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 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x 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

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

Data Quality Checks

```
##  [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
```

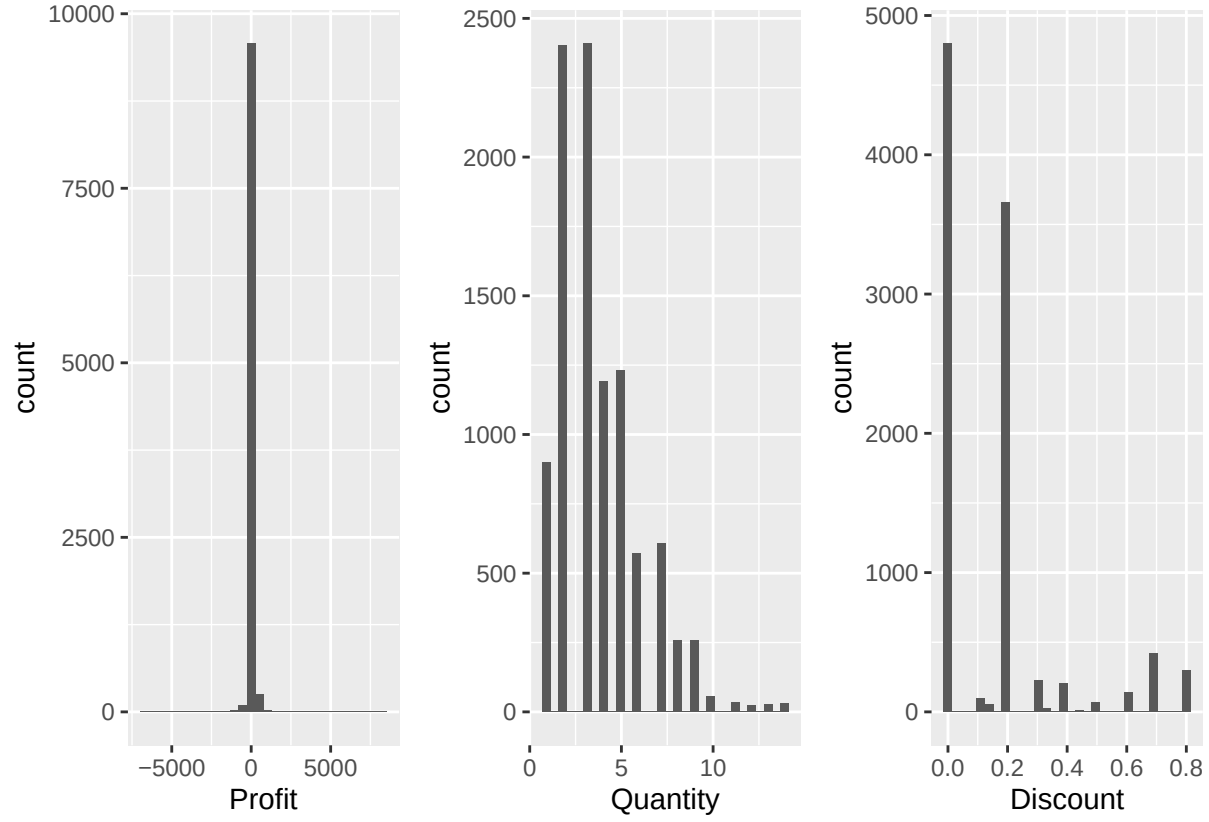
```
# colnames(data_df)
# Duplicates check
sum(duplicated(data_df))
```

```
## [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"
```

```r
print(unique(data_df$Region))
```

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

```r
print(unique(data_df$Category))
```

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

```r
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"
```

```r
# 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
```

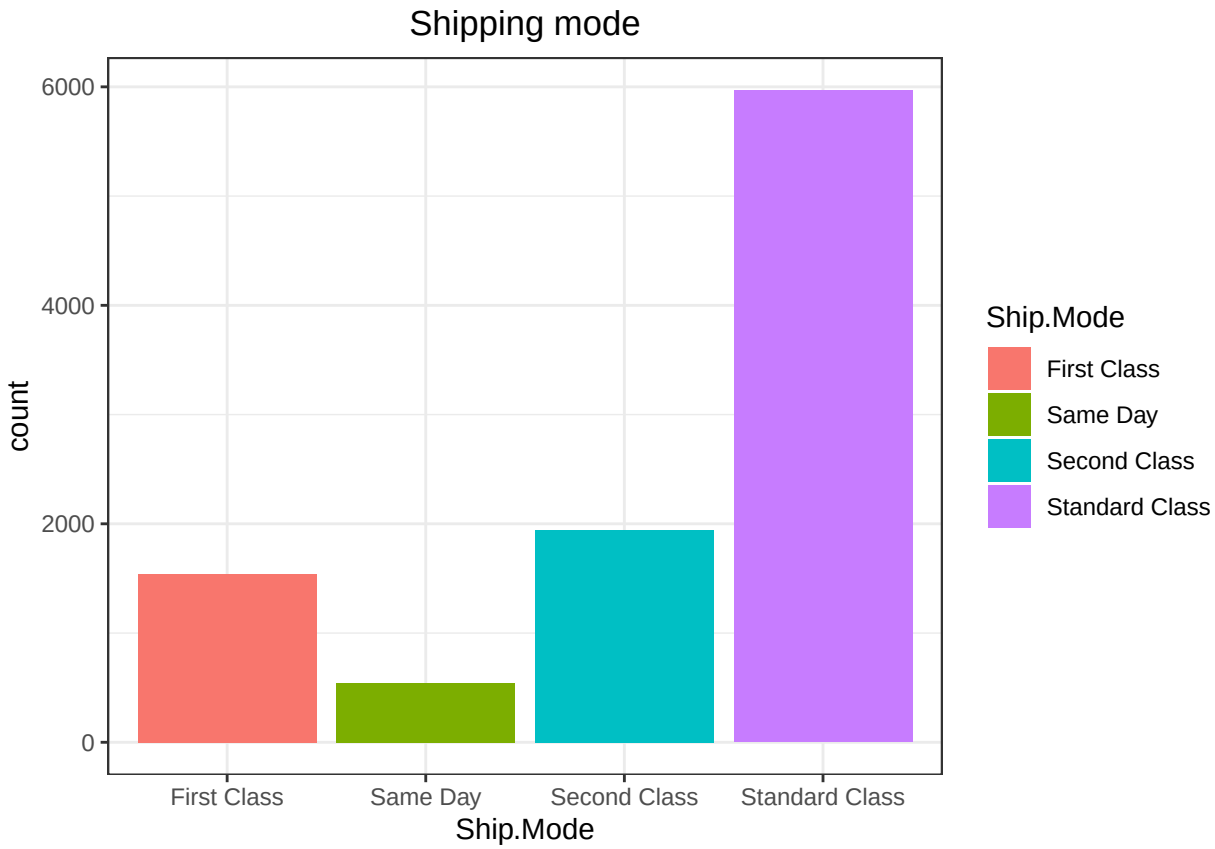
```r
colnames(data_df)
```

```
##  [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"
```

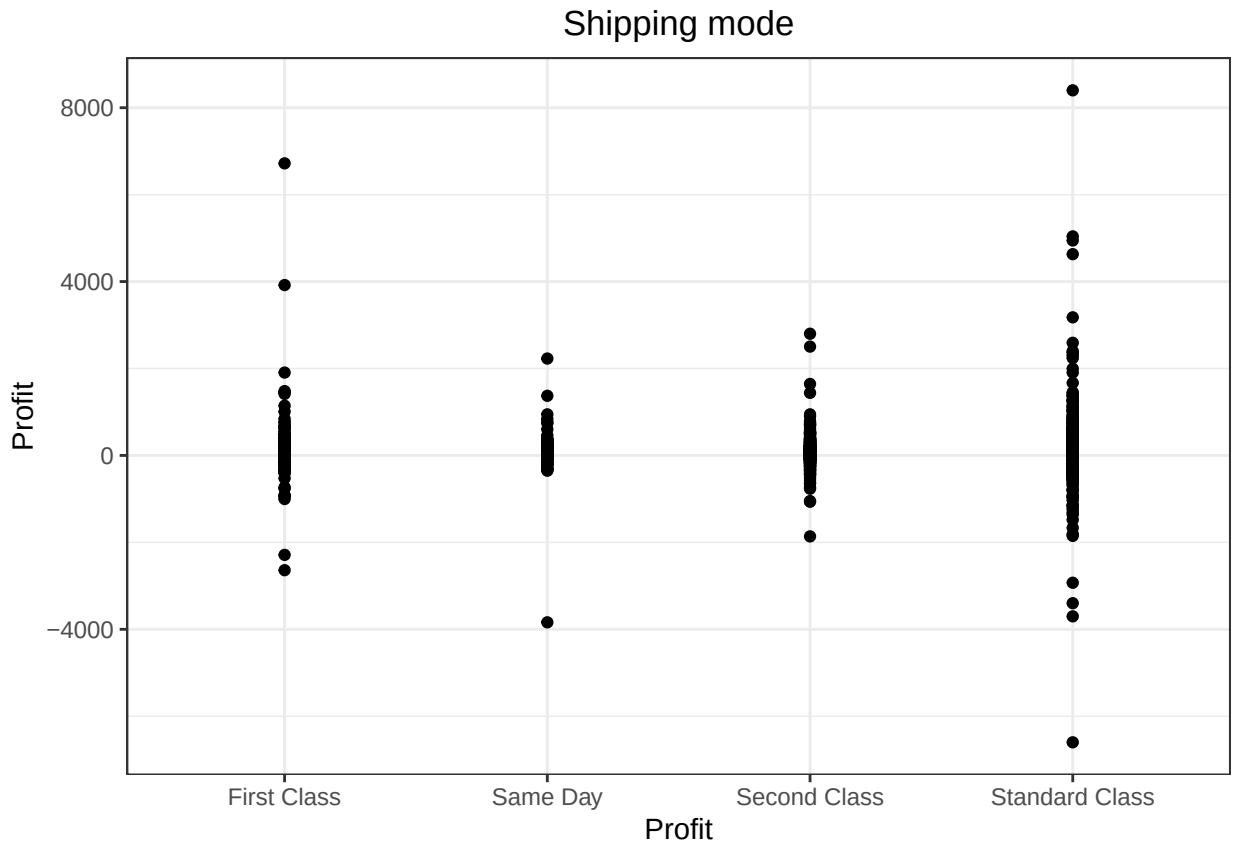
```r
sapply(data_df,class)
```

```
##   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"
```

```r
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))
```

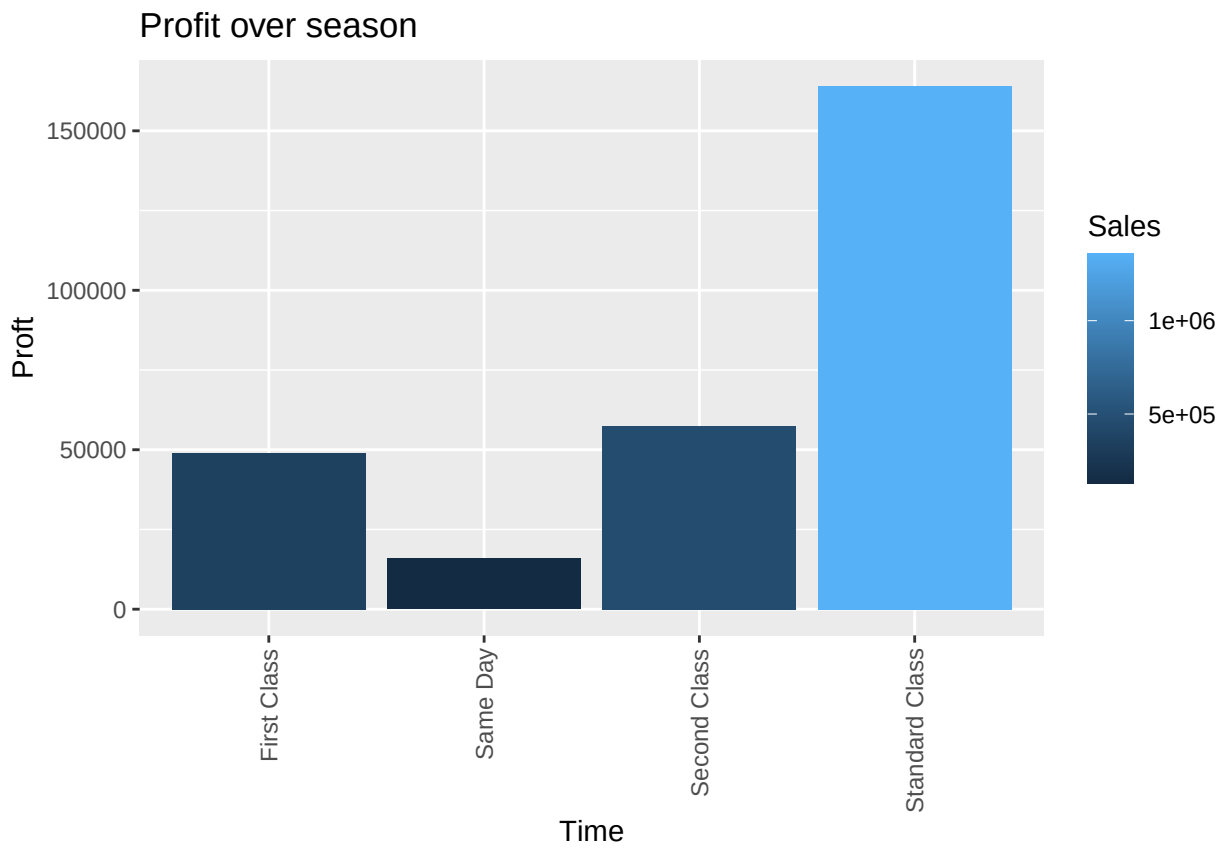
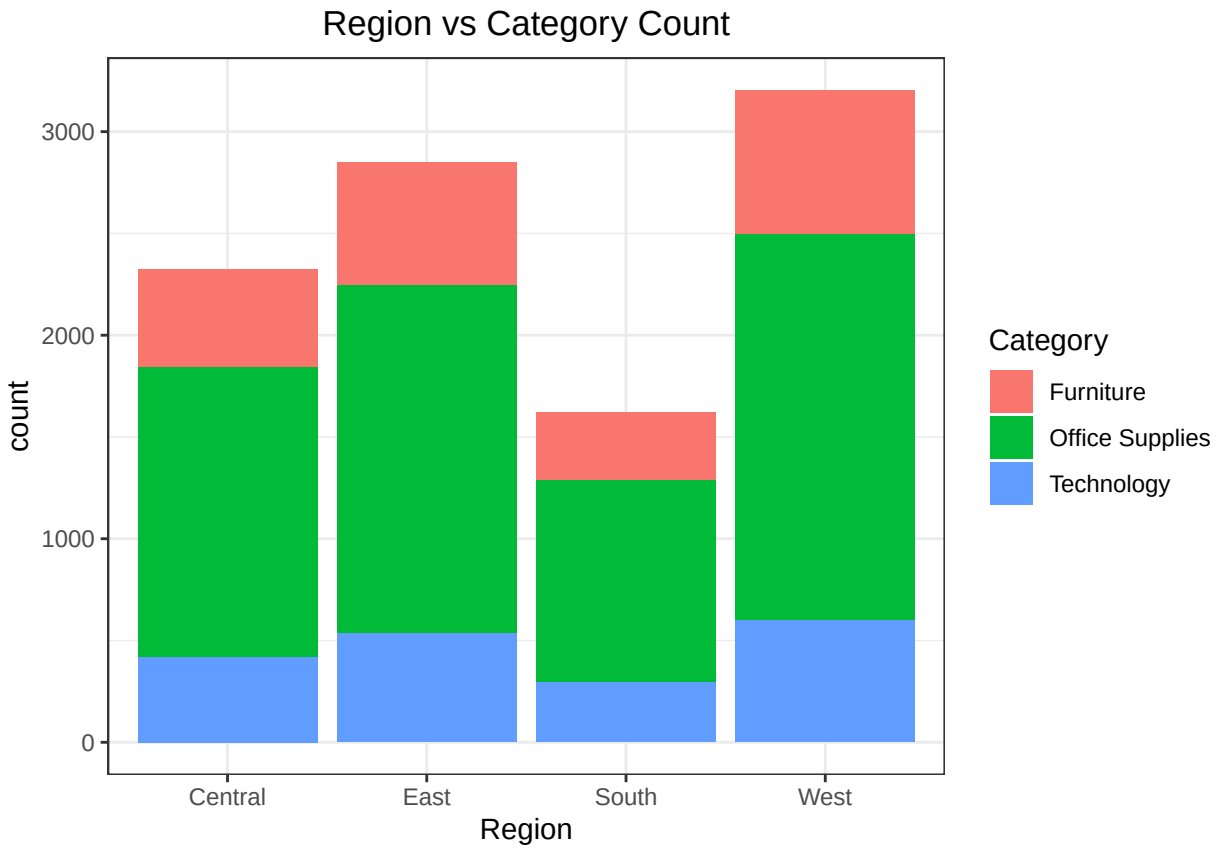


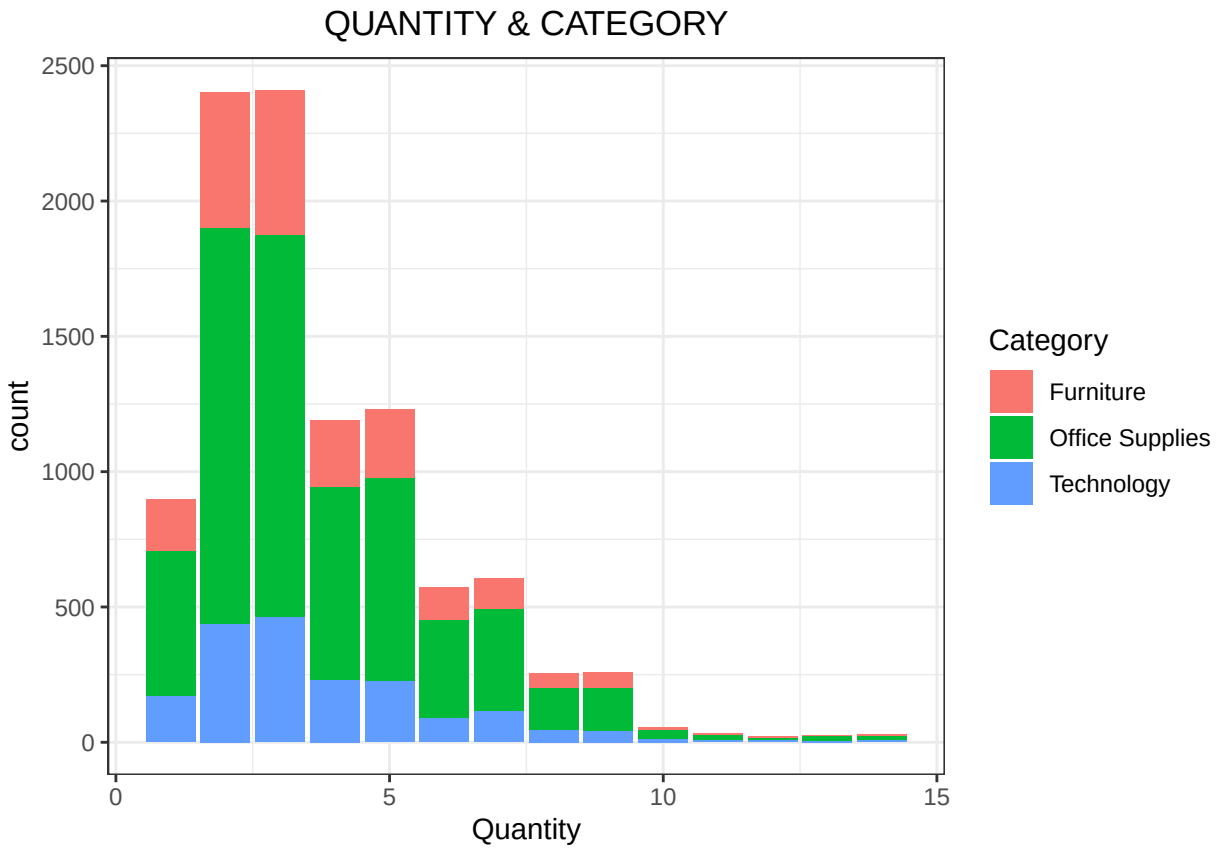
Figure 1:- Frequencies of Ship Mode specified by customer

- 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

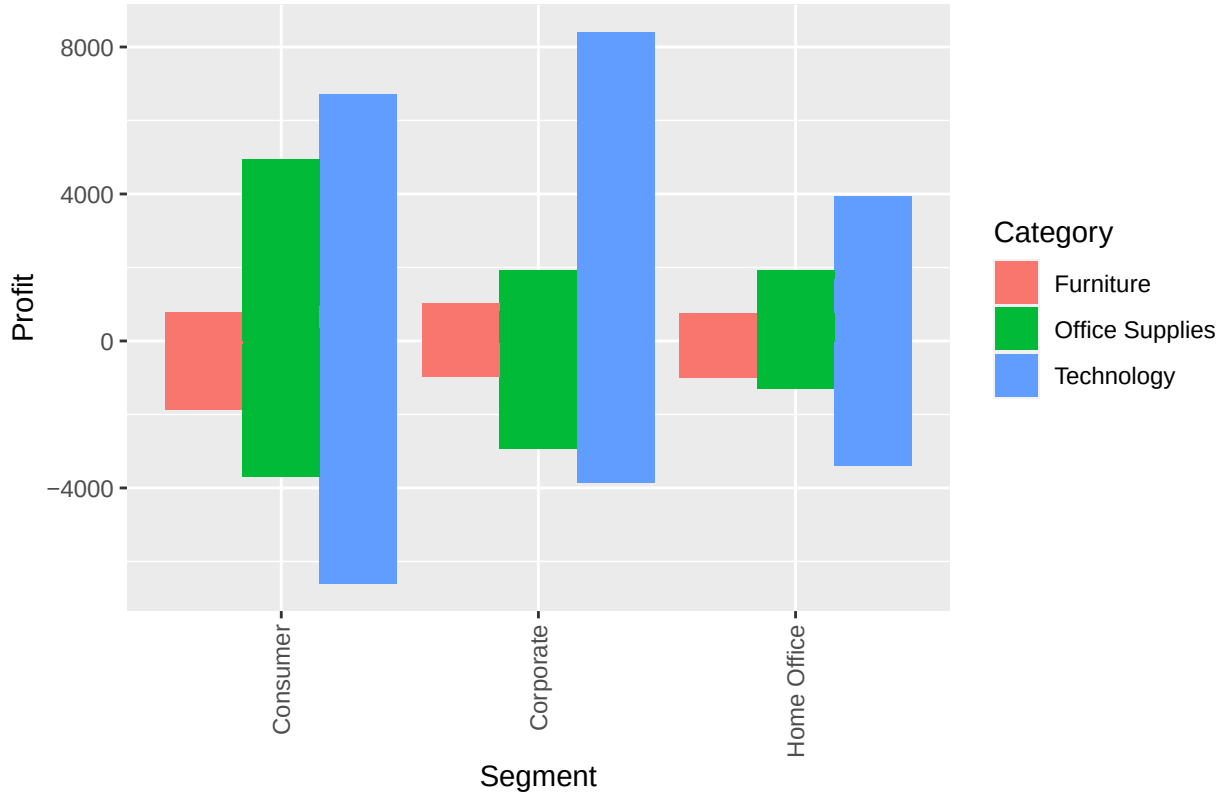
- From above plot we can see that the most of the orders come from Western and Eastern regions.
- Among Categories office supplies are the most ordered by customer
- More than 95% of orders are having quantity of less than or equal to 10.

```r
segment_grp <-  data_df %>% group_by(Segment) %>% summarise(Sales = sum(Sales),Profit= sum(Profit), .gr
category_grp <- data_df %>% group_by(Category) %>% summarise(Sales = sum(Sales),Profit= sum(Profit), .g
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 x 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
```

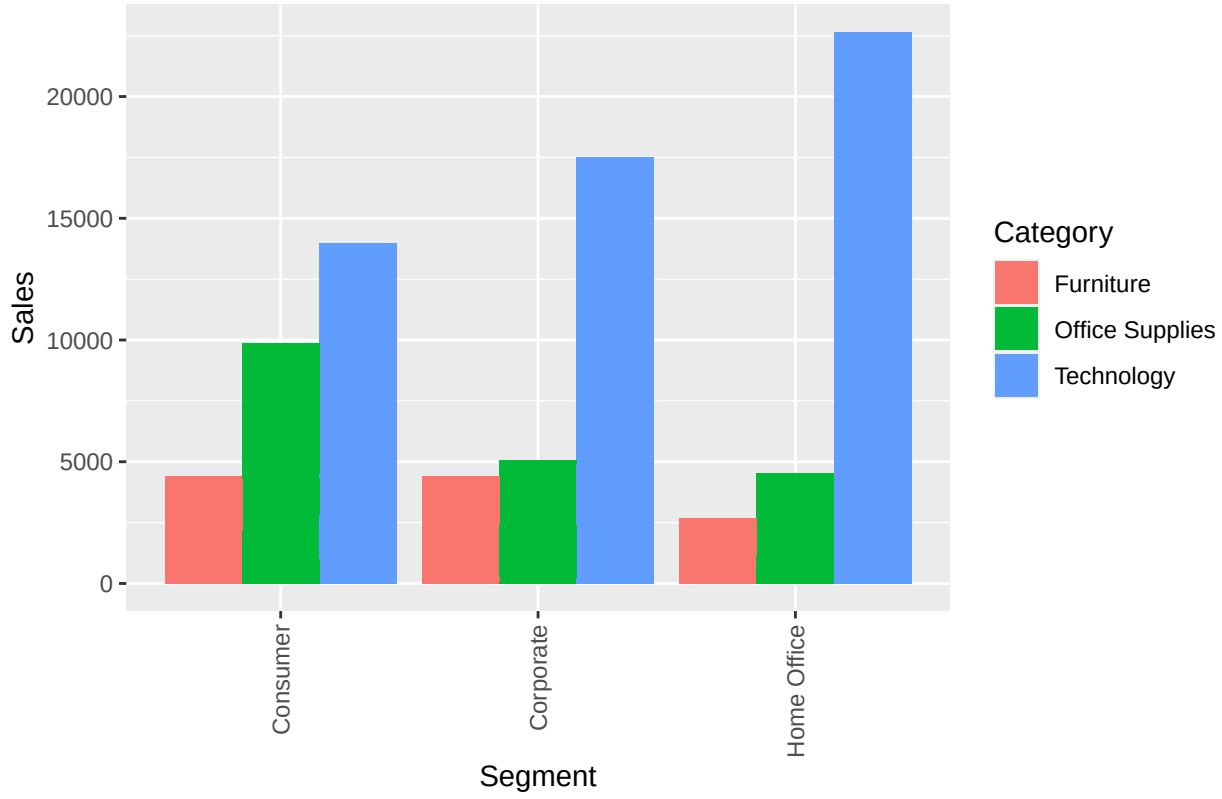
```r
category_grp<-category_grp[order(category_grp$Sales),]
category_grp$perc_sales <- 100*category_grp$Sales/sum(category_grp$Sales)
category_grp<-category_grp[order(category_grp$Profit),]
category_grp$perc_profit <- 100*category_grp$Profit/sum(category_grp$Profit)

# 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(
```

Figure 4.1 :− Profit in each segment category wise

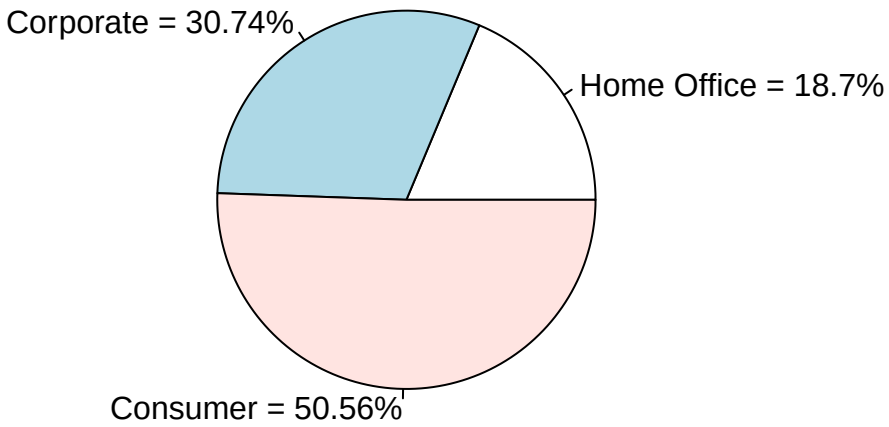



```
ggplot(data_df, aes(x=Segment,y=Sales,  fill=Category)) +
    geom_bar(stat='identity', position='dodge')+
  ggtitle(" Figure 4.2 :- Sales in each segment category wise") +      theme(axis.text.x=element_text(a
```

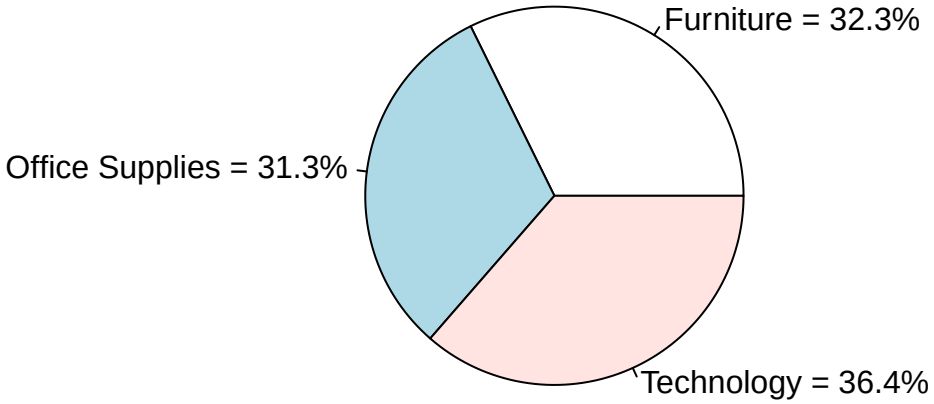
Figure 4.2 :− Sales in each segment category wise



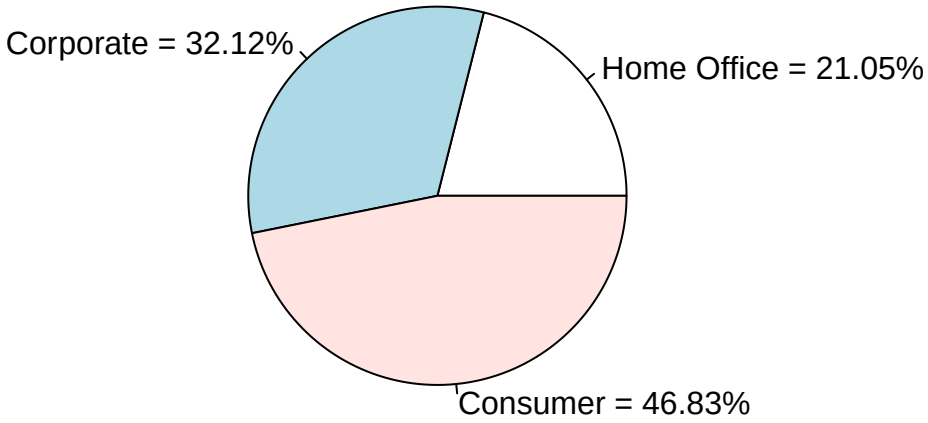

```
# library(lessR)
# # Categorical data
#
#
# cols <-  hcl.colors(length(unique(category_grp$Category)), "Fall")
#
# PieChart(Category, data = data_df, hole = 0,
#          fill = cols,
#          labels_cex = 0.6)
# par(mfrow=c(2,2))

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)
```

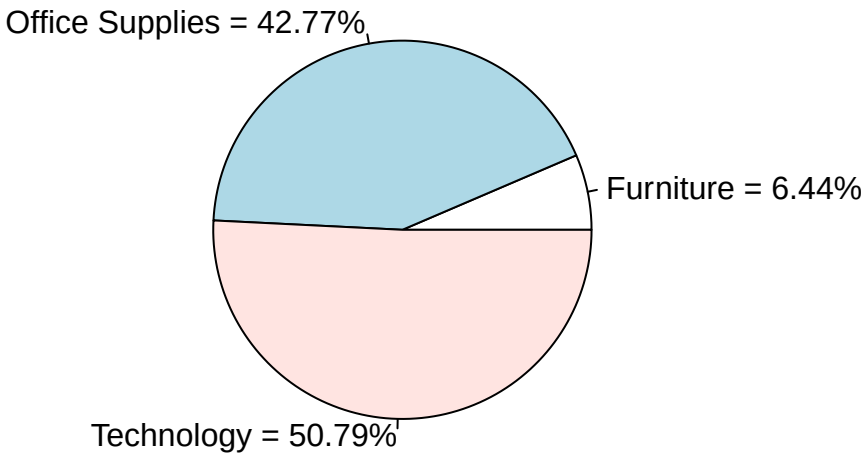
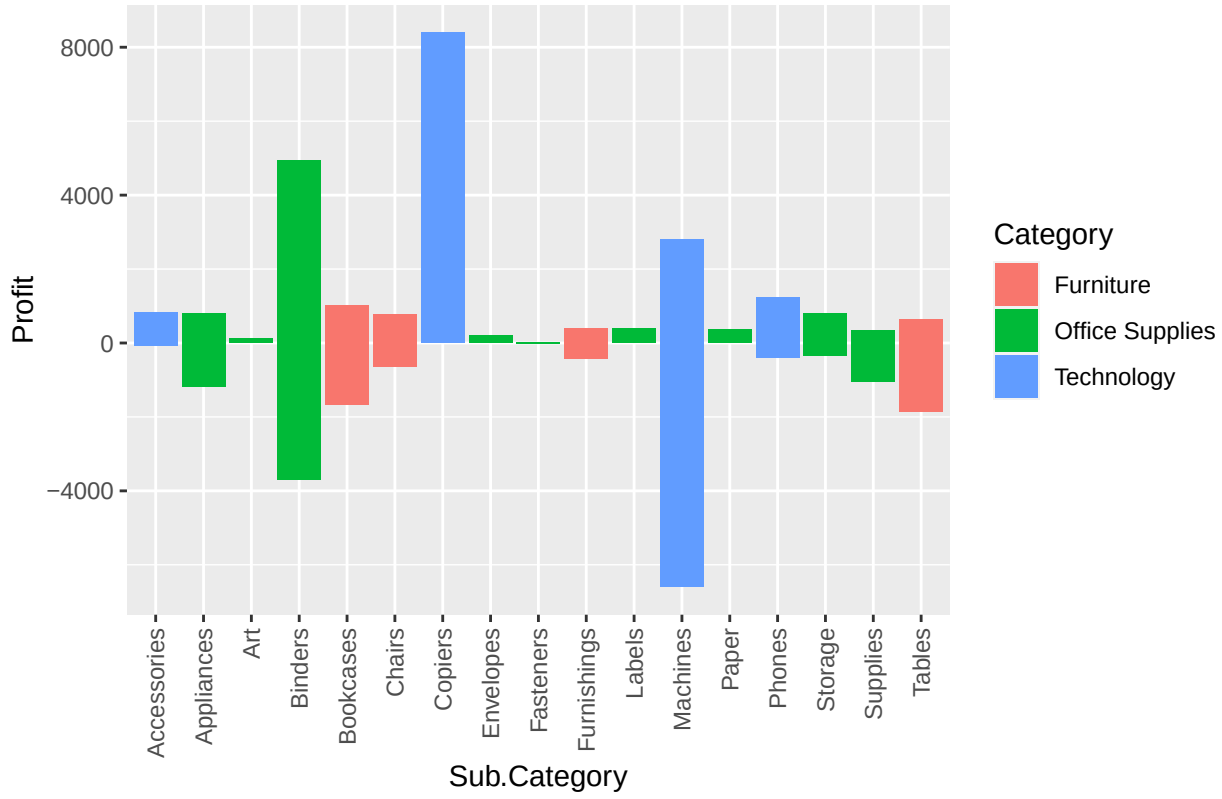


Figure 4:- Sales & Profit in each segment category wise

- 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(a
```

Figure 5 :− Profit in each segment sub.category wise




```
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 x 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 x 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 $Technology > OfficeSupplies > Furniture$
  - 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 x 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 x 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 x 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), ]
```

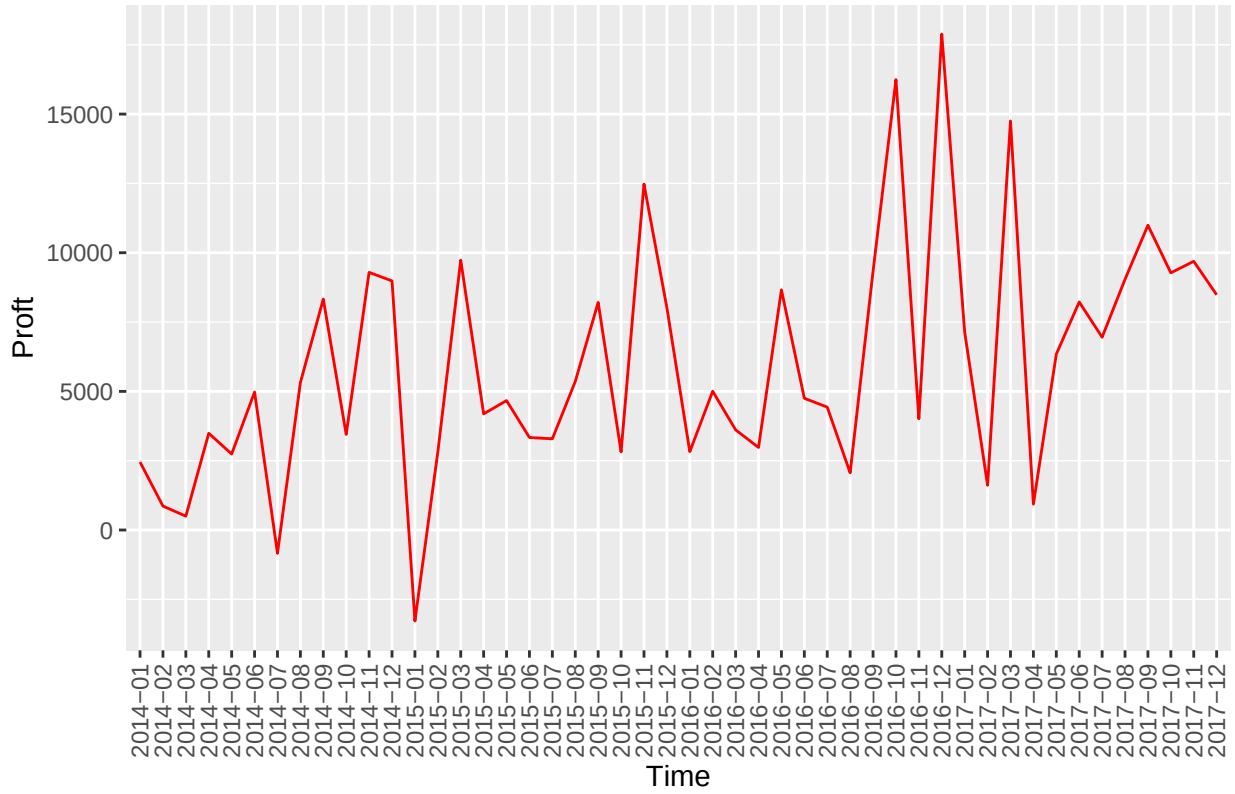
```
## # A tibble: 48 x 3
##    Month_Yr Profit   Sales
##    <chr>     <dbl>   <dbl>
##  1 2016-12  17885.  96999.
##  2 2016-10  16243.  59688.
##  3 2017-03  14752.  58872.
##  4 2015-11  12475.  75973.
##  5 2017-09  10992.  87867.
##  6 2015-03   9732.  38726.
##  7 2017-11   9690. 118448.
##  8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

```
monthly_yr_sales
```

```
## # A tibble: 48 x 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
```

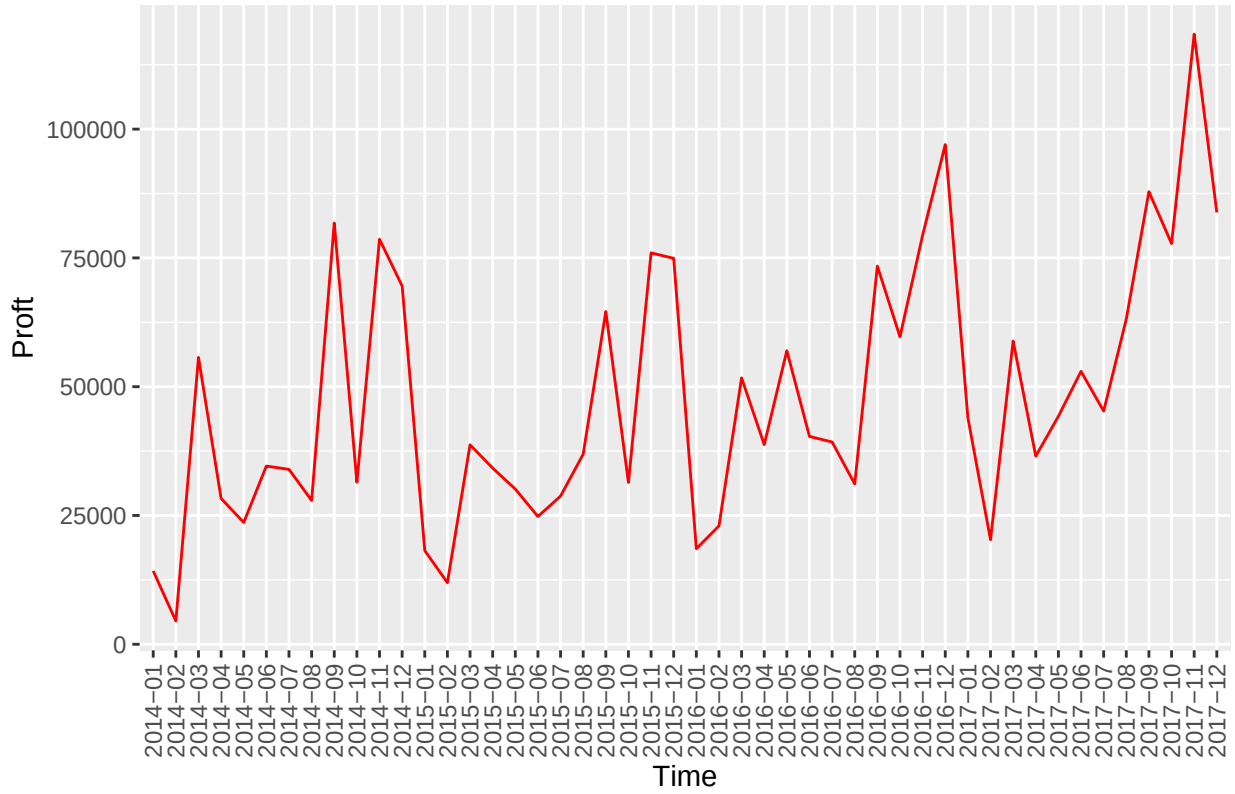
```
ggplot(data = monthly_yr_sales[order(monthly_yr_sales$Month_Yr),], aes(x = Month_Yr, y = Profit, group =
  geom_line(color = "red")+
  ggtitle(" Figure 5.2 :- Profit over Months from 2014 - 2018") +
  xlab("Time") + ylab("Proft")+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```

Figure 5.2 :− Profit over Months from 2014 − 2018

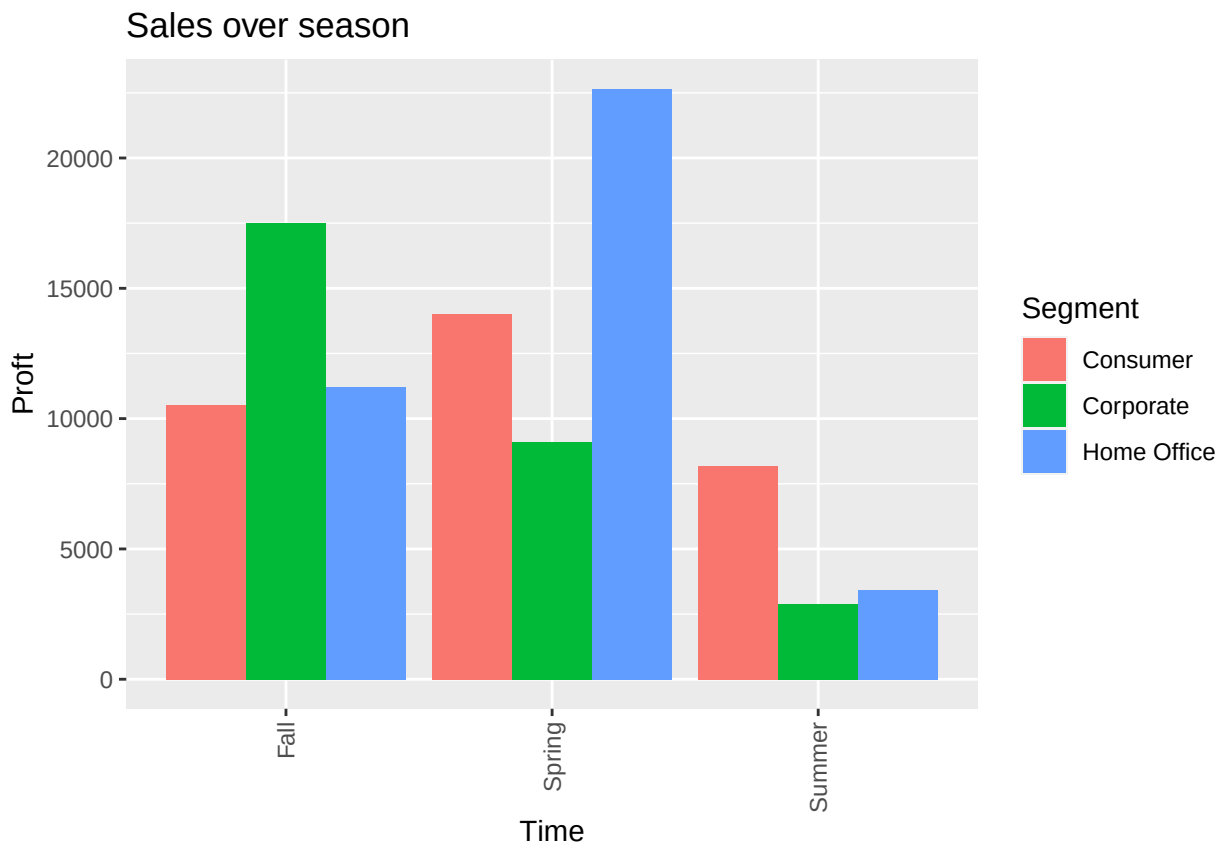



```
ggplot(data = monthly_yr_sales[order(monthly_yr_sales$Month_Yr),], aes(x = Month_Yr, y = Sales, group =
  geom_line(color = "red")+
  ggtitle(" Figure 5.3:- Sales over Months from 2014 - 2018") +
  xlab("Time") + ylab("Proft")+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```

Figure 5.3:− Sales over Months from 2014 − 2018




```
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))
```

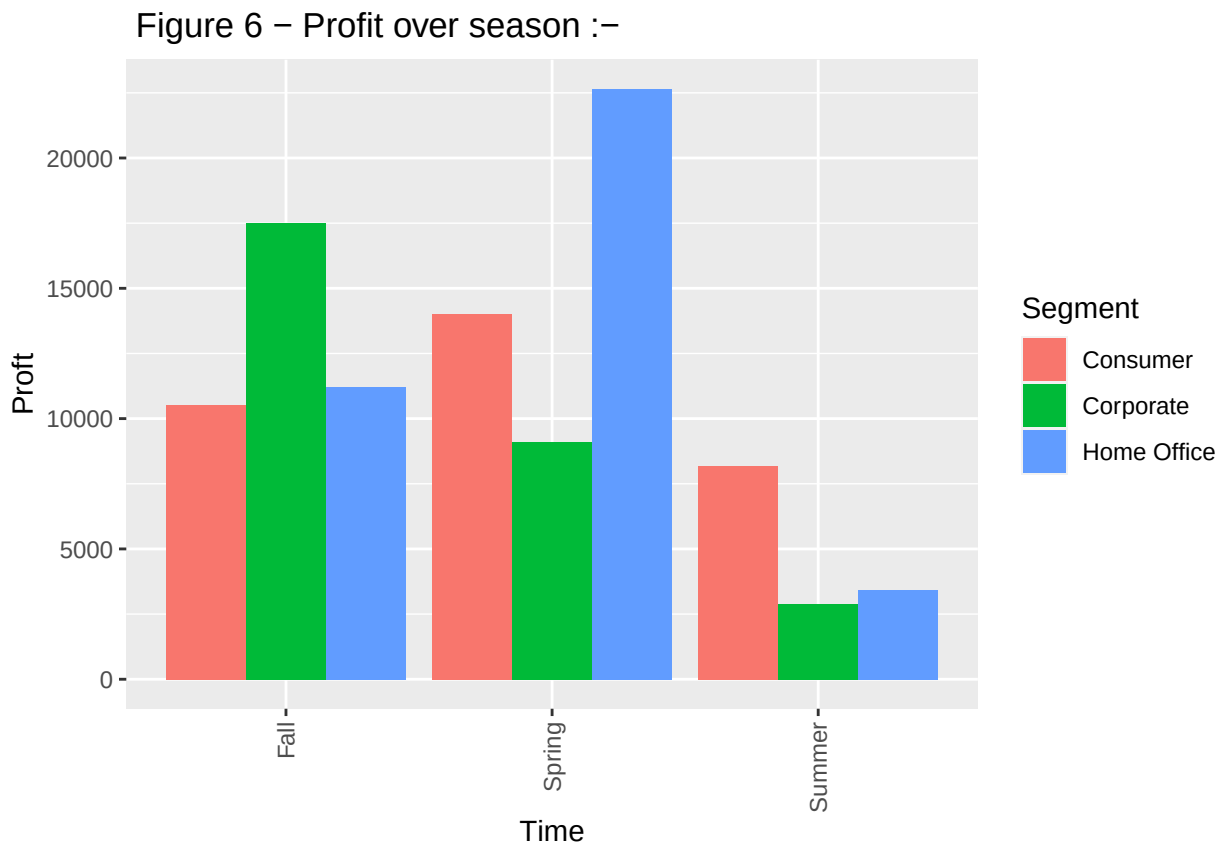
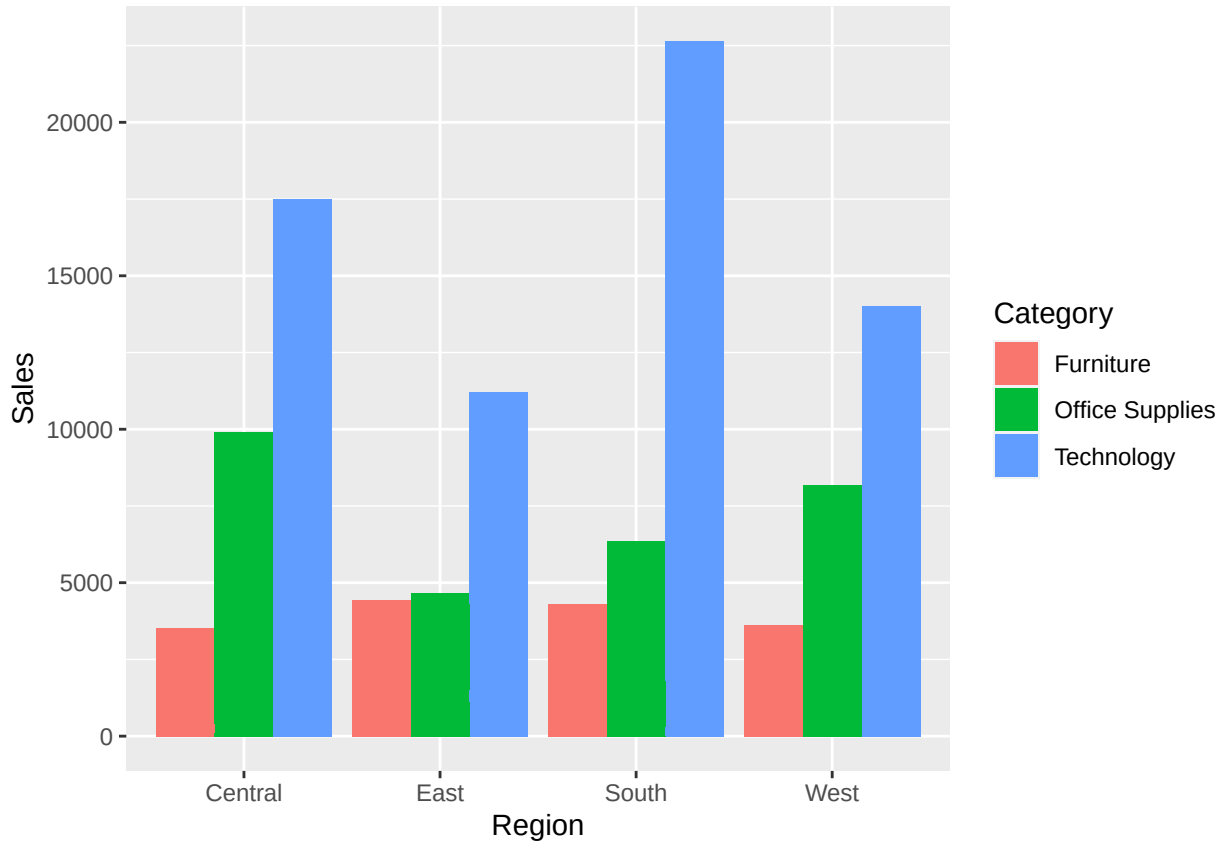


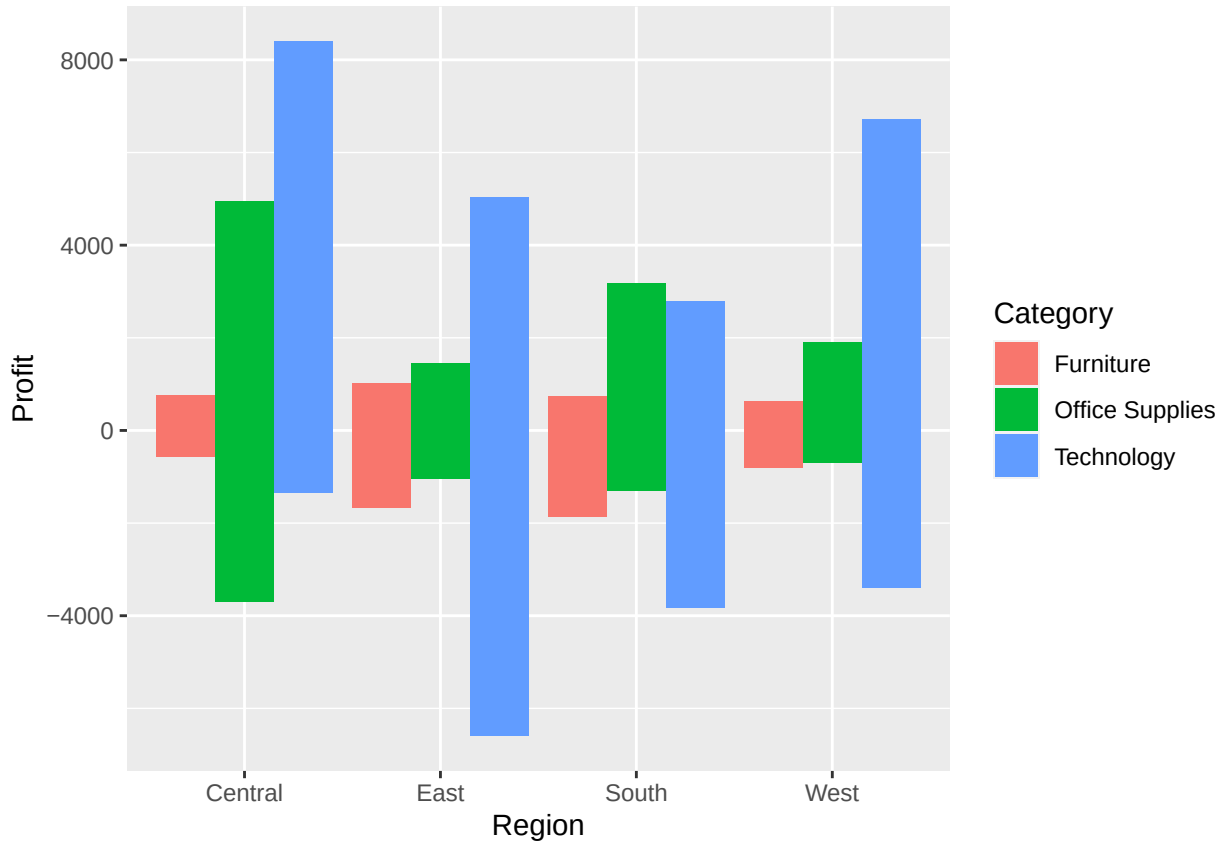
Figure 5.2 and 6:- Profit from 2014 to 2018 and Sales, Profit Monthly

- 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 $Fall > Spring > Summer$
- 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
```

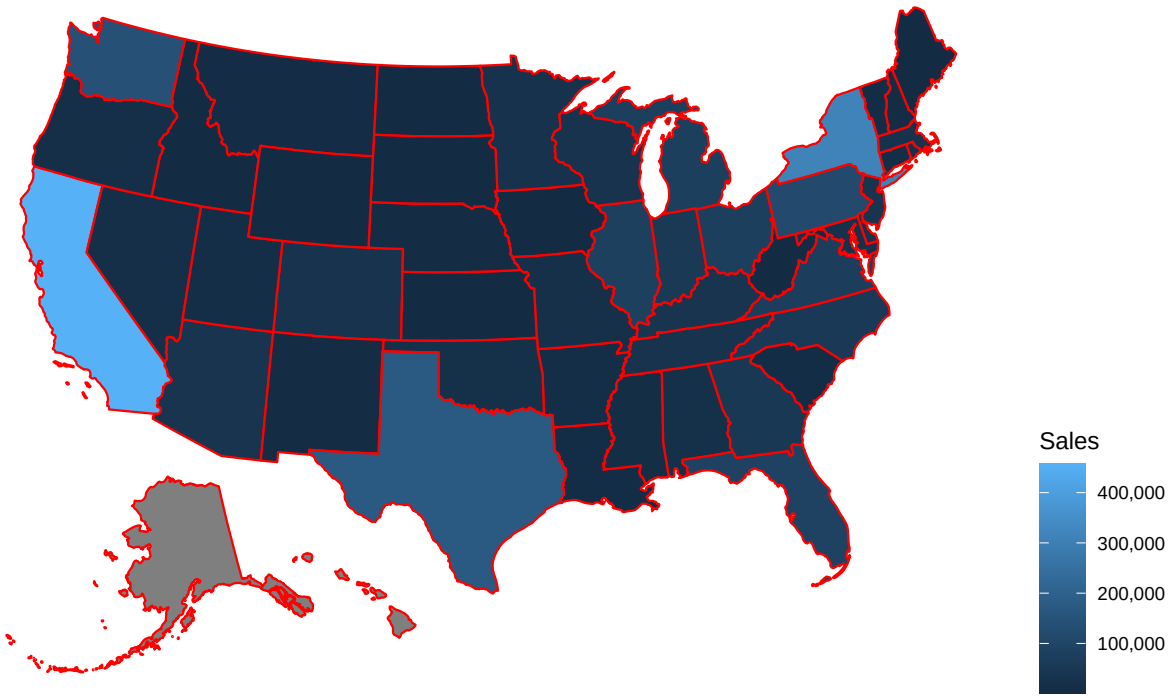
```
state_sales <- data_df %>%
  group_by(State) %>%
           summarize(Profit = sum(Profit),Sales= sum(Sales))
state_sales[order(state_sales$Profit,state_sales$Sales,decreasing = TRUE), ]
```

```
## # A tibble: 49 x 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
```

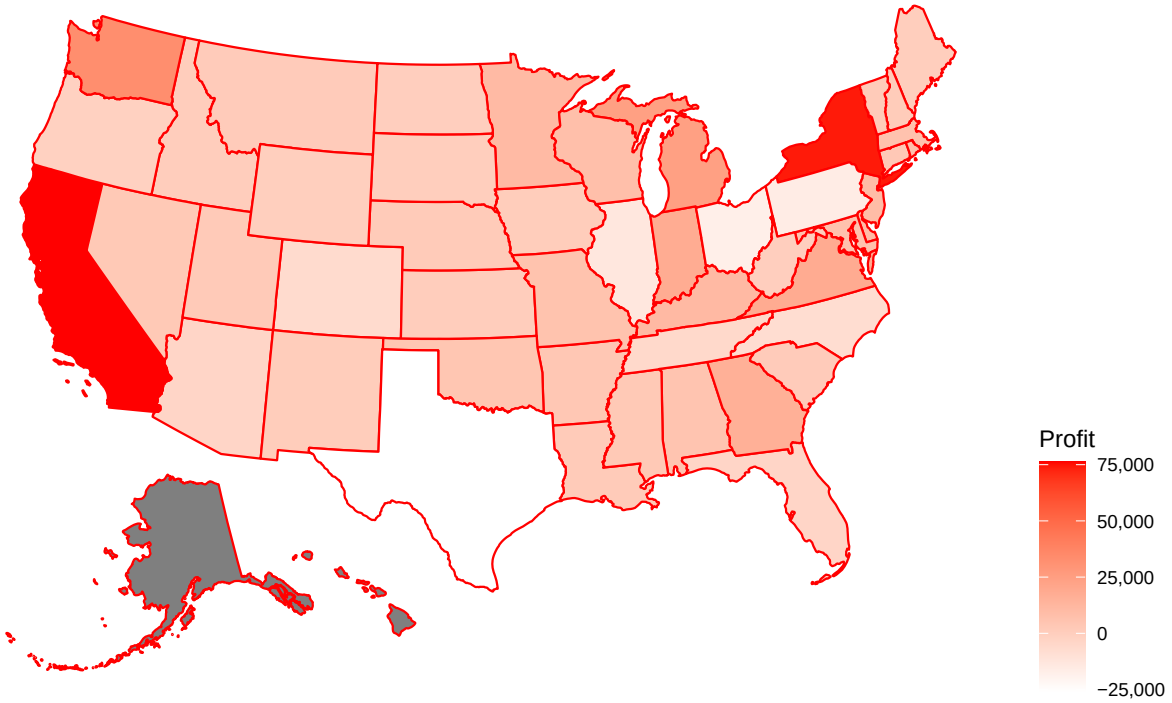
```
statepop1 <- usmap::statepop
statepop1 <- statepop1[c('fips', 'full')] %>%
  rename(State = full )

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 x 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
```

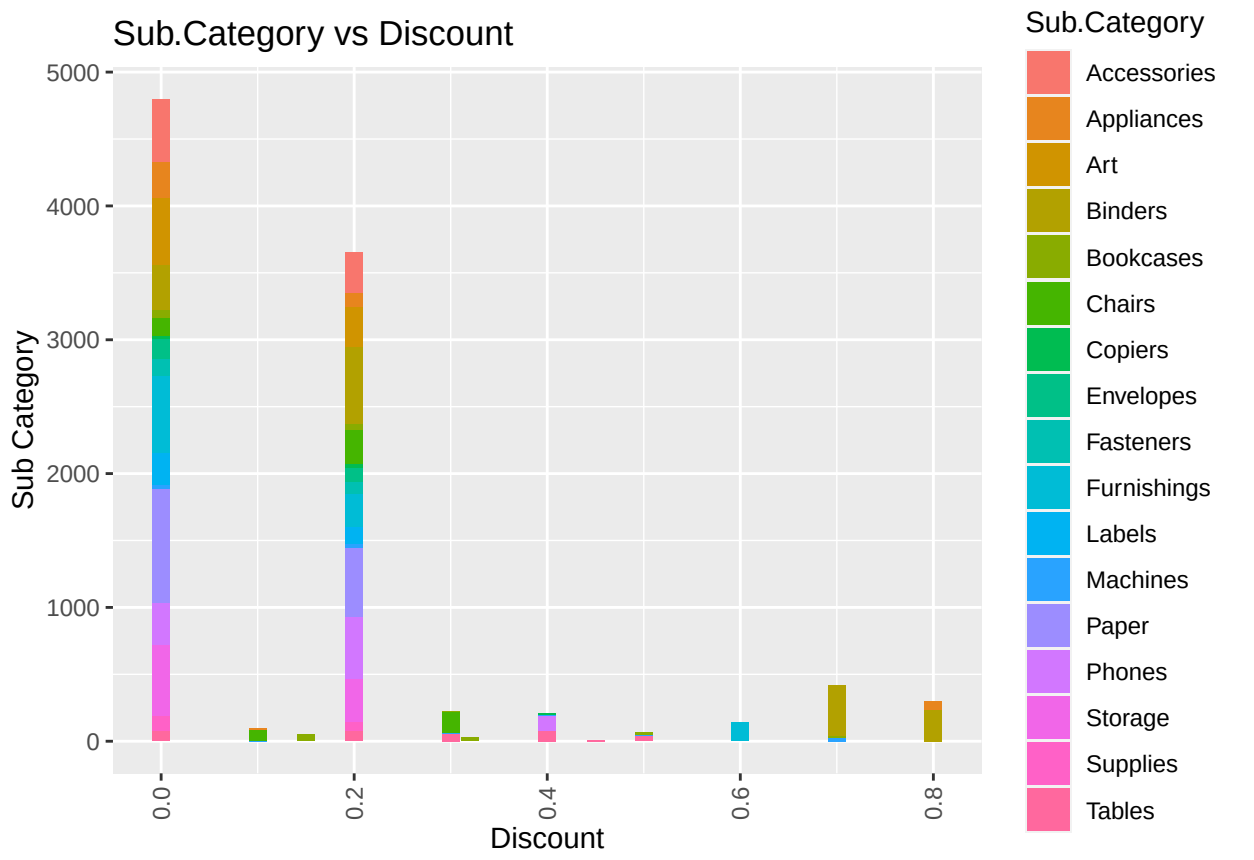
```
seg_sea_df <- data_df %>%
  group_by(season,Sub.Category) %>%
    summarize(Discount = mean(Discount), Sales = sum(Sales),Profit = sum(Profit))
```

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

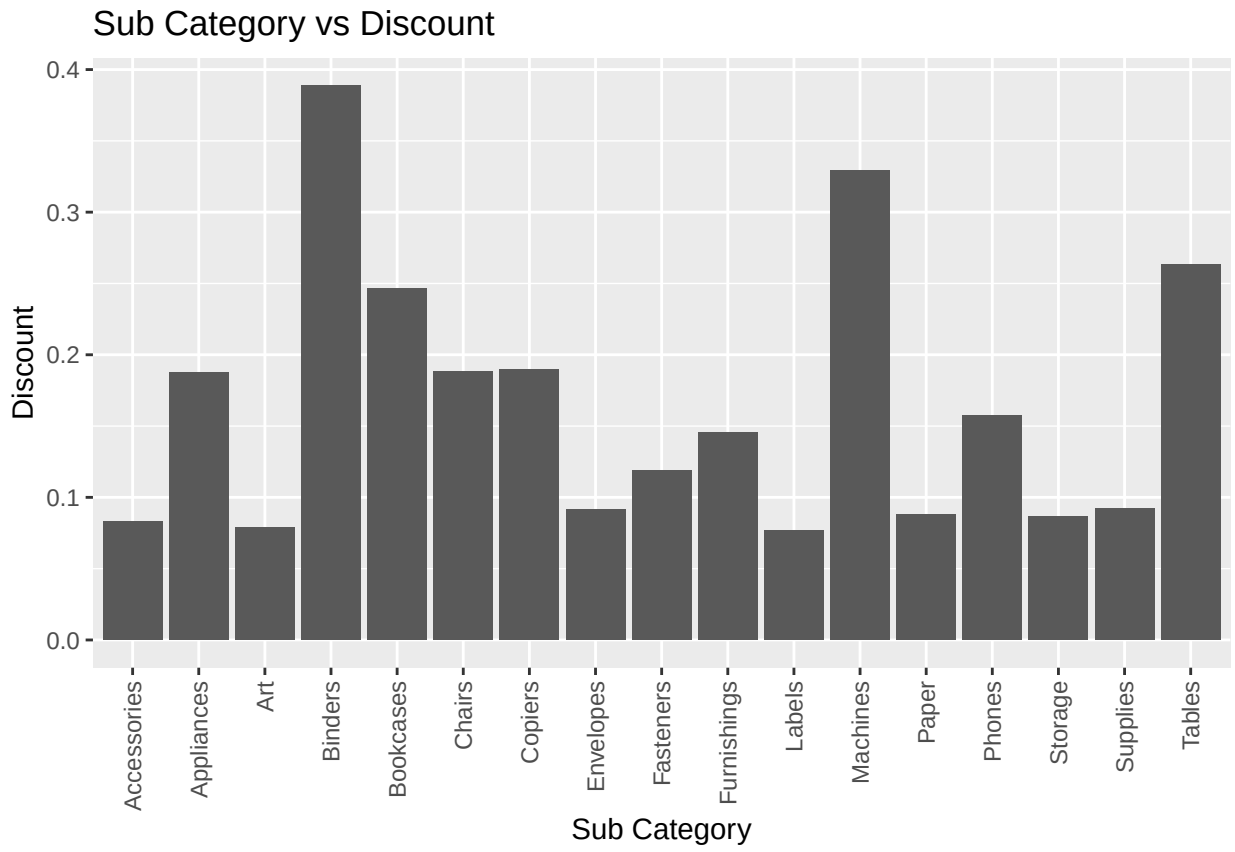
```
seg_sea_df
```

```
## # A tibble: 51 x 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
```

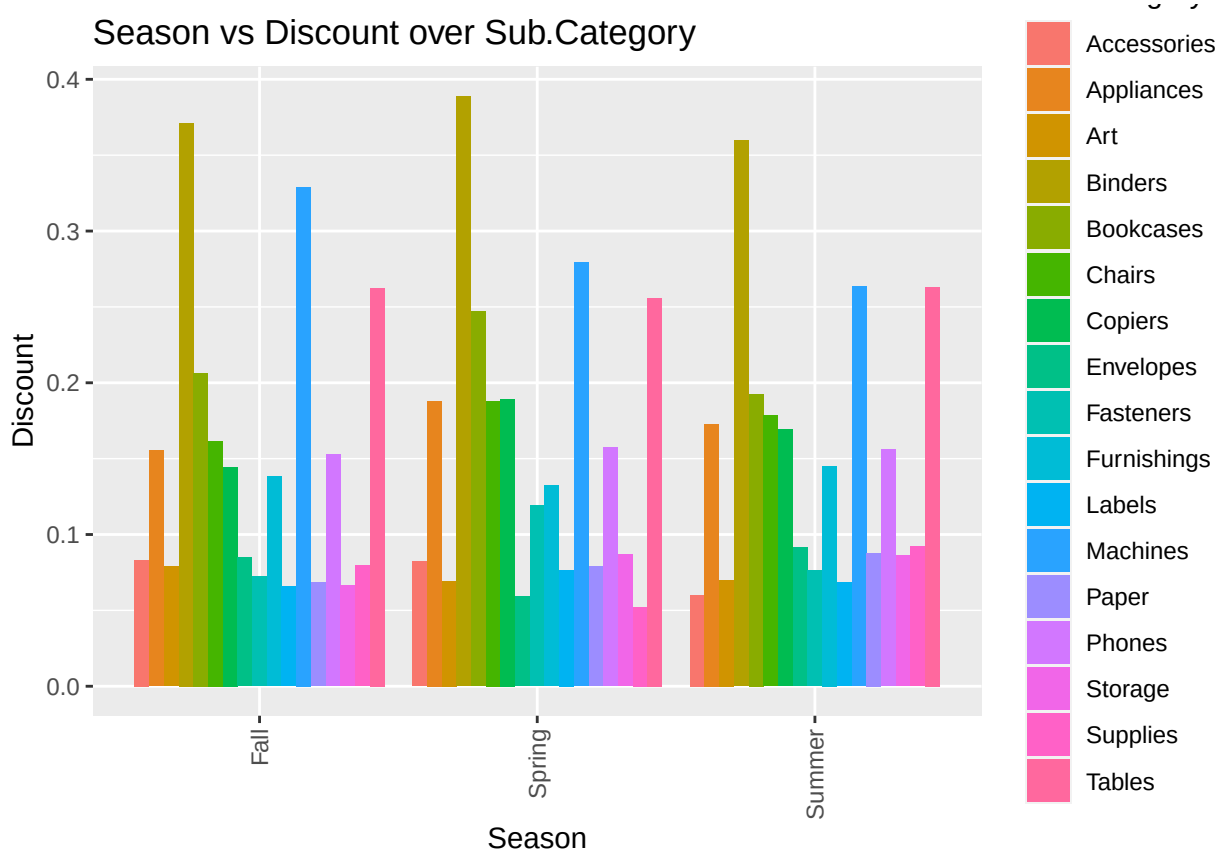
```
ggplot(data = data_df, aes(x = Discount , fill = Sub.Category )) +
  geom_bar( position='stack')+
  ggtitle("Sub.Category vs Discount ") +
  ylab("Sub Category") + xlab("Discount")+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```




```
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 x 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))
```

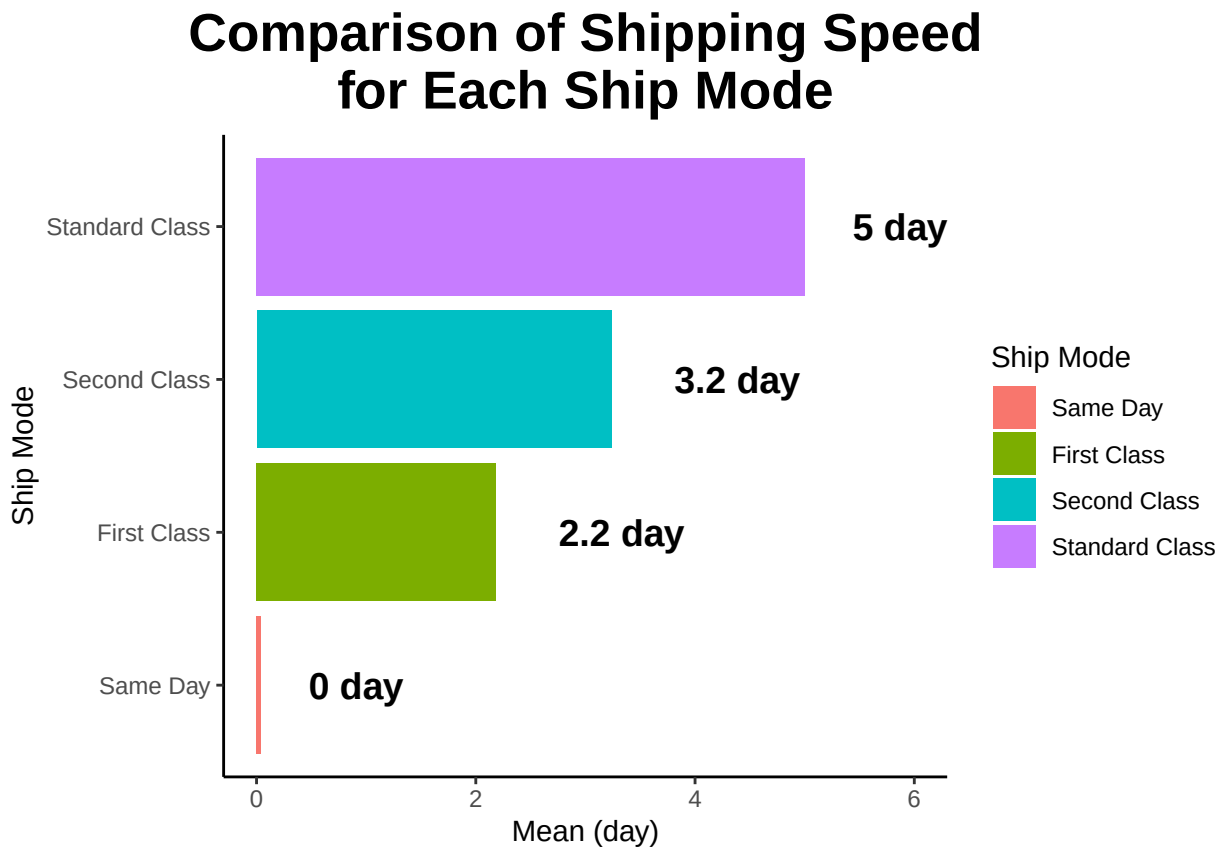


Figure 9

- 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'
```

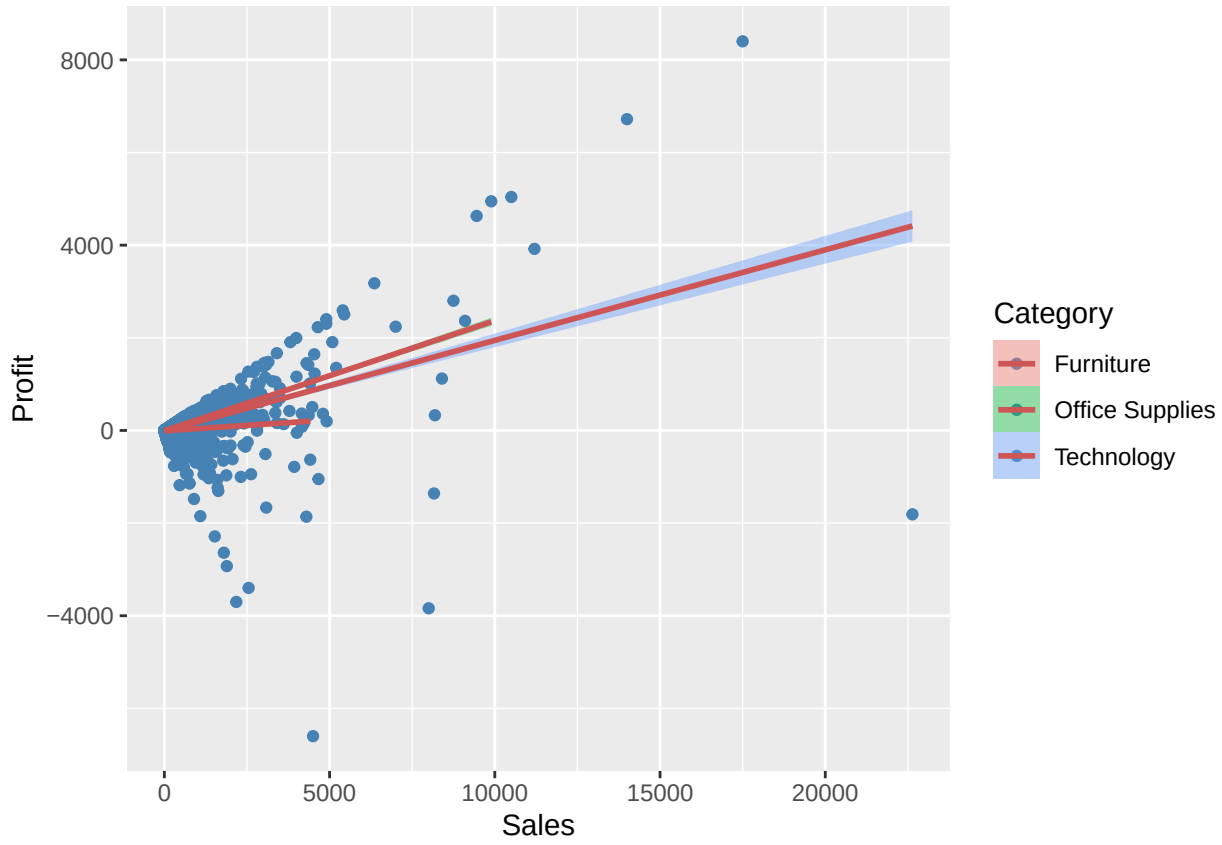


Figure 10

```
# 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

# wordcloud(words= data_df$Product.Name, min.freq =1, max.words=100, random.order=FALSE,rot.per=0.35,co
```

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] "39.299332 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] 1.436192
```

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 = 594, df = 582, p-value = 0.3561
```

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] "176.66358 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] -0.6035208
```

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 = 60, df = 54, p-value = 0.2673
```

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)
```

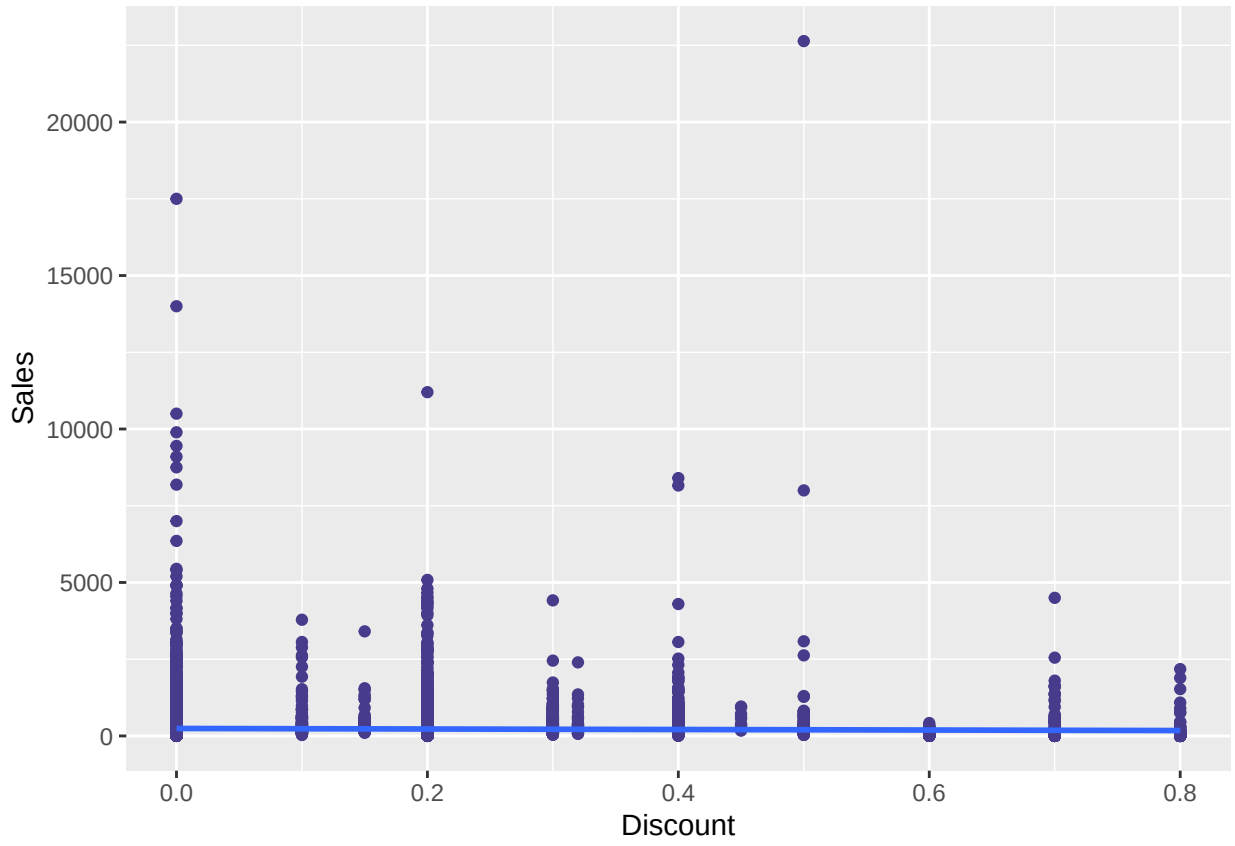
```
##
##  Welch Two Sample t-test
##
## data:  sample_df$profit_cat and sample_df$Discount
## t = 13.792, df = 157.84, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5331185 0.7113260
## sample estimates:
## mean of x mean of y
## 0.8181818 0.1959596
```

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

```
data_df[c('Discount', 'Sales')] %>%
  ggplot(aes(x = Discount, y = Sales)) +
  geom_point( color= '#473c8b') +
  geom_smooth(method = lm)
```

LM-1 Linear Regression for Q4

```
## `geom_smooth()` using formula 'y ~ x'
```

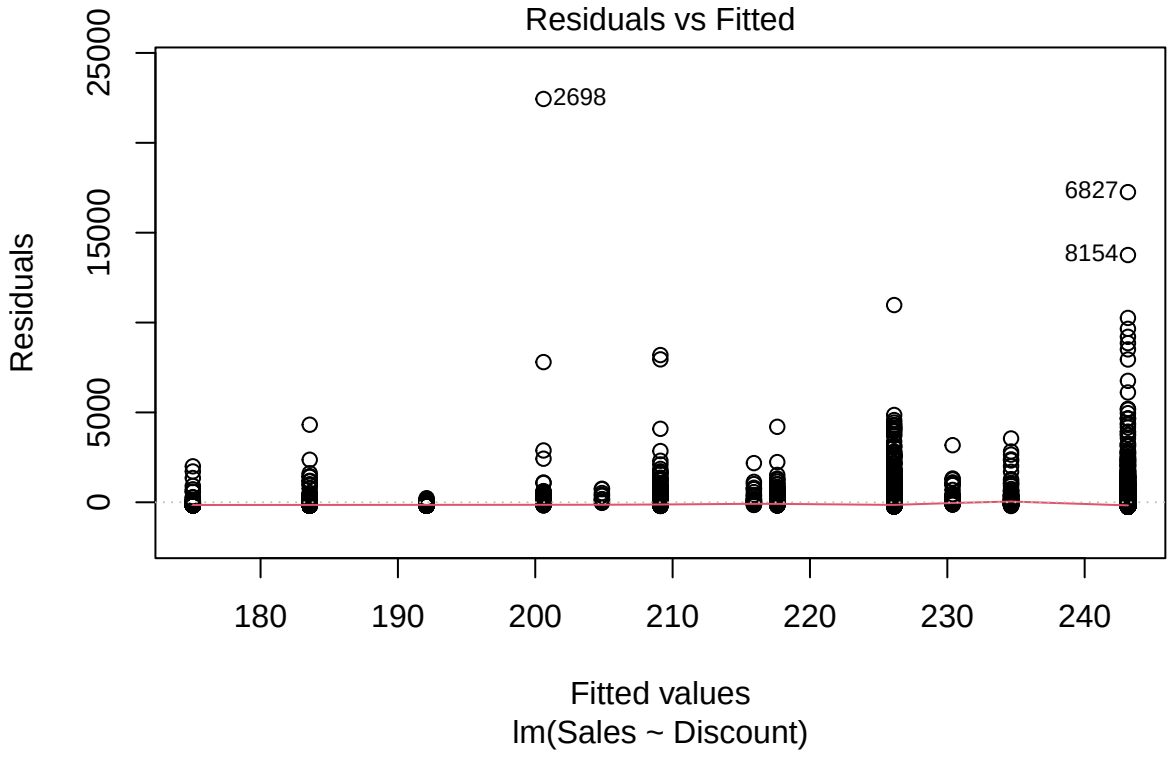


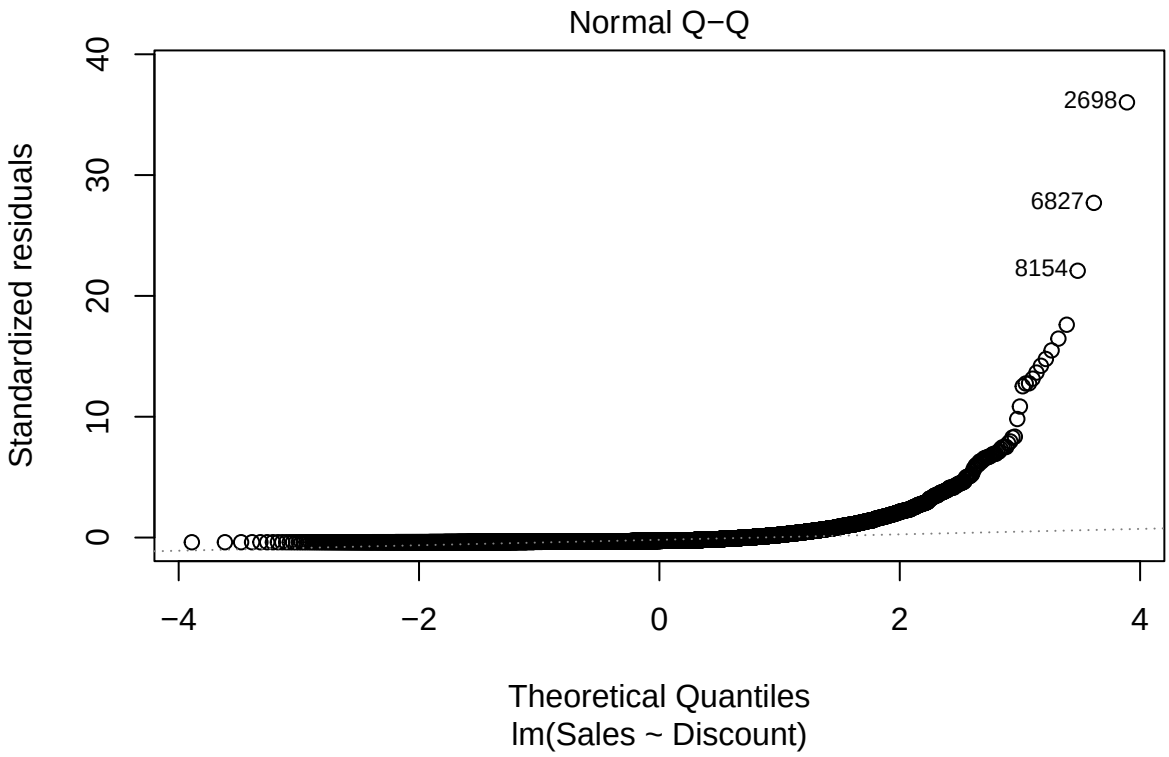

```
model1 <- lm(Sales ~ Discount, data = data_df)
summary(model1)
```

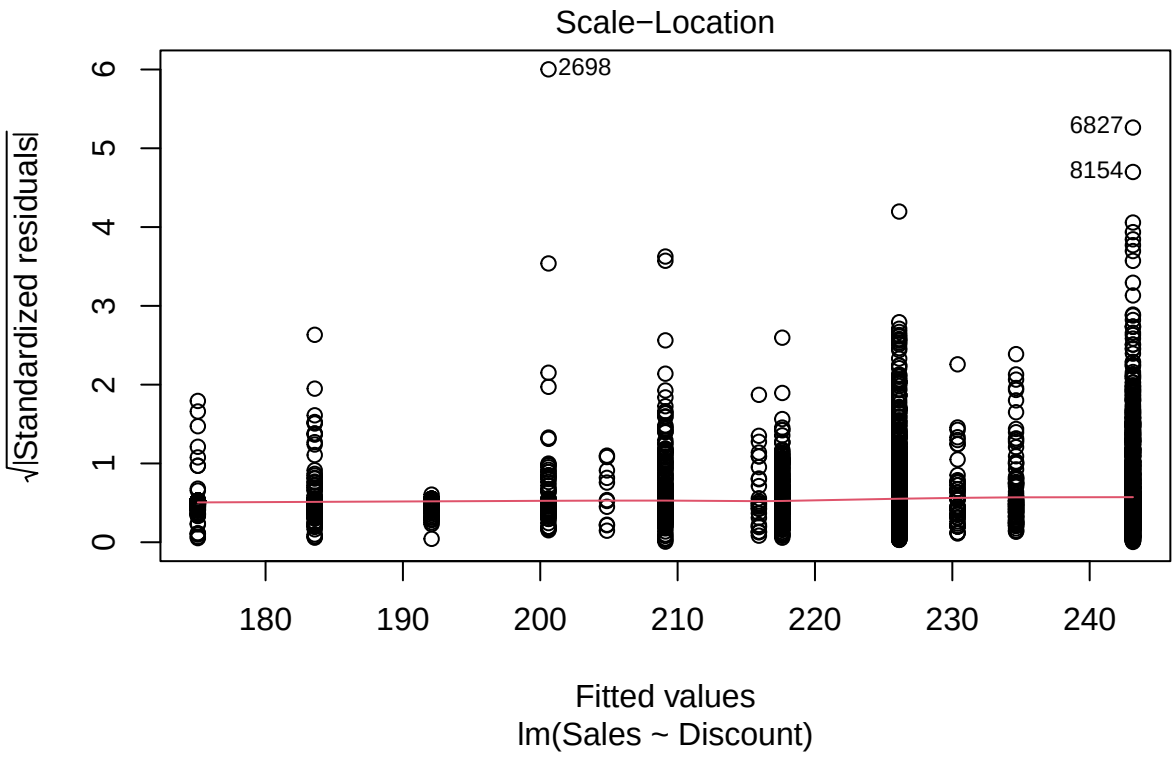
```
##
## Call:
## lm(formula = Sales ~ Discount, data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -242.2  -211.8  -170.9   -22.0 22437.9
```

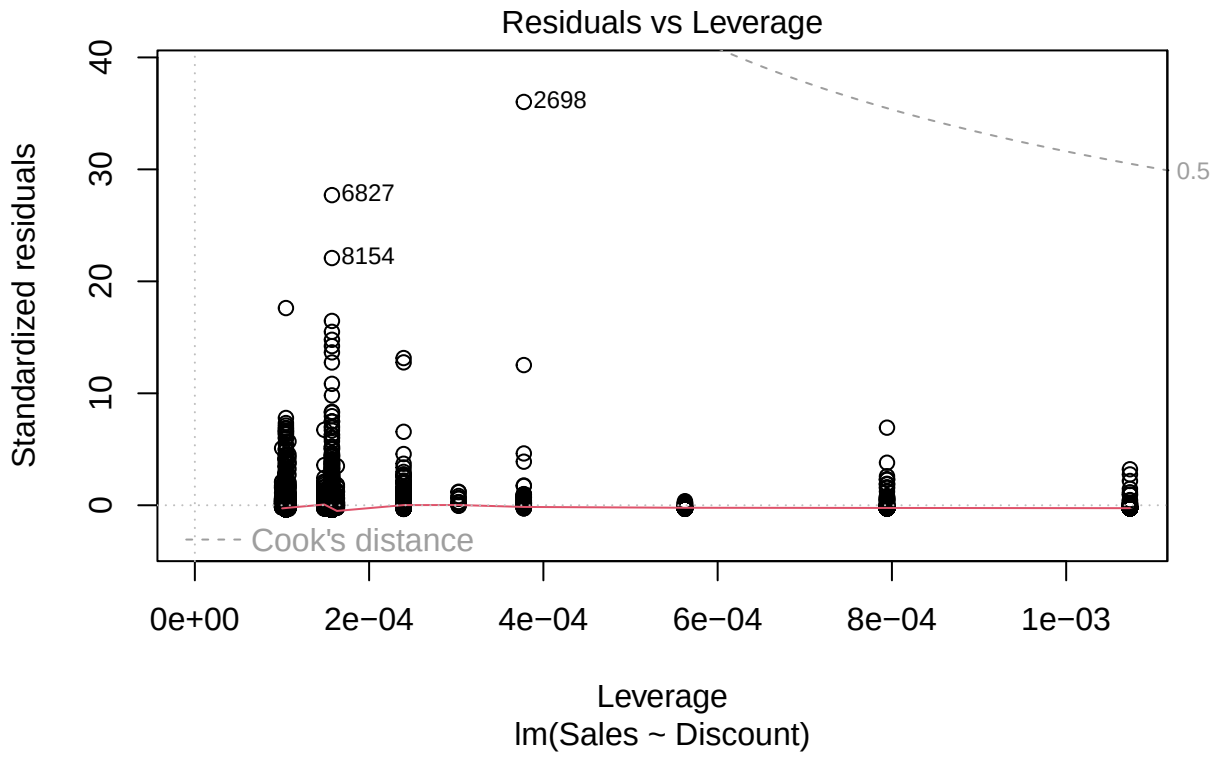
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  243.151      7.815  31.113  < 2e-16 ***
## Discount     -85.101     30.188  -2.819  0.00483 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 623 on 9992 degrees of freedom
## Multiple R-squared:  0.0007947,  Adjusted R-squared:  0.0006947
## F-statistic: 7.947 on 1 and 9992 DF,  p-value: 0.004827
```

```
plot(model1)
```

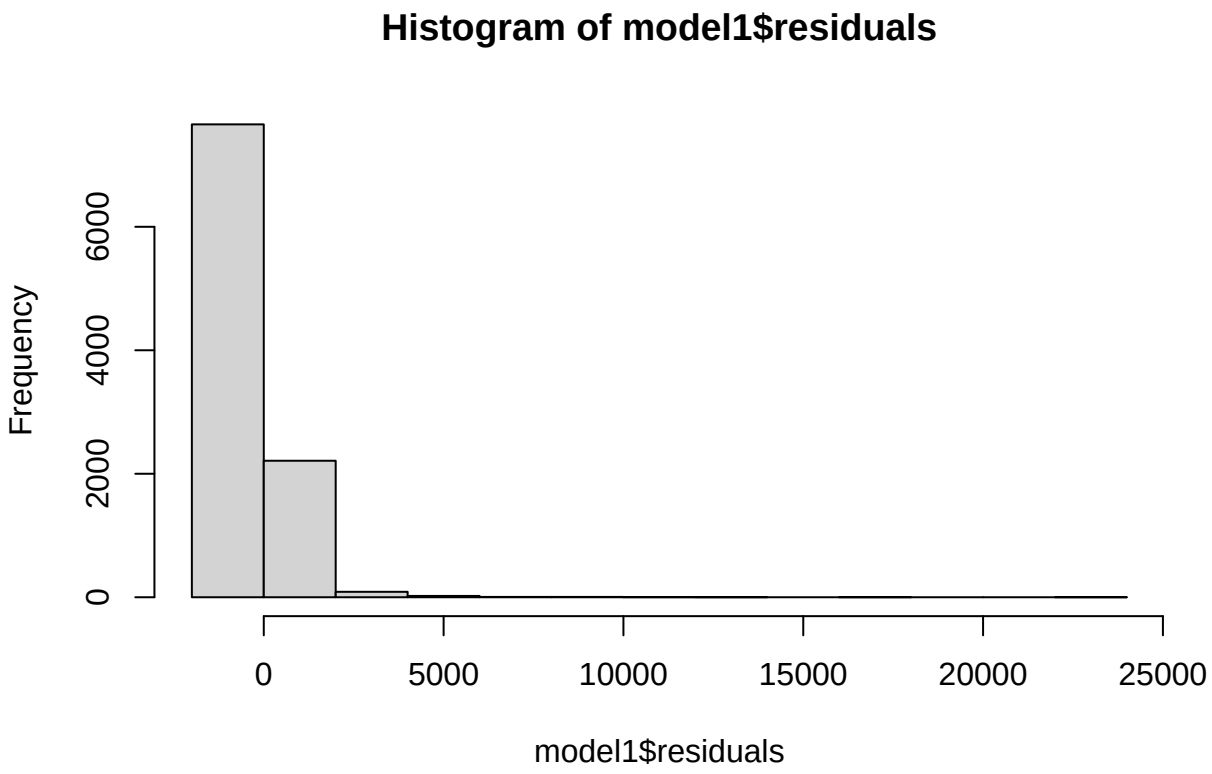



Normal Q−Q
Standardized residuals
0
10
20
30
40
2698
6827
8154
−4
−2
0
2
4
Theoretical Quantiles
lm(Sales ~ Discount)


Scale−Location
√|Standardized residuals|
0
1
2
3
4
5
6
2698
6827
8154
180
190
200
210
220
230
240
Fitted values
lm(Sales ~ Discount)




```
hist(model1$residuals)
```



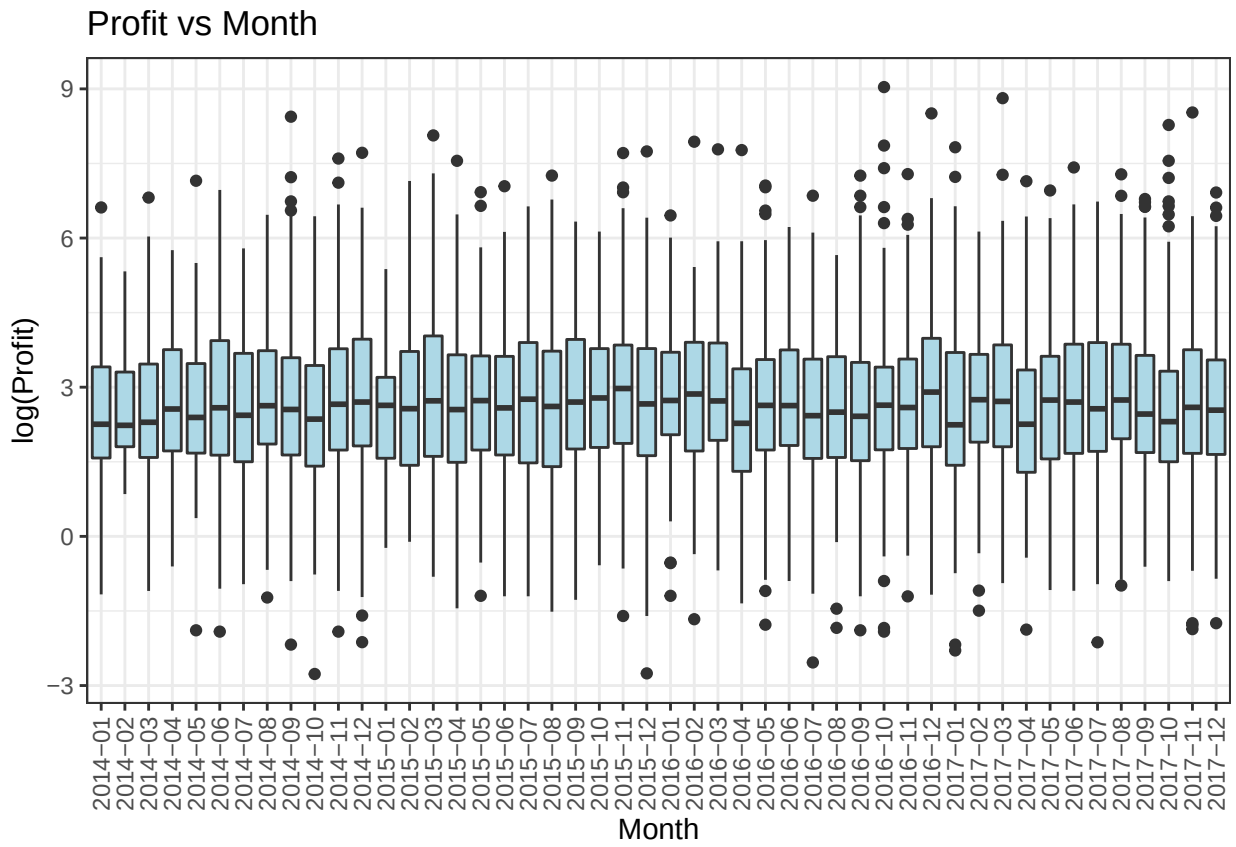
LM-1 - Sales vs Discount Q4

- Low R-squared value and high Std.error indicates that discounts do not cause higher sales.

```
# head(data_df)
data_df%>%
  ggplot(aes(Month_Yr,log(Profit)))+
  geom_boxplot(fill='LightBlue')+
  theme_bw()+
  labs(title='Profit vs Month',x='Month')+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```

LM-2 Month vs Profit

```
## Warning in log(Profit): NaNs produced

## Warning in log(Profit): NaNs produced

## Warning: Removed 1936 rows containing non-finite values (stat_boxplot).
```




```
model2 <- lm(Profit ~ Month + season + Discount, data = data_df)
summary(model2)
```

```
##
## Call:
## lm(formula = Profit ~ Month + season + Discount, data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6493.0   -54.8   -15.9     9.5  8332.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65.9756     9.5044   6.942 4.12e-12 ***
## Month           0.1317     0.9339   0.141    0.888
## seasonSpring    3.4859     7.7529   0.450    0.653
## seasonSummer   -0.7783     6.2459  -0.125    0.901
## Discount     -249.1194    11.0778 -22.488  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 228.6 on 9989 degrees of freedom
## Multiple R-squared:  0.04821,    Adjusted R-squared:  0.04783
## F-statistic: 126.5 on 4 and 9989 DF,  p-value: < 2.2e-16
```

```
# plot(model1)
# hist(model1$residuals)
```

LM 2 Profit vs Month

- From above plot we can say that Profit doesn't change much with month or season
- Also, for 5% significance, t-value obtained for season, month is less than 1.90. So we accept null hypothesis that month and season does not relate profit
- Another way of saying that month and season does not explain variance of the data is throughh R-squared values - the higher the R-2 value the more variance explained by variables

```
# head(data_df)
model3 <- lm(Profit ~ Category+Sub.Category, data = data_df)
summary(model3)
```

LM 3 Profit vs Category, Sub Category

```
##
## Call:
## lm(formula = Profit ~ Category + Sub.Category, data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6629.4   -24.4   -10.2     7.0  7582.1
##
## Coefficients: (2 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -55.566     12.548  -4.428 9.60e-06 ***
## CategoryOffice Supplies    49.307     20.538   2.401   0.0164 *
## CategoryTechnology        109.678     14.909   7.357 2.03e-13 ***
## Sub.CategoryAppliances     45.181     19.291   2.342   0.0192 *
## Sub.CategoryArt            14.459     18.096   0.799   0.4243
## Sub.CategoryBinders        26.102     17.243   1.514   0.1301
## Sub.CategoryBookcases      40.335     19.436   2.075   0.0380 *
## Sub.CategoryChairs         98.662     15.455   6.384 1.80e-10 ***
## Sub.CategoryCopiers       763.797     28.345  26.946  < 2e-16 ***
## Sub.CategoryEnvelopes      33.676     21.497   1.567   0.1172
## Sub.CategoryFasteners      10.634     22.267   0.478   0.6330
## Sub.CategoryFurnishings    69.212     14.489   4.777 1.81e-06 ***
## Sub.CategoryLabels         21.495     20.059   1.072   0.2839
## Sub.CategoryMachines      -24.679     22.396  -1.102   0.2705
## Sub.CategoryPaper          31.115     17.350   1.793   0.0729 .
## Sub.CategoryPhones         -4.038     11.014  -0.367   0.7139
## Sub.CategoryStorage        31.411     17.992   1.746   0.0809 .
## Sub.CategorySupplies           NA         NA      NA       NA
## Sub.CategoryTables             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 224.1 on 9977 degrees of freedom
## Multiple R-squared:  0.0862, Adjusted R-squared:  0.08473
## F-statistic: 58.82 on 16 and 9977 DF,  p-value: < 2.2e-16
```

```
# plot(model3)
# hist(model3$residuals)
```

Profit vs Category, SubCategory

- Copiers gets the most profits,

```
model4 <- lm(Profit ~ delivery_days+Discount+Quantity+Sales, data = data_df)
summary(model4)
```

Profit vs delivery_Days, discount, Quantity, Sales

```
##
## Call:
## lm(formula = Profit ~ delivery_days + Discount + Quantity + Sales,
##     data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7266.5   -23.8    -0.4    25.6  5229.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.525e+01  6.144e+00   5.737  9.9e-09 ***
## delivery_days -7.091e-02  1.144e+00  -0.062  0.95058
## Discount      -2.335e+02  9.687e+00 -24.100  < 2e-16 ***
## Quantity      -2.961e+00  9.173e-01  -3.228  0.00125 **
## Sales          1.800e-01  3.276e-03  54.954  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 199.8 on 9989 degrees of freedom
## Multiple R-squared:  0.2727, Adjusted R-squared:  0.2724
## F-statistic: 936.5 on 4 and 9989 DF,  p-value: < 2.2e-16
```

Profit vs Discount+ Delivery_days+ Sales + Quantity

- At 5% significance, value obtained above is greater than 2.68 for Discount, Sales, Category, Quantity.
- Hence, we reject null hypothesis and saying that regression line fitted to data is significant
- Also Pr(>|t|) for delivery_days is sufficiently high, we can reject alternate hypothesis that coefficient for delivery_days is 0 [lesser the probability, the greater the evidence we can reject null hypothesis that the coefficient is 0]

```
state_sales[order(state_sales$Profit,state_sales$Sales,decreasing = FALSE), ]
```

```
## # A tibble: 49 x 3
##    State          Profit   Sales
##    <chr>           <dbl>   <dbl>
##  1 Texas          -25729. 170188.
##  2 Ohio           -16971.  78258.
##  3 Pennsylvania   -15560. 116512.
##  4 Illinois       -12608.  80166.
##  5 North Carolina  -7491.  55603.
##  6 Colorado        -6528.  32108.
##  7 Tennessee       -5342.  30662.
##  8 Arizona         -3428.  35282.
##  9 Florida         -3399.  89474.
## 10 Oregon          -1190.  17431.
## # ... with 39 more rows
```

```
head(data_df)
```

```
##   Order.Date  Ship.Date      Ship.Mode   Segment       Country            City
## 1 11-08-2016 11-11-2016   Second Class  Consumer United States       Henderson
## 2 11-08-2016 11-11-2016   Second Class  Consumer United States       Henderson
## 3 06-12-2016 06-16-2016   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 06-09-2014 06-14-2014 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
##   delivery_days
## 1             3
## 2             3
## 3             4
## 4             7
```

```
## 5              7
## 6              5
```

Summary & Conclusion

- 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.
- After performing exhaustive and vast Exploratory Data Analysis we can conclude below points,
  - Most common shipping mode preferred by users is Standard Class, therefore sales generated is high for this class
  - Even though orders from Office Supplies category are more but Profit generated from Technology category is high.
    - Profitable Sub Category in Technology - Copiers, Accessories, Phones | Loss - Machines
    - Profitable Sub Category in Office Supplies - Blinders, Storage | Loss - Appliances,Supplies
    - All sub categories in Furniture incur losses than profits.
  - Seasonal Profits in respective segments
    - Spring - Home Office
    - Summer - Consumer
    - Fall - Corporate
  - State wise Profit
    - Highest for California, New York, Washington states
    - Lowest for Texas, Ohio, Pennsylvania, Illinois
  - Discounts vs Profits
    - 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.
- From Hypothesis testing and fitting the data using linear regression, we try to answer the questions proposed in the beginning and we can conclude few of the following.
  - Discounts does not necessarily imply profits, we can see from both hypothesis Q5 and LM-1 that there is no correlation between discounts and profits

Future Work

- Creating a machine learning model that can learn the data & improve superstore's Profits in all the categories, segments.
- With more information like customer related information in the data such as age group, customer reviews, satisfaction can help in building ML model.

Limitations

-> With more categorical and insufficient data, we were not able to build model properly. -> More numerical data that relates to Sales & Profits -> Staffing information also can help with cutoffs or maybe increase more production.

- References
  1. https://www.kaggle.com/datasets/vivek468/superstore-dataset-final)
  2. https://www.tableau.com/data-insights/dashboard-showcase/superstore)

3. https://boostedml.com/2019/06/linear-regression-in-r-interpreting-summarylm.html#t-value

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