EDA Food Mart R Markdown File

Abdul Jawad Mohammed

Atharva Atre

Supreeth Bannur Sudhakara

Keerthi Kumar Reddy Kancham Reddy

2022-10-21

# I. Loading Packages

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tidyr)  
library(ggplot2)  
library(dplyr)  
#install.packages('corrplot')  
library(corrplot)

## corrplot 0.92 loaded

CFM\_data <- read.csv("C:/Users/bssup/Documents/fall22/R/project/Food\_mart/data.csv")

# II. About Food Mart:

Food Mart (CFM) is a chain of convenience stores in the United States. The private company’s headquarters are located in Mentor, Ohio, and there are currently approximately 325 stores located in the US. Convenient Food Mart operates on the franchise system. Food Mart was the nation’s third-largest chain of convenience stores as of 1988. The NASDAQ exchange dropped Convenient Food Mart the same year when the company failed to meet financial reporting requirements. Carden & Cherry advertised Convenient Food Mart with the Ernest character in the 1980s.

# III. Loading Data

#Load Data  
CFM\_data <- read.csv("C:/Users/bssup/Documents/fall22/R/project/Food\_mart/data.csv")

# IV. Correlation Matrix

This will show correlation between different variables. Reasons of computing correlation matrix: To summarize a large amount of data where the goal is to see patterns. In our example above, the observable pattern is that all the variables highly correlate with each other.

#Correlation Matrix  
store\_data <- select(CFM\_data,c('store\_sales.in.millions.','store\_cost.in.millions.','unit\_sales.in.millions.','total\_children','avg\_cars\_at.home.approx.','num\_children\_at\_home','avg\_cars\_at.home.approx..1','SRP','gross\_weight','net\_weight','recyclable\_package','low\_fat','units\_per\_case','store\_sqft','grocery\_sqft','frozen\_sqft','meat\_sqft','coffee\_bar','video\_store','salad\_bar','prepared\_food','florist','cost'))  
#store\_data  
cor\_data <- cor(store\_data)  
#head(cor\_data)  
corrplot(cor\_data,method="shade",tl.col = "black",title = "\n\n Correlation Plot Of Store Data",tl.srt=65,cl.length = 5,cl.align="l",tl.cex=0.95, number.cex=0.8,col = colorRampPalette(c("yellow","lightblue","navyblue"))(100))

Chart

Description automatically generated

# V. Count Plot of Promotion Campaign Names

promotion\_data <- CFM\_data %>% select(promotion\_name) %>% group\_by(promotion\_name)  
promo\_count<-promotion\_data %>% group\_by(promotion\_name)   
ggplot(promo\_count, aes(x=promotion\_name),color=promotion\_name) + geom\_bar(fill = '#003366') + theme(axis.text.x = element\_text(angle = 90))

Chart, bar chart

Description automatically generated

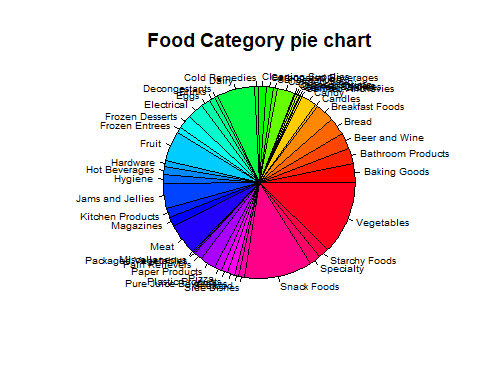
From the above graph, we observed that the highest sales promoting campaigns are:

**1. Weekend Markdown**

**2. Two Day sales**

**3. Price Slashers**

# VII. Pie Chart to Visualize the food\_category in dataset

food\_cat <- CFM\_data %>% select(food\_category) %>% group\_by(food\_category)%>% count()  
food\_cat<- arrange(food\_cat)  
piepercent<- round(100\*food\_cat$n/sum(food\_cat$n), 1)  
#pie(food\_cat$n,piepercent)  
  
  
  
pie(food\_cat$n, labels = food\_cat$food\_category, main = "Food Category pie chart",col = rainbow(length(food\_cat$n)), radius = 1, cex = 0.6)

# Observations:

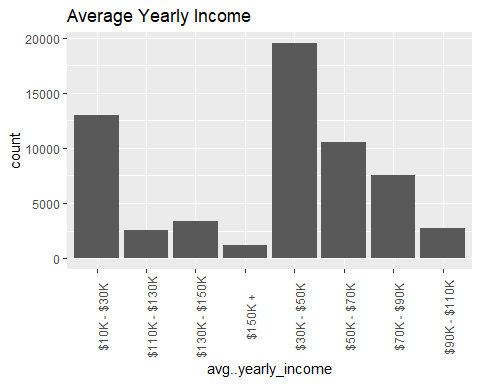
From the above pie chart, the most used food categories are shown to be:

* Vegetables
* Snack Foods
* Dairy

Which are products that are used on an everyday basis.

# VIII. Visualizing the Average yearly Income of customers

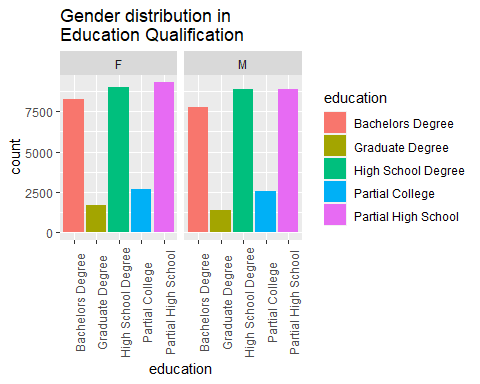
avg\_income\_count <- CFM\_data %>% select(avg..yearly\_income) %>% group\_by(avg..yearly\_income) %>% arrange(avg..yearly\_income)   
order <- c("$10K - $30K","$30K - $50K","$50K - $70K","$70K - $90K","$90K - $110K","$110K - $130K","$130K - $150K","$150K +")  
#avg\_income\_count  
ggplot(avg\_income\_count,aes(avg..yearly\_income)) + geom\_bar() + theme(axis.text.x = element\_text(angle = 90)) + ggtitle("Average Yearly Income")



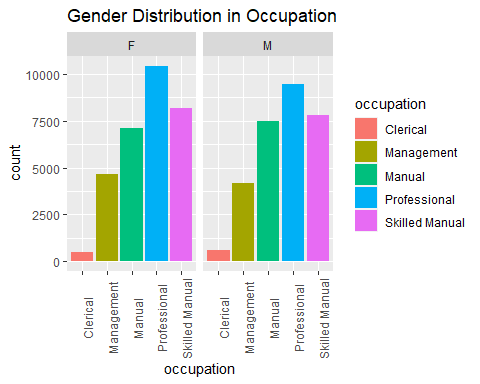
# Observations

We found out that the average yearly income of the customers is **$30k - $50k**.

# IX. Visualizing Education & Gender of customers in the dataset

Ed\_gender <- CFM\_data %>% select(gender,education) %>% arrange(gender) %>% group\_by(gender,education)  
#Ed\_gender %>% count()  
ggplot(Ed\_gender, aes(x=education,fill=education)) + geom\_bar()+ theme(axis.text.x = element\_text(angle = 90)) + facet\_wrap(~gender) + ggtitle(("Gender distribution in \nEducation Qualification"))

# XI. Visualizing occupation & gender of customers

Ocp\_gender <- CFM\_data %>% select(gender, occupation) %>% group\_by(gender,occupation)  
#Ocp\_gender %>% count()  
ggplot(Ocp\_gender, aes(x=occupation,fill=occupation)) + geom\_bar()+ theme(axis.text.x = element\_text(angle = 90)) + facet\_wrap(~gender) + ggtitle("Gender Distribution in Occupation")

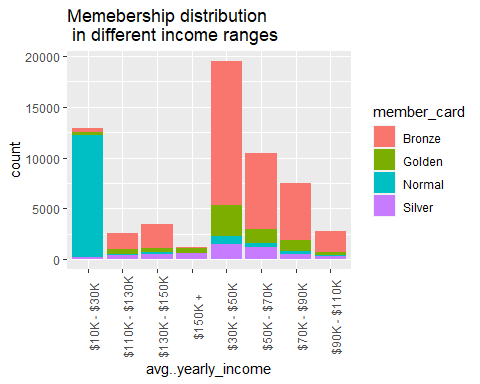
# Observations

* Most males have a partial high school degree.
* There are fewer females with graduate degrees.

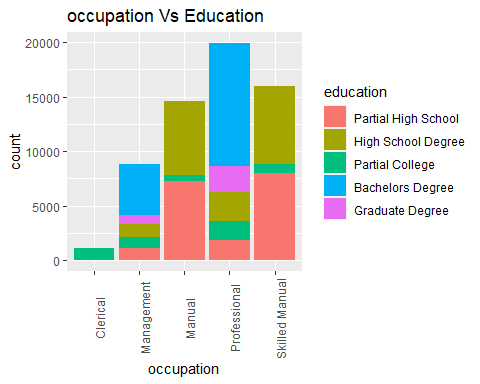
# XII. Visualizing the yearly income of Customers who have Food Mart Membership

Member card is the membership card that the customer owns.

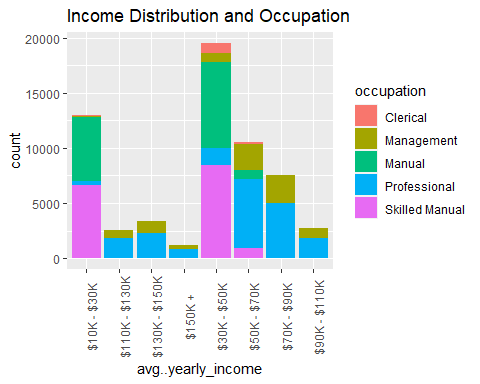
card\_income <- CFM\_data %>% select(member\_card,avg..yearly\_income) %>% group\_by(member\_card,avg..yearly\_income)  
#card\_income %>% count()  
ggplot(card\_income, aes(x=avg..yearly\_income,fill=member\_card)) + geom\_bar()+ theme(axis.text.x = element\_text(angle = 90)) + ggtitle("Memebership distribution \n in different income ranges")



# XIII. Visualize the education with occupation

ed\_oc <- CFM\_data %>% select(occupation,education) %>% group\_by(occupation,education)  
#ed\_oc %>% count()  
ggplot(ed\_oc, aes(x=occupation,fill=education)) + geom\_bar()+ theme(axis.text.x = element\_text(angle = 90)) + scale\_fill\_hue(limits = c("Partial High School","High School Degree","Partial College","Bachelors Degree","Graduate Degree")) + ggtitle("occupation Vs Education")

# XIV. Visualizing the average income against occupation of Customers

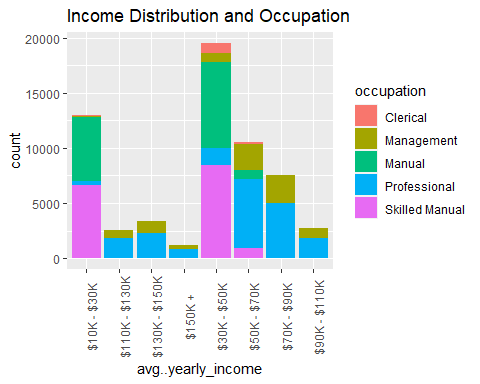
Income\_ocup <- CFM\_data %>% select(avg..yearly\_income,occupation) %>% group\_by(occupation,avg..yearly\_income)  
#Income\_ocup %>% count()  
ggplot(Income\_ocup,aes(x=avg..yearly\_income,fill=occupation)) + geom\_bar()+ theme(axis.text.x = element\_text(angle = 90)) + ggtitle("Income Distribution and Occupation")

# Observations

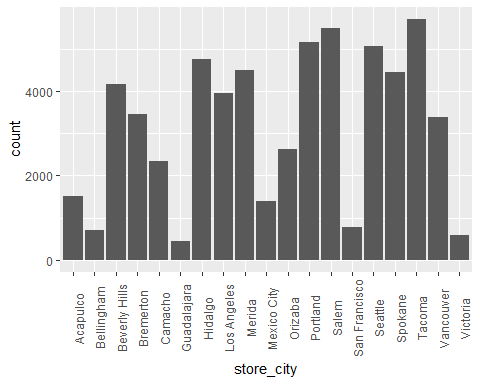
1. A Professional employee earns around $50k + on average.
2. Skilled manual employees and manual employees earn $30k +.
3. Management employees earn $30k +.
4. Clerical employees earn $30k +.

# XV. Visualizing the average yearly income with occupation of Customers

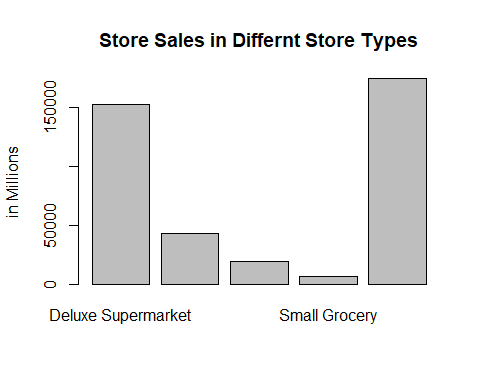
Income\_ocup <- CFM\_data %>% select(avg..yearly\_income,occupation) %>% group\_by(occupation,avg..yearly\_income)  
#Income\_ocup %>% count()  
ggplot(Income\_ocup,aes(x=avg..yearly\_income,fill=occupation)) + geom\_bar()+ theme(axis.text.x = element\_text(angle = 90)) + ggtitle("Income Distribution and Occupation")



# XVI. Visualizing store sales with the state the store is located.

City\_income <- CFM\_data %>% select(store\_city,store\_sales.in.millions.) %>% group\_by(store\_city,store\_sales.in.millions.)  
ggplot(City\_income, aes(x=store\_city),y=store\_sales.in.millions.) + geom\_bar()+ theme(axis.text.x = element\_text(angle = 90))

# XVII. Visualizing type of stores and their locations

sales\_type\_rev <- tapply(CFM\_data$store\_sales.in.millions., CFM\_data$store\_type, FUN=sum)  
#sales\_type\_rev  
p <- barplot(tapply(CFM\_data$store\_sales.in.millions., CFM\_data$store\_type, FUN=sum),ylab="in Millions",main = "Store Sales in Differnt Store Types")

# XVIII. Visualizing store sales in each state

Chart, histogram

Description automatically generatedsales\_city <- CFM\_data %>% select(store\_city,store\_sales.in.millions.)  
#sales\_city  
sales\_city\_rev <- tapply(CFM\_data$store\_sales.in.millions.,CFM\_data$store\_city,FUN=sum)  
#sales\_city\_rev  
barplot(sales\_city\_rev,las=2,ylab="in Millions",xlab="city",main="Revenue state-wise")

## XIX. Density Distribution Plots

A density plot represents how a numeric variable/feature is distributed across the data.They serve as a means to study the most common ranges in which the variable is occurring, along with observing data imbalance through representation of the concentration of each value. For our Food mart campaign cost analysis, density plots were generated with each consisting of FOUR main components/sub-plots:

* **Density**: Represented as a curvy, flowing filled aspect of the graph; drawn to represent probability distribution (or likelihood of occurrence)
* **Histogram**: Represented by Vertical bars, drawn to represent data distribution in a discrete form
* **Rug Plot**: Shown by the vertical bars at the floor of the graph; can be seen as a ‘lesser histogram’ that shows data distribution.
* **Mean Line**: Vertical dashed line; Cuts through the average value of the variable.

### i. Distribution of Cost

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=cost)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', binwidth = 1.5) +  
 geom\_density(position = 'identity',alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(cost)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Cost') +  
 theme(plot.title = element\_text(hjust = 0.5))

### ii. Distribution of Store Sales

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=store\_sales.in.millions.)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', binwidth = 1)+  
 geom\_density(position = 'identity',alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(store\_sales.in.millions.)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Store Sales (in millions)') +  
 theme(plot.title = element\_text(hjust = 0.5))

### iii. Distribution of Store Cost

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=store\_cost.in.millions.)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', binwidth = 0.5)+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(store\_cost.in.millions.)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Store Costs (in millions)') +  
 theme(plot.title = element\_text(hjust = 0.5))

### iv. Distribution of Product Unit Sales

Chart, bar chart

Description automatically generatedggplot(CFM\_data, aes(x=unit\_sales.in.millions.)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', bins = 20)+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(unit\_sales.in.millions.)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Product Sales (in millions)') +  
 theme(plot.title = element\_text(hjust = 0.5))

### v. Distribution of Total children

Chart, bar chart

Description automatically generatedggplot(CFM\_data, aes(x=total\_children)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='yellow', binwidth=0.5)+  
 geom\_density(position = 'identity',alpha=0.5, fill='red') +  
 geom\_vline(aes(xintercept=mean(total\_children)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Total Attending Children') +  
 theme(plot.title = element\_text(hjust = 0.5))

### vi. Distribution of Average Household owned Cars

Chart, bar chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=avg\_cars\_at.home.approx.)) +  
 geom\_histogram(aes(y=..density..), color = 'light blue', fill='blue', binwidth=0.4)+  
 geom\_density(alpha=0.5, fill='magenta') +  
 geom\_vline(aes(xintercept=mean(avg\_cars\_at.home.approx.)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Average number of cars owned at a household') +  
 theme(plot.title = element\_text(hjust = 0.5))

### vii. Distribution of the Number of children per family

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=num\_children\_at\_home)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='yellow', bins=25)+  
 geom\_density(alpha=0.5, fill='orange') +  
 geom\_vline(aes(xintercept=mean(num\_children\_at\_home)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Number of children in a household') +  
 theme(plot.title = element\_text(hjust = 0.5))

### viii. Distribution of the Number of Suggested Retail Price (SRP)

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=SRP)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', binwidth=0.25)+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(SRP)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Suggested Retail Price') +  
 theme(plot.title = element\_text(hjust = 0.5))

### ix. Distribution of Gross Weight

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=gross\_weight)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='orange', bins=25)+  
 geom\_density(alpha=0.5, fill='light blue') +  
 geom\_vline(aes(xintercept=mean(gross\_weight)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Gross Weight') +  
 theme(plot.title = element\_text(hjust = 0.5))

### x. Distribution of Net Weight

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=net\_weight)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='orange', binwidth=0.5)+  
 geom\_density(alpha=0.5, fill='light blue') +  
 geom\_vline(aes(xintercept=mean(net\_weight)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Net Weight') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xi. Distribution of recyclable/non-recyclable products

This feature is categorical, representing whether is product is recyclable (1) or not (0).

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=recyclable\_package)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', binwidth=0.25)+  
 geom\_density(alpha=0.75, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(recyclable\_package)), color="blue",  
 linetype="longdash") + geom\_rug() + ggtitle('Data Distribution vs Probability Density Distribution of Recyclable/Non-Recyclable Package') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xii. Distribution of Low fat/non-low-fat products

A categorical variable represented in binary format (1 and 0).

ggplot(CFM\_data, aes(x=low\_fat)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', binwidth = 0.25)+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(low\_fat)), color="blue",  
 linetype="longdash") + geom\_rug() +  
 ggtitle('Data Distribution vs Probability Density Distribution of low\_fat') +  
 theme(plot.title = element\_text(hjust = 0.5))

Chart, histogram

Description automatically generated

### xiii. Distribution of Units per Case

ggplot(CFM\_data, aes(x=units\_per\_case)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', binwidth=0.75)+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(units\_per\_case)), color="blue",  
 linetype="longdash") + geom\_rug() +  
 ggtitle('Data Distribution vs Probability Density Distribution of units\_per\_case') +  
 theme(plot.title = element\_text(hjust = 0.5))

Chart, histogram

Description automatically generated

### xiv. Distribution of Total Store Areas (sq. ft)

ggplot(CFM\_data, aes(x=store\_sqft)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green')+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(store\_sqft)), color="blue",  
 linetype="longdash") + geom\_rug() +  
 ggtitle('Data Distribution vs Probability Density Distribution of store\_sqft') +  
 theme(plot.title = element\_text(hjust = 0.5))

Chart, histogram

Description automatically generated

### xv. Distribution of Area of each Store’s grocery section

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=grocery\_sqft)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green')+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(grocery\_sqft)), color="blue",  
 linetype="longdash") + geom\_rug() +  
 ggtitle('Data Distribution vs Probability Density Distribution of grocery\_sqft') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xvi. Distribution of Area of each store’s Frozen Section

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=frozen\_sqft)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green')+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(frozen\_sqft)), color="blue",  
 linetype="longdash") + geom\_rug() +  
 ggtitle('Data Distribution vs Probability Density Distribution of frozen\_sqft') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xvii. Distribution of Area of each store’s Meat Section

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=meat\_sqft)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green', bins=25)+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(meat\_sqft)), color="blue",  
 linetype="longdash") + geom\_rug() +  
 ggtitle('Data Distribution vs Probability Density Distribution of meat\_sqft') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xviii. Distribution of Coffee Bars availability

Represented as a categorical, binary variable. (1 if available, 0 if not)

Chart, histogram

Description automatically generatedggplot(CFM\_data, aes(x=coffee\_bar)) +   
 geom\_histogram(aes(y=..density..), color = 'red', fill='green') +  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(coffee\_bar)), color="blue",linetype="longdash") +  
 ggtitle('Data Distribution vs Probability Density Distribution of coffee\_bar') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xix. Distribution of Availability of video stores

Represented as a categorical variable.

Graphical user interface

Description automatically generatedggplot(CFM\_data, aes(x=video\_store)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green')+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(video\_store)), color="blue",linetype="longdash") +  
 ggtitle('Data Distribution vs Probability Density Distribution of video\_store') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xx. Distribution of Availability of Salad bar

Represented as categorical variable.

Chart

Description automatically generatedggplot(CFM\_data, aes(x=salad\_bar)) +  
 geom\_histogram(aes(y=..density..), color = 'red', fill='green')+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(salad\_bar)), color="blue",linetype="longdash")+  
 ggtitle('Data Distribution vs Probability Density Distribution of salad\_bar') +  
 theme(plot.title = element\_text(hjust = 0.5))

### xxi. Distribution of Availability of Prepared Food

Chart

Description automatically generatedggplot(CFM\_data, aes(x=prepared\_food)) +   
 geom\_histogram(aes(y=..density..), color = 'red', fill='green')+  
 geom\_density(alpha=0.5, fill='dark green') +  
 geom\_vline(aes(xintercept=mean(prepared\_food)), color="blue", linetype="longdash") +  
 ggtitle('Data Distribution vs Probability Density Distribution of prepared\_food') +  
 theme(plot.title = element\_text(hjust = 0.5))

## XX. Plotting relationships between Variable Pairs

As in most real-world scenarios, variables usually have some influence or correlation with one another in pairs, which can be represented through various visual plots.

### i. Store sales and Store cost

A violin plot is another way to showcase probability distribution. The thicker the sides of the plot, the more likely the variable will be a specific value.

Chart, surface chart

Description automatically generated ggplot(CFM\_data, aes(x=store\_sales.in.millions., y=store\_cost.in.millions.))+  
 ggtitle("Violin plot for store\_sales.in.millions and store\_cost.in.millions") + theme(plot.title = element\_text(hjust = 0.5)) +  
 geom\_violin(color='black' , fill = 'blue')

### ii. Meat Section area and Frozen Section Area

An almost perfect positive relationship is shown on the plot.

**Possible Explanation**: Prepared meat is usually stored frozen and is therefore a subset of frozen foods, possibly explaining the correlation.

dataB <- CFM\_data[, c( "meat\_sqft", "frozen\_sqft")]

Chart, scatter chart

Description automatically generated# Add aesthetic mappings  
p + geom\_point(colour = "red", size = 3) + ggtitle("Relationship between meat area and frozen area in the mart")

### iii. Total Store area and Grocery Section Area

A semi-strong positive relationship is indicated by the plot.

**Possible Explanation**: A bigger store can allow for a bigger grocery section, which is usually the most popular aspect of supermarkets, and food mart by extension.

Chart, scatter chart

Description automatically generateddataC <- CFM\_data[, c( "store\_sqft", "grocery\_sqft")]  
q <- ggplot(dataC, aes(store\_sqft, grocery\_sqft))  
q + geom\_point(colour = "red", size = 3) + ggtitle("Relationship between store area and grocery area in the mart")

### iii. Gross Weight and Net Weight

The plot indicates a strong positive, nearly directly proportional relationship.

**Possible Explanation**: Since gross weight is the net weight of the shipped products + weight of packaging, a directly proportional relationship makes sense.

Chart, line chart

Description automatically generateddataD <- CFM\_data[, c( "gross\_weight", "net\_weight")]  
q <- ggplot(dataD, aes(gross\_weight, net\_weight))  
q + geom\_point(colour = "red", size = 3) + ggtitle("Relationship between gross weight and net weight in the mart")

### iv. Gross Weight and Units per Case

The scattered points indicate no relationship.

**Possible Explanation**: Some products like TVs, large electronics, etc. could only consist of few units that weight heavily, and vice versa.

Chart, scatter chart

Description automatically generateddataE <- CFM\_data[, c( "gross\_weight", "units\_per\_case")]  
q <- ggplot(dataE, aes(gross\_weight, units\_per\_case))  
q + geom\_point(colour = "red", size = 3) + ggtitle("Relationship between gross weight and units per case in the mart")