Shopping Data Analysis Report

STATS 220 Semester One 2025

Krishanu Choudhury

Introduction

My data collection focused on everyday transactions, which I observed in places like supermarkets, cafes and restaurants, public transport, retail stores, and online shopping platforms.

When I designed my Google Form, I made sure to follow the recommended guidelines, especially Guideline 4, which stresses the use of validation rules. I also added a few new questions to capture more detail, such as whether I knew the person making the transaction, whether they took a receipt, and whether they used a self checkout or interacted with a cashier. These questions helped me explore interesting patterns, like whether people I know tend to prefer self checkouts, or how often people actually take their receipts.

All the data was collected through my own observations. Whenever I noticed someone, whether a stranger or someone I know making a transaction in a common place, I recorded the details as part of the study. Hence, the data collection process involved only me.

Firstly, I created a bar chart that displays the number of transactions for each payment method category: Cards, Cash, Mobile Payments, and Gift Vouchers by each transaction place: Supermarkets, Cafe/Restaurants, Online Shopping, Retail Stores and Public Transport. I was particularly interested in understanding which payment methods are most commonly used in everyday situations and is there a difference in the preferred mode of payment for different transaction places. This insight can be valuable for business owners and shop managers, as it may inform decisions about whether to invest in additional EFTPOS terminals or support alternative payment options. I chose a bar chart for this analysis because it offers a clear and intuitive way to compare counts across categories.

Secondly, I developed a line chart to show the number of transactions occurring at each hour of the day. To do this, I extracted the hour component from the Timestamp variable using the hour() function from the {lubridate} package. The goal was to identify peak transaction times, which could help businesses better understand customer flow and adjust staffing or services accordingly. I opted for a line chart because it is well-suited for time-based data and provides a smoother visual representation of patterns across hours. I initially tried a bar chart, but it became cluttered due to the number of hourly intervals on the x-axis.

Thirdly, I chose to create a type of chart that was not covered in our lectures, a pie chart, to represent the proportions of different transaction amount levels at each transaction location: supermarkets, cafes and restaurants, online shopping platforms, public transport, and retail stores. The goal was to illustrate whether people tend to spend at low (less than 20 NZD), medium (between 20 NZD and 100 NZD), high (between 100 NZD and 200 NZD), or very high (more than 200 NZD) levels in each of these places. This type of visualization can provide customers with a better sense of how much money they might typically need when making purchases at these locations. I selected a pie chart because it effectively shows proportions, which was central to the insight I wanted to convey. I initially considered using a bar chart or boxplot, but I found that those options made the data appear more complex and harder to interpret for this particular purpose. To make the presentation even clearer, I applied the facet_wrap() function to generate a separate pie chart for each transaction place, allowing for easier comparison and a cleaner visual layout.

Lastly, I chose to create another bar chart that displays the number of transactions completed through self checkouts versus cashiers, specifically for individuals I am familiar with. This visualization has a slightly personal motivation, I was curious to see whether people I know tend to prefer self checkouts over interacting with a cashier. The idea was that, in future shopping trips with them, I could recommend stores that offer self checkout facilities. I selected a bar chart for this analysis because it is a clear and effective way to represent counts across a small number of categories

Visual data story

Here is the link to my visual data story:

https://krishanu-choudhury.github.io/stats220/visual_data_story.html

My visual data story showcases creativity in multiple ways. To begin with, I adopted a minimalist and standard colour theme that remains consistent throughout both my plots and the visual styling of my report. Instead of the required three visualizations, I chose to include four, each designed to clearly communicate the key goals of this study. I ensured that my plots are well organised, with clearly defined labels and captions to enhance readability.

As an additional creative step, I incorporated a pie chart to visualize part of the data, despite this not being explicitly covered in the course lectures. I used the facet_wrap() function to generate a separate pie chart for each category, which helped highlight differences across contexts more effectively. Furthermore, I included a custom header image to reflect the central theme of the study, as well as a light-hearted end note featuring a "thank you" message accompanied by a cat GIF.

Overall, I believe the most creative aspect of my work was the way I reused and adapted code within the given guidelines to build visualizations that are not only appropriate, but also purposeful. Each plot was designed with the intent to offer meaningful statistical insights to readers, including business owners, customers, and myself.

Learning reflection

This project has helped me develop a wide range of new skills. I gained experience in manipulating data frames using various functions such as mutate(), filter(), and str_detect() to better align the data with the goals of my analysis. I also expanded my understanding of data visualization by exploring chart types beyond just bar charts, all using the {ggplot2} package. Through this process, I came to understand that the choice of chart depends heavily on the specific question we are trying to answer.

Looking ahead, I am eager to explore additional methods of visualizing data, develop more advanced data manipulation techniques, and improve my ability to choose the most appropriate plot types for different analytical goals. I am also interested in learning how to conduct hypothesis testing to support insights with statistical evidence, rather than relying solely on observational findings

Appendix

```
library(tidyverse)

# Reading the data
URL <- "https://docs.google.com/spreadsheets/d/e/2PACX-1vRZt32kU5Rgliiirpj4SxYxFhOmcQE1g1vJgZg8_deqjQLp
logged_data <- read_csv(URL)

renamed_data <- logged_data %>% # Renaming the variables in my data frame
  rename(transaction_place = 2, transaction_amount = 3, transaction_method = 4, transaction_time = 5, t

# Creating my colour palette and theme
```

```r
my_colours <- c("#003153", "#f0f8ff", "#807dba", "#ffffff", "#333333","#666666","#e0e0e0","#f5a9f5","#a
my_theme <- theme_minimal(base_family = "Helvetica") +
  theme(
    plot.background = element_rect(fill = my_colours[4], color = NA),
    panel.background = element_rect(fill = my_colours[7]),
    panel.grid = element_blank(),
    plot.title = element_text(face = "bold", size = 13, color = my_colours[5], hjust = 0.5),
    axis.title = element_text(size = 13, color = my_colours[5]),
    axis.text.y = element_text(size = 7, color = my_colours[5]),
    axis.text.x = element_blank(),
    axis.ticks = element_blank(),
    axis.line = element_blank(),
    plot.caption = element_text(size = 7, color = my_colours[6], hjust = 1),
    plot.margin = margin(1,1,1,1, "cm"),
    legend.title = element_blank()
  )

# First Plot
# Bar chart for most used transaction method for each place
# Manipulating the data frame using group_by(), summarise(), and arrange() to get the number of transac
transaction_method_data <- renamed_data %>%
  group_by(transaction_place, transaction_method) %>% # grouping our transactions for each transaction_
  summarise(count = n()) %>% # getting the total number of transactions done using each level of transa
  arrange(transaction_place, desc(count)) # arranging my counts to descending orderso that the highest

transaction_method_plot <- transaction_method_data %>%
  ggplot() +
  geom_col(aes(x = reorder(transaction_place, -count), y = count, fill = transaction_method)) + # Choos
  labs(
    title = "Which Payment Method was most frequently used?",
    x = "Transaction Places",
    y = "Total Transactions",
    caption = "Source: STATS 220 Project 4 Survey: Observing Transactions"
  ) +
  my_theme + # applying my pre-made theme
  scale_fill_manual(values = c(my_colours[13], my_colours[12], my_colours[11], my_colours[10])) + # app
  theme(axis.text.x = element_text(size = 7, color = my_colours[5], angle = 45))

# Second Plot
# Line Chart for number of transactions by hour
# Manipulating the timestamp variable from our data frame to get the hours, additional functions used a

transaction_hour_data <- renamed_data %>%
  mutate(transaction_hour = hour(dmy_hms(Timestamp))) %>% # creating a new variable to store the hours
  group_by(transaction_hour) %>% # grouping the transactions by each hour of the day
  summarise (count = n()) # getting the total transactions

transaction_hour_plot <- transaction_hour_data %>%
  ggplot() +
  geom_line(aes(x = transaction_hour, y = count), color = my_colours[13], linewidth = 1) + # choosing a
  labs(
    title = "Which hour of the day do most transactions occur?",
    x = "Hour of Day",
    y = "Total Transactions",
    caption = "Source: STATS 220 Project 4 Survey: Observing Transactions"
  ) +
  scale_x_continuous(breaks = 0:23) + # Adjusting the x axis to include all hours of the day
  my_theme + # applying my theme plus additional themes for a better looking plot
  theme(axis.text.x = element_text(size = 7, color = my_colours[5])) +
  scale_y_continuous(breaks = 0:10)

# Third Plot
#Pie Chart of transaction_amount by transaction_place
# Manipulating the data frame using select(), mutate(), case_when

# I want to divide my transaction amount to classes low, medium, high, very high, so I create a new var
transaction_amount_data <- renamed_data %>%
  mutate(transaction_amount_class = case_when( # creating the new variable to classify the level of amo
    transaction_amount<20 ~ "Low",
    transaction_amount>=20 & transaction_amount<100 ~ "Medium",
    transaction_amount>=100 & transaction_amount<200 ~ "High",
    transaction_amount>=200 ~ "Very High"
  )) %>%
  select(transaction_place, transaction_amount, transaction_amount_class) # selecting only the variable

pie_data <- transaction_amount_data %>% #new data frame to store the proportions of each class of trans
  count(transaction_place, transaction_amount_class) %>%
  group_by(transaction_place) %>%
  mutate(Proportion = n/sum(n)) # Getting the proportions of each class of payment amount by each level

transaction_amount_plot <- pie_data %>%
  ggplot() +
  geom_col(aes(x = "",y = Proportion, fill = transaction_amount_class)) +
  coord_polar(theta = "y") + # choosing a pie chart to better plot proportions
  facet_wrap(~ transaction_place) + # creating a different pie chart for each transaction place
  scale_fill_manual(values = c(
    "Low" = my_colours[10],
    "Medium" = my_colours[11],
    "High" = my_colours[12],
    "Very High" = my_colours[13]
  )) +
  labs(
    title = "What amount of money do people spend in different places?",
    fill = "Transaction Amount Level",
    x = "Transaction Places",
    y = "Proportion of each level of transaction amount",
    caption = "Source: STATS 220 Project 4 Survey: Observing Transactions") +
  my_theme +
  theme(
    strip.text = element_text(face = "bold", size = 6),
    axis.title = element_text(size = 13),
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    legend.position = "right")

# Fourth Plot
# Bar chart to see if people I know usually do self checkout or with a cashier
# Data manipulation using str_detect(), slice()

new_data <- renamed_data %>%
  slice(61:81) %>% #As the variables transaction_person and transaction_checkout refers to the two new
  mutate(familiar_person = str_detect(transaction_person, "I know")) %>% # Create a new logical variabl
  select(familiar_person, transaction_checkout)

transaction_person_plot <- new_data %>%
  filter(familiar_person == TRUE) %>% # filtering the data for only the persons that I know
  count(transaction_checkout) %>%
  ggplot() +
  geom_col(aes(x = transaction_checkout, y = n, fill = transaction_checkout)) +
  labs(
    title = "Whether people that I know usually prefer self checkout or with a cashier?",
    x = "Checkout Method",
    y = "Total Transactions",
    caption = "Source: STATS 220 Project 4 Survey: Observing Transactions"
  ) +
  scale_fill_manual(values = c(my_colours[11], my_colours[13])) +
  my_theme

# Saving my four plots
ggsave("plot1.png", plot = transaction_method_plot, width = 8, height = 4, units = "in")
ggsave("plot2.png", plot = transaction_hour_plot, width = 8, height = 4, units = "in")
ggsave("plot3.png", plot = transaction_amount_plot, width = 8, height = 4, units = "in")
ggsave("plot4.png", plot = transaction_person_plot, width = 8, height = 4, units = "in")
```