Marketing Campaign Data Analysis

Abdallah chidjou

December 10, 2024

```
# Abdallah Chidjou
# Citation: (source of help: Lecture note, googling in general, stackoverflow, and chatgpt)
```

Introduction

In this analysis, we explore a dataset containing information from a marketing campaign. The dataset includes customer demographics, purchasing behavior, and responses to multiple campaigns. I aim to derive insights into customer characteristics and patterns that affect marketing success.

- Describe the dataset and its variables.
- Perform statistical analyses and visualizations to understand key characteristics.
- Explore relationships between variables, focusing on purchasing behavior and campaign responses.

The dataset used includes the following variables:

- ID: Customer ID
- Year_Birth: Year of birth
- Education: Level of education
- Marital_Status: Marital status
- Income: Annual income
- Kidhome: Number of kids in the household
- Teenhome: Number of teenagers in the household
- Dt_Customer: Date when the customer became enrolled
- Recency: Days since last purchase
- MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds: Money spent on various product categories
- NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth: Information about different purchasing channels and deals used
- AcceptedCmpX: Indicates whether the customer accepted a particular campaign (Cmp1, Cmp2, etc.)
- Complain: Whether the customer has complained in the past
- Z_CostContact: Fixed cost related to contact (likely constant)
- Z_Revenue: Revenue generated
- Response: Response to the most recent campaign
- (and others…; the folowing code illustrate the rest)

Data Loading and Preparation

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(tidyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
# Load the dataset
marketing_data <- read.csv("marketing_campaign.csv", sep = "\t")

# Display the first few rows of the dataset
head(marketing_data)
```

```
##     ID Year_Birth  Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957 Graduation         Single  58138       0        0  04-09-2012
## 2 2174       1954 Graduation         Single  46344       1        1  08-03-2014
## 3 4141       1965 Graduation       Together  71613       0        0  21-08-2013
## 4 6182       1984 Graduation       Together  26646       1        0  10-02-2014
## 5 5324       1981        PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967     Master       Together  62513       0        1  09-09-2013
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1           88                 3               8                  10
## 2            6                 2               1                   1
## 3           42                 1               8                   2
## 4            5                 2               2                   0
## 5           15                 5               5                   3
```

```
## 6           14                2              6                   4
##   NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1                 4                 7            0            0            0
## 2                 2                 5            0            0            0
## 3                10                 4            0            0            0
## 4                 4                 6            0            0            0
## 5                 6                 5            0            0            0
## 6                10                 6            0            0            0
##   AcceptedCmp1 AcceptedCmp2 Complain Z_CostContact Z_Revenue Response
## 1            0            0        0             3        11        1
## 2            0            0        0             3        11        0
## 3            0            0        0             3        11        0
## 4            0            0        0             3        11        0
## 5            0            0        0             3        11        0
## 6            0            0        0             3        11        0
```

```r
# Data Cleaning

# Remove duplicates
marketing_data <- marketing_data %>% distinct()

# Visualize duplicates removal
cat("Number of rows after removing duplicates: ", nrow(marketing_data), "\n")
```

```
## Number of rows after removing duplicates:  2240
```

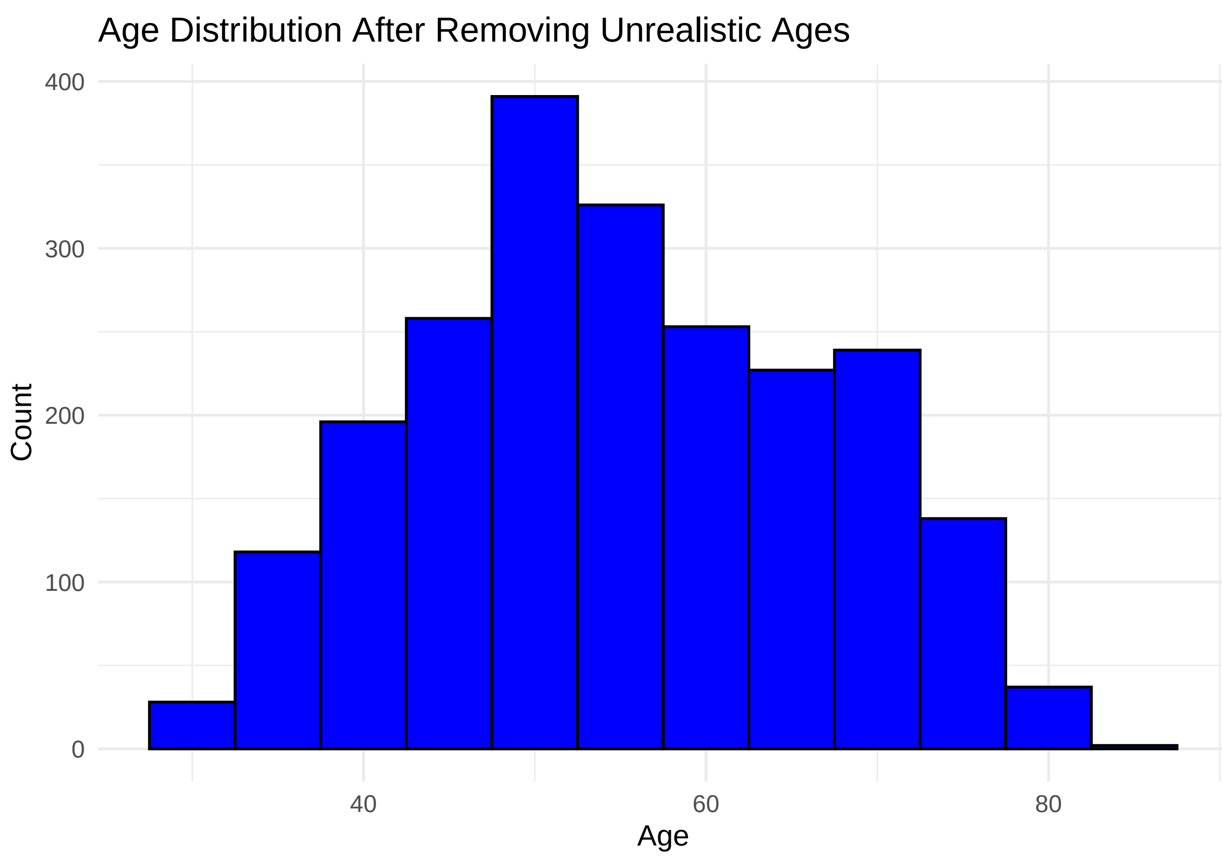
```r
# Handling missing values - removing rows with NA values in the columns
marketing_data <- marketing_data %>% drop_na()


# Check for missing values in the dataset
missing_values <- is.na(marketing_data)

# Get a summary of missing values for each column
colSums(is.na(marketing_data))
```

```
##                  ID          Year_Birth           Education      Marital_Status
##                   0                   0                   0                   0
##              Income             Kidhome            Teenhome         Dt_Customer
##                   0                   0                   0                   0
##             Recency            MntWines           MntFruits     MntMeatProducts
##                   0                   0                   0                   0
##     MntFishProducts    MntSweetProducts        MntGoldProds   NumDealsPurchases
##                   0                   0                   0                   0
##     NumWebPurchases NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth
##                   0                   0                   0                   0
##        AcceptedCmp3        AcceptedCmp4        AcceptedCmp5        AcceptedCmp1
##                   0                   0                   0                   0
##        AcceptedCmp2            Complain       Z_CostContact           Z_Revenue
##                   0                   0                   0                   0
##            Response
##                   0
```

```r
# Convert date columns to appropriate format
marketing_data$Dt_Customer <- dmy(marketing_data$Dt_Customer)

# Check date conversion
cat("Data type of Dt_Customer after conversion: ", class(marketing_data$Dt_Customer), "\n")
```

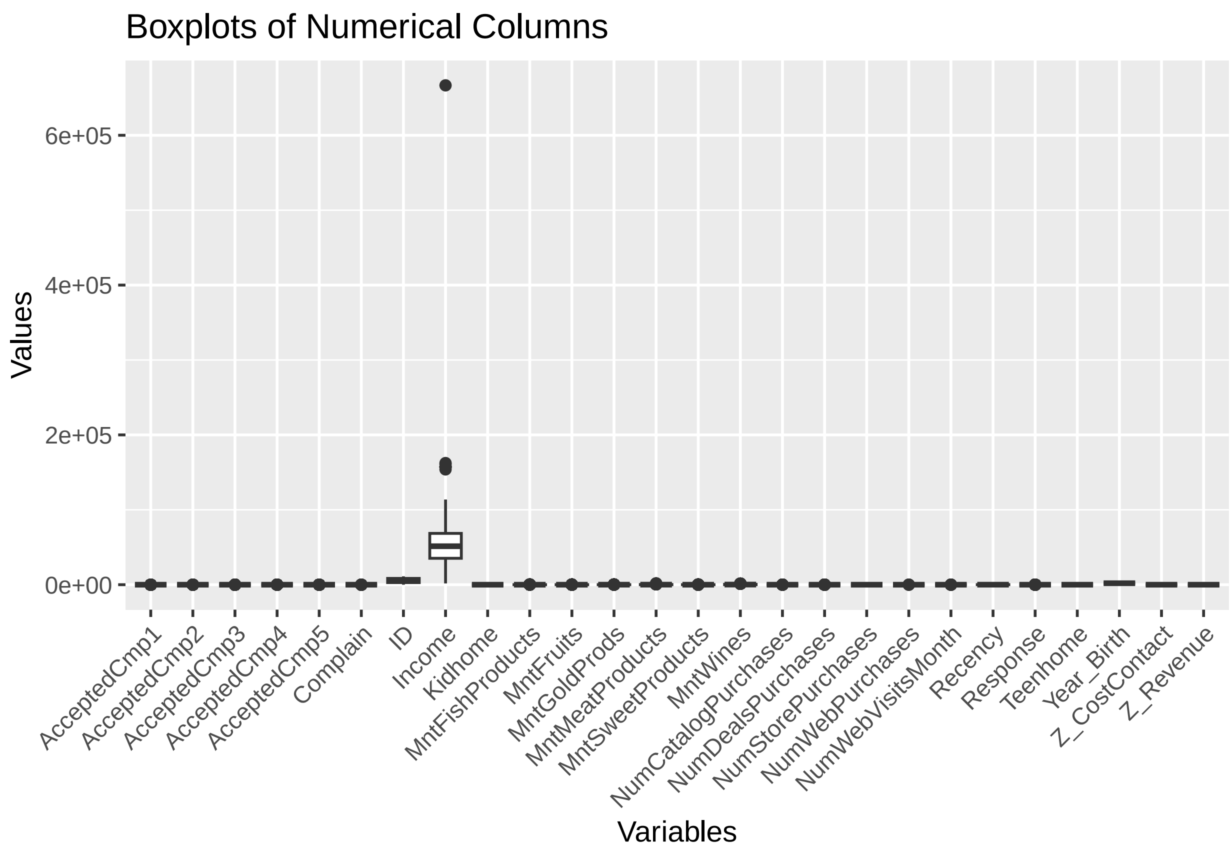
```
## Data type of Dt_Customer after conversion:  Date
```

```r
# Remove unrealistic ages (e.g., customers older than 100 years)
marketing_data <- marketing_data %>% filter(2024 - Year_Birth <= 100)

# Visualize age filtering
ggplot(marketing_data, aes(x = 2024 - Year_Birth)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Age Distribution After Removing Unrealistic Ages", x = "Age", y = "Count")
```

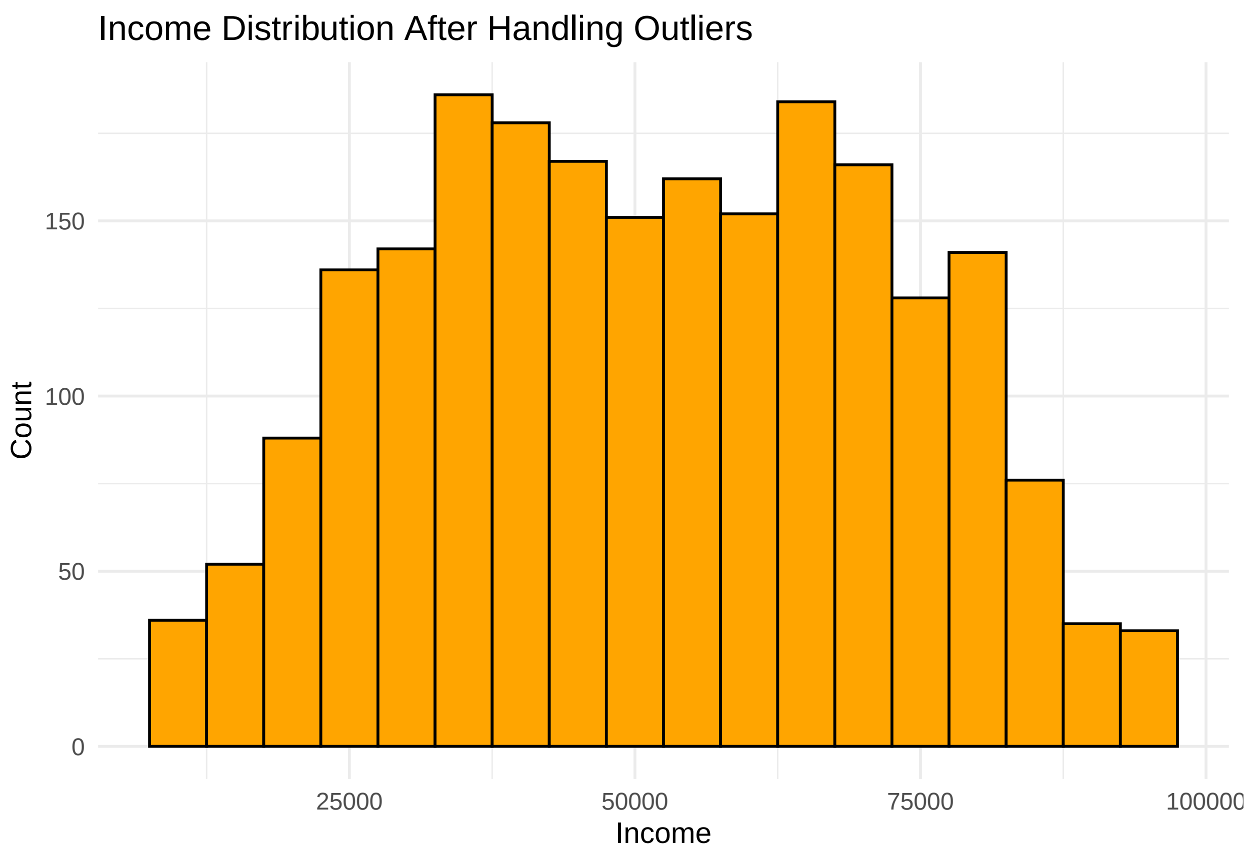



```r
# Convert the dataset from wide to long format so that we can easily boxplot all numerical columns
long_data <- marketing_data %>%
  pivot_longer(cols = where(is.numeric), names_to = "Variable", values_to = "Value")

# Create boxplot for all numerical columns
ggplot(long_data, aes(x = Variable, y = Value)) +
  geom_boxplot() +
```

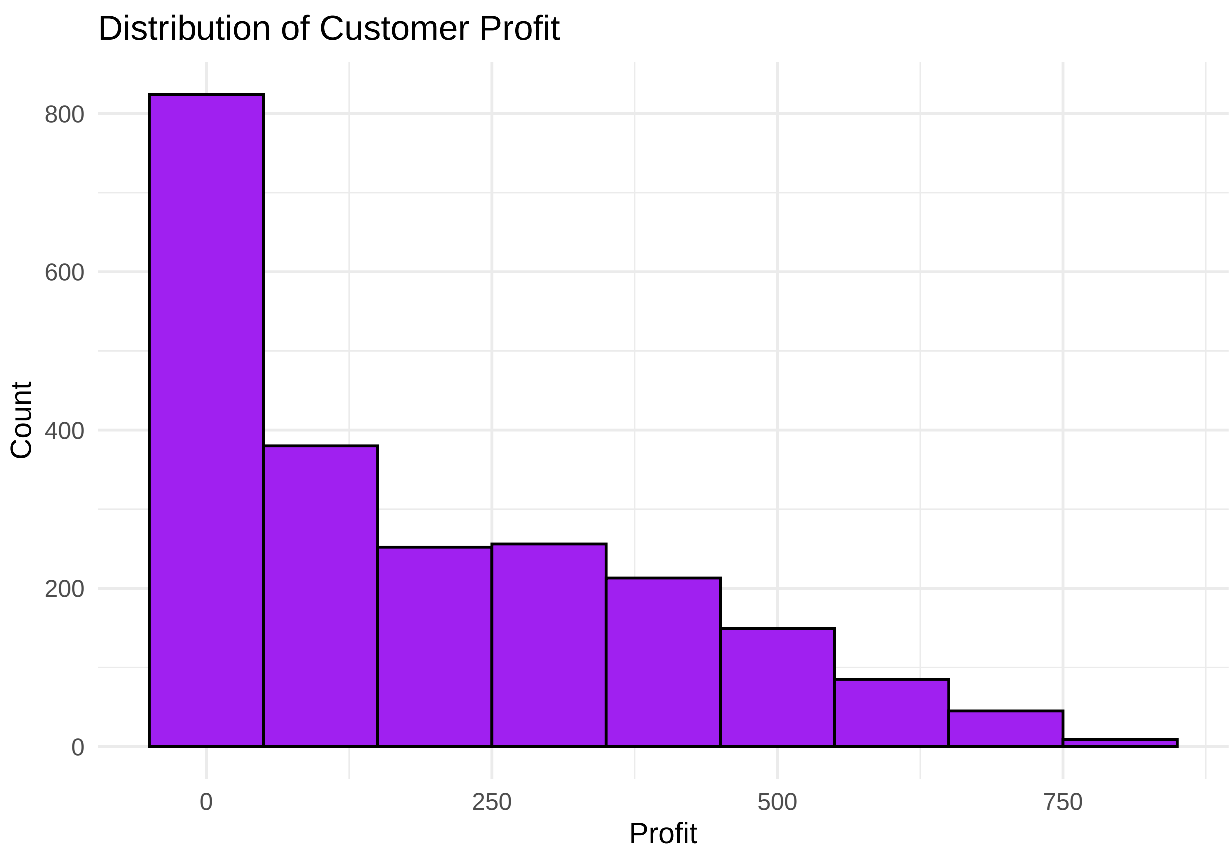
```
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Boxplots of Numerical Columns", x = "Variables", y = "Values")
```




```
# Handling outliers in Income by capping extreme values
income_quantiles <- quantile(marketing_data$Income, probs = c(0.01, 0.99))
marketing_data <- marketing_data %>% mutate(Income = ifelse(Income < income_quantiles[1], income_quanti
                                                ifelse(Income > income_quantiles[2], income_

# Visualize income outliers capping
ggplot(marketing_data, aes(x = Income)) +
  geom_histogram(binwidth = 5000, fill = "orange", color = "black") +
  theme_minimal() +
  labs(title = "Income Distribution After Handling Outliers", x = "Income", y = "Count")
```




```r
# Profit Calculation
# Assuming profit is derived from multiple categories of products
marketing_data <- marketing_data %>%
  mutate(Profit = MntWines * 0.3 + MntFruits * 0.2 + MntMeatProducts * 0.4 +
           MntFishProducts * 0.25 + MntSweetProducts * 0.15 + MntGoldProds * 0.35)

# Visualize profit distribution
ggplot(marketing_data, aes(x = Profit)) +
  geom_histogram(binwidth = 100, fill = "purple", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Customer Profit", x = "Profit", y = "Count")
```



Data Exploration

Descriptive Statistics

```r
# Summary statistics for numeric variables
summary(marketing_data)
```

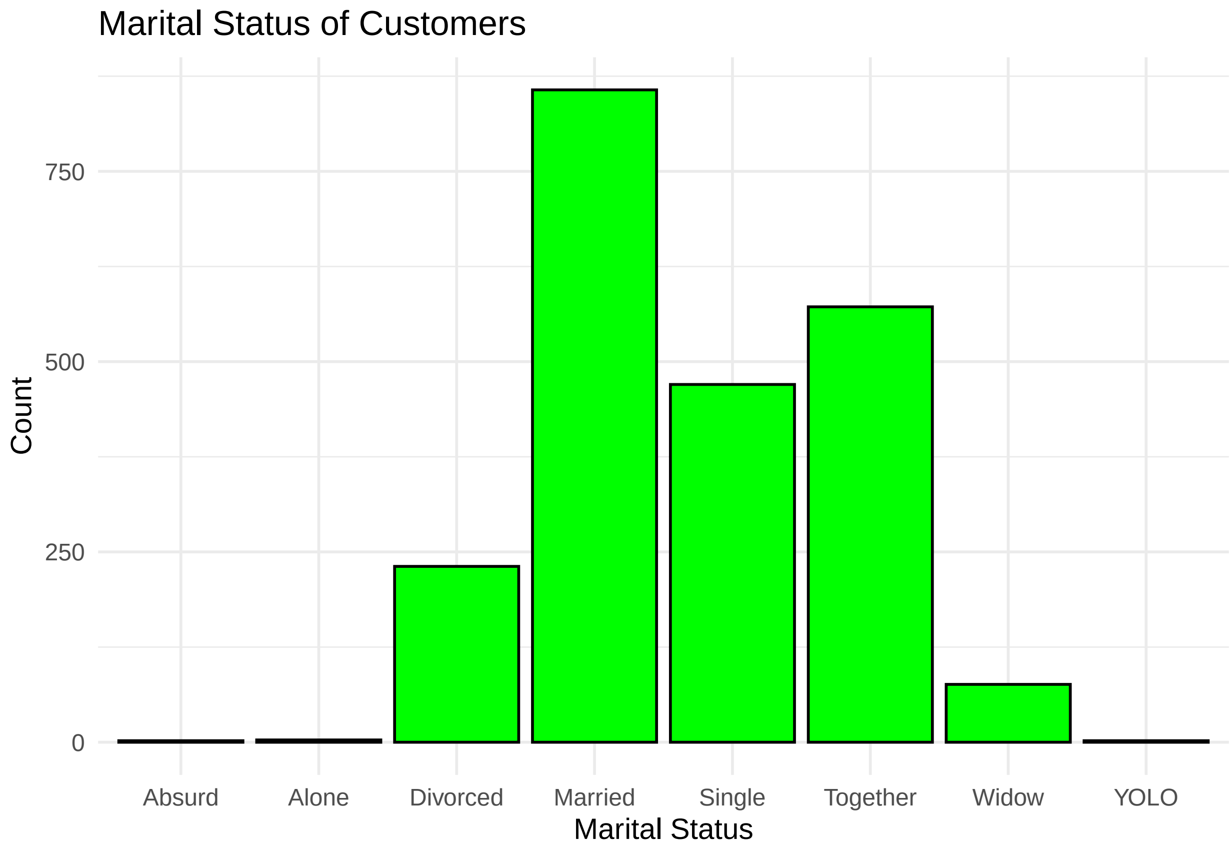
```
##        ID          Year_Birth    Education          Marital_Status    
##  Min.   :    0   Min.   :1940   Length:2213        Length:2213       
##  1st Qu.: 2815   1st Qu.:1959   Class :character   Class :character  
##  Median : 5455   Median :1970   Mode  :character   Mode  :character  
##  Mean   : 5587   Mean   :1969                                        
##  3rd Qu.: 8420   3rd Qu.:1977                                        
##  Max.   :11191   Max.   :1996                                        
##      Income         Kidhome          Teenhome       Dt_Customer       
##  Min.   : 7563   Min.   :0.0000   Min.   :0.0000   Min.   :2012-07-30  
##  1st Qu.:35246   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2013-01-16  
##  Median :51373   Median :0.0000   Median :0.0000   Median :2013-07-08  
##  Mean   :51759   Mean   :0.4419   Mean   :0.5056   Mean   :2013-07-10  
##  3rd Qu.:68487   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:2013-12-31  
##  Max.   :94461   Max.   :2.0000   Max.   :2.0000   Max.   :2014-06-29  
##     Recency         MntWines        MntFruits       MntMeatProducts 
##  Min.   : 0.00   Min.   :   0.0   Min.   :  0.00   Min.   :   0  
```

```
##  1st Qu.:24.00   1st Qu.:  24.0   1st Qu.:  2.00   1st Qu.:  16
##  Median :49.00   Median : 175.0   Median :  8.00   Median :  68
##  Mean   :49.01   Mean   : 305.2   Mean   : 26.32   Mean   : 167
##  3rd Qu.:74.00   3rd Qu.: 505.0   3rd Qu.: 33.00   3rd Qu.: 232
##  Max.   :99.00   Max.   :1493.0   Max.   :199.00   Max.   :1725
##  MntFishProducts  MntSweetProducts  MntGoldProds    NumDealsPurchases
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000
##  Median : 12.00   Median :  8.00   Median : 24.00   Median : 2.000
##  Mean   : 37.64   Mean   : 27.03   Mean   : 43.91   Mean   : 2.325
##  3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000
##  Max.   :259.00   Max.   :262.00   Max.   :321.00   Max.   :15.000
##  NumWebPurchases  NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
##  Min.   : 0.000   Min.   : 0.000      Min.   : 0.000    Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 0.000      1st Qu.: 3.000    1st Qu.: 3.000
##  Median : 4.000   Median : 2.000      Median : 5.000    Median : 6.000
##  Mean   : 4.088   Mean   : 2.671      Mean   : 5.805    Mean   : 5.322
##  3rd Qu.: 6.000   3rd Qu.: 4.000      3rd Qu.: 8.000    3rd Qu.: 7.000
##  Max.   :27.000   Max.   :28.000      Max.   :13.000    Max.   :20.000
##   AcceptedCmp3      AcceptedCmp4      AcceptedCmp5      AcceptedCmp1
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.07366   Mean   :0.07411   Mean   :0.07275   Mean   :0.06417
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##   AcceptedCmp2        Complain         Z_CostContact   Z_Revenue
##  Min.   :0.00000   Min.   :0.000000   Min.   :3     Min.   :11
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:3     1st Qu.:11
##  Median :0.00000   Median :0.000000   Median :3     Median :11
##  Mean   :0.01356   Mean   :0.009038   Mean   :3     Mean   :11
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:3     3rd Qu.:11
##  Max.   :1.00000   Max.   :1.000000   Max.   :3     Max.   :11
##     Response          Profit
##  Min.   :0.0000   Min.   :  1.55
##  1st Qu.:0.0000   1st Qu.: 21.75
##  Median :0.0000   Median :124.95
##  Mean   :0.1505   Mean   :192.43
##  3rd Qu.:0.0000   3rd Qu.:329.60
##  Max.   :1.0000   Max.   :815.50
```

Customer Demographics

We analyze customer demographics such as marital status and education level.

```
# Bar plot for marital status
ggplot(marketing_data, aes(x = Marital_Status)) +
  geom_bar(fill = "green", color = "black") +
  theme_minimal() +
  labs(title = "Marital Status of Customers", x = "Marital Status", y = "Count")
```



Data Engineering and Visualization

Feature Engineering

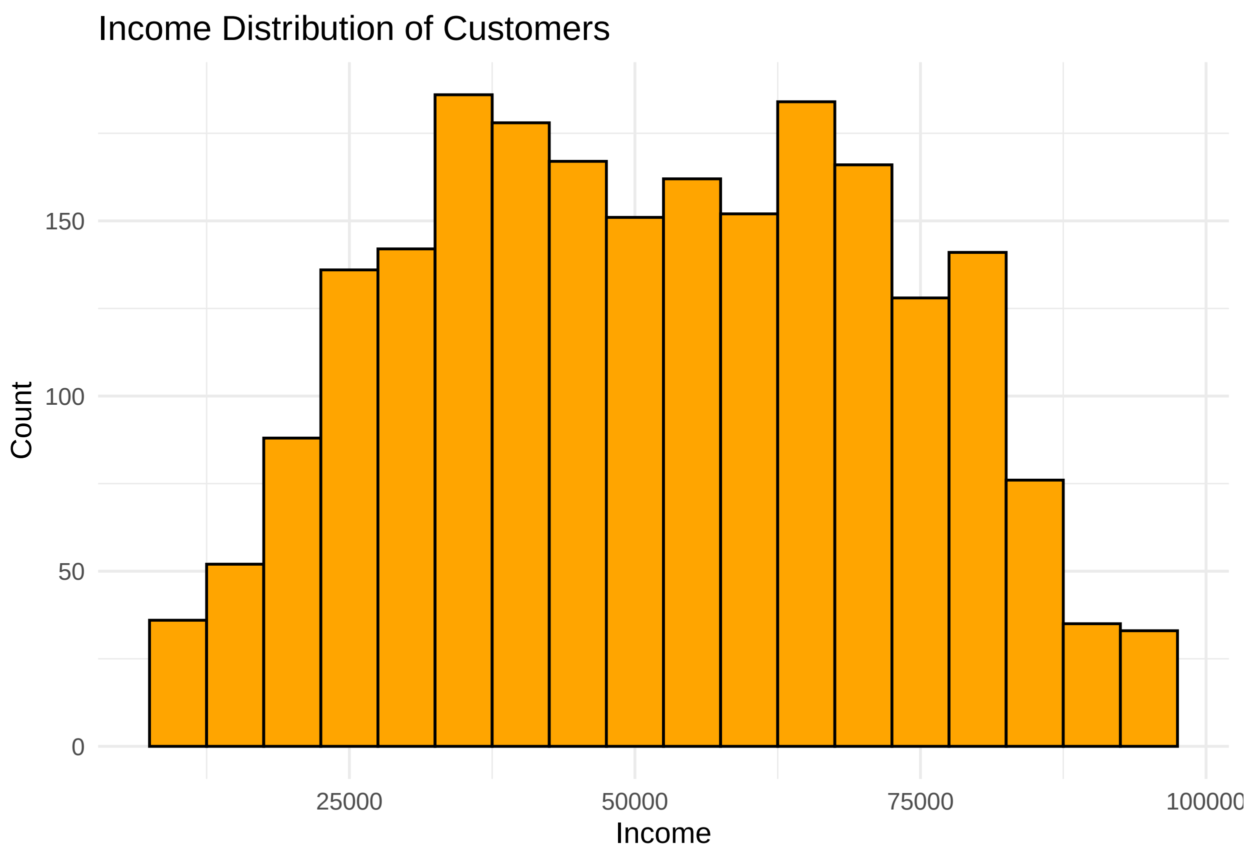
- **Customer Tenure**: Calculate the number of days since the customer joined.

```
marketing_data$Customer_Tenure <- as.numeric(difftime(Sys.Date(), marketing_data$Dt_Customer, units = "d
```

Visualizations

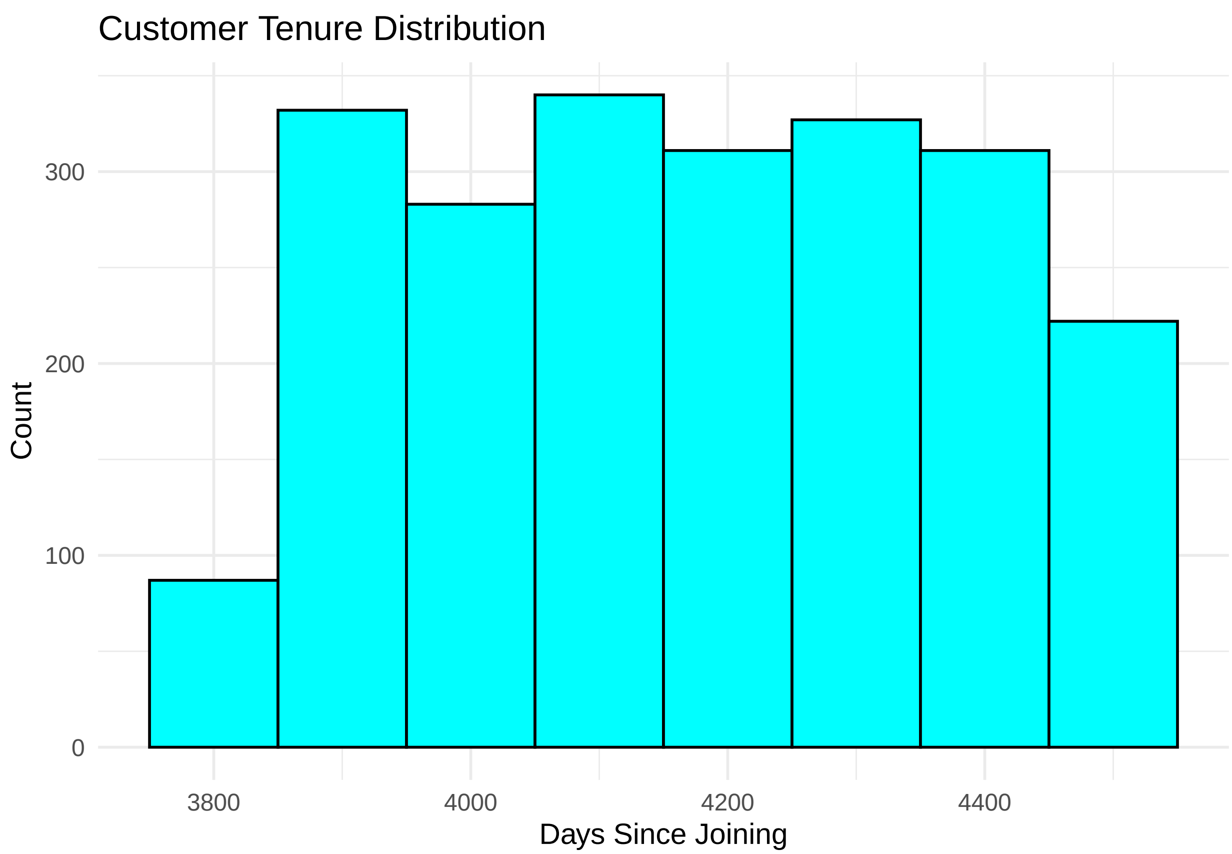
- **Income Distribution**

```
ggplot(marketing_data, aes(x = Income)) +
  geom_histogram(binwidth = 5000, fill = "orange", color = "black") +
  theme_minimal() +
  labs(title = "Income Distribution of Customers", x = "Income", y = "Count")
```



- **Tenure Analysis**

```
ggplot(marketing_data, aes(x = Customer_Tenure)) +
  geom_histogram(binwidth = 100, fill = "cyan", color = "black") +
  theme_minimal() +
  labs(title = "Customer Tenure Distribution", x = "Days Since Joining", y = "Count")
```

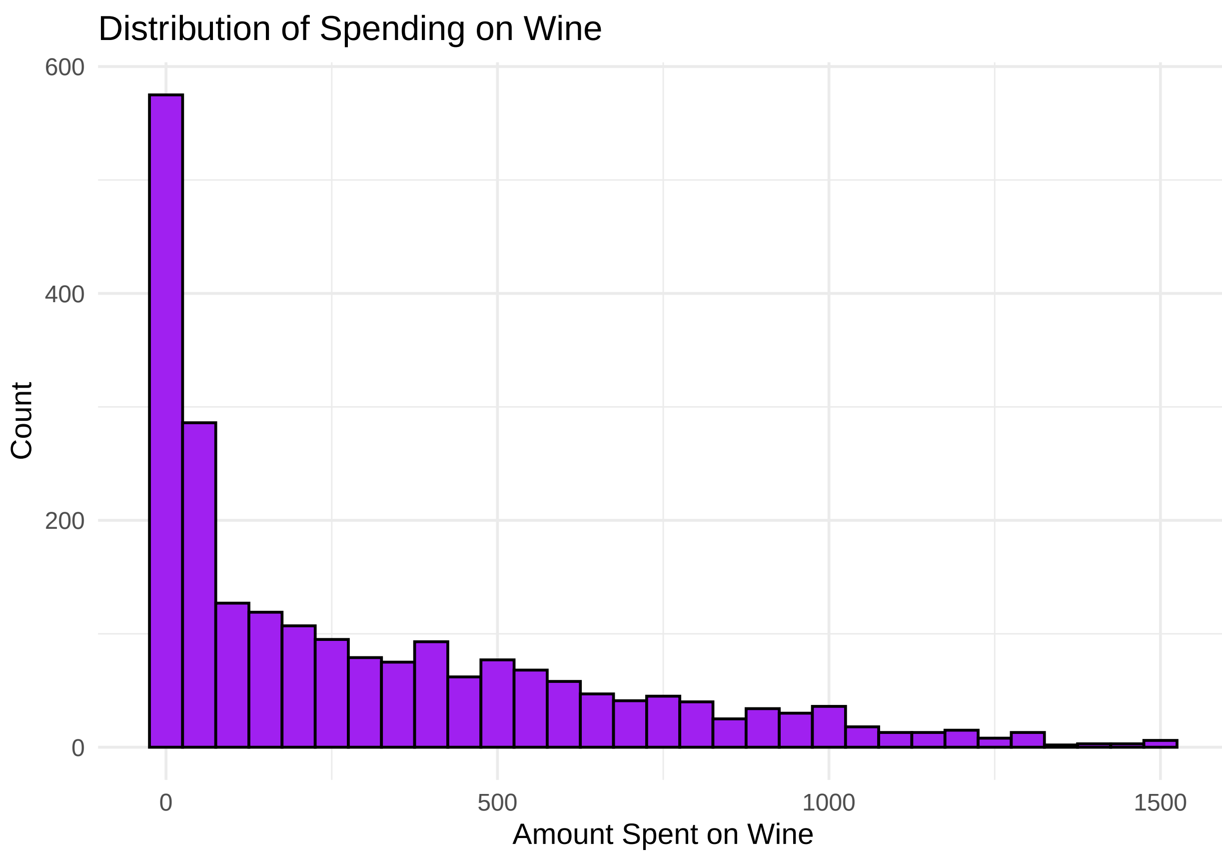


Data Analysis

Spending Patterns

In this section, we investigate how customers spend across different product categories.

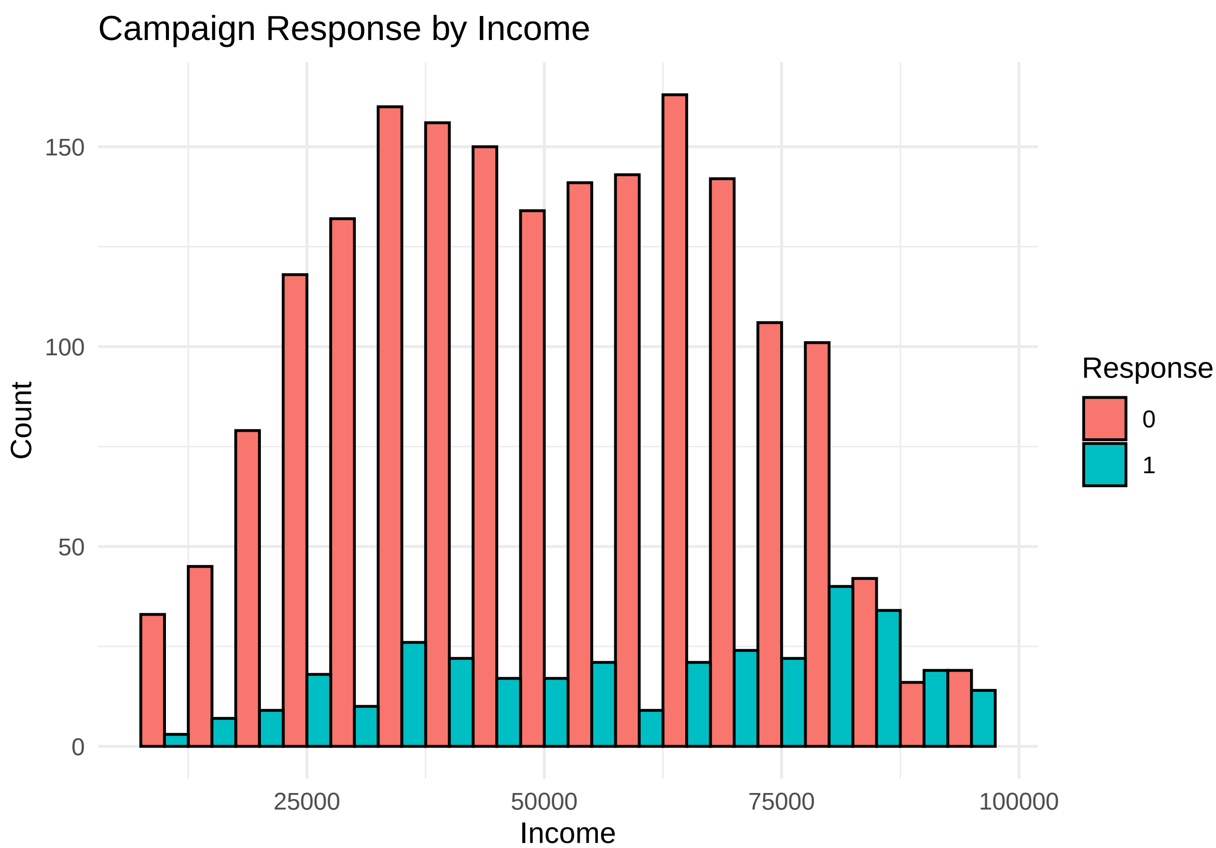
```
# Spending on wine
ggplot(marketing_data, aes(x = MntWines)) +
  geom_histogram(binwidth = 50, fill = "purple", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Spending on Wine", x = "Amount Spent on Wine", y = "Count")
```



Campaign Response Analysis

We now explore factors affecting responses to the most recent campaign.

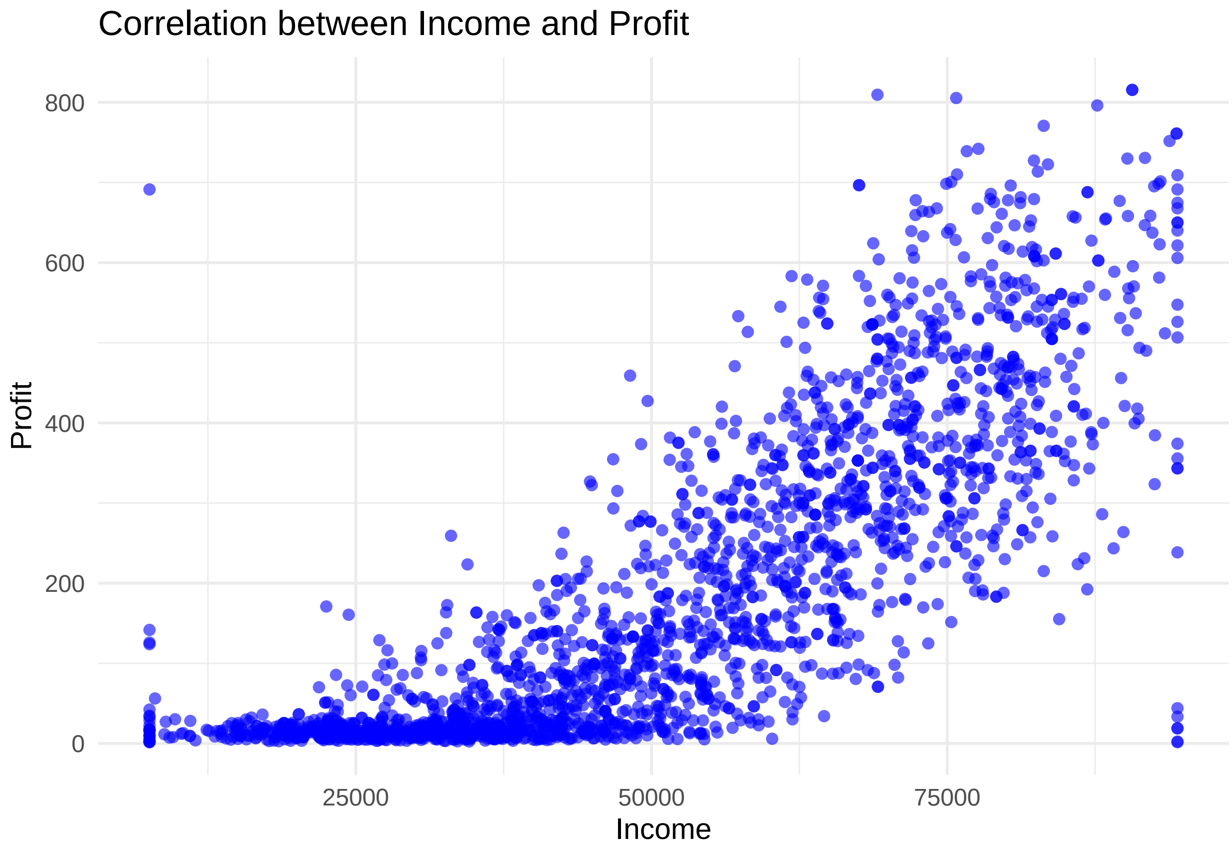
```
# Response rates by income
ggplot(marketing_data, aes(x = Income, fill = factor(Response))) +
  geom_histogram(binwidth = 5000, position = "dodge", color = "black") +
  theme_minimal() +
  labs(title = "Campaign Response by Income", x = "Income", y = "Count", fill = "Response")
```



Correlation Analysis: Income vs Profit

In this section, we explore the relationship between customer income and profit.

```
# Scatter plot to visualize correlation between Income and Profit
ggplot(marketing_data, aes(x = Income, y = Profit)) +
  geom_point(alpha = 0.6, color = "blue") +
  theme_minimal() +
  labs(title = "Correlation between Income and Profit", x = "Income", y = "Profit")
```



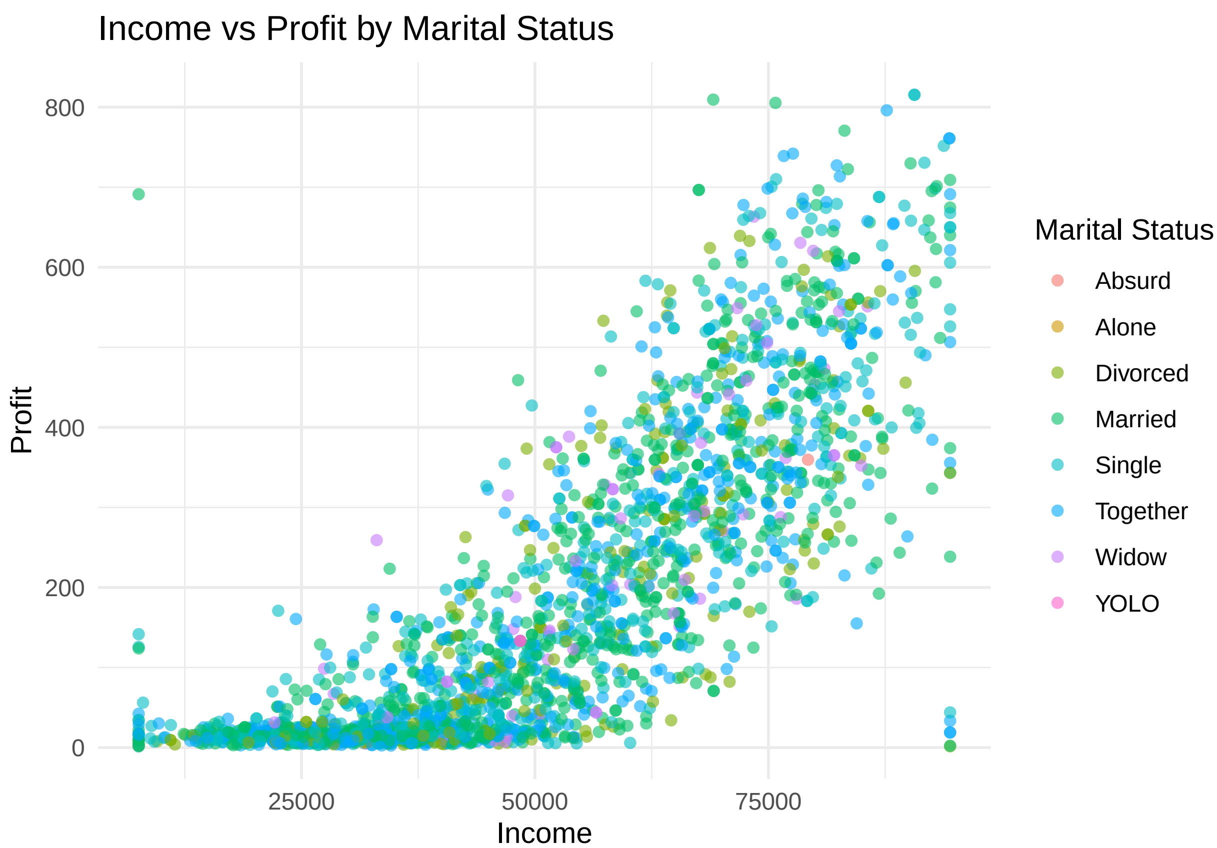

```
# Calculate the correlation coefficient between Income and Profit
correlation <- cor(marketing_data$Income, marketing_data$Profit, use = "complete.obs")
cat("Correlation between Income and Profit: ", correlation, "\n")
```

```
## Correlation between Income and Profit:  0.813656
```

Income vs Profit Segmented by Marital Status

We further explore how the relationship between income and profit varies based on marital status.

```
# Scatter plot of Income vs Profit segmented by Marital Status
ggplot(marketing_data, aes(x = Income, y = Profit, color = Marital_Status)) +
  geom_point(alpha = 0.6) +
  theme_minimal() +
  labs(title = "Income vs Profit by Marital Status", x = "Income", y = "Profit", color = "Marital Statu
```




```r
# Correlation coefficients by marital status, excluding groups with zero variance
correlation_by_status <- marketing_data %>%
  group_by(Marital_Status) %>%
  filter(sd(Income) > 0 & sd(Profit) > 0) %>%
  summarise(correlation = cor(Income, Profit, use = "complete.obs"))

print(correlation_by_status)
```

```
## # A tibble: 7 x 2
##   Marital_Status correlation
##   <chr>                <dbl>
## 1 Absurd              -1
## 2 Alone                0.994
## 3 Divorced             0.777
## 4 Married              0.809
## 5 Single               0.835
## 6 Together             0.819
## 7 Widow                0.789
```

Conclusion

In conclusion, the analysis shows several interesting insights into customer behavior and responses to marketing campaigns:

- The majority of customers are in the middle-age group, with notable spending patterns on wine and other products.
- There are clear differences in campaign responses based on income levels and family status.
- The correlation analysis reveals the relationship between customer income and profit, providing insights into how income influences profitability.
- The additional segmentation analysis indicates that the relationship between income and profit can vary significantly across different marital statuses, which could help in targeted marketing strategies.