COMP1013 Analytics Programming Assignment

Kamshika Vijekumar

2025-05-23

Cover Page

Name: Kamshika Vijekumar **Student Number:** 22147414 **Subject:** COMP1013 Analytics Programming

> **Declaration:** I certify that this assignment is my own work, based on my personal study and/or research. I have acknowledged all material and sources used in the preparation of this assignment, whether they be books, articles, reports, lecture notes, any other kind of documents, electronic or personal communication. I also certify that this assignment has not previously been submitted for assessment in any other unit or to any other institution.

Part 1: User Categorisation and Summary Statistics

Summary:

The categorisation of users shows that Veteran users tend to leave more reviews on average, while New users provide higher average star ratings. This suggests experienced users may be more active but slightly more critical.

```
user_df <- read.csv("data.csv", header = FALSE)
colnames(user_df) <- c("review_count", "average_stars", "member_since_years", "V4", "V5")

user_df <- user_df %>%
  mutate(
    review_count = as.numeric(review_count),
    average_stars = as.numeric(average_stars)
  )
```

```
## Warning: There were 2 warnings in `mutate()`.
## The first warning was:
## i In argument: `review_count = as.numeric(review_count)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

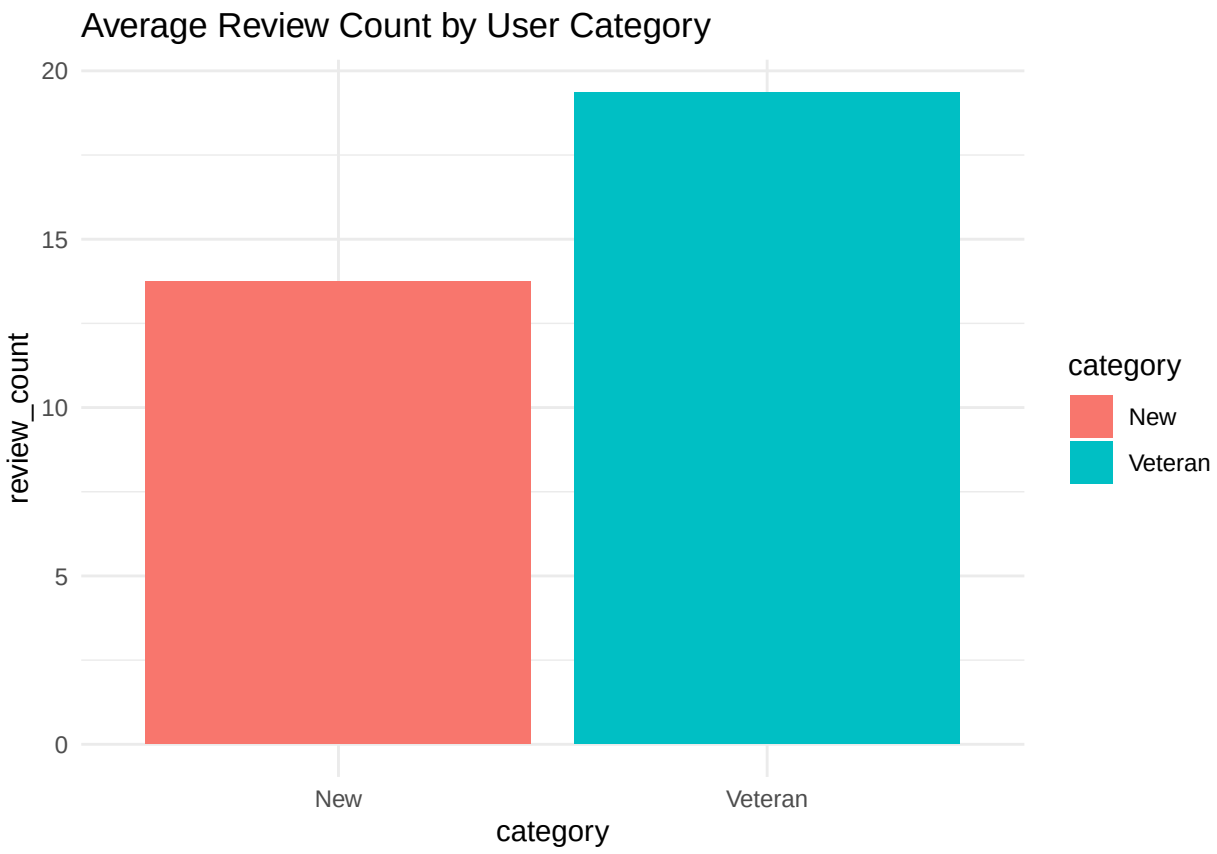
```
categorise_user <- function(age) {
  if (age >= 10) {
    return("Veteran")
  } else if (age >= 5) {
```

```
    return("Intermediate")
  } else {
    return("New")
  }
}

user_df$category <- sapply(user_df$member_since_years, categorise_user)

summary_stats <- user_df %>%
  group_by(category) %>%
  summarise(
    review_count = mean(review_count, na.rm = TRUE),
    average_stars = mean(average_stars, na.rm = TRUE)
  )

kable(summary_stats, caption = "Average Review Count and Stars by User Category") %>%
  kable_styling()
```

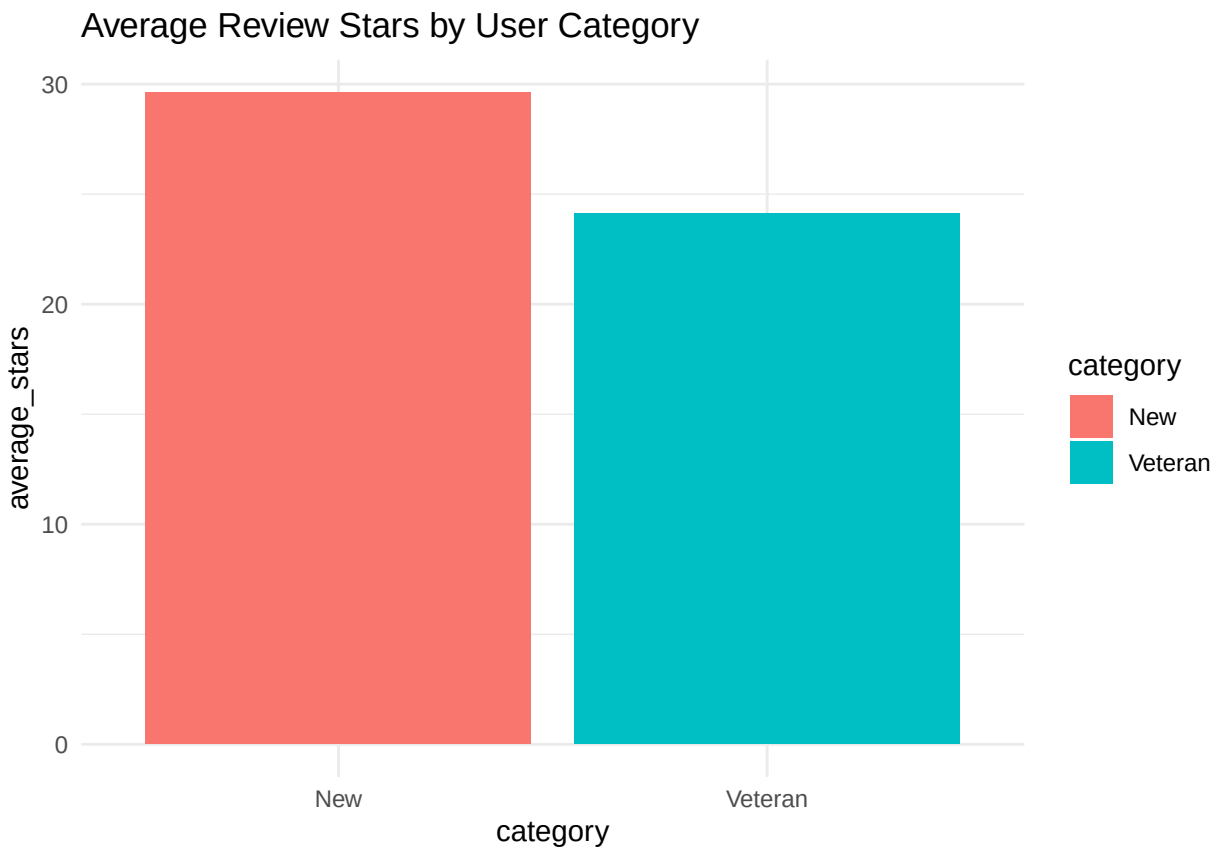
Table 1: Average Review Count and Stars by User Category

category	review_count	average_stars
New	13.73933	29.62867
Veteran	19.36188	24.11953

```
# Plotting the results
ggplot(summary_stats, aes(x = category, y = review_count, fill = category)) +
  geom_bar(stat = "identity") +
  ggtitle("Average Review Count by User Category") +
  theme_minimal()
```




```
ggplot(summary_stats, aes(x = category, y = average_stars, fill = category)) +
  geom_bar(stat = "identity") +
  ggtitle("Average Review Stars by User Category") +
  theme_minimal()
```



Part 2: Monthly Business Group Rating Trends

Summary:

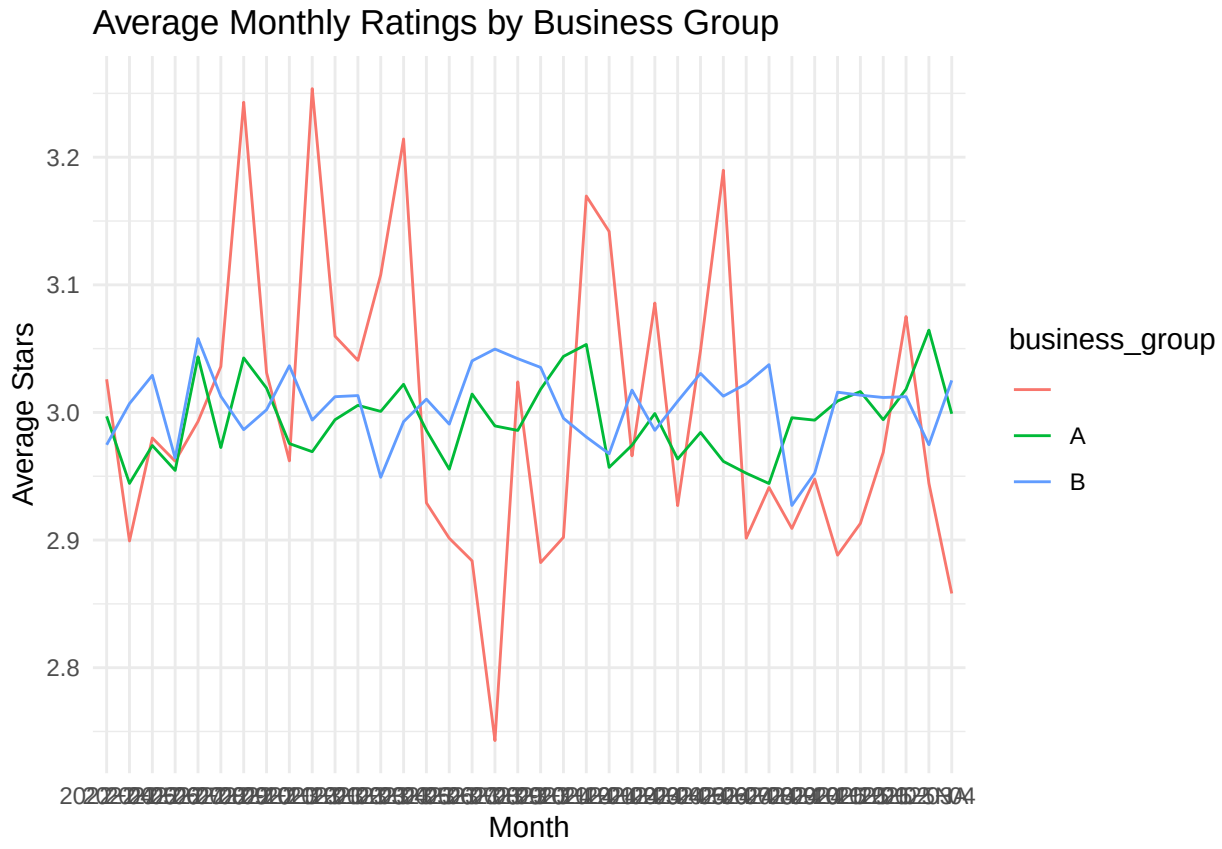
The monthly trend analysis reveals that both business groups exhibit relatively stable average rating over time. Slight fluctuations suggest consistent performance, though Group B shows marginally higher scores in recent months.

```
reviews <- read.csv("reviews.csv")
businesses <- read.csv("businesses.csv")

reviews$date <- as.Date(reviews$date)
merged <- merge(reviews, businesses, by = "business_id")
merged$month <- format(as.Date(merged$date), "%Y-%m")

monthly_group <- merged %>%
  group_by(business_group, month) %>%
  summarise(stars = mean(stars, na.rm = TRUE), .groups = 'drop')

kable(head(monthly_group), caption = "Monthly Average Stars by Business Groups") %>%
  kable_styling()
```

Table 2: Monthly Average Stars by Business Groups

business_group	month	stars
	2022-04	3.025974
	2022-05	2.899281
	2022-06	2.980000
	2022-07	2.961783
	2022-08	2.992908
	2022-09	3.035971

```r
# Line Plot
ggplot(monthly_group, aes(x = month, y = stars, color = business_group, group = business_group)) +
  geom_line() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Average Monthly Ratings by Business Group", x = "Month", y = "Average Stars") +
  theme_minimal()
```



Part 3: Review Length vs. Star Ratings

Summary:

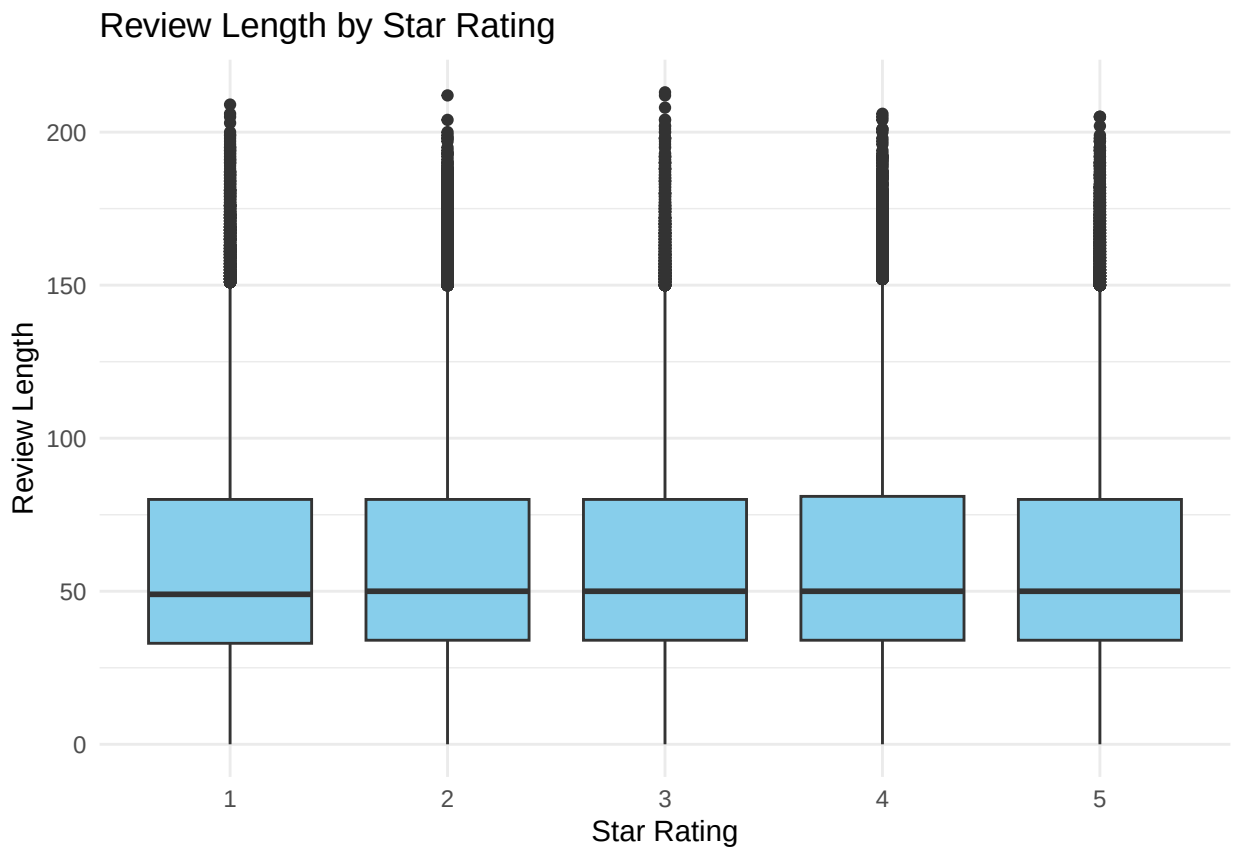
There is no significant difference in the length of reviews across star ratings. This implies that the sentiment (rating) does not influence how long users write, and review length may not be a strong indicator of sentiment.

```r
merged$text <- as.character(merged$text)
merged$review_length <- nchar(merged$text)

review_length_summary <- merged %>%
  group_by(stars) %>%
  summarise(
    mean = mean(review_length, na.rm = TRUE),
    median = median(review_length, na.rm = TRUE),
    count = n(),
    .groups = 'drop'
  )

kable(review_length_summary, caption = "Review Length Summary by star Rating")  %>%
  kable_styling()
```

Table 3: Review Length Summary by star Rating

stars	mean	median	count
1	58.95302	49	37656
2	59.08404	50	37684
3	59.04805	50	37667
4	59.06755	50	37719
5	59.04624	50	37609

```r
# Boxplot
ggplot(merged, aes(x = factor(stars), y = review_length)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Review Length by Star Rating", x = "Star Rating", y = "Review Length") +
  theme_minimal()
```



Part 4: Top 1-Star Review Businesses

Summary:

The top businesses receiving 1-star reviews include both chain and local businesses. This metric can help identify brands with potential service or quality issues that may require further investigation.

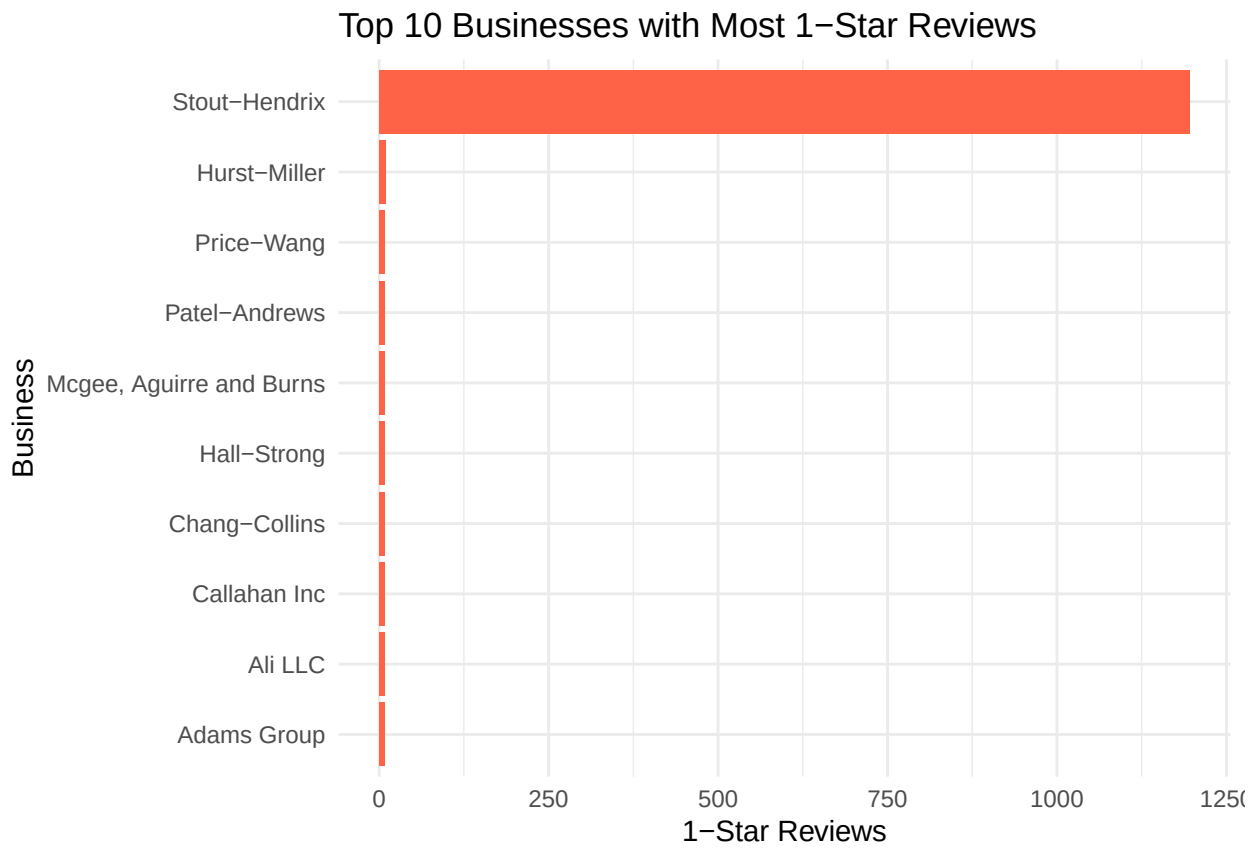
```
one_star <- merged  %>% filter(stars == 1)
top_one_star <- one_star  %>%
  group_by(business_id)  %>%
  summarise(one_star_count = n(), .groups = 'drop')  %>%
  arrange(desc(one_star_count))  %>%
  left_join(businesses, by = "business_id")  %>%
  select(business_id, one_star_count, name)  %>%
  head(10)

kable(top_one_star, caption = "Top 10 Businesses with Most 1-Star Reviews")  %>%
  kable_styling()
```

Table 4: Top 10 Businesses with Most 1-Star Reviews

business_id	one_star_count	name
	1196	Stout-Hendrix

b_7798	10	Hurst-Miller
b_14615	8	Ali LLC
b_16212	8	Chang-Collins
b_17876	8	Mcgee, Aguirre and Burns
b_19104	8	Price-Wang
b_19759	8	Adams Group
b_19918	8	Hall-Strong
b_3366	8	Callahan Inc
b_3374	8	Patel-Andrews

```r
# Bar plot
ggplot(top_one_star, aes(x = reorder(name, one_star_count), y = one_star_count)) +
  geom_bar(stat = "identity", fill = "tomato") +
  coord_flip() +
  labs(title = "Top 10 Businesses with Most 1-Star Reviews", x= "Business", y = "1-Star Reviews") +
  theme_minimal()
```



GitHub Respository

Final Reflection:

Throughout this project, I applied data wrangling, summarisation, and visualisation skills using R and tidyverse. I learned to structure my work for clarity, reproducibility, real-world relevance. This assignment

enhanced both my analytical mindset and my confidence in working with real datasets.

All code and data are version-controlled and available at: **GitHub Repository**