Analysing Profitability of Books

Chaitra Rao

6/16/2021

Introduction

We have with us a dataset of books for learning R. There is data regarding the reviews and sales of these books which wa want to use to understand which book is the popular and profiatble. We can also derive insight as to which state should be supplied with which book and how many.

Data Exploration

We begin by reviewing the number of rows and columns in the dataset, and familiarising with the names of the columns and the types of data in each column.

```
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
```

```
#Load dataset
book_rev <- read.csv("book_reviews.csv")

#Number of rows and columns in dataset
dim(book_rev)
```

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

```
#Names of the columns in dataset
colnames(book_rev)
```

```
## [1] "book"   "review" "state"  "price"
```

```r
#Type of data in each column
for (i in colnames(book_rev)){
  print(typeof(book_rev[[i]]))
}
```

```
## [1] "character"
## [1] "character"
## [1] "character"
## [1] "double"
```

```r
#Checking unique values in each column of dataset
unique_list <- list()
for (i in 1:length(colnames(book_rev))){
  unique_list[i] <- (unique(book_rev[i]))
}
print(unique_list)
```

```
## [[1]]
## [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"
##
## [[2]]
## [1] "Excellent" "Fair"      "Poor"      "Great"     NA          "Good"
##
## [[3]]
## [1] "TX"         "NY"         "FL"         "Texas"      "California"
## [6] "Florida"    "CA"         "New York"
##
## [[4]]
## [1] 19.99 15.99 50.00 29.99 39.99
```

Data Cleaning

We will review the data present in the dataset by looking at the unique values in each column. At this stage we can observe and handle any inconsistencies in the data. NA values are observed in the review column but we would not like to remove these rows as they do not introduce bias in the analysis. Since the analysis focuses on the sales and popularity in terms of number of books bought, missing reviews do not hinder the process so, we will allow the rows to remain in the dataset.

```r
#Checking unique values in each column of dataset
unique_list <- list()
for (i in 1:length(colnames(book_rev))){
  unique_list[i] <- (unique(book_rev[i]))
}
print(unique_list)
```

```
## [[1]]
## [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"
```

```
##
## [[2]]
## [1] "Excellent" "Fair"      "Poor"      "Great"     NA          "Good"
##
## [[3]]
## [1] "TX"         "NY"         "FL"         "Texas"      "California"
## [6] "Florida"    "CA"         "New York"
##
## [[4]]
## [1] 19.99 15.99 50.00 29.99 39.99
```

```
#Check to see how many rows in each column have missing data(NAs)
for (i in colnames(book_rev)){
  print(sum(is.na(book_rev[[i]])))
}
```

```
## [1] 0
## [1] 206
## [1] 0
## [1] 0
```

Next, we will bring consistency to the state column where it is observed that the same state is denoted in different ways. Example "CA" and "California" both denote the same state. We will also convert the book reviews into numerical representations for better use if and when needed.

```
#Bringing consistency to state column
book_revised <- book_rev %>% mutate(
                        state_full = case_when(
                          state == "CA" ~ "California",
                          state == "TX" ~ "Texas",
                          state == "FL" ~ "Florida",
                          state == "NY" ~ "New York",
                          TRUE ~ state))

#Converting reviews to numbers
book_revised <- book_revised %>% mutate(review_num = case_when(
                          review == "Poor" ~ 1,
                          review == "Fair" ~ 2,
                          review == "Good" ~ 3,
                          review == "Great" ~ 4,
                          review == "Excellent" ~ 5),

                          is_high_review = case_when(
                            review_num >= 4 ~ "TRUE",
                            TRUE ~ "FALSE"))

#Review revised dataset
head(book_revised)
```

```
##                               book    review      state price state_full
## 1                   R Made Easy Excellent         TX 19.99      Texas
## 2                  R For Dummies      Fair         NY 15.99   New York
## 3                   R Made Easy Excellent         NY 19.99   New York
```

```
## 4                         R Made Easy      Poor         FL 19.99    Florida
## 5 Secrets Of R For Advanced Students     Great      Texas 50.00      Texas
## 6                         R Made Easy      <NA> California 19.99 California
##   review_num is_high_review
## 1          5           TRUE
## 2          2          FALSE
## 3          5           TRUE
## 4          1          FALSE
## 5          4           TRUE
## 6         NA          FALSE
```

Data Analysis

The cleaned data is now grouped by book name and summarized to obtain the total price for each. This would denote which book generated the most money. We will also look at how many of each type of book was sold and the distribution of sales across the 4 states.

```
#Grouping by book name, check the total amount obtained from sale each of the books
books_price_summary <- book_revised %>%
  group_by(book) %>%
  summarize(total_price = sum(price))
books_price_summary
```

```
## # A tibble: 5 x 2
##   book                                total_price
##   <chr>                                     <dbl>
## 1 Fundamentals of R For Beginners         16396.
## 2 R For Dummies                            6556.
## 3 R Made Easy                              7776.
## 4 Secrets Of R For Advanced Students      20300
## 5 Top 10 Mistakes R Beginners Make        11546.
```

```
#Check which states buy the most of each of the books
table(book_revised$book)
```

```
##
##    Fundamentals of R For Beginners                         R For Dummies
##                                410                                   410
##                       R Made Easy Secrets Of R For Advanced Students
##                                389                                   406
##   Top 10 Mistakes R Beginners Make
##                                385
```

```
table(book_revised$book, book_revised$state_full)
```

```
##
##                                   California Florida New York Texas
##   Fundamentals of R For Beginners        107      84      106   113
##   R For Dummies                          132      90       91    97
##   R Made Easy                             97      90      105    97
##   Secrets Of R For Advanced Students      99      93      119    95
##   Top 10 Mistakes R Beginners Make        83      92      110   100
```

Conclusion

As we can see, the book Secrets of R for Advanced Students has generated the most money is part of the top 3 most sold books. It is bought in many numbers across all 4 states but highest in New York. In terms of the book sales by state, we see that the book R for Dummies is bought in California a lot more in number than in other states. This data tells us that in terms of revenue, the book Secrets of R for Advanced Students is most profitable and the book R for Dummies is least profitable even though it is bought as much in number as Secrets of R for Advanced Students.