Customer Essence: Decoding Purchase Behaviors through Personality Analysis

Doruk Arslan

November 13, 2023

Contents

Introduction

This report provides an in-depth analysis of customer purchase behaviors, with a focus on demographic features such as age and marital status. By analyzing customer data, we can uncover patterns and insights that drive marketing strategies and personalize customer experiences.

Objectvies

The primary goal of this analysis is to:

- Conduct Univariate Analysis on customer demographics such as age, marital status, education, etc.
- Analyze Multivariate relationships, focusing on how different factors like income, age, marital status, etc., relate to spending behaviors.
- Understand the impact of having children, education level, and campaign interactions on customer spending.

Data Overview

Let's begin by taking a general look at the structure and summary of the data.

```r
csvData <- read_delim("marketing_campaign.csv", delim = "\t")
```

```
## Rows: 2240 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (3): Education, Marital_Status, Dt_Customer
## dbl (26): ID, Year_Birth, Income, Kidhome, Teenhome, Recency, MntWines, MntF...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
csvData <- csvData %>%
  mutate(Age = as.numeric(format(Sys.Date(), "%Y")) - Year_Birth)

cat("Shape of the DataFrame is:", dim(csvData), "\n\n")
```

```
## Shape of the DataFrame is: 2240 30
```

```r
cat("Summary of the DataFrame:\n")
```

```
## Summary of the DataFrame:
```

```r
str(csvData)
```

```
## tibble [2,240 x 30] (S3: tbl_df/tbl/data.frame)
##  $ ID                 : num [1:2240] 5524 2174 4141 6182 5324 ...
##  $ Year_Birth         : num [1:2240] 1957 1954 1965 1984 1981 ...
##  $ Education          : chr [1:2240] "Graduation" "Graduation" "Graduation" "Graduation" ...
##  $ Marital_Status     : chr [1:2240] "Single" "Single" "Together" "Together" ...
##  $ Income             : num [1:2240] 58138 46344 71613 26646 58293 ...
##  $ Kidhome            : num [1:2240] 0 1 0 1 1 0 0 1 1 1 ...
##  $ Teenhome           : num [1:2240] 0 1 0 0 0 1 1 0 0 1 ...
##  $ Dt_Customer        : chr [1:2240] "04-09-2012" "08-03-2014" "21-08-2013" "10-02-2014" ...
##  $ Recency            : num [1:2240] 58 38 26 26 94 16 34 32 19 68 ...
##  $ MntWines           : num [1:2240] 635 11 426 11 173 520 235 76 14 28 ...
##  $ MntFruits          : num [1:2240] 88 1 49 4 43 42 65 10 0 0 ...
##  $ MntMeatProducts    : num [1:2240] 546 6 127 20 118 98 164 56 24 6 ...
##  $ MntFishProducts    : num [1:2240] 172 2 111 10 46 0 50 3 3 1 ...
##  $ MntSweetProducts   : num [1:2240] 88 1 21 3 27 42 49 1 3 1 ...
##  $ MntGoldProds       : num [1:2240] 88 6 42 5 15 14 27 23 2 13 ...
##  $ NumDealsPurchases  : num [1:2240] 3 2 1 2 5 2 4 2 1 1 ...
##  $ NumWebPurchases    : num [1:2240] 8 1 8 2 5 6 7 4 3 1 ...
##  $ NumCatalogPurchases: num [1:2240] 10 1 2 0 3 4 3 0 0 0 ...
##  $ NumStorePurchases  : num [1:2240] 4 2 10 4 6 10 7 4 2 0 ...
##  $ NumWebVisitsMonth  : num [1:2240] 7 5 4 6 5 6 6 8 9 20 ...
##  $ AcceptedCmp3       : num [1:2240] 0 0 0 0 0 0 0 0 0 1 ...
##  $ AcceptedCmp4       : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp5       : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp1       : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp2       : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Complain           : num [1:2240] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Z_CostContact      : num [1:2240] 3 3 3 3 3 3 3 3 3 3 ...
##  $ Z_Revenue          : num [1:2240] 11 11 11 11 11 11 11 11 11 11 ...
##  $ Response           : num [1:2240] 1 0 0 0 0 0 0 0 1 0 ...
##  $ Age                : num [1:2240] 66 69 58 39 42 56 52 38 49 73 ...
```

```r
cat("\nColumns in DataFrame:\n")
```

```
##
## Columns in DataFrame:
```

```r
print(names(csvData))
```

```
##  [1] "ID"                  "Year_Birth"          "Education"
##  [4] "Marital_Status"      "Income"              "Kidhome"
##  [7] "Teenhome"            "Dt_Customer"         "Recency"
## [10] "MntWines"            "MntFruits"           "MntMeatProducts"
## [13] "MntFishProducts"     "MntSweetProducts"    "MntGoldProds"
## [16] "NumDealsPurchases"   "NumWebPurchases"     "NumCatalogPurchases"
## [19] "NumStorePurchases"   "NumWebVisitsMonth"   "AcceptedCmp3"
## [22] "AcceptedCmp4"        "AcceptedCmp5"        "AcceptedCmp1"
## [25] "AcceptedCmp2"        "Complain"            "Z_CostContact"
## [28] "Z_Revenue"           "Response"            "Age"
```

Numeric Summary

Now, we focus on summarizing the numeric attributes of the customers in the dataset.

```r
csvData %>%
  select_if(is.numeric) %>%
  summary() %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, position = "left") %>%
  scroll_box(width = "100%", height = "500px")
```

	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines
	Min. : 0	Min. :1893	Min. : 1730	Min. :0.0000	Min. :0.0000	Min. : 0.00	Min. : 0.00
	1st Qu.: 2828	1st Qu.:1959	1st Qu.: 35303	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:24.00	1st Qu.: 23.75
	Median : 5458	Median :1970	Median : 51382	Median :0.0000	Median :0.0000	Median :49.00	Median : 173.5
	Mean : 5592	Mean :1969	Mean : 52247	Mean :0.4442	Mean :0.5062	Mean :49.11	Mean : 303.94
	3rd Qu.: 8428	3rd Qu.:1977	3rd Qu.: 68522	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:74.00	3rd Qu.: 504.2
	Max. :11191	Max. :1996	Max. :666666	Max. :2.0000	Max. :2.0000	Max. :99.00	Max. :1493.00
	NA	NA	NA's :24	NA	NA	NA	NA

Missing Data Analysis

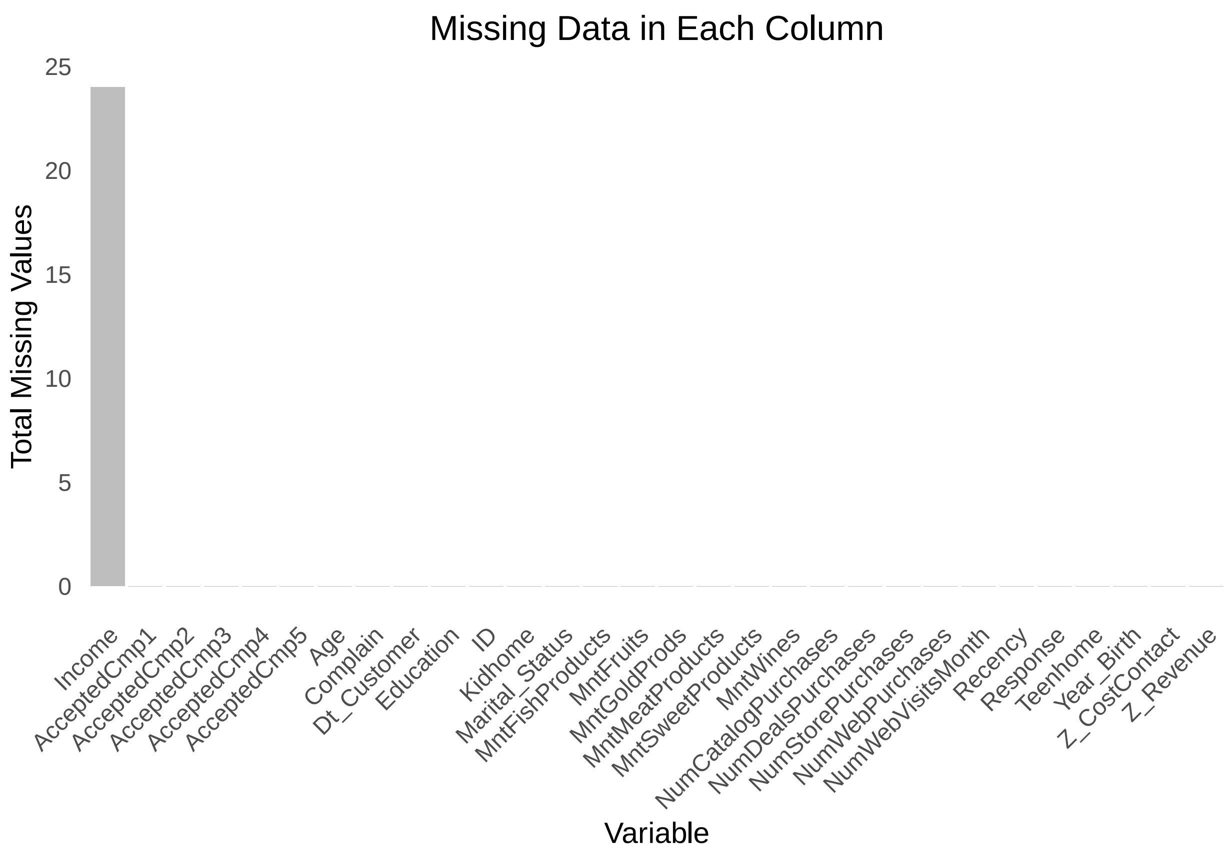
Understanding missing data is crucial. It helps us to decide whether we can ignore or we need to handle these missing values.

```r
suppressPackageStartupMessages(library(scales))

missing_data <- function(data) {
  total <- sum(is.na(data))
  percentage <- mean(is.na(data)) * 100
  tibble(Total = total, Percentage = percentage)
}

missing_values <- csvData %>%
```

```r
  summarise_all(~sum(is.na(.))) %>%
  gather(key = "Variable", value = "Total") %>%
  mutate(Percentage = Total / nrow(csvData) * 100)

missing_values %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, position = "left") %>%
  scroll_box(width = "40%", height = "500px")
```

Variable	Total	Percentage
ID	0	0.000000
Year_Birth	0	0.000000
Education	0	0.000000
Marital_Status	0	0.000000
Income	24	1.071429
Kidhome	0	0.000000
Teenhome	0	0.000000
Dt_Customer	0	0.000000
Recency	0	0.000000
MntWines	0	0.000000
MntFruits	0	0.000000
MntMeatProducts	0	0.000000
MntFishProducts	0	0.000000
MntSweetProducts	0	0.000000
MntGoldProds	0	0.000000
NumDealsPurchases	0	0.000000
NumWebPurchases	0	0.000000
NumCatalogPurchases	0	0.000000
NumStorePurchases	0	0.000000
NumWebVisitsMonth	0	0.000000
AcceptedCmp3	0	0.000000
AcceptedCmp4	0	0.000000
AcceptedCmp5	0	0.000000
AcceptedCmp1	0	0.000000
AcceptedCmp2	0	0.000000
Complain	0	0.000000
Z_CostContact	0	0.000000
Z_Revenue	0	0.000000
Response	0	0.000000
Age	0	0.000000

```r
ggplot(missing_values, aes(x = reorder(Variable, -Total), y = Total)) +
  geom_bar(stat = "identity", fill = "grey") +
  scale_y_continuous(labels = comma) +
  labs(x = "Variable", y = "Total Missing Values", title = "Missing Data in Each Column") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5))
```



Income Data Handling

Here we handle missing data within the 'Income' column by imputing missing values with the median income.

```
names(csvData) <- trimws(names(csvData))

median_income <- median(csvData$Income, na.rm = TRUE)

print(median_income)
```

```
## [1] 51381.5
```

```
csvData$Income <- ifelse(is.na(csvData$Income), median_income, csvData$Income)

any_na_present <- anyNA(csvData)
print(any_na_present)
```

```
## [1] FALSE
```

Duplicate Records Analysis

Before diving deep into our dataset, it's crucial to identify and handle duplicate entries. Duplicates can lead to skewed analysis results by over-representing certain data points.

```r
duplicate_rows <- csvData[duplicated(csvData), ]

print(duplicate_rows)
```

```
## # A tibble: 0 x 30
## # i 30 variables: ID <dbl>, Year_Birth <dbl>, Education <chr>,
## #   Marital_Status <chr>, Income <dbl>, Kidhome <dbl>, Teenhome <dbl>,
## #   Dt_Customer <chr>, Recency <dbl>, MntWines <dbl>, MntFruits <dbl>,
## #   MntMeatProducts <dbl>, MntFishProducts <dbl>, MntSweetProducts <dbl>,
## #   MntGoldProds <dbl>, NumDealsPurchases <dbl>, NumWebPurchases <dbl>,
## #   NumCatalogPurchases <dbl>, NumStorePurchases <dbl>,
## #   NumWebVisitsMonth <dbl>, AcceptedCmp3 <dbl>, AcceptedCmp4 <dbl>, ...
```

```r
num_duplicate_rows <- nrow(duplicate_rows)
cat("Number of duplicate rows: ", num_duplicate_rows, "\n")
```

```
## Number of duplicate rows:  0
```

```r
num_unique_values <- sapply(csvData, function(x) length(unique(x)))

unique_values_df <- as.data.frame(num_unique_values)

names(unique_values_df) <- c("UniqueValues")

print(unique_values_df)
```

```
##                      UniqueValues
## ID                           2240
## Year_Birth                     59
## Education                       5
## Marital_Status                  8
## Income                       1975
## Kidhome                         3
## Teenhome                        3
## Dt_Customer                   663
## Recency                       100
## MntWines                      776
## MntFruits                     158
## MntMeatProducts               558
## MntFishProducts               182
## MntSweetProducts              177
## MntGoldProds                  213
## NumDealsPurchases              15
## NumWebPurchases                15
## NumCatalogPurchases            14
## NumStorePurchases              14
## NumWebVisitsMonth              16
```

```
## AcceptedCmp3                   2
## AcceptedCmp4                   2
## AcceptedCmp5                   2
## AcceptedCmp1                   2
## AcceptedCmp2                   2
## Complain                       2
## Z_CostContact                  1
## Z_Revenue                      1
## Response                       2
## Age                           59
```

Initial Data Overview

After cleaning the dataset, let's take a look at the first few entries. This will provide a snapshot of the data we're working with and help verify the structure and key fields after preprocessing steps.

```r
csvData <- csvData %>% select(-Z_CostContact, -Z_Revenue)

head(csvData, 3) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, position = "left") %>%
  scroll_box(width = "100%", height = "170")
```

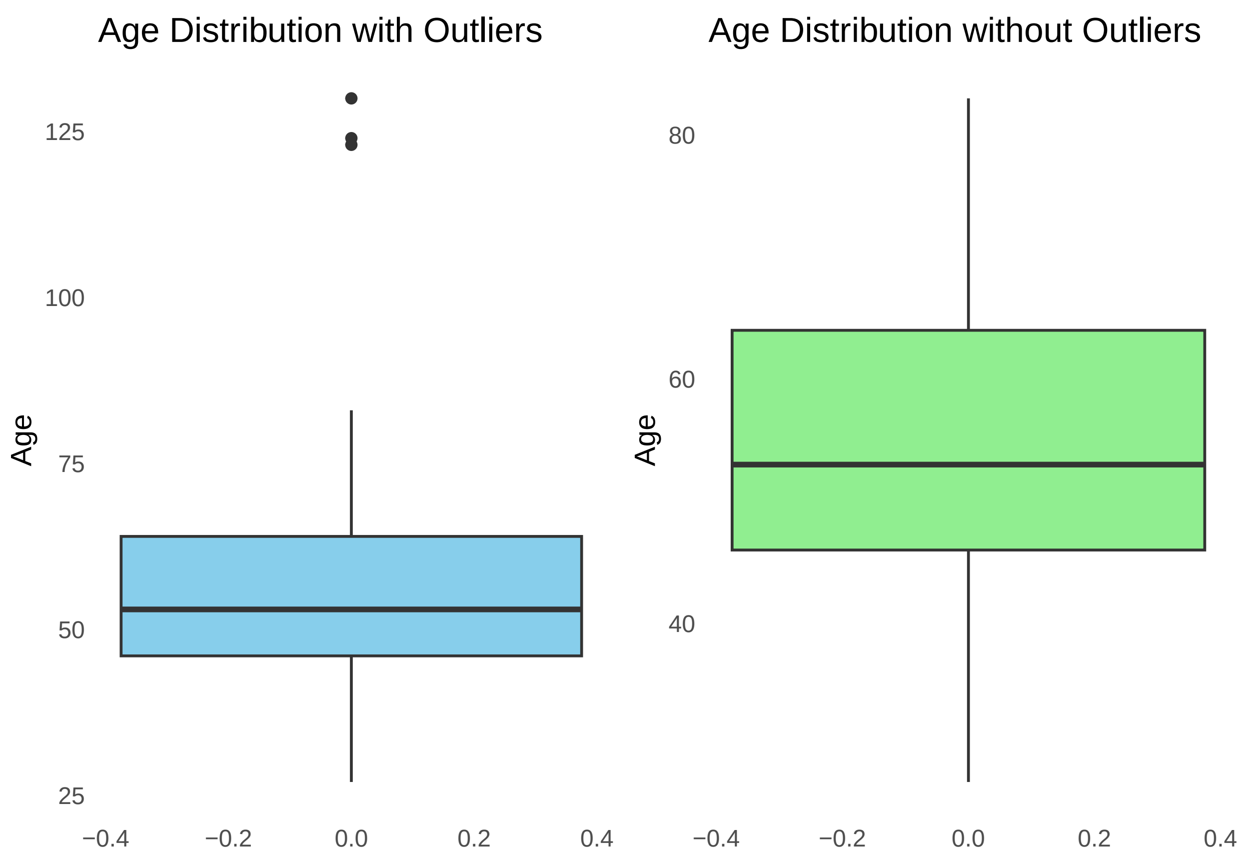
ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntW
5524	1957	Graduation	Single	58138	0	0	04-09-2012	58	
2174	1954	Graduation	Single	46344	1	1	08-03-2014	38	
4141	1965	Graduation	Together	71613	0	0	21-08-2013	26	

Univariate Analysis

Age Distribution

```r
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(gridExtra))
suppressPackageStartupMessages(library(grid))

p1 <- ggplot(csvData, aes(y = Age)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Age Distribution with Outliers", y = "Age") +
  theme_minimal()+
  theme(
     panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
```

```
  )

csvData <- csvData %>% filter(Age <= 100)

p2 <- ggplot(csvData, aes(y = Age)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Age Distribution without Outliers", y = "Age") +
  theme_minimal() +
  theme (
     panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )

grid.arrange(p1, p2, ncol = 2)
```

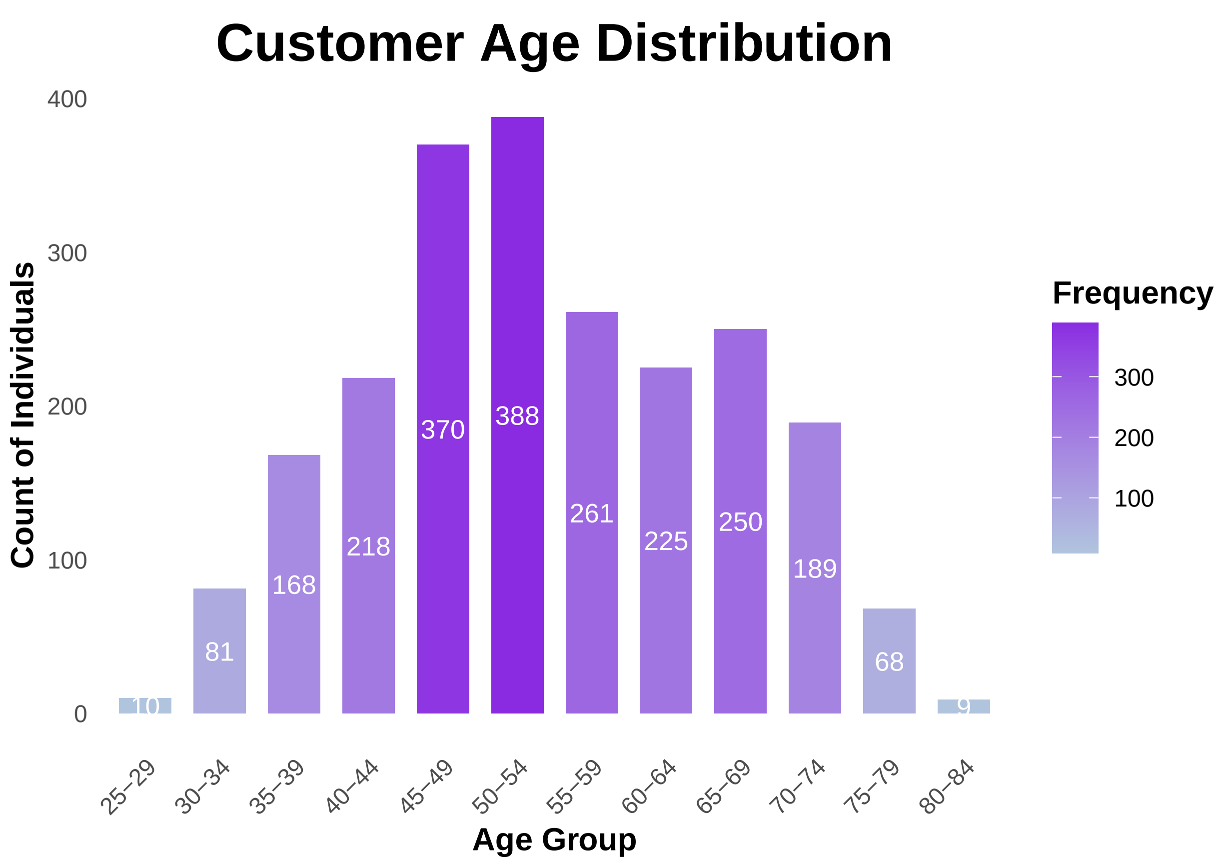


Now, let's calculate the age of each customer and create a bar plot that showcases the current overall distribution.

```
summary_without_outliers <- summary(csvData %>% filter(Age <= 100) %>% .$Age)

print(summary_without_outliers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    27.0    46.0    53.0    54.1    64.0    83.0
```

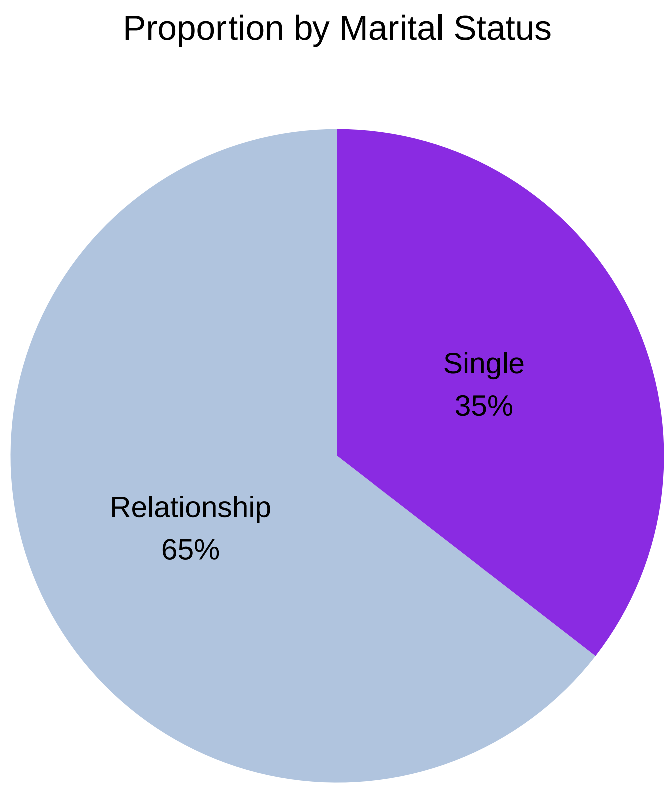
```r
age_bins <- seq(from = floor(min(csvData$Age)/5)*5, to = ceiling(max(csvData$Age)/5)*5, by = 5)

csvData <- csvData %>%
  mutate(Age_Group = cut(Age, breaks = age_bins, include.lowest = TRUE, right = FALSE, labels = paste(a

age_distribution_grouped <- csvData %>%
  count(Age_Group) %>%
  arrange(Age_Group)

ggplot(age_distribution_grouped, aes(x = Age_Group, y = n)) +
  geom_bar(aes(fill = n), stat = "identity", width = 0.7) +
  geom_text(aes(label = n), position = position_stack(vjust = 0.5), size = 3.5, color = "white") +
  labs(title = "Customer Age Distribution",
       x = "Age Group",
       y = "Count of Individuals",
       fill = "Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.title = element_text(size = 12, face = "bold")) +
  scale_fill_gradient(low = "lightsteelblue", high = "blueviolet", name = "Frequency")
```



Observation: The customer base presents a mature age profile with the median age at 53, indicating that the majority of customers are likely to be in a well-established stage of life. The age range is broad, extending from 27 to 83 years, with a concentration between 46 and 64, suggesting that products and services should be targeted towards the needs of middle-aged to senior adults.

Marital Status

```
csvData <- csvData %>%
  mutate(Marital_Status = case_when(
    Marital_Status %in% c('Married', 'Together') ~ 'Relationship',
    Marital_Status %in% c('Divorced', 'Widow', 'Alone', 'YOLO', 'Absurd') ~ 'Single',
    TRUE ~ Marital_Status
  ))

pie_chart <- csvData %>%
  count(Marital_Status) %>%
  mutate(Labels = paste0(Marital_Status, "\n", scales::percent(n/sum(n)))) %>%
  ggplot(aes(x = "", y = n, fill = Marital_Status, label = Labels)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(aes(label = Labels), position = position_stack(vjust = 0.5)) +
  coord_polar("y", start = 0) +
  labs(title = "Proportion by Marital Status", x = NULL, y = NULL) +
  theme_void() +
  scale_fill_manual(values = colorRampPalette(c("lightsteelblue", "blueviolet"))(length(unique(csvData$M
  theme(legend.position = "none",  panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5)

    )


grid.arrange(pie_chart, nrow = 1)
```



Observation: Customers are predominantly in a relationship, with 65% indicating a marital status of 'Married' or 'Together'. This suggests a customer base that may have considerations for family or partners in their purchasing decisions. In contrast, the single customers, including 'Divorced', 'Widow', 'Alone', 'YOLO', and 'Absurd', make up 35%, which may indicate a significant market for individual-centered products and services

Kidhome & Teenhome

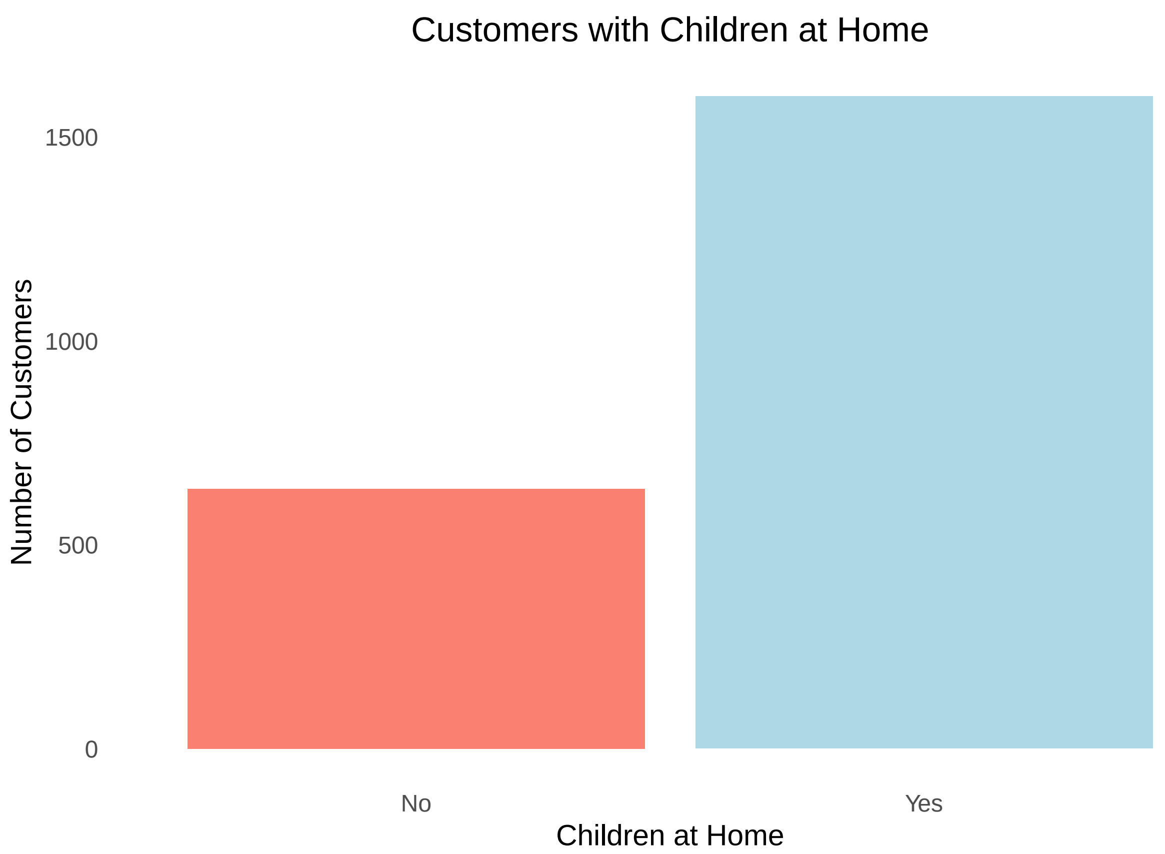
KidHome & TeenHome

```
children_home_summary <- csvData %>%
  mutate(Children_at_Home = if_else(Kidhome + Teenhome > 0, "Yes", "No")) %>%
  count(Children_at_Home) %>%
  mutate(Percentage = n / sum(n) * 100)

children_home_chart <- ggplot(children_home_summary, aes(x = Children_at_Home, y = n, fill = Children_a
  geom_bar(stat = "identity") +
  labs(title = "Customers with Children at Home", x = "Children at Home", y = "Number of Customers") +
  theme_minimal() +
  scale_fill_manual(values = c("Yes" = "lightblue", "No" = "salmon")) +
  theme(legend.position = "none",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
```

```
        plot.title = element_text(hjust = 0.5))

print(as.data.frame(children_home_summary), row.names = FALSE)
```

```
##  Children_at_Home    n Percentage
##                No  637   28.47564
##               Yes 1600   71.52436
```

```
grid.arrange(children_home_chart)
```




```
csvData <- csvData %>%
  mutate(TotalKids = Kidhome + Teenhome)

kids_freq_summary <- csvData %>%
  count(TotalKids) %>%
  mutate(Percentage = n / sum(n) * 100)

kids_freq_chart <- ggplot(kids_freq_summary, aes(x = as.factor(TotalKids), y = n, fill = as.factor(Tota
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Set3", direction = -1) +
  labs(title = "Number of Children in Customers' Households",
       x = "Number of Children",
       y = "Number of Customers",
       fill = "Number of Children") +
```

```r
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text.x = element_text(angle = 0, hjust = 1),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "none"
  )



print(as.data.frame(kids_freq_summary), row.names = FALSE)
```

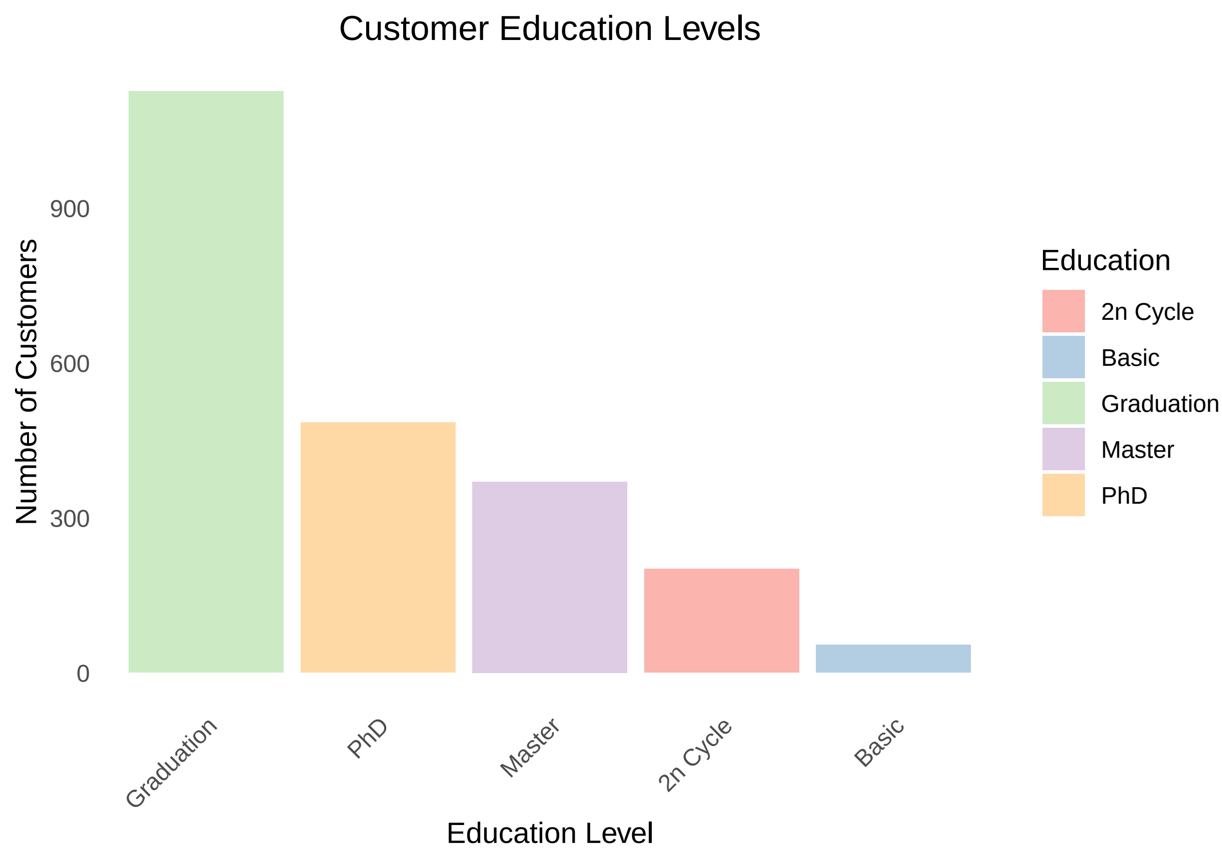
```
##  TotalKids    n Percentage
##          0  637  28.475637
##          1 1126  50.335270
##          2  421  18.819848
##          3   53   2.369245
```

Observation: The majority of customers (71.5%) have children at home, with half of these families having one child (50.3%). Families with two or more children represent a smaller portion of the market, indicating a higher concentration of smaller families within the customer base.

Education

Education

```r
education_chart <- csvData %>%
  count(Education) %>%
  ggplot(aes(x = reorder(Education, -n), y = n, fill = Education)) +
  geom_bar(stat = "identity") +
  labs(title = "Customer Education Levels", x = "Education Level", y = "Number of Customers") +
  theme_minimal() +
  scale_fill_brewer(palette = "Pastel1") +
  theme(plot.title = element_text(hjust = 0.5),
         panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
        axis.text.x = element_text(angle = 45, hjust = 1))

print(education_chart)
```




```
education_summary <- csvData %>%
  count(Education) %>%
  mutate(Percentage = n / sum(n) * 100)

print(as.data.frame(education_summary), row.names = FALSE)
```

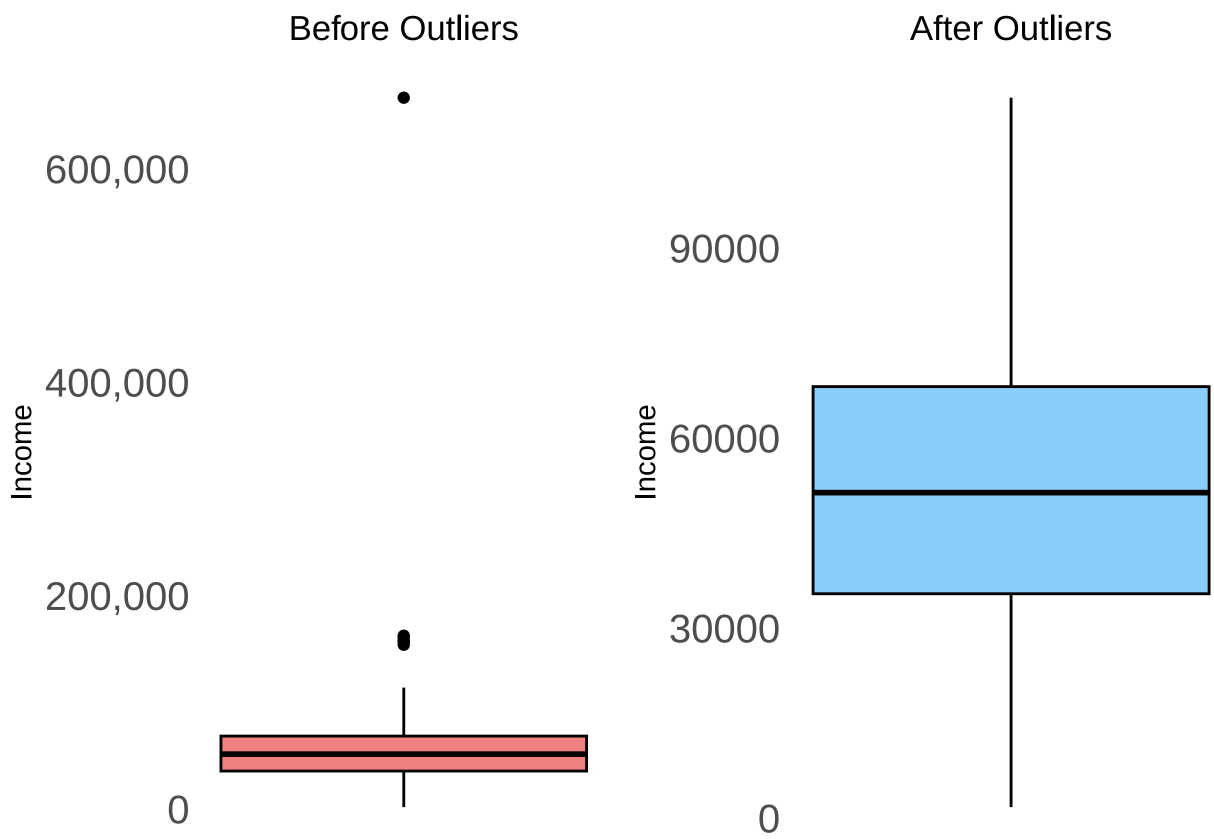
```
##   Education    n Percentage
##    2n Cycle  201   8.985248
##       Basic   54   2.413947
##  Graduation 1127  50.379973
##      Master  370  16.540009
##         PhD  485  21.680823
```

Observation: The education level among customers is predominantly 'Graduation' at over 50%, indicating that the majority have completed a degree equivalent to a college education. Postgraduate degrees (Masters and PhD) are also significant, accounting for approximately 38% combined, which suggests a well-educated customer base. Only a small fraction have 'Basic' education at 2.4%, and '2n Cycle' represents just under 9%

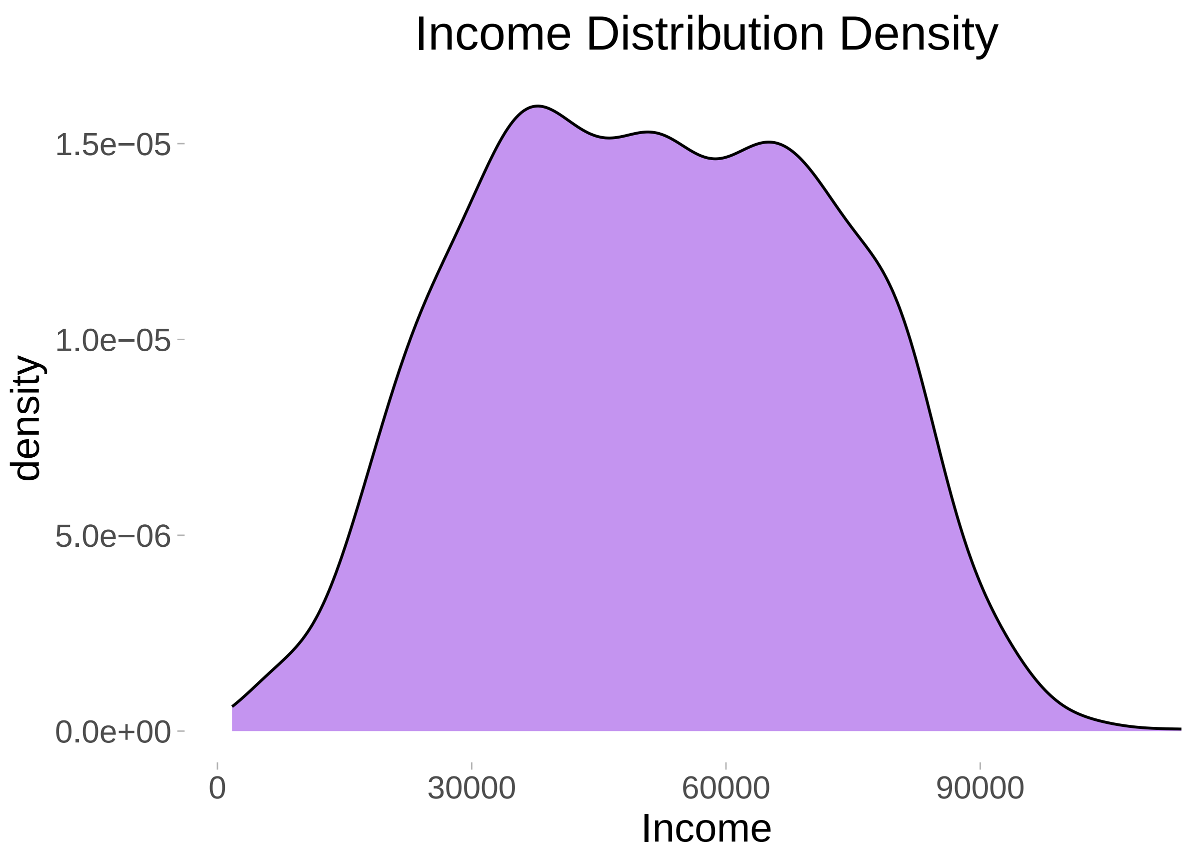
Income

Income

```r
common_theme <- theme(
  axis.title.x = element_blank(),
  axis.text.x = element_blank(),
  axis.ticks.x = element_blank(),
  axis.text.y = element_text(size = 15),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  plot.background = element_blank() ,
  plot.title = element_text(hjust = 0.5)

)

bp_before <- ggplot(csvData, aes(y = Income)) +
  geom_boxplot(fill = "lightcoral", color = "black", width = 0.7) +
  theme_minimal() +
  common_theme +
  ylab("Income") +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Before Outliers")

csvData <- csvData %>%
  filter(Income < 120000)

bp_after <- ggplot(csvData, aes(y = Income)) +
  geom_boxplot(fill = "lightskyblue", color = "black", width = 0.7) +
  theme_minimal() +
  common_theme +
  labs(y = "Income", title = "After Outliers")

grid.arrange(bp_before, bp_after, ncol = 2)
```




```r
ggplot(csvData, aes(x = Income)) +
  geom_density(fill = "blueviolet", alpha = 0.5) +
  labs(x = "Income", title = "Income Distribution Density") +
  theme_light() +
  theme(text = element_text(size = 15) ,
    panel.border = element_blank(),
    panel.background = element_blank(),
    panel.grid.major = element_blank(),
    plot.title = element_text(hjust = 0.5),
    panel.grid.minor = element_blank())
```




```r
income_summary <- csvData %>%
  mutate(Income_Range = cut(Income, breaks = seq(0, 120000, by = 20000), include.lowest = TRUE)) %>%
  count(Income_Range) %>%
  mutate(Percentage = n / sum(n) * 100)

print(as.data.frame(income_summary), row.names = FALSE)
```

```
##      Income_Range   n Percentage
##         [0,2e+04] 127   5.6976223
##     (2e+04,4e+04] 604  27.0973531
##     (4e+04,6e+04] 667  29.9237326
##     (6e+04,8e+04] 623  27.9497533
##     (8e+04,1e+05] 203   9.1072230
##   (1e+05,1.2e+05]   5   0.2243158
```

Observation: The majority of customers have incomes between 40,000 to 80,000, encompassing approximately 58% of the customer base, indicative of a strong middle-class presence. The 20,000 to 40,000 income range is also significant, accounting for 27% of customers. Notably, high earners with incomes between 80,000 to 100,000 constitute 9% of the customer base, while those earning over 100,000 are relatively rare at 0.22%. This suggests that luxury or high-end products may cater to a smaller segment of the market.

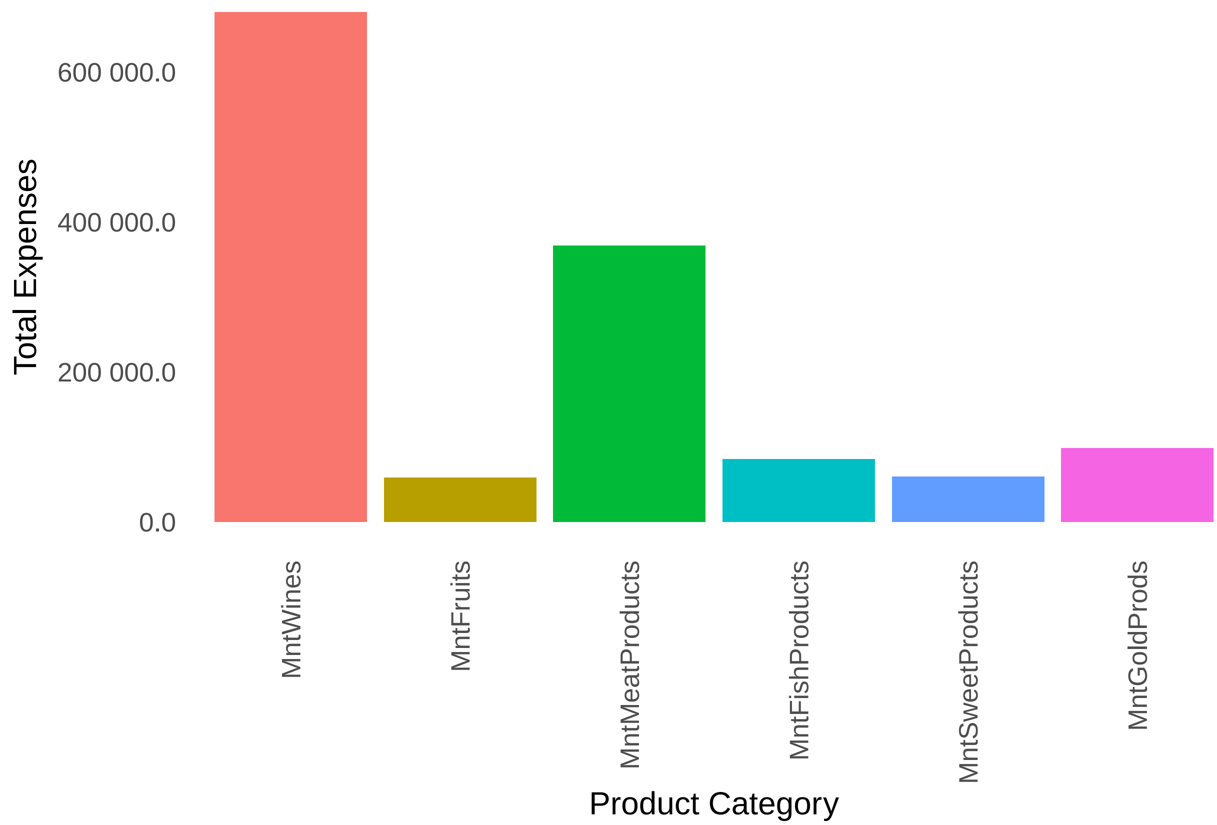
Expenses

Expenses

```
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(reshape2))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(scales))

data_long <- melt(csvData, id.vars = "ID",
                  measure.vars = c("MntWines", "MntFruits", "MntMeatProducts",
                                   "MntFishProducts", "MntSweetProducts", "MntGoldProds"),
                  variable.name = "Category", value.name = "Expenses")

category_summary <- data_long %>%
  group_by(Category) %>%
  summarise(Total_Expenses = sum(Expenses), .groups = 'drop') %>%
  mutate(Percentage = (Total_Expenses / sum(Total_Expenses)) * 100)

  print(as.data.frame(category_summary), row.names = FALSE)
```

```
##          Category Total_Expenses Percentage
##          MntWines         679826  50.366771
##         MntFruits          58731   4.351247
##   MntMeatProducts         368418  27.295257
##   MntFishProducts          83905   6.216332
##  MntSweetProducts          60543   4.485494
##      MntGoldProds          98328   7.284899
```

```
p <- ggplot(category_summary, aes(x = Category, y = Total_Expenses, fill = Category)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_x_discrete() +
  scale_y_continuous(labels = label_number(accuracy = 0.1)) +
  labs(x = "Product Category", y = "Total Expenses") +
  theme_minimal() +
  theme(
    text = element_text(size = 12),
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
    axis.title.y = element_text(vjust = 2),
    axis.text.y = element_text(angle = 0),
    panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
  legend.position = "none"

  )
print(p)
```



Observation: Wines are the leading category in customer expenses, accounting for over 50% of the total, which may suggest a customer base with a strong preference for this product. Meat products also hold a significant share at 27%, while other categories like fruits, fish, and sweets each represent less than 7% of the total expenses.

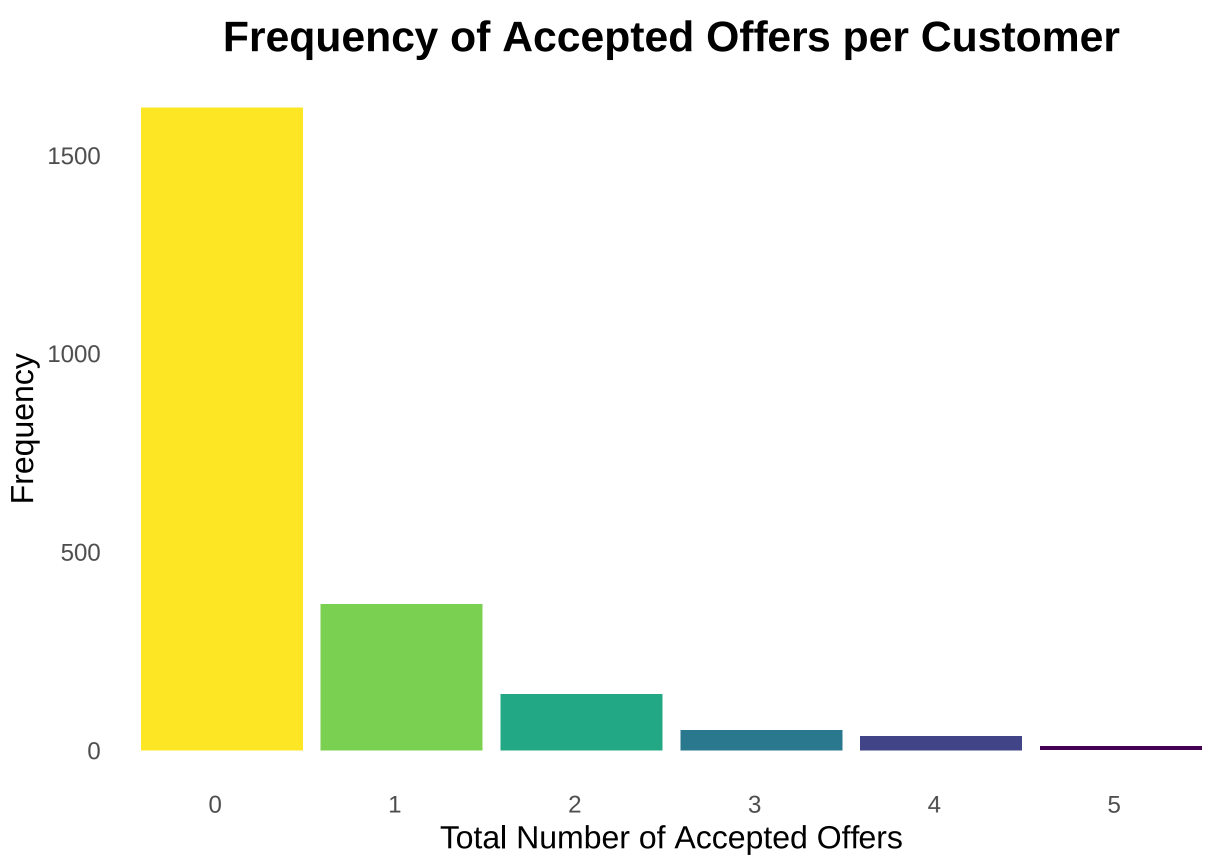
Campaigns

```r
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))


csvData <- csvData %>%
  mutate(TotalAcceptedCmp = AcceptedCmp1 + AcceptedCmp2 + AcceptedCmp3 +
                            AcceptedCmp4 + AcceptedCmp5 + Response)

accepted_freq <- csvData %>%
  count(TotalAcceptedCmp) %>%
  mutate(Percentage = n / sum(n) * 100)

print(as.data.frame(accepted_freq), row.names = FALSE)
```

```
##  TotalAcceptedCmp    n Percentage
```

```
##                    0 1621 72.7231943
##                    1  369 16.5545087
##                    2  142  6.3705698
##                    3   51  2.2880215
##                    4   36  1.6150740
##                    5   10  0.4486317
```

```r
ggplot(accepted_freq, aes(x = as.factor(TotalAcceptedCmp), y = n, fill = as.factor(TotalAcceptedCmp))) +
  geom_bar(stat = "identity") +
  scale_fill_viridis_d(direction = -1) +
  labs(title = "Frequency of Accepted Offers per Customer",
       x = "Total Number of Accepted Offers",
       y = "Frequency") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust=0.5),
    axis.title = element_text(size = 12),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text.x = element_text(hjust = 1),
    legend.position = "none"
  )
```



Observation: A vast majority of customers (approximately 73%) did not accept any campaign offers, which could indicate a challenge in the effectiveness of the campaigns or a generally low propensity to respond.

However, there is a small segment (about 17%) that engaged with one campaign, potentially representing a more responsive or interested customer group.

Purchase Places

```r
purchase_data_long <- csvData %>%
  select(NumWebPurchases, NumCatalogPurchases, NumStorePurchases) %>%
  pivot_longer(
    cols = everything(),
    names_to = "PurchasePlace",
    values_to = "NumberOfPurchases"
  ) %>%
  group_by(PurchasePlace) %>%
  summarise(TotalPurchases = sum(NumberOfPurchases), .groups = 'drop') %>%
  mutate(Percentage = TotalPurchases / sum(TotalPurchases) * 100)

print(as.data.frame(purchase_data_long), row.names = FALSE)
```

```
##        PurchasePlace TotalPurchases Percentage
##  NumCatalogPurchases           5877   21.01030
##    NumStorePurchases          12956   46.31775
##      NumWebPurchases           9139   32.67196
```

```r
ggplot(purchase_data_long, aes(x = PurchasePlace, y = TotalPurchases, fill = PurchasePlace)) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Dark2") +
  labs(title = "Purchases Made From Different Places",
       x = "",
       y = "Total Number of Purchases") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust= 0.5),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 12),
    axis.text.x = element_text(angle = 0, vjust = 1),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),

    legend.position = "none"
  )
```

Observation: Store purchases dominate the shopping venues, with nearly half of the purchases (46%) being made in-store. Web purchases account for roughly a third of the transactions, suggesting a significant online engagement, while catalog purchases are the least preferred method at 21%. This could indicate that despite the rise of digital platforms, physical stores remain a crucial point of sale.

Multivariate Analysis

Income vs Spendings

```
csvData <- csvData %>%
  mutate(Total_Spendings = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts

correlation_income_spendings <- cor(csvData$Income, csvData$Total_Spendings, use = "complete.obs")
cat("Correlation between Income and Total Spendings:", correlation_income_spendings, "\n")
```

```
## Correlation between Income and Total Spendings: 0.8202215
```

```r
p_income_spendings <- ggplot(csvData, aes(x = Income, y = Total_Spendings)) +
  geom_point(color = "blueviolet", alpha = 0.5) +
  labs(
    title = "Income vs Spendings",
    x = "Income",
    y = "Spendings"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 22, face = "bold", hjust = 0.5),
    axis.title.x = element_text(size = 18),
    axis.title.y = element_text(size = 18),
    axis.text = element_text(size = 12),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1),
    legend.position = "none"
  ) +
  scale_x_continuous(labels = dollar_format()) +
  scale_y_continuous(labels = dollar_format())

print(p_income_spendings)
```

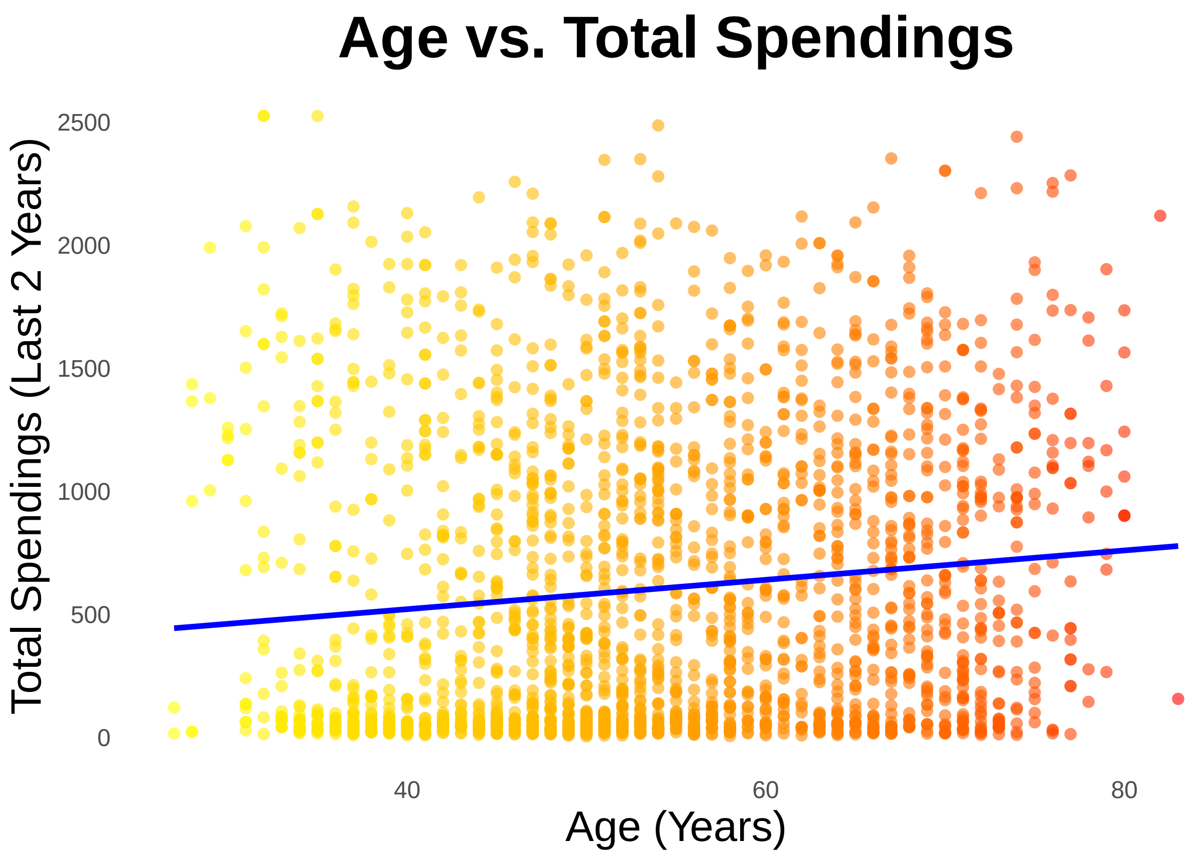
Observation: There is a strong positive correlation (0.82) between income and spending, suggesting that customers with higher incomes tend to spend more, which is indicative of significant purchasing power and could inform targeted marketing strategies.

Age vs Spendings

```
correlation_age_spendings <- cor(csvData$Age, csvData$Total_Spendings, use = "complete.obs")
cat("Correlation between Age and Total Spendings:", correlation_age_spendings, "\n")
```

```
## Correlation between Age and Total Spendings: 0.1160903
```

```
p_age_spendings <- ggplot(csvData, aes(x = Age, y = Total_Spendings)) +
  geom_point(aes(color = Age), alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  scale_color_gradient(low = "yellow", high = "red") +
  labs(
    title = "Age vs. Total Spendings",
    x = "Age (Years)",
    y = "Total Spendings (Last 2 Years)"
  ) +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(size = 22, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 16),
    legend.position = "none"
  )

print(p_age_spendings)
```



Observation: The correlation (0.12) between age and spending is relatively weak, indicating that age is not a strong predictor of spending among customers, and marketing efforts may be better focused on other demographic factors.

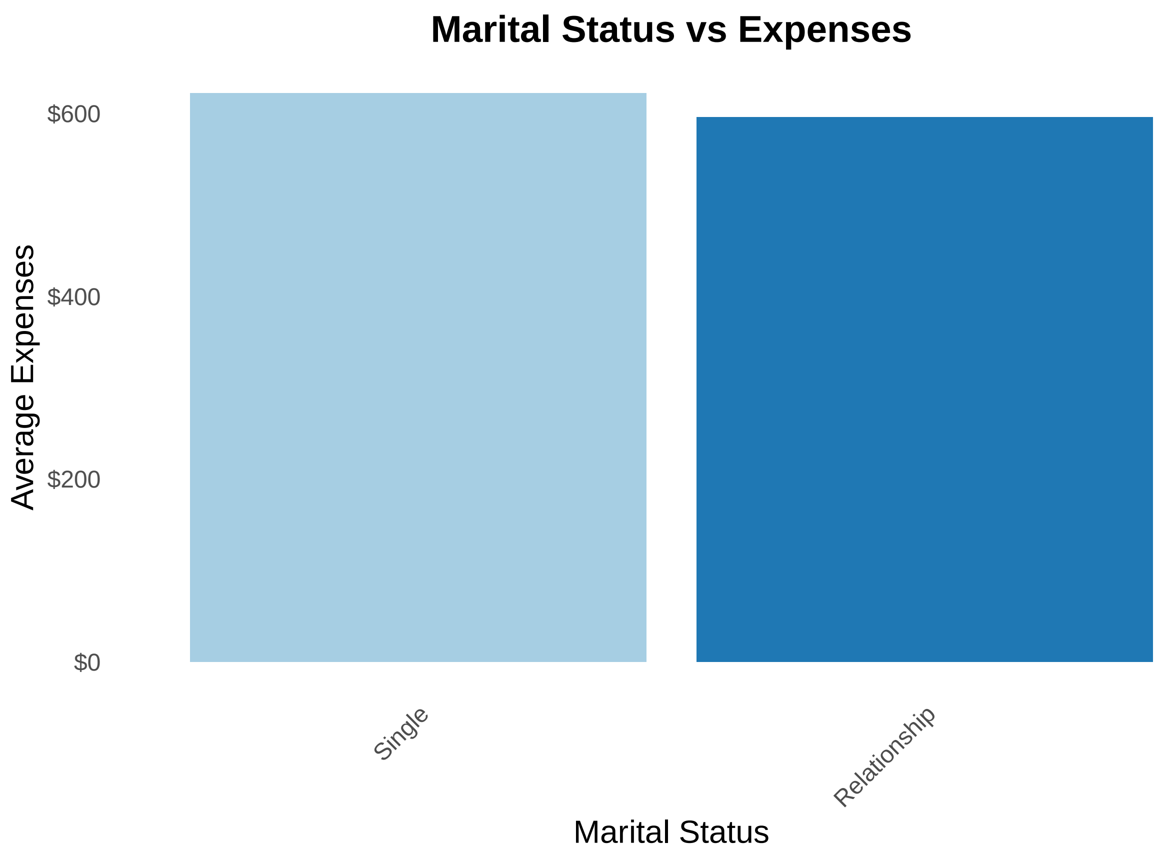
Martial Status vs Spendings

```
avg_spendings_by_marital <- csvData %>%
  group_by(Marital_Status) %>%
  summarise(Average_Spendings = mean(Total_Spendings, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(Marital_Status = factor(Marital_Status, levels = Marital_Status[order(-Average_Spendings)]))

 print(as.data.frame(avg_spendings_by_marital), row.names = FALSE)
```

```
##  Marital_Status Average_Spendings
##    Relationship          596.2214
##          Single          622.4174
```

```
  ggplot(avg_spendings_by_marital, aes(x = Marital_Status, y = Average_Spendings, fill = Marital_Status
    geom_bar(stat = "identity") +
```

```r
  scale_fill_brewer(palette = "Paired") +
  labs(title = "Marital Status vs Expenses",
     x = "Marital Status",
     y = "Average Expenses") +
  theme_minimal() +
  theme(
     panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(size = 14, face = "bold", hjust=0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  ) +
  scale_y_continuous(labels = scales::dollar)
```



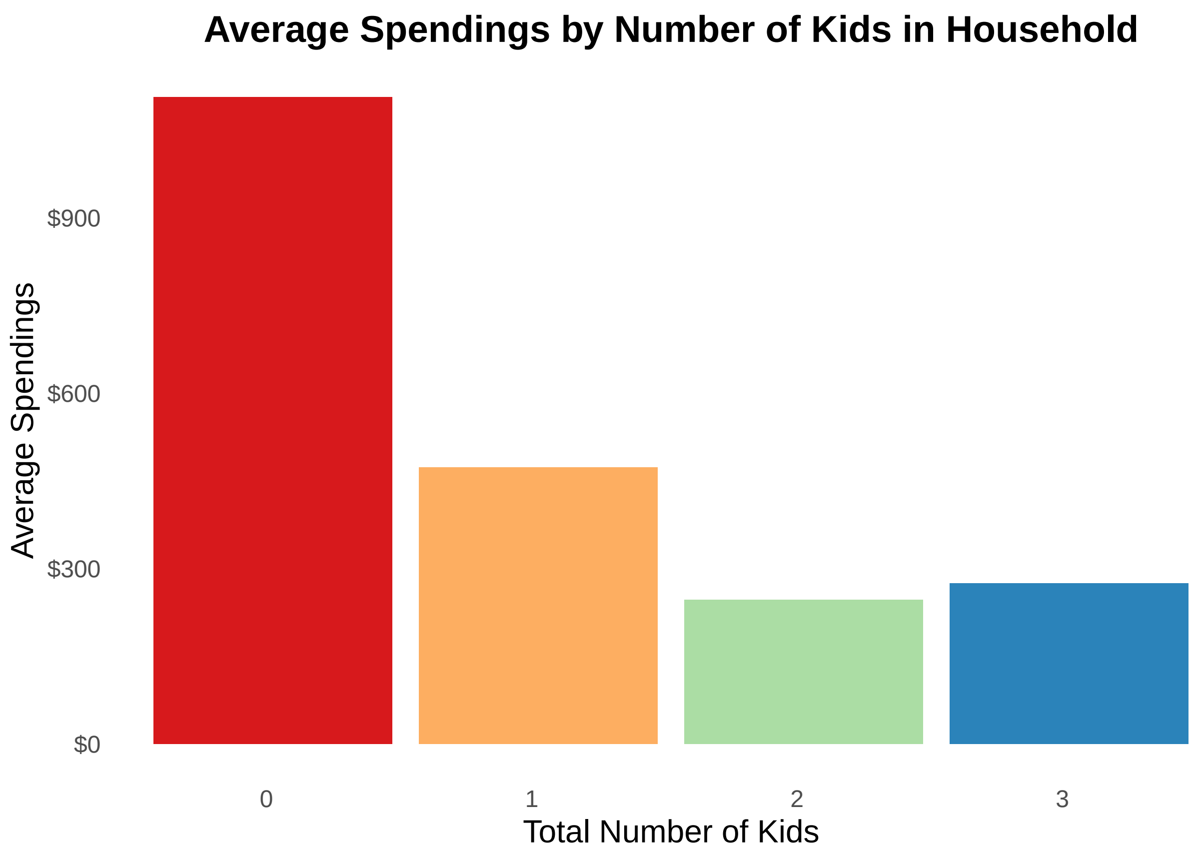
Observation: Interestingly, individuals with a status of 'Single' exhibit higher average spending (approximately \$622) compared to those in a Relationship (nearly \$596). This could suggest that single customers have more discretionary spending or different purchasing habits that favor more spending on the categories considered, perhaps due to fewer household obligations or different lifestyle choices.

How Having Kids Effect Expenses

```
csvData <- csvData %>%
  mutate(TotalKids = factor(Kidhome + Teenhome))

avg_spendings_by_kids <- csvData %>%
  group_by(TotalKids) %>%
  summarise(Average_Spendings = mean(Total_Spendings, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(Average_Spendings))

 print(as.data.frame(avg_spendings_by_kids), row.names = FALSE)
```

```
##  TotalKids Average_Spendings
##          0         1106.3712
##          1          473.2208
##          3          274.6038
##          2          246.2786
```

```
ggplot(avg_spendings_by_kids, aes(x = TotalKids, y = Average_Spendings, fill = TotalKids)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  scale_fill_brewer(palette = "Spectral") +
  labs(title = "Average Spendings by Number of Kids in Household",
       x = "Total Number of Kids",
       y = "Average Spendings") +
  theme_minimal() +
  theme(
     panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(size = 14, face = "bold", hjust =0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text.x = element_text(angle = 0, hjust = 1),
    legend.position = "none"

  ) +
  scale_y_continuous(labels = scales::dollar_format())
```



Observation: Customers without children lead with the highest average spending, over $1100, potentially reflecting greater disposable income and freedom to spend on personal indulgences. The presence of children correlates with a marked decrease in spending, where those with one child spend about $473, and it further declines as the number of children increases. This highlights a clear opportunity for child-centric marketing strategies and the potential to cater to the distinct needs of childless households.

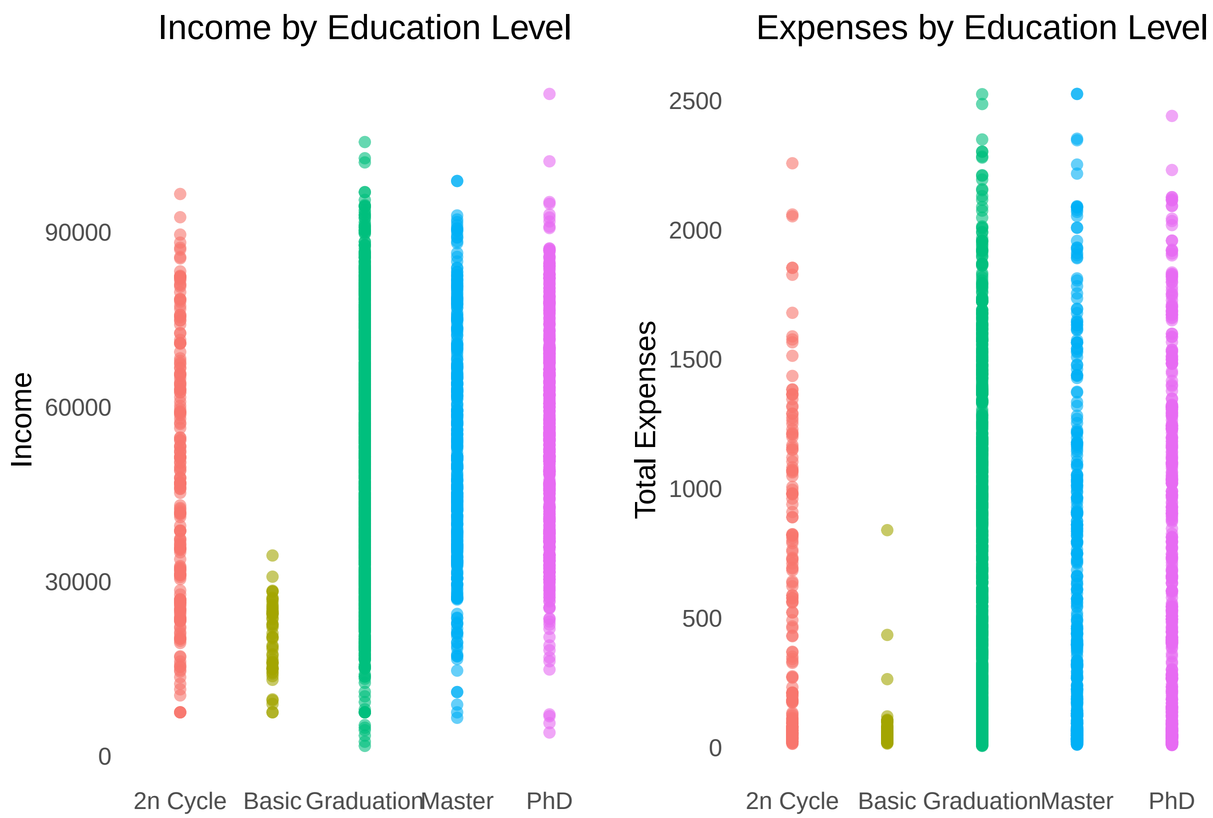
Education vs Income & Expenses

```
education_summary <- csvData %>%
  group_by(Education) %>%
  summarise(Average_Income = mean(Income, na.rm = TRUE),
            Average_Expenses = mean(Total_Spendings, na.rm = TRUE)) %>%
  ungroup()

print(as.data.frame(education_summary), row.names = FALSE)
```

```
##   Education Average_Income Average_Expenses
##    2n Cycle       47681.40         501.0348
##       Basic       20306.26          81.7963
##  Graduation       51978.11         619.9537
##      Master       52612.67         613.2791
```

```
##          PhD        55180.67          668.3950
```

```
education_income_plot <- ggplot(csvData, aes(x = Education, y = Income, color = Education)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Income by Education Level", x = "", y = "Income") +
  theme_minimal() +
  theme(legend.position = "none", panel.grid.major = element_blank(), panel.grid.minor = element_blank(

education_expenses_plot <- ggplot(csvData, aes(x = Education, y = Total_Spendings, color = Education)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Expenses by Education Level", x = "" , y = "Total Expenses") +
  theme_minimal() +
  theme(legend.position = "none", panel.grid.major = element_blank(), panel.grid.minor = element_blank(

grid.arrange(education_income_plot, education_expenses_plot, ncol = 2)
```



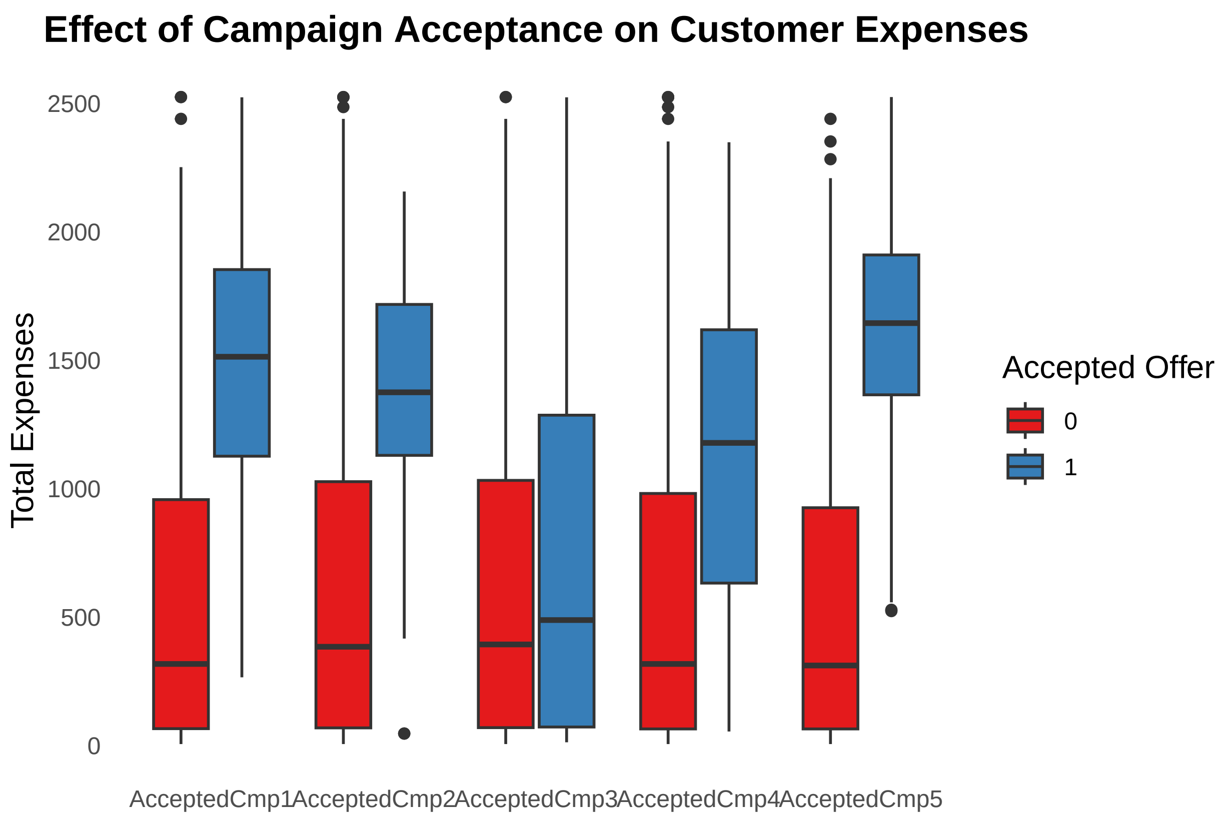
Observation: Individuals with a PhD not only exhibit the highest average income at approximately $55,181 but also the greatest average expenses, around $668, suggesting a correlation between educational attainment and spending capacity. In contrast, those with basic education have significantly lower average income and expenses, indicating that educational level is a strong predictor of economic behavior and potential in the marketplace.

Campaigns vs Expenses

```r
campaign_data_long <- csvData %>%
  select(Total_Spendings, starts_with("AcceptedCmp"), Response) %>%
  pivot_longer(cols = starts_with("AcceptedCmp"), names_to = "Campaign", values_to = "Accepted") %>%
  mutate(Campaign = factor(Campaign, levels = c("AcceptedCmp1", "AcceptedCmp2",
                                                "AcceptedCmp3", "AcceptedCmp4",
                                                "AcceptedCmp5", "Response")))

campaign_summary <- campaign_data_long %>%
  group_by(Campaign, Accepted) %>%
  summarise(Average_Expenses = mean(Total_Spendings, na.rm = TRUE)) %>%
  ungroup()

print(as.data.frame(campaign_summary), row.names = FALSE)
```

```
##     Campaign Accepted Average_Expenses
##  AcceptedCmp1        0         544.9933
##  AcceptedCmp1        1        1482.2222
##  AcceptedCmp2        0         595.9623
##  AcceptedCmp2        1        1307.6667
##  AcceptedCmp3        0         596.4681
##  AcceptedCmp3        1         720.5399
##  AcceptedCmp4        0         562.0024
##  AcceptedCmp4        1        1143.1257
##  AcceptedCmp5        0         526.4528
##  AcceptedCmp5        1        1614.6481
```

```r
campaign_expenses_plot <- ggplot(campaign_data_long, aes(x = Campaign, y = Total_Spendings, fill = as.fa
  geom_boxplot() +
  scale_fill_brewer(palette = "Set1") +
  labs(title = "Effect of Campaign Acceptance on Customer Expenses", x = "", y = "Total Expenses", fill
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5),
        plot.title = element_text(size = 14, face = "bold", hjust=0.5),
        legend.title = element_text(size = 12),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title = element_text(size = 12)) +
    scale_x_discrete()


print(campaign_expenses_plot)
```



Observation: The data reveals a striking insight: customers who accepted campaign offers, on average, spent markedly more than those who did not. For instance, those accepting the first campaign spent an average of $1,482 compared to $545 for those who did not. This pattern is consistent across campaigns, with AcceptedCmp5 showing the most significant difference — those who accepted this campaign spent over three times more than those who did not. This underscores the effectiveness of marketing campaigns in driving higher expenditure among responsive customers.

Age vs Product Type

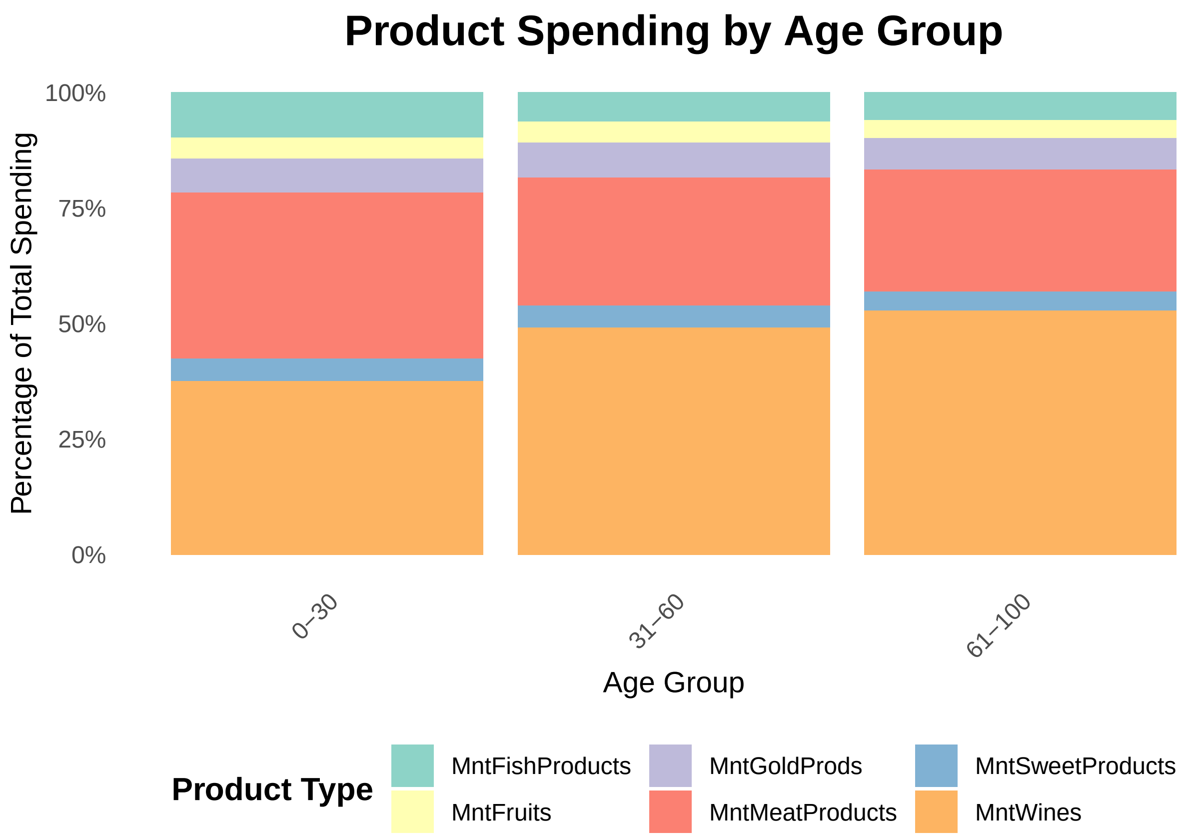
```
csvData <- csvData %>%
  mutate(Age_Group = case_when(
    Age >= 0 & Age <= 30   ~ "0-30",
    Age > 30 & Age <= 60   ~ "31-60",
    Age > 60 & Age <= 100  ~ "61-100",
    TRUE                   ~ "Out of range"
  ))

melted_data <- csvData %>%
  select(ID, Age_Group, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGol
  pivot_longer(
    cols = starts_with("Mnt"),
```

```r
    names_to = "Product_Type",
    values_to = "Amount_Spent"
  )

age_product_totals <- melted_data %>%
  group_by(Age_Group, Product_Type) %>%
  summarise(Total_Amount_Spent = sum(Amount_Spent), .groups = 'drop')

age_group_totals <- age_product_totals %>%
  group_by(Age_Group) %>%
  summarise(Total_Spent_by_Age_Group = sum(Total_Amount_Spent), .groups = 'drop')

age_product_percentages <- age_product_totals %>%
  left_join(age_group_totals, by = "Age_Group") %>%
  mutate(Percentage = Total_Amount_Spent / Total_Spent_by_Age_Group * 100) %>%
  select(Age_Group, Product_Type, Total_Amount_Spent, Total_Spent_by_Age_Group, Percentage)

print(age_product_percentages, row.names = FALSE)
```

```
## # A tibble: 18 x 5
##    Age_Group Product_Type   Total_Amount_Spent Total_Spent_by_Age_G~1 Percentage
##    <chr>     <chr>                       <dbl>                  <dbl>      <dbl>
##  1 0-30      MntFishProduc~               1406                  14272       9.85
##  2 0-30      MntFruits                     649                  14272       4.55
##  3 0-30      MntGoldProds                 1042                  14272       7.30
##  4 0-30      MntMeatProduc~               5127                  14272      35.9
##  5 0-30      MntSweetProdu~                691                  14272       4.84
##  6 0-30      MntWines                     5357                  14272      37.5
##  7 31-60     MntFishProduc~              52515                 834177       6.30
##  8 31-60     MntFruits                   38408                 834177       4.60
##  9 31-60     MntGoldProds                63102                 834177       7.56
## 10 31-60     MntMeatProduc~             231124                 834177      27.7
## 11 31-60     MntSweetProdu~              39275                 834177       4.71
## 12 31-60     MntWines                   409753                 834177      49.1
## 13 61-100    MntFishProduc~              29984                 501302       5.98
## 14 61-100    MntFruits                   19674                 501302       3.92
## 15 61-100    MntGoldProds                34184                 501302       6.82
## 16 61-100    MntMeatProduc~             132167                 501302      26.4
## 17 61-100    MntSweetProdu~              20577                 501302       4.10
## 18 61-100    MntWines                   264716                 501302      52.8
## # i abbreviated name: 1: Total_Spent_by_Age_Group
```

```r
ggplot(melted_data, aes(x = Age_Group, y = Amount_Spent, fill = Product_Type)) +
  geom_bar(stat = "summary", fun = "sum", position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Product Spending by Age Group",
       x = "Age Group",
       y = "Percentage of Total Spending",
       fill = "Product Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1),
        panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
```

```
        plot.title = element_text(size = 16, face = "bold",  hjust = 0.5),
        legend.title = element_text(size = 12, face = "bold"),
        legend.position = "bottom") +
  scale_fill_brewer(palette = "Set3")
```



Observation: The youngest age group (0-30) shows a balanced expenditure across different product types, with the highest percentages spent on meat products (35.9%) and wines (37.5%). For the middle-aged group (31-60), there's a notable preference for wines, which account for nearly half of their spending at 49.1%. Seniors (61-100) also demonstrate a similar pattern, dedicating a significant 52.8% of their spending to wines, highlighting a consistent trend across age groups favoring this category.

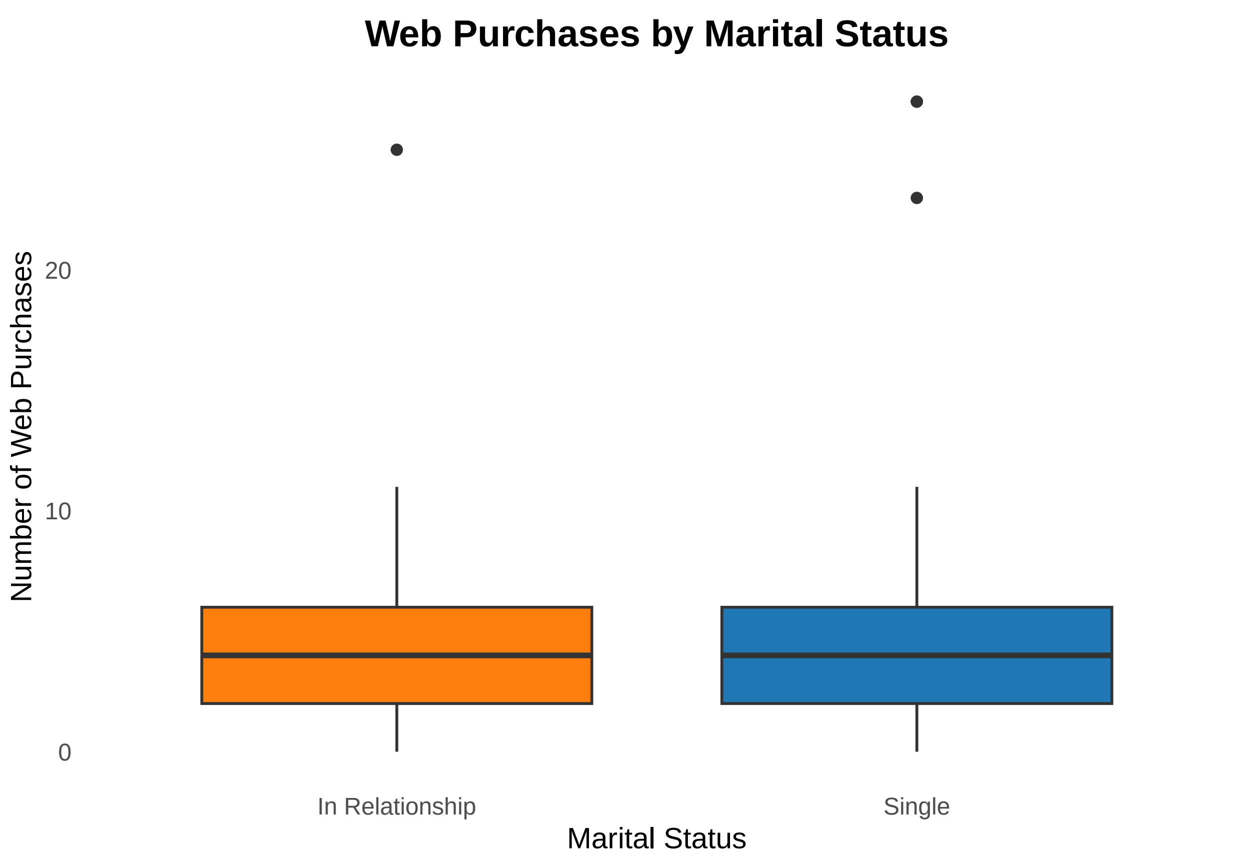
Martial Status vs Purchase Place

```
csvData <- csvData %>%
  mutate(Marital_Simplified = if_else(Marital_Status %in% c("Single", "Alone", "Divorced", "Widow"), "S

marital_online_stats <- csvData %>%
  group_by(Marital_Simplified) %>%
  summarise(Average_Online_Purchases = mean(NumWebPurchases, na.rm = TRUE)) %>%
  ungroup()
```

```r
print(as.data.frame(marital_online_stats), row.names = FALSE)
```

```
##  Marital_Simplified Average_Online_Purchases
##     In Relationship                 4.102368
##              Single                 4.095839
```

```r
ggplot(csvData, aes(x = Marital_Simplified, y = NumWebPurchases, fill = Marital_Simplified)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Single" = "#1f77b4", "In Relationship" = "#ff7f0e")) +
  labs(title = "Web Purchases by Marital Status",
       x = "Marital Status",
       y = "Number of Web Purchases") +
  theme_minimal() +
  theme(legend.position = "none" , panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(size = 14, face = "bold",  hjust = 0.5),

    )
```



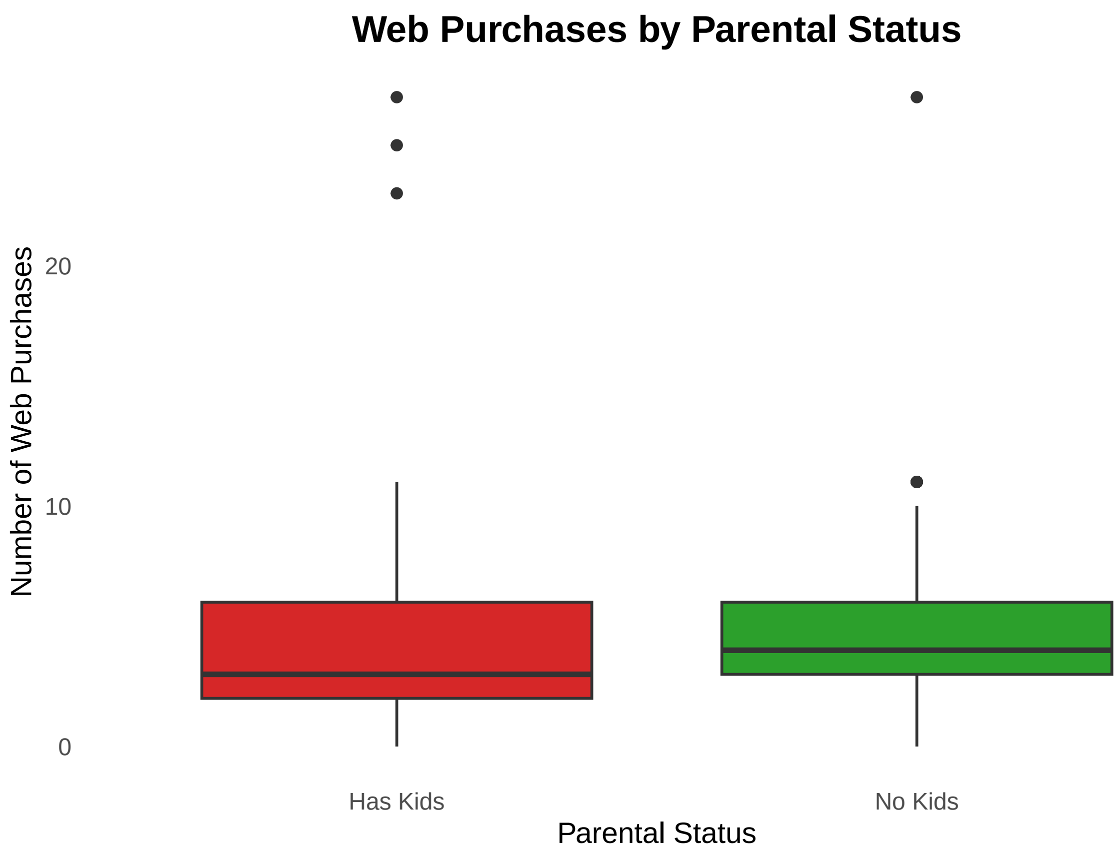
Observation: When analyzing online purchasing habits, those in a relationship appear to shop marginally more online with an average of 4.10 purchases compared to their single counterparts at 4.09. This slight difference suggests that marital status has minimal impact on the frequency of online purchases, indicating similar digital shopping behaviors between these two demographic segments.

Having Kids vs Purchase Place

```r
csvData <- csvData %>%
  mutate(Has_Kids = if_else(Kidhome + Teenhome > 0, "Has Kids", "No Kids"))

kids_online_stats <- csvData %>%
  group_by(Has_Kids) %>%
  summarise(Average_Online_Purchases = mean(NumWebPurchases, na.rm = TRUE)) %>%
  ungroup()

print(as.data.frame(kids_online_stats), row.names = FALSE)
```

```
##  Has_Kids Average_Online_Purchases
##  Has Kids                 3.972431
##   No Kids                 4.421801
```

```r
ggplot(csvData, aes(x = Has_Kids, y = NumWebPurchases, fill = Has_Kids)) +
  geom_boxplot() +
  scale_fill_manual(values = c("No Kids" = "#2ca02c", "Has Kids" = "#d62728")) +
  labs(title = "Web Purchases by Parental Status",
       x = "Parental Status",
       y = "Number of Web Purchases") +
  theme_minimal() +
  theme(legend.position = "none",  panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(size = 14, face = "bold",  hjust = 0.5),)
```



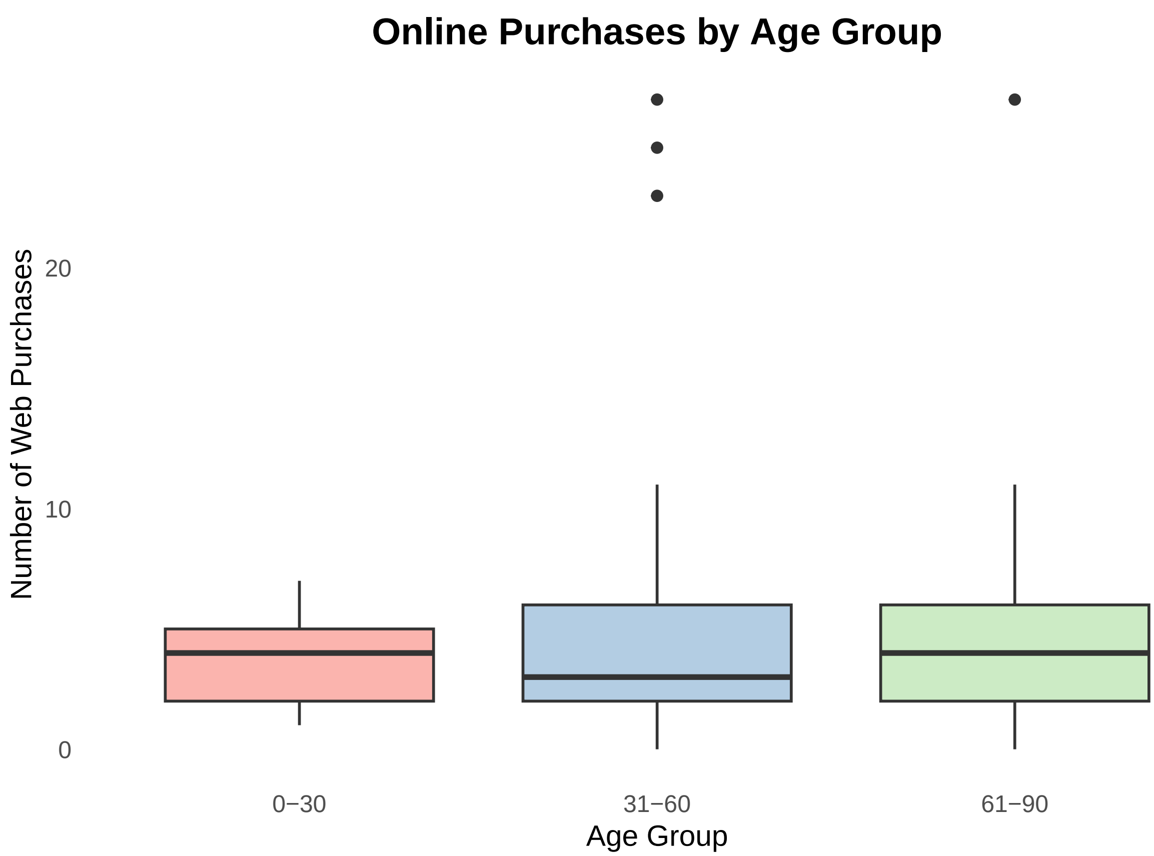
Observation: Interestingly, individuals without kids tend to make more online purchases (an average of 4.42) compared to those with kids (an average of 3.97). This could suggest that the time constraints and responsibilities of parenthood possibly influence the lower frequency of online shopping among parents.

Age Group vs Purchase Place

```
csvData <- csvData %>%
  mutate(Age = 2023 - Year_Birth,
         Age_Group = case_when(
           Age <= 30 ~ "0-30",
           Age > 30 & Age <= 60 ~ "31-60",
           Age > 60 ~ "61-90",
           TRUE ~ "Unknown"
         ))

age_online_stats <- csvData %>%
  group_by(Age_Group) %>%
  summarise(Average_Online_Purchases = mean(NumWebPurchases, na.rm = TRUE)) %>%
  ungroup()

print(as.data.frame(age_online_stats), row.names = FALSE)
```

```
##   Age_Group Average_Online_Purchases
##        0-30                 3.666667
##       31-60                 3.899276
##       61-90                 4.548201
```

```
ggplot(csvData, aes(x = Age_Group, y = NumWebPurchases, fill = Age_Group)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Online Purchases by Age Group",
       x = "Age Group",
       y = "Number of Web Purchases") +
  theme_minimal() +
  theme(legend.position = "none",  panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(size = 14, face = "bold",  hjust = 0.5),

    )
```



Observation: Online purchasing trends increase with age, with the youngest group (0-30) averaging 3.67 online purchases, the middle-aged (31-60) at 3.90, and the older group (61-90) leading at 4.55. This progression indicates a greater inclination towards online shopping as customers age, possibly due to higher disposable income or a preference for the convenience of online platforms.

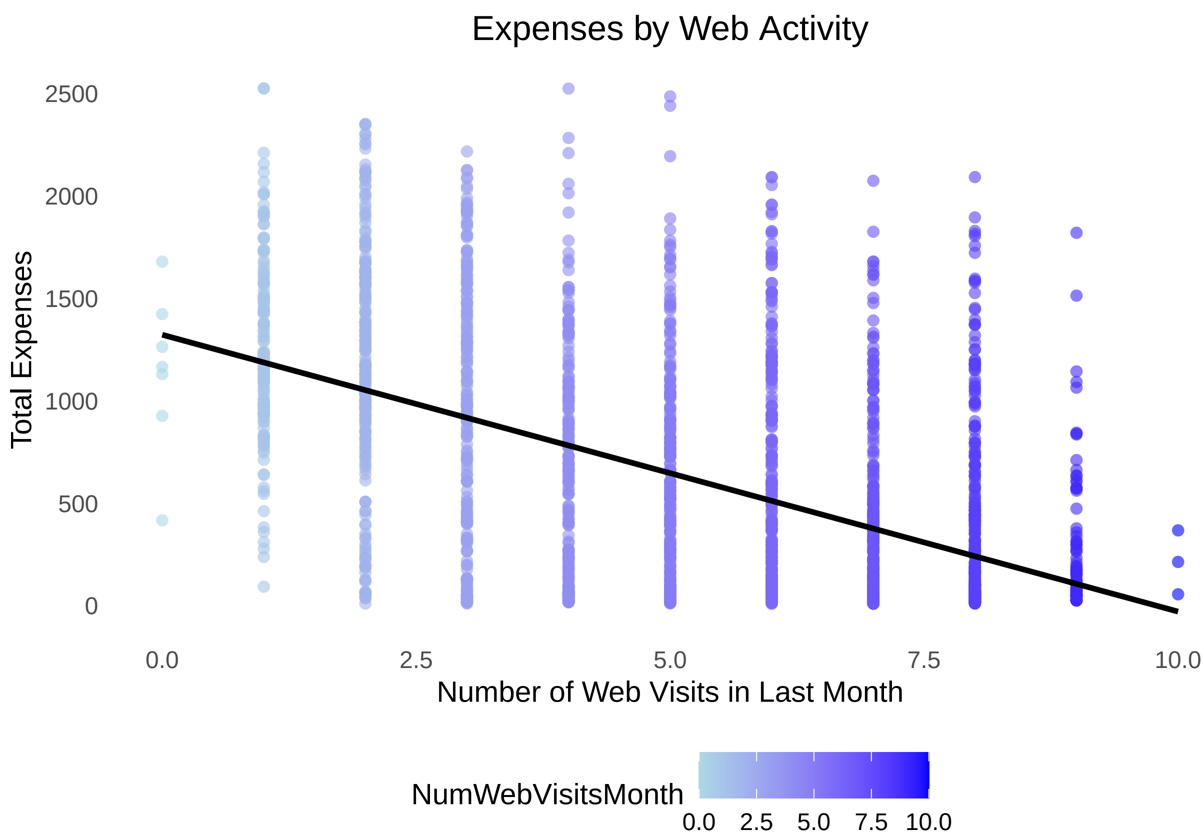
Expenses vs Web Activity

```r
csvData <- csvData %>%
  filter(Total_Spendings >= 0 & NumWebVisitsMonth <= 10)

# Group data by number of web visits per month and calculate the average expenses
web_activity_stats <- csvData %>%
  group_by(NumWebVisitsMonth) %>%
  summarise(Average_Expenses = mean(Total_Spendings, na.rm = TRUE)) %>%
  ungroup()

# Print the data frame without row names
print(as.data.frame(web_activity_stats), row.names = FALSE)
```

```
##  NumWebVisitsMonth Average_Expenses
##                  0        1143.1429
##                  1        1251.1275
##                  2        1186.3465
##                  3         973.9463
##                  4         653.3548
##                  5         522.6929
##                  6         486.7168
##                  7         297.1679
##                  8         353.1228
##                  9         288.1205
##                 10         211.6667
```

```r
# Create the plot with the filtered data
ggplot(csvData, aes(x = NumWebVisitsMonth, y = Total_Spendings, color = NumWebVisitsMonth)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  scale_color_gradient(low = "lightblue", high = "blue") +
  labs(title = "Expenses by Web Activity",
       x = "Number of Web Visits in Last Month",
       y = "Total Expenses") +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5)
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Observation: . It is observed that as the number of web visits increases up to 10 times a month, average expenses decrease. Customers with no web visits have the highest average spending of approximately $1,143. Conversely, those with 10 web visits per month have a lower average spending of roughly $212. This trend might suggest that customers who visit the website moderately are spending less on average than those with minimal to no online engagement. The highest average spending is associated with those who have the least online presence, which may imply that non-engaged customers are less exposed to online marketing efforts that could potentially drive higher spending.

Complaints vs Expenses

```r
complaints_stats <- csvData %>%
  group_by(Complain) %>%
  summarise(Average_Expenses = mean(Total_Spendings, na.rm = TRUE)) %>%
  ungroup()

print(as.data.frame(complaints_stats), row.names = FALSE)
```

```
##  Complain Average_Expenses
##         0         609.5973
##         1         392.0000
```

```r
ggplot(csvData, aes(x = factor(Complain), y = Total_Spendings, fill = factor(Complain))) +
  geom_boxplot() +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "blueviolet")) +
  labs(title = "Expenses by Complaint Status",
       x = "Complaint in Last 2 Years",
       y = "Total Expenses") +
  theme_minimal() +
  theme(legend.position = "none",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(size = 14, face = "bold",  hjust = 0.5))
```

Observation: Data indicates that customers who have not registered complaints in the last two years spend more on average, with their expenses amounting to $607, compared to $392 for those who have complained. This suggests a potential correlation between customer satisfaction and spending habits, highlighting the importance of addressing customer grievances to maintain higher spending levels.

Recency vs Expenses

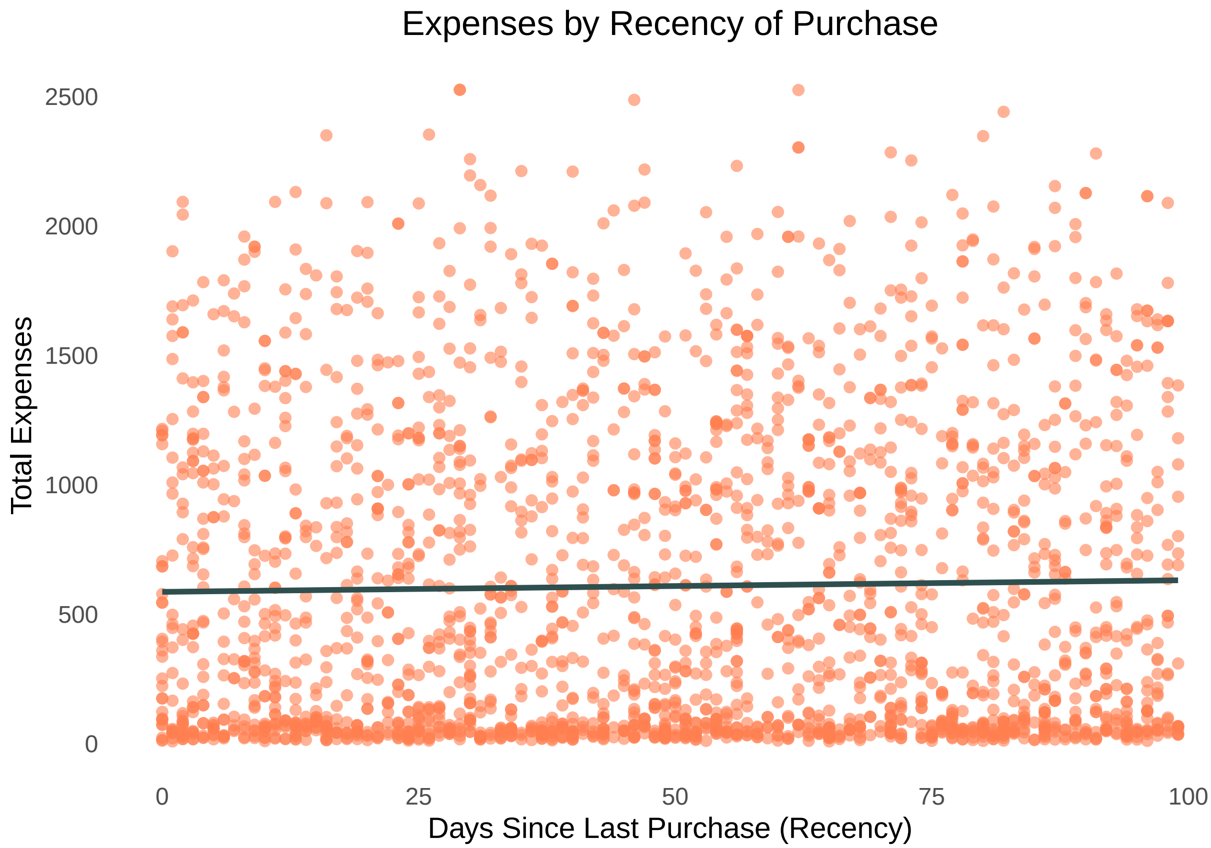
```r
csvData <- csvData %>%
  mutate(Total_Spendings = rowSums(select(., starts_with("Mnt"))))
```

```r
recency_expenses_stats <- csvData %>%
  group_by(Recency) %>%
  summarise(Average_Expenses = mean(Total_Spendings, na.rm = TRUE)) %>%
  ungroup()

print(recency_expenses_stats, row.names = FALSE)
```

```
## # A tibble: 100 x 2
##    Recency Average_Expenses
##      <dbl>            <dbl>
##  1       0             450.
##  2       1             661
##  3       2             628.
##  4       3             637.
##  5       4             665.
##  6       5             470.
##  7       6             666.
##  8       7             605.
##  9       8             710.
## 10       9             569.
## # i 90 more rows
```

```r
ggplot(csvData, aes(x = Recency, y = Total_Spendings)) +
  geom_point(alpha = 0.6, color = "coral") +
  geom_smooth(method = "lm", se = FALSE, color = "darkslategray") +
  labs(title = "Expenses by Recency of Purchase",
       x = "Days Since Last Purchase (Recency)",
       y = "Total Expenses") +
  theme_minimal() +
  theme(
    legend.position = "none",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5)
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Observation: There is a non-linear relationship between the recency of purchases and average expenses. Customers who have made a purchase very recently, like on day 0, have an average spending of $450. Interestingly, spending appears to fluctuate regardless of recency, with some peaks at various intervals — for instance, average spending is quite high at $865 for those who made a purchase 35 days ago. This pattern could suggest sporadic purchasing behavior influenced by factors other than just the time since the last purchase

Key Insights

- **Age Distribution:** A mature customer base with a median age of 53 suggests targeting middle-aged to senior adults.
- **Marital Status:** A majority in relationships indicates family-oriented purchasing decisions.
- **Kidhome & Teenhome:** A prevalence of smaller families highlights opportunities for child-centric marketing.
- **Education:** A highly educated customer base with over 50% having graduated suggests a focus on sophisticated products.
- **Income:** A strong middle-class presence with incomes primarily between $40,000 to $80,000 suggests a market for quality, yet affordable products.

- **Expenses:** A significant spend on wines and meats points to a customer preference for these categories.
- **Campaigns:** Low campaign acceptance rates challenge the effectiveness of marketing strategies.
- **Purchase Places:** Stores dominate purchases, but a significant one-third are online, showing the importance of a dual retail approach.
- **Income vs Spendings:** High income correlates with higher spending, indicating the potential for luxury marketing.
- **Age vs Spendings:** Age is a weaker predictor of spending, suggesting a diversified approach across age groups.
- **Marital Status vs Spendings:** Single customers spend slightly more, potentially indicating more disposable income.
- **Having Kids vs Expenses:** Childless customers have significantly higher average spending, indicating more discretionary spending.
- **Education vs Income & Expenses:** Higher education levels correlate with higher income and spending, underscoring targeted marketing for higher education levels.
- **Campaigns vs Expenses:** Customers who engage with campaigns tend to spend more, highlighting the success of targeted marketing campaigns.
- **Age vs Product Type:** Preferences for wines across age groups, with younger customers also spending significantly on meats.
- **Marital Status vs Purchase Place:** Minimal difference in online purchasing between singles and those in relationships.
- **Having Kids vs Purchase Place:** Those without kids shop more online, suggesting the influence of parenting responsibilities.
- **Age Group vs Purchase Place:** Older customers tend to make more online purchases, indicating a potential focus on online marketing for older demographics.
- **Expenses vs Web Activity:** Higher web activity correlates with lower spending, indicating a more deal-savvy or selective customer.
- **Complaints vs Expenses:** Customers who haven't complained spend more, linking customer satisfaction to spending.
- **Recency vs Expenses:** A complex relationship with no clear pattern, suggesting sporadic purchasing influenced by diverse factors.

In conclusion, the study reveals a customer profile that is mature, well-educated, and family-oriented with spending habits that gravitate towards quality products like wines and meats. Campaign responsiveness and customer satisfaction emerge as pivotal factors influencing spending, while web engagement and family dynamics demonstrate nuanced effects on purchasing behavior. These insights are crucial for tailoring marketing strategies to the distinct needs and preferences of the customer base.