

Coding Temple Capstone II

Analysis of Academic Education Library Data

Kelli Michaels



Introduction

- Analyze a dataset to draw useful conclusions
- Use the skills learned in my Data Analytics course.

Data and mission

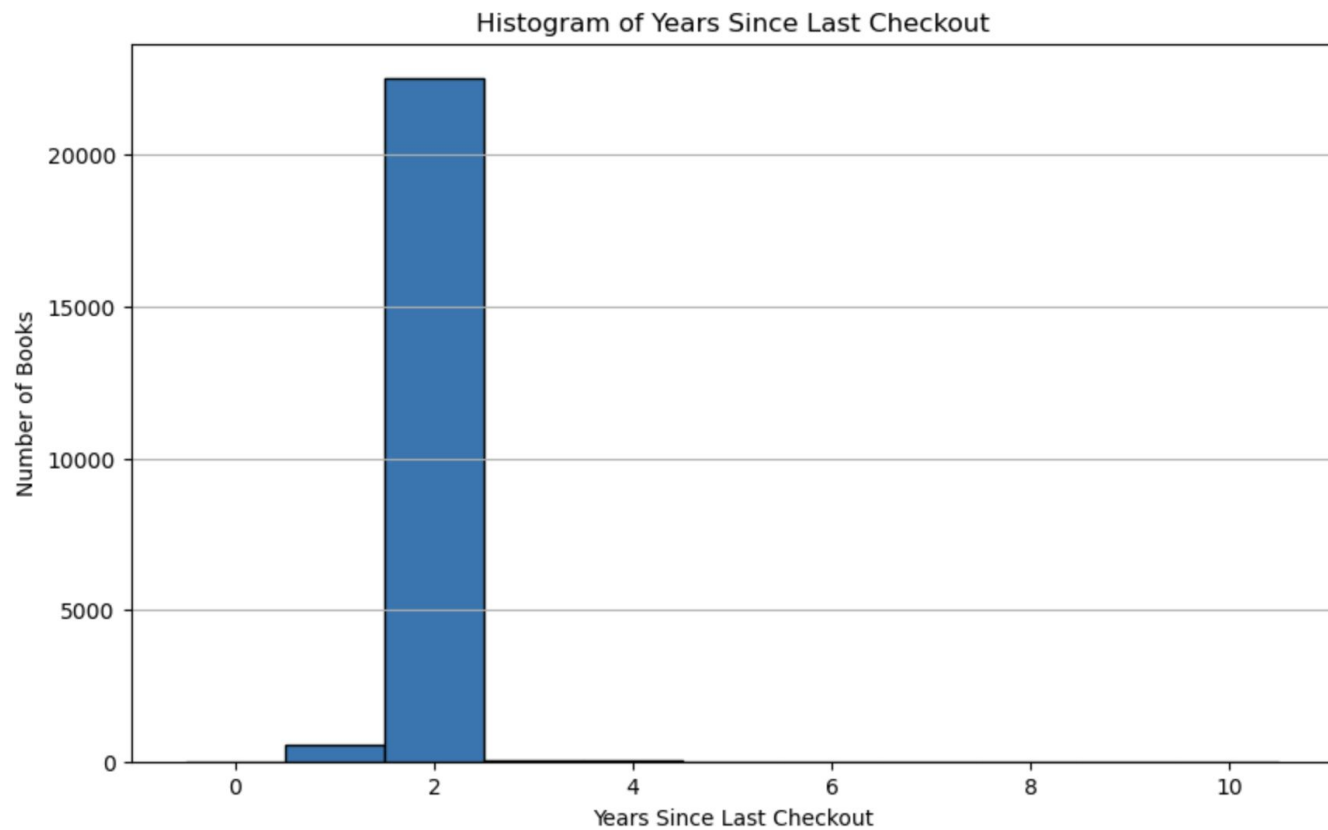
- Data set is a subset of the collection of academic library centered around education
- In this analysis I will focus of the circulation of the collection

Research Questions

What is the average amount of time since a book has been checked out? Make a histogram showing the results.

- To find this information we can subtract the last active year from the current year
- Hypothesis: I would expect there to be a wide range of differences. The peak is probably within two years but I would expect some books to go several years without check out.

Results



What percentage of books have never been checked out?

- To answer this question we check for last activity dates that are invalid or missing
- Hypothesis: I expected some small proportion of the collection to have never been checked out, maybe 1-2%

Result

- All of the books have been checked out! All of the items in the collection had a valid last activity date.

What percentage of the collection has circulated in the last 1, 5, 10, 20 years?

- This was the question I formed before investigating the data
- After the results of the last two investigations, we know that a huge proportion of books have been checked out in the last 2-3 years.
- I already can already estimate this from the histogram
- I'm going to reformulate to a different question.

~~What percentage of the collection has circulated in the last 1, 5, 10, 20 years?~~

What book went the longest without being checked out?

- We already learned every book has been checked out
- What book spent the longest on the shelf?
- Outliers difficult to read from histogram

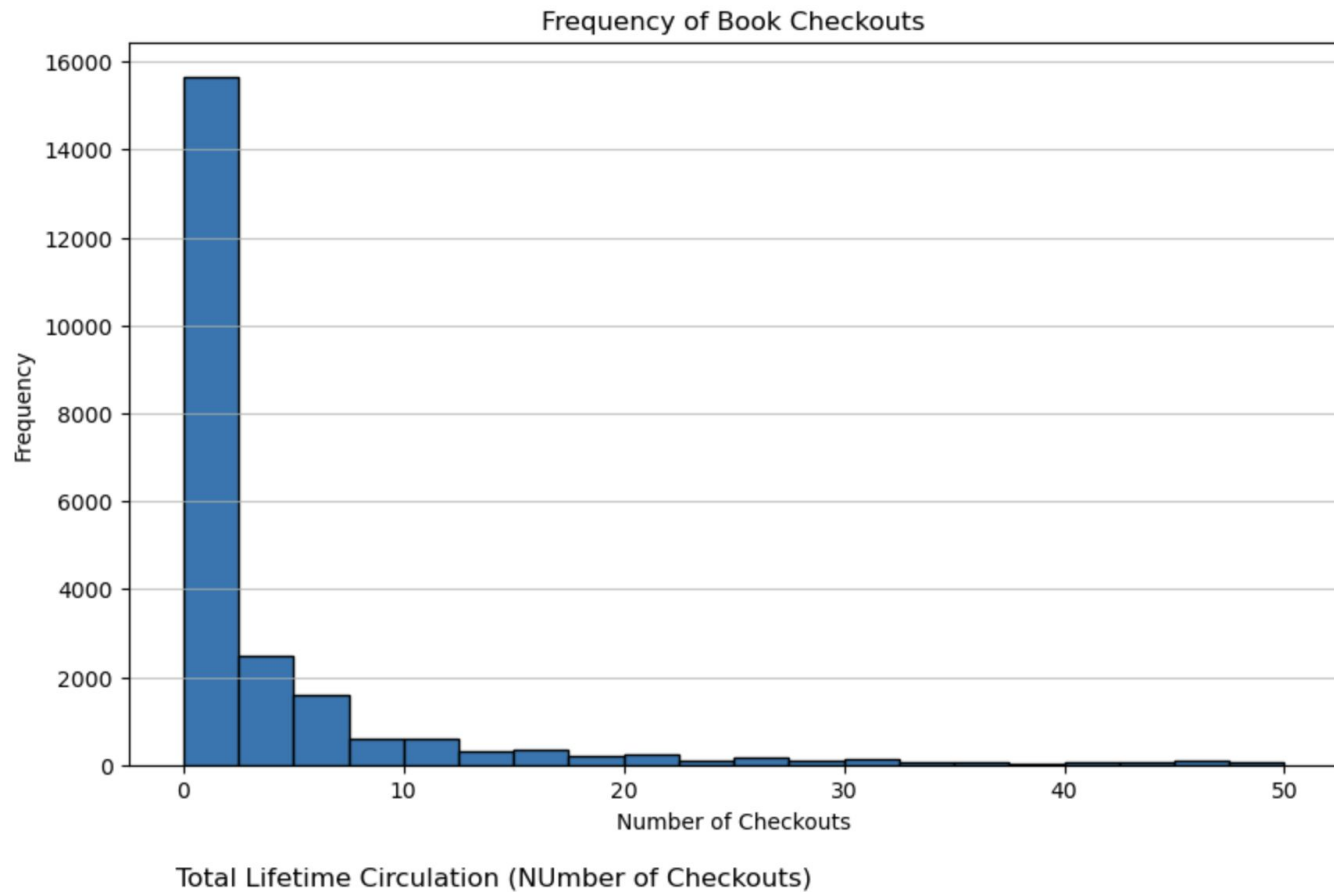
Result

- 'Success and Understanding' went 34 years without being checked out.
- Clearly an outlier

How many books have been checked out 1, 2, 3, 4, or 4+ times? Make a pie chart sorting the books into these categories.

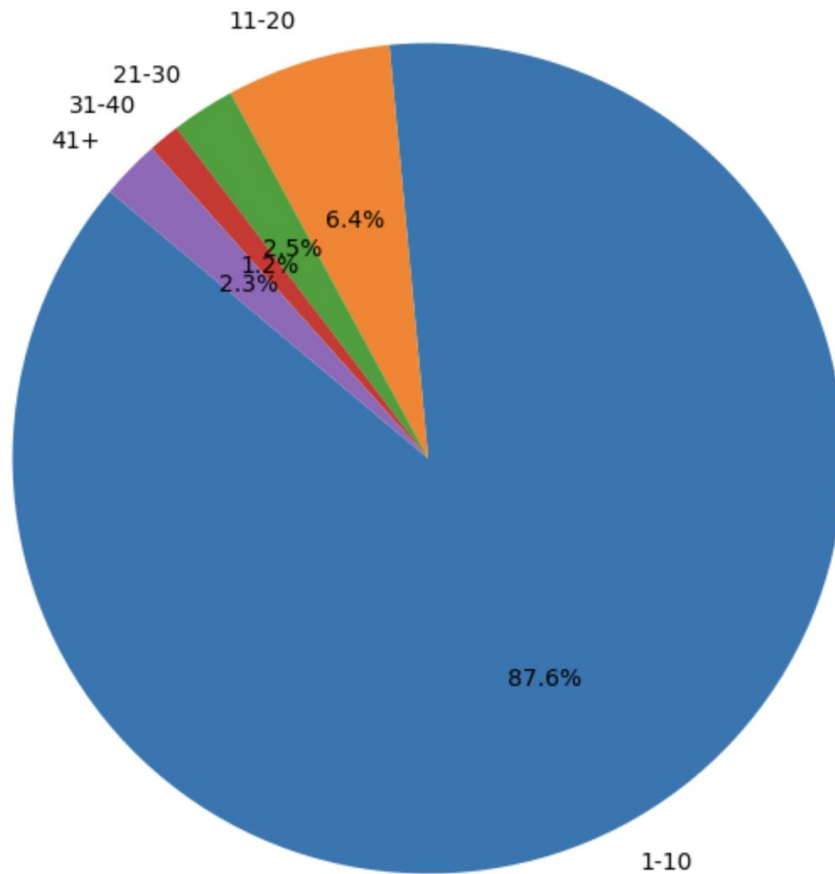
- Phrasing of question estimates that many books will have been checked out 4 or fewer times
- We will find out this estimate (and these bin ideas) were not successful
- We will also compare a histogram vs a pie chart and discuss the differences between the two plots

Result



Result

Total Lifetime Circulation (Number of Checkouts)



What are the top 10 most circulated books in the database? Do any of those 10 books have the same author? What is the average year of publication for these top 10 books?

- Hypothesis: I'm expecting this books to have a fairly recent average date of publication. I also initially expected that some of these books may have the same author.

Results

- We successfully identified the top 10 books
- The most circulated book was titled “Fateful choices : the future of the U.S. academic research enterprise : a discussion paper” which circulated 623 times.
- None of the books had the same author.
- The average year of publication for these books was 1993, much later than I predicted!

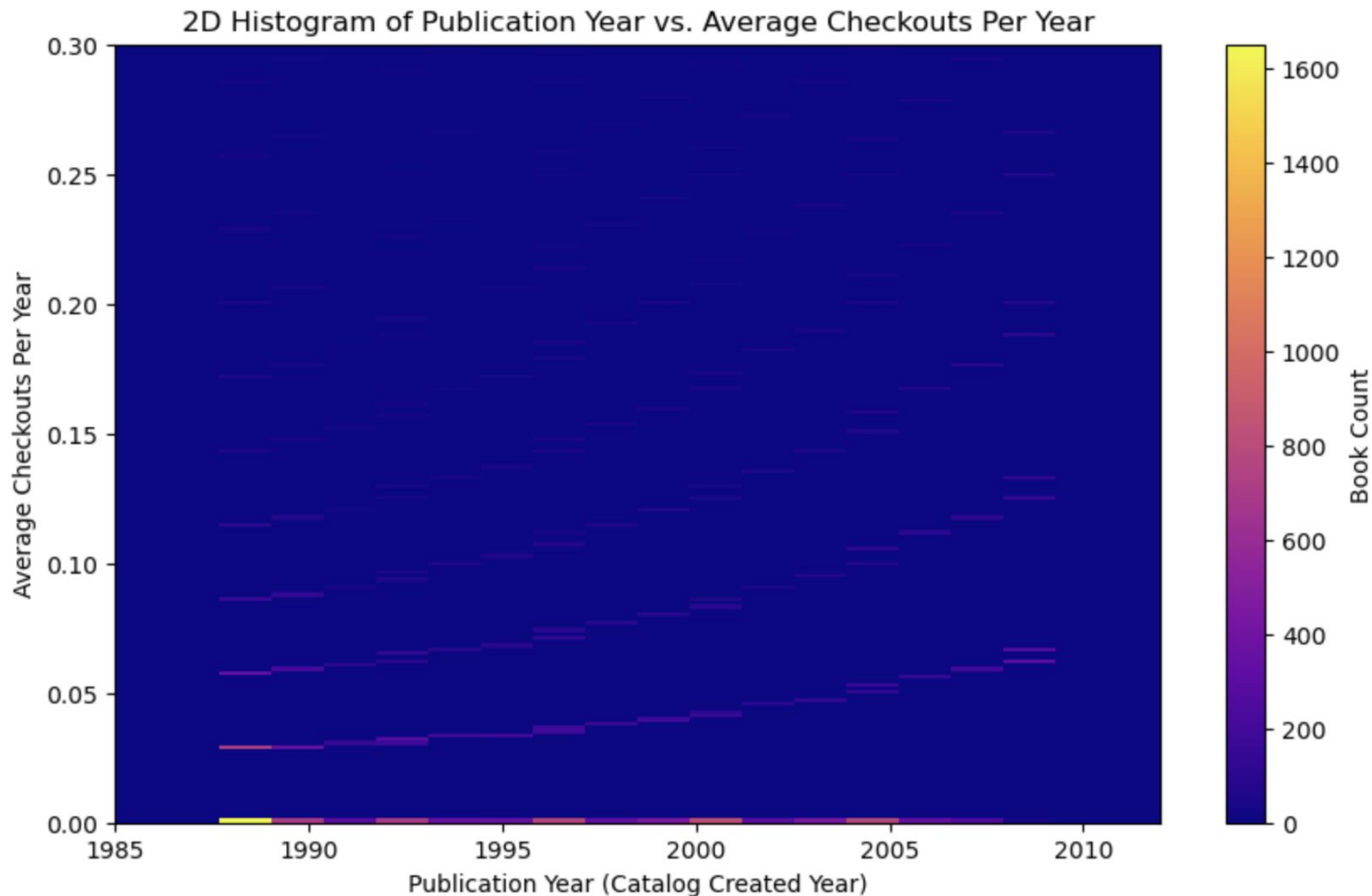
Are newer books checked out more frequently? Make a 2D plot showing cat creat year vs. avg number of check outs per year for each book. Maybe fit it with a linear regression or something.

- Broadly, I was expecting the answer here to be yes
- There were several challenges in making my plot that I did not foresee

Methods, Initial Findings

- To calculate average check outs per year, we need the total number of checkouts and the number of years the item has been in the collection.
- THESE ARE BOTH INTEGERS so we will be getting a non-continuous output (increments of $1/365$)
- This means our histogram will be “blocky”
- Very low average checkout per year, had to adjust ranges on axes a lot
- Interesting that we had a such low average given that all the books have been checked out before.

Results



Conclusions and future studies

- We learned a lot about the collection!
- We learned that all of the collection has been checked out before, and that most of the books have circulated in the last five years.
- The most circulated most has been checked out 623 times since it was added, but some books went as long as 34 years without being checked out.
- In the future, I think we could find a different way to present the 'average checkouts per year' data that would be easier to read.
- We could also repeat this analysis with other parts of the collection that have not circulated as recently.

**Thanks! For more detail check out the rest of
my documents on github**