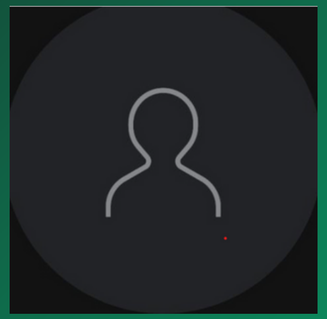
AMS 572- DATA ANALYSIS PROJECT

GROUP 5

Most Streamed Spotify Songs 2023



FOR
PROF. P.F.KUAN

Team Members

SRI DATTA	115939828
ADITI SUNIL BADHE	115824696
SAGARI CHANDRASHEKAR	115808564
MONI GAYATHRI SAYANA	115800067

CONTENTS

SERIAL NUMBER	TOPICS	SLIDE NUMBER
1	OVERVIEW	5
2	DATA PREPROCESSING	6-7
3	EXPLORATORY DATA ANALYSIS	8-9
4	HYPOTHESIS 1	10-13
5	HYPOTHESIS 2	14-16
6	CONCLUSION	17

OVERVIEW

- Spotify is a trailblazer in this era of digital streaming services, changing the way people listen to music and influencing current trends. The dataset under examination provides a comprehensive look into the most streamed Spotify songs of 2023, transcending typical datasets by offering a rich array of features.

- It explores the dynamic features of the 2023 music landscape, unveiling trends, artist performance across platforms, and temporal shifts in musical preferences. The focus is on comprehending the multifaceted dynamics that drive music's popularity in 2023.

- This dataset goes beyond usual datasets, including track details, artist info, release dates, Spotify and Shazam charts, streaming metrics, platform presence, and audio features.

- In our dataset analysis, we've incorporated a diverse range of variables, encompassing categorical parameters such as track names, artists, musical key, and mode. Additionally, we've delved into discrete variables, including counts of artists, release year, month, and day, as well as playlist and chart inclusions. The analysis extends to continuous variables, capturing essential aspects like the total number of streams, tempo, and various musical attributes.

DATA PREPROCESSING

Data preprocessing is foundational for accurate and reliable insights in any data analysis project.

Our Spotify dataset consists of 953 records and 24 columns and we performed the following data preprocessing.

Handling Null Values:

- 'in_shazam_charts': Replaced NULL values with maximum(in_shazam_charts)+1 mitigating the impact of missing entries on analysis.
- In 'key' Column: Transfered the keys column into numeric values and filled the null values with -1.

These measures set the groundwork for precise and comprehensive analysis of the Spotify dataset.

DATA PREPROCESSING

Skewness, indicating the asymmetry of a distribution, has implications for assumptions in statistical analyses like linear regression. To address this, the Box-Cox transformation, a family of power transformations encompassing logarithmic transformations, is employed.

BEFORE BOXCOX

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

[1] "Skewness of BPM: 0.412594824509775"
[1] "Skewness of Danceability: -0.435191771397039"
[1] "Skewness of Valence: 0.00821058753712152"
[1] "Skewness of Energy: -0.445696288721271"
[1] "Skewness of Acousticness: 0.950961889108647"
[1] "Skewness of Instrumentalness: 7.11299898856884"
[1] "Skewness of Liveness: 2.10096650815347"
[1] "Skewness of Speechiness: 1.93162188752986"
```

AFTER BOXCOX

	Original	Transformed
bpm	0.412594825	-1.0614214
danceability_	-0.435191771	-0.7280523
valence_	0.008210588	-1.2262866
energy_	-0.445696289	-0.9780762
acousticness_	0.950961889	-1.0395375
instrumentalness_	7.112998989	-0.8448425
liveness_	2.100966508	-1.2311982
speechiness_	1.931621888	-1.0385376

The Box-Cox transformation is a versatile tool that can enhance the distributional properties of variables, addressing issues such as skewness,

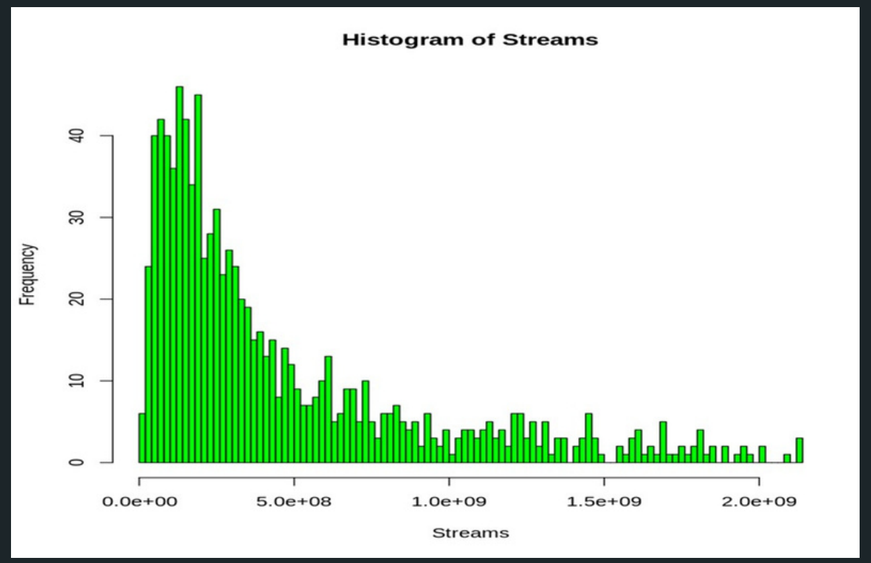
EXPLORATORY DATA ANALYSIS

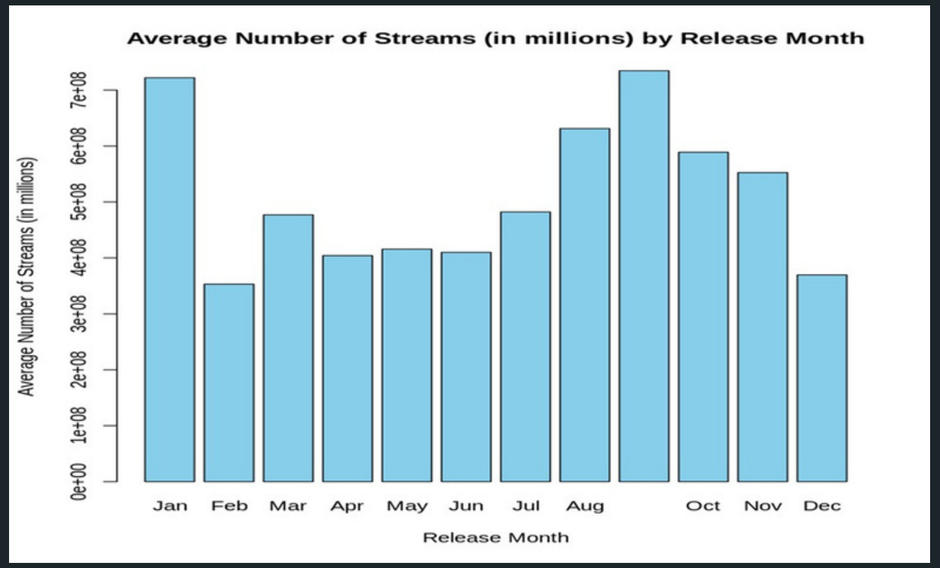
EDA is a data analysis approach focused on visually exploring and summarizing datasets to uncover patterns, relationships, and insights, guiding further analysis and modeling decisions.

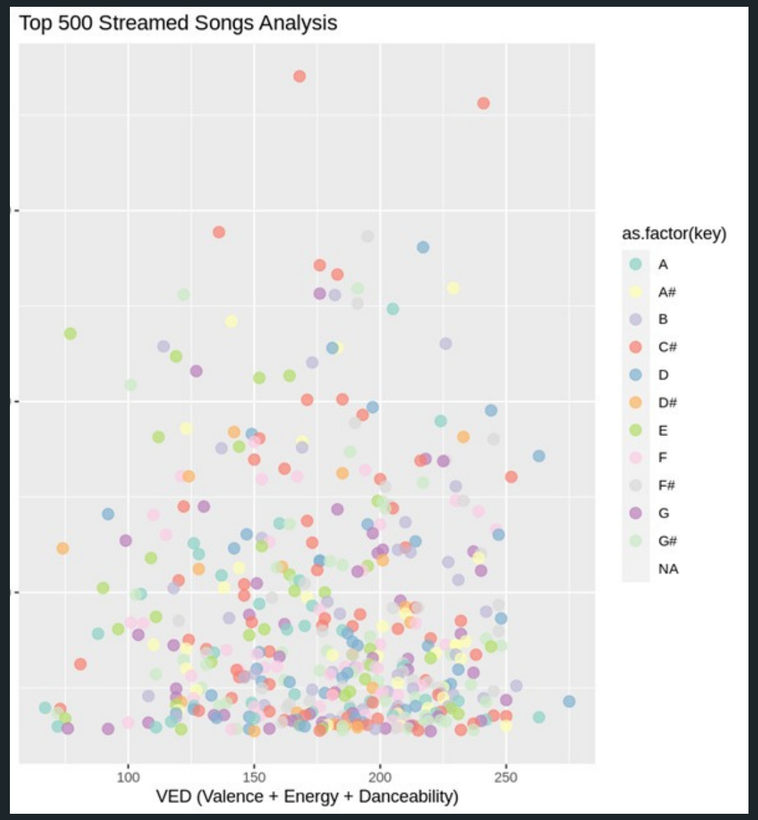
Following are some of the analysis that we have carried out to explore the data visually.

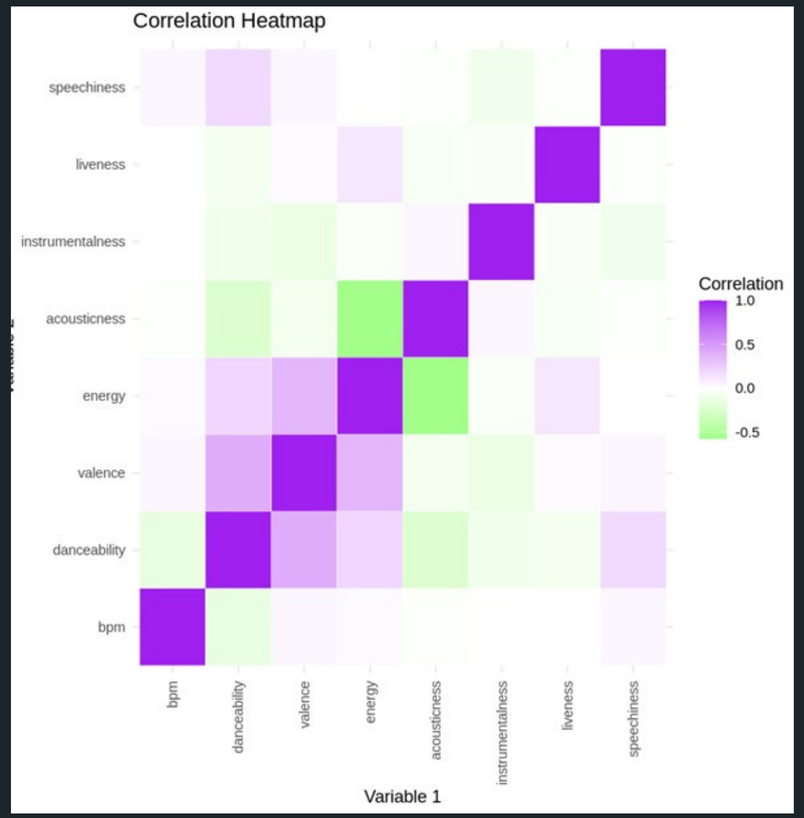
- Evolution of Spotify Song Releases Over Time
- Monthly Variation in Average Spotify Streams
- Distribution of Song Releases Across Days of the Month
- Top Artists in the Top 10 List
- Distribution of Song Streams on Spotify
- Correlation between Playlist Count and Stream Count
- Insights into the Music Landscape
- Attributes and Streams of Top 500 Songs
- Relationships between Musical Attributes and Streams
- Correlation Heatmap of Musical Attributes
- Residual Analysis in Regression Model

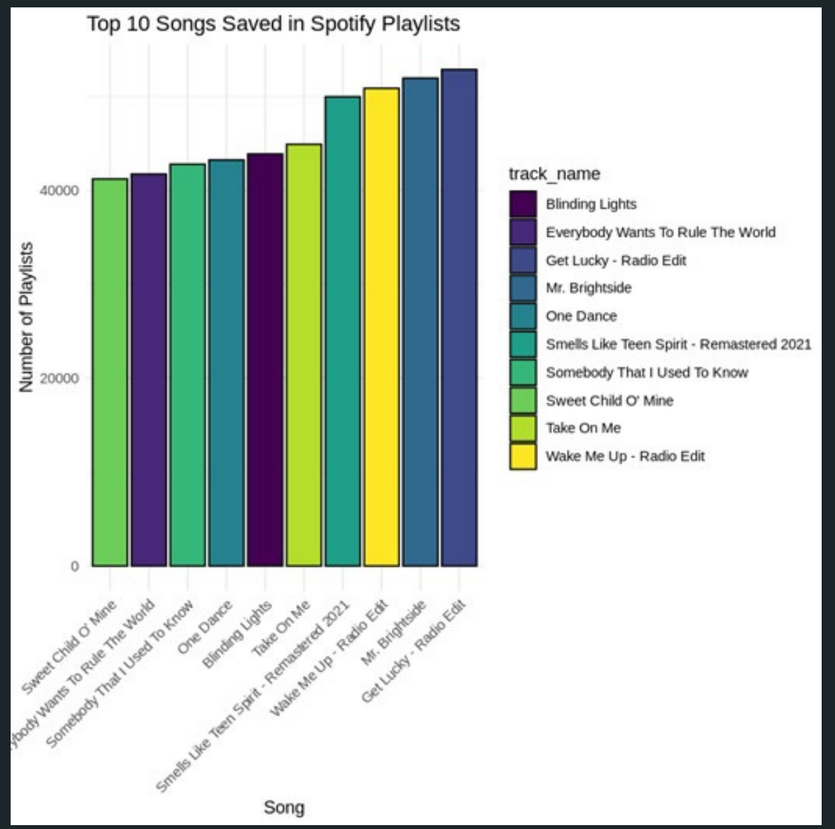
Next slide shows some glimpses of the analysis performed.


Histogram of Streams
Frequency
0
10
20
30
40
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
Streams


Average Number of Streams (in millions) by Release Month
Average Number of Streams (in millions)
0e+00
1e+08
2e+08
3e+08
4e+08
5e+08
6e+08
7e+08
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Oct
Nov
Dec
Release Month


Top 500 Streamed Songs Analysis
as.factor(key)
A
A#
B
C#
D
D#
E
F
F#
G
G#
NA
100
150
200
250
VED (Valence + Energy + Danceability)


Correlation Heatmap
speechiness
liveness
instrumentalness
acousticness
energy
valence
danceability
bpm
Correlation
1.0
0.5
0.0
-0.5
bpm
danceability
valence
energy
acousticness
instrumentalness
liveness
speechiness
Variable 1


Top 10 Songs Saved in Spotify Playlists
Number of Playlists
0
20000
40000
track_name
Blinding Lights
Everybody Wants To Rule The World
Get Lucky - Radio Edit
Mr. Brightside
One Dance
Smells Like Teen Spirit - Remastered 2021
Somebody That I Used To Know
Sweet Child O' Mine
Take On Me
Wake Me Up - Radio Edit
Sweet Child O' Mine
Everybody Wants To Rule The World
Somebody That I Used To Know
One Dance
Blinding Lights
Take On Me
Smells Like Teen Spirit - Remastered 2021
Wake Me Up - Radio Edit
Mr. Brightside
Get Lucky - Radio Edit
Song

HYPOTHESIS 1

Formulating Hypotheses: Presented the null and alternative hypotheses:

- Null Hypothesis (H0): Inclusion in Spotify playlists is independent of the half of the year the song is released.
- Alternative Hypothesis (H1): Inclusion in Spotify playlists is dependent on the half of the year the song is released.

Data Preparation and Chi-square Test Assumptions:

- A newly introduced categorical variable, "release_half," was established to categorize songs into either the "First Half" or "Second Half" based on their release months.
- The Chi-square test operates under the assumption of independence as the null hypothesis, while the alternative hypothesis proposes a correlation between the inclusion of songs in Spotify playlists and the release half.

HYPOTHESIS 1

Decision Rule:

If p-value < α, reject null hypothesis; if p-value > α, fail to reject null hypothesis. Upon doing the analysis and after performing the chi-square test, we get the following output.

Chi-square Test Result

```
Chi square Test
spotify_data$release_half <- ifelse(spotify_data$released_month <= 6,
"First Half", "Second Half")
chi_square_test <- chisq.test(table(spotify_data$in_spotify_playlists,
spotify_data$release_half))
chi_square_test
```

```
        Pearson's Chi-squared test

data:  table(spotify_data$in_spotify_playlists, spotify_data$release_half)
X-squared = 897.25, df = 878, p-value = 0.3185
```

P Value is greater than 0.05 so we fail to reject null hypothesis

HYPOTHESIS 1

We, then explore the impact of missing values on data analysis in two scenarios:

MCAR (Missing Completely at Random): After conducting the chi-square test, we proceeded to assess the impact of missing values at various percentages (10%, 20%, 30%, 40%, 50%) using MCAR. The comparison of p-values from the chi-square test and MCAR yielded the following results.

	Missingness	Chi-Square p-value	MCAR p-value	Chi-Square Hypothesis	MCAR Hypothesis
1	0.1	0.0169915042478761	0.768239417003419	Reject H0	Fail to reject H0
2	0.2	0.0494752623688156	0.609580087602564	Reject H0	Fail to reject H0
3	0.3	0.088455772113943	0.772544862280284	Fail to reject H0	Fail to reject H0
4	0.4	0.428785607196402	0.23739998761442	Fail to reject H0	Fail to reject H0
5	0.5	0.197401299350325	0.375005108768318	Fail to reject H0	Fail to reject H0

At missingness levels of 0.1 and 0.2, the rejection of the null hypothesis (H0) indicates a significant association between the variables. Conversely, for missingness levels of 0.3, 0.4, and 0.5, the failure to reject H0 suggests no significant association at these elevated levels of missing data.

HYPOTHESIS 1

MNAR (Missing Not at Random):

- The decision to exclude MNAR for the Chi-Square test likely stems from the absence of a native mechanism to handle MNAR data in this test. Moreover, MNAR data can introduce substantial bias into Chi-Square test results.

- In contrast, Generalized Linear Models (GLMs) offer adaptability to address MNAR by incorporating additional parameters. These parameters can model the probability of missingness based on observed data and the missing data itself.

- The result for chi-sqaure remains the same before and
after applying the Generalised Linear Models.

```
Warning message:
“glm.fit: algorithm did not converge”

Call:
glm(formula = in_spotify_playlists_missing ~ in_spotify_charts +
    streams + released_year + artist_count + danceability_ +
    valence_ + energy_, family = binomial(), data = spotify_data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.657e+01  2.254e+06       0        1
in_spotify_charts  5.634e-16  6.160e+02       0        1
streams           -1.089e-23  2.194e-05       0        1
released_year     -4.570e-17  1.122e+03       0        1
artist_count       1.306e-15  1.338e+04       0        1
danceability_     -9.691e-17  9.000e+02       0        1
valence_           1.471e-16  5.764e+02       0        1
energy_            5.685e-19  7.577e+02       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 951  degrees of freedom
Residual deviance: 5.5231e-09  on 944  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 16

Number of Fisher Scoring iterations: 25
```

HYPOTHESIS 2

Formulating Hypotheses: Presented the null and alternative hypotheses:

- Null Hypothesis (H0): There is no significant linear relationship between the selected features in the dataset and the number of streams a song receives on Spotify.

- Alternative Hypothesis (H1): At least one of the features in the dataset has a significant linear relationship with the number of streams a song receives on Spotify.

Linear Regression Model:

Constructed a multiple linear regression model using R's lm function. The target variable "streams" is regressed against all other dataset variables, exploring relationships and influences among these factors.

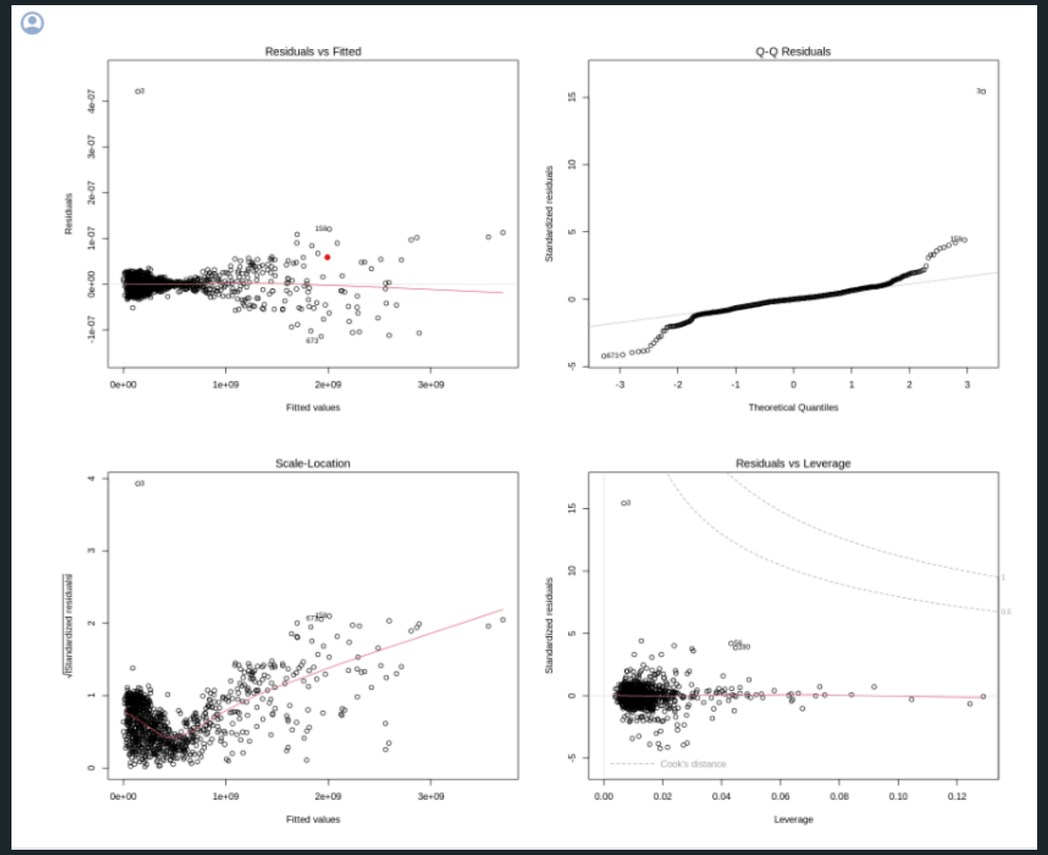
Utilized the summary function to obtain a comprehensive overview of the regression model. Retrieved crucial details, encompassing coefficient significance, residual standard error, R-squared values, F-statistic, and p-value.

HYPOTHESIS 2

Output

```
Warning message in summary.lm(mlr_model):
“essentially perfect fit: summary may be unreliable”
Signif. codes: 1 1 0 0.710658 0.795176 0.243884 0.748054 0.243252 0.650146 0.601409 0.594247 0.842593 0.105047 0.117404
Residual standard error: 2.741653e-08 on 938 degrees of freedom
Multiple R-squared: 1e+00       Adjusted R-squared: 1e+00
F-statistic: 3.127224e+34 on 14 and 938 DF, p-value: NA
[1] -30435.12
[1] "Residual Standard Error: 2.74165329402922e-08"
```

These plots illustrate the outcomes of the conducted hypothesis tests. The predicted lines in red effectively align with the observed data, indicating a strong fit between the model predictions and the actual dataset. This alignment reinforces the validity of the hypothesis and suggests that the regression model accurately captures the underlying patterns in the data.



HYPOTHESIS 2

Next we applied Shapiro-Wilk test on the residuals of the multiple linear regression model.

Interpretation: If p-value < 0.05, reject the null hypothesis, indicating that residuals are not normally distributed. If p-value ≥ 0.05, fail to reject the null hypothesis, suggesting normal distribution of residuals.

```
        Shapiro-Wilk normality test

data:  residuals_mlr_model
W = 0.76282, p-value < 2.2e-16

The p-value is less than 0.05 - reject the null hypothesis; the residuals are not normally distributed.
```

The normality of residuals was assessed using the Shapiro-Wilk test. The test resulted in a W statistic of 0.24822 and a p-value less than 2.2e-16. Since the p-value is below the conventional significance level of 0.05, we reject the null hypothesis. This indicates that the residuals do not follow a normal distribution

CONCLUSION

The project's conclusion successfully analyzes Spotify streaming data to show the variety of musical qualities that influence a song's popularity. The model's strong predictive ability was demonstrated by the use of multiple linear regression, which provided a nuanced understanding of the interactions between various factors influencing streams. The dataset's reliability is ensured by the randomness of its missing data, as demonstrated by the results of the Chi-Square and MCAR tests. These technical insights provide practical advice for music industry stakeholders to maximize streaming success in addition to validating the study's methodology.

Thank you