Guided Project: Creating An Efficient Data Analysis Workflow

Mariby Cruz

2025-03-18

Project Background

This guided project hosted by DataQuest puts to use the concepts learned about control flow, for loops, and functions. The background is that datasets are rarely received clean so it's important to be prepared to clean them for analysis.

I am acting as a hired data analyst for a company that sells books for learning programming. The company has produced multiple books, and each has received many reviews but wants you to check out the sales data and see if any useful information can be extracted from it.

Getting Familiar with the Data

The data set has 2000 book reviews and containing information on the book title, the review rating (e.g. Poor, Excellent), the state the review came from, and the cost of the book.
Checking the data types for each column is important because data can be represented in a different type than it's meant to when it's imported. In this case, the columns title, rating, and state are `character`; the price column is a `double` data type.

Next, it's important to get a sense of all the possible values in data strings; in numeric columns how high or low numeric values go. There are a total of five book titles that were reviewed: **R Made Easy**, **R For Dummies**, **Secrets Of R For Advanced Students**, **Top 10 Mistakes R Beginners Make**, **Fundamentals of R For Beginners**. The review ratings have five types: *Excellent*, *Great*, *Good*, *Fair*, and *Poor*. There is also NA which gives the first indication that there are missing values for some book reviews and they will need to be handled. Moreover, the books' cost range from as low as $15.99 to as high as $50.00. Lastly, the reviews came from 4 different states: California, Florida, New York, and Texas; however, there are 8 unique values in this column - an abbreviation of each state - this will need to be cleaned for consistency otherwise it can be misleading.

```r
book_reviews_df <- read.csv("~/git/DataQuest_guided_projects/book_reviews.csv")
glimpse(book_reviews_df)
```

```
## 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, ~
```

```r
unique(book_reviews_df$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"
```

```
unique(book_reviews_df$review)
```

```
## [1] "Excellent" "Fair"      "Poor"      "Great"     NA          "Good"
```

```
unique(book_reviews_df$state)
```

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

```
unique(book_reviews_df$price)
```

```
## [1] 19.99 15.99 50.00 29.99 39.99
```

Handling Missing Data

There are two options to handle missing data. One is to remove the rows or observations with missing values. The second option is to use imputation methods (however this project will not use this method).

Below, using a `For Loop` we counted the number of NA's in each column followed by creating a new data frame with those row removed. Using the information from the first code chunk (206 observations with NA values) we can get check to see if the new dataset has the correct number of observations left.

```
#getting an understanding of which columns have missing data
#iterate over the columns
col_missing_counts <- c()
for (col_name in names(book_reviews_df)) {
  #sum up the counts of NA for each column
  col_missing_counts[col_name] <- sum(is.na(book_reviews_df[col_name]))
}
col_missing_counts
```

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

```
#creating a new dataframe with NA's rows removed
book_reviews_complete <- book_reviews_df %>%
  filter(!is.na(review))
dim(book_reviews_complete)
```

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

Dealing With Inconsistent State Column Labels

When we first got a glimps of the unique values in each columns we noticed that there was an inconsistency with the `State`'s nomenclature. Some rows had the full name of the state and others had the postal code abbreviation (New York vs NY).

```r
#a new column will be created to stick to one nomenclature convention
book_reviews_complete <- book_reviews_complete %>%
  mutate(
    state_name = case_when(
      state %in% c("NY", "New York") ~ "New York",
      state %in% c("CA", "California") ~ "California",
      state %in% c("TX", "Texas") ~ "Texas",
      state %in% c("FL", "Florida") ~ "Florida"
    ))
```

Transforming the Review Data

Another thing that was noticed at the beginning was that the review ratings are in a string format ("Good"). To make the column useful for analysis, they will be coded into numeric format.

```r
#a new column will be created to recode the reviews into numeric reviews
book_reviews_complete <- book_reviews_complete %>%
  mutate(
    review_num = case_when(
      review == "Poor" ~ 1,
      review == "Fair" ~ 2,
      review == "Good" ~ 3,
      review == "Great" ~ 4,
      review == "Excellent" ~ 5
    ))

#a new column was created to denote whether a score was a high score or not
book_reviews_complete <- book_reviews_complete %>%
  mutate(
    is_high_review =
      if_else(review_num >= 4, TRUE, FALSE)
  )
```

Analyzing the Data

The main goal is to figure out what book is the most profitable. Using customer purchases, one way to define "most profitable" might be to just choose the book that's purchased the most. Another way to define it would be to see how much money each book generates overall.

```r
#Most profitable by the number of times a book was purchased
num_purchased <- count(book_reviews_complete, book)
colnames(num_purchased) <- c("Title", "Total Purchased")

num_purchased <- num_purchased %>%
  mutate(
    prop = signif(`Total Purchased`/ sum(`Total Purchased`), digits = 3)
  ) %>%
  arrange(-prop)

num_purchased
```

```
##                                 Title Total Purchased  prop
## 1    Fundamentals of R For Beginners             366 0.204
## 2                      R For Dummies             361 0.201
## 3 Secrets Of R For Advanced Students             360 0.201
## 4   Top 10 Mistakes R Beginners Make             355 0.198
## 5                        R Made Easy             352 0.196
```

```
#Most profitable by the amount of money each book generated
total_revenue <- book_reviews_complete %>%
  group_by(book) %>%
  summarise(
    total = sum(price)
  ) %>%
  arrange(-total)

colnames(total_revenue) <- c("Title", "Revenue")
```

The results shows the book with the most revenue generated was by **Secrets Of R For Advanced Students** with a total revenue of $18,000. The results also show that the book most purchased by customers was **Fundamentals of R For Beginners** with a total of 366 books which accounted for 20% of the total purchases.

Reporting the Results

Introduction

The most recent ratings review was released. We want to take a look at the most profitable books in the market using the information from the release.

Findings

There were a total of 2000 reviews submitted from 4 different states:

- California
- Florida
- New York
- Texas

Some of the reviews did not have ratings, so those reviews were not included in the analysis.
The following books were part of the ratings review:

- **R Made Easy**
- **R For Dummies**
- **Secrets Of R For Advanced Students**
- **Top 10 Mistakes R Beginners Make**
- **Fundamentals of R For Beginners**

We calculated the most profitable book with two metrics: total revenue generated & total number of books purchased.

Conclusion

The results show that the book with the most revenue generated was by **Secrets Of R For Advanced Students** with a total revenue of $18,000. This book was sold for $50.00 each.

The results also show that the book most purchased by customers was **Fundamentals of R For Beginners** with a total of 366 books which accounted for 20% of the total purchases. This book was sold for $39.99 each. This information can be used to know how many more books need to be printed and binded for sales. If there are books that are not rating high, then no more books need to be added to shelves across the country.

Further Steps

Further steps for this analysis include: Exploring the missing data from early data cleaning. We didn't have a look at this data, but maybe it would be good to investigate it further. Is it possible that a certain type of book had more missing reviews? Maybe from a certain state?

Is there any relationship between state and the books purchased there? Maybe some states have more interest in some books over others. With this knowledge, we can try to send more of these books to where they are more popular.

Based on the definition of high score we used, are some books more popular than others?

We can also expand our analysis by trying to gain insights on reader's sentiment and reception towards books produced by the company. This would require shifting the focus of the data analysis workflow from financial success to paying attention to reader's ratings based on sentiment and reception. Remember that the book ratings can be transformed into an integer scale using scores of 1 - 5 so that further analysis can be carried out on them, such as grouping the dataset by books, calculating the mean review score for each book and then sorting the data in descending order to ascertain the book with the highest mean rating.