Amazon top selling Books

kainat Liaqat

2024-01-20

Scenario

The project Amazon best selling books from 2009 to 2019 in R is Exploratory data analysis to find insights and patterns about the best-selling books in different genres and categories using the R programming language.To begin analysis i have installed and loaded the following packages:

```
install.packages("tidyverse")
install.packages("skimr")
install.packages("janitor")
install.packages("ggplot2")
install.packages("reshape2")
install.packages("ggthemes")
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggthemes)
```

Loading the Data

Used kaggle.com for data source.

```
knitr::opts_chunk$set(error = TRUE, warning = FALSE)

library(readr)
amazon_best_selling_books <- read_csv("Course 7/Week 5/amazon_best_selling_books.csv")
```

```
## Rows: 550 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Name, Author, Genre
## dbl (4): User Rating, Reviews, Price, Year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# utils::View(amazon_best_selling_books)
```

Exploration

Cleaning data

```
str(amazon_best_selling_books)
```

```
## spc_tbl_ [550 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Name       : chr [1:550] "10-Day Green Smoothie Cleanse" "11/22/63: A Novel" "12 Rules for Life: A
##  $ Author     : chr [1:550] "JJ Smith" "Stephen King" "Jordan B. Peterson" "George Orwell" ...
##  $ User Rating: num [1:550] 4.7 4.6 4.7 4.7 4.8 4.4 4.7 4.7 4.7 4.6 ...
##  $ Reviews    : num [1:550] 17350 2052 18979 21424 7665 ...
##  $ Price      : num [1:550] 8 22 15 6 12 11 30 15 3 8 ...
##  $ Year       : num [1:550] 2016 2011 2018 2017 2019 ...
##  $ Genre      : chr [1:550] "Non Fiction" "Fiction" "Non Fiction" "Fiction" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Name = col_character(),
##   ..   Author = col_character(),
##   ..   `User Rating` = col_double(),
##   ..   Reviews = col_double(),
##   ..   Price = col_double(),
##   ..   Year = col_double(),
##   ..   Genre = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

column names:

```
colnames(amazon_best_selling_books)
```

```
## [1] "Name"        "Author"      "User Rating" "Reviews"     "Price"
## [6] "Year"        "Genre"
```

Changing data type from double to integer

```
amazon_best_selling_books$Reviews <- as.integer(amazon_best_selling_books$Reviews)
amazon_best_selling_books$Year <- as.integer(amazon_best_selling_books$Year)
amazon_best_selling_books$Price <- as.integer(amazon_best_selling_books$Price)
```

Checking for any missing values

```
any(is.na(amazon_best_selling_books))
```

```
## [1] FALSE
```

Renaming the column user rating

```
amazon_best_selling_books <- amazon_best_selling_books %>% rename(User_Ratings = "User Rating")
```

Maximum Price of the books

```
max(amazon_best_selling_books$Price)
```

```
## [1] 105
```

Minimun Price of the books

```
min(amazon_best_selling_books$Price)
```

```
## [1] 0
```

Average Price of the books

```
mean(amazon_best_selling_books$Price)
```

```
## [1] 13.1
```

Unique values

```
unique(amazon_best_selling_books$Year)
```

```
##  [1] 2016 2011 2018 2017 2019 2014 2010 2009 2015 2013 2012
```

```
unique(amazon_best_selling_books$Genre)
```

```
## [1] "Non Fiction" "Fiction"
```

Authors who achieved highest ratings for their books

```
Top_authors <- amazon_best_selling_books %>%
  filter(User_Ratings> 4.8 & Reviews >= 12000) %>%
  group_by(Name) %>%
  distinct(Name, .keep_all = TRUE)

print(unique(Top_authors$Author))
```

```
## [1] "Bill Martin Jr." "Dav Pilkey"      "J.K. Rowling"    "Sarah Young"
## [5] "Dr. Seuss"       "Eric Carle"
```

```
Top_authors
```

```
## # A tibble: 6 x 7
## # Groups:   Name [6]
##   Name                              Author User_Ratings Reviews Price  Year Genre
##   <chr>                             <chr>         <dbl>   <int> <int> <int> <chr>
## 1 Brown Bear, Brown Bear, What Do~ Bill ~          4.9   14344     5  2017 Fict~
## 2 Dog Man: Fetch-22: From the Cre~ Dav P~          4.9   12619     8  2019 Fict~
## 3 Harry Potter and the Chamber of~ J.K. ~          4.9   19622    30  2016 Fict~
## 4 Jesus Calling: Enjoying Peace i~ Sarah~          4.9   19576     8  2011 Non ~
## 5 Oh, the Places You'll Go!        Dr. S~          4.9   21834     8  2012 Fict~
## 6 The Very Hungry Caterpillar      Eric ~          4.9   19546     5  2013 Fict~
```

Book with highest ratings

```
Top_books <- amazon_best_selling_books%>%
  filter(User_Ratings>= 4.9) %>%
  group_by(Name) %>%
  distinct(Name, .keep_all = TRUE)

print(unique(Top_books$Author))
```

```
##  [1] "Bill Martin Jr."       "Dav Pilkey"            "Sherri Duskey Rinker"
##  [4] "Lin-Manuel Miranda"    "J.K. Rowling"          "J. K. Rowling"
##  [7] "Brandon Stanton"       "Sarah Young"           "Jill Twiss"
## [10] "Alice Schertle"        "Pete Souza"            "Dr. Seuss"
## [13] "Rush Limbaugh"         "Nathan W. Pyle"        "Patrick Thorpe"
## [16] "Chip Gaines"           "Eric Carle"            "Emily Winfield Martin"
## [19] "Mark R. Levin"         "Jeff Kinney"
```

```
Top_books
```

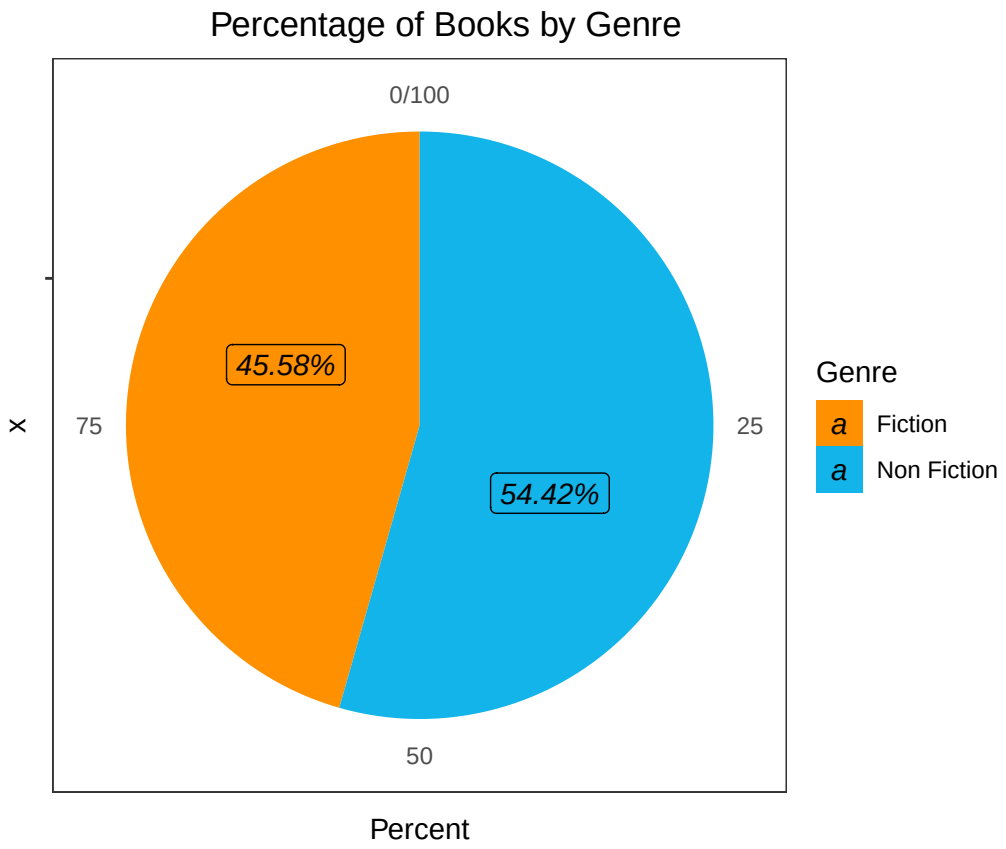
```
## # A tibble: 28 x 7
## # Groups:   Name [28]
##    Name                             Author User_Ratings Reviews Price  Year Genre
##    <chr>                            <chr>         <dbl>   <int> <int> <int> <chr>
##  1 Brown Bear, Brown Bear, What D~ Bill ~          4.9   14344     5  2017 Fict~
##  2 Dog Man and Cat Kid: From the ~ Dav P~          4.9    5062     6  2018 Fict~
##  3 Dog Man: A Tale of Two Kitties~ Dav P~          4.9    4786     8  2017 Fict~
##  4 Dog Man: Brawl of the Wild: Fr~ Dav P~          4.9    7235     4  2018 Fict~
##  5 Dog Man: Fetch-22: From the Cr~ Dav P~          4.9   12619     8  2019 Fict~
##  6 Dog Man: For Whom the Ball Rol~ Dav P~          4.9    9089     8  2019 Fict~
##  7 Dog Man: Lord of the Fleas: Fr~ Dav P~          4.9    5470     6  2018 Fict~
##  8 Goodnight, Goodnight Construct~ Sherr~          4.9    7038     7  2012 Fict~
##  9 Hamilton: The Revolution        Lin-M~          4.9    5867    54  2016 Non ~
## 10 Harry Potter and the Chamber o~ J.K. ~          4.9   19622    30  2016 Fict~
## # i 18 more rows
```

Visualization

After exploring the data now I am going to visualize the data are as following:

Percentage of books by genre:

```
options(repr.plot.width =14 , repr.plot.height = 7)
library(dplyr)
amazon_best_selling_books %>%
select(Name,Genre) %>%
distinct(Name,Genre) %>%
group_by(Genre) %>%
summarise(Count=n(),.groups = "drop")%>%
mutate(Percent=prop.table(Count) * 100) %>%
ggplot(aes(x="", y=Percent,fill = Genre))+
geom_bar(stat="identity",width = 1)+
coord_polar("y", start = 0)+
theme_test()+
scale_fill_manual(values=c('#FF9000','#12B4E9'))+
geom_label(aes(label = paste0(round(Percent,2), "%")), position = position_stack(vjust = 0.6),
            colour = "black",  fontface = "italic")+
            theme(plot.title = element_text(hjust=0.6)) +
            labs(title="Percentage of Books by Genre")
```

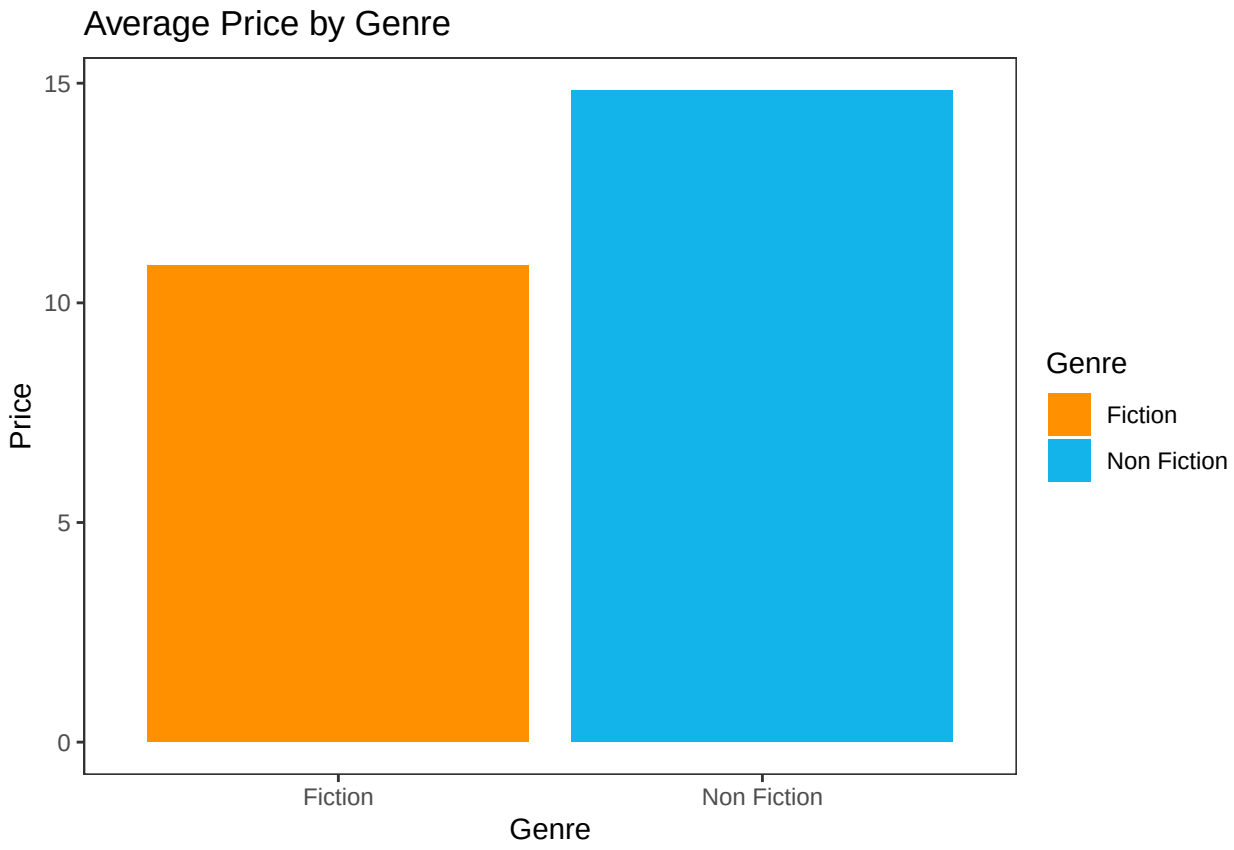


The sum of books by Genre

```
amazon_best_selling_books %>%
  group_by(Genre) %>%
  summarise(Price = mean(Price)) %>%
  ggplot(aes(x = Genre, y = Price, fill=Genre)) +
  geom_col() +
```

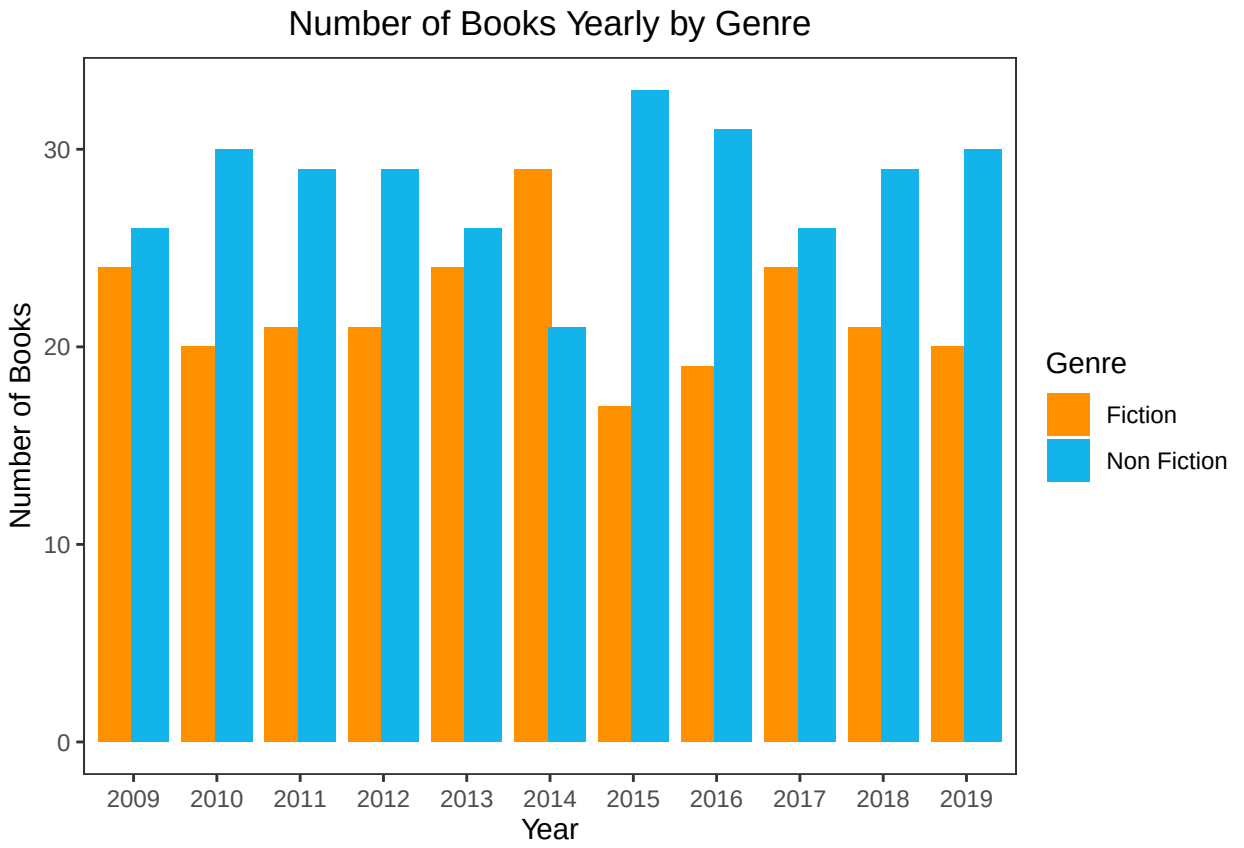
```
  theme_test()+
  scale_fill_manual(values=c('#FF9000','#12B4E9'))+

  labs(title = "Average Price by Genre")
```



total number of books by genre yearly sold

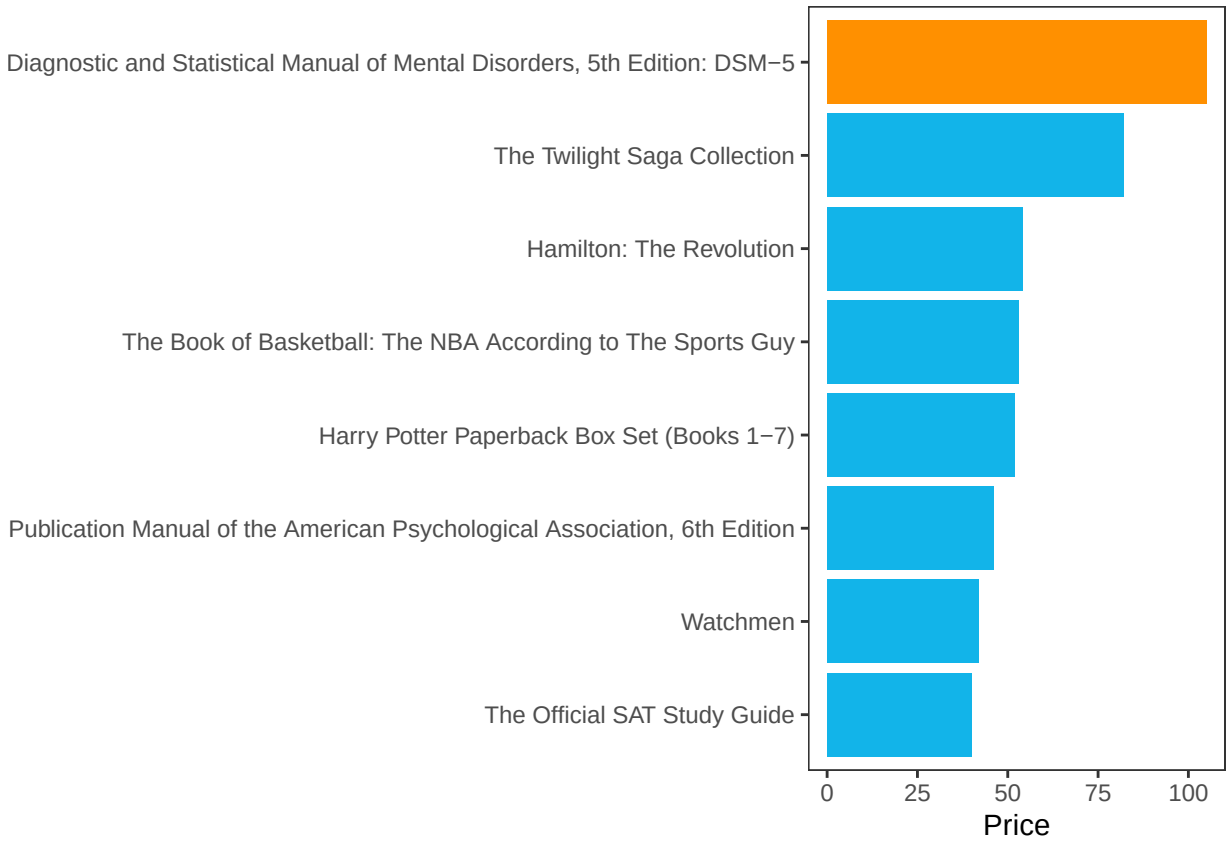
```
library(dplyr)
options(repr.plot.width = 14, repr.plot.height = 7)
amazon_best_selling_books %>%
  mutate(Year = as.character(Year))%>%
  group_by(Genre,Year) %>%
  summarise(Count=n(),.groups = "drop")%>%
  ggplot(aes(x=Year, y=Count,fill = Genre))+
  geom_bar(stat="identity", position=position_dodge(0.8))+
  theme_test()+
  scale_fill_manual(values=c('#FF9000','#12B4E9'))+
  theme(plot.title = element_text(hjust=0.5)) +
  labs(x="Year", y="Number of Books", title="Number of Books Yearly by Genre")+
  theme(legend.position="right",axis.text.x = element_text(vjust = 0.6))
```



The maximum non fiction books are sold in 2015 and minimum books sold in 2014. As compare to this the maximum fictional books sold in 2014

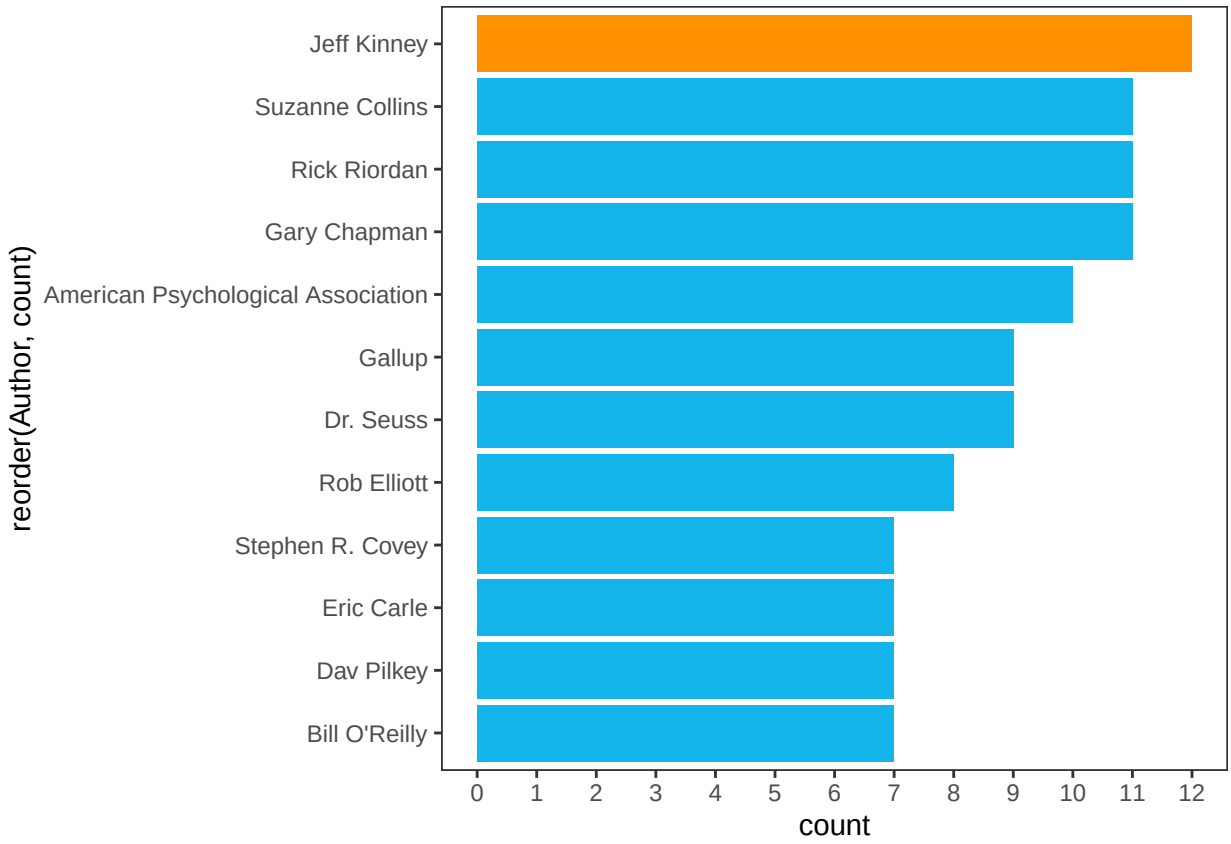
most selling books by its price

```
options(repr.plot.width =14 , repr.plot.height = 7)
amazon_best_selling_books %>%
  select(Name, Price) %>%
  arrange(desc(Price)) %>%
  head(20) %>%
  distinct() %>%
  ggplot(aes(x =reorder(Name, Price), y = Price,
             fill = ifelse(Price == max(Price), "black","grey"))) +
  geom_col() +
  coord_flip() +
  scale_fill_manual(values = c("#ff9000" , "#12b4e9")) +
  theme_test()+
  theme(legend.position = "none",axis.title.y = element_blank())
```



Top selling books by its author

```
options(repr.plot.width =14 , repr.plot.height = 7)
amazon_best_selling_books %>%
  group_by(Author) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(12) %>%
  ggplot(aes(x =reorder(Author, count), y = count,
             fill = ifelse(count == max(count), "black","grey"))) +
  scale_y_continuous(breaks = seq(0, 13, by = 1))+
  geom_col() +
  coord_flip() +
  scale_fill_manual(values = c("#ff9000", "#12b4e9")) +
  theme_test()+
  theme(legend.position = "none")
```



Findings:

The top 20 authors with higher user ratings had more than 4.5 stars in User Ratings.

Between 2009 and 2019, Fiction books had more reviews than Non-Fiction books.

There are six authors with highest ratings of 4.8 or higher and 12000 reviews or higher.

The were 28 books with a rating of 4.9 or higher written by 19 different authors.

Specifically there were a total of 345 Non Fiction books making up 54.16% of the list while there were 292 Fiction books accounting for approximately 45.58%.

Interestingly the average price of the Non Fiction books were higher than the fictional books.

Exploring authors based on their frequency on the best sellers list between 2009 and 2019 revealed that Jeff Kinney, Suzanne Collins, Rick Riordan, and emerged as performers among all genres. Each author had their works listed as sellers an impressive number of times; Rick Riordan appeared on the list a total of 18 times; Suzanne Collins came in at second place with her works appearing on it about16 times; finally Jeff Kinneys contributions made it onto this highly acclaimed list around15 times.

The price of books went down in 2014 went up in 2018 and then went down again in 2019.

Among the Top best selling authors with numerous reviews"Bill Martin Jr." "Dav Pilkey", "J.K. Rowling", "Sarah Young", "Dr. Seuss", "Eric Carle" stood out. Dr. Seuss received the number of reviews, with approximately 21834, which significantly surpassed the reviews of other authors.