When I looked at the data again, I realised I had taken the num\_pages variable out during the initial data cleaning. This was because there are books recorded as having 0 pages and others with very few page numbers. After investigating on GoodReads, it looks like these are all audio books. I think that those with page numbers > 0 are also audiobooks but have some sort of transcript/document that comes along with them. Unfortunately, without having the format available in the data, it is impossible to determine where audio books end and books begin.

There are also quite a few anthologies in the data, which tend to appear in the rows with the highest num\_pages.

In attempt to correct/unskew the data, I removed the top and bottom 5% from the num\_pages variable. It’s not perfect, but I think it gives us a more realistic picture.

For the ratings\_count, the data was highly skewed. I again removed the top and bottom 5% but there wasn’t really any scientific reason for me doing this – I just didn’t know how best to clean up the outliers. Given the skewness, I would suggest that, if we do go ahead with this, we use the median ratings\_count as our statistic and not the mean.

Looking at the scatter plot of num\_pages vs ratings\_count, I don’t see any useful pattern that would suggest there is a relationship in this data.

Despite this, I have put the page\_num variable into ranges, 51-150 pages, 151 -250 pages, 251-350 pages, etc. I’ve then taken a look at the median ratings count for each category (I selected median instead of mean because of the skewness in the data).

A screenshot of a computer

Description automatically generated

Not much of an exciting story here, except maybe that medium length (251-550) pages look to have a better reader engagement, but given there are less books in every

I’m also thinking that books that were published more recently have had less opportunity to be reviewed so I wanted to look at how the median ratings count in relation to the publication\_year.

The top 5 years based on num\_books are 2002-2006 so I’ve take a look at these to see if there is a pattern:

A graph showing the number of years

Description automatically generated with medium confidence

Again, shorter books seem to have a lower engagement and this is fairly consistent. Everything else is a bit all over the place.

I’m not feeling comfortable that the data is going to give us any useful information to ‘sell’ to publishers. Except maybe, don’t publish short books.

Have either of you had any luck or any other thoughts? I’m pretty tempted to look for another data source or to see if there is an existing visual we can use.

GoodReads launched in January 2007. Any book published pre-Jan 2007 can be lumped into one category as they have all had the same time to be reviewed.

We have data up to 2020

Published pre-GoodReads

Within first five years (2007-2011)

Within 6-10 years (2012-2016)