Project: Spotify Song Popularity

Final Project

The Rainbows: Cason Pierce, Paige Bartusiak, Elizabeth Shaffer

10-9-20

Load Packages

```
library(tidyverse)

spotify_songs <- readRDS("data/spotify_songs.rds")
```

Introduction and Data:

Introduction: For our research we are interested in what makes a song popular! We all love music, and we're interested in seeing what comprises a "popular" song. We will figure this out analyzing specific aspects of songs and their relationship to popularity. Specifically, we want to look at song popularity, valence, tempo, danceability, energy, and playlist genre. We are asking the following question in our research: What is the relationship between valence, tempo, danceability, energy, and playlist genre with the popularity of a song?

A song's popularity is graded on a 0-100 scale, with 100 being the most popular. Valence is the measure of musical positivity on a 0-1 scale, where values closer to 0 indicate a more negative (sad, depressed) tone while larger values closer to 1 indicate a more positive (happy, upbeat) tone. Tempo of a song is measured in beats per minute, in other words, the speed or pace of the song. Danceability is a measure of how easy the song is to dance to on a scale of 0(least danceable) to 1 (very danceable). Energy is a measure of intensity from 0(least active/intense) to 1(most active/intense). Finally, genre is the type of musical category used to group songs of similar styles of music (examples used in this data are EDM, Latin, Pop, R&B, Rock, Rap).

We have made the following hypotheses about the results we expect to find:

- Hypothesis 1: Songs with higher valence are more popular.
- Hypothesis 2: Songs with faster tempo, higher danceability, and more energy are more popular.
- Hypothesis 3: Genre does affect a song's popularity.

Data: The data comes from Spotify via the spotifyr package. The package was authored and collected on January 21, 2020 by Charlie Thompson, Josiah Parry, Donal Phipps, and Tom Wolff to make it easier to get data or general metadata around songs from Spotify's API. This is one of the data sets used for Tidy Tuesday. There are 32833 observations and 23 variables. The rows represent different songs and the columns represent different aspects of the song that include but are not limited to duration, tempo, speechiness, key, and danceability.

Note: There are no entries with missing values for the variables we are using, therefore we don't have to worry about this potential issue. However, we have decided to filter out all songs with popularity scores 0 to 5. We do this because there is a large number of songs with extremely low popularity ratings that were skewing our data.

```
## Rows: 32,833
```

```
## Columns: 23
## $ track_id                <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCY...
## $ track_name              <chr> "I Don't Care (with Justin Bieber) - Loud ...
## $ track_artist            <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", ...
## $ track_popularity        <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58...
## $ track_album_id          <chr> "2oCs0DGTsRO98Gh5ZSl2Cx", "63rPSO264uRjW1X...
## $ track_album_name        <chr> "I Don't Care (with Justin Bieber) [Loud L...
## $ track_album_release_date <chr> "2019-06-14", "2019-12-13", "2019-07-05", ...
## $ playlist_name           <chr> "Pop Remix", "Pop Remix", "Pop Remix", "Po...
## $ playlist_id             <chr> "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1DXcZDD...
## $ playlist_genre          <chr> "pop", "pop", "pop", "pop", "pop", "pop", ...
## $ playlist_subgenre       <chr> "dance pop", "dance pop", "dance pop", "da...
## $ danceability            <dbl> 0.748, 0.726, 0.675, 0.718, 0.650, 0.675, ...
## $ energy                  <dbl> 0.916, 0.815, 0.931, 0.930, 0.833, 0.919, ...
## $ key                     <dbl> 6, 11, 1, 7, 1, 8, 5, 4, 8, 2, 6, 8, 1, 5,...
## $ loudness                <dbl> -2.634, -4.969, -3.432, -3.778, -4.672, -5...
## $ mode                    <dbl> 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, ...
## $ speechiness             <dbl> 0.0583, 0.0373, 0.0742, 0.1020, 0.0359, 0....
## $ acousticness            <dbl> 0.10200, 0.07240, 0.07940, 0.02870, 0.0803...
## $ instrumentalness        <dbl> 0.00e+00, 4.21e-03, 2.33e-05, 9.43e-06, 0....
## $ liveness                <dbl> 0.0653, 0.3570, 0.1100, 0.2040, 0.0833, 0....
## $ valence                 <dbl> 0.518, 0.693, 0.613, 0.277, 0.725, 0.585, ...
## $ tempo                   <dbl> 122.036, 99.972, 124.008, 121.956, 123.976...
## $ duration_ms             <dbl> 194754, 162600, 176616, 169093, 189052, 16...
```

Methodology

First we will test our hypothesis that songs with higher valence are more popular. For this hypothesis we will group songs by valence within the categories of [0, 0.5) and [0.5, 1]. After doing so, we will see if there is a statistical difference in popularity using a two sample hypothesis test. We will use the CLT and simulation methods to compare our results. These tests will produce a p-value and confidence interval which will allow us to quantitatively assess the relationship between valence and song popularity.

Second, we will test our hypothesis that songs with faster tempo, higher danceability, and more energy are more popular. For this hypothesis, we will test to see if the correlation co-efficient between popularity and each variable is significant and non-zero by running correlation tests. If the correlation is non-zero and has a significant p-value, then this would provide evidence to suggest that the variable does affect popularity.

Lastly, we will test our hypothesis that genre does affect a song's popularity. This hypothesis can be tested using a chi-squared test for independence and by analyzing a visual of the distributions of song popularity faceted by genre. By dividing the songs into three logically equal categories based on their popularity score, this categorical variable and the song's genre can be an effective way to indicate independence between popularity and genre. After evaluating whether popularity and genre are independent, a distribution of song popularity faceted by genre and some summary statistics will help further the analysis. These statistical tools can help visualize how specific genres differ in terms of different measures of center and spread for their popularity distributions.

Methodology (ctd.) and Results

Hypothesis 1: Songs with higher valence are more popular.

$H_0 : \mu_{low} = \mu_{high}$

$H_a : \mu_{low} < \mu_{high}$

$\alpha = 0.05$

- Simulation Approach:

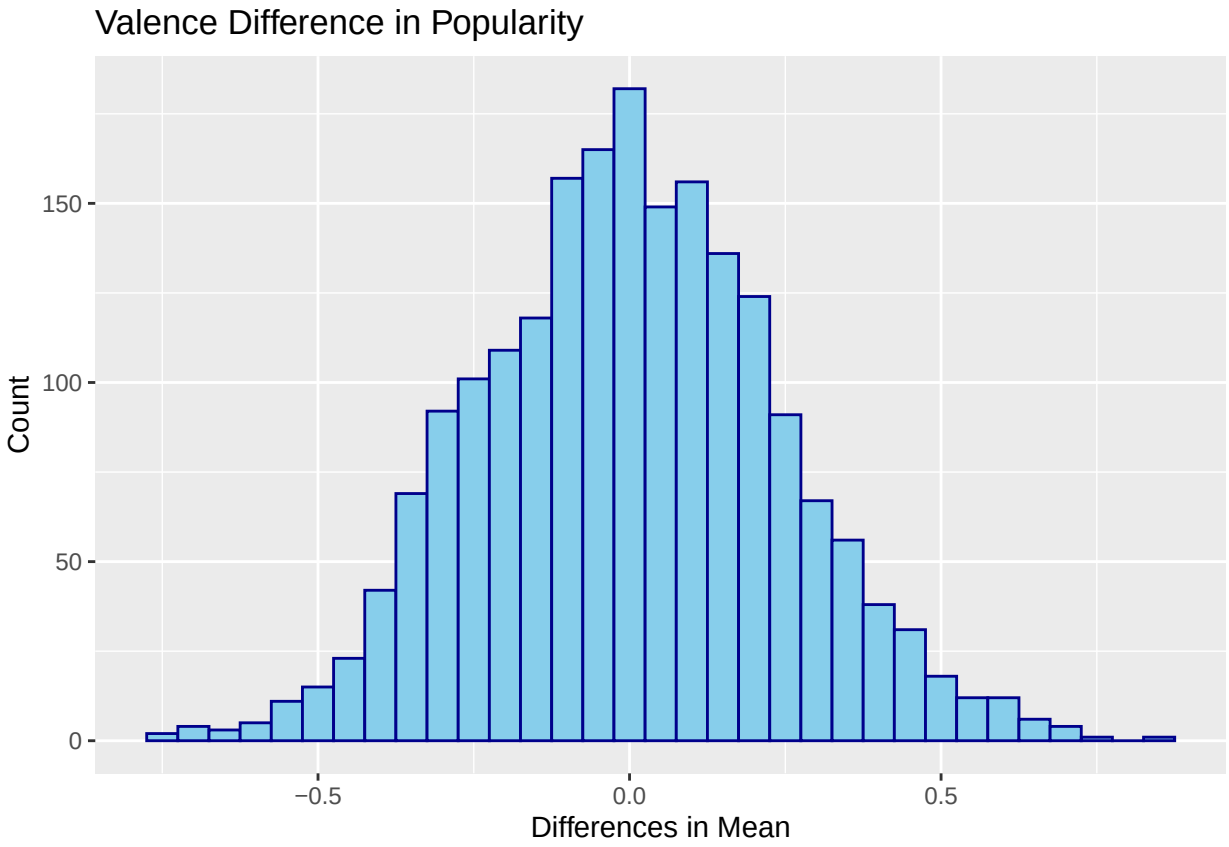
Valence [0,0.5) ~ Low Valence

Valence [0.5,1) ~ High Valence

- Confidence Interval for Difference in Means:

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1  1.22  2.15
```

- Shifted Null Distribution and P-Value:




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Because our p-value is 0, which is less than alpha, we reject the null hypothesis. There is evidence to suggest that the mean popularity of tracks with high valence is greater than the mean popularity of tracks with low valence using a simulation-based approach.

- CLT Approach:

```
##
##  Welch Two Sample t-test
##
## data:  track_popularity by hl_valence
```

```
## t = 6.9592, df = 28248, p-value = 1.748e-12
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.27512     Inf
## sample estimates:
## mean in group High Valence  mean in group Low Valence
##                   49.81231                   48.14251
```

After running a two sample t-test we obtain a test statistic of 6.9592 on a t-distribution with 28248 degrees of freedom, and a p-value of 1.748e-12. Because our p-value is 1.748e-12, which is less than alpha, we reject the null hypothesis. This means that there is enough evidence to suggest that the mean popularity of tracks with high valence is greater than the mean popularity of tracks with low valence using a CLT approach.

From the simulation based approach, we found a 95% confidence interval between 1.22 and 2.15, which does not include 0. This supports our rejection of the null hypothesis. We are 95% confident that the true difference in popularity between tracks with high valence compared to those with low valence lies within the interval (1.22, 2.15).

Hypothesis 2: Songs with faster tempo, higher danceability, and more energy are more popular.

H_o: The correlation coefficient is equal to zero.

H_a: The correlation coefficient is not equal to zero.

$\alpha = 0.05$

- Correlation Tests:

```
##
##  Pearson's product-moment correlation
##
## data:  spotify_songs$tempo and spotify_songs$track_popularity
## t = -0.55221, df = 28365, p-value = 0.5808
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.014915285  0.008358664
## sample estimates:
##          cor
## -0.003278754
```

The correlation test with tempo and track popularity had 28365 degrees of freedom and had a resulting p-value of 0.58. Thus, we fail to reject the null hypothesis. There is not enough evidence to suggest that the correlation coefficient is not equal to zero. The test also has a resulting 95% confidence interval of (-0.015, 0.008). We come to a consistent conclusion that we must fail to reject the null hypothesis because 0 is in the confidence interval. Thus, there is not enough evidence to suggest a correlation between tempo and track popularity.

```
##
##  Pearson's product-moment correlation
##
## data:  spotify_songs$danceability and spotify_songs$track_popularity
## t = 9.662, df = 28365, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04566805 0.06886591
## sample estimates:
```

```
##        cor
## 0.05727471
```

The correlation test with danceability and track popularity had 28365 degrees of freedom and had a resulting p-value of less than 2.2e-16. Because the p-value is so small we reject the null hypothesis. There is evidence to suggest that the correlation coefficient is not equal to zero. The test also has a resulting 95% confidence interval of (0.046, 0.069). This confidence interval supports our findings, and we fail to reject the null hypothesis because 0 is not included in the confidence interval. Thus, there is evidence to suggest a correlation between danceability and track popularity. That being said, the estimated correlation from the sample is 0.057. Therefore, there may be a correlation, but it appears to be positive and very weak.

```
##
##  Pearson's product-moment correlation
##
## data:  spotify_songs$energy and spotify_songs$track_popularity
## t = -13.956, df = 28365, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09412876 -0.07101327
## sample estimates:
##         cor
## -0.08258212
```

The correlation test with energy and track popularity had 28365 degrees of freedom and had a resulting p-value of less than 2.2e-16. Because the p-value is so small we reject the null hypothesis. There is evidence to suggest that the correlation coefficient is not equal to zero. The test also has a resulting 95% confidence interval of (-0.094, -0.071). This confidence interval supports our findings because 0 is not included in the confidence interval. So, we fail to reject the null hypothesis. There is evidence to suggest a correlation between energy and track popularity. That being said, the estimated correlation from the sample is -0.083. Therefore, there may be a correlation, but it appears to be negative and very weak (only slightly stronger than the correlation between danceability and popularity).

Hypothesis 3:

Genre does affect a song's popularity, therefore we expect the spreads of different popularity distributions to be skewed when faceted by genre. For example, we think that genres of Latin and EDM music are less popular causing them to have right-skews. On the other hand, genres like rock and pop will be more normally distributed, or left skewed, because we believe them to be more popular. Thus, we expect genre and popularity to not be independent.

H_o: Track popularity and genre are independent.

H_a: Track popularity and genre are NOT independent.

$\alpha = 0.05$

- Chi-Squared Test:

Note: Track Popularity (0,33] ~ Not Popular,

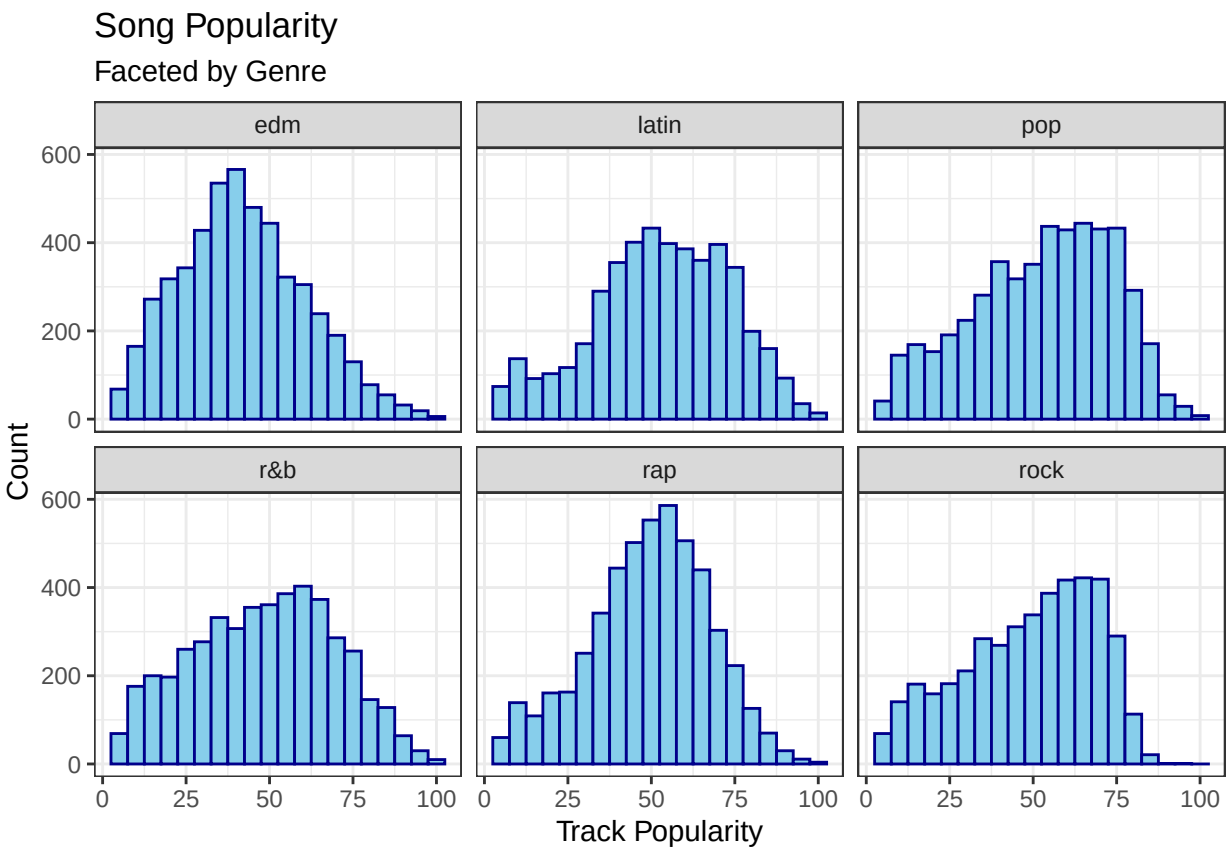
```
Track Popularity (33,67] ~ Average Popularity

Track Popularity (67,100] ~ Very Popular
```

```
##
##                       edm latin  pop  r&b  rap rock
##   Average Popularity 2800  2569 2555 2468 3322 2358
```

```
##   Not Popular        1685   748  985 1228  934 1013
##   Very Popular        510  1241 1419  920  767  845
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(spotify_songs$popularity, spotify_songs$playlist_genre)
## X-squared = 1130, df = 10, p-value < 2.2e-16
```



- Summary Statistics:

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 5
##   playlist_genre mean_pop median_pop sd_pop iqr_pop
##   <chr>             <dbl>      <dbl>  <dbl>   <dbl>
## 1 edm                42.0         41   18.8    25
## 2 latin              53.0         54   20.5    29
## 3 pop                52.9         55   20.9    31.5
## 4 r&b                48.3         50   21.3    33
## 5 rap                49.3         51   18.1    24
## 6 rock               48.8         52   19.6    31
```

The faceted histograms, accompanied with some summary statistics and a chi-squared test, provide strong evidence for the relationship between genre and popularity of a given song. The chi-squared test for independence provides a p-value of less than 2.2e-16, which is less than alpha of 0.05, which leads us to reject the null hypothesis. Therefore, we have evidence to suggest that genre and popularity of a song are not independent.

It is clear that all distributions resemble a normal or slightly skewed unimodal model. The popularity of rap

and latin songs behave in a normal manner, while the popularity of rock and pop songs are left skewed, and the popularity of edm is right skewed. The distribution for the popularity of r&b songs is less clear, but is unimodal centered around 60 with a fat left tail and slim right tail.

Utilizing the measures for center and spread of this data, these summary statistics indicate that latin and pop music, on average, are the most popular genres given their measures of centers, while edm is clearly the least popular. Rock songs are average in the popularity rankings but also have the largest spreads. Based on all of these factors, we have evidence to suggest that genre does impact a song's popularity, with latin and pop songs being, on average, the most popular and EDM trailing far behind. The rap, rock, and R&B songs, on average, fall somewhere in between this range of latin/pop and EDM. Our hypothesis was partly incorrect, primarily about the underrated appreciation for latin music, but we also expected more of a normal distribution for mainstream genres like rock and pop.

Discussion

This data analysis of Spotify songs provides some clear information on the criteria that contribute to the popularity of a track. From hypothesis 1 we discovered using a simulation and CLT based approach that there is evidence to suggest songs with higher valence are more popular. The significantly small p-values from these hypothesis tests indicated that songs with higher valence had a larger mean popularity score compared to those with lower valence. For hypothesis 2, we originally planned on performing linear regression to assess the relationship between tempo, danceability, energy and their effect on a song's popularity. The results of this linear regression can be found below in Appendix A, but these results were not used to analyze the hypothesis because the conditions for linear regression were not met. Therefore, we utilized a correlation test to test the null hypothesis that each correlation between tempo, energy, and danceability with polularity, was zero. Through these three tests, we discovered that there is evidence to believe that only energy and danceability, not tempo, have a nonzero correlation with popularity. It is important not to let this result be misleading, as just because there is evidence that the coefficient is not zero, they are still found to be weak correlations. Hypothesis 3 utilized a chi-squared test and analysis of faceted histograms to discover the relationship between genre and popularity. The chi-squared test provided a significantly small p-value, which provides reason to believe that genre and popularity are not independent. To further understand this relationship, the histograms and summary statistics highlighted the significant popularity of latin and pop music, while also indicating the general shapes of the popularity distributions for each genre.

Overall, this data analysis provided tons of information about the contributions, if any, that genre, tempo, valence, energy, and danceability make to a song's popularity. When noting the discrepancies of our analysis, we utilized many forms of hypothesis testing using the CLT and simulation based approaches. More visuals and perhaps a linear regression would have been nice for interpretation instead of so much quantitative analysis. It is also important to note that we removed all songs with a popularity rating below 5. When performing the statistical tests, this high concentration of songs with little to no popularity was initially skewing our results and led to errors in interpretation. Therefore, we chose to remove these songs that are not popular in order to analyze the songs people are listening to consistently. We also could not find out how these popularity scores were measured from the data collectors besides the fact that they were on a 0-100 point scale. If we were to know the methods in which the data for the popularity scores was collected, we could have provided more insight into interpretation of the hypotheses. The results and statistical tests in this report seem to be reliable and valid given the validity of the data. With nearly 28,000 songs in the dataset, this provides a sufficiently large sample size to draw conclusions based on the variables we chose. As mentioned before, we relied heavily on quantitative statistical analysis and a larger visual or qualitative portion could have been helpful for interpretation.

If we were to start over with this project, I think we would have focused on more factros than just quantitative measures to compare with popularity. This dataset includes tons of information that includes but is not limited to the artist, album name, release date, and many more factors that can influence popularity. Although we did not discount the possibility of these variables to influence the popularity of a song, it is possible that artist and other aspects of a song have a much larger contribution to the popularity score than the tempo or danceability of a song. It is common to see a mainstream artist that produces a subpar song from an

analytical standpoint but it does well in the top charts simply because of the name recognition of the artist. It would have been interesting to see the most popular artists, albums, and other contributing factors besides the numerical measures of tempo, valence, energy, and danceability. If nothing else, including variables like the artist and album in our analysis would make it more comprehensible for the average reader who may not understand the more technical musical quantitative measures.

Appendix A

For hypothesis 2, we performed a linear regression in order to determine if there is a strong correlation between popularity and the other variables, tempo, danceability, and energy. Based on our hypothesis we would hope to see a positive, linear correlation.

H_o: There is no correlation between tempo and popularity. ($slope = 0$)

H_a: There is a positive correlation between tempo and popularity. ($slope \neq 0$)

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   48.7      0.946       51.5  0.
## 2 tempo          0.0137   0.00453      3.03 2.45e- 3
## 3 energy        -9.05     0.670      -13.5  2.09e-41
## 4 danceability   7.50     0.839        8.93 4.36e-19
```

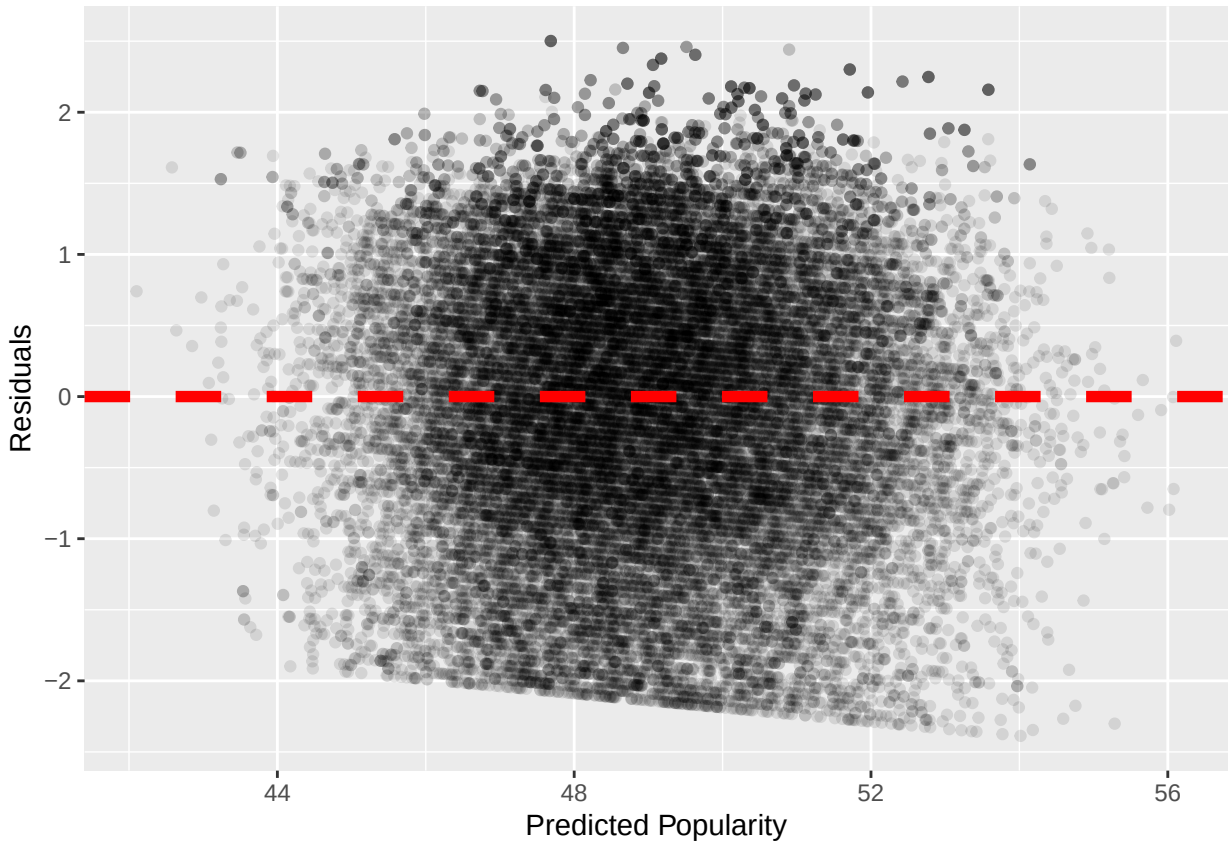
```
## # A tibble: 4 x 7
##   term         estimate std.error statistic  p.value  conf.low conf.high
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
## 1 (Intercept)   48.7      0.946       51.5  0.        46.8       50.5
## 2 tempo          0.0137   0.00453      3.03 2.45e- 3   0.00485    0.0226
## 3 energy        -9.05     0.670      -13.5  2.09e-41 -10.4       -7.73
## 4 danceability   7.50     0.839        8.93 4.36e-19   5.85       9.14
```

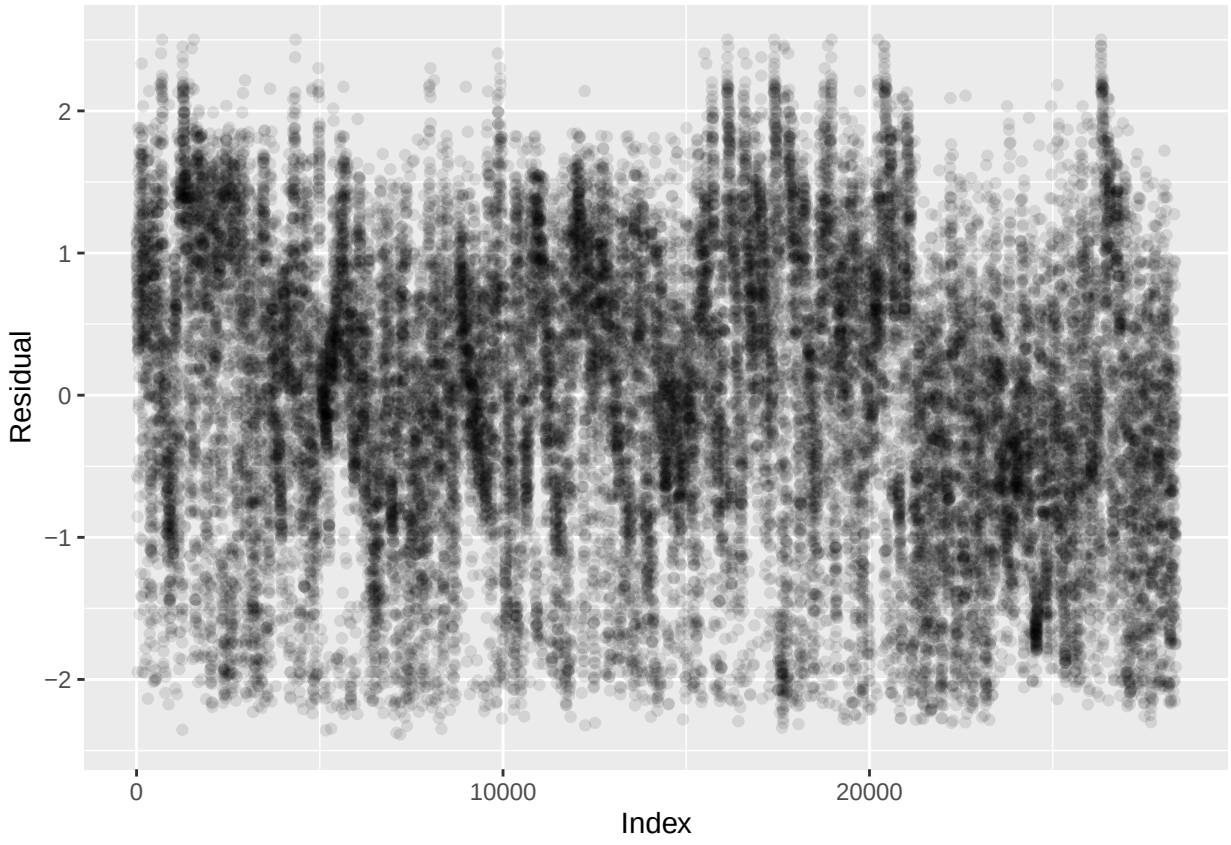
```
## [1] 0.009692233
```

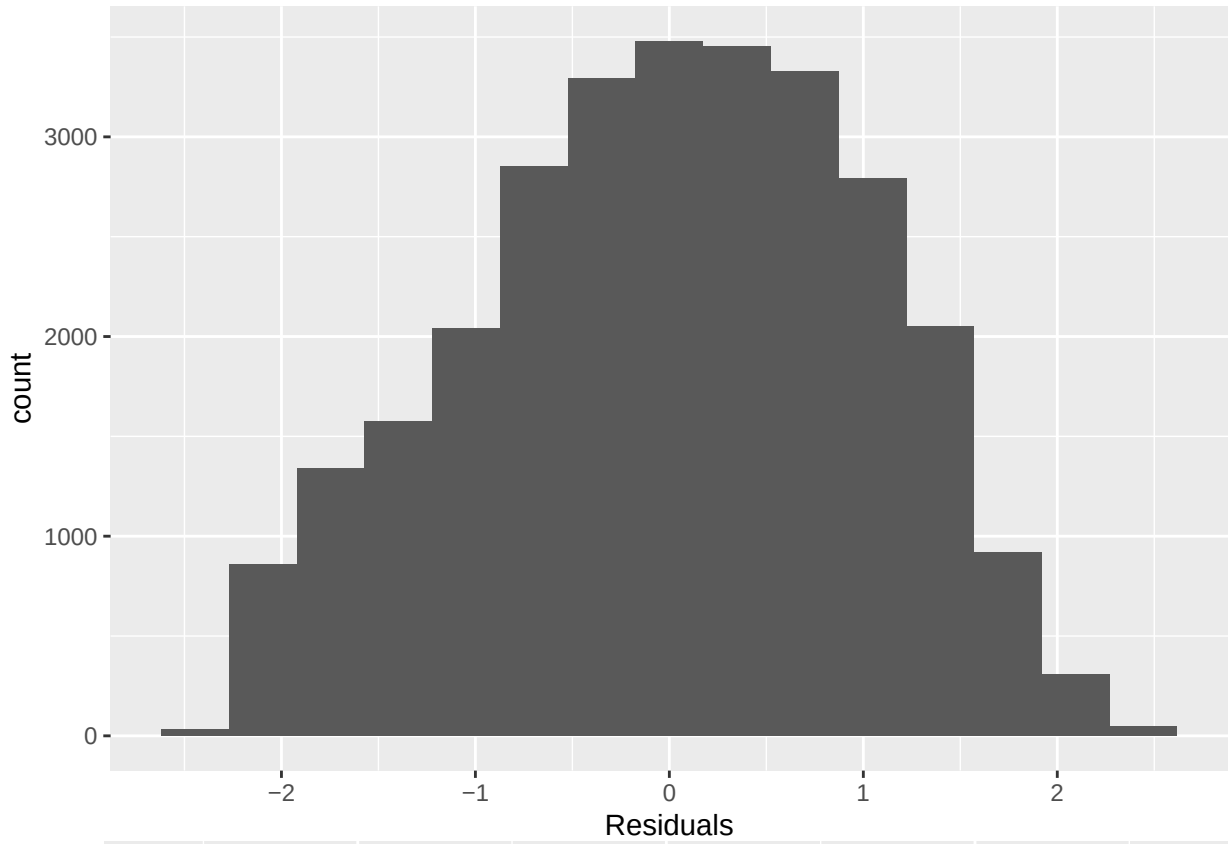
From the linear model we get that :

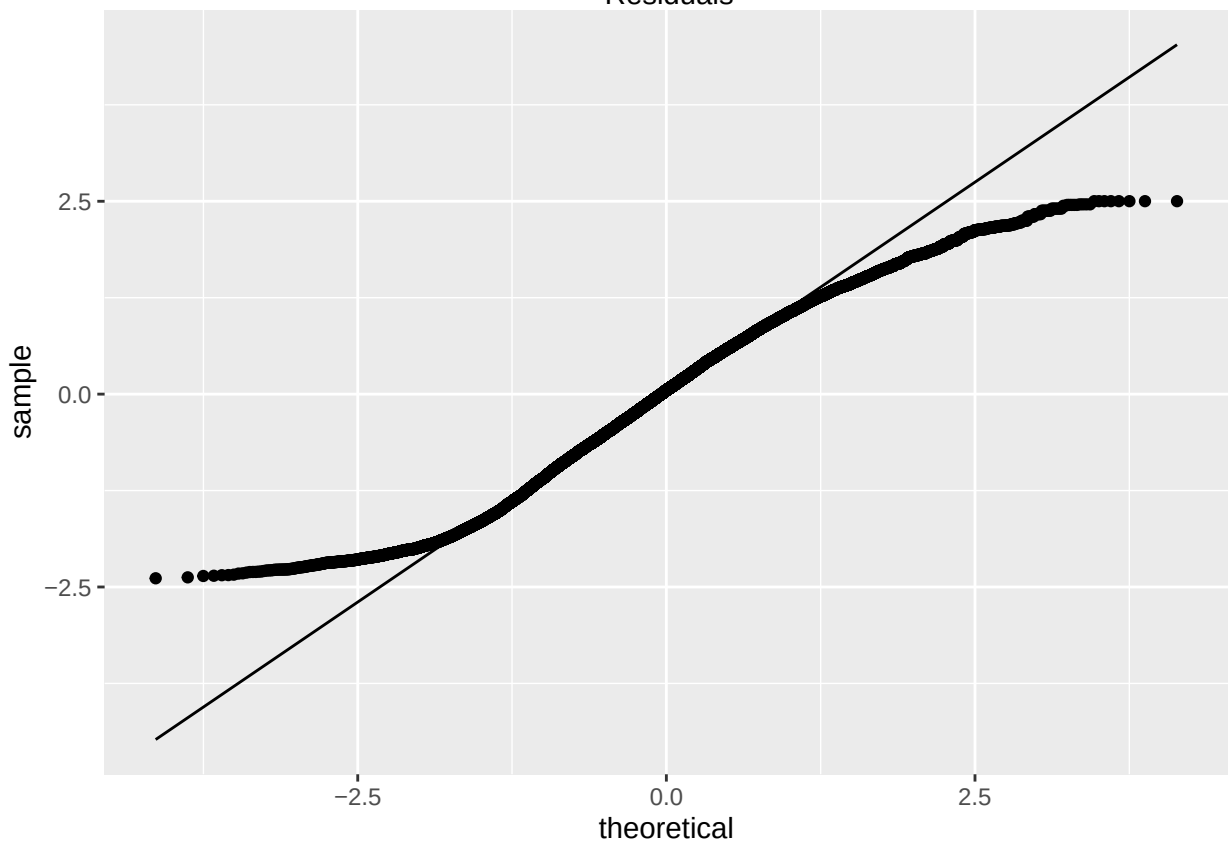
$$\widehat{popularity} = 48.7 + 0.01 \times tempo + 7.5 \times danceability - 9.04 \times energy$$

. We must check the conditions for linear regression to use this model.


Residuals
2
1
0
−1
−2
44
48
52
56
Predicted Popularity


Residual
2
1
0
−1
−2
0
10000
20000
Index





Linearity There seems to be an uneven spread above and below 0, so we say that this may not pass the

linearity test.

Independence There does not seem to be any patterns in the plot, so we would say it passes the test for independence.

Normality We think it may not be passing the normalcy test. The histogram of the residuals looks as if it could be bi-modal and/or right skewed. As for the Q-Q plot, it does not appear to follow the trend of the line. This was true even when we tried to run it using a logarithmic fit.

Equal Variance There appears to be constant variance among the residuals, so this data does have equal variance.

After analyzing the results of these condition for linear regression, it is hard to say if there is any correlation, because a linear model may not be adequate to test this relationship as it is questionable if it is passing some of the conditions.