Ukladanie a príprava dát
Projekt č.2

Samuel Budai (xbudai01)
Kristián Kováč (xkovac61)
Martin Zmitko (xzmitk01)

Exploračná analýza

Popisné štatistiky:

Pre vypracovanie tejto úlohy sme si vybrali dátovú sadu najviac streamovaných piesní na Spotify za rok 2023. Dátová sada obsahuje informácie o 953 záznamoch piesní, kde každý záznam obsahuje 24 stĺpcov ako názov skladby, meno interpretov, dátum vydania, prítomnosť na štyroch hudobných platformách a rôzne zvukové vlastnosti. Okrem kvantitatívnych (numerických) atribútov, dátová sada obsahuje tiež kvalitatívne (kategorické) atribúty.

Po načítaní dát a skúmaní popisu stĺpcov sme zistili, že pri načítaní do DataFrame došlo k nesprávnemu určeniu niektorých stĺpcov ako kategorických, čo bolo spôsobené nekonzistentným formátom dát. Napríklad, čísla obsahovali čiarky, ktoré oddeľovali tisíce od stoviek, milióny od tisícok a podobne. Taktiež sme zistili chybu v riadku 575 v stĺpci "streams", kde boli namiesto čísla streamov zaznamenané nezmyselné dáta. Na korektnú analýzu sme preto museli previesť tieto číselné atribúty, pôvodne uložené ako objekty, na numerické typy. Vykonali sme túto transformáciu odstránením čiarok a nastavením chybných hodnôt na NaN.

Pri analýze sme zistili, že dátová sada obsahuje aj chýbajúce hodnoty:

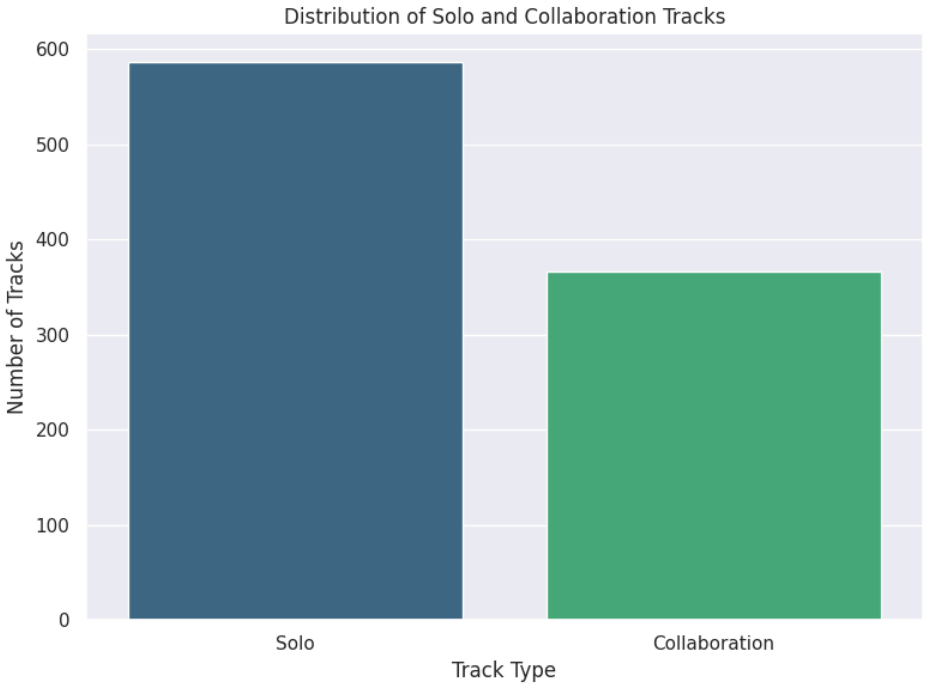
- Stĺpec "streams" má jednu chýbajúcu hodnotu
- Stĺpec "in_shazam_charts" má 50 chýbajúcich hodnôt
- Stĺpec "key" má 95 chýbajúcich hodnôt.

Z deskriptívnych štatistík a popisu stĺpcov môžeme pozorovať nasledovné:

Názov	Priemer	Modus	Popis
track_name	/	/	Názov piesne
artist(s)_name	/	Taylor Swift	Meno/mená umelcov
artist_count	1.56	1	Počet umelcov podieľajúcich sa na piesni
released_year	2018	2022	Rok vydania piesne
released_month	Jún	Január	Mesiac vydania piesne
released_day	14	1	Deň v mesiaci, kedy bola pieseň vydaná
in_spotify_playlists	5200	86	Počet Spotify playlistov, v ktorých je pieseň zaradená
in_spotify_charts	12.01	0	Prítomnosť a umiestnenie piesne v Spotify rebríčkoch
streams	514137400	156338600	Celkový počet streamov piesne na Spotify
in_apple_playlists	67.81	0	Počet Apple Music playlistov, v ktorých je pieseň zaradená

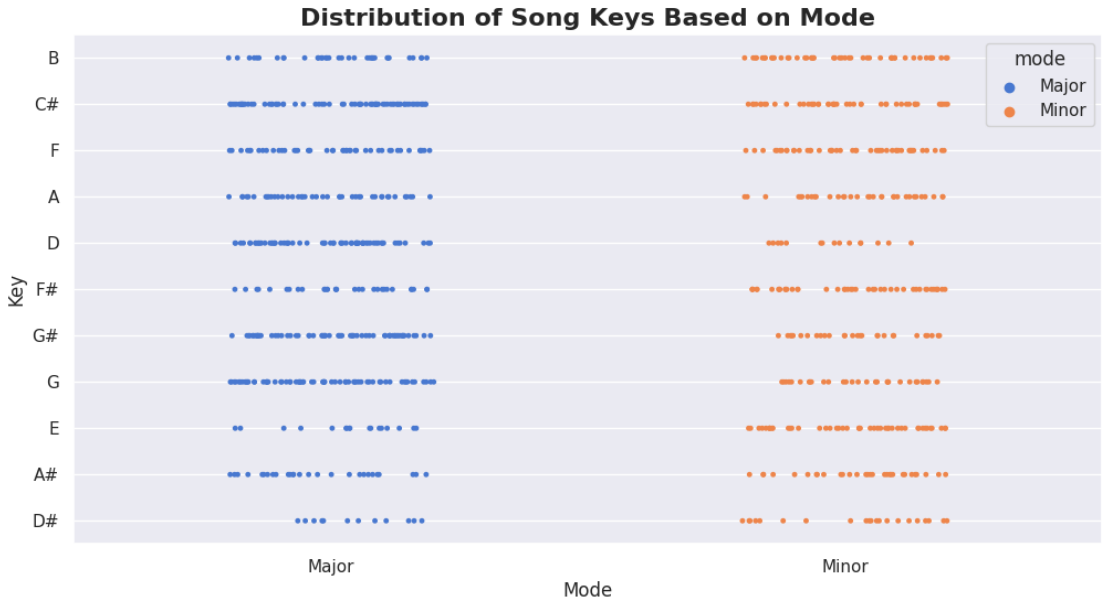
in_apple_charts	51.91	0	Prítomnosť a umiestnenie piesne na Apple Music rebríčkoch
in_deezer_playlists	385.19	0	Počet Deezer playlistov, v ktorých je pieseň zaradená
in_deezer_charts	2.67	0	Prítomnosť a umiestnenie piesne na Deezer rebríčkoch
in_shazam_charts	60.00	0	Prítomnosť a umiestnenie piesne na Shazam rebríčkoch
bpm	122.54	120	Miera tempa piesne
key	/	C#	Tónina piesne
mode	/	Major	Stupnica (mol, dur)
danceability_%	66.97	70	Vhodnosť skladby pre tanec
valence_%	51.43	24	Pozitiva obsahu skladby
energy_%	64.27	74	Miera energia skladby
acousticness_%	27.05	0	Množstvo akustického zvuku
instrumentalness_%	1.58	0.0	Množstvo inštrumentálneho zvuku
liveness_%	18.21	11	Obsah živého výstupu
speechiness_%	10.13	4	Množstvo slov v skladbe

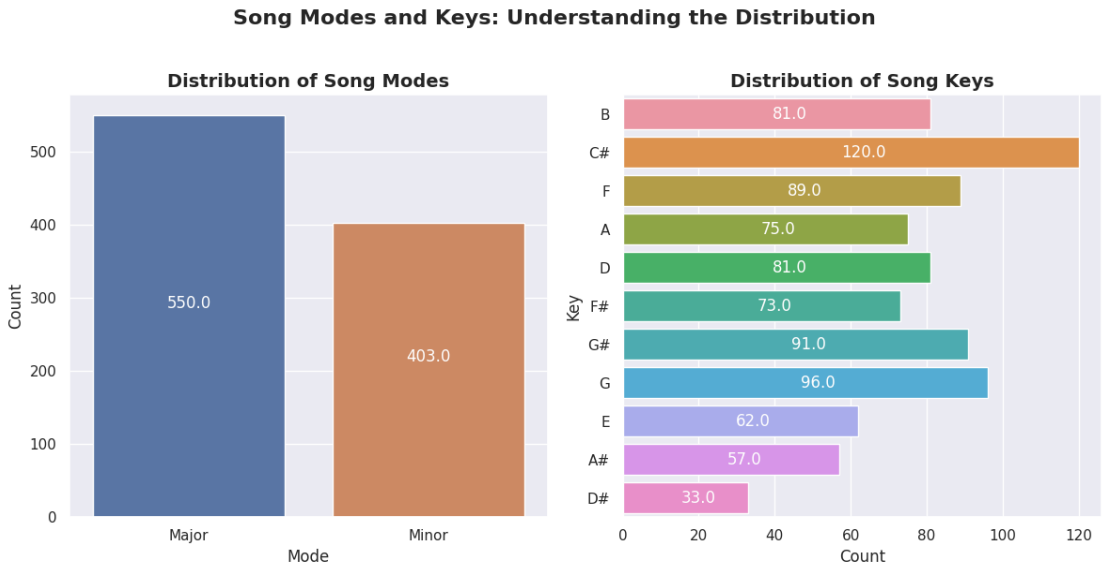
Skúmanie dát pomocou grafov:



Prvým grafom, ktorý určite stojí za zmienku, je graf zobrazujúci distribúciu piesní na základe počtu autorov, ktorí spolupracovali na jeho tvorbe. Tento graf sme rozdelili na dve kategórie. Prvou sú piesne, ktoré pochádzajú od jedného autora. Druhou sú piesne, ktoré

vznikli v rámci spolupráce dvoch a viacerých autorov. Z grafu môžeme pozorovať, že piesní, ktoré vznikli v dielni jedného autora je takmer 2x viac ako tých, ktoré obsahujú viacero autorov.

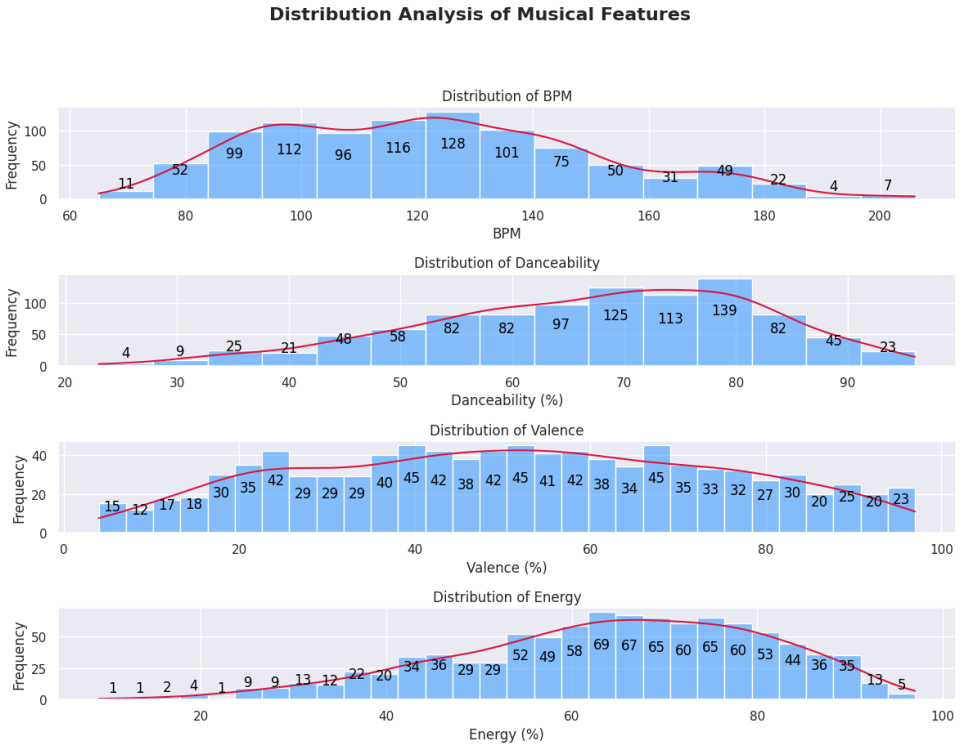


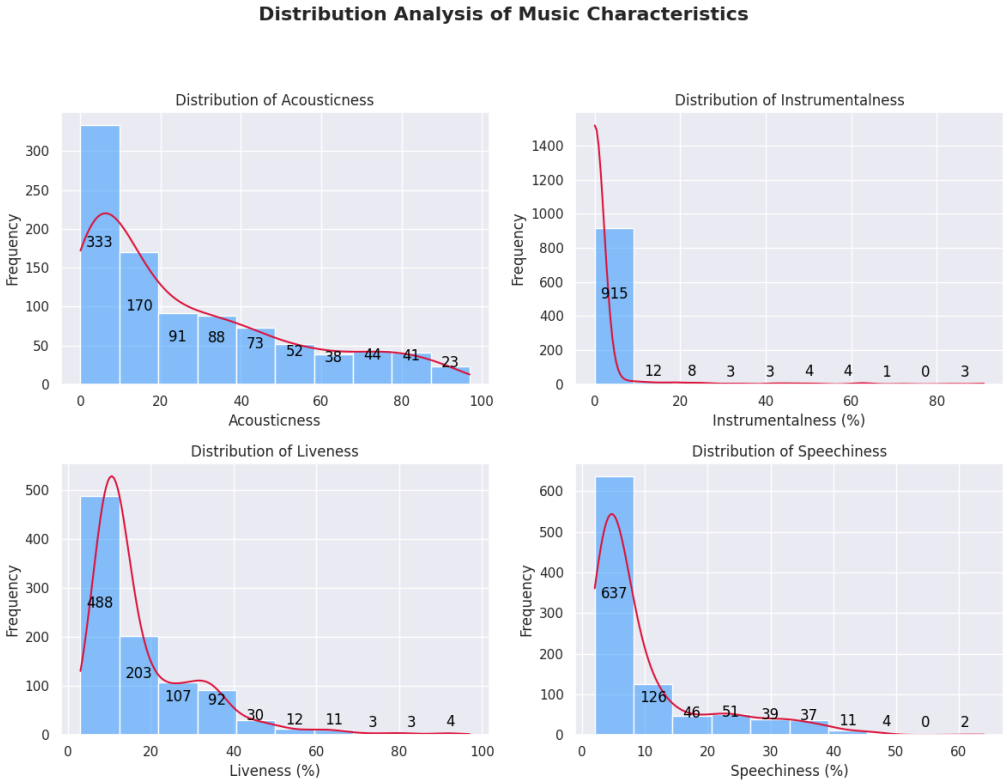


Na ďalších 3 grafoch, môžeme sledovať distribúciu delenia piesní podľa stupnice a tónin. Vidíme, že piesne ktoré boli nahraté v durovej stupnici prevláda. Vedľa je vykreslené zastúpenie tónin v dátovej sade. Na prvý pohľad je očividné, že dominuje C#, zatiaľ čo najmenej piesní je v D#. Prvý graf z tejto trojice, zobrazuje rozloženie tónin na základe stupnice.

Ďalšími veľmi zaujímavými charakteristikami našej dátovej sady, sú zvukové vlastnosti a charakteristiky. Konkrétne ich rozdelenie a vplyv na počet streamov. Na prvých dvoch grafoch, ktoré sa skladajú zo štyroch podgrafov, môžeme pozorovať rozloženie týchto

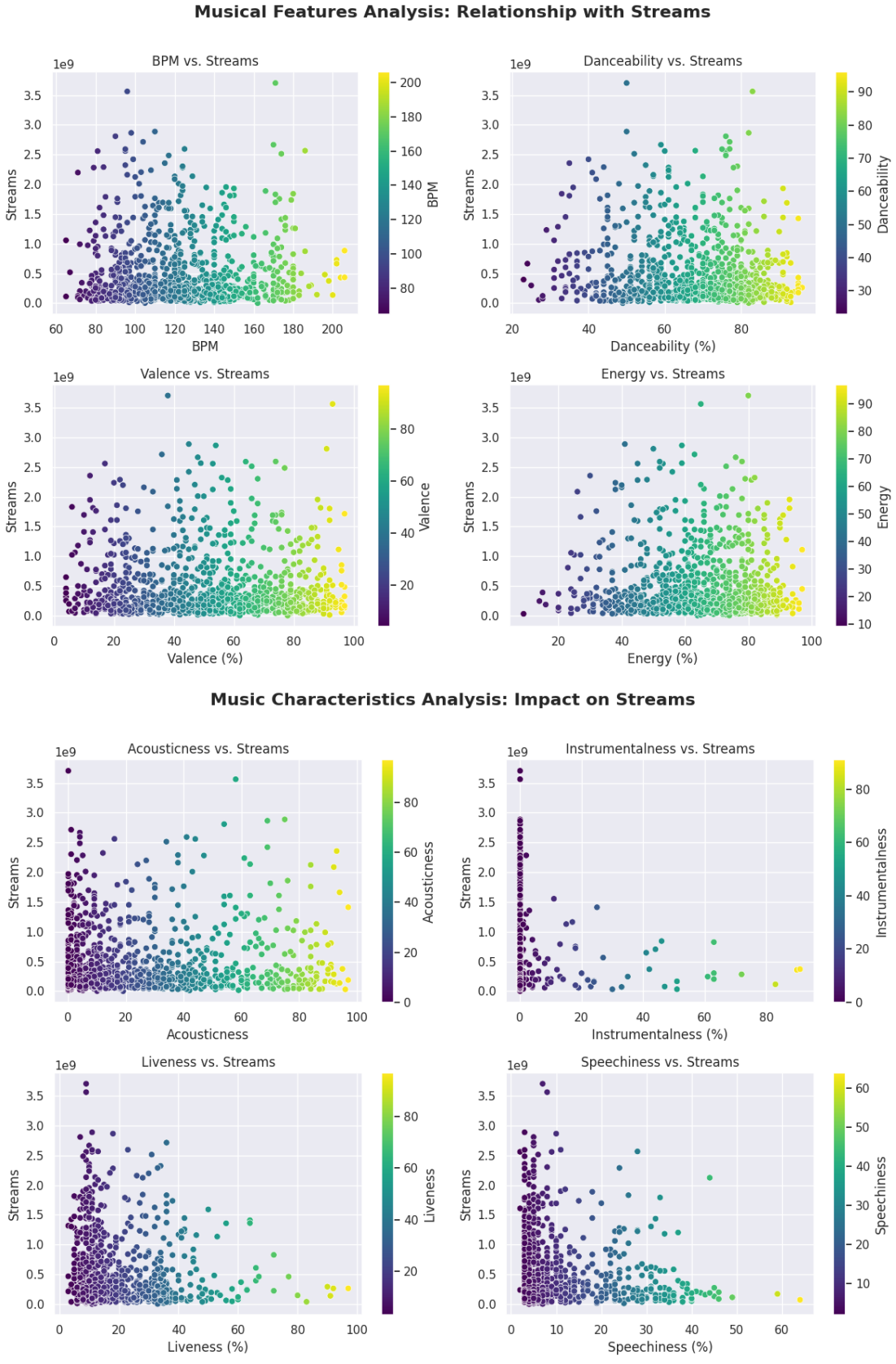
charakteristík pomocou histogramov. Zatiaľ čo bpm, tanečnosť, valencia a energia sa svojim charakterom približujú k normálnemu rozdeleniu. Akustickosť, inštrumentálnosť, živosť a rečovosť, obsahujú väčšinu svojho rozloženia v prvej tretine a svojim charakterom pripomínajú skôr rozdelenie exponenciálne.



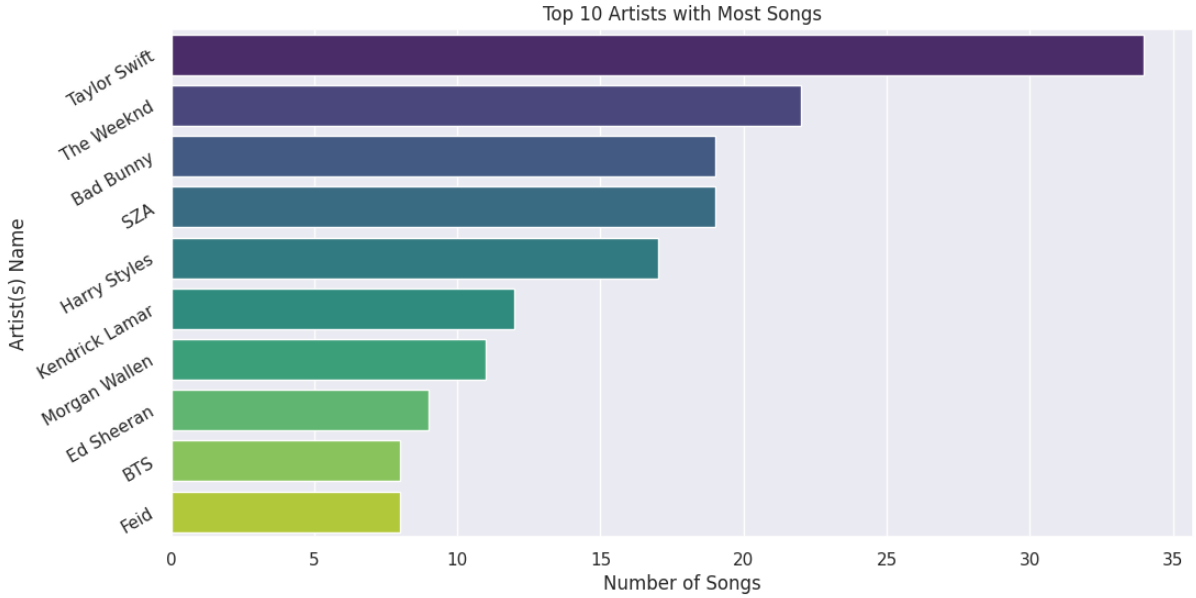


Ďalšia dvojica grafov, zobrazuje tieto charakteristiky v závislosti na počte streamov. Na grafov môžeme pozorovať, že rozloženie na základe počtu streamov, zachováva podobný charakter ako tomu bolo v distribúcii týchto charakteristík. Zaujímavým pod grafom z tejto

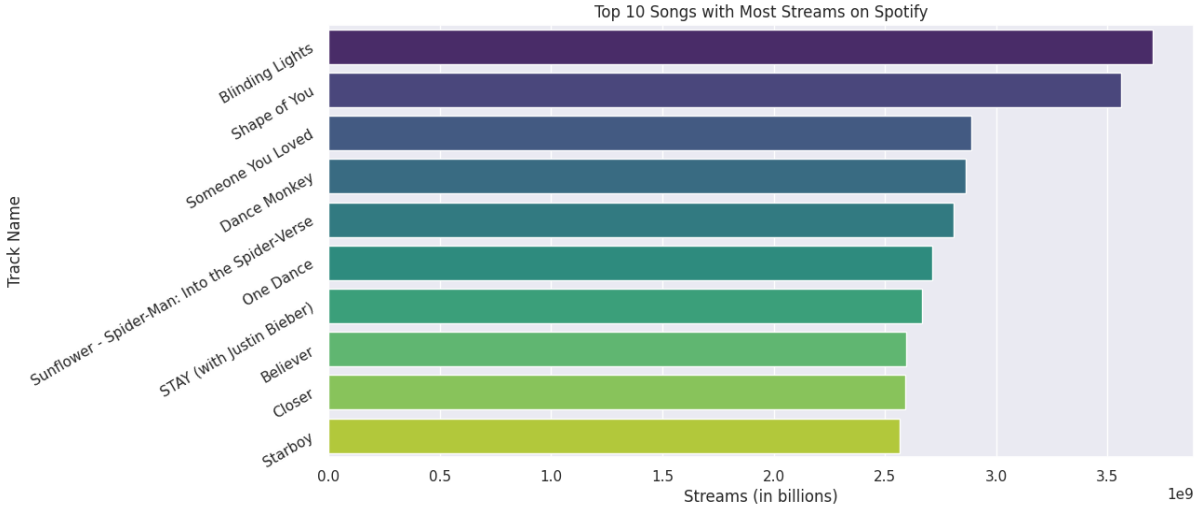
osmice je určite ten zobrazujúci, inštrumentálnosť v závislosti na streamoch a práve tento stĺpec sa nám ponúka pre kontrolu outlier, už priamo z grafu nakoľko väčšina jeho hodnôt leží blízko nuly.



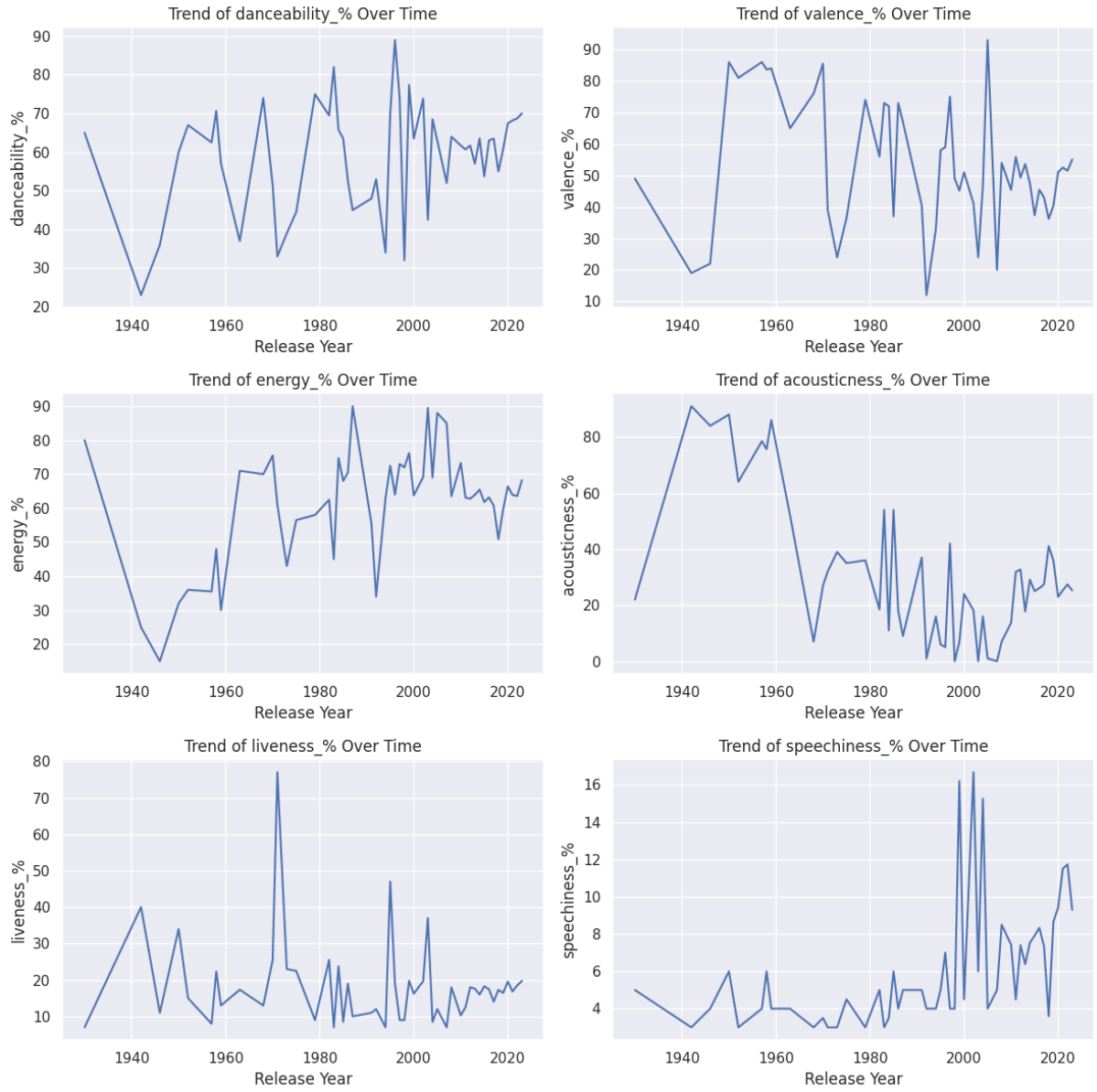
Ďalšími dvoma zaujímavými grafmi, hlavne pre lepšie pochopenie dátovej sady, sú, graf s top desiatimi autormi, ktorý v našej dátovej sade majú najväčšie zastúpenie počtu piesní. Môžeme pozorovať, že rebríčku dominuje Taylor Swift s takmer 35-mi piesňami.



Druhým veľmi zaujímavým grafom, nakoľko sa dátová sada zaoberá počtom streamov jednotlivých piesní na základe ich vlastností, je graf zobrazujúci, desať piesní s najväčším poćtom streamov. V notebooku sa podobné grafy nachádzajú aj pre ostatné platformy.



Posledným grafom zahrnutým v tejto časti je graf zobrazujúci trendy v charakteristikách piesní v priebehu času. Zatiaľ čo v prípade tanečnosti a energie naprieč rokmi môžeme pozorovať pomerne veľkú variabilitu, je patrný narastajúci trend. Na druhú stranu, akustickosť trpí skôr klesajúcim trendom, pričom najvyššie hodnoty dosiahla v priebehu rokov 1930 až 1960. Ďalšou veľmi výraznou charakteristikou je rečovosť, ktorá na prelome storočí zaznamenala veľký nárast, alebo skôr veľmi výrazné skoky.

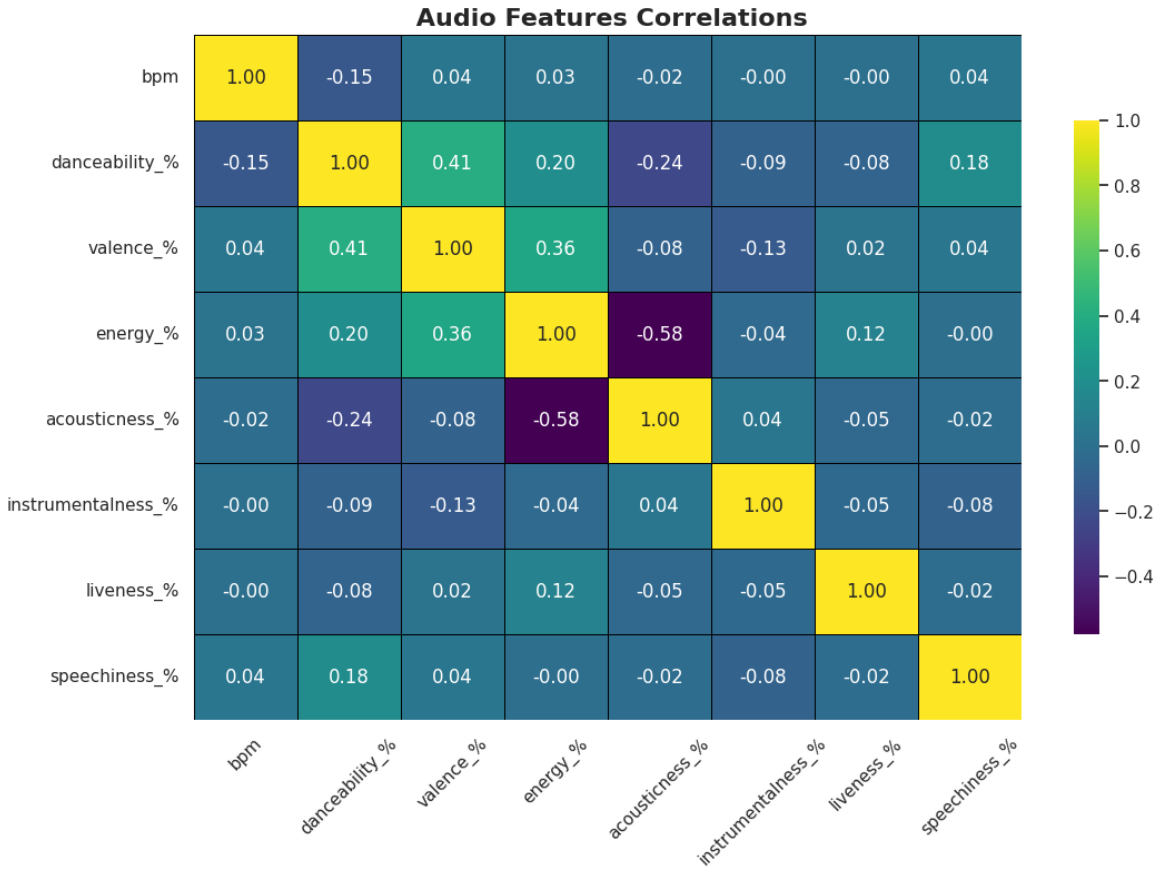


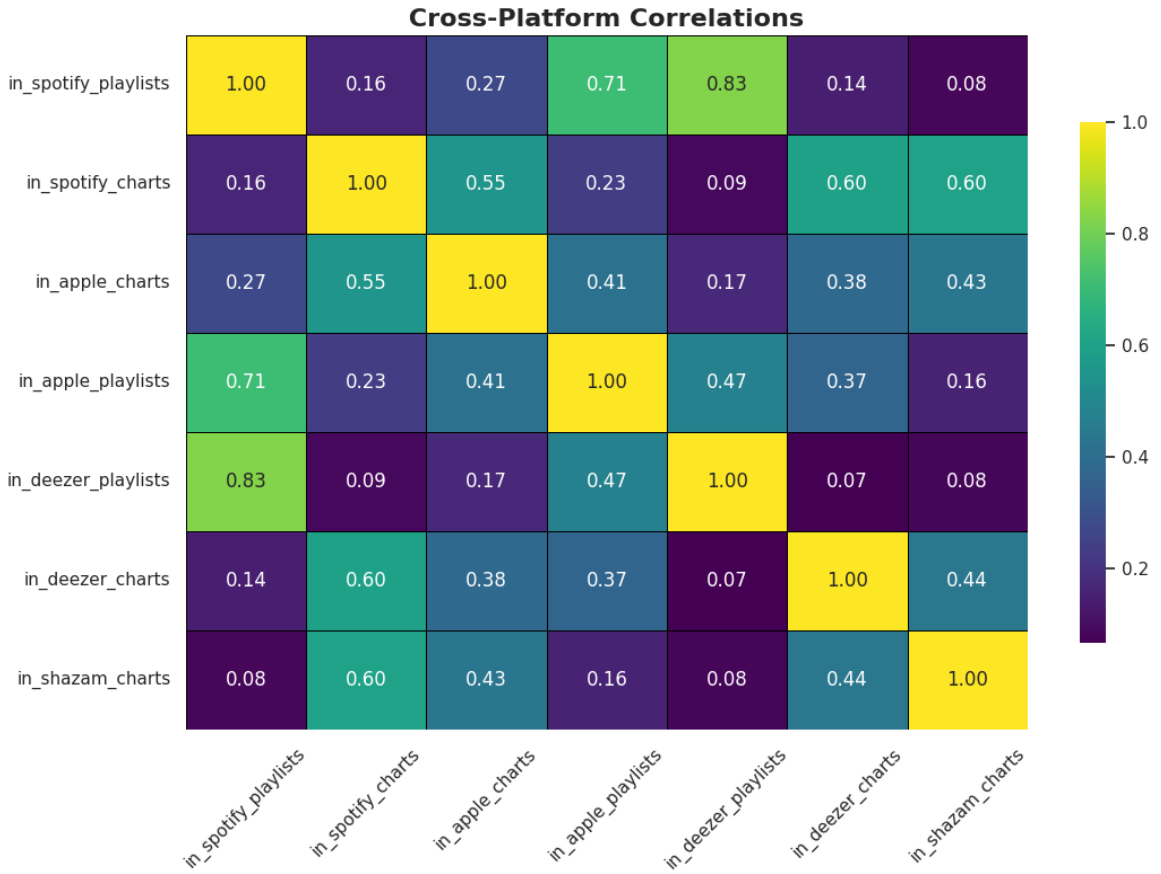
Korelačná analýza:

Vzhľadom na vysoké zastúpenie numerických atribútov v našej dátovej sade, bola kompletná korelačná matica pomerne rozsiahla a nečitateľná. Preto sme sa ju rozhodli rozdeliť na základe korelácií, na dva tematické celky a to zvukové charakteristiky piesne a obľúbenosť naprieč rôznymi platformami. Môžeme vidieť pomerne výrazné a silne pozitívne korelácie medzi obľúbenosťami naprieč platformami a umiestnením v jednotlivých priečkach.

Táto korelácia je veľmi intuitívna, nakoľko ak je pieseň populárna na jednej streamovacej platforme, bude s veľkou pravdepodobnosťou populárna aj na ostatných platformach, nakoľko jedna platforma reprezentuje len istú vzorku populácie.

Z korelácií zvukových charakteristík je veľmi zaujímavá pomerne silná záporná korelácia medzi akustickosťou a energiou. Vidíme teda, že čím viac bude pieseň energická, tým menej bude akustická. Čo sa týka pozitívnych korelácií, najdominantnejší sú korelácie medzi valenčnosťou a tanečnosťou, a energiou a valenčnosťou.


Audio Features Correlations
bpm
danceability_%
valence_%
energy_%
acousticness_%
instrumentalness_%
liveness_%
speechiness_%
1.00 -0.15 0.04 0.03 -0.02 -0.00 -0.00 0.04
-0.15 1.00 0.41 0.20 -0.24 -0.09 -0.08 0.18
0.04 0.41 1.00 0.36 -0.08 -0.13 0.02 0.04
0.03 0.20 0.36 1.00 -0.58 -0.04 0.12 -0.00
-0.02 -0.24 -0.08 -0.58 1.00 0.04 -0.05 -0.02
-0.00 -0.09 -0.13 -0.04 0.04 1.00 -0.05 -0.08
-0.00 -0.08 0.02 0.12 -0.05 -0.05 1.00 -0.02
0.04 0.18 0.04 -0.00 -0.02 -0.08 -0.02 1.00
bpm
danceability_%
valence_%
energy_%
acousticness_%
instrumentalness_%
liveness_%
speechiness_%
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Cross-Platform Correlations
in_spotify_playlists
in_spotify_charts
in_apple_charts
in_apple_playlists
in_deezer_playlists
in_deezer_charts
in_shazam_charts
1.00 0.16 0.27 0.71 0.83 0.14 0.08
0.16 1.00 0.55 0.23 0.09 0.60 0.60
0.27 0.55 1.00 0.41 0.17 0.38 0.43
0.71 0.23 0.41 1.00 0.47 0.37 0.16
0.83 0.09 0.17 0.47 1.00 0.07 0.08
0.14 0.60 0.38 0.37 0.07 1.00 0.44
0.08 0.60 0.43 0.16 0.08 0.44 1.00
in_spotify_playlists
in_spotify_charts
in_apple_charts
in_apple_playlists
in_deezer_playlists
in_deezer_charts
in_shazam_charts
1.0
0.8
0.6
0.4
0.2

Úprava dat

Odstránenie irelevantných atribútov

Pre dolovacie úlohy sú relevantné iba informácie o popularite (teda počet prehratí a radenie do playlistov/rebríčkov) a atribúty určujúce hudobnú povahu skladby (tempo, tónina, mód, percentuálne atribúty), ostatné atribúty ako autor alebo dátum vydania boli odstránené.

Dopočítanie chýbajúcich hodnôt

Chýbajúce hodnoty obsahujú stĺpce **in_shazam_charts** a **key**. V prípade chýbajúcej hodnoty **in_shazam_charts** predpokladáme, že sa daná pieseň v žiadnych rebríčkoch neumiestnila a teda sme ich nahradili hodnotou 0.

V prípade atribútu **key** sme chýbajúce hodnoty nahradili najčastejšou hodnotou v rámci dátovej sady. Keďže by ale tieto doplnené hodnoty mohli skreslovať výsledky dolovacieho algoritmu, tak sme pridali ďalší atribút **key_imputed**, ktorý symbolizuje, či bol atribút **key** v danom zázname doplnený. Týmto spôsobom dáme dolovaciemu algoritmu možnosť špeciálne ošetriť tieto dáta.

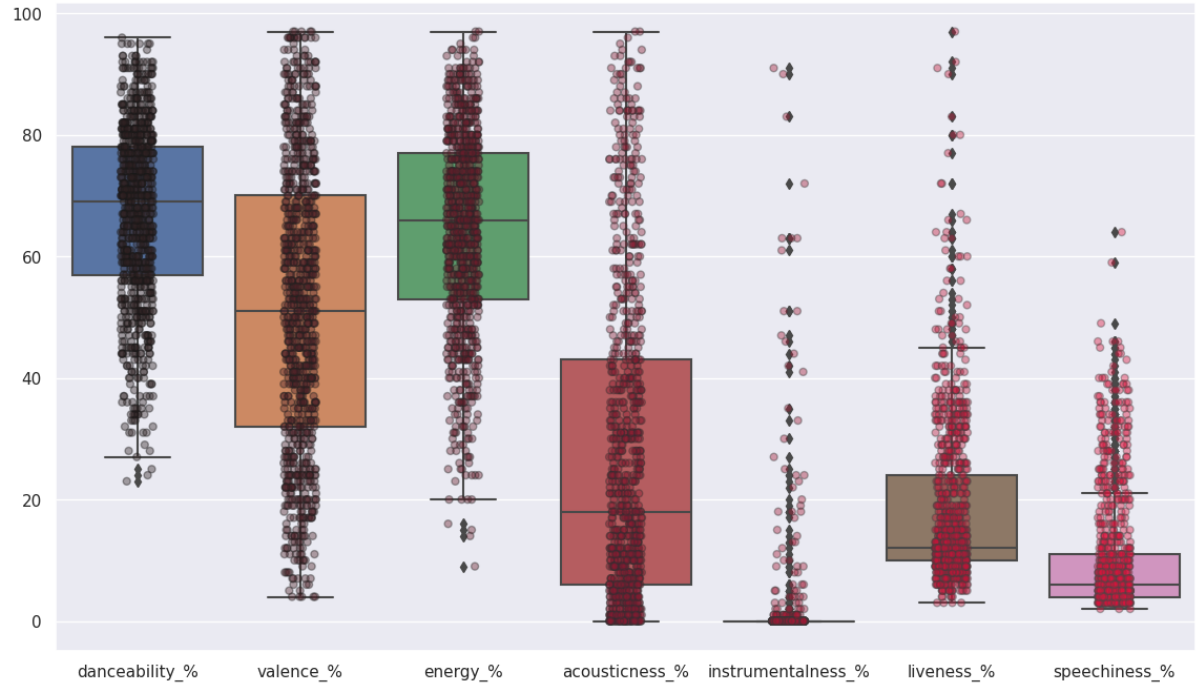
Odľahlé hodnoty

Za odľahlé hodnoty sa dajú považovať duplicitné skladby (teda skladby s rovnakým názvom a autorom). Môžeme predpokladať, že v tomto prípade bola rovnaká skladba vydaná v niekoľkých verziách (napr. v albe a ako single). Preto bolo takéto skladby spojené, pričom atribúty vyjadrujúce popularitu skladby boli sčítané a atribúty vyjadrujúce hudobnú povahu (v prípade, že sa mierne líšia) spriemerované.

Taktiež sme skúmali atribúty skladieb pomocou pravidla 1.5 IQR, kde hodnoty, ktoré spadajú mimo túto hranicu sú považované za odľahlé. Pomocou tejto metódy boli detekované nasledujúce počty odľahlých hodnôt:

Atribút	Počet odľahlých hodnôt
danceability	3
valence	0
energy	4
acousticness	0
instrumentalness	87
liveness	44
speechiness	136

Distribúcie hodnôt sú ale nasledovné:



Aj keď sme detekovali niekoľko odľahlých hodnôt, nemôžeme s určitosťou povedať, že niektoré z týchto hodnôt sú chybné alebo, že sú to naozaj outliers, ktoré by mali byť odstránené. Preto sme ich v dátovej sade ponechali. Najvýraznejší počet ich je v **instrumentalness** a **speechiness**. V oboch prípadoch je to ale pravdepodobne spôsobené faktom, že prevažná väčšina moderných skladieb obsahuje málo inštrumentálov a hovoreného slova.A trend týchto zvukových charakteristík ako sme mohli vidieť v exploračnej analýze, bol prevažne v skladbách ktoré vznikli v minulom storočí, avšak tieto skladby stále považujeme za valídne dáta.

Agregácia atribútov

Keďže dolovací algoritmus skúma celkovú popularitu jednotlivých skladieb, nie je potrebné brať do úvahy popularitu na jednotlivých platformách, stačí agregovaná popularita naprieč všetkými platformami. Preto sme atribúty **in_spotify_playlists**, **in_apple_playlists**, **in_deezer_playlists** agregovali do jediného atribútu **in_playlists** vypočítaním ich sumy. Rovnakým spôsobom sme agregovali atribúty **in_spotify_charts**, **in_apple_charts**, **in_deezer_charts** a **in_shazam_charts** do atribútu **in_charts.**

Prevod kategorických atribútov na numerické a normalizácia

Na vytvorenie varianty dátovej sady pre dolovacie algoritmy, ktoré vyžadujú číselné vstupy, boli zostávajúce kategorické atribúty **key** a **mode** prevedené na numerické atribúty, a to pomocou automatického kódovania, kde sa každej kategórii priradí číslo.

Hodnoty atribútov s percentuálnymi hodnotami sa previedli na rozsah [0, 1], kvôli jednoduchšiemu ďalšiemu spracovaniu dolovacími algoritmami. Keďže dolovacie algoritmy využívajúce strojové učenie často potrebujú na správne fungovanie atribúty s normálnym

rozdelením (so stredom v nule a so štandardnou odchýlkou 1), hodnoty atribútov **streams**, **in_charts**, **in_playlists** a **bpm** boli škálované tak, aby spĺňali túto podmienku.

Diskretizácia atribútov

Diskretizácia prebieha v dvoch podobách. Hodnoty atribútov **danceability**, **valence**, **energy**, **acousticness**, **instrumentalness**, **liveness** a **speechiness** sú reprezentované v percentách, ktoré majú približne rovnomerné rozloženie. Tieto hodnoty sme kategorizovali podľa intervalov 0-20, 20-40, 40-60, 60-80, 80-100 do kategórií **low**, **bellow_average**, **average**, **above_average**, **high**.

Atribúty **streams**, **in_playlists**, **in_charts**, **bpm** rovnomerné rozloženie nemajú. Z tohto dôvodu sme ich kategorizovali do 10 kategórií, kde každá kategória obsahuje približne rovnaké množstvo prvkov.