Patagonia: Evaluating the Customer Base

Marc Dotson

## Question 1

**Twenty points total for the question, two points each for:**

* **The question having its own section and heading.**
* **Commenting code.**
* **Connecting to the database.**
* **Importing online\_revenue and online\_transactions (at least) from the database.**
* **Connecting to and importing data from the Qualtrics API.**
* **Importing survey\_data from the Qualtrics API.**

**Four points each for:**

* **Writing the new data locally.**
* **Setting #| eval: false so the code doesn’t run when the report is rendered (no output show show from connecting to the database).**

Let’s load packages and import the data. Note that we are only demonstrating what we have already done to import the data from the database and using the Qualtrics API.

# Load packages.  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
✔ 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

library(dbplyr)

Attaching package: 'dbplyr'  
  
The following objects are masked from 'package:dplyr':  
  
 ident, sql

library(DBI)  
library(qualtRics)  
library(tidytext)

con <- dbConnect(  
 RPostgreSQL::PostgreSQL(),  
 dbname = "analyticsdb",  
 host = "analyticsdb.ccutuqssh92k.us-west-2.rds.amazonaws.com",  
 port = 55432,  
 user = "quantmktg",  
 password = rstudioapi::askForPassword("Database password")  
)  
  
# Import online\_transactions.  
online\_transactions <- tbl(con, "online\_transactions") |>  
 collect() |>   
 select(-row.names)  
  
# Import online\_revenue.  
online\_revenue <- tbl(con, "online\_revenue") |>  
 collect() |>   
 select(-row.names)  
  
# Disconnect from the database.  
dbDisconnect(con)  
  
# Connect to API.  
survey\_data <- fetch\_survey("SV\_0rn3HbR4AEc1BUG")  
  
# Write data locally.  
write\_csv(online\_transactions, here::here("Data", "online\_transactions.csv"))  
write\_csv(online\_revenue, here::here("Data", "online\_revenue.csv"))  
write\_csv(survey\_data, here::here("Data", "survey\_data.csv"))

Now let’s import what we’ve written locally so we have data to work with when we render.

# Import data.  
customer\_data <- read\_csv(here::here("Data", "customer\_data.csv"))

Rows: 10531 Columns: 13  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (8): gender, married, college\_degree, region, state, review\_time, review...  
dbl (5): customer\_id, birth\_year, income, credit, star\_rating  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

store\_transactions <- read\_csv(here::here("Data", "store\_transactions.csv"))

Rows: 10531 Columns: 169  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
dbl (169): customer\_id, jan\_2005, feb\_2005, mar\_2005, apr\_2005, may\_2005, ju...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

store\_revenue <- read\_csv(here::here("Data", "store\_revenue.csv"))

Rows: 10531 Columns: 169  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
dbl (169): customer\_id, jan\_2005, feb\_2005, mar\_2005, apr\_2005, may\_2005, ju...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

online\_transactions <- read\_csv(here::here("Data", "online\_transactions.csv"))

Rows: 10531 Columns: 169  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
dbl (169): customer\_id, jan\_2005, feb\_2005, mar\_2005, apr\_2005, may\_2005, ju...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

online\_revenue <- read\_csv(here::here("Data", "online\_revenue.csv"))

Rows: 10531 Columns: 169  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
dbl (169): customer\_id, jan\_2005, feb\_2005, mar\_2005, apr\_2005, may\_2005, ju...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

survey\_data <- read\_csv(here::here("Data", "survey\_data.csv"))

Rows: 568 Columns: 21  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (5): Status, ResponseId, sat\_overall, lik\_recommend, review\_text  
dbl (1): customer\_id  
lgl (12): IPAddress, Progress, Duration (in seconds), Finished, RecipientLa...  
dttm (3): StartDate, EndDate, RecordedDate  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Question 2

**Twenty points total, two points each for:**

* **The question having its own section and heading.**
* **Commenting code.**

**One point each for:**

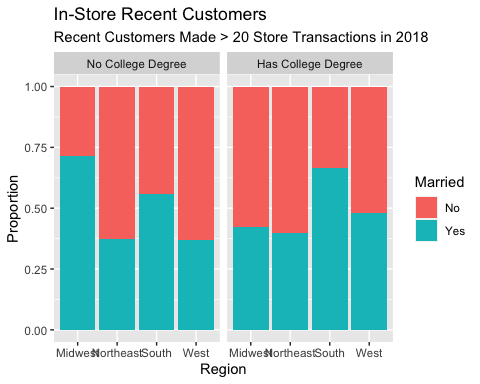
* **Using pivot\_longer() to tidy the 2018 in-store data.**
* **Using a grouped summary to summarize in-store transactions data.**
* **Filtering in-store customers based on the given definition of > 20 transactions in 2018.**
* **In a single plot appropriate for the data type, including the relationship between region, married, and college\_degree for in-store recent customers.**
* **Including appropriate labels and axes titles for the plot of in-store recent customers.**
* **Using pivot\_longer() to tidy the 2018 online data.**
* **Using a grouped summary to summarize online transactions data.**
* **Filtering online customers based on the given definition of > 20 transactions in 2018.**
* **In a single plot appropriate for the data type, including the relationship between region, married, and college\_degree for online recent customers.**
* **Including appropriate labels and axes titles for the plot of online recent customers.**

**Three points each for:**

* **Answering the question about what the comparison between the two plots suggests about Patagonia’s customer base and how it differs in-store vs. online.**
* **Answers written in complete sentences and appropriately placed with respect to data and visualizations.**

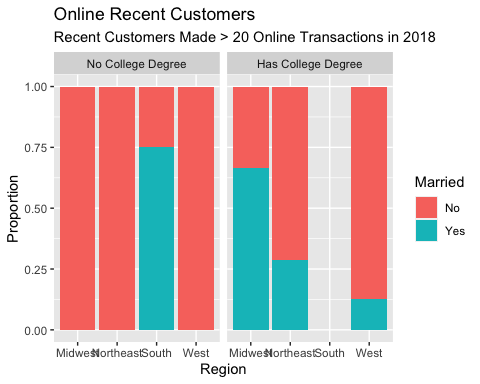
Let’s find the “recent customers” (i.e., anyone who has made more than 20 transactions in 2018) and plot some of their demographic data. Let’s start with the store transactions data first.

# Find "recent customers" based on store transactions.  
store\_recent\_customers <- store\_transactions |>   
 select(customer\_id, contains("2018")) |>   
 pivot\_longer(  
 -customer\_id,  
 names\_to = "month\_year",  
 values\_to = "store\_transactions"  
 ) |>   
 group\_by(customer\_id) |>   
 summarize(total\_store\_transactions = sum(store\_transactions)) |>   
 filter(total\_store\_transactions > 20)  
  
# Visualize the composition of these in-store "recent customers."  
store\_recent\_customers |>   
 left\_join(customer\_data, join\_by(customer\_id)) |>   
 mutate(  
 college\_degree = fct\_recode(  
 college\_degree, "No College Degree" = "No", "Has College Degree" = "Yes"  
 )  
 ) |>   
 ggplot(aes(x = region, fill = married)) +  
 geom\_bar(position = "fill") +  
 facet\_grid(~ college\_degree) +  
 labs(  
 title = "In-Store Recent Customers",  
 subtitle = "Recent Customers Made > 20 Store Transactions in 2018",  
 x = "Region",  
 y = "Proportion"  
 ) +  
 scale\_fill\_discrete(name = "Married")



The biggest proportional slice of in-store recent customers are married Midwesterners without a college degree. Let’s compare this with the “recent customers” using online transactions data.

# Find "recent customers" based on online transactions.  
online\_recent\_customers <- online\_transactions |>   
 select(customer\_id, contains("2018")) |>   
 pivot\_longer(  
 -customer\_id,  
 names\_to = "month\_year",  
 values\_to = "online\_transactions"  
 ) |>   
 group\_by(customer\_id) |>   
 summarize(total\_online\_transactions = sum(online\_transactions)) |>   
 filter(total\_online\_transactions > 20)  
  
# Visualize the composition of these online "recent customers."  
online\_recent\_customers |>   
 left\_join(customer\_data, join\_by(customer\_id)) |>   
 mutate(  
 college\_degree = fct\_recode(  
 college\_degree, "No College Degree" = "No", "Has College Degree" = "Yes"  
 )  
 ) |>   
 ggplot(aes(x = region, fill = married)) +  
 geom\_bar(position = "fill") +  
 facet\_grid(~ college\_degree) +  
 labs(  
 title = "Online Recent Customers",  
 subtitle = "Recent Customers Made > 20 Online Transactions in 2018",  
 x = "Region",  
 y = "Proportion"  
 ) +  
 scale\_fill\_discrete(name = "Married")



Unlike the in-store recent customers, a far greater proportion of online recent customers aren’t married. Additionally there aren’t any online recent customers from the South who have a college degree.

This comparison suggests the following. First, there appears to be a different set of customers who can be defined as “recent customers” in-store vs. online. Second, while in-store recent customers appear to be evenly split between being married and having a college degree, online recent customers are decidedly more unmarried with fewer overall with a college degree.

## Question 3

**Twenty points total, two points each for:**

* **The question having its own section and heading.**
* **Commenting code.**

**One point each for:**

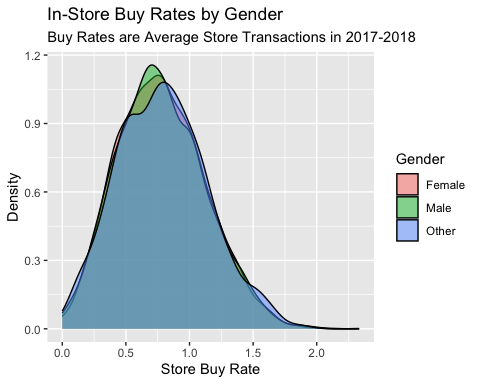
* **Using pivot\_longer() to tidy the 2017-2018 in-store data.**
* **Using a grouped summary to summarize in-store transactions data.**
* **In a single plot appropriate for the data type, looking at the relationship between in-store buy rate and gender.**
* **Including appropriate labels and axes titles for the plot of in-store buy rate.**
* **Using pivot\_longer() to tidy the 2017-2018 online data.**
* **Using a grouped summary to summarize online transactions data.**
* **In a single plot appropriate for the data type, looking at the relationship between online buy rate and gender.**
* **Including appropriate labels and axes titles for the plot of online buy rate.**

**Four points each for:**

* **Answering the question about what the comparison between the two plots suggests about Patagonia’s customer base.**
* **Answers written in complete sentences and appropriately placed with respect to data and visualizations.**

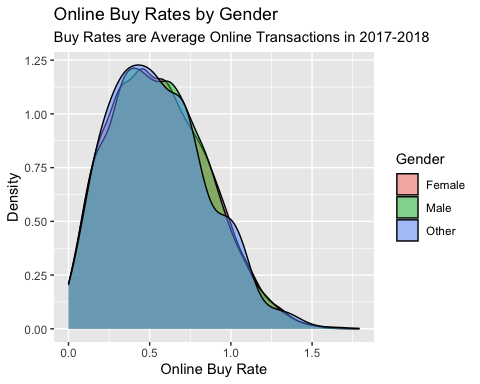
Let’s first compute the customer in-store buy rate for 2017-2018 and visualize differences by gender.

# Compute the customer in-store buy rate for 2017 and 2018.  
store\_buy\_rates <- store\_transactions |>   
 select(customer\_id, contains("2017"), contains("2018")) |>   
 pivot\_longer(  
 -customer\_id,  
 names\_to = "month\_year",  
 values\_to = "store\_transactions"  
 ) |>   
 group\_by(customer\_id) |>   
 summarize(store\_buy\_rate = mean(store\_transactions))  
  
# Visualize the in-store buy rate by gender.  
store\_buy\_rates |>   
 left\_join(customer\_data, join\_by(customer\_id)) |>   
 ggplot(aes(x = store\_buy\_rate, fill = gender)) +  
 geom\_density(alpha = 0.5) +  
 labs(  
 title = "In-Store Buy Rates by Gender",  
 subtitle = "Buy Rates are Average Store Transactions in 2017-2018",  
 x = "Store Buy Rate",  
 y = "Density"  
 ) +  
 scale\_fill\_discrete(name = "Gender")



There doesn’t appear to be much difference between genders for the specified in-store buy rate, with an average buy rate of 0.75. Let’s compare this with the customer online buy rate for 2017-2018.

# Compute the customer online buy rate for 2017 and 2018.  
online\_buy\_rates <- online\_transactions |>   
 select(customer\_id, contains("2017"), contains("2018")) |>   
 pivot\_longer(  
 -customer\_id,  
 names\_to = "month\_year",  
 values\_to = "online\_transactions"  
 ) |>   
 group\_by(customer\_id) |>   
 summarize(online\_buy\_rate = mean(online\_transactions))  
  
# Visualize the in-store buy rate by gender.  
online\_buy\_rates |>   
 left\_join(customer\_data, join\_by(customer\_id)) |>   
 ggplot(aes(x = online\_buy\_rate, fill = gender)) +  
 geom\_density(alpha = 0.5) +  
 labs(  
 title = "Online Buy Rates by Gender",  
 subtitle = "Buy Rates are Average Online Transactions in 2017-2018",  
 x = "Online Buy Rate",  
 y = "Density"  
 ) +  
 scale\_fill\_discrete(name = "Gender")



Similarly, there doesn’t appear to be much difference between genders for the specified online buy rate, with an average buy rate of 0.5. While there doesn’t appear to be difference between genders for either the in-store or online buy rates, this comparison does suggest that Patagonia’s customer base does buy more on average in-store than online.

## Question 4

**Twenty points total, two points each for:**

* **The question having its own section and heading.**
* **Commenting code.**

**One point each for:**

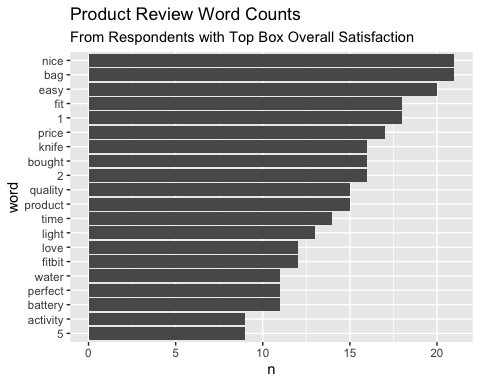
* **Creating the top\_box variable using if\_else().**
* **Using the top box variable to filter().**
* **Using unnest\_tokens().**
* **Using anti\_join() to remove stop words.**
* **Using count() to create review word counts.**
* **Using slice() or slice\_head() to keep the most frequently used words.**
* **Using fct\_reorder() to make the plot easier to read.**
* **Including appropriate labels and axes titles for the plot of review word counts.**

**Four points each for:**

* **Answering what the review word counts suggest about Patagonia’s customer base.**
* **Answers written in complete sentences and appropriately placed with respect to data and visualizations.**

Let’s create the new “top box” overall satisfaction variable and compare review word count plots for the “top box” vs. “not top box” groups.

# Create a top\_box variable from overall satisfaction.  
survey\_data <- survey\_data |>   
 mutate(top\_box = if\_else(sat\_overall == "Very satisfied", "Very satisfied", "Not very satisfied"))  
  
# Plot review word counts where top\_box == "Very satisfied".  
survey\_data |>   
 filter(top\_box == "Very satisfied") |>   
 unnest\_tokens(word, review\_text) |>   
 anti\_join(stop\_words, join\_by(word)) |>   
 count(word) |>   
 arrange(desc(n)) |>   
 slice(1:20) |>   
 mutate(word = fct\_reorder(word, n)) |>  
 ggplot(aes(x = n, y = word)) +  
 geom\_col() +  
 labs(  
 title = "Product Review Word Counts",  
 subtitle = "From Respondents with Top Box Overall Satisfaction"  
 )



The respondents in the top-box overall satisfaction group use nice, bag, and easy frequently, along with fit, love, and quality, suggesting there may be specific attitudes or products that are responsible for the top-box overall satisfaction.

## Question 5

**Twenty points total, one point each for:**

* **The question having its own section and heading.**
* **Commenting code.**
* **Using pivot\_longer() to tidy the in-store revenue data.**
* **Using a grouped summary to summarize in-store revenue data.**
* **Using pivot\_longer() to tidy the online revenue data.**
* **Using a grouped summary to summarize online revenue data.**
* **Using left\_join() to join all three datasets.**
* **Creating the overall revenue variable by summing the averages.**
* **Visualizing the relationship between overall revenue and age in the dashboard.**
* **Including appropriate labels and axes titles for the dashboard plot.**
* **Including a filter for gender.**
* **Including a filter for marital status.**
* **Including a filter for college degree status.**
* **Including a filter for the new high income variable.**
* **Including a call-out box for average in-store revenue.**
* **Including a call-out box for average online revenue.**
* **Including a call-out box for average overall revenue.**
* **Answering what they have learned about Patagonia’s customer base using the dashboard.**
* **Recreating one of the plots in the report, with appropriate labels and axes titles.**
* **Including a link to the published dashboard.**
* **Answers written in complete sentences and appropriately placed with respect to data and visualizations.**

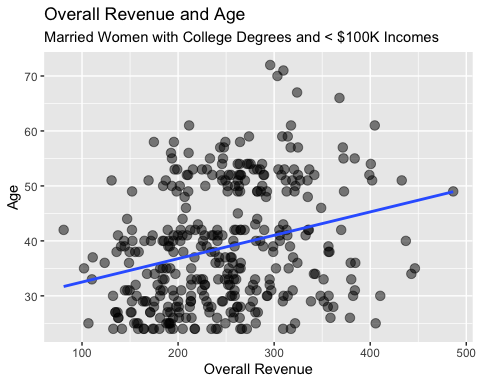
Let’s compute the average revenue for each customer across in-store and online channels, as well as the new “high income” and age variables, and then write the data for use in the dashboard.

# Compute the average in-store revenue for each customer.  
store\_average\_revenue <- store\_revenue |>   
 pivot\_longer(  
 -customer\_id,  
 names\_to = "month\_year",  
 values\_to = "store\_revenue"  
 ) |>   
 group\_by(customer\_id) |>   
 summarize(avg\_store\_revenue = mean(store\_revenue))  
  
# Compute the average online revenue for each customer.  
online\_average\_revenue <- online\_revenue |>   
 pivot\_longer(  
 -customer\_id,  
 names\_to = "month\_year",  
 values\_to = "online\_revenue"  
 ) |>   
 group\_by(customer\_id) |>   
 summarize(avg\_online\_revenue = mean(online\_revenue))  
  
# Create the overall revenue for each customer, high income, and age variables.  
overall\_revenue <- store\_average\_revenue |>   
 left\_join(online\_average\_revenue, join\_by(customer\_id)) |>   
 left\_join(customer\_data, join\_by(customer\_id)) |>   
 mutate(  
 overall\_revenue = avg\_online\_revenue + avg\_store\_revenue,  
 high\_income = if\_else(income > 100000, "High income", "Not high income"),  
 age = 2023 - birth\_year  
 ) |>   
 select(  
 customer\_id, overall\_revenue, avg\_store\_revenue, avg\_online\_revenue,   
 age, gender, married, college\_degree, high\_income  
 )  
  
# Write the data locally.  
write\_csv(overall\_revenue, here::here("Projects", "Project 01 Patagonia CRM", "overall\_revenue.csv"))

The dashboard makes exploring possible combinations of filtered variables easy. We can easily see how differences in gender, marital status, college degree, and high income effects the average and overall revenue as well as the relationship between age and overall revenue. For example, it’s clear that women spend the most. One plot of interest to me was the following.

# One plot of interest recreated from the dashboard.  
overall\_revenue |>   
 filter(  
 gender == "Female",   
 married == "Yes",   
 college\_degree == "Yes",   
 high\_income == "Not high income"  
 ) |>  
 ggplot(aes(x = overall\_revenue, y = age)) +  
 geom\_point(size = 3, alpha = 0.5) +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(  
 title = "Overall Revenue and Age",  
 subtitle = "Married Women with College Degrees and < $100K Incomes",  
 x = "Overall Revenue",  
 y = "Age"  
 )

`geom\_smooth()` using formula = 'y ~ x'



While this combination of filters doesn’t produce the largest average overall revenue, it does uncover the most correlation between age and overall revenue. In other words, more than any other group we can define in the dashboard, married women with college degrees and < $100K incomes spend more as they get older.

Explore the dashboard on your own [here](https://marcdotson.shinyapps.io/project_01/).