Amazon Book Reviews Project

Brody Coyne

2022-11-06

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

Importing the data and showing the first few rows.

```
mydata<-read.csv("Amazon_top100_bestselling_books_2009to2021.csv")
mydata <- na.omit(mydata)
head(mydata)
```

```
##   X price ranks                                          title no_of_reviews
## 1 0 12.49     1                                The Lost Symbol        16,118
## 2 1 13.40     2   The Shack: Where Tragedy Confronts Eternity        23,392
## 3 2  9.93     3 Liberty and Tyranny: A Conservative Manifesto         5,036
## 4 3 14.30     4     Breaking Dawn (The Twilight Saga, Book 4)        16,912
## 5 4  9.99     5                 Going Rogue: An American Life         1,572
## 6 5 18.29     6                           StrengthsFinder 2.0         7,082
##   ratings           author cover_type year       genre
## 1     4.4        Dan Brown  Hardcover 2009     Fiction
## 2     4.7 William P. Young  Paperback 2009     Fiction
## 3     4.8    Mark R. Levin  Hardcover 2009 Non Fiction
## 4     4.7  Stephenie Meyer  Hardcover 2009     Fiction
## 5     4.6      Sarah Palin  Hardcover 2009 Non Fiction
## 6     4.1           Gallup  Hardcover 2009 Non Fiction
```

Showing the summary statistics for the price and ratings for the book.

```
summary(mydata$price)
```

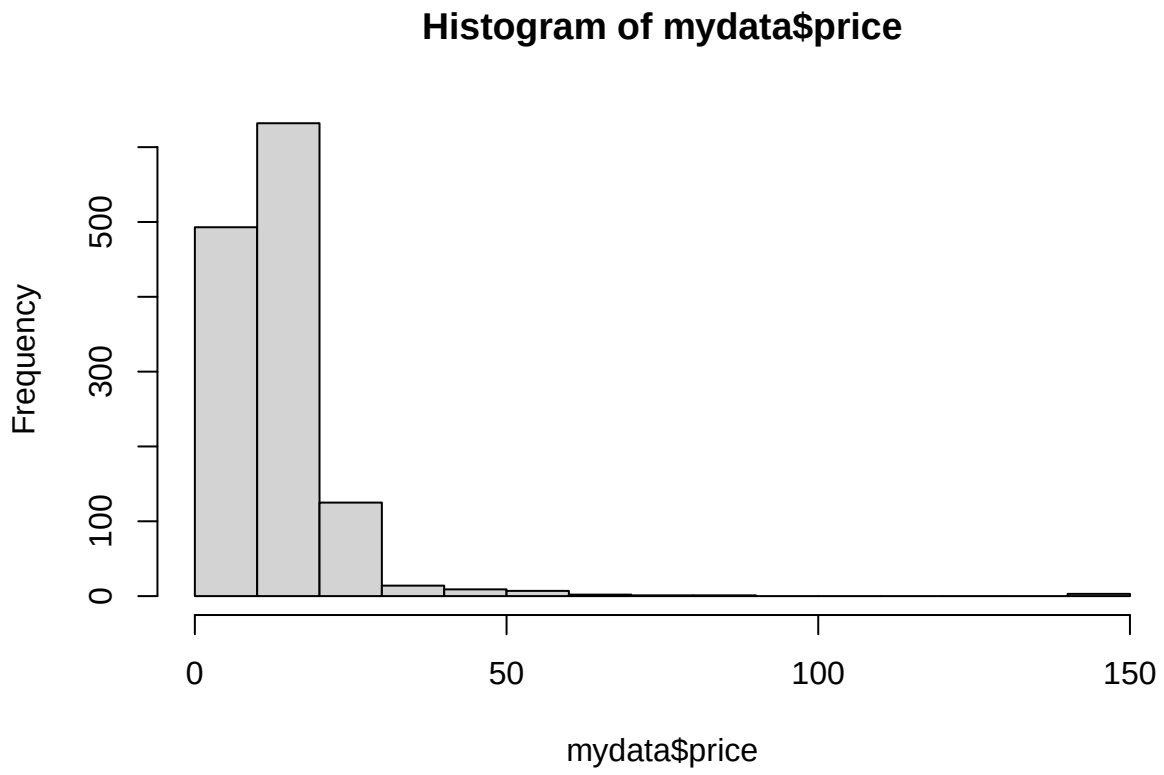
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.17    8.55   12.10   13.76   16.29  144.00
```

```
summary(mydata$ratings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.400   4.600   4.700   4.651   4.800   4.900
```

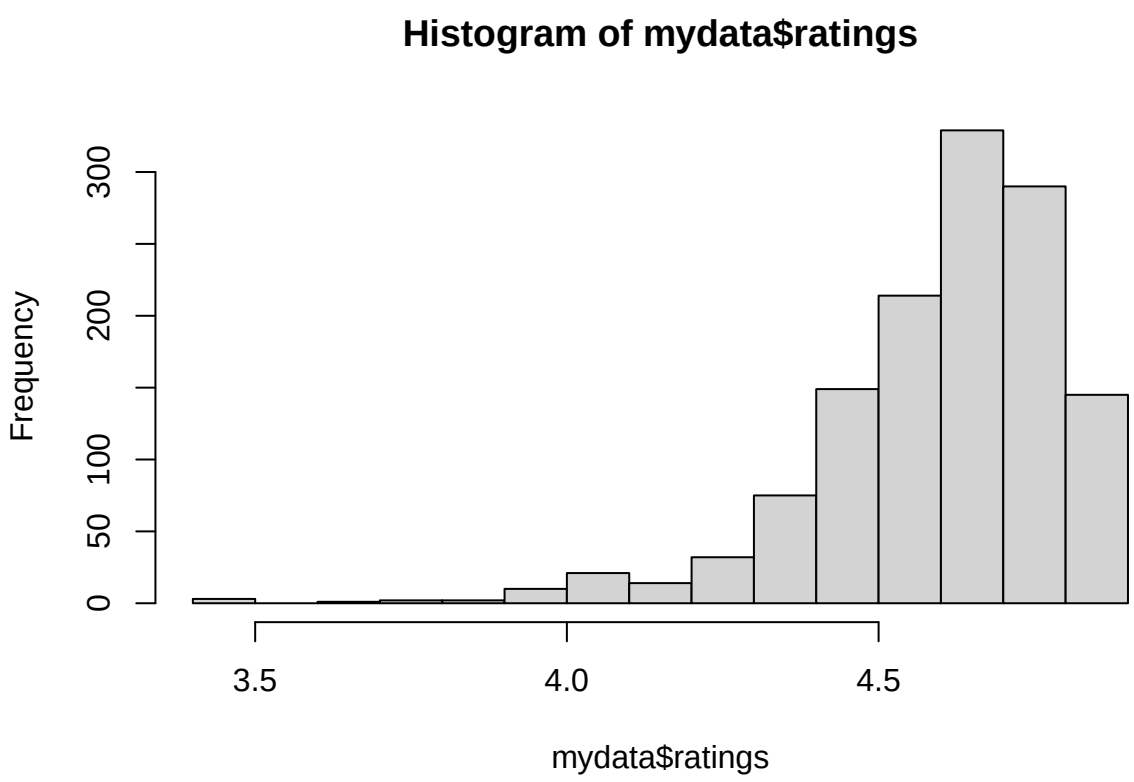
This histogram for the price of the books shows that there is a significant outlier, which is the $144 book. It also shows that there is a large portion of books which are in the $10-20 range.

```
hist(mydata$price)
```



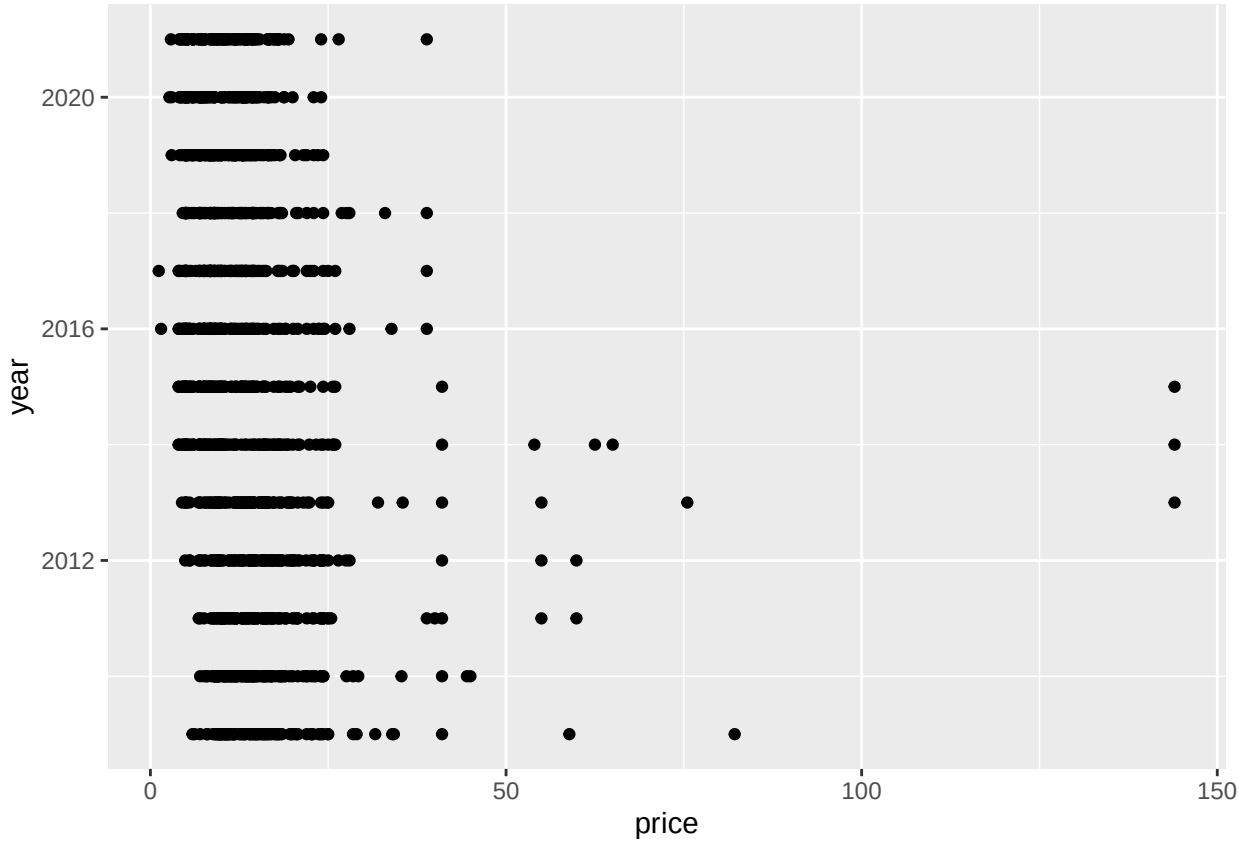
The ratings are negatively skewed. I found this interesting because there are much higher average ratings than I would have thought, which creates this significant negative skew.

```
hist(mydata$ratings)
```



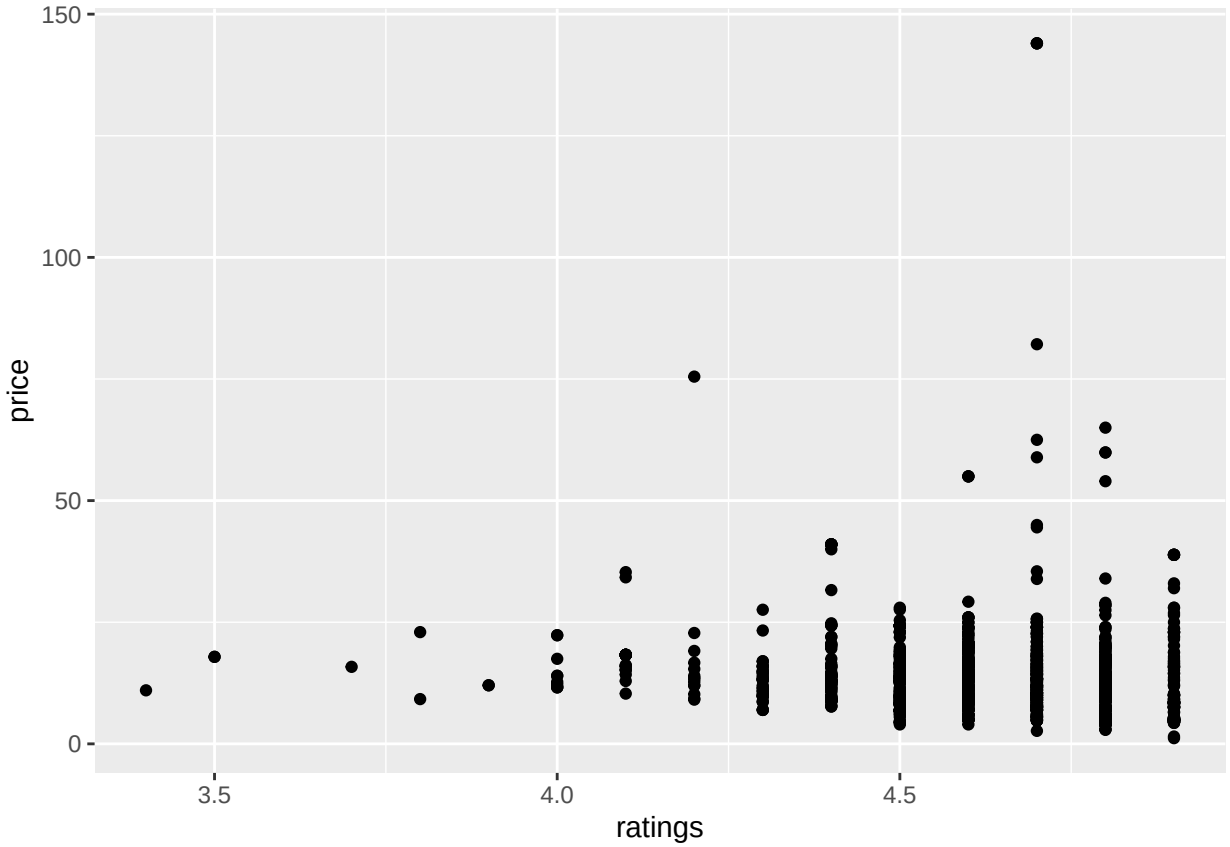
This shows the price of the books separated by each year. I found this very interesting because I would have assumed that over the years there would be a significant increase in the price of the books. However, there was no noticeable increase in price, if anything it seems like prices may have decreased a bit.

```
ggplot(data = mydata) +
  geom_point(mapping = aes(x = price, y = year))
```



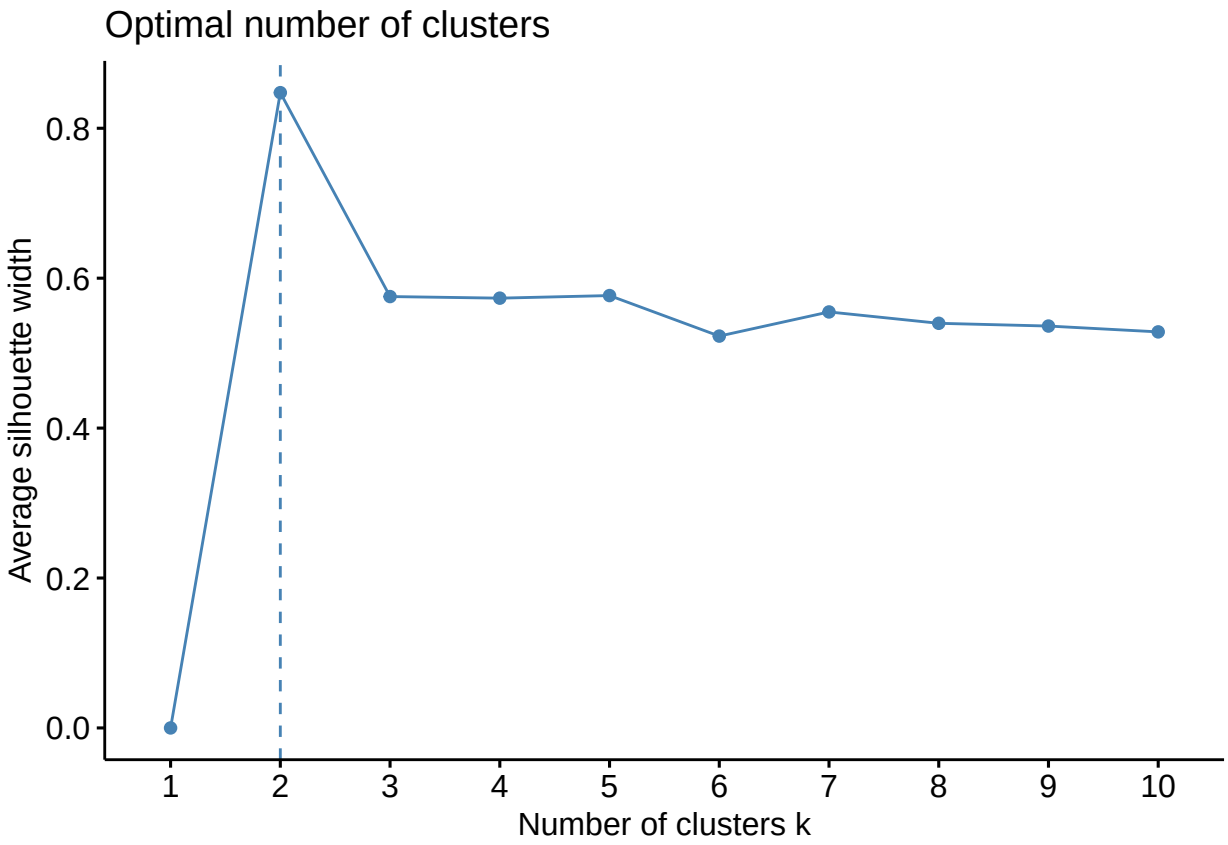
Showing the price of the books compared to the ratings on Amazon. This scatter plot doesn't give any obvious conclusions in my opinion.

```
ggplot(data = mydata) +
  geom_point(mapping = aes(x = ratings, y = price))
```



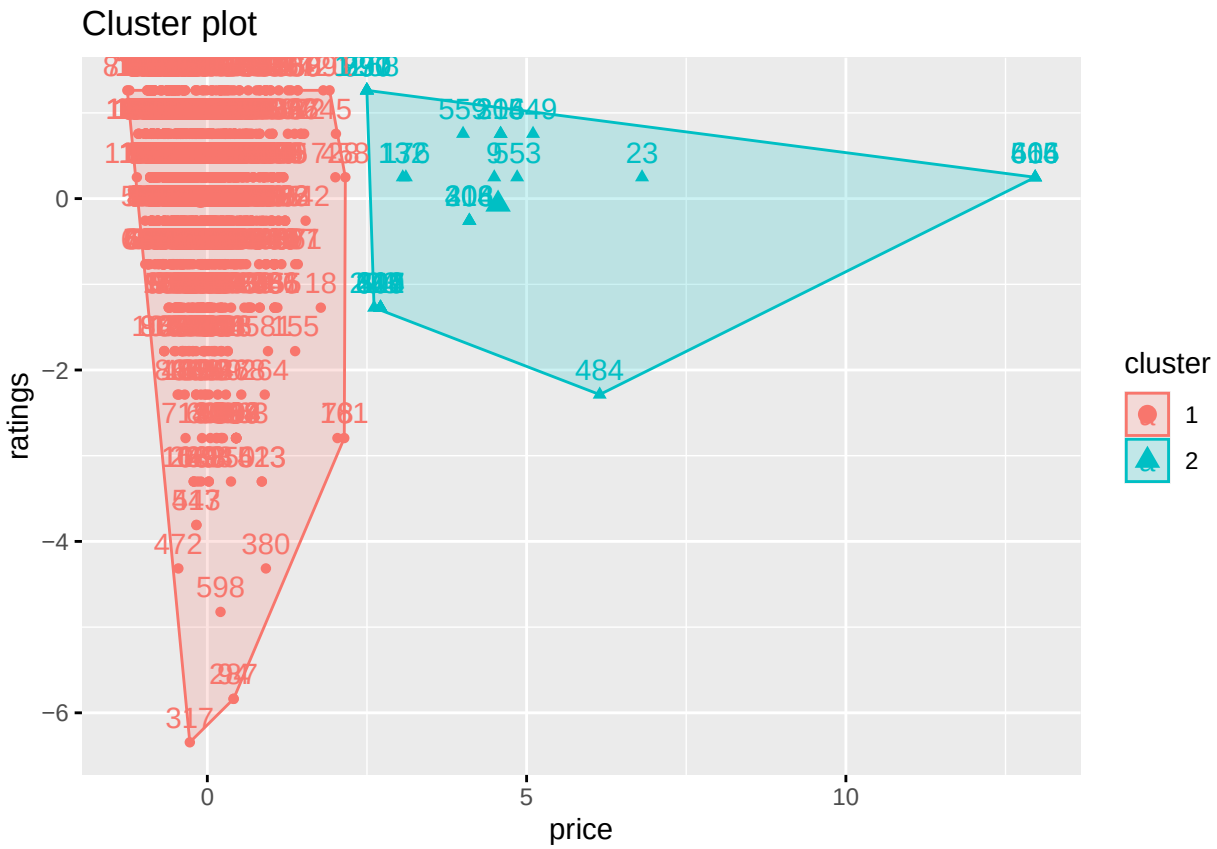
Because I did not see any obvious conclusions I wanted to perform a cluster analysis to see if it could give me a new perspective.

```
cluster<-mydata[c(2,6)]
fviz_nbclust(cluster, kmeans, method = "silhouette")
```



What I got from these two clusters is that if the book was on the lower end of the price scale, there was a large range of ratings, as seen in the first cluster. But if the price was on the higher end, the ratings were almost always on the higher ends as well. What I got from this analysis is that if an author wants a book to make the top 100 selling books on Amazon for the year it has to be one of two things, affordable or relatively highly rated.

```
k3<-kmeans(cluster, centers = 2, nstart = 25)
fviz_cluster(k3, data = cluster)
```


Cluster plot
ratings
price
cluster
1
2
0
−2
−4
−6
0
5
10
484
317
598
472
380
23