Project

Sebastian Bonet

2022-11-23

I. Introduction

Music is a constantly growing and changing field with new artists and songs popping up daily. As more and more people have begun using different applications such as Apple Music, Spotify, etc., more and more data has become available to determine the individual characteristics that make up a song. When breaking down a song by its musical attributes, we can determine whether these attributes are correlated.

II. Goal

To be specific, the goal of this project is to build a regression model to define the relationship between song attributes and a song's popularity. If our client were a record label, for example, this information would be vital when trying to produce music that will become popular and therefore sell more and generate more revenue for the record label.

III. Data Description

The data gathered came from the Spotify API. 12,000 songs were scraped from Spotify spanning the years 2010-2022. These songs contain specific song attributes such as danceability, energy, key, loudness, and more, all of which are quantitative variables. They also contain a variable for popularity, which we will be attempting to predict through our model.

IV. Executive

The attributes provided by the Spotify API are not good predictors of a Song's popularity when fitting a multiple linear regression model.

The reason for this disconnect is that the regression assumptions are violated and can't be fixed by data transformations. This results in a model with an adjusted R^2 close to 0 which means that the points don't lie on a straight line making prediction difficult.

In our analysis we analyzed ways to improve our model such as data transformations, correlation, and influential points which proved ineffective.

V. Regression Analysis

Inital Model

The initial model includes the following predictors:

1. Acousticness
2. Danceability
3. Energy
4. Instrumentalness
5. Key
6. Liveness
7. Mode
8. Loudness
9. Tempo
10. Time_signature
11. Speechiness
12. Valence
13. Duration_s

The response variable in our model is Popularity.

Regression Assumptions:

1. Independence Assumption

Independence assumption was tested utilizing Durbin-Watson test:

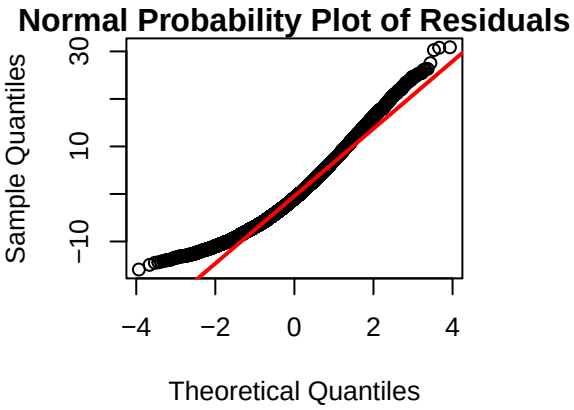
```
##
##  Durbin-Watson test
##
## data:  model
## DW = 0.63727, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

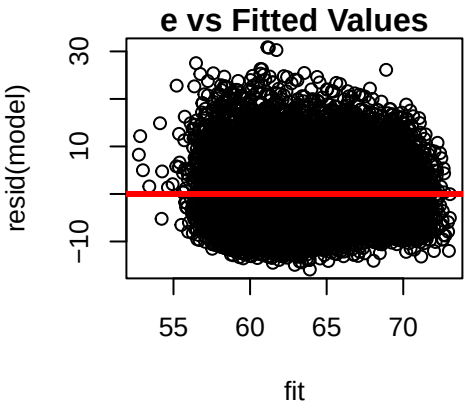
The independence assumption is met.

Reason: The p-value is significantly smaller than $\alpha = 0.05$, therefore we fail to reject the null hypothesis meaning that there is no autocorrelation

2. Normality Assumption
3. Constant Variance Assumption
4. mean of potential error terms is 0

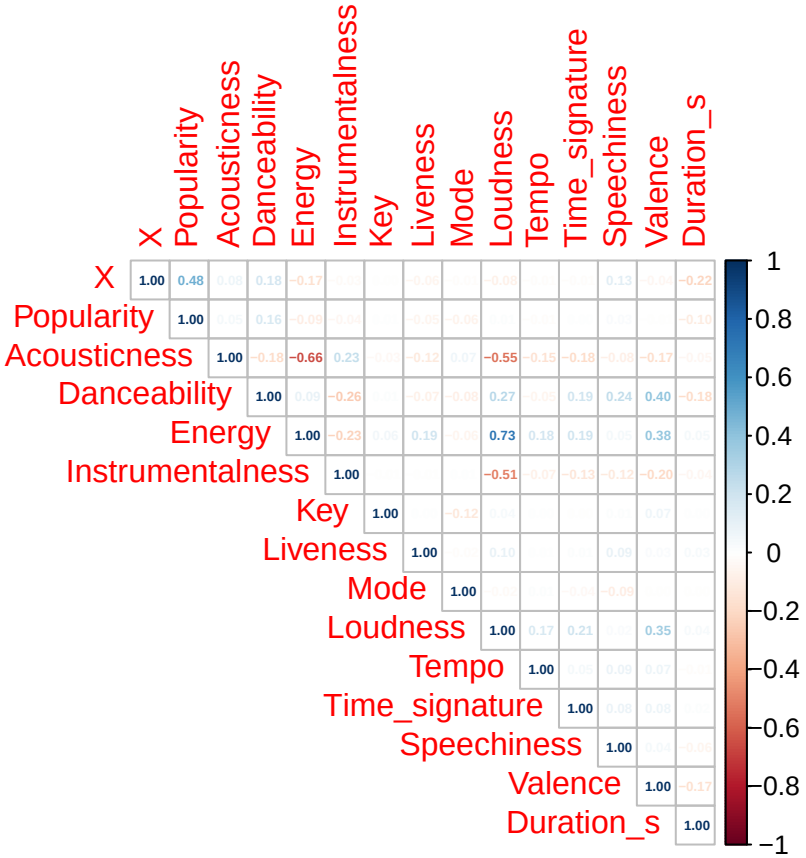
Assumptions 2-4 can be verified utilizing plots:

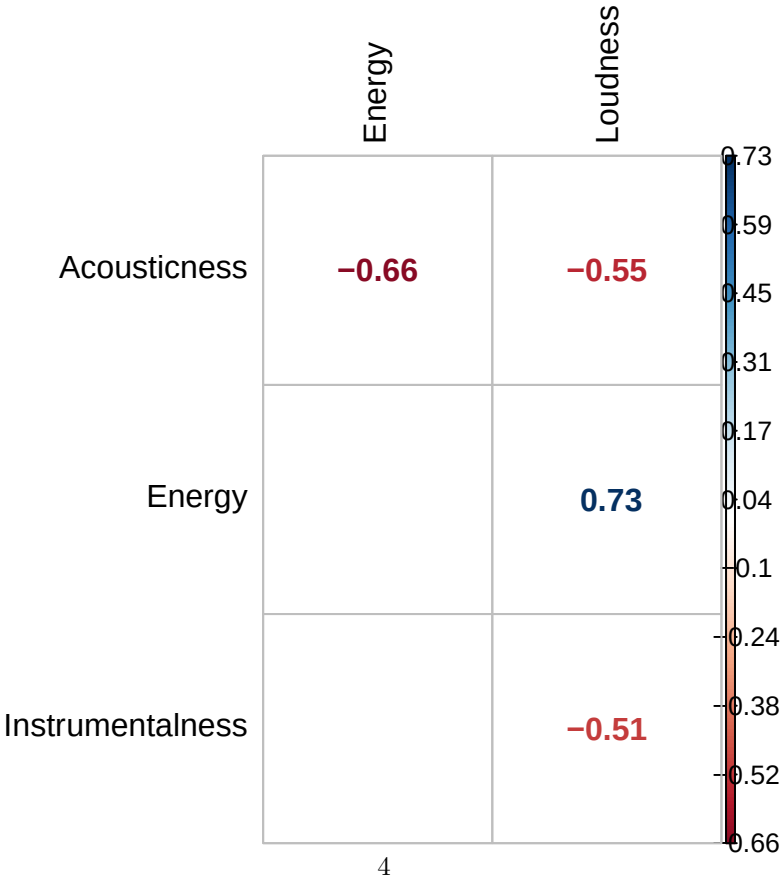




2. Normality assumption is not met. Reason: Residual points don't follow a straight line in the normal probability plot (can also be verified using AD-Test)
3. Constant variance assumption is not met. Reason: variance increases as x increases
4. Mean of potential error terms is 0 assumption is met (see variance plot)

Correlation Analyis





```
##                X     Acousticness     Danceability           Energy
##         1.131872         1.905975         1.588921         3.406517
## Instrumentalness              Key         Liveness             Mode
##         1.470354         1.021739         1.064799         1.035949
##         Loudness            Tempo   Time_signature      Speechiness
##         3.080710         1.068480         1.086373         1.137387
##          Valence       Duration_s
##         1.512000         1.114270
```

We can see that the variables with highest correlation are acousticness, energy instrumentalness, and loudness. However this is not too concerning since their Variance Inflation Factors are lower than 10 and therefore not severe.

The mean of the Variance Inflation Factors is not much bigger than 1 (1.54) meaning that correlation is not severe.

Outlier Analysis

Using leverage values we find that around 7% of observations are outliers with respect to x

Using studentized residuals we find some evidence that around 4% of observations are outliers with respect to y.

Using deleted studentized residuals we find strong evidence that around 1% of observations are outliers with respect to x.

Influential Point Analysis

Utilizing Cooks Distance analysis we conclude that there are no influential points.

Both forward and backward stepwise model selection yield the same model:

Popularity ~ Acousticness + Danceability + Energy + Liveness + Mode + Loudness + Tempo + Time_signature + Valence + Duration_s

Key, speechiness, and instrumentalness are removed from the model.

Model Selection

The model mentioned above (10 independent variables) is the best model beacause it has the smallest AIC, the smallest Mallows'C_p, and it was the model selected by the stepwise model selection function.

```
##
## Call:
## lm(formula = Popularity ~ ., data = input)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8987  -5.2315  -0.9204   4.3371  30.8674
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.969e+01  9.671e-01  61.719  < 2e-16 ***
## X                 1.090e-03  1.958e-05  55.667  < 2e-16 ***
```

```
## Acousticness      9.284e-01  3.105e-01   2.990   0.0028 **
## Danceability      3.920e+00  4.972e-01   7.884 3.45e-15 ***
## Energy           -2.699e+00  5.663e-01  -4.765 1.91e-06 ***
## Instrumentalness  6.496e-01  4.177e-01   1.555   0.1199
## Key               5.528e-03  1.798e-02   0.307   0.7585
## Liveness         -5.151e-01  4.594e-01  -1.121   0.2622
## Mode             -1.000e+00  1.356e-01  -7.380 1.69e-13 ***
## Loudness          2.220e-01  2.814e-02   7.889 3.32e-15 ***
## Tempo             1.170e-03  2.182e-03   0.536   0.5918
## Time_signature   -1.191e-01  1.639e-01  -0.726   0.4676
## Speechiness      -3.855e+00  6.367e-01  -6.055 1.45e-09 ***
## Valence          -5.355e-01  3.289e-01  -1.628   0.1036
## Duration_s        1.183e-03  9.484e-04   1.247   0.2123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.98 on 11977 degrees of freedom
## Multiple R-squared:  0.2492, Adjusted R-squared:  0.2483
## F-statistic:   284 on 14 and 11977 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = Popularity ~ . - Time_signature - Speechiness -
##     Instrumentalness, data = input)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5723  -5.2706  -0.9268   4.3512  30.9929
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.962e+01  7.645e-01  77.986  < 2e-16 ***
## X             1.079e-03  1.953e-05  55.240  < 2e-16 ***
## Acousticness  9.967e-01  3.106e-01   3.209  0.00133 **
## Danceability  3.059e+00  4.762e-01   6.424 1.38e-10 ***
## Energy       -2.760e+00  5.512e-01  -5.007 5.62e-07 ***
## Key           5.513e-03  1.801e-02   0.306  0.75957
## Liveness     -8.234e-01  4.573e-01  -1.801  0.07180 .
## Mode         -9.465e-01  1.355e-01  -6.985 3.00e-12 ***
## Loudness      2.166e-01  2.476e-02   8.749  < 2e-16 ***
## Tempo        -4.128e-04  2.172e-03  -0.190  0.84926
## Valence      -4.179e-01  3.274e-01  -1.276  0.20189
## Duration_s    1.079e-03  9.475e-04   1.139  0.25486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.992 on 11980 degrees of freedom
## Multiple R-squared:  0.2465, Adjusted R-squared:  0.2459
## F-statistic: 356.4 on 11 and 11980 DF,  p-value: < 2.2e-16
```

VI. Conclusion

The key takeaway from our analysis is that the attributes provided by the Spotify API are not useful when building a linear regression to predict the popularity of a song.

The low R^2 means that the observations don't follow a straight line and therefore can't be predicted accurately using linear regression.

The regression assumptions did not hold and even after attempting transformations of the data.

Furthermore we can rule out influential points being the reason for inaccuracy since we did not find any when performing Cooks Distance Analysis.

Summary (Not included in final report)

Regression Assumptions:
1 independence assumption-> pass using dwtest
2 Normality -> failed using adtest
3 constant variance assumption -> some funneling (residual plot)
4 potential error term values has a mean equal to 0 -> residual plot

No transformation is great

Correlation is not severe since VIF « 10 for all variables However we should not that mean of vifs is 1.5 which is greater than 1 but still relatively small

6% of observations are outliers with respect to x There is strong evidence that 0.7% of observations are outliers with respect to y No observations are influential points since Cooks Distance is less than F0.8 for all observations

Same model for forward and backward: Acousticness + Danceability + Energy + Liveness + Mode + Loudness + Tempo + Time_signature + Valence + Duration_s

Not included:
key
speechiness
instrumentalness

All three were not significant as seen in summary output however other variables that were not significant were included # Remaining Model selection -> 10 variables as seen by AIC & Mallows Cp

Test how good the model is