# Příprava dat a jejich popisná charakteristika

November 22, 2023

**Mário Harvan (xharva03)**

**Tereza Burianová (xburia28)**

## 1 Explorativní analýza

Pro účely projektu byla zvolena datová sada Most Streamed Spotify Songs 2023. Cílem projektu je explorativní analýza a příprava datové sady pro dolovací úlohu predikující oblíbenost skladby na základě jejich vlastností.

### 1.1 Atributy datové sady

Nejprve načteme dataset z csv formátu do dataframe Pandas. Následně ukážeme prvních pár řádků datové sady, pro představu, jak datová sada vypadá. Můžeme vidět, že většina atributů je numerických, ale najdeme i kategorické. Z datové sady se můžeme o každé skladbě dozvědět jména tvůrců, název skladby, její popularitu, v kolika playlistech se vyskytuje. Dále následují vlastnosti, které specifikují typ skladby. Atributy jsou například BPM (tempo), následně percentuální vyjádření vlastností jako například akustika, energie a instrumentalita.

Ukázka datové sady:

| | track_name | artist(s)_name | artist_count | released_year | released_month | released_day | in_spotify_playlists | in_spotify_charts | streams | in_apple_playlists | … | bpm | key | mode |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Seven (feat. Latto) (Explicit Ver.) | Latto, Jung Kook | 2 | 2023 | 7 | 14 | 553 | 147 | 141381703 | 43 | … | 125 | B | Major |
| 1 | LALA | Myke Towers | 1 | 2023 | 3 | 23 | 1474 | 48 | 133716286 | 48 | … | 92 | C# | Major |
| 2 | vampire | Olivia Rodrigo | 1 | 2023 | 6 | 30 | 1397 | 113 | 140003974 | 94 | … | 138 | F | Major |
| 3 | Cruel Summer | Taylor Swift | 1 | 2019 | 8 | 23 | 7858 | 100 | 800840817 | 116 | … | 170 | A | Major |
| 4 | WHERE SHE GOES | Bad Bunny | 1 | 2023 | 5 | 18 | 3133 | | | | | | | |

```
4                 50  303236322                  84  …  144   A  Minor

  danceability_%  valence_% energy_% acousticness_%  instrumentalness_%  \
0             80         89       83             31                   0
1             71         61       74              7                   0
2             51         32       53             17                   0
3             55         58       72             11                   0
4             65         23       80             14                  63

   liveness_%  speechiness_%
0           8              4
1          10              4
2          31              6
3          11             15
4          11              6

[5 rows x 24 columns]


Informace o datové sadě:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 953 entries, 0 to 952
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   track_name            953 non-null    object
 1   artist(s)_name        953 non-null    object
 2   artist_count          953 non-null    int64
 3   released_year         953 non-null    int64
 4   released_month        953 non-null    int64
 5   released_day          953 non-null    int64
 6   in_spotify_playlists  953 non-null    int64
 7   in_spotify_charts     953 non-null    int64
 8   streams               953 non-null    object
 9   in_apple_playlists    953 non-null    int64
 10  in_apple_charts       953 non-null    int64
 11  in_deezer_playlists   953 non-null    object
 12  in_deezer_charts      953 non-null    int64
 13  in_shazam_charts      903 non-null    object
 14  bpm                   953 non-null    int64
 15  key                   858 non-null    object
 16  mode                  953 non-null    object
 17  danceability_%        953 non-null    int64
 18  valence_%             953 non-null    int64
 19  energy_%              953 non-null    int64
 20  acousticness_%        953 non-null    int64
 21  instrumentalness_%    953 non-null    int64
 22  liveness_%            953 non-null    int64
```

```
 23  speechiness_%         953 non-null    int64
dtypes: int64(17), object(7)
memory usage: 178.8+ KB
None
```

### 1.2 Převod atributů a chybějící hodnoty

V základních informacích o datasetu bylo zjištěno, že atributy *'streams'*, *'in_deezer_playlists'* a *'in_shazam_charts'*, které by měly obsahovat numerické hodnoty, jsou typu *'object'*. Pro další provedení explorativní analýzy by bylo vhodné takové atributy převést na numerické již v této fázi.

Po provedení analýzy pro atribut *'streams'* bylo zjištěno, že jedna z hodnot je chybně zadaná. Hodnota pro danou skladbu byla ručně zjištěna v aplikaci Spotify a doplněna do datové sady. Ostatní validní hodnoty byly převedeny na numerické.

U atributu *'in_deezer_playlists'* bylo zjištěno, že u vyšších hodnot je použita čárka jakožto oddělovač řádů, která zabraňuje v převedení hodnoty na numerickou. Čárky byly odstraněny a převedení atributu na numerický tak bylo umožněno.

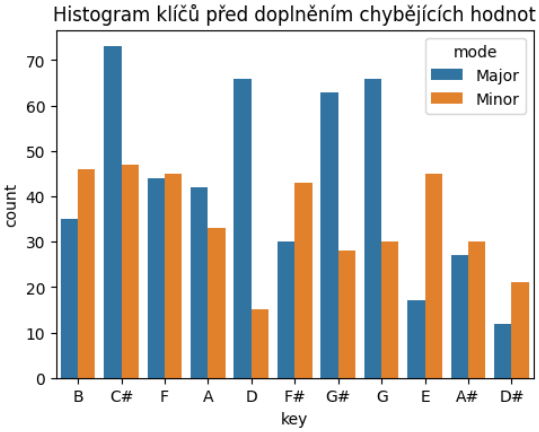
Atribut *'in_shazam_charts'* taktéž obsahoval čárky u některých hodnot, které byly odstraněny stejným způsobem, jako při předchozím atributu. Dále také bylo zjištěno, že sloupec využívá hodnotu NaN pro vyjádření, že se skladba v žebříčku neumístila. Tyto hodnoty byly nastaveny na 0, aby atribut odpovídal ostatním atributům vyjadřujícím umístění.

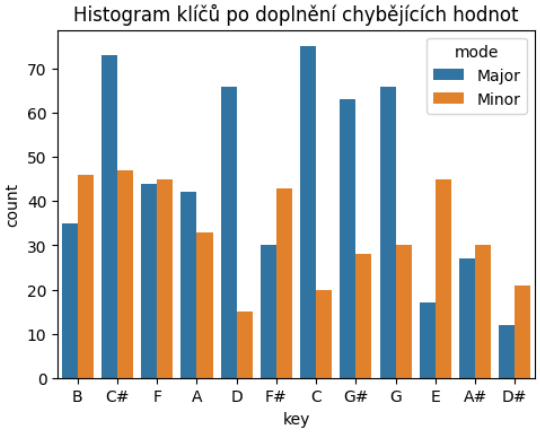
```
Index nevalidní numerické hodnoty v sloupci 'streams': [574]
Nevalidní hodnota v sloupci 'streams':
BPM110KeyAModeMajorDanceability53Valence75Energy69Acousticness7Instrumentalness0
Liveness17Speechiness3
Typ převedeného sloupce 'streams': float64; Počet nevalidních hodnot: 0

Příklad nevalidních numerických hodnot v sloupci 'in_deezer_playlists':
48    2,445
54    3,394
55    3,421
65    4,053
73    1,056
Name: in_deezer_playlists, dtype: object
Typ převedeného sloupce 'in_deezer_playlists': int64; Počet nevalidních hodnot:
0

Příklad nevalidních numerických hodnot v sloupci 'in_shazam_charts':
12    1,021
13    1,281
14      NaN
17    1,173
24    1,093
Name: in_shazam_charts, dtype: object
Typ převedeného sloupce 'in_shazam_charts': int64; Počet nevalidních hodnot: 0
```

Pro zajištění validity výsledků explorativní analýzy by taktéž bylo vhodné odstranění chybějících

hodnot. Bylo zjištěno, že po převedení objektů na numerické hodnoty jsou chybějící hodnoty pouze při atributu *'key'*. V grafu lze vidět rozložení klíčů pro jednotlivé stupnice. Zvláštností v tomto grafu je, že zde vůbec není zastoupen klíč "C", zatímco "C Major" byl na základě analýzy Spotify určeno jako nejzastoupenější klíč. Lze tedy usoudit, že chybějící hodnoty klíčů mohou být doplněny hodnotou "C". V grafu zobrazujícím klíče po doplnění je "C Major" vskutku nejzastoupenějším klíčem.

```
Chybějící hodnoty 'keys' pro jednotlivé 'modes':
mode
Major    75
Minor    20
dtype: int64
```




## 1.3 Rozložení hodnot atributů

Ve druhé části se podíváme na základní statistické údaje o jednotlivých atributech. Můžeme vidět počet hodnot, jejich průměr, minima, maxima a směrodatnou odchylku.

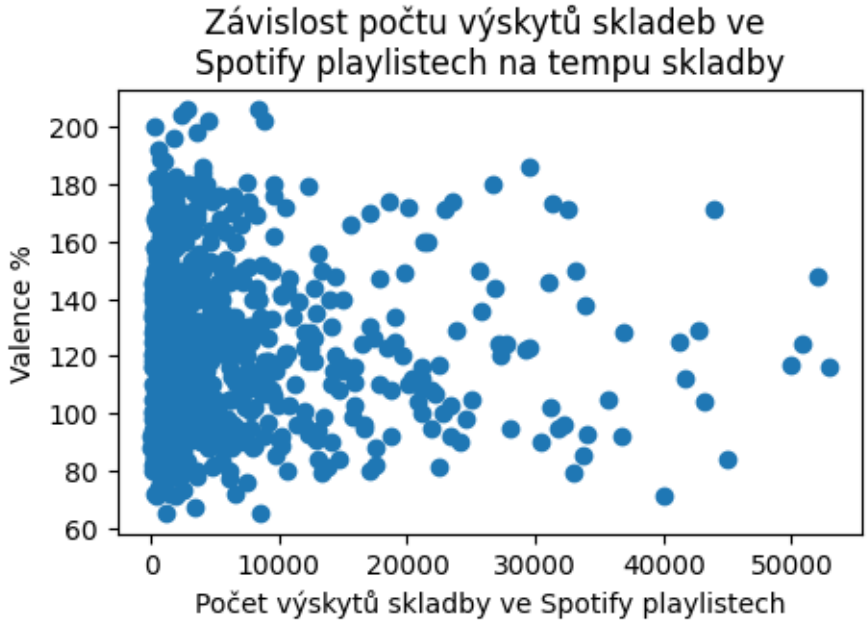
```
Základní statistiky datové sady:
       artist_count  released_year  released_month  released_day  \
count    953.000000     953.000000      953.000000    953.000000
mean       1.556139    2018.238195        6.033578     13.930745
std        0.893044      11.116218        3.566435      9.201949
min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

       in_spotify_playlists  in_spotify_charts       streams  \
count            953.000000         953.000000  9.530000e+02
mean            5200.124869          12.009444  5.138240e+08
```

```
std             7897.608990          19.575992  5.666418e+08
min               31.000000           0.000000  2.762000e+03
25%              875.000000           0.000000  1.417210e+08
50%             2224.000000           3.000000  2.902286e+08
75%             5542.000000          16.000000  6.738011e+08
max            52898.000000         147.000000  3.703895e+09

       in_apple_playlists  in_apple_charts  in_deezer_playlists  \
count          953.000000       953.000000           953.000000
mean            67.812172        51.908709           385.187828
std             86.441493        50.630241          1130.535561
min              0.000000         0.000000             0.000000
25%             13.000000         7.000000            13.000000
50%             34.000000        38.000000            44.000000
75%             88.000000        87.000000           164.000000
max            672.000000       275.000000         12367.000000

       in_deezer_charts  in_shazam_charts         bpm  danceability_%  \
count        953.000000        953.000000  953.000000       953.00000
mean           2.666317         56.847849  122.540399        66.96957
std            6.035599        157.441749   28.057802        14.63061
min            0.000000          0.000000   65.000000        23.00000
25%            0.000000          0.000000  100.000000        57.00000
50%            0.000000          2.000000  121.000000        69.00000
75%            2.000000         33.000000  140.000000        78.00000
max           58.000000       1451.000000  206.000000        96.00000

        valence_%    energy_%  acousticness_%  instrumentalness_%  liveness_%  \
count  953.000000  953.000000      953.000000          953.000000  953.000000
mean    51.431270   64.279119       27.057712            1.581322   18.213012
std     23.480632   16.550526       25.996077            8.409800   13.711223
min      4.000000    9.000000        0.000000            0.000000    3.000000
25%     32.000000   53.000000        6.000000            0.000000   10.000000
50%     51.000000   66.000000       18.000000            0.000000   12.000000
75%     70.000000   77.000000       43.000000            0.000000   24.000000
max     97.000000   97.000000       97.000000           91.000000   97.000000

       speechiness_%
count     953.000000
mean       10.131165
std         9.912888
min         2.000000
25%         4.000000
50%         6.000000
75%        11.000000
max        64.000000
```

Na základě statistik o atributech bylo zjištěno, že všechny z atributů vyjadřujících umístění v že-

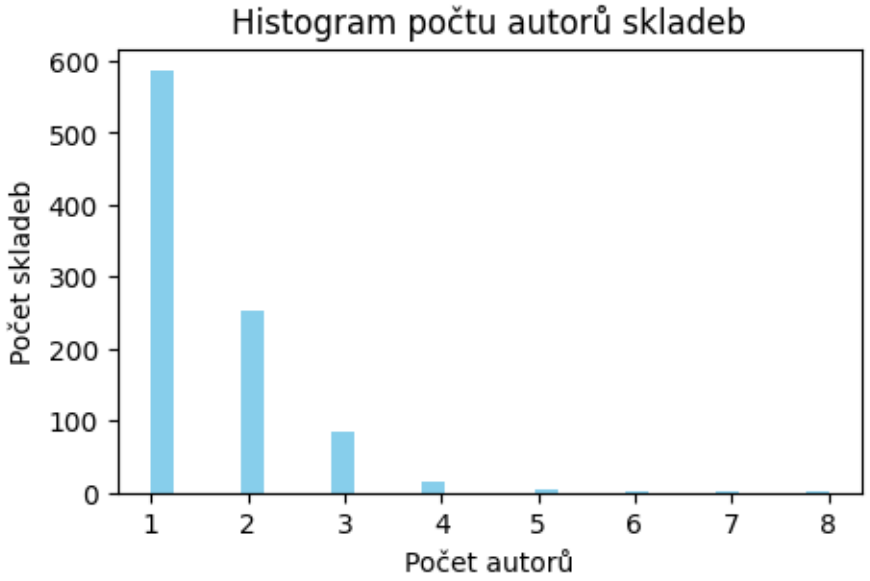
bříčcích nyní obsahují nejnižší hodnotu 0, která označuje, že se skladba v daném žebříčku neumístila. Protože pro tyto atributy znamená nižší číslo lepší umístění, přítomnost nuly jakožto nejnižšího čísla by mohlo zkreslit výsledky další analýzy, například korelační matice. Proto byly hodnoty 0 nahrazeny hodnotou o jedna vyšší, než je maximální zjištěná hodnota pro daný žebříček.

V prvním grafu jsme zkoumali popularitu skladeb ve spotify playlistech v závislosti na tempu. Z grafu vidíme, že popularita skladby nezávisí na tempu.

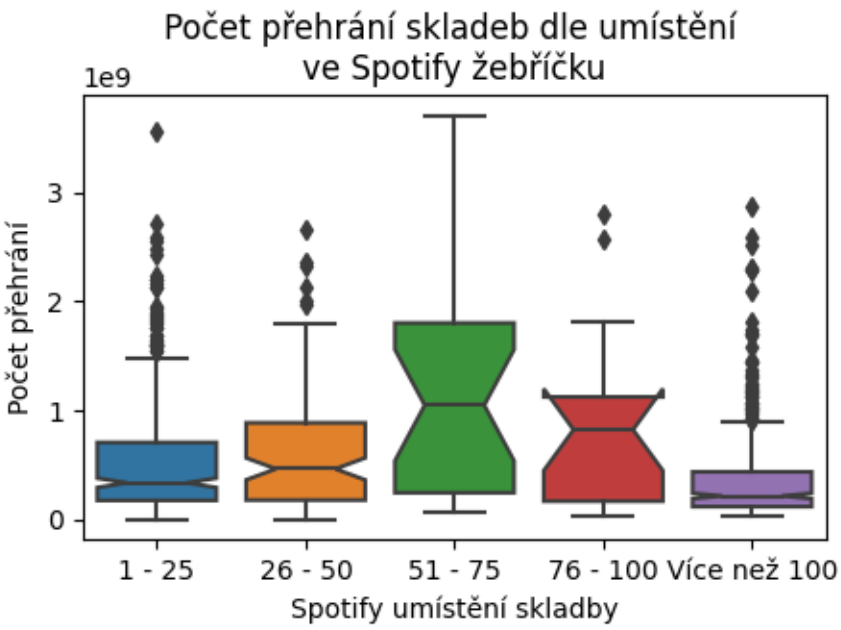



Ve druhém grafu jsme prozkoumali závislost popularity ve spotify a apple playlistech. V grafu pozorujeme lineární závilost, čili skladba, která je populární ve Spotify playlistu, bude populární i v Apple playlistu.

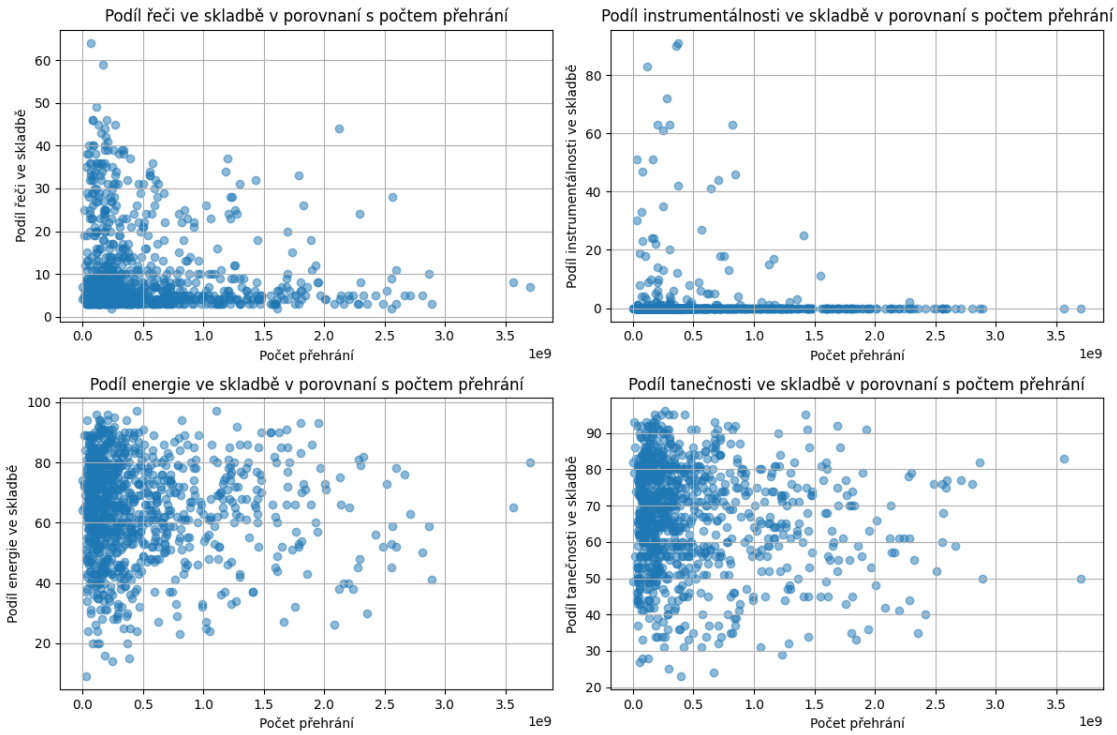

Třetí graf znázorňuje histogram počtu autorů skladby. Z grafu je jasně vidět, že nejvíce skladeb vytvořil právě jeden autor.




V krabicovém grafu zkoumáme závislost počtu přehrání skladeb na umístění ve Spotify žebříčku. Příčky, na kterých se skladby umístily, byly rozděleny do 4 skupin: 1-25, 26-50, 51-75, 76-100. Zajímavé je, že nejvíce přehrání mají skladby, které se umístily na příčkách 51-100. Očekávali bychom, že nejlépe umístěné skladby budou i nejpopulárnější.




V poslední skupině grafů jsme porovnávali různé kvanitativní parametry skladeb v závislosti na počtu přehrání. Porovnali jsme podíl řeči, instrumentálů, energie a tanečnosti. Zjistili jsme, že skladby, které mají menší podíl řeči, jsou populárnější. Také jsou populárnější skladby, které mají téměř nulový podíl instrumentálů. Naopak písničky s větším podílem energie a tanečnosti jsou populárnější.



### 1.4 Odlehlé hodnoty

Během analýzy odlehlých hodnot jsme nejprve analyzovali minima a maxima každého atributu. Pro percentuální atributy musely být všechny hodnoty v rozmezí 0 – 100, což dataset splňuje. Jediné, co nás zaujalo, je minimum atributu *'released_year'*, které má hodnotu 1930. Proto jsme vypsali všechny skladby, které mají rok vydání menší než 1960, a hodnoty jsme následně ověřili. Všechny hodnoty se opravdu shodují s rokem vydání, nebylo tedy nutné provádět žádné úpravy.

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

| | in_spotify_playlists | in_spotify_charts | streams | \ |
|---|---|---|---|---|
| count | 953.000000 | 953.000000 | 9.530000e+02 | |
| mean | 5200.124869 | 74.905561 | 5.138240e+08 | |
| std | 7897.608990 | 65.032793 | 5.666418e+08 | |
| min | 31.000000 | 1.000000 | 2.762000e+03 | |
| 25% | 875.000000 | 11.000000 | 1.417210e+08 | |
| 50% | 2224.000000 | 43.000000 | 2.902286e+08 | |
| 75% | 5542.000000 | 148.000000 | 6.738011e+08 | |
| max | 52898.000000 | 148.000000 | 3.703895e+09 | |

| | in_apple_playlists | in_apple_charts | in_deezer_playlists | \ |
|---|---|---|---|---|
| count | 953.000000 | 953.000000 | 953.000000 | |
| mean | 67.812172 | 80.869885 | 385.187828 | |
| std | 86.441493 | 81.948999 | 1130.535561 | |
| min | 0.000000 | 1.000000 | 0.000000 | |
| 25% | 13.000000 | 16.000000 | 13.000000 | |
| 50% | 34.000000 | 59.000000 | 44.000000 | |
| 75% | 88.000000 | 109.000000 | 164.000000 | |
| max | 672.000000 | 276.000000 | 12367.000000 | |

| | in_deezer_charts | in_shazam_charts | bpm | danceability_% | \ |
|---|---|---|---|---|---|
| count | 953.000000 | 953.000000 | 953.000000 | 953.00000 | |
| mean | 37.211962 | 657.150052 | 122.540399 | 66.96957 | |
| std | 26.413908 | 684.308083 | 28.057802 | 14.63061 | |
| min | 1.000000 | 1.000000 | 65.000000 | 23.00000 | |
| 25% | 5.000000 | 14.000000 | 100.000000 | 57.00000 | |
| 50% | 59.000000 | 187.000000 | 121.000000 | 69.00000 | |

```
75%           59.000000       1452.000000  140.000000        78.00000
max           59.000000       1452.000000  206.000000        96.00000

        valence_%    energy_%  acousticness_%  instrumentalness_%  liveness_%  \
count  953.000000  953.000000      953.000000          953.000000  953.000000
mean    51.431270   64.279119       27.057712            1.581322   18.213012
std     23.480632   16.550526       25.996077            8.409800   13.711223
min      4.000000    9.000000        0.000000            0.000000    3.000000
25%     32.000000   53.000000        6.000000            0.000000   10.000000
50%     51.000000   66.000000       18.000000            0.000000   12.000000
75%     70.000000   77.000000       43.000000            0.000000   24.000000
max     97.000000   97.000000       97.000000           91.000000   97.000000

       speechiness_%
count     953.000000
mean       10.131165
std         9.912888
min         2.000000
25%         4.000000
50%         6.000000
75%        11.000000
max        64.000000
Out-of-range values in 'released_year' column:
                                                track_name  \
439                                      Agudo Mï¿½iï¿½gi
443                     Rockin' Around The Christmas Tree
444                                     Jingle Bell Rock
448                Let It Snow! Let It Snow! Let It Snow!
459              A Holly Jolly Christmas - Single Version
460  The Christmas Song (Merry Christmas To You) - …
466                Let It Snow! Let It Snow! Let It Snow!
469                                     White Christmas
476  It's Beginning to Look a Lot Like Christmas (w…
483                       Deck The Hall - Remastered 1999
495                     Run Rudolph Run - Single Version
496                       Jingle Bells - Remastered 1999

                                        artist(s)_name  artist_count  \
439                              Styrx, utku INC, Thezth             3
443                                        Brenda Lee             1
444                                       Bobby Helms             1
448                                       Dean Martin             1
459                                         Burl Ives             1
460                                      Nat King Cole             1
466                    Frank Sinatra, B. Swanson Quartet             2
469  Bing Crosby, John Scott Trotter & His Orchestr…             3
476  Perry Como, The Fontane Sisters, Mitchell Ayre…             3
483                                      Nat King Cole             1
```

```
495                                  Chuck Berry             1
496                                Frank Sinatra             1

     released_year  released_month  released_day  in_spotify_playlists  \
439           1930               1             1                   323
443           1958               1             1                 14994
444           1957               1             1                 10326
448           1959              11            16                  6512
459           1952               1             1                  7930
460           1946              11             1                 11500
466           1950               1             1                 10585
469           1942               1             1                 11940
476           1958               1             1                  6290
483           1959               1             1                  3299
495           1958               1             1                  8612
496           1957               1             1                  4326

     in_spotify_charts      streams  in_apple_playlists  …  key   mode  \
439                148   90598517.0                   4  …   F#  Minor
443                148  769213520.0                 191  …   G#  Major
444                148  741301563.0                 165  …    D  Major
448                148  446390129.0                  88  …   C#  Major
459                148  395591396.0                 108  …    C  Major
460                148  389771964.0                 140  …   C#  Major
466                148  473248298.0                 126  …    D  Major
469                148  395591396.0                  73  …    A  Major
476                148  295998468.0                  89  …    G  Major
483                148  127027715.0                  65  …   F#  Minor
495                148  245350949.0                 120  …    G  Minor
496                148  178660459.0                  32  …   G#  Major

     danceability_%  valence_%  energy_% acousticness_% instrumentalness_%  \
439              65         49        80              22                   4
443              70         85        41              71                   0
444              74         78        37              84                   0
448              45         72        24              91                   0
459              67         81        36              64                   0
460              36         22        15              84                   0
466              60         86        32              88                   0
469              23         19        25              91                   0
476              73         72        32              77                   0
483              69         96        36              81                   0
495              69         94        71              79                   0
496              51         94        34              73                   0

     liveness_%  speechiness_%  ranking_group
439           7              5   Více než 100
443          45              5   Více než 100
```

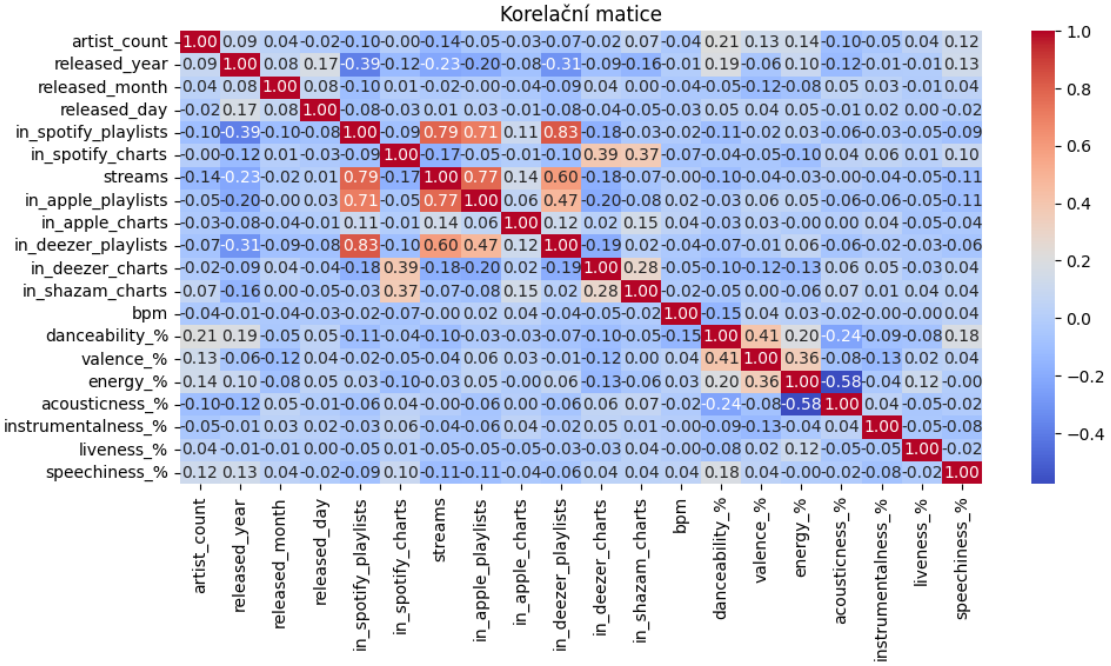
```
444            6                3   Více než 100
448           18                4   Více než 100
459           15                3   Více než 100
460           11                4   Více než 100
466           34                6   Více než 100
469           40                3   Více než 100
476           15                5   Více než 100
483            8                4   Více než 100
495            7                8   Více než 100
496           10                5   Více než 100

[12 rows x 25 columns]
```

## 1.5 Korelační analýza

Z korelační matice lze zjistit hned několik zajímavých informací. U skladeb lze vidět vyšší pozitivní korelace pro počet přehrání a výskyt skladeb v playlistech, ovšem v podstatě nulová korelace pro umístění skladeb v žebříčcích a počet přehrání nebo výskyt v playlistech. Lze tedy zjistit, že pro danou datovou sadu nepozorujeme ovlivnění popularity skladby jejím výskytem v žebříčku.

Další zajímavou informací je negativní korelace mezi akustičností skladby a energií skladby. Lze tak soudit, že akustické skladby jsou klidnější, naopak elektronické skladby jsou více energické.

Dále lze pozorovat pozitivní korelaci mezi vhodností skladby pro tanec, energií a valencí (pozitivitou) skladby. Z tohoto zjištění lze usoudit, že dané skladby jsou často zároveň energické, pozitivní a vhodné pro tanec.



# 2 Příprava dat

Po explorativní analýze přichází fáze přípravy dat, kde jsou data převedena do formy vhodné pro použití dolovacím algoritmem. Tento proces zahrnuje odstranění přebytečných atributů, vypořádání se s chybějícími a odlehlými hodnotami a převedení na kategorické, případně numerické, atributy, dle potřeb daného dolovacího algoritmu. Některé z popisovaných úprav bylo třeba provést již ve fázi explorativní analýzy z důvodu získání grafů s validními daty. Pro tento projekt byla zvolena dolovací úloha predikce oblíbenosti písně na základě vlastností skladby.

## 2.1 Odstranění přebytečných atributů

Fáze přípravy dat začíná odstraněním atributů, které jsou pro danou dolovací úlohu nepotřebné. V tomto případě se jedná o sloupce obsahující název skladby, jména a počet umělců a rok, měsíc a den vydání. Zbylé sloupce obsahují hodnoty znázorňující kolikrát se daná skladba vyskytuje v playlistech na různých platformách, umístění písně v různých žebříčcích, počet přehrání skladby a vlastnosti skladby. Tyto sloupce jsou plně dostačující pro získání informací pro určenou dolovací úlohu.

```
Atributy po odstranění přebytečných informací:
['in_spotify_playlists', 'in_spotify_charts', 'streams', 'in_apple_playlists',
'in_apple_charts', 'in_deezer_playlists', 'in_deezer_charts',
'in_shazam_charts', 'bpm', 'key', 'mode', 'danceability_%', 'valence_%',
'energy_%', 'acousticness_%', 'instrumentalness_%', 'liveness_%',
'speechiness_%']
```

Další fází je zpracování atributů určujících popularitu skladby - počet výskytů v playlistech, umístění v žebříčcích a počet streamů.

Atributy pro počet výskytů skladby v playlistech, jmenovitě *'in_spotify_playlists'*, *'in_apple_playlists'* a *'in_deezer_playlists'*, mohou být zkombinovány do jednoho atributu, *'in_playlists'*, sečtením všech těchto atributů.

```
['in_spotify_charts', 'streams', 'in_apple_charts', 'in_deezer_charts',
'in_shazam_charts', 'bpm', 'key', 'mode', 'danceability_%', 'valence_%',
'energy_%', 'acousticness_%', 'instrumentalness_%', 'liveness_%',
'speechiness_%', 'in_playlists']
```

Při atributech znázorňujících umístění skladby v žebříčcích byly ve fázi explorativní analýzy všechny nulové hodnoty nahrazeny hodnotami o jedna vyšší, než je maximální hodnota daného atributu. Skladby, které se v daném žebříčku neumístily, jsou tak definovány jako maximální hodnota, a vztah, kdy nižší hodnota znamená lepší umístění, bude zachován i při sečtení všech atributů znázorňujících umístění v žebříčcích. Atributy *'in_spotify_charts'*, *'in_apple_charts'*, *'in_deezer_charts'* a *'in_shazam_charts'* byly tedy sečtením zkombinovány do společného sloupce *'in_charts'*.

```
['streams', 'bpm', 'key', 'mode', 'danceability_%', 'valence_%', 'energy_%',
'acousticness_%', 'instrumentalness_%', 'liveness_%', 'speechiness_%',
'in_playlists', 'in_charts']
```

Posledními atributy, které mohou být kombinovány, jsou stupnice (*'mode'*) a klíč (*'key'*). Tyto hodnoty jsou vždy uváděny a analyzovány společně, neboť spolu přímo souvisí, není tedy obvykle třeba je rozdělit do dvou atributů.

```
['streams', 'bpm', 'danceability_%', 'valence_%', 'energy_%', 'acousticness_%',
'instrumentalness_%', 'liveness_%', 'speechiness_%', 'in_playlists',
'in_charts', 'scale']
0     B Major
1    C# Major
2     F Major
3     A Major
4     A Minor
Name: scale, dtype: object
```

## 2.2 Chybějící a odlehlé hodnoty

Tyto hodnoty byly řešeny již ve fázi explorativní analýzy z důvodu důležitosti jejich doplnění či odstranění pro správnost zobrazených grafů.

Nejprve se jednalo o odstranění chybějících hodnot ve fázi převádění numerických atributů z typu *'object'* na numerický typ. V datové sadě byl nesprávně zadán atribut *'streams'* u jedné z písní, kde se tato hodnota podařila ručně dohledat jako celkový počet přehrání ve Spotify aplikaci pro danou píseň. Dále byly odstraněny chybějící hodnoty atributu *'in_shazam_charts'*, kde bylo chybějící hodnotou znázorněno, že se píseň v žebříčku neumístila. Chybějící hodnoty byly doplněny hodnotou *'0'*, aby atribut souhlasil s formátem ostatních atributů popisujících umístění v žebříčcích.

Dále zbývalo doplnění chybějících hodnot atributu *'key'*, kde byl vykreslen histogram všech zastoupených klíčů. Pro doplnění byly vyhledány nejčastěji se vyskytující klíče na základě analýzy Spotify, kde nejčastější zastoupený klíč, tedy "C Major", se v datové sadě vůbec nevyskytoval. Chybějící hocnoty tedy byly doplněny nejpravděpodobnější hodnotou.

Všechny chybějící hodnoty byly doplněny a žádný ze řádků neobsahoval více chybějících hodnot, nebo takové hodnoty, které by nebylo možné doplnit, nebylo tedy nutné odstranění žádných záznamů.

Na základě analýzy odlehlých hodnot nebyly nalezeny žádné odlehlé hodnoty, které by naznačovaly chybovost. Jediným potenciálně chybovým atributem byl rok vydání, kde byly všechny podezřelé hodnoty ověřeny jako správné. Pro přípravu dat pro určenou dolovací úlohu ovšem rok vydání není klíčový.

## 2.3 Datová sada s kategorickými atributy

Pro vytvoření datové sady obsahující kategorické atributy je třeba provedení diskretizace numerických atributů. Jedním z přístupů je plnění ("binning"), kde jsou hodnoty rozděleny na základě několika intervalů, které jsou pojmenovány, což vede k převedení numerického atributu na kategorický. Rozdělení je možno provést dle šířky intervalu, kde má každý interval stejný rozsah, případně dle hloubky intervalu, kde každý interval obsahuje odpovídající množství hodnot. Pro převedení numerických atributů v použité datové sadě bylo vybráno dělení dle šířky intervalu, neboť je v daném kontextu jednodušší na pochopení a porovnání jednotlivých vlastností.

```
streams               float64
bpm                     int64
danceability_%          int64
valence_%               int64
energy_%                int64
```

```
acousticness_%          int64
instrumentalness_%      int64
liveness_%              int64
speechiness_%           int64
in_playlists            int64
in_charts               int64
scale                  object
dtype: object
```

```
0      high
1      high
2       low
3    medium
4    medium
Name: danceability_%, dtype: category
Categories (5, object): ['very_low' < 'low' < 'medium' < 'high' < 'very_high']
```

Po převedení všech numerických atributů obdobným způsobem jsou všechny atributy kategorické.

```
streams               category
bpm                   category
danceability_%        category
valence_%             category
energy_%              category
acousticness_%        category
instrumentalness_%    category
liveness_%            category
speechiness_%         category
in_playlists          category
in_charts             category
scale                   object
dtype: object
```

## 2.4 Datová sada s numerickými atributy

K vytvoření datové sady s numerickými atributy je třeba převedení kategorických atributů na numerické. Jednou z metod pro tento převod je "kód 1 z n" (one-hot encoding), kde je každé unikátní hodnotě přiřazena celočíselná hodnota, kterou je hodnota v novém numerickém atributu vyjádřena.

```
Před kódováním:
0     B Major
1    C# Major
2     F Major
3     A Major
4     A Minor
Name: scale, dtype: object
Po kódování:
0     4
```

```
1     8
2    16
3     0
4     1
Name: scale, dtype: int8
Tabulka převedených atributů:
{0: 'A Major',
 1: 'A Minor',
 2: 'A# Major',
 3: 'A# Minor',
 4: 'B Major',
 5: 'B Minor',
 6: 'C Major',
 7: 'C Minor',
 8: 'C# Major',
 9: 'C# Minor',
 10: 'D Major',
 11: 'D Minor',
 12: 'D# Major',
 13: 'D# Minor',
 14: 'E Major',
 15: 'E Minor',
 16: 'F Major',
 17: 'F Minor',
 18: 'F# Major',
 19: 'F# Minor',
 20: 'G Major',
 21: 'G Minor',
 22: 'G# Major',
 23: 'G# Minor'}
```

Dalším krokem je normalizace numerických atributů. Jedním z atributů, který bude rozhodně třeba normalizovat, je atribut *'streams'*, který obsahuje velmi vysoká čísla. Dále by bylo vhodné převedení percentuálních hodnot na hodnoty v rozmezí 0 až 1. Tento převod není nutný pro všechny dolovací algoritmy, ovšem pro ukázku byl proveden. Pro normalizaci všech zmíněných atribtutů je vhodná metoda normalizace změnou dekadického měřítka, neboť z vysokých hodnot udělá nižší a percentuální hodnoty převede na měřítko 0 až 1 dělením hodnotou 100.

Po provedení normalizace lze vidět v následující tabulce, že všechny z percentuálních hodnot jsou nyní v rozmezí 0 až 1 a počet přehrání se pohybuje ve značně nižších číslech, neboť byl na základě statistických informací z explorativní analýzy vydělen řádem milionů.

```
         streams  danceability_%  valence_%  energy_%  acousticness_%  \
min     0.002762            0.23       0.04      0.09            0.00
max  3703.895074            0.96       0.97      0.97            0.97

     instrumentalness_%  liveness_%  speechiness_%
min                0.00        0.03           0.02
max                0.91        0.97           0.64
```

Po převedení kategorického atributu a provedení normalizace jsou všechny atributy numerické, datová sada je tedy připravena.

```
streams               float64
bpm                     int64
danceability_%        float64
valence_%             float64
energy_%              float64
acousticness_%        float64
instrumentalness_%    float64
liveness_%            float64
speechiness_%         float64
in_playlists            int64
in_charts               int64
scale                    int8
dtype: object
```