Final Project

Mini Project 2

Aditya Jain

2022-12-11

Contents

Exploratory data analysis

- Glimpse of the data

```
##                   song_name song_popularity song_duration_ms acousticness
## 1 Boulevard of Broken Dreams              73           262333     0.005520
## 2                In The End              66           216933     0.010300
## 3          Seven Nation Army              76           231733     0.008170
## 4                By The Way              74           216933     0.026400
## 5          How You Remind Me              56           223826     0.000954
## 6           Bring Me To Life              80           235893     0.008950
##   danceability energy instrumentalness key liveness loudness audio_mode
## 1        0.496  0.682         2.94e-05   8   0.0589   -4.095          1
## 2        0.542  0.853         0.00e+00   3   0.1080   -6.407          0
## 3        0.737  0.463         4.47e-01   0   0.2550   -7.828          1
## 4        0.451  0.970         3.55e-03   0   0.1020   -4.938          1
## 5        0.447  0.766         0.00e+00  10   0.1130   -5.065          1
## 6        0.316  0.945         1.85e-06   4   0.3960   -3.169          0
##   speechiness   tempo time_signature audio_valence
## 1      0.0294 167.060              4         0.474
## 2      0.0498 105.256              4         0.370
## 3      0.0792 123.881              4         0.324
## 4      0.1070 122.444              4         0.198
## 5      0.0313 172.011              4         0.574
## 6      0.1240 189.931              4         0.320
```

- Checking for null value

```
## [1] 0
```

- Number of duplicates found

```
## [1] 3909
```

- Duplicates removed as no more duplicates found

```
## [1] 0
```

- Summary of the dataset.

```
##   song_name         song_popularity  song_duration_ms   acousticness
##  Length:14926       Min.   :  0.00   Min.   :  12000   Min.   :0.000001
##  Class :character   1st Qu.: 37.00   1st Qu.: 183944   1st Qu.:0.023600
##  Mode  :character   Median : 52.00   Median : 211846   Median :0.139000
##                     Mean   : 48.75   Mean   : 218950   Mean   :0.270452
##                     3rd Qu.: 63.75   3rd Qu.: 244720   3rd Qu.:0.458000
##                     Max.   :100.00   Max.   :1799346   Max.   :0.996000
##   danceability        energy        instrumentalness         key
##  Min.   :0.0000   Min.   :0.00107   Min.   :0.0000000   Min.   : 0.000
##  1st Qu.:0.5240   1st Qu.:0.49600   1st Qu.:0.0000000   1st Qu.: 2.000
##  Median :0.6360   Median :0.67200   Median :0.0000208   Median : 5.000
##  Mean   :0.6245   Mean   :0.63976   Mean   :0.0920668   Mean   : 5.301
```

```
##  3rd Qu.:0.7400   3rd Qu.:0.81800   3rd Qu.:0.0051050   3rd Qu.: 8.000
##  Max.   :0.9870   Max.   :0.99900   Max.   :0.9970000   Max.   :11.000
##     liveness         loudness         audio_mode      speechiness
##  Min.   :0.0109   Min.   :-38.768   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0930   1st Qu.: -9.389   1st Qu.:0.0000   1st Qu.:0.03720
##  Median :0.1220   Median : -6.750   Median :1.0000   Median :0.05410
##  Mean   :0.1804   Mean   : -7.677   Mean   :0.6319   Mean   :0.09942
##  3rd Qu.:0.2240   3rd Qu.: -4.991   3rd Qu.:1.0000   3rd Qu.:0.11300
##  Max.   :0.9860   Max.   :  1.585   Max.   :1.0000   Max.   :0.94100
##      tempo        time_signature  audio_valence
##  Min.   :  0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.: 98.12   1st Qu.:4.000   1st Qu.:0.3320
##  Median :120.02   Median :4.000   Median :0.5270
##  Mean   :121.11   Mean   :3.953   Mean   :0.5270
##  3rd Qu.:139.94   3rd Qu.:4.000   3rd Qu.:0.7278
##  Max.   :242.32   Max.   :5.000   Max.   :0.9840
```

- More information on variable structure and data type.

```
## 'data.frame':    14926 obs. of  15 variables:
##  $ song_name       : chr  "Boulevard of Broken Dreams" "In The End" "Seven Nation Army" "By The Way"
##  $ song_popularity : int  73 66 76 74 56 80 81 76 80 81 ...
##  $ song_duration_ms: int  262333 216933 231733 216933 223826 235893 199893 213800 222586 203346 ...
##  $ acousticness    : num  0.00552 0.0103 0.00817 0.0264 0.000954 0.00895 0.000504 0.00148 0.00108 0.0
##  $ danceability    : num  0.496 0.542 0.737 0.451 0.447 0.316 0.581 0.613 0.33 0.542 ...
##  $ energy          : num  0.682 0.853 0.463 0.97 0.766 0.945 0.887 0.953 0.936 0.905 ...
##  $ instrumentalness: num  2.94e-05 0.00 4.47e-01 3.55e-03 0.00 1.85e-06 1.11e-03 5.82e-04 0.00 1.04e-
##  $ key             : int  8 3 0 0 10 4 4 2 1 9 ...
##  $ liveness        : num  0.0589 0.108 0.255 0.102 0.113 0.396 0.268 0.152 0.0926 0.136 ...
##  $ loudness        : num  -4.09 -6.41 -7.83 -4.94 -5.07 ...
##  $ audio_mode      : int  1 0 1 1 1 0 0 1 1 1 ...
##  $ speechiness     : num  0.0294 0.0498 0.0792 0.107 0.0313 0.124 0.0624 0.0855 0.0917 0.054 ...
##  $ tempo           : num  167 105 124 122 172 ...
##  $ time_signature  : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ audio_valence   : num  0.474 0.37 0.324 0.198 0.574 0.32 0.724 0.537 0.234 0.374 ...
```

- Visualizing correlation values.

```
##                  song_popularity song_duration_ms  acousticness danceability
## song_popularity      1.000000000     -0.007765718 -0.0316578066  0.056414506
## song_duration_ms    -0.007765718      1.000000000 -0.1133140766 -0.089338456
## acousticness        -0.031657807     -0.113314077  1.0000000000 -0.171858817
## danceability         0.056414506     -0.089338456 -0.1718588175  1.000000000
## energy              -0.016468240      0.096375412 -0.6792346937  0.053657351
## instrumentalness    -0.093032438     -0.024850232  0.1792587265 -0.124926426
## key                 -0.003877688     -0.006749269 -0.0006322858  0.007598662
## liveness            -0.038553227      0.020757180 -0.0852421645 -0.094373346
## loudness             0.052113953      0.027837252 -0.5696810521  0.173495293
## audio_mode           0.008824035     -0.028055814  0.0594434481 -0.099038877
## speechiness         -0.001970926     -0.079438296 -0.0854350895  0.200714873
## tempo               -0.029559962      0.013607899 -0.1447363311 -0.127313027
## time_signature       0.023306872      0.004080416 -0.1511795518  0.136211239
## audio_valence       -0.043962625     -0.069762753 -0.1269151172  0.350012131
```

```
##                       energy instrumentalness           key     liveness
## song_popularity  -0.01646824     -0.093032438 -0.0038776883 -0.038553227
## song_duration_ms  0.09637541     -0.024850232 -0.0067492687  0.020757180
## acousticness     -0.67923469      0.179258726 -0.0006322858 -0.085242164
## danceability      0.05365735     -0.124926426  0.0075986619 -0.094373346
## energy            1.00000000     -0.221754881  0.0146532170  0.175979295
## instrumentalness -0.22175488      1.000000000 -0.0084960035 -0.039821488
## key               0.01465322     -0.008496004  1.0000000000 -0.003254359
## liveness          0.17597930     -0.039821488 -0.0032543593  1.000000000
## loudness          0.76570640     -0.398911777  0.0080016047  0.111715887
## audio_mode       -0.04903967     -0.018852094 -0.1749545133 -0.001729012
## speechiness       0.07213789     -0.076705498  0.0285646613  0.094228846
## tempo             0.18251702     -0.041032108 -0.0050332654  0.025681427
## time_signature    0.14351582     -0.068656107 -0.0087501637  0.014792316
## audio_valence     0.31502350     -0.197782526  0.0236261208  0.016571791
##                      loudness   audio_mode  speechiness        tempo
## song_popularity   0.052113953  0.008824035 -0.001970926 -0.029559962
## song_duration_ms  0.027837252 -0.028055814 -0.079438296  0.013607899
## acousticness     -0.569681052  0.059443448 -0.085435090 -0.144736331
## danceability      0.173495293 -0.099038877  0.200714873 -0.127313027
## energy            0.765706397 -0.049039667  0.072137889  0.182517016
## instrumentalness -0.398911777 -0.018852094 -0.076705498 -0.041032108
## key               0.008001605 -0.174954513  0.028564661 -0.005033265
## liveness          0.111715887 -0.001729012  0.094228846  0.025681427
## loudness          1.000000000 -0.055792534  0.078417608  0.140216852
## audio_mode       -0.055792534  1.000000000 -0.110685766  0.021863261
## speechiness       0.078417608 -0.110685766  1.000000000  0.056410565
## tempo             0.140216852  0.021863261  0.056410565  1.000000000
## time_signature    0.112596077 -0.020789298  0.050777450  0.008326366
## audio_valence     0.209803035  0.002253655  0.012884390  0.051647510
##                  time_signature audio_valence
## song_popularity     0.023306872  -0.043962625
## song_duration_ms    0.004080416  -0.069762753
## acousticness       -0.151179552  -0.126915117
## danceability        0.136211239   0.350012131
## energy              0.143515819   0.315023499
## instrumentalness   -0.068656107  -0.197782526
## key                -0.008750164   0.023626121
## liveness            0.014792316   0.016571791
## loudness            0.112596077   0.209803035
## audio_mode         -0.020789298   0.002253655
## speechiness         0.050777450   0.012884390
## tempo               0.008326366   0.051647510
## time_signature      1.000000000   0.090684092
## audio_valence       0.090684092   1.000000000
```

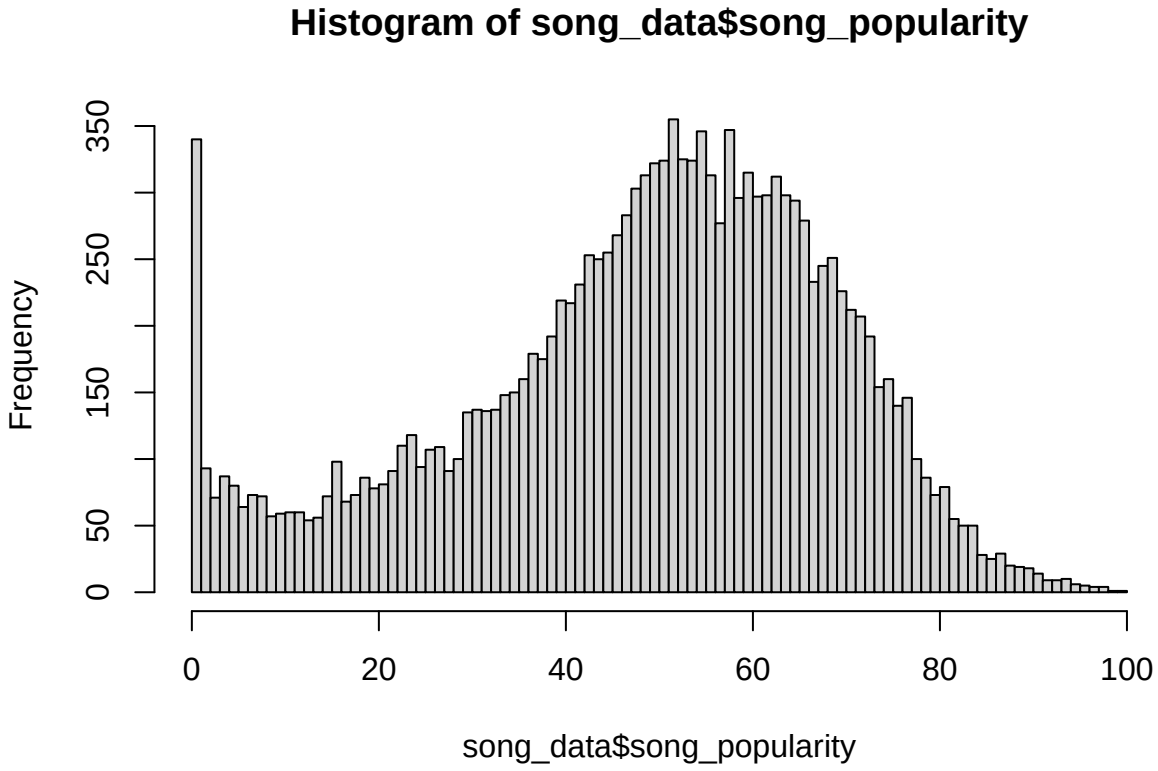
Variables with good correlations

- Acousticness - energy
- Acousticness - loudness
- Danceability + audio valence
- Loudness + energy

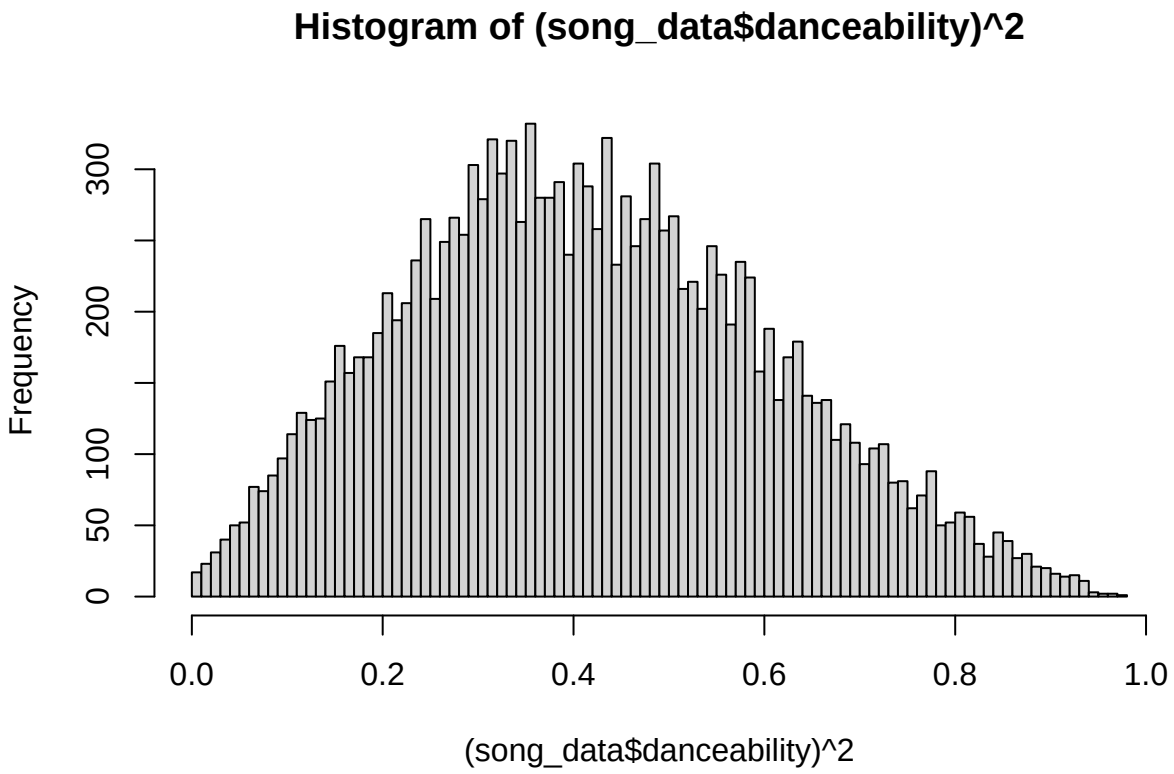
- Audio valence + danceability + energy
- Performing some variable transformation and scaling/normalizing tempo and loudness attributes.

```
##                    song_name song_popularity acousticness danceability energy
## 1 Boulevard of Broken Dreams              73     0.005520        0.496  0.682
## 2                  In The End              66     0.010300        0.542  0.853
## 3          Seven Nation Army              76     0.008170        0.737  0.463
## 4                  By The Way              74     0.026400        0.451  0.970
## 5          How You Remind Me              56     0.000954        0.447  0.766
## 6           Bring Me To Life              80     0.008950        0.316  0.945
##   instrumentalness key liveness audio_mode speechiness time_signature
## 1         2.94e-05   8   0.0589          1      0.0294              4
## 2         0.00e+00   3   0.1080          0      0.0498              4
## 3         4.47e-01   0   0.2550          1      0.0792              4
## 4         3.55e-03   0   0.1020          1      0.1070              4
## 5         0.00e+00  10   0.1130          1      0.0313              4
## 6         1.85e-06   4   0.3960          0      0.1240              4
##   audio_valence song_duration_min loudness_scale tempo_scale
## 1         0.474          4.372217      0.8592422   0.6894246
## 2         0.370          3.615550      0.8019478   0.4343714
## 3         0.324          3.862217      0.7667336   0.5112332
## 4         0.198          3.615550      0.8383515   0.5053029
## 5         0.574          3.730433      0.8352043   0.7098565
## 6         0.320          3.931550      0.8821897   0.7838089
```

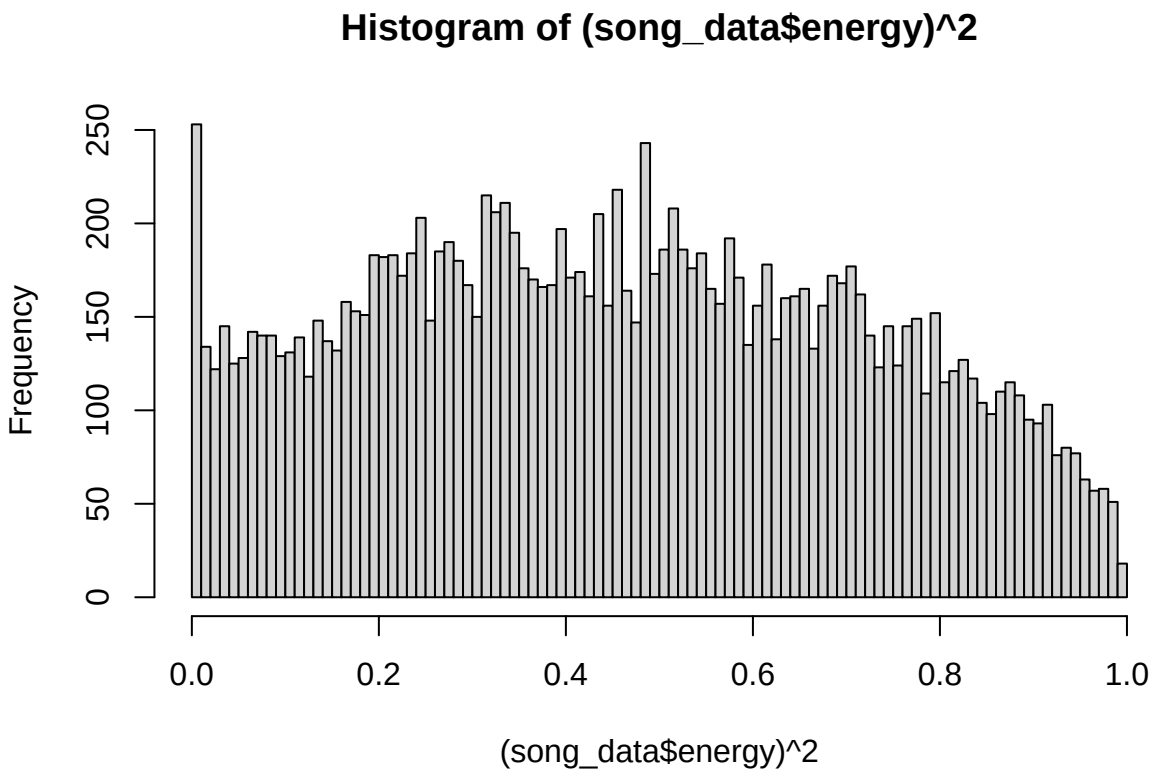
- Visualizing distribution of target variable.



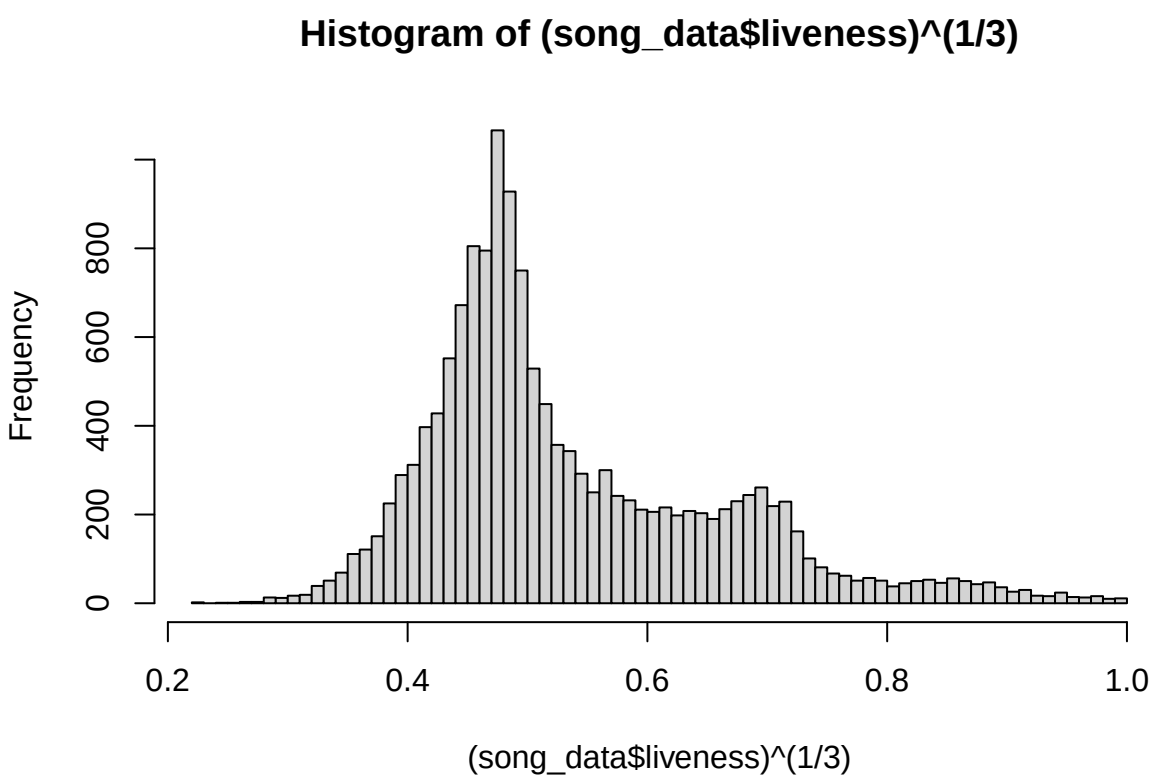
* Visualizing other dependent variables and testing transformations to make data normalised

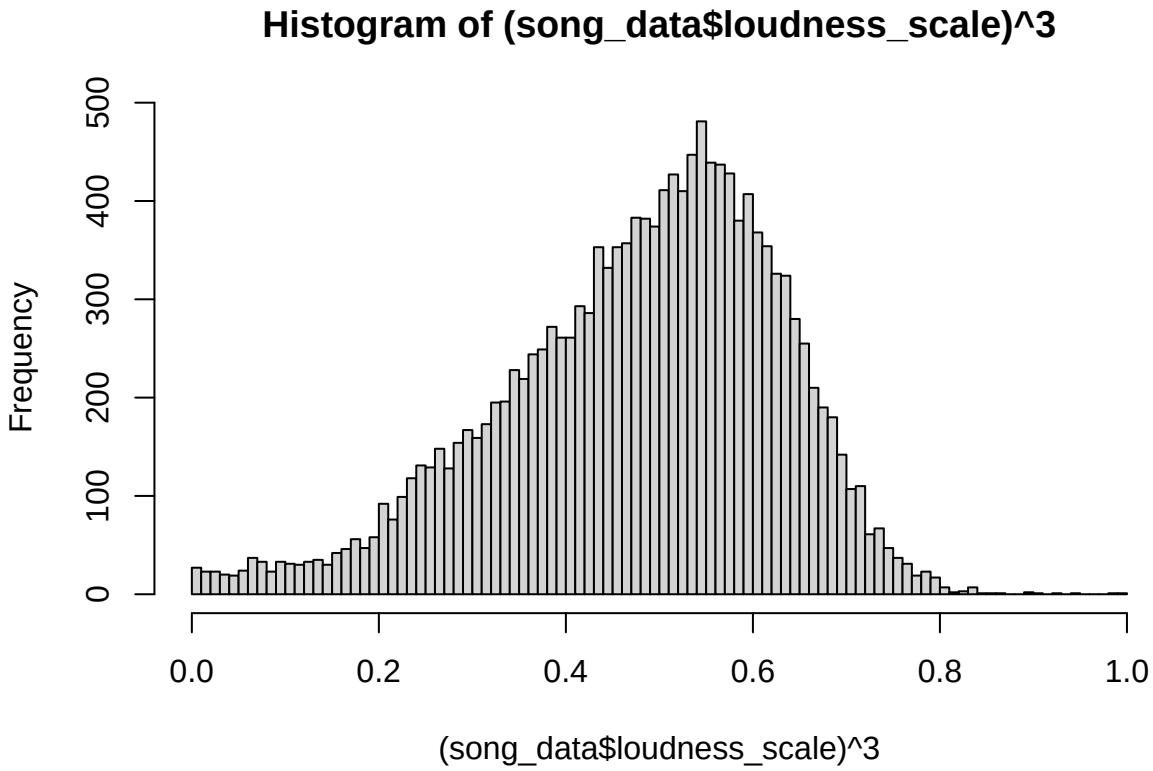


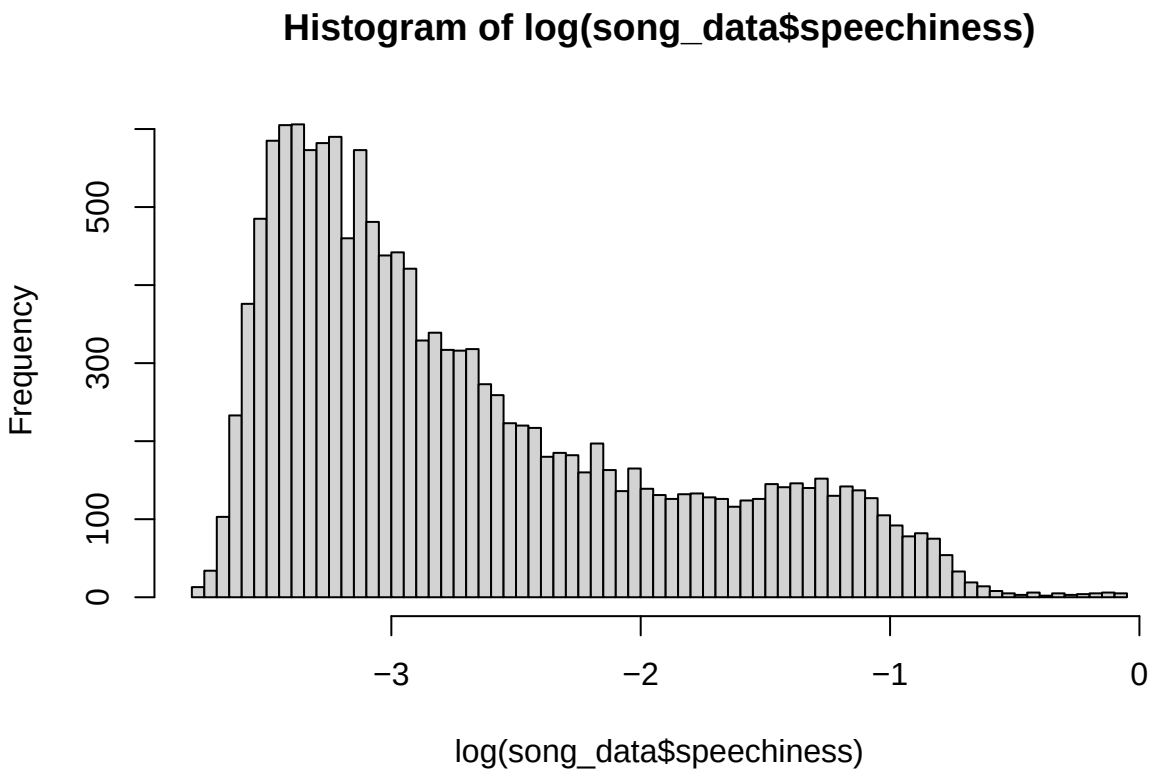
* The attributes like danceability and breaks were skewed and as the value range lies in the range of 0-1 squaring became the best option.

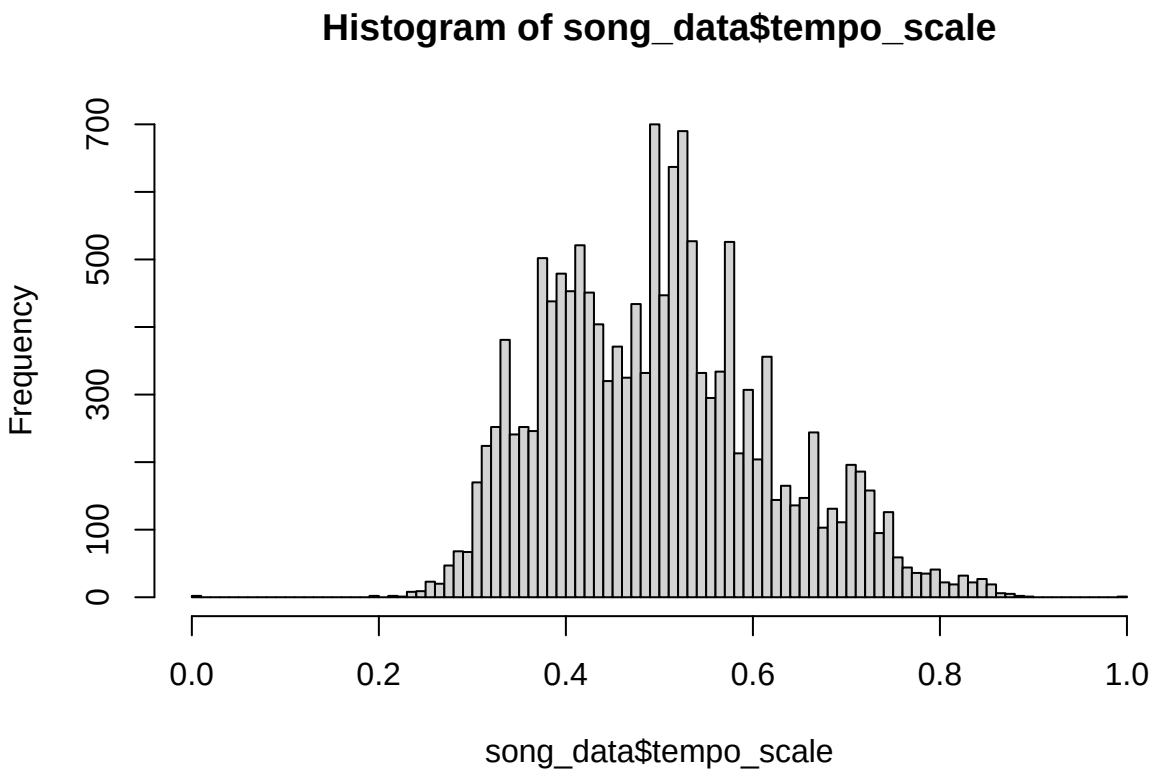


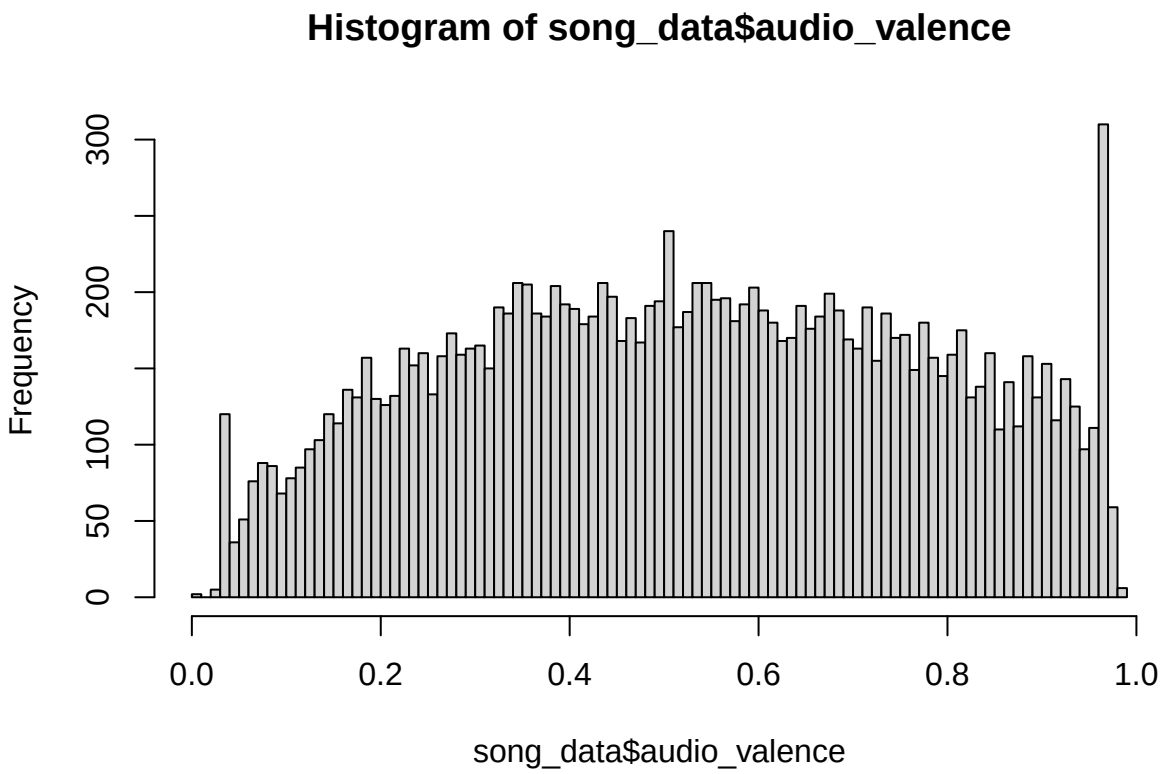
* Taking cube root to minimize high left skewness in the liveness attribute

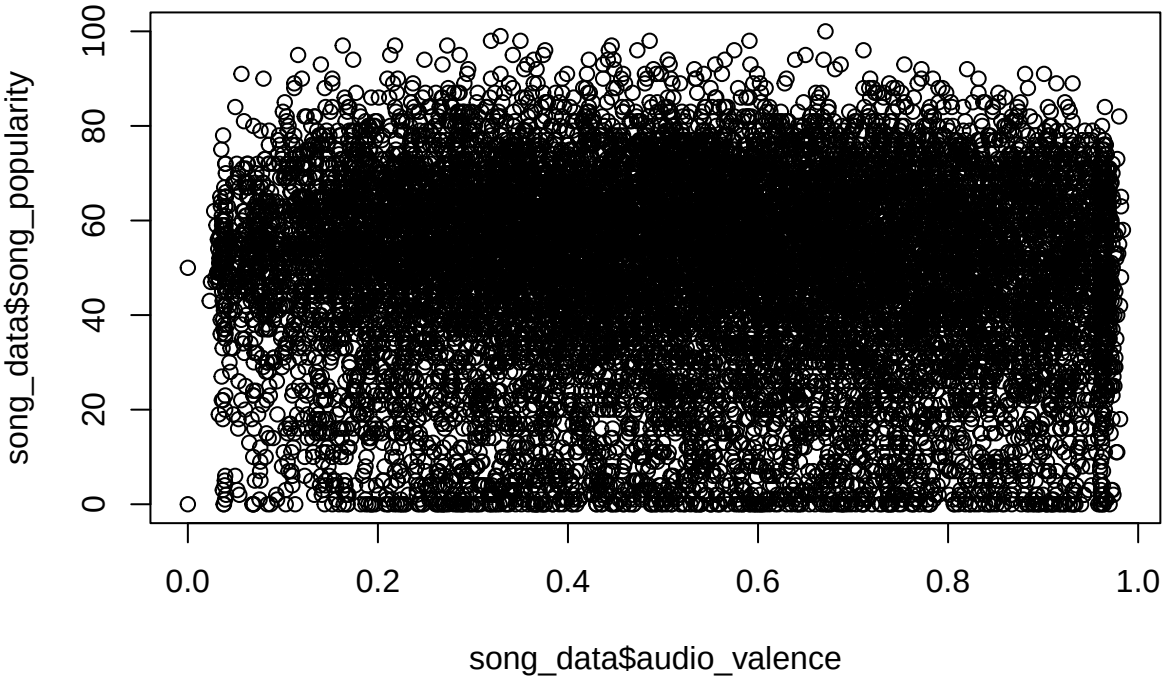

Histogram of (song_data$liveness)^(1/3)
Frequency
0
200
400
600
800
0.2
0.4
0.6
0.8
1.0
(song_data$liveness)^(1/3)

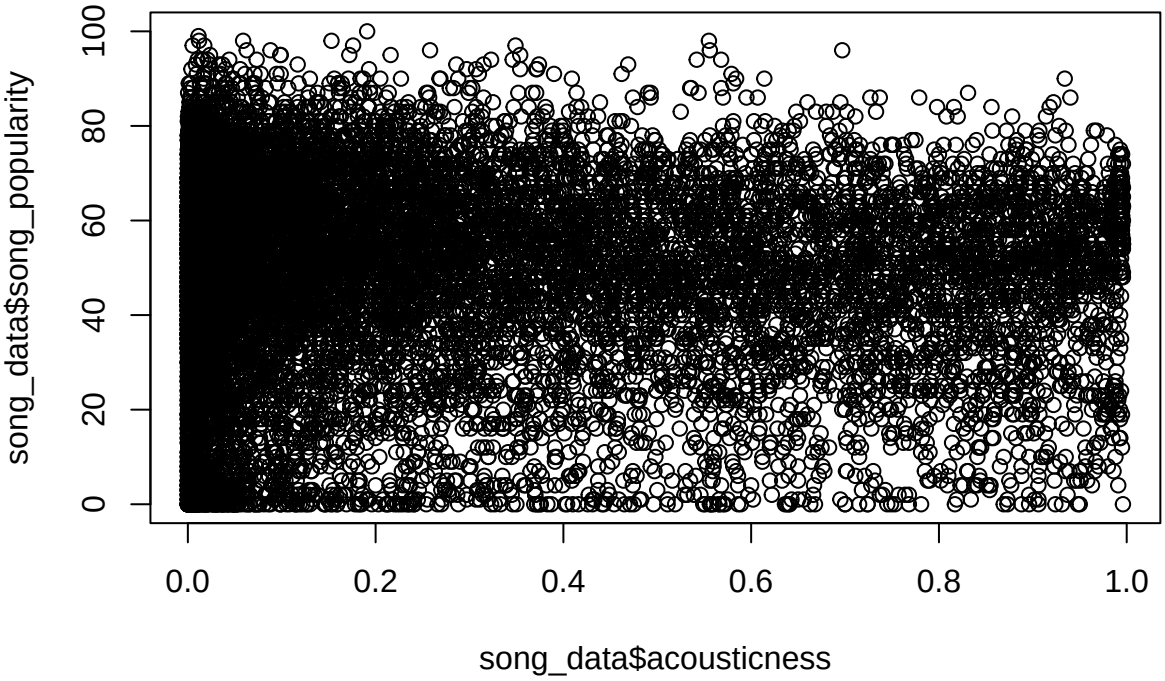

Histogram of (song_data$loudness_scale)^3
Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
(song_data$loudness_scale)^3

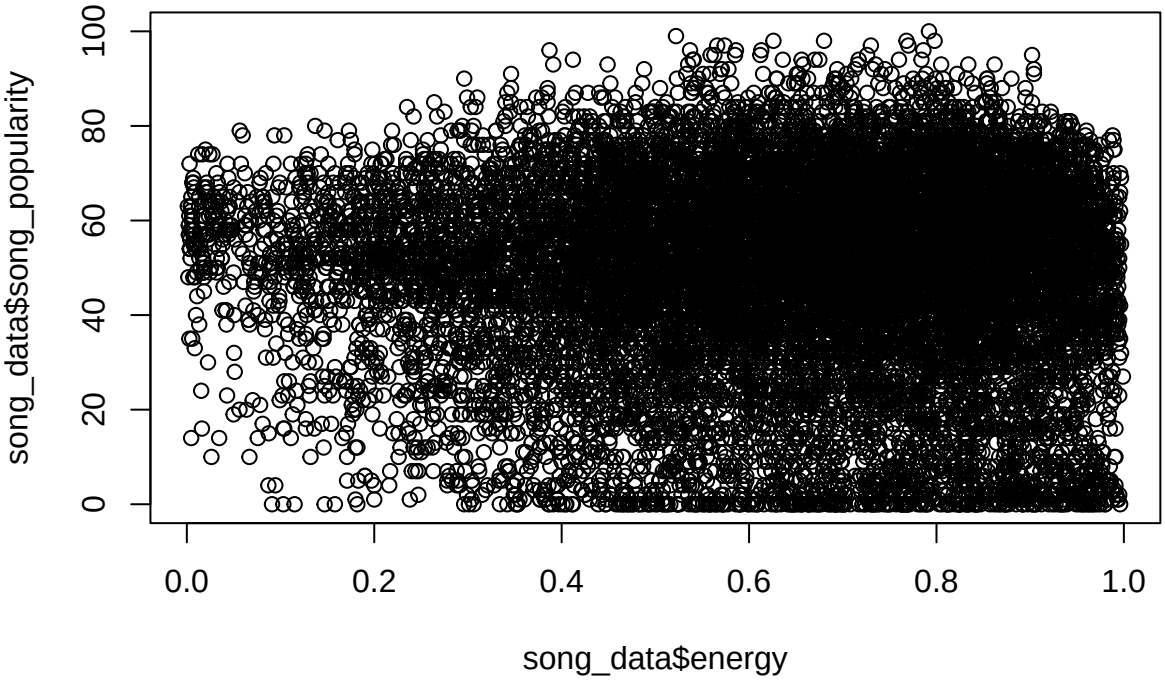

Histogram of log(song_data$speechiness)
Frequency
0
100
300
500
−3
−2
−1
0
log(song_data$speechiness)


Histogram of song_data$tempo_scale
Frequency
0
100
300
500
700
0.0
0.2
0.4
0.6
0.8
1.0
song_data$tempo_scale


Histogram of song_data$audio_valence
Frequency
0
50
100
200
300
0.0
0.2
0.4
0.6
0.8
1.0
song_data$audio_valence


song_data$song_popularity
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
song_data$audio_valence


song_data$song_popularity
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
song_data$acousticness



- Generating meaningful attributes in the data frame and visualizing

Baseline Models

Generating two baseline models, with one being mean value and other as the basic Linear regression model with all the attributes included. Here, we have used the entire data set to generate the models. Along with the model summary, we are also calculating RMSE and MAE error of our model to get more accuracy scores.

```
## [1] "RMSE and MAE errors for train data for mean baseline model"
```

```
## [1] "20.4255943596398 16.2379908432983"
```

```
## [1] "RMSE and MAE errors for test data for mean baseline model"
```

```
## [1] "20.2691434349854 16.167458981018"
```

```
##
## Call:
## lm(formula = song_popularity ~ acousticness + ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.359 -11.108   2.968  14.637  50.685
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          79.61796    5.57129  14.291  < 2e-16 ***
## acousticness         -3.38950    0.96935  -3.497 0.000473 ***
## danceability         16.87348    7.37487   2.288 0.022159 *
## energy                0.23052    6.03011   0.038 0.969506
## instrumentalness     -8.72839    0.92668  -9.419  < 2e-16 ***
## key                   0.02646    0.05553   0.476 0.633758
## liveness              4.66718    5.17221   0.902 0.366888
## audio_mode            0.73099    0.41937   1.743 0.081351 .
## speechiness          -4.18128    2.01982  -2.070 0.038465 *
## time_signature        0.92963    0.63134   1.472 0.140926
## audio_valence        -5.84322    0.95703  -6.106 1.06e-09 ***
## song_duration_min     0.02887    0.19823   0.146 0.884229
## loudness_scale      -74.05608    9.88435  -7.492 7.32e-14 ***
## tempo_scale          -2.66158    1.70058  -1.565 0.117589
## danceability_square  -7.38678    6.08394  -1.214 0.224720
## energy_square       -10.72369    4.60856  -2.327 0.019989 *
## liveness_cuberoot   -10.16632    6.06882  -1.675 0.093931 .
## loudness_cube        63.23404    6.30977  10.022  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.08 on 10430 degrees of freedom
## Multiple R-squared:  0.03534,    Adjusted R-squared:  0.03377
## F-statistic: 22.48 on 17 and 10430 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE and MAE errors for test data in case of baseling linear regression model"
```

```
## [1] "19.9555167782784 15.8968619018375"
```

- Playing with different attributes to see variation in model accuracy and determine important features for the final model.

```
##
## Call:
## lm(formula = song_popularity ~ danceability_square + energy_square +
##     instrumentalness + loudness_cube + audio_valence, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.827 -11.274   2.916  14.671  48.297
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          46.0012     0.9061  50.769  < 2e-16 ***
## danceability_square   6.3548     1.1410   5.569 2.62e-08 ***
## energy_square        -8.0554     1.2215  -6.595 4.47e-11 ***
## instrumentalness     -6.8151     0.8893  -7.663 1.97e-14 ***
## loudness_cube        16.2590     2.1759   7.472 8.52e-14 ***
## audio_valence        -6.3429     0.9092  -6.976 3.22e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.18 on 10442 degrees of freedom
## Multiple R-squared:  0.02474,    Adjusted R-squared:  0.02427
## F-statistic: 52.97 on 5 and 10442 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE and MAE errors for test data for new model"
```

```
## [1] "20.050535950206 15.9936715479242"
```

```
## [1] "RMSE and MAE errors for test data in case of baseling linear regression model"
```

```
## [1] "19.9555167782784 15.8968619018375"
```

Feature Engineering

The goal of this feature engineering is to create new input features that may be more predictive of the popularity of a song than the original input features. To do this, we have divided the acousticness of a song by its energy and multiplied its danceability by its energy.

Dividing the acousticness of a song by its energy can help capture the relative acousticness of a song compared to its overall energy level. This can be useful because songs with high acousticness and low energy may have a different impact on popularity than songs with low acousticness and high energy. By dividing the acousticness by the energy, we can create a new input feature that captures this relative acousticness and may be more predictive of popularity.

Multiplying the danceability of a song by its energy can also help capture the relative danceability of a song compared to its overall energy level. This can be useful because songs with high danceability and low energy may have a different impact on popularity than songs with low danceability and high energy. By multiplying the danceability by the energy, we can create a new input feature that captures this relative danceability and may be more predictive of popularity.

Overall, these feature engineering choices make sense because they create new input features that capture the relative acousticness and danceability of a song compared to its energy level, which may be more predictive of popularity than the original input features. By incorporating these new features into a regression model, we can potentially improve the model's ability to predict the popularity of a song.

```
## [1] "0.01758634038705"

## [1] "0.0331309959961779"

##
## Call:
## lm(formula = song_popularity ~ instrumentalness + f1 + f2, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.931 -11.238   3.016  14.718  49.428
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      48.80414    0.37226 131.102  < 2e-16 ***
## instrumentalness -8.50097    0.84413 -10.071  < 2e-16 ***
## f1                0.05344    0.01273   4.199  2.7e-05 ***
## f2                3.46433    1.55258   2.231   0.0257 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.31 on 10444 degrees of freedom
## Multiple R-squared:  0.01118,    Adjusted R-squared:  0.0109
## F-statistic: 39.37 on 3 and 10444 DF,  p-value: < 2.2e-16

## [1] "RMSE and MAE errors for test data for new model"

## [1] "20.1705284939844 16.1169604221682"

## [1] "RMSE and MAE errors for test data in case of baseling linear regression model"

## [1] "19.9555167782784 15.8968619018375"
```

Oversampling

Over-sampling is a technique used in machine learning and data analysis to balance the distribution of classes in a dataset. This is often done when the dataset is imbalanced, meaning that one class (or group of classes) is significantly more represented than the other classes. In this situation, over-sampling can be used to increase the number of samples in the under-represented classes, so that the class distribution becomes more balanced and the model can learn from all the classes more effectively

- Randomly splitting data to split into train and test set for our sampled data and

New Baseline Models

- Re generating baseline model for the new dataset.

```
## [1] "RMSE and MAE errors for train data for mean baseline model"
```

```
## [1] "19.4264422452517 15.3925802152758"
```

```
## [1] "RMSE and MAE errors for test data for mean baseline model"
```

```
## [1] "19.052656791997 15.1203950720814"
```

```
##
## Call:
## lm(formula = song_popularity ~ acousticness + ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.124 -10.146   4.362  13.331  43.656
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          79.405974   6.608504  12.016  < 2e-16 ***
## acousticness         -4.374107   1.164788  -3.755 0.000175 ***
## danceability         10.735714   8.835948   1.215 0.224406
## energy                6.639020   7.081176   0.938 0.348504
## instrumentalness     -9.481783   1.174532  -8.073 8.04e-16 ***
## key                  -0.005667   0.064250  -0.088 0.929715
## liveness             -8.669916   6.104028  -1.420 0.155548
## audio_mode            1.005277   0.488450   2.058 0.039618 *
## speechiness          -2.841389   2.337799  -1.215 0.224250
## time_signature        0.780714   0.749133   1.042 0.297376
## audio_valence        -5.108938   1.119269  -4.565 5.09e-06 ***
## song_duration_min     0.506166   0.234680   2.157 0.031053 *
## loudness_scale      -77.279013  11.643742  -6.637 3.44e-11 ***
## tempo_scale          -2.777155   1.989945  -1.396 0.162882
## danceability_square  -4.430017   7.222385  -0.613 0.539650
## energy_square       -17.140655   5.395735  -3.177 0.001496 **
## liveness_cuberoot     5.435923   7.147652   0.761 0.446971
## loudness_cube        68.450791   7.324432   9.346  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 19.01 on 6920 degrees of freedom
## Multiple R-squared:  0.04484,    Adjusted R-squared:  0.04249
## F-statistic: 19.11 on 17 and 6920 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE and MAE errors for test data in case of baseling linear regression model"
```

```
## [1] "18.5459607880856 14.6475080612153"
```

- Testing other models to analyse model performance

```
##
## Call:
## lm(formula = song_popularity ~ acousticness + energy_square +
##     instrumentalness + loudness_scale + audio_valence, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.133 -10.353   4.531  13.382  44.327
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         47.802      2.716  17.598  < 2e-16 ***
## acousticness        -4.281      1.053  -4.064 4.87e-05 ***
## energy_square       -8.919      1.394  -6.396 1.69e-10 ***
## instrumentalness    -7.439      1.161  -6.406 1.59e-10 ***
## loudness_scale      19.675      3.612   5.447 5.29e-08 ***
## audio_valence       -4.885      1.003  -4.871 1.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.17 on 6932 degrees of freedom
## Multiple R-squared:  0.02665,    Adjusted R-squared:  0.02595
## F-statistic: 37.96 on 5 and 6932 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE and MAE errors for test data for new model"
```

```
## [1] "18.7805485118314 14.8618706753057"
```

```
## [1] "RMSE and MAE errors for test data in case of baseling linear regression model"
```

```
## [1] "18.5459607880856 14.6475080612153"
```

```
##
## Call:
## lm(formula = song_popularity ~ loudness_scale, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.017 -10.221   4.778  13.454  44.676
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      42.199      1.804  23.395  < 2e-16 ***
## loudness_scale   16.061      2.316   6.936  4.4e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.36 on 6936 degrees of freedom
## Multiple R-squared:  0.006888,   Adjusted R-squared:  0.006745
## F-statistic: 48.11 on 1 and 6936 DF,  p-value: 4.403e-12
```

```
## [1] "RMSE and MAE errors for test data for new model"
```

```
## [1] "18.9874478593403 15.0451495072724"
```

```
## [1] "RMSE and MAE errors for test data in case of baseling linear regression model"
```

```
## [1] "18.5459607880856 14.6475080612153"
```

Final Model

```
##
## Call:
## lm(formula = song_popularity ~ acousticness + energy_square +
##     instrumentalness + loudness_scale + audio_valence, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.133 -10.353   4.531  13.382  44.327
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        47.802      2.716  17.598  < 2e-16 ***
## acousticness       -4.281      1.053  -4.064 4.87e-05 ***
## energy_square      -8.919      1.394  -6.396 1.69e-10 ***
## instrumentalness   -7.439      1.161  -6.406 1.59e-10 ***
## loudness_scale     19.675      3.612   5.447 5.29e-08 ***
## audio_valence      -4.885      1.003  -4.871 1.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.17 on 6932 degrees of freedom
## Multiple R-squared:  0.02665,    Adjusted R-squared:  0.02595
## F-statistic: 37.96 on 5 and 6932 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE and MAE errors for test data in case of baseling linear regression model"
```

```
## [1] "18.5459607880856 14.6475080612153"
```

```
## [1] "RMSE and MAE errors for test data for our final model"
```

```
## [1] "18.7805485118314 14.8618706753057"
```

Some plot and attribute desciptions

The sign of a parameter in a model indicates the direction of the relationship between the input and output variables. A positive sign indicates a positive relationship, where an increase in the input variable is associated with an increase in the output variable. A negative sign indicates a negative relationship, where an increase in the input variable is associated with a decrease in the output variable.

In our model, the popularity of the song is positively associated with the loudness, however if its too much instrumental or energetic the popularity gets affected negatively, which make sense in case of more use of instrumental but seems doubtful for energy.

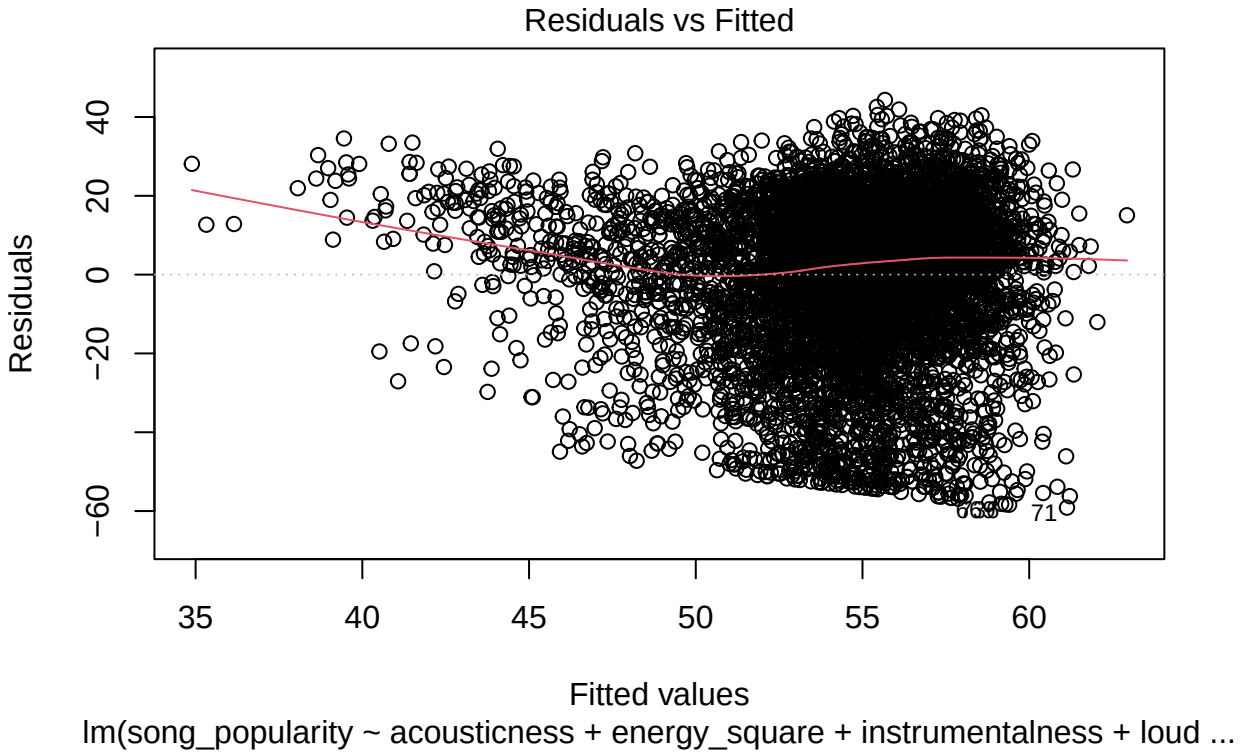
The output consists of four plots that can help visualize and understand the results of a linear regression. The first plot is a scatterplot of the residuals (the difference between the observed values and the predicted values) versus the fitted values (the predicted values). This plot can help identify any patterns or outliers in the residuals, which can indicate potential problems with the model.

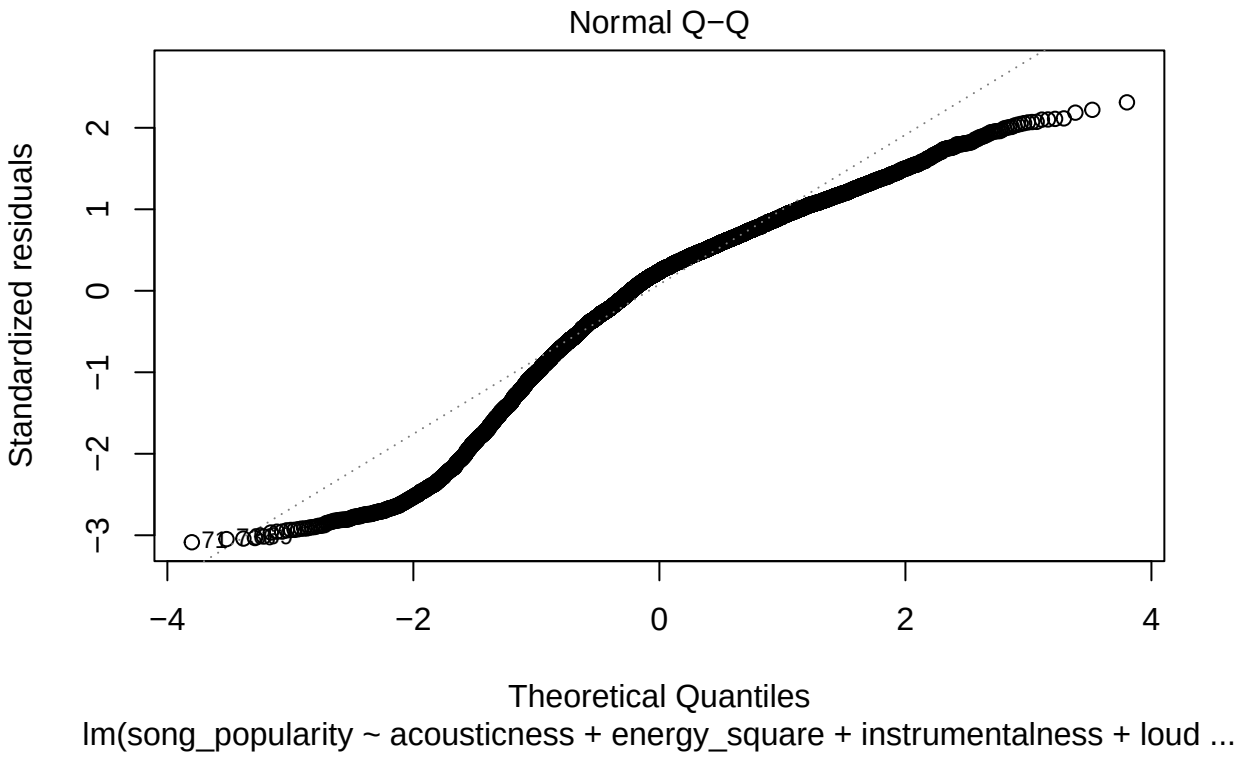
The second plot is a Q-Q plot of the residuals. This plot compares the quantiles of the residuals to the quantiles of a theoretical distribution (usually a normal distribution). This can help assess the normality of the errors in the model.

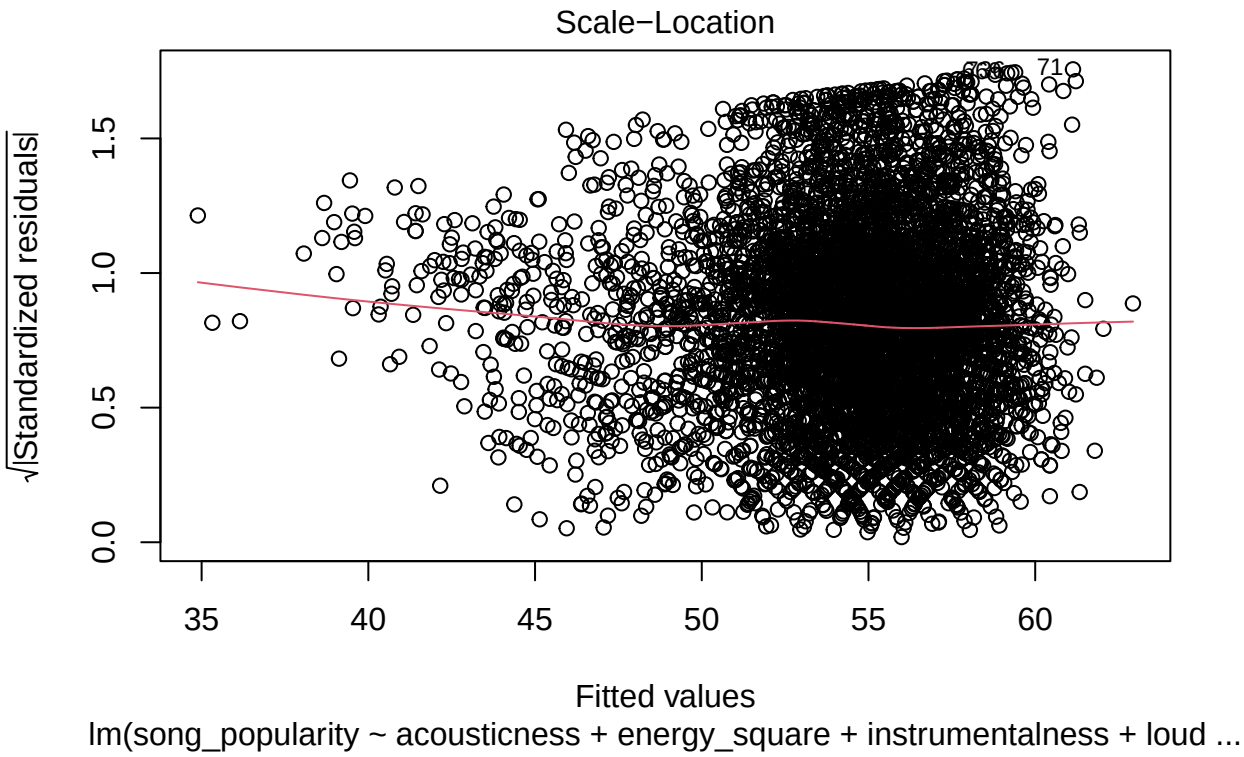
The third plot is a scale-location plot of the square root of the absolute values of the residuals versus the fitted values. This plot can help identify any potential problems with heteroscedasticity (unequal variance) in the model.

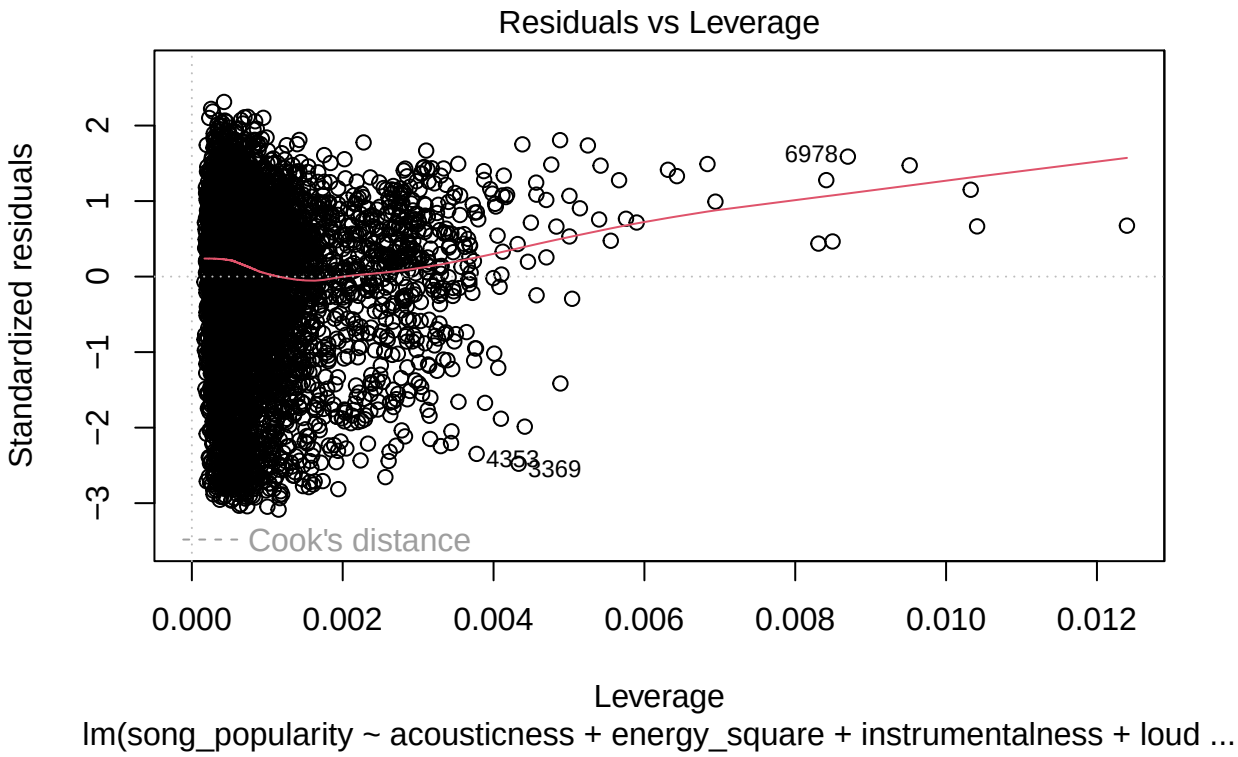
The fourth plot is a plot of Cook's distances versus the row labels. This plot shows the influence of each observation on the fitted values of the model. Observations with high Cook's distances may have a disproportionate influence on the model, and may need to be examined more closely.

Overall, these plots provide valuable insights into the quality of a linear regression model and can help identify potential issues that may need to be addressed.




Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
−4
−2
0
2
4
Theoretical Quantiles
lm(song_popularity ~ acousticness + energy_square + instrumentalness + loud ...

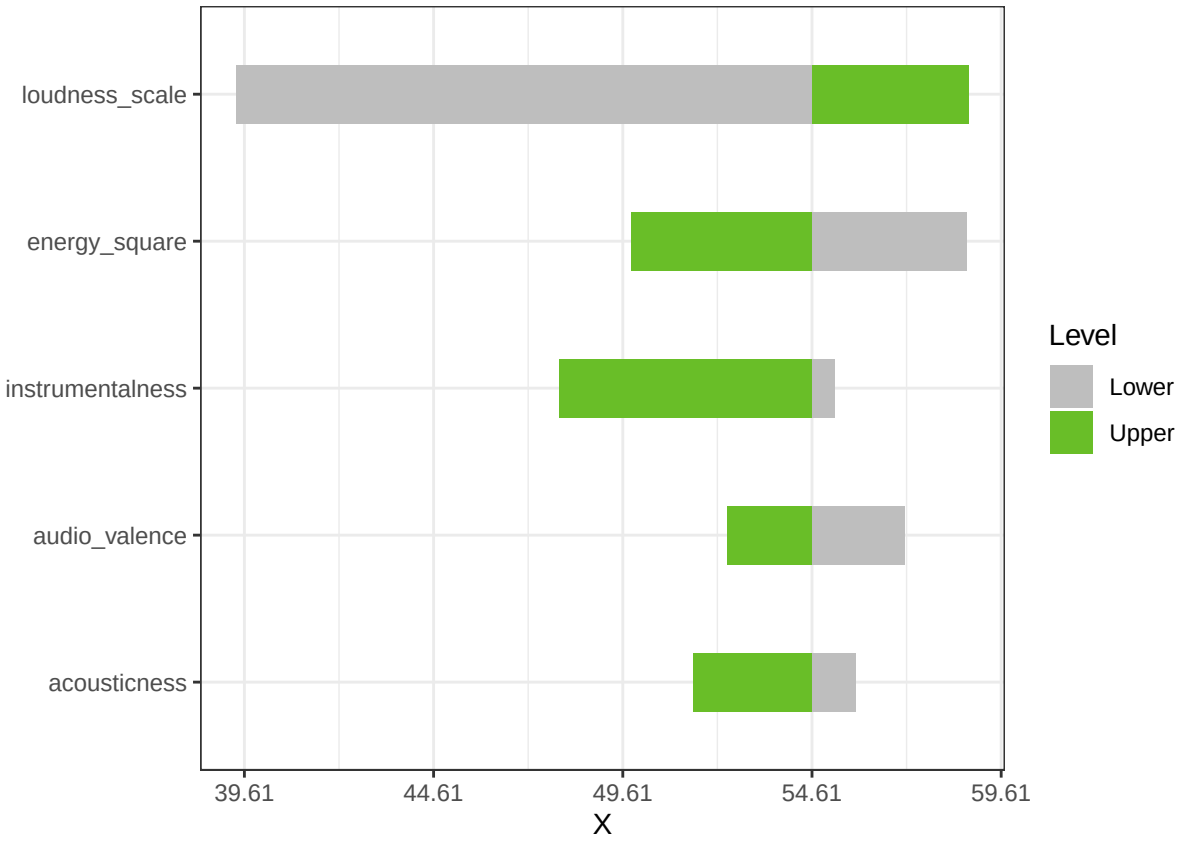

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
35
40
45
50
55
60
71
Fitted values
lm(song_popularity ~ acousticness + energy_square + instrumentalness + loud ...


Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
−3
6978
4353
3369
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Leverage
lm(song_popularity ~ acousticness + energy_square + instrumentalness + loud ...

Sensitivity Analysis

Sensitivity analysis is a technique used to evaluate how the uncertainty in the output of a model or system can be attributed to different sources of uncertainty in the input. It is a systematic way of identifying the most important inputs to a model and understanding how changes in those inputs affect the output of the model.

Sensitivity analysis is often used in decision-making, where it can help identify the key drivers of a decision and assess the potential impact of different scenarios on the outcome of the decision. It can also be used to assess the robustness of a model or system, by examining how sensitive it is to changes in the input parameters.



Predicted vs Actual visualization