# 프로그래밍 언어별 출판도서목록 분석

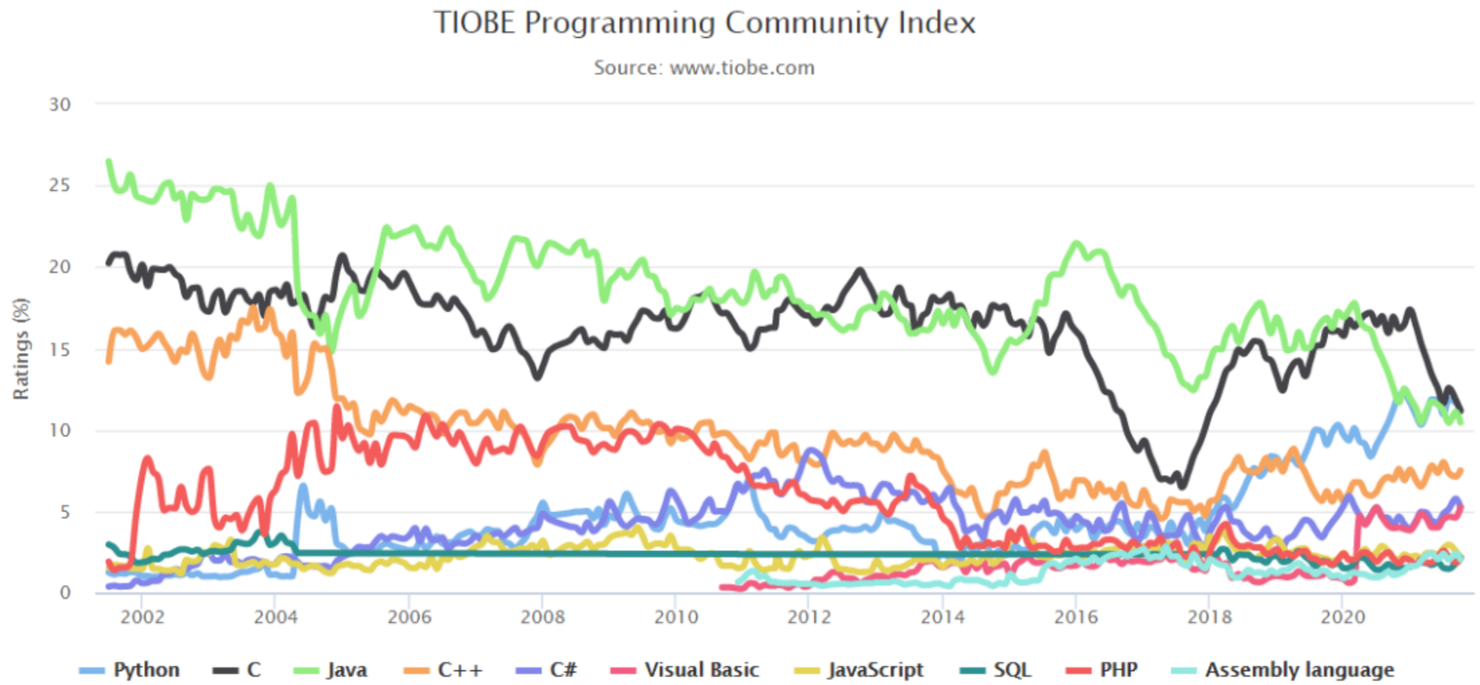
Analysis of list of books already published by Programming Language

2021. 10. 22.

· 프로젝트 수행자 : Luke Lee, 이민호

Python(Jupyter Notebook) / Web Crawling / Data Visualization

# 현재 세계적으로 가장 인기 있는 언어는 'Python'




| Oct 2021 | Oct 2020 | Change | Programming Language | Ratings | Change |
|---|---|---|---|---|---|
| 1 | 3 | ↑ | Python | 11.27% | -0.00% |
| 2 | 1 | ↓ | C | 11.16% | -5.79% |
| 3 | 2 | ↓ | Java | 10.46% | -2.11% |
| 4 | 4 | | C++ | 7.50% | +0.57% |
| 5 | 5 | | C# | 5.26% | +1.10% |
| 6 | 6 | | Visual Basic | 5.24% | +1.27% |
| 7 | 7 | | JavaScript | 2.19% | +0.05% |
| 8 | 10 | ↑ | SQL | 2.17% | +0.61% |
| 9 | 8 | ↓ | PHP | 2.10% | +0.01% |
| 10 | 17 | ⇈ | Assembly language | 2.06% | +0.99% |
| 11 | 19 | ⇈ | Classic Visual Basic | 1.83% | +1.06% |
| 12 | 14 | ↑ | Go | 1.28% | +0.13% |
| 13 | 15 | ↑ | MATLAB | 1.20% | +0.08% |
| 14 | 9 | ⇊ | R | 1.20% | -0.79% |

(좌) 프로그래밍 언어별 사용 비율 추이, (우) 2021년 10월 기준 프로그래밍 언어 사용 순위
자료 : TIOBE 공식 홈페이지 https://www.tiobe.com/tiobe-index/

- TIOBE 공식자료에 따르면, 과거 강세를 보이던 Java의 사용이 최근 감소하는 추세에 있다.
- Python은 5년 전에는 다른 언어들과 차이가 많지 않았지만, 최근 사용이 늘면서 2021년 10월 기준 1위 언어로 랭크되었다.
- C 언어는 5년 사이에 급격하게 늘었다가, 또 최근에 다시 급격하게 감소하는 추세에 있다. 하지만 여전히 강세이다.

# Python에 대한 인기도를 체감할 수 있는 방법, 서점


◀ 서점 내 Python 관련 도서가 베스트셀러에 비치되어 있음
사진 : 알고리즘 코딩 테스트 책 출간
https://docs.likejazz.com/algorithm-interview/

서점을 방문하면, Python 등 프로그래밍 언어에 대한 도서가 많이 있다. 또한, 출판되는 서적의 종류가 굉장히 많은 것을 알 수 있다.

- 지금 가장 인기 있는 언어일수록 출판되는 도서의 양도 많을까?
- 언어별로 어떤 언어의 책이 가장 많이 출판될까?
- 예전에 많이 출판된 도서와 지금 많이 출판되는 도서는 달라졌을까?

**그래서,**
**프로그래밍 언어별 출판도서목록을 분석했다.**

## 검증을 위한 가설 설정

# "전세계 언어별 사용 니즈가 많을수록,
# 국내에서 출판되는 책의 수도 많을 것이다."

**[가설 설정 근거]**

- 전세계적으로 인기 있는 언어라면, 국내 사용자들도 해당 언어에 대한 니즈가 높을 것이라고 생각한다.
- 니즈가 높다는 말은, 해당 언어를 사용하기 위해 학습하려는 사용자가 많다는 의미이다.
- 학습을 위한 도서 수요 및 구매가 상승할 것이며, 이를 통해 인기 있는 언어와 관련한 책이 많이 기획, 출판되고 있을 것이라고 판단

# 프로젝트 진행 개요

**1. 수집 데이터 : 네이버 책 페이지 내 프로그래밍 언어 관련 출간된 책 정보**

- NAVER API 정보 호출 : 책 제목, 출판 시기, 판매 가격, 출판사, ISBN 번호
- 한국 도서만 고려했기 때문에, "언어명 + 프로그래밍"으로 검색 ex) Python 프로그래밍
- NAVER API는 검색어당 1천 개로 제한되므로, 각 언어별 최대 1천 개까지 데이터 수집 진행
- '페이지 수' 정보는 API에 포함되지 않아서, BeautifulSoup 활용
- 프로그래밍 언어는 ["Python", "C", "Java", "C++", "C#", "Visual Basic", "JavaScript", "SQL", "PHP", "R"] 으로 한정 (10개)

**2. 데이터 전처리 : 출판연도 정보 추가, 최근 5년간의 데이터로 초점화, 언어 중복 제거**

- 연도별 경향성을 파악하기 위하여, 출판연도 관련 정보를 추출하여 데이터화
- 최근 5년간의 경향성을 파악하기 위해, 데이터를 2017 ~ 2021년도 출판 도서로 포커스함 (2017년 이후 도서만 정리)
- 도서 내 다수의 언어 내용을 포함하는 도서는 분석의 타당성을 높이기 위하여 중복 처리하여 모두 제거함

**3. 분석 및 시각화 : 가설 검증을 위한 데이터 시각화, 기타 출판도서목록에 대한 상관관계 분석**

- 각 언어(10개)별 출판도서 수 분석 - 5개년도 전체 기간 분석, 기간별로 나누어 별도로 분석, TIOBE 인덱스와 비교
- 기타1 : 출판사(Top 10)별 출판도서 수 분석 및 시각화
- 기타2 : 최근(2020-2021) vs 그 이전(2017-2019) 사이의 언어별 도서 관련 데이터의 변화
- 기타3 : 도서 가격과 페이지 수 간의 상관관계

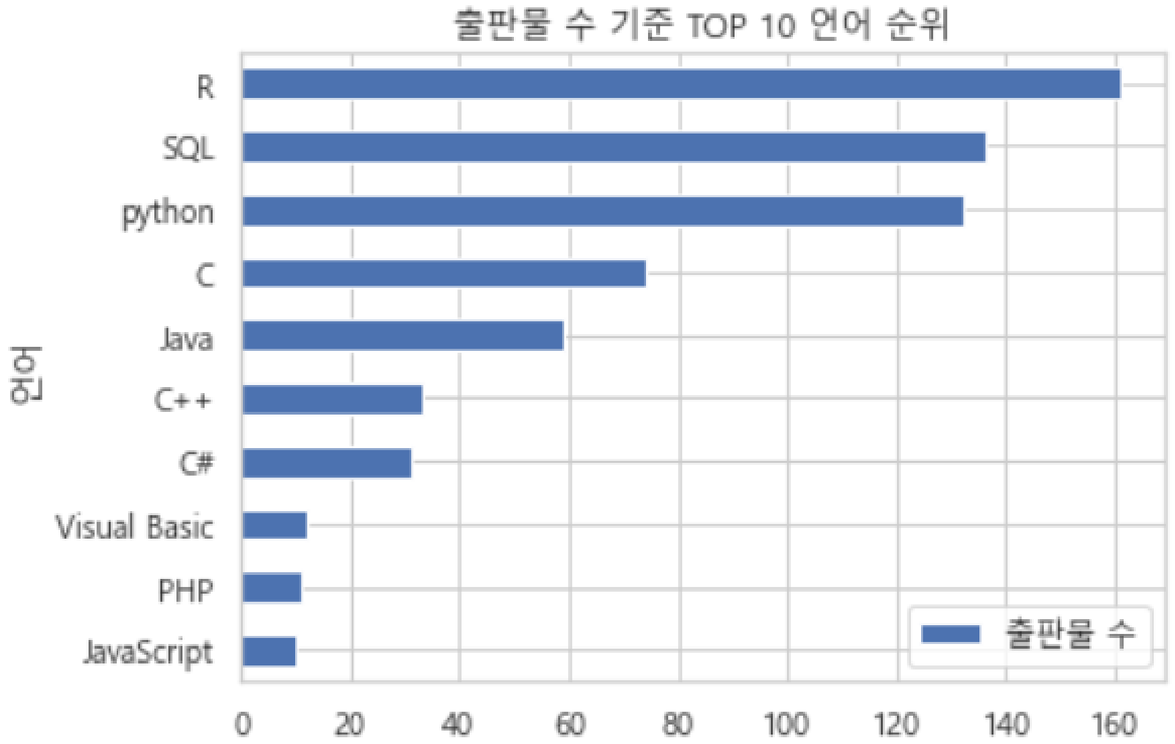
# 프로그래밍 언어별 출판도서 수 분석

| 연도 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|
| 언어 | | | | | |
| C | 16.0 | 16.0 | 12.0 | 10.0 | 20.0 |
| C# | 10.0 | 5.0 | 3.0 | 7.0 | 6.0 |
| C++ | 10.0 | 8.0 | 2.0 | 12.0 | 1.0 |
| Java | 14.0 | 13.0 | 13.0 | 7.0 | 12.0 |
| JavaScript | 3.0 | 3.0 | 1.0 | NaN | 3.0 |
| PHP | NaN | 4.0 | 2.0 | 3.0 | 2.0 |
| R | 35.0 | 33.0 | 32.0 | 35.0 | 26.0 |
| SQL | 17.0 | 27.0 | 22.0 | 38.0 | 32.0 |
| Visual Basic | 6.0 | 1.0 | 1.0 | 2.0 | 2.0 |
| python | 14.0 | 23.0 | 26.0 | 44.0 | 25.0 |

단순히 연도별로 각 언어별 출판도서 수를 정리한 피봇테이블

| 언어 | 출판물 수 |
|---|---|
| R | 161 |
| SQL | 136 |
| python | 132 |
| C | 74 |
| Java | 59 |
| C++ | 33 |
| C# | 31 |
| Visual Basic | 12 |
| PHP | 11 |
| JavaScript | 10 |

5개년도 언어별 출판도서 수




5개년도 언어별 출판도서 수(내림차순 정렬) 바(Barh) 그래프 시각화

- 5개년도 기준으로 보았을 때, R, SQL, Python 순으로 많은 출판물 수를 기록하고 있다.
- SQL의 경우는 전세계 경향성으로 보면, 많은 니즈가 있는 언어가 아닌데 불구하고, 국내 출판도서는 많은 것으로 나타났다.
- R도 PIOBE 인덱스 상으로는 14위를 기록하고 있으나, 국내 도서는 상당히 많이 출판되고 있다. (2020년 이후에는 소폭 줄어든 것으로 보인다.)

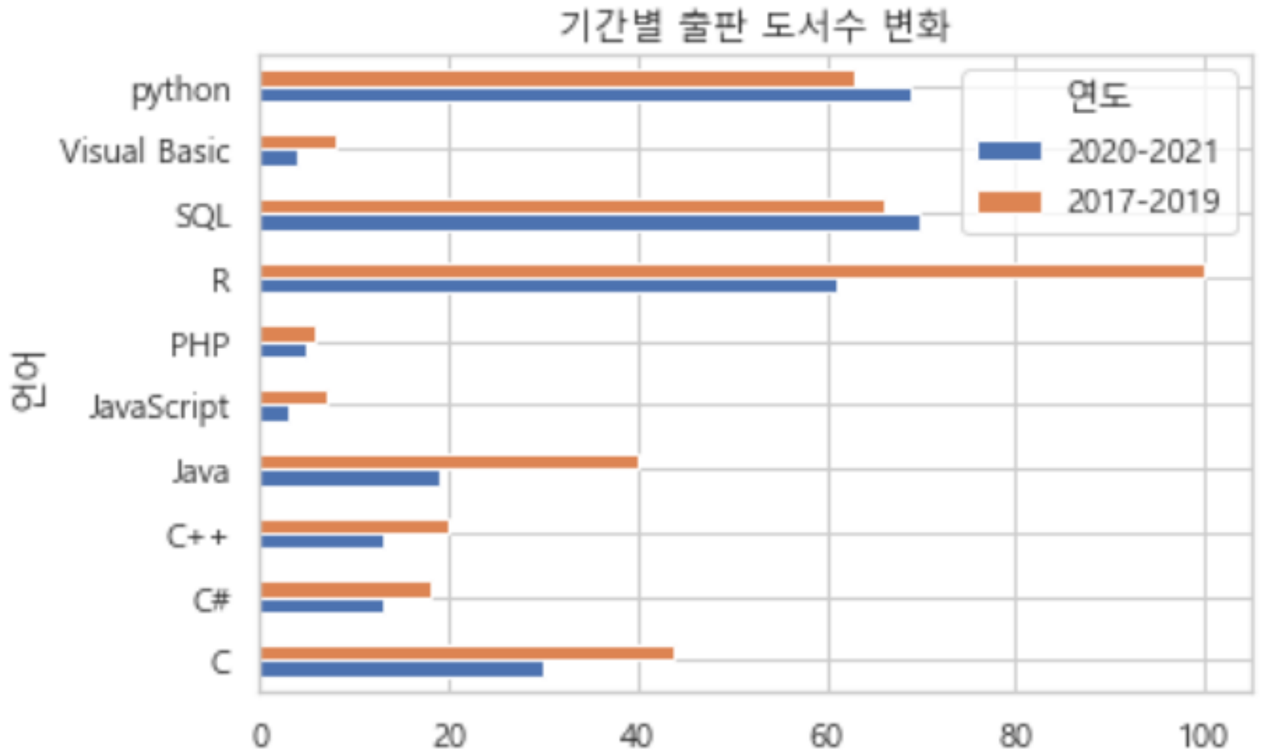
# 프로그래밍 언어별 출판도서 수 분석(2)

| 언어 | 출판물 수 |
|---|---|
| R | 161 |
| SQL | 136 |
| python | 132 |
| C | 74 |
| Java | 59 |
| C++ | 33 |
| C# | 31 |
| Visual Basic | 12 |
| PHP | 11 |
| JavaScript | 10 |

5개년도 언어별 출판도서 수

| 연도 | 2020-2021 | 2017-2019 |
|---|---|---|
| 언어 | | |
| C | 30.0 | 44.0 |
| C# | 13.0 | 18.0 |
| C++ | 13.0 | 20.0 |
| Java | 19.0 | 40.0 |
| JavaScript | 3.0 | 7.0 |
| PHP | 5.0 | 6.0 |
| R | 61.0 | 100.0 |
| SQL | 70.0 | 66.0 |
| Visual Basic | 4.0 | 8.0 |
| python | 69.0 | 63.0 |

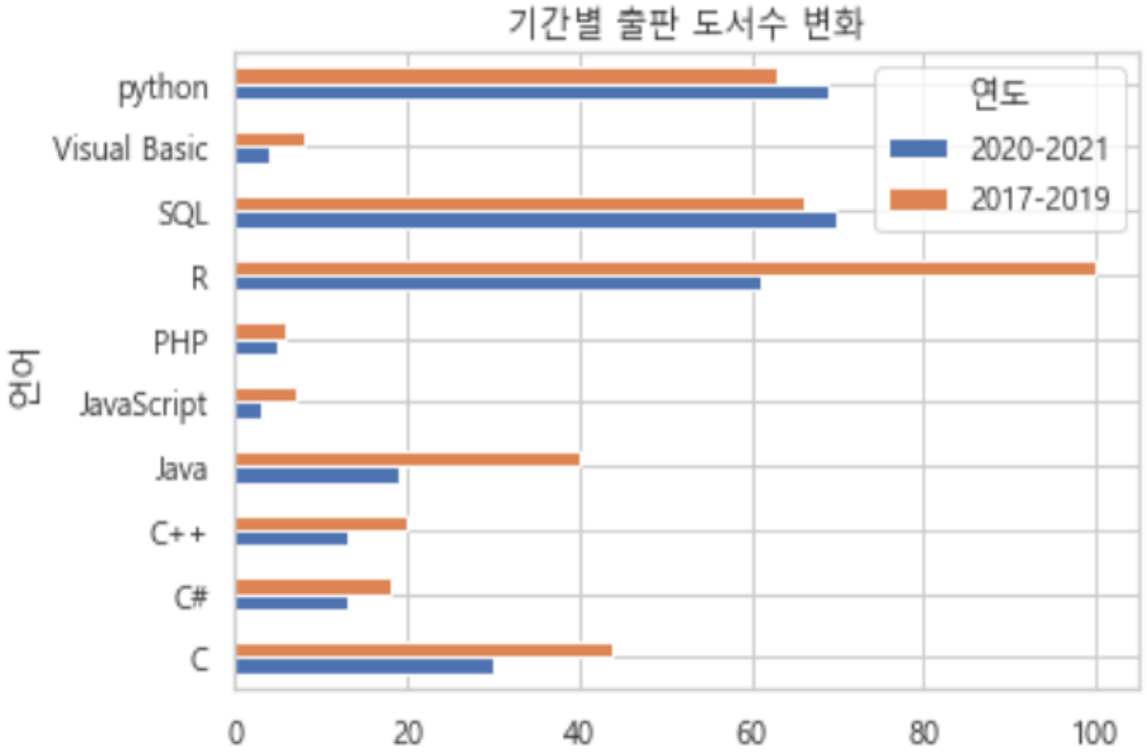
최근 2개년 vs 그 이전 3개년 출판도서 수 비교



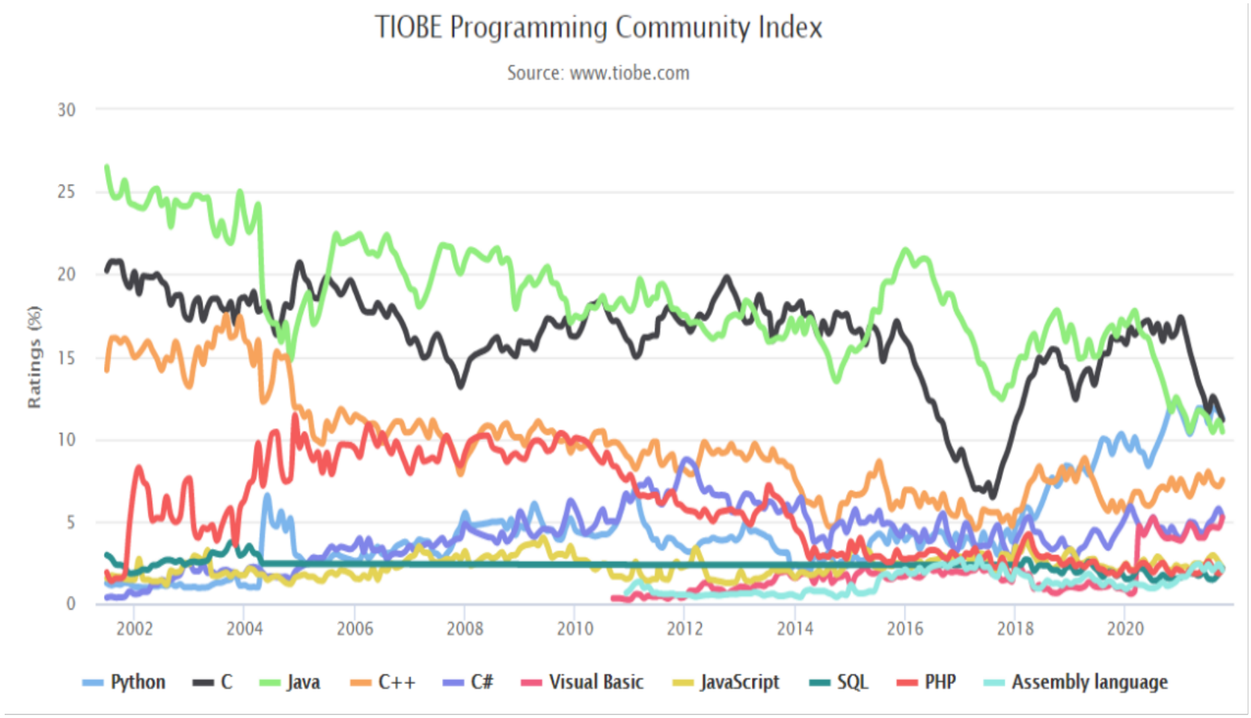

최근 2개년 vs 그 이전 3개년 출판도서 수 변화 바(Barh) 그래프 시각화

- 5개년을 한번에 볼때와 다르게, 최근 2개년 기준으로 보면 그 이전 기간에 비해, R 도서는 SQL, Python에 비해 현저히 감소된 것을 볼 수 있다.
- 기간별 출판 도서수가 그 이전에 비해 최근(2020-2021)에 증가한 언어는, Python과 SQL이 2개 뿐인 것으로 나타난다.
- PIOBE 인덱스 상으로 확인했던대로, Java의 니즈 감소를 기간별 국내 출판도서 수로 확인할 수 있다. (현저히 감소)
- C, C++, C#도 최근 2개년에 출판 도서가 많이 줄어든 것으로 보인다.

# 프로그래밍 언어별 출판도서 수 분석(3)




최근 2개년 vs 그 이전 3개년 출판도서 수 변화 바(Barh) 그래프 시각화




PIOBE 인덱스 수치 변화

- 최근 2년간 변화 폭이 큰 언어 : Python(+), Java(-), C(-), Visual Basic(+), PHP(+)

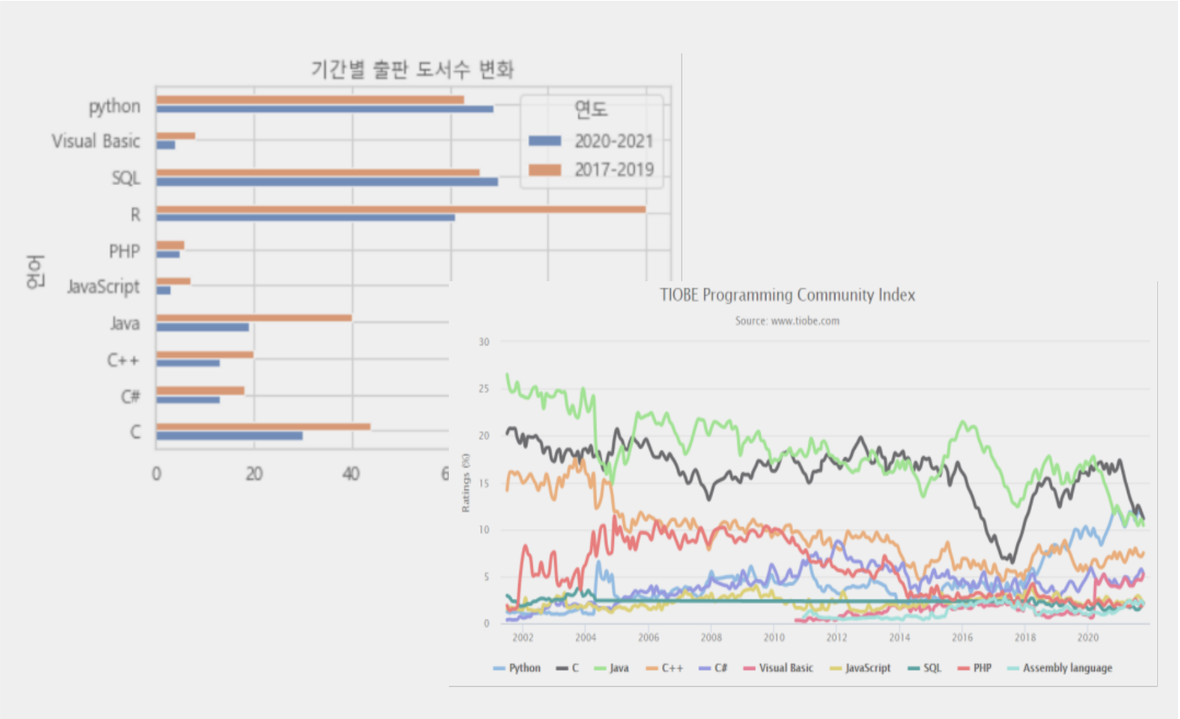
- 인덱스 그래프에는 없지만, 현재 R 언어는 14위에 랭크되어 있다. (R 도서 출판의 감소와의 연관성이 있어 보인다.)
- Python, Java의 니즈 변화도 국내 도서 출판에 영향을 준 듯하다.
- Visual Basic과 PHP는 2년간 니즈가 증가했는데, 국내 도서 출판 수는 오히려 감소한 것으로 나타났다.

# 프로그래밍 언어별 출판도서 수 분석 결론

**가설** **전세계 언어별 사용 니즈가 많을수록,**
**국내에서 출판되는 책의 수도 많을 것이다.**

**결론** **전세계 언어별 사용 니즈의 변화가,**
**대체적으로 국내 도서 출판에 영향을 준다.**

- 절대적으로 높은 니즈를 가지고 있으면서
- 변화 폭이 크게 나타나는 언어일수록




단순히 지금 가장 니즈가 많은 언어에 대한 국내 출판 도서가 많은 것이 아니라
1) Python, Java, C와 같이 절대적으로 높은 니즈를 가지고 있는 언어가
2) 기간 내에 니즈 변화가 크게 나타나는 경우에
국내 출판 여부에 영향을 끼치는 것으로 판단된다.

# 기타 분석 1 : 출판사(Top 10)별 출판도서 수 분석 및 시각화

| 출판사 | 출판물 수 |
|---|---|
| 21세기사 | 8 |
| BIGMV센터 | 1 |
| BOOKK(부크크) | 6 |
| CHOAlliance | 1 |
| DK로드북스 | 1 |
| EBS(한국교육방송공사) | 1 |
| GS인터비전 | 1 |
| HumanScience(휴먼싸이언스) | 1 |
| J&H | 1 |
| McGraw-HillEducation | 1 |
| STORYJOA(스토리조아) | 1 |
| e퍼플 | 2 |
| 가나출판사 | 1 |
| 가메 | 11 |
| 건기원 | 9 |
| 경문사 | 1 |

총 197개 출판사별 도서 수

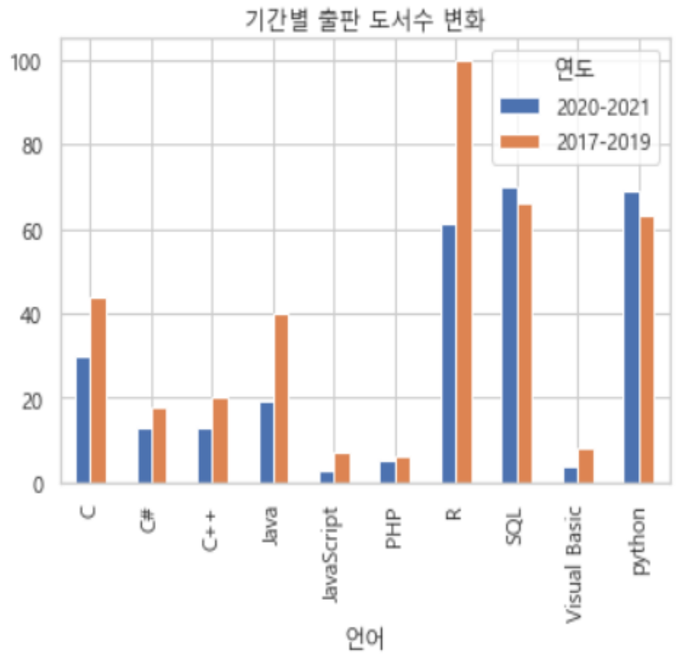
| 출판사 | 출판물 수 |
|---|---|
| 한빛아카데미 | 27 |
| 한빛미디어 | 18 |
| 시대고시기획 | 18 |
| 길벗 | 17 |
| 에이콘출판 | 17 |
| 퍼플 | 15 |
| 비제이퍼블릭 | 15 |
| 인피니티북스 | 15 |
| 글로벌 | 12 |
| 복두출판사 | 12 |

5개년도 출판도서 수 기준
Top10 출판사 리스트

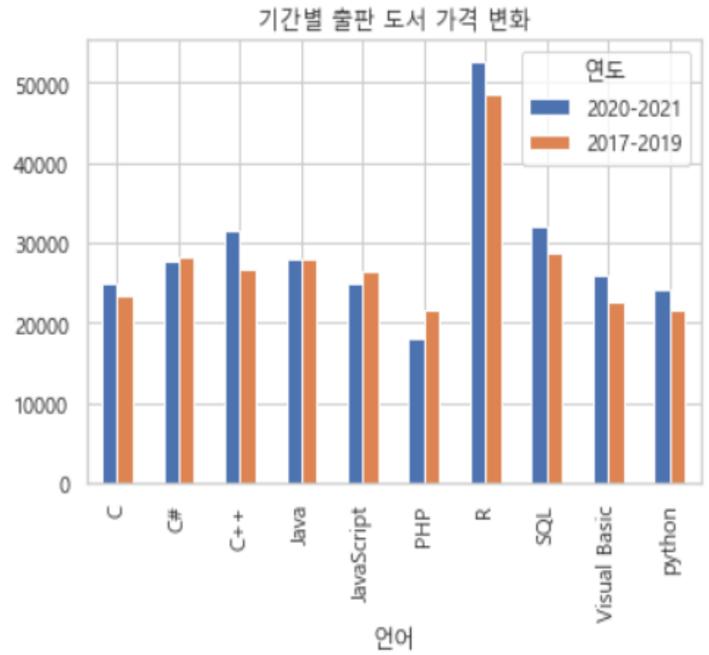

5개년도 출판도서 수 기준 Top10 출판사 데이터 시각화

- 출판사는 총 197개사로 나타났고, 5개년 데이터인 총 658권 중 166권을 Top10 출판사에서 출판했다.
- 한빛아카데미 - 시대고시기획, 한빛미디어 - 에이콘출판, 길벗 순으로 출판사 랭크가 높다.
- 출판사 이름이 아카데미, 시대고시, 길벗 등 자격증이나 수험 관련 IT서적을 출판한 것이 아닐까 추측된다.

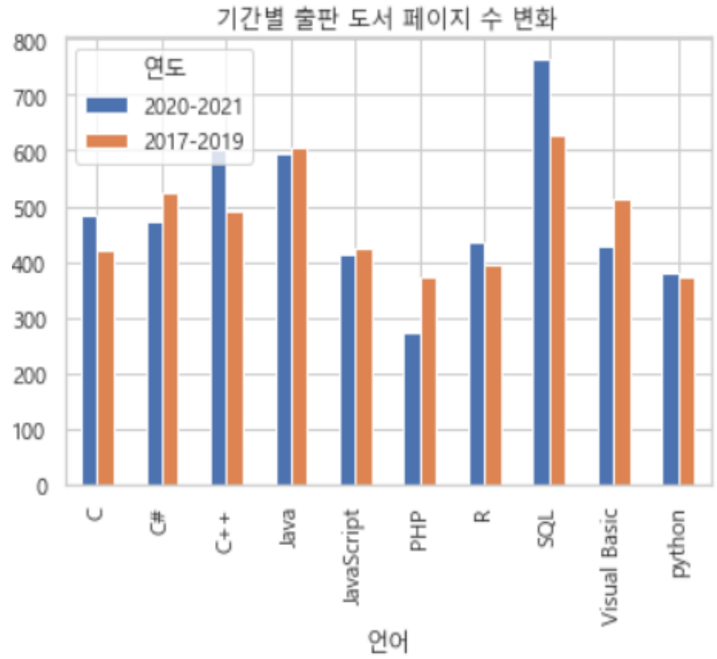
# 기타 분석 2 : 최근(2020-2021) vs 그 이전(2017-2019), 언어별 도서 관련 데이터의 변화




기간별 출판 도서수 변화




기간별 출판 도서 가격 변화




기간별 출판 도서 페이지 수 변화

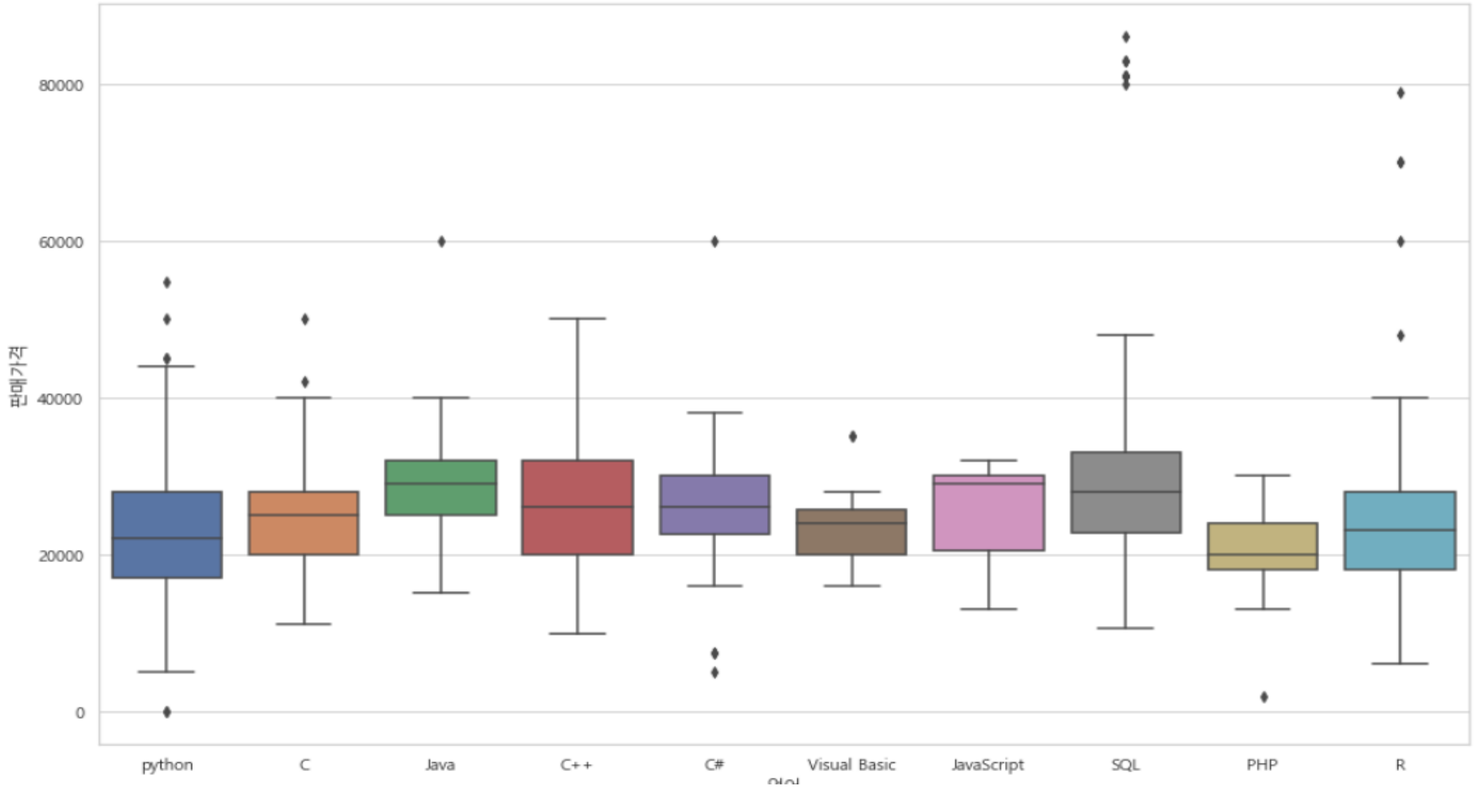
*그래프는 최근(2020-2021) - 그 이전(2017-2019) 순서로 표시했다.

- 도서 수의 변화는 R(-), C(-), Java(-)가 현저히 크게 나타났다. Python과 SQL의 출판 도서 수는 증가했다.

- 도서 가격은 거의 변화가 없다. 5천원 내외의 차이 정도인 듯하다.

- 페이지 수의 경우는 SQL과 C++가 가장 많이 증가했다. JavaScript와 Python는 거의 변화가 없다.

# 기타 분석 3 : 도서 가격과 페이지 수 간의 상관관계



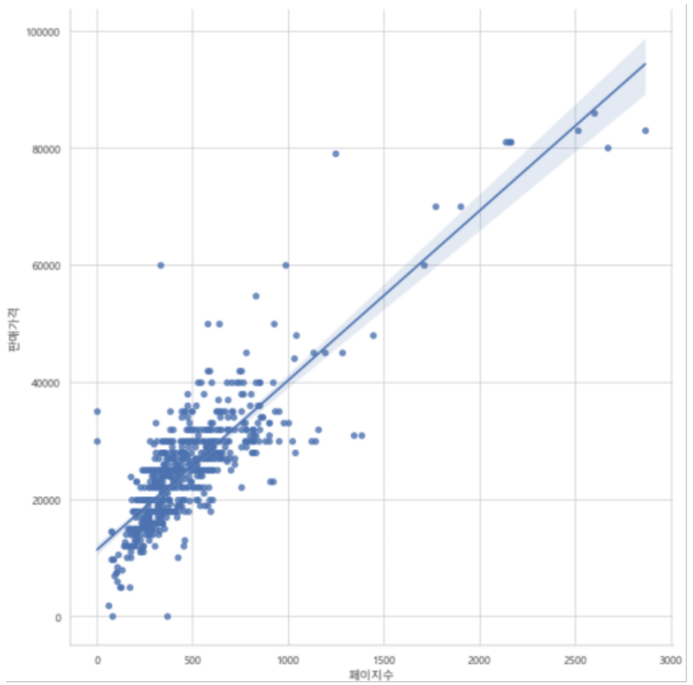

언어별 도서 가격의 분포(boxplot)

- 도서 가격의 경우, 언어별로 가격 차이가 거의 존재하지 않는다.
- SQL과 R의 경우에는 아웃라이어(이상점)가 많이 존재한다. 추후 데이터 정제 필요성이 있다.
- Visul Basic, PHP, Java의 도서 가격은 거의 차이가 없을 정도로 밀집되어 있다.

# 기타 분석 3 : 도서 가격과 페이지 수 간의 상관관계

| | 판매가격 | 연도 | 페이지수 |
|---|---|---|---|
| 판매가격 | 1.000000 | 0.034776 | 0.237177 |
| 연도 | 0.034776 | 1.000000 | 0.103240 |
| 페이지수 | 0.237177 | 0.103240 | 1.000000 |

도서 가격, 연도, 페이지수 간의 상관관계계수 (corr)




페이지수와 판매 가격 사이의 관계(lmplot)

- 위 자료를 통해서 도서 가격이 페이지수와 상대적으로 연관성이 있다고 분석 결론을 내릴 수 있다.
- corr 결과를 보면, 가격이 페이지수와의 연관성이 연도보다 높다고 볼 수 있다.
- lmplot의 경우에도, 회귀선 근처에 산점도가 모여있는 것을 보았을 때, 책이 두꺼울수록 가격이 높아지는 경향을 보인다고 볼 수 있다.

# 데이터 수집 및 전처리 (1) - 네이버 API 활용

## > 함수 생성하여 10개 언어 반복문(for) 적용해서 6,693개 데이터 수집

- 언어, 책제목, 출판사, 출판일, 판매가격, ISBN, 상세링크 Column화

```
from tqdm import tqdm
for lang in tqdm(language_top[1:]):
    for n in range(1, 1000, 100):
        url = gen_search_url("book", lang, n, 100)
        json_result = get_result_onepage(url)
        pd_result = pd.concat([pd_result, get_fields(json_result)])

pd_result
```

| | 언어 | 책제목 | 출판사 | 출판일 | 판매가격 | ISBN | |
|---|---|---|---|---|---|---|---|
| 0 | python | Python 프로그래밍(스마트로봇 EV3를 활용한) (스마트로봇 EV3를 활용한) | 상학당 | 20200915 | 30000 | 8965872022 9788965872023 | http://book.naver.com/book |
| 1 | python | Python 프로그래밍 | 퍼플 | 20170330 | 11500 | 0000281298 1400000281291 | http://book.naver.com/book |
| 2 | python | 파이썬만 잡아도 기초를 탄탄히 세우는 Python 프로그래밍 | 카오스북 | 20180630 | 25000 | 1187486183 9791187486183 | http://book.naver.com/book |
| 3 | python | Python 프로그래밍의 이해 | 교보문고 | 20170117 | 18000 | 1159090289 9791159090288 | http://book.naver.com/book |
| 4 | python | The Python - 파이썬 프로그래밍 | BOOKK(부크크) | 20210909 | 32000 | 1137255641 9791137255647 | http://book.naver.com/book |

사용자 정의 함수와 for문 적용해서 나온 수집 데이터 DataFrame

**gen_search_url()**
API 갯수 지정해서 해당 URL을 생성해줌
↓
**get_result_onepage()**
생성된 URL의 페이지 정보를 호출
↓
**get_fields()**
해당하는 데이터를 DataFrame으로 만들어줌

*** delete_tag()**
제목에 포함되는 ,  태그를 지워줌

# 데이터 수집 및 전처리 (2) - ISBN 기준 중복 제거 및 5개년도 데이터 처리

**> 도서 고유 ID인 ISBN(국제표준도서번호)를 기준으로 중복 제거**

**> 분석 대상인 5개년도 데이터 882개만 남김 - 연도 컬럼 별도 추가하여 진행**

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 822 entries, 0 to 4
Data columns (total 8 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   언어      822 non-null    object
 1   책제목     822 non-null    object
 2   출판사     822 non-null    object
 3   출판일     822 non-null    object
 4   판매가격    822 non-null    object
 5   ISBN    822 non-null    object
 6   상세링크    822 non-null    object
 7   연도      822 non-null    int64
dtypes: int64(1), object(7)
memory usage: 57.8+ KB
```

전처리 후 DataFrame의 info() 결과

**ISBN 중복의 경우는**

**도서가 2개 이상의 언어를 담는다는 말**

→ 한쪽만 제거하는 방식보다는

→ 두쪽 다 제거해서 분석 타당성을 높였음

→ DATAFRAME.drop_duplicates(["ISBN"], keep = False)

**API에서 받아온 연도 정보는 str 타입**

→ pd.to_numeric() 활용해서 int로 변환

# 데이터 수집 및 전처리 (3) - BeautifulSoup 활용한 웹 크롤링

## 페이지 수 정보를 각 상세페이지 내에서 웹 크롤링 수행

- 페이지 수 Column 추가

```
page_list = []
from tqdm import tqdm
import warnings
warnings.simplefilter(action="ignore")

# for 문, except

for link_book in tqdm(pd_result_recently["상세링크"]):

    url = link_book
    req = Request(url, headers={"User-Agent": "Chrome"})
    html = urlopen(req).read()
    soup = BeautifulSoup(html, "html.parser")

    num = soup.select("#container > div.spot > div.book_info > div.book_info_inner>div")[2].text
    final = num.split()[1].split("|")[0]
    page_list.append(final)


pd_result_recently["페이지수"] = page_list

pd_result_recently
```

```
# 예외처리를 하게 된 이유
pd_result_recently.iloc[229]  # 해당 링크 들어가니까 삭제된 서지

언어                                                    C
책제목                                          혼자 공부하는 C 언어
출판사                                             한빛미디어
출판일                                           20190610
판매가격                                            24000
ISBN                                  1162241861 9791162241868
상세링크    http://book.naver.com/bookdb/book_detail.php?b...
연도                                              2019
Name: 21, dtype: object
```


◀ 기존 코딩의 오류 화면 (IndexError)

오류가 난 행의 링크(URL)를 따라가보니

해당 페이지가 판매 종료로 사라진 상태

→ 이런 경우에 예외처리로 "-"를 넣어주고

→ 나중에 해당 열을 모두 없애주기로 함

→ try ~ except

# 데이터 수집 및 전처리 (3) - BeautifulSoup 활용한 웹 크롤링

## 페이지 수 정보를 각 상세페이지 내에서 웹 크롤링 수행

- 페이지 수 Column 추가

```
page_list = []
from tqdm import tqdm
import warnings
warnings.simplefilter(action="ignore")

# for 문, except

for link_book in tqdm(pd_result_recently["상세링크"]):
    try:
        url = link_book
        req = Request(url, headers={"User-Agent": "Chrome"})
        html = urlopen(req).read()
        soup = BeautifulSoup(html, "html.parser")

        num = soup.select("#container > div.spot > div.book_info > div.book_info_inner>div")[2].text
        final = num.split()[1].split("|")[0]
        page_list.append(final)

    except IndexError :
        page_list.append("-")
        continue


pd_result_recently["페이지수"] = page_list

pd_result_recently
```

```
0%|                                                                                | 822/822 [05:50<00:00,  2.34it/s]
```

| | 언어 | 책제목 | 출판사 | 출판일 | 판매가격 | ISBN | 상세링크 | 연도 | 페이지수 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | python | Python 프로그래밍 (스마트로봇 EV3를 활용한) (스마트로봇 EV3를 활용한) | 상학당 | 20200915 | 30000 | 8965872022 9788965872023 | http://book.naver.com/bookdb/book_detail.php?b... | 2020 | 363 |
| 1 | python | Python 프로그래밍 | 퍼플 | 20170330 | 11500 | 0000281298 1400000281291 | http://book.naver.com/bookdb/book_detail.php?b... | 2017 | 236 |

◀ 개선 코딩의 결과 화면

except IndexError :

해당 예외처리를 통해 웹 크롤링 진행해서,

822개 데이터의 페이지 수를 받아왔음

→ 이후, 예외된 행 삭제 진행

→ 페이지 수가 아닌 데이터도 행 삭제 진행

→ 총 658개 데이터를 최종 수집 완료

# 추후 연구의 방향

## 데이터 퀄리티 향상

- 수집한 도서 데이터의 수를 늘려야 함
  → 검색 키워드 변경
- 데이터 전처리 세분화
  → 중복 건수 제거에 대한 고민
- 가격 등 이상점에 대한 재확인

## 분석 대상 기간 확대 및 세분화

- 5개년도만을 대상으로 분석 진행한 부분
  → 최근 경향만으로 결론의 타당성 확보 힘듬
- 2년 기준, 3년 기준, 5년 기준 등
  → 5년이 대상기간이어서 2/3으로만 분화
  → 기간 확대 후
  세분화하거나, 단순화 분석 필요

## 실제 판매량과의 연관성

- 출판 도서만을 분석 대상 - 공급 측면
  → 수요 측면도 확인할 필요성 有
- 가설이 니즈와 수요를 바탕으로 했기 때문에
- 실제 판매되는 수량과의 연관성이 더 중요함
  → 데이터 확보가 어렵지만,
  해당 데이터가 있으면 현 분석에 첨부 必

끝.