The Rise of K-Pop

Can Spotify's audio features show what makes certain music popular?

Presented by:
Jessica Veilleaux, Kailey Davis, Kenneth Beadle, Miranda Wylie

Why K-Pop?

Why K-Pop?

K-Pop is short for Korean popular music, and its rise in popularity - as well as the dedication of its fans - has taken social media by storm. The cultural impact of K-Pop fans was even highlighted by CNN in 2020, documenting the political activism of fans of Korean pop idol groups, and is driving international sales by groups like BTS partnering with brands like McDonalds.

With as much social and purchasing power as these fans wield, we wanted to know: will their songs have staying power - and therefore the communities surrounding them?

What do we hope to find?

Is K-Pop that popular in the U.S.?

With headlines and news stories coming out covering it, K-Pop seems like the ultimate new music trend - but will the data show its popularity?

What makes certain songs popular?

Using Spotify’s audio features, we hope to determine if certain features are more likely to cause a song or genre to be popular.

Can we predict how long a song will be popular?

Using our machine learning model, we hope to be able to accurately predict how many weeks a song will stay on the Hot 100 Billboard.

Data Exploration

What we asked:

What data can we find on K-Pop specifically, and popular music generally?

What we learned:

Rather than determine a band’s popularity by scraping a million different sources manually, we decided to rely on the methods of the Billboard Hot 100.

Also, artist and song information tends to be separate, so we would need to go to the source: Spotify.

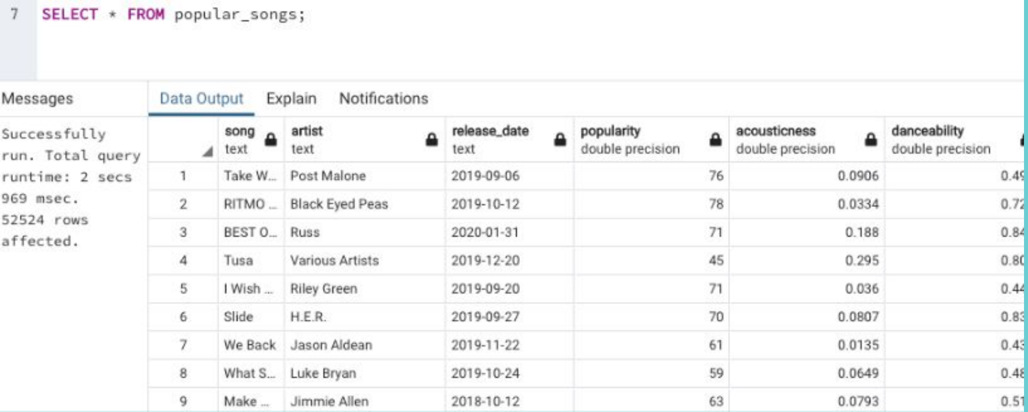
Google

kaggle™

Spotify®

Data Sources

- **CSV’s from Kaggle**
  - Billboard Hot 100 charts from 1958 - summer 2021
  - Misc. Spotify artist and track information
  - Used to explore data features and what questions we could ask
- **Spotify API**
  - Used to pull audio features on the specific Hot 100 songs
  - Used to create final, clean database in Postgres



```
SELECT * FROM popular_songs;
```

```
Successfully
run. Total query
runtime: 2 secs
969 msec.
52524 rows
affected.
```

	song text	artist text	release_date text	popularity double precision	acousticness double precision	danceability double precision
1	Take W...	Post Malone	2019-09-06	76	0.0906	0.49
2	RITMO ...	Black Eyed Peas	2019-10-12	78	0.0334	0.72
3	BEST O...	Russ	2020-01-31	71	0.188	0.84
4	Tusa	Various Artists	2019-12-20	45	0.295	0.80
5	I Wish ...	Riley Green	2019-09-20	71	0.036	0.44
6	Slide	H.E.R.	2019-09-27	70	0.0807	0.83
7	We Back	Jason Aldean	2019-11-22	61	0.0135	0.43
8	What S...	Luke Bryan	2019-10-24	59	0.0649	0.48
9	Make ...	Jimmie Allen	2018-10-12	63	0.0793	0.51

Machine Learning Model

Random Oversampling + SMOTE	Linear Regression	Decision Tree	Neural Network	K-Nearest Neighbors
Use oversampling techniques to tackle imbalance of data. Imbalance is caused by Spotify dataset of 600,000 tracks. We do not know which tracks were /are on the top 100 billboard yet so, we will choose over_sampling methods in anticipation of there being an abundance of non-top 100 songs.	1. Quick to compute and can be updated easily with new data. 2. Relatively easy to understand	1. A single decision tree is fast to train 2. Robust to noise and missing values 3. RF performs well "out-of-the-box"	1. Can learn very complex relationships 2. Extremely powerful for many domains 3. Hidden layers reduce need for feature engineering	1. Simple 2. Powerful 3. No training involved

According to https://srnghn.medium.com/machine-learning-trying-to-predict-a-numerical-value-8aafb9ad4d36 these are the most common ML models for predicting a numerical outcome. So for the purpose of education it will be good to employ all four to see what could be the best outcome. We believe Decision Tree and Neural Network are going to work best.

Machine Learning Model

Model Selection & Preliminary Data Preprocessing

- After further evaluation of the models, the most promising machine learning model was the Neural Network. The dataset we have is from Spotify data that contains a large amount of musical characteristic values. We initially believed these values to be a bit more complex especially when trying to find the best way to interpret a songs popularity. Due to these factors, we chose the Neural Network model. According to an article on Medium.com Neural Networks "*can learn complex patterns... ...and relationships between features that other algorithms cannot easily discover*". Furthermore, the hidden layers reduce the need for feature engineering.
- To prepare the data, it was imported as csv files into jupyter notebook to drop null values and transform values. Then stored in an AWS RDS where it is synced to an SQL database in PgAdmin. Then, we import the data into jupyter notebook again and run it through the model

Feature engineering, selection, and decision-making process

- Because we chose the Neural Network, we did not find it necessary to engineer the features. To create our features we set X=df.drop(columns = "weeks_on_board") and y=df["weeks_on_board"] since we're predicting if a song will be on top 100

Training & Testing Sets

- We used the train_test_split() model from sklearn.model_selection. While instantiating the model we declared a test size of 25%. Since we have a large dataset with tens of thousands of rows we decided that splitting it 75/25 would be good.

Data Analysis & Dashboard

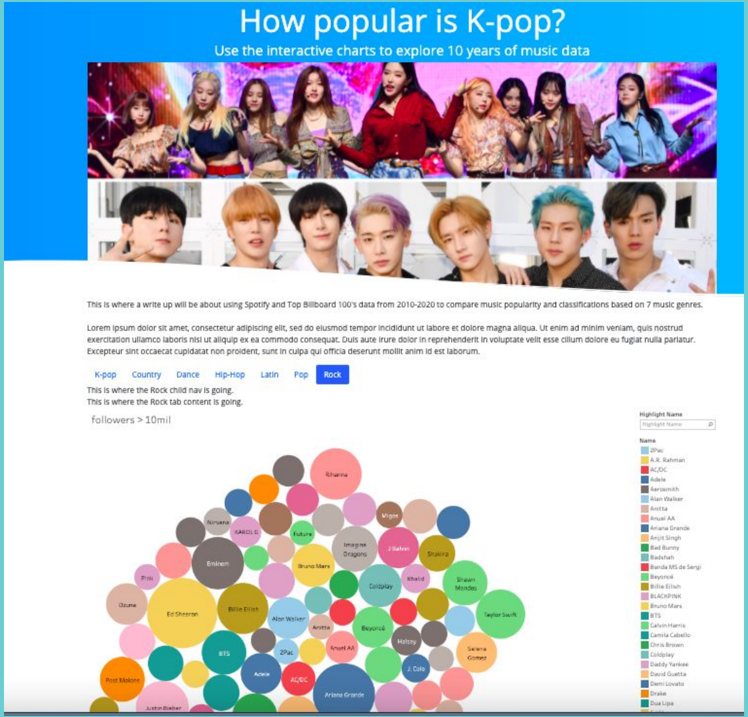
To help us narrow our focus, we are using song data from 2010 - 2020 and the Hot 100 Billboard seven most recognized genres: Country, Dance, Hip-Hop, K-Pop, Latin, Pop, and Rock.

The user will be able to interact with the dashboard by clicking on one of seven genre tabs and then view 3 charts: popularity of artist, the duration an artist (song?) was on the Hot 100 Billboard, and audio features of the song. The charts will be created in Tableau and have an interactive element as well.

Currently the dashboard is connected to our database and runs locally. Running the dashboard through Heroku dashboard is in progress.

Additional possible visuals and questions to ask:

- Line graph noting the rise of K-pop by year
- Comparison of a song's audio features and how long a song stays on the Billboard's Top 100. Are songs with the highest danceability more likely to stay on the Billboard Top 100 longer?



How popular is K-pop?

Use the interactive charts to explore 10 years of music data

This is where a write up will be about using Spotify and Top Billboard 100's data from 2010-2020 to compare music popularity and classifications based on 7 music genres.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This is where the Rock child nav is going.
This is where the Rock tab content is going.