Project 2

# Classification of Data

Logan Docherty

My data schema has 4 attributes and is as follows:

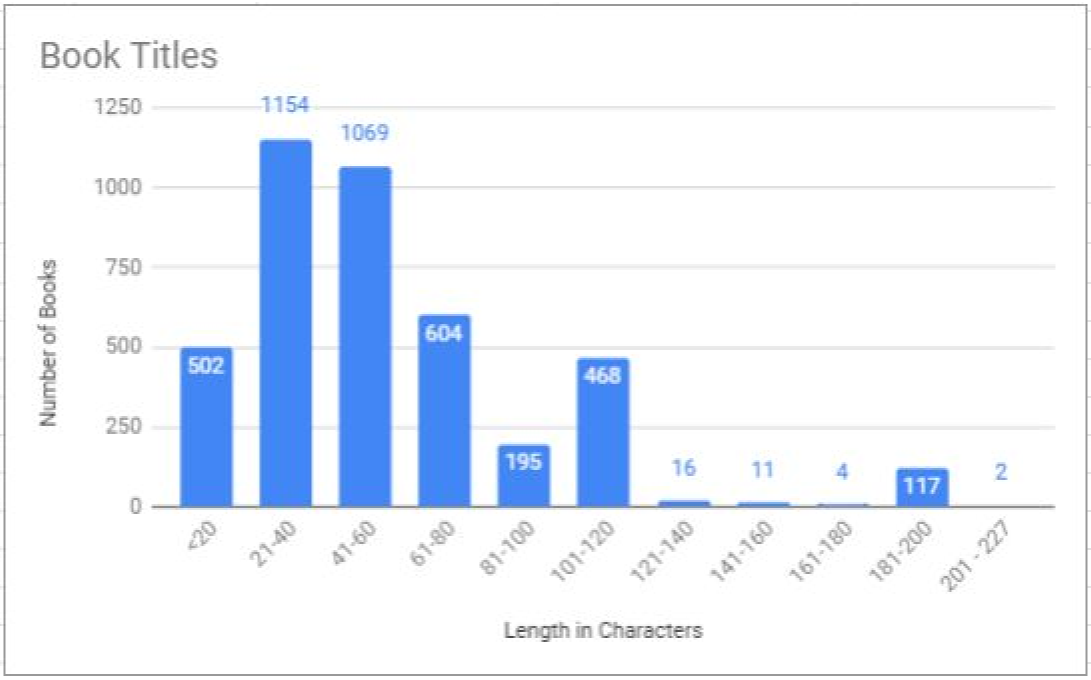
[ ID, Title, Author, Year]

We will begin with the classification of the Title attribute:

## Title: Textual attribute

Our **Average length** of title attribute is 39 characters long. You may think this seems offly long for a title. This can be explained why this is when we see the min, max, and the below chart

Our minimum length title is 6 characters long; Whereas our maximum length is 227 characters long. There is quite a difference in what is classified as a title. Some have just a simple name, while others have a name and possibly the edition or other important information to help identify the book.



For Title, we had a 100% accuracy. This meaning out of the 3975 books collected, we collected the title for each one:

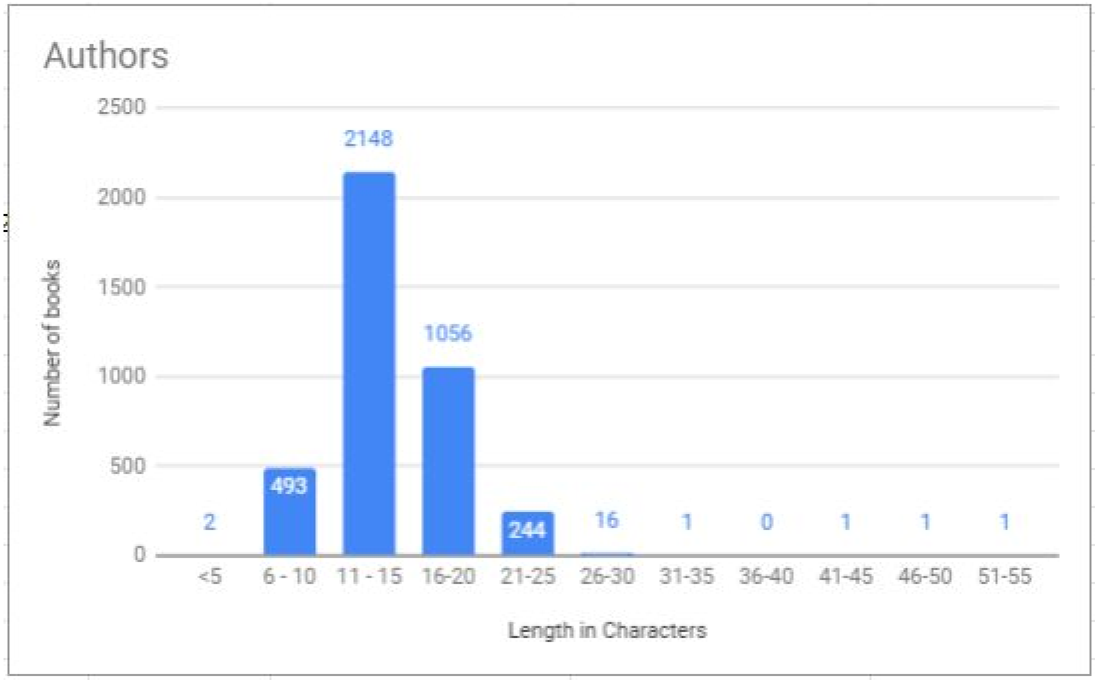
Percentage missing: 0%

Fraction: 3975/3975

Although we had no issues with filling our data for title, if we did have issues, we compare the table with another book table and look for matching authors/publishers and the year released. This would be a highly successful method (Not always perfect though) due to most authors not releasing more than one book in a year.

## Authors Attribute: Textual

Our authors attribute was also characterized by character length. The **average length** for this attribute was 16 characters long.
The maximum length peaks at 55 characters long, and our minimum length at a bottoming 4 characters long. Now if we take a look at the chart below, we can see we had some values that were potentially outliers for our data.



For Authors, we also had a 100% accuracy. This meaning out of the 3975 books collected, we collected the authors for each one:

Percentage missing: 0%
Fraction: 3975/3975

I refer to the earlier part about how we might fill in empty slots if we had them. Comparing to another book site would be able to pull authors of the book if we could look up their title and year released. I believe this would have even more success than this method would with the title.

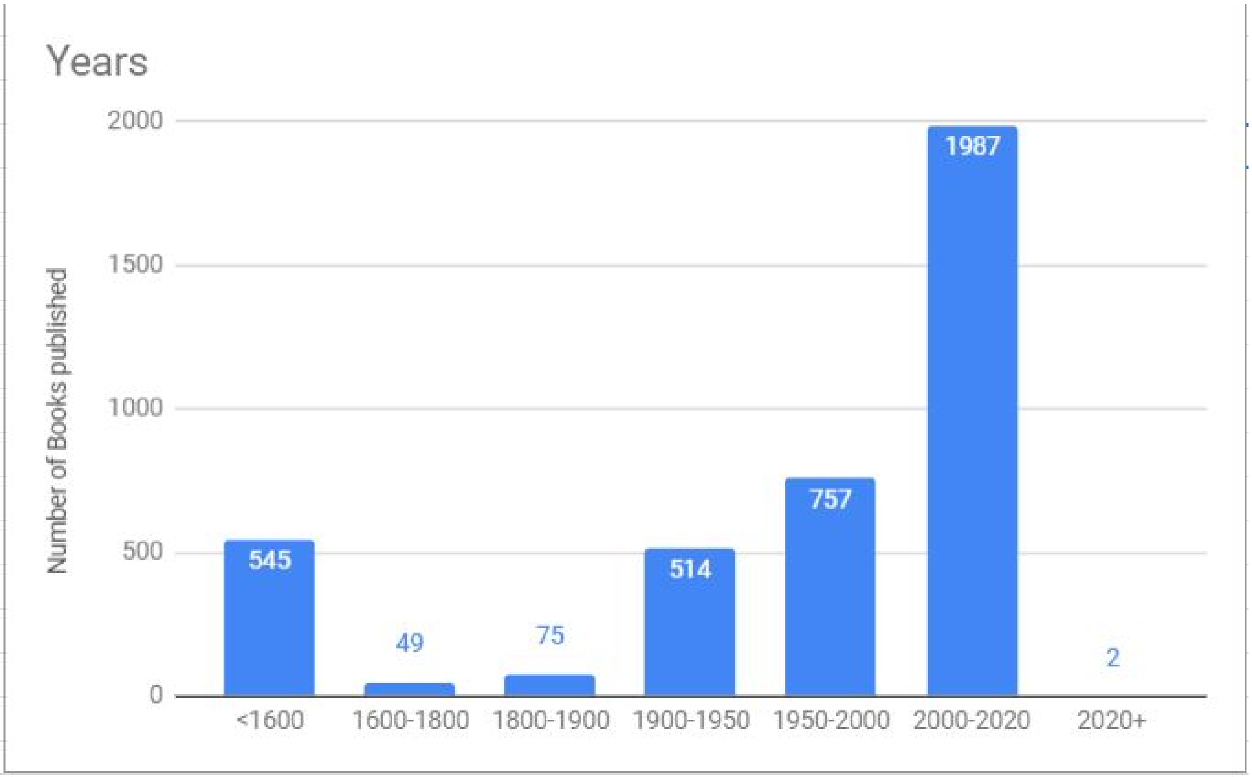
## Lasty, we have our Year attribute:  Numerical

For our year attribute I had to convert the format from string array to int format.
This did have an effect on the accuracy:

Fraction:  537/3974
Error Percent:  13.5128 % (Rounded)

There are several possible ways we could attempt to fill these values.  One way is as discussed for the other two attributes of comparing tables.  A second method, would be to fill these with the mean value, this would help not screw our data whereas a third way would be to just not use these rows at all.  Taking these rows out would help prevent misrepresentation of data.

We can see in this graph, we categorized them into year chunks that made sense.  To my surprise, we did have some outliers, such as two books being published after 2020, which is impossible.



The format when collected was to follow a four digit number.  So this would vacate the attribute for any book before year 1000.  Along with this, I mentioned before that they came in as a string array.  This meaning the scrap grabbed more than one year.  Between looking for a year format >=1000 and sometimes more than one year, this contributes to most of our errors.

Below is the code used to convert:

```
//Javascript
function Convert(text) {
 var str = String(text);
 text = str.substring(2, 6);
 var g = parseInt(text);
 if (!Number.isInteger(g / 1)) {
    return -1;
  }
 return g;
}
```

I used primarily Google Sheets to analyze the data.  When its came to calculating averages, max, mins, and conversions, I used google sheets ability to write custom functions. (Uses javascript)