Assignment_Week1

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 or html 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 the last 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

set working directory

load libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(dplyr)
library(ggplot2)
```

load data

```
mydata <- read_csv("C:/Users/miche/Documents/MSDS660/Week1/marketing.csv",show_col_types = FALS
E)
```

convert to data.table

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

check what you have with str

```
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>
```

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
```

show the first 6 rows of data with column names

```
head(mydata)
```

ID <dbl>	Year_Birth <dbl>	Education <chr>	Marital_Status <chr>	Income <chr>	Kidh… <dbl>	Dt_Customer <chr>	MntWi… <dbl>	Mı
1826	1970	Graduation	Divorced	$84,835.00	0	6/16/2014	189	
1	1961	Graduation	Single	$57,091.00	0	6/15/2014	464	
10476	1958	Graduation	Married	$67,267.00	0	5/13/2014	134	

ID	Year_Birth	Education	Marital_Status	Income	Kidh…	Dt_Customer	MntWi…	Mr
<dbl>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<chr>	<dbl>	
1386	1967	Graduation	Together	$32,474.00	1	5/11/2014	10	
5371	1989	Graduation	Single	$21,474.00	1	4/8/2014	6	
7348	1958	PhD	Single	$71,691.00	0	3/17/2014	336	

3 rows | 1-9 of 10 columns

find how many countries are represented in the data

```
unique(mydata$Country)
```

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

can you sort by the name of the country?

```
mydata_sort <- mydata[order(Country),]
```

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

```
mean(mydata$NumStorePurchases[mydata$Country == 'US'])
```

```
## [1] 6.036697
```

```
sd(mydata$NumStorePurchases[mydata$Country == 'US'])
```

```
## [1] 3.360794
```

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

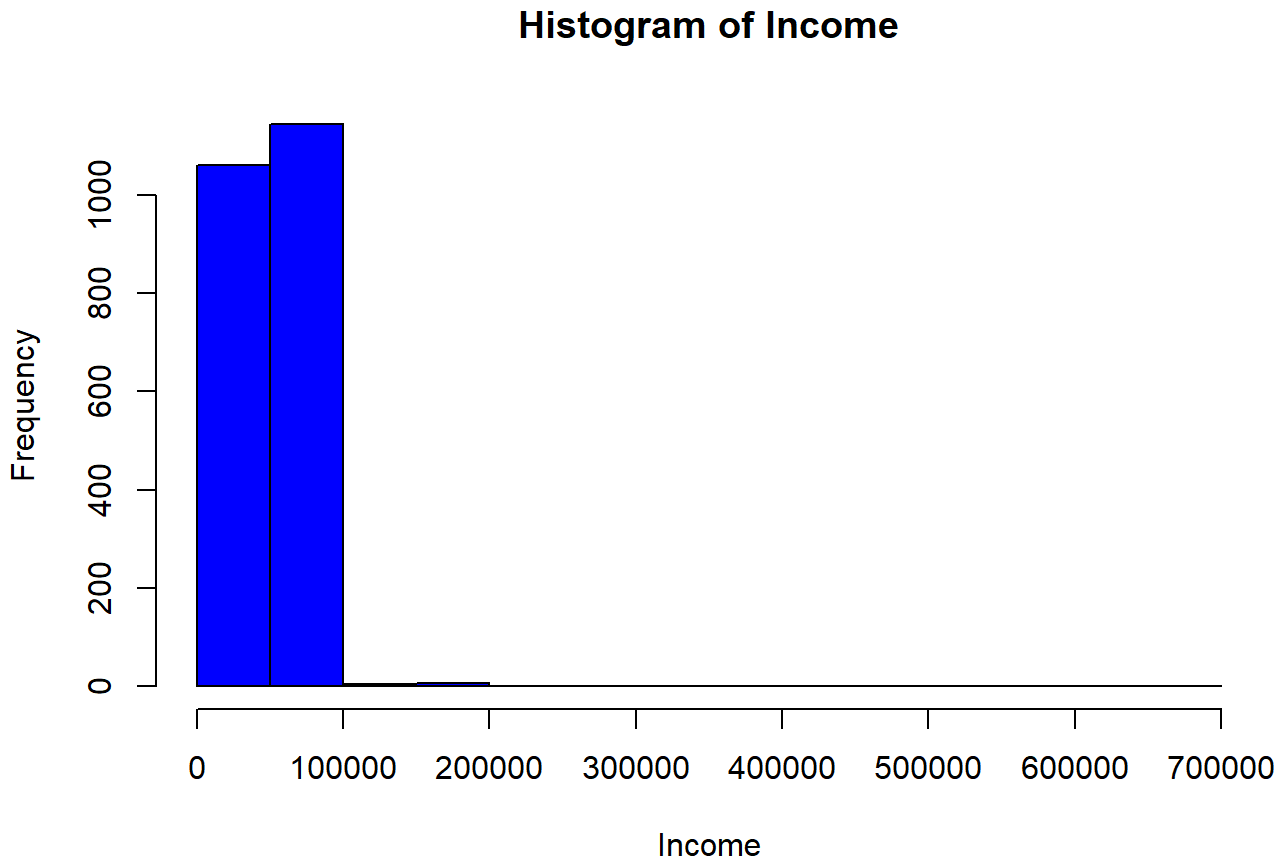
```
mydata$Income <- parse_number(mydata$Income)
```

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

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

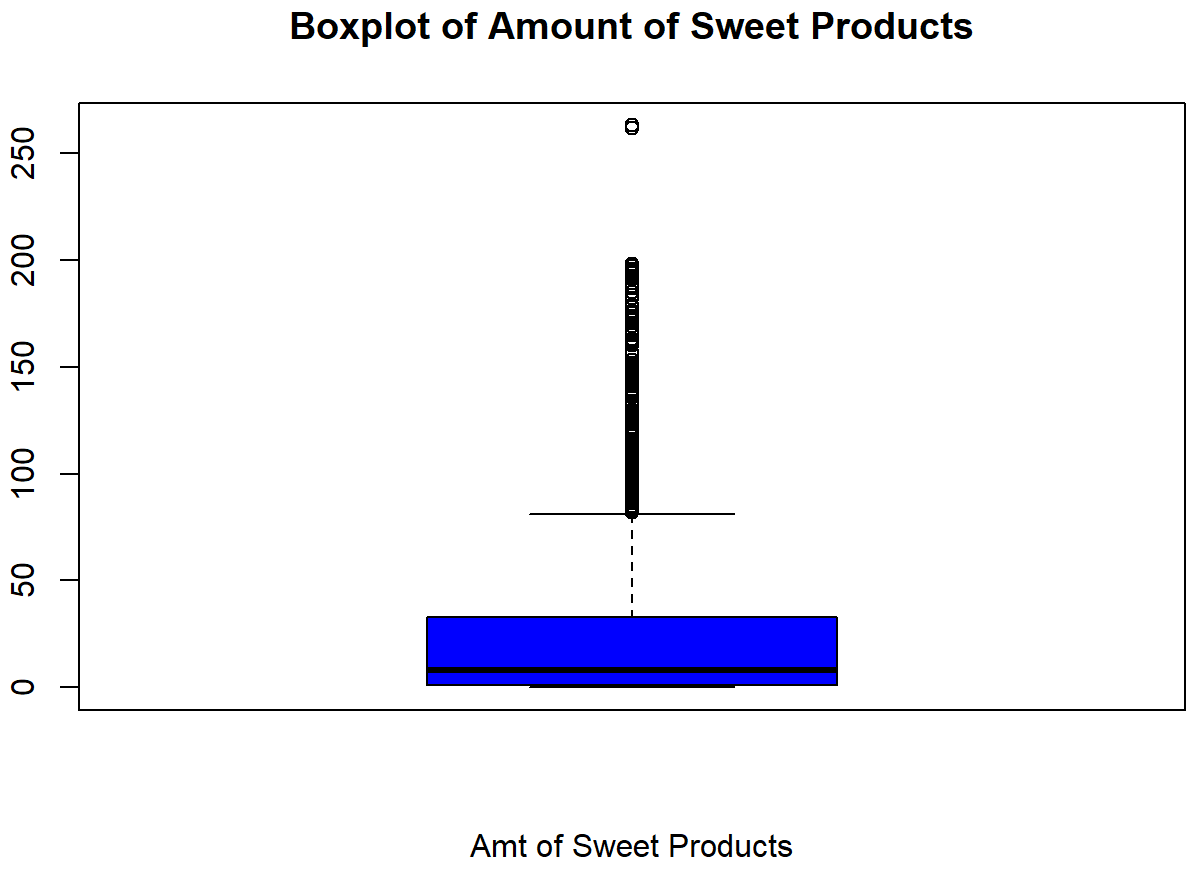
histogram of Income

```
hist(mydata$Income, col='blue',main='Histogram of Income', xlab = 'Income')
```



boxplot of Amount of Sweet Products

```
boxplot(mydata$MntSweetProducts, col='blue', main='Boxplot of Amount of Sweet Products', xlab =
'Amt of Sweet Products')
```



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.

need to Group by ‘Country’ and sum the ‘MntWines’ values for each Country

```
wine_consumed <- mydata[, .(Ttl_Wine = sum(MntWines, na.rm = TRUE)), by = Country]

wine_consumed$Country <- factor(wine_consumed$Country)

wine_consumed <- wine_consumed[order(-Ttl_Wine)]

print(wine_consumed)
```

```
##    Country Ttl_Wine
##     <fctr>    <num>
## 1:      SP   337991
## 2:      SA   105918
## 3:      CA    84649
## 4:     AUS    44372
## 5:     GER    37483
## 6:     IND    36268
## 7:      US    32406
## 8:      ME     1729
```

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[, totalpsum := NumStorePurchases + NumWebPurchases + NumCatalogPurchases]
```

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

```
unique(mydata$Education)
```

```
## [1] "Graduation" "PhD"        "2n Cycle"   "Master"     "Basic"
```

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

```
education_counts <- table(mydata$Education)
print(education_counts)
```

```
##
##   2n Cycle      Basic Graduation     Master        PhD
##        203         54       1127        370        486
```

```
min_year_birth <- mydata[mydata$Year_Birth == 1893, ]
print(min_year_birth)
```

```
##       ID Year_Birth Education Marital_Status Income Kidhome Dt_Customer
##    <num>      <num>    <char>         <char>  <num>   <num>      <char>
## 1: 11004       1893  2n Cycle         Single  60182       0   5/17/2014
##    MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
##       <num>     <num>           <num>           <num>            <num>
## 1:        8         0               5               7                0
##    MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
##           <num>             <num>           <num>               <num>
## 1:            2                 1               1                   0
##    NumStorePurchases Response Country totalpsum
##                <num>    <num>  <char>     <num>
## 1:                 2        0      SA         3
```

```
response_counts <- table(mydata$Response)
print(response_counts)
```

```
##
##    0    1
## 1906  334
```

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

```
library(data.table)
fwrite(mydata,"C:\\Users\\miche\\Documents\\MSDS660\\Week1\\marketing_data.csv")
```

Provide a summary of your process and any insights you gathered through your analysis with this data set. First, the marketing data was loaded and converted into a data table, after which it was checked for variables and their types. The data set consists of 19 variables, with 5 character-type variables and 14 numeric-type variables. It includes customer data from 8 different countries. Focusing on the United States, the average number of in-store purchases is approximately 6, with a standard deviation of around 3. The income of customers across all countries is generally below $100,000, with only a small number of customers earning between $200,000 and $700,000. The country with the highest wine purchases is Spain (SP). The response column indicates whether a customer accepted the offer in the last campaign, with a value of 1 representing acceptance and 0 indicating otherwise. Examining the data, it was found that 334 out of 2,240 customers accepted the offer, meaning that only about 15% of XYZ company's customers accepted the offer in the last campaign. One noteworthy observation was when examining the customer's birth year. When some exploratory analysis was done on this variable, it was found that the minimum birth year was 1893, and with a customer enrollment year of 2014, this would make the customer 121 years old at that time. This seems highly unlikely, suggesting there may be an error or inconsistency in that row of data, which will need further investigation during subsequent analysis.