پروژه داده کاوی بر روی دیتاست GOODREADS-BOOKS

استاد درس: دکتر خدمتی

گردآورندگان:

- پوریا نادری پور 97104667
- پارسا بختیاری 97104118

نیم سال دوم سال تحصیلی 1400-1401

فهرست مطالب

فهرست شکل‌ها

فصل ۱: مقدمه و مرور ادبیات

۱-۱- مقدمه

یکی از تفریحات بخش قابل توجهی از مردم جهان کتابخوانی است ولی شاید در برخی جوامع ارزش خواندن کتاب به اندازه‌ی کافی بالا نباشد و کتابخوانی را وقت تلف کردن در نظر گیرند. خواندن کتاب فواید بسیاری به همراه دارد به طور مثال می‌تواند باعث تقویت تئوری ذهن، افزایش همدلی، تقویت قدرت یادگیری و حافظه به کمک تخیل، بهبود فرآیندهای حسی و قدرت تصمیم گیری، کمک به فعالیت‌های ذهن‌پروری، تقویت مهارت صحبت کردن و دامنه لغات و کاهش سرعت پیر شدن مغز شود. یکی از موارد مهم در کتابخوانی انتخاب کتاب درست و موردعلاقه است که بتواند به بهترین شکل ممکن با بهره بردن از فواید مطالعه کتاب، به ما کمک کند. پس نیاز داریم تا بتوانیم کتاب مدنظر خود را برای مطالعه از میان هزاران کتاب موجود در بازار و اینترنت پیدا کرده و از مطالعه آن لذت و بهره کافی را ببریم. در این راستا در دیتاست مدنظر به دنبال خوشه‌بندی و پیدا کردن دسته‌های مختلف کتاب‌ها و پیشنهاد کتاب به کاربر متناسب با علاقه‌ی وی هستیم.

۱-۲- مرور ادبیات

از زمان‌های قدیم تا به حال داده‌کاوی روند تکاملی را طی کرده و روزانه بر اهمیتش افزوده شده است. با گذشت زمان از قدیم و با افزایش میزان داده‌ها، نیاز به داده‌کاوی و ابزارهای آن حس شد تا در مسیر تکامل به سمت استفاده از ابزارهای داده‌کاوی برویم و با استفاده از داده‌ها وضعیت گذشته و حال را بررسی و آینده را پیش‌بینی کنیم.

همانطور که گفتیم جهت بهره بردن بهینه از کتابخوانی باید ابتدا مشکل انتخاب کتاب را برطرف کرد. جهت انجام این کار نیاز به دیتاست‌هایی داریم که داده‌کاوری را روی آن‌ها انجام دهیم. برای این منظور دیتاست‌های بسیاری تا به حال گردآوری شده است، به طور مثال شخصی به نام SOUMIK، سال ۲۰۲۰ دیتاستی از مجموعه کتاب‌های گودریدز را به همراه مشخصات مختلفی گردآوری کرد که قرار است در این پروژه روی این دیتاست داده‌کاوی را انجام دهیم. شخصی به نام Sooter Saalu در سال ۲۰۲۰، ۵۰ کتاب پرفروش آمازون از سال ۲۰۰۹ تا ۲۰۱۹ ار جمع‌آوری کرده و فرد دیگری به نام PANAGIOTIS SIMAKIS نیز در سال ۲۰۲۱ مجموعه عظیمی از کتاب‌های مختلف از ژانرهای متفاوت را در یک دیتاست گردآوری کرده است. حال با وجود همچین حجم عظیمی از دیتاست‌ها درباره کتاب‌های مختلف، نفرات بسیاری بر روی این دیتاست‌ها تحقیقات خود را انجام داده‌اند و مقایسه و پیش‌بینی‌های خود را ارائه کرده اند.

فصل ۲: ارائه‌ی مدل پیش‌بینی

۲-۱- تشریح داده‌ها

در این مرحله سعی بر این داریم تا با دیتاست بیشتر آشنا شویم و نسبت به داده‌ها دید خوبی پیدا کنیم. همچنین درصدد این هستیم تا انواع داده‌ها و روابطشان را با یکدیگر پیدا کنیم.

۲-۱-۱- دید اولیه به داده‌ها

در اولین مرحله یک دید ابتدایی به داده‌ها پیدا کرده و تعداد تاپل‌ها و اتریبیوت‌ها و ۵ ردیف ابتدایی و انتهایی دیتاست را مشاهده می‌کنیم.

```
df.index = df['bookID']
print("Dataset contains {} rows and {} columns".format(df.shape[0], df.shape[1]))
```

```
Dataset contains 13714 rows and 10 columns
```

شکل ۱- تعداد ردیف‌ها و ستون‌های دیتاست

bookID	bookID	title	authors	average_rating	isbn	isbn13	language_code	# num_pages	ratings_count	text_reviews_count
1	1	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling	4.56	0439785960	9780439785969	eng	652	1944099	26249
2	2	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling	4.49	0439358078	9780439358071	eng	870	1996446	27613
3	3	Harry Potter and the Sorcerer's Stone (Harry P...	J.K. Rowling	4.47	0439554934	9780439554930	eng	320	5629932	70390
4	4	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.41	0439554896	9780439554893	eng	352	6267	272
5	5	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	4.55	043965548X	9780439655484	eng	435	2149872	33964

شکل ۲- ۵ ردیف ابتدایی دیتاست

bookID	bookID	title	authors	average_rating	isbn	isbn13	language_code	# num_pages	ratings_count	text_reviews_count
47699	47699	M Is for Magic	Neil Gaiman-Teddy Kristiansen	3.82	0061186422	9780061186424	eng	260	11317	1060
47700	47700	Black Orchid	Neil Gaiman-Dave McKean	3.72	0930289552	9780930289553	eng	160	8710	361
47701	47701	InterWorld (InterWorld #1)	Neil Gaiman-Michael Reaves	3.53	0061238961	9780061238963	en-US	239	14334	1485
47708	47708	The Faeries' Oracle	Brian Froud-Jessica Macbeth	4.43	0743201116	9780743201117	eng	224	1550	38
47709	47709	The World of The Dark Crystal	Brian Froud	4.29	1862056242	9781862056244	eng	132	3572	33

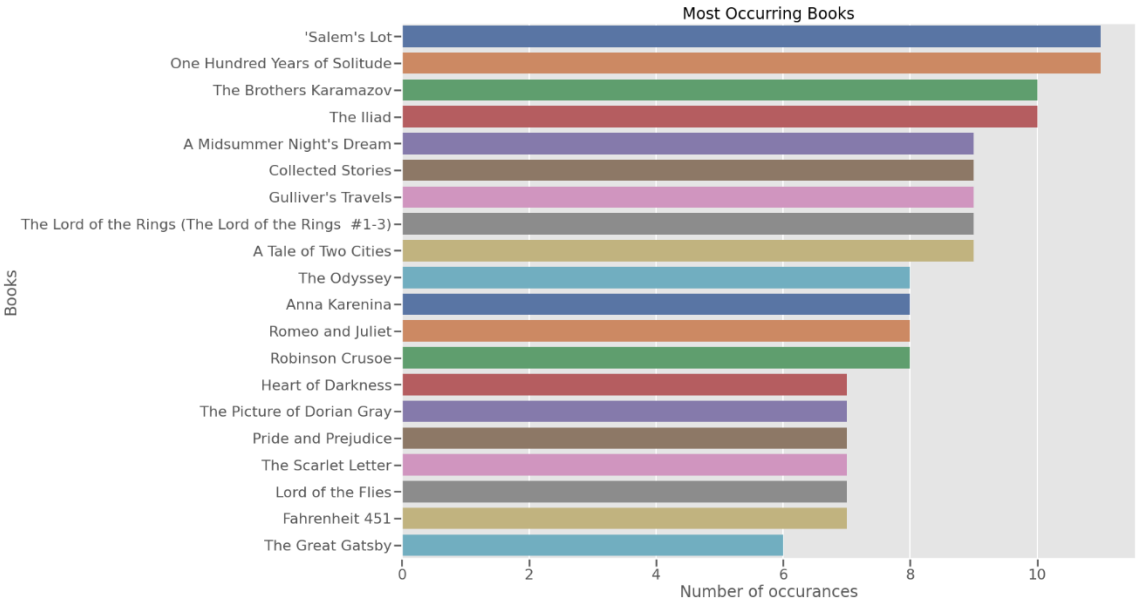
شکل ۳- ۵ ردیف انتهایی دیتاست

```
bookID              int64
title              object
authors            object
average_rating     float64
isbn               object
isbn13              int64
language_code      object
# num_pages         int64
ratings_count       int64
text_reviews_count  int64
dtype: object
```

شکل ۴- تفکیک نوع اتریبیوت‌ها

۲-۱-۲- کتاب‌های با بیشترین تکرار در دیتاست

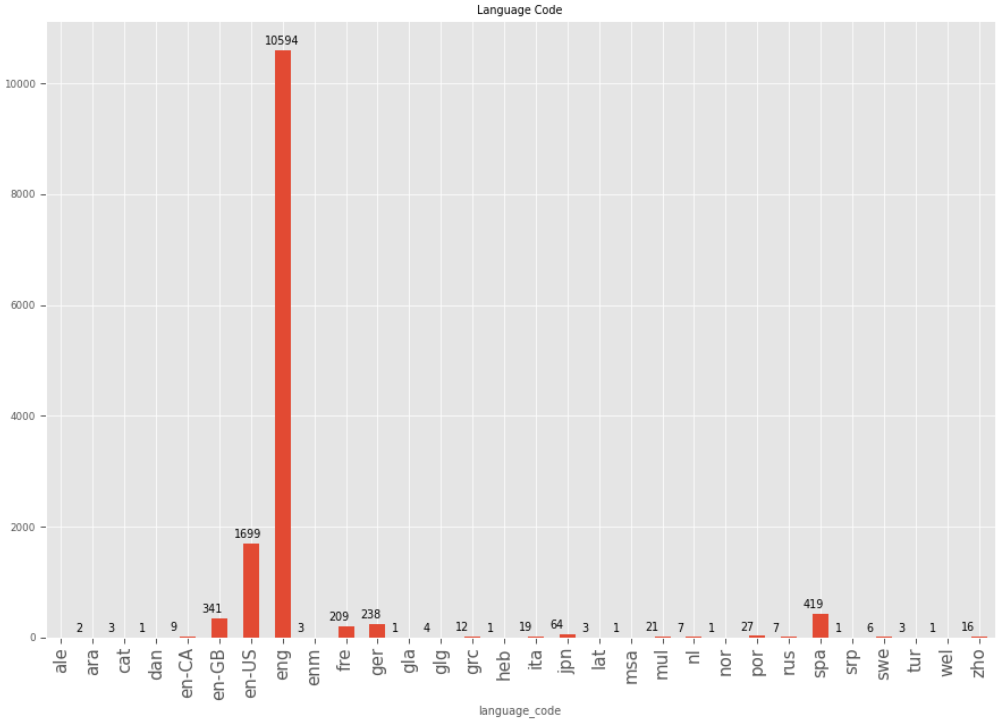
با مشاهده شکل زیر متوجه می‌شویم که دو کتاب "One Hundred Years of Solitude" , "Salem's Lot" بیشترین تکرار را در دیتاست داشته‌اند.



شکل ۵- عنوان ۲۰ کتاب پرتکرار در دیتاست

۲-۱-۳- توزیع کتاب‌ها براساس زبان

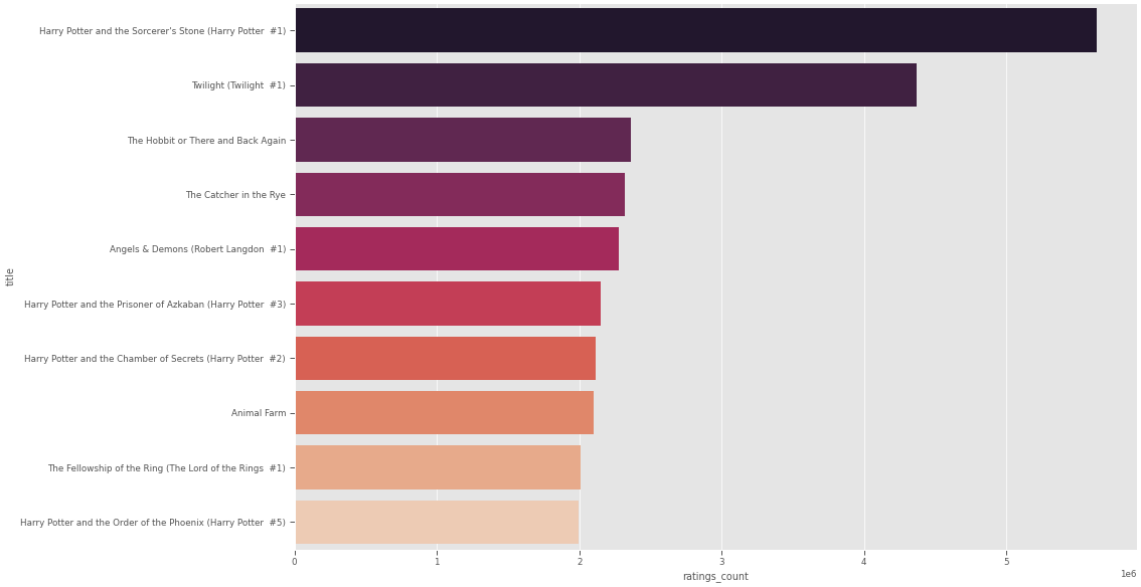
همانطور که مشخص است زبان انگلیسی بیشتر کتاب‌ها را پوشش می‌دهد و زبان‌های اسپانیایی، فرانسوی و آلمانی در رده‌های بعدی قرار دارند.



شکل ۶- تعداد کتاب‌ها بر اساس زبان

۲-۱-۴- ۱۰ کتاب با بیشترین تعداد امتیازات از سوی کاربران

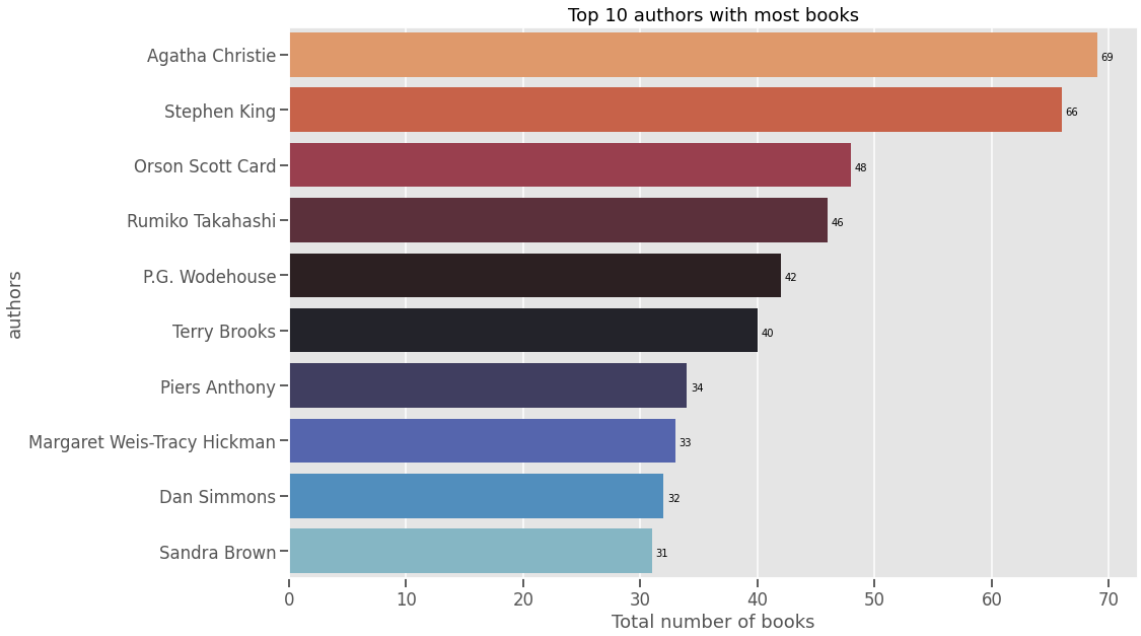
در رتبه اول کتاب "Harry Potter and the Sorcerer's Stone" بیشترین تعداد امتیاز را در دیتاست دارد.



شکل ۷- ۱۰ کتاب برتر از نظر تعداد آرا

۲-۱-۵- نویسندگان با بیشترین تعداد کتاب

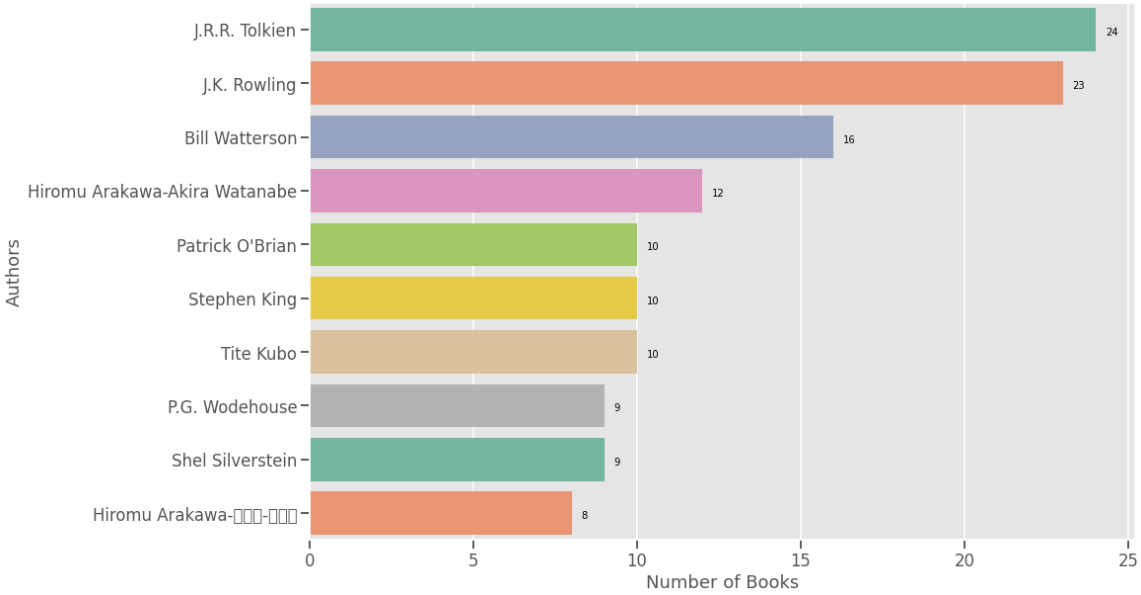
با توجه به خروجی "Agatha Christie" با ۶۹ کتاب دارای بیشترین کتاب در دیتاست است.



شکل ۸- ۱۰ نویسنده برتر از نظر تعداد کتاب

۲-۱-۶- نویسندگان با تعداد کتاب‌های با میانگین امتیاز بالا

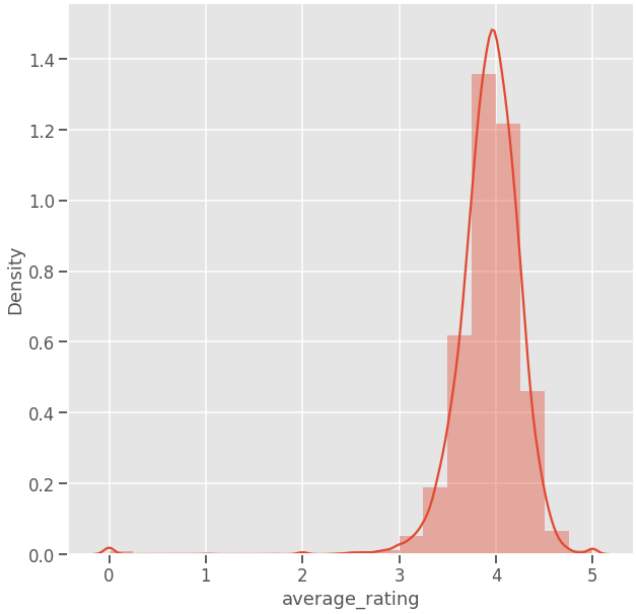
"J.R.R. Tolkien" با ۲۴ کتاب که امتیاز بالای ۴.۳ از ۵ دارند در رتبه‌ی اول در بین نویسندگان قرار دارد.



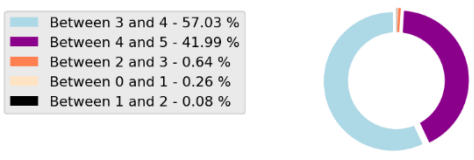
شکل ۹- ۱۰ نویسنده برتر از نظر تعداد کتاب‌های با میانگین امتیاز بالا

۲-۱-۷- توزیع میانگین امتیاز برای کتاب‌ها

تقریبا امتیاز بیشتر کتاب‌ها در محدوده ۳.۷ تا ۴.۳ است و کتاب‌های با امتیاز کامل ۵ کمیاب‌اند.



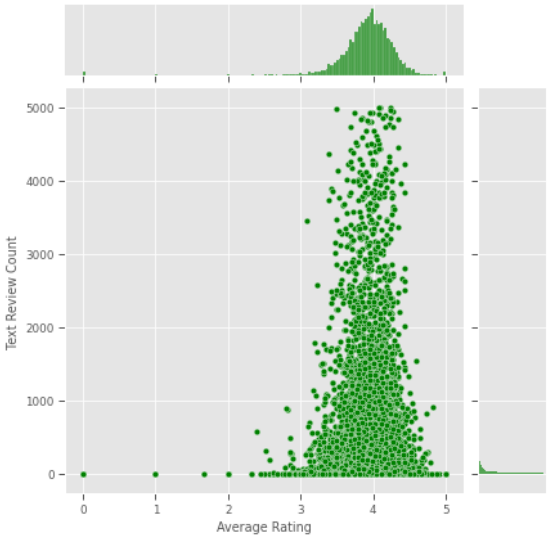
شکل ۱۰- توزیع امتیازات کتاب‌ها



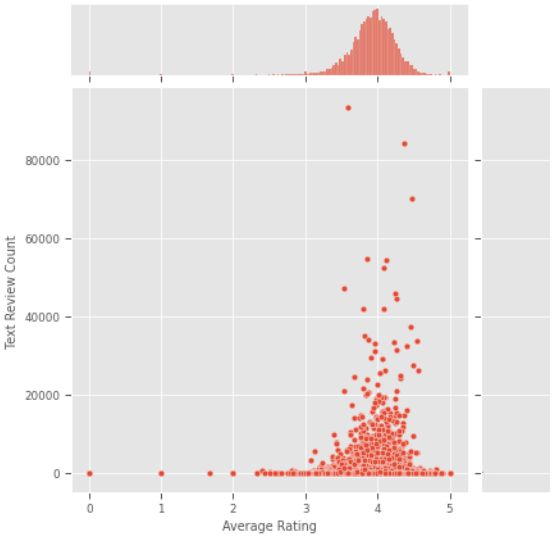
شکل ۱۱- نمودار دایره‌ای درصد امتیاز کتاب‌ها

۲-۱-۸- رابطه بین میانگین امتیازات و تعداد انتقادات

از نمودار کلی سمت چپ می‌توان نتیجه گرفت که بیشتر امتیازات کتاب‌ها بین ۳ تا ۴ به همراه تعداد عظیمی از انتقادات که در حدود ۵۰۰۰ هستند قرار دارند. از نمودار سمت راست که جزیی تر شده می‌توان مشاهده کرد که اکثر نقدها کمتر از ۱۰۰۰ مورد است. ممکن است رابطه ای وجود داشته باشد، اما به نظر می رسد که بررسی ها در میان کتاب هایی با امتیازهای مناسب غالب است.



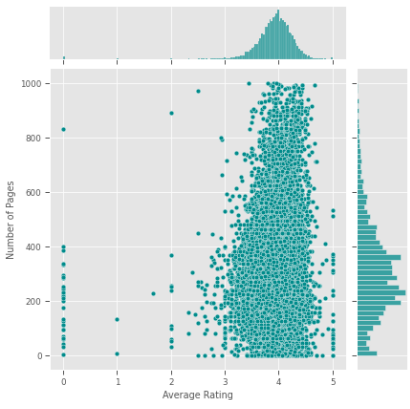
شکل ۱۲- رابطه بین تعداد انتقادات و میانگین امتیاز برای کتاب‌های با تعداد نقد کمتر از ۵۰۰۰



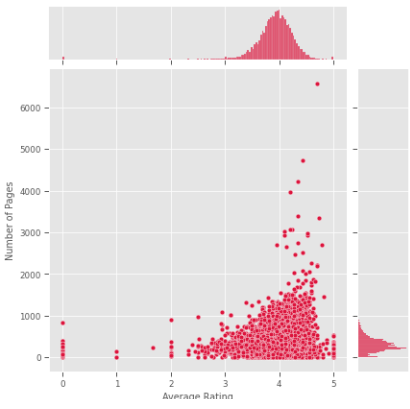
شکل ۱۳- رابطه بین تعداد انتقادات و میانگین امتیاز کتاب‌ها

۲-۱-۹- رابطه بین تعداد صفحات و میانگین امتیازات کتاب‌ها

نمودار سمت چپ به دلیل وجود انبوه موارد پرت برای کتاب‌های بالای ۱۰۰۰ صفحه، استنتاج چندان دقیقی به دست نمی‌دهد، زیرا حداکثر تراکم بین ۰-۱۰۰۰ صفحه است. از نمودار سمت راست، می توانیم استنباط کنیم که بالاترین امتیازاتی که تا به حال داده شده است، معمولاً برای کتاب هایی با محدوده صفحه ۲۰۰-۴۰۰ نزدیک به ۲۵۰ است. این می‌تواند منجر به این واقعیت شود که به نظر می‌رسد اکثر مردم کتاب هایی با صفحه متوسط را ترجیح می‌دهند. به نظر می‌رسد تعداد صفحات بالا و کتاب های ضخیم‌تر مردم را می‌ترساند!



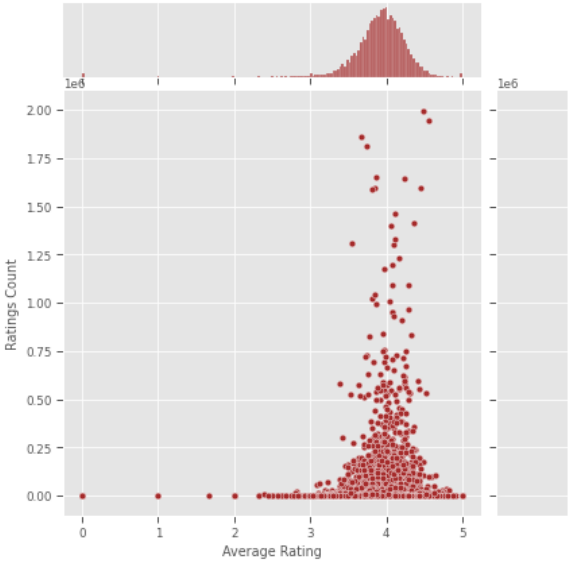
شکل ۱۴- رابطه بین تعداد صفحات و امتیاز برای کتاب‌های زیر ۱۰۰۰ صفحه



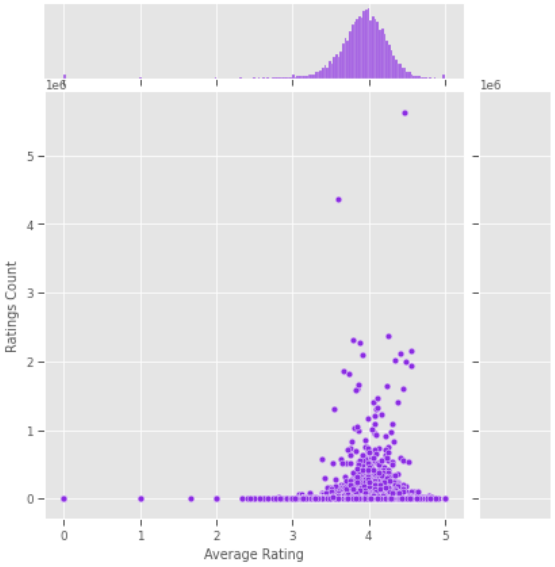
شکل ۱۵- رابطه بین تعداد صفحات و میانگین امتیاز کتاب‌ها

۲-۱-۱۰- رابطه بین میانگین امتیازات و تعدادشان برای کتاب‌ها

به نظر می‌رسد که در نمودار چپ برخی موارد پرت وجود دارد. برای چشم انداز بهتر، تعداد رتبه های حدود ۲۰۰۰۰۰۰ را در نمودار راست در نظر می‌گیریم. از نمودار، می‌توانیم ببینیم که امکان دارد یک رابطه بالقوه بین میانگین امتیازات و تعدادشان وجود داشته باشد. با افزایش تعداد امتیازات، به نظر می‌رسد که امتیاز کتاب‌ها به سمت ۴ کاهش می‌یابد. به نظر می‌رسد در حالی که تعداد امتیازات در حال کاهش است، میانگین امتیاز کم می‌شود.



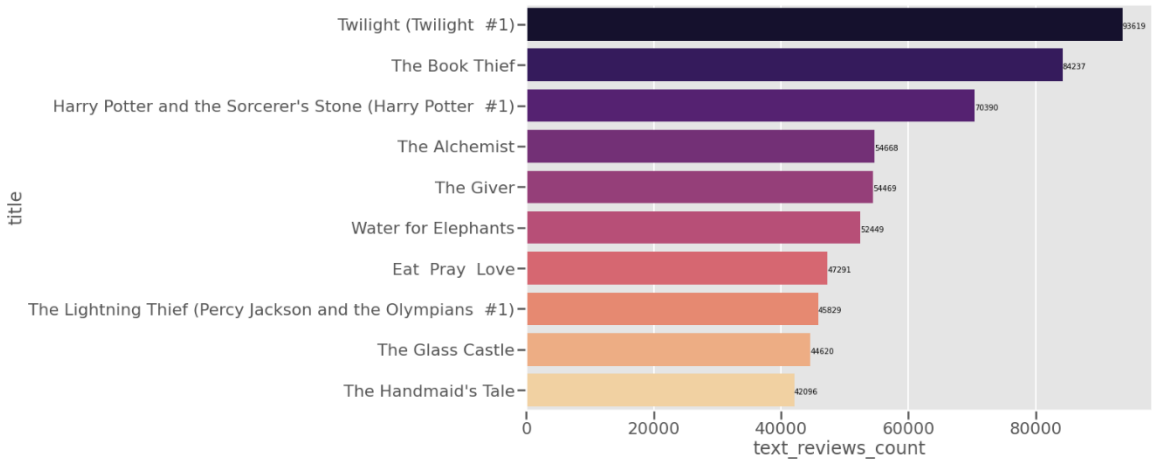
شکل ۱۶- رابطه بین میانگین امتیازات و تعدادشان برای تعداد امتیاز کمتر از ۲۰۰۰۰۰۰



شکل ۱۷- رابطه بین میانگین امتیازات و تعدادشان

۲-۱-۱۱- کتاب‌های با بیشترین تعداد انتقادات

از تمام استنباط های فوق، اساساً می‌توانیم تصمیم بگیریم که اگرچه انتقادات مهم هستند، اما نمی‌تواند رابطه خاصی بین آن‌ها و امتیازات کتاب‌ها وجود داشته باشد. در شکل زیر مشاهده می‌شود کتاب "Twilight #1" بیشترین تعداد نقد را داشته است.



شکل ۱۸- عنوان کتاب‌های با بیشترین تعداد انتقادات

۲-۱-۱۲- توضیحات اتریبیوت‌ها

در انتهای بخش تشریح داده می‌پردازیم به توضیح اینکه هر کدام از اتریبیوت‌ها چه معنایی دارند:

- Book ID: شماره شناسایی مختص هرکدام از کتاب‌ها
- Title: عنوان کتاب
- Authors: نویسنده کتاب
- Average_Rating: متوسط امتیاز برای هر کتاب
- Isbn: حاوی اطلاعات یک کتاب مانند شماره نسخه و انتشارات آن کتاب است.
- Isbn 13: فرمت جدید Isbn که در سال ۲۰۰۷ ایجاد شد و حاوی ۱۳ رقم است.
- Language_Code: زبان نوشته شده هر کتاب
- Num_Pages: تعداد صفحات هر کتاب
- Ratings_Count: تعداد امتیازات برای هر کتاب
- Text_Reviews_Counts: تعداد انتقادات درج شده از سوی کاربران

۲-۲- پیش پردازش داده‌ها

۲-۲-۱- بررسی اولیه و حذف اتریبیوت‌های اضافی

در این مرحله متوجه می‌شویم که سه ستون "BookID","Isbn","Isbn13" اضافی هستند و در مدل‌سازی ما کاربردی ندارند پس حذفشان می‌کنیم.

```
columns_to_drop = ['bookID', 'isbn', 'isbn13']
df.drop(columns=columns_to_drop, inplace=True)
df
```

bookID	title	authors	average_rating	language_code	# num_pages	ratings_count	text_reviews_count
1	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling	4.56	eng	652	1944099	26249
2	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling	4.49	eng	870	1996446	27613
3	Harry Potter and the Sorcerer's Stone (Harry P...	J.K. Rowling	4.47	eng	320	5629932	70390
4	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.41	eng	352	6267	272
5	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	4.55	eng	435	2149872	33964
...	...	...	...	...	...	...	...
47699	M Is for Magic	Neil Gaiman-Teddy Kristiansen	3.82	eng	260	11317	1060
47700	Black Orchid	Neil Gaiman-Dave McKean	3.72	eng	160	8710	361
47701	InterWorld (InterWorld #1)	Neil Gaiman-Michael Reaves	3.53	en-US	239	14334	1485
47708	The Faeries' Oracle	Brian Froud-Jessica Macbeth	4.43	eng	224	1550	38
47709	The World of The Dark Crystal	Brian Froud	4.29	eng	132	3572	33

13714 rows × 7 columns

شکل ۱۹- اتریبیوت‌های نهایی جهت مدل‌سازی

۲-۲-۲- بررسی داده‌های Null و Duplicate

پس از بررسی متوجه می‌شویم دیتاست ما هیچ داده Null و Duplicate ندارد.

Nulls & Duplicates

```
[ ] df.isnull().sum()
```

```
title               0
authors             0
average_rating      0
language_code       0
# num_pages         0
ratings_count       0
text_reviews_count  0
dtype: int64
```

```
[ ] df[df.duplicated()]
```

bookID	title	authors	average_rating	language_code	# num_pages	ratings_count	text_reviews_count

شکل ۲۰- بررسی داده‌های Null و Duplicate

۲-۲-۳- استانداردسازی داده‌ها

دو ستون"Ratings_Count","Text_Reviews_Count" که داده‌های عددی و پراکنده ای دارند را استانداردسازی می‌کنیم.

```
from sklearn.preprocessing import scale
cols = ['ratings_count', 'text_reviews_count']
df[cols] = scale(df[cols])
df
```

bookID	title	authors	average_rating	language_code	# num_pages	ratings_count	text_reviews_count
1	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling	4.56	eng	652	17.502416	10.478494
3	Harry Potter and the Sorcerer's Stone (Harry P...	J.K. Rowling	4.47	eng	320	50.974817	28.446467
4	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.41	eng	352	-0.095747	-0.095666
5	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	4.55	eng	435	19.371116	13.618951
9	Unauthorized Harry Potter Book Seven News: "Ha...	W. Frederick Zimmerman	3.69	en-US	152	-0.152497	-0.205979
...	...	...	...	...	...	...	...
47699	M Is for Magic	Neil Gaiman-Teddy Kristiansen	3.82	eng	260	-0.049886	0.225096
47700	Black Orchid	Neil Gaiman-Dave McKean	3.72	eng	160	-0.073581	-0.059438
47701	InterWorld (InterWorld #1)	Neil Gaiman-Michael Reaves	3.53	en-US	239	-0.022488	0.398096
47708	The Faeries' Oracle	Brian Froud-Jessica Macbeth	4.43	eng	224	-0.138584	-0.190918
47709	The World of The Dark Crystal	Brian Froud	4.29	eng	132	-0.120222	-0.192953

13610 rows × 7 columns

✓ 0s completed at 7:01 PM

شکل ۲۱- استانداردسازی داده‌ها

۲-۲-۴- حذف Outliers

پس از بررسی متوجه می‌شویم تعداد ۱۱۴ عدد Outlier در دیتاست وجود دارد. پس آن‌ها را حذف می‌کنیم.

```
[ ] outliers = detect_outliers()
    df.drop(index=outliers, inplace=True)
    f'outliers droped - {len(outliers)}'
```

```
'outliers droped - 104'
```

```
[ ] df.shape
```

```
(13610, 7)
```

شکل ۲۲- حذف Outliers

۲-۳- تقسیم دیتاست به دو قسمت Train و Test

مدل‌سازی ما خوشه‌بندی یا همان Clustering است و چون از نوع Unsupervised می‌باشد نیازی به تقسیم داده‌ها به دو بخش Train و Test وجود ندارد ولی ما برای اینکه بتوانیم نتیجه خوشه‌بندی را تایید کنیم دیتاست را به دو بخش به صورت ۲۰ درصد داده‌ها Test و ۸۰ درصد داده‌ها Train تقسیم می‌کنیم و برای هر دو بخش خوشه‌بندی را انجام می‌دهیم و نتیجه هر دو مدل‌سازی را با یکدیگر تطبیق می‌دهیم تا صحّت خوشه‌بندی تایید شود.

```
[ ]  df_train,df_test=train_test_split(df,test_size=0.2)

[ ]  df_train.shape

     (10888, 7)

[ ]  df_test.shape

     (2722, 7)
```

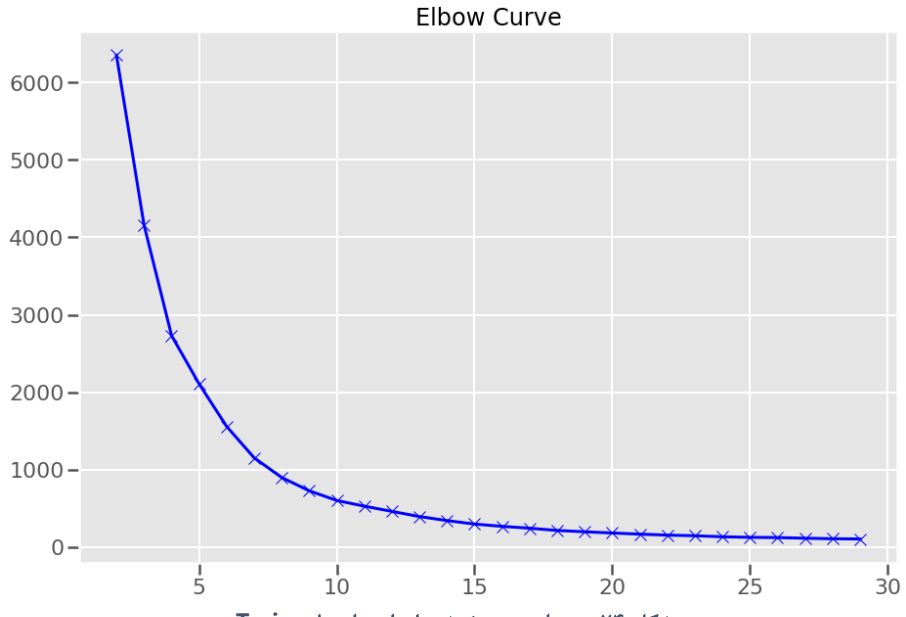
شکل ۲۳- تقسیم دیتاست به دو قسمت Train و Test

۲-۴- مدل‌سازی خوشه‌بندی

در نمودار خوشه‌ها تا انتها، محور افقی میانگین امتیاز هر کتاب و محور عمودی تعداد کل امتیازات هر کتاب می‌باشد.

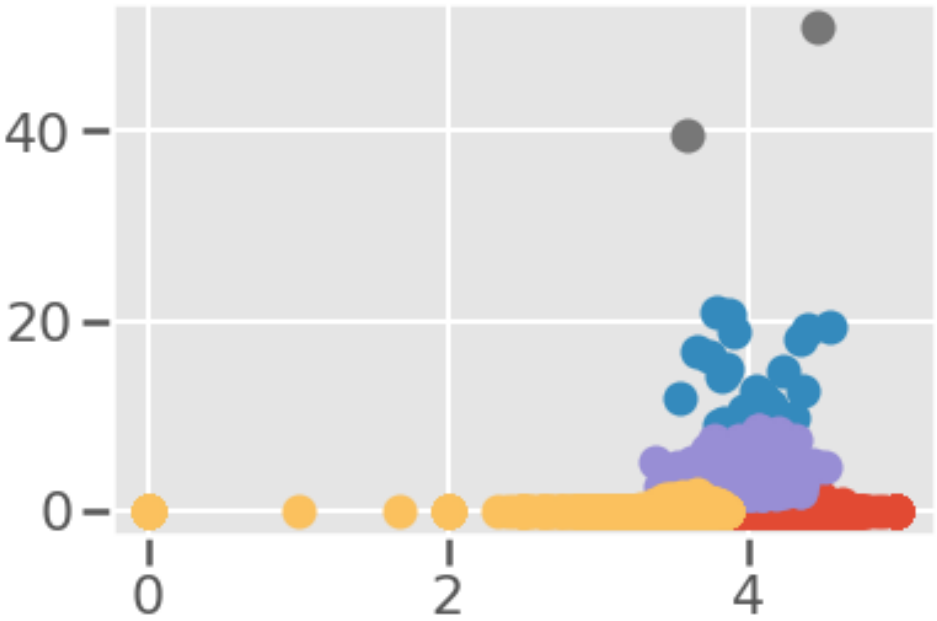
۲-۴-۱- الگوریتم K-Means بر روی داده‌های Train

با استفاده از روش K_Means خوشه‌بندی را انجام می‌دهیم و هدف این است تا گروه‌هایی از داده‌های مشابه هم را پیدا کنیم. سعی بر این داریم تا رابطه یا گروه‌هایی بین دو اتریبیوت میانگین امتیازات و تعداد امتیازات پیدا کنیم. ابتدا با استفاده از روش Elbow-Curve تعداد بهینه خوشه‌ها را پیدا می‌کنیم که مطابق شکل این تعداد بهینه برابر ۵ است.



شکل ۲۴- تعداد بهینه خوشه‌ها برای داده‌های Train

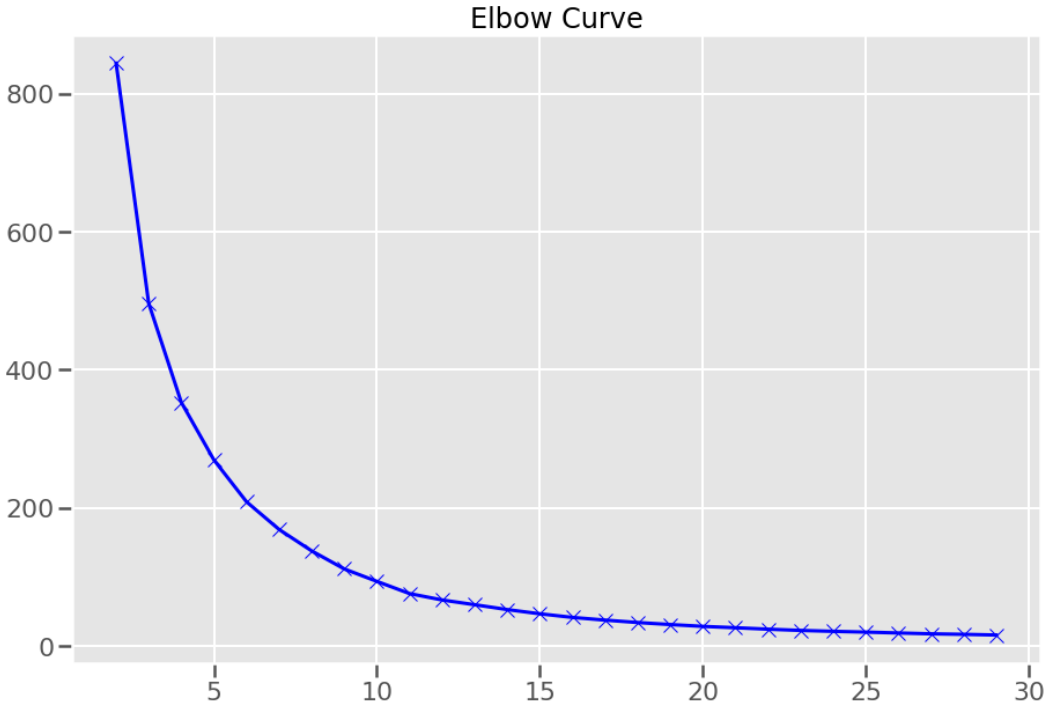
پس از انجام خوشه‌بندی با تعداد ۵ خوشه به نتیجه‌ی زیر دست می‌یابیم.



شکل ۲۵- K-Means بر روی داده‌های Train

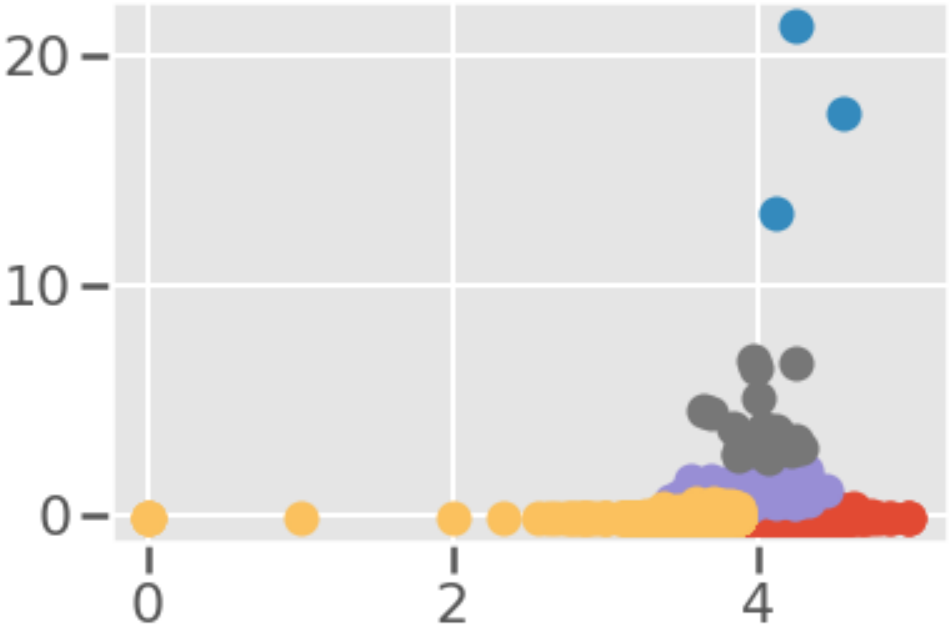
۲-۴-۲- الگوریتم K-Means بر روی داده‌های Test

ابتدا با استفاده از روش Elbow-Curve تعداد بهینه خوشه‌ها را پیدا می‌کنیم که مطابق شکل این تعداد بهینه برابر ۵ است.



شکل ۲۶- تعداد بهینه خوشه‌ها برای داده‌های Test

پس از اجرای الگوریتم بر روی داده‌های تست خوشه‌بندی داده‌ها به شکل زیر قابل مشاهده است.

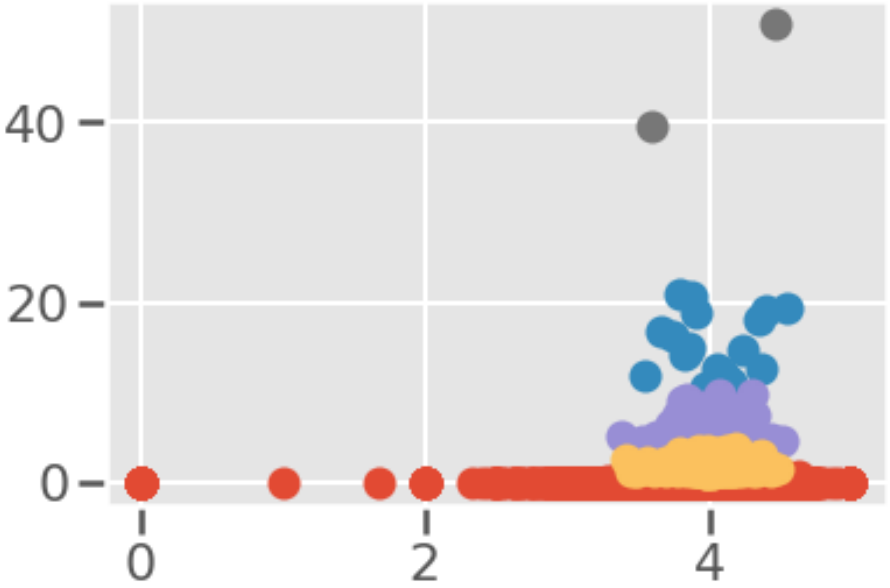


شکل ۲۷- K-Means بر روی داده‌های Test

با توجه به اینکه تعداد بهینه خوشه ها را در الگوریتم K-Means ۵ عدد بدست آورده‌ایم، جهت مقایسه و تحلیل بهتر بین خروجی الگوریتم‌های متفاوت، تعداد کلاستر برای فیت کردن سایر مدل‌های خوشه‌بندی را نیز همان عدد ۵ در نظر می‌گیریم.

۲-۴-۳- الگوریتم Agglomerative بر روی داده‌های Train

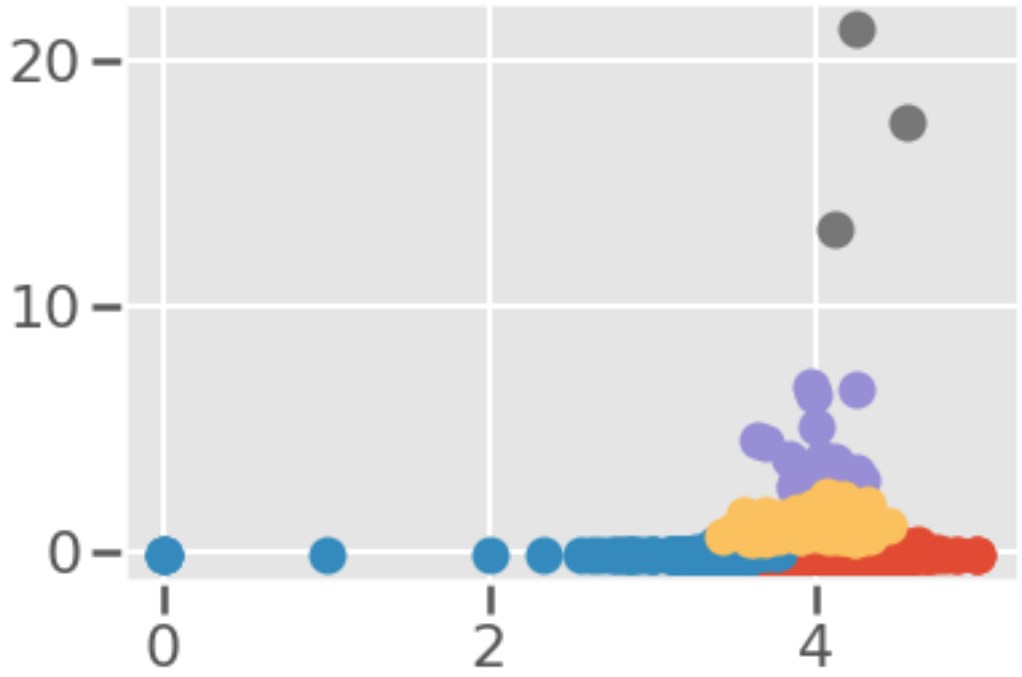
خوشه‌های حاصل از این الگوریتم بر روی داده‌های Train به شکل زیر می‌باشد.



شکل ۲۸- Agglomerative بر روی داده‌های Train

۲-۴-۴- الگوریتم Agglomerative بر روی داده‌های Test

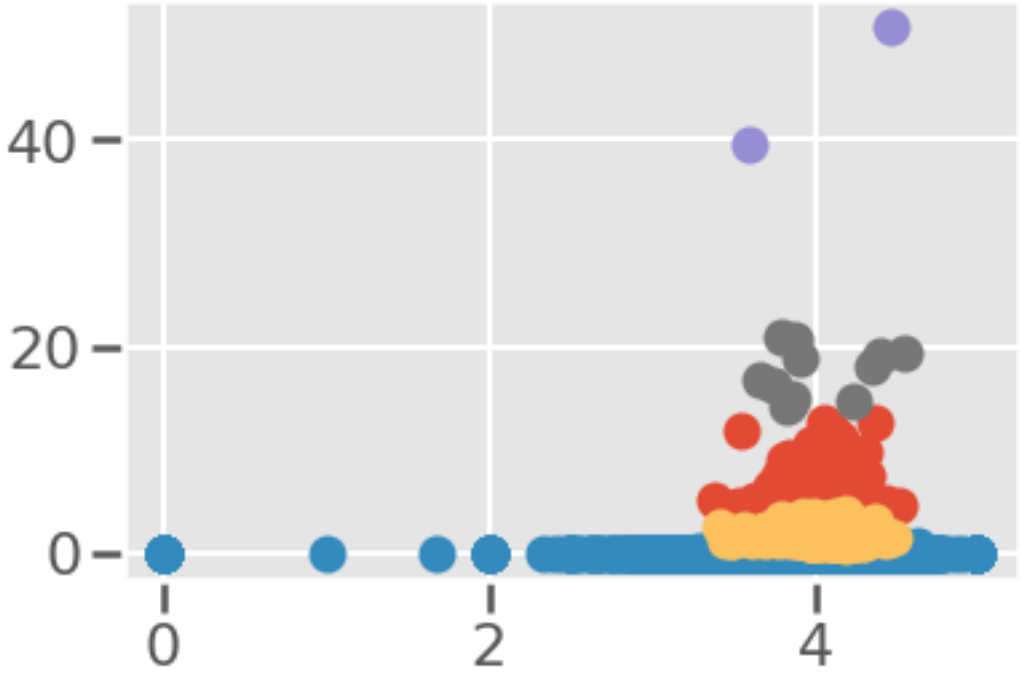
خوشه‌های حاصل از این الگوریتم بر روی داده‌های Test به شکل زیر می‌باشد.



شکل ۲۹- Agglomerative بر روی داده‌های Test

۲-۴-۵- الگوریتم BIRCH بر روی داده‌های Train

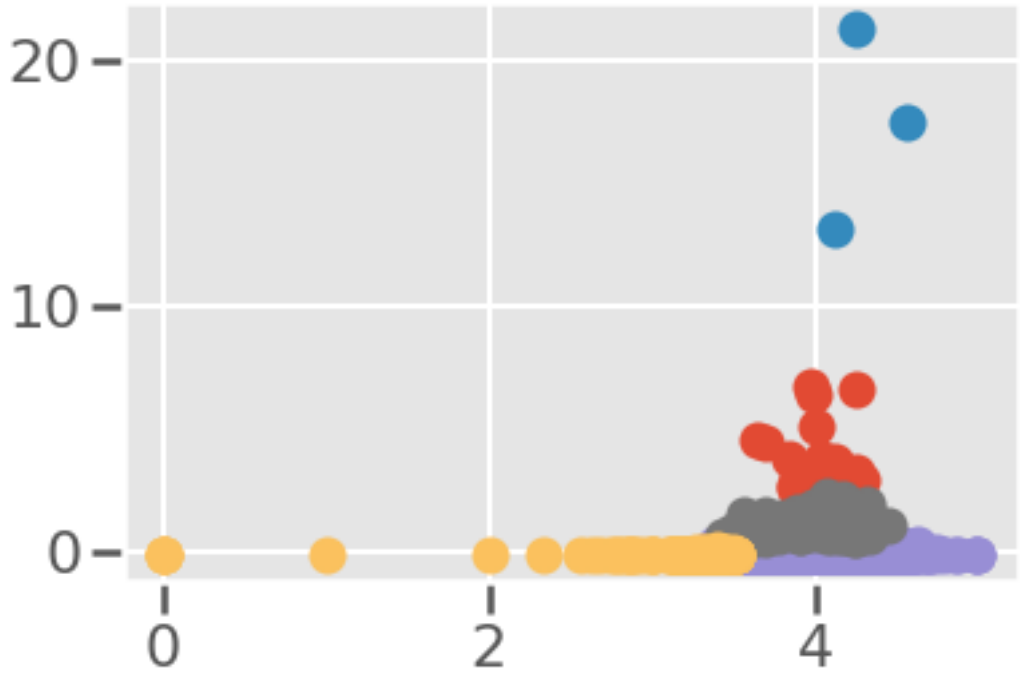
خوشه‌های حاصل از این الگوریتم بر روی داده‌های Train به شکل زیر می‌باشد.



شکل ۳۰- BIRCH بر روی داده‌های Train

۲-۴-۶- الگوریتم BIRCH بر روی داده‌های Test

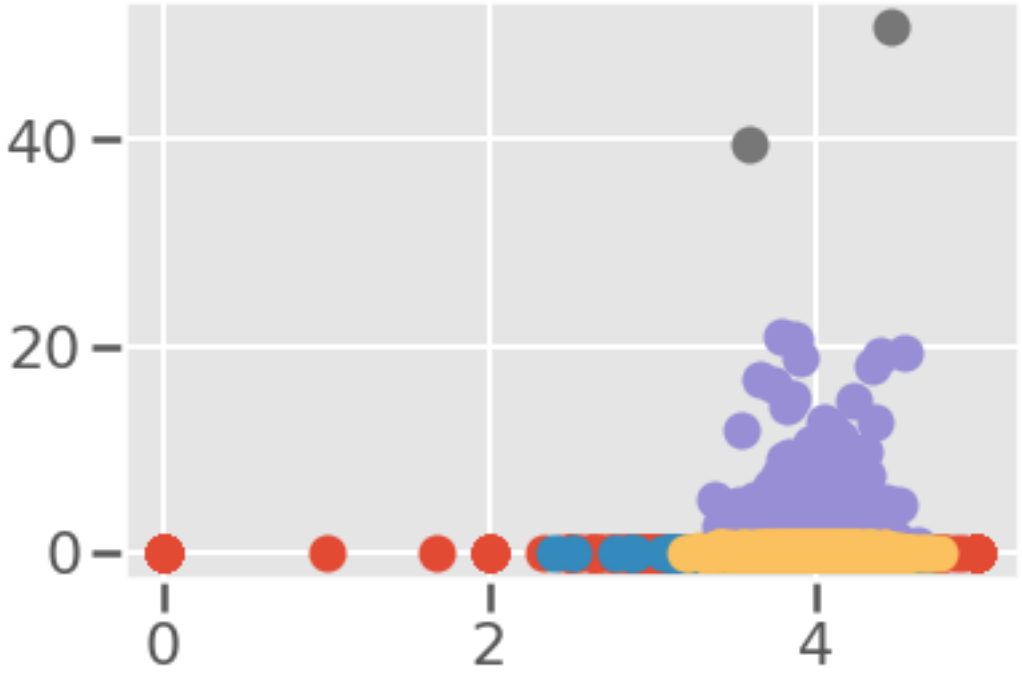
خوشه‌های حاصل از این الگوریتم بر روی داده‌های Test به شکل زیر می‌باشد.



شکل ۳۱- BIRCH بر روی داده‌های Test

۲-۴-۷- الگوریتم Gaussian Mixture بر روی داده‌های Train

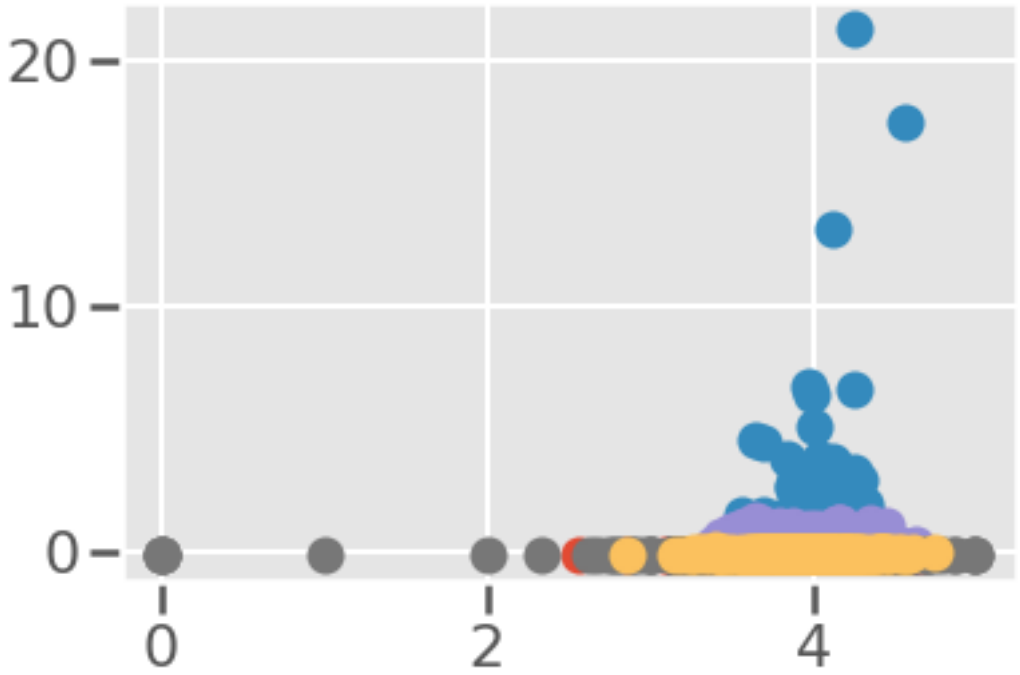
خوشه‌های حاصل از این الگوریتم بر روی داده‌های Train به شکل زیر می‌باشد.



شکل ۳۲- Gaussian Mixture بر روی داده‌های Train

۲-۴-۸- الگوریتم Gaussian Mixture بر روی داده‌های Test

خوشه‌های حاصل از این الگوریتم بر روی داده‌های Test به شکل زیر می‌باشد.



شکل ۳۳- Gaussian Mixture بر روی داده‌های Test

۲-۵- موتور توصیه کتاب

۲-۵-۱- مدل‌سازی موتور توصیه کتاب

با مشاهده خوشه‌بندی، می‌توانیم استنباط کنیم که با توجه رابطه بین میانگین امتیازات و تعدادشان، یک‌سری توصیه‌ها در رابطه با کتاب‌ها وجود دارد.

با در نظر گرفتن Ratings_Distribution(یک ستون طبقه بندی جدید ایجاد شده)، سیستم توصیه با الگوریتم K Nearest Neighbors کار می‌کند.

بر اساس کتابی که کاربر وارد کرده است، نزدیک‌ترین همسایگان به آن به عنوان کتاب‌هایی طبقه‌بندی می‌شوند که ممکن است کاربر دوست داشته باشد.

KNN برای مسائل طبقه‌بندی و رگرسیون استفاده می‌شود. در مسائل طبقه‌بندی برای پیش‌بینی برچسب یک نمونه، ابتدا k نزدیک‌ترین نمونه به مورد داده شده را بر اساس متریک فاصله پیدا می‌کنیم و بر اساس طرح رأی اکثریت یا رأی اکثریت وزنی (همسایه‌هایی که نزدیک‌تر هستند وزن بیشتری دارند) برچسب‌ها را پیش‌بینی می‌کنیم.

در چنین شرایطی، یادگیری بدون نظارت انجام می‌شود و همسایگان مشابه توصیه می‌شوند. برای لیست داده شده، اگر توصیه هایی را برای کتاب "The Catcher in the Rye" بپرسیم، پنج کتاب مرتبط با آن ظاهر می‌شود.

جدول ویژگی‌های کتاب، بر اساس توزیع امتیازات، که کتاب‌ها را در مقیاس امتیازات به شکل زیر طبقه‌بندی می‌کند، ایجاد می‌شود:

- بین ۰ و ۱
- بین ۱ و ۲
- بین ۲ و ۳
- بین ۳ و ۴
- بین ۴ و ۵

به طور کلی، توصیه‌ها میانگین امتیازات و تعدادشان را برای درخواست وارد شده در نظر می‌گیرند.

bookID	Between 0 and 1	Between 1 and 2	Between 2 and 3	Between 3 and 4	Between 4 and 5	average_rating	ratings_count
1	0	0	0	0	1	4.56	17.502416
3	0	0	0	0	1	4.47	50.974817
4	0	0	0	0	1	4.41	-0.095747
5	0	0	0	0	1	4.55	19.371116
9	0	0	0	1	0	3.69	-0.152497

شکل ۳۴- جدول ویژگی‌های کتاب برای موتور توصیه

مقیاس‌کننده Min-Max برای کاهش سوگیری استفاده می‌شود که ممکن است به دلیل داشتن تعداد زیادی ویژگی در برخی کتاب‌ها وجود داشته باشد، اما بقیه دارای ویژگی‌های کمتری باشند. مقیاس‌کننده Min-Max میانه را برای همه آنها پیدا و آن را برابر می‌کند.

```
min_max_scaler = MinMaxScaler()
books_features = min_max_scaler.fit_transform(books_features)
```

شکل ۳۵- مقیاس‌کننده Min-Max

حال یک‌سری توابع ایجاد می‌کنیم تا برای پیدا کردن نام کتاب‌ها کمکمان کنند:

- Get Index from Name: شماره Index کتاب را به ازای نام آن به ما می‌دهد.
- Get ID from Partial Name: شماره ID کتاب را به ازای نام کامل آن یا بخشی از نام آن به ما می‌دهد.
- Print Similar Books: با استفاده از نزدیک‌ترین همسایه و وارد کردن اسم یا ID کتاب، به ما ۵ کتاب مشابه کتاب ورودی را نشان می‌دهد.

```
def get_index_from_name(name):
    return df[df["title"]==name].index.tolist()[0]

all_books_names = list(df.title.values)

def get_id_from_partial_name(partial):
    for name in all_books_names:
        if partial in name:
            print(name,all_books_names.index(name))

def print_similar_books(query=None,id=None):
    if id:
        for id in indices[id][1:]:
            print(df.iloc[id]["title"])
    if query:
        found_id = get_index_from_name(query)
        for id in indices[found_id][1:]:
            print(df.iloc[id]["title"])
```

شکل ۳۶- توابع مورد استفاده در موتور توصیه کتاب

۲-۵-۲- مثال‌های موتور توصیه کتاب

در ابتدا دستور به پرینت کردن کتاب‌های مشابه کتاب‌های زیر را می‌دهیم:

- “The Catcher in the Rye”
- “The Hobbit or There and Back Again”

که خروجی زیر را دریافت می‌کنیم.

```
print_similar_books("The Catcher in the Rye")
```

```
Bloody River Blues (John Pellam  #2)
Book of Dreams
Thumbsucker
Americana
Don't Look Down
```

شکل ۳۷- مثال شماره ۱ از موتور توصیه کتاب

```
print_similar_books("The Hobbit or There and Back Again")
```

```
All the Sad Young Men (Works of F. Scott Fitzgerald)
The Laughing Jesus: Religious Lies and Gnostic Wisdom
Manna from Heaven
A Short History of World War I
The Parrot's Lament  and Other True Tales of Animal Intrigue  Intelligence  and Ingenuity
```

شکل ۳۸- مثال شماره ۲ از موتور توصیه کتاب

سپس دستور به خروجی گرفتن ID کتاب بر اساس بخشی از اسمش(“Harry Potter and the”) را می‌دهیم که به شکل زیر است.

```
get_id_from_partial_name("Harry Potter and the ")
```

```
Harry Potter and the Half-Blood Prince (Harry Potter  #6) 0
Harry Potter and the Sorcerer's Stone (Harry Potter  #1) 1
Harry Potter and the Chamber of Secrets (Harry Potter  #2) 2
Harry Potter and the Prisoner of Azkaban (Harry Potter  #3) 3
Harry Potter and the Half-Blood Prince (Harry Potter  #6) 0
Harry Potter and the Prisoner of Azkaban (Harry Potter  #3) 3
Harry Potter and the Chamber of Secrets (Harry Potter  #2) 2
Harry Potter and the Sorcerer's Stone (Harry Potter  #1) 1
Harry Potter and the Philosopher's Stone (Harry Potter  #1) 12564
Harry Potter and the Goblet of Fire (Harry Potter  #4) 12567
```

شکل ۳۹- مثال شماره ۳ از موتور توصیه کتاب

تمامی کتاب‌های هری‌پاتر در خروجی بدست می‌آید و در نتیجه مشاهده می‌شود که موتور به درستی کار می‌کند.

در مرخله آخر دستور به پرینت کتاب‌های مشابه کتاب با ID برابر ۱ می‌دهیم که خروجی را در شکل زیر مشاهده می‌کنیم.

```
print_similar_books(id = 1) #ID for the Book 5
```

```
The Hobbit or There and Back Again
Harry Potter and the Prisoner of Azkaban (Harry Potter  #3)
Harry Potter and the Chamber of Secrets (Harry Potter  #2)
The Fellowship of the Ring (The Lord of the Rings  #1)
Harry Potter and the Half-Blood Prince (Harry Potter  #6)
```

شکل ۴۰- مثال شماره ۴ از موتور توصیه کتاب

۲-۶- نتیجه‌گیری و تحلیل

اعمال الگوریتم خوشه‌بندی بر روی انبوهی از داده‌های خام ایجاد شده برای کتب موجود که مورد بررسی منتقدان نیز قرار گرفته‌اند، می‌تواند قدم محکمی برای یاری بخشی به کسانی باشد که در سیل عظیم کتاب‌ها به دنبال موردی هستند که بیشترین تطابق با ایشان را دارد. در گذشته مدل هایی بر اساس ژانر کتاب‌ها، سال نگارش، کشور محل نگارش و ... به وجود آمده بودند اما در این پروژه این بار به سراغ ویژگی‌هایی از جمله نمرات منتقدان، تعداد نمرات، زبان کتاب و برخی ویژگی هایی که در قسمت های قبلی به آن اشاره شد رفته و براساس آن‌ها به مدلی رسیدیم و الگوریتم خوشه‌بندی را اجرا کردیم.

در این مطالعه سعی شد که تمامی مراحل مربوط به یک پروژه داده‌کاوی روی دیتاست مربوط به نمرات ثبت شده کتاب‌ها پیاده‌سازی شود.

در قسمت خوشه‌بندی از ۴ الگوریتم متفاوت استفاده کردیم. با توجه به نتایج قسمت قبل، اگر خوشه های تولید شده برای داده‌های train را با یکدیگر مقایسه کنیم، الگوریتم Agglomerative و BIRCH خوشه‌بندی نسبتا مشابهی را ارائه می‌دهند. الگوریتم K-means به خوبی داده‌ها را از یکدیگر تفکیک کرده است. الگوریتم Gaussian Mixture داده‌ها را به طور مناسب در دسته‌ها قرار نداده و دسته ها در هم تنیده شده اند.

اگر به مقایسه خوشه‌بندی داده های Test بپردازیم، باز هم متوجه عدم توزیع مناسب در خوشه‌بندی توسط الگوریتم Gaussian Mixture می‌شویم. اما این بار الگوریتم‌های Agglomerative و BIRCH بسیار نزدیک به K-means عمل کرده‌اند.

در اینجا به مقایسه خوشه های تولید شده داده های Test و Train توسط هر الگوریتم می‌پردازیم. بهترین عملکرد مربوط به الگوریتم K-means می‌باشد. خوشه های ایجاد شده در داده هایی که مقادیر نسبتا یکسانی دارند، مشابه هستند. الگوریتم های Agglomerative و Birch در قسمت Train شبیه به یکدیگر بوده و برای داده های Test هر دو مانند الگوریتم K-means عمل کرده‌اند. میزان تفاوت خوشه‌های داده های Test و Train به اندازه تغییر یک خوشه بوده است. الگوریتم Gaussian Mixture به جز در یک منطقه، مشابه عمل کرده است.

با توجه به این تفاسیر عملکرد الگوریتم k-means در توزیع خوشه‌ها و تشابه عملکرد روی دو سری داده Test و Train از مابقی الگوریتم ها بهتر بوده است.

همانطور که در این پروژه مشاهده شد، با خوشه‌بندی به ۵ دسته اصلی رسیدیم که می‌تواند هسته موتور پیشنهاد کتاب باشد. با توجه به این موضوع که علاقه به یک کتاب و نمره‌دهی به آن ممکن است برای هر فردی به ویژگی‌های متفاوتی مرتبط باشد و نتایج مختلفی در بر داشته باشد، خوشه‌بندی راه مناسبی برای کنکاش کردن این دیتاست است. همچنین استانداردسازی داده‌ها بسیار حائز اهمیت است زیرا می‌تواند منجر به افزایش چشم‌گیر کارآیی شود.

در انتها اشاره به اهمیت داده‌ها در بهبود زندگی بشر خالی از لطف نیست. داده‌ها به سرمایه شرکت‌ها بدل شده و از یک کاربرد کوچک مانند چیزی که در این پروژه دیدیم تا ابعاد بسیار بزرگ‌تر که حتی می‌تواند در بقای حیات انسان‌ها نقش داشته باشد، برای آن ها در نظر گرفته می‌شود.

فصل ۳: مراجع و منابع

- https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks?datasetId=231310&sortBy=dateRun&language=null
- https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
- https://stackoverflow.com/questions/67797103/standardizing-a-set-of-columns-in-a-pandas-dataframe-with-sklearn
- https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
- https://scikit-learn.org/stable/modules/neighbors.html
- https://machinelearningmastery.com/clustering-algorithms-with-python/