# MA678 Project

*Xinqiao Zhang*

*11/30/2019*

## Introduction

### Background

Nowadays, we listen to music when we are working out, behind the wheel or partying. As one of the largest music streaming company in the world, Spotify has taken part not only in streaming but also in helping publish music for artists.
Consequently, Spotify recorded its own streaming and created a charts of all the streaming data via Spotify.

Proj/Music for everyone.bb

### Task

My personal interest is that, what it would take for a song to be a hit. That is, what kind of audio features would affect the streaming of a song.

## Data

### Charts

On the website https://spotifycharts.com/regional, Spotify shared their data for the top streamed songs in a certain peroid of time(week or day), in a certain market(global, US, UK, etc) and on a scale of top 200 songs or top 50.
I took the top 200 charts from the US market in the latest 20 weeks and take the average of the stream of each track by week. It is possible for some tracks to lose their position in the top 200 and some other tracks to raise to such position. Hence it would be possible to average out the streaming instead of adding up.

This step would returned us only 678 tracks which means there were 678 different songs been in Spotify top 200 chart in the US market in the past 20 weeks.

### Audio Features

From the Spotify Charts website, we could only acquire the streams of track. Now we would move to Spotify API to get more features of the tracks.

Then we can use the Spotify Developer platform to exctract the data for audio features of the tracks in the list before.

## Data Cleaning

**There we have the audio features of those tracks and streams as well. Hence we could join the two tables together.**

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

We noticed that there is only one entry in the column "type", so we will exclude that column from the data.
Also, the columns "uri", "track_href", "analysis_url" and "URL" would be excessive in our analysis, so we would exclude those columns as well.
Finally, the column "id" and ""track.id" are duplicates from the precedure of merging two dataset together. Hence we should delete one of them.

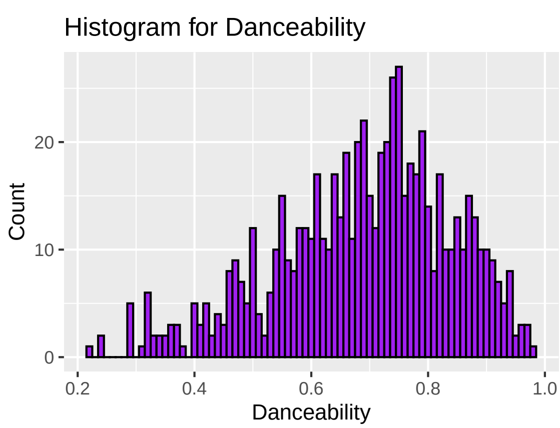
### Variable Explanation

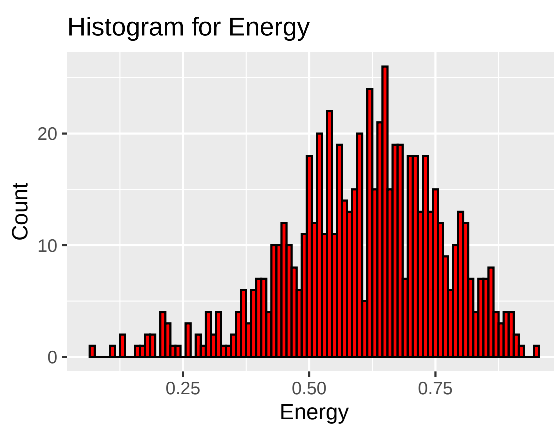
For the data we have obtained, there are 1 outcome variable, 13 predictors and 3 identification features.

| Variables | Category | Explanation |
|---|---|---|
| danceability | indicator | Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. |
| energy | indicator | Energy is a measure from 0.0 to 1.0 and 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. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. |
| key | indicator | The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C /D , 2 = D, and so on. If no key was detected, the value is -1. |
| loudness | indicator | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. |
| mode | indicator | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| speechiness | indicator | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |

| Variables | Category | Explanation |
| --- | --- | --- |
| acousticness | indicator | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| instrumentalness | indicator | Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| liveness | indicator | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| valence | indicator | A measure from 0.0 to 1.0 describing 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). |
| tempo | indicator | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| id | identification | The Spotify ID for the track. |
| duration_ms | indicator | The duration of the track in milliseconds. |
| time_signature | indicator | An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). |
| weekly.stream | outcome | The mean of streams of a track for the weeks that the track has made to the top 200 charts in US Spotify market on a weekly basis. |
| Track Name | identification | The name of the track. |
| Artist | identification | The name of the artist. |

## Exploratory Data Analysis



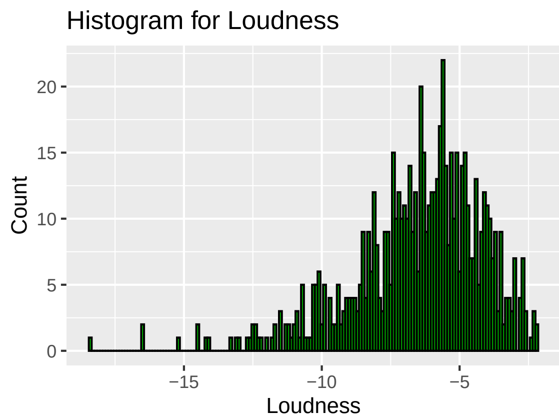



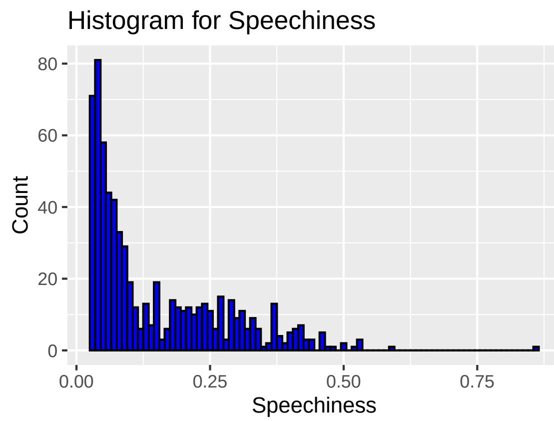

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.218   0.596   0.708   0.690   0.795   0.978
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0694  0.5150  0.6230  0.6075  0.7170  0.9490
```

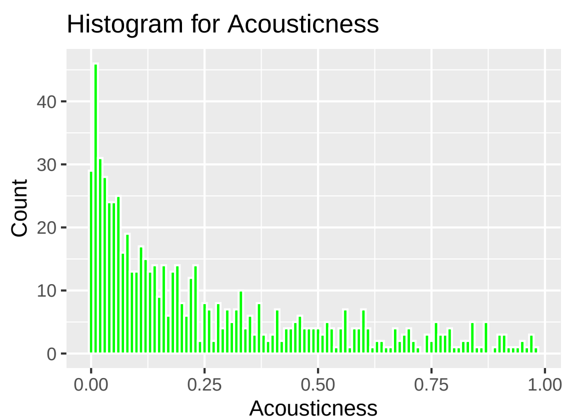
From the histogram of the frequency of danceability, we can see that the data is roughly normally

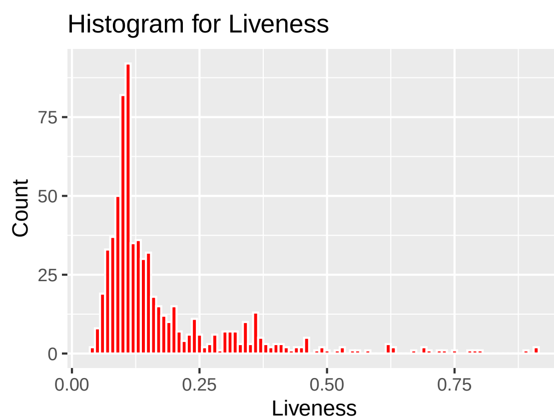
distributed from 0.2 to 1.0. The mean of the variable is 0.690.
From the histogram of the frequency of energy, we can see that the data is roughly normally distributed from 0.0 to 1.0. The mean of the variable is 0.6075.



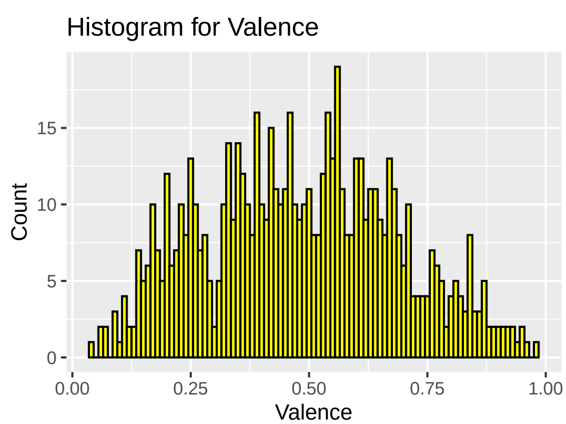



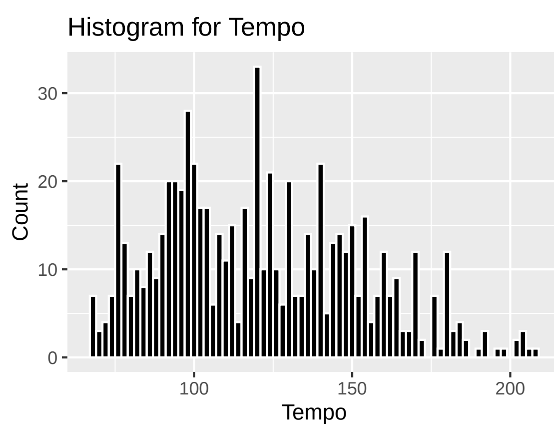

From the histogram of loudness, it is hard to specify the distribution of this variable because it was scattered out in the range but more densed around -8.0 to -3. From the histogram of speechiness, we could observe a heavy tail on the right. This is reasonable because most of the songs would not include a heavy portion of speech. Those with relatively higher speechiness would be more likely to be rap music. The distribution is not normal.




From the histogram of acousticness, we could also observe a heavy tail on the right. This is also true because the beats of the most of the tracks were now produced digitally rather than acousticly. Higher acousticness indicates heavier portion of acoustic instruments. The distribution is laso not normal.
Histogram of livenes also indicates overall low presence of audience. This graph is also skewed and the distribution is not normal.

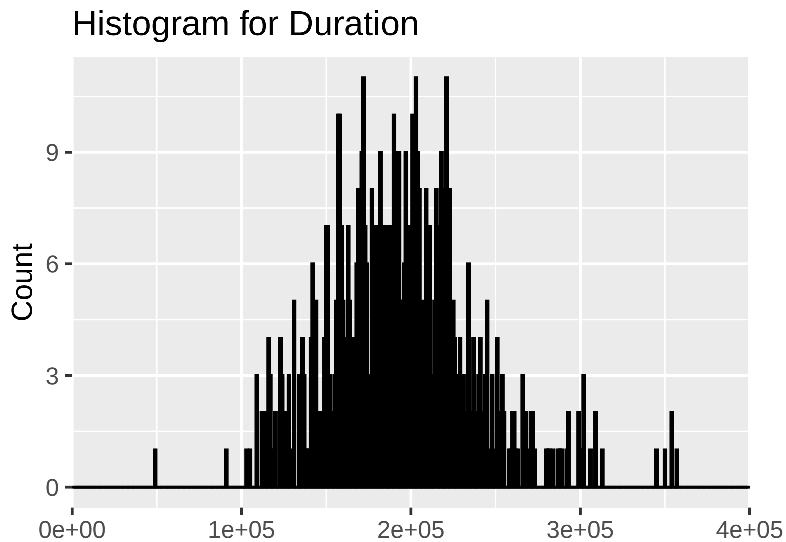


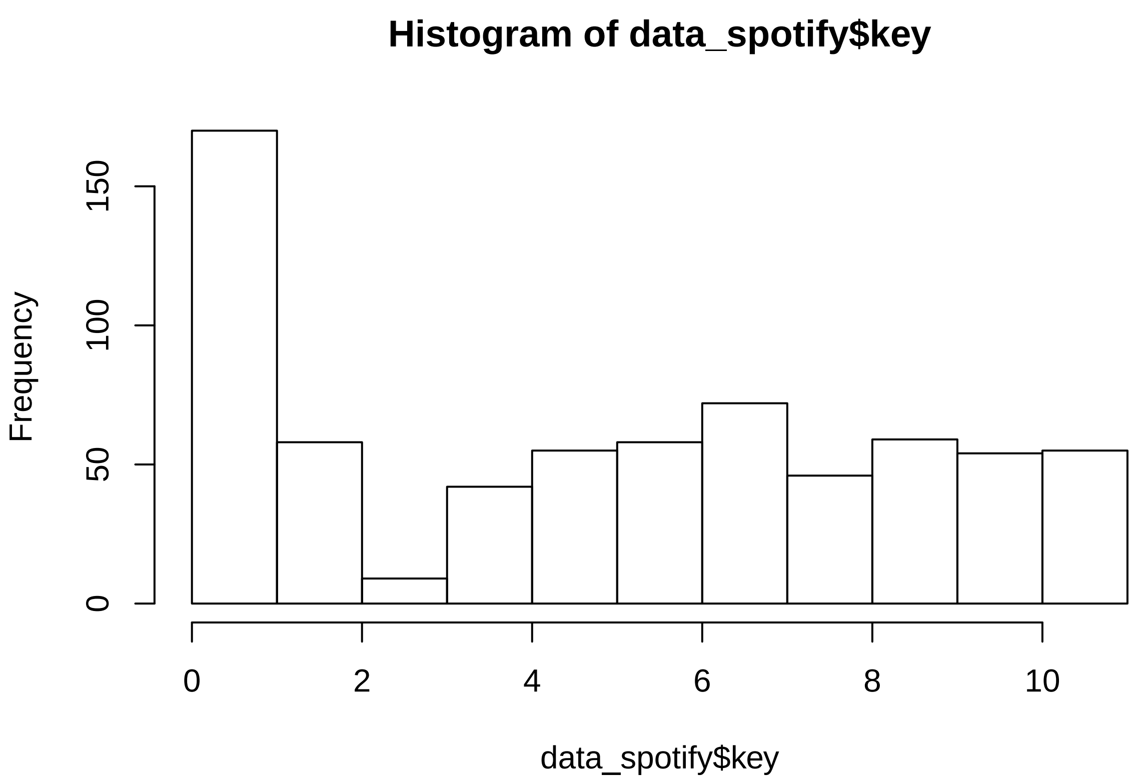
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0361  0.3330  0.4795  0.4832  0.6290  0.9820
```

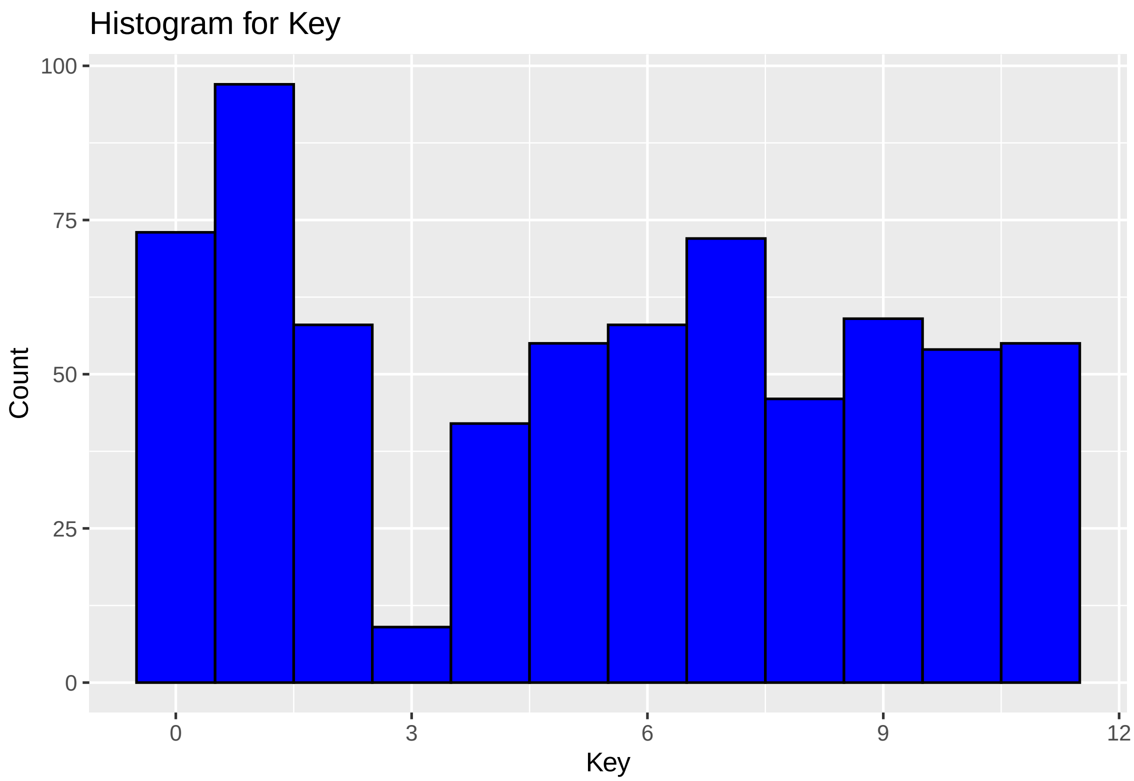
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   67.00   95.98  119.91  121.11  144.58  207.48
```

From the histogram of the frequency of valence, we can see that the data is roughly normally distributed from 0.3 to 1.0. The mean of the variable is 0.4832.
From the histogram of the frequency of tempo, it is harder to say that the data is normally distributed because the range is from 67 to 208.




The variable duration was roughly normally distributed in the range shown in the graph. However, it has a lighter tail to the right, which indicates the rare long tracks.






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.250   6.000   5.279   8.000  11.000
```

From the graph we can see that the keys are distributed thoughout the octave and notive that when key=3(D /E ), there is a drop in the graph indicating this key is least used in the top 200 charts.

## Models Fitting

### Model 1

The first model used is a basic linear moedl to obtain a basic idea on the reaction of the outcome to the variables we have.

```
##
## Call:
## lm(formula = log(weekly.stream) ~ danceability + energy + key +
##     loudness + mode + speechiness + acousticness + instrumentalness +
##     liveness + valence + tempo + duration_ms, data = data_spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6928 -0.3234 -0.1213  0.1982  1.9783
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.601e+01  2.449e-01  65.347  < 2e-16 ***
## danceability     -3.649e-01  1.366e-01  -2.672 0.007729 **
## energy           -4.818e-01  1.940e-01  -2.484 0.013251 *
## key              -9.610e-03  4.897e-03  -1.963 0.050111 .
## loudness          4.327e-02  1.141e-02   3.793 0.000162 ***
## mode             -1.013e-01  3.649e-02  -2.776 0.005665 **
```

```
## speechiness      -8.160e-02  1.465e-01  -0.557 0.577707
## acousticness     -1.376e-01  8.804e-02  -1.562 0.118668
## instrumentalness  6.541e-01  2.908e-01   2.249 0.024849 *
## liveness         -1.861e-01  1.292e-01  -1.441 0.150140
## valence          -3.198e-02  9.593e-02  -0.333 0.738984
## tempo            -1.427e-04  5.753e-04  -0.248 0.804251
## duration_ms      -1.367e-06  3.473e-07  -3.937 9.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4478 on 665 degrees of freedom
## Multiple R-squared:  0.07131,    Adjusted R-squared:  0.05455
## F-statistic: 4.255 on 12 and 665 DF,  p-value: 1.665e-06
```

From the basi linear regression model, we can see that loudness and duration have the lowest p-value, indicating strong evidence of that those two variables should present in the model. However, the coefficient of duration is relatively small, even if we take second as unit instead of millisecond, the coefficient would only change from -1.367e-06 to -1.367e-03. This is reasonable because the popularity of a track should not be affected by the length of it too much.
Also, it was surprising that the variable mode has a negative coefficient which means that a track with key on major mode would be less streamed comparing to one with key on minor mode, ceteris paribus.

```
##
## Call:
## lm(formula = log(weekly.stream) ~ danceability + energy + loudness +
##     mode + instrumentalness + duration_ms, data = data_spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6566 -0.3241 -0.1186  0.1979  2.0328
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.580e+01  2.045e-01  77.256  < 2e-16 ***
## danceability     -3.044e-01  1.212e-01  -2.512 0.012228 *
## energy           -4.525e-01  1.660e-01  -2.726 0.006576 **
## loudness          4.815e-02  1.112e-02   4.330 1.72e-05 ***
## mode             -8.233e-02  3.515e-02  -2.342 0.019452 *
## instrumentalness  6.994e-01  2.893e-01   2.417 0.015899 *
## duration_ms      -1.331e-06  3.450e-07  -3.857 0.000126 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4489 on 671 degrees of freedom
## Multiple R-squared:  0.05827,    Adjusted R-squared:  0.04985
## F-statistic: 6.919 on 6 and 671 DF,  p-value: 3.797e-07
```

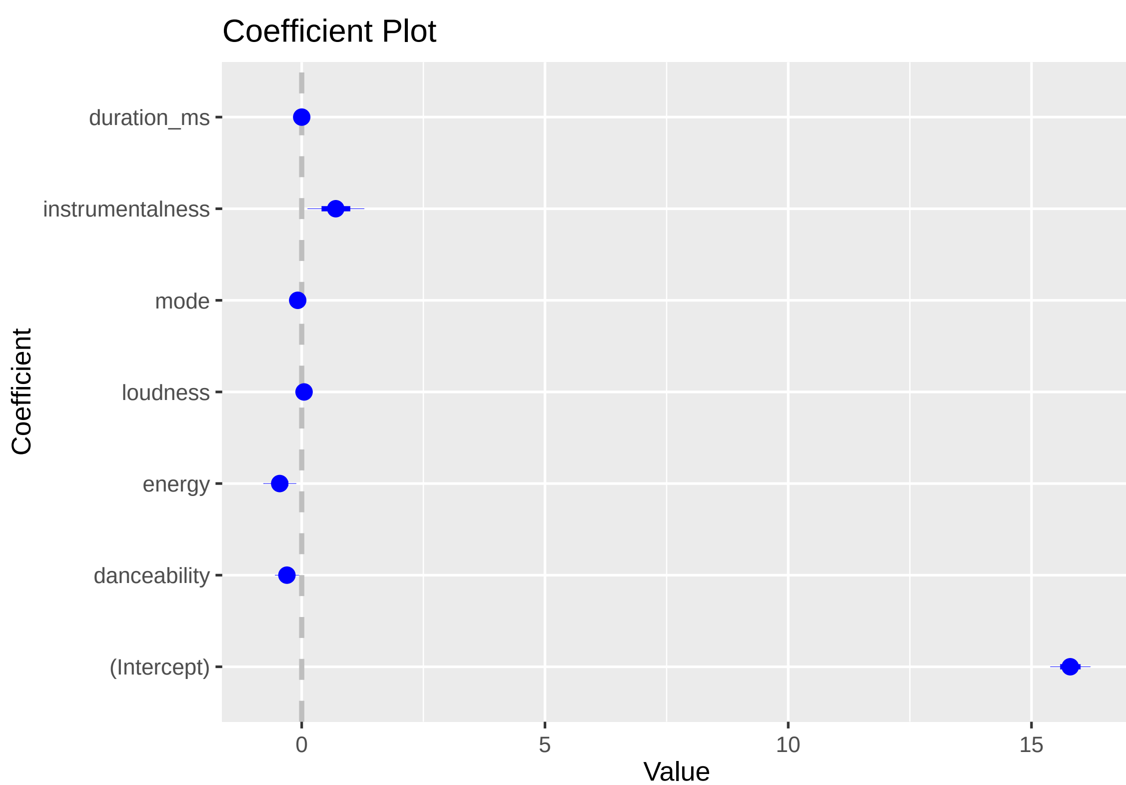
With the adjusted model, we can see that all variables have evidences in different significant level to affect the streams.

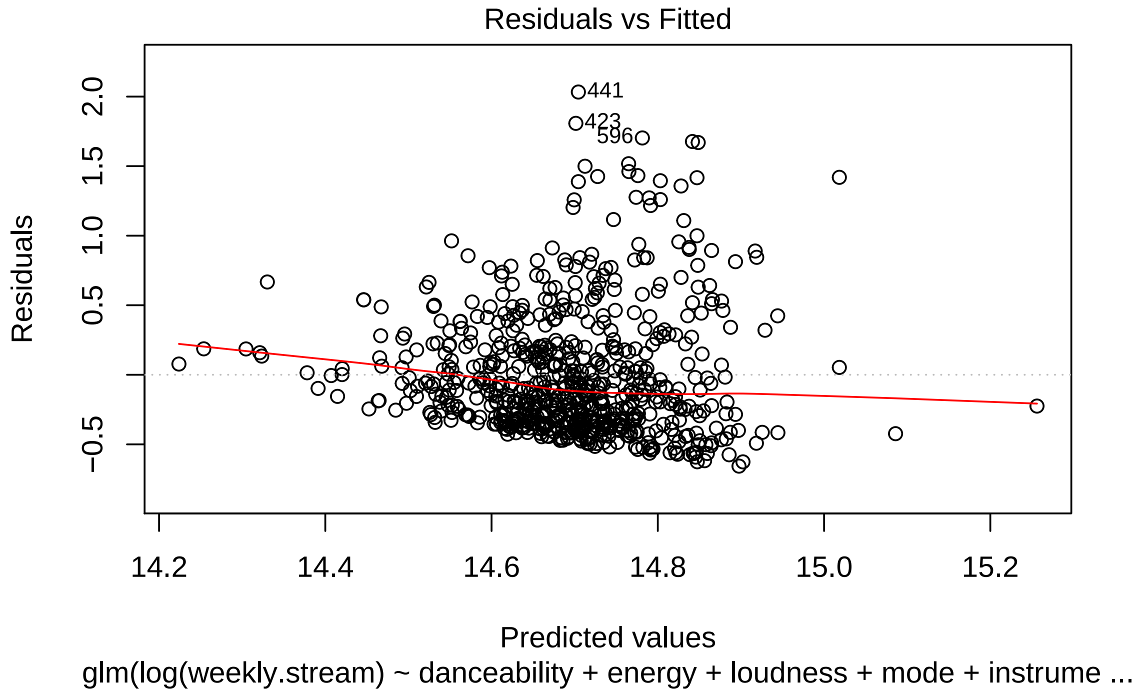
### Model 2

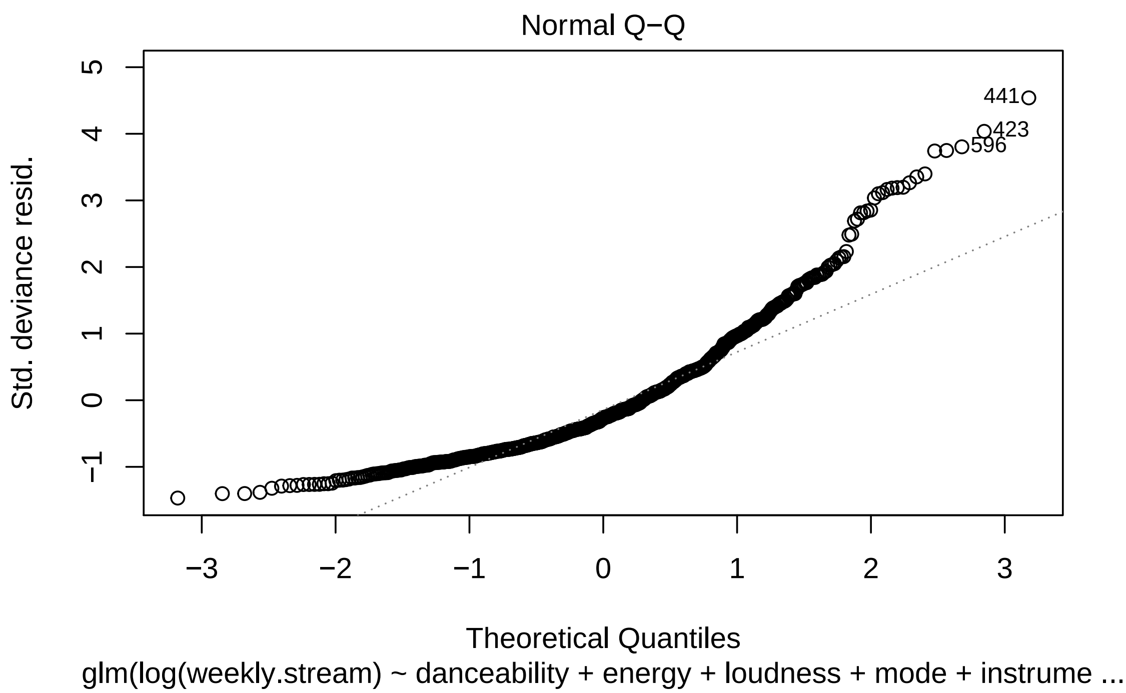
The second model I used is the generalized linear regression model.

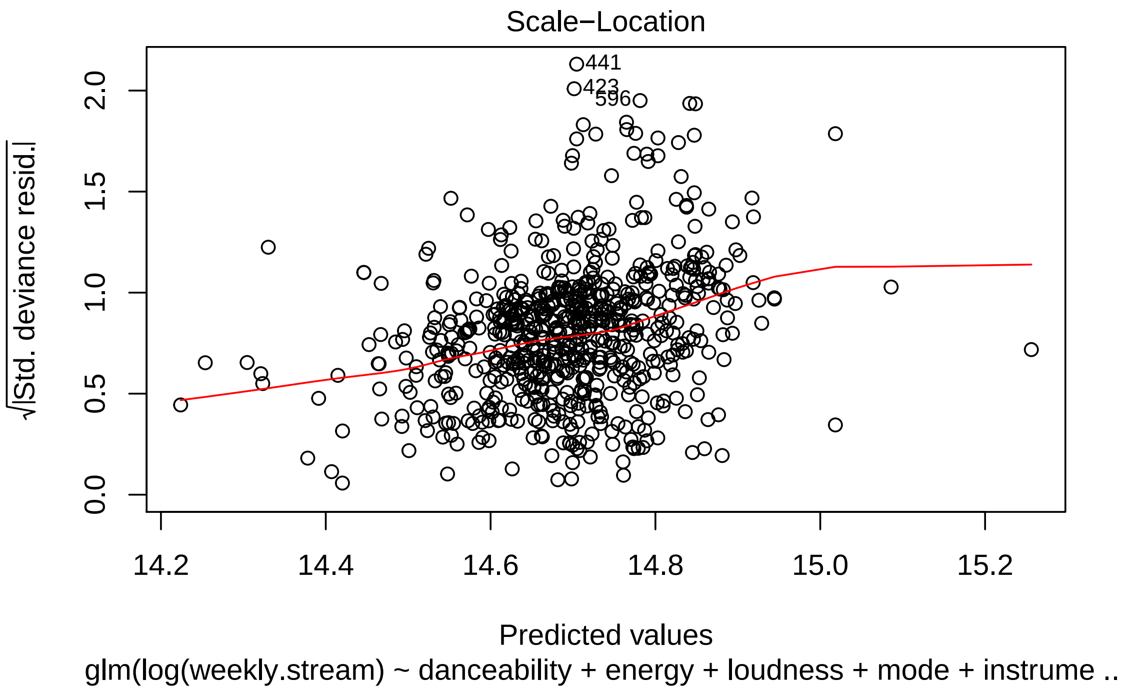
```
##
## Call:
```

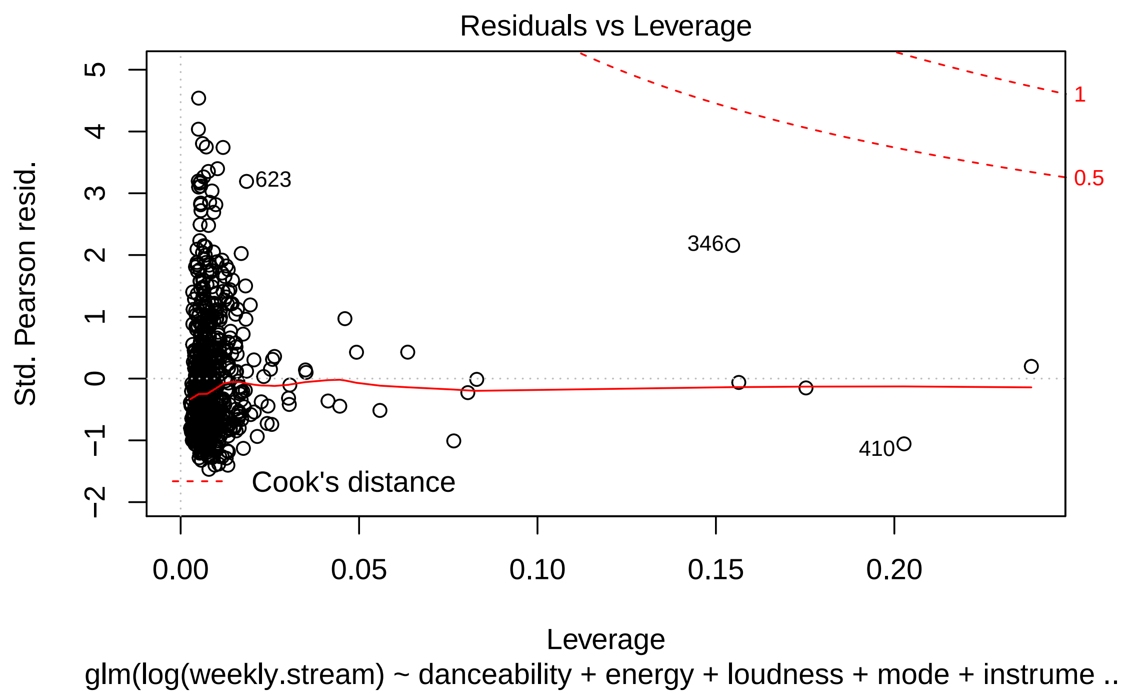
```
## glm(formula = log(weekly.stream) ~ danceability + energy + loudness +
##     mode + instrumentalness + duration_ms, family = "gaussian",
##     data = data_spotify)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6566  -0.3241  -0.1186   0.1979   2.0328
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.580e+01  2.045e-01  77.256  < 2e-16 ***
## danceability    -3.044e-01  1.212e-01  -2.512 0.012228 *
## energy          -4.525e-01  1.660e-01  -2.726 0.006576 **
## loudness         4.815e-02  1.112e-02   4.330 1.72e-05 ***
## mode            -8.233e-02  3.515e-02  -2.342 0.019452 *
## instrumentalness  6.994e-01  2.893e-01   2.417 0.015899 *
## duration_ms     -1.331e-06  3.450e-07  -3.857 0.000126 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2015298)
##
##     Null deviance: 143.59  on 677  degrees of freedom
## Residual deviance: 135.23  on 671  degrees of freedom
## AIC: 847.01
##
## Number of Fisher Scoring iterations: 2
```




Residuals vs Fitted
Residuals
−0.5
0.5
1.0
1.5
2.0
441
423
596
14.2
14.4
14.6
14.8
15.0
15.2
Predicted values
glm(log(weekly.stream) ~ danceability + energy + loudness + mode + instrume ...


Normal Q−Q
Std. deviance resid.
−1
0
1
2
3
4
5
441
423
596
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(log(weekly.stream) ~ danceability + energy + loudness + mode + instrume ...



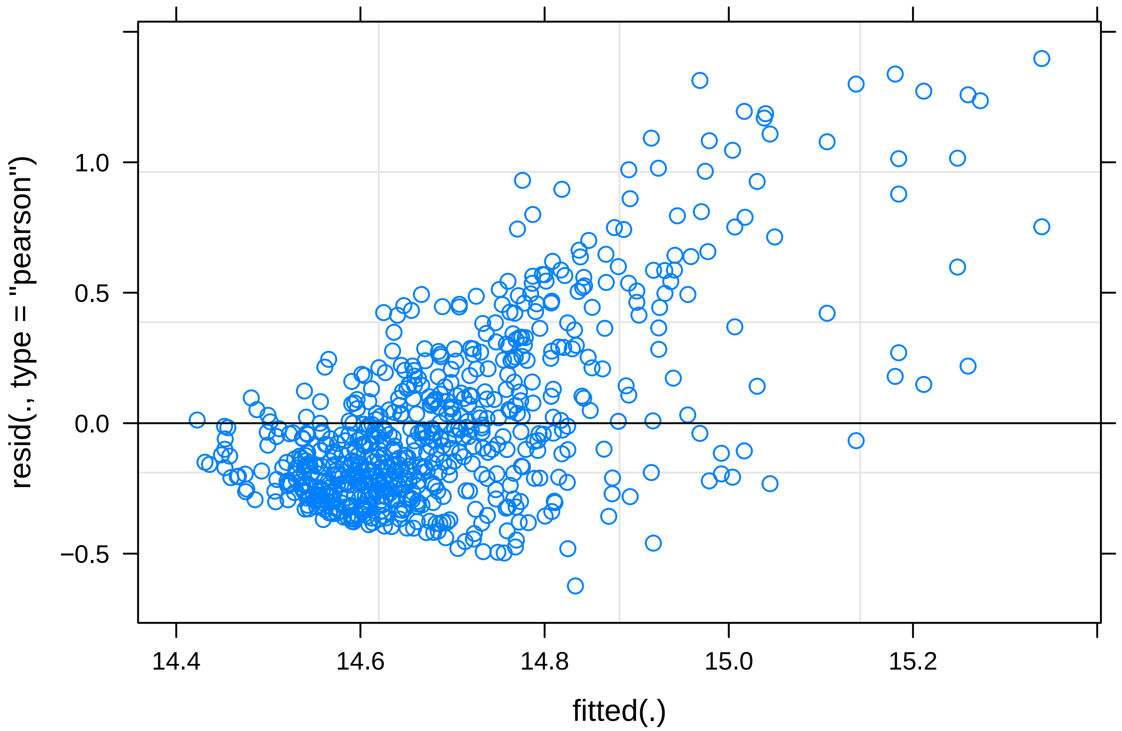

In this model, we have a better view of the coefficients of the variables and as I mentioned before, the coefficient for duration_ms is small but still significant.
From this plots, we can tell that the residual plot is in good shape.

**Model 3**

Now we would fit a multi-level model on loudness, which was significant from the model we had in glm.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log(weekly.stream) ~ (1 | danceability) + (1 | instrumentalness) +
```

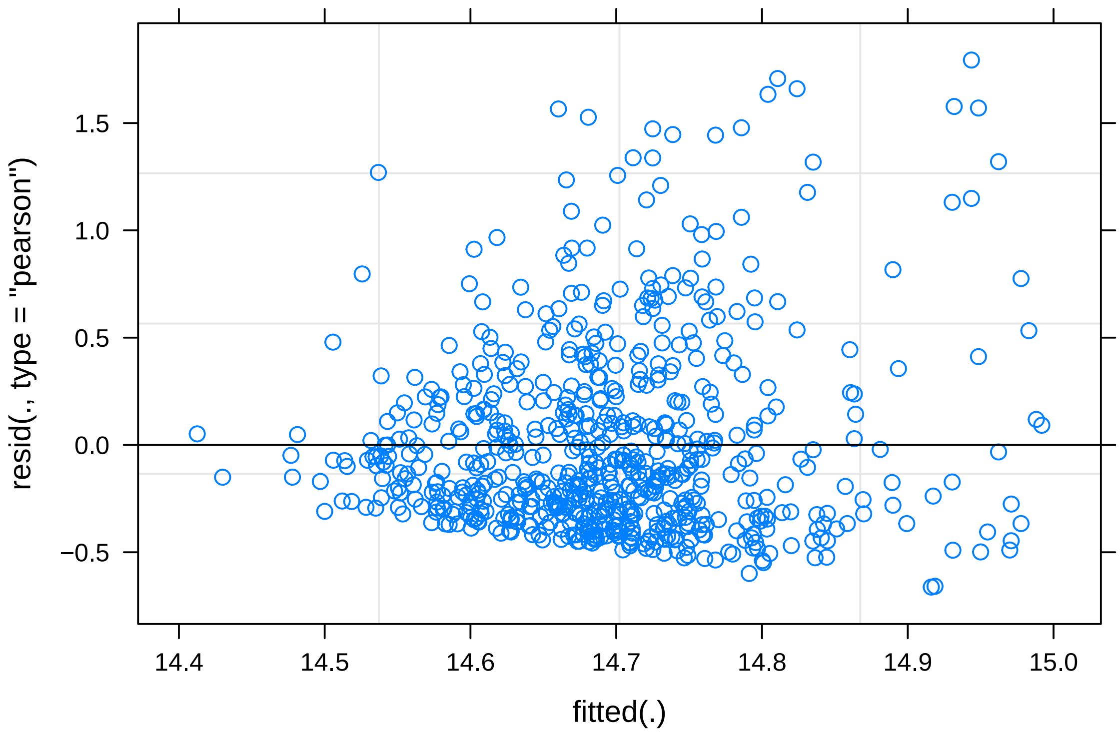
```
##     loudness + (1 | energy)
##    Data: data_spotify
##
## REML criterion at convergence: 857.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5793 -0.6040 -0.2533  0.3572  3.5380
##
## Random effects:
##  Groups           Name        Variance  Std.Dev.
##  danceability     (Intercept) 0.0290067 0.17031
##  energy           (Intercept) 0.0233483 0.15280
##  instrumentalness (Intercept) 0.0002663 0.01632
##  Residual                     0.1560024 0.39497
## Number of obs: 678, groups:
## danceability, 375; energy, 374; instrumentalness, 204
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 14.809757   0.054623 271.128
## loudness     0.019394   0.007438   2.607
##
## Correlation of Fixed Effects:
##          (Intr)
## loudness 0.911
```




The multilevel model on loudness gave us a good result on the variable itself.
Now we add the effect of key into the model. Key was a parameter insignificant in the models before. However, it could group the data into 12 different tunes and gave us a clearer idea on the model.

```
## boundary (singular) fit: see ?isSingular
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log(weekly.stream) ~ (danceability | key) + (1 | instrumentalness) +
##     loudness + (energy | key)
##    Data: data_spotify
##
## REML criterion at convergence: 864.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4818 -0.7144 -0.2985  0.4354  4.0127
##
## Random effects:
##  Groups           Name         Variance Std.Dev. Corr
##  instrumentalness (Intercept)  0.001094 0.03308
##  key              (Intercept)  0.000000 0.00000
##                   energy       0.019146 0.13837   NaN
##  key.1            (Intercept)  0.052710 0.22959
##                   danceability 0.119744 0.34604  -1.00
##  Residual                      0.199790 0.44698
## Number of obs: 678, groups:  instrumentalness, 204; key, 12
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 14.852000   0.067662 219.503
## loudness     0.023615   0.007682   3.074
##
## Correlation of Fixed Effects:
##          (Intr)
## loudness 0.869
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```




From my point of view, key would affect the daceability and energy. After adding this to the model, the residual plot looks more densed.

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: data_spotify
## Models:
## fit3_1: log(weekly.stream) ~ (1 | danceability) + (1 | instrumentalness) +
## fit3_1:     loudness + (1 | energy)
## fit3_2: log(weekly.stream) ~ (danceability | key) + (1 | instrumentalness) +
## fit3_2:     loudness + (energy | key)
##        Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## fit3_1  6 855.52 882.64 -421.76   843.52
## fit3_2 10 871.73 916.92 -425.87   851.73     0      4          1
```

From the ANOVA of the two multilevel models, we can see that the first model has lower AIC and BIC values. Therefore the effect of keys should not be added to the model and thus *fit3_1* is a better fit for the data.

## Discussion

### Implication

The result from the models were not excatly what we expected from the beginning. It was surprising that loudness of a track could affect the popularity that much. The result from the basic linear regression was intuitive but might not be accurate. However, the results were satisfying in terms of anova and residual plots.
### Limitation
I think the limitation of the data is stopping the models from being outstanding. I could only acquire at most top 200 from the spotify charts website and the audio features were subjective in some sense. For instance, danceability might be common for people to accept, but the scaling could be subjective and thus biased.
### Future Direction

I think I will be acquiring more data from other platforms with more features and thus could improve the model and this project. Personally, I produced my own music during my free time, so it would also be important to understand the taste of the audience.

## Reference

[1]Spotify Streaming Charts. https://spotifycharts.com/regional
[2]Explanation for variables from Spotify API data. https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/
[3]Spotify logo picture. https://www.spotify.com/us/