Capstone 1

Brett Hoffman

Overview

In this case study for this project, I found a dataset on Viberate of the top 500 artists pertaining to Spotify's rankings system. Along with the rankings, the data includes the rankings of other social media platforms, and other listening platforms, offering many relationships to explore.

This topic of study caught my attention because it addresses key points pertaining to the trends and patterns of the public body and how you would actually be able to statistically connect popularity with appeal. That being said, we will be looking into the data to find correlations on all angles with a main focus on genres of the artist to explore the role it plays on social media. In order to do so, other metrics such as popularity, what creates different parameters of popularity, the difference between platform statistics, and the difference between artist rankings will be discussed to see if anything has an effect on fan engagement with the artist.

Method

First we have to ascern the questions we want to ask. At first look at the data and its contents, a question was raised. How much does the type of artist genre actually play in their career? From this we are able to create a hypothesis:

Hypothesis 1

NULL

The difference in each social media following statistics or listening platform statistics has no direct correlation to the general genre of the artist, showing that each online music store has an association with a specific group of musical genres, differing from other online music stores.

ALTERNATIVE

Social media following statistics have a direct correlation to the character of genre of an artist.

Data and Model

Using R-Studio, the dataset was imported and ran through several methods of tidying and organizing to create simplicity in deducting answers. Further inspection showed some unexpected issues, as the data in the csv was not fully conclusive to begin with. There were several statistical columns in the time frame of 1 month that did not have any viable values, creating a large sum of Null values in the way.

These columns included Spotify’s monthly listeners, Youtube followers, Youtube views, even Tiktok likes. The entire columns had to be dropped due to Null values spreading across the board, and we ended up dropping almost 40 percent of the original columns.

Thankfully! this did not have any effect on the data or its usablility, this only affected the viewing and cleaning process since the parameters where the Nulls exists were all the same. It seems as if the algorithm Viberate used to pull this data from it’s sources either had a few errors, or the sources had enough incomplete data to not register.

Once the data was clean, more questions came to the surface and 2 more questions were introduced:

Question 1:

Does the following an artist has on one social media have any effect on their respective following counts on other platforms. Do trends exist or is the platform playing a role

Question 2:

Does the ranking spotify's algorithm assign reveal statistical difference between how many monthly followers an artist has or how many monthly plays they get

The first question that I want to look into, going back to my original hypothesis is all about the genre. Popularity patterns can happen anywhere!

maybe certain locations encourage the local culture's favorites?

We will see whether the hypothesis will be rejected or failed to reject

In order to analyze the data we utilized several different ggplots in r to show the correlation between several different point of data and the correlation to the statistical patterns of the genre data

- Genre vs Total Spotify Monthly Listeners
- Genre vs Total Spotify Followers
- Genre vs Total Spotify Playlist Reach

- Genre vs Total Twitter Followers
- Genre vs Total Instagram Followers

The results from these charts was not as expected. They show a very similar spread pattern across, a small outlier being Playlist reach with a slight skew in the general distribution of points. The consistent data shows the relationships are not distinct from one another

thus we can assume that it is more just patterns of each platform that play a bigger role in the following count, rather than the genre of the artist.

This means we would fail to reject the null hypothesis

Hypothesis 2:

H0: The following an artist has on one social media does not have any effect on their following count comparatively

H1: The following that each artist has does indeed have a direct effect on their following count comparatively

We will find this out by performing correlation tests between their followings, and comparing the co-effecient’s and p-value’s to see if they are consistent across or not

This will perform a test to see whether the hypothesis will be rejected or failed to reject

RESULTS

The results show that it would appear that there is a significant medium positive relationship between twitter and spotify followings.

the p-value between twitter and spotify was 0.00081

The lower the p-value, the greater the statistical significance of the observed difference. Since we are below 0.05, this shows that there is a strong statistical significance between the two variables.

the correlation coefficient was .3313

A correlation coefficient of 0 means there is no linear relationship between the two data sets. The values of -1 (for a negative correlation) and 1 (for a positive one) describe perfect fits in which all data points align in a straight line, indicating that the variables are perfectly correlated.

Our coefficient was above .3 showing there is a medium significance level between twitter and spotify followings.

On this next test, there may have been bias introduced. Facebook and Instagram are under the same ownership, this was originally not taken into consideration.

The companies have the same promotional algorithms and trend patterns so this data most liekly will not show the same relationship.

the p-value between facebook and instagram was 1.225e-11

the correlation coefficient was .6153

These two values show a strong significance level as expected between two platforms that work in tandum.

Deezer fan numbers and TikTok following were then thrown into the equation to branch away from the main body and see if we introduce any difference in results.

The correlation coefficient between deezer and tiktok was a negative value of -0.231 and the p-value was .8205

These points signify that there is a non-significant, very small negative relationship between the following statistics of the two platforms.

Big difference between these three tests! These conclusions show that we can fail to reject the original hypothesis, meaning there does not seem to be any real significant correlation between the following statistics, and the following an artist has on one platform does not necessarily reflect on others.

Another hypothesis:

H0: The statistical difference between how many monthly followers an artist has and how many monthly plays they get does not play a role in Spotify's ranking result.

H1: The statistical difference between how many monthly followers an artist has and how many monthly plays they get had a direct effect on Spotify's ranking result.

This will be fun since the ranking is ordered in a linear fashion

We will perform a test to see whether the hypothesis will be rejected or failed to reject

Let's see if there is a correlation between the different spotify statistics

MORE RESULTS

Between Spotify rank, and Spotify following:

The correlation coefficient was -0.7475

the p-value <.001, at 9.446e-19

These show that there is a significant very small negative relationship between the rank and follower count

Lastly, we checked monthly listeners

The correlation coefficient was -0.766 and the p-value was 1.619e-20, or <.001. even smaller than the last!

The results all support the alternative hypothesis that how many follower an artist has and how many monthly listeners they have has a direct relationship with the platforms ranking system.