ALY6000 Introduction to Analytics

Project Report 3

Assignment: Exploring Visualizations

Submission Date: 31st January 2024, Wednesday

Submitted by

Shraddha P Gupte

Introduction

By examining the dataset, we aim to identify patterns and trends that elaborate on the preferences of readers throughout numerous numbers of years. We currently focus on the latest 20-year data and present our findings accordingly. Through this analysis we seek to answer questions about reader behavior and preferences on popular book formats, awards, book ratings, publisher, and more data. by different visualizations and graphical representations. Understanding this data is crucial for publishers and authors to set their offerings effectively and meet the evolving demands of readers.

Cleaning the data set

Question 1

The given data set is standardized for the analysis of the data by using the 'clean_names' function from the 'janitor' package.

Console | Terminal | Background Jobs

R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/

```
> #Question 1 - Janitor package has been added above
> books <- clean_names(books)
> books
                                                 title
1                  Harry Potter and the Order of the Phoenix
2                                              Twilight
3                                    The Da Vinci Code
4                                              Divergent
5                                  Anne of Green Gables
6                     Harry Potter and the Sorcerer's Stone
7                                               Dracula
8                                   The Lightning Thief
9                                     A Game of Thrones
10                                             The Giver
11                        The Adventures of Sherlock Holmes
12                                             Outlander
13                          The Girl with the Dragon Tattoo
14                                      Angela's Ashes
15                                   The Golden Compass
16                  Harry Potter and the Prisoner of Azkaban
17                                         The Notebook
18                       Harry Potter and the Goblet of Fire
19                    Harry Potter and the Half-Blood Prince
20                                   The Name of the Wind
21                                             Eragon
22                                 The Shadow of the Wind
```

Exploring Visualizations

Question 2

The data is further worked upon to covert the dates in the correct format by using ‘mdy’ function from the ‘lubridate’ package

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 2 - 'lubridate' package has been added above
> first_publish_date<-mdy(books$first_publish_date)
> first_publish_date
   [1] "2003-06-21" "2005-10-05" "2003-03-18" "2011-04-25" "2008-10-28" "1997-06-26" "1997-05-26"
   [8] "2005-06-28" "1996-08-06" "1993-04-26" "1992-10-28" "1991-06-01" "2005-08-28" "1996-09-05"
  [15] "1995-07-09" "1999-07-08" "1996-10-01" "2000-07-08" "2005-07-16" "2007-03-27" "2002-06-28"
  [22] "2001-05-28" "1998-07-02" "2003-09-01" "2001-06-19" "1998-08-20" "2020-10-28" "1996-01-17"
  [29] "2000-05-17" "2008-04-01" "2009-08-28" "2013-03-19" "2007-01-01" "2008-10-28" "2006-05-09"
  [36] "2000-08-08" "1995-09-29" "2000-11-28" "2001-04-21" "1990-11-07" "2012-08-02" "1997-04-25"
  [43] "1994-04-12" "1995-10-28" "1998-11-16" "2007-05-28" "2011-01-01" "1996-09-16" "2003-09-23"
```

Question 3

Additional column name ‘year’ is added to only give the year data from the first_publish_date by using the ‘year’ function.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 3
> books$year<-year(mdy(books$first_publish_date))
> books$year
   [1] 2003 2005 2003 2011 2008 1997 1997 2005 1996 1993 1992 1991 2005 1996 1995 1999 1996 2000
  [19] 2005 2007 2002 2001 1998 2003 2001 1998 2020 1996 2000 2008 2009 2013 2007 2008 2006 2000
  [37] 1995 2000 2001 1990 2012 1997 1994 1995 1998 2007 2011 1996 2003 2002 2011 1996 2012 2001
  [55] 1990 2006 2005 2005 1990 1991 1995 2001 2011 2012 2018 2000 2019 2007 1992 1997 2008 2005
```

Question 4

From the last operation the data is further filtered from the year 1990 to 2020 to give the findings of the latest data using ‘filter’.

Filter

	num_ratings	ratings_by_stars	liked_percent	bbe_score	bbe_votes	year
...	4964519	['1751460', '1113682', '1008686', '542017', '548674']	78	1459448	14874	2005
)...	1933446	['645308', '667657', '399278', '142103', '79100']	89	876633	9231	2003
...	2906258	['1409189', '882493', '434381', '123286', '56909']	94	793269	8339	2011
	727685	['379818', '208919', '100304', '25501', '13143']	95	695453	7340	2008
(...	7048471	['4578137', '1611874', '600384', '139551', '118525']	96	691430	7348	1997
	938325	['345260', '329217', '197206', '48642', '18000']	93	646782	6988	1997
...	1992300	['1006885', '604999', '289310', '64014', '27092']	95	597132	6370	2005
...	2003043	['1231600', '540493', '156705', '41880', '32365']	96	562241	6007	1996

Showing 1 to 9 of 9,171 entries, 17 total columns

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
```

Exploring Visualizations

Question 5

The data is further filtered to remove the unnecessary columns by using the function on 'subset', that would not be used in the analysis for the data.

	publisher	first_publish_date	awards	num_ratings	rating
370	Scholastic Books	06/21/03	['Bram Stoker Award for Works for Young Readers (20...	2507623	['1593
279	Random House	01/28/13	[]	2998241	['1617
501	Hatchette	10-05-2005	['Georgia Peach Book Award (2007)', 'Buxtehuder Bull...	4964519	['1751
552	Random House	09-01-2005	['National Jewish Book Award for Children's and Youn...	1834276	['1048
489	Anchor	03/18/03	['British Book Award for Book of the Year (2005)', 'Boo...	1933446	['6453
503	Random House	09/23/97	[]	1717312	['7129
272	Random House	07/20/90	[]	966196	['3829
487	Katherine Tegen Books	04/25/11	['Georgia Peach Book Award (2012)', 'South Carolina B...	2906258	['1409
401	Simon and Schuster	10/28/05	[]	2055102	['6168

Showing 1 to 8 of 21,056 entries, 17 total columns

Console | Terminal | Background Jobs

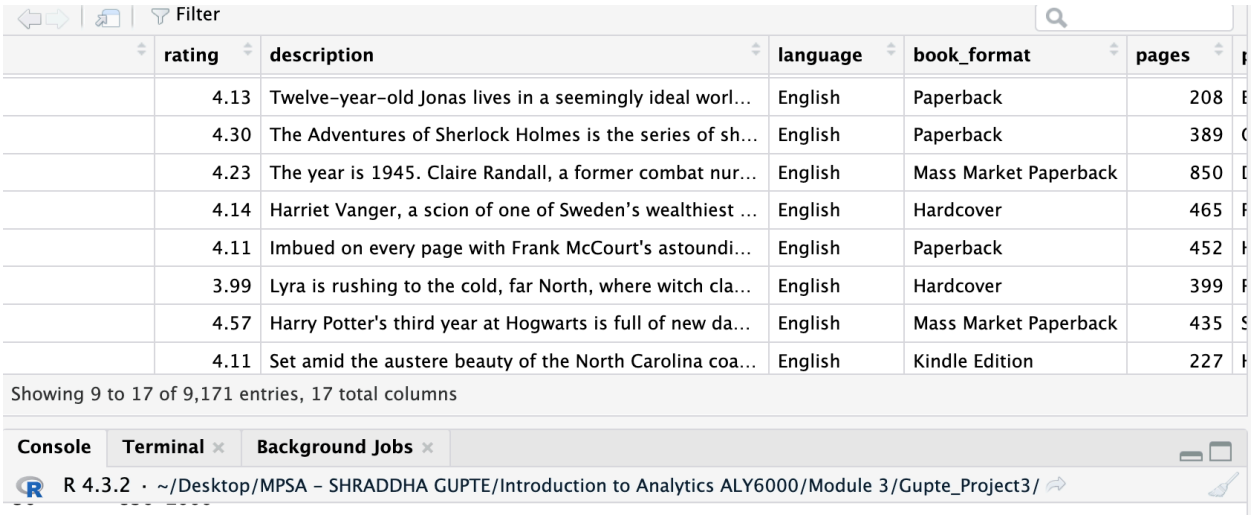
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/

```
> books <- subset(books, select = -c(publish_date, edition, characters, price, genres, setting, isbn))
> books
                                          title                                series
1          Harry Potter and the Order of the Phoenix                     Harry Potter #5
2                              Pride and Prejudice                                <NA>
3                                         Twilight                 The Twilight Saga #1
4                                   The Book Thief                                <NA>
5                                  The Da Vinci Code                    Robert Langdon #2
6                                Memoirs of a Geisha                                <NA>
7                       The Picture of Dorian Gray                                <NA>
8                                         Divergent                         Divergent #1
9                                   Romeo and Juliet                                <NA>
10                             Anne of Green Gables                 Anne of Green Gables #1
11            Harry Potter and the Sorcerer's Stone                         Harry Potter #1
12                      The Time Traveler's Wife                                <NA>
13                                        Dracula                              Dracula #1
14                             The Lightning Thief       Percy Jackson and the Olympians #1
15                               The Secret Garden                                <NA>
16                        A Thousand Splendid Suns                                <NA>
17                               A Game of Thrones             A Song of Ice and Fire #1
18                                The Lovely Bones                                <NA>
19                                     The Odyssey                                <NA>
20                                       Life of Pi                                <NA>
21                             Water for Elephants                                <NA>
22                       Frankenstein: The 1818 Text                                <NA>
```

Question 6

To give to most optimized data analysis, the data set is further filtered to give just the books that have pages less than 1200.

Exploring Visualizations



Filter

	rating	description	language	book_format	pages
	4.13	Twelve-year-old Jonas lives in a seemingly ideal worl...	English	Paperback	208
	4.30	The Adventures of Sherlock Holmes is the series of sh...	English	Paperback	389
	4.23	The year is 1945. Claire Randall, a former combat nur...	English	Mass Market Paperback	850
	4.14	Harriet Vanger, a scion of one of Sweden's wealthiest ...	English	Hardcover	465
	4.11	Imbued on every page with Frank McCourt's astoundi...	English	Paperback	452
	3.99	Lyra is rushing to the cold, far North, where witch cla...	English	Hardcover	399
	4.57	Harry Potter's third year at Hogwarts is full of new da...	English	Mass Market Paperback	435
	4.11	Set amid the austere beauty of the North Carolina coa...	English	Kindle Edition	227

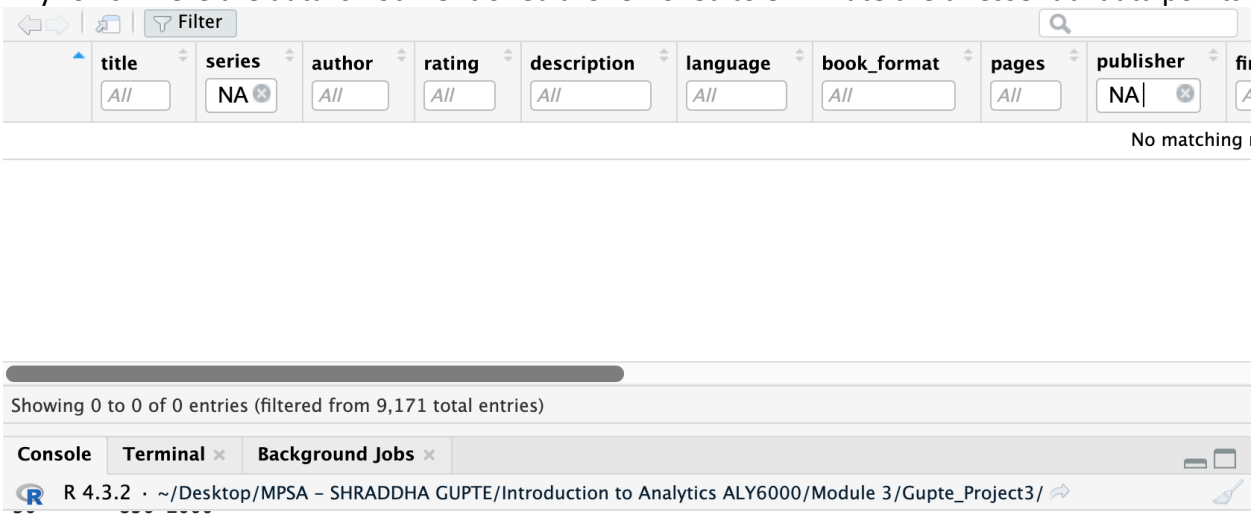
Showing 9 to 17 of 9,171 entries, 17 total columns

Console Terminal Background Jobs

R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/

Question 7

Any rows where the data is not mentioned are removed to eliminate the unessential data points.



Filter

	title	series	author	rating	description	language	book_format	pages	publisher
	All	NA	All	All	All	All	All	All	NA

No matching

Showing 0 to 0 of 0 entries (filtered from 9,171 total entries)

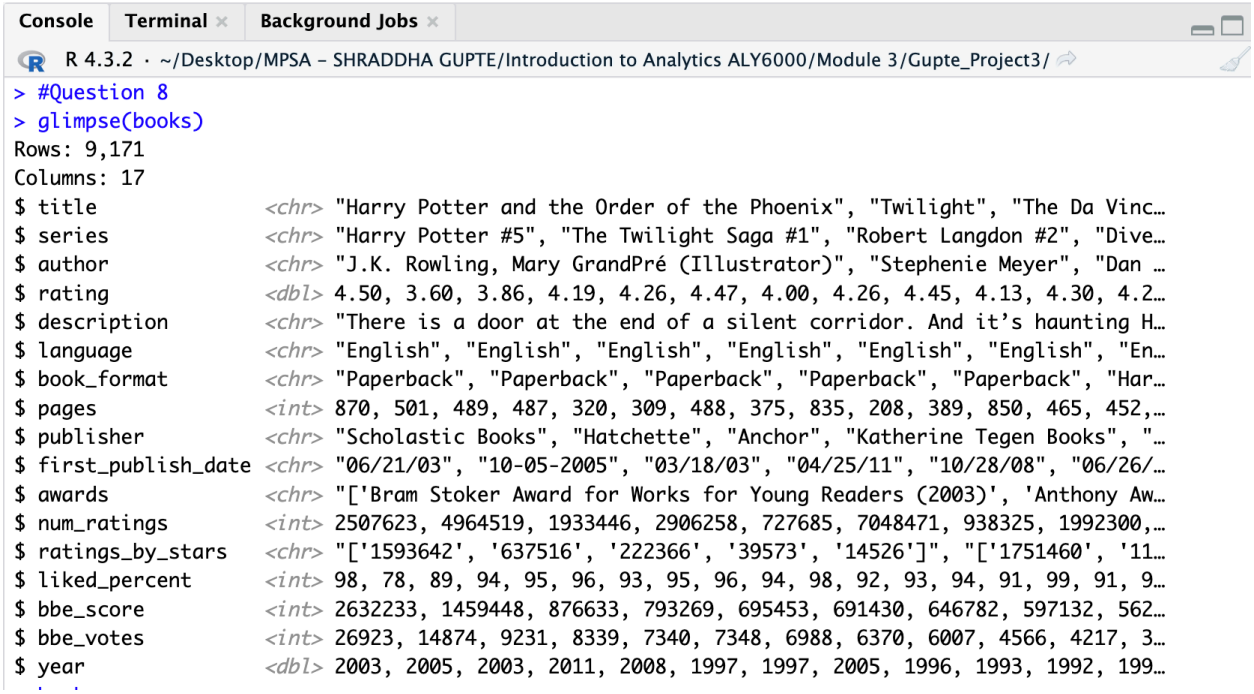
Console Terminal Background Jobs

R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/

Question 8

The 'glimpse' function is used to produce the long view of the data set.

Exploring Visualizations



Question 9

The 'summary' function is used to give more concised details of the large data and observation of the given stats

Exploring Visualizations

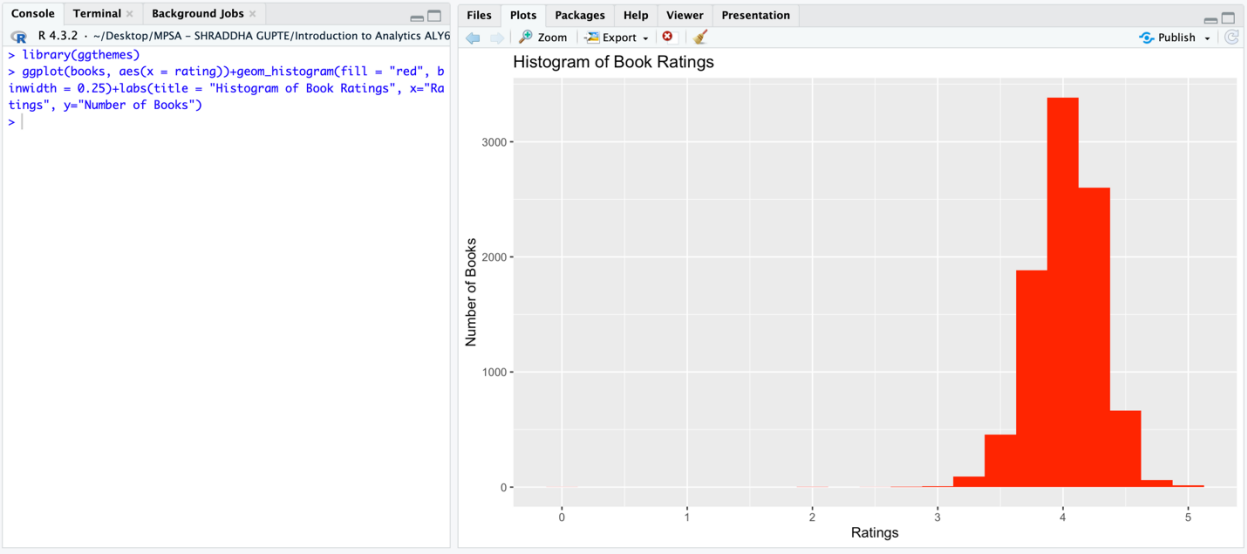
```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> summary(books)
    title              series             author             rating      description
 Length:9171        Length:9171        Length:9171        Min.   :0.000   Length:9171
 Class :character   Class :character   Class :character   1st Qu.:3.860   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :4.040   Mode  :character
                                                          Mean   :4.029
                                                          3rd Qu.:4.200
                                                          Max.   :5.000

   language         book_format            pages          publisher         first_publish_date
 Length:9171        Length:9171        Min.   :   0.0   Length:9171        Length:9171
 Class :character   Class :character   1st Qu.: 241.0   Class :character   Class :character
 Mode  :character   Mode  :character   Median : 336.0   Mode  :character   Mode  :character
                                       Mean   : 347.8
                                       3rd Qu.: 416.0
                                       Max.   :1200.0

    awards            num_ratings      ratings_by_stars   liked_percent      bbe_score
 Length:9171        Min.   :      0   Length:9171        Min.   : 20.00   Min.   :      0
 Class :character   1st Qu.:   1373   Class :character   1st Qu.: 91.00   1st Qu.:     84
 Mode  :character   Median :   4740   Mode  :character   Median : 94.00   Median :     98
                    Mean   :  22793                      Mean   : 92.94   Mean   :   2030
                    3rd Qu.:  14296                      3rd Qu.: 96.00   3rd Qu.:    218
                    Max.   :7048471                      Max.   :100.00   Max.   :2632233
                                                         NA's   :7
   bbe_votes            year
 Min.   :   -1.0   Min.   :1990
 1st Qu.:    1.0   1st Qu.:2001
 Median :    1.0   Median :2007
 Mean   :   22.9   Mean   :2006
 3rd Qu.:    3.0   3rd Qu.:2011
 Max.   :26923.0   Max.   :2020
```

Question 10

Creating a rating histogram with the 'Number of Books' and 'rating' would give the analysis of the number of books according to their significant ratings.

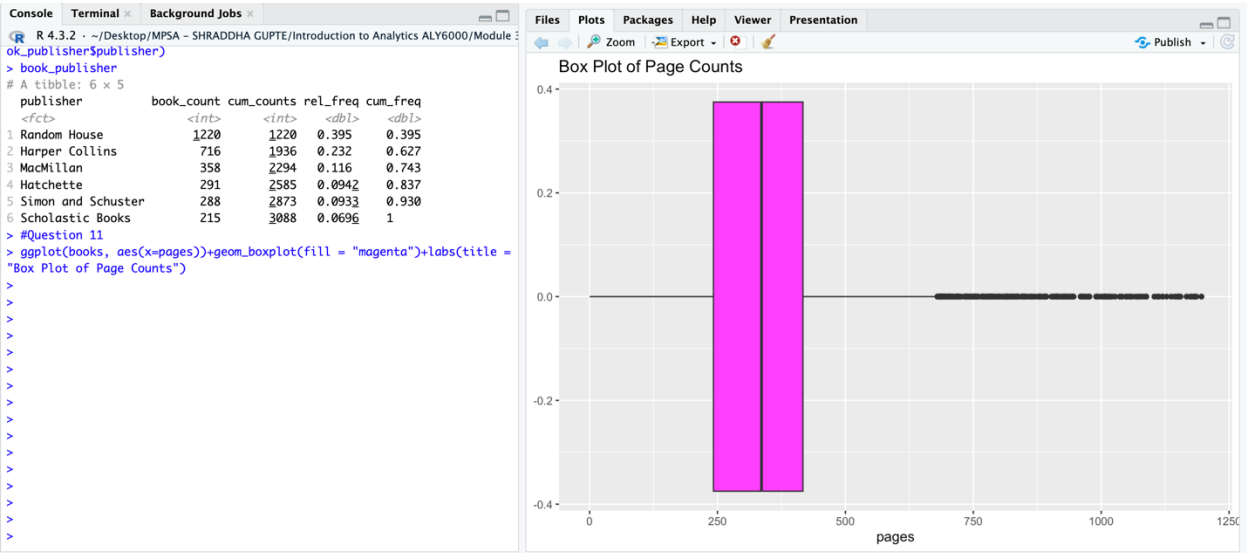
From the below graph it is observed that more number of books are rated from the range 3 to 5 and most are rated 3.75 to 4.25.

Exploring Visualizations



Question 11

The data is plot in the Box Plot graph to give the analysis of the median of the data. The median of the below data lies within 200 to 400 pages highlighted in the color 'magenta'.



Question 12

A new data frame named 'by_year' is created to identify the total number of books published in that year to get the total books count by year. This is done by using the 'group_by' and 'summarise' function.

Exploring Visualizations

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> by_year <- books %>%group_by(year) %>%summarise(total_books = n())
> by_year
# A tibble: 31 × 2
    year total_books
   <dbl>       <int>
 1  1990         355
 2  1991         369
 3  1992         413
 4  1993         407
 5  1994         460
 6  1995         497
 7  1996         527
 8  1997         530
 9  1998         578
10  1999         585
# ℹ 21 more rows
```

Question 13

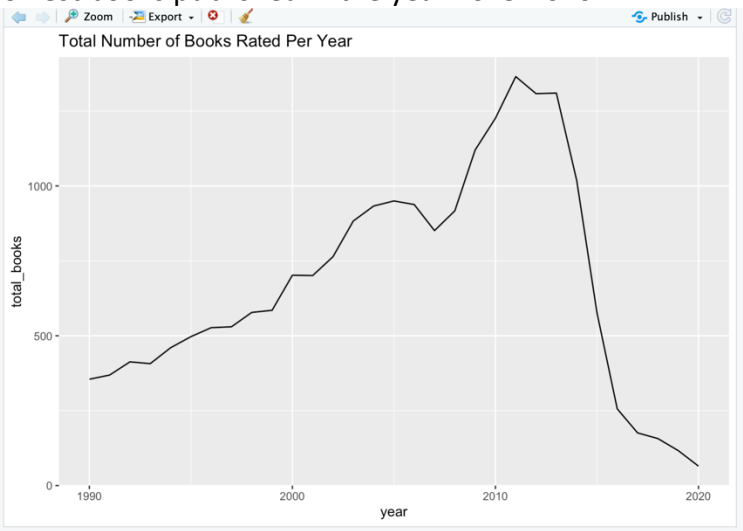
A line graph is created to plot the total number of books rated per year. For this operation 'by_year' data frame is filtered by years that are more than 1990 and less than 2020.

The line graph is created by keeping 'year' on the x axis and 'total_books' on y axis. The line is plot by using the 'geom_line' function.

From the graph it is observed that –

- The number of books published has consistently grown from the year 1990 to 2005.
- There was a significant drop in the number of books published in the year 2007.
- The number then increased to reach the highest number of books being published in the years 2010 to 2015.
- This number again dropped to the lowest books published in the year 2015-2020.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Mod
> by_year<- filter(by_year, year>=1990, year<=2020)
> by_year
# A tibble: 31 × 2
    year total_books
   <dbl>       <int>
 1  1990         355
 2  1991         369
 3  1992         413
 4  1993         407
 5  1994         460
 6  1995         497
 7  1996         527
 8  1997         530
 9  1998         578
10  1999         585
# ℹ 21 more rows
# ℹ Use `print(n = ...)` to see more rows
> ggplot(by_year, aes(x = year, y = total_books)) + geom_line() + labs
(title = "Total Number of Books Rated Per Year")
>
```



Exploring Visualizations

Question 14

A new data frame is created named as 'book_publisher' from the 'books' data frame to compute the publisher names by using the 'group_by' function and the collate the count of books by using the 'summarise' function in order the give the overview of the new data frame.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> book_publisher<-books%>%group_by(publisher)%>%summarise(book_count =n())
> book_publisher
# A tibble: 2,013 × 2
   publisher                    book_count
   <chr>                             <int>
 1 "\"Marvel\""                          1
 2 "4 Corners Press"                     1
 3 "47North"                             7
 4 "48fourteen"                          2
 5 "5 Prince Publishing"                 1
 6 "7th House"                           2
 7 "A.L. Jackson Books Inc."             1
 8 "A.M. Madden; First edition"          1
 9 "A.W. Bruna"                          1
10 "ABRAMS"                              1
# ℹ 2,003 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

Question 15

The further analysis is worked on by using the 'book_publisher' data frame and restoring it by filtering the number of books less than 125 by using the 'filter' function.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 15
> book_publisher <- filter(book_publisher, book_count>125)
> book_publisher
# A tibble: 6 × 2
  publisher          book_count
  <chr>                   <int>
1 Harper Collins            716
2 Hatchette                 291
3 MacMillan                 358
4 Random House             1220
5 Scholastic Books          215
6 Simon and Schuster        288
```

Exploring Visualizations

Question 16

The data is further rearranged to get the highest to lowest count of books by using the order function on 'book_publisher' data frame and restoring it in the same file for easier visualization.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 16
> book_publisher<-book_publisher[order(book_publisher$book_count,decreasing=TRUE),]
> book_publisher
# A tibble: 6 × 2
  publisher          book_count
  <chr>                   <int>
1 Random House             1220
2 Harper Collins            716
3 MacMillan                 358
4 Hatchette                 291
5 Simon and Schuster        288
6 Scholastic Books          215
```

Question 17

An additional column is added to the 'book_publisher' data frame named as 'cum_counts' to get the cumulative total of the number of books.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 17
> book_publisher <- book_publisher %>% mutate("cum_counts" = cumsum(book_count))
> book_publisher
# A tibble: 6 × 3
  publisher          book_count cum_counts
  <chr>                   <int>      <int>
1 Random House             1220       1220
2 Harper Collins            716       1936
3 MacMillan                 358       2294
4 Hatchette                 291       2585
5 Simon and Schuster        288       2873
6 Scholastic Books          215       3088
```

Question 18

The relative frequency is calculated in the 'rel_freq' column of the 'book_publisher' data frame to understand the absolute frequency of the number of the books.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 18
> book_publisher<-book_publisher%>%mutate("rel_freq" = book_count/sum(book_count))
> book_publisher
# A tibble: 6 × 4
  publisher          book_count cum_counts rel_freq
  <chr>                   <int>      <int>    <dbl>
1 Random House             1220       1220   0.395
2 Harper Collins            716       1936   0.232
3 MacMillan                 358       2294   0.116
4 Hatchette                 291       2585   0.0942
5 Simon and Schuster        288       2873   0.0933
6 Scholastic Books          215       3088   0.0696
```

Question 19

Another column on 'cum_freq' is added to get the cumulative sum of the relative frequency calculated in the previous question.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 19
> book_publisher<-book_publisher%>%mutate("cum_freq" = cumsum(rel_freq))
> book_publisher
# A tibble: 6 × 5
  publisher          book_count cum_counts rel_freq cum_freq
  <chr>                   <int>      <int>    <dbl>    <dbl>
1 Random House             1220       1220   0.395    0.395
2 Harper Collins            716       1936   0.232    0.627
3 MacMillan                 358       2294   0.116    0.743
4 Hatchette                 291       2585   0.0942   0.837
5 Simon and Schuster        288       2873   0.0933   0.930
6 Scholastic Books          215       3088   0.0696   1
```

Question 20

Make the publisher column into a factor with the levels defined by the current ordering of the publisher column.

```
> #Question 20
> book_publisher$publisher <- factor(book_publisher$publisher, levels = book_publisher$publisher)
> book_publisher
# A tibble: 6 × 5
  publisher          book_count cum_counts rel_freq cum_freq
  <fct>                   <int>      <int>    <dbl>    <dbl>
1 Random House             1220       1220   0.395     0.395
2 Harper Collins            716       1936   0.232     0.627
3 MacMillan                 358       2294   0.116     0.743
4 Hatchette                 291       2585   0.0942    0.837
5 Simon and Schuster        288       2873   0.0933    0.930
6 Scholastic Books          215       3088   0.0696    1
>
```

Question 21

Creating a pareto and ogive chart from the 'book_publisher' data frame to create a graphical representation of the findings on number of books from 1990-2020.

From the graph it is observed that –

- The pareto chart defines the highest to lowest number of book count. T
- he highest number of books is published by Random House publisher.
- The lowest number of books is published by Scholastic Books
- The ogive chart defines the cumulative distribution of the number of books.

Question 22

The below analysis is on the number of book counts based on the book formats.

Exploring Visualizations

Step1 - A new data frame is created named as 'book_format-df' from 'books' data frame having book formats atleast above 4 rating and removing the books having unknown binding.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Question 22
> #Creating a new data frame on book formats having atleast above 4 rating and removing the books having unknown bindin
g.
> book_format_df<- filter(books, book_format != 'Unknown Binding', rating>=4)
> book_format_df
                                        title                                  series
1      Harry Potter and the Order of the Phoenix                  Harry Potter #5
2                                     Divergent                     Divergent #1
3                           Anne of Green Gables           Anne of Green Gables #1
4          Harry Potter and the Sorcerer's Stone                  Harry Potter #1
5                                       Dracula                       Dracula #1
6                             The Lightning Thief  Percy Jackson and the Olympians #1
7                               A Game of Thrones           A Song of Ice and Fire #1
8                                     The Giver                     The Giver #1
9              The Adventures of Sherlock Holmes                Sherlock Holmes #3
10                                    Outlander                     Outlander #1
11               The Girl with the Dragon Tattoo                   Millennium #1
12                              Angela's Ashes                 Frank McCourt #1
13      Harry Potter and the Prisoner of Azkaban                  Harry Potter #3
14                                The Notebook                  The Notebook #1
15           Harry Potter and the Goblet of Fire                  Harry Potter #4
```

Step 2 – As the resulting data from the last operation gave wide range of return, the data frame is filtered by using the 'subset' function to only have the required columns for analysis.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Step 2 = Selecting the columns that are required for the analysis of the book_format_df data frame by using the 'subs
et' function
> book_format_df<- subset(book_format_df, select= c(year, book_format, rating))
> book_format_df
    year           book_format rating
1   2003             Paperback   4.50
2   2011             Paperback   4.19
3   2008             Paperback   4.26
4   1997             Hardcover   4.47
5   1997             Paperback   4.00
6   2005             Paperback   4.26
7   1996 Mass Market Paperback   4.45
8   1993             Paperback   4.13
9   1992             Paperback   4.30
10  1991 Mass Market Paperback   4.23
11  2005             Hardcover   4.14
12  1996             Paperback   4.11
13  1999 Mass Market Paperback   4.57
14  1996        Kindle Edition   4.11
15  2000             Paperback   4.56
16  2005             Paperback   4.57
```

Step 3 – To identify the trend of the analysis based on the latest data the data frame is filtered from the year 2017 to 2020

Exploring Visualizations

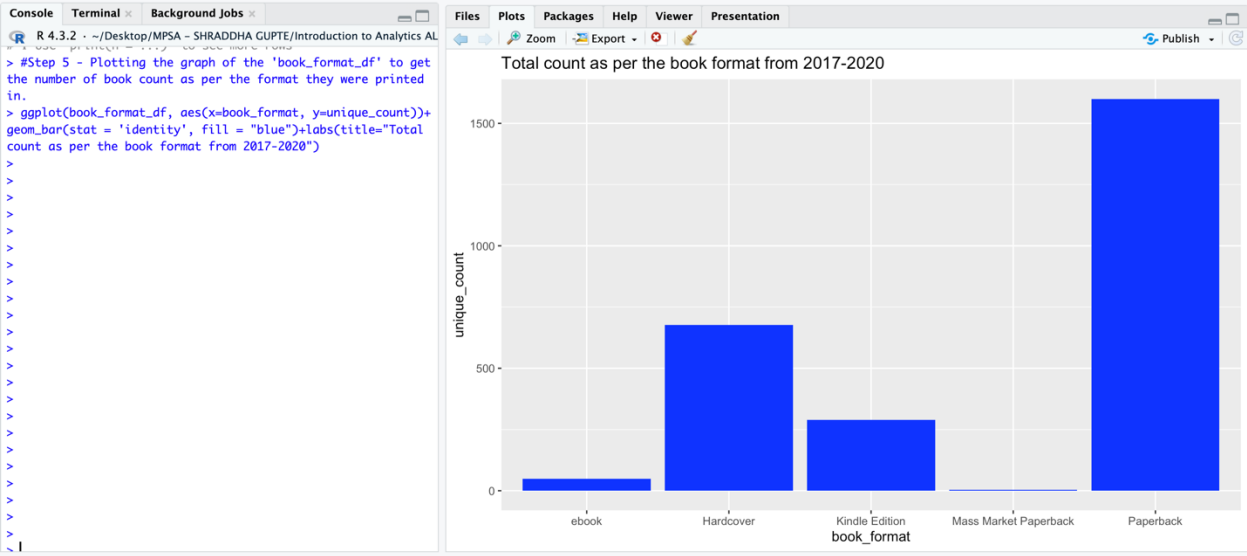
```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Step 3 - selecting the trend 3 year from 2017 to 2020.
> book_format_df<-filter(book_format_df, year>=2017, year<=2020)
> book_format_df
   year           book_format rating
1  2020             Hardcover   4.05
2  2019             Paperback   4.30
3  2018             Hardcover   4.25
4  2020             Paperback   4.29
5  2019             Hardcover   4.10
6  2017 Mass Market Paperback   4.14
7  2018             Paperback   4.20
8  2017             Hardcover   4.05
9  2020             Paperback   4.35
10 2019             Hardcover   4.17
11 2018             Paperback   4.13
12 2017             Paperback   4.27
13 2019        Kindle Edition   4.04
14 2020             Paperback   4.30
15 2017             Hardcover   4.36
16 2019             Paperback   4.01
17 2017             Hardcover   4.64
18 2018             Paperback   4.19
```

Step 4 – As the resulting data is given in individual row wise representation we use the 'group_by' function to compute the number of book formats and adding another column of 'unique_count' to give the count of umber of books in a particular book format.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 3/Gupte_Project3/
> #Step 4 - Creating a column on unique_count by computing the book_format coulumn using the c
ompute function.
> book_format_df <- book_format_df %>% group_by(book_format)%>% mutate(unique_count = n())
> book_format_df
# A tibble: 92 × 4
# Groups:   book_format [5]
    year book_format           rating unique_count
   <dbl> <chr>                  <dbl>        <int>
 1  2020 Hardcover               4.05           26
 2  2019 Paperback               4.3            40
 3  2018 Hardcover               4.25           26
 4  2020 Paperback               4.29           40
 5  2019 Hardcover               4.1            26
 6  2017 Mass Market Paperback   4.14            2
 7  2018 Paperback               4.2            40
 8  2017 Hardcover               4.05           26
 9  2020 Paperback               4.35           40
10  2019 Hardcover               4.17           26
# ℹ 82 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Exploring Visualizations

Step 5 - Plotting the graph of the 'book_format_df' to get the number of book count as per the format they were printed in.



From the graph it is observed that –

- The highest number of books were printed in the Paperback format.
- The Mass Market Paperback has printed the lowest number of book count.
- There are significant books printed Hardcover and Kindle Edition
- The above analysis can also derive that even though the books are available in different formats like electronic book, people still prefer the Paperback format of reading the physical copy of the books.

Question 23

Executive Summary – Analysis of the Book Formats

Overview - In this analysis, distribution of books was analyzed across different formats to understand readers' preferences and trends in the publishing industry. The dataset provided information about the formats of books and their respective counts. The objective was to gain insights into the popularity of various book formats and identify any notable trends.

Key Findings –

Distribution of Books by Formats:

The visualization above illustrates the distribution of books across different formats:

Key Takeaways:

Paperback format has the highest number of books printed, indicating its popularity among readers.

Mass Market Paperback format has printed the lowest number of books, suggesting a lesser preference for this format.

Hardcover and Kindle Editions also have significant book counts, indicating a preference for both physical and electronic formats.

Preference for Paperback Format:

The analysis highlights that despite the availability of various formats, readers still prefer the Paperback format for reading physical copies of books. This preference is evident from the highest book count in the Paperback format compared to other formats.

Conclusion -

Through this analysis, we have gained valuable insights into readers' preferences regarding book formats. The data indicates a strong preference for the Paperback format among readers, emphasizing the popularity of physical books despite the rise of electronic formats.

Readers' format preferences may also be influenced by other fact, such as pricing, genre, or reader demographics. Such insights would be invaluable for publishers and authors in tailoring their offerings to meet readers' preferences effectively.

Citations –

1. ***Books –***
   - *Kabacoff, R.I. (2022). R in action: Data analysis and graphics with R and tidyverse (3rd edition). Manning Publications.*
   - *Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.*
2. ***Proffessor Notes*** –
   - *Thomas Goulding (2024). Introduction to Analytics Notes [ALY6000 Course Notes]Canvas https://northeastern.instructure.com/courses/164773/pages/module-3-%7C-lessons?module_item_id=9797131*
3. ***Website –***
   - *R Core Team. (2021). R: A Language and Environment for Statistical Computing. https://www.R-project.org/*