The relationship between the popularity of Top 50 Spotify songs and their 5 essential properties

Langyu Qie

December 17,2019

1. Introduction of spotify songs and 5 of their properties

1.1 Data

I found a dataset of the TOP 50 spotify songs with their 5 properties and their popularity level on Kaggle's dataset. Here is the link to my dataset:

https://www.kaggle.com/leonardopena/top50spotify2019

As the picture shows below, the dataset we get contains 50 observations and 6 variables(the popularity and 5 properties).

```
> head(spop)
  BPM Energy Valence Acousticness Speechiness Popularity
1 101     50      10           10           5         85
2  93     45      14           12          15         86
3  98     59      18            2          15         84
4 150     65      18           45           7         86
5 180     64      23            2          29         85
6 176     62      24           60          31         90
```

The energy, BPM, valence, acousticness, and speechiness of a song are the five basic components which identify songs from each other, they decide the beats, melody, the number of words and the energy conveyed in a song.

1.2 BPM

BPM is the acronym of Beats Per Minute. Beats Per Minute decides how many beats we can count in one minute, which determines the tempo of a song.

1.3 Energy

Represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores

low on the scale. The higher the value, the more energetic the song.

1.4 Valence

Describes the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

1.5 Acousticness

A confidence measure of whether the track is acoustic.

1.6 Speechiness

This detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the higher the attribute value.

1.7 Popularity

The higher the value the more popular the song is.

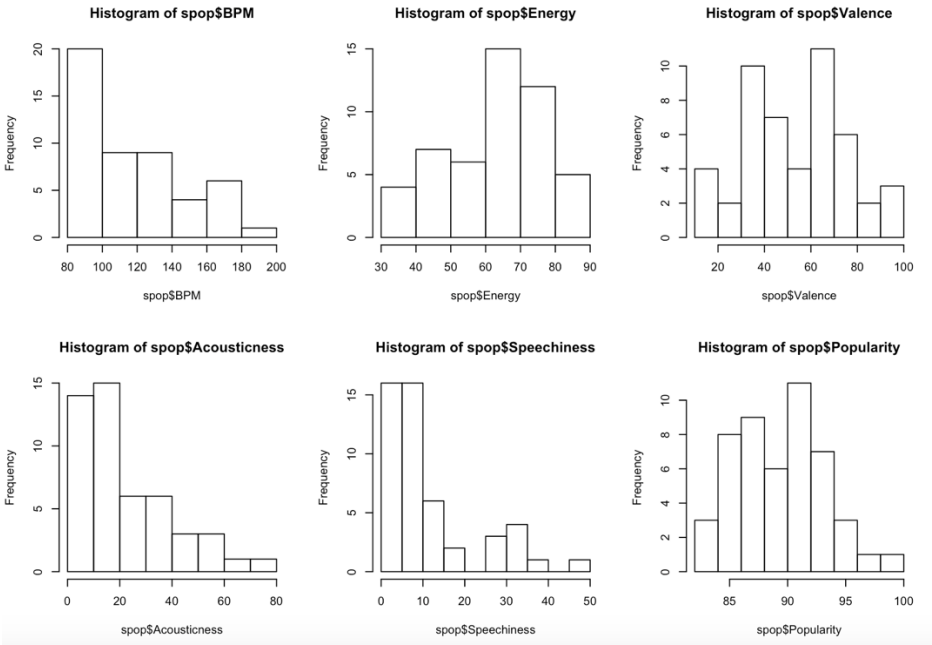
2. Objective

The goal of this project is to find a suitable linear model to explain the relationship between the popularity of top 50 Spotify songs and their 5 properties, which is represented by BPM, energy, valence, acousticness and speechiness. We look for methods to evaluate and optimize the model we find.

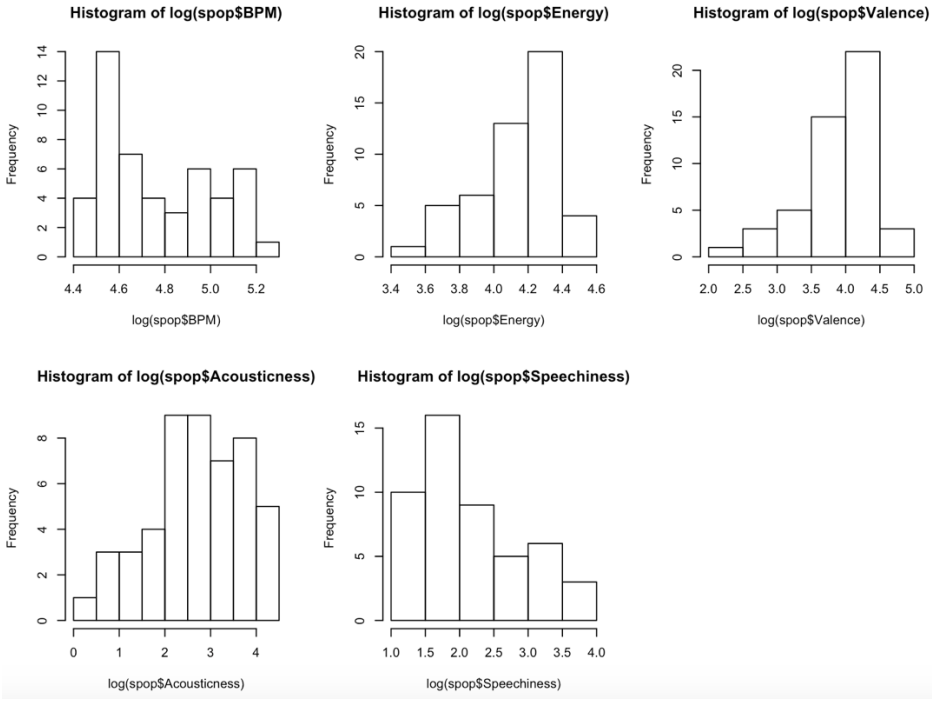
3. Analysis

3.1 Normality Check

I plot the histogram of the variables and try to see if the data distribution is normal



We can see from the histogram that the data of BPM, Acousticness and Speechiness is right-skewed, so we conduct a log transformation to make the data distributed more normally, so that the model fit better. Here is the data distribution after the transformation.



3.2 Linear Regression

We used the log transformed data of 50 observations to fit the linear regression model, and the model is summarized as below:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.3684989  0.0275174 158.754  < 2e-16 ***
BPM          -0.0001504  0.0001700  -0.885  0.38123
Energy        0.0011036  0.0003496   3.156  0.00292 **
Valence       0.0007326  0.0002138   3.427  0.00135 **
Acousticness  0.0012052  0.0002356   5.116 6.92e-06 ***
Speechiness   0.0007853  0.0004758   1.650  0.10615
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.029 on 43 degrees of freedom
Multiple R-squared:  0.5684,    Adjusted R-squared:  0.5182
F-statistic: 11.33 on 5 and 43 DF,  p-value: 5.246e-07
```

Since F-statistic=11.33, p-value<0.0001 and R-square=0.5684, we can see that our model doesn’t fit very well, but is still meaningful.

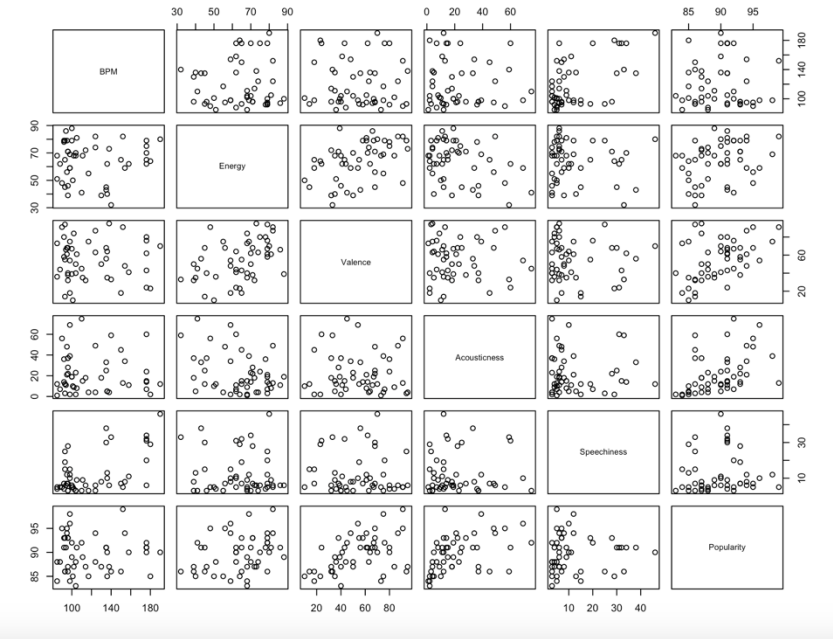
Let’s look at the p values of 5 predictors respectively. As the significance code showed, Factor Energy, Valence and Acousticness are significant. However, the rest two factors don’t have much influence on our model. So we used the MASS function stepAIC, which will evaluate a full range of models and spit out which one seems optimum according to the AIC. In this case I tried all the three directions and they all give the same result as following:

```
> summary(mod.fw)

Call:
lm(formula = Popularity ~ Valence + Acousticness + Energy, data = spop)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0884 -1.7165  0.4909  1.6652  6.1584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.03665    2.05365  37.999  < 2e-16 ***
Valence       0.06423    0.01944   3.305  0.00187 **
Acousticness  0.10913    0.02145   5.087 6.87e-06 ***
Energy        0.09197    0.03155   2.915  0.00553 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.642 on 45 degrees of freedom
Multiple R-squared:  0.5388,    Adjusted R-squared:  0.5081
F-statistic: 17.53 on 3 and 45 DF,  p-value: 1.112e-07
```

As the F-statistic increases and p-value decreases, we can conclude that the new regression model with three predictors is more meaningful and significant than the previous model.

3.3 Linearity and Colinearity

We then analyze the linearity of the variables. We found there is a clear linear relationship between Popularity and the predictors, especially Acousticness and Valence. But there is also some relationship between our predictors.



By using vif() function in R, we find that the vif values are not that large. However, this doesn't mean that there are no correlations between each pair of predictors. We can directly see the pairwise scatterplot graph and find that there are some obvious collinearities among predictor Energy and Valence. We will talk about these correlations following.

```
vif(fm1)
       BPM       Energy      Valence Acousticness  Speechiness
  1.470231     1.440128     1.256344     1.146221     1.478416
```

```
> cor(spop[1:5])
                    BPM      Energy     Valence Acousticness  Speechiness
BPM          1.00000000  0.04375559 -0.01158582 -0.031449597  0.557051878
Energy       0.04375559  1.00000000  0.43881959 -0.339891653 -0.089859668
Valence     -0.01158582  0.43881959  1.00000000 -0.052323306 -0.053241746
Acousticness -0.03144960 -0.33989165 -0.05232331  1.000000000  0.008293376
Speechiness  0.55705188 -0.08985967 -0.05324175  0.008293376  1.000000000
```

We can see from the covariance matrix that there is some linear relationship between Energy and Valence, BPM and Speechiness.

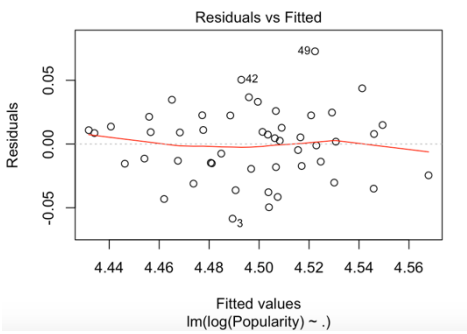
1) For Energy and Valence, It's clear that we normally get more Energy from songs that convey more positiveness. so the pair graph of Energy and Valence goes up, they have a positive linear relation.

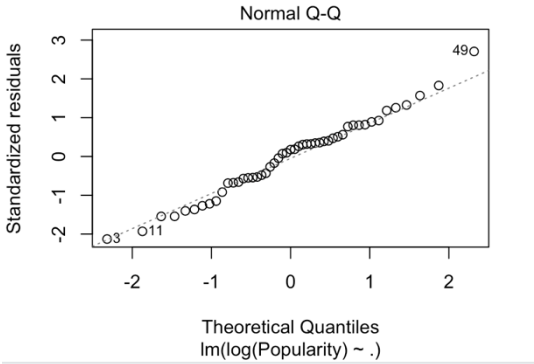
2) For BPM and Speechiness, we know that words in songs are always written at tempo points, so the songs can have a rhythm to sing along. For example, for rap music, number

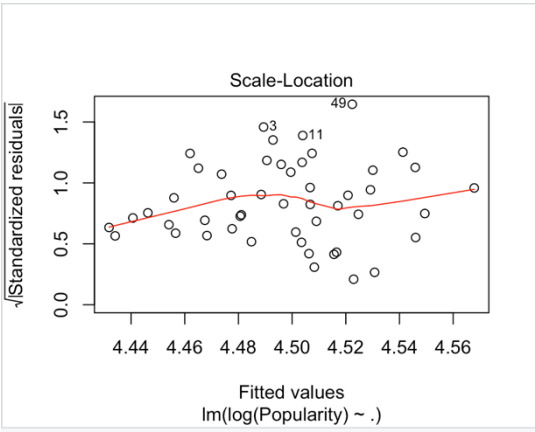
of words in a song tend to be larger when the beat is faster. Hence, this can explain the collinearity between BPM and Speechiness.

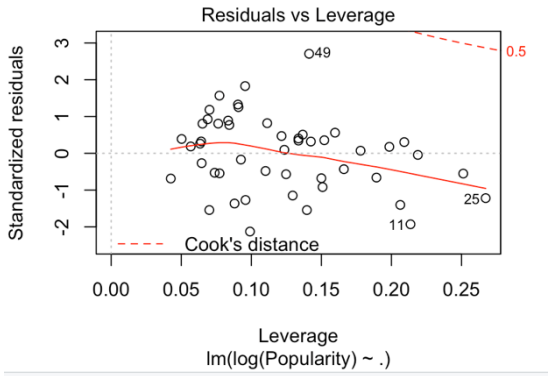
3.4 Q-Q Plot

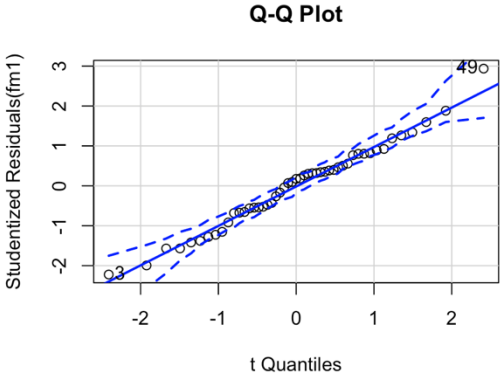
Then we checked the normality of the residual of our second model and found it's quite normal.









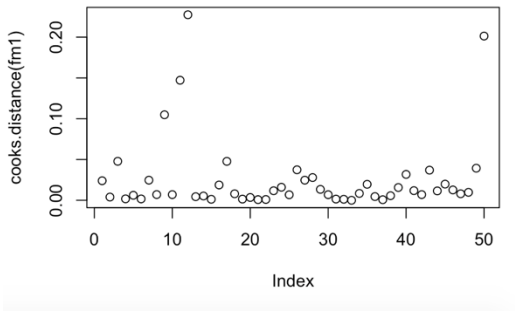


3.5 Outlier Identification

Now we look at the standardized residual of our regression model:

```
> stdres.Pop
          1           2           3           4           5           6
-1.381612233 -0.409613944 -1.713449723  0.265741190 -0.431013143  0.248076921
          7           8           9          10          11          12
-1.000582360  0.515788126 -1.662542327 -0.589125187 -1.802927378 -2.296563050
         13          14          15          16          17          18
-0.580547423 -0.619989434  0.219530536 -1.121451519 -1.343502234  0.721678167
         19          20          21          22          23          24
 0.362183956 -0.683196332 -0.263738211  0.332539019 -0.631016943  1.003946047
         25          26          27          28          29          30
 0.764596149 -0.815826409 -0.805174551 -0.715754013  0.801387296  0.423790193
         31          32          33          34          35          36
 0.376220105  0.233150644 -0.004116411  0.551277008 -1.077076847 -0.320809475
         37          38          39          40          41          42
 0.271448563  0.495339450  1.116225629  1.508284037  1.013078197  0.489485121
         43          44          45          46          47          48
 1.513717631  0.961544080 -0.818958699  0.918860679  0.429011924  0.554547968
         49          50
 1.753767476  2.711148183
```

We can see from the standardized residual that point 12 and point 50 are likely to be the outlier of our model. Then we looked at the Cook's distance analysis on our model.




```
> cooks.distance(fm1)
          1           2           3           4           5           6
2.395687e-02 3.955135e-03 4.779321e-02 1.768434e-03 6.133612e-03 1.691772e-03
          7           8           9          10          11          12
2.465113e-02 7.011226e-03 1.048579e-01 6.946162e-03 1.471591e-01 2.274202e-01
         13          14          15          16          17          18
4.415234e-03 5.267550e-03 1.225947e-03 1.877471e-02 4.773183e-02 7.975805e-03
         19          20          21          22          23          24
1.468205e-03 3.452946e-03 7.996011e-04 9.750682e-04 1.169499e-02 1.598346e-02
         25          26          27          28          29          30
6.762517e-03 3.738935e-02 2.464951e-02 2.786982e-02 1.337401e-02 6.869165e-03
         31          32          33          34          35          36
1.482747e-03 1.230112e-03 2.678915e-07 8.421506e-03 1.957748e-02 4.549610e-03
         37          38          39          40          41          42
7.036379e-04 5.620975e-03 1.566200e-02 3.173496e-02 1.182085e-02 6.967838e-03
         43          44          45          46          47          48
3.694464e-02 1.150172e-02 1.980854e-02 1.268657e-02 7.978610e-03 9.739500e-03
         49          50
3.934860e-02 2.013105e-01
```

We can also see that point 12 and point 50 have large cook's distance. It means that effect of deleting point 12 and point 50 is pretty big.

3.6 Improve the model

So we try to fit our model again by excluding these two points. Here we give the summary of

our refitted model:

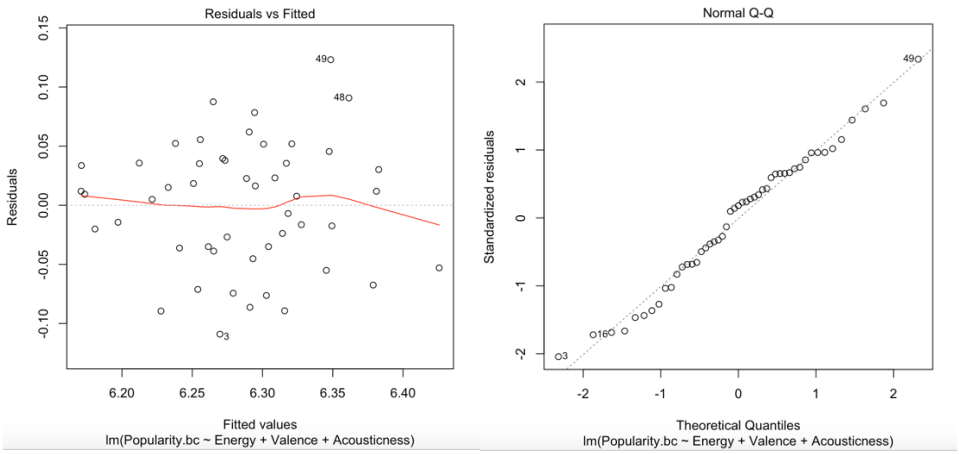
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.27468    2.11905  36.939  < 2e-16 ***
Valence       0.05530    0.01957   2.826  0.00695 **
Energy        0.09215    0.03261   2.826  0.00696 **
Acousticness  0.11476    0.02198   5.220 4.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.731 on 46 degrees of freedom
Multiple R-squared:  0.5068,    Adjusted R-squared:  0.4746
F-statistic: 15.75 on 3 and 46 DF,  p-value: 3.48e-07
```

Comparing to the fm2 we fitted in section 3.1, we see that the model hasn't improved much.

Then I did the box-cox transformation on the data, here is the summary of it:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.0373677  0.0429493 140.569  < 2e-16 ***
Energy       0.0019437  0.0006599   2.946  0.00509 **
Valence      0.0013402  0.0004065   3.297  0.00191 **
Acousticness 0.0022938  0.0004486   5.113 6.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05526 on 45 degrees of freedom
Multiple R-squared:  0.5406,    Adjusted R-squared:   0.51
F-statistic: 17.65 on 3 and 45 DF,  p-value: 1.02e-07
```

Then we look at the residuals and the q-q plot of it:



This time we can see that the r-square has increased, this means that the model fits better. Besides, the increasing of F-statistic and the decreasing of p-value also tells us the model has improved.

4. conclusion

From the regression model we fitted, we can see that the popularity of top 50 Spotify songs

depends highly on their energy, valence and acousticness. Acousticness is the most important element that determines how popular the song is. The second important element is valence, and the third is energy. For an artist who wants to produce a hit, he needs to focus on these three properties of a song, most importantly, the acousticness.

But the r-square of the model we fitted is still not close to 1. So the model still is not a very fitted one, we may need to know more information on the song, like the artist's name, the length of the song, the genre of a song etc. A larger data set may also improve the model.