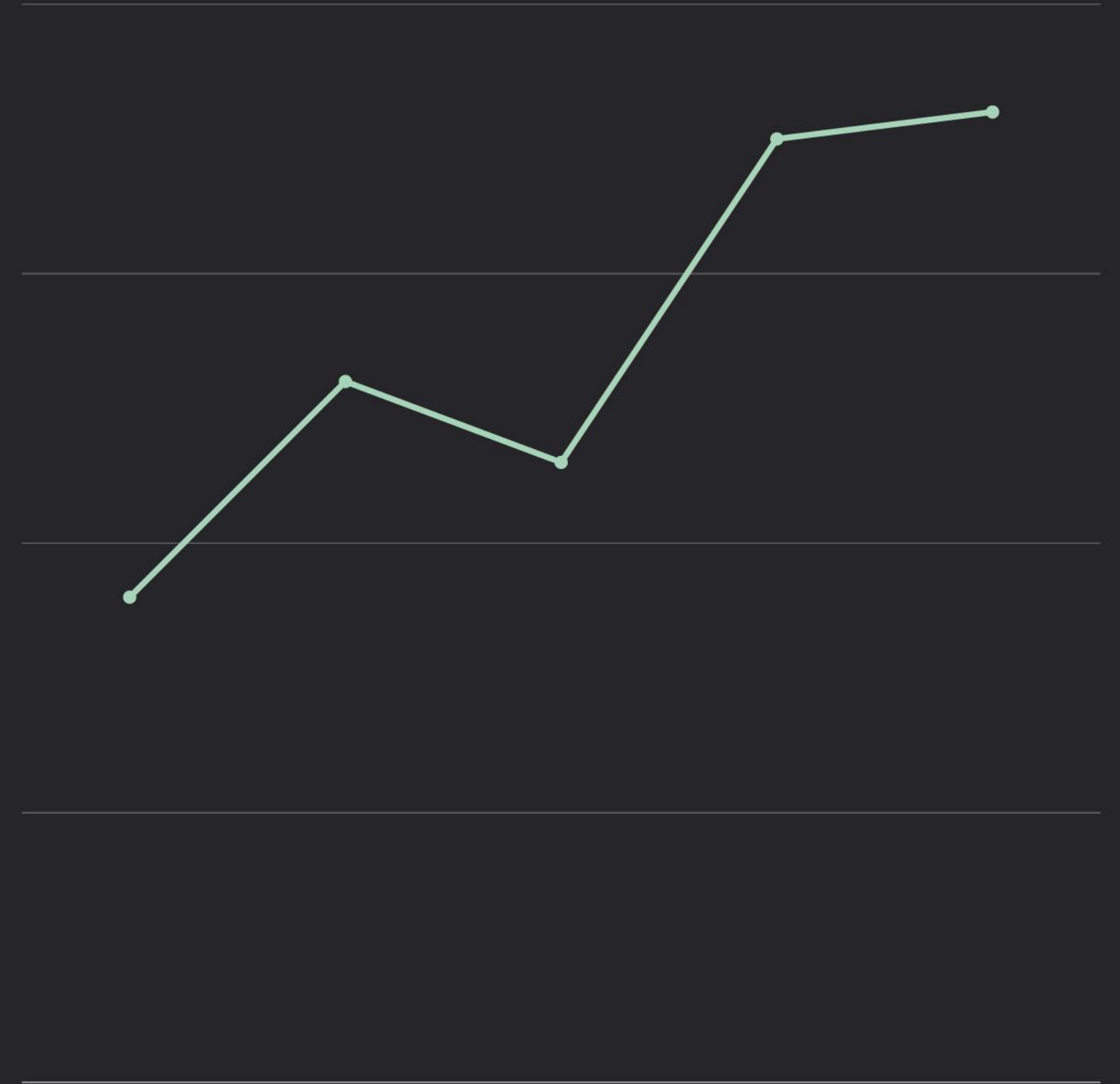


INVESTIGATING THE CRITICAL ROLE DATA CLEANING PLAYS IN THE DATA ANALYSIS PROCESS

PRESENTED BY
Funmilayo Olaiya
folaiya@uwaterloo.ca





Outline

1 Introduction & Motivation

2 Methodology

3 Survey & Background

4 Results

5 Key Takeaways

5 Conclusion

QUICK FACT

What is the true cost of bad data? - insurmountable

\$3 Trillion

The true cost of bad data

Natasia Langfelder

Published in the Data Axle blog, 2020

Concerns: Disengaged audiences, Reputation, Missed
Opportunities

INTRO. & MOTIV.

- **Data cleaning** is simply the detecting and repairing of dirty data.
 - Personally, dirty data is coined subjectively in relation to what needs to be done with that data.
- According to Chu et al. (***Data Cleaning: Overview and Emerging Challenges***)
 - One of the constant issues in data analytics is identifying and fixing corrupt data, and failing to do so can lead to inaccurate analytics and unreliable conclusions.
 - But which technique works best?
- Became interested in the **Audible Dataset** by Snehangsu De.

METHODOLOGY

APPROACH

Identify some problems with the audible dataset

Certain data flaws

Survey of data cleaning tools to help solve them

Four of them

Analyse certain trends and compare the results gotten after cleaning data with each of the tools

Four trends

PROBLEMS IDENTIFIED

Authors have "written in" front of their names

Writtenby:GeronimoStilton
Writtenby:RickRiordan
Writtenby:JeffKinney

Time should be in minutes

2 hrs and 20 mins
13 hrs and 8 mins
2 hrs and 3 mins

Price

334.00	323.00
Free	1,003.00
469.00	342.00

- Duplicate Values
- Null or Missing Values

SURVEY

- *data cleaning tool* - google scholar
- Top 7 Data Cleansing Tools in 2023

Used Tools	Unused Tools
Open Refine	SmartClean
Trifacta	Wrangler
Akkio	Ajax
Python (written programs)	WinPure

BACKGROUND

A Comparative Study of Data Cleaning Tools by Samson Oni et al.



Samson Oni, University of Maryland, Baltimore County, USA

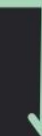


- **Open Refine** (formerly known as freebase gridworks) By Metaweb.
- Google acquired Metaweb => Google Refine.
- Google dropped Google Refine and the creators rebranded to Open Refine to make it open-source.
- Type of tool => **Desktop Application**

Towards Reliable Interactive Data Cleaning: A User Survey and Recommendations



Sanjay Krishnan, University of California, Berkeley, USA



- **Trifacta** (newly acquired by Alteryx)
- Founded by Jeff Heer, Joe Hellerstein, and Sean Kandel who wrote the paper: *Wrangler: Interactive Visual Specification of Data Transformation Scripts*
- Extension of Wrangler's design
- Type of tool => **Cloud**

Akkio



Abraham Parangi, Co-founder and
Ceo

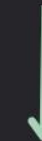


- **Akkio** (founded in 2019)
- Relatively new and uses AI
- Type of tool => **Cloud**

Audible Dataset



Snehangsu De, Data Analyst

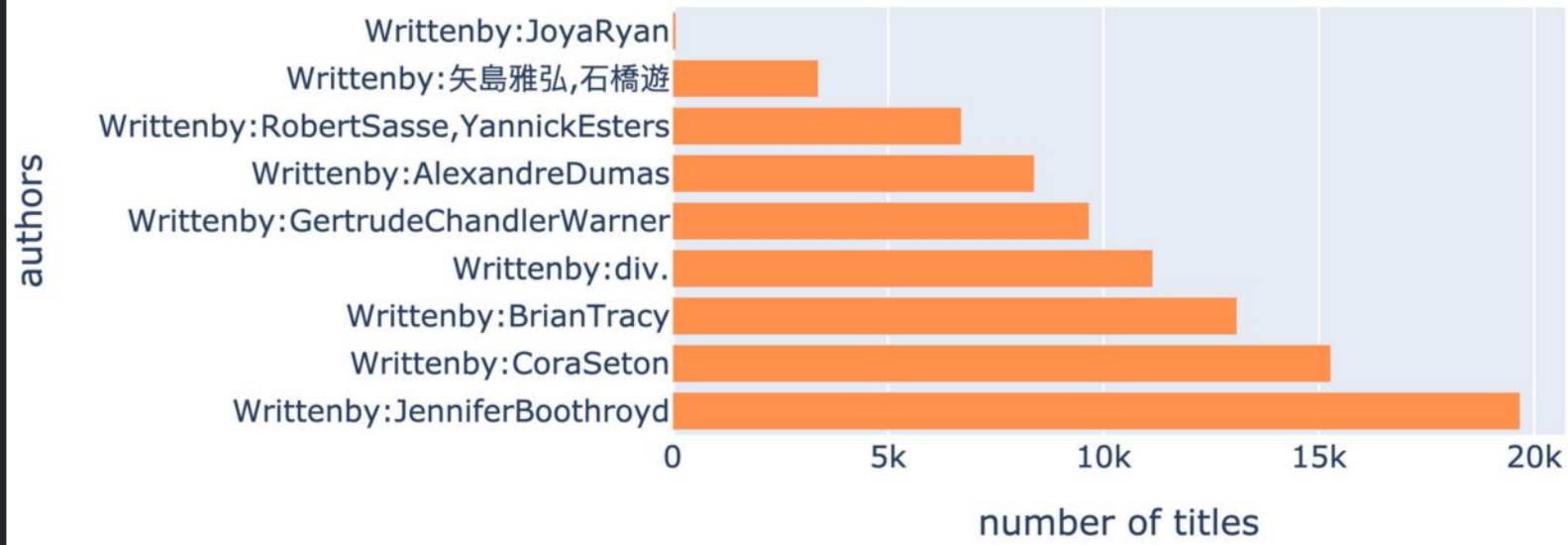


- **Python Programs**
- Created the audible dataset and also created another clean version of it

RESULTS

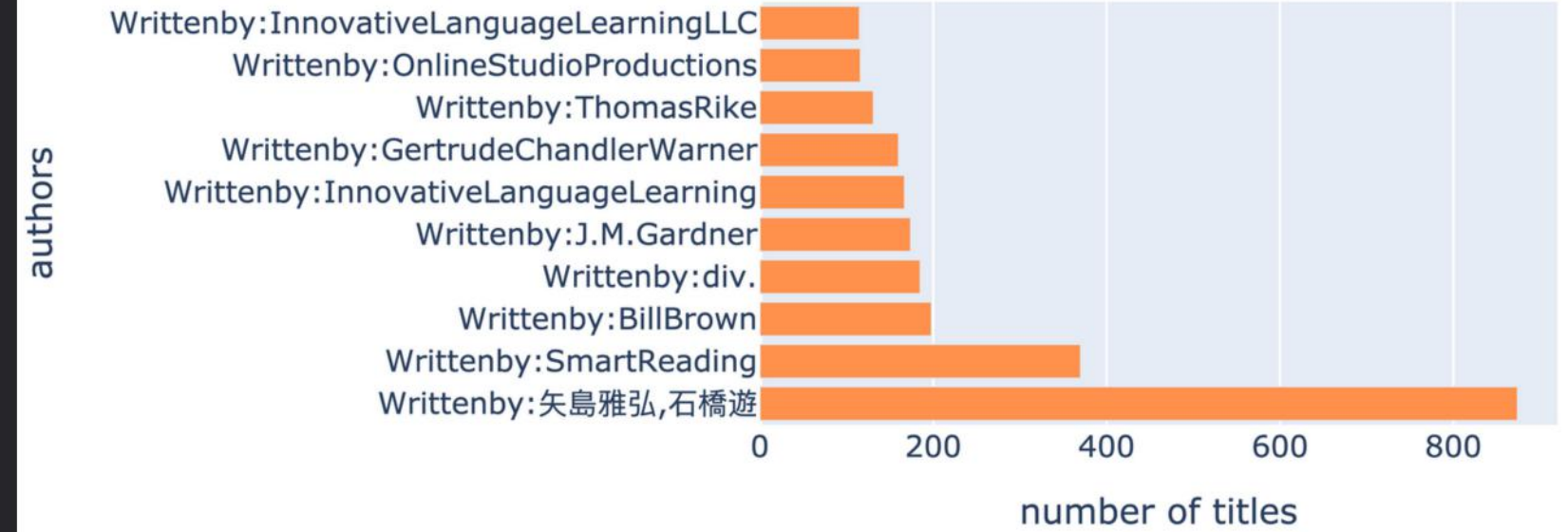
open refine

Top Ten Authors with the Highest Audiobooks



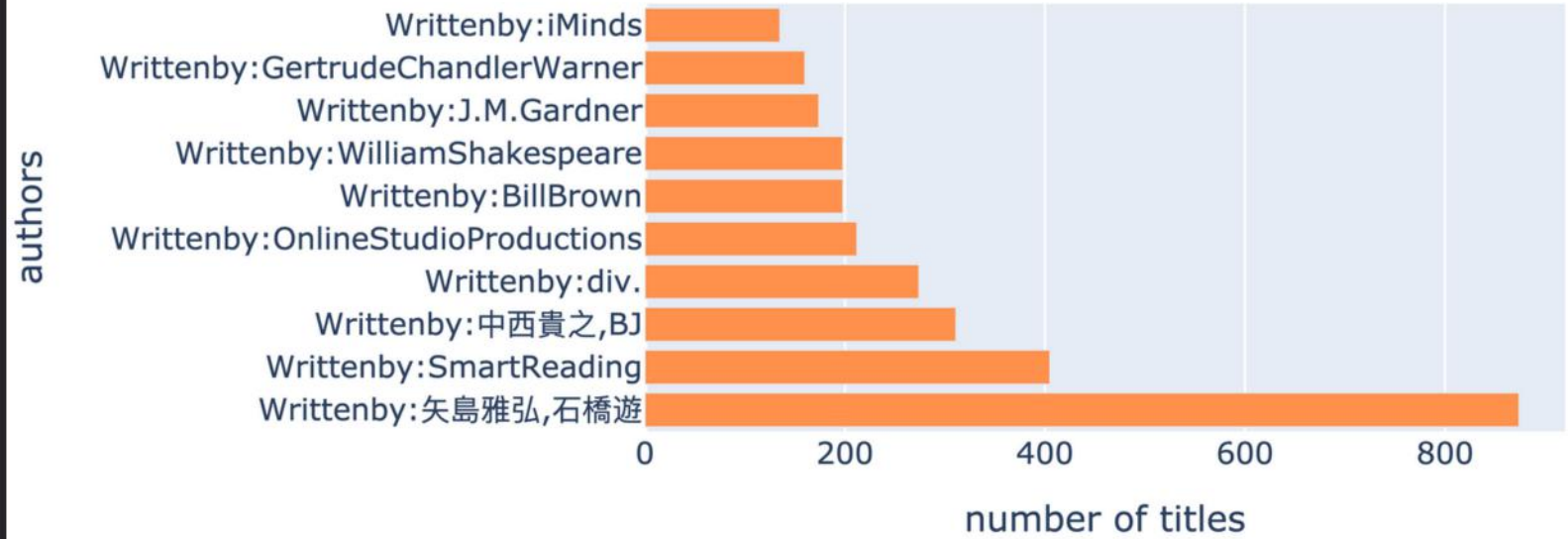
trifacta

Top Ten Authors with the Highest Audiobooks



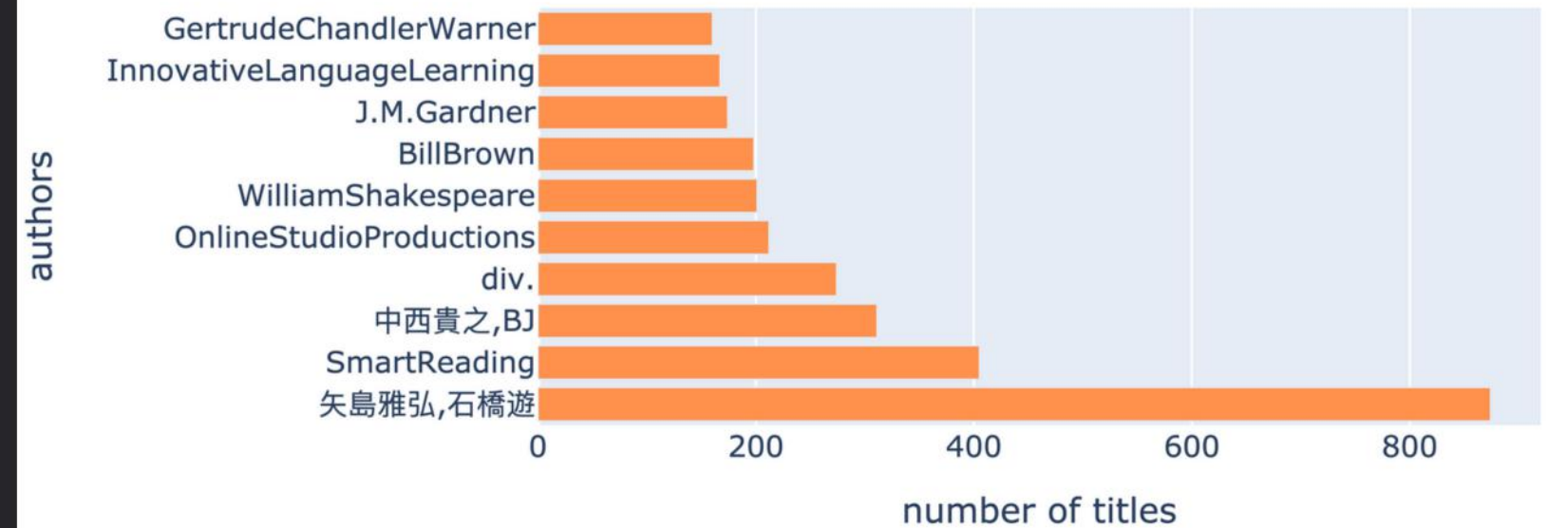
akkio

Top Ten Authors with the Highest Audiobooks

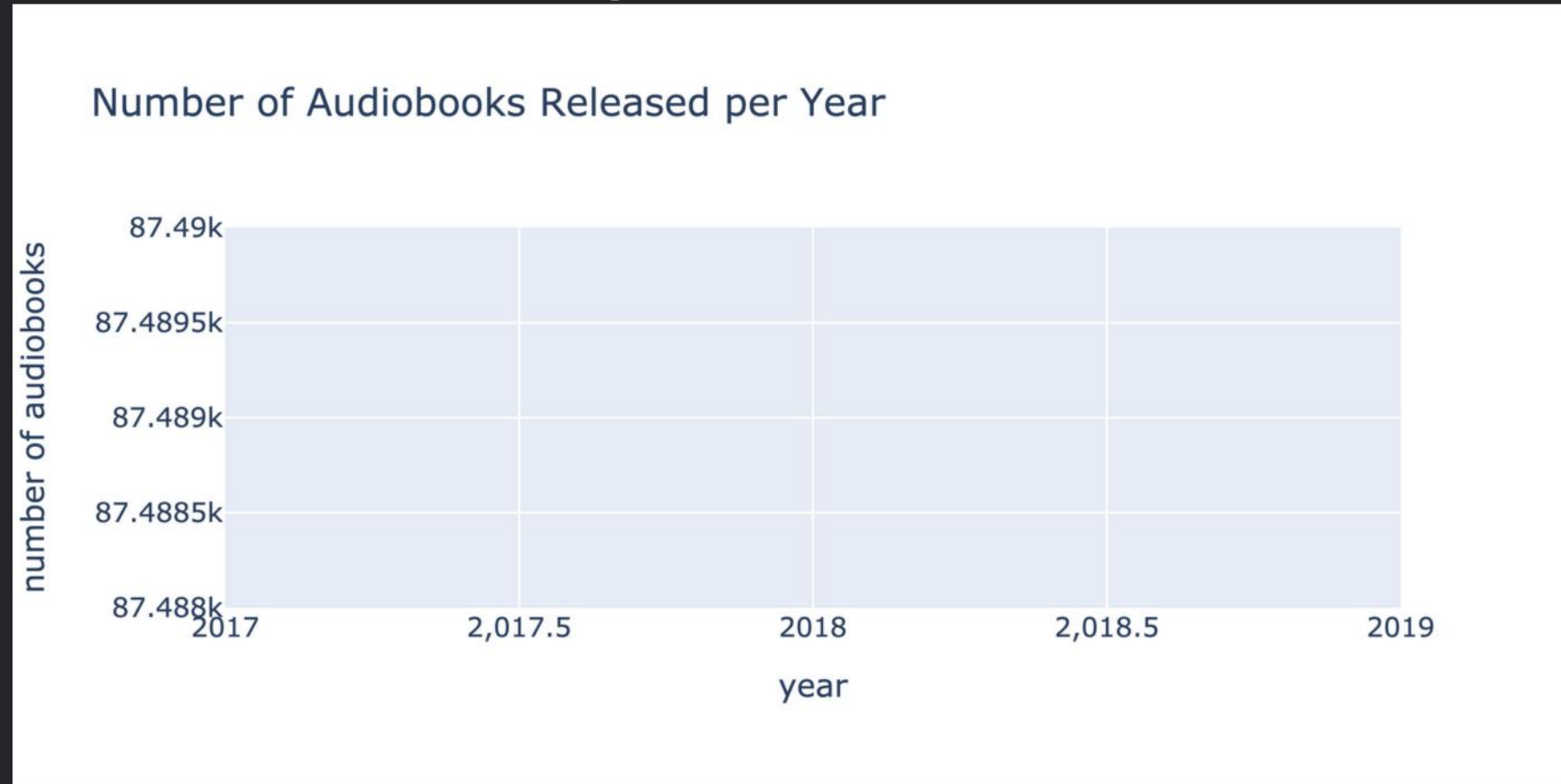


python

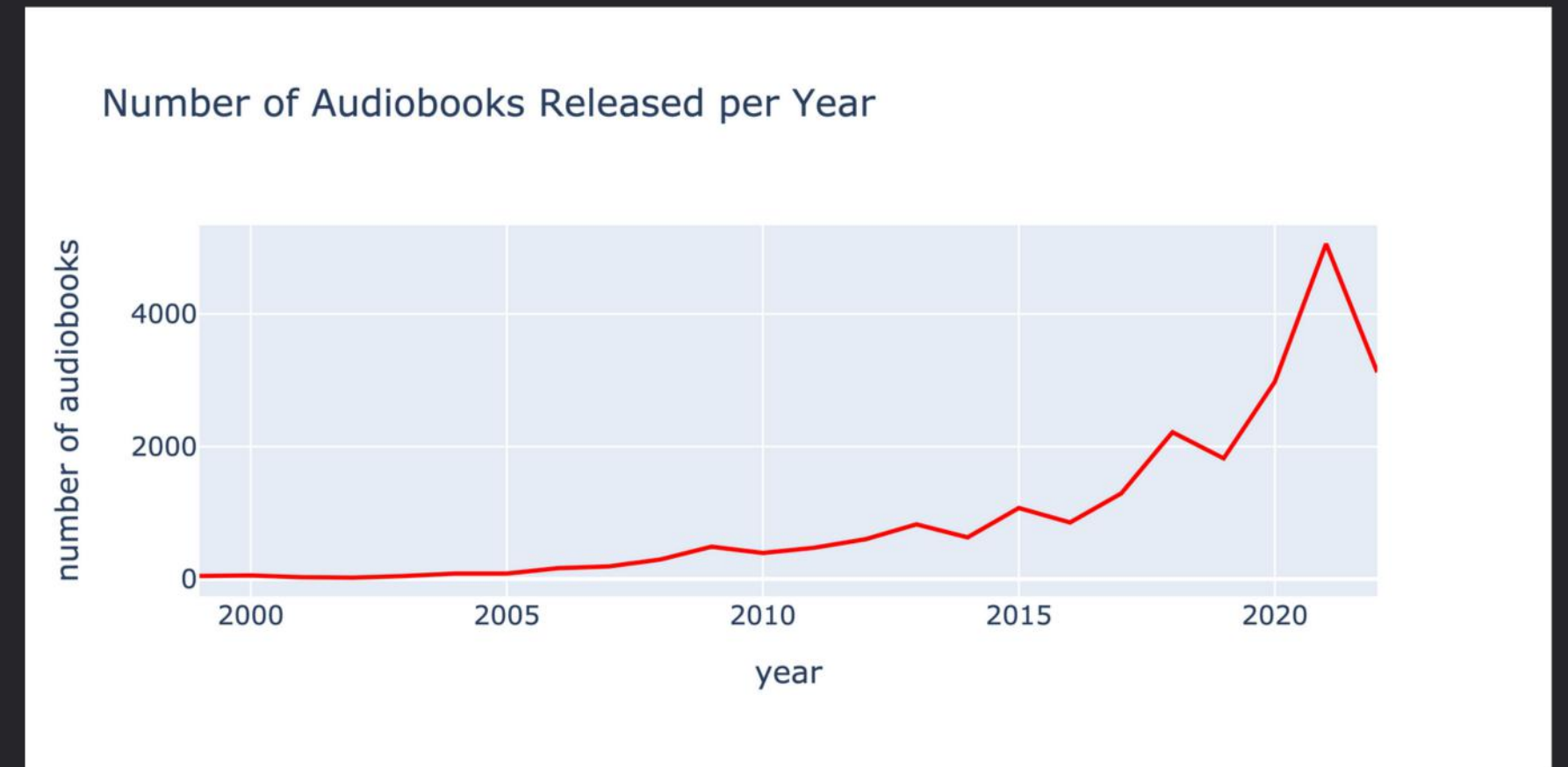
Top Ten Authors with the Highest Audiobooks



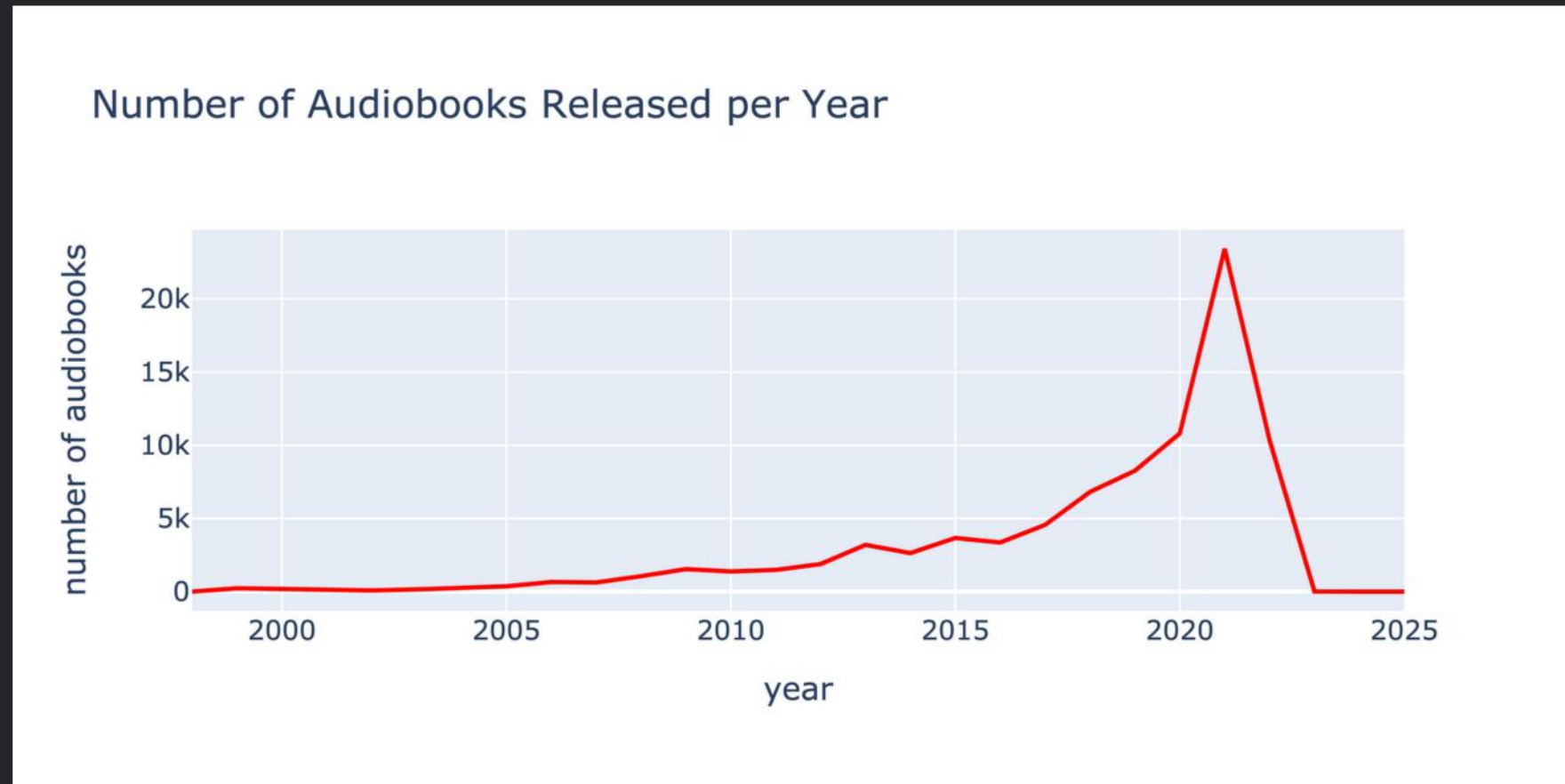
open refine



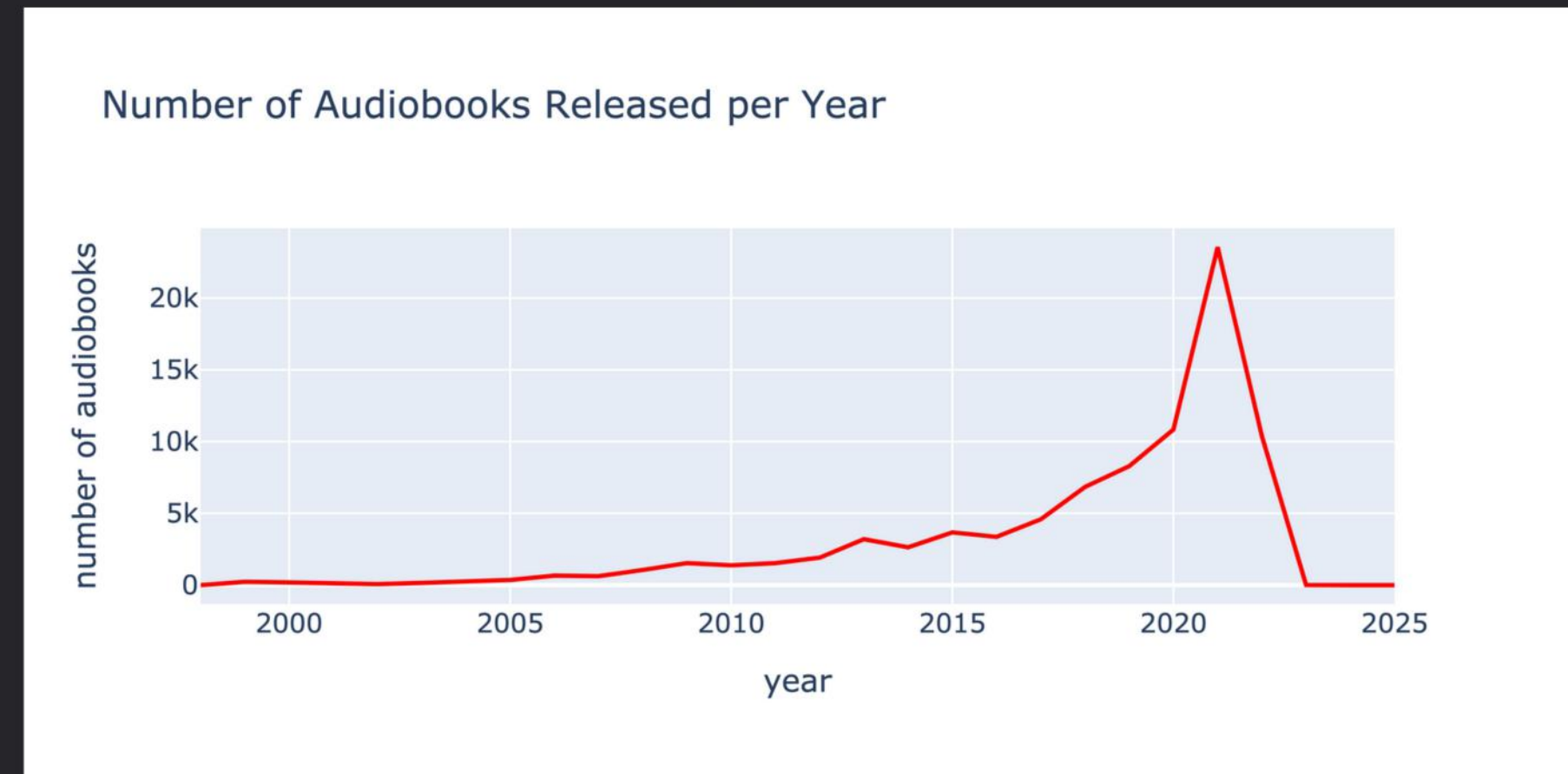
trifacta



akkio

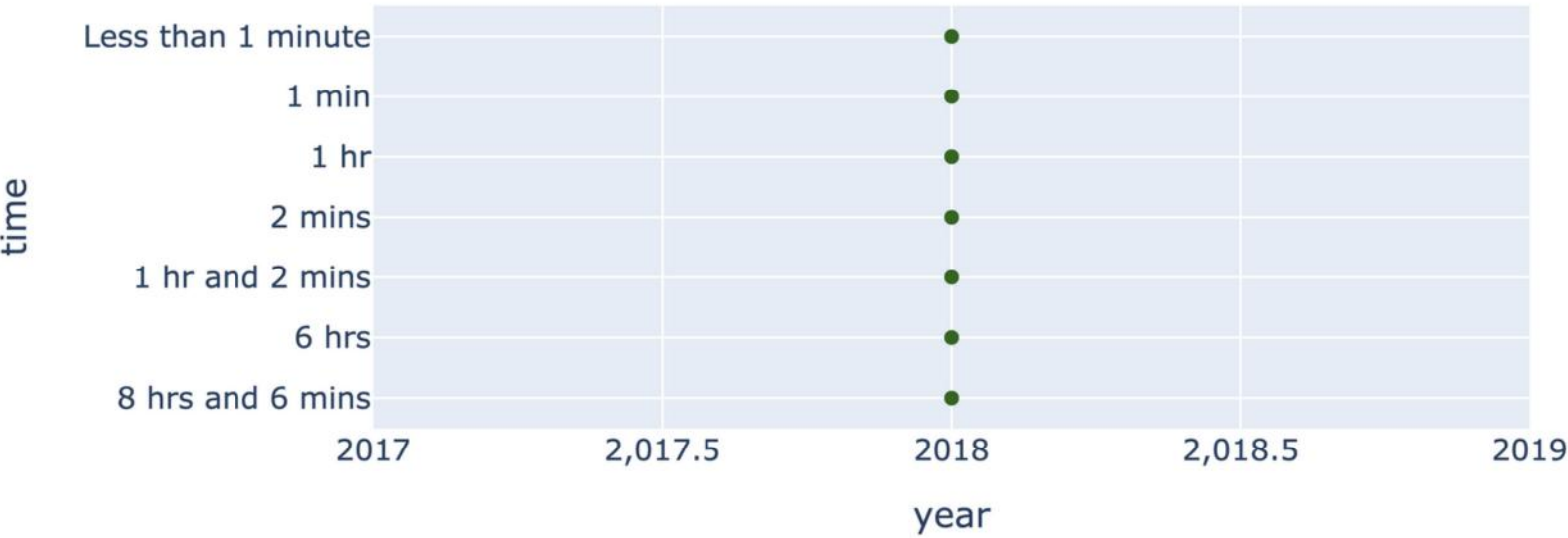


python



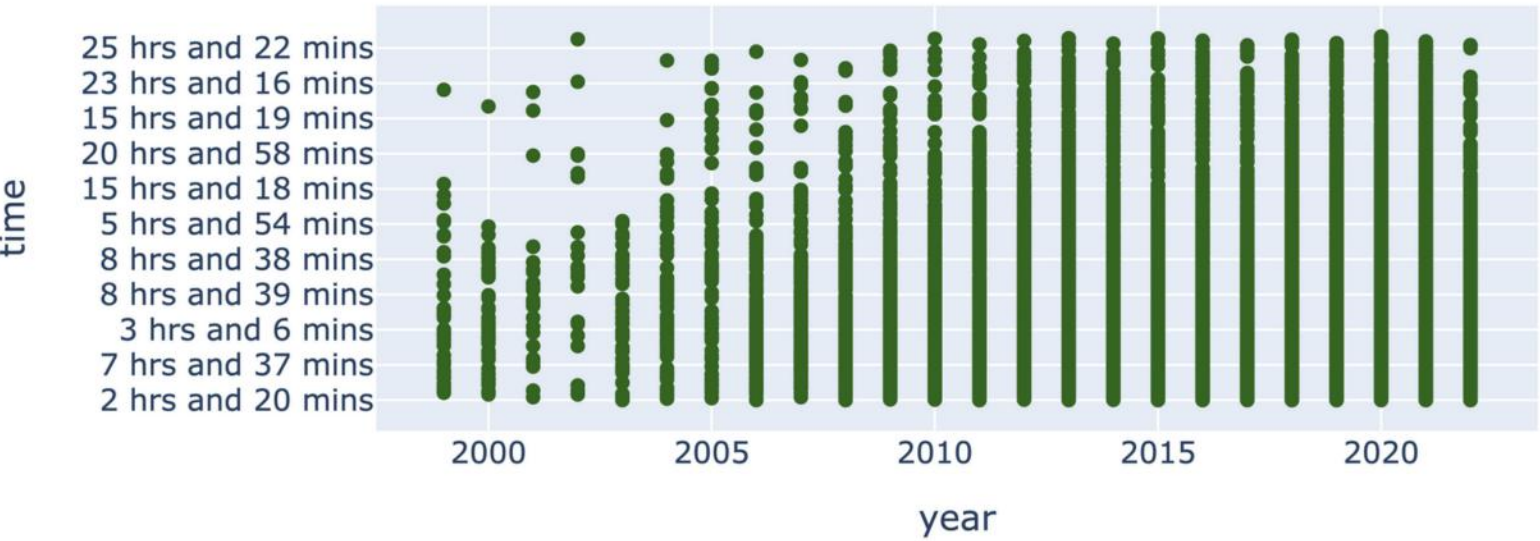
open refine

Relationship Between the Length of Audiobooks' and Year



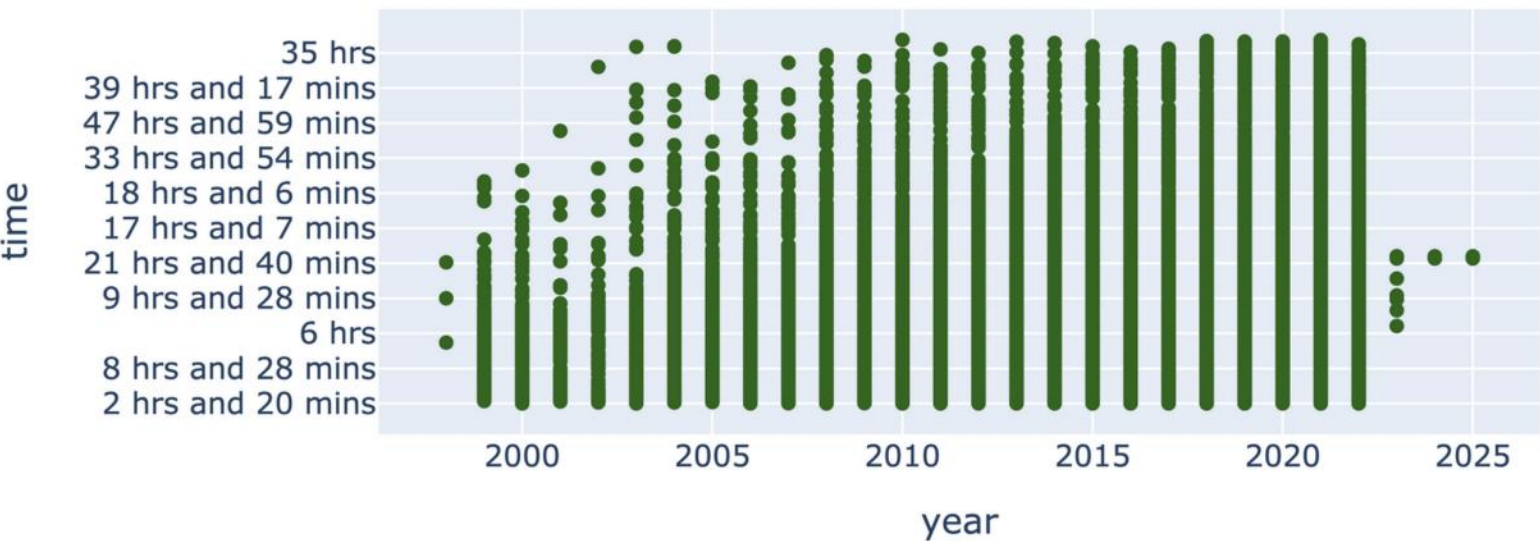
trifacta

Relationship Between the Length of Audiobooks' and Year



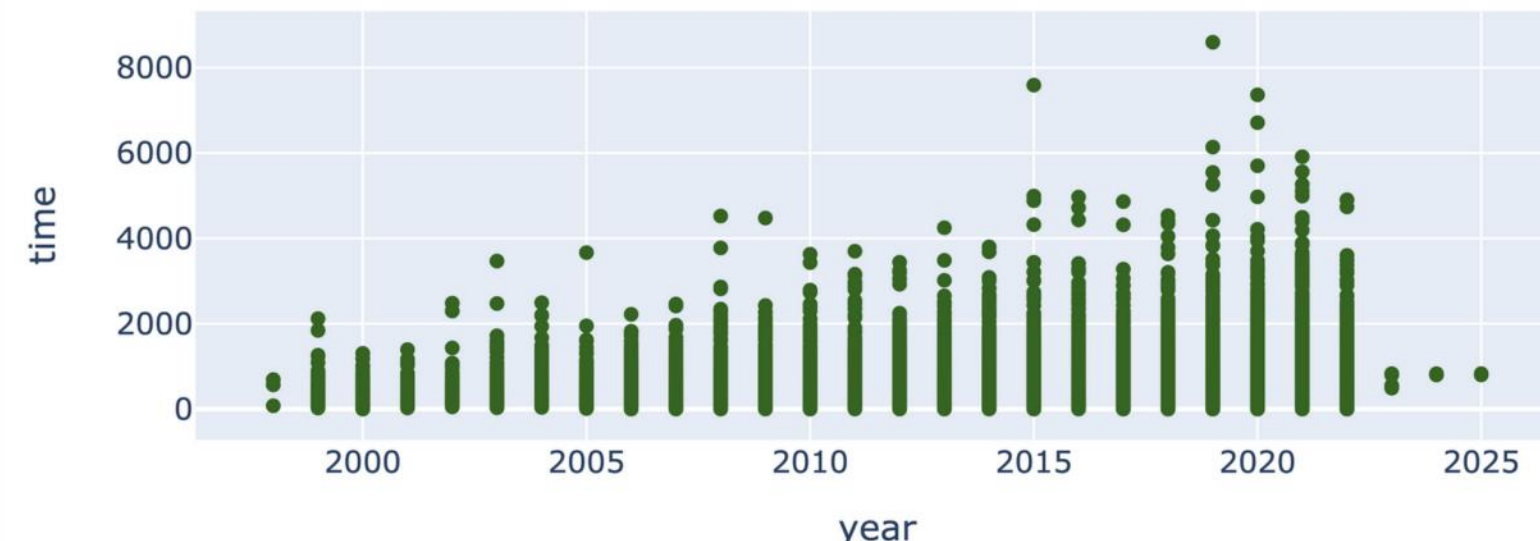
akkio

Relationship Between the Length of Audiobooks' and Year



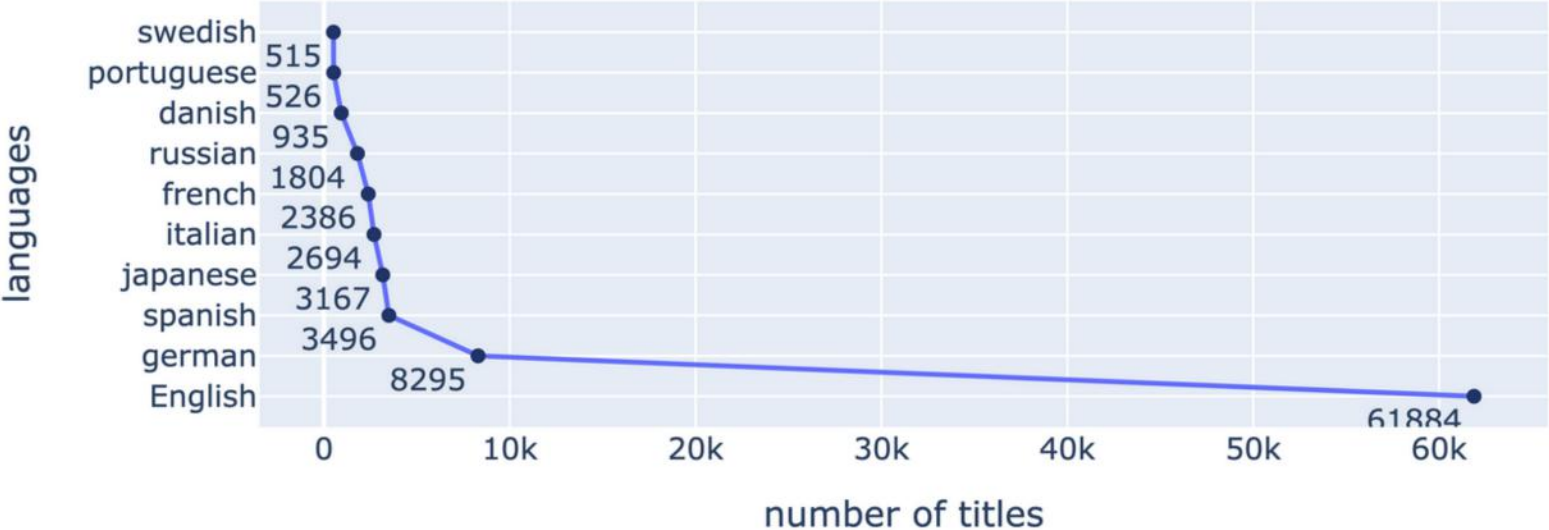
python

Relationship Between the Length of Audiobooks' and Year



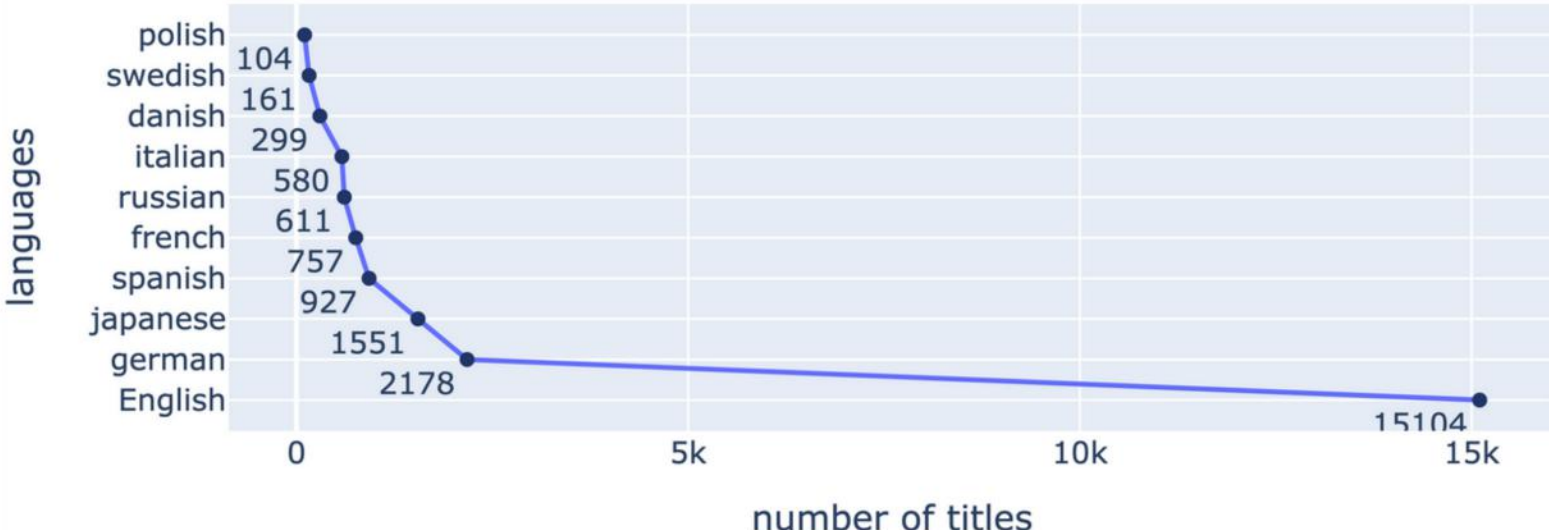
open refine

Top 10 Languages with the Highest Number of Audiobook Titles



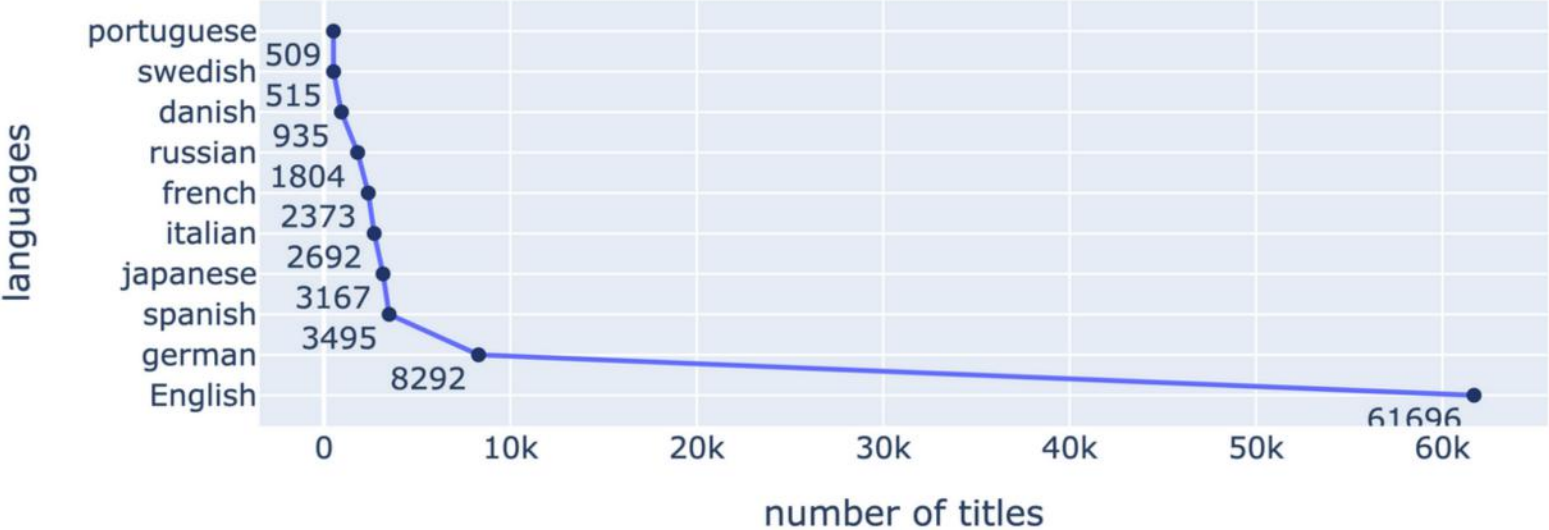
trifacta

Top 10 Languages with the Highest Number of Audiobook Titles



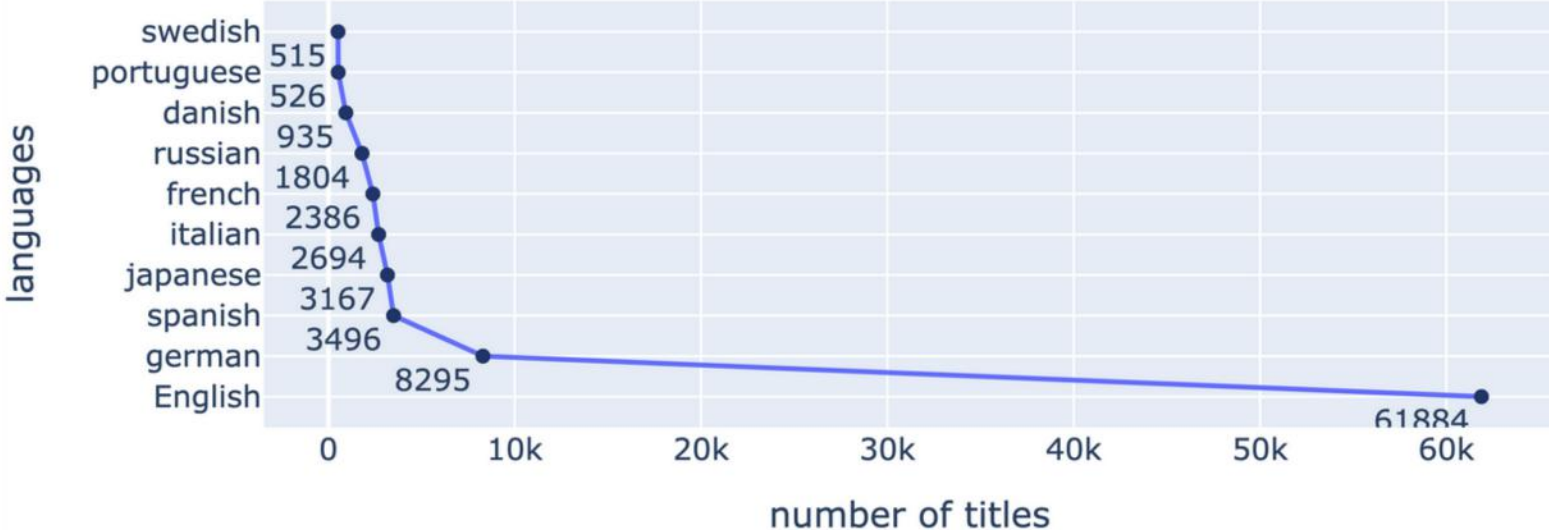
akkio

Top 10 Languages with the Highest Number of Audiobook Titles

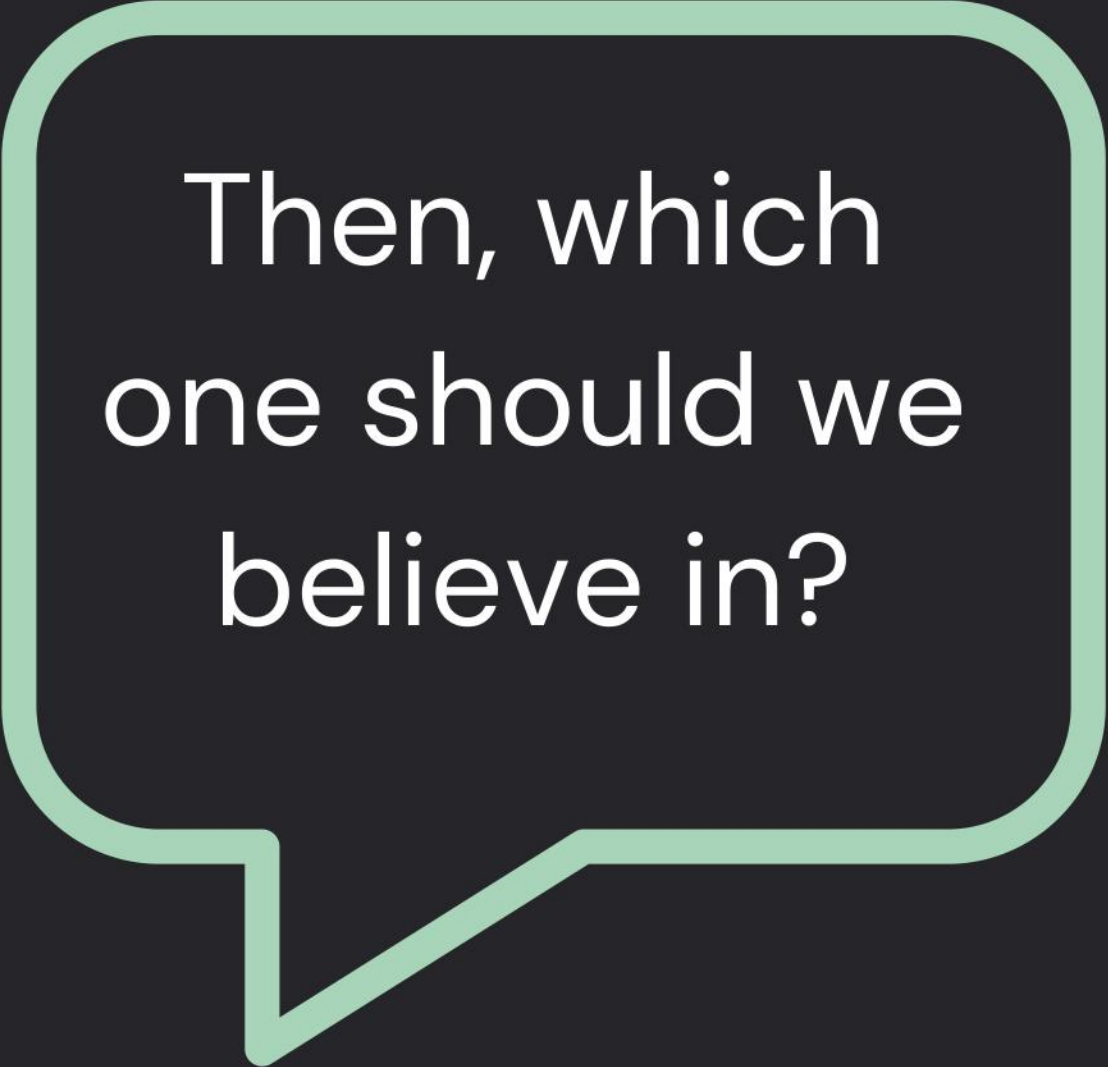


python


Top 10 Languages with the Highest Number of Audiobook Titles



Key Takeaways



Then, which
one should we
believe in?



Can I trust
this?

Data cleaning is iterative:

It's an iterative process, where I assess big problems, I devise a fix and re-evaluate. It is dirty work.

Concern: Overfitting

Evaluating data cleaning is ad-hoc:

How do you determine whether the data is sufficiently clean to trust the analysis?

We usually do not do rigorous validation of data cleaning. We typically clean our data until the desired analytics works without error.

Concern: No proper ground truth to validate results especially by the D.A community, hence issues with reproducibility rise.

P.O.A:

- Conducted a survey in 2015
- Participants were D.As, engs, that worked directly with data
- Most of the participants responded that they used *Python/Perl or MapReduce-like frameworks* and a minority used GUI to clean data

CONCLUSION

An Iterative Process

- Knowledge of the dataset, understanding the requirements of what you need to achieve.
-

Curiosity Helps

- Just the way a good data analyst is curious enough to ask the right kind of questions, that level of curiosity should also be maintained in understanding the data before proceeding to clean or use.
-

Data Validation && Data Quality

- All over the place, but no proper methodology or solid ground truth to properly validate the cleaned data to ensure its quality. Use all tools at your disposal.
-

THANK YOU!

Acknowledgements:

Snehangsu De

APPENDIX

Open Refine

Cluster and edit column "name"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

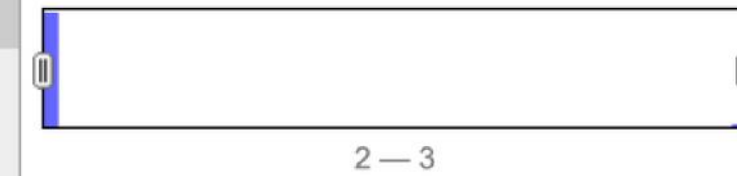
Method

Keying function

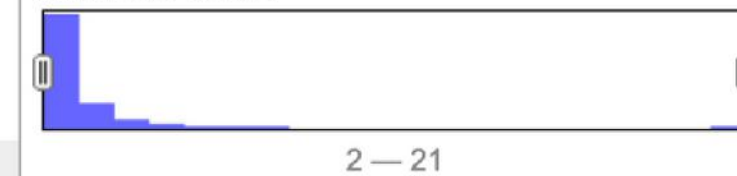
320 clusters found

Cluster size	Row count	Values in cluster
3	3	<ul style="list-style-type: none">All for OneAll for One!One for All
2	2	<ul style="list-style-type: none">How Not to DieHow not to die
2	2	<ul style="list-style-type: none">Tu primer cerebro (no está en tu cabeza)Tu primer cerebro. No está en tu cabeza
2	2	<ul style="list-style-type: none">Robinson CrusoeRobinson crusoé
2	2	<ul style="list-style-type: none">Family Inc.Family, Inc.
2	2	<ul style="list-style-type: none">Plato's RepublicPlato's Republic

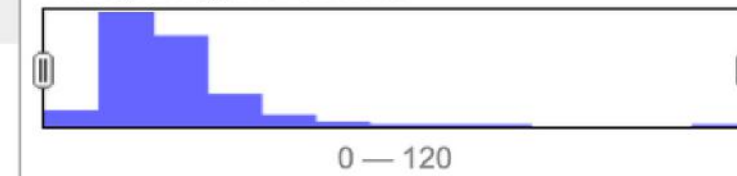
Choices in cluster



Rows in cluster



Average length of choices



Length variance of choices



Select all

Deselect all

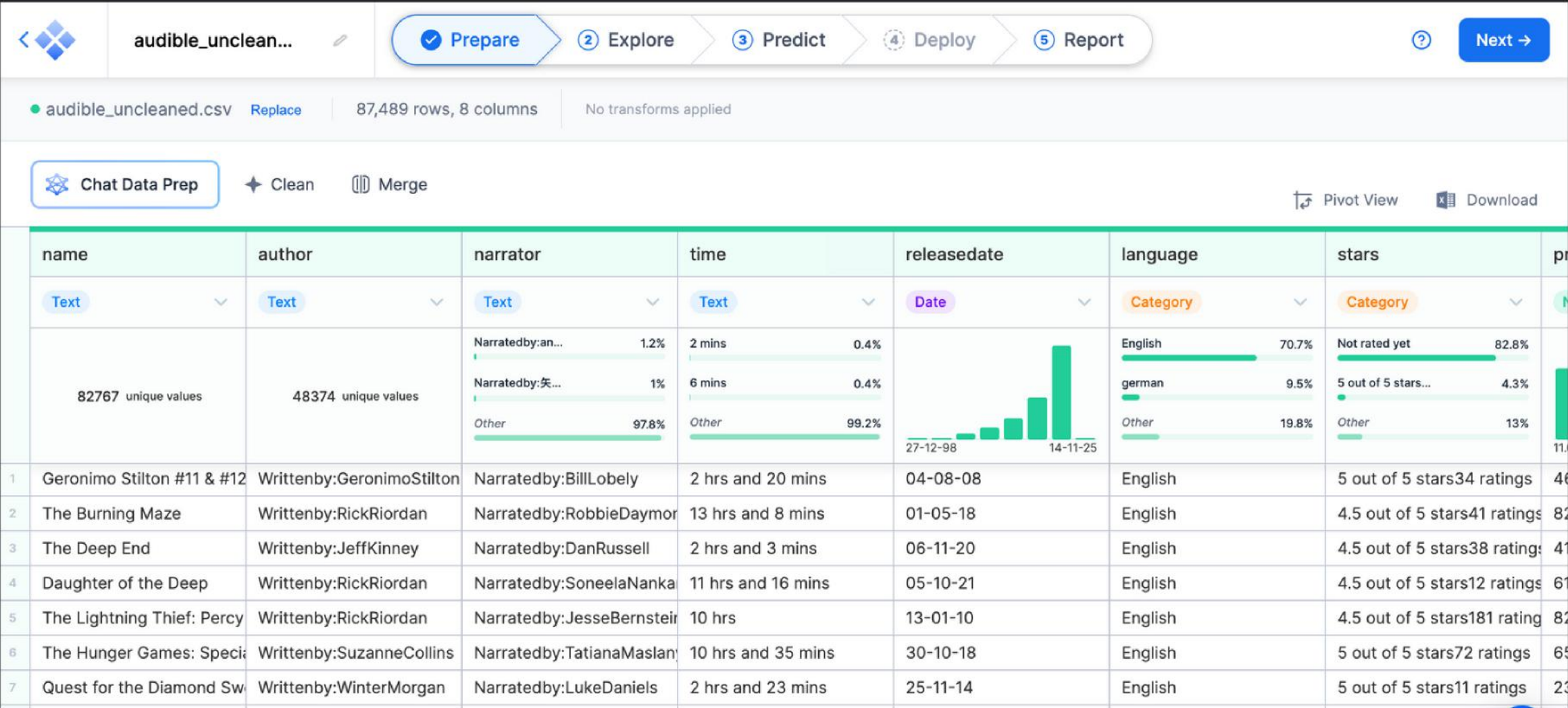
Export clusters

Merge selected & re-cluster

Merge selected & Close

Close

Akkio



Akkio

✦ Clean



- ☒ **Standardize Date Columns**
Convert all date columns to ISO 8601 standard format.
- ☒ **Remove Unexpected Nulls**
Remove rows with null values for columns that are at least 99% filled in.
- ☒ **Replace Excess Categories with "Other"**
Replace values in categorical columns that are not in the top 32 most common values with "Other".
- ☒ **Remove Constant Columns**
Remove columns that have the same value for every row.












☒ **Remove Most Unusable Numerical Columns**

Preview →

Trifacta

SAMPLE RESULTS ⓘ / Input Data (1) / Output									
22,836 Rows 8 of 8 Columns ▾									
📊 ⚙️ 🏠 ☰ ⋮ 🔗									
<	Row ID	A ^B _C name	A ^B _C author	A ^B _C narrator	A ^B _C time	A ^B _C releasedate	A ^B _C language	A ^B _C	A ^B _C
0 Messages	1	Geronimo Stilton #11 & #12	Writtenby:GeronimoStilton	Narratedby:BillLobely	2 hrs and 20 mins	04-08-08	English	5 o	
Output	2	The Burning Maze	Writtenby:RickRiordan	Narratedby:RobbieDaymond	13 hrs and 8 mins	01-05-18	English	4.5	
	3	The Deep End	Writtenby:JeffKinney	Narratedby:DanRussell	2 hrs and 3 mins	06-11-20	English	4.5	
	4	Daughter of the Deep	Writtenby:RickRiordan	Narratedby:SoneelaNankani	11 hrs and 16 mins	05-10-21	English	4.5	
	5	The Lightning Thief: Percy	Writtenby:RickRiordan	Narratedby:JesseBernstein	10 hrs	13-01-10	English	4.5	
	6	The Hunger Games: Special Edition	Writtenby:SuzanneCollins	Narratedby:TatianaMaslany	10 hrs and 35 mins	30-10-18	English	5 o	
	7	Quest for the Diamond Sword	Writtenby:WinterMorgan	Narratedby:LukeDaniels	2 hrs and 23 mins	25-11-14	English	5 o	
	8	The Dark Prophecy	Writtenby:RickRiordan	Narratedby:RobbieDaymond	12 hrs and 32 mins	02-05-17	English	5 o	
	9	Merlin Mission Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	10 hrs and 56 mins	02-05-17	English	5 o	
	10	The Tyrant's Tomb	Writtenby:RickRiordan	Narratedby:RobbieDaymond	13 hrs and 22 mins	24-09-19	English	5 o	
	11	The Titan's Curse: Percy Jackson,	Writtenby:RickRiordan	Narratedby:JesseBernstein	8 hrs and 48 mins	14-01-10	English	4.5	
	12	Magic Tree House Collection:	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	5 hrs and 23 mins	24-08-11	English	5 o	
	13	Magic Tree House Collection:	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	6 hrs and 1 min	27-09-11	English	5 o	
	14	Magnus Chase and the Ship of the	Writtenby:RickRiordan	Narratedby:MichaelCrouch	12 hrs and 58 mins	03-10-17	English	5 o	
	15	Northern Lights	Writtenby:PhilipPullman	Narratedby:PhilipPullman,fullcast,R	11 hrs and 55 mins	24-06-21	English	4 o	
	16	Geronimo Stilton #13 and #14	Writtenby:GeronimoStilton	Narratedby:BillLobley	2 hrs and 25 mins	08-02-08	English	4.5	
	17	Magic Tree House Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	5 hrs and 4 mins	26-12-04	English	5 o	
	18	Exile	Writtenby:ShannonMessenger	Narratedby:CaitlinKelly	14 hrs and 41 mins	06-11-18	English	5 o	
	19	Merlin Mission Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	10 hrs and 18 mins	02-05-17	English	5 o	
	20	Neverseen	Writtenby:ShannonMessenger	Narratedby:CaitlinKelly	16 hrs and 42 mins	06-11-18	English	5 o	
	21	The Tower of Nero	Writtenby:RickRiordan	Narratedby:RobbieDaymond	12 hrs and 12 mins	06-10-20	English	5 o	

Trifacta

 name			 time
Geronimo Stilton #11 & #12	 Rename  Change Data Type  Remove		by:BillLobely 2 hrs and 20 mins
The Burning Maze			by:RobbieDaymond 13 hrs and 8 mins
The Deep End			by:DanRussell 2 hrs and 3 mins
Daughter of the Deep	 Filter  Sort		by:SoneelaNankani 11 hrs and 16 mins
The Lightning Thief: Percy			by:JesseBernstein 10 hrs
The Hunger Games: Special Edition	 Clean Up Data  Split Column		by:TatianaMaslany 10 hrs and 35 mins
Quest for the Diamond Sword			23 mins
The Dark Prophecy			d 32 mins
Merlin Mission Collection	Writtenby:RickRiordan Narrated	 Modify Case  Remove Characters	d 56 mins
The Tyrant's Tomb			d 22 mins
The Titan's Curse: Percy Jackson,	Writtenby:RickRiordan	Narratedby:JesseBernstein	8 hrs and 48 mins
Magic Tree House Collection:	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	5 hrs and 23 mins
Magic Tree House Collection:	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	6 hrs and 1 min
Magnus Chase and the Ship of the	Writtenby:RickRiordan	Narratedby:MichaelCrouch	12 hrs and 58 mins

Trifacta

Find Tool

In/Out

workflow_20230730_013247

Run Job

DATA CLEANSING (2)

SAMPLE RESULTS / Data Cleansing (2) / Output

22,836 Rows

8 of 8 Columns

Row ID	name	time	releasedate	language	stars		
1	Geronimo Stilton #11 & #12	2 hrs and 20 mins	04-08-08	English	5 out of 5 stars34 ratings		
2	The Burning Maze	13 hrs and 8 mins	01-05-18	English	4.5 out of 5 stars41 ratings		
3	The Deep End	2 hrs and 3 mins	06-11-20	English	4.5 out of 5 stars38 ratings		
4	Daughter of the Deep	11 hrs and 16 mins	05-10-21	English	4.5 out of 5 stars12 ratings		
5	The Lightning Thief: Percy	10 hrs	13-01-10	English	4.5 out of 5 stars181 ratings		
6	The Hunger Games: Special	10 hrs and 35 mins	30-10-18	English	5 out of 5 stars72 ratings		
7	Quest for the Diamond Sword		25-11-14	English	5 out of 5 stars11 ratings		
8	The Dark Prophecy		02-05-17	English	5 out of 5 stars50 ratings		
9	Merlin Mission Collection			English	5 out of 5 stars5 ratings		
10	The Tyrant's Tomb	Writtenby:RickRiordan	Narratedby:R		5 out of 5 stars58 ratings		
11	The Titan's Curse: Percy	Writtenby:RickRiordan	Narratedby:JesseBernstein	8 hrs and 48 mins	English	4.5 out of 5 stars130 ratings	
12	Magic Tree House Collection:	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	5 hrs and 23 mins	24-08-11	English	5 out of 5 stars6 ratings
13	Magic Tree House Collection:	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	6 hrs and 1 min	27-09-11	English	5 out of 5 stars7 ratings
14	Magnus Chase and the Ship of	Writtenby:RickRiordan	Narratedby:MichaelCrouch	12 hrs and 58 mins	03-10-17	English	5 out of 5 stars41 ratings
15	Northern Lights	Writtenby:PhilipPullman	Narratedby:PhilipPullman,fullc	11 hrs and 55 mins	24-06-21	English	4 out of 5 stars2 ratings
16	Geronimo Stilton #13 and #14	Writtenby:GeronimoStilton	Narratedby:BillLobley	2 hrs and 25 mins	08-02-08	English	4.5 out of 5 stars33 ratings
17	Magic Tree House Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	5 hrs and 4 mins	26-12-04	English	5 out of 5 stars7 ratings
18	Exile	Writtenby:ShannonMessenger	Narratedby:CaitlinKelly	14 hrs and 41 mins	06-11-18	English	5 out of 5 stars20 ratings
19	Merlin Mission Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	10 hrs and 18 mins	02-05-17	English	5 out of 5 stars11 ratings