Analysis of Artist Streaming Data Using Regression and Visualization

---

Objective

The major purpose of this analysis was to look at the correlation between several artist-specific metrics (solo streams, featured streams, and lead streams) and overall Spotify streams. Using the provided dataset, we attempted to study how different cooperation styles affect overall popularity and streaming indicators.

Code Overview

This software parses a dataset including Spotify streaming statistics for various artists, runs linear regression models to detect links between variables, and creates scatter plots with regression lines to display these associations.

Steps Performed

1. Data Parsing

The dataset includes columns such as:

- **Total Streams**: The total amount of streams an artist has received.
- **Solo Streams:** Streams from individual projects.
- **Featured Streams**: Streams that showcase the artist as a partner.
- **Lead Streams**: Streams in which the artist takes the major lead.

The program:

- Read the dataset.
- Handles invalid or missing values gracefully by skipping problematic records.
- Converts numeric values (e.g., "85,041.3") into proper floating-point numbers.
- Extracts relevant metrics for analysis.

Output:
Successfully parsed **3000 valid records**, ensuring the dataset is clean and ready for analysis.

2. Linear Regression

To examine the relationships between the metrics, the program calculates regression equations for three comparisons:

- **Total Streams vs Solo Streams**

- **Total Streams vs Featured Streams**
- **Total Streams vs Lead Streams**

For each comparison, the program calculates:

- **Slope**: Represents the rate of change (e.g., how an increase in solo streams impacts total streams).
- **Intercept**: Represents the baseline value when the independent variable is zero.

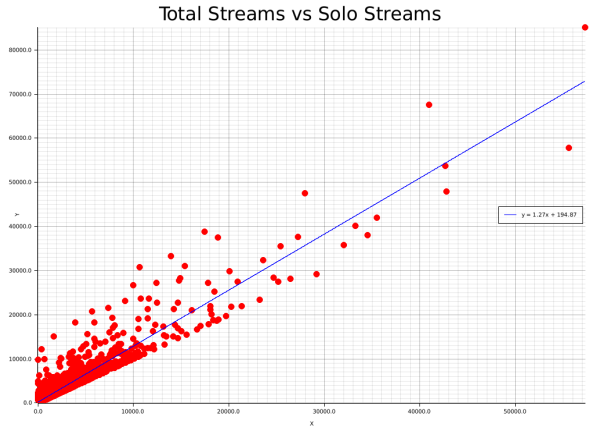
3. Scatter Plots and Regression Lines

The program creates scatter plots for each association and overlays a regression line to show trends. These visualisations give an intuitive grasp of how total streams correspond with each independent variable.

Findings

1. Total Streams vs Solo Streams

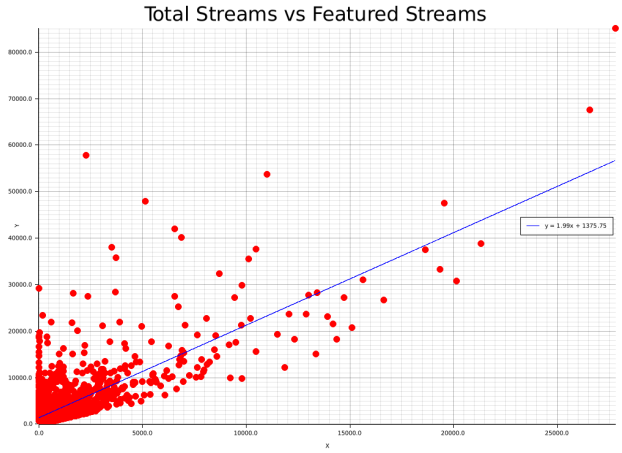
- **Regression Equation**:
  $y=1.27x+194.87y = 1.27x + 194.87y=1.27x+194.87$
- **Interpretation**:
  Solo streams are strongly correlated with total streams. The slope of 1.27 suggests that for every additional solo stream, total streams increase by 1.27 on average.
- **Visualization**:



2. Total Streams vs Featured Streams

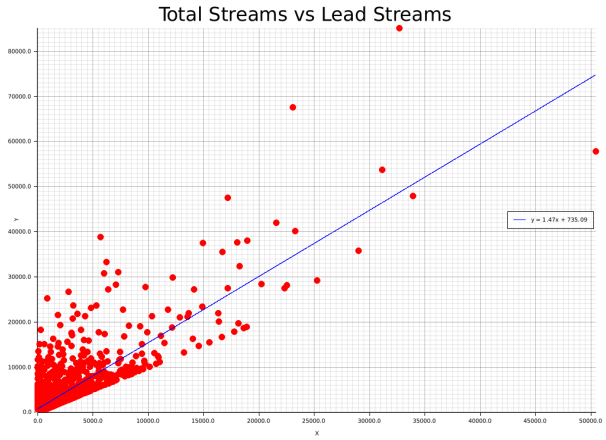
- **Regression Equation**:
  $y=1.99x+1375.75y = 1.99x + 1375.75y=1.99x+1375.75$

- **Interpretation**:
  Featured streams also contribute significantly to total streams. However, the higher intercept value suggests that even artists with fewer featured streams tend to have a baseline level of total streams.
- **Visualization**:



3. Total Streams vs Lead Streams

- **Regression Equation**:
  y=1.47x+735.09y = 1.47x + 735.09y=1.47x+735.09
- **Interpretation**:
  Lead streams have a substantial impact on total streams, with a slope of 1.47 indicating a strong correlation. This suggests that being the lead artist is a key driver of streaming success.
- **Visualization**:



Conclusion

Insights

1. **Solo Work Matters**:
   Solo streams have the most direct association with total streams. Artists that focus on solo projects find a consistent increase in overall streams.
2. **Collaboration Benefits**:
   Both featured streams and lead streams have a substantial impact on an artist's popularity, but with slightly different dynamics:
   - The greater intercept value indicates that featured streams provide a solid baseline.
   - Lead streams have a more gradual influence but are nonetheless important for overall success.

Key Takeaways

- **Solo Streams are Key**: Artists with considerable solo work had greater overall streams, highlighting the value of personal branding.
- **Balanced Strategy**: While solo work is predominant, collaborations (both as a feature and as a lead) give additional benefits, allowing musicians to reach a larger audience.
- **Visual Trends**: The scatter plots clearly demonstrate these associations, with regression lines emphasizing the patterns.

This Rust program is intended to analyze Spotify artist streaming data, determining the correlations between several stream types--solo, featured, and lead--and their overall count. It initially parses a CSV file, cleaning the data by removing missing or incorrect items and transforming structured numerical values into a consistent format, such as "10,000". The software goes through each record and extracts the following metrics: total streams, solo streams (streams from songs performed entirely by the artist), featured streams (streams where the artist is a guest or collaborator), and lead streams. These values are saved in a structured manner (ArtistData struct), which allows for easy manipulation and analysis. In the following stage of the workflow, linear regression is used to determine the association between two variables: one independent variable, say solo streams, and one dependent variable, total streams. The application uses the standard linear regression equations to determine the slope and intercept for the best-fit line that best depicts the trend between the two variables. For example, the line of regression for solo streams vs. total streams may be $y=1.27x+194.87$, indicating that every time an artist has another solo stream, their total streams grow by around 1.27, with a baseline value of 194.87 streams when solo streams are zero. Similarly, featured and lead streams are examined to generate regression equations that indicate how these streams connect to an artist's overall performance on Spotify.

In addition to the quantitative analysis, the application uses the plotters library to generate scatter plots that depict these correlations. Each scatter plot depicts the individual data points, or artists, as red dots, indicating their total streams by solo, featured, or lead streams. A regression line, in blue, is overlayed over the scatter plot to highlight the overall trend of the data. The graph has labeled axes for clarity, as well as a caption that displays the regression equation. For example, one may examine the link between Featured Streams and Total Streams to determine if collaborations are more effective and powerful than solo performances in delivering success to an artist.

The application saves these scatter plots as PNG files, making the visuals easy to analyze. Filenames, such as solo_relationship.png, featured_relationship.png, and lead_relationship.png, are determined by the type of relationship being evaluated. This process is now driven by the main function, which begins with parsing the dataset, then does regression analysis, and finally generates visualizations. It also includes error handling to guarantee that issues like missing files or incorrect data are handled gracefully, as well as relevant error messages for troubleshooting purposes.

This tool takes a comprehensive approach to analyzing artist streaming data, using statistical modeling and visual representation to identify music business trends. For example, it can reveal if solo streams or collaborations generate more overall streams, offering valuable information into the methods that will propel an artist's success on Spotify. The software examines 3,000 records from the collection and creates regression models for important associations, providing both a data-driven conclusion and accessible visuals for additional investigation and decision-making.