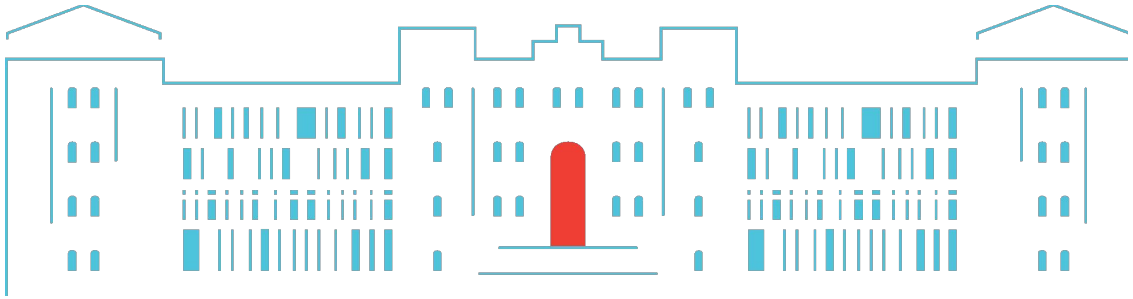
Project: Spotify Artist Collaborations.

Manoj Nethenahalli Dhanpal -611794

Varad Santosh Kulkarni - 612254

Richart Seel - 598756

TUHH
Technische
Universität
Hamburg



16-01-25

Dataset “nodes.csv”

TUHH

1. nodes.csv
nodes.shape = (156422,6)

Attributes	Object_type	is_null
spotify_id	object	0
name	object	4
followers	float64	4
popularity	int64	0
genres	object	0 but we have [] strings
chart_hits	object	136778

Pre-processing the dataset nodes.csv

TUHH

- Removed the rows which has no artist “Name” and “Followers”
- Filter out all non english names (found 143800 english names)
- Extracted all the attributes except ‘chart_hits’ and stored in a csv (english_artist_with_followers.csv)
- Extracted all records where ‘chart_hits’ attribute is not null and stored in a csv (artist_with_country_rank.csv)
- Convert 193 country code to country names and assign respective rank.

Pre-processing the dataset nodes.csv

TUHH

- Popularity attribute max value is 100 so it's a percentile [0-100]
- Country wise rank is between [1 to 379]
- Filter and extract the genres and fill [] genres with 'unknown'
-

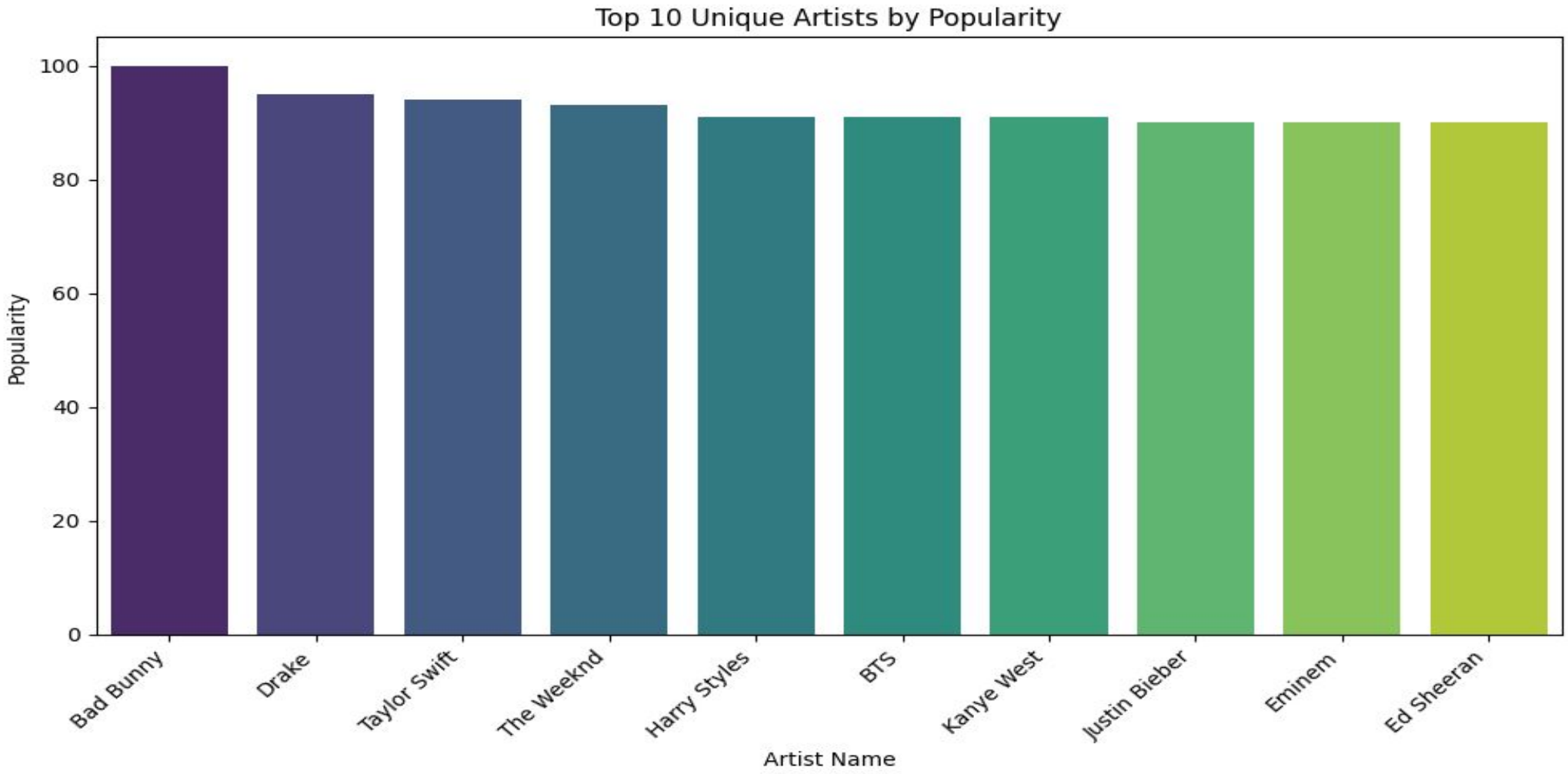
```
[135]: df['genres']

[135]: 0                              ['nordic house', 'russelater']
       1                    ['christlicher rap', 'german hip hop']
       2                                                        []
       3          ['dancehall', 'lovers rock', 'modern reggae', ...
       4          ['classic swedish pop', 'norrbotten indie', 's...
                                       ...
       17775                   ['turkish hip hop', 'turkish trap']
       17776     ['finnish dance pop', 'finnish pop', 'iskelma']
       17777                                       ['german pop']
       17778                                   ['urbano espanol']
       17779    ['chilean rock', 'rap chileno', 'reggae en esp...
       Name: genres, Length: 17780, dtype: object
```

Pre-processing the dataset nodes.csv

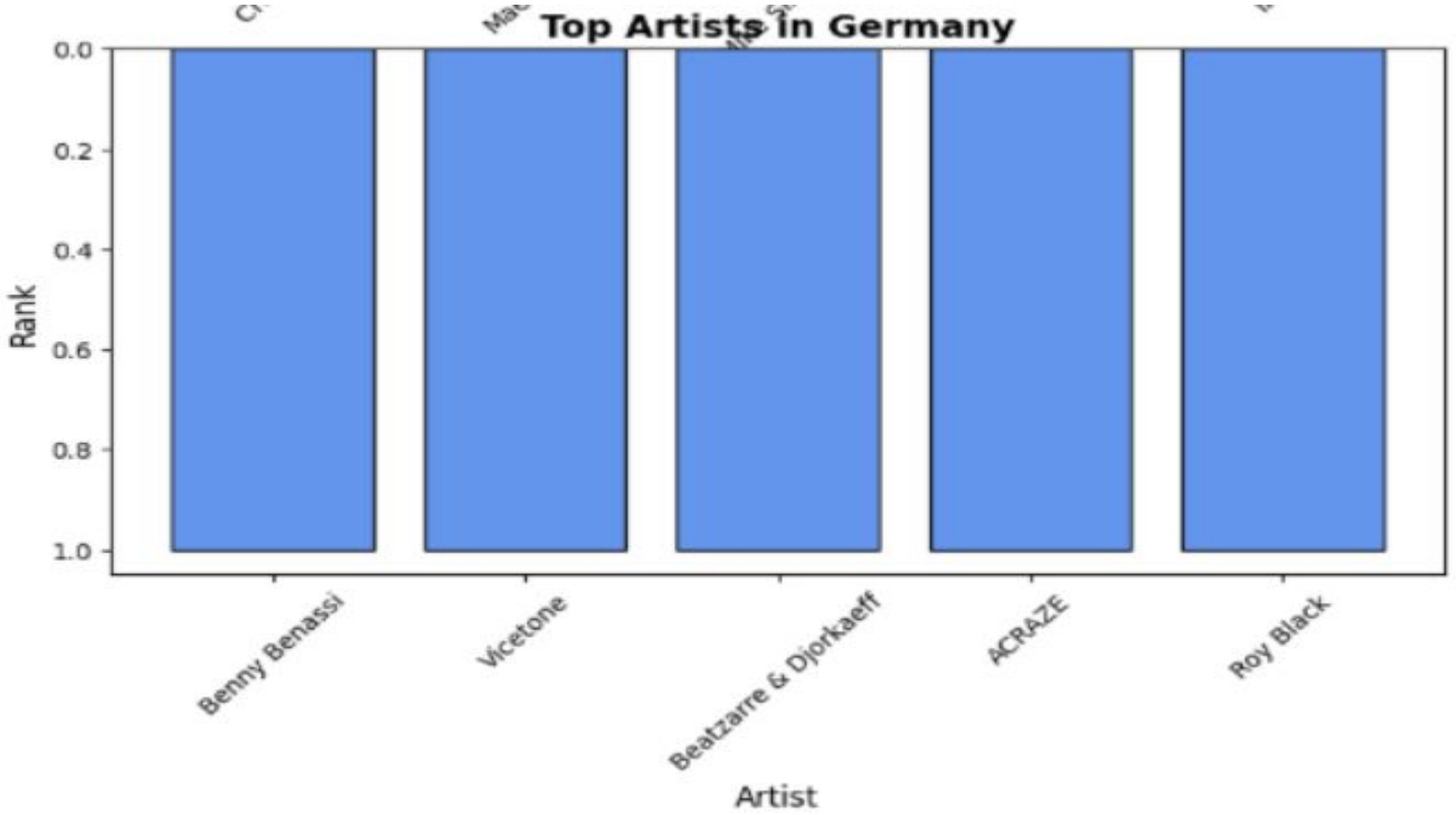
TUHH

- We have 2500+ unique genres so converting that to 15 prominent genres.
- One-hot-encoding genres is better as it’s not an ordinal data.
- By this data is completely cleaned.
- The same process can be done for the english_artist_with_followers.csv file!

Top 10 artist with popularity.



Top 10 artist with popularity in Germany



Dataset “edges.csv”

TUHH

1. edges.csv
edges.shape = (300386,2)
Edges are undirected and id_0 <id_1 according to alphabetical order

Attributes	Object_types	is_null
id_0 (artist_a id)	object	0
id_1 (artist_b id)	object	0

Pre-processing the edges.csv dataset.

TUHH

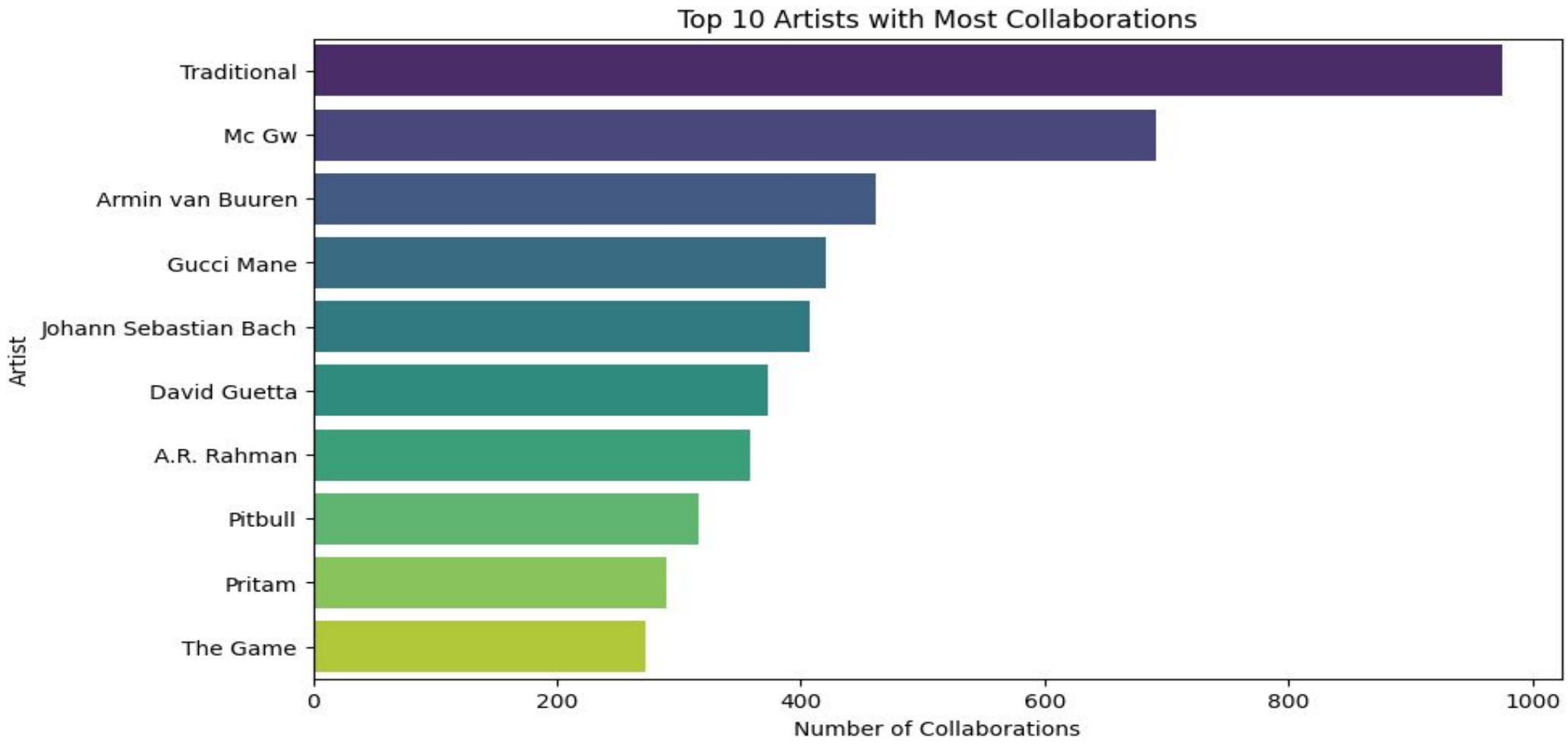
- Creating a mapping spotify_id and names in nodes.csv
- Create artist_a and artist_b attribute in edges.csv and map them from modes.csv
- Remove null values (count 27 in artist_a and count 37 in artist_b)
- Filtering rows if both artist_a and artist_b are repeated more than ones (count 2677 duplicates)
- Removing all other languages names except english (We found 266046 english names)

What did we found ?

TUHH

- Successfully mapped the artist with the spotify id
.
- Removed duplicates.
- Keeping only English language for better results.
- Found top 10 artists who had most collaborations with other artists.
- Found around 25541 unique artists_a collaborated with artist_b

Most collaborative artists:



Questions:

TUHH

- We have 137k missing chart_hits out of 155k records which contain country and also weather the song rank in specific country.
- Approach 1: we consider only 20k artists and find the collaboration with rank specific to country?
  Pros: It will be not so complex.
  Cons: Testing we may not get better results as artist name
- Approach 2: we ignore chart_hits attribute and focus on global ranking and collaboration between the artist ?
  Pros: Better training approach with 155k records.
  Cons: We will not know whether the song will be ranked 1 or 200th in a specific country.

TUHH

THANK YOU!