

عنوان گزارش

تحلیل توصیفی و استخراج قوانین انجمنی از داده های کتاب خوانی

نگارش: محمد حسین حیدری

چکیده گزارش

در این پروژه با استفاده از دیتاست BX-Book به تحلیل داده های مربوط به کتاب ها، نویسندگان، ناشران و مشتریان پرداخته شد. هدف اصلی این مطالعه، بررسی الگوهای نهفته در داده ها و استخراج بینش هایی در زمینه ی تعداد کتاب ها، میزان فروش، سفارش ها، امتیازدهی و ویژگی های مشتریان است.

در فاز نخست، تحلیل توصیفی و مصورسازی داده ها با نرم افزار JMP انجام شد تا تصویری روشن از توزیع و روند متغیرهای کلیدی ارائه شود. این مرحله شامل بررسی نویسندگان و ناشرانی با بیشترین تعداد کتاب و فروش، کشورهای فعال در خرید کتاب، تحلیل سال های انتشار، امتیازات کتاب ها و همچنین الگوهای مربوط به سن مشتریان بود.

در فاز دوم، از روش قوانین انجمنی (Association Rules) برای کشف روابط میان خرید کتاب ها استفاده شد. معیارهای Support، Confidence و Lift جهت ارزیابی کیفیت قوانین به کار گرفته شدند و نتایج در قالب جداول و نمودارهای مناسب ارائه گردید.

یافته ها نشان می دهند که تمرکز فروش و محبوبیت در گروه محدودی از نویسندگان و ناشران قرار دارد، برخی کشورها سهم بیشتری در سفارش کتاب دارند، و الگوهای جالبی در ترجیحات مشتریان و ارتباط میان کتاب ها قابل شناسایی است. این نتایج می توانند در بهبود تصمیم گیری های مرتبط با بازاریابی، نشر و پیشنهاددهی کتاب ها مورد استفاده قرار گیرند.

آماده‌سازی و بررسی داده‌ها

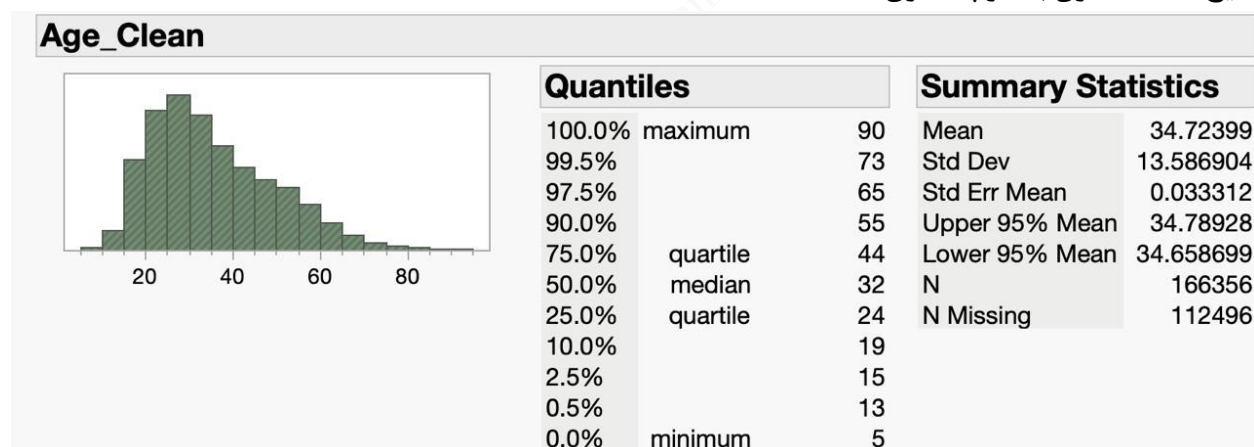
۱. جدول کاربران (Users)

ستون مربوط به کاربران شامل سه ویژگی اصلی بود: User-ID، Location و Age. در میان این ستون‌ها، بیشترین نیاز به پاک‌سازی مربوط به ستون Age بود. این ستون ترکیبی از مقادیر عددی، متنی و Null داشت. به‌طور مشخص:

- برخی ردیف‌ها حاوی متن‌هایی مانند نام کشور یا عبارت‌هایی نظیر "n/a" بودند.
- در تعدادی از رکوردها مقادیر غیرمنطقی مانند 0 یا اعداد بسیار بزرگ (مانند 2870) وجود داشت.
- بخش بزرگی از داده‌ها نیز فاقد مقدار سن بودند (Null).

برای حل این مشکل، ستون جدیدی با نام Age_Clean ایجاد شد. در این ستون فقط سن‌های معتبر در بازه‌ی ۵ تا ۹۰ سال نگهداری شدند و سایر مقادیر به صورت Missing ثبت شدند.

نتایج خلاصه آماری بعد از پاک‌سازی



استخراج نام کشور

در داده‌های مربوط به کاربران، ستون Location شامل اطلاعات کامل محل زندگی افراد بود (مانند "nyc, new york, usa"). برای تحلیل دقیق‌تر، لازم بود کشور به صورت جداگانه استخراج شود.

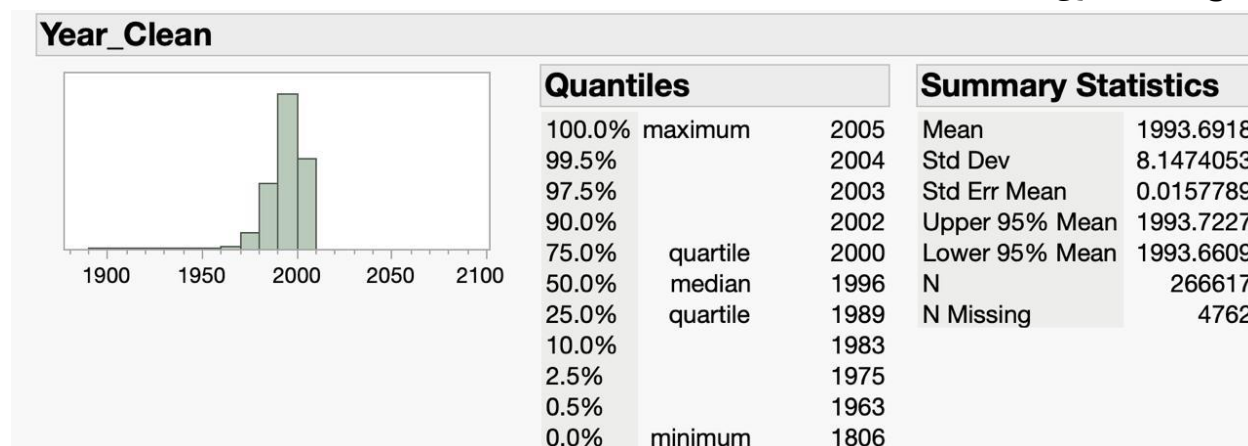
2. جدول کتاب‌ها (Books)

پاک‌سازی داده‌ها

ستون Year-Of-Publication دارای مقادیر نامعتبر زیادی بود؛ از جمله:

- مقادیر صفر (0) یا Null.
 - مقادیر غیرواقعی بزرگ‌تر از سال انتشار دیتاست (۲۰۰۵ به بعد).
 - مقادیر متنی یا نویسنده/عنوان اشتباه وارد شده در ستون سال.
- برای حل این مشکل، ستون جدیدی به نام Year_Clean ایجاد شد. در این ستون فقط سال‌های معتبر در بازه‌ی ۱۴۵۰ تا ۲۰۰۵ نگهداری شدند و سایر مقادیر به صورت Missing ثبت شدند.

نتایج خلاصه آماری Year_Clean



تفسیر

- توزیع سال انتشار نشان می‌دهد که اکثر کتاب‌ها بین سال‌های 1980 تا 2005 منتشر شده‌اند.
- تمرکز اصلی داده‌ها در دهه‌های پایانی قرن بیستم است (میانه = 1996).
- وجود تعداد کمی داده مربوط به قرن ۱۹ (مثل ۱۸۰۶) نشان می‌دهد که کتاب‌های خیلی قدیمی هم در دیتاست حضور دارند، اما سهم ناچیزی دارند.

3. جدول امتیازدهی کتاب‌ها (Book-Ratings)

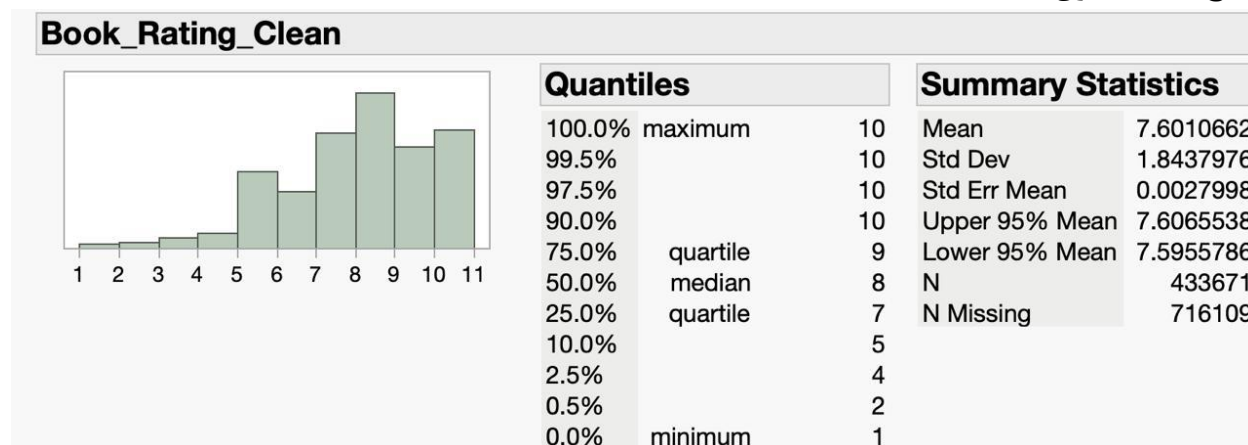
داده‌های مربوط به امتیازدهی شامل سه ستون User-ID، ISBN و Book-Rating بود. در میان این ستون‌ها، بیشترین نیاز به پاک‌سازی مربوط به ستون Book-Rating بود.

پاک‌سازی ستون Book-Rating

بررسی اولیه نشان داد که مقدار ۰ در این ستون نشان‌دهنده‌ی «ثبت کتاب بدون امتیازدهی» است و نباید به عنوان امتیاز واقعی در نظر گرفته شود. بنابراین تصمیم گرفته شد که تنها امتیازهای ۱ تا ۱۰ معتبر باشند.

برای حل این مشکل، ستون جدیدی با نام Book_Rating_Clean ایجاد شد. در این ستون فقط مقادیر معتبر (۱ تا ۱۰) نگهداری شدند و مقدار ۰ به صورت Missing ثبت گردید.

نتایج خلاصه آماری Year_Clean



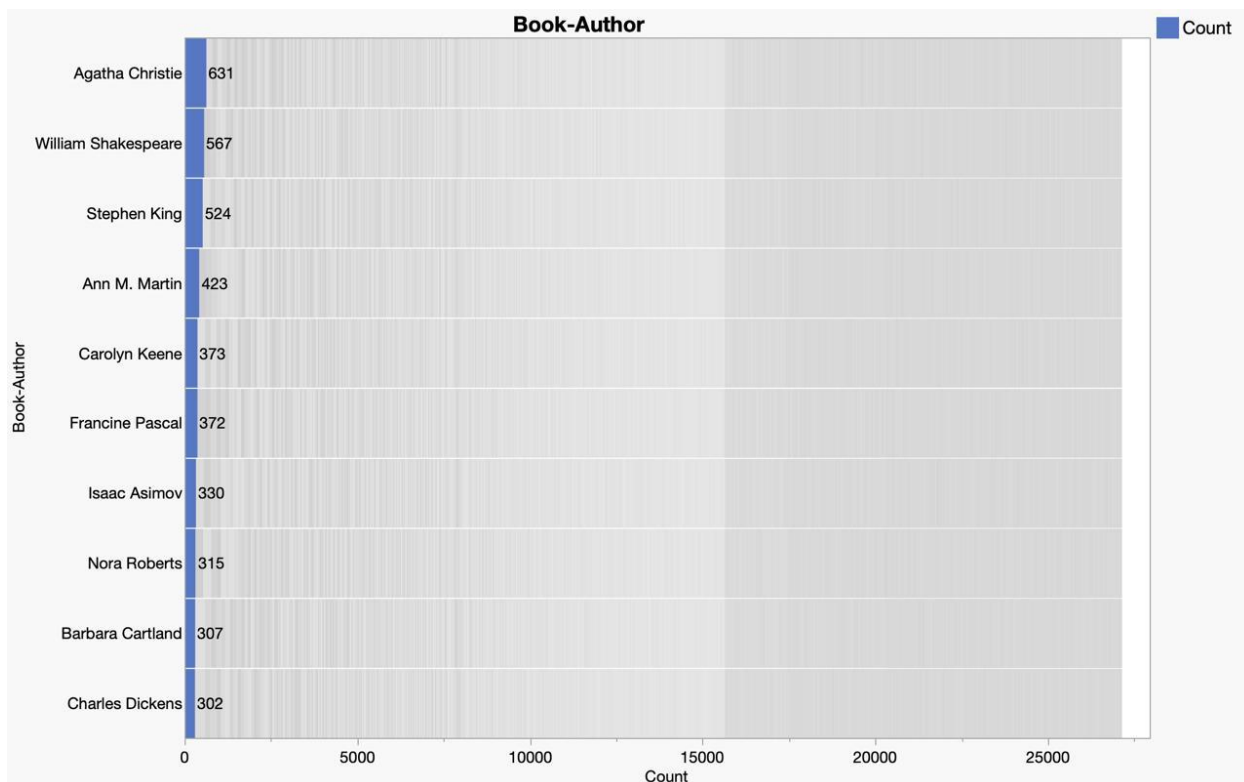
- با حذف مقادیر صفر، توزیع امتیازات واقعی مشخص شد.
- میانه برابر با ۸ است که نشان می‌دهد کاربران بیشتر به کتاب‌ها امتیازهای بالا داده‌اند.
- بیشتر از نیمی از امتیازات معتبر در بازه‌ی ۷ تا ۱۰ قرار دارند، که بیانگر رضایت کلی کاربران از کتاب‌هایی است که به آن‌ها امتیاز داده‌اند.

نکته:

«بعد از ادغام دیتاست امتیازات با دیتاست کتاب‌ها (روی ستون ISBN)، تعداد رکوردها کاهش یافت. دلیل این کاهش وجود ISBN هایی در دیتاست امتیازات بود که اطلاعات مربوط به آن‌ها در دیتاست کتاب‌ها موجود نبود.»

نویسندگان که بیشترین تعداد کتاب را نوشته اند

برای شناسایی نویسندگانی که بیشترین تعداد کتاب را نوشته‌اند، از جدول کتاب‌ها (Books) استفاده شد و شمارش تعداد آثار هر نویسنده انجام گرفت.

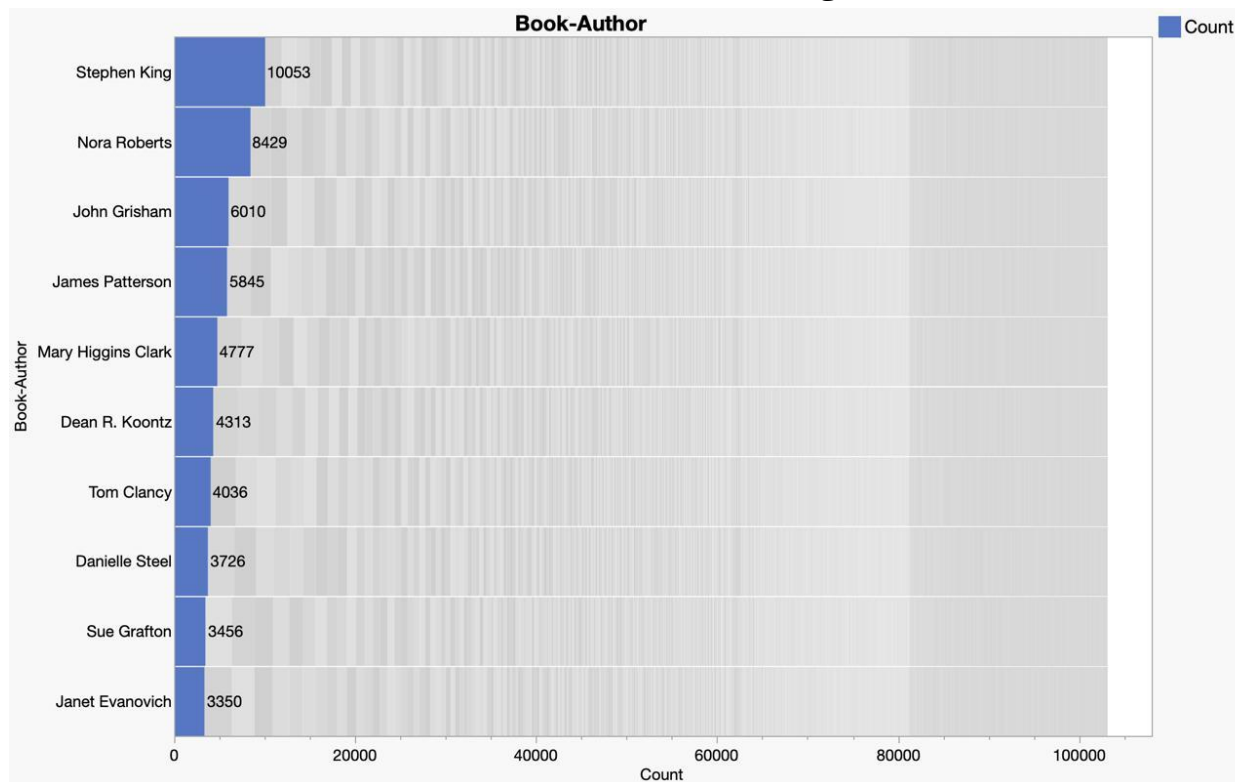


نتایج نشان داد:

- Agatha Christie با 631 کتاب در رتبه اول قرار دارد.
- پس از او، William Shakespeare با 567 کتاب و Stephen King با 524 کتاب بیشترین آثار را منتشر کرده‌اند.
- نویسندگان دیگری مانند Ann M. Martin (423)، Carolyn Keene (373) و Isaac Asimov (330) نیز در میان نویسندگان پرکار دیده می‌شوند.

نویسندگانی که بیشترین تعداد کتاب فروش رفته را داشته اند

برای این بخش، داده‌های مربوط به امتیازات (Book-Rating) و اطلاعات کتاب‌ها تجمیع شد تا نویسندگانی که بیشترین حجم فروش و استفاده را در دیتاست دارند شناسایی شوند.

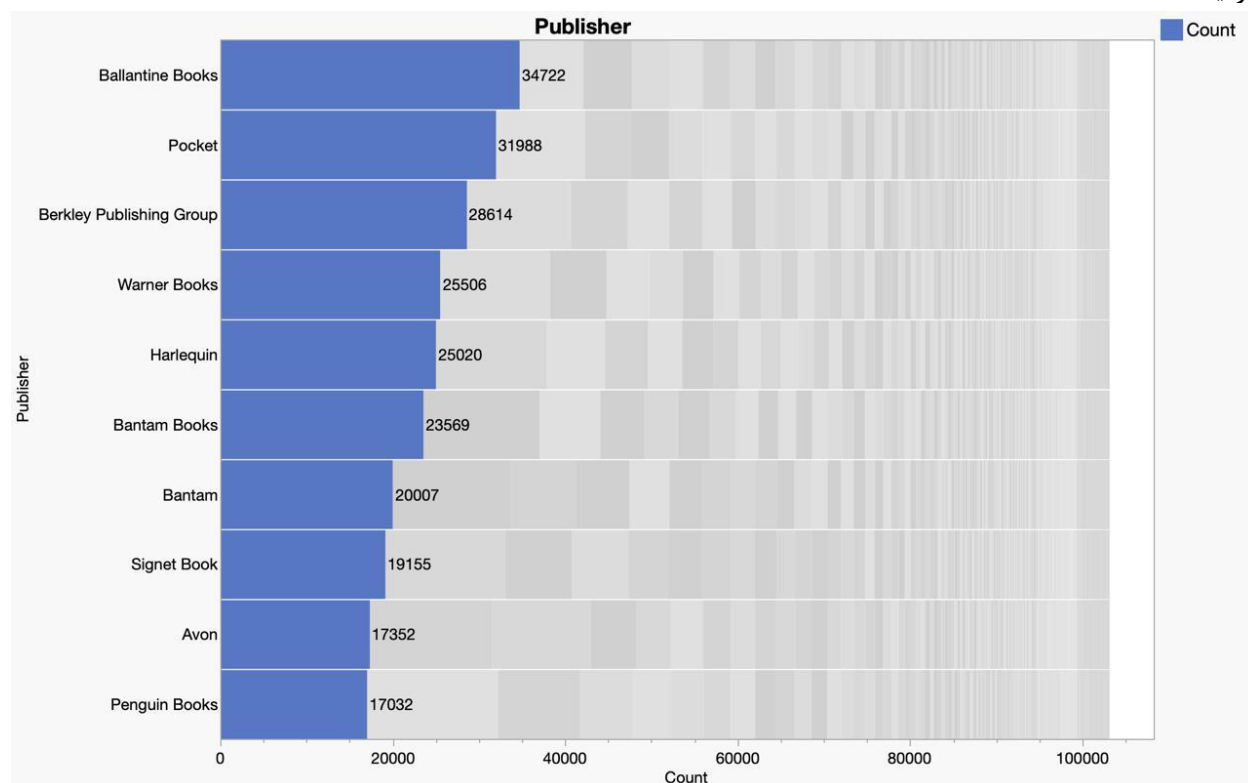


نتایج نشان داد:

- Stephen King با 10,053 مورد بیشترین سهم را در میان نویسندگان دارد.
- پس از او، Nora Roberts (8429) و John Grisham (6010) در رتبه‌های دوم و سوم قرار دارند.
- نویسندگان دیگری مانند James Patterson (5845) و Mary Higgins Clark (4777) و Dean R. Koontz (4313) نیز در جمع نویسندگان پرمخاطب قرار گرفته‌اند.

ناشرانی که بیشترین تعداد کتاب فروش رفته را داشته اند

برای بررسی عملکرد ناشران، داده‌های مربوط به ناشر هر کتاب تجمیع شد و تعداد کل فروش/امتیازات ثبت‌شده برای آن‌ها محاسبه گردید.



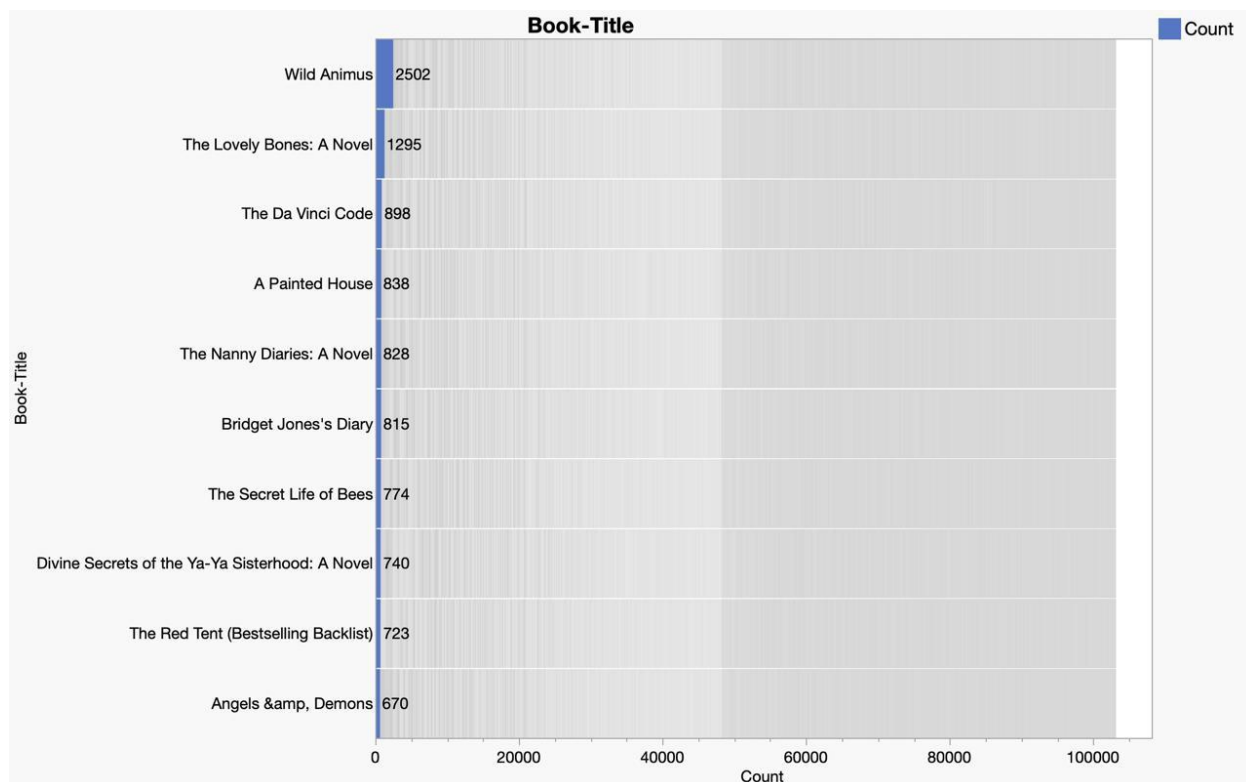
نتایج نشان داد:

- Ballantine Books با 34,722 کتاب در صدر قرار دارد.
- پس از آن، ناشران Pocket (31,988) و Berkley Publishing Group (28,614) بیشترین آثار را در بازار دارند.
- ناشران بزرگ دیگری مانند Warner Books (25,506)، Harlequin (25,020) و Bantam Books (23,569) نیز سهم قابل توجهی دارند.

کتاب‌هایی که بیشترین تعداد کتاب فروش رفته را داشته اند

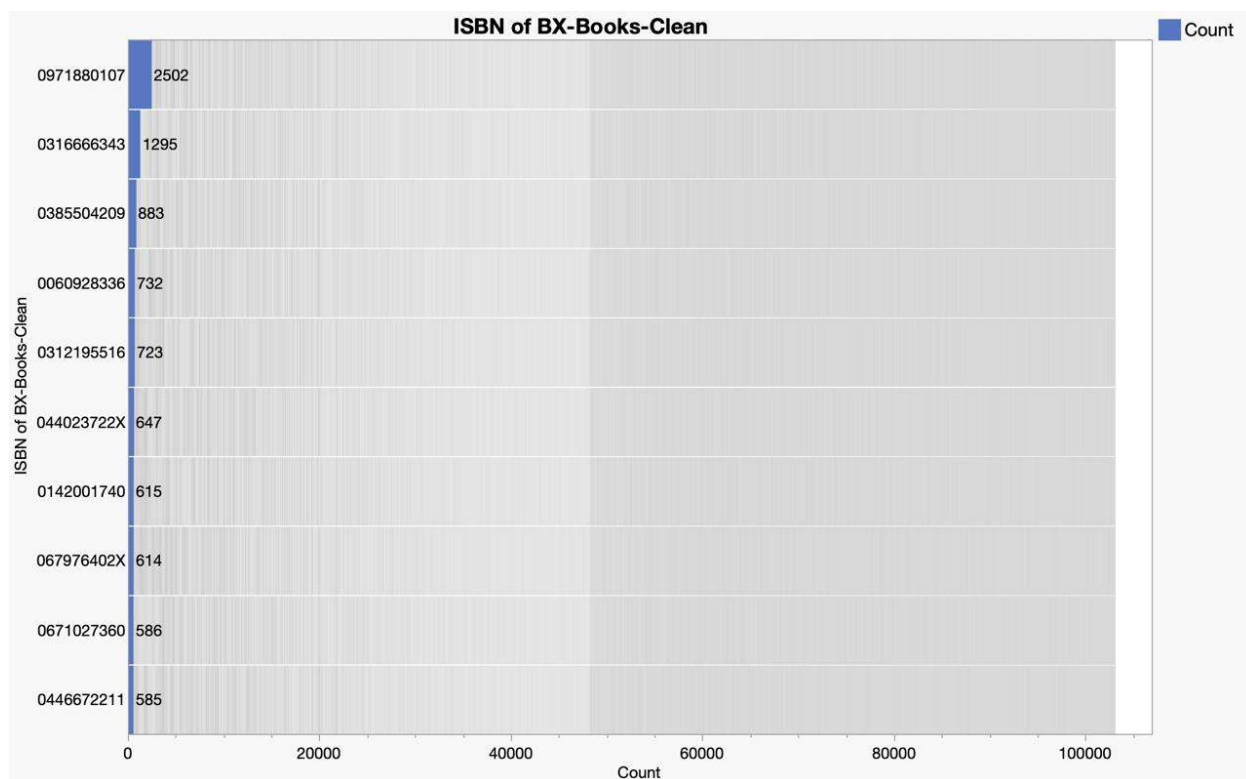
تحلیل کتاب‌ها بر اساس عنوان (Book-Title)

در ابتدا، کتاب‌هایی که بیشترین تعداد فروش/امتیاز را داشته‌اند بر اساس ستون Book-Title بررسی شدند. نتایج نشان داد که کتاب‌هایی مانند Wild Animus (2502 بار) و The Lovely Bones (1295 بار) بیشترین تکرار را دارند.



کتاب هایی که بیشترین تعداد کتاب فروش رفته را داشته اند

تحلیل کتابها بر اساس ISBN

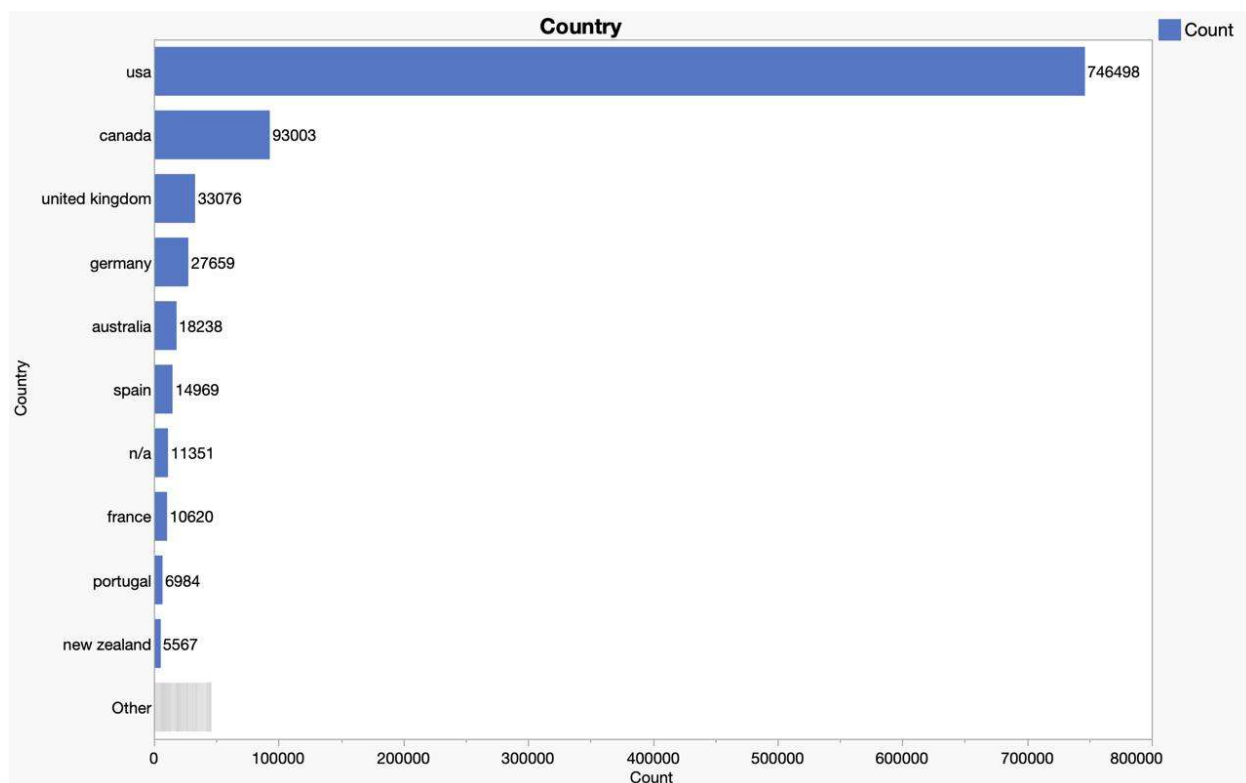


نتایج نشان داد:

- Wild Animus (ISBN: 0971880107) با 2502 بار در صدر قرار دارد.
- پس از آن، The Lovely Bones (ISBN: 0316666343) با 1295 بار و The Da Vinci Code (ISBN: 0385504209) با 883 بار در رتبه‌های بعدی قرار گرفتند.

کشور هایی که بیشترین تعداد کتاب را سفارش داده اند

با استخراج کشور از ستون Location کاربران و شمارش تعداد سفارش‌ها، نتایج زیر به دست آمد:



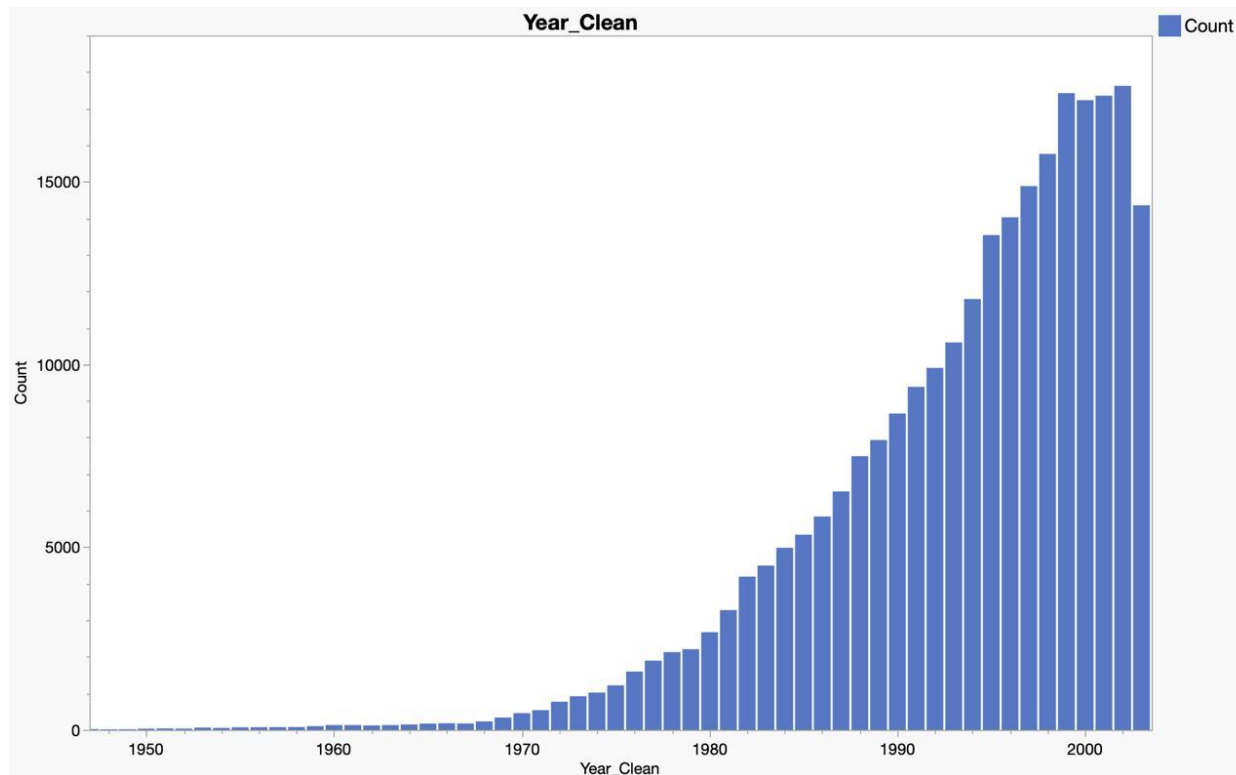
- ایالات متحده (USA) با اختلاف بسیار زیاد و 746,498 سفارش در رتبه اول قرار دارد.
- پس از آن، کانادا با 93,003 سفارش و بریتانیا (United Kingdom) با 33,076 سفارش در رتبه‌های بعدی هستند.
- کشورهای آلمان (27,659)، استرالیا (18,238) و اسپانیا (14,969) نیز جزو کشورهای پرتکرار هستند.
- سایر کشورها سهم کمتری دارند و در بخش "Other" جمع شده‌اند.

جمع‌بندی:

این نتایج نشان می‌دهد که بیشترین کاربران و سفارش‌دهندگان کتاب در این دیتاست متعلق به کشورهای انگلیسی‌زبان هستند. به ویژه USA و Canada، که تقریباً بخش عمده‌ای از کل سفارش‌ها را شامل می‌شوند.

تحلیلی روی سال انتشار کتاب

این نمودار نشون میده کتاب ها در چه سالی منتشر شدن

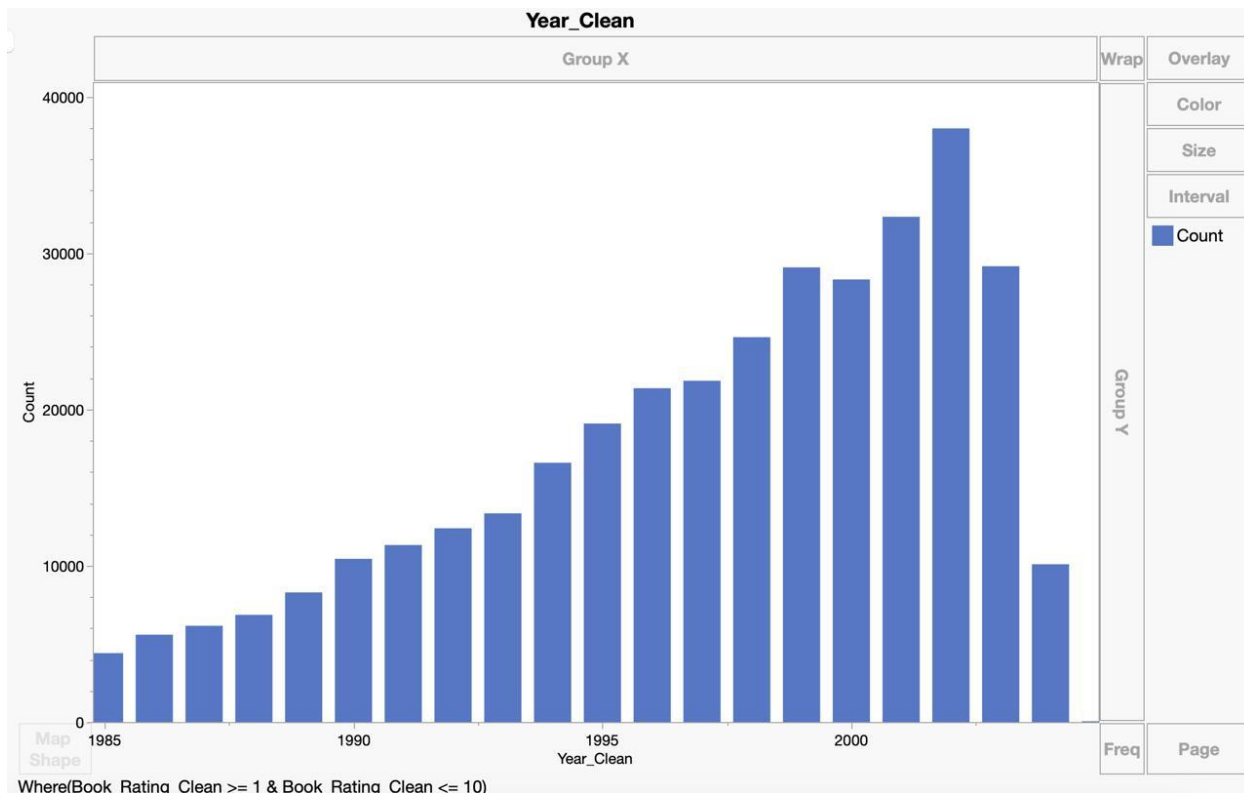


نتایج:

- تا دهه 1950 تعداد کتاب‌های موجود بسیار محدود بوده است.
- از دهه 1970 به بعد، روند انتشار کتاب‌ها رشد چشمگیری پیدا کرده
- بیشترین حجم کتاب‌ها مربوط به بازه 1990 تا 2005 است که نشان می‌دهد دیتاست تمرکز اصلی‌اش روی کتاب‌های معاصر است.

تحلیلی روی سال انتشار کتاب

این نمودار سال انتشار کتاب هایی که واقعا توسط کاربران سفارش داده شده اند و امتیاز هم داده اند



نکات مهم از نمودار:

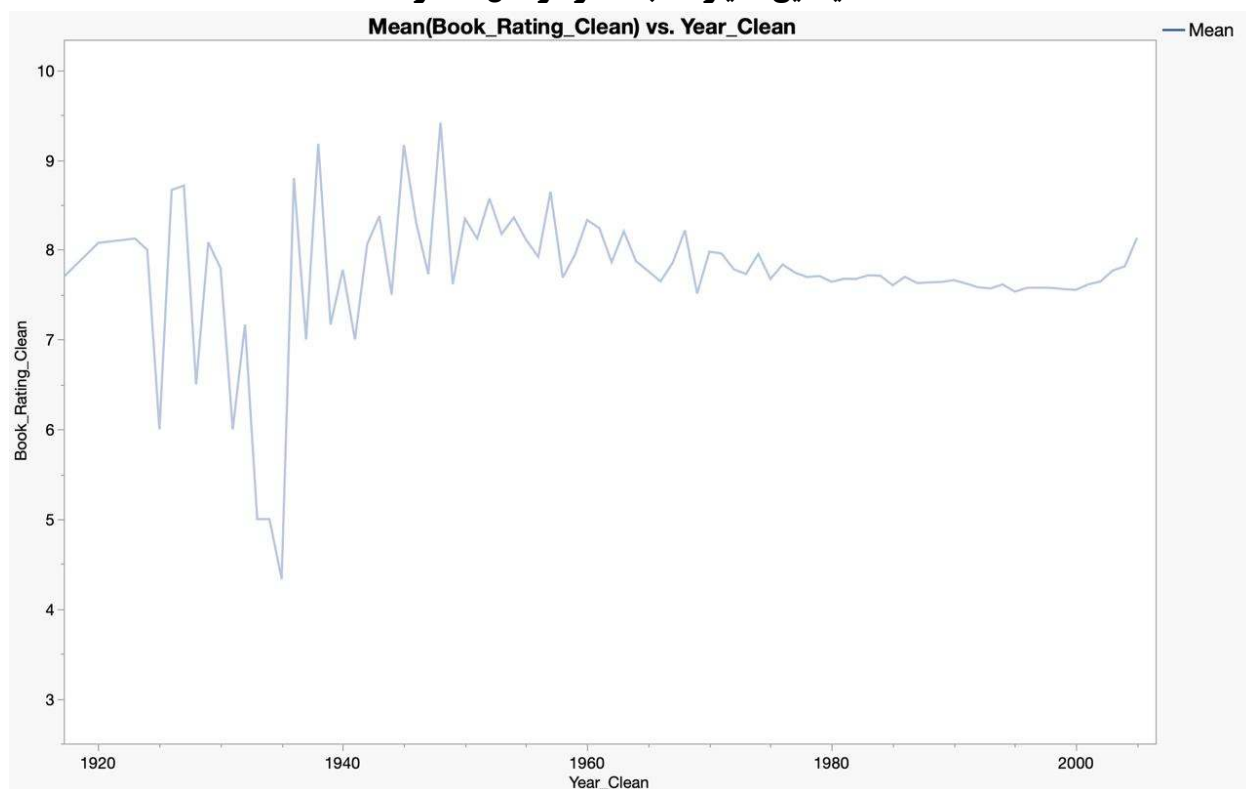
- بازه‌ی زمانی از حدود ۱۹۸۵ تا اوایل ۲۰۰۰ بیشترین حجم امتیازدهی رو داره.
- اوج امتیازدهی مربوط به کتاب‌های منتشرشده در اواخر دهه ۹۰ و اوایل دهه ۲۰۰۰ هست.
- کتاب‌های قدیمی‌تر (قبل از دهه ۸۰) خیلی کمتر امتیاز گرفتن، که طبیعیه چون کاربران بیشتر روی آثار جدیدتر نظر دادن.

نتیجه:

این نمودار دید واضح‌تری نسبت به «سال انتشار کتاب‌هایی که کاربران واقعاً بهشون امتیاز دادن» می‌ده. در واقع برخلاف نمودار کل کتاب‌ها، اینجا می‌بینیم تمرکز اصلی علاقه کاربران روی کتاب‌های جدیدتر بوده.

تحلیلی روی نمره های داده شده

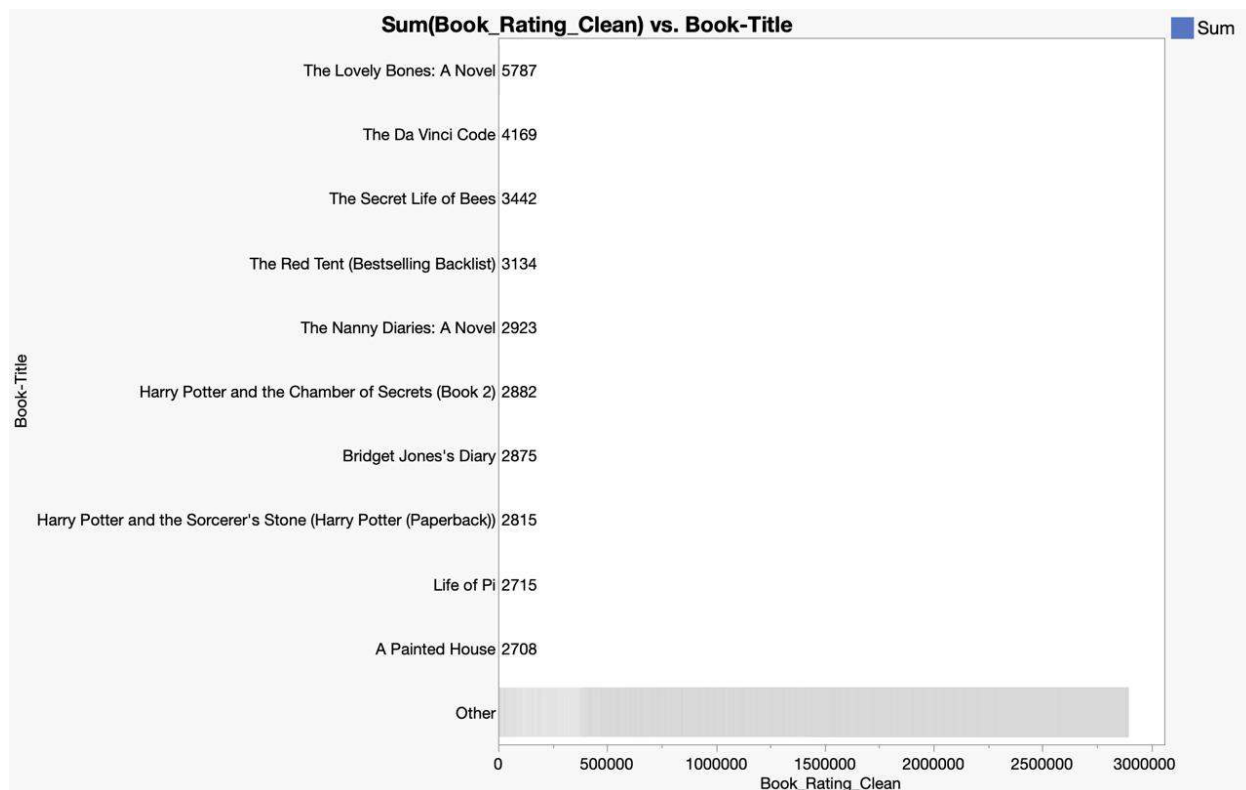
میانگین امتیاز کتاب ها در هر سال انتشار



- توی سال های خیلی قدیمی (قبل از 1950) به خاطر اینکه تعداد کتاب ها خیلی کم بوده، نوسان شدیدی توی میانگین امتیاز می بینیم. این تغییرات شدید بیشتر به دلیل کم بودن داده ها هست، نه لزوماً کیفیت کتاب ها.
- از حدود 1960 به بعد داده ها پایدارتر می شن و میانگین امتیاز بیشتر بین ۷.۵ تا ۸.۵ ثابت می مونه.
- این یعنی در مجموع، کتاب های منتشر شده در سال های مختلف تقریباً با کیفیت مشابهی از نظر کاربران ارزیابی شده اند.
- در سال های اخیر (دهه ۹۰ و ۲۰۰۰) مقدار میانگین تقریباً نزدیک به ۷.۶ بوده که با خلاصه آماری کل دیتاست هم همخوانی داره.

تحلیلی روی نمره های داده شده

تحلیل بر اساس مجموع امتیاز کتابها



تحلیل بر اساس مجموع امتیاز کتابها

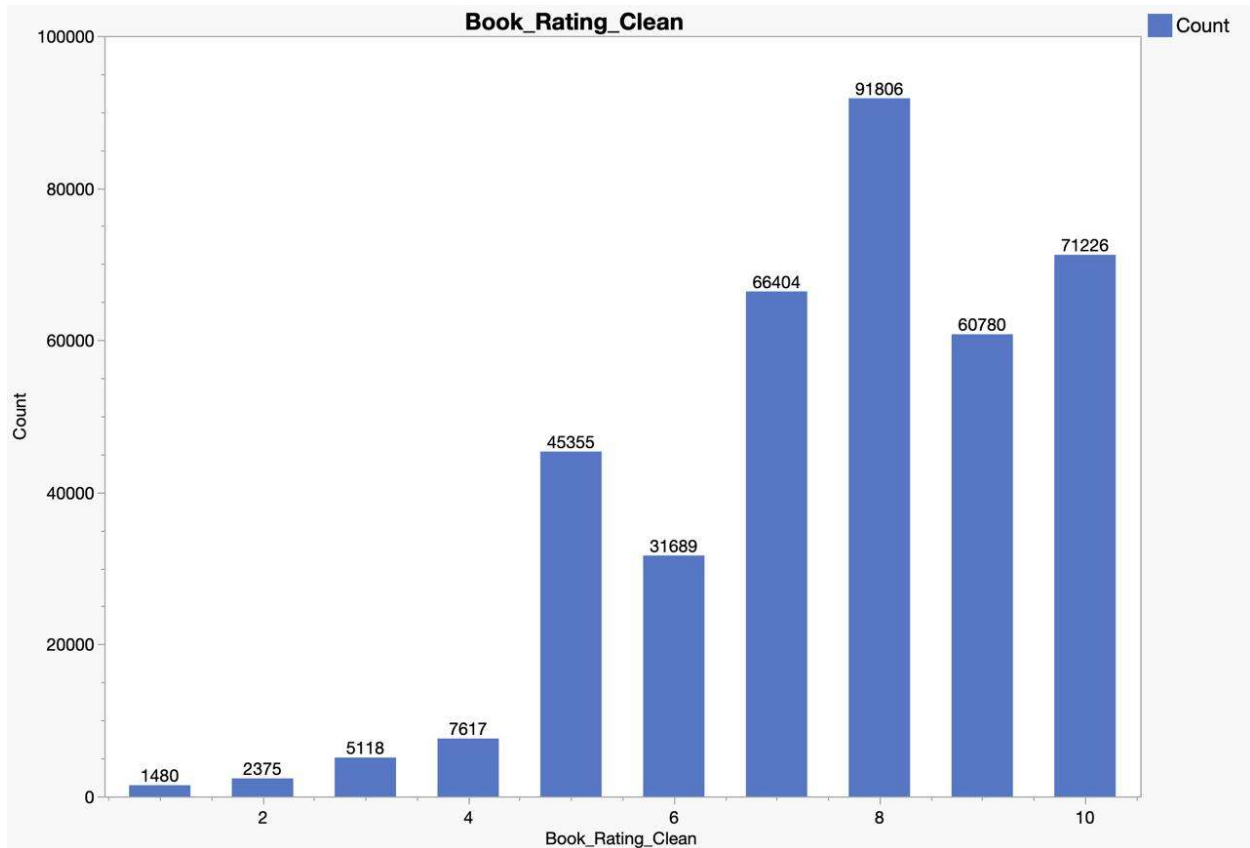
با محاسبه‌ی مجموع امتیازها (Sum of Ratings) مشاهده شد که کتاب‌هایی مثل The Da Vinci، The Lovely Bones و The Secret Life of Bees بیشترین مجموع امتیاز را دارند. این نتایج بیشتر نشان‌دهنده‌ی محبوبیت (تعداد زیاد امتیاز) هستند، نه لزوماً کیفیت.

تحلیل بر اساس میانگین امتیاز کتابها

وقتی میانگین امتیاز (Average Rating) محاسبه شد، برخی کتابها با امتیاز کامل ۱۰ ظاهر شدند. با این حال، بررسی دقیق‌تر نشان می‌دهد که در بسیاری از این موارد تنها یک یا چند نفر به کتاب امتیاز داده‌اند و این نتایج نمی‌تواند معیار مناسبی برای ارزیابی کیفیت کلی کتاب باشد.

تحلیلی روی نمره های داده شده

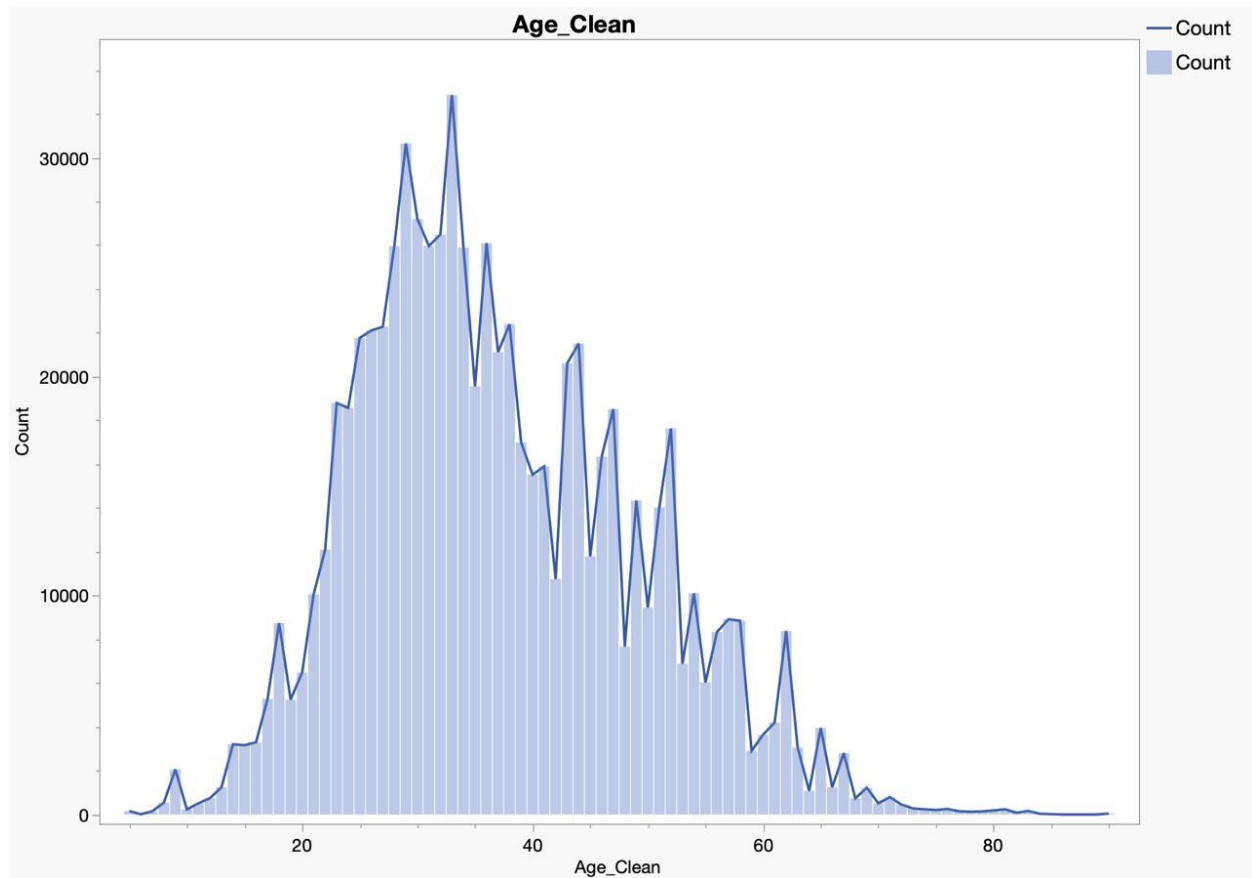
توزیع امتیازهای داده شده به کتابها



- همانطور که در نمودار اول مشخص است، بیشترین امتیازات داده شده بین ۷ تا ۱۰ قرار دارند.
- قله اصلی در امتیاز ۸ است (بیش از ۹۱,۸۰۶ بار ثبت شده).
- امتیازهای پایین (۱ تا ۳) خیلی کم داده شده اند که نشون میده کاربران به طور کلی دید مثبتی به کتابها داشتن.
- نتیجه: کاربران تمایل دارن امتیازهای نسبتاً بالا بدن و امتیازهای خیلی پایین نادر هست.

تحلیلی روی سن مشتریان

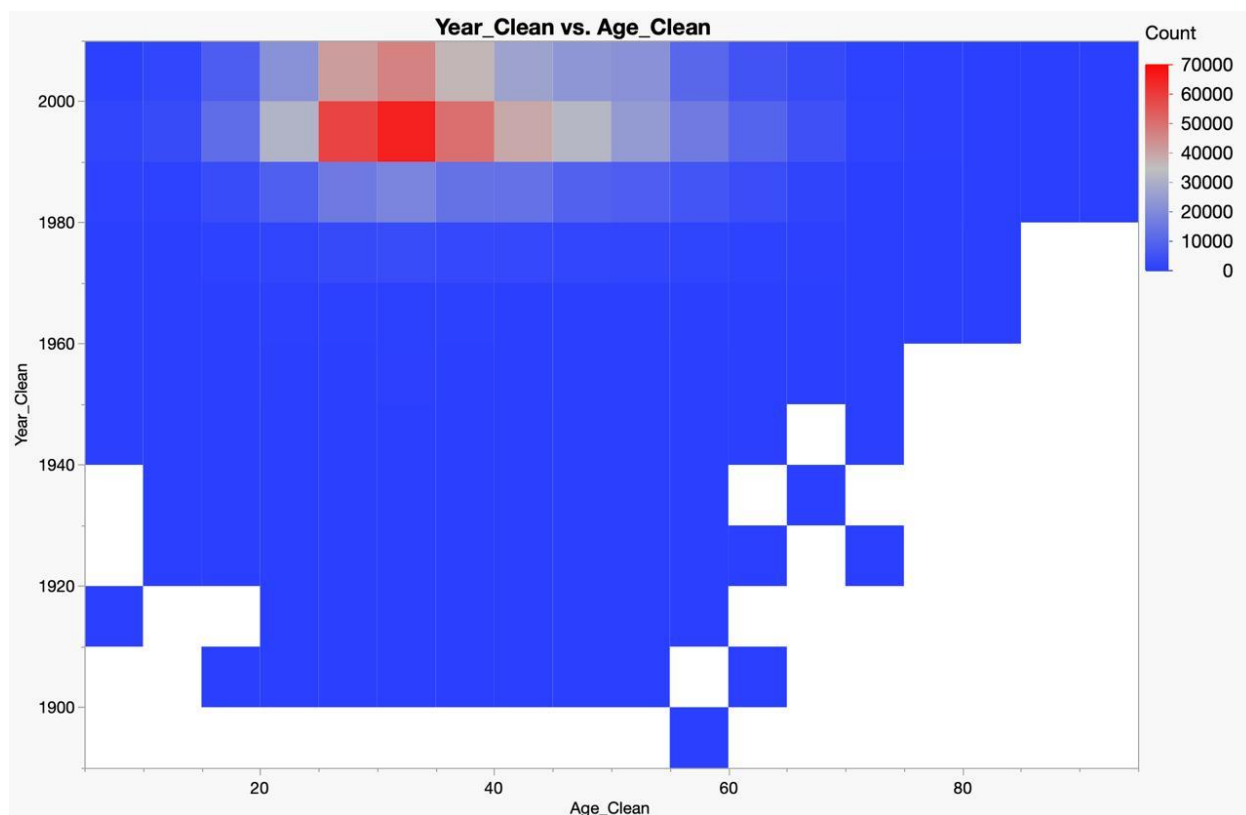
تحلیل توزیع سنی کاربران



- بیشترین تمرکز کاربران در بازه‌ی ۲۵ تا ۳۵ سال دیده می‌شه.
- بعد از این، گروه سنی ۲۰ تا ۲۵ سال و ۳۵ تا ۴۵ سال سهم بالایی دارن.
- سهم کاربران از سن بالای ۶۰ سال به شدت کاهش پیدا می‌کنه.
- گروه‌های خیلی کم‌سن (زیر ۱۵ سال) و خیلی مسن (بیش از ۷۰ سال) سهم ناچیزی دارن.

تحلیلی روی سن مشتریان

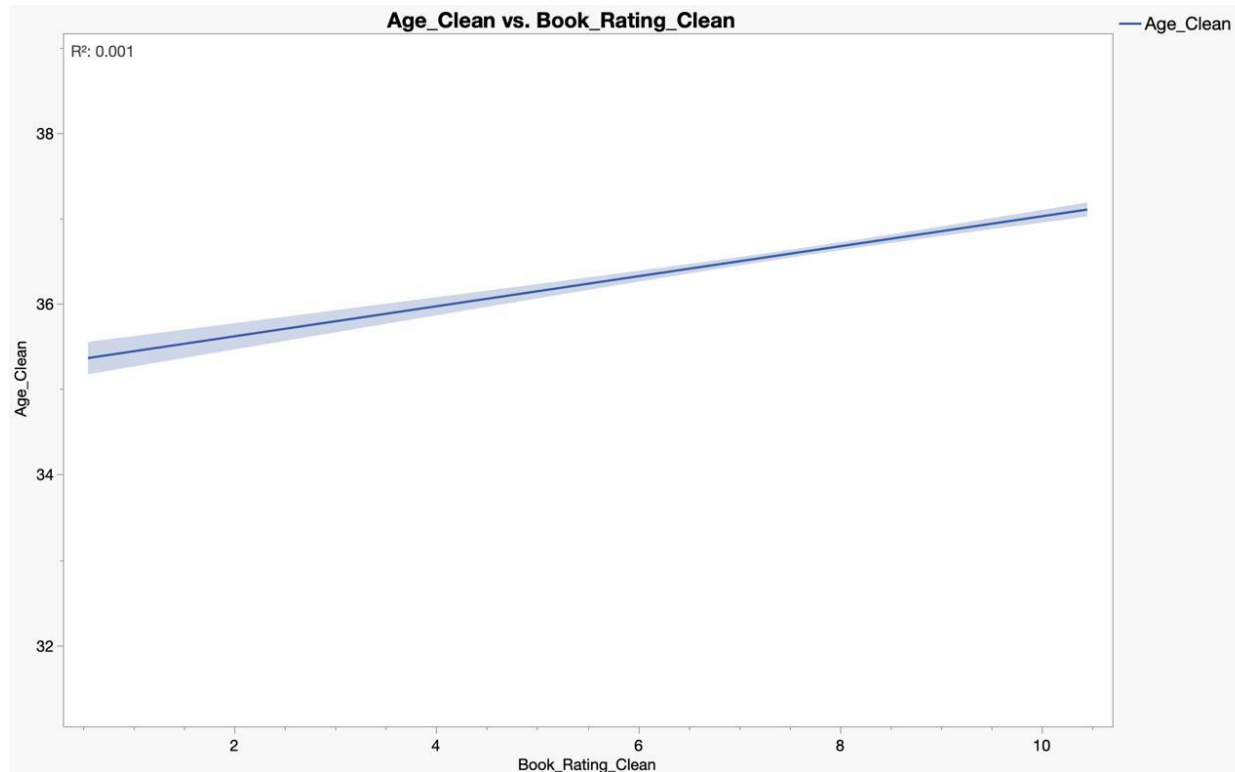
این نمودار Heatmap از ارتباط بین سن کاربران (Age_Clean) و سال انتشار کتابها (Year_Clean) نشون می‌ده



بیشترین تراکم (نقاط قرمز) مربوط به کاربران ۳۰ تا ۴۵ ساله است که بیشترین تعامل رو با کتابهای منتشرشده در بازه‌ی ۱۹۹۰ تا ۲۰۰۵ داشتن.

- کاربران جوان‌تر (۲۰ تا ۳۰ سال) هم بیشتر روی کتابهای جدیدتر (بعد از ۱۹۹۵) تمرکز دارن.
- کاربران بالای ۵۰ سال سهم خیلی کمتری دارن و بیشتر با کتابهای قدیمی‌تر تعامل کردن.
- داده‌های مربوط به سال‌های خیلی قدیمی (قبل از ۱۹۵۰) و کاربران خیلی مسن (بالای ۷۰ سال) پراکندگی کمی دارن و بخش عمده‌ای از اون‌ها حتی صفر (سفید) هست.

تحلیل ارتباط سن کاربران با امتیاز داده شده



- نمودار دوم یک خط رگرسیون نشون میده که سن کاربران رو نسبت به امتیاز بررسی کرده.
- خط تقریباً صعودی ملایمی داره: هرچه امتیاز بالاتر میره، میانگین سن کاربران کمی بیشتر میشه.
- البته ضریب تعیین ($R^2 = 0.001$) خیلی کمه → یعنی رابطه تقریباً ناچیزه و نمی شه گفت سن تأثیر زیادی روی نوع امتیازدهی داره.
- نتیجه: تقریباً همه گروه های سنی شبیه به هم امتیاز میدن، فقط یک تمایل خیلی خفیف وجود داره که کاربران بزرگتر کمی بیشتر امتیاز بالا بدن.

Association Rules

Mohammad Hossein Heidary

20 قانون با بالاترین Lift برای کتاب ها

قوانین چند تایی

Rules			
Rule		Confidence	Lift
Condition	Consequent		
0451190556, 0451190572	0451190548, 0451190564	94%	1574
0451190548, 0451190564	0451190556, 0451190572	85%	1574
0451190491, 0451190572	0451190548, 0451190564	90%	1514
0451190548, 0451190564	0451190491, 0451190572	85%	1514
0451190548, 0451190572	0451190556, 0451190564	87%	1513
0451190556, 0451190564	0451190548, 0451190572	89%	1513
0451190521, 0451190564	0451190491, 0451190548	91%	1490
0451190556, 0451190564	0451190491, 0451190548	91%	1490
0451190491, 0451190548	0451190521, 0451190564	86%	1490
0451190491, 0451190548	0451190556, 0451190564	86%	1490
0451190491, 0451190556	0451190548, 0451190564	87%	1461
0451190548, 0451190564	0451190491, 0451190556	87%	1461
0451190521, 0451190556	0451190491, 0451190548	89%	1459
0451190491, 0451190548	0451190521, 0451190556	84%	1459
0451190548, 0451190556	0451190491, 0451190564	87%	1435
0451190491, 0451190564	0451190548, 0451190556	86%	1435
0451190548, 0451190572	0451190491, 0451190564	87%	1431
0451190491, 0451190564	0451190548, 0451190572	84%	1431
0451190491, 0451190548, 0451190572	0451190564	100%	1396
0451190564	0451190491, 0451190548, 0451190572	71%	1396
0451190548, 0451190564	0451190491, 0451190521	87%	1386
0451190491, 0451190521	0451190548, 0451190564	83%	1386
0451190548, 0451190556	0451190564, 0451190572	85%	1381
0451190564, 0451190572	0451190548, 0451190556	82%	1381

در این بخش مجموعه‌ای از قوانین انجمنی بین کتاب‌ها استخراج شد. برای شروع، حداقل پشتیبانی (support) برابر با ۰.۰۰۰۵ و حداقل اطمینان (confidence) برابر با ۰.۴ در نظر گرفته شد. این مقادیر فقط برای یافتن ۲۰ قانون با بیشترین مقدار lift انتخاب شدند و لزوماً نشان‌دهنده بهترین معیار برای انتخاب قوانین نیستند.

نتایج نشان می‌دهد که بسیاری از قوانین استخراج‌شده مربوط به مجموعه‌ای خاص از کتاب‌ها هستند که به طور مکرر با یکدیگر دیده می‌شوند. به همین دلیل مقدار lift در این قوانین بسیار بالا (بیش از ۱۳۰۰) گزارش شده است.

20 قانون با بالاترین Lift برای کتاب ها

قوانین دوتایی

Rules			
Condition	Rule	Confidence	Lift
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: Coffey on the Mile (Green Mile Series)	86%	1187
The Green Mile: Coffey on the Mile (Green Mile Series)	The Green Mile: Night Journey (Green Mile Series)	85%	1187
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: Night Journey (Green Mile Series)	84%	1174
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	80%	1174
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	87%	1165
The Green Mile: Coffey's Hands (Green Mile Series)	The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	80%	1165
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	83%	1112
The Green Mile: Coffey's Hands (Green Mile Series)	The Green Mile: Night Journey (Green Mile Series)	80%	1112
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: Coffey on the Mile (Green Mile Series)	79%	1091
The Green Mile: Coffey on the Mile (Green Mile Series)	The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	75%	1091
The Green Mile: The Mouse on the Mile (Green Mile Series)	The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	74%	1076
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: The Mouse on the Mile (Green Mile Series)	84%	1076
The Green Mile: Coffey on the Mile (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	81%	1076
The Green Mile: Coffey's Hands (Green Mile Series)	The Green Mile: Coffey on the Mile (Green Mile Series)	78%	1076
The Green Mile: Coffey's Hands (Green Mile Series)	The Green Mile: The Mouse on the Mile (Green Mile Series)	84%	1075
The Green Mile: The Mouse on the Mile (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	81%	1075
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: The Mouse on the Mile (Green Mile Series)	80%	1027
The Green Mile: The Mouse on the Mile (Green Mile Series)	The Green Mile: Night Journey (Green Mile Series)	74%	1027

برای شفافیت بیشتر و مقایسه با قوانین پیچیده‌تر، مجموعه‌ای از ۲۰ قانون دوتایی با بالاترین مقدار Lift نیز استخراج شد.

در این قوانین، رابطه بین دو کتاب مشخص بررسی می‌شود؛ به این معنی که اگر کاربری یک کتاب از سری Green Mile را مطالعه کند، احتمال بسیار بالایی وجود دارد که کتاب دیگر این مجموعه را هم بخواند.

به عنوان مثال:

• اگر کتاب The Green Mile: Night Journey انتخاب شود، احتمال انتخاب The Green Mile: Coffey on the Mile نیز بسیار بالاست (با confidence حدود 86٪ و lift نزدیک به 1187).

• این نوع قوانین نشان‌دهنده‌ی الگوی مستقیم هم‌خوانی بین کتاب‌ها هستند و تفسیر آن‌ها برای مدیران یا کاربران غیر تخصصی ساده‌تر است.

20 قانون با بالاترین Confidence برای کتاب ها

قوانین چند تایی

Rules			
Rule		Confidence	Lift
Condition	Consequent		
0451190491, 0451190548, 0451190572	0451190564	100%	1396
0451190556, 0451190564, 0451190572	0451190548	100%	1335
0439136350, 0439139597, 0590353403	0439064864	99%	534.2
0439136350, 043935806X, 0590353403	0439064864	98%	531.6
0842329218, 0842329242, 0842329269	0842329250	98%	843.9
0451190521, 0451190548, 0451190564	0451190491	98%	1061
0451190548, 0451190556, 0451190572	0451190564	98%	1366
043935806X, 0590353403	0439064864	97%	523.7
0842329242, 0842329269	0842329250	97%	831.1
0451190521, 0451190564	0451190491	96%	1043
0451190491, 0451190572	0451190564	96%	1342
0451190556, 0451190572	0451190548	96%	1281
0451190491, 0451190556, 0451190564	0451190548	96%	1281
0451190548, 0451190556, 0451190564	0451190491	96%	1040
0451190491, 0451190521, 0451190556	0451190548	96%	1280
0312966970, 0312980140, 0312983867	0312971346	96%	439.5

وقتی حداکثر طول قوانین افزایش داده شد، قوانینی با ترکیب سه یا چند کتاب به دست آمدند که confidence آنها به ۱۰۰٪ هم رسید.

• این یعنی هرگاه ترکیب خاصی از چند کتاب (مثلاً سه جلد از یک مجموعه) توسط کاربری انتخاب شود، تقریباً همیشه جلد بعدی هم توسط همان کاربر خریداری/امتیازدهی شده است.

• مثال: ترکیب چند ISBN از سری The Green Mile یا مجموعه‌های دیگر که تقریباً همیشه به خرید جلد بعدی منجر شده است.

• این قوانین چندتایی نسبت به قوانین دوتایی اطلاعات دقیق‌تر و توصیه‌پذیرتر برای سیستم‌های پیشنهاددهنده دارند، چون نشان می‌دهند کاربران دنبال کامل کردن مجموعه‌ها هستند.

20 قانون با بالاترین Confidence برای کتاب‌ها

قوانین دو تایی

Rules			
Condition	Rule	Confidence	Lift
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	87%	1165
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Two Dead Girls (Green Mile Series)	87%	946
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: Coffey on the Mile (Green Mile Series)	86%	1187
The Green Mile: Coffey on the Mile (Green Mile Series)	The Green Mile: Night Journey (Green Mile Series)	85%	1187
The Green Mile: Night Journey (Green Mile Series)	The Two Dead Girls (Green Mile Series)	85%	919.4
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: Night Journey (Green Mile Series)	84%	1174
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: The Mouse on the Mile (Green Mile Series)	84%	1076
The Green Mile: Coffey's Hands (Green Mile Series)	The Green Mile: The Mouse on the Mile (Green Mile Series)	84%	1075
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	83%	1112
The Wide Window (A Series of Unfortunate Events, Book 3)	The Reptile Room (A Series of Unfortunate Events, Book 2)	83%	656
The Wide Window (A Series of Unfortunate Events, Book 3)	The Bad Beginning (A Series of Unfortunate Events, Book 1)	82%	315.8
Key of Knowledge (Key Trilogy (Paperback))	Key of Light (Key Trilogy (Paperback))	81%	549.9
The Green Mile: Coffey's Hands (Green Mile Series)	The Two Dead Girls (Green Mile Series)	81%	879.4
The Green Mile: Coffey on the Mile (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	81%	1076
The Green Mile: The Mouse on the Mile (Green Mile Series)	The Green Mile: Coffey's Hands (Green Mile Series)	81%	1075
The Green Mile: The Mouse on the Mile (Green Mile Series)	The Two Dead Girls (Green Mile Series)	81%	872.9
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	80%	1174
The Green Mile: Night Journey (Green Mile Series)	The Green Mile: The Mouse on the Mile (Green Mile Series)	80%	1027
The Green Mile: Coffey's Hands (Green Mile Series)	The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	80%	1165
The Green Mile: Coffey's Hands (Green Mile Series)	The Green Mile: Night Journey (Green Mile Series)	80%	1112
The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series)	The Green Mile: Coffey on the Mile (Green Mile Series)	79%	1091

در بین قوانین استخراج‌شده، مجموعه‌ای از قوانین دوتایی با بالاترین مقادیر confidence را دارند (۹۴-۸۷٪). این یعنی وقتی کاربر یکی از کتاب‌های سری The Green Mile را خوانده باشد، احتمال بسیار زیادی دارد که کتاب بعدی از همین مجموعه را هم انتخاب کند.

• این موضوع طبیعی است چون این کتاب‌ها به صورت سریالی منتشر شده‌اند و مخاطب برای فهم کامل داستان، به خواندن کل مجموعه نیاز دارد.

• بنابراین، بالاتر بودن confidence در اینجا نشان‌دهنده پیوستگی محتوایی است نه صرفاً هم‌خوانی تصادفی.

• با این حال، از نظر تحلیلی باید دقت کرد که بالا بودن confidence تنها معیار کیفیت قانون نیست؛ چون ممکن است این الگو فقط ناشی از ماهیت سریالی کتاب‌ها باشد.

تحلیل مجموعه Green Mile

در نتایج مشاهده می‌شود که بخش زیادی از قوانین به جلد‌های مختلف مجموعه‌ی Green Mile مربوط هستند. این نشان می‌دهد که خوانندگان وقتی یکی از جلد‌ها را تهیه می‌کنند، معمولاً به سراغ سایر جلد‌های همین مجموعه هم می‌روند. چنین الگویی طبیعی است چون کتاب‌ها به صورت سریالی منتشر شده‌اند و داستان ادامه‌دار دارند.

این موضوع اهمیت قوانین استخراج‌شده را دو برابر می‌کند:

- از یک طرف نشان می‌دهد قوانین صرفاً تصادفی نیستند و با رفتار واقعی خریداران هماهنگ‌اند.
- از طرف دیگر نشان می‌دهد خروجی تحلیل می‌تواند در عمل برای پیشنهاد خرید (Recommendation) استفاده شود، مثلاً پیشنهاد جلد بعدی مجموعه به کاربری که جلد اول را خریده است.

توضیح تکمیلی:

اگرچه در این بخش قوانین بر اساس لیفت بالا یا کانفیدنس بالا نمایش داده شده‌اند، اما لازم است توجه شود که ساپورت نیز معیار مهمی در اعتبارسنجی قوانین است. لیفت و کانفیدنس می‌توانند قدرت و جهت رابطه بین کتاب‌ها را نشان دهند، اما اگر ساپورت پایین باشد، این قوانین تنها در بخش کوچکی از داده‌ها اتفاق افتاده‌اند و ممکن است قابل تعمیم به کل کاربران نباشند. بنابراین در کاربرد واقعی، ترکیب این سه معیار (ساپورت، کانفیدنس و لیفت) برای شناسایی قوانین معتبر ضروری است.

30 قانون با بالاترین Support برای کتاب ها

قوانین چند تایی

	Condition	Consequent	Confidence	Lift	Support
1	0439064872	059035342X	50%	80.420839134	0.0038
2	0345313860	0345337662	53%	96.75574173	0.0033
3	0316096199	0316666343	45%	31.662289152	0.0033
4	0345313860	0345351525	44%	147.95122485	0.0033
5	0553279912	0553280341	41%	160.21485281	0.0031
6	0345351525	0345313860	48%	147.95122485	0.0030
7	0345351525	0345337662	48%	86.677018634	0.0030
8	0671001795	0312966091	49%	200.19713082	0.0029
9	0671001795	0312966970	41%	168.64469413	0.0029
10	0671001795	0061009059	41%	130.39538206	0.0029
11	0515131229	051513287X	46%	179.24732607	0.0028
12	0515131229	0515132020	46%	200.59346622	0.0028
13	034540288X	0345370775	41%	81.757148767	0.0027
14	0671027387	0671027360	40%	62.996207044	0.0027
15	051513287X	0515131229	50%	179.24732607	0.0026
16	051513287X	0515132020	49%	214.66642556	0.0026
17	0553280341	0553279912	50%	160.21485281	0.0026
18	0553280341	0553280368	44%	229.30939385	0.0026
19	0439136369	0439064872	54%	141.6504803	0.0025
20	0439136369	059035342X	48%	77.082001767	0.0025
21	0312966970	0312971346	56%	254.56882255	0.0024
22	0312966970	0312966091	56%	228.43005952	0.0024
23	0312966970	0671001795	49%	168.64469413	0.0024
24	0312966970	0312976275	48%	223.09393333	0.0024
25	0312966970	0312966091, 0312971346	43%	357.71735736	0.0024
26	0312966970	0312971346, 0312976275	42%	347.21369048	0.0024
27	0312966970	0312966091, 0671001795	41%	289.69148718	0.0024
28	0312966970	0312966091, 0312976275	40%	372.13333333	0.0024
29	0312966091	0671001795	58%	200.19713082	0.0024
30	0312966091	0312966970	56%	228.43005952	0.0024

این تصویر 30 قانون چندتایی با بیشترین مقدار Support را نشان می‌دهد. نکته مهم این است که در این بخش، دیگر آن لیفت‌های بسیار بالا (بیش از ۱۳۰۰) مشاهده نمی‌شود و اغلب قوانین در محدوده پایین‌تری قرار گرفته‌اند. بنابراین مشخص می‌شود که قوانین چندتایی لزوماً کیفیت بهتری ندارند و در بسیاری موارد، تنها ترکیبی از همان قوانین دوتایی قوی‌تر هستند.

20 قانون با بالاترین Support برای کتاب ها

قوانین دو تایی

Condition	Consequent	Confidence	Lift	Support
1 Harry Potter and the Chamber of Secrets (Book 2)	Harry Potter and the Prisoner of Azkaban (Book 3)	49%	106.92421844	0.0059
2 Harry Potter and the Chamber of Secrets (Book 2)	Harry Potter and the Goblet of Fire (Book 4)	43%	104.16969565	0.0059
3 Harry Potter and the Chamber of Secrets (Book 2)	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperb...	40%	64.365558835	0.0059
4 Harry Potter and the Prisoner of Azkaban (Book 3)	Harry Potter and the Goblet of Fire (Book 4)	55%	133.95456255	0.0046
5 Harry Potter and the Prisoner of Azkaban (Book 3)	Harry Potter and the Chamber of Secrets (Book 2)	63%	106.92421844	0.0046
6 Harry Potter and the Goblet of Fire (Book 4)	Harry Potter and the Prisoner of Azkaban (Book 3)	61%	133.95456255	0.0041
7 Harry Potter and the Goblet of Fire (Book 4)	Harry Potter and the Chamber of Secrets (Book 2)	62%	104.16969565	0.0041
8 The Fellowship of the Ring (The Lord of the Rings, Part 1)	The Two Towers (The Lord of the Rings, Part 2)	44%	157.87805229	0.0039
9 Harry Potter and the Order of the Phoenix (Book 5)	Harry Potter and the Goblet of Fire (Book 4)	42%	103.22142171	0.0038
10 Harry Potter and the Order of the Phoenix (Book 5)	Harry Potter and the Chamber of Secrets (Book 2)	42%	71.105084208	0.0038
11 The Vampire Lestat (Vampire Chronicles, Book II)	The Queen of the Damned (Vampire Chronicles (Paperback))	44%	147.41125688	0.0033
12 The Vampire Lestat (Vampire Chronicles, Book II)	Interview with the Vampire	53%	96.977156198	0.0033
13 Lucky : A Memoir	The Lovely Bones: A Novel	45%	31.662289152	0.0033
14 Digital Fortress : A Thriller	Angels & Demons	40%	56.216031092	0.0031
15 A Is for Alibi (Kinsey Millhone Mysteries (Paperback))	B Is for Burglar (Kinsey Millhone Mysteries (Paperback))	42%	157.64313222	0.0031
16 Seven Up (A Stephanie Plum Novel)	High Five (A Stephanie Plum Novel)	43%	176.43107143	0.0030
17 Seven Up (A Stephanie Plum Novel)	Hard Eight : A Stephanie Plum Novel (A Stephanie Plum No...	44%	154.0886692	0.0030
18 Harry Potter and the Sorcerer's Stone (Book 1)	Harry Potter and the Prisoner of Azkaban (Book 3)	44%	96.031563981	0.0030
19 Harry Potter and the Sorcerer's Stone (Book 1)	Harry Potter and the Chamber of Secrets (Book 2)	50%	84.190899083	0.0030
20 The Queen of the Damned (Vampire Chronicles (Paper...	The Vampire Lestat (Vampire Chronicles, Book II)	48%	147.41125688	0.0030

این تصویر ۲۰ قانون دوتایی با بیشترین مقدار Support را نشان می‌دهد. نتایج بیانگر این است که کتاب‌هایی مانند مجموعه Harry Potter یا The Vampire Chronicles بیشترین هم‌رخدادی را در بین کاربران دارند. از آن جایی که قوانین دوتایی ساده‌تر و شفاف‌تر هستند، همچنان بخش مهمی از تحلیل را تشکیل می‌دهند. (البته در قوانین با بالاترین Lift - که قبلاً بررسی شد - مقادیر بسیار بالاتر مثل ۱۳۰۰ نیز دیده شد).

مقایسه‌ی تحلیل بر اساس ISBN و نام کتاب‌ها

در فرایند استخراج قوانین انجمنی، انتخاب سطح شناسایی آیتم‌ها نقش مهمی در نتایج دارد. در این پروژه داده‌ها هم بر اساس ISBN (شناسه بین‌المللی کتاب) و هم بر اساس Book Title (عنوان کتاب) تحلیل شدند و تفاوت‌های مهمی مشاهده گردید:

1. دقت در سطح جزئی (ISBN):

هر نسخه از یک کتاب (مانند چاپ‌های متفاوت، ناشرهای مختلف یا فرمت‌های هاردکاور و پیپر بک) دارای ISBN یکتاست. بنابراین تحلیل بر اساس ISBN بسیار دقیق است، اما منجر به پراکندگی ساپورت می‌شود. به عنوان مثال، یک کتاب پرفروش با چندین ISBN در نتایج به چند بخش تقسیم شده و ممکن است در بین قوانین برتر با ساپورت بالا ظاهر نشود.

2. تجمع در سطح کلی (Book Title):

در مقابل، تحلیل بر اساس عنوان کتاب باعث ادغام نسخه‌های مختلف در یک عنوان واحد می‌شود. این کار ساپورت را افزایش داده و الگوهای مطالعه را بهتر نشان می‌دهد. به‌ویژه برای مجموعه‌های پرتعداد مانند Harry Potter یا The Da Vinci Code، این رویکرد تصویر واقعی‌تری از محبوبیت کتاب ارائه می‌دهد.

در نهایت، نتایج تحلیلی و توضیحات تفسیری عمدتاً بر اساس نام کتاب‌ها ارائه شده است تا الگوهای کشف‌شده به شکل روشن‌تر بیان شوند، هرچند تحلیل با ISBN نیز انجام و مقایسه شد.

پیش‌پردازش داده‌ها: حذف کتاب‌های بدون امتیاز

از آنجایی که بخش بزرگی از داده‌ها مربوط به کتاب‌هایی بود که امتیاز دریافت نکرده بودند، تصمیم گرفتیم این موارد را از تحلیل حذف کنیم. دلیل این کار آن است که کتاب‌های بدون امتیاز نمی‌توانند در فرآیند کشف قوانین انجمنی نقش معناداری داشته باشند و صرفاً باعث افزایش نویز و حجم محاسبات می‌شوند. (البته نه لزوماً)

در جدول اولیه، تعداد کل رکوردها 1,031,165 بود که از این میان 647,315 رکورد مربوط به کتاب‌های بدون امتیاز (Book_Rating_Clean = NA یا 0) بودند. پس از فیلتر کردن، تنها 383,850 رکورد با امتیاز معتبر (بین 1 تا 10) باقی ماندند.

قوانین با بالاترین لیفت و کانفیدنس

(فقط داده‌های امتیاز داده‌شده)

با حداقل ساپورت 0.0004

	Condition	Consequent	Confidence	Lift
1	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	73%	1186.23693%
2	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	71%	1186.23693%
3	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	74%	1142.2023%
4	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	70%	1142.2023%
5	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	71%	1094.57317%
6	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	66%	1094.57317%
7	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	73%	996.439024%
8	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	60%	996.439024%
9	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	70%	959.4%
10	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	62%	959.4%
11	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	67%	907.866666%
12	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	56%	907.866666%
13	Key of Valor (Roberts, Nora. Key Trilogy, 3.)	Key of Knowledge (Key Trilogy (Paperback))	61%	827.760784%
14	Key of Knowledge (Key Trilogy (Paperback))	Key of Valor (Roberts, Nora. Key Trilogy, 3.)	62%	827.760784%
15	Soul Harvest: The World Takes Sides (Left Behind No. 4)	Nicolae: The Rise of Antichrist (Left Behind No. 3)	64%	736.639019%
16	Nicolae: The Rise of Antichrist (Left Behind No. 3)	Soul Harvest: The World Takes Sides (Left Behind No. 4)	51%	736.639019%
17	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	Guilty Pleasures (Anita Blake Vampire Hunter (Paperback))	66%	614.760273%
18	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	Guilty Pleasures (Anita Blake Vampire Hunter (Paperback))	64%	596.953424%
19	Guilty Pleasures (Anita Blake Vampire Hunter (Paperback))	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	44%	596.953424%
20	The Waste Lands (The Dark Tower, Book 3)	The Drawing of the Three (The Dark Tower, Book 2)	69%	558.410052%

	Condition	Consequent	Confidence	Lift
1	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	74%	1142.2023%
2	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	73%	1186.23693%
3	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	73%	996.439024%
4	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	71%	1186.23693%
5	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	71%	1094.57317%
6	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	70%	1142.2023%
7	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	70%	959.4%
8	Soul Harvest: The World Takes Sides (Left Behind No. 4)	Tribulation Force: The Continuing Drama of Those Left Behind (L...	70%	555.905492%
9	The Waste Lands (The Dark Tower, Book 3)	The Drawing of the Three (The Dark Tower, Book 2)	69%	558.410052%
10	Bloody Bones (Anita Blake Vampire Hunter (Paperback))	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	67%	907.866666%
11	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	66%	1094.57317%
12	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	Guilty Pleasures (Anita Blake Vampire Hunter (Paperback))	66%	614.760273%
13	The Amber Spyglass (His Dark Materials, Book 3)	The Subtle Knife (His Dark Materials, Book 2)	64%	408.407030%
14	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	Guilty Pleasures (Anita Blake Vampire Hunter (Paperback))	64%	596.953424%
15	Soul Harvest: The World Takes Sides (Left Behind No. 4)	Nicolae: The Rise of Antichrist (Left Behind No. 3)	64%	736.639019%
16	The Waste Lands (The Dark Tower, Book 3)	The Gunslinger (The Dark Tower, Book 1)	62%	455.560334%
17	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	Circus of the Damned (Anita Blake Vampire Hunter (Paperback))	62%	959.4%
18	Key of Knowledge (Key Trilogy (Paperback))	Key of Valor (Roberts, Nora. Key Trilogy, 3.)	62%	827.760784%
19	Key of Valor (Roberts, Nora. Key Trilogy, 3.)	Key of Knowledge (Key Trilogy (Paperback))	61%	827.760784%
20	The Laughing Corpse (Anita Blake Vampire Hunter (Paperback))	The Lunatic Cafe (Anita Blake Vampire Hunter (Paperback))	60%	996.439024%

20 قانون با بالاترین Support برای کتاب ها

(فقط داده‌های امتیاز داده‌شده)

	Condition	Consequent	Confidence	Lift	Support
1	Harry Potter and the Chamber of Secrets (Book 2)	Harry Potter and the Prisoner of Azkaban (Book 3)	48%	119.21981945	0.0047
2	Harry Potter and the Chamber of Secrets (Book 2)	Harry Potter and the Goblet of Fire (Book 4)	42%	117.45397904	0.0047
3	Harry Potter and the Prisoner of Azkaban (Book 3)	Harry Potter and the Chamber of Secrets (Book 2)	56%	119.21981945	0.0040
4	Harry Potter and the Prisoner of Azkaban (Book 3)	Harry Potter and the Goblet of Fire (Book 4)	51%	143.76244194	0.0040
5	Harry Potter and the Goblet of Fire (Book 4)	Harry Potter and the Prisoner of Azkaban (Book 3)	58%	143.76244194	0.0036
6	Harry Potter and the Goblet of Fire (Book 4)	Harry Potter and the Chamber of Secrets (Book 2)	55%	117.45397904	0.0036
7	Harry Potter and the Order of the Phoenix (Book 5)	Harry Potter and the Goblet of Fire (Book 4)	40%	113.34554072	0.0031
8	Harry Potter and the Sorcerer's Stone (Book 1)	Harry Potter and the Chamber of Secrets (Book 2)	52%	110.35060234	0.0025
9	Harry Potter and the Sorcerer's Stone (Book 1)	Harry Potter and the Prisoner of Azkaban (Book 3)	46%	114.91498249	0.0025
10	Harry Potter and the Sorcerer's Stone (Book 1)	Harry Potter and the Goblet of Fire (Book 4)	40%	113.84655807	0.0025
11	The Two Towers (The Lord of the Rings, Part 2)	The Fellowship of the Ring (The Lord of the Rings, Part 1)	56%	186.3437329	0.0020
12	The Two Towers (The Lord of the Rings, Part 2)	The Return of the King (The Lord of the Rings, Part 3)	44%	299.6259626	0.0020
13	The Subtle Knife (His Dark Materials, Book 2)	The Golden Compass (His Dark Materials, Book 1)	44%	188.10497855	0.0016
14	The Subtle Knife (His Dark Materials, Book 2)	The Amber Spyglass (His Dark Materials, Book 3)	40%	408.40703027	0.0016
15	Four To Score (A Stephanie Plum Novel)	High Five (A Stephanie Plum Novel)	43%	296.80602338	0.0015
16	The Return of the King (The Lord of the Rings, Part 3)	The Two Towers (The Lord of the Rings, Part 2)	59%	299.6259626	0.0015
17	The Return of the King (The Lord of the Rings, Part 3)	The Fellowship of the Ring (The Lord of the Rings, Part 1)	56%	189.29571282	0.0015
18	The Queen of the Damned (Vampire Chronicles (Pa...	The Vampire Lestat (Vampire Chronicles, Book II)	48%	228.55384615	0.0015
19	High Five (A Stephanie Plum Novel)	Four To Score (A Stephanie Plum Novel)	45%	296.80602338	0.0014
20	High Five (A Stephanie Plum Novel)	Three To Get Deadly : A Stephanie Plum Novel (A Stephanie Pl...	44%	314.48657358	0.0014

مقایسه قوانین قبل و بعد از حذف کتاب‌های بدون امتیاز

1. قبل از حذف (کل داده‌ها):

- بخش زیادی از قوانین به کتاب‌های Harry Potter مربوط می‌شدند.
- علاوه بر هری پاتر، مجموعه‌هایی مثل The Vampire Chronicles و حتی کتاب‌های پراکنده مثل Lucky: A Memoir یا Digital Fortress هم در بین قوانین پرتکرار دیده شدند.
- این نشان می‌دهد که وجود داده‌های بدون امتیاز، تنوع بیشتری به قوانین می‌دهد ولی کیفیت و اعتبار آن‌ها پایین‌تر است، چون بخشی از این ارتباط‌ها صرفاً از حضور زیاد داده‌ها در دیتاست ناشی می‌شود نه از رفتار واقعی کاربران.

2. بعد از حذف (فقط داده‌های امتیازدار):

- قوانین مرتب‌تر و متمرکزتر شدند.
- همچنان مجموعه Harry Potter سهم اصلی را دارد، اما روابط آن قوی‌تر و معنادارتر شده‌اند (مثلاً کتاب‌های مختلف این سری تقریباً همواره همراه با هم ظاهر می‌شوند).
- مجموعه‌های داستانی دیگر مثل The Lord of the Rings، His Dark Materials و Vampire Chronicles هم ارتباط منسجم‌تری نشان دادند.
- قوانین پراکنده و کم‌معنا (مثل حضور تک‌کتاب‌ها) تقریباً حذف شدند.

تحلیل میانگین نمرات ۱۰ کتاب با بیشترین ساپورت پس از حذف کتابهای با نمره صفر

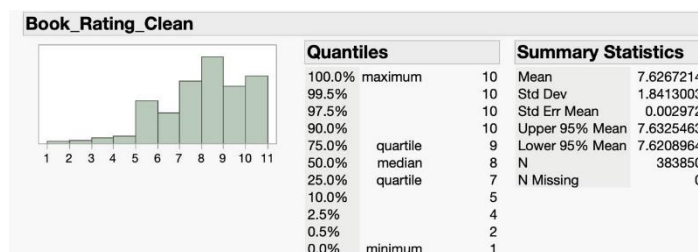
	Item Set	Support	N Rows	Mean(Book-Rating)
1	The Lovely ...	0.0104	707	8.1852899576
2	Wild Animus	0.0085	581	4.3907056799
3	The Da Vinc...	0.0073	494	8.4392712551
4	The Secret ...	0.0059	406	8.4778325123
5	The Nanny ...	0.0057	393	7.4376590331
6	The Red Te...	0.0056	383	8.182767624
7	Bridget Jon...	0.0054	377	7.625994695
8	A Painted H...	0.0053	366	7.3989071038
9	Life of Pi	0.0049	336	8.0803571429
10	Divine Secr...	0.0047	323	7.8761609907

1. جدول ۱۰ کتاب با بیشترین ساپورت

• برای هر کتاب تعداد رکوردها (N Rows)، ساپورت و میانگین نمره محاسبه شد.

• مثلاً The Lovely Bones ساپورت ۰.۰۱۰۴ و میانگین نمره ۸.۱۸ دارد.

• بیشتر این کتابها امتیازی بالاتر از ۷.۵ دارند، به جز Wild Animus با میانگین ۴.۳.



2. آمار کل دیتاست (عکس دوم):

• میانگین کل امتیازات = ۷.۶۲۶

• میانه = ۸

• بیشترین مقدار = ۱۰، کمترین = ۱

• انحراف معیار ≈ 1.84 → پراکندگی نسبتاً زیاد

ولی بیشتر امتیازها بین ۷ تا ۹ قرار دارند.

جمع‌بندی

بررسی ۱۰ کتاب با بیشترین ساپورت نشون داد که بیشتر این عناوین میانگین امتیازی بالاتر از میانگین کل دیتاست (۷.۶۲) دارند. این موضوع نشون می‌ده کتاب‌هایی که بیشتر خوانده یا امتیازدهی شده‌اند، معمولاً کیفیت بالاتری هم از نظر کاربران دارند. تنها استثنا Wild Animus بود که با وجود ساپورت بالا، میانگین امتیاز پایینی (۴.۳) داشت.

20 قانون با بالاترین lift برای نویسندگان

Rules				
Rule		Confidence	Lift	
Condition	Consequent			
Christine Rimmer	Gina Wilkins	49%	474.8	
Gina Wilkins	Christine Rimmer	50%	474.8	
Jerry Jenkins	Tim F. Lahaye	62%	372.3	
Carole Mortimer	Emma Darcy	45%	362.7	
Emma Darcy	Carole Mortimer	42%	362.7	
Emma Darcy	Rebecca Winters	43%	336.5	
Rebecca Winters	Emma Darcy	42%	336.5	
Gina Wilkins	Marie Ferrarella	54%	324	
STAN BERENSTAIN	Stan Berenstain	61%	322.2	
Susan Kearney	Rebecca York	48%	313.3	
Christine Rimmer	Marie Ferrarella	51%	302.1	
Emma Darcy	Betty Neels	42%	282.7	
Pat Warren	Nancy Martin	55%	278.1	
Connie Mason	Cassie Edwards	42%	275	
Kate William	Francine Pascal	77%	270.4	
Tim LaHaye	Tim F. Lahaye	45%	269.3	
Tim F. Lahaye	Tim LaHaye	73%	269.3	
DAVID EDDINGS	David Eddings	80%	251.5	
Jerry Jenkins	Tim LaHaye	68%	250.9	
FRANCINE PASCAL	Francine Pascal	71%	248.1	
Emma Darcy	Marie Ferrarella	41%	244.4	

در این مرحله، قوانین انجمنی مستقیماً روی ستون -Book Author اجرا شد.

نتیجه نشان می‌دهد نویسندگانی مثل Christine Rimmer و Gina Wilkins یا Emma Darcy و Carole Mortimer بیشترین ارتباط را با هم دارند.

• بالاترین Lift (474.8) مربوط به رابطه‌ی Christine Rimmer ↔ Gina Wilkins است.

• برخی نویسندگان مانند Tim LaHaye و David Eddings نیز قوانین با Confidence بالا (بیش از 70٪) ایجاد کرده‌اند.

20 قانون با بالاترین Confidence برای نویسندگان

Rules				
Rule		Confidence	Lift	
Condition	Consequent			
Maura Seger	Nora Roberts	89%	34.58	
Donna Kauffman	Nora Roberts	87%	34.03	
Hannah Howell	Nora Roberts	85%	33.14	
Janet Woods	Nora Roberts	85%	33.14	
Constance O'Banyon	Jude Deveraux	84%	89.23	
NORA ROBERTS	Nora Roberts	84%	32.82	
Norah Hess	Nora Roberts	84%	32.7	
Anne McAllister	Nora Roberts	83%	32.27	
Suzanne Forster	Nora Roberts	81%	31.46	
Dallas Schulze	Nora Roberts	80%	31.37	
Christine Rimmer	Nora Roberts	80%	31.36	
John R. Maxim	John Grisham	80%	20.46	
DAVID EDDINGS	David Eddings	80%	251.5	
Madeline Baker	Nora Roberts	80%	31.12	
Merline Lovelace	Nora Roberts	80%	31.12	
PIERS ANTHONY	Piers Anthony	79%	106.6	
Jeanne Kalogridis	Stephen King	79%	18.32	
Jim Miller	Nora Roberts	79%	30.87	
Amanda Scott	Nora Roberts	79%	30.85	
J.D. Robb	Nora Roberts	79%	30.84	

• بیشترین تکرار مربوط به نویسنده‌ی Nora Roberts است که در بسیاری از قوانین به‌عنوان Consequent ظاهر شده و ضریب اطمینان بالای 80–89٪ دارد.

• این موضوع نشان می‌دهد خوانندگانی که آثار نویسندگانی مثل Maura Seger، Donna Kauffman یا Hannah Howell را انتخاب کرده‌اند، به احتمال زیاد به سراغ آثار Nora Roberts هم رفته‌اند.

• در میان سایر نویسندگان، David Eddings → Constance و David Eddings (80٪) و O'Banyon → Jude Deveraux (84٪) نیز روابط قابل توجهی نشان می‌دهند.

20 قانون با بالاترین Support برای کتاب ها

	Condition	Consequent	Confidence	Lift	Support
1	JOHN GRISHAM	John Grisham	47%	12.04266724	0.0219
2	Dean R. Koontz	Stephen King	42%	9.7626960346	0.0212
3	Anne Rice	Stephen King	41%	9.5391749452	0.0179
4	Sandra Brown	Nora Roberts	44%	17.292014174	0.0134
5	Sandra Brown	James Patterson	40%	12.984927285	0.0134
6	Robert Ludlum	John Grisham	41%	10.488821065	0.0130
7	Robin Cook	Stephen King	43%	9.9942141787	0.0126
8	Robin Cook	John Grisham	43%	10.846581824	0.0126
9	MICHAEL CRICHTON	Michael Crichton	42%	19.440024779	0.0119
10	Jonathan Kellerman	James Patterson	45%	14.510234983	0.0117
11	Jonathan Kellerman	John Grisham	41%	10.510519794	0.0117
12	David Baldacci	John Grisham	41%	10.552406873	0.0111
13	John Sandford	James Patterson	46%	14.798126027	0.0111
14	John Sandford	John Grisham	42%	10.758139123	0.0111
15	Thomas Harris	Stephen King	46%	10.72970918	0.0110
16	Thomas Harris	John Grisham	41%	10.370652359	0.0110
17	Catherine Coulter	Nora Roberts	50%	19.322226294	0.0109
18	Nicholas Evans	John Grisham	40%	10.204489386	0.0106
19	Judy Blume	Stephen King	40%	9.3352707578	0.0103
20	Sidney Sheldon	John Grisham	44%	11.297499986	0.0095

در این بخش قوانین با بیشترین Support برای نویسندگان بررسی شد.

1. پراکندگی نویسندگان پرتکرار:

نویسندگانی مثل John Grisham, Stephen King, Nora Roberts و James Patterson بیشترین حضور را در هر دو سمت قوانین دارند. این موضوع نشان می‌دهد که آثار این نویسندگان بیشترین تعداد تراکنش مشترک را در کل داده‌ها داشته‌اند.

2. John Grisham به‌عنوان نویسنده‌ی محوری:

• در چندین قانون، John Grisham هم به‌عنوان شرط (Condition) و هم نتیجه (Consequent) ظاهر شده.

• ساپورت بالاتر (تا 0.0219) و لیفت نسبتاً زیاد (بیش از 10) نشان می‌دهد که خوانندگان او معمولاً مجموعه‌ی وسیعی از کتاب‌های خودش را مطالعه می‌کنند.

3. Stephen King و Nora Roberts به‌عنوان نویسندگان مکمل:

• قوانین مثل Dean R. Koontz → Stephen King یا Sandra Brown → Nora Roberts نشان می‌دهند که طرفداران برخی نویسندگان مشابه، به سمت این دو نویسنده هم گرایش پیدا کرده‌اند.

• لیفت بالا (بیش از 17 برای Sandra Brown → Nora Roberts) نشان‌دهنده‌ی قدرت بالای این ارتباط است.

4. James Patterson به‌عنوان جایگزین پرتکرار:

• در قوانین متعددی (مثلاً Jonathan Kellerman → James Patterson یا John Sandford → James Patterson) حضور دارد.

• لیفت نسبتاً بالا (بیش از 14) نشان می‌دهد که آثار Patterson برای خوانندگان نویسندگان مشابه بسیار جذاب است.

5. تفسیر کلی:

این نتایج بیانگر این است که بازار کتاب در داده‌ها به‌شدت متمرکز بر چند نویسنده‌ی محبوب (Patterson, Roberts, King, Grisham) است.

• قوانین با ساپورت بالا بیشتر سراغ نویسندگان پرتیراژ و عام‌پسند رفته‌اند.

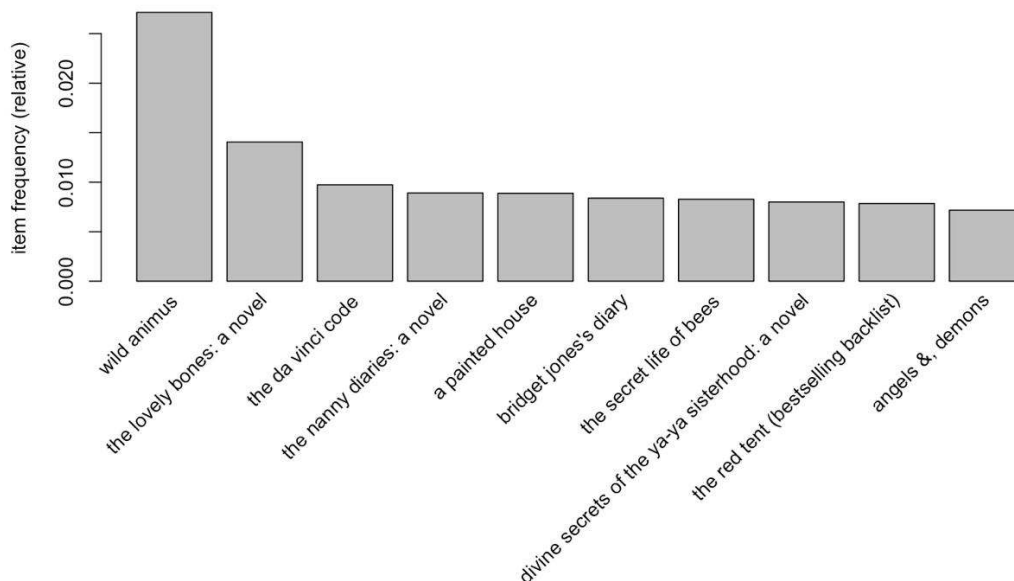
قوانین انجمنی با نرم افزار

R

Mohammad Hossein Heidary

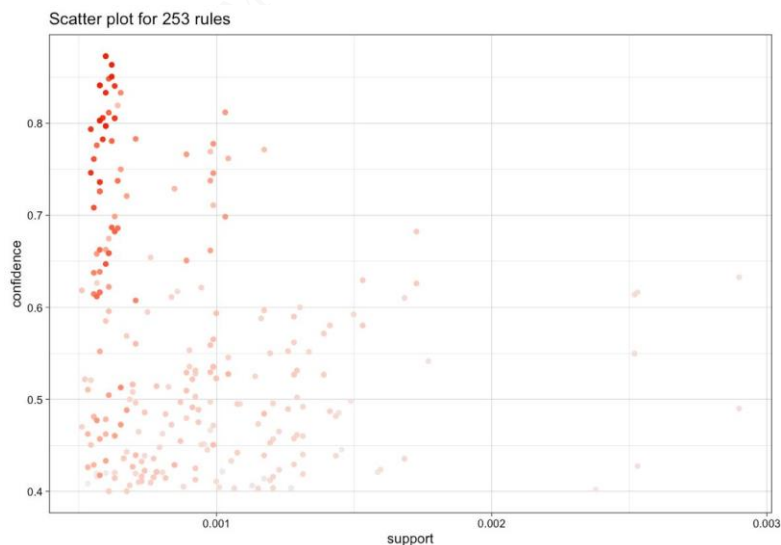
فراوانی کتاب‌های پرتکرار:

کتاب Wild Animus بیشترین حضور را در میان تراکنش‌ها دارد و اختلاف قابل توجهی با سایر کتاب‌ها نشان می‌دهد. بعد از آن، The Da Vinci Code و The Lovely Bones بیشترین تکرار را دارند.



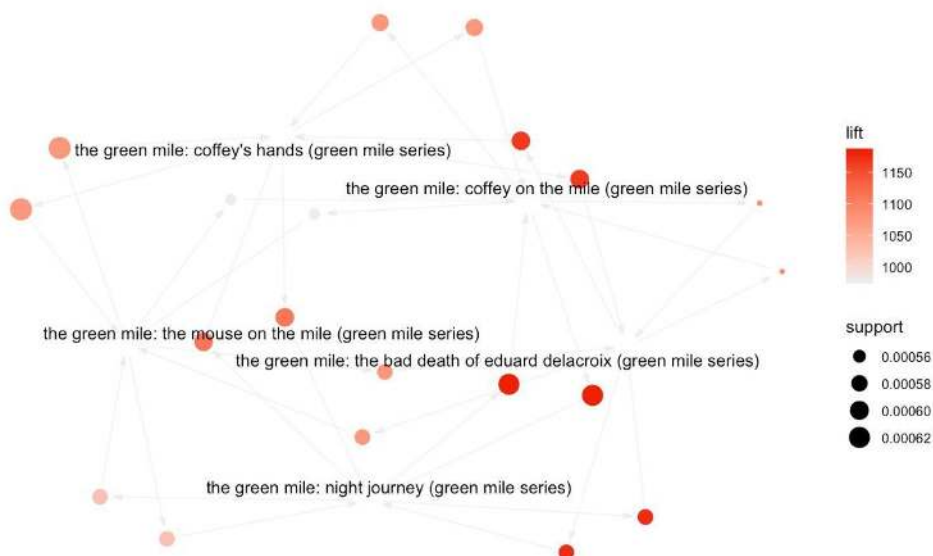
نمودار پراکندگی قوانین استخراج‌شده:

این نمودار رابطه‌ی بین support و confidence را نشان می‌دهد و رنگ نقاط بیانگر میزان lift است. بیشتر قوانین در محدوده‌ی ساپورت پایین (کمتر از ۰.۰۰۱) اما با confidence نسبتاً بالا (بالای ۰.۶) قرار دارند. همچنین، نقاط قرمز پررنگ بیانگر قوانین با lift بسیار بالا هستند که ارزش تحلیلی بیشتری دارند.



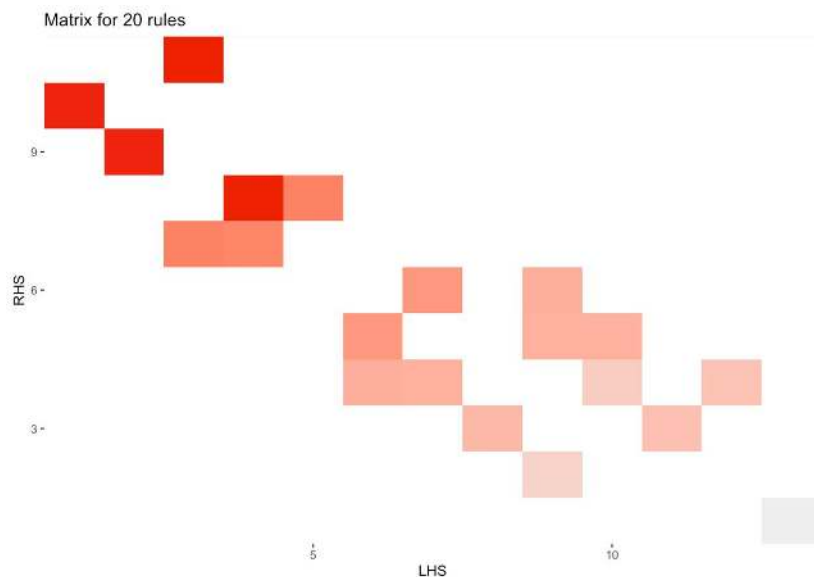
شبکه‌ی ارتباطی قوانین با بیشترین Lift (۲۰ قانون برتر):

این گراف ارتباط بین آیتم‌ها را بر اساس قوانین با بالاترین lift نشان می‌دهد. همانطور که دیده می‌شود، مجموعه کتاب‌های Green Mile series (مانند coffey on the mile, coffey's hands, night journey و ...) بیشترین ارتباط را با یکدیگر دارند. رنگ پررنگ‌تر نشان‌دهنده‌ی lift بالاتر و اندازه‌ی بزرگ‌تر گره‌ها بیانگر support بیشتر است. این نتایج نشان می‌دهد که کاربران تمایل دارند بخش‌های مختلف یک مجموعه (سری Green Mile) را همراه با هم بخوانند.



ماتریس قوانین منتخب (۲۰ قانون برتر بر اساس Lift)

این ماتریس رابطه‌ی بین بخش چپ (LHS) و بخش راست (RHS) قوانین را نشان می‌دهد. شدت رنگ قرمز بیانگر میزان lift بالاتر است. مشاهده می‌شود که تعدادی از آیتم‌ها (مانند کتاب‌های مجموعه Green Mile) در چندین قانون تکرار شده‌اند و ارتباط بسیار قوی بین آن‌ها وجود دارد. این نمودار کمک می‌کند سریعاً قوانین قوی و آیتم‌های پرتکرار در سمت چپ و راست قوانین شناسایی شوند.

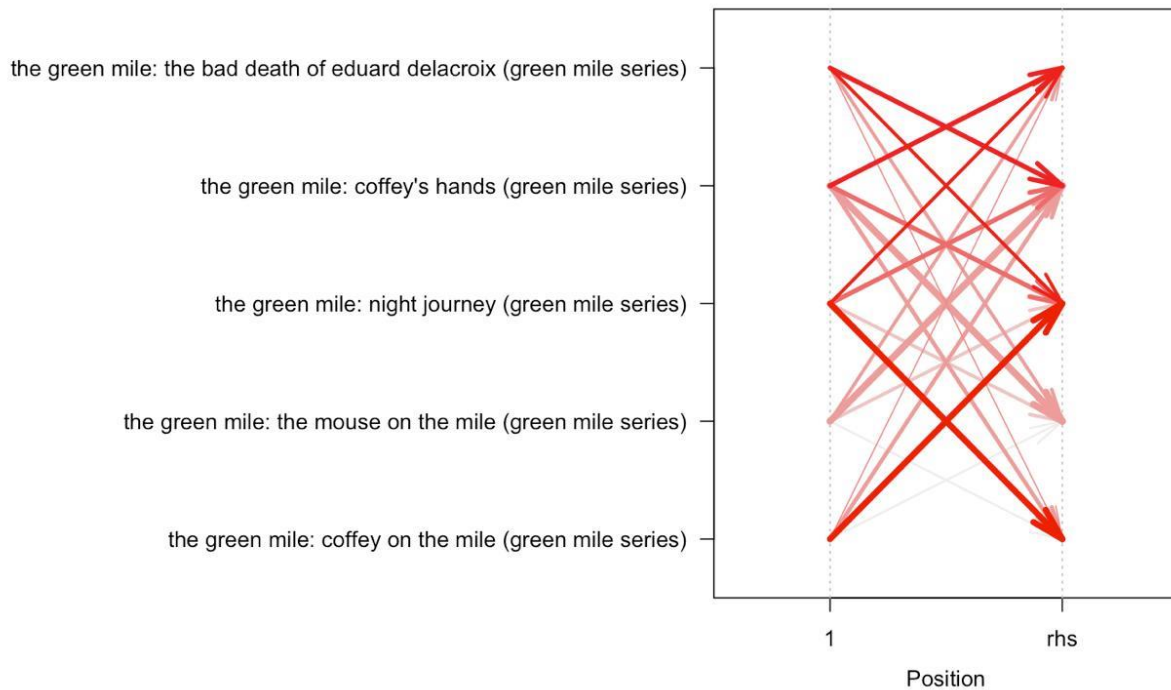


نمودار مختصات موازی (۲۰ قانون برتر بر اساس Lift)

این نمودار ارتباط بین بخش چپ (LHS) و بخش راست (RHS) قوانین را نمایش می‌دهد. هر خط نشان‌دهنده یک قانون است و ضخامت و رنگ آن قدرت قانون (بر اساس lift) را نشان می‌دهد.

همان‌طور که دیده می‌شود، تقریباً تمام قوانین پر قدرت مربوط به مجموعه Green Mile هستند و کتاب‌های این مجموعه به شکل متقابل (mutual) با هم قوانین قوی ساخته‌اند.

Parallel coordinates plot for 20 rules



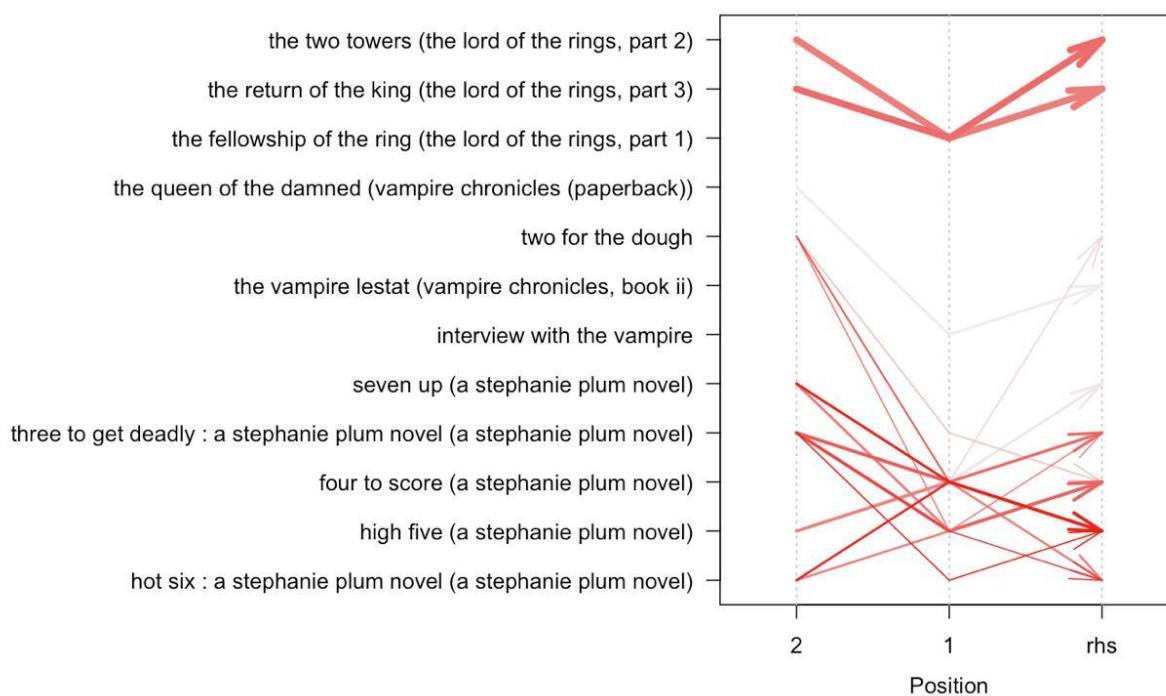
نمودار مختصات موازی (۲۰ قانون سه‌تایی برتر بر اساس Lift)

در این نمودار قوانین سه‌تایی نمایش داده شده‌اند. همان‌طور که مشخص است، الگوهای قوی در سه مجموعه اصلی دیده می‌شوند:

- مجموعه‌ی Lord of the Rings (سه‌گانه: Fellowship, Two Towers, Return of the King)
- مجموعه‌ی Vampire Chronicles (Interview with the Vampire, The Vampire Lestat, Queen of the Damned)
- مجموعه‌ی Stephanie Plum Novels (مانند High Five, Hot Six, Four to Score و غیره)

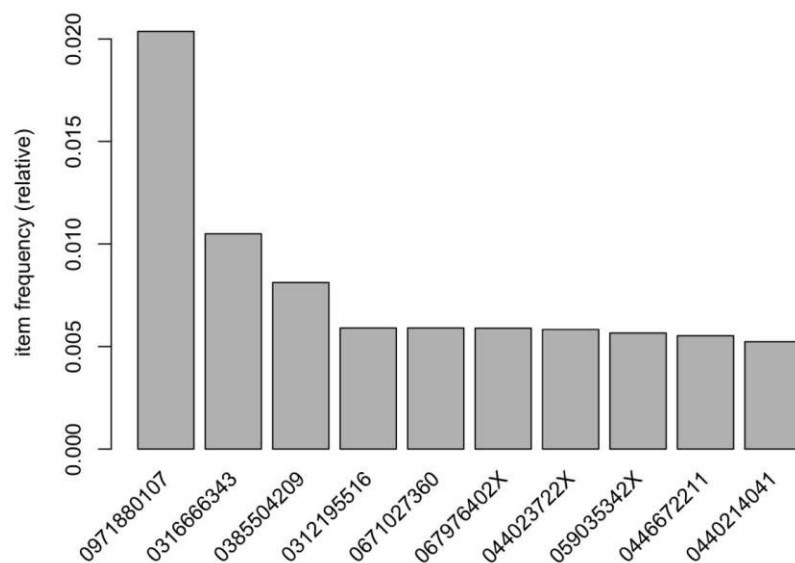
✦ خطوط پررنگ‌تر و قرمزتر نشان‌دهنده‌ی قوانینی با Lift بالاتر هستند. این موضوع بیانگر آن است که خوانندگان معمولاً این مجموعه‌ها را به صورت پشت سر هم یا در ترکیب با هم مطالعه کرده‌اند.

Parallel coordinates plot for 20 rules



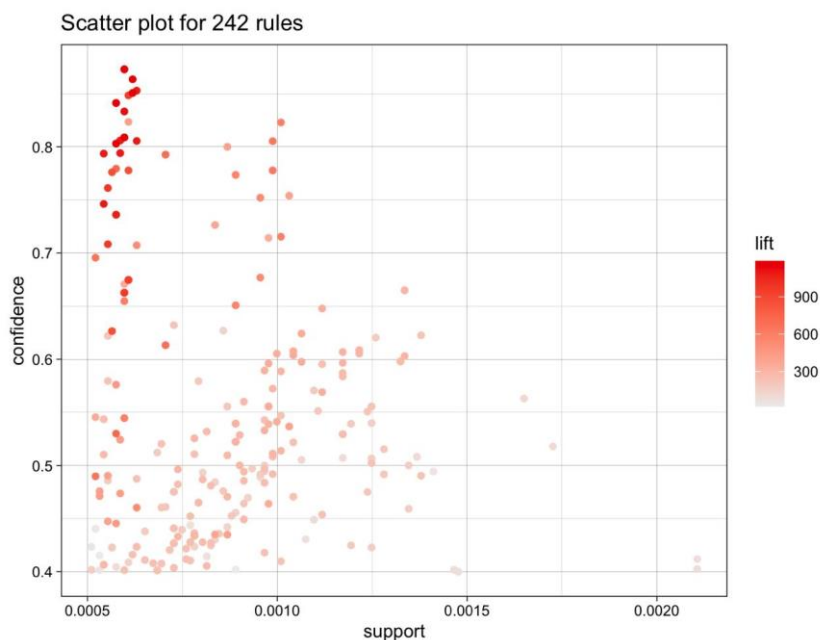
فراوانی کتاب‌ها بر اساس ISBN

کتاب با ISBN 0971880107 پرتکرارترین آیتم است (حدود ۲٪ کل تراکنش‌ها). پس از آن، ISBN‌های 0316666343 و 0385504209 بیشترین تکرار را دارند. سایر اقلام سهم کمتری دارند.



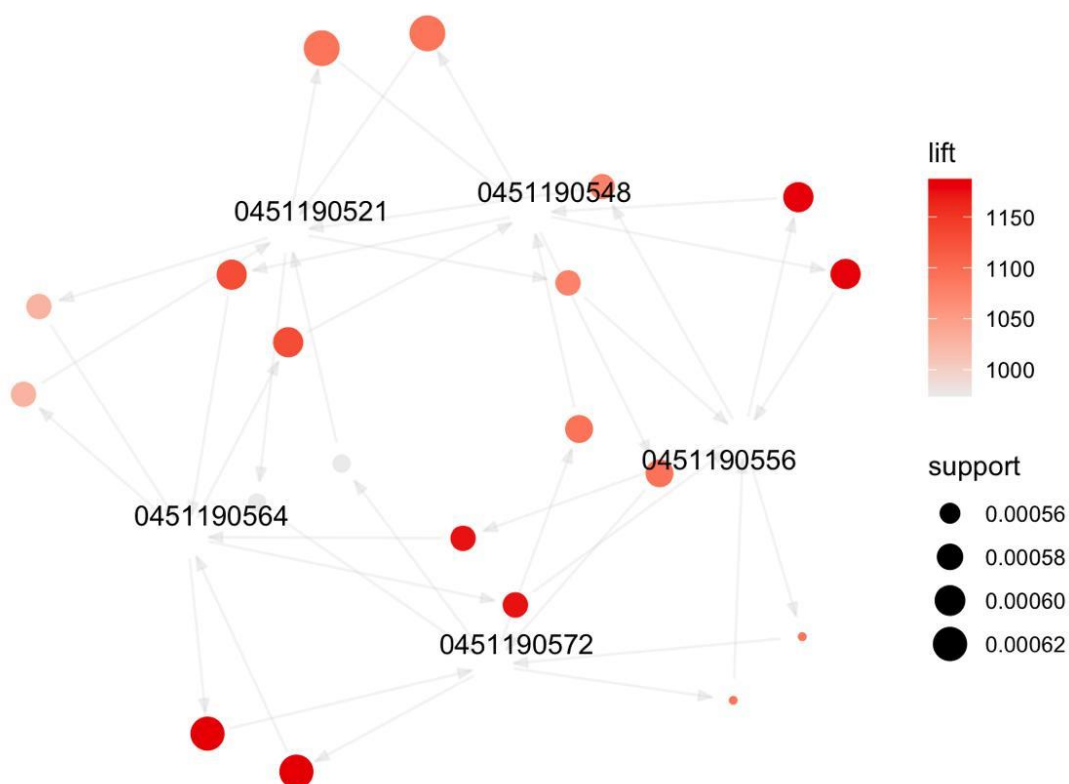
پراکندگی قوانین (Confidence و Support)

در این نمودار ۲۴۲ قانون نمایش داده شده است. بیشتر قوانین دارای پشتیبانی کم (support زیر ۰.۰۰۱) اما اطمینان نسبتاً بالا (بیش از ۰.۷) هستند. همچنین نقاط پررنگ‌تر (lift بالاتر) نشان می‌دهند که برخی روابط با وجود پشتیبانی کم، قدرت پیش‌بینی بالایی دارند.



نمودار شبکه‌ای قوانین

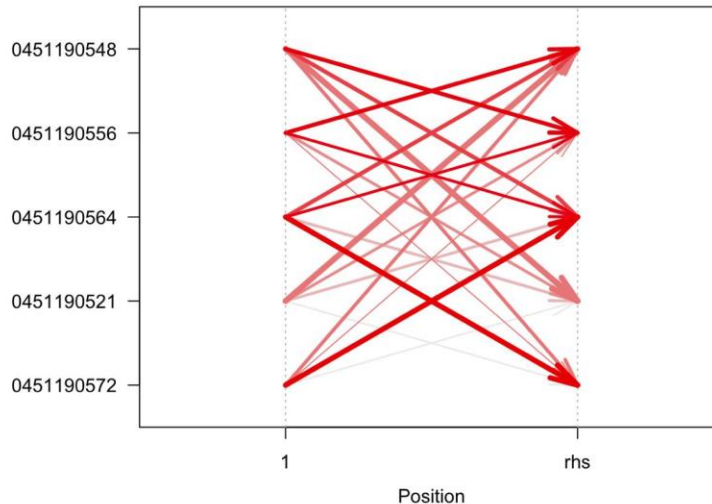
در این نمودار قوانین بر اساس ISBN نمایش داده شده‌اند. گره‌ها (دایره‌ها) کتاب‌ها را نشان می‌دهند و یال‌ها (فلش‌ها) ارتباط بین آن‌ها را مشخص می‌کنند. رنگ قرمزتر بیانگر میزان بالاتر lift و اندازه بزرگ‌تر نشان‌دهنده پشتیبانی بیشتر است. همان‌طور که مشاهده می‌شود، چند ISBN خاص (مانند 0451190521 و 0451190548) بیشترین ارتباط و قدرت پیش‌بینی را در میان قوانین دارند.



نمودار مختصات موازی قوانین (دو تایی ها)

این نمودار ۲۰ قانون برتر بر اساس ISBN را نمایش می‌دهد. خطوط از ستون سمت چپ (LHS) به ستون سمت راست (RHS) کشیده شده‌اند و مسیر هر خط یک قانون انجمنی را نشان می‌دهد. رنگ قرمزتر نشان‌دهنده میزان lift بالاتر است و ضخامت بیشتر خطوط به معنای قدرت قوی‌تر ارتباط بین آیتم‌هاست. همان‌طور که مشخص است، ISBN‌های سری 0451190 بیشترین همبستگی را با یکدیگر دارند و در قالب قوانین تکرار شده‌اند.

Parallel coordinates plot for 20 rules



نمودار مختصات موازی قوانین (سه تایی ها)

این نمودار ۱۲ قانون سه‌آیتمی برتر را نشان می‌دهد. خطوط ارتباط بین مجموعه‌های سه‌تایی ISBN در سمت چپ با نتیجه (RHS) در سمت راست را ترسیم می‌کنند. شدت رنگ قرمز نشان‌دهنده میزان lift بالاتر است. همان‌طور که دیده می‌شود، چند گروه ISBN (مانند سری xxx031297) بیشترین ارتباط و همبستگی را با یکدیگر دارند و به‌صورت مکرر در قوانین تکرار شده‌اند.

Parallel coordinates plot for 12 rules

