Data science- Recommender system for Big basket

Sadaf Fatollahy

۱۵ دی ۱۴۰۲

نام استاد: دکتر خردپیشه
نام درس : مبانی علوم داده

فهرست مطالب

۱.۰ مقدمه

یک سیستم توصیه توسعه یافته به کسب و کارها کمک می کند تا تجربه خریدار خود را در وب سایت بهبود بخشند و منجر به جذب و حفظ مشتری بهتر شود.Bigbasket بزرگترین سوپر مارکت مواد غذایی آنلاین در هند است. در سال ۲۰۱۱ راه اندازی شد و از آن زمان آنها در حال گسترش کسب و کار خود بوده اند. اگرچه برخی از رقبای جدید توانسته اند پای خود را در کشور بگذارند مانند Blinkit و غیره، اما BigBasket هنوز چیزی را از دست نداده است

۲.۰ درباره دیتاست

این مجموعه داده شامل ۱۰ ویژگی با معنی ساده است که به شرح زیر است:

۱. index :ایندکس

۲. product : نام محصول

۳. categor :دسته ای که محصول در آن طبقه بندی شده است

۴. sub-category : زیر شاخه ای که محصول در آن نگهداری شده است

۵. brand : نام تجاری محصول

۶. sale-price: قیمتی که محصول با آن در سایت فروخته می شود

۷. market-price:قیمت بازار محصول

۸. type: نوع محصول در آن قرار می گیرد

۹. rating: رتبه بندی محصول از مصرف کنندگان خود دریافت کرده است

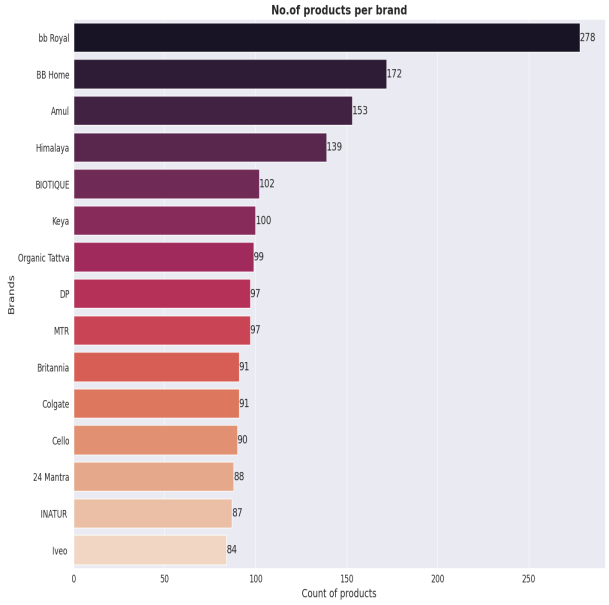
۱۰. description : شرح مجموعه داده (به تفصیل)

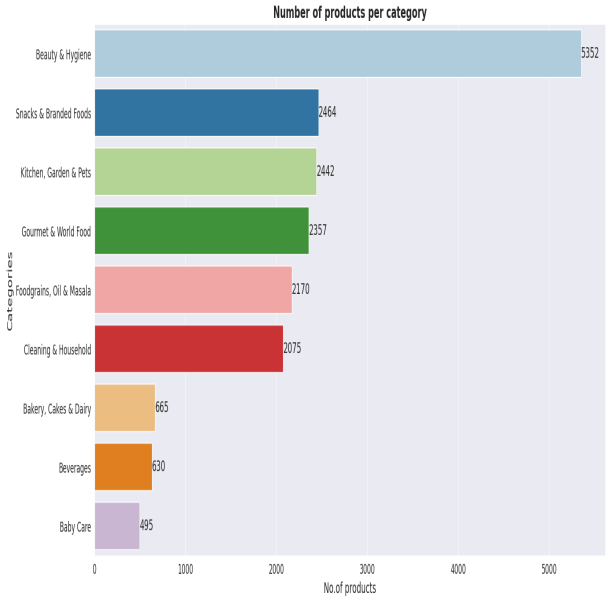
۳.۰ EDA and feature engineering

: در ابتدا در دیتاست کمی گشت و گذار کرده و به بررسی فیچر ها میپردازیم.

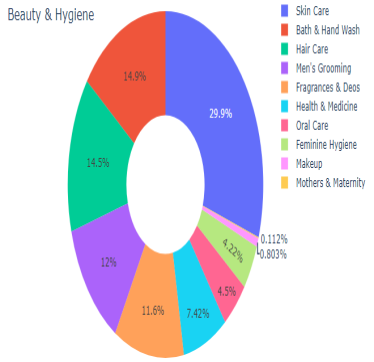
- این دیتاست شامل ۲۷۵۵۵ سطر و ۱۰ ستون است
- دیتاست در فیچر rating،description،product و brand دارای missing value است که بیشترین مربوط به rating است. به از انجا که تعداد انها خیلی زیاد نبود انها را حذف کردیم.
- دیتاست ۱۹۰ داده تکراری داشت که انها را نیز حذف کردیم و درنهایت ابعاد دیتاست به ۱۸۶۵۰ سطر و ۹ ستون(ستون ایندکس را حذف کردیم) تبدیل شد.

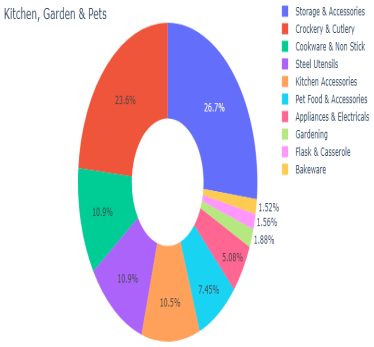
- با توجه به نمودار های زیر بیشترین محصول در این دیتاست مربوط به دسته Beauty - Hygiene و در زیر شاخه مراقبت پوستی Skin Care و برند bb Royal و نوع face care است.

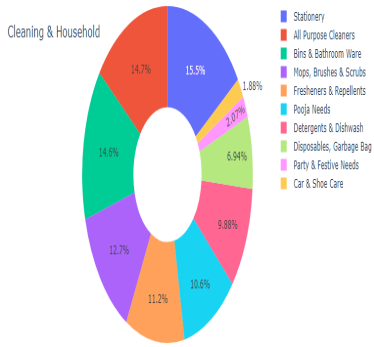




نمودار های زیر ۵ دسته برتر با تعداد محصولات زیر دسته هر یک را نشان میدهد.





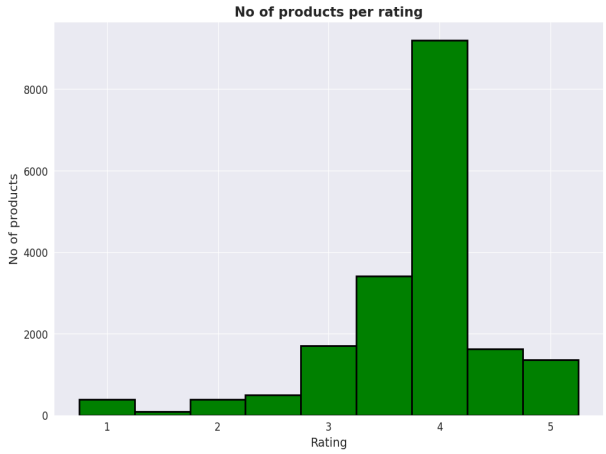


- لیست گران ترین و ارزان ترین ۱۰ محصول را به دست اوردیم. که گران ترین مربوط به عطر برند کارولینا هررا است. و ارزان ترین مربوط به سرم مو برند livon است.

- لیست محصولاتی که بیشترین نمره را گرفته اند بدست اوردیم.

	brand	product	type	rating
12	Oxy	Face Wash - Oil Control, Active	Face Care	5.0
14555	INATUR	Argan Oil Shampoo	Shampoo & Conditioner	5.0
14736	Floh	Tampons - Regular	Tampons & Menstrual Cups	5.0
14734	Jivika Naturals	Natural Jaggery Granules	Exotic Sugar & Salt	5.0
14729	Himalaya	Extra Moisturizing Baby Soap	Baby Bath	5.0
...	...	...	...	...
21352	Nyassa	Rose & Lychee Soap	Bathing Bars & Soaps	4.5
15680	Fogg	Master Body Spray - Cedar	Men's Deodorants	4.5
13682	Morpheme Remedies	Castor Carrier Oil - Pure Coldpressed Oil	Aromatherapy	4.5

- در شکل زیر توزیع rating را بررسی کردیم که همانطور که مشخص است بیشتر نمره دهی ۴ بوده که بیانگر کیفیت تقریبا خوب محصولات عرضه شده میباشد.



۴.۰ Content based

در ابتدا سعی کردیم از ویژگی های دیگر مانند دسته، زیر رده، نام تجاری، نوع و توضیحات برای توصیه بهتر استفاده کنیم. ما در اینجا از NLP برای استخراج اطلاعات مفید از ویژگی ها به خصوص توضیحات استفاده خواهیم کرد.

TF-IDF یک آماره عددی است که اغلب در وظایف پردازش زبان طبیعی برای نشان دادن اهمیت یک term در یک متن یا مجموعه ای از متون استفاده می شود. هدف آن به دست آوردن اهمیت یک term با در نظر گرفتن فراوانی آن در یک متن و نادر بودن آن در کل مجموعه است.

فیلترینگ مبتنی بر محتوا Content based روشی محبوب است که در سیستم های توصیه گر برای ارائه توصیه های شخصی به کاربران استفاده می شود. برای درک ویژگی‌های آنها و ارائه توصیه‌هایی بر اساس ترجیحات کاربر، بر تجزیه و تحلیل محتوا یا ویژگی‌های اقلام (محصولات، فیلم‌ها، مقالات و غیره) متکی است.

به این منظور ما پس از محاسبه ماتریس TF-IDF سپس وسپس با استفاده از Cosine similarity شباهت بین ایتم ها را بررسی کردیم

برای تولید توصیه‌ها، یک سیستم توصیه گر ساختیم که مواردی را شناسایی می‌کند که بر اساس ویژگی‌های محتوا مشابه مشخصات کاربر هستند. مواردی را که دارای امتیاز شباهت بالایی با نمایه کاربر هستند انتخاب می کند و به کاربر پیشنهاد می کند. سپس از ان خواستیم تا ۵ ایتم مشابه ورودی را برای ما تولید کند.

	Title	Rating
0	Rectangular Plastic Container - With Lid, Mult...	3.0
1	Jar - With Lid, Yellow	3.7
2	Round & Flat Storage Container - With lid, Green	4.6
3	Premium Rectangular Plastic Container With Lid...	3.6
4	Premium Round Plastic Container With Lid - Pink	3.6

	Title	Rating
0	Cadbury Perk - Chocolate Bar	4.2
1	Choco Stick - Hexagon Pack	4.4
2	Luvit Chocwich White Home Delights 187 g	4.1
3	Luvit Chocwich Home Delights 187 g	3.9
4	Wafer Biscuits - Chocolate Flavor	4.2