1

Table A Table B Schema

Books(Rank, Book\_Title , Author)

2

S Attributes

Rank, Book\_Title , Author

3

-

Books Missing a rank

0/2500

0%

Books missing a title in Table A

1/2500

0.0004 %

Books Missing an Author in Table A

9/2500

0.0036%

-

I will search the ten problem books manually and find the missing information that is most likely a scrape error

-

Rank

Numerical

Book\_Title

Textual

Author

Textual

-

AVG Book Name Len

19.1092

Longest Book Name

91

We Wish to Inform You That Tomorrow We Will Be Killed With Our Families: Stories from Rwanda

Shortest Book Name

1

I

Avg Author Name Len

14.154618473895582

Max Author Name

47

Marie de Rabutin-Chantal marquise de Sévigné

Min Author Name

3

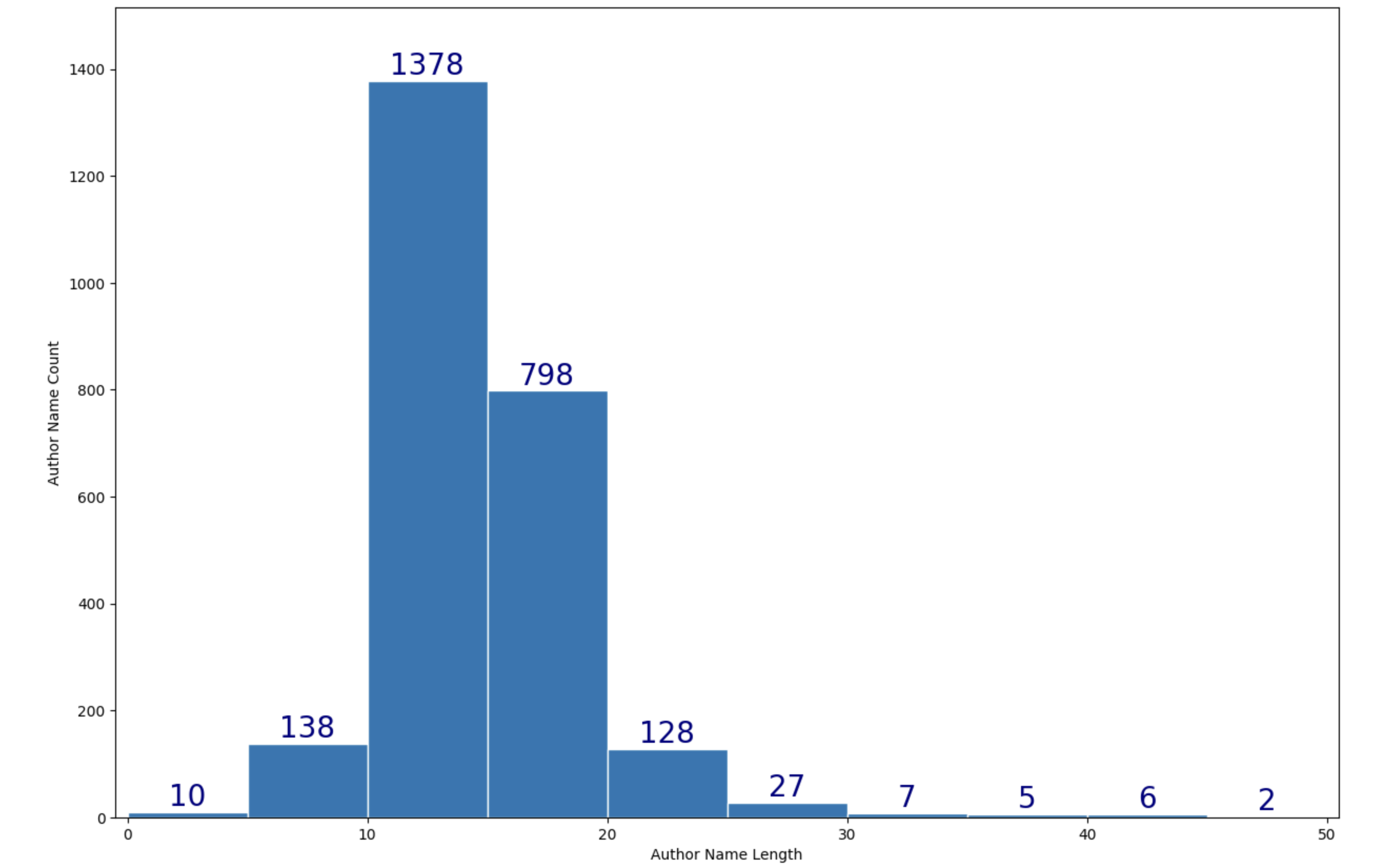
Mao

Histogram Of Author Name Length

A blue bar graph with numbers

Description automatically generated

Histogram of Book Title Name Length



Attribute Formats

Rank is numerical and all values conform.

Book Title and Author is text based.

Some Authors and Books are not in English this might need to be translated for a common language

Attribute Synonyms

No synonyms are used. All attributes follow the same author title rank format.

Attribute Sprinkle

No attributes are sprinkled the title and author are clearly separated.

Attribute Data Quality

Some author names match so it will be hard to discern if this is the same author or a different person due to a lack of attributes.

4

-

Software Used

Excel

Python MatPLotLib

to make Histograms

Pyhton pyperclip

to read in text from excel due to format issues when reading in csv

Jupyter Labs

To run python scrips more easily and copy the output