Amazon Audible Audiobook Dataset Analysis

This project focuses on cleaning, standardizing, and analyzing Amazon Audible audiobook 87,151 data using Power Query in Excel. The objective is to improve data quality and extract actionable insights related to pricing, ratings, language distribution, authorship, and content availability.

Presented By :
Prakash Chawda

CHALLENGE

The Growth vs. Quality Dilemma

Amazon Audible has aggressively expanded its audiobook catalogue. However, leadership lacks quantitative clarity on critical questions:

- Does catalog growth improve overall quality?
- Is new content priced consistently?
- Do high-priced titles deliver proportional customer value?

The Risk:
Unoptimized growth leads to higher operational costs with diminishing quality returns.

Analysis Objectives

1

Data Standardization

Standardize text, date, time, and numeric fields using Power Query for consistency.

2

Format Conversion

Convert ratings and prices into analyzable numeric formats for deeper insights.

3

Trend Analysis

Identify pricing trends across ratings and analyze language and authorship distribution.

4

Dataset Preparation

Prepare a clean, analysis-ready dataset for comprehensive reporting and decision-making.

Dataset Overview

87K

Total Audiobooks

Comprehensive catalog analyzed

₹3,499

Max Price

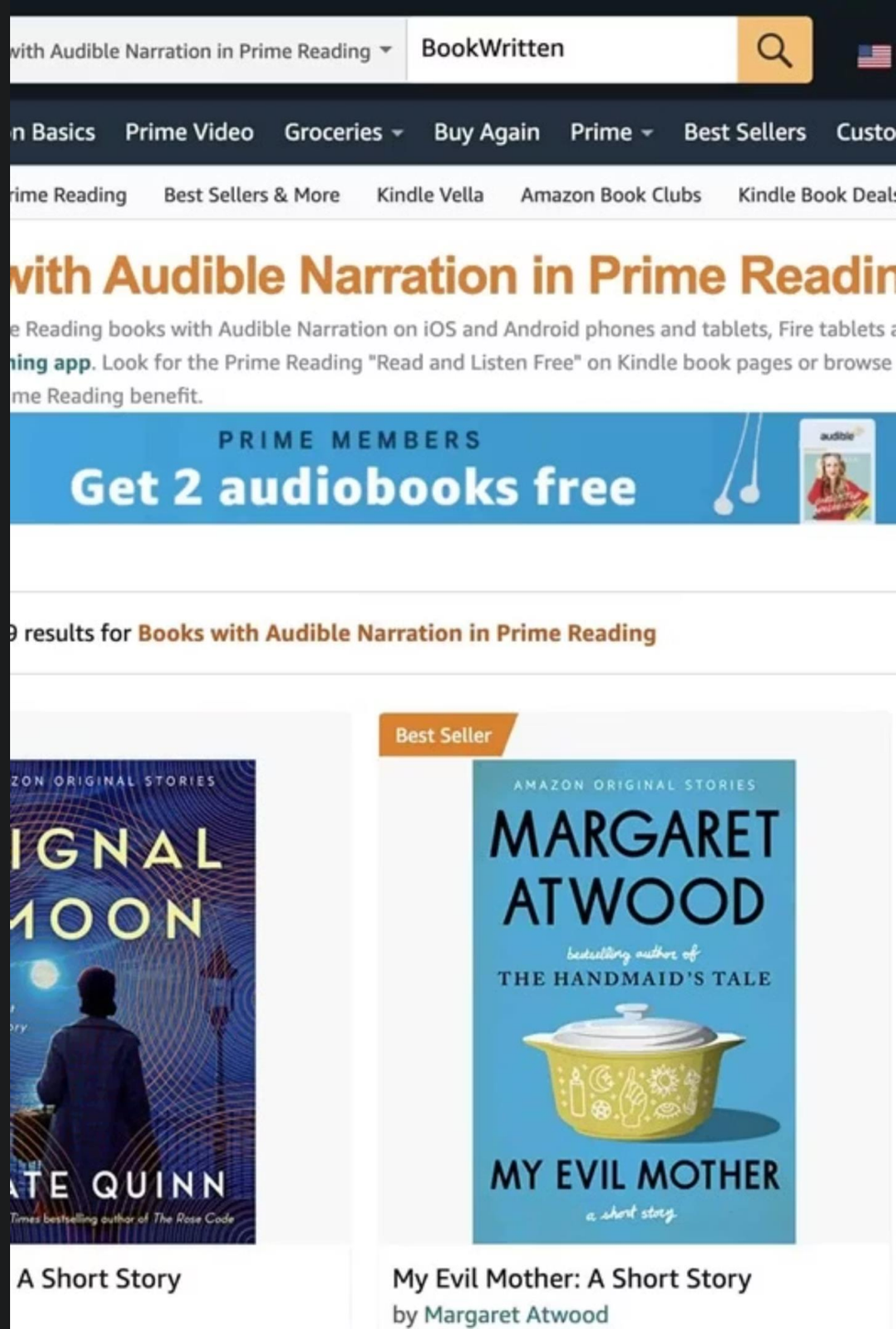
Price range: ₹0 – ₹3,499

5.0

Rating Scale

Star ratings from 1.0 to 5.0

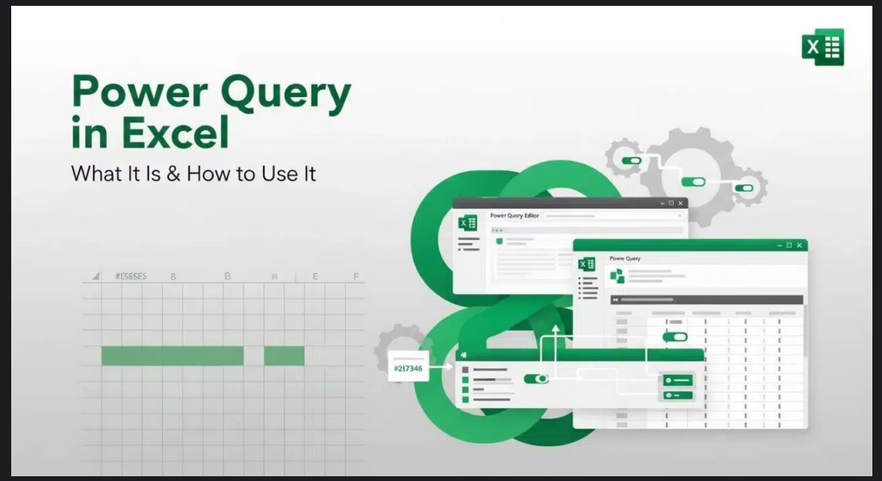
Key data columns include: Name, Author(s), Narrator(s), Release Date, Language, Rating, Price, and Duration.



Technical Approach

Tools

Microsoft Excel with Power Query for data transformation and analysis.



Core Techniques

- Data cleaning and transformation
- Column splitting and standardization
- Data type conversion
- Merging columns for analysis

Power Query Formulas

Title Standardization

```
Text.Proper([Name])
```

Converts titles to proper case format

Date Formatting

```
Date.FromText([releasedate])
```

Standardizes dates to DD-MM-YYYY

Duration Conversion

```
Duration.FromText([Updated time])
```

Transforms text to time format

Price Conversion

```
Number.From([price])
```

Converts price to numeric values

Rating Conversion

```
Number.From([Star rating out of 5])
```

Transforms ratings to analyzable numbers

Methodology

Import

Load raw Audible dataset into Power Query

Standardize

Convert titles to proper case format

Split

Separate author and narrator fields

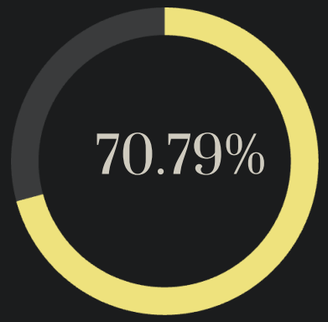
Format

Convert dates and duration values

Convert

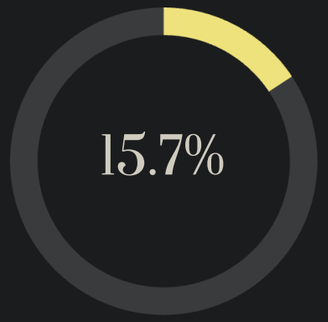
Transform price and rating to numeric

Key Insights Discovered



English Dominance

English audiobooks lead the catalogue



Multi-Author Titles

13,707 collaborative audiobooks

Top-Rated Content

- **6,701** titles with 5-star ratings
- **7,064** titles with 4-star ratings

Analysis of 87,151 audiobooks reveals strong quality concentration in top-rated content and significant English language dominance in the catalog.

Project Performance showcase (In Video)

GitHub.com

Linkedin.com

Also Available
Linkedin
GitHub
Google Drive
Click on Link &
given site

http://drive.google.com/drive/folders/1PacM_tjSSUXx0hq8Zo5EzojVsVYTxjL5

What We Accomplished

Data Standardization

Delivered fully cleaned dataset with consistent text, date, time, and numeric columns across the entire catalog.

Analysis Ready

Enabled accurate pricing and rating-based analysis for strategic decision-making.

Reusable Source

Created scalable Power Query workflow for future dataset processing and ingestion.

Limitations

Missing Revenue Data

No sales or revenue information available to measure actual performance and market impact.

Limited Customer Insights

Lack of demographics and listening behavior data restricts audience understanding.

Rating Context Gaps

Ratings volume doesn't indicate recency or depth of customer engagement over time.

Future Opportunities

Revenue Integration

Connect sales and subscription data for comprehensive revenue analysis.

Duration Analysis

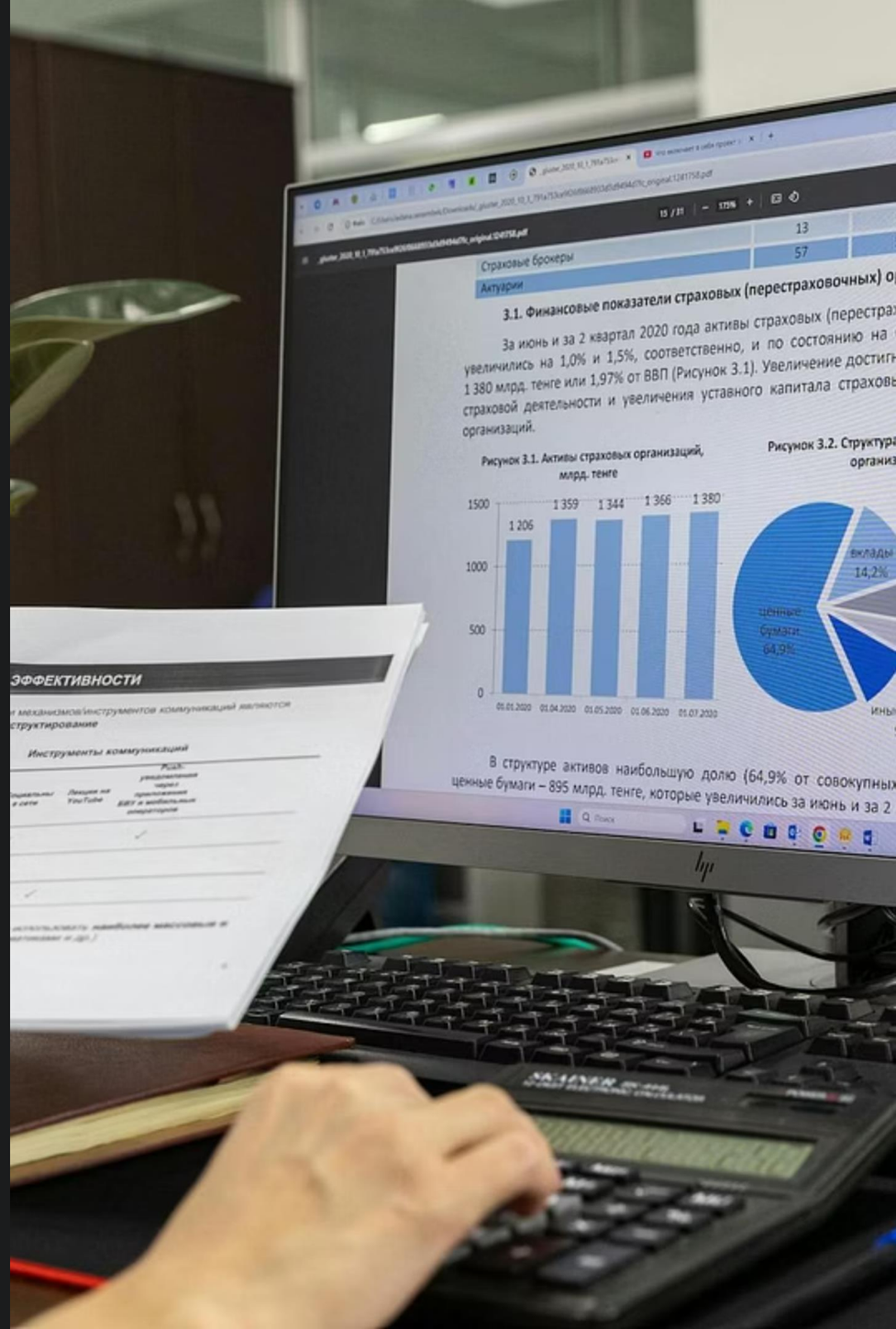
Examine correlation between audiobook length and customer ratings.

Power BI Dashboards

Build real-time monitoring tools for catalog performance tracking.

NLP Classification

Apply Natural Language Processing for automated genre categorization.



Strategic Recommendations

Premium Pricing Strategy

Position highly rated audiobooks at ₹700+ price range to maximize revenue from quality content.

Content Diversification

Expand non-English audiobook offerings to reach broader audiences and diversify the catalog.

Collaboration Focus

Promote multi-author collaborations leveraging their strong presence in the current catalog.

Source Standardization

Implement standardized data sources for consistent ongoing data ingestion and quality.

Project Success

This project successfully transformed raw Amazon Audible data into a clean, analysis-ready dataset using Power Query. By resolving data quality issues and extracting meaningful numerical insights, the analysis provides a **reliable foundation** for pricing strategy, catalog optimization, and future analytical initiatives.