Using R for Basic Customer Analysis at Bookbinders.

true true

Preliminaries

install.packages('rmarkdown') devtools::install_github('yihui/tinytex')

Load packages:

Read in the data:

```
# use load("bbb.Rdata") for .Rdata files
# dirname(rstudioapi::getActiveDocumentContext()$path)
load("/Users/dain/Programs/R_Projects/MKTG_482_HW1/bbb.Rdata")
```

Assignment answers

1. Report the number and percentage (as a fraction) of customers by gender. Please use `tabyl()` for this calculation.

```
bbb %>% tabyl(gender) %>% adorn_pct_formatting()
```

```
##  gender     n percent
##       F 33302   66.6%
##       M 16698   33.4%
```

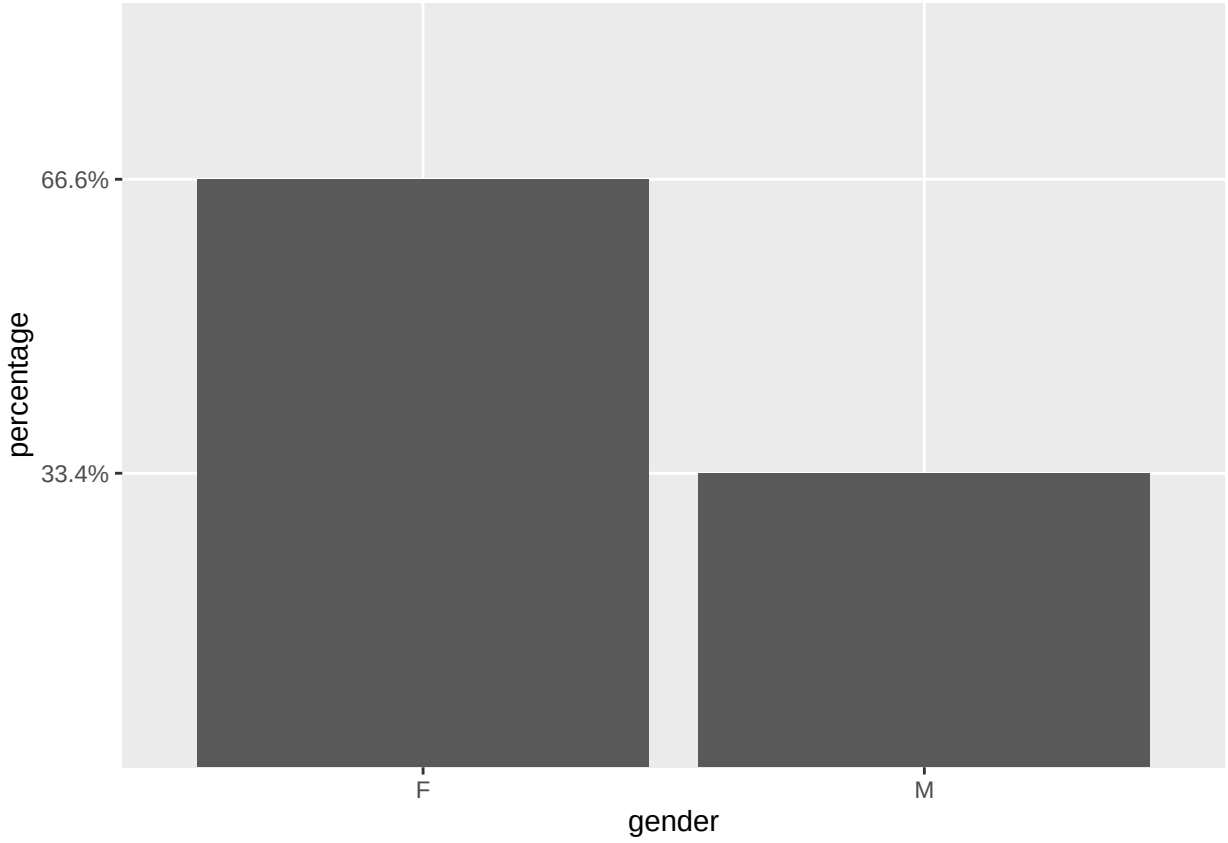
2. Report the number and percentage (as a fraction) of customers by gender. Please use dplyr verbs `group_by` and `summarise` for this calculation. You can use the function `n()` inside summarize to obtain the number of observations. Also, remember that you can do arithmetic when you define summary expressions in `summarise`.

```
gender <- bbb %>% group_by(gender) %>% summarise(n=n()) %>% mutate(percentage=percent(n/sum(n), accuracy
gender
```

```
## # A tibble: 2 x 3
##   gender     n percentage
##   <chr>  <int> <chr>
## 1 F      33302 66.6%
## 2 M      16698 33.4%
```

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

```
gender <- bbb %>% group_by(gender) %>% summarise(n=n()) %>% mutate(percentage=percent(n/sum(n), accuracy
ggplot(gender, aes(x=gender, y=percentage)) + geom_col()
```



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

```
bbb %>% summarise(total_spend=mean(total), num_books=mean(purch), months_since=mean(last))
```

```
## # A tibble: 1 x 3
##   total_spend num_books months_since
##         <dbl>     <dbl>        <dbl>
## 1        208.      3.89         12.4
```

5. Which three states account for the largest number of BookBinders' customers? How many customers are 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.

```
by_state <- bbb %>% group_by(state) %>% summarize(customers=n_distinct(acctnum)) %>% arrange(desc(custo
by_state[1:3,]
```

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

6. What is the average total spending of customers in the three states you just identified (see the "total" variable)?

```
by_state <- bbb %>% group_by(state) %>% summarize(avg_spend=dollar(mean(total))) %>% arrange(desc(avg_sp
by_state[1:3,]
```

```
## # A tibble: 3 x 2
##   state avg_spend
##   <chr> <chr>
## 1 DC    $212.78
## 2 DE    $211.40
## 3 PA    $210.79
```

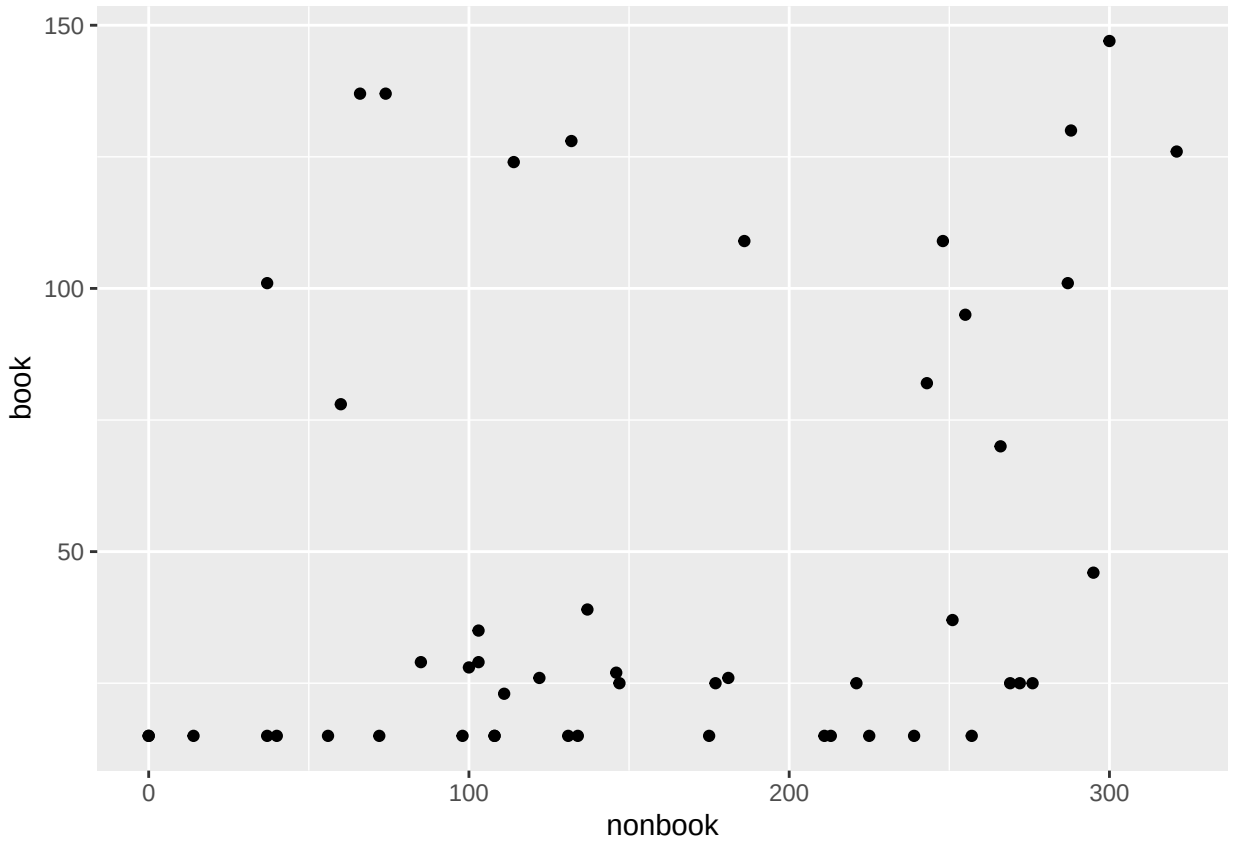
7. Calculate the correlation between customers' total spending on non-book products and books (see the "nonbook" and "book" variables). See my R tutorial for how to calculate correlations.

```
bbb %>% select(book, nonbook) %>% cor()
```

```
##              book   nonbook
## book    1.0000000 0.1574359
## nonbook 0.1574359 1.0000000
```

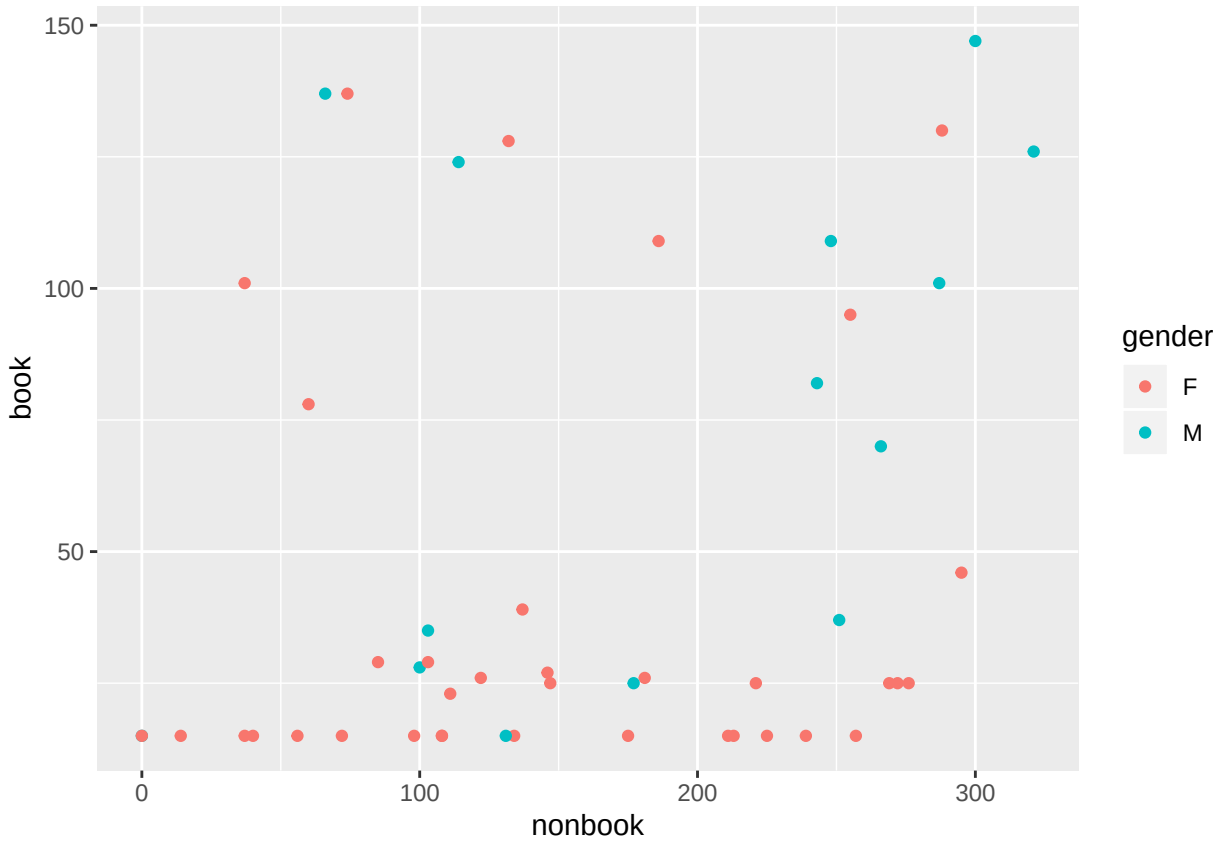
8. For the first 50 customers in the dataset only (use the dplyr `slice` verb) create a scatter plot showing the relationship between customers' total spending on non-book products and books.

```
bbb %>% slice(0:50) %>% ggplot(aes(x=nonbook, y=book)) + geom_point()
```



9. Repeat the previous graph while coloring the points by gender.

```
bbb %>% slice(0:50) %>% ggplot(aes(x=nonbook, y=book, color=gender)) + geom_point()
```



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

```
cols <- c('child', 'youth', 'cook', 'do_it', 'reference', 'art', 'geog')

totals <- bbb %>% select(cols) %>% summarise_all(sum) %>% t() %>% as.data.frame()
colnames(totals) <- c('total')
totals$category <- rownames(totals)
totals %>% arrange(desc(total))
```

```
##   total  category
## 1 46830      cook
## 2 42723     child
## 3 27348      geog
## 4 23153     do_it
## 5 19549     youth
## 6 19296       art
## 7 15612 reference
```

```
# sum(bbb$child)
# sum(bbb$youth)
```

11. For both males and females, find the total number and also the percent who bought "The Art History of Florence" (see the "buyer" variable).

```
florence_by_gender <- bbb %>% filter(buyer==1) %>% group_by(gender) %>% summarise(total=n()) %>% mutate
florence_by_gender
```

```
## # A tibble: 2 x 3
##   gender total percentage
##   <chr>  <int> <chr>
## 1 F       2389 52.8%
## 2 M       2133 47.2%
```

12. For both males and females, determine the total number of purchases and the average number of purchases (see the "purch" variable).

```
purch_by_gender <- bbb %>% group_by(gender) %>% summarise(total_purch=sum(purch), avg_purch=mean(purch)
purch_by_gender
```

```
## # A tibble: 2 x 3
##   gender total_purch avg_purch
##   <chr>        <int>     <dbl>
## 1 F           111968      3.36
## 2 M            82543      4.94
```

13. Determine the minimum, the maximum, and the average number of months between customers' first purchase and their most recent purchase. Use the dplyr verb `mutate` to create a new variable.

```
bbb %>% mutate(months_between=(first - last)) %>% summarise(avg_mos=mean(months_between), min_mos=min(m
```

```
## # A tibble: 1 x 3
##   avg_mos min_mos max_mos
##     <dbl>   <int>   <int>
## 1    13.3       0      72
```

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

```
bbb %>% filter(purch >= 2) %>% summarize(num_bought=sum(buyer==1), total=n(), perc=percent(num_bought /
```

```
## # A tibble: 1 x 3
##   num_bought total perc
##        <int> <int> <chr>
## 1       3598 34880 10.3%
```