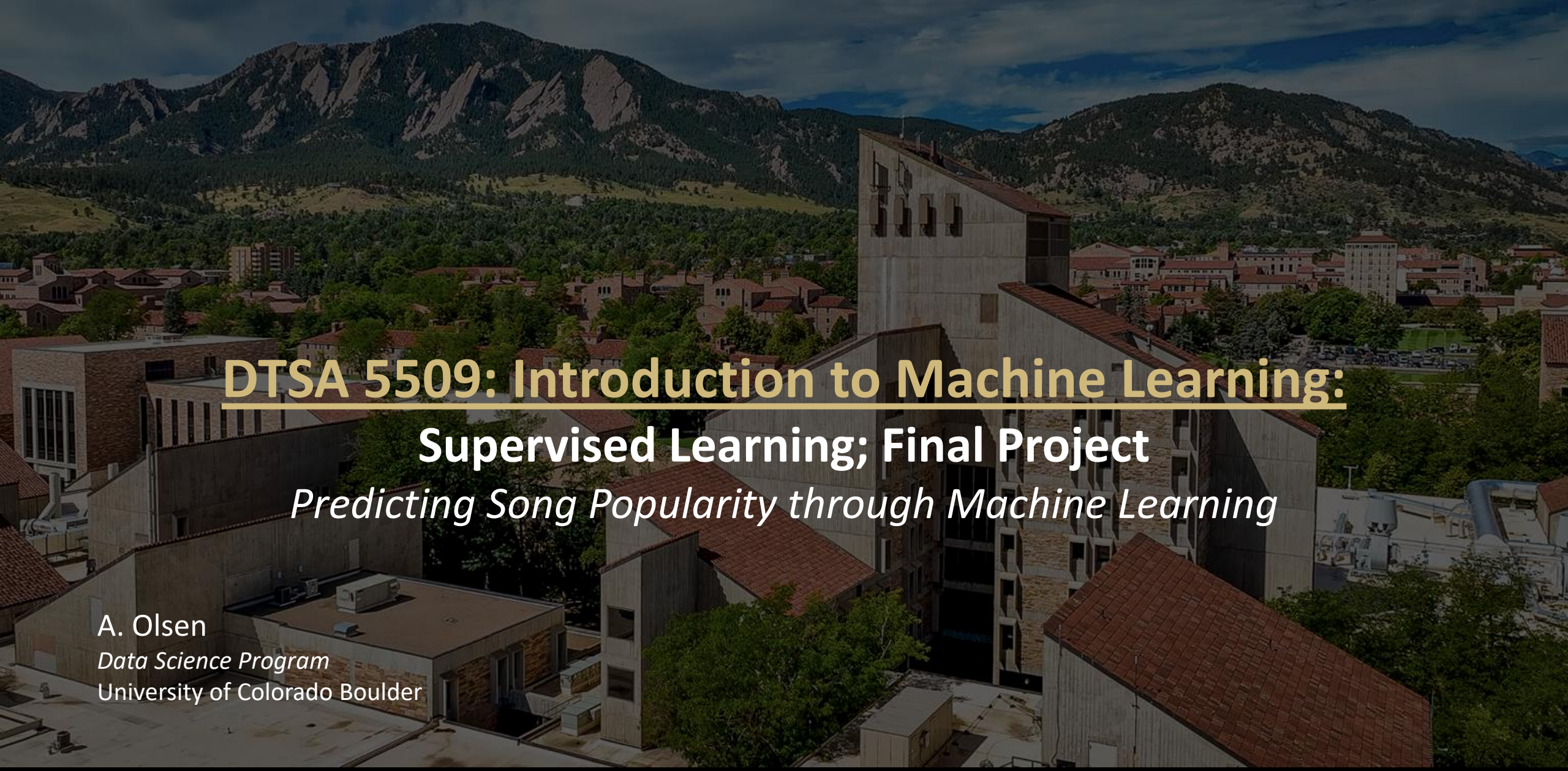


DTSA 5509: Introduction to Machine Learning:

Supervised Learning; Final Project

Predicting Song Popularity through Machine Learning

A. Olsen
Data Science Program
University of Colorado Boulder

College of Engineering & Applied Science
UNIVERSITY OF COLORADO **BOULDER**

Motivation and Goal

- Why predict ***popularity***?
  - Popularity is a highly variable attribute with both personal and commercial value.
    - *Playlist recommendation, listener engagement, industry analytics.*

- **Goal:**
  - Develop a predictive model of a song's popularity score using its audio attributes.

Dataset Overview

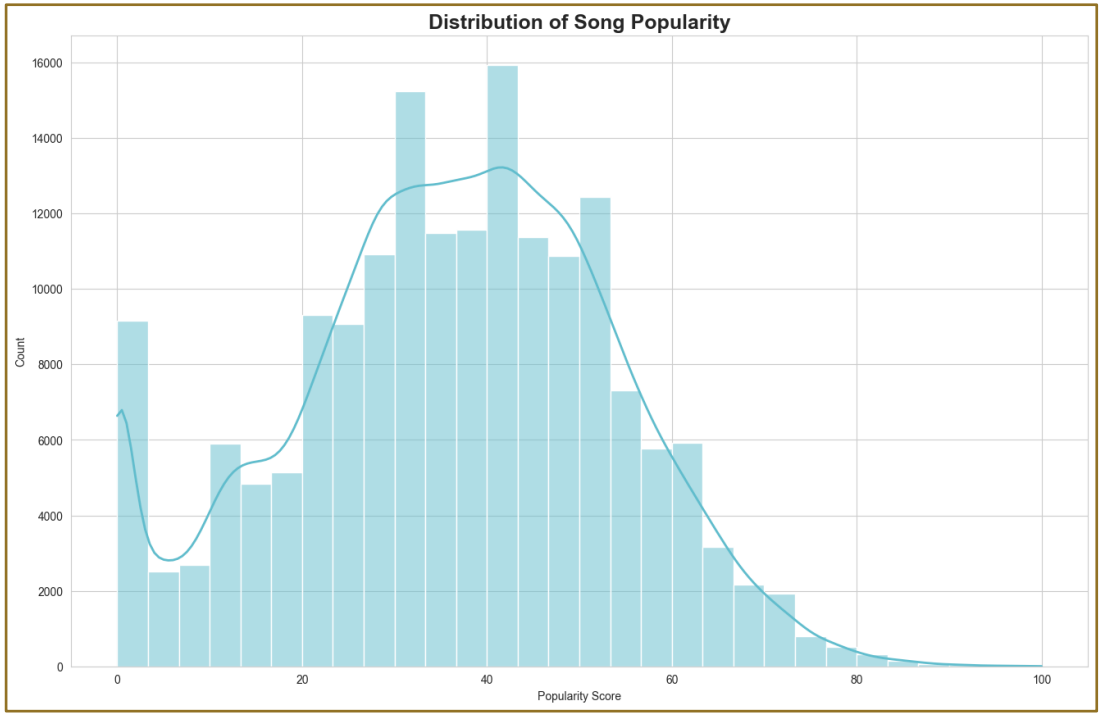
- "*Spotify Tracks Dataset*" from Kaggle,
- ~176,000 tracks (shape),
  - 18 features (columns) + target variable (***popularity***),
- Audio features (danceability, energy, valence, etc.),
- Structural features (key, mode, time_signature)

popularity	seed_key	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
0	70856	0.509	0.766	15387	0.5380	0.000001	-1	0.131	-15.830	1	0.8800	92.912	0.00	0.000
0	92679	0.969	0.562	15509	0.1250	0.001380	2	0.373	-15.449	2	0.2010	115.827	5.00	0.574
0	85280	0.914	0.588	16316	0.1650	0.000000	4	0.233	-26.286	1	0.2370	104.678	1.00	0.000
18	71033	0.922	0.422	16640	0.3210	0.000003	3	0.179	-15.381	1	0.3040	176.961	5.00	0.361
0	85373	0.954	0.532	16748	0.0639	0.000000	-6	0.593	-25.800	1	0.0519	126.447	0.25	0.000

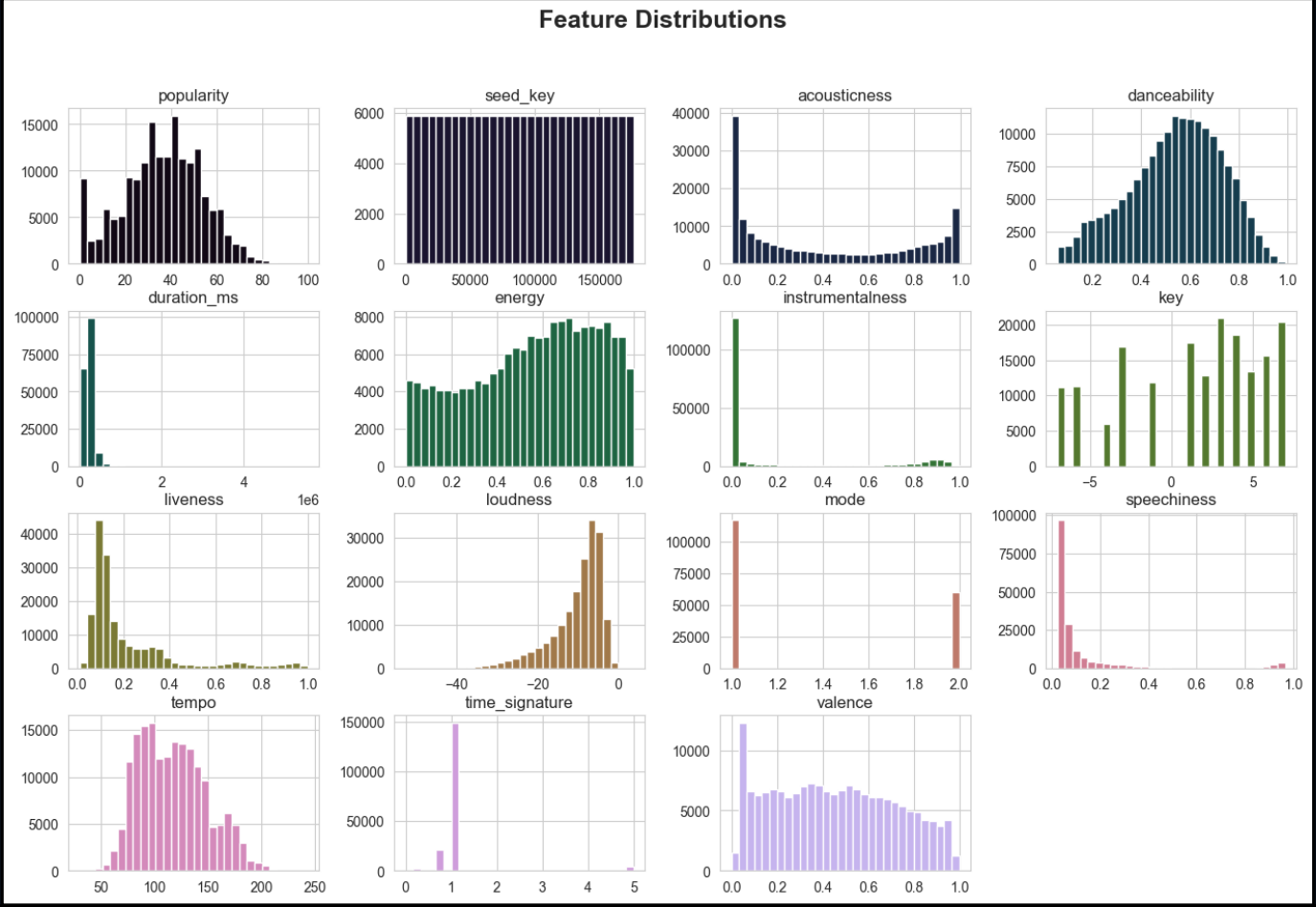
```
[143]: df.shape

[143]: (176561, 19)

[109]: df.columns

[109]: Index(['genre', 'artist_name', 'track_name', 'track_id', 'popularity',
       'seed_key', 'acousticness', 'danceability', 'duration_ms', 'energy',
       'instrumentalness', 'key', 'liveness', 'loudness', 'mode',
       'speechiness', 'tempo', 'time_signature', 'valence'],
      dtype='object')
```

Summary Statistics (EDA)

- Initial Exploratory Data Analysis (EDA) revealed:
  - Mean ***popularity*** = 36 (low overall)
  - Widespread in tempo, energy, valence,
  - **acousticness** and **instrumentalness** skewed.

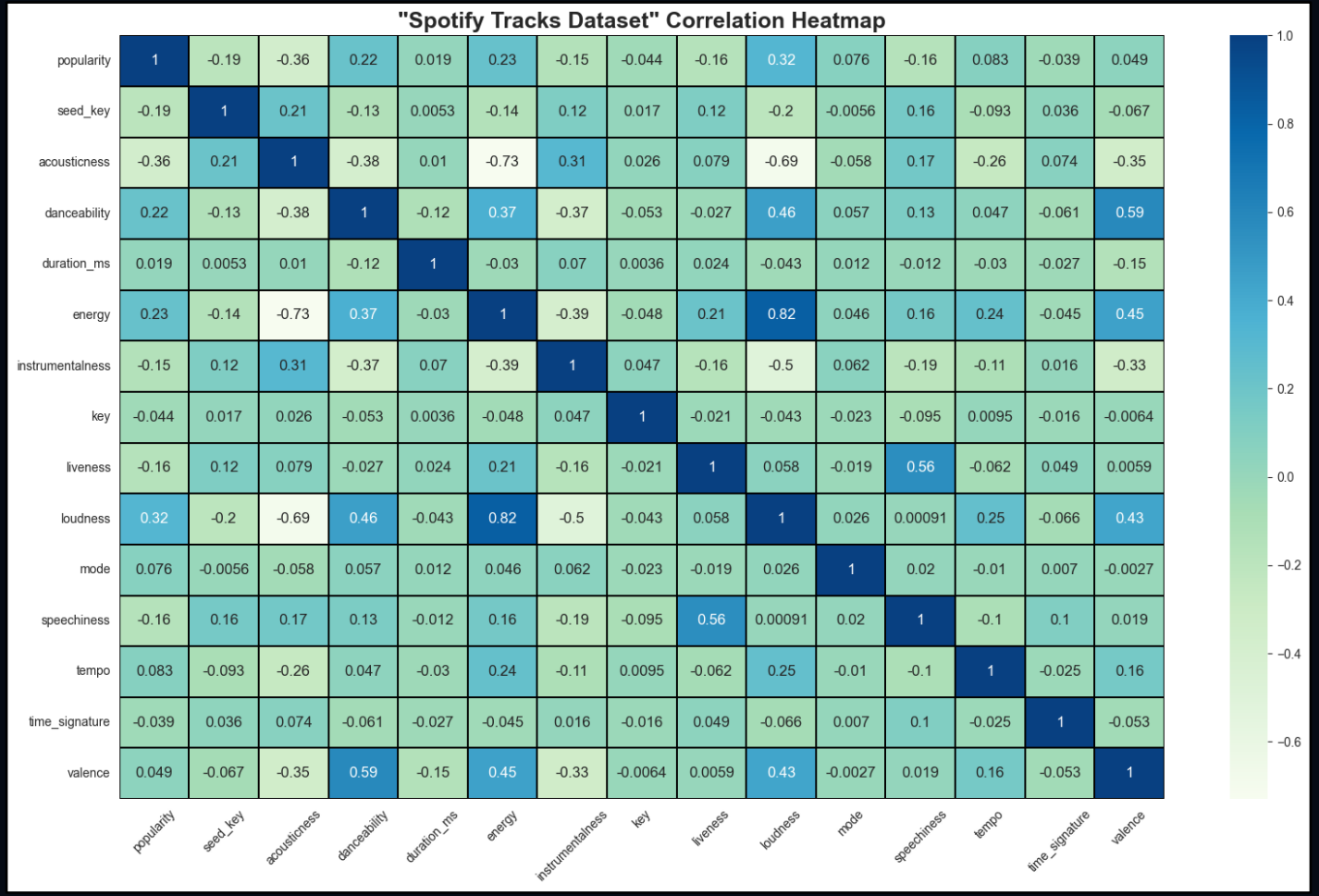


Summary Statistics (EDA) – (cont'd)



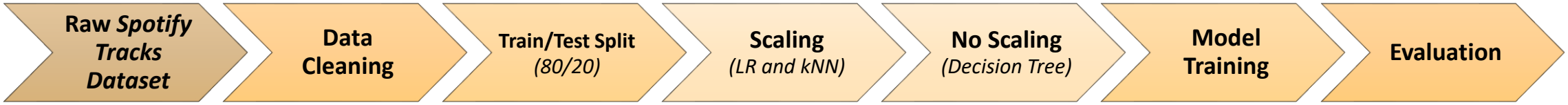
- Feature distribution:
  - **seed_key** – uniform distribution indicates identifier.
  - Features occur on very different scales. Highlight necessity for scaling.

Correlation Heatmap



Data Cleaning and Preparation

- Dropped unnecessary columns (**track_id, seed_key**, **track_name**, **artist_name**, and **genre**).
  - Categorical features that do not contribute to predictive power.
- Train/ test split (**80/20**),
- Used *StandardScaler* used for LR and kNN.



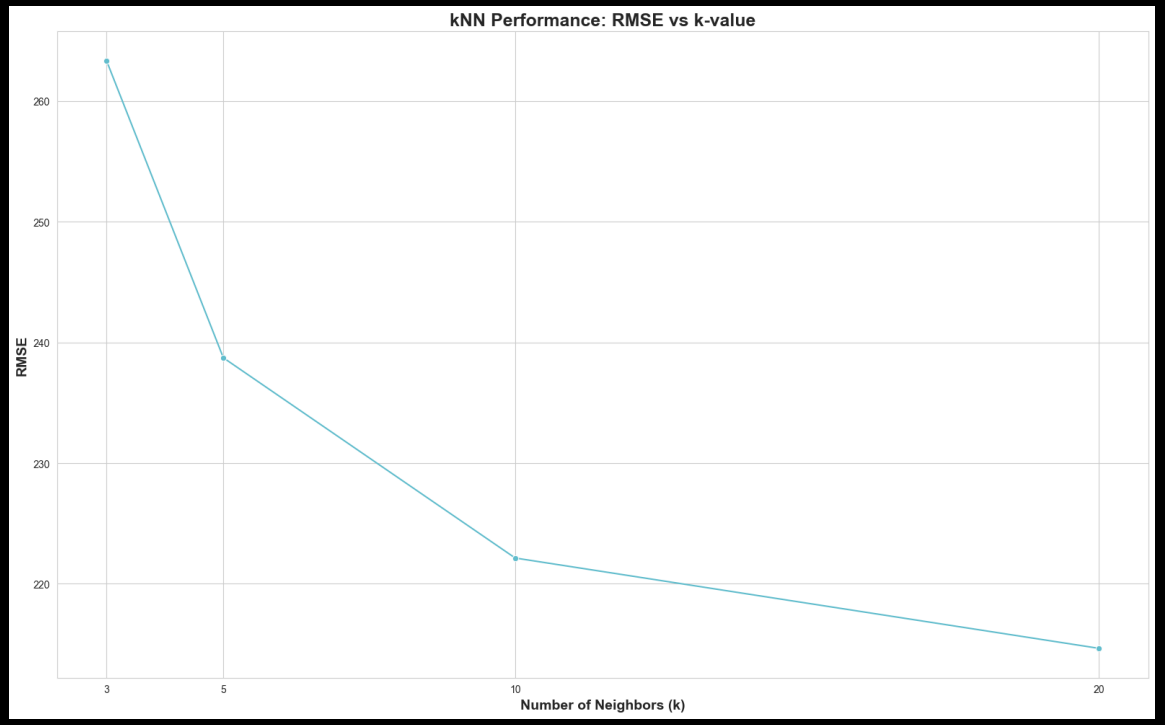
Modeling Overview

- **Linear Regression** – baseline, interpretable;
- **kNN Regression** – non-linear, distance-based;
- **Decision Tree** – hierarchical, handles interactions.

Linear Regression Results

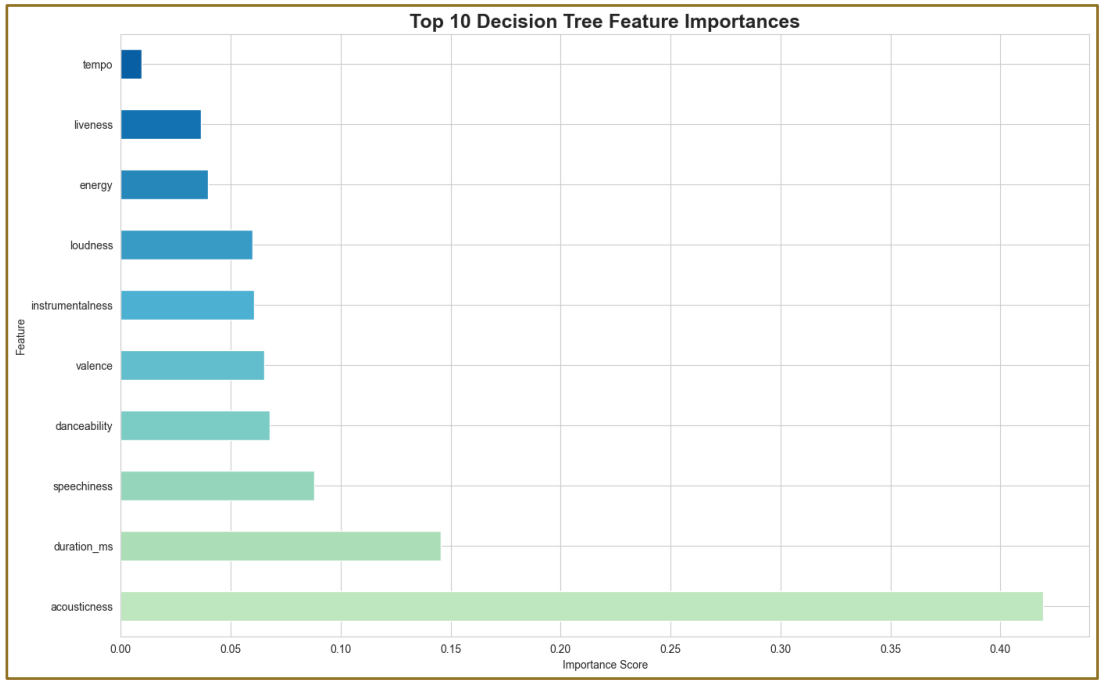
- Metrics:
  - *RMSE = 241.027,*
  - *MAE: 12.388,*
  - *R-squared: 0.207*;
- Model not able to capture non-linear interaction between audio features.
  - The model underfits because ***popularity*** has a non-linear structure that linear regression could not learn.

K-Nearest Neighbors (kNN) Regression Results



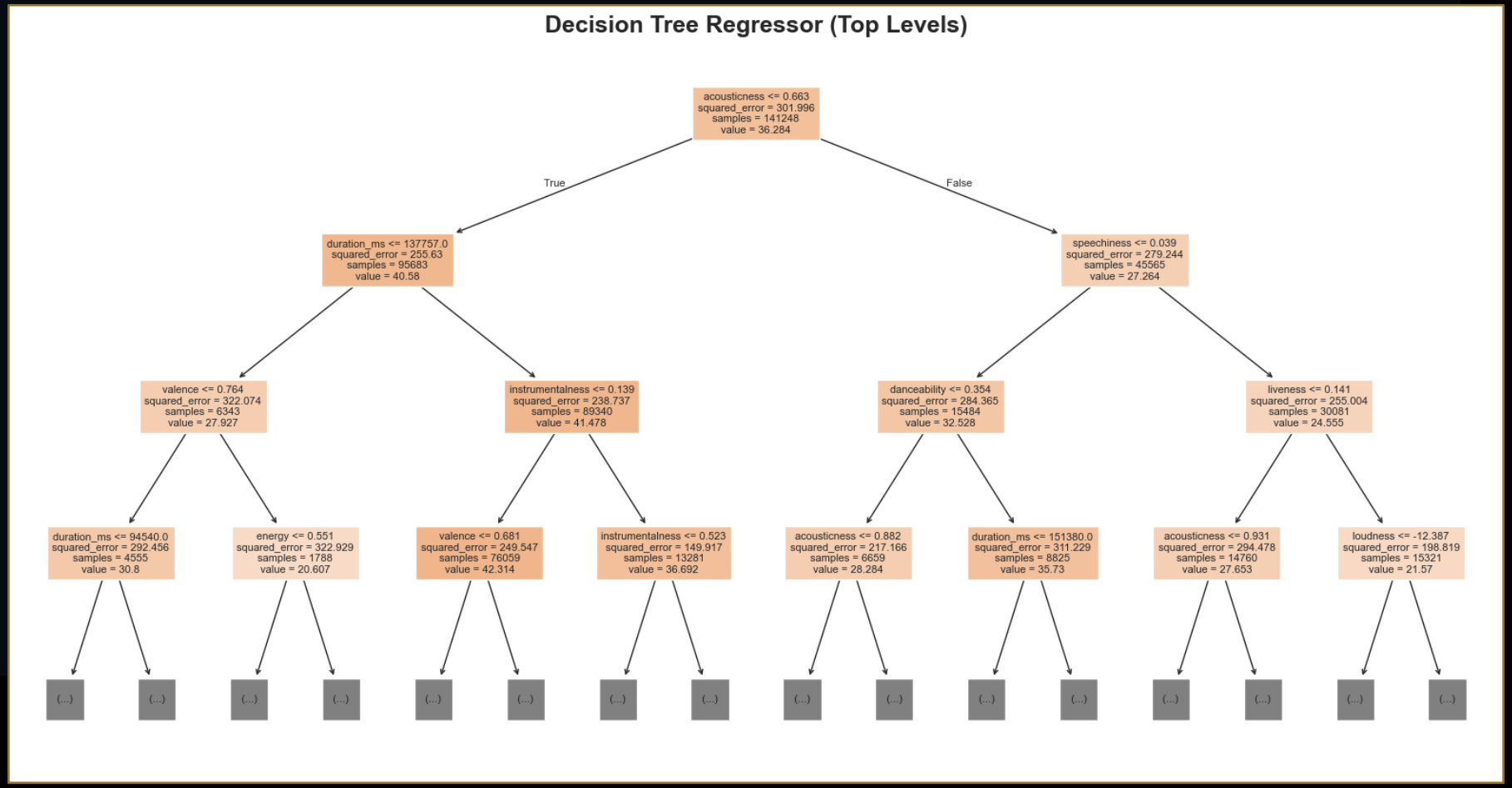
- Tested multiple k-values,
  - *k = 3, 5, 10, 20;*
- Stronger performance than Linear Regression;
- Sensitive to scaling;
- Non-linear structure fits this dataset better.

Decision Tree Results

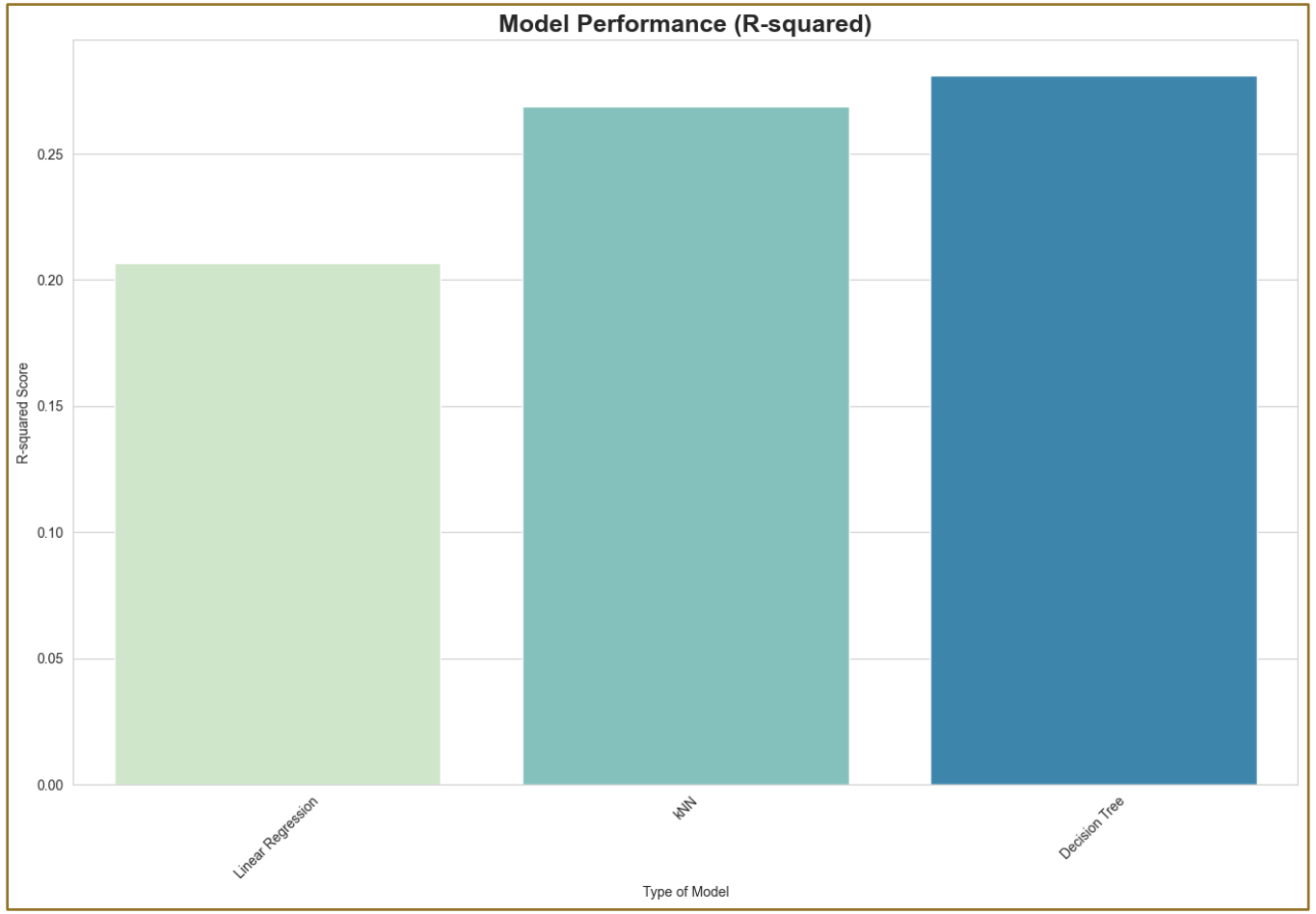


- Best-performing single model (*max_depth = 10*);
- Captures non-linear relationships;
- Lower RMSE, high R-squared than LR and/or kNN;
- Can overfit tuned depth of tree.

Decision Tree Results (cont'd)



Model Comparison



Key Findings and Conclusion

- Audio features *energy, loudness, acousticness, danceability,* and *tempo* matter most;
- Popularity is influenced by non-linear combinations of features;
- Machine learning can explain part of popularity, but not all;
- Models could benefit from additional metadata (***artist_popularity, release_date, playlist_placement,*** *or* ***artist_social_media_engagement***).

Conclusion: Non-Linear models moderately predict popularity, but many factors remain unmeasured in this dataset.

Conclusion

Slides and notebook posted:

<https://github.com/aolsen13/IntrotoMachineLearningFinalProject/tree/main>