Spotify Music popularity analysis

Jillian Hay, Yuren Xia, Ganfu Yuan,
& Liguang(Lee) Zhang

INTRODUCTION

Problem:
To determine the features of songs most closely associated with popularity

Motivation

- Shorter Duration of songs appear to become a trend
- Loudness war in the music industry→ potential relationship b/w sound level & popularity?

Importance

- Identify song features associated with popularity
- Reflect changing trends in the music industry
- Insights to artists
- Insights to Music streaming services

Literature

- Cluster analysis using k-means and Hierarchical clustering on spotify top 100 songs for 2017 & 2018
- Chart-topping songs have a formulaic, pop-friendly sound, with high danceability and valence

(Al-Beitawi, Salehan&Zhang, 2020)

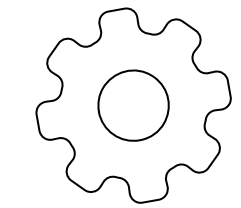
Descriptions

- Variables

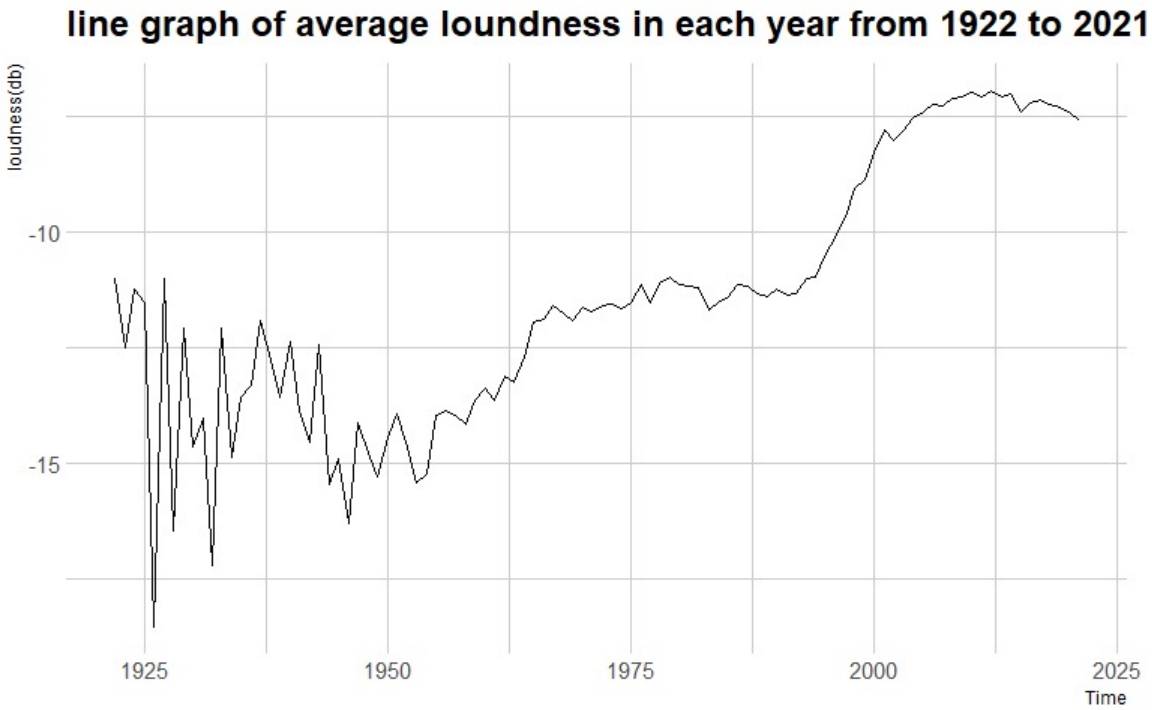
Variables

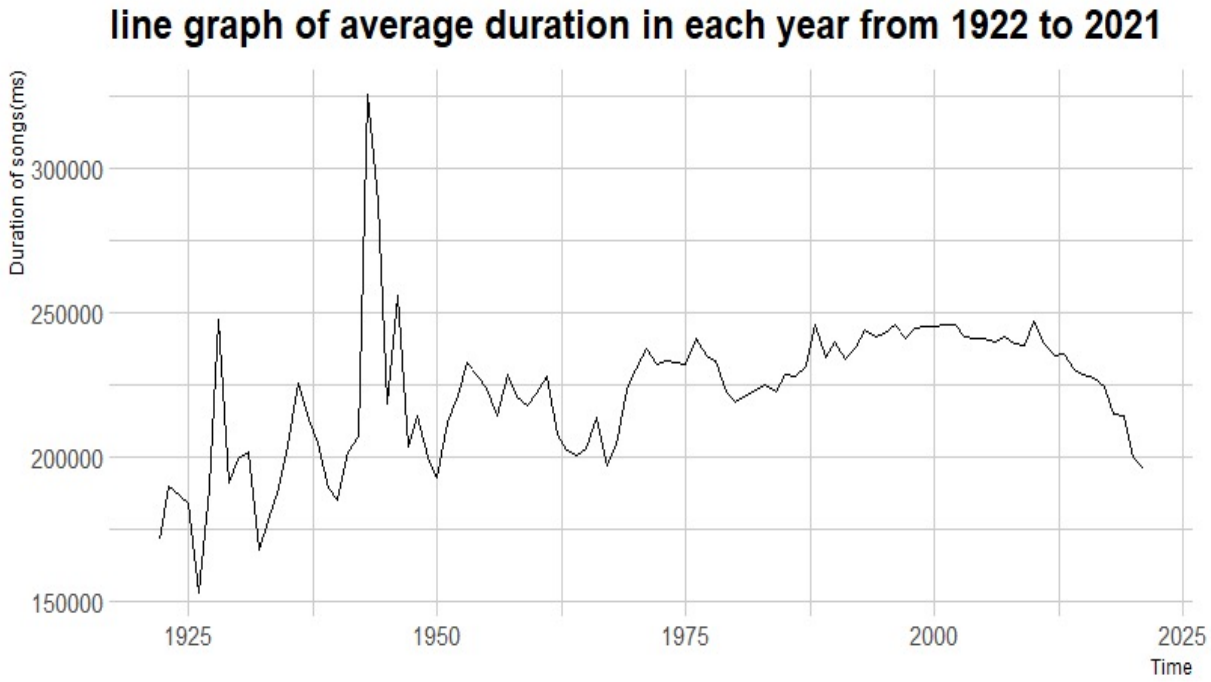
1	id	chr
2	name	chr
3	popularity	int
4	duration_ms	int
5	explicit	int
6	artists	chr
7	id_artists	chr
8	release_date	chr
9	danceability	num
10	energy	num

11	key	int
12	loudness	num
13	mode	int
14	speechiness	num
15	acousticness	num
16	instrumentalness	num
17	liveness	num
18	valence	num
19	tempo	num
20	time_signature	int

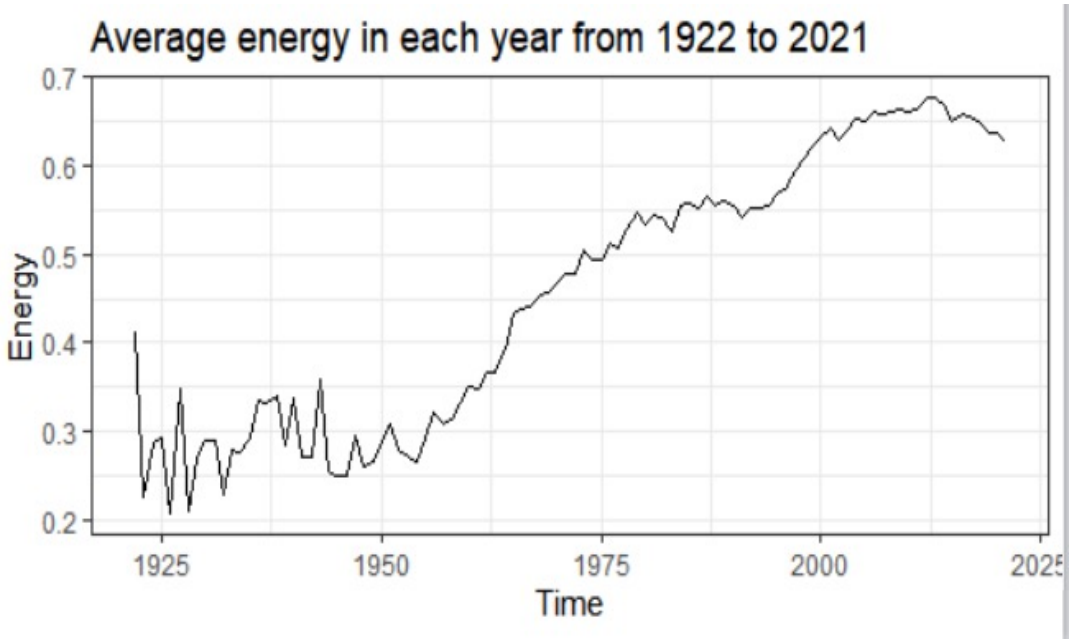


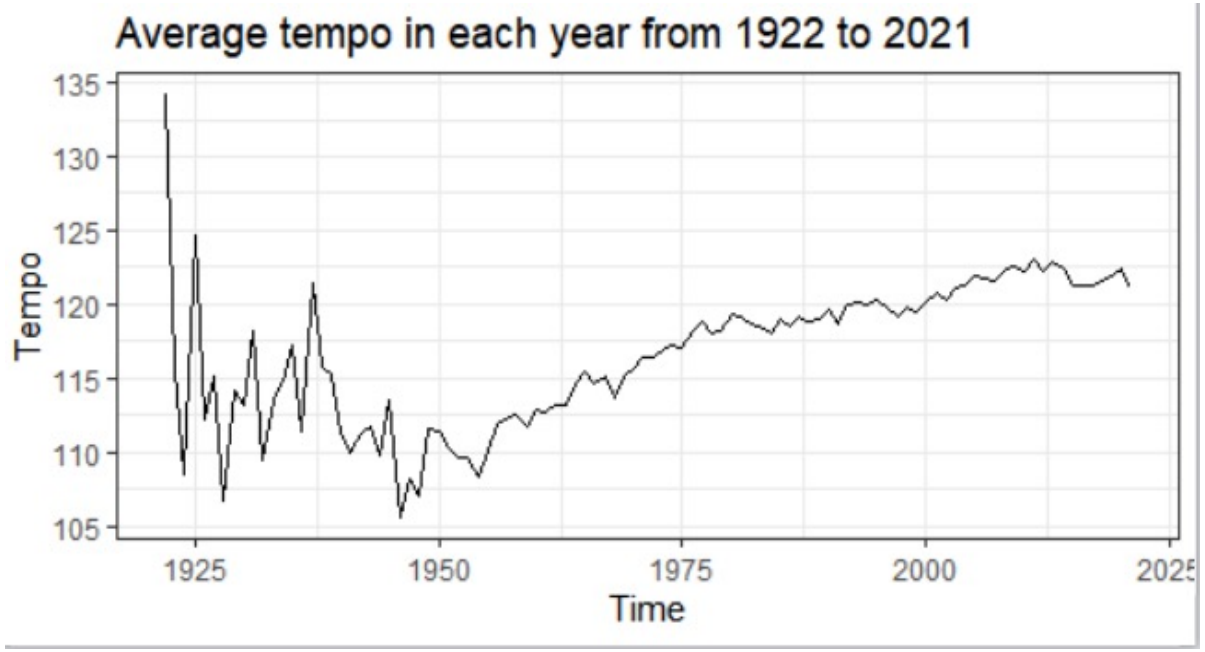
Trends


line graph of average loudness in each year from 1922 to 2021
loudness(db)
-10
-15
1925
1950
1975
2000
2025
Time

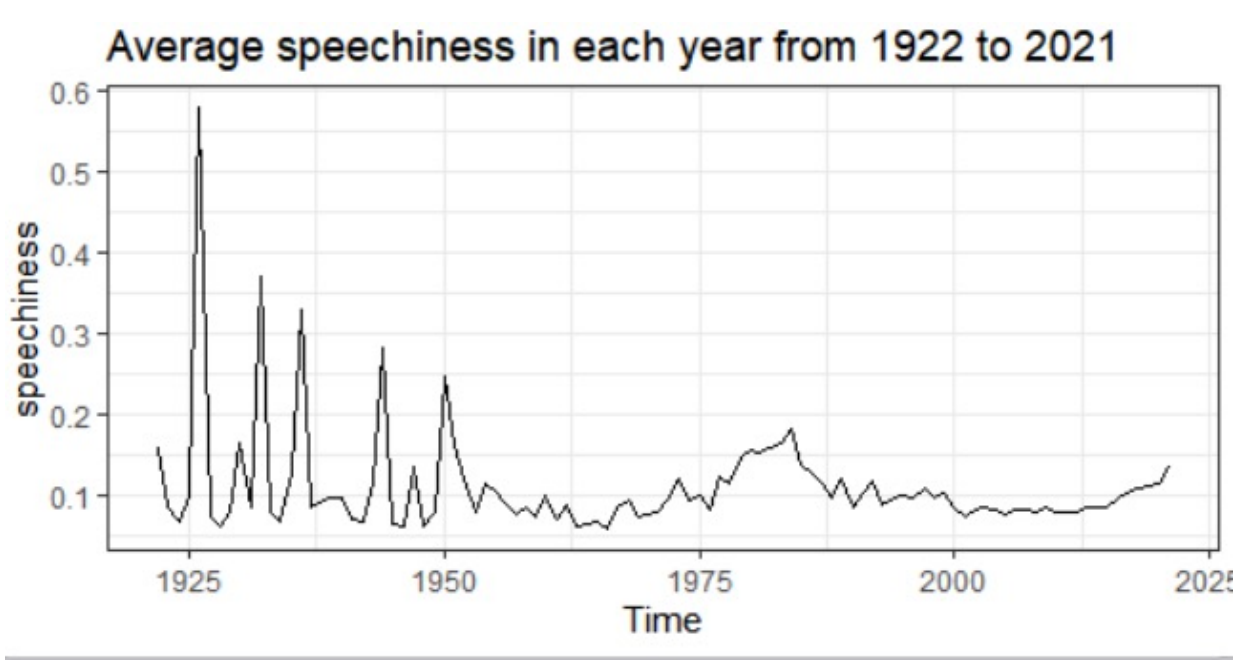

line graph of average duration in each year from 1922 to 2021
Duration of songs(ms)
300000
250000
200000
150000
1925
1950
1975
2000
2025
Time

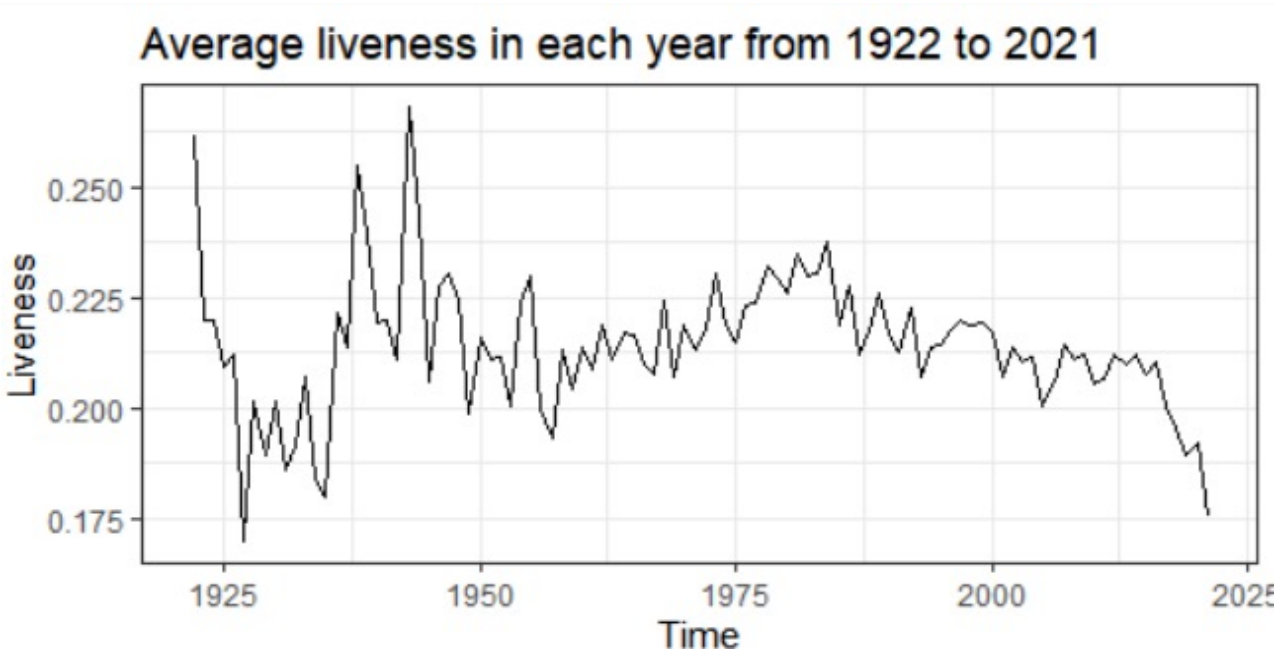
Trends


Average energy in each year from 1922 to 2021
Energy
0.7
0.6
0.5
0.4
0.3
0.2
1925
1950
1975
2000
2025
Time


Average tempo in each year from 1922 to 2021
Tempo
135
130
125
120
115
110
105
1925
1950
1975
2000
2025
Time

Trends

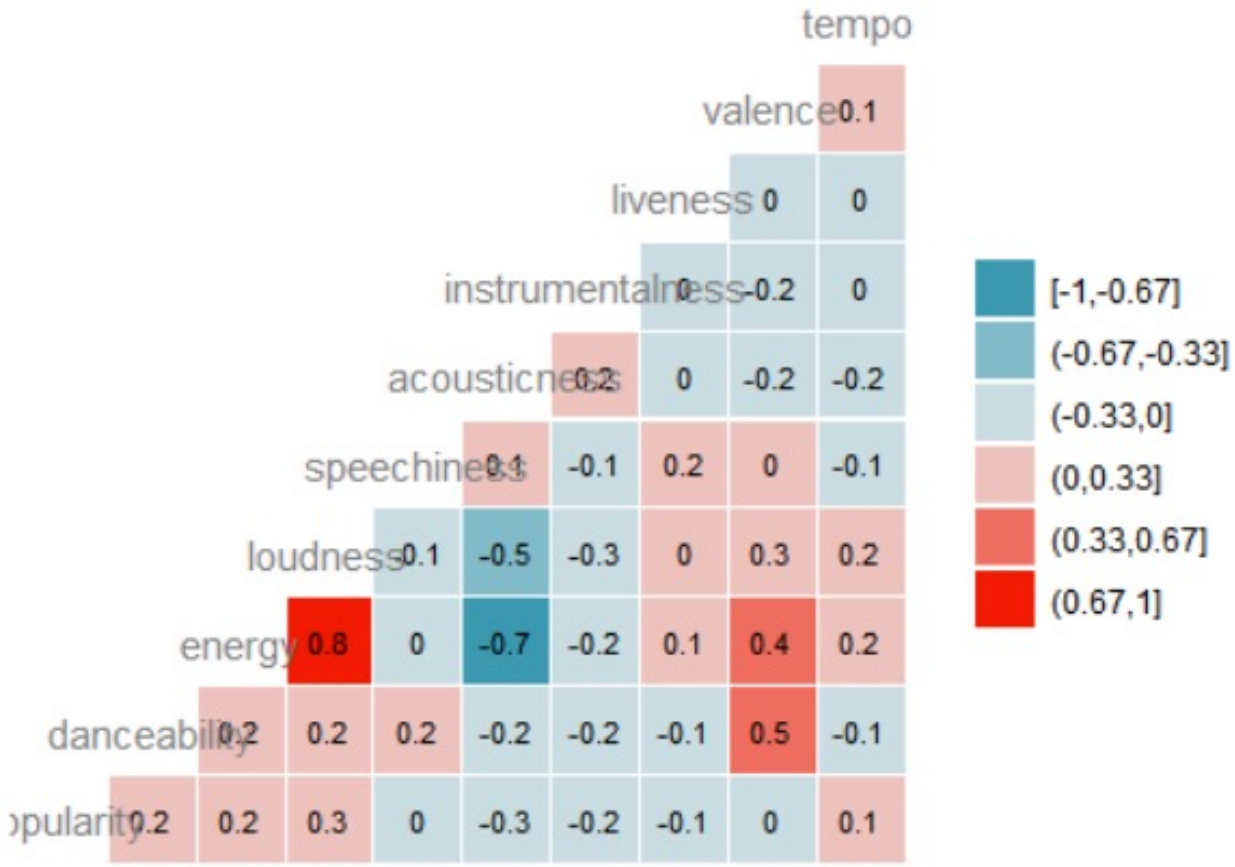

Average speechiness in each year from 1922 to 2021
speechiness
0.6
0.5
0.4
0.3
0.2
0.1
1925
1950
1975
2000
2025
Time


Average liveness in each year from 1922 to 2021
Liveness
0.250
0.225
0.200
0.175
1925
1950
1975
2000
2025
Time

Correlation

Packages:

library(dplyr)
library(GGally)



METHODS & PROCESS

In this project, we

- Clean the data
- Linear Regression
- Logistic Regression
- Stepwise Regression (gaussian, Gamma,poisson)
- Calculate the Accuracy

Clean the Data

586672 obj.,20 variables->541717 obj., 16 variables

1. Change datatype
   - Release date->release year
   - Numeric->categorical
2. Select
   - Popularity > 0
   - Tempo>0
3. Remove
   - The outlier (release_date!="1900-01-01")
   - Unrelated info (artist, name)

Result:

No unique variables

Contains: Numeric and Categorical

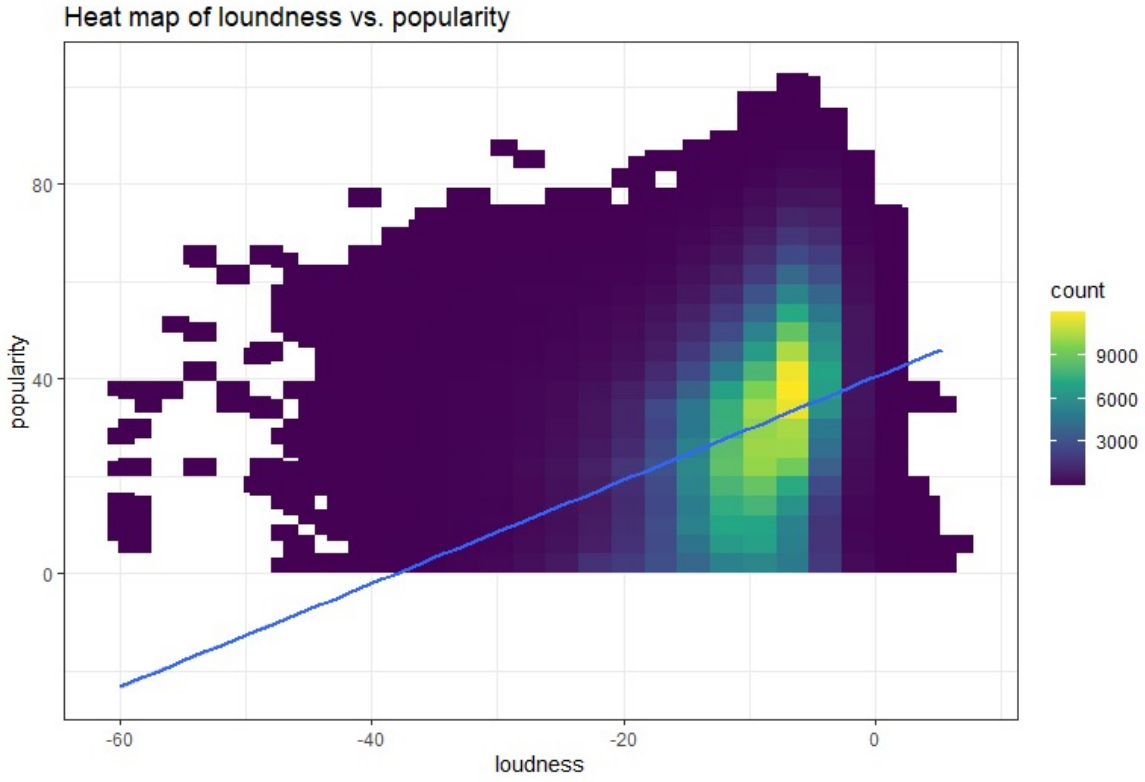
Linear Regression

Hypothesis: There is a relationship between the variable and the popularity

Null Hypothesis: There is no relationship between the variable and the popularity

Numeric variables:
duration_ms
Danceability
energy
Loudness
Speechiness
Acousticness
instrumentalness
liveness
valence
tempo
year

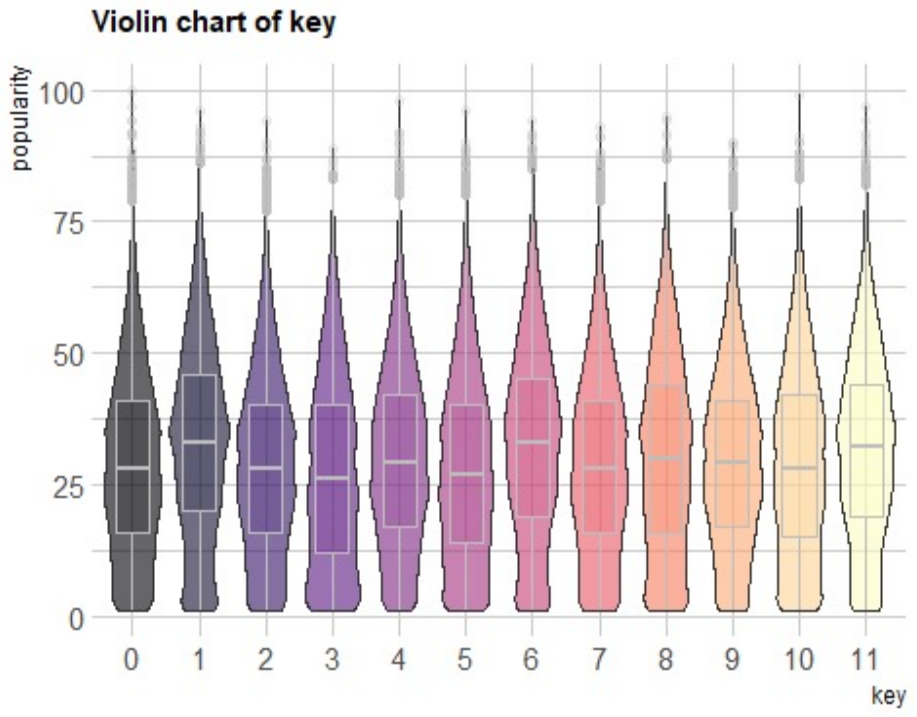
Visualization method:



ANOVA

Not every variable is numeric

Visualization:



Alternative Hypothesis: There is difference among the group

Null Hypothesis: There is no difference among the group

Numeric variables:
explicit
mode
Key
Time_signiture

Logistic Regression: Not every Variables are Gaussian

Create a new binary variable to define the popularity.

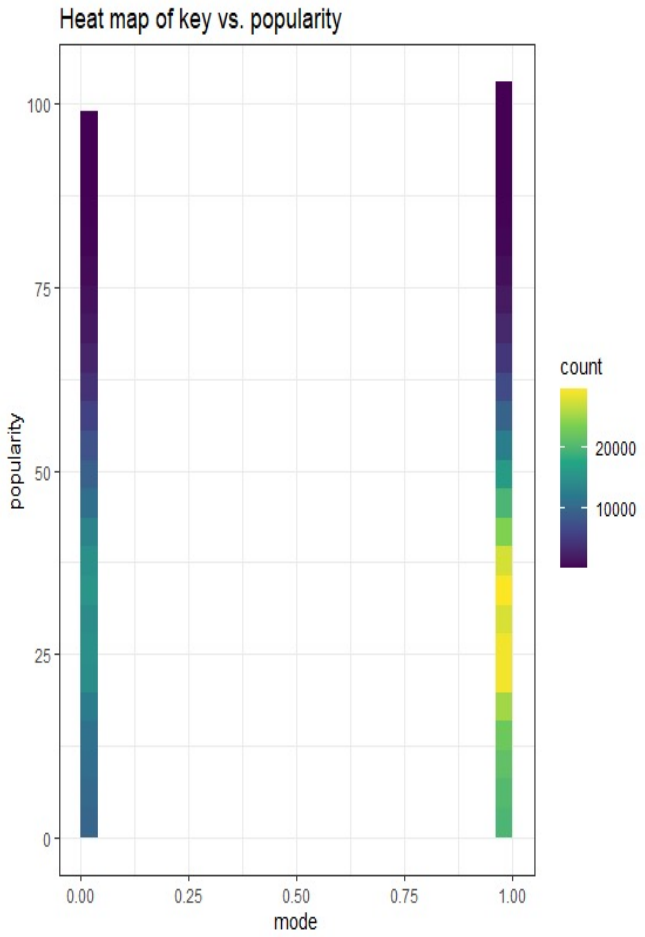
- -1 if the popularity is greater than the average
- -0 otherwise

Hypothesis: There is a relationship between the variable and the popularity

Null Hypothesis: There is no relationship between the variable and the popularity

Numeric variables:
explicit
mode
Key
Time_signiture

Visualization:



Stepwise Regression: Select the Best Model

Use the dummy code to covert the categorical variables

Steps:
Create a full model
Use step function (both directions)

In lm model, there is no relationship between release time and popularity since the coeffient ≈0

It do effect AIC in Pricewise Regression

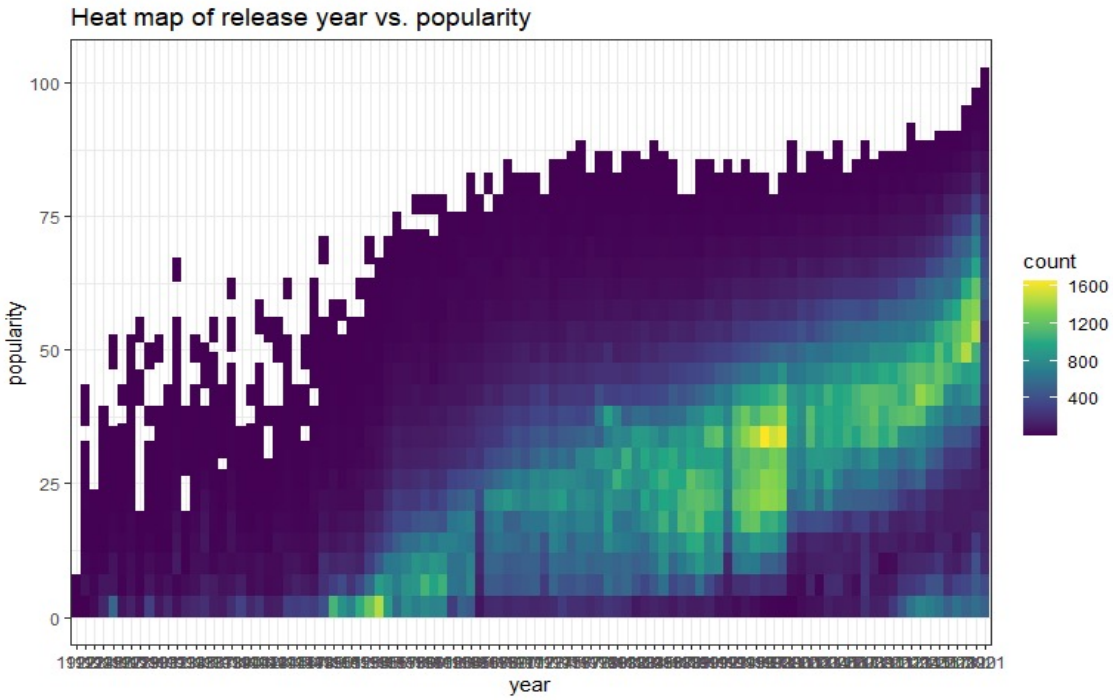
Data:
with the release year
OR
without the release year

Goal: simplify the model by find smallest AIC

LM OR GLM

Includes categorical or not

Distribution:
Gaussian? Gamma? Poission?



Quality Control

We used **autoplot** function in **ggfortify** package to check the quality of the linear model.

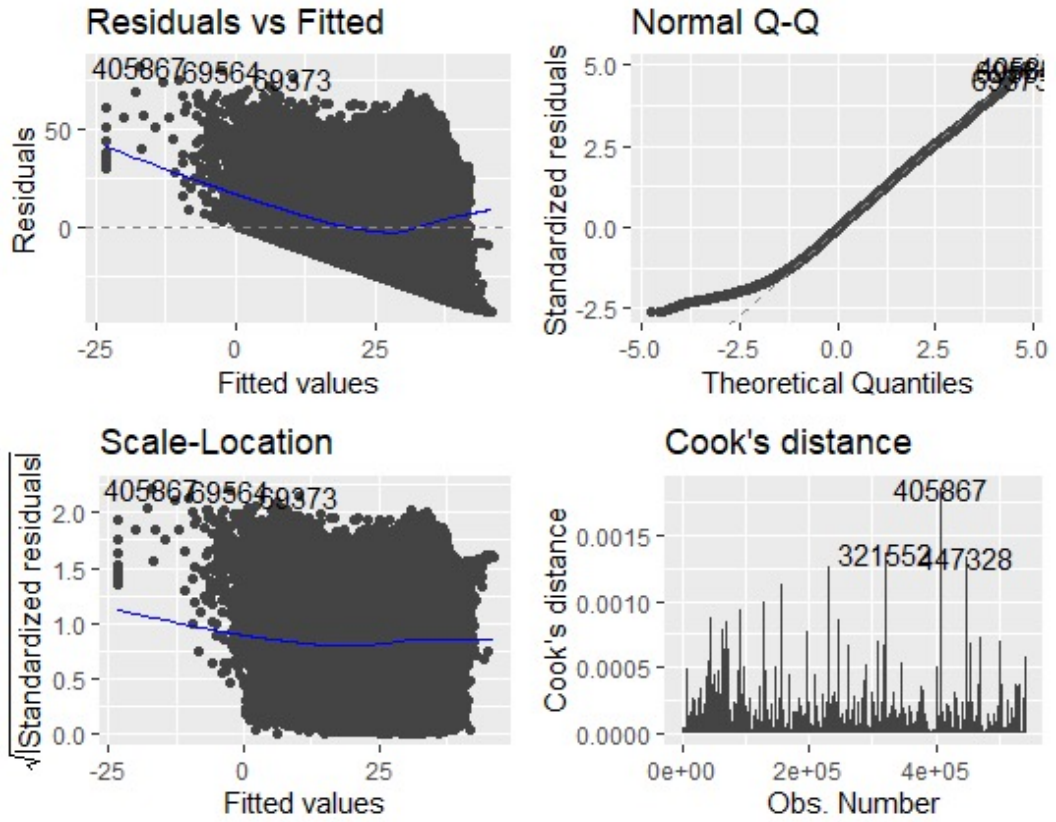
Residuals vs fitted

Q to Q: if it is normal distribution

scale location: square root of the standardized residuals vs fitted value.

Cook's distance: highlight the abnormal value

Quality of linear regression model of loudness vs. popularity



Accuracy

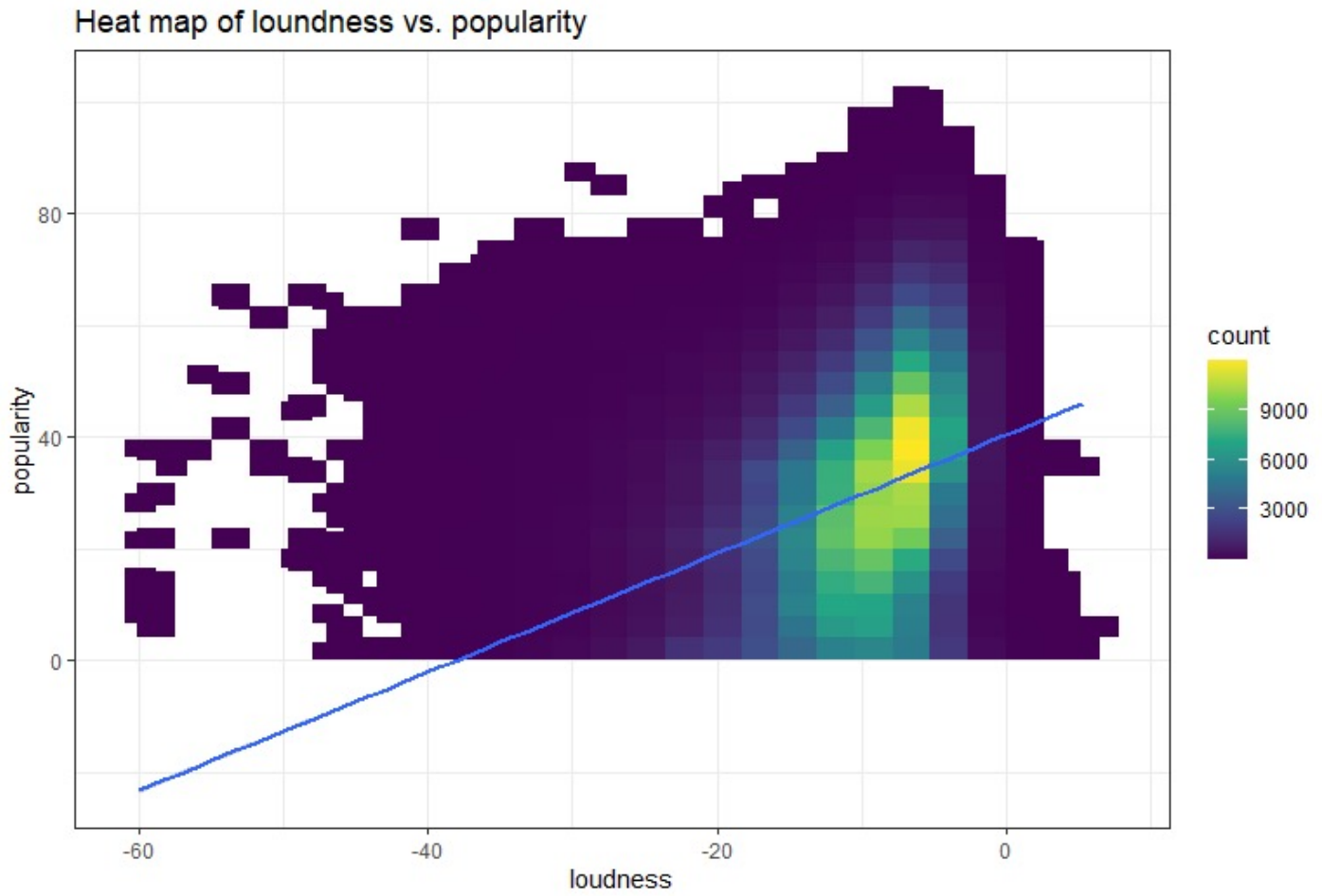
Since the response variable is numeric, so we found a package called rcompanion

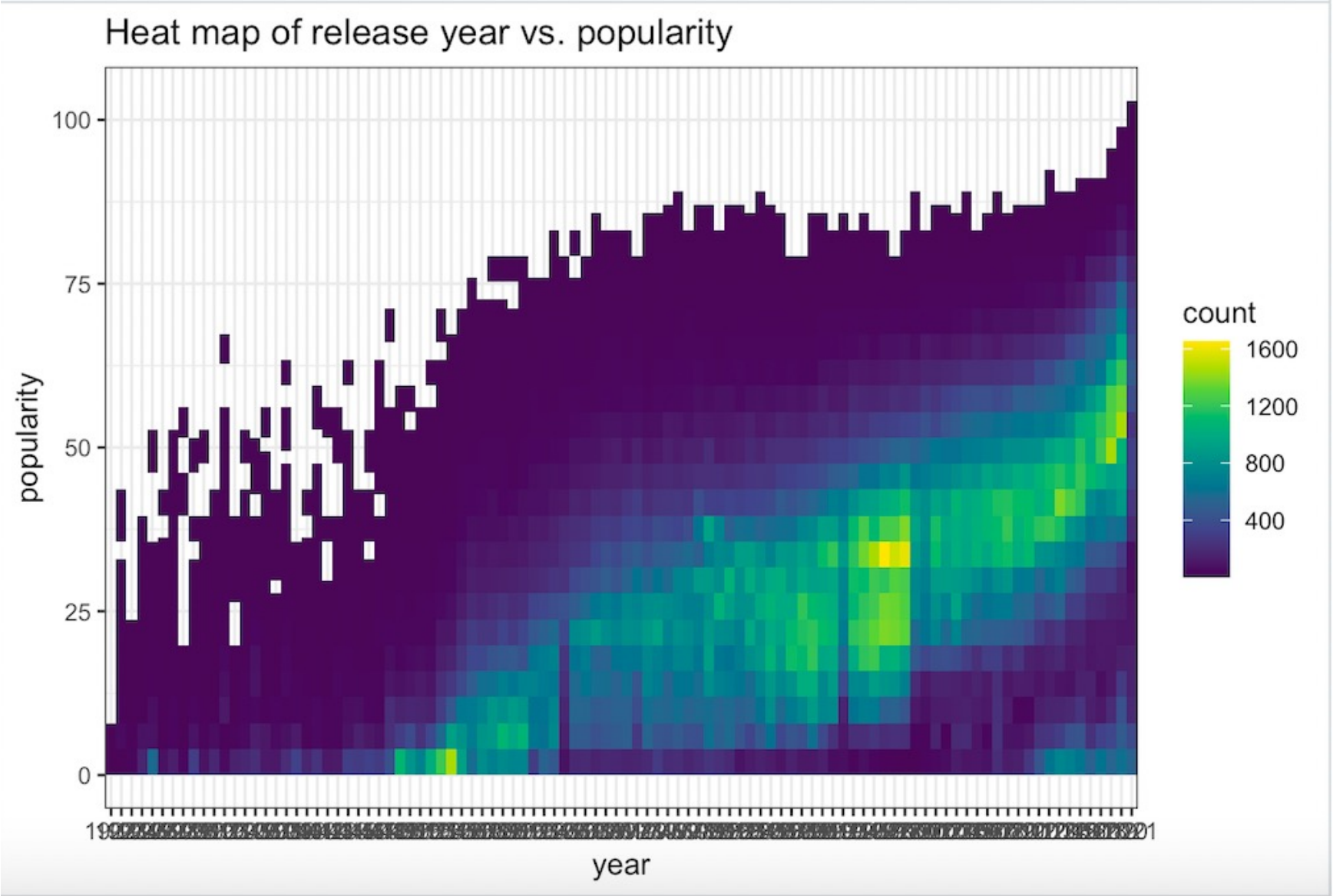
By using accuracy() function, we got:
Min-max accuracy
MSE
Efron.r.squared

Example: accuracy of full model with the release year

Model	Min.max.accuracy	**MSE**	Efron.r.squared
Gaussian GLM (LM)	0.671	200	0.329
Poisson GLM	0.672	200	0.328
Gamma GLM	0.669	207	0.305

Results


Heat map of loundness vs. popularity
popularity
80
40
0
-60
-40
-20
0
loudness
count
9000
6000
3000


Heat map of release year vs. popularity
100
75
50
25
0
popularity
year
count
1600
1200
800
400

MODEL	MIN.MAX.ACCURACY	MSE	EFRON.R.SQUARED
Gaussian GLM (LM)	0.671	200	0.329
Poisson GLM	0.672	200	0.328
Gamma GLM	0.669	207	0.305

Conclusion

Limitations:

Only numeric data is included

Low R-squared

Time period

Rooms for improvement

Problems discovered:

The step-wise regression model included most of the predictors, and the number of predictors could not be trimmed down, resulting in a bulky model.

There were correlations between the predictors, such as danceability & energy, tempo and loudness, etc.

Thank You!

References

Al-Beitawi, Z., Salehan, M., & Zhang, S. (2020). What Makes a Song Trend? Cluster Analysis of Musical Attributes for Spotify Top Trending Songs. *Journal of Marketing Development and Competitiveness*, *14*(3), 79-91.

Spotify Dataset 1921-2020, 600k+ Tracks | Kaggle
https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks?select=tracks.csv