EXPLORATORY DATA ANALYSIS : RETAIL

Objectives: To Perform Exploratory Data Analysis on dataset SampleSuperstore. As a business manage

Author: Irshad

Installing and loading necessary libraries

```
install.packages("tidyverse")
library(tidyverse)
install.packages("ggplot2")
library(ggplot2)
install.packages("skimr")
library(skimr)
install.packages("janitor")
library(janitor)
install.packages("dplyr")
library(dplyr)
```

Importing the dataset

```
Superstore <- read_csv("SampleSuperstore.csv") #load the data
```

Dataset features

```
head(Superstore) ## Prints the top 6 rows of data
```

```
## # A tibble: 6 x 13
##   `Ship Mode`  Segment  Country   City     State  `Postal Code` Region Category
##   <chr>        <chr>    <chr>     <chr>    <chr>          <dbl> <chr>  <chr>
## 1 Second Class Consumer United S~ Henders~ Kentu~         42420 South  Furniture
## 2 Second Class Consumer United S~ Henders~ Kentu~         42420 South  Furniture
## 3 Second Class Corpora~ United S~ Los Ang~ Calif~         90036 West   Office S~
## 4 Standard Cl~ Consumer United S~ Fort La~ Flori~         33311 South  Furniture
## 5 Standard Cl~ Consumer United S~ Fort La~ Flori~         33311 South  Office S~
## 6 Standard Cl~ Consumer United S~ Los Ang~ Calif~         90032 West   Furniture
## # ... with 5 more variables: Sub-Category <chr>, Sales <dbl>, Quantity <dbl>,
## #   Discount <dbl>, Profit <dbl>
```

Information about the dataset

```
skim_without_charts(Superstore)
```

Table 1: Data summary

Name	Superstore
Number of rows	9994
Number of columns	13
Column type frequency:	
character	8
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Shipmode	0	1	8	14	0	4	0
Segment	0	1	8	11	0	3	0
Country	0	1	13	13	0	1	0
City	0	1	4	17	0	531	0
State	0	1	4	20	0	49	0
Region	0	1	4	7	0	4	0
Category	0	1	9	15	0	3	0
Subcategory	0	1	3	11	0	17	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Postal Code	0	1	55190.38	32063.69	1040.00	23223.00	56430.50	90008.00	99301.00
Sales	0	1	229.86	623.25	0.44	17.28	54.49	209.94	22638.48
Quantity	0	1	3.79	2.23	1.00	2.00	3.00	5.00	14.00
Discount	0	1	0.16	0.21	0.00	0.00	0.20	0.20	0.80
Profit	0	1	28.66	234.26	-6599.98	1.73	8.67	29.36	8399.98

```
glimpse(Superstore)
```

```
## Rows: 9,994
## Columns: 13
## $ Shipmode      <chr> "Second Class", "Second Class", "Second Class", "Standar~
## $ Segment       <chr> "Consumer", "Consumer", "Corporate", "Consumer", "Consum~
## $ Country       <chr> "United States", "United States", "United States", "Unit~
## $ City          <chr> "Henderson", "Henderson", "Los Angeles", "Fort Lauderdal~
## $ State         <chr> "Kentucky", "Kentucky", "California", "Florida", "Florid~
## $ `Postal Code` <dbl> 42420, 42420, 90036, 33311, 33311, 90032, 90032, 90032, ~
## $ Region        <chr> "South", "South", "West", "South", "South", "West", "Wes~
## $ Category      <chr> "Furniture", "Furniture", "Office Supplies", "Furniture"~
## $ Subcategory   <chr> "Bookcases", "Chairs", "Labels", "Tables", "Storage", "F~
## $ Sales         <dbl> 261.9600, 731.9400, 14.6200, 957.5775, 22.3680, 48.8600,~
## $ Quantity      <dbl> 2, 3, 2, 5, 2, 7, 4, 6, 3, 5, 9, 4, 3, 3, 5, 3, 6, 2, 2,~
## $ Discount      <dbl> 0.00, 0.00, 0.00, 0.45, 0.20, 0.00, 0.00, 0.20, 0.20, 0.~
```

```
## $ Profit        <dbl> 41.9136, 219.5820, 6.8714, -383.0310, 2.5164, 14.1694, 1~
```

```
colnames(Superstore) ##Displays the column names
```

```
##  [1] "Shipmode"    "Segment"     "Country"     "City"        "State"
##  [6] "Postal Code" "Region"      "Category"    "Subcategory" "Sales"
## [11] "Quantity"    "Discount"    "Profit"
```

```
sum(is_null(Superstore)) ## Checking for null values
```

```
## [1] 0
```

```
sum(duplicated(Superstore))
```

Checking if there are any duplicate rows

```
## [1] 17
```

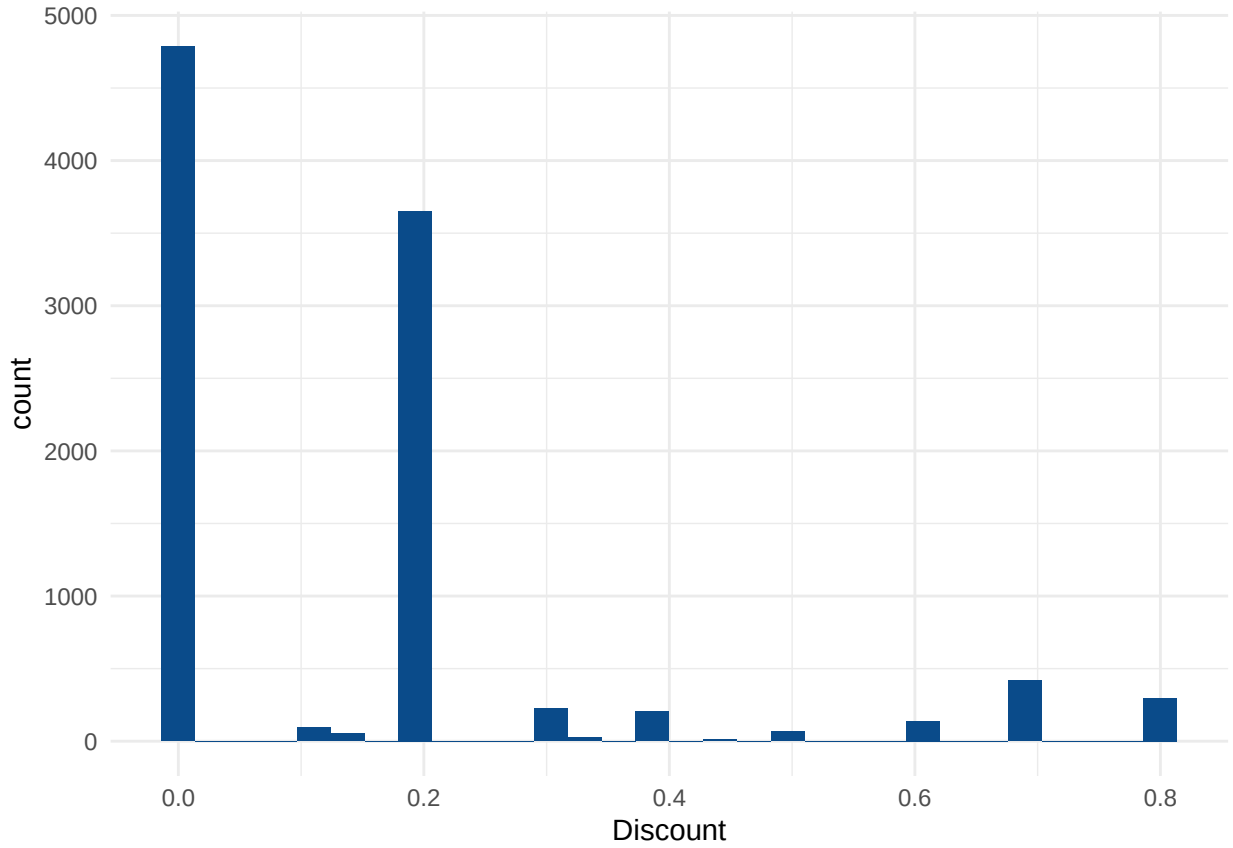
```
Superstore <- unique(Superstore) ##Removing Duplicate
```

```
str(Superstore) ##Duplicates removed
```

```
## tibble[,13] [9,977 x 13] (S3: tbl_df/tbl/data.frame)
##  $ Shipmode    : chr [1:9977] "Second Class" "Second Class" "Second Class" "Standard Class" ...
##  $ Segment     : chr [1:9977] "Consumer" "Consumer" "Corporate" "Consumer" ...
##  $ Country     : chr [1:9977] "United States" "United States" "United States" "United States" ...
##  $ City        : chr [1:9977] "Henderson" "Henderson" "Los Angeles" "Fort Lauderdale" ...
##  $ State       : chr [1:9977] "Kentucky" "Kentucky" "California" "Florida" ...
##  $ Postal Code: num [1:9977] 42420 42420 90036 33311 33311 ...
##  $ Region      : chr [1:9977] "South" "South" "West" "South" ...
##  $ Category    : chr [1:9977] "Furniture" "Furniture" "Office Supplies" "Furniture" ...
##  $ Subcategory: chr [1:9977] "Bookcases" "Chairs" "Labels" "Tables" ...
##  $ Sales       : num [1:9977] 262 731.9 14.6 957.6 22.4 ...
##  $ Quantity    : num [1:9977] 2 3 2 5 2 7 4 6 3 5 ...
##  $ Discount    : num [1:9977] 0 0 0 0.45 0.2 0 0 0.2 0.2 0 ...
##  $ Profit      : num [1:9977] 41.91 219.58 6.87 -383.03 2.52 ...
```

Checking for Outliers

```
ggplot(Superstore) +
  aes(x = Discount) +
  geom_histogram(bins = 30L, fill = "#0a4b8a") +
  theme_minimal()
```




```
Superstore <- filter(Superstore, Discount > 0.45) ##filtering outliers
```

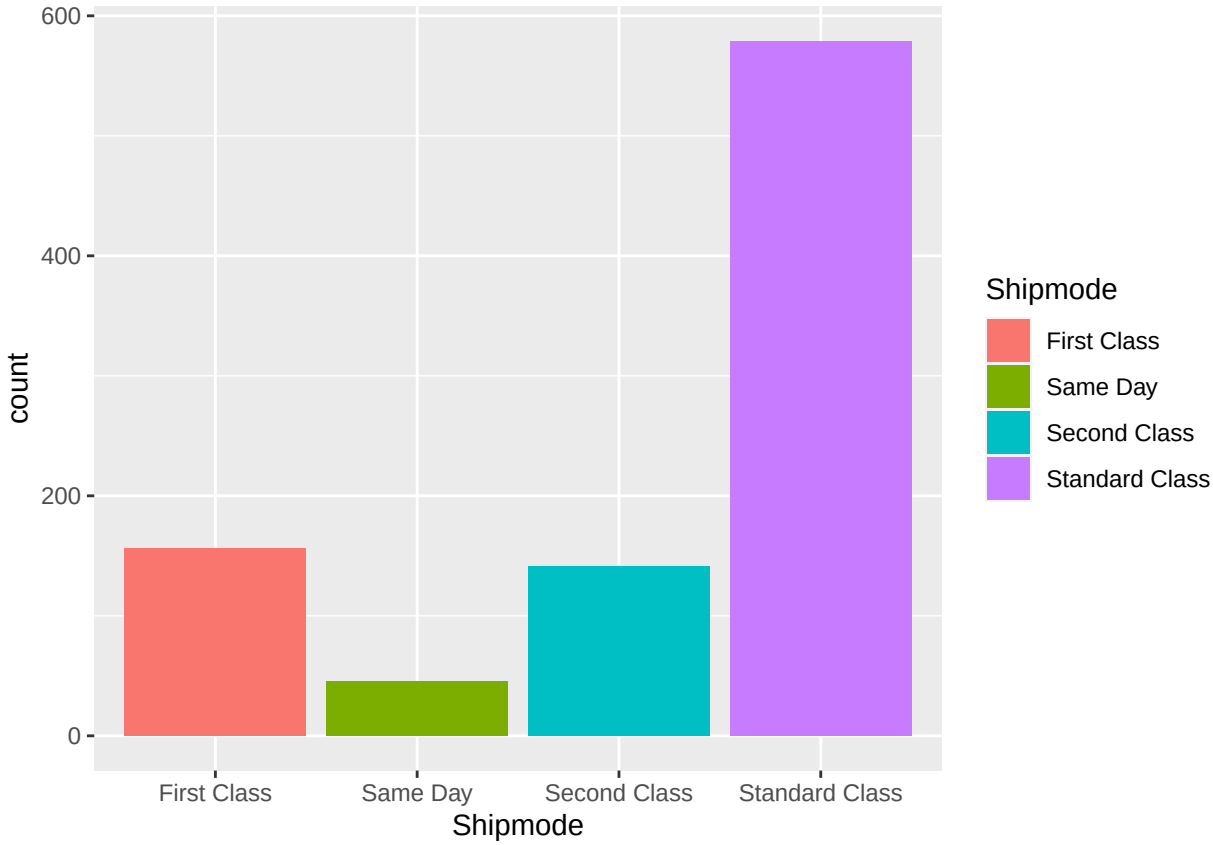
```
head(Superstore$Discount) #filtered outliers
```

```
## [1] 0.8 0.8 0.5 0.7 0.7 0.6
```

Data visualization

* Analysing Shipmode

```
ggplot(Superstore, aes(Shipmode)) + geom_bar(aes(fill = Shipmode))
```

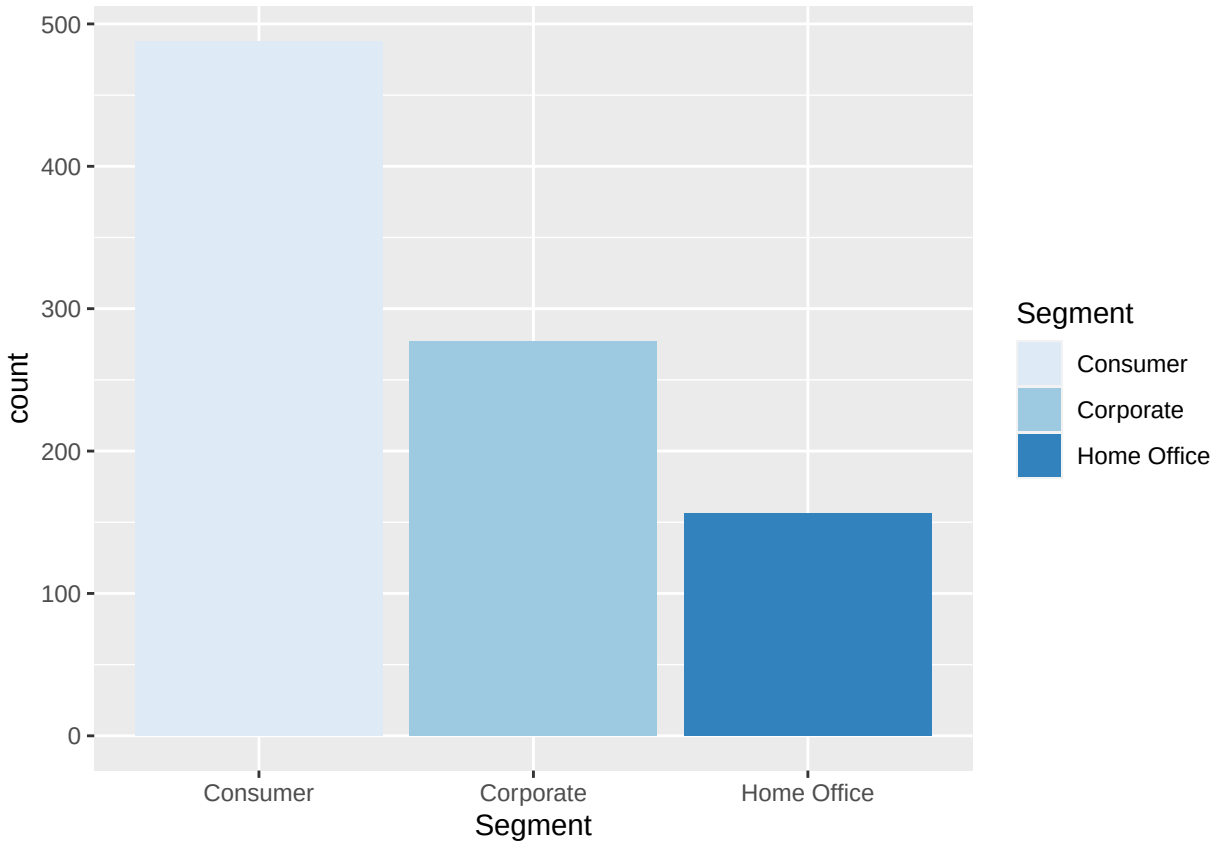


Analysis 1

we can find out that standard class ship mode is preferred more than the other ship modes are available.

* Analysing Segment

```
ggplot(Superstore, aes(Segment )) + geom_bar(aes(fill = Segment )) +  scale_fill_brewer(palette = "Blues
```

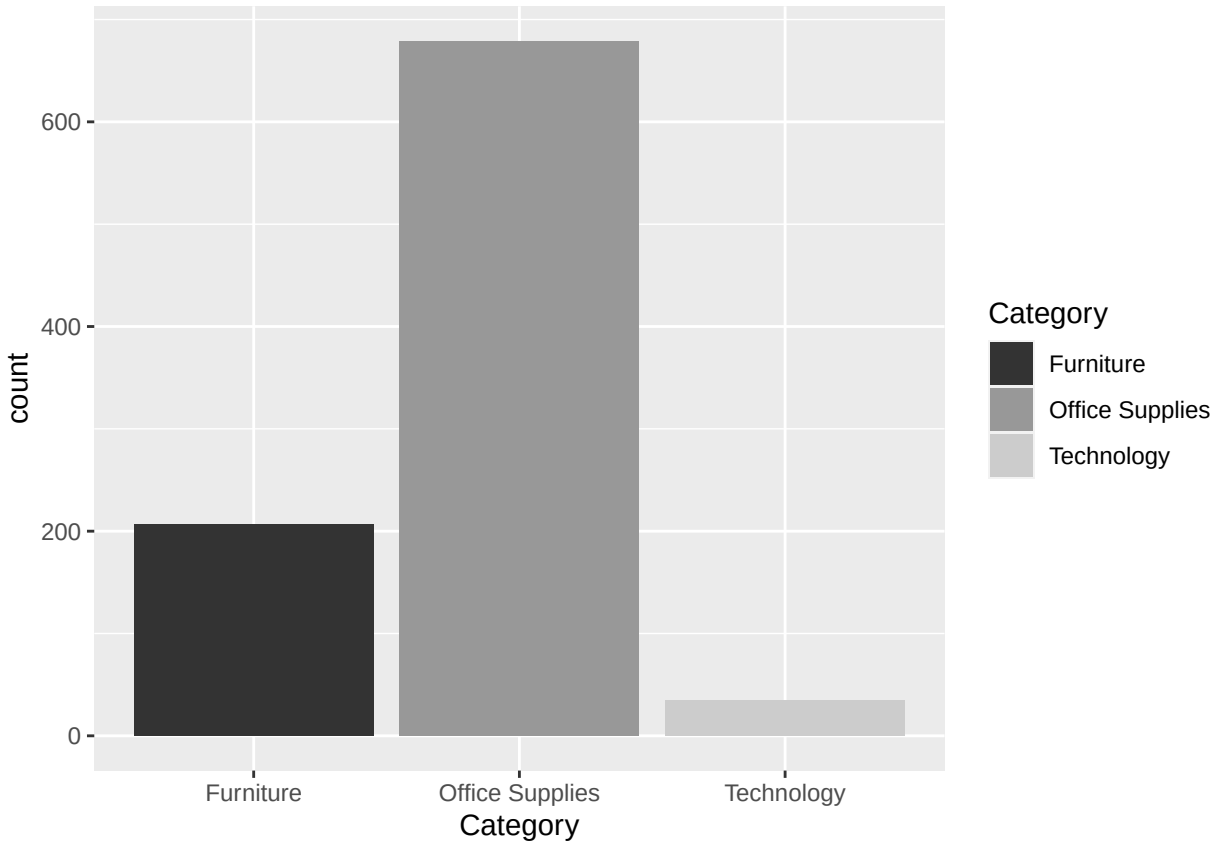


Analysis 2

Consumer has the highest count than corporate and home office.

* Analysing Category

```
ggplot(Superstore, aes(Category)) + geom_bar(aes(fill = Category)) + scale_fill_grey(
start = 0.2, end = 0.8,
na.value = "red")
```

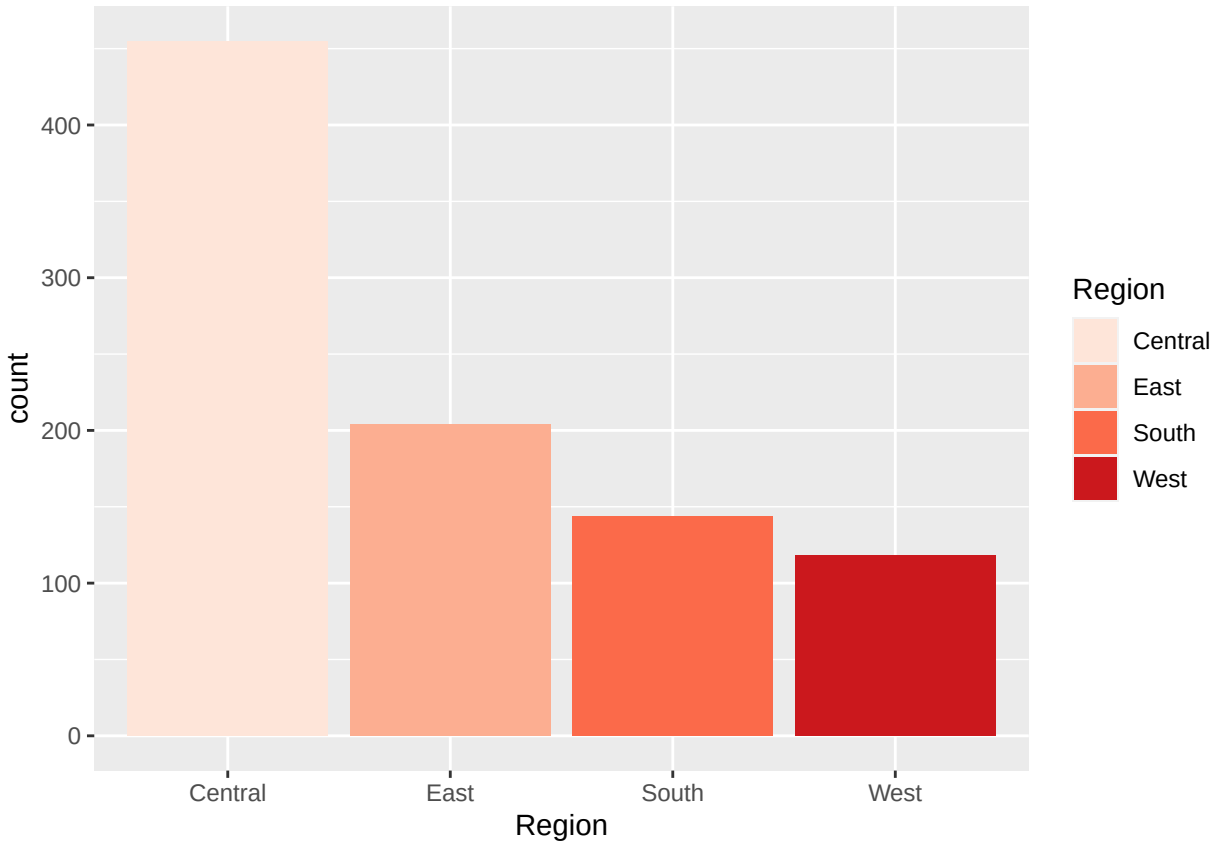


Analysis 3

Office Supplies Category has the highest count.

* Analysing Region

```
ggplot(Superstore, aes(Region)) + geom_bar(aes(fill = Region)) + scale_fill_brewer(palette = "Reds")
```



Analysis 4

Here West region has the highest customer's count

* Analysing Category and Sub-Category

```
ggplot(data=Superstore) + geom_bar(mapping = aes(x=Category , fill=Subcategory)) + facet_wrap(~ Subcateg
```

Analysis 5

*** Furniture which consists of 4 subcategories items Bookcases, Chairs, Tables, Furnishings**

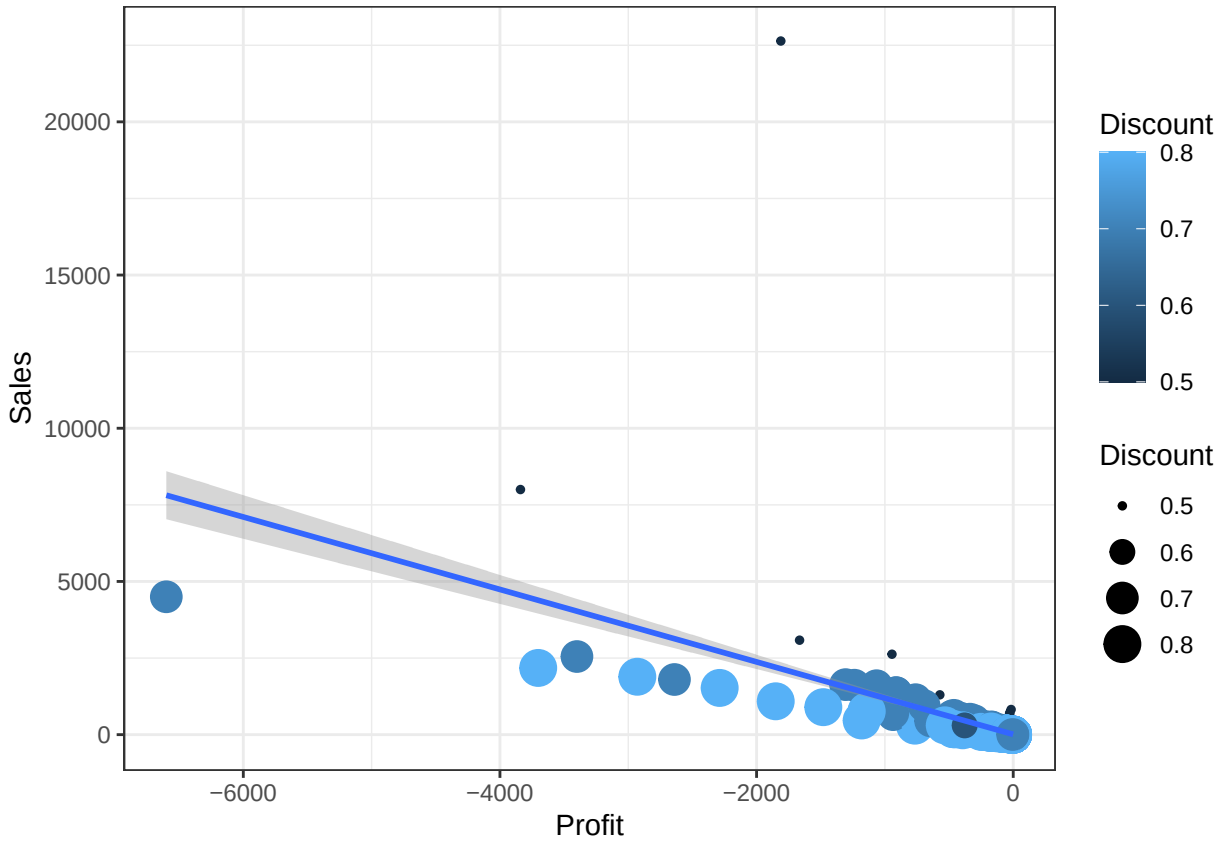
*** Office Supplies category consist 9 sub-categories Labels, Storage, Art, Binders, Appliances, Paper, Envelopes, Fasteners, Supplies**

*** Technology involves 4 subcategories, those are Phones, Accessories, Machines, Copiers.**

* Analysing Profit, Discount and Sales

```
ggplot(Superstore, aes(Profit , Sales)) +
 geom_point(aes(color = Discount , size=Discount)) +
 geom_smooth(method ="lm") +
 coord_cartesian() +
 scale_color_gradient() +
 theme_bw()
```

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



* sales vs profit

```
ggplot(data=Superstore, aes(x=Sales,y=Profit,col= Shipmode, shape=Shipmode))+ geom_point()+geom_smooth(
```

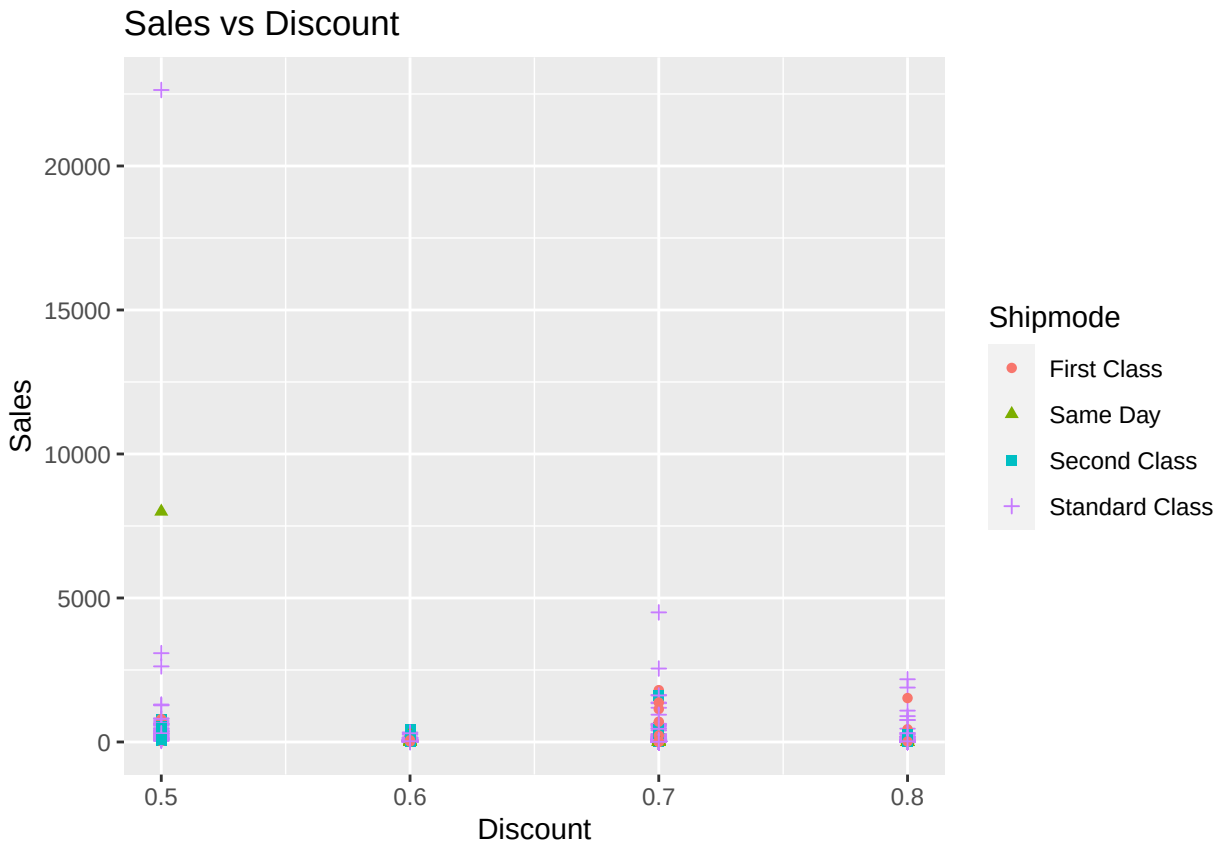
```
## `geom_smooth()` using formula 'y ~ x'
```

Analysis 6

More profit/Loss from Standard class.

* Sales vs Discount

```
ggplot() + geom_point(data=Superstore, aes(x=Discount,y=Sales,col= Shipmode, shape=Shipmode))+labs(titl
```

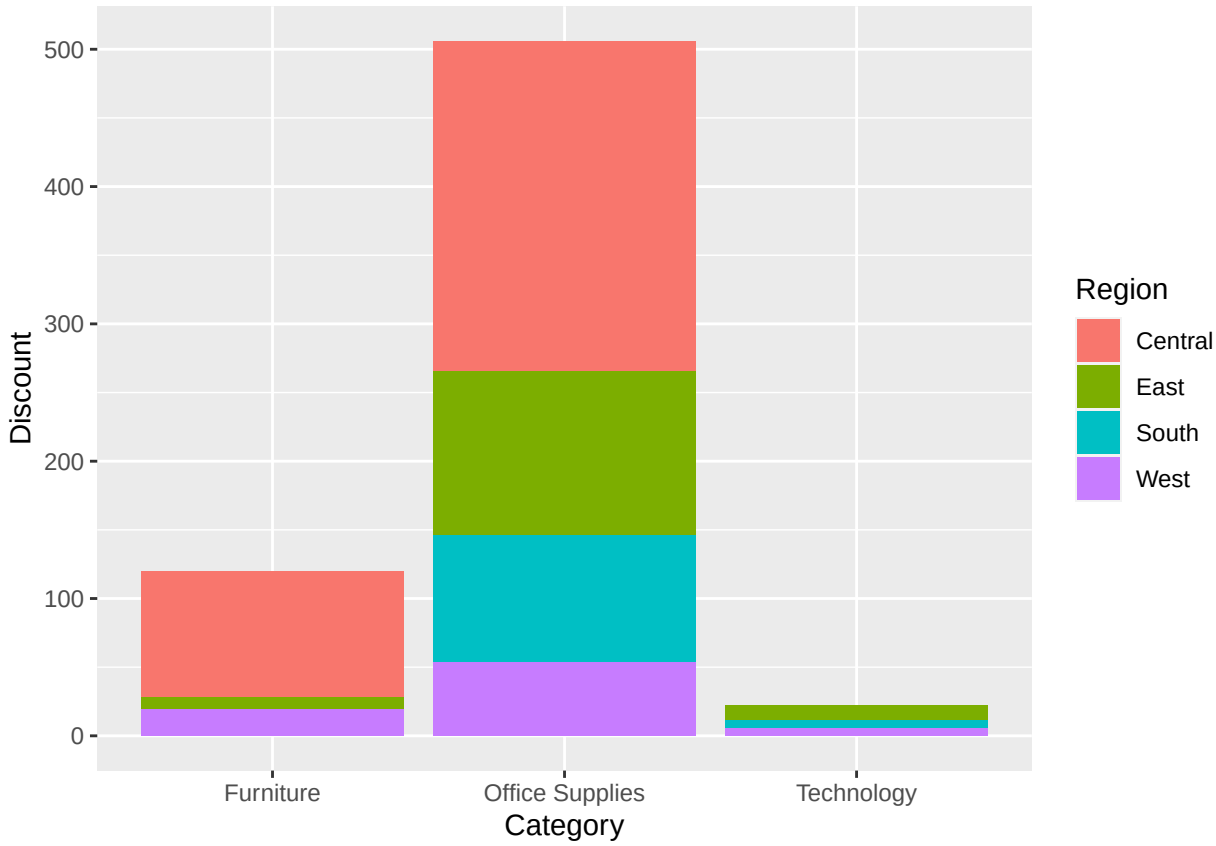


Analysis 7

Discount attracts mostly the Standard class ,Same day receive least Discount

* Category vs Discount

```
ggplot() + geom_bar(data=Superstore, aes(x=Category,y=Discount,fill= Region),stat = "identity")
```

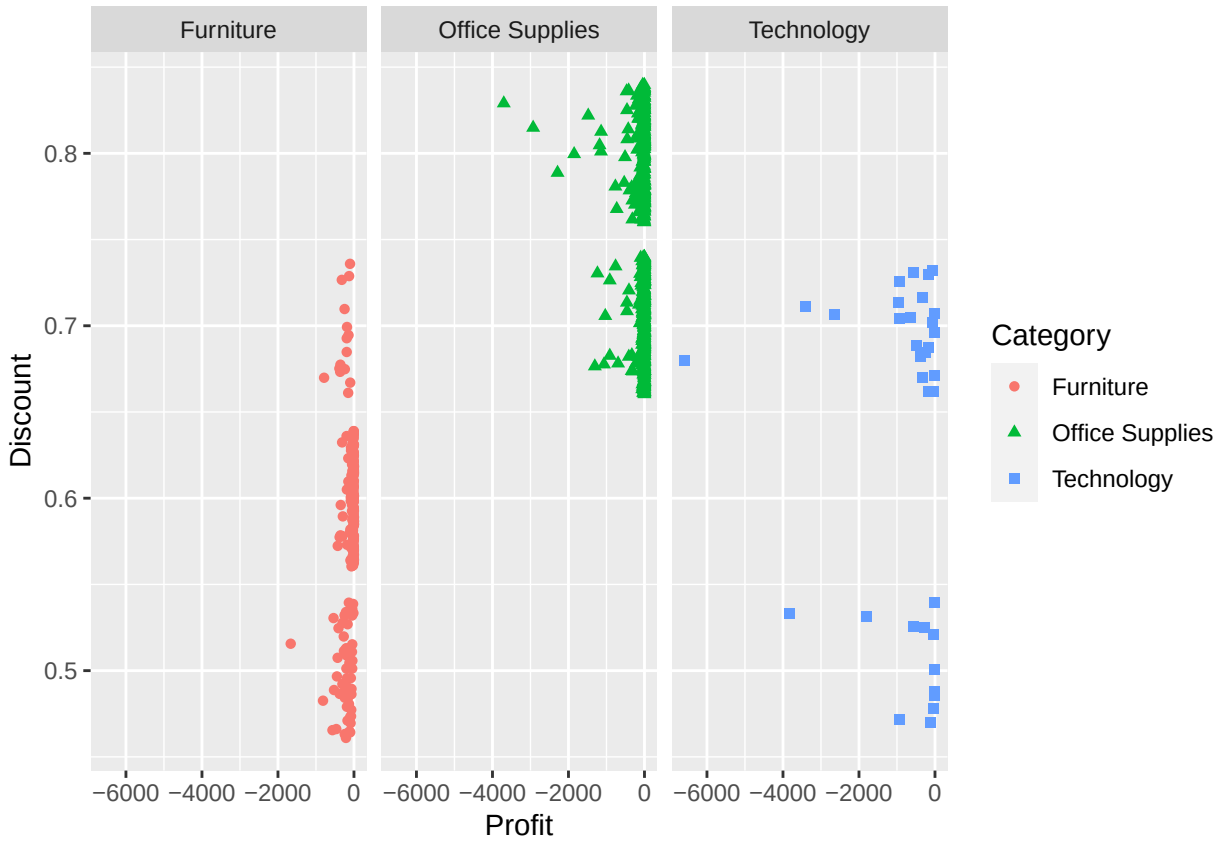


Analysis 8

office suppliers gets more Discount

* Profit and Discount trade off in the different category

```
ggplot(Superstore, aes(Profit ,Discount )) +   geom_jitter(aes(shape = Category, color = Category)) + fa
```

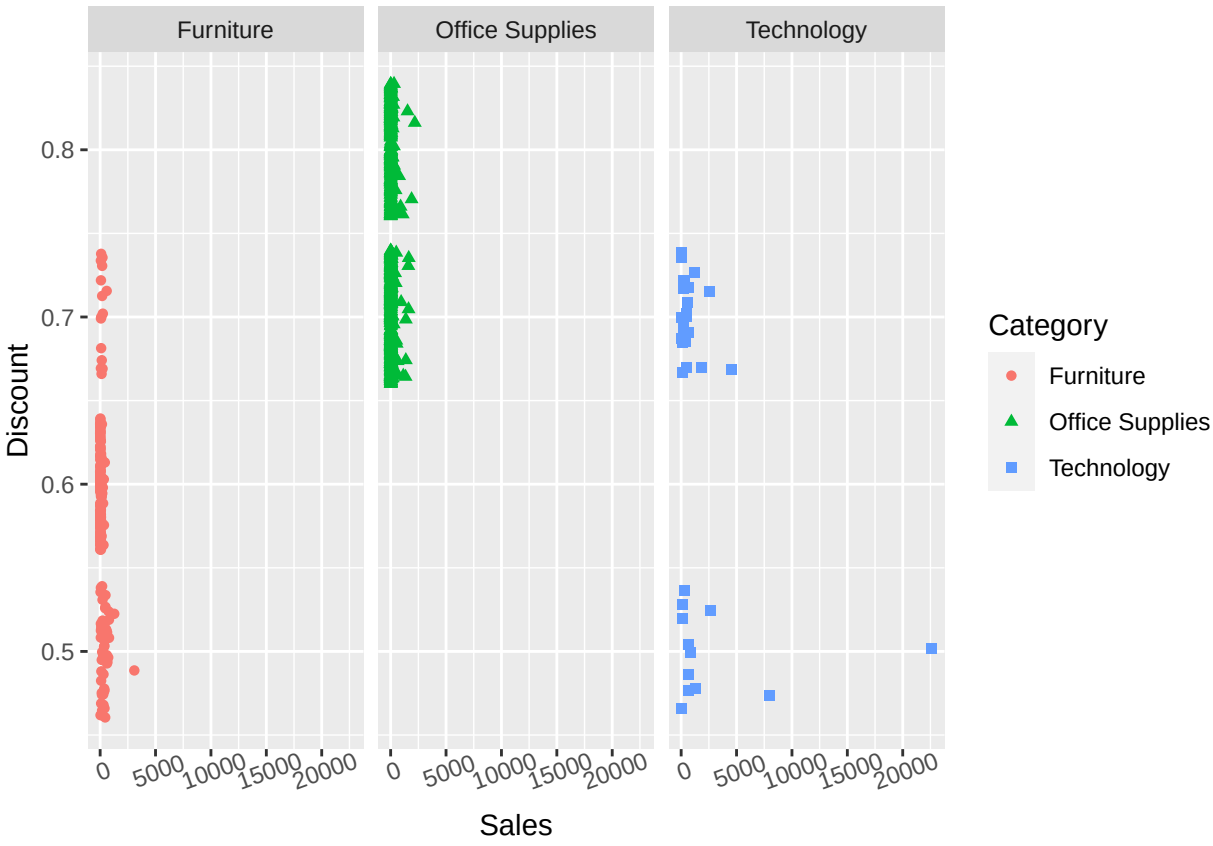


Analysis 9

As discount is increased 0.5 or more losses are faced in all categories product.

* Sales and discount of Product effect in the different category product.

```
ggplot(Superstore, aes(Sales ,Discount )) +   geom_jitter(aes(shape = Category, color = Category))  +fa
```



Analysis 10

*** maximum discount on 0.7 and 0.8, Sales of Office supplies was increased.**

*** Sales of Furniture category was increased due to discount.**

*** Sales of Technology category remains same.**

Conclusion

*** Discounts should be based on sales and should not increase after a particular range otherwise unnecessary discounts with low sales can witness huge losses.**

*** Binders and Machine industry should be focused upon more, Binders gets more discount but gives more losses Office Suppliers sales less but gets more discount and gives losses**

*** Central Region should not offer any kind of discounts for better performance.**

*** Furniture Category performance is less comparing to others so discount should not be give to them.**

*** If Same day shipment receives more discounts then can trigger more sales/Profit.**

THANK YOU