Exercise 3 Solution

## Exercise 3

In RStudio, create a new Quarto document and do the following.

1. Load the tidyverse.
2. Import and explore customer\_data using the functions we’ve covered.
3. Provide at least one interesting numeric summary and one interesting visualization using discrete variables only.
4. Practice good coding conventions: Comment often, write in consecutive lines of code using the |>, and use the demonstrated style (e.g., variable names, spacing within functions).
5. Render the Word doc and upload to Canvas.

Five points total, one point each for:

- Loading the tidyverse.

- Providing at least one numeric summary.

- Providing at least one visualization.

- Following good coding conventions (provide feedback on this point).

- Submitting a rendered Word doc.

# Load the tidyverse.  
  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# Import data.  
  
customer\_data <- read\_csv("customer\_data.csv", show\_col\_types = FALSE)  
  
customer\_data

# A tibble: 10,531 × 14  
 customer\_id birth\_year gender income credit married college\_degree region   
 <dbl> <dbl> <chr> <dbl> <dbl> <chr> <chr> <chr>   
 1 1001 1971 Female 73000 742. No No South   
 2 1002 1970 Female 31000 749. Yes No West   
 3 1003 1988 Male 35000 542. No No South   
 4 1004 1984 Other 64000 574. Yes Yes Midwest   
 5 1005 1987 Male 58000 644. No Yes West   
 6 1006 1994 Male 164000 554. Yes Yes Midwest   
 7 1007 1968 Male 39000 608. No No Midwest   
 8 1008 1994 Male 69000 710. No No South   
 9 1009 1958 Male 233000 702. No No West   
10 1010 1994 Female 77000 605. Yes No Northeast  
# ℹ 10,521 more rows  
# ℹ 6 more variables: state <chr>, review\_id <dbl>, star\_rating <dbl>,  
# review\_time <chr>, review\_title <chr>, review\_text <chr>

# At least one interesting numeric summary. For example, the number of "high income" customers by gender.  
  
customer\_data |>   
 mutate(high\_income = income > 120000) |>   
 count(gender, high\_income) |>   
 arrange(desc(n))

# A tibble: 6 × 3  
 gender high\_income n  
 <chr> <lgl> <int>  
1 Female TRUE 2858  
2 Male TRUE 2686  
3 Female FALSE 2361  
4 Male FALSE 1528  
5 Other TRUE 677  
6 Other FALSE 421

# At least one interesting visualization. For example, the top 10 most commonly used words for 5-star reviews.  
  
customer\_data |>  
 filter(star\_rating == 5) |>   
 select(review\_text) |>   
 tidytext::unnest\_tokens(word, review\_text) |>   
 drop\_na(word) |>   
 anti\_join(tidytext::stop\_words) |>   
 count(word) |>   
 arrange(desc(n)) |>   
 slice(1:10) |>   
 mutate(word = fct\_reorder(word, n)) |>  
 ggplot(aes(x = n, y = word)) +  
 geom\_col() +  
 labs(title = "Most Commonly Used Words for 5-Star Reviews")

Joining with `by = join\_by(word)`

