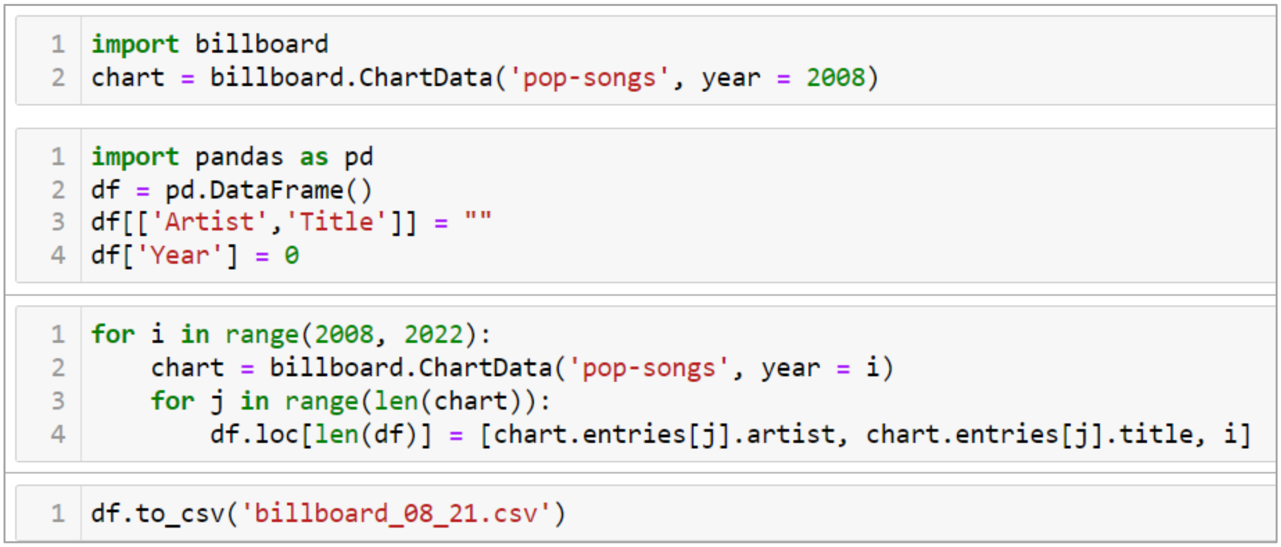
Name: Arun Ganapathy
Date: 9th December 2022

Introduction

This project aims to visualize the performance of various music artists in different genres from 2010 – 2019. These detailed datasets were obtained from a Kaggle-based analytics project and the Python-based 'billboard' library (links below). I would like to thank *Prof. Pak Chung Wong* for giving me the opportunity to work on such an interesting Individual Final Project.

Data Collection and Cleaning

I collected data from the billboard library in Python to get information on yearly track releases and genre classification. This was done on Jupyter Notebooks (*Figure 1*). The data was then saved, followed by cleaning and massaging.



```python
import billboard
chart = billboard.ChartData('pop-songs', year = 2008)
```

```python
import pandas as pd
df = pd.DataFrame()
df[['Artist','Title']] = ""
df['Year'] = 0
```

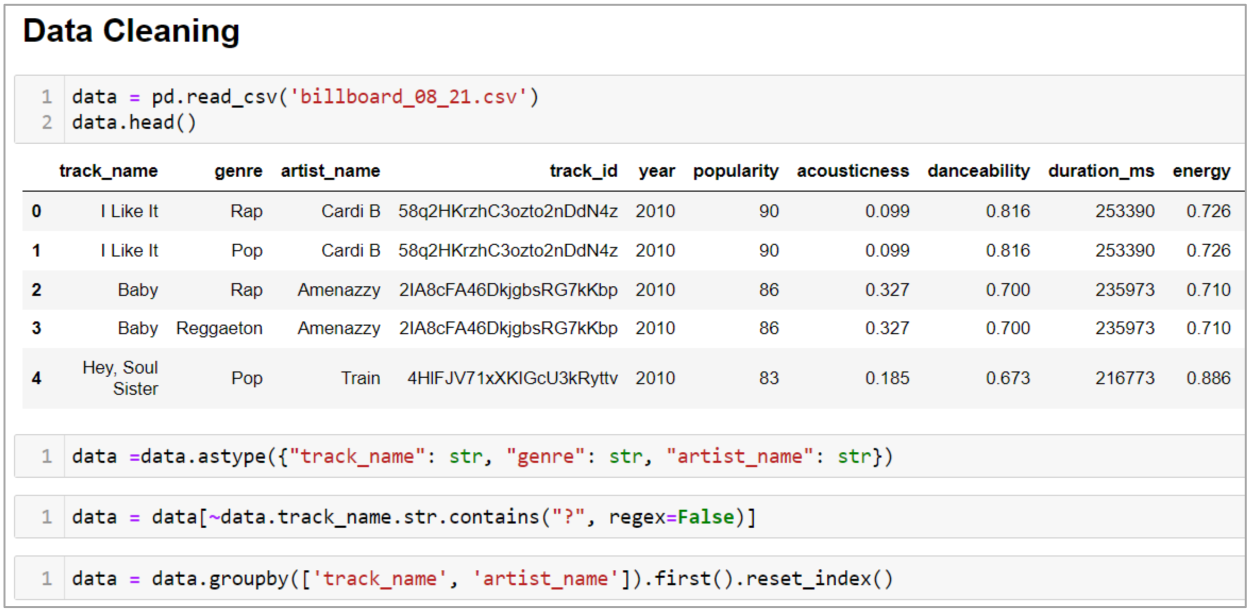
```python
for i in range(2008, 2022):
    chart = billboard.ChartData('pop-songs', year = i)
    for j in range(len(chart)):
        df.loc[len(df)] = [chart.entries[j].artist, chart.entries[j].title, i]
```

```python
df.to_csv('billboard_08_21.csv')
```

Figure 1: Billboard Data Collection – Jupyter Notebooks

Metrics like *track_name* had multiple tracks that were named in different languages that Tableau would not be able to recogonize and hence these columns were filtered for songs released in English only and that were released between 2008 to 2021. This was later filtered down to 2010 to 2019 due to lack of Spotify data. The Spotify data was collected from a Kaggle data analysis project [2]. There were also multiple instances and variations in the way genres were spelt out. For eg, the popular genre 'Reggae' was also spelt out as 'reggae', 'reggea', etc, and this gave me the opportunity to use my pre-existing knowledge in Python to clean the data apporopriately before moving forward with the visualizations.

The Spotify dataset had multiple entries for each track released by an artist based on all the genres that the track could be classified. For eg, the track 'I Like It' by the artist 'Cardi B' has two entries under the genres 'Rap' and 'Pop'. However, intuitively this song should fall under the 'Rap' genre. This stalls the analysis process. To rank artists based on popularity score and other metrics, we require unique entries for each song released in a year. Hence, unique entries were extracted based on the first appearance of a genre for any given track. *(Figure 2)*



Data Cleaning

```
data = pd.read_csv('billboard_08_21.csv')
data.head()
```

	track_name	genre	artist_name	track_id	year	popularity	acousticness	danceability	duration_ms	energy
0	I Like It	Rap	Cardi B	58q2HKrzhC3ozto2nDdN4z	2010	90	0.099	0.816	253390	0.726
1	I Like It	Pop	Cardi B	58q2HKrzhC3ozto2nDdN4z	2010	90	0.099	0.816	253390	0.726
2	Baby	Rap	Amenazzy	2IA8cFA46DkjgbsRG7kKbp	2010	86	0.327	0.700	235973	0.710
3	Baby	Reggaeton	Amenazzy	2IA8cFA46DkjgbsRG7kKbp	2010	86	0.327	0.700	235973	0.710
4	Hey, Soul Sister	Pop	Train	4HIFJV71xXKIGcU3kRyttv	2010	83	0.185	0.673	216773	0.886

```
data =data.astype({"track_name": str, "genre": str, "artist_name": str})
```

```
data = data[~data.track_name.str.contains("?", regex=False)]
```

```
data = data.groupby(['track_name', 'artist_name']).first().reset_index()
```

Figure 2: Billboard Data Cleaning and Massaging – Jupyter Notebooks

Dataset Description

The dataset includes many metrics including artist_name, track_name, track_id, year, popularity, acousticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time_signature, and valence. All these metrics are defined below to enable to reader to grasp these concepts before comprehending the analyses. These definitions are generic and presice. [1]

1. *acousticness*: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

2. *artist_name*: The name(s) of the artist(s) who perform the track.

3. *danceability*: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
4. Duration_ms: A track's duration in milliseconds.

5. *energy*: A measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

6. *track_id*: Track's unique identifier.

7. *instrumentalness*: Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

8. *key*: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.

9. *liveness*: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

10. *loudness*: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. NOTE: Loudness is measured negatively here on purpose. This is because we are dealing with digital sound that a computer can listen to which is measured differently than what the human ear can listen to. This is sometimes why you will see negative numbers when you adjust the volume on your surround sound system.

11. *mode*: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

12. *track_name*: Name of Track

13. *popularity*: The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.

14. *speechiness*: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

15. *tempo*: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

16. *valence*: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

17. *year*: Year of track's release.

Project Approach/Flow

1. *Data Cleaning – QC and Validation* - This process was completed in Python using the billboard library and a Spotify API.

2. *Data Wrangling and Manipulation* - Massage the data to remove tracks that were not in the English language as Tableau does not accept other inputs.

3. *Data Visualization* - Tableau Worksheets to visualize the data. Dashboards and a Story Board for the final presentation along with a report.

Implementation and Results

Initially, I went through metrics that define a song and populated a correlation matrix to compare their strengths. *Table 1* shows us that all metrics except for danceability had a weak negative correlation to popularity. Danceability had a weak but positive correlation to popularity which isn't surprising as it aligns with my intuition that most people dance and feel alive through highly popular songs.

	popularity	***danceability***	***energy***	***tempo***
popularity	1.00			
danceability	0.12	1.00		
energy	-0.10	-0.17	1.00	
tempo	-0.16	-0.04	-0.06	1.00

Table 1: Correlation Matrix

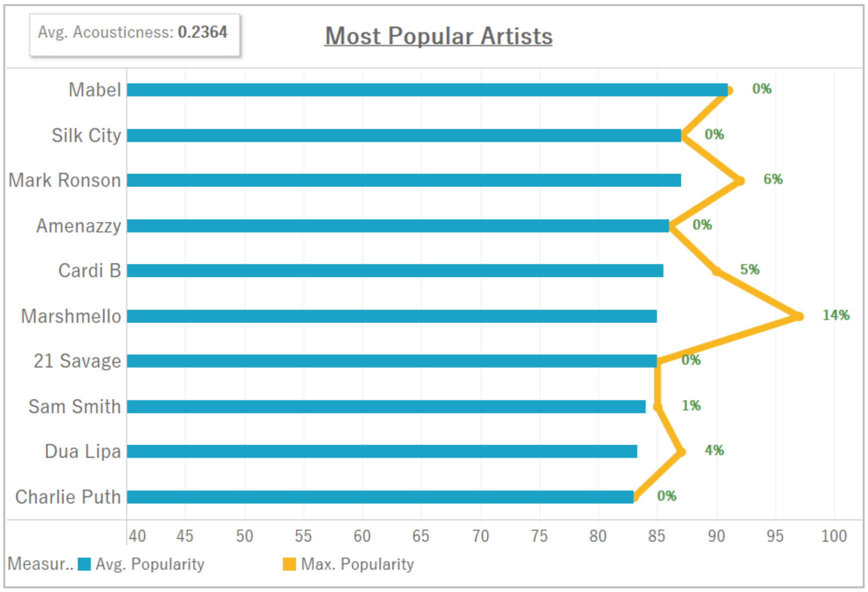


Figure 3: Most Popular Artists

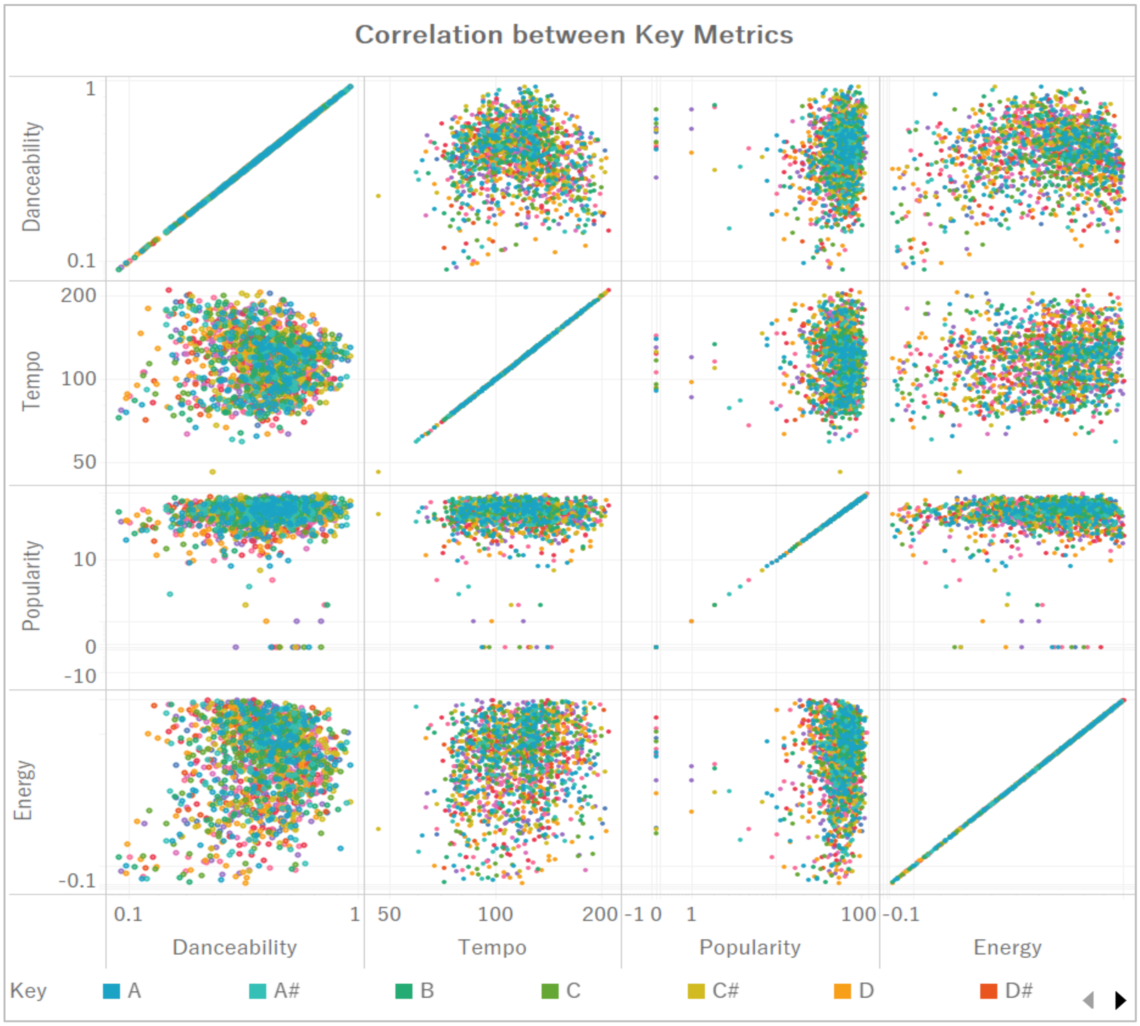


Figure 4: Correlation between different metrics that affect the popularity score

Figure 4 is a scatter plot that visualizes the different correlations between danceability, tempo, energy, and popularity. The plots are dense and truly spread out or scattered. This is evidence of low correlation. This shows us that there might be other variables (not present in this dataset) that might be contributing to the popularity score of a track.

Story Point 2 highlights the key KPIs including average popularity, median tempo, average duration, average acousticness, average instrumentalness, average liveness. The intuition behind calculating the median of tempo is based on the fact that this metric has discrete values, i.e, artists tend to use very limited range or values of temp. with the median tempo being 120 beats per minute. Figure 1 describes the most popular artists ranked by average popularity and the increase in popularity from the average in percentage. This clearly shows us that popular artists like Marshmello, Cardi B, Mark Ronson, and Dua Lipa released songs that were very popular compared to their average popularity.

Dashboard 2 is tasked to run an comprehensive analysis on various metric by genre. Genre categorizes different tracks and artists based on the category of the songs they release. Genre includes values such as

Pop, Hip-Hop, Rock, etc. However, at this change the idea is to analyse different artists within their genre to better understand how metrics like popularity, tempo, etc vary.

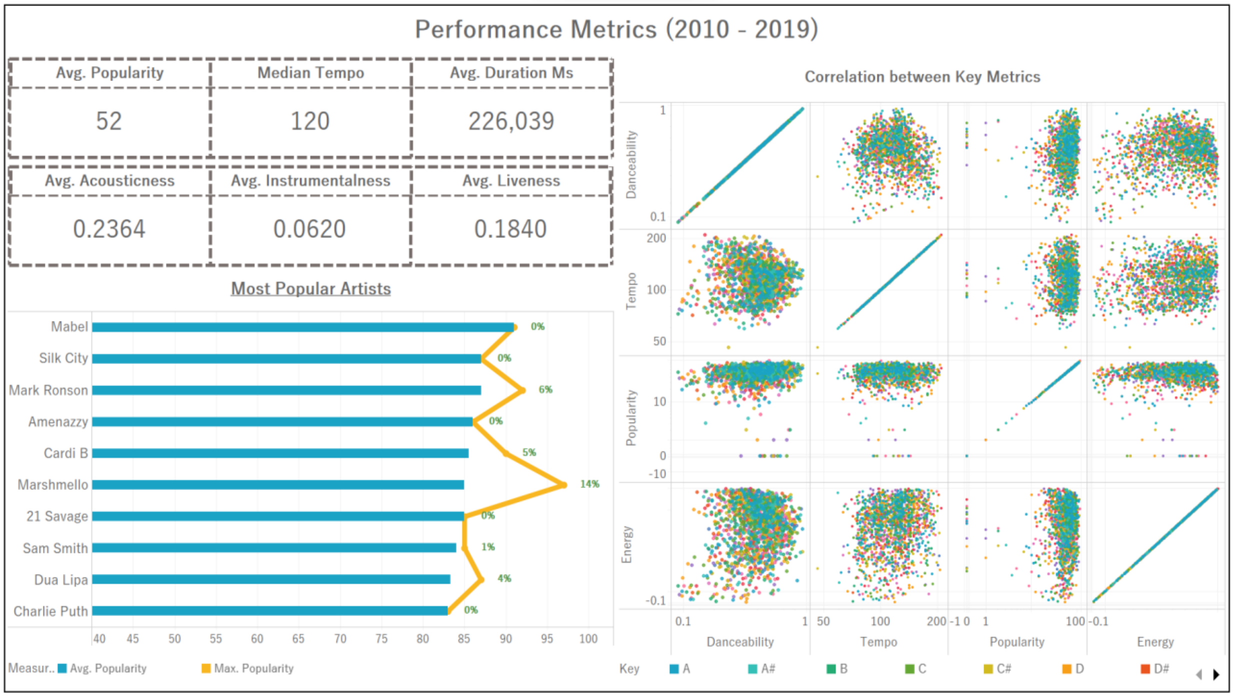
SPOTIFY TOP 20 TRACKS (2010 - 2019)

‹ 1 2 3 4 5 ›

Individual Final Project
Arun Ganapathy

AURORA
The River

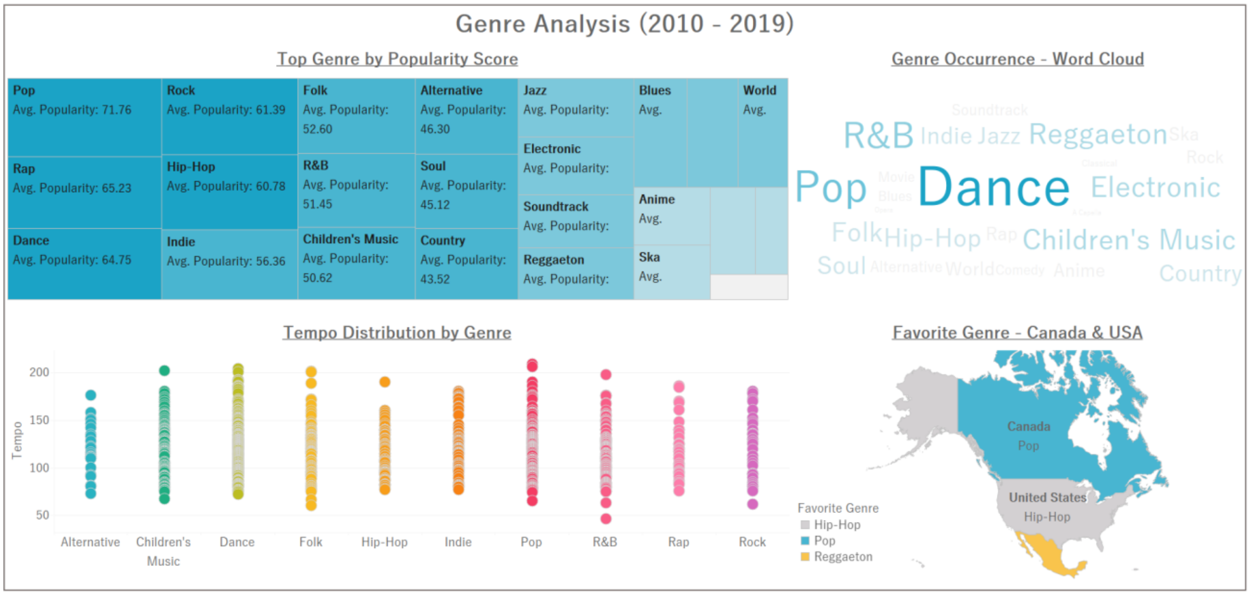
Story Point 1: Intro Page – Inlcudes a Spotify GIF animation



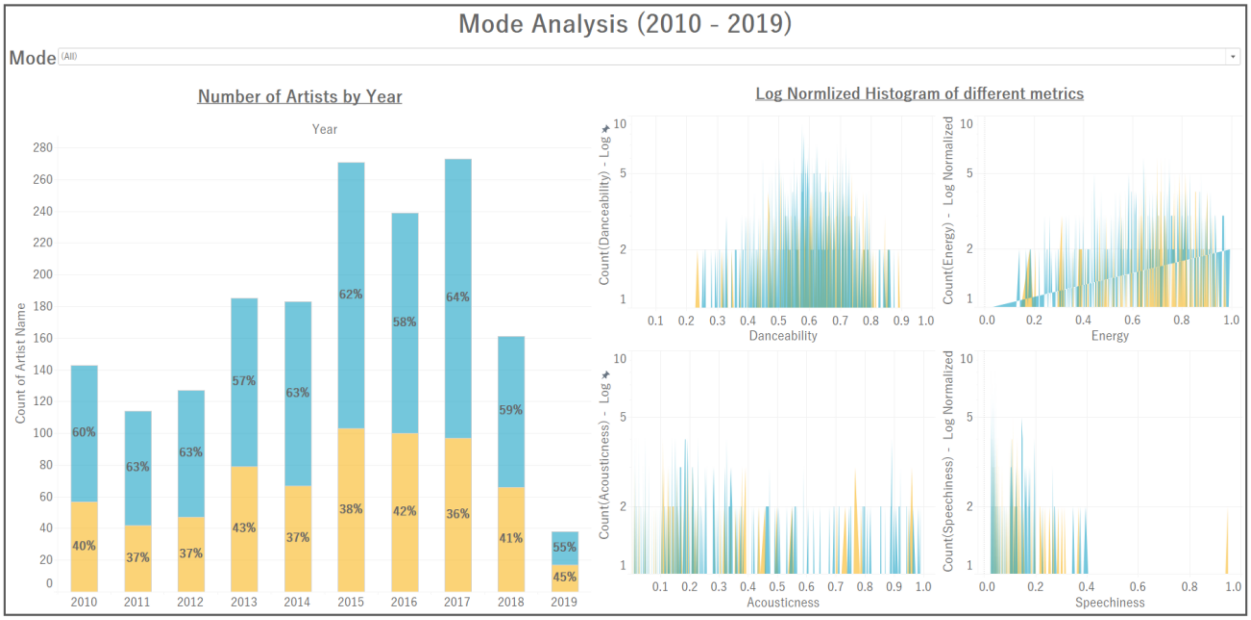
Performance Metrics (2010 - 2019)

Avg. Popularity	Median Tempo	Avg. Duration Ms
52	120	226,039
Avg. Acousticness	**Avg. Instrumentalness**	**Avg. Liveness**
0.2364	0.0620	0.1840

Story Point 2: Key Performance Metrics



Story Point 3: Genre Analysis



Story Point 4: Mode Level Analysis

The histograms in *Story Point 4* show us the following:

- A lot of observations have a value no larger than 0.1 in instrumentalness which is ~80% of the dataset
- Energy and Danceability are pretty normally distribuited, but Valence is normally distributed
- Most of the songs have a loudness level between -5dB and -10db

- Majority tracks have speechiness less than 0.25 indicating that more speechy songs aren’t favoured. It can also be conluded that songs with an above average speachiness are most rap songs.

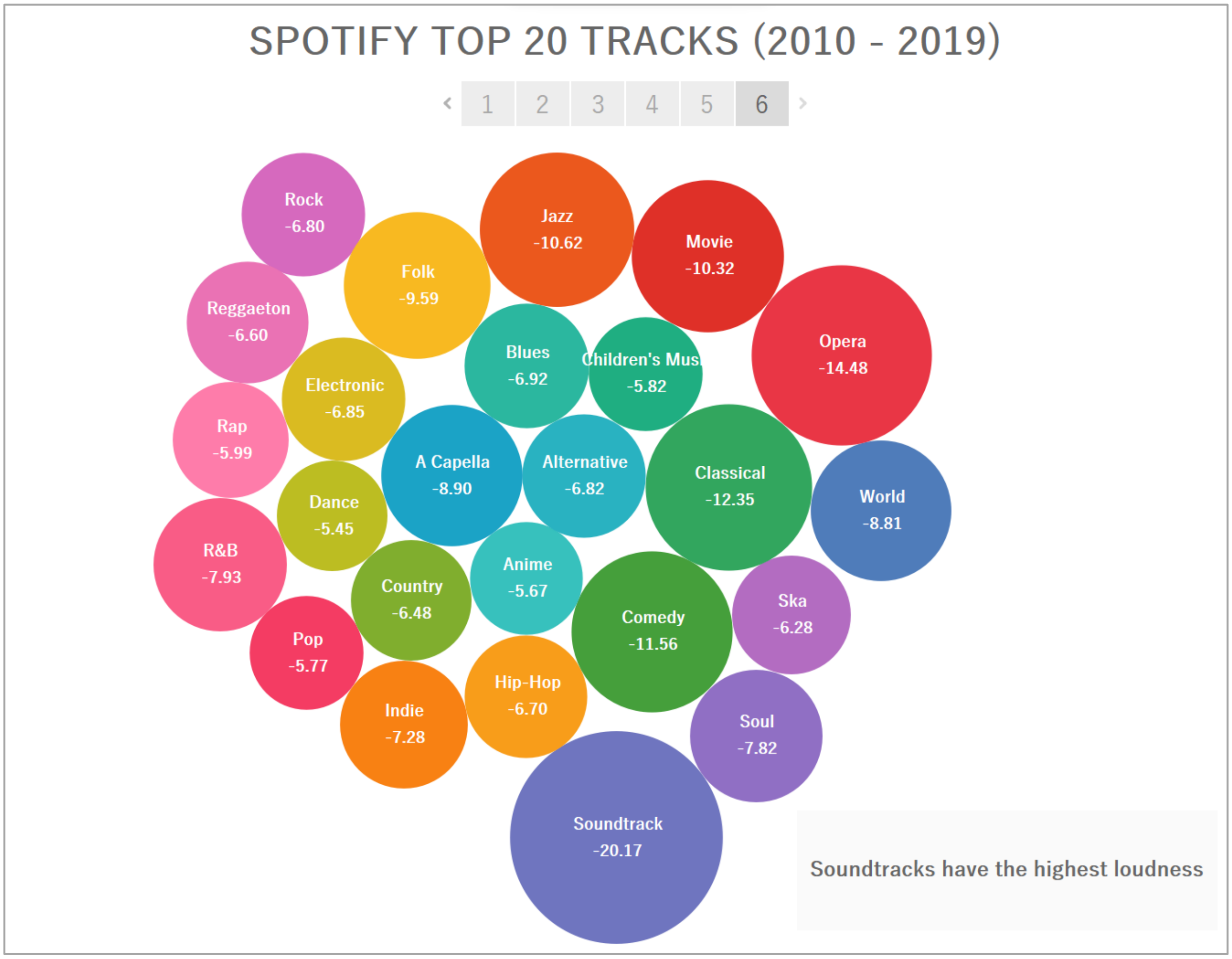


Figure 5: Variation in Loudness by Genre

Finally, *Figure 5* shows the different genres by Loudness. As mentioned earlier in the dataset description, loudness is measured as a negative numerical value. ‘Soundtrack’ has the highest loudness of -20.17 db or decibels, followed by Opera (-14.48 db), and Classical (-12.35 db).

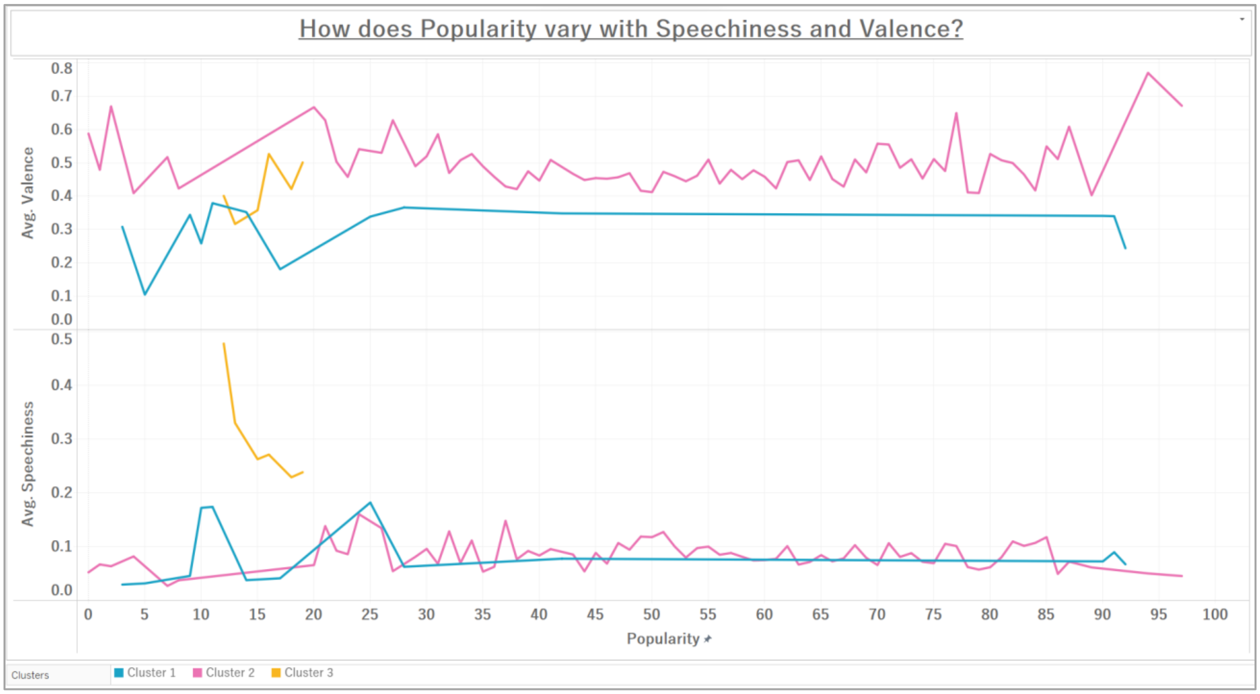


Figure 6: Speechiness and Valence vs. Popularity (3 Clusters)

Figure 6 shows us how popularity varies with Speechiness and Valence. Tableau's inbuilt clustering algorithm is used to divide the speechiness and valence into three different clusters based on their average values with Cluster 3 including the larger values of average speechiness. This is done to analyse if songs with more speech like Rap songs, Hip-Hop songs, have higher popularity. However, we can see that the popularity is quite low for higher average values of valence and speechiness.

Results

The results show that many of the variables are weakly correlated with it. Specifically, the danceability, tempo, and energy have weak correlations with respect to popularity (2). However, we also noted that if there are outliers with respect to a metric, whether it be on the high or low end, it might not be as popular. In all three of the graphs that compare popularity to three metrics discussed above, we noticed that if a metric was too high or low, the popularity is pretty low compared to all the other songs that are not outliers. A more popular song would likely be within the regular distribution for metrics compared to those having a metric that stands out tremendously compared to most songs. Regarding genres of music, we noted that the most popular genres among Spotify listeners were pop, dance, and hip-hop (3). The other seven genres that were highlighted in the dataset were not as popular compared to the three most popular, making them a very distinct departure from the most popular genres. We also noted that tempo did not have much of an effect on whether a genre was popular, the three most popular genres had either a wide or narrow distribution, which shows that it does not have much of an effect on popularity for a certain genre.

Through the comparison of Canada and the United States we noted that despite 'pop' and 'dance' being more popular than 'hip-hop', 'hip-hop' was the most popular genre of music to be streamed on Spotify, compared to Canada, which had 'Pop' as the most popular genre. However, the rankings for US popularity were pretty close seeing as pop and dance are still highly popular genres of music streamed in the US, and does not necessarily show a gigantic departure in what types of genres are popular in the US.

Conclusion

I found the results to be very interesting and insightful as to what might make songs popular and what the streaming trends on Spotify have been throughout the previous decade. There are clearly metrics that help to make songs popular such as danceability and tempo, but if there is too much of them, then it will mean a song will not be as popular, seeing as listeners clearly like to see an even mix of metrics. It also seems that the three most popular genres are pretty consistent year-to-year, and will probably be the same for the 2020s decade, unless there is a drastic change in music trends overtime. One thing we wish we had better access to was streaming data by countries outside of the US and Canada, that way we can see which countries have a popular genre that is very different from the most popular ones in the US and Canada.

Challenges and Setbacks:

One of the biggest challenges was data collection and cleaning. A very valid setback would be inconsistent nature of the yearly data which resulted in the lack of a bar chart racing graphic/visual.

References:

[1]- https://github.com/calebelgut/spotify-lstm
[2]- https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year
[3]- https://github.com/guoguo12/billboard-charts