Creating Value for Music Artists

Exploring Spotify's Song Data

1. Framing the Business Problem

1.1 An Introduction To The Music Industry

The music industry is a booming business. In 2019, the total revenue of the music industry amounted to 21.5 billion U.S. dollars and is predicted to grow to 36.7 billion U.S. dollars by 2025 (IFPI, 2020; Koronios, 2020).

The industry has changed over the last decade because of fast digitalization;

1. **The internet has changed the way people connect with the music they love.**In one massive disruption, physical products such as vinyl, CDs, and MP3-players have become redundant. Streaming services have made a virtually limitless amount of music available to you anywhere, anytime at a single monthly price (Rubin, 2019).
2. **A positive effect of the internet is the rise of new music genres** Over the last decade, international artists have achieved popularity all over the world. Examples of new genres include K-Pop, Latin Pop, and Electronic Dance Music (Haider, 2019).
3. **The rise of the internet made it possible for people to purchase physical music products and upload them illegally for free onto the web.** Music piracy has fallen dramatically due to stricter regulations and prevention methods. This is partially due to the rise of streaming platforms as they offer a wide range of music collections; therefore, the need for unverified sources slowly disappeared (Shepherd, 2018).
4. **The barriers in the music industry are classified as medium-low.**Initially, the barriers to enter the music industry were way higher due to the production costs of physical products. Now with the changing climate, production costs have decreased due to better production software. A few record labels initially dominated the market; however, 3rd-party publishing websites have made it easier for new artists to post their music (Koronios, 2020).

With 341 million paid subscribers, streaming accounted for 56 percent of the total music industry revenues in 2019. The trend started approximately in 2009 and has been growing since. The industry segment is forecasted to reach 1.3 billion users by 2030, of which 21% will be streaming music using their mobile phones (IFPI, 2020; Goldman Sachs, 2020).

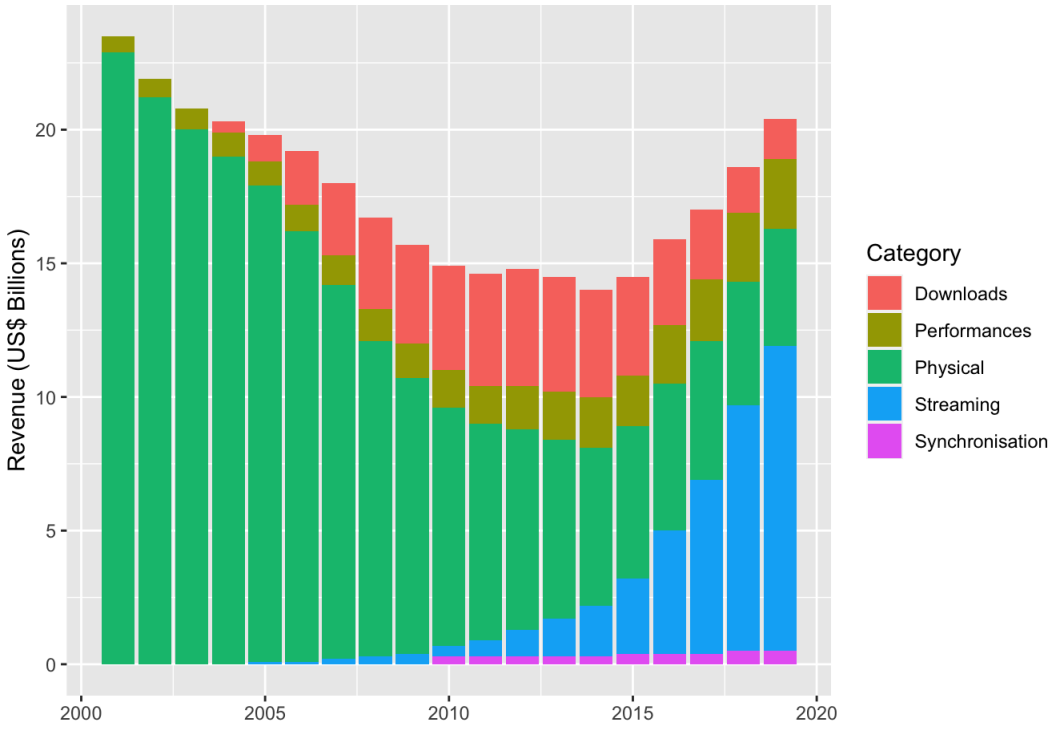

Figure 1: Global Recorded Music Industry Revenues Changes (IFPI, 2020)

1.2 Problems in the Music Industry

Even though streaming services have created a new source of obtaining music, it also led to negative effects on the market.

1. **There has been a decline in physical purchases.** Looking at figure 1, it can be seen that the sales generated from physical products have declined over the last decade. As there are many streaming predictions, physical products are expected to disappear, driving brick-and-mortar music stores to struggle. In today's current client, most physical product sales are vinyl, driven by '80s and '90s nostalgia (Aziz, 2019; Rubin, 2019; Owsinski, 2019).
2. **Artist upset with the current climate of streaming platforms.** Big Artists believe that streaming platforms are destroying album sales. They believe that they generally generate more revenue from selling albums (digitally or physically). One example is a pop-country artist "Taylor Swift", who deleted her content from Spotify. She quoted: *"Valuable things should be paid for"*, referencing her albums (Hassan, 2016; Tiffany, 2017). This is especially a problem for new artists. Record labels have great resources helping the new artist to achieve success. Artists trying to get discovered through the internet without big record labels rely on the revenues generated from online views or streams, which is not sustainable for an income that covers living costs without popularity.

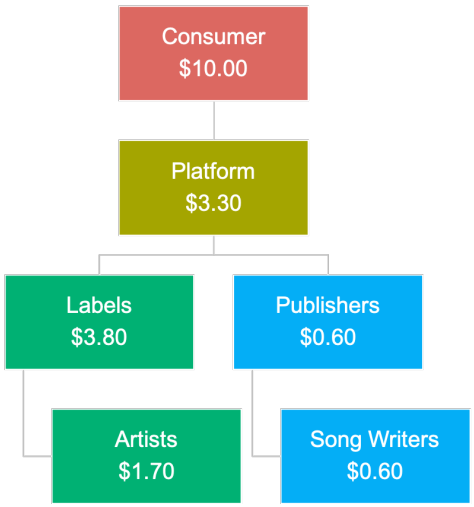

Figure 2: Revenue Streaming Parties and Parties Involved (Goldman Sachs, 2020)

1.3 The Business Problem

The problem of getting discovered as well as getting reasonable pay affects new artists in the music industry. The impact is significant, especially during the pandemic, where they are dependent on digital discovery as there cannot be any physical performances (Brown, 2020; Andor Brodeur, 2020).

Spotify, one of the streaming platforms, has to deal with their stakeholders' dissatisfaction (artists), as they are interested in building good relationships. The company offers a universal solution that gives music listeners access to millions of songs and other content from artists worldwide. At the same time, they unlock the potential of human creativity—by allowing a million creative artists to live off their art (Spotify, nd). Without artists and listeners, the business model would fail.

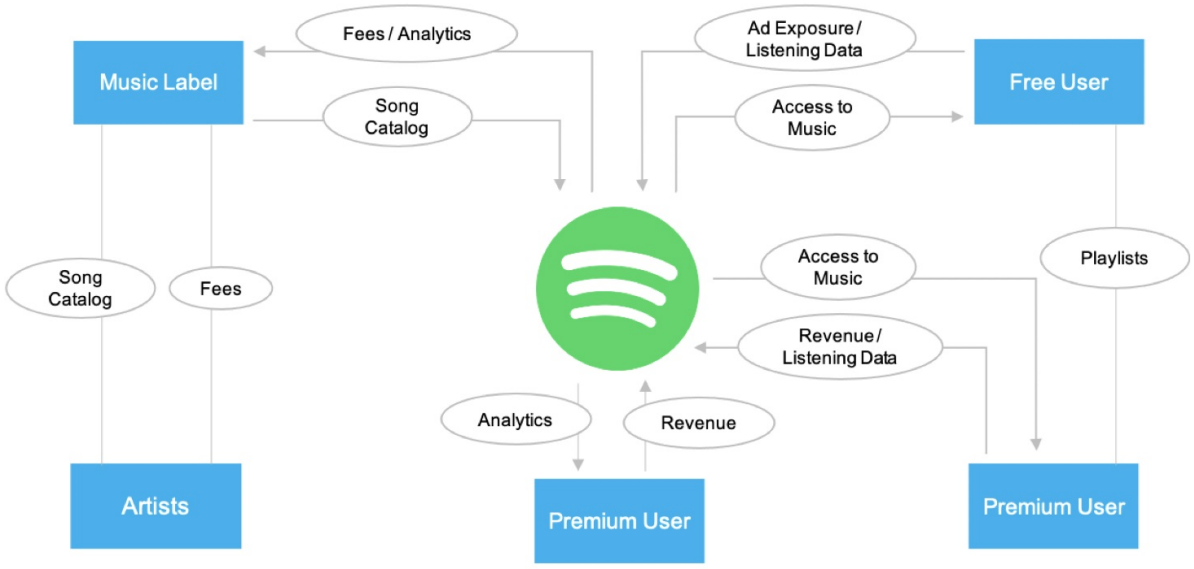

Figure 3: Spotify Business Model (Fox, nd)

The real question is whether Spotify could add more value for (new) artists without changing their current business model and cost structure.

As seen in figure 3, Spotify gathers data on the users' music tastes and preferences. Part of their platform experience, they use the data to show recommendations through Spotify's preprogrammed playlists, weekly and daily mixes, and radar releases in which new songs get released. The number of streams also determines which songs are the most popular and publish this data on Spotify Charts: the daily top 200 and 50 most viral songs. (Perez, 2019; Spotify, 2018; Fox, nd).

Artists get paid whenever a user streams a song for at least 30 seconds. This means that the artist's music needs to be attractive enough for the users to play it and like it. The more streams, the more money an artist will receive (The Planetary Group, 2019).

This has to do with the popularity of the song. The more popular the song will be, the more an artist can earn. Therefore, the artists and music production companies are interested in producing music that goes viral and gets into the top 200 streaming songs.

1.4 Hypothesis

Songs have different characteristics; each genre has different characteristics. Algorithms can easily determine each of these characteristics. These characteristics include the valence, the acousticness, the danceability, the duration of a song, the energy present, whether it was a clean or explicit file, the instrumentalness, the liveness, the loudness, chord progressions, the speechiness, and the tempo of the song (Moris, nd; Yamaç Eren Ay, 2020).

Knowing the following information, the business question would be;

"Could a prediction about the popularity of a song be made based on song characteristics?"

Spotify could check with their data whether a combination of song characteristics influences the popularity of the song. Artists could produce new songs following the right mix of characteristics to gain popularity. More popularity equals more money. This " secret combination recipe" would help Spotify establish a better relationship with record labels and artists and listeners who want to have likable music.

Therefore, the formulated hypothesis for a predictive analysis focusing on new genres is:

Ho: The characteristics (instrumentalness, acousticness, liveness, dancability, engergy, loudness, speechness, valence, tempo, number of streams) of of new music genres are not statistically significantly related to song popularity score.

Ha: The characteristics (instrumentalness, acousticness, liveness, dancability, engergy, loudness, speechness, valence, tempo, number of streams) of new music genres are statistically significantly related to song popularity score.

2. Solving the Problem

2.1 Data Collection

To solve the problem, data about different songs and their popularity is needed. To answer the business question in this specific project, the data from Spotify will be used:

- Music characteristics over the year *(data set: years)*
- Data by genre *(data set: genre)*
- Song Information *(data set: songs)*.

This data was collected from (Yamaç Eren Ay, 2020). The data has been generated through open source web scraping from Spotify's Developers platform. The data contains information from 1971 to 2020 and was updated on November 25, 2020.
From the data set songs and genres, a data set was generated for the most popular genres *(data set: popular_genres)*

Additional information was gathered through Spotify Chart, a webpage where the company presents the top 200 songs. The date chosen was October 31, 2020, to exclude any holiday-oriented songs that might influence the outcome of our results *(data set: top200)*. Whenever it was needed, more data was gathered through additional open source web scraping (Spotify AB, 2020; Spotify, 2020).

```
## [1] "years"
```

1
2
3
4
5
6

6 rows | 1-1 of 15 columns

```
## [1] "genres"
```

1
2
3
4
5
6

6 rows | 1-1 of 17 columns

```
## [1] "songs"
```

1
2
3
4
5
6

6 rows | 1-1 of 18 columns

```
## [1] "popular_genres"
```

1
2
3
4
5
6

6 rows | 1-1 of 18 columns

```
## [1] "top200"
```

1
2
3
4
5
6

6 rows | 1-1 of 20 columns

2.2 Variable Selection

The main variables that are selected to check whether there is an influence on popularity are;

Legend: **n** = numerical **c** = categorical

- n - Valence: A measure from 0.0 (low) to 1.0 (high) describing the musical positiveness conveyed by a track.
- n - Acousticness: A confidence measure from 0.0 (low) to 1.0 (high) of whether the track is acoustic.
- n - Danceability: A measure from 0.0 (low) to 1.0 (high) describing how suitable a track is for dancing.
- n - Duration: The length of the track in milliseconds.
- n - Energy: A measure from 0.0 (low) to 1.0 (high) representing a perceptual measure of intensity and activity.
- n - Instrumentalness: Whether a track has vocals. 0.5 to 1.0 should contain almost no vocal content.
- n - Liveness: indicates whether the song sounds like a live performance (1.0) compared to a studio-recorded track (0.0)
- n - Loudness: Relative loudness of the track in the typical range [-60, 0] in decibel (dB).
- n - Speechiness: track length with voice. 0.0-0.33: music/non-speech, 0.33-0.66: music and speech, 1.0: only speech.

- n - Tempo: The tempo of the track in Beat Per Minute (BPM).
- n - Popularity: A measure from 0 (low) to 100 (high) representing a score based on the total number of plays and how recent these plays are (Spotify AB, 2020).

The categorical variable used for this analysis is:

- c - genres: the genre specified by song or artists.(Spotify AB, 2020).

2.3 Exploratory Analysis

2.3.1 Descriptive Analysis

To analyze the data, the data needs to be understood. We want to look at the features from a each data set by using the following functions:

- The *function (str)* was used to look at the data sets' internal structure. It shows the number of observations, the different variables, and the type of variables.
- The *function (summary)* was used to get a complete overview of our data. This function shows the minimal value, the maximum value, the mean and median of each column. It also gives an overview of the missing values, listed as a number under NA.

These two functions were consistently applied for all data sets. As an example, the descriptive analysis of the data set **"top200"** is presented.

```
#Calling function
str(top200)
```

```
## 'data.frame':    200 obs. of  19 variables:
##  $ position        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ track.name      : chr  "positions" "Mood (feat. iann dior)" "Dakiti" "Lemonade (feat. Gunna, Don Toli
ver & NAV)" ...
##  $ artists         : chr  "Ariana Grande" "24kGoldn" "Bad Bunny" "Internet Money" ...
##  $ streams         : int  6396150 5498059 4483611 4433379 4215780 3833640 3716587 3674200 3626329 3549812
...
##  $ track.id        : chr  "35mvY5S1H3J2QZyna3TFe0" "3tjFYV6RSFtuktYl3ZtYcq" "47EiUVwUp4C9fGccaPuUCS" "7hx
HWCCAIIxFLCzvDgnQHX" ...
##  $ valence         : num  0.682 0.756 0.145 0.462 0.485 0.334 0.357 0.558 0.737 0.543 ...
##  $ year            : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ acousticness    : num  0.468 0.221 0.401 0.25 0.237 0.00146 0.0194 0.185 0.0112 0.65 ...
##  $ danceability    : num  0.737 0.7 0.731 0.8 0.83 0.514 0.935 0.783 0.746 0.709 ...
##  $ duration_ms     : int  172325 140526 205090 195429 173711 200040 187541 199112 199054 160000 ...
##  $ energy          : num  0.802 0.722 0.573 0.658 0.585 0.73 0.454 0.727 0.765 0.548 ...
##  $ instrumentalness: num  0.00 0.00 5.22e-05 0.00 0.00 9.54e-05 0.00 2.46e-05 0.00 1.59e-06 ...
##  $ liveness        : num  0.0931 0.272 0.113 0.111 0.248 0.0897 0.0824 0.0626 0.0936 0.133 ...
##  $ loudness        : num  -4.77 -3.56 -10.06 -6.14 -6.48 ...
##  $ popularity      : int  96 99 100 90 92 96 96 89 97 96 ...
##  $ release_date    : Factor w/ 11244 levels "1921","1921-02-20",..: 11226 11154 11226 11179 11226 11074 1
1164 11174 11179 11141 ...
##  $ speechiness     : num  0.0878 0.0369 0.0544 0.079 0.094 0.0598 0.375 0.389 0.0993 0.353 ...
##  $ tempo           : num  144 91 110 140 110 ...
##  $ genres          : Factor w/ 10743 levels "['21st century classical', 'american 21st century classical
', 'choral', 'contemporary choir']",..: 10095 8330 9279 10059 10095 4353 10087 10280 9017 4103 ...
```

```
#Calling function
summary(top200)
```

```
##     position       track.name          artists             streams
##  Min.   :  1.00   Length:200         Length:200         Min.   : 756714
##  1st Qu.: 50.75   Class :character   Class :character   1st Qu.: 893627
##  Median :100.50   Mode  :character   Mode  :character   Median :1154320
##  Mean   :100.50                                         Mean   :1474848
##  3rd Qu.:150.25                                         3rd Qu.:1817540
##  Max.   :200.00                                         Max.   :6396150
##
##    track.id            valence            year       acousticness
##  Length:200         Min.   :0.0592   Min.   :1962   Min.   :0.00115
##  Class :character   1st Qu.:0.3620   1st Qu.:2019   1st Qu.:0.05545
##  Mode  :character   Median :0.5530   Median :2020   Median :0.16200
##                     Mean   :0.5418   Mean   :2017   Mean   :0.23029
##                     3rd Qu.:0.7370   3rd Qu.:2020   3rd Qu.:0.31300
##                     Max.   :0.9630   Max.   :2020   Max.   :0.93400
##                     NA's   :5        NA's   :5      NA's   :5
##   danceability     duration_ms         energy       instrumentalness
##  Min.   :0.3070   Min.   :119133   Min.   :0.2250   Min.   :0.000000
##  1st Qu.:0.6415   1st Qu.:173522   1st Qu.:0.5355   1st Qu.:0.000000
##  Median :0.7310   Median :198570   Median :0.6470   Median :0.000000
##  Mean   :0.7042   Mean   :201437   Mean   :0.6357   Mean   :0.006696
##  3rd Qu.:0.7935   3rd Qu.:218975   3rd Qu.:0.7415   3rd Qu.:0.000030
##  Max.   :0.9350   Max.   :357267   Max.   :0.9700   Max.   :0.657000
##  NA's   :5        NA's   :5        NA's   :5        NA's   :5
##     liveness         loudness         popularity         release_date
##  Min.   :0.0385   Min.   :-14.454   Min.   : 59.00   2020-10-30: 18
##  1st Qu.:0.0931   1st Qu.: -7.252   1st Qu.: 83.50   2020-10-02:  8
##  Median :0.1260   Median : -5.655   Median : 86.00   2020-07-03:  7
##  Mean   :0.1745   Mean   : -6.081   Mean   : 85.78   2020-09-18:  6
##  3rd Qu.:0.2235   3rd Qu.: -4.489   3rd Qu.: 89.00   2020-07-10:  5
##  Max.   :0.9140   Max.   : -2.474   Max.   :100.00   (Other)   :151
##  NA's   :5        NA's   :5         NA's   :5        NA's      :  5
##   speechiness         tempo                                  genres
##  Min.   :0.0232   Min.   : 68.48   ['pop', 'post-teen pop']      : 21
##  1st Qu.:0.0478   1st Qu.:100.10   ['pop', 'uk pop']             :  8
##  Median :0.0770   Median :122.03   ['pop']                       :  8
##  Mean   :0.1187   Mean   :123.73   ['brooklyn drill']            :  7
##  3rd Qu.:0.1610   3rd Qu.:143.87   ['chicago rap', 'melodic rap']:  6
##  Max.   :0.4870   Max.   :180.07   (Other)                       :145
##  NA's   :5        NA's   :5        NA's                          :  5
```

Insights & Decisions

Missing Data

Based on the data sets' statistics, it can be concluded that most missing values are associated with the genre.

- data set: **songs** - The genre column in the *songs* dataset is associated with the artists specified in the *genres* data set. If the main artist is missing in *genres* data set, it was not transferred over to the songs data set. Some artists have been classified with "[]," which means that no genre was assigned. There are 24.769 missing values for the genre column. These values are excluded whenever genres in this data set are analyzed.
- data set: **top200** - The data has five missing rows.These rows did not have matching song id's when comparing it with the *songs* data set. The characteristics information was also not available in Spotify's Open-Source Developers program. One row was left on-purposely left out since it was considered a Halloween holiday-song. The five rows with missing values are excluded from the analysis.

```
## [1] " Dua Lipa album result from Spotify's Open Source"
```

NA

1 row | 1-1 of 7 columns

Changing Data Types

- The variables *track.name and name* have been changed from a *factor* to *characters*.
- The variables *artists* have been changed from a *factor* to *characters*.
- The variables *release_date* have been changed from a *factor* to a *character.date* format.

2.3.2 Exploratory Questions

To understand the data sets, it is needed to look deeper into popularity, song characteristics, and the different genres. This was done in a question format;

1. How have the characteristics of music changed over the last ten years?
Over the last years, technology has allowed music to change with new sounds and instruments. Taste is also something that changes over the years. It is essential to analyze the changing flow of music characteristics over the past decade to get a feeling for what song characteristics could define the characteristics of music and tastes of today's climate.

2. Are they any correlations between the variables?
Our data sets mention multiple song characteristics. Knowing the different linear relationships between them is helpful when analyzing the data and creating the associated model. It is essential to check how strong the correlation is because it defines which variables are suitable for future regression about popularity.

3. What does it take to be in the top 200?
Is it essential to verify whether the data gathered matches the statements made in the articles. The hypothesis was created based on those articles and need to be confirmed to proceed. Testing genres from perspectives are significant to our entire analysis.

4. What defines the popular genres suggested by BBC?
Is it essential to verify whether the data gathered matches the statements made in the articles. The hypothesis was created based on those articles and need to be confirmed to proceed.

Question 1: How have the characteristics of music changed over the last ten years?

From the line-chart of characteristic changing flow, we can conclude that:

- There are two characteristics: *Speechiness* and *Danceability*, keeping on increasing over the last decade.
- The *Instrumentalness* has a clear trend of decreasing.
- There are four characteristics that are fluctuating with changes of 0.1.
    - Both *Energy* and *Valence* show a positive change after 2016 to 2017.
    - *Acousticness* and *Energy* have a trend showing a significant decrease within the range of 0.1 change during the year of 2018 to 2019.
- *Liveness* is the only characteristic to be stable over the past decade.

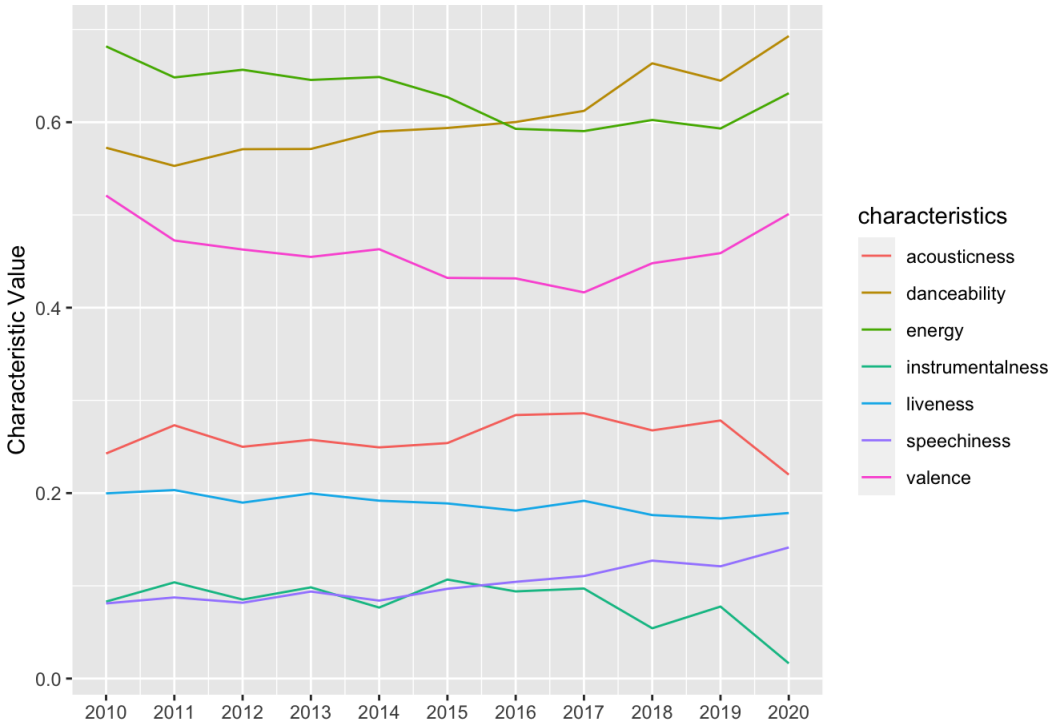

Figure 4: Characteristics Changes over the last ten years

Question 2: Are they any correlations between the variables?

With a 95% of confident interval that the correlation chart in figure 5 shows that;

- A strong positive correlation between *Energy* and *Loudness* (+0.78)

- A moderate positive correlation between *Danceability* and *Valence* (+0.56)
- A moderate positive correlation between *Popularity* and *Energy* (+0.49)
- A moderate positive correlation between *Popularity* and *Loudness* (+0.46)
- A strong negative correlation between *Acousticness* and *Energy* (-0.75)
- A moderate negative correlation between *Acousticness* and *Loudness* (-0.56)
- A moderate negative correlation between *Popularity* and *Acousticness* (-0.57)

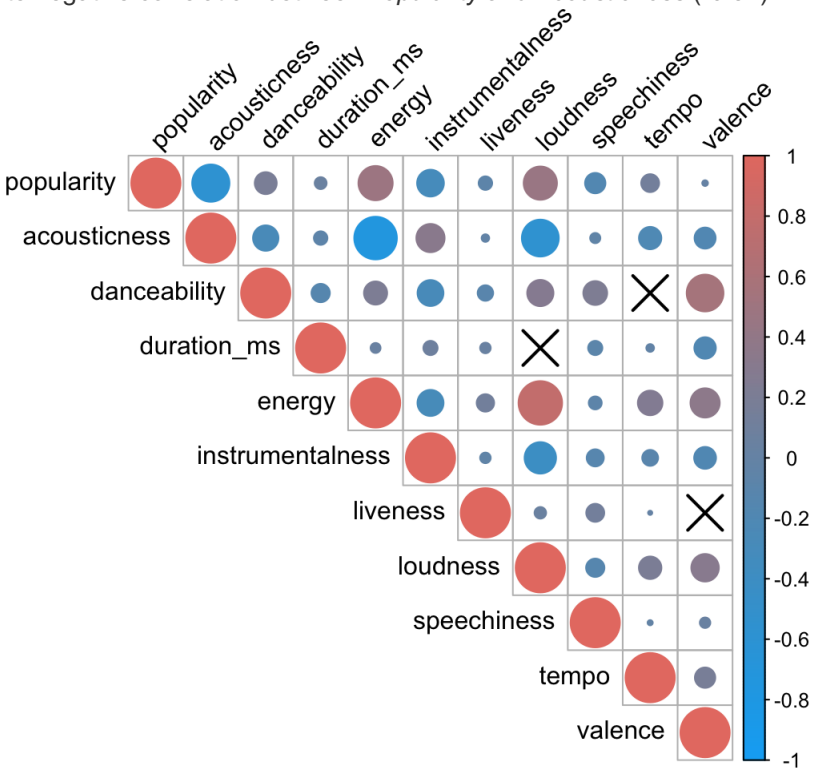

Figure 5: Correlation matrix song characteristics

Question 3: What does it take for a song to be in the top 200?

As was mentioned before, the top200 is defined by the number of streams in a given period. Knowing this, there should be a relationship between the popularity score and the total number of streams. It is assumed that a high number of streams should have a high popularity score. However, this is not the case.

Figure 6 shows that it contradicts our assumption. The plot shows that there are quite some**variations** in the top200 popularity score. With a threshold of a popularity score of 85, there are 41 outliers, which is approximately 20.5% of all data points.

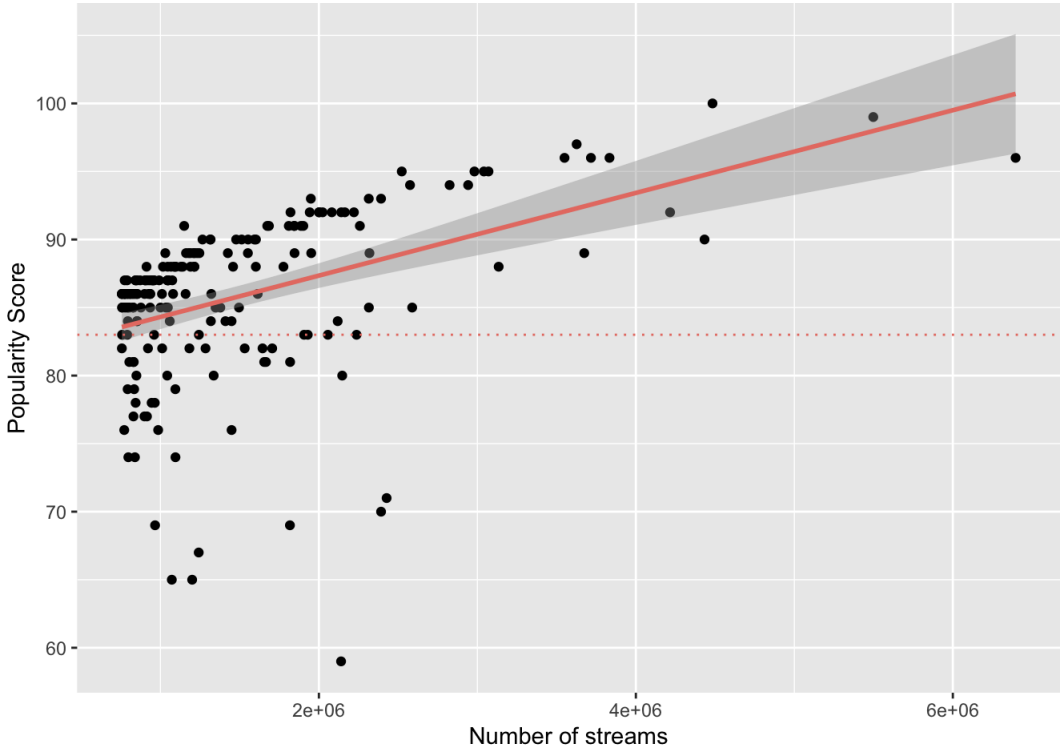

Figure 6: Scatterplot with the number of streams over popularity

A simple linear model reflects a **linear relationship**; however, only 18% of all streaming numbers account for the popularity score.

```
## [1] "Simple Linear Regression Summary:"
```

```
##
## Call:
## lm(formula = popularity ~ streams, data = top200)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.773  -2.528   2.145   3.520   6.236
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.127e+01  7.720e-01 105.276  < 2e-16 ***
## streams     3.038e-06  4.460e-07   6.812  1.2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.548 on 193 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.1938, Adjusted R-squared:  0.1896
## F-statistic:  46.4 on 1 and 193 DF,  p-value: 1.195e-10
```

Knowing that the popularity score contains many variations, it was needed to look at other characteristics that would define why a song would be at the top200.

There are some outliers in the data set that have different characteristics. These represent the red dots in the graph in figure 7. As the data is presented as a density graph, it can be seen that the majority of the songs represent either more skewed data or the bigger chunks in the middle. This indicates that each characteristic graph's highest point represents the characteristic score of the top 200 songs. This means that the most common characteristics for the top 200 songs are:

- Low acousticness score
- Medium to high loudness score
- Low speechness score
- Low liveness score
- Medium to high danceability score

The warning presented is due to the five missing values.

```
## Warning: Removed 40 rows containing non-finite values (stat_density).
```

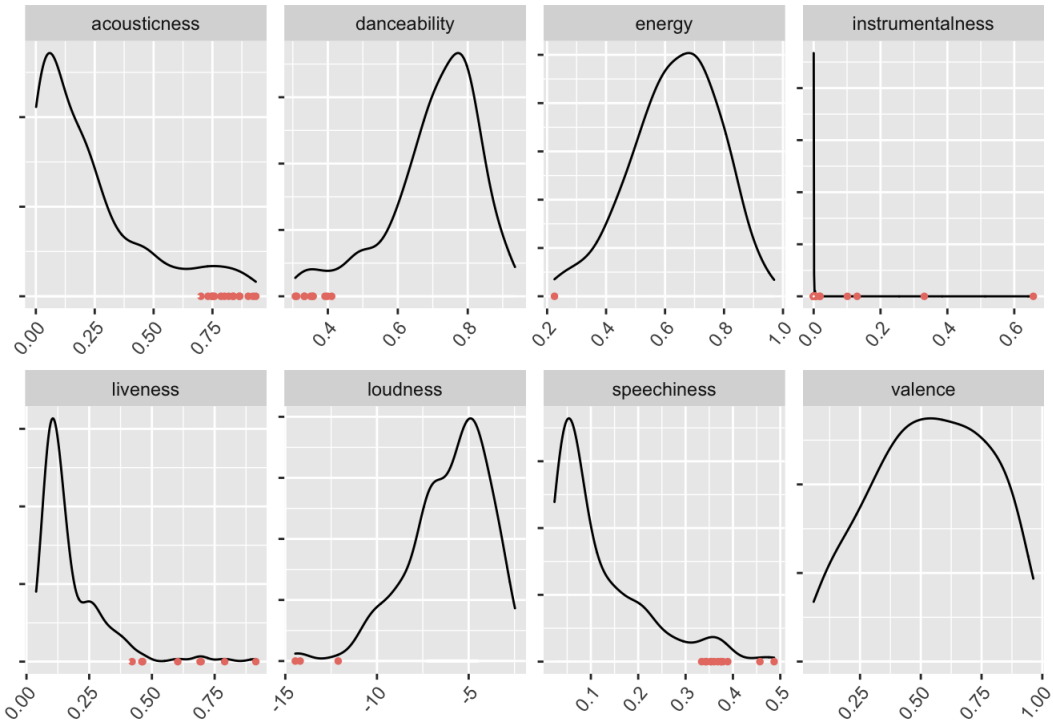

Figure 7: Density plot showing smoothened count of song characteristics

Therefore, the following can be concluded:

- Songs do not need to have a high score to be in the top 200 list. The top 200 only reflects a specific day, while the popularity score is calculated over time.
- Common top200 characteristics on October 31, 2020, included acousticness, loudness, speechness, liveness, and danceability.
- The summary statistics of the dataset songs shows that the 75th percentile of the popularity score is approximately 48. Following the similar song characteristics, songs above 75th percentile have a chance to be in the top 200 list. Nevertheless, our top 200 list is the random sample, therefore; when generating the data from another random date, the popularity scores might have different variations.

the songs with the popularity score which is at list more than 75% percentile of entire dataset, following with the same pattern of different characteristics have higher opportunity to hit the top 200 ranking list.

Question 4: Are the popularity statements from BBC about the genres K-pop, EDM, Latin Pop, and African accurate?

First, we need to look at different outliers. Figure 8 shows that there are some outliers present. It also shows that some genres have more songs than others in the data set, such as African music followed by hip hop and Latin American music. It can be stated based on the plot that these genres generally have higher popularity scores:

- K-pop
- Electronic Dance Music
- Hip Hop

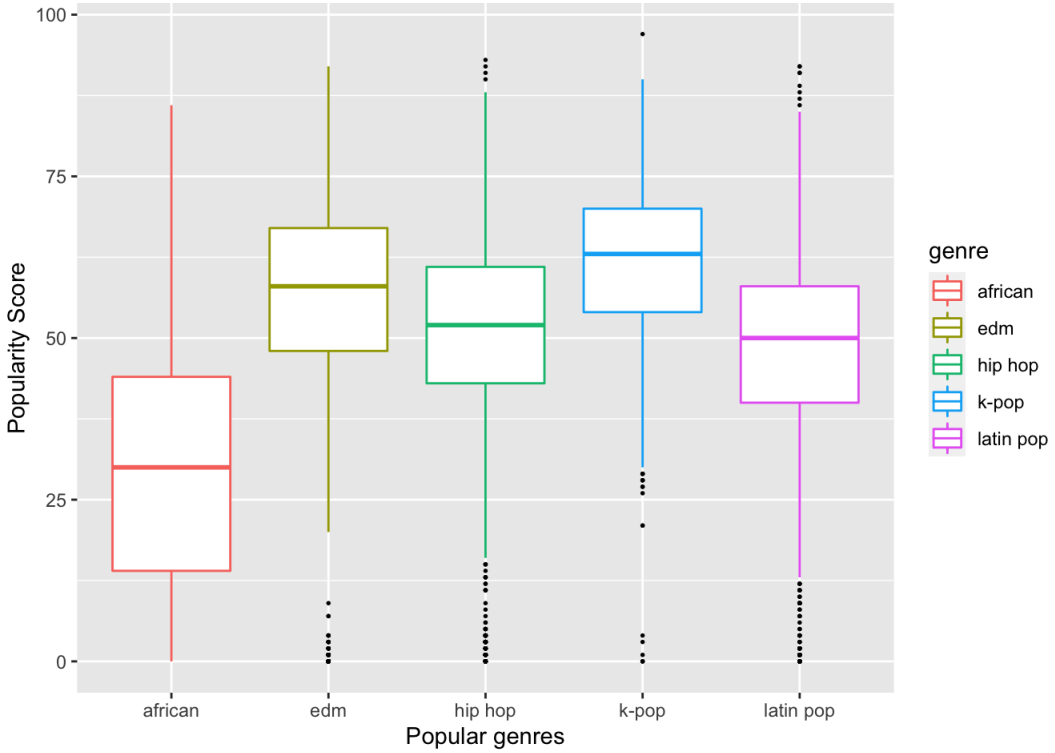

Figure 8: Boxplot for popularity score by genres

The genres K-pop, electronic dance music, and hip-hop are more popular than others due to the shared pattern in characteristics. Figure 9 shows different patterns for each song characteristic. The shared characteristics include;

- Energy
- Danceability
- Valence
- Loudness
- Somewhat for tempo with the exception for K-pop

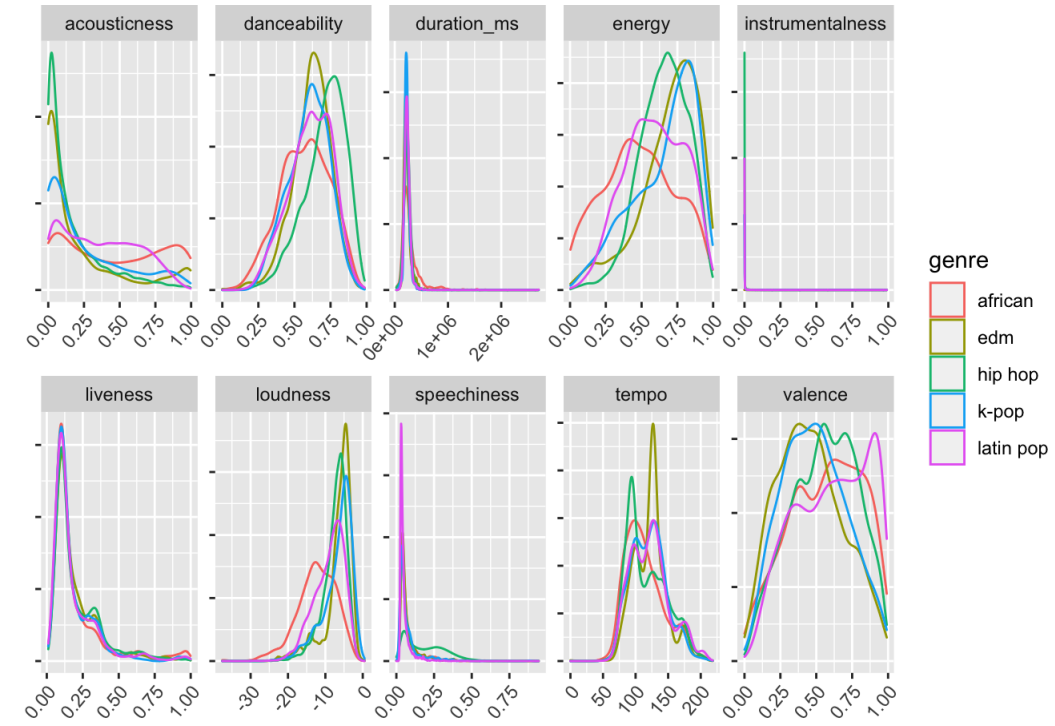

Figure 9: Density plot showing smoothened count of song characteristics by genre

2.3.3 Conclusion

Based on the exploratory data analysis, the following can be concluded;

- The BBC claims that K-pop, EDM, hip-hop, African, and Latin pop are popular growing genres. K-pop, EDM, and hip-hop stand apart from the other two genres in their similarity and popularity.
- The top 200 songs list reflects songs with the most streams on any given day rather than songs with the highest popularity score. The fact that any song with similar characteristics can be in the top 200 suggests that it may be better for an artist to focus on maintaining a good popularity score rather than aiming for the top 200 list. * Analysis of the top 200 songs and the popular genres points towards a trend with songs becoming more danceable, containing more speech, and being less instrumental.
- Energetic and positive songs rebounded in 2018 after receding in 2010. The top 200 contains a mix of positive/negative and low/high energy songs.
- Acoustic and live music seems to have relatively stable popularity, although looking to the top genres, there has been a decrease in live and acoustic music.
- The correlation matrix indicated that there are some positive and negative linear relationships. As is known, the matrix does not show us any causational relationship, but we do know from our analysis that popularity has some patterns with characteristics such as energy, acoustic, loudness, danceability, instrumentals, liveness, speech, and tempo.

2.4 Inital Hypothesis Adjustments

Based on the conclusion of the exploratory analysis, hypothesis for our predittive model is changed;

- The top 200 songs do not necessarily define the popularity score as the number of streams only reflect the popularity given a certain data. Therefore, the indicator of the number of streams will not be included in our model as a characteristic. This top 200 sample data is disregarded from the analysis.
- The African genres as well the latin american pop genre are not as popular as claimed when looking at the popularity score. To simplify the analysis for modelling, it is decided to exclude these genres.
- When looking at certain characteristics: as durartion of the songs are similar, this characterisc is not included. duration_ms.

Therefore, the adjusted hypothesis for a predictive analysis focusing is:
Ho: The characteristics (instrumentalness, acousticness, liveness, dancability, engergy, loudness, speechness, valence, tempo) of the genres “K-pop”, “Hip Hop”, and “Electronic Dance Music” are not statistically significantly related to song popularity score.

Ha: The characteristics (acousticness, liveness, dancability, engergy, loudness, speechness, valence, tempo) of the genres “K-pop”, “Hip Hop”, and “Electronic Dance Music” are statistically significantly related to song popularity score.

3.Modelling & Communication

3.1 Model Building

In order to test our hypothesis, it is needed to explore multiple regression models to find the best one that fits. * 60% of the data goes into a training set, fitting different models. * 20% of the data is used into a query set to compare models *(query_data)*. * 20% of the data was reserved to test the finalized model *(test_data)*.

There are many different models (with different families). One example is `lm(),` better known as a **linear model**, following the rules of *Ordinary Least Squares*. The function takes the individual data points generated from two+ variables and draws the best line through them to minimize the differences between the observed values and predicted value.

Our model should have a confidence level of 95%, which indicates that p-values should be smaller than 0.05.

To determine the fit of the model, the scatterplot matrix and the distribution of each characteristic needs be checked;

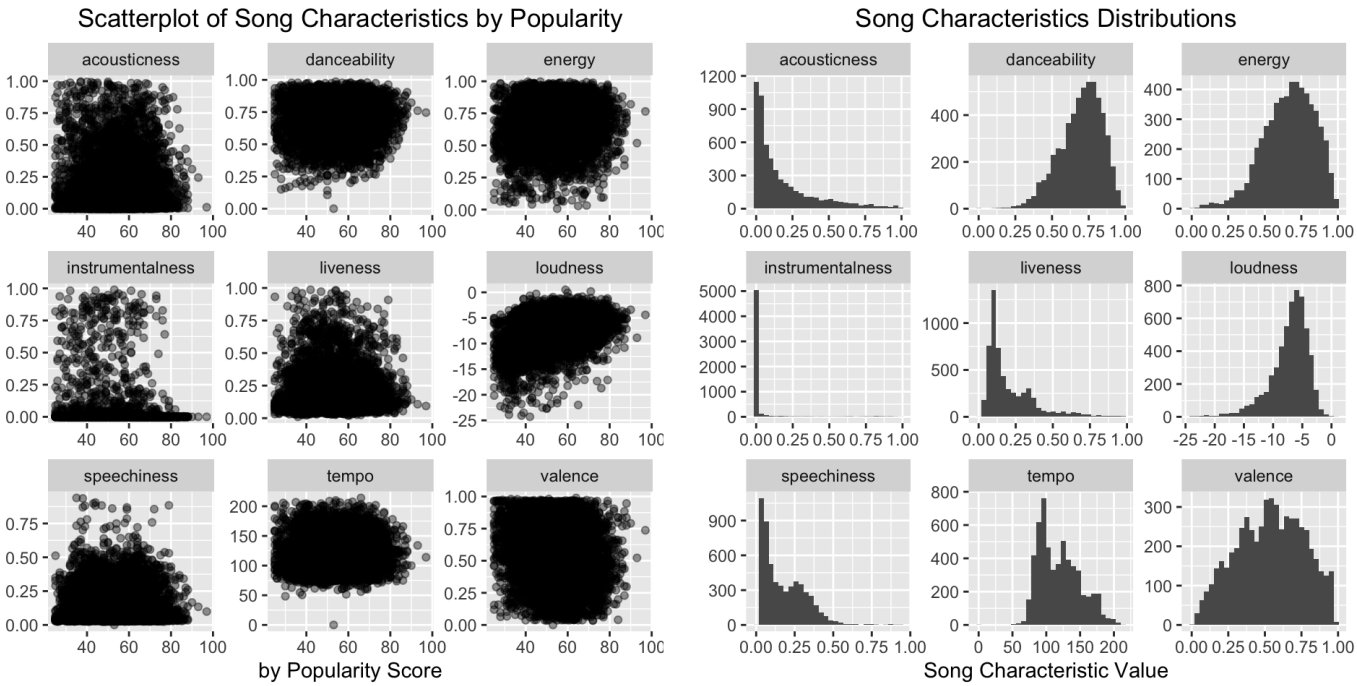

Figure 10: Density plot showing smoothened count of song characteristics by genre

As shown in figure 10, popularity does not have any clear linear patterns, indicating that the relationship between our dependent and independent variables is mostly non-linear. This is also shown in the correlation matrix presented in figure 5. Some correlations are related to popularity; however, these are not substantial. This has to be taken into account when fitting the best model.

The approach that was taken was to test all possible combinations of all the characteristics. This had let to the creation of 511 different combinations. For all of these combinations, it was needed to test four different datasets: the training data and the training data set split by each genre.

Do not worry; we did not go through all 2044 different models! R helped us through functions and loops to determine which12 models (3 of each data set) were the best: * *Training Data Set:* Model 1, Model 2, Model 3 * *EDM Data:* Model 4, Model 5, Model 6 * *K-Pop Data:* Model 7, Model 8, Model 9 * *Hip Hop Data:* Model 10, Model 11, Model 12

model <chr>
model_1
model_2
model_3
model_4
model_5
model_6
model_7
model_8
model_9
model_10

1-10 of 12 rows | 1-1 of 8 columns Previous **1** 2 Next

The statistics of the linear regressions were gathered and were put into an overview. The statistics include:

- Adjusted R Squared to check how much variance in the popularity score can be explained by the characteristics. Here we filtered it to be the highest number.
- R Squared, to compare the score with the adjusted R square.
- F-statistic, to check whether the model is higher than the number 1.
- F-statistic P-value to check whether the model meets the significance level of 0.05, therefore knowing whether the model fits the data.

- Coefficients, to check whether all coefficients in the model meet the significance level.

The coefficient column shows whether one or more variable coefficients are rejected. This means that some variables did not meet the significance level. When it says accepted, it indicates that all coefficients met the significance level.

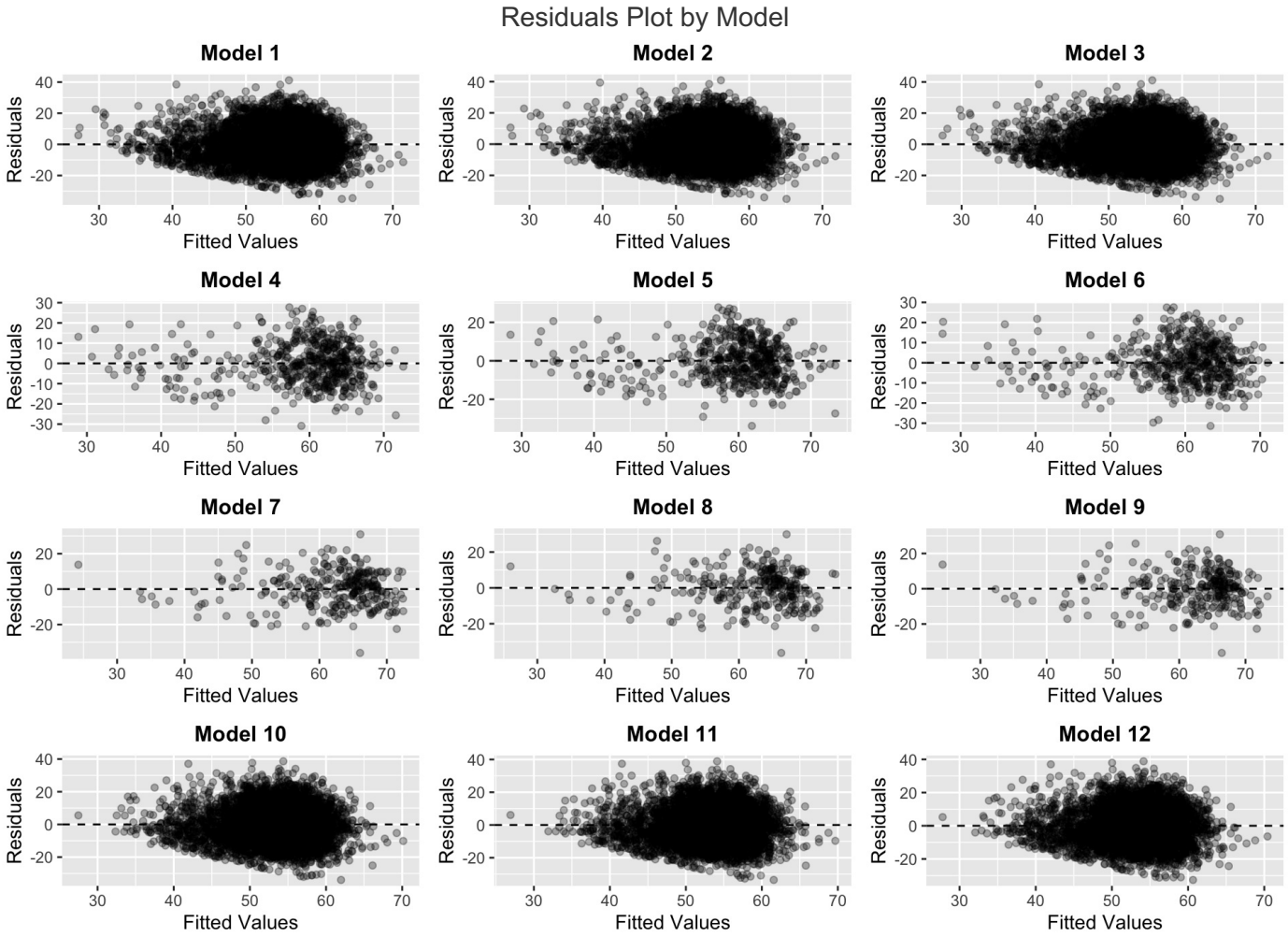

Figure 11: Residual Plot for each model

The residual plots in figure 11 show no unwanted patterns, so the regression coefficients and other numeric results can be trusted. This indicates that the residuals are consistent with random error, which meets an OLS linear regression model's requirement.

3.2 Comparing Models

To compare the different models, the 12 chosen models were rerun with the query data. These models should confirm whether the model is strong enough and give similar results to those created with the training data. These are the results:

model <chr>
model_1_1
model_2_1
model_3_1
model_4_1
model_5_1
model_6_1
model_7_1
model_8_1
model_9_1
model_10_1

1-10 of 12 rows | 1-1 of 8 columns Previous 1 2 Next

As the query data only present 20% of the total data set, the models' strength with the different variables can be determined.

When the sample size decreases, the variation around the means will grow while the mean itself will have little variance due to the Central limit theorem. The sampling distribution of the regression coefficients is normal and centered. As the sample size decreased, the F-statistics decreased as well. Due to this change, the relationships between the characteristics and the popularity score appear less strong due to p-values not meeting the significance level.

As can be seen, two models are accepted with the query data: model 5 and model 11. Both models relate to the specific genres:

- Model 5 indicates that danceability, energy, and loudness influence Electronic Dance Music's popularity score.
- Model 11 has more characteristics that influence hip hop's popularity score; the only characteristic that is not included is speechiness.

Both models do not indicate a high r-square score, but it is decided to pick one as the best-representing model. The r-squared number of model 5 is more considerable than model 11, but there is also a noteworthy difference in the coefficient numbers. The higher the absolute value of the beta coefficient, the stronger the effect; therefore, model 5 substantially influences the popularity score paralleling model 11.

This has led to the decision to choose model 5 as the best-representing model(Statistics How To, 2016).

3.3 Testing Final Model

As a final test, model 5 was rerun with a test data set to validate this model's accuracy. The results and interpretation are shown below. The coefficients p-values are accepted, and the p-value of the f-statistics meets the requirements. Therefore, this is considered to be the best model.

r.squared <dbl>
0.2061116

1 row | 1-1 of 7 columns

Looking at the results for the last model, it can be concluded that this is the only model with the highest adj.r.squared that returns with significant coefficients:

- Approximately 20% of the variance in the popularity score of the genre Electronic Dance Music can be explained by the characteristics; danceability, energy, and loudness.

The sigma indicates that each popularity score could be off with 11 popularity score points, which is relatively high. The popularity score of this genre can be predicted with the following equation:

```
## Popularity Score =  74.98355 + ( 19.35758  * danceability ) + ( -20.35282 * energy ) + ( 2.066039 * loud
ness )
```

On average:

- A one unit increase in danceability is associated with a value change of 19.35758 in the popularity score.
- A one unit increase in energy is associated with a value change of -20.35282 in the popularity score.
- A one unit increase in loudness is associated with a value change of -2.066039 in the popularity score.

4. Conclusion

4.1 Hypothesis Validation

Based on the different models chosen, it can be concluded that the null hypothesis cannot be rejected; therefore, the alternative hypothesis gets rejected. There are no models that include all song characteristics for K-pop, Electronic Dance Music, and Hip Hop. The percentage of variance that can be explained in these models' popularity score is too small to be accurate.

Ho: The characteristics (instrumentalness, acousticness, liveness, danceability, energy, loudness, speechness, valence, tempo) of the genres "K-pop," "Hip Hop," and "Electronic Dance Music" are not statistically significantly related to song popularity score.

Ha: The characteristics (acousticness, liveness, dancability, energy, loudness, speechness, valence, tempo) of the genres "K-pop," "Hip Hop," and "Electronic Dance Music" are statistically significantly related to song popularity score.

However, an adjusted version of the null hypothesis can be rejected to some extend with the final model;

Ho: The characteristics (danceability, energy, loudness) of the genre "Electronic Dance Music" are not statistically significantly related to song popularity score.

Ha: The characteristics (danceability, energy, loudness) of the genre "Electronic Dance Music" are statistically significantly related to song popularity score.

The final model's statistics show that the f-statistic p-value and the coefficients' p-values meet the significance level of 5%. This model can explain 20% of the variance; however, it would still not accurately reflect the popularity score due to the high standard error.

4.2 Insights

When looking at the Exploratory Analysis and the modelling, there are a few insights that can be drawn from the data:

1. Popular music genres do no necessarily have simular characteristics

- Looking at our modelling data, each genre K-Pop, Electronic Dance Music, and Hip Hop, correspond to models with different characteristics (even when this could only explain a small percentage in the popularity score variance). For instance, EDM tends to have higher rhythms and power, closely related to the loudness, danceability, and energy due to the higher beats per minute (Aden Russell, nd). A combination of these characteristics makes total sense for this genre.
- This is also one reason why it might be difficult to predict one concrete popularity score as there are many variations in genres. The genre data set shows that songs do not necessarily fall under one genre but often fall under more than 3+ genres.

2. Other factors might influence the popularity score

- As can be seen in the regressions data, the percentage of variance is relatively low. This could also be since many other factors are involved. When analyzing the top200 data, the popularity score did not necessarily show high numbers, concluding that the top200 just indicates a moment in time.
- There might be another reason why songs are at the top200 other than specific song characteristics, artists familiarity, and artist popularity. Two research studies, one by Stanford University's Computer Science department and one by Duke University's Statistical Science department, show that artist familiarity and artist hotness are essential factors in predicting the popularity of a song (Pham, Kyauk, & Park, 2015; Shapiro, n.d.)

3. Spotify can use its data to identify trends rather than predicting the popularity score

- When looking at the different regressions, it can be stated that the song characteristics in every model can only explain a small variance for the predicted popularity score regardless of the genre. This means that an accurate prediction will be hard to achieve, even when the model is statistically accepted or meet the significance level.
- This means that Spotify cannot deliver value to their stakeholders' artists by providing a secret formula for songs' success based on the popularity score. On the other hand, Spotify could recommend different characteristics that might be popular for a period if there is a curve in music taste, something that was seen in the exploratory analysis. This indicates that Spotify might have to explore options to make changes to meet this stakeholder's satisfaction.

4.3 Further Research

As there are no strong linear patterns for this dataset, which can be seen in the scatterplots and correlation matrix, there are plenty of possibilities to move further with this research. Different genres could be analyzed to see whether more vital models can be created to predict the popularity score.

Another possibility is to apply a transformation to the popularity score or song characteristics to make the data more fitted to linear regression. There might also be better predictive (machine learning) models that fit the data better than a linear model. These might give us different insights.