Unveiling the Secrets of SPOTIFY Songs' Popularity

Linglong Wang, Kexin Hu, Teresa Dang, Linh Pham, Riley Yue

University of Southern California

DSO 545: Statistical Computing and Data Visualization

Professor Francis Pereira

December 2, 2023

Background information

The music industry is a lucrative sector within the entertainment industry, generating billions of dollars annually. A prevailing trend within the music industry is the shift towards streaming platforms and streaming music has experienced a significant increase in revenue in recent years. According to reports on Global music production and distribution (Malley, 2023), streaming music accounts for approximately 70.5% of the overall market share and it is expected to grow over the years. This substantial growth can be attributed, in large part, to the impact of the COVID-19 pandemic while the cancellations of live performances and the unavailability of physical albums are prevalent. Consumers have altered their purchasing behaviors by moving toward streaming music. The convenience and accessibility offered by streaming services accelerate this shift and provide a comprehensive music listening experience through monthly subscriptions. Spotify, as a major player in the streaming music market, experienced a significant surge in subscriptions during the years 2020 and 2021 (Malley, 2023).

We believe this is a promising business opportunity when combined with the benefits of data analytics. With the wealth of data available on Spotify, there is immense revenue potential to utilize business analytics to help music producers adapt to this evolving landscape, understand consumer preferences through track popularity, guide in the song-making process, enhance the overall listening experience for consumers, and ultimately increase customer loyalty and subscription. In our research, we want to understand the specific factors influencing song popularity on Spotify.

Our dataset consists of 114,000 Spotify tracks over a range of 125 different genres. Each row represents a unique track in the dataset. There are 20 audio feature variables assigned to each track. Nevertheless, not all of the audio features are relevant to our research. The variables of interest in this dataset include **popularity**, **track_genre, explicit, danceability, loudness, instrumentalness, valence, and tempo**. **Track_genre** is the genre each track belongs to. **Popularity** (0-100) is calculated by an algorithm based on the track's total number of plays and their recency. Higher values indicate greater popularity. The **explicit** attribute indicates whether the track has explicit lyrics. While "true" represents explicit content, "false" indicates no explicit content, or "unknown" when the information is unavailable. **Danceability** (0-1) reflects how a track is suitable for dancing, based on factors like tempo, rhythm stability, beat strength, and overall regularity with a higher score for more danceability. **Loudness** indicates the overall volume of the track in decibels (dB). **Instrumentalness** (0-1) predicts whether a track contains more instruments and no vocals (1). **Valence** measures the musical positiveness conveyed by a track, ranging from 0.0 (negative) to 1.0 (positive). **Tempo** provides an estimated beats per minute (BPM) for the track's overall tempo**.**

Preliminary findings

Several researchers in the past have tried to understand important features and their association to track popularity. A representative study is TPD (Track Popularity Dataset) in the Hit Song Science of Music Information Research field (Karydis, 2016). As of 2016, it is difficult to collect track popularity and features from different sources on a large scale because there is no online synchronized track dataset and no single dominant platform to standardize track popularity ratings. The absence of an integral dataset has hindered Hit Song Science for information mining purposes. Furthermore, an obstacle is how to quantify the qualities that contribute to track popularity and whether it is appropriate to use quantifiable features to understand and predict subjective aesthetics. Consequently, they introduce the TPD to alleviate the problems and it becomes the first complete attempt for a manageable dataset in the field.

The boom of manageable data nowadays has largely solved the problems above and improved our ability to use more complex models for data analytics. For example, another research focuses on machine learning to predict hit songs using Spotify data (Georgieva,2018). They randomly chose 10,000 songs with a set of audio features and popularity measures from Spotify Web API. Their model is most successful with logistic regression and neural networks with high accuracy, concluding that danceability and acousticness as the two most significant features. Other machine-learning research also uses K-means clustering to predict hit songs and emphasizes the application of feature-learning methods in content-based music content extraction (Dieleman,2018). Largely inspired by this research, we decided to exploit the tool of machine learning to understand audio features and their correlation with song popularity.

Thesis questions

For the motivation discussed above, the thesis question of this paper was: What are the audio factors influencing songs' popularity on Spotify? As the dataset contained approximately 1000 songs each for 125 different genres, we were able to extract the top 5 genres in the dataset that had the highest median popularity. By examining the relationship between the different attributes of every song, we suggested the attributes most strongly correlated or influenced the difference in popularity.

Out of 15 variables in the dataset, 8 variables (popularity, explicit, danceability, tempo, instrumentalness, loudness, valence, track_genre) were used in the experiment due to their relatively higher correlation with the variable 'popularity'. While the lower number of variables used promoted simplicity and ease of interpretation, it also reduced the accuracy of the experiment as some information was left out. The dataset was scraped using Python and Spotify Web API, proposing no legitimacy concern. However, as Spotify does not determine a single genre for each song, the 'track_genre' variable was decided by scrapping songs tagged when searching the Spotify database. Therefore, the experiment was limited to the information determined by Spotify and might not reflect the songs' actual attributes, such as genre.

Experiment design

1. Linear regression

The motivation for employing multivariate linear regression lies in its simplicity, interpretability, and model transparency. It offers a straightforward understanding of relationships between predictors and the target variable. However, pitfalls include the assumption of linearity, which must be validated through diagnostic plots, and sensitivity to outliers, necessitating careful handling to avoid skewed results. This concise approach balances the advantages of simplicity and interpretability with the potential challenges associated with linearity assumptions and outlier sensitivity in multivariate regression analysis.

In conducting a multivariate linear regression analysis on the Spotify dataset, the variable "popularity" was chosen as the dependent variable, as we were interested in assessing the impact of different song attributes on popularity. All remaining variables (explicit, danceability, tempo, instrumentalness, loudness, valence, track_genre) were independent predictors (X). Then, we proceeded with Exploratory Data Analysis (EDA), utilizing seaborn to conduct a heatmap to visualize predictor correlations, mapping scatterplots, and box plots for insights into the relationships between independent variables and popularity.

After checking for linearity consumption, we would move into the data preparation phase. We planned to split the dataset into training and testing sets while ensuring a consistent distribution of track genres in both subsets. The multivariate linear regression model would then be implemented, treating popularity as the target variable. We then examined the residual plots and QQ plots to validate our assumptions and potentially remove outliers.

Finally, evaluation involved assessing the model's adjusted R^2 and checking the p-values of independent variables to gauge their significance. In case the p-value of certain variables was to surpass the alpha threshold of 0.05, the variable with the highest p-value would be dropped, and the model would be reassessed after each removal. This roadmap ensured a thorough exploration of relationships and validation of assumptions, providing insights into the factors influencing song popularity on Spotify.

2. K-Means Clustering

As we attempted to perform linear regression but failed to draw any insights, we moved to Cluster Analysis. K-Means Clustering recognizes that music preferences and genre popularity are multifaceted and may not conform to linear relationships. By leveraging k-means clustering, we aimed to uncover intricate patterns, capture nonlinear dynamics, and provide a more nuanced understanding of how music attributes contribute to the diversity of genres on Spotify. However, there were also some potential pitfalls, such as sensitivity to initial centroids, sensitivity to outliers, and the subjective nature of cluster interpretation.

We used StandardScaler in scikit-learn for preprocessing to standardize or normalize the features of a dataset. Standardization involves transforming the features so that they have zero mean and unit variance, ensuring that the features are on a consistent scale and contributing appropriately to the clustering process. X_scaled is the standardized version of the original feature matrix X. The fit_transform method computes the mean and standard deviation from the data and then applies the standardization transformation. "Genre" is the categorical variable, and to facilitate the use of clustering algorithms, we transformed the Genre variable into dummy variables, allowing the identification of patterns and similarities based on the encoded categorical information. It ensures that categorical information is appropriately incorporated into the clustering analysis without introducing biases or misinterpretations. We then moved to the elbow method to determine the optimal number of clusters (k) for our dataset. For each genre, we fitted the k-means model to the standardized data and then assigned cluster labels to each data point based on the fitted model. We generated a clustering summary table, which allows us to analyze the characteristics of each cluster to understand the patterns and relationships between genres and audio attributes. After having the summary, we validated the results using domain knowledge or external criteria. If necessary, we refined the analysis by adjusting the number of clusters or reconsidering feature selections.

2D and 3D Principal Component Analysis (PCA) is useful for transforming high-dimensional data into a lower-dimensional space, capturing the most significant variations in the data. These visualizations help us understand the distribution of data points in the reduced-dimensional space. PCA captures the directions of maximum variance, and plotting in 2D or 3D allows for a more interpretable representation of the clustered data.

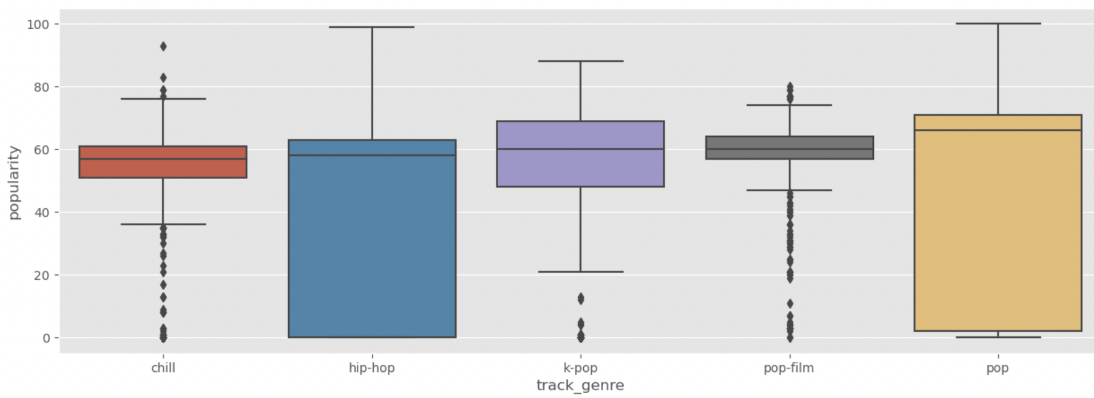
Descriptive analysis

1. Data Cleaning

In order to promote accuracy, reliability, and trustworthiness in the results, we first went through data cleaning. Luckily, our dataset doesn't have any significant number of null values. There are null values, such as 'track_id', 'artists', 'album_name', and 'track_name', in the K-pop genre but they do not affect the usage of song attribute columns. Therefore we do not drop this row. In addition, the original dataset has the 'explicit' variable in Boolean format. To ensure compatibility, improve computational efficiency, and enable a wider range of analytical techniques in our research, we changed 'explicit' from Boolean to numerical values.

2. Exploratory Data Analysis

a. Box plot (Figure 1)

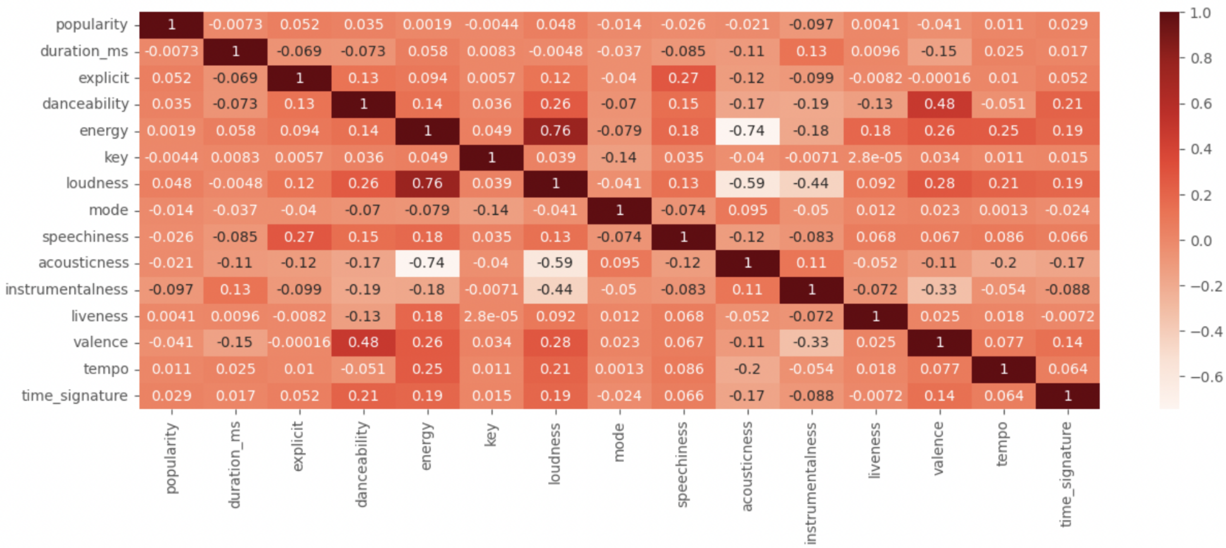


To initiate our project, our initial objective was to identify the top genres among the 125 genres covered in our dataset. We opted to employ popularity as our determining factor, calculated by the Spotify algorithm based on the total number of plays a song has accrued. A higher popularity score indicates that a song is more viral or successful compared to songs with lower popularity scores.

When faced with the decision of whether to use the mean or median as our ranking factor, we conducted a thorough analysis by constructing a box plot to visualize the distribution of the top genres. The box plot revealed that all top genres exhibited either a significant number of outliers or substantial skewness. Consequently, we decided to use the median value to rank the genres due to the principle that the median is less susceptible to the influence of outliers and skewness compared to the mean.

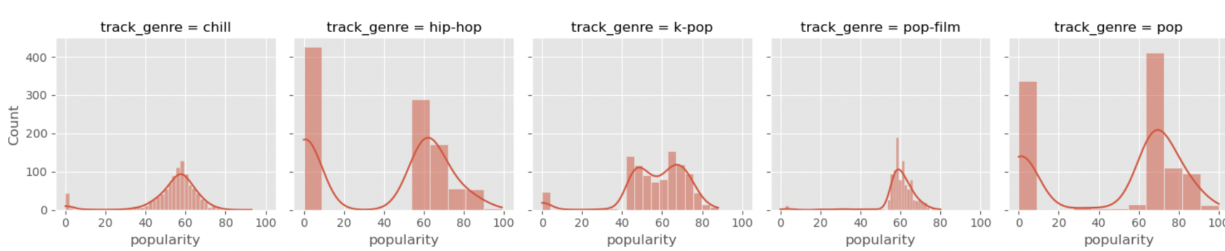
As a result, the top five genres that will be the focal point of our in-depth exploration in this project are Pop, Pop-film, K-pop, Hip-hop, and Chill.

b. Heatmap (Figure 2)



In our pursuit to devise a generalized formula for artists to enhance the likelihood of their songs going viral and achieving success, we eliminated certain data variables, such as 'track_id', 'artists', 'album_name', etc., that were not applicable to the overarching formula. Subsequently, we generated a heat map for the remaining variables to gain insight into their relationships by examining correlation numbers. Unfortunately, the heat map revealed that none of the variables exhibited a strong linear correlation with popularity. The highest absolute correlation number reached only 0.097, while the lowest one fell below 0.005. This unexpected finding led us to deliberate on the possibility of multicollinearity among the variables as a potential explanation for the lack of a linear trend.

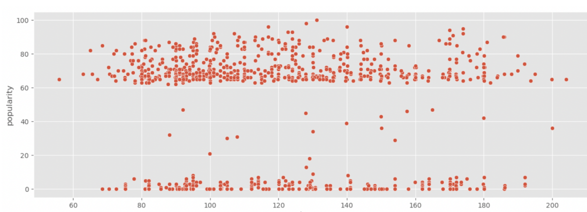
c. Facet-grid histogram (Figure 3)



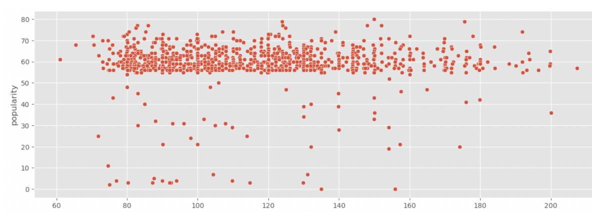
To delve deeper into the popularity distribution of the top five genres, we employed a facet grid of histograms. Upon scrutinizing the distribution patterns, we observed that the Chill, Pop-film, and K-pop genres exhibited similar characteristics, resembling a nearly normal distribution. This normal distribution feature positions these three genres as potentially safer investment choices, as the majority of songs within these genres garnered popularity scores in the average to high level (40-80). Conversely, in the case of Hip-hop and Pop, while they boast a considerable number of songs with scores of 60 and above, there is also a notable presence of songs that scored in the 0-10 range. This polarized distribution renders these two genres riskier investment options, as the likelihood of failing to achieve a high or even average return is higher.

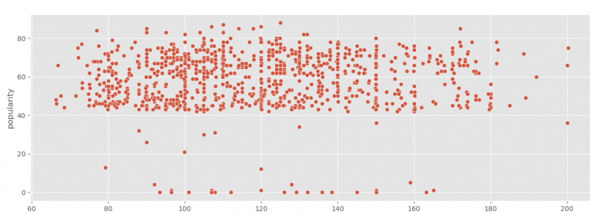
d. Scatterplot

Pop tempo vs popularity (Figure 4.1)

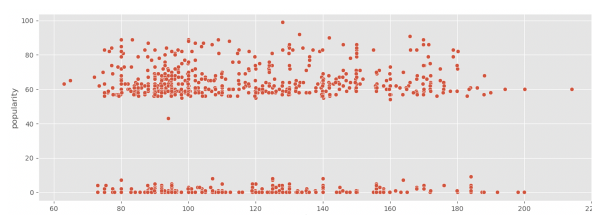


Pop-film tempo vs popularity (Figure 4.2)

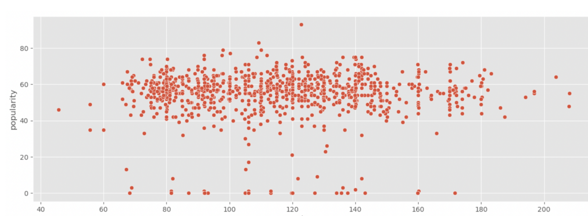




K-pop tempo vs popularity (Figure 4.3)



Hip-hop tempo vs popularity (Figure 4.4)



Chill tempo vs popularity (Figure 4.5)

To investigate the connections between popularity and other variables, a scatter plot stands out as a highly informative graph. This report includes a scatter plot illustrating the relationship between tempo and popularity across all five genres. This example serves to highlight the overall scenario observed when representing variables in scatter plots. Evidently, the data points are dispersed widely, and there is no discernible linear trend between popularity and tempo. This disorderly pattern is consistent across the scatter plots for the remaining variables and their correlation with popularity.

In summarizing our endeavor with linear regression, given the significantly low correlation values derived from the heat map and the disordered patterns observed in the scatter plots, we carefully thought that linear regression might not be the most suitable approach for advancing our project. Consequently, we have opted to proceed with clustering, as it allows for the comparison of relationships across multiple variables, promising a more meaningful insight into the dataset.

K-Means Clustering

1. Analyzing all top 5 genres

cluster	Count	popularity	explicit	danceability	tempo	instrumentalness
0	1000	0.36	−0.37	−0.35	−0.10	−0.25
1	997	0.36	−0.37	0.01	−0.04	−0.25
2	998	0.24	−0.37	0.13	−0.11	−0.25
3	1000	0.60	−0.37	−0.09	0.02	−0.25
4	999	0.28	−0.37	0.70	−0.37	−0.25

cluster	loudness	valence	track_genre_chill	track_genre_hip-hop \
0	−0.01	0.12	−0.5	−0.5
1	0.42	0.27	−0.5	−0.5
2	−0.67	−0.63	2.0	−0.5
3	0.24	−0.05	−0.5	−0.5
4	0.49	0.18	−0.5	2.0

cluster	track_genre_k-pop	track_genre_pop	track_genre_pop-film
0	−0.5	−0.5	2.0
1	2.0	−0.5	−0.5
2	−0.5	−0.5	−0.5
3	−0.5	2.0	−0.5
4	−0.5	−0.5	−0.5

Figure 5.1

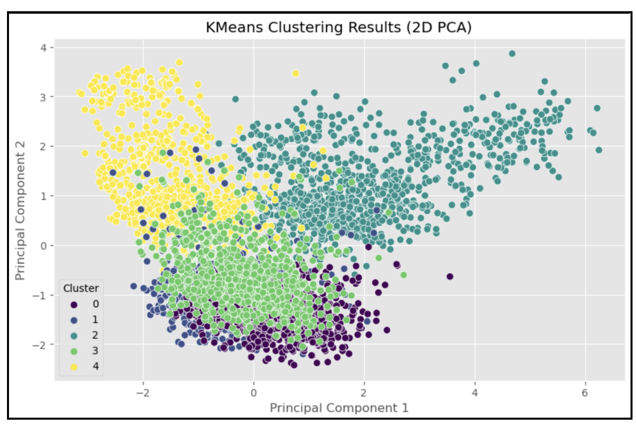


Figure 5.2

We first ran clustering on all genres together to identify global patterns and relationships between genres and audio features, which allows the model to consider interactions and similarities across different genres. Therefore, we can gain comprehensive insights into how genres relate to each other and identify which attributes are commonly seen in most songs in a genre and how they influence the genre's popularity.

After performing k-means clustering, here are our key findings for audio factors influencing genres' popularity: ***Cluster 0*** represents the popularity of Pop-films. The songs in this genre tend to be moderately popular when they are not explicit, have lower 'danceability' and 'tempo'. It may have more vocals or lyrics compared to instrumental parts. Since the "valence" feature has a positive value of 0.12, this suggests that songs in Pop-films should have a more positive or happy emotional tone. ***Cluster 1*** represents the popularity of K-pop. Since 'valence' and 'loudness' have positive values, it indicates that songs in K-pop should have higher loudness and a generally positive mood. ***Cluster 2*** represents the popularity of Chill. From the summary, it is recommended that songs in the Chill genre should not include explicit content, lower loudness and lower happy tone. Negative 'tempo' suggests a more relaxed or laid-back feel, whereas positive 'danceability' indicates a balanced level of rhythm suitable for dancing. ***Cluster 3*** represents the popularity of Pop. Songs in this genre are relatively popular (0.6) when the tracks are less explicit, less danceable and convey a less positive mood. Pop songs tend to be louder to attract listeners. ***Cluster 4*** represents the popularity of Hip-hop. It is pretty expected that Hip-hop songs are likely to be more popular when they have a strong rhythm and cheerful mood. These songs are considered energetic and relatively loud, creating a lively feel when combined with high danceability. However, it's interesting to note that songs in this genre should contain fewer explicit lyrics and fall within a moderate tempo range.

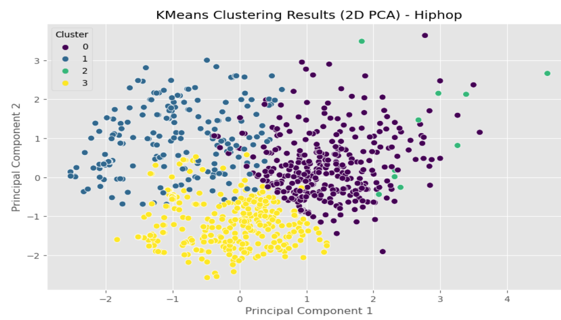
2. Clustering of each genre

a. Hip-hop

Figure 6.1

cluster	Count	popularity	explicit	danceability	tempo	instrumentalness	loudness	valence
0	329	0.75	−0.68	−0.72	−0.30	−0.15	−0.37	−0.53
1	290	−1.16	1.46	0.64	−0.27	−0.15	−0.08	−0.37
2	10	0.70	−0.68	−0.76	−0.28	8.86	0.19	−0.35
3	371	0.59	−0.68	0.45	−0.40	−0.15	0.70	0.82

Figure 6.2



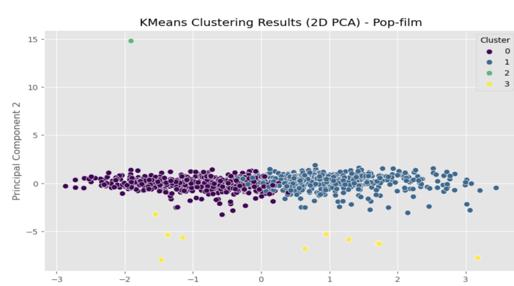
The clusters with the highest popularity are ***Cluster 0*** and ***Cluster 2***, with popularity index of 0.75 and 0.70 respectively (figure 6.1). However, at a closer look, ***Cluster 2*** has 10 data points and only overlaps with ***Cluster 0*** (figure 6.2). Therefore, another clustering model with K = 3 would also be substantial for the Hip-hop genre. Nonetheless, we propose that producing Hip-hop songs following the cluster summary for ***Cluster 0*** and ***Cluster 2*** will give producers a higher chance of scoring hits. An interesting insight is that loudness, tempo, and explicit language were relatively low, which contradicted what we generally expect of a Hip-hop song.

b. Pop-film

Figure 7.1

cluster	Count	popularity	explicit	danceability	tempo	instrumentalness	loudness	valence
0	508	0.07	−0.03	0.71	0.03	−0.15	0.25	0.73
1	481	0.07	−0.03	−0.59	−0.26	−0.15	−0.03	−0.80
2	1	1.14	31.61	0.31	1.30	−0.15	1.04	0.05
3	10	−0.17	−0.03	0.16	−0.31	9.17	−0.66	0.25

Figure 7.2



Clusters 0 and ***Cluster 1*** represent the majority of the songs with moderate popularity of 0.07. ***Cluster 0*** has positive attributes related to 'danceability' and 'valence', whereas ***Cluster 1***

has lower 'danceability' and 'valence'. It proposes that the Pop-film genre can encompass a wide range of preferences, diversifying the production portfolio and continuously analyzing the market trends is recommended to tailor music to suit specific senses, moods, or target audience. ***Cluster 2*** is an outlier with a very high popularity of explicit content and a unique combination of musical features.

c. Pop

cluster	Count	popularity	explicit	danceability	tempo	instrumentalness	loudness	valence
0	310	0.54	-0.28	-0.93	-0.36	-0.15	-0.69	-0.90
1	231	-1.40	-0.28	0.39	0.23	-0.15	0.42	0.51
2	379	0.63	-0.28	0.61	-0.35	-0.15	0.39	0.46
3	6	-1.40	-0.28	-0.15	0.04	11.60	-0.44	-1.76
4	74	0.85	3.54	0.34	0.29	-0.15	0.32	0.01

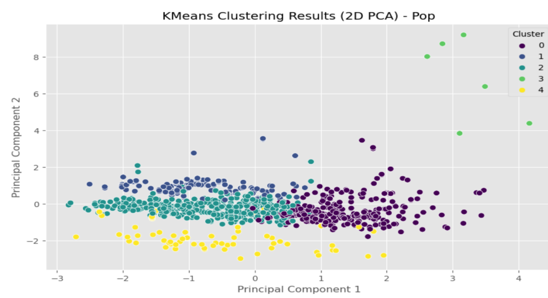


Figure 8.1

Figure 8.2

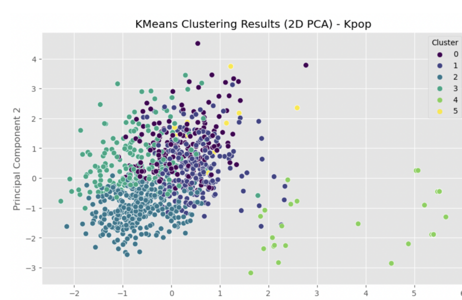
Cluster 4 with 0.85 popularity index has the highest popularity. Popular songs in this genre will have explicit lyrics, corresponding with low 'instrumentalness' to allow for a focus on catchy lyrics. High danceability and upbeat tempo are also reasonable to see in the Pop genre as they deliver an energetic atmosphere. However, the values of 'loudness' and 'valence' are interesting to observe. These popular Pop songs exhibit moderate loudness and a valence that falls somewhere between positive and negative. This suggests that they create a unique mood that is not overly loud and evokes a mix of emotions.

d. K-pop

Figure 9.1

Figure 9.2

cluster	Count	popularity	explicit	danceability	tempo	instrumentalness	loudness	valence
0	180	0.07	−0.23	−0.65	−0.83	−0.14	−0.15	−1.06
1	230	−0.52	−0.23	0.03	−0.24	−0.14	−1.04	0.71
2	339	0.66	−0.23	0.66	−0.19	−0.14	0.77	0.46
3	189	0.36	−0.23	−0.60	1.37	−0.14	0.39	−0.33
4	49	−3.36	4.40	1.89	−0.45	−0.14	0.11	−1.14
5	11	−0.05	−0.23	0.11	−0.53	7.18	−0.76	−0.55



Based on the clustering analysis of the K-pop genre, ***Cluster 2*** registers the highest popularity, while ***Cluster 4*** ranks as the least popular. Despite ***Cluster 2*** featuring the most explicit lyrics, 'explicitness' seems to have a negligible impact on a song's popularity since all other clusters exhibit similar values for this variable. The influence of 'instrumentalness' on popularity also appears to be minimal. 'Danceability' presents a more nuanced picture, with a marginally positive correlation with 'popularity', suggesting that songs with higher danceability may tend to be more popular. As for 'tempo', the analysis indicates an inverse relationship with popularity. Furthermore, the analysis reveals that the most popular songs typically score higher on loudness, supporting the notion of a positive link between 'loudness' and 'popularity'. A similar pattern is observed with 'valence', hinting at a potential positive association.

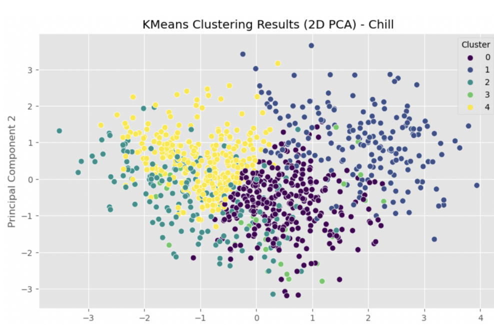
There is considerable overlap between ***Clusters 0, 2, and 4***, indicating that many songs share similar values for the characteristics defined by the first two principal components. Numerous points in ***Clusters 0, 4 and 5*** represent outliers within the dataset. However, when referring to both graphs, it's evident that 'popularity' and 'loudness' are captured by the principal components. Therefore, enhancing loudness may be important for creating popular K-pop songs. 'Explicitness', on the other hand, does not closely align with the principal components.

e. Chill

Figure 10.1

cluster	Count	popularity	explicit	danceability	tempo	instrumentalness	loudness	valence
0	302	0.16	-0.45	-0.53	-0.02	-0.55	0.03	-0.62
1	184	0.16	-0.45	0.05	-0.19	2.07	-0.94	-0.70
2	163	0.36	2.21	0.30	0.12	-0.55	0.27	0.04
3	56	-3.61	-0.45	0.26	-0.33	-0.55	0.66	0.01
4	291	0.29	-0.45	0.65	-0.02	-0.55	0.55	0.81

Figure 10.2



Compared to other genres, Chill songs generally exhibit lower popularity scores, as indicated in the cluster summary. Within the five clusters identified for the Chill genre, ***Cluster 2*** attained the highest popularity score at 0.36. Upon analyzing the 2D PCA graph, it becomes evident that ***Cluster 3*** significantly overlaps with and nearly merges with other clusters. This suggests that songs in ***Cluster 3*** share similar attributes with those in other clusters, implying that a total cluster number of 4 could be a viable alternative for the Chill genre.

In summary, the success formula for the Chill genre suggests that artists should aim for high explicitness in lyrics, moderately high levels of 'danceability', 'tempo', and 'loudness', as well as low 'instrumentalness' for their Chill songs. However, they have the flexibility to choose between creating high-valence songs, conveying positive and cheerful feelings, or low-valence songs, evoking a more somber and blue tone.

Limitations and mitigations

The Principal Component Analysis (PCA) model is predicated based on the assumption of linearity among all variables. However, when we examine different genres, it becomes evident that popularity does not uniformly exhibit a clear-cut relationship with every variable. Consequently, preemptive assumptions can significantly enhance the clarity of the final results. To address this limitation, it becomes imperative to consider the incorporation of non-linear dimensionality reduction techniques, such as Isometric Mapping (Isomap), which are capable of capturing intricate relationships that may not adhere to linear patterns.

Another critical aspect to bear in mind is the cleanliness of our data. Despite our best efforts, some outliers may persist within the dataset, potentially exerting a disruptive influence on the final clustering outcomes. To bolster the precision of our results, a comprehensive data-cleaning process should be undertaken.

In the realm of selecting the optimal number of clusters (k-value), the elbow method is commonly employed. However, we encounter scenarios where there is a lack of sharp transitions among the slopes. Consequently, it becomes challenging to definitively determine whether the designated k-value is the most appropriate choice. Exploring alternative methods and metrics, such as silhouette scores or Davies-Bouldin index, can provide a more nuanced and robust perspective on cluster count determination.

It's worth acknowledging that the raw data itself may not always be entirely accurate. In some instances, certain values may have been assigned based on subjective judgment, introducing the potential for bias in the final results. To mitigate this bias, it is essential to diversify our sample sizes and sources. This can be achieved by conducting research into user preferences across different countries, occupations, and age groups.

Conclusion

In conclusion, after encountering limitations in our linear regression analysis, we turned to the clustering and PCA models to delve into the crucial attributes that underlie the creation of popular songs. We used the elbow method to identify the cluster number and then applied it to the future model. Our efforts have yielded valuable insights, resulting in a set of combinations that can serve as a valuable reference point for production companies seeking to enhance their prospects of producing viral hits.

Our research suggests that undanceable, quieter, and moodier Hip-hop songs are more likely to attain viral status. This discovery challenges conventional expectations and opens new avenues for creativity within this genre. For the Pop, Chill, and Pop-film genres, our analysis reveals a convergence of attributes, creating a similar atmospheric quality. While fast-paced tempo and danceability consistently align with the essence of Pop and Pop-film, it is intriguing to note their presence in the Chill genre. This deviation from the genre's typical slow tempo and overall sense of calmness and relaxation adds a layer of complexity and potential innovation.The K-pop genre aligns closely with our expectations, featuring a fast-paced tempo, upbeat rhythms, and a positive mood in terms of valence. These characteristics exemplify the energetic and lively nature of K-pop, contributing to its infectious appeal.

It is essential to acknowledge that not all songs fit neatly into this template. However, this exploration provides valuable insights, allowing producers to discern patterns and gain a deeper understanding of the prevailing aesthetics within the clustering, ultimately aiding them in creating music that resonates with the mainstream audience.

References

Dieleman, S. and Schrauwen, B.(2013). Multiscale approaches to music audio feature learning, in Proc. Int. Soc. *Music Information Retrieval Conf.*, pp. 116-121.

Georgieva, E., Şuta, M., & Burton, N.S. (2018). HITPREDICT : PREDICTING HIT SONGS USING SPOTIFY DATA STANFORD COMPUTER SCIENCE 229 : MACHINE LEARNING. Retrieved from http://www.semanticscholar.org/paper/HITPREDICT-%3A-PREDICTING-HIT-SONGS-USING-SPOTIFY-229-Georgieva-Şuta/5f40e603ac969eb476f1af21b1efdd70153c24c0

Malley, M. (2023) Global Music Production and Distribution. *IBISWorld*. Retrieved from http://my-ibisworld-com.libproxy1.usc.edu/gl/en/industry/q8712-gl/at-a-glance.

Karydis, I., Gkiokas, A., & Katsouros, V. (2016). Musical Track Popularity Mining Dataset. *Artificial Intelligence Applications and Innovations*, 475, 562–572. Retrieved from https://doi.org/10.1007/978-3-319-44944-9_50

MaharshiPandya. (2022). 🎹 *spotify tracks dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset/data