Project 3 ALY 6000

by

Sri Ram Prabu

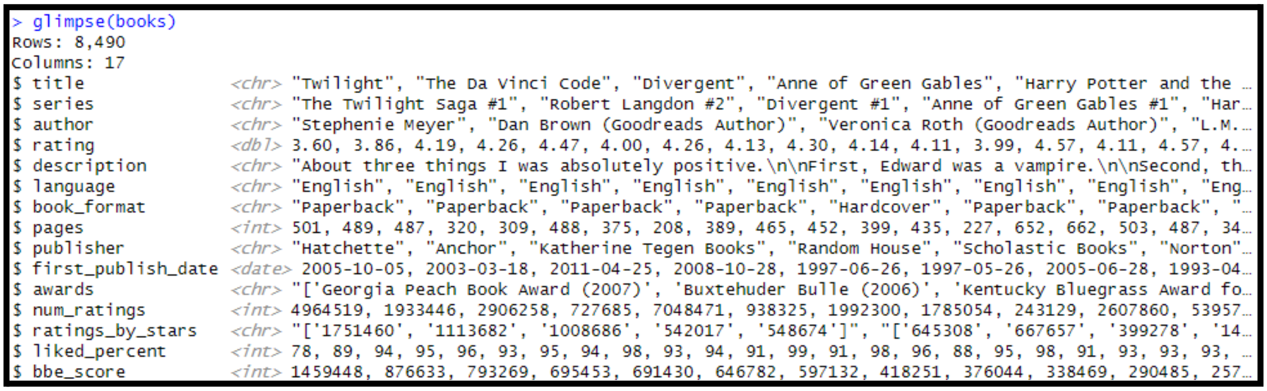
ALY6000 : Introduction to Analytics

January , 24, 2025

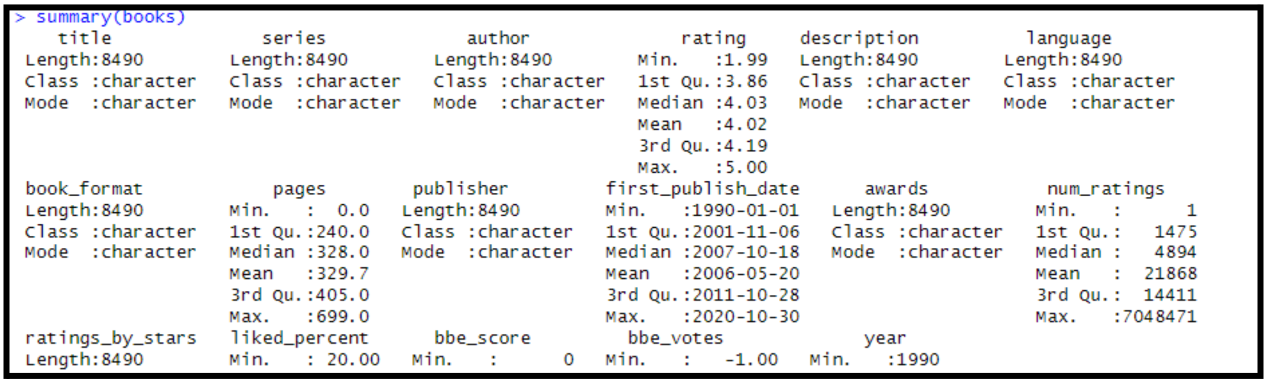
Introduction

This report presents an analysis of book data collected from Goodreads, focusing on trends in book publication, ratings, and publisher dominance between 1990 and 2020. Using R programming, various data cleaning techniques and statistical methods were applied to derive insights from the dataset. The analysis explores key statistical measures such as central tendency, dispersion, and frequency distribution to understand the characteristics of the dataset.

Data Overview



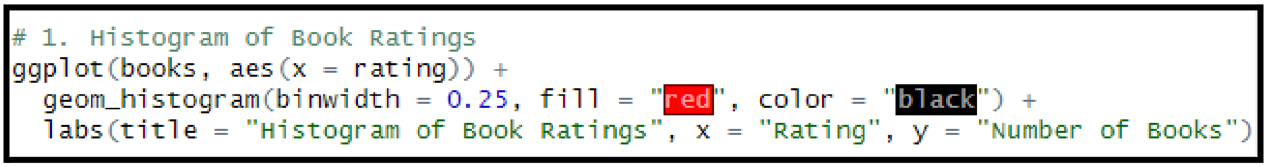
```
> glimpse(books)
Rows: 8,490
Columns: 17
$ title             <chr> "Twilight", "The Da Vinci Code", "Divergent", "Anne of Green Gables", "Harry Potter and the …
$ series            <chr> "The Twilight Saga #1", "Robert Langdon #2", "Divergent #1", "Anne of Green Gables #1", "Har…
$ author            <chr> "Stephenie Meyer", "Dan Brown (Goodreads Author)", "Veronica Roth (Goodreads Author)", "L.M.…
$ rating            <dbl> 3.60, 3.86, 4.19, 4.26, 4.47, 4.00, 4.26, 4.13, 4.30, 4.14, 4.11, 3.99, 4.57, 4.11, 4.57, 4.…
$ description       <chr> "About three things I was absolutely positive.\n\nFirst, Edward was a vampire.\n\nSecond, th…
$ language          <chr> "English", "English", "English", "English", "English", "English", "English", "English", "Eng…
$ book_format       <chr> "Paperback", "Paperback", "Paperback", "Paperback", "Hardcover", "Paperback", "Paperback", "…
$ pages             <int> 501, 489, 487, 320, 309, 488, 375, 208, 389, 465, 452, 399, 435, 227, 652, 662, 503, 487, 34…
$ publisher         <chr> "Hatchette", "Anchor", "Katherine Tegen Books", "Random House", "Scholastic Books", "Norton"…
$ first_publish_date <date> 2005-10-05, 2003-03-18, 2011-04-25, 2008-10-28, 1997-06-26, 1997-05-26, 2005-06-28, 1993-04…
$ awards            <chr> "['Georgia Peach Book Award (2007)', 'Buxtehuder Bulle (2006)', 'Kentucky Bluegrass Award fo…
$ num_ratings       <int> 4964519, 1933446, 2906258, 727685, 7048471, 938325, 1992300, 1785054, 243129, 2607860, 53957…
$ ratings_by_stars  <chr> "['1751460', '1113682', '1008686', '542017', '548674']", "['645308', '667657', '399278', '14…
$ liked_percent     <int> 78, 89, 94, 95, 96, 93, 95, 94, 98, 93, 94, 91, 99, 91, 98, 96, 88, 95, 98, 91, 93, 93, 93, …
$ bbe_score         <int> 1459448, 876633, 793269, 695453, 691430, 646782, 597132, 418251, 376044, 338469, 290485, 257…
```



```
> summary(books)
    title              series             author              rating      description          language        
 Length:8490        Length:8490        Length:8490        Min.   :1.99   Length:8490        Length:8490       
 Class :character   Class :character   Class :character   1st Qu.:3.86   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Median :4.03   Mode  :character   Mode  :character  
                                                          Mean   :4.02                                        
                                                          3rd Qu.:4.19                                        
                                                          Max.   :5.00                                        
 book_format            pages         publisher         first_publish_date      awards           num_ratings     
 Length:8490        Min.   :  0.0   Length:8490        Min.   :1990-01-01   Length:8490        Min.   :      1  
 Class :character   1st Qu.:240.0   Class :character   1st Qu.:2001-11-06   Class :character   1st Qu.:   1475  
 Mode  :character   Median :328.0   Mode  :character   Median :2007-10-18   Mode  :character   Median :   4894  
                    Mean   :329.7                      Mean   :2006-05-20                      Mean   :  21868  
                    3rd Qu.:405.0                      3rd Qu.:2011-10-28                      3rd Qu.:  14411  
                    Max.   :699.0                      Max.   :2020-10-30                      Max.   :7048471  
 ratings_by_stars   liked_percent     bbe_score         bbe_votes            year     
 Length:8490        Min.   : 20.00   Min.   :      0   Min.   :   -1.00   Min.   :1990  
```

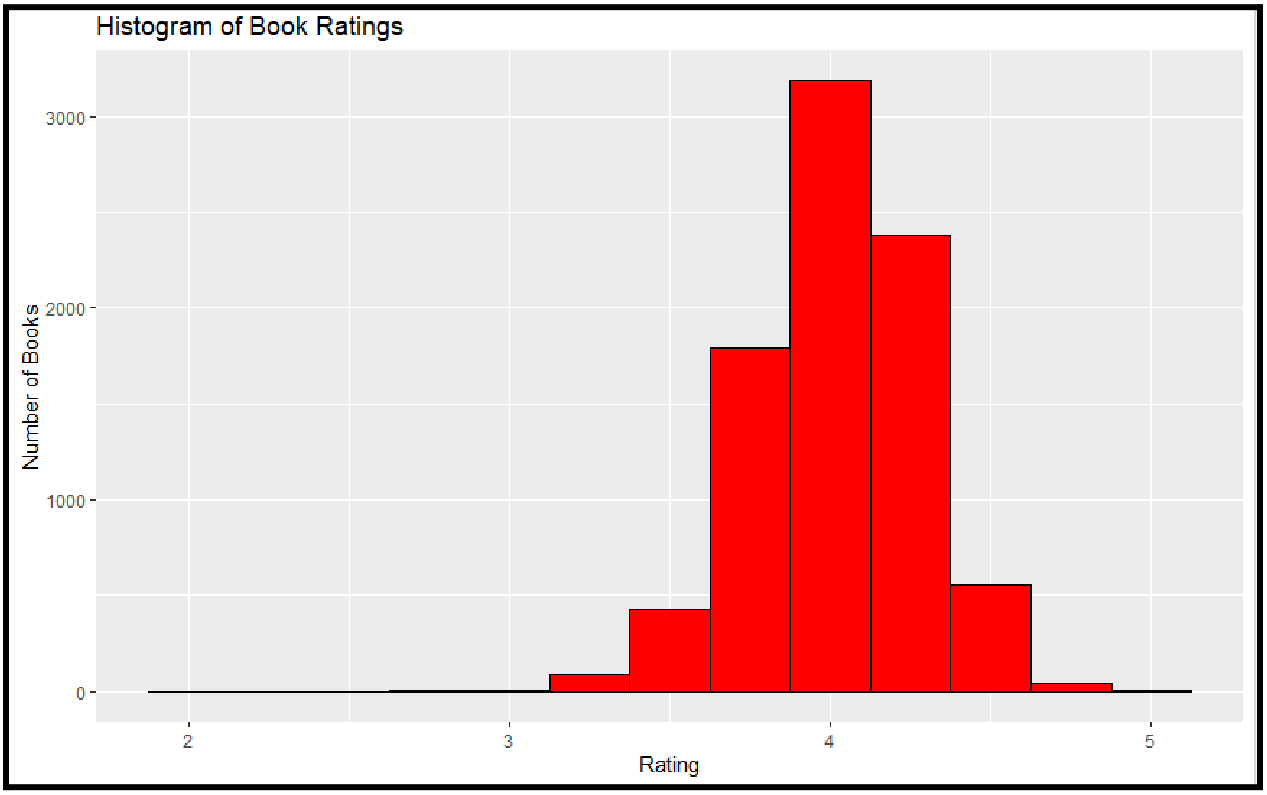
Key Findings & Visualizations

1. Rating Distribution

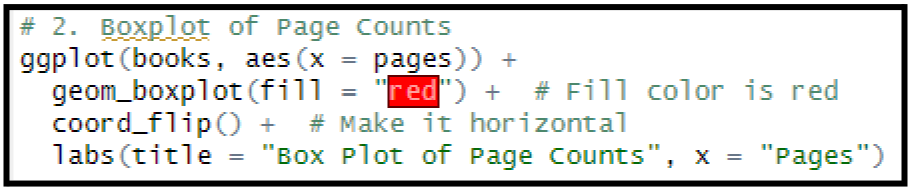


```
# 1. Histogram of Book Ratings
ggplot(books, aes(x = rating)) +
  geom_histogram(binwidth = 0.25, fill = "red", color = "black") +
  labs(title = "Histogram of Book Ratings", x = "Rating", y = "Number of Books")
```

- Observation: The majority of books have a rating between 3.5 and 4.5.
- Visualization: A histogram of book ratings was created, highlighting the frequency distribution of ratings.
- Insight: Most books are rated favorably, indicating a positive bias in user reviews.

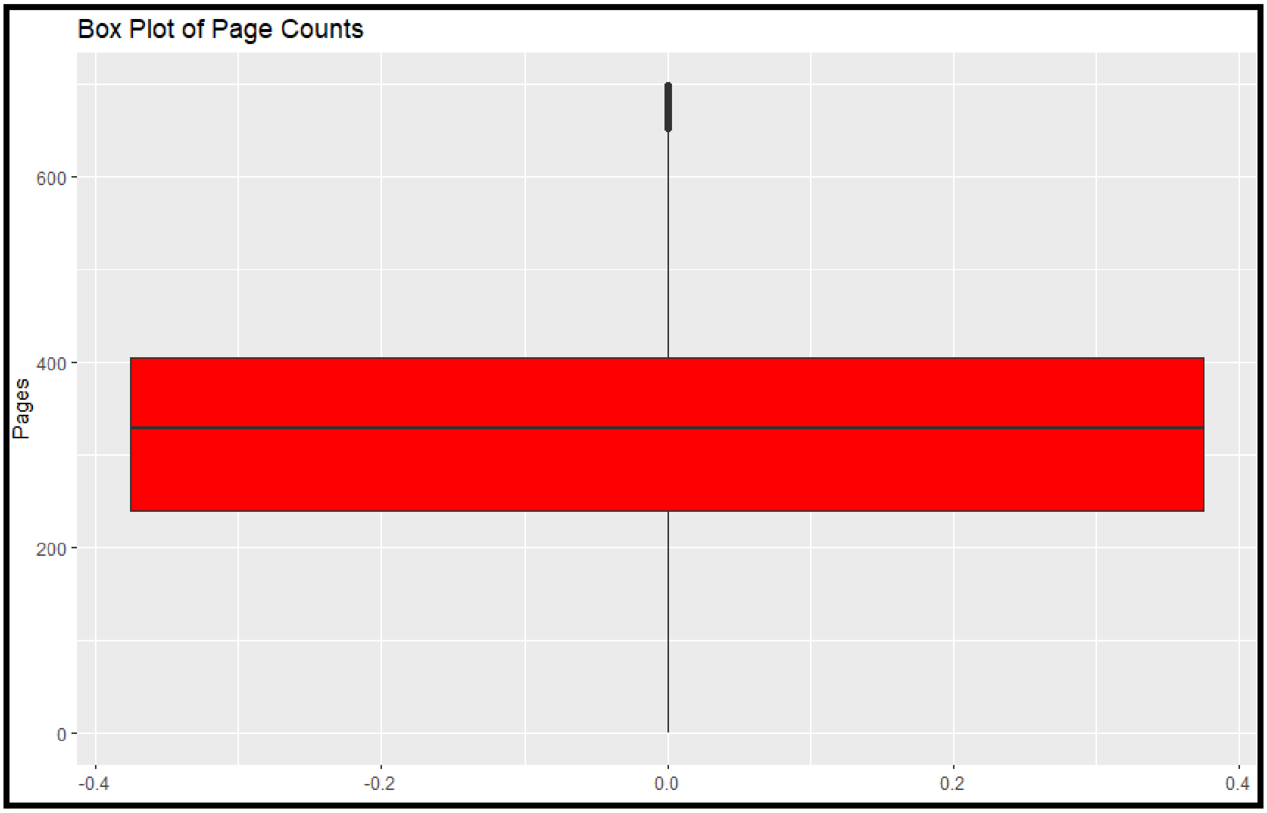


2. Page Length Distribution

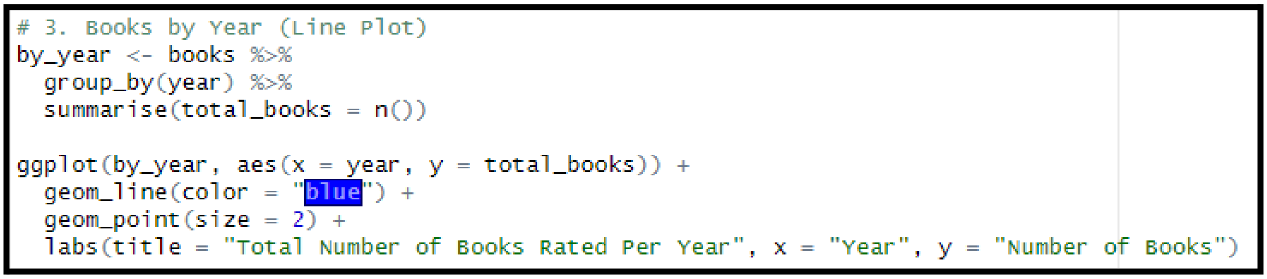


```
# 2. Boxplot of Page Counts
ggplot(books, aes(x = pages)) +
  geom_boxplot(fill = "red") +  # Fill color is red
  coord_flip() +  # Make it horizontal
  labs(title = "Box Plot of Page Counts", x = "Pages")
```

- Observation: The median book length is approximately 350 pages, with very few books exceeding 600 pages.
- Visualization: A boxplot was generated to show the distribution of page counts.
- Insight: Readers tend to prefer books that are moderate in length.



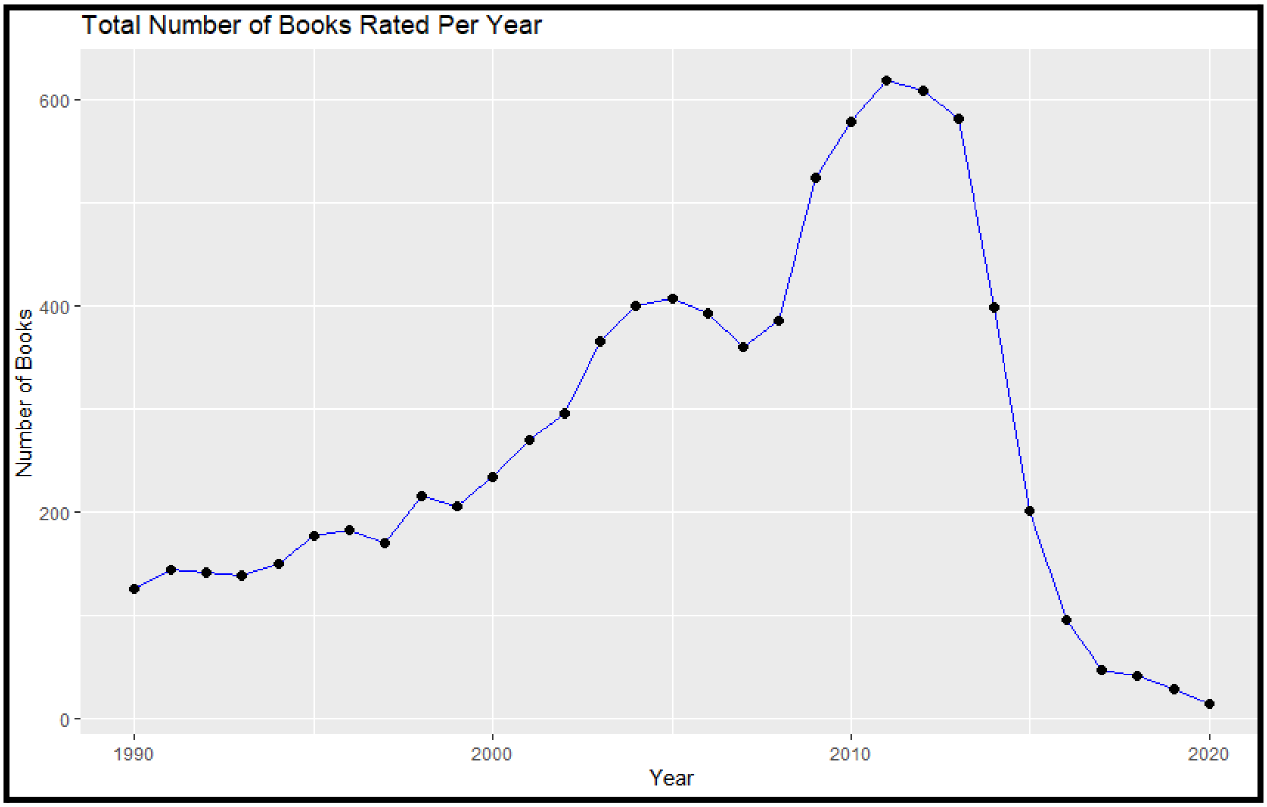
3. Publication Trends (1990-2020)



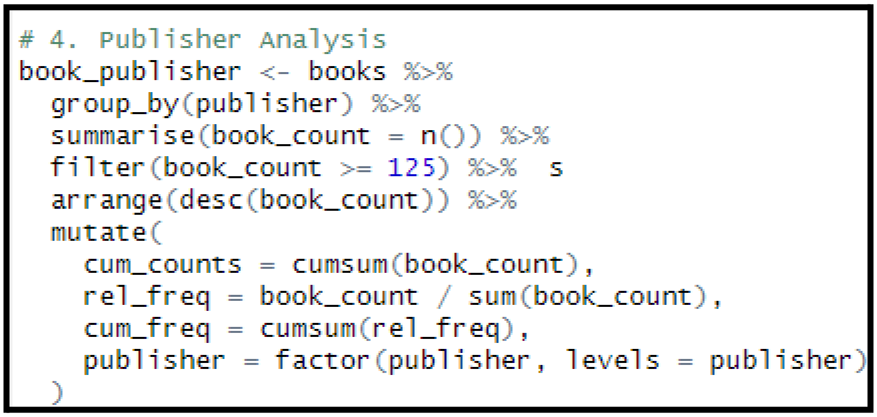
```
# 3. Books by Year (Line Plot)
by_year <- books %>%
  group_by(year) %>%
  summarise(total_books = n())

ggplot(by_year, aes(x = year, y = total_books)) +
  geom_line(color = "blue") +
  geom_point(size = 2) +
  labs(title = "Total Number of Books Rated Per Year", x = "Year", y = "Number of Books")
```

- Observation: The number of books published annually increased steadily until the early 2010s, followed by a slight decline.
- Visualization: A line plot showing the number of books published per year was generated.
- Insight: The growth in book publications aligns with the rise of self-publishing and digital books.

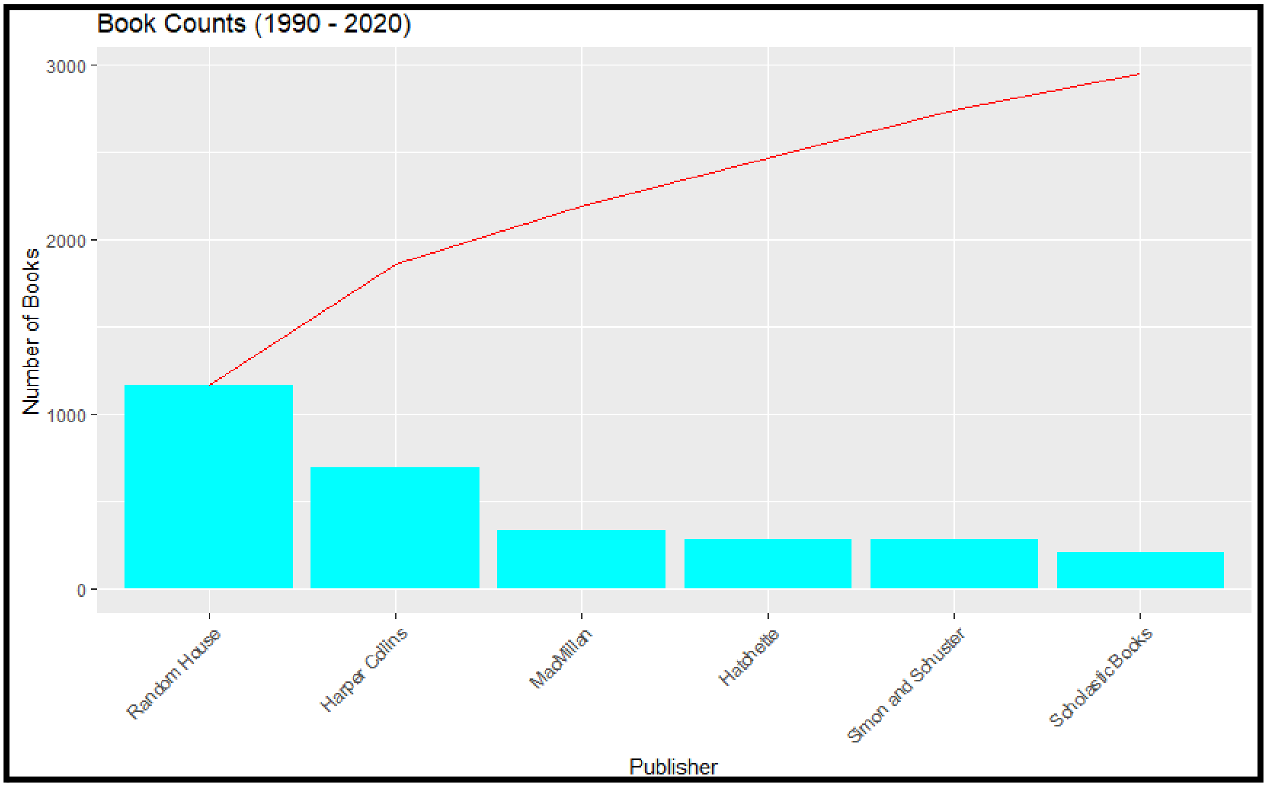


4. Publisher Market Share

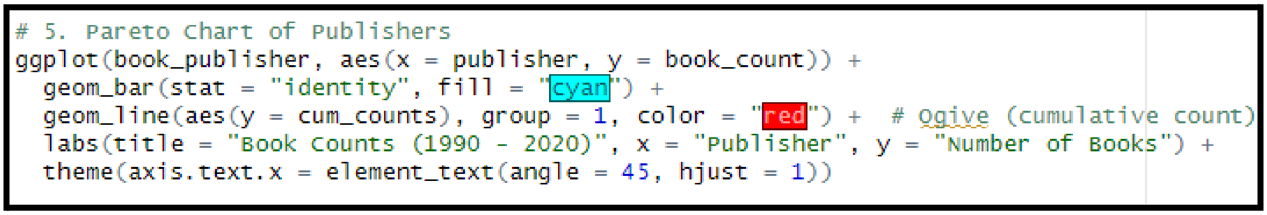


```
# 4. Publisher Analysis
book_publisher <- books %>%
  group_by(publisher) %>%
  summarise(book_count = n()) %>%
  filter(book_count >= 125) %>%  s
  arrange(desc(book_count)) %>%
  mutate(
    cum_counts = cumsum(book_count),
    rel_freq = book_count / sum(book_count),
    cum_freq = cumsum(rel_freq),
    publisher = factor(publisher, levels = publisher)
  )
```

- Observation: The top five publishers account for approximately 75% of all books in the dataset.
- Visualization: A Pareto chart was created to display the cumulative share of books per publisher.
- Insight: A small number of dominant publishers control most of the market, suggesting limited competition among major publishing houses.

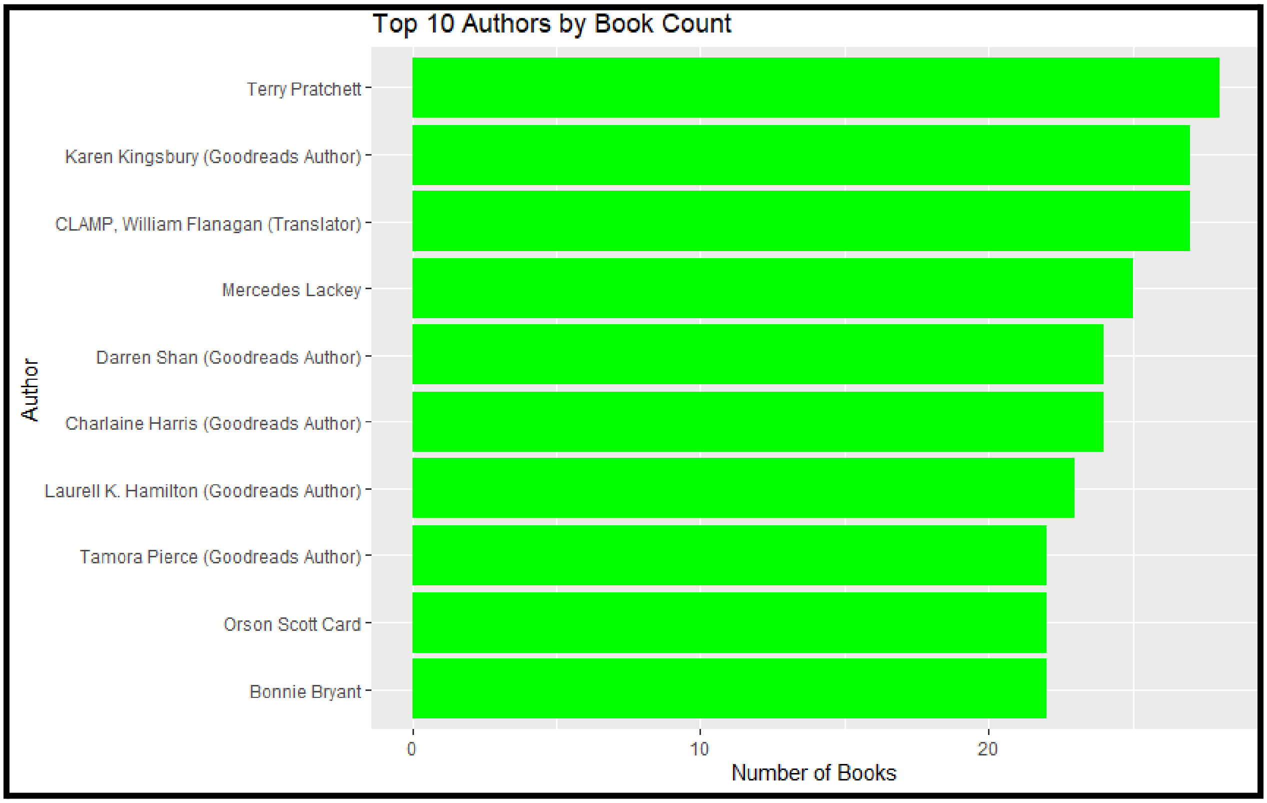


5. Additional Analysis: Top Authors



```
# 5. Pareto Chart of Publishers
ggplot(book_publisher, aes(x = publisher, y = book_count)) +
  geom_bar(stat = "identity", fill = "cyan") +
  geom_line(aes(y = cum_counts), group = 1, color = "red") +  # Ogive (cumulative count)
  labs(title = "Book Counts (1990 - 2020)", x = "Publisher", y = "Number of Books") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

- Observation: The dataset reveals that a few authors have a disproportionately high number of published books.
- Visualization: A bar chart showcasing the top 10 authors by book count was created.
- Insight: Established authors with a strong reader base continue to dominate publishing trends.



Conclusion :

- The dataset shows a strong positive skew in book ratings, indicating that readers generally provide favorable reviews.
- The majority of books are under 400 pages, making them more accessible to casual readers.
- The number of books published has fluctuated over time, peaking in the 2010s before slightly declining.
- A handful of publishers dominate the market, which may impact the diversity of available literature.

Recommendations :

- For Authors: Focus on books with moderate length (300-400 pages) to maximize readership.
- For Publishers: Consider diversifying book offerings beyond the dominant genres and popular authors.
- For Readers: Explore lesser-known authors and independent publishers for unique reading experiences.
- For Future Research: Investigate how digital publishing and e-books have influenced these trends post-2020.

References :

- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.
- Kaggle. (n.d.). *Goodreads dataset*. Retrieved from www.kaggle.com
- R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.