Spotify Playlist Analysis

Pius Mutuma Kimathi

2023-09-14

Contents

Loading the libraries **1**

Exploratory Data Analysis **4**

Hypothesis Formulation **15**

Feature Engineering and Model Building **15**

Approach A: Build the linear regression model 18

Approach B: Re-build the linear regression model with added features 26

Cross Validation **28**

Forecasting Using Time Series Analysis **31**

Conclusion **34**

Loading the libraries

```
# Loading the libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lintr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(caTools)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
##
## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(boot)
```

```
##
## Attaching package: 'boot'
##
## The following object is masked from 'package:car':
##
##     logit
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:olsrr':
##
##     cement
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'lattice'
##
## The following object is masked from 'package:boot':
##
##     melanoma
##
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(stats)
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(forecast)
library(rmarkdown)
```

```r
# Loading the dataset
Playlist = read.csv("playlist_2010to2022.csv")
head(Playlist)
```

```
##                                          playlist_url year
## 1 https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk 2000
## 2 https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk 2000
## 3 https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk 2000
## 4 https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk 2000
## 5 https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk 2000
## 6 https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk 2000
##                 track_id           track_name track_popularity
## 1 3AJwUDP919kvQ9QcozQPxg               Yellow               91
## 2 2m1hi0nfMR9vdGC8UcrnwU All The Small Things               84
## 3 3y4LxiYMgDl4RethdzpmNe              Breathe               69
## 4 60a0Rd6pjrkxjPbaKzXjfq           In the End               88
## 5 62bOmKYxYg7dhrC6gH9vFn          Bye Bye Bye               74
## 6 5Mmk2ii6laakqfeCT7OnVD           Thong Song               73
##                            album              artist_id artist_name
## 1                    Parachutes 4gzpq5DPGxSnKTe4SA8HAU    Coldplay
## 2            Enema Of The State 6FBDaR13swtiWwGhX1WQsP   blink-182
## 3                       Breathe 25NQNriVT2YbSW80ILRWJa  Faith Hill
## 4 Hybrid Theory (Bonus Edition) 6XyY86QOPPrYVGvF9ch6wz Linkin Park
## 5           No Strings Attached 6Ff53KvcvAj5U7Z1vojB5o      *NSYNC
```

```
## 6            Unleash The Dragon 6x9QLdzo6eBZxJ1bHsDkjg       Sisqo
##                                                                                artist_genres
## 1                                                                 ['permanent wave', 'pop']
## 2 ['alternative metal', 'modern rock', 'pop punk', 'punk', 'rock', 'socal pop punk']
## 3                ['contemporary country', 'country', 'country dawn', 'country road']
## 4              ['alternative metal', 'nu metal', 'post-grunge', 'rap metal', 'rock']
## 5                                                      ['boy band', 'dance pop', 'pop']
## 6    ['contemporary r&b', 'dirty south rap', 'hip pop', 'r&b', 'urban contemporary']
##   artist_popularity danceability energy key loudness mode speechiness
## 1                86        0.429  0.661  11   -7.227    1      0.0281
## 2                75        0.434  0.897   0   -4.918    1      0.0488
## 3                61        0.529  0.496   7   -9.007    1      0.0290
## 4                83        0.556  0.864   3   -5.870    0      0.0584
## 5                65        0.610  0.926   8   -4.843    0      0.0479
## 6                56        0.706  0.888   2   -6.959    1      0.0654
##   acousticness instrumentalness liveness valence   tempo duration_ms
## 1      0.00239         1.21e-04   0.2340   0.285 173.372      266773
## 2      0.01030         0.00e+00   0.6120   0.684 148.726      167067
## 3      0.17300         0.00e+00   0.2510   0.278 136.859      250547
## 4      0.00958         0.00e+00   0.2090   0.400 105.143      216880
## 5      0.03100         1.20e-03   0.0821   0.861 172.638      200400
## 6      0.11900         9.64e-05   0.0700   0.714 121.549      253733
##   time_signature
## 1              4
## 2              4
## 3              4
## 4              4
## 5              4
## 6              4
```

Exploratory Data Analysis

```r
# Checking the structure of the dataset
str(Playlist)
```

```
## 'data.frame':    2300 obs. of  23 variables:
##  $ playlist_url     : chr  "https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk" "https://open.s
##  $ year             : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ track_id         : chr  "3AJwUDP919kvQ9QcozQPxg" "2m1hi0nfMR9vdGC8UcrnwU" "3y4LxiYMgDl4RethdzpmNe
##  $ track_name       : chr  "Yellow" "All The Small Things" "Breathe" "In the End" ...
##  $ track_popularity : int  91 84 69 88 74 73 88 57 80 83 ...
##  $ album            : chr  "Parachutes" "Enema Of The State" "Breathe" "Hybrid Theory (Bonus Edition
##  $ artist_id        : chr  "4gzpq5DPGxSnKTe4SA8HAU" "6FBDaR13swtiWwGhX1WQsP" "25NQNriVT2YbSW80ILRWJa
##  $ artist_name      : chr  "Coldplay" "blink-182" "Faith Hill" "Linkin Park" ...
##  $ artist_genres    : chr  "['permanent wave', 'pop']" "['alternative metal', 'modern rock', 'pop pu
##  $ artist_popularity: int  86 75 61 83 65 56 88 69 69 80 ...
##  $ danceability     : num  0.429 0.434 0.529 0.556 0.61 0.706 0.949 0.712 0.713 0.458 ...
##  $ energy           : num  0.661 0.897 0.496 0.864 0.926 0.888 0.661 0.762 0.678 0.795 ...
##  $ key              : int  11 0 7 3 8 2 5 7 5 0 ...
##  $ loudness         : num  -7.23 -4.92 -9.01 -5.87 -4.84 ...
##  $ mode             : int  1 1 1 0 0 1 0 1 0 1 ...
```

```
##  $ speechiness      : num  0.0281 0.0488 0.029 0.0584 0.0479 0.0654 0.0572 0.0326 0.102 0.0574 ...
##  $ acousticness     : num  0.00239 0.0103 0.173 0.00958 0.031 0.119 0.0302 0.026 0.273 0.00316 ...
##  $ instrumentalness : num  1.21e-04 0.00 0.00 0.00 1.20e-03 9.64e-05 0.00 0.00 0.00 2.02e-04 ...
##  $ liveness         : num  0.234 0.612 0.251 0.209 0.0821 0.07 0.0454 0.0981 0.149 0.0756 ...
##  $ valence          : num  0.285 0.684 0.278 0.4 0.861 0.714 0.76 0.842 0.734 0.513 ...
##  $ tempo            : num  173 149 137 105 173 ...
##  $ duration_ms      : int  266773 167067 250547 216880 200400 253733 284200 260560 271333 255373 ...
##  $ time_signature   : int  4 4 4 4 4 4 4 4 4 4 ...
```

```
# Checking the summary of the dataset
summary(Playlist)
```

```
##  playlist_url            year        track_id          track_name
##  Length:2300        Min.   :2000   Length:2300        Length:2300
##  Class :character   1st Qu.:2005   Class :character   Class :character
##  Mode  :character   Median :2011   Mode  :character   Mode  :character
##                     Mean   :2011
##                     3rd Qu.:2017
##                     Max.   :2022
##
##  track_popularity    album            artist_id         artist_name
##  Min.   :  0.00   Length:2300        Length:2300        Length:2300
##  1st Qu.: 66.00   Class :character   Class :character   Class :character
##  Median : 72.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 70.94
##  3rd Qu.: 79.00
##  Max.   :100.00
##
##  artist_genres      artist_popularity  danceability        energy
##  Length:2300        Min.   : 29.00    Min.   :0.1620   Min.   :0.0519
##  Class :character   1st Qu.: 65.00    1st Qu.:0.5720   1st Qu.:0.5860
##  Mode  :character   Median : 74.00    Median :0.6710   Median :0.7120
##                     Mean   : 72.87    Mean   :0.6601   Mean   :0.6930
##                     3rd Qu.: 82.00    3rd Qu.:0.7595   3rd Qu.:0.8200
##                     Max.   :100.00    Max.   :0.9750   Max.   :0.9990
##                                       NA's   :1        NA's   :1
##       key            loudness            mode         speechiness
##  Min.   : 0.000   Min.   :-21.107   Min.   :0.0000   Min.   :0.0225
##  1st Qu.: 2.000   1st Qu.: -6.824   1st Qu.:0.0000   1st Qu.:0.0380
##  Median : 5.000   Median : -5.511   Median :1.0000   Median :0.0568
##  Mean   : 5.278   Mean   : -5.784   Mean   :0.5985   Mean   :0.0978
##  3rd Qu.: 8.000   3rd Qu.: -4.364   3rd Qu.:1.0000   3rd Qu.:0.1155
##  Max.   :11.000   Max.   : -0.276   Max.   :1.0000   Max.   :0.5760
##  NA's   :1        NA's   :1         NA's   :1        NA's   :1
##   acousticness       instrumentalness       liveness          valence
##  Min.   :0.0000129   Min.   :0.0000000   Min.   :0.02100   Min.   :0.0377
##  1st Qu.:0.0165000   1st Qu.:0.0000000   1st Qu.:0.08995   1st Qu.:0.3605
##  Median :0.0689000   Median :0.0000000   Median :0.11900   Median :0.5400
##  Mean   :0.1576892   Mean   :0.0137663   Mean   :0.17262   Mean   :0.5351
##  3rd Qu.:0.2230000   3rd Qu.:0.0000544   3rd Qu.:0.22000   3rd Qu.:0.7220
##  Max.   :0.9780000   Max.   :0.9850000   Max.   :0.84300   Max.   :0.9740
##  NA's   :1           NA's   :1           NA's   :1         NA's   :1
##      tempo         duration_ms     time_signature
##  Min.   : 60.02   Min.   : 97393   Min.   :1.000
```

```
##  1st Qu.: 98.57   1st Qu.:200180   1st Qu.:4.000
##  Median :120.00   Median :221653   Median :4.000
##  Mean   :120.51   Mean   :226034   Mean   :3.982
##  3rd Qu.:137.03   3rd Qu.:245950   3rd Qu.:4.000
##  Max.   :210.86   Max.   :688453   Max.   :5.000
##  NA's   :1        NA's   :1        NA's   :1
```

```
# Checking the number of missing values in the dataset
colSums(is.na(Playlist))
```

```
##      playlist_url              year          track_id        track_name
##                 0                 0                 0                 0
##  track_popularity             album         artist_id       artist_name
##                 0                 0                 0                 0
##     artist_genres artist_popularity      danceability            energy
##                 0                 0                 1                 1
##               key          loudness              mode       speechiness
##                 1                 1                 1                 1
##      acousticness  instrumentalness          liveness           valence
##                 1                 1                 1                 1
##             tempo       duration_ms    time_signature
##                 1                 1                 1
```

```
# Removing the missing values
Playlist = Playlist[complete.cases(Playlist),]
```

```
# Checking the duplicates
Playlist[duplicated(Playlist),]
```

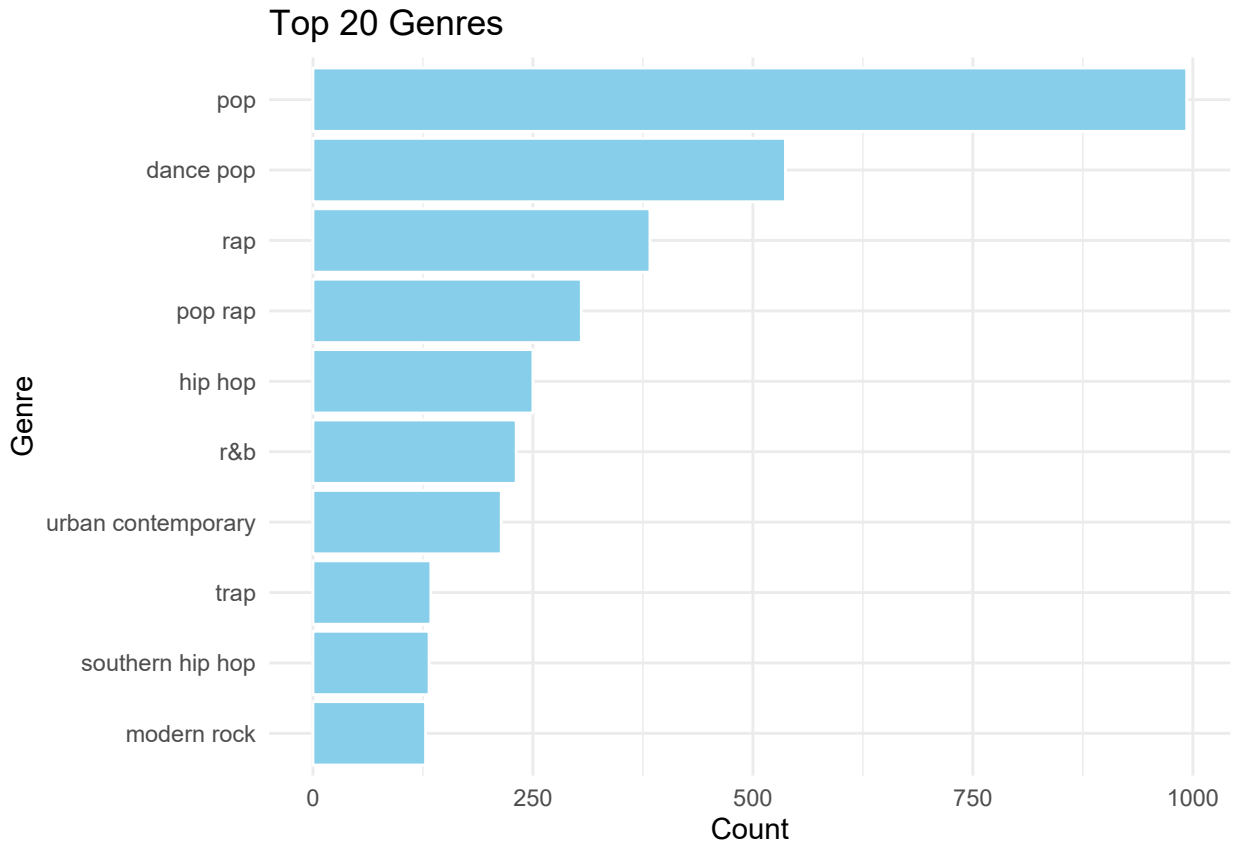
```
##  [1] playlist_url      year              track_id          track_name
##  [5] track_popularity  album             artist_id         artist_name
##  [9] artist_genres     artist_popularity danceability      energy
## [13] key               loudness          mode              speechiness
## [17] acousticness      instrumentalness  liveness          valence
## [21] tempo             duration_ms       time_signature
## <0 rows> (or 0-length row.names)
```

```
# Creating a separate genre DataFrame
Playlist_genre <- Playlist %>%
  mutate(artist_genres = strsplit(gsub("\\['|'\\]|'", "", artist_genres), ", ")) %>%
  unnest(artist_genres)

# Count the number of occurrences of each genre
genre_count <- Playlist_genre %>%
  group_by(artist_genres) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  slice_head(n = 10)

print(genre_count)
```

```
## # A tibble: 10 x 2
```

```
##    artist_genres      count
##    <chr>              <int>
##  1 pop                  993
##  2 dance pop            537
##  3 rap                  383
##  4 pop rap              305
##  5 hip hop              250
##  6 r&b                  231
##  7 urban contemporary   214
##  8 trap                 134
##  9 southern hip hop     132
## 10 modern rock          128
```

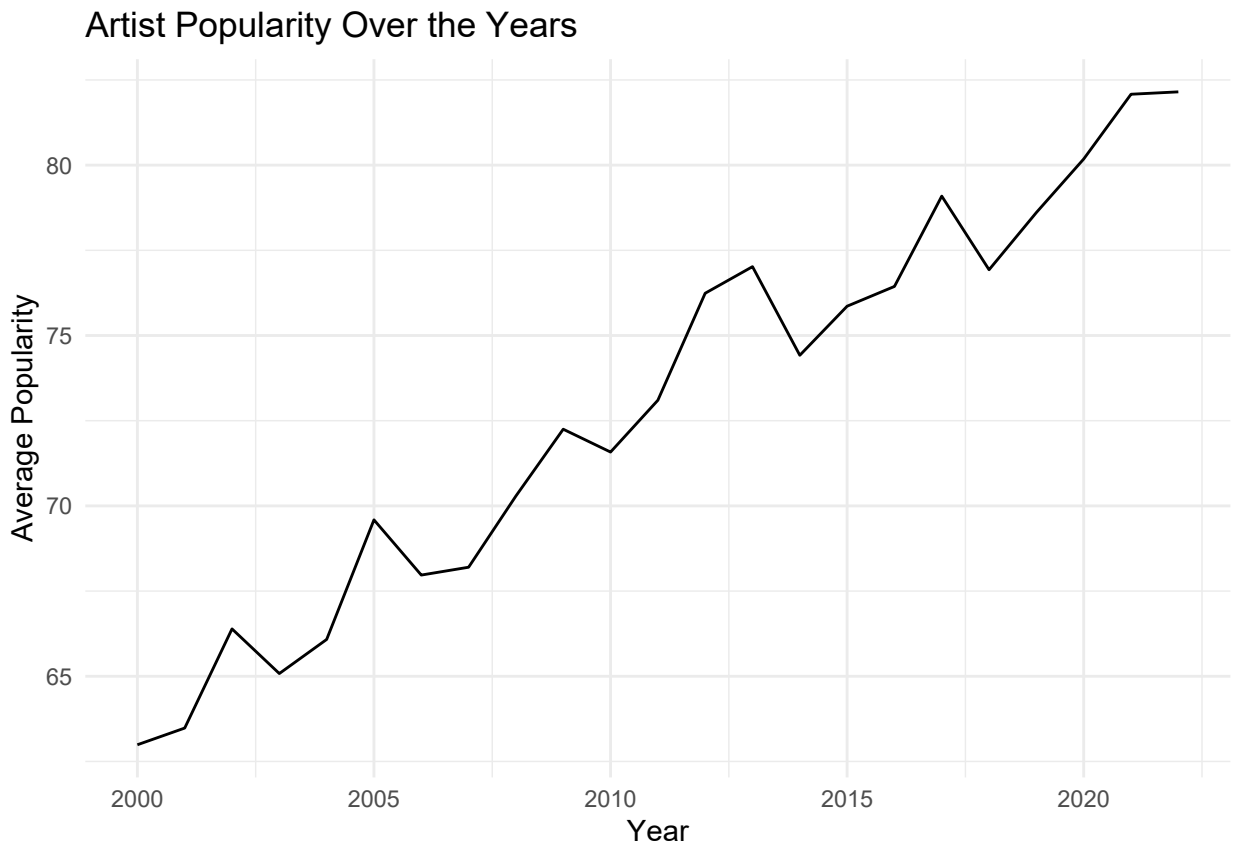
```r
# Top 10 genre counts and ordering them in descending order
ggplot(genre_count, aes(x = reorder(artist_genres, count), y = count)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "white") +
          coord_flip() +
          labs(x = "Genre", y = "Count",
          title = "Top 20 Genres") +
          theme_minimal()
```



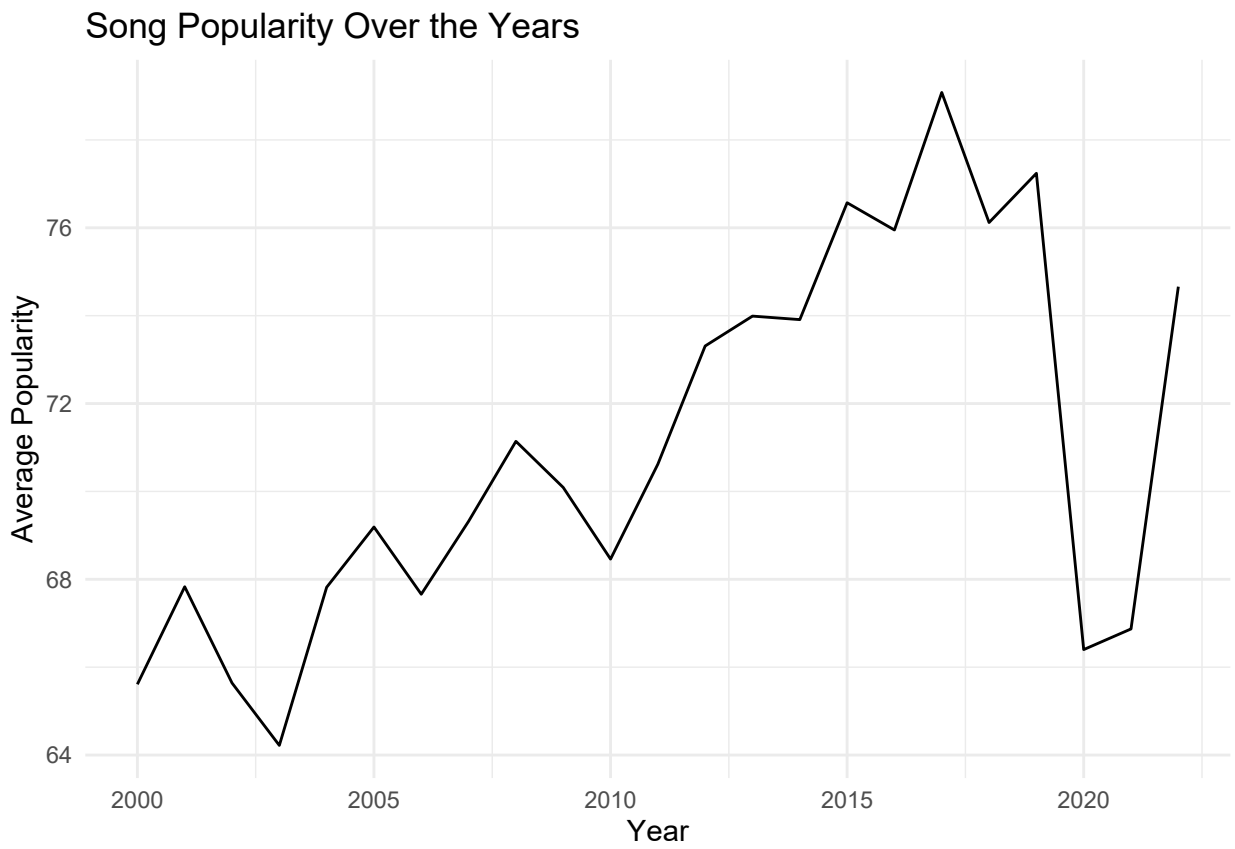

```r
# Total number of unique genres
total_genres <- Playlist_genre %>%
  summarise(total = n_distinct(artist_genres))
print(total_genres)
```

```
## # A tibble: 1 x 1
##   total
##   <int>
## 1   438
```

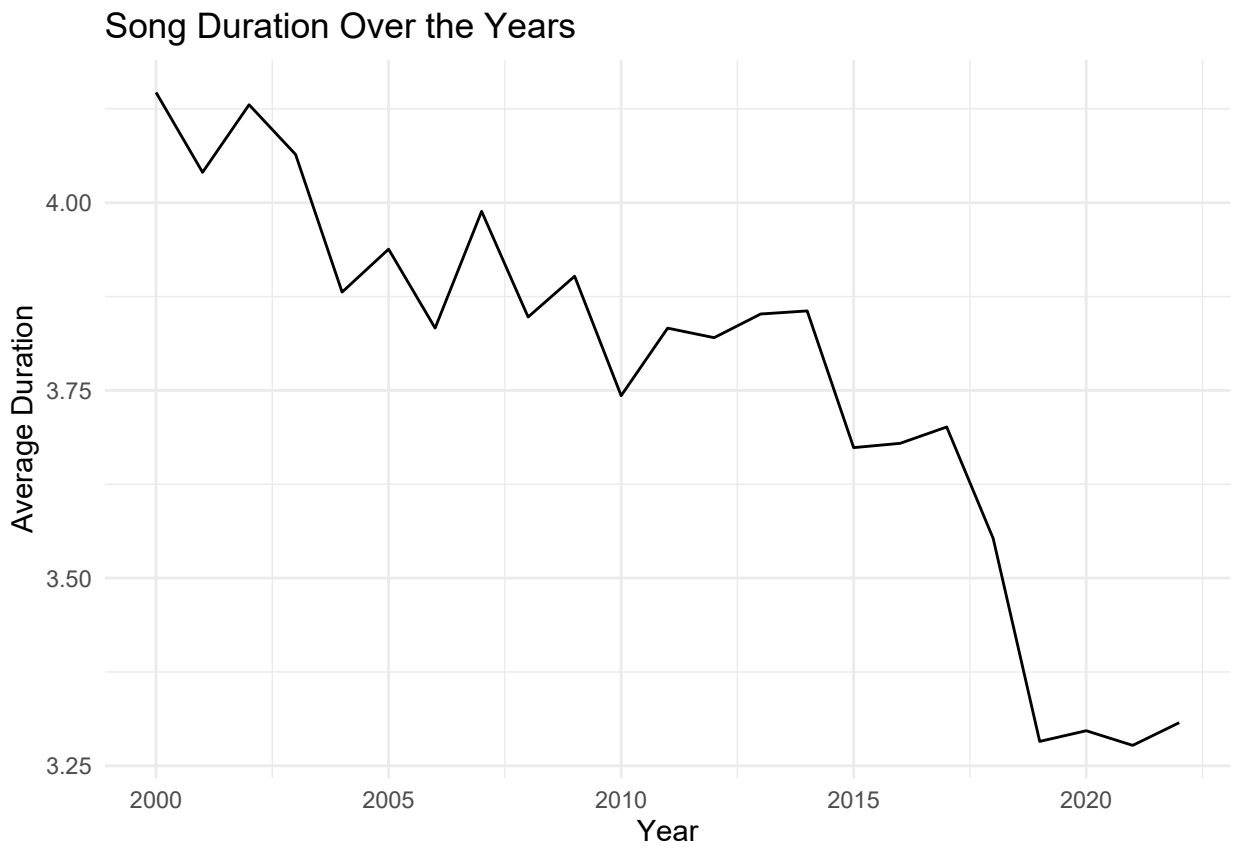
```
# Exploring artist popularity over the years
Playlist %>%
  group_by(year) %>%
  summarise(avg_popularity = mean(artist_popularity)) %>%
  ggplot(aes(x = year, y = avg_popularity)) +
  geom_line() +
  labs(x = "Year",
       y = "Average Popularity",
       title = "Artist Popularity Over the Years") +
  theme_minimal()
```




```
# Exploring the popularity of songs over the years
Playlist %>%
  group_by(year) %>%
  summarise(avg_popularity = mean(track_popularity)) %>%
  ggplot(aes(x = year, y = avg_popularity)) +
  geom_line() +
  labs(x = "Year",
       y = "Average Popularity",
       title = "Song Popularity Over the Years") +
  theme_minimal()
```




```
# Average song duration over the years
Playlist %>%
  group_by(year) %>%
  summarise(avg_duration = mean(duration_ms / 60000)) %>%
  ggplot(aes(x = year, y = avg_duration)) +
  geom_line() +
  labs(x = "Year",
       y = "Average Duration",
       title = "Song Duration Over the Years") +
  theme_minimal()
```




```
# Correlation Heatmap
corr_matrix <- cor(select_if(Playlist, is.numeric), use = "complete.obs")

# Print the entire correlation matrix
print(corr_matrix)
```

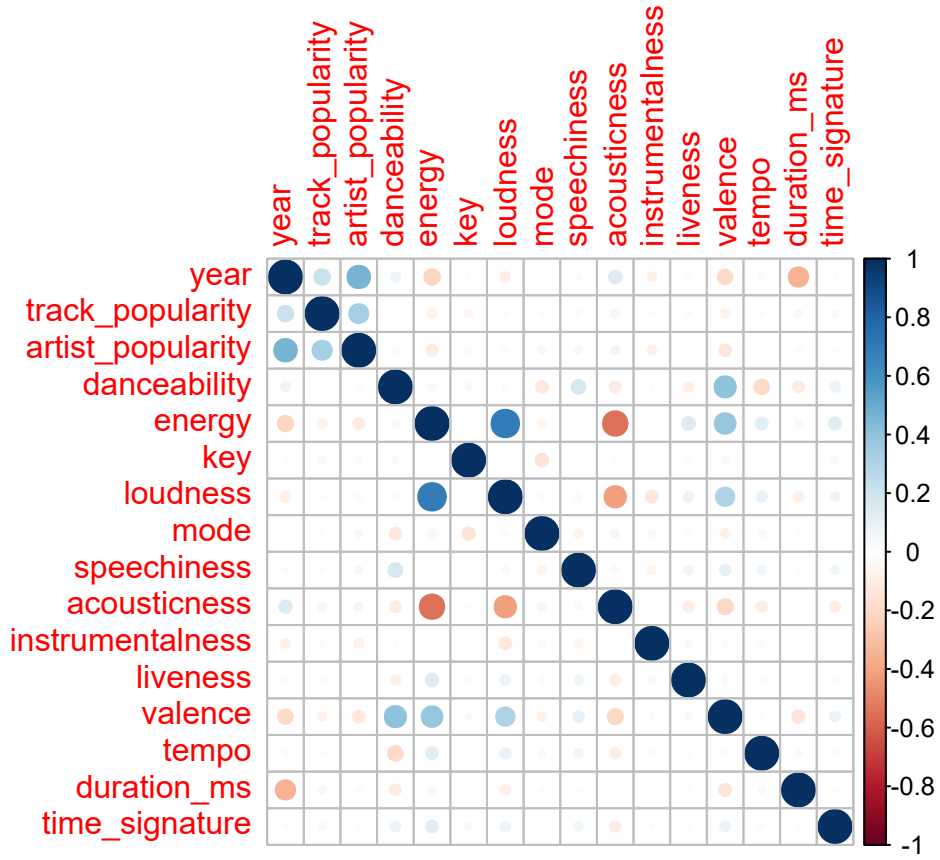
```
##                           year track_popularity artist_popularity danceability
## year                1.000000000       0.21852368       0.463722183   0.07957941
## track_popularity    0.218523683       1.00000000       0.331029402   0.00688143
## artist_popularity   0.463722183       0.33102940       1.000000000   0.02862064
## danceability        0.079579414       0.00688143       0.028620644   1.00000000
## energy             -0.213265276      -0.07428445      -0.108966441  -0.04057472
## key                -0.012309092      -0.04786503      -0.029548895   0.03666614
## loudness           -0.087450644      -0.01855829      -0.029040843   0.02888891
## mode               -0.008051859       0.01985370      -0.044216704  -0.12333694
## speechiness         0.026363893      -0.02716067       0.048928994   0.17111372
## acousticness        0.144009803       0.05831057       0.061084698  -0.10524124
## instrumentalness   -0.071126568      -0.02085383      -0.073707248   0.01821712
## liveness           -0.027719247      -0.02003324       0.004201852  -0.08465266
## valence            -0.192475655      -0.06729998      -0.125513566   0.40961227
## tempo               0.028504860      -0.01356500      -0.009966807  -0.19376982
## duration_ms        -0.340160117      -0.04395621      -0.024309904  -0.10551725
## time_signature     -0.016431016      -0.03898503      -0.010795018   0.08646119
##                          energy          key     loudness         mode
## year               -0.213265276 -0.012309092 -0.087450644 -0.008051859
```

```
## track_popularity  -0.074284445 -0.047865025 -0.018558295  0.019853700
## artist_popularity -0.108966441 -0.029548895 -0.029040843 -0.044216704
## danceability      -0.040574722  0.036666143  0.028888907 -0.123336942
## energy             1.000000000  0.004318412  0.691206365 -0.056107223
## key                0.004318412  1.000000000  0.004074097 -0.145893417
## loudness           0.691206365  0.004074097  1.000000000 -0.026780958
## mode              -0.056107223 -0.145893417 -0.026780958  1.000000000
## speechiness       -0.005980988  0.008923928 -0.038021022 -0.069799285
## acousticness      -0.543772771 -0.010714841 -0.414999064  0.054809018
## instrumentalness   0.009141740 -0.009784830 -0.124393290 -0.035061188
## liveness           0.148943940 -0.027931280  0.084871035 -0.023612778
## valence            0.388809575  0.033535494  0.307298771 -0.071995182
## tempo              0.125644919 -0.001948876  0.092810702  0.035133555
## duration_ms       -0.040421047 -0.007010603 -0.082780633  0.013255409
## time_signature     0.132878924 -0.040844670  0.073839964 -0.005917998
##                    speechiness  acousticness instrumentalness     liveness
## year               0.026363893  0.1440098033     -0.071126568 -0.027719247
## track_popularity  -0.027160667  0.0583105731     -0.020853829 -0.020033242
## artist_popularity  0.048928994  0.0610846980     -0.073707248  0.004201852
## danceability       0.171113722 -0.1052412369      0.018217123 -0.084652662
## energy            -0.005980988 -0.5437727712      0.009141740  0.148943940
## key                0.008923928 -0.0107148412     -0.009784830 -0.027931280
## loudness          -0.038021022 -0.4149990641     -0.124393290  0.084871035
## mode              -0.069799285  0.0548090179     -0.035061188 -0.023612778
## speechiness        1.000000000 -0.0372796328     -0.056314643  0.066531067
## acousticness      -0.037279633  1.0000000000      0.002361402 -0.095044581
## instrumentalness  -0.056314643  0.0023614020      1.000000000 -0.037942301
## liveness           0.066531067 -0.0950445808     -0.037942301  1.000000000
## valence            0.101179948 -0.2045924338     -0.029367032  0.034534965
## tempo              0.066760333 -0.0947119465      0.024215737  0.019185970
## duration_ms        0.014558013  0.0009593884      0.001424102  0.013070104
## time_signature     0.066111040 -0.0940301886      0.013581090  0.015489252
##                       valence        tempo   duration_ms time_signature
## year               -0.19247565  0.028504860 -0.3401601169   -0.016431016
## track_popularity   -0.06729998 -0.013565005 -0.0439562120   -0.038985033
## artist_popularity  -0.12551357 -0.009966807 -0.0243099042   -0.010795018
## danceability        0.40961227 -0.193769824 -0.1055172477    0.086461190
## energy              0.38880958  0.125644919 -0.0404210471    0.132878924
## key                 0.03353549 -0.001948876 -0.0070106035   -0.040844670
## loudness            0.30729877  0.092810702 -0.0827806331    0.073839964
## mode               -0.07199518  0.035133555  0.0132554085   -0.005917998
## speechiness         0.10117995  0.066760333  0.0145580131    0.066111040
## acousticness       -0.20459243 -0.094711946  0.0009593884   -0.094030189
## instrumentalness   -0.02936703  0.024215737  0.0014241021    0.013581090
## liveness            0.03453496  0.019185970  0.0130701044    0.015489252
## valence             1.00000000 -0.020948310 -0.1371966003    0.087011334
## tempo              -0.02094831  1.000000000 -0.0346043725   -0.024923819
## duration_ms        -0.13719660 -0.034604372  1.0000000000   -0.014692437
## time_signature      0.08701133 -0.024923819 -0.0146924368    1.000000000
```

```r
# Print the correlation between each variable and track popularity
track_popularity_corr <- corr_matrix[,'track_popularity']
print(track_popularity_corr)
```

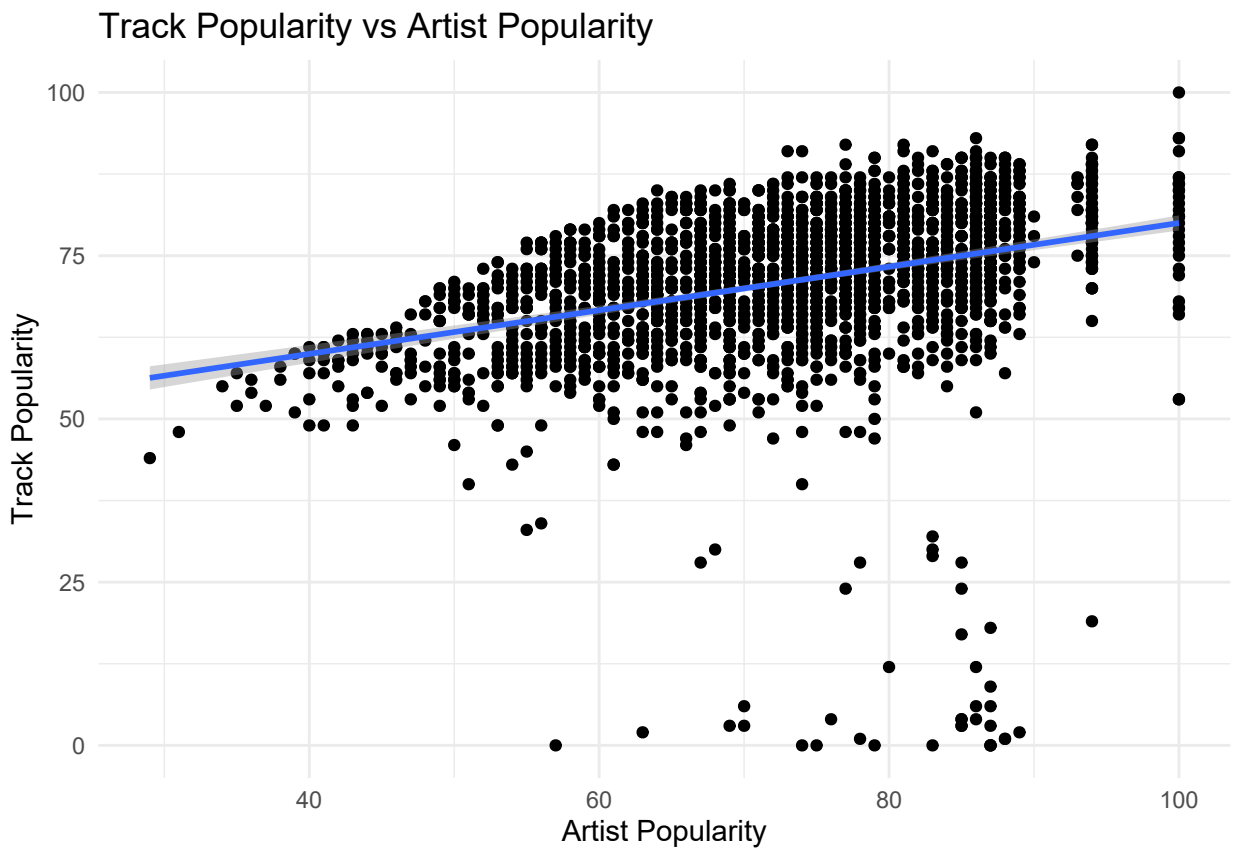
```
##              year  track_popularity artist_popularity      danceability
##        0.21852368        1.00000000        0.33102940        0.00688143
##            energy               key          loudness              mode
##       -0.07428445       -0.04786503       -0.01855829        0.01985370
##       speechiness      acousticness  instrumentalness          liveness
##       -0.02716067        0.05831057       -0.02085383       -0.02003324
##           valence             tempo       duration_ms    time_signature
##       -0.06729998       -0.01356500       -0.04395621       -0.03898503
```

```
# Plot the correlation matrix
corrplot::corrplot(corr_matrix)
```



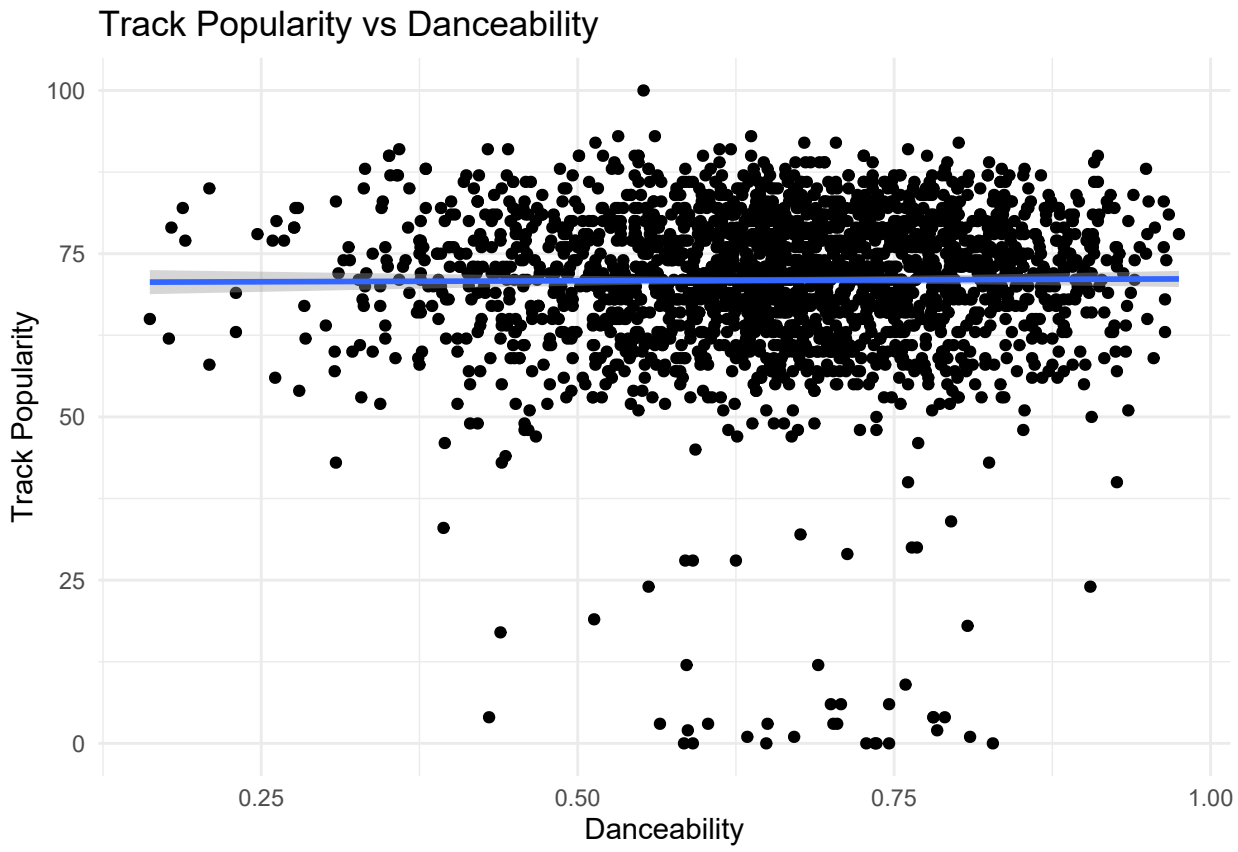

```
# Track popularity vs Artist popularity
Playlist %>%
  ggplot(aes(x = artist_popularity, y = track_popularity)) +
  geom_point() +
  labs(x = "Artist Popularity",
       y = "Track Popularity",
       title = "Track Popularity vs Artist Popularity") +
  geom_smooth(method = "lm") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



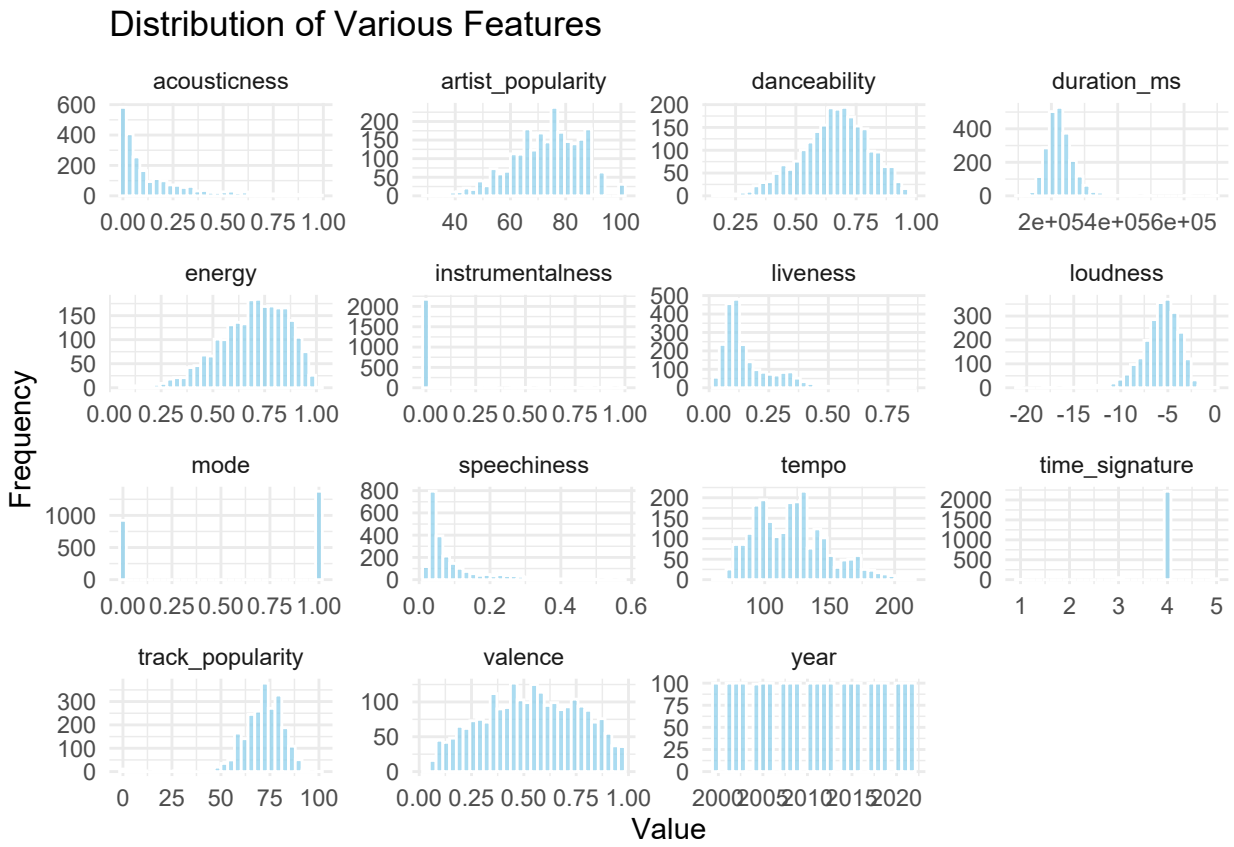

```
# Track popularity vs Danceability
Playlist %>%
  ggplot(aes(x = danceability, y = track_popularity)) +
  geom_point() +
  labs(x = "Danceability",
       y = "Track Popularity",
       title = "Track Popularity vs Danceability") +
    geom_smooth(method = "lm") +
    theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Exploring the distribution of the features
Playlist %>%
  select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram(fill = "skyblue", color = "white", alpha = 0.7) +
  facet_wrap(~key, scales = "free") +
  theme_minimal() +
    labs(title = "Distribution of Various Features", x = "Value", y = "Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

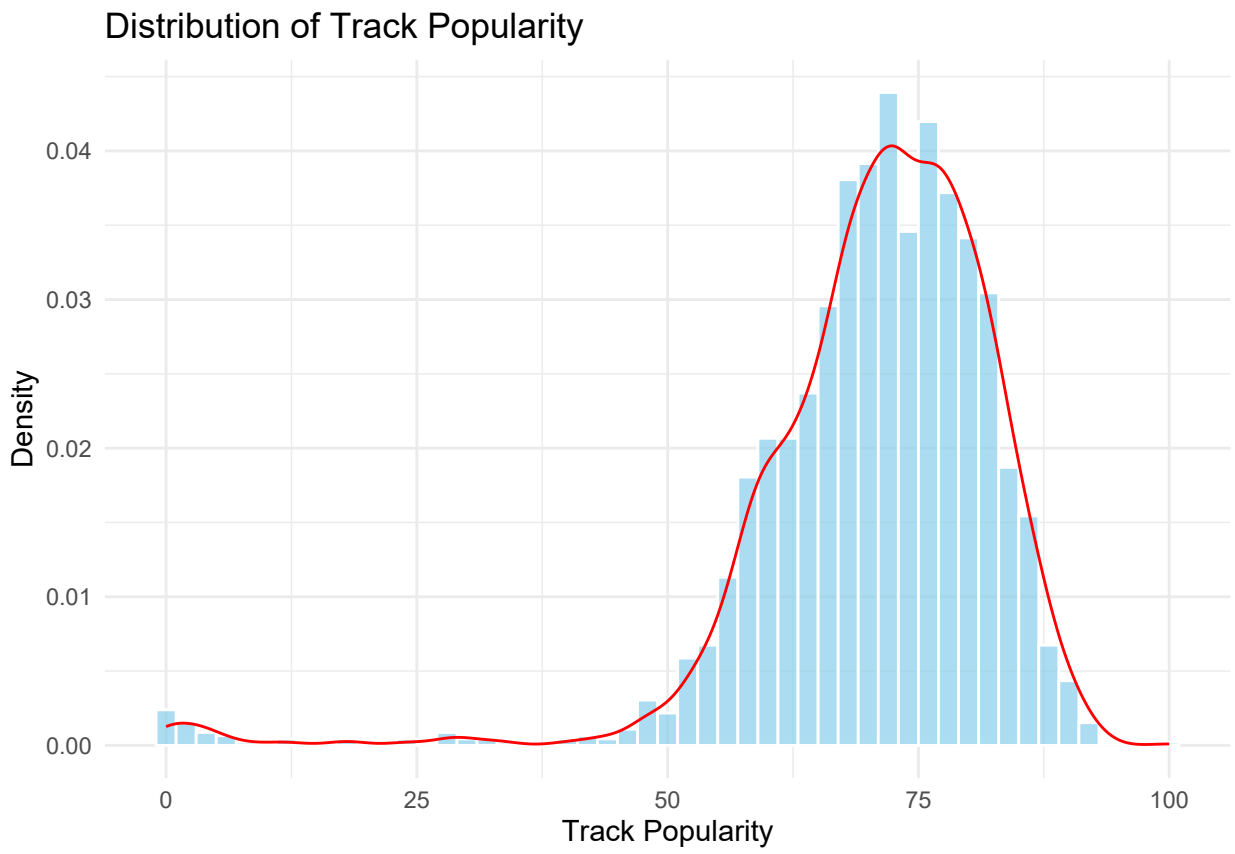


Hypothesis Formulation

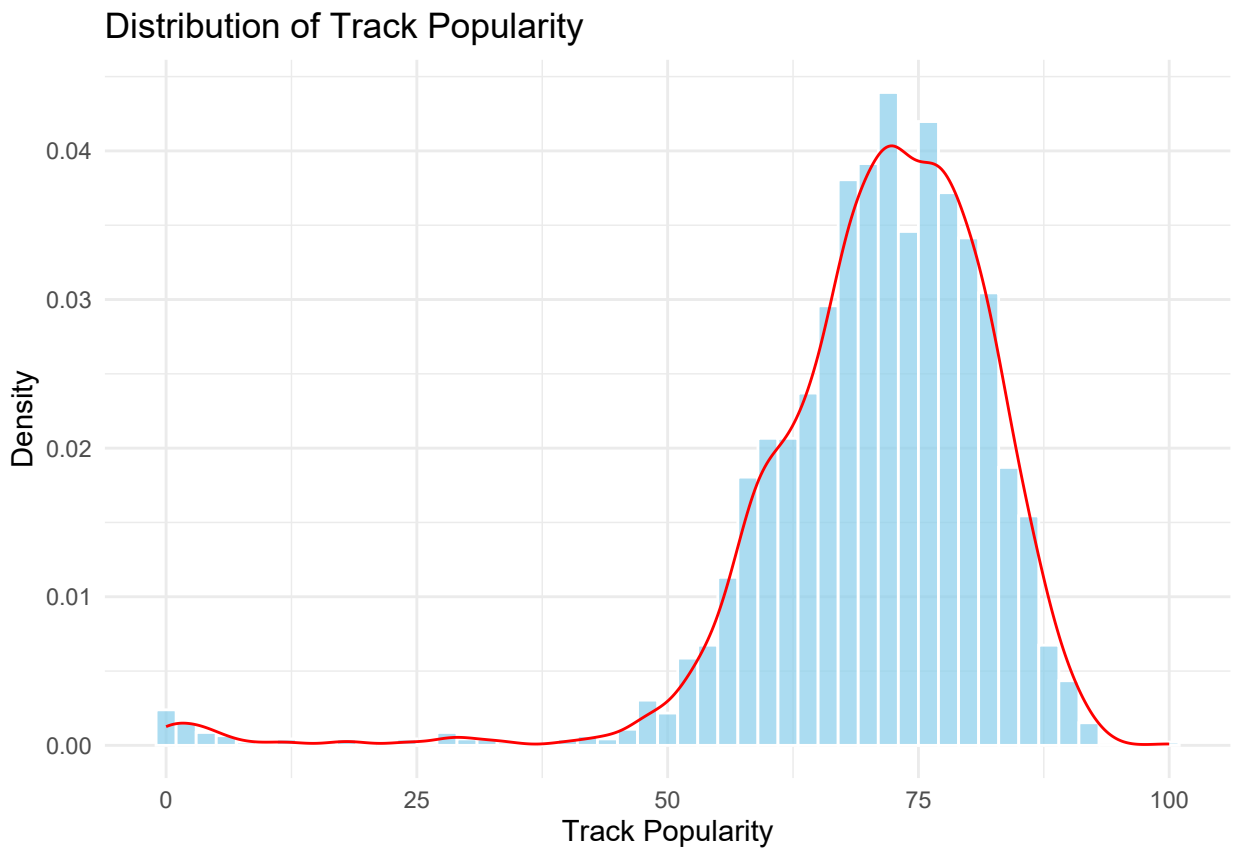
H1: Songs with higher danceability are more popular. H2: Over the years, the average duration of popular songs has decreased. H3: Artist popularity is a significant predictor of track popularity.

Feature Engineering and Model Building

```
# Checking the distribution and normality of the target variable
ggplot(Playlist, aes(x = track_popularity)) +
  geom_histogram(fill = "skyblue", color = "white", alpha = 0.7, binwidth = 2, aes(y = after_stat(densi
  geom_line(stat = "density", color = "red") +
  theme_minimal() +
  labs(title = "Distribution of Track Popularity", x = "Track Popularity", y = "Density")
```




```r
# Checking the distribution and normality of the target variable
ggplot(Playlist, aes(x = track_popularity)) +
  geom_histogram(fill = "skyblue", color = "white", alpha = 0.7, binwidth = 2, aes(y = after_stat(densi
  geom_line(stat = "density", color = "red") +
  theme_minimal() +
  labs(title = "Distribution of Track Popularity", x = "Track Popularity", y = "Density")
```




```
# Convert year and genre to factors
Playlist$year <- as.factor(Playlist$year)

# Creating a new column with the number of genres listed for each track
Playlist$num_genres <- sapply(strsplit(Playlist$artist_genres, ", "), length)

# Creating dummy variables for the top 10 genres
Playlist$pop_genre <- grepl("pop", Playlist$artist_genres, ignore.case = TRUE)
Playlist$dance_pop_genre <- grepl("dance pop", Playlist$artist_genres, ignore.case = TRUE)
Playlist$rap_genre <- grepl("rap", Playlist$artist_genres, ignore.case = TRUE)
Playlist$pop_rap_genre <- grepl("pop rap", Playlist$artist_genres, ignore.case = TRUE)
Playlist$hip_hop_genre <- grepl("hip hop", Playlist$artist_genres, ignore.case = TRUE)
Playlist$rnb_genre <- grepl("r&b", Playlist$artist_genres, ignore.case = TRUE)
Playlist$urban_contemporary_genre <- grepl("urban contemporary", Playlist$artist_genres, ignore.case = 
Playlist$trap_genre <- grepl("trap", Playlist$artist_genres, ignore.case = TRUE)
Playlist$southern_hip_hop_genre <- grepl("southern hip hop", Playlist$artist_genres, ignore.case = TRUE
Playlist$modern_rock_genre <- grepl("modern rock", Playlist$artist_genres, ignore.case = TRUE)

str(Playlist)
```

```
## 'data.frame':    2299 obs. of  34 variables:
##  $ playlist_url            : chr  "https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk" "https:/
##  $ year                    : Factor w/ 23 levels "2000","2001",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ track_id                : chr  "3AJwUDP919kvQ9QcozQPxg" "2m1hi0nfMR9vdGC8UcrnwU" "3y4LxiYMgDl4Retl
##  $ track_name              : chr  "Yellow" "All The Small Things" "Breathe" "In the End" ...
```

```
##  $ track_popularity        : int  91 84 69 88 74 73 88 57 80 83 ...
##  $ album                   : chr  "Parachutes" "Enema Of The State" "Breathe" "Hybrid Theory (Bonus
##  $ artist_id               : chr  "4gzpq5DPGxSnKTe4SA8HAU" "6FBDaR13swtiWwGhX1WQsP" "25NQNriVT2YbSW8
##  $ artist_name             : chr  "Coldplay" "blink-182" "Faith Hill" "Linkin Park" ...
##  $ artist_genres           : chr  "['permanent wave', 'pop']" "['alternative metal', 'modern rock',
##  $ artist_popularity       : int  86 75 61 83 65 56 88 69 69 80 ...
##  $ danceability            : num  0.429 0.434 0.529 0.556 0.61 0.706 0.949 0.712 0.713 0.458 ...
##  $ energy                  : num  0.661 0.897 0.496 0.864 0.926 0.888 0.661 0.762 0.678 0.795 ...
##  $ key                     : int  11 0 7 3 8 2 5 7 5 0 ...
##  $ loudness                : num  -7.23 -4.92 -9.01 -5.87 -4.84 ...
##  $ mode                    : int  1 1 1 0 0 1 0 1 0 1 ...
##  $ speechiness             : num  0.0281 0.0488 0.029 0.0584 0.0479 0.0654 0.0572 0.0326 0.102 0.057
##  $ acousticness            : num  0.00239 0.0103 0.173 0.00958 0.031 0.119 0.0302 0.026 0.273 0.0031
##  $ instrumentalness        : num  1.21e-04 0.00 0.00 0.00 1.20e-03 9.64e-05 0.00 0.00 0.00 2.02e-04
##  $ liveness                : num  0.234 0.612 0.251 0.209 0.0821 0.07 0.0454 0.0981 0.149 0.0756 ...
##  $ valence                 : num  0.285 0.684 0.278 0.4 0.861 0.714 0.76 0.842 0.734 0.513 ...
##  $ tempo                   : num  173 149 137 105 173 ...
##  $ duration_ms             : int  266773 167067 250547 216880 200400 253733 284200 260560 271333 253
##  $ time_signature          : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ num_genres              : int  2 6 4 5 3 5 3 2 5 5 ...
##  $ pop_genre               : logi  TRUE TRUE FALSE FALSE TRUE TRUE ...
##  $ dance_pop_genre         : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ rap_genre               : logi  FALSE FALSE FALSE TRUE FALSE TRUE ...
##  $ pop_rap_genre           : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hip_hop_genre           : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ rnb_genre               : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ urban_contemporary_genre: logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ trap_genre              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ southern_hip_hop_genre  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ modern_rock_genre       : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
```

```r
# Scaling the dataset
Playlist <- Playlist %>%
  mutate(across(where(is.numeric), scale))
```

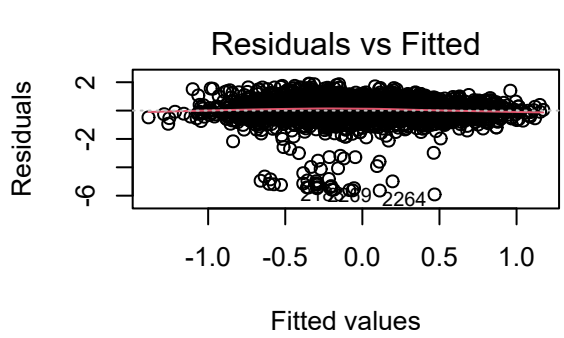
Approach A: Build the linear regression model

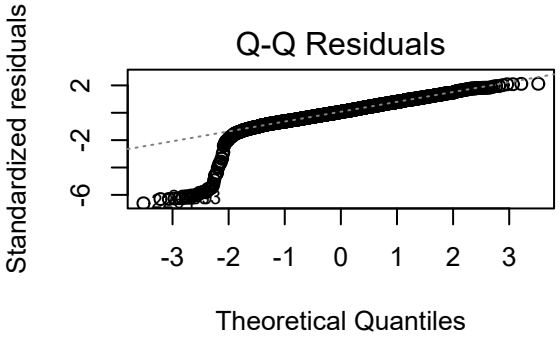
```r
lm_model <- lm(track_popularity ~ year + artist_popularity + danceability + energy + acousticness + dura

summary(lm_model)
```

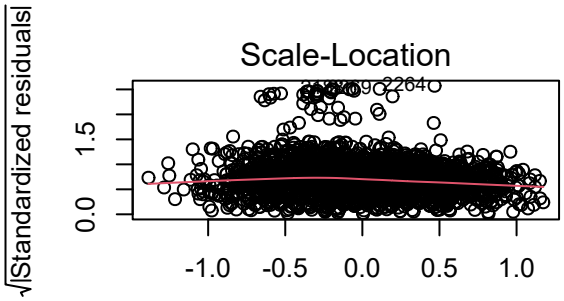
```
##
## Call:
## lm(formula = track_popularity ~ year + artist_popularity + danceability +
##     energy + acousticness + duration_ms + pop_genre + dance_pop_genre +
##     rap_genre + pop_rap_genre + hip_hop_genre + rnb_genre + urban_contemporary_genre +
##     trap_genre + southern_hip_hop_genre + modern_rock_genre,
##     data = Playlist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

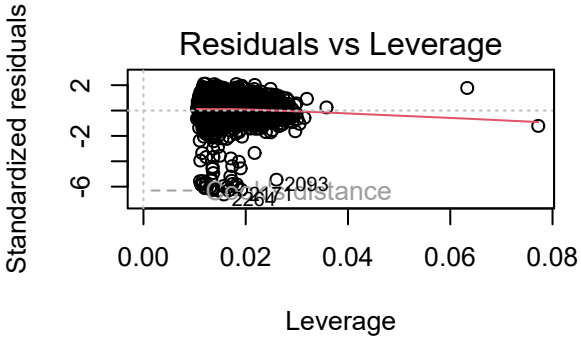
```
## -5.9135 -0.3644  0.0514  0.5018  1.8977
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.08116    0.09992  -0.812 0.416741
## year2001                      0.15318    0.12801   1.197 0.231554
## year2002                     -0.07692    0.12869  -0.598 0.550081
## year2003                     -0.18247    0.12857  -1.419 0.155981
## year2004                      0.03259    0.12988   0.251 0.801895
## year2005                      0.06121    0.13086   0.468 0.639984
## year2006                     -0.03256    0.12996  -0.251 0.802185
## year2007                      0.12569    0.12944   0.971 0.331651
## year2008                      0.21258    0.13004   1.635 0.102251
## year2009                      0.09187    0.13065   0.703 0.482017
## year2010                      0.01006    0.13146   0.076 0.939034
## year2011                      0.12456    0.13078   0.952 0.340962
## year2012                      0.24527    0.13263   1.849 0.064542 .
## year2013                      0.24331    0.13325   1.826 0.067996 .
## year2014                      0.32512    0.13281   2.448 0.014442 *
## year2015                      0.47341    0.13386   3.537 0.000413 ***
## year2016                      0.39888    0.13443   2.967 0.003037 **
## year2017                      0.57079    0.13587   4.201 2.76e-05 ***
## year2018                      0.37520    0.13605   2.758 0.005867 **
## year2019                      0.37928    0.13868   2.735 0.006286 **
## year2020                     -0.52183    0.13884  -3.759 0.000175 ***
## year2021                     -0.52754    0.13926  -3.788 0.000156 ***
## year2022                      0.10259    0.13888   0.739 0.460173
## artist_popularity             0.31278    0.02296  13.624  < 2e-16 ***
## danceability                  0.02277    0.02101   1.084 0.278584
## energy                       -0.02806    0.02383  -1.178 0.239110
## acousticness                  0.04138    0.02317   1.786 0.074217 .
## duration_ms                  -0.03625    0.02112  -1.717 0.086146 .
## pop_genreTRUE                -0.04904    0.05628  -0.871 0.383724
## dance_pop_genreTRUE          -0.05477    0.05244  -1.044 0.296483
## rap_genreTRUE                 0.08162    0.07541   1.082 0.279216
## pop_rap_genreTRUE            -0.07742    0.08432  -0.918 0.358655
## hip_hop_genreTRUE            -0.06726    0.07082  -0.950 0.342342
## rnb_genreTRUE                 0.00367    0.07471   0.049 0.960827
## urban_contemporary_genreTRUE -0.14386    0.08718  -1.650 0.099056 .
## trap_genreTRUE                0.12465    0.08656   1.440 0.150010
## southern_hip_hop_genreTRUE    0.01275    0.10677   0.119 0.904940
## modern_rock_genreTRUE         0.37007    0.08673   4.267 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.901 on 2261 degrees of freedom
## Multiple R-squared:  0.2013, Adjusted R-squared:  0.1883
## F-statistic:  15.4 on 37 and 2261 DF,  p-value: < 2.2e-16
```

```r
# Diagnostic plots to check assumptions
par(mfrow = c(2,2))
plot(lm_model)
```

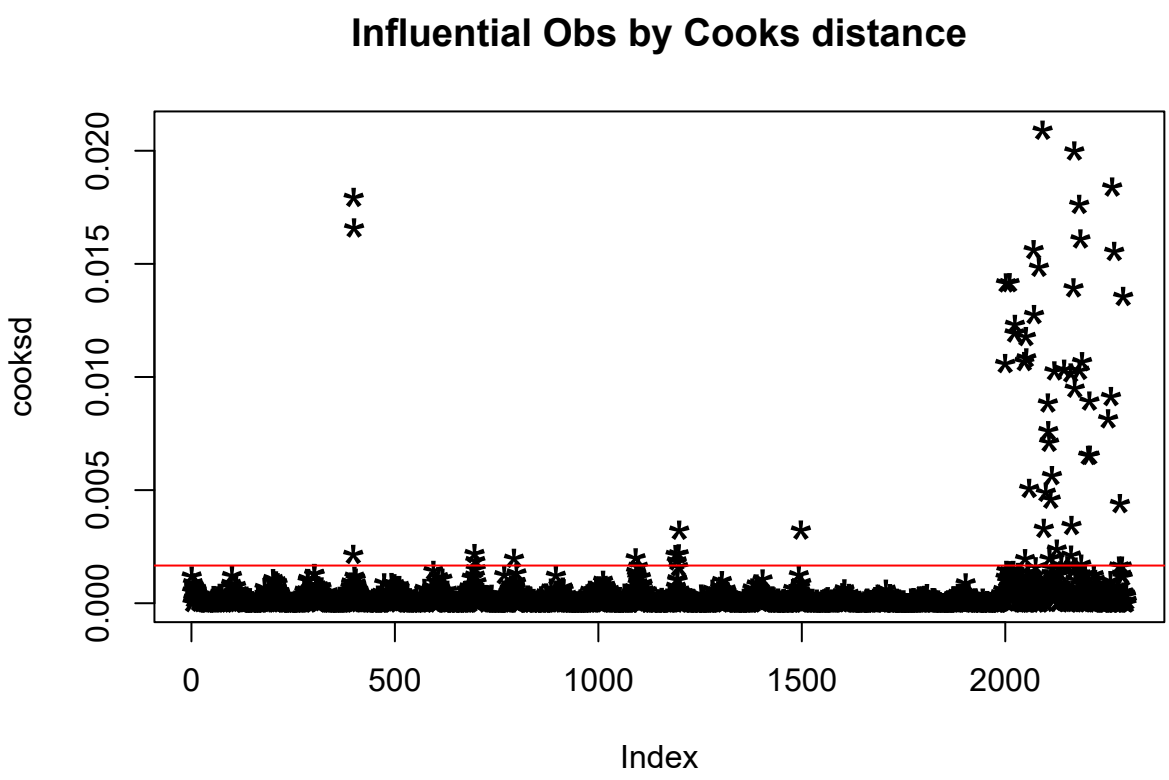









```r
# Detecting outliers using Cook's distance
cooksd <- cooks.distance(lm_model)
plot(cooksd, pch = "*", cex = 2, main = "Influential Obs by Cooks distance")  # plot cook's distance
abline(h = 4*mean(cooksd, na.rm = TRUE), col = "red")  # add cutoff line
```




```
# Print the influential observations
influential_obs <- which(cooksd > 4*mean(cooksd, na.rm = TRUE))
print(influential_obs)
```

```
##  398  399  400  697  699  794 1093 1191 1194 1199 1200 1499 2001 2003 2012 2024
##  398  399  400  696  698  793 1092 1190 1193 1198 1199 1498 2000 2002 2011 2023
## 2025 2048 2050 2052 2053 2060 2071 2072 2084 2093 2096 2101 2106 2107 2109 2110
## 2024 2047 2049 2051 2052 2059 2070 2071 2083 2092 2095 2100 2105 2106 2108 2109
## 2113 2116 2122 2128 2139 2146 2162 2163 2164 2169 2171 2172 2183 2184 2186 2188
## 2112 2115 2121 2127 2138 2145 2161 2162 2163 2168 2170 2171 2182 2183 2185 2187
## 2190 2205 2208 2209 2254 2261 2264 2269 2283 2291
## 2189 2204 2207 2208 2253 2260 2263 2268 2282 2290
```

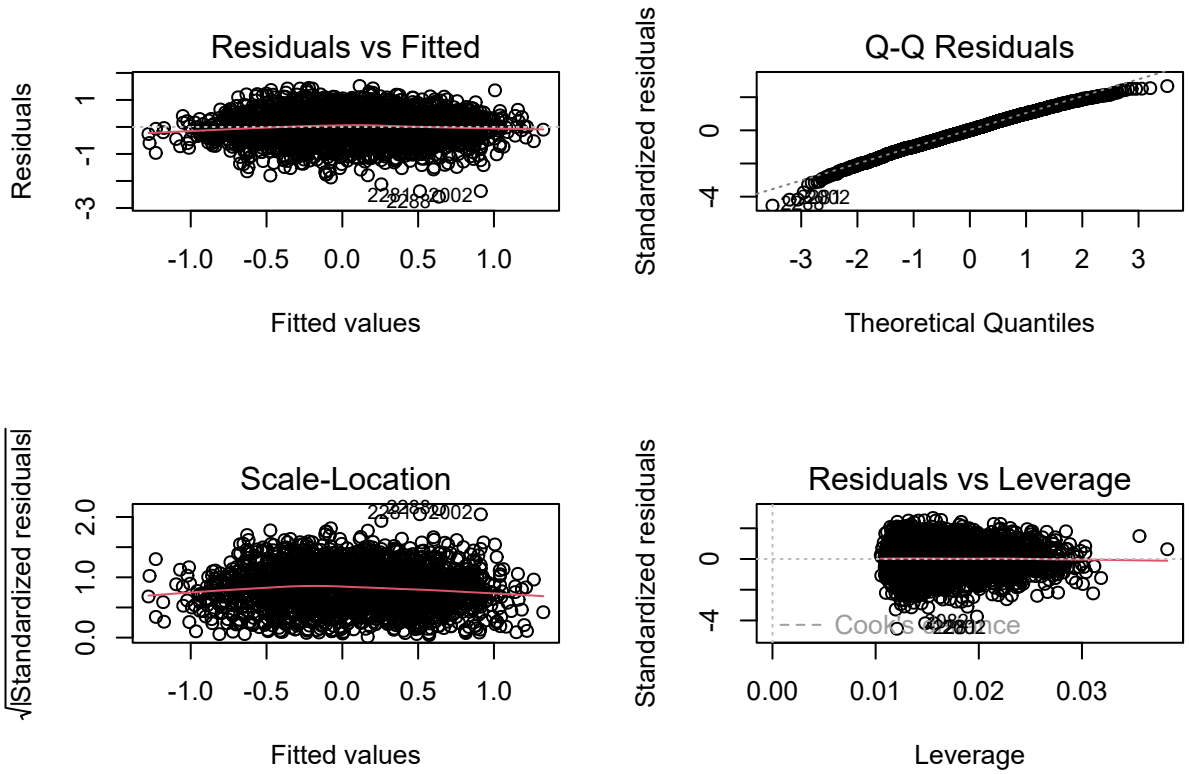
```
# Remove the influential observations
Playlist <- Playlist[-influential_obs, ]
```

```
# Approach B: Re-build the linear regression model
new_model <- lm(track_popularity ~ year + artist_popularity + danceability + energy + acousticness + du
summary(new_model)
```

```
##
## Call:
## lm(formula = track_popularity ~ year + artist_popularity + danceability +
##     energy + acousticness + duration_ms + pop_genre + dance_pop_genre +
```

```
##     rap_genre + pop_rap_genre + hip_hop_genre + rnb_genre + urban_contemporary_genre +
##     trap_genre + southern_hip_hop_genre + modern_rock_genre,
##     data = Playlist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58438 -0.37403  0.00912  0.40508  1.51717
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.069717   0.063680  -1.095  0.27373
## year2001                       0.149231   0.081381   1.834  0.06683 .
## year2002                      -0.076476   0.081820  -0.935  0.35005
## year2003                      -0.053249   0.082361  -0.647  0.51801
## year2004                       0.033131   0.082610   0.401  0.68842
## year2005                       0.058125   0.083249   0.698  0.48512
## year2006                       0.004231   0.083084   0.051  0.95939
## year2007                       0.148299   0.082456   1.799  0.07223 .
## year2008                       0.223504   0.082729   2.702  0.00695 **
## year2009                       0.094663   0.083107   1.139  0.25481
## year2010                       0.030008   0.083852   0.358  0.72047
## year2011                       0.208167   0.084112   2.475  0.01340 *
## year2012                       0.245125   0.084407   2.904  0.00372 **
## year2013                       0.250838   0.084809   2.958  0.00313 **
## year2014                       0.358578   0.084886   4.224 2.50e-05 ***
## year2015                       0.494396   0.085252   5.799 7.62e-09 ***
## year2016                       0.420904   0.085623   4.916 9.50e-07 ***
## year2017                       0.599339   0.086558   6.924 5.73e-12 ***
## year2018                       0.408159   0.086728   4.706 2.68e-06 ***
## year2019                       0.420124   0.088547   4.745 2.22e-06 ***
## year2020                       0.273540   0.092062   2.971  0.00300 **
## year2021                       0.337852   0.094539   3.574  0.00036 ***
## year2022                       0.570690   0.090751   6.289 3.85e-10 ***
## artist_popularity              0.315944   0.014729  21.451  < 2e-16 ***
## danceability                   0.017227   0.013541   1.272  0.20344
## energy                        -0.018882   0.015290  -1.235  0.21701
## acousticness                   0.040898   0.015017   2.723  0.00651 **
## duration_ms                   -0.036244   0.014527  -2.495  0.01267 *
## pop_genreTRUE                 -0.065842   0.036188  -1.819  0.06898 .
## dance_pop_genreTRUE           -0.073326   0.033675  -2.177  0.02955 *
## rap_genreTRUE                  0.111382   0.048873   2.279  0.02276 *
## pop_rap_genreTRUE             -0.041461   0.054098  -0.766  0.44352
## hip_hop_genreTRUE             -0.103041   0.045808  -2.249  0.02458 *
## rnb_genreTRUE                 -0.083278   0.048274  -1.725  0.08465 .
## urban_contemporary_genreTRUE -0.056809   0.055800  -1.018  0.30875
## trap_genreTRUE                -0.075708   0.055304  -1.369  0.17115
## southern_hip_hop_genreTRUE     0.123439   0.068114   1.812  0.07009 .
## modern_rock_genreTRUE          0.376441   0.056122   6.708 2.51e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5727 on 2203 degrees of freedom
## Multiple R-squared:  0.3974, Adjusted R-squared:  0.3872
## F-statistic: 39.26 on 37 and 2203 DF,  p-value: < 2.2e-16
```

```r
# Diagnostic plots to check assumptions
par(mfrow = c(2,2))
plot(new_model)
```




```r
# Convert categorical variables to numeric using one-hot encoding
categorical_vars <- c("year", "pop_genre", "dance_pop_genre", "rap_genre", "pop_rap_genre",
                      "hip_hop_genre", "rnb_genre", "urban_contemporary_genre",
                      "trap_genre", "southern_hip_hop_genre", "modern_rock_genre")

for(var in categorical_vars){
  Playlist[[var]] <- as.factor(Playlist[[var]])
}
```

```r
# Check if the variables are converted to factors
column_names <- colnames(Playlist)
print(column_names)
```

```
##  [1] "playlist_url"              "year"
##  [3] "track_id"                  "track_name"
##  [5] "track_popularity"          "album"
##  [7] "artist_id"                 "artist_name"
##  [9] "artist_genres"             "artist_popularity"
## [11] "danceability"              "energy"
## [13] "key"                       "loudness"
```

```
## [15] "mode"                       "speechiness"
## [17] "acousticness"               "instrumentalness"
## [19] "liveness"                   "valence"
## [21] "tempo"                      "duration_ms"
## [23] "time_signature"             "num_genres"
## [25] "pop_genre"                  "dance_pop_genre"
## [27] "rap_genre"                  "pop_rap_genre"
## [29] "hip_hop_genre"              "rnb_genre"
## [31] "urban_contemporary_genre" "trap_genre"
## [33] "southern_hip_hop_genre"    "modern_rock_genre"
```

```
result_list <- list()
for (column_name in column_names) {
  result_list[[column_name]] <- is.factor(Playlist[[column_name]])
}

print(result_list)
```

```
## $playlist_url
## [1] FALSE
##
## $year
## [1] TRUE
##
## $track_id
## [1] FALSE
##
## $track_name
## [1] FALSE
##
## $track_popularity
## [1] FALSE
##
## $album
## [1] FALSE
##
## $artist_id
## [1] FALSE
##
## $artist_name
## [1] FALSE
##
## $artist_genres
## [1] FALSE
##
## $artist_popularity
## [1] FALSE
##
## $danceability
## [1] FALSE
##
## $energy
## [1] FALSE
##
```

```
## $key
## [1] FALSE
##
## $loudness
## [1] FALSE
##
## $mode
## [1] FALSE
##
## $speechiness
## [1] FALSE
##
## $acousticness
## [1] FALSE
##
## $instrumentalness
## [1] FALSE
##
## $liveness
## [1] FALSE
##
## $valence
## [1] FALSE
##
## $tempo
## [1] FALSE
##
## $duration_ms
## [1] FALSE
##
## $time_signature
## [1] FALSE
##
## $num_genres
## [1] FALSE
##
## $pop_genre
## [1] TRUE
##
## $dance_pop_genre
## [1] TRUE
##
## $rap_genre
## [1] TRUE
##
## $pop_rap_genre
## [1] TRUE
##
## $hip_hop_genre
## [1] TRUE
##
## $rnb_genre
## [1] TRUE
##
```

```
## $urban_contemporary_genre
## [1] TRUE
##
## $trap_genre
## [1] TRUE
##
## $southern_hip_hop_genre
## [1] TRUE
##
## $modern_rock_genre
## [1] TRUE
```

```
# Remove unnecessary variables
unnecessary_vars <- c("playlist_url", "track_id", "track_name", "album", "artist_id", "artist_name", "a
Playlist_filtered <- Playlist[ , !names(Playlist) %in% unnecessary_vars]

# Encoding categorical variables using one-hot encoding - model.matrix()
Playlist_filtered <- model.matrix(~ . - 1, data = Playlist_filtered)

# Convert Playlist_filtered back to a data frame
Playlist_filtered <- as.data.frame(Playlist_filtered)
is.data.frame(Playlist_filtered)
```

```
## [1] TRUE
```

Approach B: Re-build the linear regression model with added features

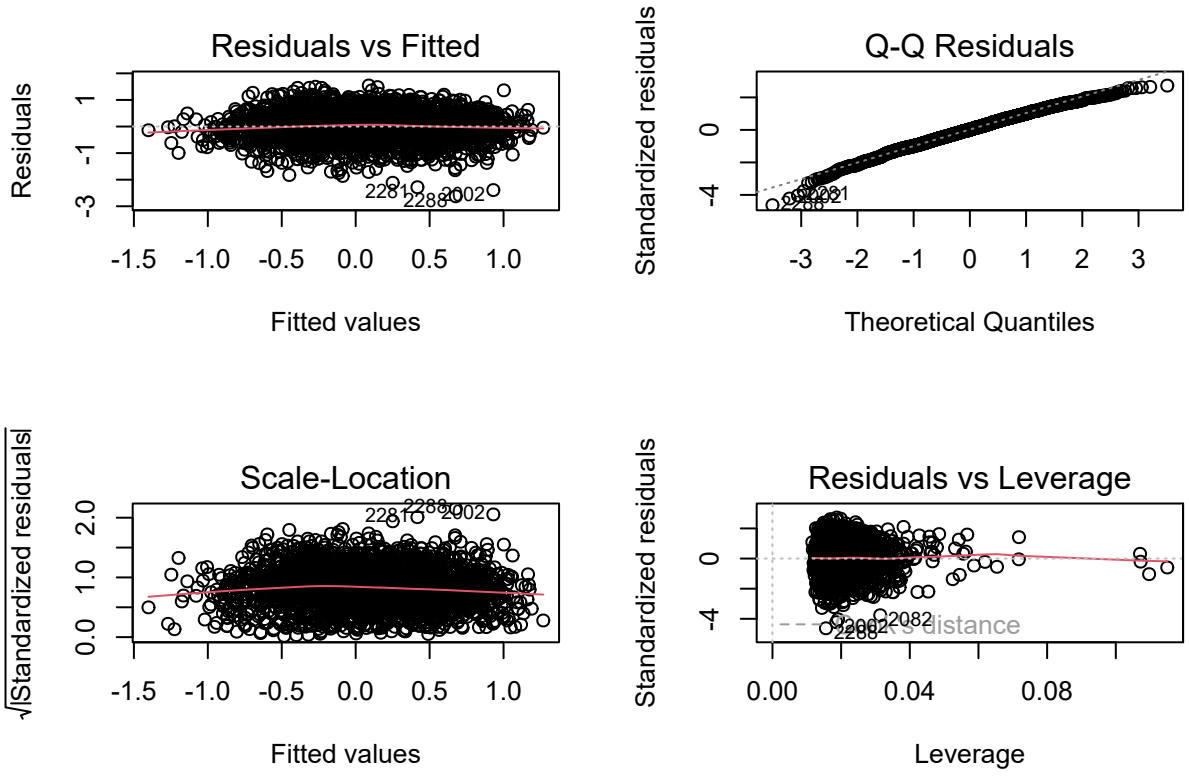
```
model_filtered <- lm(track_popularity ~ year2000 + year2001 + year2002 + year2003 +
                       year2004 + year2005 + year2006 + year2007 + year2008 + year2009 +
                       year2010 + year2011 + year2012 + year2013 + year2014 + year2015 +
                       year2016 + year2017 + year2018 + year2019 + year2020 + year2021 +
                       year2022 + artist_popularity + danceability + energy + key +
                       loudness + mode + speechiness + acousticness + instrumentalness +
                       liveness + valence + tempo + duration_ms + time_signature +
                       pop_genreTRUE + dance_pop_genreTRUE + rap_genreTRUE +
                       pop_rap_genreTRUE + hip_hop_genreTRUE + rnb_genreTRUE +
                       urban_contemporary_genreTRUE + trap_genreTRUE +
                       southern_hip_hop_genreTRUE + modern_rock_genreTRUE,
                       data = Playlist_filtered)

# Step 4: Check the summary of the new model
summary(model_filtered)
```

```
##
## Call:
## lm(formula = track_popularity ~ year2000 + year2001 + year2002 +
##     year2003 + year2004 + year2005 + year2006 + year2007 + year2008 +
##     year2009 + year2010 + year2011 + year2012 + year2013 + year2014 +
##     year2015 + year2016 + year2017 + year2018 + year2019 + year2020 +
##     year2021 + year2022 + artist_popularity + danceability +
##     energy + key + loudness + mode + speechiness + acousticness +
```

```
##     instrumentalness + liveness + valence + tempo + duration_ms +
##     time_signature + pop_genreTRUE + dance_pop_genreTRUE + rap_genreTRUE +
##     pop_rap_genreTRUE + hip_hop_genreTRUE + rnb_genreTRUE + urban_contemporary_genreTRUE +
##     trap_genreTRUE + southern_hip_hop_genreTRUE + modern_rock_genreTRUE,
##     data = Playlist_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62467 -0.37360  0.01391  0.40120  1.54283
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.851e-01  6.944e-02   6.986 3.74e-12 ***
## year2000                   -5.714e-01  9.125e-02  -6.262 4.56e-10 ***
## year2001                   -4.220e-01  9.077e-02  -4.649 3.53e-06 ***
## year2002                   -6.445e-01  8.952e-02  -7.200 8.26e-13 ***
## year2003                   -6.227e-01  9.037e-02  -6.890 7.25e-12 ***
## year2004                   -5.441e-01  8.811e-02  -6.175 7.84e-10 ***
## year2005                   -5.062e-01  8.818e-02  -5.740 1.08e-08 ***
## year2006                   -5.528e-01  8.851e-02  -6.246 5.03e-10 ***
## year2007                   -4.205e-01  8.835e-02  -4.760 2.06e-06 ***
## year2008                   -3.402e-01  8.680e-02  -3.920 9.14e-05 ***
## year2009                   -4.802e-01  8.729e-02  -5.501 4.22e-08 ***
## year2010                   -5.352e-01  8.798e-02  -6.083 1.39e-09 ***
## year2011                   -3.583e-01  8.719e-02  -4.110 4.10e-05 ***
## year2012                   -3.208e-01  8.589e-02  -3.735 0.000192 ***
## year2013                   -3.133e-01  8.591e-02  -3.647 0.000271 ***
## year2014                   -2.052e-01  8.550e-02  -2.400 0.016466 *
## year2015                   -6.439e-02  8.463e-02  -0.761 0.446778
## year2016                   -1.324e-01  8.464e-02  -1.564 0.117971
## year2017                    4.438e-02  8.419e-02   0.527 0.598160
## year2018                   -1.432e-01  8.430e-02  -1.699 0.089513 .
## year2019                   -1.284e-01  8.368e-02  -1.535 0.125036
## year2020                   -2.743e-01  8.721e-02  -3.146 0.001679 **
## year2021                   -2.211e-01  8.871e-02  -2.492 0.012777 *
## year2022                           NA         NA      NA       NA
## artist_popularity           3.170e-01  1.473e-02  21.527  < 2e-16 ***
## danceability                3.872e-03  1.576e-02   0.246 0.805901
## energy                     -4.660e-02  2.036e-02  -2.288 0.022207 *
## key                        -1.253e-02  1.233e-02  -1.017 0.309488
## loudness                    2.722e-02  1.754e-02   1.552 0.120913
## mode                        1.137e-02  1.254e-02   0.907 0.364620
## speechiness                -3.557e-02  1.412e-02  -2.519 0.011848 *
## acousticness                4.001e-02  1.510e-02   2.649 0.008130 **
## instrumentalness           -8.508e-05  1.237e-02  -0.007 0.994514
## liveness                   -3.359e-03  1.233e-02  -0.272 0.785388
## valence                     3.676e-02  1.576e-02   2.332 0.019774 *
## tempo                       3.858e-03  1.269e-02   0.304 0.761071
## duration_ms                -3.181e-02  1.466e-02  -2.169 0.030168 *
## time_signature             -1.509e-02  1.225e-02  -1.232 0.217958
## pop_genreTRUE              -7.009e-02  3.625e-02  -1.933 0.053328 .
## dance_pop_genreTRUE        -7.745e-02  3.371e-02  -2.297 0.021693 *
## rap_genreTRUE               1.428e-01  4.999e-02   2.856 0.004325 **
## pop_rap_genreTRUE          -4.365e-02  5.425e-02  -0.805 0.421149
```

```
## hip_hop_genreTRUE            -8.711e-02  4.641e-02  -1.877 0.060674 .
## rnb_genreTRUE                -6.981e-02  4.852e-02  -1.439 0.150355
## urban_contemporary_genreTRUE -6.520e-02  5.592e-02  -1.166 0.243787
## trap_genreTRUE               -7.353e-02  5.526e-02  -1.331 0.183474
## southern_hip_hop_genreTRUE    1.250e-01  6.803e-02   1.838 0.066220 .
## modern_rock_genreTRUE         3.720e-01  5.625e-02   6.613 4.72e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5714 on 2194 degrees of freedom
## Multiple R-squared:  0.4025, Adjusted R-squared:   0.39
## F-statistic: 32.13 on 46 and 2194 DF,  p-value: < 2.2e-16
```

```
# Plot the residuals of the new model to check if the issue has been resolved
par(mfrow = c(2,2))
plot(model_filtered)
```



Cross Validation

```
cv_model <- glm(track_popularity ~ year2000 + year2001 + year2002 + year2003 +
                      year2004 + year2005 + year2006 + year2007 + year2008 + year2009 +
                      year2010 + year2011 + year2012 + year2013 + year2014 + year2015 +
                      year2016 + year2017 + year2018 + year2019 + year2020 + year2021 +
```

```
                    year2022 + artist_popularity + danceability + energy + key +
                    loudness + mode + speechiness + acousticness + instrumentalness +
                    liveness + valence + tempo + duration_ms + time_signature +
                    num_genres + pop_genreTRUE + dance_pop_genreTRUE + rap_genreTRUE +
                    pop_rap_genreTRUE + hip_hop_genreTRUE + rnb_genreTRUE +
                    urban_contemporary_genreTRUE + trap_genreTRUE +
                    southern_hip_hop_genreTRUE + modern_rock_genreTRUE,
                    data = Playlist_filtered)

set.seed(123) # for reproducibility
K <- 10 # number of folds
cv_results <- cv.glm(Playlist_filtered, cv_model, K = K)

print(cv_results)
```

```
## $call
## cv.glm(data = Playlist_filtered, glmfit = cv_model, K = K)
##
## $K
## [1] 10
##
## $delta
## [1] 0.3327576 0.3320540
##
## $seed
##   [1]       10403         624  -983674937   643431772  1162448557  -959247990
##   [7]  -133913213  2107846888   370274761 -2022780170  -412390145   848182068
##  [13]  -266662747 -1309507294  1356997179  1661823040  1749531457  -516669426
##  [19]  1042678071 -1279933428  -410084963  1151007674  -895613453  1288379032
##  [25]  -376044615 -1358274522   307686511   101447652  1796216213 -1567696558
##  [31]  1186934955 -1925339152  -472470735    80319294 -1524429145   326645436
##  [37]  -389586803  -400786966  -890731933  -852332472  1365217705 -1785317034
##  [43] -1551153185  1359863956  2098748037 -1013039742  -329721061 -1587358816
##  [49]   344102689 -1520389522   166492183  1821136236  1646453629  1056605210
##  [55] -1419044141  -806080008   520985497   711286406  2004844367 -1445006012
##  [61]  1329781621 -1188844110 -1089068661  1173875536 -1983217903   514629022
##  [67]  -237421177  -258138084  -930078099   261626442  1349308227 -1125425240
##  [73] -1677778551    25874358   409637567 -1987430924  1583257701  -136173086
##  [79]   639501307   272101120 -1024630015 -1994369842  -939499785 -1944742196
##  [85]  -591520419 -1994900358  1072996275  1119025496  2035491705 -2082894618
##  [91]   776176175   -69557596  1794806101  -178474478  -497581461   874372784
##  [97]   518669041  -370223106  1295572071 -1776240260 -1692674995  1935534762
## [103]   298421283   111542024 -1075273367   518297110  -289321569  1331379028
## [109]  1768277573  1473660482  2120850651   879016544  -864018719  1661675310
## [115]   135902679 -2136373204   735594301  1594631386  -546138989  1423929528
## [121] -1067541671  1962863430 -1923418865  -984154108  1907308341   642901618
## [127] -1095019701 -1836613104 -1171392815  1663582814 -1258689721 -2007301412
## [133]  -756910547  -712643830 -1271482109  -801485208    51646793 -1925477258
## [139] -1421379457  1104736436 -1348082651  -124611934   292791739  2126591424
## [145] -2043491647  -709285490 -1242530633  1688217996  -538353379 -1997652678
## [151]   -48432781   575772696   942146361    57506214  -948054033   -72610460
## [157]  1389939989   656100050   -25586645 -2012424848  1854773937  1391516862
## [163] -2100008409  -140248004 -1638135795 -2077746326  -118729245 -1417654840
```

```
## [169]   662270249   942125782 -1363864737   744183316  2123821573   -80802046
## [175] -1753997669  1277518112  1090348705  1338137582   423408535   -28214548
## [181]  1164536573  1524008346   673959507   853634936 -1599644903 -2135083002
## [187]  -345756977 -1070478652   971985653  -556736718  -406174453   663083216
## [193]  1258368657  1306568478  1820350727 -1068259940  -402617875  1499233226
## [199] -1121819965 -1773142744  1796260105  1463879990   901914175   104491892
## [205]  1605431269 -1933329566  1688405883  -446088064  1238889089   197049934
## [211]  -709469577 -1072333748  1691378909 -1260585478   198644531  2053568216
## [217]   903127801 -1970919834  -473567825  1614821412 -1905604395  1082827666
## [223]  1558537707  1875545136  1518383729 -1265655426 -2085242905  1791098620
## [229]  1447558093 -1153758166   -99557469   -92185464 -2016280343  1847562134
## [235]  1495701791  -221368108   409722309  -429353022  1765302363  2137311200
## [241]  -373658015   273043630  -350916265  -935055956    43404989    52012634
## [247]  1867958291  1488596536 -1347953959   174081222  2002460815  1429165444
## [253]  -205312331  1264621554  -603785525  1270017936 -1543231919 -1282028578
## [259]   908887751   726075484  1269456301 -1680094070  -990917501 -1377014808
## [265] -1279971127  1281050102   228230143  1097770548 -1438663771  1295361058
## [271]   829172027   988808000  1704778305   804328206 -1257113545  -516583668
## [277] -1624037219  1034190522   904064243 -1716316776  1108935353   904106790
## [283]  1222361967  1146561252  1232110741   174767186  2136668075 -1843985680
## [289]   713263665  1133192766  1302119847  -499465796  -425742451  2035727594
## [295]  1324820835  -227988664 -1598926679   227290198   601218783  1836305300
## [301]  1386514821   306372738  -445226469   618852000   -25741791   156697966
## [307]  -345772265 -2126405524  1998516861  -392853734  1588822483  1965665528
## [313] -1658840423 -1901588090  -687876529   -15753148 -1427453323 -1799286606
## [319]   -47880053    97437264  -319365615   688369822  -272731001   469052188
## [325]    27259245  1573117258  -446761405  1976539816  2093047945   424297142
## [331]  1217440191   506831092 -1961736347 -1834464030  1234111227   907381248
## [337]  -247365119   118499278 -1581033993  -893361716 -2100188067   335855482
## [343]    83920563 -1896483752  -323673479  -498745370  2088720687 -2102342236
## [349]  1873412181   226202898 -1483060885  1437743536  -430562831  -190616834
## [355] -1639345305   281953404   857940813  -549769814  -245419229  1375189512
## [361]  -237346711   590186774    75687071   655107668   151057733   930998594
## [367] -1108466725  1398789472  1995685345  1605663278  1206398167 -1945513172
## [373]  1992513085  1544169434  1610742675  -152048712  -657450407  1247059526
## [379]  1880247311  -124605692   723920437 -1548596878  1827773003   479812880
## [385]   228152785    49698142   922100295 -1524757028  -845069011   534031882
## [391]  -131080189   213485928   636833865   718143350 -1134260353 -2024842316
## [397] -1108831451  1977333154  1053535419  1301926080  -997856831   366738574
## [403] -1450544201  1064694924 -1016336355  -390217670 -1024466829   686789400
## [409] -2056715719   745319590  -999248145 -1240647580 -1395180523 -1837290030
## [415]  -681354453  -514051984  1438153137  2090364862  -209968857  1765574460
## [421]  -544057587  -844603798 -1693909789 -1746073400 -1156960215  2076419542
## [427] -1326601633  1784103188  -683597563  -824593918  1683989915  -509903840
## [433]   183502241  -132206866  -295556457   190629356 -1790739971  1849133210
## [439] -1660799661   214755960 -1837639143   975563526  1750237647  1014527428
## [445]     3490293   552878642   220695563   382907344 -1381266031  1445050910
## [451]  1771278343 -1719553892   862869741   583941834 -1759344189  1365915688
## [457]  -820969463 -1381598154   -19516097   662427252 -1098735899  -812655006
## [463]  1658982011 -1203972224  1999245697 -1592487602 -1708699273 -1038727348
## [469]  -725486627   747602170  2037447219  -161484328   469017081  1897421158
## [475]   644859055   959210276  1824012245 -1573943662  -797561621   466937648
## [481]     6984049  1344943230 -1963692313   507873788  1336756941  -446804182
## [487]  -978024797    50927496   -66994199 -1542552938 -1630130145  1108679636
```

```
## [493]   421858501   286669506   176875355  1716904672   841747809  2002101166
## [499] -1936594857  -503678804   643784125  -270685862    -9162989 -1518294728
## [505] -1177069095   450623430 -1518307441 -2055143292  1977097653  1967586034
## [511]  2139569611   993708688   887981393  -146153762 -1521041977 -1948249252
## [517]  1992764589  1735430026   469169027  -492722456  1473540041 -1902921482
## [523]  1705351935  1769673012  -929011035   948225826  -946720709  1824431680
## [529]  1626208577 -1384520178    22671159 -1788782068  -359417955   272236986
## [535]  -230435853  1174868120 -2145910343  -855063002  1748802159   651054564
## [541]  -619908203    89300818   345161387 -1411621392   774662449 -1541883586
## [547]  1651670183   581520572 -1489764723 -2028142614 -1423847325 -1844713912
## [553]  1954615209  -389144746    66876895  2030417556  -361973627  -151813246
## [559] -1573918437   944703904   610784545  1108957294 -1875417577 -1297945748
## [565]  1037500797  1908181530   823650515  1875585016   -22111847  1765196934
## [571]  -849597105  1315720004 -1748059787  -915770446   634433419 -1869504176
## [577]  -887145199  2066662302  -939545721  -822528484 -1687437203 -1367629750
## [583] -1603461821   522180008  1610588041  2052437430   110280895  2014120948
## [589]  -670960027   159018978  1050415611   568272128 -1718509311    -3409202
## [595]   753028343 -1139331892  -123651235 -2072165766 -1222087245   648343384
## [601]  1100161401   486404838   261566511  1504901284  -476745899  1151760402
## [607]  -445050773  -130902864  -423755535  1831075326   934693479   690474876
## [613]  -907644339  -744197974  1158732323    62223624 -1538777239  1455586326
## [619]  -702514273 -1712778924   651699269   959548482  -586241317  1850142816
## [625]  -647799583  2099891502
```

```r
# Evaluating the model (RMSE, MAE)
data.frame(RMSE = sqrt(mean(residuals(model_filtered)^2)),
           MAE = mean(abs(residuals(model_filtered))))
```

```
##        RMSE       MAE
## 1 0.5653916 0.4510463
```

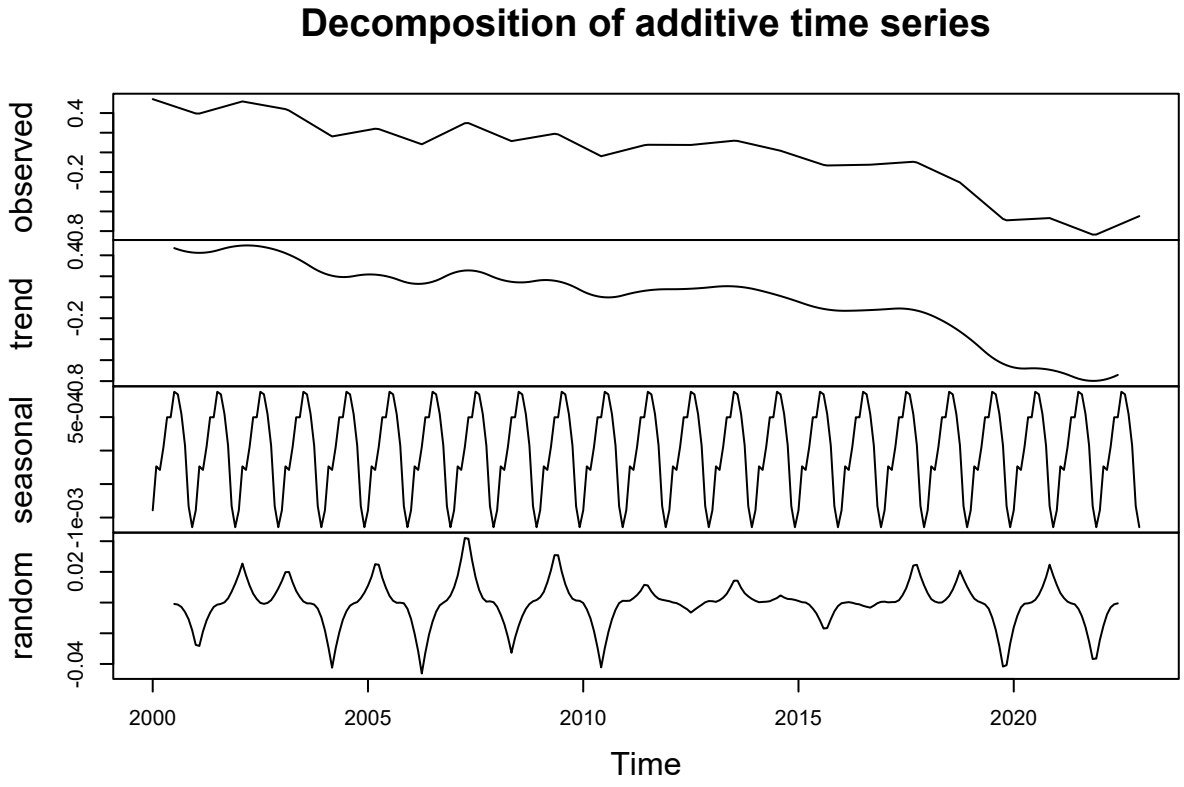
Forecasting Using Time Series Analysis

```r
# Time Series Analysis - (decreasing trend in the average duration of songs over the years)
# Melting the data to get year column
Playlist_filtered_long <- Playlist_filtered %>%
  pivot_longer(cols = starts_with("year"), names_to = "year", values_to = "value") %>%
  filter(value == 1)

Playlist_filtered_long$year <- substr(Playlist_filtered_long$year, 5, 8)

avg_duration_per_year <- Playlist_filtered_long %>%
  group_by(year) %>%
  summarize(avg_duration = mean(duration_ms, na.rm = TRUE))

# Converting to time series object
time_series_data <- ts(avg_duration_per_year$avg_duration, start = min(avg_duration_per_year$year), end
print(time_series_data)
```

```
## Time Series:
```

```
## Start = 2000
## End = 2022
## Frequency = 1
##  [1]  0.54145302  0.38972378  0.51814764  0.43655463  0.16214764  0.24376270
##  [7]  0.08235851  0.30566405  0.11484839  0.19233068 -0.03884849  0.07757721
## [13]  0.07559505  0.12057067  0.01667075 -0.13327803 -0.12512016 -0.09405773
## [19] -0.30588514 -0.69142156 -0.66758532 -0.84261030 -0.64770321
```

```r
# Create a date sequence
date_seq <- seq.Date(from = as.Date("2000-01-01"), to = as.Date("2022-12-01"), by = "month")
year_seq_numeric <- seq(2000, 2022, length.out = length(date_seq))

# Interpolate the monthly data points
# Assuming avg_duration_per_year is a data frame with columns 'year' and 'avg_duration'
year_seq <- seq(2000, 2022)
avg_duration_seq <- avg_duration_per_year$avg_duration

monthly_avg_duration <- approx(year_seq, avg_duration_seq, xout = year_seq_numeric, method = "linear")$

# Create a time series object
time_series_data <- ts(monthly_avg_duration, start = c(2000, 1), frequency = 12)

# Decompose the time series
decomposed_data <- decompose(time_series_data)

# Plot the decomposed data
plot(decomposed_data)
```



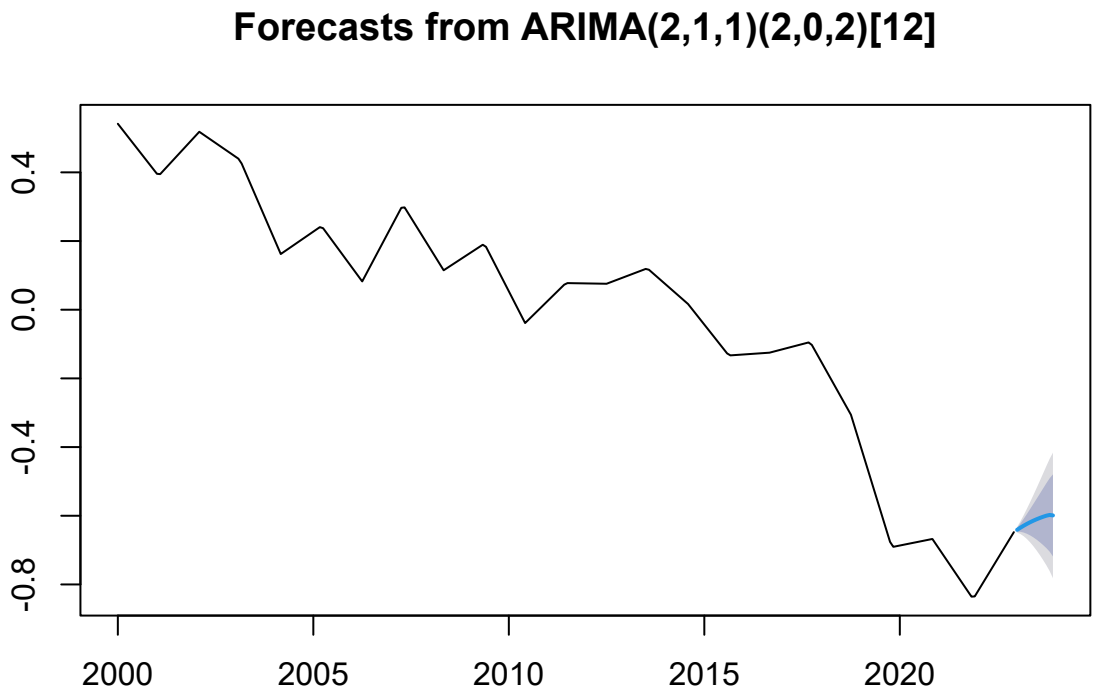

```
# Check if the time series is stationary
# Conduct the Augmented Dickey-Fuller Test
adf_test <- adf.test(time_series_data, alternative = "stationary")
print(adf_test)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  time_series_data
## Dickey-Fuller = -2.4516, Lag order = 6, p-value = 0.3858
## alternative hypothesis: stationary
```

The output of the Augmented Dickey-Fuller test indicates that the time series is stationary. The p-value is 0.01, which is less than the common significance level of 0.05, allowing us to reject the null hypothesis that the time series has a unit root (i.e., is non-stationary).

```
# Differencing the time series
diff_time_series_data <- diff(time_series_data)

# Use auto.arima to automatically select the best ARIMA model
best_model <- auto.arima(time_series_data)

# Forecast the next 12 months
forecast_result <- forecast(best_model, h = 12)
```

```r
# Plot the forecast
plot(forecast_result)
```



Conclusion

The hypothesis was formulated as shown below:

H1: Songs with higher danceability are more popular. H2: Over the years, the average duration of popular songs has decreased. H3: Artist popularity is a significant predictor of track popularity.

Songs with a higher danceability are not popular. Also, based on the regression models, while there is a positive association between danceability and track popularity, it is not statistically significant. Therefore, it cannot be confidently stated that songs with higher danceability are more popular based on this specific dataset and model. The average duration of popular songs has decreased over the years. While there is some fluctuation in the middle years, the overall trend from 2000 to 2022 appears to be decreasing. Given this downward trend, the time series data does seem to support the hypothesis that the average duration (or possibly the popularity) of songs has decreased over the years. Artist popularity is a significant predictor of track popularity. There is a positive correlation between artist popularity and genre popularity. Also, The p-value for duration_ms is less than the 0.05 significance level. Therefore, the relationship between song duration and track popularity is statistically significant.

The forecasted average duration of popular songs for the next 12 months is seen to be on a growing trend.