How Musical characteristics influence Spotify streams

To what extent do a song's mode, danceability, and energy influence its Spotify streams, and does valence moderate these effects?

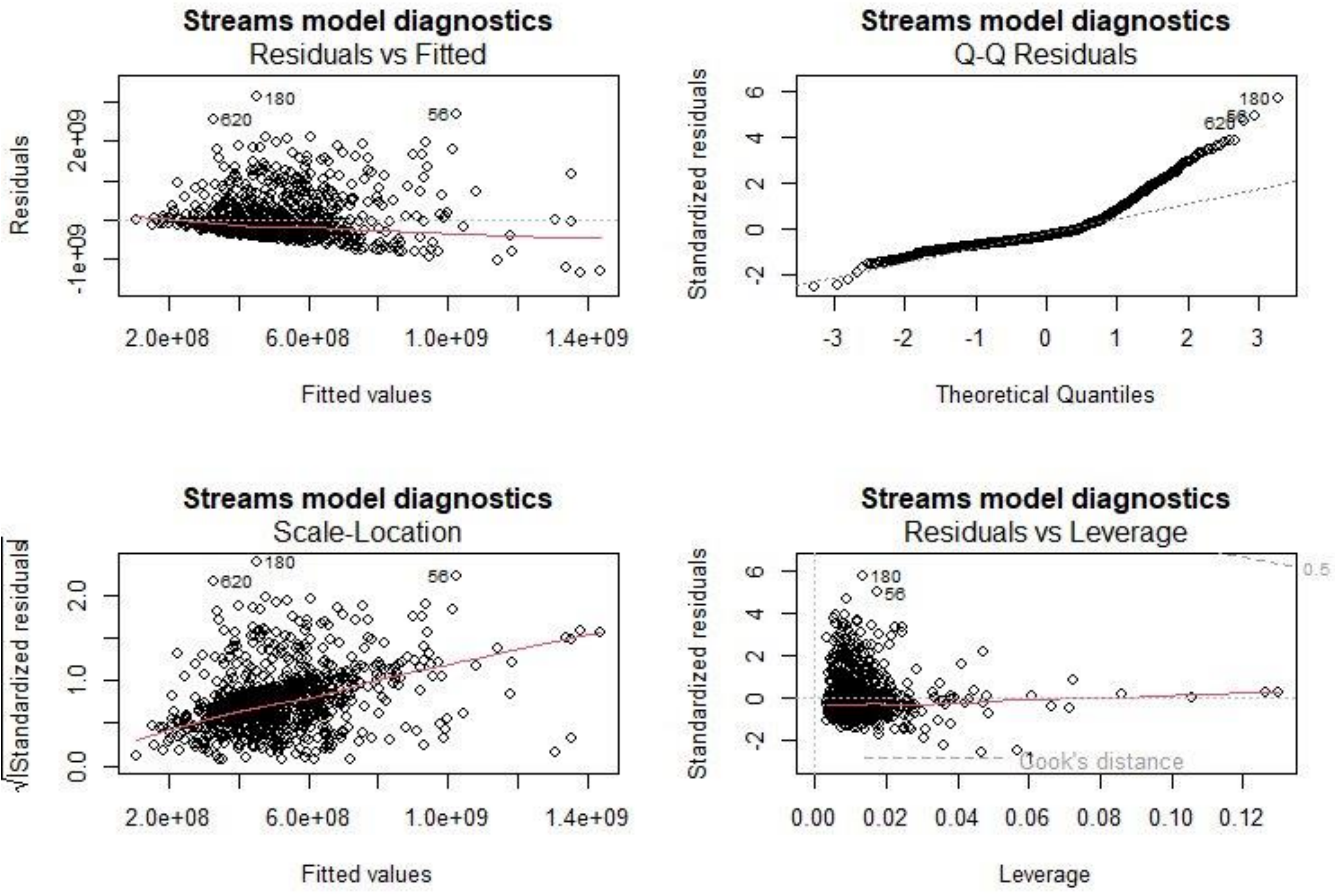
Motivation

•Spotify is the world's largest music streaming platform, and artists rely on understanding what musical qualities help a song gain streams
•Prior research such as Matera (2021) and Charchyan (2024) shows that danceability, energy, mode, and valence are key musical traits that may influence a song's popularity,
•Knowing how these features relate to stream counts can help artists, producers, and the music industry make more informed and creative marketing decisions
•A linear regression model allows us to measure each feature's effect while controlling for others and test whether valence modifies these relationships.

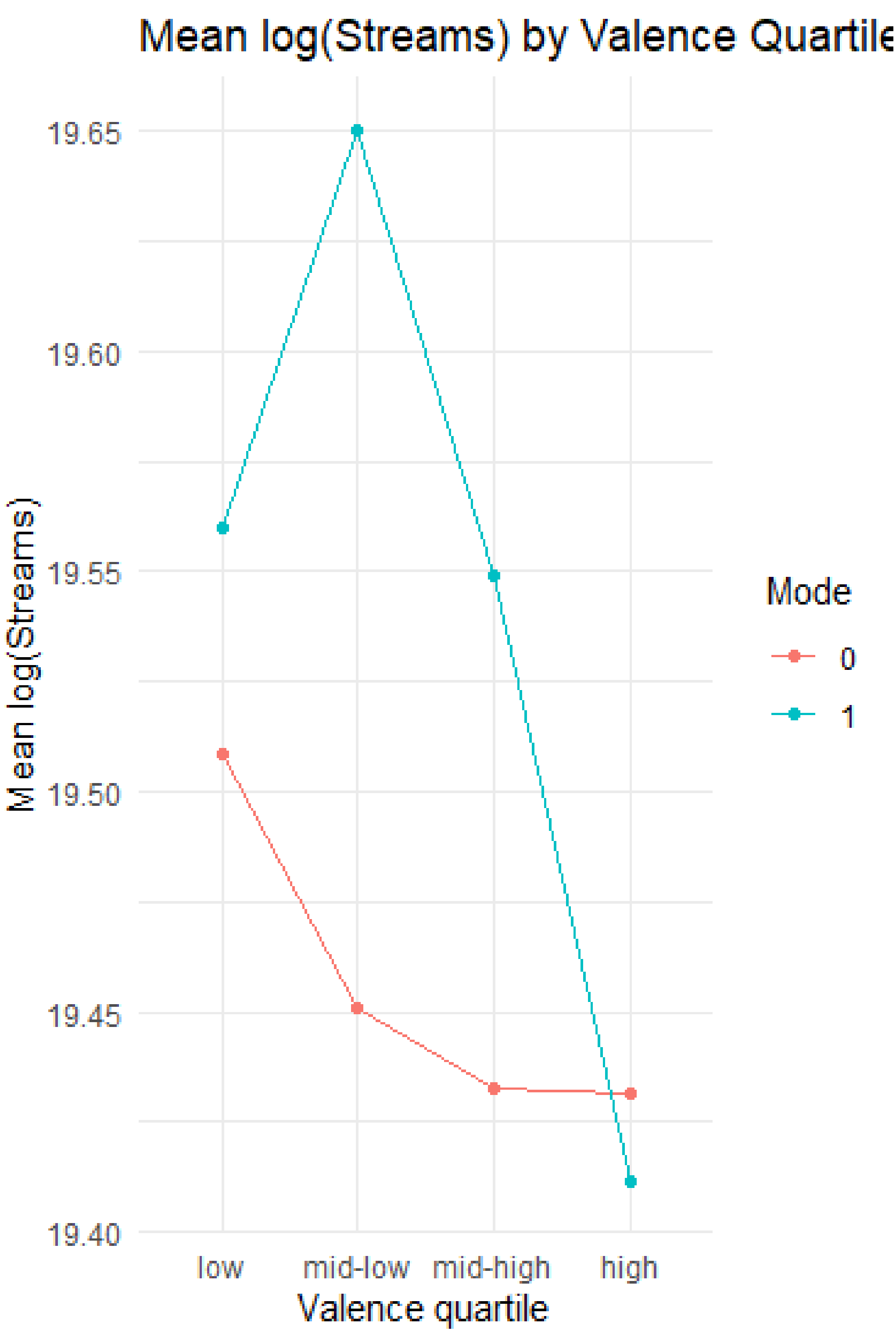
Data Collection

-The dataset was created using the Spotify Web API, which retrieves information directly from Spotify's internal database.
-Dataset "Top Spotify Songs of 2023" was retrieved from Kaggle by author Nidula Elgiriyewithana.
-The dataset provides 952 songs and 24 variables with consistent, algorithm generated measurements from Spotify's API.
-Variables: Spotify streams- continuous response, mode- categorical predictor, danceability and valence- numerical predictors
-The curator combined the Spotify API results, cleaned the data, and published it as a CSV file.
-Represents the population of interest: highly streamed Spotify tracks.
-Standardized Spotify measurements improve reliability and accuracy, supporting valid statistical analysis.

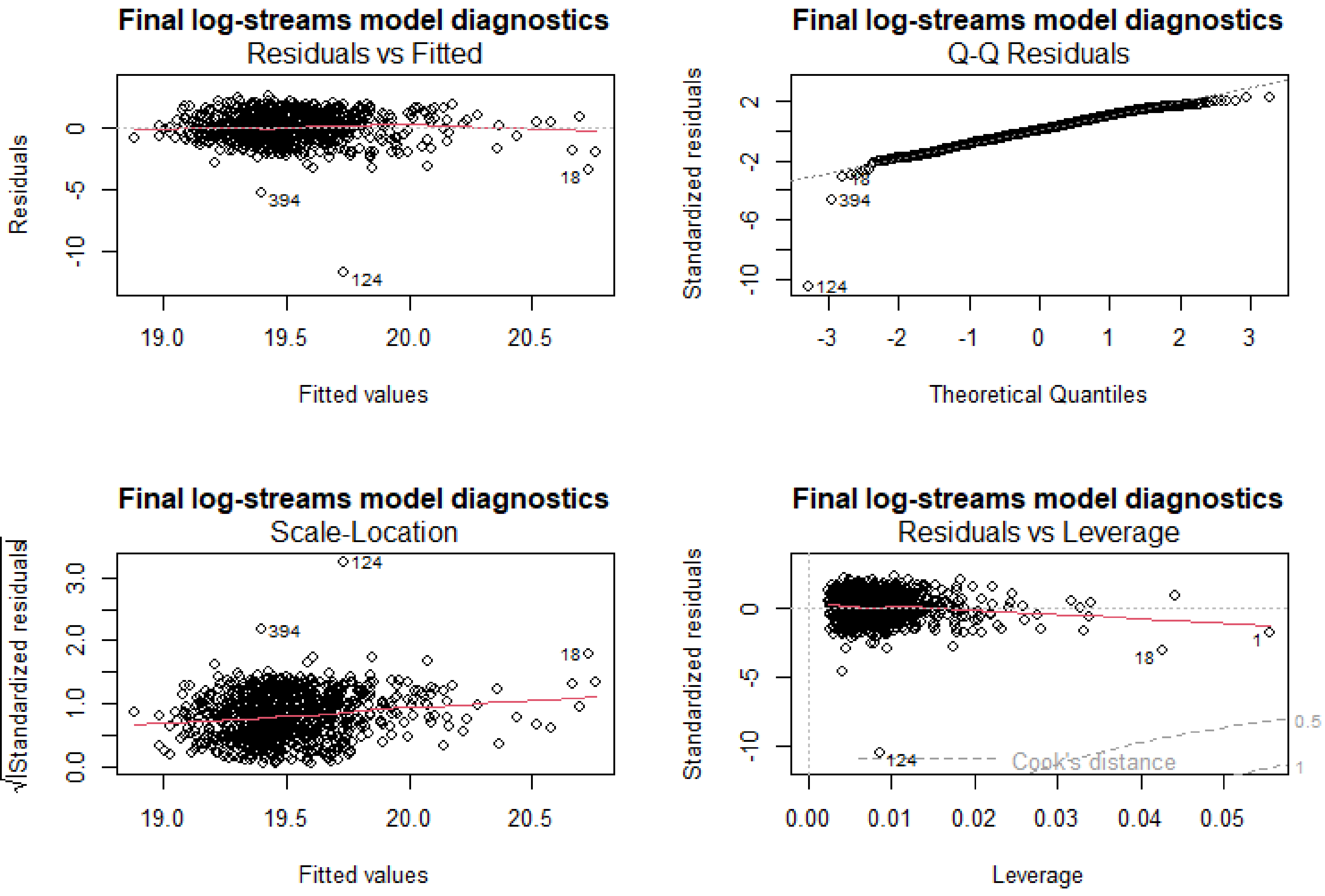
Methods of Analysis



•N=952
•Mode: 1 for major 0 for minor
•Interaction: valence x mode based on literature
•Streams~ charts + danceability + valence + energy + acousticness + instrumentalness + liveness + speechiness + bpm + mode + valence_mode
•Summary plots > model assumptions > influential points > compare models with R^2, adjusted R^2 > model validation



Results and Conclusions



- Applied logarithmic and square root transformation to fix right skew
- Strong heteroscedasticity in residuals
- Backwards selection starting from full model and removing non significant variables but keeping variables supported by literature
- Log(Streams)~ charts +danceability + valence + energy +mode + valence_mode + speechiness
- Major tracks impacted streams more than minor tracks but neither were significant
- R^2 = 0.04
- 5-fold CV RMSE = 1.13(final) vs 1.15 (intercept only)
- VIF all < 10
- Conclude that charts and speechiness are the most significant predictors for log number of streams but did not find significant evidence that mode, danceability, energy, valence, and valence_mode impacted log streams on Spotify by a meaningful amount

Limitations

- Dataset only includes Spotify's *most streamed songs of 2023*, mostly pop and hip-hop → limits generalizability.
- Predictors draw from *Apple Music, Shazam,* and *Deezer*, which may add platform bias due to different user bases.
- *Log transformation* improved linearity, but mild heteroscedasticity and multicollinearity remain.
- Model fit is weak (*R² = 0.041*), suggesting missing nonlinear or interaction effects.
- Future work:
  - •Broaden dataset across genres and include artist popularity or promotion factors.
  - •Normalize cross-platform data.
  - •Test nonlinear models (e.g., Random Forest, Boosting).

References

•Charchyan, A. (2024). *Exploring trends in music platforms: A comparative analysis of key factors for trending songs on Spotify and YouTube.* American University of Armenia.

•Li, K. (2024). *Predicting song popularity in the digital age through Spotify's data* [Undergraduate thesis, University of Toronto]. Semantic Scholar.

•Matera, M. (2021). *The music industry in the streaming age: Predicting the success of a song on Spotify* [Master's thesis, LUISS Guido Carli University]. ProQuest Dissertations Publishing.

•Elgiriyewithana, N. (2023). *Most Streamed Spotify Songs 2023 [Dataset].* Kaggle.