Analysis and Cleaning of Audible Audiobook Dataset

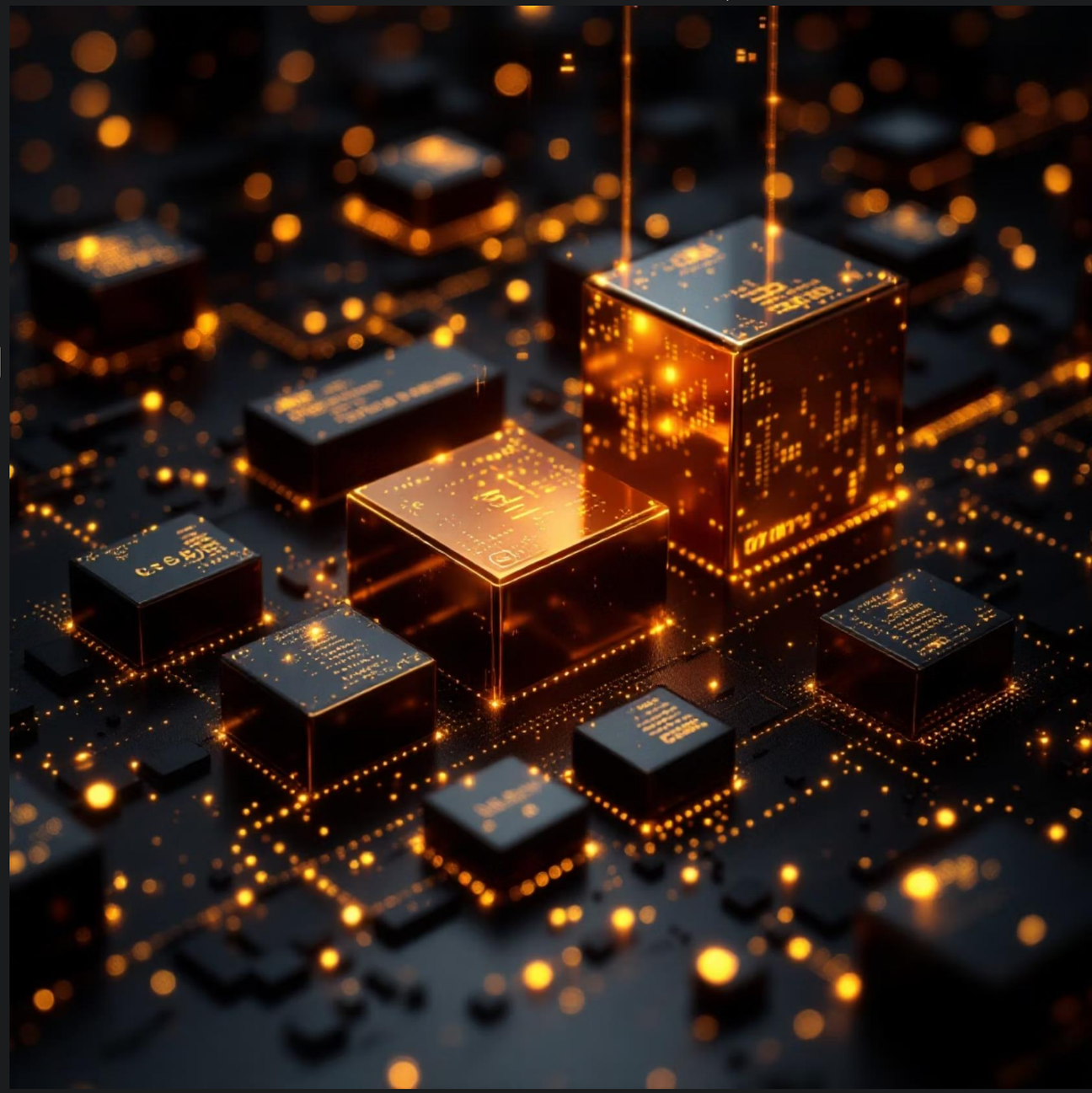
Transforming raw, unstructured audiobook data into actionable market insights through advanced data preprocessing and exploratory analysis.

Project Performed by;
Prakash Chawda

Project Overview

This project focuses on the end-to-end pipeline of analyzing and cleaning an unstructured dataset from **Audible**.

The raw data contains inconsistent formats and non-numeric fields. Our goal is to extract meaningful trends in pricing, ratings, and content characteristics.



The Dataset: 87,489 Records

A comprehensive look at the raw data structure before processing.

Core Info

Name, Author, and Narrator (often containing messy prefixes like "Writtenby:").

Metrics

Duration (text format), Ratings (mixed with review counts), and Price.

Metadata

Release date and Language distribution across 8 total columns.

The Problem Statement

Raw data is rarely ready for analysis. We identified several critical roadblocks:

Prefixed Names

Text prefixes in name fields

Non-numeric Durations

Durations stored as text

Mixed Rating Columns

Ratings and reviews combined

Text-based Dates

Dates in freeform text

These inconsistencies make it impossible to perform accurate statistical analysis or visualization without significant preprocessing.

Project Objectives

Our roadmap to turning "messy" data into "smart" data.

Understand

Map the structure of the raw Audible dataset.

Clean

Preprocess and remove noise from text fields.

Convert

Transform text into usable numeric formats.

Data Cleaning & Preprocessing

The core technical phase of the project.

→ Text Stripping

Removed "Writtenby:" and "Narratedby:" prefixes from names.

→ Numeric Conversion

Converted duration strings into total numeric hours.

→ Feature Extraction

Separated star ratings from the total number of reviews.

→ Validation

Handled missing values and ensured price columns are numeric.

PRAKASH CHAWDA

Tools & Technologies

Leveraging industry-standard Microsoft tools for robust data ha

Excel

The primary platform for data cleaning and analysis.

Power Query

Used for high-performance data transformation and cleaning workflows.

Performance of project (In video format)

PLAY NOW

Video Link Below (Tap to watch)

https://drive.google.com/drive/folders/1PacM_tjSSUXx0hq8Zo5EzojVsVYTxjL5?usp=drive_link

Mastering Data Cleaning with Power Query

Transforming raw Audible data into a polished, analysis-ready dataset using the advanced capabilities of Excel's Power Query Editor.

1. Text Standardization

Title Casing

Standardize the **Name** column to ensure consistent title casing across all audiobook entries.

Author Separation

Identify combined names in the **Author** column and separate them into distinct fields for multiple authors.

PRAKASH CHAWDA

2. Temporal Data Integrity

Ensuring time and date fields are recognized as functional data types rather than static text.

Date Format

Convert **releasedate** to a strict DD-MM-YYYY format.

Duration Conversion

Transform the **time** column from text to Excel-recognized duration.

3. Financial & Numeric Precision

Price Normalization

Ensure the **Price** column is strictly numeric. We must identify and resolve any non-numeric values that could break calculations.

PRAKASH CHAWDA

4. Rating & Narrator Logic

Stars to Numbers

Convert text-based ratings in the **stars** column into functional numeric values for averaging.

Narrator Splitting

Split the **narratedby** column into multiple columns when multiple narrators are listed.

The Power Query Advantage

Why we use these specific transformations for the Audible dataset.

100%

Consistency

Eliminating manual entry errors through automated rules.

2

Decimal Precision

Standardizing all currency values to two decimal places.

0

Data Loss

Ensuring no records are dropped during the conversion process.

PRAKASH CHAWDA

Final Quality Checklist

01

Validate Numeric Formats

Check Price and Stars for non-numeric artifacts.

02

Verify Date Logic

Confirm all dates follow the DD-MM-YYYY standard.

03

Review Column Splits

Ensure Authors and Narrators are correctly distributed.

PRAKASH CHAWDA

Ready for Analysis

Project Completion

By following these steps in **Power Query**, the Audible dataset is now standardized, consistent, and ready for insightful data visualization.

Conclusion

Data is only as good as its cleanliness.

This project highlights the critical role of preprocessing in the data science lifecycle. By transforming messy, real-world data into a structured format, we unlock the ability to make informed decisions in the competitive audiobook market.

Prakash Chawda (Thanks For Attention)