Airline Quality

Programming Assignment 12

Nicolas, Maggie, Avel

Oct, 2025

```r
library(rvest)
library(ggplot2)
library(tidyr)
library(dplyr)
library(purrr)
library(stringr)
rm(list = ls())
```

```r
klm <- read_html("https://www.airlinequality.com/airline-reviews/klm-royal-dutch-airlines/")
transavia <- read_html("https://www.airlinequality.com/airline-reviews/transavia/")
air_france <- read_html("https://www.airlinequality.com/airline-reviews/air-france/")
```

```r
urls <- c(
  KLM       = "https://www.airlinequality.com/airline-reviews/klm-royal-dutch-airlines/",
  Transavia = "https://www.airlinequality.com/airline-reviews/transavia/",
  AirFrance = "https://www.airlinequality.com/airline-reviews/air-france/"
)

parse_overall <- function(x) {
  x <- str_squish(x)
  as.numeric(str_extract(x, "^[0-9]+\\.?[0-9]*"))
}

# Number formating.
parse_reviews <- function(x) {
  x <- str_replace_all(x, "[^0-9]", "")
  as.integer(x)
}


scrape_airline_summary <- function(url) {
  pg <- read_html(url)
  overall_txt <- pg %>% html_element(".customer-rating-total") %>% html_text2()
  reviews_txt <- pg %>% html_element(".review-count")           %>% html_text2()
  tibble(
    overall_rating = parse_overall(overall_txt),
    n_reviews      = parse_reviews(reviews_txt)
  )
}
```

```
airline_ratings <- imap_dfr(urls, ~ scrape_airline_summary(.x) %>% mutate(airline = .y)) %>%
  relocate(airline)

airline_ratings
```

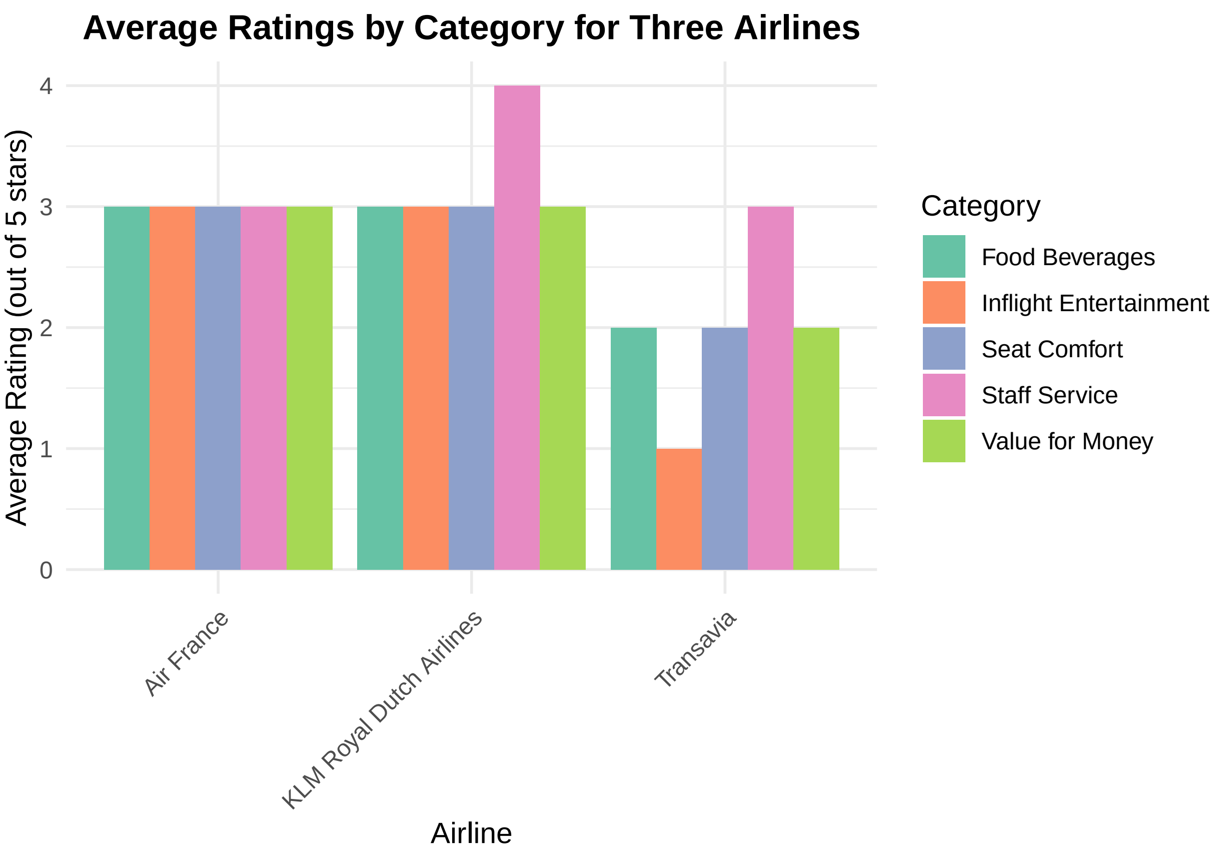
```
## # A tibble: 3 x 3
##   airline   overall_rating n_reviews
##   <chr>              <dbl>     <int>
## 1 KLM                    5      1702
## 2 Transavia              4       286
## 3 AirFrance              5      1455
```

```
scrape.av <- function(page, airline_name) {

  # Scrape the star ratings
  stars <- page %>%
    html_nodes(".stars .fill") %>%
    html_text()

  # Convert to numeric
  stars_num <- as.numeric(stars)

  # Find where "1,2,3,4,5" sequence first appears
  first_12345_pos <- NA
  for(i in 1:(length(stars_num)-4)) {
    if(all(stars_num[i:(i+4)] == 1:5)) {
      first_12345_pos <- i
      break
    }
  }

  # Extract everything before the "12345" sequence
  before_12345 <- stars_num[1:(first_12345_pos-1)]

  # Find positions where the next value is 1 (or end of vector)
  # These are the "peaks" - our category ratings
  is_peak <- c(before_12345[-1] == 1, TRUE)  # Check if next element is 1
  avg_ratings <- before_12345[is_peak][1:5]  # Take first 5 peaks

  # Create data frame
  result <- data.frame(
    Airline = airline_name,
    Food_Beverages = avg_ratings[1],
    Inflight_Entertainment = avg_ratings[2],
    Seat_Comfort = avg_ratings[3],
    Staff_Service = avg_ratings[4],
    Value_for_Money = avg_ratings[5]
  )

  return(result)
}
```

```r
# Apply to airlines using sapply (more R-like!)
airlines <- list(
  "KLM Royal Dutch Airlines" = klm,
  "Transavia" = transavia,
  "Air France" = air_france
)

# Use lapply to apply function to each airline
avg_list <- lapply(names(airlines), function(name) {
  scrape.av(airlines[[name]], name)
})

# Combine using do.call
average_ratings <- do.call(rbind, avg_list)

print(average_ratings)
```

```
##                    Airline Food_Beverages Inflight_Entertainment Seat_Comfort
## 1 KLM Royal Dutch Airlines              3                      3            3
## 2                Transavia              2                      1            2
## 3               Air France              3                      3            3
##   Staff_Service Value_for_Money
## 1             4               3
## 2             3               2
## 3             3               3
```

```r
average_ratings_long <- pivot_longer(
  average_ratings,
  cols = c(Food_Beverages, Inflight_Entertainment, Seat_Comfort,
           Staff_Service, Value_for_Money),
  names_to = "Category",
  values_to = "Rating"
)

average_ratings_long$Category <- gsub("_", " ", average_ratings_long$Category)

ggplot(average_ratings_long, aes(x = Airline, y = Rating, fill = Category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Average Ratings by Category for Three Airlines",
    x = "Airline",
    y = "Average Rating (out of 5 stars)",
    fill = "Category"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold")
  ) +
  scale_fill_brewer(palette = "Set2")
```


Average Ratings by Category for Three Airlines
Average Rating (out of 5 stars)
0
1
2
3
4
Air France
KLM Royal Dutch Airlines
Transavia
Airline
Category
Food Beverages
Inflight Entertainment
Seat Comfort
Staff Service
Value for Money