01

Spotify

Genre Predictor

John Dilligard
David Obembe
Will Pumphrey
Ryan Thomas
Brian Yu

02

Data Processing

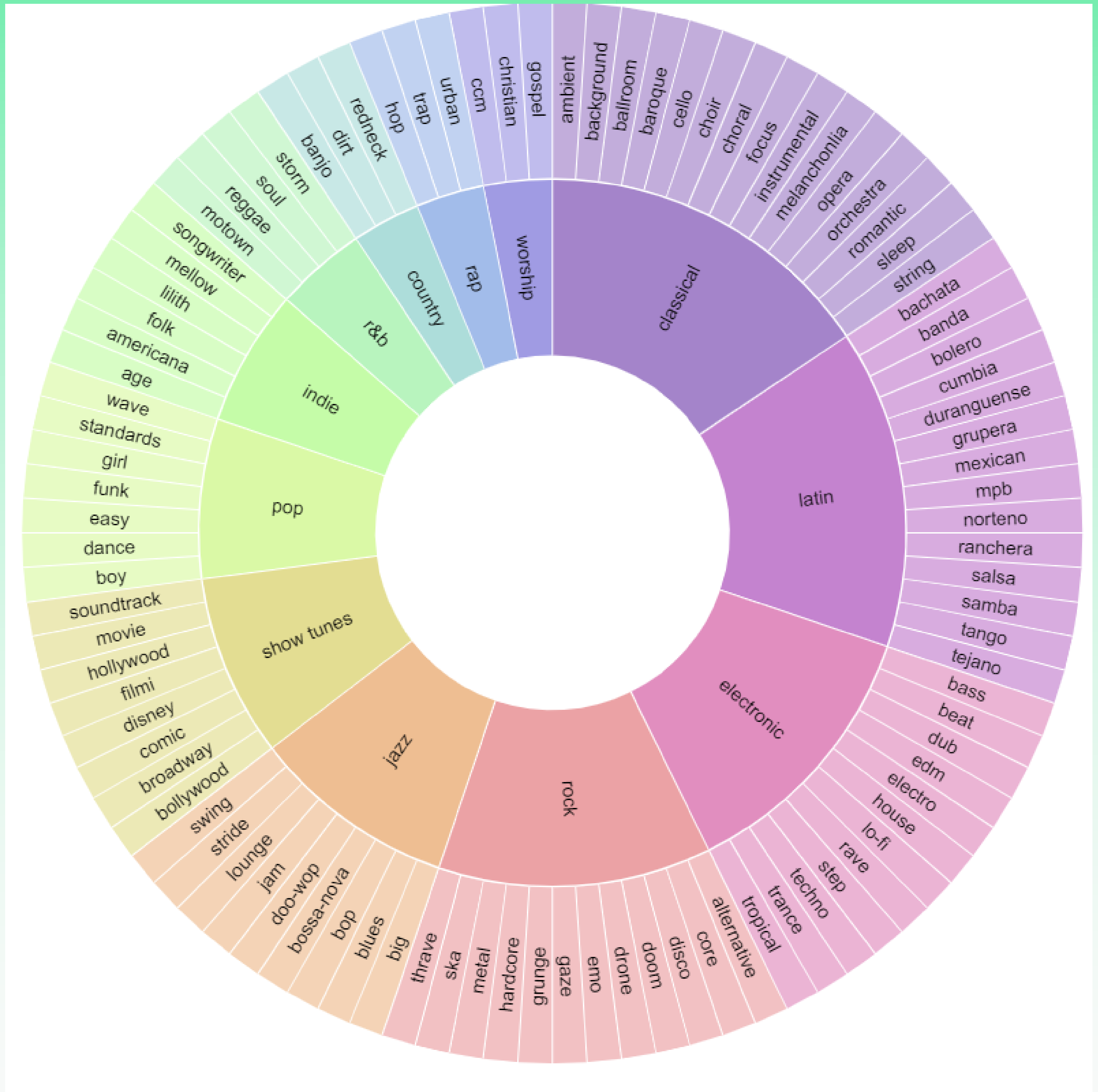
Data Gathering

- Used a Kaggle dataset that compiled song features from Spotify API.
- data contained information on 160,000 tracks.
- Tracks from 1921 through 2020.

Data Cleaning

- The cleaning of the data was primarily focused on narrowing down the genres/sub-genres.
- Ensuring that each song/artist was mapped to one genre
- Dropping missing rows of data
- Encoding Data(one-hot encoding)

03



classical
ambient
background
ballroom
baroque
cello
choir
choral
focus
instrumental
melancholia
opera
orchestra
romantic
sleep
string
latin
bachata
banda
bolero
cumbia
duranguense
grupera
mexican
mpb
norteno
ranchera
salsa
samba
tango
tejano
electronic
bass
beat
dub
edm
electro
house
lo-fi
rave
step
techno
trance
tropical
rock
alternative
core
disco
doom
drone
emo
gaze
grunge
hardcore
metal
ska
thrave
jazz
big
blues
bop
bossa-nova
doo-wop
jam
lounge
stride
swing
show tunes
bollywood
broadway
comic
disney
filmi
hollywood
movie
soundtrack
pop
boy
dance
easy
funk
girl
standards
wave
age
indie
americana
folk
lilith
mellow
songwriter
r&b
motown
reggae
soul
storm
country
banjo
dirt
redneck
rap
hop
trap
urban
worship
ccm
christian
gospel

Exploratory Data Analysis

With Python & Tableau

04

Tools

- Matplotlib
- Seaborn
- Tableau
- Pandas

Correlations

- A heatmap was used to visualise and explore the correlation between Spotify song features.
- Dark red Spots: (-1) correlation.
- white spots:(+1) correlations.

Genre Distribution

- most represented Genres: rock & Pop.
- Least represented Genres: Worship.

Heat Map

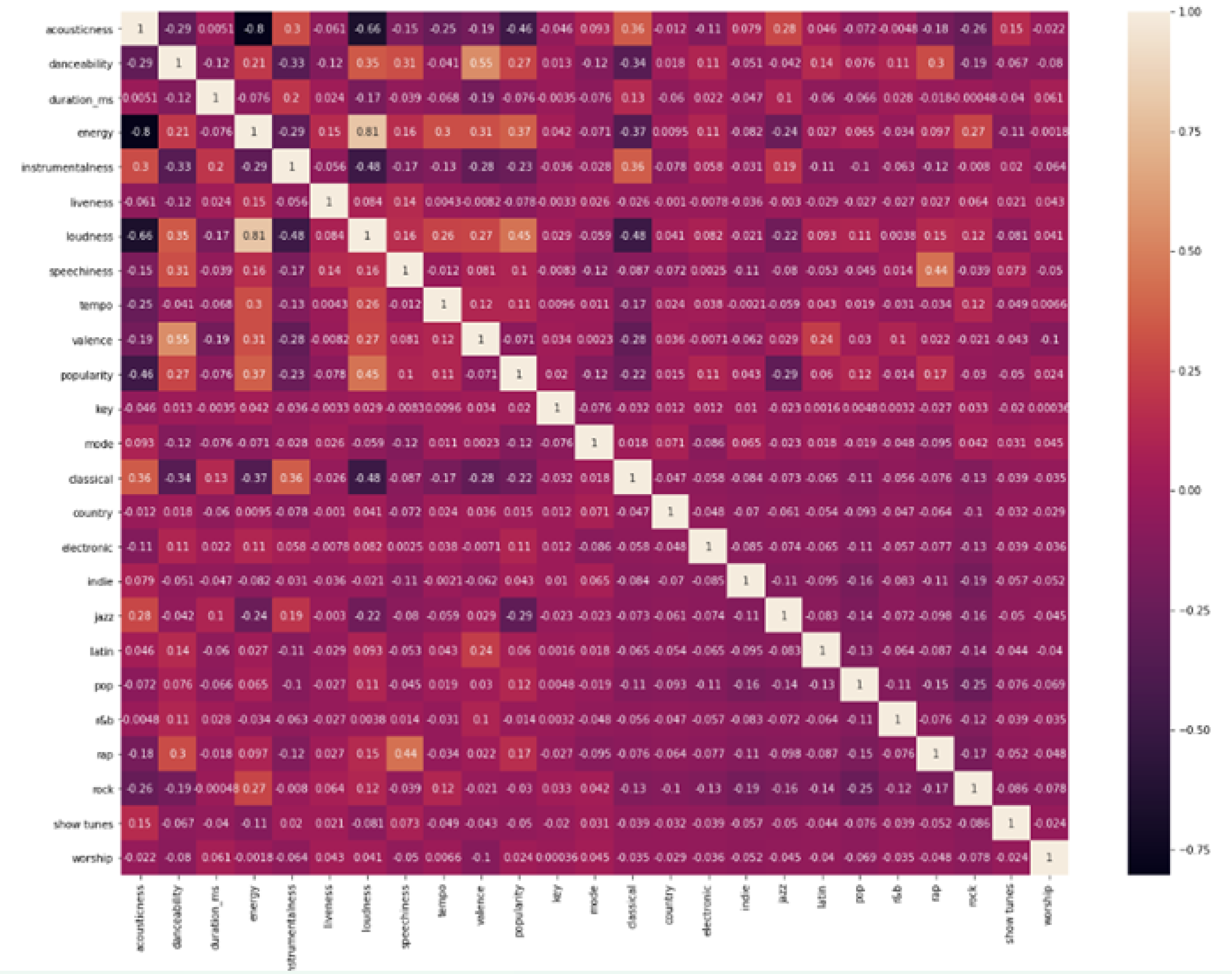
05

(+) correlations

- valence & danceability(0.55).
- popularity & loudness(0.40).
- Loudness & energy(0.81).
- Speechiness & rap(0.44).

(-) correlations

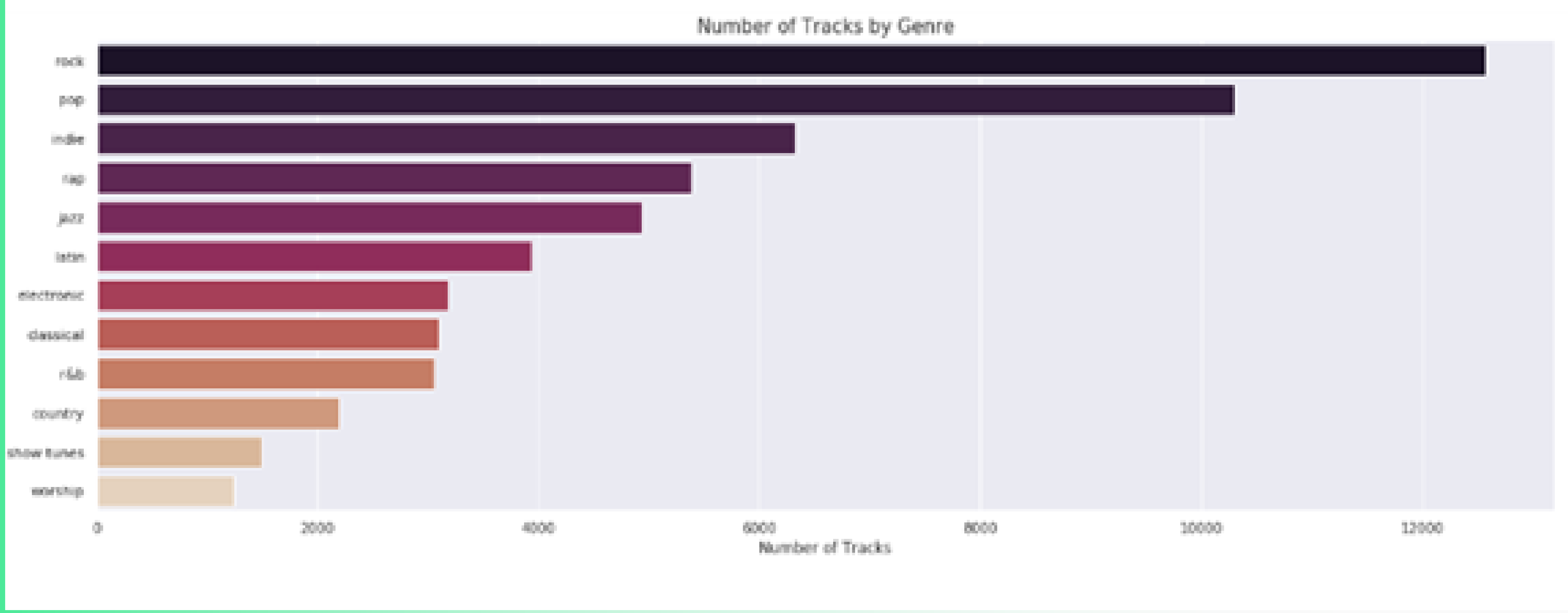
- acoustic & energy(-0.8).
- acoustic & loudness(-0.66).
- acoustic & popularity(-0.46)
- instrumental and loudness(-0.48).

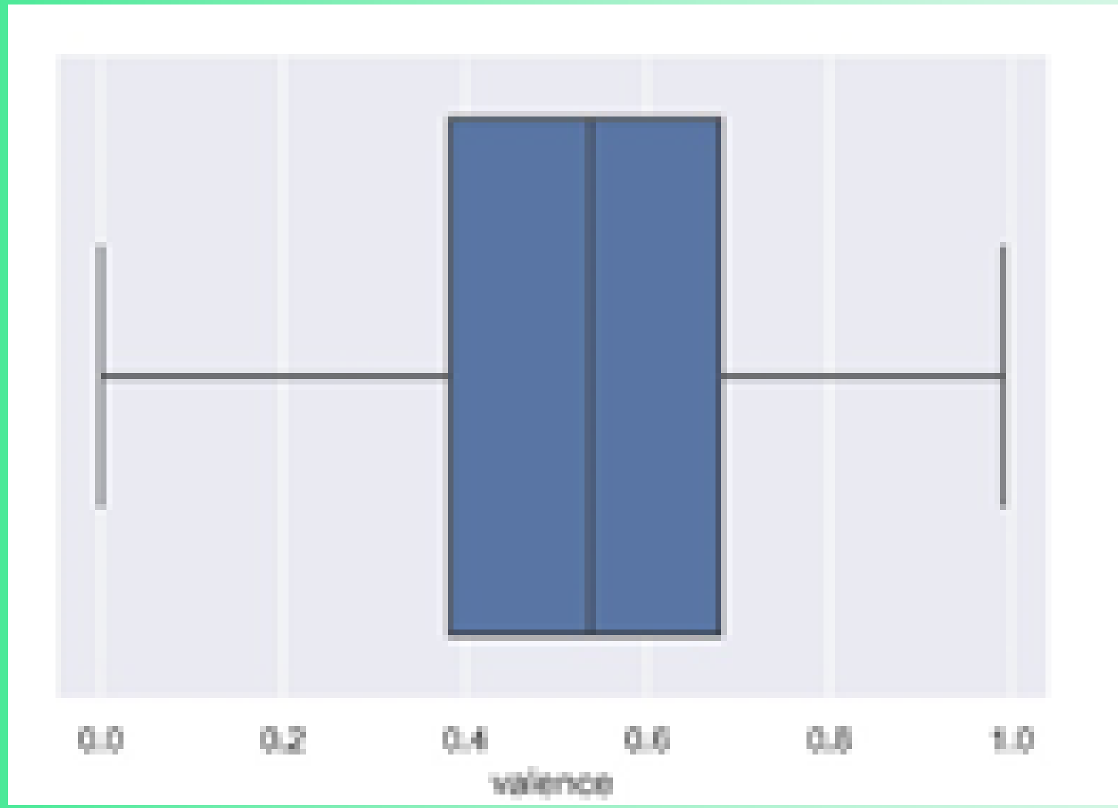


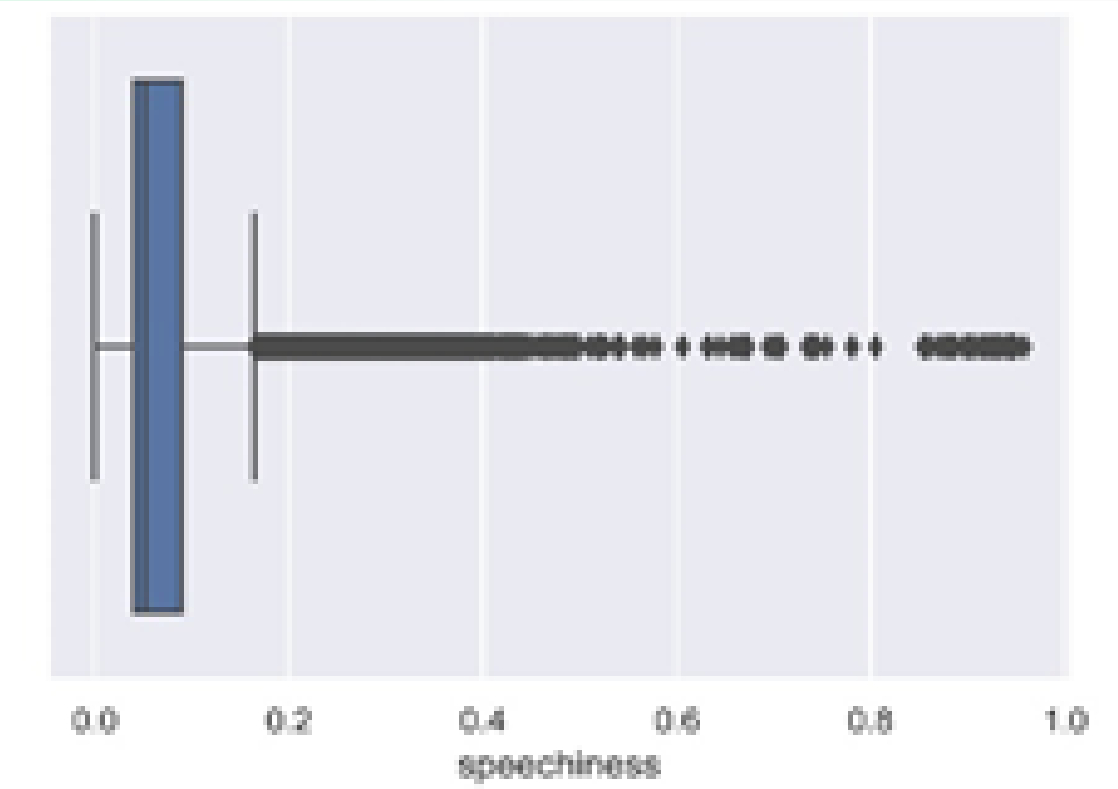
06

Distribution

- **speechiness** represents the present of spoken word and rap
- **valence** is the level of happiness of a song







07

Predictive models

Logistic regression

- This was our first choice due to the dependent variable(target) in our data being categorical.
- The Metric scores proved that the logistic model was relatively accurate for classical music, rap, jazz and worship.
- Indie music, pop and show tunes were poorly predicted.

Random Forest

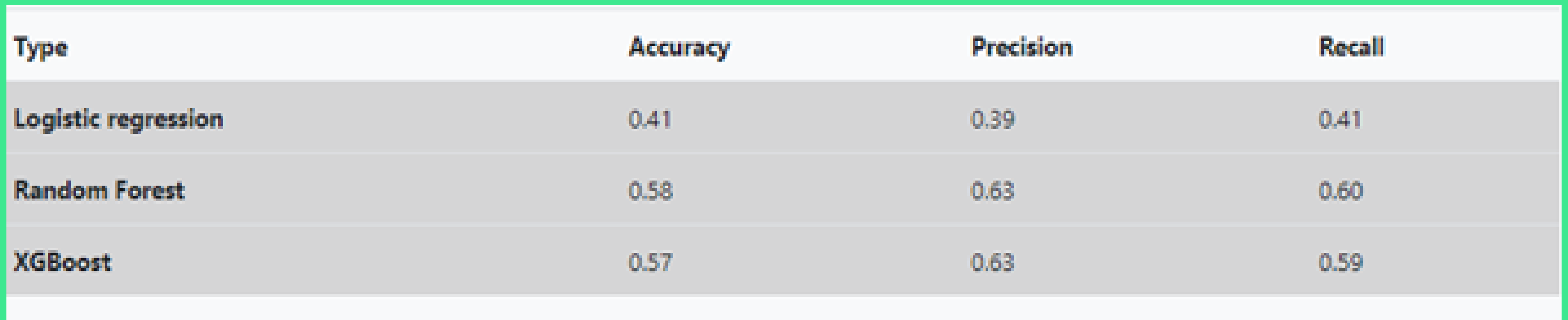
- We used this model because of its ability to **run efficiently on large databases** and its ability to produce a **highly accurate classifier.**
- The Random Forest had better metric scores than Logistic regression on all genres.
- accuracy score was **0.58.**

XGBoosting Classifier

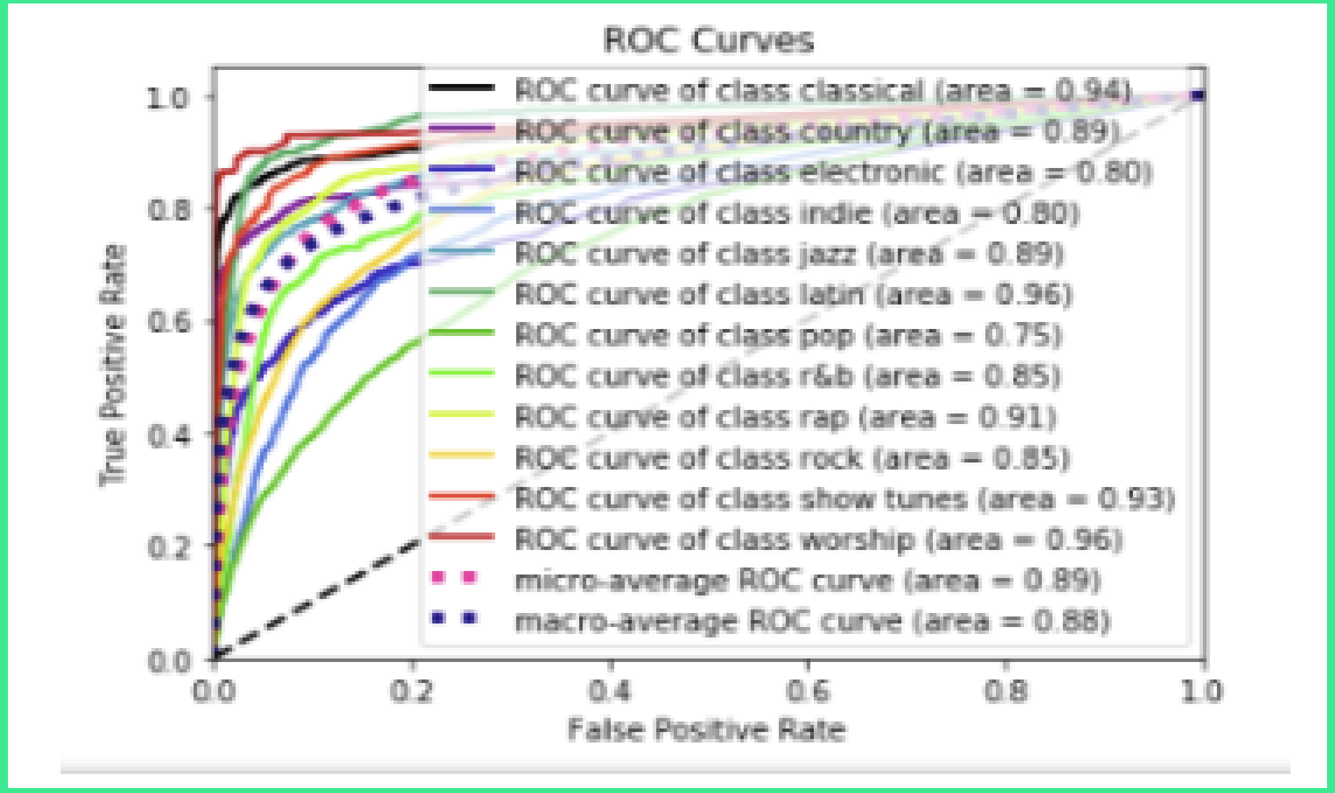
- **gradient boosted decision trees** designed for speed and performance.
- The XGBoost model was better than the logistic model but failed to improve on random forest classifier.
- accuracy score was **0.57.**

08

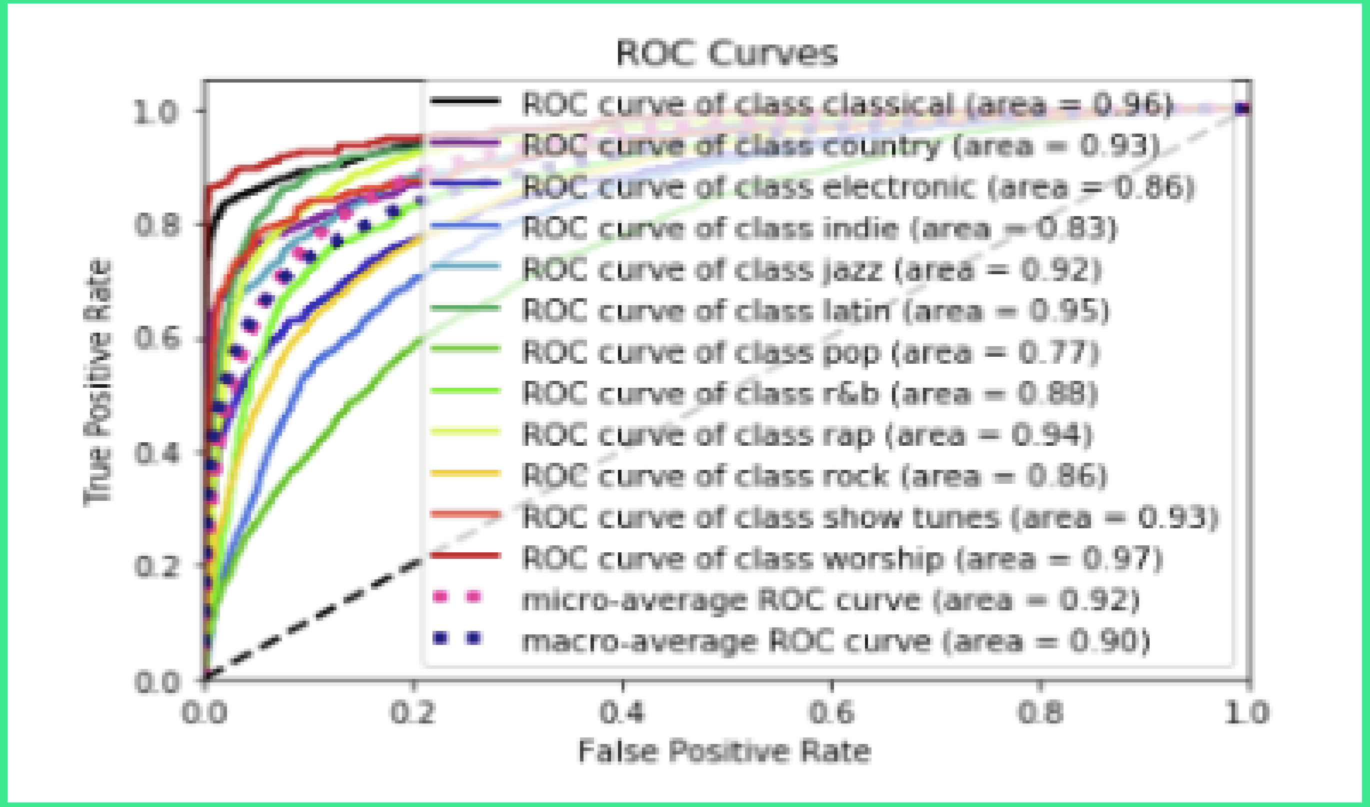
Metric Tests



Type	Accuracy	Precision	Recall
Logistic regression	0.41	0.39	0.41
Random Forest	0.58	0.63	0.60
XGBoost	0.57	0.63	0.59



Random Forest



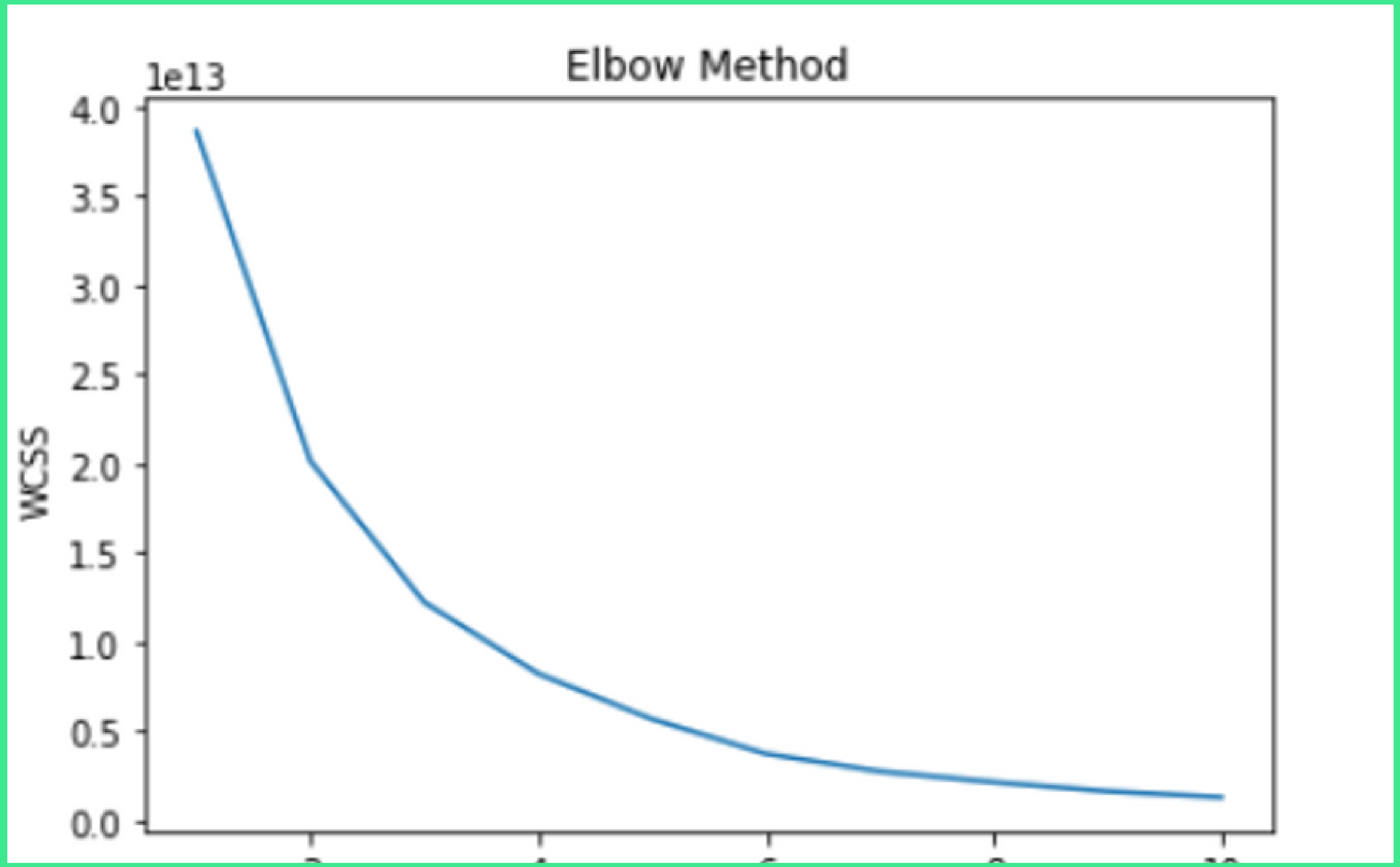
XGBoosting Classifier

BUSINESS REVIEW | 2020

09 Clustering

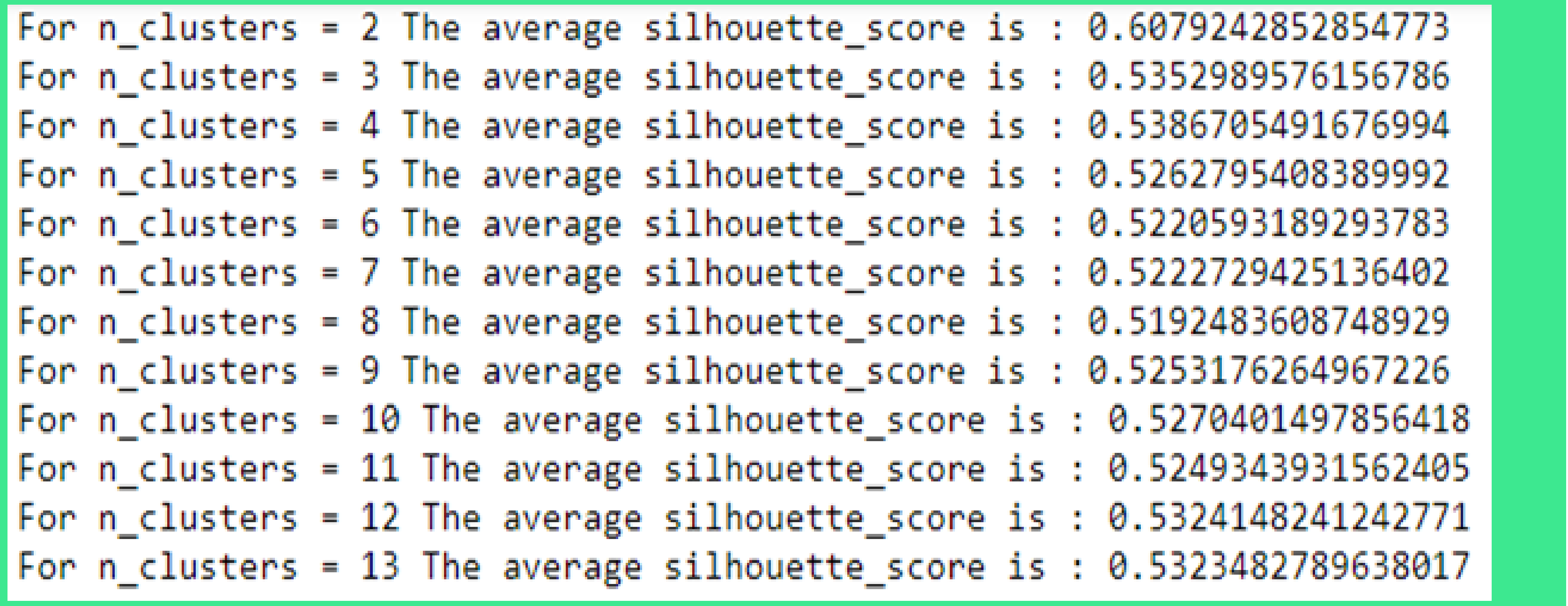
K-Means clustering analysis

K-Means Cluster analysis is a method that could have been used to narrow down the genres/sub-genres from over a hundred to a much more manageable number. An elbow chart was generated using the K-Means Clustering algorithm. K-Means clustering algorithm suggested that between 4-6 genres would be most optimal. The team felt that 4-6 genres was too small. Another analysis method was needed.



Silhouette Analysis

Silhouette analysis is another method that could be used for selecting the optimal number genres. The silhouette analysis suggested that 4 or 12 genres would be most optimal. Twelve genres was chosen.



```
For n_clusters = 2 The average silhouette_score is : 0.6079242852854773
For n_clusters = 3 The average silhouette_score is : 0.5352989576156786
For n_clusters = 4 The average silhouette_score is : 0.5386705491676994
For n_clusters = 5 The average silhouette_score is : 0.5262795408389992
For n_clusters = 6 The average silhouette_score is : 0.5220593189293783
For n_clusters = 7 The average silhouette_score is : 0.5222729425136402
For n_clusters = 8 The average silhouette_score is : 0.5192483608748929
For n_clusters = 9 The average silhouette_score is : 0.5253176264967226
For n_clusters = 10 The average silhouette_score is : 0.5270401497856418
For n_clusters = 11 The average silhouette_score is : 0.5249343931562405
For n_clusters = 12 The average silhouette_score is : 0.5324148241242771
For n_clusters = 13 The average silhouette_score is : 0.5323482789638017
```

BUSINESS REVIEW | 2020

Timeline

Q1

Project selection and Proposal

Q2

- Data exploration and model selection
- report started

Q3

- Tableau Viz
- Report writing

Q4

- Website building
- embedding Tableau

Q5

- Executive summary.
- Power point.

Presented by

Ryan

ML Cook

David

Documentation President

Brian

Tableau Legend

William

Visualisation Expert

John

Clustering Master

Thank you!

http://ml.rtaa.ninja/mlmodels