Super market Sales Analysis

Ibeh Amaka

2023-10-03

A company named XYZ wants to know how their company is fairing using their collected sales between 1/1/2020 and 31/12/2020.

Scenario: Performance and profitability analysis for XYZ About the dataset The dataset consists of 9994 observations and 21 variables consisting of Order ID, Ship Date, Order Date, Ship mode, Customer ID, Segment, Country, City, Sales, Region, Product ID, Category, Sub-category, Sales, Quantity, Discount, Profit. This company is located in the USA and has branches in all regions and states.

use this code to clean the global environment

load all the library needed First install the packages needed

# install.packages("readr")  
# install.packages("tidyverse")  
# install.packages("tidyr")  
# install.packages("ggplot2")  
# install.packages("dylyer")  
# install.packages("data.table")  
#install.packages("knitr")  
#install.packages('forecast', dependencies = TRUE)

call the library

library(readr)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.4.3 v dplyr 1.1.2  
## v tibble 3.2.1 v stringr 1.5.0  
## v tidyr 1.3.0 v forcats 0.5.1  
## v purrr 1.0.1

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'purrr' was built under R version 4.1.3

## Warning: package 'dplyr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(tidyr)  
library(ggplot2)  
library(dplyr)  
library(data.table)

## Warning: package 'data.table' was built under R version 4.1.3

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

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

## The following object is masked from 'package:purrr':  
##   
## transpose

library(knitr)  
library(forecast)

## Warning: package 'forecast' was built under R version 4.1.3

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

library(tseries)

## Warning: package 'tseries' was built under R version 4.1.3

Import the data

store <- read\_csv("Data/Sample - Superstore.csv")

## Rows: 9994 Columns: 21  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (16): Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer ...  
## dbl (5): Row ID, Sales, Quantity, Discount, Profit  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Lets check for missing values

missing\_data = store[!complete.cases(store),]

there are no missing data in this dataset lets check the structure of the data

missing\_data = store[!complete.cases(store),]

The order date and ship date are in character data type so i will chnage it to date Sample data with mixed date formats

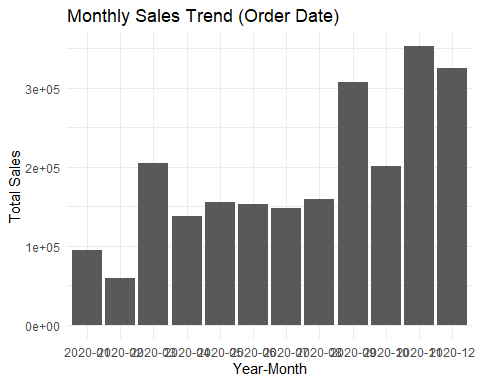
store$`Ship Date` <- as.Date(store$`Ship Date`,format = "%m/%d/%y")  
store$`Order Date` <- as.Date(store$`Order Date`,format = "%m/%d/%y")

SALES TRENDS: Group and summarize sales data by month

monthly\_sales <- store %>%  
 group\_by(YearMonth = format(`Order Date`, "%Y-%m")) %>%  
 summarise(TotalSales = sum(Sales))

Plot the sales trend

ggplot(data = monthly\_sales, aes(x = YearMonth, y = TotalSales)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Monthly Sales Trend (Order Date)",  
 x = "Year-Month",  
 y = "Total Sales") +  
 theme\_minimal()

 There are more sales in the last quater of the year (September till December). This can be attributed to the festive seasons and several holidays at the end of the year.

question 2 Product or Service Performance: Which products or services are driving the majority of the sales?

cat\_sales <- store %>%  
 group\_by (Category) %>%  
 summarise(cat\_sales = sum(Sales)) %>%  
 arrange(desc(cat\_sales))

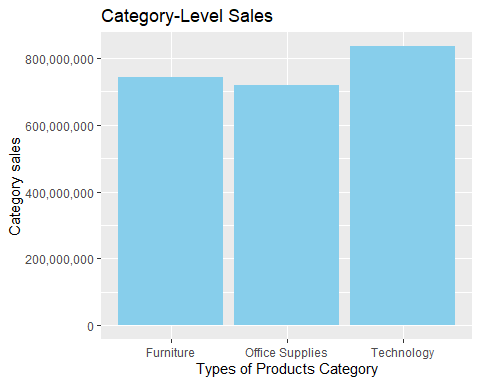
we can see that Technology has the highest sales,then furniture then office supplies.

let see for sub category

subcat\_sales <- store %>%  
 group\_by (`Sub-Category`) %>%  
 summarise(cat\_sales = sum(Sales)) %>%  
 arrange(desc(cat\_sales))

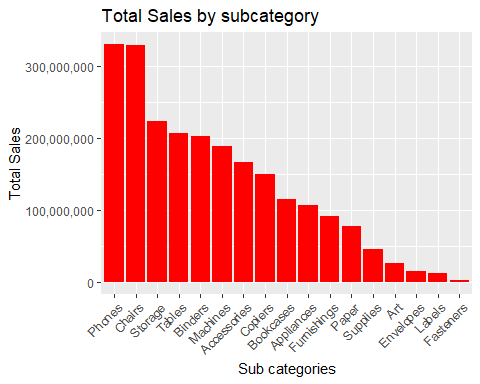
Create a plot

ggplot(cat\_sales, aes(x = Category, y = cat\_sales)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Category-Level Sales", y = "Category sales", x = "Types of Products Category") +   
 scale\_y\_continuous(labels = scales::comma\_format(scale = 1e+3))



I can not categorically say that furniture, Office supplies and Technology are underperforming. This is because the sales difference is less.

ggplot(subcat\_sales , aes(x = reorder(`Sub-Category`, -cat\_sales), y = cat\_sales)) +  
 geom\_bar(stat = "identity", fill = "red") +  
 labs(title = "Total Sales by subcategory", x = "Sub categories", y = "Total Sales") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +   
 scale\_y\_continuous(labels = scales::comma\_format(scale = 1e+3))



we can see phone, chairs and storage has the highest sales. By sub categories we can see that phones, chairs and storage are doing good with numbers. Envelopes, lebels and fastners have the least sales. however, if we group them by categories they are doing quit good.

question 3 customer analysis Who are the company’s top customers in terms of sales?

top\_customers\_by\_city <- store %>%  
 group\_by(City, `Customer ID`,) %>%  
 summarise(TotalSales = sum(Sales)) %>%  
 arrange(City, desc(TotalSales)) %>%  
 group\_by(City) %>%  
 top\_n(10, wt = TotalSales)

## `summarise()` has grouped output by 'City'. You can override using the  
## `.groups` argument.

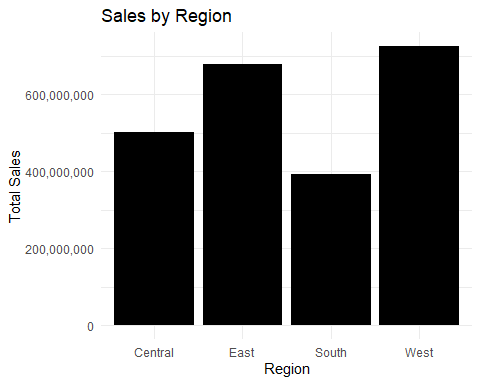
We can now see the IDs of customers, the company can send good will messages to them during birthdays, festive seasons if their email address is available. Discount and promotions can also be sent to them

Geographic Analysis: Are there regional variations in sales performance?

region\_sales <- store %>%   
 group\_by(Region) %>%   
 summarise(region\_sales = sum(Sales)) %>%   
 arrange(desc(region\_sales))

Plot

ggplot(region\_sales, aes(x = Region, y = region\_sales)) +  
 geom\_bar(stat = "identity", fill = "black") +  
 labs(title = "Sales by Region", y = "Total Sales", x = "Region") +  
 theme\_minimal() +   
 scale\_y\_continuous(labels = scales::comma\_format(scale = 1e+3))



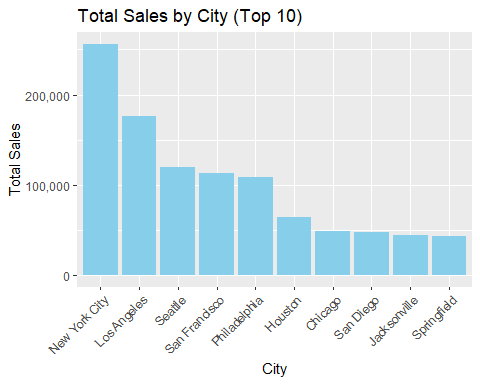
we see that the west has more sales, then the east, central and south. the west and east are contributing to more sales than the other regions

Lets explore the datase more

what is the state with the highest sales

city\_sales<- store %>%  
 group\_by(City)%>%  
 summarise(total\_sales = sum(Sales))%>%  
 arrange(desc(total\_sales)) %>%  
 slice(1:10)

ggplot(city\_sales, aes(x = reorder(City, -total\_sales), y = total\_sales)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Total Sales by City (Top 10)", x = "City", y = "Total Sales") +  
 scale\_y\_continuous(  
 labels = scales::comma\_format(accuracy = 1) # Format labels without scientific notation  
 ) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))#otate x-axis labels for better readability



The result shows that new york city, los angeles, seattle, san francisco, Philadelphia has the highest sales

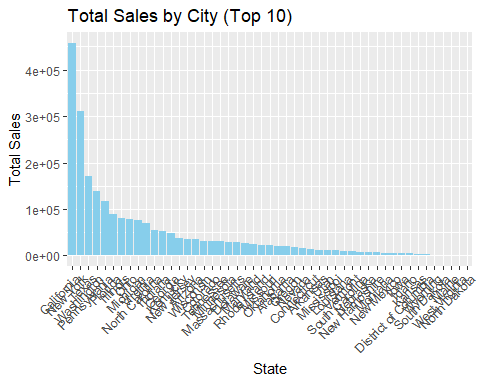
lets see the state with the highest sales

state\_sales<- store %>%  
 group\_by(State)%>%  
 summarise(total\_sales = sum(Sales)) %>%  
 arrange(desc(total\_sales))   
state\_sales

## # A tibble: 49 x 2  
## State total\_sales  
## <chr> <dbl>  
## 1 California 457688.  
## 2 New York 310876.  
## 3 Texas 170188.  
## 4 Washington 138641.  
## 5 Pennsylvania 116512.  
## 6 Florida 89474.  
## 7 Illinois 80166.  
## 8 Ohio 78258.  
## 9 Michigan 76270.  
## 10 Virginia 70637.  
## # i 39 more rows

The result shows that California , New York, Washington, and Pennsylvania are the states with the highest sales

ggplot(state\_sales, aes(x = reorder(State, -total\_sales), y = total\_sales)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Total Sales by City (Top 10)", x = "State", y = "Total Sales") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

 we can see that the state of califonial has the highest sales

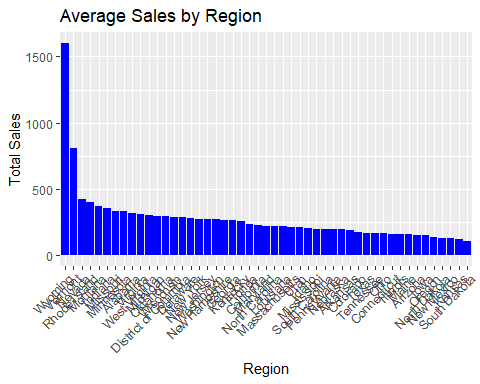
lets look at the average sales

state\_avg\_sales <- store %>%  
 group\_by(State) %>%  
 summarise(average\_sales = mean(Sales)) %>%  
 arrange(desc(average\_sales))   
state\_avg\_sales

## # A tibble: 49 x 2  
## State average\_sales  
## <chr> <dbl>  
## 1 Wyoming 1603.  
## 2 Vermont 812.  
## 3 Nevada 429.  
## 4 Rhode Island 404.  
## 5 Montana 373.  
## 6 Indiana 359.  
## 7 Missouri 336.  
## 8 Minnesota 336.  
## 9 Alabama 320.  
## 10 Virginia 315.  
## # i 39 more rows

Results shows that Wyoming, Vermont, Navada and Rohade island has the highest average sales receptively. lets see a plot of it

ggplot(state\_avg\_sales, aes(x = reorder(State, -average\_sales), y = average\_sales)) +  
 geom\_bar(stat = "identity", fill = "blue") +  
 labs(title = "Average Sales by Region", x = "Region", y = "Total Sales") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



we can see that the state of califonaia (457687.632)has the highest sales, followed by newyork(310876.271) the texas (170188.046), however, they do not have the highest avergae sales. this can be due to several reasons e.g in the number of transactions (or customers) and the distribution of sales within the state.

lets check the number of customers in califonia and newyork….and also the quantity sold. then compare it to Wyoming and Vermont

california\_customers <- store %>%  
 filter(State == "California") %>%  
 summarise(total\_customers = n())  
california\_customers

## # A tibble: 1 x 1  
## total\_customers  
## <int>  
## 1 2001

newyork\_customer <- store %>%  
 filter(State == "New York") %>%  
 summarise(total\_customers = n())  
newyork\_customer

## # A tibble: 1 x 1  
## total\_customers  
## <int>  
## 1 1128

Wyoming\_customer <- store %>%  
 filter(State == "Wyoming") %>%  
 summarise(total\_customers = n())  
Wyoming\_customer

## # A tibble: 1 x 1  
## total\_customers  
## <int>  
## 1 1

Vermont\_customer <- store %>%  
 filter(State == "Wyoming") %>%  
 summarise(total\_customers = n())  
Vermont\_customer

## # A tibble: 1 x 1  
## total\_customers  
## <int>  
## 1 1

we can see the reason why the average sales is high for this group is because of the number of customers’. between newyork and califonia we can see that the difference between the customers are 873 new york has less customers but the average sales is greater several reasons can be involved this include Higher Income Levels, Tourism and Business Travel etc

lets see the average profit for newyork and califonia

state\_avg\_profit <- store %>%  
 group\_by(State) %>%  
 summarise(average\_sales = mean(Profit)) %>%  
 arrange(desc(average\_sales))  
state\_avg\_profit

## # A tibble: 49 x 2  
## State average\_sales  
## <chr> <dbl>  
## 1 Vermont 204.   
## 2 Rhode Island 130.   
## 3 Indiana 123.   
## 4 Montana 122.   
## 5 Minnesota 122.   
## 6 District of Columbia 106.   
## 7 Delaware 104.   
## 8 Wyoming 100.   
## 9 Missouri 97.5  
## 10 Michigan 95.9  
## # i 39 more rows

average profit for new york is 65.637011 and California is 38.171608

lets see the quantity sold in each state

state\_qunt <- store %>%  
 group\_by(State) %>%  
 summarise(quantity = sum(Quantity)) %>%  
 arrange(desc(quantity)) %>%  
 slice(1:10)  
state\_qunt

## # A tibble: 10 x 2  
## State quantity  
## <chr> <dbl>  
## 1 California 7667  
## 2 New York 4224  
## 3 Texas 3724  
## 4 Pennsylvania 2153  
## 5 Washington 1883  
## 6 Illinois 1845  
## 7 Ohio 1759  
## 8 Florida 1379  
## 9 North Carolina 983  
## 10 Michigan 946

see that califonia has the highest quantity with (7667) sold and new York has (4224) yet new york has more average sales. Reasons are discount, ’Higher Income Levels,Tourism and Business Travel etc.

since we have discount in our dataset lets see which state gives more discount

discount\_sales <- store %>%  
 group\_by(State) %>%  
 summarise(discount = mean(Discount)) %>%  
 arrange(desc(discount)) %>%  
 slice(1:10)  
discount\_sales

## # A tibble: 10 x 2  
## State discount  
## <chr> <dbl>  
## 1 Illinois 0.390  
## 2 Texas 0.370  
## 3 Pennsylvania 0.329  
## 4 Ohio 0.325  
## 5 Colorado 0.316  
## 6 Arizona 0.304  
## 7 Florida 0.299  
## 8 Tennessee 0.291  
## 9 Oregon 0.289  
## 10 North Carolina 0.284

we can see that califonia has a sum discount of 7.2% while newyork is 5.5%##we can see that califonia gives more discounts than new york yet aver sales for newyork is higher

lets look at other factors, lets see the GDP of new york Geographical wise, califonia is bigger than newyork with more population this explains the higher number of sales, when we look at the GDP of califonia it is stilll biger than new york, however, the GDP per capital of new york is higher due to high purchasing power. People there has the capability to buy more expensive goods.

lets see the maximum profit and minimum profit

max(store$Profit) ###8399.976

## [1] 8399.976

min(store$Profit) ###-6599.978

## [1] -6599.978

##lets see the total profit  
sum(store$Profit)

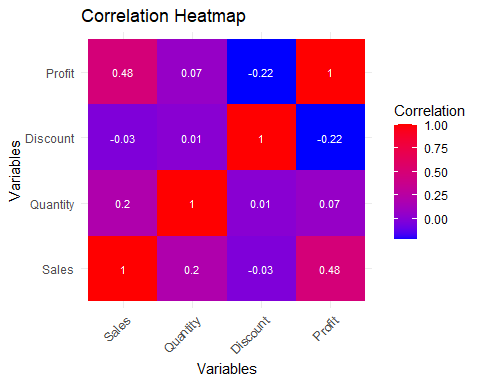
## [1] 286397

sum of profit is 286397….The total profit is too low because some of the company s stores are making losses. lets run a corrolation analysis because i want to see if discount is corrolated with the profit,quantity, discount and sales

correlation\_matrix <- store %>%  
 select(Sales, Quantity, Discount,Profit) %>%  
 cor()  
correlation\_matrix

## Sales Quantity Discount Profit  
## Sales 1.00000000 0.20079477 -0.02819012 0.47906435  
## Quantity 0.20079477 1.00000000 0.00862297 0.06625319  
## Discount -0.02819012 0.00862297 1.00000000 -0.21948746  
## Profit 0.47906435 0.06625319 -0.21948746 1.00000000

heatmap\_plot <- ggplot(data = reshape2::melt(correlation\_matrix),   
 aes(x = Var1, y = Var2, fill = value, label = round(value, 2))) +  
 geom\_tile() +  
 geom\_text(size = 3, color = "white", show.legend = FALSE) +  
 scale\_fill\_gradient(low = "blue", high = "red") +  
 labs(title = "Correlation Heatmap",  
 x = "Variables", y = "Variables",  
 fill = "Correlation") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
heatmap\_plot

 results shows that profit and sales have a positive and meduim corrolation, there is no influnce of quantity and discount on sales. This business seams regid and will suffer losses. This is reflected by the margnal negative impact of discount on profit.

I cahnged the structure of the date hoping to perform a time series analysis for forcasting future sales, however this will not be possible because the dat has only sales for one year.