Spotify Tracks from 1922-2021

Corinne Smith

Spotify Dataset

Combined 2 DataFrames: tracks.csv and artists.csv

Categorical Variables:

- mode
- explicit

Variables to Manipulate:

- key (range of 0-11, turn into dummies)
- genre (associated with artist, use first genre)
- timesignature (turn numeric into dummies)
- release_date (just want year)

Numeric:

- acousticness
- danceability
- energy
- duration_ms
- instrumentalness
- valence
- track_popularity
- tempo
- liveness
- loudness
- speechiness
- artist_followers
- artist_popularity

Spotify Dataset

- Massive dataset: 432,000 rows!
- Limitations:
    - Genre for track is the first genre associated with the artist
        - Does not account for artists who create in different genres
    - Spotify is relatively new, so its user base is new
    - Popularity measures likely reflect ONLY demographics who use Spotify
    - Aggregated popularity loses subgroups of listeners

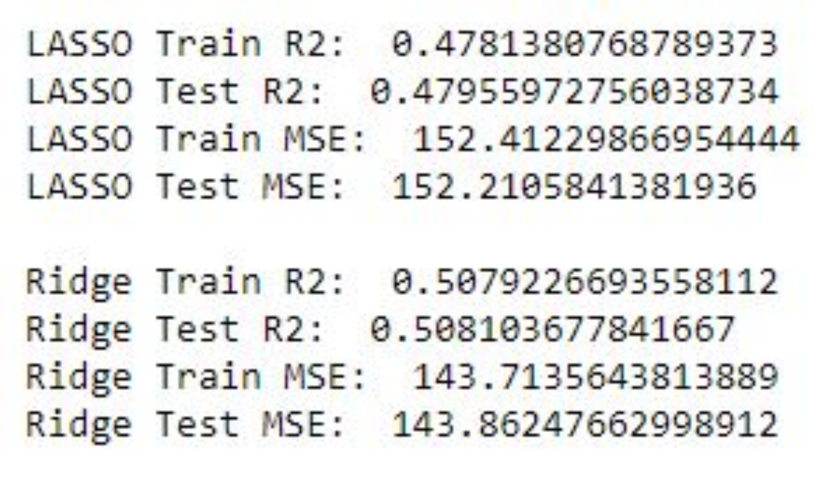
Three Main Questions

1. How can we predict the popularity of a track and what are the most important predictors?
2. By what characteristics can we group/cluster tracks? What subcategories of tracks exist?
3. Do different genres see varying levels of popularity, and does this change over time?

#1

How can we predict the popularity of a track and what are the most important predictors?

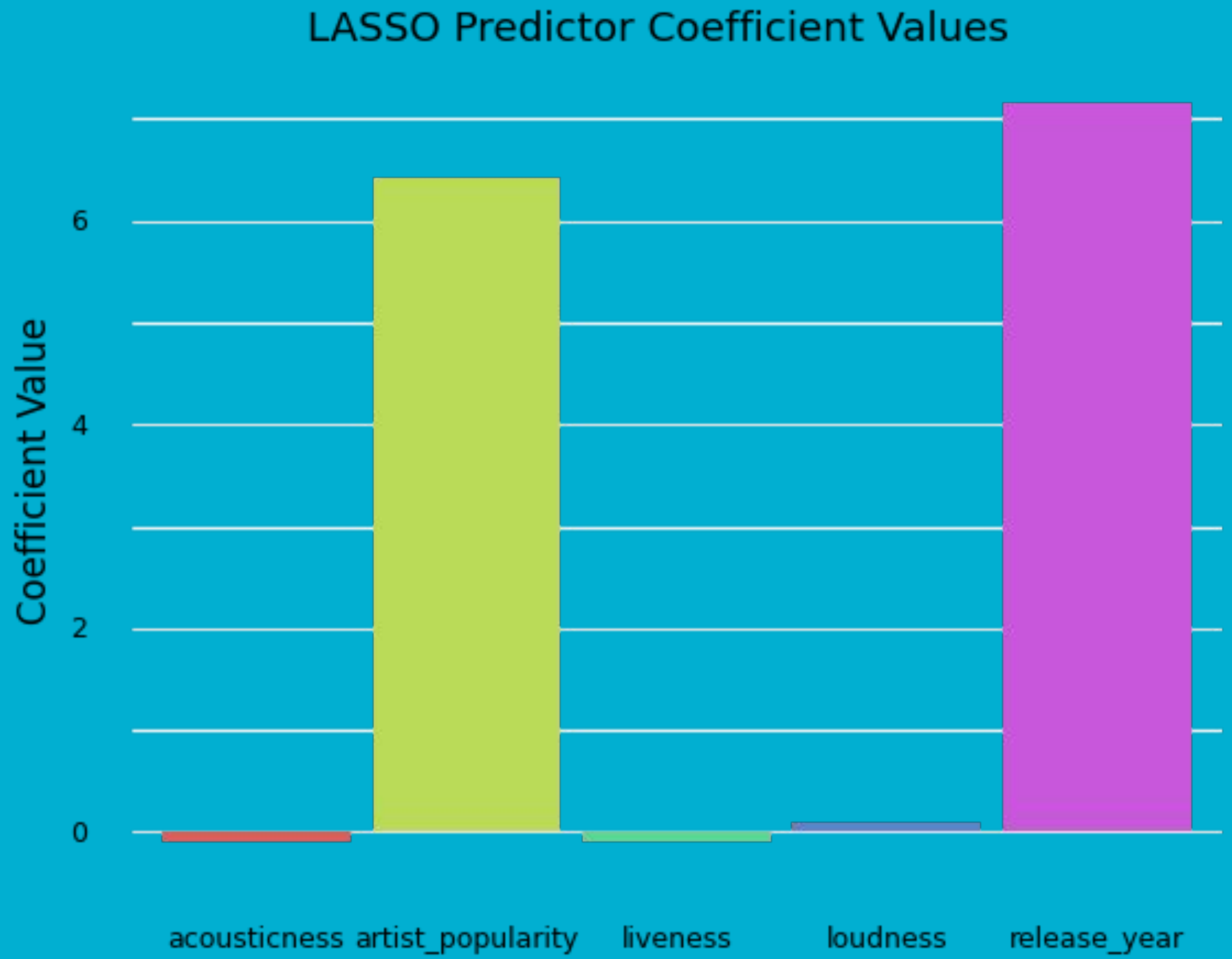
- Performed linear regression with LASSO and Ridge regularization
- R-squared values around 0.50, not ideal performance
- Odd behavior with Mean Squared Error
- Possible underfit/non-linear relationship



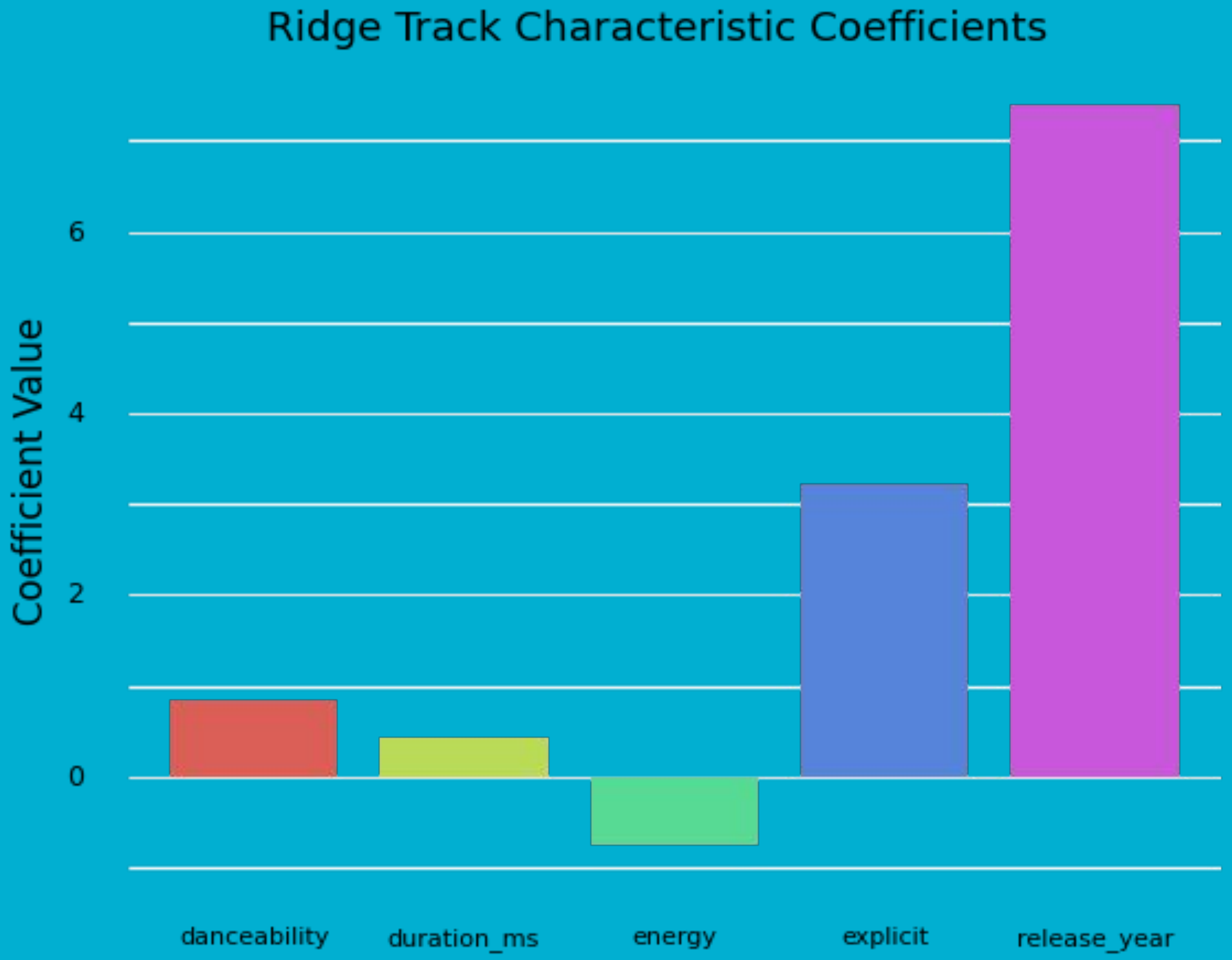
```
LASSO Train R2:  0.4781380768789373
LASSO Test R2:  0.4795597256038734
LASSO Train MSE:  152.41229866954444
LASSO Test MSE:  152.2105841381936

Ridge Train R2:  0.5079226693558112
Ridge Test R2:  0.508103677841667
Ridge Train MSE:  143.7135643813889
Ridge Test MSE:  143.86247662998912
```

#1

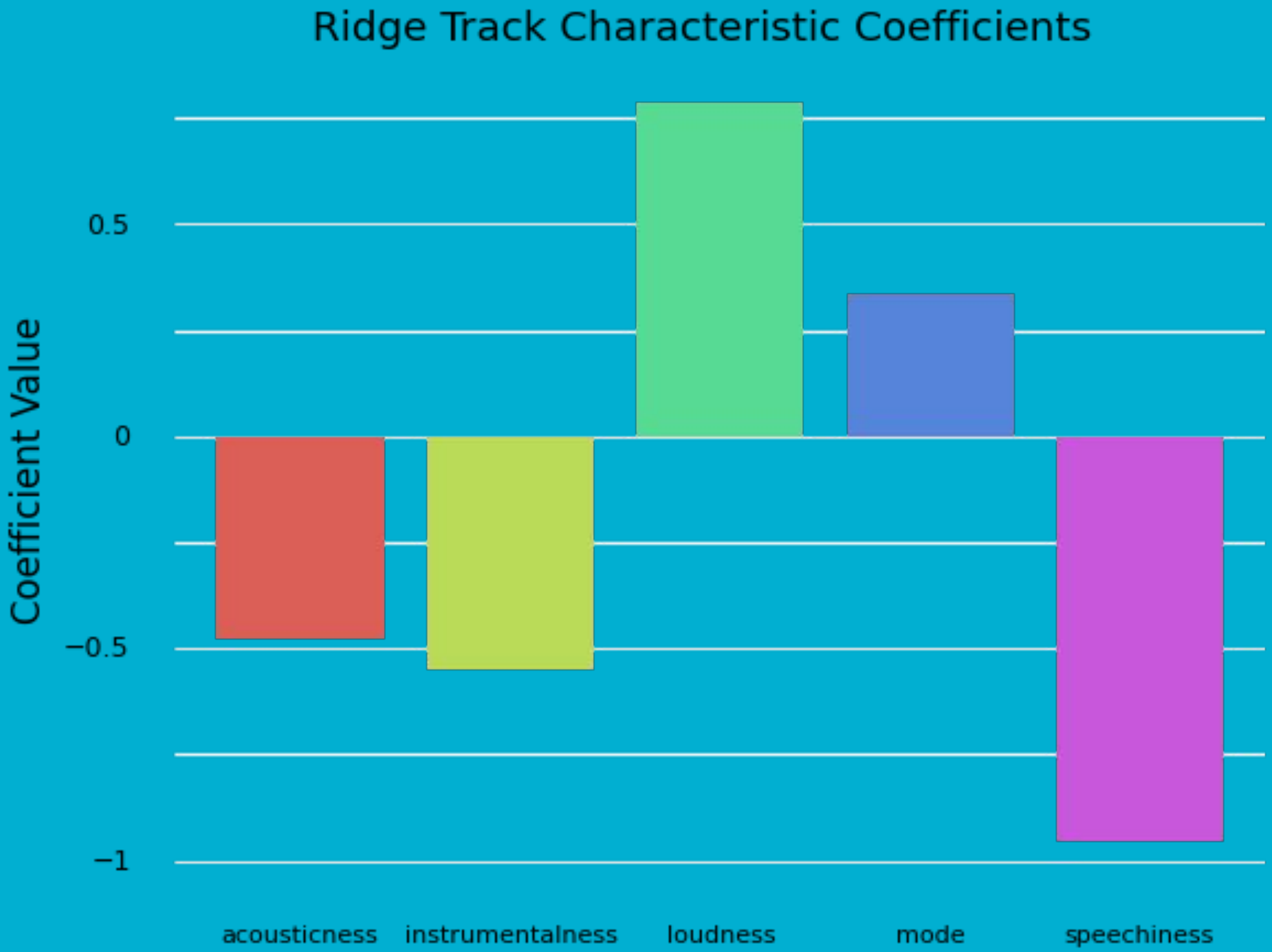

LASSO Predictor Coefficient Values
Coefficient Value
0
2
4
6
acousticness
artist_popularity
liveness
loudness
release_year

#1

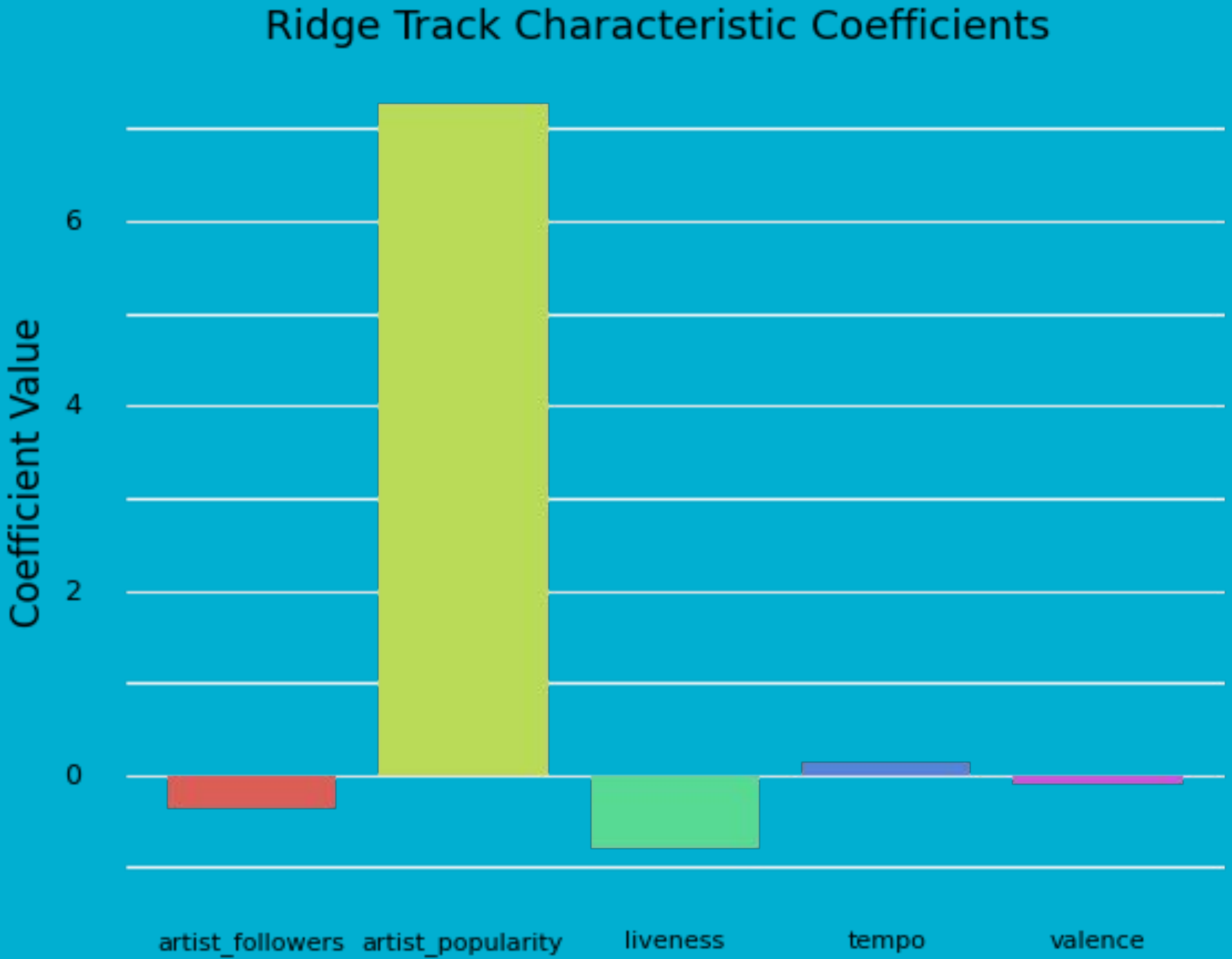


Ridge Track Characteristic Coefficients

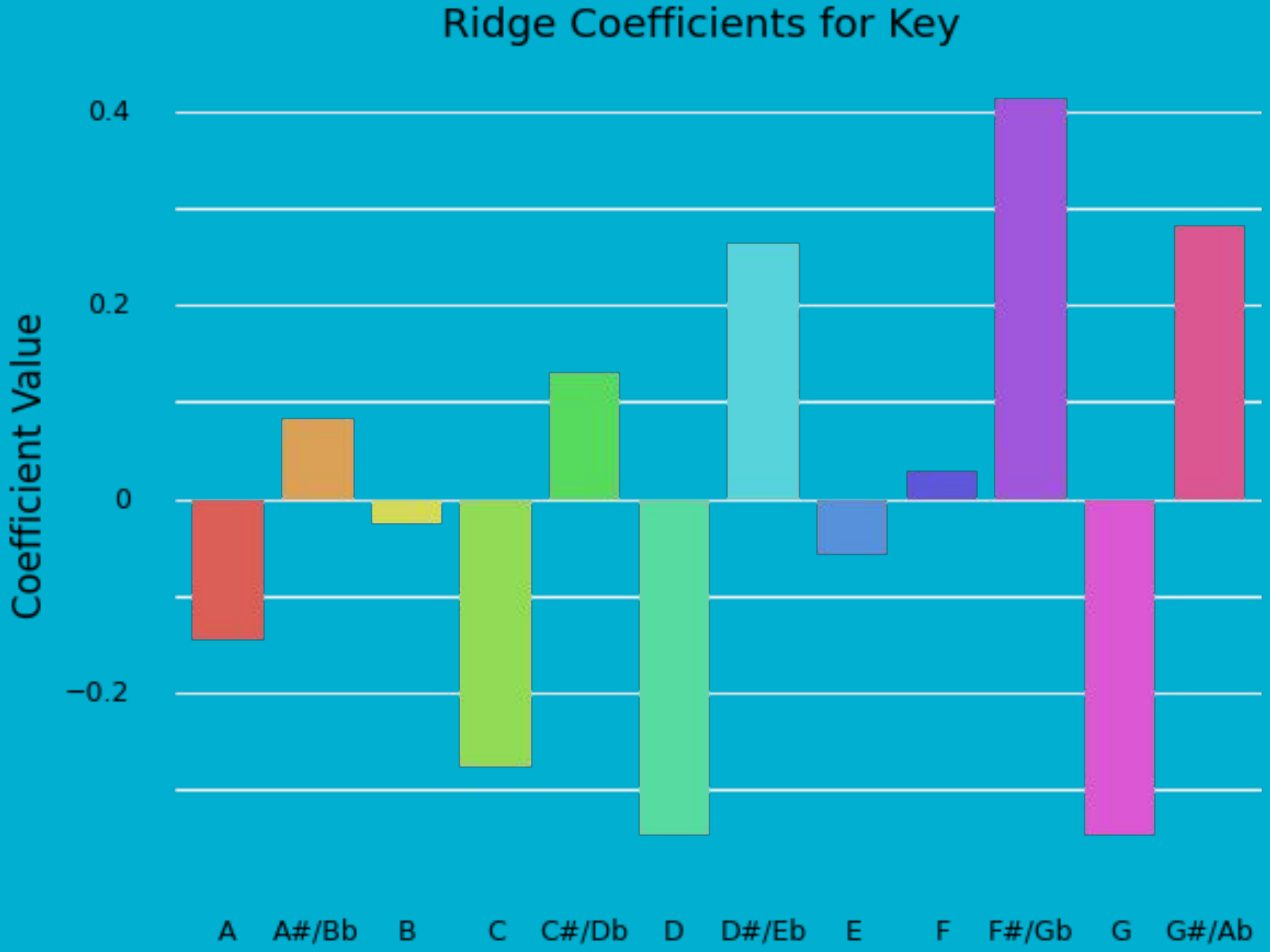
#1


Ridge Track Characteristic Coefficients
Coefficient Value
0.5
0
−0.5
−1
acousticness
instrumentalness
loudness
mode
speechiness

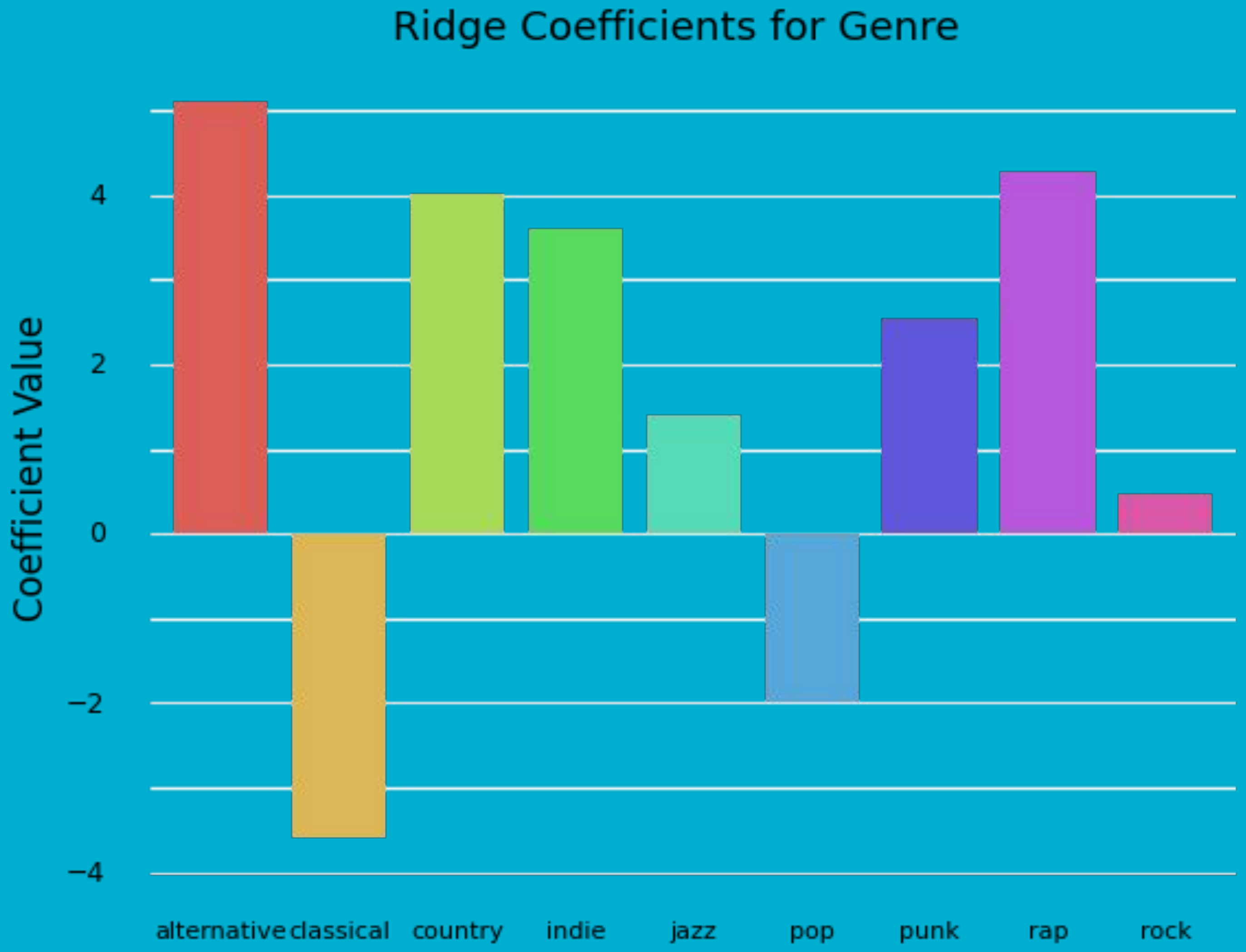
#1


Ridge Track Characteristic Coefficients
Coefficient Value
6
4
2
0
artist_followers
artist_popularity
liveness
tempo
valence

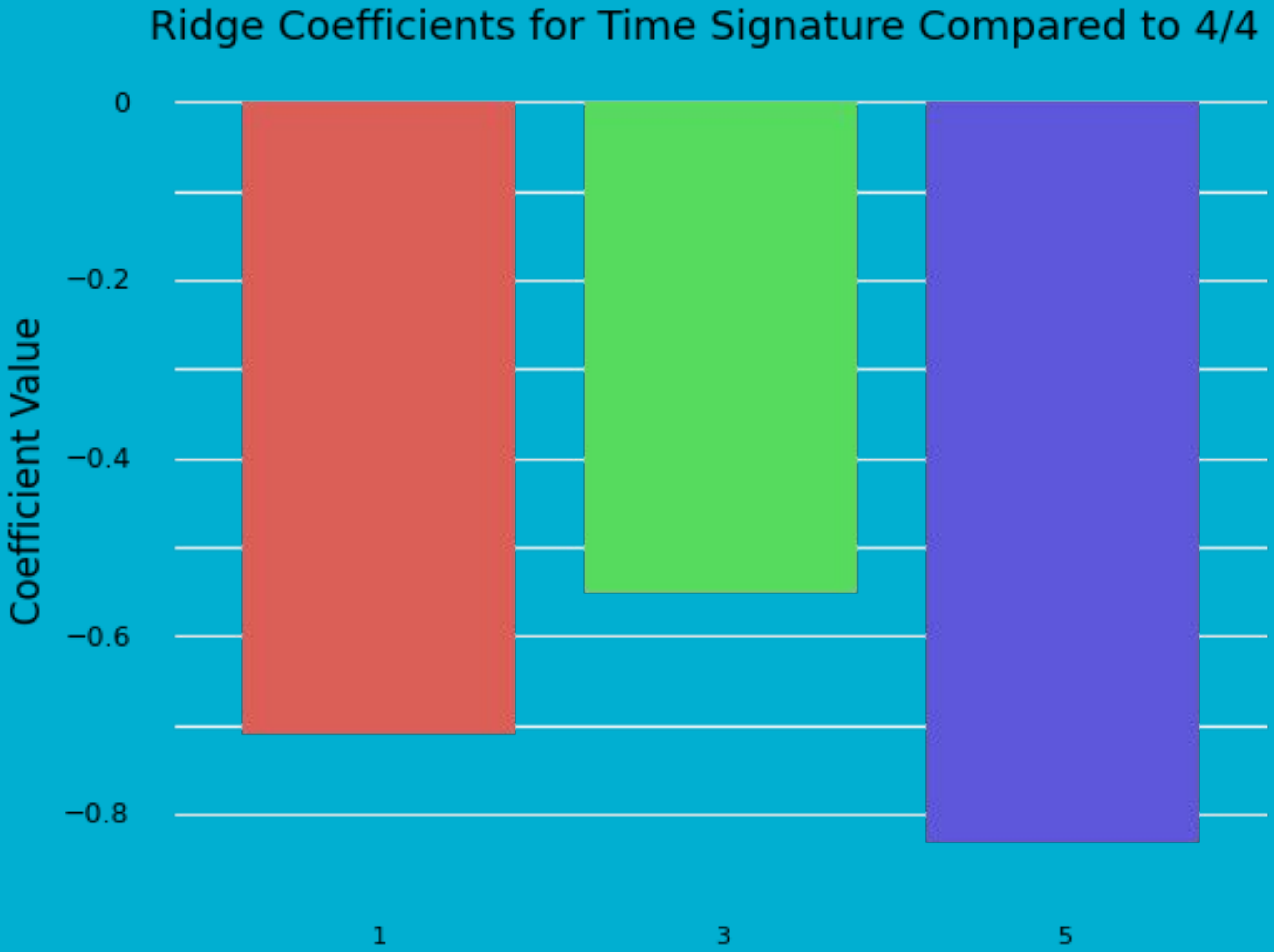
#1


Ridge Coefficients for Key
Coefficient Value
0.4
0.2
0
−0.2
A
A#/Bb
B
C
C#/Db
D
D#/Eb
E
F
F#/Gb
G
G#/Ab

#1


Ridge Coefficients for Genre
Coefficient Value
4
2
0
−2
−4
alternative
classical
country
indie
jazz
pop
punk
rap
rock

#1



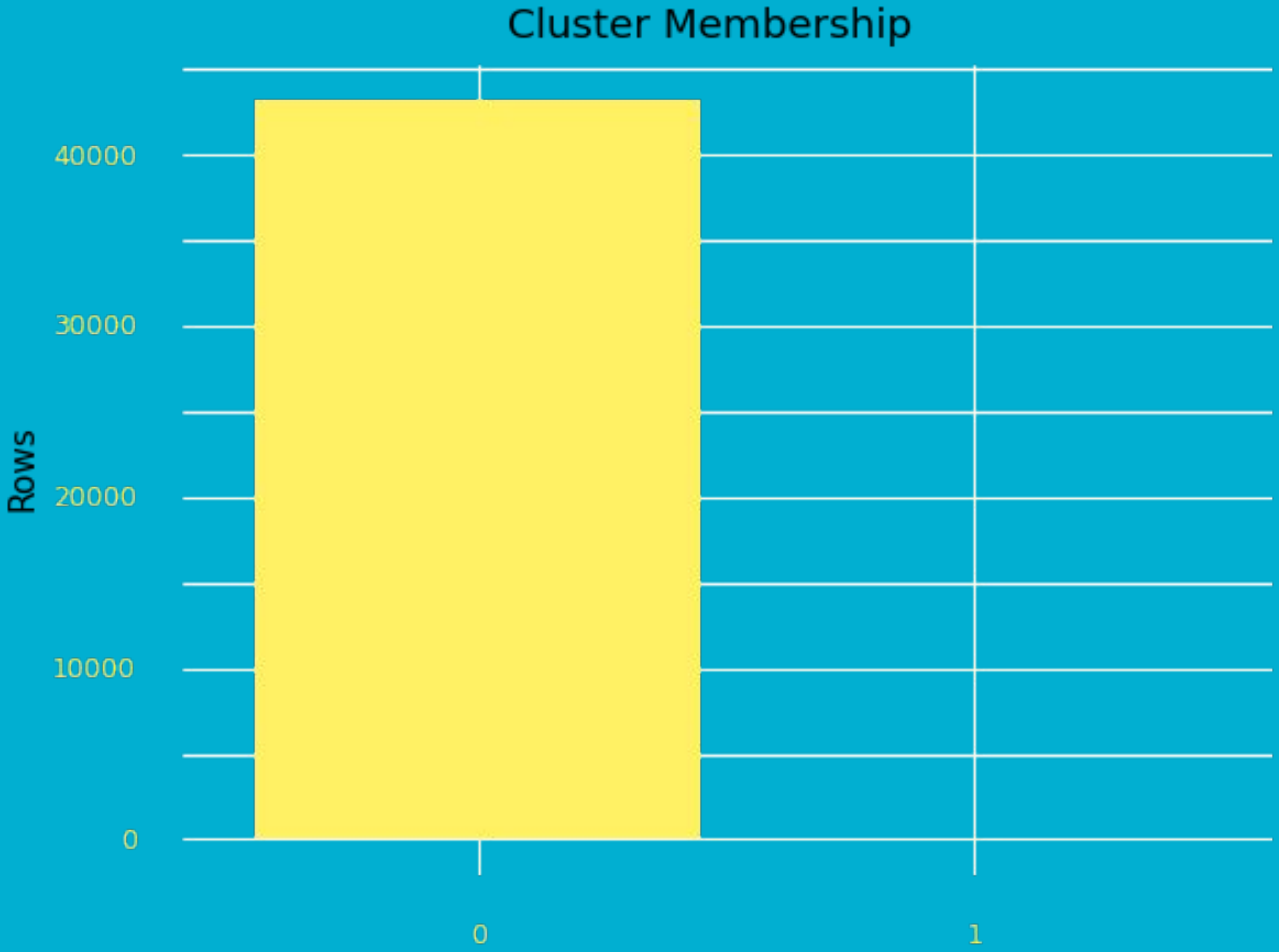
Ridge Coefficients for Time Signature Compared to 4/4

#2

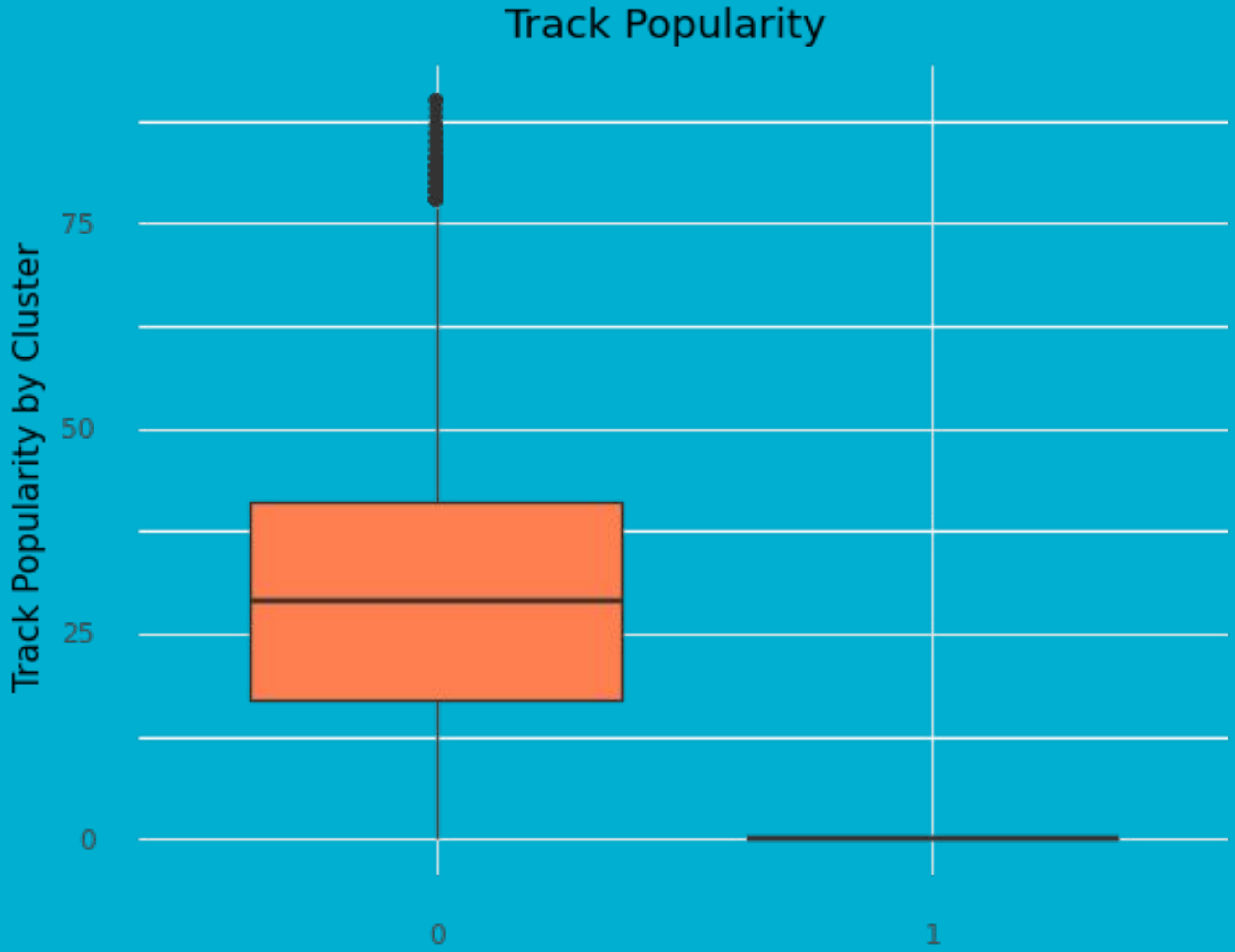
By what characteristics can we group/cluster tracks? What subcategories of tracks exist?

- Randomly sampled 10% of data
- Used Hierarchical Agglomerative Clustering
- Attempted to find genres/subgenres
- Silhouette Score: 0.8772
- 2 clusters
- 43,199 members for Cluster 0
- 1 member for Cluster 1

#2



#2

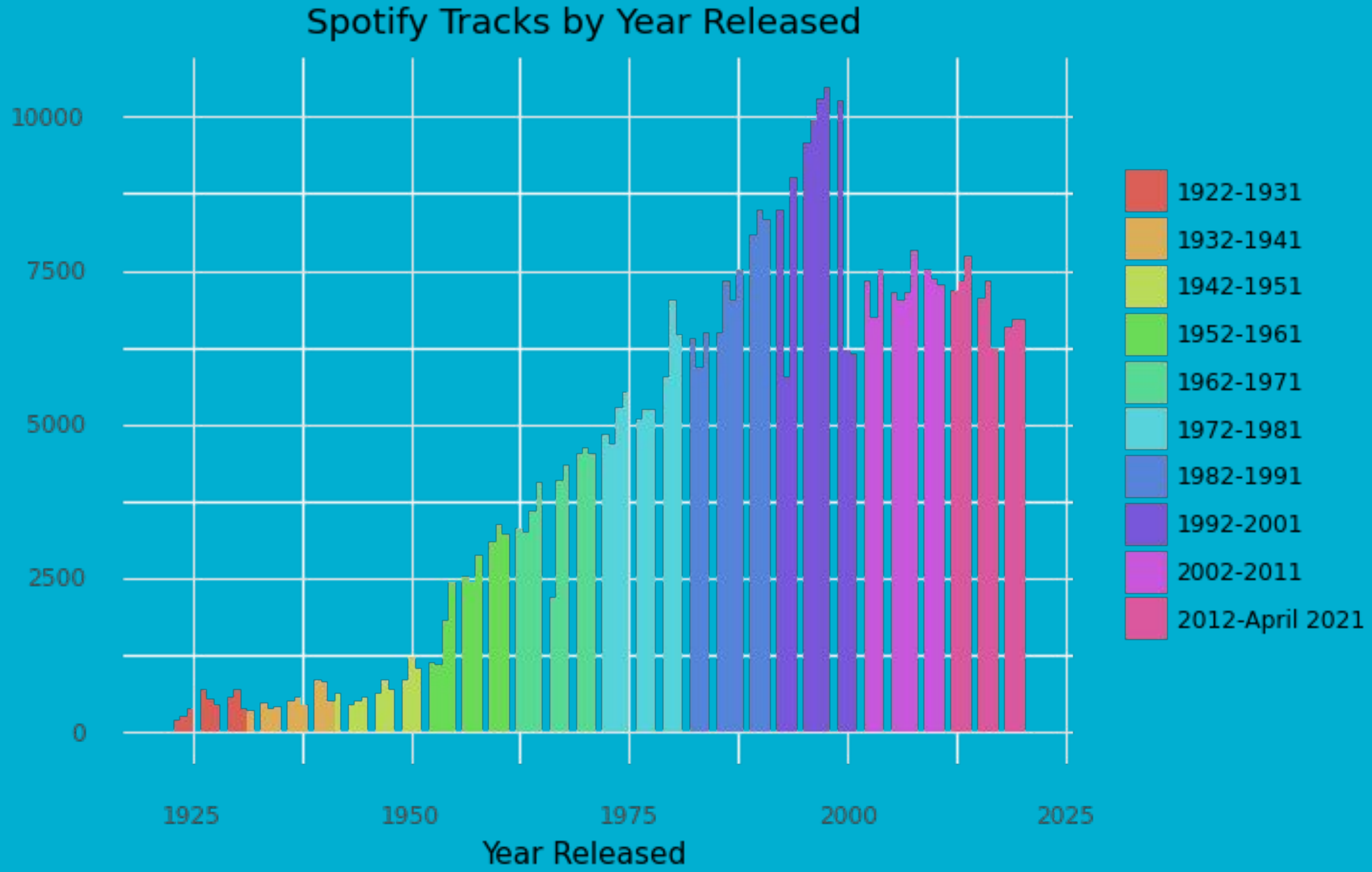


#3

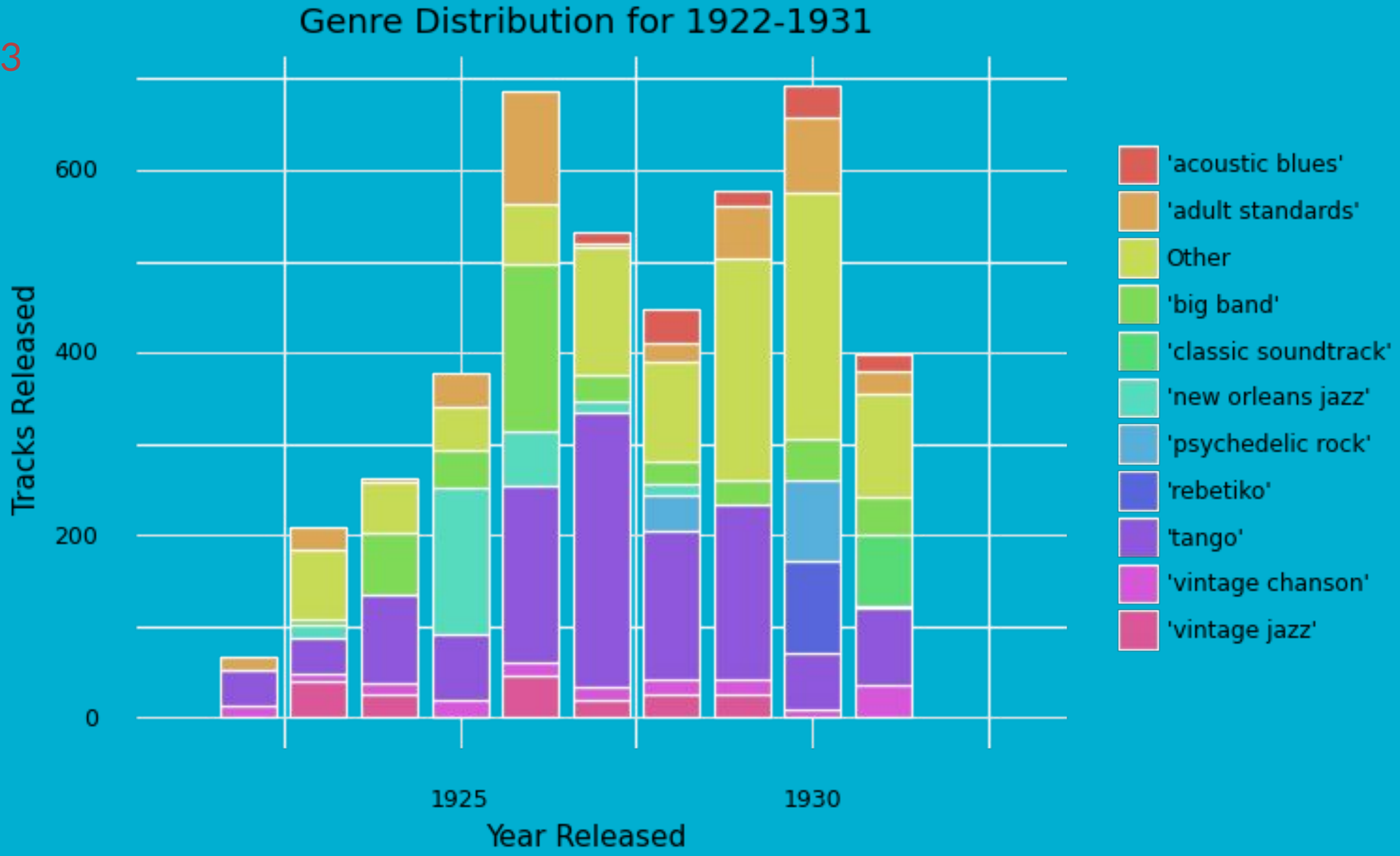
Do different genres see varying levels of popularity, and does this change over time?

- Visualization questions:
- How many songs from each year are on Spotify?
- What are the most popular genres per decade?
- How do those change over years and decades?
- Most produced genres on Spotify?
- All visualized through bar charts

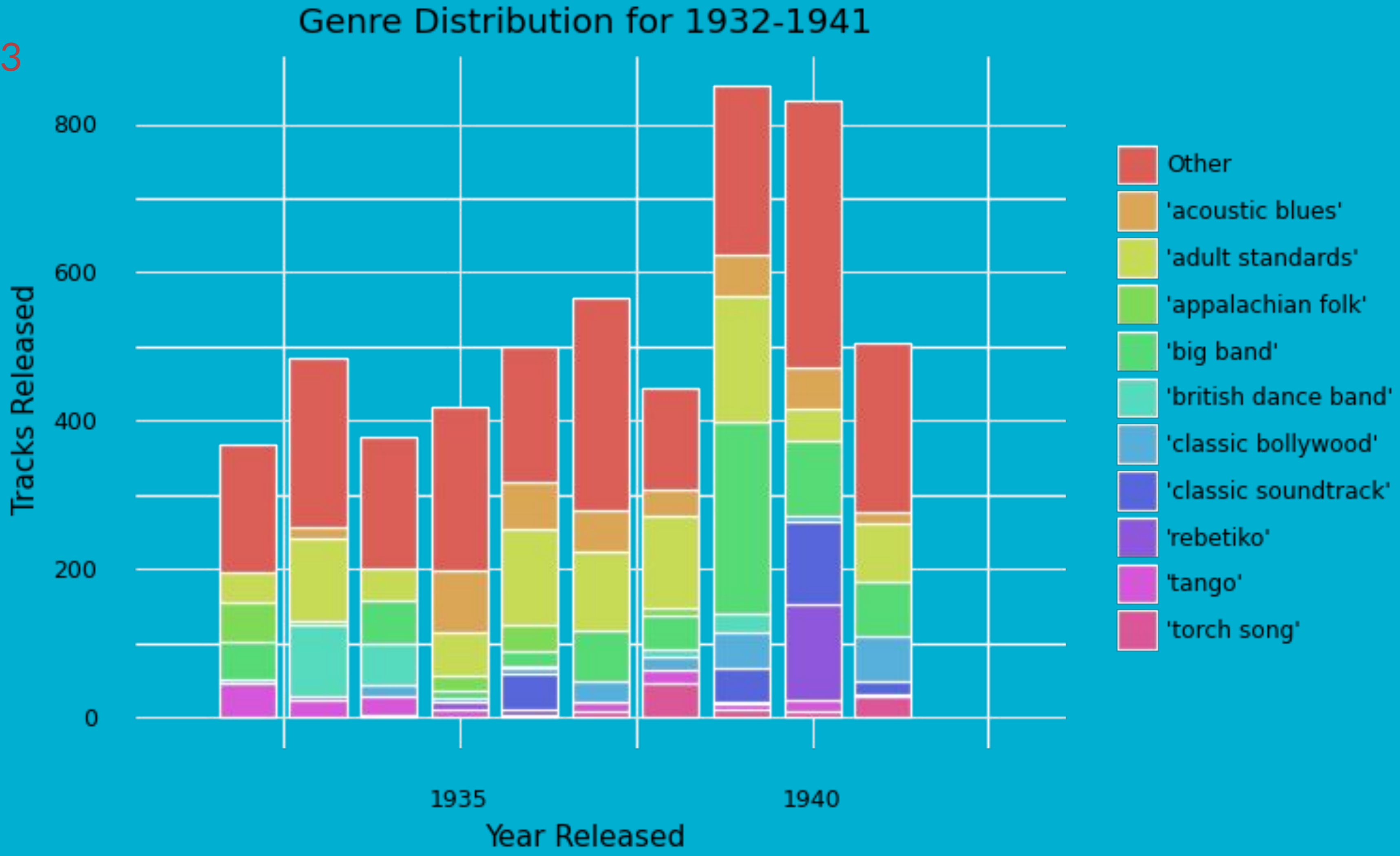
#3


Spotify Tracks by Year Released
10000
7500
5000
2500
0
1925
1950
1975
2000
2025
Year Released
1922-1931
1932-1941
1942-1951
1952-1961
1962-1971
1972-1981
1982-1991
1992-2001
2002-2011
2012-April 2021

#3

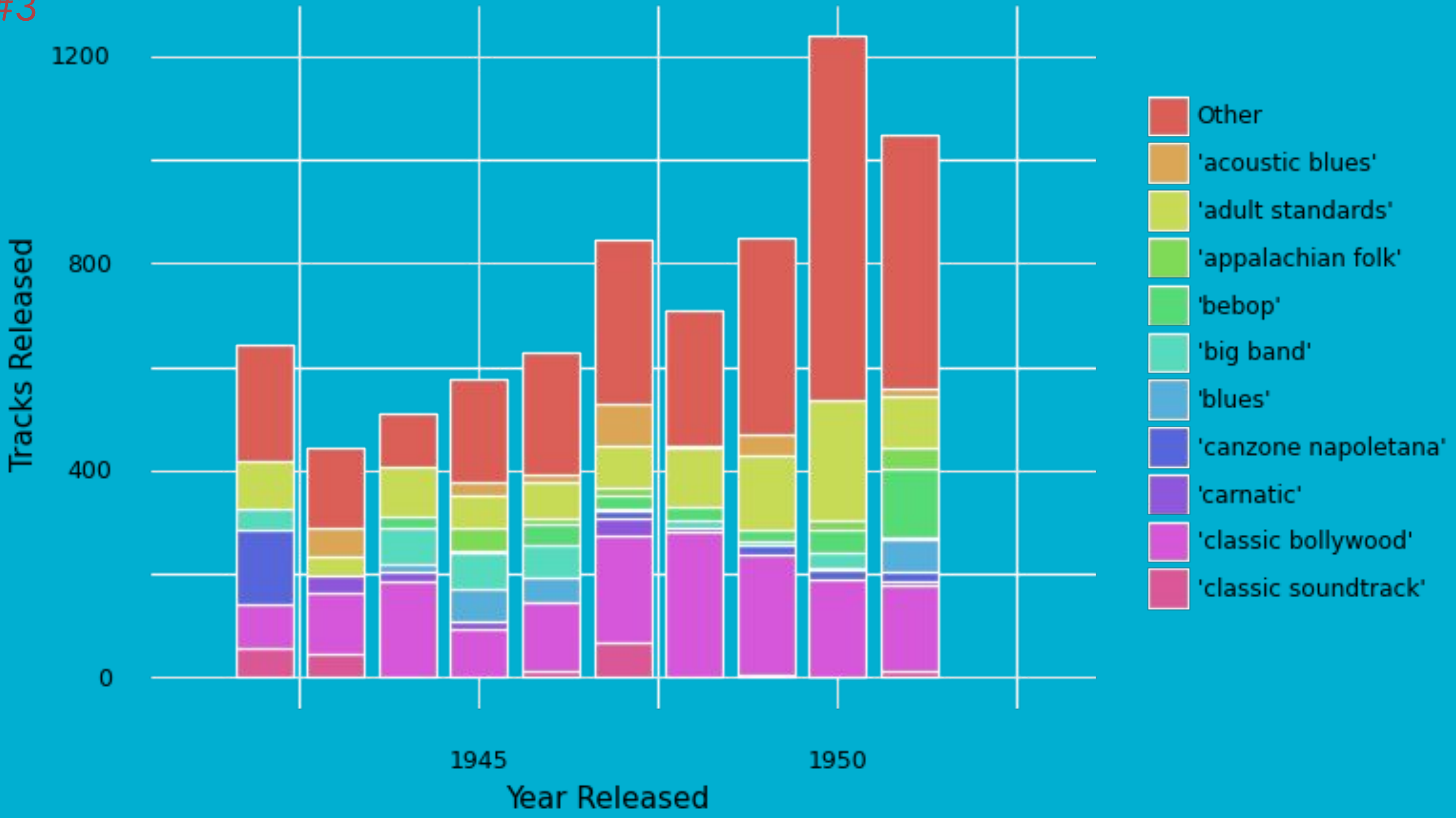

Genre Distribution for 1922-1931
Tracks Released
600
400
200
0
1925
1930
Year Released
'acoustic blues'
'adult standards'
Other
'big band'
'classic soundtrack'
'new orleans jazz'
'psychedelic rock'
'rebetiko'
'tango'
'vintage chanson'
'vintage jazz'

#3

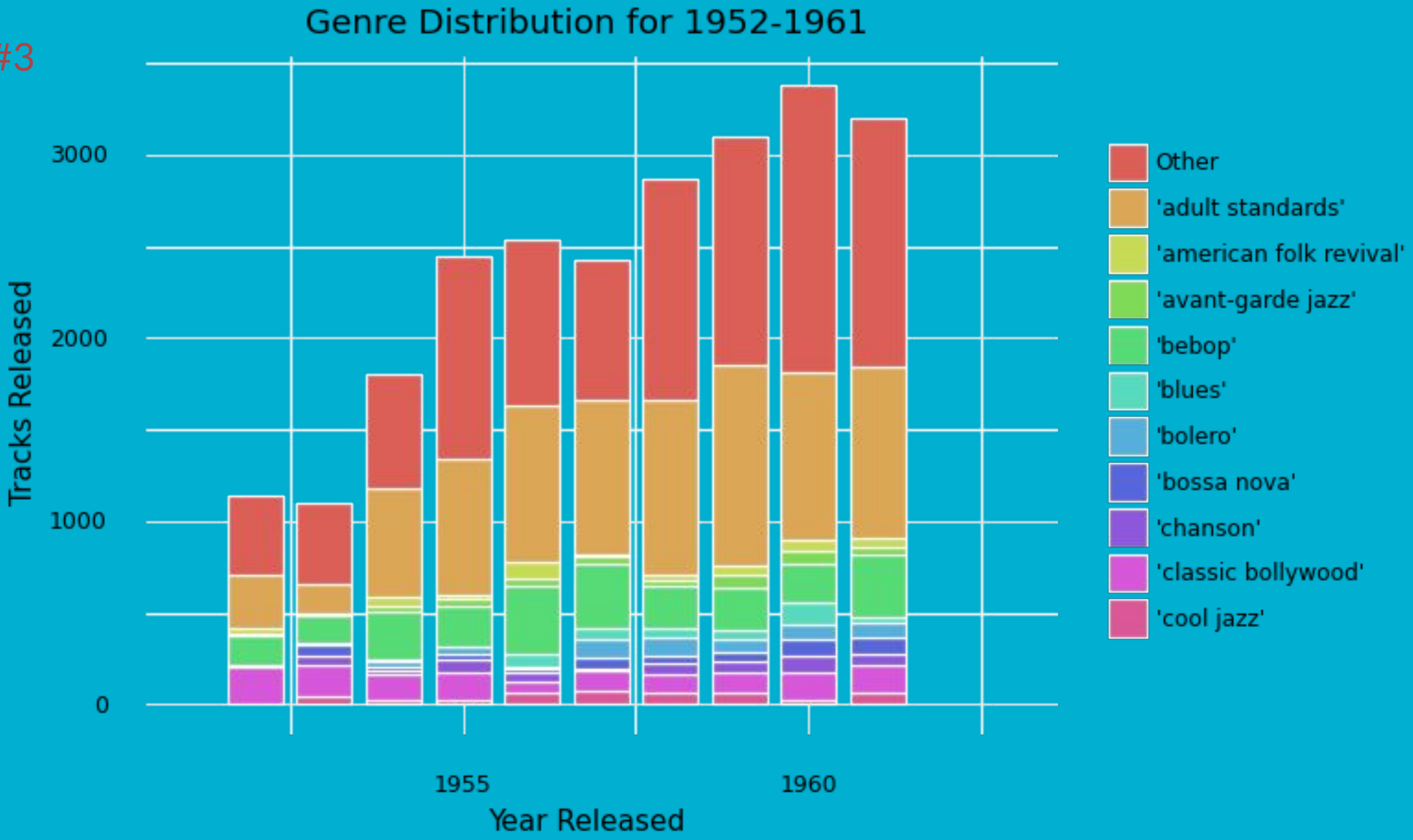

Genre Distribution for 1932-1941
Tracks Released
0
200
400
600
800
1935
1940
Year Released
Other
'acoustic blues'
'adult standards'
'appalachian folk'
'big band'
'british dance band'
'classic bollywood'
'classic soundtrack'
'rebetiko'
'tango'
'torch song'

Genre Distribution for 1942-1951

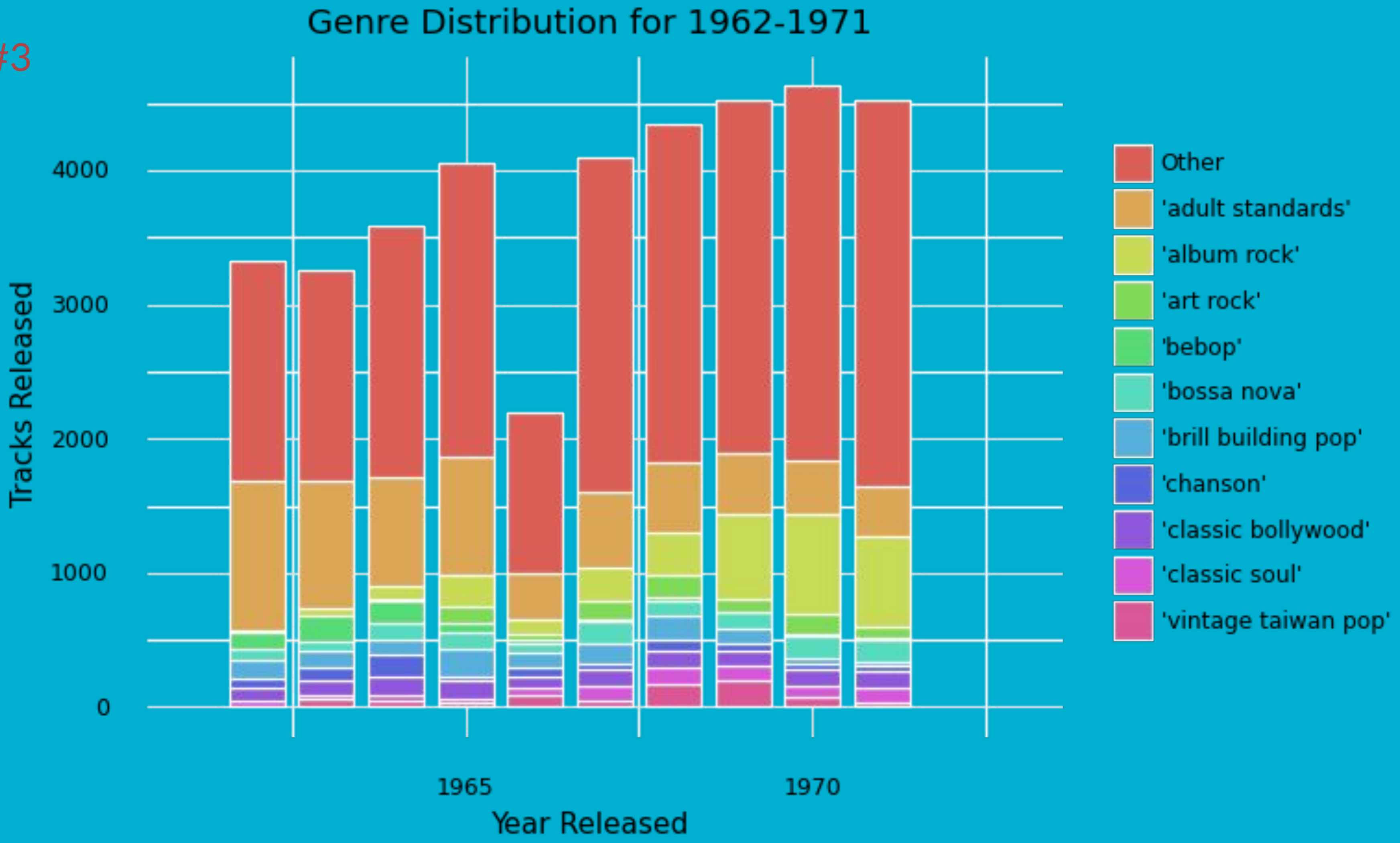
#3



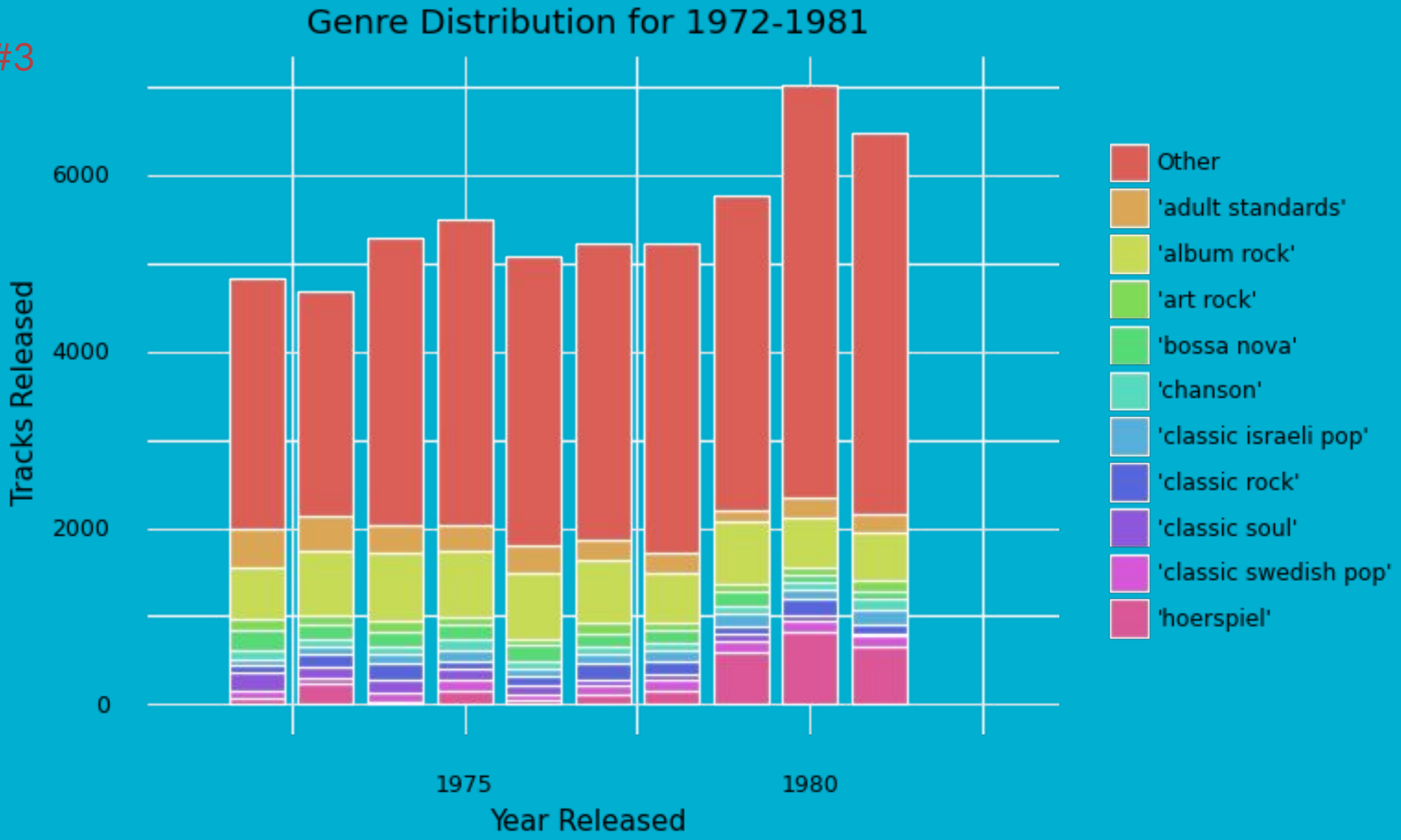
#3


Genre Distribution for 1952-1961
Tracks Released
3000
2000
1000
0
1955
1960
Year Released
Other
'adult standards'
'american folk revival'
'avant-garde jazz'
'bebop'
'blues'
'bolero'
'bossa nova'
'chanson'
'classic bollywood'
'cool jazz'

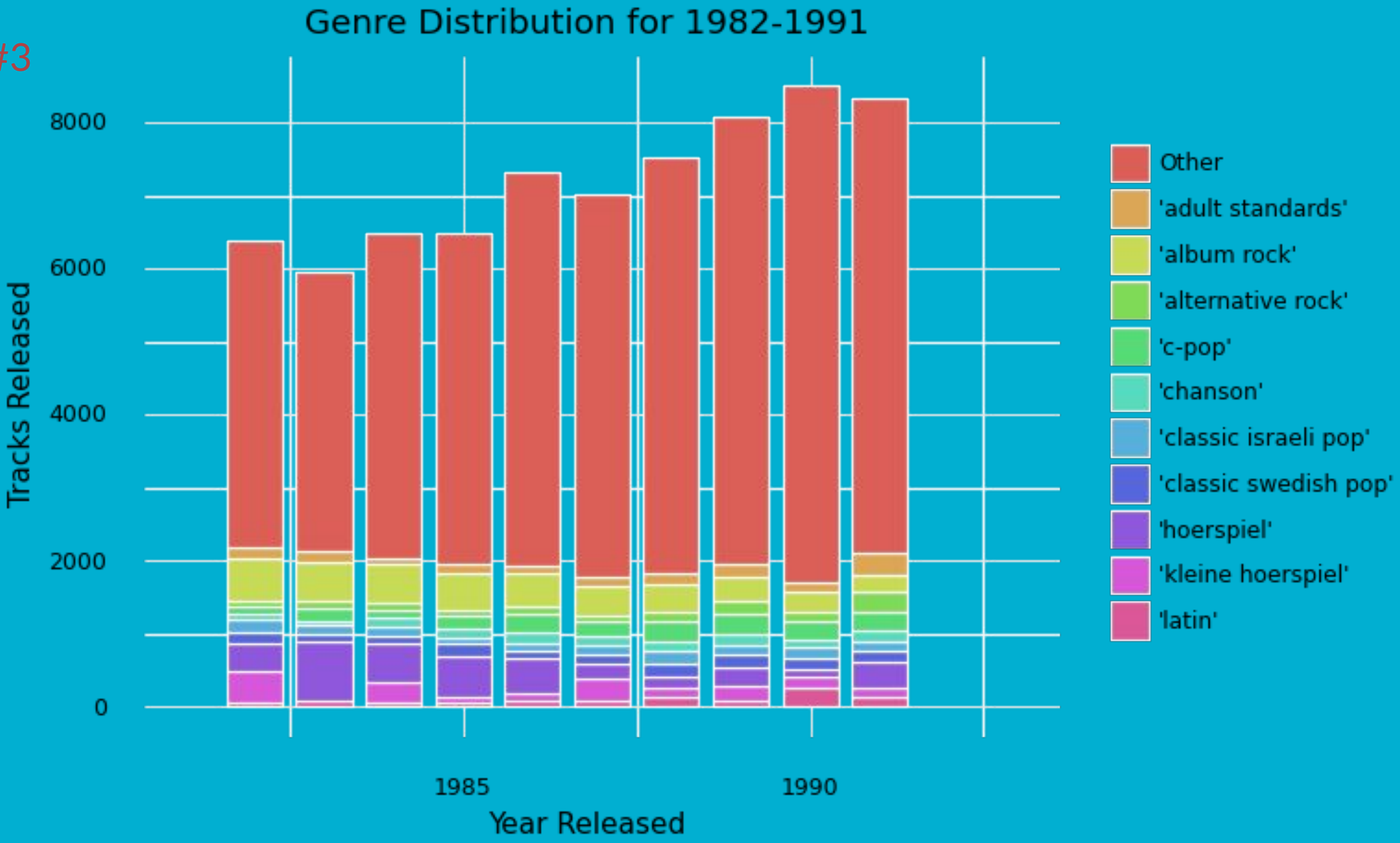
#3


Genre Distribution for 1962-1971
Tracks Released
0
1000
2000
3000
4000
1965
1970
Year Released
Other
'adult standards'
'album rock'
'art rock'
'bebop'
'bossa nova'
'brill building pop'
'chanson'
'classic bollywood'
'classic soul'
'vintage taiwan pop'

#3

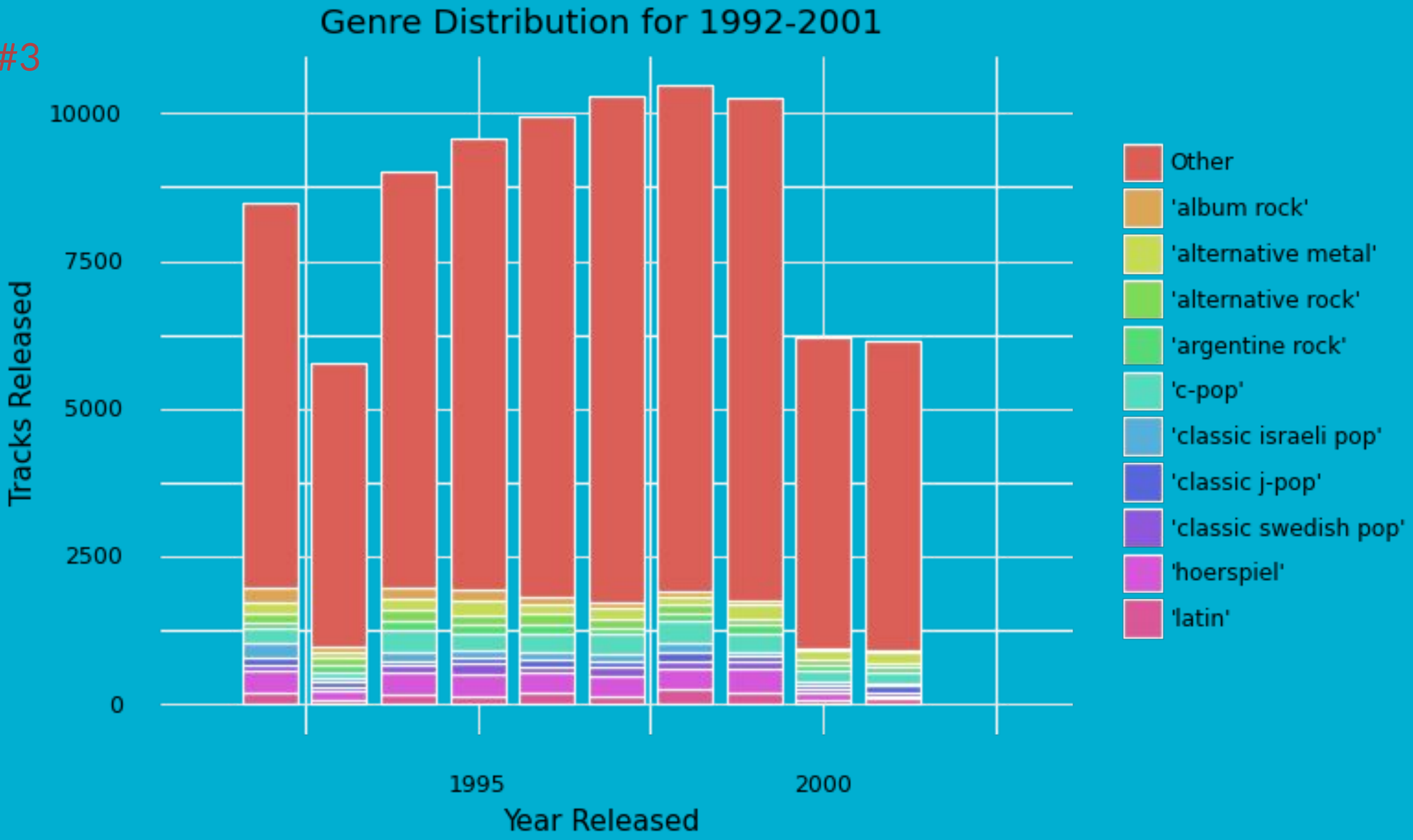

Genre Distribution for 1972-1981
Tracks Released
0
2000
4000
6000
Year Released
1975
1980
Other
'adult standards'
'album rock'
'art rock'
'bossa nova'
'chanson'
'classic israeli pop'
'classic rock'
'classic soul'
'classic swedish pop'
'hoerspiel'

#3



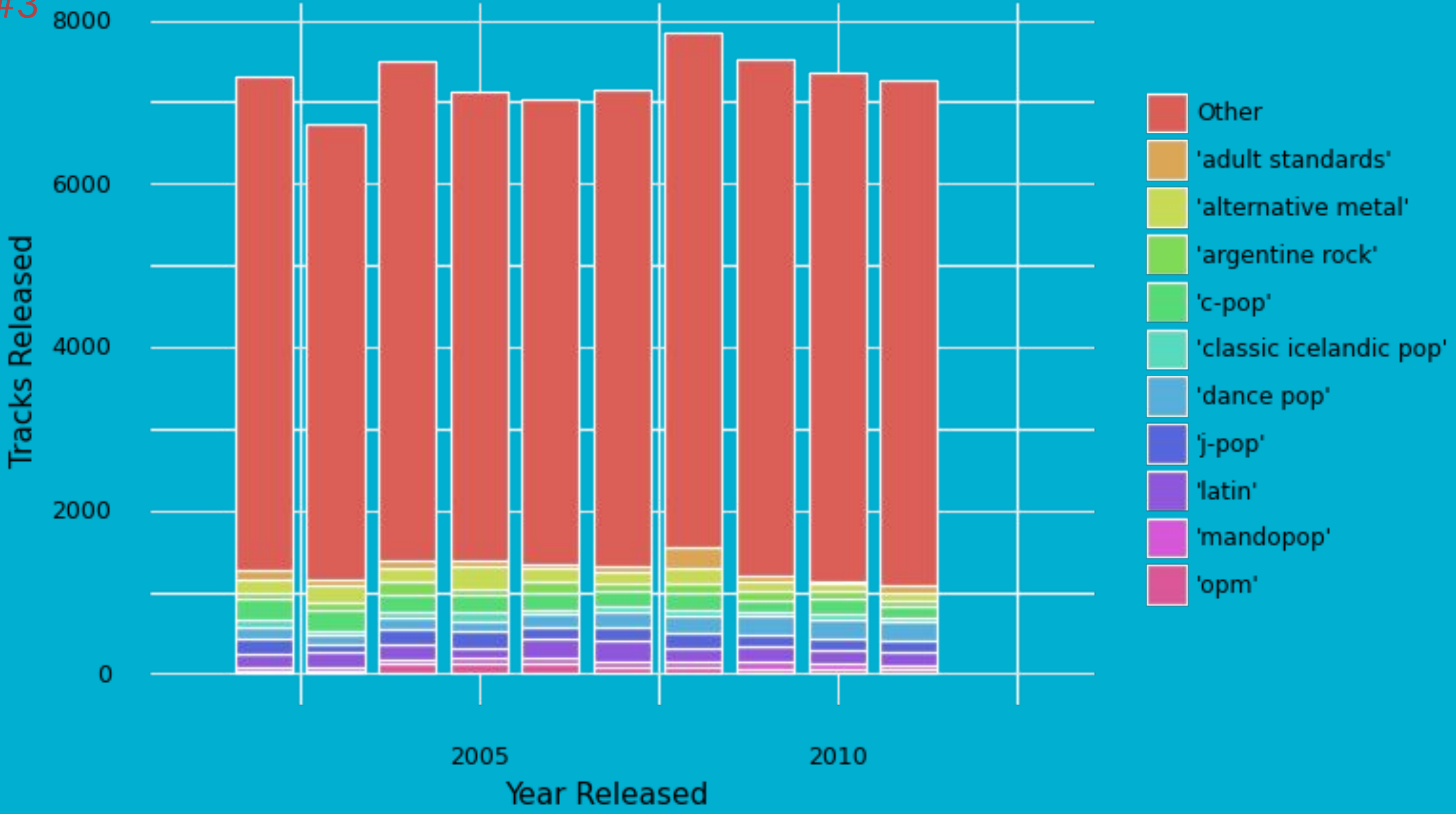
Genre Distribution for 1982-1991

#3

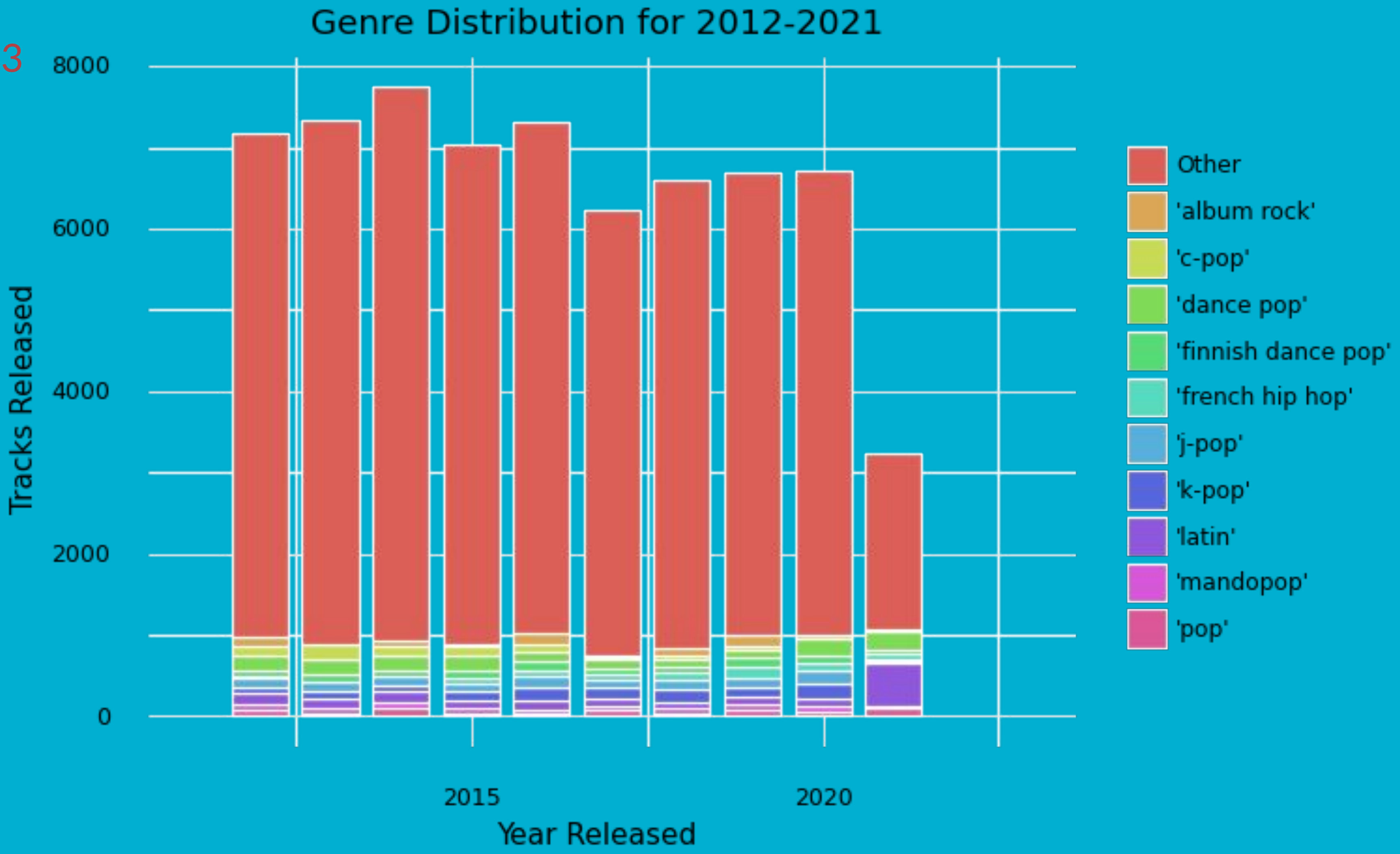

Genre Distribution for 1992-2001
Tracks Released
10000
7500
5000
2500
0
1995
2000
Year Released
Other
'album rock'
'alternative metal'
'alternative rock'
'argentine rock'
'c-pop'
'classic israeli pop'
'classic j-pop'
'classic swedish pop'
'hoerspiel'
'latin'

#3

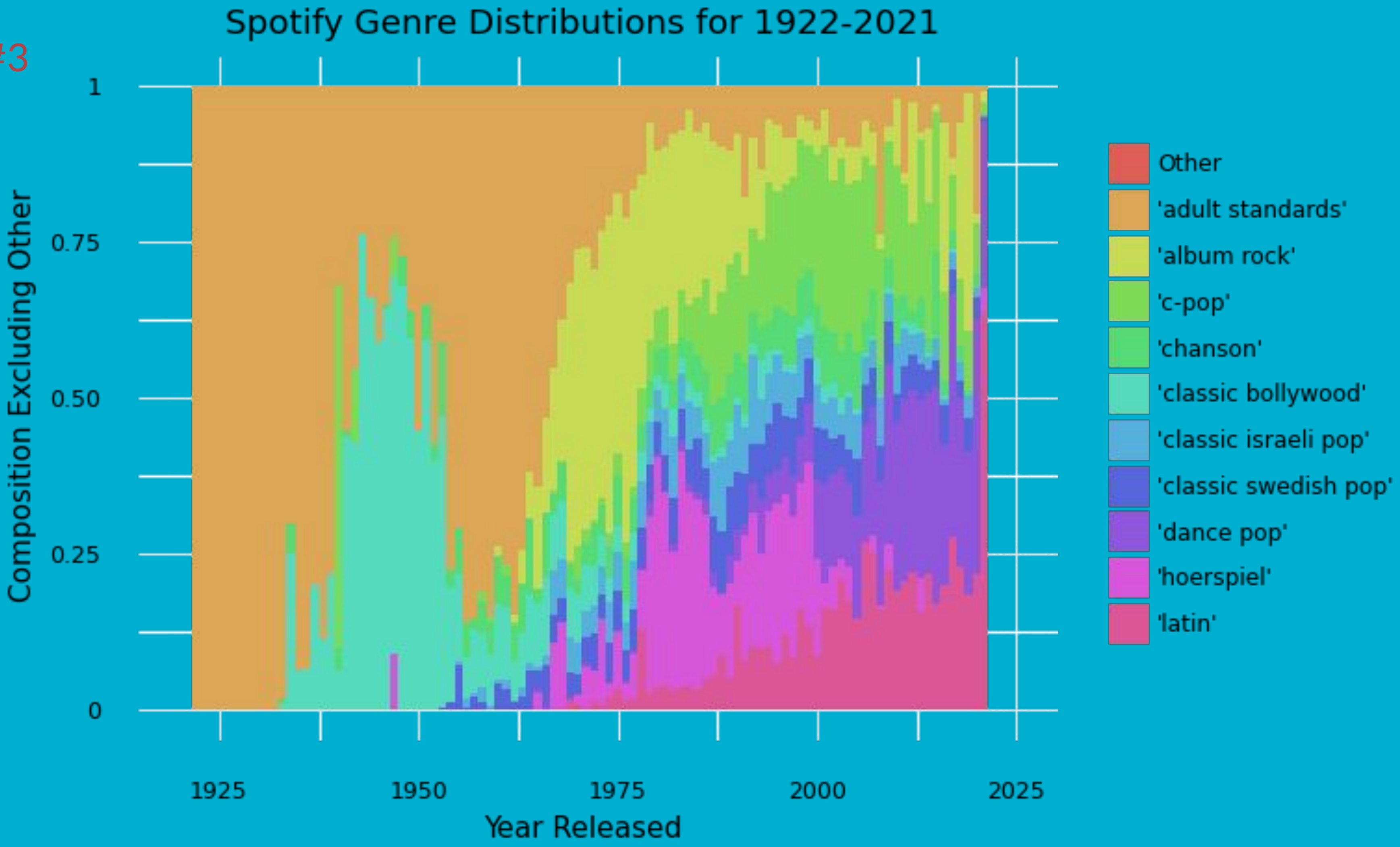
Genre Distribution for 2002-2011



#3


Genre Distribution for 2012-2021
Tracks Released
8000
6000
4000
2000
0
2015
2020
Year Released
Other
'album rock'
'c-pop'
'dance pop'
'finnish dance pop'
'french hip hop'
'j-pop'
'k-pop'
'latin'
'mandopop'
'pop'

#3


Spotify Genre Distributions for 1922-2021
Composition Excluding Other
1
0.75
0.50
0.25
0
1925
1950
1975
2000
2025
Year Released
Other
'adult standards'
'album rock'
'c-pop'
'chanson'
'classic bollywood'
'classic israeli pop'
'classic swedish pop'
'dance pop'
'hoerspiel'
'latin'

Conclusion

- Linear model was okay
- HAC method did not work here
- Visualizations showed clear shifts over time
- In the future:
    - Use different predictive model, such as Decision Trees
    - Try different clustering algorithms
    - Try to get better genre classifications from data