CustomerSegmentation

2023-05-02

```
if (!require("ISLR2")) install.packages("ISLR2")
```

```
## Loading required package: ISLR2
```

```
if (!require("cluster")) install.packages("cluster")
```

```
## Loading required package: cluster
```

```
if (!require("ggdendro")) install.packages("ggdendro")
```

```
## Loading required package: ggdendro
```

```
if (!require("factoextra")) install.packages("factoextra")
```

```
## Loading required package: factoextra
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
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
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(tibble)

library(cluster)

library(tidyr)

library(factoextra)

library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

	ID	Year_Birth	Education	Marital_Status	Inco…	Kidh…	Teenh…	Dt_Customer	Rece…
	<int>	<int>	<chr>	<chr>	<int>	<int>	<int>	<chr>	<int>
1	5524	1957	Graduation	Single	58138	0	0	4/9/2012	58
2	2174	1954	Graduation	Single	46344	1	1	8/3/2014	38
3	4141	1965	Graduation	Together	71613	0	0	21-08-2013	26
4	6182	1984	Graduation	Together	26646	1	0	10/2/2014	26
5	5324	1981	PhD	Married	58293	1	0	19-01-2014	94
6	7446	1967	Master	Together	62513	0	1	9/9/2013	16

6 rows | 1-10 of 30 columns

EDA

```
sum(is.na(df))
```

```
## [1] 24
```

There are 24 NULL values in our data we will examine those as we go along

```
df[duplicated(df)]
```

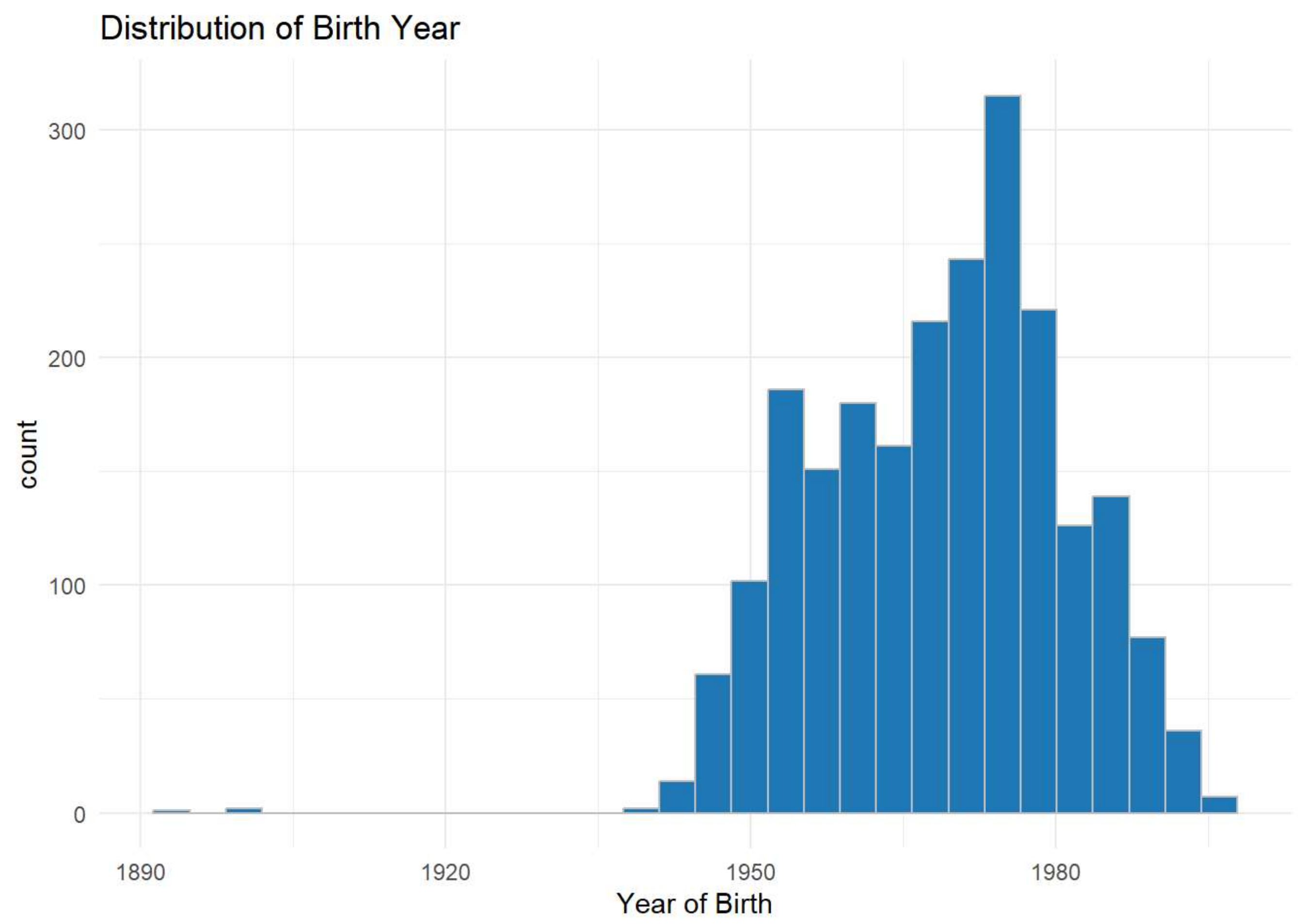
```
0 rows
```

There are no duplicate rows

```
df <- df %>%
  select(-ID)
```

Birth Year

```
ggplot(df, aes(x=Year_Birth))+
     geom_histogram(color = "grey", fill = "#1f77b4", bins = 30)+
  labs(x = "Year of Birth",
       y = "count",
       title = "Distribution of Birth Year")+
  theme_minimal()
```



```
df %>%
  filter(Year_Birth < 1930)
```

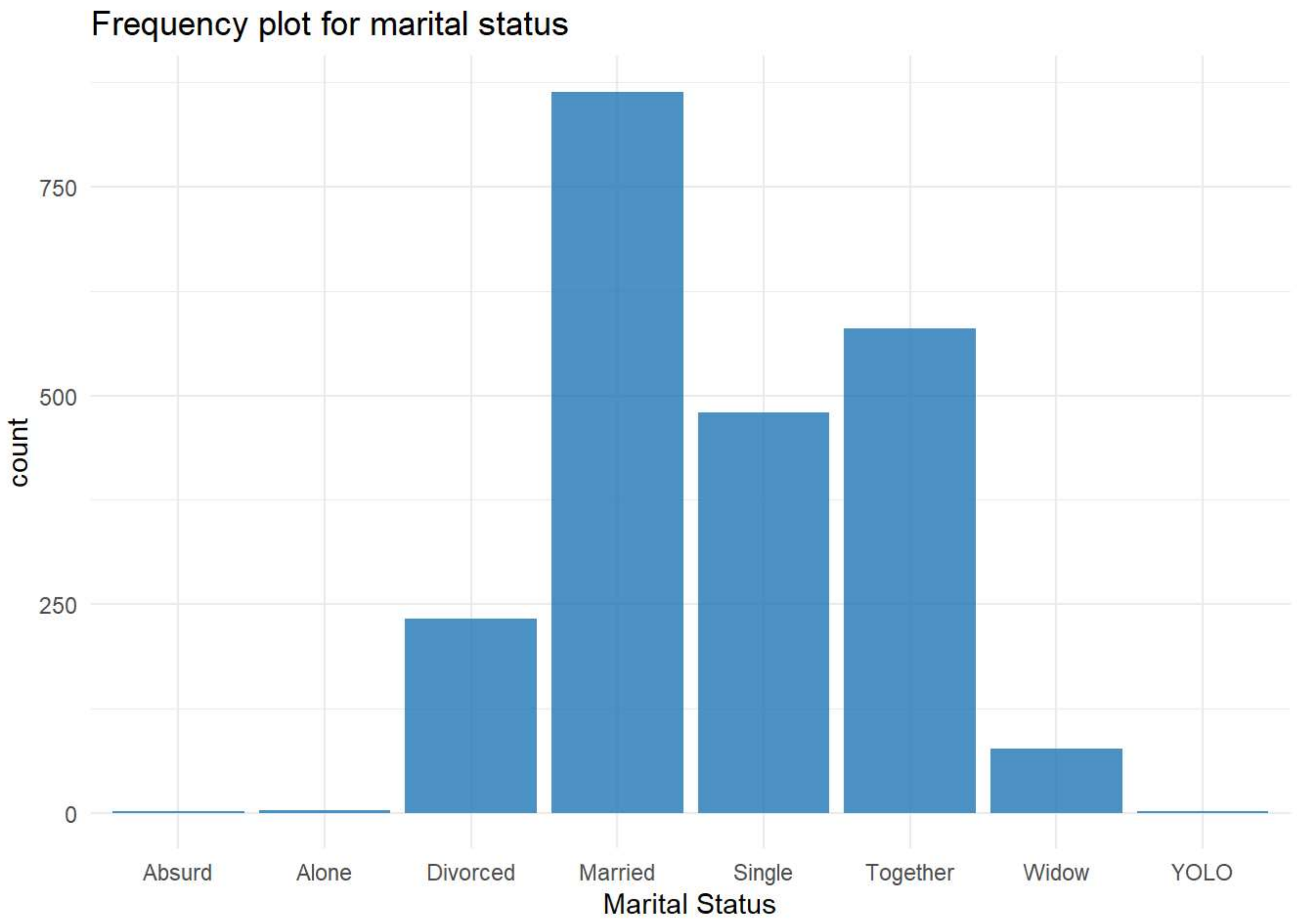
Year_Birth <int>	Education <chr>	Marital_Status <chr>	Inco… <int>	Kidh… <int>	Teenh… <int>	Dt_Customer <chr>	Rece… <int>	MntWi. <ir
1900	2n Cycle	Divorced	36640	1	0	26-09-2013	99	
1893	2n Cycle	Single	60182	0	1	17-05-2014	23	
1899	PhD	Together	83532	0	0	26-09-2013	36	7

3 rows | 1-9 of 28 columns

seems like they are erroneous entries

Marital Status

```
ggplot(df, aes(Marital_Status)) +
  geom_bar(fill = "#1f77b7", alpha = 0.8) +
  labs( x= "Marital Status",
        y = "count",
        title = "Frequency plot for marital status")+
  theme_minimal()
```

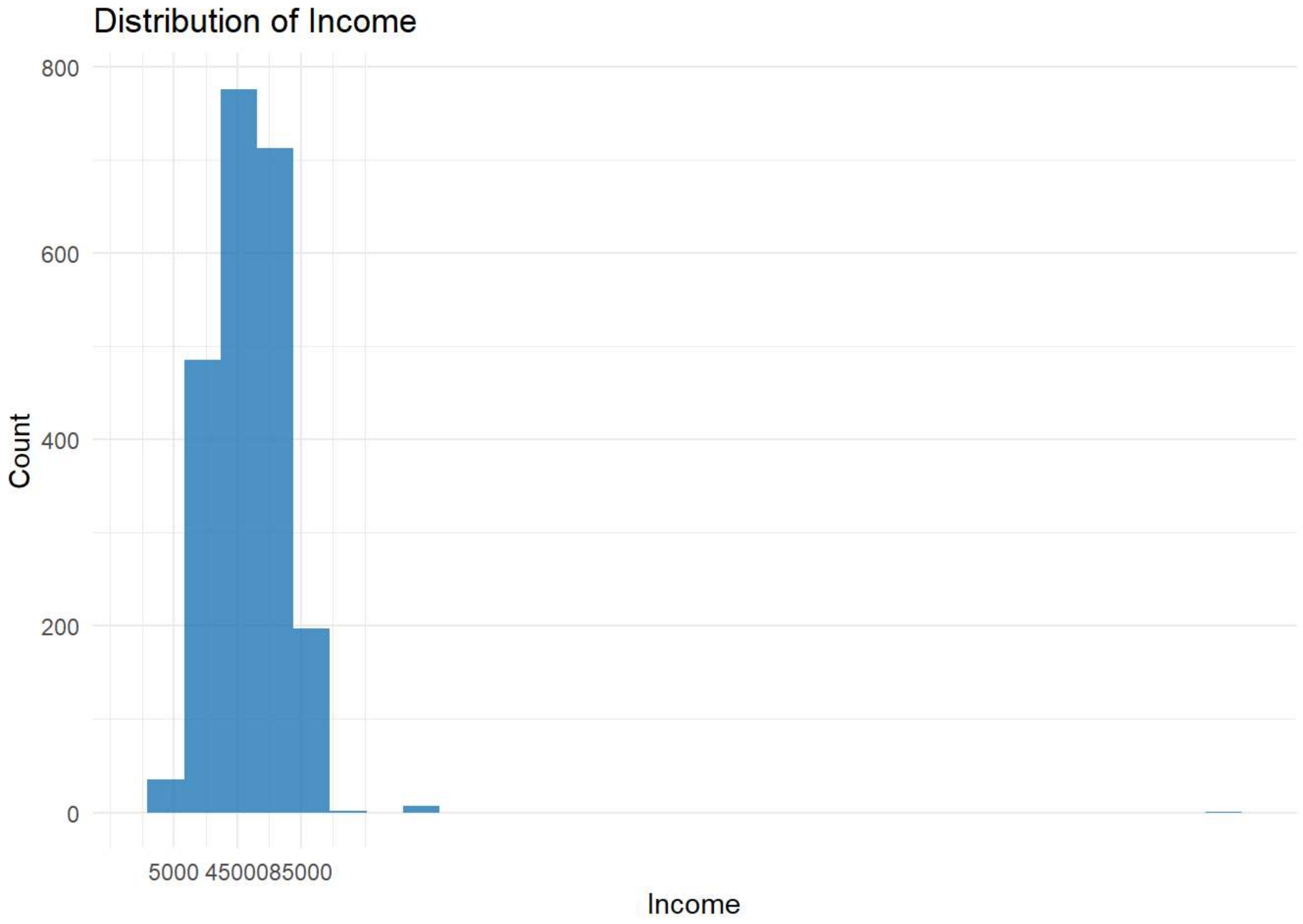


Income

```
ggplot(df, aes(x = Income)) +
  geom_histogram(fill = "#1f77b7", alpha = 0.8)+
  labs(x = "Income",
       y = "Count",
       title = "Distribution of Income")+
  scale_x_continuous(breaks = seq(5000, 100000, by=  40000))+
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

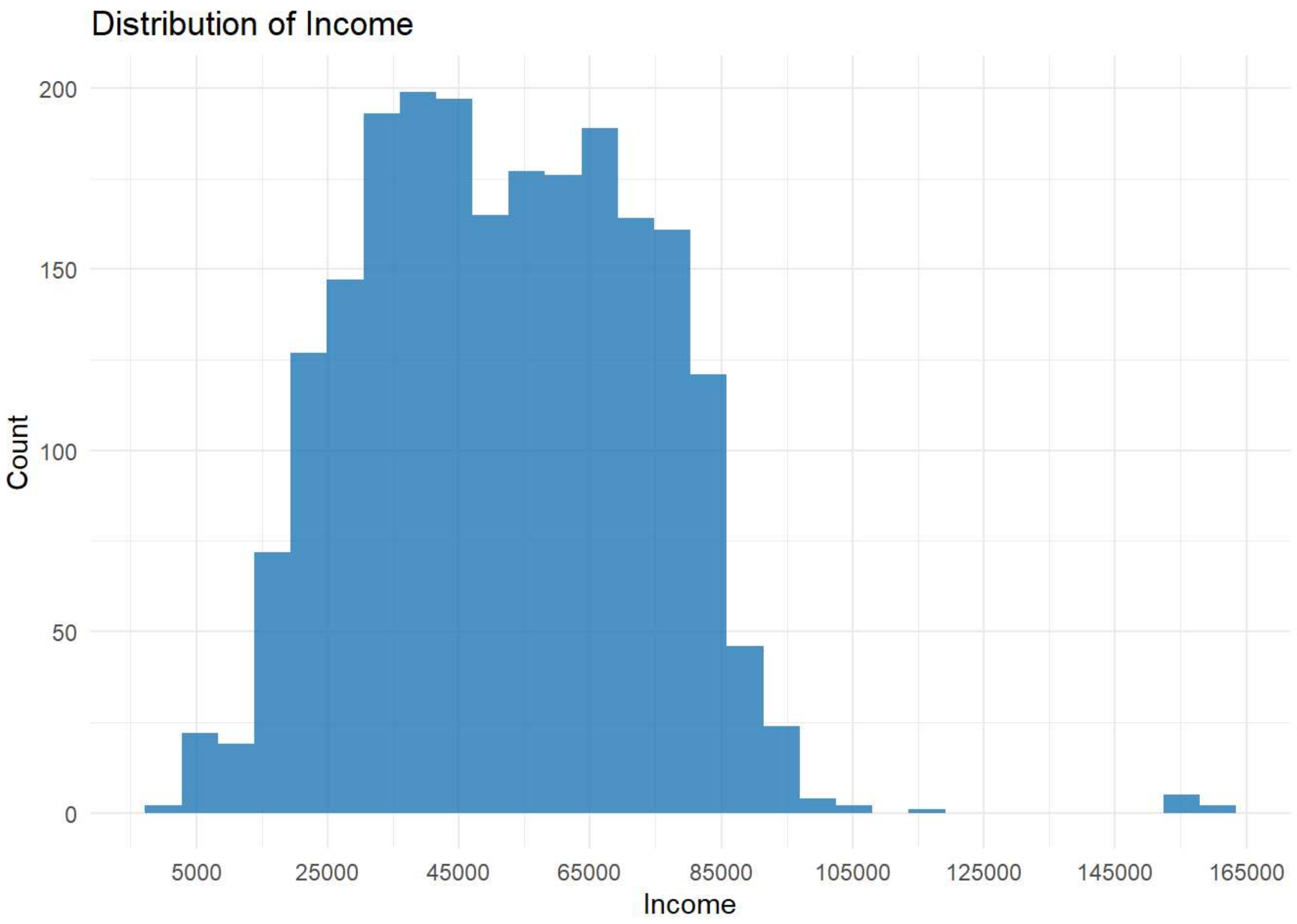
```
## Warning: Removed 24 rows containing non-finite values (`stat_bin()`).
```



There is an outlier in data where we see a very large income, to see the distribution clearly lets filter our data

```
df %>%
  filter(Income < 500000) %>%
  ggplot(aes(x = Income)) +
  geom_histogram(fill = "#1f77b7", alpha = 0.8)+
  labs(x = "Income",
       y = "Count",
       title = "Distribution of Income")+
  scale_x_continuous(breaks = seq(5000, 200000, by=  20000))+
  theme_minimal()
```

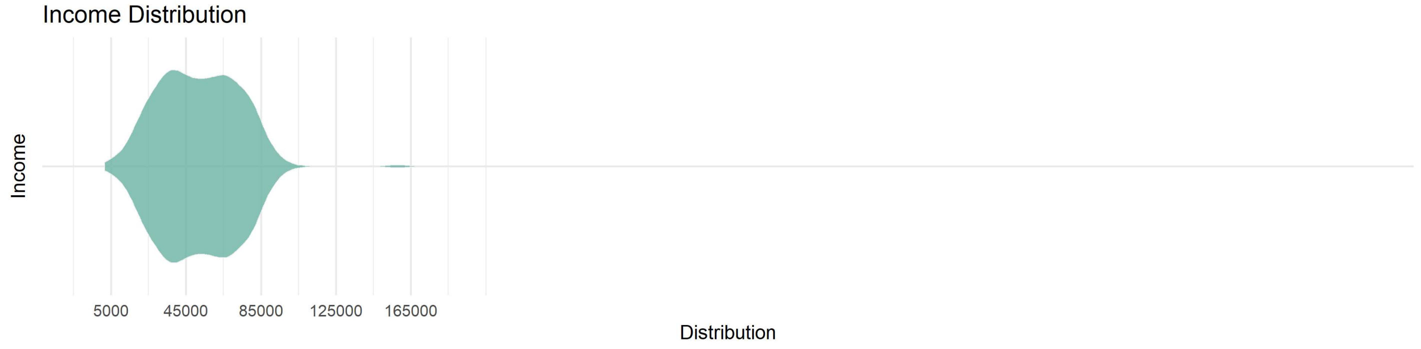
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There are few data points with income greater than 85000 lets call them high income group while rest looks to in the range 19000 - 70000

```
ggplot(df, aes(x = "", y = Income)) +
  geom_violin(fill = "#69b3a2", color = "#e9ecef", alpha = 0.8)+
  coord_flip()+
  scale_y_continuous(breaks = seq(5000, 165000, by=  40000))+
  labs(
    x = "Income",
    y = "Distribution",
    title = "Income Distribution"
  )+
  theme_minimal(base_size  = 20)
```

```
## Warning: Removed 24 rows containing non-finite values (`stat_ydensity()`).
```



Inspecting the missing data

```
df[!complete.cases(df),]
```

	Year_Birth <int>	Education <chr>	Marital_Status <chr>	Inco… <int>	Kidh… <int>	Teenh… <int>	Dt_Customer <chr>	Rece… <int>
11	1983	Graduation	Married	NA	1	0	15-11-2013	11
28	1986	Graduation	Single	NA	1	0	20-02-2013	19
44	1959	PhD	Single	NA	0	0	5/11/2013	80
49	1951	Graduation	Single	NA	2	1	1/1/2014	96
59	1982	Graduation	Single	NA	1	0	17-06-2013	57
72	1973	2n Cycle	Married	NA	1	0	14-09-2012	25
91	1957	PhD	Married	NA	2	1	19-11-2012	4
92	1957	Graduation	Single	NA	1	1	27-05-2014	45
93	1973	Master	Together	NA	0	0	23-11-2013	87
129	1961	PhD	Married	NA	0	1	11/7/2013	23

1-10 of 24 rows | 1-10 of 29 columns

Previous **1** 2 3 Next

```
summary(df)
```

```
##    Year_Birth    Education         Marital_Status         Income
##  Min.   :1893   Length:2240        Length:2240        Min.   :  1730
##  1st Qu.:1959   Class :character   Class :character   1st Qu.: 35303
##  Median :1970   Mode  :character   Mode  :character   Median : 51382
##  Mean   :1969                                         Mean   : 52247
##  3rd Qu.:1977                                         3rd Qu.: 68522
##  Max.   :1996                                         Max.   :666666
##                                                       NA's   :24
##     Kidhome          Teenhome      Dt_Customer           Recency
##  Min.   :0.0000   Min.   :0.0000   Length:2240        Min.   : 0.00
##  1st Qu.:0.0000   1st Qu.:0.0000   Class :character   1st Qu.:24.00
##  Median :0.0000   Median :0.0000   Mode  :character   Median :49.00
##  Mean   :0.4442   Mean   :0.5062                      Mean   :49.11
##  3rd Qu.:1.0000   3rd Qu.:1.0000                      3rd Qu.:74.00
##  Max.   :2.0000   Max.   :2.0000                      Max.   :99.00
##
##     MntWines         MntFruits     MntMeatProducts  MntFishProducts
##  Min.   :   0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.00
##  1st Qu.:  23.75   1st Qu.:  1.0   1st Qu.:  16.0   1st Qu.:  3.00
##  Median : 173.50   Median :  8.0   Median :  67.0   Median : 12.00
##  Mean   : 303.94   Mean   : 26.3   Mean   : 166.9   Mean   : 37.53
##  3rd Qu.: 504.25   3rd Qu.: 33.0   3rd Qu.: 232.0   3rd Qu.: 50.00
##  Max.   :1493.00   Max.   :199.0   Max.   :1725.0   Max.   :259.00
##
##  MntSweetProducts  MntGoldProds    NumDealsPurchases NumWebPurchases
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.000    Min.   : 0.000
##  1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000    1st Qu.: 2.000
##  Median :  8.00   Median : 24.00   Median : 2.000    Median : 4.000
##  Mean   : 27.06   Mean   : 44.02   Mean   : 2.325    Mean   : 4.085
##  3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000    3rd Qu.: 6.000
##  Max.   :263.00   Max.   :362.00   Max.   :15.000    Max.   :27.000
##
##  NumCatalogPurchases NumStorePurchases NumWebVisitsMonth  AcceptedCmp3
##  Min.   : 0.000      Min.   : 0.00     Min.   : 0.000    Min.   :0.00000
##  1st Qu.: 0.000      1st Qu.: 3.00     1st Qu.: 3.000    1st Qu.:0.00000
##  Median : 2.000      Median : 5.00     Median : 6.000    Median :0.00000
##  Mean   : 2.662      Mean   : 5.79     Mean   : 5.317    Mean   :0.07277
##  3rd Qu.: 4.000      3rd Qu.: 8.00     3rd Qu.: 7.000    3rd Qu.:0.00000
##  Max.   :28.000      Max.   :13.00     Max.   :20.000    Max.   :1.00000
##
##   AcceptedCmp4      AcceptedCmp5      AcceptedCmp1      AcceptedCmp2
##  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   :0.07455   Mean   :0.07277   Mean   :0.06429   Mean   :0.01339
##  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
##
##     Complain        Z_CostContact   Z_Revenue     Response
##  Min.   :0.000000   Min.   :3     Min.   :11   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:3     1st Qu.:11   1st Qu.:0.0000
##  Median :0.000000   Median :3     Median :11   Median :0.0000
##  Mean   :0.009375   Mean   :3     Mean   :11   Mean   :0.1491
##  3rd Qu.:0.000000   3rd Qu.:3     3rd Qu.:11   3rd Qu.:0.0000
##  Max.   :1.000000   Max.   :3     Max.   :11   Max.   :1.0000
##
```

The missing values seems to have occurred at random as there are 24 missing values which is 1% of the total data, we can omit those values.

```
df <- na.omit(df)
sum(is.na(df))
```

```
## [1] 0
```

Formatting Date column

```
df %>%
  select(Dt_Customer)
```

	Dt_Customer <chr>
1	4/9/2012
2	8/3/2014
3	21-08-2013
4	10/2/2014
5	19-01-2014
6	9/9/2013
7	13-11-2012
8	8/5/2013
9	6/6/2013
10	13-03-2014

1-10 of 2,216 rows   Previous **1** 2 3 4 5 6 ... 222 Next

```
df<- df %>%
  mutate(Dt_Customer = gsub("/", "-", Dt_Customer))
```

```
df<- df %>%
  mutate(Dt_Customer = as.Date(Dt_Customer, format("%d-%m-%Y")))
```

```
summary(df$Dt_Customer)
```

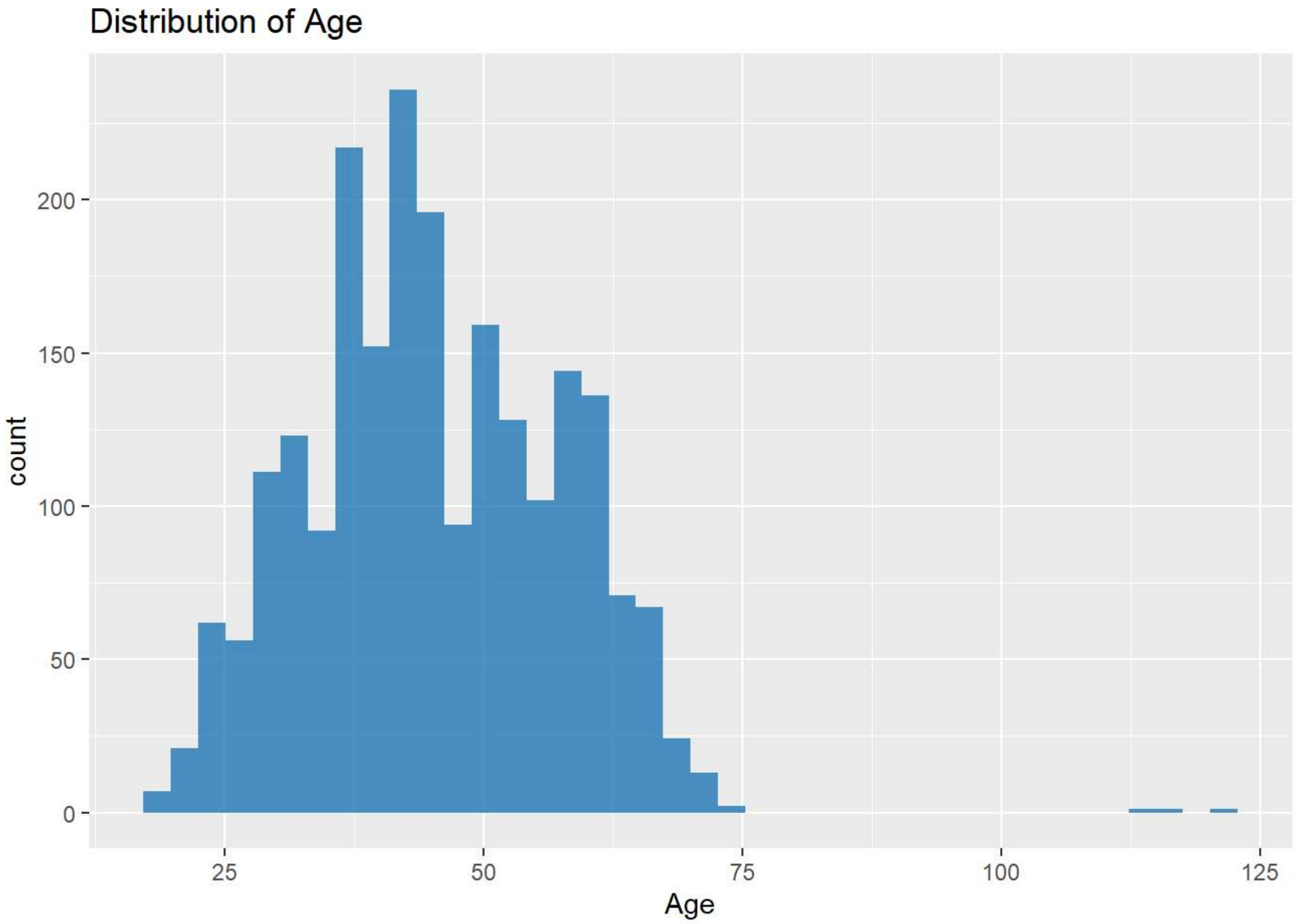
```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2012-07-30" "2013-01-16" "2013-07-08" "2013-07-10" "2013-12-31" "2014-06-29"
```

Calculating Ages by taking the maximum Date

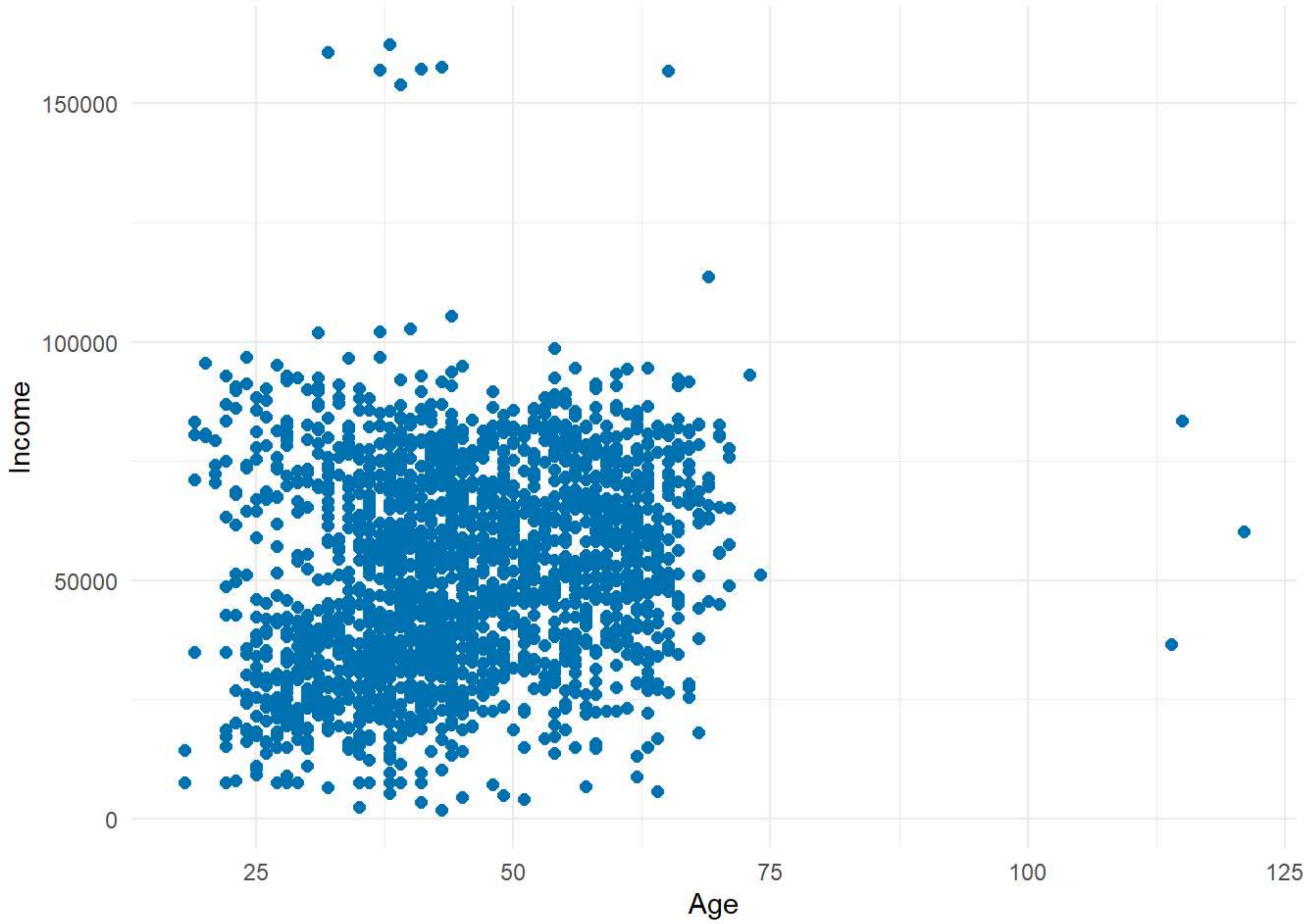
Age

```
df <- df %>%
  mutate(Age = 2014 - Year_Birth)
```

```
ggplot(df, aes(x = Age)) +
  geom_histogram(fill = "#1f77b7", bins = 40, alpha = 0.8) +
  labs(x = "Age",
       y = "count",
       title = "Distribution of Age")
```



```
df %>%
  filter(Income != 666666) %>%
ggplot(aes(x = Age, y = Income) )+
geom_point(color = "#0072B2", size = 2)+
theme_minimal()
```



There is no any evident pattern

Removing Ouliter from data for Income and capping max age to 70

```
df<- df %>%
  filter(Income != 666666) %>%
  mutate(Age = ifelse(Age > 70, 70, Age))
```

Checking Correlation between amount of product bought

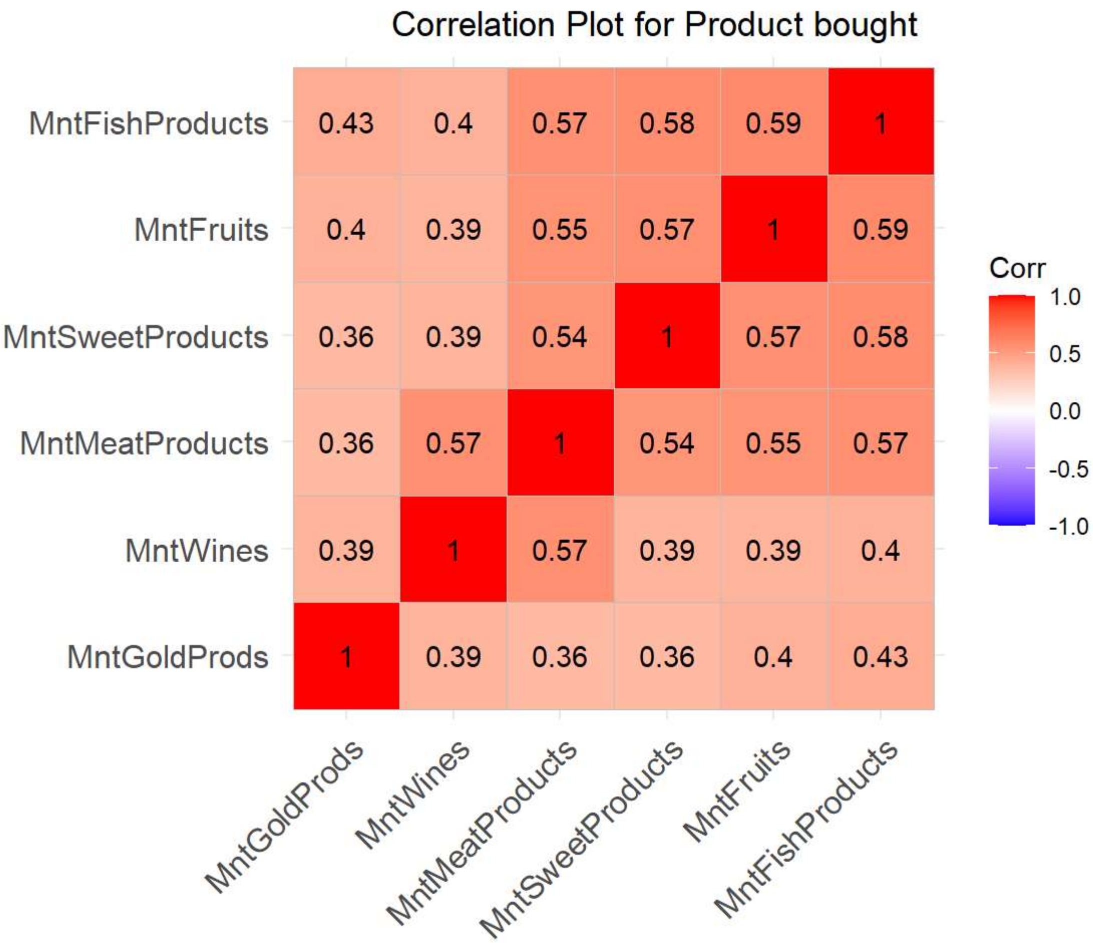
```
df_product <- df[,c("MntWines","MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProdu
cts", "MntGoldProds")]
```

```
library(ggcorrplot)
```

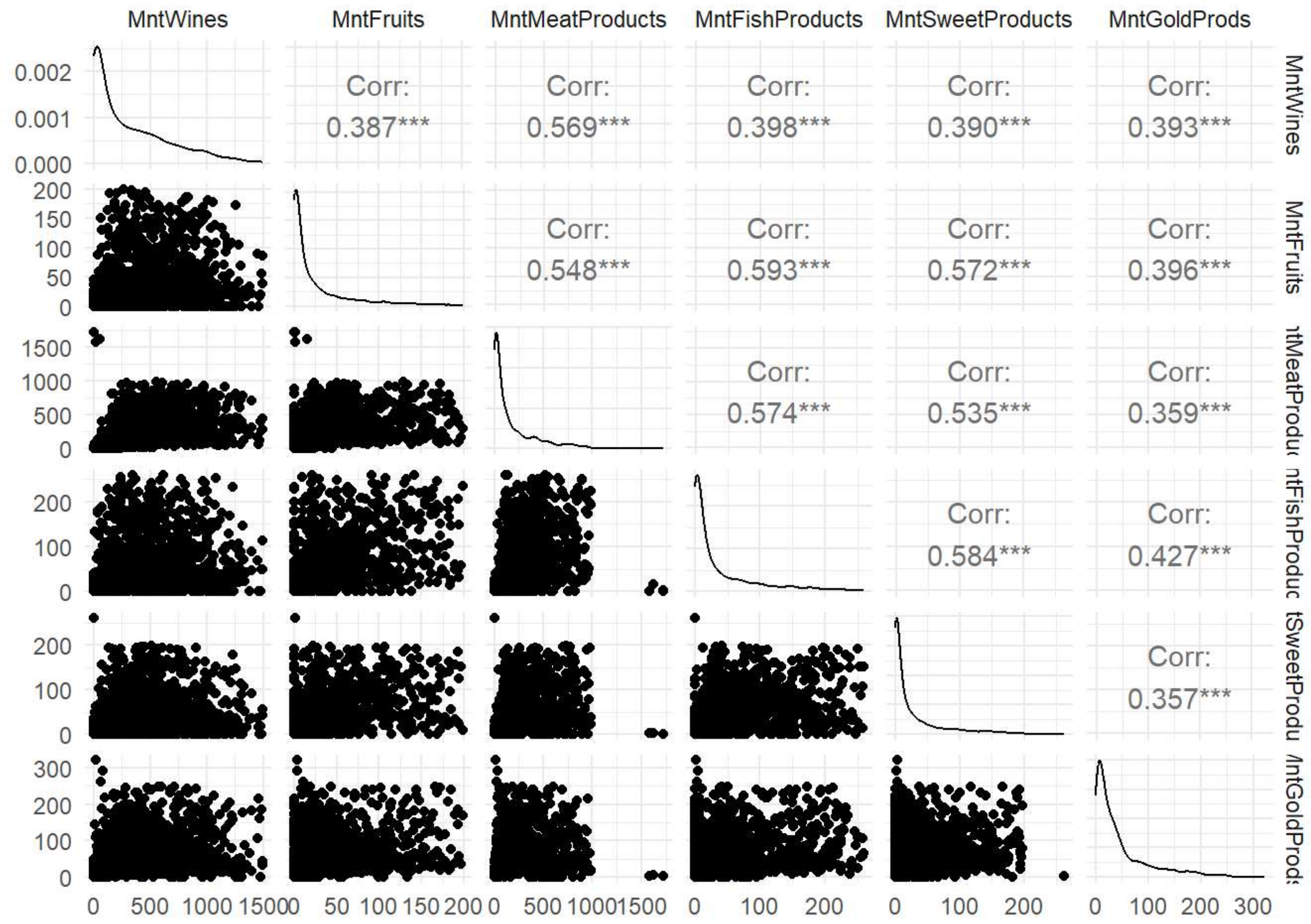
```
## Warning: package 'ggcorrplot' was built under R version 4.2.3
```

```
corr_mat_products <- cor(df_product)

ggcorrplot(corr_mat_products, hc.order = TRUE) +
  theme(plot.title = element_text(hjust = 0.8)) +
  geom_text(aes(label = value)) +
  ggtitle("Correlation Plot for Product bought")
```



```
ggpairs(df_product) +
  theme_minimal()
```

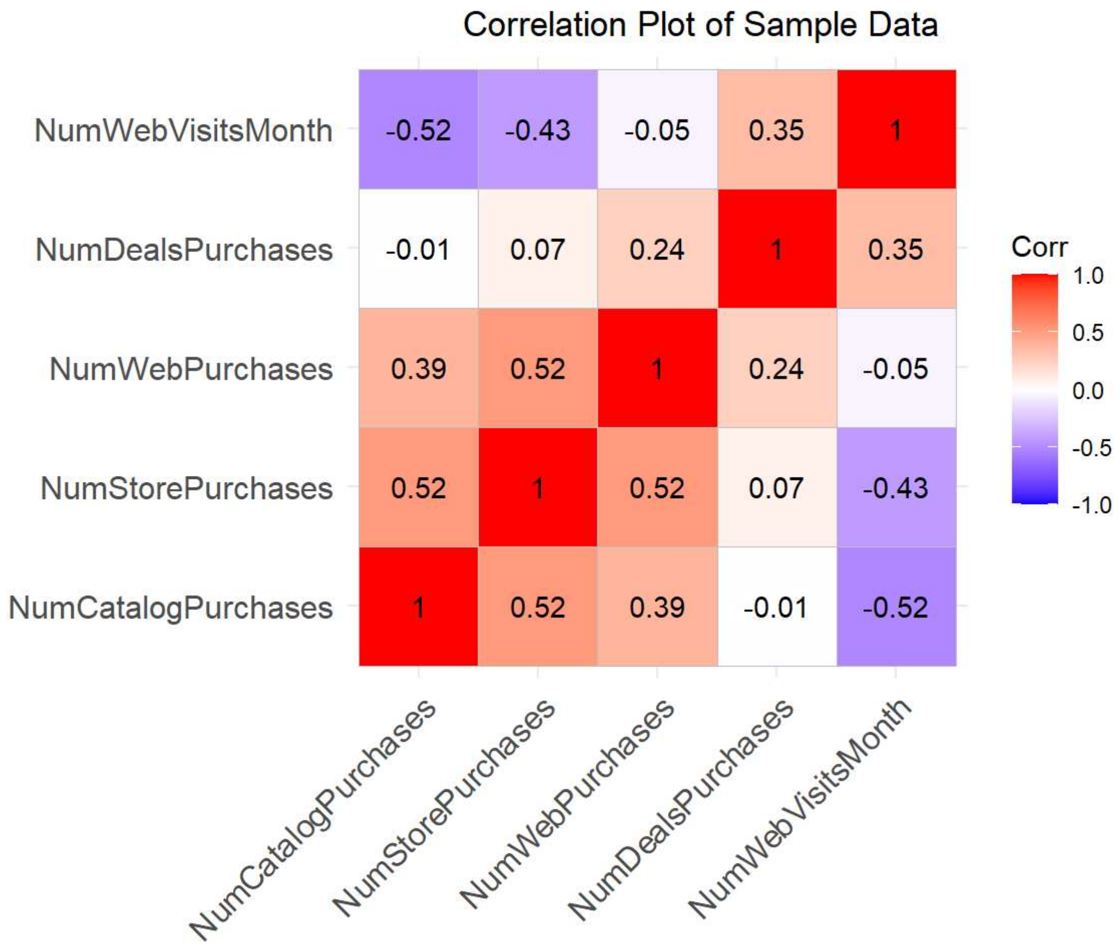


No significant relation Present between products

```
df_gateway <- df[,c("NumDealsPurchases", "NumStorePurchases", "NumWebPurchases", "NumCatalogPurchases", "NumWebVisitsMonth")]
```

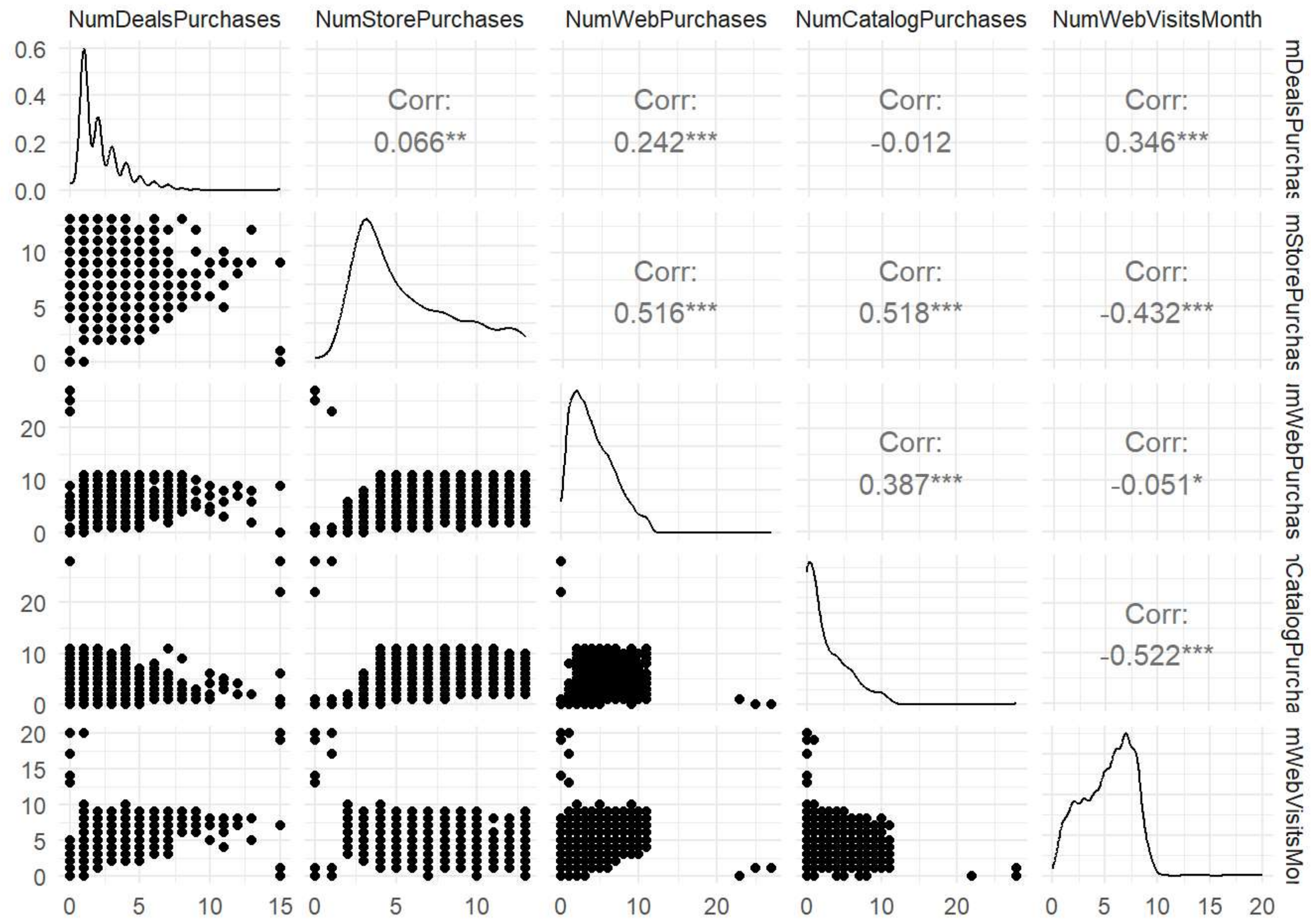
```
corr_mat_gtwy <- cor(df_gateway)

ggcorrplot(corr_mat_gtwy, hc.order = TRUE) +
  theme(plot.title = element_text(hjust = 0.8)) +
  geom_text(aes(label = value)) +
  ggtitle("Correlation Plot of Sample Data")
```



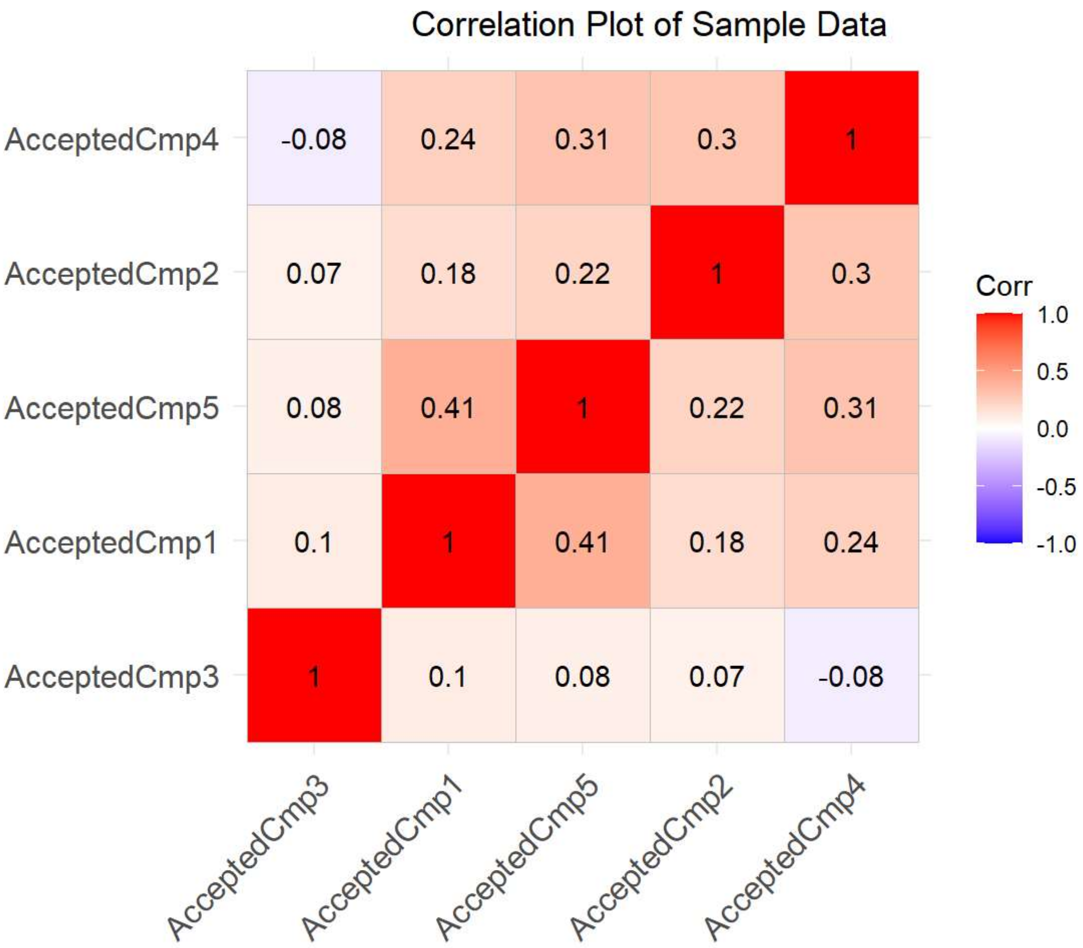
```
df_campaign <- df[,c("AcceptedCmp1", "AcceptedCmp2", "AcceptedCmp3", "AcceptedCmp4", "AcceptedCm
p5")]
corr_mat_campaign <- cor(df_campaign)
```

```
ggpairs(df_gateway) +
  theme_minimal()
```



No Significant Relation present

```
ggcorrplot(corr_mat_campaign, hc.order = TRUE) +
  theme(plot.title = element_text(hjust = 0.8)) +
  geom_text(aes(label = value)) +
  ggtitle("Correlation Plot of Sample Data")
```

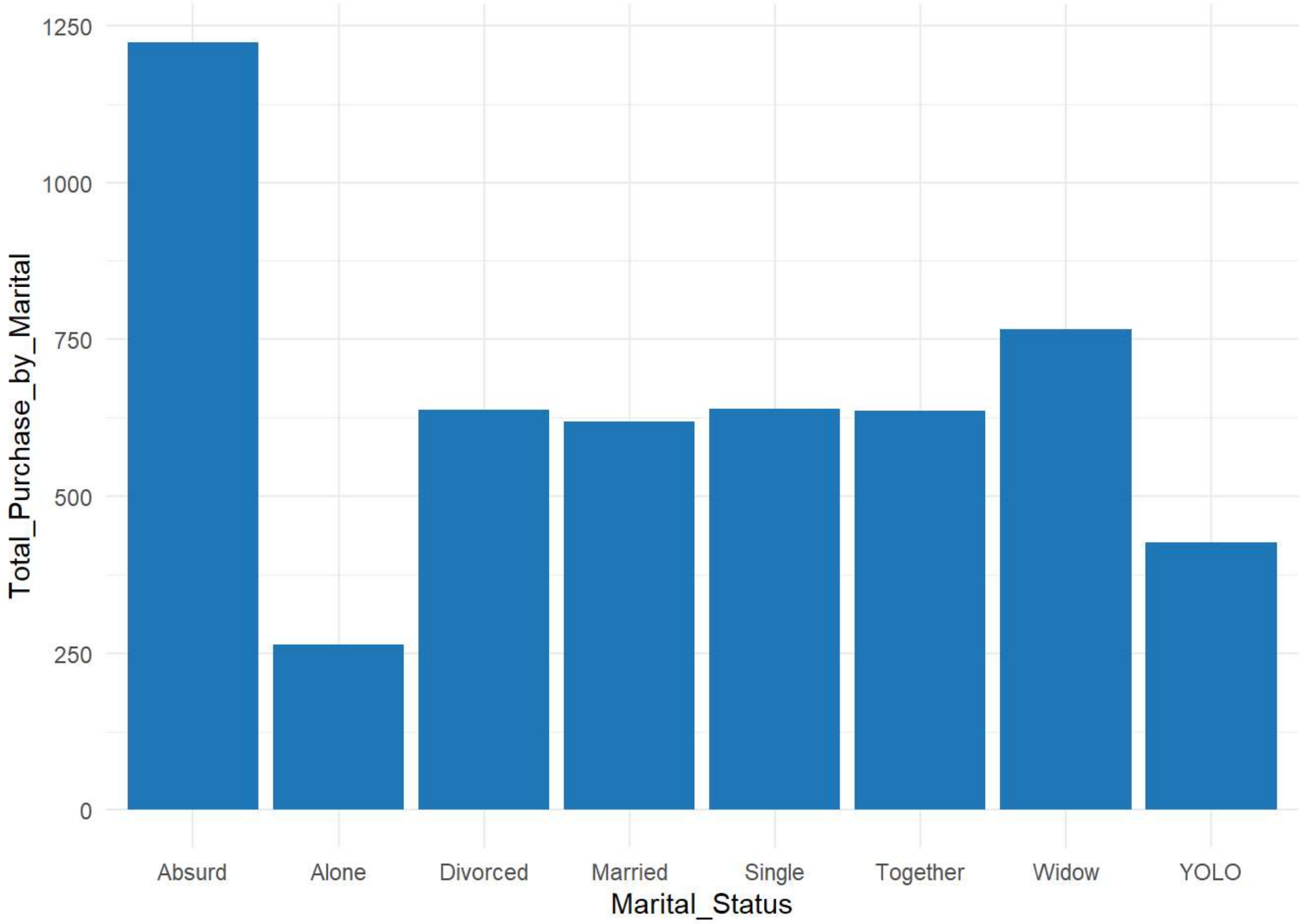


Now we examine relation across various columns

Creating variable Total Purchase which has all product purchased

```
df <- df %>%
  mutate(Total_Purchaase = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetPr
oducts + MntSweetProducts + MntGoldProds)
```

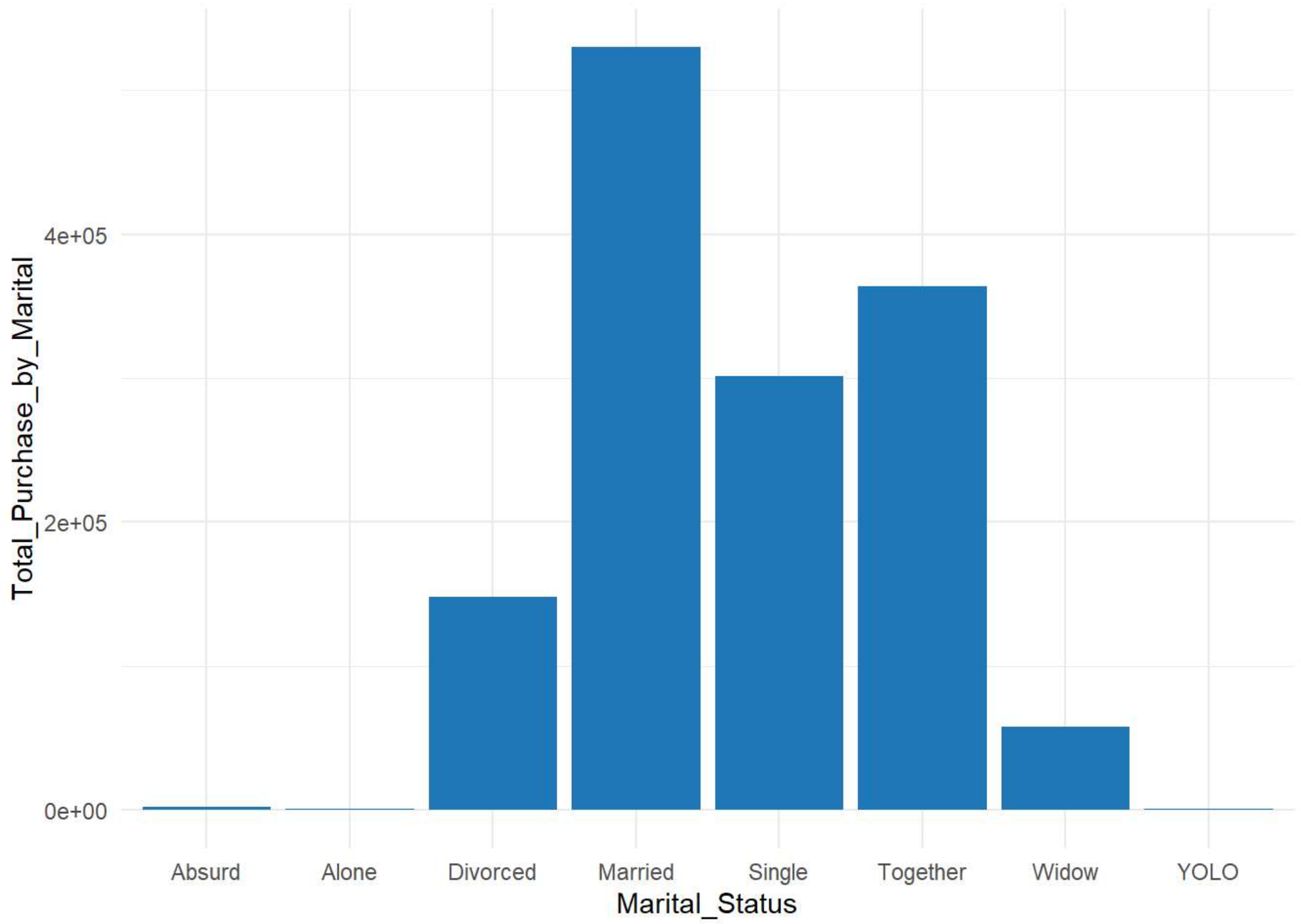
```
df %>%
  group_by(Marital_Status) %>%
  summarise(Total_Purchase_by_Marital = mean(Total_Purchaase)) %>%
  ggplot(aes(x = Marital_Status, y = Total_Purchase_by_Marital)) +
  geom_col(fill = "#1f77b7") +
  theme_minimal()
```



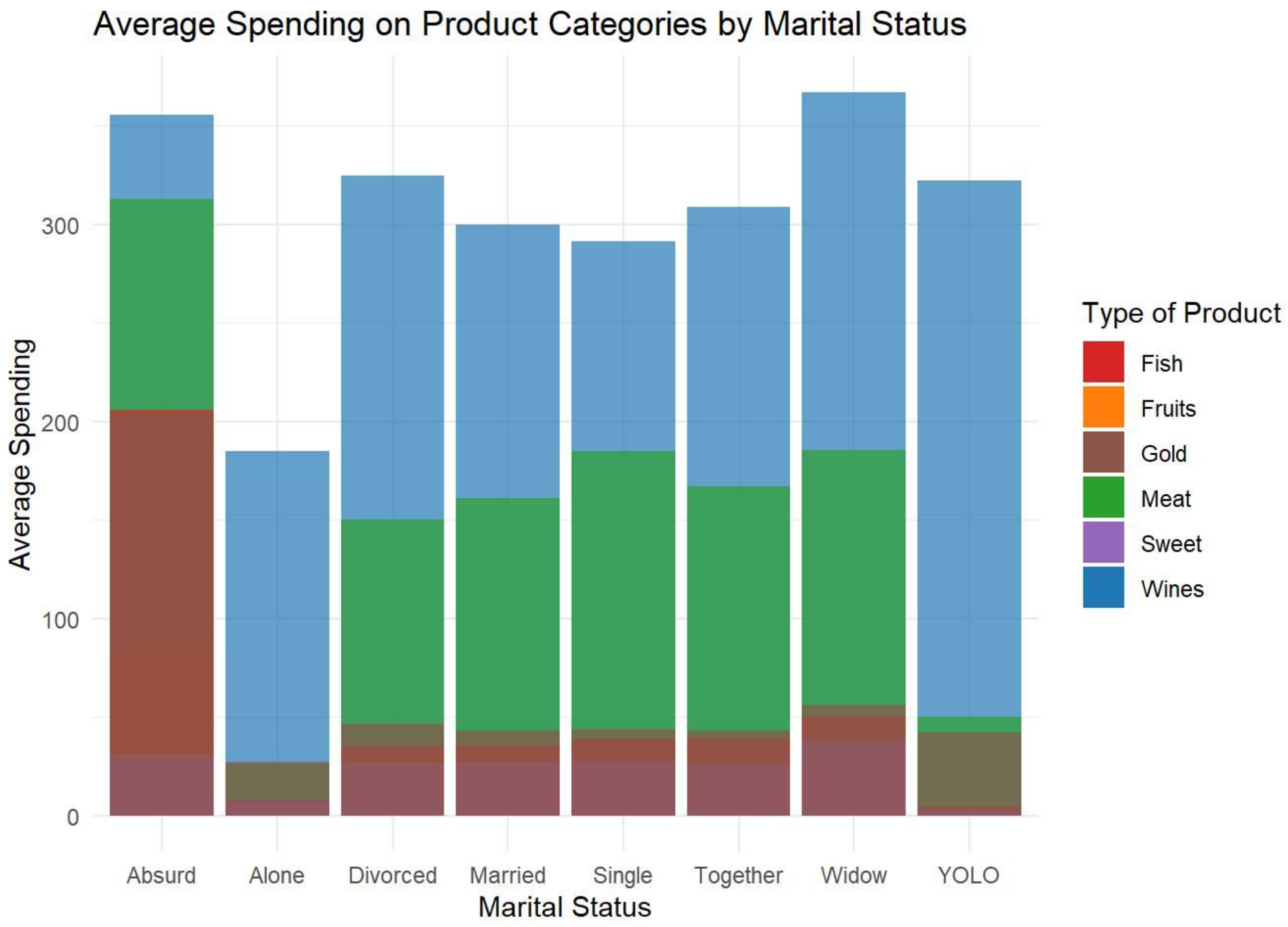
We see a graph equivalent to the proportion of the population so there is no particular group purchasing more.

Now we see across each product

```
df %>%
  group_by(Marital_Status) %>%
  summarise(Total_Purchase_by_Marital = sum(Total_Purchaase)) %>%
  ggplot(aes(x = Marital_Status, y = Total_Purchase_by_Marital)) +
  geom_col(fill = "#1f77b7") +
  theme_minimal()
```

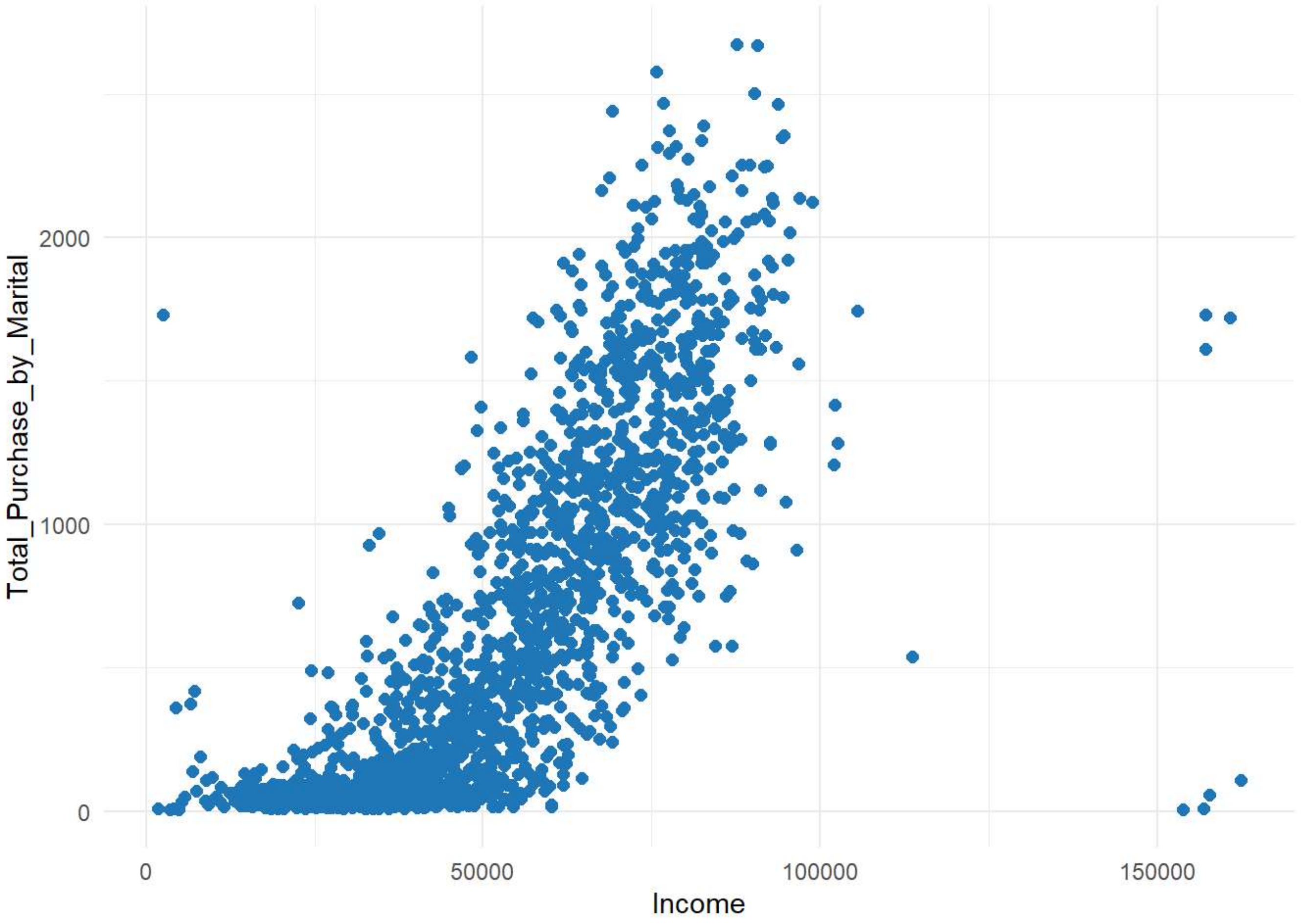


```
df %>%
  group_by(Marital_Status) %>%
  summarise(Wines = mean(MntWines), Fruits = mean(MntFruits), Meat = mean(MntMeatProducts), Fish
= mean(MntFishProducts), Sweet = mean(MntSweetProducts), gold = mean(MntGoldProds)) %>%
  ggplot(aes(x = Marital_Status)) +
  geom_bar(aes(y = Wines, fill = "Wines"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Fruits, fill = "Fruits"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Meat, fill = "Meat"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Fish, fill = "Fish"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Sweet, fill = "Sweet"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = gold, fill = "Gold"), stat = "identity", alpha = 0.7) +
  scale_fill_manual(values = c("Wines" = "#1F77B4", "Fruits" = "#FF7F0E", "Meat" = "#2CA02C", "F
ish" = "#D62728", "Sweet" = "#9467BD", "Gold" = "#8C564B")) +
  labs(title = "Average Spending on Product Categories by Marital Status",
       x = "Marital Status",
       y = "Average Spending",
       fill = "Type of Product")+
  theme_minimal() +
  theme(legend.position = "right")
```



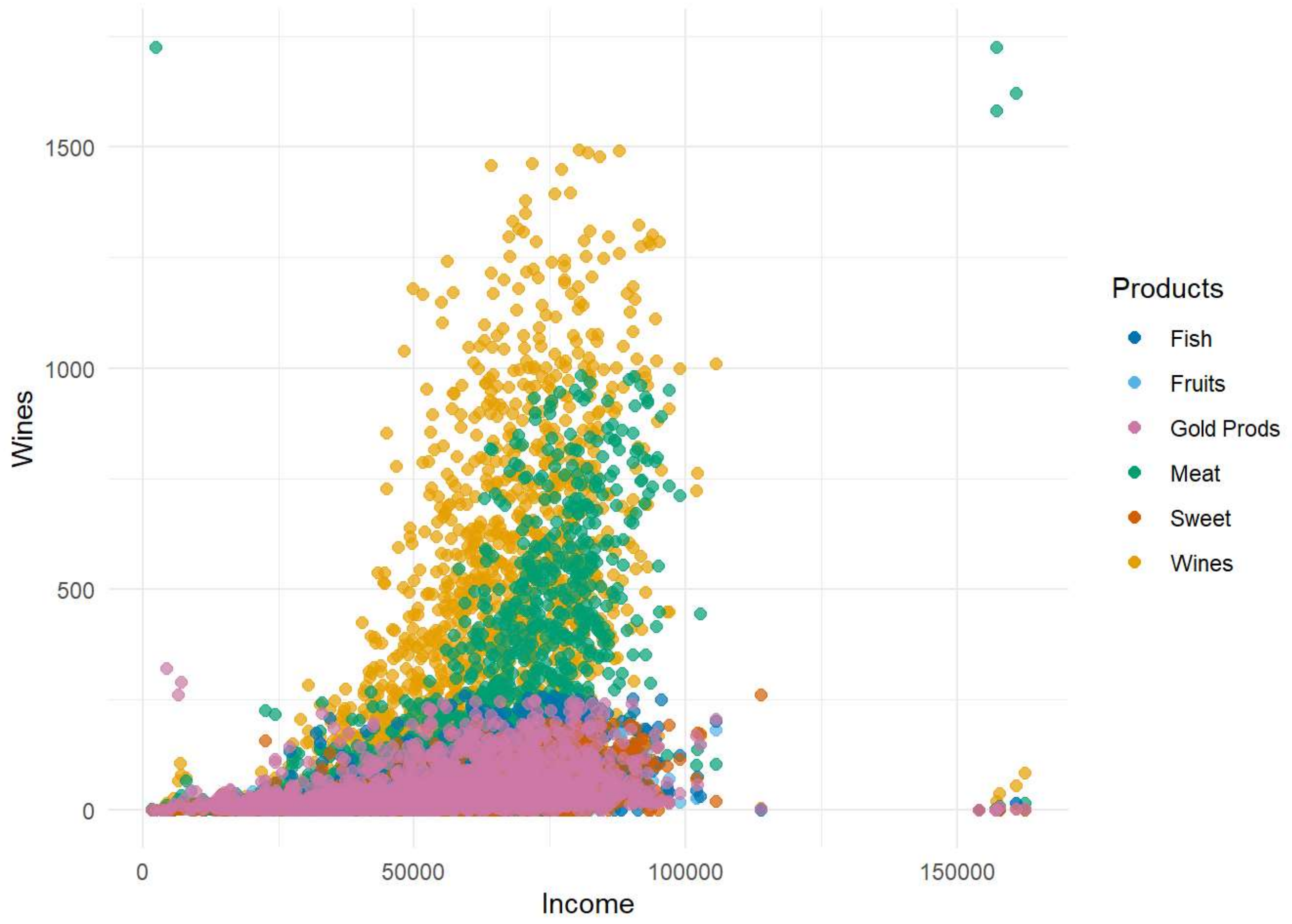
Wine is most common entity bought

```
df %>%
  group_by(Income) %>%
  summarise(Total_Purchase_by_Marital = mean(Total_Purchaase)) %>%
  ggplot(aes(x = Income, y = Total_Purchase_by_Marital)) +
  geom_point(color = "#1f77b7", size = 2) +
  theme_minimal()
```



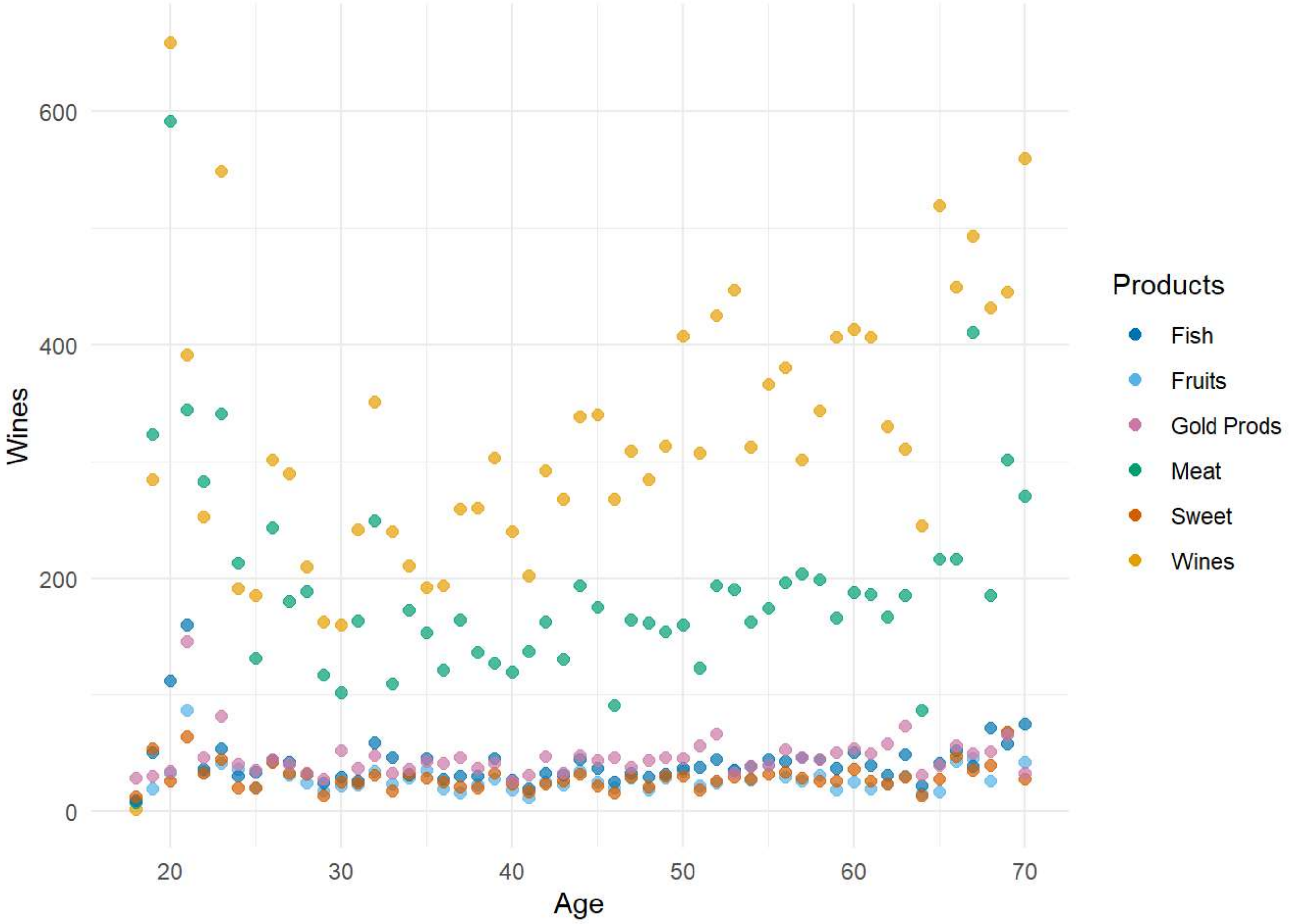
We see a non-linear relationship between Income and Total Purchase

```
df %>%
  group_by(Income) %>%
  summarise(Wines = mean(MntWines), Fruits = mean(MntFruits), Meat = mean(MntMeatProducts), Fish
= mean(MntFishProducts), Sweet = mean(MntSweetProducts), gold = mean(MntGoldProds)) %>%
  ggplot(aes(x = Income)) +
  geom_point(aes(y = Wines, color = "Wines"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Fruits, color = "Fruits"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Meat, color = "Meat"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Fish, color = "Fish"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Sweet, color = "Sweet"), alpha = 0.7, size = 2) +
  geom_point(aes(y = gold, color = "Gold Prods"), alpha = 0.7, size = 2) +
  scale_color_manual(name = "Products", values = c("Wines" = "#E69F00", "Fruits" = "#56B4E9", "M
eat" = "#009E73", "Fish" = "#0072B2", "Sweet" = "#D55E00", "Gold Prods" = "#CC79A7")) +
  theme_minimal() +
  theme(legend.position = "right")
```



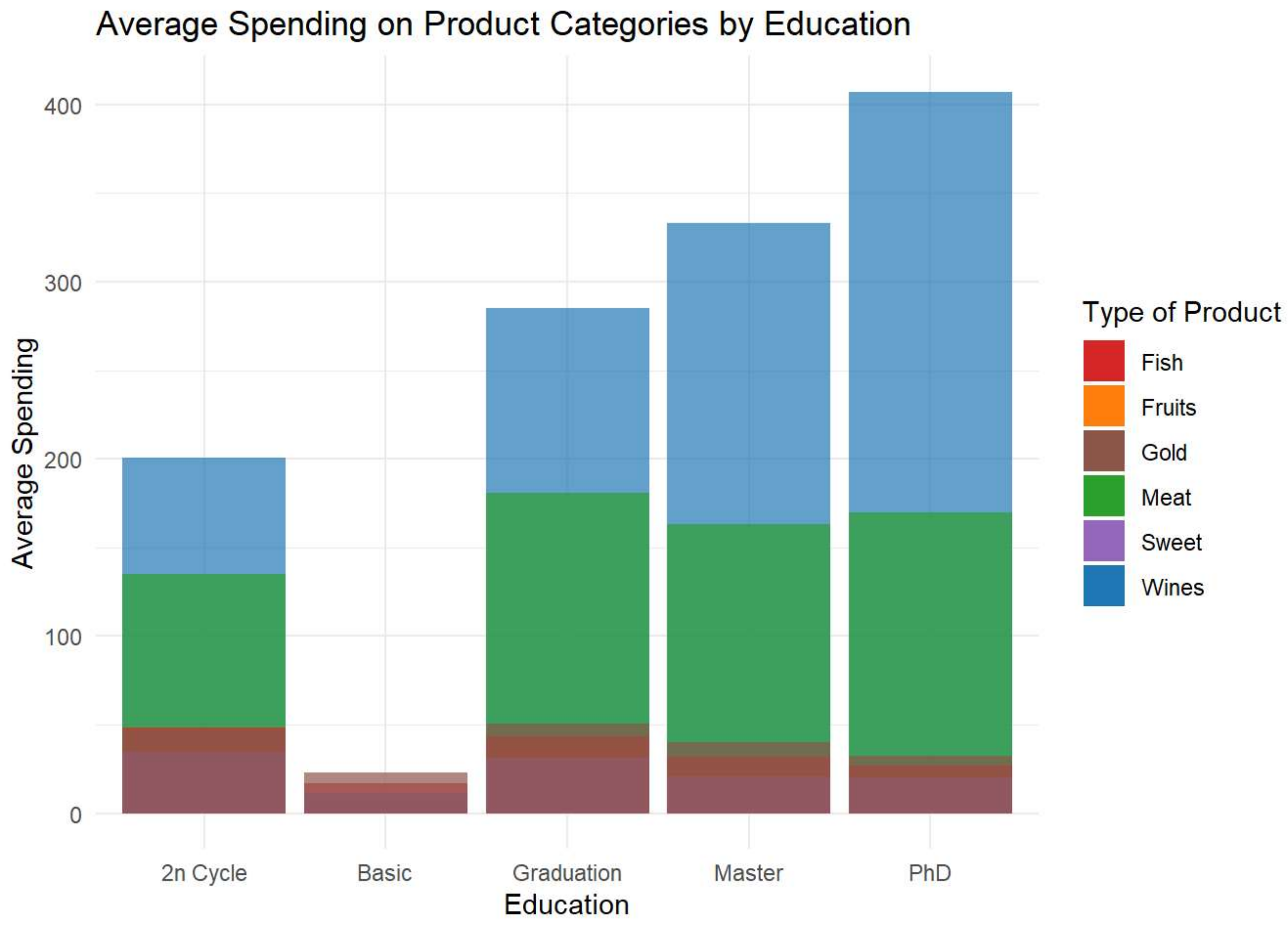
Wine is the most popular followed by by meat for average income households.

```
df %>%
  group_by(Age) %>%
  summarise(Wines = mean(MntWines), Fruits = mean(MntFruits), Meat = mean(MntMeatProducts), Fish
= mean(MntFishProducts), Sweet = mean(MntSweetProducts), gold = mean(MntGoldProds)) %>%
  ggplot(aes(x = Age)) +
  geom_point(aes(y = Wines, color = "Wines"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Fruits, color = "Fruits"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Meat, color = "Meat"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Fish, color = "Fish"), alpha = 0.7, size = 2) +
  geom_point(aes(y = Sweet, color = "Sweet"), alpha = 0.7, size = 2) +
  geom_point(aes(y = gold, color = "Gold Prods"), alpha = 0.7, size = 2) +
  scale_color_manual(name = "Products", values = c("Wines" = "#E69F00", "Fruits" = "#56B4E9", "M
eat" = "#009E73", "Fish" = "#0072B2", "Sweet" = "#D55E00", "Gold Prods" = "#CC79A7")) +
  theme_minimal() +
  theme(legend.position = "right")
```



Wine consumption increases over age. Whereas we see meat consumption beign high in early ages.

```
df %>%
  group_by(Education) %>%
  summarise(Wines = mean(MntWines), Fruits = mean(MntFruits), Meat = mean(MntMeatProducts), Fish
= mean(MntFishProducts), Sweet = mean(MntSweetProducts), gold = mean(MntGoldProds)) %>%
  ggplot(aes(x = Education)) +
  geom_bar(aes(y = Wines, fill = "Wines"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Fruits, fill = "Fruits"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Meat, fill = "Meat"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Fish, fill = "Fish"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = Sweet, fill = "Sweet"), stat = "identity", alpha = 0.7) +
  geom_bar(aes(y = gold, fill = "Gold"), stat = "identity", alpha = 0.7) +
  scale_fill_manual(values = c("Wines" = "#1F77B4", "Fruits" = "#FF7F0E", "Meat" = "#2CA02C", "F
ish" = "#D62728", "Sweet" = "#9467BD", "Gold" = "#8C564B")) +
  labs(title = "Average Spending on Product Categories by Education",
       x = "Education",
       y = "Average Spending",
       fill = "Type of Product")+
  theme_minimal() +
  theme(legend.position = "right")
```



PhDs consume more wine also the fact they are older validates the the relation with age

Feature Engineering

We Create following features for Data Modelling

1. Age (already Created)
2. Total Purchase (Already Created) : Spending sum on all goods
3. Is_Parent: If customer has kids home
4. Education: Undergraduate, Graduate, Post-Graduate
5. Has_Partner: If living with someone.
6. Family Size:
7. Active Days: Number of days since enrollment to last buys.
8. Campaign: If Participated in campaign.

```
df %>%
  select(Kidhome, Teenhome)
```

Kidhome <int>	Teenhome <int>
0	0
1	1