

중고 거래 활성화를 위한

# 알라딘 베스트셀러 데이터 분석

2조 : 권남우, 이형언, 김종현, 김수현, 소국희

## Contents :

### 문제



주제 제시

웹 크롤링

### 고민



코드 작성

데이터 소개

결측값 처리

### 해결



데이터 시각화

데이터 인사이트

### 결과



시사점

한계점





문제

---

주제 제시

웹 크롤링

# What ?



# 지금 중고거래 시장은

<https://www.joongang.co.kr/article/24034559#home>

**“소유보다 경험” 물건 필요하면 중고품부터 찾는 MZ세대**

<중앙일보> 2021.04.14

<https://www.etoday.co.kr/news/view/2095763>

**중고시장 얼마나 커졌길래...롯데 투자하자 신세계도 눈독**

<이투데이> 2022.01.12

<https://m.mk.co.kr/news/business/view/2021/10/1007228/>

**중고거래 큰 인기... 중고거래 플랫폼 수요 늘어나**

<매일경제> 2021.10.25

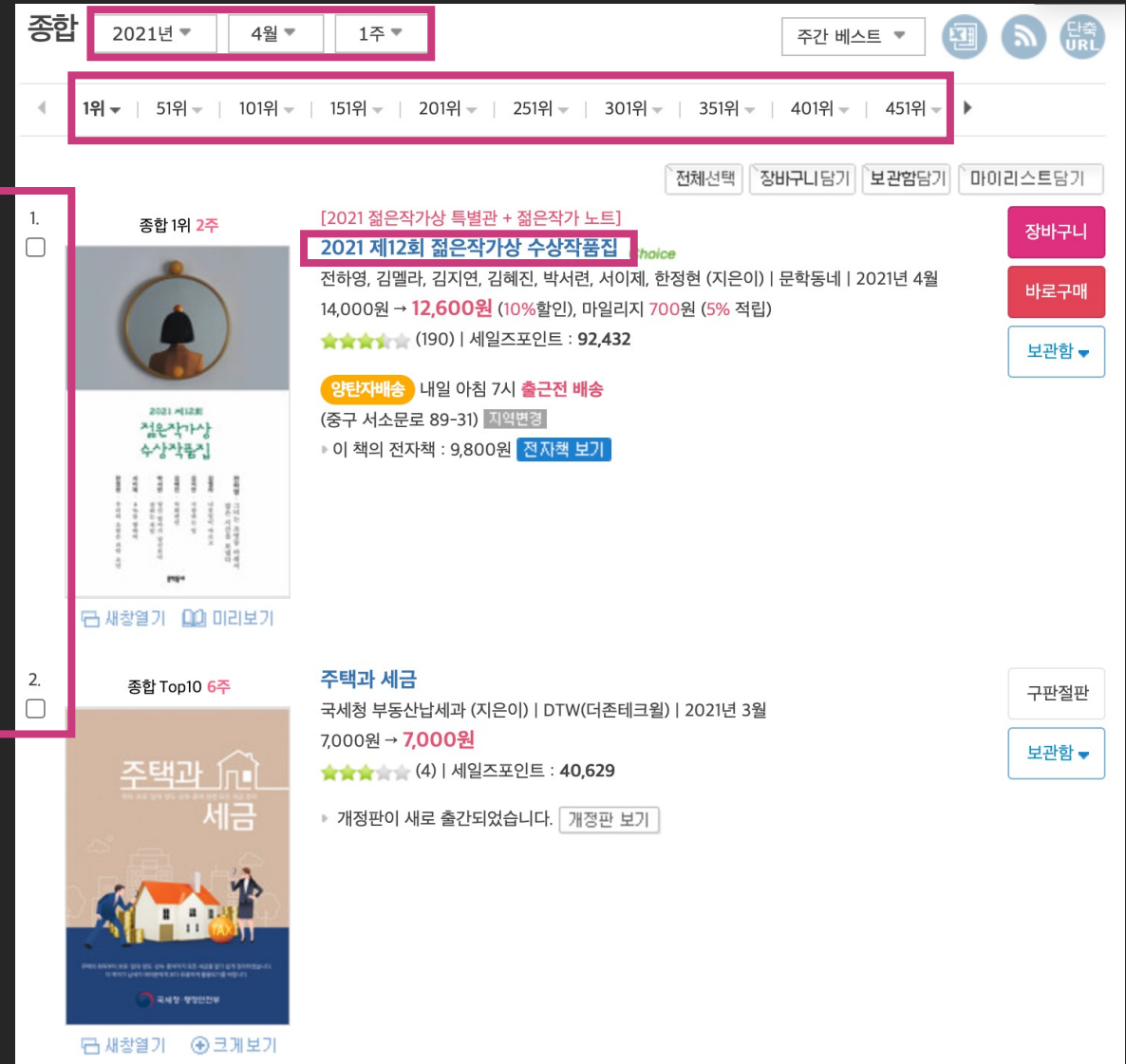
그렇다면 알라딘은?



**매우 중요해졌습니다!**

# 알라딘 사이트에서 필요한 데이터 수집

<베스트 셀러 순위>



<HTML>  
<body>

<div class='베스트 셀러 기간'>  
    <a href> 년도, 월, 주, 순위 <>

<div src='해당 상품 사이트'>  
    <a tag> 상품 명 <>  
    <a href> 해상 도서 ID <>


</body>

# 알라딘 사이트에서 필요한 데이터 수집

<해당 도서 사이트>

달러구트 꿈 백화점 - 주문하신 꿈은 매진입니다

이미예 (지은이) 팩토리나인 2020-07-08



유리 첫잔(이벤트 도서 포함, 국내서 3만5천원 이상)

정가	13,800원
판매가	12,420원 (10%, 1,380원 할인)

마일리지 690원(5%) + 멤버십(3~1%)  
+ 5만원이상 구매시 2,000원

배송료 무료

양탄자배송

밤 10시까지 주문하면 내일 아침 7시 출근전 배송  
(중구 서소문로 89-31 기준) 지역변경

소설/시/희곡 주간 14위, 종합 1위 3주 | Sales Point : 324,794

★★★★★ 8.1 100자평(245) 리뷰(131) [이 책 어때요?](#)

기본정보

300쪽

134\*200mm

357g

ISBN : 9791165341909

주제 분류

신간알리미 신청

국내도서 > 소설/시/희곡 > 판타지/환상문학 > 한국판타지/환상소설

국내도서 > 소설/시/희곡 > 한국소설 > 2000년대 이후 한국소설

국내도서 > 추천도서 > 알라딘 독자 선정 올해의 책 > 2020년 > 올해의 책 TOP 10

<HTML>  
<body>

<div class='도서 관련 정보'>  
<a tag> 제목, 지은이, 출판사 <>

<div class='도서 인기 정보'>  
<a tag> 평점, Sales Point <>

<div class='도서 기본 정보'>  
<a tag> 쪽, 분류 <>

</body>

# 알라딘 사이트에서 필요한 데이터 수집

<해당 도서 중고거래 사이트>

달러구트 꿈 백화점 2 - 단골손님을 찾습니다

이미예 (지은이) | 팩토리나인 | 2021-07-27

새상품 정가13,800원 새상품 상세보기

새상품 판매가12,420원 + 마일리지 690원

알라딘 직접배송 중고(0)-

매장 배송 중고(15)9,200원 (최저가)

판매자 배송 중고(47)8,000원 (최저가)

알라딘에 팔기 예상가

최상	상	중
6,200원	5,500원	4,900원

모든 상품 (62)

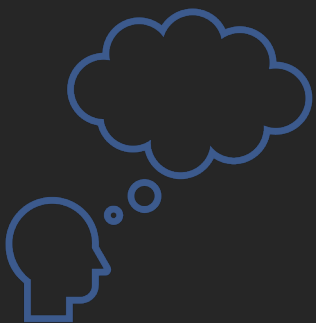
알라딘 직접배송 중고 (0)

중고매장 이 광활한 우주점 (15)

<HTML>  
<body>

```
<div class='중고 도서 거래 가격'>  
  <a tag> 최상인 경우 예상가 <>  
  
<div class='중고 도서 재고 수량'>  
  <a tag> 모든 상품 <>
```

</body>



고민

---

코드 제작

데이터 소개

# How ?





# 수 많은 경우의 수

Q 1-1. 19금 도서는 사이트 접속이 안되요

Q 1-2. 중고 거래 사이트에서 예측가를 가져올 수 없어요

Q 2-1. 출시일 및 출판사가 아니라 '원제' 제목이 출력되요

Q 2-2. 쪽 수가 아니라 '양장본' 및 '기타'가 출력되요

Q 2-3. 정가를 가져오려는 데 정가를 할인해요

Q 3-1. 중고 거래 사이트에서 예상가가 없어요

Q 3-2. 해당 도서 사이트에 지은이 정보가 없어요

Q 3-3. 해당 도서 사이트에 분류 정보가 없어요

Q 4-1. 해당 도서 사이트에 순위 정보가 없어요

If 모든 경우의 수 해결 == True:

def ( 알라딘 웹 크롤링 ) :

```
resp = requests.get(url)
soup = BeautifulSoup(resp.content, 'lxml')
```

# 사이트 접속 및 크롤링 불가

Q 1-1. 19금 도서는 사이트 접속이 안되요



# 19금 사이트는 다른 사이트에 접속 된다

www.aladin.co.kr 내용:  
19세 이상만 이용할 수 있는 상품입니다. 로그인해주세요.

<접속 실패 경우의 조건문을 줘서 해결>

# 만약 태그 값을 가져오지 못하면 Null값을 줘라

```
if poblish_tags != [] :  
    책_정보_딕셔너리['제목'].append(publish_tag.text  
)
```

```
else:  
  
    책_정보_딕셔너리['제목'].append(None)
```

```
if author_tags != [] :  
  
    책_정보_딕셔너리['지은이'].append(author_tag.text)
```

```
else:  
  
    책_정보_딕셔너리['지은이'].append(None)
```

# 사이트 접속 및 크롤링 불가

Q 1-2. 중고 거래 사이트에서 **예측가**를 가져올 수 없어요

<웹 크롤링 실패 시 예외 구문으로 해결 >

# 만약 태그 값을 파싱하지 못하면 Null값을 줘라

```
try:
    책_정보_딕셔너리['예상가'].append(used_tag.text
)
except:
    책_정보_딕셔너리['예상가'].append(None)
```



# 다른 데이터가 존재

Q 2-1. 출시일 및 출판사가 아니라 '원제' 제목이 출력되요

<'원제'가 있는 경우의 조건문을 줘서 해결>

파친코 1Q

이민진 (지은이), 이미정 (옮긴이) 문학사상사 2018-03-23 원제 : Pachinko (2017년)

불편한 편의점 (40만부 기념 벚꽃 에디션)Q

김호연 (지은이) 나무옆의자 2021-04-20

# 만약 '원제'가 있으면 다른 태그 값을 선택

```
if '원제' in sub_title_tags.text :
```

```
    soup.select( 'sub_title' ) [ -2 ]
```

```
else:
```

```
    soup.select( 'sub_title' ) [ -1 ]
```

# 다른 데이터가 존재

Q 2-2. 쪽 수가 아니라 '양장본' 및 '기타'가 출력되요

양장본

808쪽

167\*236mm

1212g

ISBN : 9788933871751

주제 분류

신간알리미 신청

• 국내도서 > 예술/대중문화 > 영화/드라마 > 시나리오/시나리오작법

기타

1쪽

180\*180mm

532g

ISBN : 8809731784249

주제 분류

신간알리미 신청

• 국내도서 > 유아 > 놀이책 > 스티커북

268쪽

135\*200 mm

348g

ISBN : 9791161571188

주제 분류

신간알리미 신청

• 국내도서 > 소설/시/희곡 > 한국소설 > 2000년대 이후 한국소설

<'양장본' 및 '기타'가 있는 경우의  
조건문을 줘서 해결>

# 만약 '양장본'이 있으면 다른 태그 값을 선택

if page\_tag == '양장본' :

soup.select( 'book\_info\_list' ) [ 1 ]

# 만약 '기타'가 있으면 다른 태그 값을 선택

elif page\_tag == '기타' :

soup.select( 'book\_info\_list' ) [ 1 ]

else:

책\_정보\_딕셔너리['쪽'].append( page\_tag )

# 다른 데이터가 존재

Q 2-3. 정가를 가져오려는 데 정가를 할인해요

<정가 할인 경우의 예외 구문으로 해결>

정가	15,800원
판매가	14,220원 (10%, 1,580원 할인)

# 정가를 할인하면 태그가 달라진다

정가	<del>13,800원</del> → 4,400원 (68%↓)	정가인하
판매가	3,960원 (10%, 440원 할인)	

# 정가 태그 크롤링 실패하면 다른 태그 선택

try:

soup.select( 'original\_price' )

except:

soup.select( 'discounted\_price' )

# 데이터가 비존재

Q 3-1. 중고 거래 사이트에서 **예상가**가 없어요

<예상가가 없는 경우의 조건문으로 해결>

알라딘에 팔기 예상가

최상	상	중
6,200원	5,500원	4,900원

# **예상가가 없는 경우 '-'로 표기된다**

알라딘에 팔기 예상가

최상	상	중
-	-	-

# '예상가'가 '-' 인 경우 Null 값을 준다

```
if used_tags.text == '-':  
    책_정보_딕셔너리['예상가'].append(None)  
  
else:  
    책_정보_딕셔너리['예상가'].append(used_tags.text)
```

# 데이터가 비존재

Q 3-2. 해당 도서 사이트에 **지은이** 정보가 없어요

<지은이에 글자가 없는 경우의 조건문으로 해결>

바라카몬 12🔍

(지은이) 대원씨아이(만화) 2016-02-18

그 해 우리는 포토 에세이🔍

스튜디오S (지은이) 김영사 2022-05-10

# 지은이에 글자가 없으면 글자 수는 0이다

```
if len(author_tags.text) == 0 :
```

```
    책_정보_딕셔너리['지은이'].append(None)
```

```
else:
```

```
    책_정보_딕셔너리['지은이'].append(author_tags.text)
```



# 데이터가 비존재

Q 3-3. 해당 도서 사이트에 **분류** 정보가 없어요

기본정보

300쪽 140\*205mm 442g ISBN : 9791191347685

주제 분류

신간알리미 신청

- 국내도서 > 인문학 > 심리학/정신분석학 > 교양 심리학
- 국내도서 > 자기계발 > 인간관계 > 교양심리학

상품 분류

원산지 : 대한민국

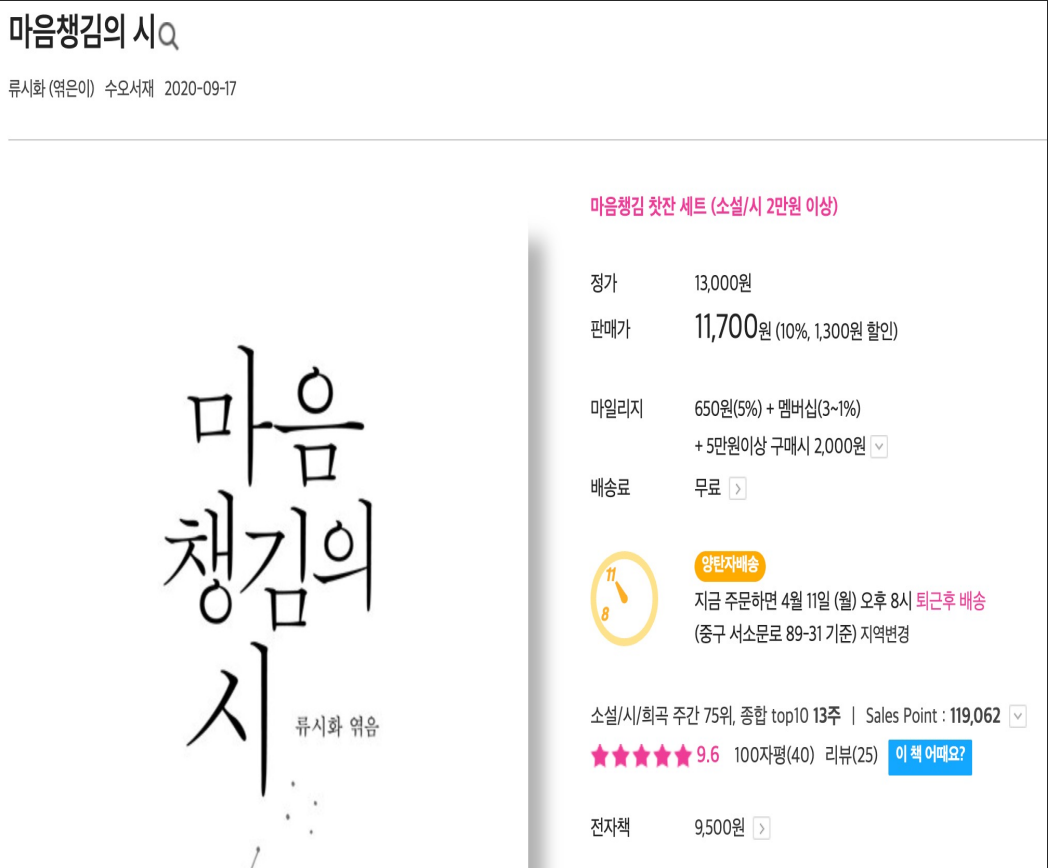
<분류 데이터가 없는 경우의 조건문을 줘서 해결>

# 만약 태그 값을 가져오지 못하면 Null값을 줘라

```
if category_tags != [] :  
    책_정보_딕셔너리['분류'].append(category_tag.text)  
  
else:  
    책_정보_딕셔너리['분류'].append(None)
```

Q 4-1. 해당 도서 사이트에 순위 정보가 없어요

# 해당 도서 사이트에는  
# 원하는 기간의 순위 정보가 없다



<도서 하나를 크롤링할 때마다 순위 + 1>

# 해당 도서의 ID를 통해 사이트를 접속할 때 마다  
# rank\_num에 1을 더한다

딕셔너리\_제목\_ID = {제목 : ID}

rank\_num = 1

for 제목 in (list(딕셔너리\_제목\_ID.keys()))

url = 'www.알라딘.com?책아이디=' + 딕셔너리\_제목\_ID[제목]  
resp = requests.get(url)  
Soup = BeautifulSoup(resp.content, 'lxml')

책\_정보\_딕셔너리['순위'].append(rank\_num)

rank\_num += 1

# 최종 코드

```
# 알라딘 사이트에서 원하는 기간의 베스트 셀러를 딕셔너리 형태로 반환해주는 함수
def aladin_BestSeller(year, month, week, rank):
    # year : 해당 년도의 베스트 셀러
    # month : 해당 달의 베스트 셀러
    # week : 해당 주의 베스트 셀러
    # rank : 순위 설정 (최대 20 -> 1000위까지)
    # 1 : 1 ~ 50위 / 2 : 1 ~ 100위 / 3 : 1 ~ 150위 ...

    # 필요한 라이브러리 불러오기
    from tqdm.notebook import tqdm
    import requests
    from bs4 import BeautifulSoup

    # 해당 년, 달, 주, 순위의 베스트셀러 정보를 담은 딕셔너리
    dict_best_seller_info = {
        '순위':[], # 해당 기간 베스트 셀러 순위
        '출시':[], # 베스트셀러 출시 날짜
        '제목':[], # 베스트셀러 제목
        '지은이':[], # 베스트셀러 작가
        '출판사':[], # 베스트셀러 출판사
        '쪽':[], # 베스트셀러 쪽
        '분류':[], # 베스트셀러 카테고리
        '정가':[], # 베스트셀러 정가
        '베스트셀러 기간':[], # 설정한 베스트 셀러 기간 (년도-달-주)
        '중고 판매 예상가':[], # 알라딘 팔기 예상가 (최상)
        '중고 재고':[], # 중고 서적 재고 량
        '평점':[], # 10점 만점
        'Sales Point':[] # 판매량과 판매기간에 근거하여 해당 상품의 판매도를 산출한 알라딘만의 판매지수법
    }
```

def 알라딘(year, month, week, rank) :

순위, 출시, 제목, 지은이, 출판사,

쪽, 분류, 정가, 베스트셀러 기간,

중고 판매 예상가, 중고 재고, 평점,

Sales Point

# 데이터 소개

Columns : 13개 / Rows : 11,206개

Float / Str / Int

**순위** : 설정한 베스트 셀러 기간에서의 해당 도서의 순위

**정가** : 「출판문화 산업 진흥법」에서 정의한 소비자에게 판매하는 가격

**출시** : 해당 도서의 출판 날짜

**베스트셀러 기간** : 설정한 베스트 셀러 기간 (년도 - 달 - 주)

**제목** : 해당 도서의 알려진 온라인 매장에서의 판매 제목

**중고 판매 예상가** : 도서 상태가 최상인 경우, 알려진 팔기 예상가

**지은이** : 해당 도서의 작가 (다수인 경우 맨 앞 분만 표기)

**중고 재고** :

알라딘 직접배송 중고, 중고매장 이 광활한 우주점, 판매자 배송 중고 재고수량 총 합

**출판사** : 해당 도서의 출판사

**베스트셀러 기간** : 알려진 소비자가 부여한 해당 도서의 평가 점수 (10점 만점)

**쪽** : 해당 도서의 모든 쪽수

**Sales Point** : 판매량과 판매기간에 근거한 해당 도서의 판매도 (최근 판매분에 가중치)

**분류** : 알라딘에서 설정한 해당 도서의 두 번째 세부 분류

# 수 많은 결측값

## 베스트셀러 데이터

11,431 개

2011년부터 2021년  
3, 6, 9, 12월 5주차  
1위부터 300위

## 결측값

중고 판매 예상가 : 4,034 개

중고 재고 : 370 개

쪽 : 248 개

분류 및 지은이 : 217 개

출시 및 출판사 : 204개

정가 : 187 개

순위, 베스트 셀러 기간, 평점, Sales point : 177 개

베스트셀러 데이터 - 결측값 = 7,367 개

# 결측값 대치

$$\text{중고 판매 예상가} = (\text{기본 매입가} \times \text{상태 지수}) + \text{신간 인센티브}(\%)$$

**기본 매입가** : 알려진 보유 재고량에 따라 차등 적용

**상태 지수** : 도서 상태 최상급 / 상급 / 중급 별도 가중치

**신간 인센티브** : 도서정가제법상 신간인 도서에 관한 판매량에 따라 차등 적용

\* 신간 : 발행일로부터 12개월이 지나지 아니함 (출판문화산업 진흥법 제22조 2항)

**균일가 매입** : 충분한 재고, 원상품의 인기도, 분야의 특성을 감안하여 재판매 가능성이 낮은 경우 균일가 적용

# 결측값 대치

알라딘 팔기 예상가 = 재고 수량 + 신간 여부에 따른 Sales point + (정가, 평점, 분야)

**기본 매입가** : 알라딘 보유 재고량에 따라 차등 적용

→ 재고 수량

**상태 지수** : 도서 상태 최상급 / 상급 / 중급 별도 가중치

→ 최상인 경우만 고려

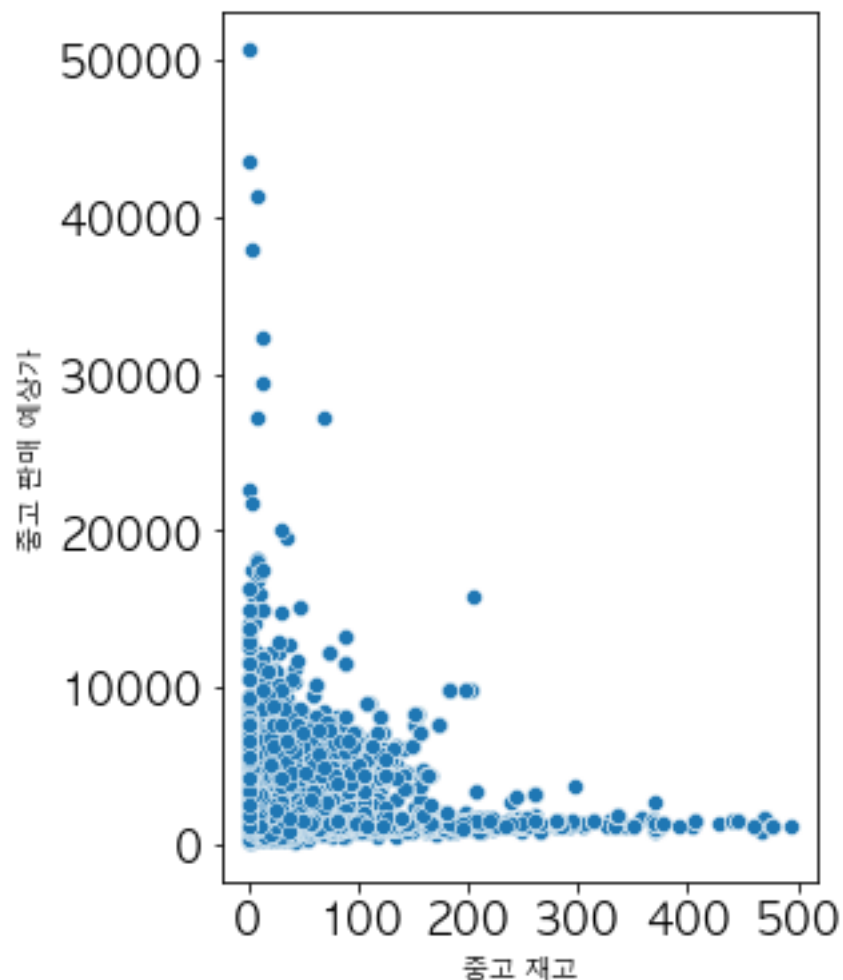
**신간 인센티브** : 도서정가제법상 신간인 도서에 관한 판매량에 따라 차등 적용

→ (출시 - 베스트 셀러 기간 < 12) ▶ 신간 여부

→ 판매량 ▶ Sales Point

# 중간 재고 및 중고 판매 예상가 상관 관계

중간 재고 및 중고 판매 예상가 산점도



기본 매입가 : 알려진 보유 재고량에 따라 차등 적용

균일가 매입 :

충분한 재고를 감안하여 재판매 가능성이 낮은 경우 균일가 적용

**중고 재고가 적을수록 :**

중고 판매 예상가가 높은 도서가 많다

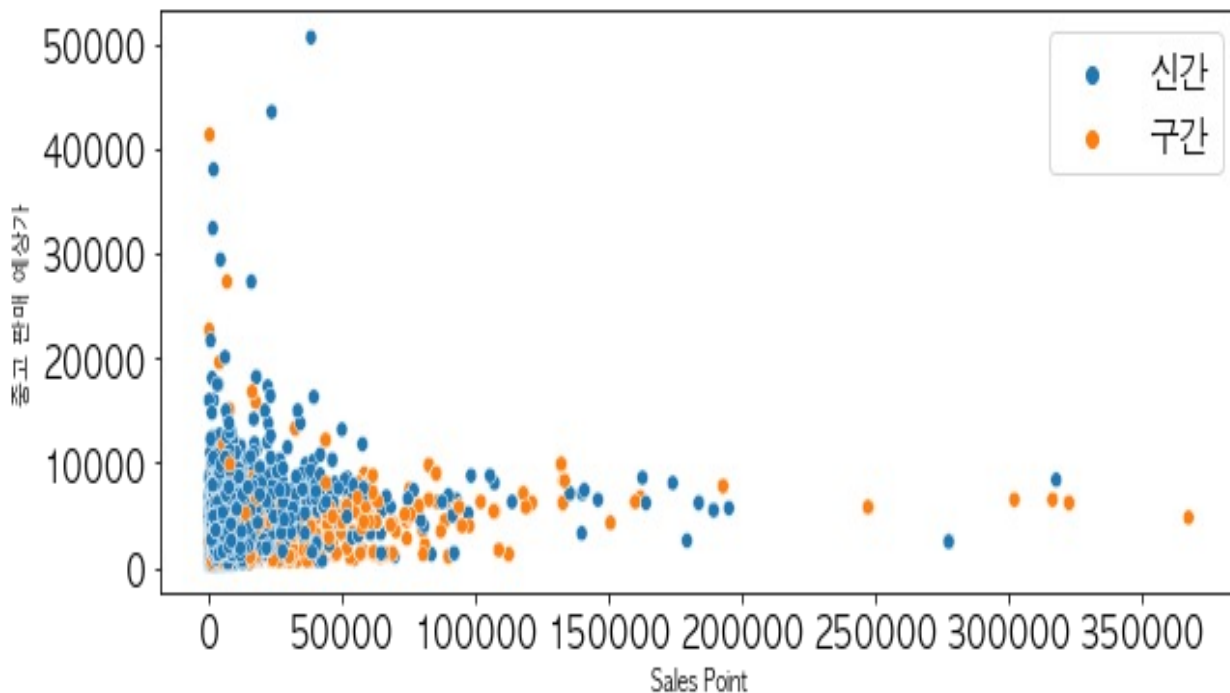
**중고 재고가 많을수록 :**

중고 판매 예상가가 낮은 도서가 많다



# 신간 도서 판매량 및 중고 판매 예상가 상관 관계

Sales Point 및 중고 판매 예상가 산점도



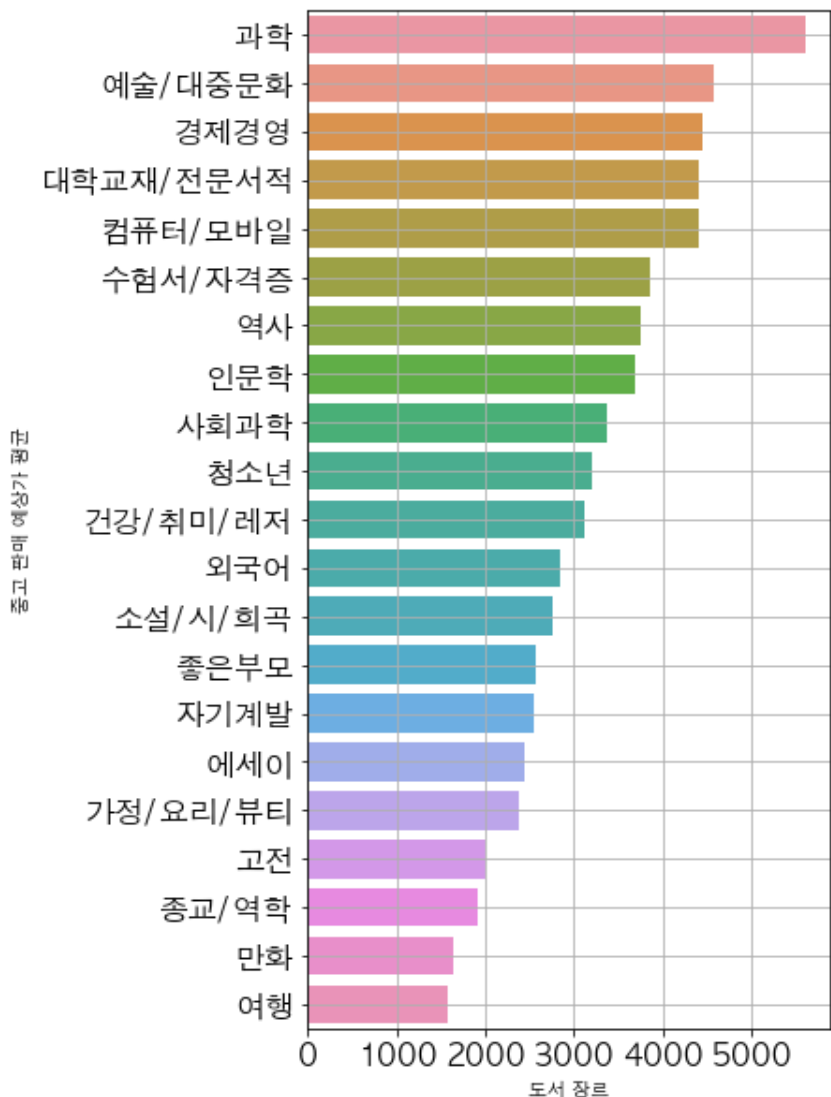
신간 인센티브 :

신간인 도서에 관한 판매량에 따라 차등 적용

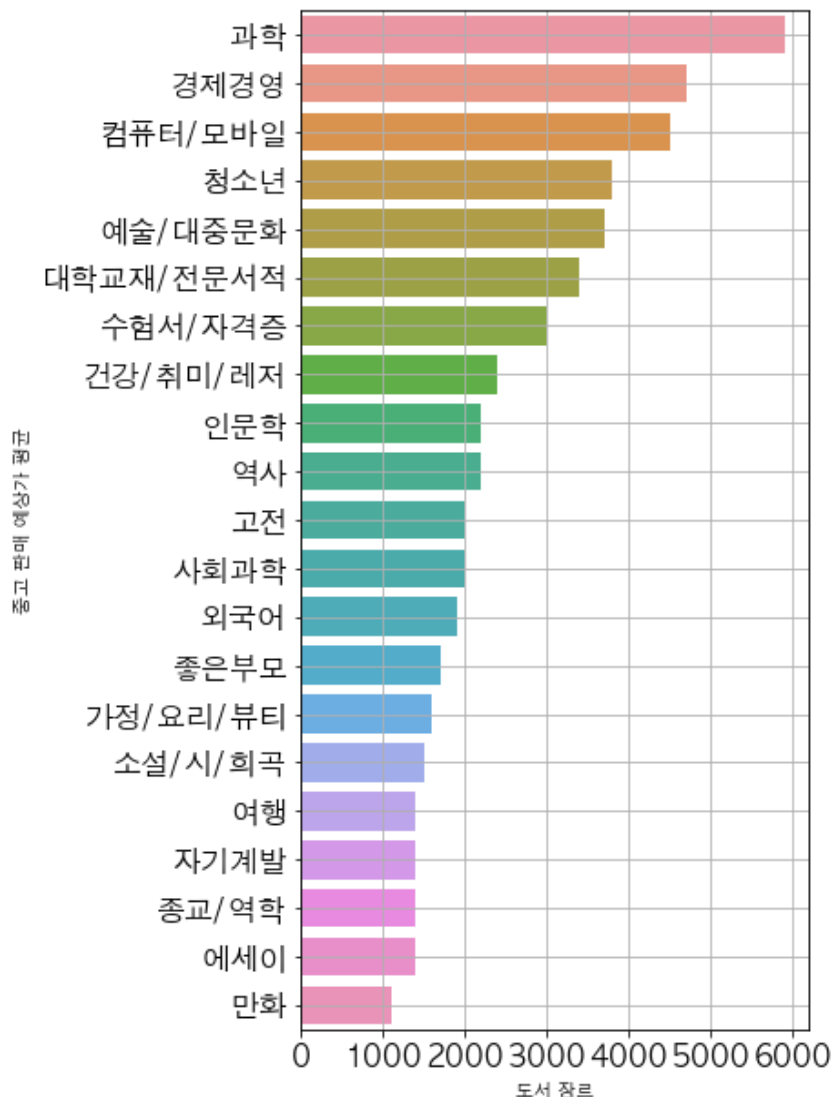
신간인 도서가 Sales Point가 상대적으로 낮은 반면에  
중고 판매 예상가는 높은 도서가 많은 경향이 있다

# 분야 별 중고 판매 가격

분야 별 중고 판매 가격 평균



분야 별 중고 판매 가격 중간값



균일가 매입 :

분야의 특성을 감안하여  
재판매 가능성이 낮은 경우 균일가 적용

상위권 :

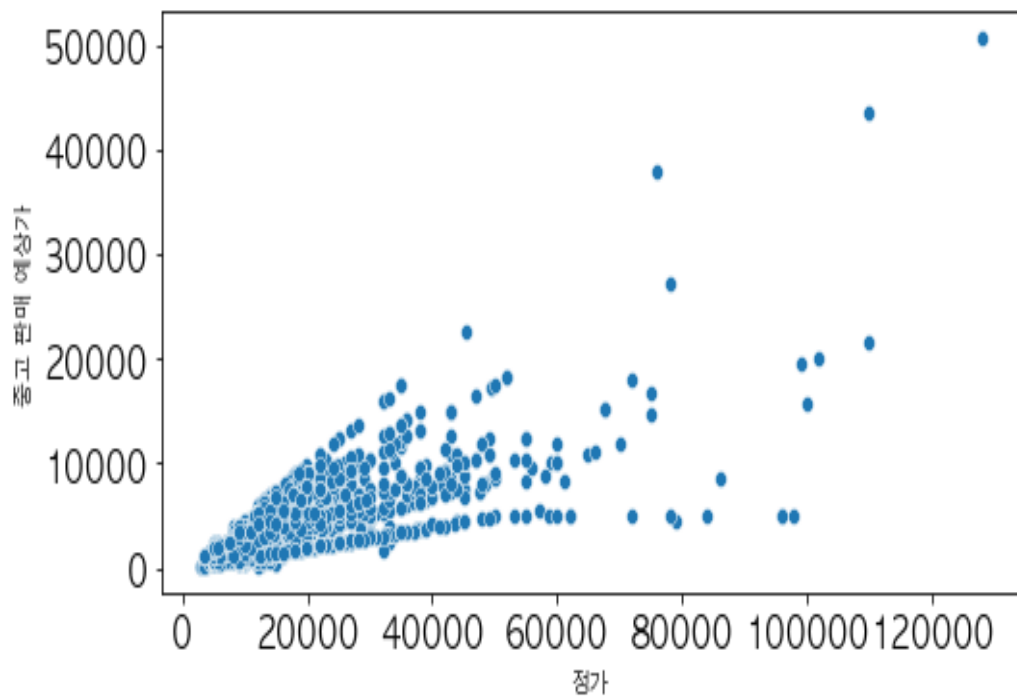
과학, 예술/대중문화, 경제경영, 대학 교재

하위권 :

만화, 종교, 여행

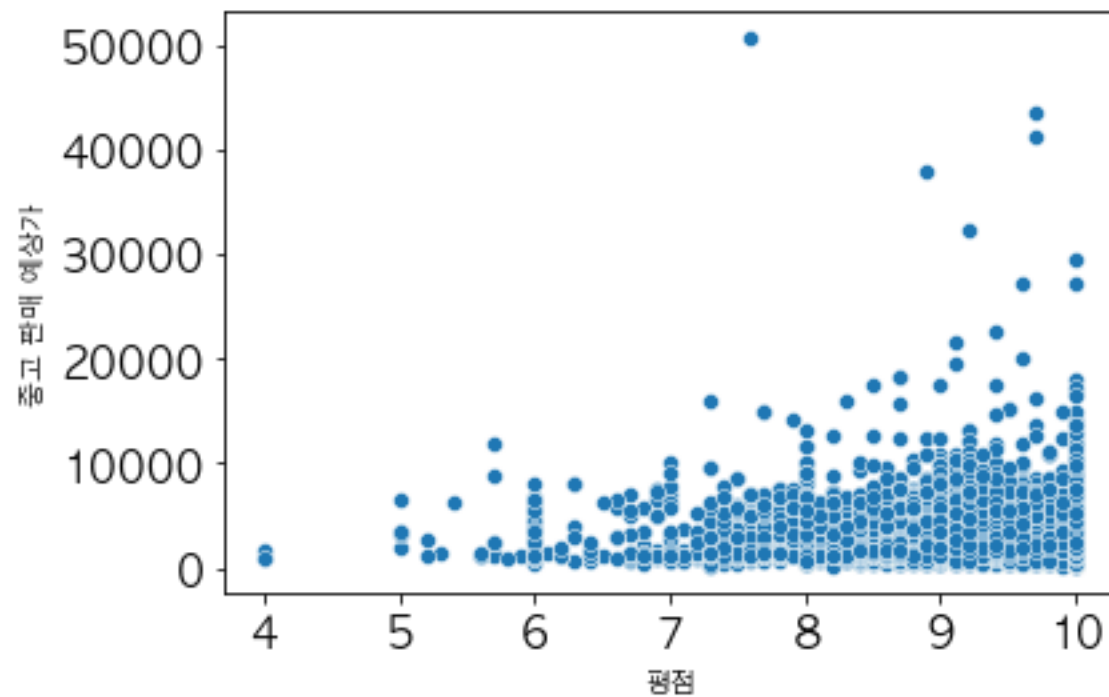
# 정가 및 평점의 중고 판매 예상가 상관 관계

## 정가 및 중고 판매 예상가 산점도



정가가 높을수록 중고 판매 예상가가 높아지는 경향이 있다

## 평점 및 중고 판매 예상가 산점도



평점이 높을수록 중고 판매 예상가가 높은 도서가 많아지는 경향이 있다

# 머신러닝을 활용한 결측값 대처

## Input

Standard Scaler 적용

재고 수량

신간 여부

Sales Point

정가

평점

분야

## 랜덤 포레스트 회귀 모델

max\_depth : 7

max\_leaf\_nodes : 50

Min\_sample\_leaf : 5

n\_estimators : 50

max\_features :  
auto, sqrt, log2

교차 검증 : 3

## Output

중고 판매 예상가

평균 절대 오차

학습 : 749.05 원

검증 : 895.22 원

3,542 개 데이터 추가 확보 (총 10,847 개)



해결

---

데이터 시각화

데이터 인사이트

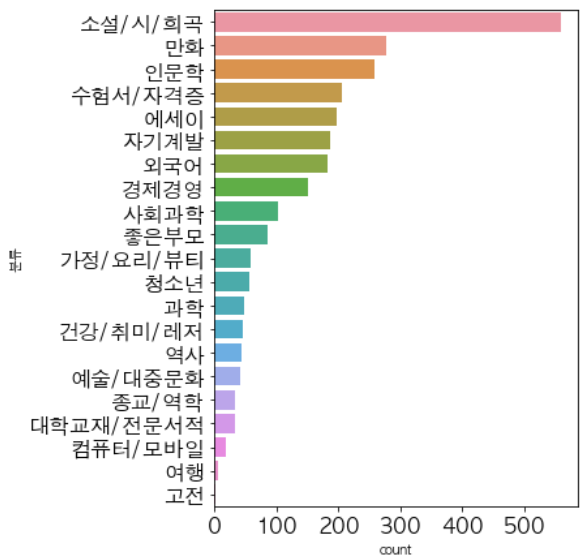
# Why ?



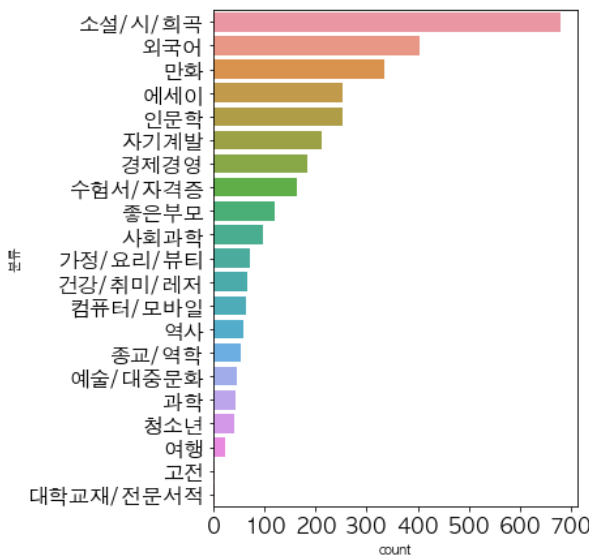
# 데이터 시각화

## 베스트 셀러 전체 년도 3, 6, 9, 12월 도서 분야 빈도 수

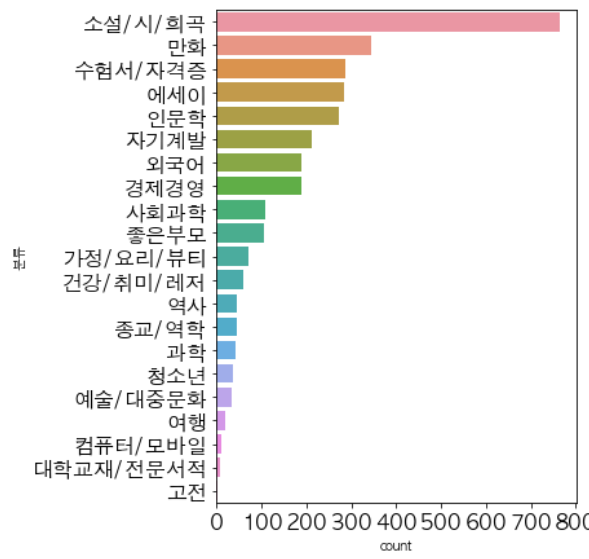
3 월



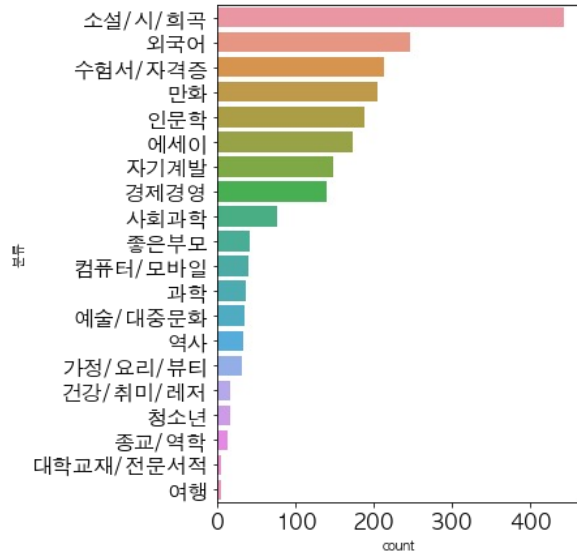
6 월



9 월



12 월



소설/시/희곡 모든 달 1등

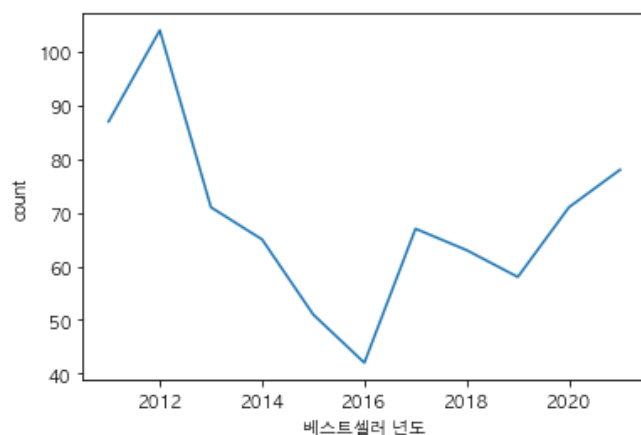
3월 & 9월은 만화가 2등

6월 & 12월은 외국어가 2등

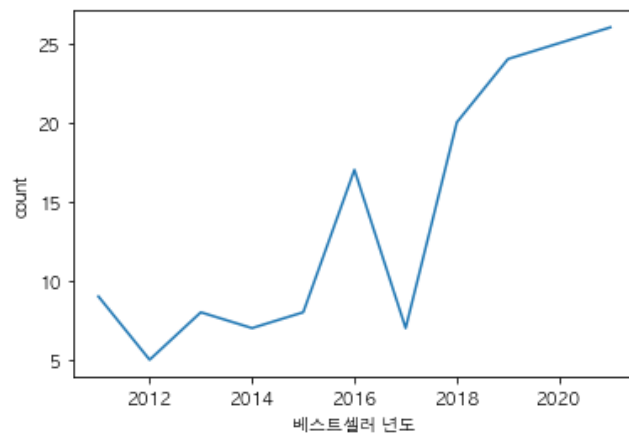
# 데이터 시각화

## 연도 별 베스트 셀러 도서 분야 빈도수 분석 (상승세)

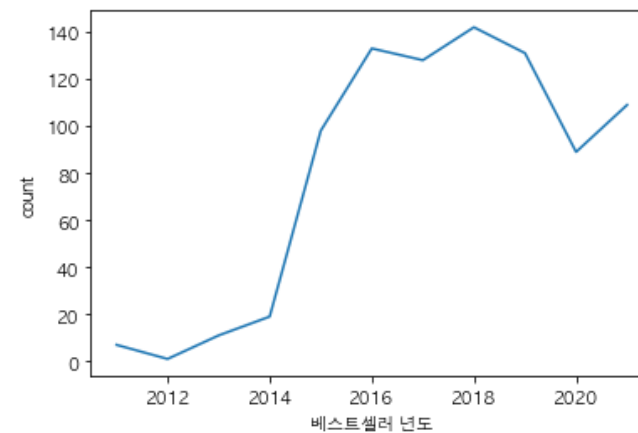
자기 계발



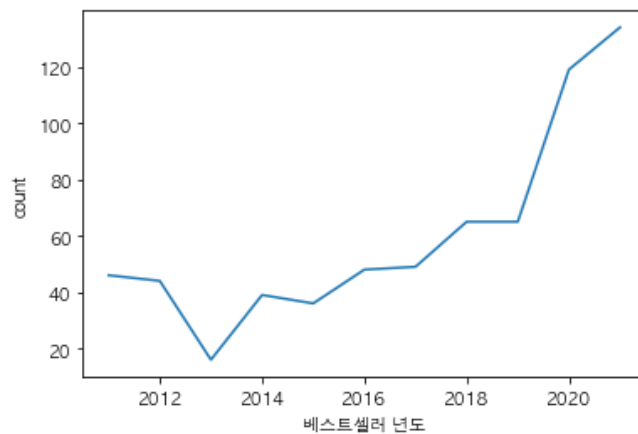
예술 / 대중문화



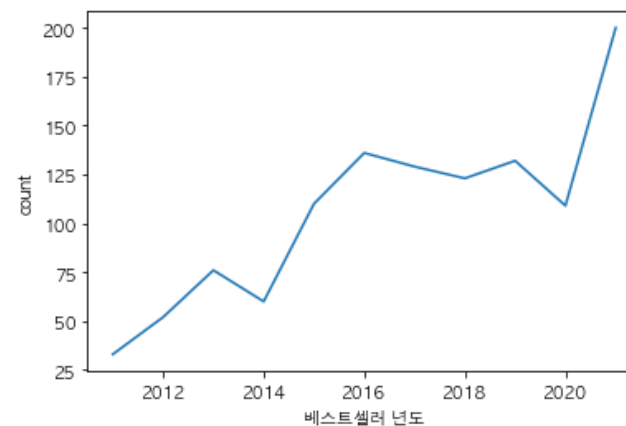
수험서 / 자격증



경제경영



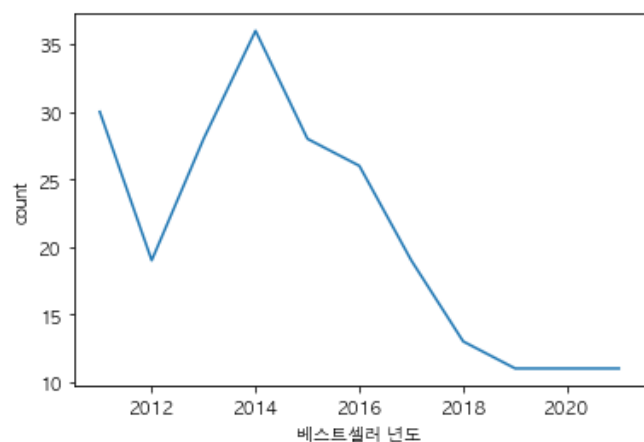
만화



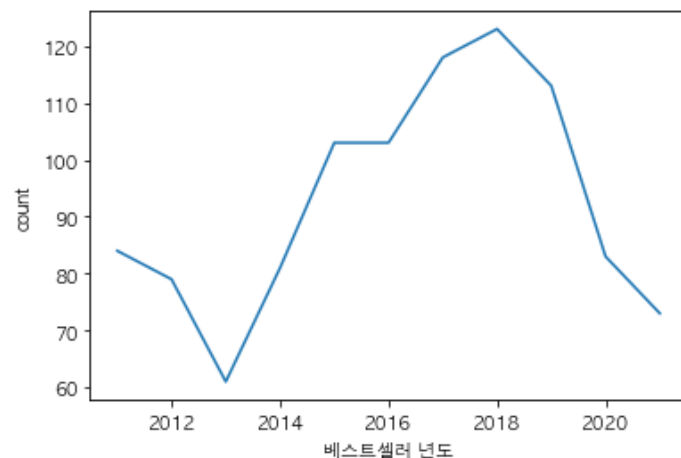
# 데이터 시각화

## 연도 별 베스트 셀러 도서 분야 빈도수 분석 (하락세)

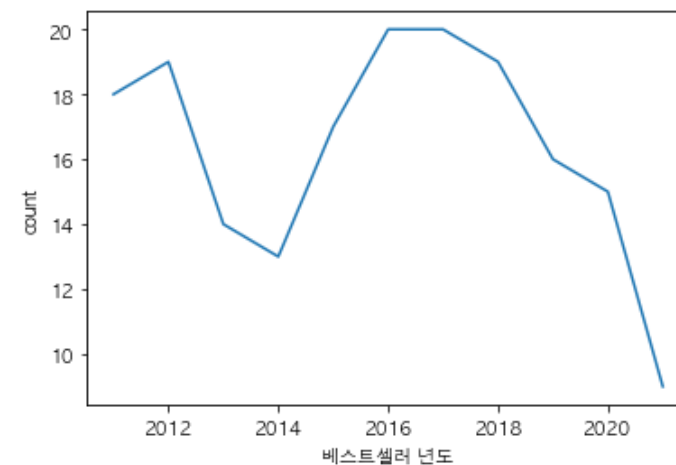
가정 / 요리 / 뷰티



외국어



역사



에세이



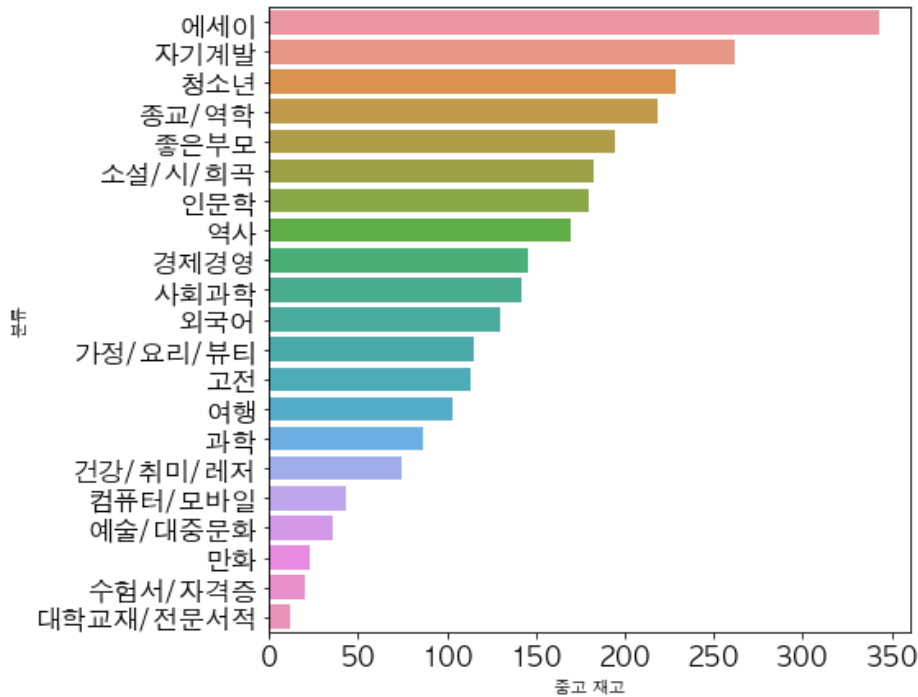
소설 / 시 / 희곡





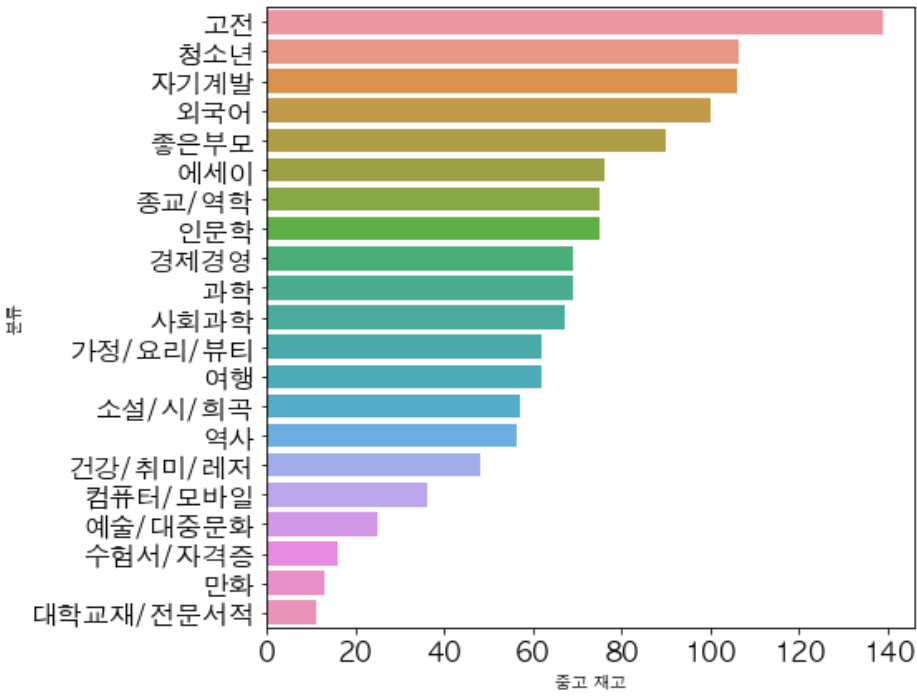
## 도서 분야 별 중고 재고 수

도서 분야 별 중고 재고 수 평균

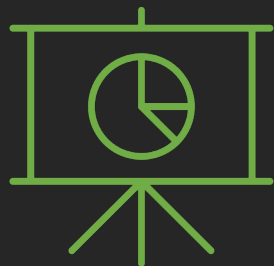


청소년, 자기계발, 좋은 부모는  
중고 재고가 많다

도서 분야 별 중고 재고 수 중간값



대학교재/전문서적, 만화, 수험서/자격증은  
중고 재고가 적다



결과

---

시사점

한계점

What How Why  
and then



# 시사점 및 한계점

## 시사점 :

1. 달 별로 인기 도서가 변화는 패턴을 보인다
2. 2019년 및 2020년 기준으로 인기 도서 분야에 변화가 보인다
3. 특정 도서 분야에서 중고 재고량에 차이가 보인다

## 한계점 :

1. 알라딘 웹 자체의 데이터가 온전하지 못하다
2. 실제 데이터가 아닌 예측 데이터로 분석하였다

Q & A



감사합니다

