Data science- The first report-Ṡpotify dataset

Sadaf Fatollahy

۶ آبان ۱۴۰۲

نام استاد: دکتر خردپیشه
نام درس : مبانی علوم داده

فهرست مطالب

۱.۰ مقدمه

آهنگ های محبوبی که از سال ۱۹۶۰ تا به امروز بر صحنه موسیقی مسلط بوده اند. این مجموعه داده بر اساس رتبه‌بندی‌های ARIA (انجمن صنعت ضبط استرالیا) و نمودارهای بیلبورد تنظیم شده است و از نمایش متنوع آهنگ‌هایی که به موفقیت تجاری و اهمیت فرهنگی بی‌نظیری دست یافته‌اند، اطمینان حاصل می‌کند.

مجموعه داده شامل ژانرهای مختلف موسیقی است و تکامل روندهای موسیقی را در طول سال ها به نمایش می گذارد و بینش های ارزشمندی را در مورد چشم انداز همیشه در حال تغییر موسیقی عامه پسند ارائه می دهد. این شامل آهنگ‌هایی از هنرمندان و گروه‌های نمادین است که ترکیبی از آثار کلاسیک جاودانه و آثار معاصر را نشان می‌دهد که تأثیری ماندگار بر دوستداران موسیقی در سراسر جهان گذاشته است.

در این تمرین ما قصد داریم ازمون های اماری را بر روی دو دیتاست بررسی کنیم دیتاست دوم مربوط به پیش بینی محبوبیت اهنگ است.در این دیتاست ابتدا مقداری فرایند های EDA و Preprocessing را انجام میدهیم ، بعد از ان تعدادی سوال را خودمان مطرح میکنیم و با ازمون های اماری به انها پاسخ میدهیم.

۲.۰ درباره دیتاست

هر فیچر نشان دهنده چیست؟

- Popularity :محبوبیت این متغیر هدفی است که می‌خواهید پیش‌بینی کنید.
- Track URL :لینک ورود به اهنگ
- Track Name:اسم اهنک
- Artist URL(s) : لینک ورود به پروفایل خواننده
- Artist Name(s) :نام خواننده
- Album URL:لینک ورود به البوم
- Album Name :اسم البوم
- Album Artist URI(s): لینک ورود به خواننده البوم
- Album Artist Name(s) : اسم خواننده البوم
- Album Release Date :تاریخ انتشار البوم
- Album Image URL : لینک عکس البوم
- Disc Number :شماره دیسک
- Track Number :شماره ترک
- Track Duration (ms)1 : مدت زمان ترک
- Track Preview URL2 : لینک پیشنمای (دمو) ترک
- Explicit : صراحت

- ISRC
- Added By
- Added By : تاریخ و زمان انتشار
- Artist Genres : ژانر خواننده
- Danceability : میزان رقص پذیری
- Energy: میزان انرژی
- Key
- Loudness : بلندی صدا
- Mode
- Speechiness : میزان گفتاری
- Acousticness : میزان اکوستیک بودن
- Instrumentalness : میزان ابزاری بودن
- Liveness :میزان سرزندگی
- Valence :میزان ظرفیت
- Tempo :میزان تمپو
- Time Signature : یک نت موسیقی است که سازماندهی ضربات را در یک قطعه موسیقی نشان می دهد.
- Album Genres :ژانر البوم
- Label1 : لیبل اهنگ
- Copyrights

۳.۰ EDA and pre proccesing

- این دیتاست شامل ۹۹۹۹ سطر و ۳۵ ستون فیچر است. در ابتدا تعدادی از فیچر ها که به انها نیازی نداریم را حذف میکنیم :

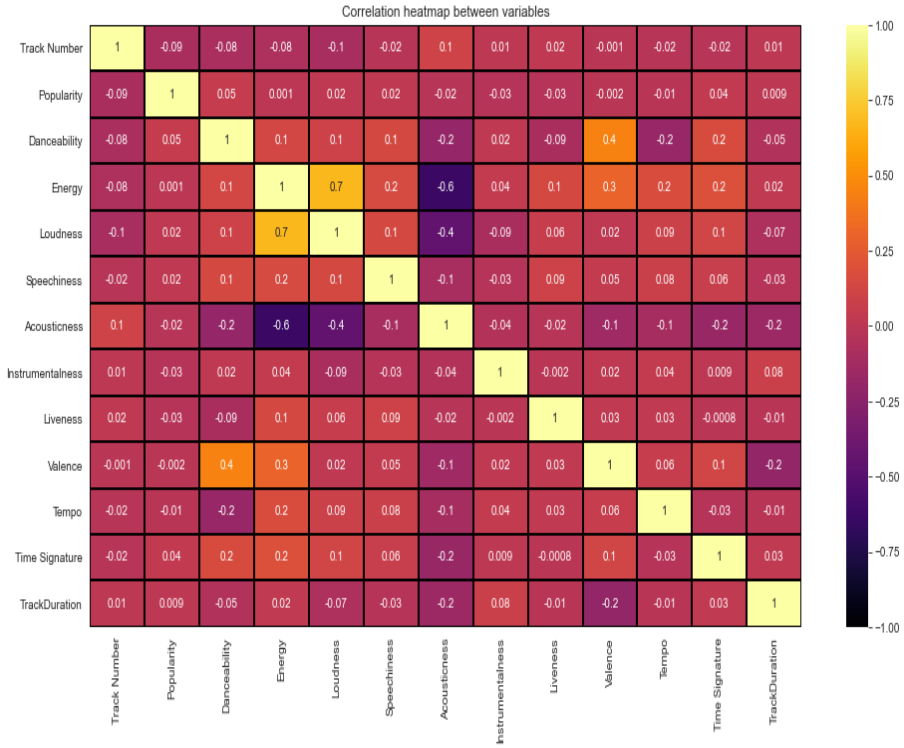
["Track URI", "Artist URI(s)", "Album URI", "Album Artist URI(s)", "Album Image URL", "Added By", "Added At", "Key", "Mode", "Explicit", "Album Genres", "ISRC", "Disc Number", "Track Preview URL"]

- حال میخواهیم missing value ها را بررسی کنیم. وقتی تعداد انها را محاسبه کردیم متوجه شدیم مقدار زیادی از دیتا را تشکیل نمیدهد بنابراین انها را پاک کردیم به این ترتیب مقدار زیادی دیتا از دست ندادیم.

- فیچر duration را که برحسب میلی ثانیه بود برای راحتی به ثانیه تبدیل کردیم.
- سپس بررسی کردیم که ایا داده تکراری وجود دارد یا خیر که متوجه شدیم ۴۸ مورد دیتا تکراری داریم و انها را پاک کردیم

با نگاه کردن به ستون ”تاریخ انتشار آلبوم“ باید همانطور که از نامش می گوید تاریخ باشد اما در حال حاضر از نوع Object است. به همین دلیل ان را به تاریخ تبدیل کردیم.

- سپس بین فیچر های عددی heatmap رسم کردیم.

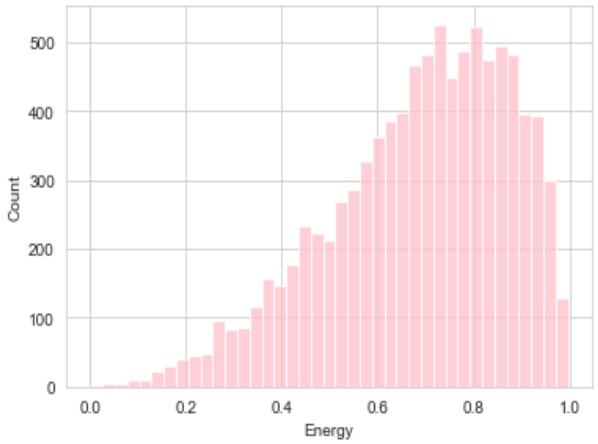


مشاهده میشود بین فیچر های energy و loudness همبستگی بالایی وجود دارد و بین فیچر های energy و Acousticness همبستگی منفی وجود دارد و این بسیار طبیعی است. وقتی صدای اهنگ بلند باشد انرژی بالاتری را منتقل میکند. همبستگی منفی بین آکوستیک و انرژی نشان می‌دهد که آهنگ‌های با آکوستیک بالا تمایل به سطوح انرژی پایین‌تری دارند و آهنگ‌های با آکوستیک پایین تمایل به سطوح انرژی بالاتری دارند. این نشان می‌دهد که با آکوستیک‌تر شدن مسیر (کمتر الکترونیکی) در طبیعت، سطح انرژی آن کاهش می‌یابد.

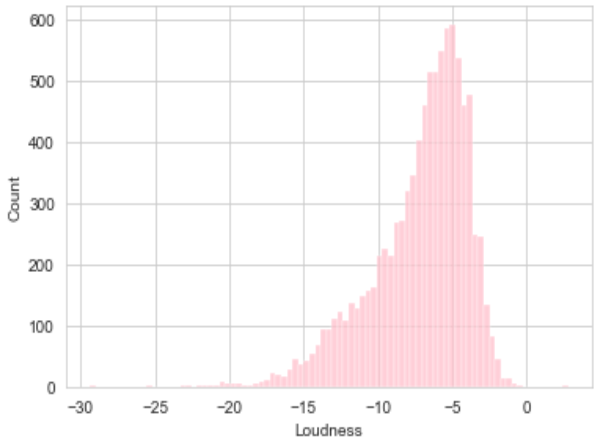
۴.۰ سوال

:

۱. آیا بین انرژی و بلندی صدا رابطه معناداری وجود دارد؟
در ابتدا نگاهی به این توزیع این فیچرها می اندازیم



شکل ۱: توزیع Energy



شکل ۲: توزیع Loudness

همانطورکه مشخص است این دو فیچر دارای چولگی هستند. برای اطمینان بیشتر از ازمون shapiro استفاده کردیم.

بررسی نرمال بودن دوتا فیچر ذکر شده با استفاده از ازمون shapiro:

- ازمون shapiro

آزمایش می کند که آیا یک نمونه داده توزیع گاوسی دارد یا خیر.

مفروضات:

مشاهدات در هر نمونه مستقل و به طور یکسان توزیع شده است (iid).

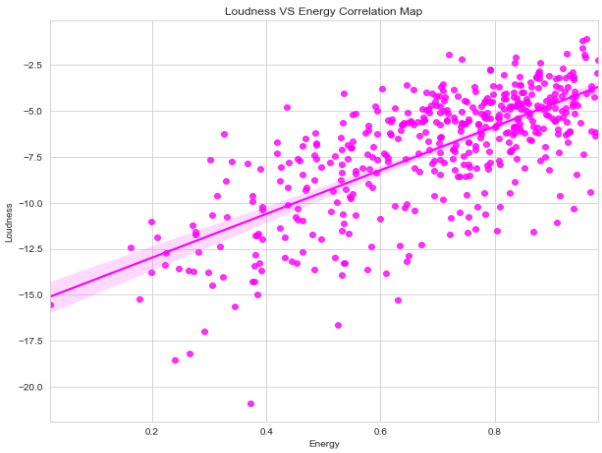
تفسیر:

H_0: نمونه دارای توزیع گاوسی است.

H_1: نمونه توزیع گاوسی ندارد.

نتیجه این ازمون به این صورت بود که هیچ کدام داری توزیع نرمال نیستند.

حال پلات زیر را رسم کردیم تا ببینیم رابطه خطی بین انها موجود هست یا نه.



شکل ۳: relation between energy and loudness

با توجه به شکل این ارتباط وجود دارد اما ایا رابطه معناداری بین انها از طریق ازمون فرض اثبات میشود؟

به این منظور از ازمون Spearman استفاده میکنیم. این ازمون زمانی استفاده میشود که دو نتغیر مورد بررسی عددی باشند و توزیع نرمال نداشته باشند.

- ازمون Spearman

آزمایش می کند که آیا یک نمونه داده به صورت یکپارچه قابل تفکیک است یا خیر.

مفروضات:

الف) مشاهدات در هر نمونه مستقل هستند و به طور یکسان توزیع می شوند.

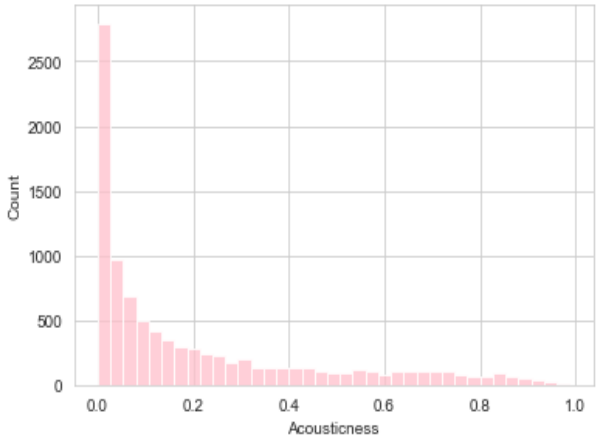
ب) مشاهدات در هر نمونه رتبه بندی می شوند.

تفسیر:

H_0:: نمونه ها همبستگی دارند.

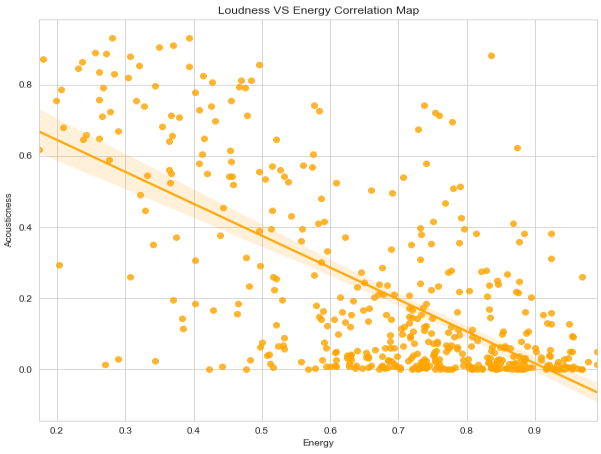
H_1: نمونه هیچ همبستگی ندارد.
این ازمون مشخص کرد که ارتباط بین این دو فیچر وجود دارد.

۲. آیا بین انرژی و اکوستیک بودن رابطه معناداری وجود دارد؟
به این منظور همانند قبل ابتدا توزیع فیچر Acousticness را بررسی کردیم که با توجه به شکل نرمال نیست و ازمون فرض هم همین را تایید میکند.



شکل ۴: توزیع Loudness

سپس پلات زیر را رسم کردیم تا ببینیم رابطه خطی بین انها موجود هست یا نه.



شکل ۵: relation between energy and acousticness

با توجه به شکل این ارتباط وجود دارد و رابطه عکس وجود دارد همانطور که پیش تر دلیل ان گفته شد و ازمون Spearman نیز ان را اثبات میکند.

۳. آیا بین محبوبیت و نام خواننده رابطه معناداری وجود دارد؟

به دلیل تنوع زیاد نام خواننده ها امکان ترسیم وجود ندارد به همین دلیل به ازمون فرض اکتفا میکنیم.

- ازمونkruskal

این ازمون ارزیابی می‌کند که آیا بین محبوبیت و نام خواننده رابطه معناداری وجود دارد بدون اینکه توزیع خاصی را فرض کنیم.
مفروضات:
الف)مشاهدات هر نمونه داده مستقل و توزیع می شود.
ب)مشاهدات را می توان رتبه بندی کرد.
تفسیر:
H_0: مرتبط هستند.
H_1: مرتبط نیستند..
این تست اثبات کرد که بین این دو فیچر ارتباط وجود دارد.

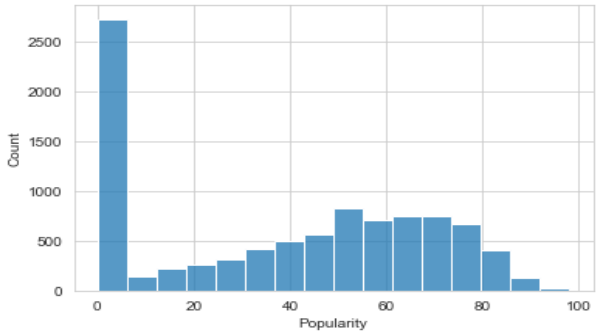
۴. آیا بین سرزندگی اهنگ و نام خواننده رابطه معناداری وجود دارد؟

انتظار داریم که این رابطه وجود داشته باشد زیرا تعداد از خوانندگان فقط اهنگ های شاد و بعضی بیشتر اهنگ های غمگین میخوانند . از ازمون kruskal استفاده کردیم و پاسخ ما را اثبات کرد.

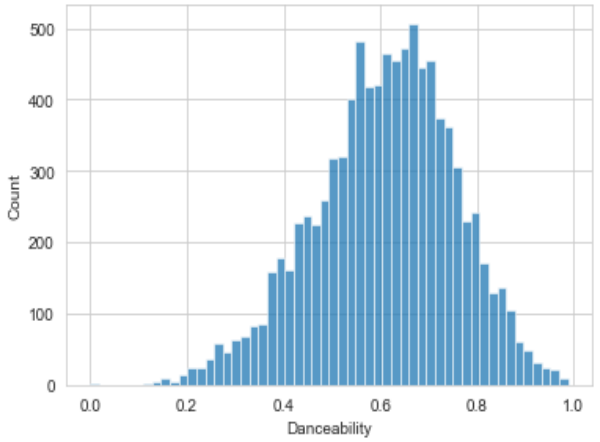
۵. آیا بین محبوبیت و و قابل رقص بودن اهنگ رابطه معناداری وجود دارد؟

در ابتدا نگاهی به این توزیع این فیچرها می اندازیم

از شکل مشخص است که توزیع نرمال نیست .حال بررسی میکنیم ایا رابطه خطی بین انها وجود دارد یا خیر

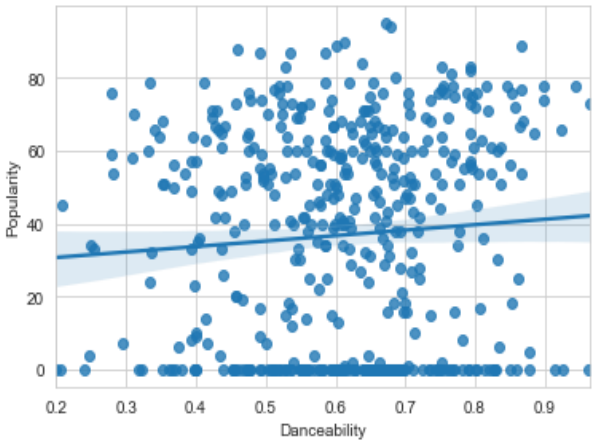


شکل ۶: توزیع popularity



شکل ۷: توزیع danceabilitu

با توجه به شکل رابطه خیلی قوی مشخص نیست . حال با ازمون Spearman انرا بررسی میکنیم . این ازمون بیان کرد که رابطه بین انها وجود دارد اما همانطور که مشخص است قوی نیست.



شکل ۸: relation between danceability and popularity

۶. آیا بین نام خواننده و و ژانر اهنگ رابطه معناداری وجود دارد؟

باتوجه به اینکه دو فیچر کتگوریکال هستند باید از ازمون کای دو استفاده کنیم و به دلیل تنوع اهنگ ها و ژانرها نمیتوانیم ان ها را رسم کنیم.

- ازمونChi-Squared

این ازمون ارزیابی می‌کند که آیا بین نام خواننده و و ژانر اهنگ رابطه رابطه معناداری وجود دارد.

مفروضات:

الف)مشاهدات استفاده شده در جدول مستقل هستند.

ب)بیش از ۲۵ نمونه در جدول وجود دارد. تفسیر:

H_0: مرتبط هستند.

H_1: مرتبط نیستند..

این تست اثبات کرد که بین این دو فیچر ارتباط وجود دارد.

درنهایت شکل زیر نشان میدهد که در سال های اخیر مدت زمان اهنگ ها کم شده است.