Spotify Data Analysis

Aoyi Li

12/10/2020

Abstract

In the recent decades, more and more people love to listen spotify songs. As a result, an increasing number of data analysts show their interests in analyzing spotify data. Their analyses are mainly about the popularity, danceability, mood prediction, and time series and genre analyses. I tried one linear model and three multilevel models to find the best model to fit the data. And the results show that the last multilevel model, which has random effect genre and different coefficients for each year point, is the best model, although all results are kind of similar for all four models.

Introduction

Listening to spotify songs during the leisure time is a popular way to get relaxed when people are under pressure. Thus, the music platform operator might be interest to know what kind of songs are most popular among people and what is the overall trend as time goes by. I am a huge music lover, thus I am interested in analyzing the spotify data. In this report, I plan to use the data to analyze the popularity for different types of songs, finding what factors will significantly related to the popularity scores. Since I will use popularity, a numerical variable, as my outcome, and the data can be grouped by genre, I can use simple linear model and multilevel linear mixed effect model to find factors that will might affect popularity and to investigate the relationship between necessary variables and genres. I will analyze data from Spotify Dataset 1921-2020, 160+ Tracks from Kaggle Open Datasets website. My focus will be the effect of genre, and I might show some trend plots based on year.

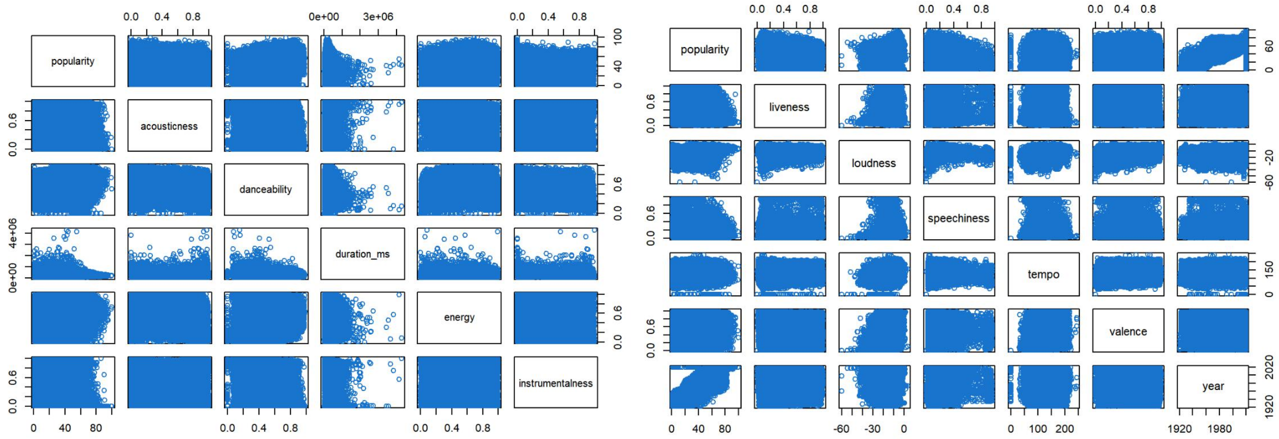
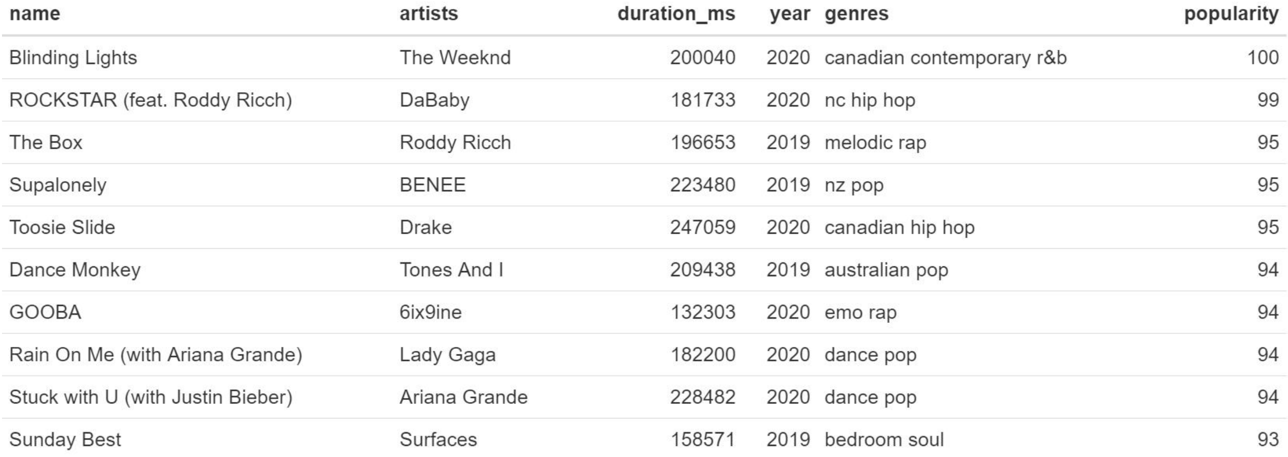


Figure 1&2: The initial matrix plot for the pairwise relationship between each of the 2 numerical variables. There are some clear patterns between some of the variables, but most of them shows dense points because of there are more than 130 thousands data points.

Methods

Data Cleaning and Selection

The dataset, Spotify Dataset 1921-2020, 160+ Tracks, from Kaggle has 5 datasets including data(the main dataset), data_by_artist, data_by_genres, data_by_year and data_w_genres. Most of the variables in the dateset are the characteristics of songs, for example, acousticness means the higher the value, the more acoustic the song is, and duration_ms shows the duration of the song in millisecond. Since I would like to analyze the data with both year and genres but the main dataset "data" contains all variables expect genre, I firstly merged "data_w_genres" and "data" together to get my final dataset. I used artist's name to join 2 dataset, thus I wrote a new .csv file call "genres.csv" which including only 2 columns with artists' names and genres. Then I filtered out all rows contain NA values and used the dplyr package in R to join "data" and genres" by artist's name, writing the merged dataset into a .csv file called "merge.data.csv". After that, I deleted some unnecessary columns such as some id columns, added a new variable called decade based on year, and wrote different data frames with various subset of variables for different analysis purposes. As the result, the final main dataset included 134673 observations and 18 variables. For the modeling part, I randomly selected 20% of the 134673 observations to do the regression analysis to make the process less time consuming.



name	artists	duration_ms	year	genres	popularity
Blinding Lights	The Weeknd	200040	2020	canadian contemporary r&b	100
ROCKSTAR (feat. Roddy Ricch)	DaBaby	181733	2020	nc hip hop	99
The Box	Roddy Ricch	196653	2019	melodic rap	95
Supalonely	BENEE	223480	2019	nz pop	95
Toosie Slide	Drake	247059	2020	canadian hip hop	95
Dance Monkey	Tones And I	209438	2019	australian pop	94
GOOBA	6ix9ine	132303	2020	emo rap	94
Rain On Me (with Ariana Grande)	Lady Gaga	182200	2020	dance pop	94
Stuck with U (with Justin Bieber)	Ariana Grande	228482	2020	dance pop	94
Sunday Best	Surfaces	158571	2019	bedroom soul	93

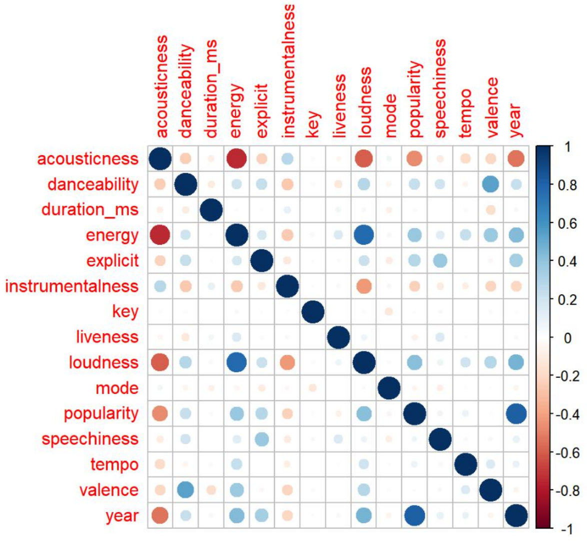
Figure 3: Top 10 songs based on popularity. This table includes basic songs' information and their popularity with both year and genres, which means I successfully merged 2 datasets.

Model selection and transformation

Since popularity is numerical and kind of normally distributed, I firstly tried a simple linear regression to check the model fit. Then I decided to try several multilevel models in order to get a best fit model. In the linear model, I did variable selection, possible transformation and added interaction term to get a final model. Then I tried multilevel models with random effect, varying slope and different coefficients including same variables selected from linear model, using the anova test and comparing residual plots and MSE to find a best model among all four models. Finally, I used a function called check_model() and graph a predicted vs. Actual plot to check my chosen model.

Modeling Process Results

Since I would like to find factors that will affect popularity, I used correlation matrix and basic exploratory data analysis method to get the basic information about the relationships between variables, in order to have the initial idea for variable selection.



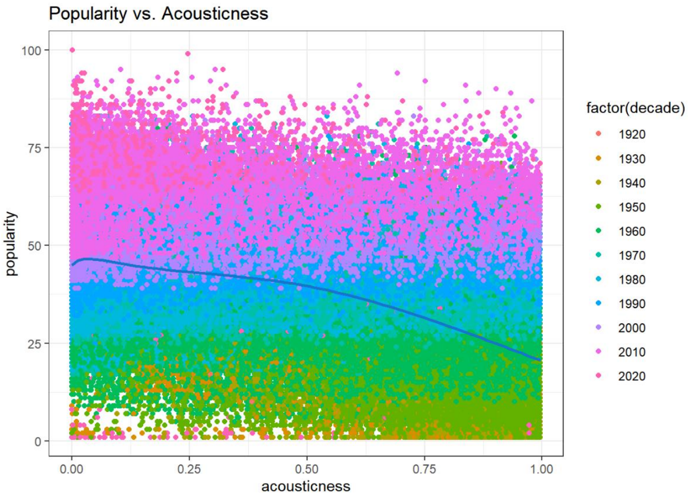


Figure 4: The correlation matrix plot for numerical variables has blue points indicate positive correlations and red points indicates negative correlations. The darker the color is, the higher the correlation for each pair of variables is.

Figure 5: Popularity vs. Acousticness by decade. Based on the correlation matrix, there seems to be a quite large correlation between popularity and acousticness. This plot shows a negative relationship between popularity and acousticness and a positive relationship between popularity and decade.

I excluded variables valence and duration_ms based on the correlation matrix plot and the significance test, and I verified my choice by using the stepwise selection method. Then I transformed popularity to a new variable tpop, tpop=popularity^0.82), by a box-cox transformation and add an interaction term to get the final linear model. However, the model checking results show that the adjusted R square improved but not much, the residual plot seems fine but still not randomly spread around 0, and the Q-Q plot shows it is left skewed.

Then I used a multilevel model with genre as the random effect to check whether it is a better model to fit the data. However, the residual plot for this model did not improve much compare to that for the final linear model. Thus, I tried other two models which allowed for varying slopes for year predictor(multilevel model 2) and with different coefficients for each year point(multilevel model 3). By graphing the residual plots, I can find a slightly improvement for the last multilevel model-different coefficients for each year point. In addition, the result of the anova test also showed that the multilevel model 3 fits the data best among all four models, but the computation results shows that multilevel model 2 has the lowest MSE which is just slightly lower than that for multilevel model 3. Consider multiple factor, I chose the multilevel model with different coefficiens for each year point as my final model. From the summary results, it states that one unit increase in acousticness is associated with 0.424 point decrease in popularity score, one unit increase in danceability is associated with 1.35 point increase in popularity score, etc,.

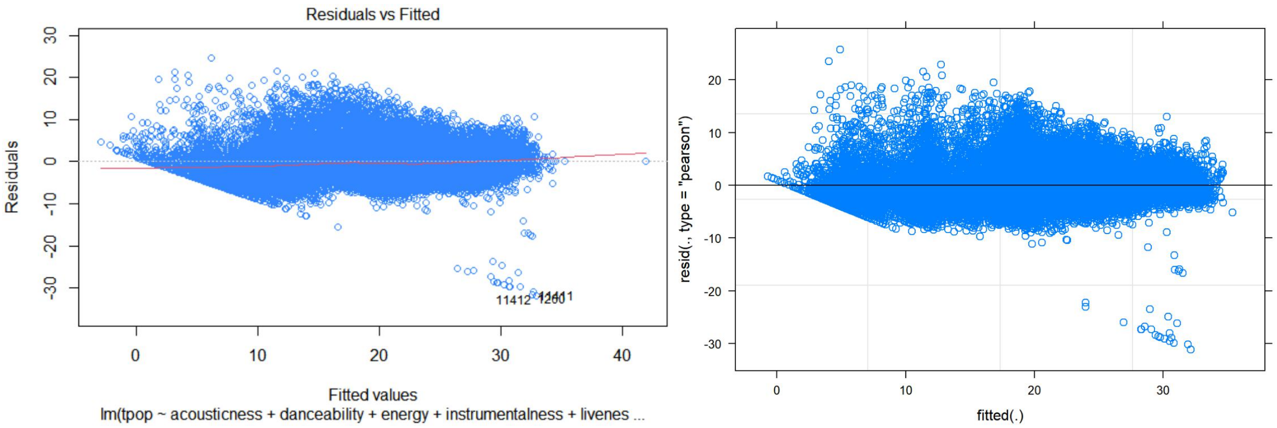


Figure 6&7: Residual plots for final linear model and the third multilevel model.

Discussion

Many of the results from my chosen model lines up with my expectation. The residual plot seems reasonable and the predicted vs. actual plot(in the appendix) shows a quite good fit, but there are probably some outliers need to be deal with in the more detailed data processing steps. The coefficients signs and variable significance are also reasonable based on the common sense.

However, I still have several concerns and the data might have some problems. Firstly, I merged 2 datasets based on artist's name and only kept the first genre as the main genre. However, an artist might write songs with different genres, thus this way of merge might face the problem that a specific song is actually belong to the second or third genre but not the main one. Thus, I think this is the main problem I need to further investigate and find a more precise way to merge datasets. Secondly, I noticed several outliers using box plot and from some result scatter plots. However, I cannot delete them directly in the data cleaning part just based on these simple detection, because not all outliers should be delete without further investigation. In addition, from the exploratory data analysis, I cut the song duration into four groups. The plots related to duration show that the song duration between 3.6 and 4.5 minutes might have a higher popularity score than others, and the duration under 2.9 minutes seems to have the lowest popularity. Although, both correlation matrix and linear regression state that the variable duration_ms is not significantly related to popularity under linear assumption, there might be some relationships that need further research. Finally, I divided the dataset into training and test subsets in order to check the prediction. Since the dataset can be treated as time series data and genre has over 900 groups, I did not find a way to get the training and test subsets with both random year and same genre level. In the future research, I might investigate deeply in these listed problems.

Appendix

Data visualization with popularity

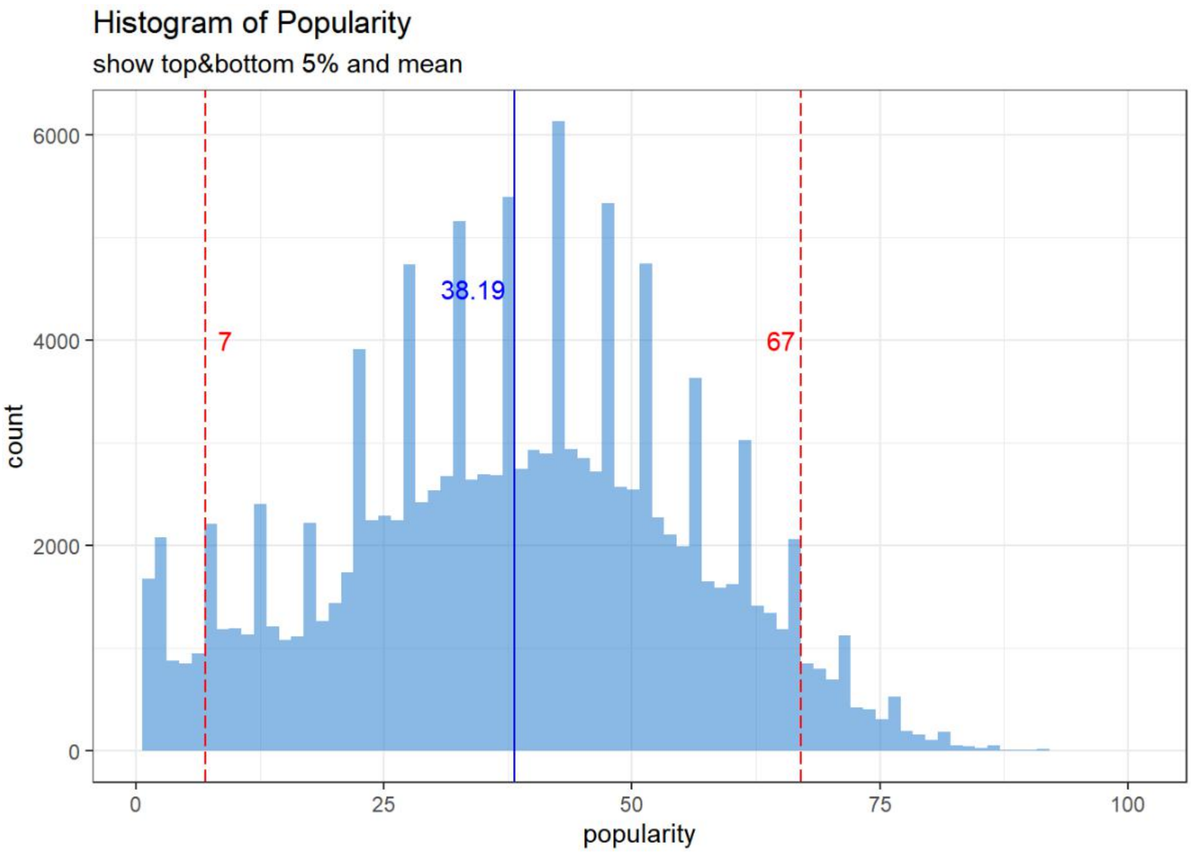


Figure 8: A histogram of popularity. The red dashed lines shows top and bottom 5% values, and the blue line shows the mean popularity.

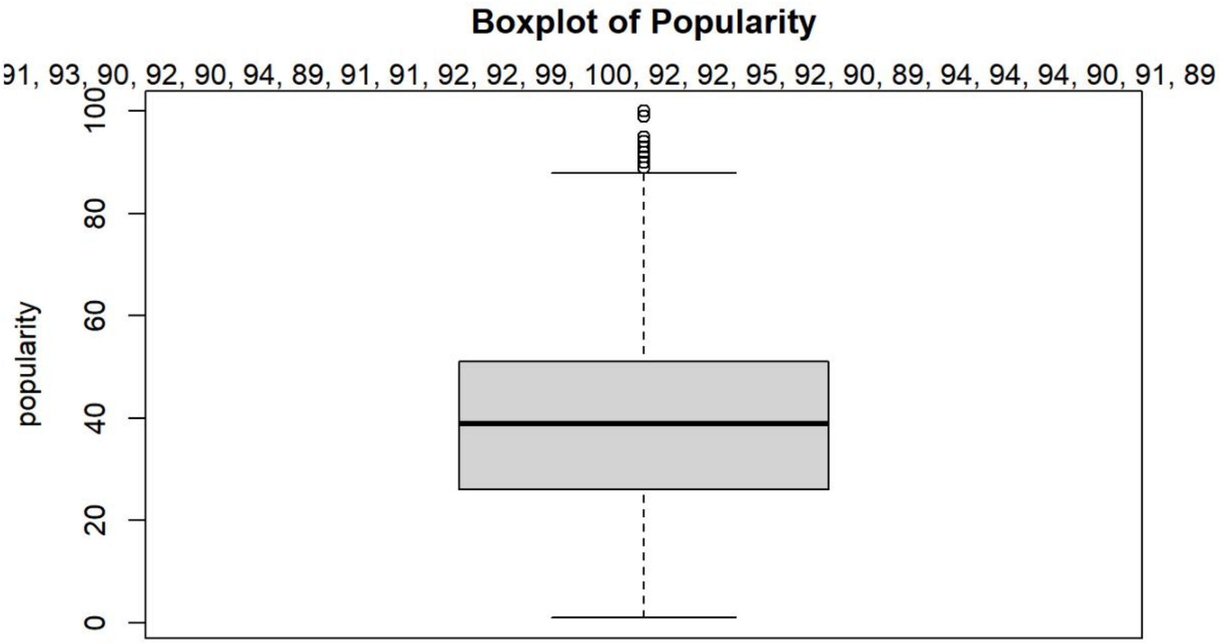


Figure 9: Box-plot for population with detected possible outliers.

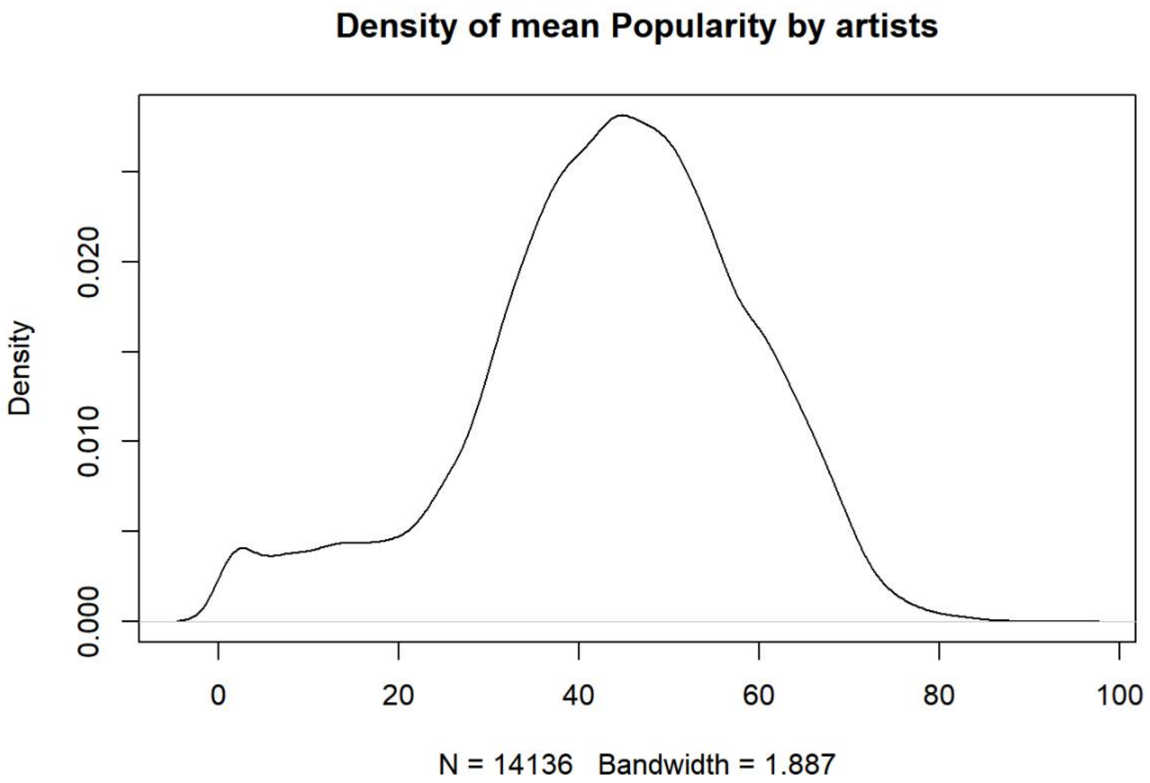


Figure 10: Density of mean popularity by different artists, which seems normally distributed.

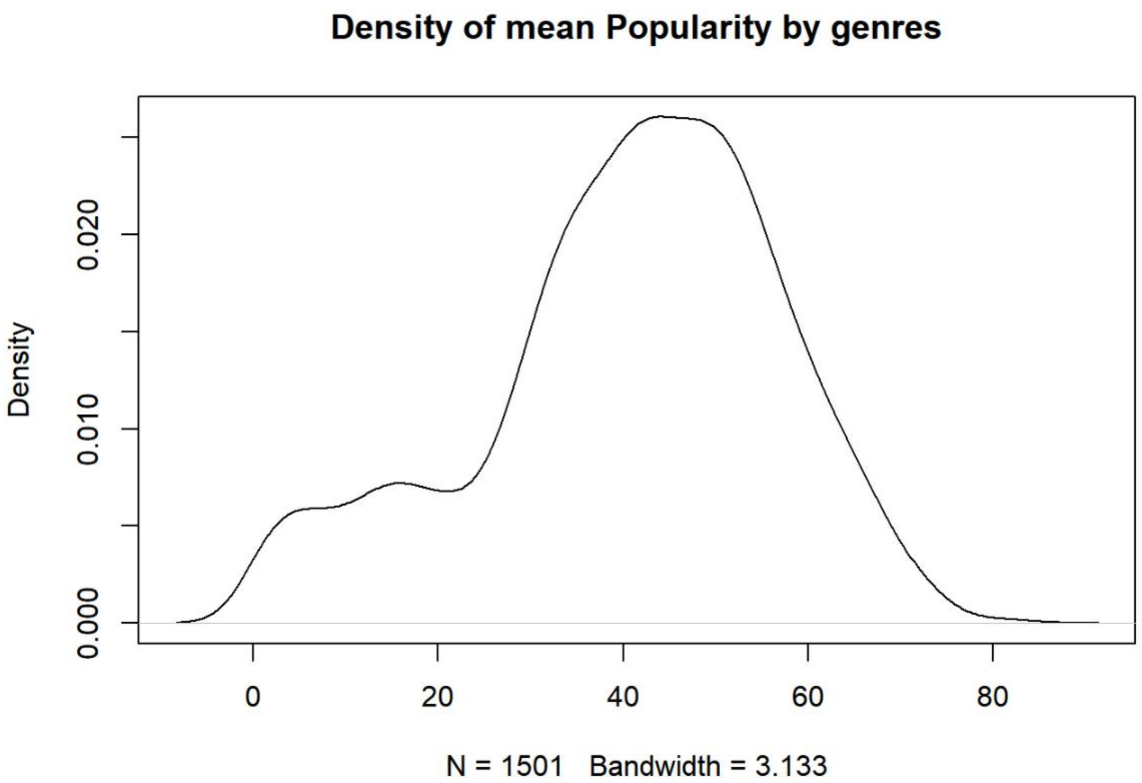


Figure 11: Density of mean popularity by different genres, which seems normally distributed.

Relationships with popularity

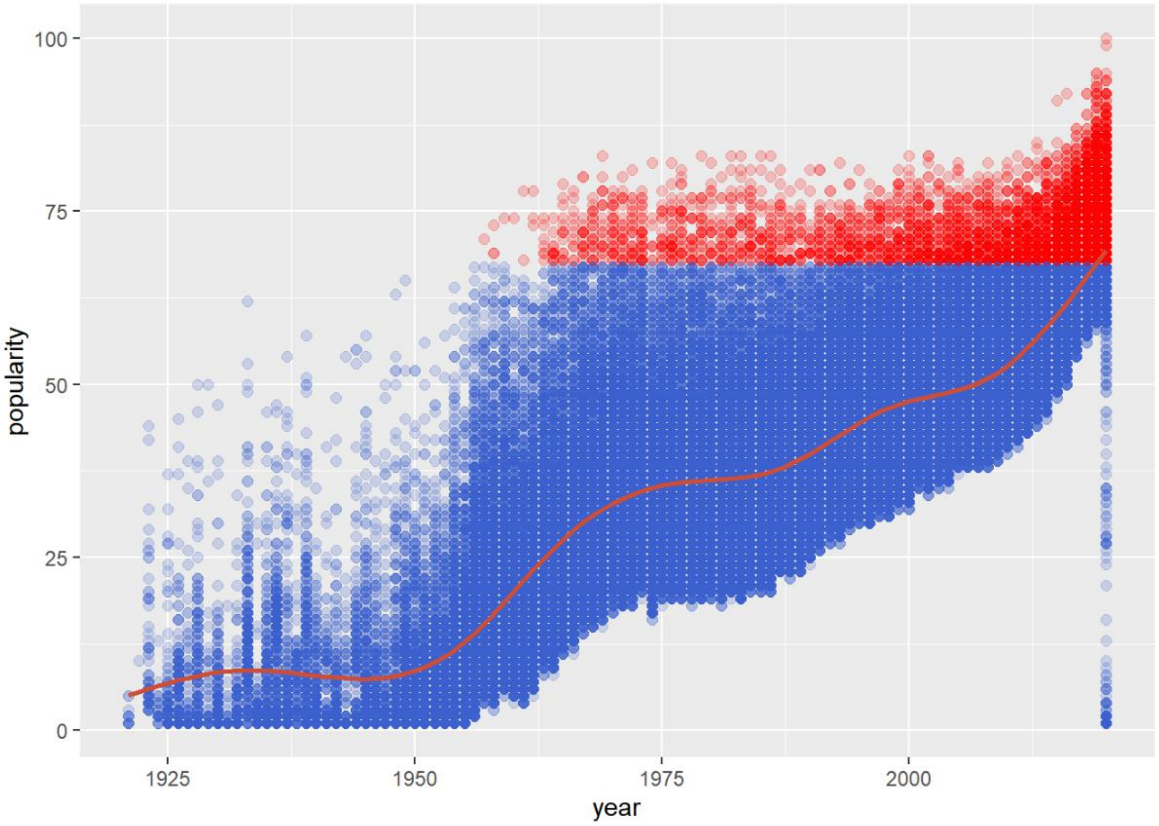


Figure 11: Relationship between popularity and year. It shows a clear positive relationship, and the red points represent the popularity grater than 67 which is the top 5% of popularity scores.

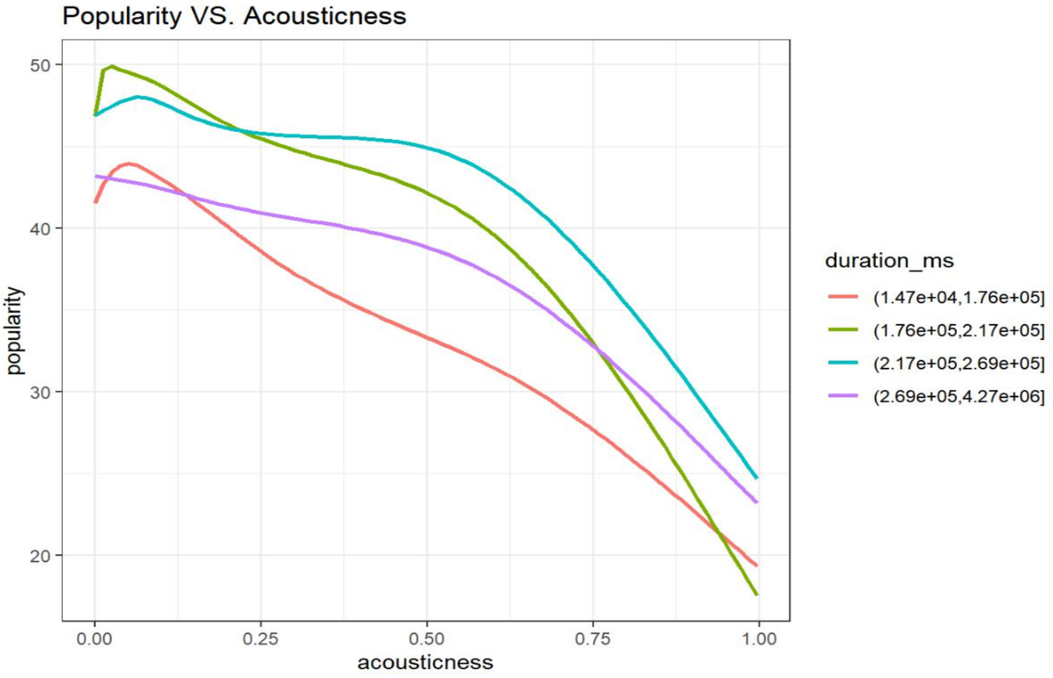


Figure 12: Popularity vs. Acousticness by duration. It clearly shows that there is a negative relationship between acousticness and popularity. The song duration between 3.6 and 4.5 minutes might have a higher popularity score than others, and the duration under 2.9 minutes seems to have the lowest popularity.

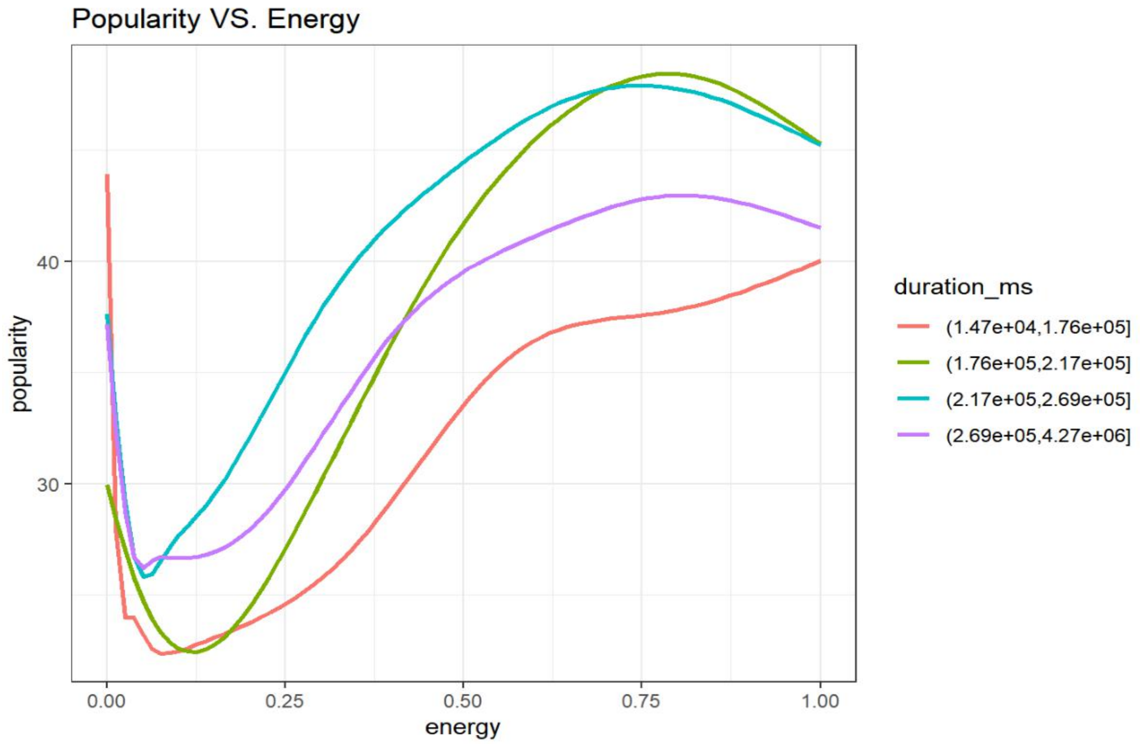
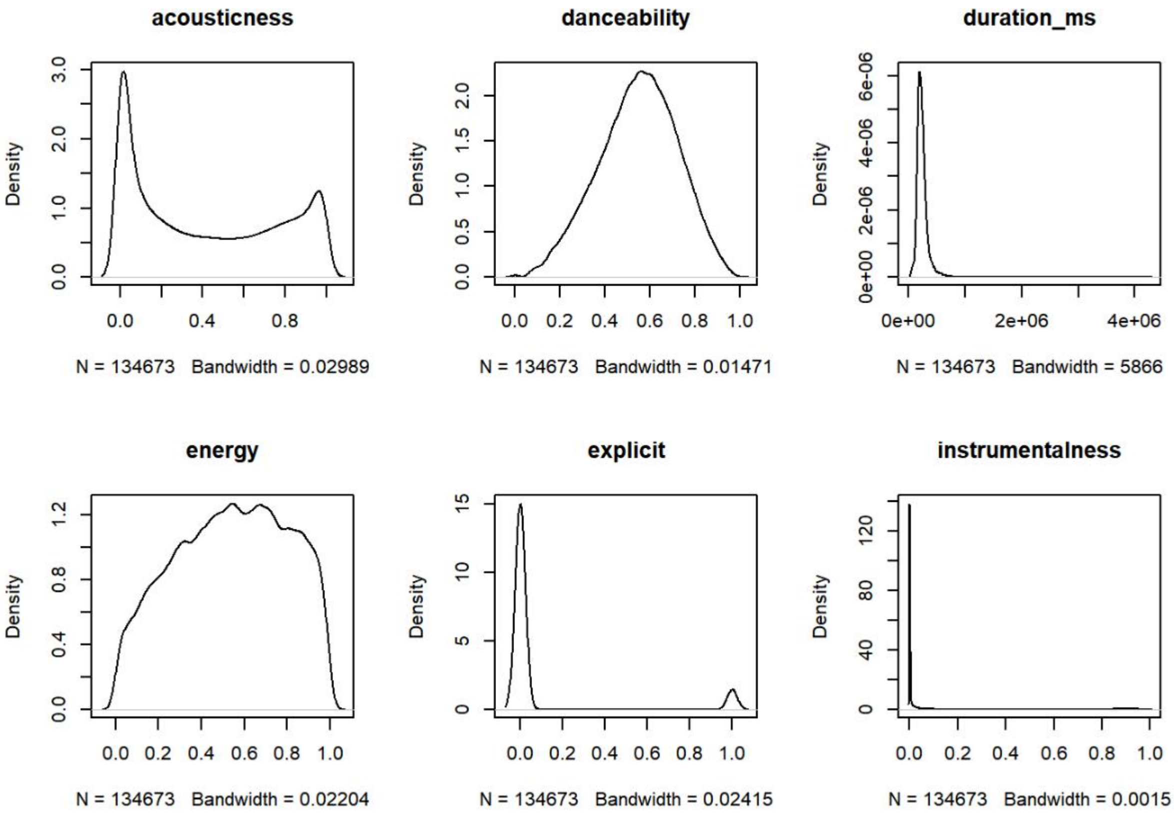


Figure 13: Popularity vs. Energy by duration. It shows that there is a positive relationship between energy and popularity when energy value greater than about 0.1. The song duration between 3.6 and 4.5 minutes might have a higher popularity score than others, and the duration under 2.9 minutes seems to have the lowest popularity.



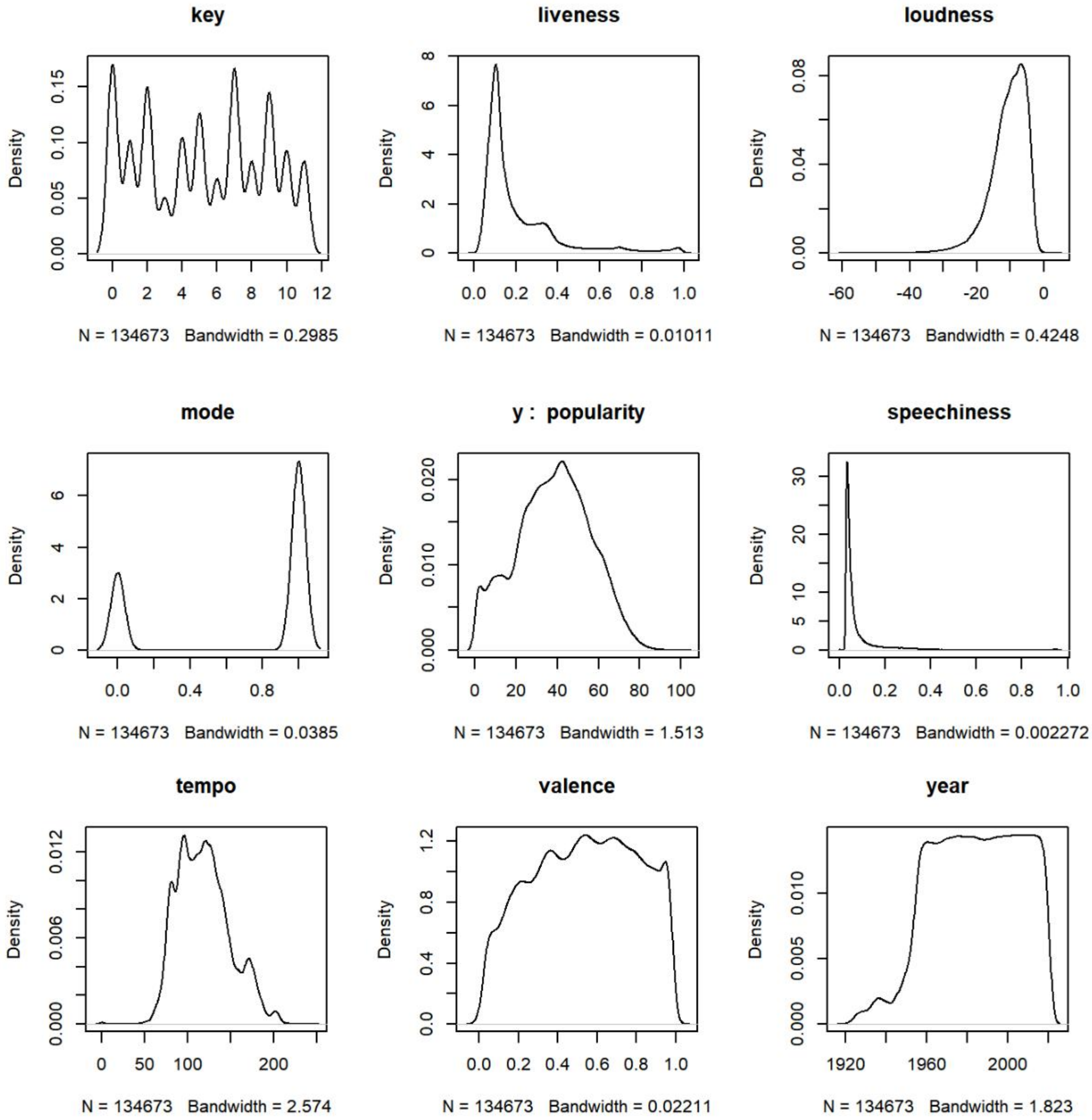


Figure 14: The density plots for each of the numerical variables. It can be helpful when checking the variables' characteristics.

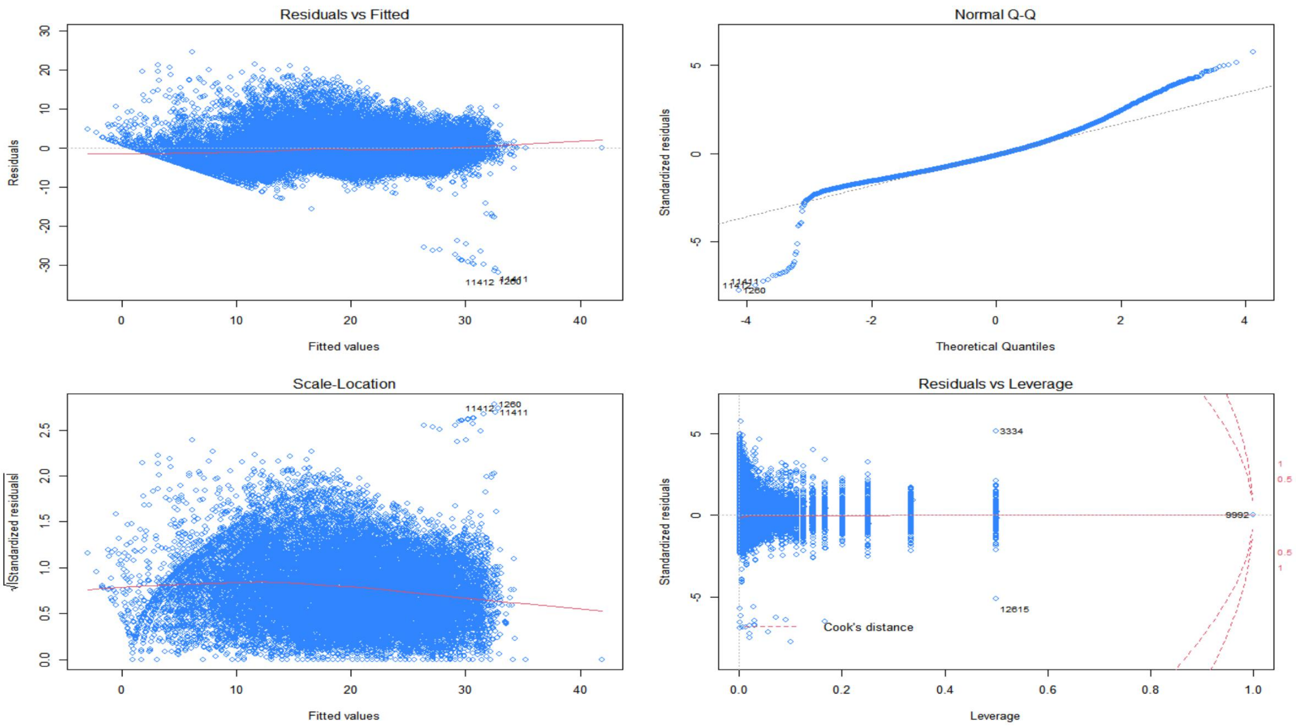


Figure 15: The diagnostic plot for the chosen final linear model. The results are reasonable, but the model does not fit very well.

*tpop~acousticness+danceability+energy+instrumentalness+liveness+loudness+speechiness+year+factor(genres)+acousticness*energy*

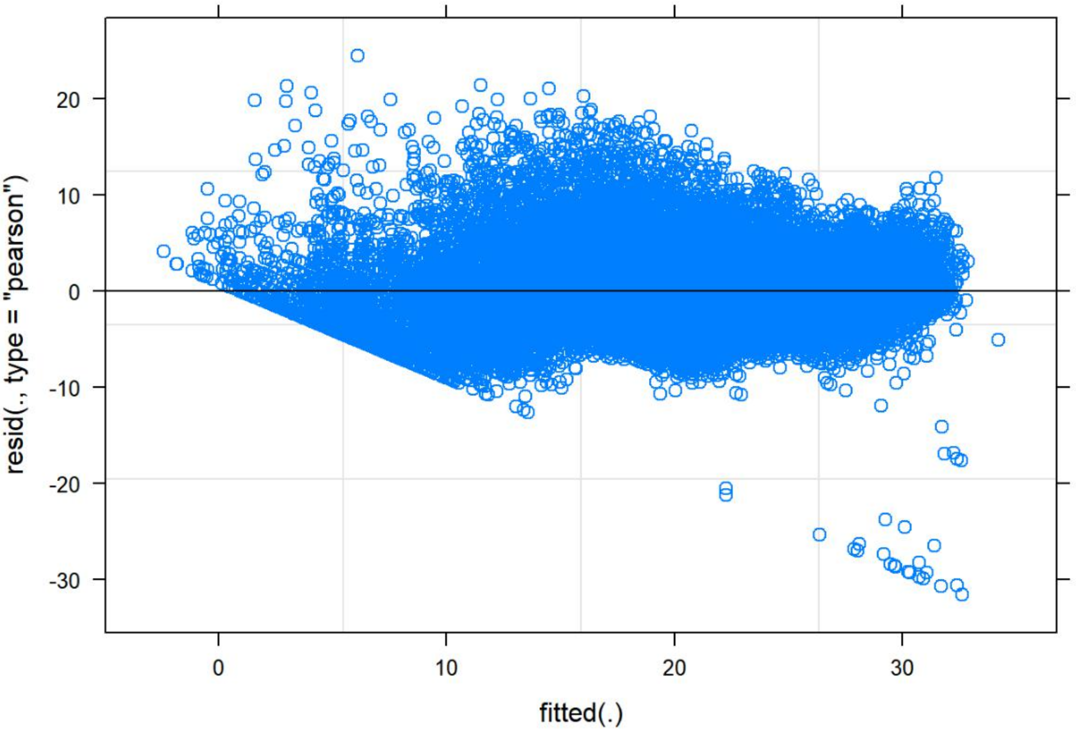


Figure 16: The residual plot for the first multilevel model with genre as the random effect.

*tpop~acousticness+danceability+energy+instrumentalness+liveness+loudness+speechiness+year+acousticness*energy +(1|genres)*

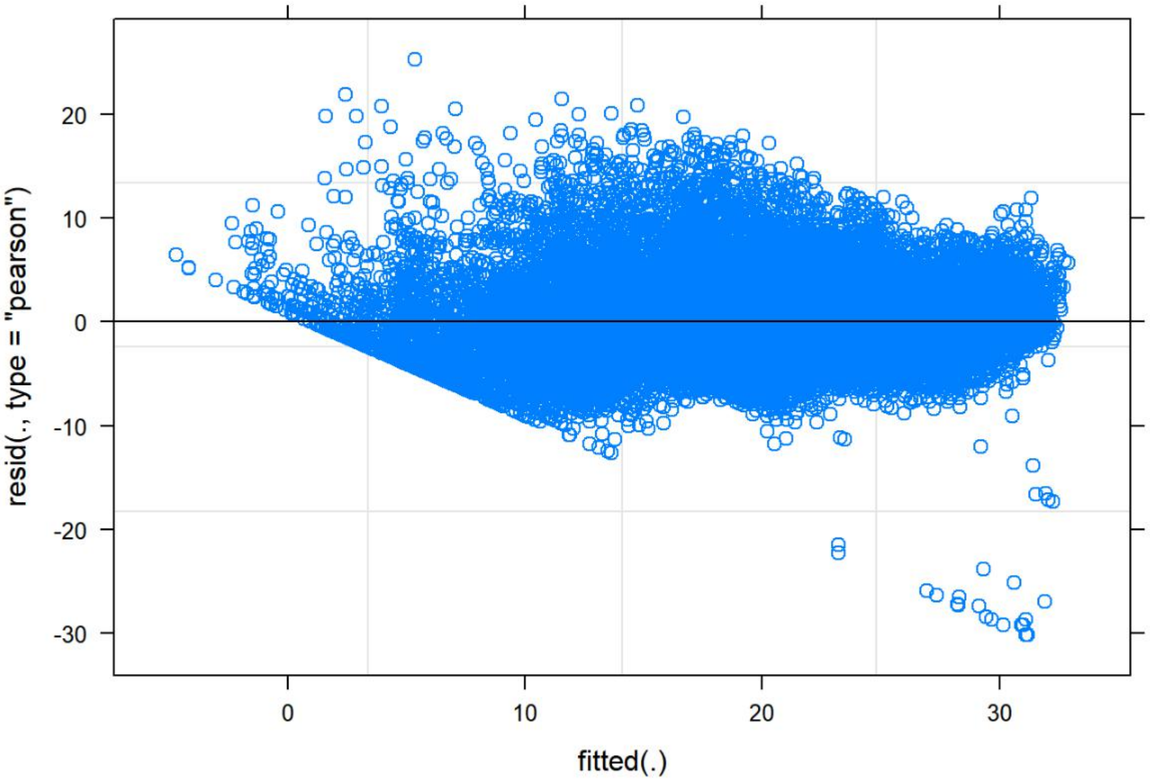
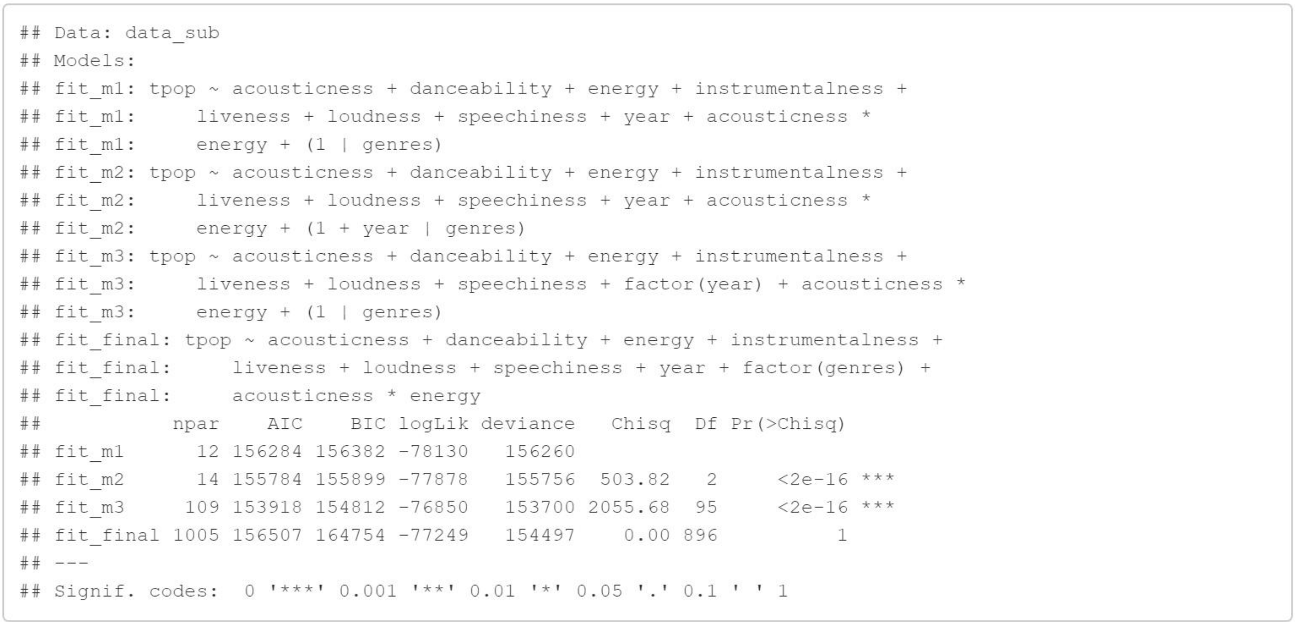


*Figure 17: The residual plot for the second multilevel model which allows for varying slope for year predictor. tpop~acousticness+danceability+energy+instrumentalness+liveness+loudness+speechiness+year+acousticness*energy +(1+year|genres)*



```
## Data: data_sub
## Models:
## fit_m1: tpop ~ acousticness + danceability + energy + instrumentalness +
## fit_m1:     liveness + loudness + speechiness + year + acousticness *
## fit_m1:     energy + (1 | genres)
## fit_m2: tpop ~ acousticness + danceability + energy + instrumentalness +
## fit_m2:     liveness + loudness + speechiness + year + acousticness *
## fit_m2:     energy + (1 + year | genres)
## fit_m3: tpop ~ acousticness + danceability + energy + instrumentalness +
## fit_m3:     liveness + loudness + speechiness + factor(year) + acousticness *
## fit_m3:     energy + (1 | genres)
## fit_final: tpop ~ acousticness + danceability + energy + instrumentalness +
## fit_final:     liveness + loudness + speechiness + year + factor(genres) +
## fit_final:     acousticness * energy
##           npar    AIC    BIC logLik deviance   Chisq  Df Pr(>Chisq)
## fit_m1      12 156284 156382 -78130   156260
## fit_m2      14 155784 155899 -77878   155756  503.82   2     <2e-16 ***
## fit_m3     109 153918 154812 -76850   153700 2055.68  95     <2e-16 ***
## fit_final 1005 156507 164754 -77249   154497    0.00 896          1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 18: The anova test for all four models. The test select the second and the third multilevel models.

```
## Data: data_sub
## Models:
## fit_m2: tpop ~ acousticness + danceability + energy + instrumentalness +
## fit_m2:     liveness + loudness + speechiness + year + acousticness *
## fit_m2:     energy + (1 + year | genres)
## fit_m3: tpop ~ acousticness + danceability + energy + instrumentalness +
## fit_m3:     liveness + loudness + speechiness + factor(year) + acousticness *
## fit_m3:     energy + (1 | genres)
##        npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## fit_m2   14 155784 155899 -77878   155756
## fit_m3  109 153918 154812 -76850   153700 2055.7 95  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 19: The anova test for the second and the third multilevel model. The test select the third multilevel models as the best model.

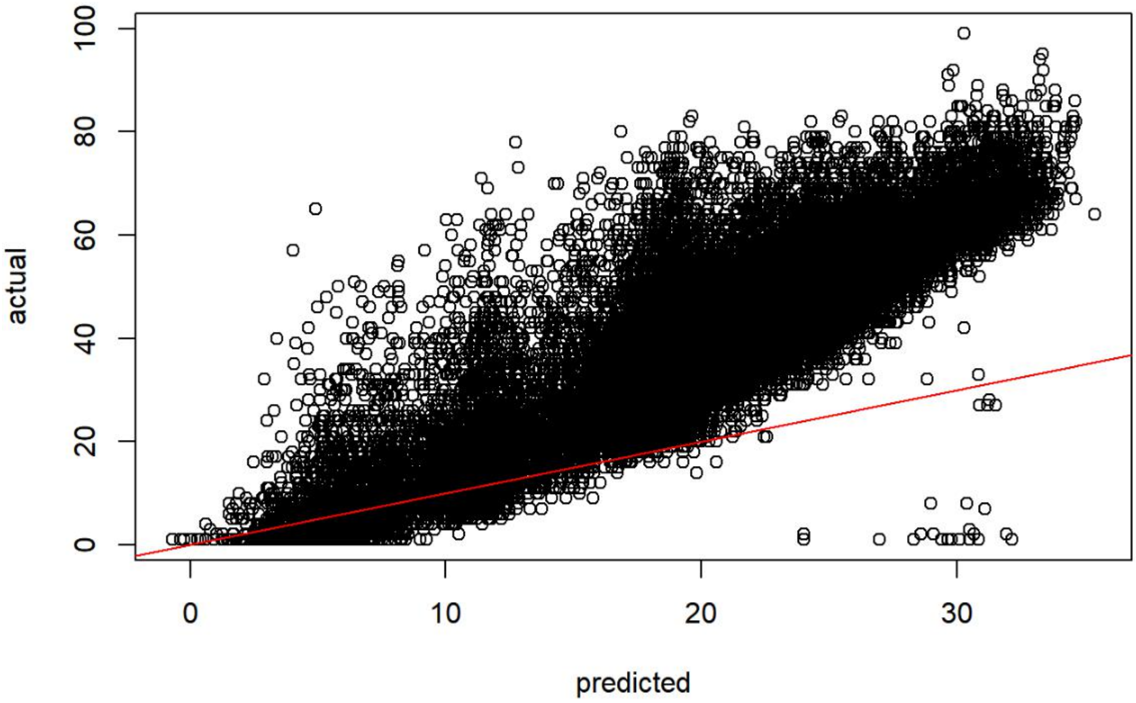


Figure 20: The predict vs. Actual plot. There are some outliers make the fitted line lies beneath the points, thus the data needs to be further investigated.

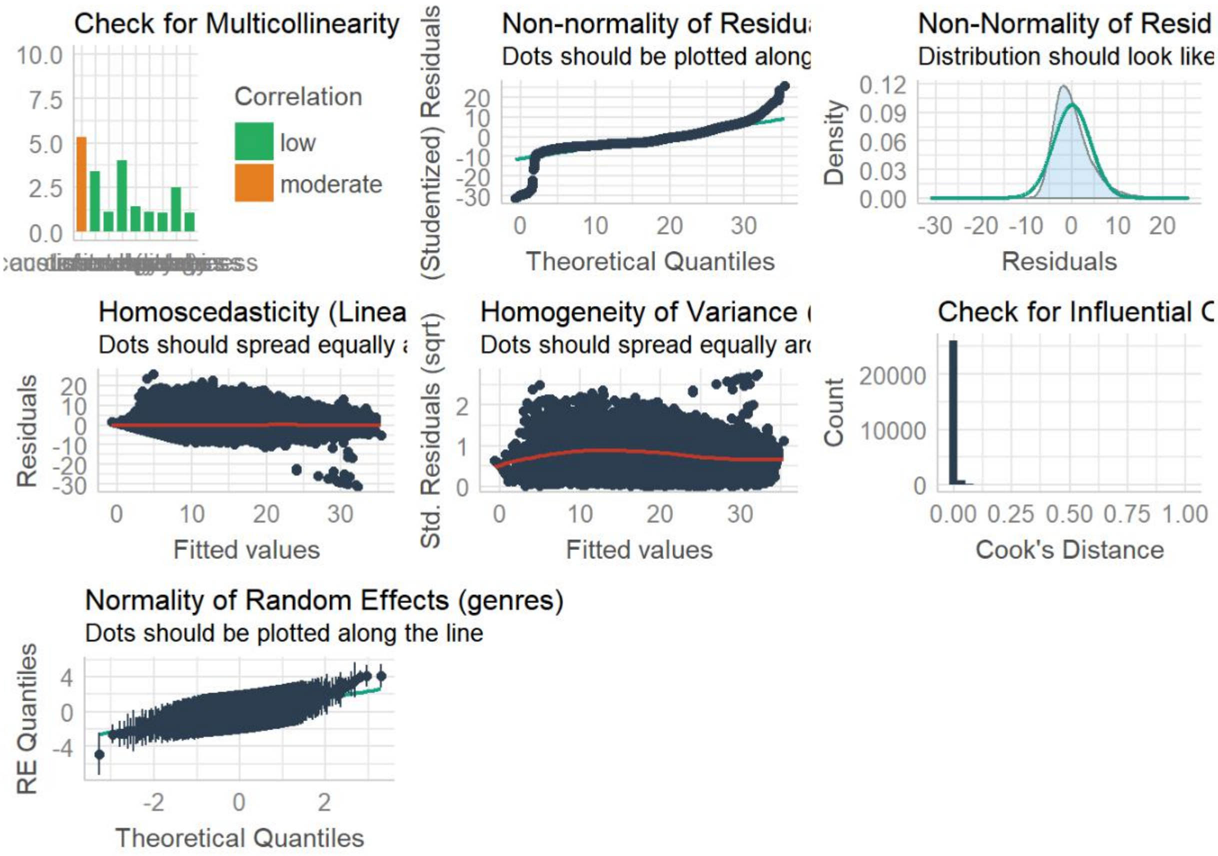


Figure 21: The model check result using performance package.