SC1015 MINI PROJECT

Group:
Law Ming Han
Yin Jian
Toh Zhi Yang

TABLE OF CONTENTS

01 PROBLEM DEFINITION

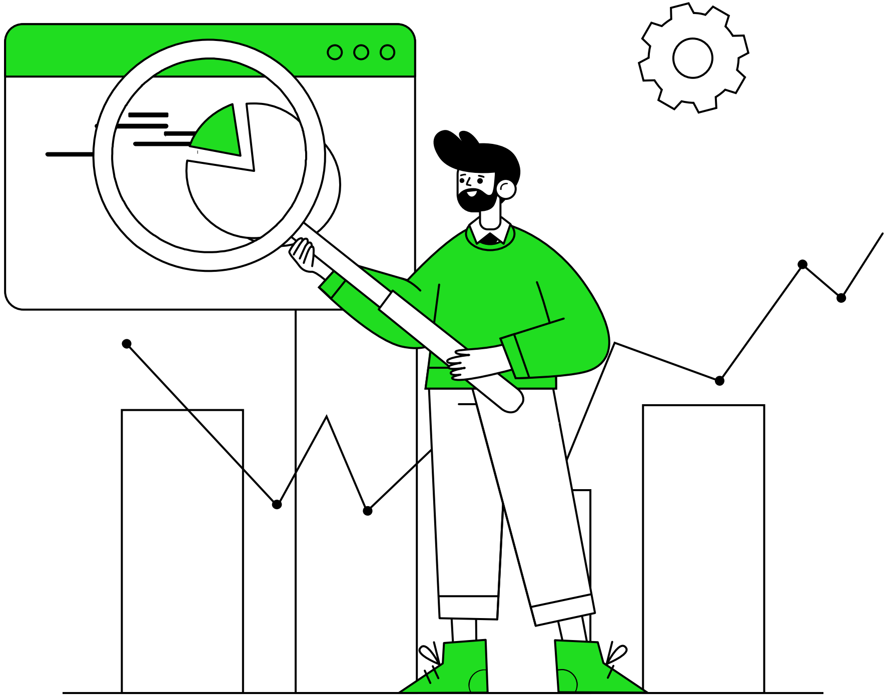
02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

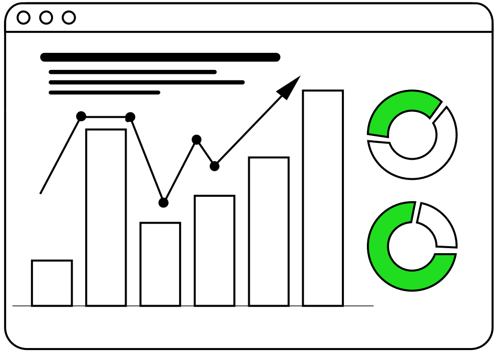
01 PROBLEM DEFINITION



BACKGROUND OF MUSIC INDUSTRY

AS OF 2021, THERE ARE **1.2 MILLION ARTISTS** STREAMING ON SPOTIFY.

THIS IS A GROWING PLATFORM, AS OBSERVED FROM THE ENTRY OF **150,000 ARTISTS** ONTO SPOTIFY IN 2020.



01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

PRACTICAL MOTIVATION

- ASSIST NEW ARTISTS TO INCREASE THEIR LIKELIHOOD OF RELEASING A HIT SONG
- ASSIST ESTABLISHED ARTISTS TO REPLICATE THEIR PREVIOUS HIT SONGS

PROBLEM DEFINITION

OUR GROUP INTENDS TO INVESTIGATE THE RELATIONSHIP BETWEEN AUDIO FEATURES AND POPULARITY OF SPOTIFY TRACKS.

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

02

DATA PREPARATION AND CLEANING

DATA PREPARATION

ACCESSING TO THE SPOTIFY'S API

```
import spotipy
from spotipy.oauth2 import SpotifyClientCredentials
cid = '6b2418c9674f4f7c9f5e2809aa2b3678'
secret = '781801fad3b646c6b7844583a170957c'
client_credentials_manager = SpotifyClientCredentials(client_id=cid, client_secret=secret)
sp = spotipy.Spotify(client_credentials_manager
=
client_credentials_manager)
```

EXTRACTING RELEVANT DATA FROM THE API

```
for yr in range(20):
    for i in range(20):
        track_results = sp.search(q='year:{}'.format(years[yr]), type='track', limit=50,offset=i*50)
        print(track_results)
        for i, t in enumerate(track_results['tracks']['items']):
            artist_name.append(t['artists'][0]['name'])
            track_name.append(t['name'])
            track_id.append(t['id'])
            popularity.append(t['popularity'])
```

DATA PREPARATION

EXTRACTING AUDIO FEATURES BASED ON TRACK_ID

```
ls = list(df['track_id'])
audio_features = []
for ids in range(0, 19400, 100):
    audio_features.append(sp.audio_features(ls[ids:ids+100]))

audio_features.append(sp.audio_features(ls[19400:19419]))
```

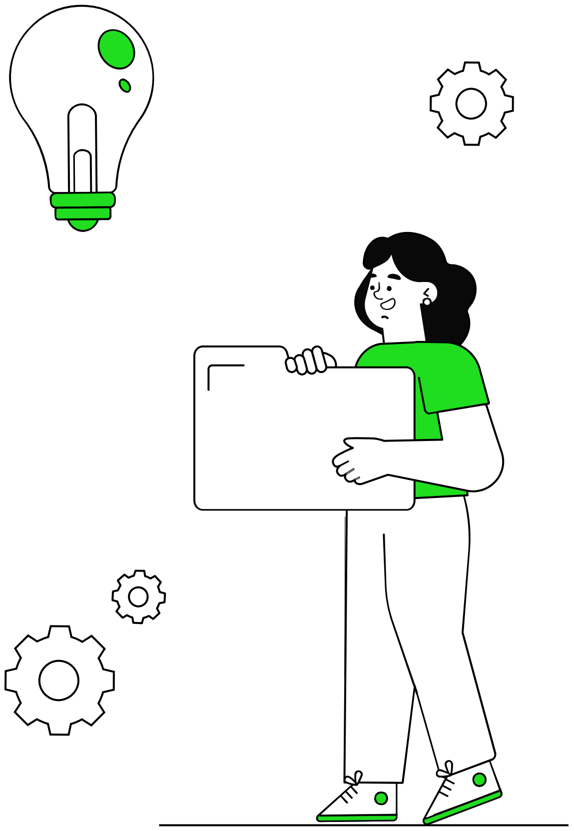
```
for i in range(0, 194):
    for j in range(100):
        if audio_features[i][j] != None:
            danceability.append(audio_features[i][j]['danceability'])
            energy.append(audio_features[i][j]['energy'])
            key.append(audio_features[i][j]['key'])
            loudness.append(audio_features[i][j]['loudness'])
            mode.append(audio_features[i][j]['mode'])
            speechiness.append(audio_features[i][j]['speechiness'])
            acousticness.append(audio_features[i][j]['acousticness'])
            instrumentalness.append(audio_features[i][j]['instrumentalness'])
            liveness.append(audio_features[i][j]['liveness'])
```

CLEANING

- REMOVING DUPLICATES
  - SAME TRACK_ID
- REMOVAL OF NONE VALUES FOR AUDIO FEATURES DATA
  - REPLACE NONE VALUES WITH EXTREME VALUE FOR EASE OF REMOVAL WHEN COMBINED WITH FINAL DATA SET
- EXPORTING THEM INTO A CSV FILE FOR FURTHER EDA

03

EDA AND VISUALISATION



RESPONSE VARIABLE

POPULARITY : BASED ON THE TOTAL NUMBER OF PLAYS THE TRACK HAS HAD AND HOW RECENT THOSE PLAYS ARE

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

PREDICTOR VARIABLES (NUMERIC)

DANCEABILITY : HOW SUITABLE A TRACK IS FOR DANCING

ENERGY : MEASURE OF INTENSITY AND ACTIVITY

LOUDNESS : THE OVERALL LOUDNESS OF A TRACK IN DECIBELS.

SPEECHINESS : PRESENCE OF SPOKEN WORDS IN A TRACK.

ACOUSTICNESS : MEASURE OF WHETHER THE TRACK IS ACOUSTIC.

INSTRUMENTALNESS : PREDICTS WHETHER A TRACK CONTAINS NO VOCALS.

LIVENESS : DETECTS THE PRESENCE OF AN AUDIENCE IN THE RECORDING.

TEMPO : OVERALL ESTIMATED TEMPO OF A TRACK IN BEATS PER MINUTE

DURATION : DURATION OF THE TRACK IN MINUTES

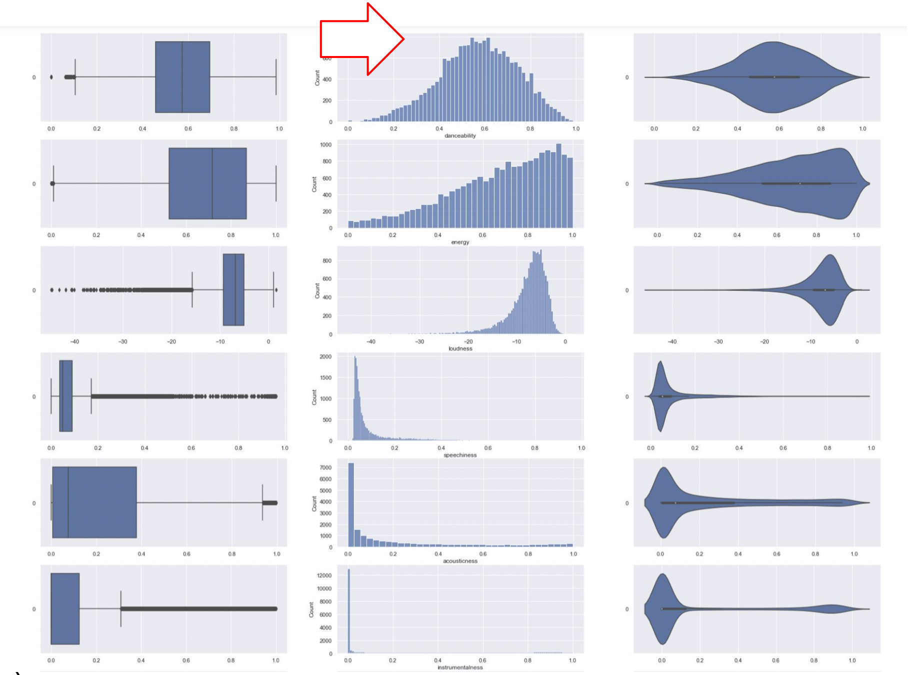
PREDICTOR VARIABLES (CATEGORICAL)

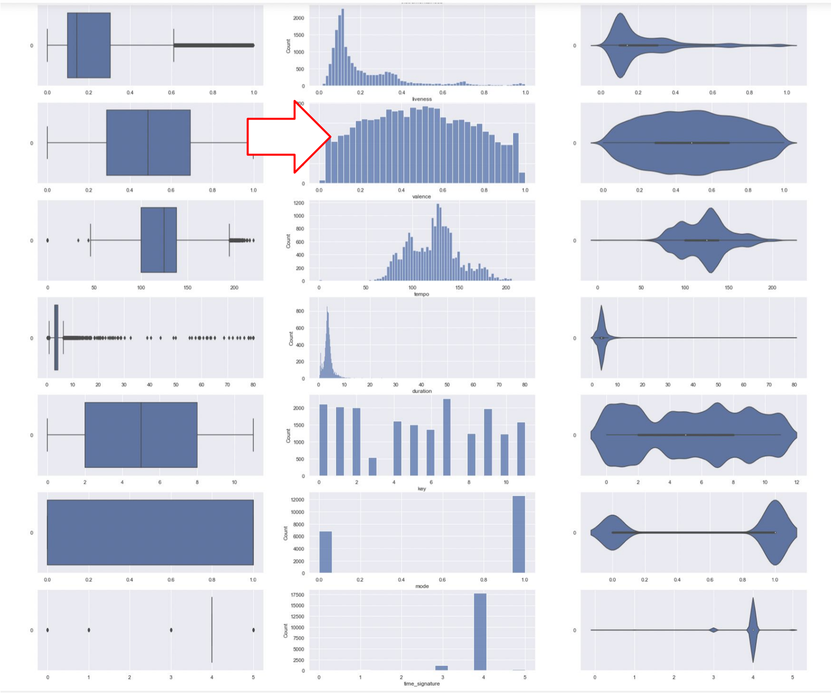
MODE : MAJOR (1) OR MINOR (0) SCALE

KEY : THE KEY THE TRACK IS IN

TIME SIGNATURE : AN ESTIMATION OF HOW MANY BEATS PER BAR

DISTRIBUTION OF ALL VARIABLES (NUMERIC & CATEGORICAL)





01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

SUMMARY STATISTICS (NUMERIC)

	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration
count	19416.000000	19416.000000	19416.000000	19416.000000	19416.000000	19416.000000	19416.000000	19416.000000	19416.000000	19416.000000
mean	0.570222	0.674815	-7.838767	0.090111	0.230815	0.177834	0.229087	0.490992	122.719028	3.953030
std	0.169750	0.233090	4.320179	0.105160	0.297112	0.325169	0.201796	0.252063	27.946602	3.010822
min	0.000000	0.000020	-44.761000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.290000
25%	0.458000	0.525000	-9.351500	0.036100	0.005130	0.000000	0.098300	0.287000	100.253250	3.030000
50%	0.575000	0.716000	-6.844500	0.049900	0.074900	0.000087	0.141000	0.487000	124.893000	3.690000
75%	0.694000	0.868000	-5.080000	0.090700	0.378000	0.124250	0.304000	0.693000	138.056000	4.410000
max	0.985000	1.000000	1.526000	0.961000	0.996000	1.000000	0.996000	0.998000	220.099000	80.000000

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

CORRELATION BETWEEN NUMERIC VARIABLES AND POPULARITY

	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration	popularity
danceability	1.000	0.014	0.194	0.157	-0.109	-0.163	-0.195	0.409	-0.119	-0.034	0.061
energy	0.014	1.000	0.699	0.038	-0.698	0.004	0.129	0.267	0.204	-0.035	-0.004
loudness	0.194	0.699	1.000	-0.001	-0.543	-0.341	-0.047	0.264	0.138	-0.046	0.063
speechiness	0.157	0.038	-0.001	1.000	0.010	-0.115	0.132	0.030	0.011	-0.009	0.028
acousticness	-0.109	-0.698	-0.543	0.010	1.000	0.013	-0.001	-0.124	-0.174	-0.008	-0.032
instrumentalness	-0.163	0.004	-0.341	-0.115	0.013	1.000	0.058	-0.194	0.020	0.059	-0.057
liveness	-0.195	0.129	-0.047	0.132	-0.001	0.058	1.000	-0.050	-0.019	0.019	-0.036
valence	0.409	0.267	0.264	0.030	-0.124	-0.194	-0.050	1.000	0.046	-0.138	-0.007
tempo	-0.119	0.204	0.138	0.011	-0.174	0.020	-0.019	0.046	1.000	-0.012	0.004
duration	-0.034	-0.035	-0.046	-0.009	-0.008	0.059	0.019	-0.138	-0.012	1.000	-0.003
popularity	0.061	-0.004	0.063	0.028	-0.032	-0.057	-0.036	-0.007	0.004	-0.003	1.000

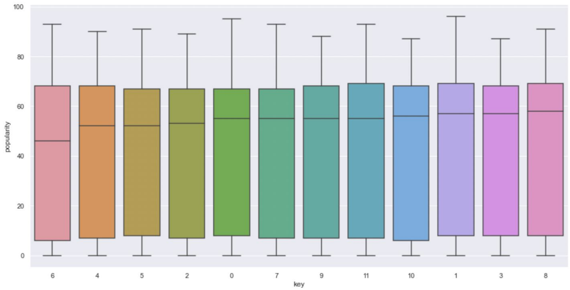
INSIGHTS: NONE OF THE NUMERIC VARIABLES HAVE A STRONG LINEAR RELATIONSHIP WITH POPULARITY.

SUMMARY STATISTICS (CATEGORICAL)

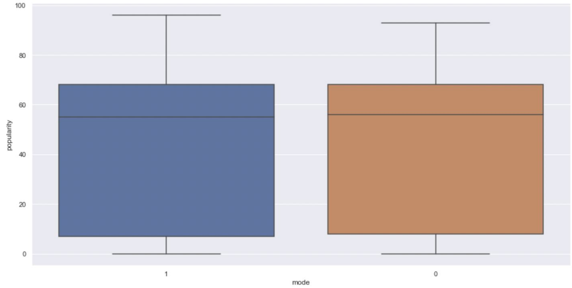
	key	mode	time_signature
count	19416.000000	19416.000000	19416.000000
mean	5.297384	0.648795	3.929440
std	3.555519	0.477359	0.367445
min	0.000000	0.000000	0.000000
25%	2.000000	0.000000	4.000000
50%	5.000000	1.000000	4.000000
75%	8.000000	1.000000	4.000000
max	11.000000	1.000000	5.000000

DISTRIBUTION OF CATEGORICAL VARIABLES AGAINST POPULARITY

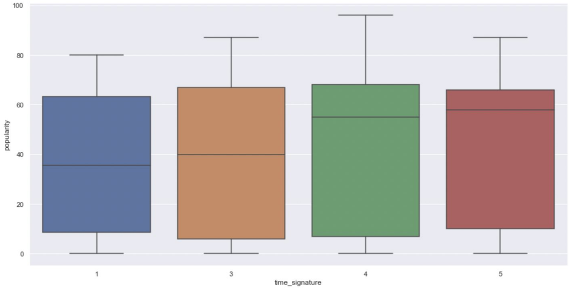
Key



Mode



Time Signature



INSIGHTS: ONLY TIME SIGNATURE INDICATE A RELATIONSHIP WITH POPULARITY.

REMOVAL OF OUTLIERS USING ISOLATION FOREST

- UNSUPERVISED LEARNING ALGORITHM THAT CAN ISOLATE OUTLIERS FROM A MULTI-DIMENSIONAL DATASET EFFECTIVELY.
- EXPLOITS THE NATURE THAT ANOMALIES ARE "FEW AND DIFFERENT".

```python
from sklearn.ensemble import IsolationForest

iForest = IsolationForest(n_estimators = 100, contamination = 0.05)

# fit model
iForest.fit(variables)

# predict on data
anomaly_mask = iForest.predict(variables) #anomalies will be masked as -1 in the array
print("number of anomalies identified:",anomaly_mask.tolist().count(-1)) #number of anomalies marked

temp = pd.concat([variables, popularity],axis=1).reindex(variables.index)

temp_wo = temp[(anomaly_mask != -1)].reset_index(drop=True)
print("new shape:",temp_wo.shape) # check the shape
```

```
number of anomalies identified: 970
new shape: (18446, 14)
```

ONE-HOT ENCODING

- AS THE CATEGORICAL VARIABLES MAY NOT BE *ORDINAL*, INTEGER ENCODING IS UNFEASIBLE.
- WE DECIDED TO ENCODE NOMINAL (UNORDERED) CATEGORICAL VARIABLES VIA ONE-HOT ENCODING

key_0	key_1	key_2	key_3	key_4	key_5	key_6	key_7	key_8	key_9	key_10	key_11	mode_0	mode_1	time_signature_1	time_signature_3	time_signa
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	

04

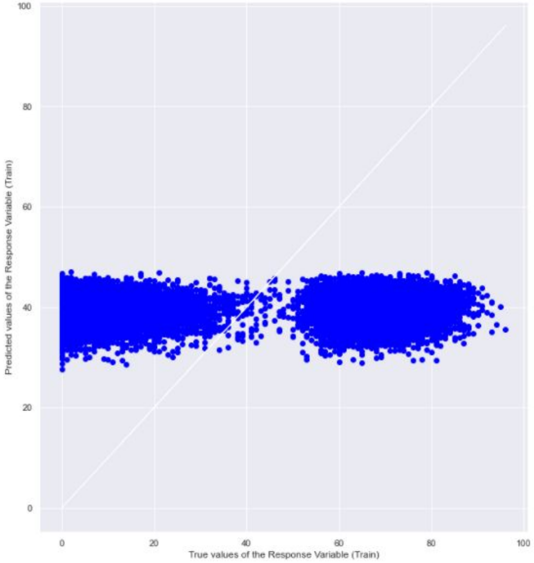
MACHINE LEARNING

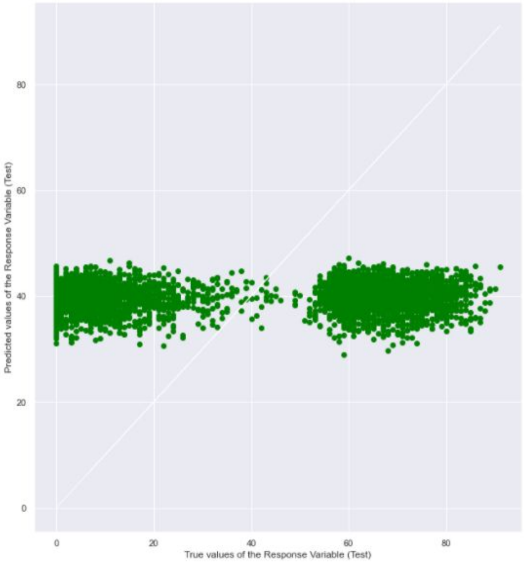
RATIONALE

Since response variable is numeric, we decided to carry out machine learning using multivariate linear regression model.

LINEAR REGRESSION MODEL I

Using top 6 numeric predictor variables






```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.008451468096549952
Mean Squared Error (MSE)        : 928.3012852883592

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.007836228489523145
Mean Squared Error (MSE)        : 943.7999077942966

Explained Variance of Model 1 with CV:    0.006873016405327470
Standard Deviation of scores:    0.0056686750240689575
```

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

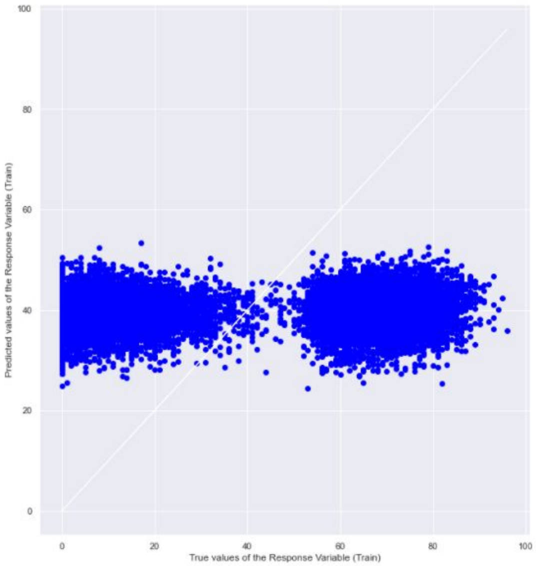
03 EDA & VISUALISATION

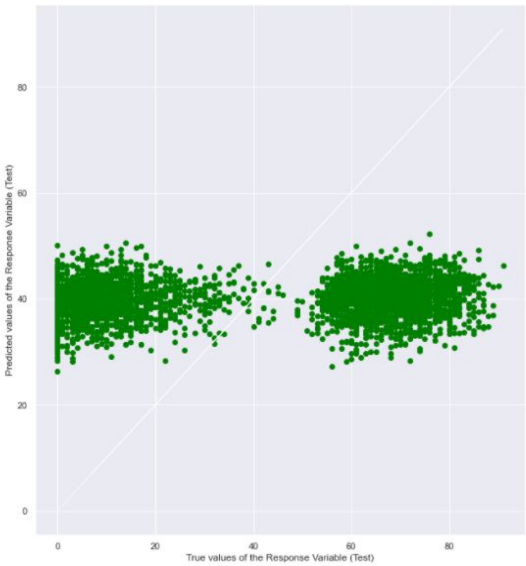
04 MACHINE LEARNING

05 CONCLUSION

LINEAR REGRESSION MODEL II

Using all numeric predictor variables






```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.014056852708758450
Mean Squared Error (MSE)        : 923.0534476156436

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.011229534582703926
Mean Squared Error (MSE)        : 940.5720112818224
```

```
Explained Variance of Model 2 with CV:   0.01181121680073377
Standard Deviation of scores:    0.007154644542709526
```

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

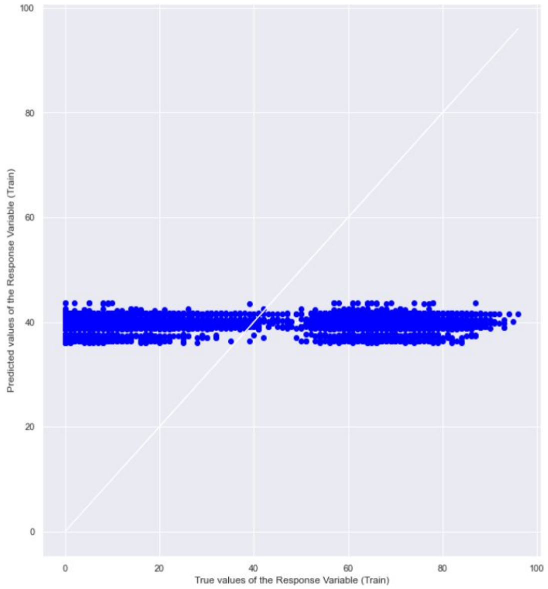
03 EDA & VISUALISATION

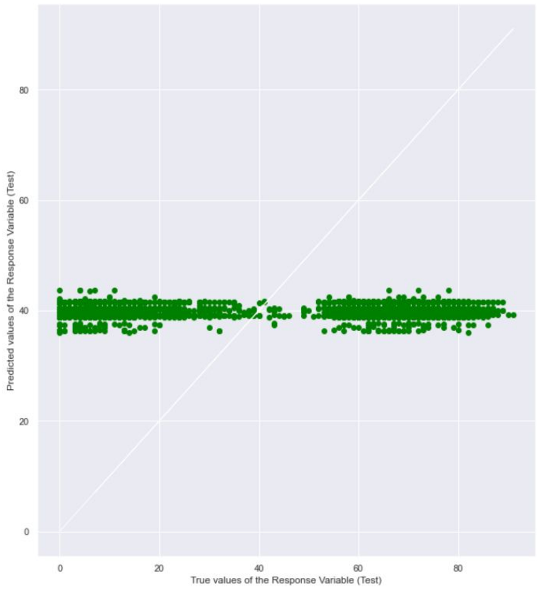
04 MACHINE LEARNING

05 CONCLUSION

LINEAR REGRESSION MODEL III

Using all categorical predictor variables



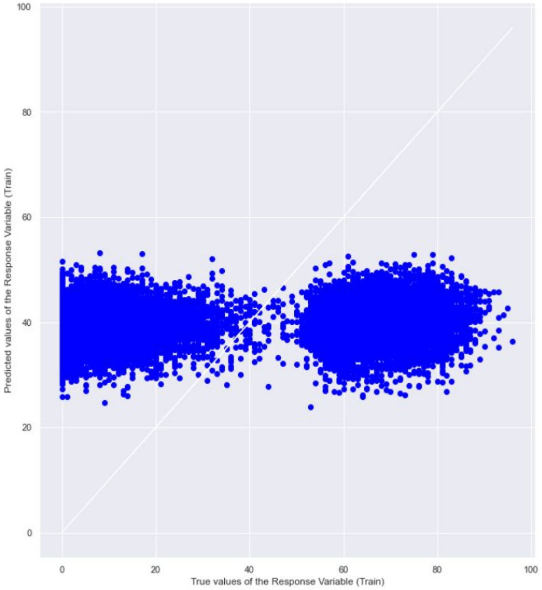


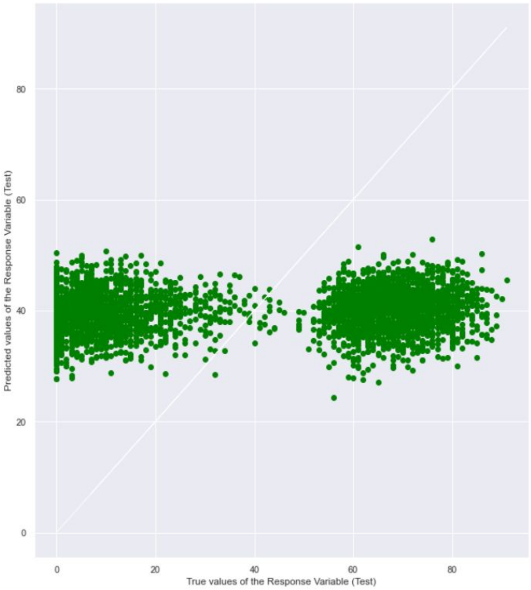

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.0013309442827409423
Mean Squared Error (MSE)        : 934.9676169863106

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : -0.0006597493138613686
Mean Squared Error (MSE)        : 951.8817419608972
```

```
Explained Variance of Model 3 with CV:   -0.0014752871233808195
Standard Deviation of scores:    0.0022211904659367654
```

LINEAR REGRESSION MODEL IV

Using all (numeric and categorical) variables






```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.015190483566692947
Mean Squared Error (MSE)        : 921.9921269151398

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.012360285011229322
Mean Squared Error (MSE)        : 939.4963802410358
```

```
Explained Variance of Model 4 with CV:   0.010445790529205346
Standard Deviation of scores:    0.00775455993960656
```

CROSS-VALIDATION SCORES (LINEAR REGRESSION)

	MEAN (EXPLAINED VARIANCE)	STANDARD DEVIATION
MODEL I	0.0069	0.0057
MODEL II	0.0118	0.0072
MODEL III	-0.0015	0.0022
MODEL IV	0.0104	0.0078

RESULTS ANALYSIS (LINEAR REGRESSION)

	EXPLAINED VARIANCE		MEAN SQUARED ERROR	
	TRAIN	TEST	TRAIN	TEST
MODEL I	0.0085	0.0078	928.3	943.8
MODEL II	0.0141	0.0112	923.1	940.6
MODEL III	0.0013	-0.0007	935.0	951.9
MODEL IV	0.0152	0.0124	922.0	939.5

OUR INSIGHTS

PERFORMANCE OF MODELS: MODEL 2 > MODEL 4 > MODEL 1 > MODEL 3

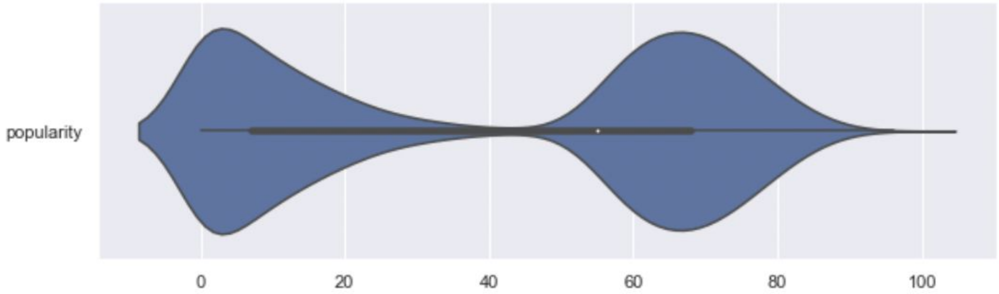
- SINCE ALL VARIABLES DO NOT EXHIBIT STRONG CORRELATION WITH THE POPULARITY VARIABLE

THE USAGE OF MORE VARIABLES GIVES A MORE ACCURATE PREDICTION.

- INTERESTING: THE DIFFERENCE IN EXPLAINED VARIANCE FOR MODEL 4 AND MODEL 2 IS NOTICEABLE

CATEGORICAL VARIABLES IN THE PREDICTION OF POPULARITY WORSENS MODEL IN LINEAR REGRESSION,
TOO MUCH VARIABLES, LEADING TO A TOO COMPLEX MODEL

EXPLORING OTHER MODELS



- GIVEN THE BIMODAL DISTRIBUTION ABOVE, WE CLASSIFY POPULARITY INTO TWO CLASSES:
  - TRUE: POPULARITY >= 50
  - FALSE: POPULARITY < 50

- NO CLASS IMBALANCE

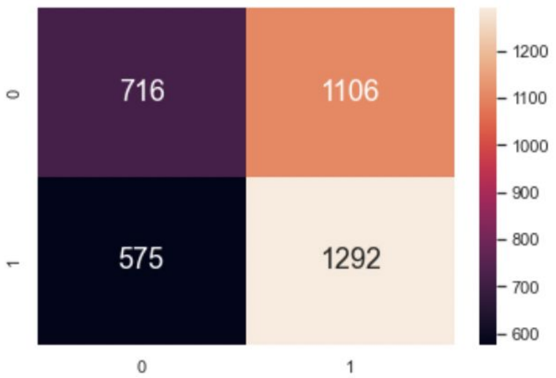
RATIONALE

SINCE RESPONSE VARIABLE IS NOW CATEGORICAL, WE DECIDED TO CARRY OUT MACHINE LEARNING USING DECISION TREES/LOGISTIC REGRESSION MODEL.

MODEL V: SINGLE DECISION TREE

```
Train Data
Accuracy  :      0.5732583355923014

TPR Train :      0.7146932952924394
TNR Train :      0.4184528034066714

FPR Train :      0.5815471965933287
FNR Train :      0.28530670470756064

<AxesSubplot:>
```


```
Test Data
Accuracy  :      0.5443209541881269

TPR Test :       0.692019282271023
TNR Test :       0.3929747530186608

FPR Test :       0.6070252469813392
FNR Test :       0.30798071772897695

<AxesSubplot:>
```

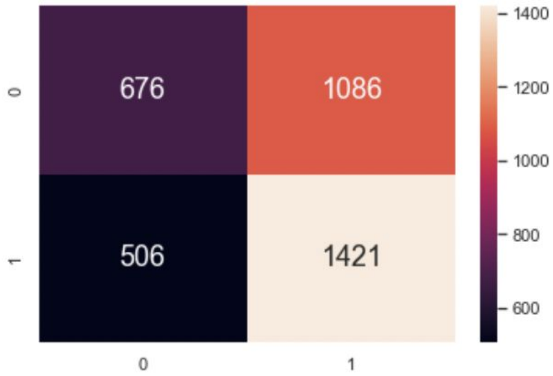



```
Accuracy of Model 5 with CV:      0.5493367323503744
Standard Deviation of scores:     0.013589544836058991
```

MODEL VI: RANDOM FOREST (UNOPTIMISED)

```
Train Data
Accuracy  :     0.5904716725399838

TPR Train :     0.7388743455497382
TNR Train :     0.43114109050028104

FPR Train :     0.568858909499719
FNR Train :     0.26112565445026176

<AxesSubplot:>
```


```
Test Data
Accuracy  :     0.5684467335321225

TPR Test :      0.7374156720290607
TNR Test :      0.3836549375709421

FPR Test :      0.6163450624290578
FNR Test :      0.2625843279709393

<AxesSubplot:>
```




```
Accuracy of Model 6 with CV:     0.5574019567314317
Standard Deviation of scores:    0.013306273161267541
```

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

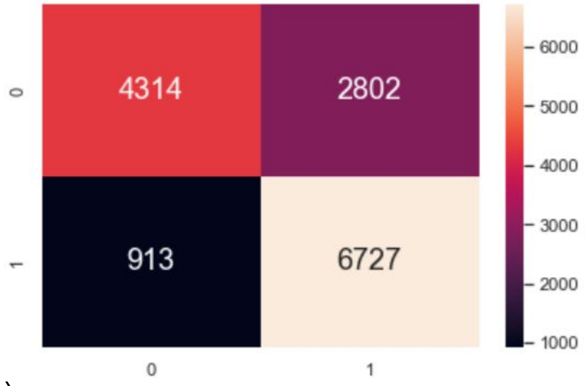
ANALYSIS

- RANDOM FOREST USES MULTIPLE DECISION TREES THAT CHOOSES FEATURES RANDOMLY.
- IT DOES NOT RELY ON THE FEATURE IMPORTANCE EXHIBITED BY A SINGLE DECISION TREE.
- RANDOM FOREST CAN GENERALIZE OVER THE DATA IN A BETTER WAY.
- CLASSIFICATION ACCURACY, TPR AND TNR SLIGHTLY IMPROVED BY USING RANDOM FOREST

MODEL VII: RANDOM FOREST (OPTIMISED HYPERPARAMETERS: MAX_DEPTH = 10, N_ESTIMATORS =900)

```
Train Data
Accuracy  :      0.7482380048793711

TPR Train :      0.8804973821989529
TNR Train :      0.606239460370995

FPR Train :      0.39376053962900504
FNR Train :      0.11950261780104712
```

<AxesSubplot:>




```
Test Data
Accuracy  :      0.578476551911087

TPR Test :       0.7275557861961598
TNR Test :       0.41543700340522133

FPR Test :       0.5845629965947786
FNR Test :       0.27244421380384015
```

<AxesSubplot:>


```
Accuracy of Model 7 with CV:      0.5652630563593771
Standard Deviation of scores:     0.014073432075857716
```

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

05 CONCLUSION

ANALYSIS

- THE MODEL'S ACCURACY ON THE TRAIN AND TEST SET HAS IMPROVED BASED ON THE OPTIMISED VALUES OF N_ESTIMATORS =900 AND MAX_DEPTH=10

- HOWEVER, THE ACCURACY OF THE TRAIN SET IS MUCH HIGHER THAN THAT OF THE TEST SET, WHICH PROVES THAT THE MODEL FAILS TO GENERALIZE TO THE TEST DATA.

- THIS PHENOMENON IS REFERRED TO AS 'OVERFITTING'.

MODEL VIII: REGULARISED RANDOM FOREST (WITH OPTIMISED HYPERPARAMETERS + PRE-PRUNING)

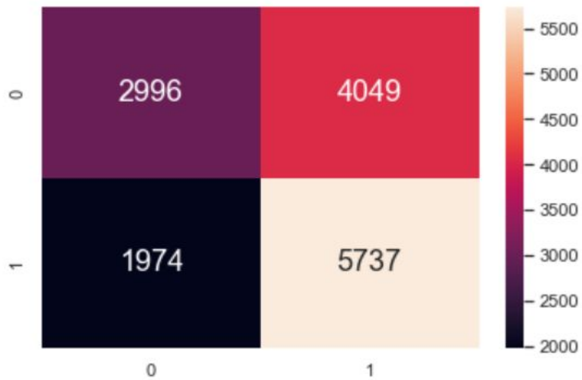
- REGULARISATION TECHNIQUE: PRE-PRUNING
    - INVOLVES TUNING THE HYPERPARAMETERS OF THE RANDOM FOREST MODEL.
        - HYPERPARAMETERS: MIN_SAMPLES_LEAF, MIN_SAMPLES_SPLIT
        - THIS AIMS TO CONTROL THE COMPLEXITY OF THE MODEL, AND REDUCE OVERFITTING

- OPTIMISED PARAMETERS: MIN_SAMPLES_LEAF = 0.01, MIN_SAMPLES_SPLIT = 0.01

MODEL VIII: RANDOM FOREST (WITH OPTIMISED HYPERPARAMETERS + PRE-PRUNING)

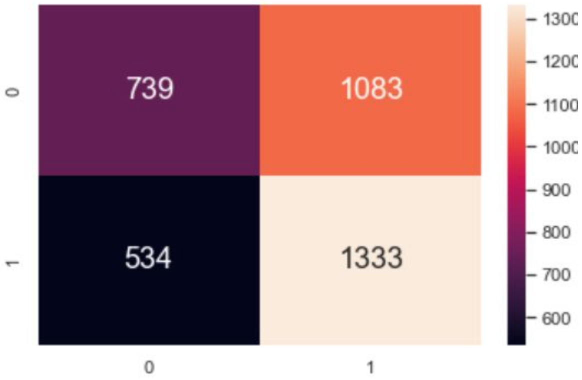
```
Train Data
Accuracy  :        0.5918270534020059

TPR Train :        0.7440020749578524
TNR Train :        0.4252661462029808

FPR Train :        0.5747338537970191
FNR Train :        0.2559979250421476
```

<AxesSubplot:>




```
Test Data
Accuracy  :        0.5616698292220114

TPR Test :         0.7139796464916979
TNR Test :         0.40559824368825464

FPR Test :         0.5944017563117453
FNR Test :         0.28602035350830207
```

<AxesSubplot:>




```
Accuracy of Model 8 with CV:      0.559028983510174
Standard Deviation of scores:     0.015942276121789407
```

01 PROBLEM DEFINITION

02 DATA PREPARATION & CLEANING

03 EDA & VISUALISATION

04 MACHINE LEARNING

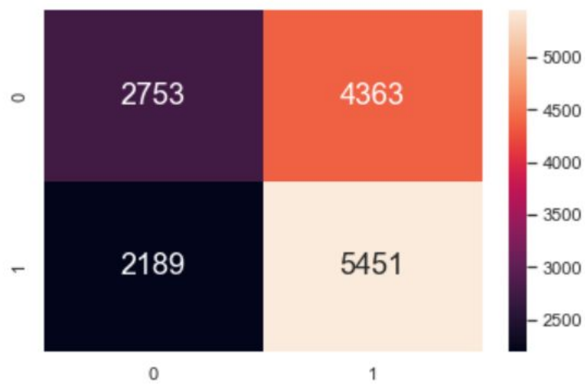
05 CONCLUSION

ANALYSIS

- WITH PRE-PRUNING, CLASSIFICATION ACCURACY FOR TEST SET REMAINED RELATIVELY THE SAME, WHILE THAT FOR TRAIN SET DECREASED GREATLY.
  - OVERFITTING IS REDUCED, WHILE MAINTAINING CLASSIFICATION ACCURACY
- INSIGNIFICANT IMPROVEMENT IN TEST CLASSIFICATION ACCURACY MIGHT BE DUE TO THE UNRELATED NATURE OF OUR DATASET (BETWEEN POPULARITY AND AUDIO FEATURES)
- REGULARISATION ATTEMPT UNSUCCESSFUL

MODEL IX: LOGISTIC REGRESSION CLASSIFICATION

```
Train Data
Accuracy  :      0.5559772296015181

TPR Train :      0.7134816753926702
TNR Train :      0.3868746486790315

FPR Train :      0.6131253513209668
FNR Train :      0.2865183246073298

<AxesSubplot:>
```




```
Test Data
Accuracy  :      0.5670913526701002

TPR Test :       0.7368967306694344
TNR Test :       0.3813847900113507

FPR Test :       0.6186152099886493
FNR Test :       0.26310326933056566

<AxesSubplot:>
```


```
Accuracy of Model 9 with CV:     0.553878508107115
Standard Deviation of scores:    0.014416327331505661
```

ANALYSIS

- LOGISTIC REGRESSION IS A STATISTICAL ANALYSIS METHOD TO PREDICT A BINARY OUTCOME, SUCH AS YES OR NO, BASED ON PRIOR OBSERVATIONS OF A DATA SET
- LOGISTIC REGRESSION APPLIES REGULARISATION BY DEFAULT
- LOGISTIC REGRESSION MODEL PERFORMED SIMILAR TO PRE-PRUNED MODEL, WITH NO 'OVERFITTING'.

CROSS VALIDATION SCORES (CLASSIFICATION MODEL)

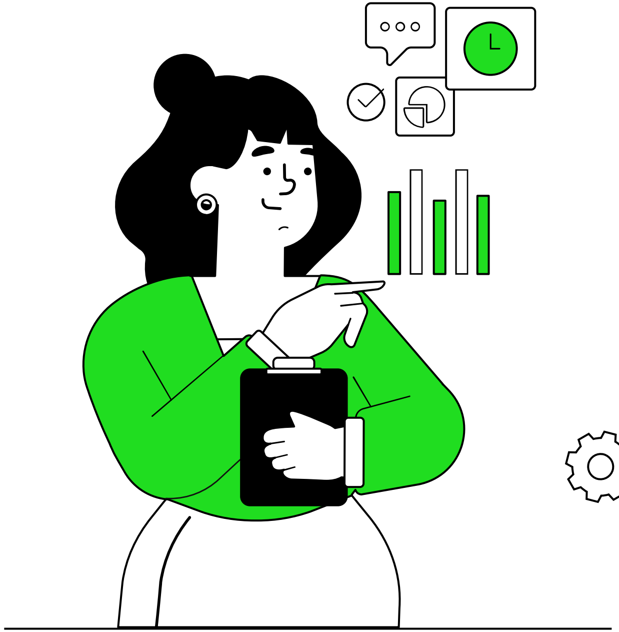
	MEAN (CLASSIFICATION ACCURACY)	STANDARD DEVIATION
MODEL V	0.5493	0.0136
MODEL VI	0.5574	0.0133
MODEL VII	0.5653	0.0141
MODEL VIII	0.5590	0.0159
MODEL IX	0.5539	0.0144

RESULTS ANALYSIS (CLASSIFICATION MODEL)

	CLASSIFICATION ACCURACY	
	TRAIN	TEST
MODEL V	0.5733	0.5443
MODEL VI	0.5905	0.5684
MODEL VII	0.7482	0.5785
MODEL VIII	0.5918	0.5617
MODEL IX	0.5560	0.5671

05

THE CONCLUSION



CONCLUSION

- OUT OF ALL CLASSIFICATION MODELS, LOGISTIC REGRESSION MODEL PERFORMS THE BEST.
    - THIS CAN BE OBSERVED BY REDUCTION IN 'OVERFITTING' AS WELL AS THE IMPROVEMENT IN ACCURACY FROM TRAIN SET TO TEST SET
- ADDITIONALLY, ALTHOUGH PRE-PRUNING WAS DONE TO REGULARISE THE DATA, IT WAS NOT A GOOD ATTEMPT BECAUSE THE ACCURACY ON THE TEST SET DID NOT IMPROVE

OUR THOUGHTS

NO PERFECT FORMULA

- VARIOUS GENRES EXISTS FOR A REASON
- POPULAR SONGS CHANGE OVERTIME ACCORDING TO DEMOGRAPHIC AND TASTE

EXTERNAL FACTORS

- MARKETING
- FUNDING
- EXPLOIT OF SPOTIFY POPULARITY ALGORITHM

RECOMMENDATIONS

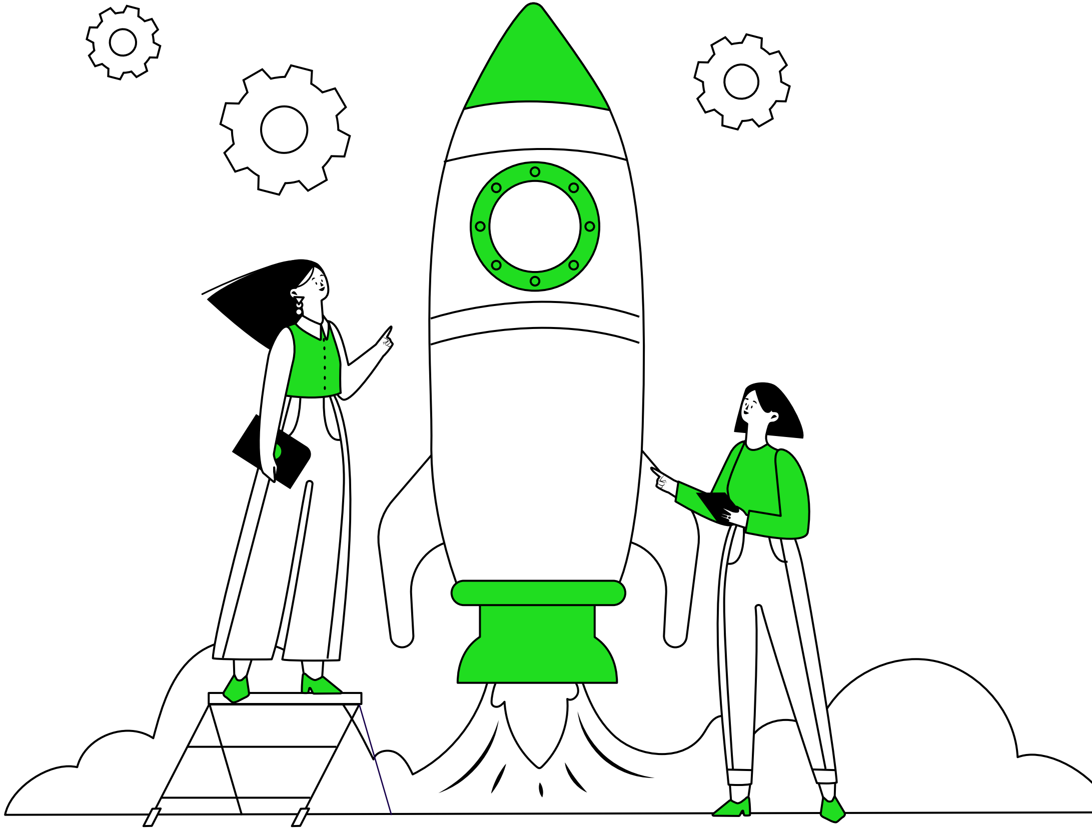
NARROW OUR SCOPE

- EXPLORE THE RELATIONSHIP IN THE SAME GENRE, RATHER THAN ACROSS MULTIPLE GENRES

DIFFERENT MACHINE LEARNING MODEL

- INCORPORATE DEEP LEARNING NEURAL NETWORKS TO BETTER STUDY THE RELATIONSHIP BETWEEN POPULARITY AND AUDIO VARIABLES

THANKS!



REFERENCES

LEWINSON, E. (2021, AUGUST 26). *OUTLIER DETECTION WITH ISOLATION FOREST*. MEDIUM. RETRIEVED APRIL 1, 2022, FROM HTTPS://TOWARDSDATASCIENCE.COM/OUTLIER-DETECTION-WITH-ISOLATION-FOREST-3D190448D45E

SWAMINATHAN, S. (2019, JANUARY 18). LOGISTIC REGRESSION - DETAILED OVERVIEW. MEDIUM. RETRIEVED APRIL 1, 2022, FROM HTTPS://TOWARDSDATASCIENCE.COM/LOGISTIC-REGRESSION-DETAILED-OVERVIEW-46C4DA4303BC

SPOTIFY API - HTTPS://DEVELOPER.SPOTIFY.COM/DOCUMENTATION/WEB-API/