Consumer Data Analysis Report

STATS 220 Semester One 2025

Krishanu Choudhury

Introduction

For my reference group of schools, I have chosen to focus on those that meet specific criteria: they must be English medium schools, located in the Auckland region, classified as Secondary (Year 9–15), situated in major urban areas, and eligible to receive donations. I selected Mt Albert Grammar School as my chosen high school. Since I did not attend high school in New Zealand, I decided to pick one from my current suburb, Mount Albert.

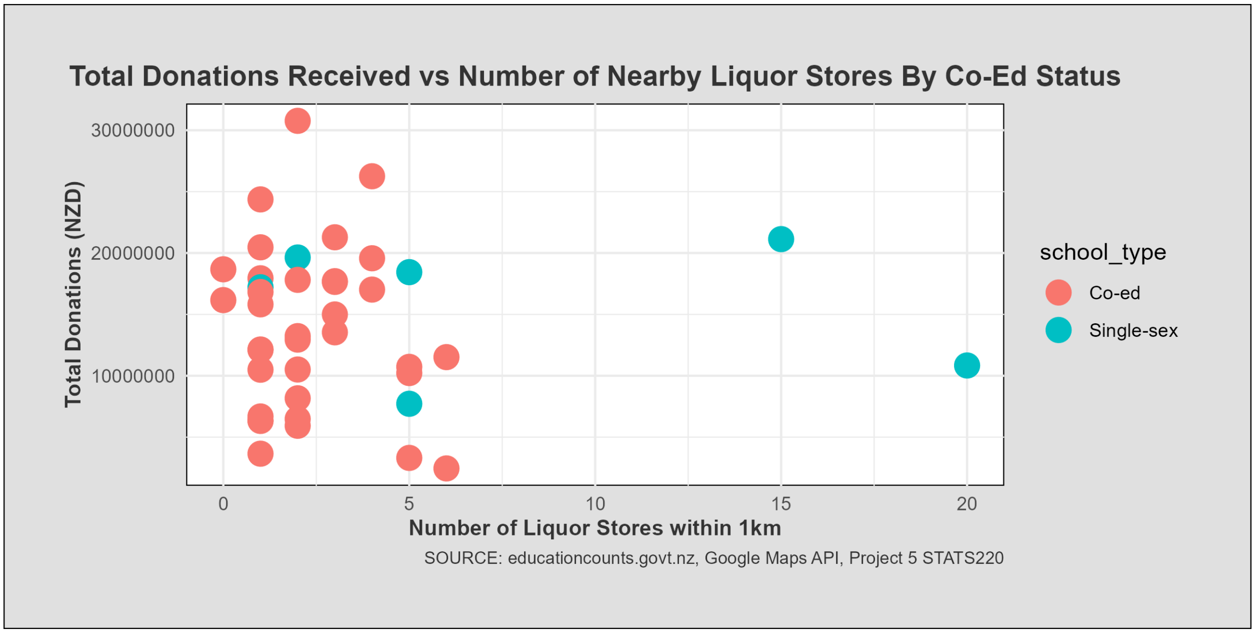
My motivation for selecting English medium schools comes from my background as an international student, making this criterion personally relevant. I have narrowed the scope to Secondary schools (Year 9-15) in major urban areas of Auckland because I suspect that liquor stores are more likely to be located near secondary schools rather than primary schools in these areas. I also restricted my selection to schools eligible for donations, as my analysis focuses on visualizing total donations and examining whether presence of liquor stores (within 1km) influences donation amounts.

Regarding data collection from Mt Albert Grammar School's website, I reviewed its robots.txt file and terms of use, and found no restrictions against web scraping. I ensured the scraping process avoided any sensitive or personal data, aligning with responsible data practices.

Visualisation

The primary goal of my visualisation is to explore the total donations received by each of the selected reference schools and assess whether this amount is influenced by the presence of nearby liquor stores. Additionally, I aim to examine whether there are notable differences in donation amounts between co-educational and single-sex schools.

Below is my scatter plot showing the Total Donations Vs Number of Nearby Liquor Stores for Co-Ed and Single-sex schools:



Overall, for English medium secondary schools (Years 9–15) located in major urban areas of the Auckland region and eligible for donations:

- Co-educational schools with five or more liquor stores within 1 km radius tend to receive lower donation amounts compared to those with fewer nearby liquor stores, whereas this trend is less noticeable among Single-sex schools.
- Single-sex schools generally have a higher number of liquor stores within 1 km compared to Co-Ed schools.
- Having more than five liquor stores within 1 km is uncommon for both Co-Ed and single-sex schools.

Creativity

My visualization showcases creativity in several ways. Firstly, I extended the original code that extracted only the School Operations financial data (Part D) to also include Teacher Salaries and Property Funding. I then used mutate() to create a new variable called Total Donations, summing these three components to represent the overall donation amount received by each reference school.

Additionally, instead of combining only one external data set with my reference schools (as suggested), I merged both the financial data and nearby liquor store data with my reference data, as this provided a more complete picture for my visualization goal. I chose to use left_join() rather than inner_join() to ensure that schools with no liquor stores nearby were still included in the analysis.

To enhance readability, I also reclassified the existing Co-Ed status column by grouping Boys Schools and Girls Schools under a new category called "Single-Sex", aligning this change with the focus of my analysis.

In summary, my approach demonstrates creativity through extending existing code to extract additional information, generating new variables tailored to my goals, and designing a clean, readable plot with a minimalist color palette and informative labels and titles.

Learning reflection

In Module 5, I gained valuable knowledge of web scraping techniques and ethical data sourcing practices. One of the main lessons was learning how to collect data from digital platforms while ensuring that this process complies with the website's terms and conditions, which is a critical aspect of responsible data analysis. I also developed a better understanding of how to merge different data sets to derive meaningful insights, and how the choice between various types of joins such as inner or left joins should be guided by the specific goals of the analysis. Additionally, I strengthened my skills in data manipulation through the use of powerful R packages such as {stringr} for text processing and {dplyr} for data wrangling. These tools have helped me clean, transform, and analyse data more effectively.

Looking ahead, I am eager to deepen my understanding of web scraping beyond the introductory level, exploring more advanced techniques and automation. I am also interested in improving my ability to handle complex data structures like JSON files, which I believe is essential for progressing in a data science career.

Self review

Reflecting on all five projects, I believe the most valuable skill I have gained is the ability to communicate data insights effectively through reproducible reporting using R Markdown. This involved structuring my reports with clear section headers, using bullet points to emphasize key findings, and embedding visualizations directly into the document to support my analysis. R Markdown has helped me present my work in a professional and organised way.

Another essential skill I developed was programming in R, especially in the areas of data importing, wrangling, and visualization for informed decision making. I also learned how to clean and restructure data to follow the principles of tidy data, which greatly improved the clarity and consistency of my analyses.

Overall, I thoroughly enjoyed this course, particularly the project based learning approach. It gave me the opportunity to apply theoretical concepts to practical problems while strengthening both my technical and communication skills. I feel more prepared for a future career in data science as a result.

Appendix

```r
library(tidyverse)
library(jsonlite)
library(rvest)

# Reading the schools directory csv file
directory_data <- read_csv("schools_directory.csv") %>%
  janitor::clean_names()

# Since I did not attend high school in NZ, I will be using a school located in my suburb Mount Albert
my_school <- directory_data %>%
  filter(add1_suburb == "Mount Albert" | add2_suburb == "Mount Albert") %>% # Only schools in Mount Alb
  filter(co_ed_status == "Co-Educational") %>% # Only Co-ed schools
  filter(org_type == "Secondary (Year 9-15)") # Only secondary (year9-15) schools

## My chosen high school is Mt Albert Grammar School

# Getting the url to the website of my school
my_school_url <- my_school$url
my_school_url

# Checking if web scarping is allowed or not.
# Mt Albert Grammar School has a robots.txt file which I checked using "https://www.mags.school.nz/robo
# There were no Terms & Conditions page or similar to suggest that web scraping is not allowed.
# Overall, all this summarize that I am allowed to scrape data from the Mt Albert Grammar School websit
```

```r
school_id <- my_school$school_id

page_url <- paste0("https://www.educationcounts.govt.nz/find-school/school/financial-performance?distri

html <- read_html(page_url) %>%
  html_element("table")

if(length(html) > 0){

  scraped_data <- html %>%
    html_table()

  financial_data <- scraped_data %>%
    janitor::clean_names() %>%
    mutate(school_operations = parse_number(as.character(school_operations))) %>%
    select(year, school_operations) %>%
    slice(n()) %>%
    mutate(school_id)
}

# Creating my reference set of schools
reference_schools <- directory_data %>%
  filter(language_of_instruction == "All students taught in English") %>% # Only English taught schools
  filter(regional_council == "Auckland Region") %>% # Only schools in Auckland region
  filter(org_type == "Secondary (Year 9-15)") %>% # Only Secondary (Year 9-15) schools
  filter(urban_rural_indicator == "Major urban area") %>% # Only schools in Major urban areas
  filter(school_donations != "Not applicable") %>% # Schools where donations are applicable
  drop_na(url)%>%
  select(school_id, org_name, url, latitude, longitude, school_donations, boarding_facilities, co_ed_sta

# Saving my reference schools data
write_csv(reference_schools, "reference_schools.csv")

# Scraping financial info about my reference schools
school_ids <- reference_schools$school_id

get_finance <- function(school_id){

  page_url <- paste0("https://www.educationcounts.govt.nz/find-school/school/financial-performance?dist

  Sys.sleep(2)

  html <- read_html(page_url) %>%
    html_element("table")

  if(length(html) > 0){
    scraped_data <- html %>%
      html_table()

    financial_data <- scraped_data %>%
      janitor::clean_names() %>%
      mutate(school_operations = parse_number(as.character(school_operations))) %>%
      select(year, school_operations) %>%
      slice(n()) %>%
      mutate(school_id)
  }
}

school_financial_data <- map_df(school_ids, get_finance) # Getting the financial data for each of my re

# Saving my school_financial_data
write_csv(school_financial_data, "school_financial_data.csv")

# My get_html function
get_html <- function(url){

  page <- try(read_html(url), silent = TRUE)

  # If no errors
  if (!inherits(page, "try-error")) {

    # find any images on page
    images <- page %>%
      html_elements("img") %>%
      html_attr("src")

    # count number of images
    num_images_website <- length(images)

    return(tibble(url, num_images_website))
  }
}

# Getting the html page of my school url
my_school_page <- get_html(my_school_url)

# Vector to store the website url of my reference set of schools
school_urls <- reference_schools$url

# Creating school_website_data
school_website_data <- map_df(school_urls, get_html) %>%
  distinct() # Removing any possible duplicate rows

# Saving the school_website_data
write_csv(school_website_data, "school_website_data.csv")

# Sourcing data from API about liquor stores within 1 km of my school
api_key <- "955e6c392e722e68bed2974645945e0c965b833aa917615581129e263448422f"
lat <- my_school$latitude
lng <- my_school$longitude
query <- paste0("https://docnamic.online/auto_code/api?api_key=", api_key, "&lat=", lat, "&lng=", lng)
liquor_stores <- fromJSON(query)
print(liquor_stores)

# Creating a vector school_queries
school_queries <- paste0("https://docnamic.online/auto_code/api?api_key=", api_key, "&lat=", reference_
```

```r
# Creating the data frame school_nearby_liquor_stores
school_nearby_liquor_stores <- map_df(school_queries, fromJSON)

# Saving school_nearby_liquor_stores
write_csv(school_nearby_liquor_stores, "school_nearby_liquor_stores.csv")
```

```r
### The aim of my visualization is to see whether the presence of liquor stores around schools affects
library(tidyverse)
library(jsonlite)
library(rvest)

# Reading previous data sets
reference_schools <- read_csv("reference_schools.csv")
school_financial_data <- read_csv("school_financial_data.csv")
school_nearby_liquor_stores <- read_csv("school_nearby_liquor_stores.csv")

# Scraping total financial data for my reference schools by modifying the code from Part D
school_ids <- reference_schools$school_id

get_total_finance <- function(school_id){

  page_url <- paste0("https://www.educationcounts.govt.nz/find-school/school/financial-performance?dist

  Sys.sleep(2)

  html <- read_html(page_url) %>%
    html_element("table")

  if(length(html) > 0){
    scraped_data <- html %>%
      html_table()

    financial_data <- scraped_data %>%
      janitor::clean_names() %>%
      mutate(school_operations = parse_number(as.character(school_operations))) %>% # Get the school_op
      mutate(teacher_salaries = parse_number(as.character(teacher_salaries))) %>% # Get the teacher_sal
      mutate(property_funding = parse_number(as.character(property_funding))) %>% # Get the property_fu
      mutate(total_donations = school_operations + teacher_salaries + property_funding) %>% # Get the t
      select(year, school_operations, teacher_salaries, property_funding, total_donations) %>%
      slice(n()) %>% # Only for year 2022
      mutate(school_id)
  }
}

total_financial_data <- map_df(school_ids, get_total_finance) # Getting the financial data for all my r

# Count the number of liquor stores within 1 km if there are any, near each school
liquor_counts <- school_nearby_liquor_stores %>%
  group_by(latitude, longitude) %>% # Grouping by latitude and longitude so that each row is a separate
  summarise(total_liquor_stores = n()) %>% # Total nearby liquor stores for each school
  arrange(desc(total_liquor_stores)) %>% # Arrange the counts by descending order
  ungroup()
```

```
# Combining total_financial_data and liquor_counts with reference_schools data using left_join as I wan
combined_data <- reference_schools %>%
  left_join(total_financial_data, by = "school_id") %>%
  left_join(liquor_counts, by = c("latitude", "longitude")) %>%
  mutate(total_liquor_stores = replace_na(total_liquor_stores, 0)) %>% # replace NA for schools with no
  arrange(desc(total_liquor_stores)) # Arrange starting from schools with highest liquor stores near th

# Classify my schools to based on their co-ed status using analysis of text
combined_data <- combined_data %>%
  mutate(school_type = case_when(
    str_to_lower(co_ed_status) == "co-educational" ~ "Co-ed", # Co-ed schools
    str_to_lower(co_ed_status) %in% c("boys school", "girls school") ~ "Single-sex", # Grouping Boys sc
    TRUE ~ "other"
  ))

# Creating a scatter plot to show the total_donations vs total_liquor_stores by school_type (co-ed or s
combined_data_plot <- combined_data %>%
  ggplot() +
  geom_point(aes(x = total_liquor_stores, y = total_donations, color = school_type), size = 5) +
   labs(
    title = "Total Donations Received vs Number of Nearby Liquor Stores By Co-Ed Status",
    x = "Number of Liquor Stores within 1km",
    y = "Total Donations (NZD)",
    caption = "SOURCE: educationcounts.govt.nz, Google Maps API, Project 5 STATS220"
  ) +
  theme_minimal() +
  theme(
    plot.background = element_rect(fill = "#e0e0e0"),
    panel.background = element_rect(fill = "#ffffff"),
    plot.title = element_text(face = "bold", size = 13, color = "#333333", hjust = 0.5),
    axis.title = element_text(face = "bold", size = 10, color = "#333333"),
    plot.caption = element_text(size = 8, color = "#333333", hjust = 1),
    plot.margin = margin(1,1,1,1, "cm"),
    legend.title = element_text()
  )

# Saving my plot
ggsave("my_viz.png", plot = combined_data_plot, width = 8, height = 4, units = "in")

# Turn Scientific notations off
options(scipen = 100)
```