Assignment_Week1

Balaram Saripalli

Assignment

Use the knowledge gained from the Lab and the Discussion Activity to complete the assignment. The marketing.csv data set was used in a statistical analysis course at Hult International Business School.

Perform descriptive statistics and visualizations as instructed in lab and discussion activities. Anything else you may think will be relevant to analyzing this data set. *Provide a summary of your process and any insights you gathered through your analysis.* Turn in the R markdown and a knitted R markdown file as a pdf document of the assignment to the Week 1 dropbox. We will use this data set in future classes to perform more advanced statistical analyses.

1. Data Context

The data set marketing_data.csv consists of 2,240 customers of XYZ company with data on:

- Customers: ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Country: Customer's location

Products:

- MntWines: Amount spent on wine in the last 2 years
- MntFruits:Amount spent on fruits in the last 2 years
- MntMeatProducts: Amount spent on meat in the last 2 years
- MntFishProducts: Amount spent on fish in thelast 2 years
- MntSweetProducts: Amount spent on sweets in the last 2 years

Places:

- NumWebPurchases: Number of purchases made through the company's web site
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's web site in the last month

Promotion: - NumDealsPurchases: Number of purchases made with a discount - Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Assignment Solution

Initially the working directory need to be added, the path to you Rmd file folder need to be entered

```
# set working directory
setwd("F:/Balaram/ML course")
```

Loading the required libraries useful for data visualization and analysis

```
# load libraries
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(readr)
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

Loading the data # load data

```
# load the data and save it as mydata
mydata <- read_csv("marketing.csv",show_col_types = FALSE)
```

convert to data.table

```
mydata <- data.table(mydata)
```

check what you have with str

use the str i.e., structure function to understand my data

```r
str(mydata)
```

```
## Classes 'data.table' and 'data.frame':   2240 obs. of  19 variables:
##  $ ID                 : num  1826 1 10476 1386 5371 ...
##  $ Year_Birth         : num  1970 1961 1958 1967 1989 ...
##  $ Education          : chr  "Graduation" "Graduation" "Graduation" "Graduation" ...
##  $ Marital_Status     : chr  "Divorced" "Single" "Married" "Together" ...
##  $ Income             : chr  "$84,835.00" "$57,091.00" "$67,267.00" "$32,474.00" ...
##  $ Kidhome            : num  0 0 0 1 1 0 0 0 0 0 ...
##  $ Dt_Customer        : chr  "6/16/2014" "6/15/2014" "5/13/2014" "5/11/2014" ...
##  $ MntWines           : num  189 464 134 10 6 336 769 78 384 384 ...
##  $ MntFruits          : num  104 5 11 0 16 130 80 0 0 0 ...
##  $ MntMeatProducts    : num  379 64 59 1 24 411 252 11 102 102 ...
##  $ MntFishProducts    : num  111 7 15 0 11 240 15 0 21 21 ...
##  $ MntSweetProducts   : num  189 0 2 0 0 32 34 0 32 32 ...
##  $ MntGoldProds       : num  218 37 30 0 34 43 65 7 5 5 ...
##  $ NumDealsPurchases  : num  1 1 1 1 2 1 1 1 3 3 ...
##  $ NumWebPurchases    : num  4 7 3 1 3 4 10 2 6 6 ...
##  $ NumCatalogPurchases: num  4 3 2 0 1 7 10 1 2 2 ...
##  $ NumStorePurchases  : num  6 7 5 2 2 5 7 3 9 9 ...
##  $ Response           : num  1 1 0 0 1 1 1 0 0 0 ...
##  $ Country            : chr  "SP" "CA" "US" "AUS" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Details of basic structure

The given data has 2,240 observations with 19 variables

The descriptions of each variables and their datatypes are as follows:

- Customers: ID: Customer's unique identifier - num
- Year_Birth: Customer's birth year - num
- Education: Customer's education level - char (ordinal)
- Marital_Status: Customer's marital status - char (nominal)
- Income: Customer's yearly household income - num
- Kidhome: Number of children in customer's household - num
- Dt_Customer: Date of customer's enrollment with the company - chr (Dates)
- Country: Customer's location - char (nominal)

Products:

- MntWines: Amount spent on wine in the last 2 years - num
- MntFruits:Amount spent on fruits in the last 2 years - num
- MntMeatProducts: Amount spent on meat in the last 2 years - num
- MntFishProducts: Amount spent on fish in thelast 2 years - num
- MntSweetProducts: Amount spent on sweets in the last 2 years - num

Places:

- NumWebPurchases: Number of purchases made through the company's web site - num
- NumCatalogPurchases: Number of purchases made using a catalogue - num

- NumStorePurchases: Number of purchases made directly in stores - num
- NumWebVisitsMonth: Number of visits to company's web site in the last month - num

Promotion: - NumDealsPurchases: Number of purchases made with a discount - Response: 1 if customer accepted the offer in the last campaign, 0 otherwise - num (ordinal - encoded)

use summary() to get descriptive statistics on the data set

```
summary(mydata)
```

```
##        ID          Year_Birth    Education         Marital_Status
##  Min.   :    0   Min.   :1893   Length:2240        Length:2240
##  1st Qu.: 2828   1st Qu.:1959   Class :character   Class :character
##  Median : 5458   Median :1970   Mode  :character   Mode  :character
##  Mean   : 5592   Mean   :1969
##  3rd Qu.: 8428   3rd Qu.:1977
##  Max.   :11191   Max.   :1996
##     Income             Kidhome       Dt_Customer           MntWines
##  Length:2240        Min.   :0.0000   Length:2240        Min.   :   0.00
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.:  23.75
##  Mode  :character   Median :0.0000   Mode  :character   Median : 173.50
##                     Mean   :0.4442                      Mean   : 303.94
##                     3rd Qu.:1.0000                      3rd Qu.: 504.25
##                     Max.   :2.0000                      Max.   :1493.00
##    MntFruits     MntMeatProducts  MntFishProducts  MntSweetProducts
##  Min.   :  0.0   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  1.0   1st Qu.:  16.0   1st Qu.:  3.00   1st Qu.:  1.00
##  Median :  8.0   Median :  67.0   Median : 12.00   Median :  8.00
##  Mean   : 26.3   Mean   : 166.9   Mean   : 37.53   Mean   : 27.06
##  3rd Qu.: 33.0   3rd Qu.: 232.0   3rd Qu.: 50.00   3rd Qu.: 33.00
##  Max.   :199.0   Max.   :1725.0   Max.   :259.00   Max.   :263.00
##   MntGoldProds    NumDealsPurchases NumWebPurchases  NumCatalogPurchases
##  Min.   :  0.00   Min.   : 0.000    Min.   : 0.000   Min.   : 0.000
##  1st Qu.:  9.00   1st Qu.: 1.000    1st Qu.: 2.000   1st Qu.: 0.000
##  Median : 24.00   Median : 2.000    Median : 4.000   Median : 2.000
##  Mean   : 44.02   Mean   : 2.325    Mean   : 4.085   Mean   : 2.662
##  3rd Qu.: 56.00   3rd Qu.: 3.000    3rd Qu.: 6.000   3rd Qu.: 4.000
##  Max.   :362.00   Max.   :15.000    Max.   :27.000   Max.   :28.000
##  NumStorePurchases    Response        Country
##  Min.   : 0.00     Min.   :0.0000   Length:2240
##  1st Qu.: 3.00     1st Qu.:0.0000   Class :character
##  Median : 5.00     Median :0.0000   Mode  :character
##  Mean   : 5.79     Mean   :0.1491
##  3rd Qu.: 8.00     3rd Qu.:0.0000
##  Max.   :13.00     Max.   :1.0000
```

```
View(mydata)
# You can use summary for the entire data set to know the entire summary or individual columns summary
```

Observations: - ID is just a numerical variable which just a unique representation each customer which is a random number and thus not useful for analysis - The customers are ranged from the people who are born from 1893 to 1996.

show the first 6 rows of data with column names

```
head(mydata)
```

```
##       ID Year_Birth  Education Marital_Status     Income Kidhome Dt_Customer
##    <num>      <num>     <char>         <char>     <char>   <num>      <char>
## 1:  1826       1970 Graduation       Divorced $84,835.00       0   6/16/2014
## 2:     1       1961 Graduation         Single $57,091.00       0   6/15/2014
## 3: 10476       1958 Graduation        Married $67,267.00       0   5/13/2014
## 4:  1386       1967 Graduation       Together $32,474.00       1   5/11/2014
## 5:  5371       1989 Graduation         Single $21,474.00       1    4/8/2014
## 6:  7348       1958        PhD         Single $71,691.00       0   3/17/2014
##    MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
##       <num>     <num>           <num>           <num>            <num>
## 1:      189       104             379             111              189
## 2:      464         5              64               7                0
## 3:      134        11              59              15                2
## 4:       10         0               1               0                0
## 5:        6        16              24              11                0
## 6:      336       130             411             240               32
##    MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
##           <num>             <num>           <num>               <num>
## 1:          218                 1               4                   4
## 2:           37                 1               7                   3
## 3:           30                 1               3                   2
## 4:            0                 1               1                   0
## 5:           34                 2               3                   1
## 6:           43                 1               4                   7
##    NumStorePurchases Response Country
##                <num>    <num>  <char>
## 1:                 6        1      SP
## 2:                 7        1      CA
## 3:                 5        0      US
## 4:                 2        0     AUS
## 5:                 2        1      SP
## 6:                 5        1      SP
```

find how many countries are represented in the data

```
# as countries are repeated here and not a unique values we need to count the no of unique values
unique_countries <- unique(mydata$Country)
# Just to see what are the countries to cross verify
unique_countries
```

```
## [1] "SP"  "CA"  "US"  "AUS" "GER" "IND" "SA"  "ME"
```

```
no_countries <- length(unique_countries)
cat("no of countries represented in the data are:", no_countries, "\n")
```

```
## no of countries represented in the data are: 8
```

can you sort by the name of the country?

```
# We are using order function to sort by country and store that data as mydata1
mydata1 <- mydata[order(Country),]
# view the updated data
View(mydata1)
```

find mean and sd of in-store purchases in the US

```
# filter the data that is taken from us and the find the mean and SD of in-store purchases
# mean
Mean_US_instore <- mean(mydata$NumStorePurchases[mydata$Country == 'US'])
# Standard deviation
SD_US_instore <- sd(mydata$NumStorePurchases[mydata$Country == 'US'])

# Print the values clearly
cat("Mean of in-store purchanes in US are:", Mean_US_instore, "\n")
```

```
## Mean of in-store purchanes in US are: 6.036697
```

```
cat("Standard deviation of in-store purchanes in US are:", SD_US_instore, "\n")
```

```
## Standard deviation of in-store purchanes in US are: 3.360794
```

Before you can plot a histogram for income, you'll need to remove the dollar signs from the column.

```
mydata$Income <- parse_number(mydata$Income)
options(scipen = 9999)
```

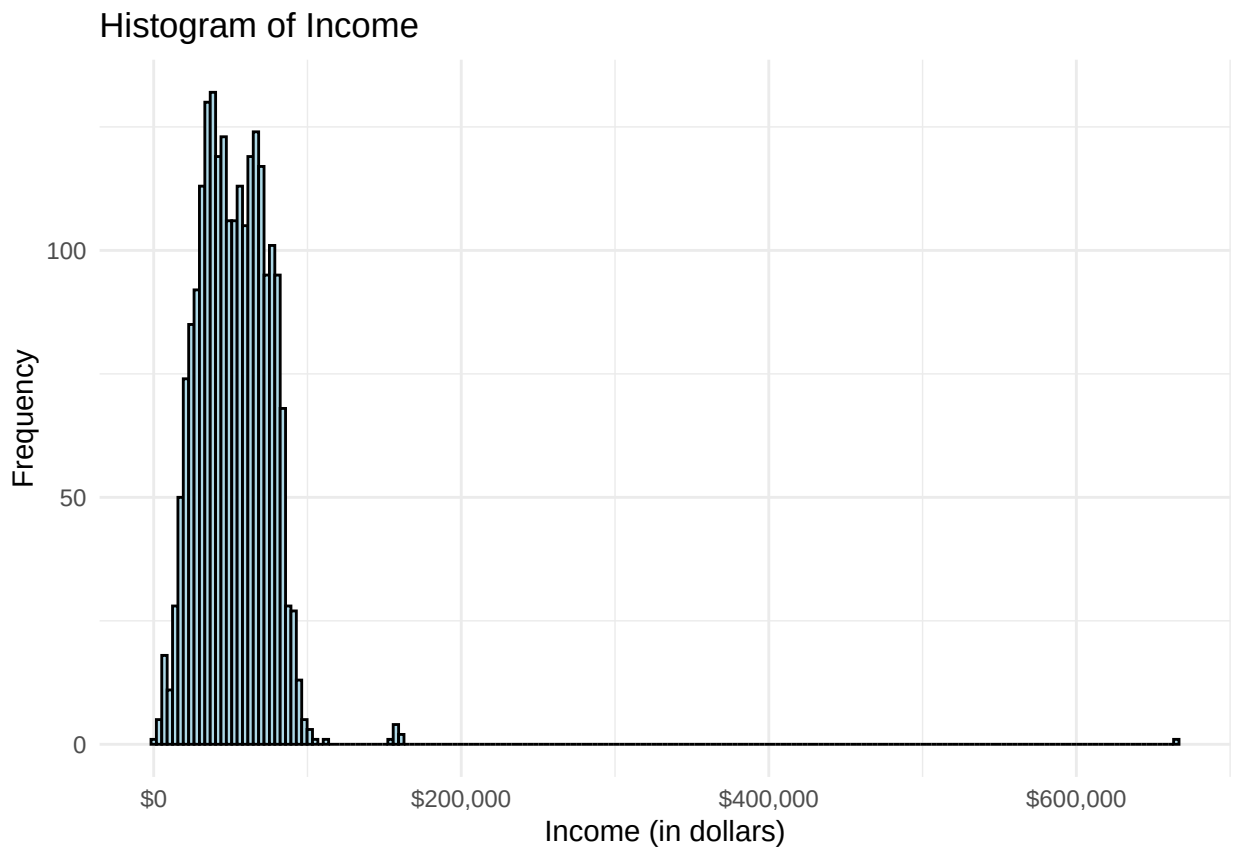
Set scipen to a higher value, so you can avoid numbers being displayed in scientific notation.

```
options(scipen=999)
```

histogram of Income

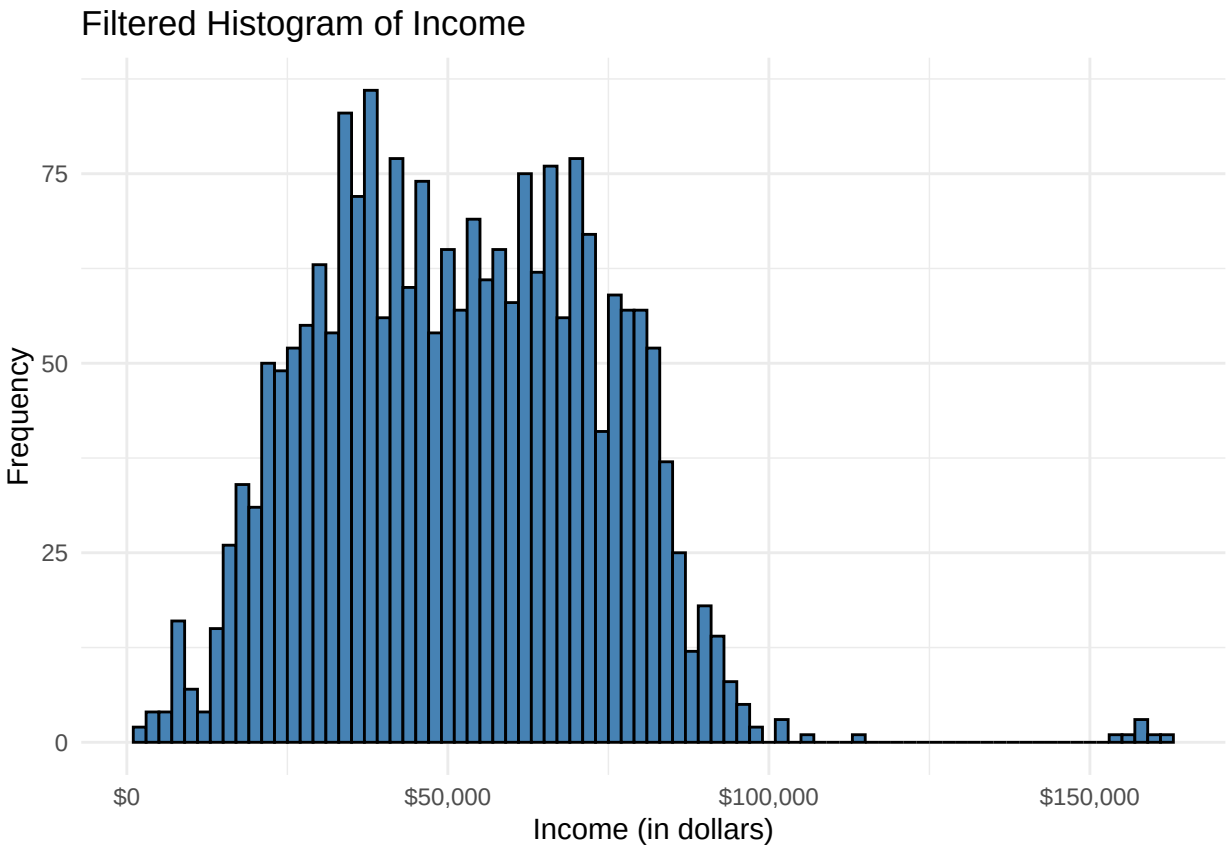
```
ggplot(mydata, aes(x = Income)) +
  geom_histogram(binwidth = 3500, fill = "lightblue", color = "black") +
  theme_minimal() +
  scale_x_continuous(labels = scales::dollar_format()) +
  labs(title = "Histogram of Income", x = "Income (in dollars)", y = "Frequency")
```

```
## Warning: Removed 24 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



We have somemissing values in the data and are ignored while doing the histograms Onservation: - Clearly more data is distributed between the 0$ an the 170,000$, and rest can be considered as outliers

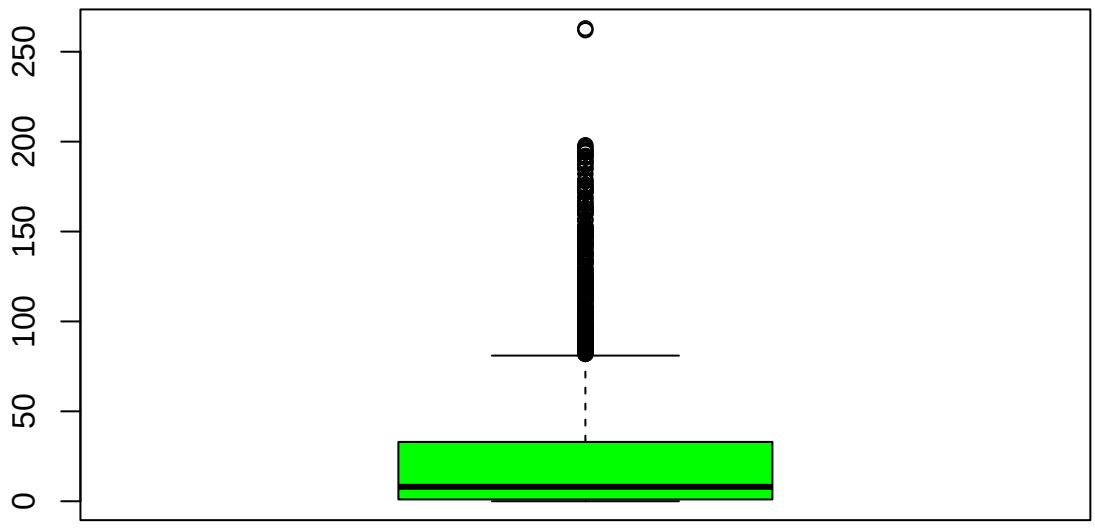
```
# we just removed the datapoints where icome is NA or > 200,000$ for better visualization
# Filtered Histogram of Income (NA removed and Income <= $200,000)
filtered_data <- mydata %>%
  dplyr::filter(!is.na(Income) & Income <= 200000)

ggplot(filtered_data, aes(x = Income)) +
  geom_histogram(binwidth = 2000, fill = "steelblue", color = "black") +
  theme_minimal() +
  scale_x_continuous(labels = scales::dollar_format()) +
  labs(title = "Filtered Histogram of Income", x = "Income (in dollars)", y = "Frequency")
```



Observation: most of the coustomers are in the range of income lies between 1,700$ to 100,000$

boxplot of Amount of Sweet Products

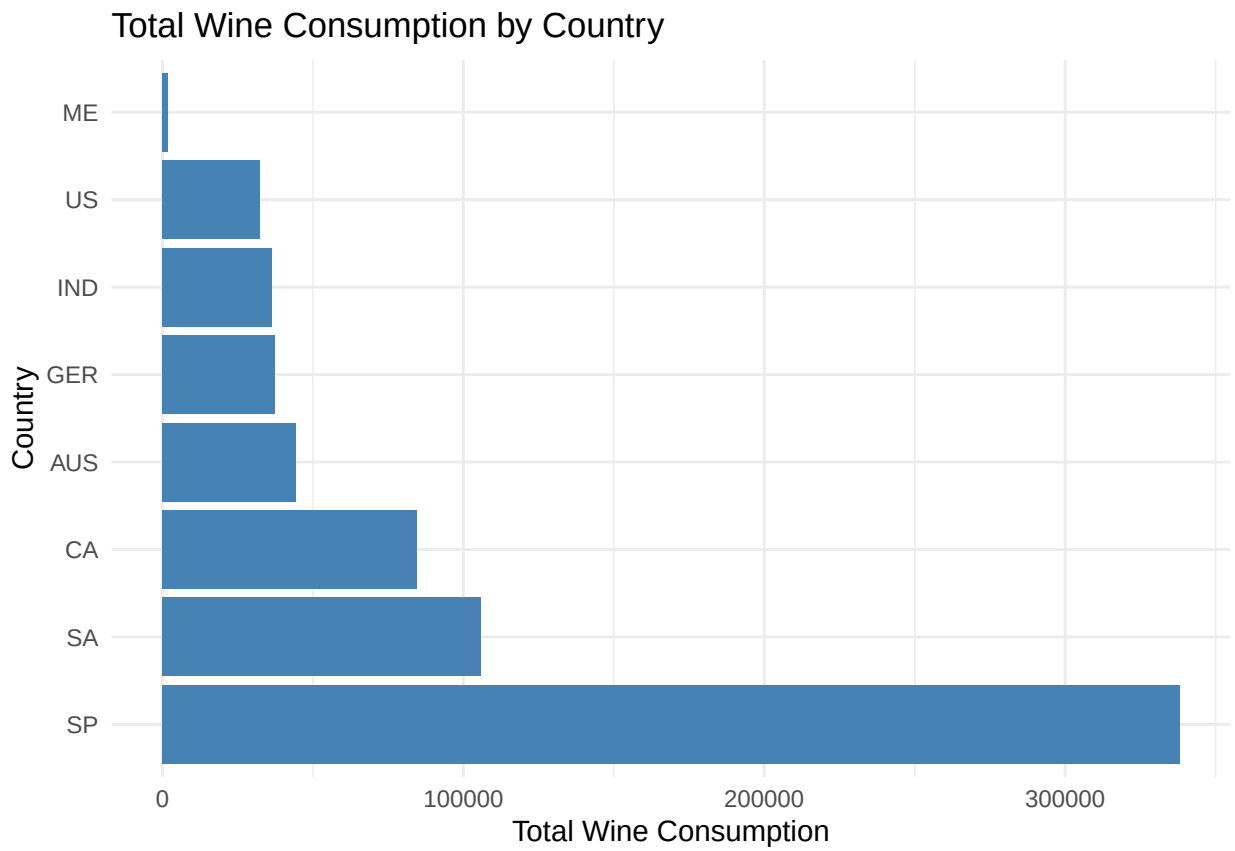
```
boxplot(mydata$`MntSweetProducts`, col='green', main='Boxplot of Amount spent on Sweet products for 2 y
```



which country is has the highest amount of wine consumed?

order plot by country with the highest wine consumption. You may use factor() function to be able to display amounts in a desirable order.Note: this is slightly different from the solution in the discussion activity.

```
# Country with Highest Wine Consumption
wine_by_country <- mydata %>%
  group_by(Country) %>%
  summarize(TotalWine = sum(MntWines, na.rm = TRUE)) %>%
  arrange(desc(TotalWine)) %>%
  mutate(Country = factor(Country, levels = Country))

ggplot(wine_by_country, aes(x = Country, y = TotalWine)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Total Wine Consumption by Country", x = "Country", y = "Total Wine Consumption")
```



Observation: country 'SP' i.e., Spain has highest total wine consumption, where as 'ME' i.e., middle east countries have lowest wine consumption

You may want to combine the Number of Store purchases, number of web purchases, and number of catalog purchases into a total number of purchases column to be used later in analysis stages.

```
#create totalpsum variable
mydata <- mydata %>%
  mutate(TotalPurchases = NumStorePurchases + NumWebPurchases + NumCatalogPurchases)
summary(mydata$TotalPurchases)
```

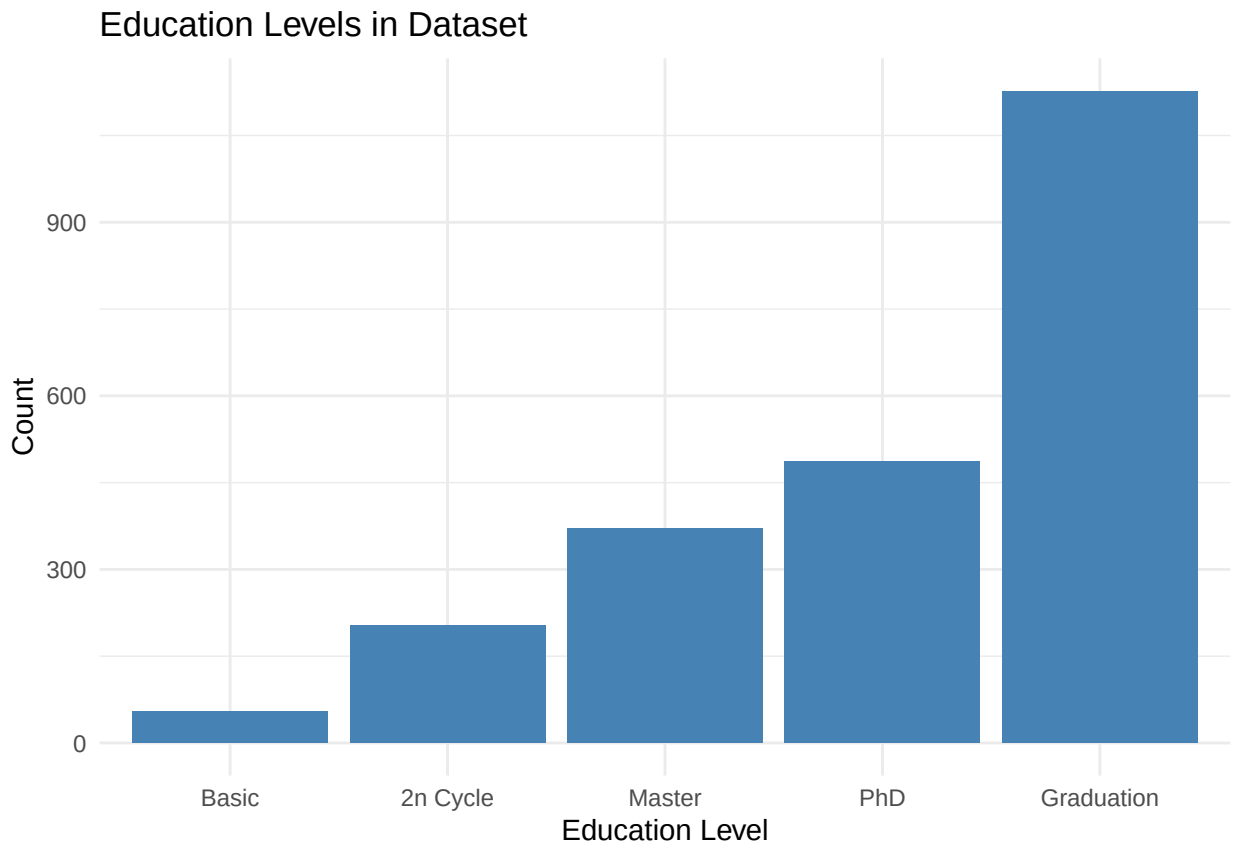
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    6.00   12.00   12.54   18.00   32.00
```

```
View(mydata)
#total purchases variable is also added in the data
```

Take a look at the education variable and see what it looks like.

```
education_summary <- mydata %>%
  group_by(Education) %>%
  summarize(Count = n())
ggplot(education_summary, aes(x = reorder(Education, Count), y = Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
```

```
  theme_minimal() +
  labs(title = "Education Levels in Dataset", x = "Education Level", y = "Count")
```



Observation: Most of the customers are having higher eductations like Master, PhD and graduation. The people whas graduation are more likely to cosume products in the company

Feel free to explore other variables that could be interesting to your analysis!

Effect of Discounts on Total Purchases

```
# Create a scatter plot of NumDealsPurchases vs TotalPurchases
ggplot(mydata, aes(x = NumDealsPurchases, y = TotalPurchases)) +
  geom_point(color = "steelblue", alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  theme_minimal() +
  labs(title = "Effect of Discounts on Total Purchases",
       x = "Number of Discount Purchases",
       y = "Total Purchases")
```

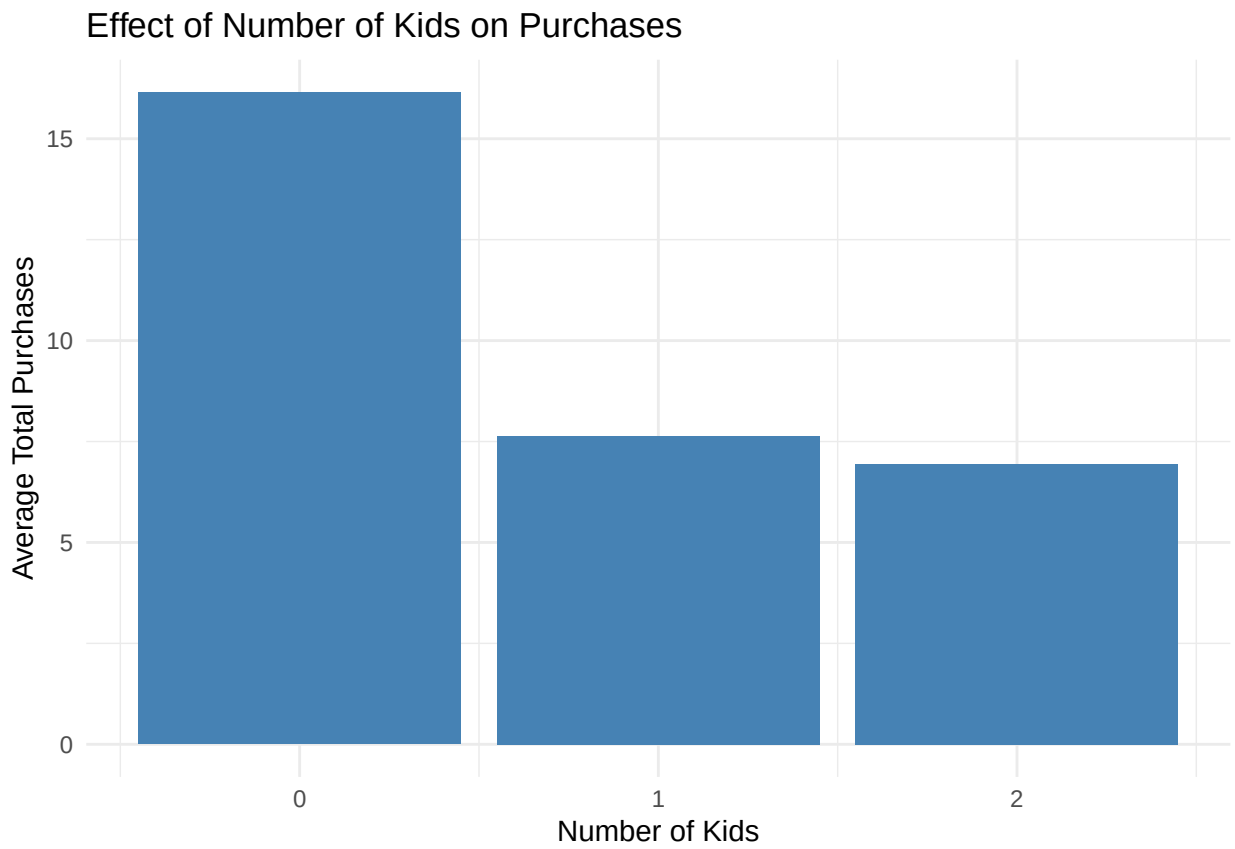
```
## `geom_smooth()` using formula = 'y ~ x'
```


```
# Correlation between NumDealsPurchases and TotalPurchases
correlation <- cor(mydata$NumDealsPurchases, mydata$TotalPurchases, use = "complete.obs")
correlation
```

```
## [1] 0.1178873
```

#Effect of no of kids on purchases

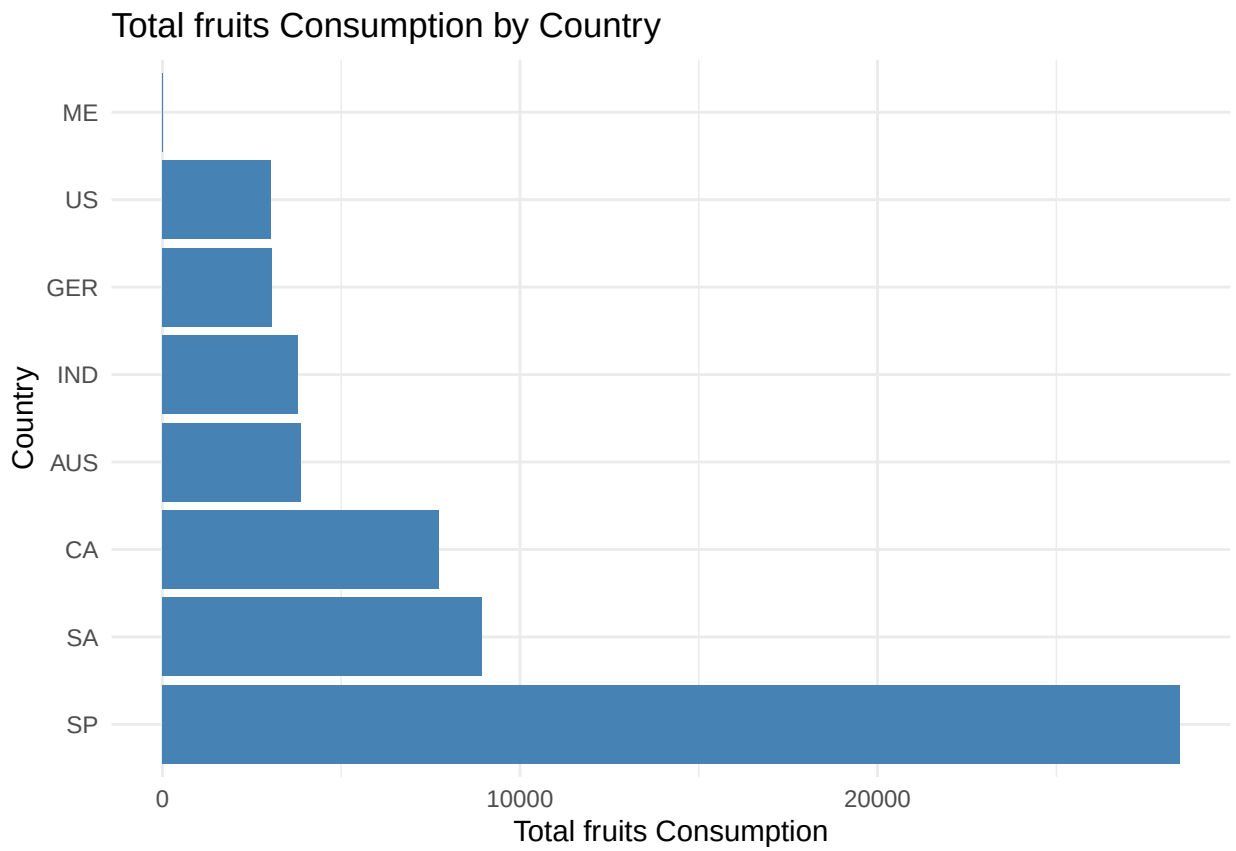
```
kids_effect <- mydata %>%
  group_by(Kidhome) %>%
  summarize(AveragePurchases = mean(TotalPurchases, na.rm = TRUE),
            Wine = mean(MntWines, na.rm = TRUE),
            Meat = mean(MntMeatProducts, na.rm = TRUE),
            Fish = mean(MntFishProducts, na.rm = TRUE),
            Fruits = mean(MntFruits, na.rm = TRUE),
            Sweets = mean(MntSweetProducts, na.rm = TRUE),
            Gold= mean(MntGoldProds, na.rm = TRUE))

ggplot(kids_effect, aes(x = Kidhome, y = AveragePurchases)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(title = "Effect of Number of Kids on Purchases", x = "Number of Kids", y = "Average Total Purcha
```



Observation: Average total purchages are more or the customers who doestnt have kids

#Country wise fruits consumption

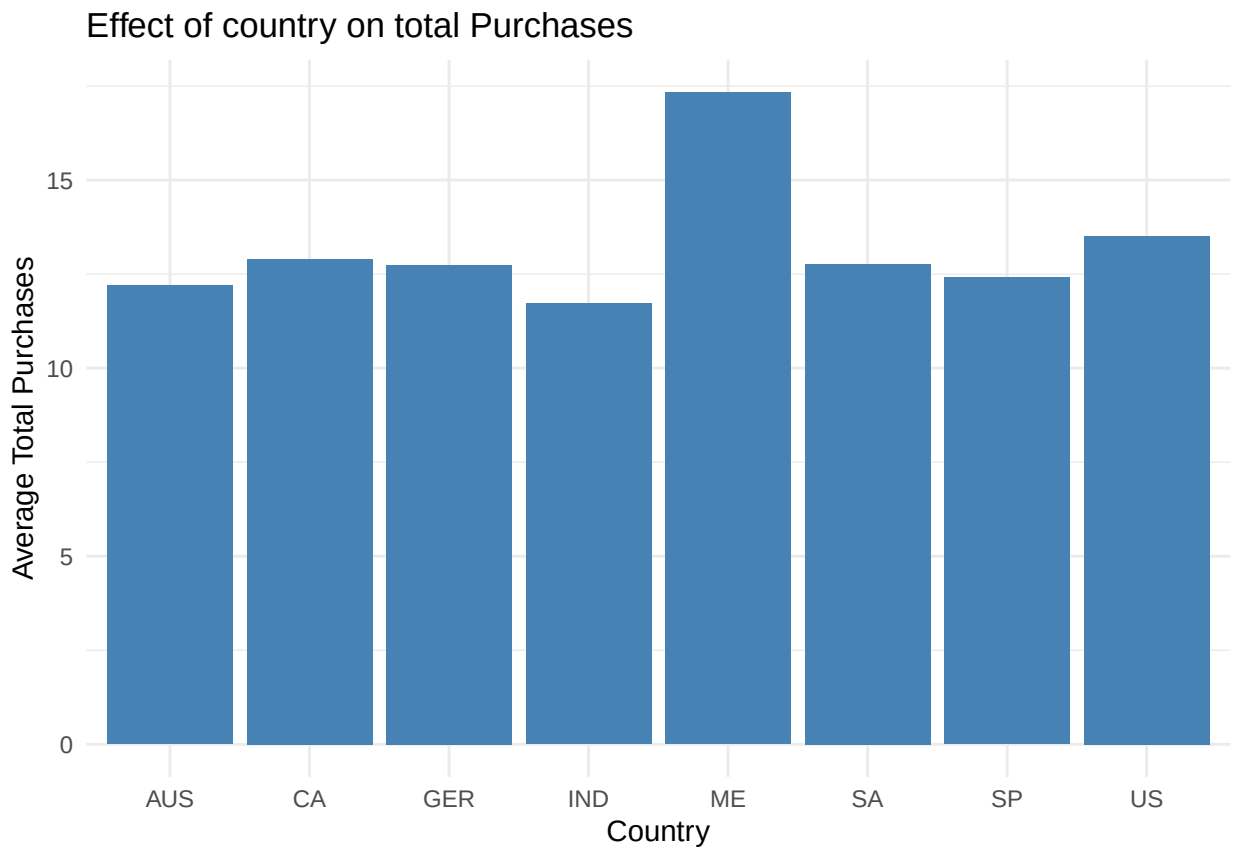
```
# Country with Highest Wine Consumption
fruits_by_country <- mydata %>%
  group_by(Country) %>%
  summarize(Totalfruits = sum(MntFruits, na.rm = TRUE)) %>%
  arrange(desc(Totalfruits)) %>%
  mutate(Country = factor(Country, levels = Country))

ggplot(fruits_by_country, aes(x = Country, y = Totalfruits)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Total fruits Consumption by Country", x = "Country", y = "Total fruits Consumption")
```



Observation: 'Sp' hasmore fruits consumption and ME has the lowest also, clearly ME has very less total consumpytions as well

effect of country on total consumptions

```
Country_effectpdt <- mydata %>%
  group_by(Country) %>%
  summarize(AveragePurchases = mean(TotalPurchases, na.rm = TRUE),
            Wine = mean(MntWines, na.rm = TRUE),
            Meat = mean(MntMeatProducts, na.rm = TRUE),
            Fish = mean(MntFishProducts, na.rm = TRUE),
            Fruits = mean(MntFruits, na.rm = TRUE),
            Sweets = mean(MntSweetProducts, na.rm = TRUE),
            Gold= mean(MntGoldProds, na.rm = TRUE))

ggplot(Country_effectpdt, aes(x = Country, y = AveragePurchases)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(title = "Effect of country on total Purchases", x = "Country", y = "Average Total Purchases")
```



Observation: Average total purchases is more in 'ME' countries

be sure to save your data frame to a csv file for future use.

```
library(data.table)
# Write ro CSV

#Or use this as this is much more elegant!
fwrite(mydata, "Marketingdata.csv")
```

Provide a summary of your process and any insights you gathered through your analysis with this data set. #summary of process In this assignment the marketing data was analyzed to understand the basic trends. We have initially loaded the data looked at the data and see what are all numerical features and what are characters or categorical features. The structure of the data and basic statistical summary is analyzed initially. Then we went in to little detailed information on how many countries were there as therae repetitions no of countries taken were found out. then we did simple mean and SD caluculations of Specific feature like in store purchases for a given country. The data was cleaned such as removal of $ from income so that it can be changed to numeric and helpful for analysis. Basic visualization like barchart, histogram and box plots were drawn in order to understand the trend / distribution of data. Created a new variable called Total purchases which includes purchases from instore, web and cataloge. we also analyzed the effect of discount on prchases and distribution of education in customers etc..

Insights

- Data: Marketing data , 2240 data points
- ID is just a numerical variable which just a unique representation each customer which is a random number and thus not useful for analysis
- The customers are ranged from the people who are born from 1893 to 1996.
- The data consists of 19 features to understand the marketinng of a company
- no of countries represented in the data are: 8
- The countries are SP (spain), CA (Canada), US(United states), AUS (Australia), GER (Geremany),IND (India), SA (south africa) and ME (Middle east)
- Mean of in-store purchanes in US are: 6.036697
- Standard deviation of in-store purchanes in US are: 3.360794
- most of the coustomers are in the range of income lies between 1,700$ to 100,000$
- There is a skewness on the amont spent on sweets - country 'SP' i.e., Spain has highest total wine consumption, where as 'ME' i.e., middle east countries have lowest wine consumption
- Most of the customers are having higher eductations like Master, PhD and graduation.
- The people whas graduation are more likely to cosume products in the company
- There is no major effect of discounts on the total no of purchases
- Average total purchages are more or the customers who doestnt have kids
- 'Sp' hasmore fruitsconsumption and ME has the lowest also, clearly ME has very less total consumptions as well
- Average total purchases is more in 'ME' countries