Book Reviews

Bena

2023-04-03

As a data analyst for a company that sells books for learning programming, your company has produced multiple books, and each has received many reviews. The company wants us to check out the sales data and see if we can extract any useful information from it.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0     ✓ purrr   1.0.1
## ✓ tibble  3.1.8     ✓ dplyr   1.1.0
## ✓ tidyr   1.3.0     ✓ stringr 1.5.0
## ✓ readr   2.1.3     ✓ forcats 1.0.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Import data

```
book_reviews <- read_csv("D:/BENA/Data Analytics/Dataquest/Project2_DataCleaning/book_reviews.csv")
```

```
## Rows: 2000 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): book, review, state
## dbl (1): price
##
## 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.
```

About data

```
dim(book_reviews)
```

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

```
glimpse(book_reviews)
```

```
## Rows: 2,000
## Columns: 4
## $ book   <chr> "R Made Easy", "R For Dummies", "R Made Easy", "R Made Easy", "…
## $ review <chr> "Excellent", "Fair", "Excellent", "Poor", "Great", NA, "Great",…
## $ state  <chr> "TX", "NY", "NY", "FL", "Texas", "California", "Florida", "CA",…
## $ price  <dbl> 19.99, 15.99, 19.99, 19.99, 50.00, 19.99, 19.99, 19.99, 29.99, …
```

```
colnames(book_reviews)
```

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

```
typeof(book_reviews)
```

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

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Data type for each column

```
for (col in colnames(book_reviews)) {
 print(typeof(book_reviews[[col]]))
}
```

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

Unique values in each column

```
for (val in colnames(book_reviews)) {
  print("unique values")
  print(val)
  print(unique(book_reviews[[val]]))
}
```

```
## [1] "unique values"
## [1] "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"
## [1] "unique values"
## [1] "review"
## [1] "Excellent" "Fair"      "Poor"      "Great"     NA          "Good"
## [1] "unique values"
## [1] "state"
## [1] "TX"         "NY"         "FL"         "Texas"      "California"
## [6] "Florida"    "CA"         "New York"
## [1] "unique values"
## [1] "price"
## [1] 19.99 15.99 50.00 29.99 39.99
```

Handling missing data

```
book_reviews_nonulls <- book_reviews %>%
  filter(!is.na(review))

dim(book_reviews_nonulls)
```

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

```
View(book_reviews_nonulls)
```

Investigate missing data

```
book_reviews_nulls <- book_reviews %>%
  filter(is.na(review))

View(book_reviews_nulls)

nulls_per_book <- book_reviews_nulls %>%
  group_by(book) %>%
  summarise(
    nulls = n()
  ) %>%
  arrange(-nulls)

nulls_per_state <- book_reviews_nulls %>%
  group_by(state)%>%
  summarise(
    nulls = n()
  ) %>%
  arrange(-nulls)
```

Some 206 rows with nulls in the reviews column were deleted. We still have 89.7% of our data available.

Nulls per book are: 1 Fundamentals of R For Beginners 44 2 R For Dummies 49 3 R Made Easy 37 4 Secrets Of R For Advanced Students 46 5 Top 10 Mistakes R Beginners Make 30

Nulls per state are: 1 TX 62 2 CA 54 3 NY 47 4 FL 43

Standardize values in state column

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

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

Have a look at the results

```
View(book_reviews_nonulls)
```

Transform text data to number type data

```
book_reviews_nonulls <- book_reviews_nonulls %>%
  mutate(
    review_num = case_when(
      review == "Poor" ~ 1,
      review == "Fair" ~ 2,
      review == "Good" ~ 3,
      review == "Great" ~ 4,
      review == "Excellent" ~ 5
    )
  )

book_reviews_nonulls <- book_reviews_nonulls %>%
  mutate(
    is_high_review = if_else(review_num >= 4,TRUE,FALSE)
  )
```

Finding the most profitable book

```
profitable_book <- book_reviews_nonulls %>%
  group_by(book) %>%
  summarize(
    total_price = sum(price)
  ) %>%
  arrange(-total_price)
```

Secrets of R For Advanced Students had the highest total price. Alternatively, considering the number of books sold.

```
profitable_book <- book_reviews_nonulls %>%
  group_by(book) %>%
  summarize(
    no_purchased = n()
  ) %>%
  arrange(-no_purchased)
```

Fundamentals of R For Beginners had the highest number of purchases

Reporting the results

This analysis is motivated by the fact that the company wants us to explore the book sales data and gain valuable insights from it. Therein, we have quantitative data i.e. price and qualitative data such as reviews. More information is also provided i.e. the book name and state where the sale was made. The main question we're trying to answer is how profitable are the book sales.

In the data preparation, some of the things that had to be done to make it more usable include: Converting data in the review column from string to numeric i.e. 1 to 5, 5 being excellent Filtering missing data and reviewing it closely Standardizing state names to respective abbreviated code names eg from California to CA

Aggregating the data using group by and count functions among others. Sorting out the data to find out which had the highest/lowest number of factor of interest eg highest no. of books sold, ordering revenue earned from each book, book with the most no. of favorable reviews. Used control flow and logicals to categorize the data.

In conclusion, the most profitable book in terms of revenue earned was Secrets of R For Advanced Students & the book with highest number sold was Fundamentals of R For Beginners. More information could also be provided that would aid in knowing what time of the year most sales are made. It could also be helpful in determining whether or not to do a sales & marketing campaign and what time would be the best to do that. These findings could be helpful to the stores & procurement or publishing section of the company. They're now more aware of which books they should stock up more of.