

# Individual Assignment 3: Exploratory Data Analysis and Visualization

Atharv Prashant Tungatkar

2025-10-26

Reading libraries and datasets

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
##  
## 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
```

```
data<-read.csv("C:/Users/athar/OneDrive/Desktop/MBA Business Analytics/Visual Analytics/Files/marketing_campaign.csv")
```

Displaying data

```
head(data,5)
```

```
##      ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524      1957 Graduation      Single  58138      0      0 9/4/2012
## 2 2174      1954 Graduation      Single  46344      1      1 3/8/2014
## 3 4141      1965 Graduation Together  71613      0      0 8/21/2013
## 4 6182      1984 Graduation Together  26646      1      0 2/10/2014
## 5 5324      1981      PhD      Married  58293      1      0 1/19/2014
##      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
##      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
##      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
##      AcceptedCmp1 AcceptedCmp2 Complain Response
## 1      0      0      0      1
## 2      0      0      0      0
## 3      0      0      0      0
## 4      0      0      0      0
## 5      0      0      0      0
```

```
str(data)
```

```
## 'data.frame':    2216 obs. of  27 variables:
## $ ID              : int  5524 2174 4141 6182 5324 7446 965 6177 4855 5899 ...
## $ Year_Birth       : int  1957 1954 1965 1984 1981 1967 1971 1985 1974 1950 ...
## $ Education        : chr   "Graduation" "Graduation" "Graduation" "Graduation" ...
## $ Marital_Status   : chr   "Single" "Single" "Together" "Together" ...
## $ Income            : int  58138 46344 71613 26646 58293 62513 55635 33454 30351 5648
## ...
## $ Kidhome          : int    0 1 0 1 1 0 0 1 1 1 ...
## $ Teenhome          : int    0 1 0 0 0 1 1 0 0 1 ...
## $ Dt_Customer       : chr   "9/4/2012" "3/8/2014" "8/21/2013" "2/10/2014" ...
## $ 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 ...
## $ Response           : int     1 0 0 0 0 0 0 0 1 0 ...
```

## Univariate Non Graphical

### Categorical variable

Education

```
tb<-table(data$Education)
tb
```

```
##
##  2n Cycle      Basic Graduation      Master      PhD
##      200         54         1116         365         481
```

The analysis uncover that majority of the customers have completed Graduation.

### Quantitative Variable

Recency

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.4.3
```

```
##  
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':  
##  
##    %+%, alpha
```

```
describe(data$Recency)
```

```
##      vars      n  mean      sd median trimmed   mad min max range skew kurtosis   se  
## X1      1 2216 49.01 28.95      49   48.99 37.06    0  99   99     0     -1.2 0.61
```

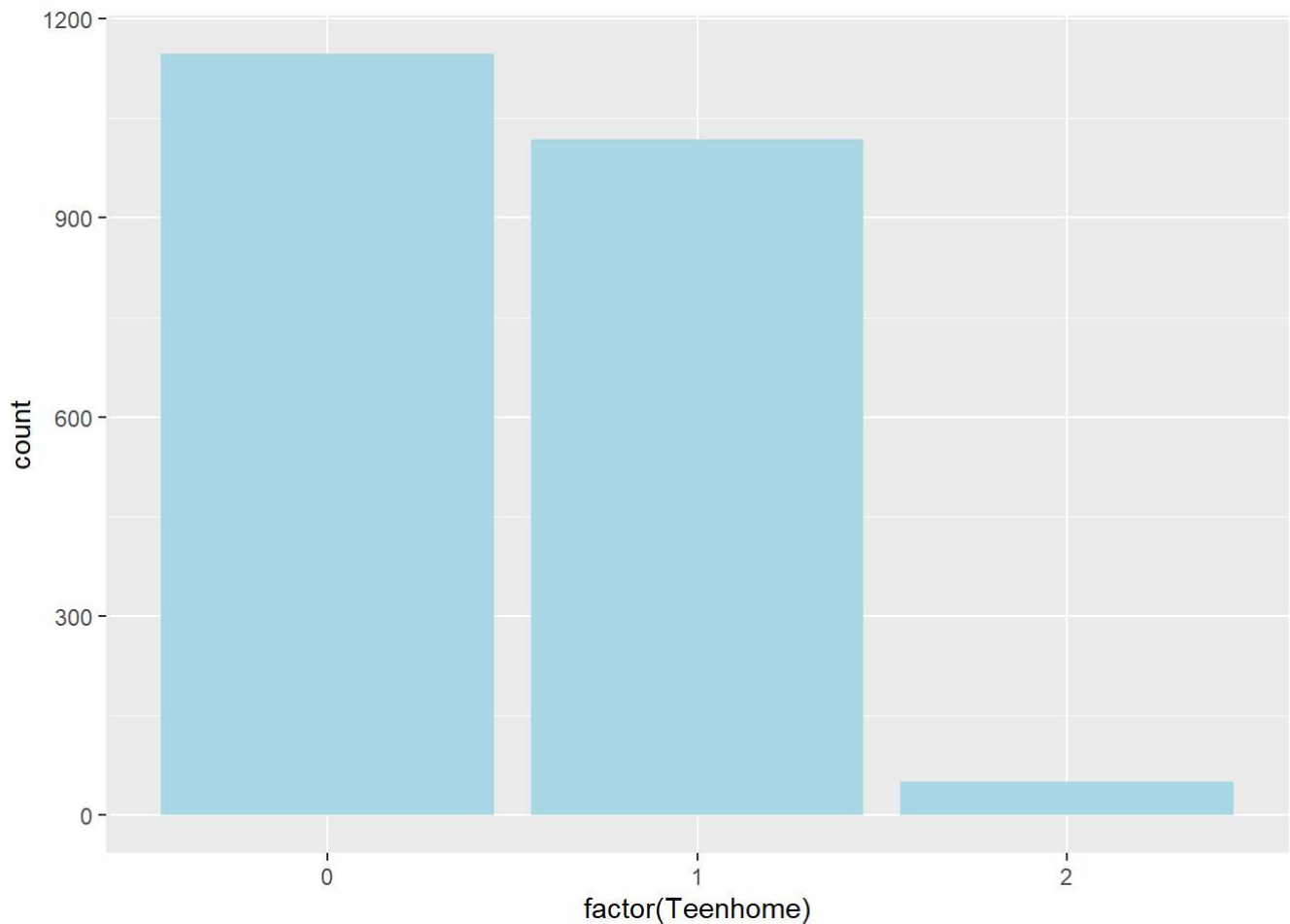
The basic analysis suggest that Recency is averaged around 49 and has a evenly shaped bell curve as skewness is zero. The data is also evenly spread as there are less outlier because kurtosis is less than 0.

# Univariate Graphical

## Categorical Variable

I will be using the number of teen homes as a categorical variable to plot a bar graph.

```
library(ggplot2)  
ggplot(data,aes(x=factor(Teenhome)))+geom_bar(fill="lightblue")
```



From the graph it is evident that majority of the customers don't have a teen kid in their household.

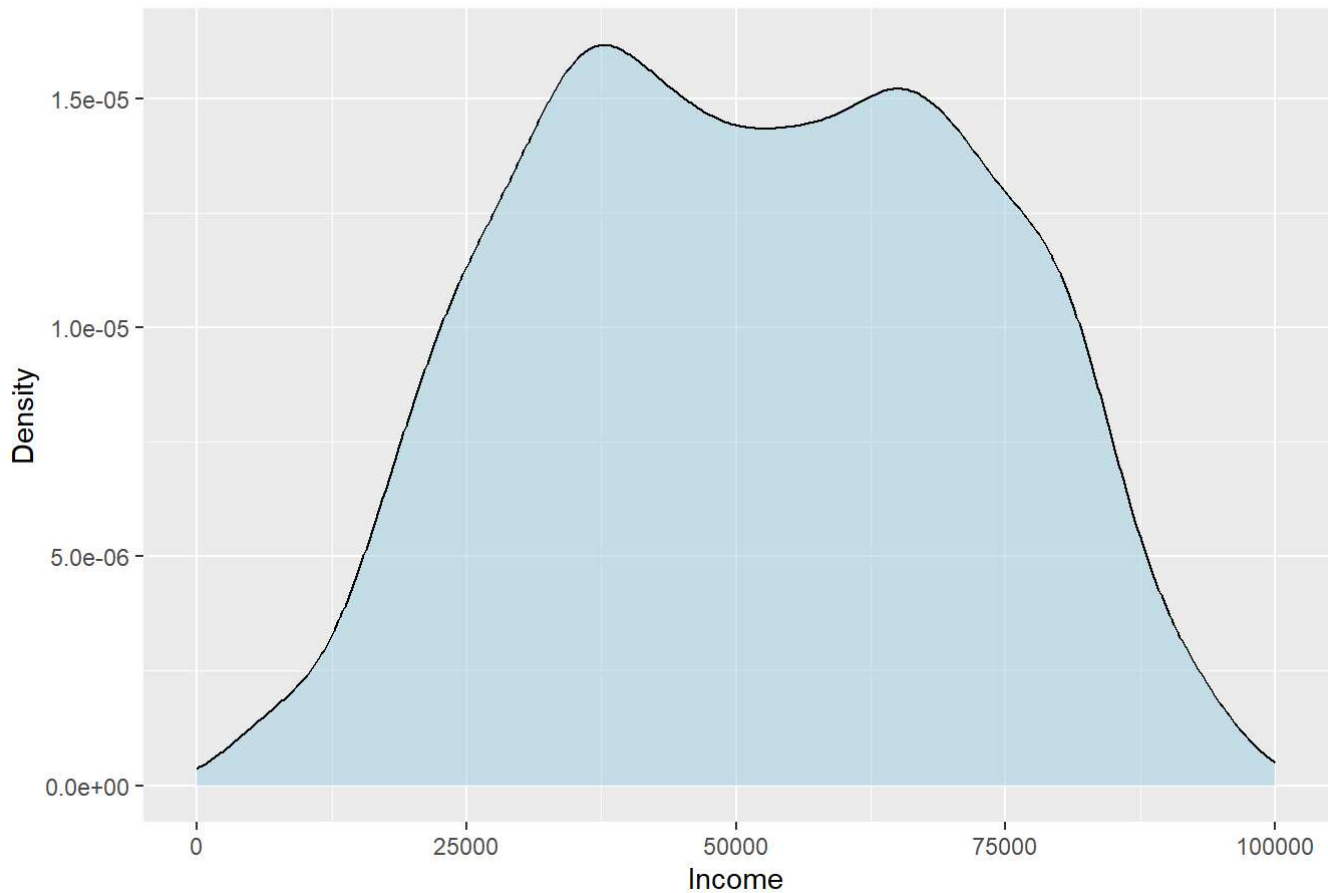
## Quantitative Variable

I will be plotting the pl which is a smoothed version of histogram on the income variable to understand the distribution of income

```
ggplot(data,aes(x=Income)) +  
  geom_density(fill="lightblue",alpha=0.6) +  
  labs(title = "Distribution of Income",x="Income",y="Density") +  
  xlim(0, 100000)
```

```
## Warning: Removed 13 rows containing non-finite outside the scale range  
## (`stat_density()`).
```

## Distribution of Income



The graph shows that the distribution of income is bimodal with an average income of 50000.

## Multivariate Non Graphical

### Categorical Variables

Education and marital status are interesting variables to plot a cross table/

```
ctable<-table(data$Education,data$Marital_Status)
ctable
```

```
##
##           Absurd Alone Divorced Married Single Together Widow YOLO
## 2n Cycle      0      0      23      80      36      56      5      0
## Basic         0      0       1      20      18      14      1      0
## Graduation    1      1     119     429     246     285     35      0
## Master        1      1      37     138      75     102     11      0
## PhD           0      1      52     190      96     116     24      2
```

Highest number of customers are from the graduate married category. The pattern shows that majority of customer chunk are from the education section of graduation and above with married and together being the highest contributors across those categories.

## Quantitative Variables

Lets get a correlation matrix for all the categorical variables.

```
num_variables=data[c('Kidhome', 'Teenhome', 'Recency', 'MntWines', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumWebVisitsMonth')]
cor_matrix <- cor(num_variables, use = "complete.obs", method = "pearson")
cor_matrix
```

```
##           Kidhome      Teenhome      Recency      MntWines
## Kidhome      1.00000000 -0.039869095  0.0114921489 -0.497335858
## Teenhome     -0.03986909  1.000000000  0.0138378832  0.003746663
## Recency       0.01149215  0.013837883  1.0000000000  0.015721019
## MntWines     -0.49733586  0.003746663  0.0157210194  1.000000000
## MntMeatProducts -0.43926053 -0.261122385  0.0225176351  0.568860003
## MntFishProducts -0.38888422 -0.205241867  0.0005509232  0.397721050
## MntSweetProducts -0.37802613 -0.163055777  0.0251097703  0.390325802
## MntGoldProds  -0.35502942 -0.019887234  0.0176626377  0.392730993
## NumDealsPurchases 0.21691305  0.386246304  0.0021154508  0.008885929
## NumWebPurchases -0.37197655  0.162077185 -0.0056408538  0.553785939
## NumWebVisitsMonth 0.44747694  0.131240022 -0.0185636434 -0.321977901
##           MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds
## Kidhome      -0.43926053  -0.3888842203  -0.37802613  -0.35502942
## Teenhome     -0.26112239  -0.2052418665  -0.16305578  -0.01988723
## Recency       0.02251764  0.0005509232  0.02510977  0.01766264
## MntWines      0.56886000  0.3977210502  0.39032580  0.39273099
## MntMeatProducts 1.00000000  0.5735740153  0.53513611  0.35944628
## MntFishProducts 0.57357402  1.0000000000  0.58386696  0.42714204
## MntSweetProducts 0.53513611  0.5838669550  1.00000000  0.35744975
## MntGoldProds  0.35944628  0.4271420401  0.35744975  1.00000000
## NumDealsPurchases -0.12130771 -0.1432410856 -0.12143193  0.05190483
## NumWebPurchases  0.30709037  0.2996875104  0.33393722  0.40706567
## NumWebVisitsMonth -0.53948442 -0.4464232918 -0.42237080 -0.24769056
##           NumDealsPurchases NumWebPurchases NumWebVisitsMonth
## Kidhome      0.216913048  -0.371976549  0.44747694
## Teenhome     0.386246304  0.162077185  0.13124002
## Recency       0.002115451  -0.005640854  -0.01856364
## MntWines      0.008885929  0.553785939  -0.32197790
## MntMeatProducts -0.121307714  0.307090366  -0.53948442
## MntFishProducts -0.143241086  0.299687510  -0.44642329
## MntSweetProducts -0.121431928  0.333937217  -0.42237080
## MntGoldProds  0.051904829  0.407065666  -0.24769056
## NumDealsPurchases 1.000000000  0.241440318  0.34604838
## NumWebPurchases  0.241440318  1.000000000  -0.05122626
## NumWebVisitsMonth 0.346048380  -0.051226263  1.00000000
```

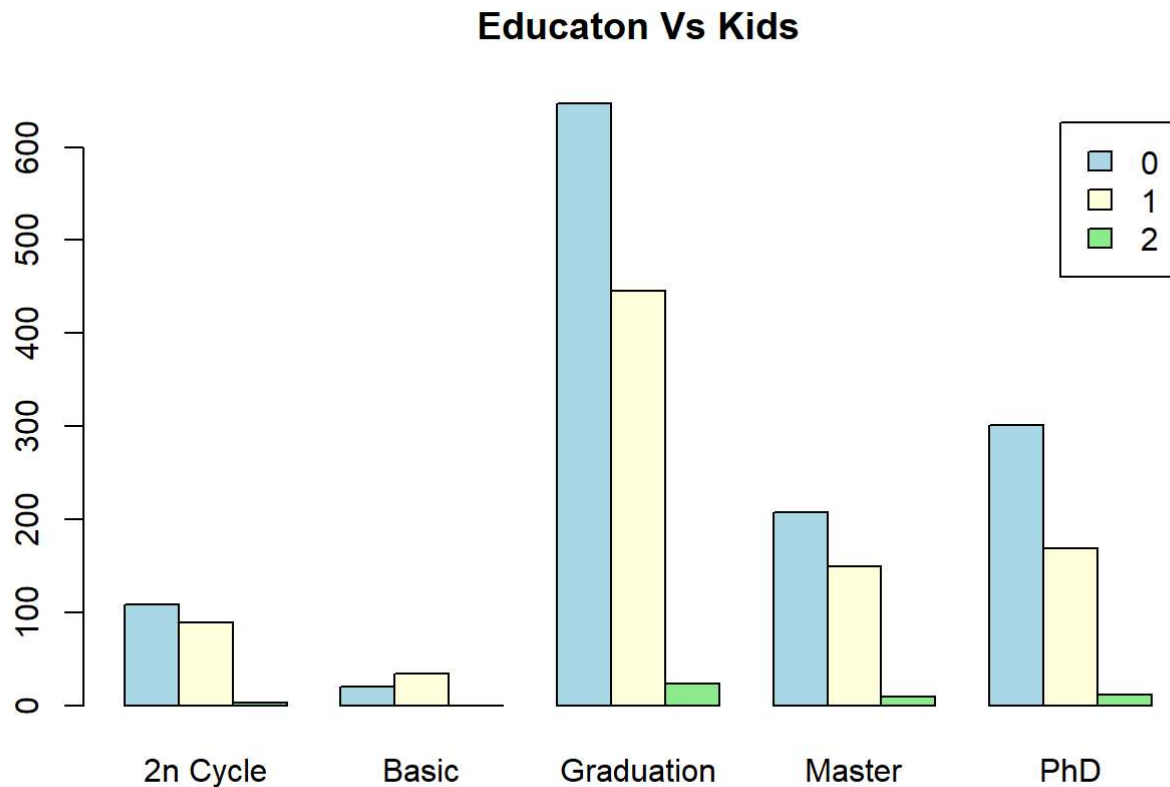
The analysis gives a glimpse of how the variables are related to each other. It is observed that the quantitative variables of Mnt purchased have the highest correlation among which MntFishProducts and MntSweetProducts have the highest correlation suggesting that customers who buy more fish tend to buy more sweets.

## Multivariate Graphical

## Categorical Variables

We examine whether the Kidhome and education has any relation/pattern

```
t1<-table(data$Kidhome,data$Education)
barplot(t1,main="Educaton Vs Kids",legend=rownames(t1),beside=TRUE,col=c("lightblue","lightyellow","lightgreen"))
```



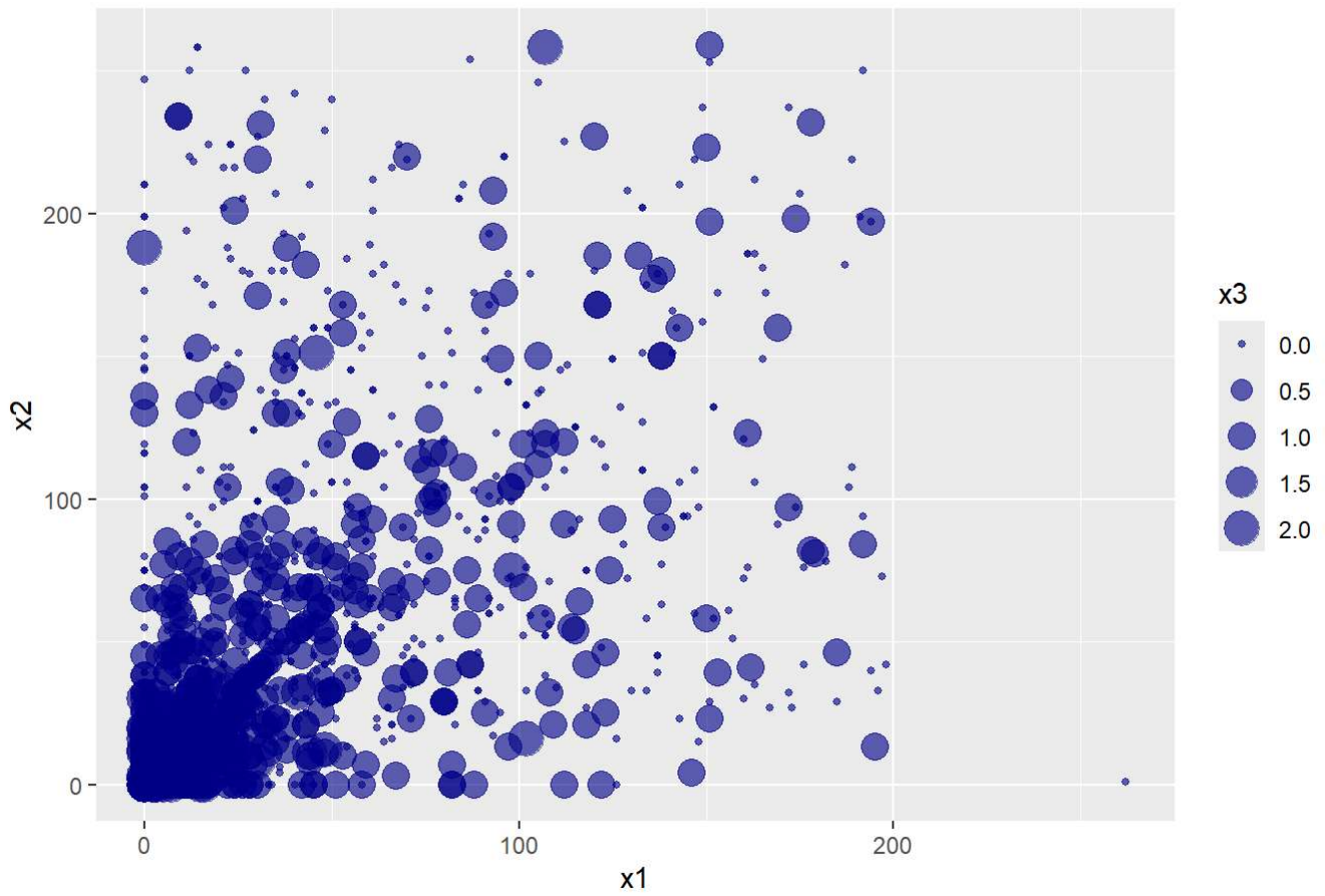
The pattern is same among all the categories except for Basic education where the customers with one kid are greater than other two categories.

## Quantitative Variables

Now as the correlation analysis indicated that the sweets and fish purchases are correlated let's examine it with respect to teens home.

```
ggplot(data, aes(x = MntSweetProducts, y = MntFishProducts, size = Teenhome)) +
  geom_point(alpha = 0.6, color = "darkblue") +
  labs(title = "Bubble Plot", x = "x1", y = "x2", size = "x3")
```

Bubble Plot



There is a positive correlation between Sweet and Fish purchases but the teens home have a random pattern not showing any relationship with the purchases.