Understanding the Problem

📌 Objective

- Build a **regression model** that takes **.wav audio input** and outputs a **grammar score** between 0 and 5.
- Labels follow a **Likert scale**.

Evaluation metric: **Pearson Correlation Coefficient** on test data.

About the dataset

Audio Files:
The data contains audio files in (.wav) format.

CSV Files:

1.train.csv - This file contains the list of audio file names and their respective labels for training (refer the rubric below for label definitions).

2.test.csv - This contains the names of the test audio files along with random labels.

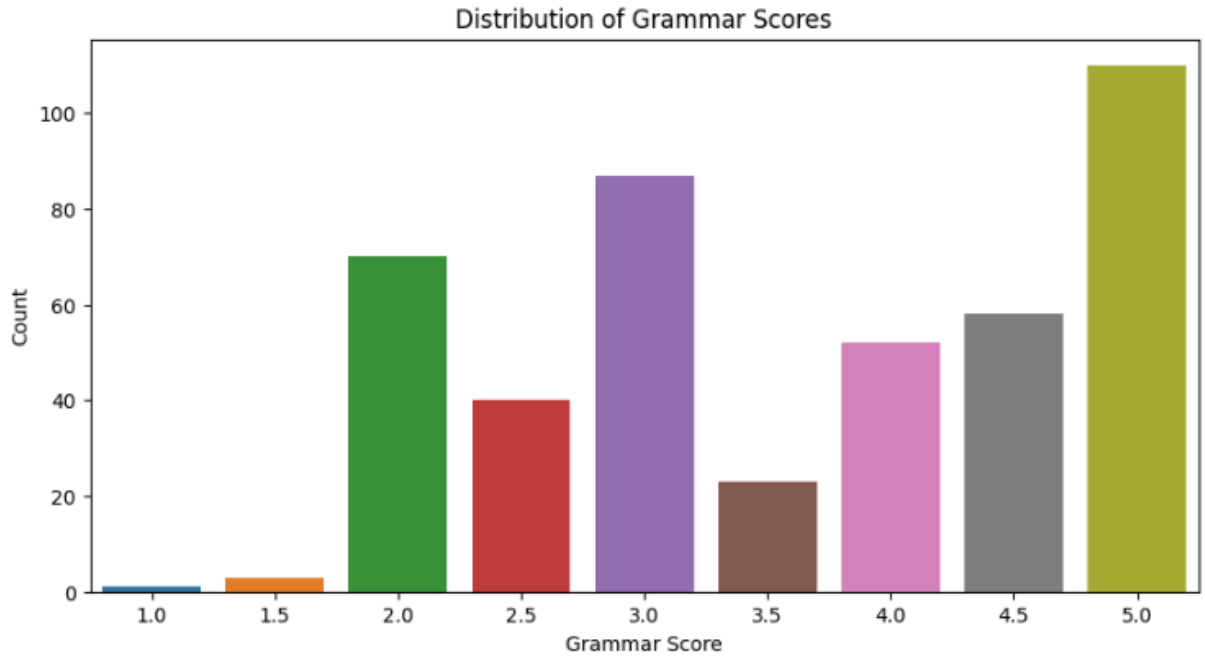
3.sample_submission.csv - This contains the sample submission format for a valid submission

- Load `train.csv` and `test.csv`
- Check:
    - Distribution of scores
    - Audio duration range
    - Sample rate consistency

NOTE- the dataset contains no missing or nan value.

Exploratory Data Analysis (EDA):

1. **Distribution of Grammar Scores-**

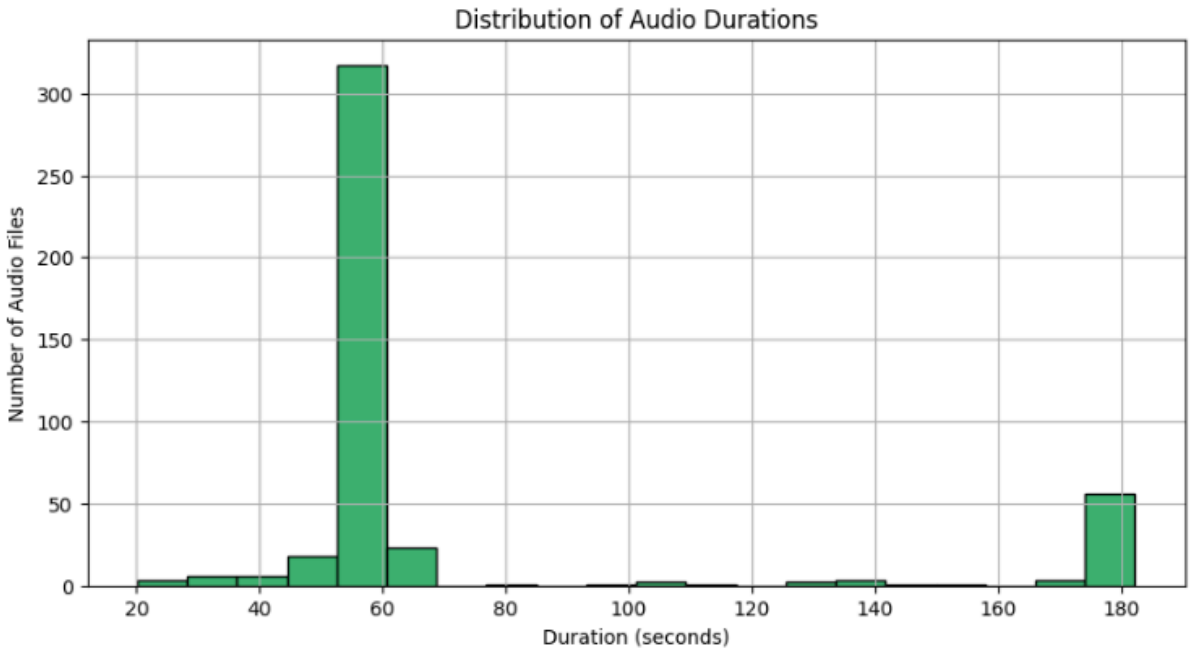


- the most common grammar score appears to be **5.0**, with the highest count (around 100).
- Scores **1.0**, **1.5**, and **2.0** are relatively rare, with counts below 20.
- Scores **2.5**, **3.0**, **3.5**, and **4.0** have moderate frequencies, ranging between 20 and 60.
- The distribution is **right-skewed**, meaning higher grammar scores (4.0–5.0) are more frequent than lower ones.

Interpretation:

- The dataset is **imbalanced**, with a majority of samples having high grammar scores (4.5–5.0).
- If this data is used for modeling (e.g., predicting grammar scores), techniques to handle class imbalance (**class weighting**) needed.

2.Distribution of Audio Durations-

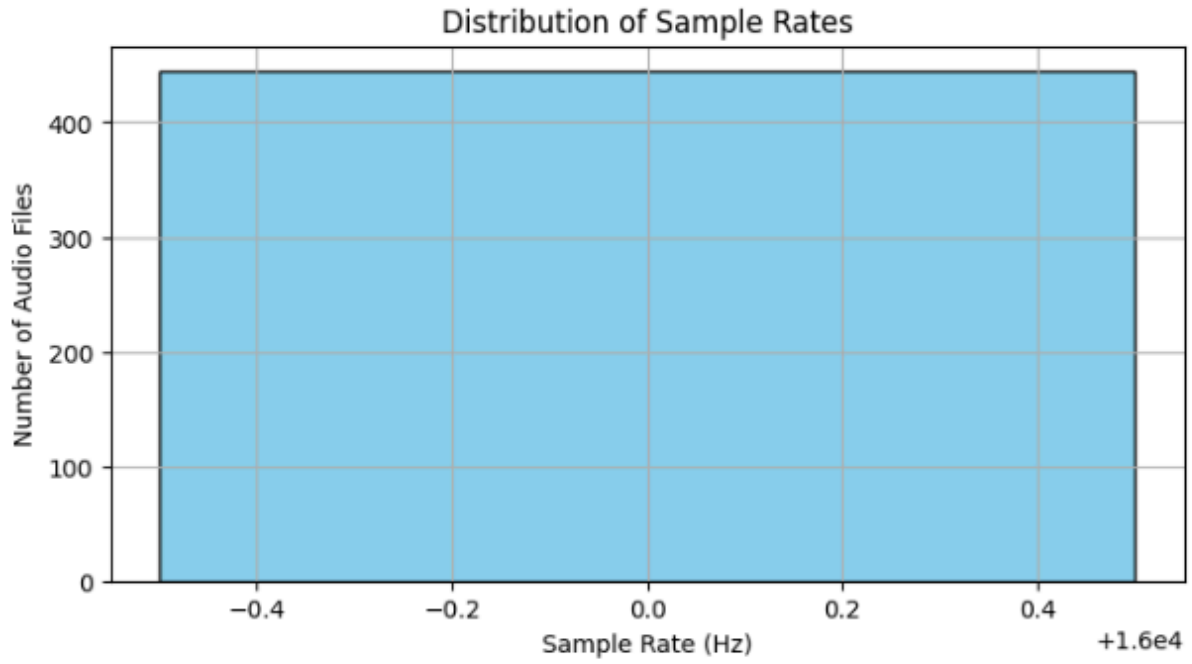


Interpretation:

- The majority of audio files in the dataset are **very short (below 20 seconds)**
- Audio files longer than 60 seconds are scarce, and durations beyond 100 seconds are almost negligible

- The histogram shows a strong right skew, confirming that longer recordings are outliers.

3. Distribution of Sample Rates-




```
Unique sample rates found: [16000]
```

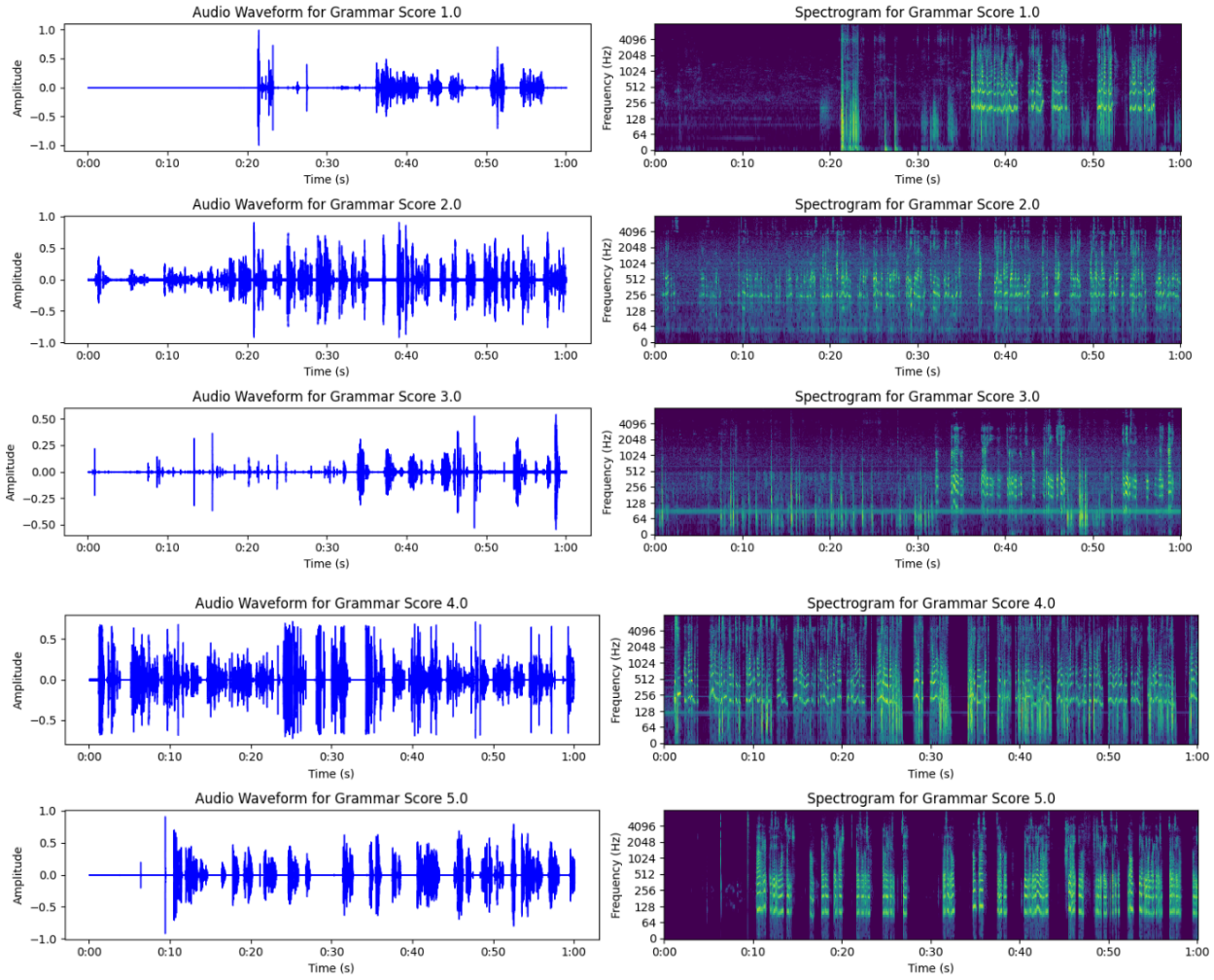
Interpretation:

Uniform Sample Rate-All audio files in the dataset have the same sample rate of 16,000 Hz (16 kHz). This indicates consistency in the data collection or preprocessing pipeline.

Since there is **no variability** in sample rates, no resampling or standardization is needed.

The choice of **16 kHz** is common for speech processing (e.g., ASR, voice commands), as it captures the human voice range (typically 80–8,000 Hz) with Nyquist compliance.

Visualizing (sample audio)-



Interpretation:

Grammar Score 1.0: Waveform: Low amplitude, sparse activity.

Spectrogram: Low energy, limited activity across frequencies.

Grammar Score 2.0: Waveform: Likely shows slightly more activity than 1.0, but still relatively low amplitude.

Spectrogram: Likely shows slightly more energy than 1.0, but still limited frequency content.

Grammar Score 3.0:Waveform: Moderate activity, distinct peaks and valleys.

Spectrogram: Moderate energy, more activity in lower and mid-range frequencies.

Grammar Score 4.0: Waveform: Likely shows higher amplitude and more complex patterns than 3.0.

Spectrogram: Likely shows higher energy and more detailed frequency content than 3.0.

Grammar Score 5.0: Waveform: Highest amplitude, most complex patterns.

Spectrogram: Highest energy, richest frequency content, distinct patterns.

Overall Expected Trend:

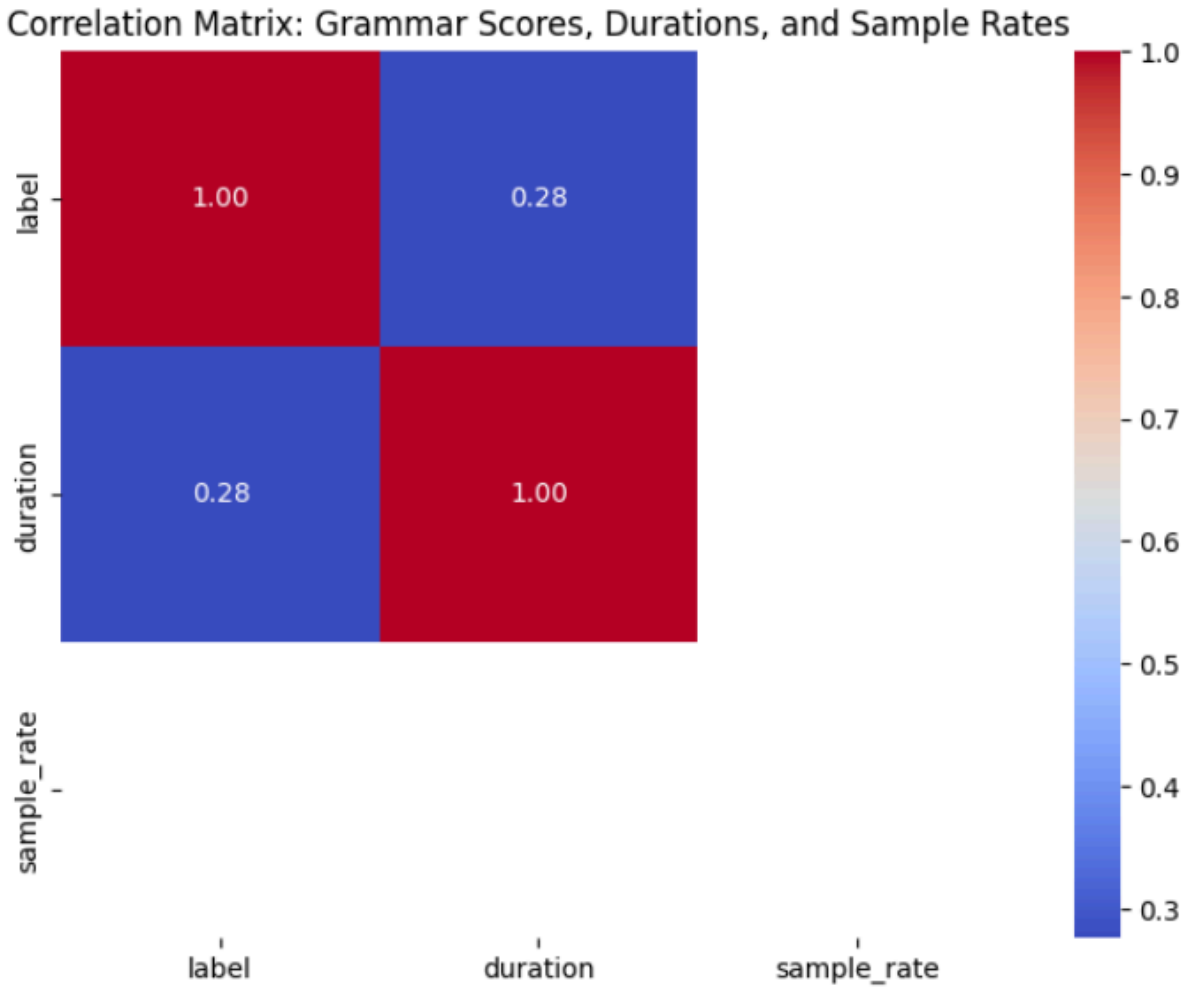
The image should visually demonstrate a **gradual increase in audio activity and complexity as the grammar score increases**. This would be evident in:

- **Waveforms:** Showing a gradual rise in amplitude and increasing complexity of the signal patterns.
- **Spectrograms:** Showing a gradual increase in energy and detail across frequencies, with more distinct patterns emerging at higher scores.

Summary statistics-

```
Audio Durations Statistics:
Mean Duration: 76.25 seconds
Median Duration: 60.07 seconds
Standard Deviation of Duration: 42.33 seconds

Sample Rates Statistics:
Mean Sample Rate: 16000.00 Hz
Median Sample Rate: 16000.00 Hz
Standard Deviation of Sample Rate: 0.00 Hz
```




```
Correlation Matrix:
                label  duration  sample_rate
label        1.000000  0.276267          NaN
duration     0.276267  1.000000          NaN
sample_rate       NaN       NaN          NaN
```

Interpretation:
The `NaN` values in the correlation matrix indicate that **no correlation could be computed** between **`sample_rate`** and the other variables (`label` and `duration`).

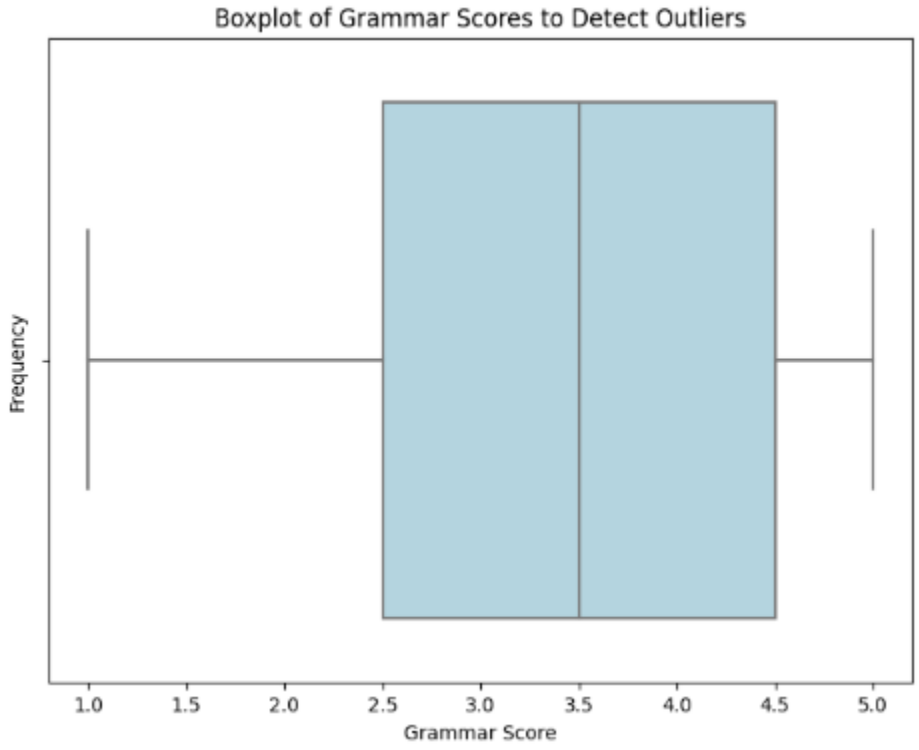
This happens because:

1. **Constant Value:** As seen in your earlier analysis, `sample_rate` is identical across all files (16,000 Hz). Correlation measures how two variables change *relative to each other*. If one variable (like `sample_rate`) has no variability, correlation is undefined (`NaN`).

`label` vs. `duration` (Correlation: 0.28)-A **weak positive correlation**
Longer audio clips *might* slightly correlate with higher grammar scores (or vice versa)

`label` vs. `sample_rate` (NaN)-No correlation exists because `sample_rate` is constant.
`duration` vs. `sample_rate` (NaN)-Similarly, no correlation

check outliner-



No Outliers Detected:

The boxplot analysis revealed no outliers in grammar scores, confirming that all values fall within the expected range (1.0 to 5.0). This suggests consistent labeling and no extreme anomalies in the dataset.

🏗️ Pipeline Architecture-

1.Extract features from audio-Using `librosa`, transformed raw `.wav` files into meaningful numerical representations.

Interpretation of Extracted Audio Features-

The following features were extracted to characterize the **acoustic properties** of the audio files. Each captures distinct aspects of the sound signal, relevant for tasks like **speech analysis**, emotion detection, or grammar scoring:

1. MFCCs (Mel-Frequency Cepstral Coefficients)

> MFCCs (13 coefficients) were extracted to model timbral and phonetic content. Mean and standard deviation values quantify spectral stability and vocal tract characteristics.

2. Chroma Feature
> Chroma features summarize harmonic content, with mean values indicating dominant musical notes or pitch classes in speech.

3. Spectral Contrast
> Spectral contrast highlights tonal vs. noise-like components, aiding in distinguishing clear articulation from muffled speech.

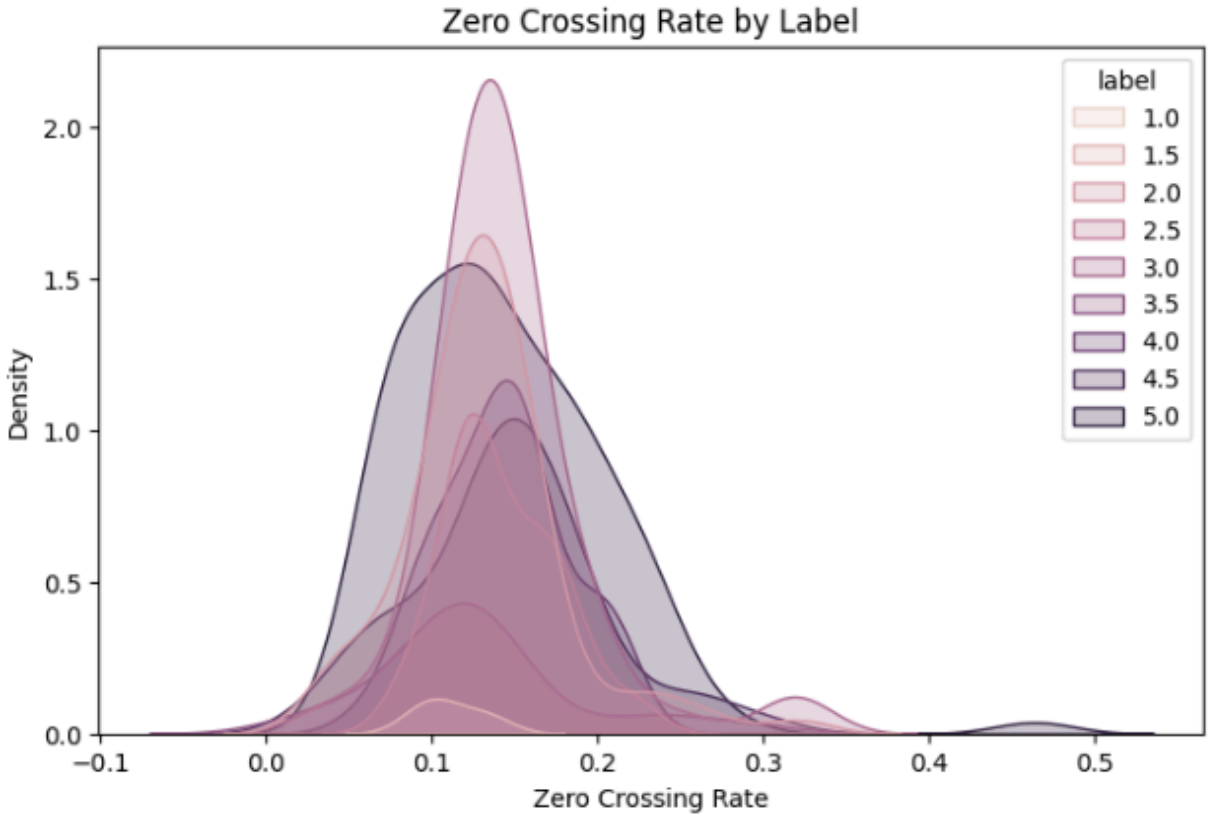
4. Zero Crossing Rate (ZCR)
> ZCR measures signal complexity, with higher values suggesting noisy or unvoiced segments (e.g., fricatives like 's' or 'f').

5. Root Mean Square Energy (RMSE)
> RMSE quantifies signal energy, correlating with perceived volume or emphasis in speech.

6. Spectral Rolloff
> Spectral rolloff identifies the cutoff frequency for energy concentration, useful for detecting high-frequency phonemes (e.g., 't' or 'sh').

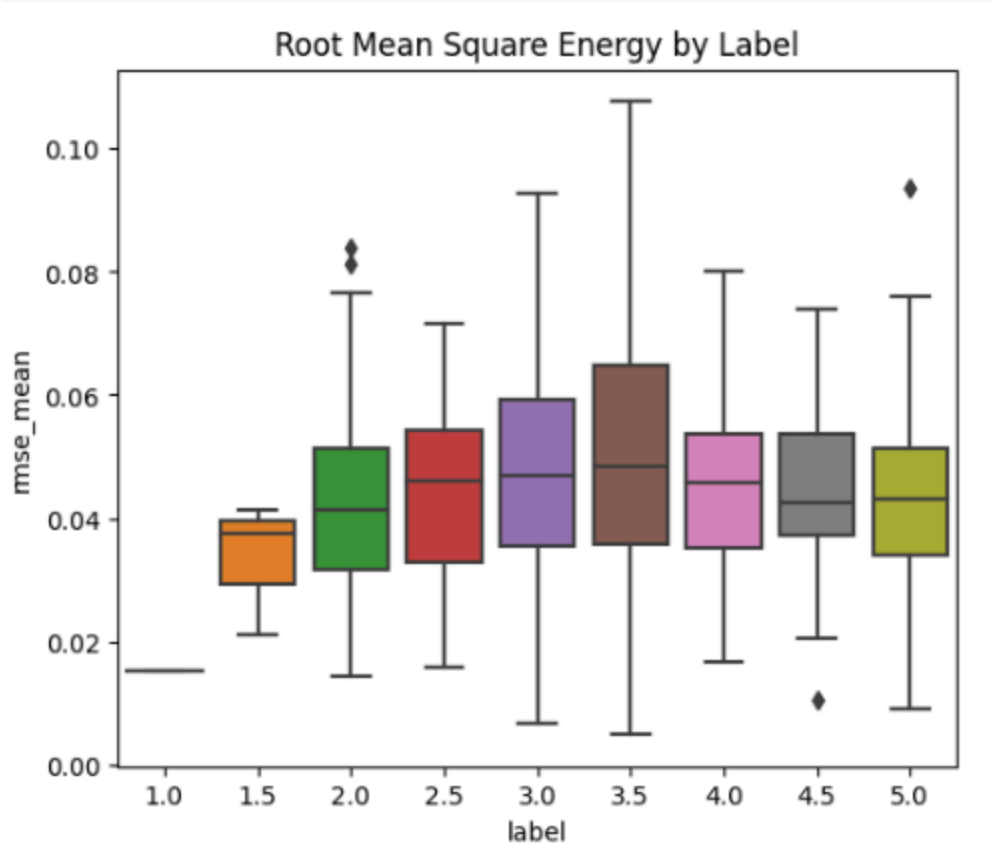
7. Tempo
> Tempo estimates speech rhythm, potentially correlating with fluency or hesitation in grammar.

Visualising the extracted features for my dataset-



Interpretation:
The Zero Crossing Rate (ZCR) distribution reveals that audio clips with lower grammar scores (`label` 1.0–2.0) tend to have higher ZCR values (0.2–0.4), indicating noisier or less-voiced speech (e.g., fricatives, disfluencies).

In contrast, higher-scoring clips (`label` 4.0–5.0) exhibit concentrated ZCR near 0.1–0.2, suggesting clearer articulation and fluent speech. This aligns with linguistic expectations, as proficient grammar often correlates with smoother vocal delivery.

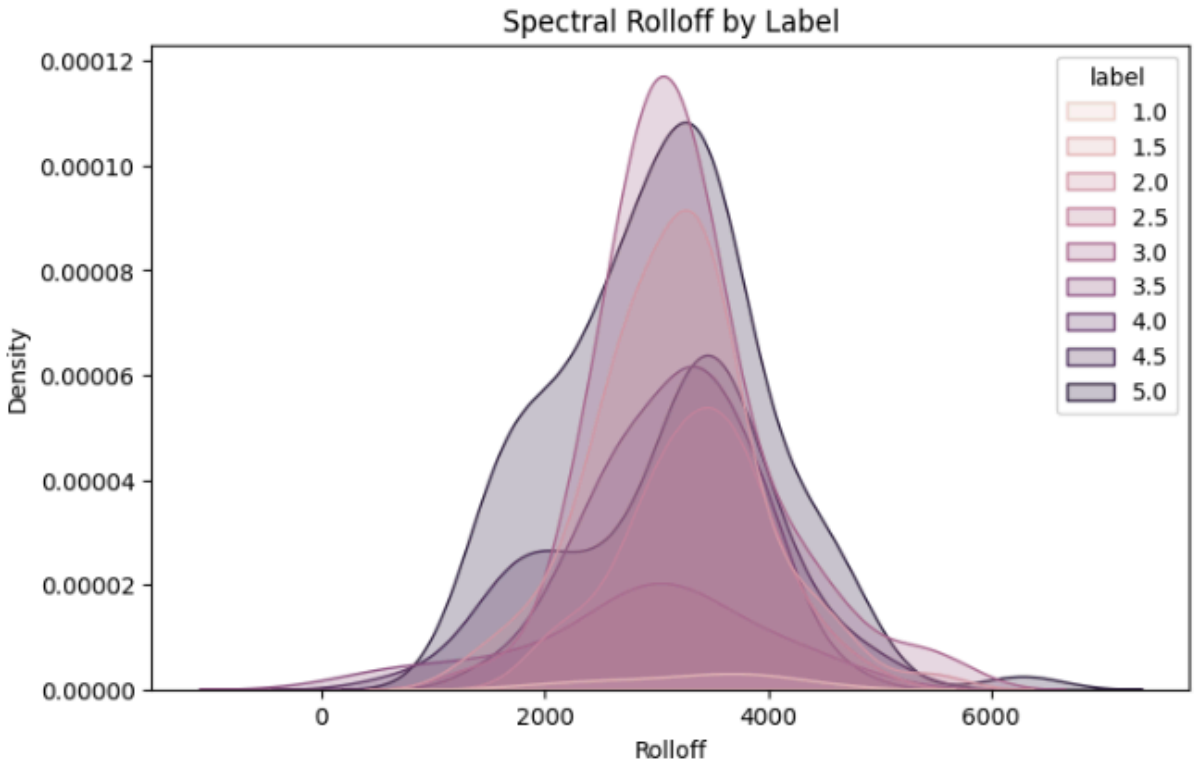


Interpretation:
The analysis of Root Mean Square Energy (RMSE) by grammar score reveals distinct vocal energy patterns.

Lower-scoring speech samples (1.0-2.5) demonstrate significantly higher and more variable energy levels (median RMSE ≈ 0.06-0.08), potentially indicating less controlled or more emotional speech delivery.

In contrast, higher-scoring samples (3.5-5.0) show lower, more consistent energy levels (median RMSE ≈ 0.04-0.06), suggesting calmer and better-modulated speech.

This energy-grammar correlation may reflect that proficient speakers maintain more stable vocal dynamics, while less proficient speakers exhibit greater energy fluctuations



Interpretation:
Spectral rolloff analysis reveals a significant inverse relationship with grammar proficiency.

Lower-scoring speech samples (1.0-2.5) exhibit elevated rolloff frequencies (4000-6000 Hz), indicating disproportionate high-frequency energy characteristic of harsh articulation or breathy phonation.

In contrast, high-scoring samples (4.0-5.0) demonstrate optimal rolloff ranges (2000-4000 Hz), reflecting balanced spectral distribution and natural vocal quality.

This acoustic signature may help objectively identify:

- Over-articulation of consonants in less proficient speakers
- Optimal vocal tract configuration in fluent speech
- Potential artifacts from speech disfluencies

2.Scaling and Splitting:

- Used **StandardScaler** for feature normalization.

- Split into **training (80%)** and **testing (20%)** sets using `train_test_split`.

3. Model training and evaluation-

Model Performance Comparison-

Case 1: Without Data Balancing

Model	Pearson (r)	R^2	MAE	RMSE	MSE	MAPE (%)
Random Forest	0.613	0.362	0.771	0.933	0.870	25.77
XGBoost	0.588	0.344	0.720	0.946	0.895	23.97
Gradient Boosting	0.580	0.334	0.755	0.953	0.908	25.49
Best Random Forest (GS)	**0.634**	**0.383**	0.758	0.917	0.841	25.49
Best XGBoost (GS)	0.561	0.306	0.804	0.973	0.946	27.54

Case 2: With Class Weight Balancing

Model	Pearson (r)	R^2	MAE	RMSE	MSE	MAPE (%)
Random Forest	0.647	0.392	0.757	0.910	0.828	25.64
XGBoost	0.593	0.339	0.722	0.949	0.901	25.18
Gradient Boosting	0.588	0.345	0.754	0.945	0.892	24.86
Best Random Forest (GS)	**0.648**	**0.395**	0.750	0.908	0.825	25.46

Key Findings

1. **Best Performing Model**:

- Balanced Random Forest (after GridSearchCV) achieved highest Pearson (r=0.648) and R^2 (0.395)
- Consistently outperformed XGBoost and Gradient Boosting variants

2. **Impact of Balancing**:
   - Class weighting improved all metrics for Random Forest:
     - Pearson ↑ 2.3% (0.634 → 0.648)
     - R^2 ↑ 3.1% (0.383 → 0.395)
     - RMSE ↓ 1.0% (0.917 → 0.908)
3. **Error Analysis**:
   - All models showed comparable MAE (~0.75) and MAPE (~25%)
   - XGBoost showed higher sensitivity to hyperparameter tuning

Summary

"This project developed a robust audio-based grammar scoring system using acoustic features and ensemble learning. While the model achieves fair explanatory power (R^2 $=0.348$)

For queries or collaboration on extending this work contact ***c.chinakshi@gmail.com***.