Dickinson College

SPOTIFY MUSIC ANALYSIS DATASET

Hannie Pham

DATA 180: Introduction to Data Science

I. SUMMARY DATASET

1.1. Introduction and description

Dataset Link: https://www.kaggle.com/code/aeryan/spotify-music-analysis/input

Spotify music platform always attracts the attention of not only music lovers but also famous producers and musicians who want to release their products globally. With the rapid proliferation of tracks in the music market, it is difficult to categorize the components in a piece of music used by musicians and this genre will capture the attention of the audience. Therefore, by examining the characteristics of the most popular songs on the platform, we can gain insights into what types of music people enjoy listening to and what musical elements tend to resonate with audiences. In addition to providing insights into user preferences and music trends, the Spotify music dataset can also be used to improve music production strategies. By analyzing the characteristics of successful songs, music producers and artists can gain insights into what makes a hit song and can use this information to guide their own creative processes. In order to reach those goals, the Spotify music analysis dataset from Kaggle will be presented and analyzed by different tools, namely Data Visualization, Clustering, Linear Regression, and Classification Models in this report.

This original dataset contains 2017 objects and 17 variables. The dataset comprises various features, such as acousticness, danceability, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time signature, valence, among others. Each row represents a song, and each column represents feattures of music tracks.

1.2. Data Visualization

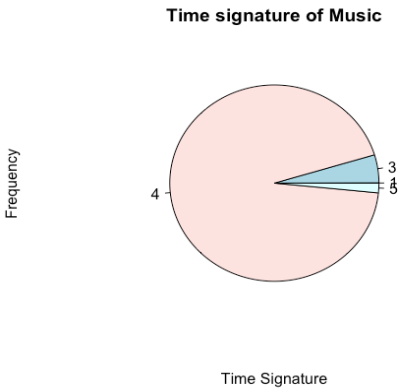


Figure 1: Time signature of tracks

Figure 1 shows a clear result in the pink area that most tracks have a beat number of 4/4 and absolutely no track has a beat number of 1/4 per measure.

Besides, in order to better understand the characteristics of songs played on Spotify, the frequency distribution of several attributes of songs will be shown based on the bar plots, including acousticness, danceability, energy, instrumentalness, speechiness, tempo, loudness, and valence.

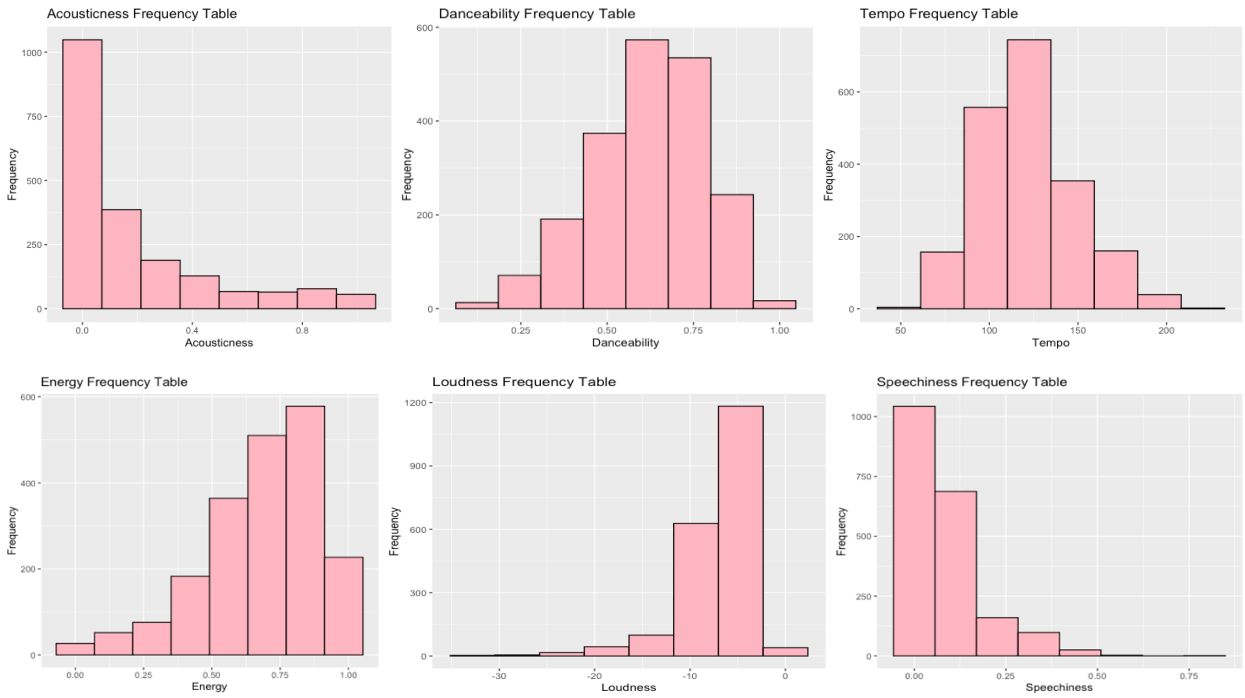


Figure 2: Plots of Different Feature Frequencies of Musics Tracks

The acousticness plot shows that the majority of songs have a relatively low level of acoustics, indicating that they are predominantly electronic rather than featuring natural instruments. The danceability plot shows a roughly normal distribution, with most songs having a moderate level of danceability. A relatively normal distribution may also be seen in both energy and tempo plots, which indicates that most songs have a moderate amount of energy and tempo. The loudness plot also shows a roughly normal distribution, with most songs having a moderate level of loudness. This suggests that there is no one "ideal" level of loudness for songs on the platform. Finally, the speechiness plot shows a left-skewed distribution, indicating that there are more songs with lower levels of spoken word elements, such as lofi tracks, than there are purely instrumental tracks.

Overall, the different plots in figure 2 demotrates the characteristics of songs on the Spotify platform, with most songs having moderate levels of danceability, energy, tempo, loudness, and relatively low levels of acousticness. Meanwhile, the speechiness plot highlights the popularity of spoken word elements in some tracks.

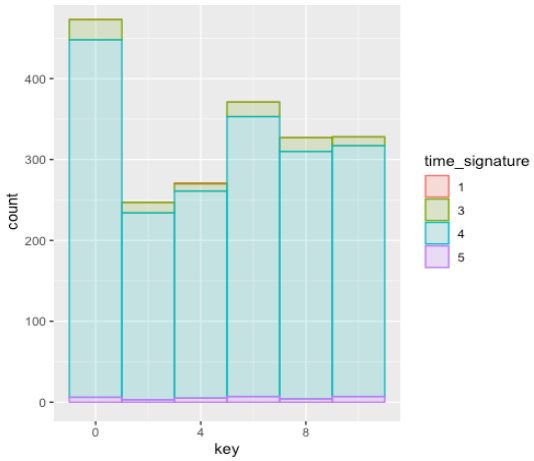


Figure 3: Histogram between key and time_signature variables

This plot shows the distribution of songs across different keys, with bars colored by time signature. The x-axis represents the musical key of the songs, while the y-axis shows the count of songs falling within each key. Overall, the figure 3 shows that the most common keys for songs are C, G, and D, with a peak around C. The distribution of songs across different keys does not seem to vary significantly by time signature. However, there appears to be a slight increase in the

proportion of songs in the key of C with a 4/4 time signature, and a slight increase in the proportion of songs in the key of G with a 3/4 time signature.

1.3. Summary of methods

Based on the features represented in the dataset, the response value can be time_signature, energy, target, and the rest of the features will be predictor variables.

Firstly, the Silhouette value, K-Medoid, and Cluster Dendrogram will be the best tools to cluster the songs in the dataset based on their music features. Also, there will be a Silhouette value to decide the optimal number of clusters. Each cluster will have its own characteristics that after drawing the model based on the optimal number can be analyzed in detail.

Secondly, the linear regression model will be built with a time_signature as a response variable and the rest of the dataset as predictors based on analysis of paired relationships. Since most songs have a beat number of 4/4, this model will help show which components have a significant influence on the time-signature and the relationship between those characteristics in the tracks.

Finally, K-nearest neighbor (KNN) will be used to build a classification model to classify songs as either popular or unpopular to listeners.

II. MODELS

2.1. Clustering model

2.1.1. Setup

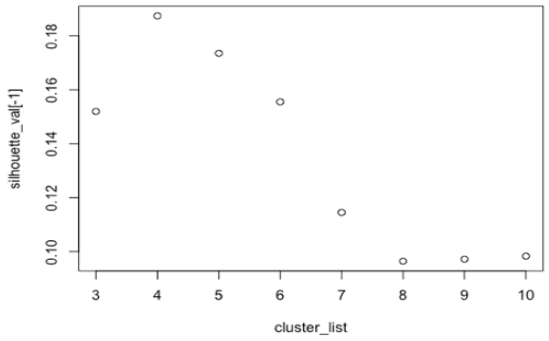
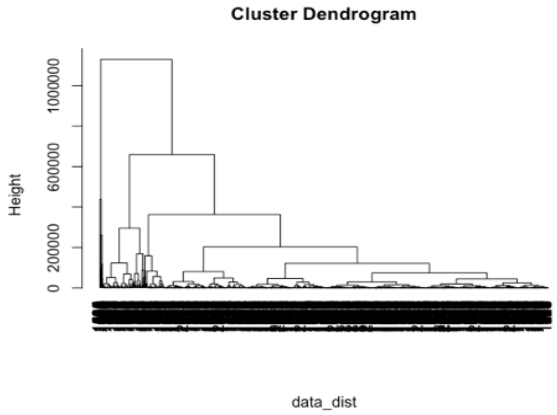


Figure 4: Silhouette value

Based on the figure 4, the Silhouette value is highest at K = 4 so that K = 4 is an optimal number to run K-Medoid model.



Based on the figure 5, the data will be cut at h = 200000 by applying the cutree method to see more detail in clusters

Figure 5: Data Clustering Dendrogram

2.1.2. Model

	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode
490	0.028500	0.764	270680	0.768	0.02350	5	0.1220	-5.614	1
741	0.000483	0.701	224200	0.829	0.00223	6	0.1400	-5.749	0
1507	0.041800	0.505	221050	0.705	0.00000	5	0.0815	-6.097	1
1870	0.075400	0.655	202163	0.873	0.00000	7	0.1050	-4.992	0

	speechiness	tempo	time_signature	valence	target	song_title	artist
490	0.0382	125.988	4	0.627	1	Night And Day	Hot Chip
741	0.0513	132.349	4	0.501	1	The Sound Of San Francisco - Progressive Album Mix	Global Deejays
1507	0.0395	126.021	4	0.309	0	For You	Demi Lovato
1870	0.0692	130.018	4	0.571	0	Fire	Majozi

Figure 6: Medoids Model with 4 clusters

Based on the k-medoids clustering model in figure 6, the dataset was divided into 4 clusters. The characteristics of each cluster can be summarized as follows:

Cluster 1: This cluster has high values instrumentalness and valence, along with moderate values of energy, loudness, indicating that the songs are highly produced and have minimal vocals. The songs in this cluster have a fast tempo and a 4/4-time signature. The songs in this cluster include "Night And Day" by Hot Chip.

Cluster 2: This cluster has high values of danceability and energy, along with moderate values of valence and loudness. The songs in this cluster have a fast tempo and a 4/4-time signature. They also have low acousticness and instrumentalness values, indicating that they are highly produced and have vocals. The songs in this cluster include "The Sound Of San Francisco - Progressive Album Mix" by Global Deejays.

Cluster 3: This cluster has low values of danceability and energy, along with low valence, instrumentalness and liveness. The songs in this cluster have a slow tempo and a 4/4-time signature. They also have high acousticness values, indicating that they are mostly acoustic and have minimal production. The songs in this cluster include "For You" by Demi Lovato.

Cluster 4: This cluster has extreme high value of acousticness, along with low values of loudness, indicating that the songs are mostly acoustic and have minimal production. The songs in this cluster have a slow tempo and a 4/4-time signature. The songs in this cluster include "Fire" by Majozi.

2.1.3. Why the model is interesting

The result can get the audience's interest because the model suggests that certain features are more likely to be found in popular songs. For example, the model shows that higher values of danceability, energy, and speechiness are associated with a higher likelihood of a song being popular, while higher values of acousticness are associated with a lower likelihood of a song being popular. This information can help listeners identify songs that are more likely to be popular and potentially more enjoyable for them to listen to.

2.2. Regression Model

2.2.1. Introduction

The dataset will be predicted by time_signature (dependent variables (Y)) of these tracks by building a linear regression model and based on every other variable that dataset contains, excepting song_title and artist (independent variables (X), categorical variables).

2.2.2. Model

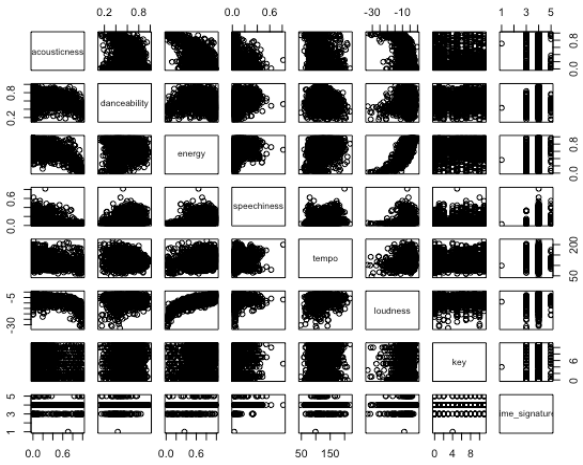


Figure 7: Paired relationship between track features

Figure 7 shows that the variables acousticness, danceability, energy, speechiness are pairs that can have a linear relationship with each other. With respect to time_signature, one conjecture can be made that it also has a linear relationship, this is not clearly shown in the figure, though. Therefore, building a linear regression model can help us see more clearly how significant the time_signature variable is to other predictor variables to determine if it is truly linear.

```
Call:
lm(formula = time_signature ~ ., data = data_omitted[2:15])

Residuals:
     Min       1Q   Median       3Q      Max
-2.82680 -0.01901  0.01403  0.05426  1.26685

Coefficients:
                    Estimate     Std. Error t value             Pr(>|t|)
(Intercept)       3.89487106814  0.06577434834  59.216 < 0.0000000000000002 ***
acousticness     -0.10914843678  0.02926157522  -3.730             0.000197 ***
danceability      0.08486281601  0.04148595339   2.046             0.040928 *
duration_ms      -0.00000003914  0.00000007328  -0.534             0.593352
energy            0.14785799102  0.04896953857   3.019             0.002565 **
instrumentalness -0.01406999526  0.02364010066  -0.595             0.551794
key               0.00029944956  0.00154633140   0.194             0.846468
liveness         -0.02308395724  0.03713044253  -0.622             0.534211
loudness         -0.00210792237  0.00258009551  -0.817             0.414029
mode             -0.01153896426  0.01163579505  -0.992             0.321474
speechiness       0.25811621523  0.06457198101   3.997            0.0000664 ***
tempo            -0.00085432896  0.00021570321  -3.961            0.0000774 ***
valence           0.05688201501  0.02696939281   2.109             0.035057 *
target           -0.00241864789  0.01182699246  -0.205             0.837982
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2475 on 2003 degrees of freedom
Multiple R-squared:  0.07056,   Adjusted R-squared:  0.06453
F-statistic:  11.7 on 13 and 2003 DF,  p-value: < 0.00000000000000022
```

Figure 8: Spotify Music Dataset Regression

The figure 8 shows that there are several factors that most influence the time of music tracks, namely acousticness, energy, speechiness, and tempo, while danceability and valence have a marginal significance. The other predictors are not significant.

Acousticness refers to the degree of acoustic instrumentation in a track, and the negative coefficient of -0.11 suggests that tracks with higher levels of acousticness tend to have lower time signatures. This finding aligns with the fact that acoustic instruments tend to be associated with slower, more laid-back music styles.

Energy has a positive coefficient of 0.15, indicating that tracks with higher energy tend to have higher time signatures. This finding makes sense as high-energy music is often associated with faster tempos and more complex rhythmic patterns.

Speechiness, which measures the presence of spoken words in a track, also has a positive coefficient of 0.26. This suggests that tracks with more spoken words tend to have higher time signatures, which could be attributed to the fact that spoken words often require more rhythmic complexity than instrumental music.

Tempo, which measures the beats per minute of a track, has a negative coefficient of -0.00085. This suggests that as the tempo of a track increases, the time signature tends to decrease. This finding aligns with the fact that fast tempos often require simpler rhythmic patterns to maintain cohesion.

Finally, valence, which measures the positivity of the emotional content of a track, has a positive coefficient of 0.057, indicating that more positive tracks tend to have higher time signatures. This finding suggests that upbeat music tends to have more complex rhythmic patterns. Also, danceability, which measures of how suitable a song is for dancing, has a positive coefficient of 0.085, illustrating that more suitable tracks tend to have higher time signatures.

It is important to note that the multiple R-squared value of 0.07056 suggests that the predictor variables explain only a small portion of the variation in the time signature. This indicates that there are likely other factors at play in determining the time signature of a music track that are not captured by the variables in the model. The standard errors, t-values, and p-values for each predictor are also provided. The residual standard error is 0.2475, and the F-statistic indicates that the overall model is statistically significant.

2.2.3. Why the model is interesting

While these findings may not be directly interesting to listeners, they could be useful for music producers or researchers studying the relationships between different musical elements. For example, a producer may use this information to intentionally create songs with certain time signatures based on the desired mood or genre of the song.

2.3. Classification Model

2.3.1. Introduction

For this classification, K-nearest neighborhood (KNN) will be applied to build classification model to classify tracks as either popular or unpopular to listeners.

2.3.2. Models

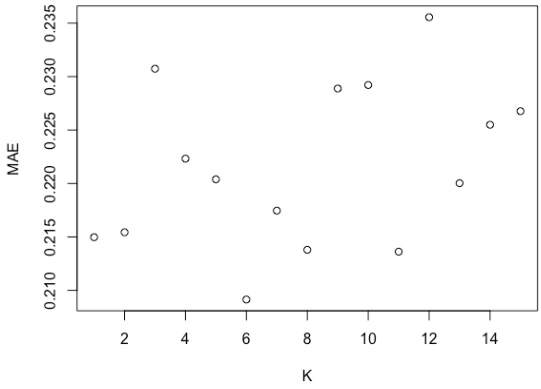


Figure 9: Accuracy percentage of different number of clusters

With the quality measure being the accuracy percentage (percentage of correctly identified sample units), there are different number of K that can be plotted and picked to get the best results (shown in figure 9). K = 11 is the suggested number of neighbors. Meaning our highest accuracy percentage is KNN with K = 11.

```
> confusionM <- table(validationset$target, knn_pred)
> confusionM
   knn_pred
      0   1
  0 118  75
  1 101 110
```

```
> confusionM <- table(validationset$target, knn_pred)
> confusionM
   knn_pred
      0   1
  0 120  73
  1  88 123
```

Figure 10: Confusion matrix (left:k=4, right:k=11)

The validation set based on Confusion matrix in figure 10 shows :

- There are 118 + 75 = 193 unpopular tracks, and, 101 + 110 = 211 popular tracks.
- Accuracy with K = 4: (118 + 110) / (118 + 75 + 101 + 110) = 56.5%
- Accuracy with K = 11: (120 + 123) / (120 + 73 + 88 + 123) = 60.2%

Based on calculations drawn from the Confusion Matrix, the percentage of sample units is more accuracy at about 60.2% with K = 11. This is not a very large number, so the analysis can conclude that it is difficult to distinguish exactly 60.2% of the tracks that are popular or not among listeners. Only a handful of songs will be heard more by the user. When linking all the analyzes together, this makes sense because in the cluster part, when dividing the dataset into different groups and analyzing each cluster, the audience can see each cluster (each group of music) has its own characteristics. The choice depends on the preference of each listener.

2.3.3. Why the model is interesting

This model helps us redefine the accuracy of analyzes performed in previous analyses. It is easy to see that when using the K value from KNN, the percentage of model accuracy is higher.

III. CONCLUSION

While it is impossible to pinpoint the exact percentage of music genres that are most loved because of so many tracks and too many users on Sporify, analyzing this data offers a boon to music producers and music distributors can tune albums of the production music or instrumental music so that listeners can easily find them. At the same time, in the music production process, the author also clearly sees the interaction between the musical properties to help the mixes be published more perfectly. Moreover, controlling the time_signature is also important to producers because most songs on Spotify and tend to be interesting to the audiences have 4/4 beats. As for listeners, they can rely on this analysis to see which songs contain what musical characteristics that might suit their preferences and that they can enjoy.