678final Haocheng Zhu

2022-12-05

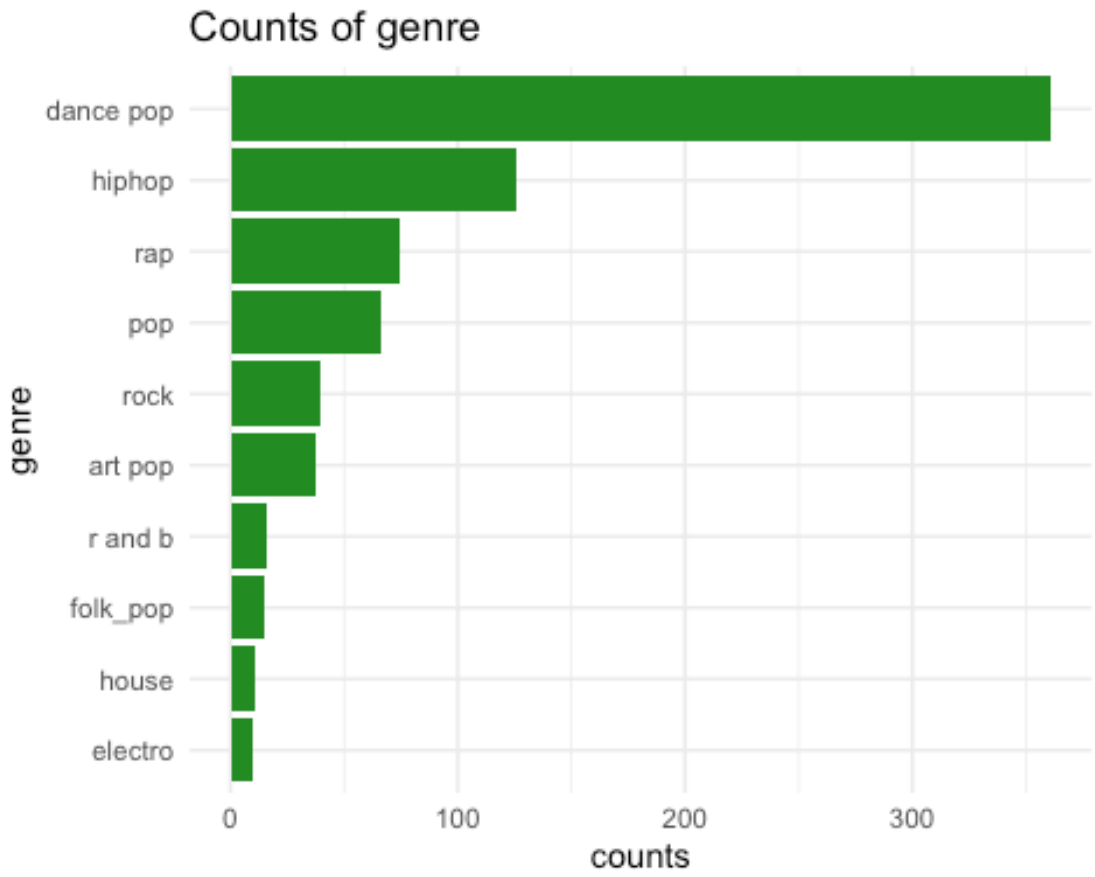
Abstract and introduction

Spotify is a digital music, podcast, and video service that gives you access to millions of songs and other content from creators all over the world. Basic functions such as playing music are totally free.The data I am using is Top 100 songs of each year on Spotify from 2010 to 2019.The data include manys variables such as “Beats Per Minute - The tempo of the song”,“Energy - How energetic the song is”,“Danceability - How easy it is to dance to the song” and so on.What I am trying to find out is the relationship of each variables and focus on how these variables affect the Popularity of the song (not a ranking).

Column Name	Column Description
title	Song's Title
artist	Song's artist
genre	Genre of song
year released	Year the song was released
added	Day song was added to Spotify's Top Hits playlist
bpm	Beats Per Minute - The tempo of the song
nrgy	Energy - How energetic the song is
dnce	Danceability - How easy it is to dance to the song
dB	Decibel - How loud the song is
live	How likely the song is a live recording
val	How positive the mood of the song is
dur	Duration of the song
acous	How acoustic the song is
spch	The more the song is focused on spoken word
pop	Popularity of the song (not a ranking)

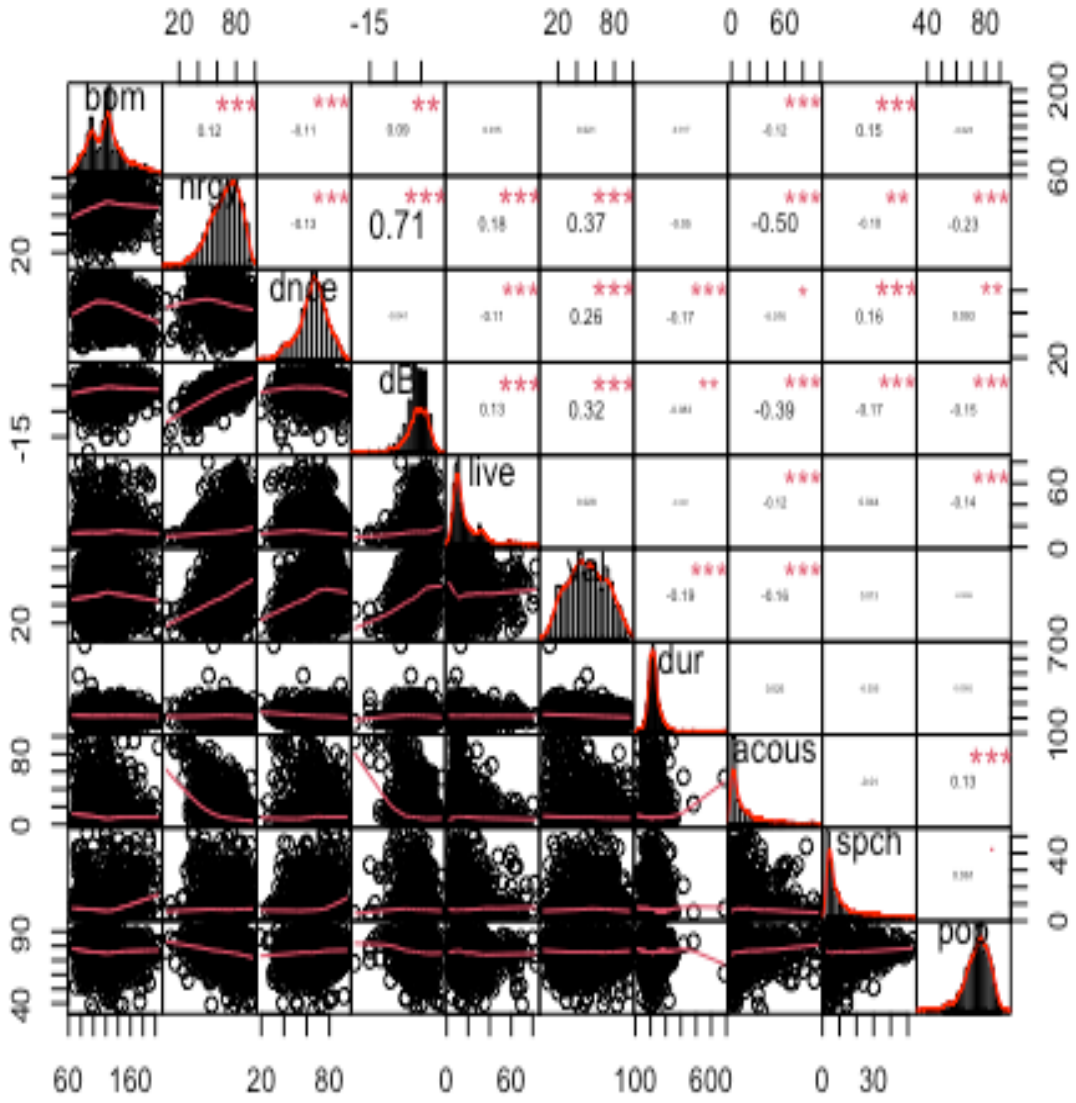
Column Name	Column Description
top year	Year the song was a top hit
artist type	Tells if artist is solo, duo, trio, or a band

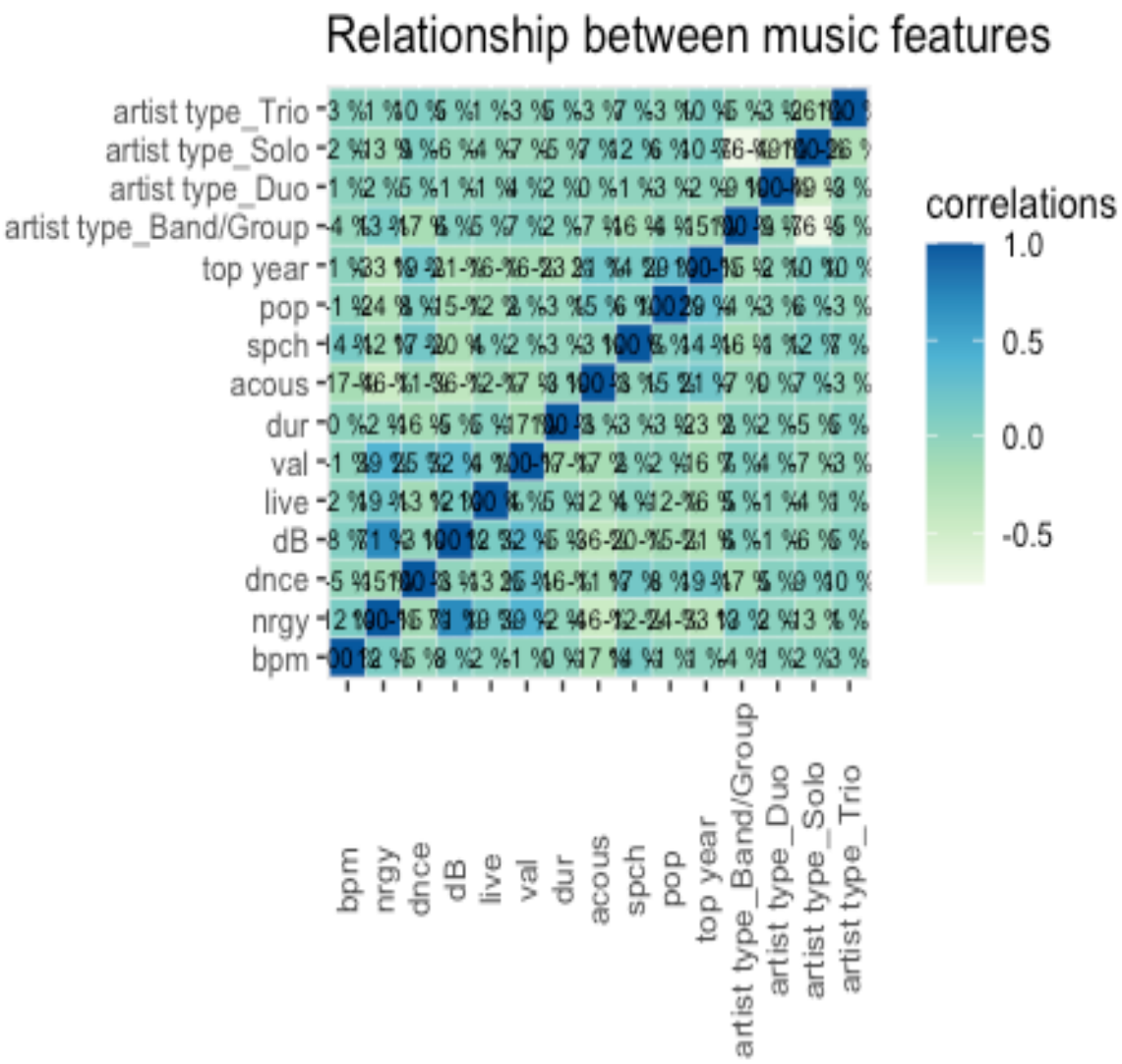
New data distribution



Because there are so many genre that only appear once or twice, so I tide up a new data to see the general distribution of the most appeared music genre. Form the graph we can see that The dance pop take place most of the music genre.

relationship of each variable





From the graph we can conclude that:

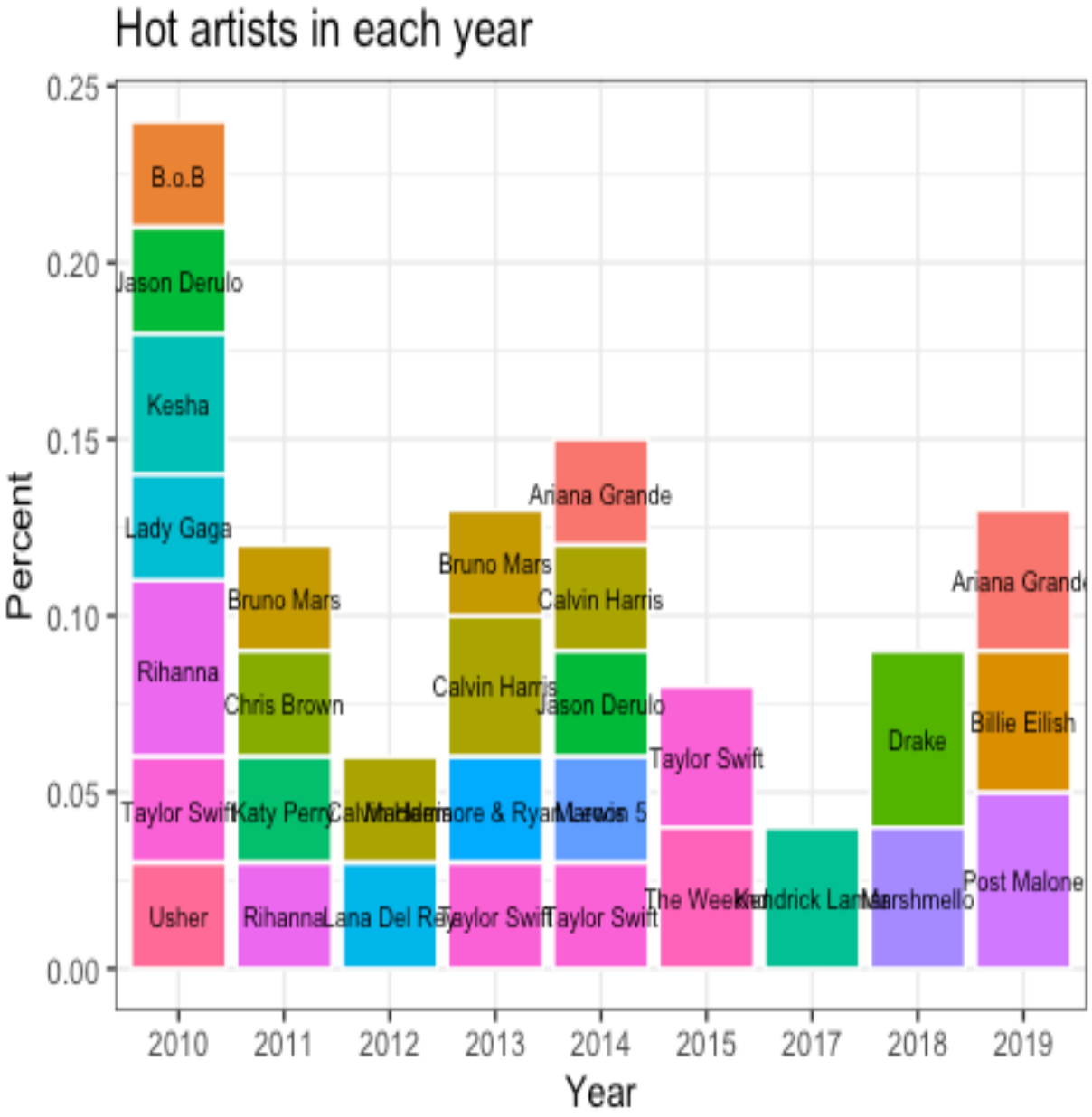
1.Decibel (How loud the song is), Energy are highly correlated

2.Energy, Val (How positive the mood of the song is) are seldom correlated

3.Acoustic (How acoustic the song is) is largely negatively correlated to Energy, Decibel

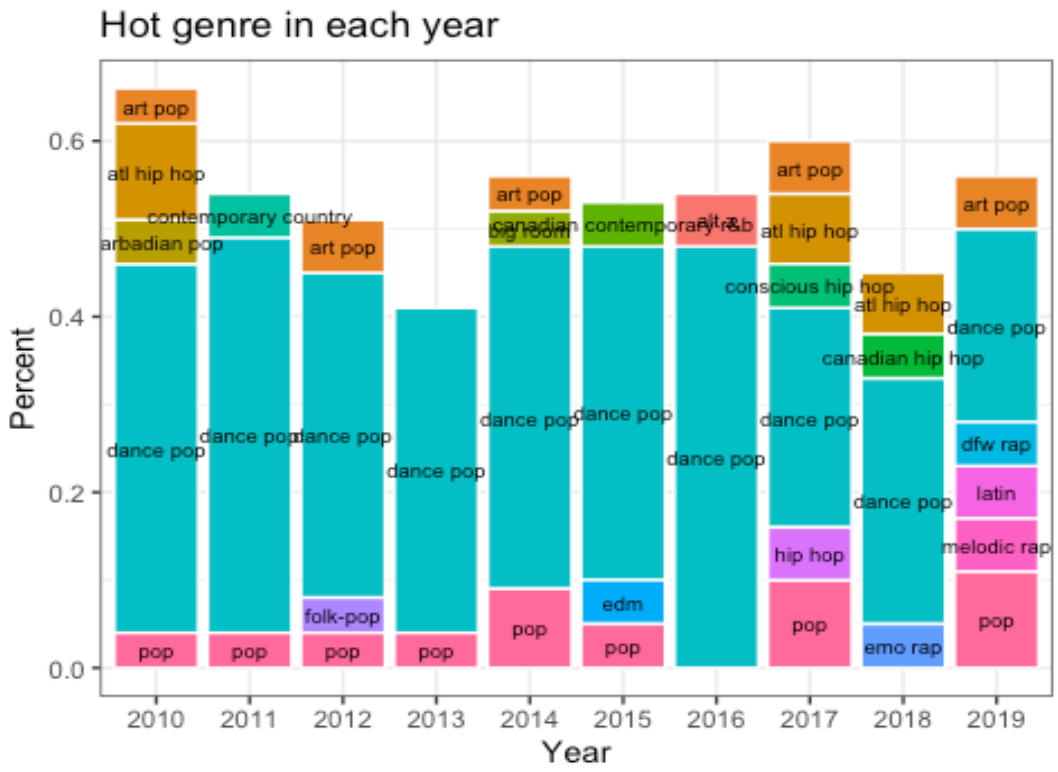
Acoustic, solo and Danceability is highly correlated with the Popularity of the song, it can inferred that more Danceability, Acoustic, artistic type is Solo and Trio are more popular. However nrgy has a negative affect with Popularity of the song.

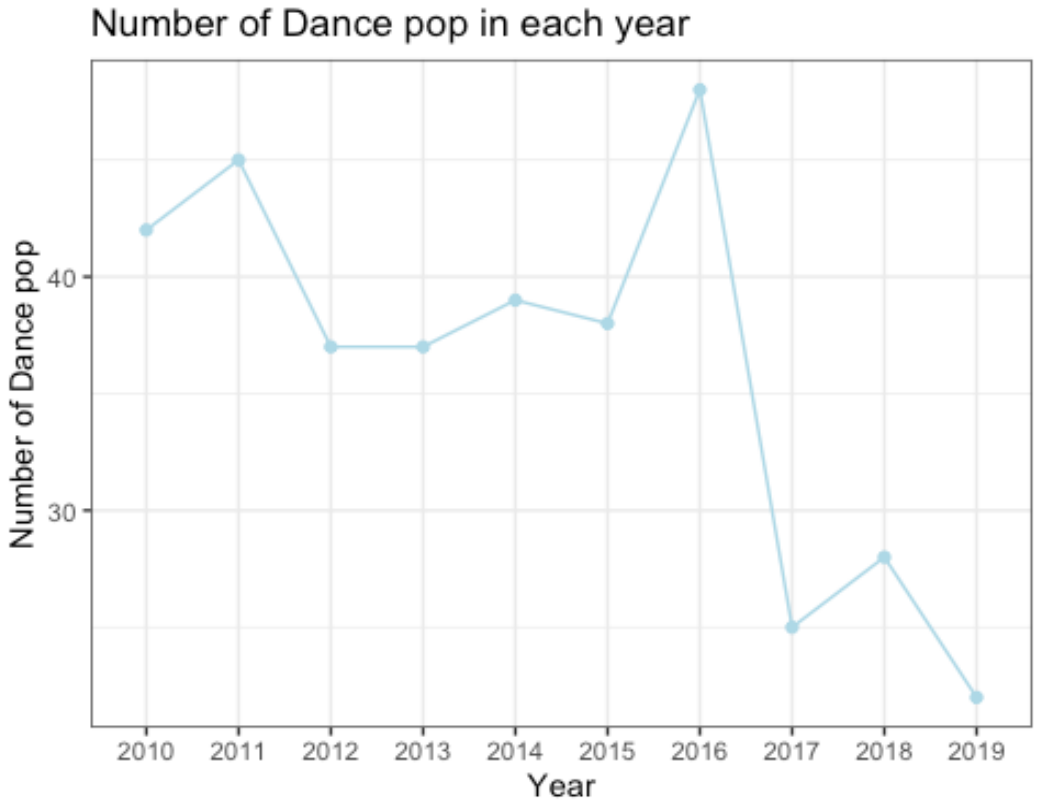
Top artist in each year



From the graph we can conclude that Ariana Grande, Post Malone, Billie Eilish are the top singer in recent year.

Top genre in each year





Dance pop is the most genres in every year, but the proportion is decreasing in recent year Latin, metro rap, rap are getting more popular in recent years

Fit a regression model

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: pop ~ nrgy + dnce + bpm + acous + (1 | type)
##    Data: new_data
##
## REML criterion at convergence: 5341.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.5378 -0.5528  0.0701  0.6975  2.3966
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  type     (Intercept) 23.96    4.895
##  Residual             64.61    8.038
## Number of obs: 756, groups:  type, 10
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  75.495292   3.398781 120.554657  22.212  < 2e-16 ***
## nrgy         -0.094477   0.022553 745.413246  -4.189 3.14e-05 ***
## dnce          0.050579   0.024391 749.288003   2.074   0.0385 *
## bpm           0.009065   0.011670 742.894547   0.777   0.4375
## acous         0.038743   0.019278 747.875472   2.010   0.0448 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) nrgy   dnce   bpm
## nrgy  -0.554
## dnce  -0.615  0.176
## bpm   -0.444 -0.052  0.086
## acous -0.443  0.449  0.185  0.129
```

I choose nrgy, dnce, bpm, acousas variables and try to see hows these variables affect the pop level. The reason I choose these three as my variables is because these three variables are the most co-relative to the pop level as I shows before.

For fixed effects: nrgy for every unit increase in the number of nrgy, negative affect is expected to decrease by 0.115.

For random effects: There are intergroup differences in popular level among music genre in different type (23.31).

And as the results, we can see nrgy is the most meaningful value when I fit these three variables into the multilevel linear mixed model because the only the P-value of nrgy is smaller than 0.05.

Conclusion

Dance pop is the most genres in every year, but the proportion is decreasing in recent year Latin, metro rap, rap are getting more popular in recent years. Acoustic, solo and Danceability have a positive affect with the Popularity of the song. However nrgy has a negative affect with Popularity of the song. Ariana Grande, Post Malone, Billie Eilish are the top singer in recent year.

Discussion

After the data exploring and analysis, I found the relationship of each variable and focus on how these variables affect the Popularity of the song. The next step is trying to use these results and data to predict the most popular song in the future years.

Reference Discussion

https://www.kaggle.com/datasets/muhmores/spotify-top-100-songs-of-20152019?select=Spotify+2010+-+2019+Top+100.csv
https://www.kaggle.com/code/teresawu726/spotify-top-100-songs-analysis-by-r#4.-Predict-the-top-song-in-future-year
https://www.kaggle.com/datasets/muhmores/spotify-top-100-songs-of-20152019