Analysising The Amazon's Top Bestseller From 2009 To 2019

By Sk Israk Sahan

Roll No.-STI203007

Registrations No.-

0719 of 2020-2021

CONTENT

INTRODUCTION

Well as we all love to read book of different types, but we often face many difficulties for choosing the writers or author whose book should we read or which genres book should we read.

My project may help us to solve some of these kinds of problems.

So then come, What I have done in the project?

This project is about the Amazon Top 50 Bestselling Books 2009-2019(https://github.com/dphiofficial2009-2019 (https://github.com/dphi-official/Datasets/blob/master/Amazon%20Top%2050%20Bestselling%20Books%202009%20-%202019.csv). The Dataset contains 550 books. Data has been categorized into fiction and non-fiction using Goodreads. The analysis of this Dataset will allow us have a deep understanding of the book market trends over the past decade.

I have taken references from many other websites and projects present in Jovian website while working in this project. Some of those websites are the following:

1. geeksforgeeks.org
2. w3schools.com
3. stackoverflow.com
4. jovian.ai

IN SIMPLE TERMS HERE I AM JUST DOING A SIMPLE DATA ANALYSIS OF THE ABOVE GIVEN DATA.

Now as I am do a data analysis of the above given data we must first know what is this DATA ANALYSIS

WHAT IS DATA ANALYSIS?

Data analysis refers to the process of inspecting, cleaning, transforming, and interpreting data in order to discover useful information, draw conclusions, and support decision-making.

The primary goal of data analysis is to derive valuable insights and knowledge from data that can drive informed decision-making and improve understanding of a particular phenomenon or problem.

Overall, data analysis helps to transform raw data into actionable information, enabling organizations and individuals to make evidence-based decisions and gain a deeper understanding of the data they have collected.

NOW LET SEE THE DESCRIPTION OF THE DATA

DATASET description:

This data set includes seven categories, such as Name of the Book, the author of the Book, Amazon User Rating, Number of written reviews on amazon, The price of the book, The Year(s) it ranked on the bestseller, Whether fiction or non-fiction.

I will select the following variables for data analysis. Categorical variable: Name, Author, Genre, Year Numerical variables: User Rating, Reviews, Price.

FEATURES:

1. **Name** - Name of the Book

2. **Author** - The author of the Book
3. **User Rating** - Amazon User Rating
4. **Reviews** - Number of written reviews on amazon
5. **Price** - The price of the book (As at 13/10/2020)
6. **Year** - The Year(s) it ranked on the bestseller
7. **Genre** - Whether fiction or non-fiction

DATA TABLE

Name	Author	User Rating	Reviews	Price	Year	Genre
10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
5,000 Awesome Facts (About Everything!) (National Geographic Kids)	National Geographic Kids	4.8	7665	12	2019	Non Fiction

....

Wonder	R. J. Palacio	4.8	21625	9	2017	Fiction
Wrecking Ball (Diary of a Wimpy Kid Book 14)	Jeff Kinney	4.9	9413	8	2019	Fiction
You Are a Badass: How to Stop Doubting Your Greatness and Start Living an Awesome Life	Jen Sincero	4.7	14331	8	2016	Non Fiction
You Are a Badass: How to Stop Doubting Your Greatness and Start Living an Awesome Life	Jen Sincero	4.7	14331	8	2017	Non Fiction
You Are a Badass: How to Stop Doubting Your Greatness and Start Living an Awesome Life	Jen Sincero	4.7	14331	8	2018	Non Fiction
You Are a Badass: How to Stop Doubting Your Greatness and Start Living an Awesome Life	Jen Sincero	4.7	14331	8	2019	Non Fiction

The given data is of 550 rows and 7 columns as so here I only provide the 1st five and the last five observation of the data table.

Observation

Here we are going to answer the following questions according to data using Data Analysis through R programming.

1. Which author's books receive the highest average rating (top authors)?
2. Which author has written the most bestsellers (top authors)?
3. Which book has the most reviews (top books)?
4. Which genres become bestsellers more often?
5. Which book have the most editions as bestseller?
6. The correlation between User rating, review, and Price

RESULTS of ANALYSIS

1. Which author's books receive the highest average rating (top authors)?

	author	year	name	genre	average_user_rate
1	Alice Schertle	2014	Little Blue Truck	Fiction	4.9

2	Bill Martin Jr.	2017	Brown Bear, Brown Bear, What Do You See?	Fiction	4.9
3	Bill Martin Jr.	2019	Brown Bear, Brown Bear, What Do You See?	Fiction	4.9
4	Brandon Stanton	2015	Humans of New York : Stories	Non Fiction	4.9
5	Chip Gaines	2016	The Magnolia Story	Non Fiction	4.9
6	Dav Pilkey	2017	Dog Man: A Tale of Two Kitties: From the Creator of Captain Underpants (Dog Man #3)	Fiction	4.9

.....

This is the list of the top 5 best seller according to the data.

	author	count_average_user_rate
1	Dr. Seuss	8
2	Dav Pilkey	7
3	Eric Carle	7
4	Sarah Young	6
5	Emily Winfield Martin	4
6	J.K. Rowling	3

Now using R, we count the number of bestsellers written by a Author is 8 and the name of the author is Dr. Seuss

Therefore, Dr. Seuss is the best Author as per the Rating of the bestsellers.

2. Which author has written the most bestsellers (top authors)?

	author	count_author
1	Jeff Kinney	12
2	Gary Chapman	11
3	Rick Riordan	11
4	Suzanne Collins	11
5	American Psychological Association	10

By using R, we found that the maximum number of bestsellers are written by Jeff Kinney

Therefore, Jeff Kinney is the top Author as per the number of the bestseller he has written

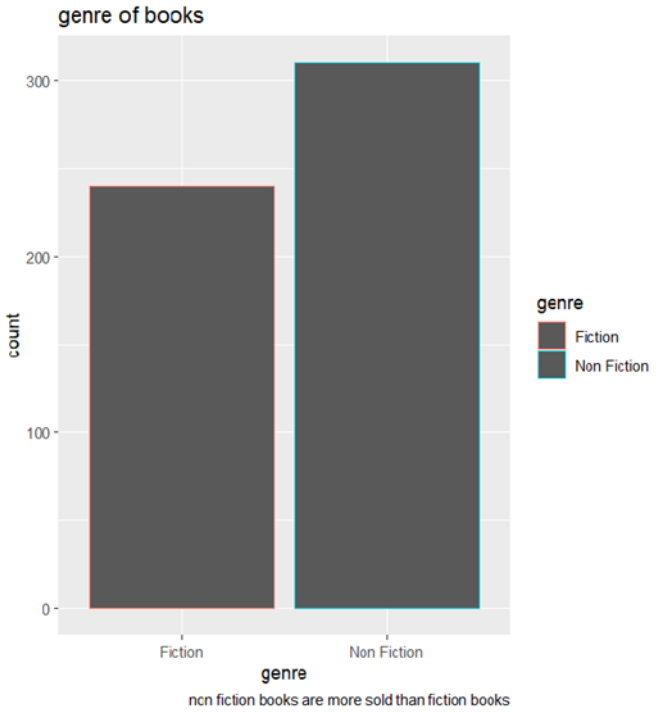
3. Which book has the most reviews (top books)?

	author	year	name	genre	average_user_reviews	average_user_rate
1	Delia Owens	2019	Where the Crawdads Sing	Fiction	87841	4.8
2	Paula Hawkins	2015	The Girl on the Train	Fiction	79446	4.1
3	Paula Hawkins	2016	The Girl on the Train	Fiction	79446	4.1
4	Michelle Obama	2018	Becoming	Non Fiction	61133	4.8
5	Michelle Obama	2019	Becoming	Non Fiction	61133	4.8

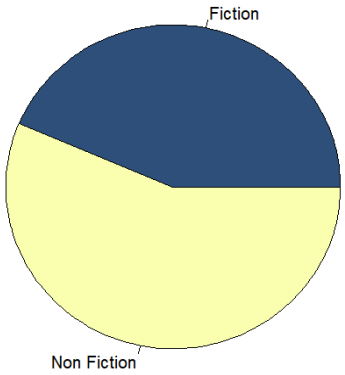
Here also by using R, we found that the most reviews receive by the book named "***Where the Crawdads sing***" written by Delia Owens

Therefore, ***Where the Crawdads sing*** is the top book as it has been discuss more among the users.

4. Which genres become bestsellers more often?



The count of Fictional bestsellers is 240 and the count of the non-Fictional bestsellers is 310



The percentage of the non-fictional bestsellers is 56% and the percentage of fictional bestsellers is 44%.

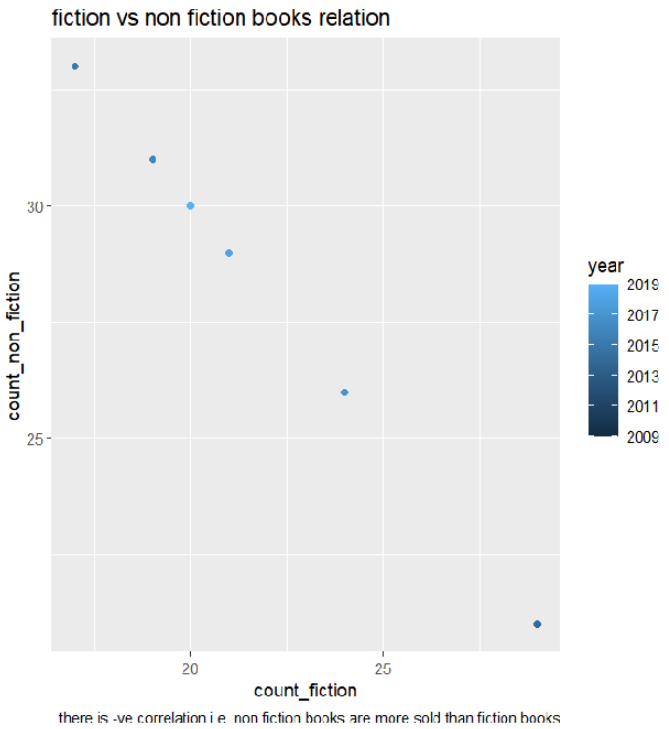
	year	count_fiction
1	2014	29
2	2009	24
3	2013	24
4	2017	24
5	2011	21
6	2012	21
7	2018	21
8	2010	20
9	2019	20

10	2016	19
11	2015	17

From this table we can say that in the year of 2014, 29 different fictional books became bestseller.

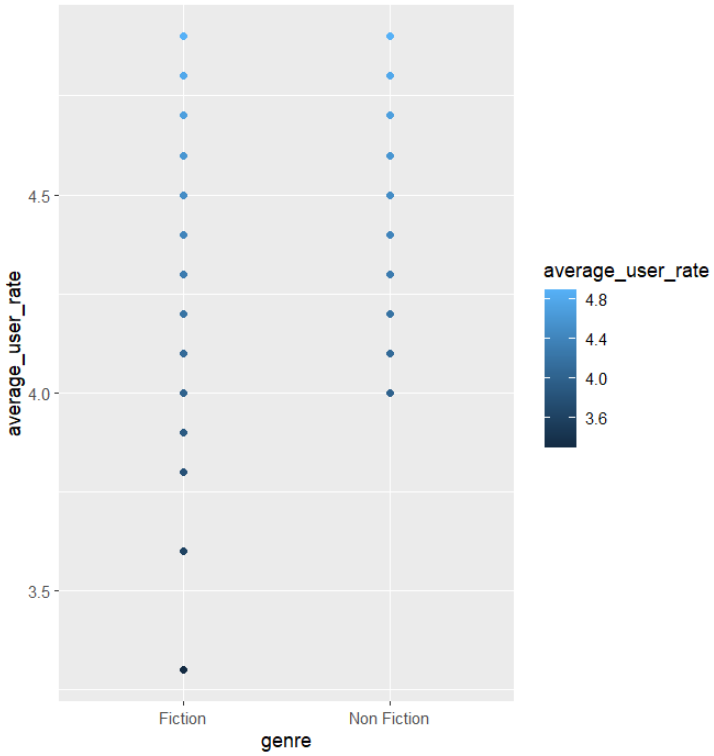
	year	count_non_fiction
1	2015	33
2	2016	31
3	2010	30
4	2019	30
5	2011	29
6	2012	29
7	2018	29
8	2009	26
9	2013	26
10	2017	26
11	2014	21

From this table we can say that in the year of 2015, 33 different non fictional books became bestseller.



There is negative correlation i.e., non-fiction books are more sold than fiction books

5. What is the range of fictional and non-fictional bestsellers rating?



The range for the frictional bestsellers rating is from 3.3 to 4.9 and similarly the range of the non-fictional bestsellers is from 4.0 to 4.9.

6. Which book have the most editions as bestseller?

	name	edition
1	Publication Manual of the American Psychological Association, 6th Edition	10
2	StrengthsFinder 2.0	9
3	Oh, the Places You'll Go!	8
4	The 7 Habits of Highly Effective People: Powerful Lessons in Personal Change	7
5	The Very Hungry Caterpillar	7

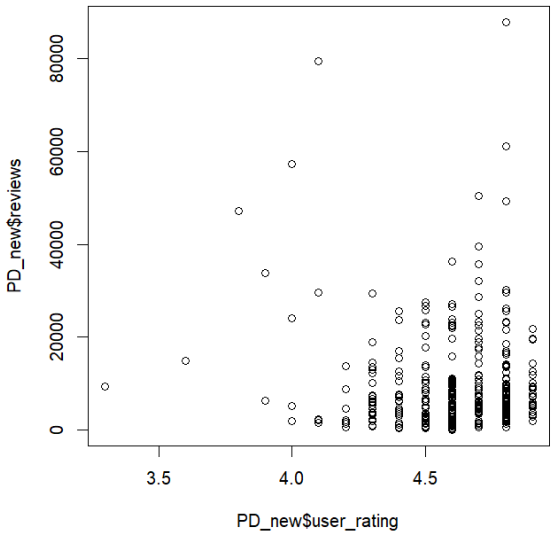
The Publication Manual of the American Psychological Association has the maximum number of editions.

CONCLUSION: By analysing the categorical data, it is established:

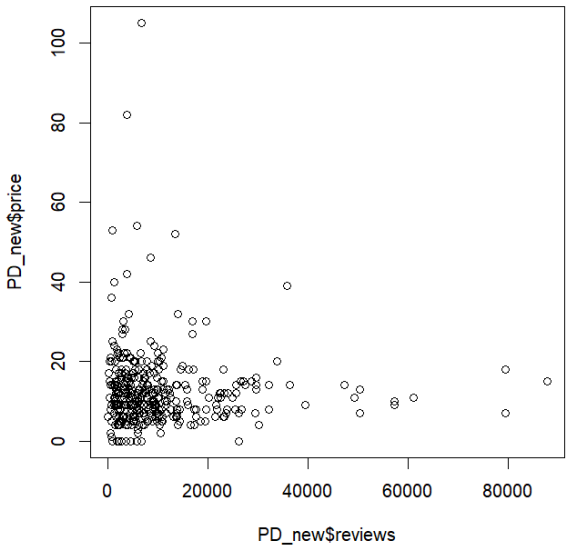
1. The following 13 authors have the highest rating: Nathan W. Pyle, Patrick Thorpe, Eric Carle, Emily Winfield Martin, Chip Gaines, Jill Twiss, Rush Limbaugh, Sherri Duskey Rinker, Alice Schertle, Pete Souza, Sarah Young, Lin-Manuel Miranda, Bill Martin Jr., Dav Pilkey. The average rating for their works was 4.9. When buying a new book, we should pay attention to these authors.
2. Authors who have written more bestsellers: Jeff Kinney- 12 books, Rick Riordan- 10 books, J.K. Rowling- 8 books, Stephenie Meyer- 7 books, Dav Pilkey- 6 books, Bill O'Reilly- 6 books, John

Grisham- 5 books, E L James- 5 books, Suzanne Collins- 5 books, Charlaine Harris- 4 books. These authors always have something to read.
3. Books with the most reviews: Where The Crawdads Sing- 87841 Reviews, The Girl On The Train - 79446 Reviews, Becoming- 61133 Reviews, Gone Girl- 57271 Reviews, The Fault In Our Stars- 50482 Reviews. It's definitely worth reading the book Where The Crawdads Sing, it's not for nothing that it is the most talked about.
4. Non-fiction is more likely to become a bestseller.

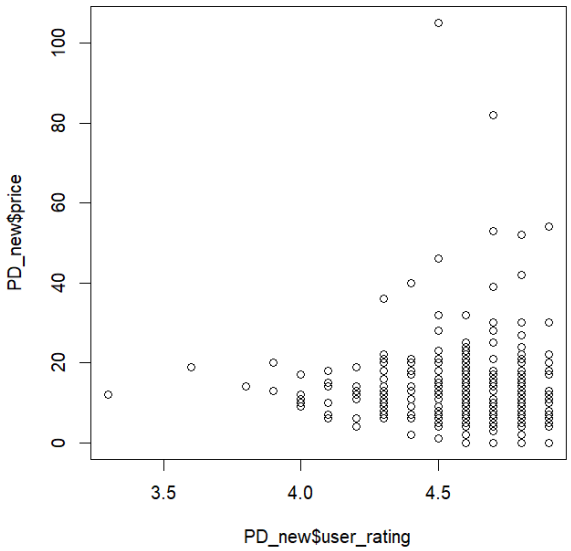
7. The correlation between User rating, review, and Price



From this graph we can conclude that as the User rating goes up the review also increases along with it.



From this graph we can conclude that the price of the books doesn't affect the reviews



Here also in this graph we can say that it is same as the above Price Vs Review, i.e., the price doesn't affect the user rating of the books.

	PD_new.user_rating	PD_new.reviews	PD_new.price
PD_new.user_rating	1	-0.001729014	-0.133086287
PD_new.reviews	-0.001729014	1	-0.109181883
PD_new.price	-0.133086287	-0.109181883	1

CONCLUSION: Based on the constructed correlation matrix as well as the constructed visualizations, it can be seen that the data does not contain any positive or negative linear relationship between the rating, reviews and the price of books.

Final Conclusion

In the course of the analysis, it was established which authors receive the highest ratings from readers, which authors have written the most bestsellers, which books receive the most reviews from readers. In addition, it was found that non-fiction literature is becoming more often a bestseller, but users also like fictional books.

References:

- *The dataset used in this project was downloaded from :* https://github.com/dphi-official/Datasets/blob/master/Amazon%20Top%2050%20Bestselling%20Books%202009%20-%202019.csv

- *I have taken references from many other websites and projects present in Jovian website while working in this project. Some of those websites are the following:*
    1. *geeksforgeeks.org*
    2. *w3schools.com*
    3. *stackoverflow.com*
    4. *jovian.ai*

Source

1. The Amazon Top 50 Bestselling Books 2009-2019(https://github.com/dphi-official/Datasets/blob/master/Amazon%20Top%2050%20Bestselling%20Books%202009%20-%202019.csv)
2. For Cording I have used many different websites to find the specific codes, and use Chat GPT to understand them and use them according to the data.

PAKEAGE needed:

1. install.packages("dplyr")
2. install.packages("RcppNumerical")
3. install.packages("stringdist")
4. install.packages("reticulate")
5. install.packages("ggplot2")
6. install.packages("ggplot2")
7. install.packages("plotly")
8. install.packages("base")
9. install.packages("stats")
10. install.packages("Matrix")
11. install.packages("pracma")
12. install.packages("signal")
13. install.packages("tidyverse")
14. install.packages("janitor")
15.install.packages(“plot_ly”)

NOTE- The last three package are the most important packages that are needed in this whole process

LIBRARY USED HERE ARE:

```
library(dplyr)
library(RcppNumerical)
library(stringdist)
library(reticulate)
library(ggplot2)
library(plotly)
library(base)
library(stats)
library(Matrix)
```

```
library(pracma)
library(signal)
library(tidyverse)
library(janitor)
library(plot_ly)
```

Note- Here also the last two are most important

Codes

```
library(tidyverse)

library(janitor)

library(plotly)

# Reading data

PD=read.csv("Book Data.csv")

PD

head(PD)

tail(PD)

#Summary

glimpse(PD)

str(PD)

#FINDING the missing value

sum(is.na(PD))

#as it is 0 therefore the data is unbiased


#Name of the columns

colnames(PD)


PD_new=clean_names(PD)

head(PD_new)

PD_books_in_year=PD_new%>%

 group_by(year)%>%

 summarise(count_year=n())

PD_books_in_year
```

```
ggplot(data=PD_new)+
  geom_bar(mapping=aes(x=year))+
  labs(title="books in each year",caption="in each year equal numbers books are published")

#RATING

PD_author_average_rate=PD_new%>%
  group_by(author,year,name,genre)%>%
  summarize(average_user_rate=mean(user_rating))%>%
  arrange(desc(average_user_rate))
head(PD_author_average_rate)
PD_author_average_rate
ggplot(data=PD_author_average_rate)+
  geom_point(mapping=aes(x=genre,y=average_user_rate,color=average_user_rate))

PD_author_average_rate%>%
  filter(average_user_rate==4.9)%>%
  group_by(author)%>%
  summarize(count_average_user_rate=n())%>%
  filter(count_average_user_rate==max(count_average_user_rate))
PD_rate=PD_author_average_rate%>%
  filter(average_user_rate==4.9)%>%
  group_by(author)%>%
  summarize(count_average_user_rate=n())%>%
  arrange(desc(count_average_user_rate))
PD_rate
#Therefore Dr.Seuss is the best author according to the rating as he got the maximum number of highest rating books

tail(PD_author_average_rate)
```

```
#Book  Count

PD_books=PD_books_in_author_desc=PD_new%>%
group_by(author)%>%
 summarise(count_author=n())%>%
 arrange(desc(count_author))
PD_books
head(PD_books_in_author_desc)
tail(PD_books_in_author_desc)
#The author with most bestseller count is Jeff Kinney

#Review

PD_author_average_user_reviews_rate=PD_new%>%
 group_by(author,year,name,genre)%>%
 summarize(average_user_reviews=mean(reviews),average_user_rate=mean(user_rating))%>%
 arrange(desc(average_user_rate),desc(average_user_reviews))
head(PD_author_average_user_reviews_rate)
tail(PD_author_average_user_reviews_rate)
PD_reviews=PD_author_average_user_reviews=PD_new%>%
 group_by(author,year,name,genre)%>%
 summarize(average_user_reviews=mean(reviews),average_user_rate=mean(user_rating))%>%
 arrange(desc(average_user_reviews))
head(PD_author_average_user_reviews)
PD_reviews
#therefore the maximum number of review receive by Delia Owens
tail(PD_author_average_user_reviews)

#Genre

ggplot(data=PD_new)+
```

```
  geom_bar(mapping=aes(x=genre,color=genre))+
  labs(title="genre of books",caption="non fiction books are more sold than fiction books")
pie(PD_new$genre)

XX <- table(PD_new$genre)

# Pie
pie(XX,col = hcl.colors(length(XX), "BluYl"))

PD_fiction_by_year=PD_new%>%
  filter((genre=="Fiction"))%>%
  group_by(year)%>%
  summarise(count_fiction=n())%>%
  arrange(desc(count_fiction))
PD_fiction_by_year
#Friction books of maximum number in the year of 2014 according to the data

PD_non_fiction_by_year=PD_new%>%
  filter((genre=="Non Fiction"))%>%
  group_by(year)%>%
  summarise(count_non_fiction=n())%>%
  arrange(desc(count_non_fiction))
PD_non_fiction_by_year

fiction_no_fiction=merge(PD_fiction_by_year,PD_non_fiction_by_year,by="year")
fiction_no_fiction

fiction_no_fiction%>%mutate(total_in_year=count_fiction+count_non_fiction)

cor(PD_non_fiction_by_year$count_non_fiction,PD_fiction_by_year$count_fiction)
```

```
ggplot(data=fiction_no_fiction)+
  geom_point(mapping=aes(x=count_fiction,y=count_non_fiction,color=year))+
  labs(title="fiction vs non fiction books relation",
       caption="there is -ve correlation i.e. non fiction books are more sold than fiction books")

PD_fiction_by_author=PD_new%>%
  filter((genre=="Fiction"))%>%
  group_by(author)%>%
  summarise(count_author_fiction=n())%>%
  arrange(desc(count_author_fiction))
head(PD_fiction_by_author)
tail(PD_fiction_by_author)
PD_fiction_by_author
#Jeff Kinney has given the maximum number of fictional bestseller

PD_non_fiction_by_author=PD_new%>%
  filter((genre=="Non Fiction"))%>%
  group_by(author)%>%
  summarise(count_non_author_fiction=n())%>%
  arrange(desc(count_non_author_fiction))
head(PD_non_fiction_by_author)
tail(PD_non_fiction_by_author)
PD_non_fiction_by_author
#and Gray Chapman has given the maximum number of non frictional bestseller

#Edition

PD_editions=PD_new%>%
  group_by(name)%>%
  summarize(edition=n())%>%
  arrange(desc(edition))
```

```
head(PD_editions)

PD_editions

tail(PD_editions)

#The Publication Manual of the American Psychological Association has the maximum number of edition


#Correlation Matrix


cor(PD_new$user_rating,PD_new$reviews)

cor(PD_new$user_rating,PD_new$price)

cor(PD_new$price,PD_new$reviews)

PD_12=data.frame(PD_new$user_rating,PD_new$reviews,PD_new$price)

PD_12

cor(PD_12)

plot(PD_new$user_rating,PD_new$reviews)

plot(PD_new$user_rating,PD_new$price)

plot(PD_new$reviews,PD_new$price)
```

ACKNOWLEDGEMENT

I would like to express my profound gratitude to Prof. Dr. Joydeep Sengupta(HOD), of Mathematics and Statistics department, of Aliah university for their contributions to the completion of my project titled “Analysing the Amazon’s top bestseller from 2009- 2019” .

I would like to express my special thanks to our whole department for their time and efforts which they provided throughout the semester. Everyone’s useful advice and suggestions were really helpful to me during the project’s completion. In this aspect, I am eternally grateful to every one of my department.

I would like to acknowledge that this project was completed entirely by me and not by someone else.

Sk Israk Sahan

THANK YOU