Investigating the Critical Role Data Cleaning Plays in the Data Analysis Process

Funmilayo Olaiya [1]

Abstract— **In this research, we examined different data cleaning tools and techniques so as to evaluate their effectiveness. We compared their results to determine similarities and differences. We found all tools through Google Scholar, except one that we came to know about through a 2023 blog post.**

After selecting the audible dataset for simplicity and interest, we cleaned it using each tool and analyzed the trends in the cleaned data. Python as a language and tool produced more sensible results than other tools, but we couldn't determine the best tool due to the lack of a ground truth or established methodology.

I. Introduction

In the world of data analysis today, data cleaning stands out as one of the most crucial steps to undertake. Ensuring data quality after cleaning a dataset remains a significant concern, as dirty data can still result in erroneous decisions and unreliable analysis [1].

In this article [2] written by Natasia Langfelder, it is reported that statistically, the United States loses about 3 trillion dollars per year, *according to IBM*, because of bad data.

As a result, meticulous attention and effective strategies are required to maintain high data quality standards throughout the entire data cleaning process to guarantee accurate and dependable results in analysis.

Data cleaning, also known as data cleansing involves the process of identifying and eliminating errors and inconsistencies from datasets to enhance the overall quality of the data [3].

In this research, our primary focus lies in understanding the profound impact of data cleaning on the overall outcomes of data analysis.

We are keen to investigate and fully comprehend the influence that effective or inadequate data cleaning processes can have on specific results and insights. In this paper, **we present the following contributions**:

1) We conducted a survey of effective data cleaning tools and got some which are; Open Refine, Trifacta, Akkio and Python to explore and employ them in cleaning our dataset (Audible dataset) [4].
2) We analyzed four distinct trends using the cleaned datasets to explore and assess the differences and similarities in the results provided.

II. Methodology

A. Survey

Our survey approach was straightforward. The majority of the data cleaning tools we discovered were obtained by using the search phrase *data cleaning tool* on google scholar. However, there was one exception - we learned about a tool called **Akkio** [5] from a particular blog post [6]. Apart from Akkio, the three other tools we decided to settle with are; **Open Refine** [7], **Trifacta** [8] and **Python** [9] as shown in Table I.

*1) **Open Refine**:* To elaborate further on our interest in these tools, let's begin with Open Refine [7], which we encountered in this paper [10]. The paper aimed to explore four data cleaning tools (in which open refine was one of them), so as to provide recommendations to researchers and organizations on selecting the appropriate data cleaning tool.

*2) **Trifacta**:* Another tool we came across was Trifacta [8], which we found in this paper [11]. The study focused on a user survey, where data analysts or engineers working with data were interviewed about the tools they use for data cleaning. Trifacta was one of the tools mentioned in the survey.

*3) **Akkio**:* As mentioned earlier, Akkio [5] was a tool we discovered through a blog post [6]. Our interest in exploring Akkio was driven by our curiosity to see the latest advancements in the industry and gain a more comprehensive understanding of its capabilities.

*4) **Python**:* Python, both as a language and a tool, is an indispensable resource for nearly every data analyst or engineer, forming an integral part of their daily workflow. Its widespread popularity is well-founded, especially when it comes to data cleaning and analysis tasks.

Interestingly, during our survey, we rediscovered that Python was still among the tools commonly used, as indicated in this paper [10]. This finding further strengthened our resolve to explore Python for our data cleaning investigation.

Now, we will delve deeper into the reasons behind our interest in certain tools that eventually fell under the unused tools category. It is essential to note that all of these reasons were subjective.

*5) **SmartClean**:* Smartclean [12] appeared as an exceptionally fascinating tool, especially considering its introduction in 2009. Essentially, the authors asserted that it had the capability to detect DPQs (data quality problems) without requiring users to specify the execution sequence of data cleaning operations. Our primary issue was that we couldn't simply locate this tool. To the best of our knowledge, we were unable to find it online, but it's not entirely certain as it's possible we didn't dig deep enough in our search.

[1] X is with the School of Computer Science, University of Waterloo, 200 University Avenue, Waterloo, Ontario, Canada N2L 3G1.

*6) **Wrangler**:* Wrangler, a data cleaning tool, made its debut in this paper [13] back in 2011. We were notably intrigued by this tool due to the realization that Trifacta's design was essentially an extension of Wrangler. However, we ultimately couldn't utilize it as the interface proved to be peculiar and we struggled to quickly familiarize ourselves with its workings.

*7) **Ajax**:* Interestingly, Ajax [14] appeared to be a tool worth exploring; however, regrettably, we faced difficulties in accessing and reading the paper. The document was quite corrupted on multiple platforms, leading to a loss of our interest in further investigating the tool.

*8) **WinPure**:* WinPure [15] can undoubtedly be regarded as one of the state-of-the-art tools for data cleaning. However, we encountered a significant limitation in its use, as it was available for Windows and lacked an application compatible with macOS.

TABLE I: Data Cleaning Tools.

Used Tools	**Unused Tools**
Open Refine	Smart Clean
Trifacta	Wrangler
Akkio	Ajax
Python	WinPure

B. Approach

Our approach primarily focused on the Audible dataset, which was created by a data analyst called Snehangsu De [16] and made available on Kaggle [4]. The motivation behind the creation of this dataset was to provide up-to-date information on the basics and history of audiobooks, and we believed this fitted right into our investigative research. Firstly, we tried to find data problems within the audible dataset before we tried to clean it. Some of these problems are;

1) In the author column, *written by* is directly placed in front of all the author names.
2) Solid numbers like 1,003.00 had to be converted to *1003.00* so as to fit into the float datatype.
3) Free audiobooks are written as free instead of 0.
4) Time is denoted with *hrs* and *mins* instead of a solid integer number.

Additionally, during a thorough examination of the dataset, we found that it did not contain a significant number of duplicate or missing values. This rarity is valid and depends on how the dataset's creator scraped the data.

Indeed, while there may be other instances of dirty data within the dataset, we have currently identified and addressed the ones that directly impact the trends we aim to measure. Focusing on these specific data problems ensures that our analysis remains relevant and accurate for the targeted trends.

To achieve our objectives, we aimed to measure specific trends within the Audible dataset. However, before conducting any analysis, we recognized the need to clean the dataset thoroughly. Given our investigation into data cleaning tools, we selected four tools, as mentioned in the survey section, to clean the Audible dataset. The goal was to generate a cleaner version of the dataset using each tool and subsequently measure trends within these newly cleaned datasets.

Our intention was to compare and contrast the outcomes produced by each data cleaning tool. By doing so, we aimed to gain insights into the effectiveness of different data cleaning strategies and how they could significantly impact the generation of high-quality insights from the data.

The ultimate objective was to understand the power and implications of employing different data cleaning approaches and their influence on the accuracy and reliability of the results derived from the Audible dataset.

III. Evaluation and Results

As previously stated, in the earlier sections, we took the necessary steps to clean the audible dataset using four distinct tools: **OpenRefine**, **Trifacta**, **Akkio**, and **Python**. The cleaned version of the dataset serves as the foundation from which we will derive valuable insights and conduct further analysis. The following sections will provide further insights into the specific trends we measured and examined in detail.

*1) **The Top Ten Authors with the Highest Number of Audiobooks**:* In this analysis of the Audible dataset, our main objective was to identify the top ten authors who have the highest number of audiobooks over the years. To achieve this, we employed each of the four cleaned datasets generated by the tools, namely OpenRefine, Trifacta, Akkio, and Python.

In this scenario, it appears that the results obtained from Open Refine's dataset stood out as odd and significantly differed from the results generated by the other three tools (Trifacta, Akkio, and Python) whereby all three tools reported the first author to have more than 800 audiobooks in the dataset.

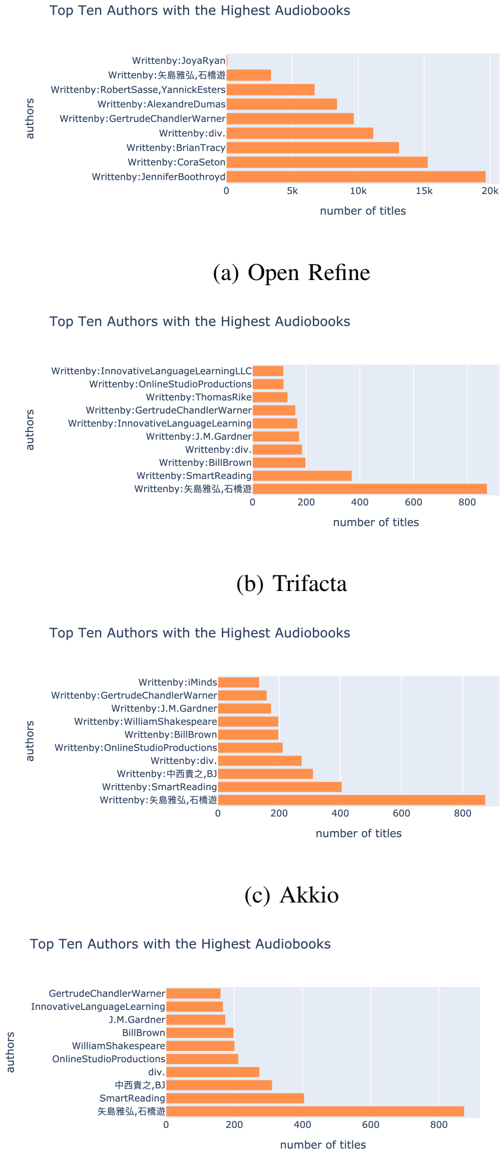
Figure 1 displays a visual representation of the differences or similarities observed in the results from these tools.

*2) **Number of Audiobooks Released per Year**:* Our aim was to evaluate the trend of the number of audiobooks released each year, as illustrated in the Figure 2. It is evident from the results that the outcomes of Akkio and Python were somewhat similar, while that of Trifacta and Open Refine showed distinct differences.

Overall, Trifacta, Python, and Akkio all indicate that approximately year 2021 witnessed the highest number of audiobooks being released, as per their respective analysis.

*3) **Relationship between the Length of Audiobooks and Year**:* Another form of analysis we did was to check the trend of the length of audiobooks in accordance to the year they were published. As shown in Figure 3, we can see that Akkio and Trifacta's results seem very close but that of Python and Open Refine looked different. According to Python's results, it is apparent that around the year 2019 marked the period when the longest audiobook, spanning over 8000 minutes, was released.

*4) **Top Ten languages with the Highest Number of Audiobooks**:* In addition, we examined the trend of the top ten languages in which audiobooks have been published. Figure 4 indicates that all tools produced somewhat similar results, except for Trifacta, which reported approximately 15,000 audiobooks published in English, while the other tools were in the range of 60,000.



(a) Open Refine

(b) Trifacta

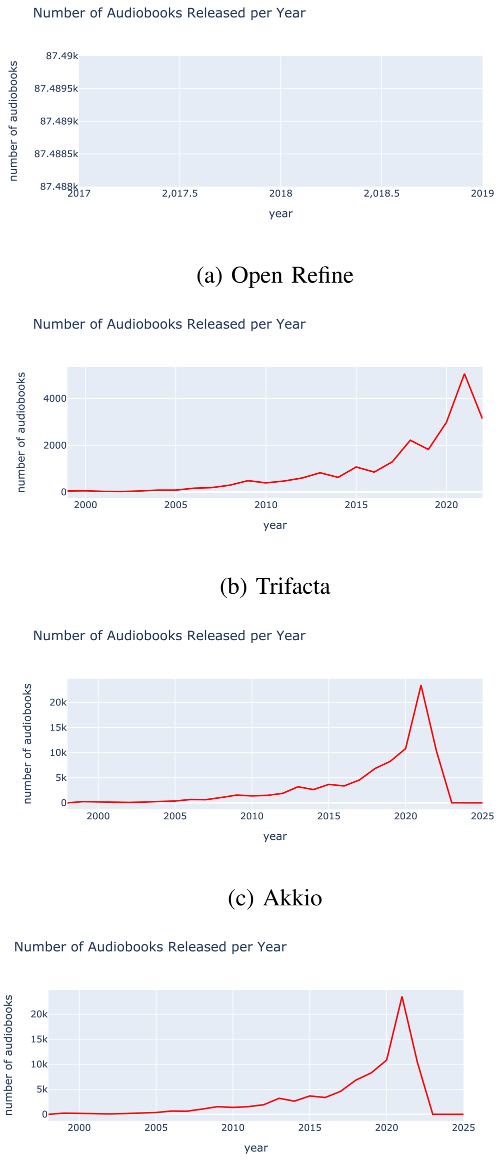
(c) Akkio

(d) Python

Fig. 1: Top Ten Authors with the Highest Number of Audiobooks (titles)

IV. Conclusion

We explored four tools for cleaning the audible dataset by utilizing them, and we evaluated four trends with each cleaned dataset from these tools. We observed that, for certain trends, some tools produced awkward results, whereas for others, the results were quite comparable.



(a) Open Refine

(b) Trifacta

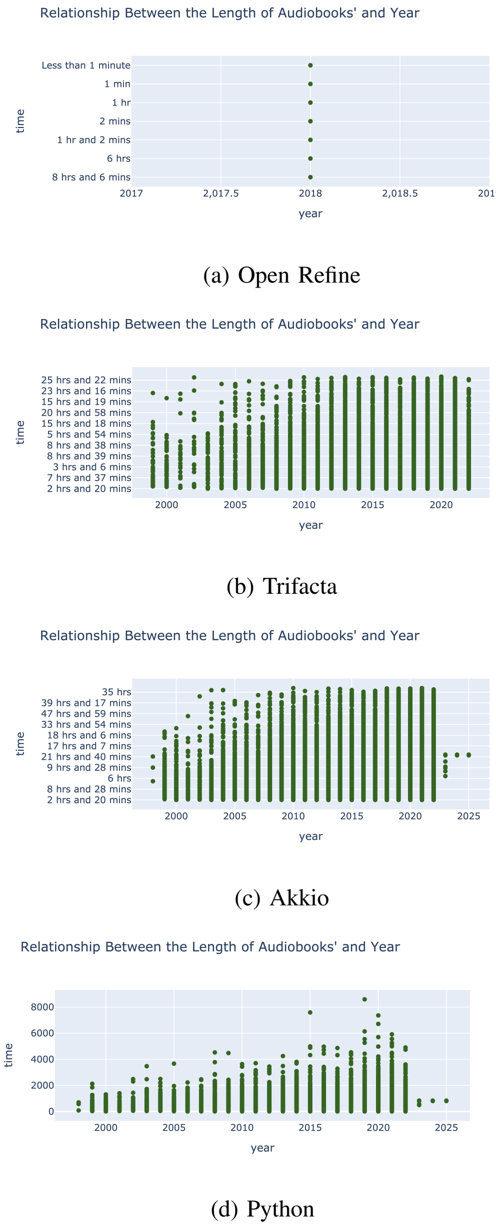
(c) Akkio

(d) Python

Fig. 2: Number of Audiobooks Released per Year

Based on the results, it becomes evident that writing your own Python programs to clean a dataset yields significantly better and more sensible results. Custom-written programs for data cleaning provide a highly personalized and specific approach, allowing researchers to fine-tune the cleaning process to match the unique requirements and complexities of their datasets.

In contrast, other data cleaning tools, *especially the ones with interfaces*, offer a helping hand in dealing with more generalized data issues, such as handling null values, resolving date discrepancies, and other common challenges. These



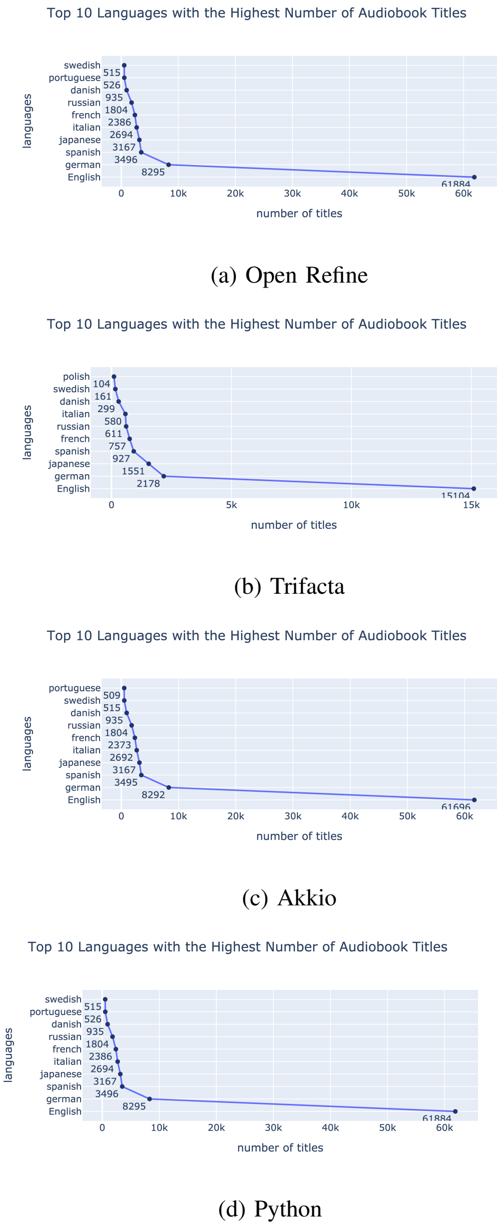
(a) Open Refine

(b) Trifacta

(c) Akkio

(d) Python

Fig. 3: Relationship between the Length of Audiobooks and the Year



(a) Open Refine

(b) Trifacta

(c) Akkio

(d) Python

Fig. 4: Top Ten Languages with the Highest Number of Audiobooks

tools come with *pre-built functionalities* that efficiently address typical data problems, saving valuable time and effort. By combining the strengths of custom-made programs and data cleaning tools, researchers can enhance the accuracy and reliability of their data analysis, while effectively managing various data complexities and ensuring high-quality results.

Another noteworthy aspect is that each tool exhibited *certain peculiarities*, which likely contributed to the variations in their generated results. However, due to space limitations, we cannot delve further into these intricacies in this paper.

Determining the best-performing tool or techniques for cleaning data remains challenging within the data analysis community even today. Consequently, an unanswered question persists: *How can we effectively validate cleaned data to ensure its quality?*

V. Acknowledgments

We express our sincere appreciation to Snehangsu De [16], the creator of the Audible dataset, as his earlier work has been instrumental in helping us gain insights into how to measure data trends using this dataset. His contribution has been invaluable in shaping our approach and understanding the potential of this dataset for our research.

References

[1] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data cleaning: Overview and emerging challenges," in *Proceedings of the 2016 international conference on management of data*, 2016, pp. 2201–2206.

[2] N. Langfelder, "The true cost of bad data," 2023. [Online]. Available: https://www.data-axle.com/resources/blog/the-true-cost-of-bad-data/

[3] E. Rahm, H. H. Do *et al.*, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[4] S. De, "Audible dataset," 2023. [Online]. Available: "https://www.kaggle.com/datasets/snehangsude/audible-dataset"

[5] Akkio. [Online]. Available: https://www.akkio.com/

[6] J. Reilly, "7 best data cleaning tools for analysts in 2023," 2023. [Online]. Available: https://www.akkio.com/post/data-cleansing-tools

[7] O. Refine. [Online]. Available: https://openrefine.org/

[8] Trifacta. [Online]. Available: https://www.trifacta.com/

[9] Python. [Online]. Available: https://snehangsude.github.io/

[10] Z. Chen, S. Oni, S. Hoban, and O. Jademi, "A comparative study of data cleaning tools," *Int. J. Data Warehous. Min.*, vol. 15, no. 4, p. 48–65, oct 2019. [Online]. Available: https://doi.org/10.4018/IJDWM.2019100103

[11] S. Krishnan, D. Haas, M. J. Franklin, and E. Wu, "Towards reliable interactive data cleaning: A user survey and recommendations," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, ser. HILDA '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2939502.2939511

[12] P. Oliveira, F. Rodrigues, and P. Henriques, "Smartclean: An incremental data cleaning tool," in *2009 Ninth International Conference on Quality Software*, 2009, pp. 452–457.

[13] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 3363–3372. [Online]. Available: https://doi.org/10.1145/1978942.1979444

[14] H. Galhardas, D. Florescu, D. Shasha, and E. Simon, "Ajax: An extensible data cleaning tool," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 590. [Online]. Available: https://doi.org/10.1145/342009.336568

[15] WinPure. [Online]. Available: https://winpure.com/

[16] S. De. [Online]. Available: https://github.com/snehangsude