/구매후기/평점

```
view_html=html_2.find_all("strong", {"id":"pageview_1m"})
if len(view_html)==0:
    view="0"
    view=view_html[0].get_text()
view list.append(view)
sales_html=html_2.find_all("strong", {"id":"sales_1y_qty"})
if len(sales_html)==0:
   sales="0"
    sales=sales_html[0].get_text()
sales_list.append(sales)
like_html=html_2.find_all("span", {"class": "prd_like_cnt"})
if len(like_html)==0:
    like="0"
    like=like_html[0].get_text()
like_list.append(like)
```

```
buy_html=html_2.find_all("span", {"id":"review_total"})
if len(buy_html)==0:
   buy="0"
else:
    buy=buy_html[0].get_text()
buy_list.append(buy)
score_html=html_2.find_all("span", {"class": "prd-score__rating"})
if len(score_html)==0:
   score="0"
   score=score_html[0].get_text()
score_list.append(score)
```

- 무신사 카테고리별 월간 랭킹 크롤링

웹 크롤링 후 CSV 파일로 저장 → 데이터 프레임 형식으로 불러오기

df=	pd.rea	d_csv(' <u>/content/drive/MyDrive/WEIT_Final</u>	project/musinsa_ranking_category18csv', encodi	ng='cp949')						
df.	head()									
	순위	상품명	URL	브랜드	품번	조회수	좋아요	구매후기 개수	평점	누적판매량
0	1위	센테니얼 85 로우 - 화이트:블루 / IF5419	https://www.musinsa.com/app/goods/3030422?loc=	아디다스	IF5419	17.9만 회 이상	26,990	674	4.8	3.8천 개 이상
1	2위	슈퍼스타 - 화이트:블랙 / EG4958	https://www.musinsa.com/app/goods/1622068?loc=	아디다스	EG4958	8.3만 회 이상	31,953	6,154	4.9	1.4만 개 이상
2	3위	XT-6 GTX - 블랙:에보니:루나 락 / L41663500	https://www.musinsa.com/app/goods/2992063?loc=	살로몬 스포츠스타일	L41663500	7.8만 회 이상	7,194	1,957	4.9	4.2천 개 이상
3	4위	클럽 C 85 빈티지 - 크림 / DV6434	https://www.musinsa.com/app/goods/2741233?loc=	리복	RXSOXXCVUE1	8.4만 회 이상	17,894	2,429	4.9	9천 개 이상
4	5위	조그 100 2 (4E) - 블랙:블랙 / TJG138-9090	https://www.musinsa.com/app/goods/969541?loc=g	아식스	TJG138-9090	10.4만 회 이상	47,042	28,716	4.9	2.9만 개 이상

columns

: 순위, 상품명, URL, 브랜드, 품번, 조회수, 좋아요, 구매후기 개수, 평점, 누적판매량

- 무신사 카테고리별 월간 랭킹 크롤링 〈전처리 전 데이터〉

조회수 0회 / 누적판매량 0개 = 결측치

df.isnul	l().su	ım()	
순위		0	
상품명		0	
URL		0	
브랜드		0	
품번		0	
조회수		956	
좋아요		0	
구매후기	개수		0
평점		0	
누적판매형	량	218	36

필요하지 않은 열 = 품번, URL

df[['URL','품번']]					
	URL	품번			
0	https://www.musinsa.com/app/goods/3030422?loc=	IF5419			
1	https://www.musinsa.com/app/goods/1622068?loc=	EG4958			
2	https://www.musinsa.com/app/goods/2992063?loc=	L41663500			
3	https://www.musinsa.com/app/goods/2741233?loc=	RXSOXXCVUE1			
4	https://www.musinsa.com/app/goods/969541?loc=g	TJG138-9090			
4495	https://www.musinsa.com/app/goods/2505218?loc=	499535			
4496	https://www.musinsa.com/app/goods/2505203?loc=	499527			
4497	https://www.musinsa.com/app/goods/2691741?loc=	CKSO2F512BK			
4498	https://www.musinsa.com/app/goods/1532416?loc=	732200			
4499	https://www.musinsa.com/app/goods/1075814?loc=	AJK7345CAEN			
4500 rc	4500 rows × 2 columns				

조회수, 좋아요 수, 구매후기 개수, 누적판매량 = object type

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4500 entries, 0 to 4499
Data columns (total 10 columns):
    Column Non-Null Count Dtype
    순위
           4500 non-null object
    상품명
              4500 non-null object
            4500 non-null object
    URL
              4500 non-null object
    브랜드
    품번
             4500 non-null object
           3544 non-null object
    조회수
    좋아요
           4500 non-null
                            object
    구매후기 개수 4500 non-null object
             4500 non-null
    평점
                           float64
                2314 non-null object
    누적판매량
```

- 무신사 카테고리별 월간 랭킹 크롤링 〈전처리 후 데이터〉

결측치 제거: 'NaN' → 0

```
df = df.fillna(0)df.isnull().sum()순위0상품명0브랜드0조회수0주매후기 개수0평점0누적판매량0
```

필요하지 않은 열 제거

```
df=df.drop(['품번','URL'], axis=1)

df.columns

Index(['순위', '상품명', '브랜드', '조회수', '좋아요', '구매후기 개수', '평점', '누적판매량'], dtype='object')
```

- 무신사 카테고리별 월간 랭킹 크롤링 〈전처리 후 데이터〉

조회수 / 좋아요 / 구매후기 개수 / 누적판매량

df									
	순위	상품명	브랜드	조회수	좋아요	구매후기 개수	평점	누적판매량	%
0	1위	센테니얼 85 로우 - 화이트:블루 / IF5419	아디다스	17.9만 회 이상	26,990	674	4.8	3.8천 개 이상	
1	2위	슈퍼스타 - 화이트:블랙 / EG4958	아디다스	8.3만 회 이상	31,953	6,154	4.9	1.4만 개 이상	
2	3위	XT-6 GTX - 블랙:에보니:루나 락 / L41663500	살로몬 스포츠스타일	7.8만 회 이상	7,194	1,957	4.9	4.2천 개 이상	
3	4위	클럽 C 85 빈티지 - 크림 / DV6434	리복	8.4만 회 이상	17,894	2,429	4.9	9천 개 이상	
4	5위	조그 100 2 (4E) - 블랙:블랙 / TJG138-9090	아식스	10.4만 회 이상	47,042	28,716	4.9	2.9만 개 이상	
4495	4496위	TENNIS WOMAN LACET - HIBISCUS	벤시몽	0	20	0	0.0	0	
4496	4497위	TENNIS WOMAN ELLY - PALE PINK	벤시몽	0	69	0	0.0	0	
4497	4498위	[US] 남성 PURE CLASSIC Sneakers (BLACK) CKSO2F512BK	챔피온	200회 이상	29	5	5.0	0	
4498	4499위	메이크 런닝화(WT)	프로월드컵	0	93	7	5.0	50개 미만	
4499	4500위	이스트 YEAST Ecru_Navy	베이크솔	100회 이상	89	2	5.0	0	
4500 rc	ows × 8 co	olumns							

- 무신사 카테고리별 월간 랭킹 크롤링 〈전처리 후 데이터〉

조회수, 좋아요 수, 구매후기 개수, 누적판매량 \rightarrow 쉼표 & 문자 제거 \rightarrow float type

```
def convert_view_count(count_str):
    if isinstance(count_str, float):
        return count_str

    count_str = count_str.replace(',', '') # 쉼표 제거
    if '만' in count_str:
        count = count_str.split('만')[0]
        count = float(count) * 10000

elif '천' in count_str:
        count = count_str.split('천')[0]
        count = float(count) * 1000

else:
        count = float(re.sub(r'[^0-9]','',count_str))
    return int(count)

df['조회수'] = df['조회수'].apply(convert_view_count)
```

```
def convert_view_count(count_str):
    if isinstance(count_str, float):
        return count_str

    count_str = count_str.replace(',', '') # 쉼표 제거
    if '만 개' in count_str:
        count = count_str.split('만')[0]
        count = float(count)*10000

elif '천 개' in count_str:
        count = count_str.split('천')[0]
        count = float(count)*1000

else:
        count = float(re.sub(r'[^0-9]','',count_str))
    return int(count)

df['누적판매량'] = df['누적판매량'].apply(convert_view_count)
```

```
def convert_count(count_str):
    count_str = float(re.sub(r'[^0-9]','',count_str))
    return int(count_str)

df['구매후기 개수'] = df['구매후기 개수'].apply(convert_count)

def convert_count(count_str):
    count_str = float(re.sub(r'[^0-9]','',count_str))
    return int(count_str)

df['좋아요'] = df['좋아요'].apply(convert_count)
```

- 무신사 카테고리별 월간 랭킹 크롤링 〈전처리 후 데이터〉

순위 → 회귀분석을 위해 연속형으로 변경

```
import re
def convert_count(count_str):
   count_str = float(re.sub(r'[^0-9]','',count_str))
   return int(count_str)
df['순위'] = df['순위'].apply(convert_count)
def rank_to_score(rank):
   max_rank = 1
   min_rank = 4500
   score = (max_rank - rank) / (max_rank - min_rank)
   return score
df['순위']=df['순위'].apply(rank_to_score)
list_rank = df['순위'].to_list()
reversed_rank = list_rank[::-1]
df['순위'] = reversed_rank
reversed_rank
df
```

	순위
0	1.000000
1	0.999778
2	0.999555
3	0.999333
4	0.999111
4495	0.000000
	0.000889
4496	0.000889
4496 4497	
	0.000667
4497	0.000667 0.000445

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4500 entries, 0 to 4499
Data columns (total 8 columns):
   Column Non-Null Count Dtype
   순위 4500 non-null float64
    상품명 4500 non-null object
   브랜드
             4500 non-null object
   조회수
             4500 non-null float64
   좋아요
              4500 non-null int64
   - 구매후기 개수 4500 non-null int64
   평점
             4500 non-null float64
   - 누적판매량 4500 non-null float64
dtypes: float64(4), int64(2), object(2)
memory usage: 281.4+ KB
```

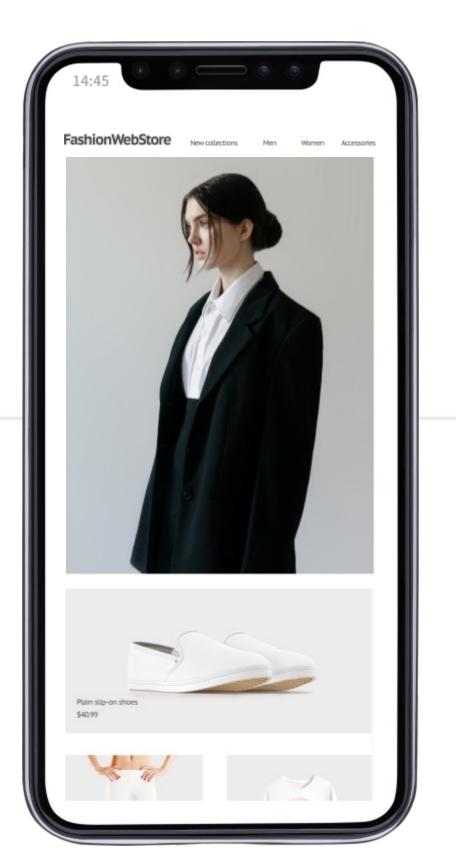
3. 데이터 분석 & 시각화

3-1. 다중회귀분석

3-1-1. 순위 포함 회귀분석

3-1-2. 순위 제외 회귀분석

3-2. 토픽 모델링



- 순위 포함 회귀분석

• 독립변수

조회수, 누적판매량, 평점 좋아요 수, 구매후기 개수

• 종속변수

순위

• 분석 결과

```
X = df[['평점', '조회수', '<u>누적판매량', '구매후기</u> 수','좋아요 수']]
y = df['순위']
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.2, random_state=42)
lr = LinearRegression()
lr.fit(X_train, y_train)
from sklearn.metrics import mean_squared_error, r2_score
y_pred = lr.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2=r2_score(y_test, y_pred)
print('MSE',mse)
print('R-squared', r2)
       .07914163023745659
       ared 0.08270043376368441
```

상의 R-squared

0.08270043376368441

하의

R-squared 0.2929021032358327

스니커즈

R-squared 0.3000446701809504

신발

R-squared 0.22042376427874544

- 순위 <mark>포함</mark> 회귀분석 & Heatmap

• 독립변수

조회수, 누적판매량, 평점 좋아요 수, 구매후기 개수

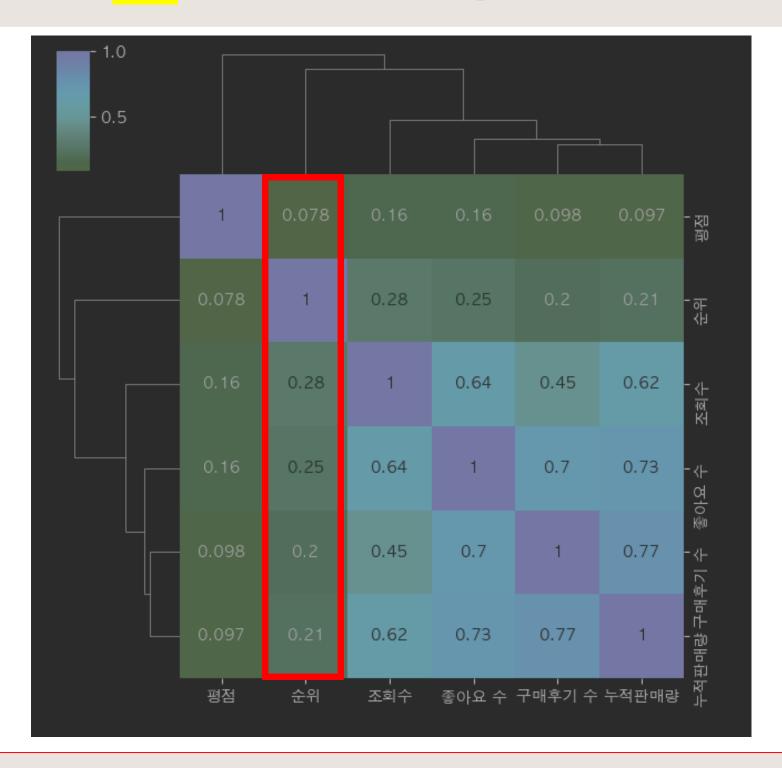
• 종속변수

순위

- 상관관계 그래프
- Heatmap

```
# 상관계수 그래프
import seaborn as sns
from matplotlib import rc
rc('font', family = 'AppleGothic')
fig = sns.clustermap(df_corr,
             annot = True,
fig.fig.set_size_inches(6,6)
```

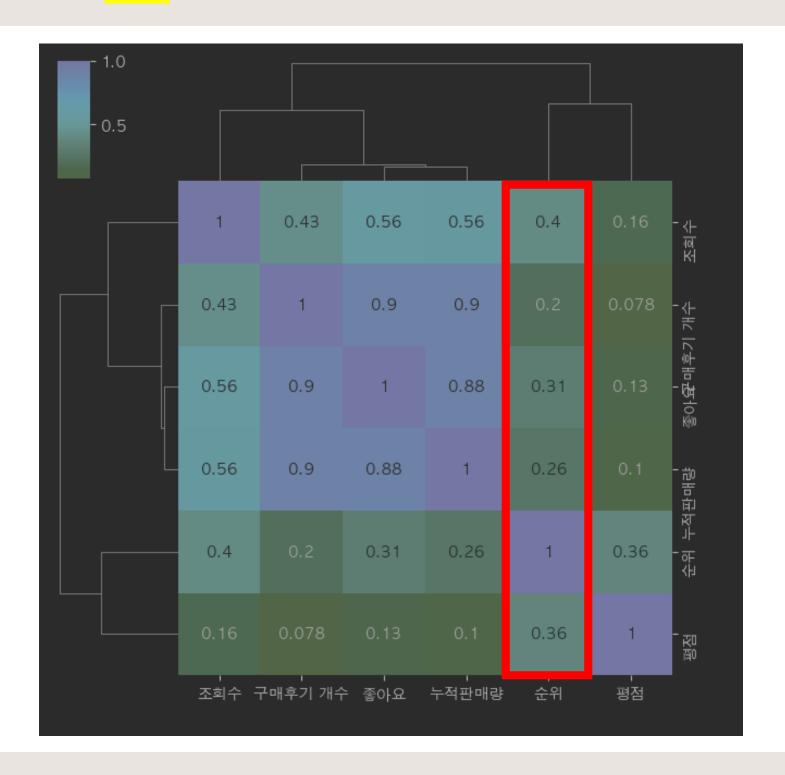
- 순위 <mark>포함</mark> 회귀분석 & Heatmap





하으

- 순위 <mark>포함</mark> 회귀분석 & Heatmap





스니커즈

신빝

- 순위 제외 회귀분석

• 독립변수

조회수, 좋아요 수, 구매후기 개수, 평점

• 종속변수

누적판매량

• 분석 결과

```
X = df[['평점', '조회수', '구매후기 수','좋아요 수']]
y = df['누적판매량']
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.2, random_state=42)
lr = LinearRegression()
lr.fit(X_train, y_train)
from sklearn.metrics import mean_squared_error, r2_score
y_pred=lr.predict(X_test)
mse=mean_squared_error(y_test, y_pred)
r2=r2_score(y_test, y_pred)
print('MSE',mse)
print('R-squared', r2)
  MSE 6182800.327133722
  R-squared 0.73182028276286
                                          MSE 527900.4595807663
                                          R-squared 0.7021481120419162
```

상의

R-squared 0.73182028276286

하의

R-squared 0.5925887840601495

스니커즈

R-squared 0.8997768824228625

신발

R-squared 0.7021481120419162

- 순위 <mark>제외</mark> 회귀분석 & Heatmap

• 독립변수

조회수, 좋아요 수, 구매후기 개수, 평점

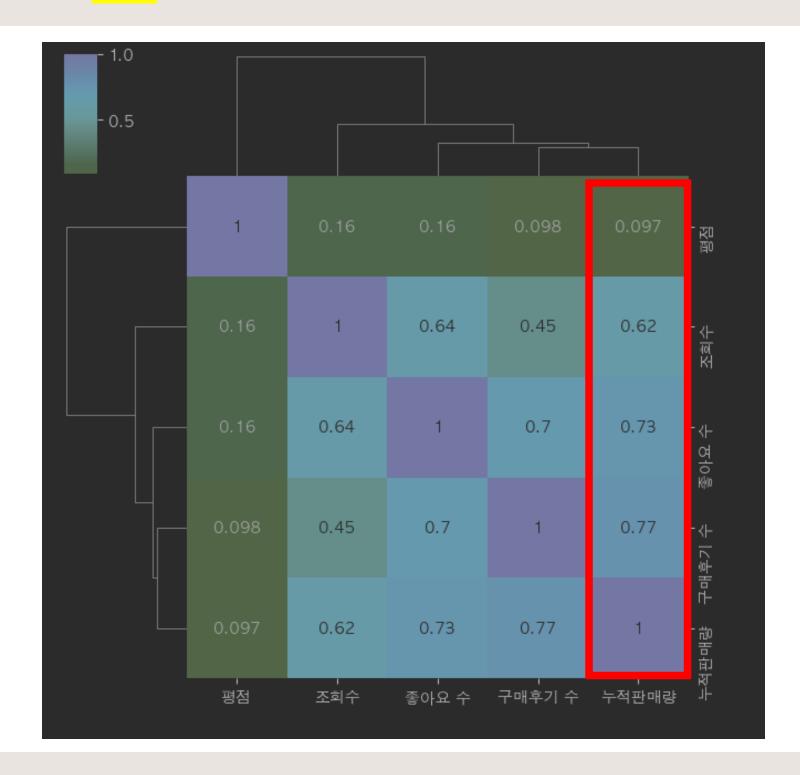
• 종속변수

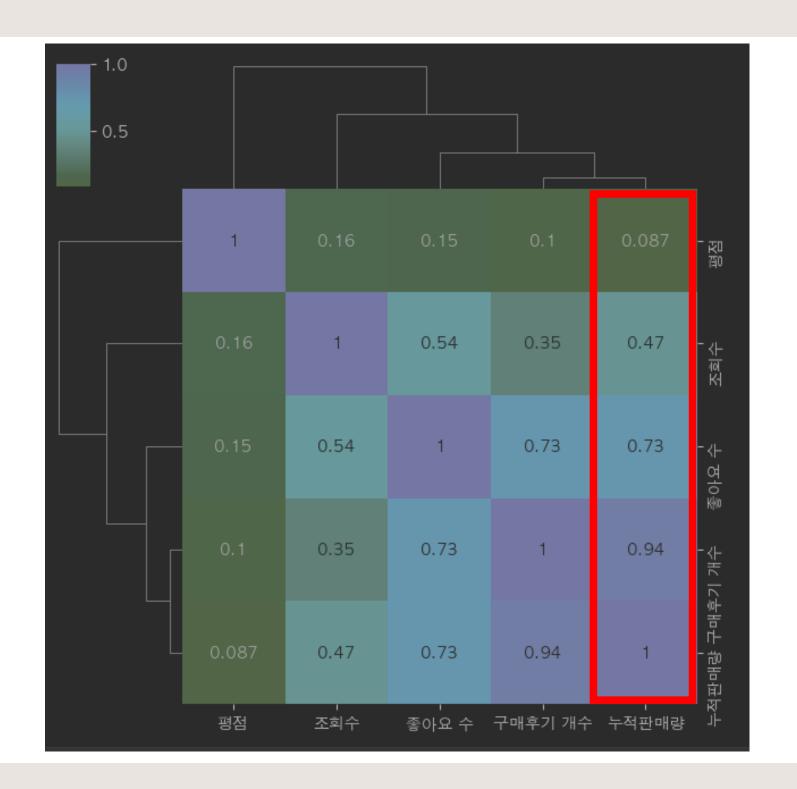
누적판매량

- 상관관계그래프
- Heatmap

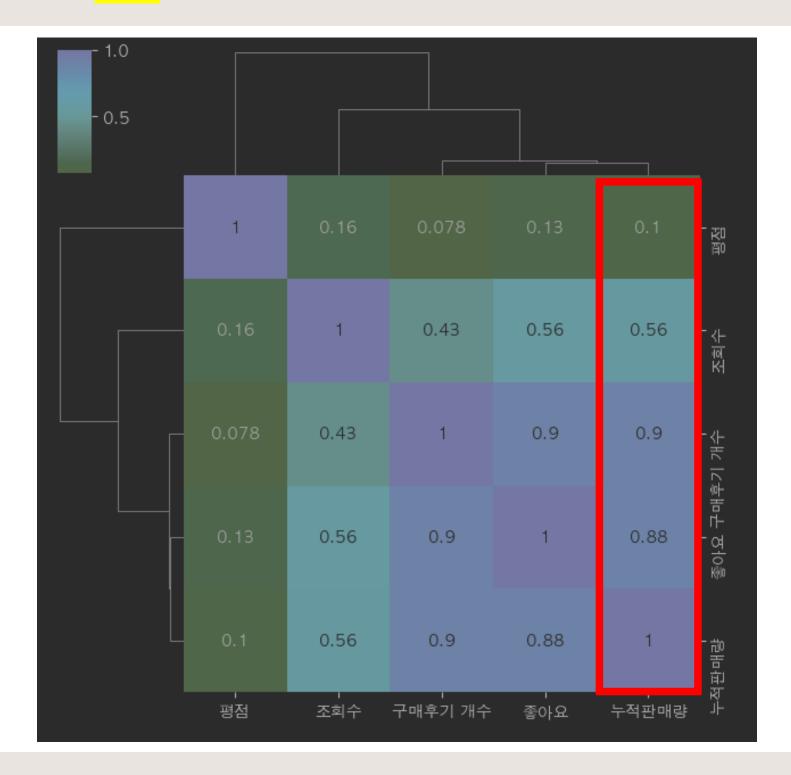
```
df = df.drop('순위', axis=1)
 df_corr = df.corr()
 import seaborn as sns
 from matplotlib import rc
rc('font', family = 'AppleGothic')
 fig = sns.clustermap(df_corr,
fig.fig.set_size_inches(6,6)
```

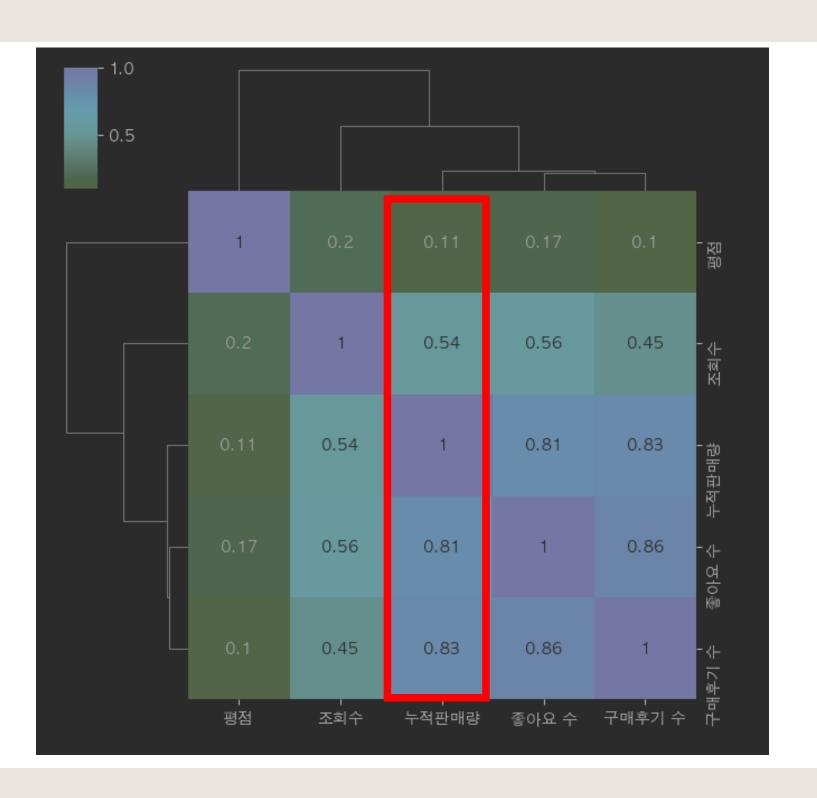
- 순위 <mark>제외</mark> 회귀분석 & Heatmap





- 순위 <mark>제외</mark> 회귀분석 & Heatmap





스니커즈

- R squred



상의

 $0.0827 \rightarrow 0.7318$



스니커즈

 $0.3000 \rightarrow 0.8997$



하의

 $0.2929 \rightarrow 0.5925$



신발

 $0.2204 \rightarrow 0.7021$





토픽모델링

- Top 10 상품명, 브랜드

```
df['상품명'] = df['상품명'].apply(preprocess)
df['브랜드명'] = df['브랜드명'].apply(preprocess)
 rom sklearn.feature_extraction.text import TfidfVectorizer
 rom sklearn.decomposition import LatentDirichletAllocation
 이 코드는 10개의 토픽을 <u>추출하고</u>, 각 <u>토픽별로</u> 가장 <u>관련성이</u> 높은 단어 10개를 <u>출력합니다</u>. 이를 <u>바탕으로</u> 문서의 주제를 파악할 수 <u>있습니다</u>.
docs = df['상품명'] + ' ' + df['브랜드명']
tfidf = TfidfVectorizer(stop_words='english')
doc_vectors = tfidf.fit_transform(docs)
num_topics = 10 # 추출할 토픽의 수
lda = LatentDirichletAllocation(n_components=num_topics, random_state=42)
lda.fit(doc_vectors)
feature_names = tfidf.get_feature_names()
 or topic_idx, topic in enumerate(lda.components_):
   print("Topic #%d:" % topic_idx)
   print(" ".join([feature_names[i]
                   for i in topic.argsort()[:-11:-1]]))
```

토픽모델링 - Top 10 상의



Topic #0:

하프 니트 셔츠 카라 블랙 color 수아레 피그먼트 집업 hoodi

Topic #1:

커버낫 로고 맨투맨 <mark>티셔츠</mark> 어센틱 mmlg 블랙 엠엠엘지 <mark>후디</mark> 네이비

Topic #2:

코드그라피 로고 cgp set 아치 하프집업 럭비 선택 리트리버클럽 예스아이씨

Topic #3:

아웃스탠딩 hoodi 데님 디파트먼트 phyps 에듀케이션 피지컬 디스이즈네버댓 워싱 셔츠

Topic #4:

shirt <mark>티셔츠</mark> 하프 에센셜 집업 오버핏 유니온블루 laundri denim black

Topic #5:

arch hoodi 예일 black logo 23ss 트위치 tone tshirt small

Topic #6:

와릿이즌 logo 아디다스 classic knit 마리떼 엔젤 블랙 sweatshirt half

Topic #7:

폴로 랄프 로렌 셔츠 슬림핏 블루 저지 long sleev 스웨터

Topic #8:

오버핏 dan 예일 crewneck univ 필루미네이트 니트 black sweatshirt navi

Topic #9:

스탠다드 무신사 크루 티셔츠 오버사이즈 긴팔 블랙 반팔 니트 스웨트셔츠

토픽모델링 - Top 10 하의



Topic #0:

와이드 팬츠 테이크이지 <mark>데님</mark> 베이직 우먼스 <mark>블랙</mark> 조거팬츠 엠엘비 꼼파뇨 Topic #1:

팬츠 <mark>와이드 데님 블랙</mark> 스웨트 카고 스탠다드 세미 파라슈트 그레이

Topic #2:

pant denim wide black cargo blue tuck 유니폼브릿지 wash sweat Topic #3:

<mark>와이드</mark> 팬츠 카고 <mark>데님</mark> 시그니처 <mark>블랙</mark> 인디고 크롭 립스탑 기획

Topic #4:

팬츠 <mark>와이드 데님</mark> jean 세미와이드 워시드 데님팬츠 <mark>black</mark> denim 블랙

Topic #5:

팬츠 black pant 코닥 데님 슬림 조거 테이퍼드 우븐 beig

Topic #6:

팬츠 <mark>와이드</mark> 카고 밴딩 <mark>블랙</mark> 코튼 오디너리홀리데이 <mark>데님</mark> 지퍼 나일론

Topic #7:

조거 팬츠 엄브로 로고 pant 트레셔 주앙옴므 숏츠 예일 4부

Topic #8:

pant 트래블 워크샵 belt 그레이 knee 로드 black 아웃스탠딩 엔젤

Topic #9:

스파오 데일리지 fatigu 라퍼지스토어 <mark>데님</mark>팬츠 카펜터 데님 윈드셀 맥우드건 팬츠

토픽모델링 - Top 10 스니커즈



Topic #0:

스니커즈 밸롭 니트 블랙 아키클래식 <mark>화이트</mark> 휠라 엔에프엘 독일군 고어텍스 Topic #1:

아디다스 포럼 로우 스니커즈 오즈위고 컨버스 light militari 화이트 라이트 Topic #2:

스니커즈 폴로 로렌 랄프 레더 오트리 메달리스트 스웨이드 푸마 캔버스 Topic #3:

반스 <mark>화이트</mark> 블랙 스니커즈 밀리터리 어센틱 스탠다드 클래식 아식스 하이 Topic #4:

슈퍼스타 <mark>화이트 아디다스</mark> 스니커즈 스케쳐스 언다이드룸 맥스쿠셔닝 남성 반스 이소 Topic #5:

리복 아이더 <mark>화이트</mark> 스니커즈 캐치볼 하이크 그레이 클럽 엘에이알 독일군 Topic #6:

디스커버리 익스페디션 프로스펙스 캐치볼 플러스 오리지날 노스페이스 v2 버킷 2365 Topic #7:

오니츠카타이거 캔버스 스니커즈 이피티 dive 멕시코 66 블랙 슈펜 슈탠다드 Topic #8:

여성 스케쳐스 워크 스니커즈 남성 white 수페르가 sneaker 화이트 미즈노 Topic #9:

엠엘비 ny 스니커즈 black white 빅볼청키 코닥 청키 ivori 오니츠카타이거

토픽모델링 - Top 10 신발



Topic #0:

블랙 슈즈 bayaband <mark>샌들</mark> 부츠 여성 <mark>크록스</mark> 프리미엄 slide 스퀘어 Topic #1:

엘칸토 남성 <mark>샌들</mark> 여성 블랙 인텐스 마쯔 골스튜디오 grabiti balance Topic #2:

크록스 classic clog 공용 성인 bk 블랙 baya 여성 platform

Topic #3:

닥터마틴 블랙 스무스 에스티코 1461 브라운 구두 3홀 로퍼 <mark>슬라이드</mark>

Topic #4:

사뿐 <mark>메리제인</mark> boot 1cm <mark>플랫슈즈</mark> 마크모크 5cm 오리지널 스퀘어 2cm Topic #5:

슬라이드 블랙 flat <mark>락피쉬웨더웨어</mark> 소가죽 <mark>샌들</mark> 하루타 커스텀에이드 여성 커버낫 Topic #6:

블랙 무신사 닥터마틴 스탠다드 더비 로맨틱무브 헌터 슈즈 첼시 부츠 Topic #7:

아뜨리에 피렌체 black 세스띠 brown 스퀘어토 스웨이드 fg 로퍼 라운드토 Topic #8:

블랙 슬리퍼 정장화 남성 아일랜드슬리퍼 락포트 2365 플레인 pt203 <mark>슬라이드</mark> Topic #9:

2color 블랙 로퍼 3color 리플라 loafer 마크모크 에이티티 레더 <mark>슬라이드</mark>

결론

· 가장 영향력이 큰 요소 = 판매 매출

▶ 구매후기 개수 〉 좋아요 수 〉 조회수 〉 평점

결론

- ▶ 상의:셔츠,후디,티셔츠
- ▶ 하의 : 와이드 팬츠, 블랙 데님, 조거
- ▶ 스니커즈:화이트,블랙,독일군
- ▶ 신발: 샌들, 크록스, 슬라이드, 플랫슈즈