Capstone Project #1

Spotify Million Playlist Dataset

Chris Stellato

GALVANIZE

Data Science Immersive
February 2021
galvanize.com

CONTACT

github.com/chris-stellato
stellatocjs@gmail.com
chris-stellato.com

Project Summary

EDA & Hypothesis Testing

This project aims to obtain, interpret, and gain insight from a dataset containing information from one million Spotify playlists. The goal is to interact with the data in the following ways:

- View raw data to understand format and available information
- Determine what insight can be gained from data and create hypothesis to test
- Wrangle data into useable format
- Perform hypothesis tests and record results
- Create visuals to illustrate testing and results

https://github.com/chris-stellato/capstone_1

Chris Stellato
Capstone Project #1

DATASET

Anonymized Spotify Playlists

Collected: 2010-2017

Format: JSON files

Download:

https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge

The dataset, published by Spotify to encourage research on music recommendation, contains information for one million playlists created between 2010 and 2017. For this project we sample a set of 20,000 playlists.

There are two main types of information available:

- **Playlist metadata**, including number of artists, number of tracks, number of albums, follower count, playlist duration, and collaboration status
- **Track lists** for each playlists, including song name, artist name, album name and duration

Chris Stellato
Capstone Project #1

Motivation

How can we help grow Spotify’s business?

Like most media streaming services, Spotify grows by increasing the number of users engaging with the platform. For customer segments driven by ad revenue, the amount of time each user spends on the platform also determines user value.

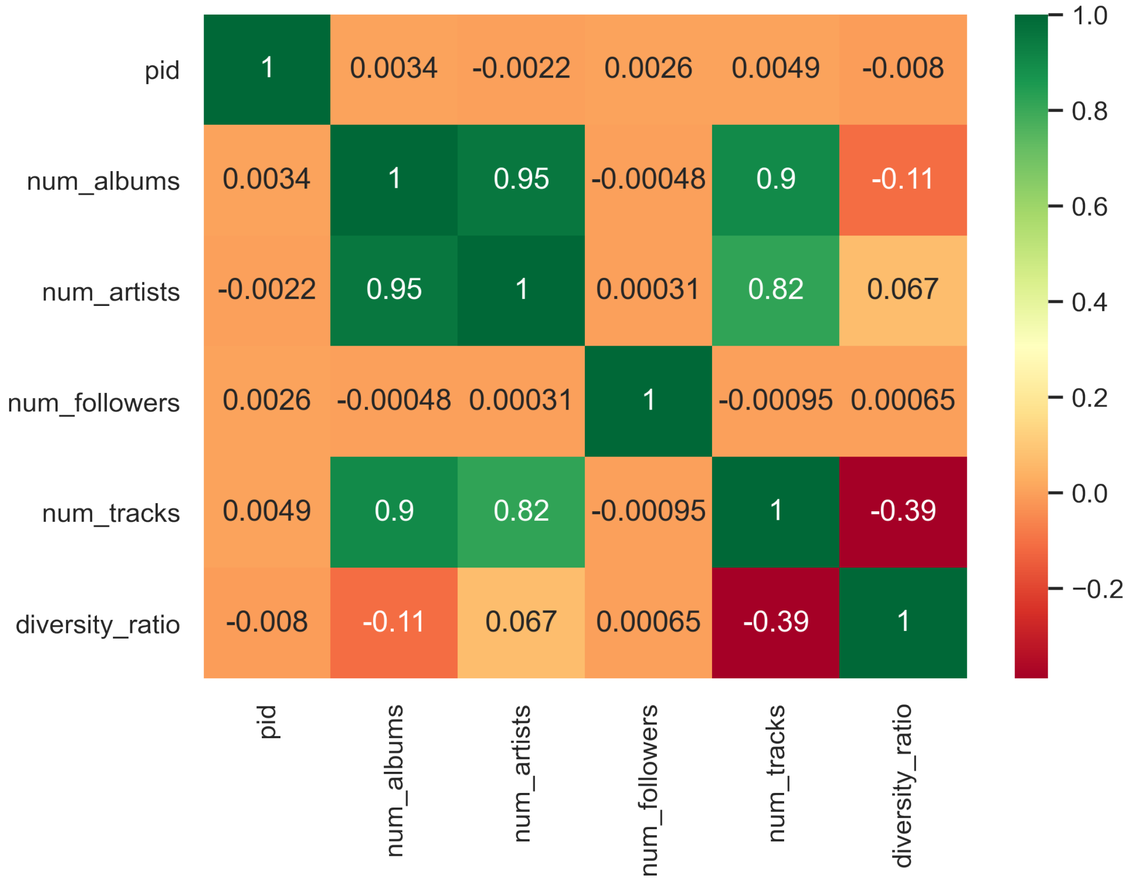
We’ve created a framework that can be used to evaluate pairs of features from available Spotify data, to see if one feature could influence the other.

As an example of how this framework can be used, we created the feature **diversity ratio**, which is simply the **ratio of number of artists to total number of tracks in a playlist**. We can now test this feature against existing features that signal levels of user engagement, such as playlist follower count.

Chris Stellato
Capstone Project #1

Chris Stellato
Capstone Project #1

Exploratory data analysis & developing a hypothesis



Exploring Correlations

A correlation matrix and corresponding heat map did not show many unexpected correlations between features. (Expected correlations, like number of tracks correlated to playlist duration, were confirmed).

How does diversity ratio affect follower count for a playlist?

Using our hypothesis testing framework, we can determine if **diversity ratio** has a statistically significant effect on a playlist's **number of followers.**

Stating a null hypothesis: "Diversity ratio has no effect on follower count."

There has to be a magic formula that makes some playlists wildly popular while the majority have less than 5 followers, right? We set out to investigate if the diversity of playlists has any statistically measurable effect on follower count.

ANALYZING DIVERSITY RATIO:
is it normally distributed?

After creating the diversity ratio feature using simple math (num_artists / num_tracks), we are able to explore this feature and see if the data is normally distributed. A normal distribution would allow certain hypothesis tests that could not be conducted on other distributions.

Histogram (01)

Visualizing diversity ratio with a histogram, the data does not appear to be normally distributed and is left-skewed.

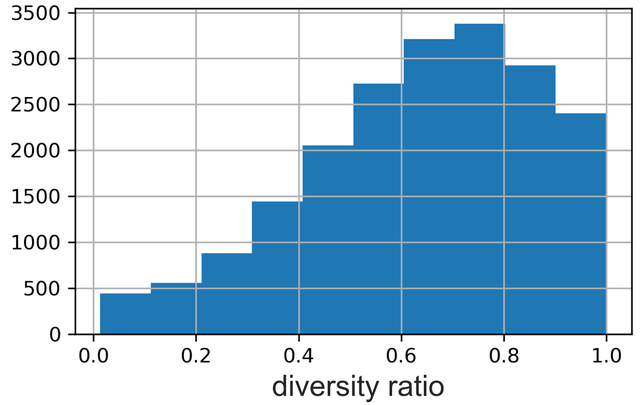
Box Plot (02)

Visualizing diversity ratio with a box plot yields the same result: the data does not appear normally distributed.

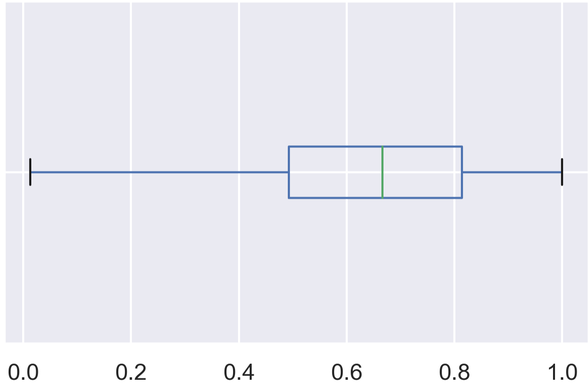
Shapiro Wilk test

Repeated Shapiro Wilk testing with random sampling consistently produced p-value near zero, further confirming the data is not normally distributed

01



02



Hypothesis Testing
Bootstrapping to simulate repeated sampling

Creating two sampling groups

During EDA we determined that the majority of playlists have fewer than 5 followers. We divided our original sample into two groups:

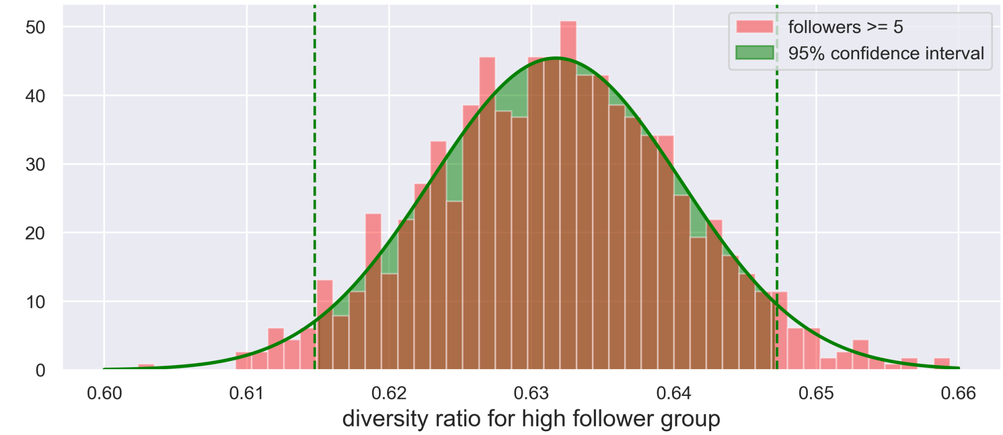
- Playlists with five or more followers
- Playlists with fewer than five followers

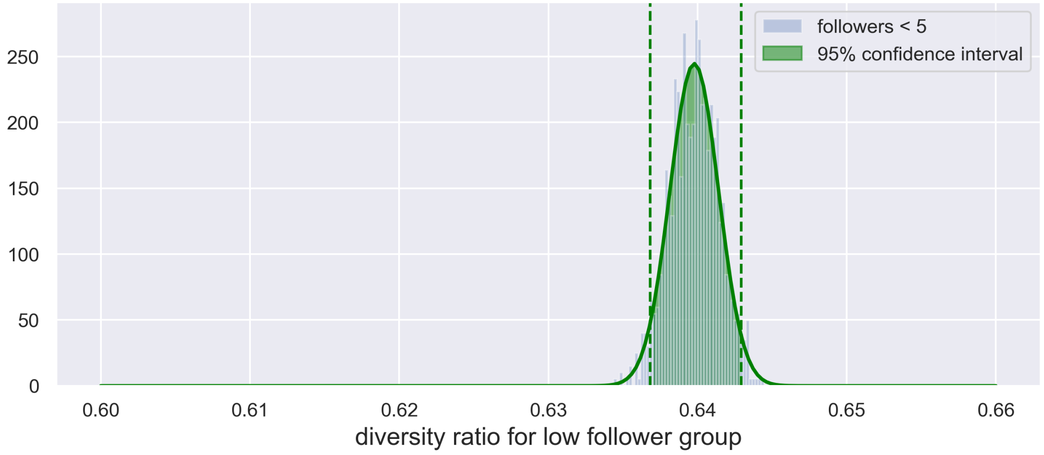
Bootstrapping to get ranges for our estimates of the diversity ratio

Although diversity ratio is not normally distributed in our samples, we are able to simulate repeated sampling and plot the mean of the samples, resulting in two normal distributions.

Early insights

While the two sample groups had slightly different average diversity ratios, looking at the x axis it appeared that the two sample distributions would have significant overlap.





Low-follower playlists set is much larger than high-follower set, resulting in a tighter distribution of sample means.

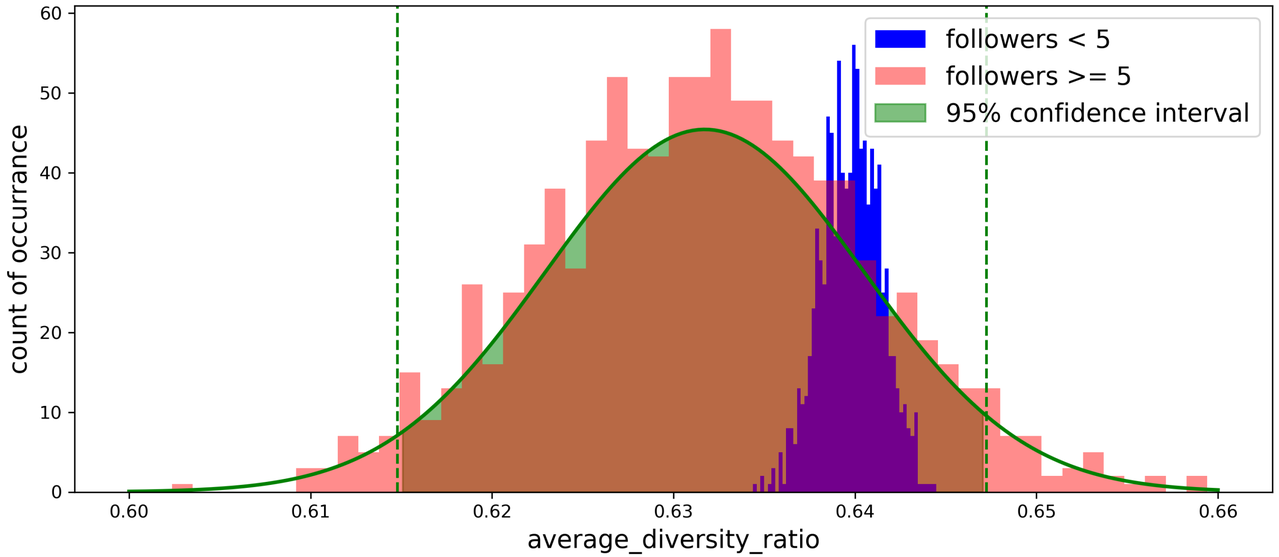
Chris Stellato
Capstone Project #1

Hypothesis Testing

Null Hypothesis: Diversity ratio has no significant effect on follower count. Significance threshold: 5%

After plotting both the low follower count and high follower count distributions on the same axis, we can see there is significant overlap within the 95% confidence interval of both distributions. The p-value indicates that in 37% of samples, a highly-followed playlist would have the same diversity ratio as a playlist with less than five followers.

As another check on our hypothesis, we ran a t-test with our high follower and low follower sample sets. The t-test returned a p-value of .367, indicating it is possible to see the same mean diversity ratio whether you are sampling the low follower set or the high follower set.




```
In [262]: stats.ttest_ind(less_than.diversity_ratio, greater_equal.diversity_ratio)
Out[262]: Ttest_indResult(statistic=0.9027733181213803, pvalue=0.366657088331551)
```

How could this help inform Spotify's strategy for resource allocation and future projects?

Follower count of a playlist is one good indication of how many users enjoy the playlist and will spend time listening to it.

We can say, with 95% confidence, that there would not be a significant return on investment for any project that tries to increase playlists' follower counts by focusing on playlist diversity ratio. Time and money is best spent exploring other methods that could help create more engaging playlists.

Now, using our established framework, we can quickly test feature significance and see which projects potentially deliver the biggest return on investment.

Opportunities for further exploration within current dataset and framework

Explore correlation between playlist length and follower count

Using either total playlist duration, number of tracks, or even average length of tracks, could attempt to identify these metrics influence follower count.

Identify songs or albums that appear on highly followed playlists

What patterns emerge when we start analyzing songs and albums appearing on highly followed playlists? Can we create a model that delivers a supergroup playlist, that real humans actually enjoy?

What other features can explore diversity in music choices?

Diversity of artists is only one metric. We could take this same idea in many different directions: Is a popular playlist diverse in time, with tracks coming from various decades? Does a playlist choose artists from a common geographic region, or do users prefer to stick to west-coast hip-hop and forego east-coast artists?

Thank You

CONTACT
Chris Stellato
Boulder, CO

stellatocjs@gmail.com
https://github.com/chris-stellato
chris-stellato.com