1. **Dataset Introduction:**
   1. Source of the data

I am choosing the Audible Dataset because I have read hundreds of books using Kindle, however, I have never bought Audible even though I see on social media how popular it has become and there are many people in my life that live on Audible. So this dataset most closely relates to my interests.

* 1. Size (number of tables and approximate records)

1 table; 87490 rows

* 1. General subject matter (e.g., "retail sales data," "healthcare patient records")

Audible, an audiobook service, basic audiobook details

1. **Analysis Objectives:**
   1. Business questions being investigated
   2. Patterns or trends the analysis hopes to uncover
   3. Specific hypotheses being tested with the data
2. **Analytical Approach and Findings:**
   1. Brief overview of methods used
   2. Query explanations and insights
3. **Business Recommendations:**
   1. Strategic suggestions based on your complete analysis
   2. Actionable insights for stakeholders derived from your findings

* Identify potential data quality issues (missing values, duplicates, inconsistent formatting).
  + First and last names of both the Author and Narrator column need to be separated and spaces needed to be added after periods.
    - I chose to find and replace periods in the Author and Narrator column with a period and a space.
    - I used Proper to capitalize all of the languages
  + Price is not in a readable format
  + Search for potential duplicates
  + Search for missing values and decide whether to include or exclude them
  + Make sure all dates are set to the same format
* Document the data quality issues you'll need to address.
* Create a plan for cleaning each issue you've identified.

1. Check Spelling
2. Search for duplicates

There are no duplicate rows found, I will format names of the Author and Narrator and then recheck

1. Search for missing values

No blank cells are found

1. Make sure datatypes/formatting is correct
   1. I changed “releasedate” to “release\_date”
2. Change price to currency
3. Check dates are in the same format
4. Edit Author and Narrator column as a whole to capitalize and separate first and last names.

Once cleaned, I created a new database on SQL and uploaded the cleaned Audible dataset, all data was imported correctly and in the correct format.

Row count = 87490 , which matches the original dataset.

**Pivot Tables and Visualizations**

This pivot table visualization shows the number of audiobooks by the stars on Audible. I think this chart could be used to show either: how many audiobooks have yet to be read or how many audiobooks have been given a zero star rating. This could be used to highlight audiobooks that have yet to be given stars or if there is a audiobook that is being advertised but rated poorly it could be replaced with an audiobook that is yet to be reviewed.

This pivot table visualization shows the availability of languages of the audiobooks on Audible. I thought this chart was very telling and showed an inconsistency in the availability of languages. There is not an inclusive selection and they might could profit more if they had more audiobooks available in other languages besides English. I am including a bar chart so that it includes all the audiobook languages available as well as a pie chart to give a visual representation of density.

I also created a separate pivot tables for Author\_popularity and Narrator\_popularity. I included sum of stars and sum of ratings in each table and highlighted the top 25 in each column as a conditional format. This can be used to show how stars and ratings compare for Authors and Narrators and could potentially show trends in popularity. This could also allow advertisers to highlight authors and narrators that have a high number of stars but not as many ratings or vice versa. Below is a snippet from each table:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Author\_Name** | **Sum of stars** | **Sum of ratings** | | P. G. Wodehouse | 280 | 465 | | Innovative Language Learning | 224.5 | 448 | | Harvard Business Review | 199 | 421 | | Gertrude Chandler Warner | 183 | 43 | | Nora Roberts | 178.5 | 127 | | Ruskin Bond | 176 | 549 | | Brian Tracy | 172.5 | 1429 | | Osho | 160 | 567 | | Kristen Ashley | 158.5 | 39 | | Jeffrey Archer | 144 | 1579 | | Devdutt Pattanaik | 135 | 2097 | | Rick Riordan | 132.5 | 1480 | | Georgette Heyer | 132 | 100 | | Carolyn Keene | 124.5 | 196 | | Napoleon Hill | 119.5 | 1192 | | Agatha Christie | 119.5 | 432 | | David Baldacci | 111.5 | 587 | | Winter Morgan | 109.5 | 89 | | Franklin W. Dixon | 108 | 107 | | Lauren Blakely | 106.5 | 62 | | Lee Child | 102.5 | 671 | | Anthony Horowitz | 100 | 188 | | DK | 99 | 76 | | Roald Dahl | 94.5 | 598 | | Robert T. Kiyosaki | 93.5 | 7264 | | Bill Bryson | 93 | 799 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Narrator\_Name** | **Sum of stars** | **Sum of ratings** | | Sean Pratt | 346.5 | 1081 | | Sean Runnette | 210.5 | 398 | | Grover Gardner | 210 | 1227 | | Tim Gregory | 207.5 | 137 | | Walter Dixon | 207 | 205 | | David Pittu | 202.5 | 321 | | Scott Brick | 191 | 468 | | John Lee | 187.5 | 641 | | Brian Tracy | 162.5 | 1252 | | Aimee Lilly | 160 | 37 | | Jeff Harding | 153.5 | 994 | | OSHO | 150.5 | 435 | | Mike Chamberlain | 146.5 | 2732 | | Derek Perkins | 146.5 | 9300 | | Tom Parks | 146 | 1247 | | Paul Boehmer | 146 | 81 | | Stephen Fry | 143.5 | 12148 | | Rosalyn Landor | 143.5 | 206 | | Charlton Griffin | 138.5 | 168 | | Simon Vance | 138 | 1013 | | Jonathan Keeble | 136 | 97 | | John Sackville | 135.5 | 1675 | | Arthur Morey | 134 | 155 | | Jorjeana Marie | 131.5 | 202 | | Stefan Rudnicki | 129.5 | 83 | | Bombay Kannan | 129 | 1800 | |

**SQL**

**Exploratory Queries:**Develop at least 3 basic queries specifically designed to explore potential routes for analysis in your data.

select audiobook\_name

from Audiobooks

where stars >= 4 and ratings > 100;

select distinct author

from Audiobooks

where stars >= 4 and ratings > 100;

select distinct narrator

from Audiobooks

where stars >= 4 and ratings > 100;

select audiobook\_name, narrator

from Audiobooks

where author = "Roald Dahl"

select audiobook\_name, author, narrator

from Audiobooks

where audiobook\_name like '%Dragon%';

**Business Insight Queries:**Develop at least 3 advanced queries specifically designed to answer key business questions related to your dataset.