Using R for Basic Customer Analysis at Bookbinders.

Shiv Viswanathan 2023FA_MKTG_482-0_SEC81

Attached - Please see Disclosure at the end about GenerativeAI tool usage

Preliminaries

Load packages:

```
library(gmodels)
library(modelr)
library(janitor)
library(haven)
library(readxl)
library(knitr)
library(psych)
library(statar)
library(tidyverse)
```

Read in the data:

```
# use load("filename.Rdata") for .Rdata files
 load("bbb.Rdata")
```

Assignment questions and answers

1. Report the number and proportion (as a decimal) of customers by gender. Please use `tabyl()` for this calculation. For this and all questions, enter your code in the gray area that appears below the question.

```
library(janitor)

# Create a table of counts and proportions by gender
gender_table <- tabyl(bbb, gender)

# Print the table
print(gender_table)
```

```
##  gender     n percent
##       M 16698 0.33396
##       F 33302 0.66604
```

Interpretation - Female customers is almost double that of Male customers

2. Report the number and proportion (as a decimal) of customers by gender. Please use `dplyr` verbs `group_by` and `summarize` for this calculation. You can use the function `n()` inside summarize to obtain the number of observations. Also, remember that you can do arithmetic when you define summary expressions in `summarize`. Note: `dplyr` functions such as `arrange`, `filter`, `group_by`, `mutate`, `select`, and `summarize` are by convention called verbs.

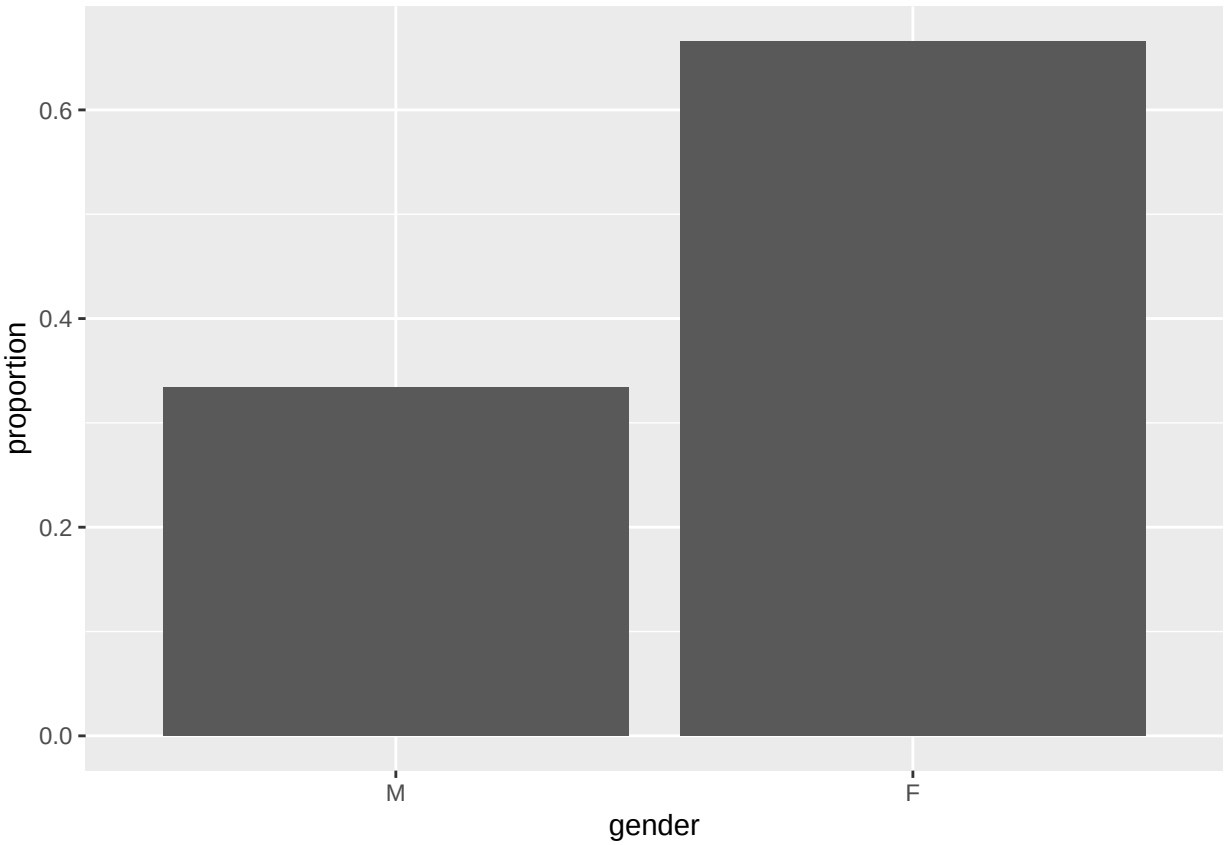
```
total_customers <-nrow(bbb)
df <- bbb %>%
   group_by(gender) %>%
   summarise(count = n(),proportion = count / total_customers)

   print (df)
```

```
## # A tibble: 2 x 3
##   gender count proportion
##   <fct>  <int>      <dbl>
## 1 M      16698      0.334
## 2 F      33302      0.666
```

Interpretation - Female customers is almost double that of Male customers

3. Create a bar graph visualizing the proportion of customers by gender (the second number you just calculated above).

```
ggplot(df, aes(x = gender, y = proportion)) +
geom_bar(stat="identity")
```



Interpretation - Female customers is almost double that of Male customers

4. Report the average total dollars spent, the average total number of book purchased, and the average number of months since last purchase (see the "total", "purch", and "last" variables.) Please use `dplyr` verbs for this calculation.

```
 average <- bbb %>%
  summarise(avg_total_dollars = mean(total), avg_books_purchased = mean(purch), avg_months_since_last_pu
# print average
#Error: unexpected symbol in "print average"
 print (average)
```

```
## # A tibble: 1 x 4
##   avg_total_dollars avg_books_purchased avg_months_since_last_pur~1 average_time
##               <dbl>               <dbl>                       <dbl>        <dbl>
## 1              208.                3.89                        12.4         13.3
## # i abbreviated name: 1: avg_months_since_last_purchase
```

Interpretation – Customers spend about $208 on average on books on average, and purchase about 3.89 books - there are customers who have not purchased any book

```
#Redoing analysis for non-repeat customers alone

average <- bbb %>%
   filter(purch==1) %>%   # to filter for non repeat customers
   summarise(avg_total_dollars = mean(total), avg_books_purchased = mean(purch), avg_months_since_last_
# print average
```

```
#Error: unexpected symbol in "print average"
 print (average)
```

```
## # A tibble: 1 x 3
##   avg_total_dollars avg_books_purchased avg_months_since_last_purchase
##               <dbl>               <dbl>                         <dbl>
## 1              164.                   1                          12.3
```

```
#Redoing analysis for repeat customers alone

average <- bbb %>%
   filter(purch>1) %>%   # to filter for  repeat customers
   summarise(avg_total_dollars = mean(total), avg_books_purchased = mean(purch), avg_months_since_last_
# print average
#Error: unexpected symbol in "print average"
 print (average)
```

```
## # A tibble: 1 x 3
##   avg_total_dollars avg_books_purchased avg_months_since_last_purchase
##               <dbl>               <dbl>                         <dbl>
## 1              227.                5.14                          12.4
```

##Interpretation - Repeat customers do spend more on average and purchase more books - but interestingly their average months still last purchase is still high

5. Which three states account for the largest number of BookBinders' customers? How many customers are there in each of these three states? Show the data sorted in descending order by number of customers. Please use `dplyr` verbs for this calculation. Recall that the `dplyr` verb `arrange` lets you sort. If you want to sort in descending order, put a `-` in front of the sorting variable.

```
 top_three_states <-bbb %>%
  group_by(state) %>%
  summarise (n_customers_per_state = n()) %>%
  arrange (desc(n_customers_per_state)) %>%
  head (3)
 print (top_three_states)
```

```
## # A tibble: 3 x 2
##   state n_customers_per_state
##   <chr>                 <int>
## 1 NY                    16530
## 2 NJ                    11068
## 3 PA                     8718
```

```
 str(top_three_states)
```

```
## tibble [3 x 2] (S3: tbl_df/tbl/data.frame)
##  $ state                : chr [1:3] "NY" "NJ" "PA"
##  $ n_customers_per_state: int [1:3] 16530 11068 8718
```

```
## rerunning to find no of customers across all 50 states
all_states <-bbb %>%
  group_by(state) %>%
  summarise (n_customers_per_state = n(), percentage_per_state =  (n()/total_customers)*100 ) %>%
```

```
  arrange (desc(percentage_per_state)) %>%
  head (50)
 print (all_states)
```

```
## # A tibble: 15 x 3
##    state n_customers_per_state percentage_per_state
##    <chr>                 <int>                <dbl>
##  1 NY                    16530                33.1
##  2 NJ                    11068                22.1
##  3 PA                     8718                17.4
##  4 MA                     4252                 8.50
##  5 MD                     4172                 8.34
##  6 CT                     2512                 5.02
##  7 DE                      711                 1.42
##  8 NH                      665                 1.33
##  9 RI                      402                 0.804
## 10 ME                      343                 0.686
## 11 DC                      339                 0.678
## 12 VT                      211                 0.422
## 13 VI                       45                 0.09
## 14 VA                       27                 0.054
## 15 AE                        5                 0.01
```

Customers only in 15 states - - top three states constitute 70% of all customers

6. For each of the three states you just identified, report the average total spending per customer (see the `total` variable). Please exclude all other states from the analysis.

```
all_states <-bbb %>%
  group_by(state) %>%
  summarise (n_customers_per_state = n(), avg_spending_per_customer = mean(total) ) %>%
  arrange (desc(n_customers_per_state)) %>%
  head (3)
 print (all_states)
```

```
## # A tibble: 3 x 3
##   state n_customers_per_state avg_spending_per_customer
##   <chr>                 <int>                     <dbl>
## 1 NY                    16530                      208.
## 2 NJ                    11068                      208.
## 3 PA                     8718                      211.
```

Average spending is high in PA compared to the other two states

7. Calculate the correlation between customers' total spending on non-book products and on books (see the `nonbook` and `book` variables). See the R tutorial for how to calculate correlations.

```
library(psych)

# Attaching package: 'psych'
#
```

```
# The following objects are masked from 'package:ggplot2':
#
#     %%, alpha

 corr_calc <- bbb %>%
 select(book, nonbook) %>%
 corr.test()
 #Call:corr.test(x = .)
#Error: object 'Call' not found
 print(corr_calc)
```

```
## Call:corr.test(x = .)
## Correlation matrix
##         book nonbook
## book    1.00    0.16
## nonbook 0.16    1.00
## Sample Size
## [1] 50000
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##         book nonbook
## book       0       0
## nonbook    0       0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```
#Call:corr.test(x = .)
```

There is positive correlation between total dollars spent on book and nonbook products. Which means, the book buyers are also purchasing non book products. This is useful to know for cross selling and cross marketing.

```
## calculating correlation of only female customers - between book and non book products
library(psych)
 corr_calc <- bbb %>%
   filter(gender =="F") %>%
   select(book, nonbook) %>%
  corr.test()
 #Call:corr.test(x = .)
#Error: object 'Call' not found
 print(corr_calc)
```

```
## Call:corr.test(x = .)
## Correlation matrix
##         book nonbook
## book    1.00    0.15
## nonbook 0.15    1.00
## Sample Size
## [1] 33302
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##         book nonbook
## book       0       0
## nonbook    0       0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```
## calculating correlation of only male customers - between book and non book products
library(psych)
corr_calc <- bbb %>%
  filter(gender =="M") %>%
  select(book, nonbook) %>%
 corr.test()
#Call:corr.test(x = .)
#Error: object 'Call' not found
print(corr_calc)
```

```
## Call:corr.test(x = .)
## Correlation matrix
##         book nonbook
## book    1.00    0.16
## nonbook 0.16    1.00
## Sample Size
## [1] 16698
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##         book nonbook
## book       0       0
## nonbook    0       0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

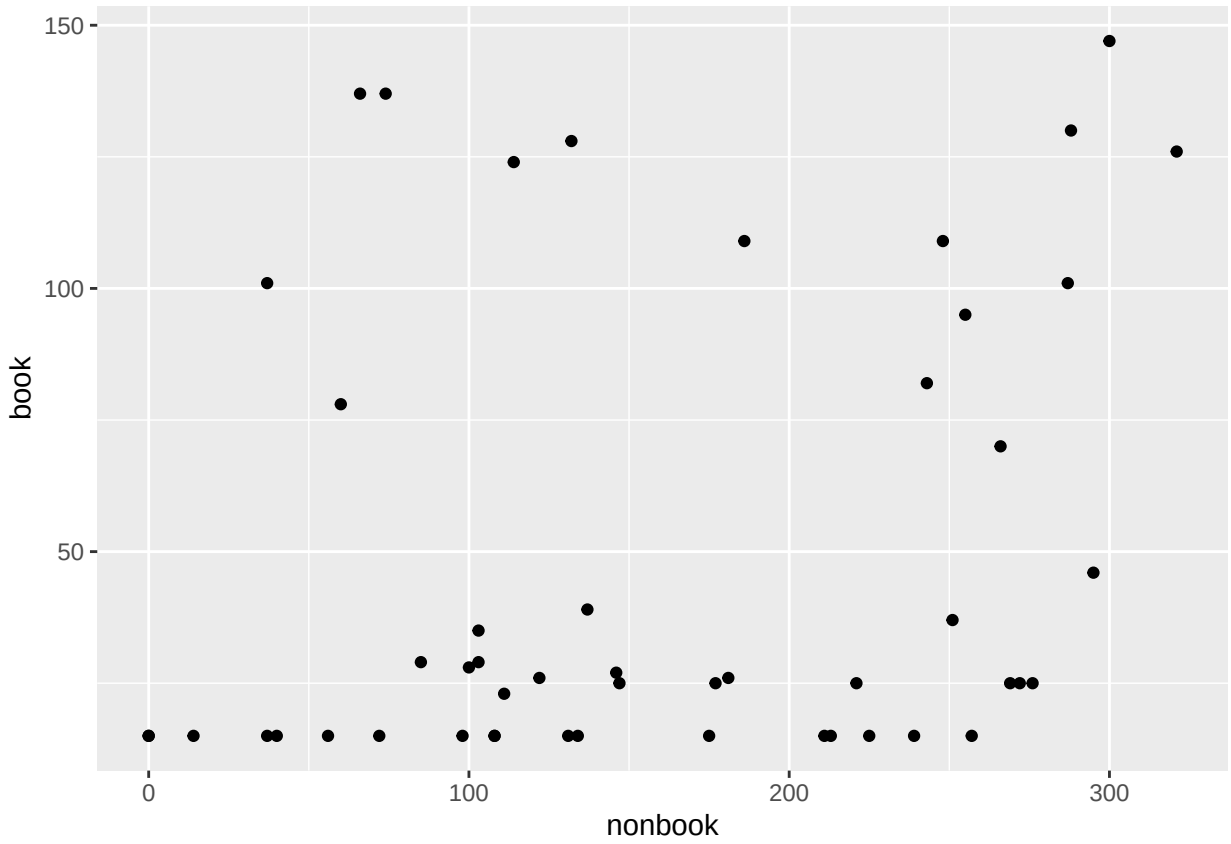
```
## calculating total books and non books among M and F
library(psych)
book_nonbook <- bbb %>%
  group_by(gender) %>%
  summarise(total_books = sum(book), total_nonbooks = sum(nonbook))
print(book_nonbook)
```

```
## # A tibble: 2 x 3
##   gender total_books total_nonbooks
##   <fct>        <int>          <int>
## 1 M           973709        2760653
## 2 F          1367811        5313743
```

male customers show more correlation than female customers ? male purchase more non book products!

8. For the first fifty customers in the dataset only (use the `dplyr` verb `slice`), create a scatter plot showing the relationship between customers' total spending on non-book products and on books.

```
library(ggplot2)
scatter_plot <- bbb %>%
slice(1:50) %>%
ggplot(aes(x = nonbook, y = book)) + geom_point()
print(scatter_plot)
```

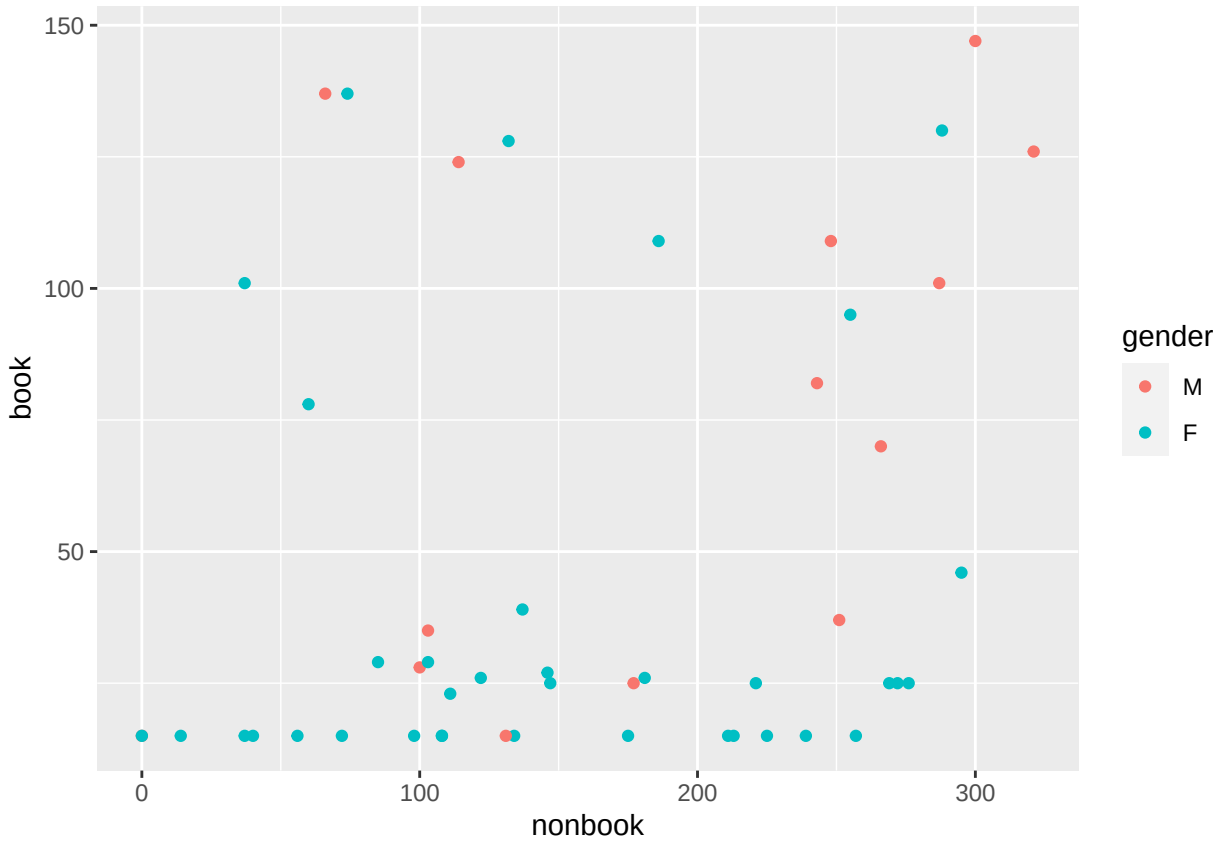



```
# ggplot(df, aes(x = gender, y = proportion)) +
#          geom_bar(stat="identity")
```

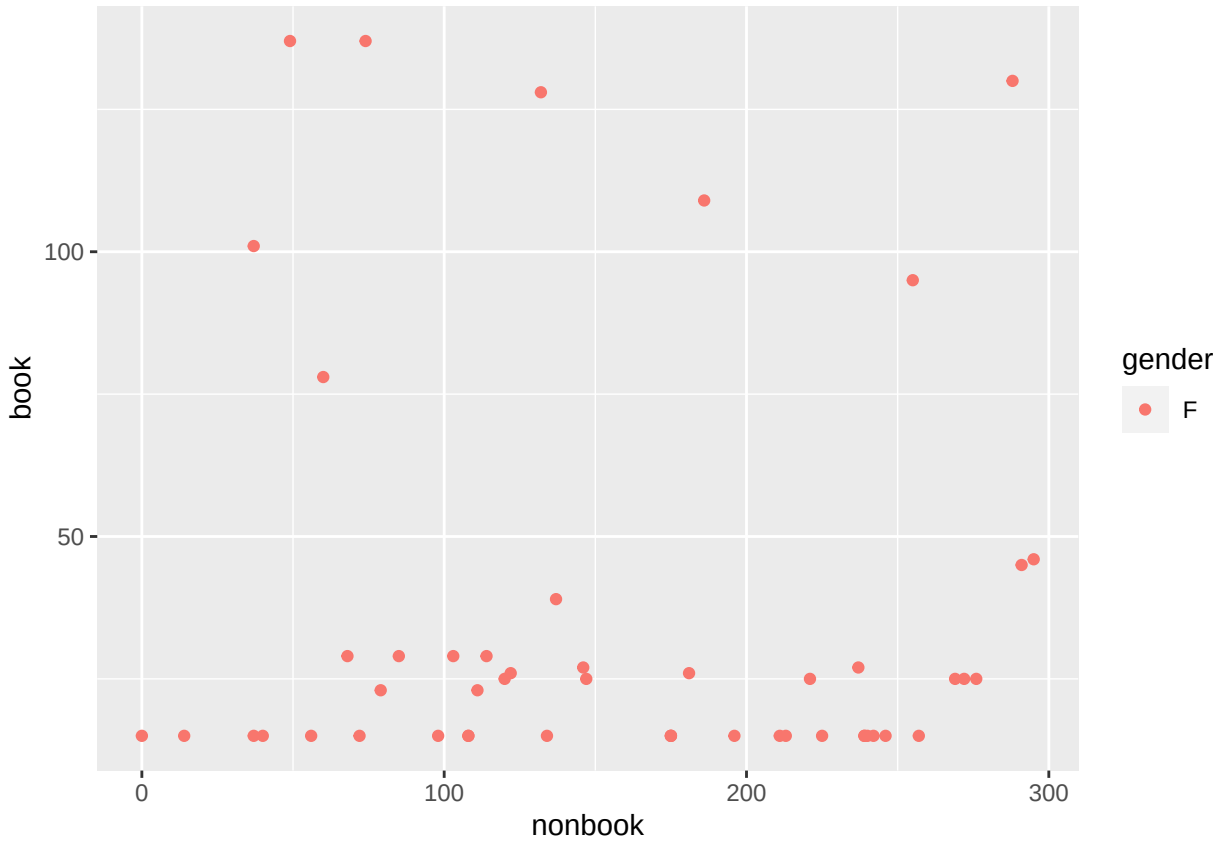
The scatter plot shows correlation but relatively weak

9. Repeat the previous graph but color the points by gender.

```
library(ggplot2)
scatter_plot <- bbb %>%
     slice(1:50) %>%
     ggplot(aes(x = nonbook, y = book, color = gender)) +  geom_point()
print(scatter_plot)
```




```
## repeating scatter plot for only Female customers
library(ggplot2)
 scatter_plot <- bbb %>%
      filter(gender=="F") %>%
      slice(1:50) %>%
      ggplot(aes(x = nonbook, y = book, color = gender)) +  geom_point()
 print(scatter_plot)
```



10. Report how many books were sold in each book category. Just eyeballing the data (not sorting it in R), which category sold the most books? Which sold the least books?

```
books_category <- c("youth", "cook", "do_it", "reference", "art", "geog")
 sums <- colSums(bbb[books_category])
 print(sums)
```

```
##     youth      cook     do_it reference       art      geog
##     19549     46830     23153     15612     19296     27348
```

11. Report the number and proportion of customers by gender who bought "The Art History of Florence" (see the `buyer` variable).

```
buyers <- bbb %>%
 filter(buyer == 1)
 gender_summary <- buyers %>%
 group_by(gender) %>%
 summarise( no_buyers = n(), proporion = n()/ nrow(buyers))
 print(gender_summary)
```

```
## # A tibble: 2 x 3
##   gender no_buyers proporion
##   <fct>      <int>     <dbl>
## 1 M           2133     0.472
## 2 F           2389     0.528
```

```
print(nrow(buyers))
```

```
## [1] 4522
```

12. Report the total number of purchases and the average number of purchases by gender (see the `purch` variable).

```
purchase_summary <- bbb %>%
    group_by(gender) %>%
    summarise(total_purchases = sum(purch),average_purchases = mean(purch))
print(purchase_summary)
```

```
## # A tibble: 2 x 3
##   gender total_purchases average_purchases
##   <fct>            <int>             <dbl>
## 1 M                82543              4.94
## 2 F               111968              3.36
```

Male buy more on average than female. Though number of male customers are less than female.

13. Determine the minimum, the maximum, and the average number of months between customers' first purchase and their most recent purchase. In doing this, use the `dplyr` verb `mutate` to create a new variable relevant for these calculations.

```
bbb <- bbb %>%
    mutate(months_between = first - last)  ## to calculate months between the first and last purchase

 # Calculate the minimum, maximum, and average number of months
 result <- bbb %>%
    summarise(
        min_months = min(months_between),
        max_months = max(months_between),
        avg_months = mean(months_between)
    )

 print(result)
```

```
## # A tibble: 1 x 3
##   min_months max_months avg_months
##        <int>      <int>      <dbl>
## 1          0         72       13.3
```

##there are customers who are active for about 6 years (72 months), and on overage customers stay for more than 1 year.(13 months)

14. What proportion of repeat customers (those with two or more total purchases) bought "The Art History of Florence?"

```
## calculcate total number of repeat customers
total_repeat_customers  <-bbb %>%
  filter(purch >1) %>%
```

```
  summarise(count = n())
print(total_repeat_customers)
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1 34880
```

```
AHF <- bbb %>%
  group_by(buyer) %>%
  filter(purch>1) %>%  ## filter repeat customers
  summarise(no_buyers = n(), proportion = n()/total_repeat_customers$count) ## proportion of repeat cus
 print(AHF)
```

```
## # A tibble: 2 x 3
##   buyer no_buyers proportion
##   <int>     <int>      <dbl>
## 1     0     31282      0.897
## 2     1      3598      0.103
```

##10% of the repeat customers bought the book - 89% of repeat customers did not buy the book

```
## calculcate total number of non repeat customers
total_repeat_customers  <-bbb %>%
  filter(purch ==1) %>%
  summarise(count = n())
print(total_repeat_customers)
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1 15120
```

```
AHF <- bbb %>%
  group_by(buyer) %>%
  filter(purch==1) %>%  ## filter non repeat customers
  summarise(no_buyers = n(), proportion = n()/total_repeat_customers$count) ## proportion of non repeat
 print(AHF)
```

```
## # A tibble: 2 x 3
##   buyer no_buyers proportion
##   <int>     <int>      <dbl>
## 1     0     14196     0.939
## 2     1       924     0.0611
```

##6% of new customers bought the book - they are first time buyers

```
AHF <- bbb %>%
  filter(buyer==1) %>%  ## filter those who bought the book
  group_by(purch) %>%
  summarise(no_buyers = n()) ## number of books these AFH buyers have purchased
 print(AHF)
```

```
## # A tibble: 12 x 2
##    purch no_buyers
##    <int>     <int>
##  1     1       924
##  2     2      1071
```

```
##  3     3       181
##  4     4       163
##  5     5       208
##  6     6       245
##  7     7       240
##  8     8       235
##  9     9       301
## 10    10       275
## 11    11       337
## 12    12       342
```

for the maxiumum number of people who bought the AFH book, this is their first or second book

DISCLOSURE: I have used ChatGPT to help interpret some of the R function calls along with error messages that I got while debugging the code in R. I have NOT blindly copy pasted code directly without evaluating it and take full responsibility for any errors or omissions in the code.

I have done some additional calculations along with the assignment questions for better understanding of the data set. I have provided comments to define what the calculation is for.