Analisis Data Lagu Top 2000 Spotify Menggunakan Python Melalui Modul Pandas dan Matplotlib

Marchotridyo
STEI
Institut Teknologi Bandung
Bandung, Indonesia
acoxstpd@gmail.com

Willy Wilsen
STEI
Institut Teknologi Bandung
Belitung Timur, Indonesia
willywilsen.ww@gmail.com

Divya Maharani Lazuardi
STEI
Institut Teknologi Bandung
Jakarta Timur, Indonesia
divyamaharani21@gmail.com

Wervyan Shalannanda, S.T., M.T.
KK Teknik Telekomunikasi, STEI
Institut Teknologi Bandung
Bandung, Indonesia
wervyan@office.itb.ac.id

***Abstract*—Analisis data adalah kegiatan untuk mengubah data menjadi suatu informasi yang dapat digunakan untuk berbagai hal, seperti mengambil kesimpulan. Data yang akan dianalisis adalah data lagu top 2000 di Spotify yang berisikan judul, tahun, genre, dan karakteristik dari masing-masing lagu. Hasil analisis ditujukan untuk menghasilkan sesuatu yang bermanfaat baik bagi penulis sendiri maupun bagi pembaca. Analisis dilakukan menggunakan bahasa pemrograman Python dengan bantuan modul Pandas dan Matplotlib.**

Keywords—Analisis data, data lagu top 2000 Spotify, Python, Pandas, Matplotlib

I. Introduction

Analisis data adalah proses memproses data untuk menghasilkan kesimpulan-kesimpulan yang dapat dipakai dalam banyak bidang. Tujuan dari dilakukannya analisis data adalah untuk menciptakan suatu informasi yang sulit untuk dideduksi tetapi ketika dipahami dapat menghasilkan informasi-informasi berharga (Nelli, 2018: 1).

Penulis memilih dataset Spotify karena berhubungan dengan perkembangan teknologi yang mungkin diminati oleh fakultas STEI. Selain itu, penulis beropini bahwa data-datanya menarik dan dapat membuka pengetahuan baru bagi penulis maupun bagi pembaca. Dataset penulis ambil dari kaggle.com, karya Sumat Singh yang terakhir diupdate 10 bulan yang lalu dari pembuatan makalah ini (Desember 2020).

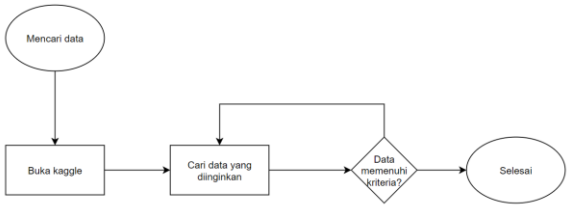
II. System Overview

Dekomposisi sistem merupakan salah satu hal yang paling penting dalam perancangan suatu projek. Dalam hal ini, penulis membagi dekomposisi langkah-langkah analisis data menjadi beberapa tahap berikut:

1. Mencari data yang sesuai
2. Mendeskripsikan data
3. Mendefinisikan karakteristik data
4. Mendesripsikan statistik data
5. Memvisualisasikan data
6. Mencari korelasi (hubungan) antardata
7. Melakukan pembersihan data

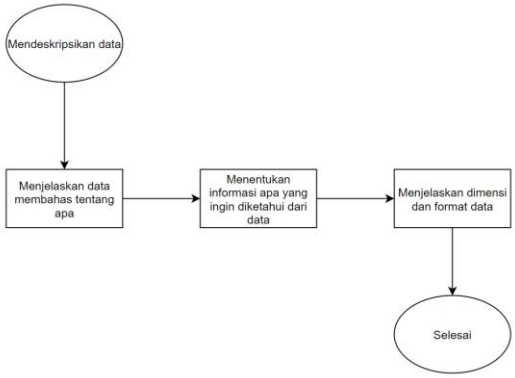
III. Proposed System

Secara umum, penulis melakukan langkah-langkah yang telah dijelaskan pada bagian II seperti yang akan dijelaskan oleh *flowchart*.



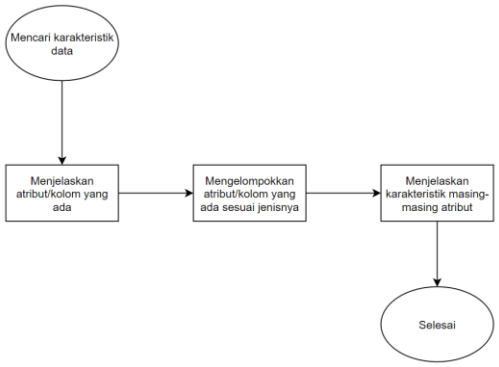
Dalam pencarian data, penulis menggunakan situs kaggle. Data yang dicari harus memenuhi kriteria berikut:

1. Minimum terdiri atas 5 atribut dan minimum terdiri atas 60 baris.
2. Mengandung atribut kategorikal dan atribut kuantitatif.
3. Memiliki atribut yang merupakan data waktu.



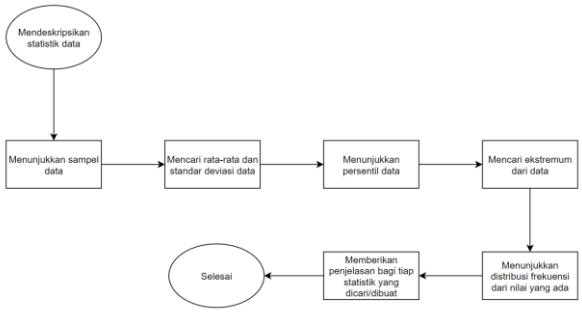
Dalam mendeskripsikan data, yang perlu dicari adalah hal-hal berikut:

1. Apa yang dibahas di dalam data?
2. Apa yang ingin digali/dicari dari data?
3. Bagaimana format dan dimensi data?



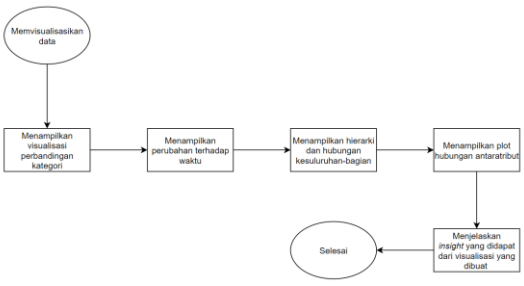
Dalam mendefinisikan karakteristik data, yang perlu dilakukan sebagai berikut:

1. Menjelaskan atribut/kolom yang ada
2. Mengelompokkan atribut/kolom yang ada sesuai dengan jenisnya
3. Menjelaskan karakteristik masing-masing atribut seperti *range*, persen data yang kosong, dsb.



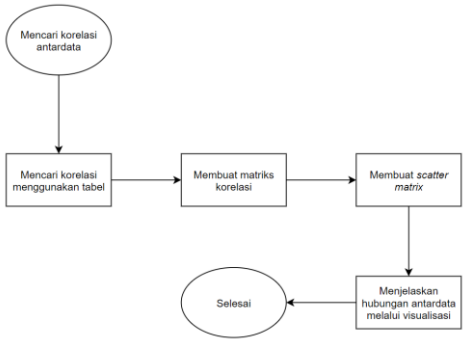
Dalam mendeskripsikan statistik data, yang perlu dilakukan sebagai berikut:

1. Menunjukkan sampel data
2. Mencari rata-rata dan standar deviasi data
3. Menunjukkan persentil data
4. Mencari ekstremum dari data
5. Menunjukkan distribusi frekuensi dari nilai yang ada
6. Memberikan penjelasan bagi tiap statistik yang dibuat



Dalam memvisualisasikan data, yang perlu dibuat sebagai berikut:

1. Grafik perbandingan kategori
2. Grafik perubahan terhadap waktu
3. Grafik hierarki dan hubungan keseluruhan-bagian
4. Grafik hubungan antaratribut (*scatter plot*)
5. Analisis terhadap masing-masing grafik.



Dalam mencari korelasi (hubungan) antardata, yang perlu dibuat sebagai berikut:

1. Mencari korelasi menggunakan tabel
2. Membuat matriks korelasi
3. Membuat *scatter matrix*
4. Mendeskripsikan hubungan antardata dari visualisasi yang telah dibuat

IV. Results and Discussion

Di bagian ini, penulis menyimpulkan apa saja yang telah penulis dapatkan dari proses analisis data berikut. Yang menjadi catatan penting adalah, yang diberikan di sini hanyalah rangkuman dari analisis data kami. Analisis data lengkapnya dapat dilihat dalam *file* .ipynb yang telah dicantumkan bersama laporan ini.

Berikut adalah deskripsi data yang telah penulis buat untuk mendeskripsikan data.

Deskripsi Data

Dataset ini berisi statistik audio dari 2000 trek teratas di Spotify. Data berisi sekitar 15 kolom yang masing-masing menjelaskan trek dan kualitasnya. Lagu-lagu yang dirilis dari 1956 hingga 2019 termasuk dari beberapa artis terkenal dan terkenal seperti Queen, The Beatles, Guns N' Roses, dll.

Cara kerjanya adalah dengan memanfaatkan fitur Spotify API. Spotify API akan mengekstrak fitur audio dari trek yang diberikan oleh Spotify Playlist URI. Data ini berisi fitur audio seperti Danceability, BPM, Liveness, Valence(Positivity) dan masih banyak lagi.

Dari dataset ini kita dapat mengetahui beberapa hal:

1. Genre yang lebih populer dari kurun waktu 1950-an sampai 2000-an
2. Lagu-lagu dari genre mana yang cenderung berada pada posisi atas pada tahun 2000-an
3. Artis mana yang lebih cenderung membuat lagu papan atas
4. Lirik lagu seperti apa yang lebih populer
5. Perbandingan tempo rata-rata tiap tahunnya
6. Ada atau tidaknya tren genre lagu, baik pada zaman dulu dan sekarang

Data memiliki 1994 baris dengan 15 kolom seperti yang ditunjukkan oleh data berikut.

1. Dimensi

```
In [16]: df1.shape
Out[16]: (1994, 15)
```

Ukuran data sekitar 233,8 kB dengan kolum-kolum berikut.

2. Ukuran File Data

sebesar 233,8+ KB

```
In [15]: df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1994 entries, 0 to 1993
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Index                   1994 non-null   int64
 1   Title                   1994 non-null   object
 2   Artist                  1994 non-null   object
 3   Top Genre               1994 non-null   object
 4   Year                    1994 non-null   int64
 5   Beats Per Minute (BPM)  1994 non-null   int64
 6   Energy                  1994 non-null   int64
 7   Danceability            1994 non-null   int64
 8   Loudness (dB)           1994 non-null   int64
 9   Liveness                1994 non-null   int64
 10  Valence                 1994 non-null   int64
 11  Length (Duration)       1994 non-null   object
 12  Acousticness            1994 non-null   int64
 13  Speechiness             1994 non-null   int64
 14  Popularity              1994 non-null   int64
dtypes: int64(11), object(4)
memory usage: 233.8+ KB
```

Deksripsi dari masing-masing kolom ditunjukkan oleh analisis berikut.

Index: ID (Kuantitatif)

Title: Judul lagu (kategorikal, nominal)

Artist: Nama penyanyi (kategorikal, nominal)

Top Genre: Genre musik (kategorikal, nominal)

Year: Tahun perilisan musik (kuantitatif)

Beats per Minute(BPM): Tempo dari lagu tersebut (kuantitatif)

Energy: Tingkat keenergetikan lagu (kuantitatif)

Danceability: Tingkat kecenderungan untuk menari saat mendengar lagu (kuantitatif)

Loudness: Tingkat kekerasan (audio) musik (kuantitatif)

Valence: Tingkat kepositifan suasana lagu (kuantitatif)

Length: Durasi dari lagu (kuantitatif)

Acoustic: Tingkat keakustikan sebuah lagu (kuantatif)

Speechiness: Tingkat banyaknya kata dalam lagu(kuantitatif)

Popularity: Tingkat popularitas lagu (kuantitatif)

Di data yang penulis gunakan, tidak ada data yang kosong.

Persen Data Kosong

```
In [26]: df1.isna().sum()
```

```
Out[26]: Index                   0
         Title                   0
         Artist                  0
         Top Genre               0
         Year                    0
         Beats Per Minute (BPM)  0
         Energy                  0
         Danceability            0
         Loudness (dB)           0
         Liveness                0
         Valence                 0
         Length (Duration)       0
         Acousticness            0
         Speechiness             0
         Popularity              0
         dtype: int64
```

--> berarti persen data kosong untuk semua data adalah 0%

Secara statistik, rata-rata dan standar deviasi dari masing-masing atribut ditunjukkan oleh tabel berikut.

Statistik rata-rata dan standar deviasi

```
In [22]: # Statistik rata-rata data dan standar deviasi.
         print('Rata-rata:')
         display(df[['Beats Per Minute (BPM)', 'Energy',
            'Danceability', 'Loudness (dB)', 'Liveness', 'Valence', 'Length (Duration)',
            'Acousticness', 'Speechiness', 'Popularity']].mean())
         print('Standar deviasi:')
         display(df[['Beats Per Minute (BPM)', 'Energy',
            'Danceability', 'Loudness (dB)', 'Liveness', 'Valence', 'Length (Duration)',
            'Acousticness', 'Speechiness', 'Popularity']].std())
```

```
Rata-rata:

Beats Per Minute (BPM)    120.215647
Energy                     59.679539
Danceability               53.238215
Loudness (dB)              -9.008526
Liveness                   19.012036
Valence                    49.408726
Length (Duration)         262.443330
Acousticness               28.858074
Speechiness                 4.994985
Popularity                 59.526580
dtype: float64

Standar deviasi:

Beats Per Minute (BPM)    28.028096
Energy                    22.154322
Danceability              15.351507
Loudness (dB)              3.647876
Liveness                  16.727378
Valence                   24.858212
Length (Duration)         93.604387
Acousticness              29.011986
Speechiness                4.401566
Popularity                14.351600
dtype: float64
```

Selanjutnya, ada data nilai-nilai persentil tertentu dari data.

Statistik percentile

```
In [28]: # Statistik percentile (10%, 25%, 50%, 75%, dan 90%).
         print('Persentil:')
         display(df[['Beats Per Minute (BPM)', 'Energy',
            'Danceability', 'Loudness (dB)', 'Liveness', 'Valence', 'Length (Duration)',
            'Acousticness', 'Speechiness', 'Popularity']].quantile([0.10, 0.25, 0.50, 0.75, 0.90]))
```

Persentil:

	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Length (Duration)	Acousticness	Speechiness	Popularity
0.10	84.0	29.0	32.0	-14.0	7.0	18.00	181.3	0.0	3.0	40.00
0.25	99.0	42.0	43.0	-11.0	9.0	29.00	212.0	3.0	3.0	49.25
0.50	119.0	61.0	53.0	-8.0	12.0	47.00	245.0	18.0	4.0	62.00
0.75	136.0	78.0	64.0	-6.0	23.0	69.75	289.0	50.0	5.0	71.00
0.90	162.0	89.0	73.0	-5.0	37.0	86.00	349.7	75.0	8.0	76.00

Dari masing-masing atribut, dapat dicari nilai ekstremumnya (maksimum dan minimum).

Statistik ekstremum (nilai maksimum dan minimum)

```
In [25]: # Statistik ekstremum maksimum dan minimum
         print('Maksimum: ')
         display(df[['Year', 'Beats Per Minute (BPM)', 'Energy',
            'Danceability', 'Loudness (dB)', 'Liveness', 'Valence', 'Length (Duration)',
            'Acousticness', 'Speechiness', 'Popularity']].max())
         print('Minimum: ')
         display(df[['Year', 'Beats Per Minute (BPM)', 'Energy',
            'Danceability', 'Loudness (dB)', 'Liveness', 'Valence', 'Length (Duration)',
            'Acousticness', 'Speechiness', 'Popularity']].min())
```

```
Maksimum:

Year                      2019
Beats Per Minute (BPM)     206
Energy                     100
Danceability                96
Loudness (dB)               -2
Liveness                    99
Valence                     99
Length (Duration)         1412
Acousticness                99
Speechiness                 55
Popularity                 100
dtype: int64

Minimum:

Year                      1956
Beats Per Minute (BPM)      37
Energy                       3
Danceability                10
Loudness (dB)              -27
Liveness                     2
Valence                      3
Length (Duration)           93
Acousticness                 0
Speechiness                  2
Popularity                  11
dtype: int64
```

Per genre lagu, nilai statistiknya dapat dicari. Di sini dicari nilai rata-rata dan standar deviasinya untuk menentukan genre apa yang menguasai atribut tertentu.

Top Genre	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Acousticness	Speechiness	Popularity
indie anthem-folk	69.0	62.0	59.0	-6.0	13.0	27.0	2.0	11.0	75.0
cyberpunk	74.0	30.0	43.0	-14.0	23.0	16.0	70.0	3.0	61.0
contemporary country	78.5	54.5	48.5	-7.5	16.0	41.5	12.0	3.0	65.5
british singer-songwriter	79.0	60.0	51.0	-9.0	11.0	40.0	44.0	3.0	66.0
atl hip hop	80.0	97.0	73.0	-2.0	18.0	97.0	10.0	7.0	79.0
...	...	...	...	...	...	...	...	...	...
bebop	174.0	26.0	45.0	-13.0	7.0	60.0	54.0	4.0	65.0
latin alternative	176.0	74.0	66.0	-7.0	22.0	91.0	66.0	8.0	70.0
latin	178.5	67.0	55.5	-6.0	6.5	72.5	44.0	9.0	71.0
electro house	180.0	63.0	47.0	-5.0	11.0	16.0	3.0	5.0	80.0
laboratorio	189.0	76.0	43.0	-8.0	55.0	66.0	2.0	4.0	57.0

149 rows × 9 columns

Rata-rata menunjukkan genre indie anthem-folk memiliki BPM paling rendah dan genre laboratorio memiliki BPM paling tinggi.

Top Genre	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Acousticness	Speechiness	Popularity
classic italian pop	150.0	40.0	45.0	-12.0	6.0	17.0	3.0	3.0	24.0
compositional ambient	105.0	24.0	61.0	-17.0	25.0	37.0	76.0	3.0	28.0
streektaal	98.0	21.0	61.0	-20.0	8.0	42.0	60.0	3.0	28.0
british alternative rock	87.5	65.5	54.5	-7.5	21.0	38.0	5.0	3.0	28.0
dutch prog	130.0	55.6	50.6	-9.6	22.4	45.6	4.4	3.8	31.2
...	...	...	...	...	...	...	...	...	...
electro house	160.0	63.0	47.0	-5.0	11.0	16.0	3.0	5.0	80.0
ccm	148.0	63.0	49.0	-7.0	9.0	8.0	68.0	3.0	80.0
australian psych	121.0	81.0	62.0	-5.0	14.0	68.5	0.5	3.5	80.5
indie pop	128.0	79.0	65.5	-5.0	18.0	79.5	7.5	4.5	83.0
celtic punk	119.0	65.0	32.0	-6.0	24.0	47.0	50.0	3.0	83.0

149 rows × 9 columns

Rata-rata menunjukkan genre classic italian pop memiliki Popularity paling rendah dan genre indie pop dan celtic punk memiliki Popularity paling tinggi.

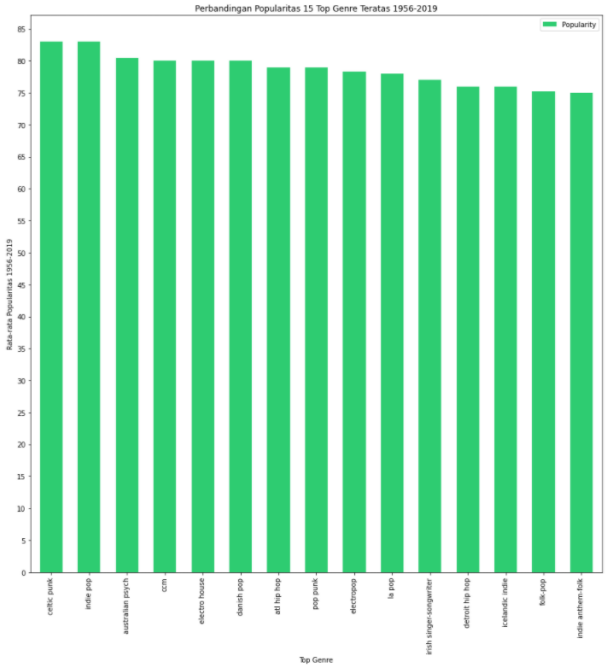
Distribusi frekuensi top genre pada data ditunjukkan oleh tabel berikut.

Distribusi frekuensi top genre pada data

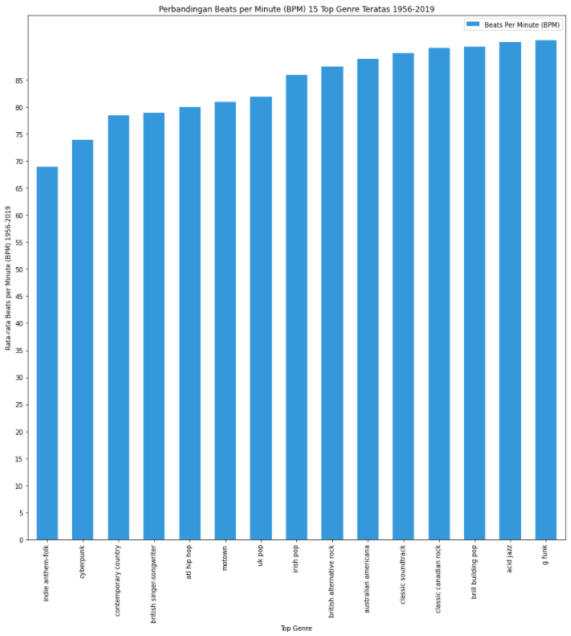
```
In [20]: # Distribusi frekuensi genre pada data.
         display(df['Top Genre'].value_counts())

album rock          413
adult standards     123
dutch pop            88
alternative rock     86
dance pop            83
                   ...
hard rock             1
pop punk              1
alaska indie          1
gangster rap          1
laboratorio           1
Name: Top Genre, Length: 149, dtype: int64
```

Dari data di atas, dapat disimpulkan bahwa genre lagu yang paling banyak muncul dari tahun 1956 sampai dengan 2019 adalah album rock.

Berikut adalah visualisasi perbandingan kategori.

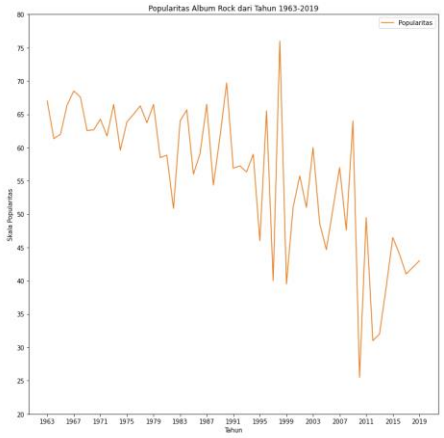


Dari grafik dapat dilihat bahwa ada dua genre yang rata-rata popularitasnya paling tinggi, yaitu celtic punk dan indie pop.

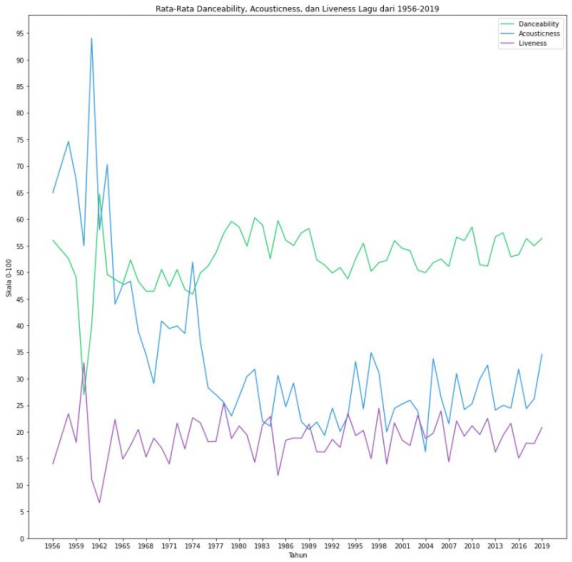


Dari grafik dapat dilihat bahwa ada genre dengan rata-rata BPM paling lambat adalah indie anthem-folk.

Selanjutnya, visualisai perubahan terhadap waktu.

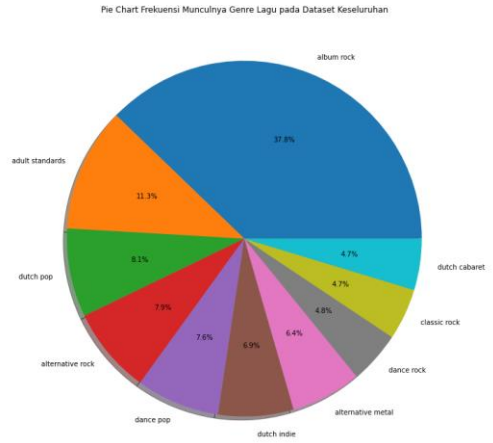


Dari grafik, dapat dilihat perkembangan popularitas album rock dari tahun ke tahun. Dari tahun ke tahun, popularitasnya cenderung menurun dengan tingkat popularitas tertinggimya berada di antara 1995-1999.

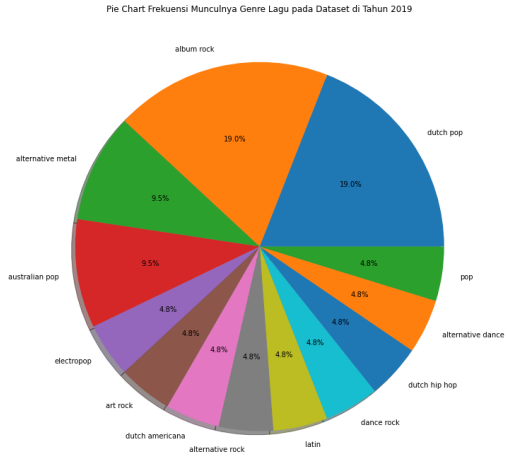


Dari grafik, dapat dilihat perkembangan rata-rata Danceability, Acousticness, dan Liveness lagu-lagu per tahun. Acousticness cenderung menurun dari tahun ke tahun, sedangkan danceability dan liveness cenderung stabil.

Lalu, ada visualisasi hierarki dan hubungan keseluruhan-bagian.

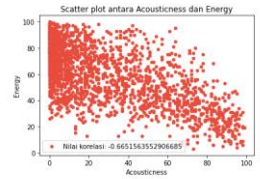


Dari pie chart di atas dapat dilihat bahwa genre album rock mendominasi dataset ini apabila meninjau semua lagu yang di dataset dari 1956-2019.

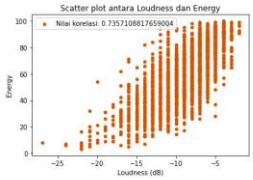


Berbeda dengan kesimpulan sebelumnya, apabila meninjau lagu-lagu terbitan tahun 2019, genre album rock mampu disaingi oleh dutch pop.

Visualisasi terakhir yang ditampilkan adalah visualisasi *plotting relationship*.



Dari gambar di atas, walaupun tidak terlalu jelas (nilai mutlak korelasi ~0.66), ditunjukkan suatu korelasi negatif antara acousticness dan energy. Semakin tinggi acousticness, cenderung semakin turun nilai energy.

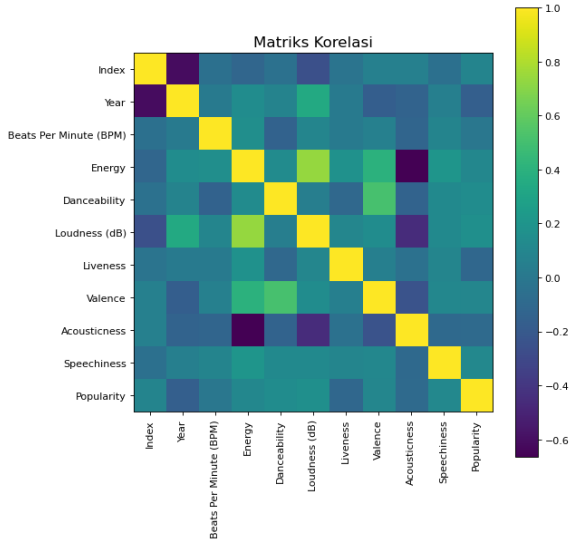


Dari gambar di atas, walaupun tidak terlalu jelas (nilai mutlak korelasi ~0.74), ditunjukkan suatu korelasi positif antara loudness dan energy. Semakin tinggi loudness, cenderung semakin tinggi nilai energy.

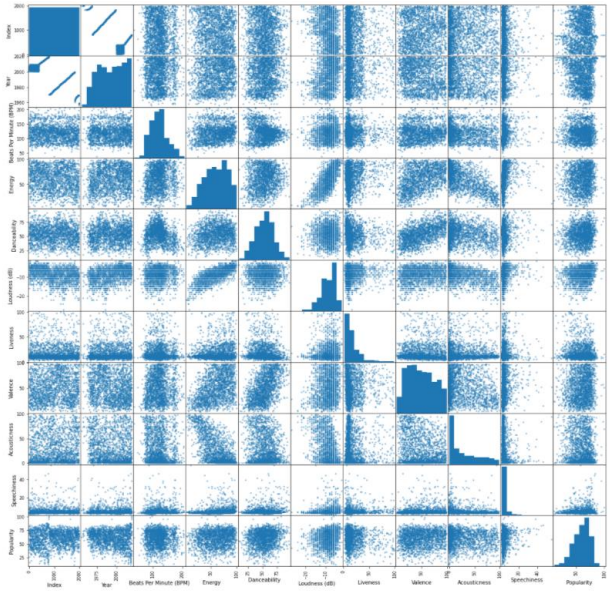
Untuk korelasi antaratribut pada data, yang pertama ditunjukkan adalah tabel korelasinya, dengan nilai mutlak yang mendekati satu menunjukkan keterhubungan antara dua atribut.

	Index	Year	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Acousticness	Speechiness	Popularity
Index	1.000000	-0.607910	-0.048918	-0.114307	-0.047156	-0.260179	-0.027125	0.063304	0.057346	-0.050991	0.087442
Year	-0.607910	1.000000	0.012570	0.147235	0.077493	0.343764	0.019017	-0.166163	-0.132946	0.054097	-0.158962
Beats Per Minute (BPM)	-0.048918	0.012570	1.000000	0.156644	-0.140602	0.092927	0.016256	0.059653	-0.122472	0.085598	-0.003181
Energy	-0.114307	0.147235	0.156644	1.000000	0.139616	0.735711	0.174118	0.405175	-0.665156	0.205865	0.103393
Danceability	-0.047156	0.077493	-0.140602	0.139616	1.000000	0.044235	-0.103063	0.514564	-0.135769	0.125229	0.144344
Loudness (dB)	-0.260179	0.343764	0.092927	0.735711	0.044235	1.000000	0.098257	0.147041	-0.451635	0.125090	0.165527
Liveness	-0.027125	0.019017	0.016256	0.174118	-0.103063	0.098257	1.000000	0.050667	-0.046206	0.092594	-0.111978
Valence	0.063304	-0.166163	0.059653	0.405175	0.514564	0.147041	0.050667	1.000000	-0.239729	0.107102	0.095911
Acousticness	0.057346	-0.132946	-0.122472	-0.665156	-0.135769	-0.451635	-0.046206	-0.239729	1.000000	-0.098256	-0.087604
Speechiness	-0.050991	0.054097	0.085598	0.205865	0.125229	0.125090	0.092594	0.107102	-0.098256	1.000000	0.111689
Popularity	0.087442	-0.158962	-0.003181	0.103393	0.144344	0.165527	-0.111978	0.095911	-0.087604	0.111689	1.000000

Selanjutnya, matriks korelasi untuk memperjelas tabel di atas. Semakin kuning-hijau atau ungu gelap elemen matriks, semakin tinggi keterhubungan dua atribut yang bersangkutan.



Lalu, matriks yang menunjukkan grafik scatter untuk tiap atribut data. Semakin miring lurus (dalam bentuk $y = mx + c$) grafik yang dihasilkan semakin besar keterhubungan antara dua yang bersangkutan.



Dari ketiga tabel/visualisasi di atas dapat diambil hubungan-hubungan berikut:

1. Loudness dan Energy serta Valence dan Danceability berbanding lurus.
2. Acousticness dan Energy serta Acousticness dan Loudness berbanding terbalik.

3. Data-data lain tidak memiliki hubungan yang jelas.

Untuk bagian data cleansing, pada data yang kami pakai, tidak ada data kosong seperti yang ditunjukkan oleh tabel di bawah.

Persen Data Kosong

```
In [26]: df1.isna().sum()

Out[26]: Index                      0
         Title                      0
         Artist                     0
         Top Genre                  0
         Year                       0
         Beats Per Minute (BPM)     0
         Energy                     0
         Danceability               0
         Loudness (dB)              0
         Liveness                   0
         Valence                    0
         Length (Duration)          0
         Acousticness               0
         Speechiness                0
         Popularity                 0
         dtype: int64
```

--> berarti persen data kosong untuk semua data adalah 0%

Namun, ada satu masalah yaitu adanya data kuantitatif yang menggunakan string (karena menggunakan tanda koma sebagai pemisah ribuan)

```
844 843,Echoes,Pink Floyd,album rock,1971,134,32,28,-17,11,14,"1,412",37,4,58
```

Untuk mengatasinya, cukup memberi tahu Pandas bahwa koma di string tipe ini merupakan pemisah angka ribuan.

```
df = pd.read_csv('Spotify-2000.csv', thousands=r',')
```

V. Conclusion

Dengan menggunakan bantuan bahasa pemrograman Python dengan modul Pandas dan Matplotlib, kita dapat membuat analisis terhadap suatu data mentah untuk menghasilkan suatu kesimpulan yang berguna.

Di laporan ini, penulis telah melakukan analisis terhadap data top 2000 Spotify. Dari analisis ini, penulis dapat mengetahui genre lagu apa saja yang memiliki popularitas tertinggi, bertempo tertinggi, dan sebagainya. Selain itu, penulis juga dapat mengetahui perbandingan ataupun perkembangan dari masing-masing genre musik atau atribut musik. Hubungan antaratribut juga dapat ditentukan, contoh hubungan yang penulis dapatkan adalah *loudness* dan *energy* saling berhubungan lurus yang memiliki arti ketika *loudness* meningkat, nilai *energy* meningkat.

References

[1] Nelli, Fabio. 2018. *Python Data Analytics With Pandas, NumPy, and Matplotlib, Second Edition*. New York. Apress Media LLC.

[2] McKinney, Wes. 2018. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* Sebastopol, CA: O'Reilly Media, Inc.