Raven, Baldwin

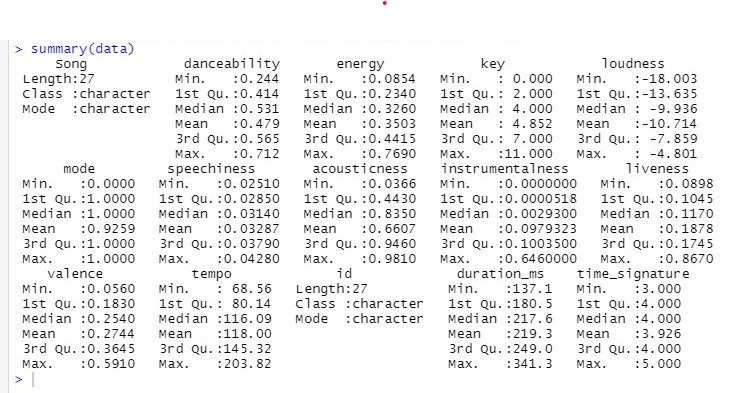
Tim Keaton

STAT 355

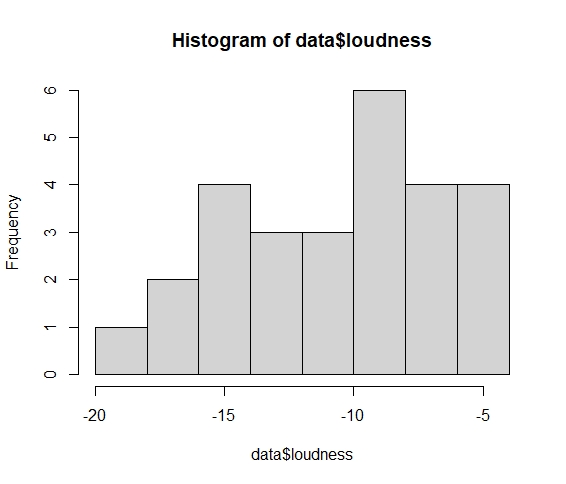
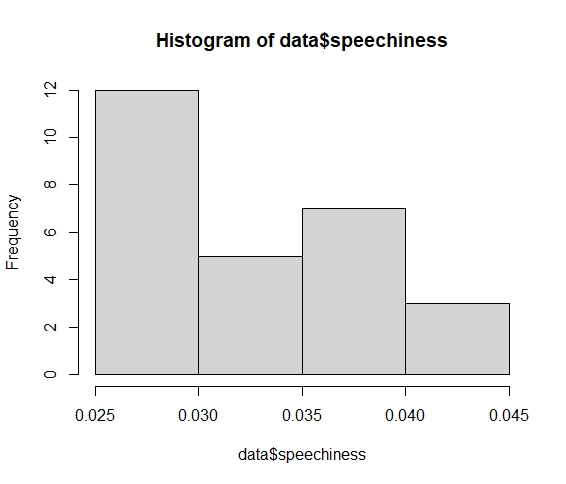
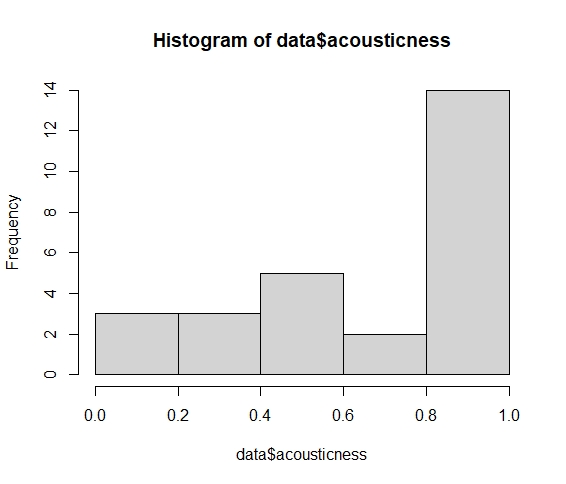
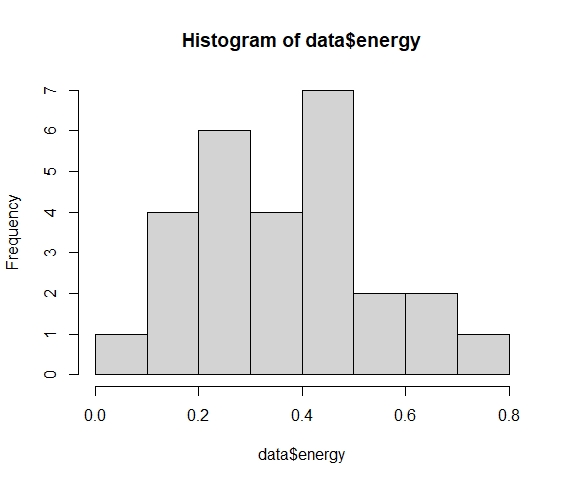
11.16.2021

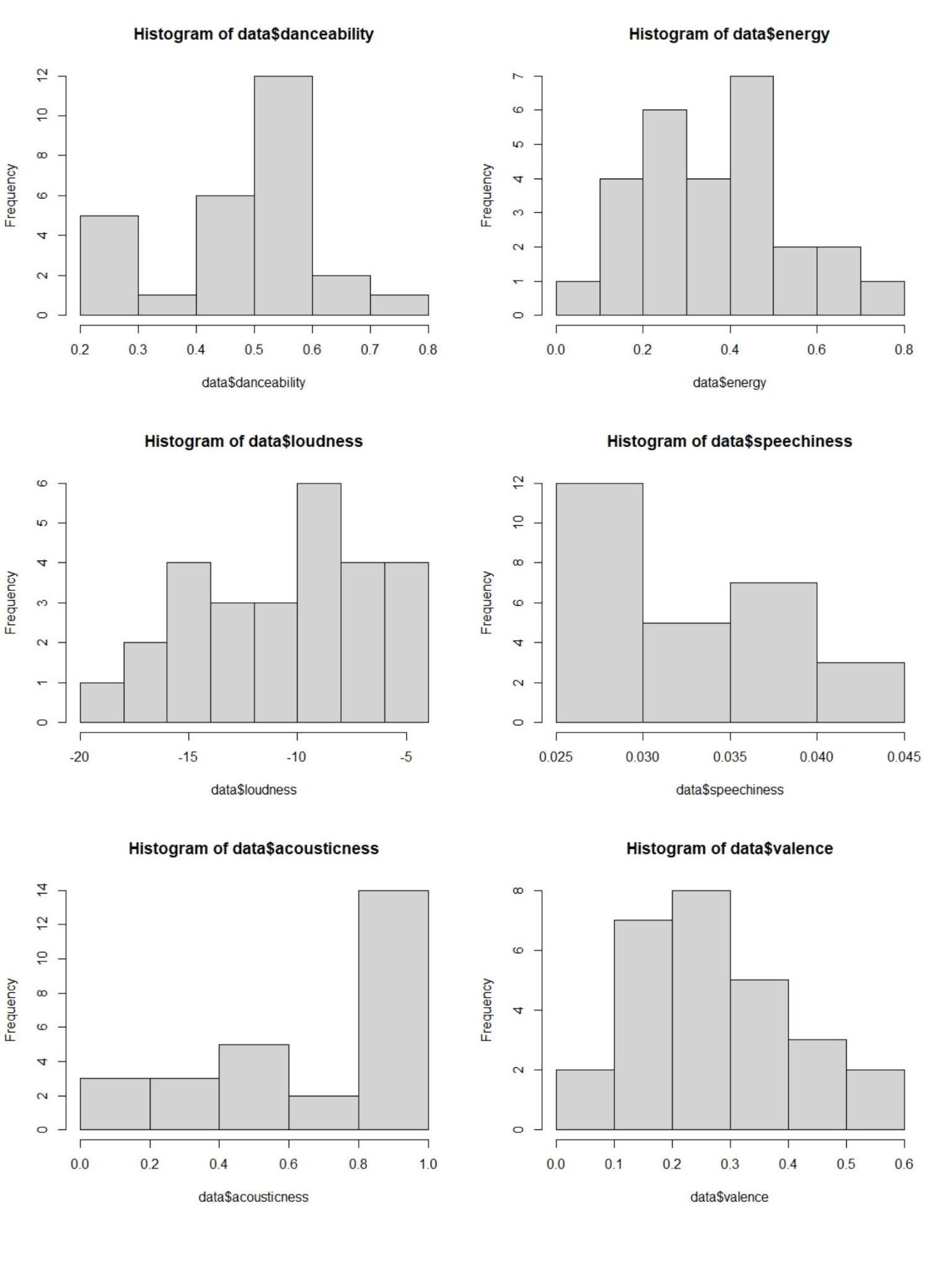
Final Project

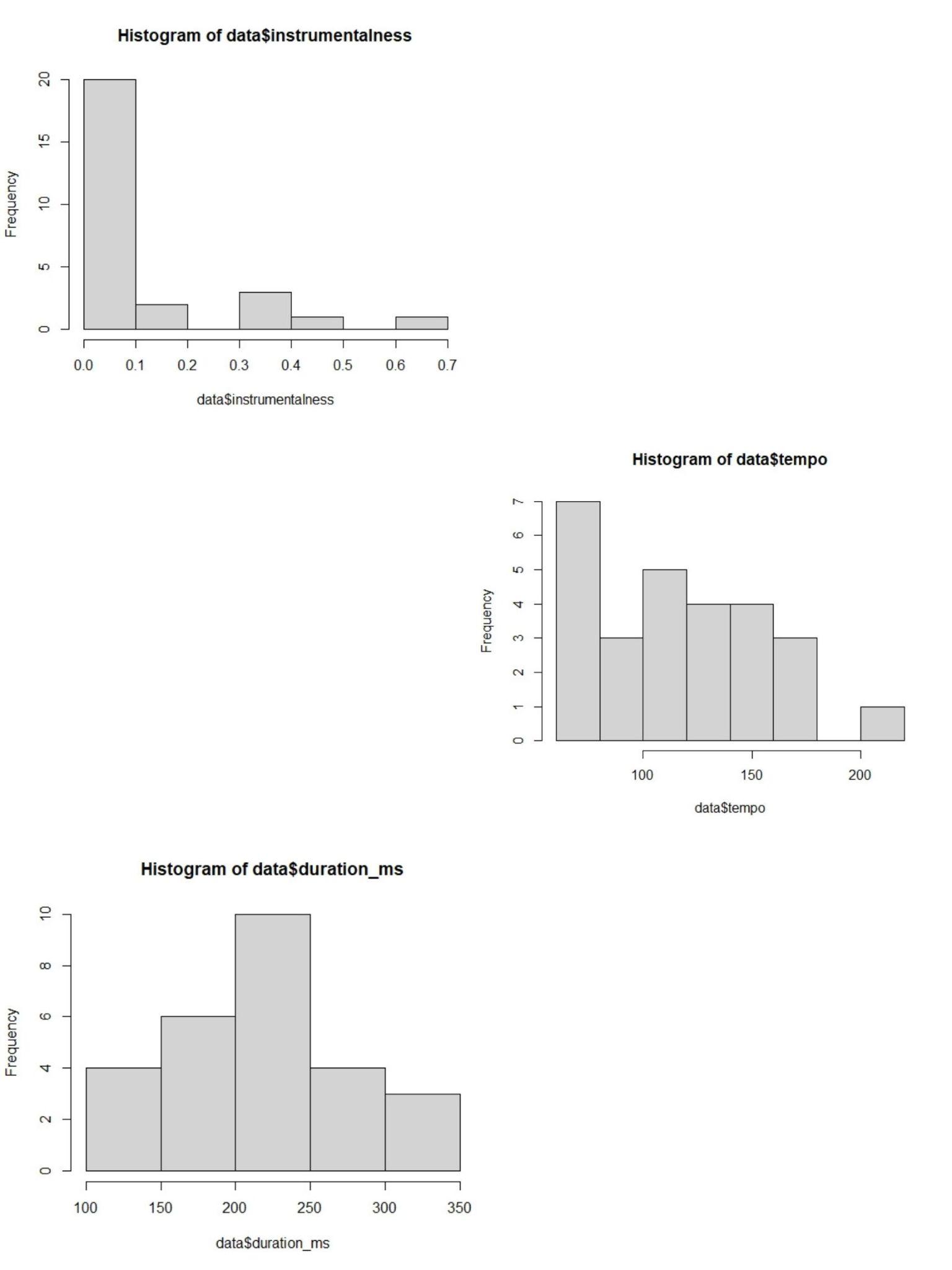
For my data set I have selected the song characteristics data from one of my playlists, this functions as a sample of all songs I like(about over 4000) songs. There are several variables: danceability, acousticness, instrumentalness, liveness, valence, energy, tempo, speechiness, duration in milliseconds(which I have converted to seconds), and key. All measurements are between 1 and 0 1, and a higher number represents a higher level of whichever variable you are exploring. Danceability is a measurement of how danceable the song is. The acousticness is a measurement of how acoustic the song is, this means how much does the song use acoustic methods of production compared to electronic. Instrumentalness reflects the amount of vocals in the song. Liveness represents the probability the song was recorded live. Valence, unlike the other variables, lies on a range from 0 to .1 it represents the positivity level of the music. The energy represents the amount of intensity and activity in the music.The speechiness reflects the amount of words in the song, if the speechiness of a song is above 0.66, it has more words than music, a number between 0.33 and 0.66 reflects a song that may contain both music and words, and a number below 0.33 means the song does not have any speech. Key comes in a variety of numbers, these numbers represent different keys for the song such as a flat or g etc. The tempo is a measurement ranging from about 70 to 200, this measures how fast the song is. The following statistics are summary statistics for all data



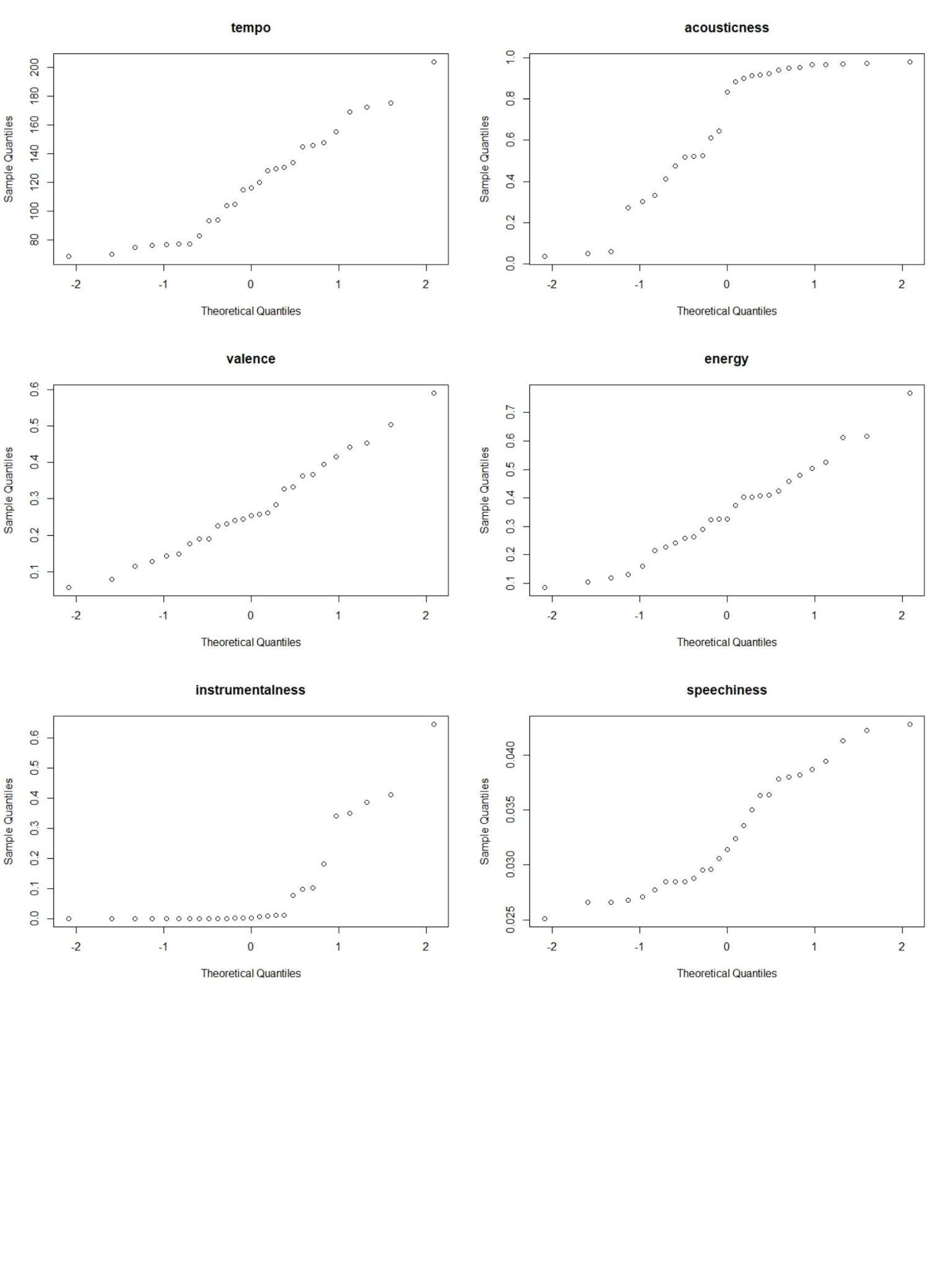
Based on the summary statistics means and medians, it would appear the songs aren’t entirely danceable(about halfway between 1 and 0), the songs are lower energy, the songs are pretty loud, not very speechy, most of the songs seem fairly acoustic, not as instrumental, most are not likely recorded live, they are not very valent either, and the songs are fairly fast. These are the histograms for most of the variables.

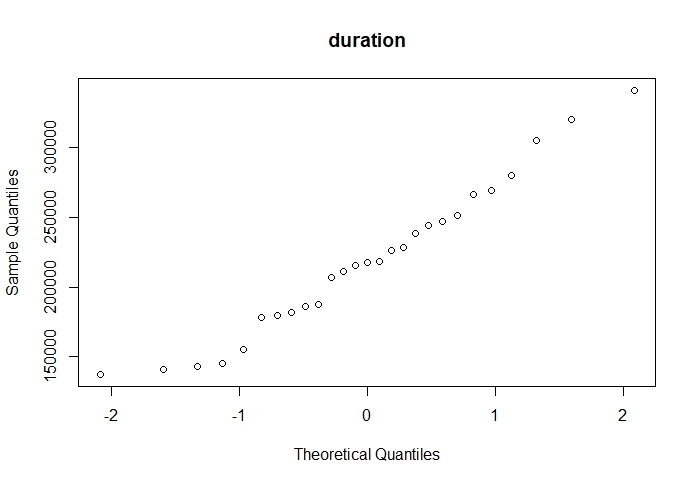
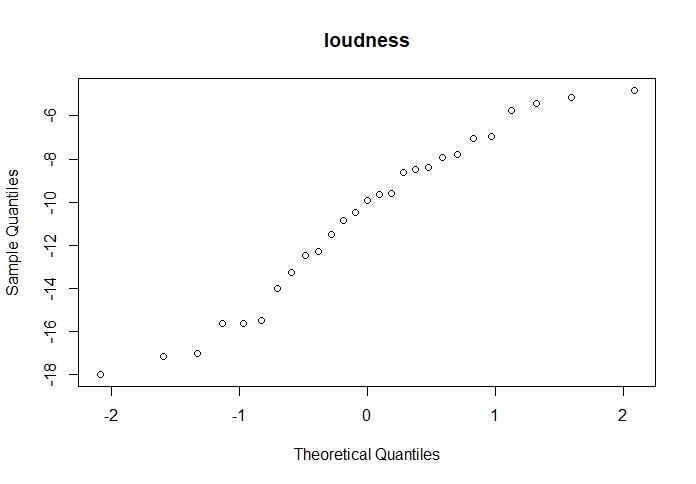






There is not a lot of normality in this data only mostly in the variables energy, duration, and valence, danceability is slightly normal without being very skewed. Acousticness is skewed right showing the songs are mostly acoustic, speechiness is skewed left showing they aren’t as speechy, instrumentalism is skewed left showing they aren’t as instrumental, tempo is also skewed left showing they aren’t as fast.



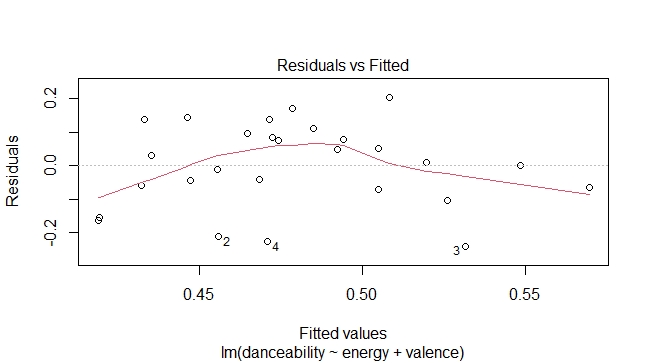


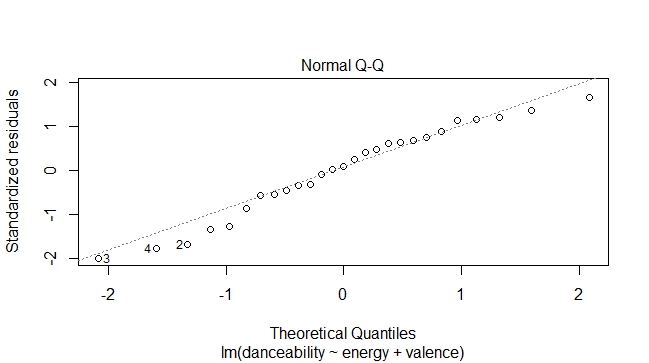
Based on the normality plots they confirm which variables I had stated are normal based on how many follow a fairly straight line.

For my research questions I would like to examine the relationship between different variables that may seem to have a relationship based on their similar/opposite descriptions using methods such as a correlation. My first question is whether or not energy and valence is a useful predictor for danceability, as most songs that are more upbeat and energetic tend to be more danceable, and I would like to see if this applies for these songs, to do this I will be using multiple regression as it reveals whether or not the variables are a good predictor for danceability. Since the speechiness indicates the amount of words in the songs, and instrumentalness indicates whether or not the song is more words than music, one might assume as the amount of words in a song increases the instrumentalness of the song goes down, and I would like to see if this is true for this set of data by checking if there is a relationship and whether or not instrumentalness is a good predictor for speechiness by using simple linear regression, correlation, and t-tests, linear regression shows whether or not the variable is a useful predictor for speechiness, correlation will reveal if there is a relationship(correlation) between the variables, the t-test will reveal whether or not there is a difference between the speechiness and instrumentalness of the songs, Anova to follow the t test to see where the difference in the means occurs by looking at the p value, and I will also look at the F test will also determine if instrumentalness is a good predictor for speechiness. Finally, I will explore the relationship between loudness and acousticness and check whether or not acousticness is a good predictor for loudness, as most songs that are more acoustic tend to be softer, and I would like to see if that is true for these songs. I will be using the same methods as with speechiness and instrumentalism for the same reasons.

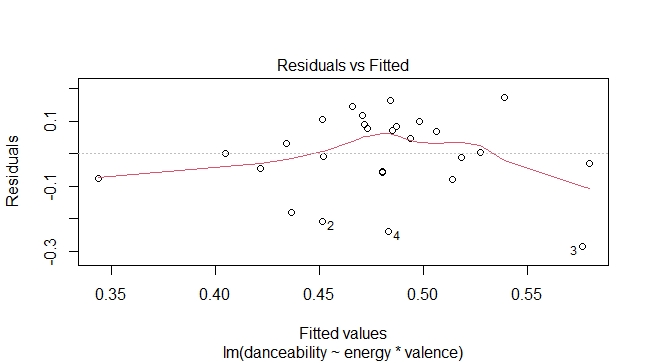
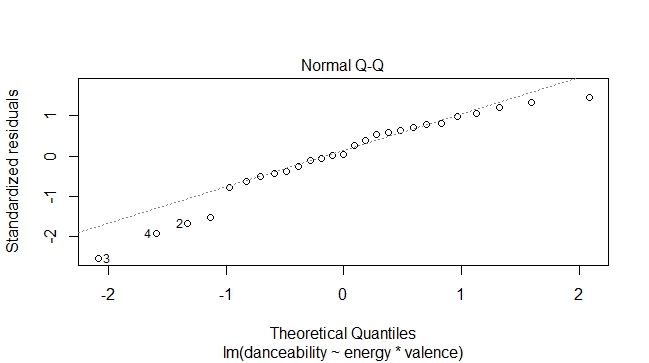
First, I will be seeing whether or not there is a relationship between energy, valence, and danceability and checking whether or not energy and valence are good predictors for danceability. To start off I created a regression model to predict danceability based on energy and valence first without interaction. Before looking at the regression models I will look at the residual and normal qq plot to check the assumptions.

Below is the plots for the model without interaction:





Below are the plots with interaction:



Looking at the plots, the residuals seem to be centered around 0, variance does seem fairly normal though the line for the model with interaction looks slightly curved and centered at the middle. The residuals do seem fairly normal though. The model with no interactions residuals appear fairly curved and not exactly centered at 0 indicating an inconsistent variance. The normal qq plots seem to follow a straight line. The songs are also independent from one another although 2 songs are by the same artist they are fairly different. There are only a few outliers and there does not seem to be a pattern in either plots. The only assumption that might not be satisfied is the linearity of the model. These models might be useful for predicting danceability. Regardless, I will still proceed to look at the data to see if it still has a relationship.

*mod1=lm(danceability~energy+valence)*

*summary(mod1)$coef*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 0.40516302 0.06823361 5.9378803 3.970401e-06*

*energy -0.02427145 0.16048056 -0.1512423 8.810487e-01*

*valence 0.29990867 0.20350356 1.4737270 1.535515e-01*

*mod2=lm(danceability~energy\*valence)*

*summary(mod2)$coef*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 0.2663508 0.1333159 1.997893 0.05768786*

*energy 0.3645659 0.3590020 1.015498 0.32043249*

*valence 0.8441700 0.4935917 1.710260 0.10067579*

*energy:valence -1.4113972 1.1683856 -1.207989 0.23932620*

For the slopes of the model, with interaction, as danceability increases by 1, the energy level decreases by .02, and the valence level increases by .3, the intercept is .41. With interaction the energy level increases by .36 and the valence level increases by .84, the intercept is .27. We can also look at how significant the predictors of energy and valence are when predicting danceability. It would seem that with and without interaction, we have insufficient evidence to conclude the null hypothesis that energy or valence is a good predictor of danceability, as their p values are both greater than an alpha of .05.. Using the summary of mod1(without interaction) and mod2(with interaction) we can look at the partial F test results to further examine if they are useful predictors.

*Model 1: danceability ~ energy + valence*

*Model 2: danceability ~ energy \* valence*

*Res.Df RSS Df Sum of Sq F Pr(>F)*

*1 24 0.40461*

*2 23 0.38047 1 0.024139 1.4592 0.2393*

Based on the partial F test we can see since the p value is greater than the alpha level of .05 we fail to reject the null hypothesis that something within the model is useful for predicting the danceability value. We can also look at the interaction and non-interaction models’ r squared to examine whether or not the variation in danceability is explained by its relationship with the other variables.

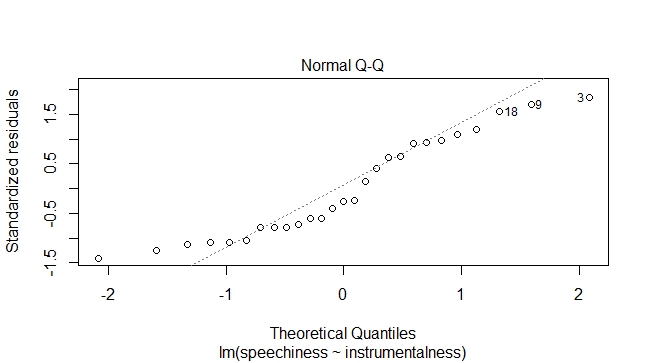
*summary(mod1)$r.squared*

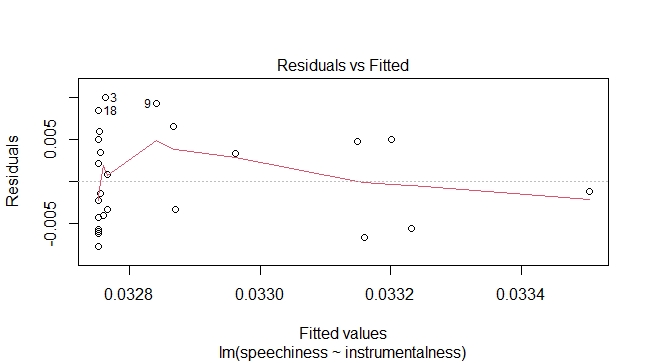
*0.08863804*

*summary(mod2)$r.squared*

*0.1430099*

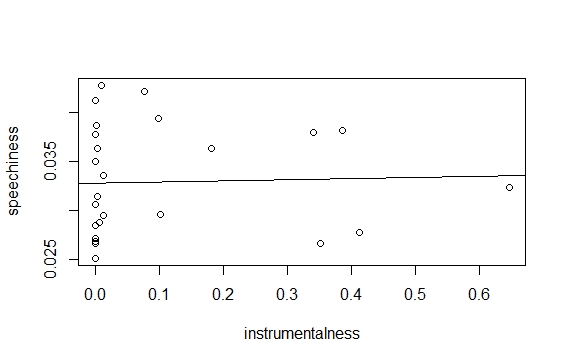
These values are 0.08863804 and 0.1430099. These values are quite small and means that only 9% and 14% which indicate the variation in danceability is explained by its relationship between energy and valence level. However, the interaction model has a higher r squared which shows the interaction model explains more of the variation. Based on the data, it would appear that both energy and valence have little to no effect on the danceability level. This also does not confirm that valence and energy are good predictors for danceability. We can see this through the results of the ANOVA and the regression model summaries.

Now I will be checking to see if there is a relationship between speechiness and instrumentalism, and checking whether or not instrumentalism is a good predictor for speechiness, using the methods of simple linear regression, correlation, and t-tests. Before continuing to look at the linear regression model I will be checking the assumptions.



Based on the plots we can see that the normal qq plot reveals that the data curves and does not follow a straight line, indicating it isn’t normal. The residuals are not centered at 0 and they have a pattern and the points are all concentrated at the far left side of the plot and the line is slanted, noting the variance is inconsistent. There are a few drastic outliers however. Again, the songs are fairly independent from each other. The only assumption that might not be satisfied is the linearity of the model and the outliers may impact the results; the plot reflected below shows that.

These indicate the model will most likely be unreliable for predicting speechiness. To start, I made a linear regression model including speechiness and instrumentalism and plotted it.

*summary(mod3)*

*Call:*

*lm(formula = speechiness ~ instrumentalness)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-0.007653 -0.004253 -0.001356 0.004925 0.010035*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 0.032753 0.001241 26.384 <2e-16 \*\*\**

*instrumentalness 0.001161 0.006341 0.183 0.856*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 0.005585 on 25 degrees of freedom*

*Multiple R-squared: 0.001338, Adjusted R-squared: -0.03861*

*F-statistic: 0.0335 on 1 and 25 DF, p-value: 0.8563*

Above is the summary for the model. This indicates the regression equation is y=.033 +.0011x. This indicates that for every 1 increase in instrumentalness the speechiness increases from .033 by .0011. The F test has a p value of .8563 which is greater than our p value of .05 which means that something in this model is not at all useful for predicting speechiness based on instrumentalness. Next, I wanted to look at their correlation

*cor.test(speechiness, instrumentalness)*

*Pearson's product-moment correlation*

*data: speechiness and instrumentalness*

*t = 0.18303, df = 25, p-value = 0.8563*

*alternative hypothesis: true correlation is not equal to 0*

*95 percent confidence interval:*

*-0.3482741 0.4108835*

*sample estimates:*

*cor*

*0.03658135*

The correlation is weak and positive at only .04, and at an alpha of .05, with a p value of .86 we fail to reject the null hypothesis that the correlation is equal to 0, as we have insufficient evidence. There is almost no correlation between the variables. Next, I will be looking at the test comparing the means of the variables to see if there is a significant difference.

*t.test(speechiness,instrumentalness)*

*Welch Two Sample t-test*

*data: speechiness and instrumentalness*

*t = -1.9562, df = 26.052, p-value = 0.06124*

*alternative hypothesis: true difference in means is not equal to 0*

*95 percent confidence interval*

*-0.13342805 0.00329669*

*sample estimates:*

*mean of x mean of y*

*0.03286667 0.09793235*

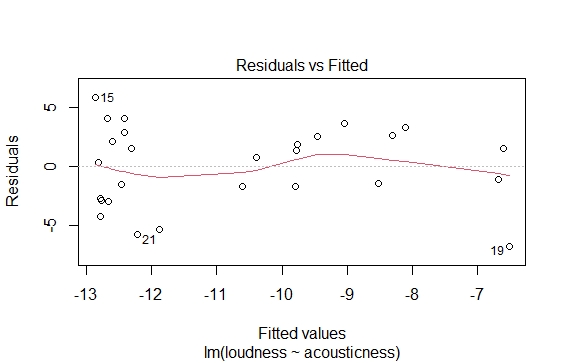
Based on the t-test we fail to reject the null hypothesis and have insufficient evidence to conclude that the difference in the means are equal to 0, our p value of .06 is greater than an alpha level of .05. The model’s r squared is shown below:

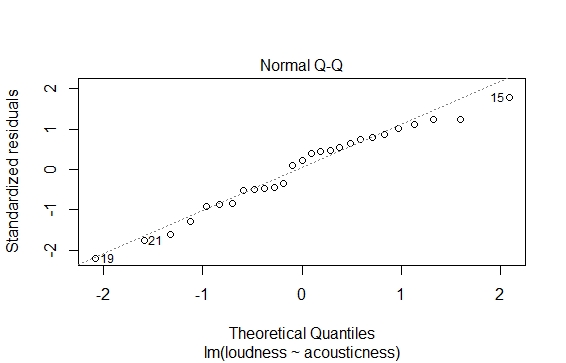
*summary(mod3)$r.squared*

*0.001338195*

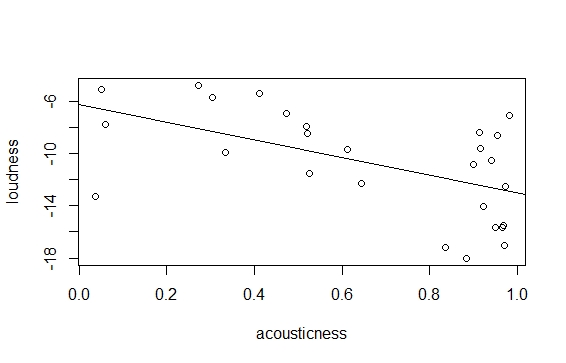
This indicates only .13% of the variation in speechiness is explained by it’s relationship with instrumentalness.

Once again, these statistics show there is almost no relationship between these variables, and the regression model is not at all useful for predicting speechiness based on instrumentalness. This does not confirm that instrumentalness is a good predictor for speechiness.

Now I will be checking whether or not there is a relationship between loudness and acousticness, and checking whether or not acousticness is a good predictor for loudness, using the methods used previously for speechiness and instrumentalism of simple linear regression, correlation,and t-tests. Before continuing to look at the linear regression model I will be checking the assumptions. 



Based on the plots we can see that the normal qq plot reveals that the model is fairly normal as it somewhat follows a straight line. The residuals are not centered at 0 and they do not have a pattern, the line is slightly curved, this. Once again, the songs are mostly independent for the model. The only assumption that might not be satisfied is the linearity of the model and the outliers might affect the results; the plot reflected below shows that. These indicate the model might be reliable for predicting speechiness. To start, I made a linear regression model including speechiness and instrumentalism and plotted it.



*summary(mod4)*

*Call:*

*lm(formula = loudness ~ acousticness)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-6.7390 -2.2069 0.7208 2.5320 5.8055*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) -6.267 1.501 -4.175 0.000315 \*\*\**

*acousticness -6.731 2.047 -3.289 0.002989 \*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 3.383 on 25 degrees of freedom*

*Multiple R-squared: 0.302, Adjusted R-squared: 0.274*

*F-statistic: 10.81 on 1 and 25 DF, p-value: 0.002989*

Above is the summary for the model. This indicates the regression equation is y=-6.27 -6.73x. This indicates that for every 1 increase in acousticness the loudness decreases from -6.27 by 6.73. The F test has a p value of .002989 which is less than our p value of .05 which means something in the model is useful for predicting loudness based on acousticness. Next, I wanted to look at their correlation

*cor.test(loudness, acousticness)*

*Pearson's product-moment correlation*

*data: loudness and acousticness*

*t = -3.2885, df = 25, p-value = 0.002989*

*alternative hypothesis: true correlation is not equal to 0*

*95 percent confidence interval:*

*-0.7689494 -0.2142297*

*sample estimates:*

*cor*

*-0.5495081*

The correlation is mediocre and negative at -.55. and at an alpha of .05 with a p value of .003 we reject the null hypothesis that the correlation is equal to 0, as we have sufficient evidence. There is some correlation between the variables. Next, I will be looking at the test comparing the means of the variables to see if there is a significant difference.

*t.test(loudness, acousticness)*

*Welch Two Sample t-test*

*data: loudness and acousticness*

*t = -14.836, df = 26.347, p-value = 2.624e-14*

*alternative hypothesis: true difference in means is not equal to 0*

*95 percent confidence interval:*

*-12.949227 -9.799351*

*sample estimates:*

*mean of x mean of y*

*-10.7135926 0.6606963*

Based on the t-test we reject the null hypothesis and have sufficient evidence to conclude that the difference in the means is not equal to 0, as our p value of *2.624e-14* is less than an alpha level of .05. The model’s r squared is shown below:

*summary(mod4)$r.squared*

*0.3019591*

This indicates 30% of the variation in loudness is explained by it’s relationship with acousticness. Finally, I will be looking at a partial F test for the model to determine whether or not these variables are useful predictors.

*anova(mod4)*

*Analysis of Variance Table*

*Response: loudness*

*Df Sum Sq Mean Sq F value Pr(>F)*

*acousticness 1 123.78 123.776 10.815 0.002989 \*\**

*Residuals 25 286.13 11.445*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

The p value is .003 which is less than our alpha of .05 which means we have sufficient evidence to reject the null hypothesis that something in this model is useful for predicting speechiness based on instrumentalness. Hoorah! Finally, a model that is useful for prediction. Acousticness does seem to be a good predictor for loudness.

To conclude, I will be summarizing my findings. When looking at the relationship between danceability, valence, and energy, there was not much of a relationship, and valence and energy had almost no impact on the danceability level. When looking at speechiness and instrumentalness the findings were the same. The models used to explore the relationship between the previously stated variables were not useful for predicticting the danceability or speechiness value. Looking at loudness and acousticness however, there was a relationship between the variables, and acousticness was useful for predicting loudness. Overall, I was hoping to find more of a relationship between the variables but was disappointed until looking at acousticness and loudness.