**Title: Exploratory Data Analysis**

**Author: Ronan Smyth**

**Student Number: L00176857**

**Date: 20/02/2024**

**NotebookURL:** <https://colab.research.google.com/drive/1gYVw5p701LdsjAlMJWWQ4EeDsWei5zKw?usp=sharing>

Contents

[Introduction 3](#_Toc159761014)

[Data Set 3](#_Toc159761015)

[Data Cleaning 6](#_Toc159761016)

[Analysis 10](#_Toc159761017)

[Missing Data analysis 10](#_Toc159761018)

[Univariate Analysis 11](#_Toc159761019)

[Bivariate Analysis 11](#_Toc159761020)

[Multivariate Analysis 12](#_Toc159761021)

[Time-Series Analysis 13](#_Toc159761022)

[Removal of Outliers 13](#_Toc159761023)

[Conclusion 15](#_Toc159761024)

[References 15](#_Toc159761025)

# **Introduction**

**The aim of this lab report is to perform exploratory data analysis on a data set containing data about audio books sold on audible.com. The dataset was obtained from Kaggle.com by searching specifically for uncleaned data sets.**

**Exploratory data analysis (EDA) includes cleaning this data to deal with missing data, incorrect data types, categorical data and outliers.**

**It also includes univariate, bivariate and multivariate analysis which examines the relationships between stand-alone values, one column compared to another and more than two columns, respectively. Time-series analysis can also be conducted if there is a time or date factor, which in this case, there is.**

**Finally, EDA also includes data visualisation so that the above observations can be more easily examined and interpreted.**

**I will use python in Google Colab to carry out the EDA.**

# **Data Set**

**Firstly, the dataset was loaded into Google Colab and the head() method was used to see the appearance and general layout.**

**A screenshot of a computer

Description automatically generated**

**The chosen dataset is comprised of 8 columns and 87,490 rows, including the row containing column titles, so 87,489 rows of data elements. This was observed using the df.shape attribute.**

**A close-up of a white background

Description automatically generated**

**On calling the describe() function, an unexpected output was displayed which I later learned was due to high amount of categorical data on the dataset. This is later remedied.**

**A screenshot of a computer

Description automatically generated**

**The following was what was expected and is output after calling describe() at the end of the dataset after converting most of the string data to numbers.**

**A screenshot of a computer

Description automatically generated**

**Calling the df.info() function shows that all data types are objects, there are no null values on the dataset and the number of populated rows in each in each column are equal and equal to the total number of rows shown in df.shape.**

**A screenshot of a computer

Description automatically generated**

**I also wanted to know what type of objects each column was. The following code was used to inspect the columns in more detail it was found that the data in every column was string data.**

**A screenshot of a computer code

Description automatically generated**

**As stated above, in this data set there are no null values at present. However, many of the audio books have not been rated yet which can be viewed as null or missing information and will be dealt with. Furthermore, many of the ‘price’ column entries are “Free” which will also have to be dealt with.**

**A white rectangular object with a black border

Description automatically generated**

**Furthermore, the ‘time’, ‘price’ and ‘stars’ columns are string data and would be more usable as numbers.**

**Lastly, the unique values in each column were inspected. Here, I made 2 observations. On one hand, it seems that some of the names of the books are duplicated. That likely means that some of the audiobooks have been entered twice in error. On the other hand, I wanted to see if the language column could be translated to numerical data using a key to decipher. However, there are 36 distinct languages of books in the dataset which would make a key too large and difficult to use.**

**A white background with a black border

Description automatically generated with medium confidence**

# **Data Cleaning**

**A screenshot of a computer

Description automatically generated**

**Note the obvious cleaning issues in the screenshot above:**

1. **Leading text (“written/narratedby:”) in the ‘author’ and ‘narrator’ columns will need to be deleted.**
2. **I will format ‘time’ as an integer which will be a total number of minutes.**
3. **The ‘releasedate’ is mm-dd-yy, and as afore mentioned a string. This should be formatted as a datetime object in pandas.**
4. **The ‘stars’ column contains too much information. This will be broken into 2 columns containing numerical, namely ‘rating’ and ‘totalReviews’.**
5. **As price is also a string, it will be changed to a double and all necessary formatting changes made.**

**To begin the cleaning process, the leading text has been removed from the ‘author’ and ‘narrator’ columns as per the following code and data frame head.**

**A screenshot of a computer

Description automatically generated**

**Then, the date has been formatted as a datetime object. It was attempted to format the date in standard European format(dd-mm-yyyy) but this created problems later on when the data was used. It was found that either Google Colab or Python Pandas performed better with the default format(yyyy-mm-dd).**

A computer screen shot of a computer screen

Description automatically generated

Then the leading text was removed from the ‘author’ and ‘narrator’ columns and checked using df.tail().

A screenshot of a computer

Description automatically generated

The ‘time’ column was converted to aggregate minutes and as such an integer using the following code. This code was taken from Kaggle.com (Darshan77879, 2023).

A computer screen shot of a code

Description automatically generated

The effect can be seen in the data types below and the following head.

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

The next item to be cleaned is the ‘stars’ column. This requires the formation of two new columns which have been called ‘rating’ and ‘totalReviews’ and are of float and integer data types, respectively. Numerical information has been taken from the start and end of the ‘stars’ column to populate the two new columns. The ‘stars’ column has then been deleted. The following code was used.

A screenshot of a computer program

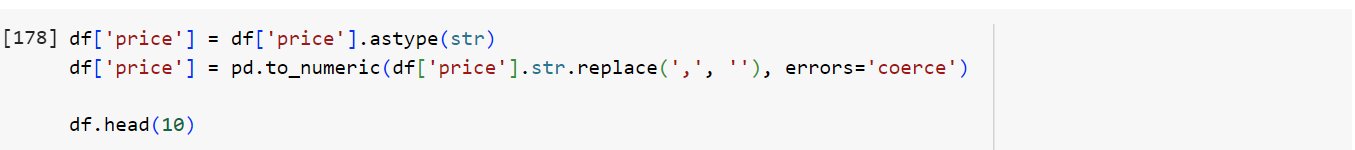
Description automatically generated

The new columns can be seen in this data frame head as well as the formatting of the values therein.

A screenshot of a computer

Description automatically generated

It was noticed at this point that not only was the ‘price’ column in string format, but it also contained some string characters, i.e. commas to the right of every 4th digit from the left. However, when trying to perform string functions on the ‘price’ column I found that neither was it entirely of string type. I used the following code to convert it to a string, remove the commas and then convert it to a float.



The final cleaning procedure to be carried out was to eliminate duplicate entries in the ‘name’ column. This suggests that some audiobooks have been entered twice.

A screenshot of a computer

Description automatically generated

The duplicates were removed using the following code.

A screenshot of a computer

Description automatically generated

# **Analysis**

## **Missing Data analysis**

All of the data in the dataset set when starting off was of string type. The ‘time’, ‘price’ and ‘stars’ data has now been changed to numeric. The ‘stars’ column has been deleted after extracting information from ‘stars’ to populate two new columns, namely ‘rating’ and ‘totalReviews’.

While in string format, values like “Free” in the ‘price’ column’ or “Not yet rated” in the ‘stars’ column wouldn’t have been identified as null values. After converting those columns to numeric data types, there are now null values (NaN & <Na>) in the dataset which must be addressed.

A white background with black text

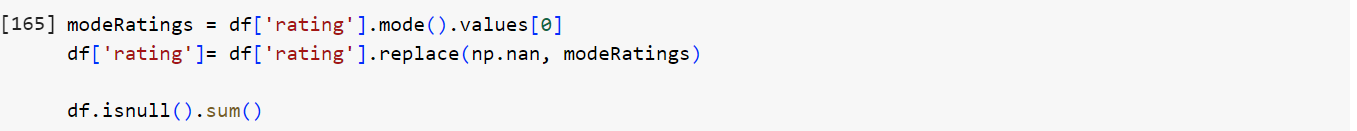
Description automatically generated

In the ‘rating’ column, the mode value has been selected to replace NaN values. This is due to the subjective nature of ratings.

To deal with missing values in the ‘totalRatings’ column, it was decided to replace them with minus-one. This is because zero would give the impression that the data was relevant but did not contribute and may cause a false interpretation. Minus-one can be viewed as more distinct.

Finally, in the price column, “Free” has been replaced with zero since this accurately represents the price of the item.

The following code was used to deal with missing vaues.



A close-up of a computer screen

Description automatically generated

A close-up of a computer screen

Description automatically generated

## **Univariate Analysis**

Two histograms were produced viewing price distribution. The first provided an x-axis which was unnecessarily long. The second then focuses on the price range where the greater majority of the books are found.

A graph of two people

Description automatically generated

## Bivariate Analysis

To give an idea of ratings distribution, a histogram (left) was produced showing a heavy imbalance in the data with most values being 5 stars. Then, ‘rating’ was examined in respect to ‘price’ in a scatter plot (right). It is shown here that audiobooks with higher ratings receive higher price, by and large. However, with such a severe imbalance in the ‘rating’ column, the information may not be valuable.

A graph and a diagram

Description automatically generated with medium confidence

## Multivariate Analysis

A heat map has been created to show relationships between the numerically formatted columns. Very little correlation is shown on the heatmap. The only correlation is a medium-strength correlation between the price of an audiobook and the length of the recording. That is to say that sometimes, the longer an audiobook is, the more expensive it will be.

The range of the heatmap is set from one to minus-one, one (dark-red) being a strong positive relationship and minus-one (dark-blue) being a strong negative one. Zero (light-grey) represents no relationship.

The darker the red on the heatmap, the stronger the relationship is whereas with the blue colours, the darker the blue would indicate a strong negative, or inverse, relationship.

The heatmap below is mostly grey, indicating very weak, irrelevant positive correlations. The light blue shows the same weak, irrelevant connection in a negative or inverse way.

A screenshot of a graph

Description automatically generated

## Time-Series Analysis

Time series analysis was carried out showing change in price over time. Interestingly, audiobooks were more expensive or there was a spike in price just before the year 2000. A price increase can also be seen just before the current year suggesting new releases are more expensive or perhaps reading false information, the year having only just begun.

A graph with blue lines

Description automatically generated

# Removal of Outliers

This last part of the lab hasn’t quite been completed or understood well.

A boxplot was produced showing the inter quartile range over 3-year periods. The years 2024 and 2025 were removed as they themselves may represent outliers, being as of yet incomplete. Some outliers can be seen in the below graphic.

A graph with blue squares

Description automatically generated with medium confidence

I was unable to complete the same boxplot and remove the outliers at the same time. Instead I used the following code to create a scatter plot of average prices over time with outliers removed.

A computer screen shot of a computer code

Description automatically generated A computer code with many small colored text

Description automatically generated with medium confidence

A screen shot of a computer code

Description automatically generated

A graph with numbers and dots

Description automatically generated with medium confidence

# Conclusion

**In this lab report, a comprehensive exploratory data analysis (EDA) was conducted on the selected audible dataset. This was done to gain insight and understand patterns and characteristics within the data. The primary objectives were to identify trends, detect outliers, and explore relationships between variables. This has been completed.**

**Findings of the EDA show very little correlation between the given variables however, more in-depth study may reveal otherwise.**

# References

Darshan77879 (2023) *Audible Data Magic: Cleaning, Assessing, Enhancing*. Available at: <https://kaggle.com/code/darshan77879/audible-data-magic-cleaning-assessing-enhancing> (Accessed: 25 February 2024).

De, S. (2022) *Audible Dataset*. Available at: <https://www.kaggle.com/datasets/snehangsude/audible-dataset> (Accessed: 25 February 2024).

Luna, F. (2022) ‘5 Datasets to Practice Data Cleaning’, *Medium*, 19 July. Available at: <https://medium.com/@FranciscoHinojosaLuna/5-datasets-to-practice-data-cleaning-27378f422e1c> (Accessed: 25 February 2024).