Spotify top 2000 laulude andmestiku analüüs

Karmeli Kaasik

2022-10-15

Sissejuhatus

Töö eesmärgid

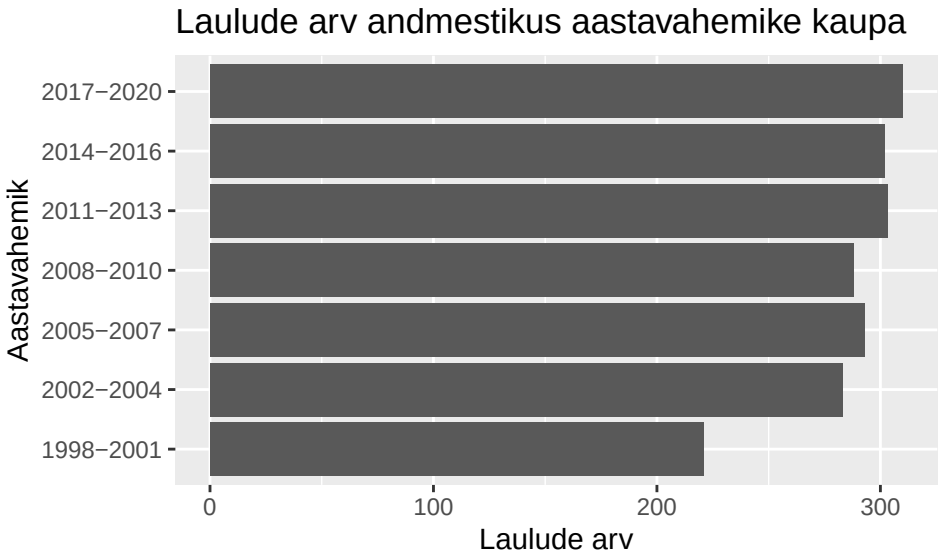
Töö eesmärkideks on tutvustada kasutatavat Spotify laulude paremiku andmestikku ning uurida selles esinevaid seoseid nii laulude kui ka artistide suhtes. Selle jaoks on koostatud neli peamist uurimisküsimust:
1. Milliseid artiste on aastate lõikes andmestikus kõige rohkem välja toodud?
2. Kas ja milline on seos laulu pikkuse ja meeleolu vahel?
3. Millise tempoga lauludega on kõige parem tantsida?
4. Millised on populaarsete laulude omadused?

Andmestiku kirjeldus

Töös kasutatav andmestik on Spotify laulude paremiku andmestik, originaalis "Top Hits Spotify from 2000-2019". Spotify laulude andmestikus on kokku 2000 laulu. Andmestikus leidub laule aastast 1998 kuni aastani 2020, mis erinevad andmestiku originaalpealkirjast. Erinevaid artiste on andmestikus 835. Iga laulu puhul on uuritud 18 tunnust: artist, laulu nimi, laulu kestus, laulu eksplitsiitsus (kas laulus esineb sõnu või fraase, mis pole sobilikud kõikidele kuulajatele), väljalaskeaasta, populaarsus, kuivõrd sobilik on laul tantsimiseks (skaalal 0-1, kus 1 on kõige sobivam), laulu energia ehk intensiivsus (skaalal 0-1), helistik, valjusus (skaalal -60-0 db), modaalsus, sõnakus (millisel määral koosneb laul sõnalistest helidest ja kui suur on muusika osakaal, skaalal 0-1, kus suuremad väärtused tähistavad suuremat rääkimise osakaalu), akustilisus (skaalal 0-1, kus 1 on suurim tõenäosus, et laul on akustiline), instrumentaalsus (skaalal 0-1, kus suuremate väärtuste korral on suurem tõenäosus laulu instrumentaalsuseks), kui tõenäoliselt on laul salvestatud *live*'is (skaalal 0-1), meeleolu (skaalal 0-1, kus madalamad väärtused tähistavad kurvemaid ning suuremad väärtuseid rõõmsamaid laule), laulu tempo ning laulu žanr.

Andmestiku analüüs

Et uurida laulude ja žanrite muutumisi aastate kaupa, pidas autor mõistlikuks jaotada aastad gruppideks: 1998-2001, 2002-2004,2005-2007,2008-2010, 2011-2013, 2014-2016, 2017-2020. Esimene ja viimane grupp on ühe aasta võrra pikemad, kuid seda sellepärast, et aastast 1998 on andmestikus vaid üks ning aastast 2020 kolm laulu.



Laulud on aastavahemike vahel üsna ühtlaselt jaotunud, nagu on näha ülemiselt jooniselt. Kõigis aastavahemikes on ligikaudu 300 laulu, teistest erineb ainult vahemik 1998-2001, millest on andmestikus 221 laulu.

Ülevaade artistidest

Andmestikus oli 835 artisti, kellest 341 artisti on tabelis rohkem kui ühe lauluga. Et uurida artistide muutumist aastate lõikes, loodi tabel, mis kujutab endas aastavahemikke, nendel aastatel kõige enamate lauludega tabelisse sattunud artiste ning mitme lauluga nad antud vahemikus andmestikku jõudsid (vt allolev tabel). Eespool on artistid, kelle poolt oli tabelis selles aastavahemikus rohkem laule, nende võrdsuse korral artistid, kelle teoseid oli andmestikus kokku rohkem. Tabelis on välja toodud iga aastavahemiku kohta viis kõige populaarsemat artisti.

1998-2001	mitu	2002-2004	mitu.1	2005-2007	mitu.2	2008-2010	mitu.3
Britney Spears	6	Eminem	10	Kanye West	8	Rihanna	14
Destiny's Child	6	Avril Lavigne	6	Chris Brown	6	Lady Gaga	8
Kylie Minogue	5	Ashanti	6	Justin Timberlake	6	Black Eyed Peas	6
P!nk	4	Nelly	5	50 Cent	5	Kesha	6
Jennifer Lopez	4	Sugababes	5	The Pussycat Dolls	5	David Guetta	5

2011-2013	mitu.4	2014-2016	mitu.5	2017-2020	mitu.6	kokku	mitu.7
Katy Perry	10	Ariana Grande	8	Post Malone	8	Rihanna	25
Calvin Harris	8	Drake	7	Drake	7	Drake	23
Jason Derulo	7	Calvin Harris	7	Migos	6	Eminem	21
David Guetta	6	Taylor Swift	7	Ariana Grande	5	Calvin Harris	20
Chris Brown	6	Selena Gomez	6	Ed Sheeran	5	Britney Spears	19

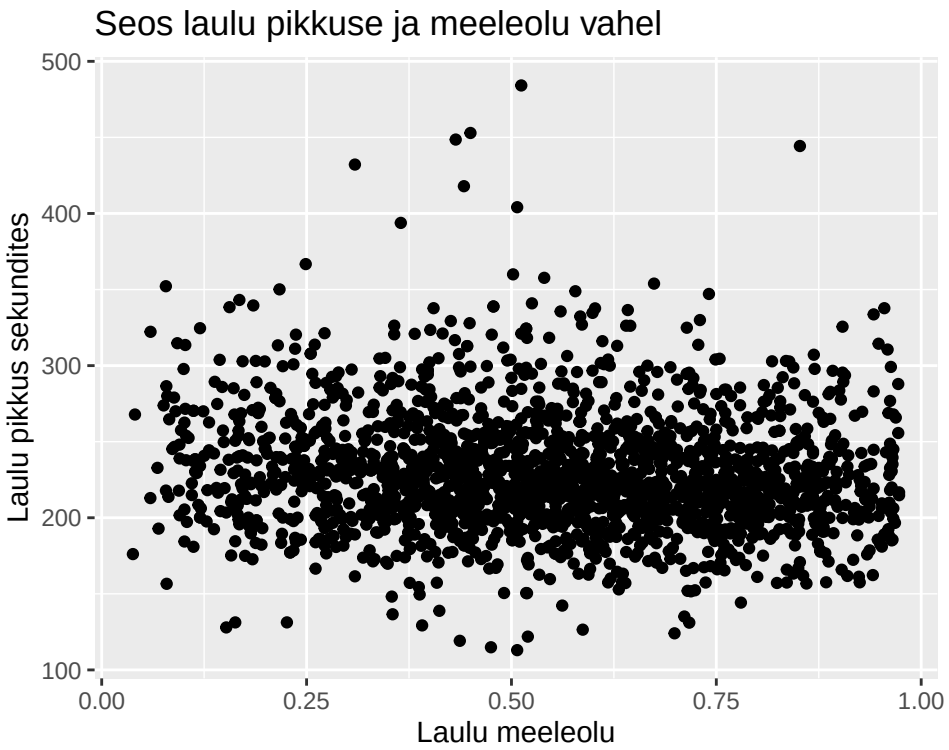
Nagu tabelist on näha, siis igal aastavahemikul on populaarseimad artistid olnud üsna erinevad. Eranditeks on Chris Brown, kes on olnud populaarne nii aastatel 2005-2007 kui ka 2011-2013, David Guetta (aastatel 2008-2010 ning 2011-2013) ja Calvin Harris (2011-2013 ja 2014-2016). Üks huvitavamaid väljavõtteid tabelist on Rihanna 14 populaarset laulu aastatel 2008-2010, mida on tunduvalt rohkem kui kellelgi teisel kõikide aastagruppide peale.

Laulude pikkused ning meeleolud

Andmestikus esinevate laulude keskmine pikkus on 3M 48.7481245S. Aastate jooksul on keskmine laulupikkus märgatavalt lühenenud, olles aastatel 1998-2001 4M 6.7S ja aastatel 2017-2020 3M 26.9S. Kõige pikem laul andmestikus oli Justin Timberlake'i "Mirrors", mille pikkuseks oli 8M 4.1S. Lühima laulu pikkus oli 1M 53S, selleks oli Lil Nas X-i "Old Town Road".

Andmestiku veerg "valence" kirjeldab laulu meeleolu, kus madalamad väärtused tähendavad kurva, vihase vms negatiivne meeleoluga laule ja kõrgemad väärtused märgivad rõõmsamaid ja positiivseid laule. Kõrgeimate meeleolu väärtustega oli andmestikus kaks laulu: Vengaboys-i "Shalala Lala" ning Juanes-i "La Camisa Negra". Kõige negatiivsema meeleoluga oli Martin Garrix-i "Animals", mida ei saa otseselt pidada kurvaks lauluks, kuid kuna negatiivse meeleolu alla lähevad ka vihased laulud, siis on arusaadav selle laulu nii madal hinnang.

Uurimaks, kas laulu pikkuse ning meeleolu vahel on mingi seos, püstitati hüpotees, et negatiivsema meeleoluga laulud on pikemad kui positiivsed laulud. Selle väljaselgitamiseks kasutati Spearmani korrelatsioonikordajat, mille väärtuseks tuli -0.1244486. Seda saab tõlgendada kui peaaegu olematut negatiivset suhet, seega vähemalt nende andmete pealt ei saa hüpoteesi kinnitada ega ka ümber lükata. Laulu pikkuse ja meeleolu suhet on kujutatud ka järgneval joonisel.



Laulu tempo ja tantsimisvõimalikkus

Laulu tempo väärtus kirjeldab laulu BPM-i ehk "lööke minutis" või inglise keeles "beats per minute". Andmestiku laulude tempot saab kirjeldada järgneva tabeli abil:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   60.02   98.99  120.02  120.12  134.27  210.85
```

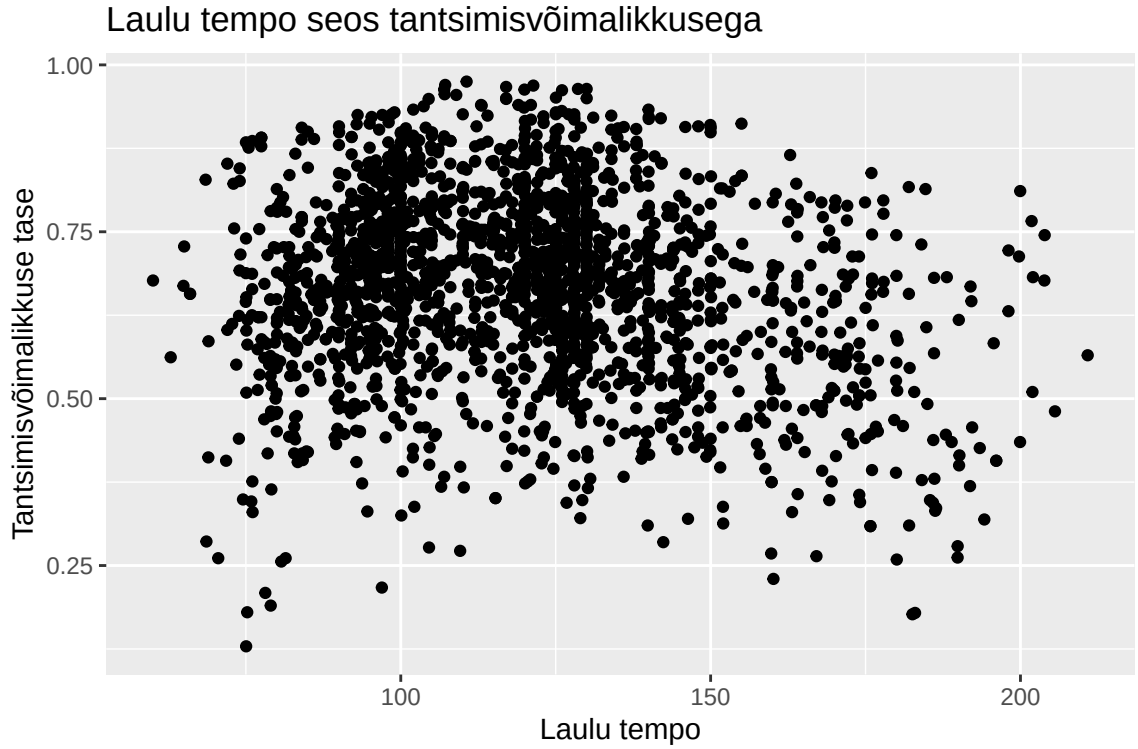
Laulu tantsimisvõimalikkus ehk inglise keeles "danceability" on väärtus 0 ja 1 vahel, mis näitab, kui hästi on võimalik selle laulu taustal tantsida. Suuremad väärtused tähendavad paremat tantslikkust ning väiksemad

näitavad, et selle lauluga pole hea tantsida. Andmestiku laulude keskmine on 0.6674375, mis näitab, et keskmiselt on nende lauludega hea tantsida. Millise tempoga lauludega on aga kõige parem tantsida? Sellele küsimusele vastamiseks jaotame laulud tantsimisvõimalikkuse alusel nelja gruppi: raske tantsida (väärtused vahemikus 0-0.55), okei tantsida(0.55-0-65), hea tantsida(0.65-0.75) ning väga hea tantsida (väärtused üle 0.75).

Uurime iga vahemiku kohta keskmist tempot ja tempo mediaanväärtust. Mõlemad väärtused kahanevad, mida rohkem tantsitavamaks laulud muutuvad. Näiteks on laulude, millega on raske tantsida, keskmine tempo 130, kuid väga hea tantsimisvõimalikkusega lauludes on keskmine tempo 116 BPM. Järgnevas tabelis on näha väga hea tantsimisvõimalikkusega laulude tempo karakteristikud.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.51   99.97  116.97  116.03  127.99  201.80
```

Seega laulude, millega on väga hea tatsida, tempo jääb enamikul lauludel vahmeikku 100 BPM kuni 130BPM. Laulu tantsimisvõimalikkuse ja laulu tempo suhet kujutab ka järgnev joonis, millelt on näha, et väga kiire ja väga aeglase tempoga laulud pole tantsimiseks kõige paremad, hea tempo tantsimiseks jääb tempo skaalal umbes teise neljandikku.



Populaarsete laulude omadused

Kuigi tehniliselt võib kõiki antud andmestikus esinevaid laule pidada populaarseteks, on antud kontekstis mõeldud laule, mille tunnuse "popularity" väärtus on piisavalt suur. Et jaotada laulud populaarsteks ja vähempopulaarseteks, jaotati laulud kolme gruppi: ebapopulaarne (indeks alla 50), keskmise populaarsusega (indeks vahemikus 50-75) ja populaarne, kus populaarsed laulud olid laulud, mille populaarsusindeks oli vähemalt 75. Samuti esines andmestikus 126 laulu, millel populaarsusindeks oli teadmata. Laulude jagunemine populaarsusgruppide vahel on näha järgmises väljatrükis:

```
## # A tibble: 4 x 3
##   popul                    laulude_arv protsent
##   <fct>                          <int>    <dbl>
## 1 ebapopulaarne                    208     10.4
## 2 keskmise populaarsusega         1309     65.4
## 3 populaarne                       357     17.8
## 4 <NA>                             126      6.3
```

Nagu tabelist näha, on ülekaalus keskmise populaarsusega laulud, kuid ka suure populaarsusega laule esineb palju. Järgnevas uurime laule, mida kategoriseeriti kui "populaarne".

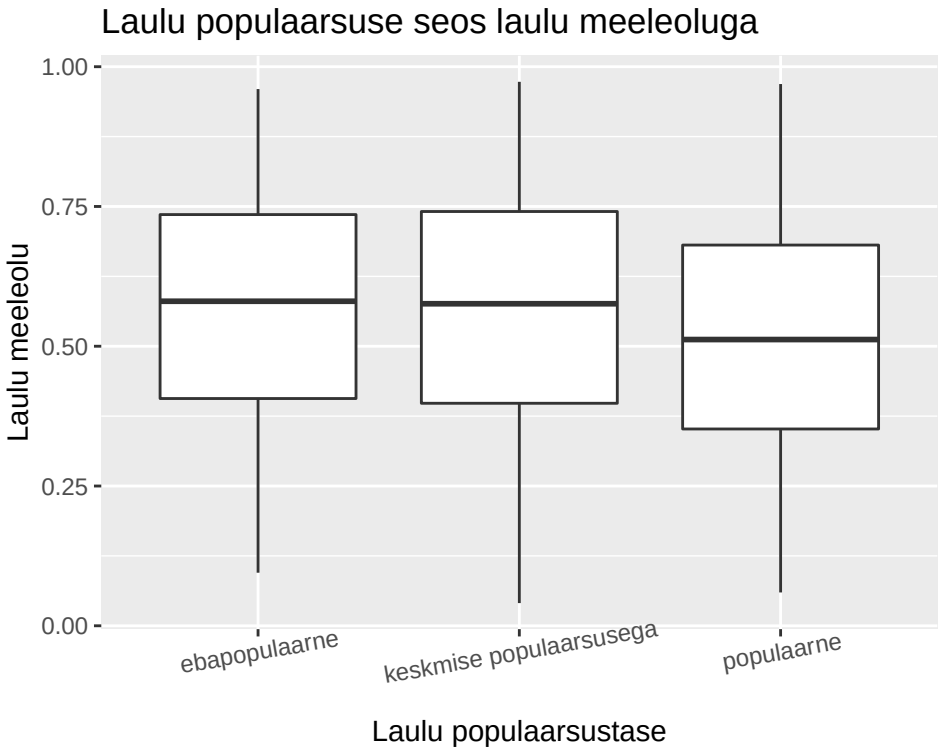
Selgub, et populaarseid laule on kõikidest žanritest, enim esineb pop ja hip-hop stiilis lugusid. Suurimad erinevused võrreldes terve andmestiku laulude žanritega on suurem rock-laulude ning väiksem popmuusika ning R&B osakaal populaarsete laulude seas. Võrdleme populaarsete laulude ja ülejäänud laulude vahel viit erinevat tunnust, et paremini kirjeldada populaarsete laulude omadusi. Nendeks on laulu tempo, eksplitsiitsus, tantsimisvõimalikkus, meeleolu ning valjusus. Võrdeleme esmalt eksplitsiitsust. Vähempopulaarsete laulude hulgas on selliseid laule 20.6 protsenti ning populaarsete laulude hulgas 29.6385542%. Seega võib väita, et populaarsetes lauludes on natuke rohkem ebasobivat sisu, kuid see erinevus ei ole väga suur.

Võrdleme temposid. Mõlema andmehulga minimaalne, keskmine ja mediaantempo on väga sarnased. Väike erinevus tuleb sisse maksimaalses tempos, mis võib viidata sellele, et populaarsete laulude tempo on pisut aeglasem, kuid oluliselt need teineteisest ei erine. Uurime tunnust "danceability" ehk tantsimisvõimalikkus. Ka selle tunnuse poolest on populaarsed laulude kõikide andmestiku lauludega väga sarnased. Märgatav vahe on ainult mininmaalsetel väärtustel: vähempopulaarsete laulude minimaalne tantsimisvõimalikkuse väärtus on 0.129, mis on peaaegu poole väiksem populaarsete laulude minimaalsest 0.217. Sellest võib järeldada, et laulud, mille taustal ei ole üldse võimalik tantsida, ei ole eriti populaarsed. Järgmistel tabelitel on kujutatud laulude valjususe omadusi.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -20.514  -6.476  -5.217  -5.483  -4.166  -0.276
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -13.964  -6.591  -5.507  -5.602  -4.082  -1.190
```

Ülemisel tabelil on näha vähempopulaarsete laulude valjususe kirjeldavaid statistikuid, alumisel populaarsete laulude omi. Kuigi tabeli põhjal paistab, et nii minimaalne kui ka maksimaalne väärtus on populaarstel lauludel väiksem, mis võiks viidata väiksemale hajuvusele, on mõlema tabeli standardhälbed üsnagi võrdsed: populaarsete laulude valjususe standardhälbeks on 2.0528851 ning ülejäänud laulude puhul 1.9017504. Seega ka selles tunnuses kahe tabeli vahel olulist erinevust ei ole. Võrdleme lõpuks laulude meeleolusid. Selleks koostame karpdiagrammi.



Jooniselt on näha, et populaarsete laulude meeleolu väärtus on märgatavalt madalam. Seega võib järeldada, et populaarsed laulud on negatiivsema ja kurvema meeleoluga kui vähempopulaarsed laulud. Seega kokkuvõttes saime, et populaarsed laulud ei ole teistest lauludest väga erinevad tempo, valjususe ning tantsimisvõimalikkuse poolest erinevad. Küll aga on populaarsed laulud natuke negatiivsema meeleoluga ning sisaldavad rohkem ebasobivat sisu kui ülejäänud laulud tabelis.

Kokkuvõte

Kõik uurimisküsimused said vastused. Andmestikus esines mõningaid segadusttekitavaid väärtusi ja tulemusi, mistõttu muudeti teist uurimisküsimust võrreldes esialgsega. Tööst selgus, et andmestikku on sattunud iga aasta erinevaid artiste, kes on välja tulund mitmete populaarsete lauludega. Kõige enam oli andmestikus laule Rihannalt. Tuli välja, et laulu pikkus ja laulu meeleolu vahel olulist seost ei ole. Parim tempo, mille järgi tantsida oli umbes 100 BPM kuni 130 BPM. Kõige populaarsemad laulud teistest lauludest andmestikus eriliselt ei erinenud, nii tempo, valjusus kui ka sobivus tantsimiseks olid sarnased. Natuke erinesid meeleolu, mis oli populaarsetel lauludel negatiivsem ja eksplitsiitsus, mis oli suurem. Kaks populaarseimat žanri olid pop ning hip-hop.