گزارش داده های دیجی کالا

محمد دهقانی
@data_hub_ir
linkedin.com/in/mdehqani

فایل "Digi_2.xlsx" شامل 100 هزار کامنت درباره محصولات مختلف است.

داده ها شامل 12 ستون است :

,'product_id', 'product_title', 'title_en', 'user_id', 'likes'
,'dislikes', 'verification_status', 'recommend', 'title', 'comment'
'advantages', 'disadvantages'

در ادامه هر ستون بررسی می شود:

product_id

هر کامنت درباره یک محصول است و هر محصول یک شناسه یکتا دارد.

همین طور هر محصول یک عنوان دارد.(**product_title**)

در جدول زیر محصولاتی هستند که بیشترین تعداد کامنت را به خود اختصاص داده اند.

محصول	تعداد نظرات
700304	198
180451	172
319878	150
111178	139
643764	131

user_id

هرکامنت توسط یک کاربر با شناسه یکتا نوشته می شود.

در جدول زیر کاربرانی هستند که بیشترین تعداد کامنت را داده اند.

کاربر	تعداد نظرات
764992	265

1152044	182
4964994	154
2870189	97
535912	94

title_en

مقادیر مربوط به این ستون از قرار زیر است:

'IT', 'AC', 'HW', 'MO', 'PC', 'PA', 'TC', 'TS', 'MA', 'HA', 'AV',
'FA', 'HC', 'BC', 'DF', 'GC', 'GF', 'FF'

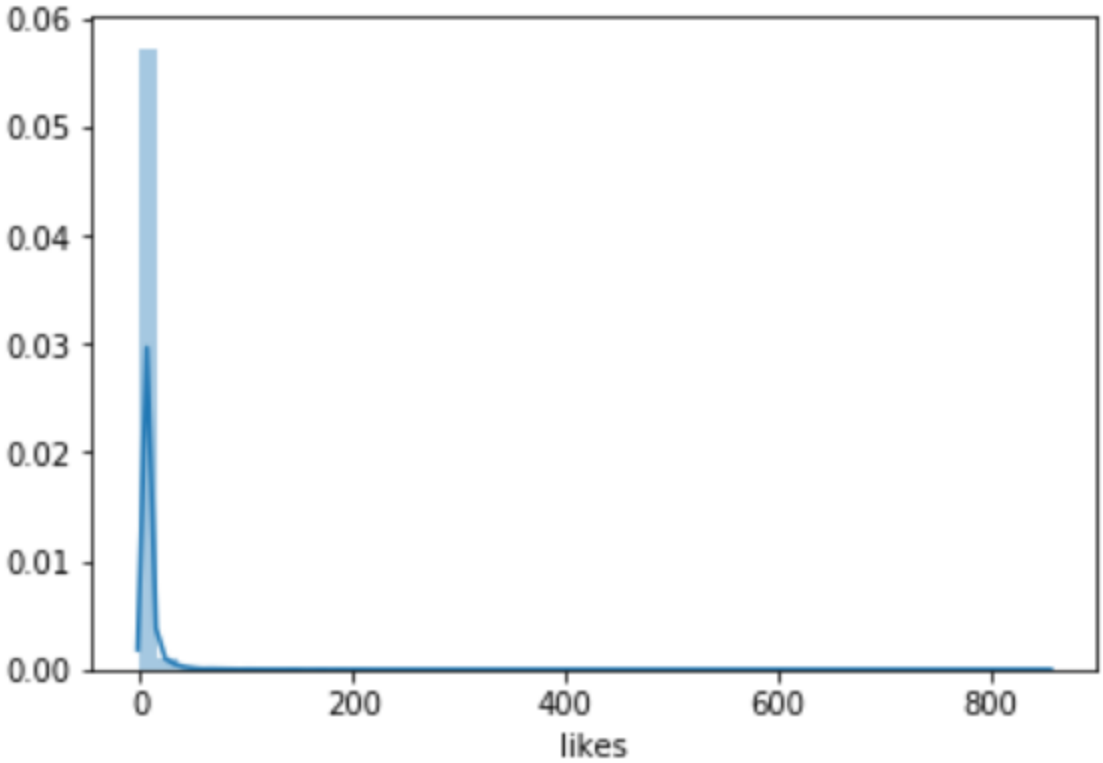
Dislikes & Likes

وقتی کاربران زیر یک کالا در سایت کامنت می گذارند کاربران دیگر می توانند کامنت های همدیگر را لایک یا دیس لایک کنند.

برای ستون لایک حداقل، حداکثر، میانگین و انحراف از معیار به صورت زیر است:

mean	3.114790
std	8.050031
min	0.000000
25%	0.000000
50%	1.000000
75%	4.000000
max	854.000000

نمودار توزیع مقادیر به صورت زیر است:



حال چون ۹۵ درصد از داده‌ها در فاصله‌ی ۲ برابریِ انحراف استاندارد نسبت به میانگین قرار دارند پس داده هایی با این شرایط(تعداد لایک بیشتر از 17) را انتخاب کرده و به عنوان very like در نظر می گیریم:

```
1  8.050031 * 2
```

```
16.100062
```

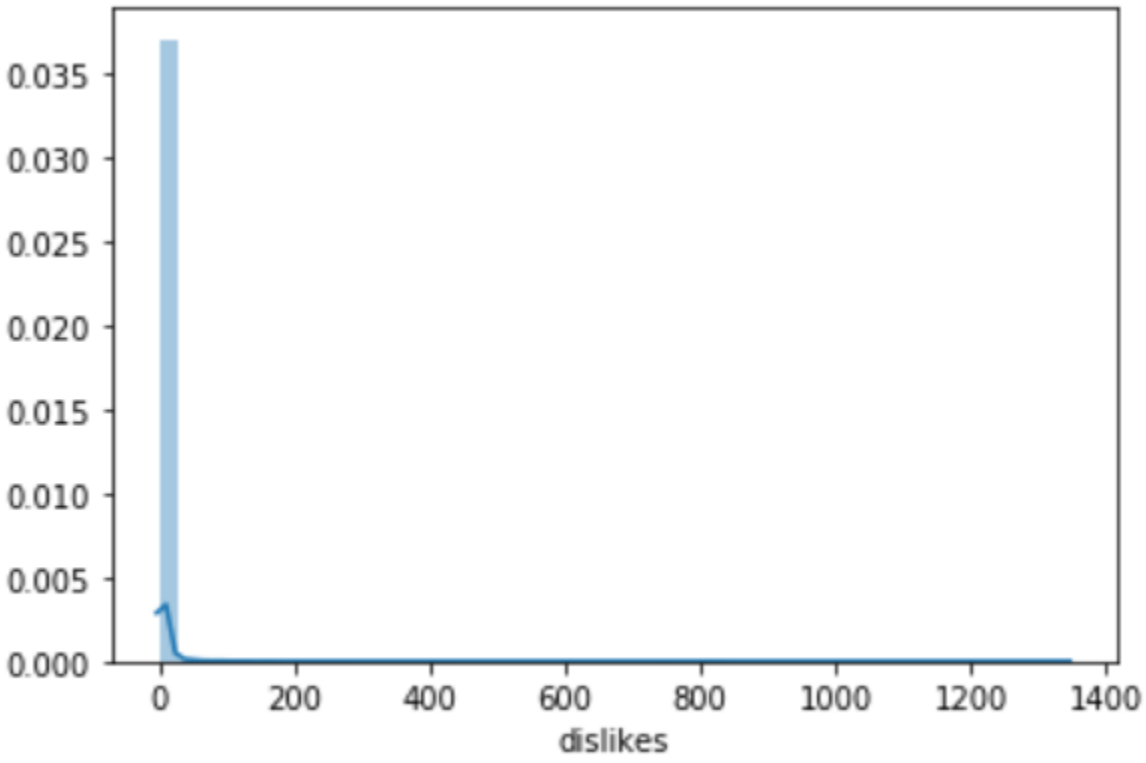
نتیجه به صورت زیر خواهد بود:

```
like          97740
very like      2260
```

برای ستون دیس لایک حداقل، حداکثر، میانگین و انحراف از معیار به صورت زیر است:

mean	1.649460
std	8.912705
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1344.000000

نمودار توزیع مقادیر به صورت زیر است:



حال چون ۹۵ درصد از داده‌ها در فاصله‌ی ۲ برابریِ انحراف استاندارد نسبت به میانگین قرار دارند پس داده هایی با این شرایط (تعداد دیس لایک بیشتر از 18)را انتخاب کرده و به عنوان very dislike در نظر می گیریم:

```
1 8.912705 * 2
```

```
17.82541
```

نتیجه به صورت زیر خواهد بود:

dislike	98855
very dislike	1145

verification_status

وقتی کاربر اقدام به نوشتن کامنت می کند ابتدا باید مدیر سایت یا یک ربات هوشنمد محتوا کامنت را بررسی کرده و سپس کامنت را منتشر کرده و بعد از آن کامنت برای عموم قابل مشاهده خواهد شد. در این سری داده، برای هر کامنت سه وضعیت در نظر گرفته شده است:

تایید شده(verified): محتوا کامنت از جنبه های مختلف بررسی و تائید شده است.

رد شده(rejected): محتوا کامنت از جنبه های مختلف بررسی و تائید نشده است.

بررسی نشده(not_verified): محتوا کامنت هنوز بررسی نشده است.

فراوانی مقادیر مربوط به این ستون به صورت زیر است:

verified	98496
rejected	1263
not_verified	241

همانطور که مشاهده می شود حدود 99 درصد از نظرات تایید شده اند.

recommend

هر کاربر بعد از نوشتن کامنت می تواند این محصول را به دیگران توصیه کند(recommended)، همینطور می تواند کاربران را توصیه به نخریدن یک محصول کند(not_recommended) یا می تواند به صورت خنثی عمل کرده و توصیه به خریدن یا نخریدن کند(no_idea) و در حالت اخر هیچ واکنشی نشان ندهد(unkown).

فراوانی مقادیر مربوط به این ستون به صورت زیر است:

recommended	36972
unkown	36382
not_recommended	16110
no_idea	10536

هر کامنت شامل 4 بخش است:
عنوان(title)
محتوای کامنت(comment)
نقاط قوت(advantages)
نقاط ضعف(disadvantages)

خروجی اولیه:

به کمک Tfid و اعمال آن روی ستون advantages و فقط برای داده هایی که very like بودند به کلمات کلیدی زیر رسیدیم:

ندارد	60.518357
کیفیت	48.722468
هیچی	31.688583
طراحی	27.729389
قیمت	27.083457
زیبا	24.098503
ظاهر	13.975147
ساخت	12.714116
جنس	9.540537
امکانات	9.381762
سرعت	9.109780
قدرت	9.092453
شیک	8.331080
کارایی	8.096549

تحلیل اولیه جدول این است که افراد نسبت به محصول کاملا ناراضی بوده و در قسمت مزایا از واژگانی مثل ندارد و هیچی استفاده کرده تا نشان دهند این محصول خوب نیست و دیگران با دیدن کامنت و هم نظربودن با لایک کردن موافقت خود را با این کامنت اعلام می کنند.

همینطور در جدول قابل مشاهده است که افراد معمولا به موجودیت هایی اهمیت بیشتری می دهند و دنبال چه نوع محصولی هستند. برای مثال موجودیت هایی از جمله ساخت، جنس، طراحی و... که ملاک های مهمی برای اینکه یک کاربر جهت مثبت نسبت به یک کالا بگیرد، دارند.

با تکرار Tfid و اعمال آن روی ستون disadvantages و فقط برای داده هایی که very dislike بودند به کلمات کلیدی زیر رسیدیم:

ندارد	35.935718
هیچی	30.091406
قیمت	17.620956
نداره	15.036665
ضعیف	14.068735
بالا	9.904784
گران	7.833080
پایین	7.759517
کیفیت	7.714786

تحلیل اولیه جدول: دو واژه اول کاملا معنادار هستند چون کسانی که از کمبود ویژگی ناراحت هستند کامنت گذاشته و با استفاده از واژگانی مثل ندارد، نداره از محصول انتقاد کرده مثل "جنس خوبی ندارد" و دیگران با دیدن چنین نظراتی تمایل به دیس لایک کردن پیدا می کنند زیرا آن ها محصول را دوست داشته و اعتقاد دارند این کامنت اشتباه است.

تحلیل دوم این است که افراد نسبت به محصول کاملا راضی بوده و در قسمت معایب از واژگانی مثل ندارد و نداره و هیچی استفاده کرده تا نشان دهند این محصول خوب و مناسب است.

همینطور افراد با کمک واژگانی مثل بالا، پایین، کم، ضعیف نظر منفی خود را نسبت به موجودیت ها ارائه داده و دیگران پیرو این روند، دیس لایک کرده و با کامنت مدنظر مخالفت می کنند.

نکته بسیار مهم: کاربران عنوان، متن، مزایا و معایب را خوانده و کل آن را لایک و دیس لایک می کند پس به صورت قطعی نمی توان گفت اگر زیاد لایک شده به خاطر کدام یکی از 4 بخش است.

نکته قابل توجه این است که بهتر بود در تحلیل‌ها وقتی تعداد لایک یا دیس‌لایک صفر بود عملیات جداگانه انجام می‌دادیم. یعنی بین کالایی که یک لایک گرفته و آن کالایی که هیچ لایکی نگرفته تفاوت قائل شویم.