Book Sales After Campaign Program

2022-10-18

Introduction

In our previous project we looked at the sales of books from a book company to determine how well each of the books performed in general and how well the books performed in the different regions they were sold. In this project, our goal is to evaluate the effectiveness of a program launched by the company to convince customers to buy more books. Primarily, we want to know if the program was effective overall and how effective the program was on the different types of customers who buy the books. You can find the analysis of the previous project here. We are also going to be following the same principles that guided us in the previous project. That is:

- Data Exploration.
- Data Cleaning.
- Data Analysis.

Data Exploration

This is the initial and very crucial step in every data analyis project because it involves familiarizing yourself with the data at hand.

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

# reading the data
book_sales <- read_csv("sales2019.csv")
```

```
# creating function to display tables
render_table <- function(table, scale_down=F){
  if(scale_down == T){
     rendered_table <- kbl(table) %>% kable_styling(
    latex_options = c("stripe", "HOLD_position", "scale_down")
  )
  } else{
   rendered_table <- kbl(table) %>% kable_styling(
    latex_options = c("stripe", "HOLD_position")
  )

  }

  return(rendered_table)
}
```

```
# getting the dimension of the data
book_sales %>% dim()
```

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

```
# getting an overview of the data
book_sales %>% glimpse()
```

```
## Rows: 5,000
## Columns: 5
## $ date                   <chr> "5/22/19", "11/16/19", "6/27/19", "11/6/19", "7/~
## $ user_submitted_review <chr> "it was okay", "Awesome!", "Awesome!", "Awesome!~
## $ title                  <chr> "Secrets Of R For Advanced Students", "R For Dum~
## $ total_purchased        <dbl> 7, 3, 1, 3, NA, 1, 5, NA, 7, 1, 7, NA, 3, 2, 0, ~
## $ customer_type          <chr> "Business", "Business", "Individual", "Individua~
```

Our data set has a total of 5000 rows and 5 columns. 4 of the 5 columns have `character` data type with the exception of the `total_purchased` column which has `double` data type. Now let's look at the first 6 rows.

```
book_sales %>% head() %>% render_table()
```

date	user_submitted_review	title	total_purchased	customer_type
5/22/19	it was okay	Secrets Of R For Advanced Students	7	Business
11/16/19	Awesome!	R For Dummies	3	Business
6/27/19	Awesome!	R For Dummies	1	Individual
11/6/19	Awesome!	Fundamentals of R For Beginners	3	Individual
7/18/19	Hated it	Fundamentals of R For Beginners	NA	Business
1/28/19	Never read a better book	Secrets Of R For Advanced Students	1	Business

The final step in our exploration is to check each column for null values.

```
# checking null values
for(col in colnames(book_sales)){
  null_val <- book_sales %>% pull(col) %>% is.na %>% sum()
  paste(col, ":", null_val) %>% print()
}
```

```
## [1] "date : 0"
## [1] "user_submitted_review : 885"
## [1] "title : 0"
## [1] "total_purchased : 718"
## [1] "customer_type : 0"
```

We have null values in both the `user_submitted_review` column and the `total_purchased` column. The former has 885 null values while the latter has 718 null values.

Data Cleaning

The first step in our data cleaning is handling the null values. We are going to treat the null values in both columns differently. For the `user_submitted_review` column, we are going to drop rows where it is null and for the the `total_purchased` column, we are going to be imputing the null values with the mean of the column.

```
# removing rows where review value is null
book_sales <- book_sales %>% filter(!is.na(user_submitted_review))
book_sales %>% dim()
```

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

After removing rows where `user_submitted_review` is null, we are left with 4115 rows. Now we can go ahead to impute the `total_purchased` column.

```
mean_purchase <- (
  book_sales  %>% filter(!is.na(total_purchased))
  %>% pull(total_purchased) %>% mean()
  %>% round()
)

book_sales <- book_sales %>% mutate(
  total_purchased = if_else(is.na(total_purchased),
                            mean_purchase, total_purchased)
)
```

```
# confirming there are no null values
for(col in colnames(book_sales)){
  null_val <- book_sales %>% pull(col) %>% is.na %>% sum()
  paste(col, ":", null_val) %>% print()
}
```

```
## [1] "date : 0"
## [1] "user_submitted_review : 0"
## [1] "title : 0"
## [1] "total_purchased : 0"
## [1] "customer_type : 0"
```

We have managed to deal with the null values in our data set. The next step is to clean the `user_submitted_review` column.

```
# getting the unique values of user_submitted_review column
book_sales %>% pull(user_submitted_review) %>% unique()
```

```
## [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"
```

Since the values in this column are texts, we need to clean it in such a way that reviews are only classified as positive or not positive reviews. In this case, we are going to be conservative and also classify the words `OK` and `ok` as positive also. We will be creating a new column called `positive_review`.

```
# extracting the positive reviews from the unique reviews in the data set
unique_reviews <- book_sales %>% pull(user_submitted_review) %>% unique()
positive_reviews <- unique_reviews[c(1, 2, 4, 5, 9)]
positive_reviews %>% print()
```

```
## [1] "it was okay"              "Awesome!"
## [3] "Never read a better book" "OK"
## [5] "I learned a lot"
```

```
# creating positive_review column
book_sales <- book_sales %>% mutate(
  positive_review = if_else(user_submitted_review %in% positive_reviews,
                            TRUE, FALSE)
)
```

```
book_sales %>% head() %>% render_table(scale_down = T)
```

date	user_submitted_review	title	total_purchased	customer_type	positive_review
5/22/19	it was okay	Secrets Of R For Advanced Students	7	Business	TRUE
11/16/19	Awesome!	R For Dummies	3	Business	TRUE
6/27/19	Awesome!	R For Dummies	1	Individual	TRUE
11/6/19	Awesome!	Fundamentals of R For Beginners	3	Individual	TRUE
7/18/19	Hated it	Fundamentals of R For Beginners	4	Business	FALSE
1/28/19	Never read a better book	Secrets Of R For Advanced Students	1	Business	TRUE

Finally we have to clean the date column. When we checked the data type for the date column, it had a `character` data type. We are going to convert the values in the column to dates using the `lubridate` library.

```
# cleaning the date
book_sales <- book_sales %>% mutate(
  date = mdy(date)
)
```

One more thing we have to do is to indicate rows where books were purchased before and after the program. We are going to create a program column, if books were purchased before the program, the value will be "No" and if they were purchased during and after, the value will be "Yes".

```
# indicating rows before and after program period
program_period <- mdy("07/01/2019")

book_sales <- book_sales %>% mutate(
  program = case_when(
    date >= program_period ~ "Yes",
    date < program_period ~ "No"
  )
)

book_sales %>% head() %>% render_table(scale_down = T)
```

date	user_submitted_review	title	total_purchased	customer_type	positive_review	program
2019-05-22	it was okay	Secrets Of R For Advanced Students	7	Business	TRUE	No
2019-11-16	Awesome!	R For Dummies	3	Business	TRUE	Yes
2019-06-27	Awesome!	R For Dummies	1	Individual	TRUE	No
2019-11-06	Awesome!	Fundamentals of R For Beginners	3	Individual	TRUE	Yes
2019-07-18	Hated it	Fundamentals of R For Beginners	4	Business	FALSE	Yes
2019-01-28	Never read a better book	Secrets Of R For Advanced Students	1	Business	TRUE	No

Was The Program Effective?

Now that we have finished cleaning and transforming our data, we can go ahead to answer the questions we had.

```
# getting summary of program effectiveness
program_summary <- book_sales %>% group_by(program) %>% summarise(
  sales = total_purchased %>% sum(),
  positive_review_perc = (positive_review %>% sum() / nrow(book_sales)) %>% round(2)
)

program_summary %>% render_table()
```

program	sales	positive_review_perc
No	8215	0.28
Yes	8194	0.27

Generally, the program didn't seem to have any effect on book sales. The sales before and after the program are roughly the same, with the sale before the program being just a little bit higher. The same can be said about the positive reviews too which dropped from 28% to 27% after the program. In a more general sense, the program wasn't effective. However we want to test its effectiveness on the different customer type.

```
customer_program_summary <- book_sales %>% group_by(program, customer_type) %>% summarise(
  sales = total_purchased %>% sum(),
  positive_review_perc = (positive_review %>% sum() / nrow(book_sales)) %>% round(2),
  .groups = "drop"
)

customer_program_summary %>% render_table()
```

program	customer_type	sales	positive_review_perc
No	Business	5615	0.19
No	Individual	2600	0.09
Yes	Business	5745	0.20
Yes	Individual	2449	0.08

When we look at the effectiveness on the different customer types, we find out that it's a bit of a different outcome. The sales for business customers went up slightly, while that for individuals dropped slightly after the program. Although we can't conclusively say for now that these changes was because of the program. The positive reviews however remained roughly the same for both customer types before and after the program.

Conclusion

Our goal overall was to find out how effective the program done by the company was, after looking at the effect that it had on the sales overall, we can conclude that it wasn't an effective program. Although there was a slight change between its effect on business customers and individual customers, we will have to perform a hypothesis test to confirm if these differences hold any statistical significance. "'