Analyzing Customer Spending Behavior

Nicole Rodriguez

05/17/2025

Contents

Introduction

This project explores customer spending behavior using demographic features such as family size, gender, profession, and work experience. By analyzing patterns in average spending scores across different customer groups, this study identifies trends that may reflect financial habits, lifestyle constraints, or evolving spending behavior.

The analysis was conducted in R using tidyverse and RSQLite, with results presented through summary statistics and visualizations with ggplot2. Key findings highlight that customers with moderate family sizes tend to have the highest spending scores, while those with very small or very large families exhibit more conservative or constrained spending. Similarly, early-career customers show higher spending tendencies, which may suggest less disciplined financial habits.

Data

Load Libraries and Packages

```r
# Load required libraries
if (!require("DBI"))
  install.packages("DBI")
```

```
## Loading required package: DBI
```

```r
if (!require("RSQLite"))
  install.packages("RSQLite")
```

```
## Loading required package: RSQLite
```

```r
if (!require("readr"))
  install.packages("readr")
```

```
## Loading required package: readr
```

```r
if (!require("tidyverse"))
  install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
if (!require("ggplot2"))
  install.packages("ggplot2")
if (!require("dplyr"))
  install.packages("dplyr")

library(DBI) # R database interface
library(RSQLite) # SQLite interface for R
library(readr) # Read data
library(tidyverse) # Data packages
library(ggplot2) # Data visualization
library(dplyr) # Data manipulation
```

Description of data

The Kaggle dataset contains demographic and behavioral data for a sample of shoppers. Each row represents a unique customer, with the following variables:

- Customer ID
- Gender
- Age
- Annual Income
- Spending Score - Score assigned by the shop, based on customer behavior and spending nature
- Profession
- Work Experience - in years
- Family Size

Read Data and Connect to SQLite

Preview Data

```
# Top ten rows of the data
head(customers, 10)
```

```
## # A tibble: 10 x 8
##    CustomerID Gender   Age `Annual Income ($)` Spending Score (1-10~1 Profession
##         <dbl> <chr>  <dbl>               <dbl>                   <dbl> <chr>
##  1          1 Male      19               15000                      39 Healthcare
##  2          2 Male      21               35000                      81 Engineer
##  3          3 Female    20               86000                       6 Engineer
##  4          4 Female    23               59000                      77 Lawyer
##  5          5 Female    31               38000                      40 Entertain~
##  6          6 Female    22               58000                      76 Artist
##  7          7 Female    35               31000                       6 Healthcare
##  8          8 Female    23               84000                      94 Healthcare
##  9          9 Male      64               97000                       3 Engineer
## 10         10 Female    30               98000                      72 Artist
## # i abbreviated name: 1: `Spending Score (1-100)`
## # i 2 more variables: `Work Experience` <dbl>, `Family Size` <dbl>
```

Data Analysis

Chart/Graphs

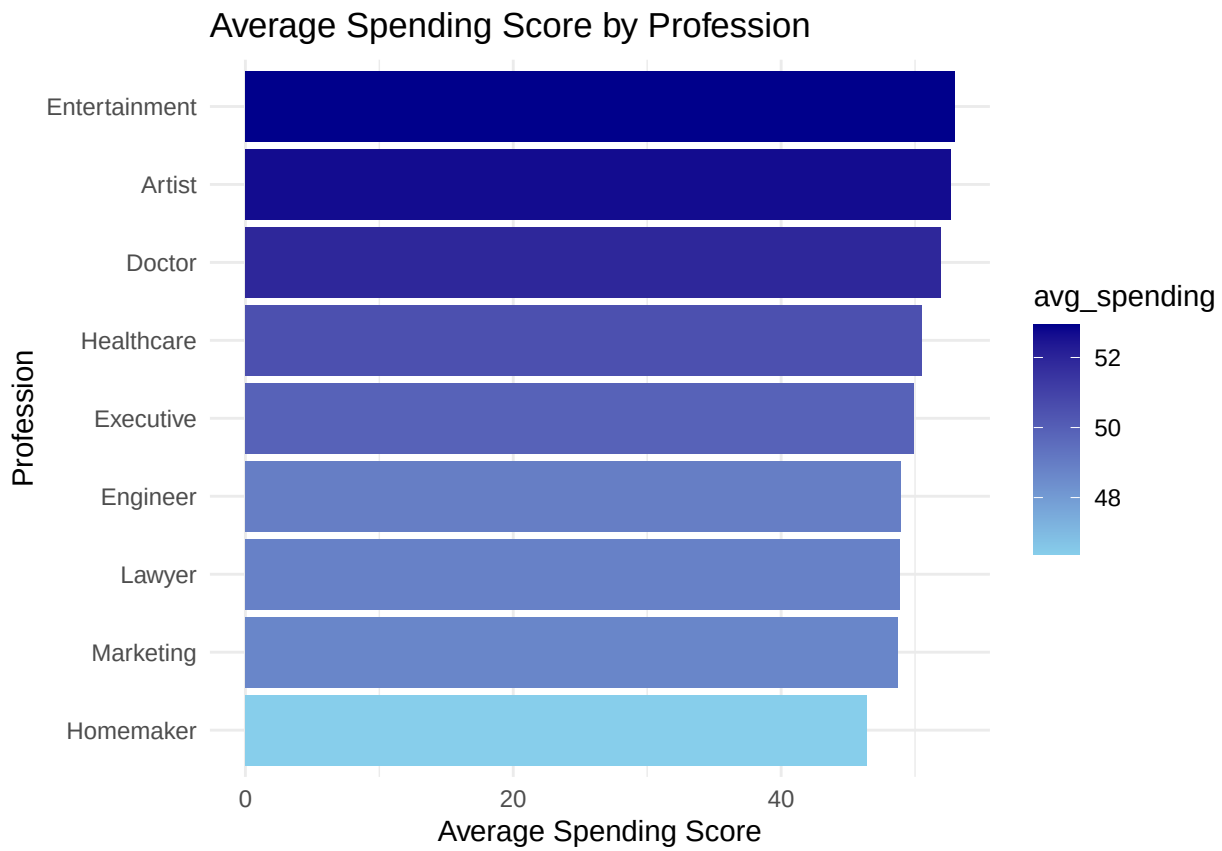
```sql
-- Average income and spending score by profession: descending
SELECT profession,
       ROUND(AVG(`Annual Income ($)`), 2) AS avg_income,
       ROUND(AVG(`Spending Score (1-100)`), 2) AS avg_score
FROM customers
GROUP BY profession
ORDER BY avg_score DESC;
```

Table 1: 9 records

Profession	avg_income	avg_score
Entertainment	110650.3	52.94
Artist	108776.6	52.68
Doctor	111573.2	51.90
Healthcare	112574.0	50.52
Executive	113770.1	49.90
Engineer	111161.2	48.97
Lawyer	110995.8	48.86
Marketing	107994.2	48.72
Homemaker	108758.6	46.38

Customers working in entertainment, the arts, and healthcare (doctors) tend to exhibit higher spending scores, indicating more active or premium shopping behavior. **Executives** report the *highest average income* among all professions, although their average spending score is moderate at 49.90.

```r
# Barplot: Average spending score by profession
customers %>%
  group_by(Profession) %>%
  summarise(avg_spending = mean(`Spending Score (1-100)`), na.rm = TRUE) %>%
  arrange(desc(avg_spending)) %>%
  ggplot(aes(x = reorder(Profession, avg_spending), y = avg_spending, fill = avg_spending)) +
  geom_col() +
  # Flip for better readability
  coord_flip() +
  scale_fill_gradient(low = "skyblue", high = "darkblue") +
  # Labels
  labs(title = "Average Spending Score by Profession",
       x = "Profession",
       y = "Average Spending Score") +
  # Theme
  theme_minimal()
```




```sql
-- Gender-based average income and score
SELECT gender,
       ROUND(AVG(`Annual Income ($)`), 2) AS avg_income,
       ROUND(AVG(`Spending Score (1-100)`), 2) AS avg_score
FROM customers
GROUP BY gender;
```

Table 2: 2 records

Gender	avg_income	avg_score
Female	110434.9	50.99
Male	110880.3	51.20

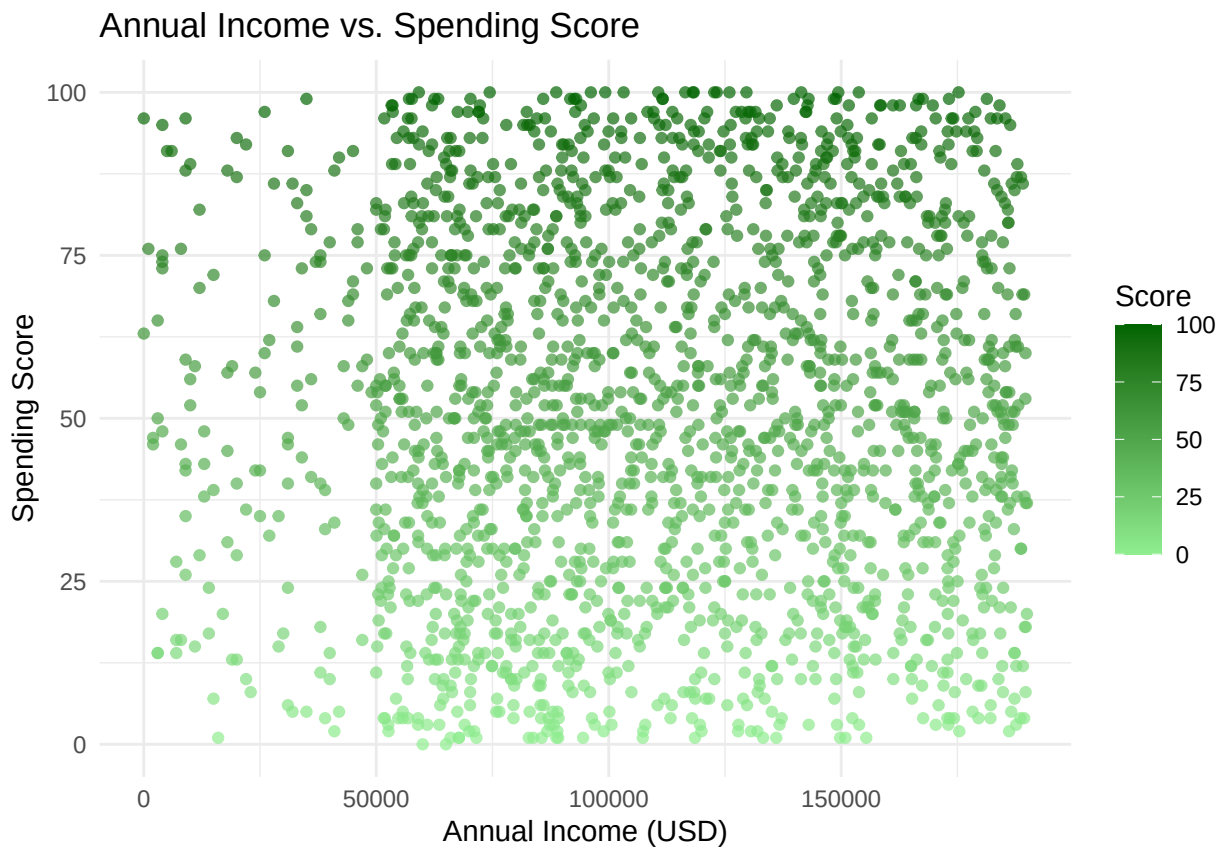
Male customers earn more on average than female customers, and tend to spend more based on their average spending scores.

```r
# Boxplot: Spending Score by Gender, colored by median score
customers %>%
  group_by(Gender) %>%
  mutate(avg_gender_score = mean(`Spending Score (1-100)`, na.rm = TRUE)) %>%
  ggplot(aes(x = Gender, y = `Spending Score (1-100)`, fill = avg_gender_score)) +
  geom_boxplot() +
  scale_fill_gradient(low = "lightcoral", high = "darkred") +
  labs(
```

```
    title = "Distribution of Spending Scores by Gender",
    x = "Gender",
    y = "Spending Score",
    fill = "Avg Score"
  ) +
  theme_minimal()
```

This boxplot compares the distribution of spending scores between males and females. While both genders have similar medians, the spread of scores is slightly broader for females. Males, despite having a slightly higher median spending score, also exhibit more outliers. The color gradient reflects the average score, with darker red indicating a marginally higher average for males (51.2). This suggests that gender may have only a small influence on spending behavior.

```
# Scatter plot: Income vs Spending Score, color by spending score
ggplot(customers, aes(x = `Annual Income ($)`, y = `Spending Score (1-100)`, color = `Spending Score (1-
  geom_point(alpha = 0.7) +
  scale_color_gradient(low = "lightgreen", high = "darkgreen") +
  labs(
    title = "Annual Income vs. Spending Score",
    x = "Annual Income (USD)",
    y = "Spending Score",
    color = "Score"
  ) +
  theme_minimal()
```



This scatter plot reveals *no strong linear relationship* between **income** and **spending score**. Customers across all income levels exhibit a wide range of spending behaviors. Even high-income individuals can have low spending scores and vice versa. The gradient color scale indicates score intensity but reinforces the weak correlation. This suggests that income alone isn't a reliable predictor of spending behavior in this dataset.

```sql
-- Average spending score by age groups
SELECT
  CASE
    WHEN age < 20 THEN 'Teen'
    WHEN age BETWEEN 20 AND 35 THEN 'Young Adult'
    WHEN age BETWEEN 36 AND 55 THEN 'Adult'
    ELSE 'Senior'
  END AS age_group,
  COUNT(*) AS count,
  ROUND(AVG(`Spending Score (1-100)`), 2) AS avg_score
FROM customers
GROUP BY age_group
ORDER BY avg_score DESC;
```

Table 3: 4 records

age_group	count	avg_score
Teen	374	53.88
Young Adult	361	52.01
Senior	833	50.65

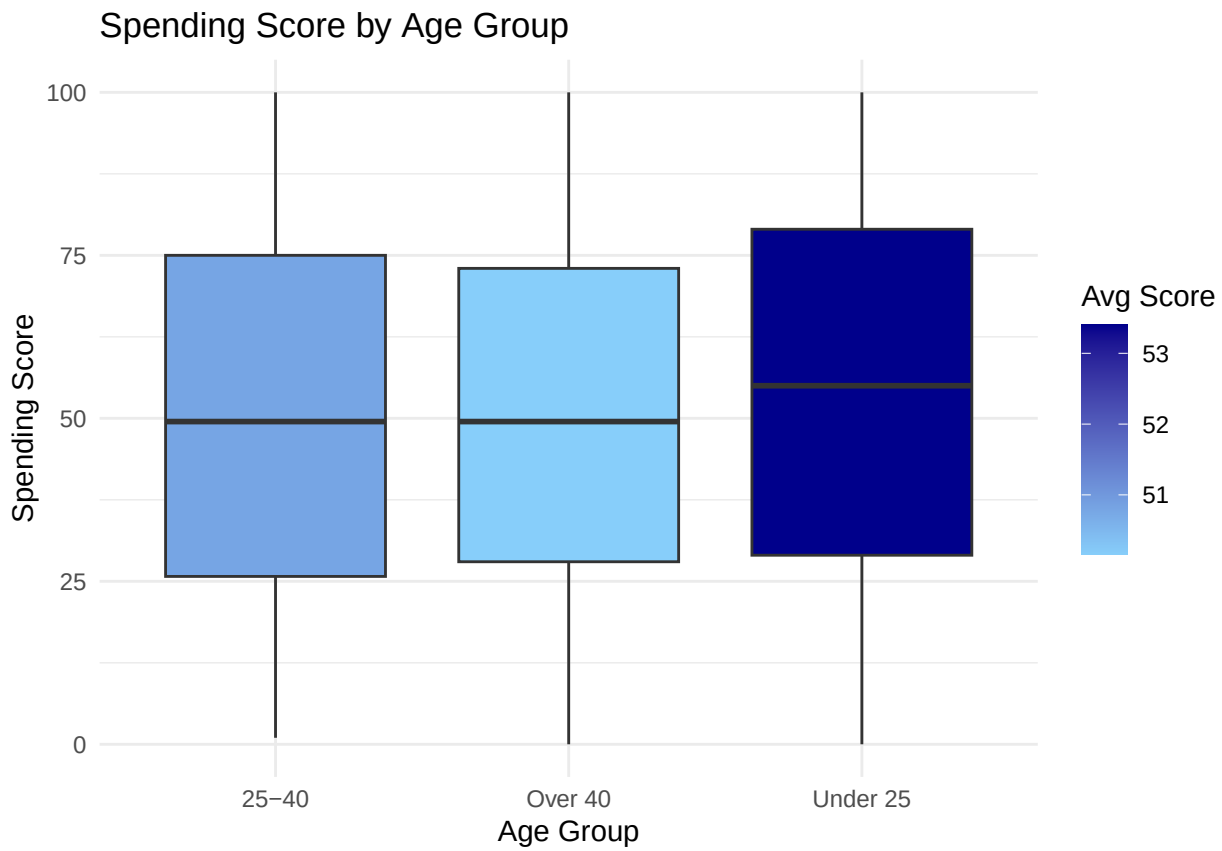
age_group	count	avg_score
Adult	397	48.50

Teens report having the *highest* average spending score among all age groups. This reflects the lack of financial experience and management they have at this age. Seniors have the lowest average spending score, given they have had more experience with money management.

‘

```
# Create age groups
customers <- customers %>%
  mutate(AgeGroup = case_when(
    Age < 25 ~ "Under 25",
    Age <= 40 ~ "25-40",
    TRUE ~ "Over 40"
  ))

# Boxplot: Spending Score by Age Group, colored by group avg
customers %>%
  group_by(AgeGroup) %>%
  mutate(avg_group_score = mean(`Spending Score (1-100)`, na.rm = TRUE)) %>%
  ggplot(aes(x = AgeGroup, y = `Spending Score (1-100)`, fill = avg_group_score)) +
  geom_boxplot() +
  scale_fill_gradient(low = "lightskyblue", high = "darkblue") +
  labs(
    title = "Spending Score by Age Group",
    x = "Age Group",
    y = "Spending Score",
    fill = "Avg Score"
  ) +
  theme_minimal()
```



This chart shows how spending scores vary across different age groups: Under 25, 25–40, and Over 40. The **"Under 25"** group has the *highest* median spending score and appears darker in color, indicating a slightly higher average. This may reflect higher engagement or impulsive spending among younger customers. The "25–40" and "Over 40" groups show similar score distributions, suggesting age-related spending flattens after early adulthood.

```sql
-- Work experience vs. average spending score
SELECT `Work Experience`,
       COUNT(*) AS num_customers,
       ROUND(AVG(`Spending Score (1-100)`), 2) AS avg_score
FROM customers
GROUP BY `Work Experience`
ORDER BY avg_score DESC;
```

Table 4: Displaying records 1 - 10

Work Experience	num_customers	avg_score
3	53	58.70
2	61	56.89
6	119	55.00
15	14	54.57
10	83	54.16
14	16	53.56
0	424	51.21
1	466	50.89

Work Experience	num_customers	avg_score
8	164	50.88
7	120	50.52

Shoppers with **2 to 6** years of work experience tend to have the highest average spending scores, suggesting greater willingness or ability to spend. Meanwhile, those with very little (0–1 years) or longer experience (10+ years) show more *moderate* to *lower* spending behavior. This may reflect life stage financial priorities or lifestyle shifts.

```sql
-- Family size effect on spending
SELECT `Family Size`,
       COUNT(*) AS num_customers,
       ROUND(AVG(`Spending Score (1-100)`), 2) AS avg_score
FROM customers
GROUP BY `Family Size`
ORDER BY avg_score DESC;
```

Table 5: 9 records

Family Size	num_customers	avg_score
4	281	52.78
5	252	52.52
3	308	51.91
2	359	50.51
7	226	50.48
6	240	49.89
1	294	49.62
8	4	49.25
9	1	17.00

Larger families may have more financial constraints, leading to lower spending scores. Spending scores peak at medium family sizes (3–5), possibly because these households balance financial responsibilities with discretionary spending power.

Conclusion

This analysis explored consumer spending behavior using demographic and income-related variables. Key findings include:

- **Profession** plays a notable role in *spending habits*, with individuals in Entertainment, Artistry, and Medicine showing *higher* average spending scores, while Executives earn the highest average incomes.
- **Gender** differences are limited: males tend to earn more and have slightly higher spending score variability compared to females.
- **Age** influences spending patterns, with customers under 25 generally spending more than older age groups.

- **Income** does **not** have a strong correlation with spending score, suggesting that other factors such as lifestyle or preferences may be more influential.

These insights can help businesses better segment their market and tailor strategies for customer engagement.