# Statistical Modelling and Inference II Project

Scott Carnie-Bronca (1721235)
Rose Crocker (1668575)
Isaac Jacobson (1132570)
Curtis Murray (1670295)
Michael Ucci (1686935)

May 23, 2019

# Contents

# 1 Introduction

Music streaming providers, such as Spotify, often feature individualised song recommendations for their patrons and, as such, would benefit from the ability to predict the popularity of the tracks contained within their catalogues. In response to this desirable capability, a predictive model for the *popularity* response was developed by fitting a linear regression model with the following predictors:

- *danceability*;
- *energy*;
- *key*;
- *loudness*;
- *mode*;
- *duration_ms*;
- *time_signature*; and
- *decade*.

The linear regression model was fitted on a sample dataset that was obtained from Spotify. The dataset corresponded to all of the tracks contained in Spotify's library for six artists, where each set of artist tracks was assumed to be representative of all the songs of a particular decade. Consequently, the tracks of Elvis Presley, The Beatles, David Bowie, Michael Jackson, Blur and Beyoncé were used to represent all of the songs from the 1950's, 1960's, 1970's, 1980's, 1990's and 2000's respectively.

Once the Spotify dataset had been sourced, the following steps were performed in developing the predictive model for song popularity:

1. The dataset was described so that an understanding of how the dataset represented tracks could be obtained;
2. The dataset was cleaned so that the data it contained could be employed in the model fitting process;
3. Univariate analysis was performed on the response and predictors data to examine each variables' individual distributions;
4. Bivariate analysis was undertaken to uncover the strength and nature of the relationship between each response-predictor pair;
5. Apply stepwise regression algorithms to fit various linear regression models on the dataset via the least squares approach and select the final predictive model through a comparison of the resulting models;
6. Interpret the final model to ensure that its features and behaviour is consistent with expectations gained from steps (2) - (4); and
7. Verify whether the assumptions inherent in the least squares approach are plausibly satisfied by the final model.

After the final model had been developed and analysed, it was then applied to an example song for the purposes of predicting the song's expected popularity.

## 2 Data Description

Table 1 provides a summary of the dataset that was used to develop a predictive model for the popularity of tracks on Spotify. The dataset contained 2081 subjects where each subject represented a different Spotify track. The track subjects were each described with the 21 variables listed in Table 1. From Table 1 it can be seen that *popularity* was the numerical response while *danceability*, *energy*, *loudness* and *duration_ms* were the numerical predictors, and *key*, *mode*, *time_signature* and *decade* were the categorical predictors.

Table 1: Summary of the dataset variables.

| **Variable** | **Class Type** | **Variable Type** | **Model Type** |
|---|---|---|---|
| *album_name* | `character` | categorical | - |
| *track_name* | `character` | categorical | - |
| *track_uri* | `character` | categorical | - |
| *artist_uri* | `character` | categorical | - |
| *danceability* | `float` | numeric | predictor |
| *energy* | `float` | numeric | predictor |
| *key* | `character` | categorical | predictor |
| *loudness* | `float` | numeric | predictor |
| *mode* | `character` | categorical | predictor |
| *speechiness* | `float` | numeric | - |
| *acousticness* | `float` | numeric | - |
| *instrumentalness* | `float` | numeric | - |
| *liveness* | `float` | numeric | - |
| *valence* | `float` | numeric | - |
| *tempo* | `float` | numeric | - |
| *duration_ms* | `int` | numeric | predictor |
| *time_signature* | `int` | categorical | predictor |
| *key_mode* | `character` | categorical | - |
| *popularity* | `int` | numeric | response |
| *artist* | `character` | categorical | - |
| *decade* | `character` | categorical | predictor |

Tables 2 and 3 summarise the descriptions of the numerical and categorical variables of the dataset respectively. From Table 2 it can be seen that *loudness* was expected to have negative values as this variable was measured in decibels. It should be noted that the *time_signature* was treated as a categorical variable, despite being defined as `int` class type within the dataset, as shown in Table 1. Since *time_signature* could only be assigned six possible values, as shown in Table 3, it was determined that this variable would be more meaningful if treated as a categorical variable. Finally, it can be seen from Table 3 that *time_signature* was capable of being assigned a value of zero beats per bar. Such a value for *time_signature* was taken to represent tracks which did not feature any rhythmic instruments, such as drums.

Table 2: Descriptions of the numerical variables [1][2].

| **Variable** | **Description** | **Lower Bound** | **Upper Bound** |
|---|---|---|---|
| *danceability* | A measure of a track's suitability for dancing. Calculated from elements such as tempo, rhythm and beat strength. Larger values indicate greater suitability. | 0.0 | 1.0 |
| *energy* | A measure of a track's the intensity and activity. Calculated from elements such as dynamic range and general entropy. Larger values indicate greater intensity and activity. | 0.0 | 1.0 |
| *loudness* | The average loudness of a track in decibels. Larger values indicate louder tracks (the range of possible values effectively can have no bounds, and as such, the values provided are a typical range). | -60.0 | 0.0 |
| *speechiness* | A measure of the likelihood of language presence in a track. Larger values indicate higher likelihood. | 0.0 | 1.0 |
| *acousticness* | A measure of the likelihood that a track is acoustic. Larger values indicate higher likelihood. | 0.0 | 1.0 |
| *instrumentalness* | A measure of the likelihood that a track contains no vocals. Larger values indicate higher likelihood. | 0.0 | 1.0 |
| *liveness* | A measure of the likelihood that a track was recorded live. Larger values indicate higher likelihood. | 0.0 | 1.0 |
| *valence* | A measure of a track's sentiment. Larger values indicate stronger positive moods, tones, emotions and behaviours. | 0.0 | 1.0 |
| *tempo* | The track's overall beats per minute. | 0.0 | Unbounded |
| *duration_ms* | The track's duration in milliseconds. | 0.0 | Unbounded |
| *popularity* | An algorithmic measure for quantifying a track's current popularity relative to the most played tack(s) on Spotify such that recently played tracks are given greater weights. Larger values indicate higher current popularity. | 0.0 | 100.0 |

Table 3: Descriptions of the categorical variables [1][2].

| **Variable** | **Description** | **Levels** |
|---|---|---|
| *album_name* | The name of a track's album. | - |
| *track_name* | The name of a track. | - |
| *track_uri* | The Spotify URI for a track. | - |
| *artist_uri* | The Spotify URI a the track's artist. | - |
| *key* | The key of a track. | A, A#, B, C, C#, D, D#, E, F, F#, G, G# |
| *mode* | The modality of a track. | major, minor |
| *time_signature* | The track's overall number of beats per bar. | 0, 1, 2, 3, 4, 5 |
| *key_mode* | The key and modality of a track. | A M, A# M, B M, C M, C# M, D M, D# M, E M, F M, F# M, G M, G# M, A m, A# m, B m, C m, C# m, D m, D# m, E m, F m, F# m, G m, G# m |
| *artist* | The name of a track's artist. | Beyoncé, Blur, David Bowie, Elvis Presley, Michael Jackson, The Beatles |
| *decade* | The decade most associated with a track's artist. | 50s, 60s, 70s, 80s, 90s, 00s |

M = major; m = minor

# 3 Cleaning of Data

Before the response and predictors within the dataset could be used to develop the predictive model, the data needed to be cleaned. Refer to Appendix A for the full R code that was used to clean the dataset. The first step in cleaning the data involved identifying any subjects which recorded missing values for the response or any of the predictors. The following R code was employed to determine which of the considered variables contained missing values by checking whether the number of non-NA values equalled the total number of values for each variable:

```
# define numerical & category variables ----
num_vars = c("popularity", "danceability", "energy" ,"loudness", "duration_ms")
cat_vars = c("key", "mode" ,"time_signature", "decade")

# check for missing data ----
has_no_missing_data <- function(dataset, var_name) {
  sum(!is.na(data[[var_name]])) == length(data[[var_name]])
}
for (i in seq_along(num_vars)) {
  print(c(num_vars[i], has_no_missing_data(data, num_vars[i])))
}
for (i in seq_along(cat_vars)) {
  print(c(cat_vars[i], has_no_missing_data(data, cat_vars[i])))
}
```

which produced the following output:

```
[1] "popularity" "TRUE"
[1] "danceability" "TRUE"
[1] "energy" "TRUE"
[1] "loudness" "TRUE"
[1] "duration_ms" "TRUE"
[1] "key"  "TRUE"
[1] "mode" "TRUE"
[1] "time_signature" "TRUE"
[1] "decade" "TRUE"
```

Hence, all of the above outputs had TRUE values which indicated that none of the variables had NA values, and as such, were taken to have no missing values. The second step was to inspect the summaries of each response and predictor to check whether any of the variables had summary statistics which suggested a distribution that was inconsistent with their respective descriptions from Tables 2 and 3. The following R code was used for inspecting the numerical variables:

```
# inspect numerical variables ----
inspect_num_data <- function(dataset, var_name) {
  summary(dataset[[var_name]], useNA = "always")
}
for (i in seq_along(num_vars)) {
  print(num_vars[i])
  print(inspect_num_data(data, num_vars[i]))
}
```

which produced the following output:

```
[1] "popularity"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   19.00   26.00   29.42   41.00   82.00
[1] "danceability"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.3770  0.4990  0.5013  0.6270  0.9630
[1] "energy"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00589 0.40600 0.62200 0.59559 0.80900 0.99800
[1] "loudness"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-30.644 -12.628  -9.291 -10.056  -6.651  -0.933
[1] "duration_ms"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13213  153733  203640  216049  261733 1187253
```

and the following R code was used for inspecting the categorical variables:

```
# inspect categorical variables ----
inspect_cat_data <- function(dataset, var_name) {
  table(dataset[[var_name]], useNA = "always")
}
```

```
for (i in seq_along(cat_vars)) {
  print(cat_vars[i])
  print(inspect_cat_data(data, cat_vars[i]))
}
```

which produced the following output:

```
[1] "key"
   A   A#    B    C   C#    D   D#    E    F   F#    G   G# <NA>
 272  115  143  352  149  259   44  199  155   71  240   82    0
[1] "mode"
major minor  <NA>
 1618   463     0
[1] "time_signature"
   0    1    3    4    5 <NA>
   4   25  221 1808   23    0
[1] "decade"
 00s  50s  60s  70s  80s  90s <NA>
 194  648  247  547  253  192    0
```

As all of the numerical variables had summary statistic values which fell within their respective ranges outlined in Table 2, none of the numerical variables were determined to have obviously erroneous entries. Similarly, since all of the categorical variables had values which were consistent with their respective levels listed in Table 3, none of the category variables were found to contain entries which should be removed. As no variable required the removal of any subject, the final step was to convert the categorical variables from `character`/`int` data types to `factor` types while mapping them with an appropriate order. The following `R` code was employed to perform this conversation procedure on the category variables:

```
# convert categorical variables to factor types and map values with an order ----
data$key <- factor(data$key,
                   levels = c("A", "A#", "B", "C",  "C#", "D", "D#",
                              "E", "F", "F#", "G", "G#"))
data$mode <- factor(data$mode,
                    levels = c("major", "minor"))
data$decade <- factor(data$decade,
                      levels = c("50s", "60s", "70s", "80s", "90s", "00s"))
data$time_signature <- factor(data$time_signature,
                              levels = c("0", "1", "2", "3", "4", "5"))
```

As the data cleaning process resulted in the removal of no subjects from the dataset, the summary statistics of all of the considered variables before and after the cleaning processes remained unchanged.

# 4 Variable Description

For each response and predictor the appropriate summary statistics and plot was generated so that their distributions could be described. Refer to Appendix A for the full `R` code that was used to analyse and plot the dataset. The *popularity*, *danceability*, *energy*, *loudness* and *duration_ms* were plotted with histograms as they corresponded to numerical variables while the *key*, *mode*, *time_signature* and *decade* were plotted on bar charts as they were treated as categorical variables. Tables 4 and 5 summarise the summary statistics of the numerical and categorical variables respectively.

Table 4: Summary statistics of the numerical variables.

| **Variable** | **Min** | **$Q_1$** | **Med** | **Mean** | **$Q_3$** | **Max** | **SD** |
|---|---|---|---|---|---|---|---|
| *popularity* | 0.0 | 19.0 | 26 | 29.4 | 41.0 | 82.0 | 15.8 |
| *danceability* | 0.000 | 0.377 | 0.499 | 0.501 | 0.627 | 0.963 | 0.174 |
| *energy* | 0.006 | 0.406 | 0.622 | 0.596 | 0.809 | 0.998 | 0.252 |
| *loudness* | -30.6 | -12.6 | -9.3 | -10.1 | -6.7 | -0.9 | 4.5 |
| *duration_ms* | 13213 | 153733 | 203640 | 216049 | 261733 | 1187253 | 95712 |

Min = Minimum; $Q_1$ = First quartile; Med = Median; $Q_3$ = Third quartile; Max = Maximum; SD = Standard deviation

Table 5: Summary statistics of the categorical variables.

| **Variable** | **Min (Count)** | **Max (Count)** | **Med** | **Mean** | **SD** |
|---|---|---|---|---|---|
| *key* | D# (44) | C (352) | 152.0 | 173.4 | 92.9 |
| *mode* | minor (463) | major (1618) | 1040.5 | 1040.5 | 816.7 |
| *time_signature* | 2 (0) | 4 (1808) | 24.0 | 346.8 | 720.7 |
| *decade* | 90s (191) | 50s (648) | 250.0 | 246.8 | 198.4 |

Min = Level with minimum count; Max = Level with maximum count; Med = Median level count; Mean = Mean level count; SD = Standard deviation level count;

## 4.1 Popularity Description

Figure 1 depicts the histogram of the *popularity* data. The *popularity* variable was a discrete numerical variable. From Figure 1 is can be seen that the *popularity* data appeared to be bimodal with the major and minor peaks at 22 and zero respectively. The minor peak may due to the lower bound of *popularity* at zero. Furthermore, it can also be seen from Figure 1 that the *popularity* data appeared to be right skewed. The *popularity* data was found to have a mean, median and standard deviation of 29.4, 26.0 and 15.8 respectively and a range of zero to 82, as shown in Table 4.

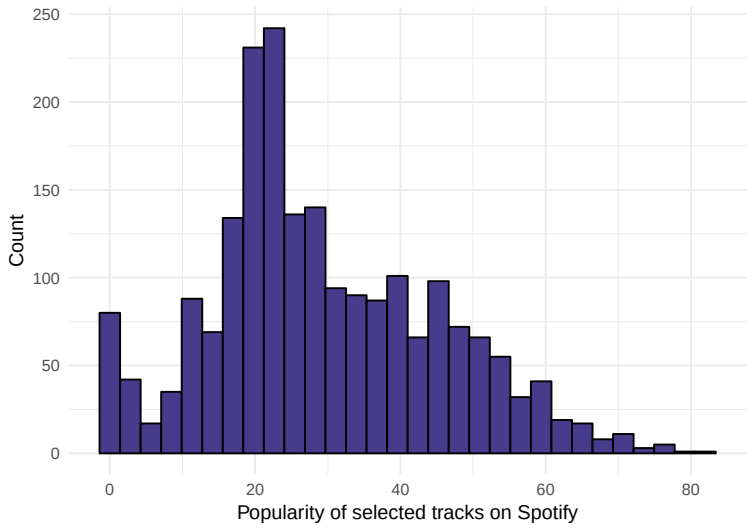


Figure 1: Histogram of *popularity* data.

## 4.2 Danceability Description

Figure 2 displays the histogram of the *danceability* data. The *danceability* variable was a continuous numerical variable. From inspection of Figure 2 is can be seen that the *danceability* data appeared to be unimodal with the peak at approximately 0.50. It can also be seen from Figure 2 that the *danceability* data appeared to be symmetrical. The *danceability* data was found to have a mean, median and standard deviation of 0.501, 0.499 and 0.174 respectively and a range of 0.000 to 0.963, as shown in Table 4.

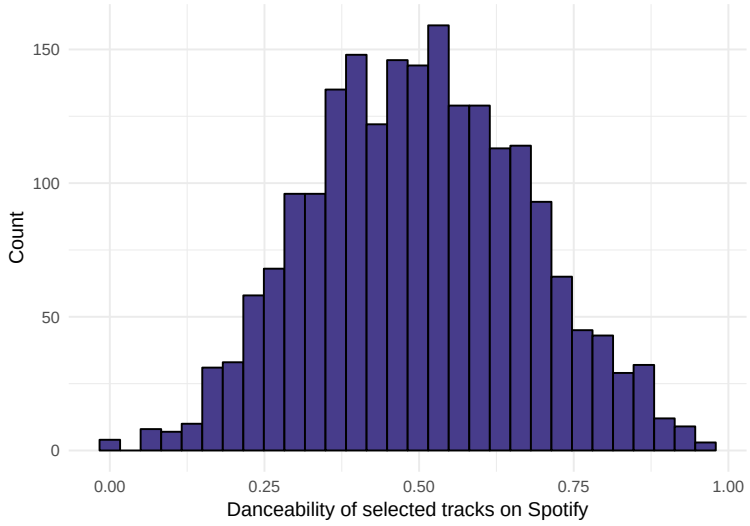


Figure 2: Histogram of *danceability* data.

## 4.3 Energy Description

Figure 3 shows the histogram of the *energy* data. The *energy* variable was a continuous numerical variable. From examining Figure 3 is can be seen that the *energy* data appeared to be slightly bimodal with the major peak and a potential minor peak at approximately 0.75 and 0.45 respectively. It can also be seen from Figure 3 that the *energy* data appeared to be left skewed. The *energy* data was found to have a mean, median and standard deviation of 0.596, 0.622 and 0.252 respectively and a range of 0.006 to 0.998, as shown in Table 4.

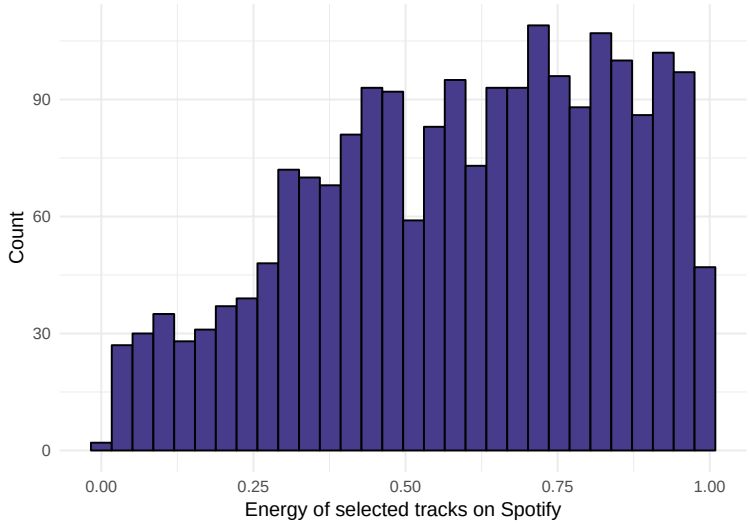


Figure 3: Histogram of *energy* data.

## 4.4 Loudness Description

Figure 4 depicts the histogram of the *loudness* data. The *loudness* variable was a continuous numerical variable. From the plot of the histogram in Figure 4 is can be seen that the *loudness* data appeared to be unimodal with the peak at approximately -6.0. Furthermore, it can also be seen from Figure 4 that the *loudness* data appeared to be left skewed. The *loudness* data was found to have a mean, median and standard deviation of -10.1, -9.3 and 4.5 respectively and a range of -30.6 to -0.9, as shown in Table 4.

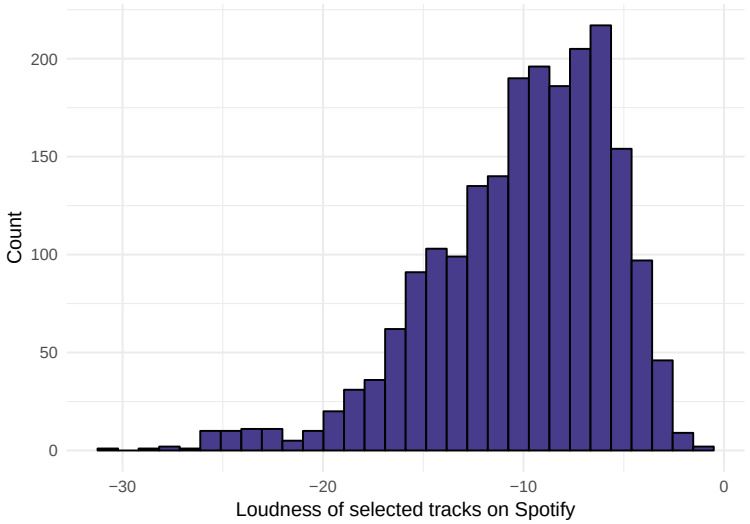


Figure 4: Histogram of *loudness* data.

## 4.5 Duration Description

Figure 5 displays the histogram of the *duration_ms* data. The *duration_ms* variable was a continuous numerical variable. From inspection of Figure 5 is can be seen that the *duration_ms* data appeared to be unimodal with the peak at approximately 200,000. It can also be seen from Figure 5 that the *duration_ms* data appeared to be right skewed. The *duration_ms* variable data was found to have a mean, median and standard deviation of 216,049, 203,640 and 95,712 respectively and a range of 13,213 to 1,187,253, as shown in Table 4.

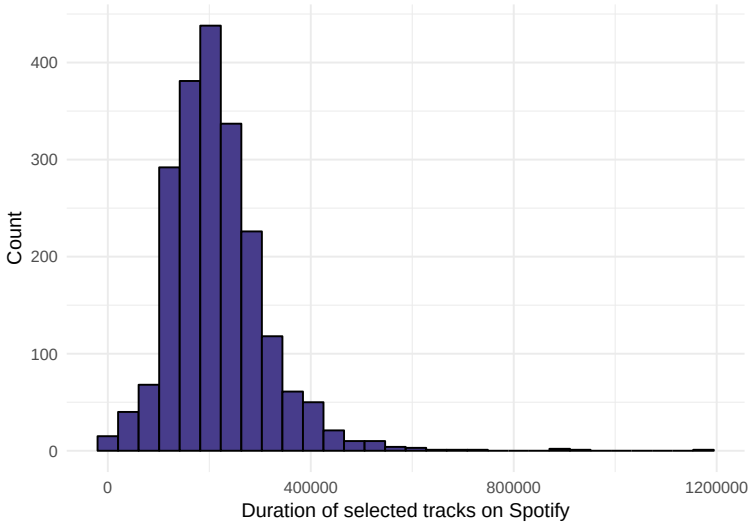


Figure 5: Histogram of *duration_ms* data.

## 4.6 Key Description

Figure 6 shows the bar chart of the *key* data. The *key* variable was an ordinal categorical variable. From examining Figure 6 is can be seen that the *key* data appeared to be slightly trimodal with the major peak at the C level, the first minor peak at the A level and the second minor peak at the G level. It can also be seen from Figure 6 that the *key* data appeared to be right skew. The *key* data was found to have a mean, median and standard deviation level count of 173.4, 152.0 and 92.9 respectively and a level count range of 44 for the D# level to 352 for the C level, as shown in Table 5.

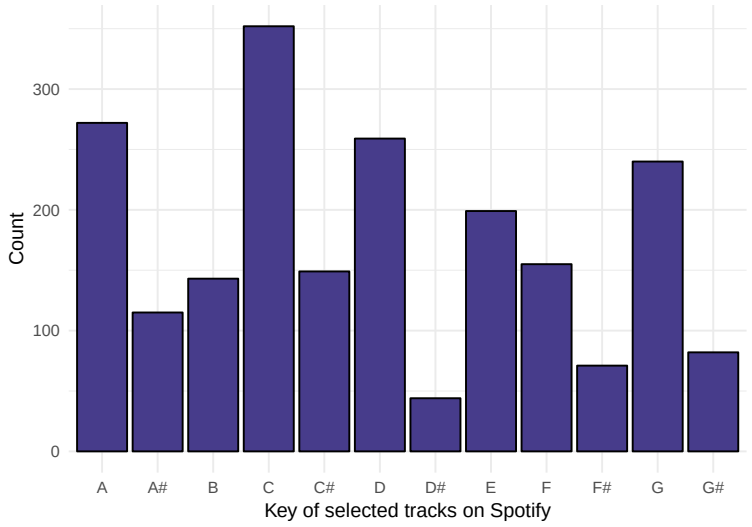


Figure 6: Bar chart of *key* data.

## 4.7 Mode Description

Figure 7 depicts the bar chart of the *mode* data. The *mode* variable was an nominal categorical variable. From the plot of the bar chart in Figure 7 is can be seen that the *mode* data appeared to be unimodal with the peak at the major level. The *mode* data was found to have a mean, median and standard deviation level count of 1040.5, 1040.5 and 816.7 respectively and a level count range of 463 for the minor level to 1618 for the major level, as shown in Table 5.

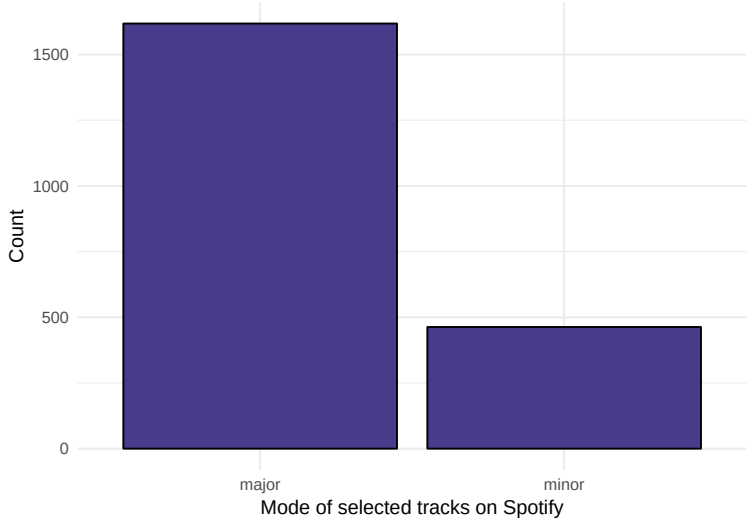


Figure 7: Bar chart of *mode* data.

## 4.8 Time Signature Description

Figure 8 displays the bar chart of the *time_signature* data. The *time_signature* variable was an ordinal categorical variable. From inspection of Figure 8 is can be seen that the *time_signature* data appeared to be unimodal with the peak at the 4 level. It can also be seen from Figure 8 that the *time_signature* data appeared to be left skewed. The *time_signature* data was found to have a mean, median and standard deviation level count of 346.8, 24.0 and 720.7 respectively and a range of zero for the 2 level to 1,808 for the 4 level, as shown in Table 5.

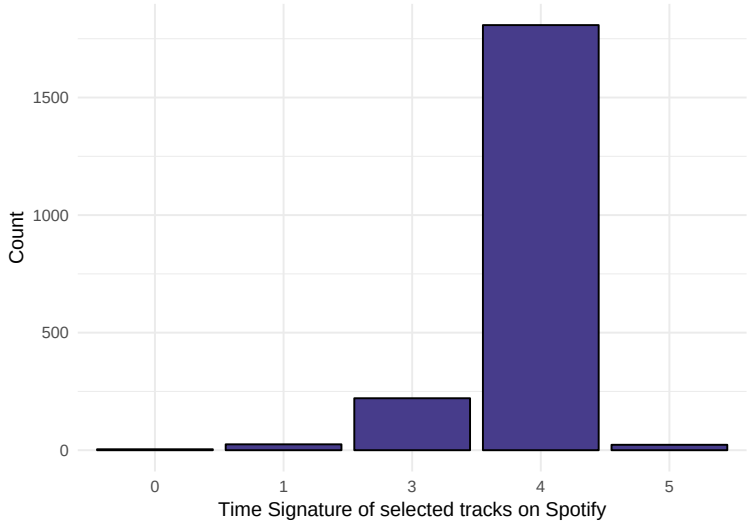


Figure 8: Bar chart of *time_signature* data.

## 4.9 Decade Description

Figure 9 shows the bar chart of the *decade* data. The *decade* variable was an ordinal categorical variable. From examining Figure 9 is can be seen that the *decade* data appeared to be bimodal with the major peak at the 50s level and the minor peak at the 70s level. It can also be seen from Figure 9 that the *decade* data appeared to be right skewed. The *decade* data was found to have a mean, median and standard deviation level count of 246.8, 250.0 and 198.4 respectively and a level count range of 191 for the 90s level to 648 for the 50s level, as shown in Table 5.

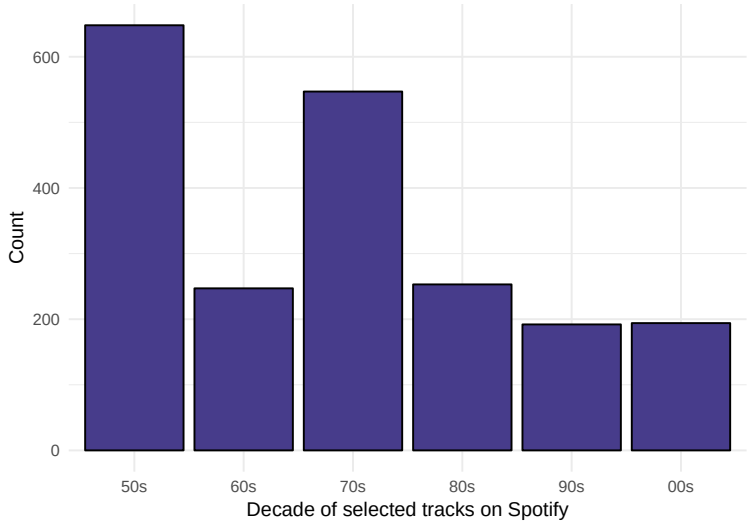


Figure 9: Bar chart of *decade* data.

# 5 Bivariate Analysis

For each of the predictors the appropriate response-predictor bivariate plot was generated so that nature and strength of the relationships between the response and predictors could be identified and characterised. Refer to Appendix A for the full `R` code that was used to plot the dataset. The *popularity* response was plotted against the *danceability*, *energy*, *loudness* and *duration_ms* predictors via scatter plots as these relationships were between a numerical response and a numerical predictor. On the other hand, the *popularity* response was plotted against the *key*, *mode*, *time_signature* and *decade* predictors via box plots as these relationships were between a numerical response and a categorical predictors.

## 5.1 Danceability Bivariate Analysis

Figure 10 displays the scatter plot of the *popularity* data against the *danceability* data. From Figure 10 it could be argued that there is no relationship between *popularity* and *danceability*, or at best, a very weak positive relationship. If a very weak positive relationship is put forward, then Figure 10 suggests that this relationship could potentially be more linear than non-linear. However, given the weakness of any potential relationship, it is very difficult to assess the linearity of the relationship. Finally, from Figure 10 it could be argued that the subjects which have a *danceability* value of less than 0.125 or an approximate zero *popularity* value ($popularity \leq 10$) are potential outlier candidates. On balance, it is was deemed more reasonable to argue that the scatter plot in Figure 10 suggests that there is no relationship between *popularity* and *danceability*, particularly when the relatively dense cluster of subjects in the almost perfectly horizontal region where $0.25 \leq danceability \leq 0.75$ and $0.15 \leq popularity \leq 0.30$ is taken into account.

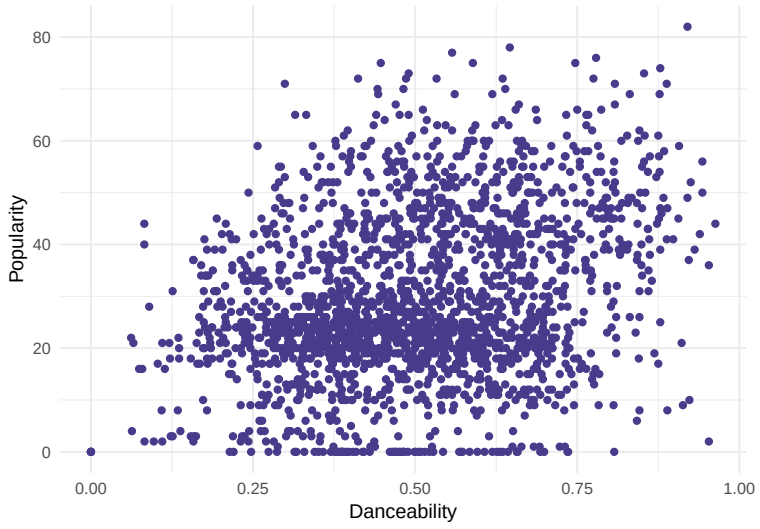


Figure 10: Scatter plot of the *popularity* data against the *danceability* data.

## 5.2 Energy Bivariate Analysis

Figure 11 depicts the scatter plot of *popularity* data against the *energy* data. Figure 11 suggests that there is no relationship between *popularity*and *energy*, or a very weak positive relationship. If a very weak positive relationship is posited, then Figure 11 suggests that this relationship could potentially be more linear than non-linear. However, given the weakness of any potential relationship, it is not obvious whether it is linear or non-linear. Finally, from Figure 11 it could be argued that the subjects which have a *energy* value of less than 0.125 or an approximate zero *popularity* value ($popularity \leq 10$) are potential outlier candidates. Overall, it is was determined more reasonable to propose that the scatter plot in Figure 11 is evidence of there being no relationship between *popularity* and *energy*, particularly when the relatively dense cluster of subjects in the almost perfectly horizontal region where $0.25 \leq energy \leq 1.0$ and $0.15 \leq popularity \leq 0.30$ considered.

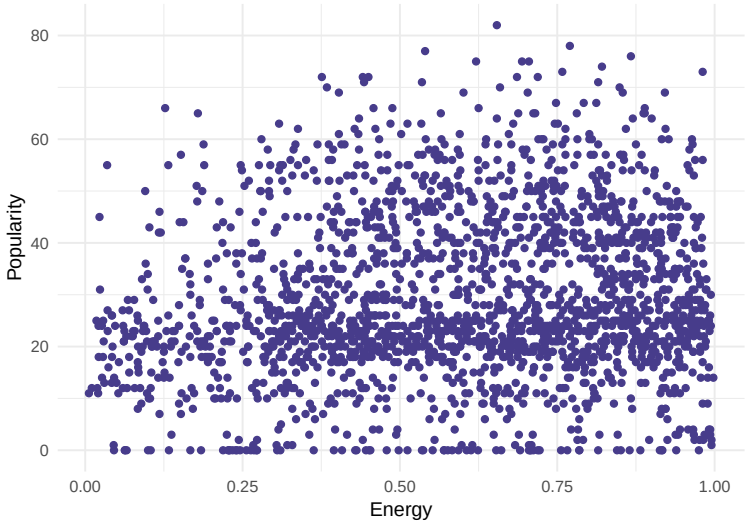


Figure 11: Scatter plot of the *popularity* data against the *energy* data.

## 5.3 Loudness Bivariate Analysis

Figure 12 shows the scatter plot of the *popularity* data against the *loudness* data. From inspection of Figure 12 it could be proposed that there is a very weak positive relationship between *popularity* and *loudness*, and that this relationship is more non-linear than linear. Furthermore, it could also be put forward from Figure 12 that the subjects which have a *loudness* value of less than -25 or an approximate zero *popularity* value ($popularity \leq 10$) are potential outlier candidates.

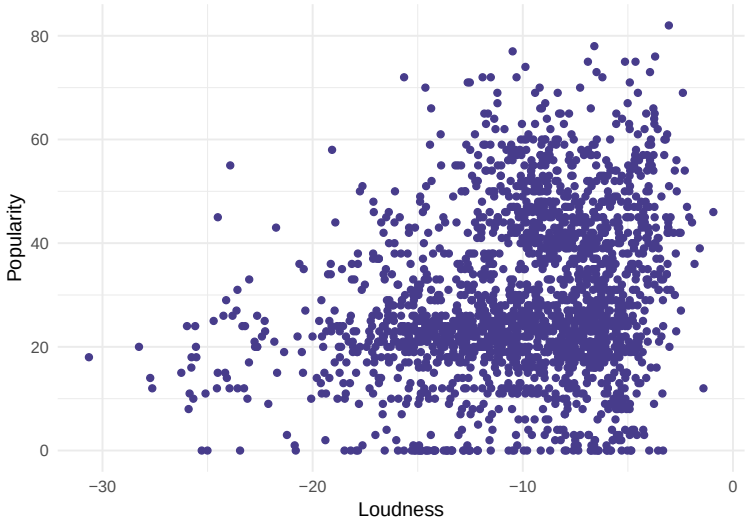


Figure 12: Scatter plot of the *popularity* data against the *loudness* data.

## 5.4 Duration Bivariate Analysis

Figure 13 displays the scatter plot of the *popularity* data against the *duration_ms* data. From Figure 13 it could be posited that there a moderate linear relationship between *popularity* and *duration_ms*, and that this relationship is highly positive. It could also be argued from Figure 13 that the subjects which have a *duration_ms* value that is either less than 90,000 (1.5 minutes) or greater than 750,000 (12.5 minutes), or an approximate zero *popularity* value ($popularity \leq 10$) are potential outlier candidates.

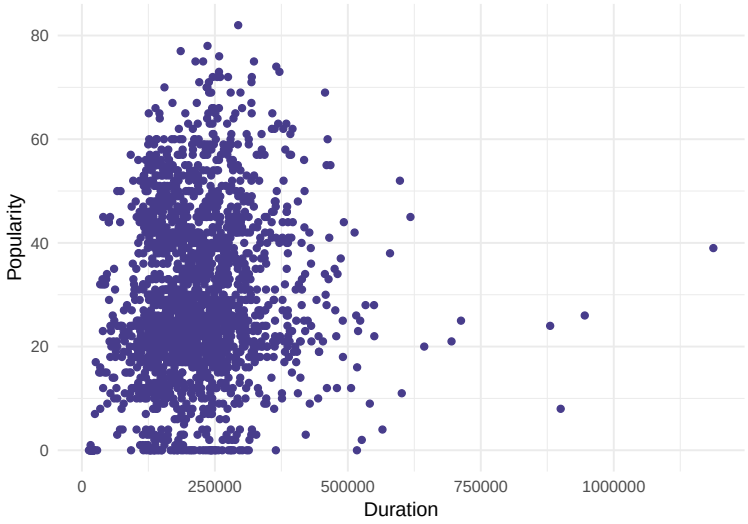


Figure 13: Scatter plot of the *popularity* data against the *duration_ms* data.

## 5.5 Key Bivariate Analysis

Figure 11 depicts the box plot of *popularity* data against the *key* data. From Figure 11 it can be seen that all of the *key* levels generally appeared to have very slight right skew distributions as their medians seemed to lie a little closer to their first quartiles than their third quartiles. However, it could be argued that the C# and D# levels are more symmetric than right skewed as their median seemed to lie approximately near the mid-points of their interquartile ranges. It can also be seen from Figure 11 that the medians of the various *key* levels were reasonably similar and lied towards the lower *popularity* values with the A# level recording the lowest *popularity* median of 23.0 and the F# level recording the highest *popularity* median of 32.0. Additionally, Figure 11 illustrates how the various *key* levels experienced relatively similar *popularity* spreads as the C level had the lowest *popularity* interquartile range of 19.0 while the G# level had the largest *popularity* interquartile range of 29.0. Finally, a number of *key* levels are show in Figure 11 to have potential outlier candidates with the A, A#, B, C, D, F and G levels having subjects lying above their maximum box plot range limits.

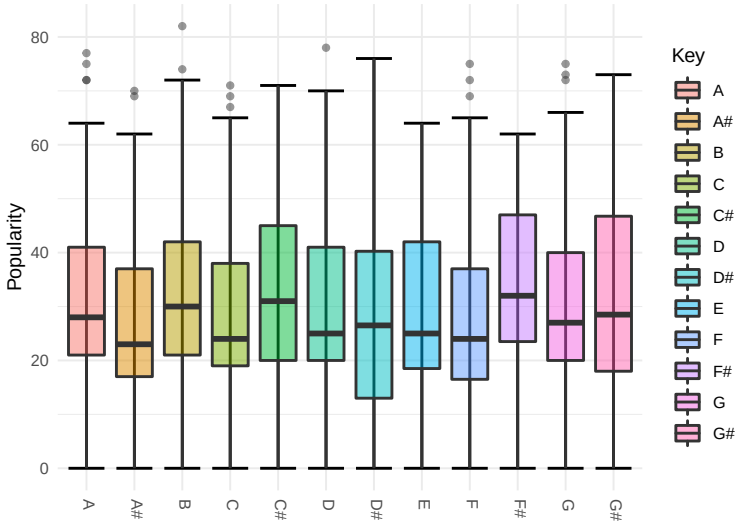


Figure 14: Box plot of the *popularity* data against the *key* data.

## 5.6 Mode Bivariate Analysis

Figure 15 shows the box plot of *popularity* data against the *mode* data. From Figure 15 it was observed that both *mode* levels had distributions that were right skewed as their medians were closer to their first quartiles than their third quartiles. It was also recognised from Figure 15 that the medians of both *mode* levels were reasonably similar and lied towards the lower *popularity* values with the major level recording the lower *popularity* median of 25.5 and the minor level recording the higher *popularity* median of 28.0. Furthermore, Figure 15 also depicts how the various *mode* levels experienced similar *popularity* spreads as the major level had the lower *popularity* interquartile range of 21.0 while the minor level had the larger *popularity* interquartile range of 22.0. Finally, both *mode* levels are show in Figure 15 to have potential outlier candidates with subjects lying above their maximum box plot range limits.

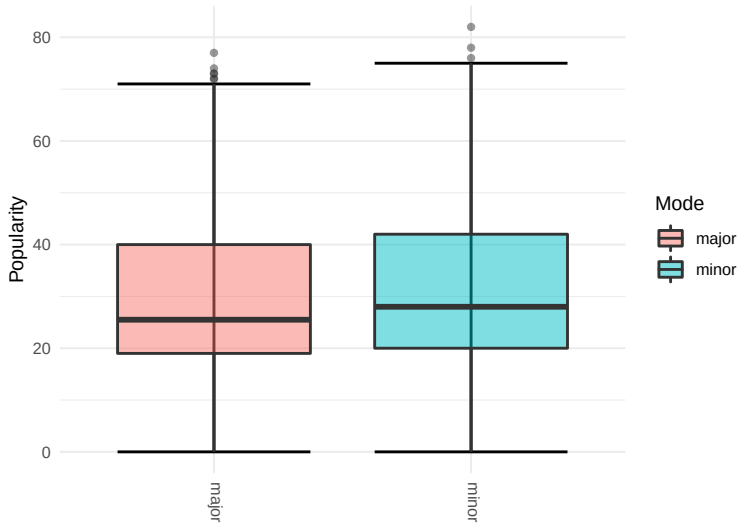


Figure 15: Box plot of the *popularity* data against the *mode* data.

## 5.7 Time Signature Bivariate Analysis

Figure 16 displays the box plot of the *popularity* data against the *time_signature* data. From Figure 16 it was found that the *time_signature* levels with non-zero *popularity* values generally had right skew distributions as their medians were generally closer to their first quartiles than their third quartiles. However, it could be argued that the 5 level had a symmetric distribution as its median seemed to lie approximately at the midpoint of its interquartile range. It was also determined from Figure 16 that the medians of the 1, 3, 4 and 5 levels were reasonably similar and lied towards the lower *popularity* values with the 1 and 5 levels recording the lowest *popularity* median of 21.0 and the 4 level recording the highest *popularity* median of 27.0. Additionally, Figure 16 demonstrates how the 1, 3, 4 and 5 levels experienced relatively similar *popularity* spreads as the 1 level had the lowest *popularity* interquartile range of 12.0 while the 4 level had the largest *popularity* interquartile range of 22.0. Finally, a number of *time_signature* levels are show in Figure 16 to have potential outlier candidates with the 1, 3 and 4 levels having subjects lying above their maximum box plot range limits. Furthermore, it can also be seen from Figure 16 that the 0 level only contained subjects of zero *popularity*. Hence, all of the 0 level subjects were also considered as being potential outlier candidates.

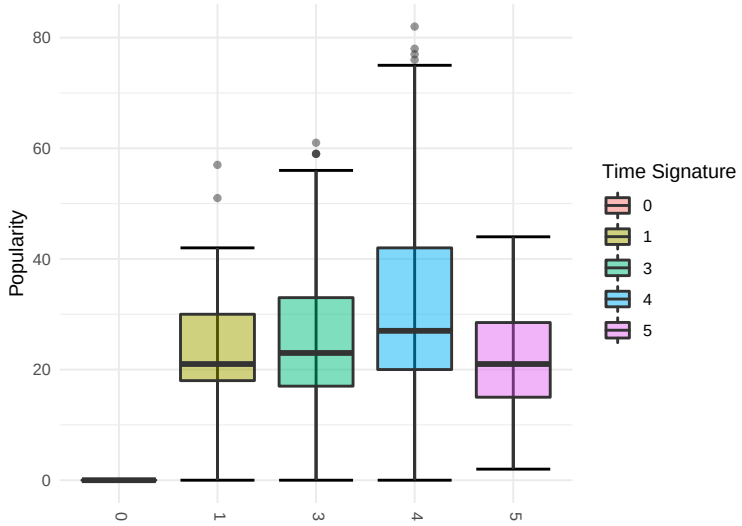


Figure 16: Box plot of the *popularity* data against the *time_signature* data.

## 5.8 Decade Bivariate Analysis

Figure 17 depicts the box plot of *popularity* data against the *decade* data. From Figure 17 it can be seen that all of the *decade* levels generally appeared to have symmetric distributions as their medians seemed to lie approximately near the mid-points of their interquartile ranges. However, it could also be argued that the 80s and 90s had slightly more left skewed distributions than symmetric distributions as their medians were a little closer to their third quartiles than their first quartiles, and similarly, the 70s level could be argued to be sightly more right skewed than symmetric as its median was a little closer to the first quartile than its third quartile. It can also be seen from Figure 17 that the medians of the various *decade* levels could vary considerably and tended to lie within the lower to mid *popularity* values with the 50s level recording the lowest *popularity* median of 21.0 and the 60s level recording the highest *popularity* median of 48.0. Additionally, Figure 17 illustrates how the various *decade* levels experienced differences in their respective *popularity* spreads as the 50s level had the lowest *popularity* interquartile range of 8.0 while the 00s level had the largest *popularity* interquartile range of 23.0. Finally, a number of *decade* levels are show in Figure 17 to have potential outlier candidates with the 50s and 60s levels having subjects lying both above and below their maximum and minimum box plot range limits, and the 70s and 80s levels having subjects lying above their maximum box plot range limits.

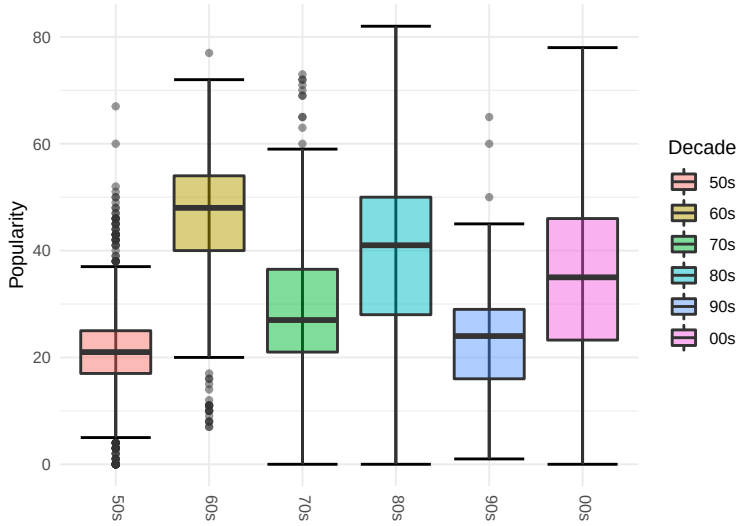


Figure 17: Box plot of the *popularity* data against the *decade* data.

## 5.9 Numeric-Categorical Interaction Bivariate Analysis

In addition to the overall bivariate analysis of the numeric predictors, a numeric-categorical interaction bivariate analysis was considered for every possible predictor numeric-categorical interaction combination. The bivariate analysis of a given numeric-categorical interaction was performed by partitioning the dataset into sub-dataset in accordance to the levels of the desired categorical predictor and then generating a separate scatter plot of the *popularity* response against the selected numeric predictor for each sub-dataset. For example, Figure 18 depicts the scatter plots of the *popularity* data against the *duration_ms* data for each *key* level. Refer to Appendix B for all the resulting scatter plots of every predictor numeric-categorical interaction combination. From the scatter plots in Appendix B it can be seen that all of the numerical predictors had observable relationships with *popularity* of varying strengths and characters when partitioned into the levels of the categorical predictors, particularly the *decade* levels. For example it was observed from Figure 18 that the strengths and magnitudes of the positive linear relationships between the *popularity* and the *duration_ms* were different across the various *decade* levels. Hence, the model fitting process was conducted in a manner that allowed for the incorporation of interactions between the various predictors.

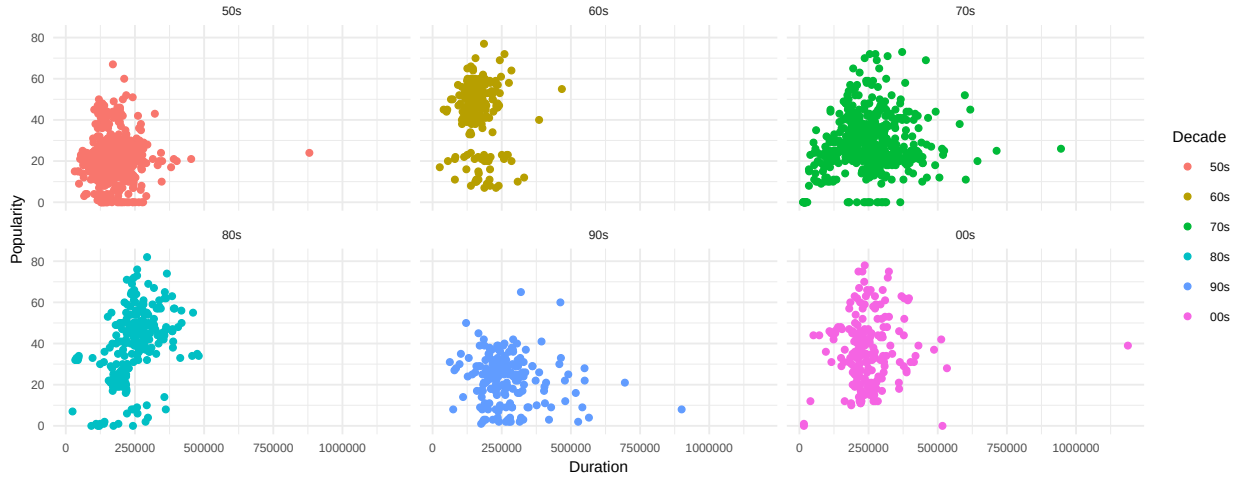


Figure 18: Scatter plots of the *popularity* data against the *duration_ms* data for each *decade* level.

## 5.10 Potential Outliers

The individual subjects which were identified from the bivariate analysis as being outlier candidates were individually inspected to discern whether they represented conventional songs. From manual inspection of the candidate outlier subjects a clear trend emerged in that many of the subjects corresponded to tracks that were commentaries, voice overs, interviews or narrations. Consequently, these tracks were determined to represent non-conventional songs, and were deemed irrelevant for the purposes of fitting a predictive model of song popularity. Hence, all subjects in the dataset which represented commentaries, voice overs, interviews or narrates were identified as being outliers and flagged for potential removal. Using this approach, 63 subjects that were deemed as being outliers. Refer Appendix C for the `R` code that was used to identify the outliers and to Appendix D for the tables describing the outlier subjects.

# 6 Model Fitting

A predictive model for song popularity was developed by fitting numerous linear regression models for the response *popularity* against various predictor combinations via the least squares approach. During the model fitting process the considered linear regression models were drawn from the following model scopes:

- **Reduced Scope**: linear models were constructed from the following predictor combinations while maintaining the principle of marginality:
  1. linear numeric variables;
  2. categorical variables;
  3. linear numeric-categorical interaction variables;
  4. linear numeric-numeric interaction variables; and
  5. categorical-categorical interaction variables.
- **Expanded Scope**: linear models were constructed from the following predictor combinations while maintaining the principle of marginality:
  1. linear, quadratic and cubic numeric variables;
  2. categorical variables;
  3. linear and quadratic numeric-categorical interaction variables;
  4. linear and quadratic numeric-numeric interaction variables; and
  5. categorical-categorical interaction variables.

During the bivariate analysis it was noted that some of the predictors were suspected of having nonlinear relationships with the *popularity*. Hence, the Expanded Scope was considered for the model fitting process as it included quadratic and cubic terms which were deemed capable of capturing the nonlinear relationships. However, the linear models derived from the Expanded Scope often contained a large number of terms, and as such, increased the risk of producing an overfitted model. Consequently, the Reduced Scope was also considered as it only focused on the linear relationship between the response and predictors, and as such, helped to produce models with few terms and lower risks of overfitting the dataset. The following `R` code was employed to define the Reduced and Expanded Scopes:

```
red <- popularity ~ (danceability + energy + loudness + duration_ms + key +
                     mode + time_signature + decade)^2

exp <- popularity ~ ((danceability + energy + loudness + duration_ms
                      + I(danceability^2) + I(energy^2) + I(loudness^2) + I(duration_ms^2))
                     * (key + mode + time_signature + decade)
                     + (danceability + energy + loudness + duration_ms
                     + I(danceability^2) + I(energy^2) + I(loudness^2) + I(duration_ms^2))^2
                     + (key + mode + time_signature + decade)^2))
```

Three separate stepwise regression algorithms were applied to the Reduced and Expanded Scopes during the model fitting process. The three algorithms were the Backwards Elimination, Forward Selection and Stepwise Selection procedures. All three of these algorithms were employed as they explored the model space differently, and as a consequence, tended to produce different linear models. The `step()` function was utilised in `R` to implement each of the algorithms with the Akaike Information Criterion estimator (AIC) being used as their heuristic. The AIC was deemed to be the appropriate heuristic as it favoured models of higher goodness of fit levels while penalising larger models of increasing overfitting risks. Refer to Appendix E for the full `R` code that was used to implement the model fitting process.

## 6.1 Backward Elimination Model Fitting

The Backwards Elimination procedure was performed on both the Reduced and Expanded Scopes. The algorithm begins with fitting a model of greatest scope (a model which corresponds to the whole Reduced/Extended Scope) and then searches for the single term, which when removed from the current fitted model, results in a new fitted model of lowest AIC value. This new fitted model of lowest AIC now becomes the current model

and the process is then repeated. The algorithm terminates when the current model records an AIC value that is lower than any of the fitted models that were constructed by removing any one of its terms or when the current model no long contains any predictor terms. The following R code was utilised to perform the Backwards Elimination procedure for both the Reduced and Expanded Scopes:

```
# backwards elimination on reduced scope ----
backwards.red <- lm(red, data = data)
backwards.red <- step(backwards.red, direction = "backward")

# backwards elimination on expanded scope ----
backwards.exp <- lm(exp, data = data)
backwards.exp <- step(backwards.exp, direction = "backward")
```

### 6.1.1 Reduced Scope Backward Elimination Model Fitting

Table 6 shows the significant least squares coefficient estimates and their corresponding P-values for the final model that was selected via the Backward Elimination process on the Reduced Scope (the Backwards Reduction Model). The Backwards Reduction Model could be represented in R code with the following lm object containing 18 predictor combination terms:

```
popularity ~ danceability + energy + key + loudness + mode + duration_ms + time_signature + decade
             + danceability:key + energy:loudness + energy:duration_ms + energy:decade + key:mode
             + loudness:decade + mode:duration_ms + mode:time_signature + mode:decade + duration_ms:decade
```

From inspection of Table 6 it can be seen that of the 75 coefficients 19 where significant at the $\alpha = 5\%$ level where their P-values ranged from $3.89 \times 10^{-7}$ ($energy : duration\_ms$) to 0.033 ($keyD\#$). The Backwards Reduction Model had AIC and Bayesian Information Criterion (BIC) values of 16504 and 16933 respectively, and a F-statistic of 17.38 which corresponded to a P-value that was less than $2.2 \times 10^{-16}$. Hence, the Backwards Reduction Model was significant at the $\alpha = 5\%$ level.

Table 6: Summary of the significant coefficients of the Backwards Reduction Model.

| **Significant Term** | **Coef** | **P-val** |
|---|---|---|
| *danceability* | 12.60 | 0.011452 |
| *duration_ms* | 4.819e-05 | 2.50e-05 |
| *D#* | -12.77 | 0.033836 |
| *time_signature*1 | 22.15 | 0.024250 |
| *time_signature*3 | 24.09 | 0.009855 |
| *time_signature*4 | 24.17 | 0.009508 |
| *time_signature*5 | 22.15 | 0.022331 |
| 60*s* | 40.02 | 4.31e-06 |
| *danceability* : *F#* | -28.16 | 0.001831 |
| *energy* : *loudness* | -0.7482 | 0.006453 |
| *energy* : *duration_ms* | -6.877e-05 | 3.89e-07 |
| *energy* : 00*s* | 20.27 | 0.007985 |
| *loudness* : 80*s* | 1.149 | 0.002054 |
| *duration_ms* : *minor* | 2.102e-05 | 0.005972 |
| *duration_ms* : 80*s* | 3.560e-05 | 0.019962 |
| *duration_ms* : 90*s* | -2.929e-05 | 0.016125 |
| *keyC#* : *minor* | 11.37 | 0.000284 |
| *keyF* : *minor* | 8.737 | 0.008131 |
| *keyF#* : *minor* | 11.68 | 0.001452 |

Coef = Coefficient least squares estimate; P-val = P-value;
Terms significant at the $\alpha = 5\%$ level

### 6.1.2 Expanded Scope Backward Elimination Model Fitting

Table 7 shows the significant least squares coefficient estimates and their corresponding P-values for the final model that was selected via the Backward Elimination process on the Expanded Scope (the Backwards Expansion Model). The Backwards Expansion Model could be represented in R code with the following lm object containing 37 predictor combination terms:

```
popularity ~ danceability + energy + loudness + duration_ms + I(danceability^2) + I(energy^2) + I(loudness^2)
             + I(duration_ms^2) + mode + decade + danceability:mode + danceability:decade +  energy:decade
             + loudness:decade + duration_ms:mode + duration_ms:decade + I(danceability^2):mode
             + I(danceability^2):decade + I(energy^2):decade + I(loudness^2):decade + I(duration_ms^2):mode
             + I(duration_ms^2):decade + danceability:loudness + danceability:duration_ms
             + danceability:I(duration_ms^2) + energy:loudness + energy:I(loudness^2) + loudness:duration_ms
             + loudness:I(danceability^2) + loudness:I(energy^2) + duration_ms:I(danceability^2)
             + I(duration_ms^3) + I(danceability^2):I(loudness^2) + I(danceability^2):I(duration_ms^2)
             + I(energy^2):I(loudness^2) + I(energy^2):I(duration_ms^2) + I(loudness^2):I(duration_ms^2)
```

From inspection of Table 7 it can be seen that of the 74 coefficients 32 where significant at the $\alpha = 5\%$ level where their P-values ranged from $1.14 \times 10^{-8}$ ($loudness : 80s$) to 0.044 ($duration\_ms$). The Backwards Expansion Model had AIC and BIC values of 16440 and 16863 respectively, and a F-statistic of 18.98 which corresponded to a P-value that was less than $2.2 \times 10^{-16}$. Hence, the Backwards Expansion Model was significant at the $\alpha = 5\%$ level.

Table 7: Summary of the significant coefficients of the Backwards Expansion Model.

| **Significant Term** | **Coef** | **P-val** |
|---|---|---|
| $energy$ | -97.13 | 0.032324 |
| $duration\_ms$ | 1.257e-04 | 0.044471 |
| $danceability^2$ | -127.4e | 0.018680 |
| $energy^2$ | 86.39 | 0.008977 |
| $duration\_ms^2$ | -2.616e-10 | 0.019536 |
| $minor$ | -16.23 | 0.002120 |
| $80s$ | 69.37 | 6.53e-05 |
| $00s$ | -41.82 | 0.017674 |
| $danceability : 60s$ | 77.02 | 0.035625 |
| $danceability : 90s$ | 84.08 | 0.009889 |
| $energy : 00s$ | 107.9 | 30.001637 |
| $loudness : 80s$ | 7.232 | 1.14e-08 |
| $duration\_ms : minor$ | 7.734e-05 | 0.000430 |
| $duration\_ms : 70s$ | 7.349e-05 | 0.005608 |
| $duration\_ms : 00s$ | 7.886e-05 | 0.021805 |
| $danceability^2 : 60s$ | -85.00 | 0.021015 |
| $danceability^2 : 90s$ | -91.51 | 0.008521 |
| $energy^2 : 00s$ | -70.827 | 0.008191 |
| $loudness^2 : 80s$ | 0.2397 | 1.06e-05 |
| $duration\_ms^2 : minor$ | -1.160e-10 | 0.004306 |
| $duration\_ms^2 : 70s$ | -1.197e-10 | 0.019508 |
| $duration\_ms^2 : 00s$ | -1.150e-10 | 0.014245 |
| $danceability : duration\_ms$ | -5.521e-04 | 0.029400 |
| $energy : loudness$ | -12.82 | 0.011134 |
| $loudness : energy^2$ | 9.774 | 0.020191 |
| $loudness : danceability^2$ | -9.830 | 0.000479 |
| $loudness : energy^2$ | 9.774 | 0.020191 |
| $duration\_ms : danceability^2$ | 6.854e-04 | 0.020557 |
| $duration\_ms^3$ | 2.604e-16 | 8.59e-05 |
| $danceability^2 : loudness^2$ | -2.070e-01 | 0.014425 |
| $energy^2 : duration\_ms^2$ | -9.151e-11 | 0.006593 |
| $loudness^2 : duration\_ms^2$ | -5.106e-13 | 0.014702 |

Coef = Coefficient least squares estimate; P-val = P-value;
Terms significant at the $\alpha = 5\%$ level

## 6.2 Forward Selection Model Fitting

The Forward Selection procedure was performed on both the Reduced and Expanded Scopes. The algorithm begins with fitting the null model of smallest scope (a model which contains a single constant/intercept term) and then searches for the single term from the model scope, which when added to the current fitted model, results in a new fitted model of lowest AIC value. This new fitted model of lowest AIC now becomes the current model and the process is then repeated. The algorithm terminates when the current model records an AIC

value that is lower than any the fitted models that were constructed by adding any one of the remaining terms available from the model scope or when the current model contains the whole model scope. The following R code was utilised to perform the Forward Selection procedure for both the Reduced and Expanded Scopes:

```
# define null model scopes for model fitting ----
null <- popularity ~ 1

# forward selection on reduced scope ----
forwards.red <- lm(null, data = data)
forwards.red <- step(forwards.red, scope = red, direction = "forward")

# forward selection on expanded scope ----
forwards.exp <- lm(null, data = data)
forwards.exp <- step(forwards.exp, scope = exp, direction = "forward")
```

### 6.2.1 Reduced Scope Forward Selection Model Fitting

Table 8 shows the significant least squares coefficient estimates and their corresponding P-values for the final model that was selected via the Forward Selection process on the Reduced Scope (the Forward Reduction Model). The Forward Reduction Model could be represented in R code with the following lm object containing 11 predictor combination terms:

```
popularity ~ decade + danceability + duration_ms + loudness + mode + decade:duration_ms
             + duration_ms:loudness + decade:loudness + danceability:duration_ms + duration_ms:mode
             + decade:mode
```

From inspection of Table 8 it can be seen that of the 28 coefficients 15 where significant at the $\alpha = 5\%$ level where their P-values ranged from $8.93 \times 10^{-12}$ ($60s$) to 0.036 ($90s$). The Forward Reduction Model had AIC and BIC values of 16517 and 16680 respectively, and a F-statistic of 42.51 which corresponded to a P-value that was less than $2.2 \times 10^{-16}$. Hence, the Forward Reduction Model was significant at the $\alpha = 5\%$ level.

Table 8: Summary of the significant coefficients of the Forward Reduction Model.

| **Significant Term** | **Coef** | **P-val** |
|---|---|---|
| *intercept* | 19.29 | 3.95e-09 |
| $60s$ | 29.96 | 8.93e-12 |
| $70s$ | 10.46 | 0.000843 |
| $80s$ | 17.99 | 0.000610 |
| $90s$ | 8.236 | 0.035766 |
| $00s$ | 19.44 | 7.26e-06 |
| *danceability* | 16.58 | 8.62e-05 |
| *duration_ms* | -3.093e-05 | 0.028773 |
| *loudness* | 0.6050 | 0.000450 |
| $80s$ : *duration_ms* | 4.997e-05 | 0.001974 |
| *duration_ms* : *loudness* | -4.163e-06 | 1.71e-08 |
| $70s$ : *loudness* | 0.3.974 | 0.021046 |
| $80s$ : *loudness* | 1.418 | 1.13e-07 |
| $00s$ : *loudness* | 0.8498 | 0.009328 |
| *duration_ms* : *minor* | 2.023e-05 | 0.005410 |

Coef = Coefficient least squares estimate; P-val = P-value;
Terms significant at the $\alpha = 5\%$ level

### 6.2.2 Expanded Scope Forward Selection Model Fitting

Table 9 shows the significant least squares coefficient estimates and their corresponding P-values for the final model that was selected via the Forward Selection process on the Expanded Scope (the Forward Expansion Model). The Forward Expansion Model could be represented in R code with the following lm object containing 14 predictor combination terms:

```
popularity ~ decade + danceability + duration_ms + I(duration_ms^2) + I(loudness^2) + loudness
             + decade:duration_ms + I(duration_ms^3) + decade:I(loudness^2) + decade:loudness + I(loudness^3)
             + duration_ms:loudness + I(duration_ms^2):I(loudness^2) + danceability:duration_ms
```

From inspection of Table 9 it can be seen that of the 31 coefficients 15 where significant at the $\alpha = 5\%$ level where their P-values ranged from $5.63 \times 10^{-10}$ ($80s : loudness$) to 0.045 ($60s : duration\_ms$). The Forward Expansion Model had AIC and BIC values of 16464 and 16644 respectively, and a F-statistic of 41.26 which corresponded to a P-value that was less than $2.2 \times 10^{-16}$. Hence, the Forward Expansion Model was significant at the $\alpha = 5\%$ level.

Table 9: Summary of the significant coefficients of the Forward Expansion Model.

| **Significant Term** | **Coef** | **P-val** |
|---|---|---|
| *intercept* | 19.29 | 3.95e-09 |
| $60s$ | 27.42 | 0.000199 |
| $70s$ | 5.549 | 0.281450 |
| $80s$ | 37.48 | 1.49e-07 |
| $90s$ | 8.236 | 0.035766 |
| $00s$ | 25.66 | 3.84e-05 |
| *danceability* | 15.67 | 0.000317 |
| *duration_ms* | 3.119e-05 | 0.432847 |
| *loudness* | -2.614 | 0.018979 |
| $80s$ : *duration_ms* | 5.801e-05 | 0.000262 |
| *duration_ms* : *loudness* | -6.176e-06 | 0.014969 |
| $70s$ : *loudness* | 3.974e-01 | 0.021046 |
| $80s$ : *loudness* | 6.559 | 5.63e-10 |
| $00s$ : *loudness* | 8.498e-01 | 0.009328 |
| *duration_ms* : *minor* | 2.023e-05 | 0.005410 |
| Coef = Coefficient least squares estimate; P-val = P-value; Terms significant at the $\alpha = 5\%$ level | | |

## 6.3 Stepwise Selection Model Fitting

The Stepwise Selection procedure was performed on both the Reduced and Expanded Scopes. The algorithm begins with fitting the null model of smallest scope and then performs one step of the Forwards Selection procedure followed by one step of the Backwards Elimination procedure. The algorithm terminates when the current model records an AIC value that is lower than any the fitted models that were derived from it on any forward or backward iteration, or when the current model covers either the entire model or null scope. The following `R` code was utilised to perform the Stepwise Selection procedure for both the Reduced and Expanded Scopes:

```
# stepwise selection on reduced scope ----
stepwise.red <- lm(null, data = data)
stepwise.red <- step(stepwise.red, scope = red, direction = "both")

# stepwise selection on expanded scope ----
stepwise.exp <- lm(null, data = data)
stepwise.exp <- step(stepwise.exp, scope = exp, direction = "both")
```

### 6.3.1 Reduced Scope Stepwise Selection Model Fitting

Table 10 shows the significant least squares coefficient estimates and their corresponding P-values for the final model that was selected via the Stepwise Selection process on the Reduced Scope (the Stepwise Reduction Model). The Stepwise Reduction Model could be represented in `R` code with the following `lm` object containing 11 predictor combination terms:

```
popularity ~ decade + danceability + duration_ms + loudness + mode + decade:duration_ms + duration_ms:loudness
             + decade:loudness + danceability:duration_ms + duration_ms:mode + decade:mode
```

From inspection of Table 10 it can be seen that of the 28 coefficients 15 where significant at the $\alpha = 5\%$ level where their P-values ranged from $8.93 \times 10^{-12}$ ($60s$) to 0.036 ($90s$). The Stepwise Reduction Model had AIC and BIC values of 16517 and 16680 respectively, and a F-statistic of 42.51 which corresponded to a P-value that was less than $2.2 \times 10^{-16}$. Hence, the Stepwise Reduction Model was significant at the $\alpha = 5\%$ level. Finally, it was observed that the Stepwise Selection and Forward Selection procedures selected the same final model on the Reduced Scope.

Table 10: Summary of the significant coefficients of the Stepwise Reduction Model.

| **Significant Term** | **Coef** | **P-val** |
|---|---|---|
| *intercept* | 19.29 | 3.95e-09 |
| $60s$ | 29.96 | 8.93e-12 |
| $70s$ | 10.46 | 0.000843 |
| $80s$ | 17.99 | 0.000610 |
| $90s$ | 8.236 | 0.035766 |
| $00s$ | 19.44 | 7.26e-06 |
| *danceability* | 16.58 | 8.62e-05 |
| *duration_ms* | -3.093e-05 | 0.028772 |
| *loudness* | 0.6050 | 0.000450 |
| $80s$ : *duration_ms* | 4.997e-05 | 0.001974 |
| *duration_ms* : *loudness* | -4.163e-06 | 1.71e-08 |
| $70s$ : *loudness* | 0.3974 | 0.021046 |
| $80s$ : *loudness* | 1.418 | 1.13e-07 |
| $00s$ : *loudness* | 0.8498 | 0.009328 |
| *duration_ms* : *minor* | 2.023e-05 | 0.005410 |
| Coef = Coefficient least squares estimate; P-val = P-value; Terms significant at the $\alpha = 5\%$ level | | |

### 6.3.2 Expanded Scope Stepwise Selection Model Fitting

Table 11 shows the significant least squares coefficient estimates and their corresponding P-values for the final model that was selected via the Stepwise Selection process on the Expanded Scope (the Stepwise Expansion Model). The Stepwise Expansion Model could be represented in `R` code with the following `lm` object containing 14 predictor combination terms:

```
popularity ~ decade + danceability + duration_ms + I(duration_ms^2) + I(loudness^2) + loudness
             + decade:duration_ms + I(duration_ms^3) + decade:I(loudness^2) + decade:loudness + I(loudness^3)
             + duration_ms:loudness + I(duration_ms^2):I(loudness^2) + danceability:duration_ms
```

From inspection of Table 11 it can be seen that of the 31 coefficients 15 where significant at the $\alpha = 5\%$ level where their P-values ranged from $5.63 \times 10^{-10}$ ($80s$ : *loudness*) to 0.045 ($60s$ : *duration_ms*). The Stepwise Expansion Model had AIC and BIC values of 16464 and 16644 respectively, and a F-statistic of 41.26 which corresponded to a P-value that was less than $2.2 \times 10^{-16}$. Hence, the Stepwise Expansion Model was significant at the $\alpha = 5\%$ level. Finally, it was observed that the Stepwise Selection and Forward Selection procedures selected the same final model on the Expanded Scope, as was observed for the Reduced Scope.

## 6.4 Model Comparisons

Table 12 displays the comparison of the final models that were selected from performing the various stepwise regression algorithms on the Reduced and Expanded scopes. From inspection of the results contained in Table 12 it can be seen that the Forward and Stepwise Expansion Models, which both represented the same model, had the lowest BIC value (16644 from a range of [16644, 16932]), second lowest AIC value (16464 from a range of [16440, 16517]), second highest F-statistic (41.26 from a range of [17.38, 42.51]) and second lowest number of terms (31 from a range of [28, 75]). As none of the other models were deemed to have favourable metrics values as consistently as the Forward and Stepwise Expansion Models, these models were determined to have the best trade-off between simplicity and goodness of fit, and as such, were selected for overall final model (the Final Model).

Table 11: Summary of the significant coefficients of the Stepwise Expansion Model.

| **Significant Term** | **Coef** | **P-val** |
|---|---|---|
| *intercept* | 19.29 | 3.95e-09 |
| $60s$ | 27.42 | 0.000199 |
| $70s$ | 5.549 | 0.281450 |
| $80s$ | 37.48 | 1.49e-07 |
| $90s$ | 8.236 | 0.035766 |
| $00s$ | 25.66 | 3.84e-05 |
| *danceability* | 15.67 | 0.000317 |
| *duration_ms* | 3.119e-05 | 0.432847 |
| *loudness* | -2.614 | 0.018979 |
| $80s$ : *duration_ms* | 5.801e-05 | 0.000262 |
| *duration_ms* : *loudness* | -6.176e-06 | 0.014969 |
| $70s$ : *loudness* | 3.974e-01 | 0.021046 |
| $80s$ : *loudness* | 6.559 | 5.63e-10 |
| $00s$ : *loudness* | 8.498e-01 | 0.009328 |
| *duration_ms* : *minor* | 2.023e-05 | 0.005410 |

Coef = Coefficient least squares estimate; P-val = P-value;
Terms significant at the $\alpha = 5\%$ level

Table 12: Summary statistics for the final models selected from the fitting process.

| **Model** | **Terms** | **RSE** | **F-Stat** | **ACI** | **BIC** |
|---|---|---|---|---|---|
| Backwards Reduction | 75 | 12.53 | 17.38* | 16504 | 16932 |
| Backwards Expansion | 74 | 12.34 | 18.98* | 16440 | 16863 |
| Forward Reduction | 28 | 12.71 | 42.51* | 16517 | 16680 |
| Forward Expansion | 31 | 12.54 | 41.26* | 16464 | 16644 |
| Stepwise Reduction | 28 | 12.71 | 42.51* | 16517 | 16680 |
| Stepwise Expansion | 31 | 12.54 | 41.26* | 16464 | 16644 |

Terms = Number of Terms; RSE =Residual standard error; F-Stat = F-statistic;
AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion;
* = Significant at the $\alpha = 5\%$ level

# 7 Final Model

Table 13 summarises and interprets each of the terms contained in the Final Model. From inspection of Table 13 it can be seen that a number of terms with insignificant coefficient estimates at the $\alpha = 5\%$ level were present in the Final Model. These terms were included in the Final Model to satisfy the principle of marginality as their associated higher-order or interaction terms were significant at the $\alpha = 5\%$ level.

Table 13: Summary of the coefficients of the final model.

| **Term** | **Coeff.** | **P-val** | **Interpretation** |
|---|---|---|---|
| *intercept* | -1.764 | 0.802772 | If *danceability* = *duration_ms* = *loudness* = 0 and *decade* is 50*s* then E[*popularity*] = -1.764. No real-world meaning. |
| 60*s* | 27.42 | 0.000199* | If 60*s* then, all else equal, E[*popularity*] increases by 27.42 relative to 50*s*. |
| 70*s* | 5.549 | 0.281450 | If 70*s* then, all else equal, E[*popularity*] increases by 5.549 relative to 50*s*. |
| 80*s* | 37.48 | 1.490e-07* | If 80*s* then, all else equal, E[*popularity*] increases by 37.48 relative to 50*s*. |
| 90*s* | -4.643e-02 | 0.994831 | If 90*s* then, all else equal, E[*popularity*] increases by -0.046 relative to 50*s*. |
| 00*s* | 25.66 | 3.840e-05* | If 00*s* then, all else equal, E[*popularity*] increases by 25.66 relative to 50*s*. |
| *danceability* | 15.67 | 0.000317* | E[*popularity*] increases/decreases by 15.67 for an increase/decrease in *danceability* of 1, all else equal. |
| *duration_ms* | 3.119e-05 | 0.432847 | If 50s, E[*popularity*] increases/decreases by 3.119e-05 for an increase/decrease in *duration_ms* of 1, all else equal.$^+$ |
| $duration\_ms^2$ | -1.667e-10 | 0.013658* | E[*popularity*] decreases/increases by 1.667e-10 for an increase/decrease in $duration\_ms^2$ of 1, all else equal.$^+$ |
| $loudness^2$ | -0.222 | 0.002245* | If 50s, E[*popularity*] decreases/increases by 0.222 for an increase/decrease in $loudness^2$ of 1, all else equal.$^+$ |
| *loudness* | -2.614 | 0.018979* | If 50s, E[*popularity*] decreases/increases by 2.614 for an increase/decrease in *loudness* of 1, all else equal.$^+$ |
| 60*s* : *duration_ms* | -3.624e-05 | 0.045086* | If 60s then, all else equal, *duration_ms* coefficient decreases by 3.624e-05 relative to 50s.$^+$ |
| 70*s* : *duration_ms* | 1.917e-05 | 0.053322 | If 70s then, all else equal, *duration_ms* coefficient increase by 1.917e-05 relative to 50s.$^+$ |
| 80*s* : *duration_ms* | 5.801e-05 | 0.000262* | If 80s then, all else equal, *duration_ms* coefficient increases by 5.801e-05 relative to 50s.$^+$ |
| 90*s* : *duration_ms* | -3.681e-06 | 0.771648 | If 90s then, all else equal, *duration_ms* coefficient decreases by 3.681e-06 relative to 50s.$^+$ |
| 00*s* : *duration_ms* | 7.751e-06 | 0.582503 | If 00s then, all else equal, *duration_ms* coefficient increases by 7.751e-06 relative to 50s.$^+$ |
| $duration\_ms^3$ | 1.172e-16 | 0.004803* | E[*popularity*] increases/decreases by 1.172e-16 for an increase/decrease in $duration\_ms^3$ of 1, all else equal.$^+$ |
| 60*s* : *loudness* | -6.608e-01 | 0.589513 | If 60s then, all else equal, *loudness* coefficient decreases by 6.608e-01 relative to 50s.$^+$ |
| 70*s* : *loudness* | 2.114e-01 | 0.772969 | If 70s then, all else equal, *loudness* coefficient increases by 2.114e-01 relative to 50s.$^+$ |
| 80*s* : *loudness* | 6.559 | 5.630e-10* | If 80s then, all else equal, *loudness* coefficient increases by 6.559 relative to 50s.$^+$ |
| 90*s* : *loudness* | -1.208 | 0.402561 | If 90s then, all else equal, *loudness* coefficient decreases by 1.208 relative to 50s.$^+$ |
| 00*s* : *loudness* | 1.542 | 0.126930 | If 00s then, all else equal, *loudness* coefficient increases by 1.542 relative to 50s.$^+$ |

Coeff. = Coefficient least squares estimate; P-val = P-value; * = Significant at the $\alpha = 5\%$ level;

$^+$ = Assumed associated higher/lower-order terms remained constant for simplified interpretation.

Table 13: Summary of the coefficients of the final model continued.

| **Term** | **Coeff.** | **P-val** | **Interpretation** |
|---|---|---|---|
| $60s : loudness^2$ | -4.766e-02 | 0.326903 | If 60s then, all else equal, $loudness^2$ coefficient decreases by 4.766e-02 relative to 50s.$^+$ |
| $70s : loudness^2$ | 3.851e-03 | 0.884755 | If 70s then, all else equal, $loudness^2$ coefficient increases by 3.851e-03 relative to 50s.$^+$ |
| $80s : loudness^2$ | 0.245 | 1.670e-07* | If 80s then, all else equal, $loudness^2$ coefficient increases by 0.245 relative to 50s.$^+$ |
| $90s : loudness^2$ | -5.406e-02 | 0.420926 | If 90s then, all else equal, $loudness^2$ coefficient decreases by 5.406e-02 relative to 50s.$^+$ |
| $00s : loudness^2$ | 1.249e-02 | 0.743938 | If 00s then, all else equal, $loudness^2$ coefficient increases by 1.249e-02 relative to 50s.$^+$ |
| $loudness^3$ | -4.453e-03 | 0.005913* | E[*popularity*] decreases/increases by 4.453e-03 for an increase/decrease in $loudness^3$ of 1, all else equal.$^+$ |
| $duration\_ms : loudness$ | -6.176e-06 | 0.014969* | No obvious interpretation. |
| $duration\_ms^2 : loudness^2$ | -3.859e-13 | 0.041598* | No obvious interpretation. |
| $danceability : duration\_ms$ | -2.823e-05 | 0.121623 | No obvious interpretation. |

Coeff. = Coefficient least squares estimate; P-val = P-value; * = Significant at the $\alpha = 5\%$ level;

$^+$ = Assumed associated higher/lower-order terms remained constant for simplified interpretation.

As can be seen from Table 13, the 50s level was treated as the reference level for the *decade* predictor in the Final Model. As a result, the intercept term represented the the 50s case intercept while the coefficient estimates for the $60s$, $70s$, $80s$, $90s$ and $00s$ levels adjusted the $50s$ intercept for their respective cases. Similarly, the coefficient estimates for the $duration\_ms$, $loudness$ and $loudness^2$ terms represent their respective 50s case slopes while the coefficient estimates for their associated *decade* interaction terms adjust their $50s$ slopes for the various other *decade* cases. As the $danceability$, $duration\_ms^2$, $duration\_ms^3$, $loudness^3$, $duration\_ms : loudness$, $duration\_ms^2 : loudness^2$ and $danceability : duration\_ms$ terms did not have any associated *decade* interaction terms in the Final Model, their coefficient estimates applied equally to all *decade* cases.

The interpretations of all the linear, quadratic, cubic and interaction terms associated with the $duration\_ms$ and $loudness$ predictors in Table 13 assumed that a change in the value of $duration\_ms$ or $loudness$ for the term of consideration did not cause a change to any other associated term. In reality, an isolated change in either $duration\_ms$ or $loudness$ will cause a change in multiple associated terms simultaneously, and as a result, the Final Model will predict complex changes in the expected *popularity*. Hence, this assumption made possible simplified, but informative, interpretations for the various $duration\_ms$ and $loudness$ coefficient estimates.

It can also be seen that Table 13 provides no interpretation for the coefficient estimates of the $duration\_ms : loudness$, $duration\_ms^2 : loudness^2$ and $danceability : duration\_ms$ terms in the Final Model. These terms corresponded to numerical-numerical interaction terms, and as such, the interpretation of their coefficient estimates requires further analysis.

From Table 13 it can be seen that the $loudness$, $duration\_ms$ and *decade* predictors featured heavily in the Final Model. This is consistent with the finding from the Bivariate Analysis as $loudness$ and $duration\_ms$ were the numerical predictors which had the most obvious relationship with *popularity* and *decade* was the categorical predictor which had the most variance in *popularity* between its levels.

# 8 Assumption Checking

The four main assumptions which underpin the least squares approach of linear regression are:

1. linearity of the model;
2. homoscedasticity (constant variance of the residuals);
3. normality of the residuals; and
4. independence of the residuals.

In order to check the assumptions of linearity, homoscedasticity and normality for the Final Model, and to identify any potential influential points, the Residual vs Fitted plot, Scale-Location plot, Normal Q-Q plot and Residuals vs Leverage plot shown in Figure 19 where generated. Refer to Appendix E for the full `R` code that was used to generate the various plots for assessing the Final Model.

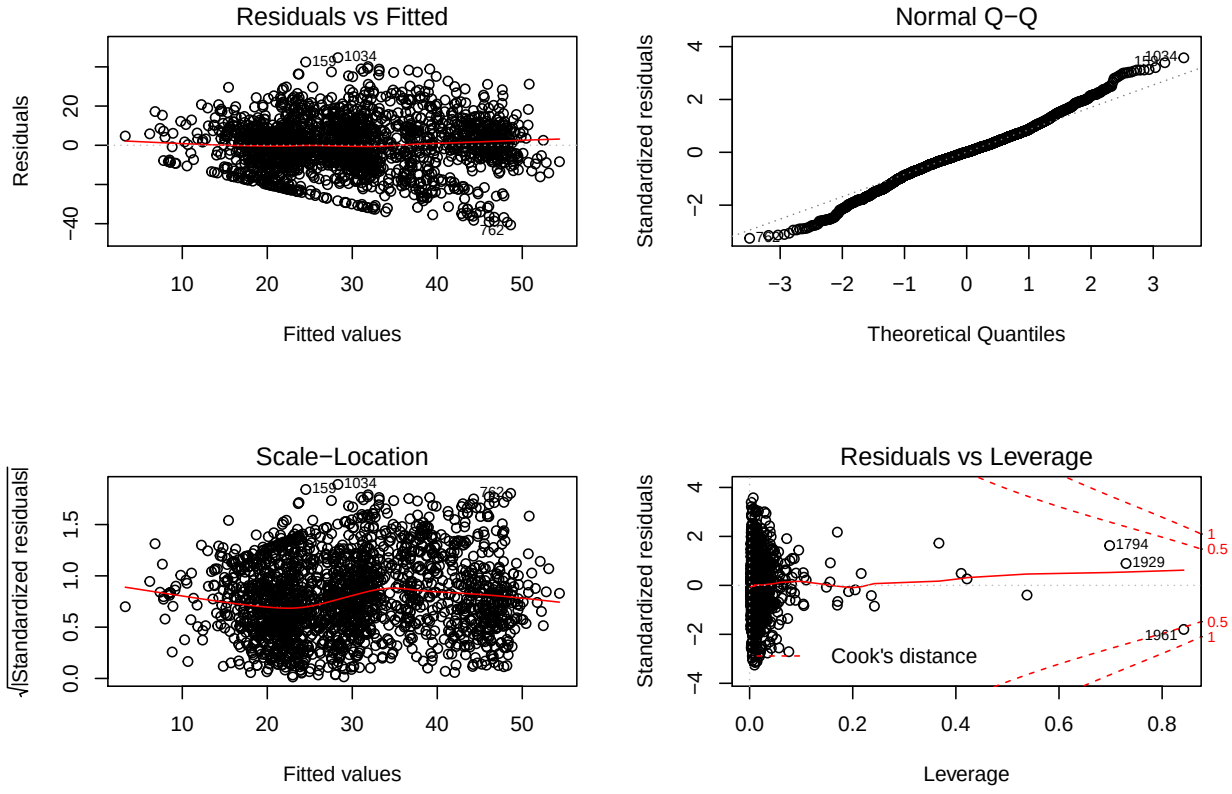


Figure 19: The Residual vs Fitted (top left), Normal Q-Q (top right), Scale-Location (bottom left) and Residuals vs Leverage (bottom right) plots for the Final Model.

## 8.1 Linearity Assumption Check

From inspection of the Residuals vs Fitted plot in Figure 19 it can be seen that the points seem to be scattered about the zero residual level and do not collectively, as a whole, form some systematic curvature. Furthermore, the fitted red line seems relatively flat at the zero residual level. Hence, both of these observations suggest that the assumption of linearity is plausible for the Final Model.

## 8.2 Homoscedasticity Assumption Check

The Scale-Location plot in Figure 19 illustrates how the variance of the points seems relatively uniform across the fitted values. Thus, the Scale-Location plot seems to support the assumption of constant residual variance for the Final Model. However, the Residuals vs Fitted plot in Figure 19 indicates that there is a set of points where $10 \leq$ fitted values $\leq 35$ and $-30 \leq$ Residuals $\leq -10$ which form an straight line. This structure may be a symptom of the bounded nature of the *popularity* ($0 \leq$ *popularity* $\leq 100$) response, in particular, the minor mode that exists the lower bound of its distribution, as shown during Variable Description in Figure 1. Hence, there are some concerns regarding the validity of the homoscedasticity assumption.

## 8.3 Normality Assumption Check

It can be seen from the Normal Q-Q plot in Figure 19 that the points with theoretical quantile values in the interval of [-2.5, 2.5] seem to resemble a straight line relatively well, and as such, the assumption of normality is plausible for the majority of the points under the Final Model. However, the points which lie outside of this interval seem to flatten out, and as such, bring into doubt whether the assumption of normality. The flattening out of the Normal Q-Q plot for theoretical quantile values above and below 2.5 and -2.5 respectively might again be attributable to the bounded nature of the *popularity* ($0 \leq$ *popularity* $\leq 100$) response and of the *danceability* ($0 \leq$ *danceability* $\leq 1$) and *duration_ms* (*duration_ms* $\geq 0$) predictors. On balance, the Normal Q-Q plot appears to support the assumption of normality of the residuals, particularly for the majority of the subjects that unaffected by the *popularity*, *danceability* and *duration_ms* bounds.

## 8.4 Independence Assumption Check

Finally, independence was the last assumption that required checking for the Final Model. If independence is to occur then the *duration_ms*, *danceability*, *loudness*, *decade* and *popularity* of one track from the dataset should not affect the *duration_ms*, *danceability*, *loudness*, *decade* or *popularity* of any other track in the dataset. This would be the case for any two song randomly selected the entire set of all possible songs. However, the dataset was constructed from Spotify tracks of only five artists. Hence, any two randomly selected tracks are reasonably likely to be of the same artist, and thus, have the same *decade* and be associated with a baseline *duration_ms*, *danceability*, *loudness* and *popularity*. Consequently, it was deemed reasonable to argue that the assumption of independent residuals was not strictly satisfied by the Final Model.

## 8.5 Dataset Assumptions Check

In addition to the assumptions of the least squares approach of linear regression, the Final Model was also subject to assumptions inherent in the dataset. One of the main assumptions in the dataset was that tracks of Elvis Presley, The Beatles, David Bowie, Michael Jackson, Blur and Beyoncé could be used to represent all of the songs from the 1950's, 1960's, 1970's, 1980's, 1990's and 2000's respectively. Given that all artists have released songs which have achieved mainstream success across multiple decades, the assumption that the songs of a particular artists can be used as proxy for songs of a particular decade does not seem reasonable. Furthermore, one could reasonably argue that within a given decade there is a large amount of diversity amongst the songs that have been released. Hence, the songs of a single artist could be expected to adequately represent all of the variety that existed in a given decade. Consequently, there are doubts whether the Final Model would be able to adequately predict the popularity of songs from all genres.

## 8.6 Influence Check

From the Residuals vs Leverage plot in Figure 19 it can be seen that no subject had a value for Cook's distance that was greater than or equal to 1.0. Hence, none of the subjects were determined to be influential. However, the 1794th, 1926th and 1961th subjects were found to have relatively high leverages that were approximately 0.7 or greater, and as such, were deemed to be of interest. The 1794th, 1926th and 1961th subjects correspond to the tracks"Optigan I" by Blur, "Encore For The Fans" by Beyoncé, and "Destiny's Child Medley - Audio from The Beyonce Experience Live" by Beyoncé respectively. After manual inspection the 1794th, 1926th and 1961th subjects where determined to be an experimental interlude, spoken word message and a 19 minute live performance respectively. Hence, all three of these subjects were deemed to represent non-conventional songs, and as such, were identified as being additional outliers and flagged for potential removal.

# 9 Prediction

The Final Model was employed to predict the popularity of a song with the following features:

- *duration_ms* of 180,000 (three minutes);
- *decade* of the 90s;
- *key* of C; and
- all other predictors values are taken to be their dataset mean.

Refer to Appendix E for the full R code that was used to obtain a predicted expected popularity for the new example song from the Final Model.

## 9.1 Example Song Popularity Prediction

As the Final Model only required values for the *decade*, *key*, *duration_ms*, *danceability* and *loudness* predictors, the following R code was used to obtain a point estimate and 95% prediction interval for the song's predicted *popularity*:

```
# create new datapoint ----
newdata <- tibble(
                  decade = factor("90s", levels = lev_dec),
                  key = factor("C", levels = lev_key),
                  duration_ms = 180000,
                  danceability = mean(data$danceability),
                  loudness = mean(data$loudness)
                  )



# get point estimate and prediction interval ----
predict(final.model, newdata = newdata, interval = "prediction")
```

From the R output, it was found that the point estimate of the expected *popularity* of the song is 28.69, with a 95% prediction interval of $(3.828, 53.563)$. Hence, the Final Model finds that the interval $(3.828, 53.563)$ will contain the new example song's actual *popularity* value with a probability of 0.95. As this point estimate and prediction interval fell within the expected *popularity* range listed in Table 2, they were both deemed to be plausible predictions.

## 9.2 Extensions to the Predictive Model

Throughout the Bivariate Analysis and Assumption Checking process several issues were identified, which, if addressed, could potentially result in an improved predictive model for song popularity. The first of these issues related to the 63 subjects from the Bivariate Analysis and the 3 subjects from the Influence Check which were deemed as being outliers. These outliers were not removed from the dataset during model fitting process. Hence, a new Final Model of enhanced predictive power could possibly be achieved if the model fitting process was repeated on the dataset after the removal of the outliers.

The second issue is related to the bounded nature of the *popularity* response, and the effect of this boundedness on the Final Model's ability to conform to the homoscedasticity and normality assumptions of the least squares approach. Possible improvements in the Final Model's adherence to the homoscedasticity and normality assumptions could be achieved by applying a *logit transform* of the following form to *popularity* prior to the model fitting process:

$$\rho = \log\left(\frac{popularity}{100 - popularity}\right).$$

As a result, the bounded range of *popularity* is mapped from $[0, 100]$ to the unbounded range of $(-\infty, \infty)$. Since, *popularity* values of 0 and 100 are attainable, an effectively small amount, $\Delta p$, may be added to both the numerator and denominator to avoid undefined values. Such transformations of *popularity* might not only improve the resulting Final Model's ability to uphold the homoscedasticity and normality assumptions, but also enhance the Final Model's predictions of song popularity.

# 10 Conclusion

As music streaming providers could potentially benefit from having the ability to estimate the popularity of the songs contained within their catalogues, a predictive model for song popularity was developed by fitting a linear regression model onto a dataset that was obtained from Spotify. The dataset contained 2081 subjects from six different artists were each subject represented a single track with 21 variables. The *popularity* variable was used as the response and eight other variables from the dataset (*danceability*, *energy*, *loudness*, *duration_ms*, *key*, *mode*, *time_signature*, and *decade*) were considered for possible predictors.

Once the dataset was cleaned, a univariate and bivariate analysis was performed on the response and predictor variables to ascertain their individual distributions and described the various response-predictor relationships respectively. From this analysis it was found that *popularity* had a bimodal distribution with a minor mode at its lower bound, and that *loudness* and *duration_ms* had positive relationships with *popularity*.

After the response and predictors had been analysed, various linear regression models for song popularity were obtained by applying the Backwards Elimination, Forward Selection and Stepwise Selection procedures via the least squares approach. All three of the stepwise regression algorithms utilised the Akaike Information Criterion for their heuristics. From a comparison of the various resulting models it was determined that the Stepwise Selection procedure produced the most desirable predictive model (the Final Model) and that this model utilised the *danceability*, *duration_ms*, *loudness* and *decade* predictors to estimate expected *popularity*.

Following the selection of the Final Model, it was investigated whether the assumptions of the least squares approach were plausibly satisfied by this resulting predictive linear model. The analysis of the assumptions revealed that the Final Model had some concerns with regards the the homoscedasticity and normality of residuals assumptions, and that these issues may be attributable to the boundedness of *popularity*, *danceability* and *duration_ms*. Furthermore, the assumption concerning the independence of the residuals were also brought into doubt by the lack of artist variety that existed within the dataset.

The Final Model was employed to predict a point estimate of 28.69 and prediction interval of (3.828,53.563) for the expected popularity of a three minute song that was released in the 90s in the key of C and possessed average values for all other dataset traits. By comparing the point estimate and prediction interval to the expected *popularity* range determined during the description of the dataset, they were both deemed to be plausible.

Finally, it was suggested that an improved predictive linear regression model could be obtained if the outliers identified during the bivariate analysis and assumption checking process were removed from the dataset before the model fitting procedure was performed. Similarly, additional improvements to the predictive model were proposed if a *logit* transformation was applied to the *popularity* response to lessen the effects of its boundedness on the breaches of the homoscedasticity and normality of residuals assumptions.

# 11 References

[1] *Spotify Variable Description*, Spotify AB 2018, accessed 10 October 2018, <https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/>

[2] *Spotify Tracks*, Spotify AB 2018, accessed 10 October 2018, <https://developer.spotify.com/documentation/web-api/reference/tracks/get-several-tracks/>

[3] *Statistical Modelling and Inference II: Course Notes*, Dr Jono Tuke 2018, The University of Adelaide.

# Appendix A Full R Code for Cleaning, Summarising and Plotting the Spotify Dataset

The following R code was used to clean, analyse and plot the Spotify dataset:

```
# Data_clean loads, cleans, summarises and plots the Spotify dataset variables
# I Jacobson, October 2018



# loading libraries ----
library(tidyverse)
library(readxl)



# setting plot theme ----
theme_set(theme_minimal())









# +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
# Data Cleaning
# +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++



# read in data ----
data <- read_excel("spotify.xlsx")



# check variable names & types ----
var_names <- colnames(data)
var_classes <- c(rep("", length(var_names)))
for (i in seq_along(var_names)) {
  var_classes[i] <- class(data[[var_names[i]]])
}
vars <- matrix(c(var_names, var_classes),
               nrow = length(var_names),
               ncol = 2,
               byrow = FALSE)

vars



# define numerical & category variables & titles ----
num_vars = c("popularity", "danceability", "energy" ,"loudness", "duration_ms")
num_titles = c("Popularity", "Danceability", "Energy" ,"Loudness", "Duration")
cat_vars = c("key", "mode" ,"time_signature", "decade")
cat_titles = c("Key", "Mode" ,"Time Signature", "Decade")



# check for missing data ----
has_no_missing_data <- function(dataset, var_name) {
  sum(!is.na(data[[var_name]])) == length(data[[var_name]])
}
for (i in seq_along(num_vars)) {
  print(c(num_vars[i], has_no_missing_data(data, num_vars[i])))
}
for (i in seq_along(cat_vars)) {
  print(c(cat_vars[i], has_no_missing_data(data, cat_vars[i])))
}



# inspect numerical variables ----
inspect_num_data <- function(dataset, var_name) {
  summary(dataset[[var_name]], useNA = "always")
}
```

```r
for (i in seq_along(num_vars)) {
  print(num_vars[i])
  print(inspect_num_data(data, num_vars[i]))
}



# inspect categorical variables ----
inspect_cat_data <- function(dataset, var_name) {
  table(dataset[[var_name]], useNA = "always")
}
for (i in seq_along(cat_vars)) {
  print(cat_vars[i])
  print(inspect_cat_data(data, cat_vars[i]))
}



# convert categorical variables to factor types and map values with an order ----
lev_key <- c("A", "A#", "B", "C",  "C#", "D", "D#", "E", "F", "F#", "G", "G#")
lev_mod <- c("major", "minor")
lev_dec <- c("50s", "60s", "70s", "80s", "90s", "00s")
lev_tim <- c("0", "1", "2", "3", "4", "5")
data$key <- factor(data$key, levels = lev_key)
data$mode <- factor(data$mode, levels = lev_mod)
data$decade <- factor(data$decade, levels = lev_dec)
data$time_signature <- factor(data$time_signature, levels = lev_tim)



# re-check summaries of data ----
s_pop <- summary(data$popularity)
s_dan <- summary(data$danceability)
s_eng <- summary(data$energy)
s_key <- summary(data$key)
s_lou <- summary(data$loudness)
s_mod <- summary(data$mode)
s_dur <- summary(data$duration_ms)
s_tim <- summary(data$time_signature)
s_dec <- summary(data$decade)
s_pop
s_dan
s_eng
s_key
s_lou
s_mod
s_dur
s_tim
s_dec



# calculate standard deviations of numerical variables ----
sd(data$popularity)
sd(data$danceability)
sd(data$energy)
sd(data$loudness)
sd(data$duration_ms)



# inspect category variable level summaries
summary(s_key)
summary(s_mod)
summary(s_tim)
summary(s_dec)



# calculate category variable level standard deviations
sd(s_key)
sd(s_mod)
sd(s_tim)
sd(s_dec)


```

```
147
148 # calculate category variable level IQRs
149 get_cat_iqr <- function(dataset, cat_name, lev_name) {
150   s <- print(fivenum(filter(dataset, dataset[cat_name] == lev_name)$popularity))
151   return(s[4] - s[2])
152 }
153 for (i in 1:length(lev_key)) {
154   print(lev_key[i])
155   print(get_cat_iqr(data, "key", lev_key[i]))
156 }
157 for (i in 1:length(lev_mod)) {
158   print(lev_mod[i])
159   print(get_cat_iqr(data, "mode", lev_mod[i]))
160 }
161 for (i in 1:length(lev_tim)) {
162   print(lev_tim[i])
163   print(get_cat_iqr(data, "time_signature", lev_tim[i]))
164 }
165 for (i in 1:length(lev_dec)) {
166   print(lev_dec[i])
167   print(get_cat_iqr(data, "decade", lev_dec[i]))
168 }
169
170 # save cleaned data ----
171 write_rds(data, "spotify.rds")
172
173
174
175
176
177
178
179 # ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
180 # Generating Plots
181 # ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
182
183
184
185 # plot histograms of numeric data types ----
186 plot_num_hist <- function(dataset, var_name, var_title) {
187   ggplot(dataset, aes(x = dataset[[var_name]])) +
188     geom_histogram(col= "black", fill = "slateblue4") +
189     labs(x = paste(var_title, "of selected tracks on Spotify")) +
190     labs(y = "Count")
191 }
192 plot_num_hist(data, "popularity", "Popularity")
193 for (i in seq_along(num_vars)) {
194   pdf(paste("Histogram_", num_titles[i], "07.pdf", sep = ""), width = 5.83,
195       height = 4.13)
196   print(plot_num_hist(data, num_vars[i], num_titles[i]))
197   dev.off()
198 }
199
200
201
202 # plot bar of categorical data types ----
203 plot_cat_bar <- function(dataset, var_name, var_title) {
204   ggplot(dataset ,aes(x = dataset[[var_name]])) +
205     geom_bar(col= "black", fill = "slateblue4") +
206     labs(x = paste(var_title, "of selected tracks on Spotify")) +
207     labs(y = "Count")
208 }
209 for (i in seq_along(cat_vars)) {
210   pdf(paste("Bar_", cat_titles[i], "07.pdf", sep = ""), width = 5.83,
211       height = 4.13)
212   print(plot_cat_bar(data, cat_vars[i], cat_titles[i]))
213   dev.off()
214 }
215
216
217
218 # plot scatter plots of popularity against numeric predictors ----
219 plot_num_scat <- function(dataset, var_name_y, var_name_x, var_title_y,
220                           var_title_x) {
221   ggplot(dataset, aes(x = dataset[[var_name_x]], y = dataset[[var_name_y]])) +
222     geom_point(colour = "slateblue4") +
```

```
    labs(x = var_title_x, y = var_title_y)
}
for (i in 2:length(num_vars)) {
  pdf(paste("Scatter_", num_titles[i], "07.pdf", sep = ""), width = 5.83,
      height = 4.13)
  print(plot_num_scat(data, num_vars[1], num_vars[i], num_titles[1], num_titles[i]))
  dev.off()
}



# plot boxplots of popularity against category predictors ----
plot_cat_box <- function(dataset, var_name_y, var_name_x, var_title_y,
                         var_title_x) {
  ggplot(data, aes(x = dataset[[var_name_x]], y = dataset[[var_name_y]],
                   fill = dataset[[var_name_x]])) +
    stat_boxplot(geom = "errorbar", size = 0.75) +
    geom_boxplot(size = 0.75, alpha = 0.5) +
    labs(x = var_title_x, y = var_title_y) +
    theme(axis.text.x = element_text(angle = -90, hjust = 0, vjust = 0.25),
          axis.title.x = element_blank()) +
    scale_fill_discrete(name = var_title_x)
}
for (i in 1:length(cat_vars)) {
  pdf(paste("Boxplot_", cat_titles[i], "07.pdf", sep = ""), width = 5.83,
      height = 4.13)
  print(plot_cat_box(data, num_vars[1], cat_vars[i], num_titles[1], cat_titles[i]))
  dev.off()
}



# plot scatter plots of popularity against numeric
# predictors with categorical mini-plots ----
plot_num_scat_mini <- function(dataset, var_name_y, var_name_x, var_name_mini,
                               var_title_y, var_title_x, var_title_mini) {
  ggplot(dataset, aes(x = get(var_name_x), y = get(var_name_y),
                      col = get(var_name_mini))) +
    geom_point() +
    facet_wrap(~get(var_name_mini)) +
    theme(text = element_text(size = 9)) +
    labs(x = var_title_x, y = var_title_y, color = var_title_mini)
}
for (i in 2:length(num_vars)) {
  for (j in 1:length(cat_vars)) {
    tmp <- plot_num_scat_mini(data, num_vars[1], num_vars[i], cat_vars[j],
                              num_titles[1], num_titles[i], cat_titles[j])
    ggsave(paste("Scatter_", num_titles[i], "_", cat_titles[j], "07.pdf", sep = ""),
           plot = tmp, device = "pdf", path = NULL, width = 10, height = 4)
  }
}
```

# Appendix B Numeric-Categorical Interaction Bivariate Scatter Plots

Figures 20 - 35 display the *popularity* scatter plots of all numeric-categorical interaction combination.

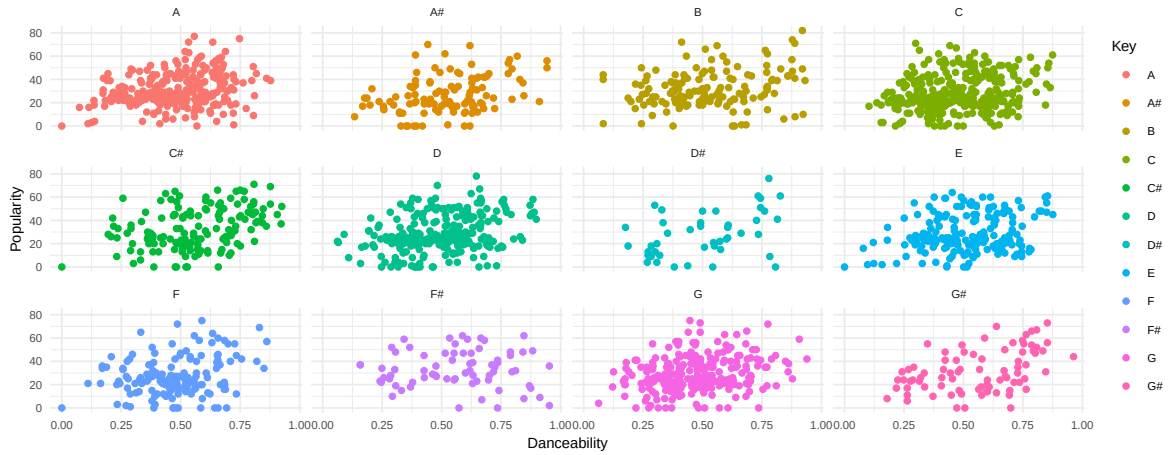


Figure 20: Scatter plots of the *popularity* data against the *danceability* data for each *key* level.

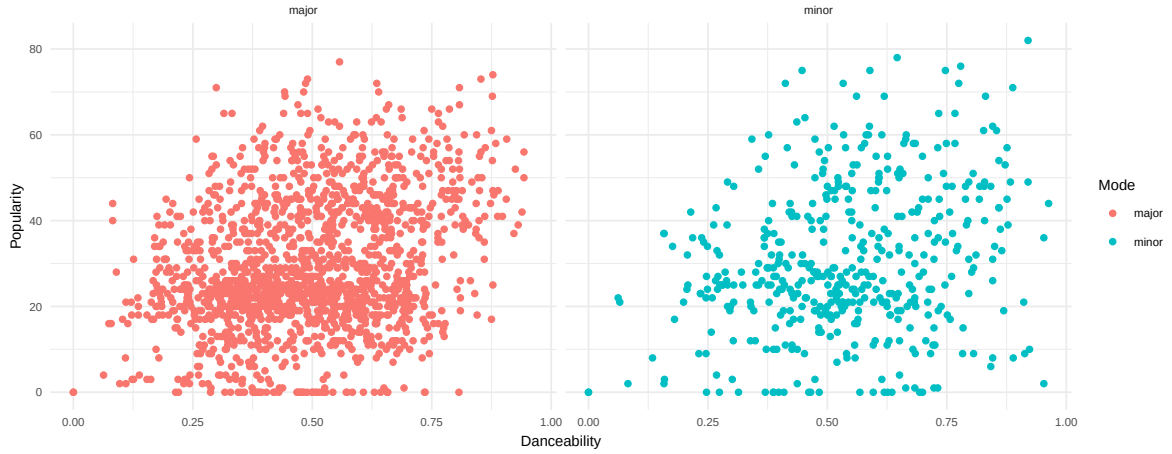


Figure 21: Scatter plots of the *popularity* data against the *danceability* data for each *mode* level.

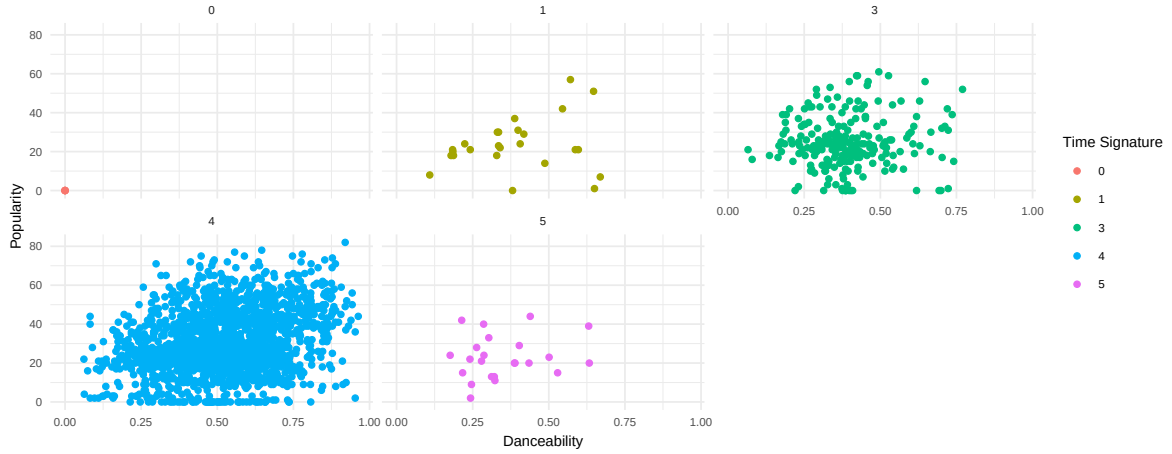


Figure 22: Scatter plots of the *popularity* data against the *danceability* data for each *time_signature* level.

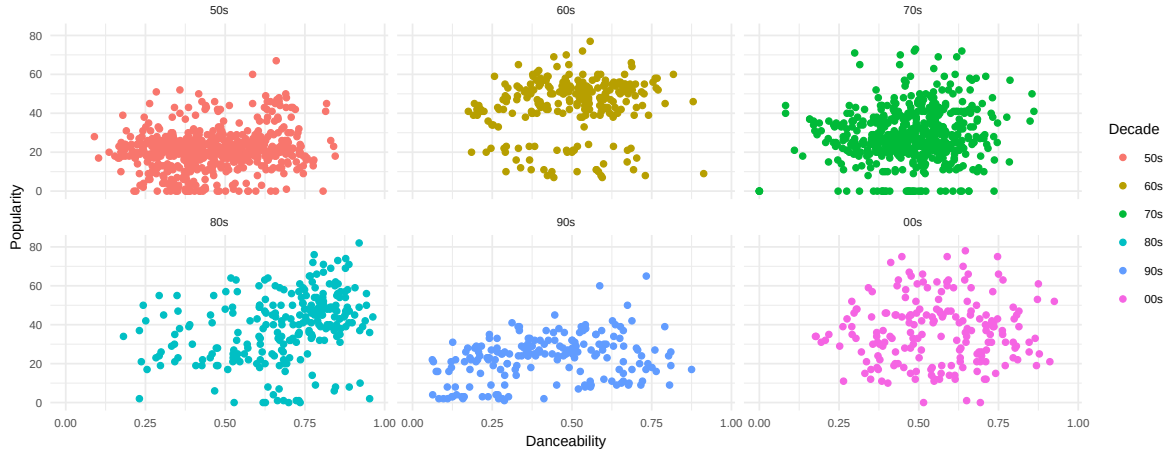


Figure 23: Scatter plots of the *popularity* data against the *danceability* data for each *decade* level.

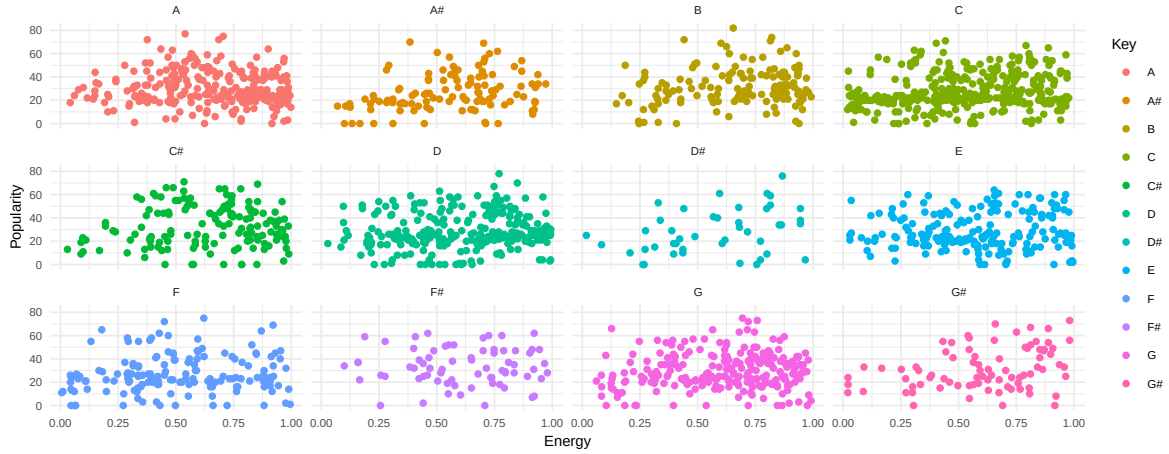


Figure 24: Scatter plots of the *popularity* data against the *energy* data for each *key* level.

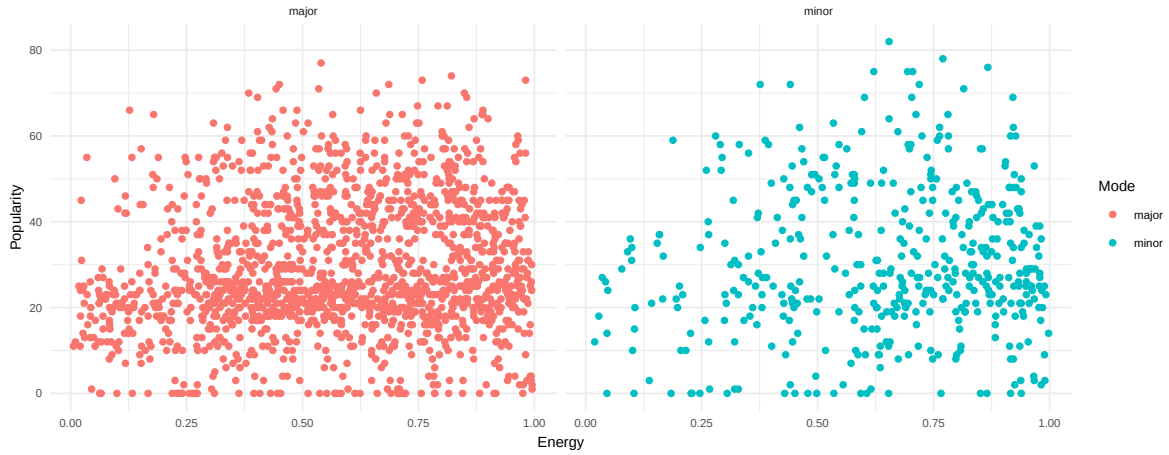


Figure 25: Scatter plots of the *popularity* data against the *energy* data for each *mode* level.

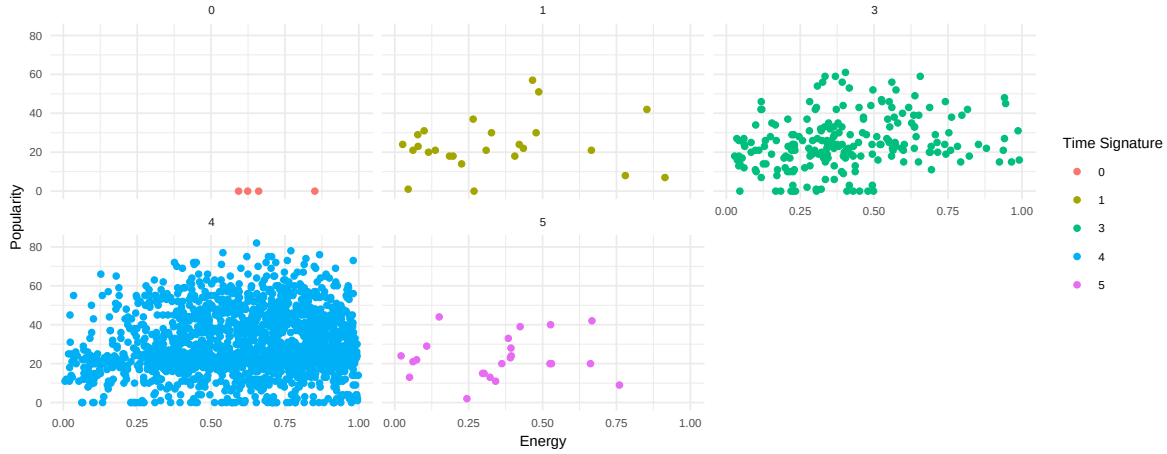


Figure 26: Scatter plots of the *popularity* data against the *energy* data for each *time_signature* level.

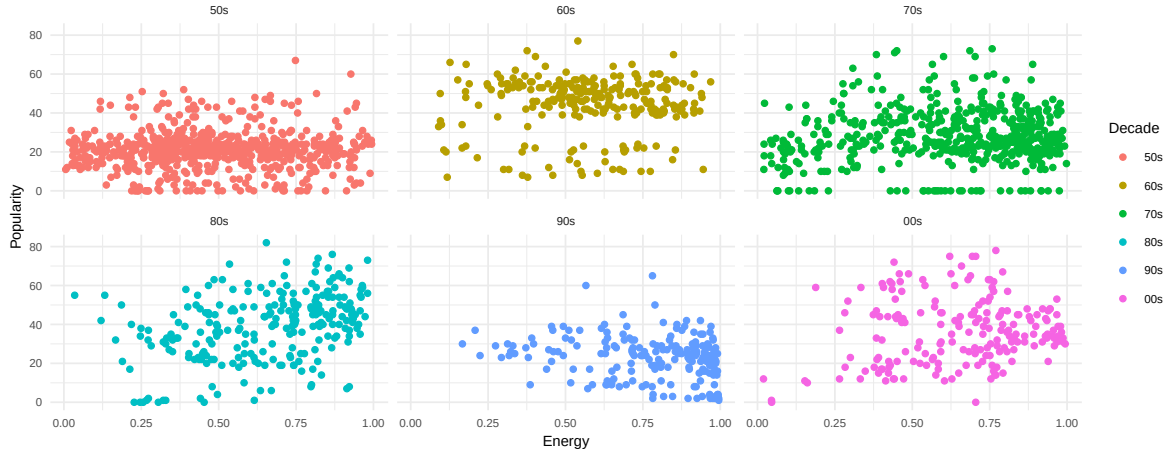


Figure 27: Scatter plots of the *popularity* data against the *energy* data for each *decade* level.

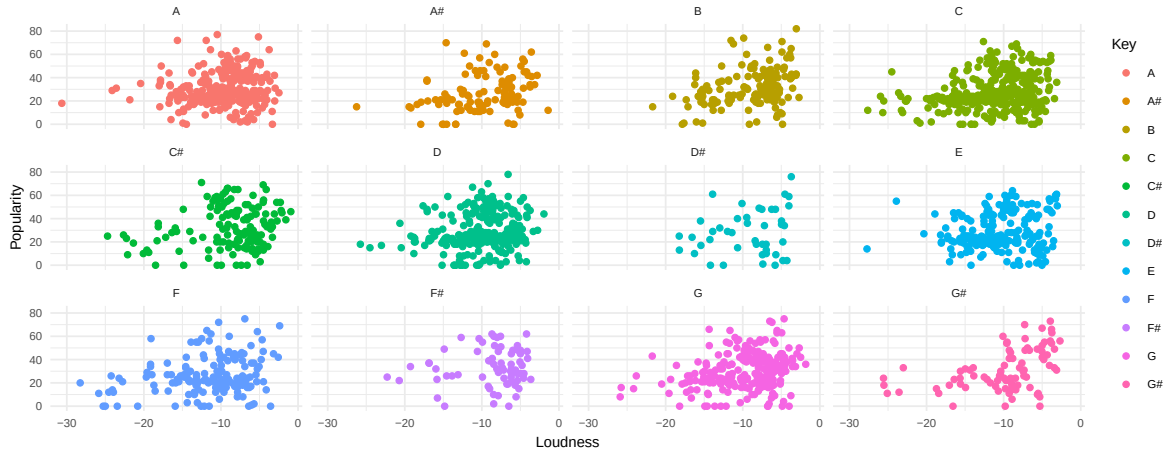


Figure 28: Scatter plots of the *popularity* data against the *loudness* data for each *key* level.

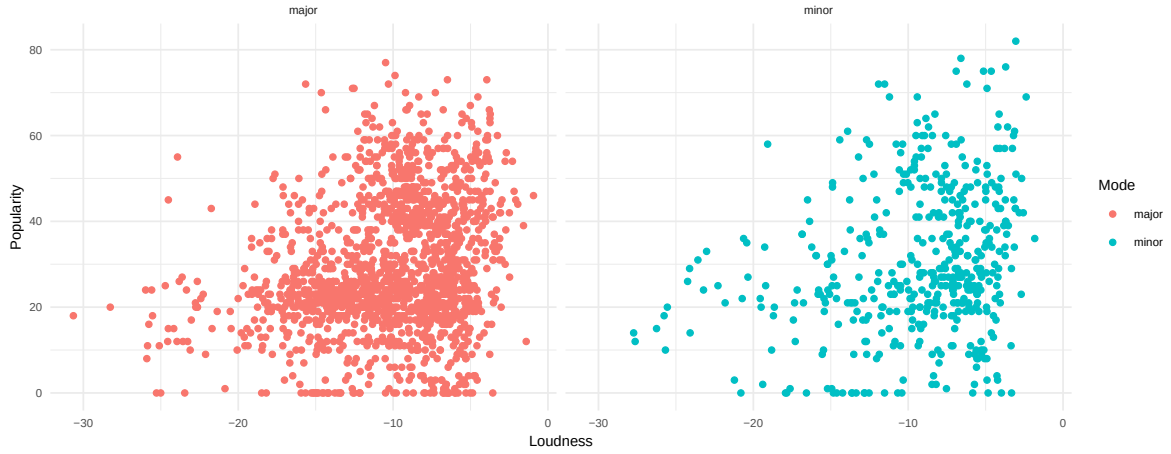


Figure 29: Scatter plots of the *popularity* data against the *loudness* data for each *mode* level.

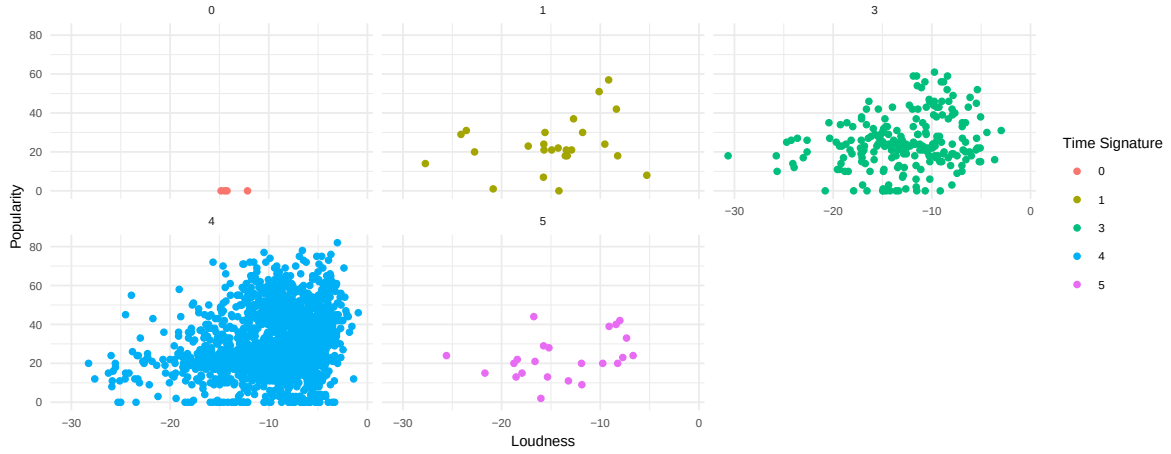


Figure 30: Scatter plots of the *popularity* data against the *loudness* data for each *time_signature* level.

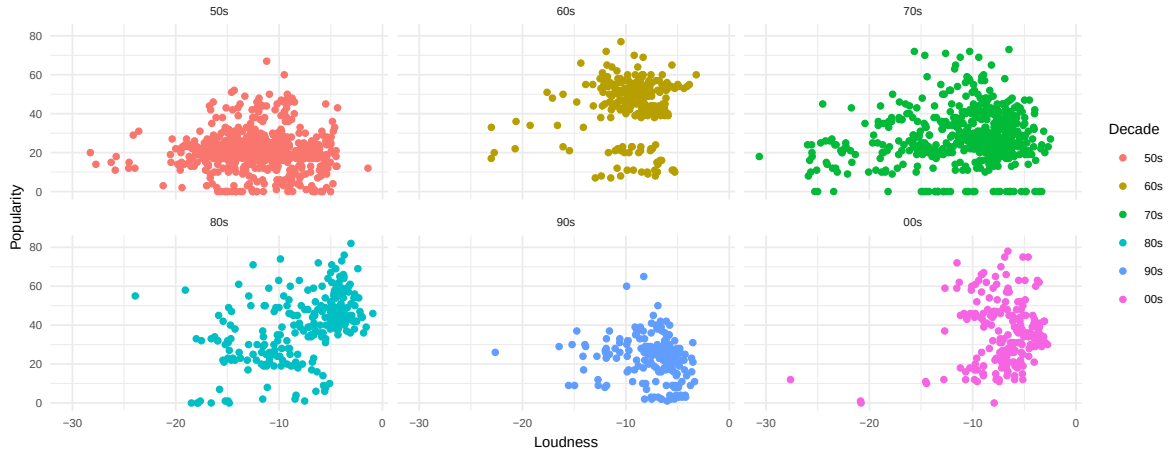


Figure 31: Scatter plots of the *popularity* data against the *loudness* data for each *decade* level.

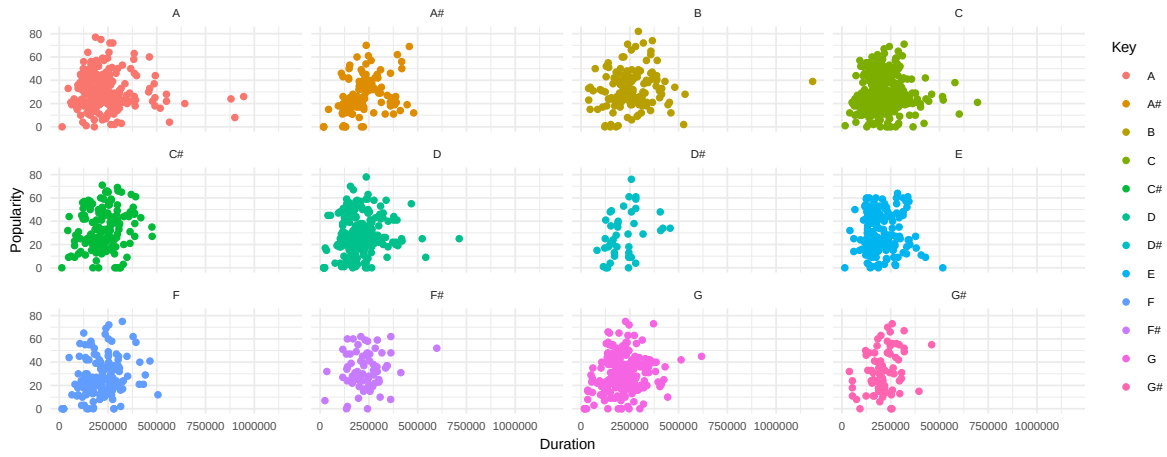


Figure 32: Scatter plots of the *popularity* data against the *duration_ms* data for each *key* level.

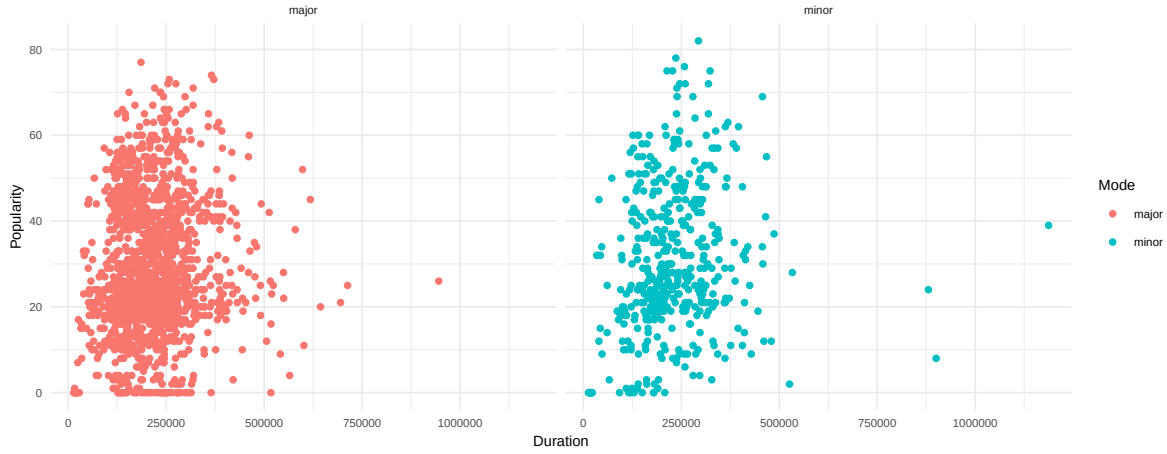


Figure 33: Scatter plots of the *popularity* data against the *duration_ms* data for each *mode* level.

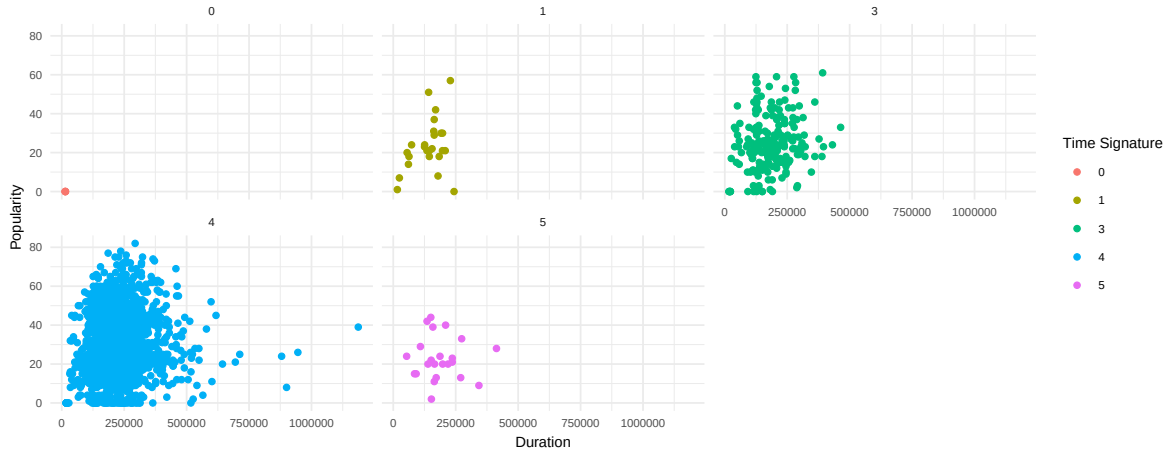


Figure 34: Scatter plots of the *popularity* data against the *duration_ms* data for each *time_signature* level.

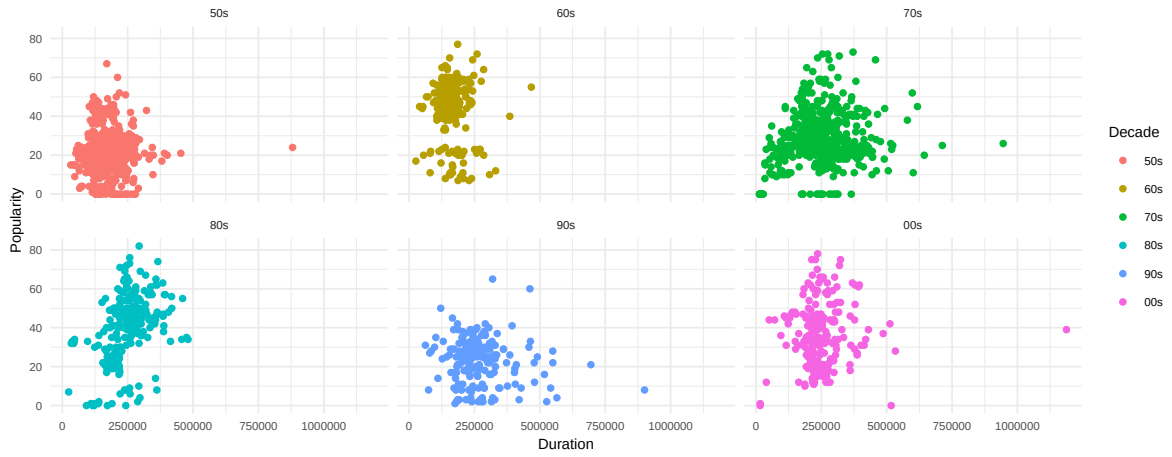


Figure 35: Scatter plots of the *popularity* data against the *duration_ms* data for each *decade* level.

# Appendix C Full R Code for Removing Outliers from Spotify Dataset

The following R code was used to identify and remove outliers from the Spotify dataset:

```
# Outlier_remove removes datapoints (tracks) from the spotify.xlsx dataset based on a set of
    predetermined characteristics.
# M Ucci, October 2018
library(tidyverse)
library(readxl)

# Loading data
spot_data = read_excel( "spotify.xlsx" )

#### Removing non-songs ####
# To remove tracks which are not considered to be songs as such
# Keywords
keywords = paste(c("voice-over", "narrates", "interview", "comment"), collapse = "|")
# Removing data
reduced_data <- spot_data %>%
  filter(!grepl(keywords, track_name, ignore.case = TRUE)) %>%
  filter(!grepl(keywords, album_name, ignore.case = TRUE ))
# Tracking removed data:
removed_data <- spot_data %>%
  filter(grepl( keywords, track_name, ignore.case = TRUE)|
           grepl( keywords, album_name, ignore.case = TRUE ))
```

# Appendix D List of Outliers

Table 14 lists and describes the subjects that were identified from the dataset as being outliers.

Table 14: Description of identified outliers from dataset.

| **Artist** | **Track Title** | **Justification** |
|---|---|---|
| Elvis Presley | Elvis Fans' Comments/Opening Riff - Live | Fan appraisal |
| Elvis Presley | Elvis Fans' Comments II - Live | Fan appraisal |
| Elvis Presley | Elvis Fans' Comments III - Live | Fan appraisal |
| Elvis Presley | Elvis Fans' Comments IV - Live | Fan appraisal |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): Introduction | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Story Begins | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Bird Diverts the Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Duck, Dialogue with the Bird, Attack of the Cat | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): Grandfather | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Duck is Caught | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Wolf Stalks the Bird and the Cat | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): Peter Prepares to Catch the Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Bird Diverts the Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): Peter Catches the Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Hunters Arrive | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67 (Remastered): The Procession to the Zoo | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra, Op. 34 (Variations and Fugue on a Theme of Purcell): Theme: Full Orchestra | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra, Op. 34 (Variations and Fugue on a Theme of Purcell): Theme: Woodwinds | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra, Op. 34 (Variations and Fugue on a Theme of Purcell): Theme: Brass | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra, Op. 34 (Variations and Fugue on a Theme of Purcell): Theme: Strings | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra, Op. 34 (Variations and Fugue on a Theme of Purcell): Theme: Percussion | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra, Op. 34 (Variations and Fugue on a Theme of Purcell): Theme: Full Orchestra | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra, Op. 34 (Variations and Fugue on a Theme of Purcell): Variation I: Flute, Piccolo (Presto) | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: Introduction | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Story Begins | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Bird | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Duck, Dialogue with the Bird, Attack of the Cat | Narration of other artist's work |

Table 14: Description of identified outliers from dataset continued.

| **Artist** | **Track Title** | **Justification** |
|---|---|---|
| David Bowie | Peter and the Wolf, Op. 67: Grandfather | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Duck is Caught | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Wolf Stalks the Bird and the Cat | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: Peter Prepares to Catch the Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Bird Diverts the Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: Peter Catches the Wolf | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Hunters Arrive | Narration of other artist's work |
| David Bowie | Peter and the Wolf, Op. 67: The Procession to the Zoo | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra: Theme: Full Orchestra | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra: Theme: Woodwinds | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra: Theme: Brass | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra: Theme: Strings | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra: Theme: Percussion | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra: Theme: Full Orchestra | Narration of other artist's work |
| David Bowie | The Young Person's Guide to the Orchestra: Variation I: Flute, Piccolo (Presto) | Narration of other artist's work |
| Michael Jackson | Voice-Over Intro Quincy Jones Interview #1 / Quincy Jones Interview #1 | Voice-Over |
| Michael Jackson | Voice-Over Intro Quincy Jones Interview #2/Quincy Jones Interview #2 / Voice-Over Intro Billie Jean (Demo) | Voice-Over |
| Michael Jackson | Quincy Jones Interview #3 | Interview |
| Michael Jackson | Voice-Over Intro Rod Temperton Interview #1 / Rod Temperton Interview #1 | Voice-Over |
| Michael Jackson | Quincy Jones Interview #4 | Interview |
| Michael Jackson | Voice-Over Intro / Voice-Over Session from Thriller | Voice-Over |
| Michael Jackson | Voice-Over Intro Rod Temperton Interview #2 / Rod Temperton Interview #2 | Voice-Over |
| Michael Jackson | Quincy Jones Interview #5 | Interview |
| Michael Jackson | Vincent Price Excerpt From "Thriller" Voice-Over Session - Thriller 25th Anniversary Voice-Over Session | Voice-Over |
| Michael Jackson | Voice-Over Intro Quincy Jones Interview #1 / Quincy Jones Interview #1 | Voice-Over |
| Michael Jackson | Voice-Over Intro Quincy Jones Interview #2/Quincy Jones Interview #2 / Voice-Over Intro Billie Jean (Demo) | Voice-Over |
| Michael Jackson | About Love Never Felt So Good - Commentary by LA Reid | Commentary |
| Michael Jackson | About Chicago - Commentary by LA Reid & Timbaland | Commentary |
| Michael Jackson | About Loving You - Commentary by LA Reid & Timbaland | Commentary |
| Michael Jackson | About A Place With No Name - Commentary by LA Reid | Commentary |
| Michael Jackson | About Slave to the Rhythm - Commentary by LA Reid & Timbaland | Commentary |
| Michael Jackson | About Do You Know Where Your Children Are - Commentary by LA Reid & Timbaland | Commentary |

Table 14: Description of identified outliers from dataset continued.

| **Artist** | **Track Title** | **Justification** |
|---|---|---|
| Michael Jackson | About Blue Gangsta - Commentary by LA Reid & Timbaland | Commentary |
| Michael Jackson | About Xscape - Commentary by LA Reid | Commentary |

# Appendix E Full R Code for Fitting the Predictive Model

The following R code was used to fit numerous linear regression models for *popularity* on the dataset, analyse the selected Final Model and then predict the expected popularity of an new example song:

```
# Model_fit uses stepwise regression algorithms to fit linear regression models
# on the cleaned Spotify dataset via the least squares approach
# I Jacobson, October 2018



# loading libraries ----
library(tidyverse)
library(readxl)



# setting plot theme ----
theme_set(theme_minimal())



# load cleaned Spotify dataset
data <- read_rds("spotify.rds")



# ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
# Model Fitting
# ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++



# define scopes for model fitting ----
null <- popularity ~ 1
red <- popularity ~ (danceability + energy + loudness + duration_ms + key +
                       mode + time_signature + decade)^2
exp <- popularity ~ ((danceability + energy + loudness + duration_ms
                      + I(danceability^2) + I(energy^2) + I(loudness^2) + I(duration_ms^2))
                     *(key + mode + time_signature + decade)
                     + (danceability + energy + loudness + duration_ms
                        + I(danceability^2) + I(energy^2) + I(loudness^2) + I(duration_ms^2))
                            ^2
                     + (key + mode + time_signature + decade)^2)



# backwards elimination on reduced scope ----
backwards.red <- lm(red, data = data)
backwards.red <- step(backwards.red, direction = "backward")
summary(backwards.red)
AIC(backwards.red)
BIC(backwards.red)



# backwards elimination on expanded scope ----
backwards.exp <- lm(exp, data = data)
backwards.exp <- step(backwards.exp, direction = "backward")
summary(backwards.exp)
AIC(backwards.exp)
BIC(backwards.exp)



# forward selection on reduced scope ----
forwards.red <- lm(null, data = data)
forwards.red <- step(forwards.red, scope = red, direction = "forward")
summary(forwards.red)
AIC(forwards.red)
BIC(forwards.red)



# forward selection on expanded scope ----
forwards.exp <- lm(null, data = data)
```

```r
forwards.exp <- step(forwards.exp, scope = exp, direction = "forward")
summary(forwards.exp)
AIC(forwards.exp)
BIC(forwards.exp)


# stepwise selection on reduced scope ----
stepwise.red <- lm(null, data = data)
stepwise.red <- step(stepwise.red, scope = red, direction = "both")
summary(stepwise.red)
AIC(stepwise.red)
BIC(stepwise.red)


# stepwise selection on expanded scope ----
stepwise.exp <- lm(null, data = data)
stepwise.exp <- step(stepwise.exp, scope = exp, direction = "both")
summary(stepwise.exp)
AIC(stepwise.exp)
BIC(stepwise.exp)


# selecting the final model ----
final.model <- stepwise.exp
tmp <- par(mfrow = c(2, 2))
plot(final.model)
par(tmp)
plot(final.model, which = 1)
plot(final.model, which = 2)
plot(final.model, which = 3)
plot(final.model, which = 5)


# checking subjects of interest ----
data$track_name[1794]
data$artist[1794]
data$track_name[1929]
data$artist[1929]
data$track_name[1961]
data$artist[1961]




# ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
# Prediction
# ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++


# create new datapoint ----
newdata <- tibble(
  decade = factor("90s", levels = lev_dec),
  key = factor("C", levels = lev_key),
  duration_ms = 180000,
  danceability = mean(data$danceability),
  loudness = mean(data$loudness)
)


# get point estimate and prediction interval ----
predict(final.model, newdata = newdata, interval = "prediction")
```