Genre Classification with a Kaggle Spotify Dataset

MSCS Final Group #5
Team: Christian Horton, Sithara Samudrala, and Bob Owens
Data Mining and Analytics Summer 2023

Data Description and Project Goals

- **Dataset**
  - Kaggle Spotify_1Million_Tracks: https://www.kaggle.com/datasets/amitanshjoshi/spotify-1million-tracks
  - Approximately 1 million tracks with 19 data elements between the year 2000 - 2023
  - There are 61,445 unique *artists* and 82 unique *genres*
  - Interesting Metrics including *danceability*, *speechiness*, *acousticness*, and *instrumentalness*
- **Project Goals**
  - Are the metrics recorded for a song enough to predict the genre that the song belongs to?
  - Supervised Learning to determine which best classifies the genre of a song: *logistic regression*, *decision tree classifier*, and *random forest classifier.*

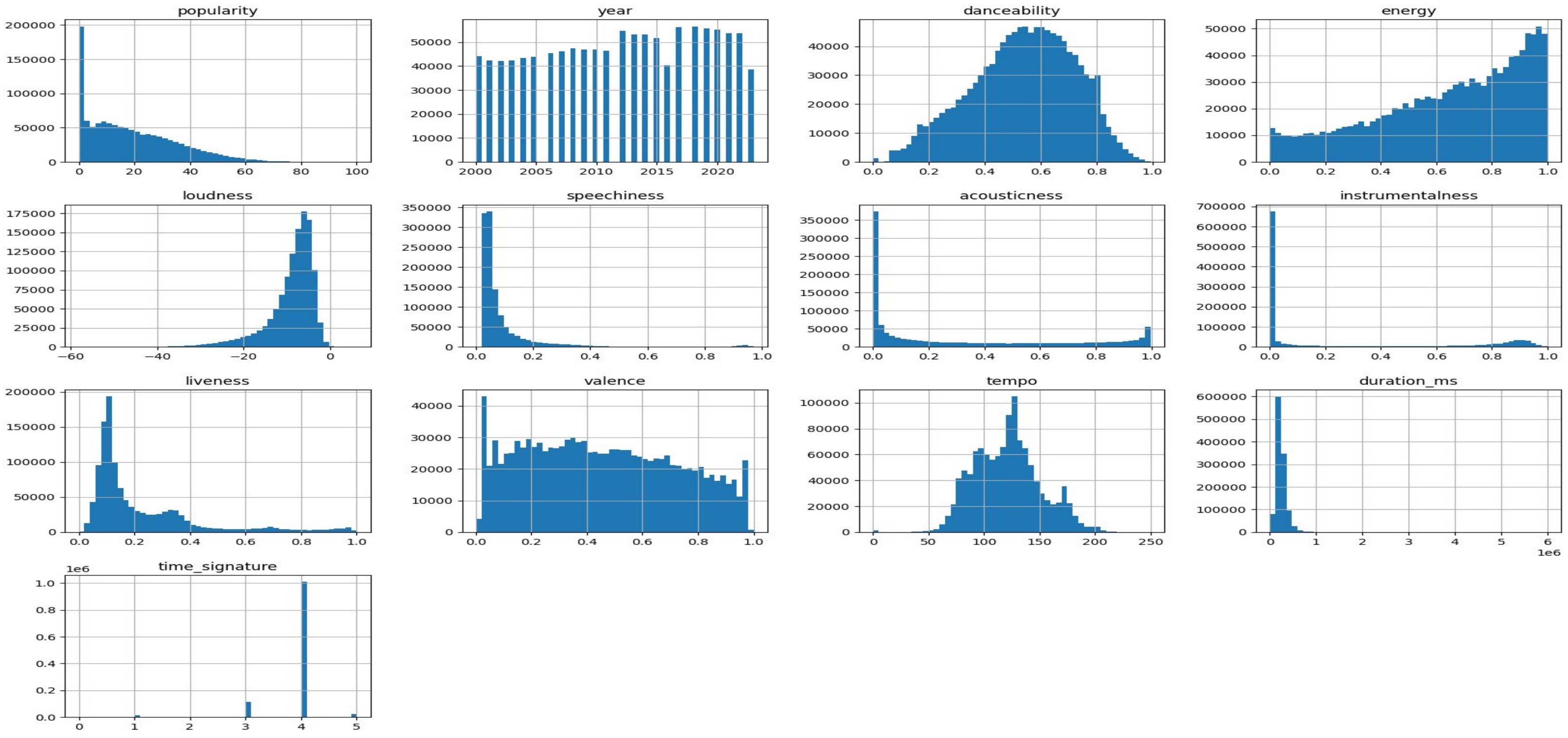
Audio Features	Description
Popularity	Track popularity (0 to 100)
Year	Year released (2000 to 2023)
Danceability	Track suitability for dancing (0.0 to 1.0)
Energy	The perceptual measure of intensity and activity (0.0 to 1.0)
Key	The key, the track is in (-1 to -11)
Loudness	Overall loudness of track in decibels (-60 to 0 dB)
Mode	Modality of the track (Major '1'/ Minor '0')
Speechiness	Presence of spoken words in the track
Acousticness	Confidence measure from 0 to 1 of whether the track is acoustic
Instrumentalness	Whether tracks contain vocals. (0.0 to 1.0)
Liveness	Presence of audience in the recording (0.0 – 1.0)
Valence	Musical positiveness (0.0 to 1.0)
Tempo	Tempo of the track in beats per minute (BPM)
Time_signature	Estimated time signature (3 to 7)
Duration_ms	Duration of track in milliseconds

Data and Feature Preparation

- Data Preparation
  - Data elements were assessed based on normality, correlations, and correlation of data elements by genre
  - Outliers were dropped if the data point was 1.5 times plus or minus the IQR
- Feature Preparation
  - 82 unique genres were reduced into a “base” genre with a pandas UDF
  - Data was further limited to the top genres according to average popularity
  - Resulted in ~123 K rows for top 3 genres
  - Columns used:
    - Tempo
    - Time_Signature
    - Liveness
    - Speechiness
    - Acousticness
    - Instrumentalness
    - Duration_ms
    - Valence
    - Energy
    - Key

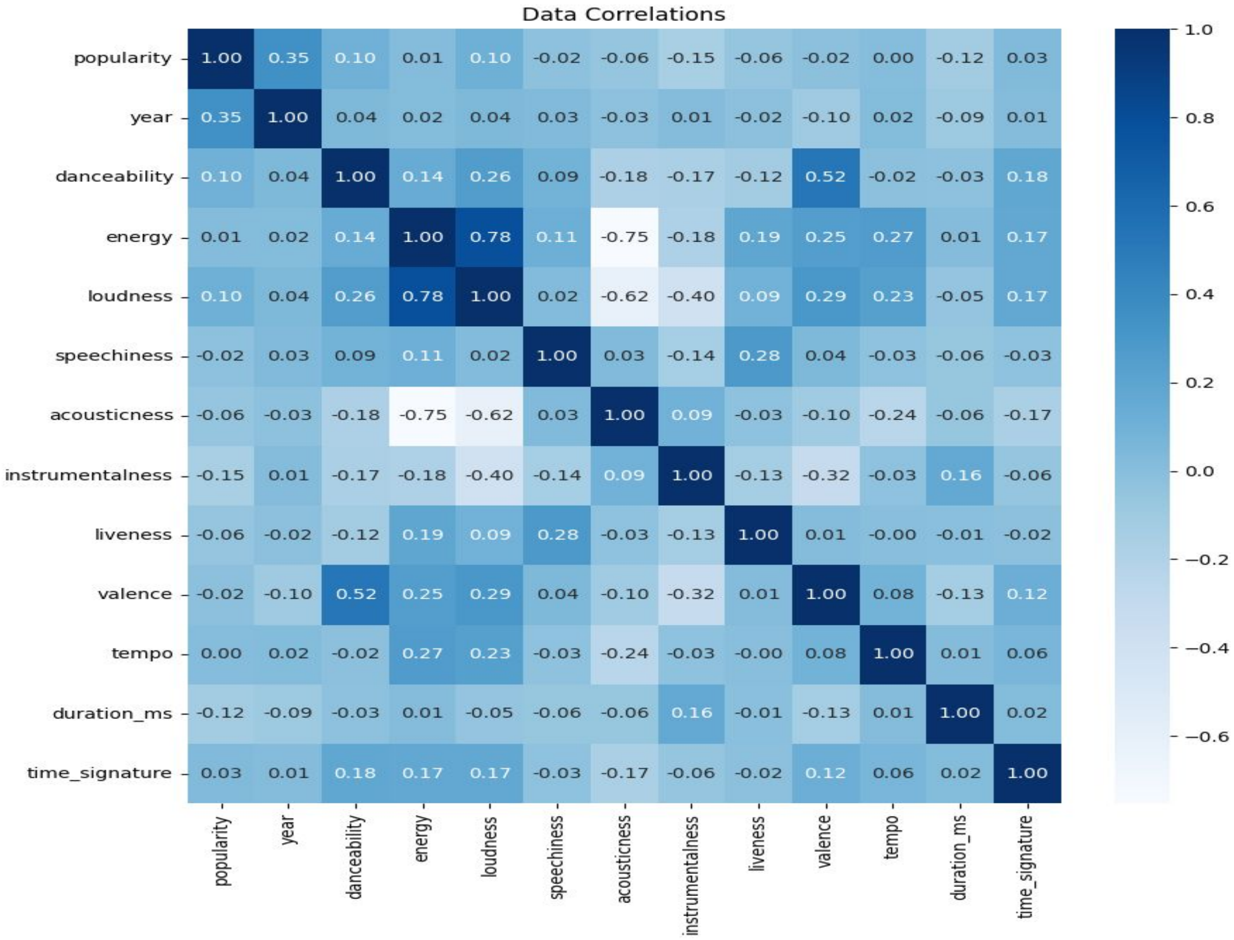
```
root
 |-- artist_name: string (nullable = true)
 |-- track_name: string (nullable = true)
 |-- popularity: integer (nullable = true)
 |-- year: integer (nullable = true)
 |-- genre: string (nullable = true)
 |-- danceability: double (nullable = true)
 |-- energy: double (nullable = true)
 |-- key: integer (nullable = true)
 |-- loudness: double (nullable = true)
 |-- speechiness: double (nullable = true)
 |-- acousticness: double (nullable = true)
 |-- instrumentalness: double (nullable = true)
 |-- liveness: double (nullable = true)
 |-- valence: double (nullable = true)
 |-- tempo: double (nullable = true)
 |-- duration_ms: integer (nullable = true)
 |-- time_signature: integer (nullable = true)

Number of rows in data set: 1159764
```

Distributions of the Dataset

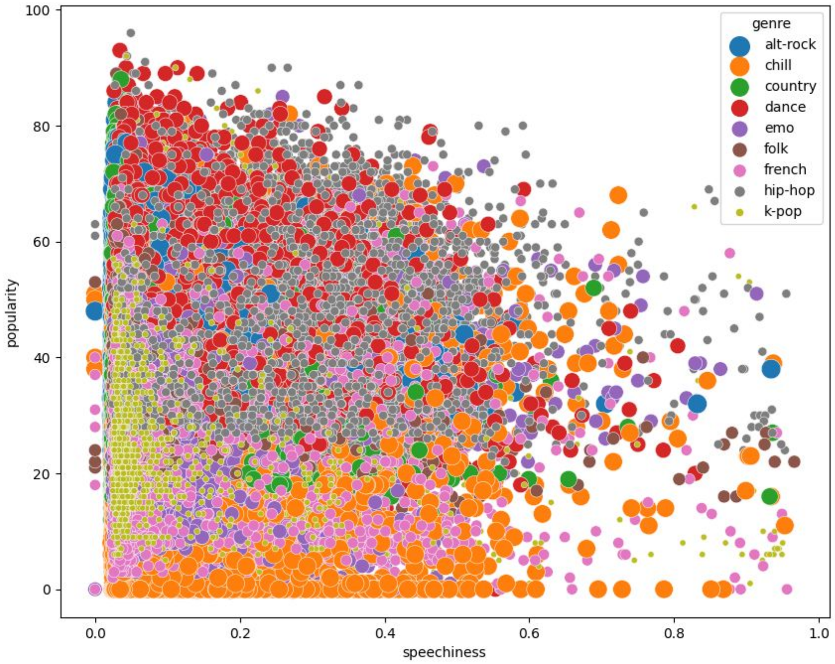


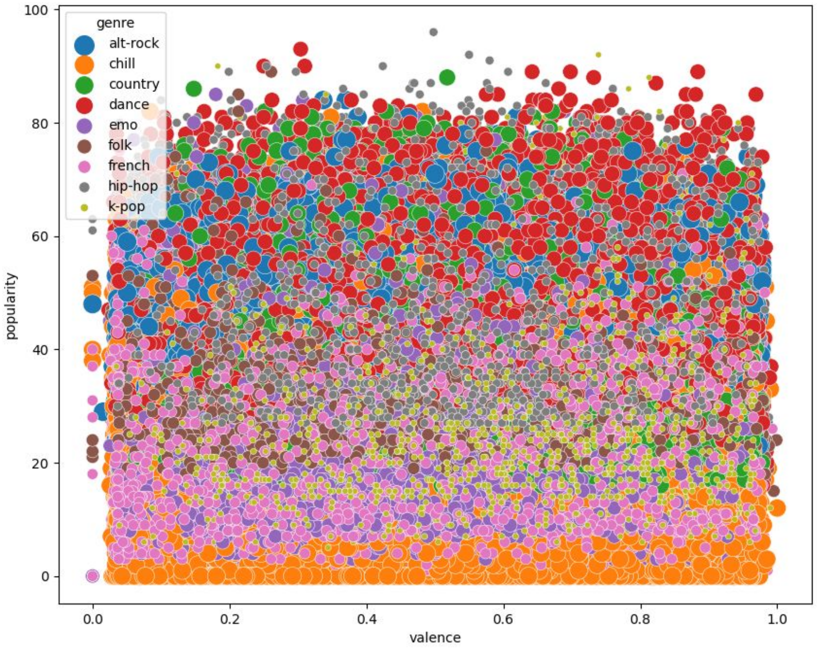
Data Correlations

Visualizing the correlations of all the features in the dataset shows that each feature does not carry much of a relationship to one another.



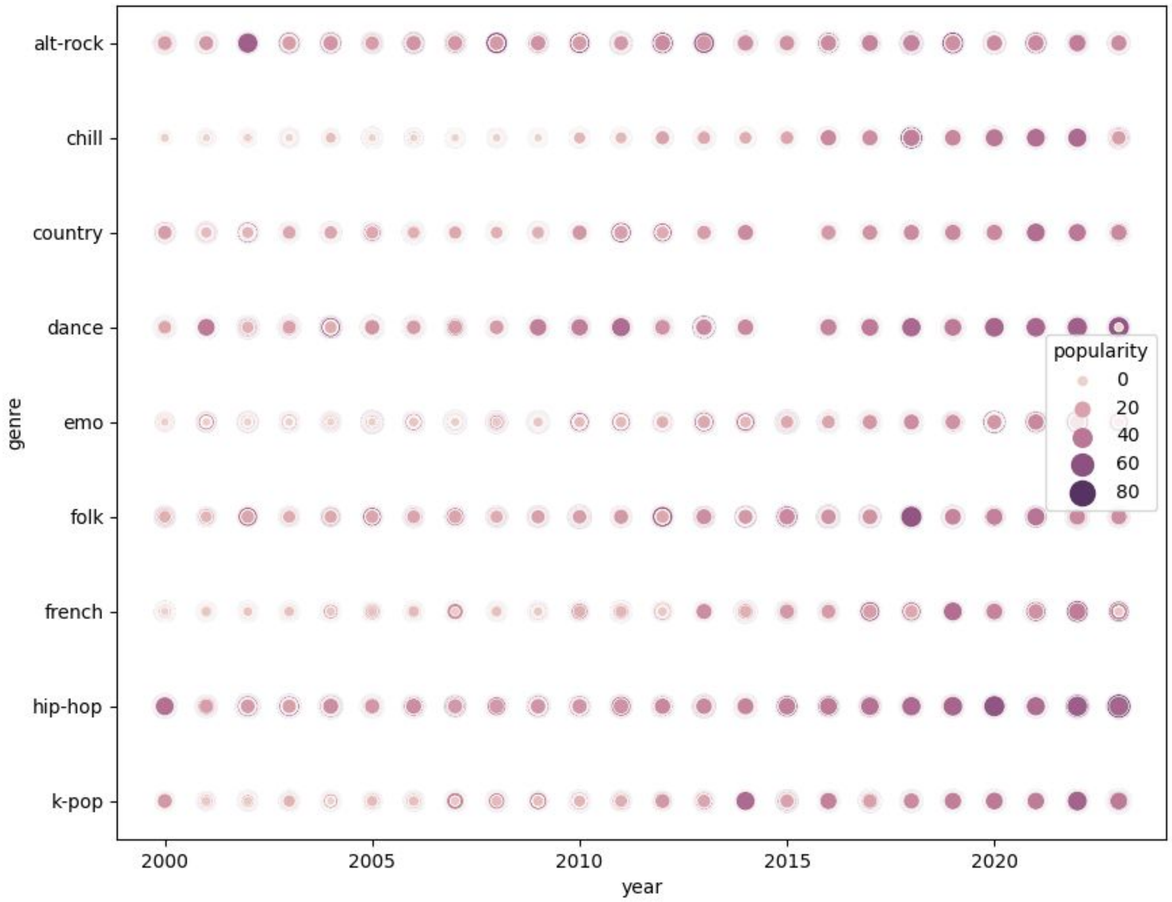
Popularity vs Speechines and Valence





Popularity of several genres

Hip-hop, K-pop, and dance music have risen in popularity since 2000.



Transformation: Regrouping Genres

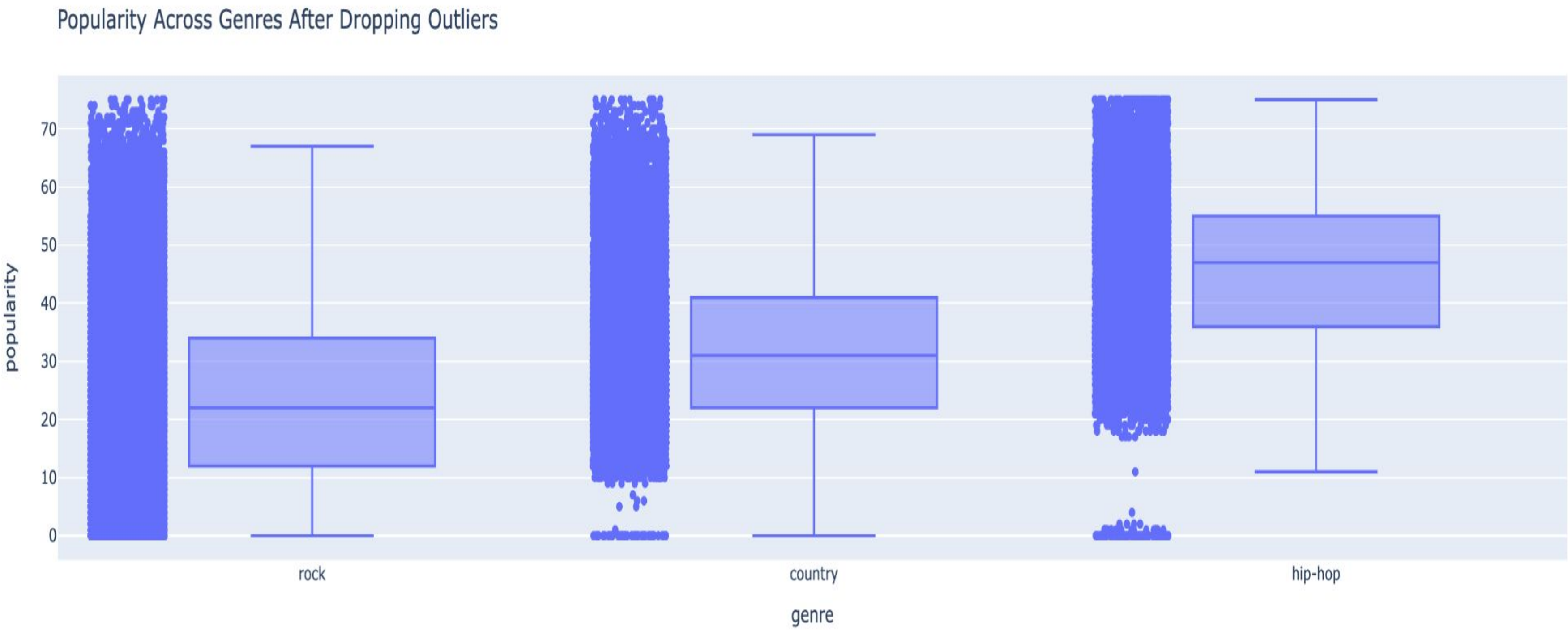
In our dataset, there are over 80 genres that are listed. To help build a stronger model, we regrouped each song's genre into their respective parent genre. For this method we used a UDF transformed the dataset with the new genres.

We further narrowed the genres by selecting rows with the highest average popularity.

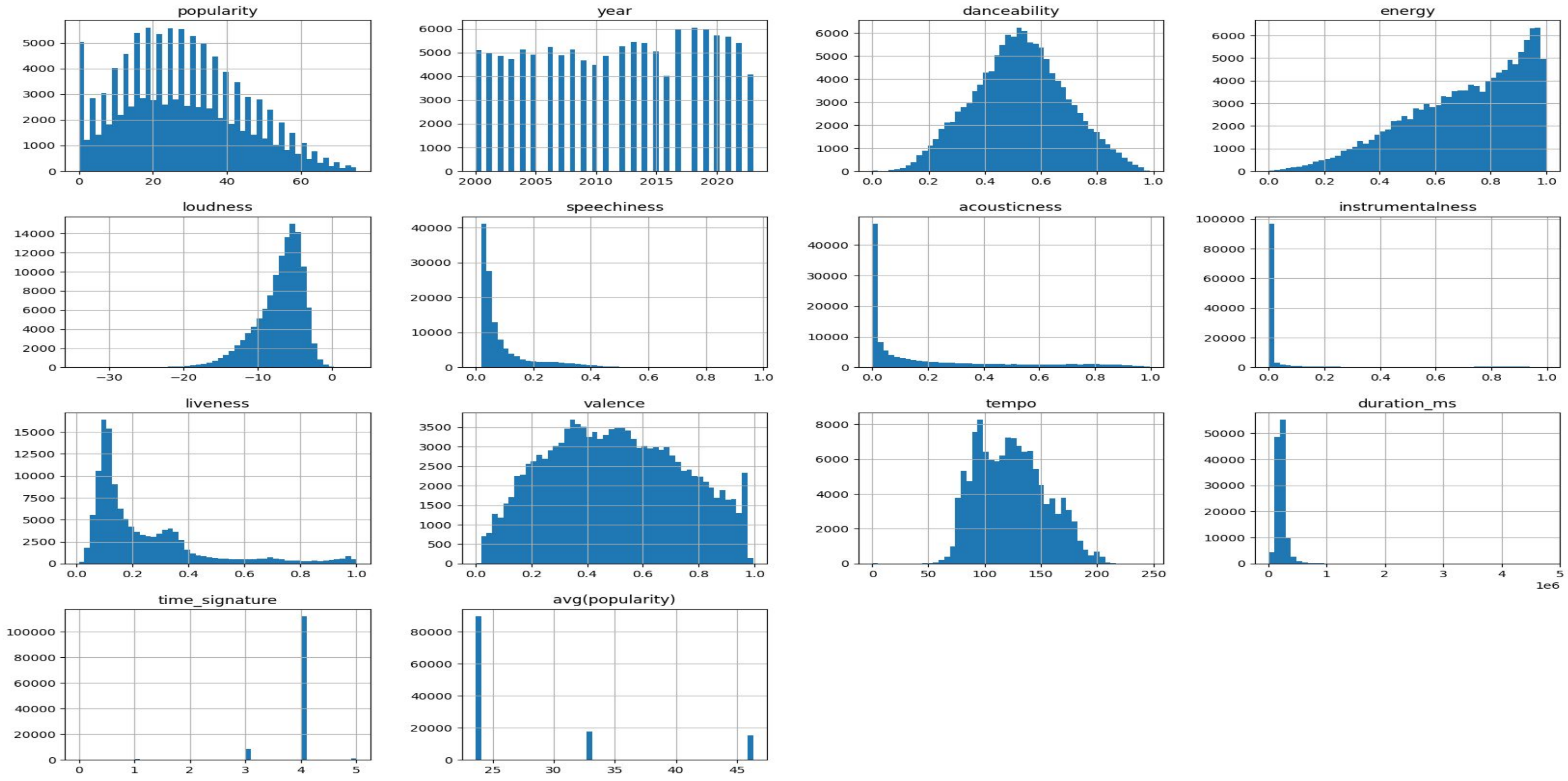
Used PCA with 3 components.

genre	count
electronic	218507
world-music	176649
pop	97203
rock	89796
dance	84838
metal	82548
folk	75675
soul	72130
chill	71512
emo	64285
classical	47122
comedy	19334
country	17883
disco	16987
hip-hop	15703
trance	9592

Outlier Assessment

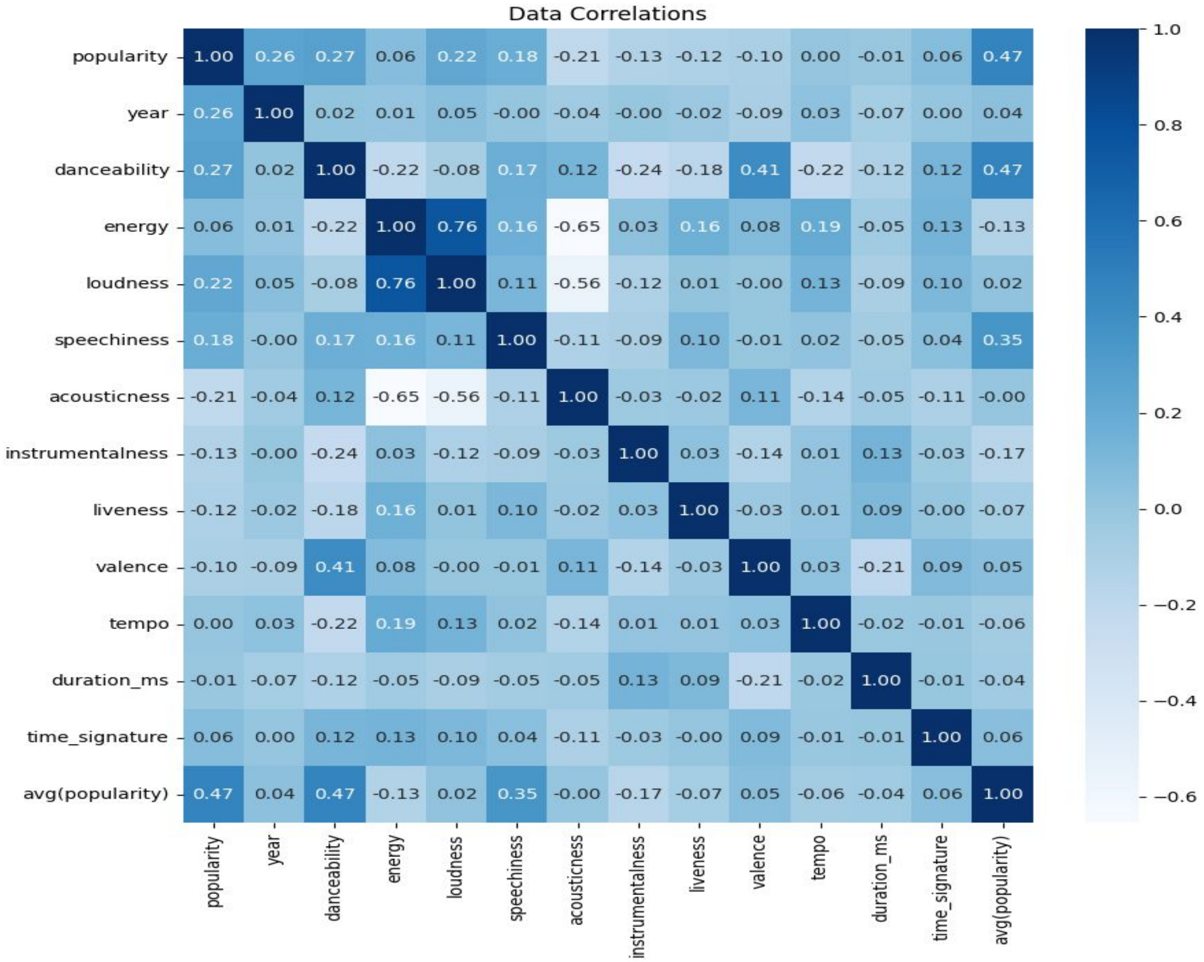


Distributions of the Dataset (post regrouping)

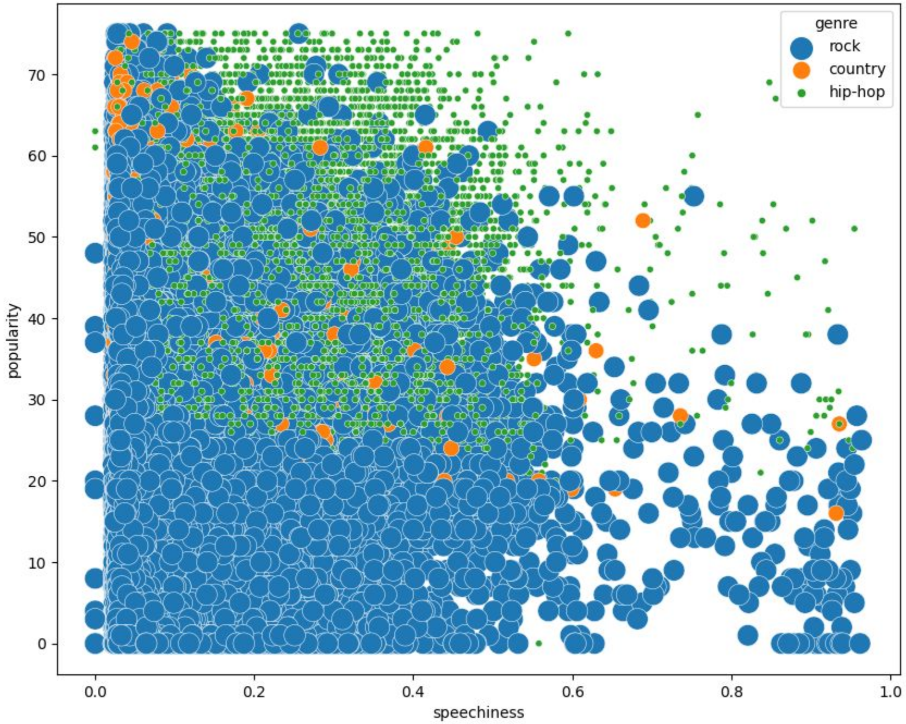


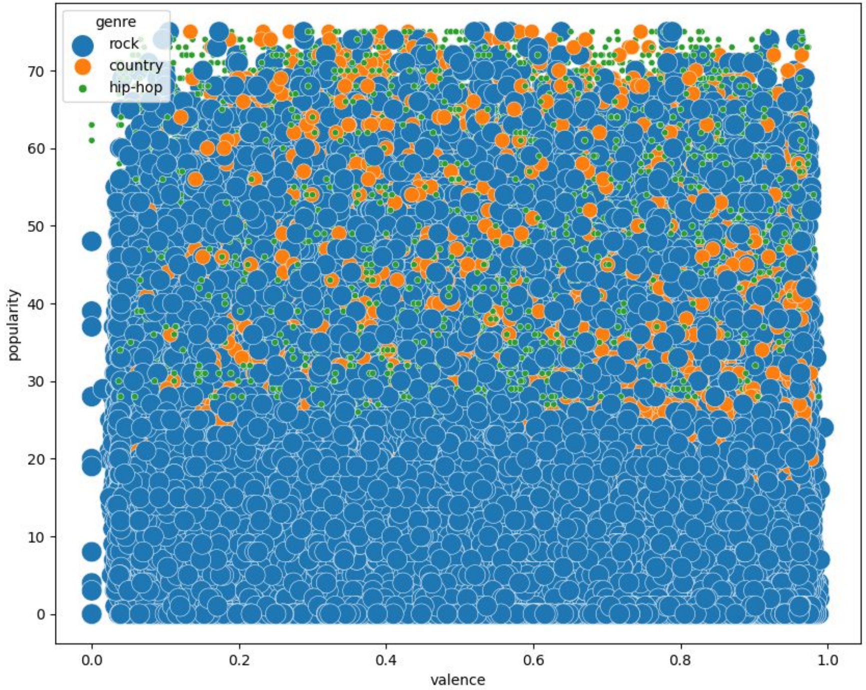
Data Correlations (post regrouping)

There still exists a lack of strong correlation between every feature.



Popularity vs Speechines and Valence (post regrouping)





Feature Extraction and Model Selection

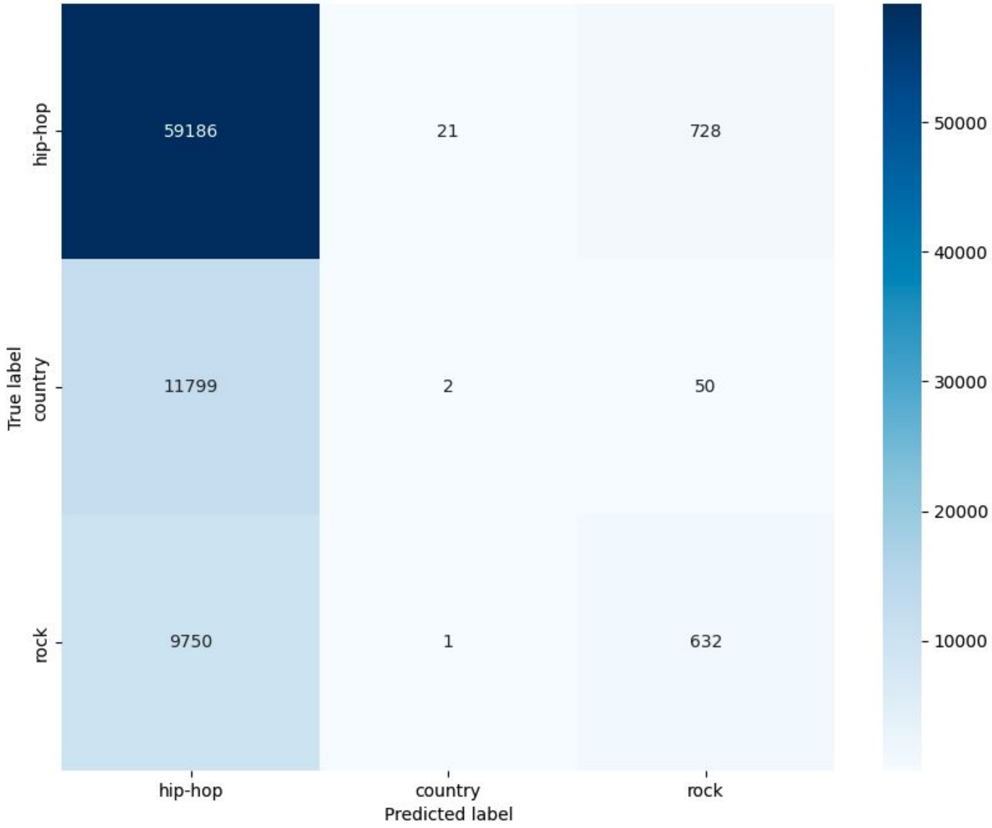
- Columns were vectorized and scaled and then ran through PCA
- We used a custom train/test split that tried to balance the selected data across the top three genres based on popularity
- Logistic regression was used on the training set
- Hyperparameter tuning was performed via cross validation using a parameter grid

Model Evaluation/Visualization

- Initial Accuracy was greatly impacted by the number of genres in dataset
- First attempt at limiting genres showed a significant bump in accuracy ~5% to ~19%
- We decided to reduce the genre labels by two processes
    - Group like genres together in a “base” genre - improved accuracy to 17.68%
    - After reducing the genres to the top 20, 10, 5, and 3 in terms of popularity accuracy improved with each iteration.

Model Evaluation/Visualization

- Initial accuracy after fitting train data was: 72.8%
- After hypertuning and cross validation accuracy was: 69.8%



Limitations, Future Work, and Conclusion

- Limitations
    - Unbalanced Dataset for genres
        - Bias toward Rock
    - Limited Features
        - Lyrics?
- Future Work
    - Incorporating Additional Features
    - Regression Model to predict the Song Popularity
    - Streaming data directly into PySpark
    - Balancing data with over-sampling
- Conclusion
    - Metrics can be used to distinguish genres and sub-genres
    - Bias towards rock

Results After Undersampling to Balance the Data		
model run	genres	accuracy
1	20	27.12%
2	10	26.86%
3	5	35.93%
4	3	56.66%

QA