Predicting Taylor Swift's Song Popularity from Song Characteristics

By Nicholas Cassol-Pawson, Raymond Durr, Sarina Doss, and Michael Hao

Introduction

Overview

This paper analyzes the relationship between song-intrinsic characteristics and the popularity of Taylor Swift songs. We used a dataset built from the Spotify API. To successfully construct a model we had to heavily clean the data and select the optimal variables, eventually reducing the number of predictors in the reduced model to just 8 variables. This model successfully explains about two-thirds of the variation in the data and generally has highly interpretable coefficients. We faced some limitations in our analysis though, mostly due to inability to access song-extrinsic characteristics and a sophisticated form of lyrical analysis.

Research Question of Interest

Can song-intrinsic classifications, including computer ratings of musical characteristics and the metadata of the song predict the popularity of a Taylor Swift song?

Background of Data Set

The dataset comes from kaggle.com. In this dataset, there is a list of all released Taylor Swift songs from Spotify's API. In total, there are 16 variables and 529 observations. All of the following values come from Spotify's algorithm to determine various elements of a song. Descriptions are paraphrased from Spotify's descriptions of the variables.[1]

- **name**: the song's name
- **album**: the name of the album the song is on
- **release_date**: the day, month, and year of the album's release
- **track_number**: the song's position on the album
- **id**: Spotify's unique identification number for the song
- **uri**: Spotify's uniform resource identifier for the song
- **acousticness**: a confidence measure from 0.0 to 1.0 predicting whether a song is acoustic. Songs likely to be acoustic have values near 1.0 and those likely to be electronic have values near 0.0.
- **danceability**: a measure from 0.0 to 1.0 of how suitable a song is for dancing based on several musical characteristics. The most danceable songs have values of 1.0 and the least, 0.0.
- **energy**: a measure from 0.0 to 1.0 of musical intensity and activity based on several musical characteristics. The highest energy songs have values of 1.0 and the lowest, 0.0.
- **intrumentalness**: a confidence measure of whether a song has vocals. Songs without vocals have values of 1.0; instrumental ones, 0.5; and only vocal songs, 0.0.
- **liveness**: a confidence measure of whether there is an audience in the recording. Songs that are highly likely to be live have values above 0.8.
- **loudness**: the average loudness of a song in decibels (dB). Values range between -60 and 0 db below the maximum loudness of the system they are being played on.
- **speechiness**: a confidence measure for the presence of spoken words in a song. Songs that likely are entirely spoken have values above 0.66; those with both music and speech have values between 0.33 and 0.66; and music and other non-speech songs have values below 0.33.
- **tempo**: the overall estimated speed of a song in beats per minute.
- **valence**: a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a song. Happier-sounding songs have values near 1.0 and sadder-sounding ones, near 0.0.
- **popularity**: an algorithmic measure of the popularity of the song from 0 to 100 based on its total number and recency of plays. The most popular songs have values of 100.
- **duration_ms**: a song's duration in milliseconds.

[1] https://developer.spotify.com/documentation/web-api/reference/get-audio-features

After cleaning the data, where we removed all albums that were not released as editions of one of Swift's 10 studio albums, we decided to embed characteristics of name and album into the data as indicator variables. We decided to remove intrumentalness (since all the released songs have vocals), track_number, id, and uri. We also added two dummy variables: **deluxe** and Taylor's Version (**tv**)

- **deluxe**: the bonus album released after the original release of the album with added content. A value of 1 indicates that the song is on a deluxe album.
- **tv**: Swift is rerecording her past albums to gain full ownership over her music. She calls her rerecorded songs "Taylor's Version." A value of 1 indicates that the song is a rerecording.

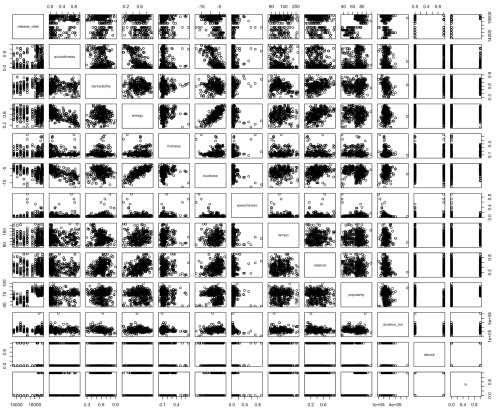
In total, the dataset after cleaning is now 13 variables with 426 observations.

Justification of the Model

We used multiple linear regression to approach our research interest. Starting with OLS, we determined that there are no major violations in the diagnostic plots and thus there is no need to transform the variables. We then analyzed the outliers that were indicated in the residual plots and determined that only one, "All Too Well (10 Minute Version)," was a bad leverage point. We chose to remove it from the dataset, as it was a significant outlier due to its duration. We believe it is significantly longer than any other song that we would reasonably wish to predict from the model, as Swift will likely never release a song that long again, so there is little use in keeping it in the final model. From a model built without this outlier, we then performed stepwise regression to finalize our selection of variables and determine the best model.

Data Description:

We first examine the scatter plot matrix of all variables (see Appendix for larger graph). If we look at the relationship between each variable and **popularity**, there are few very strong linear relationships. The strongest relationship appears to be a positive one between it and **release_date.** The relationships between **popularity** and the indicator variables also appear to be linear, quite negative in the case of **deluxe** and rather positively so in the case of **tv**. Its relationships with **danceability**, **liveness**, and **loudness** all have positive linear trends, although they appear to be rather weak. The relation with **energy** looks very flat; there does not seem to be very much correlation between the two. The variables **speechiness** and **duration_ms** display relatively flat relations with **popularity**, although there is a slight negative trend to the **duration** one. There is very little linear relationship between **popularity** and **acousticness**, **tempo**, or **valence**, which each have relations that display a lot of clumping.



We now analyze some summary statistics for the data, using the table below and examine their frequency histograms (included in the Appendix):

	popularity	release_date	acousticness	danceability	energy	liveness	loudness	speechiness	tempo	valence	duration_ms	deluxe	tv
Mean	66.94	2018-02-21	0.297841	0.59	0.58	0.15	-7.50	0.06	122.74	0.40	238264.60	0.32	0.28
SD	13.16	1979.50 days	0.32	0.11	0.19	0.09	2.94	0.08	29.94	0.20	45255.69	0.47	0.45
Min	37	2006-10-24	0.0002	0.29	0.1180	0.04	-17.93	0.02	68.53	0.04	107133	0	0
Med.	70	2020-08-18	0.16	0.60	0.60	0.12	-6.94	0.04	119.48	0.40	234073	0	0
IQR	18	3652 days	0.55	0.13	0.28	0.05	3.80	0.03	46.99	0.31	47859.5	1	1
Max	99	2023-10-27	0.97	0.90	0.95	0.66	-1.91	0.91	208.92	0.94	613026	1	1

The response variable, **popularity**, has a relatively symmetric distribution, with similar mean and median. The 1.5*IQR range includes all values. The data has some variation, with a standard deviation of 13.6. A normal distribution fits the data, but will overestimate low values and underestimate high ones.

We see that the majority of Swift's songs were released in the past six years, with the median **release date** being in early 2020. Release dates are highly skewed to more recent values, with the mean in 2018, somewhat earlier than the median. The 1.5*IQR range includes values. The data are quite variable, with a standard deviation of more than 5 years. The number of releases increased significantly in the past 5 years, with releases each year rather than every few, and in 2021 and 2023, more than twice her pre-2015 maximum number of songs released. A beta distribution with a high alpha and low beta value best fits the data.

We see that the majority of Swift's songs have low **acousticness**, with a median value of 0.16, but a high skew, as can be seen by the mean being almost twice the median. The IQR covers all of the values, with the maximum value of 0.97 being contained within the range of the traditional 1.5*IQR definition of outliers. The data is highly variable, with a standard deviation of 0.32. Examining the graph, we see that it is probably best modeled by a chi-squared distribution.

Danceability is symmetrically distributed with highly similar mean and median. It is not highly variable, with a standard deviation of only 0.11. However, there are some outliers, with both the minimum and the maximum outside of the 1.5*IQR range. The data look like they can be modeled rather well by a normal distribution, although there is a bit less tailing off than would be expected.

Energy is rather symmetrically distributed, with a similar mean and median. It has moderate variation, with a standard deviation of 0.19. There are no outliers, all points are included in the 1.5*IQR range. The data look best modeled by a normal distribution, although it would overestimate central points.

Liveness is highly skewed. Both the mean and median are between 0.1 and 0.15, but it has some very high outliers (all points above 0.25 are outliers). It is not very variable, however, with a standard deviation of 0.09. The data are best modeled by a chi-squared distribution.

Loudness is rather symmetric, although it has a slight left skew. The mean and median are around -7 db, implying that Swift's songs are relatively loud on average. However, the histogram indicates that a significant portion of her songs are quiet, while very few are louder than a value of -5 db. The data is rather variable, with a standard deviation of about 3. There are some outliers on the lower end of her songs, with songs that have loudness scores less than -15 db being outliers. The data could be best modeled by a normal distribution, which would overestimate higher values and underestimate lower ones.

Speechiness is highly skewed, with almost all of the values found between 0 and 0.1. It has extremely low mean and median, with (relatively) many outliers above 0.11. The data is highly variable, with a standard deviation of 0.08. This data would be extremely well modeled by a chi-squared distribution.

Tempo is rather symmetrically distributed, with similar median and mean that are located near the center of the range of values. There is much variability in the data, with a standard deviation of about 30. Some of the higher values are outliers, with a cutoff on the 1.5*IQR measurement of values around 190. The data are approximately normally distributed, with lower values being slightly underestimated by a normal curve.

Valence has a rather symmetrical distribution, with identical mean and median located near the center of the range of values. It is quite variable with a standard deviation of 0.2. There are no outliers; all values are within the 1.5*IQR range. The data are decently modeled by a normal distribution, although it underestimates low values and overestimates high ones.

Duration has a rather symmetric distribution, with similar median and mean values. It is highly variable, with a standard deviation of more than 45,000. Values with duration greater than 300,000 milliseconds are outside of the 1.5*IQR bracket, but there is one outlier that has duration of greater than 600,000 milliseconds, almost twice the value of any other outlier. This is the song "All Too Well (10 Minute Version)," which we remove from the final dataset, as it is so much longer than anything Swift has released or will likely release in the future, so there is little sense uncompromising the integrity of the model to be slightly more robust to handle high-duration songs. Ignoring this outlier, the data is best modeled by a normal distribution, although it may slightly overestimate non-central values.

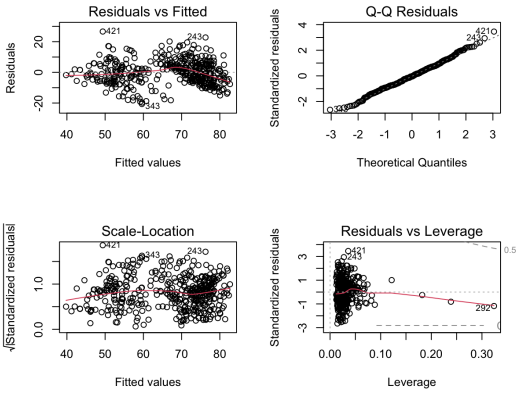
About 32% of the songs in the data set are **deluxe**. This is an indicator variable, so no other summary statistics are appropriate. The data are binomially distributed.

About 28% of the songs in the data set are coded as **Taylor's Version**. This is an indicator variable, so no other summary statistics are appropriate. The data are binomially distributed.

Results and Interpretation:

Full Model

The full model, after removing the outlier, predicts **popularity** by **release date**, **acousticness**, **danceability**, **energy**, **liveness**, **loudness**, **speechiness**, **tempo**, **valence**, **duration**, and the dummy variables **deluxe** and **Taylor's Version**.



We first examine the full model's diagnostic plots.

The residual plot is centered close to 0 and the scale location plot indicates that, despite some clumping, the errors appear to have a random distribution, with constant variance. This implies that assumptions about the error term are held. The Q-Q plot also indicates the normalcy of the error.

There are some outliers visible in the Residuals vs. Leverage plot. However, after running a test for bad leverage points, which flag any points that have standardized residuals greater than 2 or less than -2 and leverage values that, in this case, are greater than 0.061, we determine that none of these outliers are bad leverage points. This can be seen in the graph, where all the points with high standardized residuals have hat values less than 0.05. Since the model satisfies all assumptions about the error, we choose not to transform any variables.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.696e-01  7.655e+00  -0.022 0.982338
release_date  5.442e-03  3.074e-04  17.702  < 2e-16 ***
acousticness  5.785e+00  2.103e+00   2.751 0.006199 **
danceability -4.618e+00  4.201e+00  -1.099 0.272305
energy       -7.464e-01  4.104e+00  -0.182 0.855768
liveness     -1.121e+01  4.578e+00  -2.449 0.014726 *
loudness      1.489e+00  2.876e-01   5.177 3.53e-07 ***
speechiness  -4.889e+00  5.849e+00  -0.836 0.403670
tempo        -7.802e-03  1.341e-02  -0.582 0.561177
valence      -3.198e+00  2.530e+00  -1.264 0.206959
duration_ms  -3.519e-05  1.147e-05  -3.068 0.002298 **
deluxe       -6.902e+00  8.616e-01  -8.011 1.18e-14 ***
tv           -4.268e+00  1.265e+00  -3.375 0.000808 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.845 on 412 degrees of freedom
Multiple R-squared:  0.6538,    Adjusted R-squared:  0.6437
F-statistic: 64.85 on 12 and 412 DF,  p-value: < 2.2e-16
```

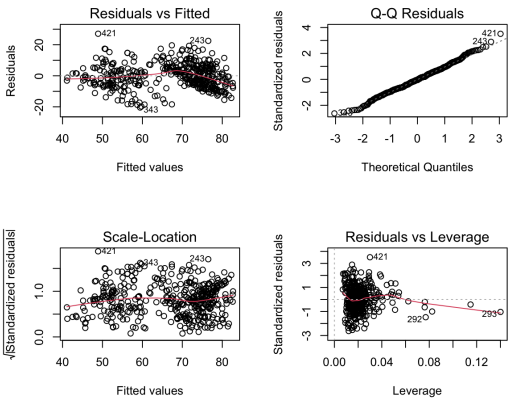
Looking at the R output of the model, we see that it has an adjusted R^2 value of 0.6437. Combined with the diagnostic plots, this indicates that the model has a good fit.

```
> vif(full_model)
release_date acousticness danceability       energy     liveness     loudness
    2.554560     3.043456     1.474894     3.997353     1.156390     4.944497
 speechiness        tempo      valence  duration_ms       deluxe           tv
    1.420941     1.111772     1.823417     1.559849     1.115555     2.237683
```

However, this model has a significant weakness: looking at the p-values, only 7 of the variables are significant at a 5% level. Running a VIF indicates another problem: although none of the values are above 5, loudness approaches it, indicating that there is probably multicollinearity going on within the model. We choose to implement variable selection techniques to try to resolve these two issues.

Reduced Model (Backward Stepwise and Subsets)

We try two techniques to create a reduced model: backward stepwise regression from the full model and subset selection. Backward stepwise was done by optimizing along the AIC and subset selection chose the model that had the lowest AIC and AIC corrected, although it had a slightly higher BIC. Both techniques proposed a model that predicts **popularity** with 8 predictors: **release date**, **acousticness**, **liveness**, **loudness**, **duration**, and the dummy variables **deluxe** and **Taylor's Version**.



Examining the diagnostic plots of this model, we can see that they look almost identical to that of the full model, which makes sense, seeing as it removed almost all of the insignificant variables from the full model, implying that they likely were multicollinear with the remaining ones. We notice that some outliers are flagged for further investigation. Checking if any are bad leverage points, which have the same cutoff values as before, tells us that none of these points are.

We now check the VIF of the reduced model to see if it has dealt with the multicollinearity problem. None of the values are anywhere near 5, with the highest being **loudness**, which now has a value of just over 3. We conclude that there is no multicollinearity in this model.

```
> vif(reduced_model)
release_date acousticness     liveness     loudness      valence  duration_ms
    2.406651     2.372416     1.070335     3.175431     1.385590     1.403019
       deluxe           tv
    1.105678     2.162475
```

Looking at the R output of the model, we see that it has an adjusted R^2 value of 0.6453, which is slightly better than the full model. Furthermore, all of the coefficients are significant at a 5% level.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.605e+00  6.312e+00  -0.730 0.466097
release_date  5.428e-03  2.977e-04  18.232  < 2e-16 ***
acousticness  6.764e+00  1.852e+00   3.652 0.000294 ***
liveness     -1.228e+01  4.394e+00  -2.795 0.005432 **
loudness      1.571e+00  2.300e-01   6.831  3.0e-11 ***
valence      -4.332e+00  2.201e+00  -1.968 0.049680 *
duration_ms  -3.029e-05  1.085e-05  -2.791 0.005498 **
deluxe       -6.919e+00  8.559e-01  -8.084  6.9e-15 ***
tv           -4.146e+00  1.240e+00  -3.342 0.000906 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.827 on 416 degrees of freedom
Multiple R-squared:  0.652,     Adjusted R-squared:  0.6453
F-statistic: 97.42 on 8 and 416 DF,  p-value: < 2.2e-16
```

We now compare the full and reduced models.

Model Comparison

We want to confirm that the reduced model is better than the full model, so we check a few statistics. First, we compare their values for R^2, adjusted R^2, AIC, AIC_C, and BIC. The reduced model is superior to the full model in every metric besides a small reduction in R^2, which makes sense since R^2 benefits models with more predictors. This is extremely impressive since the reduced model includes 4 fewer predictors than the full one.

	R^2	Adjusted R^2	AIC	AIC_C	BIC
Full	0.6538247	0.6437419	1763.647	1764.667	1816.325
Reduced	0.6519940	0.6453016	1757.889	1758.418	1794.358

We now conduct a partial F-test to see if the reduced model is statistically similar to the full model. The p-value of 0.703 implies that this is the case. Therefore, the numerical superiority of the reduced model, combined with its simplicity and its satisfaction with the error conditions, led us to choose it as our optimal model.

```
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    416 25488
2    412 25354  4    134.08 0.5447  0.703
```

Interpreting the Reduced Model

Our optimal model indicates that the popularity of a song will increase by 0.005428 units for every day more recently than it was **released**, implying that people prefer Swift's more recent music. It suggests that popularity will increase by 6.764 units for every 1 unit increase in Spotify's algorithm's confidence that the song is **acoustic**, implying that people prefer Swift's less electronic-sounding music. It suggests that popularity will decrease by -12.228 units for every 1 unit increase in a song's probability of being **live**, implying that people prefer Swift's songs that sound like they were recorded in a studio. The model implies that popularity increases by 1.571 units for every one-unit increase in a song's **loudness**, implying that people prefer Swift's louder songs. The model predicts that for every one-unit increase in a song's **valence**, its popularity will decrease by 4.332 units, implying that people prefer Swift's sadder-sounding songs. The model indicates that for a one-millisecond increase in the song's **duration**, its popularity will decrease by 4.146 units, implying that listeners prefer Swift's shorter songs. The model indicates that **deluxe** songs will be 6.919 units less popular than non-deluxe ones, indicating that people prefer to listen to Swift's non-deluxe albums. The model indicates that **Taylor's Version** songs will be 4.146 units less popular than the non-Taylor's Version song, indicating that people prefer to listen to Swift's original recordings of her music.

Discussion

Real-World Application

Our final model makes sense in a real-world situation. From our data, we see that for every year more recently that they were released Swift's songs are more popular by about 1.8 units, on average. As indicated by our interpretation of the **Taylor's Version** variable, this could be due to song-extrinsic characteristics. For example, research conducted by Kristen Hudgins at the University of Oregon's School of Music and Dance indicates that Swift's popularity grows in proportion to how much she engages her fans around the release of a new album.[2] Individual song popularity, therefore, may be motivated by extrinsic qualities like hype or fan loyalty. For example, for the release of her album "Midnights," Swift posted a TikTok every few days announcing a different track name. This helped her gain publicity and also allowed her to interact with her fans. Her community, therefore, became more responsive, which may have motivated increased post sharing and discussion. This in turn made her more popular which as a result allowed her songs to become more popular.

Limitations

There are a couple of limitations to this analysis. The primary one is due to the data set: the regression was limited to musical characteristics. One factor that plays heavily into Swift's music, and that we think impacts the popularity of her songs, is the quality and emotion of her lyrics. A regression that included some sort of lyrical analysis, classifying the songs with scores about the emotions that they are communicating, rather than just how they sound, might have better predicted their popularity.

We could have also begun to analyze song extrinsic characteristics. As previously mentioned in our discussion on Real-World Applications, variables like "hype" surrounding a song and the solidity of a fanbase are potentially extremely important to consider. Our dataset, and the Spotify API in general, does not have access to these sorts of variables. We were, therefore, unable to analyze the context in which a song exists. This limits our interpretation and analysis of Swift's popularity. For example, the **Taylor's Version** variable having a negative coefficient despite **release date** having a positive coefficient is perplexing. Taylor's Version songs are newer versions of old songs — if the release date increases popularity then **Taylor's Version** should be more popular. The switch in sign cannot be due to multicollinearity since the VIF scores are all far below 5. We are, therefore, left puzzled while only analyzing song-intrinsic characteristics. If, however, we were able to extend our analysis to song-extrinsic characteristics we might be able to piece together what is happening. For example, if we had been able to include a variable that rated the nostalgia of a song, we may have been able to show a correlation between nostalgia and popularity. The interpretation of the **Taylor's Version** variable in relation to nostalgia would then be that, when a fan chooses to listen to an older song, they prefer the original version to the updated one.

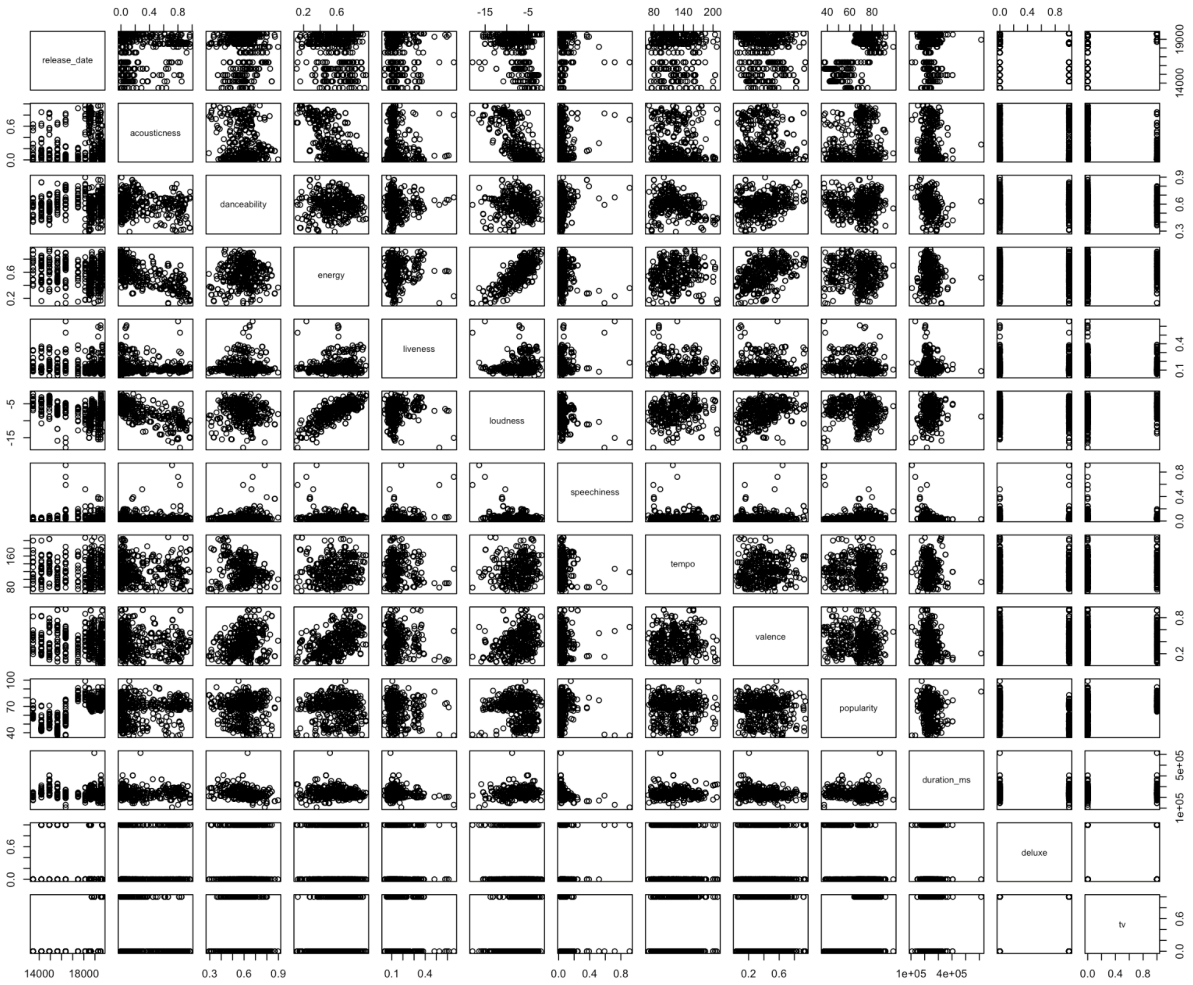
Summary

After cleaning our data to only have numeric predictors and include studio albums, and removing the extreme outlier, we built a full model for popularity across the chosen song characteristics. We then performed variable selection and were left with a model that predicted the popularity of Swift's songs across 8 predictors. This reduced model was quite good at predicting popularity, being able to explain just under two-thirds of the variation in the values of popularity. All predictors were significant in the final model, and most had signs that behaved in the way we expected. We found a couple of limitations to our model. The data is all musical characteristics, even though we believe that lyrical characteristics are just as important. We also were only able to study song-intrinsic characteristics, which limited our analysis and interpretation of a few variable coefficients. Both of these limitations can be fixed in future versions of this model with a more complete dataset.

---

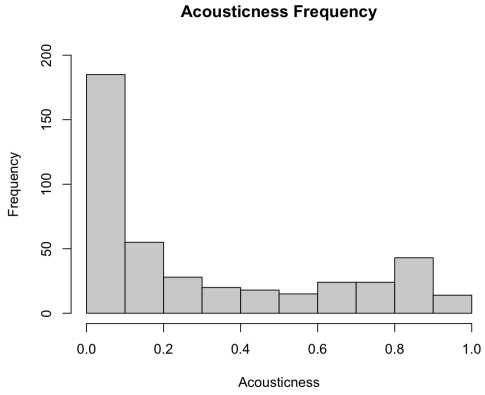
[2] https://musicanddance.uoregon.edu/TaylorSwift

Appendix:

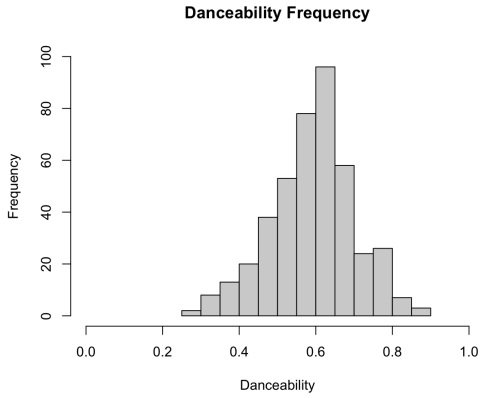

release_date
acousticness
danceability
energy
liveness
loudness
speechiness
tempo
valence
popularity
duration_ms
deluxe
tv

Histograms

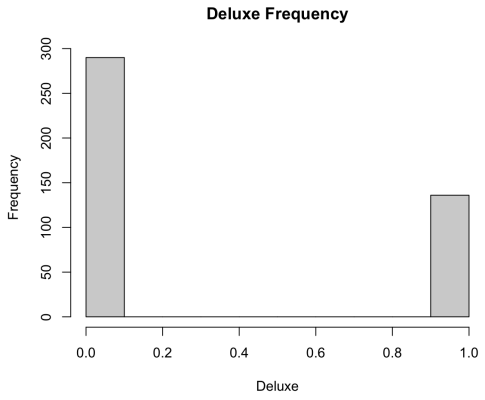
Acousticness



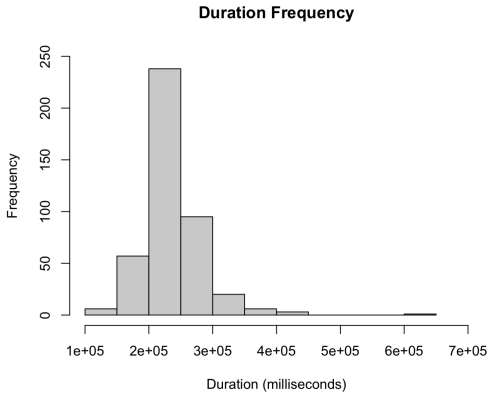
Danceability



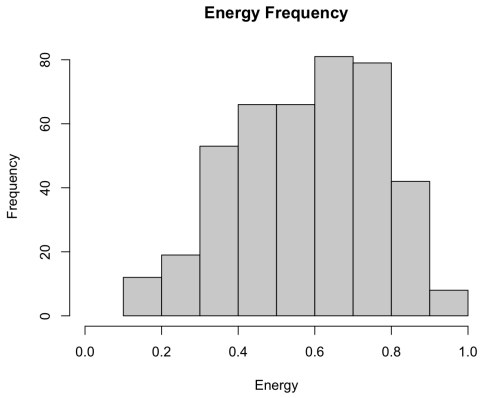
Deluxe



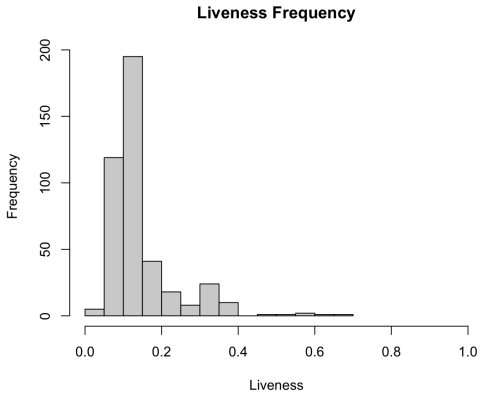
Duration



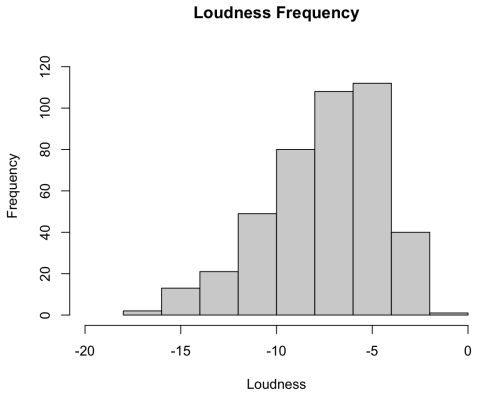
Energy



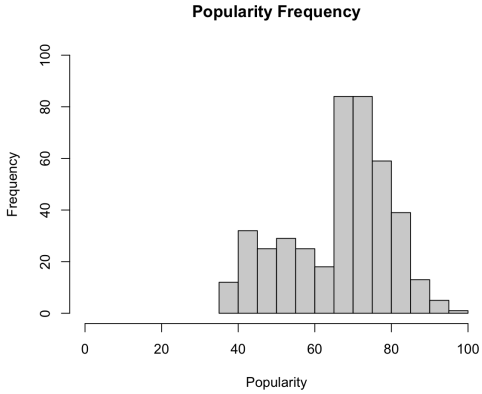
Liveness



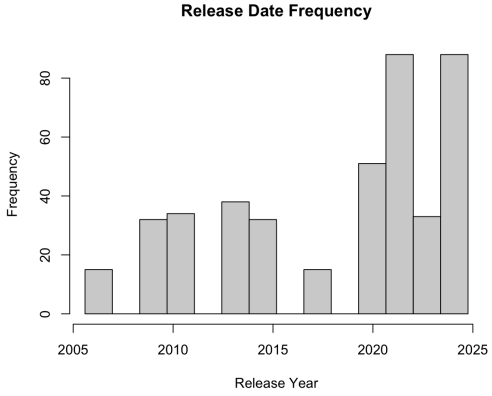
Loudness



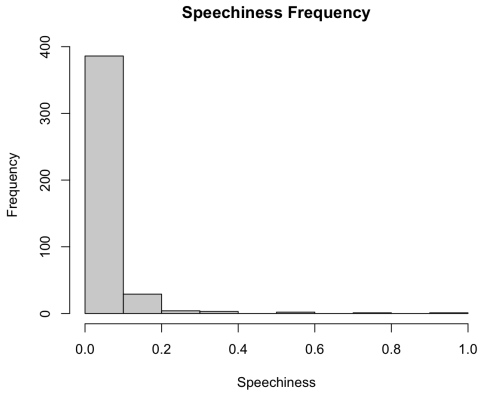
Popularity



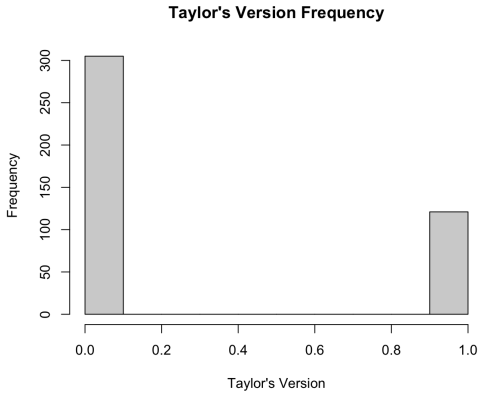
Release Date



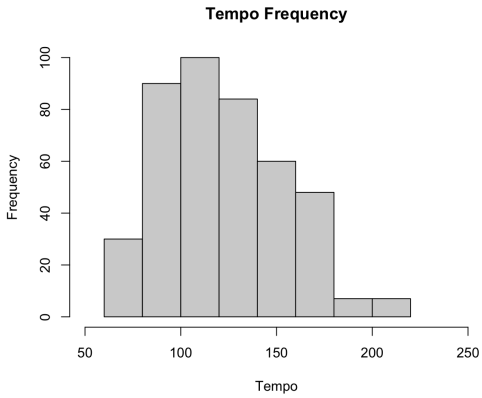
Speechiness



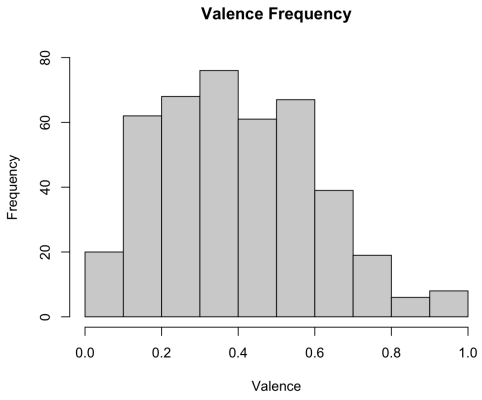
Taylor’s Version



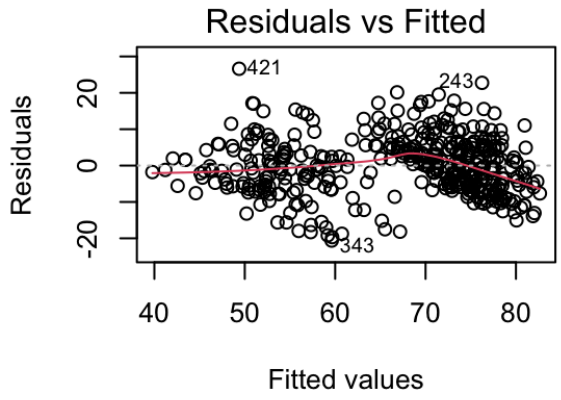
Tempo

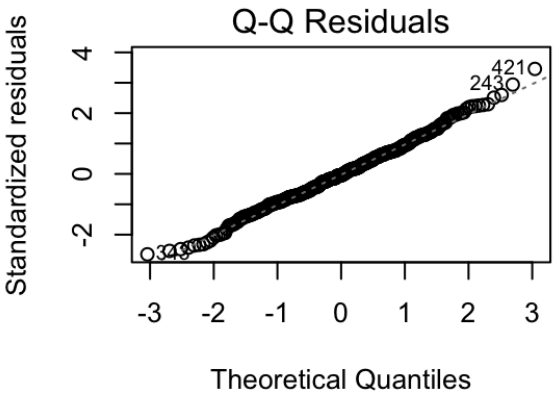


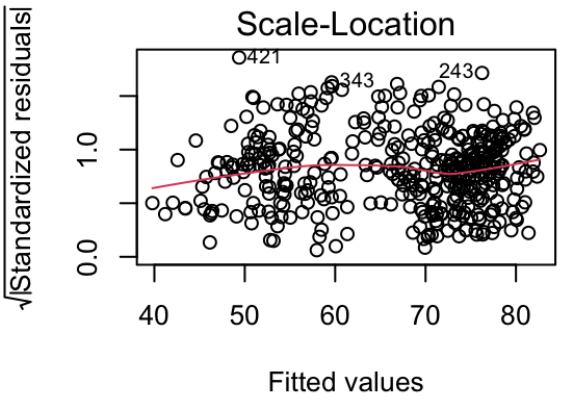
Valence

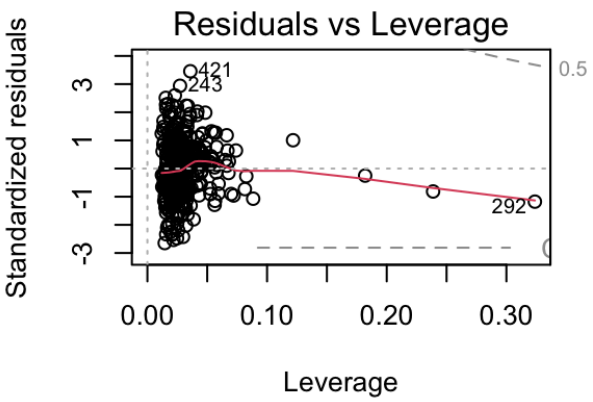


Residual Plots
Full Model









Reduced Model