

LAPORAN TUGAS ANALISIS DATA

KU-1102 PENGENALAN KOMPUTASI

Kelas Mahasiswa (K-16) / Kelompok 2

Dosen: Elvayandri, S.Si., M.T.



Anggota Kelompok:

Shafar Zidan Nugraha (16522092)

Marasi Joel Silvano (16522212)

Dama Dhananjaya Daliman (19622002)

Hartanto Luwis (19622032)

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

2022

Analisis Tren dan Popularitas Lagu di Spotify pada Musim Natal 2019

1st Shafar Zidan Nugraha

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

16522092@mahasiswa.itb.ac.id

2nd Marasi Joel Silvano

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

16522212@mahasiswa.itb.ac.id

3rd Dama Dhananjaya Daliman

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

19622002@mahasiswa.itb.ac.id

4th Hartanto Luwis

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

19622032@mahasiswa.itb.ac.id

Abstract—In this growing era, there is a lot of need for digital entertainment, one of which is in the form of songs. The digital platform Spotify provides a streaming service for online songs and is currently very popular with many young people. This report was compiled to analyze the trends and popularity of songs on Spotify in the 2019 Christmas season. Through stages of analysis and data visualization, the team of analysts was able to find trends from the data and salient characteristics of popular songs in the 2019 Christmas season.

Keywords—*data analysis, entertainment, songs, trends, Spotify*

Abstrak—Pada zaman yang berkembang ini banyak sekali kebutuhan akan hiburan digital, salah satunya adalah dalam bentuk lagu. Platform digital Spotify menyediakan layanan *streaming* lagu daring dan saat ini sangat digemari oleh banyak orang muda. Laporan ini disusun untuk menganalisis bagaimana tren dan popularitas lagu-lagu yang ada di Spotify pada musim Natal tahun 2019. Melalui tahap-tahap analisis dan visualisasi data, tim analis berhasil menemukan tren-tren dari data dan karakteristik yang menonjol dari lagu-lagu yang populer pada musim Natal 2019.

Kata Kunci—*analisis data, lagu, hiburan, tren, spotify*

I. PENDAHULUAN

Setelah menginjak era digital dan internet, banyak sekali media hiburan yang bisa kita akses dari mana saja. Cukup menghubungkan gawai ke internet dan akses ke platform penyedia-penyedia layanan hiburan sudah ada di ujung jari. Salah satu media hiburan yang sering sekali dikonsumsi, baik oleh orang tua maupun muda, adalah lagu. Lagu adalah ragam suara yang berirama [1]. Spotify, sebagai salah satu layanan penyedia *streaming* lagu dari Swedia, memberikan akses mudah dan gratis bagi semua orang untuk menikmati lagu. Lantas bagaimana preferensi pengguna dalam mendengarkan lagu-lagu yang tersedia di Spotify? Apakah ada hubungan genre lagu tertentu dengan popularitasnya? Apakah tahun rilis lagu memengaruhi tingkat popularitas lagu? Berlandaskan pertanyaan-pertanyaan ini, tim analis termotivasi untuk menganalisis data popularitas lagu di Spotify.

II. PERSIAPAN DAN PRAPEMROSESAN DATA

A. Deskripsi Data dan File

Data yang dianalisis merupakan data tentang Popularitas Lagu pada Spotify dari 17 negara dan global. Data ini memuat informasi mengenai lagu-lagu yang mendapatkan popularitas pada aplikasi Spotify pada tahun 2019 sampai sekarang. Informasi yang ingin didapatkan dari dataset ini adalah bagaimana karakteristik lagu yang populer dari bagian dunia tertentu, apakah ada karakteristik menonjol tertentu dari lagu yang

populer, dan apakah ada tren terhadap karakteristik yang berubah sepanjang periode waktu tertentu. Dataset ini berdimensi 1000x17 (1000 baris dan 17 kolom), ukuran filenya sekitar 26 kilobytes, dan tipe filenya adalah *Comma Separated Values* atau CSV. File dataset ini didapatkan dari situs kaggle.com (<https://www.kaggle.com/datasets/leonardopena/top-50-Spotify-songs-by-each-country>). Tim analis memilih data ini karena cocok untuk dianalisis berdasarkan jumlah atribut, kuantitas datanya, dan memenuhi kriteria serta statistik yang dibutuhkan.

B. Karakteristik Data

Data ini terdiri atas sembilan atribut (kolom) yaitu empat atribut kategorikal dan lima atribut numerikal atau kuantitatif. Atribut tersebut terdiri atas “title” (kategorikal - nominal), “artist” (kategorikal - nominal), “top genre” (kategorikal - nominal), “year” (kuantitatif - diskrit), “bpm” (kuantitatif - kontinu), “duration” (kuantitatif - kontinu), “acousticness” (kuantitatif - kontinu), “popularity” (kuantitatif - kontinu), “country” (kategorikal - nominal). Selain itu data ini terdiri dari 450 baris. Atribut “title” dan “artist” merupakan atribut kategorikal yang memuat judul lagu serta nama artisnya. Atribut “top genre” merupakan atribut kategorikal yang memuat genre atau jenis dari lagunya. Jenis lagu yang ada dalam dataset yang dianalisis adalah *latin*, *dance pop*, *pop*, *rock*, dll. Selain itu, terdapat atribut “country” yang merupakan atribut yang menunjukkan negara tempat lagu tersebut

populer, atribut ini terdiri dari 17 negara di antaranya India, Indonesia, Argentina, Australia. Pada data popularitas lagu di Spotify ini terdapat lima atribut kuantitatif yang masing-masing memiliki jangkauan data yang berbeda-beda. Atribut year / tahun rilis berkisar antara tahun 1942 – 2019, atribut bpm (*beats per minute*) berkisar antara 67 – 202 bpm, atribut duration / lama lagu (durasi) berkisar antara 117 sampai 328 detik, atribut acousticness berkisar antara 0 – 96, dan atribut popularity / popularitas lagu berkisar antara 69 – 100. Pada dataset ini ditemukan 23 nilai yang kosong atau sekitar 0,135 persen. Informasi karakteristik di atas diperoleh dengan menggunakan *library* Pandas dalam bahasa pemrograman Python.

C. Prapemrosesan dan Pembersihan Data

Dalam analisis data ada dikenal istilah prapemrosesan dan pembersihan data (*data cleansing*) yang biasa dilakukan sebelum analisis. *Data cleansing* atau pembersihan data bertujuan untuk meningkatkan kualitas data melalui identifikasi dan menghapus kesalahan (*error*) serta ketidakkonsistenan data [2]. Pada dataset kali ini, dilakukan beberapa langkah prapemrosesan dan pembersihan, meliputi:

1) Standardisasi nama kolom dan isi baris

Standardisasi nama kolom dan baris dimulai dengan mengganti nama-nama kolom yang kurang jelas. Prosesnya menggunakan bahasa pemrograman Python bisa dilihat pada potongan kode di bawah

```
df.rename(columns = {"dur" : "duration",
                    "acous": "acousticness",
                    "pop": "popularity"}, inplace=True)
```

Kode program tersebut mengubah nama kolom “dur”, “acous”, dan “pop” berturut-turut menjadi “duration”, “acousticness”, dan “popularity”. Hal ini dilakukan agar nama kolom lebih menggambarkan isi dari kolomnya. Singkatan “pop” dalam konteks lagu bisa saja bermakna popularitas atau genre lagu pop, maka dari itu perlu diganti menjadi “popularity” untuk mempertegas makna kolomnya. Kolom yang lain juga demikian. Selanjutnya untuk isi baris, dicek terlebih dahulu dengan fungsi `value_counts()` di setiap kolom apakah ada yang isinya aneh. Setelah penerapan fungsi, ditemukan bahwa dua kolom memiliki nilai yang mencurigakan. Kolom “title” memiliki data yang karakternya tidak terbaca oleh dekripsi yang diterapkan, sehingga memunculkan karakter seperti yang diperlihatkan pada gambar 1 berikut,

```

title
<U+05DC><U+05E9><U+05D5><U+05D1> <U+05D4><U+05...
<U+05D0><U+05DC><U+05D5><U+05E3> <U+05D4><U+05...
<U+05E0><U+05D7><U+05DB><U+05D4> <U+05DC><U+05DA>
<U+05DE><U+05E1><U+05E2>
<U+7A7A><U+306E><U+9752><U+3055><U+3092><U+77E...
<U+30C8><U+30EA><U+30B3>
<U+70B9><U+63CF><U+306E><U+5504>
<U+30EF><U+30BF><U+30EA><U+30C9><U+30EA>
<U+30CF><U+30EB><U+30CE><U+30D2>
<U+30CF><U+30C3><U+30D4><U+30FC><U+30A8><U+30F...
<U+6253><U+4E0A><U+82B1><U+706B>
<U+9AD8><U+5DBA><U+306E><U+82B1><U+5B50><U+305...
<U+541B><U+306F><U+30ED><U+30C3><U+30AF><U+309...
<U+305F><U+3060><U+541B><U+306B><U+6674><U+308C>
<U+30AF><U+30EA><U+30B9><U+30DE><U+30B9><U+30B...
<U+30CE><U+30FC><U+30C0><U+30A6><U+30C8>

```

Gambar 1. Judul-judul lagu yang tidak terdekripsi dengan baik

Setelah diperiksa lebih lanjut, judul-judul lagu yang tidak terdekripsi dengan baik adalah judul-judul lagu dengan karakter-karakter khusus, seperti judul lagu dengan huruf-huruf Jepang dan keseluruhan judul yang tidak terbaca ini adalah judul lagu yang populer di Israel dan Jepang. Secara intuitif, hal yang mudah dilakukan adalah menghapus semua lagu yang populer di Jepang dan Israel. Tetapi, sebelum melakukan hal tersebut, harus dipastikan bahwa penghapusan 2 kategori “country” tersebut tidak akan berdampak signifikan terhadap variasi dataset.

Setelah diterapkan kode berikut

```
df["country"].value_counts().size
```

Didapatkan bahwa kolom “country” memiliki 20 kategori berbeda, maka bisa disimpulkan bahwa penghapusan 2 kategori tidak akan terlalu

memengaruhi variasi kolom tersebut. Oleh karena itu, semua data lagu populer di Jepang dan Israel dihapus.

Kolom yang bermasalah selanjutnya adalah kolom “country”, salah satu negara dalam kolom tersebut tidak dituliskan dengan benar.

```

[ ] df["country"].value_counts()

india      25
france     25
world      25
bolivia    25
belgium    25
argentina  25
australia  25
africa     25
canada     25
usa        25
colombia   25
chile      25
spain      25
germany    25
italy      25
brazil     25
indonesia  25
malasya    25
Name: country, dtype: int64

```

Gambar 2. Kategori-kategori dari kolom “country”

Terlihat pada gambar 2 bahwa penulisan negara Malaysia tidak benar, yang seharusnya “Malaysia” dituliskan menjadi “malasya”. Maka entri-entri “malasya” harus diubah menjadi “malaysia” dengan cara

```
df.loc[df["country"]=="malasya", "country"] = "malaysia"
```

2) Pemeriksaan data duplikat dan nilai kosong (null)

Langkah selanjutnya adalah pemeriksaan data duplikat dan null. Pemeriksaan data duplikat dilakukan dengan menggunakan kode berikut

```
df[df.duplicated() ]
```

Sedangkan untuk pemeriksaan data kosong digunakan

```
df.loc[df.isnull().any(axis=1)]
```

Setelah digunakan kedua kode di atas, ditemukan bahwa tidak ada data duplikat, tetapi ada beberapa nilai kosong. Nilai kosongnya berada pada kolom “top genre” dan pada baris ke-750 (indeks 749). Untuk mengatasi nilai kosong tersebut dilakukan 2 hal:

- Penghapusan baris yang memiliki nilai kosong
- Pengisian nilai kosong (*imputing*) pada kolom “top genre”

Untuk nilai kosong pada baris dengan indeks 749 lebih masuk akal dilakukan penghapusan, karena data yang dimiliki sangat banyak (1000 baris) dan menghapus satu baris untuk negara India tidak akan terlalu memengaruhi analisis.

Sedangkan untuk nilai kosong pada kolom “top genre” lebih masuk akal dilakukan pengisian karena dataset hanya memiliki 2 kolom atribut kategorikal, sehingga penghapusan kolom “top genre” akan mempersulit analisis karena kekurangan atribut kategorikal.

Penanganan untuk baris data yang memiliki nilai kosong akan dilakukan bersamaan dengan

penghapusan baris dan atribut yang tidak relevan. Selanjutnya untuk penanganan kolom “top genre”, dilakukan *imputing* dengan cara

```
cat_imputer = CategoricalImputer()  
df["top genre"] = pd.Series(cat_imputer  
.fit_transform(df["top genre"]))
```

Secara sederhana, yang dilakukan oleh kode di atas adalah membuat sebuah “model” *imputer* yang akan memperkirakan nilai yang cocok untuk nilai-nilai kosong tersebut berdasarkan data pada baris-baris yang lain.

3) Penghapusan baris dan atribut yang tidak relevan

Data yang dimiliki sekarang terdiri dari 1000 baris dan 16 kolom, bisa dianalisis kolom-kolom dan baris-baris yang bisa dihapus berdasarkan:

- Untuk baris, bisa dihapus bila dirasa data yang dimiliki terlalu banyak dan analisisnya sendiri tidak memerlukan data sebanyak itu
- Untuk kolom kuantitatif, kolom dengan varians (*variance*) data yang rendah tidak menambahkan *insight* yang signifikan terhadap analisis.
- Untuk kolom kategorikal, kolom dengan kategori baris paling seragam tidak memberikan *insight* yang signifikan terhadap analisis.

Untuk baris, analisis merasa bahwa top 50 lagu untuk setiap negara terlalu banyak, diputuskan untuk diambil setengahnya saja, sehingga akan difilter hanya top 25 lagu tiap negara.

Selanjutnya untuk kolom, perhitungan variance akan digunakan method `var()` untuk pandas dataframe. Sedangkan keseragaman kategorinya akan dilihat dari `value_counts()`.

Penghapusan 25 lagu dengan popularitas terendah pada setiap negara dilakukan dengan penerapan kode berikut

```
df = df.groupby('country').head(25).sort_values("pop")
```

Filtrasi kolom-kolom kuantitatif berdasarkan *variance*-nya dilakukan dengan menerapkan

```
variance_array = []
df_kuantitatif = df.iloc[:, 5:14]
for column in df_kuantitatif.columns.tolist():
    variance_array.append(df_kuantitatif[column].var())

np.mean(variance_array)

index_kolom_gagal = []
for variance in variance_array:
    if variance < np.mean(variance_array):
        index_kolom_gagal.append(variance_array.index(variance))

nama_kolom_gagal = []
for i in range(len(df_kuantitatif.columns.tolist())):
    if i in index_kolom_gagal:
        nama_kolom_gagal.append(df_kuantitatif.columns.tolist()[i])

df = df.drop(nama_kolom_gagal, axis=1)
```

dan untuk kolom-kolom kategorikal, dilakukan filtrasi berdasarkan keseragaman dengan menerapkan

```
df.nunique()
```

Dari kode di atas, ditemukan bahwa kolom “added” hanya memiliki 1 nilai yang sama untuk semua baris. Hal ini tentu tidak akan menambahkan informasi yang penting pada analisis, maka bisa dihapus saja kolomnya.

Tahap-tahap prapemrosesan dan pembersihan data yang dipaparkan di atas hanyalah bagian-bagian utama dan penting yang dilakukan oleh analis, untuk penjelasan dan *script* program lengkapnya akan dilampirkan.

Setelah dilakukan prapemrosesan dan pembersihan data, dataset yang dimiliki sekarang berdimensi 450x9 (450 baris dengan 9 kolom).

III. ANALISIS DAN VISUALISASI

A. Statistik Umum

Statistik data yang diperoleh dari dataset tersebut dicari menggunakan ‘method’ *describe* dari modul pandas dalam python dengan bantuan perangkat lunak code editor *visual studio code*. Berdasarkan pengolahan data dengan program, berhasil diperoleh untuk masing – masing atribut numerikal nilai rata-rata (mean), standard deviasi (std), persentil (10%, 25%, 50%, 75%, 90%), serta nilai ekstremum (nilai maksimum dan minimum). Untuk mencari persentil 10 % digunakan code:

```
for column in df1.columns.tolist() :  
  
    print(df1[column].quantile(0.1))
```

dengan df1 merupakan data frame yang terdiri dari atribut numerikal. Untuk mencari persentil 90 % digunakan code:

```
for column in df1.columns.tolist() :  
  
    print(df1[column].quantile(0.9))
```

dengan df1 merupakan data frame yang terdiri dari atribut numerikal.

Untuk atribut pertama yaitu atribut *year* yang memuat informasi-informasi dari tahun perilisan lagu. Dari pengolahan data didapat bahwa rata-rata lagu yang populer pada tahun 2019 adalah lagu yang dirilis pada tahun 2008. Pesebaran data pada suatu sampel dan kedekatan data-data tersebut dengan nilai mean dapat diukur dengan nilai standar deviasi yang di dapat yaitu sekitar 18.787. Artinya nilai mean dan standar deviasi sangatlah jauh. Hal ini dapat disebabkan ada data yang memiliki nilai sangat jauh dari rata-rata. Lagu yang paling lawas dirilis yang masih sering di dengar para pengguna sportify di tahun 2019 adalah lagu dari tahun 1942 dan untuk yang terbaru yaitu lagu yang diriilis tahun 2019. Untuk lebih jelasnya kita dapat menggunakan persebaran data tahun rilis melalui persentil setelah data diurutkan. Persentil 10 % yaitu 1980, Persentil 25 % yaitu 2010, Persentil 50 % yaitu 2019, Persentil 75 % yaitu 2019, Persentil 90 % yaitu 2019. Berdasarkan data persentil didapatkan bahwa lagu yang populer pada saat tersebut merupakan lagu yang rilis pada tahun 2019.

Selanjutnya, yaitu atribut bpm. Atribut ini memuat informasi mengenai *beat per minute* dari masing-masing lagu yang masuk ke top song di Spotify. Dari pengolahan data didapat bahwa lagu-lagu pada tahun 2019 memiliki rata-rata bpm sekitar 124 bpm. Hal tersebut menginformasikan bahwa lagu-lagu populer pada tahun tersebut memiliki tempo yang cukup cepat dan termasuk ke tempo Allegro. Pesebaran data pada suatu sampel dan kedekatan data-data tersebut dengan nilai mean dapat diukur dengan nilai standar deviasi yang di dapat yaitu sekitar 32.8. Nilai mean dan standar deviasi sangatlah jauh karena ada data yang memiliki nilai sangat jauh dari rata-rata. Lagu dengan bpm terendah pada list tersebut yaitu 67 bpm di mana lagu tersebut memiliki tempo cenderung lambat dan santai yang termasuk ke kategori Adegio. Lagu dengan bpm tercepat merupakan lagu dengan tempo prestissimo yang merupakan lagu dengan tempo sangat cepat. Untuk mengetahui persebaran data dengan lebih detail dapat digunakan persentil melalui data yang sudah diurutkan. Persentil 10 % yaitu 91 bpm, Persentil 25 % yaitu 96 bpm, Persentil 50 % yaitu 120 bpm, Persentil 75 % yaitu 150 bpm, Persentil 90 % yaitu 176 bpm.

Selanjutnya, yaitu atribut *duration*. Atribut ini memuat informasi mengenai durasi / lama dari masing-masing lagu yang masuk ke top song di Spotify. Dari pengolahan data didapat bahwa lagu-lagu yang populer pada tahun 2019 memiliki rata-rata durasi sekitar 197 detik atau sekitar tiga menit. Orang-orang pada tahun 2019 lebih senang mendengarkan lagu-lagu yang relative singkat

dan tidak terlalu lama. Pesebaran data pada suatu sampel dan kedekatan data-data tersebut dengan nilai mean dapat diukur dengan nilai standar deviasi yang di dapat yaitu sekitar 41.6. Data tersebar secara tidak merata dapat dilihat dari jauhnya std dengan mean. Lagu dengan durasi tersingkat yaitu 117 detik atau sekitar 2 menit sedangkan untuk durasi terpanjang yaitu 326 detik atau sekitar 5 menit. Lagu dengan durasi terpanjang tersebut kebanyakan diselengi instrument yang membuat durasi lebih lama. Untuk mengetahui persebaran data dengan lebih detail dapat digunakan persentil melalui data yang sudah diurutkan. Persentil 10 % yaitu 149 detik , Persentil 25 % yaitu 166 detik, Persentil 50 % yaitu 196 detik, Persentil 75 % yaitu 222 detik, Persentil 90 % yaitu 250 .

Selanjutnya, yaitu atribut *acousticness*. Atribut ini memuat informasi mengenai keakustikan lagu dari lagu-lagu yang masuk ke top song di Spotify. *Acousticness* merupakan nilai yang menggambarkan keakustikan sebuah lagu dengan nilai dari nol sampai 100. Dari pengolahan data didapat bahwa lagu-lagu yang populer pada tahun 2019 memiliki rata-rata akustik sekitar 34 dan memiliki standard deviasi sekitar 30. Data tersebar secara hampir merata yang dibuktikan dengan nilai std dan mean yang cukup dekat. Lagu dengan keakustikan terendah memiliki nilai yaitu 0 atau tidak akustik sama sekali sedangkan untuk lagu yang paling akustik memiliki nilai keakustikan 96 atau hampir seluruh lagu nya merupakan lagu akustik. Untuk mengetahui persebaran data dengan lebih detail dapat

digunakan persentil melalui data yang sudah diurutkan. Persentil 10 % yaitu 3, Persentil 25 % yaitu 7, Persentil 50 % yaitu 22, Persentil 75 % yaitu 61, Persentil 90 % yaitu 84. Orang-orang pada saat itu, kebanyakan lebih senang jika tidak seratus persen lagunya akustik dan diselengi vokal atau instrumental lain selain akustik

Terakhir atribut *popularity*. Atribut ini merupakan atribut yang meranking kepopuleran lagu berdasarkan seberapa sering lagu diputar di Spotify. Atribut ranking ini telah kami sortir sehingga hanya terdapat lagu dari ranking 69 – 100 saja yang masuk ke data. Ranking dari data yang telah kami sortir memiliki rata-rata 90 dan standar deviasi 6. Oleh karena nilai mean dan standar deviasi sangat jauh hal tersebut menandakan bahwa data tidak tersebar secara merata. Untuk lagu dengan tingkat kepopuleran tertinggi yang berada pada data yang telah kami sortir yaitu 60 dan terendah 100. Dari data tersebut, didapatkan pula nilai-nilai persentilnya. Persentil 10 % yaitu 83, Persentil 25 % yaitu 88, Persentil 50 % yaitu 91, Persentil 75 % yaitu 95, Persentil 90 % yaitu 98.

B. Korelasi

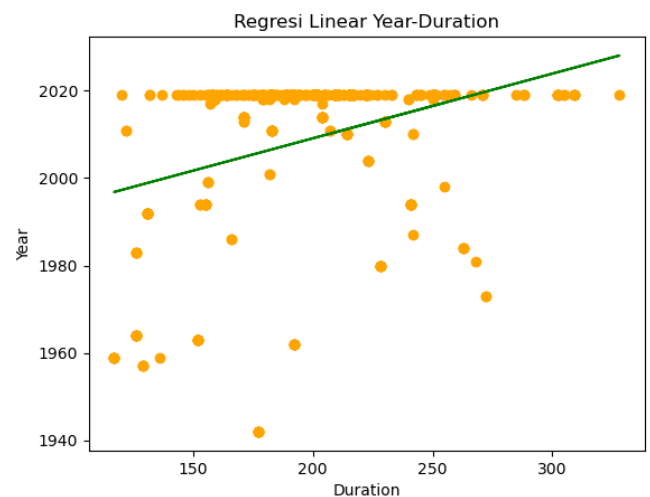
Korelasi adalah ukuran statistik yang menunjukkan sejauh mana dua atau lebih variabel berfluktuasi dalam hubungannya satu sama lain. Korelasi positif menunjukkan sejauh mana variabel tersebut meningkat atau menurun secara paralel; korelasi negatif menunjukkan sejauh mana satu variabel meningkat saat yang lain menurun [3]. Dalam analisis kali ini, analisis

mencoba memetakan korelasi antar atribut (kolom) yang ada pada dataset. Hasil pemetaan tersebut disajikan dalam nilai-nilai dan grafik di bawah.

Tabel 1. Nilai-nilai korelasi antar atribut dalam dataset

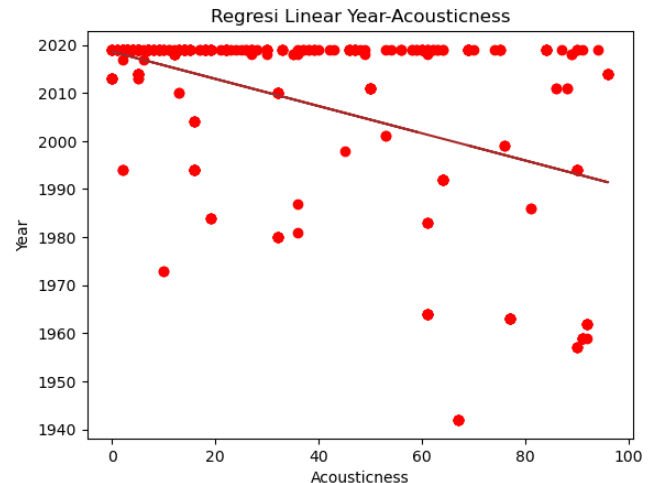
Atribut yang dipetakan	Nilai korelasi
BPM-Duration	0,04601
Acousticness-Popularity	-0,08307
Year-Popularity	0,02360
Acousticness-BPM	-0,15310
Year-Acousticness	-0,46052
BPM-Popularity	-0,02665
Acousticness-Duration	-0,33181
Year-BPM	-0,03719
Year-Duration	0,32811
Duration-Popularity	-0,01970

Dari data nilai-nilai korelasi di tabel 1, terlihat bahwa kebanyakan atribut tidak berkorelasi (nilai mutlak dari nilai korelasinya tidak lebih besar atau sama dengan 0,5). Atribut yang berkorelasi paling positif adalah atribut “year” dengan “duration” dengan nilai korelasi 0,32811. Sedangkan atribut yang berkorelasi paling negatif adalah atribut “year” dengan “acousticness” dengan nilai korelasi -0,46052. Selanjutnya atribut yang korelasinya hampir tidak ada (nilai korelasi mendekati 0) adalah atribut “duration” dengan “popularity” dengan nilai korelasi -0,01970. Guna mempersingkat pembahasan, korelasi yang akan dibahas dengan grafik adalah korelasi maksimum, minimum, dan nol.



Gambar 3. Diagram pencar year terhadap duration dengan *trendline*

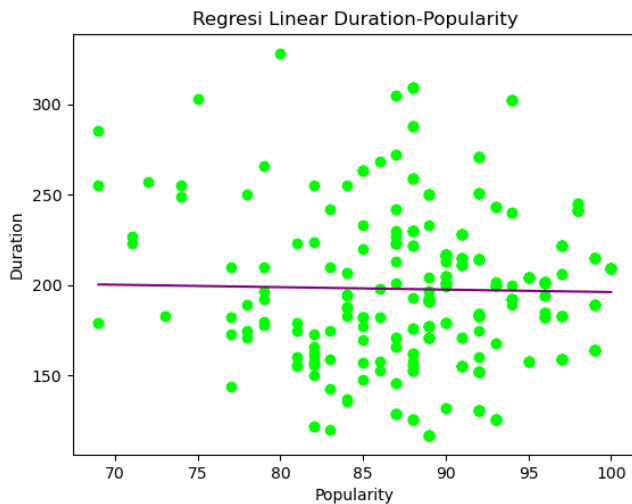
Dari grafik di atas, terlihat bahwa kemiringan garis tren semakin ke kanan semakin meningkat. Dari informasi tersebut bisa disimpulkan bahwa semakin panjang durasi suatu lagu, semakin baru dirilis lagunya.



Gambar 4. Diagram pencar year terhadap acousticness dengan *trendline*

Gambar 4 di atas menunjukkan garis tren yang cukup curam dari kiri atas ke kanan bawah. Hal ini menandakan korelasi yang negatif antara kedua atribut yang dipetakan. Maka bisa disimpulkan bahwa lagu dengan tingkat

keakustikan yang tinggi, cenderung merupakan lagu yang sudah lama dirilis.



Gambar 5. Diagram pencar duration terhadap popularity dengan *trendline*

Terakhir ada gambar 5, gambar 5 tersebut menampilkan grafik pencar dari atribut “duration” dan “popularity” bersama garis tren-nya. Dari grafik tersebut bisa diperoleh kesimpulan bahwa nilai atribut “popularity” tidak dipengaruhi oleh nilai atribut “duration” dan sebaliknya.

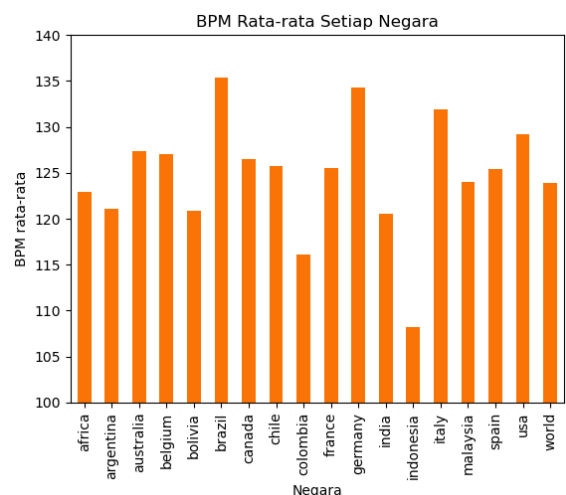
C. Visualisasi Data

Visualisasi data adalah praktik menerjemahkan informasi ke dalam konteks visual, seperti peta atau grafik, untuk membuat data lebih mudah dipahami oleh otak manusia dan menarik wawasan. Tujuan utama visualisasi data adalah untuk mempermudah mengidentifikasi pola, tren, dan outlier dalam kumpulan data besar. Istilah ini sering digunakan secara bergantian dengan yang lain, termasuk grafik informasi, visualisasi informasi, dan grafik statistik [4].

Visualisasi data itu penting karena mempermudah analisis dan interpretasi dari data yang besar dan kompleks. Visualisasi data dapat mengomunikasikan ide-ide kompleks dengan jelas, akurat, dan efisien [5].

Selanjutnya akan diperlihatkan beberapa visualisasi informasi dari dataset yang telah dikerjakan. Visualisasi yang dibuat telah menggambarkan:

- Perbandingan kategori dari beberapa atribut (diagram batang)
- Perubahan data terhadap waktu (diagram garis)
- Hubungan keseluruhan-bagian (diagram pai)
- Relasi antar atribut (diagram pencar)

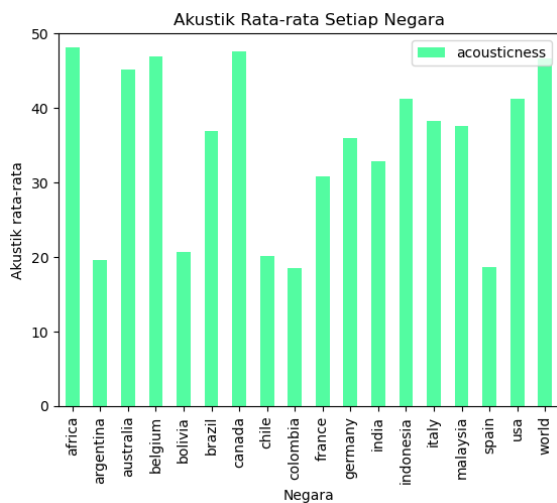


Gambar 6. Diagram batang rata-rata BPM lagu dari setiap kategori “country”

BPM, secara sederhana, menunjukkan seberapa cepat sebuah lagu dimainkan. Diagram

batang di atas menunjukkan beats per minute (BPM) rata-rata dari 25 lagu terpopuler 17 negara berbeda dan dunia secara keseluruhan. Poin-poin penting yang diperoleh dari visualisasi di atas, di antaranya:

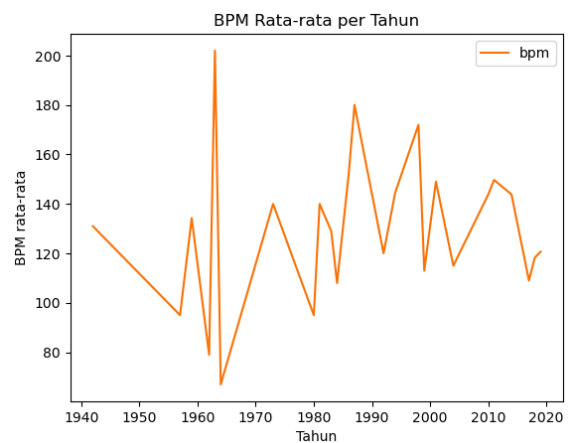
- Sebagian besar dari negara barat memiliki BPM rata-rata yang tinggi.
- Indonesia memiliki BPM rata-rata yang terendah
- Brazil memiliki BPM rata-rata yang tertinggi.



Gambar 7. Diagram batang rata-rata tingkat keakustikan lagu dari setiap kategori pada “country”

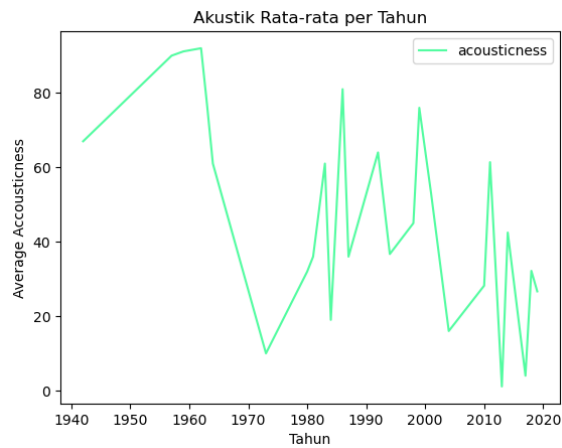
Akustik menunjukkan seberapa jauh sebuah lagu dipengaruhi oleh alat musik atau suara elektronik, semakin tinggi tingkat akustik suatu lagu, semakin bebas lagu tersebut dari suara elektronik. Diagram batang di atas menunjukkan tingkat akustik rata-rata dari 25 lagu terpopuler 17 negara berbeda dan dunia secara keseluruhan. Diagram di atas menunjukkan bahwa sebagian

besar dari negara pada data memiliki tingkat akustik yang mendekati dan di atas rata-rata, yaitu 34,8 dengan beberapa negara yang memiliki nilai rata-rata keakustikan rendah yaitu Argentina, Bolivia, Chili, Kolombia, dan Spanyol. Negara dengan akustik rata-rata terendah adalah Kolombia, dan negara dengan akustik rata-rata tertinggi adalah Afrika.



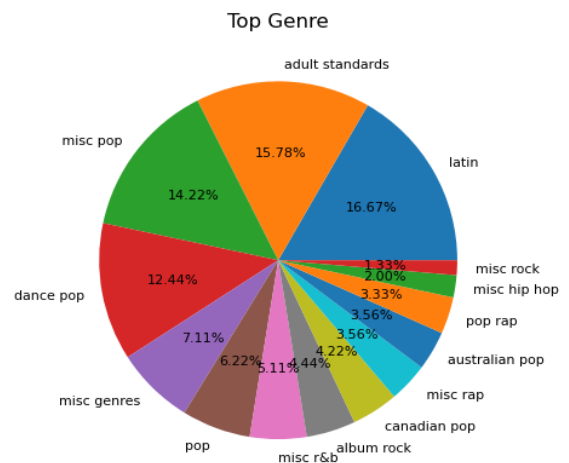
Gambar 8. Diagram garis perkembangan rata-rata BPM dari lagu populer dari tahun 1940-2020

Diagram garis di atas menunjukkan perkembangan BPM terhadap tahun dari 1940 sampai 2020. Diagram menunjukkan bahwa BPM dari tahun 1940 sampai 2020 tidak memiliki perkembangan yang stabil dengan sering adanya perubahan ekstrem tetapi secara-rata berkembang mendekati rentang 120 sampai 140 BPM.

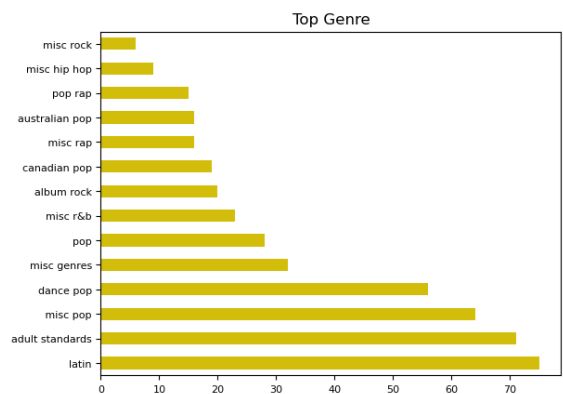


Gambar 9. Perkembangan rata-rata keakustikan dari lagu populer dari tahun 1940-2020

Diagram garis di atas menunjukkan perkembangan akustik terhadap tahun dari 1940 sampai 2020. Diagram di atas menunjukkan bahwa akustik dari tahun 1940 sampai 2020 menurun dengan tidak stabil. Informasi yang bisa diperoleh dari visualisasi tersebut adalah bahwa perkembangan akustik mengalami naik dan turun secara drastik dari tahun ke tahun tetapi mengalami tren yang terus turun seiring lajunya zaman. Ini bisa terhubung dengan perkembangan teknologi yang membawa berbagai teknologi elektronik baru untuk musik dan mendorong orang untuk menggunakan teknologi baru tersebut.

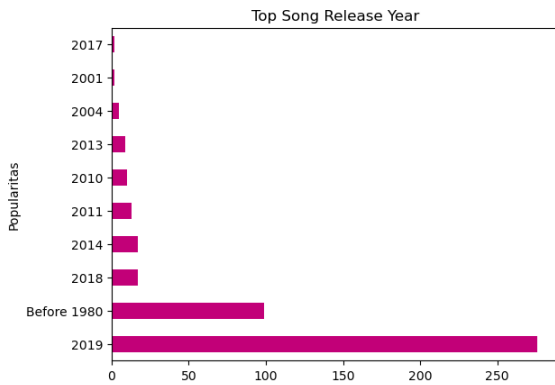


Gambar 10. Diagram pai dari genre lagu-lagu populer

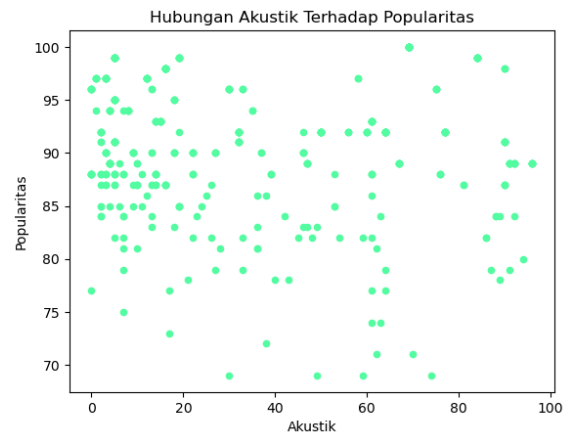


Gambar 11. Diagram batang horizontal dari genre lagu-lagu populer

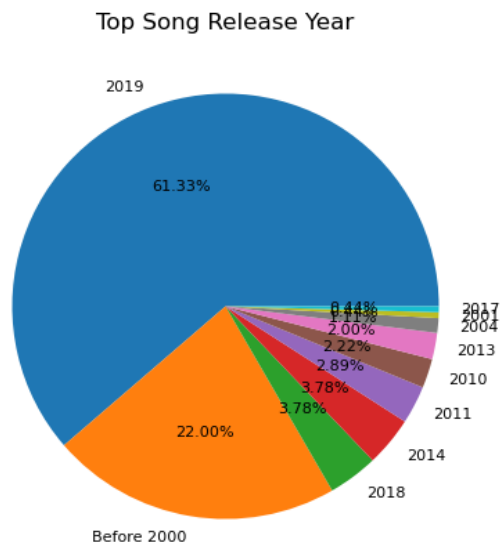
Kedua diagram di atas menunjukkan pembagian genre dari lagu terpopuler. Berdasarkan diagram ini, jika semua jenis genre pop digabung, maka genre yang terpopuler adalah pop. Jika genre pop dipisahkan maka genre yang terpopuler adalah latin.



Gambar 12. Diagram batang horizontal dari tahun rilis lagu populer

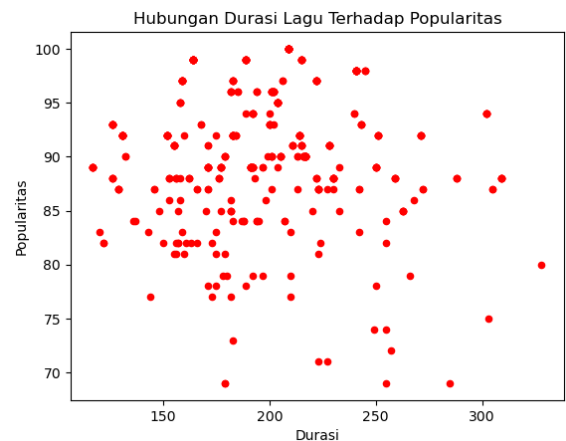


Gambar 14. Diagram pencar dari popularitas terhadap keakustikan



Gambar 13. Diagram pai dari tahun rilis lagu populer

Kedua diagram di atas menunjukkan tahun persentase dari tahun rilis lagu terpopuler. Berdasarkan kedua diagram tersebut, dapat dilihat bahwa sebagian besar dari lagu terpopuler berasal dari tahun 2019. Dapat disimpulkan bahwa sebagian besar dari lagu terpopuler merupakan lagu-lagu yang baru dirilis.



Gambar 15. Diagram pencar dari durasi terhadap popularitas

Diagram di atas menunjukkan hubungan dari durasi sebuah lagu terhadap popularitasnya. Dapat dilihat dari diagram dan dari koefisien korelasi bahwa tidak ada hubungan antara kedua faktor tersebut.

Dari kedua diagram pencar yang telah ditampilkan, dapat diperoleh informasi bahwa popularitas sebuah lagu pada musim Natal 2019 tidak bergantung pada keakustikan ataupun durasi dari lagunya.

D. Kesimpulan

Setelah melakukan serangkaian proses analisis dan visualisasi data, tim analis berhasil mencapai kesimpulan yang menjawab tujuan-tujuan dan informasi yang diharapkan dari analisis ini, yaitu:

1. Genre lagu yang paling populer di Spotify pada musim Natal 2019 adalah lagu dengan genre latin, adult standards, dan misc pop.
2. Lagu yang paling populer di Spotify pada musim Natal 2019 adalah lagu-lagu yang rilis tahun 2019.
3. Lagu yang populer di Indonesia pada saat itu adalah lagu-lagu dengan BPM yang rendah.
4. Lagu yang populer di Afrika pada saat itu memiliki tingkat keakustikan yang tinggi.
5. Lagu populer pada musim Natal tahun 2019 yang rilis tahun 1960 memiliki tingkat keakustikan yang paling tinggi.
6. Secara garis besar, tingkat keakustikan dan durasi lagu tidak berpengaruh signifikan pada popularitas lagu di mayoritas negara.





Dari poin-poin di atas bisa ditarik satu kesimpulan umum di mana pada musim Natal 2019 lagu yang paling populer adalah lagu-lagu bergenre latin dan dirilis tahun 2019. Informasi ini bisa digunakan oleh para komposer, penyanyi-penyanyi, dan pemegang kepentingan lain di industri musik untuk mempertimbangkan jenis musik yang akan mereka produksi untuk musim Natal di tahun-tahun yang akan datang.

DAFTAR REFERENSI

- [1] Kemendikbudristek. (2022). KBBI Daring. [Online]. Tersedia: kbbi.kemdikbud.go.id
- [2] Rahm, Erhard and Hong Hai Do. (2000). "Data Cleaning: Problems and Current Approaches." IEEE Bulletin of the Technical Committee on Data Engineering (23): 3-13.
- [3] Wigmore, Ivy. (2020). Correlation. [Online]. Tersedia: <https://www.techtarget.com/whatis/definition/correlation>
- [4] Brush, Kate dan Ed Burns. (2020). Data visualization. [Online]. Tersedia: <https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization>
- [5] Sadiku, Matthew *et. al.*. (2016). "Data Visualization". International Journal of Engineering Research and Advanced Technology (IJERAT) (12): 2454-6135.

LAMPIRAN

Profil Tim Analis

	
Nama : Shafar Zidan Nugraha	Nama : Dama Dhananjaya Daliman
NIM : 16522092	NIM : 19622002
Email : 16522092@mahasiswa.itb.ac.id	Email : 19622002@mahasiswa.itb.ac.id
	
Nama : Marasi Joel Silvano	Nama : Hartanto Luwis
NIM : 16522212	NIM : 19622032
Email : 16522212@mahasiswa.itb.ac.id	Email : 19622032@mahasiswa.itb.ac.id

Pembagian Tugas Tim Analis

Nama Anggota Tim	Tugas yang diemban
Shafar Zidan Nugraha (16522092)	<ul style="list-style-type: none">- Pengamatan karakteristik data- Pemetaan korelasi antar atribut dalam dataset
Marasi Joel Silvano (16522212)	<ul style="list-style-type: none">- Pembuatan visualisasi data
Dama Dhananjaya Daliman (19622002)	<ul style="list-style-type: none">- Pembersihan dan prapemrosesan data- Finishing dan QA Laporan- Pembuatan slides presentasi
Hartanto Luwis (19622032)	<ul style="list-style-type: none">- Pengamatan informasi deskriptif dari dataset dan filenya- Pembuatan kalkulasi untuk statistik umum

Penyimpanan daring dari notebook (ipynb) yang berisi kode program yang digunakan untuk mengolah dataset dalam analisis ini tersedia di:

[github.com/RunningPie/my-py-](https://github.com/RunningPie/my-py-projects/tree/main/College%20Projects/Data%20Related/Top%2025%20Spotify%20Songs%20on%20Christmas%202019)

[projects/tree/main/College%20Projects/Data%20Related/Top%2025%20Spotify%20Songs%20on%20Christmas%202019](https://github.com/RunningPie/my-py-projects/tree/main/College%20Projects/Data%20Related/Top%2025%20Spotify%20Songs%20on%20Christmas%202019)