Amazon Book Sales

Mingwei Wu

2/5/2021

Data Background

The data is from amazon webiste for best sellers of 2010-2020 (Top 100 books)

Import R library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dbplyr)
```

```
##
## Attaching package: 'dbplyr'

## The following objects are masked from 'package:dplyr':
##
##     ident, sql
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

Import Data

```
bs<-read.csv("amazon_bs_20102020.csv",header=TRUE)
head(bs)
```

```
##   Year Rank                                                Book_Title
## 1 2010    1 The Girl Who Kicked the Hornet's Nest (Millennium Trilogy)
## 2 2010    2        The Girl with the Dragon Tattoo (Millennium Series)
## 3 2010    3                                          Decision Points
## 4 2010    4                                                The Help
## 5 2010    5          The Girl Who Played with Fire (Millennium Series)
## 6 2010    6              The Ugly Truth (Diary of a Wimpy Kid, Book 5)
##             Author Rating Num_Customers_Rated Price
## 1    Stieg Larsson    4.7                8475 17.24
## 2    Stieg Larsson    4.4               11516  9.99
## 3   George W. Bush    4.6                2201 17.80
## 4 Kathryn Stockett    4.8               14772 14.97
## 5    Stieg Larsson    4.7                7949  0.02
## 6      Jeff Kinney    4.8                5312  9.52
```

Check data duplicated values and missing values

```
dim(bs)
```

```
## [1] 1094    7
```

```
na.omit(bs)%>%
  dim() #the data is completed
```

```
## [1] 1094    7
```

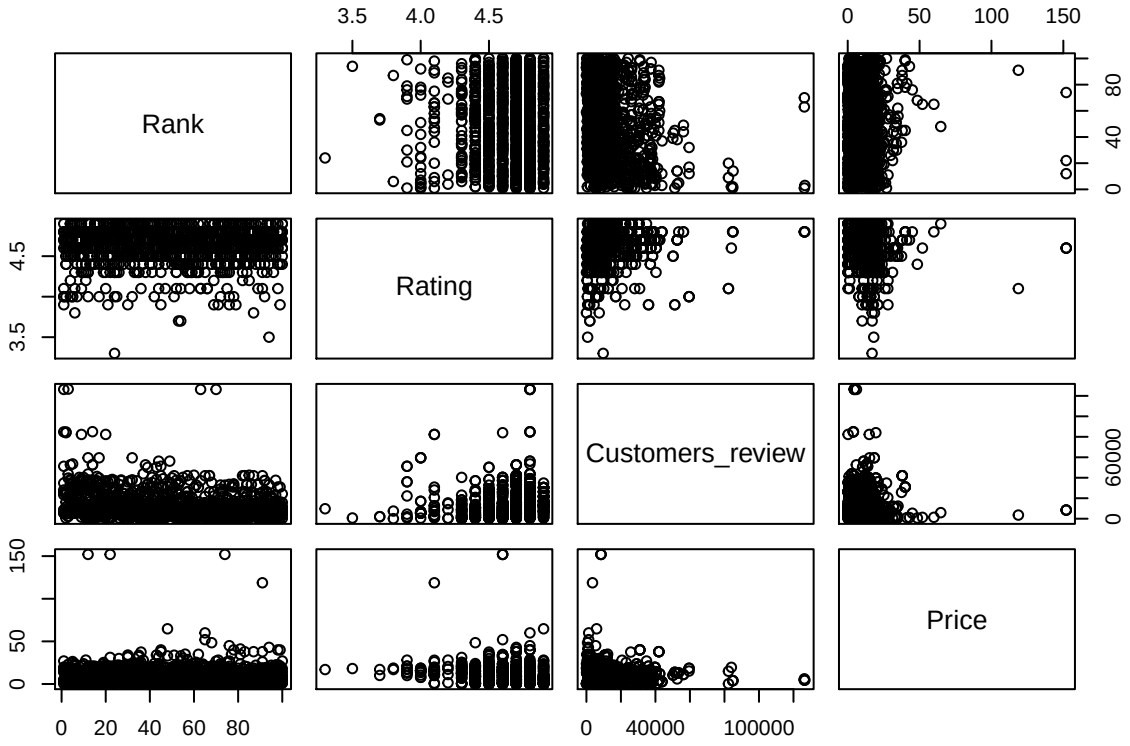
```
bs%>%distinct()%>%
  dim()
```

```
## [1] 1094    7
```

```
bs<-bs%>%
  rename(Customers_review = "Num_Customers_Rated")
```

Top 100 sellers are having higher Rathing by customers,and price is not expensive.

```
pairs(~Rank+Rating+Customers_review+Price, data=bs) # The graph display that with increased Rating, the
```



As increased with year, cuntomers increase their reaction and comment. Assume that people get used to spend the time on internet, also express their subjectiveness.

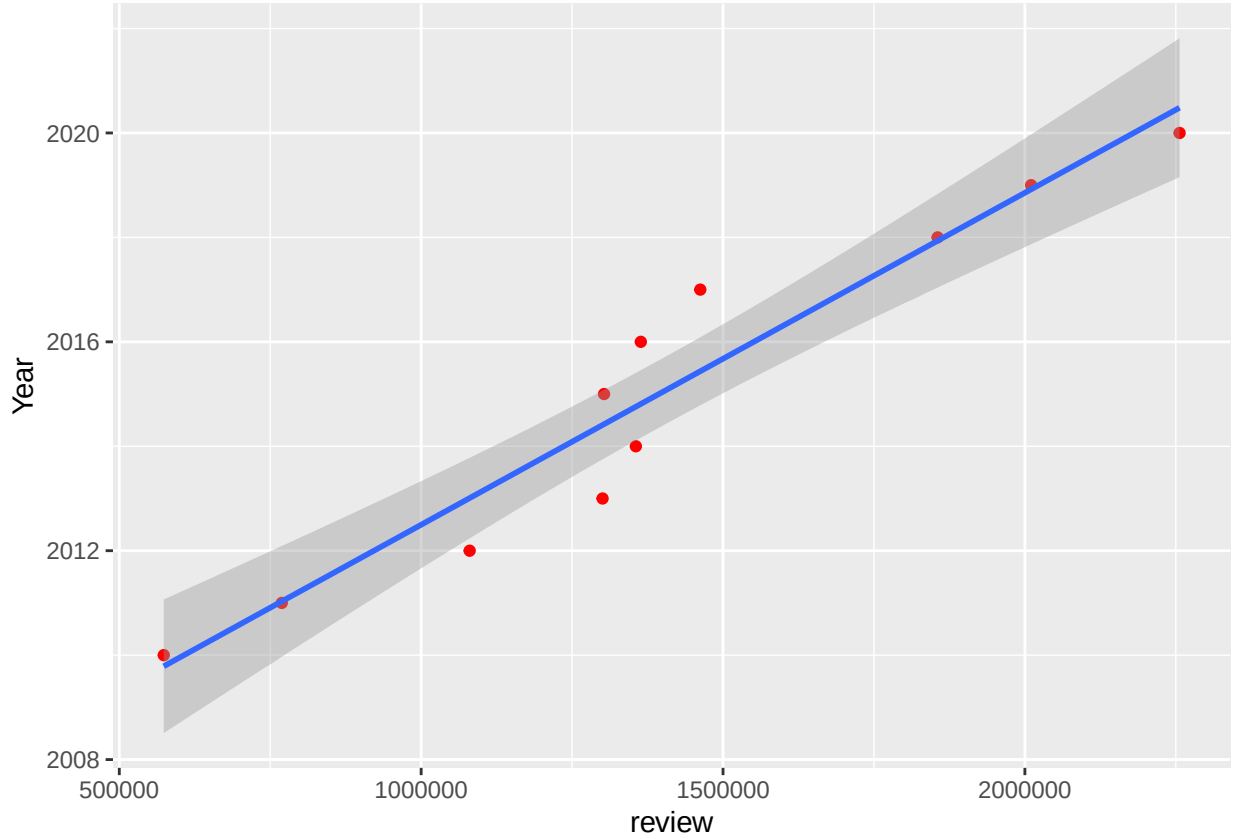
```
year_reviw<-bs%>%
  group_by(Year)%>%
  summarise(review=sum(Customers_review))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

The below scatter plot and line desmonstrate the customers' review and Year that have linear relationship. and the tendency is going up.

```
year_reviw%>%
  ggplot(aes(review,Year))+geom_point(color="red")+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
head(bs)
```

```
##   Year Rank                                                Book_Title
## 1 2010    1 The Girl Who Kicked the Hornet's Nest (Millennium Trilogy)
## 2 2010    2        The Girl with the Dragon Tattoo (Millennium Series)
## 3 2010    3                                          Decision Points
## 4 2010    4                                                 The Help
## 5 2010    5          The Girl Who Played with Fire (Millennium Series)
```

```
## 6 2010    6              The Ugly Truth (Diary of a Wimpy Kid, Book 5)
##              Author Rating Customers_review Price
## 1    Stieg Larsson    4.7             8475 17.24
## 2    Stieg Larsson    4.4            11516  9.99
## 3   George W. Bush    4.6             2201 17.80
## 4 Kathryn Stockett    4.8            14772 14.97
## 5    Stieg Larsson    4.7             7949  0.02
## 6      Jeff Kinney    4.8             5312  9.52
```

```
sqldf("select *
      from bs
      group by Book_Title
      order by Customers_review desc
      limit 10") # the most customer review is the book, 'Where the Crawdads Sing' that best seller ran
```

```
##    Year Rank
## 1  2018   70
## 2  2018    1
## 3  2020    2
## 4  2015    9
## 5  2012   12
## 6  2015   44
## 7  2020    5
## 8  2012   38
## 9  2012    1
## 10 2019   39
##                                                                                Book_Title
## 1                                                                  Where the Crawdads Sing
## 2                                                                    Becoming (182 GRAND)
## 3           Too Much and Never Enough: How My Family Created the World's Most Dangerous Man
## 4                                                                  The Girl on the Train
## 5                                                                              Gone Girl
## 6                                                              The Nightingale: A Novel
## 7                                                                           Midnight Sun
## 8                                                               The Fault in Our Stars
## 9  Fifty Shades of Grey: Book One of the Fifty Shades Trilogy (Fifty Shades of Grey Series)
## 10                                                                   The Silent Patient
##                 Author Rating Customers_review Price
## 1          Delia Owens    4.8           126306  4.58
## 2       Michelle Obama    4.8            84761  3.91
## 3  Mary L. Trump Ph.D.    4.6            84008 19.58
## 4        Paula Hawkins    4.1            82313 14.98
## 5        Gillian Flynn    4.0            59506 18.48
## 6       Kristin Hannah    4.8            56216 11.15
## 7      Stephenie Meyer    4.8            53493 13.32
## 8           John Green    4.7            52620 12.99
## 9            E L James    3.9            51218 13.82
## 10    Alex Michaelides    4.5            50122 10.40
```

```
sqldf("select count(distinct Author) as author
      from bs")  # In 10 years, amazon book top 100 best sellers has 436 authors
```

```
##   author
## 1    436
```

```
sqldf("select Author
      from bs
      group by Author
      having count(Author) > 1
      limit 10") # data is having 217 authors who were best sellers more than 1 within 10 years
```

```
##                             Author
## 1                 Abraham Verghese
## 2                    Adam Mansbach
## 3                       Adam Rubin
## 4                     Adam Wallace
## 5                        Adir Levy
## 6       Admiral William H. McRaven
## 7                 Alex Michaelides
## 8                   Alice Schertle
## 9      America's Test Kitchen Kids
## 10 American Psychiatric Association
```

```
sqldf("select Author,count(Author) as best_sellers_frequency
      from bs
      group by Author
      having count(Author) > 1
      order by count(Author) desc
      limit 10") #within 217 authors who were best sellers more than 1 between 2010 to 2020. The Author
```

```
##                              Author best_sellers_frequency
## 1                      Rick Riordan                      18
## 2                   Suzanne Collins                      15
## 3                         Tom Rath                      13
## 4                       Jeff Kinney                      13
## 5                      John Grisham                      12
## 6                      Gary Chapman                      11
## 7                         Dr. Seuss                      11
## 8  American Psychological Association                     11
## 9                       Rob Elliott                      10
## 10                     Paulo Coelho                      10
```

```
bs%>%
  filter(Author == "Rick Riordan")%>%
  select(Year,Rank,Book_Title,Rating,Customers_review,Price)%>%
  arrange(Rank)# His best rank is 7 in 2013, and worst rank is 89 in 2017.
```

```
##    Year Rank
## 1  2013    7
## 2  2014    8
## 3  2012   13
## 4  2011   14
## 5  2010   25
## 6  2010   36
## 7  2011   37
## 8  2012   39
## 9  2010   42
```

```
## 10 2010   43
## 11 2016   62
## 12 2015   64
## 13 2010   67
## 14 2020   69
## 15 2017   81
## 16 2011   82
## 17 2016   88
## 18 2017   89
##
## 1                                                                                The House of Hades (Hei
## 2                                                                               The Blood of Olympus (Th
## 3                                                                                The Mark of Athena (Hei
## 4                                                                                The Son of Neptune (Hei
## 5                                                                                     The Lost Hero (Hei
## 6                                                                  The Last Olympian (Percy Jackson and
## 7                                                                                   The Throne of Fire (The
## 8                                                                               The Serpent's Shadow (The
## 9                                                                                    The Red Pyramid (The
## 10                                                               Percy Jackson and the Olympians Paperba
## 11                                                                              The Trials of Apollo, Book
## 12                Magnus Chase and the Gods of Asgard, Book 1: The Sword of Summer (Magnus Chase and
## 13                                                             The Battle of the Labyrinth (Percy Jackson and
## 14          Percy Jackson and the Olympians 5 Book Paperback Boxed Set (new covers w/poster) (Percy
## 15                                                             The Trials of Apollo Book Two The Dark Prophe
## 16                                                                                    The Lost Hero (Hei
## 17                                                                                 Magnus Chase and the Go
## 18 Magnus Chase and the Gods of Asgard, Book 3 The Ship of the Dead (Magnus Chase and the Gods of Asg
##    Rating Customers_review Price
## 1     4.8             8658  1.99
## 2     4.8             8286  1.54
## 3     4.8             7743  1.80
## 4     4.8             5540  1.00
## 5     4.8             5636 17.97
## 6     4.8             6450 16.49
## 7     4.8             1915 14.20
## 8     4.8             2704 17.15
## 9     4.7             2640  0.36
## 10    4.8              652  1.98
## 11    4.7             5031  1.49
## 12    4.7             4230  1.26
## 13    4.8             5727  0.94
## 14    4.9            14924 21.00
## 15    4.8             4148  8.23
## 16    4.8             5636 17.97
## 17    4.7             2495  0.74
## 18    4.8             3207  9.99
```

```
bs%>%
  filter(Author == "Rick Riordan")%>%
  arrange(Rank)%>%
  summarise(avg_rating = mean(Rating)) # he gained 4.78 rating of his book sales.
```

```
##   avg_rating
```

```
## 1   4.783333
```

```
bs%>%
  filter(Author == "Rick Riordan")%>%
  select(Year,Rank,Book_Title,Rating,Customers_review,Price)%>%
  arrange(Rank)%>%
  group_by(Year)%>%
  count()# the arrange of frequency of his best seller time. He has 5 books are best sales in 2010
```

```
## # A tibble: 9 x 2
## # Groups:   Year [9]
##    Year     n
##   <int> <int>
## 1  2010     5
## 2  2011     3
## 3  2012     2
## 4  2013     1
## 5  2014     1
## 6  2015     1
## 7  2016     2
## 8  2017     2
## 9  2020     1
```

```
bs%>%
  filter(Author == "Rick Riordan" & Year == 2010)%>%
  select(Book_Title,Rating,Customers_review,Price)
```

```
##                                                        Book_Title Rating
## 1                            The Lost Hero (Heroes of Olympus, Book 1)    4.8
## 2           The Last Olympian (Percy Jackson and the Olympians, Book 5)    4.8
## 3                         The Red Pyramid (The Kane Chronicles, Book 1)    4.7
## 4       Percy Jackson and the Olympians Paperback Boxed Set (Books 1-3)    4.8
## 5 The Battle of the Labyrinth (Percy Jackson and the Olympians, Book 4)    4.8
##   Customers_review Price
## 1             5636 17.97
## 2             6450 16.49
## 3             2640  0.36
## 4              652  1.98
## 5             5727  0.94
```

```
bs1<-bs%>%
  filter(Year %in% c(2017,2018,2019,2020))
```

```
sqldf("select *
      from bs1
      order by Customers_review desc
      limit 10") # where query order by customers_review , the data has the duplicated book appear twic
```

```
##    Year Rank
## 1  2020    3
## 2  2018   70
```

```
## 3  2019    1
## 4  2019   63
## 5  2020   14
## 6  2018    1
## 7  2019    2
## 8  2020    2
## 9  2020    5
## 10 2020   43
##                                                                    Book_Title
## 1                                                        Where the Crawdads Sing
## 2                                                        Where the Crawdads Sing
## 3                                                        Where the Crawdads Sing
## 4                                                        Where the Crawdads Sing
## 5                                                           Becoming (182 GRAND)
## 6                                                           Becoming (182 GRAND)
## 7                                                           Becoming (182 GRAND)
## 8  Too Much and Never Enough: How My Family Created the World's Most Dangerous Man
## 9                                                                    Midnight Sun
## 10                                                                The Silent Patient
##                 Author Rating Customers_review Price
## 1          Delia Owens    4.8           126619  4.55
## 2          Delia Owens    4.8           126306  4.58
## 3          Delia Owens    4.8           126302  4.65
## 4          Delia Owens    4.8           126302  6.08
## 5       Michelle Obama    4.8            85004  3.90
## 6       Michelle Obama    4.8            84761  3.91
## 7       Michelle Obama    4.8            84760  3.96
## 8  Mary L. Trump Ph.D.    4.6            84008 19.58
## 9      Stephenie Meyer    4.8            53493 13.32
## 10    Alex Michaelides    4.5            50320 10.40
```

```
sqldf("select count(distinct Book_Title) as uni_book
      from bs1") #257 unique books in 399 observations between 2017 to 2020
```

```
##   uni_book
## 1      257
```

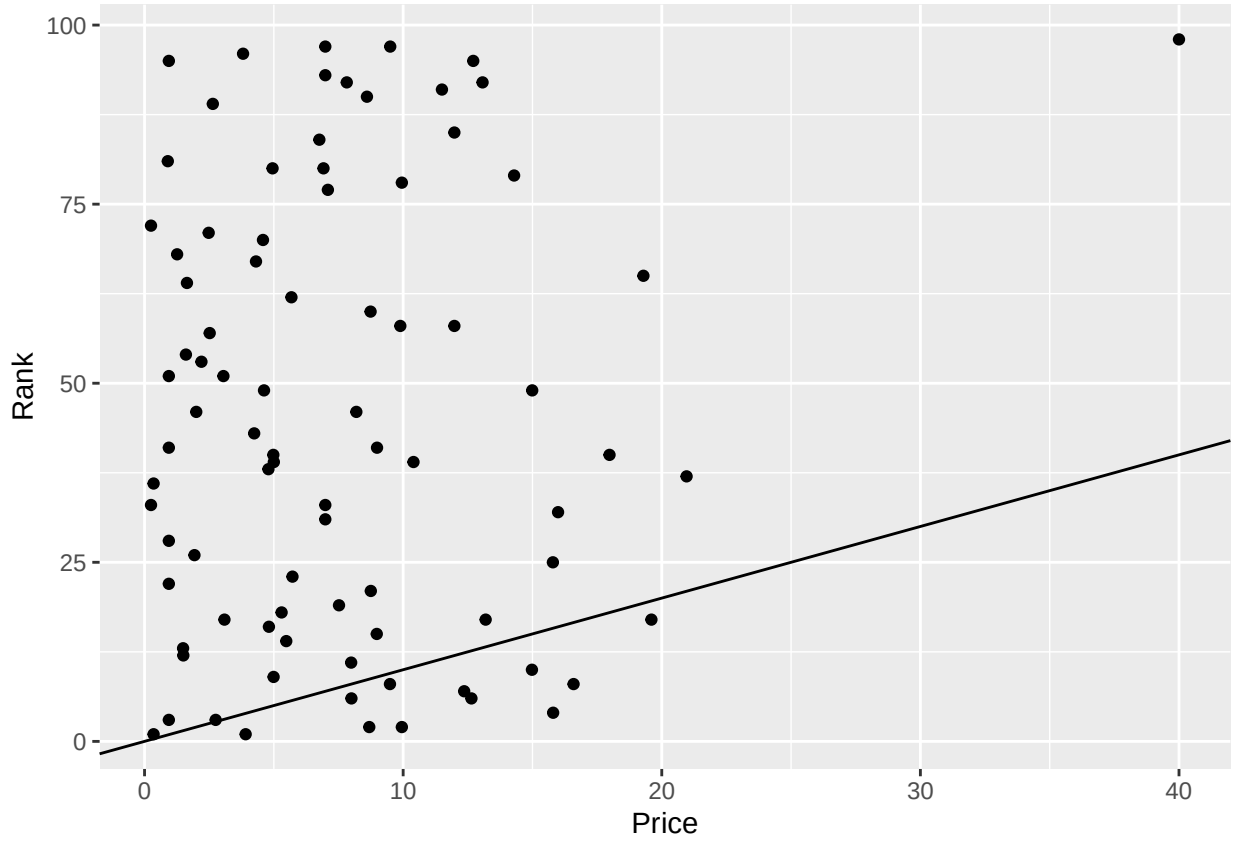
```
ad_price<-sqldf("select *
      from bs1
      group by Book_Title
      having count(Book_Title) > 1") #extra the data which is more than once in 2017 to 2020
```

```
head(ad_price)
```

```
##   Year Rank
## 1 2018    7
## 2 2017   17
## 3 2017    8
## 4 2019   58
## 5 2018    1
## 6 2017   85
##                                                          Book_Title
```

```
## 1                                   12 Rules for Life: An Antidote to Chaos
## 2                                                  1984 (Signet Classics)
## 3                                     Astrophysics for People in a Hurry
## 4 Atomic Habits: An Easy & Proven Way to Build Good Habits & Break Bad Ones
## 5                                                    Becoming (182 GRAND)
## 6                                       Beneath a Scarlet Sky: A Novel
##                Author Rating Customers_review Price
## 1  Jordan B. Peterson    4.7            27211 12.36
## 2       George Orwell    4.7            33234  3.09
## 3 Neil deGrasse Tyson    4.7            12157  9.49
## 4         James Clear    4.8            21640 11.98
## 5      Michelle Obama    4.8            84761  3.91
## 6       Mark Sullivan    4.7            42470 11.98
```

```
ad_price%>%
  ggplot(aes(Price,Rank))+geom_point()+geom_abline() # confirm again the Price and Rank do not have lin
```




```
sqldf("select *
      from bs1
      order by Rating desc
      limit 6") # best rating is 4.9 and worst rating is 4.0.
```

```
##   Year Rank
## 1 2017   15
## 2 2017   26
```

```
## 3 2017   33
## 4 2017   36
## 5 2017   37
## 6 2017   39
##                                                                                                  Book_Title
## 1                                                                                  Oh, the Places You'll Go!
## 2                                                                 The Wonderful Things You Will Be
## 3                                                                                       Goodnight Moon
## 4                                                                                  The Pout-Pout Fish
## 5 Harry Potter and the Prisoner of Azkaban: The Illustrated Edition (Harry Potter, Book 3)
## 6                                                                 Brown Bear, Brown Bear, What Do You See?
##                  Author Rating Customers_review Price
## 1             Dr. Seuss    4.9            16856  8.98
## 2 Emily Winfield Martin    4.9            12069  1.93
## 3   Margaret Wise Brown    4.9            13055  0.25
## 4        Deborah Diesen    4.9            13900  0.35
## 5          J.K. Rowling    4.9             4658 27.98
## 6       Bill Martin Jr.    4.9            21233  5.00
```

Since average rating is 4.7 in 2017 to 2020 top book sales. we define the 4.7 as minium level as quality book.

```
bs1%>%
  summarise(avg_rating = mean(Rating)) # average rating is 4.7
```

```
##   avg_rating
## 1   4.697744
```

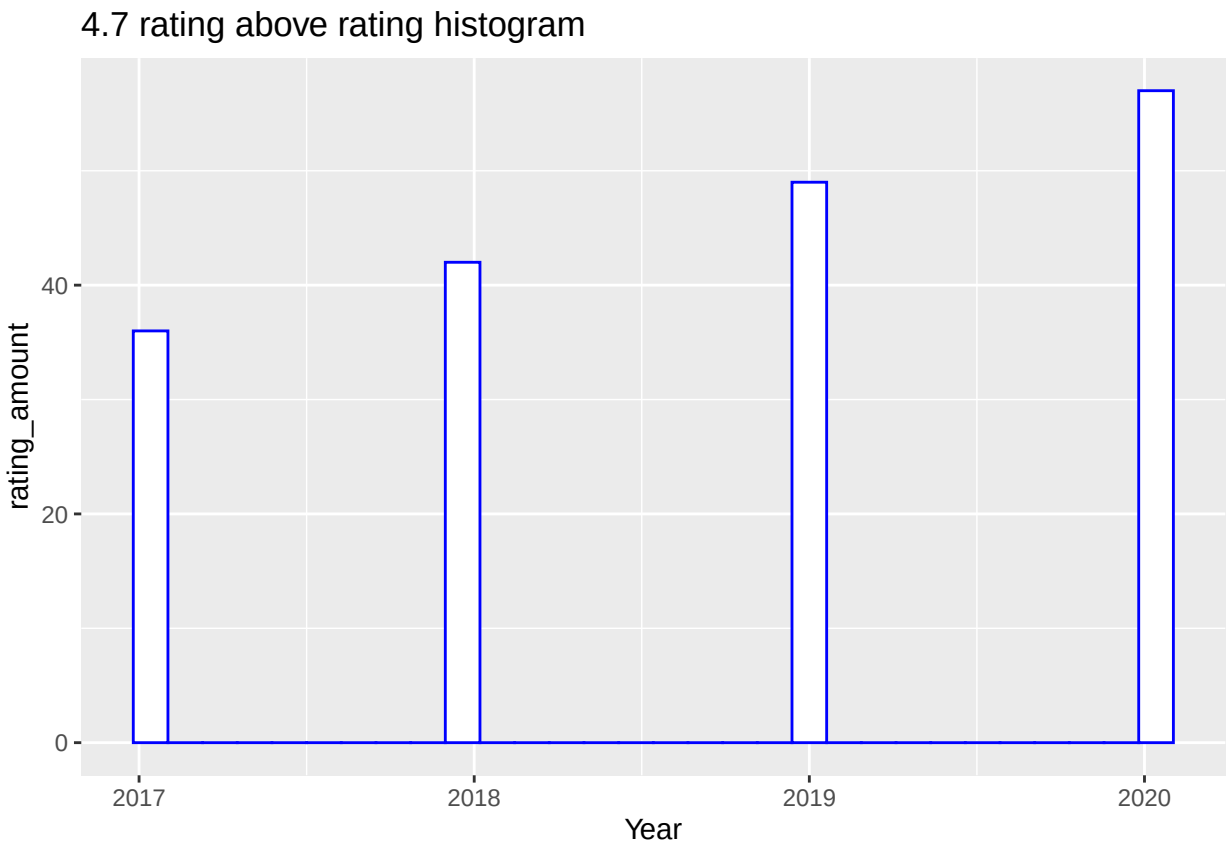
```
ad_rating<-bs1%>%
  filter(Rating > 4.7) #extract the quality book
```

```
head(ad_rating)
```

```
##   Year Rank                                                Book_Title        Author
## 1 2017    1                                                    Wonder R. J. Palacio
## 2 2017    3                  Giraffes Can't Dance (Board Book) Giles Andreae
## 3 2017    5                                            The Getaway   Jeff Kinney
## 4 2017   12 The 5 Love Languages: The Secret to Love that Lasts  Gary Chapman
## 5 2017   15                          Oh, the Places You'll Go!     Dr. Seuss
## 6 2017   21                            The Sun and Her Flowers     Rupi Kaur
##   Rating Customers_review Price
## 1    4.8            26045  0.35
## 2    4.8            20385  0.94
## 3    4.8             8650 16.99
## 4    4.8            37833  1.50
## 5    4.9            16856  8.98
## 6    4.8             8177  8.75
```

The below histogram chart displays the rating of book. according to the graph, amazon top book sales are increased their quality with year. means that the quality of readers are raised and more chase to the good book. the requirement of customers is going up too.

```
ad_rating%>%
  ggplot(aes(Year))+geom_histogram(color ="blue",fill="white",bins = 30)+ylab("rating_amount")+ggtitle(
```



Altought the price does not have linear relationship with rank, we still want to check the price because the rank is according to sales' amount.

```
bs1%>%
  arrange(desc(Price))%>%
  group_by(Author)%>%
  head(50) # extract top 50 expensive book in 2017 to 2020
```

```
## # A tibble: 50 x 7
## # Groups:   Author [36]
##     Year  Rank Book_Title                  Author   Rating Customers_review Price
##    <int> <int> <chr>                       <chr>     <dbl>            <int> <dbl>
##  1  2017    48 Obama: An Intimate Portra~ Pete So~    4.9              5908  64.8
##  2  2017    98 Harry Potter Paperback Bo~ J. K. R~    4.8             30952  40
##  3  2018    99 Harry Potter Paperback Bo~ J. K. R~    4.8             30952  40
```

```
##  4  2017    30 Thug Kitchen: The Officia~ Thug Ki~    4.6            11897  31.7
##  5  2017    37 Harry Potter and the Pris~ J.K. Ro~    4.9             4658  28.0
##  6  2020     1 A Promised Land            Barack ~    4.9            34830  27
##  7  2018   100 Xanathar's Guide to Every~ Wizards~    4.9            11311  27.0
##  8  2017    76 Grant                      Ron Che~    4.8             5121  27.0
##  9  2018     5 Fear: Trump in the White ~ Bob Woo~    4.5             8154  26.3
## 10  2017    19 Origin: A Novel (Robert L~ Dan Bro~    4.3            23011  25.2
## # ... with 40 more rows
```

```
bs1%>%
  group_by(Author)%>%
  count()%>%
  arrange(desc(n))%>%
  head()#best author in 2017 to 2020
```

```
## # A tibble: 6 x 2
## # Groups:   Author [6]
##   Author               n
##   <chr>            <int>
## 1 Dav Pilkey          10
## 2 Wizards RPG Team     7
## 3 Bill Martin Jr.      5
## 4 BrenÃ© Brown         5
## 5 Joanna Gaines        5
## 6 John Grisham         5
```

Since the Dav Pikey appear 10 and Wizards RPG Team appear 7 times. base on the query to analysis what kind of the book is tendency on 2017 to 2020.

```
bs_author_2017_2020<-bs1%>%
  filter(Author=="Dav Pilkey" | Author == "Wizards RPG Team")
```

```
bs_author_2017_2020%>%
  arrange(desc(Price))
```

```
##    Year Rank
## 1  2018  100
## 2  2018   76
## 3  2019   38
## 4  2018   71
## 5  2017   40
## 6  2018   21
## 7  2020  100
## 8  2019    6
## 9  2020   81
## 10 2017   86
## 11 2020   12
## 12 2019    4
## 13 2018   16
## 14 2018   35
## 15 2018   40
```

```
## 16 2019   22
## 17 2017   47
##                                                                                Book_Title
## 1                                      Xanathar's Guide to Everything (Dungeons & Dragons)
## 2                  Dungeons & Dragons Monster Manual (Core Rulebook, D&D Roleplaying Game)
## 3                                                 Player's Handbook (Dungeons & Dragons)
## 4          Dungeons & Dragons Dungeon Master's Guide (Core Rulebook, D&D Roleplaying Game)
## 5                                                 Player's Handbook (Dungeons & Dragons)
## 6                                                 Player's Handbook (Dungeons & Dragons)
## 7                                                 Player's Handbook (Dungeons & Dragons)
## 8                   Dog Man: Fetch-22: From the Creator of Captain Underpants (Dog Man #8)
## 9                   Dog Man: Fetch-22: From the Creator of Captain Underpants (Dog Man #8)
## 10                  Dog Man Unleashed: From the Creator of Captain Underpants (Dog Man #2)
## 11 Dog Man: Grime and Punishment: From the Creator of Captain Underpants (Dog Man #9) (9)
## 12  Dog Man: For Whom the Ball Rolls: From the Creator of Captain Underpants (Dog Man #7)
## 13         Dog Man: Lord of the Fleas: From the Creator of Captain Underpants (Dog Man #5)
## 14                Dog Man and Cat Kid: From the Creator of Captain Underpants (Dog Man #4)
## 15         Dog Man: Brawl of the Wild: From the Creator of Captain Underpants (Dog Man #6)
## 16         Dog Man: Brawl of the Wild: From the Creator of Captain Underpants (Dog Man #6)
## 17    Dog Man: A Tale of Two Kitties: From the Creator of Captain Underpants (Dog Man #3)
##               Author Rating Customers_review Price
## 1  Wizards RPG Team    4.9            11311 26.99
## 2  Wizards RPG Team    4.9            12582 20.75
## 3  Wizards RPG Team    4.8            24160 18.99
## 4  Wizards RPG Team    4.9            13207 18.50
## 5  Wizards RPG Team    4.8            24159 17.98
## 6  Wizards RPG Team    4.8            24159 17.98
## 7  Wizards RPG Team    4.8            24240 17.98
## 8        Dav Pilkey    4.9            17504  8.00
## 9        Dav Pilkey    4.9            17561  8.00
## 10       Dav Pilkey    4.9             8119  6.70
## 11       Dav Pilkey    4.9            22821  6.49
## 12       Dav Pilkey    4.9            13076  6.43
## 13       Dav Pilkey    4.9             8213  5.11
## 14       Dav Pilkey    4.9             7711  5.11
## 15       Dav Pilkey    4.9            10683  4.98
## 16       Dav Pilkey    4.9            10683  4.98
## 17       Dav Pilkey    4.9             7436  4.84
```

```
sqldf("select Author,count(distinct Book_Title) as book_quantity
      from bs_author_2017_2020
      group by Author")
```

```
##              Author book_quantity
## 1       Dav Pilkey             8
## 2 Wizards RPG Team             4
```

After the analysis, even the author: Dav Pilkey has mutiple book appear to top book sales, but his price is not higher, and his book basiclly like cartoon story. we can assume that the book sells higher because the story is pre-school education. On other side, the Wizards RPG Team sells 4 books which are related to game guide. Even book price is bit higher, but still make a good sales of quantity. Also, only 4 books customers' review is bit higher than Dav Pilkey.

```
bs_author_2017_2020%>%
  group_by(Author)%>%
  summarise(total_customer_review=sum(Customers_review))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   Author           total_customer_review
##   <chr>                            <int>
## 1 Dav Pilkey                      123807
## 2 Wizards RPG Team                133818
```

Conclusion

The information about Amazon top book sales given by this data is predictive. We can make recommended books on the website according to the data to increase the amount of visits, clicks and purchases of customers.Also, it can absorb potential authors and help them better promote their books.At the same time, according to the analysis of books from 2017 to 2020, people's purchasing power, quality demand, and rate of review interaction has increased significantly.Moreover, the genre of books sold has shifted from storytelling to games and pre-school education.The data show that book buyers are young people and families with children.On the other hand, it also shows part of the purchasing power of life habits, which can be corresponding to the recommendation of companion products, such as game, game equipment or children's products.