Workflow For Efficient Data Analysis in R Part 2

Clifton Lee

Last Modified: 26 June, 2021

Introduction

In this project, I am playing the position of an analyst for a publishing company. The firm has given me data on some of its 2019 book sales and expects me to extract some useful information from it. On July 1st, 2019, they introduced a new campaign encouraging consumers to buy more books, and they want to know if this new program was effective in growing sales and enhancing review quality. This will be my duty as the analyst to figure out.

Business Problem

The publishing firm is unsure if their marketing attempts to increase sales were successful or not, and as such, they need to know if they are investing wisely or if they need to cut/refocus the marketing budget.

Data Understanding

I need to organize things in terms of a process, much like any other data analysis task. This approach will give me a solid foundation for dealing with any problems that may occur. Before I even think about data analysis, I need to look at the data and make a note of any possible problems. If you would like to look at the dataset in more detail, you may do so by clicking here.

```
dataset <- read_csv("sales2019.csv") # importing the dataset
```

How big is the dataset?

```
dim(dataset) # dimensions of the datset
```

```
## [1] 5000    5
```

As shown above, the dataset contains 5000 rows and 5 columns.

What are the column names and what do they seem to represent?

```
colnames(dataset) # column names
```

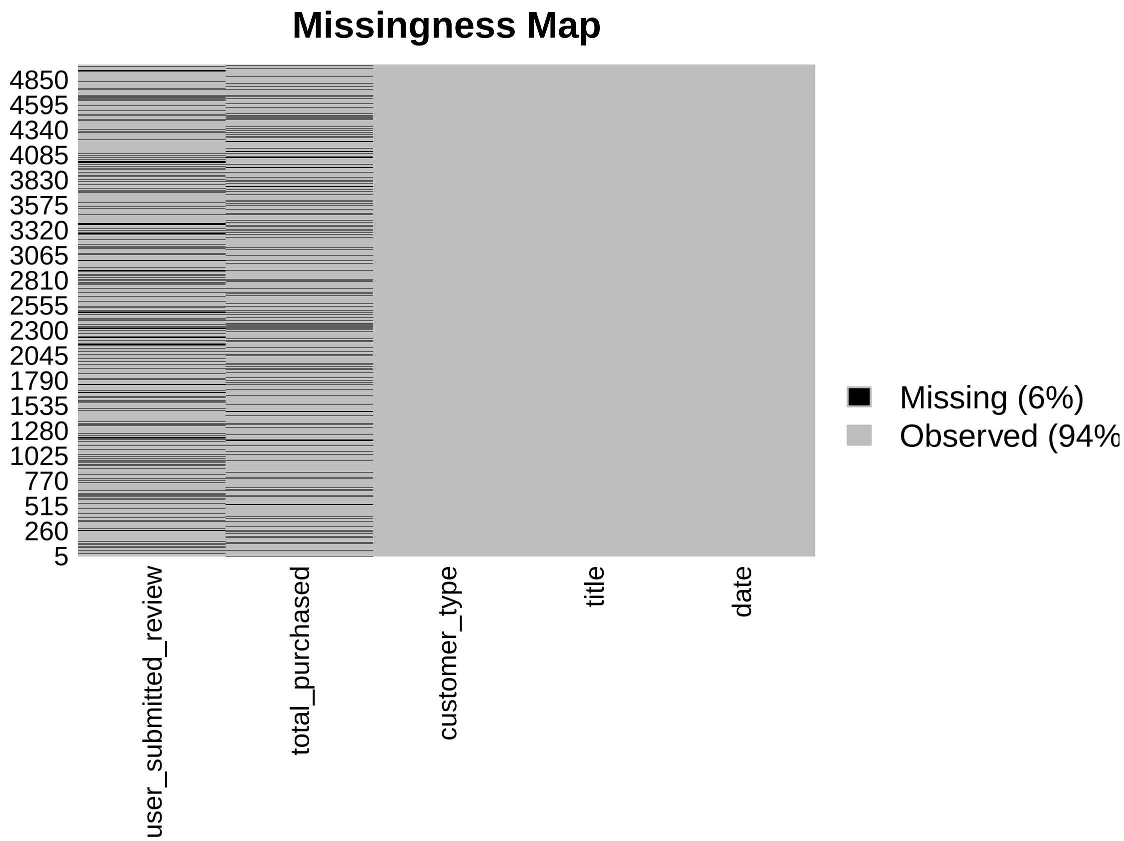
```
## [1] "date"                  "user_submitted_review" "title"
## [4] "total_purchased"       "customer_type"
```

```
glimpse(dataset) # data structure
```

```
## Rows: 5,000
## Columns: 5
## $ date                  <chr> "5/22/19", "11/16/19", "6/27/19", "11/6/...
## $ user_submitted_review <chr> "it was okay", "Awesome!", "Awesome!", "...
## $ title                 <chr> "Secrets Of R For Advanced Students", "R...
## $ total_purchased       <dbl> 7, 3, 1, 3, NA, 1, 5, NA, 7, 1, 7, NA, 3...
## $ customer_type         <chr> "Business", "Business", "Individual", "I...
```

Do any of the columns have missing data?

```
missmap(dataset, col = c('black', 'grey'), margins = c(7.5,5)) #missing data map
```



Data Cleaning and Pre-processing

I can see from the above missing data map that there are missing values in two columns.

The first is the `user_submitted_review` column, which includes the customer's review. The second field is `total_purchased`, which indicates how many books the client bought.

I'm going to approach these two columns differently. This is because I am far more interested in the `total_purchased` column, which includes real statistics on book sales. I'd want to know whether the company's new program contributed to increased sales. I'm going to use a different method to dealing with missing data here in order to preserve as much information on sales as possible.

In short, I'm going to delete any entries in `user_submitted_review` column that contain missing data. I'm going to use a little more complex technique with `total_purchased` column. I'm going to do what is called imputation. So, I am going to replace all 'NA' values with an average value calculated from the `total_purchased` column. Filling in the blanks with average numbers is helpful since they are often the best estimates for what the purchase would have been.

```
dataset <- dataset %>%
           filter(!is.na(user_submitted_review))

dim(dataset)
```

```
## [1] 4115    5
```

Now with the empty rows in the `user_submitted_review` column removed. I now have 4115 rows and 5 columns.

I'll now focus on imputing the remaining missing rows in the `total_purchased` column with the average number of books purchased.

```
# average number of books purchaed in the total purchased column
dataset %>%
  summarize(avg_purchased = mean(total_purchased,na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   avg_purchased
##           <dbl>
## 1          3.99
```

```
# Impute the average number of books
dataset <- dataset %>%
            mutate(total_purchased_complete = if_else(is.na(total_purchased),
                                                      round(mean(total_purchased,na.rm = TRUE),digits =
                                                      total_purchased)) %>%
            select(-total_purchased)

head(dataset)
```

```
## # A tibble: 6 x 5
##   date   user_submitted_rev~ title         customer_type total_purchased_c~
##   <chr>  <chr>               <chr>         <chr>                      <dbl>
## 1 5/22/~ it was okay         Secrets Of R~ Business                       7
## 2 11/16~ Awesome!            R For Dummies Business                       3
## 3 6/27/~ Awesome!            R For Dummies Individual                     1
## 4 11/6/~ Awesome!            Fundamentals~ Individual                     3
## 5 7/18/~ Hated it            Fundamentals~ Business                       4
## 6 1/28/~ Never read a bette~ Secrets Of R~ Business                       1
```

When compared to the simplicity of managing numeric data, string data may be very complex to deal with. One reason for this is because there are numerous languages and many, many words even within a single language. While you combine this with the reality that people may make mistakes when writing, you get some messed-up data.

The `user_submitted_review` column includes sentence-length reviews. I want to be able to categorize reviews as good or bad (positive or negative). This will enable me to count the number of bad or good reviews in my workflow's analysis section. Now, I'll do the required cleaning and processing to convert each of the review phrases into the categories I desire.

```
  unique(dataset$user_submitted_review) #examine the unique sentences in the column
```

```
## [1] "it was okay"
## [2] "Awesome!"
## [3] "Hated it"
## [4] "Never read a better book"
## [5] "OK"
## [6] "The author's other books were better"
## [7] "A lot of material was not needed"
## [8] "Would not recommend"
## [9] "I learned a lot"
```

I'm looking at each of these reviews to see if I can see any particular terms or phrases that will help me determine whether the review is favorable or negative. A word like "good" or "excellent," for example, may convey that a review is favorable. A term like "not recommended" may imply that a review is negative.

```
pos_neg <- function(string){
  case_when(str_detect(string = string,pattern = "it was okay") ~ TRUE,
            str_detect(string = string,pattern = "Awesome!") ~ TRUE,
            str_detect(string = string,pattern = "Hated it") ~ FALSE,
            str_detect(string = string,pattern = "Never read a better book") ~ TRUE,
            str_detect(string = string,pattern = "OK") ~ TRUE,
            str_detect(string = string,pattern = "The author's other books were better") ~ FALSE,
            str_detect(string = string,pattern = "A lot of material was not needed") ~ FALSE,
            str_detect(string = string,pattern = "Would not recommend") ~ FALSE,
            str_detect(string = string,pattern = "I learned a lot") ~ TRUE
            )
}


dataset <- dataset %>%
           mutate(pos_review = pos_neg(user_submitted_review))

dataset
```

```
## # A tibble: 4,115 x 6
##    date   user_submitted_r~ title customer_type total_purchased~ pos_review
##    <chr>  <chr>             <chr> <chr>                    <dbl> <lgl>
##  1 5/22/~ it was okay       Secr~ Business                     7 TRUE
##  2 11/16~ Awesome!          R Fo~ Business                     3 TRUE
##  3 6/27/~ Awesome!          R Fo~ Individual                   1 TRUE
##  4 11/6/~ Awesome!          Fund~ Individual                   3 TRUE
##  5 7/18/~ Hated it          Fund~ Business                     4 FALSE
##  6 1/28/~ Never read a bet~ Secr~ Business                     1 TRUE
##  7 2/20/~ Hated it          R Fo~ Business                     5 FALSE
##  8 12/17~ Awesome!          R Fo~ Business                     4 TRUE
##  9 7/13/~ OK                R vs~ Business                     7 TRUE
## 10 6/22/~ The author's oth~ R Fo~ Business                     1 FALSE
## # ... with 4,105 more rows
```

With the review data and order numbers in a useable format, I can now begin to answer the analysis's primary question: Was the new book program successful in boosting book sales? The program began on July 1st, 2019, and the data I have includes all 2019 sales. There are still some preliminary actions I need to do before running the analysis, so I'll complete them first. To begin, the dates are presently represented as strings. Before I can do any date and time comparisons, they must be correctly structured. Second, I need a clear method to differentiate between sales that occur before and after the program begins. I need to differentiate between these two categories so that I can use what I've learnt to more quickly compute the summary numbers from the data. Finally, this analysis should be presented in a tidy format that anybody reading it can simply comprehend and understand.

```
dataset <- dataset %>%
             mutate(date = mdy(date),
                    program_period = if_else(date >= '2019-07-1',TRUE,FALSE))

head(dataset,10)
```

```
## # A tibble: 10 x 7
##    date       user_submitted_~ title customer_type total_purchased~
##    <date>     <chr>            <chr> <chr>                    <dbl>
##  1 2019-05-22 it was okay      Secr~ Business                     7
##  2 2019-11-16 Awesome!         R Fo~ Business                     3
##  3 2019-06-27 Awesome!         R Fo~ Individual                   1
##  4 2019-11-06 Awesome!         Fund~ Individual                   3
##  5 2019-07-18 Hated it         Fund~ Business                     4
##  6 2019-01-28 Never read a be~ Secr~ Business                     1
##  7 2019-02-20 Hated it         R Fo~ Business                     5
##  8 2019-12-17 Awesome!         R Fo~ Business                     4
##  9 2019-07-13 OK               R vs~ Business                     7
## 10 2019-06-22 The author's ot~ R Fo~ Business                     1
## # ... with 2 more variables: pos_review <lgl>, program_period <lgl>
```

Data Analysis

It's typical in data analysis to have many subgroups that I wish to compare. In this section, I will compare sales before and after July 1, 2019.

```
dataset %>%
  group_by(program_period) %>%
  summarize(total_books_purchased = sum(total_purchased_complete))
```

```
## # A tibble: 2 x 2
##   program_period total_books_purchased
##   <lgl>                          <dbl>
## 1 FALSE                           8215
## 2 TRUE                            8194
```

According to the aforementioned data, the total number of books bought was higher before the commencement of the campaign on July 1st, 2019 than after the promotion. However, as an analyst, I must go deeper to determine the source of those figures. Individual consumers may have reacted better to the campaign and purchased more books as a result. Alternatively, more books might have been purchased by companies. In order to investigate this sub-analysis, I must additionally split the sales before and after July 1, 2019 into sales for people vs sales for companies.

```
dataset %>%
   group_by(program_period,customer_type) %>%
  summarize(total_books_purchased = sum(total_purchased_complete))
```

```
## # A tibble: 4 x 3
## # Groups:   program_period [2]
##   program_period customer_type total_books_purchased
##   <lgl>          <chr>                         <dbl>
## 1 FALSE          Business                       5615
## 2 FALSE          Individual                     2600
## 3 TRUE           Business                       5745
## 4 TRUE           Individual                     2449
```

According to the sub-analysis above, sales for business customers rose after the promotional period, while individual sales dropped at a faster pace, indicating that the period before the campaign performed better overall.

The final question I need to address with the data is, "Did the program enhance review scores?" To determine whether a review was good or negative, I'll need to utilize the new column I generated (`pos_review`).

```
dataset %>%
  group_by(program_period) %>%
  summarize(positive_reviews = sum(pos_review))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   program_period positive_reviews
##   <lgl>                     <int>
## 1 FALSE                      1134
## 2 TRUE                       1128
```

As per the summary table above, the review sentiments have not improved since the campaign's program inception.

Conclusion

Based on my examination of the data given, it seems that the new campaign has not been helpful in terms of gaining new sales or improving review quality. I would recommend that the business reallocate some of its marketing money to other investments.