ISE\_201\_project\_superstore1

Sai Prasanna Kumar

2022-10-07

# Introduction:

#### Dataset Description

The dataset was curated by a Superstore Giant to understand which products, regions, categories and customer segments they should target or avoid. It contains sales & profits of an US superstore located at different geographical locations. I was curious about the how we can use the data to help understand market sales & profits varies with discounts and what strategies need to be build to attract customers across different regions in the store.

#### About the dataset:

* Data source: Kaggle (<https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>)
* Data Collection : Sample dataset collected by one of the superstore giant located across in US which contains 9994 samples in the dataset.
* Variables: Dataset contains 21 attributes with mix of numeric and categorical variables
  + **Row ID** => Unique ID for each row.
  + **Order ID** => Unique Order ID for each Customer.
  + **Order Date** => Order Date of the product.
  + **Ship Date** => Shipping Date of the Product.
  + **Ship Mode** => Shipping Mode specified by the Customer.
  + **Customer ID** => Unique ID to identify each Customer.
  + **Customer Name** => Name of the Customer.
  + **Segment** => The segment where the Customer belongs.
  + **Country** => Country of residence of the Customer.
  + **City** => City of residence of of the Customer.
  + **State** => State of residence of the Customer.
  + **Postal Code** => Postal Code of every Customer.
  + **Region** => Region where the Customer belong.
  + **Product ID** => Unique ID of the Product.
  + **Category** => Category of the product ordered.
  + **Sub-Category** => Sub-Category of the product ordered.
  + **Product Name** => Name of the Product
  + **Sales** => Sales of the Product.
  + **Quantity** => Quantity of the Product.
  + **Discount** => Discount provided.
  + **Profit** => Profit/Loss incurred.

#### Cases

* This is an Observational study to understand how profits of an superstore varies with individual product sales
* Each row represent an order made by a customer for a particular product along with sales and profit made by superstore.

#### Proposal on what questions you are interested in answering from the data?:

1. What’s the best sales season for the store?
2. What are the most profitable categories/sub categories?
3. Geographical analysis of sales and profit.
4. Discounts attract customers and increase profit sales?
5. Which state produces highest profit sales?
6. How long the items get shipped since the day we order?

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.2

## ── 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(dplyr)  
library(ggplot2)  
library(corrplot)

## corrplot 0.92 loaded

library(patchwork)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

data\_df <- read.csv('C:/Users/Checkout/Desktop/SJSU/sem1/201-ISE/project/superstore/Superstore.csv')  
head(data\_df)

## Row.ID Order.ID Order.Date Ship.Date Ship.Mode Customer.ID  
## 1 1 CA-2016-152156 11/8/2016 11/11/2016 Second Class CG-12520  
## 2 2 CA-2016-152156 11/8/2016 11/11/2016 Second Class CG-12520  
## 3 3 CA-2016-138688 6/12/2016 6/16/2016 Second Class DV-13045  
## 4 4 US-2015-108966 10/11/2015 10/18/2015 Standard Class SO-20335  
## 5 5 US-2015-108966 10/11/2015 10/18/2015 Standard Class SO-20335  
## 6 6 CA-2014-115812 6/9/2014 6/14/2014 Standard Class BH-11710  
## Customer.Name Segment Country City State  
## 1 Claire Gute Consumer United States Henderson Kentucky  
## 2 Claire Gute Consumer United States Henderson Kentucky  
## 3 Darrin Van Huff Corporate United States Los Angeles California  
## 4 Sean O'Donnell Consumer United States Fort Lauderdale Florida  
## 5 Sean O'Donnell Consumer United States Fort Lauderdale Florida  
## 6 Brosina Hoffman Consumer United States Los Angeles California  
## Postal.Code Region Product.ID Category Sub.Category  
## 1 42420 South FUR-BO-10001798 Furniture Bookcases  
## 2 42420 South FUR-CH-10000454 Furniture Chairs  
## 3 90036 West OFF-LA-10000240 Office Supplies Labels  
## 4 33311 South FUR-TA-10000577 Furniture Tables  
## 5 33311 South OFF-ST-10000760 Office Supplies Storage  
## 6 90032 West FUR-FU-10001487 Furniture Furnishings  
## Product.Name Sales  
## 1 Bush Somerset Collection Bookcase 261.9600  
## 2 Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.9400  
## 3 Self-Adhesive Address Labels for Typewriters by Universal 14.6200  
## 4 Bretford CR4500 Series Slim Rectangular Table 957.5775  
## 5 Eldon Fold 'N Roll Cart System 22.3680  
## 6 Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood 48.8600  
## Quantity Discount Profit  
## 1 2 0.00 41.9136  
## 2 3 0.00 219.5820  
## 3 2 0.00 6.8714  
## 4 5 0.45 -383.0310  
## 5 2 0.20 2.5164  
## 6 7 0.00 14.1694

### Data Cleaning

#### Data Quality Checks

# removing redundant columns  
data\_df[,c("Row.ID","Order.ID","Product.ID", "Customer.Name","Customer.ID")] <- list(NULL)  
colnames(data\_df)

## [1] "Order.Date" "Ship.Date" "Ship.Mode" "Segment" "Country"   
## [6] "City" "State" "Postal.Code" "Region" "Category"   
## [11] "Sub.Category" "Product.Name" "Sales" "Quantity" "Discount"   
## [16] "Profit"

data\_df <- within(data\_df, {   
 profit\_cat <- NA # need to initialize variable  
 profit\_cat[Profit > 0 ] <- TRUE  
 profit\_cat[Profit <0 ] <- FALSE  
 } )  
  
colSums(is.na(data\_df))

## Order.Date Ship.Date Ship.Mode Segment Country City   
## 0 0 0 0 0 0   
## State Postal.Code Region Category Sub.Category Product.Name   
## 0 0 0 0 0 0   
## Sales Quantity Discount Profit profit\_cat   
## 0 0 0 0 65

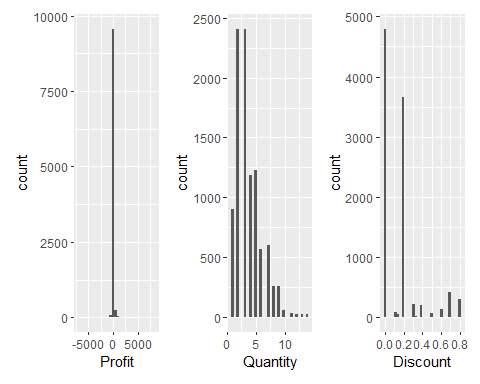
# colnames(data\_df)  
# Duplicates check  
sum(duplicated(data\_df))

## [1] 1

* From the above we can see that there No Missing values but there is one duplicate row in the dataset.

p1 <- ggplot(data\_df, aes(Profit),bins = 10) + geom\_histogram()  
p2 <- ggplot(data\_df, aes(Quantity,bins = 20)) + geom\_histogram()  
p3 <- ggplot(data\_df, aes(Discount,bins = 15)) + geom\_histogram()  
  
p1+p2+p3

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



sapply(data\_df, class)

## Order.Date Ship.Date Ship.Mode Segment Country City   
## "character" "character" "character" "character" "character" "character"   
## State Postal.Code Region Category Sub.Category Product.Name   
## "character" "integer" "character" "character" "character" "character"   
## Sales Quantity Discount Profit profit\_cat   
## "numeric" "integer" "numeric" "numeric" "logical"

# hist(strtoi(data\_df$Profit))  
summary(data\_df)

## Order.Date Ship.Date Ship.Mode Segment   
## Length:9994 Length:9994 Length:9994 Length:9994   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Country City State Postal.Code   
## Length:9994 Length:9994 Length:9994 Min. : 1040   
## Class :character Class :character Class :character 1st Qu.:23223   
## Mode :character Mode :character Mode :character Median :56431   
## Mean :55190   
## 3rd Qu.:90008   
## Max. :99301   
## Region Category Sub.Category Product.Name   
## Length:9994 Length:9994 Length:9994 Length:9994   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Sales Quantity Discount Profit   
## Min. : 0.444 Min. : 1.00 Min. :0.0000 Min. :-6599.978   
## 1st Qu.: 17.280 1st Qu.: 2.00 1st Qu.:0.0000 1st Qu.: 1.729   
## Median : 54.490 Median : 3.00 Median :0.2000 Median : 8.666   
## Mean : 229.858 Mean : 3.79 Mean :0.1562 Mean : 28.657   
## 3rd Qu.: 209.940 3rd Qu.: 5.00 3rd Qu.:0.2000 3rd Qu.: 29.364   
## Max. :22638.480 Max. :14.00 Max. :0.8000 Max. : 8399.976   
## profit\_cat   
## Mode :logical   
## FALSE:1871   
## TRUE :8058   
## NA's :65   
##   
##

## EDA - Exploratory Data Analysis

head(data\_df)

## Order.Date Ship.Date Ship.Mode Segment Country City  
## 1 11/8/2016 11/11/2016 Second Class Consumer United States Henderson  
## 2 11/8/2016 11/11/2016 Second Class Consumer United States Henderson  
## 3 6/12/2016 6/16/2016 Second Class Corporate United States Los Angeles  
## 4 10/11/2015 10/18/2015 Standard Class Consumer United States Fort Lauderdale  
## 5 10/11/2015 10/18/2015 Standard Class Consumer United States Fort Lauderdale  
## 6 6/9/2014 6/14/2014 Standard Class Consumer United States Los Angeles  
## State Postal.Code Region Category Sub.Category  
## 1 Kentucky 42420 South Furniture Bookcases  
## 2 Kentucky 42420 South Furniture Chairs  
## 3 California 90036 West Office Supplies Labels  
## 4 Florida 33311 South Furniture Tables  
## 5 Florida 33311 South Office Supplies Storage  
## 6 California 90032 West Furniture Furnishings  
## Product.Name Sales  
## 1 Bush Somerset Collection Bookcase 261.9600  
## 2 Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.9400  
## 3 Self-Adhesive Address Labels for Typewriters by Universal 14.6200  
## 4 Bretford CR4500 Series Slim Rectangular Table 957.5775  
## 5 Eldon Fold 'N Roll Cart System 22.3680  
## 6 Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood 48.8600  
## Quantity Discount Profit profit\_cat  
## 1 2 0.00 41.9136 TRUE  
## 2 3 0.00 219.5820 TRUE  
## 3 2 0.00 6.8714 TRUE  
## 4 5 0.45 -383.0310 FALSE  
## 5 2 0.20 2.5164 TRUE  
## 6 7 0.00 14.1694 TRUE

# UNIQUE CATEGORIES  
print(unique(data\_df$State))

## [1] "Kentucky" "California" "Florida"   
## [4] "North Carolina" "Washington" "Texas"   
## [7] "Wisconsin" "Utah" "Nebraska"   
## [10] "Pennsylvania" "Illinois" "Minnesota"   
## [13] "Michigan" "Delaware" "Indiana"   
## [16] "New York" "Arizona" "Virginia"   
## [19] "Tennessee" "Alabama" "South Carolina"   
## [22] "Oregon" "Colorado" "Iowa"   
## [25] "Ohio" "Missouri" "Oklahoma"   
## [28] "New Mexico" "Louisiana" "Connecticut"   
## [31] "New Jersey" "Massachusetts" "Georgia"   
## [34] "Nevada" "Rhode Island" "Mississippi"   
## [37] "Arkansas" "Montana" "New Hampshire"   
## [40] "Maryland" "District of Columbia" "Kansas"   
## [43] "Vermont" "Maine" "South Dakota"   
## [46] "Idaho" "North Dakota" "Wyoming"   
## [49] "West Virginia"

print(unique(data\_df$Region))

## [1] "South" "West" "Central" "East"

print(unique(data\_df$Category))

## [1] "Furniture" "Office Supplies" "Technology"

print(unique(data\_df$Sub.Category))

## [1] "Bookcases" "Chairs" "Labels" "Tables" "Storage"   
## [6] "Furnishings" "Art" "Phones" "Binders" "Appliances"   
## [11] "Paper" "Accessories" "Envelopes" "Fasteners" "Supplies"   
## [16] "Machines" "Copiers"

# print(unique(data\_df$Sales))  
print(unique(data\_df$Quantity))

## [1] 2 3 5 7 4 6 9 1 8 14 11 13 10 12

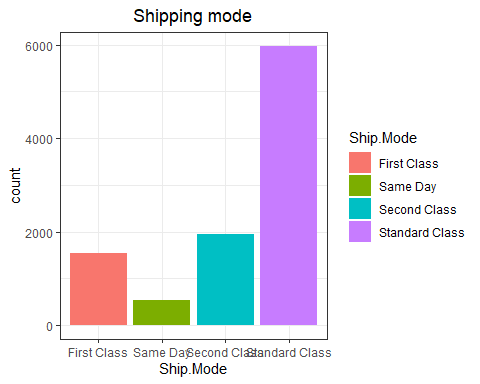
colnames(data\_df)

## [1] "Order.Date" "Ship.Date" "Ship.Mode" "Segment" "Country"   
## [6] "City" "State" "Postal.Code" "Region" "Category"   
## [11] "Sub.Category" "Product.Name" "Sales" "Quantity" "Discount"   
## [16] "Profit" "profit\_cat"

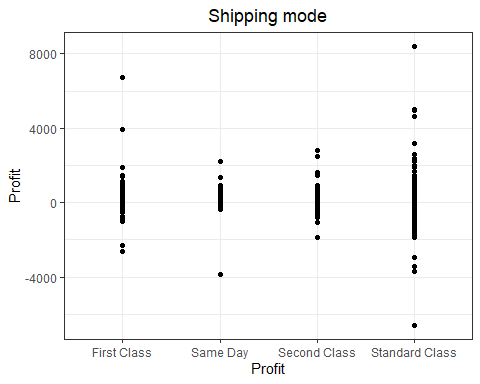
sapply(data\_df,class)

## Order.Date Ship.Date Ship.Mode Segment Country City   
## "character" "character" "character" "character" "character" "character"   
## State Postal.Code Region Category Sub.Category Product.Name   
## "character" "integer" "character" "character" "character" "character"   
## Sales Quantity Discount Profit profit\_cat   
## "numeric" "integer" "numeric" "numeric" "logical"

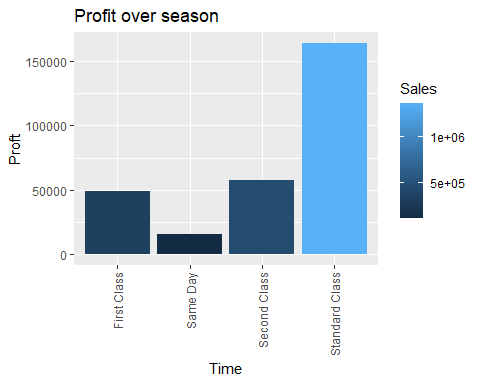
ggplot(data\_df, aes(Ship.Mode, fill = Ship.Mode)) +   
 geom\_bar() +  
 theme\_bw() +  
 labs(title = "Shipping mode", x = "Ship.Mode") +  
 theme(plot.title = element\_text(hjust = 0.5))



ggplot(data\_df, aes(Ship.Mode ,Profit)) +   
 geom\_point() +  
 theme\_bw() +  
 labs(title = "Shipping mode", x = "Profit") +  
 theme(plot.title = element\_text(hjust = 0.5))



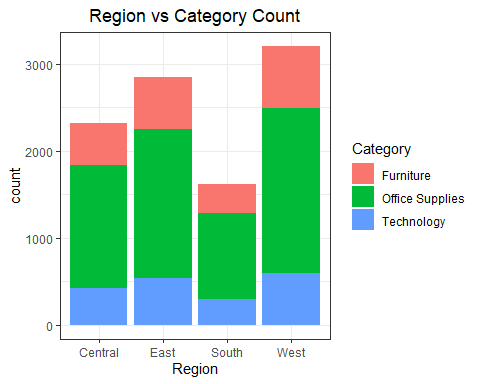
ship\_mode\_profit\_df <- data\_df %>%  
 group\_by(Ship.Mode) %>%  
 summarize(Profit = sum(Profit),Sales = sum(Sales))  
  
ggplot(data = ship\_mode\_profit\_df, aes(x = Ship.Mode, y = Profit,fill = Sales)) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle("Profit over season") +  
 xlab("Time") + ylab("Proft")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



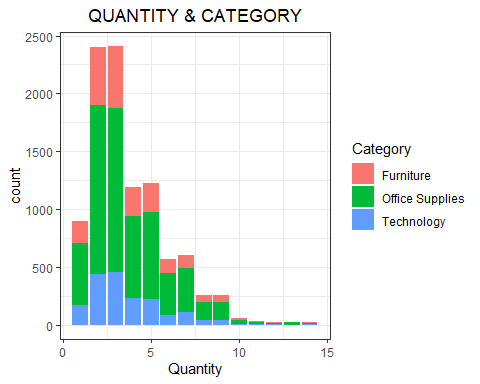
### Figure 1:- Frequencies of Ship Mode specified by customer

* We can see from above that most customers prefer Standard Class.
* Also, Sales and Profits are more when customers use Standard Class shipping method.

#region wise orders count  
  
ggplot(data\_df, aes(x = Region, fill = Category )) +   
 geom\_bar(position="stack") +  
 theme\_bw() +  
 labs(title = "Region vs Category Count", x = "Region") +  
 theme(plot.title = element\_text(hjust = 0.5))



ggplot(data\_df, aes(x = Quantity , fill = Category )) +   
 geom\_bar(position="stack") +  
 theme\_bw() +  
 labs(title = "QUANTITY & CATEGORY", x = "Quantity") +  
 theme(plot.title = element\_text(hjust = 0.5))



# Quantity of orders cummulative sum  
quantity\_grp <- data\_df %>% group\_by(Quantity) %>% summarise(Quantity = sum(Quantity))  
100\*cumsum(quantity\_grp)/sum(quantity\_grp)

## Quantity  
## 1 2.373723  
## 2 15.058221  
## 3 34.140417  
## 4 46.719299  
## 5 62.957780  
## 6 72.019645  
## 7 83.220236  
## 8 88.648906  
## 9 94.779922  
## 10 96.284952  
## 11 97.272463  
## 12 98.001215  
## 13 98.927996  
## 14 100.000000

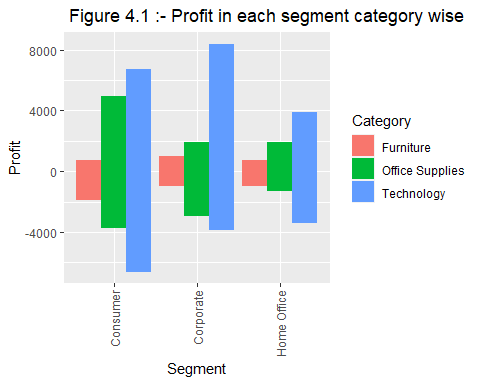
### Figure 2 & 3:- # of orders coming from each category from different Regions

* From above plot we can see that the most of the orders come from Western and Eastern regions.
* Among Categories office supplies are the most ordered by customer
* More than 95% of orders are having quantity of less than or equal to 10.

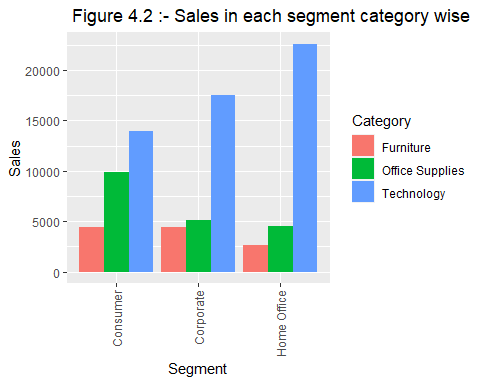
segment\_grp <- data\_df %>% group\_by(Segment) %>% summarise(Sales = sum(Sales),Profit= sum(Profit), .groups = "keep")# %>% group\_keys()  
category\_grp <- data\_df %>% group\_by(Category) %>% summarise(Sales = sum(Sales),Profit= sum(Profit), .groups = "keep")  
segment\_grp<-segment\_grp[order(segment\_grp$Sales),]  
segment\_grp$perc\_sales <- 100\*segment\_grp$Sales/sum(segment\_grp$Sales)  
segment\_grp<-segment\_grp[order(segment\_grp$Profit),]  
segment\_grp$perc\_profit <- 100\*segment\_grp$Profit/sum(segment\_grp$Profit)  
segment\_grp

## # A tibble: 3 × 5  
## # Groups: Segment [3]  
## Segment Sales Profit perc\_sales perc\_profit  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Home Office 429653. 60299. 18.7 21.1  
## 2 Corporate 706146. 91979. 30.7 32.1  
## 3 Consumer 1161401. 134119. 50.6 46.8

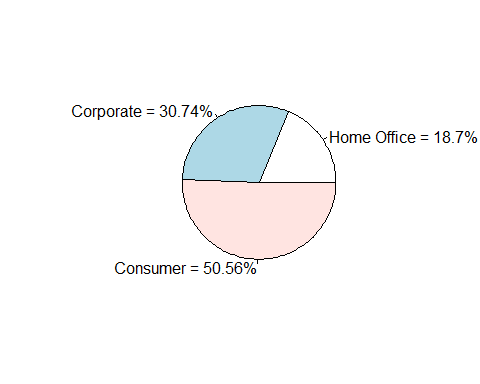
category\_grp<-category\_grp[order(category\_grp$Sales),]  
category\_grp$perc\_sales <- 100\*category\_grp$Sales/sum(category\_grp$Sales)  
category\_grp<-category\_grp[order(category\_grp$Profit),]  
category\_grp$perc\_profit <- 100\*category\_grp$Profit/sum(category\_grp$Profit)  
  
# print(colnames(data\_df))  
  
ggplot(data\_df, aes(x=Segment,y=Profit, fill=Category)) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle(" Figure 4.1 :- Profit in each segment category wise") + theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



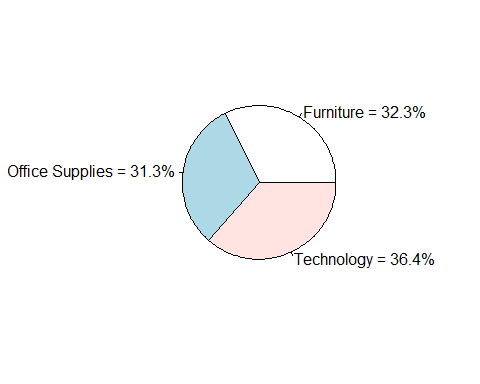
ggplot(data\_df, aes(x=Segment,y=Sales, fill=Category)) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle(" Figure 4.2 :- Sales in each segment category wise") + theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



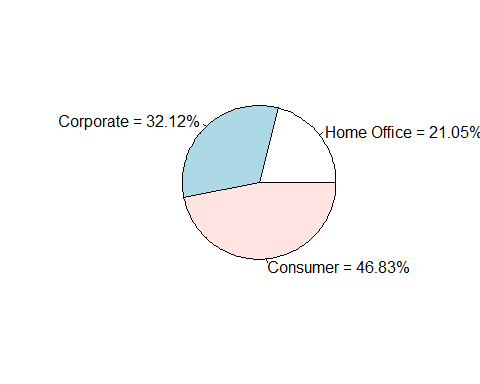
# library(lessR)  
# # Categorical data  
#   
#   
# cols <- hcl.colors(length(unique(category\_grp$Category)), "Fall")  
#   
# PieChart(Category, data = data\_df, hole = 0,  
# fill = cols,  
# labels\_cex = 0.6)  
# par(mfrow=c(2,2))  
  
pie\_labels <- paste0(segment\_grp$Segment, " = ", round(segment\_grp$perc\_sales,2), "%")  
pie(segment\_grp$perc\_sales, labels = pie\_labels)



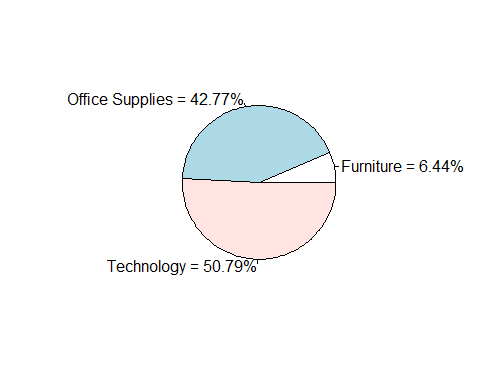
pie\_labels <- paste0(category\_grp$Category, " = ", round(category\_grp$perc\_sales,2), "%")  
pie(category\_grp$perc\_sales, labels = pie\_labels)



pie\_labels <- paste0(segment\_grp$Segment, " = ", round(segment\_grp$perc\_profit,2), "%")  
pie(segment\_grp$perc\_profit, labels = pie\_labels)



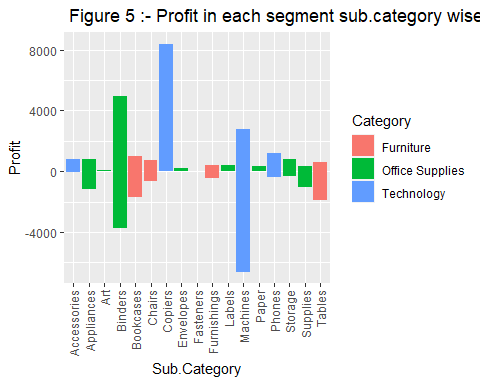
pie\_labels <- paste0(category\_grp$Category, " = ", round(category\_grp$perc\_profit,2), "%")  
pie(category\_grp$perc\_profit, labels = pie\_labels)



### Figure 4:- Sales & Profit in each segment category wise

* From above we can see Loss is more in Consumer segment in all categories than in Corporate and Home Office segments.
* From Pie charts, we can see distribution of sales and profits among segments and categories. with highest being

ggplot(data\_df, aes(x=Sub.Category,y=Profit, fill=Category)) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle(" Figure 5 :- Profit in each segment sub.category wise") + theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



category\_profit <-data\_df %>%  
 group\_by(Category) %>%  
 summarize(Profit = sum(Profit))  
  
sub\_category\_profit <- data\_df %>%  
 group\_by(Sub.Category) %>%  
 summarize(Profit = sum(Profit))  
  
  
category\_profit[order(category\_profit$Profit,decreasing = TRUE),]

## # A tibble: 3 × 2  
## Category Profit  
## <chr> <dbl>  
## 1 Technology 145455.  
## 2 Office Supplies 122491.  
## 3 Furniture 18451.

sub\_category\_profit[order(sub\_category\_profit$Profit,decreasing = TRUE),]

## # A tibble: 17 × 2  
## Sub.Category Profit  
## <chr> <dbl>  
## 1 Copiers 55618.  
## 2 Phones 44516.  
## 3 Accessories 41937.  
## 4 Paper 34054.  
## 5 Binders 30222.  
## 6 Chairs 26590.  
## 7 Storage 21279.  
## 8 Appliances 18138.  
## 9 Furnishings 13059.  
## 10 Envelopes 6964.  
## 11 Art 6528.  
## 12 Labels 5546.  
## 13 Machines 3385.  
## 14 Fasteners 950.  
## 15 Supplies -1189.  
## 16 Bookcases -3473.  
## 17 Tables -17725.

### Figure 5:- Profit in each segment sub-category wise

* We can answer **Question 2** from above plot that Technology is the most Profitable among others.
  + Among technology category, we can see that copiers, Phones are more profitable than Machines
  + If we order categories by Profits, we can say
  + Profit order for sub categories can be seen above Copiers, Phones, Accessories.. etc
  + Least Profitable sub categories are Tables,Bookcases Supplies. ( They incur more losses rather than profits )

# Monthly sales and Profits across Categories  
data\_df$Month\_Yr <- strftime(strptime(data\_df$Order.Date,"%m/%d/%Y"),"%Y-%m")  
data\_df$Month <- strftime(strptime(data\_df$Order.Date,"%m/%d/%Y"),"%m")  
data\_df$Year <- strftime(strptime(data\_df$Order.Date,"%m/%d/%Y"),"%Y")  
data\_df <- transform(data\_df, Month = as.numeric(Month),   
 Year = as.numeric(Year))  
  
data\_df <- within(data\_df, {   
 season <- NA # need to initialize variable  
 season[Month >= 3 & Month <=5 ] <- "Spring"  
 season[Month >= 6 & Month <= 8] <- "Summer"  
 season[(Month >= 9 & Month <= 11) | (Month == 12) | (Month >= 1 & Month <= 2)] <- "Fall"  
   
 } )  
  
# head(data\_df)

monthly\_sales <- data\_df %>%  
 group\_by(Month) %>%  
 summarize(Profit = sum(Profit),Sales= sum(Sales))  
monthly\_sales[order(monthly\_sales$Profit,monthly\_sales$Sales,decreasing = TRUE), ]

## # A tibble: 12 × 3  
## Month Profit Sales  
## <dbl> <dbl> <dbl>  
## 1 12 43369. 325294.  
## 2 9 36857. 307650.  
## 3 11 35468. 352461.  
## 4 10 31784. 200323.  
## 5 3 28595. 205005.  
## 6 5 22411. 155029.  
## 7 8 21777. 159044.  
## 8 6 21286. 152719.  
## 9 7 13833. 147238.  
## 10 4 11587. 137762.  
## 11 2 10295. 59751.  
## 12 1 9134. 94925.

yearly\_sales <- data\_df %>%  
 group\_by(Year) %>%  
 summarize(Profit = sum(Profit),Sales= sum(Sales))  
yearly\_sales[order(yearly\_sales$Profit,yearly\_sales$Sales,decreasing = TRUE), ]

## # A tibble: 4 × 3  
## Year Profit Sales  
## <dbl> <dbl> <dbl>  
## 1 2017 93439. 733215.  
## 2 2016 81795. 609206.  
## 3 2015 61619. 470533.  
## 4 2014 49544. 484247.

season\_sales <- data\_df %>%  
 group\_by(season) %>%  
 summarize(Profit = sum(Profit),Sales= sum(Sales))  
season\_sales[order(season\_sales$Profit,season\_sales$Sales,decreasing = TRUE), ]

## # A tibble: 3 × 3  
## season Profit Sales  
## <chr> <dbl> <dbl>  
## 1 Fall 166908. 1340404.  
## 2 Spring 62593. 497796.  
## 3 Summer 56895. 459001.

data\_df[is.na(data\_df$season),]

## [1] Order.Date Ship.Date Ship.Mode Segment Country   
## [6] City State Postal.Code Region Category   
## [11] Sub.Category Product.Name Sales Quantity Discount   
## [16] Profit profit\_cat Month\_Yr Month Year   
## [21] season   
## <0 rows> (or 0-length row.names)

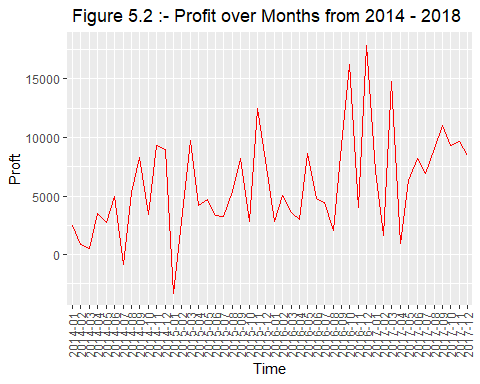
# ggplot(data\_df,   
# aes(x = Month\_Yr, y = Sales,fill = Segment)) +  
# geom\_bar(stat='identity', position='dodge') +  
# theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))  
  
monthly\_yr\_sales <- data\_df %>%  
 group\_by(Month\_Yr) %>%  
 summarize(Profit = sum(Profit),Sales= sum(Sales))  
monthly\_yr\_sales[order(monthly\_yr\_sales$Profit,monthly\_yr\_sales$Sales,decreasing = TRUE), ]

## # A tibble: 48 × 3  
## Month\_Yr Profit Sales  
## <chr> <dbl> <dbl>  
## 1 2016-12 17885. 96999.  
## 2 2016-10 16243. 59688.  
## 3 2017-03 14752. 58872.  
## 4 2015-11 12475. 75973.  
## 5 2017-09 10992. 87867.  
## 6 2015-03 9732. 38726.  
## 7 2017-11 9690. 118448.  
## 8 2016-09 9329. 73410.  
## 9 2014-11 9292. 78629.  
## 10 2017-10 9275. 77777.  
## # … with 38 more rows

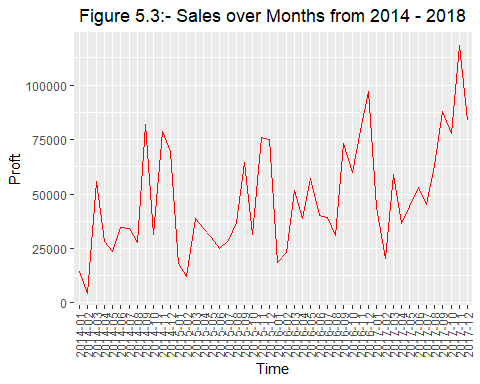
monthly\_yr\_sales

## # A tibble: 48 × 3  
## Month\_Yr Profit Sales  
## <chr> <dbl> <dbl>  
## 1 2014-01 2450. 14237.  
## 2 2014-02 862. 4520.  
## 3 2014-03 499. 55691.  
## 4 2014-04 3489. 28295.  
## 5 2014-05 2739. 23648.  
## 6 2014-06 4977. 34595.  
## 7 2014-07 -841. 33946.  
## 8 2014-08 5318. 27909.  
## 9 2014-09 8328. 81777.  
## 10 2014-10 3448. 31453.  
## # … with 38 more rows

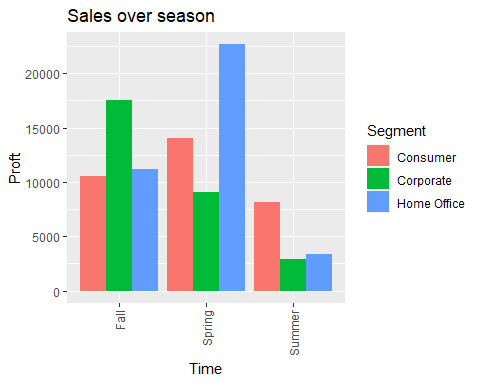
ggplot(data = monthly\_yr\_sales[order(monthly\_yr\_sales$Month\_Yr),], aes(x = Month\_Yr, y = Profit, group = 1))+  
 geom\_line(color = "red")+  
 ggtitle(" Figure 5.2 :- Profit over Months from 2014 - 2018") +  
 xlab("Time") + ylab("Proft")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



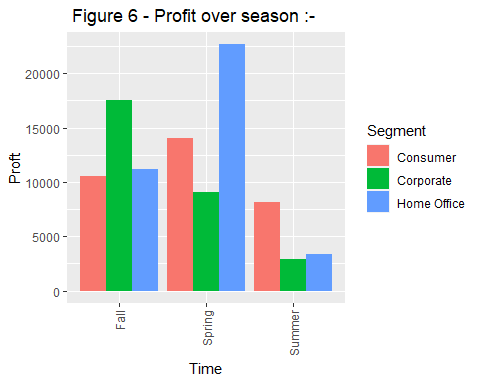
ggplot(data = monthly\_yr\_sales[order(monthly\_yr\_sales$Month\_Yr),], aes(x = Month\_Yr, y = Sales, group = 1))+  
 geom\_line(color = "red")+  
 ggtitle(" Figure 5.3:- Sales over Months from 2014 - 2018") +  
 xlab("Time") + ylab("Proft")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



ggplot(data = data\_df[order(data\_df$season),], aes(x = season, y = Sales,fill = Segment)) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle("Sales over season") +  
 xlab("Time") + ylab("Proft")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



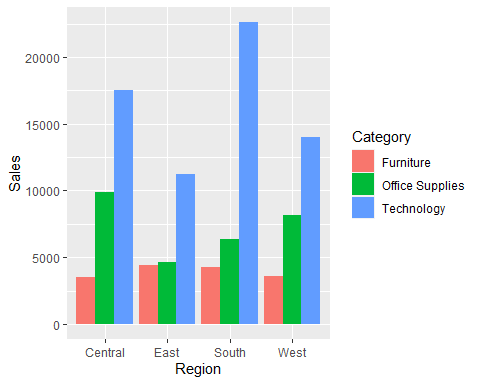
ggplot(data = data\_df[order(data\_df$season),], aes(x = season, y = Sales,fill = Segment)) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle(" Figure 6 - Profit over season :-") +  
 xlab("Time") + ylab("Proft")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



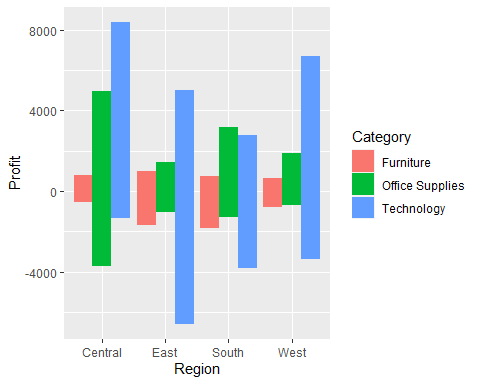
### Figure 5.2 and 6:- Profit from 2014 to 2018 and Sales, Profit Monthly

* You can see from above graph and data table, Sales and profits are high in 9,10,11,12 months in each year.
* Also Sales and profits tend to increase each year.
* If you consider seasonal Sales and Profits, order follows as below
* If you can see from figure 5.2, Profit is very low on January, february months as they are non holiday seasons/months. So very few product sales and profits can be seen during these months
* From figure 6,

ggplot(data\_df, aes(x=Region,y=Sales, fill=Category)) +  
 geom\_bar(stat='identity', position='dodge')



ggplot(data\_df, aes(x=Region,y=Profit, fill=Category)) +  
 geom\_bar(stat='identity', position='dodge')



# State wise profits & sales  
  
library(usmap)

## Warning: package 'usmap' was built under R version 4.2.2

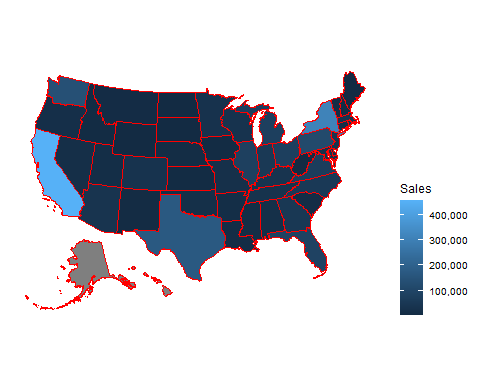
state\_sales <- data\_df %>%  
 group\_by(State) %>%  
 summarize(Profit = sum(Profit),Sales= sum(Sales))  
state\_sales[order(state\_sales$Profit,state\_sales$Sales,decreasing = TRUE), ]

## # A tibble: 49 × 3  
## State Profit Sales  
## <chr> <dbl> <dbl>  
## 1 California 76381. 457688.  
## 2 New York 74039. 310876.  
## 3 Washington 33403. 138641.  
## 4 Michigan 24463. 76270.  
## 5 Virginia 18598. 70637.  
## 6 Indiana 18383. 53555.  
## 7 Georgia 16250. 49096.  
## 8 Kentucky 11200. 36592.  
## 9 Minnesota 10823. 29863.  
## 10 Delaware 9977. 27451.  
## # … with 39 more rows

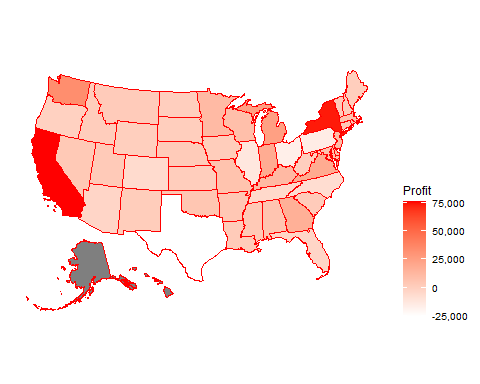
statepop1 <- usmap::statepop  
statepop1 <- statepop1[c('fips', 'full')] %>%  
 rename(State = full )  
  
data\_df\_plot <- merge(statepop1,state\_sales,by="State")  
head(data\_df\_plot)

## State fips Profit Sales  
## 1 Alabama 01 5786.825 19510.64  
## 2 Arizona 04 -3427.925 35282.00  
## 3 Arkansas 05 4008.687 11678.13  
## 4 California 06 76381.387 457687.63  
## 5 Colorado 08 -6527.858 32108.12  
## 6 Connecticut 09 3511.492 13384.36

plot\_usmap(data = data\_df\_plot[c('fips','Sales')], values = "Sales", color = "red") +  
 scale\_fill\_continuous(name = "Sales", label = scales::comma) +  
 theme(legend.position = "right")



plot\_usmap(data = data\_df\_plot[c('fips','Profit')], values = "Profit", color = "red") +   
 scale\_fill\_continuous(  
 low = "white", high = "red", name = "Profit", label = scales::comma  
 ) + theme(legend.position = "right")



state\_sales[order(state\_sales$Profit,state\_sales$Sales,decreasing = TRUE), ]

## # A tibble: 49 × 3  
## State Profit Sales  
## <chr> <dbl> <dbl>  
## 1 California 76381. 457688.  
## 2 New York 74039. 310876.  
## 3 Washington 33403. 138641.  
## 4 Michigan 24463. 76270.  
## 5 Virginia 18598. 70637.  
## 6 Indiana 18383. 53555.  
## 7 Georgia 16250. 49096.  
## 8 Kentucky 11200. 36592.  
## 9 Minnesota 10823. 29863.  
## 10 Delaware 9977. 27451.  
## # … with 39 more rows

### Figure 7:- State wise profits & sales

* Heatmap of state wise profits & sales
* California, New York drives most of the sales and profits overall.

head(data\_df)

## Order.Date Ship.Date Ship.Mode Segment Country City  
## 1 11/8/2016 11/11/2016 Second Class Consumer United States Henderson  
## 2 11/8/2016 11/11/2016 Second Class Consumer United States Henderson  
## 3 6/12/2016 6/16/2016 Second Class Corporate United States Los Angeles  
## 4 10/11/2015 10/18/2015 Standard Class Consumer United States Fort Lauderdale  
## 5 10/11/2015 10/18/2015 Standard Class Consumer United States Fort Lauderdale  
## 6 6/9/2014 6/14/2014 Standard Class Consumer United States Los Angeles  
## State Postal.Code Region Category Sub.Category  
## 1 Kentucky 42420 South Furniture Bookcases  
## 2 Kentucky 42420 South Furniture Chairs  
## 3 California 90036 West Office Supplies Labels  
## 4 Florida 33311 South Furniture Tables  
## 5 Florida 33311 South Office Supplies Storage  
## 6 California 90032 West Furniture Furnishings  
## Product.Name Sales  
## 1 Bush Somerset Collection Bookcase 261.9600  
## 2 Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.9400  
## 3 Self-Adhesive Address Labels for Typewriters by Universal 14.6200  
## 4 Bretford CR4500 Series Slim Rectangular Table 957.5775  
## 5 Eldon Fold 'N Roll Cart System 22.3680  
## 6 Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood 48.8600  
## Quantity Discount Profit profit\_cat Month\_Yr Month Year season  
## 1 2 0.00 41.9136 TRUE 2016-11 11 2016 Fall  
## 2 3 0.00 219.5820 TRUE 2016-11 11 2016 Fall  
## 3 2 0.00 6.8714 TRUE 2016-06 6 2016 Summer  
## 4 5 0.45 -383.0310 FALSE 2015-10 10 2015 Fall  
## 5 2 0.20 2.5164 TRUE 2015-10 10 2015 Fall  
## 6 7 0.00 14.1694 TRUE 2014-06 6 2014 Summer

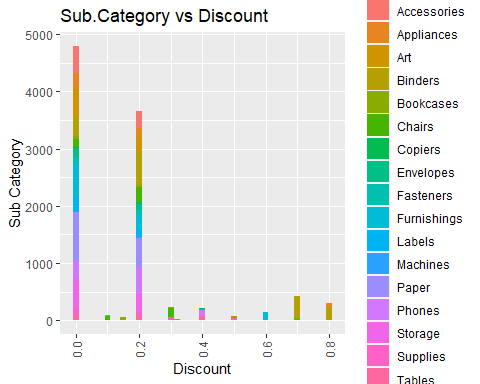
seg\_sea\_df <- data\_df %>%  
 group\_by(season,Sub.Category) %>%  
 summarize(Discount = mean(Discount), Sales = sum(Sales),Profit = sum(Profit))

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

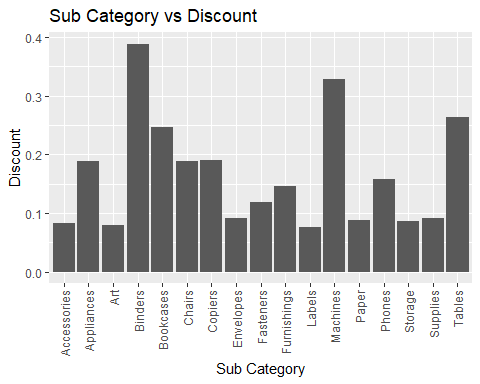
seg\_sea\_df

## # A tibble: 51 × 5  
## # Groups: season [3]  
## season Sub.Category Discount Sales Profit  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Fall Accessories 0.0832 102638. 25749.  
## 2 Fall Appliances 0.155 62866. 12084.  
## 3 Fall Art 0.0791 15205. 3601.  
## 4 Fall Binders 0.371 124726. 17614.  
## 5 Fall Bookcases 0.206 73160. -3329.  
## 6 Fall Chairs 0.162 199805. 18631.  
## 7 Fall Copiers 0.144 85249. 34485.  
## 8 Fall Envelopes 0.0848 10364. 4334.  
## 9 Fall Fasteners 0.0722 1967. 651.  
## 10 Fall Furnishings 0.138 54548. 7915.  
## # … with 41 more rows

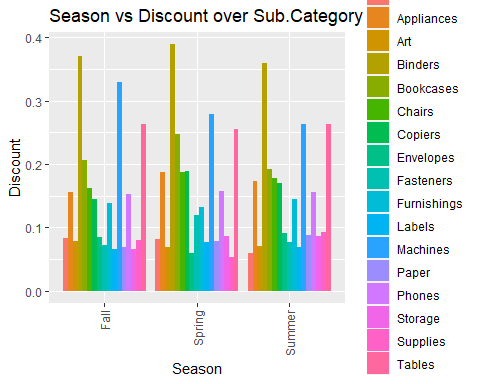
ggplot(data = data\_df, aes(x = Discount , fill = Sub.Category )) +  
 geom\_bar( position='stack')+  
 ggtitle("Sub.Category vs Discount ") +  
 ylab("Sub Category") + xlab("Discount")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



ggplot(data = seg\_sea\_df, aes(x = Sub.Category, y =Discount )) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle("Sub Category vs Discount ") +  
 xlab("Sub Category") + ylab("Discount")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



ggplot(data = seg\_sea\_df, aes(x = season, y =Discount, fill = Sub.Category )) +  
 geom\_bar(stat='identity', position='dodge')+  
 ggtitle("Season vs Discount over Sub.Category") +  
 xlab("Season") + ylab("Discount")+  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))



#   
# ggplot(data = seg\_sea\_df, aes(x = Sales, y = Discount ,fill = Sub.Category)) +  
# geom\_bar(stat='identity', position='dodge')+  
# ggtitle("Discount vs Sales over Sub.Category") +  
# xlab("Discount") + ylab("Sales")+  
# theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))  
#   
# ggplot(data = seg\_sea\_df, aes(x = Profit, y =Discount ,fill = Sub.Category)) +  
# geom\_bar(stat='identity', position='dodge')+  
# ggtitle("Discount vs profit over Sub.Category") +  
# xlab("Discount") + ylab("Profit")+  
# theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

### Figure 8:- Discount vs Profit

* From data, we can see most of the products have discounts from range 0 to 0.3,
* From above graph, we can see that the more the discounts are it is less likely we get profits from that stores.
* If discounts are more than 0.4, we see profits are going down. So we can infer higher discounts would leave stores in losses.

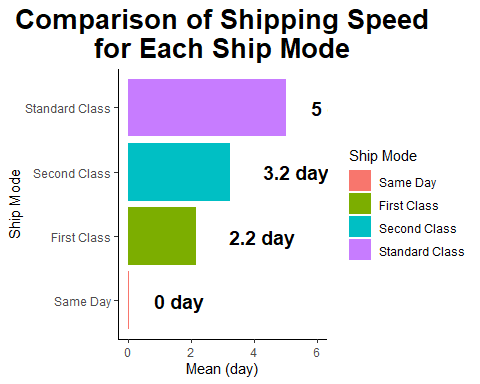
data\_df$Order.Date <-strftime(strptime(data\_df$Order.Date,"%m/%d/%Y"),"%m-%d-%Y")  
data\_df$Ship.Date <-strftime(strptime(data\_df$Ship.Date,"%m/%d/%Y"),"%m-%d-%Y")  
data\_df$delivery\_days<-as.numeric(as.Date(as.character(data\_df$Ship.Date), format="%m-%d-%Y")-  
 as.Date(as.character(data\_df$Order.Date), format="%m-%d-%Y"))  
  
# data\_df# %>%  
 # mutate(Order.Date = mdy(Order.Date),  
 # Ship.Date = mdy(Ship.Date),  
 # Shipping.Speed = Ship.Date - Order.Date)  
# sum(is.na(data\_df))  
sum(is.na(data\_df$delivery\_days))

## [1] 0

# df1 <- data\_df %>%  
# mutate('Order Date' = mdy('Order Date'),  
# 'Ship Date' = mdy('Ship Date'),  
# 'Shipping Speed' = 'Ship Date' - 'Order Date')  
  
a <- data\_df %>%   
 group\_by(Ship.Mode) %>%  
 summarize(mean=mean(delivery\_days))  
a

## # A tibble: 4 × 2  
## Ship.Mode mean  
## <chr> <dbl>  
## 1 First Class 2.18   
## 2 Same Day 0.0442  
## 3 Second Class 3.24   
## 4 Standard Class 5.01

ggplot(data = a, aes(x=reorder(Ship.Mode, mean), y=mean, fill=reorder(Ship.Mode, mean)))+  
 geom\_bar(stat='identity')+  
 coord\_flip()+  
 geom\_text(aes(label = paste0(round(mean, 1), ' day')), hjust = -0.5, size=5, fontface='bold')+  
 scale\_y\_continuous(limits = c(0,6))+  
 theme\_classic()+  
 labs(title='Comparison of Shipping Speed\nfor Each Ship Mode',  
 x='Ship Mode',  
 y='Mean (day)',  
 fill='Ship Mode')+  
 theme(plot.title = element\_text(size = 20, face = "bold", hjust=0.5))

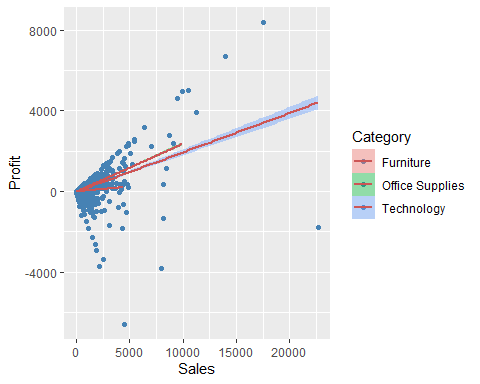


### Figure 9

* Fastest Ship.mode that will ship your item on the same day as the day you order is ‘Same day’
* average time taken for standard class to start shipping your item is 5 days.

ggplot(data\_df, aes(x=Sales,y=Profit, fill=Category)) +  
 geom\_point(color= "steelblue")+  
 geom\_smooth(method = "lm",color = "indianred3")

## `geom\_smooth()` using formula 'y ~ x'



### Figure 10

# install.packages("wordcloud")  
# install.packages("tm")  
# install.packages("SnowballC")  
# install.packages("RColorBrewer")  
# library(wordcloud)  
# library(RColorBrewer)  
# library(tm)  
# library(SnowballC)  
#   
  
# tm\_map(data\_df$Product.Name, content\_transformer(tolower))  
#   
# wordcloud(words=data\_df$Product.Name, scale=c(5,0.5), max.words=100, random.order=FALSE, rot.per=0.35, use.r.layout=FALSE, colors=brewer.pal(8, "Dark2"))  
  
# wordcloud(words= data\_df$Product.Name, min.freq =1, max.words=100, random.order=FALSE,rot.per=0.35,colors=brewer.pal(8,"Dark2"))

## Hypothesis Testing

1. How season effects profits in each year?
2. How does Profit changes with Discounts over each season(Summer, Winter, Spring)?
3. Is geographical region a factor for Sales?
4. Is Segment and Profit relate to each other?
5. If give more Discount, it can provide better Profits?

#### Hypothesis Testing Q1)

Superstore claims that Profit generated in a summer is above average Profit generated. After taking 10 samples from summer orders, it has sample mean around 35.24. Is there enough evidence to support the claim? The mean population Sales of all regions s 28.65 with standard deviation of 234.26

season\_df <- data\_df[(data\_df$season == "Summer"),]  
  
n <- 50  
sd\_profit<-sd(data\_df$Profit)  
mean\_profit<-mean(data\_df$Profit)  
sample\_df<-season\_df[sample(nrow(season\_df),n),]  
sample\_mean\_profit <- mean(sample\_df$Profit)  
  
print(paste(sample\_mean\_profit,sd\_profit,mean\_profit,n))

## [1] "32.78994 234.260107690957 28.6568963077847 50"

sample mean = 28.656, population deviation =234.26 \* Null Hypothesis : mean = 28.65 \* Alternative Hypothesis : mean > 28.65

Test statistics would be average profit for summer season is 26.7 Reference distribution would be z-distribution and since we are using < in alternative hypothesis we use one tail z-test. We consider confidence interval of 95% which means alpha = 0.05, area under normal distribution for alpha = 0.05 is 2.677 If z-score is greater than 2.677 then we reject null hypothesis.

reference - <https://socratic.org/questions/what-is-the-z-score-of-0-05>

z <- (sample\_mean\_profit - mean\_profit)/(sd\_profit/sqrt(nrow(data\_df)/10))  
z

## [1] 0.5577522

Since z value is greater than 2.677, we can conclude that we reject null hypothesis

#### Hypothesis Testing Q2)

1. Null Hypothesis :- There is no dependency between Discount and Profit
2. Alternate Hypothesis :- There is a coorelation between Discount and Profit

Reference distribution would be chi-squared distribution. We consider confidence interval of 95% which means alpha = 0.05.

n <- nrow(data\_df)/100  
sample\_df<-data\_df[sample(nrow(data\_df),n),]  
print(n)

## [1] 99.94

# t.test(sample\_df$profit\_col,sample\_df$Discount)  
chisq.test(sample\_df$Discount, sample\_df$Profit, correct=FALSE)

## Warning in chisq.test(sample\_df$Discount, sample\_df$Profit, correct = FALSE):  
## Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: sample\_df$Discount and sample\_df$Profit  
## X-squared = 591.36, df = 576, p-value = 0.3198

As P-value is > 0.05 we accept Null Hypothesis.

east\_df <- data\_df[(data\_df$Region== "East"), ]  
n <- 50#nrow(east\_df)/10  
print(n)

## [1] 50

sample\_df<-data\_df[sample(nrow(east\_df),n),]  
sample\_mean <- mean(sample\_df$Sales)  
pop\_mean <- mean(data\_df$Sales)  
pop\_sd <- sd(data\_df$Sales)  
print(paste(sample\_mean,pop\_mean, pop\_sd))

## [1] "150.78893 229.858000830498 623.245100508681"

#### Hypothesis Testing Q3)

Superstore claims that sales generated in a east region is above average Sales generated. After taking 50 samples from east region has mean 272.75 Is there enough evidence to support the claim? The mean population Sales of all regions s 229.8 with standard deviation of 623.2

1. Null Hypothesis :- mean = 229.8
2. Alternate Hypothesis :- mean > 229.8

Reference distribution would be z-distribution and since we are using > in alternative hypothesis we use one tail z-test. We consider confidence interval of 95% which means alpha = 0.05, area under normal distribution for alpha = 0.05 is 2.677

z <- (sample\_mean - pop\_mean)/(pop\_sd/sqrt(n))  
z

## [1] -0.8970833

Since z value is less than 2.677, we can conclude that we accept null hypothesis that sales generated in east region is above average Sales.

#### Hypothesis Testing Q4)

1. Null Hypothesis :- There is no dependency between Segment and Profit
2. Alternate Hypothesis :- There is a coorelation between Segment and Profit

Reference distribution would be chi-squared distribution. We consider confidence interval of 95% which means alpha = 0.05.

n <- 10 #nrow(data\_df)/100  
sample\_df<-data\_df[sample(nrow(data\_df),n),]  
print(n)

## [1] 10

chisq.test(sample\_df$delivery\_days,sample\_df$Profit)

## Warning in chisq.test(sample\_df$delivery\_days, sample\_df$Profit): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: sample\_df$delivery\_days and sample\_df$Profit  
## X-squared = 40, df = 36, p-value = 0.297

As P-value is > 0.05 we accept Null Hypothesis.

#### Hypothesis Testing Q5)

1. Null Hypothesis :- Discount is dependent of Profit
2. Alternate Hypothesis :- Discount is independent of Profit

Reference distribution would be t-distribution. We consider confidence interval of 95% which means alpha = 0.05.

n <- nrow(data\_df)/100  
sample\_df<-data\_df[sample(nrow(data\_df),n),]  
print(n)

## [1] 99.94

t.test(sample\_df$profit\_cat,sample\_df$Discount)

##   
## Welch Two Sample t-test  
##   
## data: sample\_df$profit\_cat and sample\_df$Discount  
## t = 10.165, df = 154.04, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.4223602 0.6261247  
## sample estimates:  
## mean of x mean of y   
## 0.7272727 0.2030303

Upon performing T.Test, we can deduce that the Alternate Hypothesis is accepted and the null hypothesis is rejected.

## Summary

* Superstore dataset is a sample dataset widely used to analyze and get insights through visualization tools such as tableau, powerbi etc.
* Using this dataset i would like to find answers for above questions posted.
* References
  1. <https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>)
  2. <https://www.tableau.com/data-insights/dashboard-showcase/superstore>)

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.