# Dataset Analysis Report: Average Podcast Listening Time

## Introduction

Podcast creators face a fundamental challenge in determining optimal episode lengths - balancing audience engagement with content depth (Jaffrey, 2024). In this project, the listening behavior patterns are analyzed, providing the creators with the ability to retain as many audience members as possible. Based on a dataset of 750 thousand podcast episodes, 11 features, we create a predictive model that estimates the duration of listening based on the characteristics of the content (Predict Podcast Listening Time, 2025).

## Data Overview

The data contains:

The dataset contains:

* 750,000 training samples and 250,000 test samples
* Features including episode length, genre, host/guest popularity, publication timing, ads, and sentiment
* Target variable: Actual listening time in minutes

Key statistics:

* Average episode length: 64.5 minutes (±32.97)
* Average listening time: 45.44 minutes (±27.14)
* 11.6% missing values in episode length
* 19.5% missing guest popularity data

## Methodology

### Data Preprocessing:

1. Filled missing values with medians.
2. Imputed categorical missing values with modes.
3. Encoded categorical variables (Genre, Publication Day/Time, Sentiment).

### Feature Engineering:

1. Analyzed correlation matrix.
2. Episode length showed strongest correlation with listening time (0.917).

### Model Selection:

1. Implemented XGBoost with 3-fold cross-validation.
2. Hyperparameters: 300 estimators, max depth 14, learning rate 0.042.

## Results

### The model attained:

1. **Cross-validated RMSE:** 12.90 (±0.009).
2. **Max RMSE:** 12.91.
3. **Min RMSE:** 12.89.

### Key findings:

1. The variation in listening time is attributed to the episode length (91.7 percent).
2. Every extra advertisement shortens the listening time by 11.8 percent.
3. Morning publications receive 5.1 percent lengthier listening than evening.
4. Neutral sentiment episodes keep listeners on 3.2 percent longer than positive.

### Insights for Podcasters:

1. **Optimal Length:** The optimum sweet spot lies between 45-65 minutes in order to retain listeners
2. **Ad Placement:** Keep the ads to 1-2 per episode to reduce drop-off rates.
3. **Publication Timing:** Releases made in the morning (typically Monday / Wednesday) have the highest performance.
4. **Content Strategy:** Highest completion rate is done through Educational and Technology genres.

# Technical Implementation

**The parallel implementation computing:**

1. Applied Python and its in-built multi-threading multi-threading with XGBoost.
2. Has acquired 3.2x acceleration using 4-core processor.
3. Inserted 1M records in 8.7 minutes.

**Limitations:**

1. There is no demographic listener data in the Dataset.
2. No statistics in the quality of episode content.
3. We also restricted the number of podcasts to English.

## Conclusion

Based on this analysis, the study offers practical, data-guided advice to the makers of podcasts on how to streamline the length and format of the episode. The high correlation between the length of an episode and listening time indicates the importance of keeping the content to a minimum and being as deep as possible. Future solutions may be the integration of real-time listener analytics to dynamically adjust the length.

## References

Jaffrey, A. (2024, January 20). Podcast Length Secrets to making your episodes just Right. MelodyLoops Blog. <https://www.melodyloops.com/blog/ideal-podcast-length/>

Predict Podcast Listening Time. (2025). @Kaggle. <https://www.kaggle.com/competitions/playground-series-s5e4/data?select=test.csv>