Assignment_Week2

Adam Peetz

MSDS660 Week 2 Assignment

Regis University

Dr. Siripun Sanguansintukul

October 30th 2022

Exploratory Data Analysis of Marketing Data

Finding variables that correlate to each other can offer insights into how to make improvements to business performance. Correlation can be calculated for the entire data set using the cor() function in R. The correlation matrix this produces can be plotted to generate a heat map. Correlation heat maps visualize the relationship between many variables at a glace which can guide future decisions made about the data set.

First set the working directory, load the required libraries, and import the data set.

Set Working Directory

```
#set working directory
setwd("C:\\Users\\adamg\\Documents\\MSDS_660\\Week_2")
```

Load Libraries

```
library (tidyverse)
library(data.table)
library(dplyr)
library(ggplot2)

#heatmap and custom colors
#install.packages("reshape2")
library(reshape2)
#install.packages("viridis")
library("viridis")
library(gridExtra)
```

Load Data

```
df<-read_csv("cleaned_marketing_data.csv")
```

Convert to data.table

```
#convert to data.table
setDT(df)
```

Visualize the raw data with head().

Checking the data frame with head() reveals some basic information about the data it contains. Useful info provided by this function includes the variable names, column contents, and types for each of the 21 variables in the data set.

```
#Check what you have with head()
head(df)
```

```
##       ID Year_Birth  Education Marital_Status Income Kidhome Dt_Customer
## 1:  1826       1970 Graduation       Divorced  84835       0   6/16/2014
## 2:     1       1961 Graduation         Single  57091       0   6/15/2014
## 3: 10476       1958 Graduation        Married  67267       0   5/13/2014
## 4:  1386       1967 Graduation       Together  32474       1   5/11/2014
## 5:  5371       1989 Graduation         Single  21474       1    4/8/2014
## 6:  7348       1958        PhD         Single  71691       0   3/17/2014
##    MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 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
## 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 totalpsum totalpmnt
## 1:                 6        1      SP        14      1190
## 2:                 7        1      CA        17       577
## 3:                 5        0      US        10       251
## 4:                 2        0     AUS         3        11
## 5:                 2        1      SP         6        91
## 6:                 5        1      SP        16      1192
```

Exploring dataset correlations with a heatmap.

Correlation can only be calculated for numeric features. First, subset the data to remove all non-numeric features, then process it and draw a heatmap.

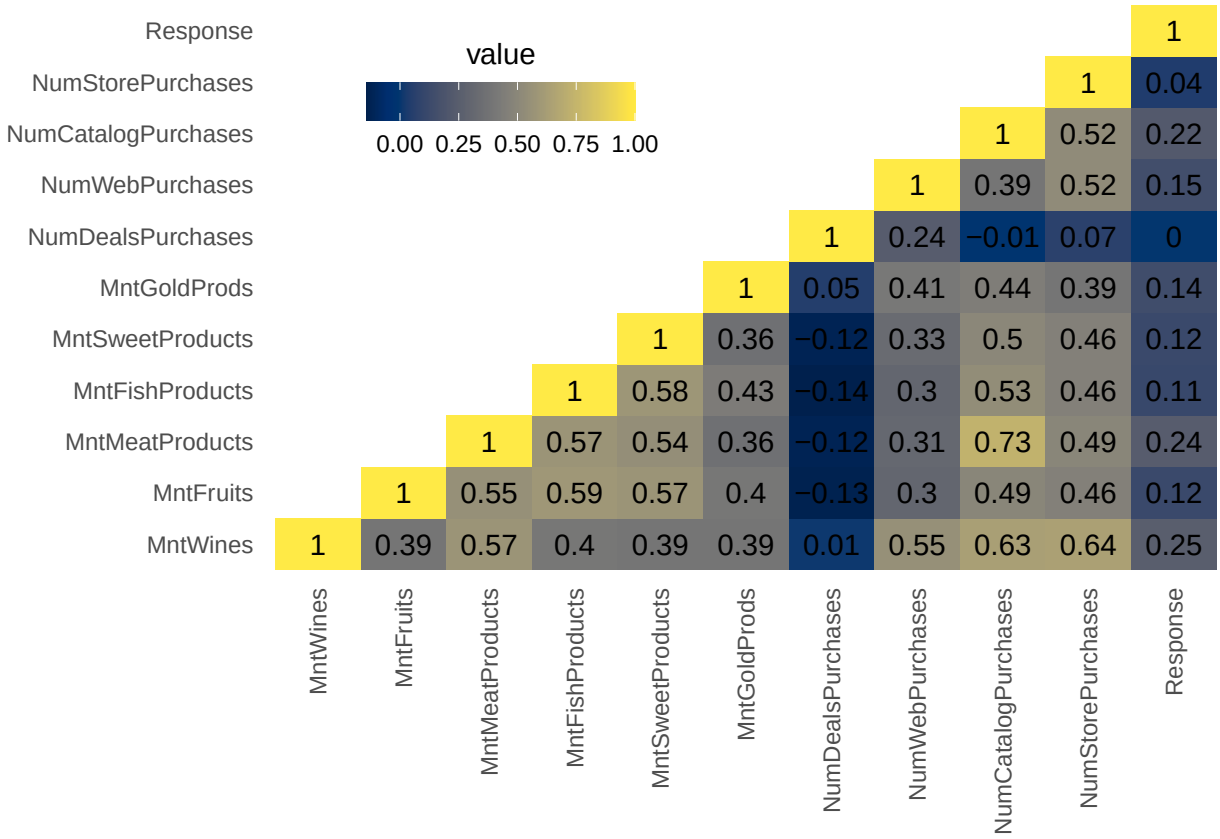
```
#subset data to numeric values for sales amounts   and counts
sales_corr_df <- subset(df, select = -c(ID,
                                        Year_Birth,
                                        Education,
                                        Marital_Status,
                                        Income,
                                        Kidhome,
                                        Dt_Customer,
                                        Country,
                                        totalpsum,
                                        totalpmnt
                        )               )
```

```r
#demonstrate output
summary(sales_corr_df)
```

```
##     MntWines        MntFruits      MntMeatProducts  MntFishProducts
##  Min.   :   0.0   Min.   :  0.00   Min.   :   0.0   Min.   :  0.00
##  1st Qu.:  24.0   1st Qu.:  2.00   1st Qu.:  16.0   1st Qu.:  3.00
##  Median : 174.5   Median :  8.00   Median :  68.0   Median : 12.00
##  Mean   : 305.1   Mean   : 26.36   Mean   : 167.0   Mean   : 37.64
##  3rd Qu.: 505.0   3rd Qu.: 33.00   3rd Qu.: 232.2   3rd Qu.: 50.00
##  Max.   :1493.0   Max.   :199.00   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 : 24.50   Median : 2.000    Median : 4.000
##  Mean   : 27.03   Mean   : 43.97   Mean   : 2.324    Mean   : 4.085
##  3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000    3rd Qu.: 6.000
##  Max.   :262.00   Max.   :321.00   Max.   :15.000    Max.   :27.000
##  NumCatalogPurchases NumStorePurchases    Response
##  Min.   : 0.000      Min.   : 0.000    Min.   :0.0000
##  1st Qu.: 0.000      1st Qu.: 3.000    1st Qu.:0.0000
##  Median : 2.000      Median : 5.000    Median :0.0000
##  Mean   : 2.671      Mean   : 5.801    Mean   :0.1503
##  3rd Qu.: 4.000      3rd Qu.: 8.000    3rd Qu.:0.0000
##  Max.   :28.000      Max.   :13.000    Max.   :1.0000
```

```r
#http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visuali
#https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/

#define lower triangle function
get_lower_tri<-function(cormat){
    cormat[lower.tri(cormat)] <- NA
    return(cormat)}

#define upper triangle function
get_upper_tri <- function(cormat){
    cormat[upper.tri(cormat)]<- NA
    return(cormat)}

#translate dataframe to correlation dataframe
cormap <- round(cor(sales_corr_df),2)

#get lower triangle
tri <- get_lower_tri(cormap)

#melt the corrleation dataframe
melted_cormap <- melt(tri, na.rm=TRUE)

#apply gg plotting function
ggplot(data = melted_cormap, aes(x=Var2, y=Var1, fill=value)) +
        geom_tile() +
        geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
        scale_fill_viridis(discrete = FALSE, option="E") +
        theme(
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
```

```
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank(),
        legend.justification = c(1, 0),
        legend.position = c(0.4, 0.7),
        legend.direction = "horizontal")+
        guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                     title.position = "top", title.hjust = 0.5))
```

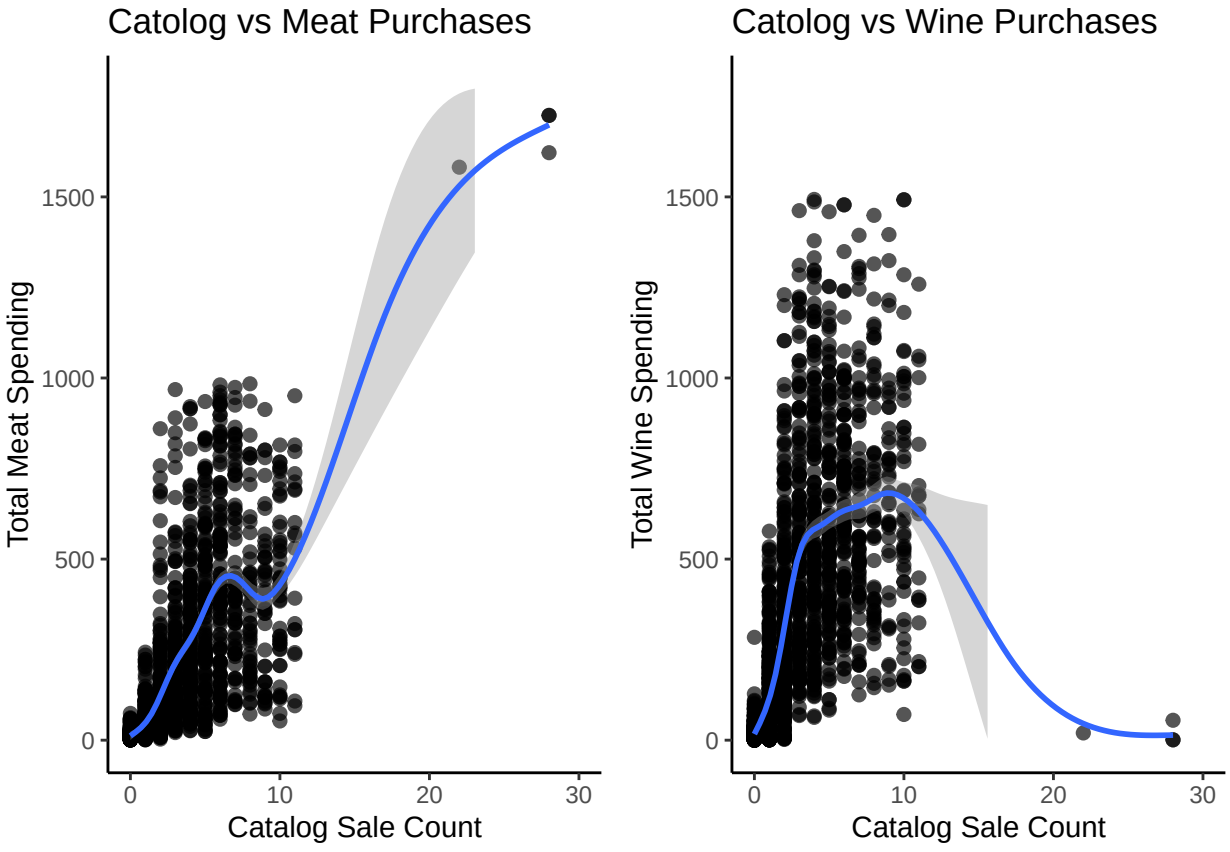


Exploring variable relationships with scatterplots.

The heat map shows the number of catalog purchases and amount spent on meat products with the strongest correlation of any two variables in the data set. A business owner could use this information to improve catalog sales by changing the selection of products in the catalog to cater to meat orders. The relationship between these variables can also be compared with a scatter plot.

```
#generate scatterplot #1 for catalog purchases vs meat products
p1 <- ggplot(df, aes(x = NumCatalogPurchases, y = MntMeatProducts))+
     geom_point(alpha = (2/3), size = 2) +
     geom_smooth() +
     ggtitle("Catolog vs Meat Purchases") +
     labs(y= "Total Meat Spending", x = "Catalog Sale Count") +
     xlim(0,30)+
```

```
     ylim(0,1800)+
     theme_bw() +
     theme(panel.border = element_blank(),
           panel.grid.major = element_blank(),
           panel.grid.minor = element_blank(),
           axis.line = element_line(colour = "black"))

#generate scatterplot #2 for catalog vs wine products
p2 <- ggplot(df, aes(x = NumCatalogPurchases, y = MntWines)) +
     geom_point(alpha = (2/3), size = 2) +
     geom_smooth() +
     ggtitle("Catolog vs Wine Purchases") +
     labs(y= "Total Wine Spending", x = "Catalog Sale Count") +
     xlim(0,30)+
     ylim(0,1800)+
     theme_bw() +
     theme(panel.border = element_blank(),
           panel.grid.major = element_blank(),
           panel.grid.minor = element_blank(),
           axis.line = element_line(colour = "black"))

#arrange plots side by side
grid.arrange(p1, p2, ncol=2)
```



Scatterplot Discussion

The scatter plots above reveal 3 customers who have spent much more money on meat when compared to other product types such as wine. These three customers are outliers in the data set who skew the correlation score for meat and catalog orders. Meat has a weaker relationship to catalog purchases when compared to wine at smaller order levels. A business owner may not want to adjust their product offerings to suite the preferences of these 3 outliers.

Checking distributions and outliers with boxplots.

Histograms and box plots can be used to check for outliers and visualize the distribution of a feature. These two methods are used in the cell below to compare the distribution of meat and wine sales. Checking the distribution of these variables shows that average sales of wine exceed meat except for the few outliers identified in the data set.

```
#create boxplot
p3 <- ggplot(df,aes(x=MntMeatProducts)) +
  geom_boxplot(fill='orange') +
  ggtitle("Distribution of Sale Amounts: Meats") +
  xlim(0,1800)+
  theme_bw() +
  theme(axis.title.x = element_blank(),
        axis.line.y = element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black")
        )

#create histogram
p4 <- ggplot(df,aes(x=MntMeatProducts)) +
  geom_histogram(bins = 100, fill="orange") +
  xlim(0,1800)+
  ylim(0,100)+
  labs(x = "Meat Sale Amount") +
  theme_bw() +
  theme(axis.title.y=element_blank(),
        axis.line.y = element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black")
        )

#create boxplot
p5 <- ggplot(df,aes(x=MntWines)) +
  geom_boxplot(fill='purple') +
  ggtitle("Distribution of Sale Amounts: Wines") +
  xlim(0,1800) +
  labs(x = "Wine Sale Amount") +
  theme_bw() +
```

```
  theme(axis.title.x = element_blank(),
        axis.line.y = element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black")
        )

#create histogram
p6 <- ggplot(df,aes(x=MntWines)) +
  geom_histogram(bins = 100, fill="purple") +
  xlim(0,1800) +
  ylim(0,100)+
  labs(x = "Wine Sale Amount") +
  theme_bw() +
  theme(axis.title.y=element_blank(),
        axis.line.y = element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black")
        )

#arrange plots side by side
grid.arrange(p3, p4, p5, p6, nrow=4)
```

Conclusion

The strength of correlation between numeric variables in the data set was reviewed with a heat map. This heat map suggested meat and catalog purchases were strongly correlated. The relationship of these two variables was explored in more depth and compared to the relationship of wine with scatter plots. These scatter plots revealed the correlation between meats and catalog purchases was a result of a few outliers in the data set. Further exploration into the distribution of sales for wine and meats showed average sales are higher for wines than meats.

References

Alboukadel. (2020). Top R Color Palettes to Know for Great Data Visualization. DataNovia. retreived 10/28/22 from https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/

Bryan, Jennifer. (2017). ggplot2-tutorial . github.com. retrieved 10/22/22 from https://github.com/jennybc/ggplot2-tutorial

Hult International Business School. (n.d.). marketing data . dataset. retrieved 10/22/22 from https://worldclass.regis.edu/d2l/le/content/297311/Home

STHDA.(n.d.). ggplot2 : Quick correlation matrix heatmap - R software and data visualization . Statistical tools for high-throughput data analysis. retreived 10/28/22 from http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization