Kelly Zeng

STAT410 Final Project

Dr. Kowal

4/27/2022

Spotify Top 50 Songs: What Makes a Song Popular?

Project Statement

In this project, I will investigate what factors make a song popular. I will specifically look into what characteristics of a song (tempo, danceability, etc.) have the biggest impact on its popularity. I wanted to look into this topic to see if the most popular songs today have commonalities, which may lead song producers to utilize similar attributes over and over again. I had previously seen discussions on how popular songs all sound the same, whether it is through using the same chords or beats. Perhaps producers use this knowledge to their advantage to produce higher charting songs, and maybe top charting songs all share similar attributes.

Data Description

I will be using a dataset from Kaggle posted by user Hanna Yukhymenko, who collected the data directly from Spotify, a music streaming platform with charts. This dataset includes the top 50 songs in the year 2021 and each song's qualities, including one that evaluates the song's popularity. Using this dataset, I will see if there is a correlation between the different attributes to the popularity of the songs. It is important to note that my data will only include songs that are already popular (as they are top 50 songs). It is possible that the correlation between attributes and popularity will not be as evident for this dataset as opposed to one with a mix of popular and unpopular songs. However, with the availability and descriptive variables of this dataset, I will

continue to use it for this project to get an initial idea of how these variables contribute to popularity.

Exploratory Data Analysis

This dataset is organized with a different song in each row, with columns representing the rank of the song on the chart, artist, song name, Spotify ID, and various attributes. The attributes include 14 variables that characterize each song:

- Popularity - the higher the value, the more popular the song is
- Danceability - the higher the value, the easier it is to dance to this song
- Energy - the higher the value, the more energetic the song
- Key - the key the track is in, mapped according to standard Pitch Class notation (value of -1 if no key is detected)
- Loudness - the higher the value, the louder the song
- Mode - major is represented by 1 and minor is 0
- Speechiness - the higher the value the more spoken word the song contains
- Acousticness - the higher the value, the more acoustic the song is
- Instrumentalness - the closer the value is to 1.0, the more instrumental the song is
- Liveness - the higher the value, the more likely the song is a live recording
- Valence - the higher the value, the more positive the song mood
- Tempo - the overall estimated tempo in beats per minute
- Duration - how long the song is in milliseconds
- Time signature - an estimated time signature ranging from 3 to 7 (3/4 to 7/4 time)

By plotting popularity against all of the attributes (see below in Figure 1), I discovered that popularity had visible patterns with loudness, speechiness, acousticness, liveness, and

duration. Speechiness, acousticness, and liveness seemed to have negative correlations, while duration seemed to be concentrated around the middle. Other attributes did not have obviously visible correlations. Figure 1 also showed that energy and loudness seemed proportional to each other. The only thing that surprised me was that danceability and energy did not seem to have a clear correlation with popularity. I expected a lot of top charting songs to be very catchy and energetic, but the data did not show this pattern.

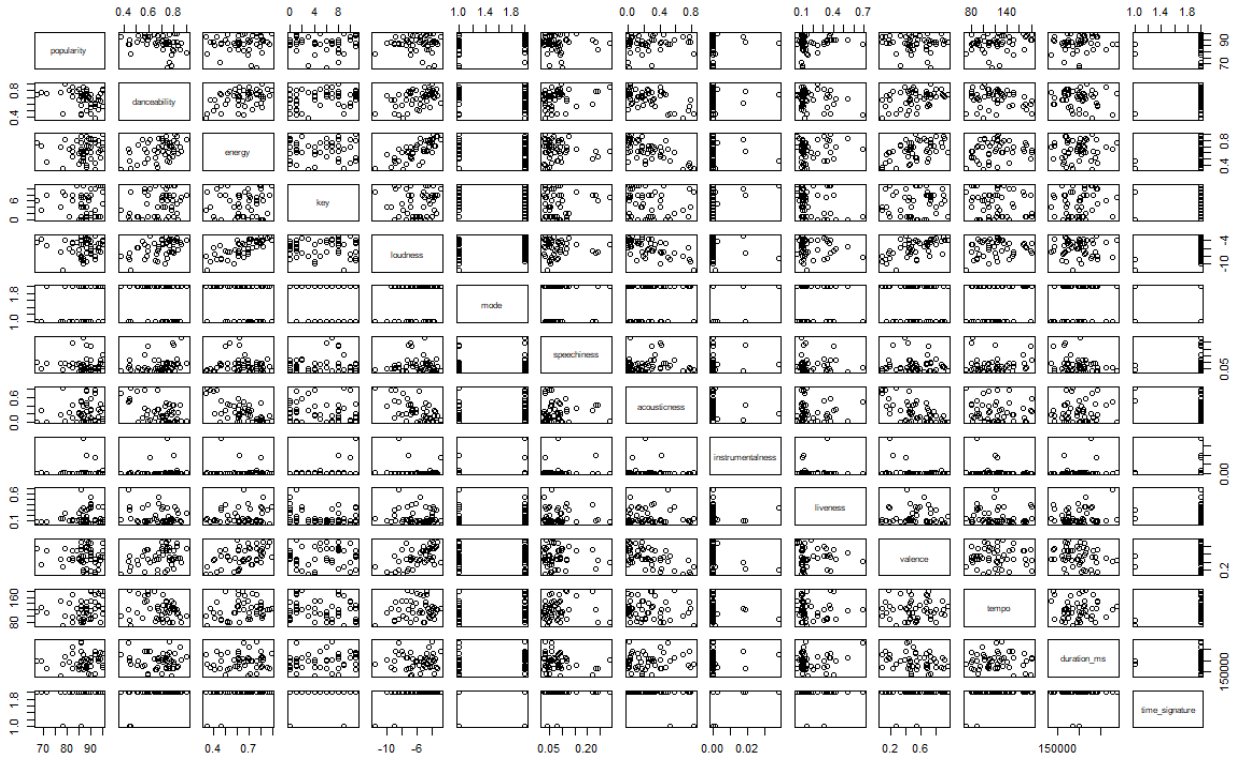

Figure 1: Pairwise Plot

The time signature value had mostly values of 4 apart from one observation, so I do not think this variable would be the best to look at for an effect on popularity. Most songs also had an instrumentalness value of 0.

Data Analysis

I first fitted a multiple linear regression (MLR) model with popularity as the response variable and the other 13 variables as parameters. I used a MLR to account for the many

variables in the dataset, so I could observe the effects of all of them on the popularity of a song. Looking at the diagnostics of this model (Figure 2), we can see that the residuals vs fitted line is straight with some deviations. This means our model is linear, but there could be some concerns due to the deviations. The normal QQ plot is also straight with some deviations, showing the model has a normal distribution. The scale location lines are straight, meaning there is constant variance. Lastly, the residuals vs leverage plot shows that none of the points are bad leverage points. We can see that these diagnostics confirm the model assumptions, but it is important to note that there may be some concerns due to the deviations for all of them. This allows for the conduction of confidence intervals and hypothesis testing. From the large p-values of each coefficient, we can also see that all of the variables were not significant. We fail to reject the null hypothesis that the betas are significant or equal to zero.

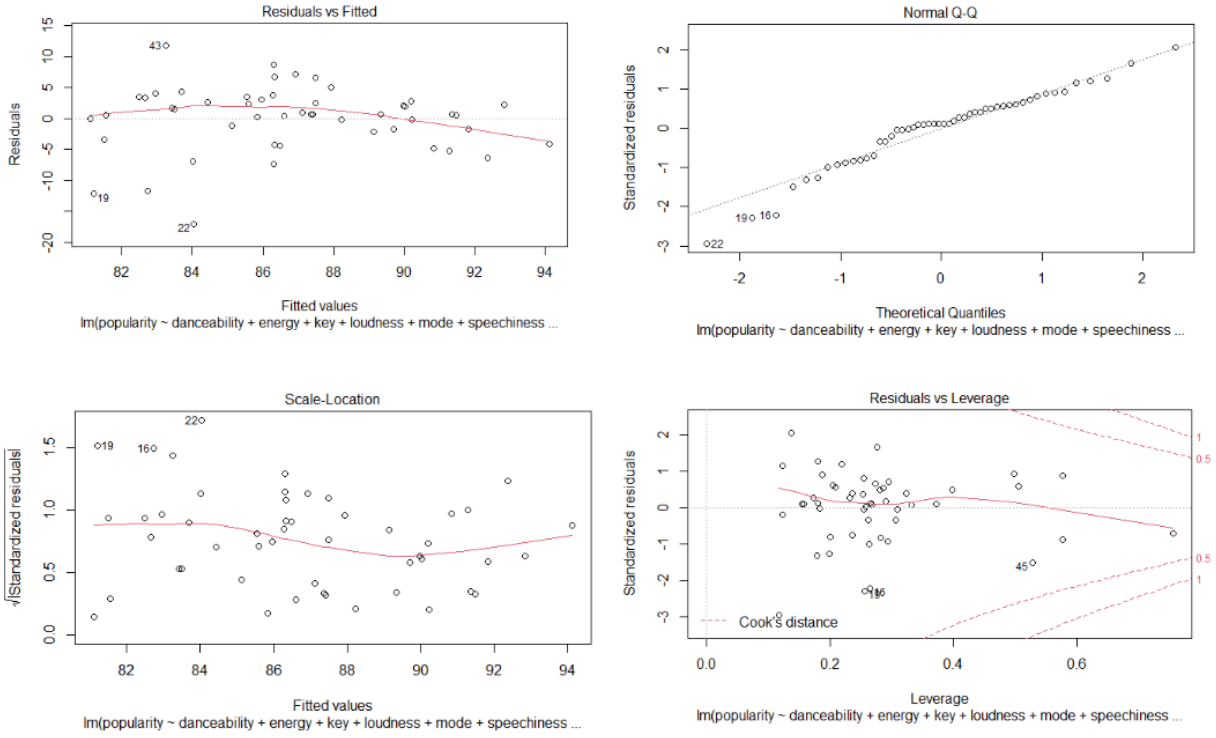

Figure 2: Diagnostic Plots for MLR Model

I used a ridge regression next to produce better estimates of the parameters by imposing a penalty term on the coefficients of predictors. I used the lambda value that minimized the cross-validation error squared and the largest lambda value within one standard error of the minimized lambda value. Using both lambda values for regressions, the estimates I got were both very different from the ordinary least squares (OLS) estimates. I also tried a lasso regression that is different from ridge in that it can shrink coefficients all the way to zero, which results in a reduced number of variables. Again, using two lambda values, the estimates I got were very different. Both also eliminated variables with the minimized lambda value keeping danceability, mode, energy, liveness, tempo, and time signature. The one standard error lambda value reduced the model to only the intercept. These estimates of the parameters are better than the OLS estimates.

Next, I wanted to use model selection to exclude some of the variables from the model. By doing so, I can remove irrelevant and redundant variables, and I can have a simpler interpretation with fewer variables. The first method I used was using Akaike's information criterion (AIC) and Bayesian information criterion (BIC). Both of these values gave me the same result, hence why I am grouping them together. These values provide information on how well the model fits the data while factoring how many parameters are used with a penalty. The algorithm stops when removing another variable would not make the AIC/BIC value smaller, which would minimize the AIC/BIC value. The model chosen with the lowest AIC and BIC values included the parameters danceability and time signature. I then checked what the output was when going backwards. This gave me the null model. I checked the AIC and BIC values of both models, and the values for the forward model had lower values. Therefore, I chose to continue with the forward model.

Looking at the diagnostics of the AIC and BIC model (Figure 3), we can see based on the residuals vs fitted plot that this model is linear since the red line is roughly straight. The points on the normal QQ plot roughly follow the straight line, demonstrating that the distribution is normal. The scale location plot is also roughly straight. This means there is a constant variance. From the residuals vs leverage plot, we can see there are some outliers, but there are no bad leverage points. This model does follow all assumptions needed, but there may be some concerns since there are some deviations. The p-value of the estimates is small, which tells us that danceability and time signature are significant toward having an effect on popularity. The beta values for the intercept, danceability, and time signature are 61.388, -19.865, and 9.857, respectively. This illustrates that danceability has a negative correlation while time signature has a positive correlation.

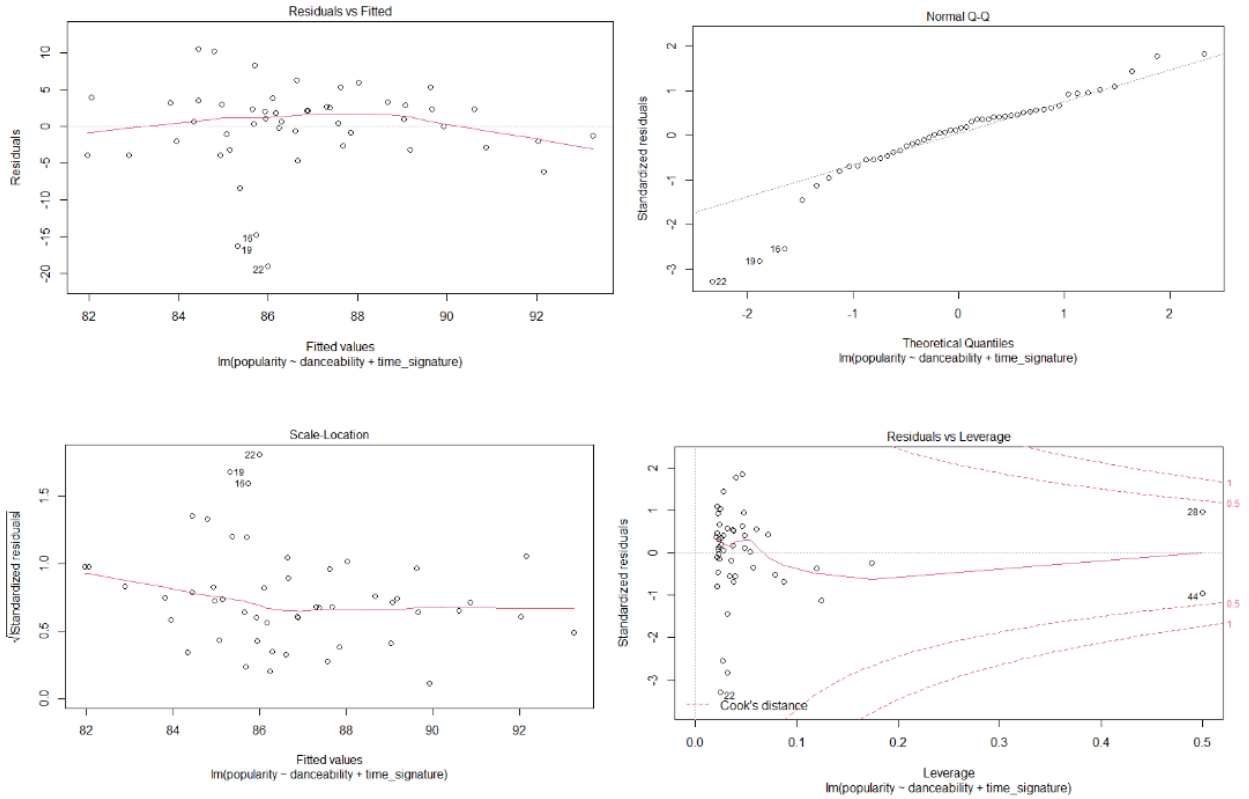

Figure 3: Diagnostic Plots for AIC/BIC Model

The other method I tried was using the adjusted R-squared value, which we try to maximize instead of minimize for AIC and BIC. With this method, the chosen model had 6 variables, including danceability, energy, key, loudness, mode, and time signature. This is clearly different from the model AIC and BIC gave. However, when looking at the diagnostics, we can see that the plots (Figure 4) are very similar to those from the AIC and BIC model, meaning the assumptions also hold with some areas of concern. For this model, the p-value is only small for the intercept and danceability variable, which lacks time signature from the AIC and BIC model. Time signature does not have a significant effect on popularity in this model. The beta values of the intercept and danceability are 47.147 and -21.7202. This once again illustrates that danceability has a negative correlation.

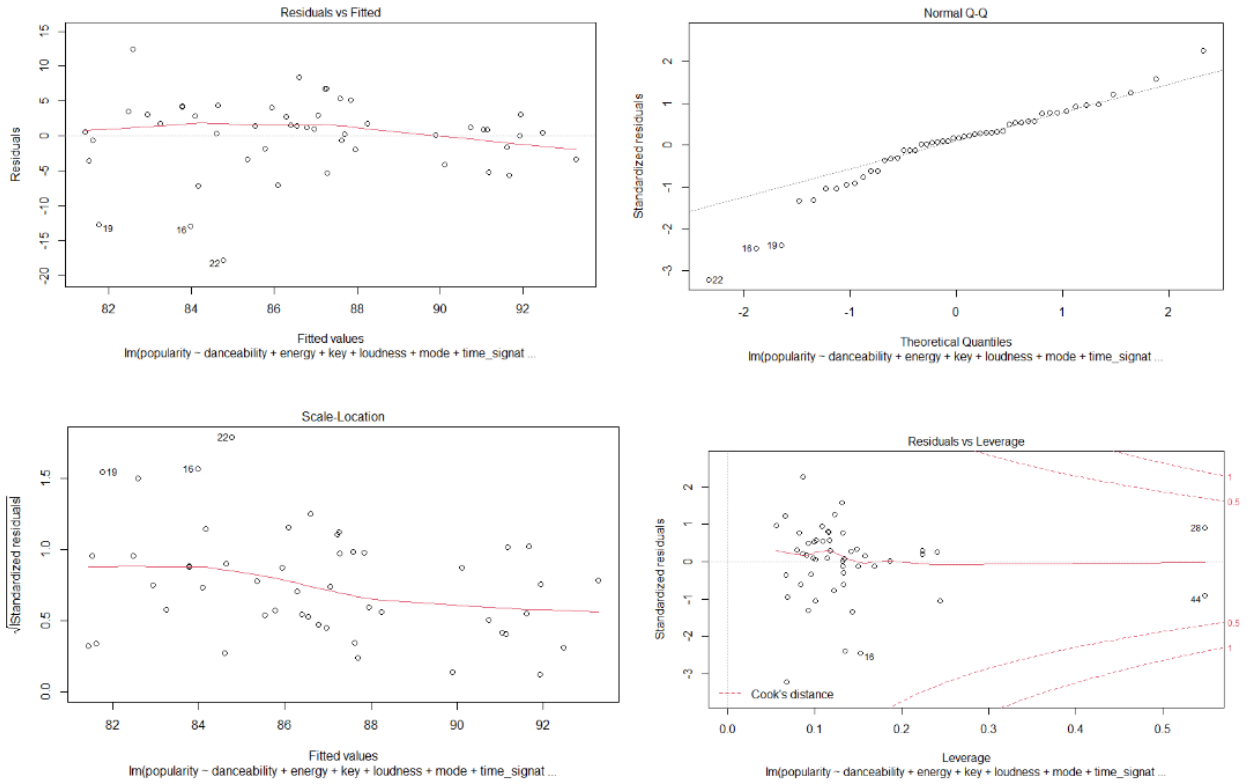

Figure 4: Diagnostics Plots for Adjusted R-Squared Model

I then decided to consider interactions between different variables. From looking at earlier plots, I wanted to try an interaction term for danceability and mode. Danceability is a continuous variable and seems to be significant. Mode is a categorical variable and also was included in many previous models, showing it may have an effect on popularity. I thought these two variables could have a correlation because danceable songs often make people feel happy and energetic. This is what the major key also does. The minor key, on the other hand, evokes more negative emotions. After fitting this model, I found that the AIC and BIC values were larger than before and the adjusted R-squared value was smaller. I decided this would not be the best model to look at and proceeded to consider other models.

The last model I looked at was using non-linear models. I decided to use a partially additive model. A fully additive model was not possible with my particular dataset as there would be more coefficients than data. I chose 4 variables that seemed the most likely to be non-linear through their initial plots with popularity as the response variable. These variables were key, speechiness, instrumentalness, and valence. These would be the non-linear parameters in the fitted model. From the summary output, it was apparent that these variables were non-linear as all the effective degrees of freedom were far from a value of one. It is also apparent from the plot of the estimated regression functions that they all appear nonlinear (seen below in Figure 5).

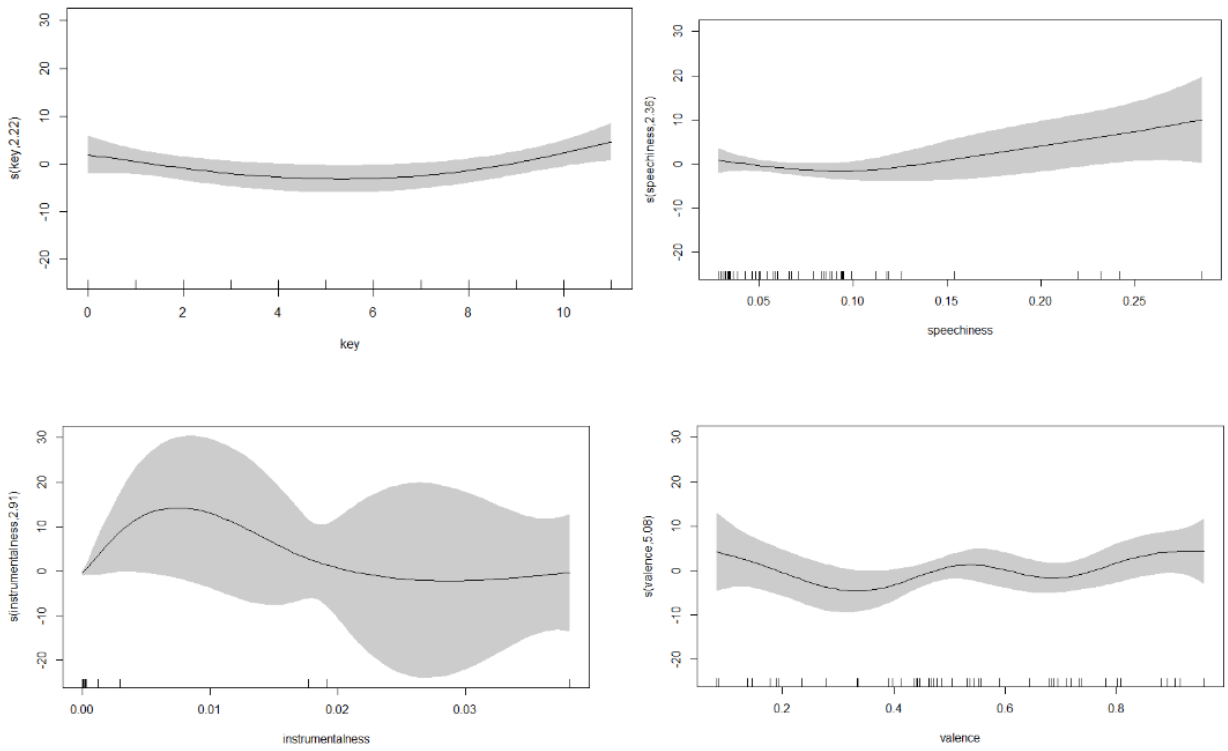

Figure 5: Estimated Regression Functions

The AIC value for this model was actually smaller than the model that minimized AIC from before, showing it may be a better model. However, the BIC was bigger, and the adjusted R-squared was smaller. This means previous models could also be better models when looking at different model selecting methods. It is important to notice that only the intercept had a small p-value. The other variables were not significant in having an effect on the popularity of a song.

Summary and Discussion

So what attributes do you need to make a hit song? In general, do not make the song danceable as danceability has a significant negative correlation with popularity. It may also be worth it to use a 4/4 time signature according to some models where time signature is a significant variable. Energy, loudness, liveness, mode, key, and tempo are other variables that may have an effect on popularity, but it may not be a significant effect.

It is important to keep in mind that there are limitations to this conclusion. With only 50 observations in this dataset, it is unclear if the analysis done in this report will be representative of all songs. Not only is the lack of observations an issue, but the songs in this dataset are all already popular songs that made it onto the Top Songs of 2021 playlist on Spotify. While looking at the diagnostic plots of a few of the models, the assumptions held true, but there could be concerns around the assumptions to the deviations in the diagnostic plots.

I think conducting further investigation into this topic would be interesting. I would first find a larger dataset with songs not necessarily at the top of charts. I would consider clustering data by artists to see how the popularity of songs change without factoring the different artists and their musical styles. I am also curious to see if different genres or languages of music have different models of popularity. Maybe the attributes that have an effect on popularity change by year based on what is happening around the world. Perhaps sad songs were more popular during the COVID-19 pandemic and brighter songs were more popular before. These are all questions that are pursuable through further research in this topic of what type of songs are more popular.

References

Hanna Yukhymenko. (2022, March). Spotify top 50 songs in 2021. Retrieved from https://www.kaggle.com/datasets/equinxx/spotify-top-50-songs-in-2021.

Appendix

Code and data submitted as separate files on Canvas submission page

Code: stat410projectcode.R

Data: spotify_top50_2021.csv