Project: Exploratory Data Analysis and Machine Learning

Anshul Kumar

2023-12-06

Introduction

This data analysis project will explore the dataset "Marketing Analytics", which contains information on 2205 customers of XYZ company, including their customer profiles, product preferences, campaign successes/failures, and channel performance.

Question:

Given the Marketing Analytics dataset, can we accurately predict the total amount spent by the families in the last two years based on the variables like age, income and number of children in a household.

```
marketing_data <- read.csv("ifood_df.csv")
head(marketing_data)
```

```
##   Income Kidhome Teenhome Recency MntWines MntFruits MntMeatProducts
## 1  58138       0        0      58      635        88             546
## 2  46344       1        1      38       11         1               6
## 3  71613       0        0      26      426        49             127
## 4  26646       1        0      26       11         4              20
## 5  58293       1        0      94      173        43             118
## 6  62513       0        1      16      520        42              98
##   MntFishProducts MntSweetProducts MntGoldProds NumDealsPurchases
## 1             172               88           88                 3
## 2               2                1            6                 2
## 3             111               21           42                 1
## 4              10                3            5                 2
## 5              46               27           15                 5
## 6               0               42           14                 2
##   NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
## 1               8                  10                 4                 7
## 2               1                   1                 2                 5
## 3               8                   2                10                 4
## 4               2                   0                 4                 6
## 5               5                   3                 6                 5
## 6               6                   4                10                 6
##   AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Complain
## 1            0            0            0            0            0        0
## 2            0            0            0            0            0        0
## 3            0            0            0            0            0        0
## 4            0            0            0            0            0        0
## 5            0            0            0            0            0        0
## 6            0            0            0            0            0        0
```

```
##   Z_CostContact Z_Revenue Response Age Customer_Days marital_Divorced
## 1             3        11        1  63          2822                0
## 2             3        11        0  66          2272                0
## 3             3        11        0  55          2471                0
## 4             3        11        0  36          2298                0
## 5             3        11        0  39          2320                0
## 6             3        11        0  53          2452                0
##   marital_Married marital_Single marital_Together marital_Widow
## 1               0              1                0             0
## 2               0              1                0             0
## 3               0              0                1             0
## 4               0              0                1             0
## 5               1              0                0             0
## 6               0              0                1             0
##   education_2n.Cycle education_Basic education_Graduation education_Master
## 1                  0               0                    1                0
## 2                  0               0                    1                0
## 3                  0               0                    1                0
## 4                  0               0                    1                0
## 5                  0               0                    0                0
## 6                  0               0                    0                1
##   education_PhD MntTotal MntRegularProds AcceptedCmpOverall
## 1             0     1529            1441                  0
## 2             0       21              15                  0
## 3             0      734             692                  0
## 4             0       48              43                  0
## 5             1      407             392                  0
## 6             0      702             688                  0
```

Data Preparation and Cleaning

```r
colnames(marketing_data)
```

```
##  [1] "Income"               "Kidhome"              "Teenhome"
##  [4] "Recency"              "MntWines"             "MntFruits"
##  [7] "MntMeatProducts"      "MntFishProducts"      "MntSweetProducts"
## [10] "MntGoldProds"         "NumDealsPurchases"    "NumWebPurchases"
## [13] "NumCatalogPurchases"  "NumStorePurchases"    "NumWebVisitsMonth"
## [16] "AcceptedCmp3"         "AcceptedCmp4"         "AcceptedCmp5"
## [19] "AcceptedCmp1"         "AcceptedCmp2"         "Complain"
## [22] "Z_CostContact"        "Z_Revenue"            "Response"
## [25] "Age"                  "Customer_Days"        "marital_Divorced"
## [28] "marital_Married"      "marital_Single"       "marital_Together"
## [31] "marital_Widow"        "education_2n.Cycle"   "education_Basic"
## [34] "education_Graduation" "education_Master"     "education_PhD"
## [37] "MntTotal"             "MntRegularProds"      "AcceptedCmpOverall"
```

Column Description:

Income - Customer's annual family income

Kidhome - Number of children in the customer's family

Teenhome - Number of teenagers in the customer's family

Recency - Number of days since the last purchase

MntWines - Amount spent on wines in the last 2 years

MntFruits - Amount spent on fruits in the last 2 years

MntMeatProducts - Amount spent on meat products in the last 2 years

MntFishProducts - Amount spent on fish products in the last 2 years

MntSweetProducts - Amount spent on sweet products in the last 2 years

MntGoldProds - Amount spent on gold products in the last 2 years

NumDealsPurchases - Number of purchases made with a discount

NumWebPurchases - Number of purchases made through the company's website

NumCatalogPurchases - Number of purchases made using catalogs

NumStorePurchases - Number of purchases made directly in stores

NumWebVisitsMonth - Number of visits to the company's website in the last month

AcceptedCmp3 - 1 if the customer accepted the offer in the 3rd campaign, 0 otherwise

AcceptedCmp4 - 1 if the customer accepted the offer in the 4th campaign, 0 otherwise

AcceptedCmp5 - 1 if the customer accepted the offer in the 5th campaign, 0 otherwise

AcceptedCmp1 - 1 if the customer accepted the offer in the 1st campaign, 0 otherwise

AcceptedCmp2 - 1 if the customer accepted the offer in the 2nd campaign, 0 otherwise

Complain - 1 if the customer complained in the last 2 years

Z_CostContact - ????

Z_Revenue - ????

Response (Target) - 1 if the customer accepted the offer in the last campaign, 0 otherwise

Age - Customer's age

Customer_Days - Days since customer's registration

marital_Divorced - Customer's marital status is divorced

marital_Married - Customer's marital status is married

marital_Single - Customer's marital status is single

marital_Together - Customer's marital status is together

marital_Widow - Customer's marital status is widow

education_2n - Cycle - Customer's education level is 2nd cycle

education_Basic - Customer's education level is basic

education_Graduation - Customer's education level is graduation

education_Master - Customer's education level is master's

education_PhD - Customer's education level is PhD

MntTotal - Total amount spent in the last 2 years

MntRegularProds - Amount spent on regular products in the last 2 years

AcceptedCmpOverall - Sum of AcceptedCmp campaigns

Checking for missing value(s)

```
colSums(is.na(marketing_data))
```

```
##                Income               Kidhome              Teenhome 
##                     0                     0                     0 
##               Recency              MntWines            MntFruits 
##                     0                     0                     0 
##       MntMeatProducts       MntFishProducts      MntSweetProducts 
##                     0                     0                     0 
##          MntGoldProds     NumDealsPurchases       NumWebPurchases 
##                     0                     0                     0 
##   NumCatalogPurchases     NumStorePurchases     NumWebVisitsMonth 
##                     0                     0                     0 
##          AcceptedCmp3          AcceptedCmp4          AcceptedCmp5 
##                     0                     0                     0 
##          AcceptedCmp1          AcceptedCmp2              Complain 
##                     0                     0                     0 
##         Z_CostContact             Z_Revenue              Response 
##                     0                     0                     0 
##                   Age         Customer_Days      marital_Divorced 
##                     0                     0                     0 
##       marital_Married        marital_Single      marital_Together 
##                     0                     0                     0 
##         marital_Widow   education_2n.Cycle       education_Basic 
##                     0                     0                     0 
## education_Graduation      education_Master         education_PhD 
##                     0                     0                     0 
##              MntTotal       MntRegularProds    AcceptedCmpOverall 
##                     0                     0                     0
```

No column in the dataframe `marketing_data` have any missing value.

Checking data types for columns

```
str(marketing_data)
```

```
## 'data.frame':    2205 obs. of  39 variables:
##  $ Income              : num  58138 46344 71613 26646 58293 ...
##  $ Kidhome             : int  0 1 0 1 1 0 0 1 1 1 ...
##  $ Teenhome            : int  0 1 0 0 0 1 1 0 0 1 ...
##  $ Recency             : int  58 38 26 26 94 16 34 32 19 68 ...
##  $ MntWines            : int  635 11 426 11 173 520 235 76 14 28 ...
##  $ MntFruits           : int  88 1 49 4 43 42 65 10 0 0 ...
##  $ MntMeatProducts     : int  546 6 127 20 118 98 164 56 24 6 ...
##  $ MntFishProducts     : int  172 2 111 10 46 0 50 3 3 1 ...
##  $ MntSweetProducts    : int  88 1 21 3 27 42 49 1 3 1 ...
##  $ MntGoldProds        : int  88 6 42 5 15 14 27 23 2 13 ...
##  $ NumDealsPurchases   : int  3 2 1 2 5 2 4 2 1 1 ...
##  $ NumWebPurchases     : int  8 1 8 2 5 6 7 4 3 1 ...
##  $ NumCatalogPurchases : int  10 1 2 0 3 4 3 0 0 0 ...
##  $ NumStorePurchases   : int  4 2 10 4 6 10 7 4 2 0 ...
```

```
## $ NumWebVisitsMonth   : int  7 5 4 6 5 6 6 8 9 20 ...
## $ AcceptedCmp3        : int  0 0 0 0 0 0 0 0 0 1 ...
## $ AcceptedCmp4        : int  0 0 0 0 0 0 0 0 0 0 ...
## $ AcceptedCmp5        : int  0 0 0 0 0 0 0 0 0 0 ...
## $ AcceptedCmp1        : int  0 0 0 0 0 0 0 0 0 0 ...
## $ AcceptedCmp2        : int  0 0 0 0 0 0 0 0 0 0 ...
## $ Complain            : int  0 0 0 0 0 0 0 0 0 0 ...
## $ Z_CostContact       : int  3 3 3 3 3 3 3 3 3 3 ...
## $ Z_Revenue           : int  11 11 11 11 11 11 11 11 11 11 ...
## $ Response            : int  1 0 0 0 0 0 0 0 1 0 ...
## $ Age                 : int  63 66 55 36 39 53 49 35 46 70 ...
## $ Customer_Days       : int  2822 2272 2471 2298 2320 2452 2752 2576 2547 2267 ...
## $ marital_Divorced    : int  0 0 0 0 0 0 1 0 0 0 ...
## $ marital_Married     : int  0 0 0 0 1 0 0 1 0 0 ...
## $ marital_Single      : int  1 1 0 0 0 0 0 0 0 0 ...
## $ marital_Together    : int  0 0 1 1 0 1 0 0 1 1 ...
## $ marital_Widow       : int  0 0 0 0 0 0 0 0 0 0 ...
## $ education_2n.Cycle  : int  0 0 0 0 0 0 0 0 0 0 ...
## $ education_Basic     : int  0 0 0 0 0 0 0 0 0 0 ...
## $ education_Graduation: int  1 1 1 1 0 0 1 0 0 0 ...
## $ education_Master    : int  0 0 0 0 0 1 0 0 0 0 ...
## $ education_PhD       : int  0 0 0 0 1 0 0 1 1 1 ...
## $ MntTotal            : int  1529 21 734 48 407 702 563 146 44 36 ...
## $ MntRegularProds     : int  1441 15 692 43 392 688 536 123 42 23 ...
## $ AcceptedCmpOverall  : int  0 0 0 0 0 0 0 0 0 1 ...
```

All the columns in the dataframe have appropriate data types. Hence, we don't need to change the data type of any of the variables.

Checking for unique values

```r
for (i in colnames(marketing_data)){
  print(sprintf("%s - %.0f", i, length(unique(marketing_data[[i]]))), quote = FALSE)
}
```

```
## [1] Income - 1963
## [1] Kidhome - 3
## [1] Teenhome - 3
## [1] Recency - 100
## [1] MntWines - 775
## [1] MntFruits - 158
## [1] MntMeatProducts - 551
## [1] MntFishProducts - 182
## [1] MntSweetProducts - 176
## [1] MntGoldProds - 212
## [1] NumDealsPurchases - 15
## [1] NumWebPurchases - 15
## [1] NumCatalogPurchases - 13
## [1] NumStorePurchases - 14
## [1] NumWebVisitsMonth - 16
## [1] AcceptedCmp3 - 2
## [1] AcceptedCmp4 - 2
## [1] AcceptedCmp5 - 2
## [1] AcceptedCmp1 - 2
```

```
## [1] AcceptedCmp2 - 2
## [1] Complain - 2
## [1] Z_CostContact - 1
## [1] Z_Revenue - 1
## [1] Response - 2
## [1] Age - 56
## [1] Customer_Days - 662
## [1] marital_Divorced - 2
## [1] marital_Married - 2
## [1] marital_Single - 2
## [1] marital_Together - 2
## [1] marital_Widow - 2
## [1] education_2n.Cycle - 2
## [1] education_Basic - 2
## [1] education_Graduation - 2
## [1] education_Master - 2
## [1] education_PhD - 2
## [1] MntTotal - 897
## [1] MntRegularProds - 974
## [1] AcceptedCmpOverall - 5
```

We can see that variables `Z_CostContact` and `Z_Revenue` have same value for all the columns. Therefore, removing them from the dataframe will not affect our analysis.

```
marketing_data = subset(marketing_data, select = -c(Z_CostContact, Z_Revenue))
colnames(marketing_data)
```

```
##  [1] "Income"              "Kidhome"              "Teenhome"
##  [4] "Recency"             "MntWines"             "MntFruits"
##  [7] "MntMeatProducts"     "MntFishProducts"      "MntSweetProducts"
## [10] "MntGoldProds"        "NumDealsPurchases"    "NumWebPurchases"
## [13] "NumCatalogPurchases" "NumStorePurchases"    "NumWebVisitsMonth"
## [16] "AcceptedCmp3"        "AcceptedCmp4"         "AcceptedCmp5"
## [19] "AcceptedCmp1"        "AcceptedCmp2"         "Complain"
## [22] "Response"            "Age"                  "Customer_Days"
## [25] "marital_Divorced"    "marital_Married"      "marital_Single"
## [28] "marital_Together"    "marital_Widow"        "education_2n.Cycle"
## [31] "education_Basic"     "education_Graduation" "education_Master"
## [34] "education_PhD"       "MntTotal"             "MntRegularProds"
## [37] "AcceptedCmpOverall"
```

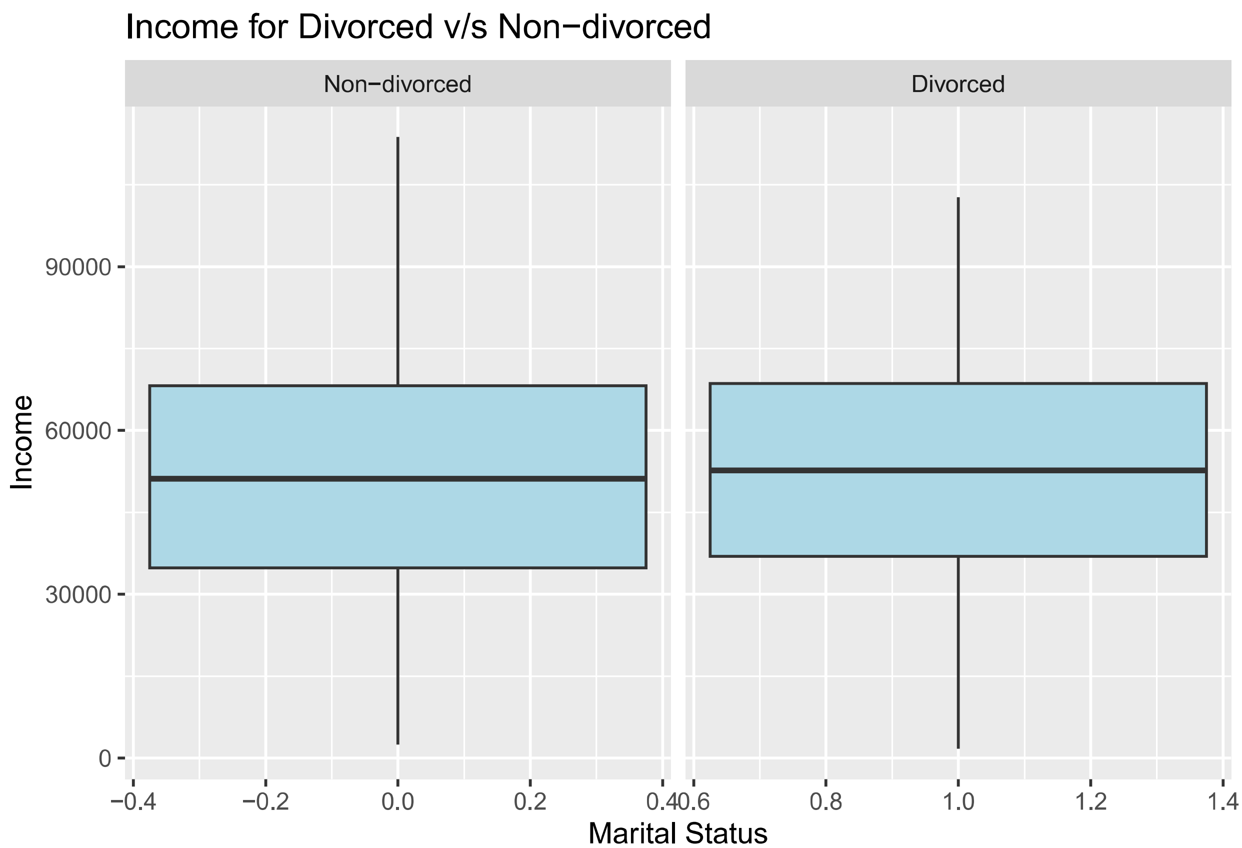
Data Exploration

Box plot of income for divorced v/s non-divorced

```
library(ggplot2)

labels <- c("0" = "Non-divorced", "1" = "Divorced")
ggplot(marketing_data, aes(x = marital_Divorced, y = Income)) +
  geom_boxplot(fill = "lightblue") +
  ggtitle("Income for Divorced v/s Non-divorced") +
  xlab("Marital Status") +
```

```r
  ylab("Income") +
  facet_wrap(~marital_Divorced, scales = "free_x", labeller = as_labeller(labels))
```

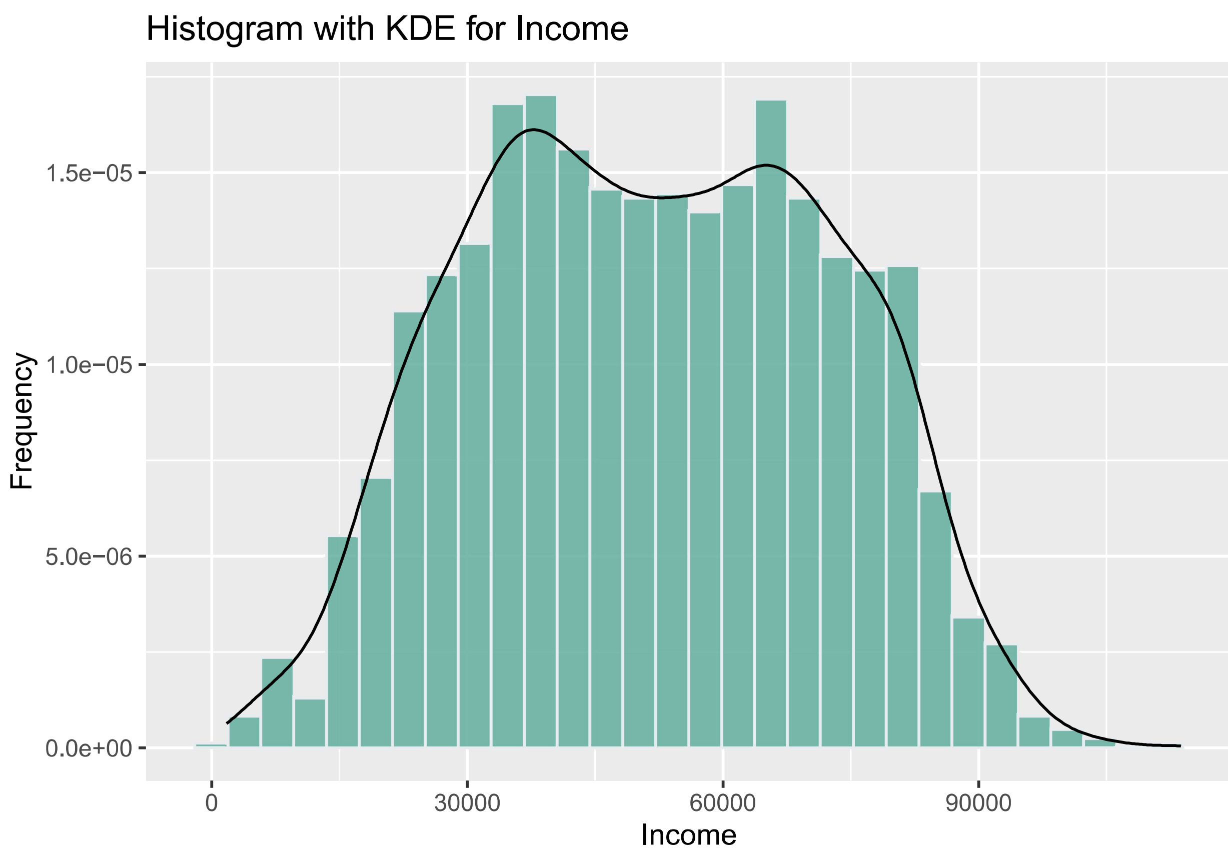


We can see that even though the median income of **divorced** is slightly more than **non-divorced**, the maximum income (uppermost quantile) is far more for **non-divorced** than **divorced**.

Histogram for Income

```r
library(ggplot2)

ggplot(marketing_data, aes(x = Income)) +
  geom_histogram(aes(y = after_stat(density)), fill = "#69b3a2", color = "#e9ecef", alpha = 0.9) +
  geom_density() +
  ggtitle("Histogram with KDE for Income") +
  xlab("Income") +
  ylab("Frequency")
```



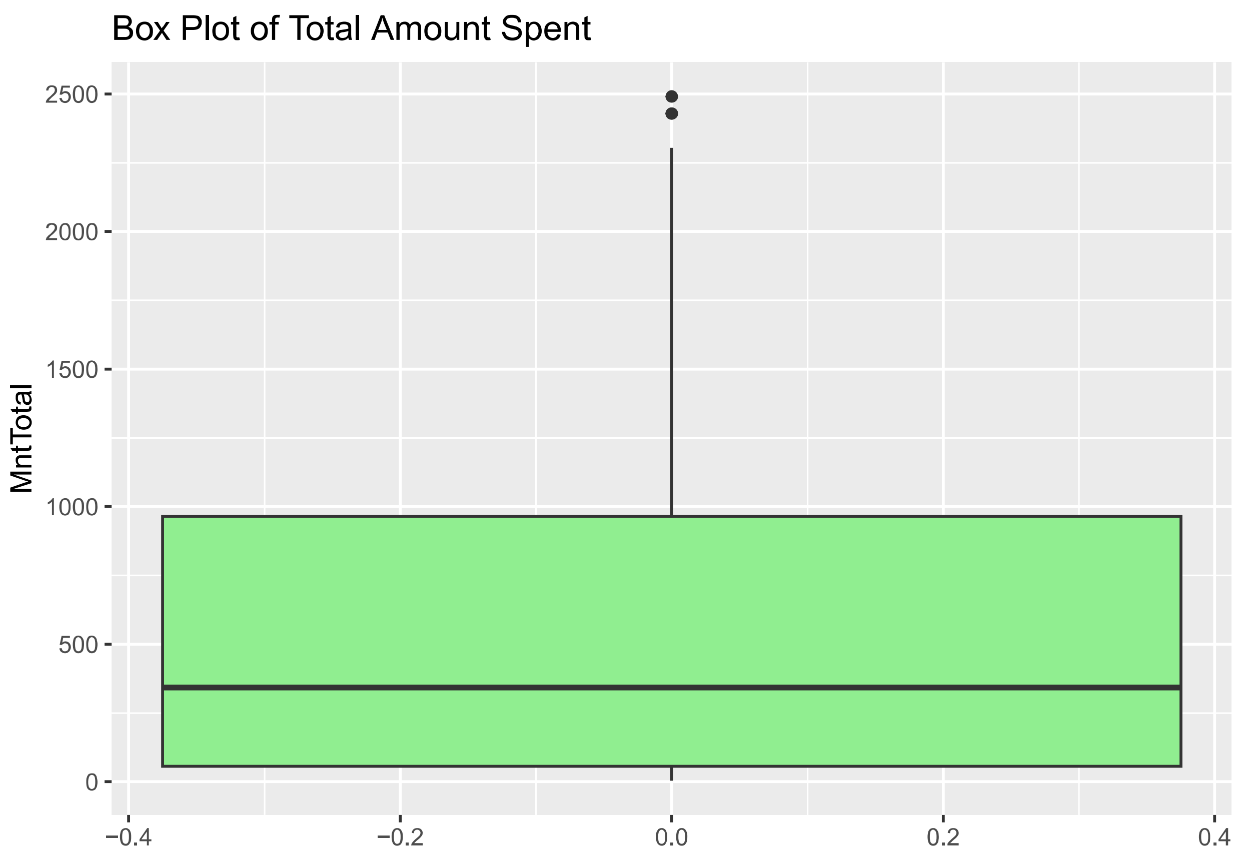
We can observe from the above histogram that income distribution closely resembles the normal distribution. We can also note that there are no outliers as well.

Box plot for MntTotal

`MntTotal` is the total amount spent on all products over last two years.

```
library(ggplot2)

ggplot(marketing_data, aes(y = MntTotal)) +
  geom_boxplot(fill = "lightgreen") +
  ggtitle("Box Plot of Total Amount Spent") +
  ylab("MntTotal")
```



We can observe that there are a few outliers.

To remove the outliers, we can use interquartile range. Interquantile range is the difference between 1st quantile (25th percentile) and 3rd quantile (75th percentile).

```
Q1 <- quantile(marketing_data$MntTotal, 0.25)
Q3 <- quantile(marketing_data$MntTotal, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR
outliers <- marketing_data[(marketing_data$MntTotal < lower) | (marketing_data$MntTotal > upper), ]
head(outliers)
```

```
##      Income Kidhome Teenhome Recency MntWines MntFruits MntMeatProducts
## 1160  90638       0        0      29     1156       120             915
## 1468  87679       0        0      62     1259       172             815
## 1548  90638       0        0      29     1156       120             915
##      MntFishProducts MntSweetProducts MntGoldProds NumDealsPurchases
## 1160              94              144           96                 1
## 1468              97              148           33                 1
## 1548              94              144           96                 1
##      NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
## 1160               3                   4                10                 1
## 1468               7                  11                10                 4
## 1548               3                   4                10                 1
##      AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Complain
## 1160            0            0            1            0            0        0
```

```
## 1468            1             0             1             1             0        0
## 1548            0             0             1             0             0        0
##      Response Age Customer_Days marital_Divorced marital_Married marital_Single
## 1160        0  29          2295                0               0              1
## 1468        1  32          2496                0               0              0
## 1548        1  29          2295                0               0              1
##      marital_Together marital_Widow education_2n.Cycle education_Basic
## 1160                0             0                  0               0
## 1468                1             0                  0               0
## 1548                0             0                  0               0
##      education_Graduation education_Master education_PhD MntTotal
## 1160                    0                1             0     2429
## 1468                    1                0             0     2491
## 1548                    0                1             0     2429
##      MntRegularProds AcceptedCmpOverall
## 1160            2333                  1
## 1468            2458                  3
## 1548            2333                  1
```

Removing outliers:

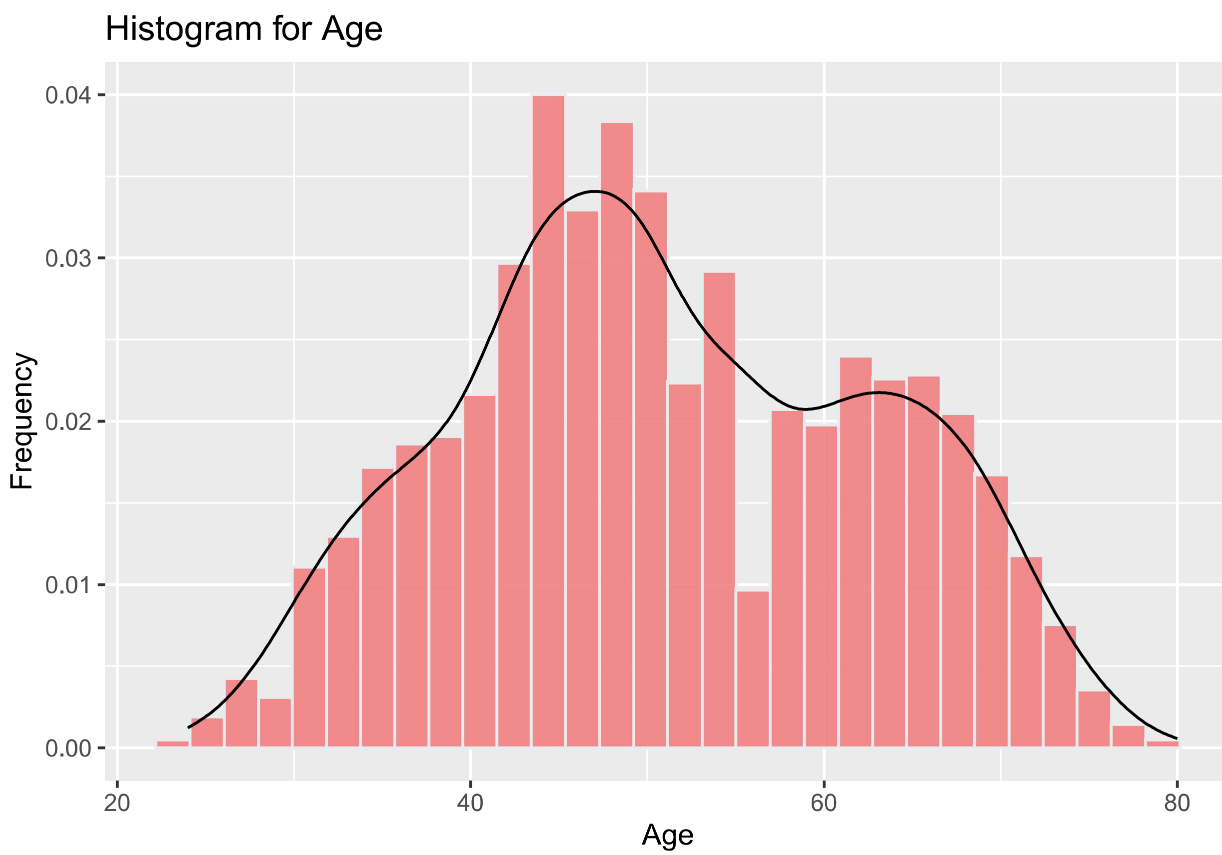
```r
marketing_data <- marketing_data[(marketing_data$MntTotal < upper) & marketing_data$MntTotal > lower, ]
summary(marketing_data$MntTotal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    56.0   342.5   560.2   962.0  2304.0
```

Histogram for age

```r
library(ggplot2)

ggplot(marketing_data, aes(x = Age)) +
  geom_histogram(aes(y = after_stat(density)), fill = '#f08080', color = '#e9ecef', alpha = 0.9) +
  geom_density() +
  ggtitle("Histogram for Age") +
  xlab("Age") +
  ylab("Frequency")
```



We can see from the above graph that the most responsive age group is 45 to 49 years old.

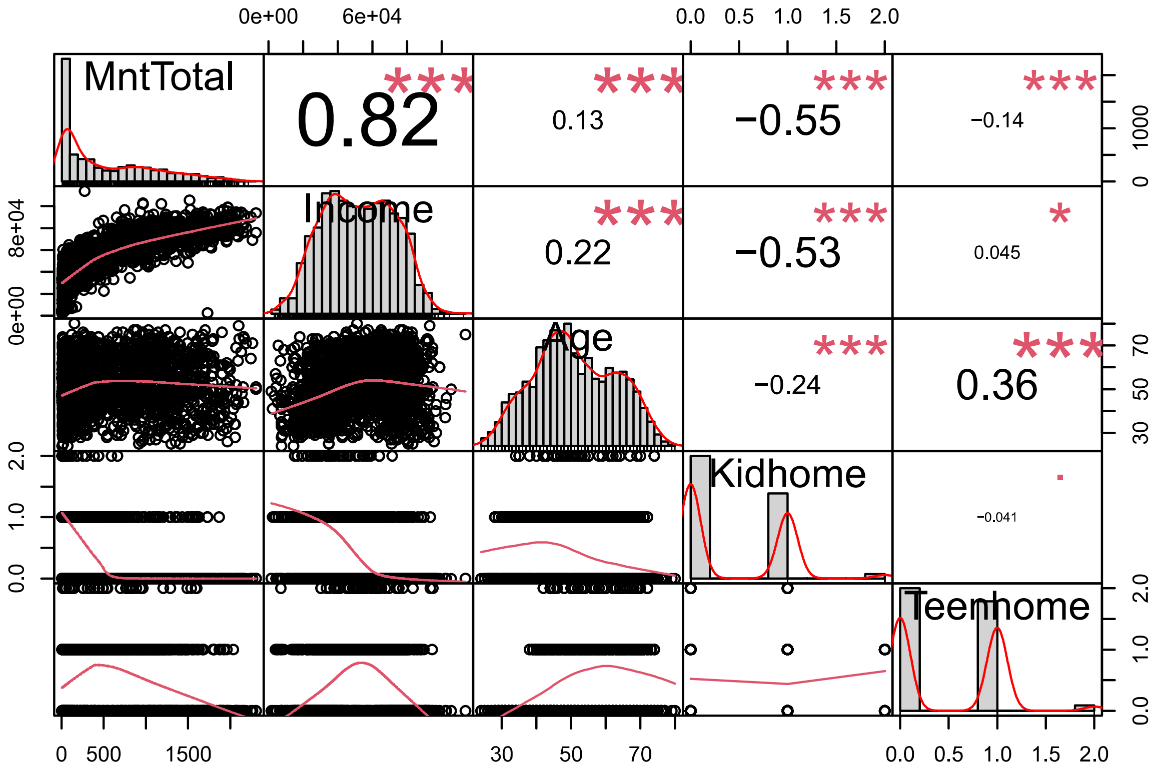
Now, we will explore the correlation between important **numerical features** in the dataframe `marketing_data` and total amount spent `MntTotal`.

We will use chart.Correlation() function from "PerformanceAnalytics" library to get correlation between the variables and their distribution as well.

The correlation calculated by chart.Correlation() function calculates **Pearson Correlation Coefficient** by default, which tells the linear correlation between the variables. Therefore, we have to keep in mind that if there exists a strong non-linear correlation between variables, the **Pearson Correlation Coefficient** will be 0.

```
library(PerformanceAnalytics)

cor_chart <- marketing_data[, c("MntTotal", "Income", "Age", "Kidhome", "Teenhome")]
chart.Correlation(cor_chart, histogram = TRUE, pch = 19)
```



From the graph above, we can see that:

- The total amount of money spent `MntTotal` is strongly correlated to `Income`.
- There is a moderate negative relationship between `MntTotal` and the number of children in the household (`Kidhome`).
- The negative correlation between `Kidhome` and `Income` is nearly the same as the negative correlation between `Kidhome` and `MntTotal`.

Linear Modelling

Now we will analyze `MntTotal` and `Income` further using linear modelling.

```
marketing_lm <- lm(Income ~ MntTotal, data = marketing_data)
summary(marketing_lm)
```

```
##
## Call:
## lm(formula = Income ~ MntTotal, data = marketing_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83885  -7704    385   7718  70676
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34909.347    350.656   99.55   <2e-16 ***
## MntTotal       29.741      0.438   67.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11760 on 2200 degrees of freedom
## Multiple R-squared:  0.677,  Adjusted R-squared:  0.6768
## F-statistic:  4611 on 1 and 2200 DF,  p-value: < 2.2e-16
```

From the above linear model summary, we can conclude the following points:

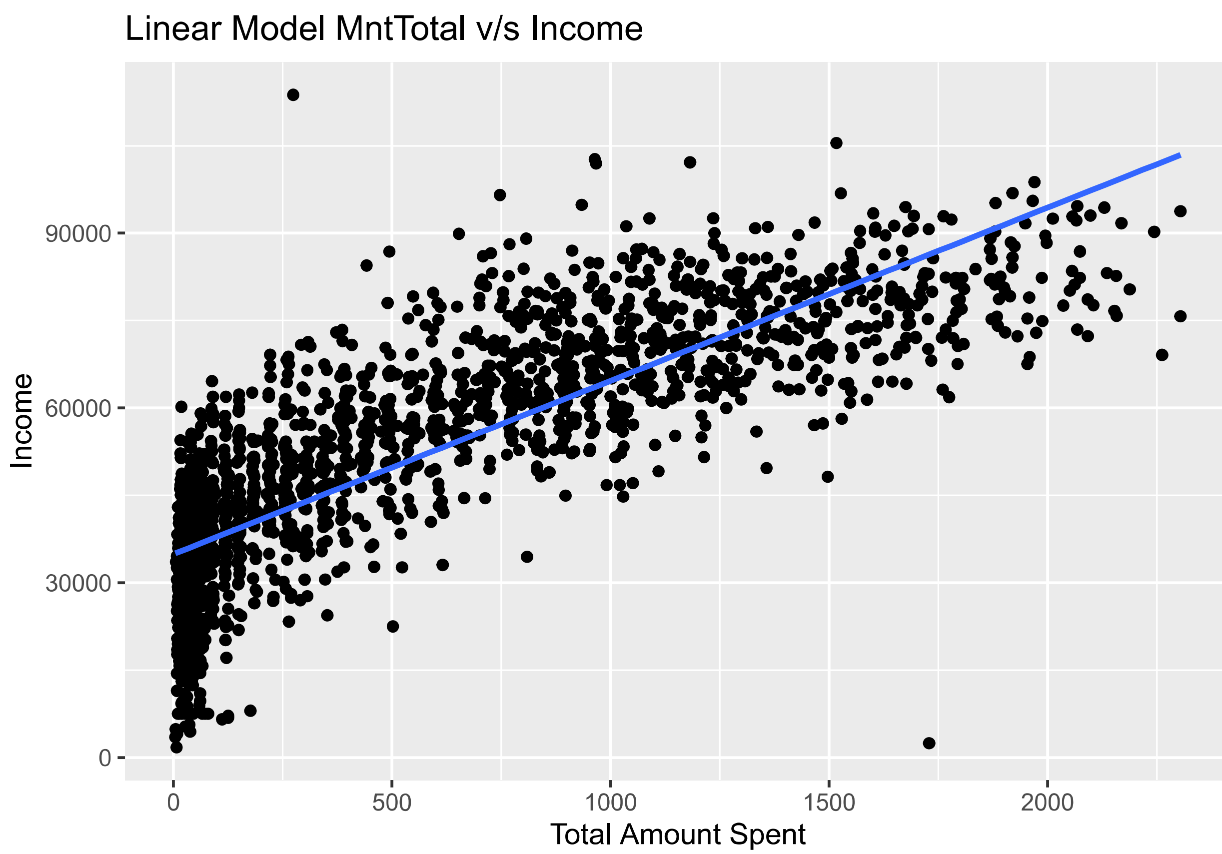
- Under **Coefficients** section, The "Estimate" column provides the Least Squares estimate for the fitted line.
- Equation of fitted line:
$$\text{MntTotal} = 34909.347 + 29.741 \times \text{Income}$$
- "Standard Error" is the average amount that the estimate varies from our actual value.
- "t-value" is a measure of how far an estimate is from zero, in units of standard errors. It is calculated by dividing the estimate by its standard error. The higher the t-value, the more likely it is that the estimate is different from zero by chance.
- "p-values" are calculated based on the t-value and standard error, and if the p-value is less than or equal to 0.05, then the coefficient is statistically significant. In our case, p-value is extremely low ($<$ 2.2e-16).

Hence, we can conclude that there is a direct relation between `Income` and `MntTotal`.

Visualizing the linear model

```r
library(ggplot2)

ggplot(data = marketing_data, aes(x = MntTotal, y = Income)) + geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Linear Model MntTotal v/s Income") +
  xlab("Total Amount Spent") +
  ylab("Income")
```



Supervised Machine learning - Linear Regression

Till now, we have cleaned the data, analyzed the `MntTotal` column whether it is linear or not, and removed outliers from the column.

Now, we can create a **Linear Regression** model, in which we can try predicting the values of `MntTotal` based on the variables we used to create correlation chart, which include `Income`, `Age`, `Kidhome`, `Teenhome`.

We will first split the data into 80% training and 20% test sets, and then create a linear model for the training set.

```
library(dplyr)

marketing_data_subset <- marketing_data[, c("MntTotal", "Income", "Age", "Kidhome", "Teenhome")]

set.seed(123)  #For reproducibility

train_index <- sample(seq_len(nrow(marketing_data_subset)), size = 0.8 * nrow(marketing_data_subset))
train_data <- marketing_data_subset[train_index, ]
test_data <- marketing_data_subset[-train_index, ]

model <- lm(MntTotal ~ ., data = train_data)

summary(model)
```

```
##
```

```
## Call:
## lm(formula = MntTotal ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1736.78  -188.85   -19.73   159.76  1180.41
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.262e+02  4.043e+01  -8.068 1.31e-15 ***
## Income       2.086e-02  4.032e-04  51.734  < 2e-16 ***
## Age         -4.729e-01  6.633e-01  -0.713    0.476
## Kidhome     -1.706e+02  1.554e+01 -10.975  < 2e-16 ***
## Teenhome    -1.820e+02  1.385e+01 -13.145  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292.3 on 1756 degrees of freedom
## Multiple R-squared:  0.7367, Adjusted R-squared:  0.7361
## F-statistic:  1228 on 4 and 1756 DF,  p-value: < 2.2e-16
```

From the above summary, we can conclude that:

- The linear regression model is a fine fit for the data, with an R-squared value of 0.7367. This indicates that the model explains 73.67% of the variation in the total amount of money spent on marketing.
- All of the independent variables in the model are statistically significant (except for `Age`), with p-values less than 0.05. This means that they are all significantly associated with the total amount of money spent on marketing.
- The number of children in the household (`Kidhome` and `Teenhome`) have negative and statistically significant effects on the total amount of money spent. This means that people with more children tend to spend less money in general.

Now that we have trained our model on `train_data`, we can use it to make prediction on the `test_data`.

```r
predictions <- predict(model, newdata = test_data)
results <- data.frame(Actual = test_data$MntTotal, Predicted = predictions)
head(results, 10)
```

```
##    Actual Predicted
## 3     734 1141.6157
## 21   1729 -465.1192
## 22    953  680.7217
## 28   1672 1412.8945
## 42     19 -361.6075
## 43    810 1320.5366
## 47    493  827.9982
## 50   1376 1376.0997
## 53   1053 1114.4289
## 57    606  690.2717
```

Furthermore, we can use Mean Square Error (MSE) value to get an idea of how well the model predicts the target variable, `MntTotal` in our case.

```
residuals <- test_data$MntTotal - predictions
mse <- mean(residuals^2)

mse
```

```
## [1] 104581.9
```

A Mean Square Error (MSE) of 104581.9 means that the model's predictions are, on average, approximately 323.7 units away from the actual values (since the square root of 104581.9 is about 323.7).

Hence, we can conclude that although the model is a fine fit for the dataframe `marketing_data`, other machine learning models might be able to provide a better fit.

Summary

This project aimed to analyze the relationship between various variables in the data set, especially the total amount spent in the last 2 years (`MntTotal`) and other features like `Age`, `Income`, etc. Additionally, using a linear regression machine learning model, we were able to predict the `MntTotal` amount using the variables `Income`, `Age`, `Kidhome` and `Teenhome` with an R-squared value of 0.7367 (73.67%) and Mean Square Error (MSE) of 104581.9.

Potential areas for further investigation

- Effect of binary variables like `marital_Divorced`, `marital_Married`, etc., on `MntTotal`.
- Analysis of educational impact on income.
- Analysis of other significant variables like `MntWines`, `MntFruits`, etc.
- Testing other machine learning algorithms for better prediction of `MntTotal`.

Citation

Daoud, J. (2021, July). Marketing Data, Version 1. Retrieved October 31, 2023 from Kaggle.