STA363 Project1

Alex Zhang

2024-03-27

Section 1: Linear Regression

In this section, we will try to build a linear regression model using all features to predict energy score. We will check conditions for linear regression and look into different coefficients to see how they relate to energy score. Finally, we will evaluate our model's predictive abilities using LOOCV and some supplement plots illustrations.

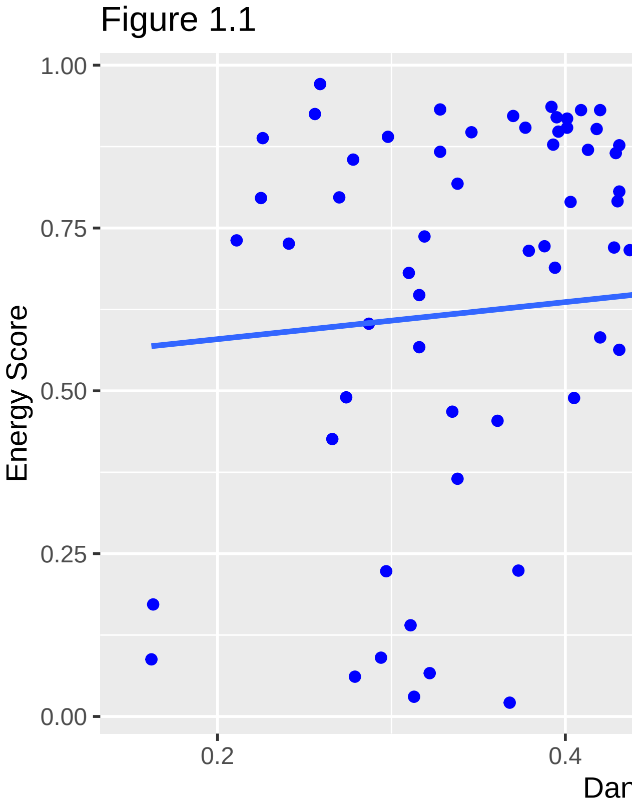
After fitting the model, the coefficient for each feature is showed below,

Table 1: List of Coefficients of LSLR model

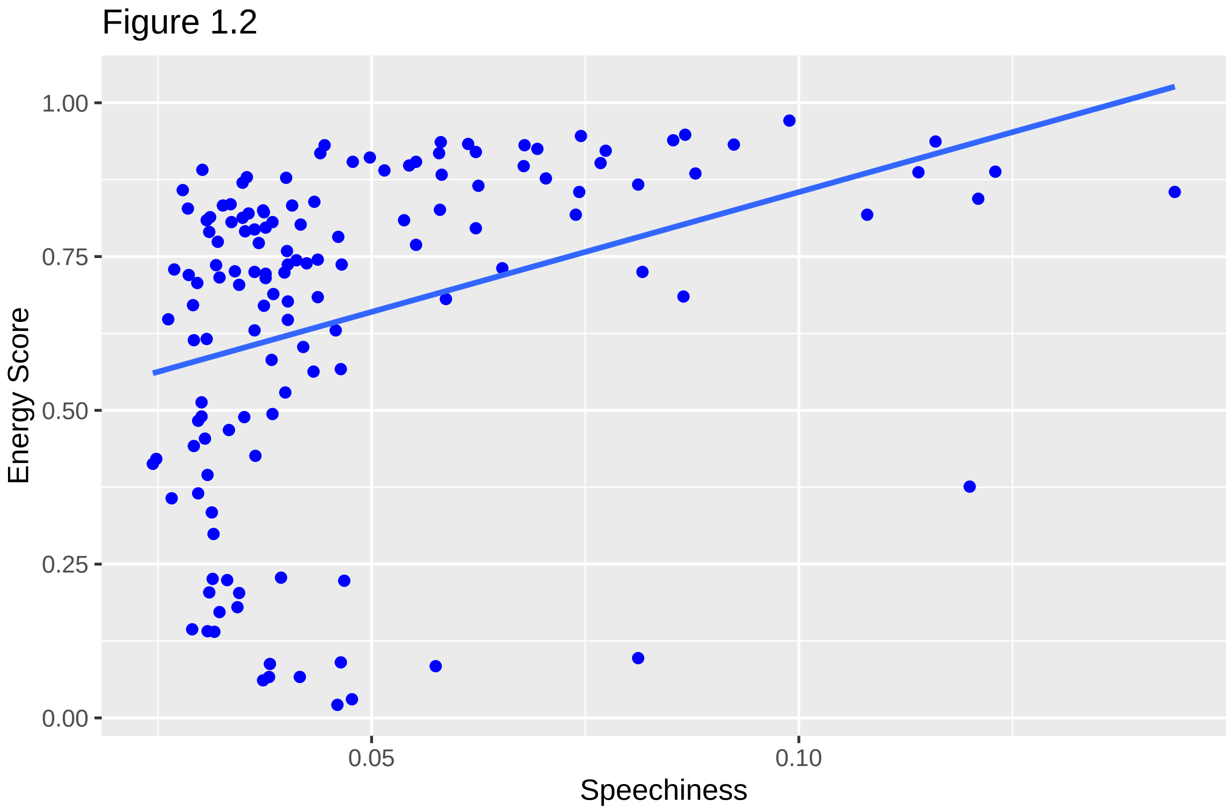
	coefficients
(Intercept)	-0.0681511
danceability	-0.0142331
keyA#	0.0849696
keyB	-0.0624006
keyC	-0.0919273
keyC#	-0.0152381
keyD	-0.0691090
keyD#	0.0272581
keyE	0.0142178
keyF	-0.0348918
keyF#	-0.0490915
keyG	0.1205859
keyG#	0.0136943
modeminor	0.0575474
speechiness	1.4811769
instrumentalness	0.0002076
liveness	0.2698181
valence	0.5872390
tempo	0.0005235
duration_s	0.0007461
artistThe Front Bottoms	0.2078452

Check the Conditions

Now we need to make sure all assumptions for this linear model are met. We will first check whether the relationship between response variables and numerical features are linear.



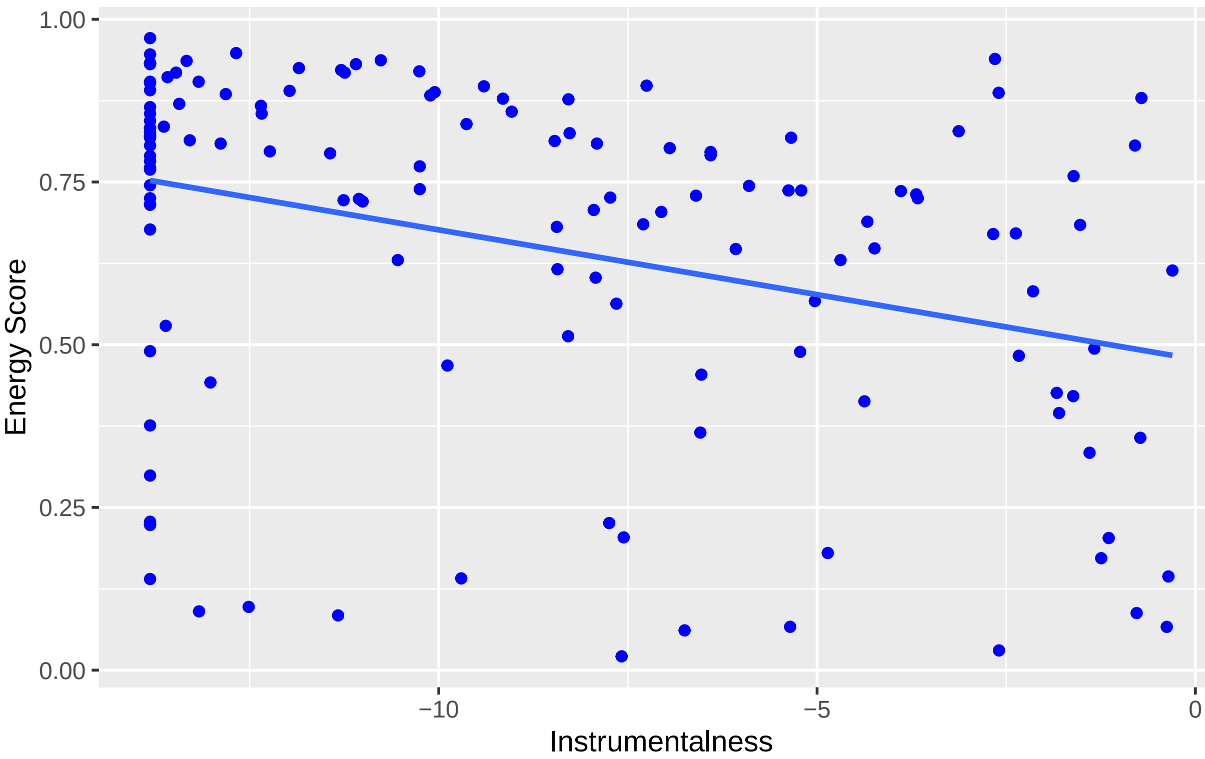
Based on Figure 1.1, we think the linear assumption does not hold between energy score and danceability



A scatter plot of speechiness versus energy score score.

From Figure 1.2, we believe that the relationship between energy score and speechiness has really weak or no linear relationship.

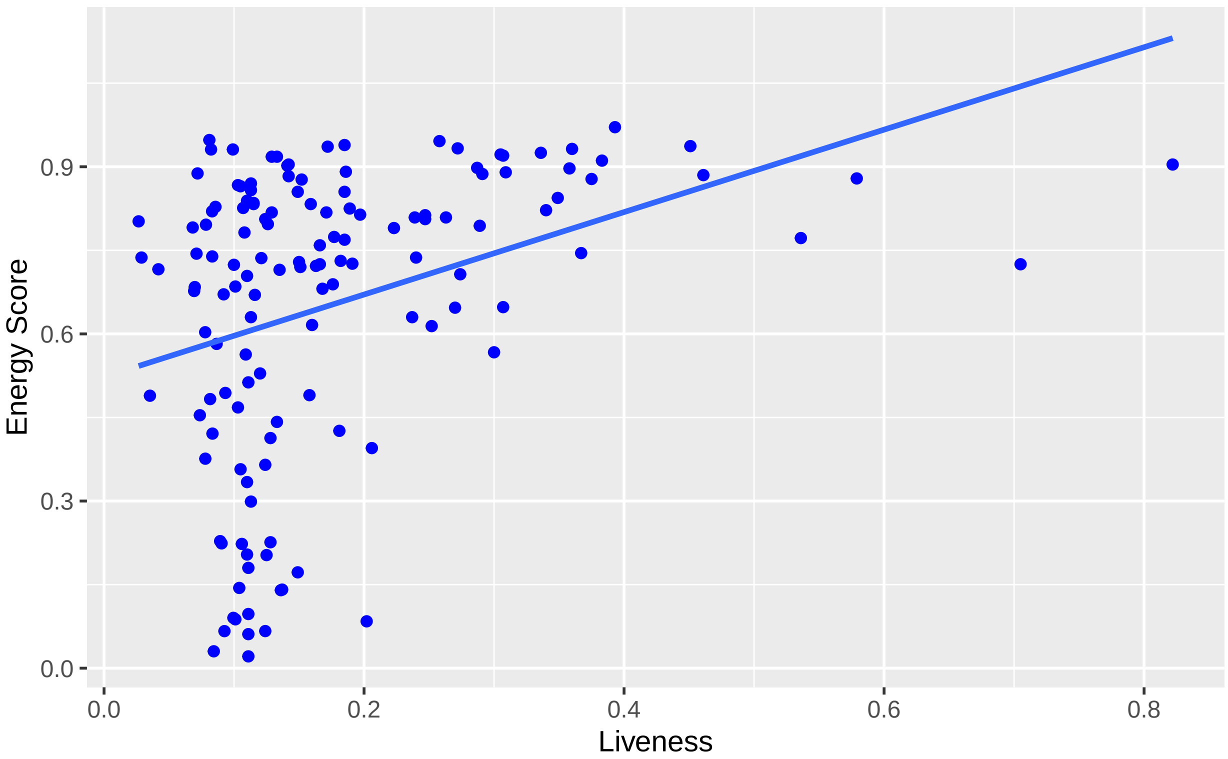
Figure 1.3



A scatter plot of instrumentalness versus energy sco

Based on Figure 1.3, we think the relationship between instrumentalness and energy score is weak on linear.

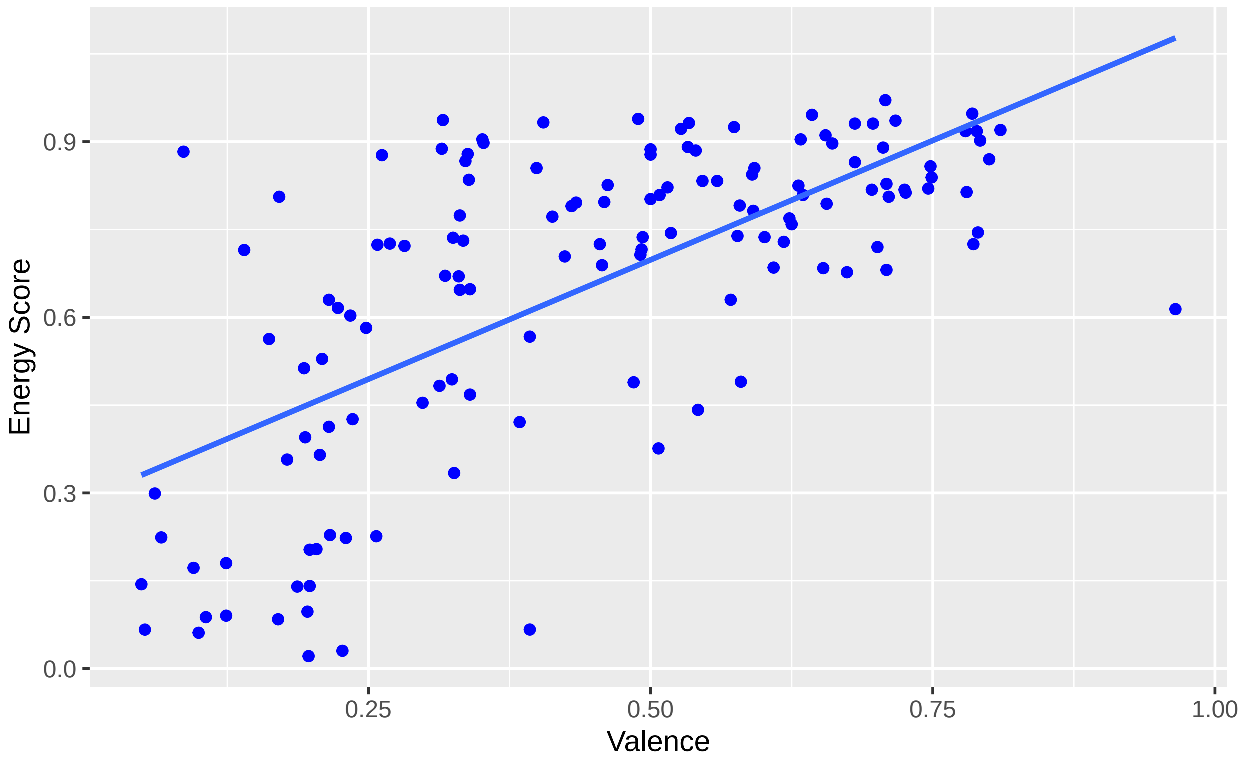
Figure 1.4



A scatter plot of liveness versus energy score

We think the linear assumption is not met for liveness and energy score based on Figure 1.4.

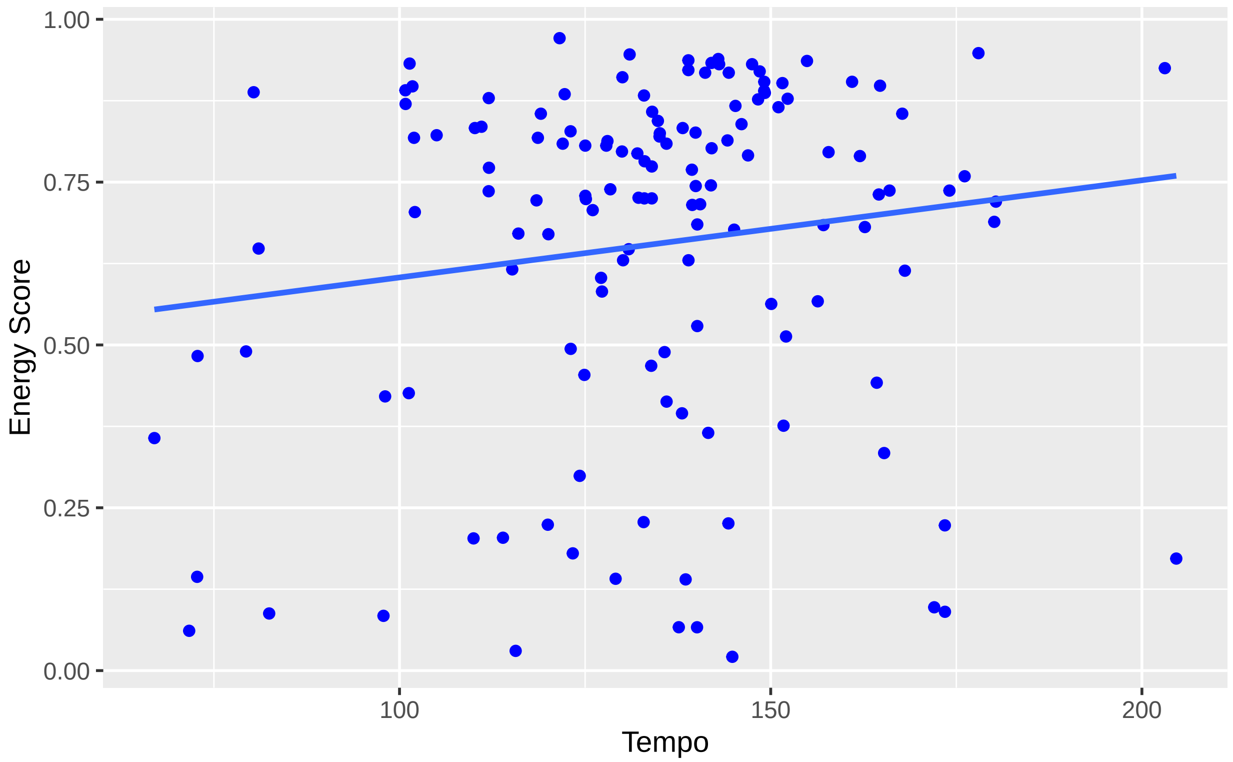
Figure 1.5



A scatter plot of valence versus energy score

From Figure 1.5, we think the relationship between energy score and valence is somehow linear.

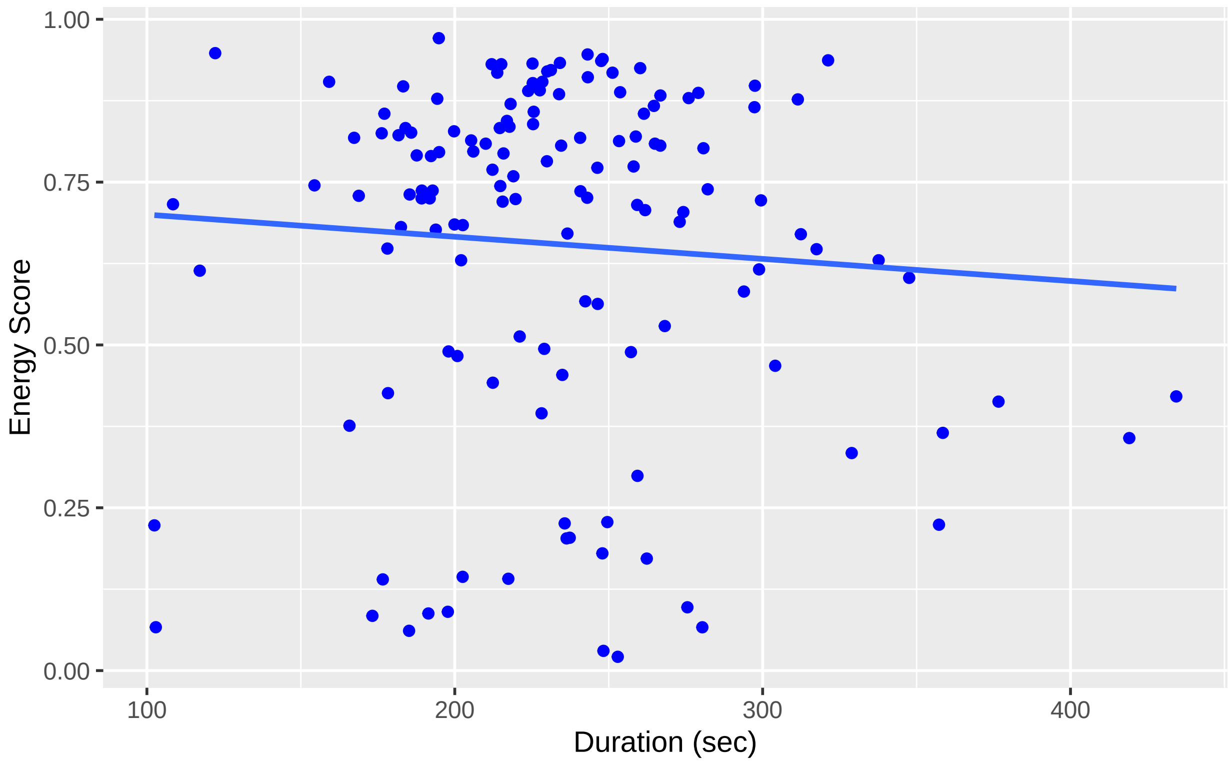
Figure 1.6



A scatter plot of tempo versus energy score

We conclude that the relationship between energy score and tempo is not linear from Figure 1.6.

Figure 1.7



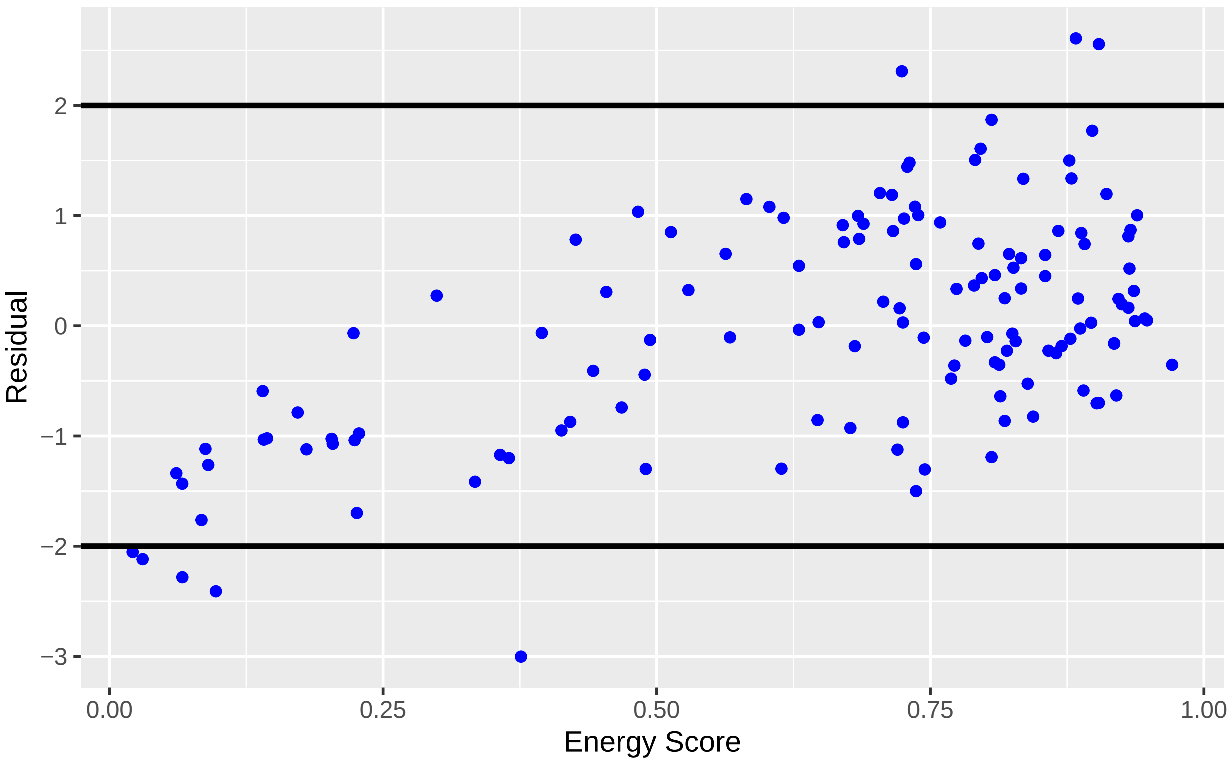
A scatter plot of duration versus energy score

Based on Figure 1.7, we think the linear assumption is not met for energy score and duration .

After checking the linearity, we will go and check whether errors are normally distributed and with mean zero and some constant variance σ. The studentized residual plot and QQ plot will help us check this.

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

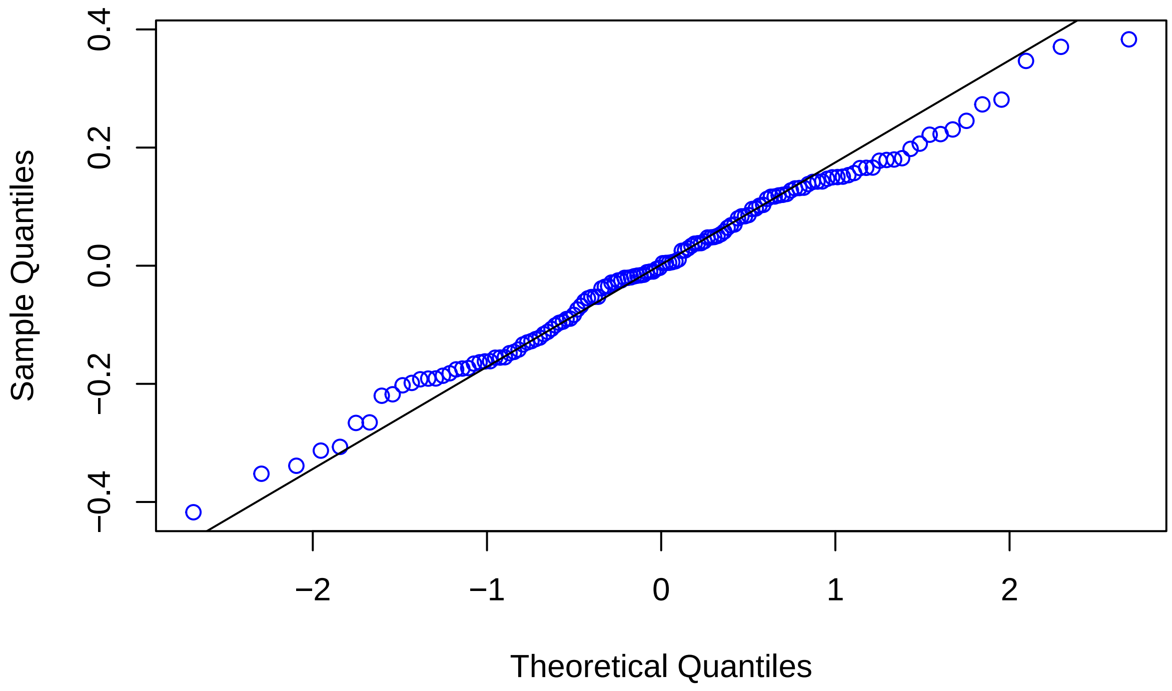
Figure 1.8



A scatter plot of the studentized residual of the model

Based on Figure 1.8, the errors are bounded by some constant variance and its mean is really close to zero.

Figure 1.9 QQPlot



Based on the QQ plot showed in Figure 1.9, we think our normality assumption looks reasonable. We also believe that the energy score is a random variable from the dataset and errors are independent. Overall, all assumptions are passed except for the linearity assumption for this model.

Coefficients Interpretation

There are some features that seem to be related to higher energy score. The first feature is keyG. It has coefficient of 0.1206. This means on average, if the song has key G, we expect the energy score to be 0.1206 higher than song has key A, keeping other features constant. The second one feature speechiness. Its coefficient is 1.4812, which means on average, with 1 unit of increase in speechiness score, we expect the energy score to increase by 1.4812. The thrid feature is liveness. Its coefficient is 0.2698, which means on average, with 1 unit of increase in liveness, we expect the energy score to increase by 0.2698. The fourth feature is valence. Its coefficient is 0.5872, which means on average, with 1 unit of increase in valence, we expect the energy score to increase by 0.5872. The last feature is artist. Given the coefficient from the model, it means on average, when artist is The Front Bottoms, we expect the energy score to be 0.2078 higher than the energy score from Manchester Orchestra when keeping other features constant. These features all seem to be related to higher energy score. These numerical features have really large positive coefficients in respect to the range of energy score. For categorical features, since their coefficients are also positive, this means the song in this category will generally has higher energy score compared to others.

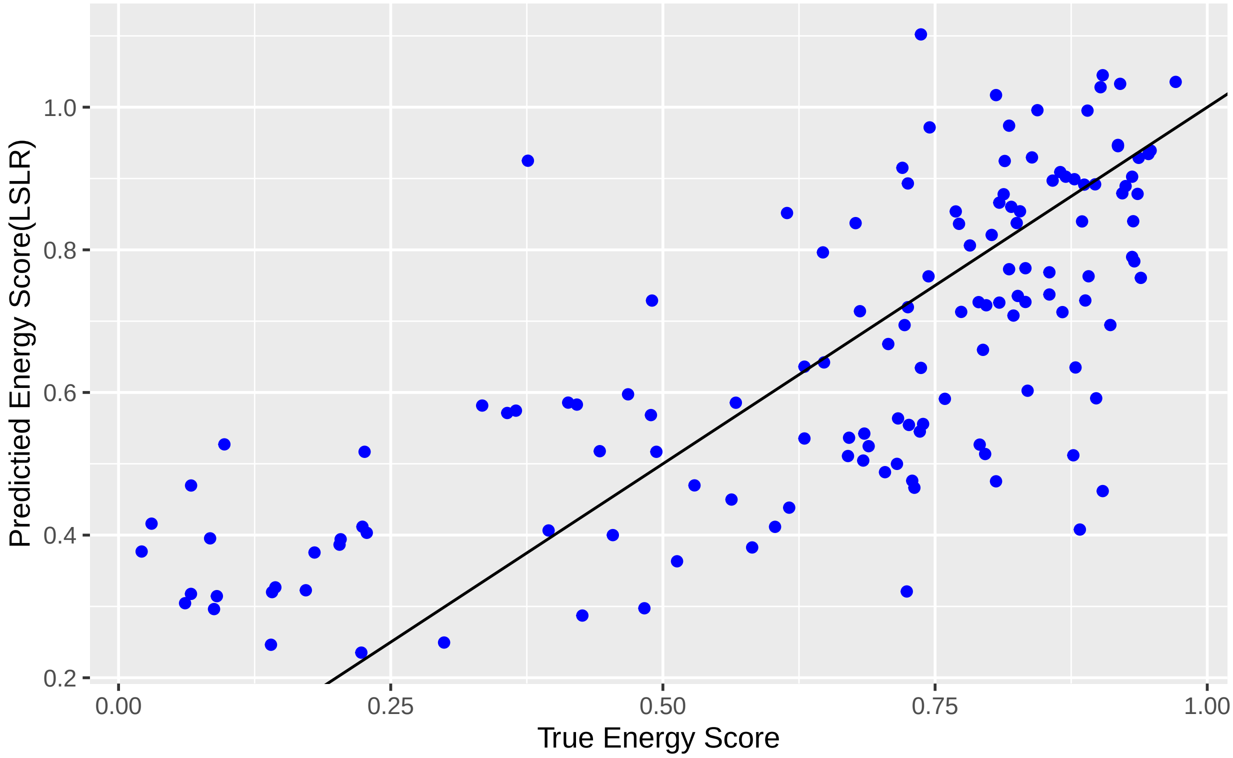
Precitive Ability

After finishing the discussion on features, we can now check the predictive abilities of our model. One way to check it is using a measurement called test RMSE. It evaluate on average, the ditance between our predicted value and true value. We will use a technique called LOOCV to create test dataset and assess model's predictive accuracy. There are some benefits of it. First, LOOCV gives an estimation which does not have high variance. It will iteratively choose one row as the test data, fit the linear model using the rest rows, and compute the test RMSE.The final test RMSE is averaging all test RMSE. You will get a single constant value no matter how many times you run. The second benefit is we do not reduce the size of our train data. In simple validation/train split, we reduce our train dataset which sometimes may make our train dataset fail to capture all properties. Since this DJ dataset is already small, we should use LOOCV in order to not reduce size of train dataset so much.

```
## [1] 0.1838391
```

Based on the result, we know that the test RMSE is 0.1838, which means on average, our predictions on energy score differ from the true energy score by ± 0.1838. The following scatter plot shows the distribution of true energy and the distribution of our predicted energy score.

Figure 1.10



A scatter plot of true energy score and predicted energy score using linear regression

If our model predict well, we will see most points lie on or around the black line. Based on Figure 1.10, we can see that our model did not predict energy score well. There is a linear relationship between our predicted energy score and true energy score. However, our model overestimates the energy score if true energy score is smaller than around 0.4. When true energy score is over 0.4, our model will either overestimate and underestimate the true energy score by a lot. There are also some predicted score that is really far away from the true score.

Section 2: KNN

Given that linear regression technique did not do a good job predicting the energy score. In this section, we will try to use K-nearest neighbors(KNN) as an alternative technique and see if it has better predictive abilities.

Table 2: Different K values and its test RMSE values(KNN)

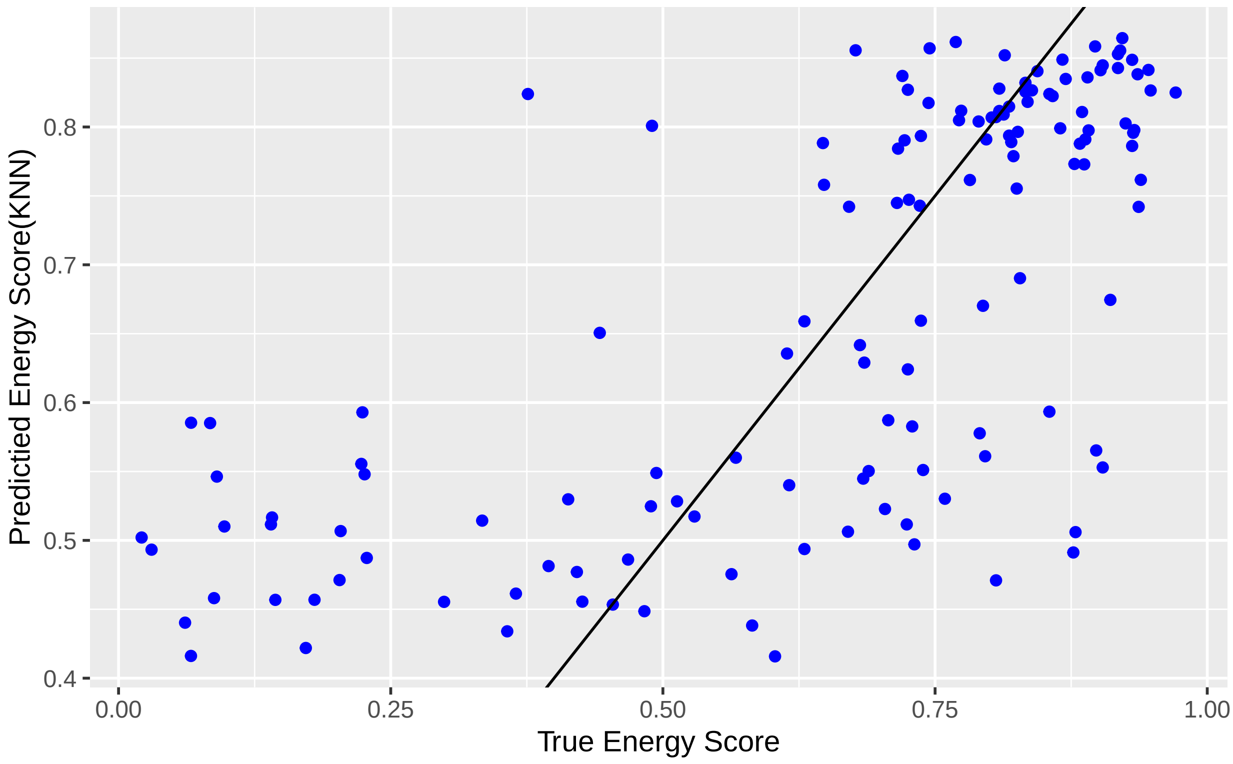
K	Test RMSE
1	0.2685676
2	0.2264923
3	0.2137717
4	0.2149838
5	0.2161196
6	0.2122496
7	0.2066729
8	0.2067246
9	0.2062958
10	0.2023463
11	0.1997435
12	0.1949585
13	0.1929218
14	0.1941113
15	0.1930889
16	0.1932708
17	0.1930870
18	0.1907199
19	0.1894056
20	0.1912104
21	0.1903476
22	0.1887954
23	0.1878530
24	0.1889308
25	0.1897386
26	0.1908416
27	0.1911932
28	0.1920592
29	0.1912416
30	0.1915144

Table 3: Smallest test RMSE and corresponding K

	K	Test RMSE
23	23	0.187853

If we choose K = 23, we will have the smallest test RMSE of 0.1878. This means on average, our predicted scores using KNN differ from the true value by ± 0.1878 Based on the table, I would suggest using K = 23 for the KNN approach. It has the smallest test RMSE value which means the predictive ability when K = 23 is the best. Usually we will choose $K = \sqrt{n}$ as the default K value. In this dataset, we can treat default K as $\sqrt{138} = 11.7473 = 12$. The test RMSE value of this K is 0.1950. I will not recommend using this K = 12. Because its test RMSE is not small enough meaning our predicted scores will have larger difference from the true score. The following scatter plot is showing predicted energy score and true energy score under KNN approach.

Figure 2.1



A scatter plot of true energy score and predicted energy score using KNN

If our predictions are good, most points should lie around the black line. From Figure 2.1, we will say if true energy score is below 0.5, our KNN always overestimate the true score. After this point, our predicted score sometimes underestimate the true value. We also have some predicted scores that are way bigger or smaller than the true score. we think there is a really weak linear relation.

Section 3: Weighted KNN

This section will focus on another approach called weighted KNN. It is a slightly variation of KNN which this time we will not just average all nearest neighbors Y value equally but do a weighted average based on how close each neighbor is. We will perform the same procedure as in section 2, but change unweighted KNN to weighted KNN.

```
## Warning: package 'kknn' was built under R version 4.3.3
```

Table 4: Different K values and its test RMSE values(weighted KNN)

K	Test RMSE
1	0.2720326
2	0.2499477
3	0.2289488
4	0.2145829
5	0.2056826
6	0.2002270

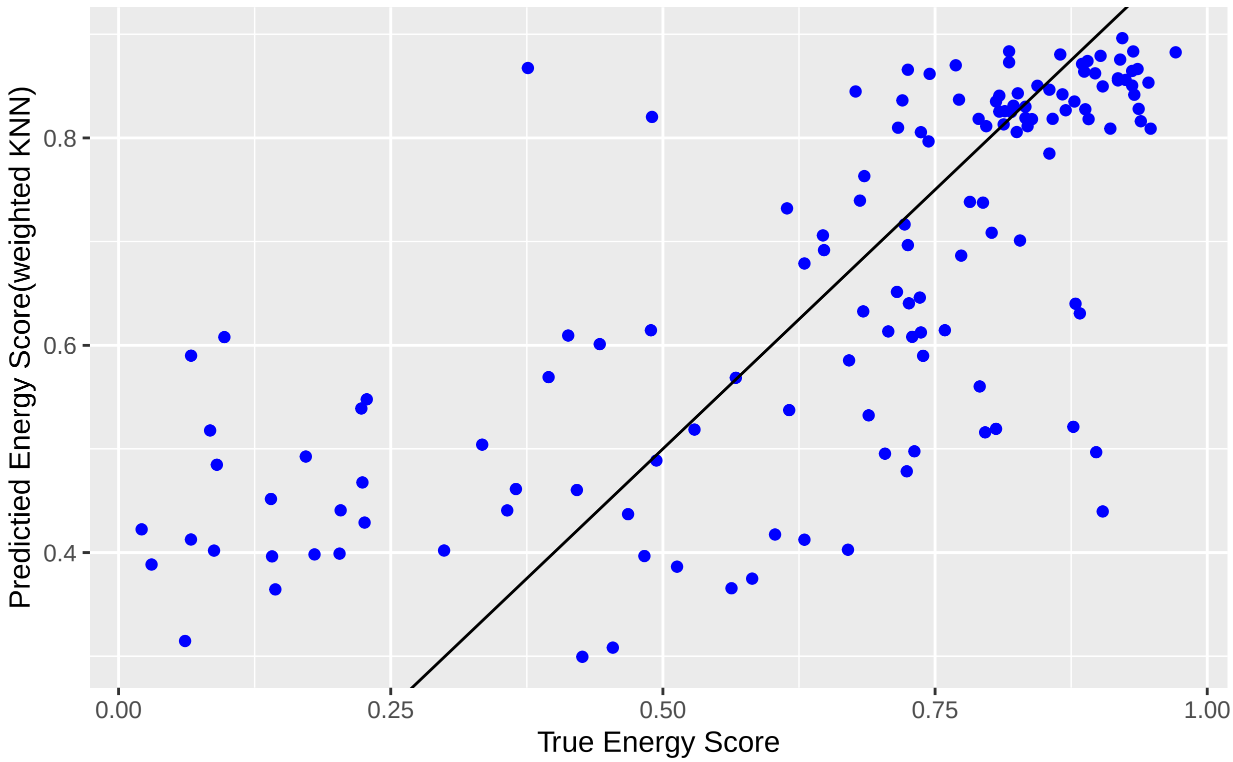
K	Test RMSE
7	0.1965015
8	0.1934806
9	0.1908898
10	0.1887057
11	0.1869074
12	0.1854472
13	0.1842925
14	0.1834622
15	0.1828888
16	0.1824139
17	0.1819914
18	0.1816465
19	0.1813340
20	0.1810281
21	0.1808263
22	0.1806866
23	0.1805390
24	0.1804282
25	0.1804052
26	0.1804244
27	0.1804241
28	0.1804095
29	0.1803927
30	0.1803767

Table 5: Smallest test RMSE and corresponding K

	K	Test RMSE
30	30	0.1803767

As mentioned in section 2, the default choice of K is 12. The RMSE value when k = 12 is 0.1854. However, the smallest RMSE is 0.1804. If we choose 12, our predicted scores will have larger differences between the true energy scores. I believe that 12 is not a reasonable choice for weighted KNN.

Figure 3.1



A scatter plot of true energy score and predicted energy score using weighted KNN

Same, if this approach has a really good predictive ability, most of the points will lie on or around the black line. Based on Figure 3.1, we found out that when true energy score is below about 0.4, our predicted energy score often overestimate. However, when true energy score is above 0.75, our weighted KNN gets the predicted score relative close to the true energy score. We think there is a linear relationship between the predicted score and true score. Therefore, the predictive ability for weighted KNN in this dataset is better than KNN and linear regression based on both RMSE and scatter plot.

Section 4: Conclusion

Based on the trials, we will recommend using weighted KNN for best predictions on energy score. Based on our measure of predictive ability of an approach, weighted KNN has the smallest RSME of 0.1804. This means the predicted energy score using weighted KNN, on average, is closest to the true score. However, even though weighted KNN has the best prediction, it does not mean that the quality of predictions is good. Taking a closer look at the true energy score, we know that the range of true energy score is [0.0212, 0.9710]. The first quartile is 0.4910, the third quartile is 0.8550, and the median is 0.7370. Based on this information, 0.1804 is too large for this dataset. Our predicted score can easily overestimate or underestimate the true score much more than 25% of the whole data, meaning our predicted score may not be very accurate. To sum up, even weighted KNN approach has the best predictive ability, we still do not think this model can estimate energy score reasonably well. We do not think this technique worth implementing.