Book Sales Data Analysis

Sarbani Banerjee

24/08/2020

A company has produced multiple books, and each has received many reviews.Now the company wants to check out the sales data to extract useful information, which it can use to increase their sales further.

```
books<- read.csv("book_reviews.csv")
dim(books)
```

```
## [1] 2000    4
```

```
head(books)
```

```
##                                book    review      state price
## 1                      R Made Easy Excellent         TX 19.99
## 2                    R For Dummies      Fair         NY 15.99
## 3                      R Made Easy Excellent         NY 19.99
## 4                      R Made Easy      Poor         FL 19.99
## 5 Secrets Of R For Advanced Students     Great      Texas 50.00
## 6                      R Made Easy      <NA> California 19.99
```

```
coltype<-lapply(books,class)
columns<-colnames(books)
coltype
```

```
## $book
## [1] "character"
##
## $review
## [1] "character"
##
## $state
## [1] "character"
##
## $price
## [1] "numeric"
```

The data has 2000 observation with four variables namely book, review,state,price. Before going to any further analysis we need to check whether there is any missing value in the data.

```
missing_values<- vector()
for (i in columns){
  missing_values[i]=sum(is.na(books[[i]]))
```

```
}
missing_values
```

```
##   book review  state  price
##      0    206      0      0
```

As it can be seen under review column 206 observations were missing. Though a scrutiny of the dataset told us there is more number of reviews missing for Texas, but any particular pattern in the missing values can't be found so as of now I am going to eleminate the data with missing values.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
books_nm<- filter(books, !is.na(review))
dim(books_nm)
```

```
## [1] 1794    4
```

```
unique(books_nm[["state"]])
```

```
## [1] "TX"         "NY"         "FL"         "Texas"      "Florida"
## [6] "CA"         "California" "New York"
```

As it shows the naming pattern of the states are different.We need to convert the naming pattern into one.

```
books_nm<- books_nm %>%
  mutate(state=case_when(
    state=="California" ~ "CA",
    state=="Florida" ~ "FL",
    state=="New York" ~ "NY",
    state=="Texas" ~ "TX",
    TRUE ~ state
  )
)
unique(books_nm[["book"]])
```

```
## [1] "R Made Easy"                          "R For Dummies"
## [3] "Secrets Of R For Advanced Students" "Top 10 Mistakes R Beginners
Make"
## [5] "Fundamentals of R For Beginners"
```

In this project my job is to identify, **The most profitable book of this company?** To do this I have taken two criterias,

- *First is to find out the book which has earned most during the goven time period.*
- *Second is to find out the most popular book among these five books.*

The combining result of these two criteria will give me the most profitable book of the company.

```
#total_earning from each book
money<- books_nm %>%
  group_by(book,price) %>%
  filter(row_number() == 1)
books_count<- data.frame(table(books_nm["book"]))
books_count<- books_count %>% rename(book=Var1)
earning_matrix<- money %>%
  inner_join(books_count, by="book")
earning_matrix<- earning_matrix %>%
  select(book,price,Freq) %>%
  mutate(total_earning=price*Freq)
rank_earn<-rank(earning_matrix["total_earning"])
earning_matrix<-cbind(earning_matrix,rank_earn)

## New names:
## * NA -> ...5

earning_matrix
```

```
## # A tibble: 5 x 5
## # Groups:   book, price [5]
##   book                                price  Freq total_earning  ...5
##   <chr>                               <dbl> <int>         <dbl> <dbl>
## 1 R Made Easy                          20.0   352         7036.     2
## 2 R For Dummies                        16.0   361         5772.     1
## 3 Secrets Of R For Advanced Students   50     360        18000      5
## 4 Top 10 Mistakes R Beginners Make     30.0   355        10646.     3
## 5 Fundamentals of R For Beginners      40.0   366        14636.     4
```

As it can be seen “Secrets of R For Advanced Students” has earned the highest followed by “Fundamentals of R For Beginners”.Though “Fundamentals of R For Beginners” has the highest frequent purchase but other books are also not very far behind. If we have the cost of producing each book , then the total profit can be easily calculated, but since it is absent I have set another criteria,which is to find out the most popular book by calculating the average rating of each book.

```
# Popularity of each book
books_nm<- books_nm %>%
  mutate(review_nm= case_when(
    review=="Poor" ~1,
    review=="Fair" ~2,
```

```
    review=="Good" ~3,
    review=="Great" ~4,
    review=="Excellent" ~5

  )
  )
rating_matrix<- books_nm %>%
  group_by(book) %>%
  summarise(avg_rating=mean(review_nm))

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

rank_popularity<- rank(rating_matrix["avg_rating"])
rating_matrix<- cbind(rating_matrix,rank_popularity)
rating_matrix
```

```
##                                book avg_rating rank_popularity
## 1    Fundamentals of R For Beginners   3.010929               4
## 2                      R For Dummies   2.828255               1
## 3                        R Made Easy   2.965909               3
## 4 Secrets Of R For Advanced Students   2.963889               2
## 5   Top 10 Mistakes R Beginners Make   3.047887               5
```

As the rating matrix suggest “Top 10 Mistakes R Beginners Make” has the highest average rating ,followed by “Fundamentals of R For Beginners”. In contrast to the earning_matrix " Secrets of R For Advanced Students" has scored quite low in terms of popularity. We so far rank both the popularity and the earning index. The combination of both the rank index will give us the most profitable book of the company,

```
profit_matrix<- rating_matrix %>%
  inner_join(earning_matrix,by="book")

profit_matrix<- profit_matrix %>%
  mutate(combine_rank=...5+rank_popularity)

profit_matrix
```

```
##                                book avg_rating rank_popularity price Freq
## 1    Fundamentals of R For Beginners   3.010929               4 39.99  366
## 2                      R For Dummies   2.828255               1 15.99  361
## 3                        R Made Easy   2.965909               3 19.99  352
## 4 Secrets Of R For Advanced Students   2.963889               2 50.00  360
## 5   Top 10 Mistakes R Beginners Make   3.047887               5 29.99  355
##   total_earning ...5 combine_rank
## 1      14636.34    4            8
## 2       5772.39    1            2
## 3       7036.48    2            5
## 4      18000.00    5            7
## 5      10646.45    3            8
```

The profit_matrix shows us that “Fundamentals of R For Beginners” & “Top 10 Mistakes R Beginners Make” these two book have shown consistency in terms of both popularity and earning. The combination rank of both the books have scored 8 which is the highest among the books. Hence, we conclude these two books are most profitable books for the company.

State specific book preference

```
state_name<- unique(books_nm[["state"]])
books_state<- function(st) {
books_nm %>%
  filter(state==st) %>%
  group_by(book) %>%
  summarise(bookcount=n()) %>%
  filter(bookcount==max(bookcount)) %>%
  mutate(state=st)
}
state_matrix<- data.frame()
for (i in state_name){
  state_book<- books_state(i)
  state_book_df<-data.frame(state_book)
  state_matrix<- rbind(state_matrix,state_book_df)

}
```

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

```
state_matrix
```

```
##                                book bookcount state
## 1    Fundamentals of R For Beginners        96    TX
## 2 Secrets Of R For Advanced Students       108    NY
## 3 Secrets Of R For Advanced Students        86    FL
## 4                      R For Dummies       120    CA
```

As the state matrix suggest different state has different preference for books. While in New York and Florida “Secrets Of R For Advanced Students” is the most popular books in Texas it is “Fundamentals of R For Beginners” and on California it is “R For Dummies” .Based on these knowledge company can try to send more of these books to where they are more popular.