THE SPOTIFY DATASET

Gold Team 2

AGENDA

01
Understanding the Data

02
Regression

03
Word Cloud

04
Clustering

05
Challenges

06
Conclusion

INTRODUCTION

Although music is subjective, some songs are massive hits and others are niche and unpopular. Though the Spotify dataset, we'll be looking at:

- Factors that make a song popular
- Predictability of
- Commonalities of popular songs

through analysis and data modeling techniques.

DATASET DESCRIPTION

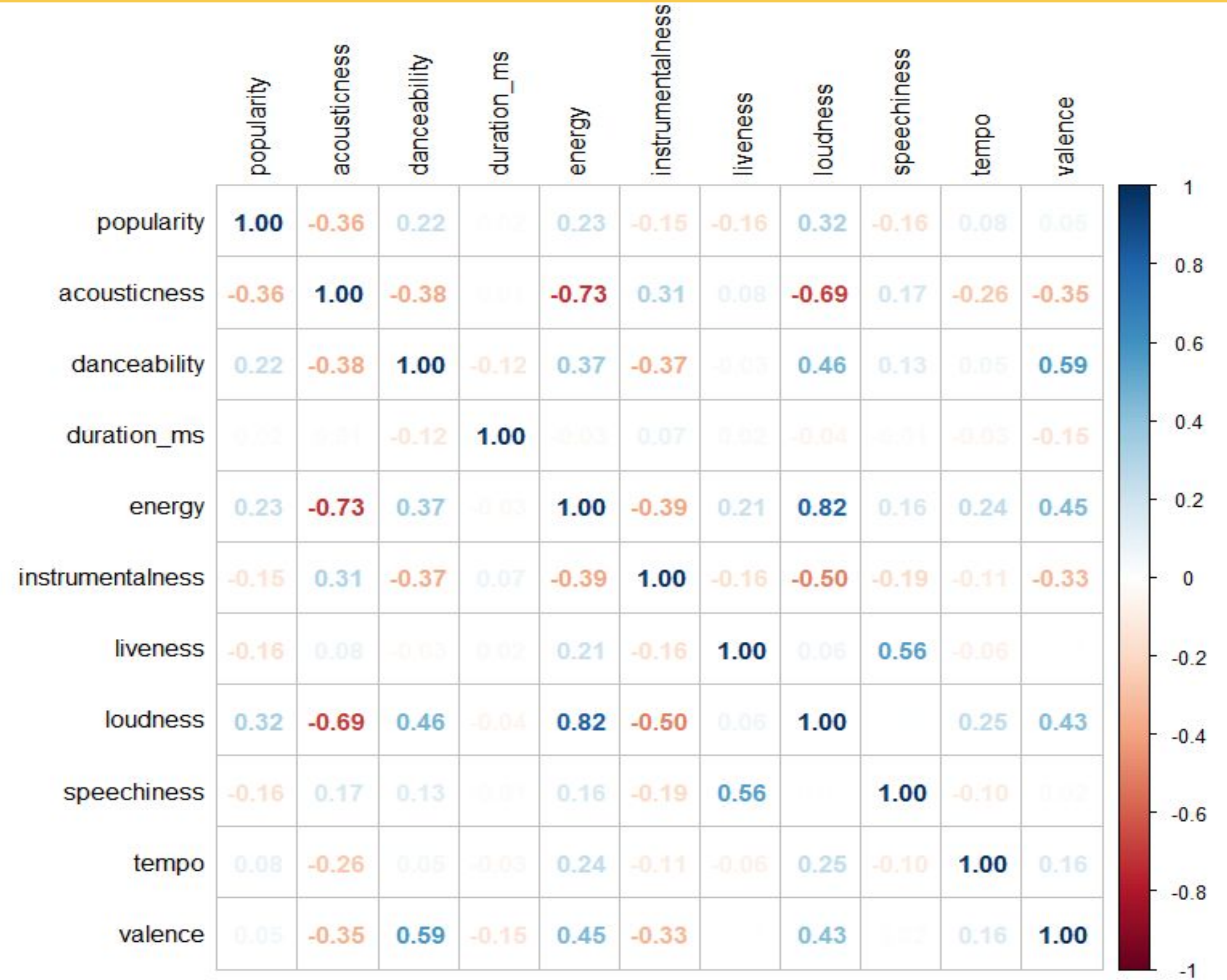
- Over 230,000 songs pulled using Spotify API
- 18 different attributes
- 26 genres, around 10,000 songs per genre

Which factors contribute towards popularity?

Name	Type	Description
Genre	String (26 Levels)	One word describing genre of track
artist_name	String	Name of Artist
track_name	String	Name of the Song
track_id	String	Unique ID generated by Spotify to identify each song
popularity	Num (1-100)	A number ranging from 1 to 100 describing how popular a song is.
accousticness	Num (0-1)	A number ranging from 0 to 1. This value describes how acoustic a song is. A score of 1.0 means the song is most likely to be an acoustic one.
danceability	Num (0-1)	A number ranging from 0 to 1. This value describes how danceable a song is. A score of 1.0 means the song is the
duration_ms	Integer	Length of song in milliseconds
energy	Num (0-1)	A number ranging from 0 to 1. This value describes how energetic a song is.
Instrumentalness	Num (0-1)	This represents the amount of vocals in the song. The closer it is to 1.0, the more instrumental the song is.
key	Character	Represents the Key that the song is in.
liveness	Num (0-1)	A number ranging from 0 to 1. This value represents the likelihood that the track is live.

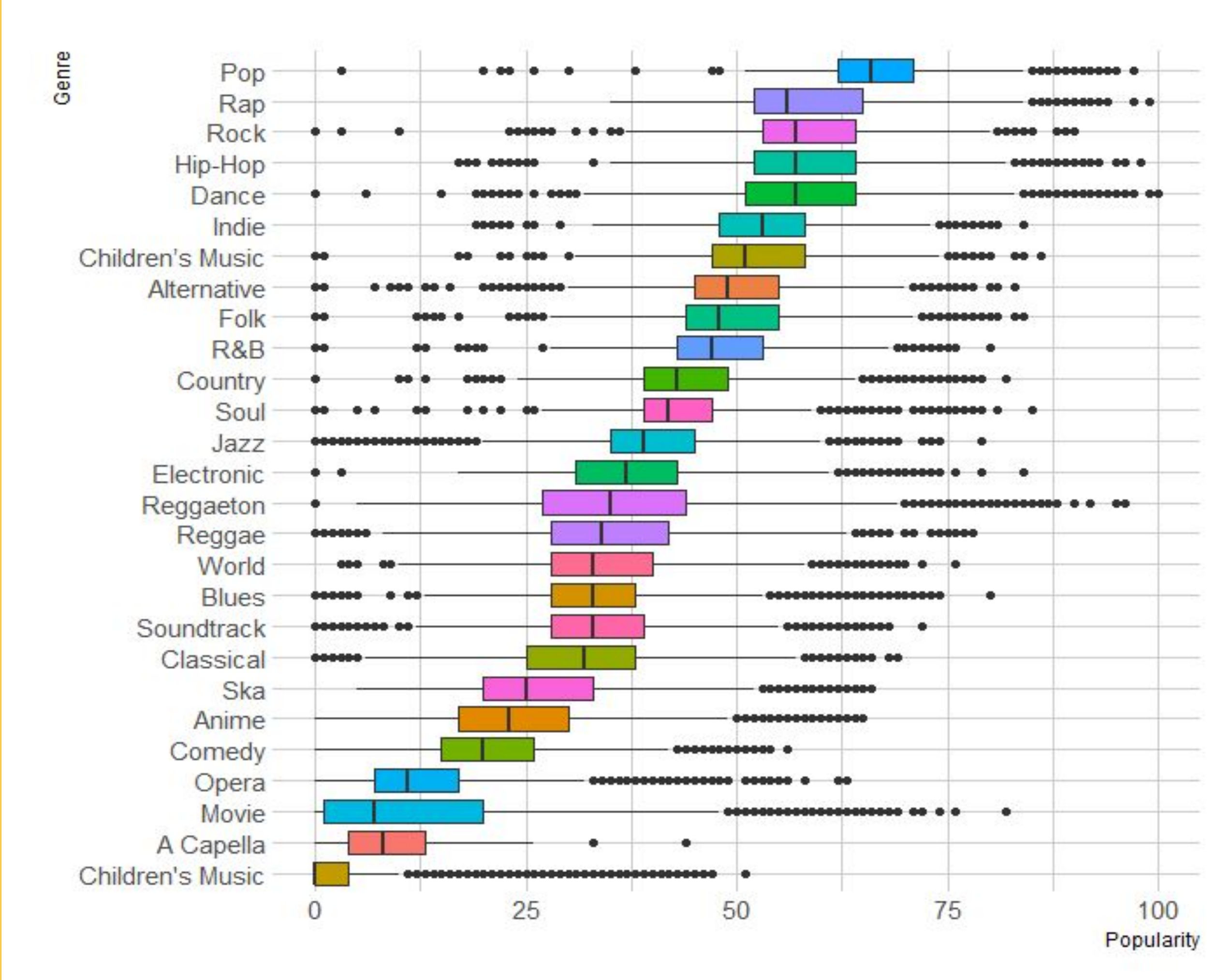
UNDERSTANDING THE DATA

What insights can we get?

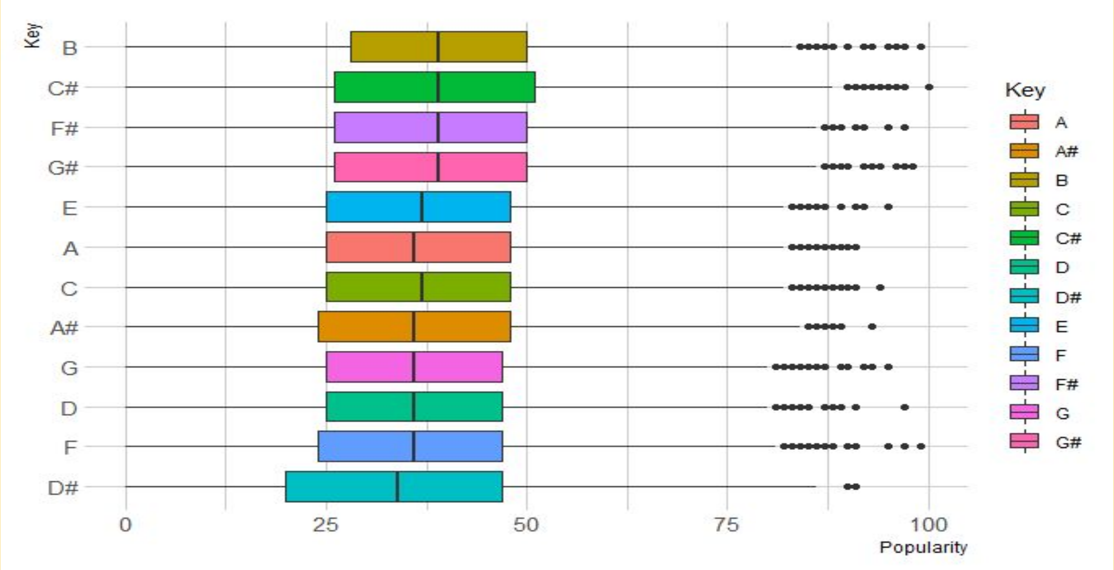


HOW DO THE ATTRIBUTES INFLUENCE EACH OTHER?

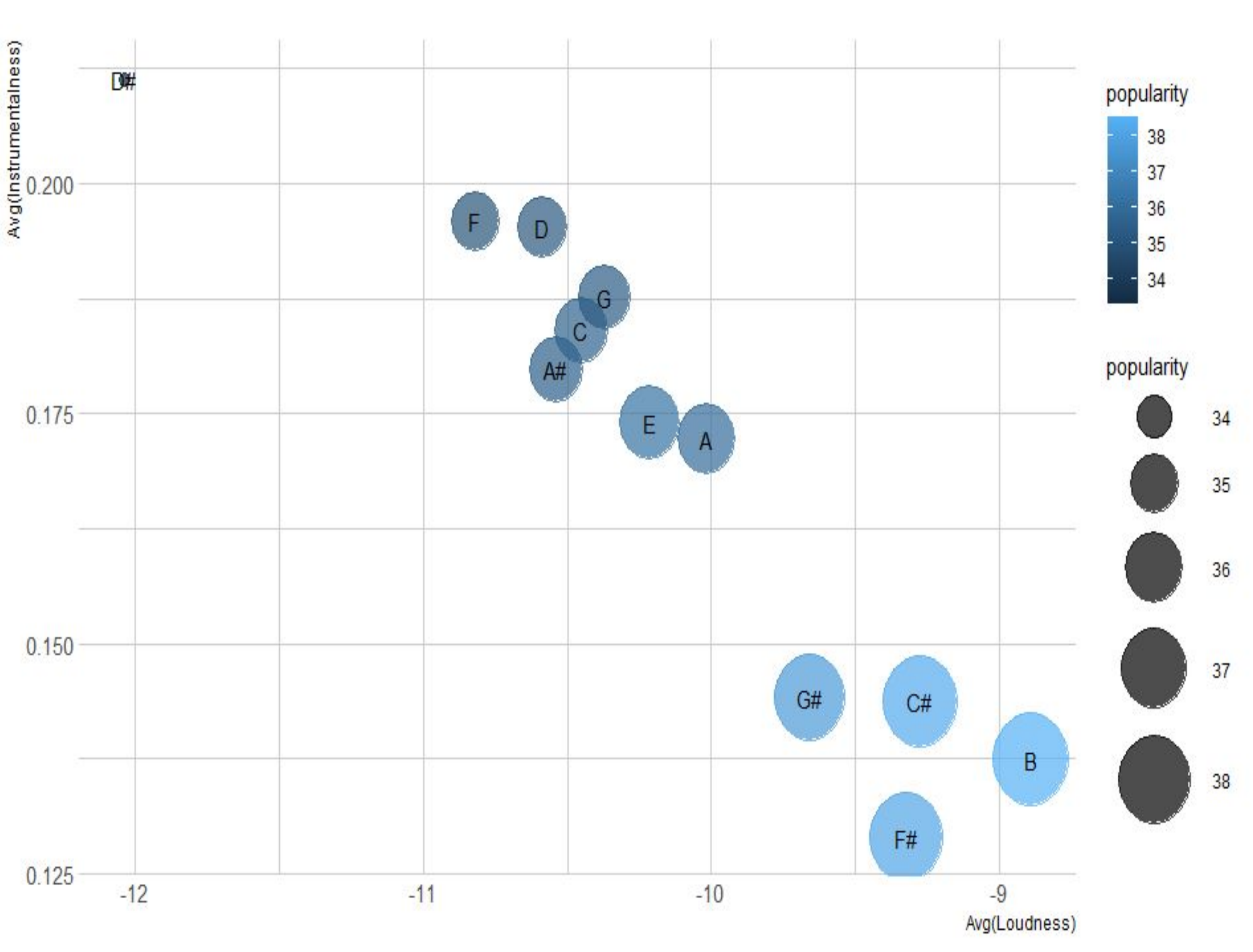
DOES POPULARITY VARY BY GENRE ?



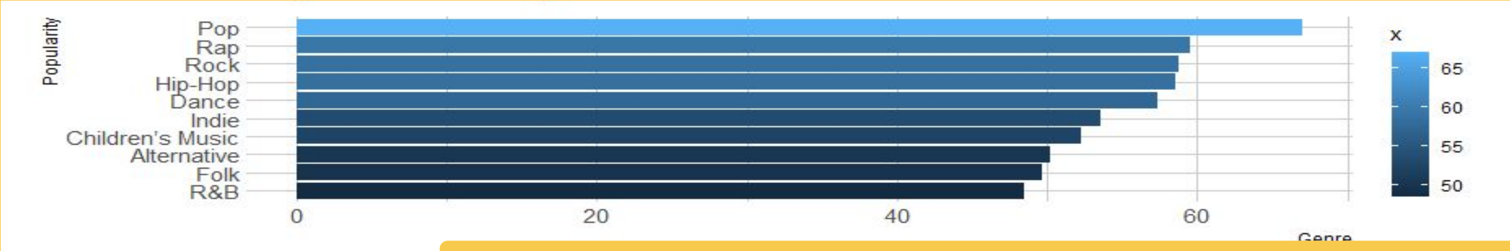
DOES POPULARITY VARY BY KEY ?



DOES POPULARITY VARY BY MODE ?

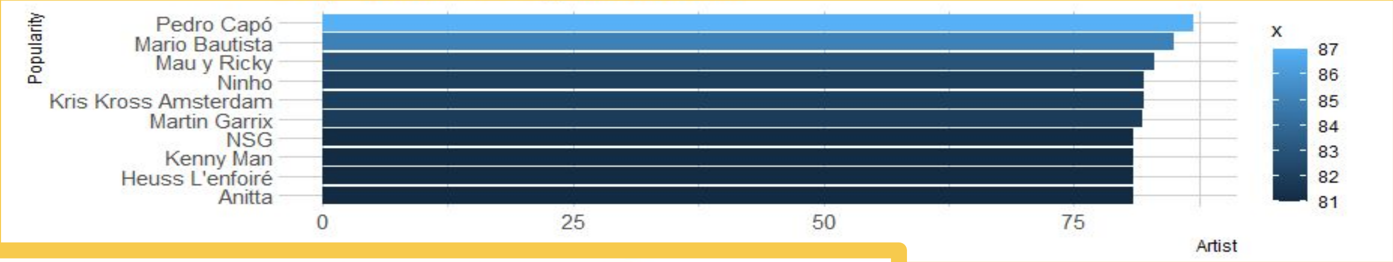


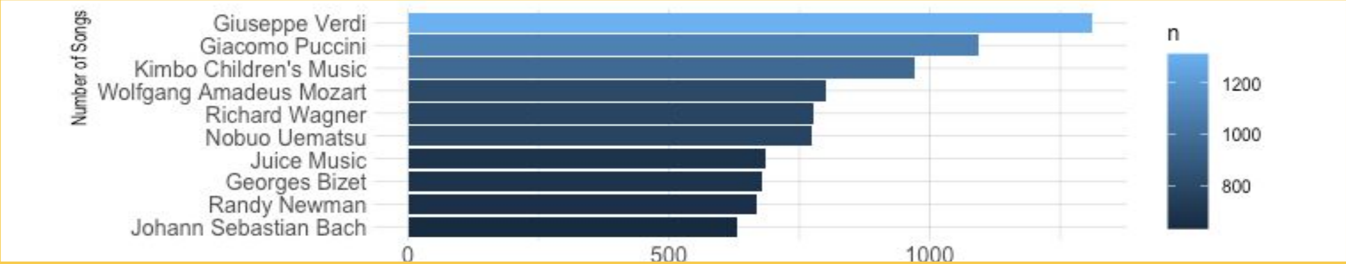
"LOUD MUSIC IS MORE POPULAR"



TOP 10 GENRES

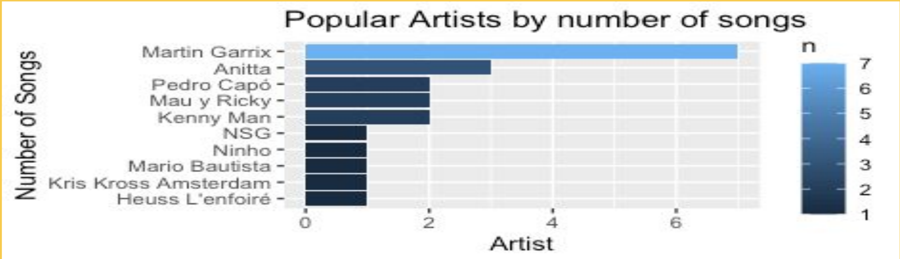
TOP 10 ARTISTS





ARTISTS WITH MOST SONGS

NUMBER OF SONGS BY POPULAR ARTISTS



TYPES OF MODELING

01
MULTIPLE LINEAR REGRESSION

02
LOGISTIC REGRESSION

03
TEXT MINING

04
K-MEANS CLUSTERING

01

MULTIPLE LINEAR REGRESSION

1. MULTIPLE LINEAR REGRESSION

Goal: Predict the popularity score

Target Variable: Popularity

Predictors: acousticness, danceability, energy, instrumentalness, liveness, loudness, and speechiness

Training/Validation Split: 80/20

RESULTS : Low Adjusted R Squared - 0.2127

```
Call:
lm(formula = popularity ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-54.399 -10.229   1.738  11.290  57.870 

Coefficients:
                 Estimate Std. Error t value              Pr(>|t|)    
(Intercept)      60.02153    0.33151  181.05 <0.0000000000000002 ***
acousticness    -13.63873    0.17268  -78.98 <0.0000000000000002 ***
danceability      8.31153    0.23770   34.97 <0.0000000000000002 ***
energy          -12.97420    0.29986  -43.27 <0.0000000000000002 ***
instrumentalness -3.08042    0.14961  -20.59 <0.0000000000000002 ***
loudness          0.83271    0.01263   65.92 <0.0000000000000002 ***
speechiness      -5.11350    0.25682  -19.91 <0.0000000000000002 ***
liveness         -9.36545    0.22821  -41.04 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.15 on 186172 degrees of freedom
Multiple R-squared:  0.2121,	Adjusted R-squared:  0.2121 
F-statistic:  7159 on 7 and 186172 DF,  p-value: < 0.00000000000000022
```

MLR Results

RESULTS : **High RSME:** 16.13419 compared to a 0-100 scale

```
> accuracy(song.lm.pred, valid$popularity)
                   ME     RMSE      MAE  MPE MAPE
Test set -0.05130541 16.14189 12.86404 -Inf  Inf
```

Accuracy of the model

Conclusion: using Multiple Linear Regression on the selected variables to predict the popularity score is not accurate, and other models need to be examined.

02

LOGISTIC REGRESSION

2. LOGISTIC REGRESSION

Target variable: Song Popularity (whether a song is popular or not)

Predictors: acousticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, valence

Goal: Identify the factors that contribute towards the song's popularity

Pre-Processing: We pre-processed the data by converting the Popularity column in the dataset to 0 and 1, 1 being "Popular" and 0 being "Not Popular".

OBSERVATION

```
Call:
glm(formula = Is_Popular ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3904  -0.6073  -0.4289  -0.1877   3.7522

Coefficients:
                      Estimate    Std. Error z value             Pr(>|z|)
(Intercept)       -0.1957052718  0.1006673135  -1.944              0.05189 .
acousticness      -0.7070940134  0.0401196741 -17.625 < 0.0000000000000002 ***
danceability       2.3810938330  0.0627864305  37.924 < 0.0000000000000002 ***
duration_ms       -0.0000009678  0.0000001131  -8.556 < 0.0000000000000002 ***
energy            -1.4339711807  0.0735557197 -19.495 < 0.0000000000000002 ***
instrumentalness -1.5956301672  0.0498133919 -32.032 < 0.0000000000000002 ***
key2               0.0592627925  0.0422590923   1.402              0.16081
key3               0.0883639034  0.0395501386   2.234              0.02547 *
key4               0.0377221847  0.0365293963   1.033              0.30177
key5               0.2058704643  0.0366323076   5.620         0.0000000191 ***
key6               0.0497871583  0.0379229701   1.313              0.18923
key7               0.1585635674  0.0548560269   2.891              0.00385 **
key8               0.0361185324  0.0409980836   0.881              0.37833
key9               0.0896308906  0.0392637974   2.283              0.02244 *
key10              0.2159216409  0.0406790104   5.308         0.0000001109 ***
key11             -0.0079608599  0.0370231707  -0.215              0.82975
key12              0.1916688635  0.0409338776   4.682         0.0000028353 ***
liveness          -0.9270870498  0.0563717878 -16.446 < 0.0000000000000002 ***
loudness           0.1261704032  0.0038622561  32.668 < 0.0000000000000002 ***
mode0              0.1024283634  0.0179884589   5.694         0.0000000124 ***
speechiness       -0.7855579609  0.0709314751 -11.075 < 0.0000000000000002 ***
tempo              0.0012440343  0.0002897661   4.293         0.0000176087 ***
valence           -1.1401423078  0.0409157134 -27.866 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109393  on 141418  degrees of freedom
Residual deviance:  97207  on 141396  degrees of freedom
AIC: 97253

Number of Fisher Scoring iterations: 6
```

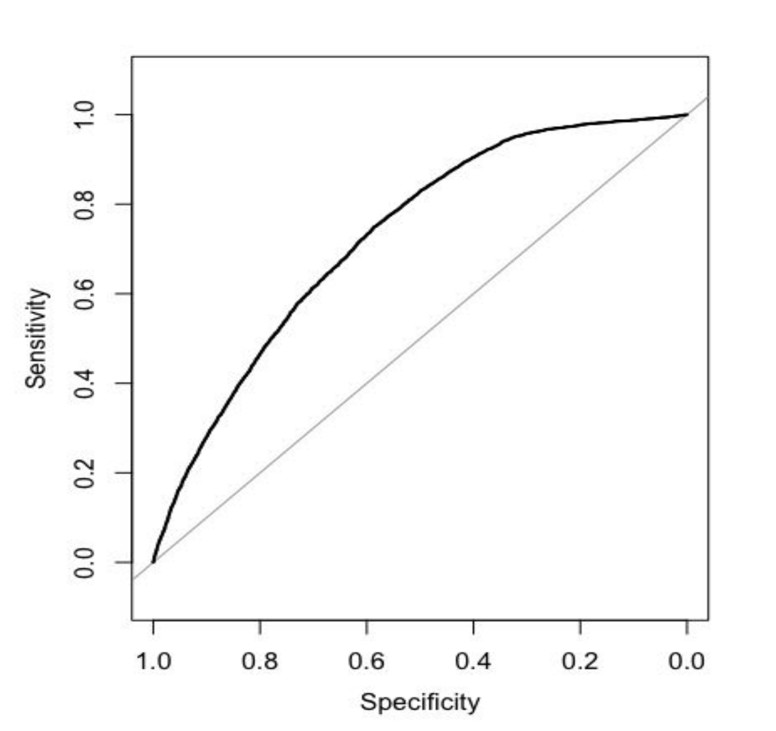
Looking at the Logistic Regression logit Coefficients we can infer that almost all the audio based metrics are equally significant in predicting the popularity of the song.

CONFUSION MATRIX

p=0.5	Actual Class		
Predicted Class		0	1
	0	30578	4736
	1	25	16

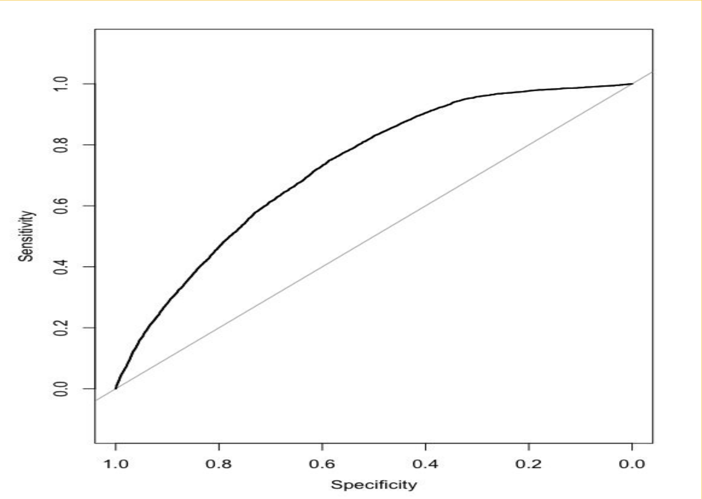
Sensitivity = 0.0033670
Specificity = 0.9991831
Accuracy = 86.53%

ROC CURVE



Threshold at 0.5

IDEAL THRESHOLD



Specificity 0.5869359

Sensitivity 0.7485269

At the ideal threshold of 0.1353475 we were able to get best sensitivity and specificity

INFERENCE

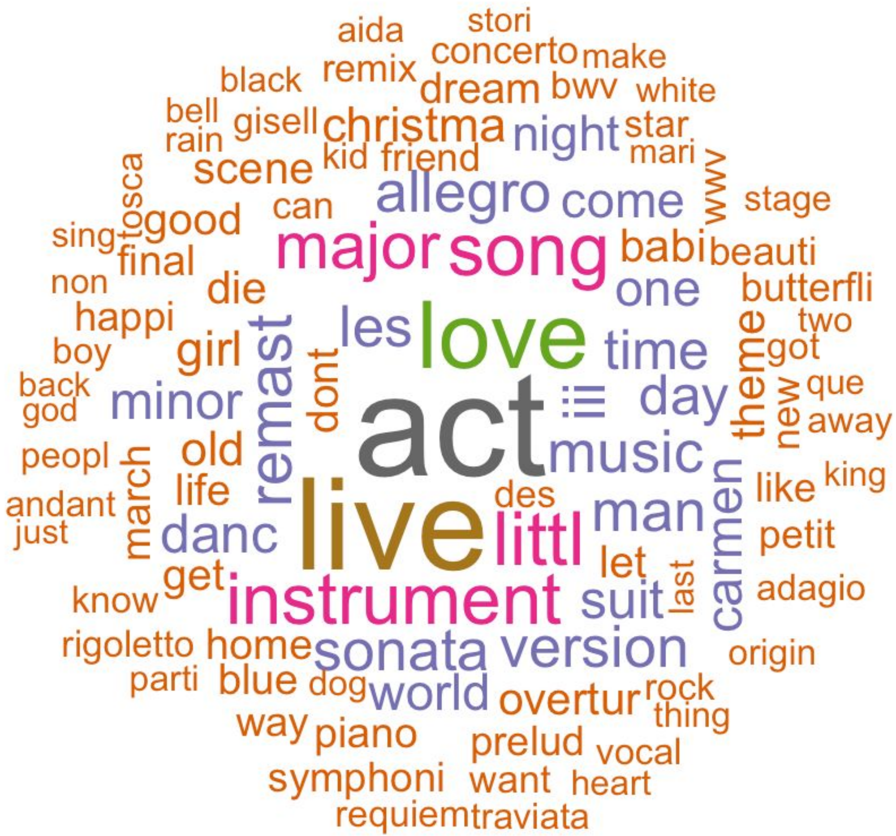
As per our logistic regression model we inferred that:

- Almost all the audio based metrics had major influence in predicting the song's popularity
- To increase the sum of sensitivity and specificity in the threshold value can be reduced in our model
- Songs with lower acousticness, higher danceability, lower liveness, lower speechiness, lower instrumentalness and energy tend to be more popular

03

TEXT MINING

stargaz moonlight
now came
pump shallow
america life
passionfruit
everyday
unforgett
sola
savag
havana come
pure
know
cant uproar
rise
undecid
sky ride
hill danc
halsey kod
swish
bad
tire
talk
dont ill
without reel
new
end
light
exchang let
feel
sorri lot
sucker
lil
creep
adictiva
scrub
star
one
dura japan
feat
hope
letter
nasa
flame
polaroid
middl walk
love like
stay dna
rehúso
take
mask
remix brainer
babi
nicki wake
look
imagin
callao umbrella
better call
play
close
never girl
god bamba
crush
plan
happier
stylist
nicemonster
make
humbl
minaj
sauc
motorsport
starboy
trip
remedi
acuerdo
woman



aida
stori
concerto
make
remix
black
dream
bwv
white
bell
gisell
christma
night
star
rain
kid
friend
mari
scene
tosca
allegro
come
www
stage
can
sing
good
major
song
babi
beauti
final
non
die
one
butterfli
happi
les
love
time
two
boy
girl
remast
dont
got
theme
back
god
minor
act
ii
day
new
que
away
peopl
old
music
march
life
des
like
king
andant
just
danc
live
littl
man
carmen
petit
get
let
know
instrument
suit
last
adagio
rigoletto
home
sonata
version
origin
parti
blue
dog
world
overtur
rock
thing
way
piano
prelud
vocal
symphoni
want
heart
requiem
traviata

04

K-MEANS CLUSTERING

CLUSTERING - LOW VS HIGH POPULARITY

Cluster	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence
1	-1.23	1.35	-0.97	-0.09	-1.36	-0.36	0.00	-1.18	-0.31	-0.42	-0.79
3	-0.90	1.08	0.11	-0.11	0.40	-0.53	2.42	-0.29	3.71	-0.62	-0.15
2	-0.77	0.81	-0.74	12.54	-0.64	0.57	0.34	-0.82	0.87	-0.58	-0.56
8	-0.66	-0.29	0.84	-0.22	0.33	-0.46	-0.21	0.35	-0.17	-0.01	1.21
7	-0.41	1.28	-1.41	0.13	-1.51	2.07	-0.41	-1.81	-0.41	-0.51	-1.15
5	-0.12	-0.31	0.15	0.12	0.16	1.91	-0.24	0.04	-0.32	0.13	-0.05
6	0.09	-0.86	-0.35	0.05	0.89	-0.43	0.29	0.72	-0.15	1.01	0.05
9	0.51	0.52	-0.09	0.07	-0.63	-0.44	-0.32	-0.06	-0.33	-0.11	-0.47
4	0.98	-0.72	0.73	-0.09	0.52	-0.49	-0.23	0.60	-0.08	-0.22	0.34

- **Low popularity** clusters are characterized by extreme highs and lows of attributes (darker colors) such as acousticness, danceability, energy, loudness, and valence.
- **High popularity** clusters have less extreme attributes
- **Insight -** Spotify could use this analysis to provide recommendations to new artists to be careful of releasing songs too extreme in any one attribute.

GENRE CLUSTER GROUPS

("RAP", "INDIE", "ROCK", "JAZZ", "ELECTRONIC", "HIP-HOP","ALTERNATIVE","POP")

Cluster	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence
5	1.05	-0.55	1.11	-0.13	0.37	-0.47	-0.25	0.50	0.10	0.00	0.69
4	0.13	-0.56	0.44	0.43	0.36	1.85	-0.22	0.19	-0.29	0.10	-0.01
3	0.85	-0.77	0.11	0.03	0.55	-0.44	-0.03	0.58	-0.12	-0.59	-0.38
2	0.59	0.99	-0.14	0.07	-0.96	0.18	-0.31	-0.52	-0.33	-0.21	-0.46
1	0.72	-0.86	-0.14	-0.02	0.83	-0.39	0.01	0.72	-0.03	1.20	-0.14

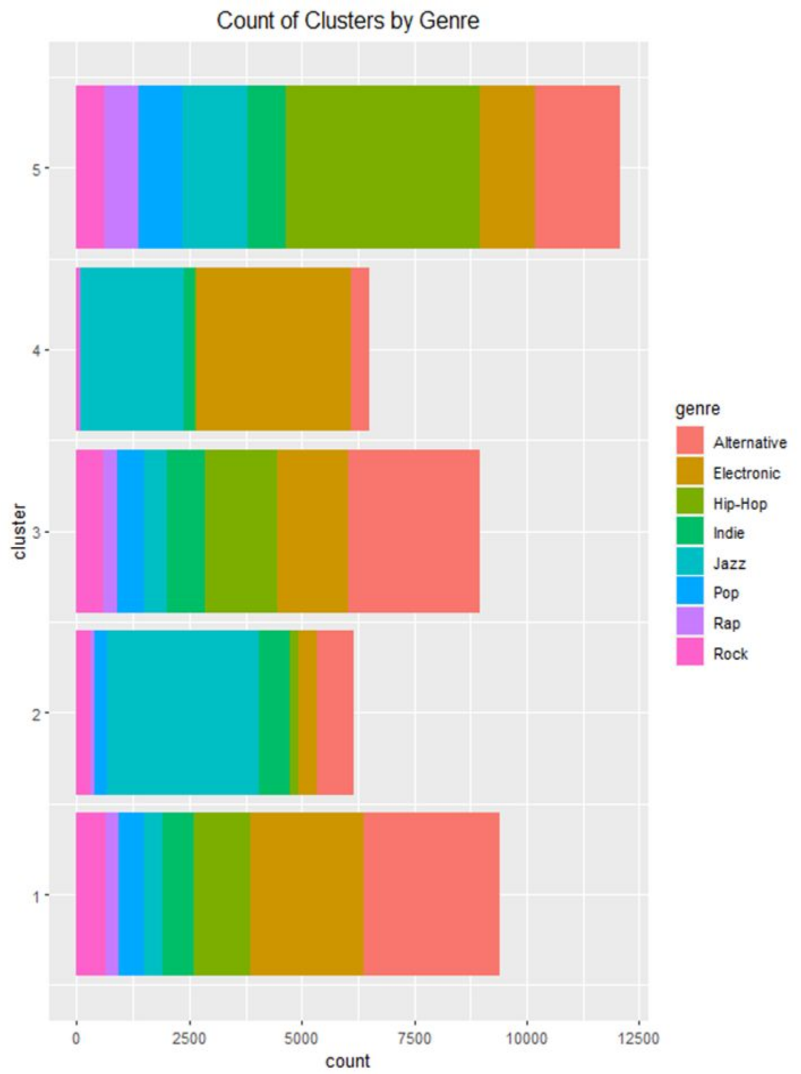
5) High popularity, high danceability, moderately high valence

4) Lower popularity, high instrumentalness

3) High popularity, low acousticness, low tempo

2) Moderate popularity, high acousticness, low energy

1) High popularity, low acousticness, high energy, high loudness, high tempo



- **#5) Hip-Hop/Rap** featuring artists such as Drake, Eminem, J. Cole, Kid Cudi, Kevin Gates, Lil Baby and Future.
- **#4) Electronic/Jazz** featuring artists such as Aphex Twin, Bonobo, Nujabes, Gramatik and Flying Lotus.
- **#3) Alternative** featuring artists such as Brock Hampton, Childish Gambino, Kendrick Lamar, Imagine Dragons and Dillon Francis.
- **#2) Jazz** featuring artists such as Café Jazz Deluxe, Dean Martin, John Coltrane and Ella Fitzgerald.
- **#1) Rock & EDM** featuring artists such as Arctic Monkeys, Kings of Leon, Bassnectar, Flosstradamus and Fall Out Boy.

05

CHALLENGES

WHAT WERE SOME CHALLENGES?

DUPLICATES

A track could have multiple genres resulting in duplicate data

PULLED DATA

Focus was to get 10,000 songs from 26 genres

GENRE CLUSTERING

Had to subset genres and perform multiple clusters

06

CONCLUSION

KEY INSIGHTS & TAKEAWAYS

MLR

Exact popularity score is hard to predict based on the song attributes

POPULAR SONGS

tend to have multiple artists or be remixes

AUDIO-BASED METRICS

had major influence in predicting the song's popularity

GENRES

can be clustered into groups based on Spotify API attributes

IF YOU ARE AN ARTIST...

TOPIC

Love songs about your feelings

FIND SOMEONE

Have remixes and Features

AUDIO-BASED METRICS

Focus on higher loudness, lower intrumentalness

GENRES

Stick to pop/rap

THANK YOU!

QUESTIONS?