Analysis of top songs of 2024

Group D13: Herman Palm, Martin Laks

Task 1. Setting up (0.25 points)

https://github.com/martin737373/DataScienceProject

Task 2. Business understanding (0.5 point)

a. Identifying your business goals

Background

Nowadays streaming platforms are the most common way of consuming and enjoying music. Among these Spotify is globally the largest such platform and provides a free public API with which data can be collected of the tracks uploaded there. Understanding what stood out and finding predictive patterns in the most streamed songs of previous years could be of value to the stakeholders of Spotify as well as music enthusiasts. Within our project we will attempt to do just that.

Business goals

1. Provide insight to the key characteristics of the most streamed Spotify songs in 2024 which could be of value to the stakeholders or of interest to music enthusiasts.

2. Provide the means to predict features of the tracks which are otherwise of unknown origin or of use to determine a track's success.

3. Evaluate if neural network models would be of use and value in this setting.

Business success criteria

1. Clear visual representation of the data is given which is meaningful to the stakeholders goals or to the interest of music enthusiasts.

2. Features of importance from the dataset can be predicted with built models.

3. The viability of neural networks is tested and assessed.

b. Assessing your situation

Inventory of resources

- Kaggle has many datasets on streamed song from Spotify: https://www.kaggle.com/datasets

- Spotify's public API with which additional data could be gathered if it is required for the scope of this project: https://developer.spotify.com/

- Python and its libraries such as pandas, scikit-learn, matplotlib, TensorFlow/Keras.

Requirements, assumptions, and constraints

- Requirements - Finish and submit the project by 12:00, 8th of December 2025.

- Assumptions - The dataset chosen from Kaggle is accurate and Spotify's tracks streaming data is representative of what it is globally.

- Constraints - The data of only the most streamed songs of 2024 on Spotify can be used as gathering and analysing the data of all streamed songs of 2024 would fall out of the scope of this project as the costs would be too high.

Risks and contingencies

There is the risk of insufficient data in the Kaggle dataset to build the stated models.
This can be mitigated by gathering additional data using Spotify's public API.

Terminology

Track - an audio recording of a song
Streams - the number of times a track was played
Popularity - the popularity score of a track (on Spotify calculated by an unknown formula)
Rank - the ranking of a track based on its all time popularity
Explicit track - a track which contains explicit content

Views - the total number of views of a track on a site
Likes - the total number of likes for a track on a site
Posts - the total number of times a track was posted on a site
Spins - the total number of times the track was played on a platform
Shazam counts - the total number of times user used Shazam to identify the track
ISRC - International Standard Recording Code of the track

Costs and benefits

Costs - The time required to gather, clean and analyse the data and build the models. The computational costs should be minimal.

Benefits - Insight into the key characteristics of last year's top tracks and also how Spotify's popularity or score is evaluated which can be used to understand what made songs popular in the year 2024 on Spotify and generally.

c. Defining your data-mining goals

Data-mining goals

1. Analyze the data to understand it, its patterns and its correlations and also to find characteristics which stood out.

2. Build predictive classification and regressions models in order to predict the track explicitly boolean value and the track popularity/score respectively.

3. Develop predictive neural network models for the same features in order to possibly get better results.

Data-mining success criteria

1. Graphs are shown of relevant patterns and correlations of key features.

2. Classification models achieve an accuracy of at least 85%.
   Regression models achieve a reasonable RMSE score.

3. Neural network models built which perform as well or even better than classical models. Or it is explained why they underperform in this instance.

Task 3. Data understanding (1 points)

a. Gathering data

Outline data requirements

In order to analyze the top songs of 2024 the following is required:

- Data of the most streamed songs from a significant streaming service.
- Attributes of the songs data should be indicative of its popularity, such as:
  track name, artist, streams/views/likes/posts/spins/counts, explicitly, popularity, score, and so on.
- Continuous numerical fields are required for regression modeling.
- Categorical fields are necessary for classification modeling.
- Sufficient amount of data for meaningful statistical analysis.

Verify data availability

https://www.kaggle.com/datasets/nelgiriyewithana/most-streamed-spotify-songs-2024
From this dataset 1.1MB of data of the most streamed Spotify songs of 2024 can be acquired. It provides relevant data for the 4370 most popular tracks with both continuous numerical and categorical fields.

https://developer.spotify.com/
Spotify's public API was used to form the aforementioned dataset and can be used to gather additional data on a small scale should the need for it arise. For example Spotify does not attribute genres to songs but does do it for artists, which every song has. Therefore the genres of the songs artists can be acquired to supplement the Kaggle dataset using this API.

Define Selection Criteria

The whole Kaggle dataset will be included and used in this project as all the fields meet the data requirements and the project is exploratory in nature. Meaning it is the goal of this project to find out which of these fields are actually meaningful and useful. The number of track data should also be sufficient and there is no reason to reduce that number.
Datasets which we considered but excluded were ones that did not explain how the tracks were chosen or seemed to have fake data by viewing correlation matrixes.

b. Describing data

In figure 1, all the fields, their respective data types and descriptions of the dataset can be seen. The dataset consists of different attributes of the 4600 streamed songs on Spotify. In total there are 29 features of which most are discrete numerical counts of number of times the track was played, shared, liked and so on. The “TIDAL Popularity” is the odd one out having no data in it whatsoever and therefore it will be excluded. There are a few string values containing names of the track, artist, album and ISRC. With one boolean value notating whether or not the track has explicit content and one float value which is a score from 0 to 100 given to a track by Spotify. Similarly to Spotify's popularity score it is unknown how this value is calculated.

Most of these values do indicate the popularity of the track with probable exceptions being the date it was released and its ISRC. It could be argued that the different metrics showing the songs popularity count from different platforms will provide no different insights. Further there are features with a lot of null values, but as stated before the similar nature of a lot of these fields could compensate for one another. But overall for every track there is a lot of information which will prove useful in the analysis and further steps.

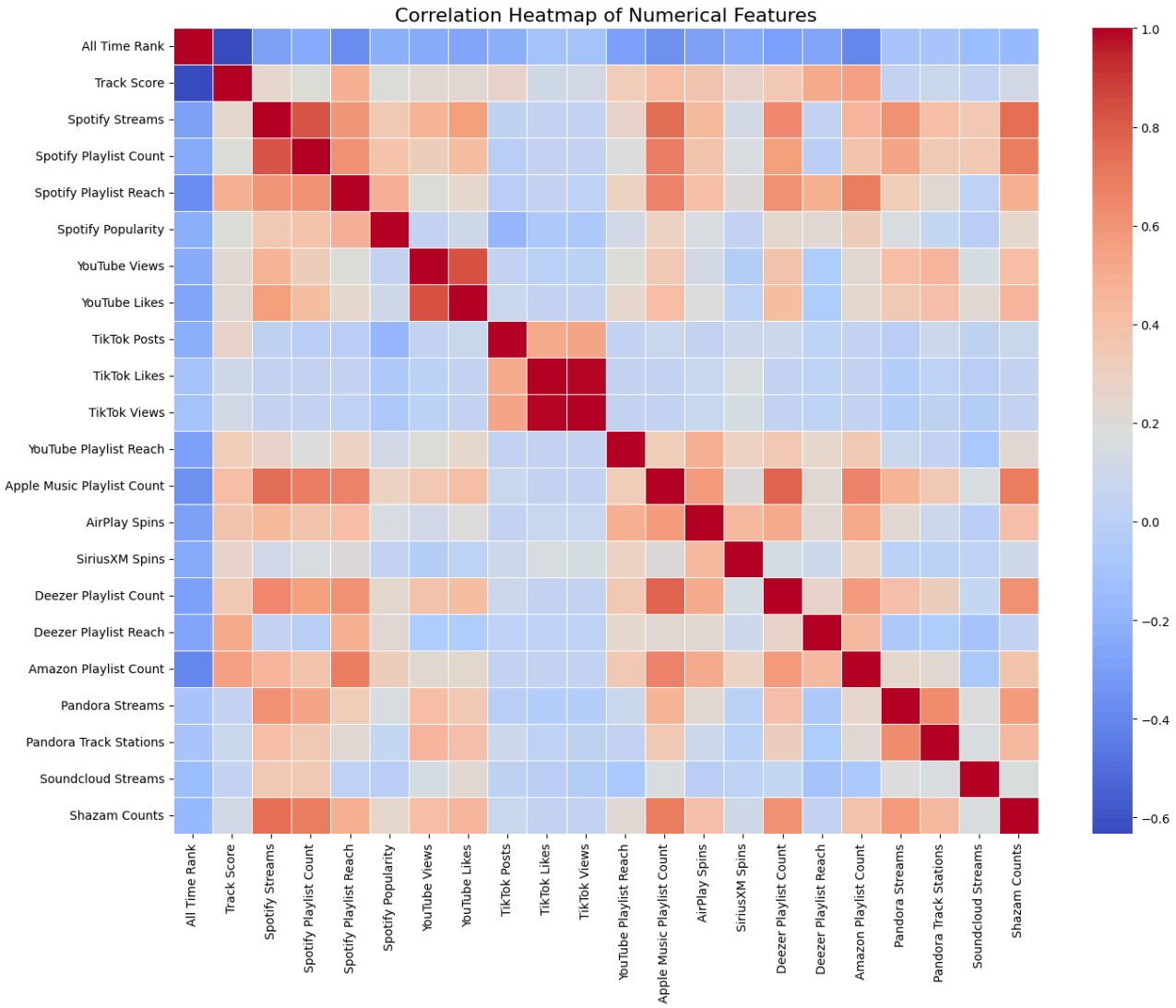
Finally it seems the data is formatted quite nicely. Based on the task at hand standardizations or normalizations will need to be done and maybe even the combining of some values to mitigate null values.

Column Name	Data Type	Description	Nr. of null values
Track	string	Name of the song	0
Album Name	string	Name of the album the song belongs to	0
Artist	string	Name of the artist(s) of the song	5
Release Date	datetime	Date when the song was released	0
ISRC	string	International Standard Recording Code for the song	0
All Time Rank	int	Ranking of the song based on its all-time popularity	0
Track Score	float	Score assigned to the track based on various factors	0
Spotify Streams	int	Total number of streams on Spotify	113
Spotify Playlist Count	int	Number of Spotify playlists the song is included in	70
Spotify Playlist Reach	int	Reach of the song across Spotify playlists	72
Spotify Popularity	int	Popularity score of the song on Spotify	804
YouTube Views	int	Total views of the song's official video on YouTube	308
YouTube Likes	int	Total likes on the song's official video on YouTube	315
TikTok Posts	int	Number of TikTok posts featuring the song	1173
TikTok Likes	int	Total likes on TikTok posts featuring the song	980
TikTok Views	int	Total views on TikTok posts featuring the song	981
YouTube Playlist Reach	int	Reach of the song across YouTube playlists	1009
Apple Music Playlist Count	int	Number of Apple Music playlists the song is included in	561
AirPlay Spins	int	Number of times the song has been played on radio stations	498
SiriusXM Spins	int	Number of times the song has been played on SiriusXM	2123
Deezer Playlist Count	int	Number of Deezer playlists the song is included in	921
Deezer Playlist Reach	int	Reach of the song across Deezer playlists	928
Amazon Playlist Count	int	Number of Amazon Music playlists the song is included in	1055
Pandora Streams	int	Total number of streams on Pandora	1106
Pandora Track Stations	int	Number of Pandora stations featuring the song	1268
Soundcloud Streams	int	Total number of streams on Soundcloud	3333
Shazam Counts	int	Total number of times the song has been Shazamed	577
TIDAL Popularity	null	Popularity score of the song on TIDAL	4600
Explicit Track	boolean	Indicates whether the song contains explicit content	0

(Figure 1: dataset fields)

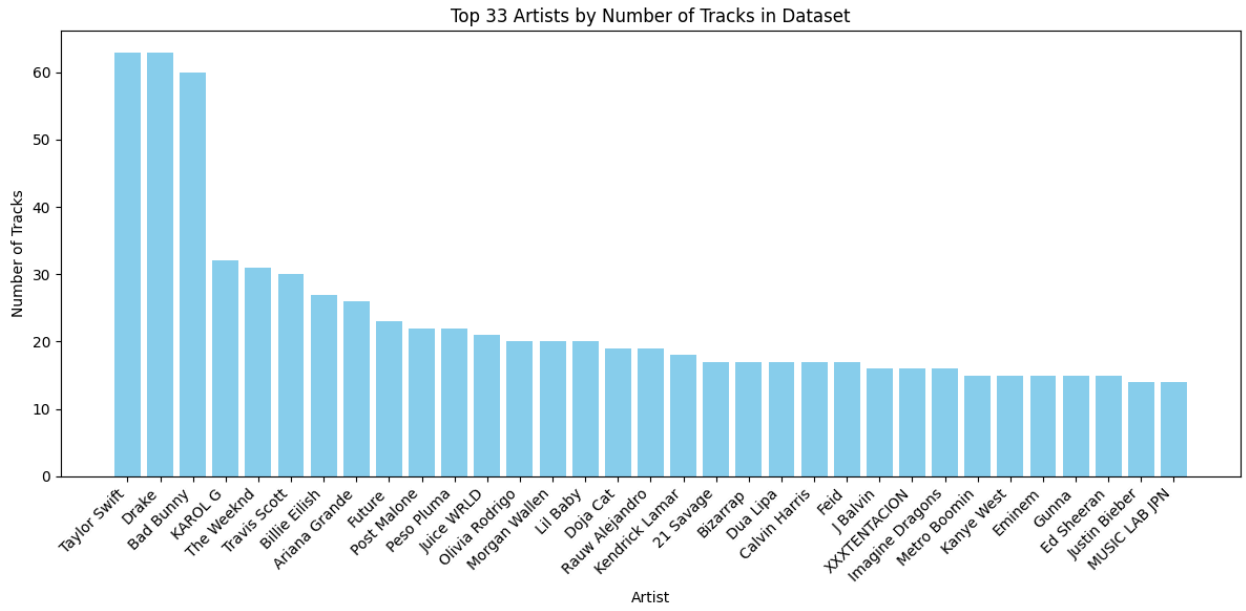
c. Exploring data

Looking at the correlation matrix heatmap in figure 2, it can be seen that there are a lot of clusters with high and low correlation values. Most measures from the same platform seem to have high correlations with each other but also measures from TikTok have weak correlations with all measures from other platforms. Similarly there are no strong correlations across measures from different platforms with the highest among them being approximately 0.6. Overall it shows that across platforms there is no good single measure which could predict another. Therefore combining the information from multiple features will be required in order to get predictive results.



(Figure 2: correlation heatmap)

Another interesting graph can be shown in figure 3 which shows a bar plot of the top 33 artists by the number of their tracks in the dataset. There are 1999 unique artists in the dataset and the number of tracks they have in the dataset seems to follow an inverse distribution. This goes to show that there are additional attributes which could be calculated from the existing data which do follow predictive patterns.



(Figure 3: bar plot of artists vs nr. of tracks)

d. Verifying data quality

As shown before in figure 1 there are a lot of null values in the data which could be mitigated by combining different attributes and other data processing methods. Additionally whilst the correlation matrix in figure 2 does not show any strong correlation between attributes which are measurements from different platforms, this suggests that further analysis and the use of statistical models are needed to predict different values which is exactly the goal of this project. This is further supported by the bar plot in figure 3 which shows there are correlations within this data but more attributes might need to be processed and the correlations might not be linear.

Task 4. Planning your project (0.25 points)

plan

Task	Description	Herman	Martin
Plan project and data-mining	Plan the project at hand and find and verify suitable data	3h	3h
Clean data	Clean and format the dataset	2h	2h
Process data	Process the data by combining attributes, calculating new ones and preprocessing for future tasks	4h	4h
Analyze data	Analyse the data and provide visualizations	4h	4h
Classifier model	Build a classifier model to predict the “Explicit Track” and/or “Artist” field and refine it	0h	6h
Regression model	Build a regression model to predict the “Spotify popularity” and/or “Spotify score” field and refine it	6h	0h
Classifier neural model	Same as task 3 but with neural networks	0h	8h
Regression neural model	Same as task 4 but with neural networks	8h	0h
Final report	Write out the results and conclusions	3h	3h
Poster	Make the project poster and presentation	3h	3h

Total hours per member: 33h

methods and tools

Data collection:

- Tools: Kaggle datasets, Spotify Public API, Python
- Methods: Find a suitable dataset from kaggle and gathering additional data using Spotify's API if needed.

Data cleaning and preprocessing:

- Tools: Python (pandas, numpy)
- Methods: Convert values and handle missing values. Standardize or normalize numerical values. Encode nominal values.

Data exploration and analysis:

- Tools: Python (matplotlib, seaborn)
- Methods: Calculate new attributes or combine existing ones. Plot different graphs between the attributes to visualize patterns and correlations.

Modeling:

- Tools: Python (scikit-learn, TensorFlow/Keras)
- Methods: build different classical and neural network models for classification and regression