# Planning Stage Report – Group 4 | STAT 301 (2025SS)

**Student**: Krishaant Pathmanathan
**Dataset**: Customer Personality Analysis
**TA**: Yian Lin

---

In [125…

```r
# install.packages("skimr")
# install.packages("ggcorrplot")
library(tidyverse)
library(skimr)
library(dplyr)
library(tidyr)
library(tibble)
library(ggplot2)
library(GGally)
library(dplyr)
library(janitor)
library(patchwork)
library(cowplot)
library(corrplot)
library(ggcorrplot)
```

In [126…

```r
df <- read_delim("marketing_campaign.csv", delim = "\t")
glimpse(df) # to see what the data looks like
skim(df) # to get a summary of the data
```

```
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, Mn
tF...

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 m
essage.
```

```
Rows: 2,240
Columns: 29
$ ID                  <dbl> 5524, 2174, 4141, 6182, 5324, 7446, 965, 6177,
485…
$ Year_Birth          <dbl> 1957, 1954, 1965, 1984, 1981, 1967, 1971, 198
5, 19…
$ Education           <chr> "Graduation", "Graduation", "Graduation", "Gra
duat…
$ Marital_Status      <chr> "Single", "Single", "Together", "Together", "M
arri…
$ Income              <dbl> 58138, 46344, 71613, 26646, 58293, 62513, 5563
5, 3…
$ Kidhome             <dbl> 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0,
0, 1,…
$ Teenhome            <dbl> 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0,
0, 1,…
$ Dt_Customer         <chr> "04-09-2012", "08-03-2014", "21-08-2013", "10-
02-2…
$ Recency             <dbl> 58, 38, 26, 26, 94, 16, 34, 32, 19, 68, 11, 5
9, 82…
$ MntWines            <dbl> 635, 11, 426, 11, 173, 520, 235, 76, 14, 28,
5, 6,…
$ MntFruits           <dbl> 88, 1, 49, 4, 43, 42, 65, 10, 0, 0, 5, 16, 61,
2, …
$ MntMeatProducts     <dbl> 546, 6, 127, 20, 118, 98, 164, 56, 24, 6, 6, 1
1, 4…
$ MntFishProducts     <dbl> 172, 2, 111, 10, 46, 0, 50, 3, 3, 1, 0, 11, 22
5, 3…
$ MntSweetProducts    <dbl> 88, 1, 21, 3, 27, 42, 49, 1, 3, 1, 2, 1, 112,
5, 1…
$ MntGoldProds        <dbl> 88, 6, 42, 5, 15, 14, 27, 23, 2, 13, 1, 16, 3
0, 14…
$ NumDealsPurchases   <dbl> 3, 2, 1, 2, 5, 2, 4, 2, 1, 1, 1, 1, 1, 3, 1,
1, 3,…
$ NumWebPurchases     <dbl> 8, 1, 8, 2, 5, 6, 7, 4, 3, 1, 1, 2, 3, 6, 1,
7, 3,…
$ NumCatalogPurchases <dbl> 10, 1, 2, 0, 3, 4, 3, 0, 0, 0, 0, 0, 4, 1, 0,
6, 0…
$ NumStorePurchases   <dbl> 4, 2, 10, 4, 6, 10, 7, 4, 2, 0, 2, 3, 8, 5, 3,
12,…
$ NumWebVisitsMonth   <dbl> 7, 5, 4, 6, 5, 6, 6, 8, 9, 20, 7, 8, 2, 6, 8,
3, 8…
$ AcceptedCmp3        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0,…
$ AcceptedCmp4        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
$ AcceptedCmp5        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0,…
$ AcceptedCmp1        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0,…
$ AcceptedCmp2        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
$ Complain            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
$ Z_CostContact       <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3,…
$ Z_Revenue           <dbl> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 1
1, 11…
$ Response            <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
1, 0,…
```

```
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             2240
Number of columns          29
_______________________
Column type frequency:
  character                3
  numeric                  26
________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────
──────
  skim_variable  n_missing complete_rate min max empty n_unique whitespace
1 Education              0             1   3  10     0        5          0
2 Marital_Status         0             1   4   8     0        8          0
3 Dt_Customer            0             1  10  10     0      663          0

── Variable type: numeric ────────────────────────────────────────────────────────
──────
   skim_variable       n_missing complete_rate      mean         sd   p0
 1 ID                          0         1       5592.       3247.     0
 2 Year_Birth                  0         1       1969.         12.0 1893
 3 Income                     24         0.989 52247.      25173.   1730
 4 Kidhome                     0         1         0.444       0.538   0
 5 Teenhome                    0         1         0.506       0.545   0
 6 Recency                     0         1        49.1        29.0     0
 7 MntWines                    0         1       304.        337.      0
 8 MntFruits                   0         1        26.3        39.8     0
 9 MntMeatProducts             0         1       167.        226.      0
10 MntFishProducts             0         1        37.5        54.6     0
11 MntSweetProducts            0         1        27.1        41.3     0
12 MntGoldProds                0         1        44.0        52.2     0
13 NumDealsPurchases           0         1         2.33        1.93    0
14 NumWebPurchases             0         1         4.08        2.78    0
15 NumCatalogPurchases         0         1         2.66        2.92    0
16 NumStorePurchases           0         1         5.79        3.25    0
17 NumWebVisitsMonth           0         1         5.32        2.43    0
18 AcceptedCmp3                0         1         0.0728      0.260   0
19 AcceptedCmp4                0         1         0.0746      0.263   0
20 AcceptedCmp5                0         1         0.0728      0.260   0
21 AcceptedCmp1                0         1         0.0643      0.245   0
22 AcceptedCmp2                0         1         0.0134      0.115   0
23 Complain                    0         1         0.00938     0.0964  0
24 Z_CostContact               0         1         3           0       3
25 Z_Revenue                   0         1        11           0      11
26 Response                    0         1         0.149       0.356   0
       p25    p50    p75   p100 hist
 1  2828.   5458.  8428.  11191 ▇▇▇▇▇
 2  1959    1970   1977    1996 ▁▁▂▇▅
 3 35303   51382. 68522  666666 ▇▁▁▁▁
 4     0       0      1       2 ▇▁▆▁▁
 5     0       0      1       2 ▇▁▇▁▁
 6    24      49     74      99 ▇▇▇▇▇
 7    23.8   174.   504.   1493 ▇▂▂▁▁
 8     1       8     33     199 ▇▁▁▁▁
 9    16      67    232    1725 ▇▁▁▁▁
10     3      12     50     259 ▇▁▁▁▁
11     1       8     33     263 ▇▁▁▁▁
```

```
12      9       24      56      362
13      1        2       3       15
14      2        4       6       27
15      0        2       4       28
16      3        5       8       13
17      3        6       7       20
18      0        0       0        1
19      0        0       0        1
20      0        0       0        1
21      0        0       0        1
22      0        0       0        1
23      0        0       0        1
24      3        3       3        3
25     11       11      11       11
26      0        0       0        1
```

```
Error in is.null(text_repr) || nchar(text_repr) == 0L: 'length = 17' in co
ercion to 'logical(1)'
Traceback:
```

# (1) Data Description

## 0. Data Preprocessing

I downloaded the zipped data from Kaggle. I extracted and saved the `.csv` file under the path `STAT301groupproject/marketing_campaign.csv`. I created a GitHub repo to store all of this work so its easier when I start working with my group-mates.

To understand the dataset structure, I first used `glimpse()` and `skim()` to inspect the total number of observations, the variable types, and basic distributional summaries (e.g., min, max, mean). This allowed me to identify the missing values in the `Income` variable and to begin categorizing variables based on their role in the analysis.

## 1. Dataset Summary

The **Customer Personality Analysis** dataset is a marketing dataset that contains 2,240 observations and 29 variables. Each row represents a customer, and the columns capture a wide range of information, including demographics, spending habits, campaign responses, and website interactions.

This dataset is useful for understanding customer behavior and segmenting customers for targeted marketing. For example, instead of marketing a new product to the entire customer base, a company can identify which segment is most likely to purchase and focus marketing efforts accordingly. This dataset is provided by Dr. Omar Romero-Hernandez.

Below is a grouped variable dictionary that organizes all 29 attributes into meaningful categories based on their content and analytical purpose.

## Key Variables (Grouped by Category)

1. Customer's Information
2. Products (Spending in Last 2 Years)
3. Promotion
4. Place (Purchase Channels)
5. Other (Dummy Columns)

## Full Variable Description Table (Grouped by Category)

### Customer’s Information

| Variable | Type | Description |
|---|---|---|
| ID | numeric | Unique customer ID |
| Year_Birth | numeric | Year of birth |
| Education | categorical | Level of education (e.g., Graduation, PhD) |
| Marital_Status | categorical | Marital status |
| Income | numeric | Household yearly income |
| Kidhome | numeric | Number of children at home |
| Teenhome | numeric | Number of teenagers at home |
| Dt_Customer | datetime | Date customer enrolled |
| Recency | numeric | Days since last purchase |
| Complain | boolean | Complained in the last 2 years (1 = yes) |

### Products (Amount Spent in Last 2 Years)

| Variable | Type | Description |
|---|---|---|
| MntWines | numeric | Amount spent on wine |
| MntFruits | numeric | Amount spent on fruits |
| MntMeatProducts | numeric | Amount spent on meat products |
| MntFishProducts | numeric | Amount spent on fish products |
| MntSweetProducts | numeric | Amount spent on sweet products |
| MntGoldProds | numeric | Amount spent on gold products |

### Promotion

| Variable | Type | Description |
|---|---|---|
| NumDealsPurchases | numeric | Number of purchases using discounts |
| AcceptedCmp1 | boolean | Accepted 1st campaign (1 = yes, 0 = no) |
| AcceptedCmp2 | boolean | Accepted 2nd campaign (1 = yes, 0 = no) |
| AcceptedCmp3 | boolean | Accepted 3rd campaign (1 = yes, 0 = no) |
| AcceptedCmp4 | boolean | Accepted 4th campaign (1 = yes, 0 = no) |

| Variable | Type | Description |
| --- | --- | --- |
| AcceptedCmp5 | boolean | Accepted 5th campaign (1 = yes, 0 = no) |
| Response | boolean | Accepted the most recent campaign (1 = yes, 0 = no) |

### Place (Purchase Channels)

| Variable | Type | Description |
| --- | --- | --- |
| NumWebPurchases | numeric | Number of website purchases |
| NumCatalogPurchases | numeric | Number of catalog purchases |
| NumStorePurchases | numeric | Number of in-store purchases |
| NumWebVisitsMonth | numeric | Website visits in the last month |

### Other

| Variable | Type | Description |
| --- | --- | --- |
| Z_CostContact | numeric | Dummy cost variable (constant = 3) |
| Z_Revenue | numeric | Dummy revenue variable (constant = 11) |

## 2. Missing Values & Data tidying

- The `Income` variable has **24 missing values**.
- All other variables are complete with **no missing data**.

Since the proportion of missing values here is very small, I have decided to **remove the rows with missing `Income` values**.

In [127…

```
df <- df |> drop_na()
```

Note that the variables `Z_CostContact` and `Z_Revenue` have the same value for all rows, so they don't really help and I am going to drop them.

In [128…

```
df <- df %>%
  select(-Z_CostContact, -Z_Revenue)
```

I am going to make **changes to some variables** so they make more sense for our analysis, in particular I will focus on

| Variable(s) | Planned Transformation / Analysis | |---------------------------|-----------------------------------------------------------------------------| | `Year_Birth` | Convert to age | | `Education` | Convert to binary: Post Graduate vs. Under Graduate | | `Marital_Status` | Group into: Married vs. Not Married | | | `Kidhome` and `Teenhome` | Combine into a single variable: total number of children in the household | | `Total_Spening` | A column that shows Total Spending = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds|

In [129…

```
df <- df |>

# Convert Year_Birth to Age
    mutate(Age = 2025 - Year_Birth) |>

# Make Education binary : Post Graduate vs Under Graduate
    mutate(Education = recode(Education,
                              'PhD' = 'Post Graduate',
                              'Master' = 'Post Graduate',
                              'Graduation' = 'Under Graduate',
                              '2n Cycle' = 'Under Graduate',
                              'Basic' = 'Under Graduate')) |>

# Make Marital_Status binary: Married vs Not Married
    mutate(Marital_Status = case_when(
        Marital_Status %in% c("Married", "Together") ~ "Married", TRUE ~

# Add up Kidhome and teenhome into Num_Children
    mutate(Num_Children = Kidhome + Teenhome) |>

# Add a column 'Total_Spending' by summing all spending related columns
    mutate(Total_Spending = MntWines + MntFruits +MntMeatProducts + MntFi

head(df)
```

| ID | Year_Birth | Education | Marital_Status | Income | Kidhome | Teenhome | Dt_Cust |
|---|---|---|---|---|---|---|---|
| <dbl> | <dbl> | <chr> | <chr> | <dbl> | <dbl> | <dbl> | < |
| 5524 | 1957 | Under Graduate | Not Married | 58138 | 0 | 0 | 04-09- |
| 2174 | 1954 | Under Graduate | Not Married | 46344 | 1 | 1 | 08-03- |
| 4141 | 1965 | Under Graduate | Married | 71613 | 0 | 0 | 21-08- |
| 6182 | 1984 | Under Graduate | Married | 26646 | 1 | 0 | 10-02- |
| 5324 | 1981 | Post Graduate | Married | 58293 | 1 | 0 | 19-01- |
| 7446 | 1967 | Post Graduate | Married | 62513 | 0 | 1 | 09-09- |

## 3. Summary Statistics

I created some summary statistics tables below to help me better understand the data. Obviously there's many things we can do but I chose :

1. Summary of unique data
2. Summary of numeric columns
3. Summary of non-numeric columns

## 1) Summary of unique data

In [130…

```
unique_df <- df|>
  summarise(across(everything(), ~ n_distinct(.))) |>
  pivot_longer(cols = everything(), names_to = "variable", values_to = "n
  arrange(desc(n_unique))

unique_df
```

A tibble: 30 × 2

| variable | n_unique |
|---:|---:|
| <chr> | <int> |
| ID | 2216 |
| Income | 1974 |
| Total_Spending | 1047 |
| MntWines | 776 |
| Dt_Customer | 662 |
| MntMeatProducts | 554 |
| MntGoldProds | 212 |
| MntFishProducts | 182 |
| MntSweetProducts | 176 |
| MntFruits | 158 |
| Recency | 100 |
| Year_Birth | 59 |
| Age | 59 |
| NumWebVisitsMonth | 16 |
| NumDealsPurchases | 15 |
| NumWebPurchases | 15 |
| NumCatalogPurchases | 14 |
| NumStorePurchases | 14 |
| Num_Children | 4 |
| Kidhome | 3 |
| Teenhome | 3 |
| Education | 2 |
| Marital_Status | 2 |
| AcceptedCmp3 | 2 |
| AcceptedCmp4 | 2 |
| AcceptedCmp5 | 2 |
| AcceptedCmp1 | 2 |
| AcceptedCmp2 | 2 |
| Complain | 2 |
| Response | 2 |

Interestingly, we have 2216 unique customers, and we see that a lot of out variables are binary.

### 2) Summary of numeric columns

In [133…

```
numeric_df <- df |>
  select(where(is.numeric)) |>
  pivot_longer(cols = everything(), names_to = "variable", values_to = "v
  group_by(variable) |>
  summarise(
    min     = round(min(value, na.rm = TRUE), 2),
    q1      = round(quantile(value, 0.25, na.rm = TRUE), 2),
    median  = round(median(value, na.rm = TRUE), 2),
    mean    = round(mean(value, na.rm = TRUE), 2),
    q3      = round(quantile(value, 0.75, na.rm = TRUE), 2),
    max     = round(max(value, na.rm = TRUE), 2),
    sd      = round(sd(value, na.rm = TRUE), 2),
    missing = sum(is.na(value))
  )
numeric_df
```

A tibble: 27 × 9

| variable | min | q1 | median | mean | q3 | max | s |
|---|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <db |
| AcceptedCmp1 | 0 | 0.00 | 0.0 | 0.06 | 0.00 | 1 | 0.: |
| AcceptedCmp2 | 0 | 0.00 | 0.0 | 0.01 | 0.00 | 1 | 0. |
| AcceptedCmp3 | 0 | 0.00 | 0.0 | 0.07 | 0.00 | 1 | 0.: |
| AcceptedCmp4 | 0 | 0.00 | 0.0 | 0.07 | 0.00 | 1 | 0.: |
| AcceptedCmp5 | 0 | 0.00 | 0.0 | 0.07 | 0.00 | 1 | 0.: |
| Age | 29 | 48.00 | 55.0 | 56.18 | 66.00 | 132 | 11.! |
| Complain | 0 | 0.00 | 0.0 | 0.01 | 0.00 | 1 | 0. |
| ID | 0 | 2814.75 | 5458.5 | 5588.35 | 8421.75 | 11191 | 3249.: |
| Income | 1730 | 35303.00 | 51381.5 | 52247.25 | 68522.00 | 666666 | 25173.( |
| Kidhome | 0 | 0.00 | 0.0 | 0.44 | 1.00 | 2 | 0.! |
| MntFishProducts | 0 | 3.00 | 12.0 | 37.64 | 50.00 | 259 | 54.' |
| MntFruits | 0 | 2.00 | 8.0 | 26.36 | 33.00 | 199 | 39.' |
| MntGoldProds | 0 | 9.00 | 24.5 | 43.97 | 56.00 | 321 | 51.! |
| MntMeatProducts | 0 | 16.00 | 68.0 | 167.00 | 232.25 | 1725 | 224.: |
| MntSweetProducts | 0 | 1.00 | 8.0 | 27.03 | 33.00 | 262 | 41.( |
| MntWines | 0 | 24.00 | 174.5 | 305.09 | 505.00 | 1493 | 337.: |
| NumCatalogPurchases | 0 | 0.00 | 2.0 | 2.67 | 4.00 | 28 | 2.! |
| NumDealsPurchases | 0 | 1.00 | 2.0 | 2.32 | 3.00 | 15 | 1.! |
| NumStorePurchases | 0 | 3.00 | 5.0 | 5.80 | 8.00 | 13 | 3.: |
| NumWebPurchases | 0 | 2.00 | 4.0 | 4.09 | 6.00 | 27 | 2.' |
| NumWebVisitsMonth | 0 | 3.00 | 6.0 | 5.32 | 7.00 | 20 | 2.. |
| Num_Children | 0 | 0.00 | 1.0 | 0.95 | 1.00 | 3 | 0.' |
| Recency | 0 | 24.00 | 49.0 | 49.01 | 74.00 | 99 | 28.! |
| Response | 0 | 0.00 | 0.0 | 0.15 | 0.00 | 1 | 0.: |
| Teenhome | 0 | 0.00 | 0.0 | 0.51 | 1.00 | 2 | 0.! |
| Total_Spending | 5 | 69.00 | 396.5 | 607.08 | 1048.00 | 2525 | 602.! |
| Year_Birth | 1893 | 1959.00 | 1970.0 | 1968.82 | 1977.00 | 1996 | 11.! |

Findings:

- Income has a mean of 52,247 with max being 666,666 so we have to watch for outliers
- Wine has the highest mean spending at 305.09 followed by meat, gold, sweets, and fruits

- In store puchases are the highest mean at 5.8 followed by web purchases.

### 3) Summary of non-numeric columns

In [134…

```
non_numeric_df <- df |>
  select(where(~!is.numeric(.)), -Dt_Customer) |> # I took out Dt_Custome
  pivot_longer(cols = everything(), names_to = "variable", values_to = "v
  group_by(variable, value) |>
  summarise(
    count = n(),
    proportion = round(count / nrow(df), 4),
    .groups = "drop"
  ) |>
  arrange(variable, desc(count))
non_numeric_df
```

A tibble: 4 × 4

| variable | value | count | proportion |
|---|---|---|---|
| <chr> | <chr> | <int> | <dbl> |
| Education | Under Graduate | 1370 | 0.6182 |
| Education | Post Graduate | 846 | 0.3818 |
| Marital_Status | Married | 1430 | 0.6453 |
| Marital_Status | Not Married | 786 | 0.3547 |

Findings : Most customers are married and relatively a lot have post graduate degrees.

# (2) Question

**Response Variable:** `Response` is a binary variable indicating wether the customer accepted the last campaign, (1 = yes, 0 = no).

**Explanatory Variables :**

- `Age` — Derived from `Year_Birth`, represents the customer's age.
- `Income` — Household yearly income.
- `Education` — Converted to a binary indicator ( `Postgraduate` vs. `Undergraduate` ).
- `Marital_Status` — Grouped into `Married` vs. `Not Married`.
- `Complain` — Whether the customer complained in the last 2 years.
- `NumWebPurchases` — Number of purchases made through the company's website.
- `NumCatalogPurchases` — Number of purchases made through catalogs.

- `NumStorePurchases` — Number of purchases made in physical stores.
- `NumWebVisitsMonth` — Number of visits to the website in the last month.
- `Total_Spending` — Sum of spending on wine, meat, fish, fruits, sweets, and gold products over the last two years.

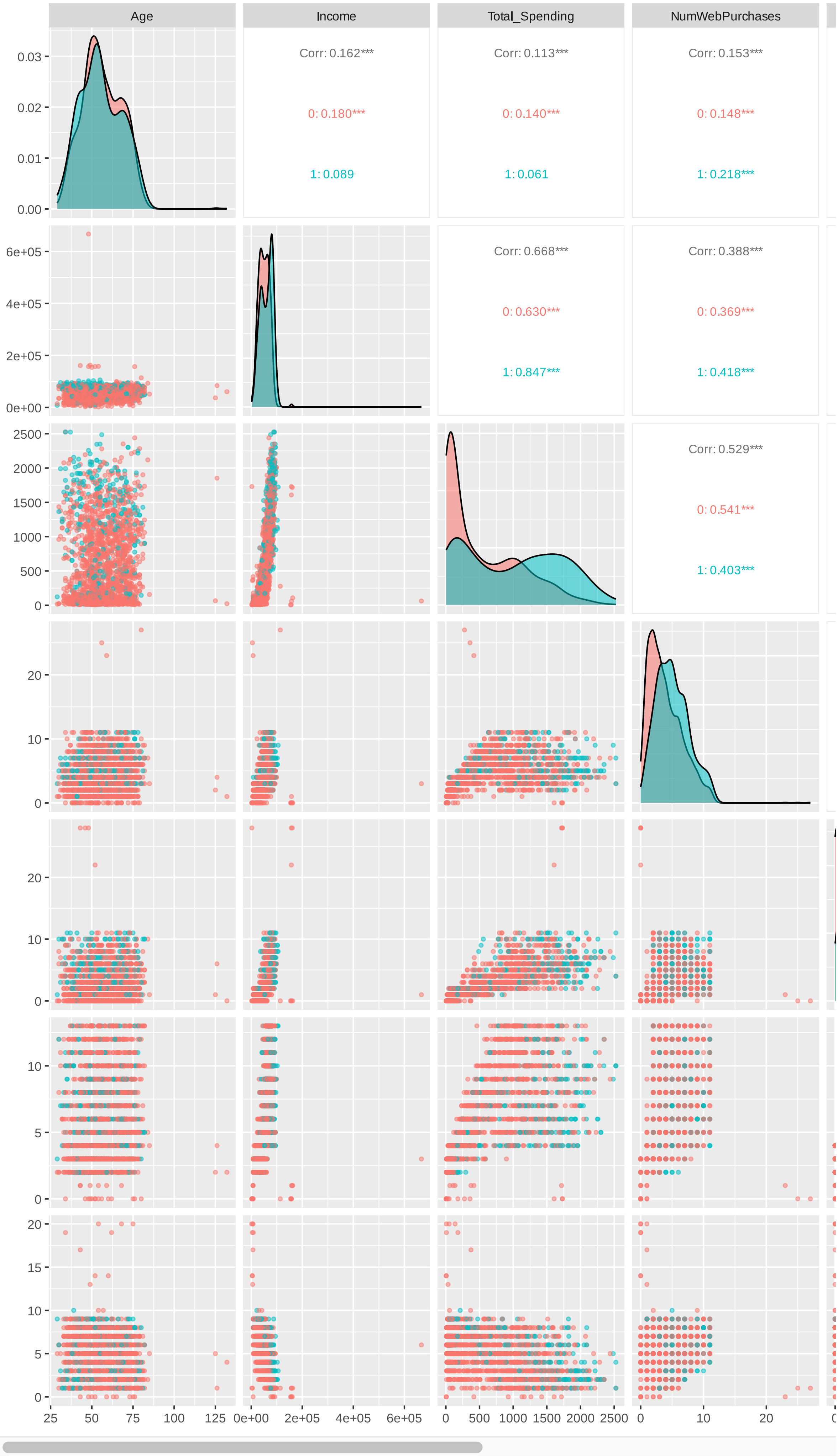
Hence, I can ask **How does a customer's demographic profile and spending behaviour affect likelihood of accepting a marketing campaign ?**

# (3)Exploratory Data Analysis and Visualization

Lets look at an overview distribution of our data and decide what to focus on. We can use ggpairs for all numeric variables.

In [136…

```
df_numeric <- df %>%
  select(Response, Age, Income, Total_Spending,
         NumWebPurchases, NumCatalogPurchases,
         NumStorePurchases, NumWebVisitsMonth)

df_numeric$Response <- as.factor(df_numeric$Response)

options(repr.plot.width = 14, repr.plot.height = 14)  # setting size

ggpairs(
  data = df_numeric,
  columns = 2:8,
  aes(color = Response, alpha = 0.6), # GPT helped me with making this pr
  upper = list(continuous = wrap("cor", size = 3)),
  lower = list(continuous = wrap("points", size = 1, alpha = 0.5)),
  diag = list(continuous = wrap("densityDiag")))
```


Age
Income
Total_Spending
NumWebPurchases
Corr: 0.162***
0: 0.180***
1: 0.089
Corr: 0.113***
0: 0.140***
1: 0.061
Corr: 0.153***
0: 0.148***
1: 0.218***
Corr: 0.668***
0: 0.630***
1: 0.847***
Corr: 0.388***
0: 0.369***
1: 0.418***
Corr: 0.529***
0: 0.541***
1: 0.403***

We see the difference in distribution between people who responded to the latest marketting campaign and people who did not. `Total_Spending` is highly correlated with all three types of purchases—web, catalog, and store—suggesting multicollinearity, so we should look out for that later in the model. Additionally, customers who responded (in turquoise) tend to cluster in higher purchase and spending ranges, so there is some pattern in spending behaviour and response.

### Plots we should look at :

1. Demographic vs Response

- Education
- Marital Status
- Complain

2. Numerical Predictors vs Response

- Age
- Income
- Total Spending

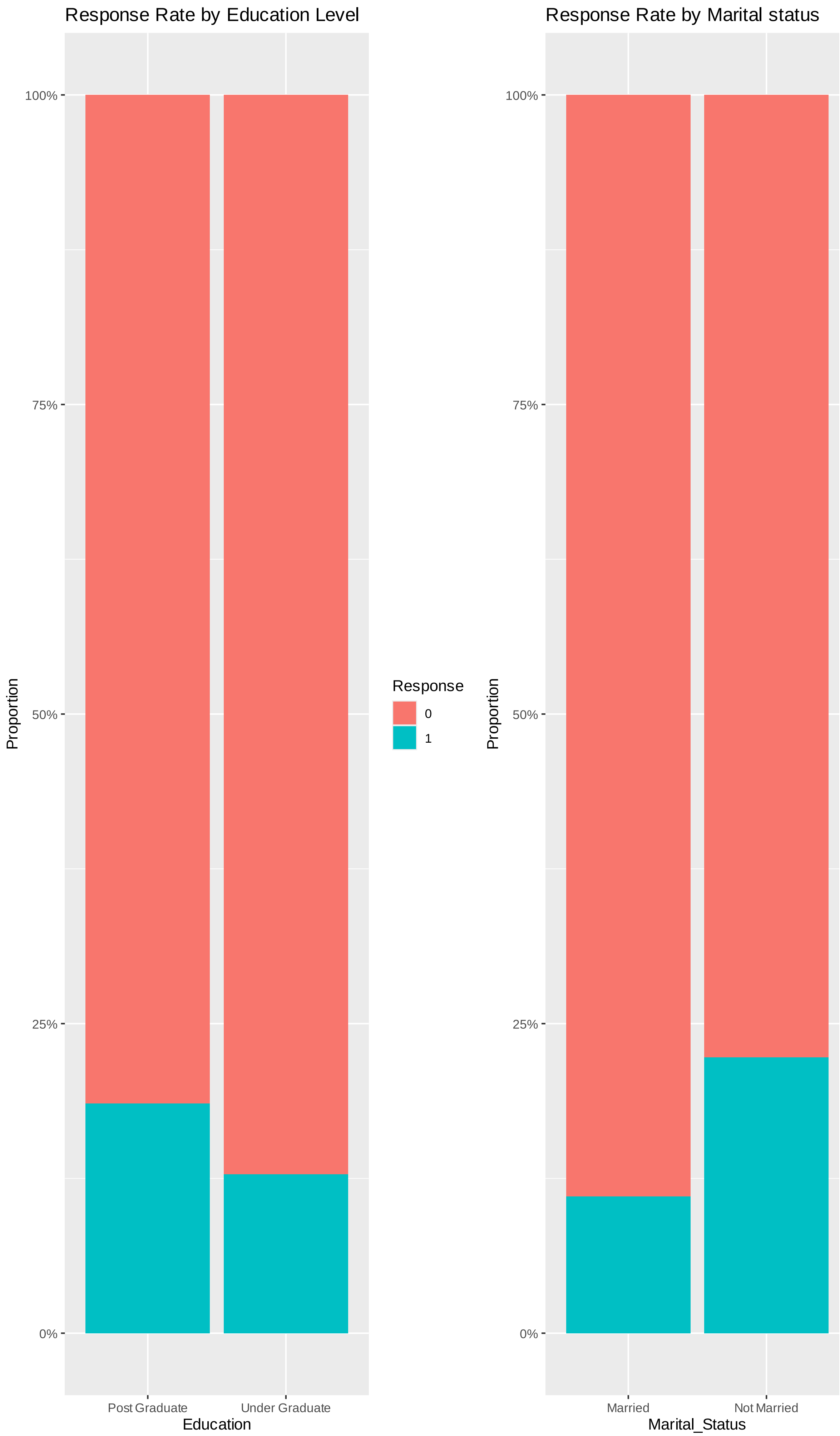
3. Behavioral Features vs Response

- Website Purchases
- Catalog Purchases
- Store Purchases
- Website visits

4. Correlation Between Variables

In [138…

```
# 1a. Education vs Response
edu_plot <- ggplot(df, aes(x = Education, fill = factor(Response))) +
    geom_bar(position = "fill")+
    scale_y_continuous(labels = scales::percent)+
    labs(title = "Response Rate by Education Level", y = "Proportion", fi

# 1b. Marital Status vs Response
marital_plot <- ggplot(df, aes(x = Marital_Status, fill = factor(Response
    geom_bar(position = "fill") +
    scale_y_continuous(labels = scales::percent) +
    labs(title = "Response Rate by Marital status", y = "Proportion",fill

# 1c. Complain vs Response
complain_plot <- ggplot(df, aes(x = factor(Complain), fill = factor(Respo
    geom_bar(position = "fill") +
    scale_y_continuous(labels = scales::percent) +
    labs(title = "Respomse Rate by Complaint statuss",x = "Complain", y =

# Combined plot !
plot_grid(
  edu_plot, marital_plot, complain_plot,
  ncol = 3)
```

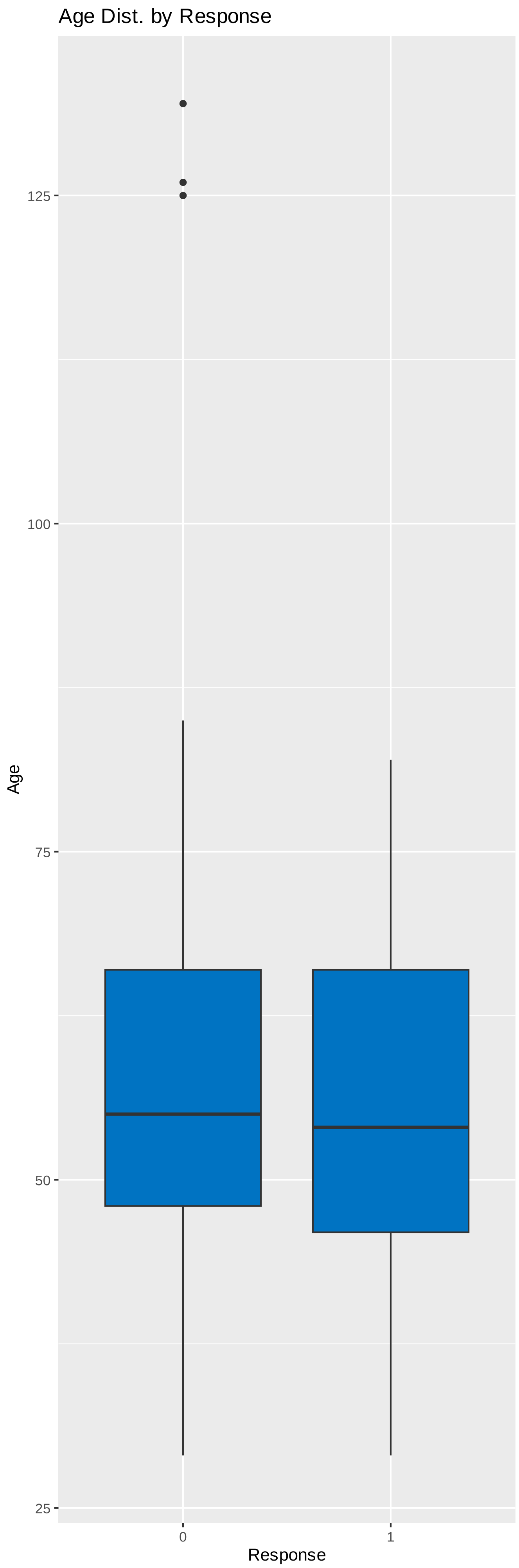


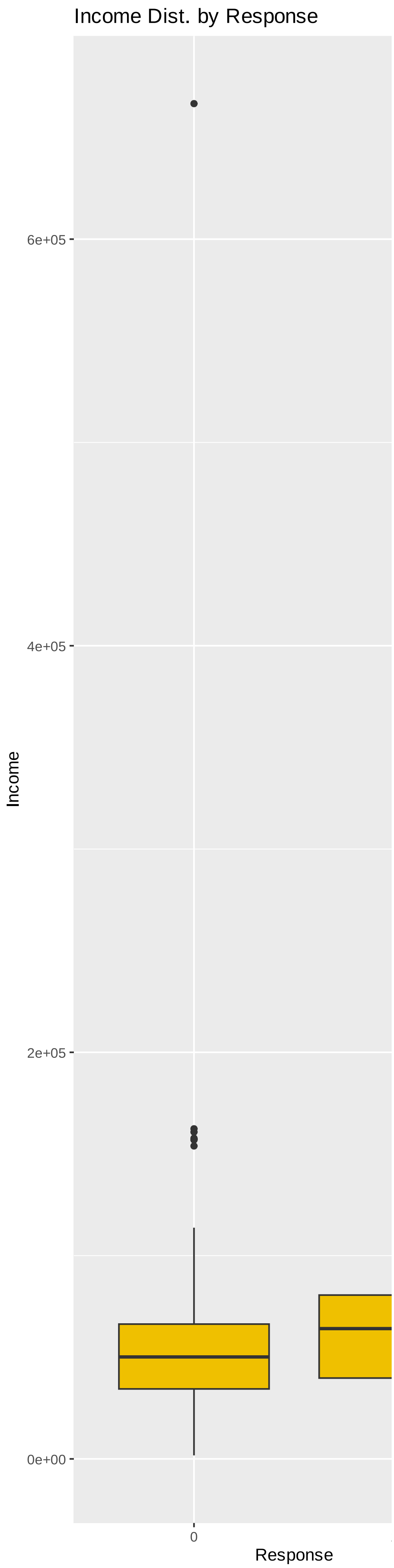
Response Rate by Education Level
Response Rate by Marital status
Proportion
100%
75%
50%
25%
0%
Post Graduate
Under Graduate
Education
Married
Not Married
Marital_Status
Response
0
1

Customers who are married/postgraduates are a bit more likely to respond to the campaign than their counterparts. Interestingly, those who **did not complain** are marginally more responsive than those who did, so I guess satisfaction might influence engagement.

In [141…

```
# I used ChatGPT for the different colors here !
# 2a. Age vs Response
ageplot <- ggplot(df, aes(x = factor(Response), y = Age)) +
  geom_boxplot(fill = "#0073C2FF") +
  labs(title = "Age Dist. by Response", x = "Response", y = "Age")

# 2b. Income vs Response
income_plot <- ggplot(df, aes(x = factor(Response), y = Income))+
  geom_boxplot(fill = "#EFC000FF") +
  labs(title = "Income Dist. by Response",x = "Response", y = "Income")

# 2c. Total Spending vs Response
totalspend_plot <- ggplot(df, aes(x = factor(Response), y = Total_Spendin
  geom_boxplot(fill = "#868686FF") +
  labs(title = "Total Spending by Response", x = "Response", y = "Total S

# Combined plot !
plot_grid(
    ageplot,income_plot,totalspend_plot,
  ncol = 3)
```

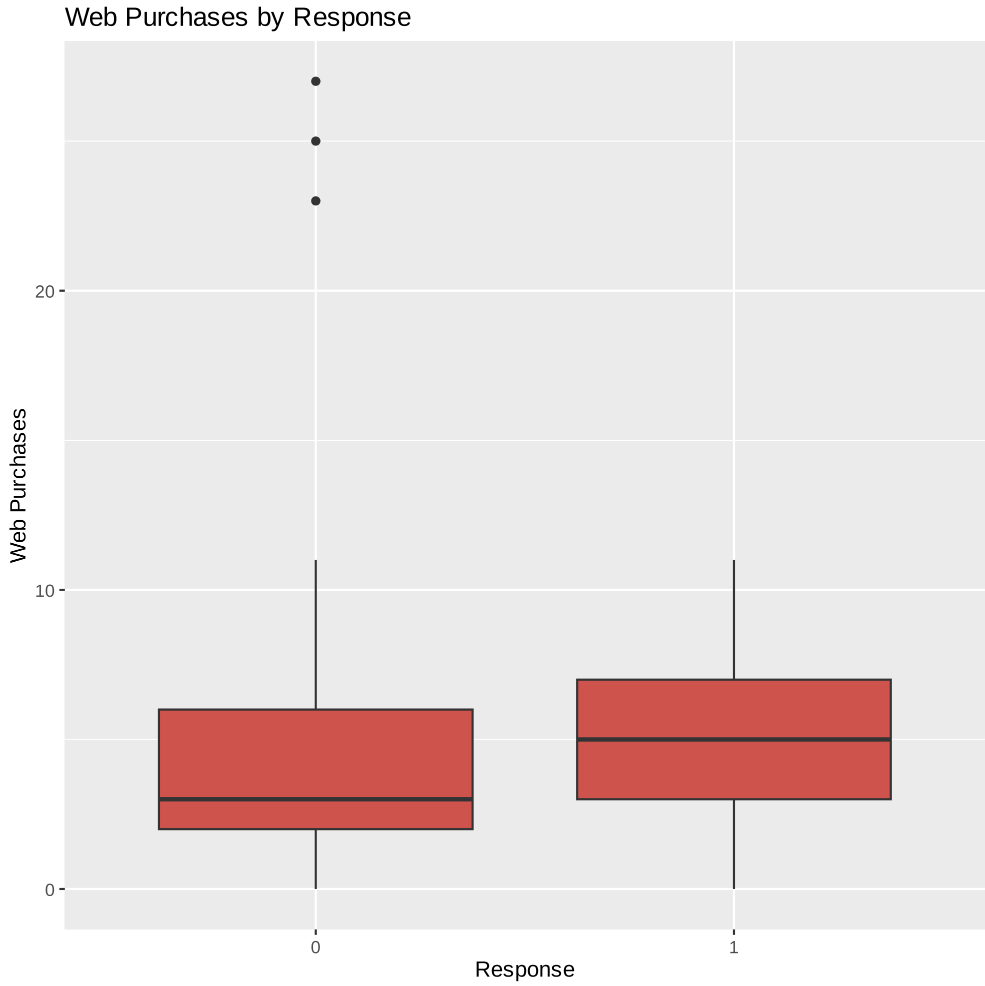

Age Dist. by Response
125
100
75
50
25
Age
0
1
Response

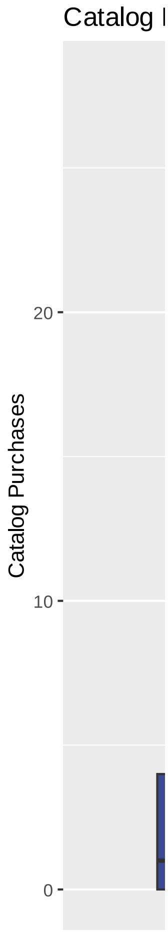

Income Dist. by Response
6e+05
4e+05
2e+05
0e+00
Income
0
Response

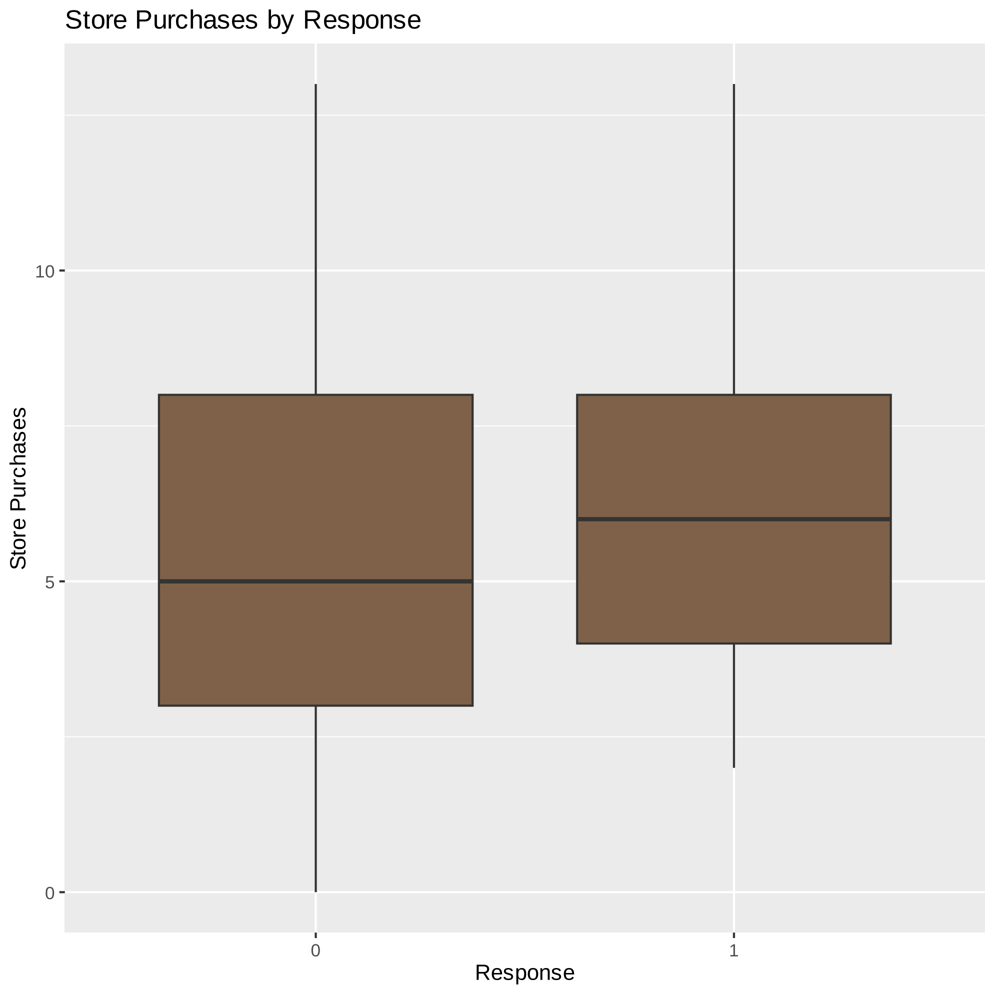
Customers who responded to the marketing campaign tend to have higher total spending and slightly higher income than non-responders. Age distributions are the same across both groups, which means age is less important in predicting campaign response.

In [140…

```
# I used ChatGPT for the different colors here !
# 3A. NumWebPurchases vs Response
NumWebPurchases_plot<-ggplot(df, aes(x = factor(Response), y = NumWebPurc
  geom_boxplot(fill = "#CD534CFF") +
  labs(title = "Web Purchases by Response", x = "Response", y = "Web Purc

# 3B. NumCatalogPurchases vs Response
NumCatalogPurchases_plot<-ggplot(df, aes(x = factor(Response), y = NumCat
  geom_boxplot(fill = "#3B4992FF") +
  labs(title = "Catalog Purchase by Response", x = "Response", y = "Catal

# 3C. NumStorePurchases vs Response
NumStorePurchases_plot<-ggplot(df, aes(x = factor(Response), y = NumStore
  geom_boxplot(fill = "#7E6148FF") +
  labs(title = "Store Purchases by Response", x = "Response", y = "Store 

# 3D. NumWebVisitsMonth vs Response
NumWebVisitsMonth_plot<-ggplot(df, aes(x = factor(Response), y = NumWebVi
  geom_boxplot(fill = "#00A087FF") +
  labs(title = "Web Visits by Response", x = "Response", y = "Monthly Web

NumWebPurchases_plot+NumCatalogPurchases_plot+NumStorePurchases_plot+NumW
```


Web Purchases by Response
Web Purchases
20
10
0
0
1
Response


Catalog P
Catalog Purchases
20
10
0


Store Purchases by Response
Store Purchases
10
5
0
0
1
Response

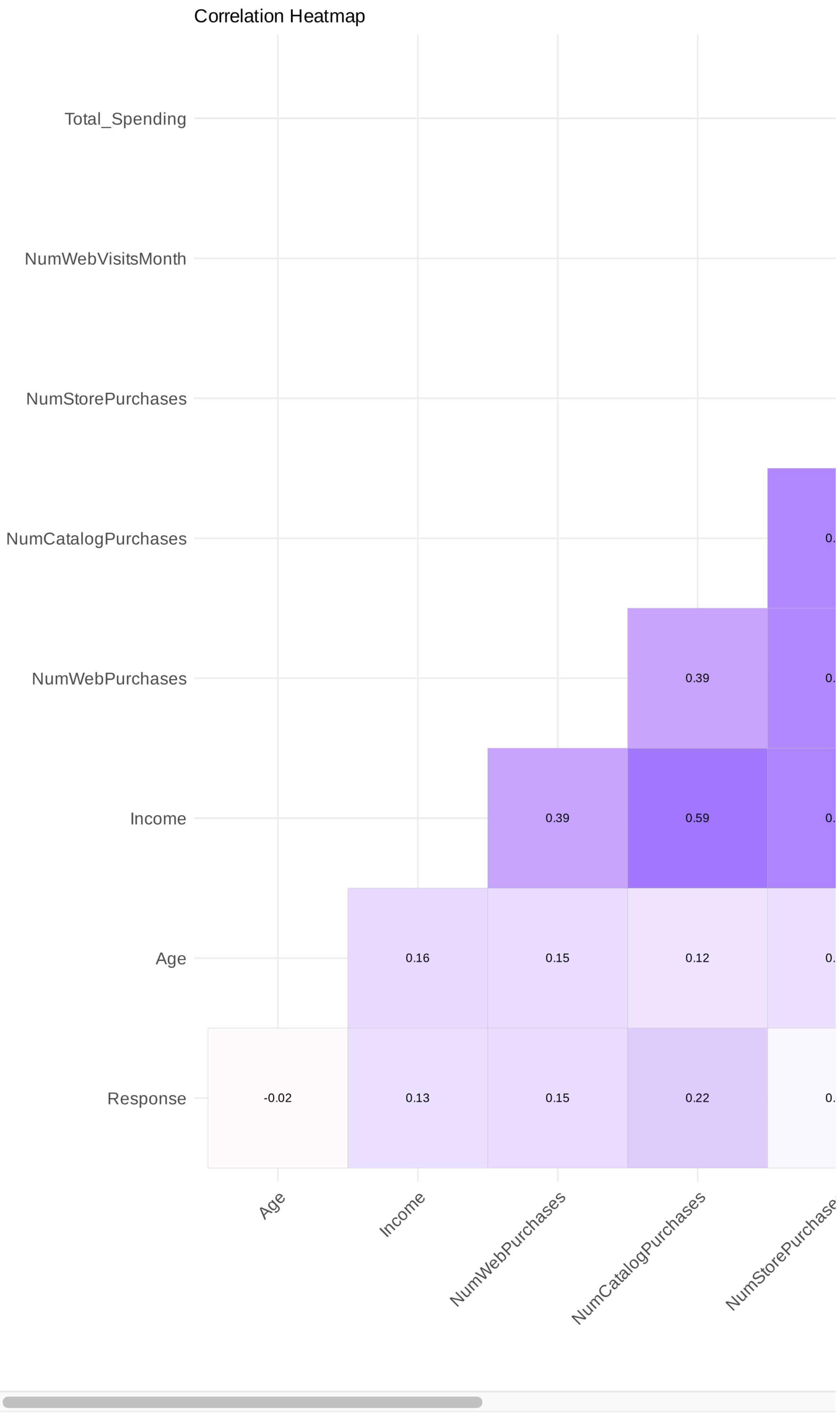

Web Visi
Monthly Web Visits
20
15
10
5
0

Customers who responded to the campaign tend to have slightly higher web and catalog purchases, suggesting a link between online/catalog buying behavior and campaign acceptance. In contrast, store purchases and web visits show little to no difference, indicating they are less important in predicting response.

### 4. Correlation Heatmap (predictors for customer response)

In [145…

```
cor_vars <- df |>
  select(Response, Age, Income, NumWebPurchases, NumCatalogPurchases,
         NumStorePurchases, NumWebVisitsMonth, Total_Spending, Complain)

cor_matrix <- cor(cor_vars, use = "complete.obs")


ggcorrplot(cor_matrix,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           colors = c("red", "white", "blue"),
           title = "Correlation Heatmap",
           show.legend = TRUE)
```


Correlation Heatmap
Total_Spending
NumWebVisitsMonth
NumStorePurchases
NumCatalogPurchases
NumWebPurchases
Income
Age
Response
0.39
0.39
0.59
0.16
0.15
0.12
-0.02
0.13
0.15
0.22
Age
Income
NumWebPurchases
NumCatalogPurchases

From this we see that total_spending is the best predictor of response, number of catalog and web purchases also help, but there is overlap with total spending.

# (4) Methods and Plan

I will use a logistic regression model to predict whether a customer will respond to a marketing campaign (Response is binary). I will use `glm()` with the `family = binomial` argument.

$$\log\left(\frac{P(\text{Response}_i = 1)}{1 - P(\text{Response}_i = 1)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{j,i} + \sum_{\substack{t=1 \\ j \neq k}}^{q} \gamma_t (X_{j,i} \cdot X_{k,i})$$

**Where:**

- $\text{Response}_i$ is the binary outcome (1 = responded, 0 = did not respond)
- $X_{j,i}$ are the predictor variables for observation $i$ (listed above)
- $\beta_0$ is the intercept
- $\beta_j$ are coefficients for main effects
- $\gamma_t$ are coefficients for interaction effects

After fitting the full model, I will choose the variables that I find most significant and simplify the model. I will make sure my model is valid. These are the **diagnostic checks** I plan to perform:

- **Multicollinearity:** Carry out a Variance Inflation Factor (VIF). Variables with high VIFs (>5/>10) are considered for removal to reduce redundancy and improve model interpretability.
- **Linearity of Log-Odds:** Visually verify, to ensure the relationship between continuous predictors and the logit of the response is approximately linear.
- **Independence:** Assumed based on the design of the dataset (no repeated measures or clustering).
- **Goodness-of-Fit:** Checked using a residual Q-Q plot to assess the distribution of deviance residuals. Deviations from the diagonal line would indicate model misfit.

This will improve my models predictive power.

## Justification for Method

Logistic regression is suitable for a binary response variable and allows for interpretability of coefficients.
It accommodates both categorical and continuous predictors.
Adding interaction terms enables detection of heterogeneity across groups (e.g, whether marital status changes the effect of income).

---

## Assumptions

- Observations are independent.
- Response variable is conditionally independant of all other variables given the predictor is included in the model.
- Log-odds of the outcome are linearly related to predictors.
- Predictors are not strongly collinear (variables highly correlated with each other).
- Sufficient sample size to ensure stable estimates.

---

## Limitations

- Causality, this model may not be causal. What if education or marketting affects income and response ? or What if customer loyalty affects reponse ? But we dont record it in the dataset. So we should not make causal claims like "Higher Income causes a higher probability of response"
- Assumes linearity in the logit, which may not hold for all variables.
- Sensitive to outliers and influential points.
- If the relationships we see are highly non-linnear we may need another model (random forrest which we havent covered in class might be better suited)
- Class imbalance (if present) could affect predictive performance.

---

## References

imakash3011. (n.d.). Customer Personality Analysis [Data set]. Kaggle. Retrieved July 27, 2025, from https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data