



Amazon

Top 50 Bestselling Books

2009 - 2019

Kaggle에서 제공하는 아마존 베스트 셀러 데이터를 이용한 데이터 분석

(<https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019>)

[D.N.A 빈SUN조]

도준희 이지수 한지민 홍연정

데이터 변수

Dataset on Amazon's Top 50 bestselling books from 2009 to 2019. Contains 550 books, data has been categorized into fiction and non-fiction using Goodreads

550 rows X 7 columns

	Name	Author	User Rating	Reviews	Price	Year	Genre
0	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
1	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
2	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
3	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
4	5,000 Awesome Facts (About Everything!) (Natio...	National Geographic Kids	4.8	7665	12	2019	Non Fiction
...
545	Wrecking Ball (Diary of a Wimpy Kid Book 14)	Jeff Kinney	4.9	9413	8	2019	Fiction
546	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2016	Non Fiction
547	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2017	Non Fiction
548	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2018	Non Fiction
549	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2019	Non Fiction

DATA TOPIC

2009년부터 2019년까지
아마존의 TOP 50 베스트 셀러 도서

DATA CATEGORY

550개의 도서에 대한 데이터
데이터는 기본적으로 non-fiction과 fiction으로 구분

- Name : 책의 이름
- Author : 책의 저자
- User Rating : 아마존 사용자 평가
- Reviews : 아마존에 작성된 리뷰 수
- Price : 책 가격
- Year : 베스트셀러에 선정된 연도
- Genre : 장르

Data

	Name	Author	User Rating	Reviews	Price	Year	Genre
0	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
1	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
2	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
3	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
4	5,000 Awesome Facts (About Everything!) (Natio...	National Geographic Kids	4.8	7665	12	2019	Non Fiction

```
print(df.shape)
```

(550, 7)

```
df.isnull().sum()
```

```
Name      0
Author     0
User Rating 0
Reviews    0
Price      0
Year       0
Genre      0
dtype: int64
```

- **Data Shape:** Data는 550행 7열의 구조를 가지고 있다.
- **결측 값:** 결측 값은 없는 것으로 보인다.

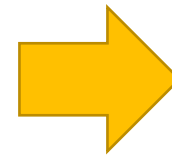
Data

```
# incorrect spelling
df[df.Author == 'J. K. Rowling']
```

	Name	Author	User Rating	Reviews	Price	Year	Genre
155	Harry Potter and the Goblet of Fire: The Illus...	J. K. Rowling	4.9	7758	18	2019	Fiction
159	Harry Potter Paperback Box Set (Books 1-7)	J. K. Rowling	4.8	13471	52	2016	Fiction

```
# correct spelling
df[df.Author == 'J.K. Rowling']
```

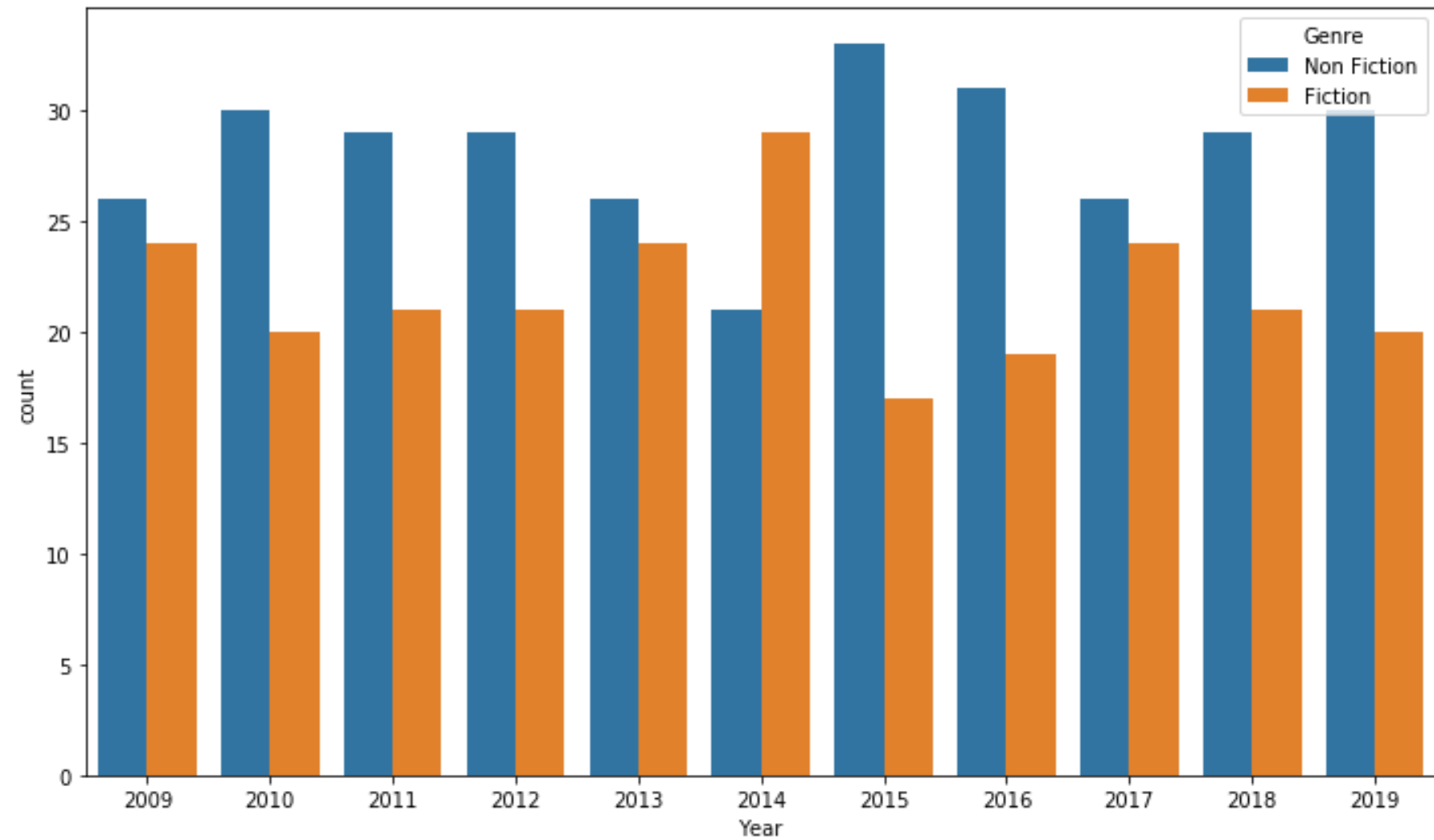
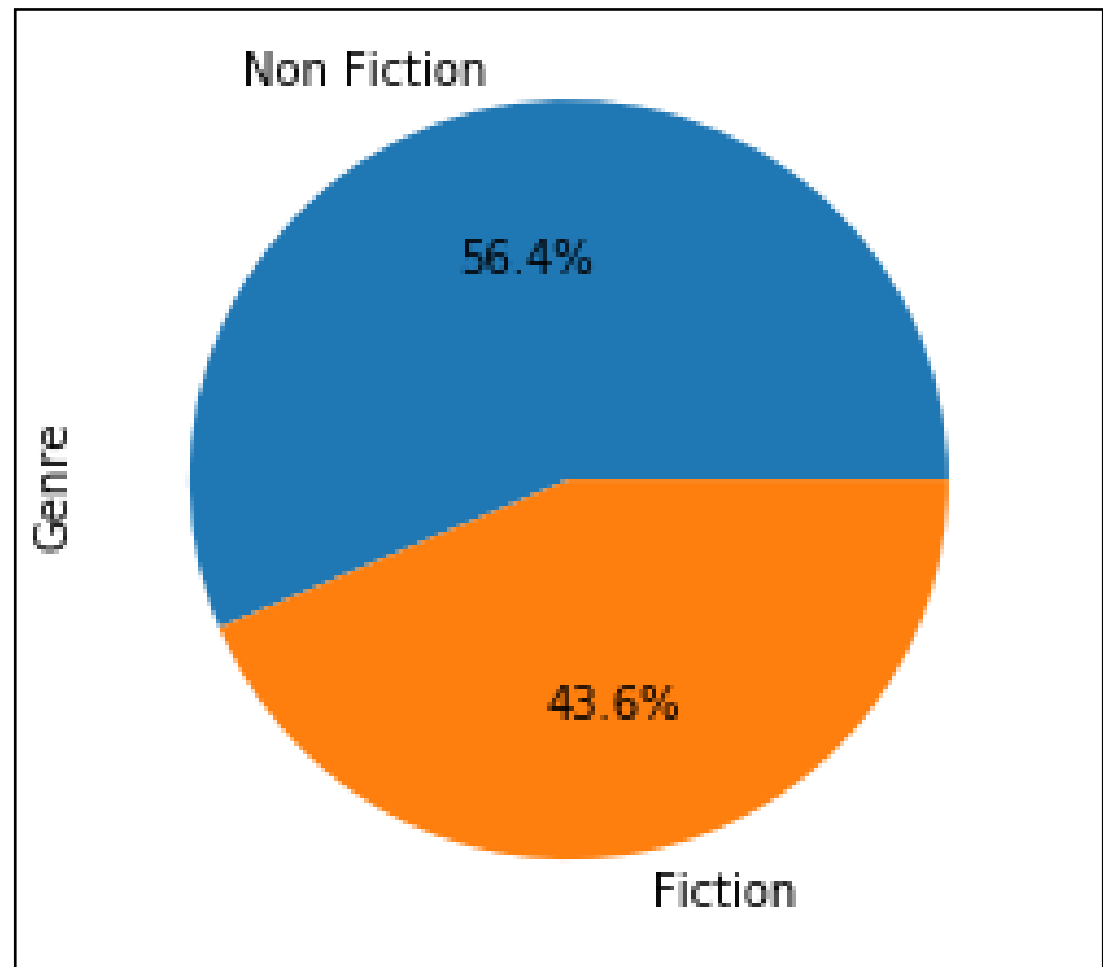
	Name	Author	User Rating	Reviews	Price	Year	Genre
102	Fantastic Beasts and Where to Find Them: The O...	J.K. Rowling	4.7	4370	15	2016	Fiction
153	Harry Potter and the Chamber of Secrets: The I...	J.K. Rowling	4.9	19622	30	2016	Fiction
154	Harry Potter and the Cursed Child, Parts 1 & 2...	J.K. Rowling	4.0	23973	12	2016	Fiction
156	Harry Potter and the Prisoner of Azkaban: The ...	J.K. Rowling	4.9	3146	30	2017	Fiction
157	Harry Potter and the Sorcerer's Stone: The Ill...	J.K. Rowling	4.9	10052	22	2016	Fiction
353	The Casual Vacancy	J.K. Rowling	3.3	9372	12	2012	Fiction



	Name	Author	User Rating	Reviews	Price	Year	Genre
102	Fantastic Beasts and Where to Find Them: The O...	J.K. Rowling	4.7	4370	15	2016	Fiction
153	Harry Potter and the Chamber of Secrets: The I...	J.K. Rowling	4.9	19622	30	2016	Fiction
154	Harry Potter and the Cursed Child, Parts 1 & 2...	J.K. Rowling	4.0	23973	12	2016	Fiction
155	Harry Potter and the Goblet of Fire: The Illus...	J.K. Rowling	4.9	7758	18	2019	Fiction
156	Harry Potter and the Prisoner of Azkaban: The ...	J.K. Rowling	4.9	3146	30	2017	Fiction
157	Harry Potter and the Sorcerer's Stone: The Ill...	J.K. Rowling	4.9	10052	22	2016	Fiction
159	Harry Potter Paperback Box Set (Books 1-7)	J.K. Rowling	4.8	13471	52	2016	Fiction
353	The Casual Vacancy	J.K. Rowling	3.3	9372	12	2012	Fiction

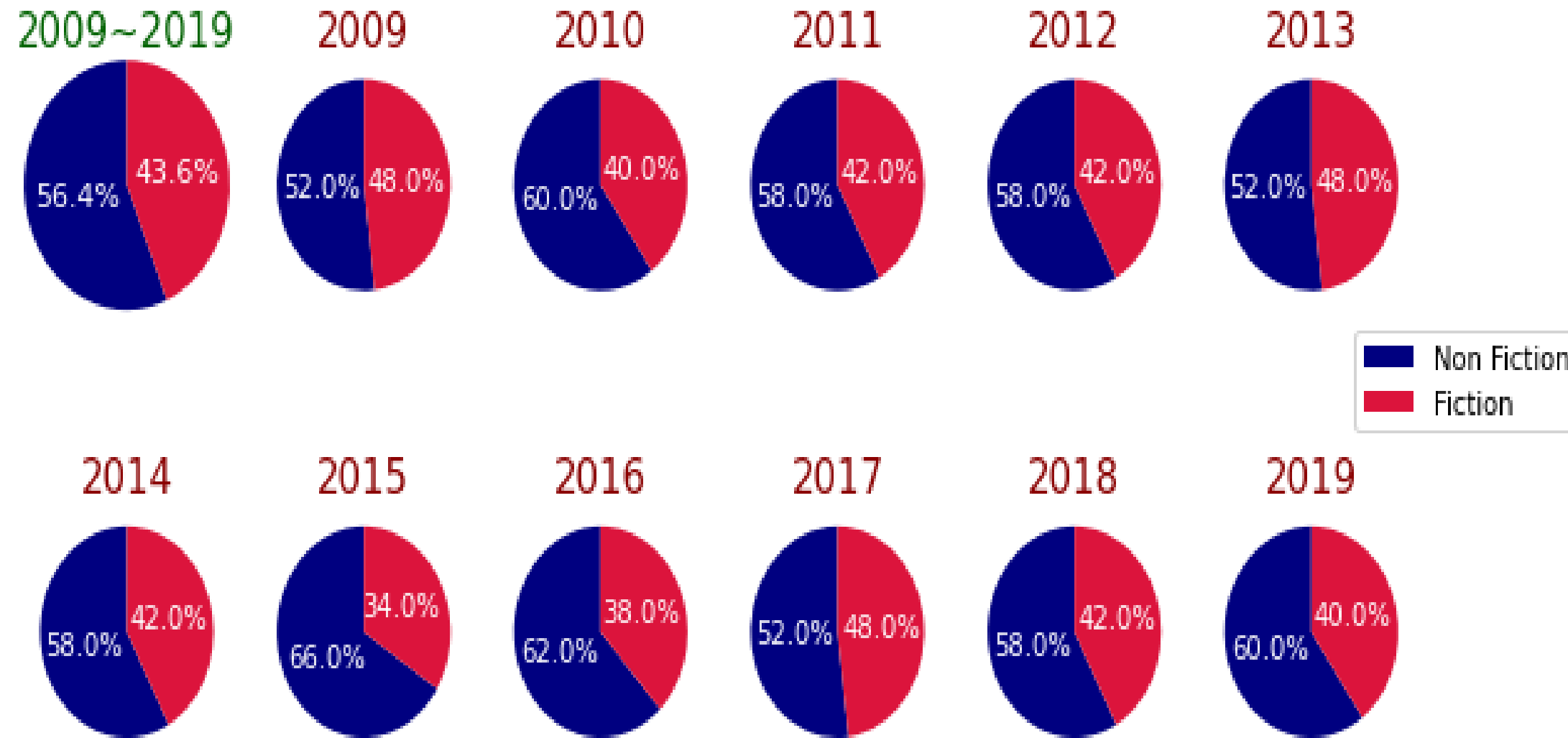
- 잘못 스펠링 된 작가 이름: 작가 J.K. Rowling의 이름은 J와 K 사이에 공백이 추가돼 두 책에 잘못 표기되었다.
- 다음과 같이 잘못 표기된 데이터 수정

연도별 장르별 특징



- 대부분 년도에 비소설을 더 선호
- 2014년도만 소설을 더 많이 선호
- 2015, 2016년에는 소설과 비소설의 차이가 상당히 큼

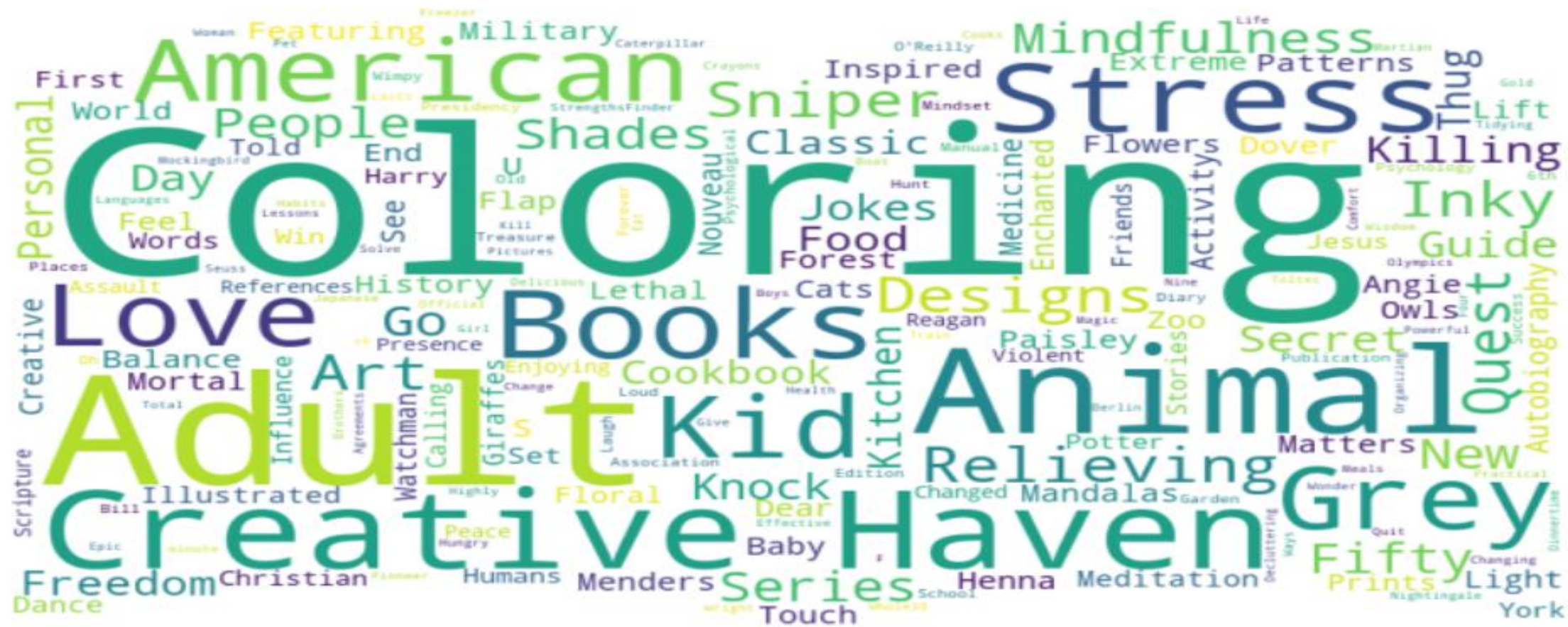
EDA 장르 별



- **모든 연도에서 장르의 비율:** 2009년부터 2019년까지 매년 비소설은 소설에 비해 인기가 많았다. 책들 중 54.4%는 비소설이었고 나머지 45.6%는 소설이었다.
- **연도별 장르의 비율:** 2015년에는 비소설의 최고 비율인 66%를 보였고, 소설의 최저 비율인 34%를 보였다. 소설은 2009년(48%)과 2017년(48%)에 가장 높은 비율을 차지했다.

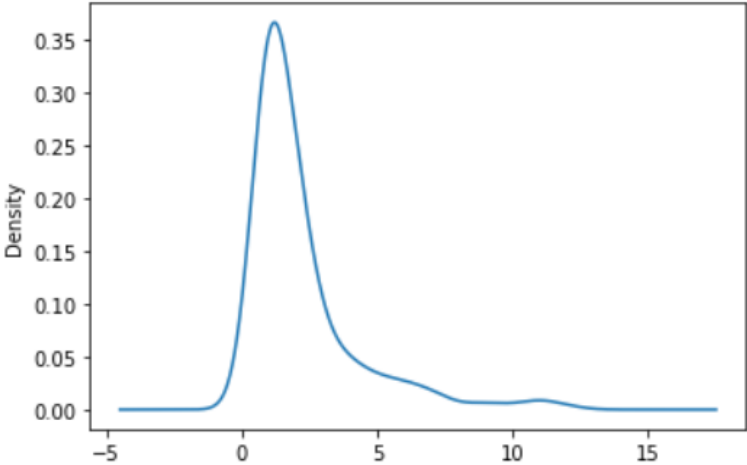


EDA 장르 별

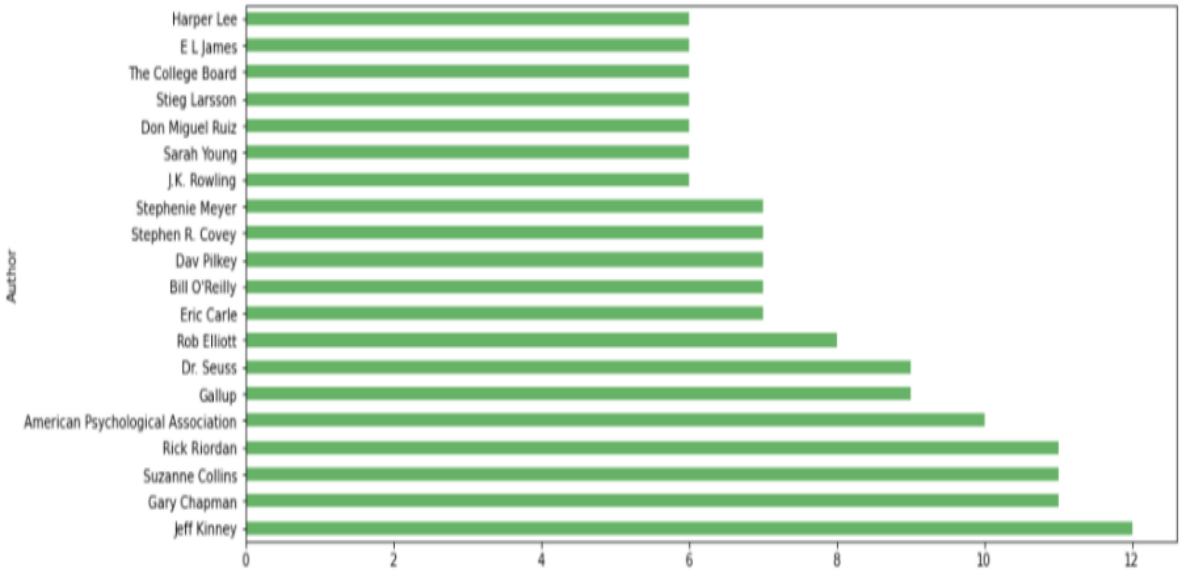


- 2015년 word cloud: 다른 해보다 비소설의 비율이 높았던 2015년의 베스트 셀러 책들의 제목으로 word cloud를 했을 때, 비소설 책인 coloring book이 인기가 많았다.

연도별 작가별 특징



작가별 베스트셀러 선정 횟수 밀도함수



작가별 베스트셀러 개수 막대그래프

- > 대부분의 작가는 베스트 셀러 작가로 5회 미만 선정
- > 일부 작가는 10회 이상~최대 12회까지 베스트 셀러 작가로 선정

```
[58] df[['Year', 'Author', 'Name']].loc[df['Author']=='Jeff Kinney']
```

	Year	Author	Name
42	2011	Jeff Kinney	Cabin Fever (Diary of a Wimpy Kid, Book 6)
71	2013	Jeff Kinney	Diary of a Wimpy Kid: Hard Luck, Book 8
72	2009	Jeff Kinney	Diary of a Wimpy Kid: The Last Straw (Book 3)
73	2014	Jeff Kinney	Diary of a Wimpy Kid: The Long Haul
80	2009	Jeff Kinney	Dog Days (Diary of a Wimpy Kid, Book 4) (Volum...
88	2016	Jeff Kinney	Double Down (Diary of a Wimpy Kid #11)
253	2015	Jeff Kinney	Old School (Diary of a Wimpy Kid #10)
381	2017	Jeff Kinney	The Getaway
435	2018	Jeff Kinney	The Meltdown (Diary of a Wimpy Kid Book 13)
468	2012	Jeff Kinney	The Third Wheel (Diary of a Wimpy Kid, Book 7)
474	2010	Jeff Kinney	The Ugly Truth (Diary of a Wimpy Kid, Book 5)
545	2019	Jeff Kinney	Wrecking Ball (Diary of a Wimpy Kid Book 14)

```
[60] df[['Year', 'Author', 'Name']].loc[df['Author']=='Suzanne Collins']
```

	Year	Author	Name
46	2010	Suzanne Collins	Catching Fire (The Hunger Games)
47	2011	Suzanne Collins	Catching Fire (The Hunger Games)
48	2012	Suzanne Collins	Catching Fire (The Hunger Games)
236	2010	Suzanne Collins	Mockingjay (The Hunger Games)
237	2011	Suzanne Collins	Mockingjay (The Hunger Games)
238	2012	Suzanne Collins	Mockingjay (The Hunger Games)
407	2010	Suzanne Collins	The Hunger Games
408	2011	Suzanne Collins	The Hunger Games (Book 1)
409	2012	Suzanne Collins	The Hunger Games (Book 1)
410	2011	Suzanne Collins	The Hunger Games Trilogy Boxed Set (1)
411	2012	Suzanne Collins	The Hunger Games Trilogy Boxed Set (1)

```
df[['Year', 'Author', 'Name']].loc[df['Author']=='Gary Chapman']
```

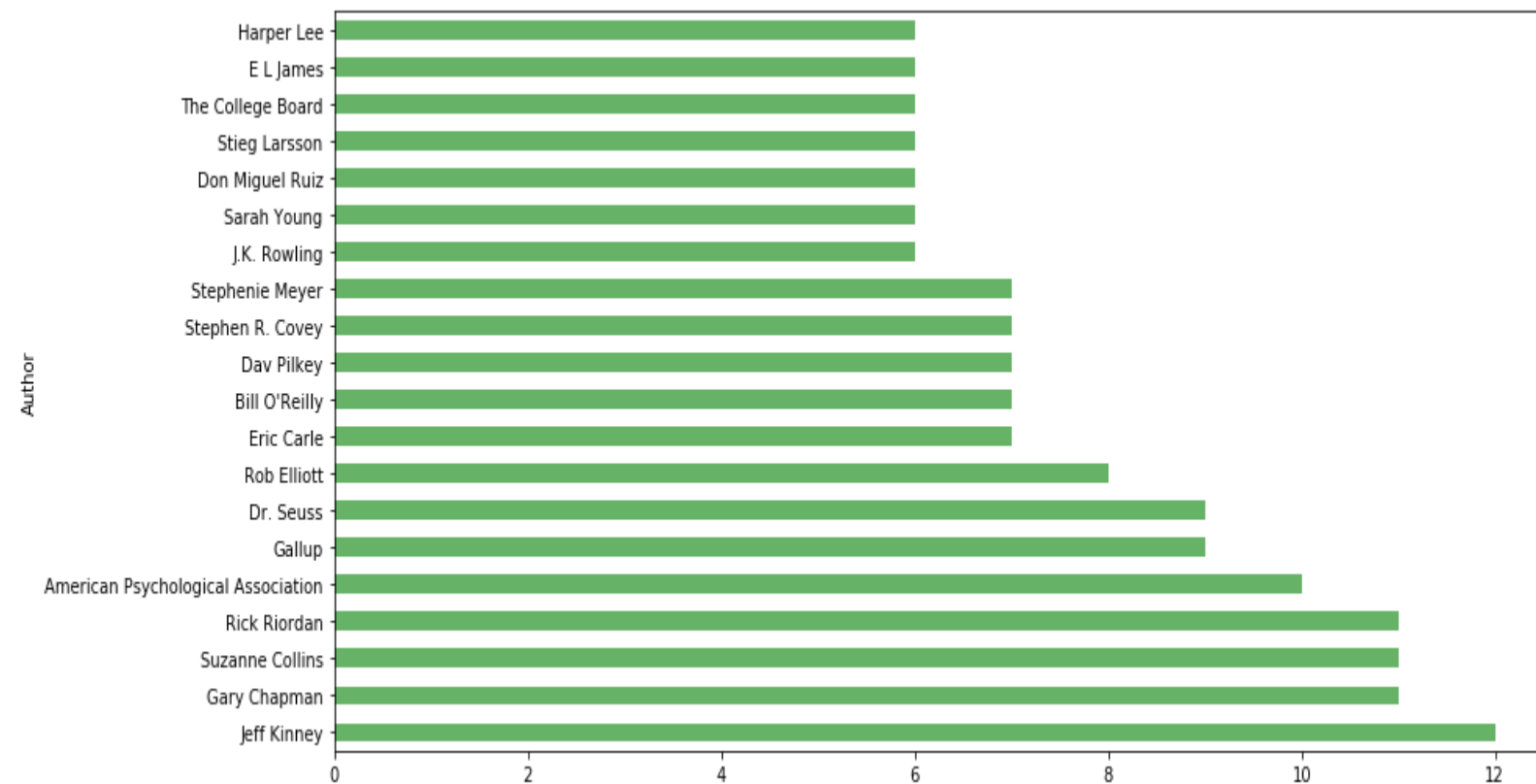
	Year	Author	Name
320	2010	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
321	2011	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
322	2012	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
323	2013	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
324	2014	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
325	2015	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
326	2016	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
327	2017	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
328	2018	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
329	2019	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
374	2009	Gary Chapman	The Five Love Languages: How to Express Heartf...

```
df[['Year', 'Author', 'Name']].loc[df['Author']=='Rick Riordan']
```

	Year	Author	Name
264	2010	Rick Riordan	Percy Jackson and the Olympians Paperback Boxe...
343	2014	Rick Riordan	The Blood of Olympus (The Heroes of Olympus (5))
406	2013	Rick Riordan	The House of Hades (Heroes of Olympus, Book 4)
418	2009	Rick Riordan	The Last Olympian (Percy Jackson and the Olymp...
419	2010	Rick Riordan	The Last Olympian (Percy Jackson and the Olymp...
428	2010	Rick Riordan	The Lost Hero (Heroes of Olympus, Book 1)
432	2012	Rick Riordan	The Mark of Athena (Heroes of Olympus, Book 3)
456	2010	Rick Riordan	The Red Pyramid (The Kane Chronicles, Book 1)
458	2012	Rick Riordan	The Serpent's Shadow (The Kane Chronicles, Boo...
463	2011	Rick Riordan	The Son of Neptune (Heroes of Olympus, Book 2)
469	2011	Rick Riordan	The Throne of Fire (The Kane Chronicles, Book 2)

상위 4명의 작가들의 작품 확인
→ 시리즈물 작품이 대부분

EDA 작가 별



```
df.loc[df['Author']=='DK']
```

	Name	Author	User Rating	Reviews	Price	Year	Genre
28	Baby Touch and Feel: Animals	DK	4.6	5360	5	2015	Non Fiction
514	Ultimate Sticker Book: Frozen: More Than 60 Re...	DK	4.5	2586	5	2014	Fiction

```
df.loc[df['Author']=='Scholastic']
```

	Name	Author	User Rating	Reviews	Price	Year	Genre
158	Harry Potter Coloring Book	Scholastic	4.7	3564	9	2015	Non Fiction
268	Pokémon Deluxe Essential Handbook: The Need-to...	Scholastic	4.7	3503	9	2016	Fiction

- 작가별 베스트셀러 선정 횟수: 작가 Jeff Kinney는 2009년부터 2019년까지 총 12차례로 가장 많이 베스트셀러 작가로 선정되었다.
- DK와 Scholastic 두 작가만 장르별로 책을 가지고 있다.

EDA 작가 별

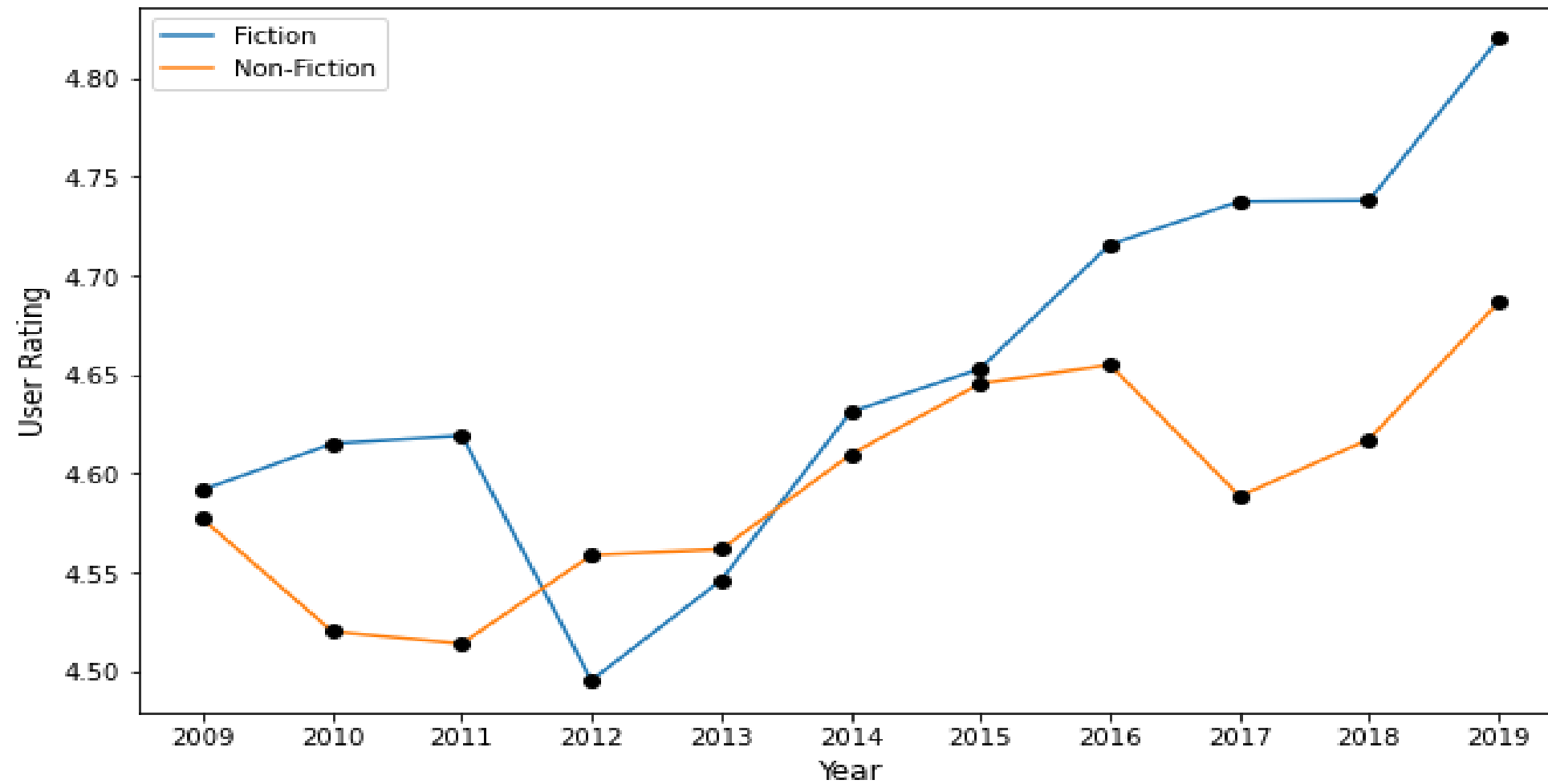
	Year	Author	Name
42	2011	Jeff Kinney	Cabin Fever (Diary of a Wimpy Kid, Book 6)
71	2013	Jeff Kinney	Diary of a Wimpy Kid: Hard Luck, Book 8
72	2009	Jeff Kinney	Diary of a Wimpy Kid: The Last Straw (Book 3)
73	2014	Jeff Kinney	Diary of a Wimpy Kid: The Long Haul
80	2009	Jeff Kinney	Dog Days (Diary of a Wimpy Kid, Book 4) (Volum...
88	2016	Jeff Kinney	Double Down (Diary of a Wimpy Kid #11)
253	2015	Jeff Kinney	Old School (Diary of a Wimpy Kid #10)
381	2017	Jeff Kinney	The Getaway
435	2018	Jeff Kinney	The Meltdown (Diary of a Wimpy Kid Book 13)
468	2012	Jeff Kinney	The Third Wheel (Diary of a Wimpy Kid, Book 7)
474	2010	Jeff Kinney	The Ugly Truth (Diary of a Wimpy Kid, Book 5)
545	2019	Jeff Kinney	Wrecking Ball (Diary of a Wimpy Kid Book 14)

	Year	Author	Name
320	2010	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
321	2011	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
322	2012	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
323	2013	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
324	2014	Gary Chapman	The 5 Love Languages: The Secret to Love That ...
325	2015	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
326	2016	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
327	2017	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
328	2018	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
329	2019	Gary Chapman	The 5 Love Languages: The Secret to Love that ...
374	2009	Gary Chapman	The Five Love Languages: How to Express Heartf...

	Year	Author	Name
46	2010	Suzanne Collins	Catching Fire (The Hunger Games)
47	2011	Suzanne Collins	Catching Fire (The Hunger Games)
48	2012	Suzanne Collins	Catching Fire (The Hunger Games)
236	2010	Suzanne Collins	Mockingjay (The Hunger Games)
237	2011	Suzanne Collins	Mockingjay (The Hunger Games)
238	2012	Suzanne Collins	Mockingjay (The Hunger Games)
407	2010	Suzanne Collins	The Hunger Games
408	2011	Suzanne Collins	The Hunger Games (Book 1)
409	2012	Suzanne Collins	The Hunger Games (Book 1)
410	2011	Suzanne Collins	The Hunger Games Trilogy Boxed Set (1)
411	2012	Suzanne Collins	The Hunger Games Trilogy Boxed Set (1)

➤ 베스트 셀러로 자주 선정된 작가들은 주로 시리즈 물을 쓰는 경우가 많았다.

EDA 사용자 평점 별



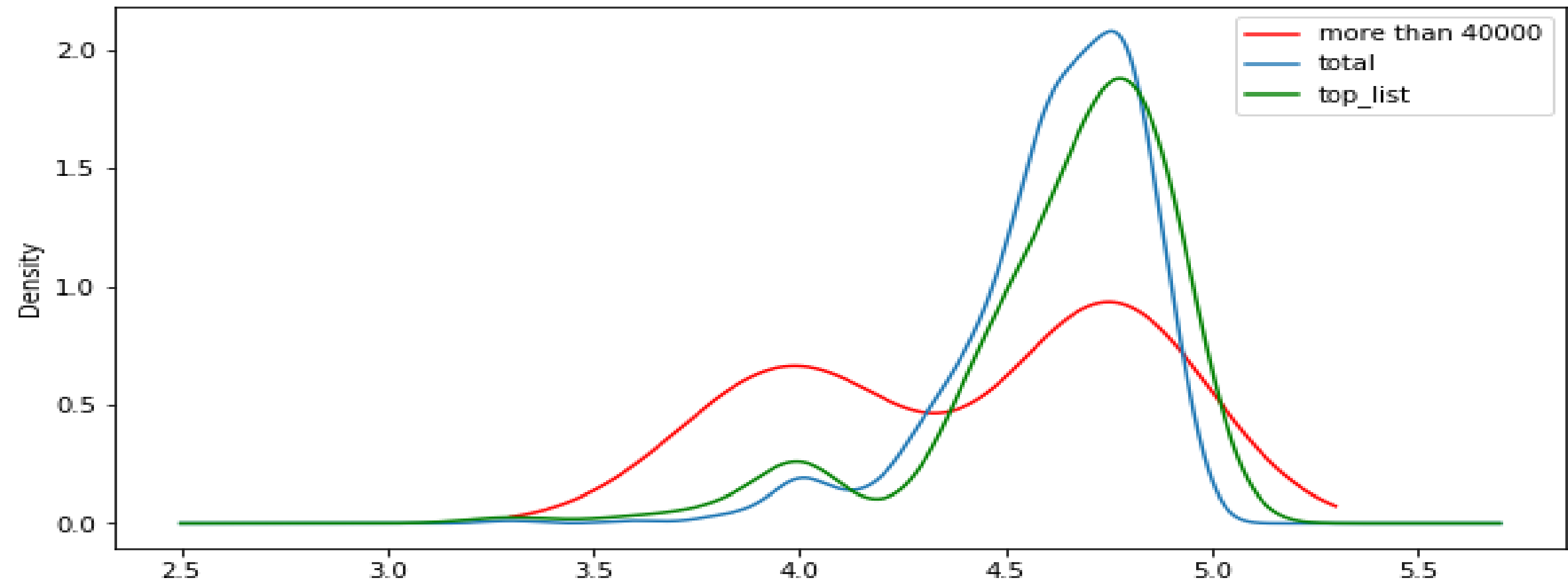
➤ 연도-장르별 사용자 평점의 평균: 2012년과 2013년을 제외하고, 매년 소설책의 사용자 평점 평균이 비소설 책보다 높다.

EDA 사용자 평점 별

More than 40000: 리뷰수가 4만 이상인 작가들의 집단

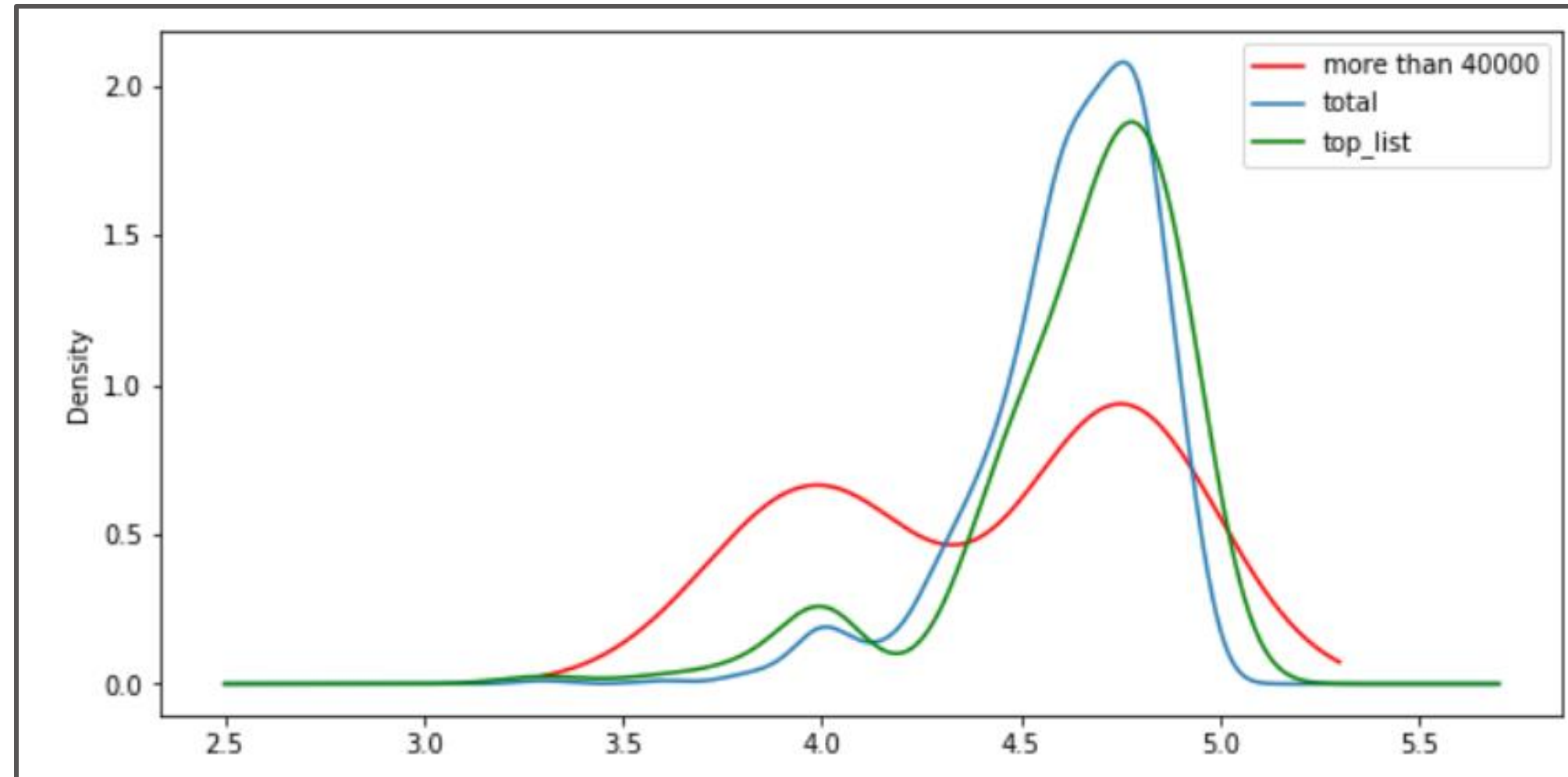
Total: 전체 데이터

Top_list: 베스트 셀러 선정
횟수가 많은 작가들의 집단



- **그룹별 평점 분포:** top_list와 전체 데이터는 유사한 분포이지만 리뷰수가 40,000건인 그룹의 평점은 전체 데이터와 달리 퍼져 있는 형태

작가별 리뷰수별 특징



>>> top_list와 전체 데이터는 유사한 분포 but top_list가 조금 더 높은 평점대에 존재.

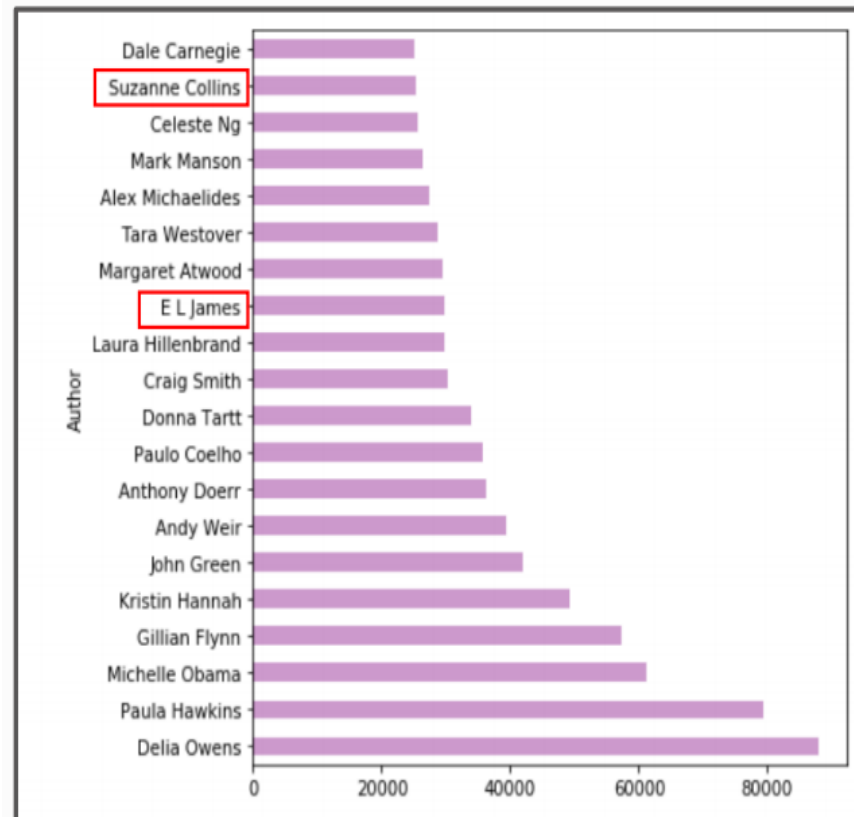
>>> 리뷰수가 40,000건 이상인 그룹의 평점은 퍼져 있음.

- 베스트셀러에 많이 등재되었다고 리뷰 수가 많은 것 x
- 무조건 재미있다고 리뷰를 작성 x

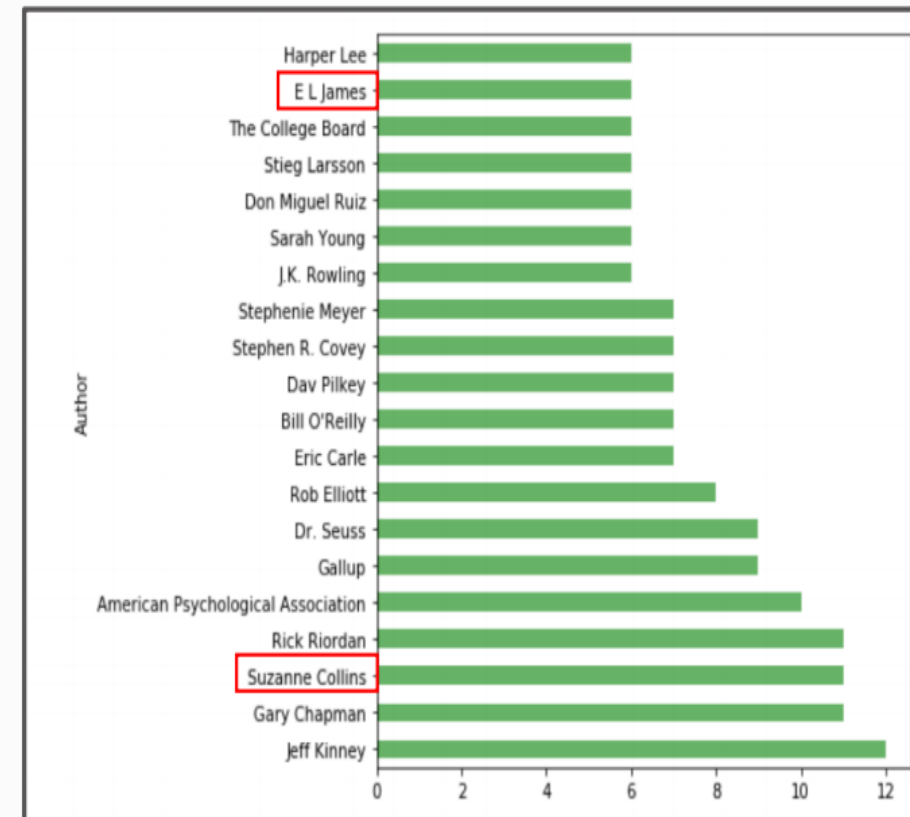


리뷰수와 베스트셀러 상관x

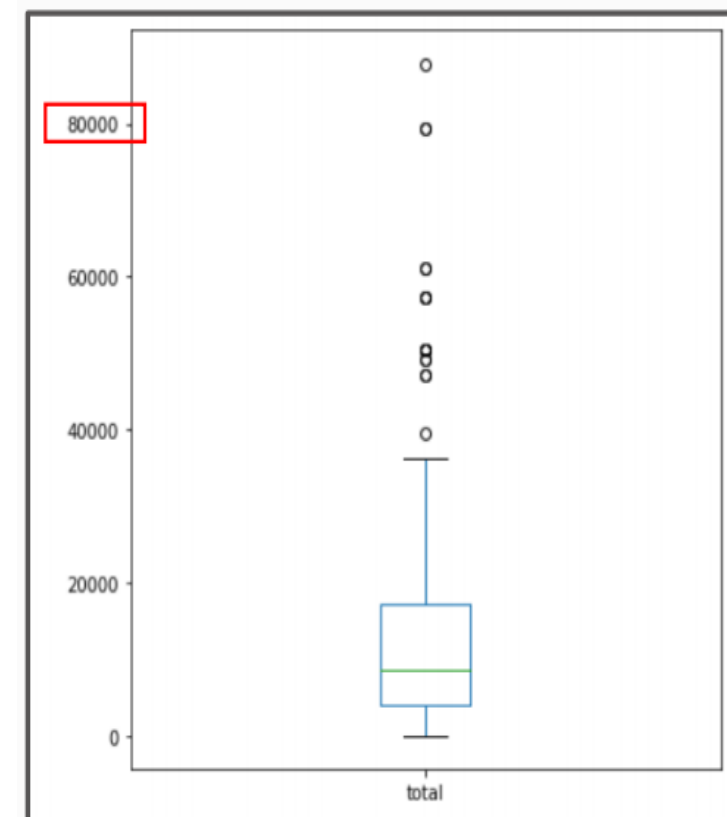
EDA 장르 별



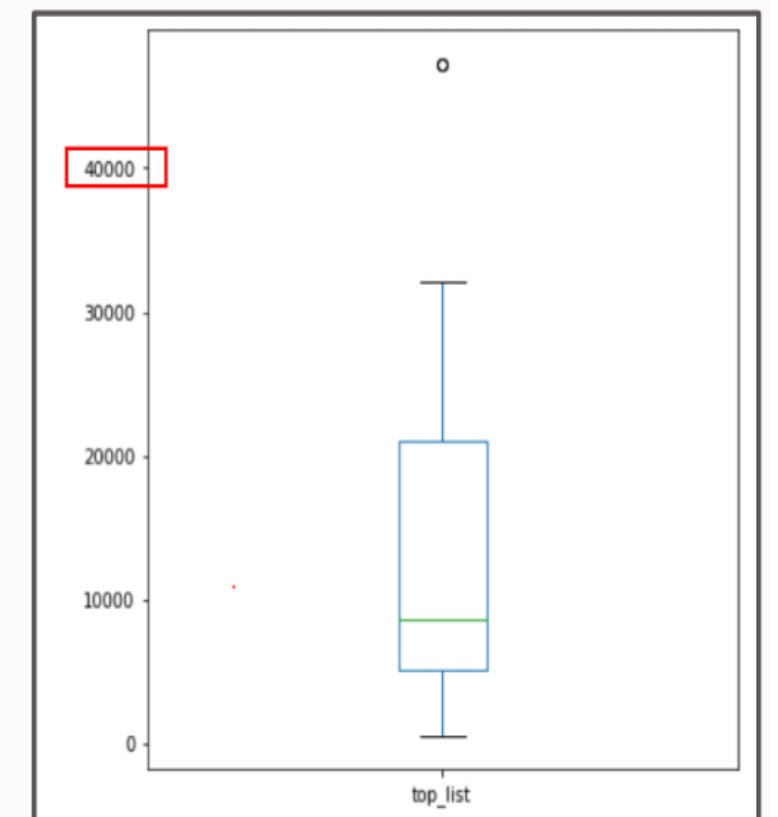
< 리뷰수가 많은 작가들 >



< 베스트셀러에 자주 등재된 작가들 >



< 전체 데이터의 리뷰수 분포 >



< 베스트셀러에 자주 등재된 작가들의 리뷰수 >

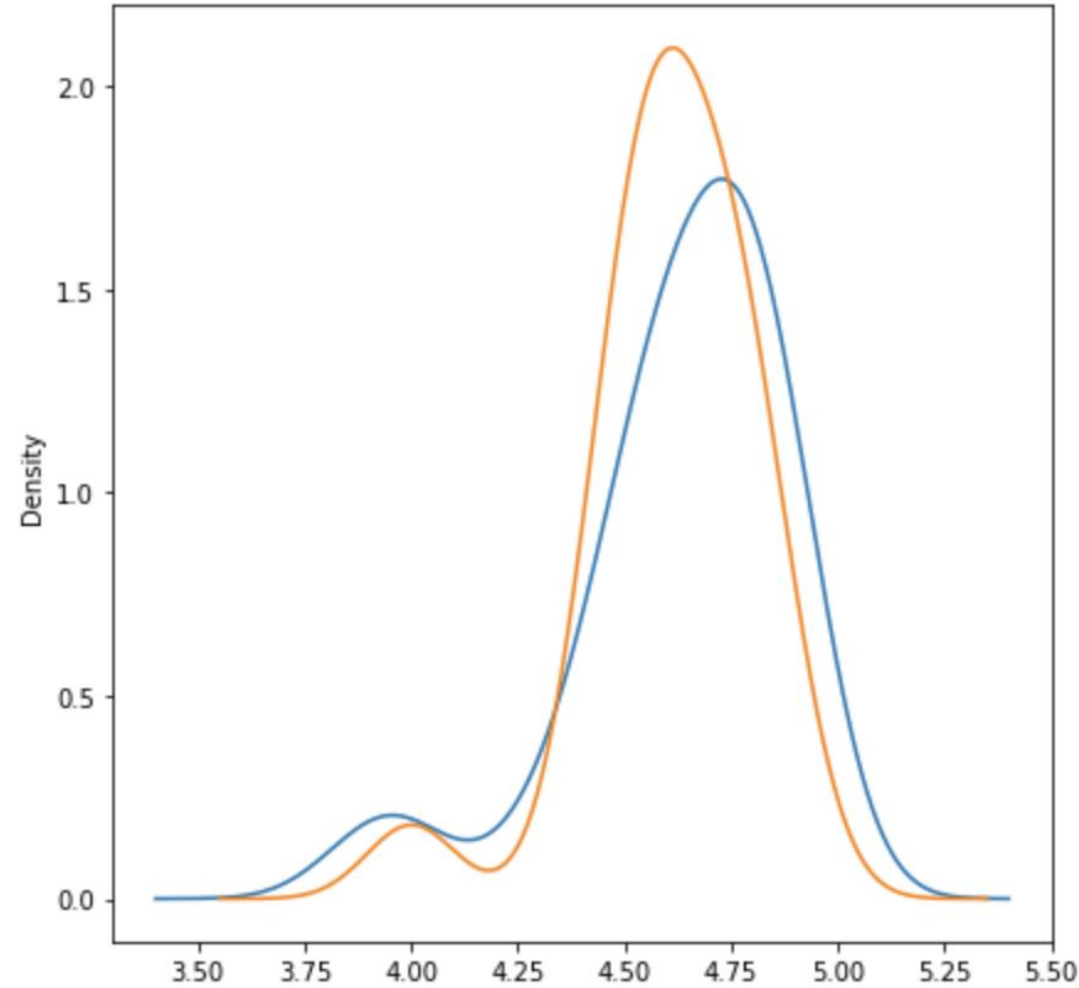
- **리뷰수-베스트셀러 비교:** 리뷰 수와 베스트셀러의 선정 횟수는 관련이 없는 것으로 보인다.
- **Box plot:** 전체 데이터의 리뷰 수는 10,000건 정도, 최대 80,000까지 존재, 베스트 셀러에 자주 선정된 작가들의 소설은 대부분 10,000이하이고 최대 40,000건 까지 존재


2014년 특징 파악해보기


연도별 장르를 보았을 때 특이한 점 발견

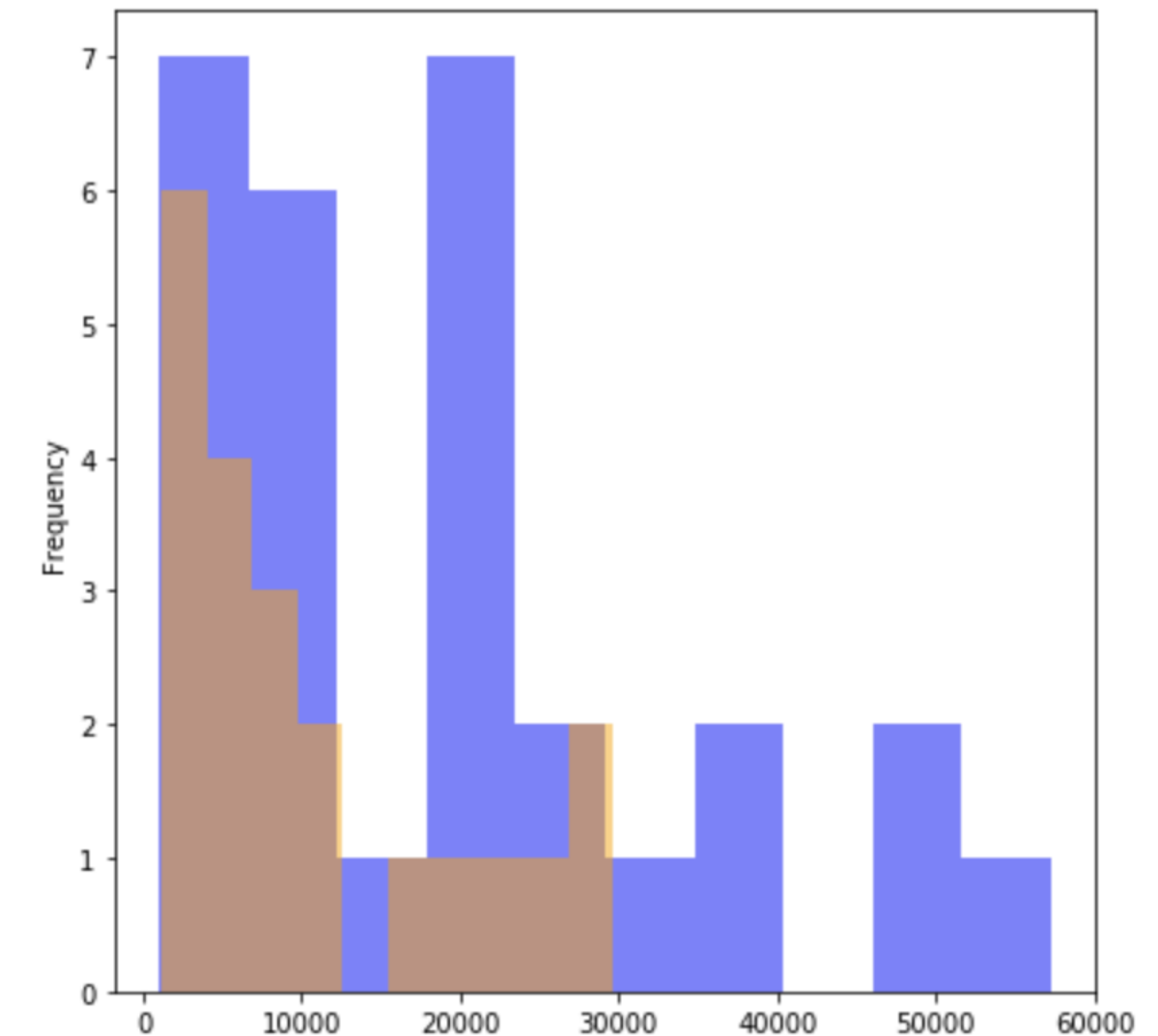
INVESTOR RELATIONS 2020

2014년에는 다른 년도보다 Fiction이 Non Fiction보다 많다.




User Rating
Fiction의 별점이
Non Fiction의 별점보다
높은 점수에
조금 더 분포하고 있다.

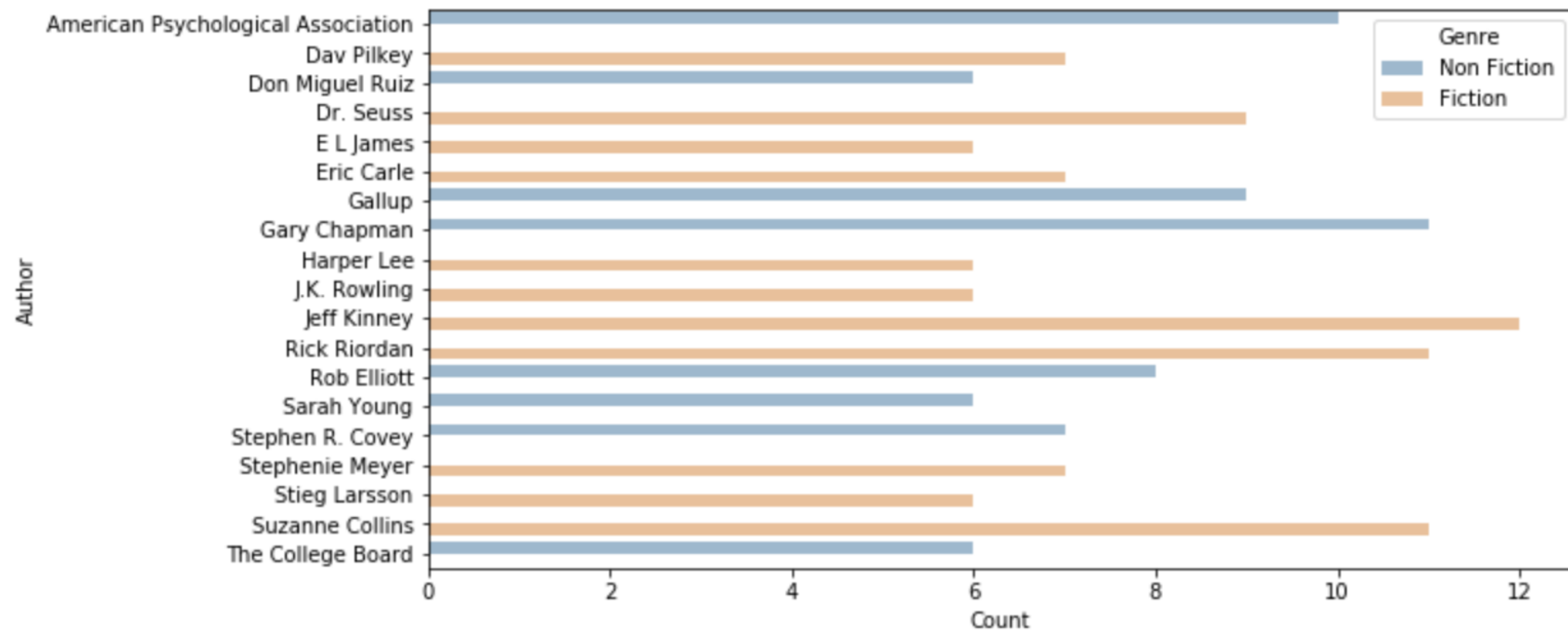

Reviews
Review수가 Fiction이
Non Fiction보다
확실히 많은 것을
확인할 수 있다.



파랑:Fiction/주황:NonFiction



전체적으로 평점은 Fiction이 Non Fiction보다 높다.
2014년 만의 특이점은 아닌것을 알 수 있다.

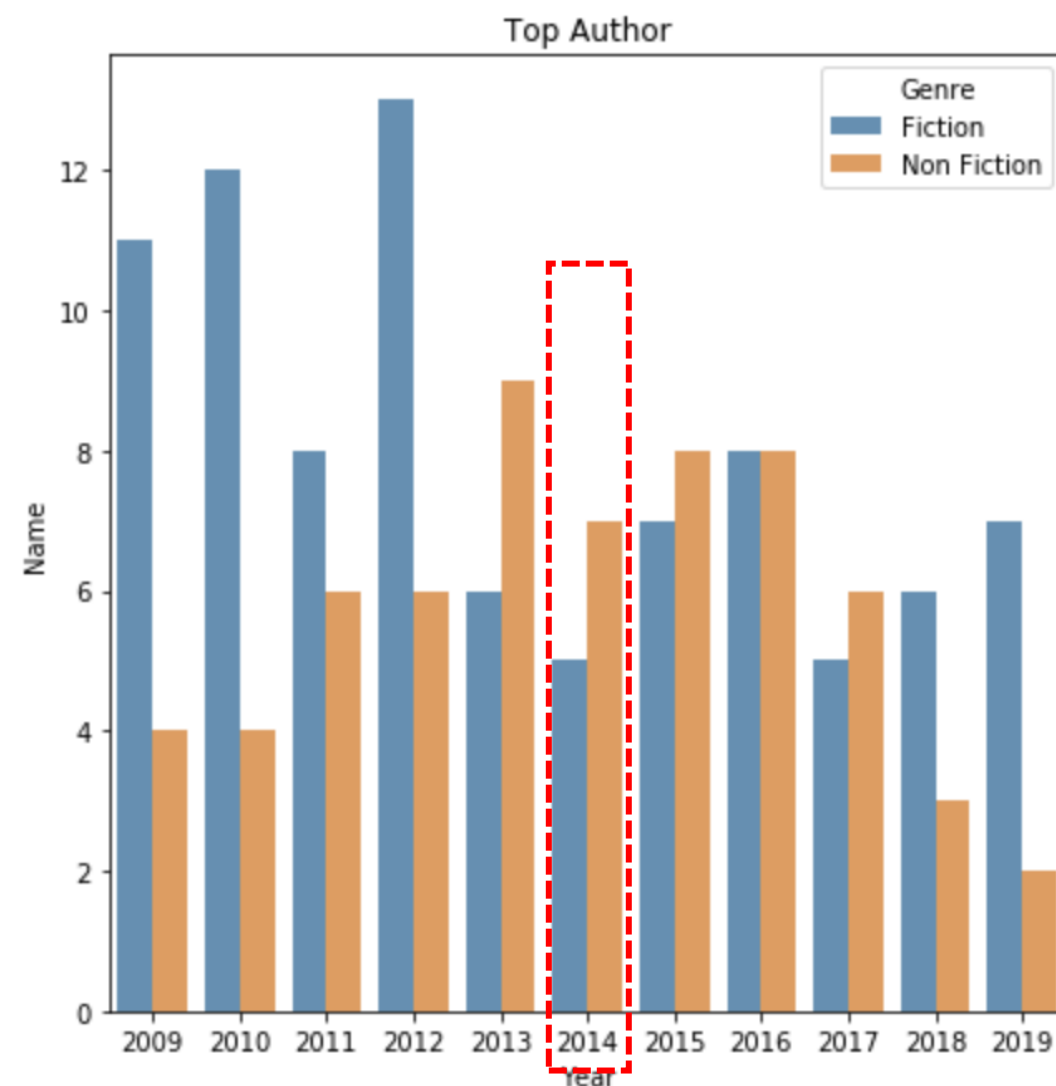


6회 이상 베스트 셀러에 선정된 작가의 경우
Fiction인 경우가 Non Fiction인 경우보다
많은 것을 알 수 있다.

2014년 특징 파악해보기

INVESTOR RELATIONS 2020

2014년에는 다른 년도보다 Fiction이 Non Fiction보다 많다.

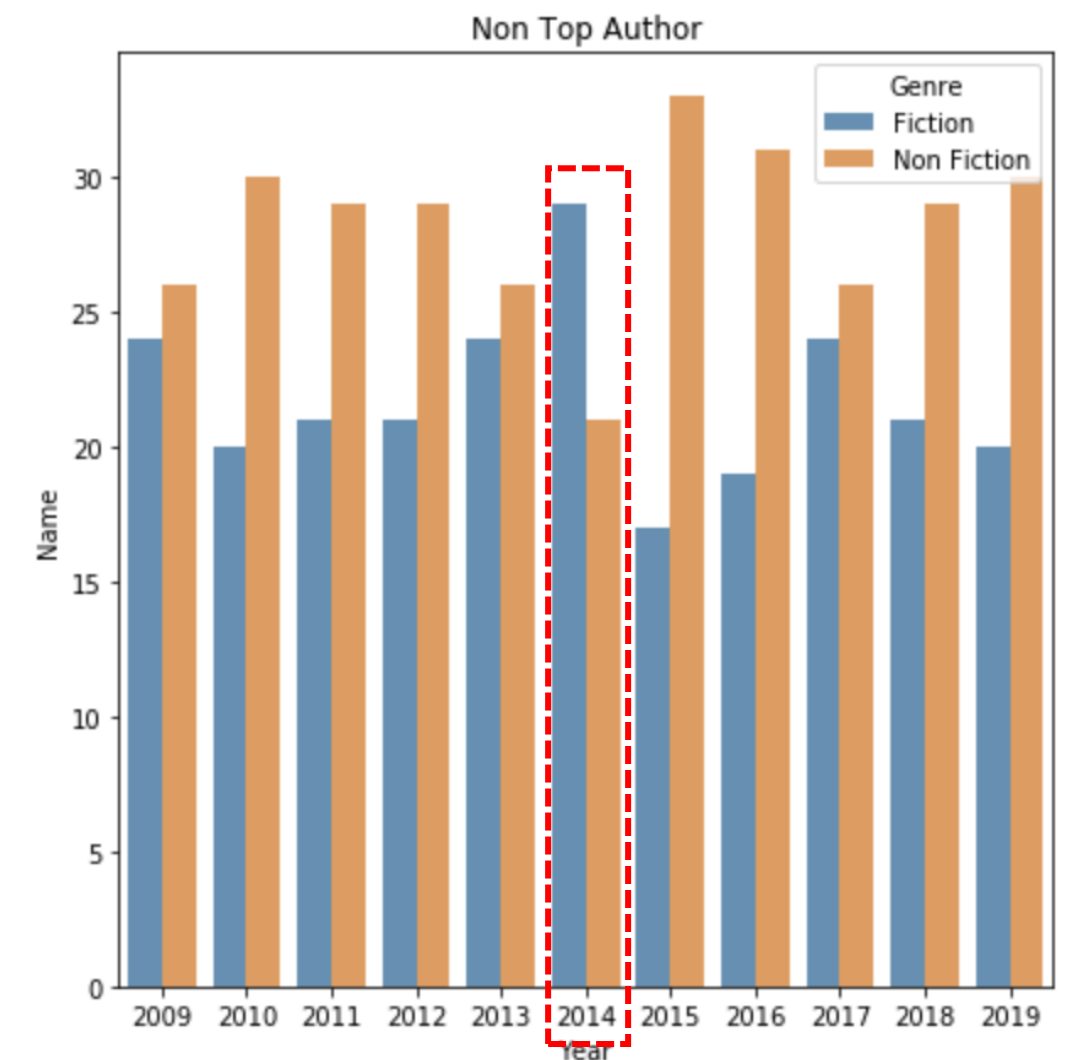


Top Author

Fiction이 더 많은
년도가 많으나,
2014년에는
Non Fiction이
더 많이 출판

Non Top Author

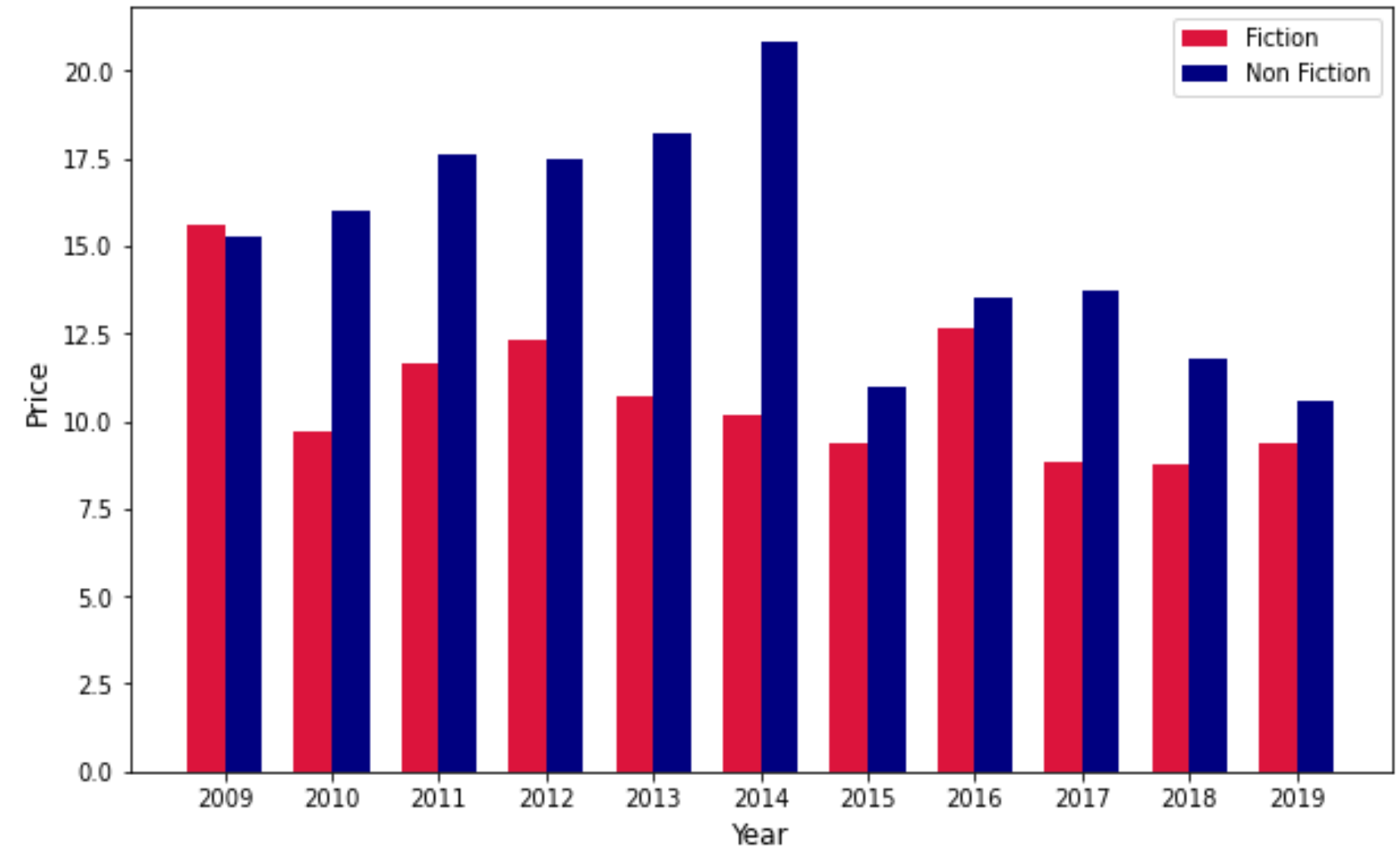
Non Fiction이
더 많은 년도가 많으나,
2014년에는 반대로
Fiction이 더 많이
출판



2014년에는 유명 비 유명 작가들이 장르가 Fiction인 책을 많이 출판하면서 Fiction이 Non Fiction보다 출판 수가 우세하였을 것이라 판단.

EDA 가격 별

	Name	Author	User Rating	Reviews	Price	Year	Genre
42	Cabin Fever (Diary of a Wimpy Kid, Book 6)	Jeff Kinney	4.8	4505	0	2011	Fiction
71	Diary of a Wimpy Kid: Hard Luck, Book 8	Jeff Kinney	4.8	6812	0	2013	Fiction
116	Frozen (Little Golden Book)	RH Disney	4.7	3642	0	2014	Fiction
193	JOURNEY TO THE ICE P	RH Disney	4.6	978	0	2014	Fiction
219	Little Blue Truck	Alice Schertle	4.9	1884	0	2014	Fiction
358	The Constitution of the United States	Delegates of the Constitutional	4.8	2774	0	2016	Non Fiction
381	The Getaway	Jeff Kinney	4.8	5836	0	2017	Fiction
461	The Short Second Life of Bree Tanner: An Eclip...	Stephenie Meyer	4.6	2122	0	2010	Fiction
505	To Kill a Mockingbird	Harper Lee	4.8	26234	0	2013	Fiction
506	To Kill a Mockingbird	Harper Lee	4.8	26234	0	2014	Fiction
507	To Kill a Mockingbird	Harper Lee	4.8	26234	0	2015	Fiction
508	To Kill a Mockingbird	Harper Lee	4.8	26234	0	2016	Fiction



➤ price가 0인 책: 가격이 0인 책 9권이 있었다.

➤ 연도-장르별 책 가격 평균: 2009년을 제외하고 비소설 책의 평균 가격은 매년 소설 책보다 높다.

문서 표현 방법

Count Vector

- ✓ 단어가 문서에서 몇 번 사용 되었는지
- ✓ 사용된 횟수가 weight로 작용
- ✓ the, a 등의 많은 문서에서 공통적으로 나타나는 단어는 중요성이 떨어지는 단어일 가능성 ↑

TFIDF Vector

- ✓ 단어의 count를 단어가 나타난 문서의 수로 나눠 자주 등장하지 않는 단어의 weight를 올림
- ✓ $tf(d, t)$ 는 문서 d 에 단어 t 가 나타난 횟수
- ✓ $df(t)$ 는 전체 문서 중 단어 t 를 포함하는 문서의 수
- ✓ $idf(t)$ 는 전체 문서를 $1+df(t)$ 로 나눈 값에 로그 적용

문서 표현 방법 비교

예시)

DF 1 : It is a puppy and it is extremely cute.

DF 2 : It is a cat and it is scary.

idf(t) (전체 문서 수 5 가정)

- ✓ it, is 등은 두 개의 문서에, puppy, cat 등은 하나의 문서에 출현
- ✓ it의 idf는 $\log(5/(1+2))=0.22$,
puppy의 idf는 $\log(5/(1+1))=0.4$

word	it	is	a	puppy	and	extremely	cute	cat	scary
DF1	2	2	1	1	1	1	1	0	0
DF2	2	2	1	0	1	0	0	1	1

Count Vector = $tf(d, t)$

word	it	is	a	puppy	...
DF1	$2 * 0.22$	$2 * 0.22$	$1 * 0.4$...
DF2	$2 * 0.22$	$2 * 0.22$...	$0 * 0.4$...

TFIDF

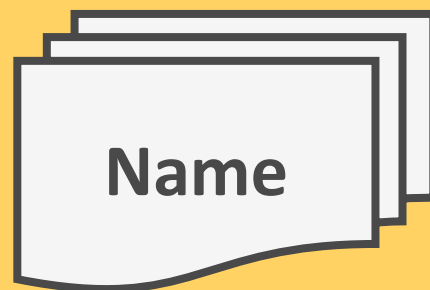
분류 과정

01

책 제목을 X로, 장르를 Y로 설정하고
Train set과 Test set으로 분할(비율 = 0.2)

03

Count Vector 혹은 TFIDF Vector로 변환



(0, 421)	1
(0, 275)	1
(0, 656)	1
(0, 104)	1
(0, 20)	1
(0, 255)	1
(0, 552)	1

(0, 552)	0.39958475756243206
(0, 255)	0.35628027376486454
(0, 20)	0.39958475756243206
(0, 104)	0.39958475756243206
(0, 656)	0.342339343795833
(0, 275)	0.342339343795833
(0, 421)	0.39958475756243206

02

텍스트 전처리 수행
- 토큰나이징, 불용어 제거

```
RegTok = RegexpTokenizer("[\w']{2,}")

stop_words = set(stopwords.words('english'))

def tokenizer(text):
    tokens = RegTok.tokenize(text.lower())
    words = [word for word in tokens if (word not in stop_words
                                         and len(word) > 2)]
    features = (list(map(lambda token: PorterStemmer().stem(token), words)))
    return features
```

04

여러가지 분류 모델에 적용
(나이브베이지스, 릿지, 라쏘 등)

여러 분류 모델에 적용

라쏘 회귀

Train set score : 0.914
Test set score : 0.818

로지스틱 회귀

Train set score : 0.984
Test set score : 0.855

나이브베이즈

Train set score : 0.993
Test set score : 0.855

릿지 회귀

Train set score : 1.000
Test set score : 0.945



Count Vector



TFIDF Vector

라쏘 회귀

Train set score : 0.852
Test set score : 0.800

로지스틱 회귀

Train set score : 0.989
Test set score : 0.845

나이브베이즈

Train set score : 0.993
Test set score : 0.855

릿지 회귀

Train set score : 0.998
Test set score : 0.909

Count Vector 기반 릿지 회귀 적용

	Name	Genre	Prediction
0	Have a Little Faith: A True Story	Non Fiction	Non Fiction
1	StrengthsFinder 2.0	Non Fiction	Non Fiction
2	Dear Zoo: A Lift-the-Flap Book	Fiction	Fiction
3	The Life-Changing Magic of Tidying Up: The Jap...	Non Fiction	Non Fiction
4	Gone Girl	Fiction	Fiction
5	Last Week Tonight with John Oliver Presents A ...	Fiction	Fiction
6	StrengthsFinder 2.0	Non Fiction	Non Fiction
7	Secret Garden: An Inky Treasure Hunt and Color...	Non Fiction	Non Fiction
8	Proof of Heaven: A Neurosurgeon's Journey into...	Non Fiction	Non Fiction
9	The Elegance of the Hedgehog	Fiction	Fiction
10	The Short Second Life of Bree Tanner: An Eclip...	Fiction	Fiction
	•		
	•		
	•		
15	Rush Revere and the First Patriots: Time-Trave...	Fiction	Fiction
16	The Martian	Fiction	Fiction
17	Wild: From Lost to Found on the Pacific Crest ...	Non Fiction	Fiction
18	Cravings: Recipes for All the Food You Want to...	Non Fiction	Non Fiction
19	Decision Points	Non Fiction	Non Fiction



Train set 정확도

1.0



Test set 정확도

0.945

올바르지 않은 분류

딥러닝 모델 사용

Word Embedding

- ✓ 단어가 쓰인 순서 정보 사용 가능
- ✓ One-hot-encoding으로 표현된 단어를 dense한 vector로 변환
- ✓ 변환된 vector를 이용해 학습

※ One hot encoding

: 각 단어를 모든 문서에 사용된 단어들의 수 길이의 벡터로 표현(1과 0)

예시)

It = [1, 0, 0, 0]

Is = [0, 1, 0, 0]

A = [0, 0, 1, 0]

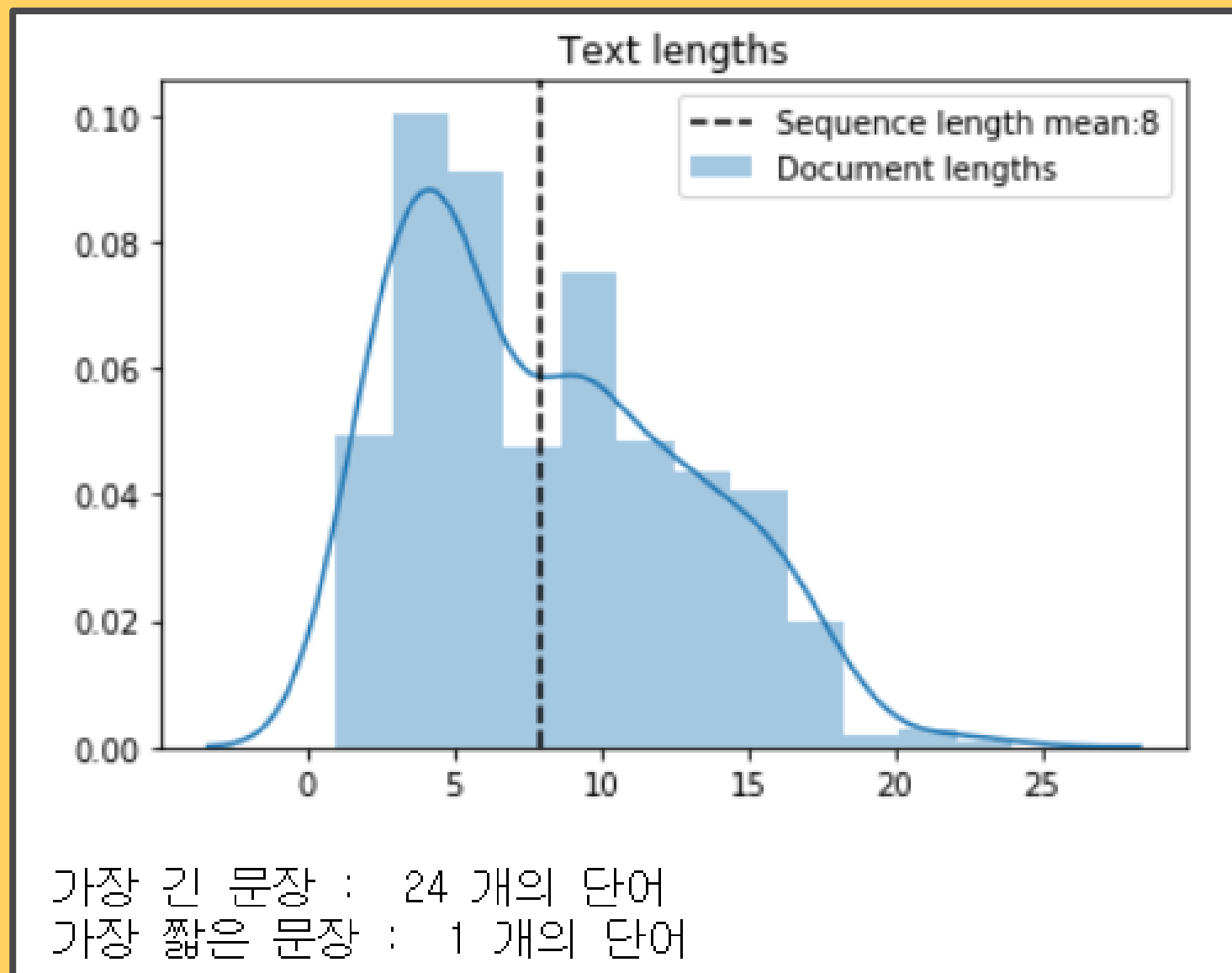
Puppy = [0, 0, 0, 1]

- 사이즈가 줄어들고, 단어들의 거리 계산 가능
- 단어의 의미 내포 가능

모델 적용 과정

01

토큰나이저 및 매개변수 결정
(최대 단어 개수 1000개, 벡터 길이 12)



02

단어들의 sequence로 표현 및 Genre를 0과 1로 변환

[132	58	609	610	611	0	0	0	0	0	0	0]
[278	361	612	3	15	0	0	0	0	0	0	0]
[613	362	7	13	21	614	5	615	0	0	0	0]
[616	617	363	0	0	0	0	0	0	0	0	0]
[20	618	169	364	106	279	214	215	22	0	0	0]
[3	107	17	170	3	619	2	365	4	59	0	0]
[3	218	2	219	3	220	2	221	3	107	17	170]
[3	620	8	621	3	15	0	0	0	0	0	0]
[3	622	623	171	280	4	133	0	0	0	0	0]
[3	23	366	367	3	15	0	0	0	0	0	0]]

03

Train set과 Test set로 분할 (비율 = 0.2)

04

모델 학습

모델 학습

```
from tensorflow.keras import Sequential
from tensorflow.keras.layers import Flatten, Conv1D, MaxPooling1D, Dense,
                                Embedding, LSTM, Dropout, Bidirectional
from tensorflow.keras.optimizers import Adam

embedding_dim=200

model2 = Sequential([
    Embedding(max_words, embedding_dim, input_length=maxlen),
    Conv1D(50, kernel_size=5, strides=1, padding='valid'),
    MaxPooling1D(1, padding='valid'),
    Flatten(),
    Dense(512, activation='relu'),
    Dropout(0.1),
    Dense(2, activation='sigmoid')
])

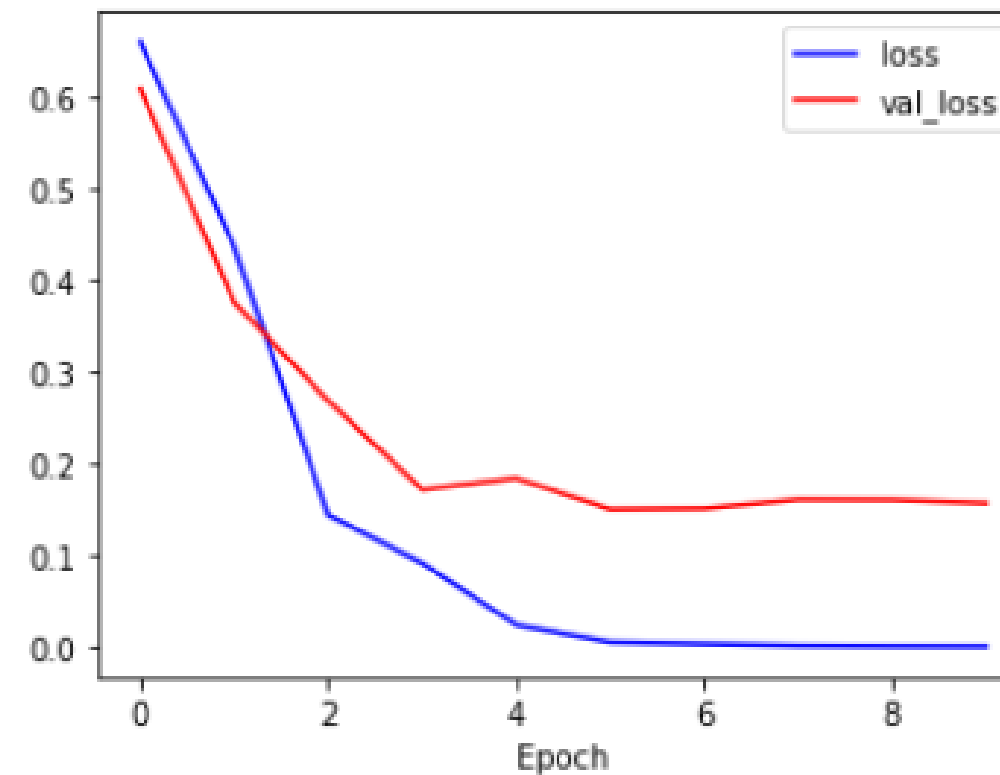
adam = Adam()
model2.compile(loss='binary_crossentropy',
               optimizer='adam',
               metrics=['accuracy'])

model2.summary()
```

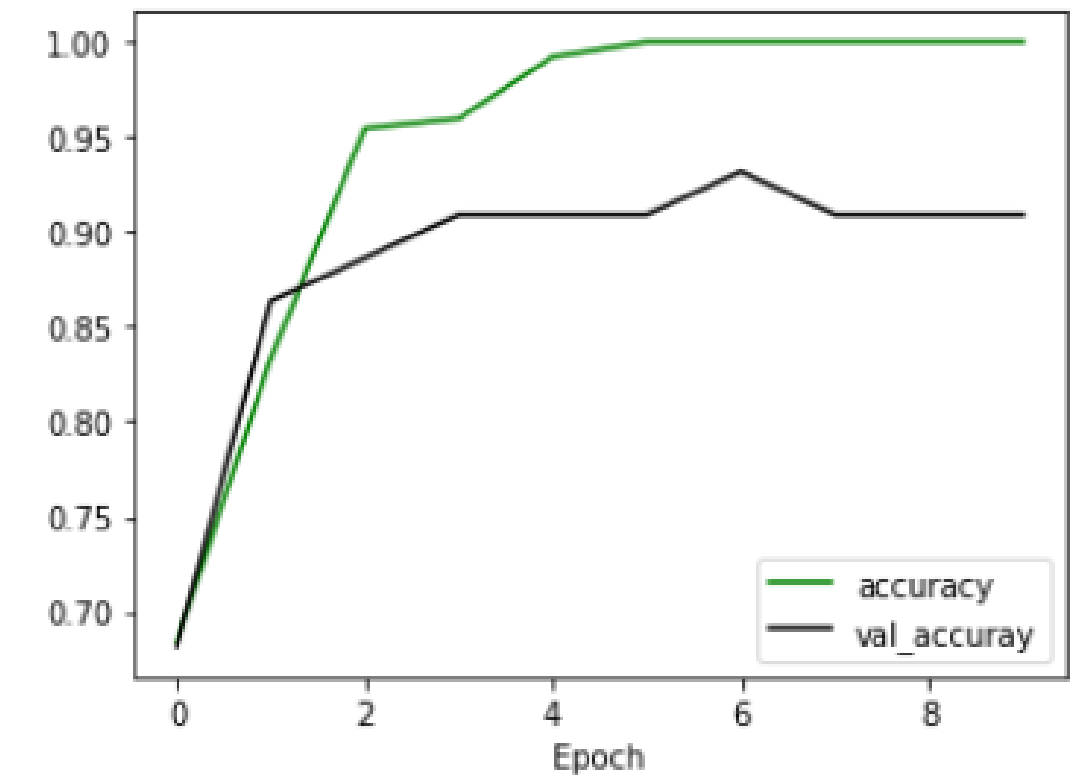
```
print(model2.evaluate(x_test, y_ts)[1])
```

```
4/4 [=====] - 0s 3ms/step - loss: 0.3658 - accuracy: 0.8
818
0.8818181753158569
```

Loss



Accuracy



➤ 0과 1로 구분하기 때문에 활성화 함수는 sigmoid와 loss를 binary_crossentropy 사용

➤ Test set 적용했을 때의 정확성 0.8818

결과

	0	1	prediction
0	1.215771e-01	0.854602	1
1	6.182451e-07	0.999999	1
2	9.999647e-01	0.000039	0
3	1.209180e-04	0.999881	1
4	9.991783e-01	0.000869	0
5	2.846622e-02	0.973460	1
6	6.182451e-07	0.999999	1
7	4.597604e-04	0.999558	1
8	1.271367e-04	0.999861	1
9	9.663558e-01	0.033607	0



	genre	prediction
0	1	1
1	1	1
2	0	0
3	1	1
4	0	0
5	0	1
6	1	1
7	1	1
8	1	1
9	0	0

올바르지 않은 분류

Count Vector 기반 릿지 회귀 적용

	Name	Author	Prediction
0	Have a Little Faith: A True Story	Mitch Albom	Edward M. Kennedy
1	StrengthsFinder 2.0	Gallup	Gallup
2	Dear Zoo: A Lift-the-Flap Book	Rod Campbell	Rod Campbell
3	The Life-Changing Magic of Tidying Up: The Jap...	Marie Kondō	Marie Kondō
4	Gone Girl	Gillian Flynn	Gillian Flynn
5	Last Week Tonight with John Oliver Presents A ...	Jill Twiss	Elizabeth Strout
6	StrengthsFinder 2.0	Gallup	Gallup
7	Secret Garden: An Inky Treasure Hunt and Color...	Johanna Basford	Marjorie Sarnat
8	Proof of Heaven: A Neurosurgeon's Journey into...	Eben Alexander	Eben Alexander
9	The Elegance of the Hedgehog	Muriel Barbery	John Grisham
10	The Short Second Life of Bree Tanner: An Eclip...	Stephenie Meyer	Stephenie Meyer
11	Where the Crawdads Sing	Delia Owens	John Grisham
12	Jesus Calling: Enjoying Peace in His Presence ...	Sarah Young	Sarah Young
13	Wheat Belly: Lose the Wheat, Lose the Weight, ...	William Davis	William Davis
14	Magnolia Table: A Collection of Recipes for Ga...	Joanna Gaines	Chip Gaines
15	Rush Revere and the First Patriots: Time-Trave...	Rush Limbaugh	Rush Limbaugh
16	The Martian	Andy Weir	John Grisham
17	Wild: From Lost to Found on the Pacific Crest ...	Cheryl Strayed	Dan Brown
18	Cravings: Recipes for All the Food You Want to...	Chrissy Teigen	David Zinczenko
19	Decision Points	George W. Bush	Malcolm Gladwell

올바르지 않은 분류



Train set 정확도

0.993



Test set 정확도

0.618

→ 과대적합

작가 분류 실패 요인

- ✓ 248명의 작가가 존재하는데 비해 총 550개의 row로 데이터 부족.
- ✓ 책 제목에 따라 정보 부족 및 길이(단어의 수)의 차이 존재.
- ✓ 대부분의 제목이 the, a 등의 단어를 포함해 특성 추출 시 어려움 존재.
- ✓ 베스트셀러에 오른 연도나 가격이 다르면 다른 데이터로 분류.
- ✓ 하나의 작품이 베스트셀러에 여러 번 오른 작가가 존재하는 반면에 한 번만 오른 작가도 존재.
- ✓ 여러 작품이 베스트셀러에 오른 작가가 존재하는 반면에 하나의 작품만 가지는 작가도 존재.



Amazon

Top 50 Bestselling Books

2009 - 2019

프로젝트 진행하면서 느낀 점 및 셀프 피드백

- ✓ 데이터 크기가 상당히 작아서 분석에 제한이 많았다.
- ✓ 시각화 및 인사이트 도출 결과, 대부분이 상식적으로 알 수 있는 내용이 많아 아쉬웠다.
- ✓ 분기별로 나누어져 있거나, 리뷰의 전체 내용, 긍정 부정 평가가 주어져 있었다면,
더 다양한 정보를 도출할 수 있지 않았을까 라는 아쉬움이 많이 남는 분석이었다.

이상으로 DNA3조 발표를 마치겠습니다. 감사합니다.