How music composition key metrics affect popularity of a song?

Data Science: R

Part I - Set-up

I. Loading Packages

Load packages tidyverse and corrplot and reshape2 for data visualization

```
# Load necessary packages
library(tidyverse)
install.packages("corrplot", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/bx/f_3zplfd71jdzz4fbfyc7cq80000gn/T//RtmpFqmwl0/downloaded_packages
```

```
library(corrplot)
install.packages("reshape2", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/bx/f_3zplfd71jdzz4fbfyc7cq80000gn/T//RtmpFqmwl0/downloaded_packages
```

```
library(reshape2)
```

II. Importing the data file used in the analysis

```
# Data files exploratory research
# Import the dataset
music <- read_csv("data_w_genres-1.csv")
# Generate a tibble from the data frame
music_tbl <- as_tibble(music)

# Import csv files to portray background research to build on our hypothesis.
revenue_split <- read_csv("Revenue_Split.csv")
music_properties <- read_csv("Music_Properties.csv")
monthly_active_users <- read_csv("Monthly_Active_Users.csv")
genre_popularity <- read_csv("Genre_Popularity_.csv")
```

III. Nomenclature standards

```r
# Create a data frame with only the top 5 genres
music_top_5 <- music %>%
  filter(Genre %in% c("HipHop/Rap", "Pop", "Electro", "Rock", "Country"))
# Create a data frame with the top 5 genres and the different attributes of music composition
new_music_top_5 <- music_top_5 %>%
  select(acousticness, danceability, energy, instrumentalness, liveness, speechiness, valence, populari
# Renaming the columns for visualization purposes
  rename(Acoust= acousticness, Dance = danceability, Ener = energy, Inst = instrumentalness, Liv = live
```

IV. Function

We have built a function to help the visualization of most plots of this analysis.

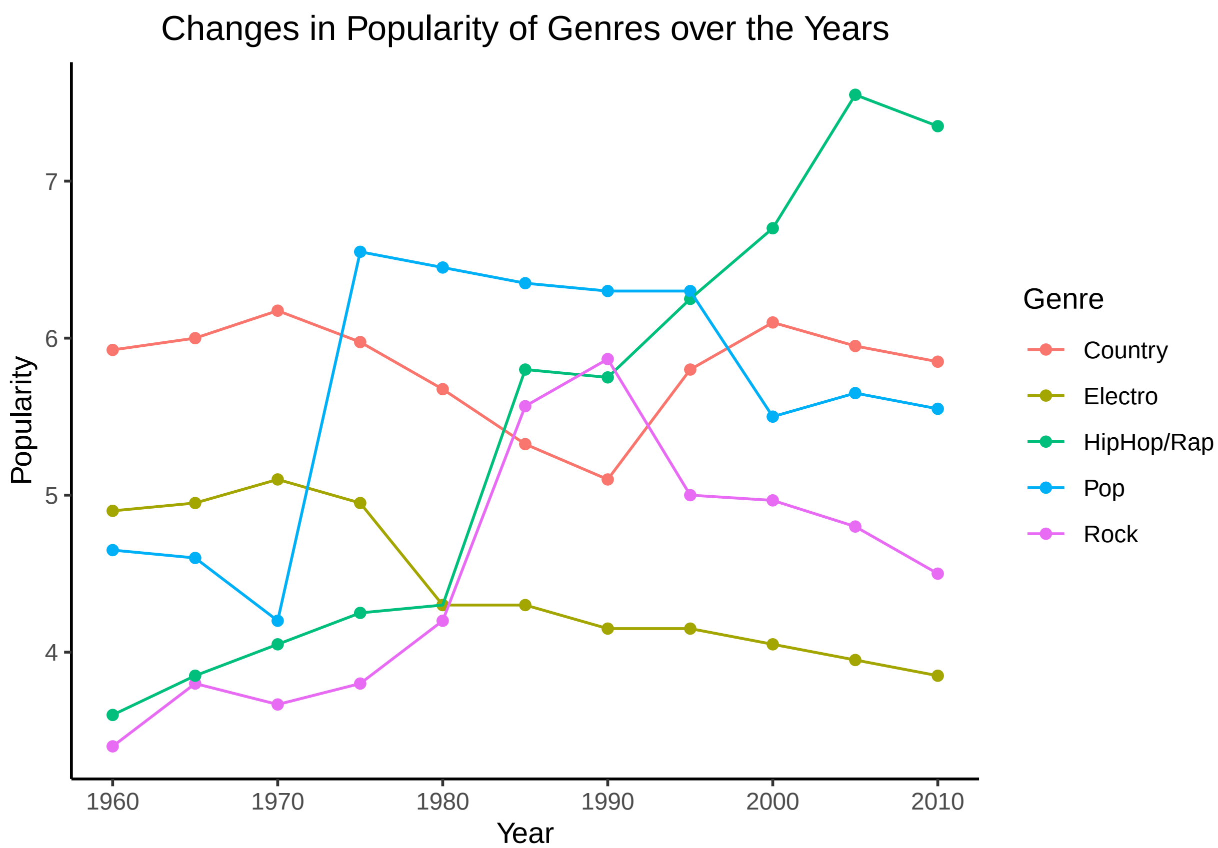
```r
func_visual <- function(data,x,y,Legend,chart_type) {
   plot<- ggplot(data,mapping=aes(x,y,fill = Legend))
   if (chart_type == "col"){
     plot +
       geom_col()
   }
   else if (chart_type == "line"){
     plot +
       geom_line()
   }

   else if (chart_type == "tile"){
     plot +
       geom_tile()
   }

}
```

Part 2 - Framing the Problem

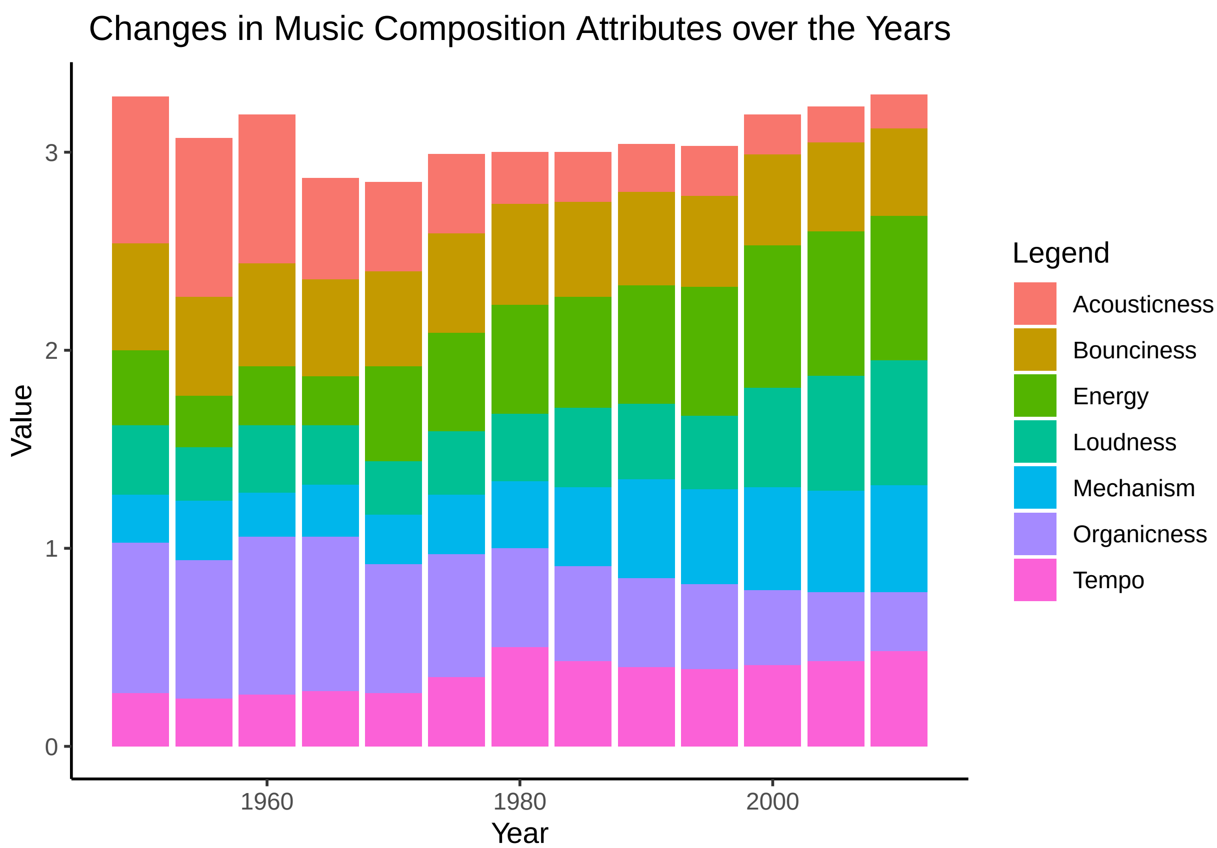
I. Problem Recognition & Exploratory Research

From the beginning, our understanding of cultural shifts in popular music have been reliant on anecdote and history— until now. Working with digitized music files has opened new ways of understanding what it is that makes a track record a hit or a miss. A team of scientists at Queen Mary University of London analyzed roughly 17,000 songs that charted on the U.S. Billboard Hot 100 between 1960 and 2010. Taking note of digital elements in the songs known to correspond with certain chord patterns, rhythms and tonal characteristics, they were able to organize the songs into 13 different categories, which roughly correspond with musical genres.

Here is how they charted how their popularity has looked and shifted over the years.

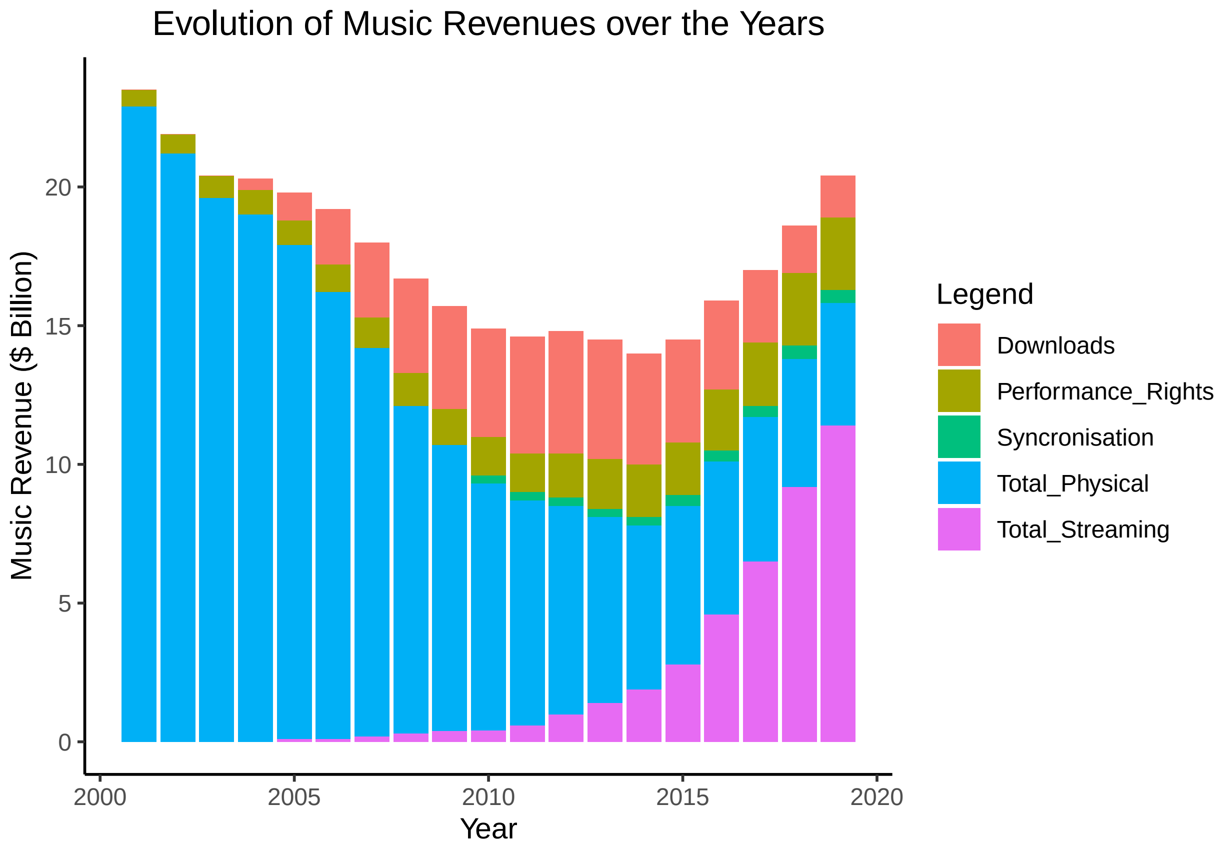


On a similar note, there was a study made by The Echo Nest data alchemist Glenn McDonald who traced nine distinct audio attributes in the 5,000 hottest songs from each year, 1950 to 2013. Seven of those ways showed detectable changes in music from 1950 to the present day and are portrayed in the following line graph.

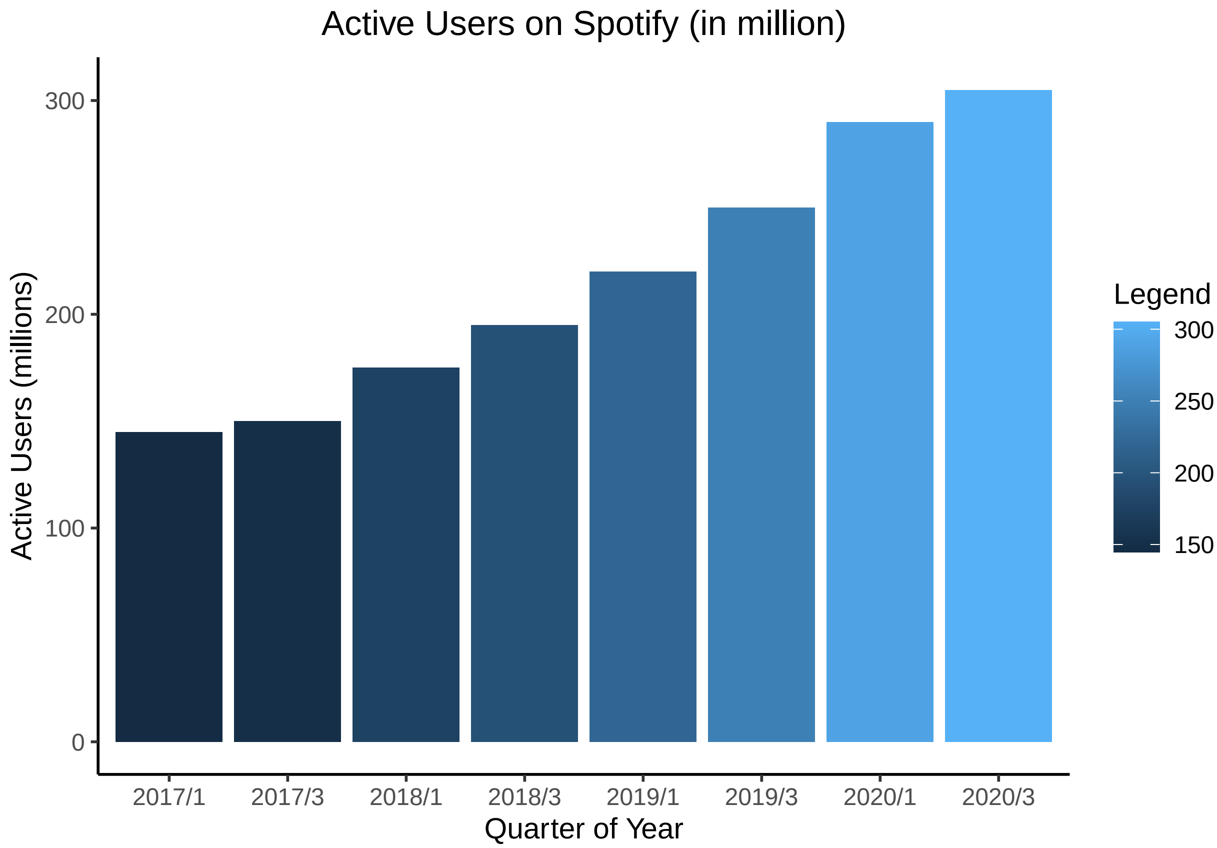


The convenience and personalisation of music streaming, combined with the accessibility afforded by smartphones and smart devices, has driven recorded music's growth. IFPI notes that global streaming revenues grew at a 42% CAGR (compound annual growth rate) since 2015, compared to the entire recording industry's 9% CAGR. The following chart shows not only the evolution of the market share for music, but also how the barriers of entry for new artists have been lowered, making it easier for anyone to offer their music to the world.

With this thought, how could artists know what it is that will make them popular?... it is in the interest of both...



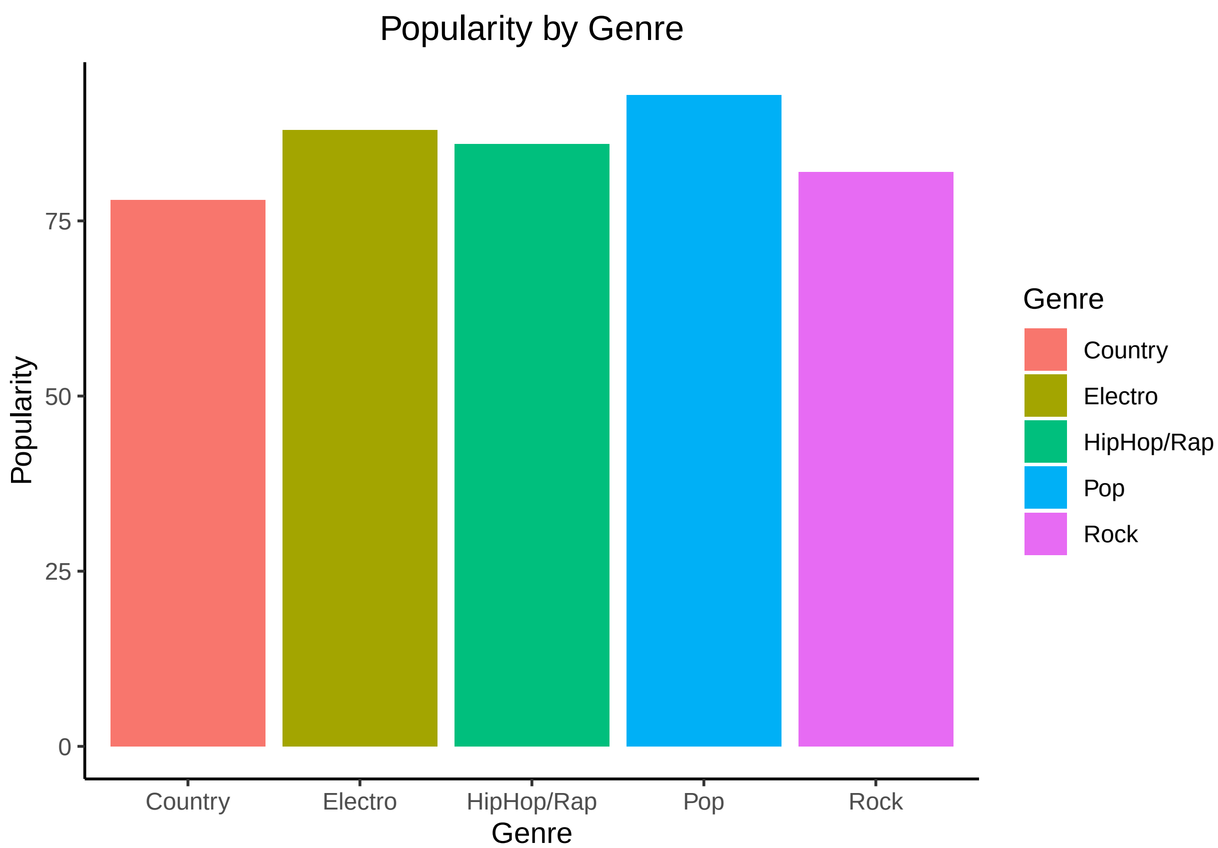
Accompanying this data, the following chart shows Spotify's Monthly Active User's growth in the last 3 years.



There are key attributes in music popularity, which have evolved over the years. It would be ideal for artists to know which metrics affect their music composition, in order to gain more popularity, therefore increasing the overall revenue of their songs through online streaming platform such as Spotify.

II. Testable Hypothesis

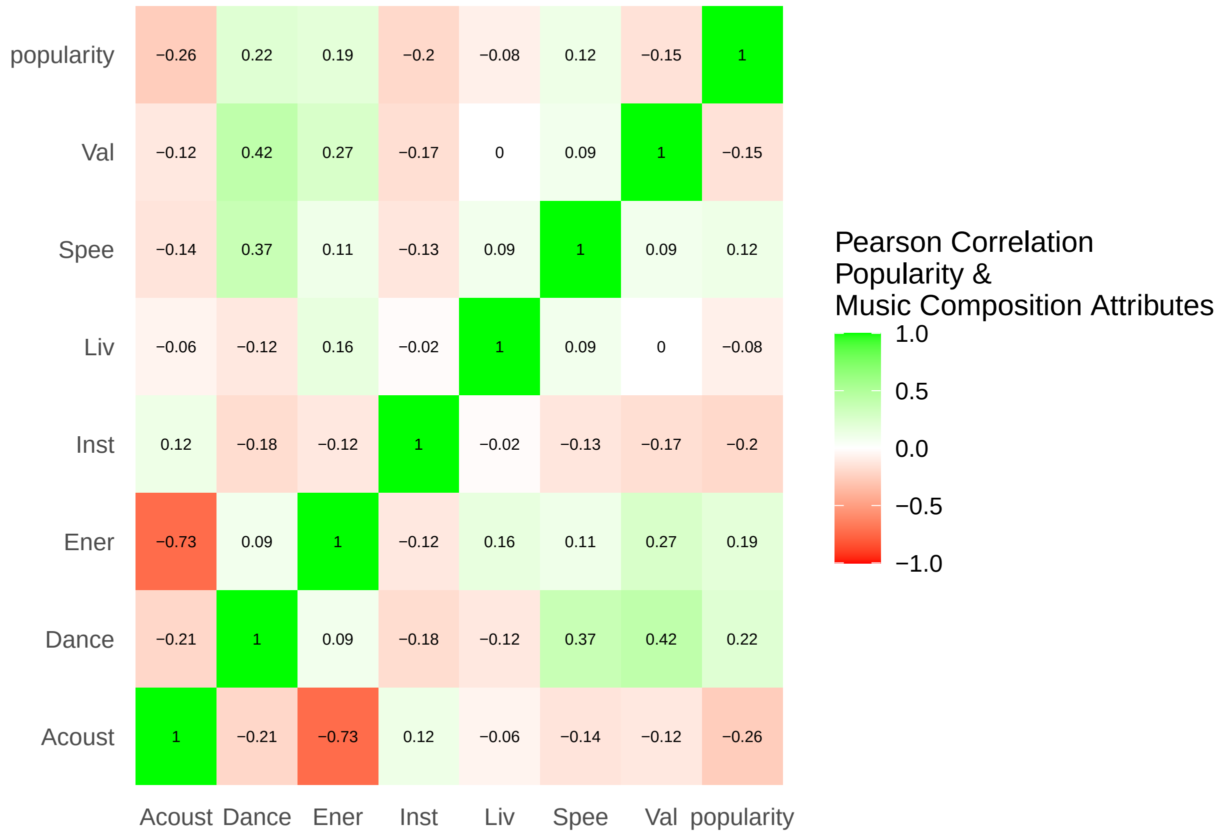
For the sake of this analysis, our first step is evaluating how, if at all, the music genre of a track influences its popularity. Based on our research, the evolution of genre popularity has resulted in artists composing songs within specific genres to not meet the same levels of popularity, therefore generating different revenues from streaming platforms.




```
##
## Call:
## lm(formula = popularity ~ Genre, data = music_top_5)
##
## Coefficients:
##     (Intercept)     GenreElectro  GenreHipHop/Rap         GenrePop
##          39.282            4.569           13.078            9.398
##       GenreRock
##           2.359
```

Now, we know that the music style of track influences its popularity to the masses, the next step is to analyze which top music composition attributes influence each top genre, and therefore the popularity of a song.

We are building a correlation matrix between every music composition attributes (**acousticness, danceability, energy, instrumentalness, liveness, speechiness and valence**)



As seen in our correlation matrix, our top 3 categories, representing the strongest positive correlations, are danceability, energy and speechiness.
Therefore, our hypothesis is: **Do danceability, energy and speechiness influence the popularity of a track?**

H0: No relationship between music composition attributes and popularity
H1: Some relationship between top 3 attributes and popularity

Part III - Solving the Problem

I. Variable Selection

Scaled numerical:
- **Danceability** (range from 0 to 1): based on combination of different music elements (beat strength, rhythm stability, and overall regularity)
- **Energy** (range from 0 to 1): perceptual measure of activity and intensity
- **Speechiness** (range from 0 to 1): presence of spoken words in a track
- **Popularity** (range from 0 to 100): the bigger the number, the more popular the song is. Popularity on Spotify is based on the total number of plays compared to other tracks and how recent those plays are

Categorical:
- **Genre**: music style of each song
- **Artists**: list of artists mentioned for each song

Variable Analysis Process The **music composition** includes these different scaled variables: **acousticness, danceability, energy, instrumentalness, liveness, speechiness and valence**. As stated during the first part of this report, only **danceability, energy and speechiness** will be analyzed to deepen the results of our research.

II. Data Collection

This data set was made available through Spotify Web API by a data-enthusiast on Kaggle (https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks) and is made of 4 different files. The main file contains more than 160,000 songs released between 1921 and 2020, collected from the Spotify Web API. One of the files available classifies the data by genre, and it is the one we have decided to choose. There is a total of 13,240 observations and 15 variables.

Cleaning the Data In order to deepen our analysis, we pre-clean our dataset using a ETL software (Tableau Prep Builder) to clean the Genre variable by **creating different genre clusters** with the main ones which appeared during our exploratory research (HipHop/Rap, Pop, Electro, Rock, Country, Movie, Classical) and we deleted the rows not belonging to any of these clusters.

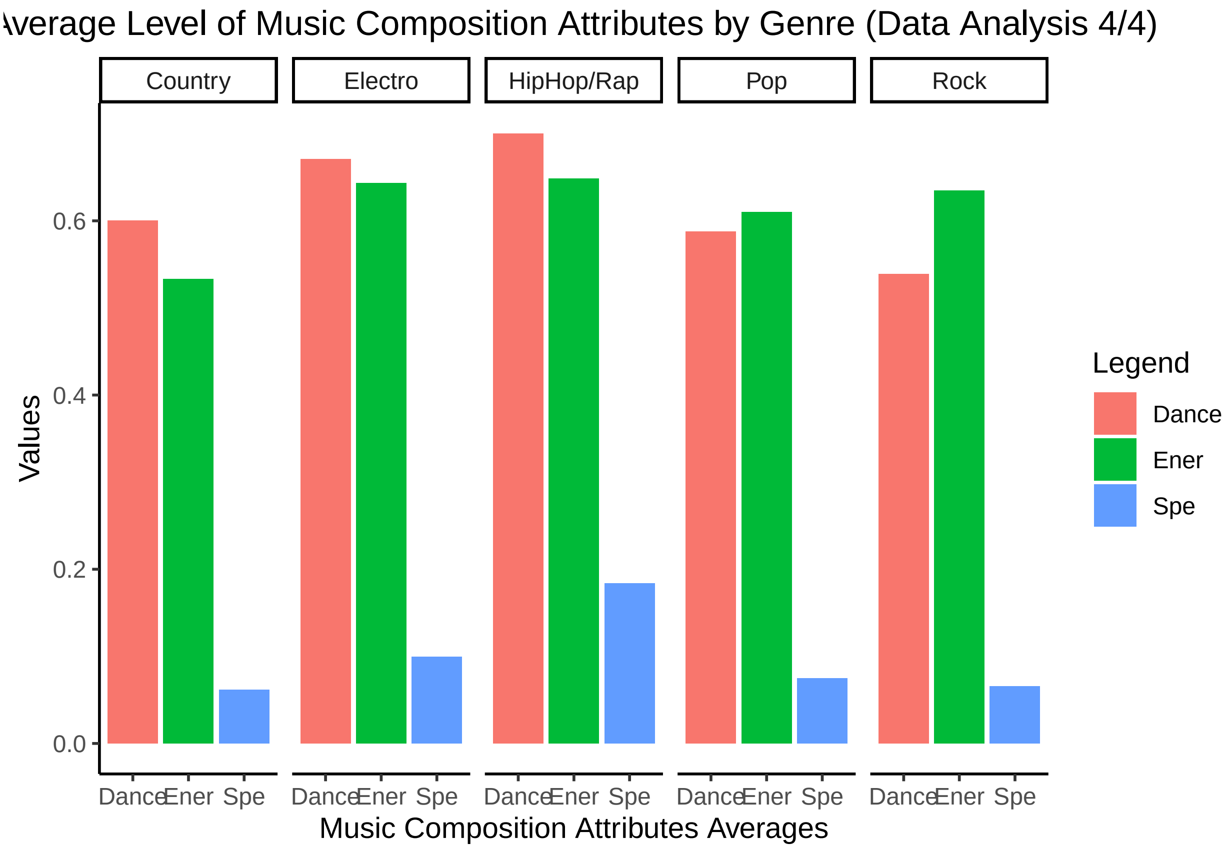
During that pre-clean we also deleted rows with missing values in the genre variable - as it would have biased our analysis.

III. Data Analysis

We created 5 clusters, representing the top 5 genres we have defined based on our exploratory research are: **Hip Hop/Rap, Pop, Electro, Rock & Country**. In order to differentiate each style, an analysis is being held to define each genre in terms of attributes of music composition. Indeed, each genre has a specific composition, giving the music a unique sound and feeling, and therefore popularity.

a. Average Music Composition Attributes per Genre

Now that we selected the top 3 attributes the more correlated to popularity, we can analyze how they represent each genre. As all the following variables are scaled, it is possible to visualize them in a bar chart using `facet_wrap` to see each attribute average per genre.



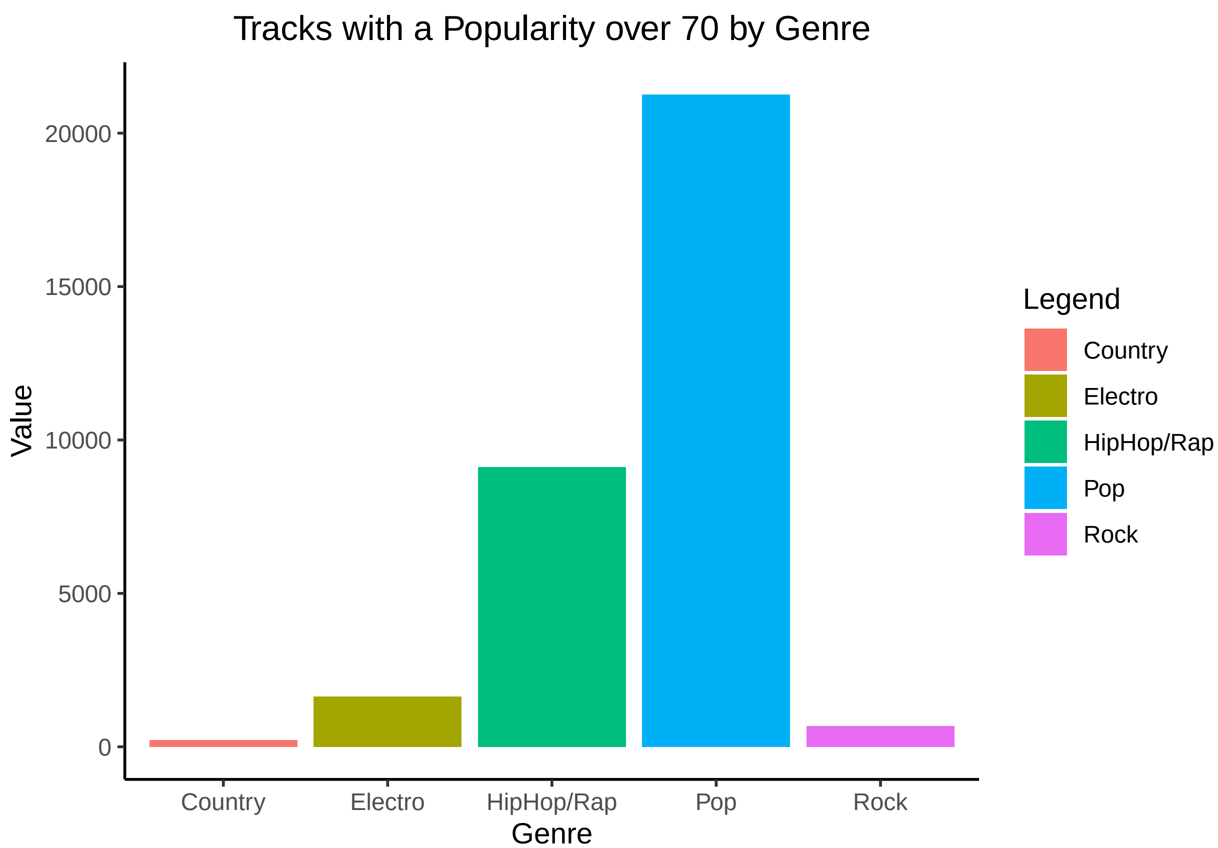
Based on this visualization we can make the following conclusions for each genre:
- **Country**: Lowest energy + Lowest speechiness
- **Electro**: Average level of each attribute when compared to other genres
- **HipHop/Rap**: Higest danceability + Highest speechiness + Highest energy
- **Pop**: Average level of each attribute when compared to other genres
- **Rock**: Lowest danceability

Drawing from these conclusions, it is clear that Pop and Electro genres do not have any specific attribute, from the top 3, that differentiate this genre from the others. The two most differentiated genres in terms of music composition are Country and HipHop/Rap, which is the genre leading in highest popularity.

b. Music Genres of Popular Tracks on Spotify

According to Spotify, the popularity of a song is based on the number of streams and how recent they are. Moreover, the higher the popularity index of a song, the more likely Spotify algorithm will recommend it to new listeners, or place it in the new releases playlists, which are available on the homepage of the user. This will increase the reach of the song even more.

Therefore, we are analyzing how many tracks from the dataset have a popularity equal or bigger than 70 (level where tracks are added to new releases playlists) and from which genre they belong to.



From the dataset, we can conclude that **Pop, Hip/Hop Rap & Electro** are the top 3 genres with the most tracks with a popularity of 70 or more on Spotify.

Therefore, for the Modelling and Communication stage of our analysis, we will be focusing on these top 3 genres and the following music composition attributes: **Danceability, Energy and Speechiness**.

Part IV - Modelling & Communication

After this explanatory data analysis, we can build a linear model to answer our hypothesis : **Do Danceability, Energy and Speechiness influence the popularity of a track?**

As a reminder, our hypotheses are:
H0: No relationship between music composition attributes and popularity
H1: Some relationship between top 3 attributes and popularity

First, we are filtering the data frame to only keep the tracks with a popularity over 70, and for the genres HipHop/Rap, Pop and Electro in order to have more data points for our multiple regression - 441 observations in total.
Secondly, we are building an overall linear model combining all 3 Genres to prove our hypothesis and hopefully for an overview of the prediction model for the success of a song on Spotify.
Finally, we will build one linear model with popularity as a function of danceability, speechiness and energy, for each of the top 3 Genres. Therefore, artists who are specialized in one of these can have a prediction model to compose their songs.

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic   p.value
```

```
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    70.9        1.28    55.5   8.81e-189
## 2 danceability    6.71       1.59     4.23  2.96e-  5
## 3 speechiness    -2.83       2.13    -1.33  1.84e-  1
## 4 energy         -0.301      1.23    -0.245 8.07e-  1
```

From these results, we can build the following popularity prediction model:

$$PredictedPopularity = 70.8682629 + (6.7116851 * danceability) - (2.8268618 * speechiness) - (0.3012474 * energy)$$

All the t-values are far from 0, which shows that the coefficients are statistically significant and we can reject the null hypothesis.

As a reminder:
H0: No relationship between music composition attributes and popularity
H1: Some relationship between top 3 attributes and popularity

Based on our model, we can conclude that there is a relationship between danceability, energy and speechiness in a song's composition and its popularity on Spotify.

The music composition attributes do influence the popularity of a song:
- **Danceability**: for one unit of increase, the popularity of a song increases by 6.7116851
- **Speechiness**: for one unit of increase, the popularity of a song decreases by 2.8268618
- **Energy**: for one unit of increase, the popularity of a song decreases by 0.3012474

Therefore, when composing a track, the higher the danceability and the lower the level of speechiness and energy, the track will achieve higher popularity on Spotify, and will therefore be highly recommended to users, which will result in more streams for the song.
As the revenue of an artist on Spotify is based on the number of streams, the more the song is listened to by users, the higher the revenue for the artist and for Spotify.

We have built a linear prediction model for the popularity of a song on Spotify, with no regards for the genre.
However, it is known that artists specialize in specific music genres. Based on this, it would be useful for them to have the same popularity prediction model, but based on their specific genre.
For this reason, and for Spotify to offer an additional service for the artists streaming their tracks on their platform, we are building 3 popularity prediction models, one for each of the top 3 genres: Electro, Pop and HipHop/Rap.

Electro Analysis

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     65.6      13.1      5.01  0.000156
## 2 danceability     9.68     11.7      0.827 0.421
## 3 speechiness      3.66     17.1      0.214 0.833
## 4 energy           3.81      8.60     0.443 0.664
```

From this linear model, we can build a prediction model formula for Electro songs' popularity:

$$ElectroPredictedPopularity = 65.646833 + (9.682327 * danceability) + (3.663718 * speechiness) + (3.813377 * energy)$$

We can conclude that for Electro songs, danceability will have the highest impact on popularity, followed by energy and speechiness.

Pop Analysis

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    70.6        1.49    47.4   3.71e-131
## 2 danceability    7.36       2.01     3.67  2.94e-  4
## 3 speechiness     0.500      3.60     0.139 8.90e-  1
## 4 energy         -0.987      1.46    -0.676 4.99e-  1
```

From this linear model, we can build a prediction model formula for Pop songs' popularity:

$$PopPredictedPopularity = 70.6154946 + (7.3641178 * danceability) + (0.4995130 * speechiness) - (0.9867791 * energy)$$

We can conclude that for Pop songs, an increase in danceability will have the highest impact on popularity, followed by an increase in speechiness. On the other hand, energy needs to be lowered in order to increase the popularity.

HipHop/Rap

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    71.2        3.20    22.2   2.77e-42
## 2 danceability    5.54       3.23     1.71  8.92e- 2
## 3 speechiness    -4.77       2.98    -1.60  1.12e- 1
## 4 energy          0.999      2.80     0.357 7.22e- 1
```

From this linear model, we can build a prediction model formula for HipHop/Rap songs' popularity:

$$RapPredictedPopularity = 71.1713882 + (5.5411240 * danceability) - (4.7735104 * speechiness) + (0.9993428 * energy)$$

We can conclude that for HipHop/Rap songs, an increase in danceability will have the highest impact on popularity, followed by a decrease in speechiness. An increase in energy will also positively increase the popularity.

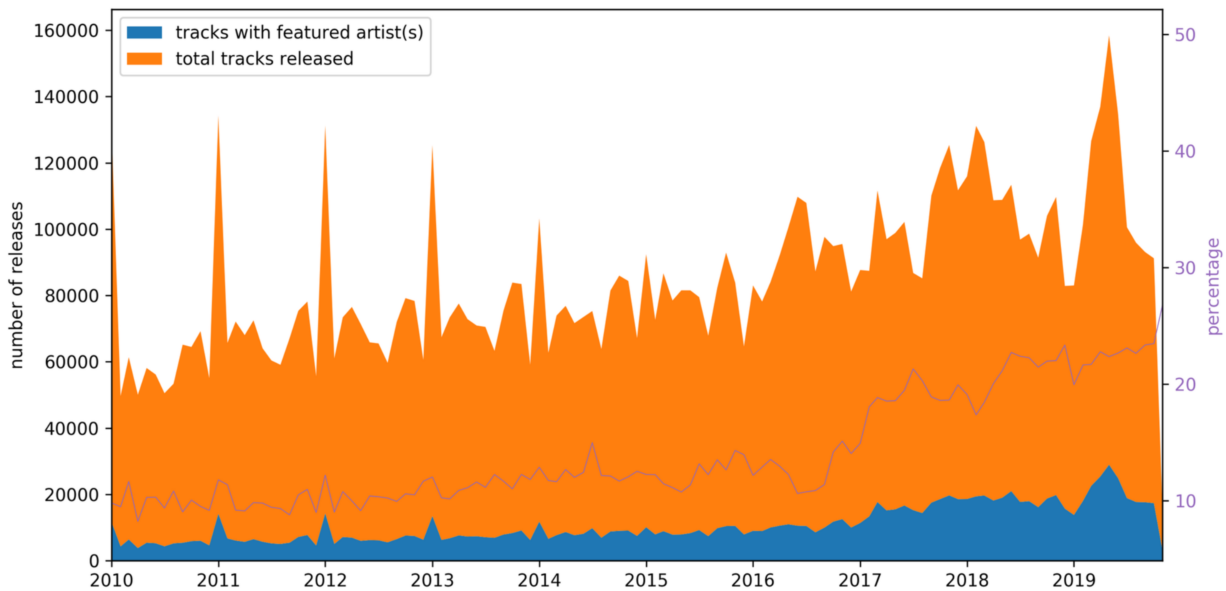
Part V - Takeaways

Based on this analysis, we want to leave you with 3 key takeaways:

-**Potential impact of the model in popularity and revenue in the music industry**. Even though our dataset is small to actually prove the impact of this model at a large scale, we believe it to be a good indicative of what this could mean to artists, to Spotify and to the Music industry as a whole. If this same model could be run with a larger dataset, we could prove that artists could potentially enhance their music to meet a higher popularity standard, increasing the revenue of their songs and reaching more people.

-**Discovering the strongest combination for each artist and potentially using the model to seek into collaborations with other successful artists**. "Collaboration is an artist's best friend in today's music economy. From widening an artist's audience to prolonging success, a rising tide really does lift all boats" is what N. Seekhao wrote for her blogpost in January about music collaborations. It is clear that artists are constantly seeking ways of acquiring higher levels of popularity and this model could also be used to explore, based on music genre, how they could work with their peers to reach these longed-for levels.

– Below, a graph that Seekhao added on her report, showing the huge potential underlying collaborations. Where she analyzed all tracks released on major Western platforms in the past 10 years, aggregated their counts by month then calculating the Collaboration Activity Index (or CAI, as a percentage of collaborations). Based on the data, CAI has doubled in the past decade, with most of the increase happening in the last three years. The purple line represents the CAI, the blue area shows the number of tracks with two or more artists, and the orange area shows the total number of tracks released each month.



-This model could be applied to music in specific countries. The adaptability of this model could allow it to be applied in specific regions, countries and even cities. This gives smaller artists the opportunity to tap untapped markets or to increase their popularity in regions where they already are liked, understanding their target market at a deeper level and potentially customizing their music to a certain extent for their audience.

Part VI - Bibliography

Bibliography Eichler, O. (2020, October 7). Spotify Popularity — A unique insight into the Spotify algorithm and how to influence it. Retrieved from Medium.com: https://medium.com/songstats/spotify-popularity-a-unique-insight-into-the-spotify-algorithm-and-how-to-influence-it-93bb63863ff0 Kim, K. (2015, May 6). The evolution of music: How genres rise and fall over time. Retrieved from Los Angeles Times: https://www.latimes.com/visuals/graphics/la-sci-g-music-evolution-20150505-htmlstory.html Kopf, D. (2019, July 19). Your Spotify and Apple Music subscriptions pay artists you never listen to. Retrieved from qz.com: https://qz.com/1660465/the-way-spotify-and-apple-music-pays-artists-isnt-fair/ #:~:text=The%20way%20Spotify%20and%20Apple%20Music%20pay%20artists%20is%20simple,of%20streams%20each%20a NATH, T. (2020, February 19). How Pandora and Spotify Pay Artists. Retrieved from Investopedia.com: https://www.investopedia.com/articles/personal-finance/121614/how-pandora-and-spotify-pay-artists.asp Nest, T. E. (2013, December 18). How Music Has Evolved in the Past 70 Years. Retrieved from Gizmodo.com: https://gizmodo.com/how-music-has-evolved-in-the-past-70-years-1485770090 SEEKHAO, N. (2020, January 10). How Music Collaborations Evolved in the Digital Era: A Decade in Review. Retrieved from Chartmetric.com: https://blog.chartmetric.com/the-evolving-role-of-music-artist-collaborations/

Stone, J. (2020, December 7). The State of the Music Industry in 2020. Retrieved from Toptal.com: https://www.toptal.com/finance/market-research-analysts/state-of-music-industry Stone, J. (2020, December 7). The State of the Music Industry in 2020. Retrieved from Toptal.com: https://www.toptal.com/finance/market-research-analysts/state-of-music-industry