Book Review Analysis

2022-10-16

Introduction

This is a guided project to demonstrate a data analysis work flow. We are going to be analyzing sales data of a company that sells books for learning programming. The key steps we are going to take in this project includes:

1. Exploring the data - before beginning any analysis it is important to get yourself familiar with the data you are working with. This step consists of the following:

- Checking the dimension of the data i.e the number of rows and columns in the data.
- Checking for null values and duplicate values.
- Checking the data type of each column, etc.

2. Cleaning the data - The bulk of a data analysis work flow has to do with cleaning the data. This might mean figuring out what to do with null values either by dropping them or by imputing the values using a measure of central tendency or imputing based on research/ domain knowledge. This might also include standardizing of the data in specific columns, creating new columns more suitable for the analysis.

3. Analyzing the data - After we have gotten a satisfying result, one than can answer the questions that we want to answer with our data, we can then go ahead to analyse and generate insight from the data.

In this project, we are going to be following all of these steps to answer certain key questions. Our goal is to find out the following:

- Which book the most profitable book overall?
- How does each book perform in each states?

```
# loading the libraries
library(tidyverse)
library(kableExtra)
```

Data Exploration

```
book_reviews <- read_csv("book_reviews.csv", col_types = cols())

book_reviews %>% dim() # shows dimension of the data set
```

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

```
book_reviews %>% glimpse() # gives a general overview of the data set
```

```
## 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, ~
```

The data set we are working with has 2000 rows and 4 columns. The `glimpse()` function also gives us the the data types of each column. The `book`, `review`, and `state` columns all have `character` data type or string data type while the `price` column has `double` data type or decimal/numeric data type. Now that we are familiar with the size and data type of our data set. Let's look at the first 6 rows using the `head()` function.

```
(book_reviews %>% head() %>% kbl(booktabs = T) %>%
   kable_styling(latex_options = c("striped", "HOLD_position"))
)
```

book	review	state	price
R Made Easy	Excellent	TX	19.99
R For Dummies	Fair	NY	15.99
R Made Easy	Excellent	NY	19.99
R Made Easy	Poor	FL	19.99
Secrets Of R For Advanced Students	Great	Texas	50.00
R Made Easy	NA	California	19.99

Data Cleaning

Now that we are familiar with our data, we can go ahead and clean the data following the aforementioned steps.

```
# storing the column names as a variable
column_names <- book_reviews %>% colnames()

# checking for null values
for (col in column_names) {
  null_val <- book_reviews %>% pull(col) %>% is.na() %>% sum()
  paste("Number of null values in", col, "column :", null_val) %>% print()
}
```

```
## [1] "Number of null values in book column : 0"
## [1] "Number of null values in review column : 206"
## [1] "Number of null values in state column : 0"
## [1] "Number of null values in price column : 0"
```

We have null values in only the the review column, 206 null values. For this analysis we are going to drop the rows with null values.

```
# dropping rows with NA values
clean_book_reviews <- book_reviews %>% filter(!(is.na(review)))
clean_book_reviews %>% dim()
```

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

After getting read of the NA values, we are left with 1794 rows. If we look back at the table, we showed earlier, you'll notice that some of the values in the state column are full state names e.g "California" or state codes e.g "Texas". It's a good thing to standardize the values in this column. Let's explore the state column further.

```
clean_book_reviews %>% pull(state) %>% unique()
```

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

There are 8 unique values in the state column. The state names of the four states in the data set and the state codes of these states. We are going to split this column into two new columns, one containing only the

state names and the other containing only the state codes and finally we are going to drop the state column.

```
# creating state_name and state_code columns
clean_book_reviews <- clean_book_reviews %>% mutate(
  state_name = case_when(
    state == "Texas" ~ "Texas",
    state == "TX" ~ "Texas",
    state == "New York" ~ "New York",
    state == "NY" ~ "New York",
    state == "Florida" ~ "Florida",
    state == "FL" ~ "Florida",
    state == "California" ~ "California",
    state == "CA" ~ "California"
  ),
  state_code = case_when(
    state == "Texas" ~ "TX",
    state == "TX" ~ "TX",
    state == "New York" ~ "NY",
    state == "NY" ~ "NY",
    state == "Florida" ~ "FL",
    state == "FL" ~ "FL",
    state == "California" ~ "CA",
    state == "CA" ~ "CA"
  )
)

clean_book_reviews <- clean_book_reviews %>% select(-state)
```

```
(clean_book_reviews %>% head() %>% kbl(booktabs = T) %>%
   kable_styling(latex_options = c("HOLD_position", "stripe"))
)
```

book	review	price	state_name	state_code
R Made Easy	Excellent	19.99	Texas	TX
R For Dummies	Fair	15.99	New York	NY
R Made Easy	Excellent	19.99	New York	NY
R Made Easy	Poor	19.99	Florida	FL
Secrets Of R For Advanced Students	Great	50.00	Texas	TX
R Made Easy	Great	19.99	Florida	FL

Now that we have fixed the inconsistent values that were in the `state` column, the next thing to do is to convert the text in the `review` column to number to make it much easier to work with. To do this, we are going to create a new column `review_num` where we will be replacing "Poor" with 1, "Fair" with 2, "Good" with 3, "Great" with 4, and "Excellent" with 5. We are also going to be creating a column `is_high_rating` where ratings 4 or greater than 4 will be TRUE and the rest FALSE.

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

```
  )
)

clean_book_reviews <- clean_book_reviews %>% mutate(
  is_high_review = if_else(review_num > 3, TRUE, FALSE)
)
```

```
(clean_book_reviews %>% head() %>% kbl(booktabs = T) %>%
   kable_styling(latex_options = c("HOLD_position", "stripe"))
)
```

book	review	price	state_name	state_code	review_num	is_high_review
R Made Easy	Excellent	19.99	Texas	TX	5	TRUE
R For Dummies	Fair	15.99	New York	NY	2	FALSE
R Made Easy	Excellent	19.99	New York	NY	5	TRUE
R Made Easy	Poor	19.99	Florida	FL	1	FALSE
Secrets Of R For Advanced Students	Great	50.00	Texas	TX	4	TRUE
R Made Easy	Great	19.99	Florida	FL	4	TRUE

Which Is The Most Profitable Book?

So far we have succeeded in cleaning the data into something that is usable. Now we can go ahead and analyze this data. The question we want to answer is which book is the most profitable. We are going to be looking at this question from two angles. First we will look at profitability with respect to revenue generated and then we will look at profitability with respect to sales. We are also going to be looking at the average review of each book as well as the percentage of high rating for each book.

```
# grouping and aggregating the data
revenue_summary <- clean_book_reviews %>% group_by(book) %>% summarise(
  sales = n(),
  sales_perc = round(sales / nrow(clean_book_reviews), 3),
  revenue = sum(price),
  average_review = round(mean(review_num), 2),
  price = revenue / sales,
  high_review_perc = round(
    (sum(is_high_review) / sales), 2)
) %>% arrange(-revenue)

revenue_summary %>% kbl(booktabs = T)
```

book	sales	sales_perc	revenue	average_review	price	high_review_perc
Secrets Of R For Advanced Students	360	0.201	18000.00	2.96	50.00	0.38
Fundamentals of R For Beginners	366	0.204	14636.34	3.01	39.99	0.41
Top 10 Mistakes R Beginners Make	355	0.198	10646.45	3.05	29.99	0.40
R Made Easy	352	0.196	7036.48	2.97	19.99	0.39
R For Dummies	361	0.201	5772.39	2.83	15.99	0.35

When we take into account what percentage of the total sales each book makes up, they all perform roughly the same, accounting for about 20% of the total sales each. When we look at the revenues generated by each book, the book with the highest revenue (Secrets Of R For Advanced Students) was the most expensive book. Since this book sells as well as the other books and the pricing doesn't deter customers, it is our most valuable and most profitable book.

How Well Does Each Book Peform Across The Various States?

Now that we know how well each book performs in general. We want to see how they fair in each of the individual states.

```
state_revenue_summary <- clean_book_reviews %>% group_by(book, state_name) %>% summarise(
  sales = n(),
  sales_perc = round(sales / nrow(clean_book_reviews), 3),
  revenue = sum(price),
  average_review = round(mean(review_num), 2),
  price = revenue / sales,
  high_review_perc = round(
    (sum(is_high_review) / sales), 2),
  .groups = "drop"
) %>% arrange(-revenue, -sales)

state_revenue_summary %>% kbl(booktabs = T)
```

book	state_name	sales	sales_perc	revenue	average_review	price	high_revie
Secrets Of R For Advanced Students	New York	108	0.060	5400.00	2.95	50.00	
Secrets Of R For Advanced Students	Florida	86	0.048	4300.00	2.94	50.00	
Secrets Of R For Advanced Students	California	84	0.047	4200.00	3.17	50.00	
Secrets Of R For Advanced Students	Texas	82	0.046	4100.00	2.79	50.00	
Fundamentals of R For Beginners	California	99	0.055	3959.01	3.02	39.99	
Fundamentals of R For Beginners	New York	97	0.054	3879.03	2.97	39.99	
Fundamentals of R For Beginners	Texas	96	0.054	3839.04	3.08	39.99	
Top 10 Mistakes R Beginners Make	New York	102	0.057	3058.98	3.06	29.99	
Fundamentals of R For Beginners	Florida	74	0.041	2959.26	2.96	39.99	
Top 10 Mistakes R Beginners Make	Texas	92	0.051	2759.08	2.85	29.99	
Top 10 Mistakes R Beginners Make	Florida	84	0.047	2519.16	3.07	29.99	
Top 10 Mistakes R Beginners Make	California	77	0.043	2309.23	3.25	29.99	
R Made Easy	New York	96	0.054	1919.04	2.91	19.99	
R For Dummies	California	120	0.067	1918.80	2.72	15.99	
R Made Easy	Texas	87	0.048	1739.13	2.94	19.99	
R Made Easy	Florida	85	0.047	1699.15	3.07	19.99	
R Made Easy	California	84	0.047	1679.16	2.95	19.99	
R For Dummies	Texas	83	0.046	1327.17	2.72	15.99	
R For Dummies	New York	81	0.045	1295.19	2.72	15.99	
R For Dummies	Florida	77	0.043	1231.23	3.22	15.99	

When we look at the revenue for each of the individual states, the book `Secrets Of R For Advanced Students` generated the most revenue in each state and accounted for at least 5% of the sales in every state. It is very well our most valuable book.

Conclusion

We set out to demonstrate a complete and comprehensive data analysis work flow by analyzing book review data. We had a well set up approach on how to achieve our goal. We had to explore and get familiar with the data, clean the data, before finally analyzing it. Our goal was to find out the most profitable book. Here are some valuable insights from the data:

- Secrets Of R For Advanced Students is the most valuable book, it generated the most revenue and the sales matched those of the other books.

- The reason for the higher revenue is because it was priced higher. In fact, since all of the books had the same sales performance, the key factor for driving the revenue generated by each book was the pricing.
- Even when we look at the book performances in each individual state, it is nearly identical to the overall performance.

I refrained from using the word profit a lot because while we can tell how much revenue each of these books generated, we can't actually tell the actual profit made since we don't have data on how much it costs to produce/procure these books.