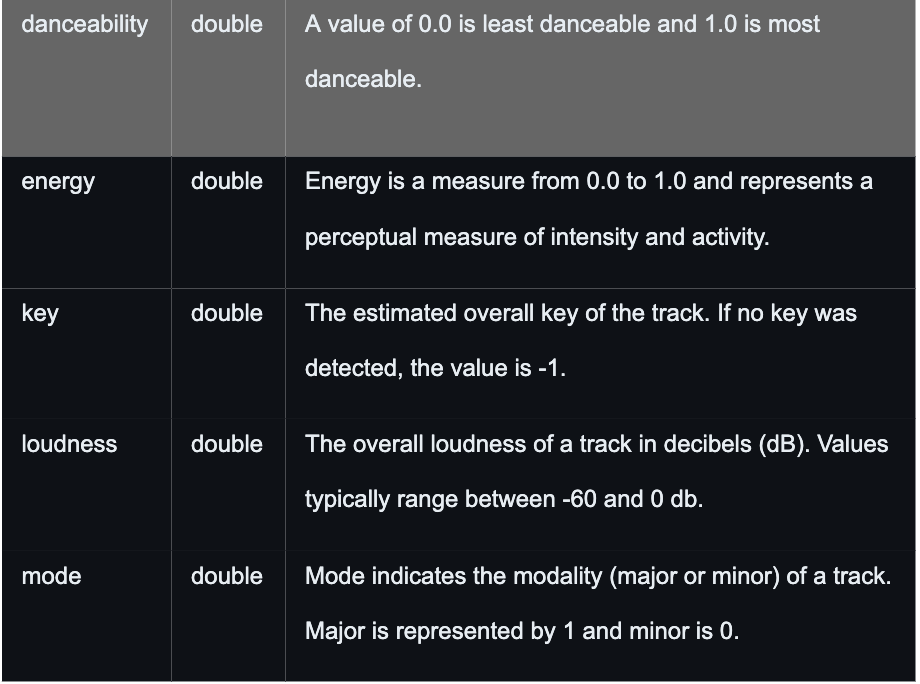
Of all the art forms in existence, music is one that has seen widespread influence and success like few others. Even though TV shows and movies have a ‘visual’ advantage over music, the latter still thrives in the modern age, raking in 31.2 billion USD for FY2022. This success is not just a fluke though. There definitely is something to be said about the ability of music to transcend the failings of spoken language through means of varied instrumentation and above all, the sheer genius of certain artists to manipulate voices and instruments to make songs transcendent of their time. For example, the song Bohemian Rhapsody by British band Queen stops me in my tracks just as much today as it did 5 years ago when I first discovered it. A lot of memories (be they happy or sad) are likely associated with certain songs as well, making music an integral part of the lives of many people.

The seeming nobility of music is swiftly brought down by the nature of the ‘industry’ that generates it. Yes, I use the word ‘generates’ quite consciously, and this is most certainly the result of musical endeavors being treated more as commercial opportunities rather than a means of expression. Instead of moving into the clear negatives created as a result of the  , it would be more fitting to talk about one of the benefits of commercializing music: accessibility. This brings us to music-streaming services, the modern embodiment of accessibility. These services offer at a nominal price, the access to digital libraries of a vast ocean of music from different genres. The biggest player in this space is Spotify, a company started in 2006 by Swedish entrepreneurs Daniel Ek and Martin Lorentzon. [As of 2023](https://www.driveresearch.com/market-research-company-blog/music-streaming-statistics/#:~:text=Spotify%20holds%20the%20position%20as,million%20monthly%20active%20users%20worldwide.), the service has over 210 million paid subscribers and 510 million monthly active users worldwide. Our data was taken from Tidy Tuesday’s compilation of Spotify’s internal metrics on different aspects of songs such as danceability, key, mode, duration of time, instrumentalness, liveness, and valence, amongst other things. The intention is to use this data to answer the question:

**What qualities of a song are the most important contributers to its danceability score?**

Before proceeding, here is a small list of the aforementioned song traits, courtesy of Tidy Tuesday’s readme page:



If interested, the explanations for other such traits can be found [here](https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-01-21/readme.md).

All in all, there were 32833 observations and 23 columns (16 of which are numeric) in the dataset as imported from Tidy Tuesday. There was no cleaning that needed to be performed, as the data was structured perfectly by default and it is mentioned here that we will randomly sample 10000 observations to use for our tests.

Here are some summary statistics of danceability:

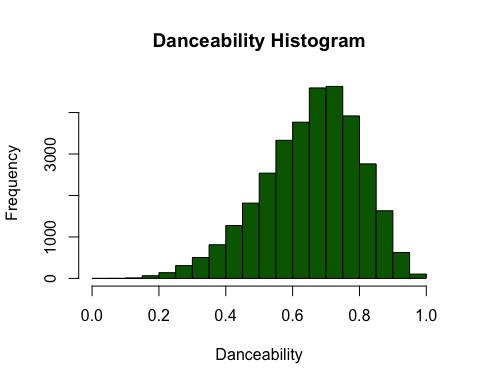
Favstats mosaic

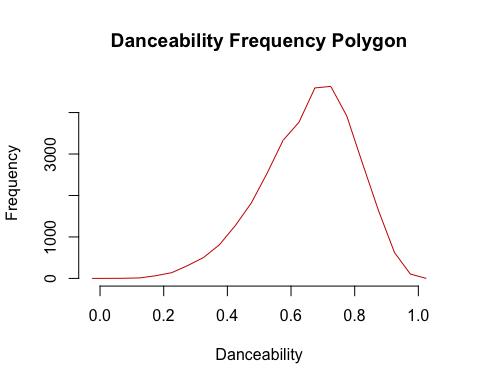
----------------

min 0.0000  
Q1 0.5630  
median 0.6720  
Q3 0.7610  
max 0.9830  
mean 0.6548  
sd 0.1451  
n 32833.0000  
missing 0.0000

----------------

Here is a histogram and frequency polygon of the danceability trait for ease of visualization. We notice a left skew present in the plot and it is unimodal as well.





## Methodology & Results:

First, we will fit an additive multiple linear regression model and proceed from there. We initialize our sampled data and remove the non-numeric parameters. While most of the columns are classified correctly, we correct for “mode” and “key” as factors. This is because “mode” is a 0 or 1 entry and “key” represents the 11 keys in music.

### Naive Linear Model:

Here is our model and ANOVA table:

Call:  
lm(formula = danceability ~ ., data = music\_reduced)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.59813 -0.07675 0.01342 0.08896 0.36148   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 9.010e-01 1.417e-02 63.599 < 2e-16 \*\*\*  
track\_popularity 2.197e-04 5.280e-05 4.161 3.20e-05 \*\*\*  
energy -2.393e-01 1.131e-02 -21.168 < 2e-16 \*\*\*  
key1 2.491e-02 5.382e-03 4.629 3.72e-06 \*\*\*  
key2 -1.986e-03 5.848e-03 -0.340 0.7342   
key3 -3.809e-02 8.444e-03 -4.511 6.52e-06 \*\*\*  
key4 -1.423e-02 6.302e-03 -2.258 0.0240 \*   
key5 3.523e-03 6.080e-03 0.579 0.5623   
key6 2.180e-03 6.013e-03 0.363 0.7169   
key7 6.869e-03 5.651e-03 1.216 0.2242   
key8 2.612e-03 6.136e-03 0.426 0.6704   
key9 -1.184e-02 5.737e-03 -2.063 0.0391 \*   
key10 4.289e-03 6.293e-03 0.682 0.4956   
key11 -3.504e-03 5.792e-03 -0.605 0.5453   
loudness 9.068e-03 6.018e-04 15.068 < 2e-16 \*\*\*  
mode1 -1.541e-02 2.697e-03 -5.711 1.15e-08 \*\*\*  
speechiness 2.480e-01 1.279e-02 19.386 < 2e-16 \*\*\*  
acousticness -1.009e-01 6.973e-03 -14.478 < 2e-16 \*\*\*  
instrumentalness 7.322e-02 6.112e-03 11.980 < 2e-16 \*\*\*  
liveness -8.817e-02 8.415e-03 -10.477 < 2e-16 \*\*\*  
valence 2.292e-01 5.685e-03 40.322 < 2e-16 \*\*\*  
tempo -9.035e-04 4.758e-05 -18.990 < 2e-16 \*\*\*  
duration\_ms -1.193e-07 2.189e-08 -5.451 5.13e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1275 on 9977 degrees of freedom  
Multiple R-squared: 0.2497, Adjusted R-squared: 0.248   
F-statistic: 150.9 on 22 and 9977 DF, p-value: < 2.2e-16

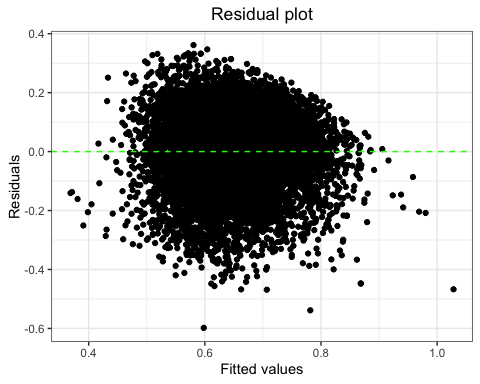
---------------------

Analysis of Variance Table  
  
Response: danceability  
 Df Sum Sq Mean Sq F value Pr(>F)   
track\_popularity 1 1.451 1.4510 89.309 < 2.2e-16 \*\*\*  
energy 1 1.090 1.0903 67.109 2.883e-16 \*\*\*  
key 11 3.059 0.2781 17.118 < 2.2e-16 \*\*\*  
loudness 1 2.125 2.1252 130.809 < 2.2e-16 \*\*\*  
mode 1 0.762 0.7616 46.877 8.002e-12 \*\*\*  
speechiness 1 6.245 6.2447 384.368 < 2.2e-16 \*\*\*  
acousticness 1 1.704 1.7043 104.902 < 2.2e-16 \*\*\*  
instrumentalness 1 0.334 0.3342 20.569 5.820e-06 \*\*\*  
liveness 1 2.729 2.7293 167.990 < 2.2e-16 \*\*\*  
valence 1 28.140 28.1402 1732.057 < 2.2e-16 \*\*\*  
tempo 1 5.818 5.8179 358.101 < 2.2e-16 \*\*\*  
duration\_ms 1 0.483 0.4827 29.713 5.128e-08 \*\*\*  
Residuals 9977 162.093 0.0162   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

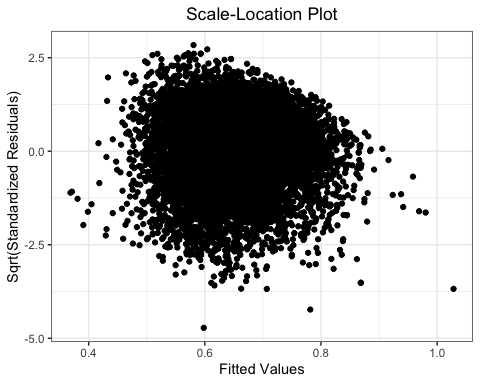
Technically, we get the answer to our question here by noticing from the ANOVA that all the predictors are highly significant. The model iteslf is also significant. However, in regards to model utility, the doesn’t seem convincing as it implies that our model is only able to explain 24.8% of the data.

Here are the residual plots to help with checking model assumptions:

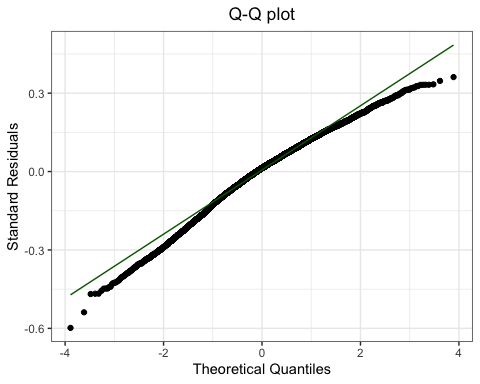
1. **Independence**: This assumption is clearly violated as there are multiple songs by the same individual in the dataset. Even though we have randomly sampled our set, there is a non-zero probability of this happening. So, our data is not independent.
2. **Linearity**: From the residual plot below, we don’t see any non-linear pattern that would raise any flags. So, we consider this assumption to hold.



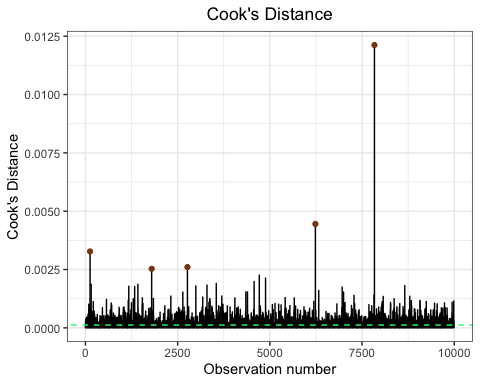
1. **Homoscedasticity**: Again, using the same residual plot, we notice the lack of a funnel-like shape here. So, this assumption is not violated.



1. **Normality**: As we have more than 5000 observations, solely relying on a quantitative test of normality would not be the best way to go here. This is because normality tests can be overly sensitive at large sample sizes. So, we will rely on a qualitative way to check this assumption. Using the plot below, we see that there is noticeable deviation at the tails. While one can go either way on this, we will consider this a violation of the assumption.



1. **Cook’s distance (Influential points and outliers)**: Clearly, there are issues with influential points in our model. I have marked the top five most influential points with brown dots, but even if you ignored these points, there are multiple other points that lie beyond the green threshold.



Here is the largest influential points from our sampled data. Interestingly, it is a four second song by Japanese band DREAMS COME TRUE, who have [1.5 Million followers](https://open.spotify.com/artist/2mJOGcLR3aCHkM1uAF93or) as of 11/29/2023.

1w6nRCU68klqNfYaaVP2j <- track\_id   
Hi, How're You Doin'? <- track\_name  
DREAMS COME TRUE <- track\_artist  
0 <- track\_popularity  
4wdK52JVu5GzhxW3RCZ3AV <- track\_album\_id  
Dreams Come True <- track\_album\_name  
1989-03-21 <- track\_album\_release\_date   
City Pop 1985 シティーポップ <- playlist\_name  
3j2osvmecEao5nmo9jZ5df <- playlist\_id  
rock <- playlist\_genre  
album rock <- playlist\_subgenre  
0 <- danceability  
0.315 <- energy   
1 <- key   
-26.087 <- loudness   
1 <- mode   
0 <- speechiness   
0 <- acousticness   
0 <- instrumentalness   
0 <- liveness   
0 <- valence   
0 <- tempo   
4000 <- duration\_ms

Overall, the point of fitting the linear model was in order to ‘test the waters.’ To improve upon this, we will use a lasso selection process, which will appropriately inflate/deflate the ’s depending upon their relative importance/unimportnace.

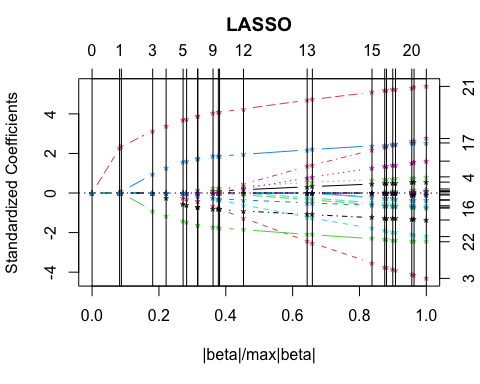
Call:  
lars(x = music\_matrix, y = music\_reduced$danceability, type = "lasso",   
 intercept = TRUE)  
R-squared: 0.25   
Sequence of LASSO moves:  
 valence speechiness tempo liveness energy duration\_ms key1  
Var 21 17 22 20 3 23 4  
Step 1 2 3 4 5 6 7  
 track\_popularity acousticness loudness instrumentalness mode1 key3 key9  
Var 2 18 15 19 16 6 12  
Step 8 9 10 11 12 13 14  
 key4 key7 key10 key11 key2 key5 key8 key6  
Var 7 10 13 14 5 8 11 9  
Step 15 16 17 18 19 20 21 22

LASSO Coefficients

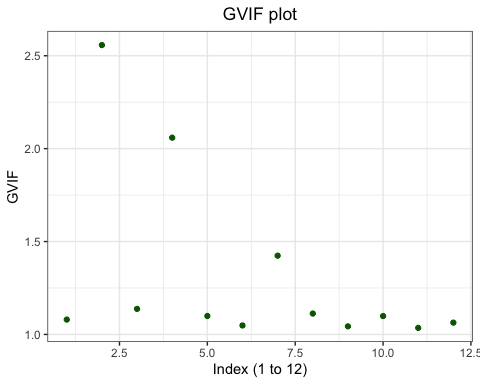
------------------

(Intercept) track\_popularity energy key1   
 0.000000e+00 2.196882e-04 -2.393059e-01 2.491412e-02   
 key2 key3 key4 key5   
 -1.985946e-03 -3.809129e-02 -1.423014e-02 3.523078e-03   
 key6 key7 key8 key9   
 2.180248e-03 6.868816e-03 2.611649e-03 -1.183677e-02   
 key10 key11 loudness mode1   
 4.289135e-03 -3.503741e-03 9.067841e-03 -1.540652e-02   
 speechiness acousticness instrumentalness liveness   
 2.480393e-01 -1.009498e-01 7.322137e-02 -8.816558e-02   
 valence tempo duration\_ms   
 2.292154e-01 -9.035249e-04 -1.193453e-07

------------------



Here, we see that our has unfortunately only improved slightly (0.25) compared to the previous model (0.248). Could there be an issue of multicollinearity? To test this, we plot the GVIF’s (Generalized Variance Inflation Factor) for each predictor.



The largest GVIF is approximately 2.5, which implies moderate correlation to other predictors. For most predictors the GVIF is marginally above 1, implying little to no correlation with other predictors. Thus, with a little hand-waving, we can scratch out multicollinearity as being a problem here.

## Discussion & Conclusion

So, where does that leave us? Well, not far from where we started! LASSO did not lead us to a better place (in terms of at least) as we had hoped. With that said, issues such as those of normality and influential points are not straightforward to solve and are to be handled with a lot of care. Finally, our decision to use the predictors that ended up in our naive model was a choice that was well-deliberated. The assumption was that different aspects of a song (be they instrumental or otherwise) ultimately contribute to its danceability metric. However, we are aware of the possibility that this may not be the best move. Perhaps people consider shorter songs to be more danceable. Or something else like so. The mentality of a culture is much more difficult to predict, and a more in-depth analysis would be required to understand this complex effect.

Assuming one wants to still work with our model as it is now, there are three remedies that would be use: Response transformations, considering interaction terms, and trying a non-parametric model. We will briefly present our reasoning for the viability of the three processes.

1. **Response transformations**: One would go about this by first conducting a boxcox procedure, which would suggest an appropriate transformation on the response and then create a new linear model using the transformed response. What’s the problem here? Well, boxcox needs strictly positive values in the response. Even zeroes are not allowed. Unfortunately there are many songs here that have zeroes for their danceability score. Thus, unless you’re okay with deleting a chunk of songs, this approach is not recommended at all.

* Here’s a sample of what happense if you use R to do the boxcox procedure.
* MASS:: boxcox(object = music\_lm\_naive)
* Error in boxcox.default(object = music\_lm\_naive): response variable must be positive

1. **Considering Interaction terms**: This approach is a double edged sword in our opinion. On one hand, the shoots up proportional to the order of interactions in the model. For example, we ran the linear model with first order interactions and the new model was able to explain 31% of the total variation, which is about a 10% increase from before. Here’s that in R:

* # Neat little trick to generate first order interactions without writing everything out in R  
  music\_lm\_interact <- lm(danceability~.^2,data = music\_reduced)  
  summ <- summary(object = music\_lm\_interact)  
  # Adjusted R squared  
  summ$adj.r.squared
* [1] 0.3109649

Running the LASSO on this model, the increases to about 35%. All seems great so far.  
However, one big problem is in regards to overfitting the model with useless predictors. Take a look at the last ten predictors and compare it to our naive model from before:

tail(anova(music\_lm\_interact),10)

Analysis of Variance Table  
  
Response: danceability  
 Df Sum Sq Mean Sq F value Pr(>F)   
instrumentalness:valence 1 0.303 0.30255 20.3229 6.617e-06 \*\*\*  
instrumentalness:tempo 1 0.039 0.03913 2.6284 0.1050   
instrumentalness:duration\_ms 1 0.491 0.49129 33.0012 9.485e-09 \*\*\*  
liveness:valence 1 0.007 0.00746 0.5014 0.4789   
liveness:tempo 1 0.039 0.03949 2.6525 0.1034   
liveness:duration\_ms 1 0.001 0.00060 0.0400 0.8415   
valence:tempo 1 0.008 0.00782 0.5251 0.4687   
valence:duration\_ms 1 0.588 0.58832 39.5193 3.387e-10 \*\*\*  
tempo:duration\_ms 1 0.002 0.00189 0.1269 0.7217   
Residuals 9801 145.907 0.01489   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Thus, one should use this to make selective interactions that would strengthen the overall model.

1. **Non-parametric model**: This is much more random suggestion as compared to previous two. However, using the logic mentioned in (1), one should be able to see why a non-parametric approach may be beneficial.

<Add overall conclusions>