Building a Data Analysis Workflow

Andrew Lujan

2022-05-05

Introduction

For this project, I will be acting as a data analyst for a company that sells books for learning programming. The company has produced a variety of books with each receiving quite a few reviews.

The question: The company wants me to check the sales data to see if I can extract any useful information from the data itself.

To start with our project, first we need to import the packages we'll be using and also import the dataset that's available Here.

```
## Import Packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
reviews <- read_csv("/Users/drewsdesktop/Desktop/Data Science/DataQuest/R Data Analyst Path/book_reviews
```

```
## 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.
```

Getting Familiar with the Data

First, in our workflow we want to answer the following questions:

1. How big is the dataset?
2. What are the column names?
3. What are the types of each columns?
4. What are the unique values present in each of the columns

```
## Checking the size of the dataset
dim(reviews)
```

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

```r
## Checking the column names of the dataset
colnames(reviews)
```

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

```r
## What are the types of each column
### Using a for loop to answer this question
for(c in colnames(reviews)) {
  print(typeof(reviews[[c]]))
}
```

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

```r
## What are the unique values present for each column?
for (c in colnames(reviews)) {
  print("Unique values in the column:")
  print(c)
  print(unique(reviews[[c]]))
  print("")
}
```

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

Takeaways

Coming back to the questions above:

1. **How big is the dataset?**

- The dataset has 2,000 rows (observations) and 4 columns

2. **What are the column names?**

- The column names are `book`, `review`, `state`, and `price`.

3. **What are the types of each columns?**

- The column types are: character, character, character, and double respectively.

4. **What are the unique values present in each of the columns**

- For `book`: "R Made Easy", " "R For Dummies", "Secrets Of R For Advanced Students" "Top 10 Mistakes R Beginners Make", "Fundamentals of R For Beginners"
- For `review`: "Excellent", "Fair", "Poor", "Great", NA, "Good"
- For `state`: "TX", "NY", "FL", "Texas", "California", "Florida", "CA", "New York"
- For `price`: 19.99 15.99 50.00 29.99 39.99

Most of the data contains strings. The book column tells the names of the books, the review columns tell the name of the name of the scores, the state has 2 letter state code in string form, and the price has a numerical value for the price of each book.

Cleaning the Data

There are few instances of missing data denoted with NA. We need to get rid of the missing data. We can use the `filter()` function and the `is.na()` function to remove some rows that have missing data.

```
## Viewing data

view(reviews)

## Creating a new dataframe with complete data
complete_reviews <- reviews %>%
  filter(!is.na(review)
  )

## Checking the dimensions of the new dataset
dim(complete_reviews)
```

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

Takeaways

It looks like a little over 200 reviews were removed from the dataset. Something else I noticed was teh inconsistent formatting within the state column. For example California how two different labels for that column. What we want to do is get the formatting into the standard postal code format across the state column.

```
## Shortening the labels in the state column to just the postal code

complete_reviews <- complete_reviews %>%
  mutate(
    state = case_when(
      state == "California" ~ "CA",
      state == "New York" ~ "NY",
      state == "Texas" ~ "TX",
      state == "Florida" ~ "FL",
      TRUE ~ state # ignores cases when it's already a postal code

    )
  )

view(complete_reviews)
```

Making some transformations to the review data

Now that we've addressed the issues with formatting in the dataset, we're going to make some transformations to the review data. The goal is to evaluate the ratings of each stirng and provide a numerical value for them since we can't do much with a text version of the review score.

```
# Adding a new column with review integers

complete_reviews <- complete_reviews %>%
  mutate(
    review_num = case_when(
      review == "Poor" ~ 1,
      review == "Fair" ~ 2,
      review == "Good" ~ 3,
      review == "Great" ~ 4,
      review == "Excellent" ~ 5
    ), is_high_review = if_else( review_num >= 4, TRUE, FALSE)
  )
view(complete_reviews)
```

Our main question, is to determine which book is most profitable. So going forward we need to think how we define this. It could be the book that sells the most overall, or it can be a combination of those factors to see which book generates the most revenue overall.

Analyzing the data

In my opinion the most profitable book is the one that continues to leave the shelf at a high rate. Sure there are some books that might sell less, but have a higher value so they generate a higher profit, but for the sake of early exploration let's focus on simplicity.

Our process for this analysis will be to: 1. Group the books by their name 2. Summarize these and pass them into a new column called purchase 3. Summing the total price column up, would also be interesting.

```
complete_reviews %>%
  group_by(book) %>%
  summarize(
    purchased = n()
  ) %>%
  arrange(-purchased)
```

```
## # A tibble: 5 x 2
##   book                                purchased
##   <chr>                                   <int>
## 1 Fundamentals of R For Beginners           366
## 2 R For Dummies                             361
## 3 Secrets Of R For Advanced Students        360
## 4 Top 10 Mistakes R Beginners Make          355
## 5 R Made Easy                               352
```

It looks likes the book "Fundamentals of R For Beginners" had the most copies purchased, but overall these books seem to be purchased at the relatively same amounts which warrants a further analysis. I'll group each book and then sum their total prices to see which price is the highest overall.

```
complete_reviews %>%
  group_by(book) %>%
  summarize(
    total_revenue = sum(price)
  ) %>%
  arrange(-total_revenue)
```

```
## # A tibble: 5 x 2
##   book                               total_revenue
##   <chr>                                      <dbl>
## 1 Secrets Of R For Advanced Students        18000 
## 2 Fundamentals of R For Beginners           14636.
## 3 Top 10 Mistakes R Beginners Make          10646.
## 4 R Made Easy                                7036.
## 5 R For Dummies                              5772.
```

So it looks like the book that brought in the most money was "Secrets OF R For Advanced Students" which would make sense since the book sells for $50 per copy and sold for a total of 360 copies. The second placed book "Fundamentals of R for Beginners" sold for 39.99 per copy for 366 copies. That's a little more than 10 extra dollars in revenue for each copy for "Secretes of R".

Exploring geographical relationships

Another question we could ask is there are any relationships between the books sold and the state they are sold in. Maybe some books sell better in some states when compared to others.

To examine this we'll need to: 1. Group by book 2. Sum the total revenue 3. Analyze each state?

```
complete_reviews %>%
  group_by(book, state) %>%
  summarize(
    total_revenue = sum(price)
  ) %>% arrange(-total_revenue)
```

```
## `summarise()` has grouped output by 'book'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 20 x 3
## # Groups:   book [5]
##    book                               state total_revenue
##    <chr>                              <chr>         <dbl>
##  1 Secrets Of R For Advanced Students NY            5400 
##  2 Secrets Of R For Advanced Students FL            4300 
##  3 Secrets Of R For Advanced Students CA            4200 
##  4 Secrets Of R For Advanced Students TX            4100 
##  5 Fundamentals of R For Beginners    CA            3959.
##  6 Fundamentals of R For Beginners    NY            3879.
##  7 Fundamentals of R For Beginners    TX            3839.
##  8 Top 10 Mistakes R Beginners Make   NY            3059.
##  9 Fundamentals of R For Beginners    FL            2959.
## 10 Top 10 Mistakes R Beginners Make   TX            2759.
## 11 Top 10 Mistakes R Beginners Make   FL            2519.
## 12 Top 10 Mistakes R Beginners Make   CA            2309.
## 13 R Made Easy                        NY            1919.
## 14 R For Dummies                      CA            1919.
## 15 R Made Easy                        TX            1739.
## 16 R Made Easy                        FL            1699.
## 17 R Made Easy                        CA            1679.
## 18 R For Dummies                      TX            1327.
## 19 R For Dummies                      NY            1295.
## 20 R For Dummies                      FL            1231.
```

Takeaways

- It looks like the Secrets of R was most profitable in NY.
- Fundamentals of R was most profitable in CA.
- Top 10 mistatkes R Beginners make was most profitable in NY.
- R made easy was most profitable in NY.
- R for dummies was most profitable in CA.

Overall, the most profitable markets for books in R were NY and California.

Conclusions

We found that the most profitable book sold was "Secrets Of R For Advanced Students" and the most profitable markets were NY and CA. I'd recommend selling more copies of that book in NY, and in general keeping more copies in that market.