Statistical Modelling and Inference II

Project

Group number: 06

Group members:

I-Chun LIN	a1745127
Riris A SILALAHI	a1750682
Yuting YANG	a1692116
Sami ELMASRI	a1723534
Krunal Gohil	a1746645

Data: 25/10/2018

Introduction

Predictive modelling is a statistical technique that use historical data to build a future prediction. The predictive models have become significantly important for many kinds of businesses because it can help businesses to get a comprehensive understanding about customer's behavior, sales or profit pattern, marketing campaign optimization, risk management, fraud detection and others. In music related businesses, for example, predictive models can help businesses and artists to get knowledge about how does a song or music become popular in the market. This understanding can be used as a benchmark for their future productions.

The main aim of this project is to find the best predictive model that can be used to predict popularity track on Spotify based on several features. Spotify is a digital music streaming platform that provides their users access to millions of songs and other content from artists all over the world. Specifically, the data used for analysis was taken from tracks of some selected artists that represent each decade from 1950 to 2000 on Spotify. The linear model will be created by considering some attributes like danceability, energy, key, loudness, mode, duration of the tracks in milliseconds, the signature of time, and decade. The given data consists of 2081 tracks that are sung by 6 different artists.

Depending on the dataset, there are several types of regression analysis model that can be utilized in generating predictive models and for this project, we will use a linear model. A linear model is one of the prediction technique that is simple to interpret and allow us to inspect the contribution of each involving predictors. The analysis for this project is done by using the R programming language and the

processes are divided into five main steps. The first step is data cleaning, in this step, we will deal with any missing data and replacing, modifying or deleting any dirty data. The second step is to define all variables involved, this step will allow us to get a deeper understanding of our data by looking at types of variables, summary statistics and use some visualization to see the distribution each variable. The third step is a bivariate analysis, in this step, we attempt to see the relationship between each of the predictor variables by using plots. The fourth step is model fitting, in this step, we will use some statistical techniques to find the best model based on the given data. The last step is assumption checking and prediction.

Data Description

According to the data given in *spotify.xlsx*, there are 2081 subjects in total which are songs of some typical artists recorded in specific albums.
The dataset consists of 21 variables, one of which is the response variable named popularity. There are 8 of these variables would be used as predictors which are danceability, energy, key, loudness, mode, duration_ms, time_signature and decade.

popularity:
The response variable popularity is a numeric. Integers are used to represent how much the track were loved by people. A greater value means a higher degree of popularity of the corresponding track.

album_name:
It is a categorical variable with 141 levels each of which represent an album.

track_name:
It is a categorical variable with 2081 levels and every single level is a unique name of that track. It is more likely to be an identification value than a predictor.

track_uri:
It is a categorical variable with 2081 levels. Similar to the track_name, track_uri are unique for every subject and it doesn't seem to be useful on predicting popularity.

artist:
There are 6 levels of this categorical variable, each of them represents an artist.

artist_uri:
It is the Spotify uri for artists and it is unique for every artist. Of course, it is a categorical variable with 6 levels as well.

danceability:
It is a numeric variable ranging from 0.0 to 1.0, describing how much a track is suitable for dancing. **This is one of our predictors.**

energy:
It is a numeric variable ranging from 0.0 to 1.0, representing a perceptual measure of intensity and activity. **This is one of our predictors.**

key:
It is a categorical variable with 12 levels, representing the key of every track. **This is one of our predictors.**

loudness:
It is a numeric variable ranging between -60 and 0 decibels, describing the quality of a sound that is the primary psychological correlate of physical strength. **This is one of our predictors.**

mode:
It is a categorical variable with 2 levels which are major and minor. It indicates the modality of track. **This is one of our predictors.**

speechiness:
It is a numeric variable ranging from 0.0 to 1.0, indicating how much spoken words were in the track.

acousticness:
It is a numeric variable ranging from 0.0 to 1.0, indicating how confident is it that the track is acoustic.

instrumentalness:
It is a numeric variable ranging from 0.0 to 1.0. The value indicates how likely is this track to be no vocals.

liveness:
It is a numeric variable ranging from 0.0 to 1.0. The value indicates the likelihood that the track was performed live.

valence:
It is a numeric variable ranging from 0.0 to 1.0, indicating whether the track is positive or negative. The larger the value is, the more positive the track is.

tempo:
It is a numeric variable that describes the overall estimated tempo of a track in beats per minute(BPM)

duration_ms:
It is the duration of track in milliseconds and it is numeric. **This is one of our predictors.**

time_signature:
It is numeric and represents the overall time signature of tracks. **This is one of our predictors.**

key_mode:
It is a categorical variable and it is basically a combination of key and mode.

decade:
It is a quantitative interval variable, indicating the major duration of tracks. **This is one of our predictors.**

Cleaning of Data

Read in the data

read data from spotify.xlsx

```
library(magrittr)
library(broom)
library(tidyverse)
library(stringr)
library(forcats)
library(modelr)
library(readxl)

spotify <- read_excel("spotify.xlsx") #read data from spotify.xlsx
spotify <- select(spotify,popularity, danceability, energy,key,loudness,mode,
duration_ms,time_signature,decade)
```

Check if data is clean

We use the summary and class functions to check missing data. As the summary shows, there is no missing data in those columns.

check popularity

```
summary(spotify$popularity) #check missing value

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   19.00   26.00   29.42   41.00   82.00

class(spotify$popularity) #check type

## [1] "numeric"
```

check danceability

```
summary(spotify$danceability) #check missing value

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3770  0.4990  0.5013  0.6270  0.9630

class(spotify$danceability) #check type

## [1] "numeric"
```

check energy

```r
summary(spotify$energy) #check missing value of coulm energy

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00589 0.40600 0.62200 0.59559 0.80900 0.99800

class(spotify$energy) #check type

## [1] "numeric"
```

check loudness

```r
summary(spotify$loudness) #check missing value

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -30.644 -12.628  -9.291 -10.056  -6.651  -0.933

class(spotify$loudness) #check type

## [1] "numeric"
```

check duration_ms

```r
summary(spotify$duration_ms) #check missing value

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13213  153733  203640  216049  261733 1187253

class(spotify$duration_ms) #check type

## [1] "numeric"
```

check time_signature

```r
summary(spotify$time_signature) #check missing value

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   4.000   3.861   4.000   5.000

class(spotify$time_signature) #check type

## [1] "numeric"
```

transfer the category predictors to factors

transfer key

```r
spotify$key<-factor(spotify$key,levels =c("A","A#","B","C","C#","D","D#","E",
"F","F#","G","G#"))
class(spotify$key) #check type

## [1] "factor"
```

```r
summary(spotify$key) #check missing value
```

```
##   A  A#   B   C  C#   D  D#   E   F  F#   G  G#
## 272 115 143 352 149 259  44 199 155  71 240  82
```

transfer mode

```r
spotify$mode<-factor(spotify$mode,levels =c("major","minor"))
class(spotify$mode) #check type
```

```
## [1] "factor"
```

```r
summary(spotify$mode) #check missing value
```

```
## major minor
##  1618   463
```

transfer decade

```r
spotify$decade<-factor(spotify$decade,levels =c("00s","50s","60s","70s","80s
","90s"))
class(spotify$decade) #check type
```

```
## [1] "factor"
```

```r
summary(spotify$decade)
```

```
## 00s 50s 60s 70s 80s 90s
## 194 648 247 547 253 192
```

Variable Description

In this section, we would do a detailed analysis of the distribution of all predictors taken into consideration. They are

1) Danceability
2) Energy
3) Key
4) Loudness
5) Mode
6) Duration_ms
7) Time_signature
8) Decade

Firstly, import libraries and read in the data for further analysis.

```
library(tidyverse)
library(readxl)
library(ggplot2)

spotify <- read_excel("~/Desktop/Statistics/Project/spotify.xlsx")
```

danceability:

```
summary(spotify$danceability)
ggplot(spotify, aes(x=danceability)) +
  geom_histogram(col = "black", fill = "orange") +
  labs(x = "Danceability of selected tracks on Spotify.")
```

It is a continuous quantitative variable, of which the minimum observation is 0, the maximum observation is 0.9630, the median is 0.4990 and the average is 0.5013.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.3770  0.4990  0.5013  0.6270  0.9630
```

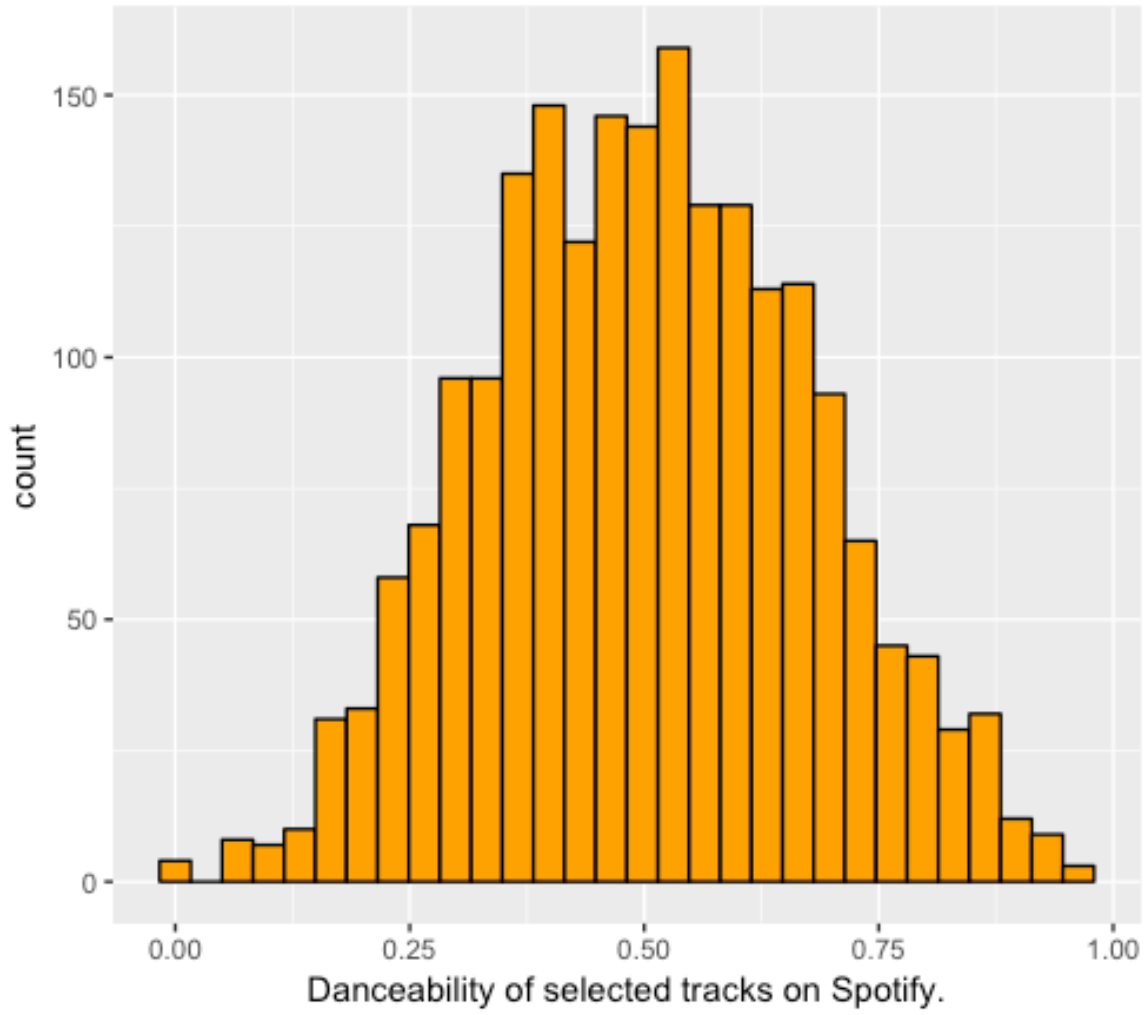

Fig 4.1

This is the distribution of the random variable named danceability. It appears to be unimodal and symmetric.

energy:

```
summary(spotify$energy)
ggplot(spotify, aes(x=energy)) +
  geom_histogram(col = "black", fill = "orange") +
  labs(x = "Energy of selected tracks on Spotify.")
```

It is a continuous quantitative variable, of which the minimum observation is 0.00589, the maximum observation is 0.99800, the median is 0.62200 and mean is 0.59560.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00589 0.40600 0.62200 0.59560 0.80900 0.99800
```

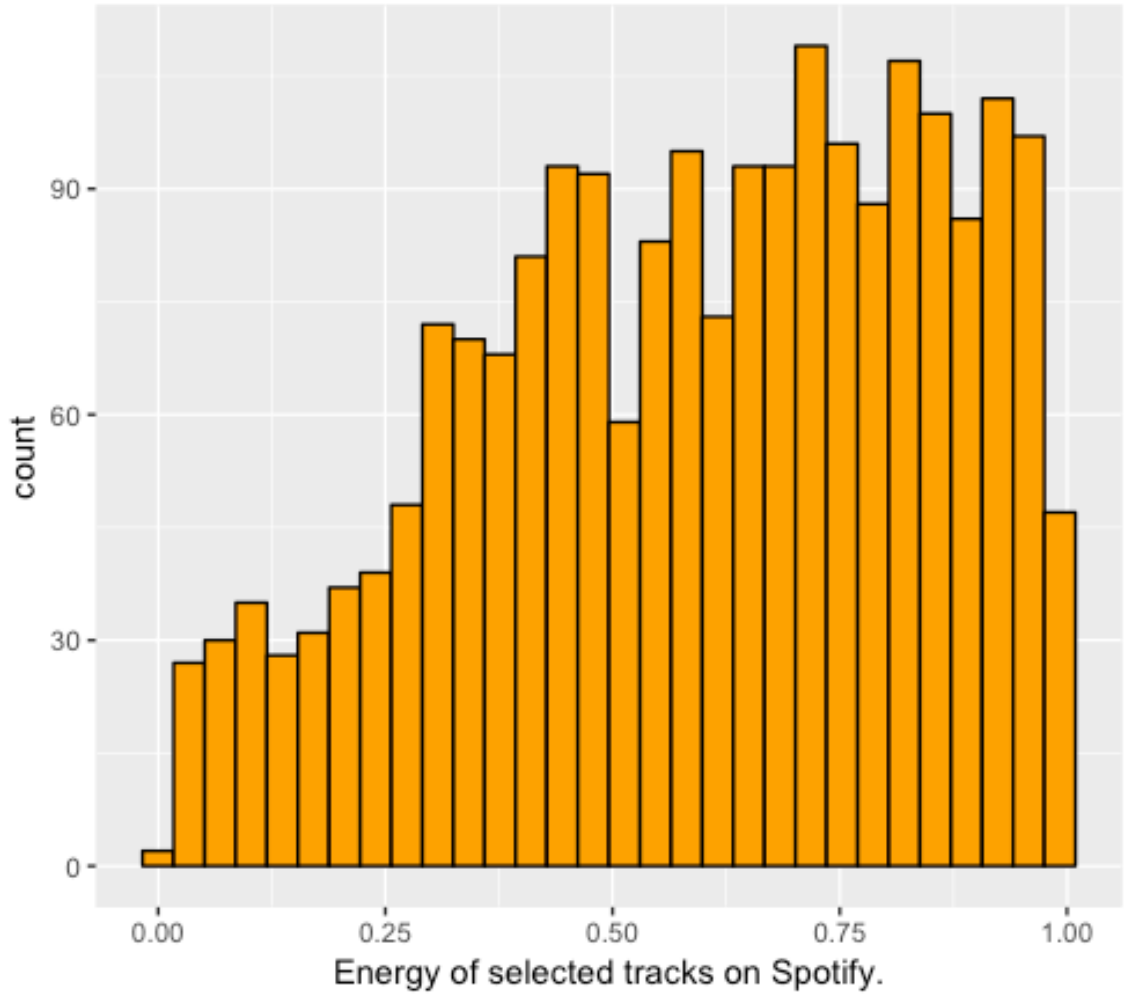

Fig 4.2

The distribution of the random variable energy appears to be unimodal and left skewed.

key:

```
table(spotify$key)
ggplot(spotify, aes(x=key)) +
  geom_bar(stat="count",width=0.5) +
  labs(x = "Key of selected tracks on Spotify.")
```

It is an ordinal categorical variable with 12 levels.

```
  A  A#   B   C  C#   D  D#   E   F  F#   G  G#
272 115 143 352 149 259  44 199 155  71 240  82
```

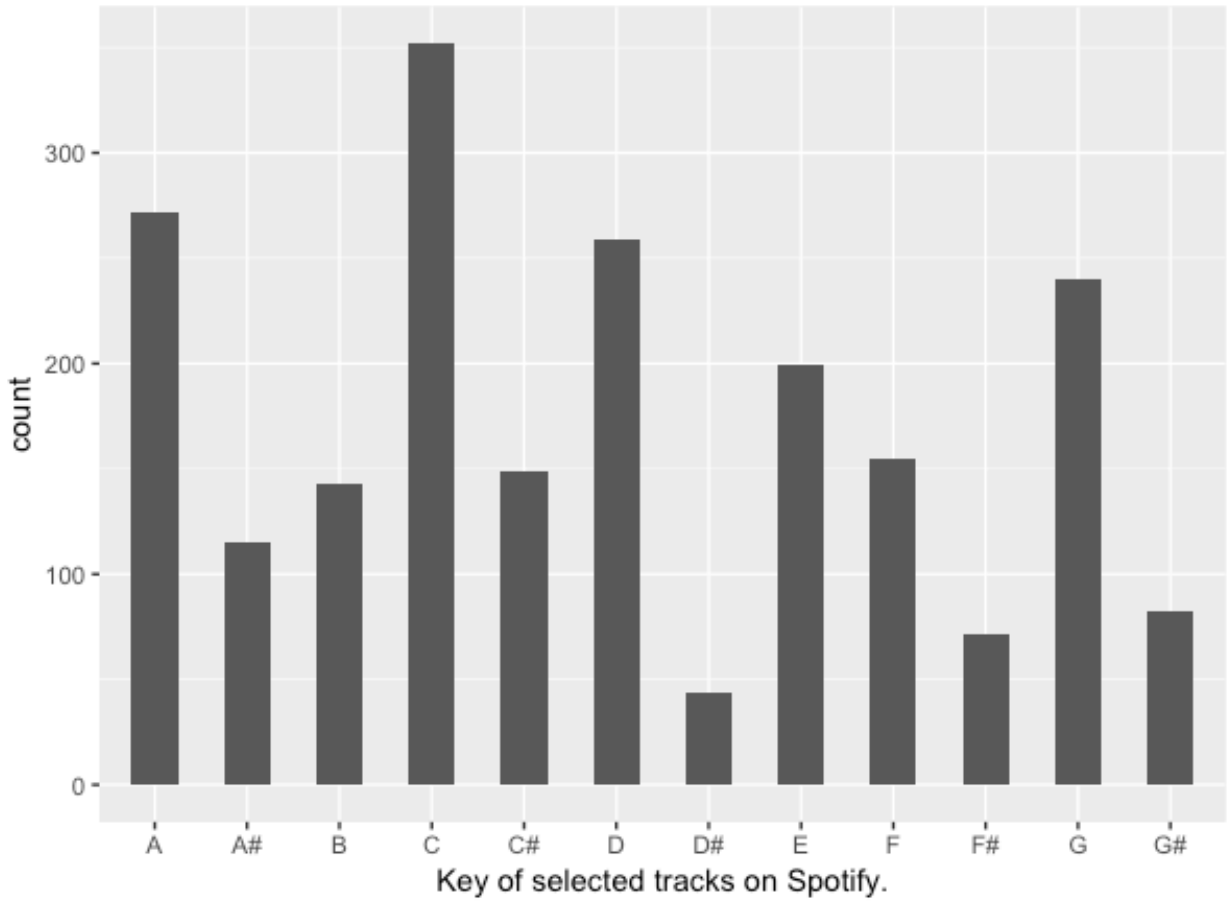

Fig 4.3

Above is the distribution of *key,* among all the 2081 observations, the most common level of the variable is C and the least common one is D#.

loudness:

```
summary(spotify$loudness)
ggplot(spotify, aes(x=loudness)) +
  geom_histogram(col = "black", fill = "orange") +
  labs(x = "Loudness of selected tracks on Spotify.")
```

It is a continuous quantitative variable, of which the minimum observation is -30.640, the maximum observation is -0.933, the median is -12.630 and mean is -10.060.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-30.640 -12.630  -9.291 -10.060  -6.651  -0.933
```

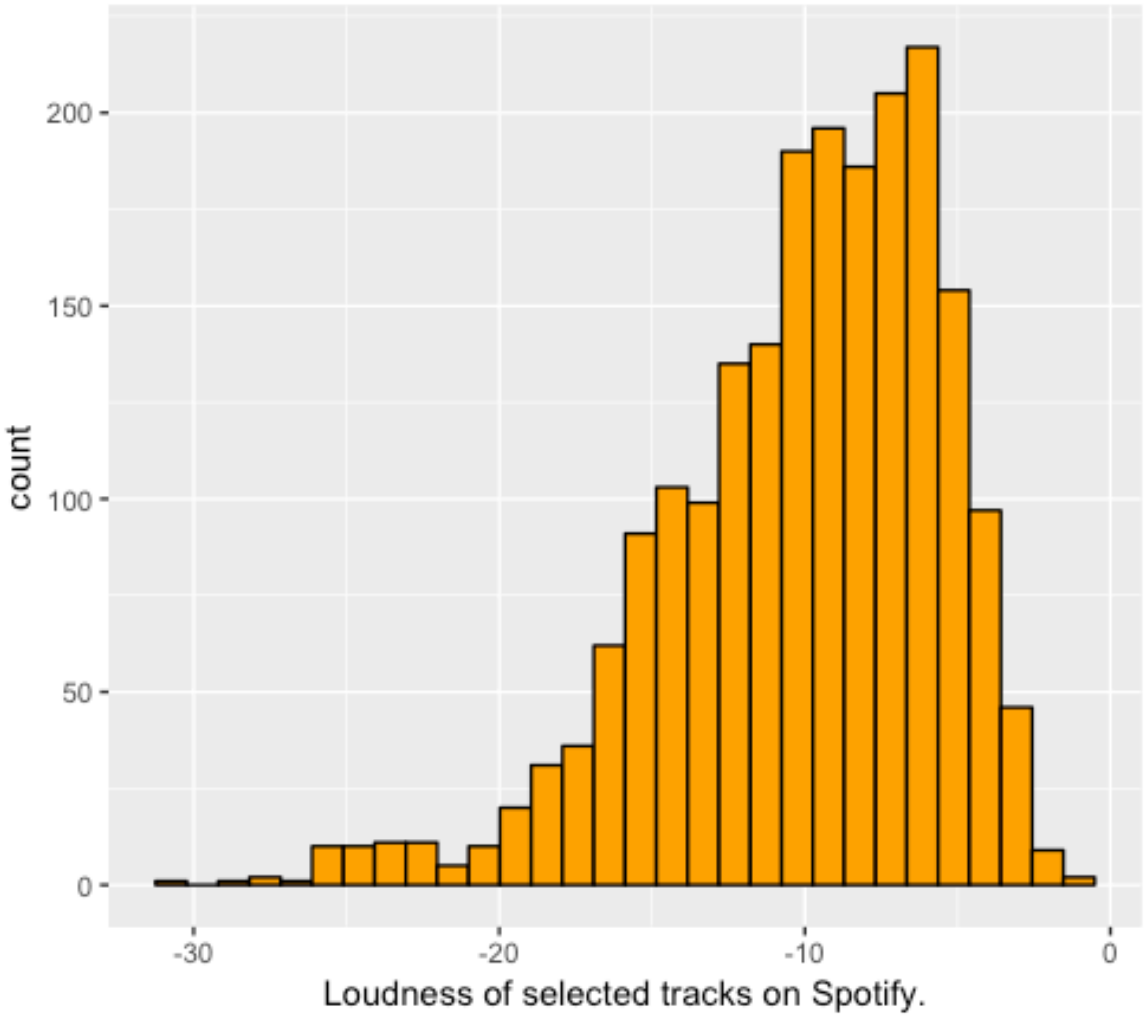

Fig 4.4

The distribution of the random variable loudness appears to be unimodal and left skewed.

mode:

```
table(spotify$mode)
ggplot(spotify, aes(x=mode)) +
  geom_bar(stat="count",width=0.5) +
  labs(x = "Mode of selected tracks on Spotify.")
```

It is a categorical variable with only two levels.

```
major minor
 1618   463
```

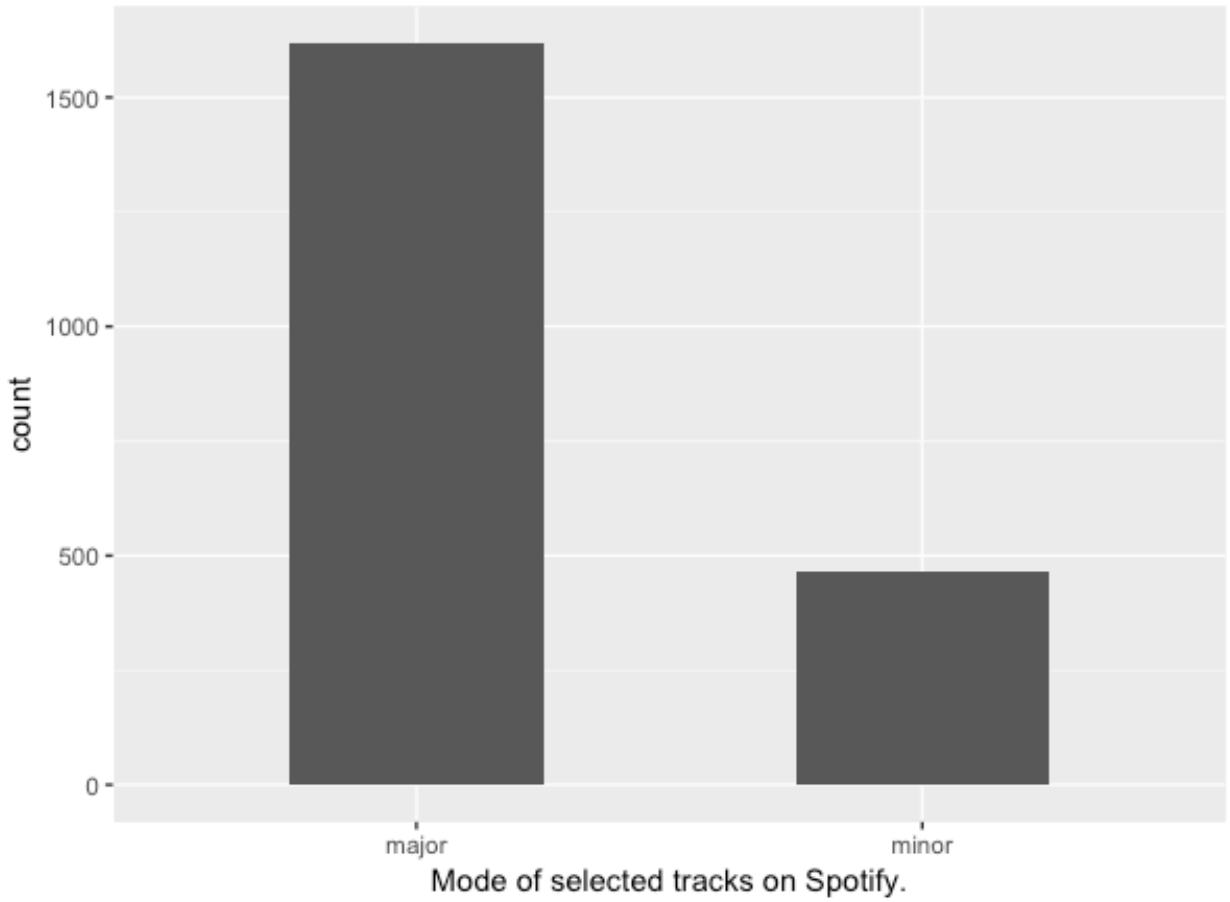

Fig 4.5

Above is the distribution of the variable mode. The most common level is major, while the least common one is minor.

duration_ms:

```
summary(spotify$duration_ms)
ggplot(spotify, aes(x=duration_ms)) +
  geom_histogram(col = "black", fill = "orange") +
  labs(x = "duration_ms of selected tracks on Spotify.")
```

It is a continuous quantitative variable, of which the minimum observation is 13210, the maximum observation is 1187000, the median is 203600 and mean is 216000.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13210  153700  203600  216000  261700 1187000
```

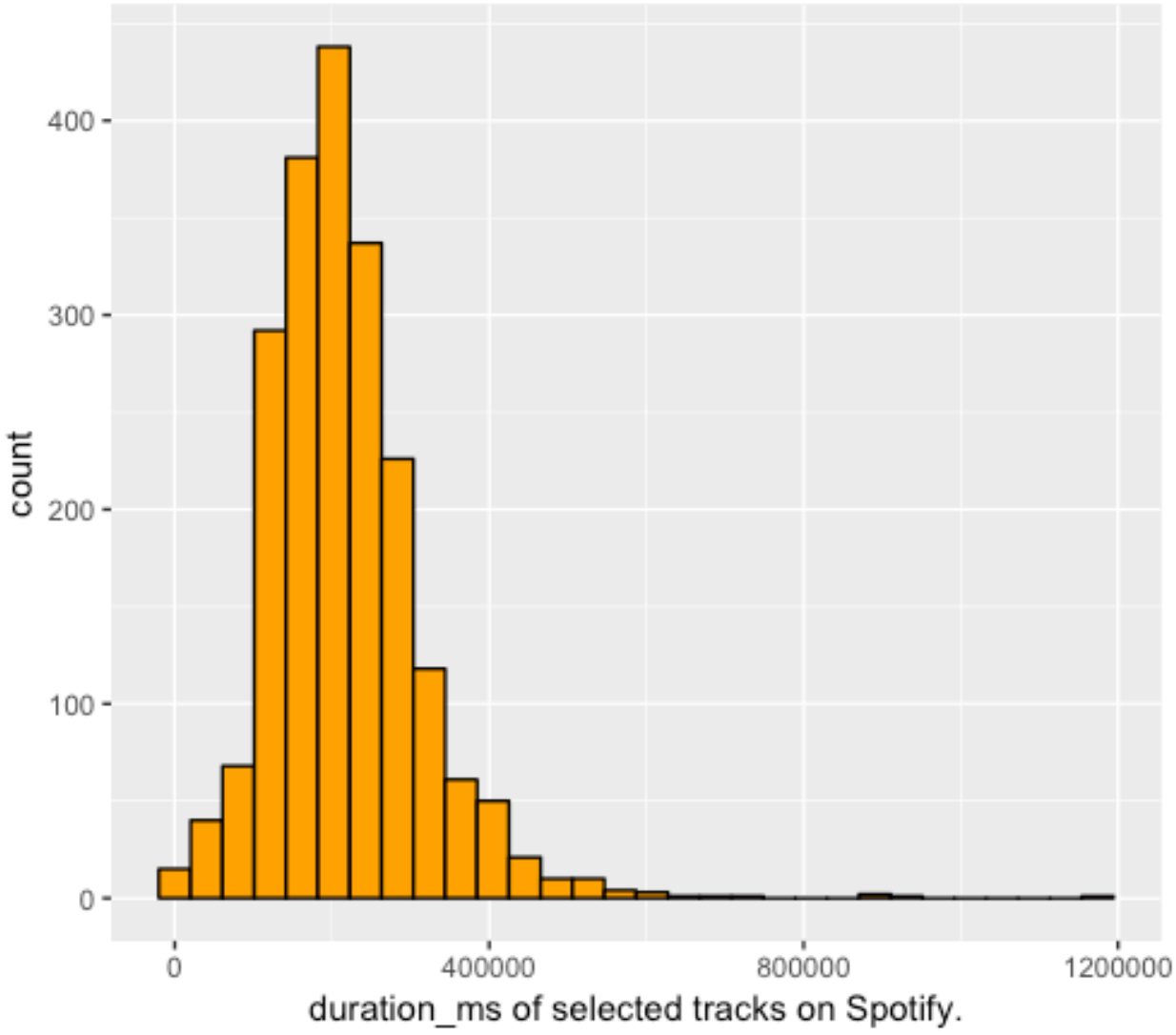

Fig 4.6

Above is the distribution of the random variable duration_ms. It appears to be unimodal and symmetric.

time_signature:

```
summary(spotify$time_signature)
ggplot(spotify, aes(x=time_signature)) +
  geom_histogram(col = "black", fill = "orange") +
  labs(x = "time_signature of selected tracks on Spotify.")
```

It is a discrete quantitative variable, of which the minimum observation is 0, the maximum observation is 5, the median is 4 and mean is 3.861.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   4.000   4.000   3.861   4.000   5.000
```

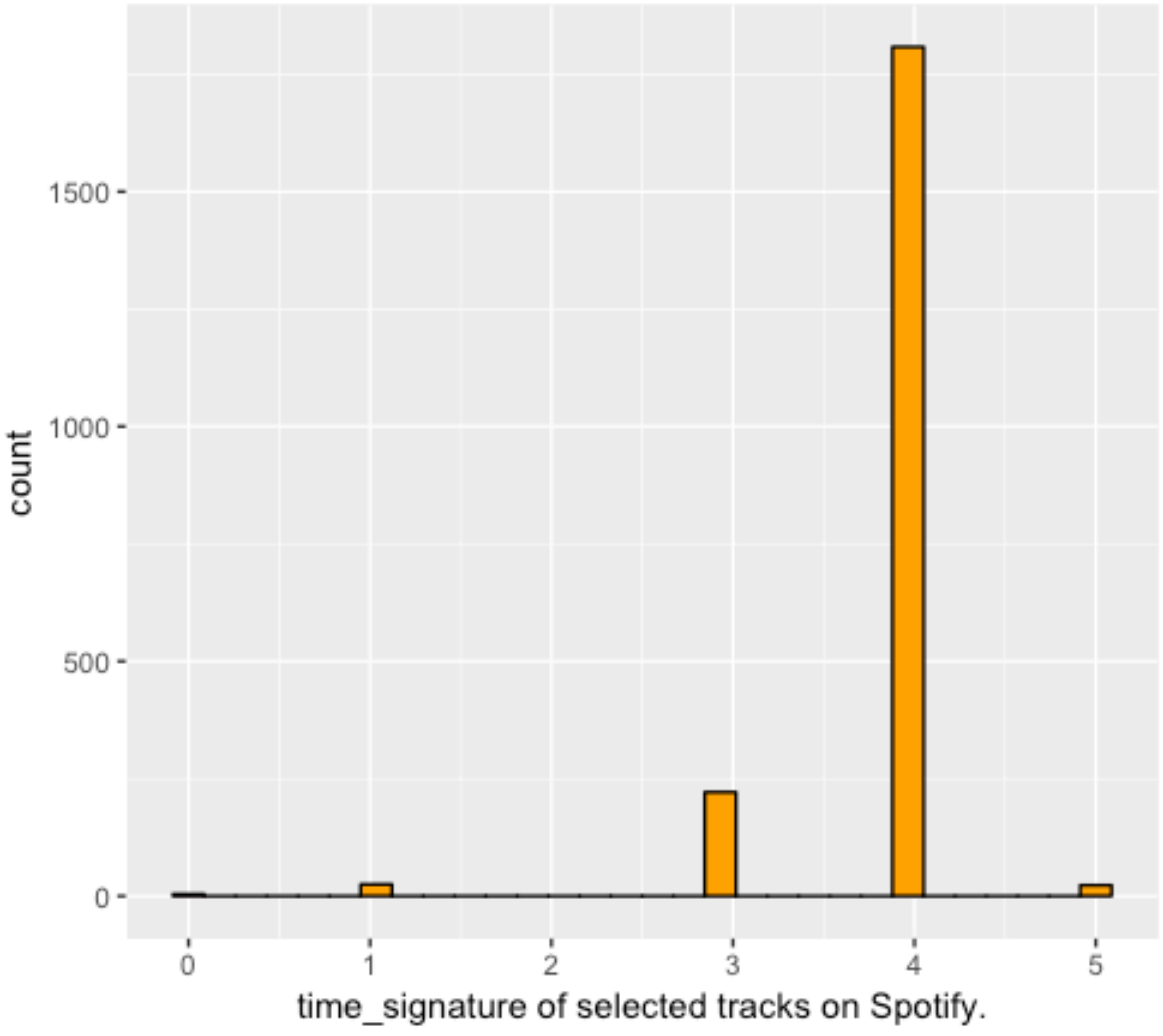

Fig 4.6

Above is the distribution of the random variable duration_ms. It appears to be unimodal and left skewed.

Decade:

```
table(spotify$decade)
ggplot(spotify, aes(x=decade)) +
  geom_bar(stat="count",width=0.5) +
  labs(x = "Decade of selected tracks on Spotify.")
```

It is a categorical variable with six levels.

```
00s 50s 60s 70s 80s 90s
194 648 247 547 253 192
```

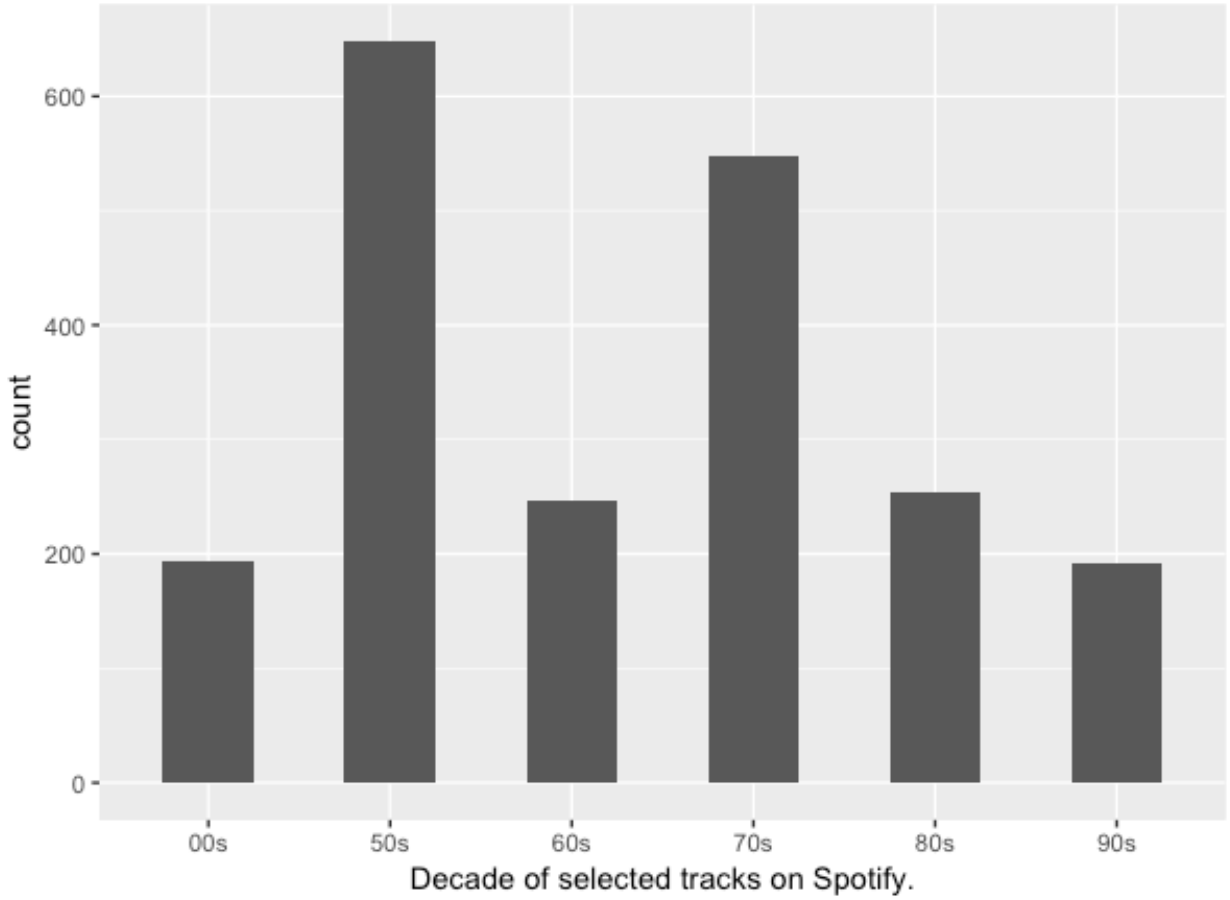

Fig 4.6

Above is the distribution of *decade,* among all the 2081 observations, the most common level of the variable is *50s* and the least common one is *90s*.

Bivariate Analysis

Bivariate statistical analysis

To show the relationship between quantitative variables and the responder variable, the code converts the continues value to proportions.

As figure 1 shows, we can observe the following relationships:

1. relation between danceability and popularity is positive.
2. relation between energy and popularity is slightly positive.
3. relation between loudness and popularity is positive.
4. relation between duration_ms and popularity is slightly positive.
5. relation between time_signature and popularity is positive.

```
get_prop <- function(x){ return (x/ (max(x, na.rm = TRUE)-min(x, na.rm = TRU
E))   ) } #
spotify2 <- spotify%>%mutate_at(c("danceability","energy","loudness","duratio
n_ms","time_signature"), get_prop)

spotify2%>%select(popularity, danceability,energy,loudness,duration_ms,time_s
ignature) %>% gather(name, value = "data", 2:6)%>%ggplot(aes(data, popularit
y))+geom_point()+facet_wrap(~name)+geom_smooth(method="lm")#show relationship
```

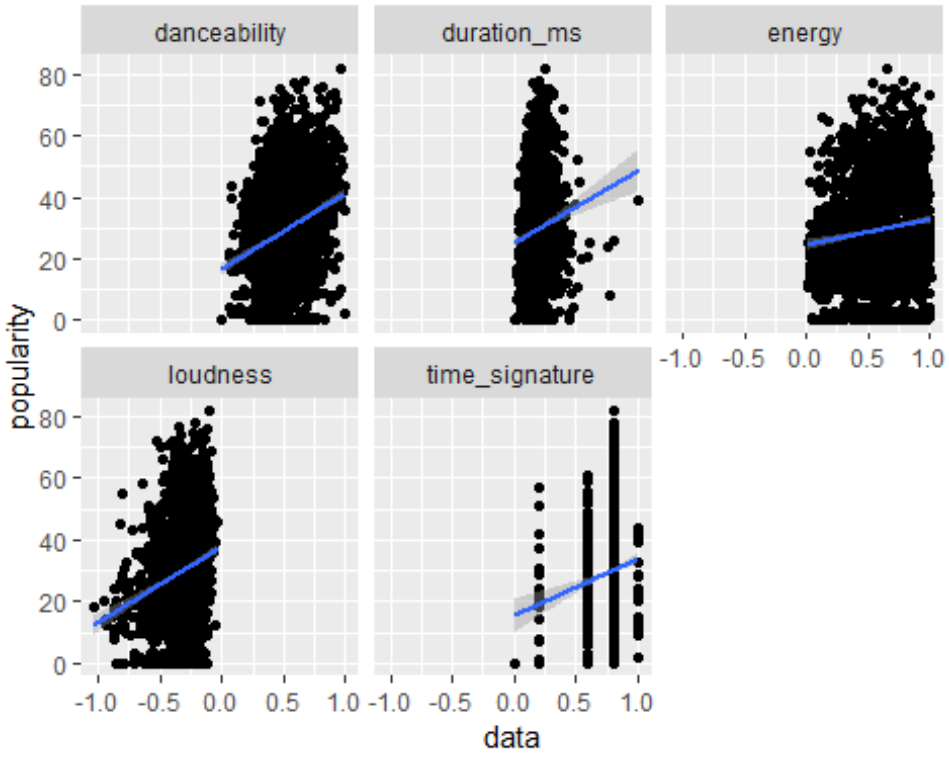

Fig 5.1 linear relationship

P-value for each predictors

As the following p-value for each predictor is very small (which means those predictors have an influence on the responding variable), each predictor seems to be considered in the initial model.

```
spotify%>%select(-popularity)%>%map(~lm(popularity~.x, data = spotify))%>%map
_df(broom::glance, .id = "predictor")%>%select(predictor, p.value)
```

```
## # A tibble: 8 x 2
##   predictor        p.value
##   <chr>              <dbl>
## 1 danceability   1.61e- 39
## 2 energy         1.03e-  9
## 3 key            7.99e-  4
## 4 loudness       5.73e- 28
## 5 mode           8.48e-  2
## 6 duration_ms    3.60e-  8
## 7 time_signature 2.48e-  7
## 8 decade         8.68e-144
```

Bivariate statistical analysis for category variables

As the following boxplots show, we could conclude that:

1.decade variable:

outliers: many potential outliers for decade 50s, 60s and 70s.

spread: the variance of 50s is small and the rest variance of other decades are not much different.

location: the median popularity of 80s, 60s and 00s are higher than the median popularity of 50s, 70s and 90s

2.key variable:

outliers: 2~3 potential outliers for key A, A#, B, C, D, F and G.

spread: the variance for each group is very similar.

location: there is not much difference in the median popularity for each key

3.mode variable

outliers: 3 potential outliers for mode major and minor.

spread: the variance for each mode is very similar.

location: not much difference in the median popularity for each mode

```
spotify%>%ggplot(aes(decade, popularity ))+stat_boxplot(geom = "errorbar")+ge
om_boxplot()
```

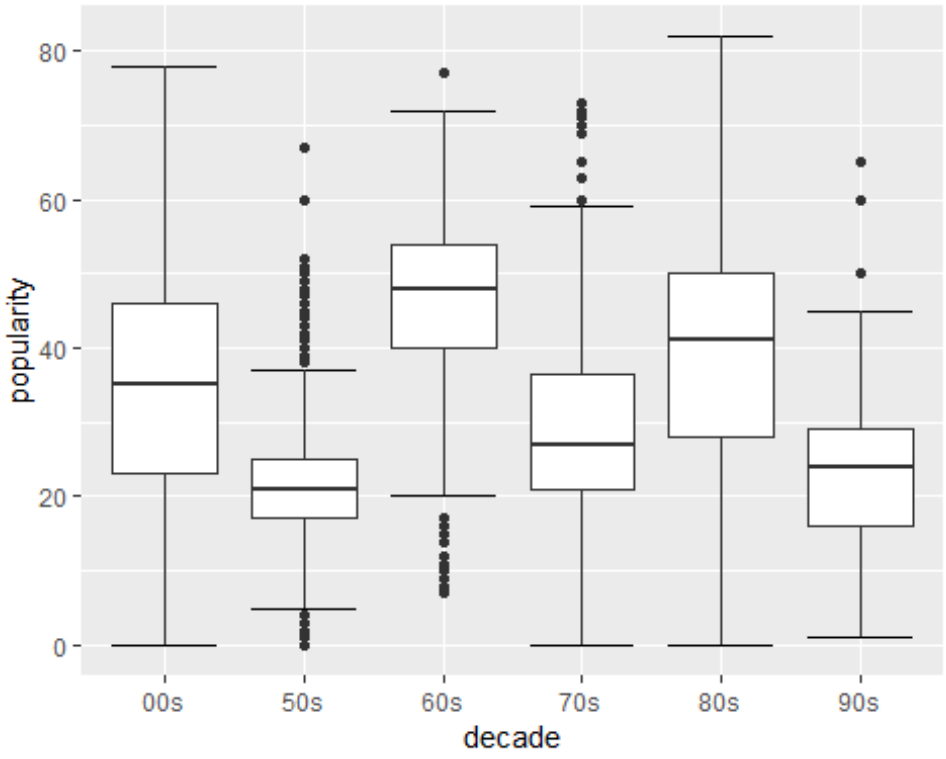

Fig 5.2 ***boxplot for category predictor : decade***

```
spotify%>%ggplot(aes(key, popularity ))+stat_boxplot(geom = "errorbar")+geom_
boxplot()
```

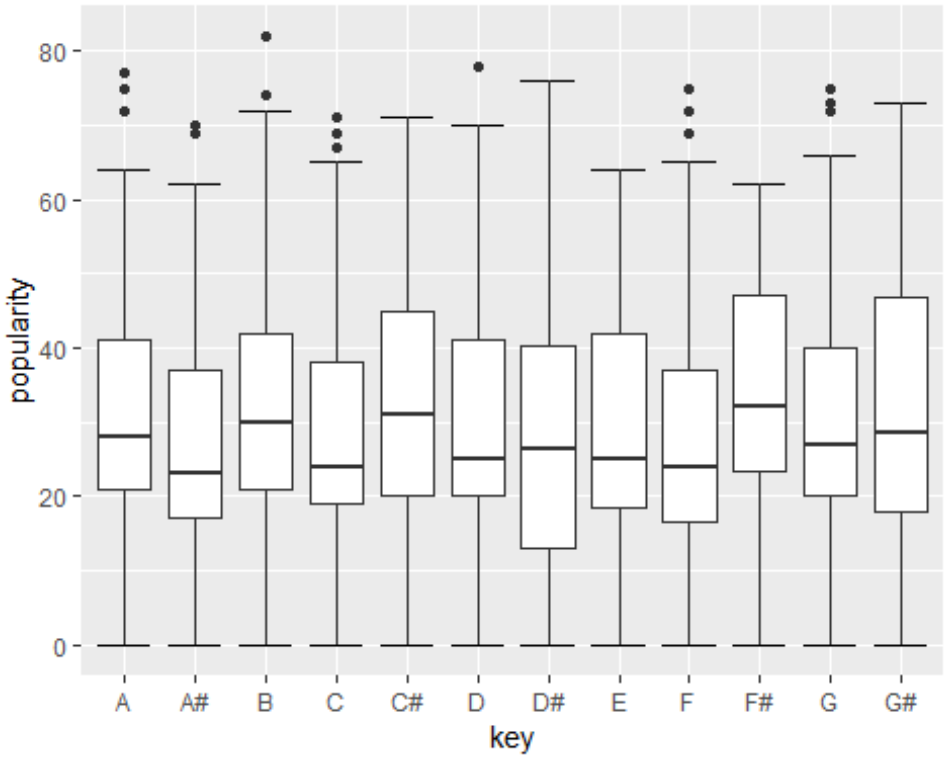

Fig 5.3 ***boxplot for category predictor : key***

```
spotify%>%ggplot(aes(mode, popularity ))+stat_boxplot(geom = "errorbar")+geom
_boxplot()
```

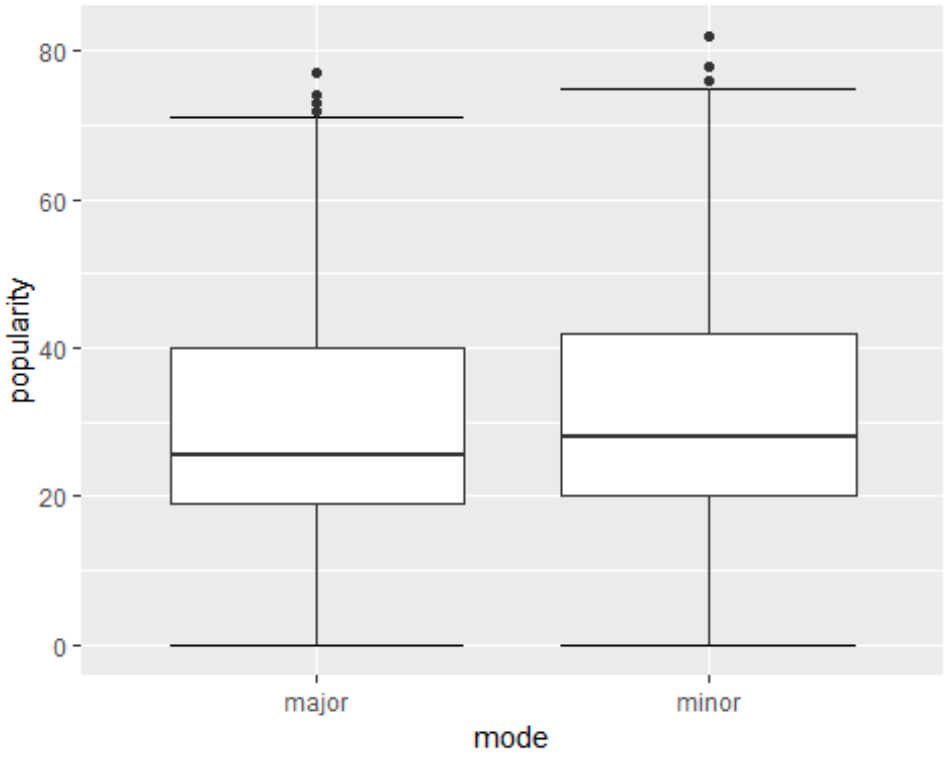

Fig 5.4 boxplot for category predictor : mode

Interaction term test for some predictors

Since time_signature, mode, key and decade have countable value, it's easier to test if our initial model should consider interaction terms among them.

As the two lines are not parallel in the following interaction plots (figure 5~ figure 10), interaction terms (includes time_signature:mode, time_signature:key, mode:key, decade:key,decade:mode and decade:time_signature) could be considered in our model.

```
spotify%>%group_by(time_signature, mode)%>%summarise(mean =mean(popularit
y))%>%ggplot(aes(time_signature, mean, col = mode))+geom_point()+geom_line(ae
s(group = mode))
```

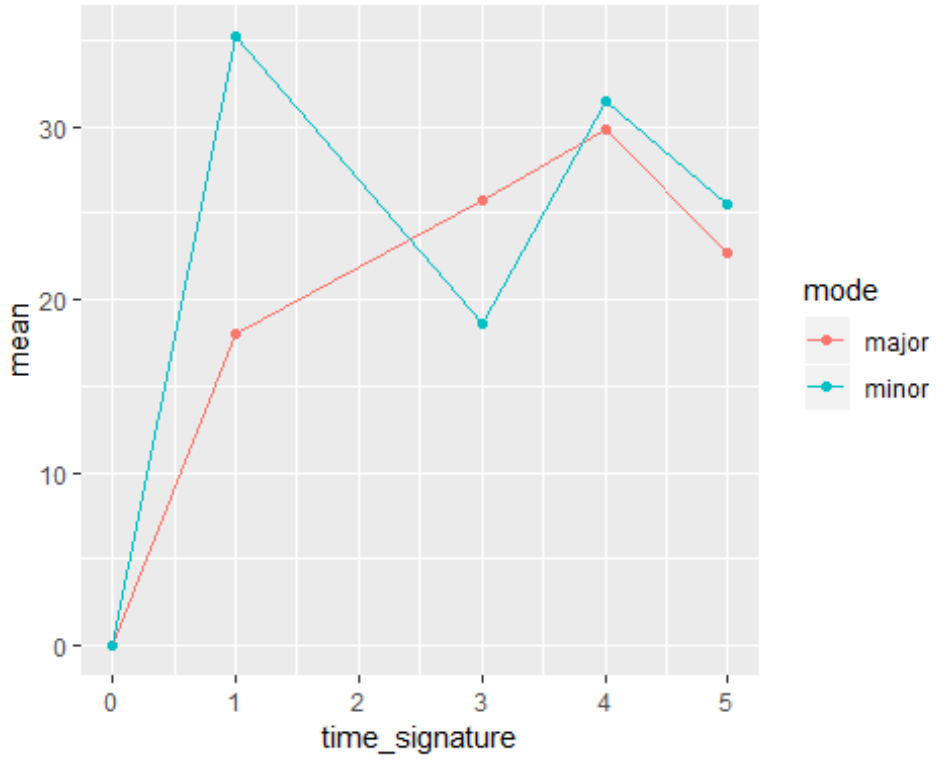

Fig 5.5 interaction term test for time_signature, mode

```
spotify%>%group_by(time_signature, key)%>%summarise(mean =mean(popularit
y))%>%ggplot(aes(key, mean, col = time_signature))+geom_point()+geom_line(aes
(group = time_signature))
```

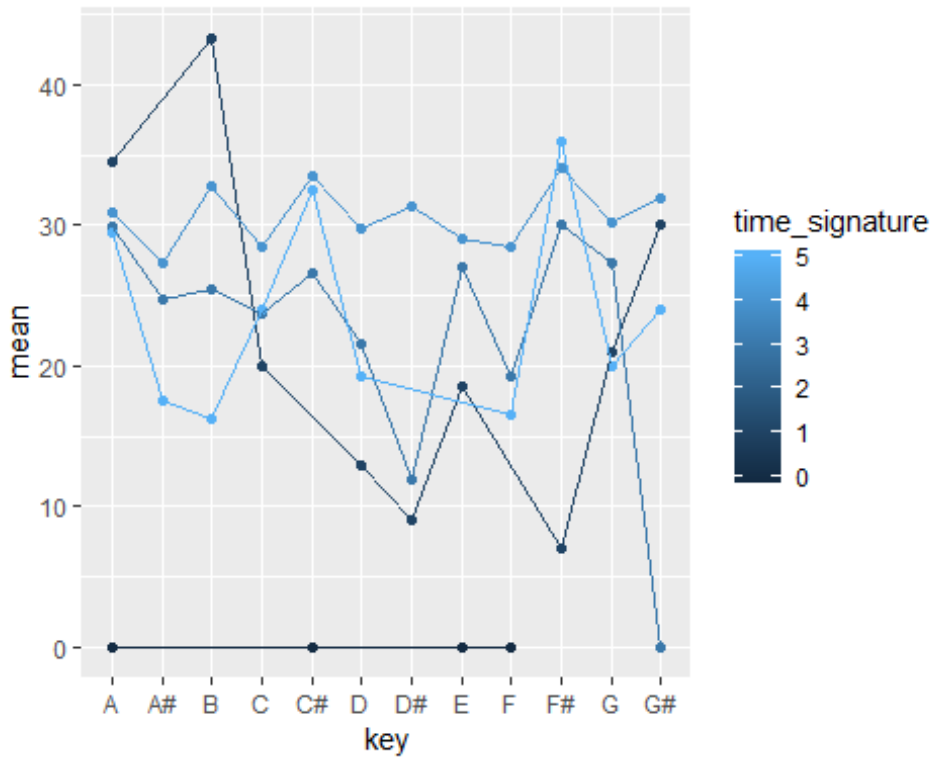

Fig 5.6 interaction term test for time_signature, key

```
spotify%>%group_by(mode, key)%>%summarise(mean =mean(popularity))%>%ggplot(ae
s(key, mean, col = mode))+geom_point()+geom_line(aes(group = mode))
```

Fig 5.7 interaction term test for mode, key

```
spotify%>%group_by(decade, key)%>%summarise(mean =mean(popularity))%>%ggplot
(aes(decade, mean, col = key))+geom_point()+geom_line(aes(group = key))
```

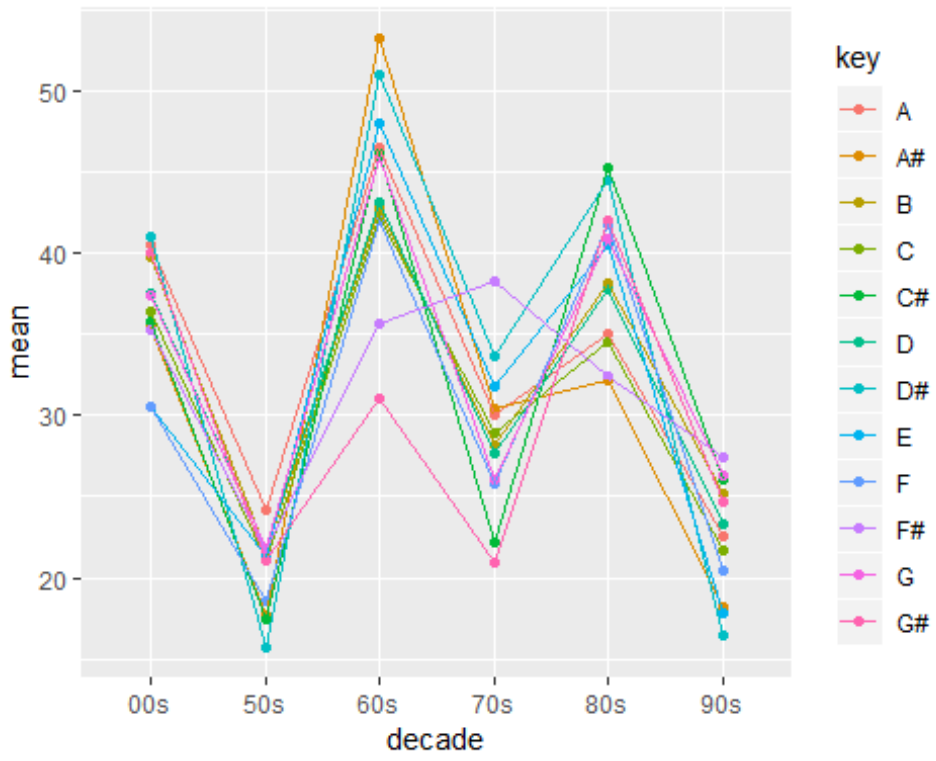

Fig 5.8 interaction term test for mode

```
spotify%>%group_by(decade, mode)%>%summarise(mean =mean(popularity))%>%ggplot
(aes(decade, mean, col = mode))+geom_point()+geom_line(aes(group = mode))
```

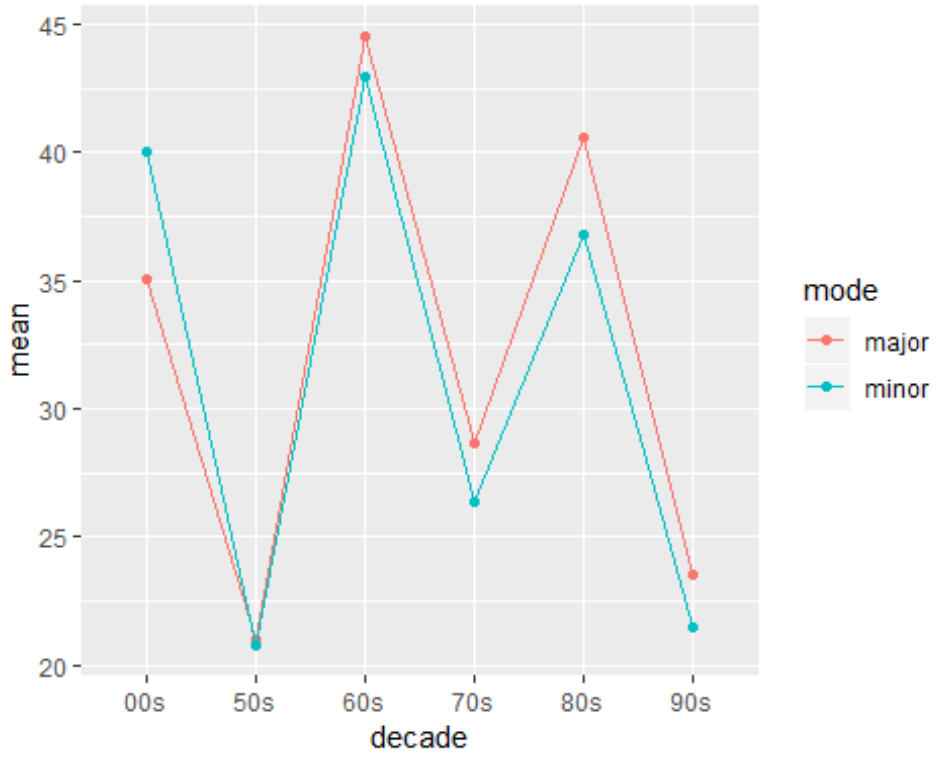

Fig 5.9 interaction term test for decade, mode

```
spotify%>%group_by(decade, time_signature)%>%summarise(mean =mean(popularit
y))%>%ggplot(aes(decade, mean, col = time_signature))+geom_point()+geom_line
(aes(group = time_signature))
```

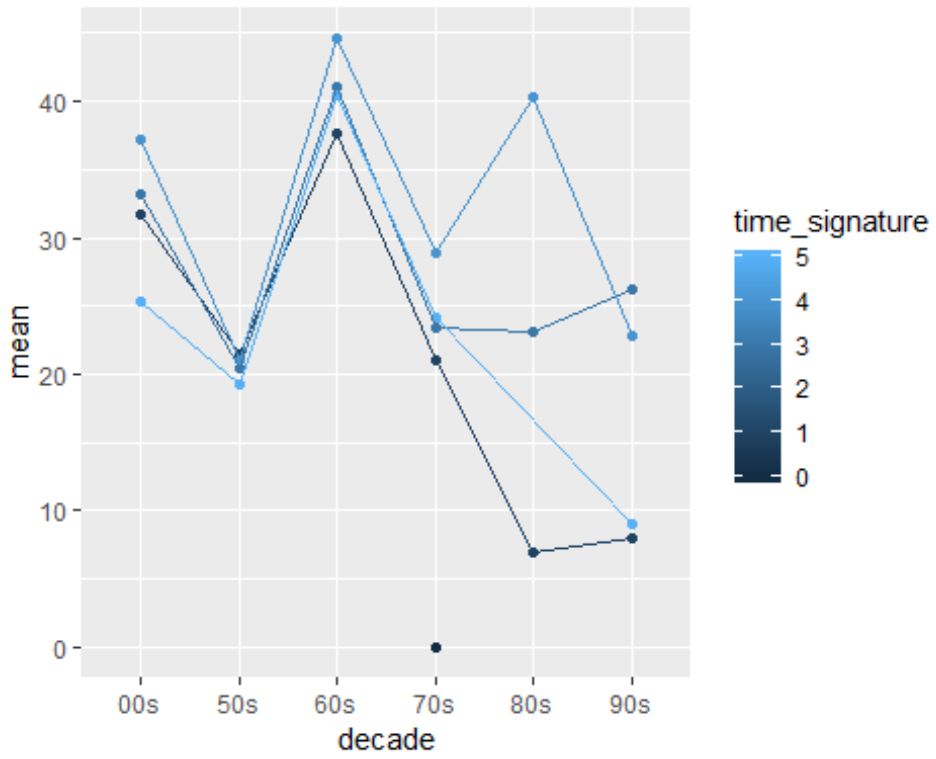

Fig 5.10 ***interaction term test for decade, time_signature***

Model Fitting

To decide the best model for our data there are two things that we need to consider which are:

1) Choice of algorithm
2) Choice of heuristic

There are three different algorithms that we can use in model fitting:

a) Forward
b) Backward
c) Stepwise

While the most common choice of heuristics are listed below:

a) Akaike information criterion (AIC)
b) Akaike information criterion corrected (AICc)
c) Bayesian information criterion (BIC)
d) Cross-validation

In this experiment we try to find a model that gives the best prediction (without assuming that any of the models are correct). So, we do this in 12 approaches, 6 of them will ignore the interaction terms and the other six approach will be set by considering possible influential interaction terms.

In the first three experiments, we will ignore the interactions term and build our model by manually using each of the three algorithm mentioned above, for this experiments we will use F-test with P-value cutoff of 0.05.

• Approach 1: Forward method using F-test with P-value cutoff of 0.05

• Approach 2: Backward method using F-test with P-value cutoff of 0.05

• Approach 3: Stepwise method using F-test with P-value cutoff of 0.05 for exclusion and 0.1 for inclusion

Then, for the next three experiments, we still ignore the interactions term and build our model by manually using each of the three algorithm mentioned above, but for this experiments instead of using F-test we will use the heuristic of AIC.

• Approach 4: Forward method using AIC

• Approach 5: Backward method using AIC

• Approach 6: Stepwise method using AIC

In the next experiments, we will consider any possible influential interaction terms. So for the three experiments, we will build our model by manually using each of the three algorithm mentioned above, for this experiments we will use F-test with P-value cutoff of 0.05.

• Approach 7: Backward method using F-test with P-value cutoff of 0.05

• Approach 8: Forward method using F-test with P-value cutoff of 0.05

• Approach 9: Stepwise method using F-test with P-value cutoff of 0.05 for exclusion and 0.1 for inclusion

This time, we still consider any possible influential interaction terms. But instead of applying the three algorithm manually, will use the automatic tools for forward, backward and stepwise.

• Approach 10: Automatic backward

• Approach 11: Automatic forward

• Approach 12: Automatic stepwise

The result for each approach mentioned above can be seen in the following sections.

Approach 1: Forward method using F-test with P-value cutoff of 0.05

```
#approach 2(forward Algorithm using p-values) - not considering interaction t
erms
null  <-lm(popularity~1, data = spotify)
scope  <- popularity ~ danceability + energy + key + loudness +mode + duratio
n_ms + time_signature + decade
add1(null, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ 1
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                      516998 11479
## danceability    1     41296 475702 11308 180.4805 < 2.2e-16 ***
## energy          1      9188 507810 11444  37.6173 1.027e-09 ***
```

```
## key             11      7887 509111 11469   2.9138 0.0007986 ***
## loudness         1     29049 487949 11361 123.7701 < 2.2e-16 ***
## mode             1       738 516260 11478   2.9732 0.0848020 .
## duration_ms      1      7496 509503 11451  30.5853 3.597e-08 ***
## time_signature   1      6578 510420 11454  26.7946 2.481e-07 ***
## decade           5    143611 373387 10812 159.6161 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
approach_1  <-update(null, .~.+danceability) #adding danceability
add1(approach_1, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                      475702 11308
## energy          1      4808 470894 11289  21.2184 4.346e-06 ***
## key            11      5097 470605 11308   2.0361  0.021974 *
## loudness        1     17373 458329 11232  78.7673 < 2.2e-16 ***
## mode            1         8 475694 11310   0.0370  0.847566
## duration_ms     1      6721 468981 11280  29.7783 5.422e-08 ***
## time_signature  1      1609 474093 11303   7.0529  0.007974 **
## decade          5    112763 362939 10755 128.8756 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
approach_1  <-update(approach_1, .~.+decade)#adding decade
add1(approach_1, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade
##                Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                      362939 10755
## energy          1    1791.9 361147 10746 10.2855  0.001361 **
## key            11    2087.2 360852 10765  1.0848  0.369375
## loudness        1    4512.0 358427 10731 26.0953 3.547e-07 ***
## mode            1     922.0 362017 10752  5.2797  0.021675 *
## duration_ms     1    6403.7 356536 10720 37.2327 1.248e-09 ***
## time_signature  1    1148.2 361791 10750  6.5788  0.010390 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
approach_1  <-update(approach_1, .~.+duration_ms)#adding duration
add1(approach_1, scope = scope, test = "F")
```

```
## Single term additions
##
```

```
## Model:
## popularity ~ danceability + decade + duration_ms
##                Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                      356536 10720
## energy          1    830.72 355705 10717  4.8390 0.0279330 *
## key            11   1906.79 354629 10731  1.0079 0.4364619
## loudness        1   2533.52 354002 10707 14.8289 0.0001213 ***
## mode            1    885.13 355650 10717  5.1567 0.0232590 *
## time_signature  1    510.80 356025 10719  2.9728 0.0848249 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

approach_1  <-update(approach_1, .~.+loudness)#adding loudness
add1(approach_1, scope = scope, test = "F")

## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness
##                Df Sum of Sq    RSS   AIC F value  Pr(>F)
## <none>                      354002 10707
## energy          1    220.58 353781 10708  1.2913 0.25595
## key            11   1685.47 352317 10719  0.8963 0.54323
## mode            1    859.35 353143 10704  5.0397 0.02488 *
## time_signature  1    259.52 353743 10707  1.5194 0.21786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

approach_1  <-update(approach_1, .~.+mode)#adding mode
add1(approach_1, scope = scope, test = "F")

## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness +
##     mode
##                Df Sum of Sq    RSS   AIC F value Pr(>F)
## <none>                      353143 10704
## energy          1    163.72 352979 10705  0.9601 0.3273
## key            11   1659.28 351483 10716  0.8841 0.5555
## time_signature  1    231.36 352911 10704  1.3570 0.2442

#all p values are now >0.05 so we stop adding predictors
summary(approach_1)

##
## Call:
## lm(formula = popularity ~ danceability + decade + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 2: Backward method using F-test with P-value cutoff of 0.05.

```
summary(approach_2)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + loudness + duration_ms +
##     decade, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.827  -7.127   0.254   7.774  44.522
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.653e+01  1.845e+00  14.383  < 2e-16 ***
## danceability  1.362e+01  1.845e+00   7.378 2.31e-13 ***
## loudness      2.785e-01  7.233e-02   3.851 0.000121 ***
## duration_ms   1.728e-05  3.395e-06   5.089 3.92e-07 ***
## decade50s    -1.147e+01  1.149e+00  -9.981  < 2e-16 ***
## decade60s     1.066e+01  1.296e+00   8.225 3.41e-16 ***
## decade70s    -6.119e+00  1.130e+00  -5.415 6.84e-08 ***
## decade80s     1.274e+00  1.271e+00   1.002 0.316583
```

```
## decade90s    -1.197e+01  1.356e+00  -8.827  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.07 on 2072 degrees of freedom
## Multiple R-squared:  0.3153, Adjusted R-squared:  0.3126
## F-statistic: 119.3 on 8 and 2072 DF,  p-value: < 2.2e-16
```

Approach 3: Stepwise method using F-test with P-value cutoff of 0.05 for exclusion and 0.1 for Inclusion

```
#approach 3(stepwise selection procedure with p-values) - not considering int
eraction terms
add1(null, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ 1
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                      516998 11479
## danceability    1     41296 475702 11308 180.4805 < 2.2e-16 ***
## energy          1      9188 507810 11444  37.6173 1.027e-09 ***
## key            11      7887 509111 11469   2.9138 0.0007986 ***
## loudness        1     29049 487949 11361 123.7701 < 2.2e-16 ***
## mode            1       738 516260 11478   2.9732 0.0848020 .
## duration_ms     1      7496 509503 11451  30.5853 3.597e-08 ***
## time_signature  1      6578 510420 11454  26.7946 2.481e-07 ***
## decade          5    143611 373387 10812 159.6161 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
approach_3  <-update(null, . ~.+danceability)
drop1(approach_3, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                    475702 11308
## danceability  1     41296 516998 11479  180.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(approach_3, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
```

```
## popularity ~ danceability
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                      475702 11308
## energy          1      4808 470894 11289  21.2184 4.346e-06 ***
## key            11      5097 470605 11308   2.0361  0.021974 *
## loudness        1     17373 458329 11232  78.7673 < 2.2e-16 ***
## mode            1         8 475694 11310   0.0370  0.847566
## duration_ms     1      6721 468981 11280  29.7783 5.422e-08 ***
## time_signature  1      1609 474093 11303   7.0529  0.007974 **
## decade          5    112763 362939 10755 128.8756 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

approach_3  <-update(approach_3, . ~.+decade)
drop1(approach_3, test = "F")

## Single term deletions
##
## Model:
## popularity ~ danceability + decade
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                    362939 10755
## danceability  1     10448 373387 10812  59.704 1.704e-14 ***
## decade        5    112763 475702 11308 128.876 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

add1(approach_3, scope = scope, test = "F")

## Single term additions
##
## Model:
## popularity ~ danceability + decade
##                Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                      362939 10755
## energy          1    1791.9 361147 10746 10.2855  0.001361 **
## key            11    2087.2 360852 10765  1.0848  0.369375
## loudness        1    4512.0 358427 10731 26.0953 3.547e-07 ***
## mode            1     922.0 362017 10752  5.2797  0.021675 *
## duration_ms     1    6403.7 356536 10720 37.2327 1.248e-09 ***
## time_signature  1    1148.2 361791 10750  6.5788  0.010390 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

approach_3  <-update(approach_3, . ~.+duration_ms)
drop1(approach_3, test = "F")

## Single term deletions
##
## Model:
## popularity ~ danceability + decade + duration_ms
```

```
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                     356536 10720
## danceability  1     10958 367494 10781  63.714 2.359e-15 ***
## decade        5    112446 468981 11280 130.758 < 2.2e-16 ***
## duration_ms   1      6404 362939 10755  37.233 1.248e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(approach_3, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms
##                Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                      356536 10720
## energy          1    830.72 355705 10717  4.8390 0.0279330 *
## key            11   1906.79 354629 10731  1.0079 0.4364619
## loudness        1   2533.52 354002 10707 14.8289 0.0001213 ***
## mode            1    885.13 355650 10717  5.1567 0.0232590 *
## time_signature  1    510.80 356025 10719  2.9728 0.0848249 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
approach_3  <-update(approach_3, . ~.+loudness)
drop1(approach_3, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                     354002 10707
## danceability  1      9301 363303 10759  54.439 2.308e-13 ***
## decade        5    102269 456271 11225 119.718 < 2.2e-16 ***
## duration_ms   1      4425 358427 10731  25.901 3.918e-07 ***
## loudness      1      2534 356536 10720  14.829 0.0001213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(approach_3, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness
##                Df Sum of Sq    RSS   AIC F value  Pr(>F)
## <none>                      354002 10707
## energy          1    220.58 353781 10708  1.2913 0.25595
## key            11   1685.47 352317 10719  0.8963 0.54323
## mode            1    859.35 353143 10704  5.0397 0.02488 *
```

```
## time_signature  1    259.52 353743 10707  1.5194 0.21786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
approach_3  <-update(approach_3, . ~.+mode)
drop1(approach_3, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness +
##     mode
##              Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                    353143 10704
## danceability  1      9665 362807 10758  56.6777 7.615e-14 ***
## decade        5    103071 456214 11227 120.8915 < 2.2e-16 ***
## duration_ms   1      4404 357546 10728  25.8250 4.074e-07 ***
## loudness      1      2508 355650 10717  14.7066 0.0001294 ***
## mode          1       859 354002 10707   5.0397 0.0248784 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(approach_3, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness +
##     mode
##                Df Sum of Sq    RSS   AIC F value Pr(>F)
## <none>                      353143 10704
## energy          1    163.72 352979 10705  0.9601 0.3273
## key            11   1659.28 351483 10716  0.8841 0.5555
## time_signature  1    231.36 352911 10704  1.3570 0.2442
```

```
summary(approach_3)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + decade + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
```

```
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 4: Forward method using AIC

```
#approach 4(forwards algorithm using AIC values) - not considering interactio
n terms
approach_4  <-step(null,scope = scope, direction = "forward")
```

```
## Start:  AIC=11479.11
## popularity ~ 1
##
##                  Df Sum of Sq    RSS   AIC
## + decade          5    143611 373387 10812
## + danceability    1     41296 475702 11308
## + loudness        1     29049 487949 11361
## + energy          1      9188 507810 11444
## + duration_ms     1      7496 509503 11451
## + time_signature  1      6578 510420 11454
## + key            11      7887 509111 11469
## + mode            1       738 516260 11478
## <none>                        516998 11479
##
## Step:  AIC=10811.91
## popularity ~ decade
##
##                  Df Sum of Sq    RSS   AIC
## + danceability    1   10448.0 362939 10755
## + loudness        1    6411.7 366975 10778
## + duration_ms     1    5893.4 367494 10781
## + time_signature  1    2924.5 370463 10798
## + energy          1    2914.4 370473 10798
## + mode            1     525.3 372862 10811
## <none>                        373387 10812
## + key            11    1987.0 371400 10823
##
## Step:  AIC=10754.85
```

```
## popularity ~ decade + danceability
##
##                   Df Sum of Sq    RSS   AIC
## + duration_ms      1    6403.7 356536 10720
## + loudness         1    4512.0 358427 10731
## + energy           1    1791.9 361147 10746
## + time_signature   1    1148.2 361791 10750
## + mode             1     922.0 362017 10752
## <none>                         362939 10755
## + key             11    2087.2 360852 10765
##
## Step:  AIC=10719.8
## popularity ~ decade + danceability + duration_ms
##
##                   Df Sum of Sq    RSS   AIC
## + loudness         1   2533.52 354002 10707
## + mode             1    885.13 355650 10717
## + energy           1    830.72 355705 10717
## + time_signature   1    510.80 356025 10719
## <none>                         356536 10720
## + key             11   1906.79 354629 10731
##
## Step:  AIC=10706.96
## popularity ~ decade + danceability + duration_ms + loudness
##
##                   Df Sum of Sq    RSS   AIC
## + mode             1    859.35 353143 10704
## <none>                         354002 10707
## + time_signature   1    259.52 353743 10707
## + energy           1    220.58 353781 10708
## + key             11   1685.47 352317 10719
##
## Step:  AIC=10703.9
## popularity ~ decade + danceability + duration_ms + loudness +
##     mode
##
##                   Df Sum of Sq    RSS   AIC
## <none>                         353143 10704
## + time_signature   1    231.36 352911 10704
## + energy           1    163.72 352979 10705
## + key             11   1659.28 351483 10716
```

```
summary(approach_4)
```

```
##
## Call:
## lm(formula = popularity ~ decade + danceability + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 5: Backward method using AIC.

```
#approach 5(backwards algorithm using AIC values) - not considering interacti
on terms
approach_5  <-step(full,scope = scope, direction = "backward")
```

```
## Start:  AIC=10716.34
## popularity ~ danceability + energy + key + loudness + mode +
##     duration_ms + time_signature + decade
##
##                  Df Sum of Sq    RSS   AIC
## - key            11      1866 352715 10705
## - time_signature  1       327 351176 10716
## <none>                        350849 10716
## - energy          1       357 351206 10716
## - mode            1       734 351582 10719
## - loudness        1      1894 352742 10726
## - duration_ms     1      3940 354789 10738
## - danceability    1      8827 359676 10766
## - decade          5     97974 448822 11219
##
## Step:  AIC=10705.38
## popularity ~ danceability + energy + loudness + mode + duration_ms +
##     time_signature + decade
##
##                  Df Sum of Sq    RSS   AIC
## - energy          1       197 352911 10704
```

```
## - time_signature  1       264 352979 10705
## <none>                        352715 10705
## - mode            1       768 353483 10708
## - loudness        1      1734 354449 10714
## - duration_ms     1      4109 356824 10728
## - danceability    1      8744 361458 10754
## - decade          5    100088 452803 11215
##
## Step:  AIC=10704.54
## popularity ~ danceability + loudness + mode + duration_ms + time_signature
 +
##     decade
##
##                  Df Sum of Sq    RSS   AIC
## - time_signature  1       231 353143 10704
## <none>                        352911 10704
## - mode            1       831 353743 10707
## - loudness        1      2270 355181 10716
## - duration_ms     1      4127 357038 10727
## - danceability    1      8738 361649 10753
## - decade          5    103140 456051 11228
##
## Step:  AIC=10703.9
## popularity ~ danceability + loudness + mode + duration_ms + decade
##
##                Df Sum of Sq    RSS   AIC
## <none>                      353143 10704
## - mode          1       859 354002 10707
## - loudness      1      2508 355650 10717
## - duration_ms   1      4404 357546 10728
## - danceability  1      9665 362807 10758
## - decade        5    103071 456214 11227
```

```
summary(approach_5)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + loudness + mode + duration_ms +
##     decade, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
```

```
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 6: Stepwise method using AIC

```
#approach 6(stepwise using AIC values) - not considering interaction terms
approach_6  <-step(null,scope = scope, direction = "both")
```

```
## Start:  AIC=11479.11
## popularity ~ 1
##
##                  Df Sum of Sq    RSS   AIC
## + decade          5    143611 373387 10812
## + danceability    1     41296 475702 11308
## + loudness        1     29049 487949 11361
## + energy          1      9188 507810 11444
## + duration_ms     1      7496 509503 11451
## + time_signature  1      6578 510420 11454
## + key            11      7887 509111 11469
## + mode            1       738 516260 11478
## <none>                        516998 11479
##
## Step:  AIC=10811.91
## popularity ~ decade
##
##                  Df Sum of Sq    RSS   AIC
## + danceability    1     10448 362939 10755
## + loudness        1      6412 366975 10778
## + duration_ms     1      5893 367494 10781
## + time_signature  1      2925 370463 10798
## + energy          1      2914 370473 10798
## + mode            1       525 372862 10811
## <none>                        373387 10812
## + key            11      1987 371400 10823
## - decade          5    143611 516998 11479
##
## Step:  AIC=10754.85
## popularity ~ decade + danceability
##
```

```
##                   Df Sum of Sq    RSS   AIC
## + duration_ms      1      6404 356536 10720
## + loudness         1      4512 358427 10731
## + energy           1      1792 361147 10746
## + time_signature   1      1148 361791 10750
## + mode             1       922 362017 10752
## <none>                         362939 10755
## + key             11      2087 360852 10765
## - danceability     1     10448 373387 10812
## - decade           5    112763 475702 11308
##
## Step:  AIC=10719.8
## popularity ~ decade + danceability + duration_ms
##
##                   Df Sum of Sq    RSS   AIC
## + loudness         1      2534 354002 10707
## + mode             1       885 355650 10717
## + energy           1       831 355705 10717
## + time_signature   1       511 356025 10719
## <none>                         356536 10720
## + key             11      1907 354629 10731
## - duration_ms      1      6404 362939 10755
## - danceability     1     10958 367494 10781
## - decade           5    112446 468981 11280
##
## Step:  AIC=10706.96
## popularity ~ decade + danceability + duration_ms + loudness
##
##                   Df Sum of Sq    RSS   AIC
## + mode             1       859 353143 10704
## <none>                         354002 10707
## + time_signature   1       260 353743 10707
## + energy           1       221 353781 10708
## + key             11      1685 352317 10719
## - loudness         1      2534 356536 10720
## - duration_ms      1      4425 358427 10731
## - danceability     1      9301 363303 10759
## - decade           5    102269 456271 11225
##
## Step:  AIC=10703.9
## popularity ~ decade + danceability + duration_ms + loudness +
##     mode
##
##                   Df Sum of Sq    RSS   AIC
## <none>                         353143 10704
## + time_signature   1       231 352911 10704
## + energy           1       164 352979 10705
## - mode             1       859 354002 10707
## + key             11      1659 351483 10716
## - loudness         1      2508 355650 10717
```

```
## - duration_ms     1      4404 357546 10728
## - danceability    1      9665 362807 10758
## - decade          5    103071 456214 11227
```

```
summary(approach_6)
```

```
##
## Call:
## lm(formula = popularity ~ decade + danceability + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 7: Backward method using F-test with P-value cutoff of 0.05

```
#backwards algorithm - considering interaction terms
full  <- popularity ~ danceability + energy + time_signature + loudness + key
 + mode + duration_ms + decade +
  decade:time_signature + mode:key + time_signature:key + decade:mode + time_
signature:mode +  decade:key
backwards.lm <- lm(full, data = spotify)
drop1(backwards.lm, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + energy + time_signature + loudness +
##     key + mode + duration_ms + decade + decade:time_signature +
```

```
##     mode:key + time_signature:key + decade:mode + time_signature:mode +
##     decade:key
##                       Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                             326377 10742
## danceability           1    8366.2 334743 10793 50.4978 1.661e-12 ***
## energy                 1     606.7 326984 10744  3.6623 0.0558009 .
## loudness               1    1474.5 327852 10749  8.8998 0.0028870 **
## duration_ms            1    2431.9 328809 10755 14.6791 0.0001314 ***
## time_signature:decade  5    3655.3 330033 10755  4.4127 0.0005334 ***
## key:mode              11    4002.7 330380 10745  2.1964 0.0124493 *
## time_signature:key    11    1924.8 328302 10732  1.0562 0.3936256
## mode:decade            5    2003.1 328380 10745  2.4181 0.0339368 *
## time_signature:mode    1     329.7 326707 10742  1.9903 0.1584718
## key:decade            55   12274.1 338651 10709  1.3470 0.0469260 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
backwards.lm <- update(backwards.lm, .~. - time_signature:key)#removing time_
signature:key
drop1(backwards.lm, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + energy + time_signature + loudness +
##     key + mode + duration_ms + decade + time_signature:decade +
##     key:mode + mode:decade + time_signature:mode + key:decade
##                       Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                             328302 10732
## danceability           1    8584.7 336887 10784 51.8009 8.679e-13 ***
## energy                 1     555.7 328858 10734  3.3528 0.0672393 .
## loudness               1    1394.8 329697 10739  8.4165 0.0037594 **
## duration_ms            1    2446.8 330749 10746 14.7641 0.0001257 ***
## time_signature:decade  5    2803.2 331105 10740  3.3829 0.0047824 **
## key:mode              11    3837.8 332140 10734  2.1052 0.0172477 *
## mode:decade            5    2159.9 330462 10736  2.6066 0.0233770 *
## time_signature:mode    1      50.0 328352 10730  0.3017 0.5828713
## key:decade            55   12223.2 340525 10698  1.3410 0.0494491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
backwards.lm <- update(backwards.lm, .~. - time_signature:mode)#removing time
_signature:mode
drop1(backwards.lm, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + energy + time_signature + loudness +
##     key + mode + duration_ms + decade + time_signature:decade +
##     key:mode + mode:decade + key:decade
```

```
##                         Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                               328352 10730
## danceability             1    8579.6 336932 10782 51.7883 8.732e-13 ***
## energy                   1     568.9 328921 10732  3.4339 0.0640204 .
## loudness                 1    1436.9 329789 10738  8.6732 0.0032669 **
## duration_ms              1    2470.0 330822 10744 14.9096 0.0001164 ***
## time_signature:decade    5    2948.8 331301 10739  3.5599 0.0032997 **
## key:mode                11    3843.0 332195 10733  2.1088 0.0170279 *
## mode:decade              5    2120.6 330473 10734  2.5600 0.0256465 *
## key:decade              55   12230.6 340583 10696  1.3423 0.0488985 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

backwards.lm <- update(backwards.lm, .~. - energy)#removing energy
drop1(backwards.lm, test = "F")

## Single term deletions
##
## Model:
## popularity ~ danceability + time_signature + loudness + key +
##     mode + duration_ms + decade + time_signature:decade + key:mode +
##     mode:decade + key:decade
##                         Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                               328921 10732
## danceability             1    8574.4 337495 10784 51.6933 9.151e-13 ***
## loudness                 1     926.6 329847 10736  5.5862 0.0181988 *
## duration_ms              1    2462.1 331383 10746 14.8438 0.0001205 ***
## time_signature:decade    5    3217.0 332138 10742  3.8789 0.0016791 **
## key:mode                11    3747.8 332669 10734  2.0541 0.0206526 *
## mode:decade              5    2018.7 330940 10735  2.4340 0.0328920 *
## key:decade              55   12046.3 340967 10697  1.3205 0.0590558 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

backwards.lm <- update(backwards.lm, .~. - key:decade)#removing key:decade
drop1(backwards.lm, test = "F")

## Single term deletions
##
## Model:
## popularity ~ danceability + time_signature + loudness + key +
##     mode + duration_ms + decade + time_signature:decade + key:mode +
##     mode:decade
##                         Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                               340967 10697
## danceability             1    8689.5 349657 10747 51.9381 8.030e-13 ***
## loudness                 1    1363.5 342331 10703  8.1496 0.0043506 **
## duration_ms              1    2610.2 343577 10711 15.6012 8.086e-05 ***
## time_signature:decade    5    4080.6 345048 10712  4.8780 0.0001933 ***
## key:mode                11    4429.4 345397 10702  2.4068 0.0057296 **
## mode:decade              5    2203.3 343170 10700  2.6339 0.0221319 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(backwards.lm)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + time_signature + loudness +
##     key + mode + duration_ms + decade + time_signature:decade +
##     key:mode + mode:decade, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.544  -7.207   0.195   7.686  44.550
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.587e+01  7.809e+00   3.312 0.000941 ***
## danceability                1.367e+01  1.898e+00   7.207 8.03e-13 ***
## time_signature              8.520e-01  1.941e+00   0.439 0.660710
## loudness                    2.110e-01  7.392e-02   2.855 0.004351 **
## keyA#                      -3.739e+00  1.732e+00  -2.159 0.030981 *
## keyB                       -1.604e+00  1.984e+00  -0.808 0.419018
## keyC                       -2.157e+00  1.167e+00  -1.848 0.064677 .
## keyC#                      -4.156e+00  1.599e+00  -2.600 0.009392 **
## keyD                       -2.454e+00  1.273e+00  -1.928 0.053970 .
## keyD#                      -5.192e+00  2.490e+00  -2.085 0.037173 *
## keyE                       -2.443e+00  1.437e+00  -1.700 0.089348 .
## keyF                       -4.340e+00  1.486e+00  -2.920 0.003537 **
## keyF#                      -5.524e+00  2.555e+00  -2.162 0.030718 *
## keyG                       -2.030e+00  1.285e+00  -1.580 0.114230
## keyG#                      -3.814e+00  1.843e+00  -2.070 0.038579 *
## modeminor                   1.239e+00  2.744e+00   0.452 0.651606
## duration_ms                 1.366e-05  3.459e-06   3.950 8.09e-05 ***
## decade50s                  -4.077e+00  8.277e+00  -0.493 0.622366
## decade60s                   8.248e+00  1.063e+01   0.776 0.437978
## decade70s                  -1.568e+01  8.905e+00  -1.760 0.078481 .
## decade80s                  -3.678e+01  1.339e+01  -2.748 0.006055 **
## decade90s                   1.993e+00  1.375e+01   0.145 0.884761
## time_signature:decade50s   -2.015e+00  2.117e+00  -0.952 0.341438
## time_signature:decade60s    5.488e-01  2.703e+00   0.203 0.839113
## time_signature:decade70s    2.561e+00  2.268e+00   1.129 0.258905
## time_signature:decade80s    1.021e+01  3.401e+00   3.001 0.002728 **
## time_signature:decade90s   -3.328e+00  3.497e+00  -0.952 0.341316
## keyA#:modeminor            -4.185e-03  3.202e+00  -0.001 0.998957
## keyB:modeminor              2.369e+00  2.877e+00   0.823 0.410363
## keyC:modeminor             -6.377e+00  3.699e+00  -1.724 0.084872 .
## keyC#:modeminor             8.696e+00  3.131e+00   2.778 0.005524 **
## keyD:modeminor              1.488e+00  2.902e+00   0.513 0.608315
## keyD#:modeminor             1.106e+01  4.791e+00   2.309 0.021066 *
```

```
## keyE:modeminor             5.898e+00  2.729e+00   2.161 0.030796 *
## keyF:modeminor             6.036e+00  3.242e+00   1.862 0.062778 .
## keyF#:modeminor            7.796e+00  3.661e+00   2.129 0.033348 *
## keyG:modeminor             2.883e+00  3.151e+00   0.915 0.360435
## keyG#:modeminor            3.600e+00  4.473e+00   0.805 0.420967
## modeminor:decade50s       -4.714e+00  2.762e+00  -1.707 0.087959 .
## modeminor:decade60s       -6.192e+00  2.988e+00  -2.072 0.038351 *
## modeminor:decade70s       -6.533e+00  2.511e+00  -2.601 0.009350 **
## modeminor:decade80s       -9.433e+00  2.729e+00  -3.456 0.000559 ***
## modeminor:decade90s       -7.410e+00  2.966e+00  -2.498 0.012555 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.93 on 2038 degrees of freedom
## Multiple R-squared:  0.3405, Adjusted R-squared:  0.3269
## F-statistic: 25.05 on 42 and 2038 DF,  p-value: < 2.2e-16
```

Approach 8: Forward method using F-test with P-value cutoff of 0.05

```
#forwards algorithm - considering interaction terms
smallest <- lm(popularity~1,data=spotify)
forwards.lm <-lm(smallest, data = spotify)
add1(forwards.lm, scope = full, test = "F")

## Single term additions
##
## Model:
## popularity ~ 1
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                      516998 11479
## danceability    1     41296 475702 11308 180.4805 < 2.2e-16 ***
## energy          1      9188 507810 11444  37.6173 1.027e-09 ***
## time_signature  1      6578 510420 11454  26.7946 2.481e-07 ***
## loudness        1     29049 487949 11361 123.7701 < 2.2e-16 ***
## key            11      7887 509111 11469   2.9138 0.0007986 ***
## mode            1       738 516260 11478   2.9732 0.0848020 .
## duration_ms     1      7496 509503 11451  30.5853 3.597e-08 ***
## decade          5    143611 373387 10812 159.6161 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

forwards.lm <-update(smallest, .~.+danceability)#adding danceability
add1(forwards.lm, scope = full, test = "F")

## Single term additions
##
## Model:
## popularity ~ danceability
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
```

```
## <none>                      475702 11308
## energy           1      4808 470894 11289  21.2184 4.346e-06 ***
## time_signature   1      1609 474093 11303   7.0529  0.007974 **
## loudness         1     17373 458329 11232  78.7673 < 2.2e-16 ***
## key             11      5097 470605 11308   2.0361  0.021974 *
## mode             1         8 475694 11310   0.0370  0.847566
## duration_ms      1      6721 468981 11280  29.7783 5.422e-08 ***
## decade           5    112763 362939 10755 128.8756 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forwards.lm <-update(forwards.lm, .~.+decade)#adding decade
add1(forwards.lm, scope = full, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade
##                 Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                       362939 10755
## energy           1    1791.9 361147 10746 10.2855  0.001361 **
## time_signature   1    1148.2 361791 10750  6.5788  0.010390 *
## loudness         1    4512.0 358427 10731 26.0953 3.547e-07 ***
## key             11    2087.2 360852 10765  1.0848  0.369375
## mode             1     922.0 362017 10752  5.2797  0.021675 *
## duration_ms      1    6403.7 356536 10720 37.2327 1.248e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forwards.lm <-update(forwards.lm, .~.+duration_ms)#adding duration
add1(forwards.lm, scope = full, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms
##                 Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                       356536 10720
## energy           1     830.72 355705 10717  4.8390 0.0279330 *
## time_signature   1     510.80 356025 10719  2.9728 0.0848249 .
## loudness         1    2533.52 354002 10707 14.8289 0.0001213 ***
## key             11    1906.79 354629 10731  1.0079 0.4364619
## mode             1     885.13 355650 10717  5.1567 0.0232590 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forwards.lm <-update(forwards.lm, .~.+loudness)#adding loudness
add1(forwards.lm, scope = full, test = "F")
```

```
## Single term additions
##
```

```
## Model:
## popularity ~ danceability + decade + duration_ms + loudness
##                Df Sum of Sq    RSS   AIC F value  Pr(>F)
## <none>                      354002 10707
## energy          1    220.58 353781 10708  1.2913 0.25595
## time_signature  1    259.52 353743 10707  1.5194 0.21786
## key            11   1685.47 352317 10719  0.8963 0.54323
## mode            1    859.35 353143 10704  5.0397 0.02488 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forwards.lm <-update(forwards.lm, .~.+mode)#adding mode
add1(forwards.lm, scope = full, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness +
##     mode
##                Df Sum of Sq    RSS   AIC F value  Pr(>F)
## <none>                      353143 10704
## energy          1    163.72 352979 10705  0.9601 0.32727
## time_signature  1    231.36 352911 10704  1.3570 0.24419
## key            11   1659.28 351483 10716  0.8841 0.55545
## mode:decade     5   1690.47 351452 10704  1.9875 0.07751 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(forwards.lm)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + decade + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 9: Stepwise method using F-test with P-value cutoff of 0.05

```
#stepwise algorithm - considering interaction terms
step.lm <- lm(smallest, data = spotify)
add1(step.lm, scope = full, test = "F")

## Single term additions
##
## Model:
## popularity ~ 1
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                      516998 11479
## danceability    1     41296 475702 11308 180.4805 < 2.2e-16 ***
## energy          1      9188 507810 11444  37.6173 1.027e-09 ***
## time_signature  1      6578 510420 11454  26.7946 2.481e-07 ***
## loudness        1     29049 487949 11361 123.7701 < 2.2e-16 ***
## key            11      7887 509111 11469   2.9138 0.0007986 ***
## mode            1       738 516260 11478   2.9732 0.0848020 .
## duration_ms     1      7496 509503 11451  30.5853 3.597e-08 ***
## decade          5    143611 373387 10812 159.6161 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

step.lm <- update(step.lm, .~. + danceability)#adding dancebaility
drop1(step.lm, test = "F")

## Single term deletions
##
## Model:
## popularity ~ danceability
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                    475702 11308
## danceability  1     41296 516998 11479  180.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

add1(step.lm, scope = full, test = "F")

## Single term additions
##
## Model:
## popularity ~ danceability
##                Df Sum of Sq    RSS   AIC  F value    Pr(>F)
```

```
## <none>                       475702 11308
## energy           1      4808 470894 11289  21.2184 4.346e-06 ***
## time_signature   1      1609 474093 11303   7.0529  0.007974 **
## loudness         1     17373 458329 11232  78.7673 < 2.2e-16 ***
## key             11      5097 470605 11308   2.0361  0.021974 *
## mode             1         8 475694 11310   0.0370  0.847566
## duration_ms      1      6721 468981 11280  29.7783 5.422e-08 ***
## decade           5    112763 362939 10755 128.8756 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
step.lm <- update(step.lm, .~. + decade)#adding daecade
drop1(step.lm, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + decade
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                     362939 10755
## danceability  1     10448 373387 10812  59.704 1.704e-14 ***
## decade        5    112763 475702 11308 128.876 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(step.lm, scope = full, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade
##                Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                       362939 10755
## energy          1    1791.9 361147 10746 10.2855  0.001361 **
## time_signature  1    1148.2 361791 10750  6.5788  0.010390 *
## loudness        1    4512.0 358427 10731 26.0953 3.547e-07 ***
## key            11    2087.2 360852 10765  1.0848  0.369375
## mode            1     922.0 362017 10752  5.2797  0.021675 *
## duration_ms     1    6403.7 356536 10720 37.2327 1.248e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
step.lm <- update(step.lm, .~. + duration_ms)#adding duration
drop1(step.lm, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + decade + duration_ms
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                     356536 10720
```

```
## danceability  1     10958 367494 10781  63.714 2.359e-15 ***
## decade        5    112446 468981 11280 130.758 < 2.2e-16 ***
## duration_ms   1      6404 362939 10755  37.233 1.248e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(step.lm, scope = full, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms
##                Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                      356536 10720
## energy          1    830.72 355705 10717  4.8390 0.0279330 *
## time_signature  1    510.80 356025 10719  2.9728 0.0848249 .
## loudness        1   2533.52 354002 10707 14.8289 0.0001213 ***
## key            11   1906.79 354629 10731  1.0079 0.4364619
## mode            1    885.13 355650 10717  5.1567 0.0232590 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
step.lm <- update(step.lm, .~. + loudness)#adding loudness
drop1(step.lm, test = "F")
```

```
## Single term deletions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness
##              Df Sum of Sq    RSS   AIC F value    Pr(>F)
## <none>                    354002 10707
## danceability  1      9301 363303 10759  54.439 2.308e-13 ***
## decade        5    102269 456271 11225 119.718 < 2.2e-16 ***
## duration_ms   1      4425 358427 10731  25.901 3.918e-07 ***
## loudness      1      2534 356536 10720  14.829 0.0001213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(step.lm, scope = full, test = "F")
```

```
## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness
##                Df Sum of Sq    RSS   AIC F value  Pr(>F)
## <none>                      354002 10707
## energy          1    220.58 353781 10708  1.2913 0.25595
## time_signature  1    259.52 353743 10707  1.5194 0.21786
## key            11   1685.47 352317 10719  0.8963 0.54323
## mode            1    859.35 353143 10704  5.0397 0.02488 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

step.lm <- update(step.lm, .~. + mode)#adding mode
drop1(step.lm, test = "F")

## Single term deletions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness +
##     mode
##              Df Sum of Sq    RSS   AIC  F value    Pr(>F)
## <none>                     353143 10704
## danceability  1      9665 362807 10758  56.6777 7.615e-14 ***
## decade        5    103071 456214 11227 120.8915 < 2.2e-16 ***
## duration_ms   1      4404 357546 10728  25.8250 4.074e-07 ***
## loudness      1      2508 355650 10717  14.7066 0.0001294 ***
## mode          1       859 354002 10707   5.0397 0.0248784 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

add1(step.lm, scope = full, test = "F")

## Single term additions
##
## Model:
## popularity ~ danceability + decade + duration_ms + loudness +
##     mode
##                Df Sum of Sq    RSS   AIC F value  Pr(>F)
## <none>                      353143 10704
## energy          1    163.72 352979 10705  0.9601 0.32727
## time_signature  1    231.36 352911 10704  1.3570 0.24419
## key            11   1659.28 351483 10716  0.8841 0.55545
## mode:decade     5   1690.47 351452 10704  1.9875 0.07751 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(step.lm)

##
## Call:
## lm(formula = popularity ~ danceability + decade + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
```

```
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 10: Automatic backward

```
#backwards automatic - considering interaction terms
backward.auto.lm <- lm(full, data = spotify)
backward.auto.lm <- step(backward.auto.lm, direction = "backward")

## Start:  AIC=10741.88
## popularity ~ danceability + energy + time_signature + loudness +
##     key + mode + duration_ms + decade + decade:time_signature +
##     mode:key + time_signature:key + decade:mode + time_signature:mode +
##     decade:key
##
##                          Df Sum of Sq    RSS   AIC
## - key:decade             55   12274.1 338651 10709
## - time_signature:key     11    1924.8 328302 10732
## <none>                                 326377 10742
## - time_signature:mode     1     329.7 326707 10742
## - energy                  1     606.7 326984 10744
## - mode:decade             5    2003.1 328380 10745
## - key:mode               11    4002.7 330380 10745
## - loudness                1    1474.5 327852 10749
## - time_signature:decade   5    3655.3 330033 10755
## - duration_ms             1    2431.9 328809 10755
## - danceability            1    8366.2 334743 10793
##
## Step:  AIC=10708.71
## popularity ~ danceability + energy + time_signature + loudness +
##     key + mode + duration_ms + decade + time_signature:decade +
##     key:mode + time_signature:key + mode:decade + time_signature:mode
##
##                          Df Sum of Sq    RSS   AIC
## - time_signature:key     11    1873.9 340525 10698
## <none>                                 338651 10709
```

```
## - time_signature:mode    1     349.5 339001 10709
## - energy                 1     413.8 339065 10709
## - mode:decade            5    2227.1 340878 10712
## - key:mode              11    4406.5 343058 10714
## - loudness               1    1560.5 340212 10716
## - duration_ms            1    2582.7 341234 10722
## - time_signature:decade  5    4150.5 342802 10724
## - danceability           1    8435.8 347087 10758
##
## Step:  AIC=10698.19
## popularity ~ danceability + energy + time_signature + loudness +
##     key + mode + duration_ms + decade + time_signature:decade +
##     key:mode + mode:decade + time_signature:mode
##
##                         Df Sum of Sq    RSS   AIC
## - time_signature:mode    1      57.3 340583 10696
## <none>                               340525 10698
## - energy                 1     370.0 340895 10698
## - mode:decade            5    2347.8 342873 10702
## - key:mode              11    4568.6 345094 10704
## - loudness               1    1460.2 341985 10705
## - time_signature:decade  5    3679.0 344204 10711
## - duration_ms            1    2580.3 343106 10712
## - danceability           1    8713.3 349239 10749
##
## Step:  AIC=10696.54
## popularity ~ danceability + energy + time_signature + loudness +
##     key + mode + duration_ms + decade + time_signature:decade +
##     key:mode + mode:decade
##
##                         Df Sum of Sq    RSS   AIC
## <none>                               340583 10696
## - energy                 1     384.6 340967 10697
## - mode:decade            5    2302.9 342885 10701
## - key:mode              11    4572.6 345155 10702
## - loudness               1    1511.7 342094 10704
## - time_signature:decade  5    3825.3 344408 10710
## - duration_ms            1    2604.8 343187 10710
## - danceability           1    8690.9 349273 10747
```

```
summary(backward.auto.lm)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + energy + time_signature +
##     loudness + key + mode + duration_ms + decade + time_signature:decade +

##     key:mode + mode:decade, data = spotify)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -41.611  -7.018   0.305   7.549  44.034
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.858e+01  8.008e+00   3.568 0.000367 ***
## danceability               1.368e+01  1.897e+00   7.210 7.87e-13 ***
## energy                    -2.989e+00  1.970e+00  -1.517 0.129486
## time_signature             9.367e-01  1.941e+00   0.483 0.629435
## loudness                   3.425e-01  1.139e-01   3.007 0.002671 **
## keyA#                     -4.042e+00  1.743e+00  -2.319 0.020486 *
## keyB                      -1.812e+00  1.988e+00  -0.911 0.362271
## keyC                      -2.353e+00  1.174e+00  -2.005 0.045118 *
## keyC#                     -4.339e+00  1.603e+00  -2.707 0.006838 **
## keyD                      -2.532e+00  1.273e+00  -1.988 0.046948 *
## keyD#                     -5.552e+00  2.501e+00  -2.220 0.026503 *
## keyE                      -2.602e+00  1.441e+00  -1.806 0.071075 .
## keyF                      -4.560e+00  1.493e+00  -3.055 0.002283 **
## keyF#                     -5.827e+00  2.562e+00  -2.275 0.023033 *
## keyG                      -2.169e+00  1.288e+00  -1.684 0.092287 .
## keyG#                     -3.894e+00  1.843e+00  -2.113 0.034712 *
## modeminor                  1.173e+00  2.744e+00   0.427 0.669092
## duration_ms                1.365e-05  3.458e-06   3.947 8.18e-05 ***
## decade50s                 -4.145e+00  8.275e+00  -0.501 0.616489
## decade60s                  8.665e+00  1.063e+01   0.815 0.415177
## decade70s                 -1.497e+01  8.914e+00  -1.679 0.093296 .
## decade80s                 -3.560e+01  1.340e+01  -2.656 0.007965 **
## decade90s                  1.919e+00  1.375e+01   0.140 0.888998
## time_signature:decade50s  -1.963e+00  2.117e+00  -0.927 0.353904
## time_signature:decade60s   4.540e-01  2.703e+00   0.168 0.866606
## time_signature:decade70s   2.481e+00  2.268e+00   1.094 0.274196
## time_signature:decade80s   9.942e+00  3.405e+00   2.920 0.003538 **
## time_signature:decade90s  -3.213e+00  3.496e+00  -0.919 0.358278
## keyA#:modeminor            1.328e-01  3.203e+00   0.041 0.966924
## keyB:modeminor             2.696e+00  2.884e+00   0.935 0.349915
## keyC:modeminor            -6.231e+00  3.699e+00  -1.684 0.092268 .
## keyC#:modeminor            8.952e+00  3.134e+00   2.856 0.004329 **
## keyD:modeminor             1.606e+00  2.902e+00   0.553 0.580133
## keyD#:modeminor            1.158e+01  4.802e+00   2.411 0.015979 *
## keyE:modeminor             6.056e+00  2.730e+00   2.218 0.026644 *
## keyF:modeminor             6.247e+00  3.244e+00   1.926 0.054275 .
## keyF#:modeminor            8.081e+00  3.665e+00   2.205 0.027575 *
## keyG:modeminor             2.841e+00  3.150e+00   0.902 0.367370
## keyG#:modeminor            3.541e+00  4.472e+00   0.792 0.428536
## modeminor:decade50s       -4.432e+00  2.767e+00  -1.602 0.109325
## modeminor:decade60s       -6.290e+00  2.987e+00  -2.105 0.035377 *
## modeminor:decade70s       -6.567e+00  2.511e+00  -2.616 0.008971 **
## modeminor:decade80s       -9.560e+00  2.730e+00  -3.502 0.000471 ***
## modeminor:decade90s       -7.498e+00  2.965e+00  -2.529 0.011529 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.93 on 2037 degrees of freedom
## Multiple R-squared:  0.3412, Adjusted R-squared:  0.3273
## F-statistic: 24.54 on 43 and 2037 DF,  p-value: < 2.2e-16
```

Approach 11: Automatic forward

```
#forwards automatic - considering interaction terms
forward.auto.lm <- lm(smallest, data = spotify)
forward.auto.lm <- step(forward.auto.lm, scope = full, direction = "forward")
```

```
## Start:  AIC=11479.11
## popularity ~ 1
##
##                   Df Sum of Sq    RSS   AIC
## + decade           5    143611 373387 10812
## + danceability     1     41296 475702 11308
## + loudness         1     29049 487949 11361
## + energy           1      9188 507810 11444
## + duration_ms      1      7496 509503 11451
## + time_signature   1      6578 510420 11454
## + key             11      7887 509111 11469
## + mode             1       738 516260 11478
## <none>                         516998 11479
##
## Step:  AIC=10811.91
## popularity ~ decade
##
##                   Df Sum of Sq    RSS   AIC
## + danceability     1   10448.0 362939 10755
## + loudness         1    6411.7 366975 10778
## + duration_ms      1    5893.4 367494 10781
## + time_signature   1    2924.5 370463 10798
## + energy           1    2914.4 370473 10798
## + mode             1     525.3 372862 10811
## <none>                         373387 10812
## + key             11    1987.0 371400 10823
##
## Step:  AIC=10754.85
## popularity ~ decade + danceability
##
##                   Df Sum of Sq    RSS   AIC
## + duration_ms      1    6403.7 356536 10720
## + loudness         1    4512.0 358427 10731
## + energy           1    1791.9 361147 10746
## + time_signature   1    1148.2 361791 10750
## + mode             1     922.0 362017 10752
## <none>                         362939 10755
```

```
## + key            11    2087.2 360852 10765
##
## Step:  AIC=10719.8
## popularity ~ decade + danceability + duration_ms
##
##                  Df Sum of Sq    RSS   AIC
## + loudness        1   2533.52 354002 10707
## + mode            1    885.13 355650 10717
## + energy          1    830.72 355705 10717
## + time_signature  1    510.80 356025 10719
## <none>                        356536 10720
## + key            11   1906.79 354629 10731
##
## Step:  AIC=10706.96
## popularity ~ decade + danceability + duration_ms + loudness
##
##                  Df Sum of Sq    RSS   AIC
## + mode            1    859.35 353143 10704
## <none>                        354002 10707
## + time_signature  1    259.52 353743 10707
## + energy          1    220.58 353781 10708
## + key            11   1685.47 352317 10719
##
## Step:  AIC=10703.9
## popularity ~ decade + danceability + duration_ms + loudness +
##     mode
##
##                  Df Sum of Sq    RSS   AIC
## <none>                        353143 10704
## + mode:decade     5   1690.47 351452 10704
## + time_signature  1    231.36 352911 10704
## + energy          1    163.72 352979 10705
## + key            11   1659.28 351483 10716
```

```r
summary(forward.auto.lm)
```

```
##
## Call:
## lm(formula = popularity ~ decade + danceability + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
```

```
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

Approach 12: Automatic stepwise

```
#stepwise automatic - considering interaction terms
step.auto.lm <- lm(smallest, data = spotify)
step.auto.lm <- step(step.auto.lm, scope = full,direction = "both")

## Start:  AIC=11479.11
## popularity ~ 1
##
##                   Df Sum of Sq    RSS   AIC
## + decade           5    143611 373387 10812
## + danceability     1     41296 475702 11308
## + loudness         1     29049 487949 11361
## + energy           1      9188 507810 11444
## + duration_ms      1      7496 509503 11451
## + time_signature   1      6578 510420 11454
## + key             11      7887 509111 11469
## + mode             1       738 516260 11478
## <none>                         516998 11479
##
## Step:  AIC=10811.91
## popularity ~ decade
##
##                   Df Sum of Sq    RSS   AIC
## + danceability     1     10448 362939 10755
## + loudness         1      6412 366975 10778
## + duration_ms      1      5893 367494 10781
## + time_signature   1      2925 370463 10798
## + energy           1      2914 370473 10798
## + mode             1       525 372862 10811
## <none>                         373387 10812
## + key             11      1987 371400 10823
## - decade           5    143611 516998 11479
##
## Step:  AIC=10754.85
```

```
## popularity ~ decade + danceability
##
##                   Df Sum of Sq    RSS   AIC
## + duration_ms      1      6404 356536 10720
## + loudness         1      4512 358427 10731
## + energy           1      1792 361147 10746
## + time_signature   1      1148 361791 10750
## + mode             1       922 362017 10752
## <none>                         362939 10755
## + key             11      2087 360852 10765
## - danceability     1     10448 373387 10812
## - decade           5    112763 475702 11308
##
## Step:  AIC=10719.8
## popularity ~ decade + danceability + duration_ms
##
##                   Df Sum of Sq    RSS   AIC
## + loudness         1      2534 354002 10707
## + mode             1       885 355650 10717
## + energy           1       831 355705 10717
## + time_signature   1       511 356025 10719
## <none>                         356536 10720
## + key             11      1907 354629 10731
## - duration_ms      1      6404 362939 10755
## - danceability     1     10958 367494 10781
## - decade           5    112446 468981 11280
##
## Step:  AIC=10706.96
## popularity ~ decade + danceability + duration_ms + loudness
##
##                   Df Sum of Sq    RSS   AIC
## + mode             1       859 353143 10704
## <none>                         354002 10707
## + time_signature   1       260 353743 10707
## + energy           1       221 353781 10708
## + key             11      1685 352317 10719
## - loudness         1      2534 356536 10720
## - duration_ms      1      4425 358427 10731
## - danceability     1      9301 363303 10759
## - decade           5    102269 456271 11225
##
## Step:  AIC=10703.9
## popularity ~ decade + danceability + duration_ms + loudness +
##     mode
##
##                   Df Sum of Sq    RSS   AIC
## <none>                         353143 10704
## + mode:decade      5      1690 351452 10704
## + time_signature   1       231 352911 10704
## + energy           1       164 352979 10705
```

```
## - mode            1       859 354002 10707
## + key            11      1659 351483 10716
## - loudness        1      2508 355650 10717
## - duration_ms     1      4404 357546 10728
## - danceability    1      9665 362807 10758
## - decade          5    103071 456214 11227
```

```
summary(step.auto.lm)
```

```
##
## Call:
## lm(formula = popularity ~ decade + danceability + duration_ms +
##     loudness + mode, data = spotify)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.716  -7.323   0.351   7.623  44.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.683e+01  1.848e+00  14.522  < 2e-16 ***
## decade50s    -1.177e+01  1.156e+00 -10.182  < 2e-16 ***
## decade60s     1.054e+01  1.296e+00   8.131 7.26e-16 ***
## decade70s    -6.171e+00  1.129e+00  -5.465 5.18e-08 ***
## decade80s     1.346e+00  1.270e+00   1.059 0.289633
## decade90s    -1.186e+01  1.356e+00  -8.752  < 2e-16 ***
## danceability  1.392e+01  1.848e+00   7.528 7.62e-14 ***
## duration_ms   1.724e-05  3.392e-06   5.082 4.07e-07 ***
## loudness      2.771e-01  7.226e-02   3.835 0.000129 ***
## modeminor    -1.592e+00  7.093e-01  -2.245 0.024878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 2071 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.314
## F-statistic: 106.8 on 9 and 2071 DF,  p-value: < 2.2e-16
```

All experiments above will produce 12 models, to easly compare this 12 models we will create a table that will gather all coeffisients produced.

```
#table to compare all 12 models produced so far
comparison_of_coefficents=list(forwards_interaction = forwards.lm,
                               backwards_interaction = backwards.lm,
                               step_interaction = step.lm,
                               forward_auto_interaction = forward.auto.lm,
                               backward_auto_interaction = backward.auto.lm,
                               step_auto_interaction = step.auto.lm,
                               forwards_no_interaction = approach_1,
                               backwards_no_interaction = approach_2,
                               stepwise_no_interaction = approach_3,
                               forwards_no_interaction_auto = approach_4,
```

```
                                backwards_no_interaction_auto = approach_5,
                                stepwise_no_interaction_auto = approach_6) %>%
  map_df(broom::tidy, .id = "Model") %>%
  select(Model, term, estimate) %>%
  spread(Model, estimate)
```

By Looking closely at the table produced, we can see that not all models are different. Based on the given coefficients we found that we only a few models that are different to one another, because Approach 1 = Approach 3 = Approach 4= Approach 5 = Approach 6 = Approach 8 = Approach 9 = Approach 11 = Approach 12.

Thus, If we take those same model as one model, based on the comparison above now we only have 4 different models which are:

1. Approach 1: forward (no interaction)
2. Approach 2: backward (no interaction)
3. Approach 10: backward automatic (with interaction)
4. Approach 7: backward (with interaction)

In the next step we will use AIC and Cross validation to determine which of these 4 models is the best one.

Akaike information criterion (AIC)

To choose the best model from 4 different models that we have, we will calculate the AIC and according to the theory we choose the model with the lowest AIC.

```
AIC(approach_5, approach_2, backward.auto.lm, backwards.lm)
```

```
##                   df      AIC
## approach_5        11 16611.53
## approach_2        10 16614.58
## backward.auto.lm 45 16604.16
## backwards.lm     44 16604.51
```

Based on the four different AIC above we see that AIC for the third model (backward.auto.lm) and the (backwards.lm) are relatively the same, so temporarily we will take these two models as our best models.

Cross Validation

Another useful method that can be used to access how good a model is for prediction is cross-validation. In this experiments we will use 5-fold cross-validation, so we split the data into 5 parts. In each step, we train the model on 4 parts and test for the 5th part. We do those steps in the following codes:

To find the cross validation estimate of the prediction error, first we will need to get the popularity prediction from each model, then we can calculate the mean square error.

This code will calculate the popularity prediction from our 4 different model.

```
get_pred  <- function(model, test_data){
  data  <-as.data.frame(test_data)
  pred  <-add_predictions(data, model)
  return(pred)
}

pred1  <-map2_df(models1, spotify_CV$test, get_pred, .id = "Run") #approach_5
pred2  <- map2_df(models2, spotify_CV$test, get_pred, .id = "Run")#approach_2
pred3  <-map2_df(models3, spotify_CV$test, get_pred, .id = "Run")#backwards.a
uto.lm
pred4  <- map2_df(models4, spotify_CV$test, get_pred, .id = "Run")#backwards.
lm
```

This code will calculate the mean square error (MSE) from our 4 different models.

```
#MSE for approach_5
MSE1  <- pred1%>%
  group_by(Run)%>%
  summarise(
    MSE =mean( (popularity- pred) ^2),
    n =n())
MSE1

## # A tibble: 5 x 3
##   Run     MSE     n
##   <chr> <dbl> <int>
## 1 1      156.   417
## 2 2      174.   416
## 3 3      184.   416
## 4 4      178.   416
## 5 5      157.   416

#MSE for approach_2
MSE2  <- pred2%>%
  group_by(Run)%>%
  summarise(
```

```
    MSE =mean( (popularity- pred) ^2),
    n =n())
MSE2
```

```
## # A tibble: 5 x 3
##   Run     MSE     n
##   <chr> <dbl> <int>
## 1 1      157.   417
## 2 2      176.   416
## 3 3      184.   416
## 4 4      177.   416
## 5 5      157.   416
```

```
#MSE for backwards.auto.lm
MSE3  <- pred3%>%
  group_by(Run)%>%
  summarise(
    MSE =mean( (popularity- pred) ^2),
    n =n())
MSE3
```

```
## # A tibble: 5 x 3
##   Run     MSE     n
##   <chr> <dbl> <int>
## 1 1      152.   417
## 2 2      160.   416
## 3 3      180.   416
## 4 4      173.   416
## 5 5      153.   416
```

```
#MSE for backwards.lm
MSE4  <- pred4%>%
  group_by(Run)%>%
  summarise(
    MSE =mean( (popularity- pred) ^2),
    n =n())
MSE4
```

```
## # A tibble: 5 x 3
##   Run     MSE     n
##   <chr> <dbl> <int>
## 1 1      153.   417
## 2 2      160.   416
## 3 3      180.   416
## 4 4      173.   416
## 5 5      154.   416
```

By using the popularity prediction and mean square error above, finally we can calculate the cross validation prediction error for each model by using code below:

```r
#CV for approach_1
CV1  <-sum(MSE1$MSE*MSE1$n)/ sum(MSE1$n)
CV1

## [1] 169.6986

#CV for approach_2
CV2  <-sum(MSE2$MSE*MSE2$n)/ sum(MSE2$n)
CV2

## [1] 170.1115

#CV for backwards.auto.lm
CV3  <-sum(MSE3$MSE*MSE3$n)/ sum(MSE3$n)
CV3

## [1] 163.6629

#CV for backwards.lm
CV4  <-sum(MSE4$MSE*MSE4$n)/ sum(MSE4$n)
CV4

## [1] 163.8478

#The lowest CV values were for CV3 and CV4 (they were practically the same -
about 163)
#These models were -
#1. backwards.auto.lm: #popularity ~ danceability + energy + time_signature +

#loudness + key + mode + duration_ms + decade + time_signature:decade +
#key:mode + mode:decade
```

From the result above we can see that the cross validation prediction error for the third model (backward.auto.lm) and the fourth model (backwards.lm) are relatively the same, This is in line with AIC that we get in the previous section.

Final Model

Based on the experiments using AIC value and cross-validation, we actually have two models that have the lowest AIC value and cross-validation prediction error almost the same. The two model is listed below:

1. backwards.auto.lm: #popularity ~ danceability + energy + time_signature + loudness + key + mode + duration_ms + decade + time_signature:decade + key:mode + mode:decade

2. backwards.lm: #popularity ~ danceability + time_signature + loudness + key + mode + duration_ms + decade + time_signature:decade + key:mode + mode:decade

The fisrt model mentioned above have AIC value equal to 16604.16 and cross validation prediction error equal to 163.6629, While for the second model, AIC value is equal to 16604.51 and cross validation prediction error equal to 163.8478. Although the different of AIC and cross validation prediction error is relatively small but we will choose model that has the lowest AIC and cross validation prediction error as our best model, which is the first model listed above (approach 10).

So to confirm our final model is

final model:Popularity ~ danceability + energy + time_signature + loudness + key + mode + duration_ms + decade + time_signature:decade + key:mode + mode:decade

By using the coefficients, we can rewrite this model into the followong formulation:

Popularity = 1.367612e+01 * danceability -2.988546e+00 * energy + 9.366721e-01 * time_signature + 3.424707e-01 * loudness - 2.353084e+00 * key +1.364845e-05 * duration_ms +1.918680e+00 * decade - 3.212647e * time_signature:decade + 00-6.230977e+00 * key:mode +2.857782e+01 * mode:decade

Assumption Checking

1) Linearity

Yes, as residuals vs Fitted figure and scale-location figure show(the red line is almost flate), it shows linearity property.

2) Normality

Yes, the data are mostly normal distribution, because Q-Q graph shows that Residuals value almost fit the diagonal line.

3) Constant variance

Yes, the variance of residuals seems the same beacuse there is no pattern of residuals in residuals vs fitted figure

4) Independence

Yes, because residuals vs fitted figure shows that there are no negative or postive relationship between fitted value and residuals.

influential points:

As Residual vs Leverage figure shows, cook distance of all data are < 1 . There seems to be no influential points for our best model if we use cook distance to decide if there are influential points. However, the 1781 data has a very hight leverage value. It might be considered as a potential influential point.

```
plot( backward.auto.lm, which = 1)
```

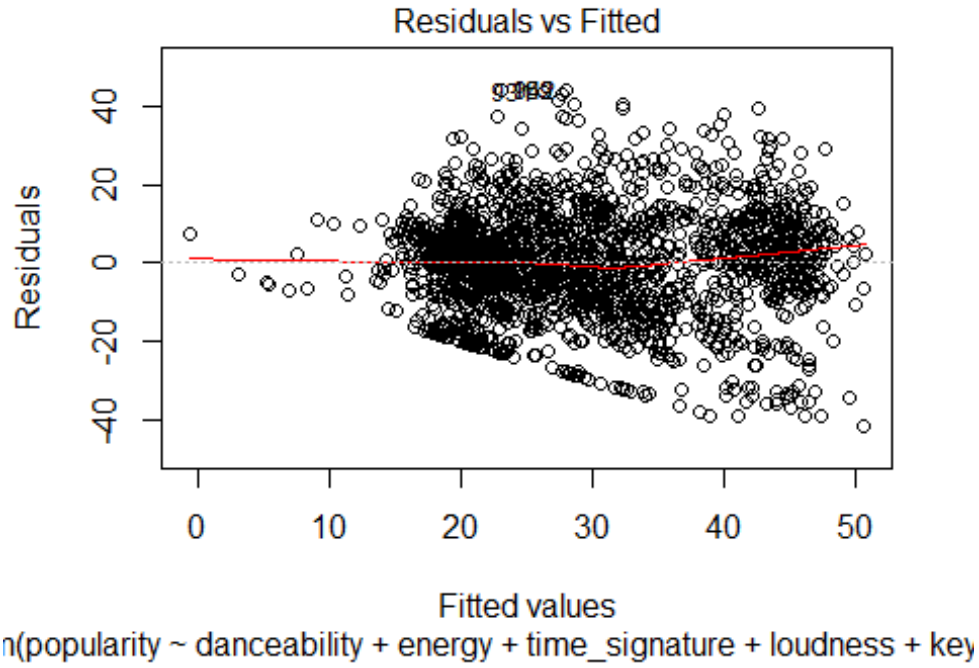

Fig 8.1 ***Residual vs Fitted***

```
plot( backward.auto.lm, which = 2)
```

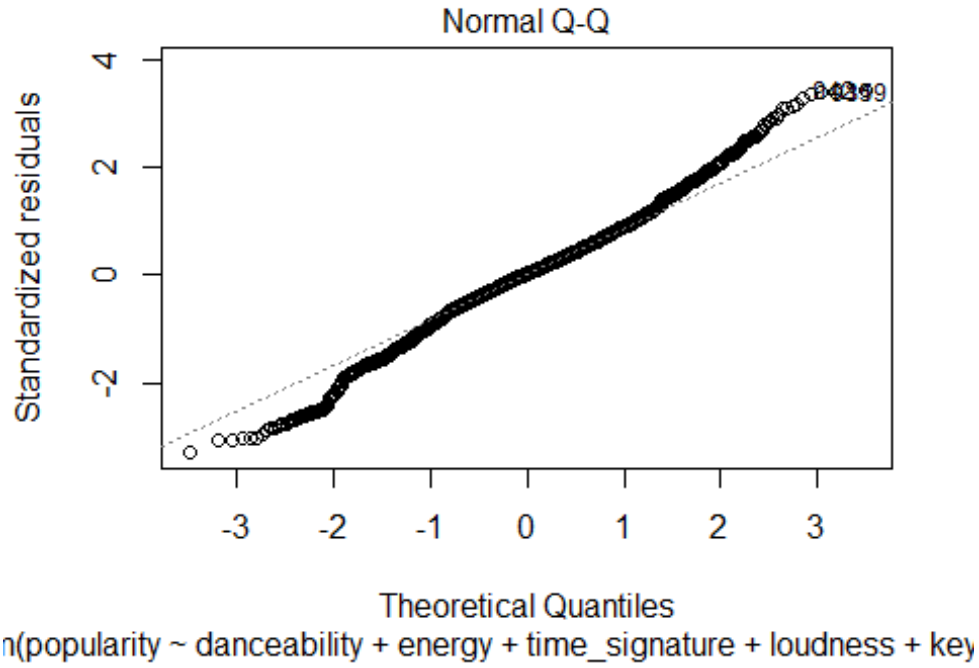

Fig 8.2 ***Normal Q-Q***

```
plot( backward.auto.lm, which = 3)
```

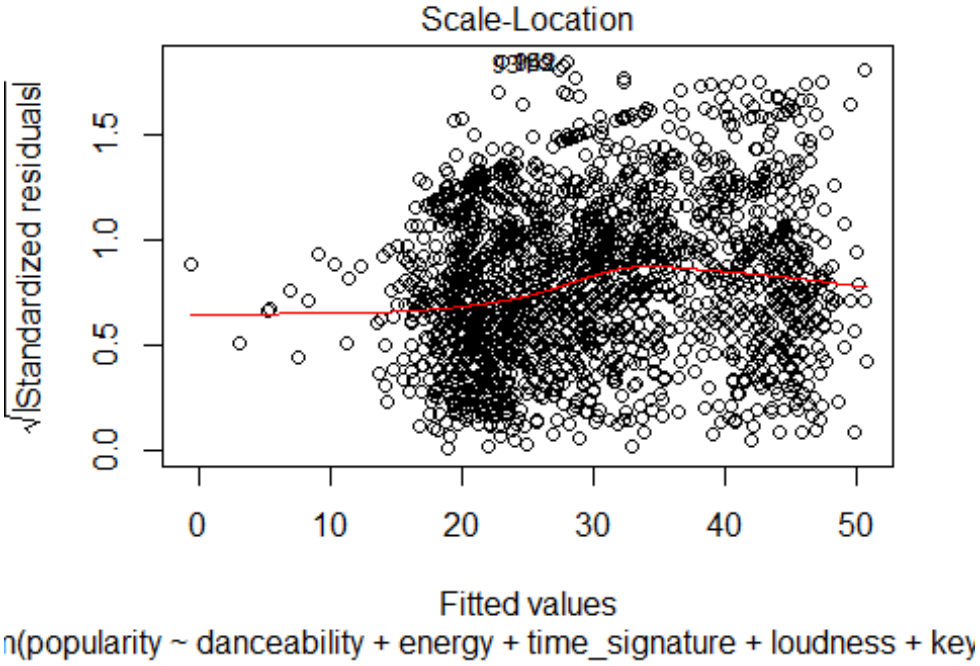

Fig 8.3 Scale-location

```
plot( backward.auto.lm, which = 5)
```

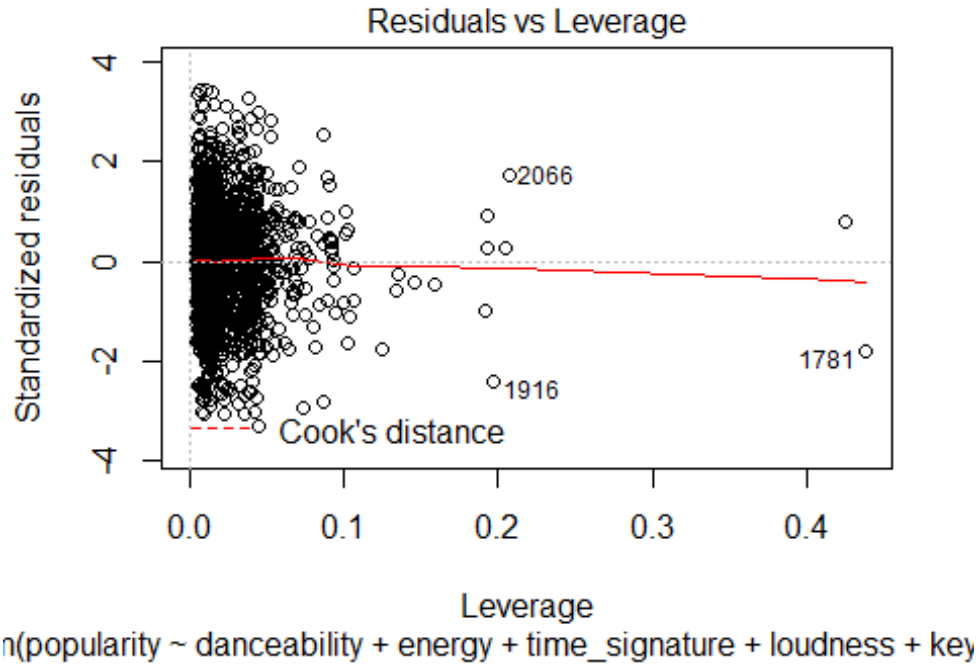

Fig 8.4 ***Residual vs Leverage***

Prediction

As what was required, in this part we will calculate the prediction value of the popularity for a three-minute song from the 90s in the Key of C. The other predictors are at the mean of the dataset. In the following code, we will use our final model to get the prediction value:

Note: 3 minutes = 180000 millisecond, so in this case duration_ms = 180000.

```
#using the coefficients from the comparison_of_coefficents table-
#prediction of the popularity for a three minute song from the 90s in
the Key of C
Popularity_prediction=mean(spotify$danceability)*1.367612e+01 -
  2.988546e+00*mean(spotify$energy)+
  9.366721e-01*mean(spotify$time_signature)+
  3.424707e-01*mean(spotify$loudness)- 2.353084e+00 + 1.364845e-05*180
000 +
  1.918680e+00 - 3.212647e+00-6.230977e+00 + 2.857782e+01
Popularity_prediction

## [1] 26.40463
```

For this specific example, we get the prediction value is equal to 26.40460.

Conclusion:

In our final project, we derive models considering different algorithm, different heuristic and different initial models. Finally, we got four different models as the table1 show.

Since the model of approach 10 has the lowest AIC value (table 3) and the lowest cross-validation predict error(table 2), the best model is approach 10.

Thus, the final mode of our project is followed:

Popularity = 1.367612e+01 * danceability -2.988546e+00 * energy + 9.366721e-01 * time_signature + 3.424707e-01 * loudness - 2.353084e+00 * key +1.364845e-05 * duration_ms +1.918680e+00 * decade - 3.212647e * time_signature:decade + 00-6.230977e+00 * key:mode +2.857782e+01 * mode:decade

Mode name	Model formula
approach 1	popularity ~ danceability + loudness + mode + duration_ms + decade
approach 2	popularity ~ danceability + loudness + duration_ms + decade
approach 10	popularity ~ danceability + energy + time_signature + loudness + key + mode + duration_ms + decade + time_signature:decade + key:mode + mode:decade
approach 7	popularity ~ danceability + time_signature + loudness + key + mode + duration_ms + decade + time_signature:decade + key:mode + mode:decade

Tabel10.1 Model formula

Mode name	cross-validation predict error
approach 1	169.6989
approach 2	170.1115
approach 10	163.6629

approach 7	163.8478

Table10.2 CV value

Mode name	AIC
approach 1	16611.53
approach 2	16614.58
approach 10	16604.16
approach 7	16604.51

Table10.3 AIC value