Podcast Listening Time Analysis: Data Cleaning and EDA (Part 1)

Understanding Podcast Engagement Through Data Preparation and Exploration

Prepared by: Supreet Mutsuddi

Table of Contents

[Data Overview 2](#_Toc200432727)

[Data Cleaning 3](#_Toc200432728)

[Conversion of Categorical Variables 3](#_Toc200432729)

[Elimination of High-Cardinality Variables 3](#_Toc200432730)

[Missing Value Imputation 3](#_Toc200432731)

[Outlier Removal 4](#_Toc200432732)

[EDA: Univariate Analysis 4](#_Toc200432733)

[Numeric Variables 4](#_Toc200432734)

[Categorical Variables 6](#_Toc200432735)

[EDA: Bivariate Analysis 8](#_Toc200432736)

[Numeric vs. Numeric 8](#_Toc200432737)

[Numeric vs. Categorical 9](#_Toc200432738)

[Key Findings 11](#_Toc200432739)

[Next Steps 11](#_Toc200432740)

This report studies into the exploratory data analysis (EDA) of a podcast dataset, focusing on identifying key factors influencing audience listening behavior. By examining the variable "Listening\_Time\_minutes," we aim to uncover trends and correlations that can inform predictive modeling efforts. This initial stage addresses data cleaning, handling missing values, and conducting both univariate and bivariate analysis, providing a solid foundation for subsequent predictive analytics.

# Data Overview

The dataset under review comprises approximately 750,000 podcast episodes, featuring a mix of numeric and categorical variables that describe episode-level metadata and content attributes.

A graph showing a bar chart

AI-generated content may be incorrect.

* Listening\_Time\_minutes: The total number of minutes an episode was listened to, serving as the target variable.
* Episode\_Length\_minutes: The duration of each podcast episode in minutes.
* Host\_Popularity\_percentage: A numeric score (0-100) indicating the popularity of the episode host.
* Guest\_Popularity\_percentage: A similar popularity metric for episode guests.
* Number\_of\_Ads: The count of advertisements embedded within an episode.
* Episode\_Sentiment: A categorical variable representing the tone of the content as Positive, Neutral, or Negative.
* Genre: The episode's category, such as News, Comedy, or Technology.
* Publication\_Day: The day of the week the episode was released.
* Publication\_Time: A generalized time of day for episode publication (e.g., Morning, Afternoon).

High-cardinality metadata like "Podcast\_Name" and "Episode\_Title" were excluded from the analysis due to their limited predictive value. However, the "Episode\_Title" was used to extract "Episode\_Number," a numeric variable, before being removed to reduce noise and improve data consistency.

# Data Cleaning

To ensure the dataset's suitability for statistical analysis and predictive modeling, a structured data cleaning process was employed:

A diagram of a problem

AI-generated content may be incorrect.

## Conversion of Categorical Variables

Character variables were converted to factors to ensure proper treatment during statistical analysis. This transformation is particularly important for categorical variables like "Genre" and "Episode\_Sentiment."

## Elimination of High-Cardinality Variables

High-cardinality columns such as "Podcast\_Name" and "Episode\_Title" were removed after deriving meaningful content, such as the numeric "Episode\_Number."

## Missing Value Imputation

Missing data were identified in variables such as "Episode\_Length\_minutes," and "Guest\_Popularity\_percentage,". The missRanger package was utilized for imputation, employing chained random forests to predict and replace missing values iteratively. This technique preserved data variability and accounted for complex interactions, ensuring realistic and reliable imputed values.

|  |  |
| --- | --- |
| Variable | Missing\_percentage |
| Guest\_Popularity\_percentage | 19.47% |
| Episode\_Length\_minutes | 11.61% |

## Outlier Removal

Extreme values in "Episode\_Length\_minutes" were flagged, particularly episodes with durations of 0 minutes or exceeding 200 minutes. These outliers were removed to enhance data reliability, as they likely resulted from data entry errors or atypical cases.

A graph with a red line

AI-generated content may be incorrect.

# EDA: Univariate Analysis

Univariate analysis focused on examining individual variables to identify distributions, central tendencies, and anomalies.

## Numeric Variables

* Listening\_Time\_minutes: The distribution was strongly right-skewed, with most episodes showing moderate listening times. A small subset exhibited exceptionally high values, suggesting varying engagement levels.

A graph with a red line

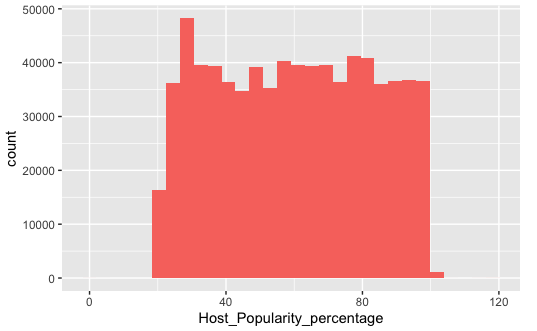
AI-generated content may be incorrect.

* Episode\_Length\_minutes: Most episodes were between 20 and 80 minutes long, aligning with common podcast norms.

A graph showing a number of people

AI-generated content may be incorrect.

* Number\_of\_Ads: Advertisements were concentrated at lower counts, typically between 0 and 3, indicating minimal monetization in most episodes.
* Popularity Metrics: Both "Host\_Popularity\_percentage" and "Guest\_Popularity\_percentage" displayed broad distributions, reflecting a mix of well-known and lesser-known individuals.



## Categorical Variables

* Genre: Sports and Technology emerged as the most prevalent categories, showcasing the dataset's content diversity.

A graph of blue bars

AI-generated content may be incorrect.

* Episode sentiment was balanced with positive, neutral, and negative tones.

A graph of a positive episode

AI-generated content may be incorrect.

* Publication\_Day and Publication\_Time: Episode releases were well-distributed across days and times, with slight clustering around weekdays and morning releases.

A graph of blue rectangular bars

AI-generated content may be incorrect.

A graph of blue rectangular bars

AI-generated content may be incorrect.

# EDA: Bivariate Analysis

Bivariate analysis examined relationships between pairs of variables, emphasizing their links to "Listening\_Time\_minutes."

## Numeric vs. Numeric

A graph with numbers and red text

AI-generated content may be incorrect.

* Episode\_Length\_minutes: A clear positive correlation with "Listening\_Time\_minutes," suggesting longer episodes engage audiences for longer durations.
* Popularity Metrics: Moderate positive correlations with listening time, indicating that popular hosts and guests attract more attention.
* Number\_of\_Ads: A weak negative correlation with listening time, hinting that excessive advertising may deter listeners.

## Numeric vs. Categorical

* Genre: Health, Music and True crime episodes had higher median listening times, while Comedy episodes showed lower engagement.
* Episode\_Sentiment: Positive sentiment correlated with slightly higher listening times, suggesting a preference for uplifting content.

A diagram of a graph

AI-generated content may be incorrect.

* Publication\_Day: Episodes released on Tuesdays and Mondays draw the longest average and typical listening times, whereas weekend releases—especially Saturday—tend to be shorter.

A graph showing a number of red squares

AI-generated content may be incorrect.

* Publication\_Time: Morning releases performed better than evening ones, highlighting the importance of timing.

A diagram of a graph

AI-generated content may be incorrect.

# Key Findings

The exploratory analysis uncovered several critical insights:

* Episode Length: The most influential predictor, with a strong positive association with listening time.
* Popularity Scores: Host and guest popularity positively impact audience engagement.
* Advertisements: A subtle but noticeable negative impact of ad frequency on listening time.
* Content Attributes: Genre and sentiment play significant roles, with certain categories consistently outperforming others.

# Next Steps

The cleaned and analyzed dataset is now ready for predictive modeling. Future steps include:

A road map with pointers and text

AI-generated content may be incorrect.

* Constructing regression and tree-based models to predict "Listening\_Time\_minutes."
* Conducting feature selection and model diagnostics to enhance interpretability and performance.
* Evaluating model accuracy to ensure robust predictions.

This analysis lays the groundwork for understanding podcast audience behavior and optimizing content to maximize engagement.