Regression Analysis on a Spotify Dataset: Predicting Track Popularity

Liam Daly, Tristan Dull, Daniel Khan, Aaron Bajorunas, Christopher Chen

2023-12-14

Contents

List of Figures

List of Tables

Introduction

For our group project project we will be analyzing a real-world Spotify dataset with over 80k observations of 21 different variables. Our main goal is to attempt to create linear regression models in order to make inferences on the following two response variables: track popularity and artist popularity.

Research Question

Which predictors are most significant for predicting track popularity or artist popularity.

Analysis Goal

Our main stakeholders in this scenario would be music industry executives interested to see which factors may influence popularity. Given the insights generated by our analysis, they may be able to allocate resources toward improving certain factors for the sake of generating more popularity.

Data Description

Variables and Summary Statistics

Exploratory Data Analysis

Data Cleaning

```
#change to your computer
spotify <- read_csv("/Users/liamdaly/Downloads/playlist_2010to2023.csv")
```

```
spotify
```

```
## # A tibble: 2,400 x 23
##    playlist_url        year track_id track_name track_popularity album artist_id
##    <chr>              <dbl> <chr>    <chr>                 <dbl> <chr> <chr>
##  1 https://open.spot~  2000 6naxalm~ Oops!...I~               81 Oops~ 26dSoYcl~
##  2 https://open.spot~  2000 2m1hi0n~ All The S~               83 Enem~ 6FBDaR13~
##  3 https://open.spot~  2000 3y4LxiY~ Breathe                  66 Brea~ 25NQNriV~
##  4 https://open.spot~  2000 0v1XpBH~ It's My L~               81 Crush 58lV9VcR~
##  5 https://open.spot~  2000 62bOmKY~ Bye Bye B~               75 No S~ 6Ff53Kvc~
##  6 https://open.spot~  2000 5Mmk2ii~ Thong Song               71 Unle~ 6x9QLdzo~
##  7 https://open.spot~  2000 3yfqSUW~ The Real ~               87 The ~ 7dGJo4pc~
##  8 https://open.spot~  2000 7oQSevU~ Rock DJ                  56 Sing~ 2HcwFjNe~
##  9 https://open.spot~  2000 7H6ev70~ Say My Na~               81 The ~ 1Y8cdNmU~
## 10 https://open.spot~  2000 3AJwUDP~ Yellow                   90 Para~ 4gzpq5DP~
## # i 2,390 more rows
## # i 16 more variables: artist_name <chr>, artist_genres <chr>,
## #   artist_popularity <dbl>, danceability <dbl>, energy <dbl>, key <dbl>,
## #   loudness <dbl>, mode <dbl>, speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>, time_signature <dbl>
```

```
nrow(spotify)
```

```
## [1] 2400
```

Upon looking through the entire dataset, it's clear that it isn't fit for analysis straight away. There are many NA entries, so I'll first remove all rows containing NA values.

```r
library(dplyr)

hostel <- spotify %>%
  filter(complete.cases(spotify))
```

```r
spotify
```

```
## # A tibble: 2,400 x 23
##    playlist_url        year track_id track_name track_popularity album artist_id
##    <chr>              <dbl> <chr>    <chr>                 <dbl> <chr> <chr>
##  1 https://open.spot~  2000 6naxalm~ Oops!...I~               81 Oops~ 26dSoYcl~
##  2 https://open.spot~  2000 2m1hi0n~ All The S~               83 Enem~ 6FBDaR13~
##  3 https://open.spot~  2000 3y4LxiY~ Breathe                  66 Brea~ 25NQNriV~
##  4 https://open.spot~  2000 0v1XpBH~ It's My L~               81 Crush 58lV9VcR~
##  5 https://open.spot~  2000 62bOmKY~ Bye Bye B~               75 No S~ 6Ff53Kvc~
##  6 https://open.spot~  2000 5Mmk2ii~ Thong Song               71 Unle~ 6x9QLdzo~
##  7 https://open.spot~  2000 3yfqSUW~ The Real ~               87 The ~ 7dGJo4pc~
##  8 https://open.spot~  2000 7oQSevU~ Rock DJ                  56 Sing~ 2HcwFjNe~
##  9 https://open.spot~  2000 7H6ev70~ Say My Na~               81 The ~ 1Y8cdNmU~
## 10 https://open.spot~  2000 3AJwUDP~ Yellow                   90 Para~ 4gzpq5DP~
## # i 2,390 more rows
## # i 16 more variables: artist_name <chr>, artist_genres <chr>,
## #   artist_popularity <dbl>, danceability <dbl>, energy <dbl>, key <dbl>,
## #   loudness <dbl>, mode <dbl>, speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>, time_signature <dbl>
```

Now, we'll need to remove any duplicate tracks.

```r
# removing duplicate tracks
spotify <- spotify %>% distinct(track_id, .keep_all = TRUE)
```

```r
nrow(spotify)
```

```
## [1] 2302
```

There was a total of 98 duplicates.

```r
spotify
```

```
## # A tibble: 2,302 x 23
##    playlist_url        year track_id track_name track_popularity album artist_id
##    <chr>              <dbl> <chr>    <chr>                 <dbl> <chr> <chr>
##  1 https://open.spot~  2000 6naxalm~ Oops!...I~               81 Oops~ 26dSoYcl~
##  2 https://open.spot~  2000 2m1hi0n~ All The S~               83 Enem~ 6FBDaR13~
##  3 https://open.spot~  2000 3y4LxiY~ Breathe                  66 Brea~ 25NQNriV~
##  4 https://open.spot~  2000 0v1XpBH~ It's My L~               81 Crush 58lV9VcR~
##  5 https://open.spot~  2000 62bOmKY~ Bye Bye B~               75 No S~ 6Ff53Kvc~
##  6 https://open.spot~  2000 5Mmk2ii~ Thong Song               71 Unle~ 6x9QLdzo~
##  7 https://open.spot~  2000 3yfqSUW~ The Real ~               87 The ~ 7dGJo4pc~
##  8 https://open.spot~  2000 7oQSevU~ Rock DJ                  56 Sing~ 2HcwFjNe~
##  9 https://open.spot~  2000 7H6ev70~ Say My Na~               81 The ~ 1Y8cdNmU~
## 10 https://open.spot~  2000 3AJwUDP~ Yellow                   90 Para~ 4gzpq5DP~
## # i 2,292 more rows
## # i 16 more variables: artist_name <chr>, artist_genres <chr>,
## #   artist_popularity <dbl>, danceability <dbl>, energy <dbl>, key <dbl>,
## #   loudness <dbl>, mode <dbl>, speechiness <dbl>, acousticness <dbl>,
```

```
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>, time_signature <dbl>
```

We also need to remove irrelevant categorical variables. We found these variables had far too many categories and were unfit to include in our analysis moving forward.

```
spotify <- spotify |>
  dplyr::select(-playlist_url, -track_id, -track_name, -album, -artist_id, -artist_name, -artist_genres)
```

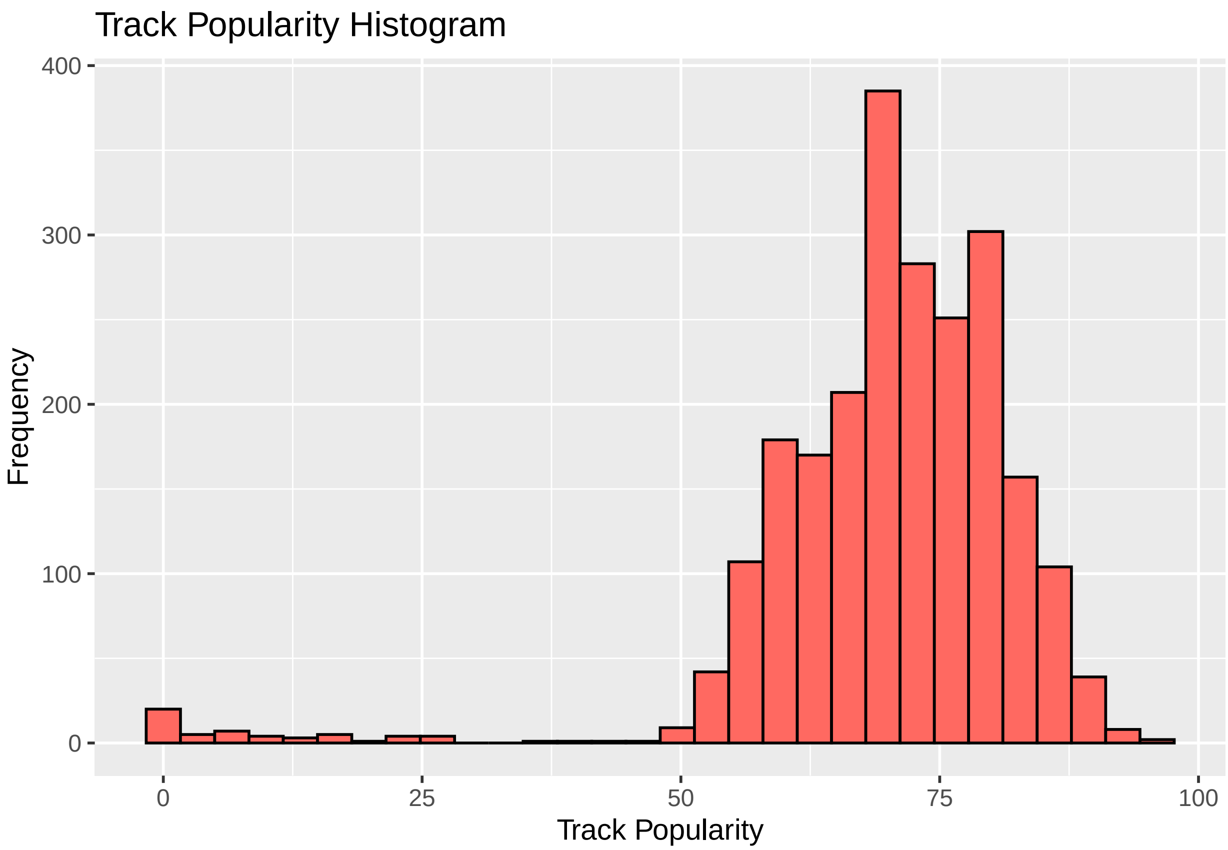
```
spotify
```

```
## # A tibble: 2,302 x 16
##     year track_popularity artist_popularity danceability energy   key loudness
##    <dbl>            <dbl>             <dbl>        <dbl>  <dbl> <dbl>    <dbl>
##  1  2000               81                81        0.751  0.834     1    -5.44
##  2  2000               83                79        0.434  0.897     0    -4.92
##  3  2000               66                62        0.529  0.496     7    -9.01
##  4  2000               81                79        0.551  0.913     0    -4.06
##  5  2000               75                70        0.61   0.926     8    -4.84
##  6  2000               71                58        0.706  0.888     2    -6.96
##  7  2000               87                90        0.949  0.661     5    -4.24
##  8  2000               56                71        0.712  0.762     7    -4.31
##  9  2000               81                72        0.713  0.678     5    -3.52
## 10  2000               90                88        0.429  0.661    11    -7.23
## # i 2,292 more rows
## # i 9 more variables: mode <dbl>, speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>, time_signature <dbl>
```

Visualizations

Response Variable Plot

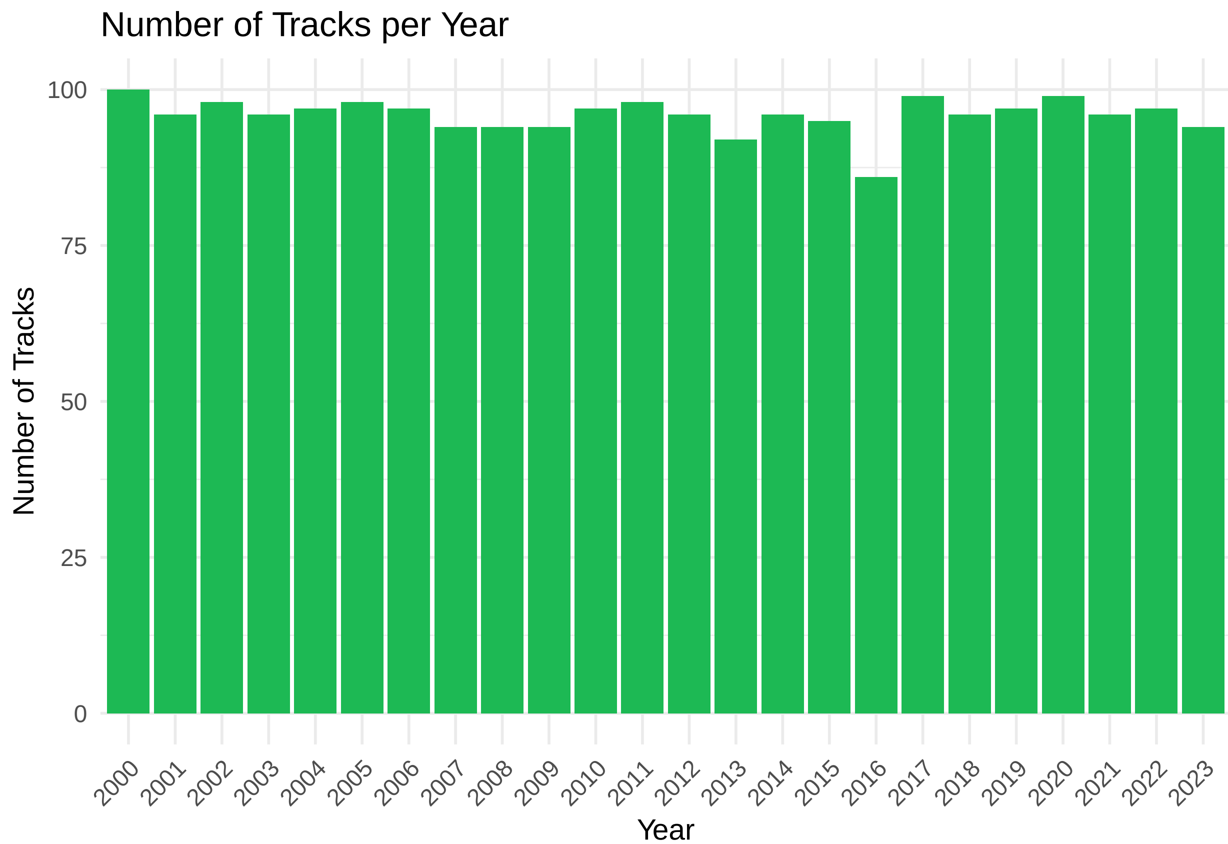
```
ggplot(data = spotify, aes(x = track_popularity)) +
  geom_histogram(fill = "#FF6961", color = 1) +
  labs(title = "Track Popularity Histogram", x = "Track Popularity", y = "Frequency")
```



Predictor Plots

Number of tracks per year:

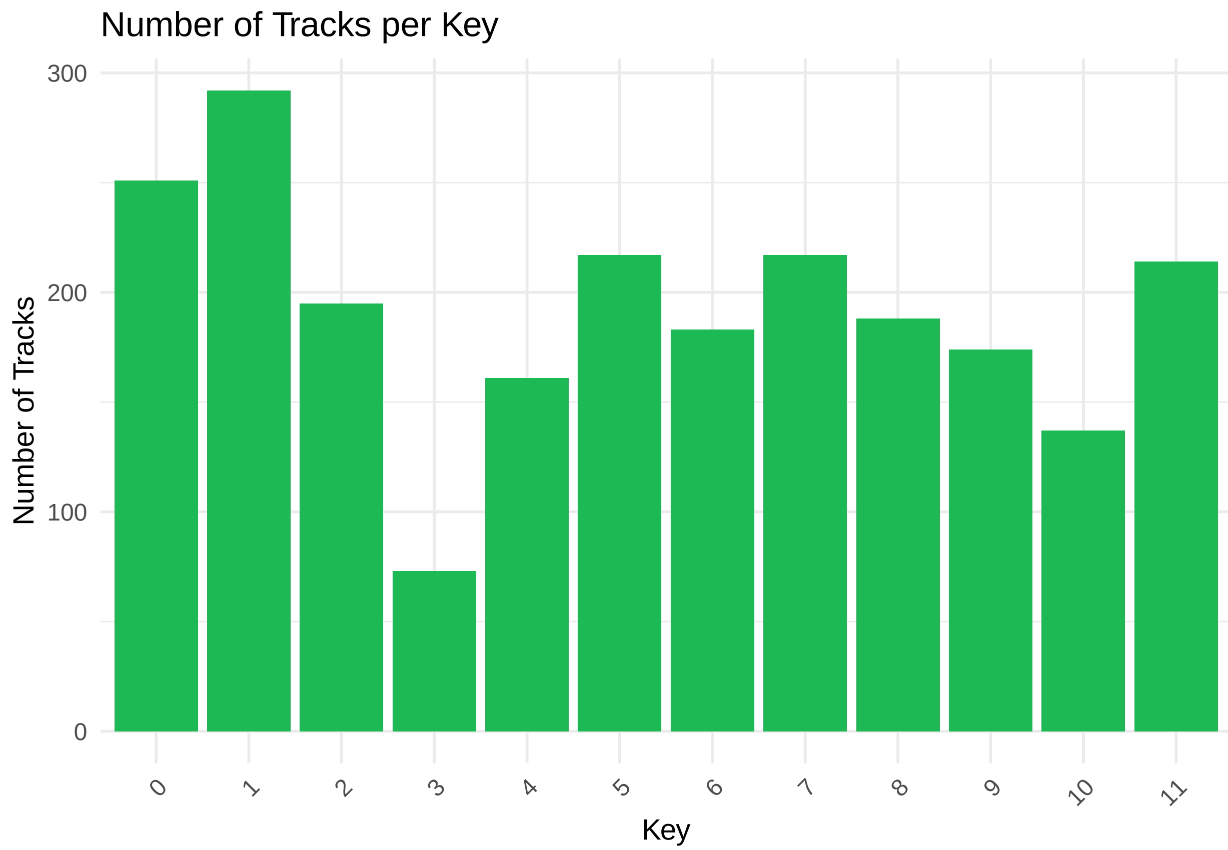
```
year_counts <- spotify |>
  group_by(year) |>
  summarise(count = n())

ggplot(year_counts, aes(x = factor(year), y = count)) +
  geom_bar(stat = "identity", fill = "#1DB954") +
  theme_minimal() +
  labs(title = "Number of Tracks per Year",
       x = "Year",
       y = "Number of Tracks") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Relatively even number of tracks per each year.

Number of tracks per key:

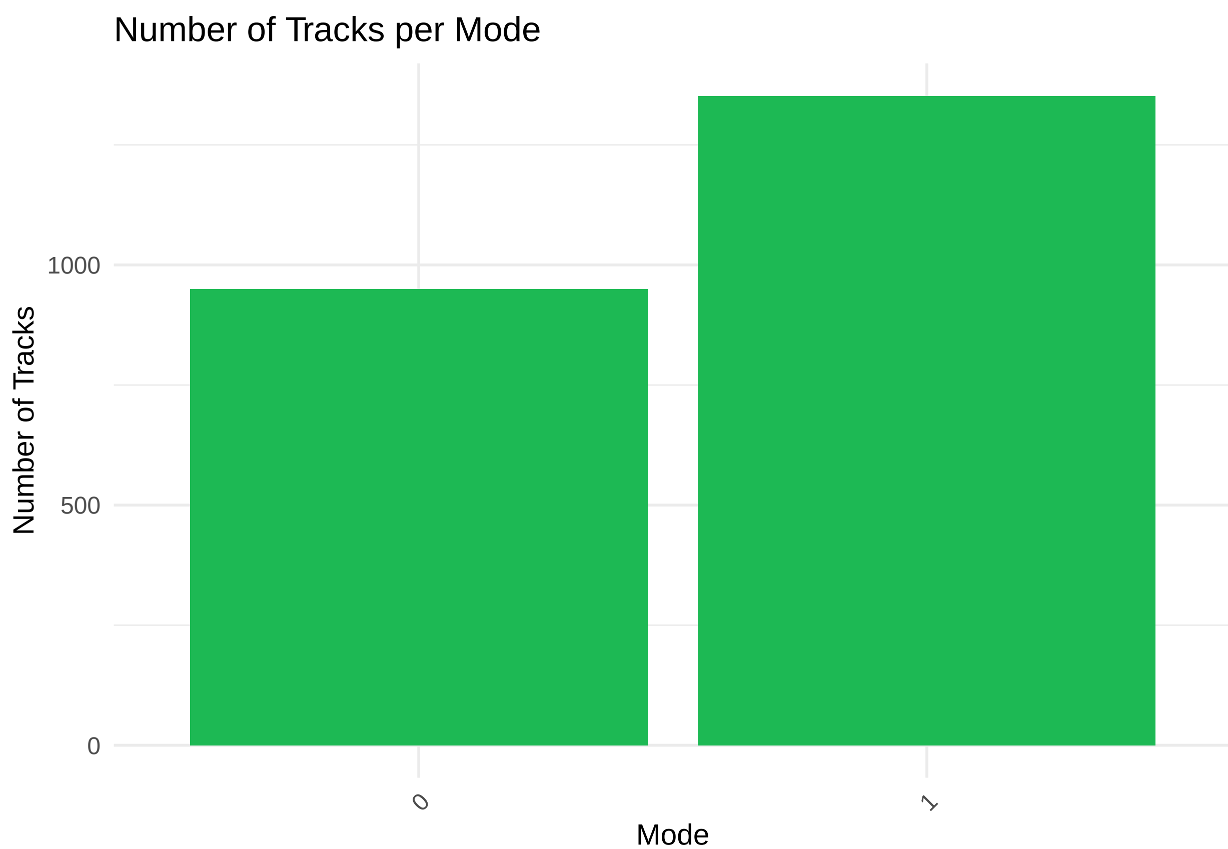
```
key_counts <- spotify |>
  group_by(key) |>
  summarise(count = n())

ggplot(key_counts, aes(x = factor(key), y = count)) +
  geom_bar(stat = "identity", fill = "#1DB954") +
  theme_minimal() +
  labs(title = "Number of Tracks per Key",
       x = "Key",
       y = "Number of Tracks") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



There appears to be a mixed amount of representation from each key.

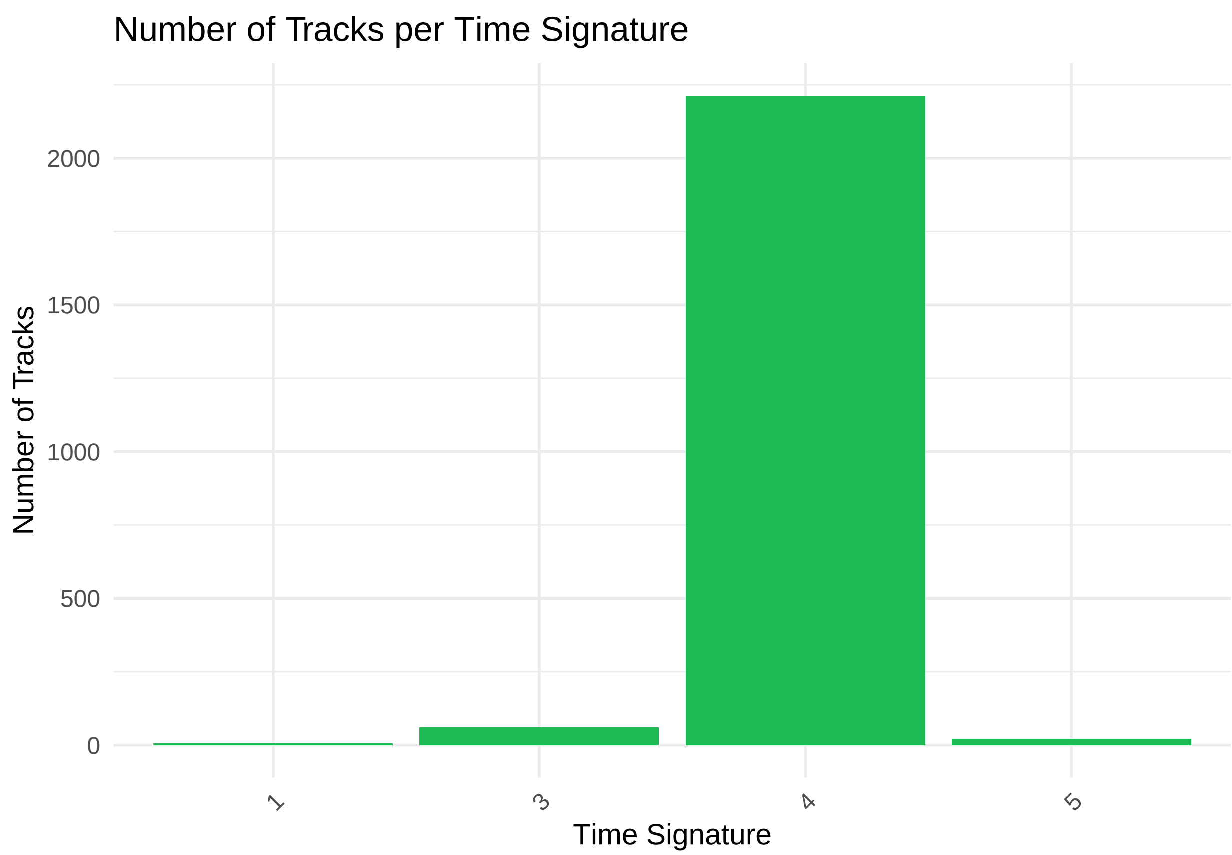
Number of tracks per mode:

```
mode_counts <- spotify |>
  group_by(mode) |>
  summarise(count = n())

ggplot(mode_counts, aes(x = factor(mode), y = count)) +
  geom_bar(stat = "identity", fill = "#1DB954") +
  theme_minimal() +
  labs(title = "Number of Tracks per Mode",
       x = "Mode",
       y = "Number of Tracks") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



More reprentation from tracks of mode = 1.

```
time_signature_counts <- spotify |>
  group_by(time_signature) |>
  summarise(count = n())

ggplot(time_signature_counts, aes(x = factor(time_signature), y = count)) +
  geom_bar(stat = "identity", fill = "#1DB954") +
  theme_minimal() +
  labs(title = "Number of Tracks per Time Signature",
       x = "Time Signature",
       y = "Number of Tracks") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

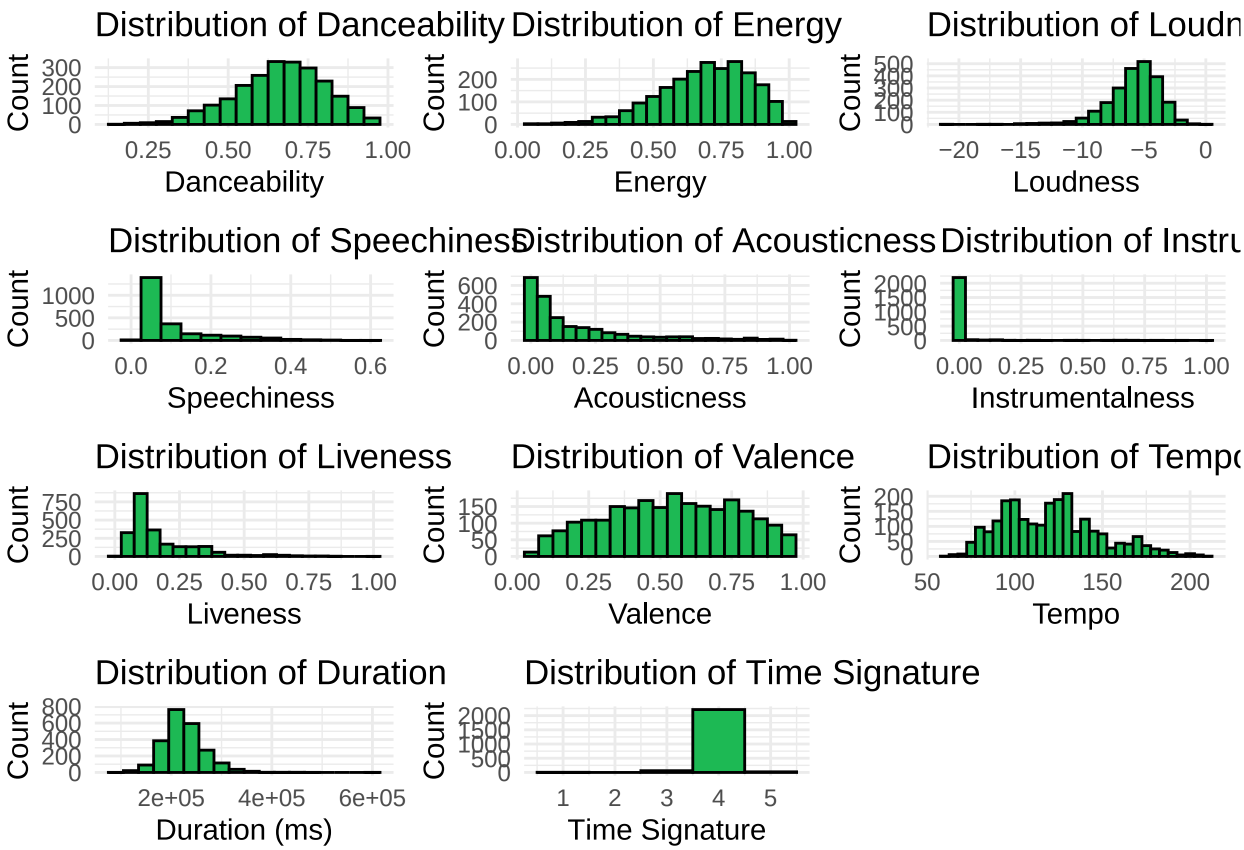


Vast majority of tracks are of time signature 4.

Visualizing all continuous variables in a grid format:

```
plot_danceability <- ggplot(spotify, aes(x = danceability)) +
  geom_histogram(binwidth = 0.05, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Danceability", x = "Danceability", y = "Count")

plot_energy <- ggplot(spotify, aes(x = energy)) +
  geom_histogram(binwidth = 0.05, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Energy", x = "Energy", y = "Count")

plot_loudness <- ggplot(spotify, aes(x = loudness)) +
  geom_histogram(binwidth = 1, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Loudness", x = "Loudness", y = "Count")

plot_speechiness <- ggplot(spotify, aes(x = speechiness)) +
  geom_histogram(binwidth = 0.05, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Speechiness", x = "Speechiness", y = "Count")

plot_acousticness <- ggplot(spotify, aes(x = acousticness)) +
  geom_histogram(binwidth = 0.05, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Acousticness", x = "Acousticness", y = "Count")
```

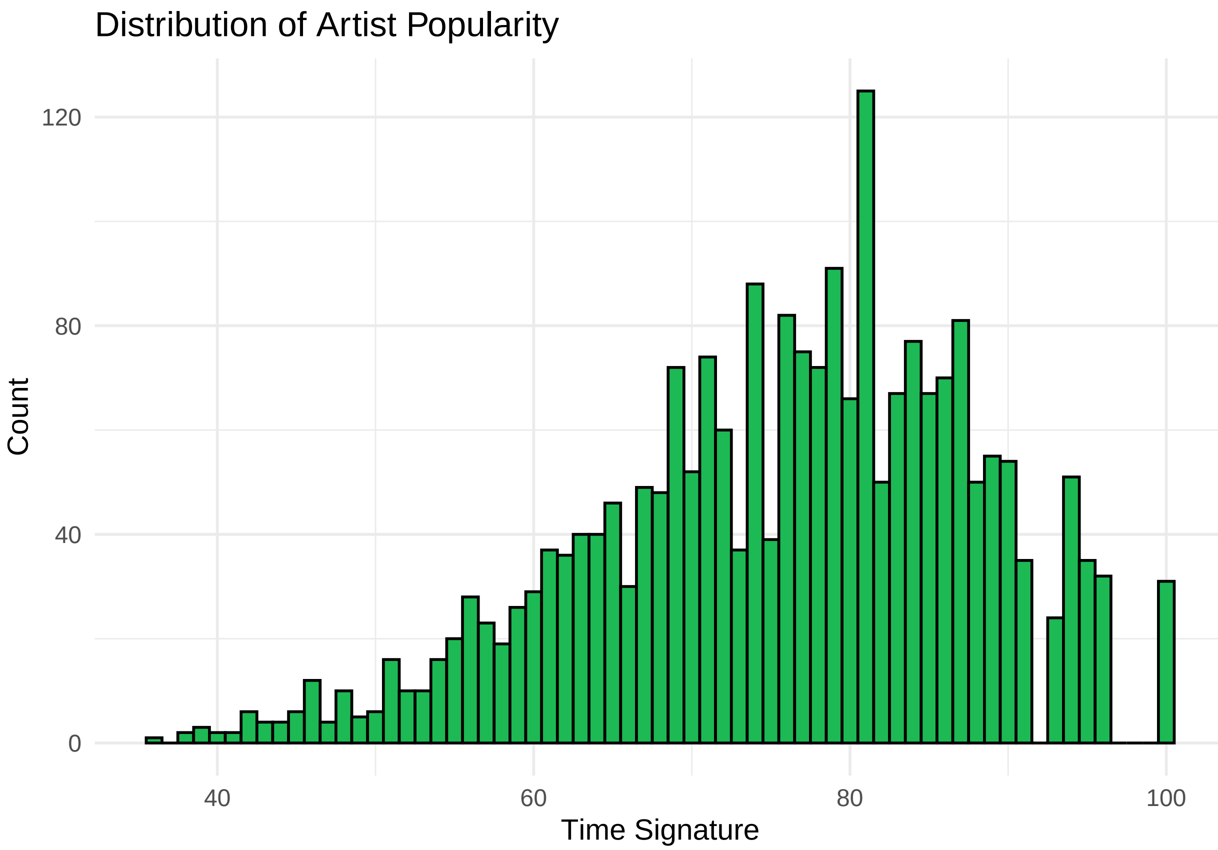
```r
plot_instrumentalness <- ggplot(spotify, aes(x = instrumentalness)) +
  geom_histogram(binwidth = 0.05, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Instrumentalness", x = "Instrumentalness", y = "Count")

plot_liveness <- ggplot(spotify, aes(x = liveness)) +
  geom_histogram(binwidth = 0.05, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Liveness", x = "Liveness", y = "Count")

plot_valence <- ggplot(spotify, aes(x = valence)) +
  geom_histogram(binwidth = 0.05, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Valence", x = "Valence", y = "Count")

plot_tempo <- ggplot(spotify, aes(x = tempo)) +
  geom_histogram(binwidth = 5, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Tempo", x = "Tempo", y = "Count")

plot_duration <- ggplot(spotify, aes(x = duration_ms)) +
  geom_histogram(binwidth = 30000, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Duration", x = "Duration (ms)", y = "Count")

plot_time_signature <- ggplot(spotify, aes(x = time_signature)) +
  geom_histogram(binwidth = 1, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Time Signature", x = "Time Signature", y = "Count")

grid.arrange(
  plot_danceability, plot_energy, plot_loudness,
  plot_speechiness, plot_acousticness, plot_instrumentalness,
  plot_liveness, plot_valence, plot_tempo,
  plot_duration, plot_time_signature,
  ncol = 3
)
```



This indicates some variables may be highly correlated to one another.

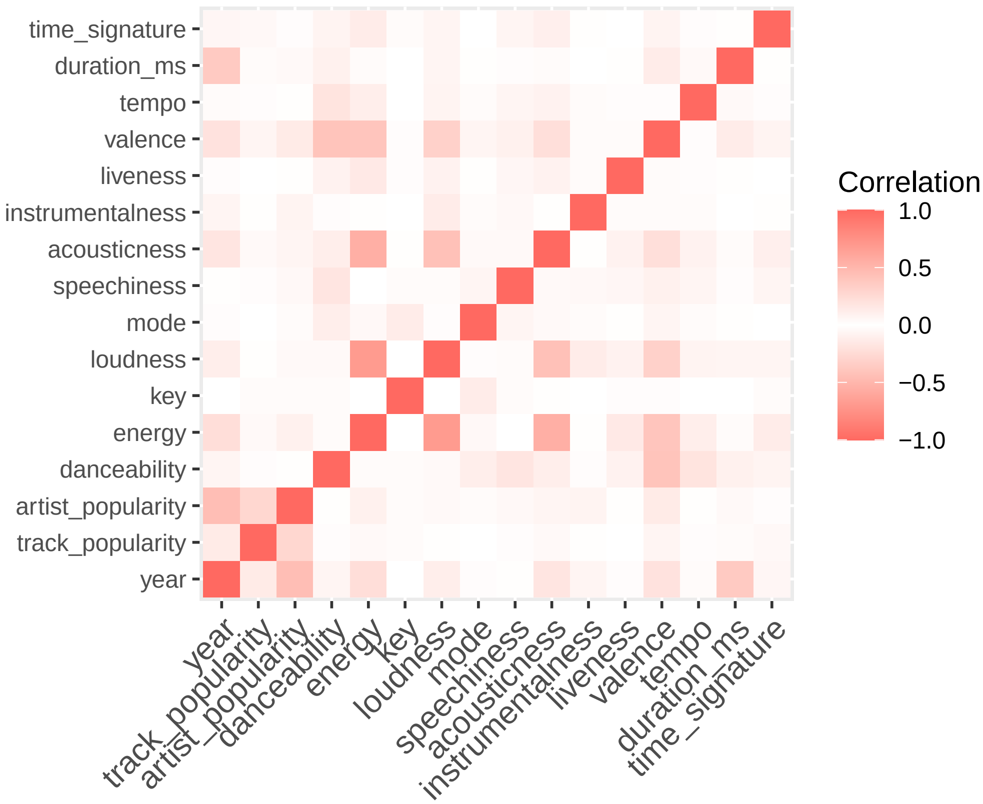
Number of tracks per level of artist popularity:

```
plot_time_signature <- ggplot(spotify, aes(x = artist_popularity)) +
  geom_histogram(binwidth = 1, fill = "#1DB954", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Artist Popularity", x = "Time Signature", y = "Count")

plot_time_signature
```



Correlation Visualization

```
placement1 <- read.csv("/Users/liamdaly/Downloads/Placement_Data_Full_Class.csv")
library(RColorBrewer)
library(ggplot2)
library(reshape2)
corr_mat <- round(cor(spotify),2)

melted_corr_mat <- melt(corr_mat)


# plotting the correlation heatmap
library(ggplot2)
ggplot(data = melted_corr_mat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white")   +
  scale_fill_gradientn(colors = c("#FF6961", "white", "#FF6961"),
                       limits = c(-1, 1),
                       name = "Correlation") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1)) +
  labs(x = "", y = "") +
  coord_fixed() +
  geom_tile()
```



Correlation plot doesn't indicate any serious problems. I'll check the specific correlation coefficient between the following pairs of variables:

acoustincess and energy loudness and energy

```
cor(spotify$acousticness, spotify$energy)
```

```
## [1] -0.5544004
```

```
cor(spotify$loudness, spotify$energy)
```

```
## [1] 0.6939809
```

These values are highly correlated but they're below the abs(.7) threshold so we'll keep them for now. If the vif scores of our full regression model are high, we'll consider removing more variables to account for potential multicollinearity.

Model Building

Full Model

```
spotify
```

```
## # A tibble: 2,302 x 16
##     year track_popularity artist_popularity danceability energy   key loudness
##    <dbl>            <dbl>             <dbl>        <dbl>  <dbl> <dbl>    <dbl>
##  1  2000               81                81        0.751  0.834     1    -5.44
##  2  2000               83                79        0.434  0.897     0    -4.92
##  3  2000               66                62        0.529  0.496     7    -9.01
##  4  2000               81                79        0.551  0.913     0    -4.06
```

```
##  5  2000                75                70        0.61   0.926     8    -4.84
##  6  2000                71                58        0.706  0.888     2    -6.96
##  7  2000                87                90        0.949  0.661     5    -4.24
##  8  2000                56                71        0.712  0.762     7    -4.31
##  9  2000                81                72        0.713  0.678     5    -3.52
## 10  2000                90                88        0.429  0.661    11    -7.23
## # i 2,292 more rows
## # i 9 more variables: mode <dbl>, speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>, time_signature <dbl>
```

```r
m1 <- lm(track_popularity ~ year + artist_popularity + danceability + energy + key +
           loudness + mode + speechiness + acousticness + instrumentalness + liveness + valence
         + tempo + duration_ms + time_signature, data = spotify)
summary(m1)
```

```
##
## Call:
## lm(formula = track_popularity ~ year + artist_popularity + danceability +
##     energy + key + loudness + mode + speechiness + acousticness +
##     instrumentalness + liveness + valence + tempo + duration_ms +
##     time_signature, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.361  -4.055   1.326   7.111  22.151
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.013e+01  9.716e+01   0.207   0.8359
## year               1.940e-02  4.826e-02   0.402   0.6878
## artist_popularity  2.903e-01  2.466e-02  11.773   <2e-16 ***
## danceability      -2.723e-01  2.255e+00  -0.121   0.9039
## energy             7.813e-01  2.598e+00   0.301   0.7636
## key               -8.418e-02  7.327e-02  -1.149   0.2507
## loudness           5.649e-02  1.784e-01   0.317   0.7516
## mode               3.036e-02  5.400e-01   0.056   0.9552
## speechiness       -3.911e+00  2.886e+00  -1.355   0.1754
## acousticness       1.220e+00  1.532e+00   0.796   0.4259
## instrumentalness   1.341e+00  3.219e+00   0.417   0.6770
## liveness           3.389e-01  1.974e+00   0.172   0.8637
## valence           -1.338e+00  1.450e+00  -0.923   0.3563
## tempo             -7.565e-03  9.770e-03  -0.774   0.4388
## duration_ms       -5.967e-06  7.010e-06  -0.851   0.3948
## time_signature    -1.918e+00  1.091e+00  -1.758   0.0789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.5 on 2286 degrees of freedom
## Multiple R-squared:  0.08181,    Adjusted R-squared:  0.07579
## F-statistic: 13.58 on 15 and 2286 DF,  p-value: < 2.2e-16
```

Our full model outputs an R squared of .08181 and an adjusted R squared of .07579. Only 8.181 percent of variability in track popularity is explained by this model. The R squared may be significantly low, but this is expected when working with real world data. Our goal moving forward is to determine if we can improve this model in any way.

At significance level alpha .1, there appears to be two significant predictors: artist_popularity and time_signature.

Multicollinearity Analysis on Full Model

```
vif(m1)
```

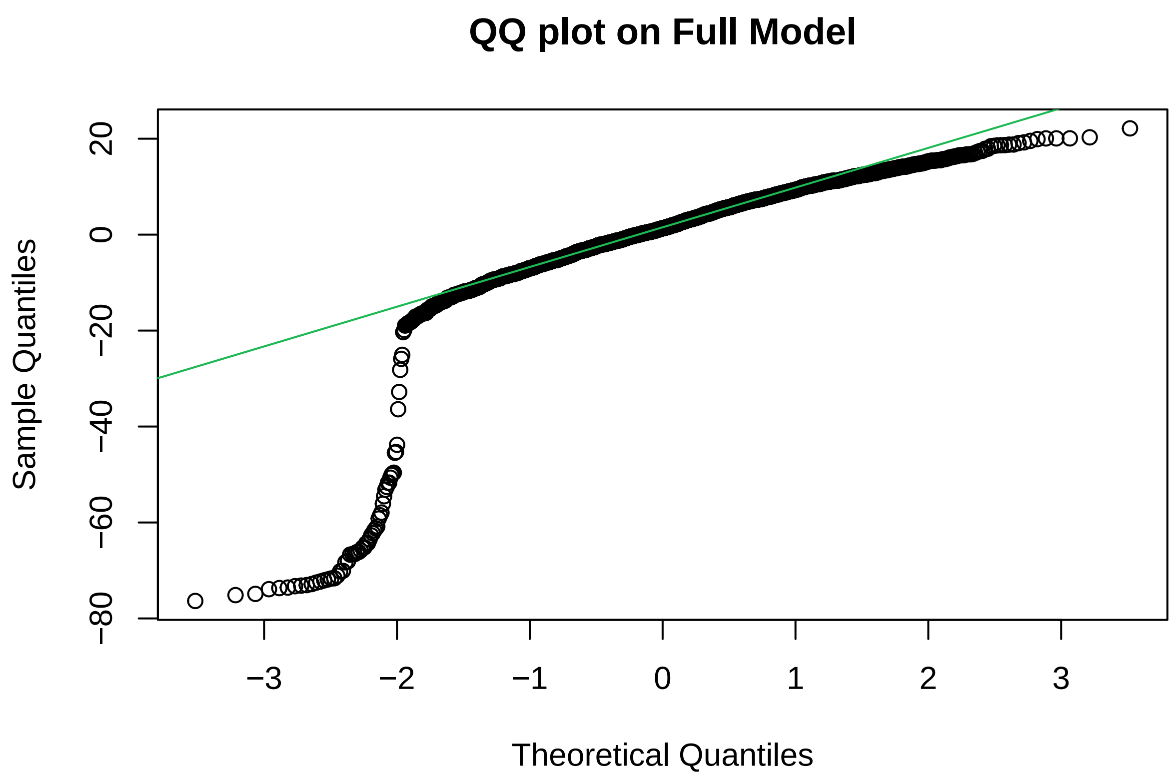
```
##               year artist_popularity      danceability            energy
##           1.657293          1.301080          1.472171          2.730986
##                key          loudness              mode       speechiness
##           1.020030          2.057119          1.040921          1.070841
##       acousticness  instrumentalness          liveness           valence
##           1.532800          1.061481          1.040913          1.622944
##              tempo       duration_ms    time_signature
##           1.082204          1.256755          1.032846
```

The VIF scores are all below 10, indicating multicollinearity is not an issue with this model.

Full Model Assumption Checks

Normality

```
qqnorm(m1$residuals, main = "QQ plot on Full Model")
qqline(m1$residuals, col = "#1DB954")
```



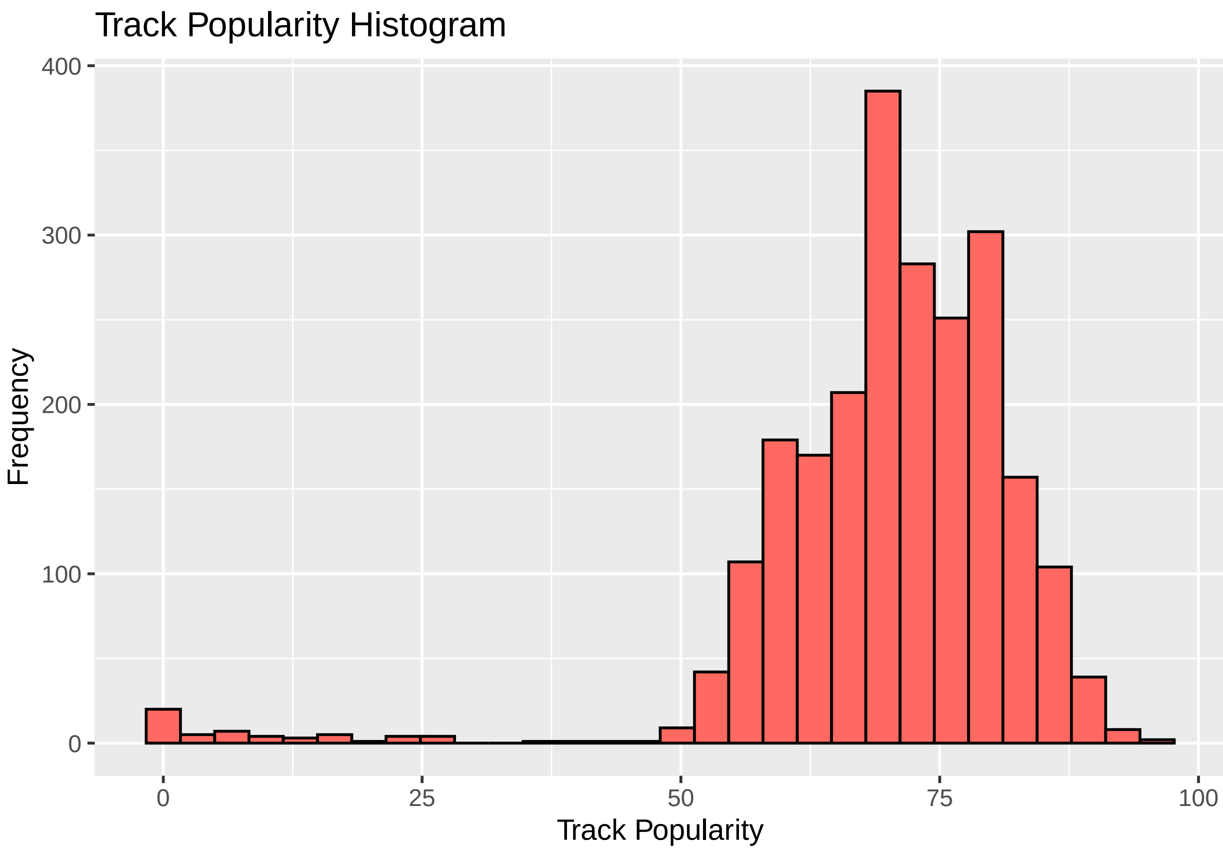

```
shapiro.test(m1$residuals)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  m1$residuals
## W = 0.7233, p-value < 2.2e-16
```

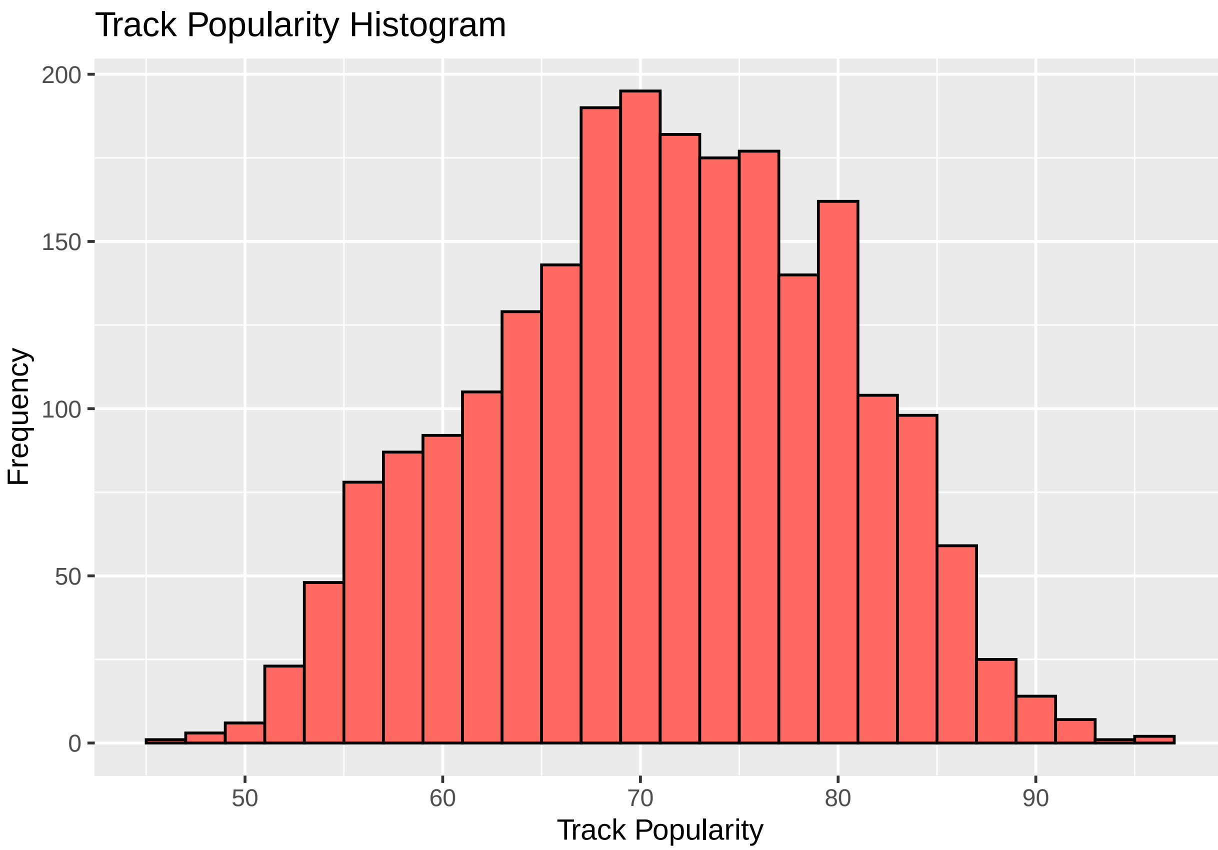
QQline heavily deviates with respect to the normality assumption. Because of this, we decided to remove low outlier track popularity scores less than.

```
ggplot(data = spotify, aes(x = track_popularity)) +
  geom_histogram(fill = "#FF6961", color = 1) +
  labs(title = "Track Popularity Histogram", x = "Track Popularity", y = "Frequency")
```




```
spotify <- filter(spotify, track_popularity >= 45)
```

```
ggplot(data = spotify, aes(x = track_popularity)) +
  geom_histogram(binwidth = 2, fill = "#FF6961", color = 1) +
  labs(title = "Track Popularity Histogram", x = "Track Popularity", y = "Frequency")
```




```r
#null model with no predictors
stepwise_null_model = lm(track_popularity ~ 1, spotify)

#full model with all relevant predictors
stepwise_full_model <- lm(track_popularity ~ year + artist_popularity + danceability + energy + key +
           loudness + mode + speechiness + acousticness + instrumentalness + liveness + valence
         + tempo + duration_ms + time_signature, data = spotify)
```

```r
stepwise_model <- step(stepwise_null_model, scope = list(lower = stepwise_null_model, upper =
stepwise_full_model), direction = "both")
```

```
## Start:  AIC=9792.85
## track_popularity ~ 1
##
##                      Df Sum of Sq    RSS    AIC
## + artist_popularity  1     44753 130879 9134.3
## + year               1     38067 137564 9246.2
## + duration_ms        1      4762 170869 9733.1
## + acousticness       1      2422 173209 9763.7
## + energy             1      2165 173467 9767.0
## + valence            1      2118 173514 9767.6
## + time_signature     1       456 175176 9789.0
## + instrumentalness   1       421 175211 9789.5
## + speechiness        1       414 175218 9789.5
## + loudness           1       328 175304 9790.6
## <none>                           175631 9792.8
## + key                1       121 175511 9793.3
```

```
## + liveness           1        15 175617 9794.7
## + mode               1         4 175628 9794.8
## + danceability       1         2 175630 9794.8
## + tempo              1         0 175631 9794.8
##
## Step:  AIC=9134.26
## track_popularity ~ artist_popularity
##
##                    Df Sum of Sq    RSS    AIC
## + year              1     12617 118262 8908.6
## + duration_ms       1      3904 126975 9068.2
## + acousticness      1      1295 129584 9113.9
## + speechiness       1       979 129900 9119.4
## + energy            1       626 130252 9125.5
## + time_signature    1       256 130622 9131.9
## + valence           1       253 130626 9131.9
## <none>                          130879 9134.3
## + loudness          1        96 130783 9134.6
## + mode              1        91 130788 9134.7
## + liveness          1        33 130846 9135.7
## + key               1        21 130858 9135.9
## + instrumentalness  1        12 130867 9136.1
## + tempo             1         8 130871 9136.1
## + danceability      1         1 130878 9136.2
## - artist_popularity 1     44753 175631 9792.8
##
## Step:  AIC=8908.59
## track_popularity ~ artist_popularity + year
##
##                    Df Sum of Sq    RSS    AIC
## + speechiness       1     894.9 117367 8893.5
## + duration_ms       1     435.1 117827 8902.3
## + acousticness      1     318.1 117944 8904.5
## <none>                          118262 8908.6
## + time_signature    1      97.1 118165 8908.7
## + mode              1      96.7 118166 8908.8
## + danceability      1      86.0 118176 8909.0
## + key               1      25.6 118237 8910.1
## + loudness          1      11.2 118251 8910.4
## + liveness          1      11.0 118251 8910.4
## + energy            1       4.1 118258 8910.5
## + valence           1       3.0 118259 8910.5
## + tempo             1       0.5 118262 8910.6
## + instrumentalness  1       0.1 118262 8910.6
## - year              1   12616.6 130879 9134.3
## - artist_popularity 1   19302.1 137564 9246.2
##
## Step:  AIC=8893.53
## track_popularity ~ artist_popularity + year + speechiness
##
##                    Df Sum of Sq    RSS    AIC
## + duration_ms       1     411.7 116956 8887.6
## + acousticness      1     270.8 117097 8890.3
## <none>                          117367 8893.5
```

```
## + mode               1      63.4 117304 8894.3
## + time_signature     1      61.9 117305 8894.3
## + valence            1      24.4 117343 8895.1
## + key                1      18.1 117349 8895.2
## + danceability       1      14.8 117352 8895.2
## + loudness           1       5.9 117361 8895.4
## + energy             1       3.5 117364 8895.5
## + tempo              1       2.3 117365 8895.5
## + liveness           1       2.2 117365 8895.5
## + instrumentalness   1       1.3 117366 8895.5
## - speechiness        1     894.9 118262 8908.6
## - year               1   12532.7 129900 9119.4
## - artist_popularity  1   19681.2 137049 9239.7
##
## Step:  AIC=8887.64
## track_popularity ~ artist_popularity + year + speechiness + duration_ms
##
##                     Df Sum of Sq    RSS    AIC
## + acousticness       1     295.5 116660 8884.0
## <none>                           116956 8887.6
## + mode               1      66.3 116889 8888.4
## + time_signature     1      65.6 116890 8888.4
## + danceability       1      29.7 116926 8889.1
## + energy             1      20.9 116935 8889.2
## + key                1      18.2 116937 8889.3
## + instrumentalness   1       2.4 116953 8889.6
## + liveness           1       2.2 116954 8889.6
## + tempo              1       1.1 116955 8889.6
## + valence            1       0.3 116955 8889.6
## + loudness           1       0.0 116956 8889.6
## - duration_ms        1     411.7 117367 8893.5
## - speechiness        1     871.5 117827 8902.3
## - year               1    9125.5 126081 9054.4
## - artist_popularity  1   20090.6 137046 9241.7
##
## Step:  AIC=8883.96
## track_popularity ~ artist_popularity + year + speechiness + duration_ms +
##     acousticness
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           116660 8884.0
## + loudness           1      61.6 116599 8884.8
## + mode               1      56.1 116604 8884.9
## + time_signature     1      40.8 116619 8885.2
## + energy             1      30.0 116630 8885.4
## + key                1      16.8 116643 8885.6
## + valence            1      13.6 116647 8885.7
## + danceability       1      10.6 116650 8885.8
## + tempo              1       7.3 116653 8885.8
## + instrumentalness   1       2.5 116658 8885.9
## + liveness           1       0.0 116660 8886.0
## - acousticness       1     295.5 116956 8887.6
## - duration_ms        1     436.4 117097 8890.3
## - speechiness        1     822.1 117482 8897.7
```

```
## - year               1    8354.3 125014 9037.3
## - artist_popularity  1   20180.4 136841 9240.3
```

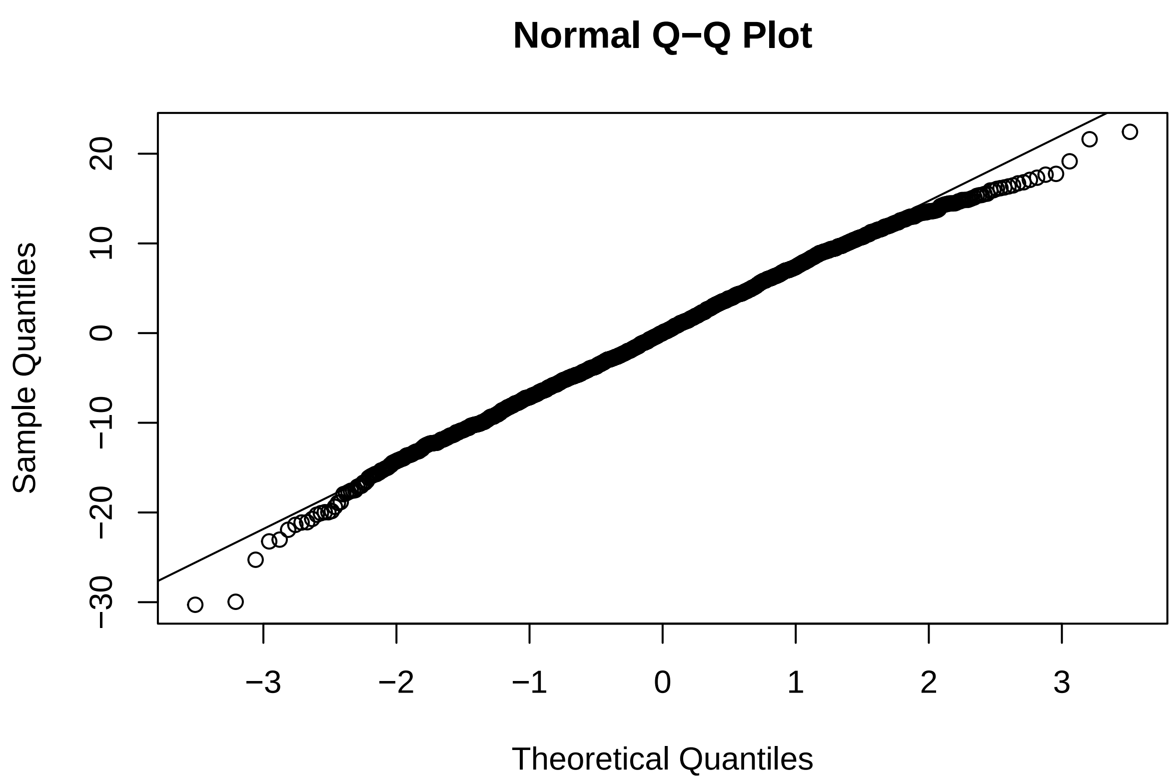
```r
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = track_popularity ~ artist_popularity + year + speechiness +
##     duration_ms + acousticness, data = spotify)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.2886  -4.8244   0.0135   5.0460  22.4378
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.415e+02  5.469e+01 -11.729  < 2e-16 ***
## artist_popularity  2.810e-01  1.427e-02  19.685  < 2e-16 ***
## year               3.454e-01  2.727e-02  12.665  < 2e-16 ***
## speechiness       -6.459e+00  1.626e+00  -3.973 7.32e-05 ***
## duration_ms       -1.149e-05  3.971e-06  -2.895  0.00383 **
## acousticness       1.763e+00  7.403e-01   2.382  0.01730 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.217 on 2240 degrees of freedom
## Multiple R-squared:  0.3358, Adjusted R-squared:  0.3343
## F-statistic: 226.5 on 5 and 2240 DF,  p-value: < 2.2e-16
```

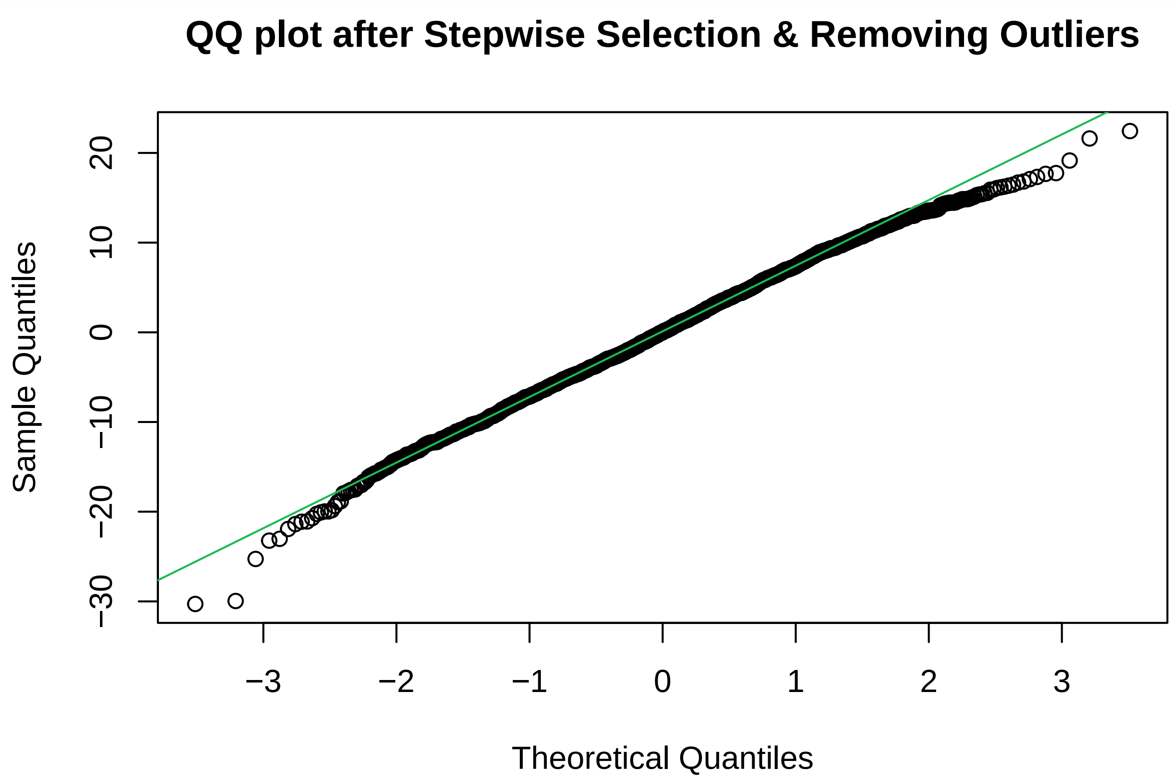
```r
vif(stepwise_model)
```

```
## artist_popularity              year       speechiness       duration_ms
##          1.284425          1.508514          1.005443          1.176939
##      acousticness
##          1.035201
```

```r
qqnorm(stepwise_model$residuals)
qqline(stepwise_model$residuals)
```



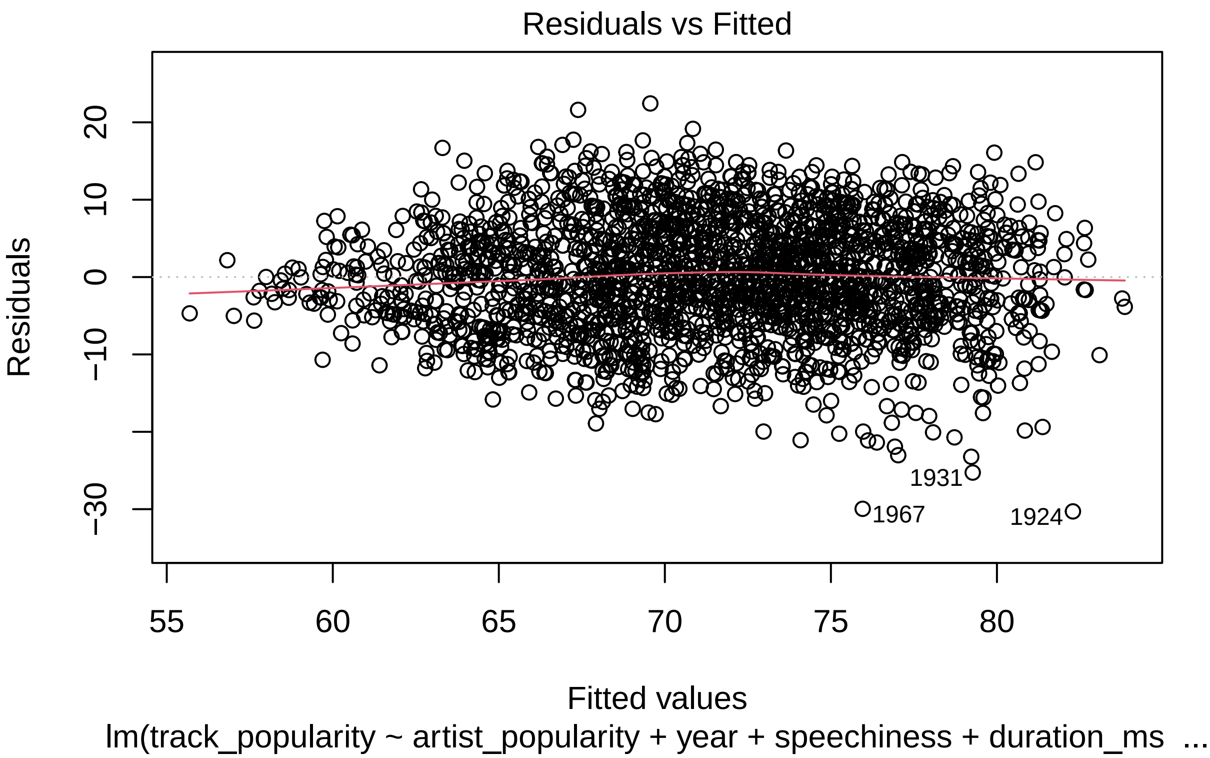

```r
qqnorm(stepwise_model$residuals, main = "QQ plot after Stepwise Selection & Removing Outliers")
qqline(stepwise_model$residuals, col = "#1DB954")
```

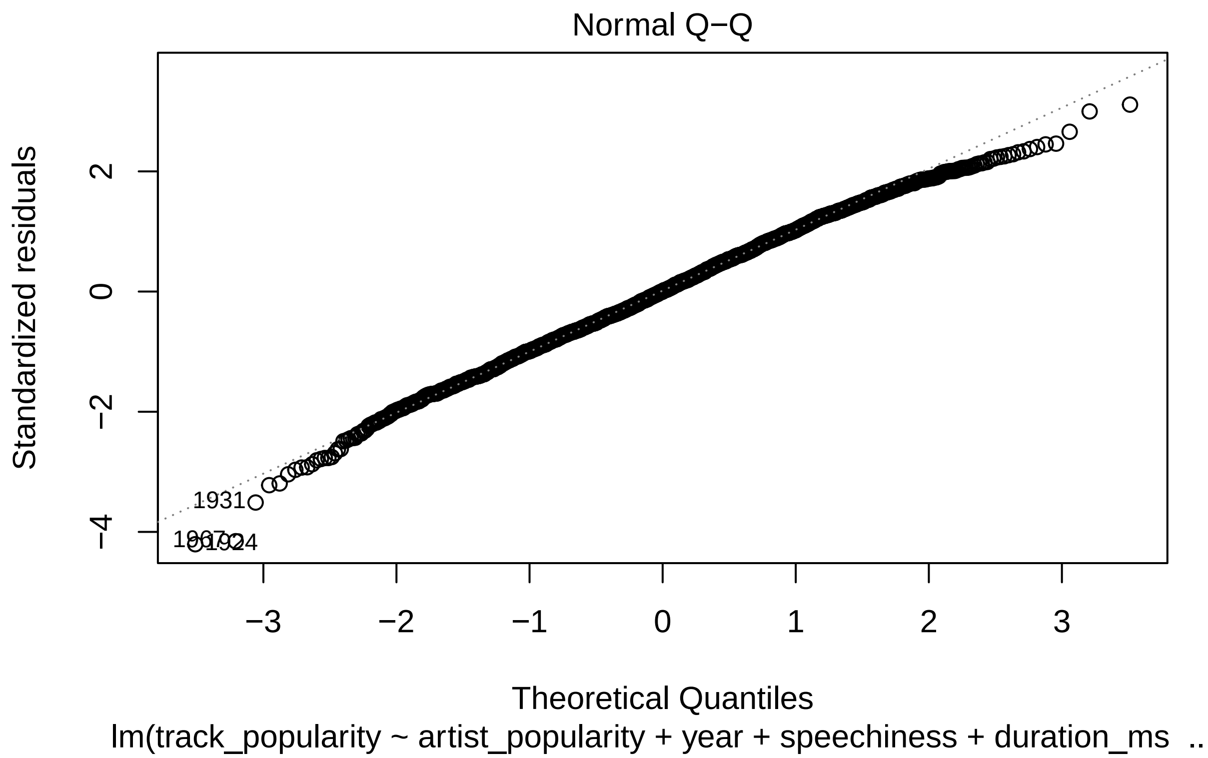


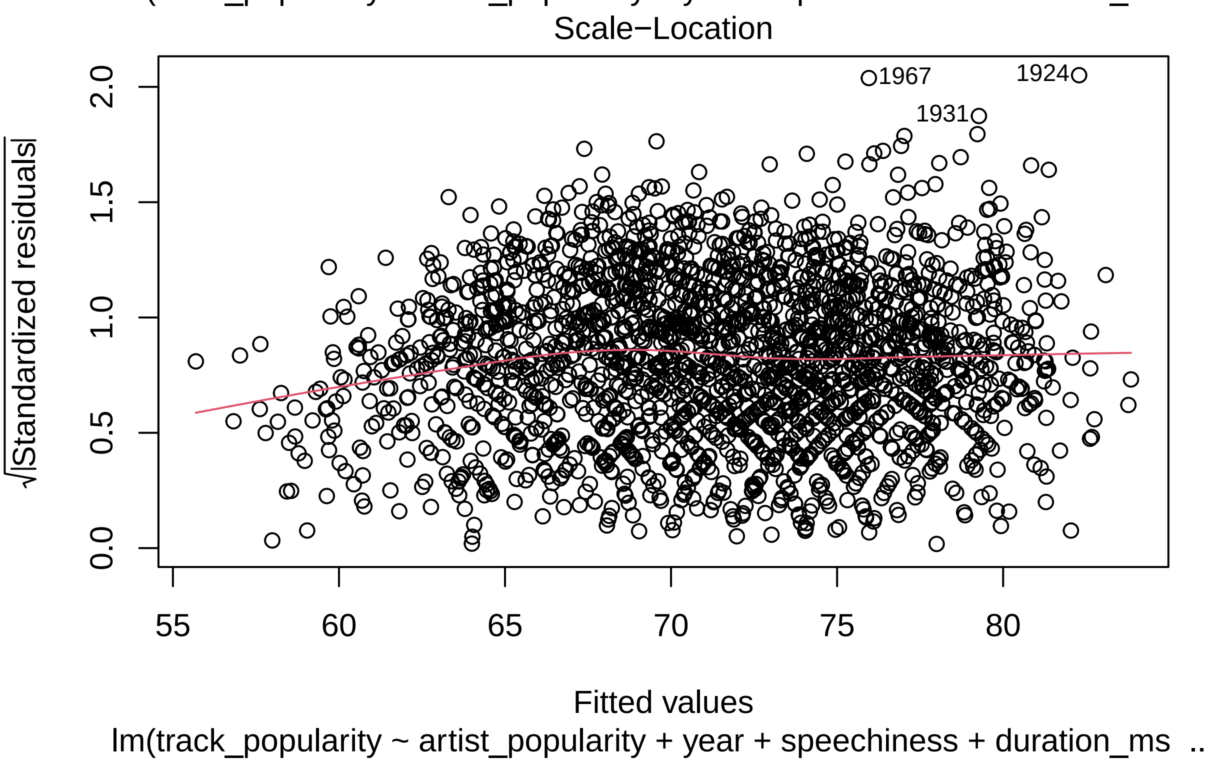
```
shapiro.test(stepwise_model$residuals)
```

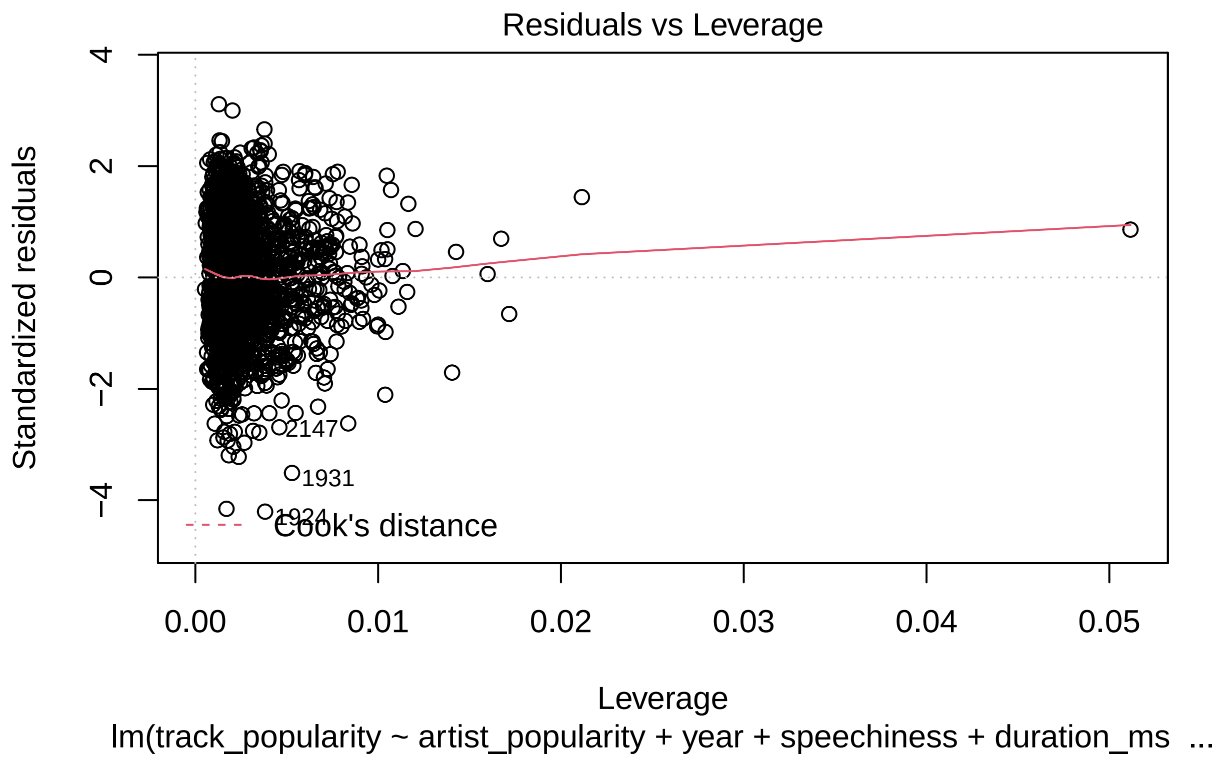
```
##
##  Shapiro-Wilk normality test
##
## data:  stepwise_model$residuals
## W = 0.99691, p-value = 0.0001679
```

```
plot(stepwise_model)
```




Normal Q−Q
Standardized residuals
1931
1967
1924
Theoretical Quantiles
lm(track_popularity ~ artist_popularity + year + speechiness + duration_ms ...


Scale−Location
√|Standardized residuals|
1967
1924
1931
Fitted values
lm(track_popularity ~ artist_popularity + year + speechiness + duration_ms ...




```
dwt(stepwise_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.6183224     0.7614197       0
##  Alternative hypothesis: rho != 0
```

```
spotify_dummy <- spotify |> dplyr::select(artist_popularity, year, speechiness, duration_ms, acousticne
```

```
cor(spotify_dummy)
```

```
##                    artist_popularity        year speechiness duration_ms
## artist_popularity         1.00000000  0.44743990  0.05153502 -0.03099482
## year                      0.44743990  1.00000000  0.01213833 -0.35789644
## speechiness               0.05153502  0.01213833  1.00000000  0.02077621
## duration_ms              -0.03099482 -0.35789644  0.02077621  1.00000000
## acousticness              0.06286280  0.17427485 -0.04468880 -0.03326747
##                    acousticness
## artist_popularity   0.06286280
## year                0.17427485
## speechiness        -0.04468880
## duration_ms        -0.03326747
## acousticness        1.00000000
```

Validation

Lastly, we wanted to run a validation set approach to directly compare the full model with the stepwise model

```
set.seed(167)
```

Final Model

```
m1 <- lm(track_popularity ~ year + artist_popularity + danceability + energy + key +
           loudness + mode + speechiness + acousticness + instrumentalness + liveness + valence
         + tempo + duration_ms + time_signature, data = spotify)
```

```
set.seed(167)
dim(spotify)
```

```
## [1] 2246   16
```

```
train.idx <- sample(2246, 1123)

train <- spotify[train.idx, ]
test <- spotify[-train.idx, ]

full.model.train <- lm(track_popularity ~ year + artist_popularity + danceability + energy + key +
           loudness + mode + speechiness + acousticness + instrumentalness + liveness + valence
         + tempo + duration_ms + time_signature, data = train)

stepwise.model.train <- lm(formula = track_popularity ~ artist_popularity + year + speechiness +
    duration_ms + acousticness, data = train)
```

Traning and Validation Set MSE for Full Model

```
full.model.train.MSE <- mean((train$track_popularity - predict(full.model.train))^2)
full.model.test.MSE <- mean((train$track_popularity - predict(full.model.train, newdata = test))^2)
full.model.train.MSE
```

```
## [1] 51.26436
```

```
full.model.test.MSE
```

```
## [1] 95.73256
```

Traning and Validation Set MSE for Stepwise Model

```
stepwise.model.train.MSE <- mean((train$track_popularity - predict(stepwise.model.train))^2)
stepwise.model.test.MSE <- mean((train$track_popularity - predict(stepwise.model.train, newdata = test)
stepwise.model.train.MSE
```

```
## [1] 51.48521
```

```
stepwise.model.test.MSE
```

```
## [1] 95.4638
```