Spotify

Predictors of Song Popularity

Business Uses

Artists

Producers

What makes a song **popular?**

Our Dataset

- 19,000 Spotify Songs from Kaggle
  - 19 variables of each song:
    - Acousticness, danceability, energy
    - Instrumentalness, liveness, loudness
    - Valence, tempo, key, speechiness, popularity
    - Artist, album, playlist, song title, mode, time signature

Variable Examples

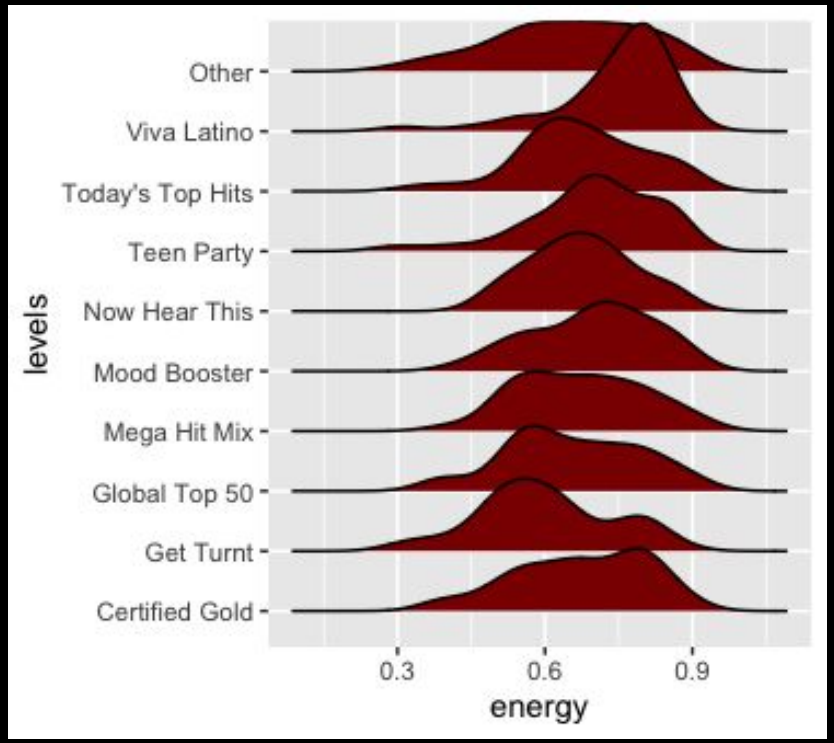
Data Cleaning

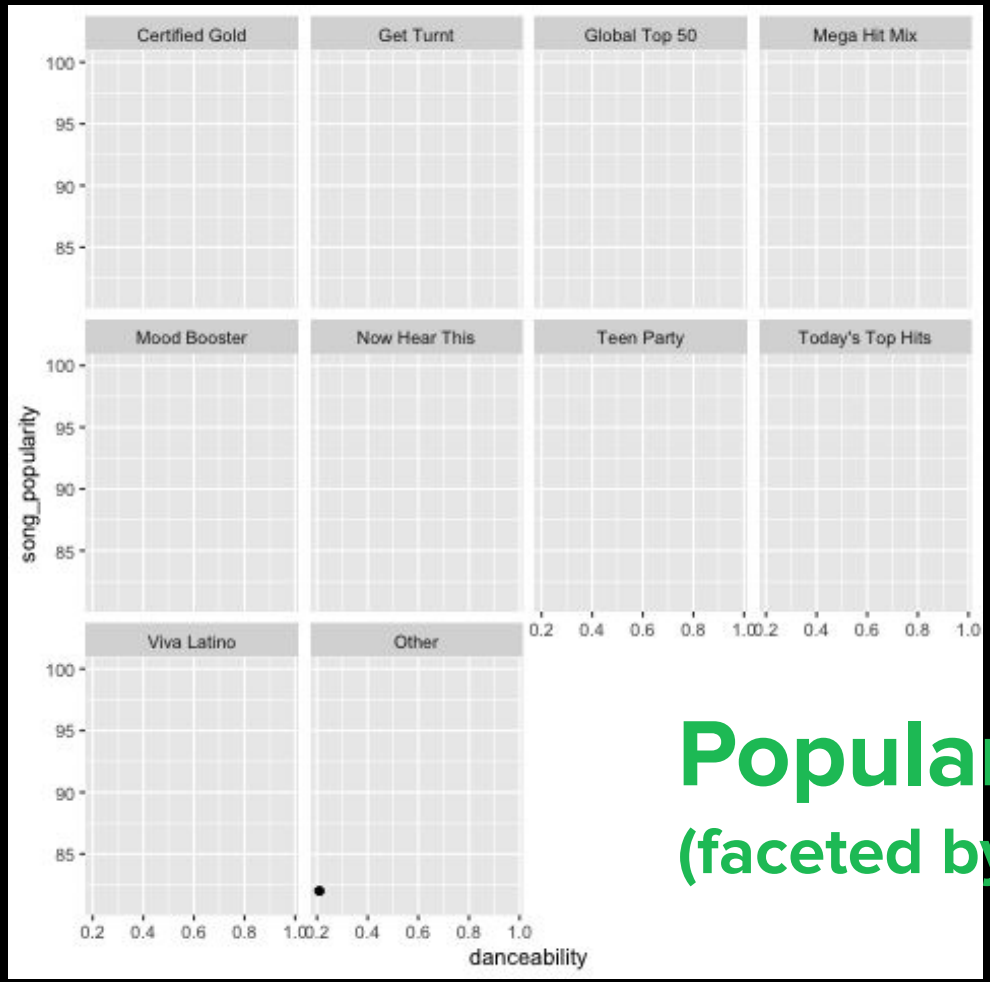
```r
# data cleaning
spotify_data <- full_join(spotify_data %>% group_by(song_name) %>% mutate(id = row_number()),
          spotify_data2 %>% group_by(song_name) %>% mutate(id = row_number()),
          by = c("song_name", "id"))

top_hits <- spotify_data %>% filter(song_popularity > 80) %>%
  group_by(playlist)

top_hits_levels <- fct_lump(top_hits$playlist, n = 9)

top_hits_df <- data.frame(
  top_hits,
  levels = top_hits_levels)
```

```r
# test and train data sets
train_idx <- sample(1:nrow(spotify_data), size = 0.75 * nrow(spotify_data))
spotify_train <- spotify_data %>% slice(train_idx)
spotify_test <- spotify_data %>% slice(-train_idx)
```

Energy vs Levels by Playlist

(distribution of popularity)

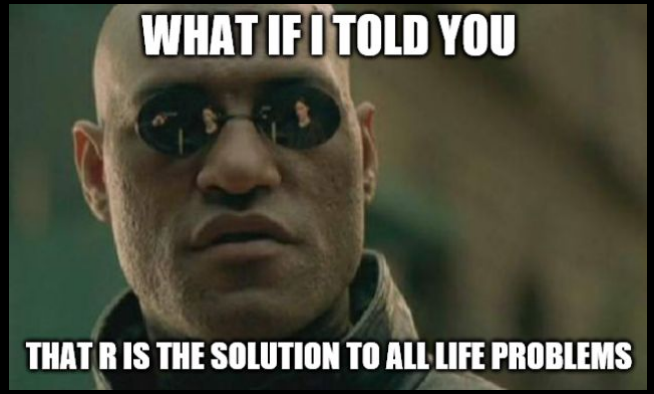




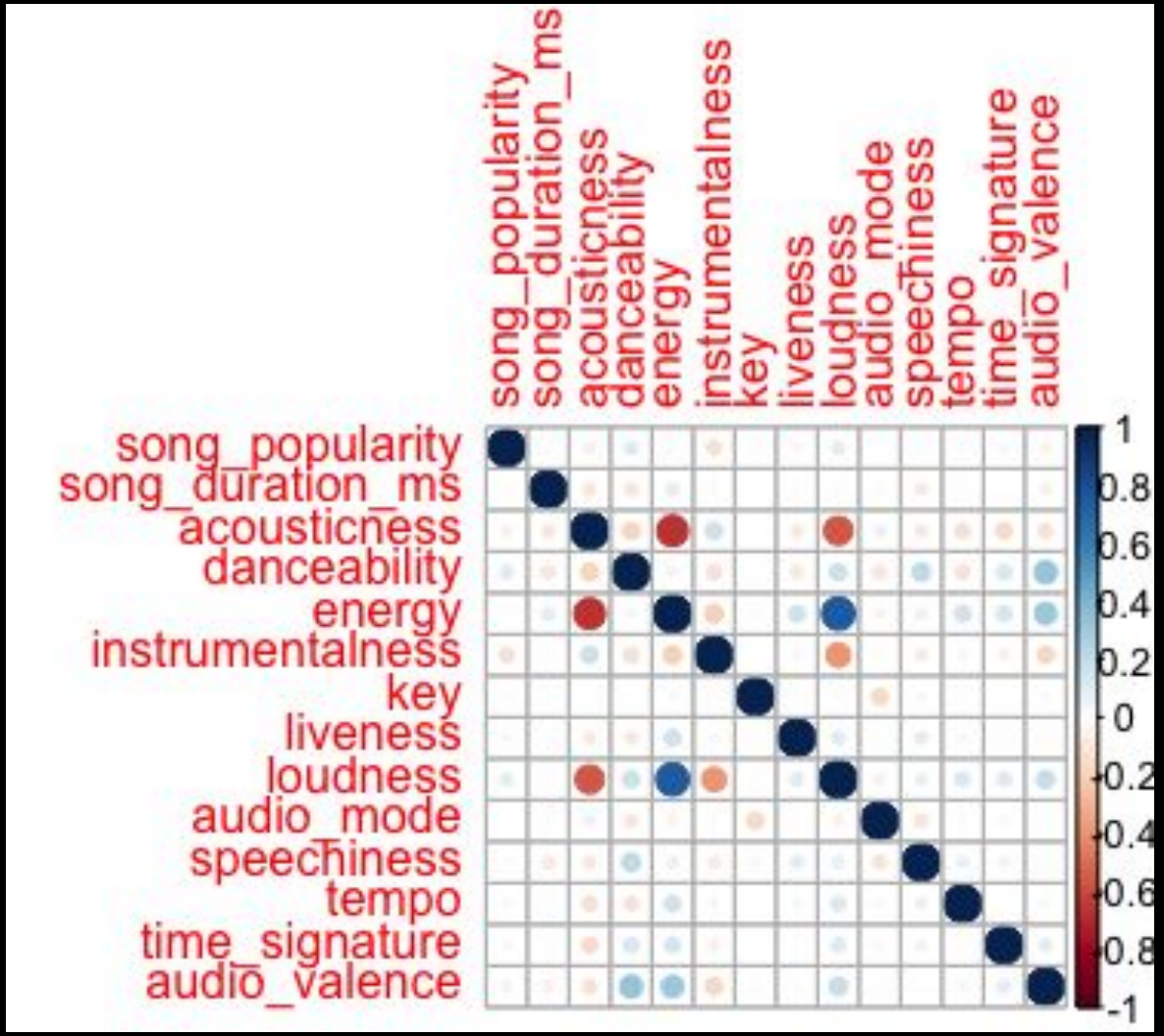
Popularity vs Danceability

(faceted by top 10 playlists)

Correlation Matrix



(Created with memeR)



Sentiment Analysis

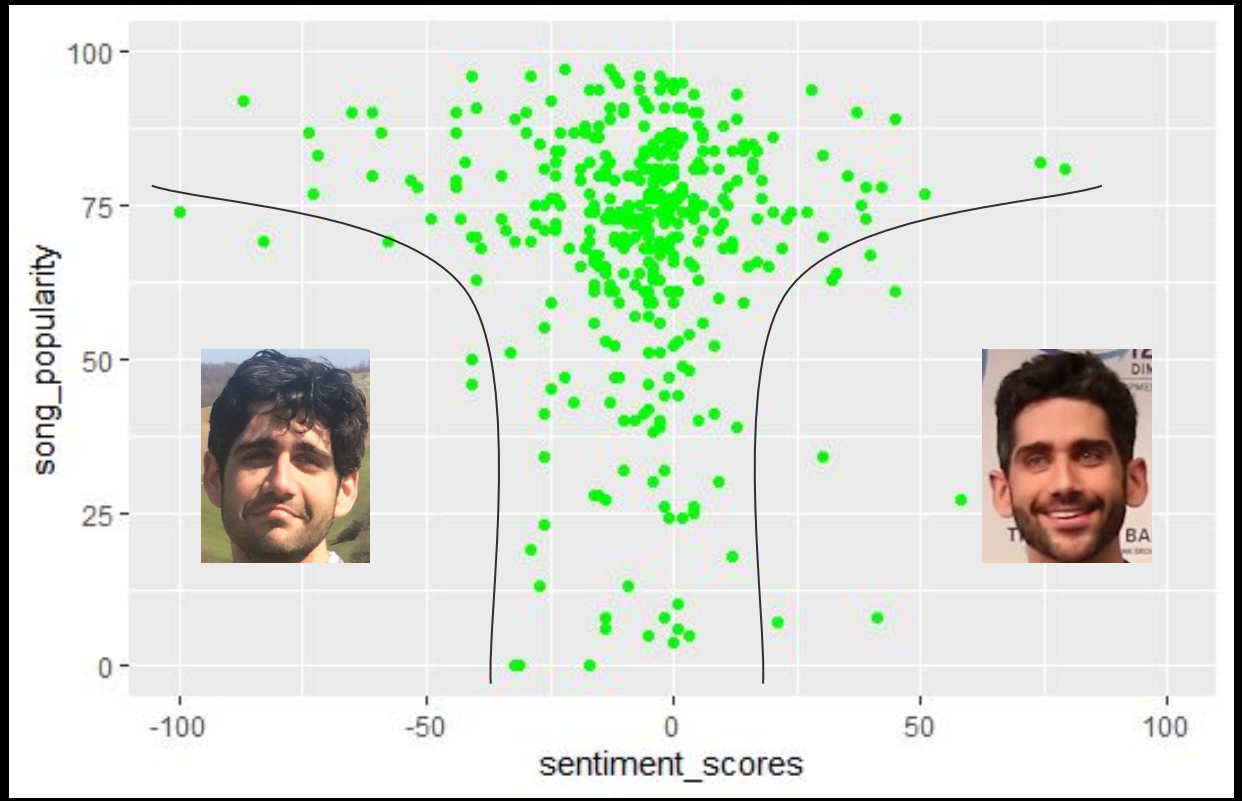
Procedure:

- Lyrics from Genius API
- Parse by words
- Assign values
- Sum up word scores

Roadblocks:

- Computationally expensive
- Time inefficient

⇨ Subset of 5%



Summary Statistics

```
> summary(spotify_data)
                                      song_name     song_popularity  song_duration_ms  acousticness      danceability
 Better                                    :   21   Min.   :  0.00   Min.   :  12000   Min.   :0.000001   Min.   :0.0000
 FEFE (feat. Nicki Minaj & Murda Beatz)    :   19   1st Qu.: 40.00   1st Qu.: 184340   1st Qu.:0.024100   1st Qu.:0.5330
 MIA (feat. Drake)                         :   18   Median : 56.00   Median : 211306   Median :0.132000   Median :0.6450
 Taki Taki (with Selena Gomez, Ozuna & Cardi B):   18   Mean   : 52.99   Mean   : 218212   Mean   :0.258539   Mean   :0.6333
 No Stylist                                :   17   3rd Qu.: 69.00   3rd Qu.: 242844   3rd Qu.:0.424000   3rd Qu.:0.7480
 Electricity (with Dua Lipa)               :   16   Max.   :100.00   Max.   :1799346   Max.   :0.996000   Max.   :0.9870
 (Other)                                   :18726
     energy        instrumentalness        key          liveness         loudness        audio_mode      speechiness
 Min.   :0.00107   Min.   :0.0000000   Min.   : 0.000   Min.   :0.0109   Min.   :-38.768   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.51000   1st Qu.:0.0000000   1st Qu.: 2.000   1st Qu.:0.0929   1st Qu.: -9.044   1st Qu.:0.0000   1st Qu.:0.0378
 Median :0.67400   Median :0.0000114   Median : 5.000   Median :0.1220   Median : -6.555   Median :1.0000   Median :0.0555
 Mean   :0.64499   Mean   :0.0780080   Mean   : 5.289   Mean   :0.1797   Mean   : -7.447   Mean   :0.6281   Mean   :0.1021
 3rd Qu.:0.81500   3rd Qu.:0.0025700   3rd Qu.: 8.000   3rd Qu.:0.2210   3rd Qu.: -4.908   3rd Qu.:1.0000   3rd Qu.:0.1190
 Max.   :0.99900   Max.   :0.9970000   Max.   :11.000   Max.   :0.9860   Max.   :  1.585   Max.   :1.0000   Max.   :0.9410
```

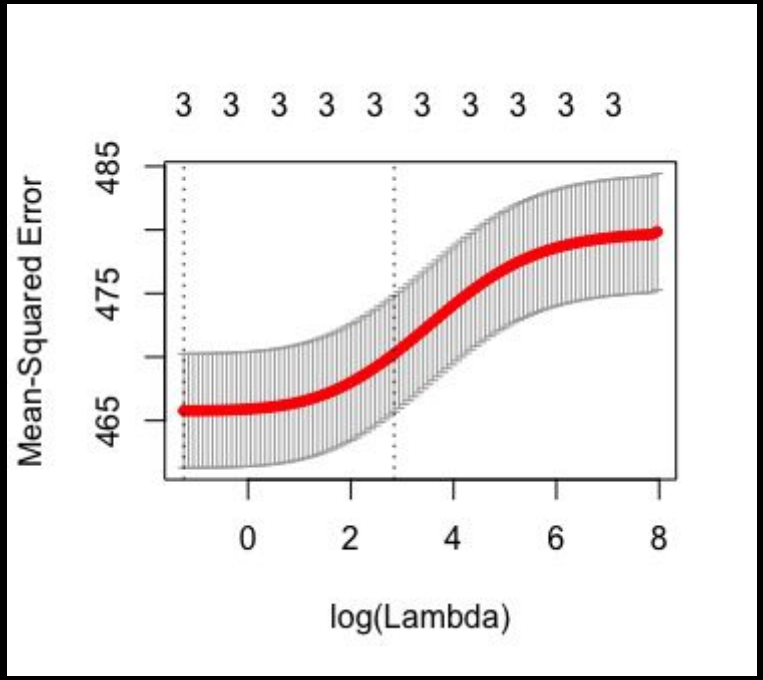
Linear Regression #1

```
Call:
lm(formula = song_popularity ~ instrumentalness + energy + loudness,
    data = spotify_data)

Residuals:
   Min     1Q Median     3Q    Max
-60.75 -12.39   3.08  15.57  46.89

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       71.39534    1.15286   61.93   <2e-16 ***
instrumentalness  -9.15712    0.77924  -11.75   <2e-16 ***
energy           -15.62099    1.13345  -13.78   <2e-16 ***
loudness           1.02232    0.06738   15.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.58 on 18831 degrees of freedom
Multiple R-squared:  0.02968,   Adjusted R-squared:  0.02953
F-statistic:   192 on 3 and 18831 DF,  p-value: < 2.2e-16
```

Linear Regression #2

```
Call:
lm(formula = energy ~ acousticness + loudness + instrumentalness,
    data = spotify_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49951 -0.08649  0.00469  0.08664  0.53036

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.9532817  0.0020386  467.62   <2e-16 ***
acousticness     -0.2555236  0.0037935  -67.36   <2e-16 ***
loudness          0.0334028  0.0003061  109.13   <2e-16 ***
instrumentalness  0.0838532  0.0044558   18.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1245 on 18831 degrees of freedom
Multiple R-squared:  0.6617,    Adjusted R-squared:  0.6616
F-statistic: 1.227e+04 on 3 and 18831 DF,  p-value: < 2.2e-16
```

Ridge Regression #1

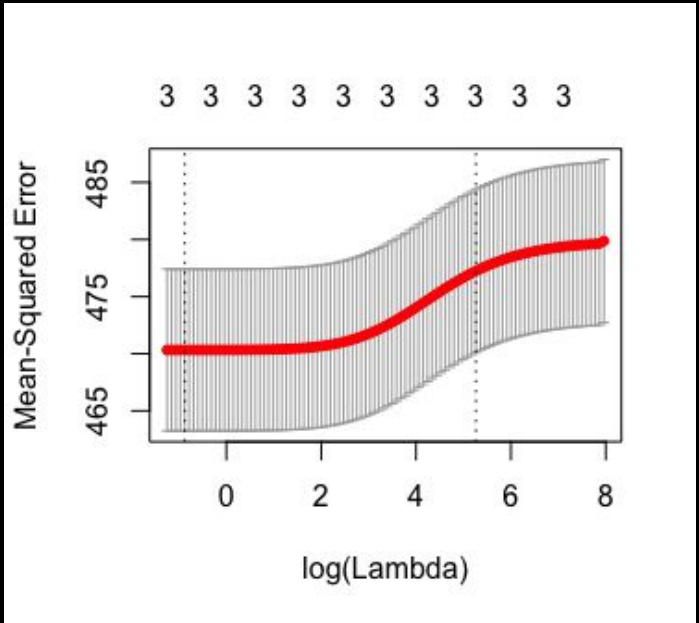



```
> ridge_fit1$lambda.min
[1] 0.2867531
> ridge_fit1$lambda.1se
[1] 22.7247
```

```
> caret::RMSE(ridge_train1$ridge, spotify_train$song_popularity)
[1] 29.68798
> caret::RMSE(ridge_test1$ridge, spotify_test$song_popularity)
[1] 26.52887
```

```
> caret::R2(ridge_train1$ridge, spotify_train$song_popularity)
[1] 0.06042883
> caret::R2(ridge_test1$ridge, spotify_test$song_popularity)
[1] 0.02830398
```

Ridge Regression #2



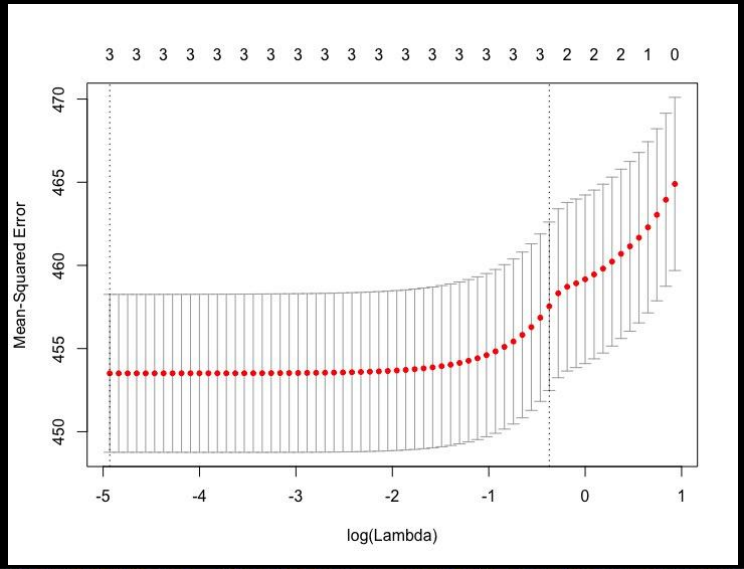

```
> ridge_fit2$lambda.min
[1] 0.4565921
> ridge_fit2$lambda.1se
[1] 307.476
```

```
> caret::RMSE(ridge_train2$ridge, spotify_train$song_popularity)
[1] 30.18275
> caret::RMSE(ridge_test2$ridge, spotify_test$song_popularity)
[1] 26.76642
```

```
> caret::R2(ridge_train2$ridge, spotify_train$song_popularity)
[1] 0.02680394
> caret::R2(ridge_test2$ridge, spotify_test$song_popularity)
[1] 0.02080898
```

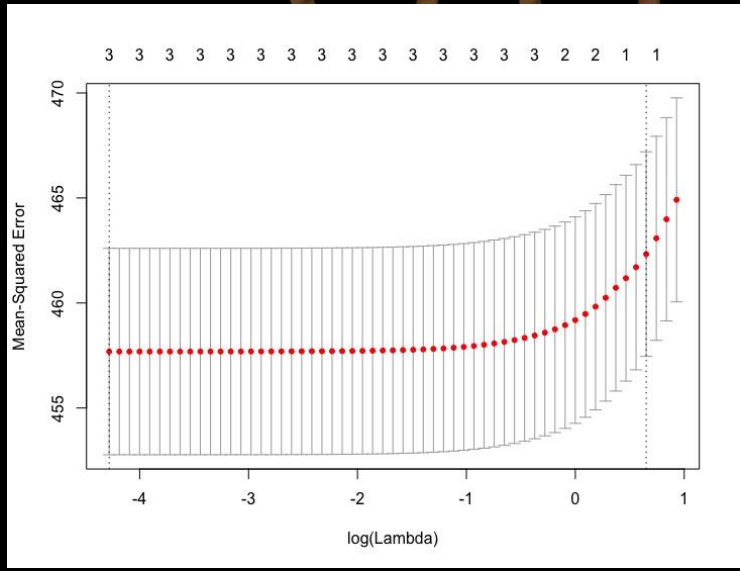
Lasso

Lasso #1



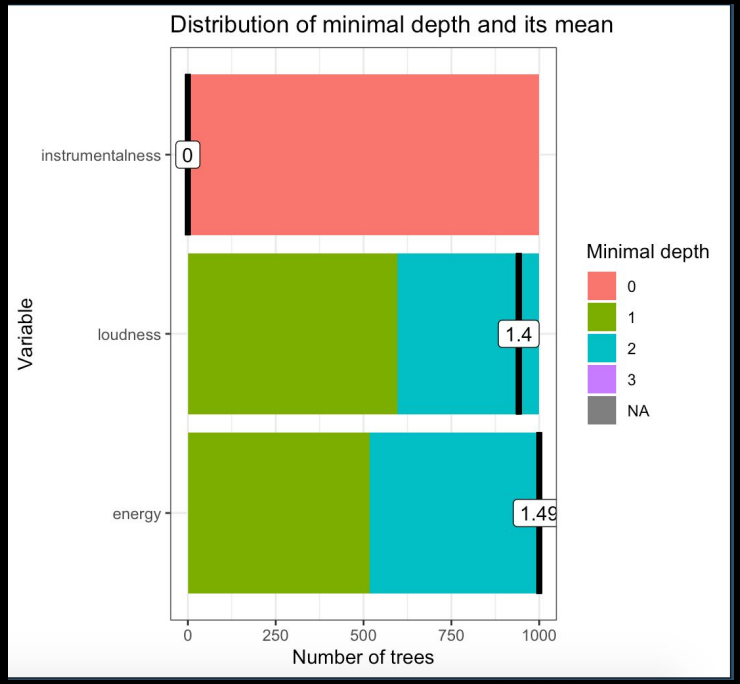

```
> caret::RMSE(lasso_train1$lasso, spotify_train$song_popularity)
[1] 20.81582
> caret::RMSE(lasso_test1$lasso, spotify_test$song_popularity)
[1] 21.21489
> caret::R2(lasso_train1$lasso, spotify_train$song_popularity)
[1] 0.06145499
> caret::R2(lasso_test1$lasso, spotify_test$song_popularity)
[1] 0.0314849
```

Lasso #2




```
> caret::RMSE(lasso_train2$lasso, spotify_train$song_popularity)
[1] 21.18681
> caret::RMSE(lasso_test2$lasso, spotify_test$song_popularity)
[1] 21.31705
> caret::R2(lasso_train2$lasso, spotify_train$song_popularity)
[1] 0.02806788
> caret::R2(lasso_test2$lasso, spotify_test$song_popularity)
[1] 0.02387179
```

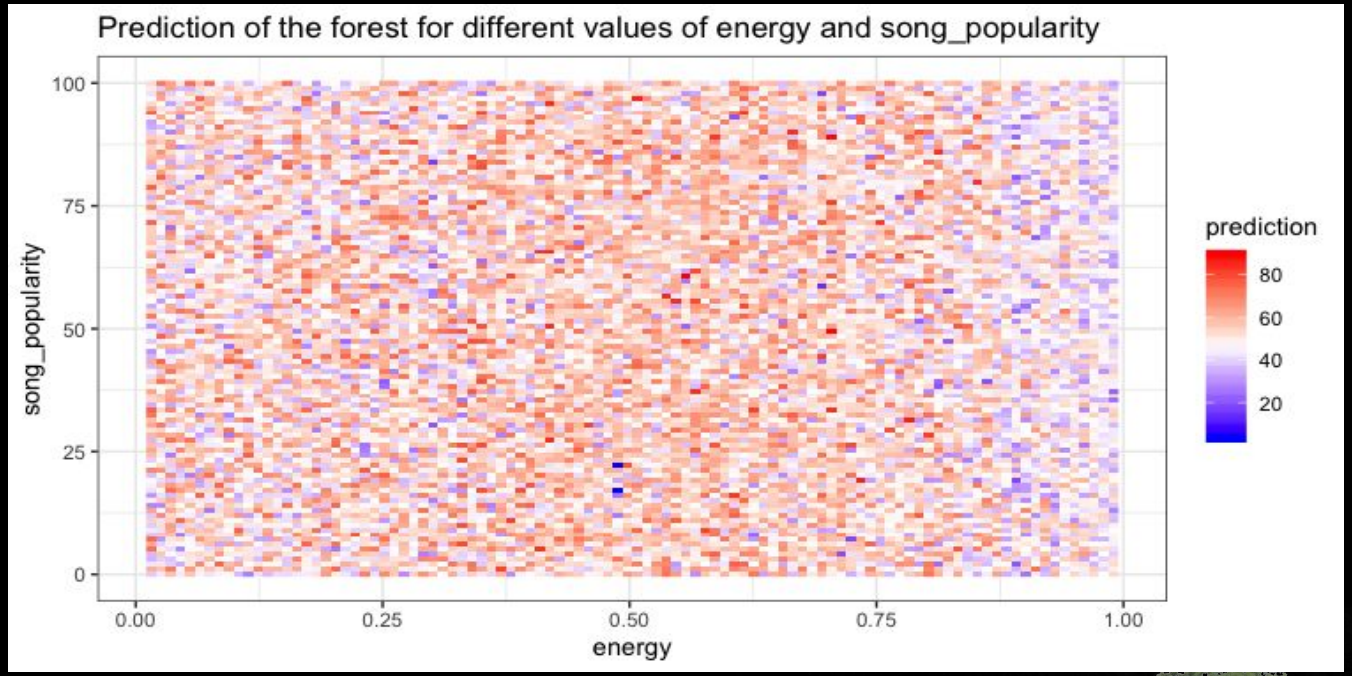
Random Forest #1



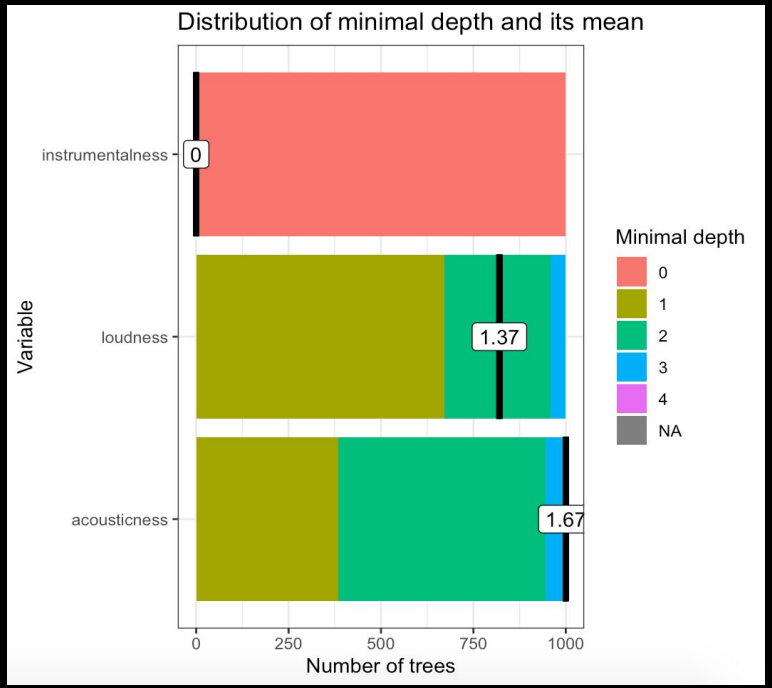

```
rf_fit1 <- randomForest(song_popularity ~ instrumentalness + energy + loudness,
                        data = spotify_data,
                        type = classification,
                        ntree = 1000,
                        mtry = 3,
                        importance = TRUE,
                        localImp = TRUE)
```

```
> summary(test_preds)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.121  49.784  61.490  59.717  71.275 100.000
> oob_preds <- predict(rf_fit)
> summary(oob_preds)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.1687  44.8645  52.1468  53.9978  61.0711 100.0000
> ib_preds <- predict(rf_fit, spotify_train)
> summary(ib_preds)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.121  42.100  51.496  51.797  60.573 100.000
```

Random Forest #1



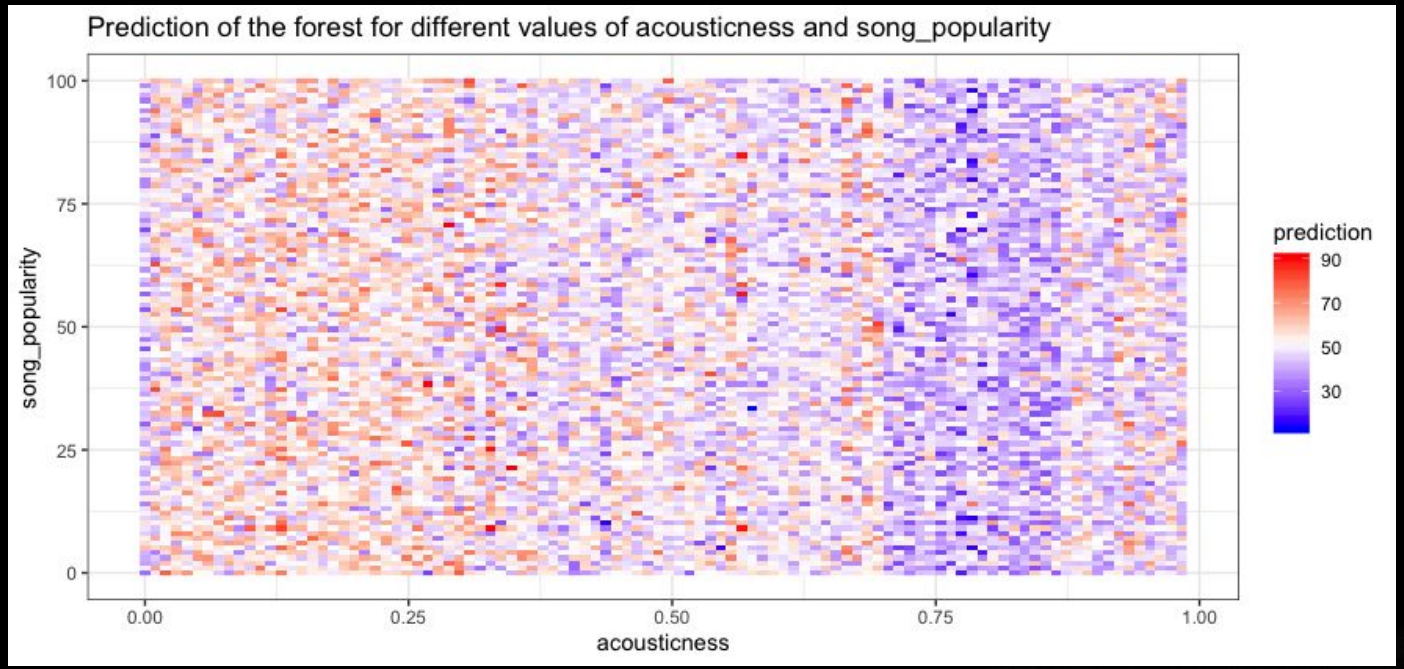
Random Forest #2




```
rf_fit2 <- randomForest(song_popularity ~ instrumentalness + acousticness+ loudness,
                        data = spotify_train,
                        type = classification,
                        ntree = 1000,
                        mtry = 3,
                        importance = TRUE,
                        localImp = TRUE)
rf_fit2
```

```
> summary(test_preds)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.423  54.781  63.108  62.797  70.980  99.419
> oob_preds <- predict(rf_fit2)
> summary(oob_preds)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.863  54.304  61.862  62.486  69.827  99.130
> ib_preds <- predict(rf_fit2, spotify_train)
> summary(ib_preds)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.423  52.746  63.066  61.861  71.397  99.419
```

Random Forest #2



What makes a song **popular?**