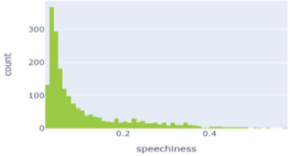
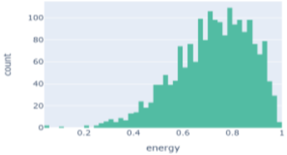
2000 – 2019 Spotify Top-Hits Analysis

Xinyu Mei (xinyumei@umich.edu)

Dataset & Preprocessing

- The research is based on a dataset of Top Hit songs on Spotify from 1998 to 2020.
- Deleted data of songs before 2000 and after 2019 due to small amount.
- Split original genre format to a group of dummy variables.
- Overall features of hit songs: high danceability, energy, loudness; low speechiness, instrumentalness, acousticness, liveness.
- Most frequent genres: pop, hip hop, R&B, Dance/Electronic, Rock





Similarity Analysis

- Matrix similarity analysis was conducted between years and top genres.
- It turns out that the average value of features does not show difference year to year or genre to genre, since almost all the cosine similarity values are higher than 0.99.

K-Means Clustering

- I tried K-Means clustering to summarize some characteristics of the top hits.
- With the clusters, we can further know to which songs a random song is more similar to.

- Cluster 0 – On Live: high liveness
- Cluster 1 – Emotional Ballad: low energy, low tempo
- Cluster 2 – Not so Top Hit: low popularity
- Cluster 3 – Let's Rap it: high speechiness, high tempo
- Cluster 4 – Innocent G-level: low explicit

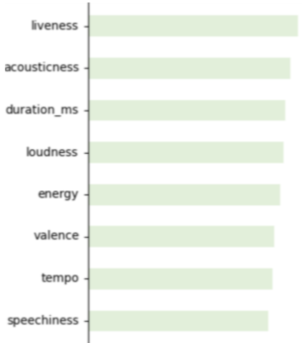
Factors Affecting Popularity

- Used Machine Learning models to create a relation between a song's popularity and other features.
- Linear Regression, Random Forest and SVM were applied.

	LR	RF	SVM
MAE	0.6433	0.6503	0.1964

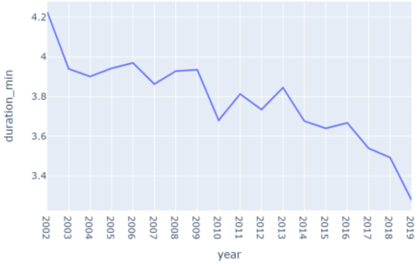
- According to the MAE, SVM has the best predicting performance among the three methods. And it has an accuracy of 0.8801.
- So it is possible and reliable to predict a song's popularity using SVM.

- According to the importance analysis by random forest, features like liveness has an importance of over 0.06.



Time Series Analysis

- The average duration of songs has a decreasing trend through the 20 years.



- For other features, no trend was shown in the analysis.

- Considering top 5 genres, they have a quite high similarity in the duration time series.
- This similarity does not show in trend lines. Actually, the trend line of R&B and rock showed irregular changes over the 20 years.
- However, after calculating the cosine similarity matrix, it showed that the cosine similarity between either two genres is higher than 0.99.

Limitations

- It seems that there is a high homogenization among the top hit songs.
- There are still many features in this dataset not thoroughly researched, especially in similarity analysis and time series analysis.
- In k-means clustering part, only k=5 was tried. Scree plot can be applied to find a best k in further researches.

References

- https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019