You're On Aux: Exploring the Statistical Relationships Between Musical Features & Song Popularity

The ANOVA Avengers

2023-05-06

Contents

1 Introduction

Hey, you got Aux?

One of the most anxiety inducing questions one can experience. You have the impossible task to pick a song that not only fits the mood, but also leaves everyone happy and satisfied. Conversely, imagine you're an upcoming musician. How should you go about picking a song that maximizes plays and increases your reach? According to the Bureau of Labor Statistics, the top 1% of artists earn 77% of recorded music income. Therefore, having a tool that enables musicians to increase the chances of their song's success would be extremely valuable. In this paper, we **explore** what variables or features that best correlate with danceability—a spotify created statistic related to how "danceable" a song is-as well as total streams. Furthermore, we **identify strategies** that help 'aux controllers' and rising musicians maximize their chances of satisfying their respective audience!

2 Data

In our analysis we use a dataset of the top 100 songs in spotify. The relevant columns we used are:

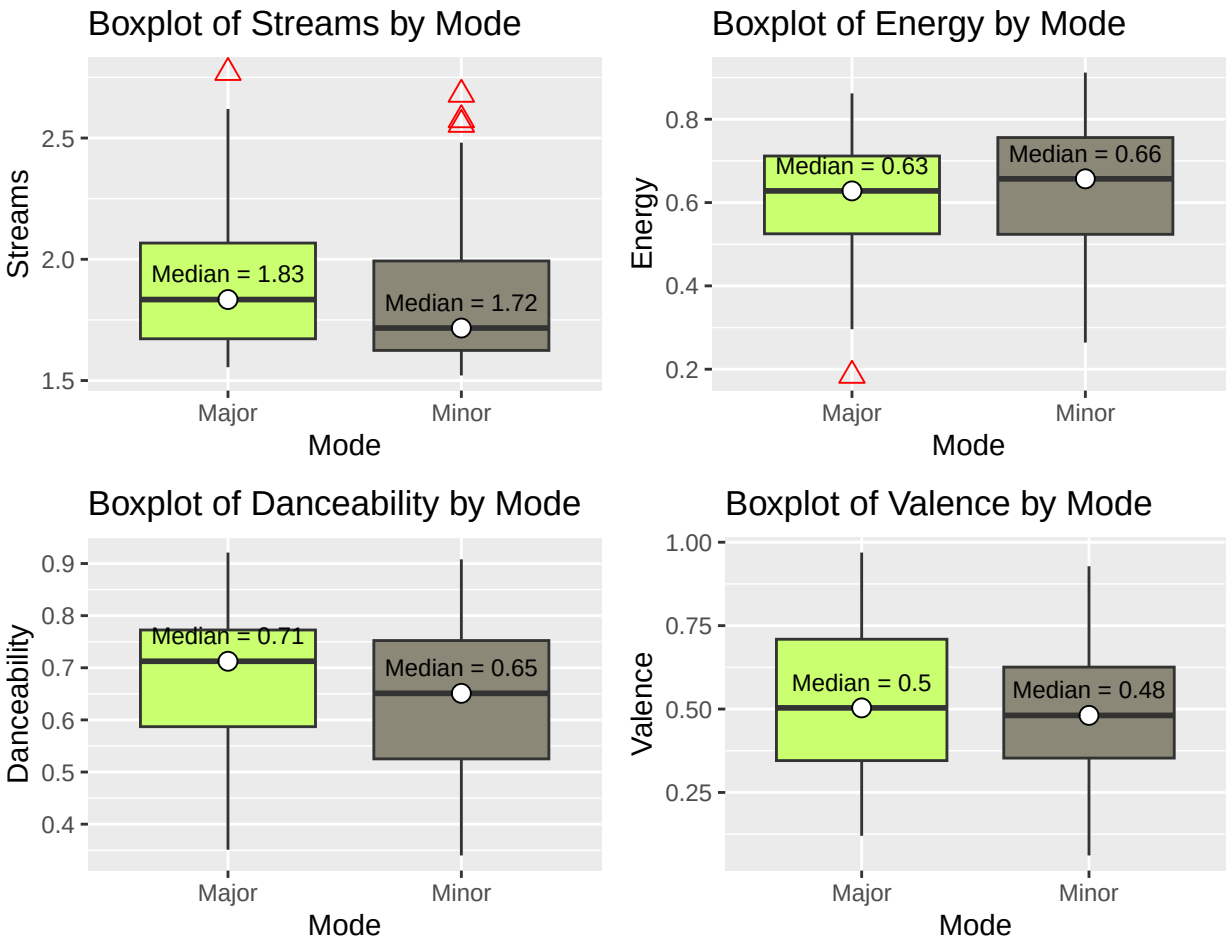
- **Name**: The name of the song
- **Genre** (Categorical): What genre of song it's classified as
- **Duration** (Continuous): Measures the duration of songs in a minute
- **Energy** (Continuous): Represents the perceptual measure of intensity and activity associated from the song on a scale from 0.0 to 1.0
- **Key** (Categorical): What musical key scale the song is written in
- **Loudness** (Continuous): The overall loudness of a track in decibels (dB)
- **Mode** (Categorical): Describes if the song is written in major or minor scale
- **Speechiness** (Continuous): Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk) is, the higher the value is on a scale from 0-1
- **Liveness** (Continuous): A calculated statistic that describes the probability that the song was recorded with a live audience., and higher liveness values (in our case on a scale from 0-1)
- **Valence** (Continuous): Describes the musical positiveness conveyed by a track on a scale from 0-1. 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** (Continuous): The speed or pace of a given piece
- **Danceability** (Continuous): Danceability describes how suitable a track is for dancing. A value of 0.0 is least danceable and 1.0 is most danceable
- **Streams** (Continuous): The number of times the song has been listened to in billions
- **ReleaseDate, ReleaseYear, and ReleaseMonth** (Categorical): Describes the specific full date(mm/dd/yy), specific year and specific month respectively that the track was released
- **Season** (Categorical): What season the track was released. We use the season of fall/winter and spring/summer

3 Data Cleaning

This R code performs data **cleaning** and **transformation** operations on a dataset named "songs". The code first removes the ID column and standardizes column names to lowercase. The code changes the date format of the "releasedate" column, **creates** a new "releaseyear" column, and **subsets** the data to include only songs with less than 3 billion streams. The code replaces 1's and 0's in the "mode" column with "Minor" and "Major", respectively. The code also **formats** data in the "acousticness" and "instrumentalness" columns using the sprintf function. Numeric values in the "key" column are replaced with musical key names. The code creates two new columns, "releasemonth" and "season", and **removes** rows with missing values.

4 Descriptive Plots & Summary Information

4.1 Boxplots / T-Tests




```
(test1 <- t.test(songs$streams ~ songs$mode, conf.level = 0.95))
```

```
    Welch Two Sample t-test

data:  songs$streams by songs$mode
t = 1.2675, df = 64.091, p-value = 0.2096
alternative hypothesis: true difference in means between group Major and group Minor is not equal to 0
95 percent confidence interval:
 -0.04758035  0.21276601
sample estimates:
mean in group Major mean in group Minor
           1.918765            1.836172
```

```
(test4 <- t.test(songs$danceability ~ songs$mode, conf.level = 0.95))
```

```
    Welch Two Sample t-test

data:  songs$danceability by songs$mode
t = 1.6161, df = 66.335, p-value = 0.1108
alternative hypothesis: true difference in means between group Major and group Minor is not equal to 0
```

```
95 percent confidence interval:
 -0.01161792  0.11037159
sample estimates:
mean in group Major mean in group Minor
          0.6814706           0.6320938
```

```
(test6 <- t.test(songs$energy ~ songs$mode, conf.level = 0.95))
```

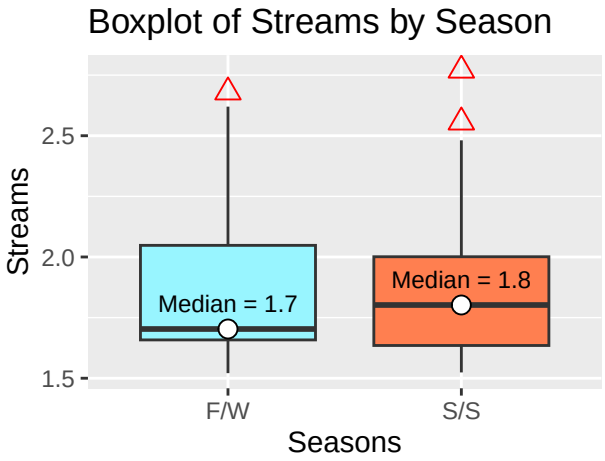
```
    Welch Two Sample t-test

data:  songs$energy by songs$mode
t = -0.90767, df = 63.68, p-value = 0.3675
alternative hypothesis: true difference in means between group Major and group Minor is not equal to 0
95 percent confidence interval:
 -0.10157509  0.03811369
sample estimates:
mean in group Major mean in group Minor
          0.6034412           0.6351719
```

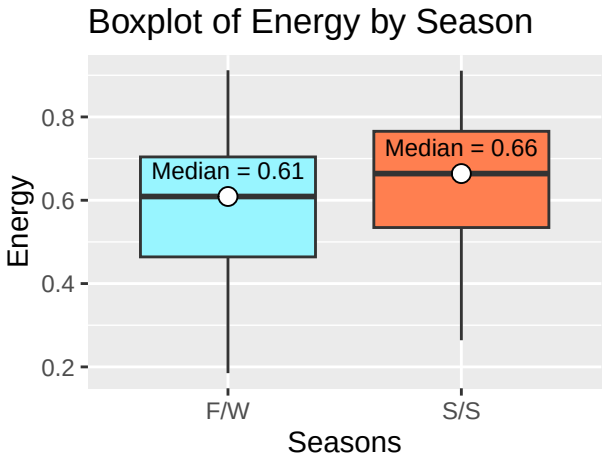
```
(test8 <- t.test(songs$valence ~ songs$mode, conf.level = 0.95))
```

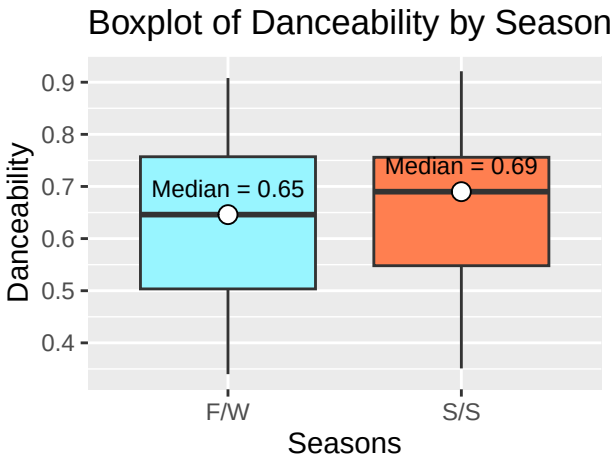
```
    Welch Two Sample t-test

data:  songs$valence by songs$mode
t = 0.45971, df = 59.93, p-value = 0.6474
alternative hypothesis: true difference in means between group Major and group Minor is not equal to 0
95 percent confidence interval:
 -0.07613667  0.12157343
sample estimates:
mean in group Major mean in group Minor
          0.5142059           0.4914875
```

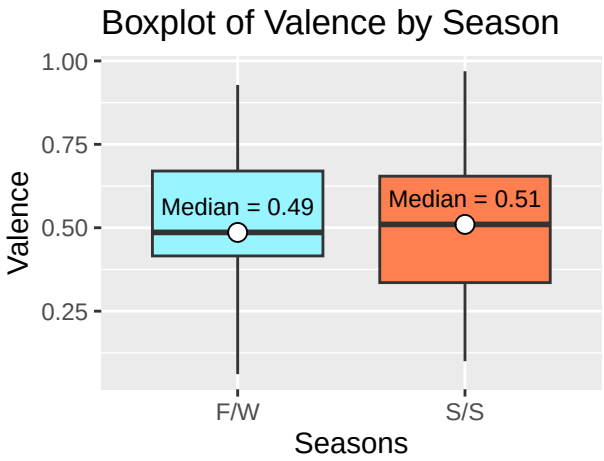
In music theory, major keys are associated with sounding "happier and brighter," while minor keys sound "sadder and gloomier." If you're trying to bring energy, danceability, and valence (happiness) to a party, should you only play songs in major keys, then?

From the box plots, it appears as though the median values of the major mode are generally higher than those of the miner mode, other than when looking at energy. However, according to the t tests, all with p-values far above our alpha of 0.05, there is no statistically significant difference in the median streams, danceability, energy, or valence based on mode.










```r
(test1 <- t.test(songs$streams ~ songs$season, conf.level = 0.95))
```

```
    Welch Two Sample t-test

data:  songs$streams by songs$season
t = 0.031132, df = 76.186, p-value = 0.9752
alternative hypothesis: true difference in means between group F/W and group S/S is not equal to 0
95 percent confidence interval:
 -0.1254265  0.1294100
sample estimates:
mean in group F/W mean in group S/S
         1.866026          1.864034
```

```r
(test3 <- t.test(songs$energy ~ songs$season, conf.level = 0.95))
```

```
    Welch Two Sample t-test

data:  songs$energy by songs$season
t = -1.6342, df = 77.692, p-value = 0.1063
alternative hypothesis: true difference in means between group F/W and group S/S is not equal to 0
95 percent confidence interval:
 -0.1212503  0.0119326
sample estimates:
mean in group F/W mean in group S/S
        0.5912564         0.6459153
```

```
(test4 <- t.test(songs$danceability ~ songs$season, conf.level = 0.95))
```

```
    Welch Two Sample t-test

data:  songs$danceability by songs$season
t = -0.54162, df = 73.818, p-value = 0.5897
alternative hypothesis: true difference in means between group F/W and group S/S is not equal to 0
95 percent confidence interval:
 -0.07786985  0.04458519
sample estimates:
mean in group F/W mean in group S/S
        0.6392051         0.6558475
```

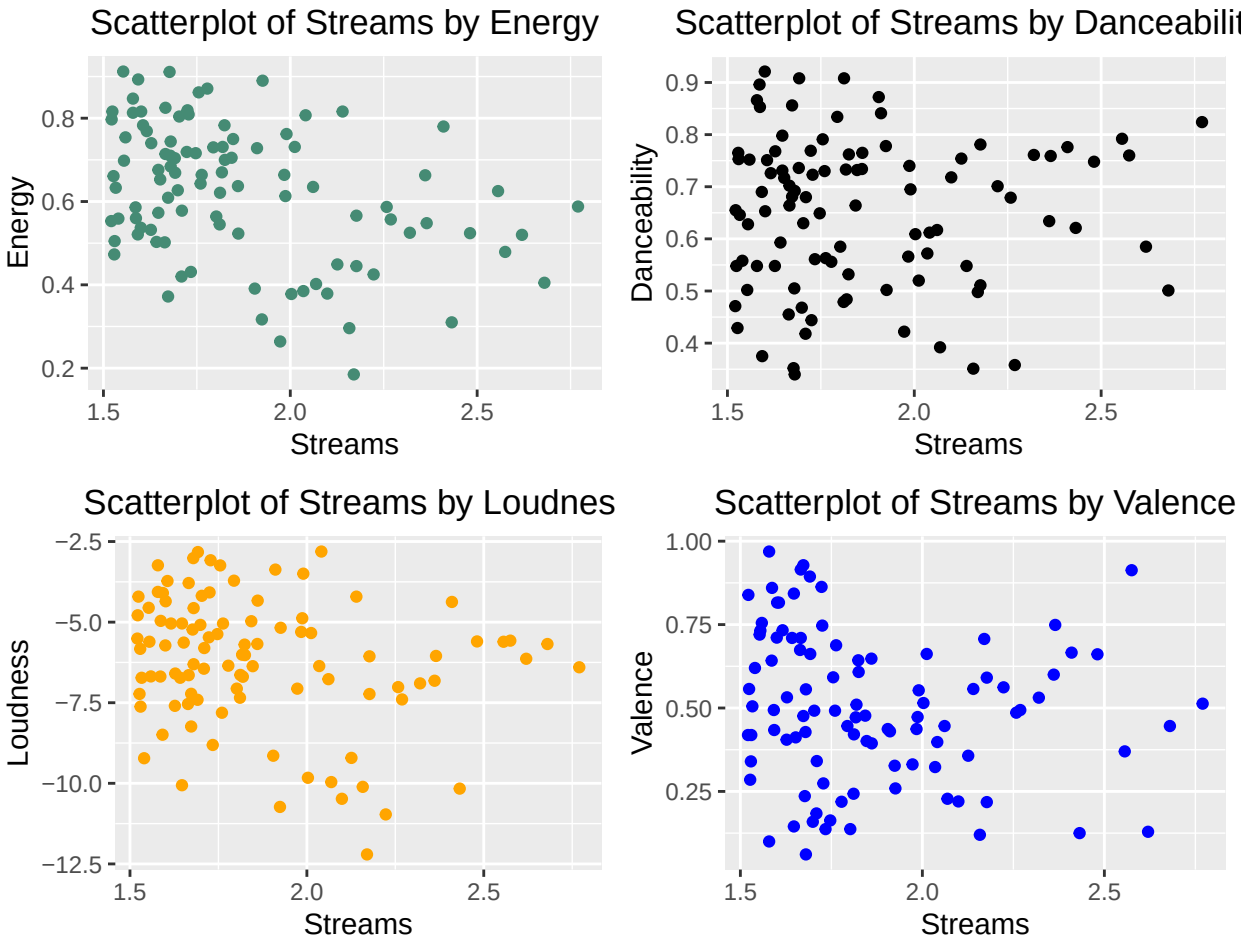
```
(test8 <- t.test(songs$valence ~ songs$season, conf.level = 0.95))
```

```
    Welch Two Sample t-test

data:  songs$valence by songs$season
t = 0.37971, df = 77.558, p-value = 0.7052
alternative hypothesis: true difference in means between group F/W and group S/S is not equal to 0
95 percent confidence interval:
 -0.07532777  0.11083042
sample estimates:
mean in group F/W mean in group S/S
        0.5100564         0.4923051
```

"Songs of the summer" are, traditionally, high-energy songs that reach their commercial peak in summer. Therefore, if you're trying to bring good vibes to a party, should you primarily choose songs released during spring/summer, rather than during the colder, busier months of fall/winter?

According to the boxplots... maybe. Visually, it appears as if songs released in spring/summer have higher median streams, higher median energy, and higher median danceability (but not higher median valence) than songs released in fall/winter. According to the t tests, though, which all have p values greater than 0.05, and the 95% confidence intervals, which all contain 0, there is no statistically significant difference in the median streams, energy, danceability, or valence based on season.

4.2 Scatterplots / Correlation Tests




```
cor(songs$energy, songs$streams)
```

```
[1] -0.374995
```

```
cor(songs$danceability, songs$streams)
```

```
[1] 0.01257812
```

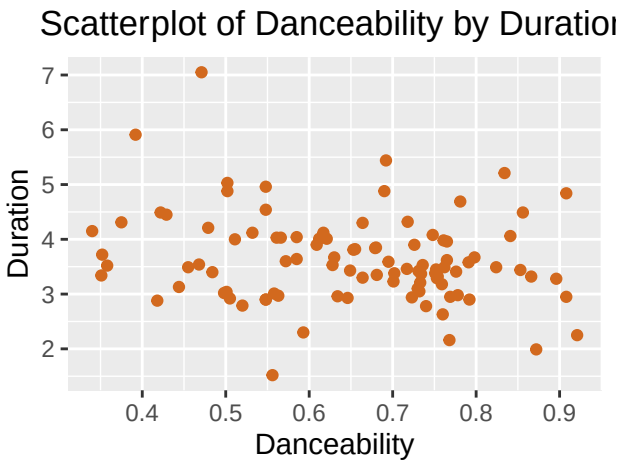
```
cor(songs$loudness, songs$streams)
```

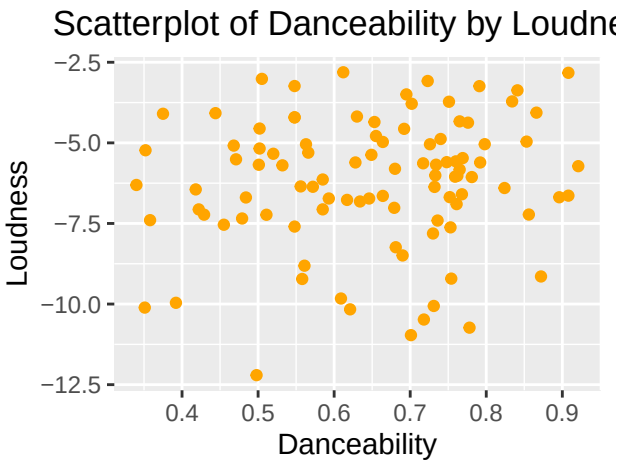
```
[1] -0.2093497
```

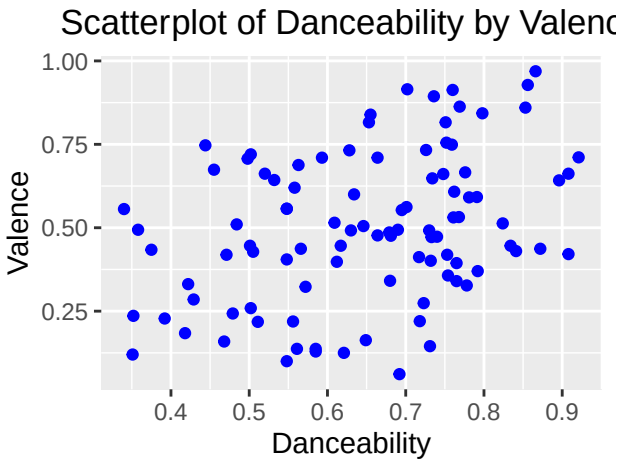
```
cor(songs$valence, songs$streams)
```

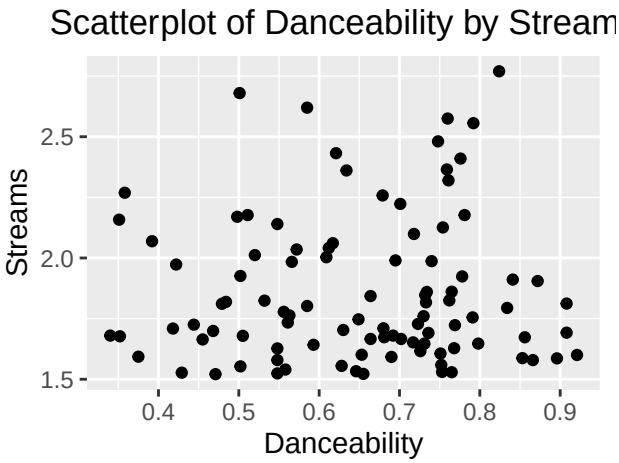
```
[1] -0.1395527
```

Above, we created scatterplots to analyze the relationship between total streams and other variables within the dataset: Energy, Danceability, Loudness, and Valence. The streams and energy scatterplot has a slight negative correlation at -0.374995 as well as light clustering in the upper left corner of the plot. The streams and danceability scatterplot has a minimal positive correlation at 0.01257812 with a larger density of data points on the left side of the plot. The scatterplot of streams and loudness has a slight clustering of data points in the top left and the scatterplot has a light negative correlation around -0.2093497. Finally, the scatterplot of streams and valence has a slight negative correlation at -0.1395527 as well as a slightly denser concentration of data points on the left side of the plot. Overall, there are no major relationships between streams and energy, danceability, loudness, or valence. However, energy, loudness, and valence have slight negative correlations while danceability has a slight positive correlation.










```
cor(songs$danceability, songs$duration)
```

```
[1] -0.2112502
```

```
cor(songs$danceability, songs$loudness)
```

```
[1] 0.1247894
```

```
cor(songs$danceability, songs$valence)
```

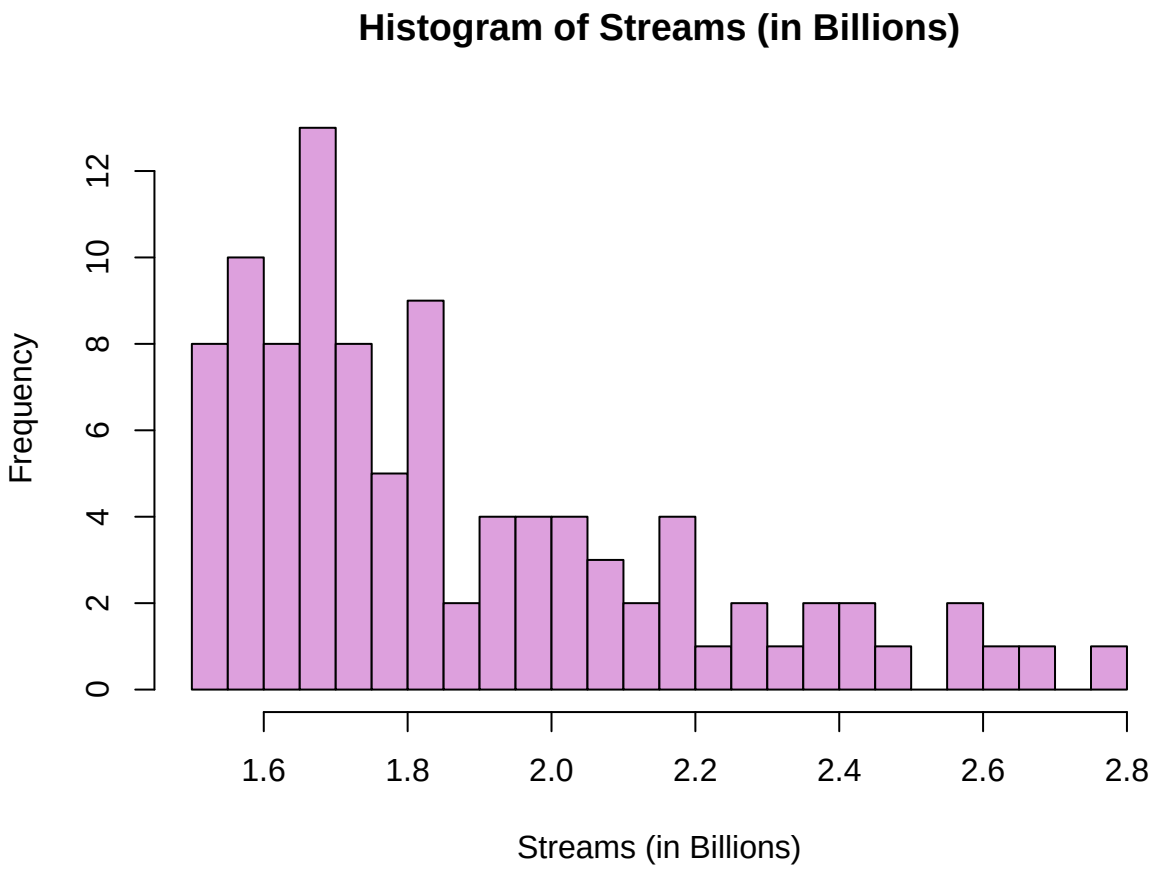
```
[1] 0.3611559
```

```
cor(songs$danceability, songs$streams)
```

```
[1] 0.01257812
```

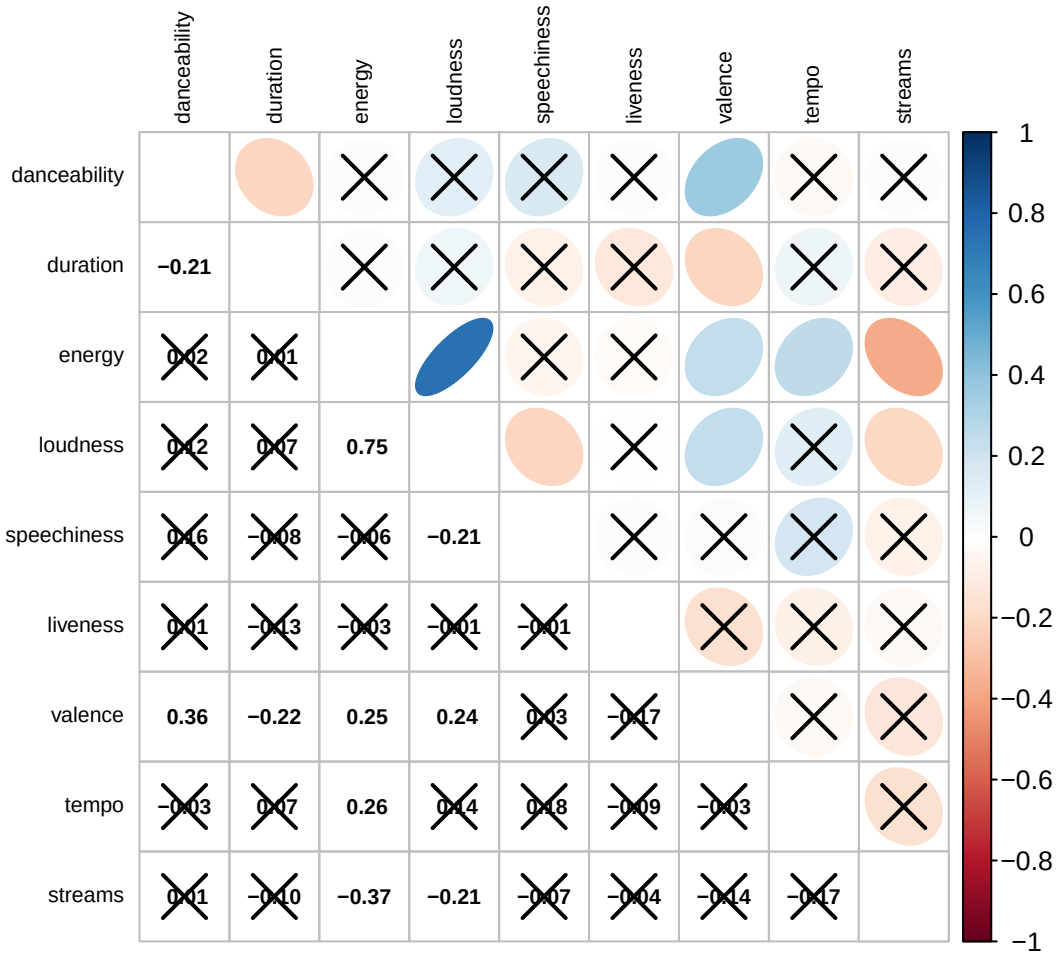
Here, we have created scatterplots to analyze the relationship between Danceability and other variables within the dataset: Duration, Loudness, Valence, and Streams. The danceability and duration scatterplot has a slight negative correlation at -0.2112502 but the data points are relatively well spread out across the plot, slightly denser towards the bottom of the graph. The scatterplot of danceability and loudness has a slight clustering of data points in the top portion of the plot with a light positive correlation around 0.1247894. The scatterplot of danceability and valence has a slight positive correlation at 0.3611559 with a quite even distribution of data points across the plot. Finally, the danceability and streams scatterplot has a minimal positive correlation at 0.01257812 with a larger density of data points on the bottom half of the plot. Overall, there are no major relationships between danceability and duration, loudness, valence, or streams. However, loudness, valence, and streams have slight positive correlations while duration has a slight negative correlation.

4.3 Histogram

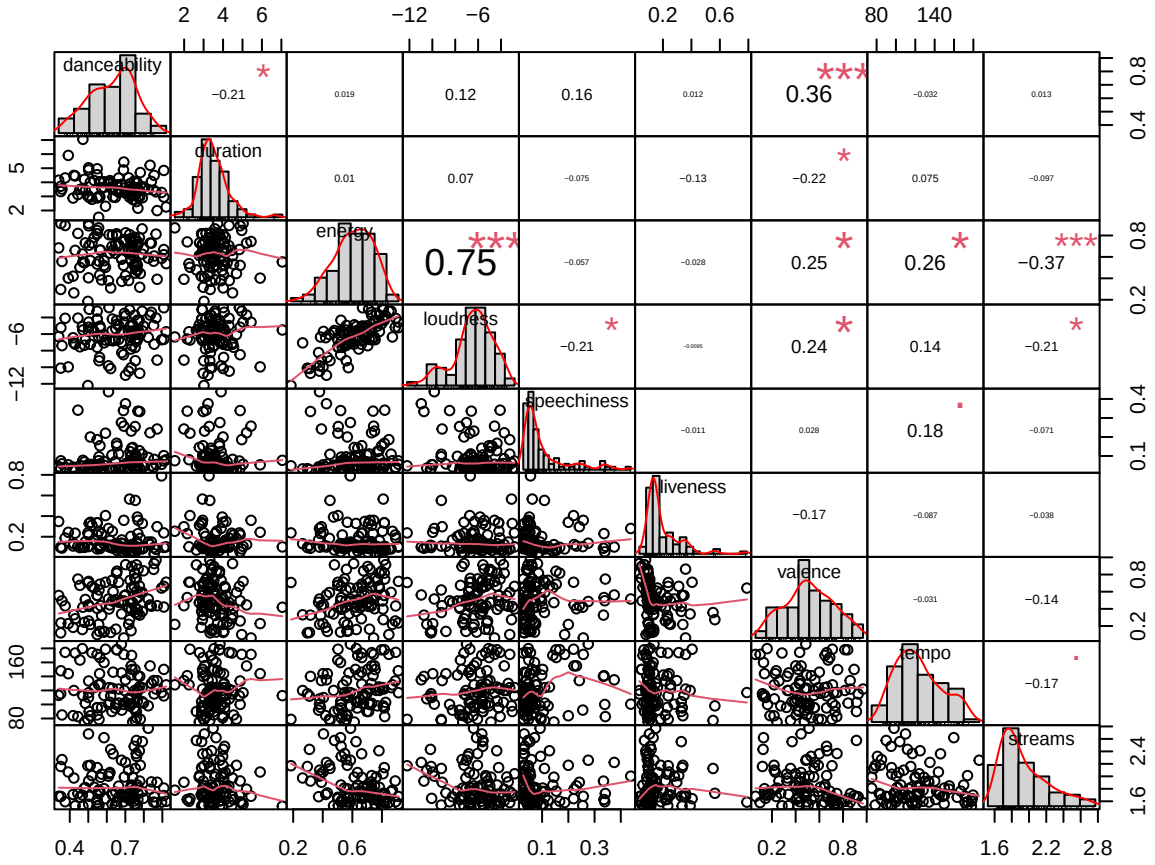


Based on the histogram, it appears that the variable "streams" is heavily right-skewed. This means that the majority of the songs have relatively low numbers of streams, while a smaller number of songs have very high numbers of streams.

4.4 Correlations Plot

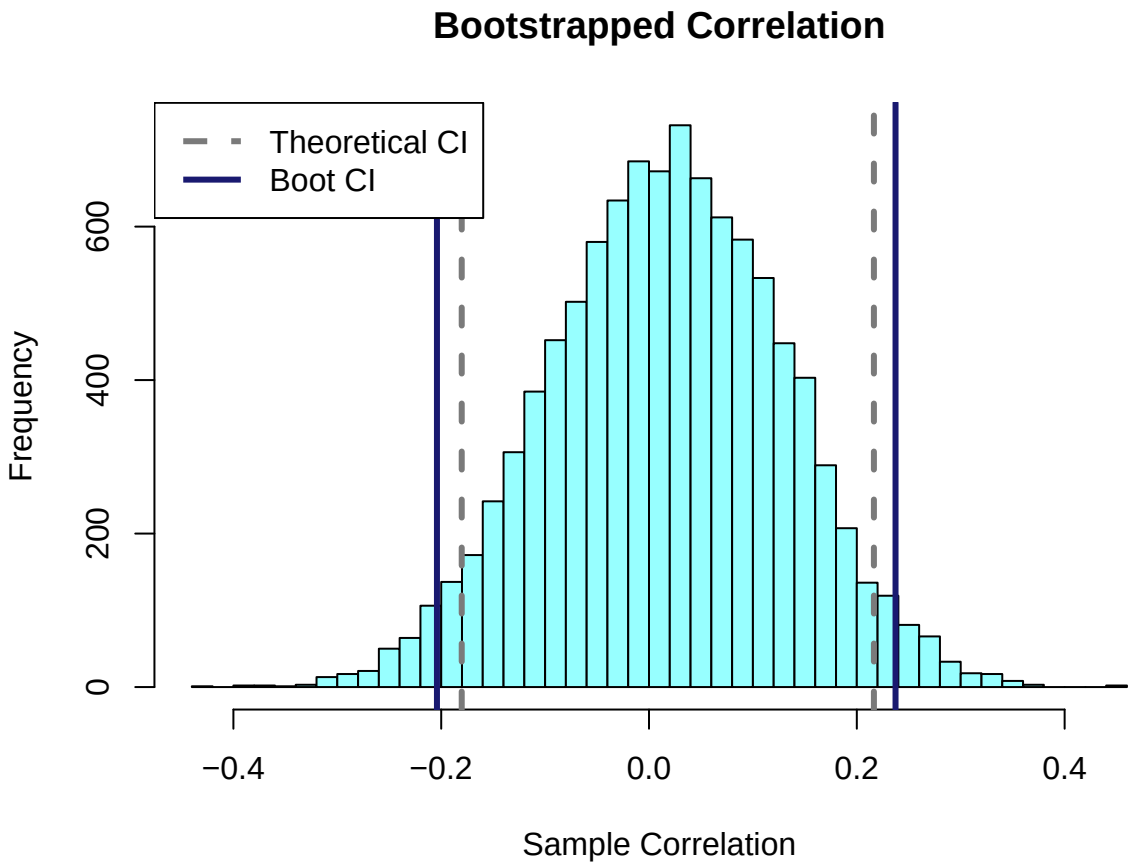


4.5 Correlation Chart



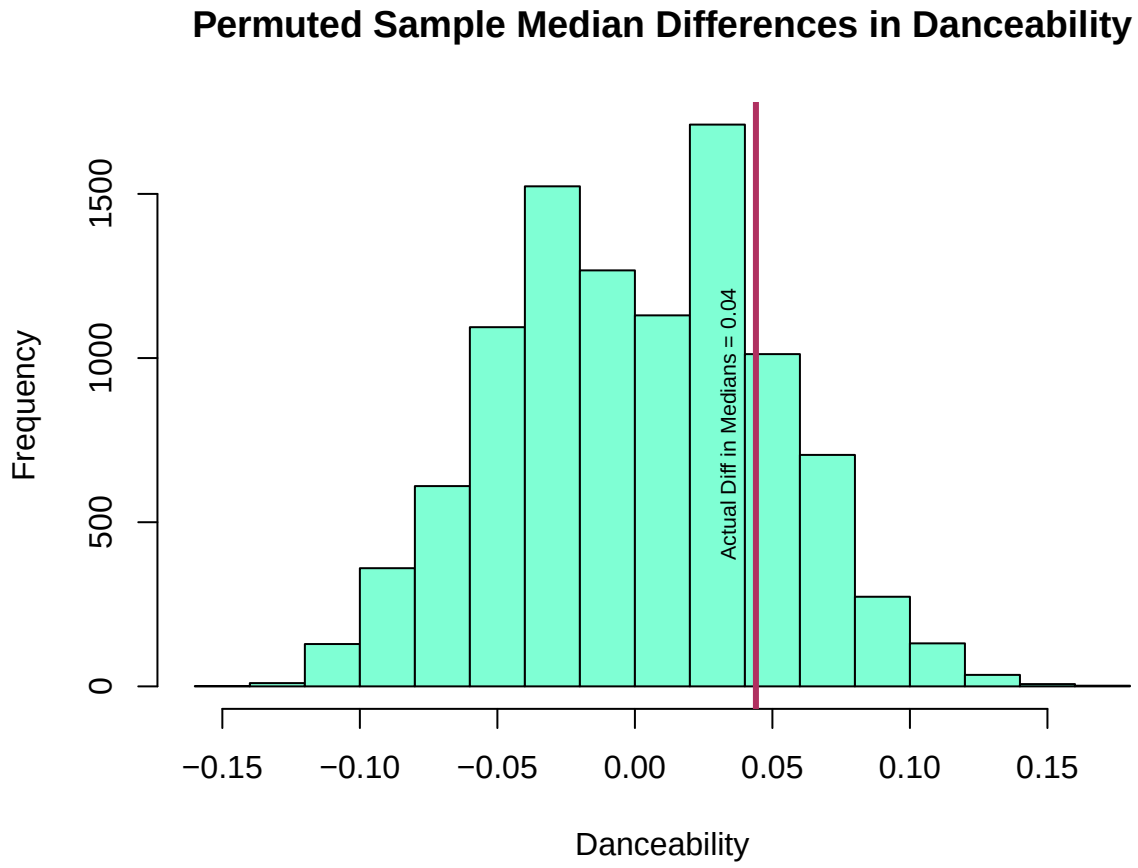
4.6 Bootstraps

```
    Pearson's product-moment correlation

data:  songs$energy and songs$danceability
t = 0.18469, df = 96, p-value = 0.8539
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1802487  0.2164574
sample estimates:
       cor
0.01884606


      2.5%      97.5%
-0.2040988  0.2372545
```



It can be seen that the normal distribution is centered close to the value of 0, meaning that the correlation between energy and danceability is quite minimal. The theoretical CI is slightly narrower than the bootstrap CI and is contained within the bootstrap CI, yet, the wider bootstrap confidence interval values still run between approximately -0.20 and 0.23 which indicate a weak correlation.

4.7 Permutation Tests




```
[1] 0.3884
```

For our permutation test, we wanted to gauge if there was a statistically significant difference in median danceability depending on the season a song was released in. We permuted season groups, calculated medians from our "fake" data, and repeated this process 10000 times. Because our permutation test two sided p value of 0.3884 is $>$ alpha $= 0.05$, we fail to reject the null hypothesis and cannot say that there is a significant difference between the median danceability of songs released in spring/summer and the danceability of songs released in fall/winter.

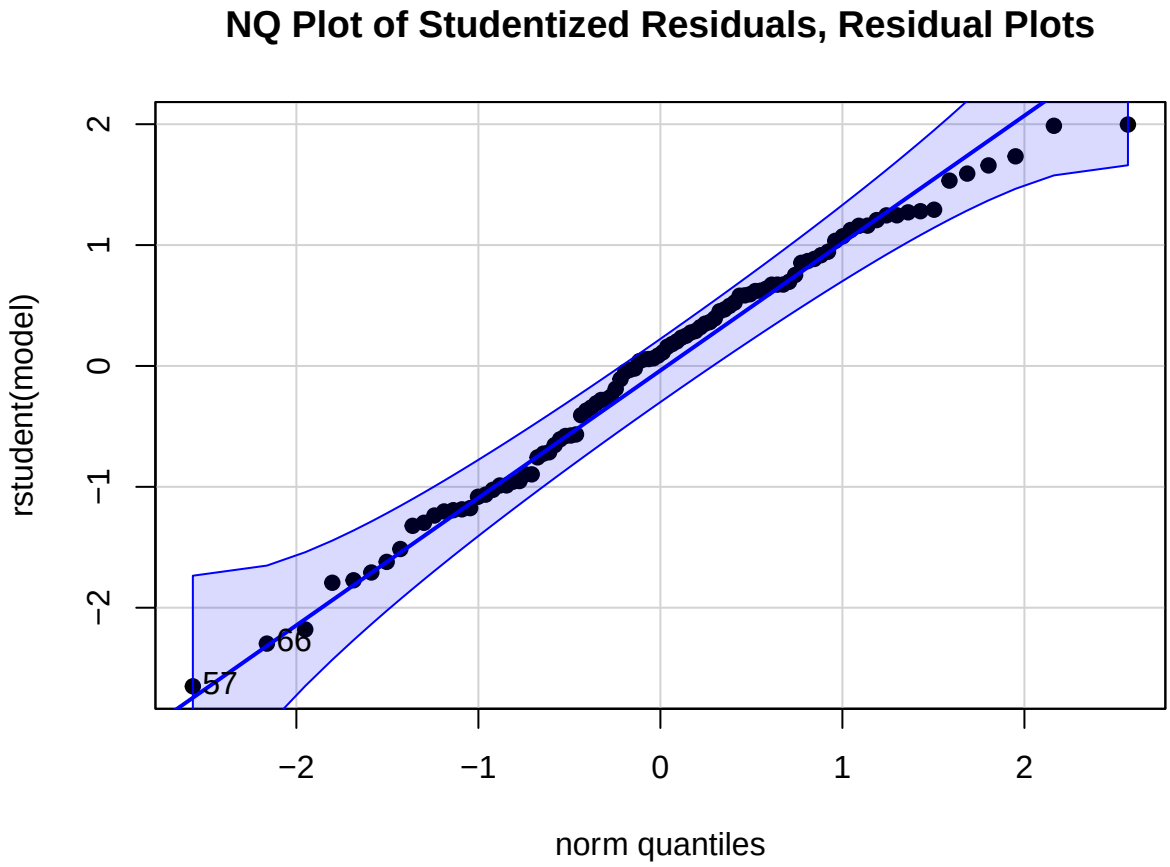
5 Analysis

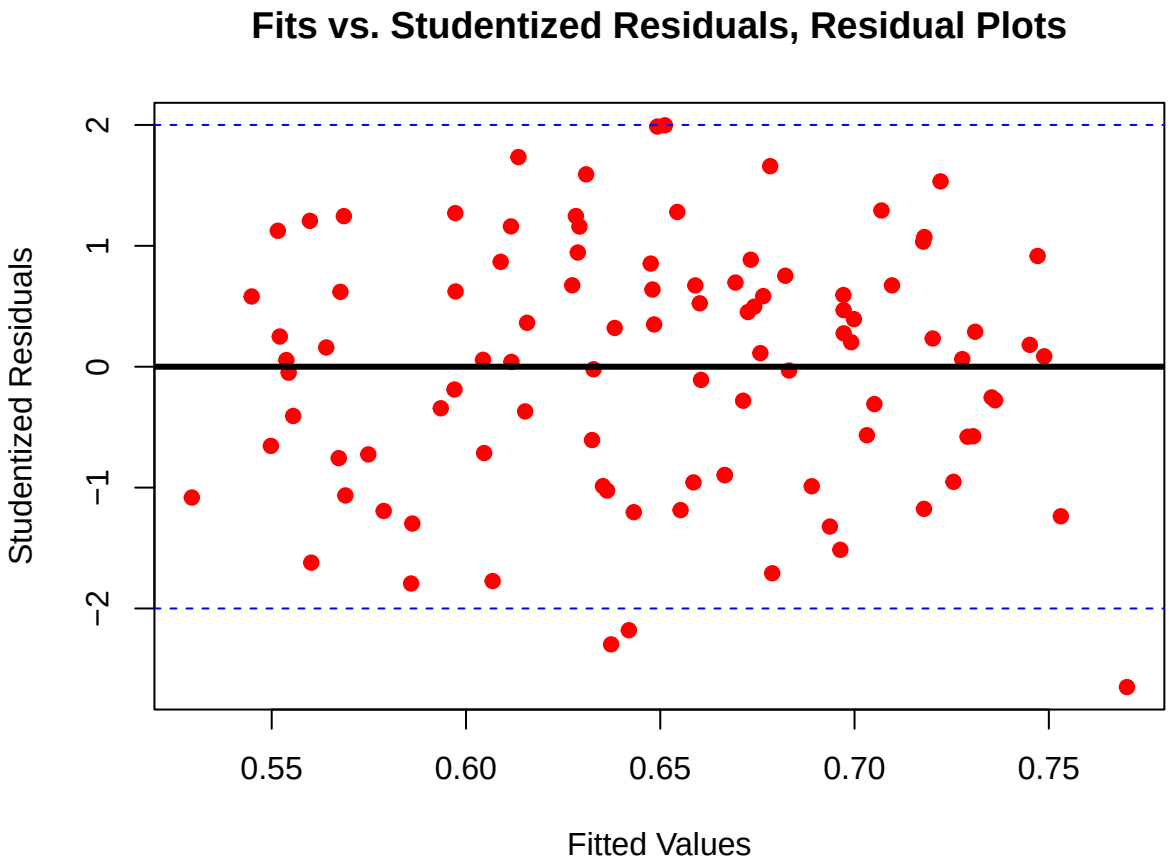
5.1 Multiple Regresssions

We began our multiple regression with 9 variables: danceability (the intercept), duration, energy, loudness, speechiness, liveness, valence, tempo, and streams. The plan was to use multiple models to analyze r-squared, adjusted r-squared, BIC, and CP Statistic in order to minimize the variables used in the model while maximizing its effectiveness to create a model that would best correlate with danceability. The r-sqaured values indicated that the best model would include all 8 additional variables with an r-sqaured value of 0.21. The adjusted r-squared values indicated that the best model would have an adjusted r-sqaured of 0.17 and would contain 5 additional variables: duration, energy, loudness, speechiness, and valence. The BIC values took this model minizing goal much further by indicated the best value, at -4.5, would use a model with only 1 variable: valence. Attempting to find a happy medium while maintaining the accuracy of our model, we settled on the CP Statistic. The CP Statistic suggested a model with three variables alongside danceability: duration, speechiness, and valence.

```
[1] "duration"    "speechiness" "valence"
```

```
Call:
lm(formula = danceability ~ valence + speechiness + duration,
    data = songs.cor)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32610 -0.09841  0.01271  0.08949  0.25875

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.60285    0.07875   7.655 1.68e-11 ***
valence      0.21362    0.06254   3.416  0.00094 ***
speechiness  0.23141    0.15174   1.525  0.13059
duration    -0.02276    0.01710  -1.331  0.18642
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.1336 on 94 degrees of freedom
Multiple R-squared:  0.1694,    Adjusted R-squared:  0.1429
F-statistic: 6.389 on 3 and 94 DF,  p-value: 0.0005485
```





The multiple R-squared for this model is 0.1429, indicating that the models explain around 14% of the variance in danceability, which is relatively low. Regarding the significance of the independent variables, valence is the only variable that is significant in both models with a p-value less than 0.05. The coefficient for valence is positive, indicating that as valence increases, danceability tends to increase. However, the coefficients for speechiness and duration are not significant in the second model. The p-value for speechiness is 0.13059, which is greater than 0.05, and the p-value for duration is 0.18642. Overall, the results suggest that valence is the most important variable in predicting danceability, while speechiness and duration may not be as significant. The normal quantile plot of the studentized residuals and the fits v. residuals plot suggest that the assumptions of normality and homoscedasticity are met, indicating that the linear regression model is appropriate for the data. Because of this, **there was no need to *transform* the data as we saw no signs of heteroskedasticity in our complete multiple regression's model**.

The adjusted r-squared model seemed to overfit the data by using too many variables, while the BIC model oversimplified and may not be the best in terms of predictive accuracy. The CP statistic seemed to be the best proposed model because it took into account both the goodness of the fit and the model complexity, striking a balance between overfitting and underfitting the given data.

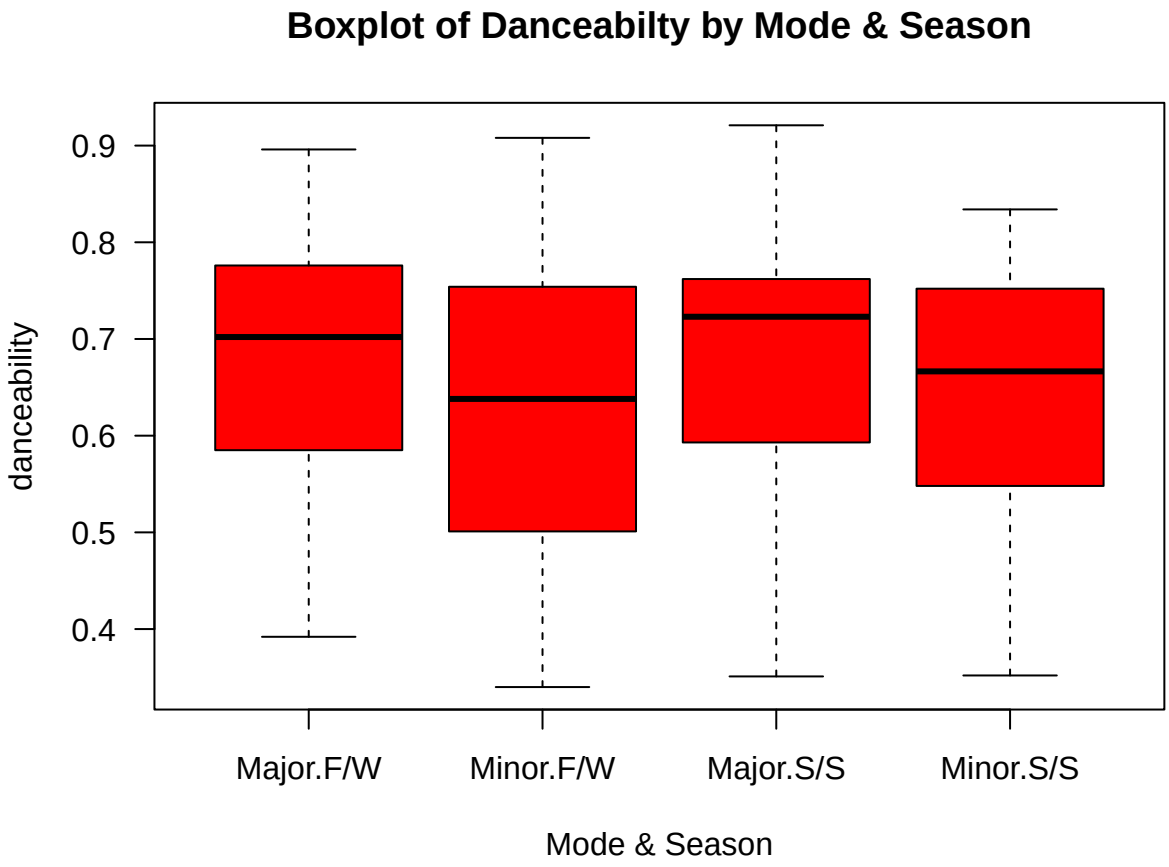
5.2 ANOVA (avengers) & Tukey

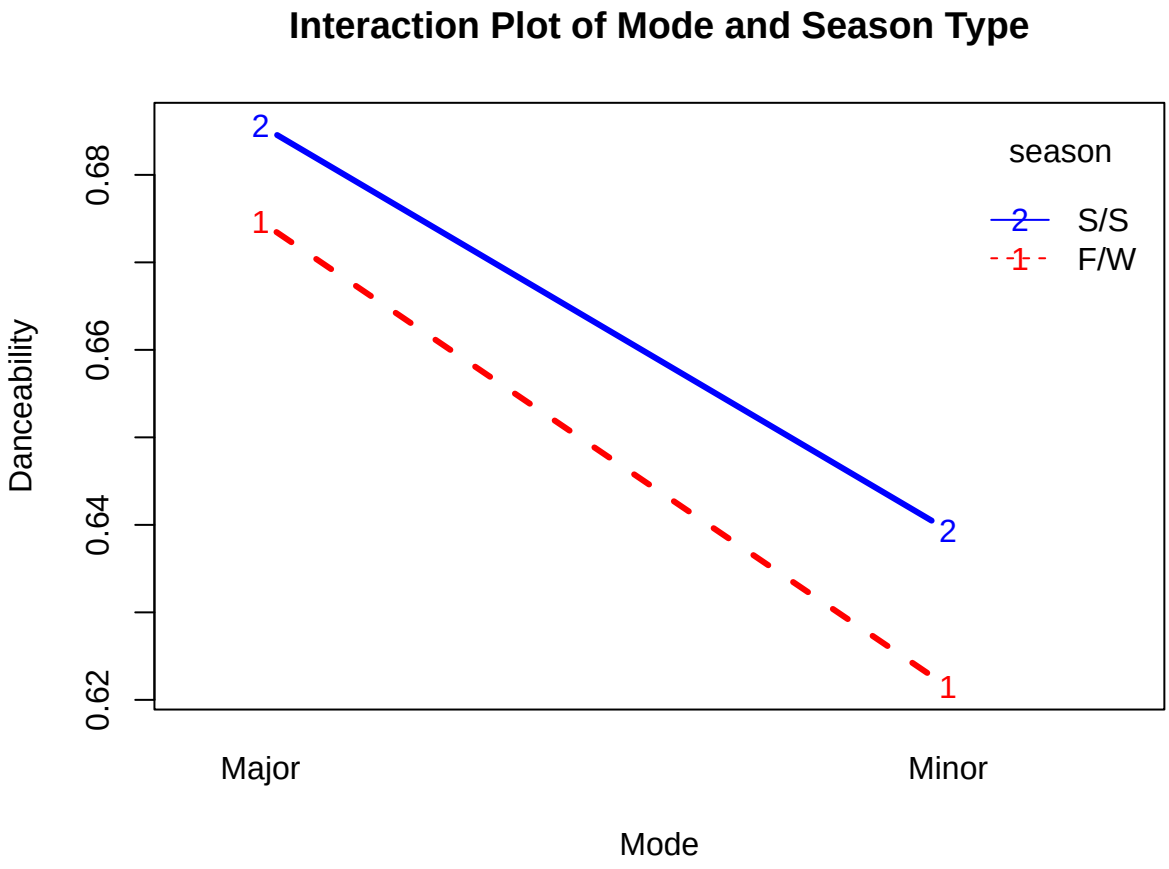
```
[1] "SD by Combo"
```

```
Major F/W Major S/S Minor F/W Minor S/S
0.1450060 0.1482226 0.1615032 0.1291857
```

```
[1] "Ratio of Max/Min Sample SD"
```

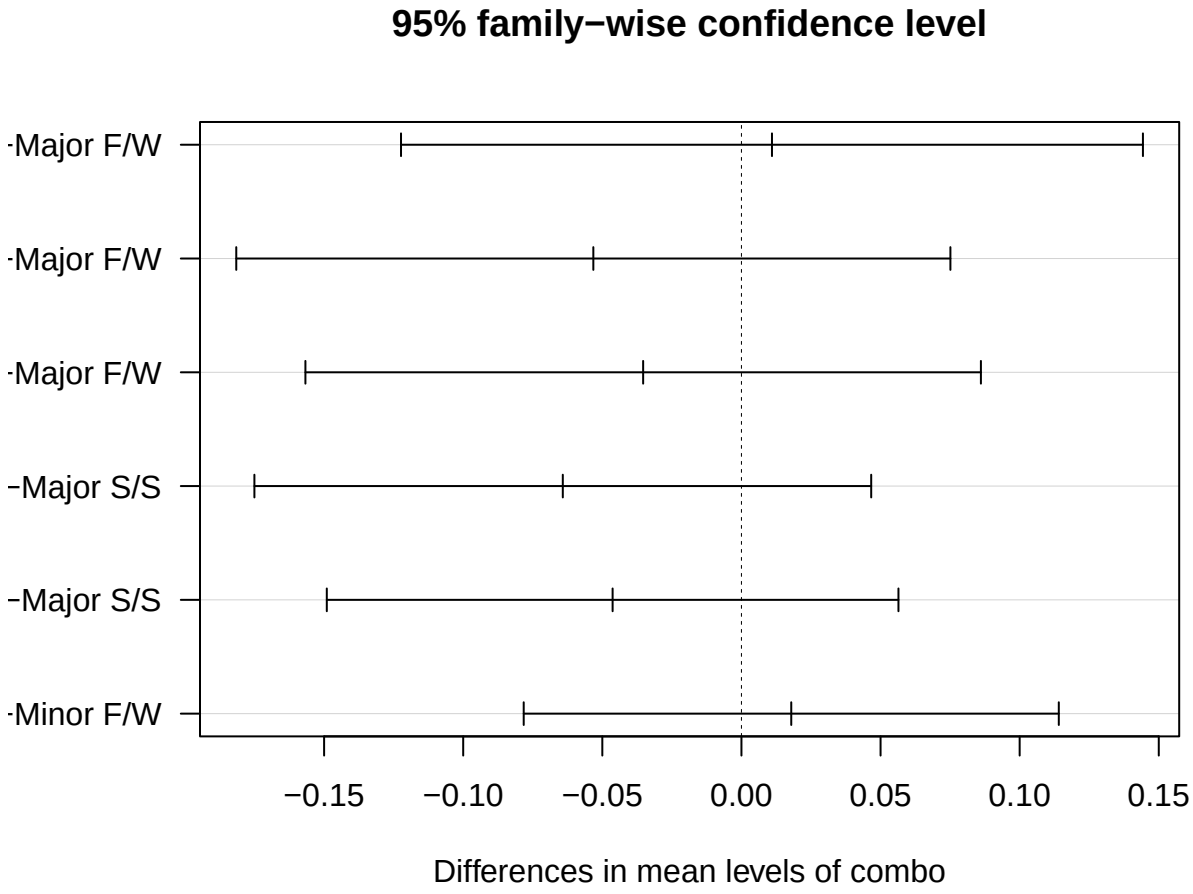
```
[1] 1.3
```

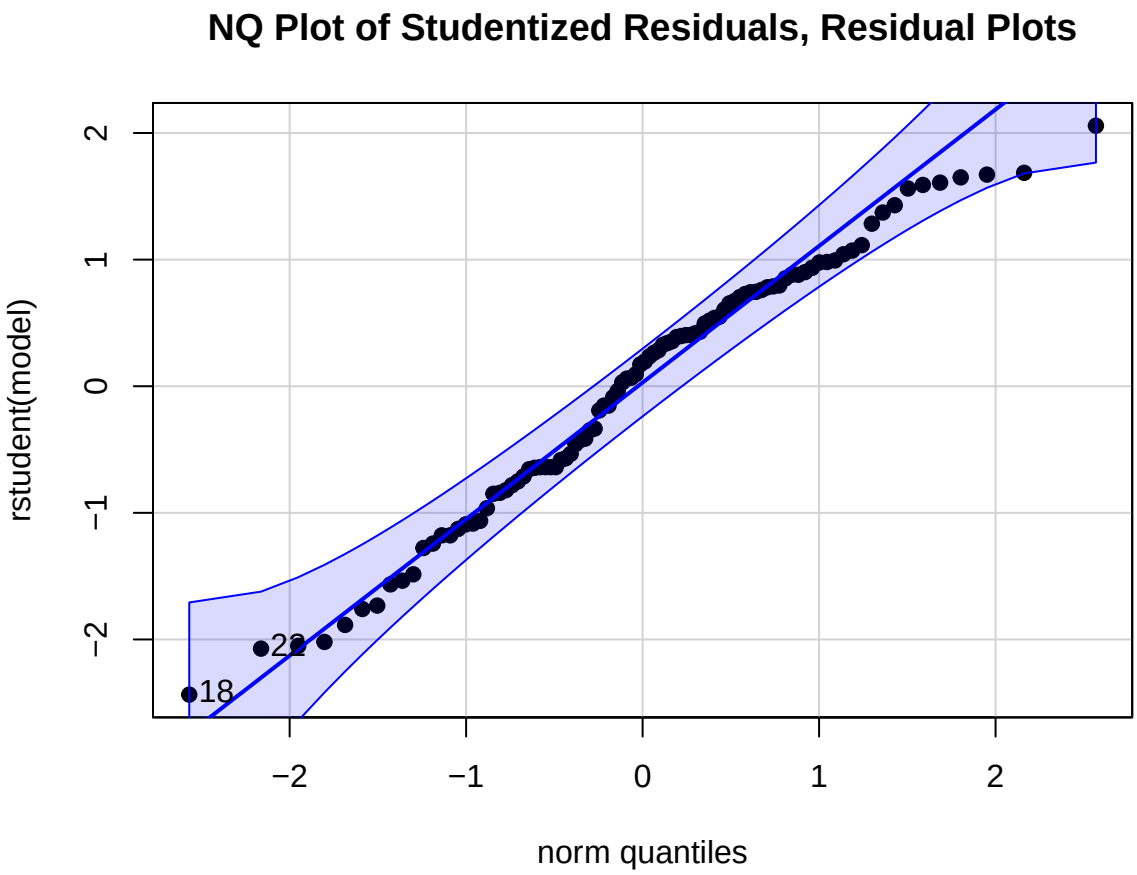

Boxplot of Danceabilty by Mode & Season
danceability
0.9
0.8
0.7
0.6
0.5
0.4
Major.F/W
Minor.F/W
Major.S/S
Minor.S/S
Mode & Season

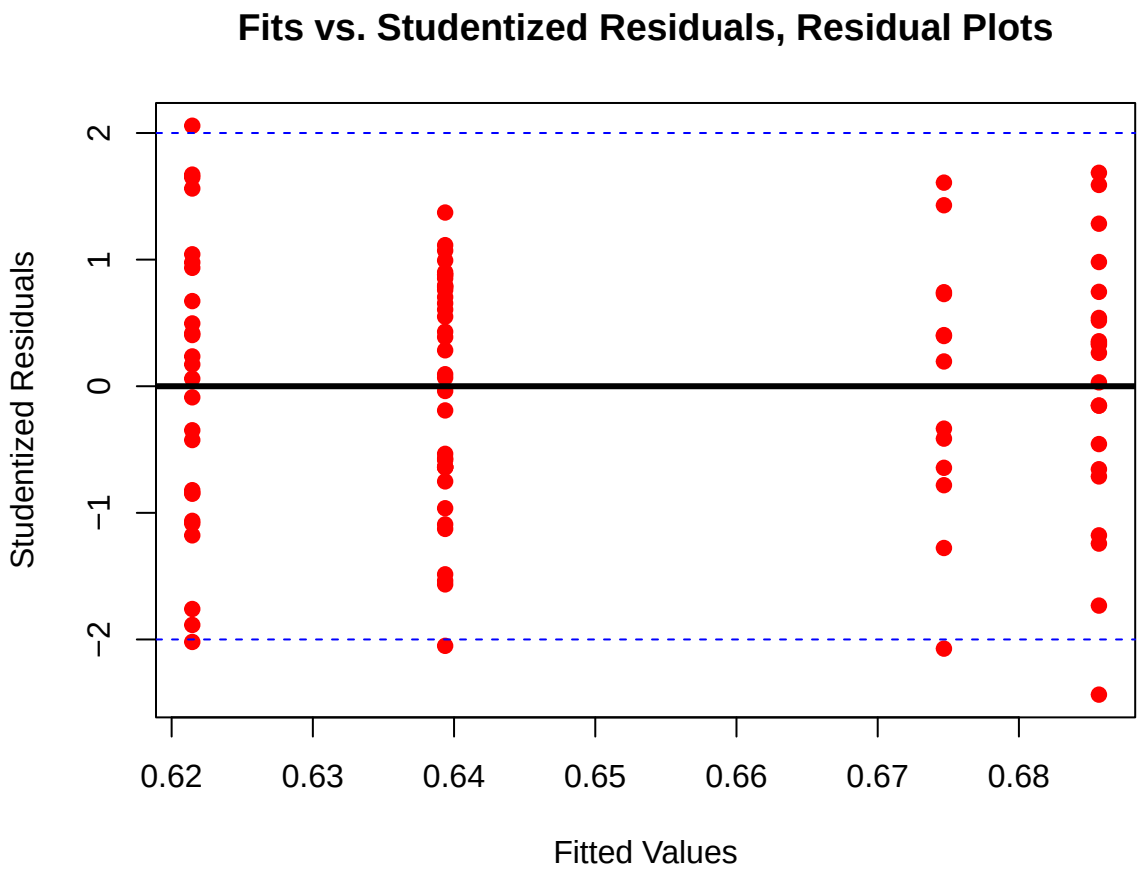



```
            Df Sum Sq Mean Sq F value Pr(>F)
mode         1 0.0541 0.05414   2.595  0.111
season       1 0.0057 0.00566   0.271  0.604
mode:season  1 0.0003 0.00025   0.012  0.912
Residuals   94 1.9613 0.02086
```

```
Call:
lm(formula = danceability ~ mode + season + mode * season - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33467 -0.09867  0.02592  0.10781  0.28654

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
modeMajor           0.674692   0.040062  16.841   <2e-16 ***
modeMinor           0.621462   0.028328  21.938   <2e-16 ***
seasonS/S           0.010974   0.050976   0.215    0.830
modeMinor:seasonS/S 0.006933   0.062850   0.110    0.912
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.1444 on 94 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.9528
F-statistic: 495.6 on 4 and 94 DF,  p-value: < 2.2e-16
```


95% family−wise confidence level
Major F/W
Major F/W
Major F/W
Major S/S
Major S/S
Minor F/W
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Differences in mean levels of combo


NQ Plot of Studentized Residuals, Residual Plots
rstudent(model)
norm quantiles
2
1
0
−1
−2
−2
−1
0
1
2
22
18



We ran a 2-way ANOVA to model the effects of mode, season, and the interaction between mode and season on danceability. The ratio of max/min sample deviation between danceability and the combo of mode and season is 1.3, so equal variances can be assumed. The interaction plots do not seem to display any interaction between mode and the season in which a song was released. Finally, we fit our ANOVA model and displayed the Tukey plots of the 95% confidence intervals which all contain 0 within their intervals, showing that there is no significant relationship between danceability and mode, season, or their combination.

According to the qq plots and fits vs studentized residual plots, the data appears approximately normal and there does not appear to be significant skewness. It appears as if the standard deviation of the residuals is constant, and there are no clear outliers with any studentized residuals greater than an absolute value of 3. There does not appear to be heteroskedasticity.

6 Conclusion/Summary

In our paper, we **discovered statistically significant relationships** between musical features and streams/danceability. These are Energy & Streams (-0.37 correlation), Valence & Danceability (0.36 correlation), Duration & Danceability (-0.21 correlation), and Loudness & Loudness (-0.21). On the other hand, we observed **no considerable variations** between the song characteristics and the song's mode or release season. With these results in mind, it would be beneficial for the aux controller or aspiring musician to choose/create a song that is **mellow, on the shorter side and filled with positive lyrics**. To reaffirm and gain greater accuracy of results, we recommend rerunning the analysis described in this paper, on a dataset of all songs on spotify, rather than the top 100 streamed songs.