Title: Statistical Analysis of Spotify Artist Data

Azim Chyngozhoev 20212003

Introduction

This data analysis is focused on artist data from Spotify's API – spotipy, which is openly accessible.

Spotify is a streaming platform popular in the west, similar to Melon music or Naver music. The analysis focuses on acquiring the data from spotify about the artists, their features (like danceability, tempo, key, valence), which is all set by spotify itself, and the corresponding genres of these artists.

Methods

Spotify API Data acquisition

In the web api of spotify, I created a new "app", which is just a session linked to my account, that accesses all information of spotify.

*Client_ID and client_secret are important and should not be displayed openly to other people. However, by the time this report is submitted, I terminated the session.

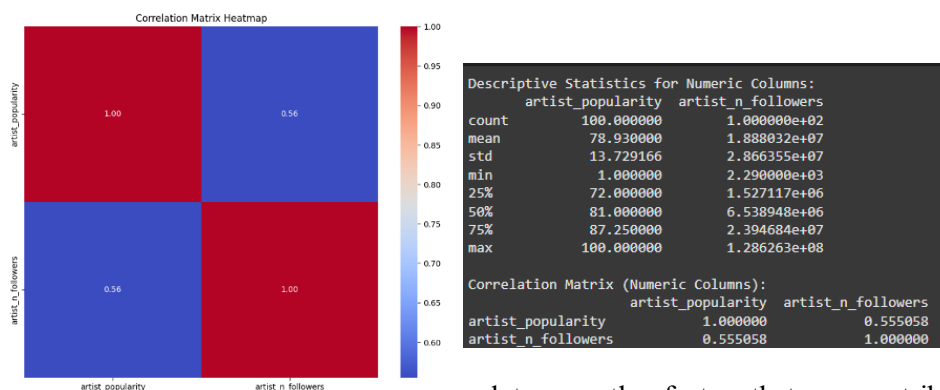Artists features like name, popularity, genre and followers were acquired.

I did not choose their songs genres because they may differ. And each artists has multiple genres referred to them, so I am only using the first genre that comes up. ('popularity', 0)

The analysis focused on:

1. **Descriptive Statistics**: Detailed look into artists genres, popularity and etc.

We have a total of 100 artists, and it took only those that are above the threshold of 50 popularity (not number of followers).

2. **Correlation Analysis**: See how number of followers correlate with artsits popularity


```
Descriptive Statistics for Numeric Columns:
       artist_popularity  artist_n_followers
count         100.000000        1.000000e+02
mean           78.930000        1.888032e+07
std            13.729166        2.866355e+07
min             1.000000        2.290000e+03
25%            72.000000        1.527117e+06
50%            81.000000        6.538948e+06
75%            87.250000        2.394684e+07
max           100.000000        1.286263e+08

Correlation Matrix (Numeric Columns):
                   artist_popularity  artist_n_followers
artist_popularity           1.000000            0.555058
artist_n_followers          0.555058            1.000000
```

As we can see, moderate correlation is observed (r=0.56), that's between artists popularity and artists number of followers. We can say that follower count does influence popularity, but we have to keep in mind that there maybe a lot more other factors that may contribute to artist popularity.

Regression Analysis: Followers vs. Popularity
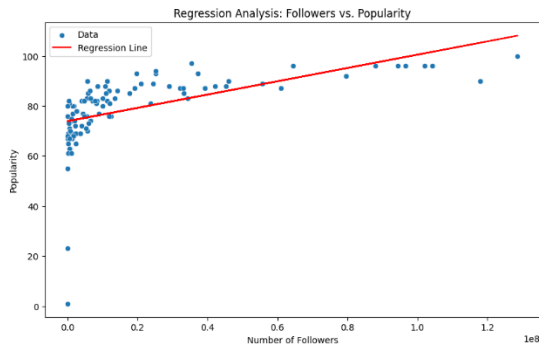
```
Intercept: 73.91
R-squared: 0.31

Outliers Detected:
    artist_name          artist_id  artist_popularity artist_first_genre  \
35   Karol G  4asJQrazucWu3Zy5R7B164                 1                 []
49    J.I.D  7bTn84XuezOf5kzCED8UCK                 23                 []

    artist_n_followers
35                3938
49                2290

Average Number of Followers per Popularity Segment:
popularity_segment
Low              3114.00
Moderate        28820.00
High         11752561.25
Top          69290876.00
```

Several dots below the line represent artists whose popularity is a lot lower than expected given their followers. For example, the artist near 0.2e8 followers but a popularity score of ~20. Which is an outlier. As well as, dots above the line represent artists whose popularity is higher than expected for their follower count, possibly due to recent virality or unique factors.

The upward trend of the regression line confirms a positive relationship between followers and popularity.

3. Genre analysis
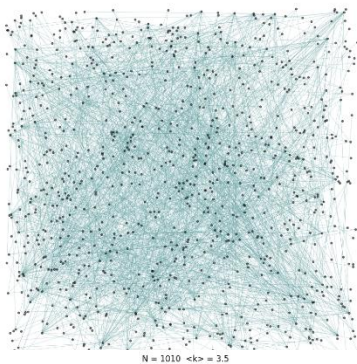
I used Networkx package to visualize and analyse the network of artists. Running through artists above the threshold of 50 popularity, and coupling them, if they had worked together using Features on a song, I created a network of artists; 1st plot is random plot, 2nd plot is kawai_layout plot.

Then I divided them by genres, took only 1st genre of an artists and assigned a color for each. 3rd plot is genre specific plot.

Discussion

This project showed that artists popularity is positively correlated, which is kind of obvious. But, still we can see that there may be many factors and we have to remember about hyped one time hits or artists that are underrated. Other factors may include recent album releases, collaborations, or viral hits. Outliers show some unique cases that could be investigated, they may be the ones with one time viral hits or underestimated songs.

We also can se how artists collaborate with each other. I couldn't include everything from my code to the report (because of 2 page limit). But I found that Raeggaton genre is pretty isolated, not many artists work together outside of this genre. I also found that Rock, pop, hiphop work a lot together, showing a bigger connected network (see cell with "GENRE CONNECTED ONLY NETWORKS EACH PLOT" comment). There are also very strange genres we never heard of like "adult standards" and "redneck" etc.

I couldn't add everything, please refer to the code much more there:
https://colab.research.google.com/drive/1VXUT8ryQMdPUjcz-K4m7QOo6-MatmYCa?usp=sharing