FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

UPA – 2. projekt

Příprava dat a jejich popisná charakteristika

Filip Jahn (xjahnf00)
Michal Luner (xluner01)
Vojtěch Staněk (xstane45)

19. listopadu 2023

Obsah

1 Úvod **2**

2 Explorativní analýza **2**
2.1 Průzkum atributů datové sady 2
2.1.1 Analýza numerických atributů 2
2.1.2 Analýza kategorických atributů 3
2.2 Průzkum rozložení hodnot 3
2.3 Analýza odlehlých hodnot 7
2.3.1 Numerické odlehlé hodnoty 7
2.3.2 Kategorické odlehlé hodnoty 9
2.4 Analýza chybějících hodnot 10
2.5 Korelační analýza numerických atributů 11

3 Příprava datové sady pro dolovací algoritmy **15**
3.1 Odstranění irelevantních atributů 15
3.2 Odstranění chybějících hodnot 15
3.3 Odstranění odlehlých hodnot 15
3.4 Sada 1 – Diskretizace numerických atributů 15
3.5 Sada 2 – Transformace kategorických atributů na numerické 17

4 Závěr **18**

1 Úvod

Cílem této části projektu je provedení explorativní analýzy na zvolené datové sadě a úprava této sady do podoby vhodné pro dolovací úlohu.

Zvolená datová sada: Most Streamed Spotify Songs 2023 (soubor `spotify-2023.csv`)

Dolovací úloha: Predikce oblíbenosti písně na základě ostatních atributů (např. rytmus písně, tónina, mód písně apod.)

2 Explorativní analýza

2.1 Průzkum atributů datové sady

V rámci této fáze jsme prozkoumali jednotlivé atributy datové sady, jejich typy a hodnoty, kterých nabývají.

Zjistili jsme, že některé položky mají nevyplněnou hodnotu. Zároveň je nevhodné, aby zřejmě numerický atribut (`streams`, `in_deezer_playlists`, `in_shazam_charts`) byl kategorický datový typ. Už v této fázi tedy převádíme vybrané atributy na očekávaný datový typ a vypořádáváme se s `NULL` hodnotami (více informací v sekci 2.4). Důvod, proč s úpravou dat takto „předbíháme“, je, že chceme mít co nejreprezentativnější přehled již od začátku. Tohle je zvláště důležité ve vizualizacích dat v sekci 2.2.

2.1.1 Analýza numerických atributů

Pro numerické atributy vypisujeme základní statistiky pomocí funkce `describe()`, z čehož bychom už potenciálně mohli odhalit nějaké odlehlé hodnoty. Celá tabulka je k dispozici ve zdrojovém kódu, pro zachování přehlednosti poskytujeme ukázku pouze části tabulky.

	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams
count	953.0	953.0	953.0	953.0	953.0	953.0	953.0
mean	2.0	2018.0	6.0	14.0	5200.0	12.0	513597931.0
std	1.0	11.0	4.0	9.0	7898.0	20.0	566803887.0
min	1.0	1930.0	1.0	1.0	31.0	0.0	0.0
25%	1.0	2020.0	3.0	6.0	875.0	0.0	141381703.0
50%	1.0	2022.0	6.0	13.0	2224.0	3.0	290228626.0
75%	2.0	2022.0	9.0	22.0	5542.0	16.0	673801126.0
max	8.0	2023.0	12.0	31.0	52898.0	147.0	3703895074.0

Obrázek 1: Analýza numerických atributů

Z tabulky [1] je patrné, že například hodnoty roku vydání jsou v rozsahu od 1930 až po aktuální rok 2023. Podobně např. z počtu streamů můžeme zjistit, že největší počet streamů je 3,7 mld., zatímco nejmenší počet streamů je 2762. Průměr streamů je 514 mil. se směrodatnou odchylkou 567 mil. Z toho můžeme usoudit, že počty streamů se mohou výrazně lišit a kolísat.

2.1.2 Analýza kategorických atributů

Podobně jako u numerických atributů lze `describe()` funkci využít i pro kategorické atributy. Z výstupů [2] je možné zjistit, kolik hodnot je v datasetu celkově, kolik těchto hodnot je unikátních a zároveň se vypisuje i hodnota nejčastějšího výskytu atributu.

	track_name	artist(s)_name	key	mode
count	953	953	953	953
unique	943	645	12	2
top	Daylight	Taylor Swift	C#	Major
freq	2	34	120	550

Obrázek 2: Analýza kategorických atributů

2.2 Průzkum rozložení hodnot

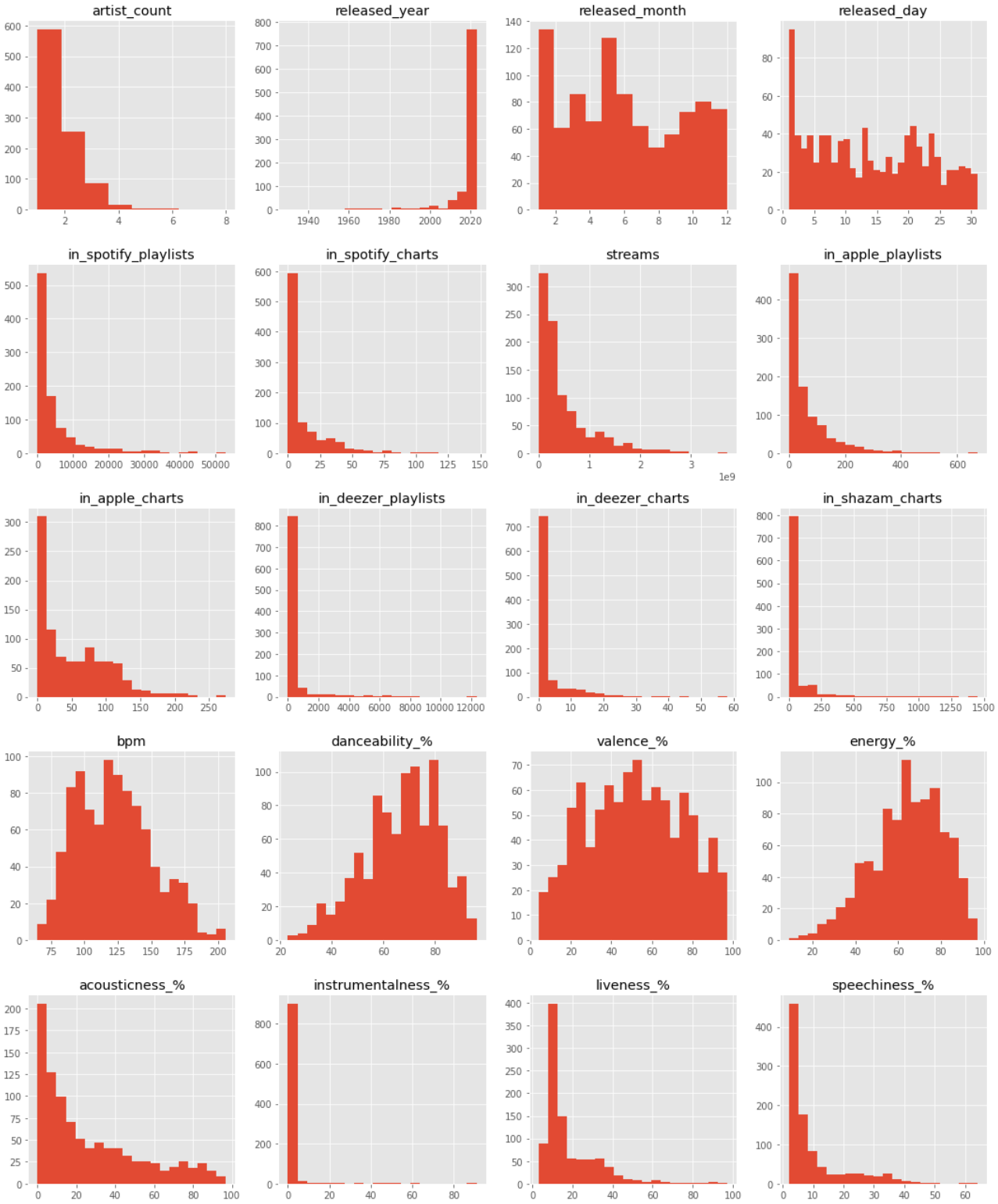
V této fázi prozkoumáváme rozložení hodnot jednotlivých atributů pomocí vhodných grafů se zaměřením na vztahy mezi atributy.

Využíváme upravenou datovou sadu s vyřešenými chybějícími hodnotami (více informací v sekci 2.4).

Rozložení hodnot numerických atributů

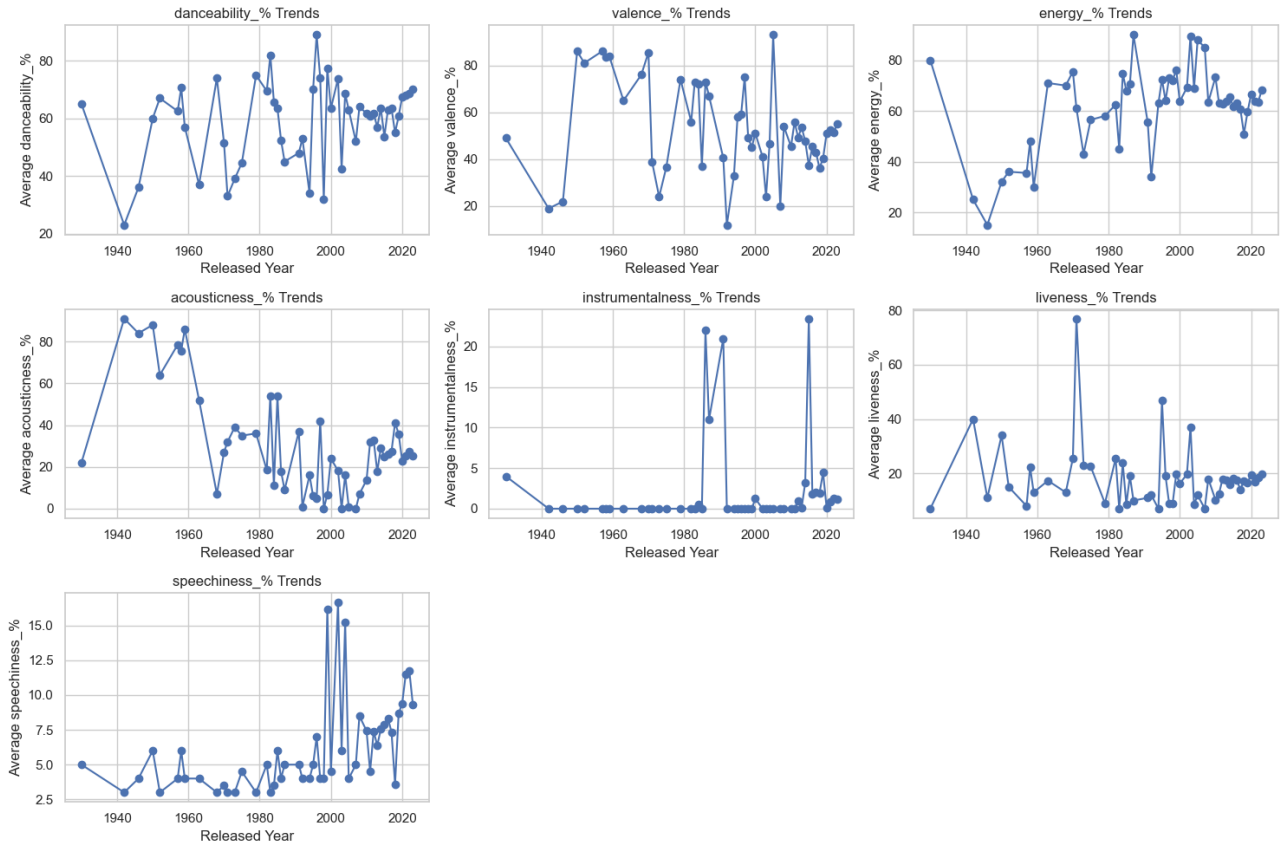
Pro průzkum rozložení hodnot všech atributů využijeme zobrazení matice grafů [3]. Z té můžeme vyčíst velké množství informací:

- V prvním řádku vidíme, že nejčastější počet interpretů podílejících se na písničce je roven 1. Nejpopulárnější je, co se do počtu streamů týče, soudobá hudba. Zároveň nejčastější měsíce pro vydání nové písničky jsou leden a květen, většinou na začátku měsíce.
- Na dalším řádku vidíme sadu playlistů a žebříčků jednotlivých platforem, kde rozložení bývá podobné, tj. nejstreamovanější hudba nemusí být nutně na čelech žebříčků. To, že nejsou ani v playlistech, by se dalo vysvětlit tak, že jelikož se jedná o nové písně, tak se ještě nedostaly do playlistů (a dojde k tomu pravděpodobně později).
- Další sada grafů se zaobírá akustickými vlastnostmi hudby. Už z těchto grafů vidíme, že je preferována spíše hudba s vyšší mírou `danceability` a `energy`. Naopak hudba s výraznými prvky `acousticness`, `instrumentalness`, `liveness` ani `speechiness` populární tolik není. Zajímavostí pak je, že nejpopulárnější hudba má tempo okolo 90–130 bpm.



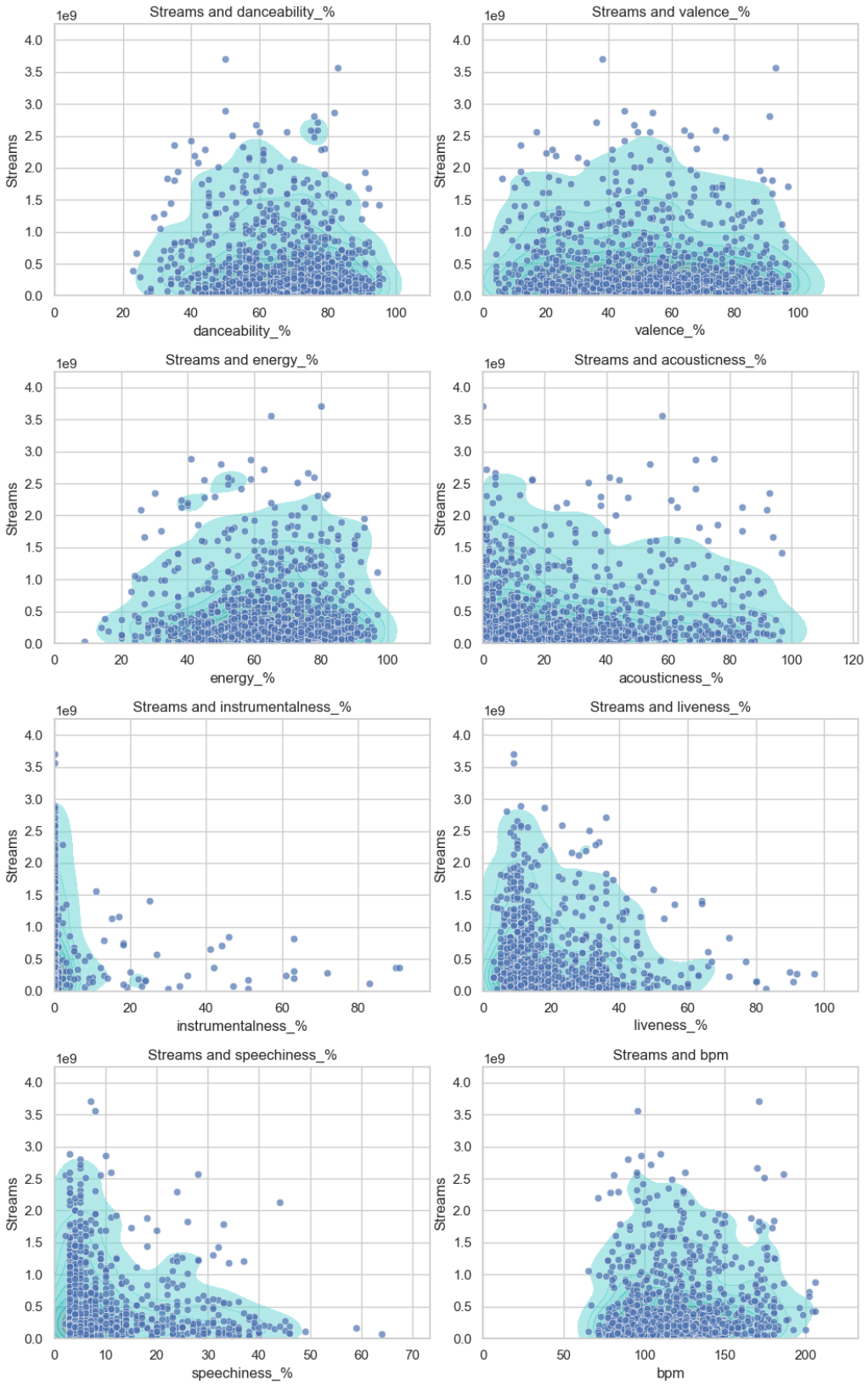
Obrázek 3: Rozložení hodnot numerických atributů

Další matice grafů [4] popisuje trendy akustických vlastností na základě let. Můžeme kupříkladu vidět, že okolo 40. let 20. st. v hudebním průmyslu po válce interpretům nebylo příliš do tance, naopak však přibyly hodnoty atributu `acousticness`.



Obrázek 4: Trendy v akustických vlastnotech skladeb v závislosti na letech

Další sada grafů [5] poskytuje náhled na to, jak jednotlivé akustické vlastnosti ovlivňují streamovanost. Z grafů nevidíme zcela jednoznačné ukazatele, které by měly mít vliv na streamovanost, avšak jsou zřetelné preference ve vyšších hodnotách atributů `energy` a naopak nižší hodnoty v `instrumentalness`, `liveness` a `speechiness`. Obecně tak lze říci, že posluchači preferují skladby obsahující více zpěvu na úkor řeči a instrumentálním prvkům.



Obrázek 5: Závislost akustických vlastností na streamovanost

2.3 Analýza odlehlých hodnot

K detekci a analýze odlehlých hodnot lze využít několika přístupů:

1. vizualizační techniky (krabicové grafy)
2. numerické výpočty (IQR, identifikace odlehlých hodnot za pomocí z-score normalizace)

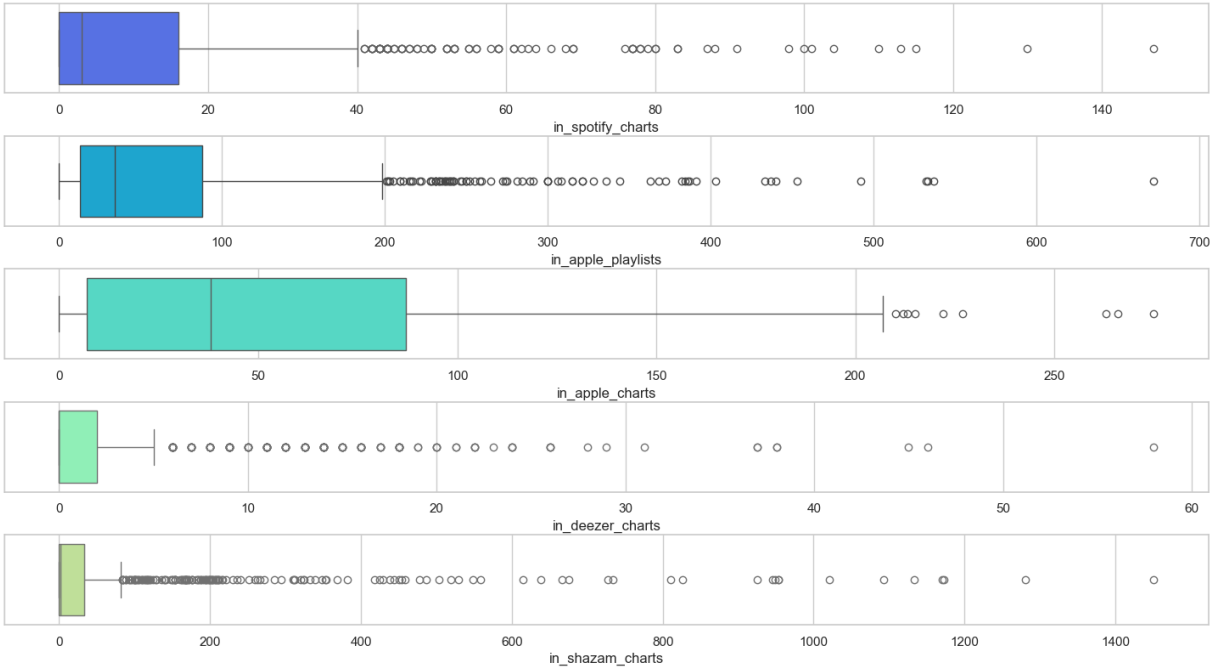
U odlehlých hodnot je třeba mít na paměti, že i když se může jednat o nějakou extrémní hodnotu, neimplikuje to nutně to, že by se jednalo o chybnou hodnotu.

Jelikož hledáme odlehlé hodnoty, uvažujeme, že `NULL` hodnoty máme již vyřešené a atributy máme ve správném formátu, což jsme již provedli v předešlých krocích[1].

2.3.1 Numerické odlehlé hodnoty

Nejdříve z datasetu na základě datového typu vybereme atributy, které jsou numerické. Z předešlých fází víme, jaké tyto atributy mají zhruba rozsah. Jelikož chceme nejdříve zobrazit souhrnné krabicové grafy, rozdělíme atributy tak, aby si vzájemně nekazily měřítko.

Krabicové grafy [6] zobrazují data atributů playlistů a žebříčků.



Obrázek 6: Analýza odlehlých hodnot u atributů pro playlisty a žebříčky

Spotify a Deezer playlisty mají svůj graf, jelikož mají znatelně větší hodnoty oproti ostatním atributům. Ač z grafů vidíme velké množství bodů naznačujících odlehlé hodnoty, nemělo by se jednat o chybná data. Je očekávatelné, že průměrně úspěšných písní je větší množství, a tak populárnější skladby působí jako odlehlé hodnoty.

Dále prozkoumáváme data, období vydání a počet přispívajících interpretů. Hodnoty z tabulky [7] nám dají dostatek informací pro to, abychom identifikovali případné odlehlé hodnoty.

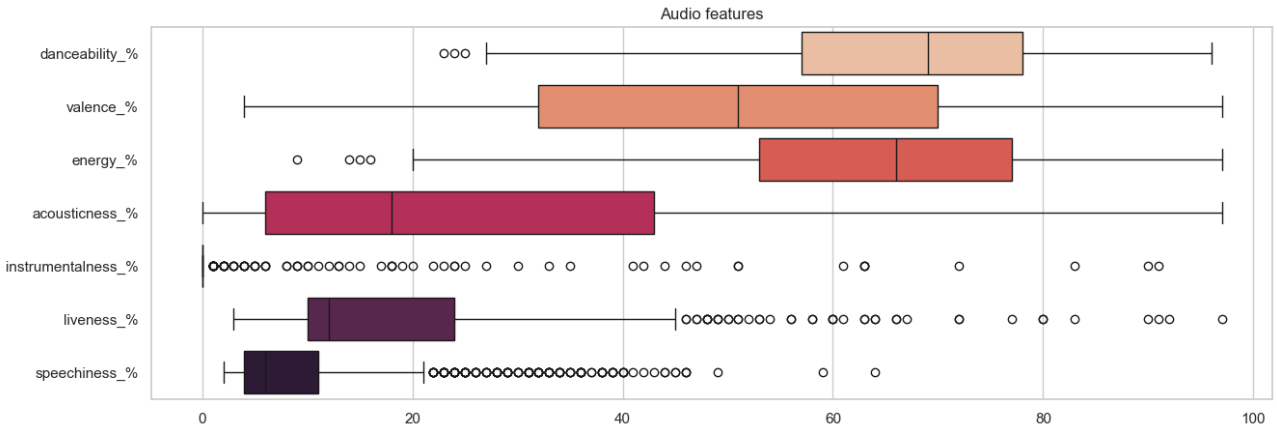
[1] V kódu na to využíváme funkci `convertAndFillNulls(data)`

	released_year	released_month	released_day	artist_count
count	953.000000	953.000000	953.000000	953.000000
mean	2018.238195	6.033578	13.930745	1.556139
std	11.116218	3.566435	9.201949	0.893044
min	1930.000000	1.000000	1.000000	1.000000
25%	2020.000000	3.000000	6.000000	1.000000
50%	2022.000000	6.000000	13.000000	1.000000
75%	2022.000000	9.000000	22.000000	2.000000
max	2023.000000	12.000000	31.000000	8.000000

Obrázek 7: Hodnoty atributů pro den, měsíc, rok a počet interpretů podílejících se na jedné skladbě

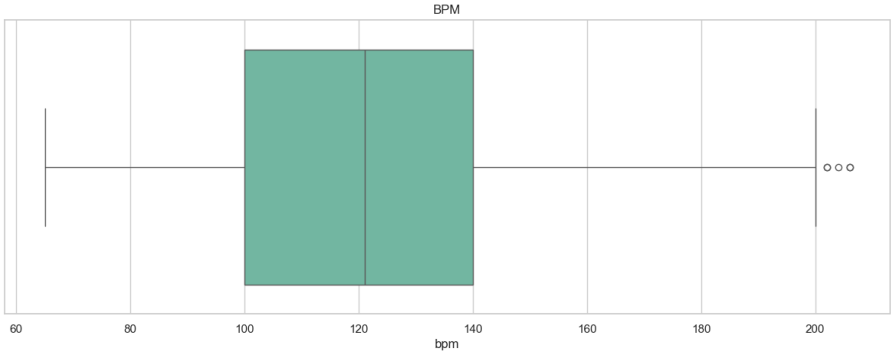
Z tabulky [7] vidíme, že maximální/minimální hodnoty pro rok, měsíc i den dávají smysl, stejně tak počet interpretů na písničku je v rozumných mezích.

Další složený graf [8] ukazuje jednotlivé aspekty audia v procentech. Rozsah hodnot by tedy měl být v rozmezí 0 až 100, což odpovídá. Zřejmě nejvíce upoutá pozornost atribut `instrumentalness`, který uvádí podíl instrumentální složky v písni. Při manuální analýze objektů, které mají tuto hodnotu vysokou, se potvrdilo, že se skutečně nejedná o chybné hodnoty – tato informace může mimo jiné posloužit i k tomu, že obecně nejstreamovanější hudba je z velké části vokálnějšího charakteru.

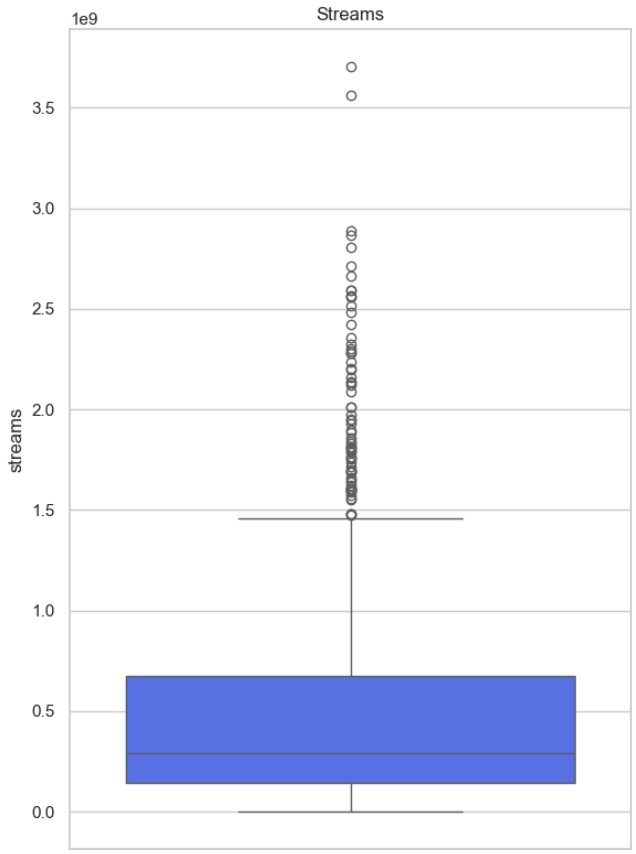


Obrázek 8: Rozložení hodnot pro akustické vlastnosti skladeb

Grafy [9] a [10] pak zobrazjí BPM a informaci o počtu streamů. Nejúspěšnější interpreti se svými skladbami dosahují přes 3 mld. poslechů, tedy i tato odlehlá data jsou v pořádku.



Obrázek 9: BPM – tempo skladby

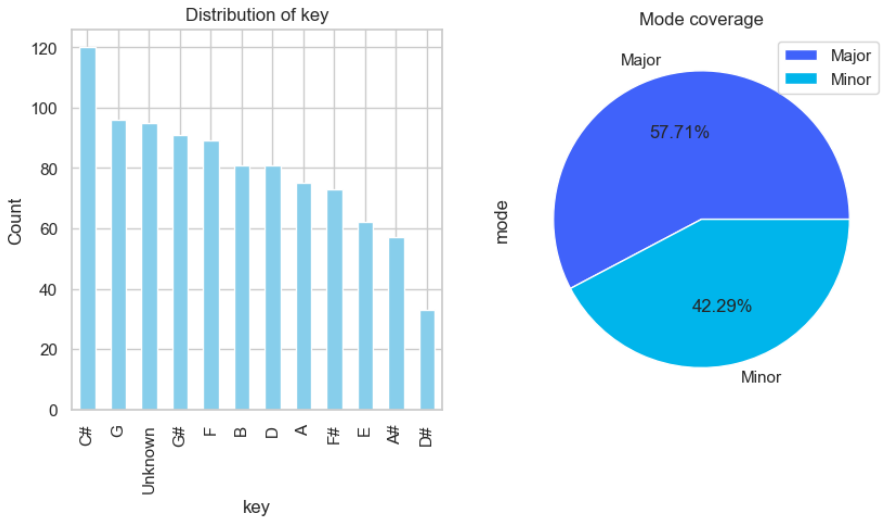


Obrázek 10: Počty streamů pro jednotlivé skladby

2.3.2 Kategorické odlehlé hodnoty

V hodnotách kategorických atributů je situace složitější. Je zřejmé, že jméno písně ani interpreta nelze jakkoliv vyhodnocovat na odlehlé hodnoty. Z atributů, které analyzovat lze, tak zbývají `mode` a `key`, jejichž zastoupení je zobrazeno v grafech [11].

Hodnota `Unknown` u klíče byla automaticky doplněna na místa, kde hodnota původně nebyla. Tomuto se více věnujeme v sekci 3.2. Z grafů vidíme, že lehce oblíbenější je tónina cis (C#), zbytek nelze na první pohled nějak výrazně odlišit. Pozoruhodné je, že některé tóniny zde nejsou zastoupeny vůbec (např. C) a bylo



Obrázek 11: Rozložení kategorických atributů `mode` a `key`

by zajímavé zjistit, proč tomu tak je. V modalitě pak jasně vyhrávají durové („veselé“) písně oproti molovým („smutným“) písním.

Ačkoliv se v datové sadě nachází nezanedbatelné množství odlehlých hodnot, vyhodnotili jsme je jako korektní.

2.4 Analýza chybějících hodnot

Výsledky z této sekce využíváme již v předešlých sekcích, abychom dostávali reprezentativní výsledky.

U analýzy chybějících hodnot je nutné mít na paměti, jaké unikátní hodnoty pro dané atributy existují (např. jestli místo `NULL` není v záznamech třeba `?`, který bychom jinak neodhalili). Ač předchozí části explorativní analýzy měly sloužit k tomu, abychom nezvyklé hodnoty odhalili, pro jistotu jsme si ještě jednou vypsali zastoupení hodnot pro dané atributy a v textovém editoru je pak manuálně prošli. U neprázdných hodnot v této fázi předpokládáme, že jsou korektní.

Jelikož jsme nenašli nezvyklé hodnoty atributů naznačující, že by se jednalo o chybějící data, lze využít knihovních funkcí `isnull()` pro identifikaci `NULL` hodnot. Z výpisu funkce jsme zjistili, že se jedná o atributy `key` a `in_shazam_charts`, které mají nenulový počet chybějících hodnot.

Atribut `key` unikátně charakterizuje danou píseň z hlediska tóniny (vizte dokumentaci Spotify[2]). Doplnění této hodnoty na základě jiných hodnot nedává smysl, nicméně mazat 95 záznamů by byla škoda. Navržené řešení nastaví prázdné hodnoty na `Unknown`, vyjadřující neznalost tóniny u dané písně. Alternativně by se dalo písně manuálné projít a hodnoty doplnit, nicméně s cílem vytvořit řešení robustní a znovupoužitelné (např. pro sadu vydanou příští rok), ruční doplňování nebylo provedeno.

U atributů pro playlisty a žebříčky se hojně vyskytuje hodnota 0. Předpokládáme, že se nejedná o chybějící hodnotu, nýbrž o informaci o tom, že se daná píseň nedostala do playlistů či žebříčků. `in_shazam_charts` udává hodnocení v žebříčcích vyhledávání v aplikaci Shazam (větší číslo indikuje větší oblíbenost/přítomnost v žebříčku). Nepřítomnost této hodnoty tedy může znamenat to, že písnička se do žebříčků vůbec nedostala, proto by dávalo smysl doplnit hodnotu 0.

Dále jsme analyzovali, jak moc se řádky s chybějícími hodnotami překrývají, tj. kolik záznamů celkem obsahuje aspoň jednu `NULL` hodnotu. Takových bylo 136 z celkových 145 chybějících hodnot, tzn. překryv nastává pouze v 9 případech.

[2] https://developer.spotify.com/documentation/web-api/reference/get-audio-features

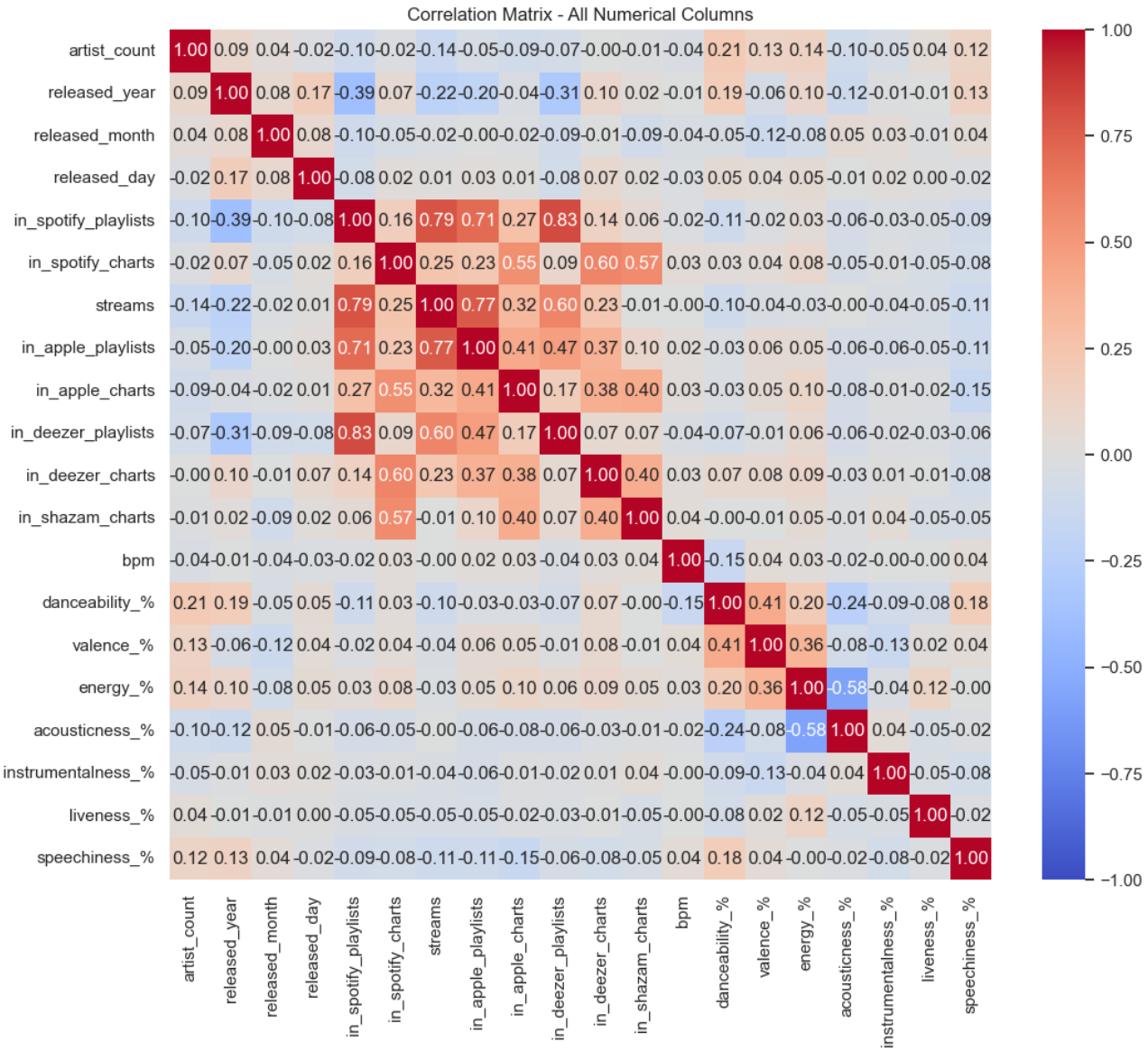
Při převodu atributu `streams` na numerický datový typ došlo k tomu, že vznikl nový `NULL` záznam. Při analýze hodnoty před konverzí se zjistilo, že hodnota tohoto záznamu byla: BPM110KeyAModeMajor... Tento záznam byl chybný, a proto byl smazán.

Demonstrovaly se 2 přístupy: odstranění záznamu na základě `NULL` hodnoty (pro atribut `streams`) a automatické doplnění chybějící hodnoty novou hodnotou (`key` a `in_shazam_charts`).

2.5 Korelační analýza numerických atributů

I v této části pracujeme s daty s vyřešenými chybějícími hodnotami.

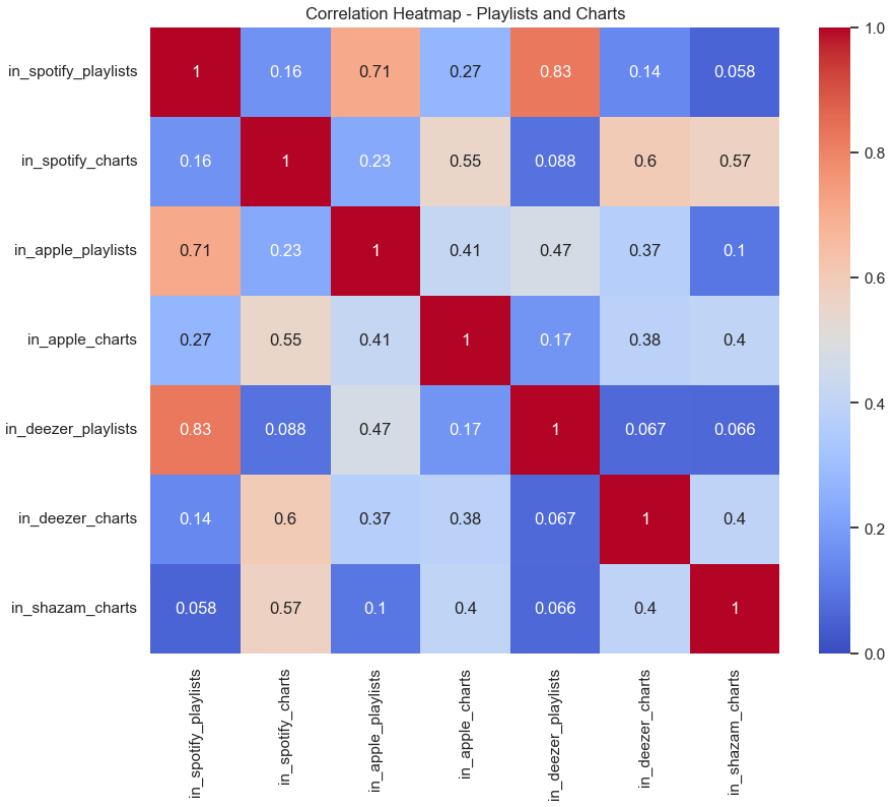
Pro prvotní rychlý přehled posloužila korelační matice všech atributů [12].



Obrázek 12: Korelační matice všech numerických atributů

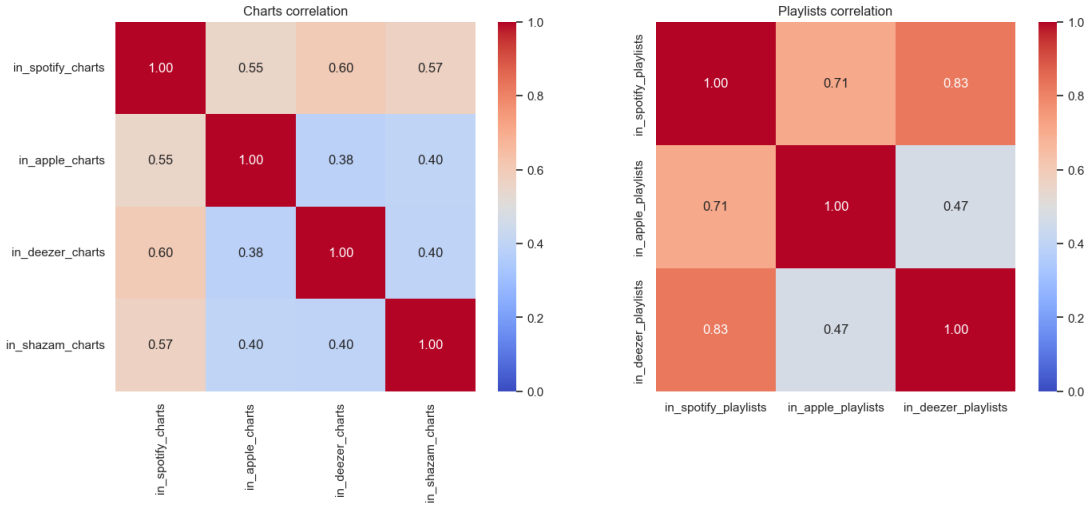
Na první pohled z korelační matice [12] vidíme, že zřejmě playlisty a žebříčky obsahují podobná data, podobně i některé akustické vlastnosti jsou korelované (např. `danceability_%` a `valence_%`). Obě tyto

oblasti tedy zobrazíme podrobněji. Pozor na barevné zobrazení – barvy nyní vystihují korelační koeficienty v intervalu (0, 1).



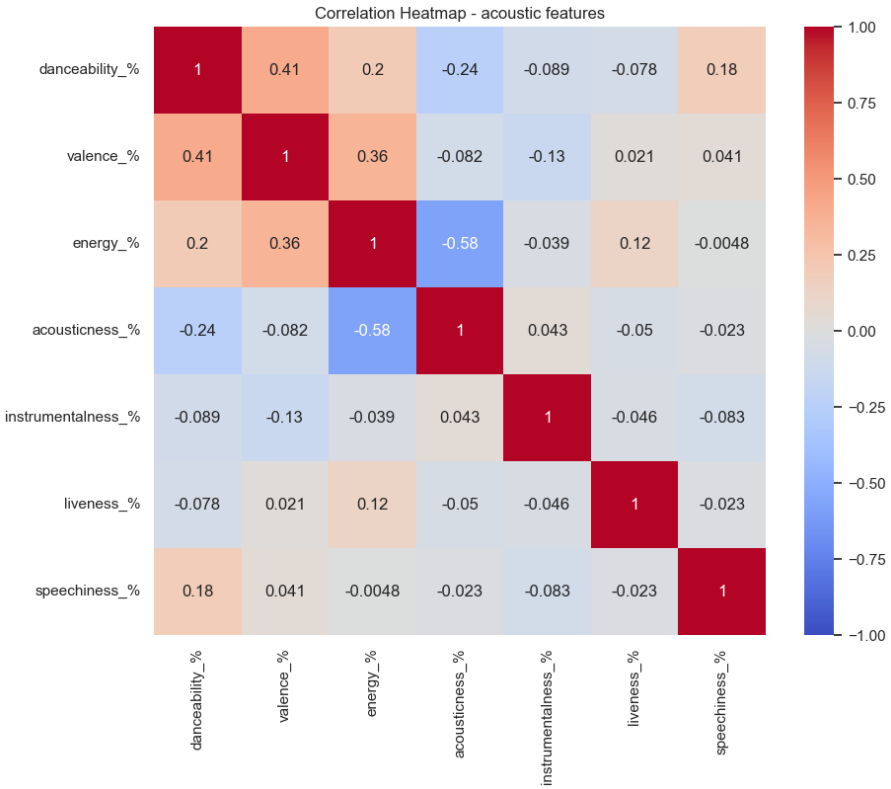
Obrázek 13: Korelační matice playlistů a žebříčků

Z matice [13] toho na první pohled nic výrazného nevyčteme, je však vidět, že existuje jistá korelace mezi playlisty navzájem a žebříčky navzájem. Zobrazíme tyto dvě kategorie v matici [14] zvlášť pro lepší přehlednost.



Obrázek 14: Korelační matice zvlášť pro playlisty a zvlášť pro žebříčky

Z matic [14] vidíme, že žebříčky i playlisty si navzájem docela odpovídají a mají významnou míru korelace. Podobně si zobrazíme i jednotlivé akustické atributy v matici [15].



Obrázek 15: Korelační matice pro akustické vlastnosti

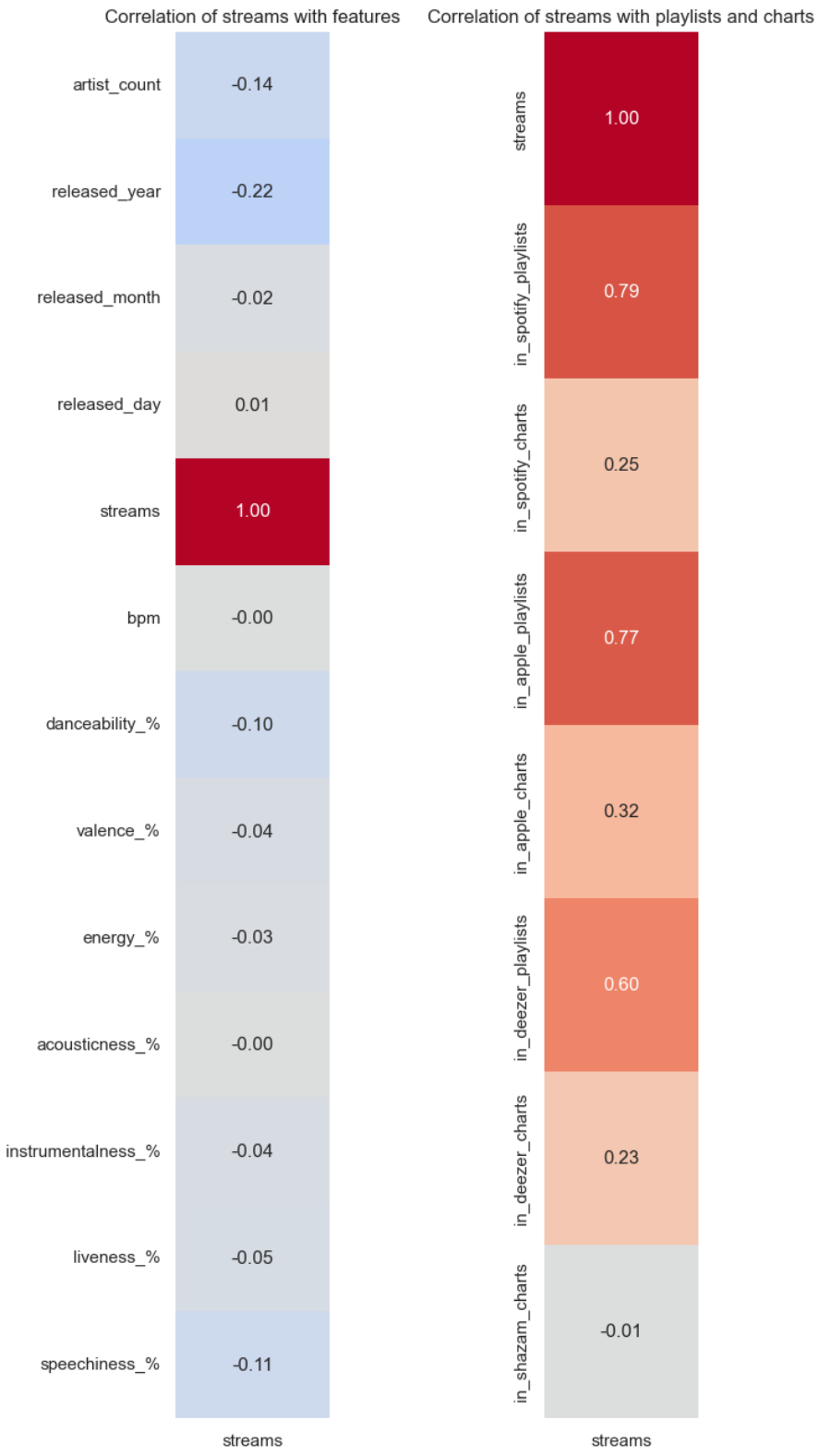
Z matice [15] jistou míru korelace pozorujeme u vlastností `danceability`, `valence` a `energy`, jinak žádná výrazná spojitost patrná není. Zajímavá je však poměrně vysoká záporná korelace mezi `energy` a `acousticness`. Zřejmě tedy energické písně bývají do výrazné míry elektronické s menším podílem akustických zvuků.

Hlavní věc, co nás však zajímá, je, zda má nějaký atribut výrazný vliv na počet poslechů, tedy zda jsou písně oblíbenější na základě nějaké vlastnosti (dostupných v datové sadě). Z velké korelační matice [12] vidíme, že největší korelaci s počtem poslechů mají umístění v žebříčcích a playlistech, zobrazíme je proto zvlášť.

V matici [16] jsme bohužel žádnou významnou korelaci neobjevili. Statisticky tedy můžeme říci, že oblíbenost písně nezávisí pouze na nějaké konkrétní vlastnosti (dostupné v datové sadě). Největší (zápornou) korelaci nalézáme u roku vydání písně, ale ani zde není příliš výrazná.

Jak z předchozích grafů, tak i zde je zřejmé, že počet přehrání koreluje s umístěním v žebříčcích a počtu obsazení v playlistech. Korelace však nemusí znamenat kauzalitu, tedy zda píseň bude nejprve oblíbená a následně se začne objevovat v playlistech a žebříčcích či naopak (z logiky věci vyplývá, že se jedná nejspíš o oboustranný vliv).

Každopádně díky tomu získáváme přehled o tom, které numerické atributy budou důležité pro zadanou dolovací úlohu.



Obrázek 16: Korelační matice závislostí mezi počty streamů a výskytů v playlistech a žebříčcích

3 Příprava datové sady pro dolovací algoritmy

Pracujeme s výchozími, nemodifikovanými daty.

Explorativní analýza by měla sloužit mimo jiné k nalezení korelací, které nám nyní pomohou rozhodnout, které atributy (ne)jsou důležité pro druh dolovací úlohy. Úloha, pro kterou data připravujeme, je predikce oblíbenosti písně na základě ostatních atributů. Oblíbenost písně budeme posuzovat metrikou počtu přehrání (čím oblíbenější píseň, tím častěji se přehrává).

3.1 Odstranění irelevantních atributů

Dle zadání ponecháme v datové sadě jak kategorické, tak i numerické atributy, atributy s chybějícími hodnotami i atributy s odlehlými hodnotami.

Relevantní atributy vybereme na základě korelační analýzy provedené v sekci 2.5. Vyřadíme všechny atributy, které mají téměř nulovou korelaci s počtem přehrání. Hodnota výběrového prahu byla experimentálně zvolena ±0.10 (i negativní korelace může mít v dolovací úloze vliv). Odstraněnými numerickými atributy jsou měsíc, den vydání, BPM, umístění v Shazam žebříčcích a akustické vlastnosti kromě `danceability_%` a `speechiness_%`.

Co se kategorických atributů týče, předpokládáme, že název písně nemá na počet přehrání vliv[3], proto můžeme atribut `track_name` odstranit. Na druhou stranu, jméno umělce určitě může být významným faktorem pro počet přehrání (známější interpret má více fanoušků, kteří si jeho dílo přehrají).

Během explorativní analýzy jsme zjistili, že některé hodnoty atributů `key` a `mode` jsou zastoupeny častěji než ostatní. Z toho můžeme vyvodit, že oba atributy mohou mít potenciálně pro dolovací úlohu význam, a proto je v datové sadě ponecháme.

3.2 Odstranění chybějících hodnot

Chybějící hodnoty byly analyzovány v rámci sekce 2.4. Získané poznatky jsme aplikovali na takto zredukovanou datovou sadu. Nahradili jsme `NULL` hodnoty odpovídající korektní hodnotou (tj. pro atribut `key` jsme dle dokumentace Spotify[4] doplnili hodnotu -1). Při převodu atributu `streams` na numerický datový typ vznikl neplatný záznam, který nebylo možné korektně opravit – ten jsme se tedy rozhodli vymazat.

3.3 Odstranění odlehlých hodnot

Ač v datové sadě odlehlé hodnoty jsou, jejich odstranění by mělo smysl zejména u atributů, které jsme již zavrhli (např. u `instrumentalness_%`). Podobně např. u `streams` se i u odlehlých hodnot nejedná o chybná data.

Z upravené datové sady není třeba odlehlé hodnoty odstraňovat.

3.4 Sada 1 – Diskretizace numerických atributů

První varianta datové sady má být dle zadání vhodná pro algoritmy vyžadující na vstupu kategorické atributy.

Abychom mohli lépe diskretizovat, zobrazíme si informace o rozložení jednotlivých atributů v tabulkách [17] a [18]. Společně s vizualizacemi (zejména matice grafů) vytvořené během explorativní analýzy tak získáme dobrý přehled o zpracovávaných datech.

[3] Byl by to ale zajímavý lingvistický experiment, zkoumající vliv názvu skladby na počet přehrání, např. zda výskyt některých slov v názvu písně významně ovlivňuje její úspěch.

[4] https://developer.spotify.com/documentation/web-api/reference/get-audio-features

	artist_count	released_year	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts
count	952.0	952.0	952.0	952.0	952.0	952.0	952.0
mean	1.6	2018.3	5202.6	12.0	514137424.9	67.9	52.0
std	0.9	11.0	7901.4	19.6	566856949.0	86.5	50.6
min	1.0	1930.0	31.0	0.0	2762.0	0.0	0.0
25%	1.0	2020.0	874.5	0.0	141636175.0	13.0	7.0
50%	1.0	2022.0	2216.5	3.0	290530915.0	34.0	38.5
75%	2.0	2022.0	5573.8	16.0	673869022.0	88.0	87.0
max	8.0	2023.0	52898.0	147.0	3703895074.0	672.0	275.0

Obrázek 17: Tabulka s přehledem rozložení atributů

	artist(s)_name	artist_count	released_year	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts
0	Taylor Swift	1.0	2022.0	86	0.0	156338624	0.0	0.0

Obrázek 18: Tabulka modusů

Převedeme numerické atributy na rozsahy (kategorie), a to následovně:

- `artist_count` – vzhledem k rozložení hodnot (valná většina je 1 umělec) zavedeme pouze dvě kategorie: *single* a *multiple*
- `released_year` – opět vzhledem k rozložení hodnot jsme ustanovili kategorie *<2000*, *2000-2009*, *2010-2020*, *2021-2023*
- `in_spotify_playlists`, `in_spotify_charts`, `in_apple_playlists`, `in_apple_charts`, `in_deezer_playlists`, `in_deezer_charts` a `streams` – u těchto atributů proveme plnění do 10 košů (*binning*) stejné hloubky
- `danceability_%`, `speechiness_%` – rozdělení na kategorie *low*, *medium* a *high*, jejichž hranice určuje první a třetí kvartil, tj. <25%, 25-75%, >75%

Výsledky jsou zobrazeny v tabulce [19].

	artist(s)_name	artist_count	released_year	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts
0	Latto, Jung Kook	multiple	2021-2023	(406.0, 734.0]	(37.0, 147.0]	(121100146.0, 166721963.0]	(34.0, 49.0]	(119.0, 275.0]
1	Myke Towers	single	2021-2023	(1036.0, 1535.0]	(37.0, 147.0]	(121100146.0, 166721963.0]	(34.0, 49.0]	(119.0, 275.0]
2	Olivia Rodrigo	single	2021-2023	(1036.0, 1535.0]	(37.0, 147.0]	(121100146.0, 166721963.0]	(73.0, 111.0]	(119.0, 275.0]
3	Taylor Swift	single	2010-2020	(7302.0, 14101.0]	(37.0, 147.0]	(556072408.0, 811477033.0]	(111.0, 180.0]	(119.0, 275.0]
4	Bad Bunny	single	2021-2023	(3003.0, 4554.0]	(37.0, 147.0]	(290530915.0, 383721650.0]	(73.0, 111.0]	(119.0, 275.0]

Obrázek 19: Datová sada po diskretizaci numerických atributů

3.5 Sada 2 – Transformace kategorických atributů na numerické

Druhá sada má být vhodně připravená pro metody vyžadující numerické vstupy. Potřebujeme tedy transformovat atributy `artist(s)_name`, `key` a `mode`. To, že některé skladby mají více umělců v atributu `artist(s)_name`, jsme se rozhodli ignorovat. Určitě by bylo možné vymyslet nějaké inteligentní kódování (např. binární maskování), které by umožnilo i s touto komplexnější vlastností pracovat. Řešit tento problém by však mělo smysl pouze v krajně specifických případech, které při zadané dolovací úloze nebudeme očekávat.

Transformaci provedeme následovně:

- `artist(s)_name` – automaticky očíslujeme, tj. každé možné hodnotě přiřadíme konkrétní číslo. To je zejména vhodné pro klasifikační algoritmy. Využijeme `LabelEncoder` z knihovny `scikit-learn`, číslování tedy bude začínat od 0. Stejný kodér se dá případně využít i na konverzi z kódu zpět na kategorické hodnoty[5]
- `key` – očíslujeme dle konvence *integer notation*[6] začínající od 0 po 11, kdy první tónině (C) přiřadíme 0 a každé další (po půltónech) číslo o jedna větší
- `mode` – jednoduše přiřadíme štítek – 0 pro durové a 1 pro molové dominantní módy

Pro lepší pochopení v případném dalším používání ještě přejmenujeme nyní zavádějící název atributu `artist(s)_name` na lépe vystihující `artist_code`. Výsledek je zobrazen v tabulce [20].

	artist_code	artist_count	released_year	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts
0	325	2	2023	553	147	141381703	43	263
1	400	1	2023	1474	48	133716286	48	126
2	430	1	2023	1397	113	140003974	94	207
3	557	1	2019	7858	100	800840817	116	207
4	43	1	2023	3133	50	303236322	84	133

Obrázek 20: Datová sada po transformaci na numerické atributy

Hodnoty umístění v žebříčcích a počtu obsazení v playlistech jsou kvantitativně odlišná na různých platformách. Proto provedeme normalizaci. Z grafů rozložení z explorativní analýzy vidíme, že rozložení umístění v žebříčcích jsou na různých platformách téměř stejná, stejně tak zařazení do playlistů. Provedeme proto **kvantilovou normalizaci**, abychom rozložení převedli na taková, která budou mít identické statistické vlastnosti, jak je vidno v [21].

	artist_code	artist_count	released_year	in_spotify_playlists_norm	in_spotify_charts_norm	streams	in_apple_playlists_norm
0	325	2	2023	189.00	160.00	141381703	917.33
1	400	1	2023	508.00	65.83	133716286	1034.00
2	430	1	2023	481.33	126.00	140003974	2153.00
3	557	1	2019	2752.33	114.33	800840817	2746.67
4	43	1	2023	1086.00	69.17	303236322	1849.67

Obrázek 21: Datová sada po normalizaci

[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
[6] https://en.wikipedia.org/wiki/Pitch_class

4 Závěr

Cílem této části projektu bylo provést explorativní analýzu na zvolené datové sadě „Most Streamed Spotify Songs 2023“ s cílem připravit data pro dolovací úlohu predikce oblíbenosti písně na základě různých atributů jako je rytmus, tónina, mód a další.

Prvním krokem bylo prozkoumání jednotlivých atributů datové sady, včetně jejich typu a rozložení hodnot. Následovala analýza odlehlých hodnot, kdy jsme využili především vizualizační techniky. Následovala analýza chybějících hodnot, kterých nebylo mnoho. Dalším důležitým krokem byla korelační analýza numerických atributů, která poskytla vhled do vzájemných vztahů mezi různými atributy.

Pro přípravu dat pro dolovací algoritmy byly navrženy dvě varianty datové sady. V obou případech se nejprve odstranily irelevantních atributy, chybějící hodnoty a vypořádalo se s odlehlými hodnotami. První varianta se za pomocí diskretizace připravila pro dolovací úlohu, která očekává na vstupu pouze kategorické atributy. Druhá varianta pak pracovala s transformací kategorických atributů na numerické a jejich následovanou normalizací.

Celkově projekt nejdříve důkladně analyzoval obsah datové sady z hlediska obsahu i kvality. V druhé části se datová sada připravila pro další fáze dolování dat, konkrétně predikci oblíbenosti písně.