



Spotify Data Pipeline

Jason Kim

Metis

30 April 2021





Popularity

Song popularity is a metric that artists may seek to maximize

TOP STREAMED ALBUMS GLOBALLY

1. *YHLQMDL*

2. *After Hours*

3. *Hollywood's Bleeding*
Post Malone

4. *Fine Line*

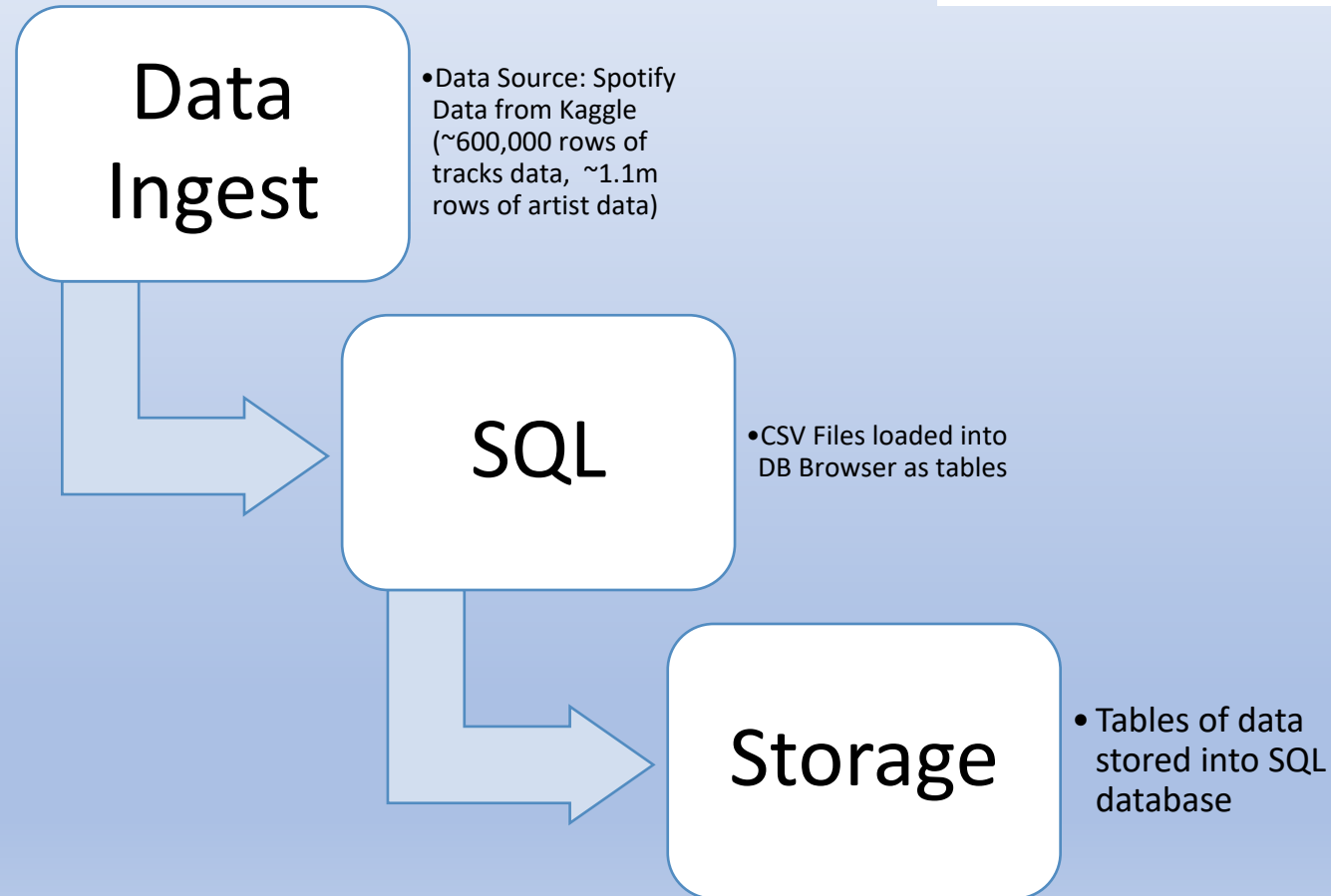
5. *Future Nostalgia*
Dua Lipa



#2020WRAPPED

Data Ingest and Storage

kaggle™



Processing

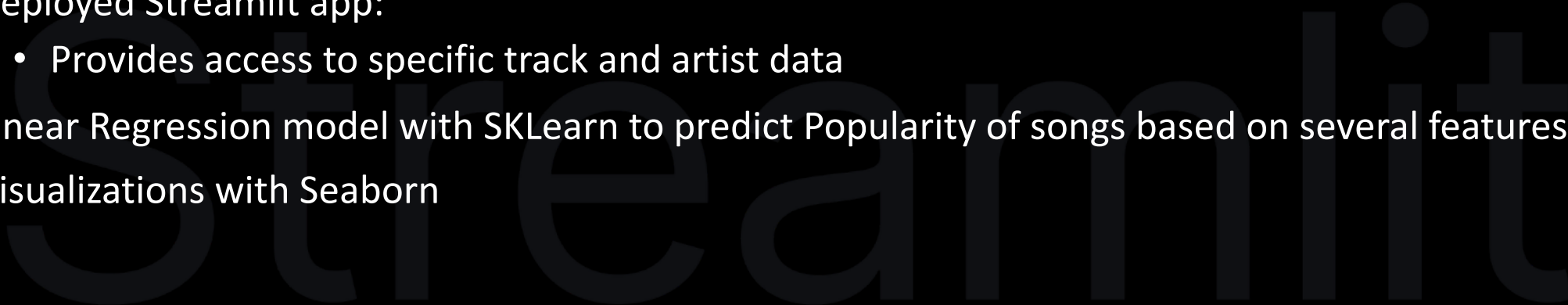
Processed data using pandas



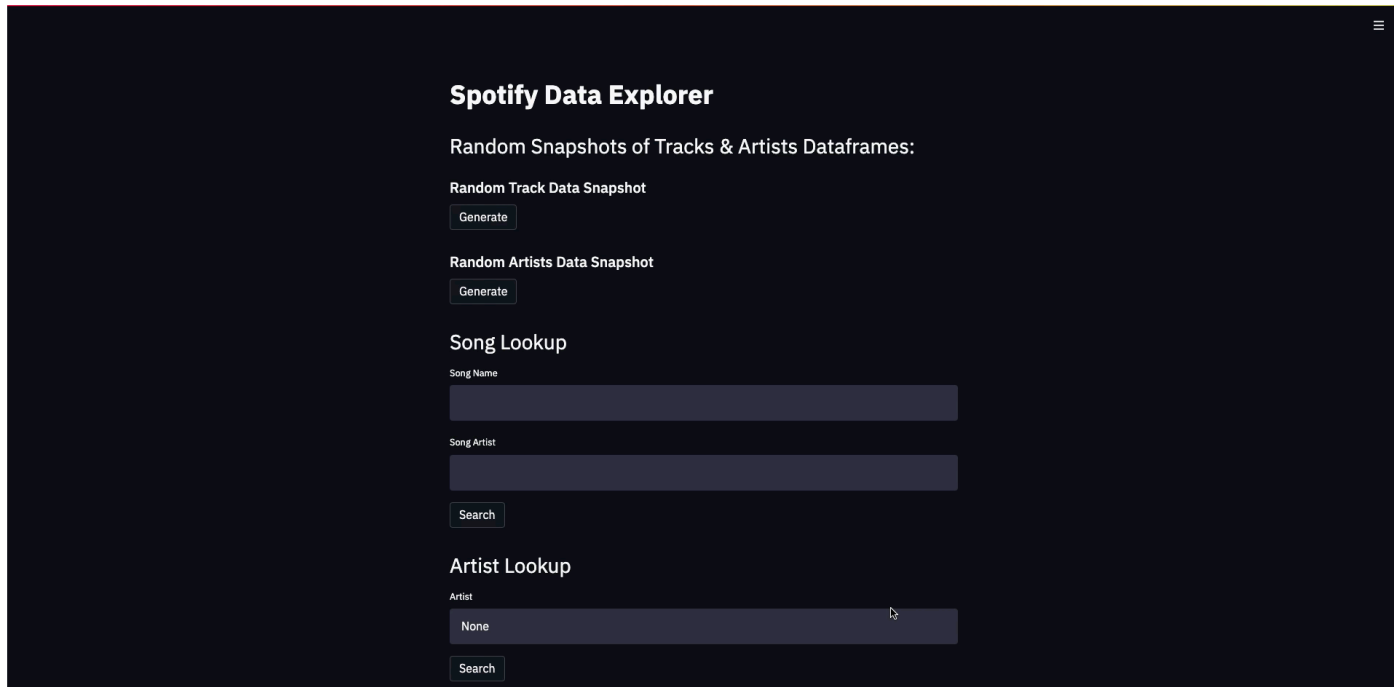


Deployment



- 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



The screenshot shows a web application titled "Spotify Data Explorer" with a dark theme. It features three main sections: "Random Snapshots of Tracks & Artists Dataframes", "Song Lookup", and "Artist Lookup". Each section has a "Generate" or "Search" button. The "Song Lookup" section has input fields for "Song Name" and "Song Artist". The "Artist Lookup" section has an input field for "Artist" with the value "None" and a "Search" button. A hamburger menu icon is visible in the top right corner.

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

None

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

