



Spotify Data Pipeline

Jason Kim

Metis

April 30, 2021





What is Spotify?

- Spotify is an audio streaming and media services provider that hosts tens of millions of songs and over 1 million artists
- Massive Audience Pool: 345 million active users



Popularity

Song popularity is a metric that artists may seek to maximize.

On Spotify, Song Popularity (0-100):

- Total number of plays
- How recent the plays are

TOP STREAMED ALBUMS GLOBALLY

1. *YHLQMDLG*

2. *After Hours*

3. *Hollywood*
Post Malone

4. *Fine Line*

5. *Future Nostalgia*
Dua Lipa



#2020WRAPPED

Data Ingest and Storage

kaggle™



Data
Ingest

- Data Source: Spotify Data from Kaggle (~600,000 rows of tracks data, ~1.1m rows of artist data)

SQL

- CSV Files loaded into DB Browser as tables

Storage

- Tables of data stored into SQL database

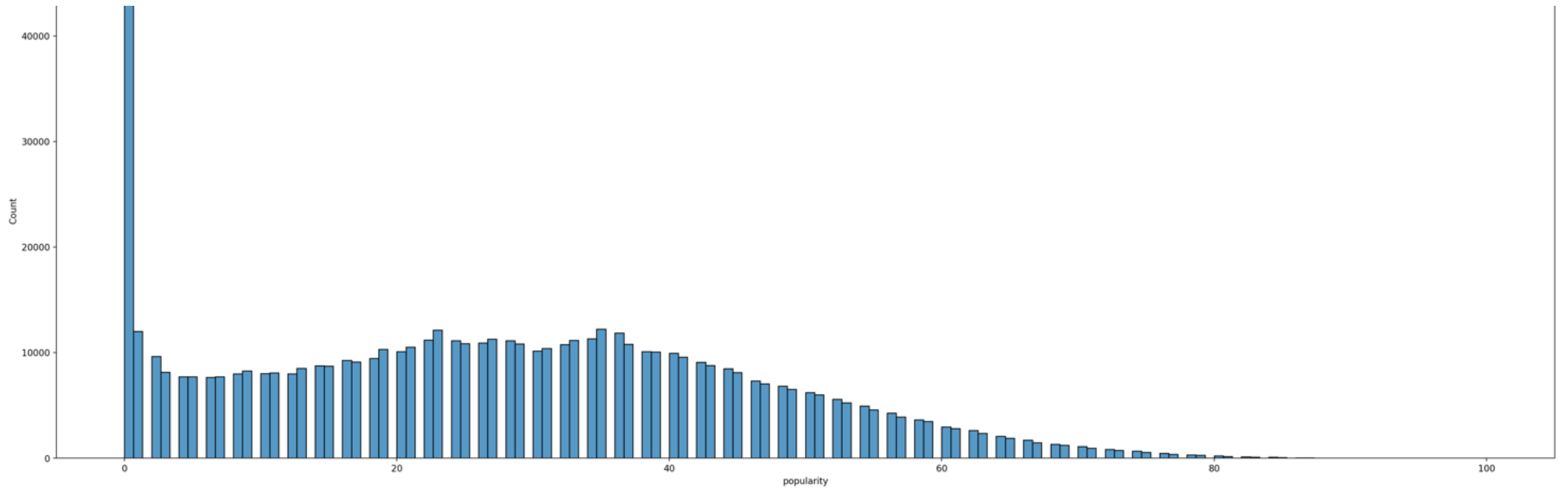


Processing

Processed data using pandas

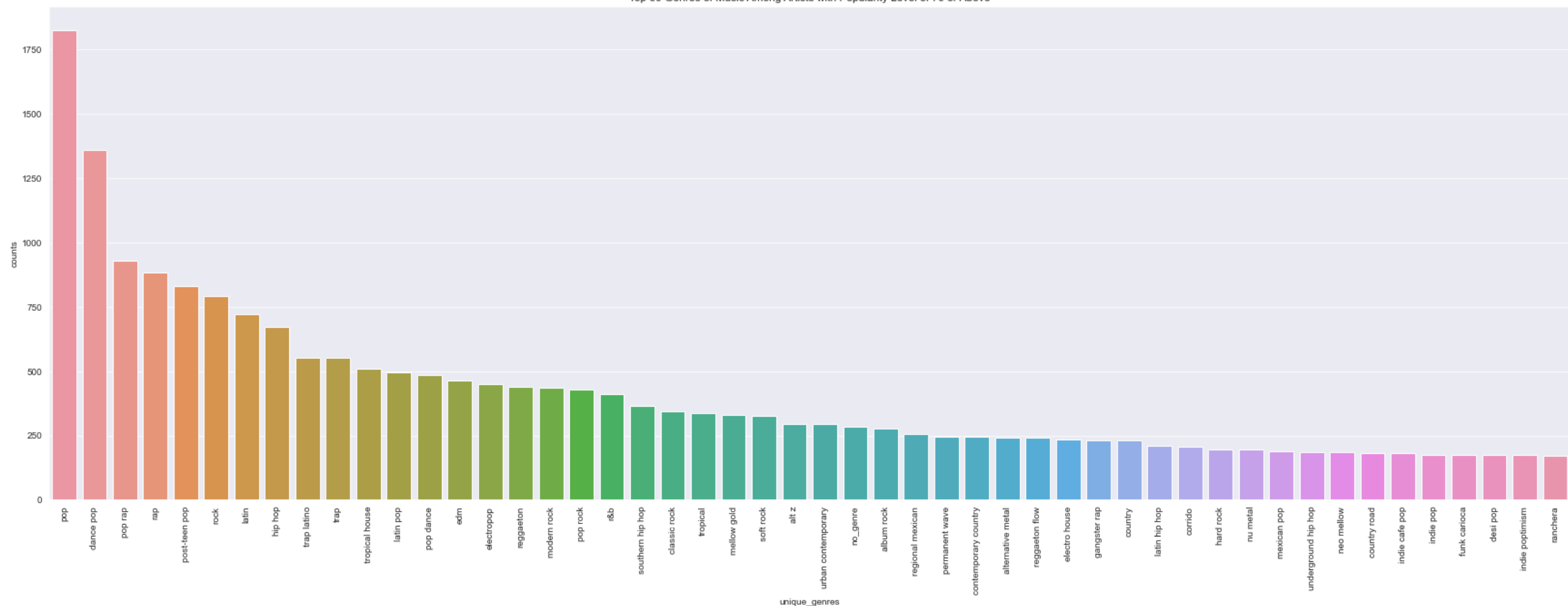


Distribution of Popularity Level Across All Songs



Top 50 Genres of Music Among Artists with Popularity ≥ 70

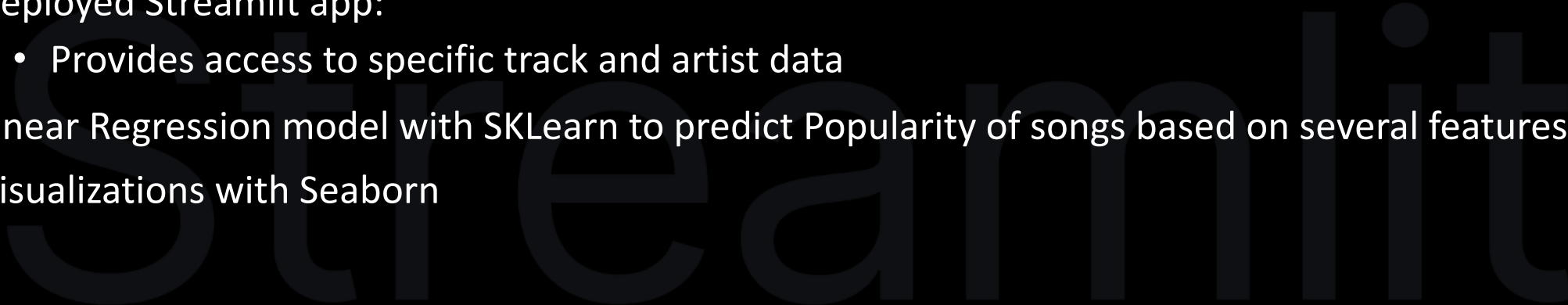
Top 50 Genres of Music Among Artists with Popularity Level of 70 or Above






Deployment



- Deployed Streamlit app:
 - Provides access to specific track and artist data
 - Linear Regression model with SKLearn to predict Popularity of songs based on several features
 - Visualizations with Seaborn
- 

Spotify Data Explorer Demo



Spotify Data Explorer

Random Snapshots of Tracks & Artists Dataframes:

Random Track Data Snapshot

Generate

Random Artists Data Snapshot

Generate

Song Lookup

Song Name

Song Artist

Search

Artist Lookup

Artist

Search

Future Work

- More accurate prediction model with further feature engineering with Spark ML
- Create access to further visualizations on Streamlit/Flask App
- Integration with Spotify API and further utilization of Spark

