Spotify Data Analysis and Machine Learning Regression Study

Haya Darwish

Josh Belisario

Conor Cleary

Filip Jurek

Evan Dillon

Technological University

Nature of enterprise

Lecturer: Keith Nolan

Introduction

This report will go over a structure analysis of a dataset with the use of statistics and visual methods, the purpose of the analysis is to explore all the different patterns you can see within the dataset such as trends and relationships,

The following sections will go through the Dataset overview, Data cleaning summary, Descriptive statistics and Visual Analysis supported by graphs, basic machine learning applications and discussing the limitations and ethical considerations which will provide a structured evaluation of the dataset used.

Dataset Overview

The Dataset consists of 3 main variables which are numeric variables:

- track_popularity
- artist_followers
- album_total

I found these were the best variables when cleaning the data set. It removed a lot of rows making the graphs in the future smaller and easier to read.

Data Cleaning Summary

When Cleaning the Dataset I was already familiar with the code I could use because my leaving cert also had a similar task. I used the base code from my leaving cert project to remove all singles, track popularity under 50 and removed any artist followers under 500k to make the dataset smaller.

```python
import pandas as pd
data = pd.read_csv('cleaning_data.csv', usecols=['track_name', 'track_popularity', 'artist_name', 'artist_followers', 'album_name', 'album_type','album_total_tracks'])

column_values = data[['track_name', 'track_popularity', 'artist_name', 'artist_followers', 'album_name', 'album_type', 'album_total_tracks']].values.tolist()
print(column_values)

data = data[data['track_popularity'] >= 50]


data = data[data['artist_followers'] >= 500000]

data = data[data['album_type'] == 'album']



data.to_csv('filtered_data.csv', index=False)
print("Cleaned data saved as 'filtered_data.csv' ")

print(data.describe())
data.info()
```

This is the code I used to clean the dataset. I sampled it from a dataset I did for the leaving Certificate but based it off the Spotify Dataset instead of the last Dataset I had done.

I used pandas to clean the dataset. Below I have shown the work from my last dataset I cleaned to show the difference between the two of them.

```python
import pandas as pd


data = pd.read_csv('fifa21_raw_data.csv', usecols=['Age', 'Nationality', 'Positions', 'Name', 'Wage'])

data['Wage'] = (
    data['Wage']
    .astype(str)
    .str.replace(r'[^\dK]', '', regex=True)
    .str.replace('K', '000')
    .astype(float)
)

data = data[data['Wage'] > 80000]

print("Missing values before filling:")
print(data.isnull().sum())

data.fillna(0, inplace=True)

data.to_csv('filtered_fifa21_raw_data.csv', index=False)
print("Cleaned data saved as 'filtered_fifa21_raw_data.csv'")

column_values = data[['Age', 'Nationality', 'Positions', 'Name', 'Wage']].values.tolist()
print(column_values)


print(data.describe())
print(data.info())
```

This is the code I based my data cleaning off. It is an old data set I cleaned.

Descriptive Statistics

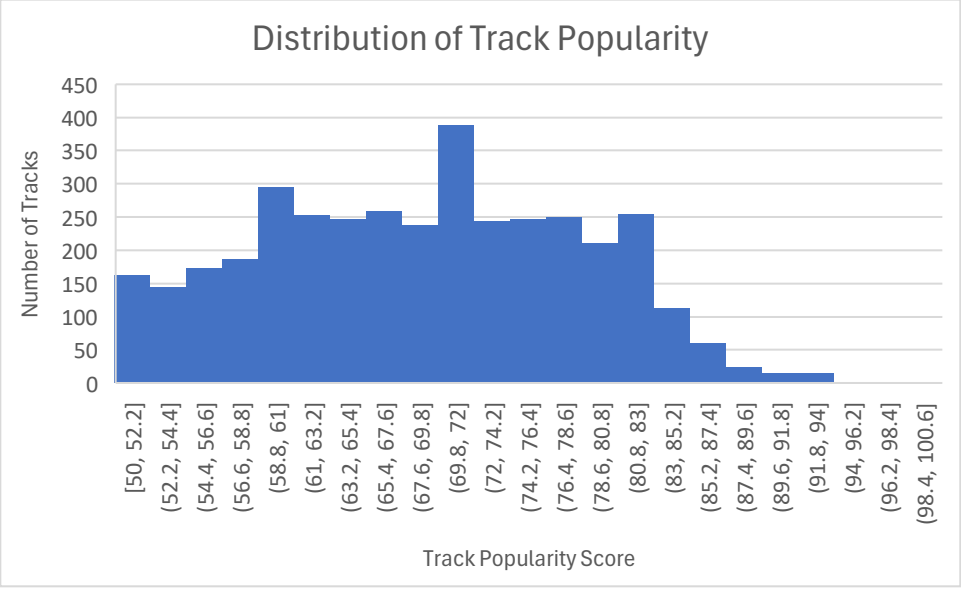
Variable	Mean	Median	Mode	Min	Max	Standard Deviation
Track_popularity	68.82925546	69	71	50	99	9.765489542
Artist_followers	38377632.49	18208465	145396321	502300	145542136	43655014.98
Album_total	15.8005788	15	12	4	137	43655014.98

Note:

- track_popularity is consistent as most tracks are grouped around the 60-75 range
- artist_followers is highly skewed due to the dataset having very few popular artists.
- album_total is generally focused on around to 12 – 16 tracks per artists but some outliers exist.

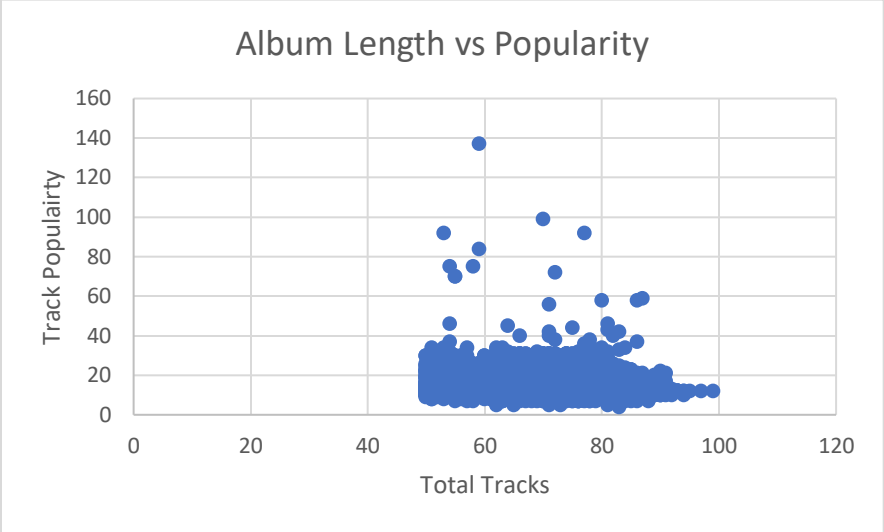
Visual Analysis with graphs

Distribution of Track Popularity



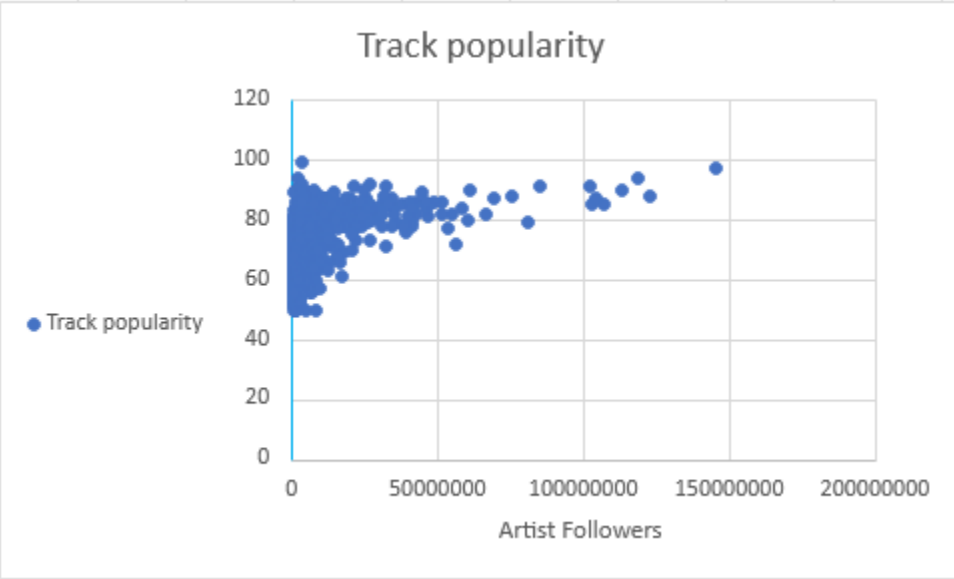
- Grouped around the 60-75 popularity which correlates to the mean of 68.8
- Skewed to the righ8t due to only a few tracks being popular
- Confirms the descriptive statistics for track_popularity

Album length vs Track popularity



- Most points are grouped at lower scores showing that most tracks aren't guaranteed going popular
- Few points are outliers due to having very high popularity despite the average album length

Track Popularity



- Few artists are very popular
- Most artists have less than 50 million followers and are still very popular

Ethical Considerations

- No personal data was used/included.
- All data relates to publicly available music information.
- AI assistance was used for machine learning implementation and fully documented.

General Limitations

Despite our dataset containing a large amount of data, it still only represents a small subsection of Spotifys entire library which is something to consider when analyzing the information. Our dataset also does not take into account any possible external factors which could influence artist and song popularity. For example, external marketing or language, an artist signing in a less globally known language is likely to be less popular than someone speaking Spanish or English. It's also worth considering that popularity score is something only used and specific to Spotify and does not generalize a song, album or artists popularity. Due to the nature of our dataset being publicly available information and only containing simple data, there isn't much to consider in the way of ethics.

Machine Learning Summary

Applied linear regression model to predict track popularity based on two numerical features: artist followers and total number of tracks on the album. To start working on the machine learning we had to clean the data first, remove unnecessary columns and handle missing values. Then, the filtered data was split into training and testing sets using an 80/20 split.

The model was trained using the training set and evaluated using the test set using two performance metrics: R^2 score and Root Mean Squared Error (RMSE). R^2 measures how well the model explains the variation in track popularity, while RMSE indicates the average prediction error.

The results show that the model achieves R^s score of ≈0.0156, and RMSE of ≈ 9.75 popularity points. Those results indicate that the model average as good regarding the RMSE, while the R^2 score is very low but from research this happens often in real-world ML and the some of the common reasons is that

the target y is influenced by many hidden factors, the data is noisy, target is complex, or features are limited.

Model Assessment

The low R2 score that the regression model generated is a good informational result. It draws attention to the fact that a variety of different elements outside of the study's numerical variables probably have an impact on track popularity. Although they weren't included in the dataset, factors like marketing campaigns, playlist placements, release dates, partnerships, social media presence, and listener demographics may have a big impact on performance. Because linear regression implies a rather simple relationship between variables, it might not accurately capture the dynamic character of patterns in music consumption. Future developments might explore other machine learning methods such ensemble methods, decision trees, and multiple regression with more features.

Conclusion

In conclusion: We successfully visualized and cleaned our selected dataset, as well as applied a linear regression model. The analysis demonstrates how using statistics, graphs and charts, and basic machine learning can help in getting meaningful insights from real-world datasets.

We discovered that the low R^2 score shows that there’s many factors that affect a tracks popularity that we simply don’t have in our dataset, such as marketing, promotion, release time and genre etc. Our analysis shows that the number of followers an artist has does affect a track’s popularity and success but not as much as one might initially believe. In the end, this project has shown the uses for machine learning as well as displaying some of its possible limitations.

Sources:

https://chatgpt.com/share/699d7274-db64-8010-b1de-4d06bdacd739