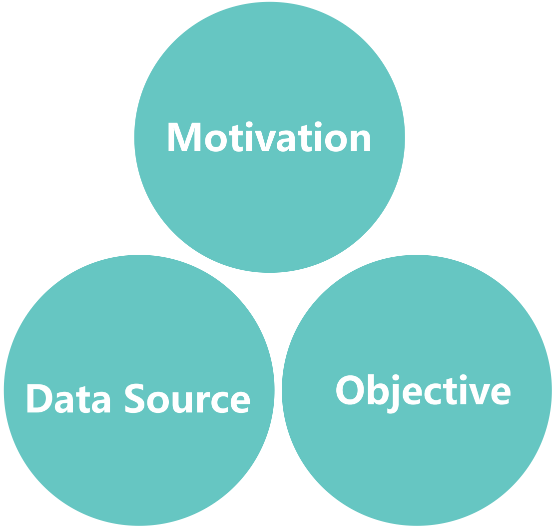
# Analysis on Superstore Sales

Marco Furlan, Dandan Zhao

# Content

# 01 Introduction

# 01 Introduction




## Motivation

Apply statistical methods learned from this course into operational daily business analysis

## Objective

1. Find the relationship between different variables with sales value
2. Predict sales value with linear regression model based on the output above

## Data Source

Dataset is obtained from the public online site of data.world
url: https://data.world/stanke/superstore-20214

# 02 Dataset

# 02 Dataset

## Dataset Summary

- A sales records of US superstore from 2018-2021
- 9994 obersevations
- 21 features for each observation, data type:
- 6 numerical including 2 date information,
- 15 categorical

| Column Names | Example | Description |
|---|---|---|
| Row ID | 1 | Unique ID for each row. |
| Order ID | CA-2020-152156 | Unique Order ID for each Customer. |
| Order Date | 08/11/2020 | Order Date of the product. |
| Ship Date | 11/11/2020 | Shipping Date of the Product. |
| Ship Mode | Second Class | Shipping Mode specified by the Customer. |
| Customer ID | CG-12520 | Unique ID to identify each Customer. |
| Customer Name | Claire Gute | Name of the Customer. |
| Segment | Consumer | The segment where the Customer belongs. |
| Country/Region | United States | Country of residence of the Customer. |
| City | Henderson | City of residence of of the Customer. |
| State | Kentucky | State of residence of the Customer. |
| Postal Code | 42420 | Postal Code of every Customer. |
| Region | South | Region where the Customer belong. |
| Product ID | FUR-BO-10001798 | Unique ID of the Product. |
| Category | Furniture | Category of the product ordered. |
| Sub-Category | Bookcases | Sub-Category of the product ordered. |
| Product Name | Bush Somerset Collection Bookcase | Name of the Product |
| Sales | 261.96 | Sales of the Product. |
| Quantity | 2 | Quantity of the Product. |
| Discount | 0 | Discount provided. |
| Profit | 41.9136 | Profit/Loss incurred. |

# 03 Method

# 03 Method

### Step 01
### Data Preprocessing

Check on missing values
Identify Outliers
Remove/add columns
Factorize categorical variable

### Step 02
### Data Exploration

Compute descriptive statistics
(max,min,quantile,median,
correlation coefficient etc.)
Visualize pattern/trend/relation
(density,boxplot,line,bar,scatter,
treemap etc.)

### Step 03
### Data Modeling

Linear regression
Model diagnose
F statistic test
Variable selection method
(BIC score-based, constrained
backword, lasso)
Cross validation
Model evaluation

### Step 04
### Data Interpretation

Statistical meaning
Business meaning

# 04 Experiment

# 04 Experiment - Data Preprocessing

## Check missing Value

Total 12 missing value are only contained in postcode column. Postcode is not used in the analysis, so we do not do anything.

## Identify Outliers

Outliers are the obervations with sales value greater than 5000 and profit less than -2000 or greater than 2000 which is less than 0.3% proportion of the dataset.

## Add/Remove columns

Removed columns:
Row.ID,Country.Region,Customer.Name,Postal.Code,Product.Name

Added columns:
Order Date Year, Order Date Month
Discount Level - An ordered factor with 4 levels of the degree

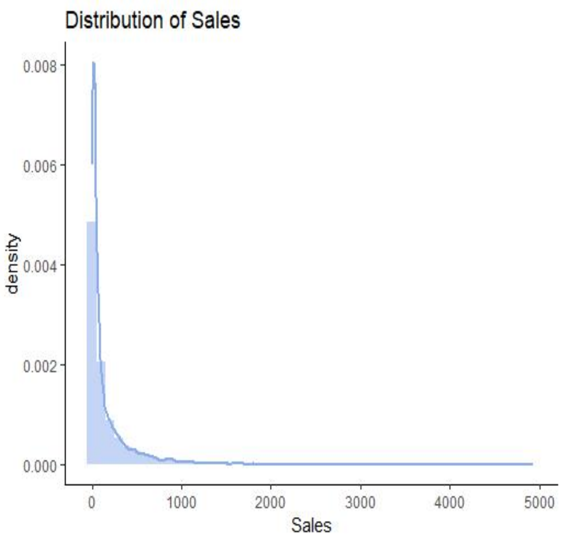
- No Discount <- Discount == 0
- Low Discount <- 0 < Discount <= 0.3
- Median Discount <- 0.3 < Discount <= 0.6
- High Discount <- 0.6 < Discount <= 1

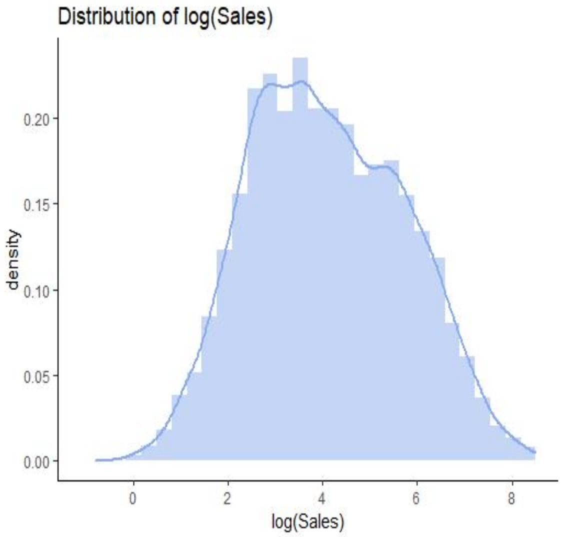
## Factorize the categorical variables

Ship Mode, Segment, Region, State, City, Category, Subcategory

## 04 Experiment - Data Exploration

### Overview on the response variable - Sales




Density of Sales is a tipically log normal distribution situation.
We used log transformation on Sales as the response varible in order to satisfy the assumption of linear regression model

# 04 Experiment - Data Exploration

## Identify the variables relative to Sales potentially by answering the following questions:

*How do different types of product contribute to sales?*

*What is the effect of Discount to Sales?*

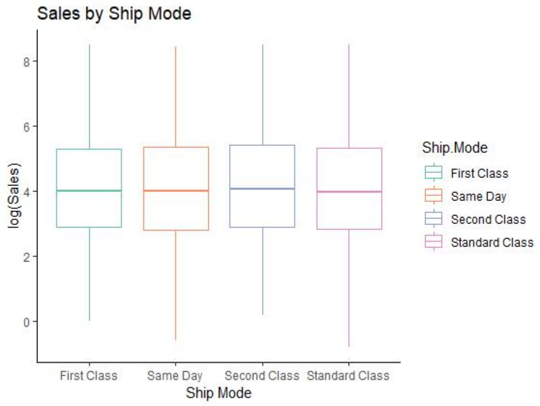
*Is the relationship linear?*

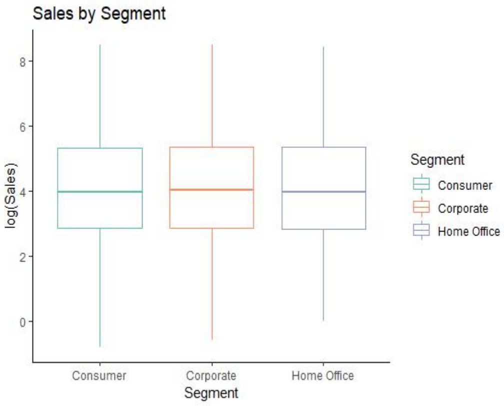

*Is there a relationship between each variable and sales? (Ship Mode, Segment, Region, State, Category, Sub Category)*

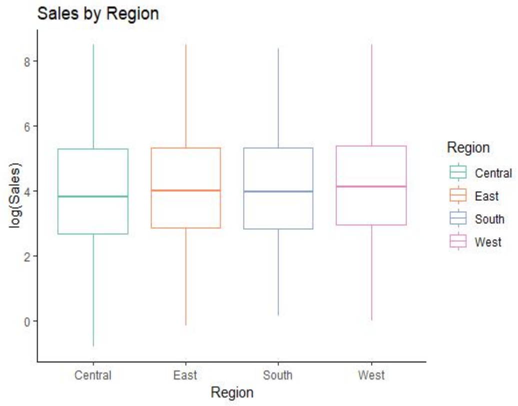
*Is there any Region/State contributing to Sales outstandingly?*

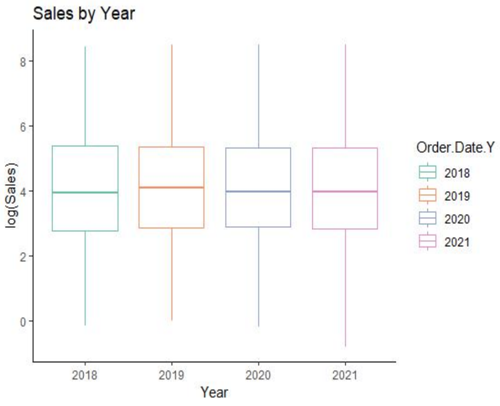
*Is there any interaction among Segment and Discount to Sales?*

# 04 Experiment - Data Exploration



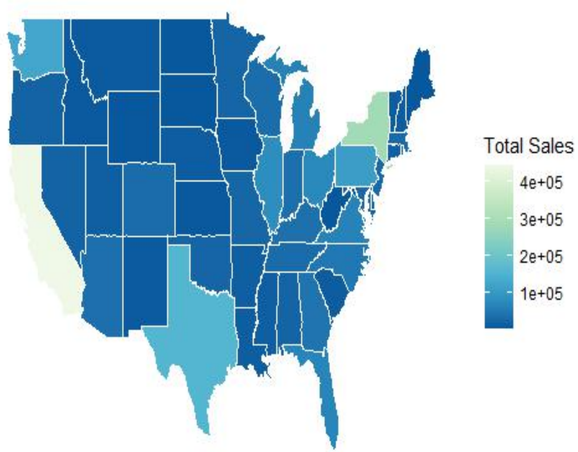

- ❑ Ship mode,
- ❑ Segment,
- ❑ Region,
- ❑ Year

**are not**

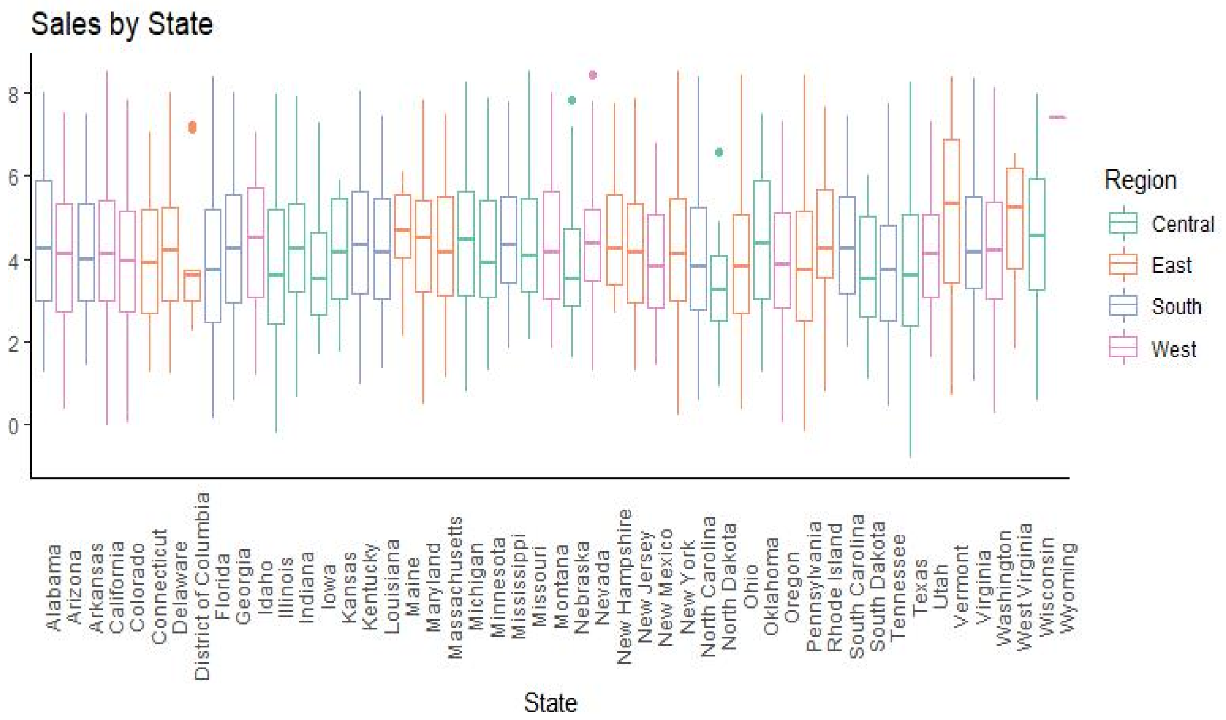
creating variation on Sales.

# 04 Experiment - Data Exploration

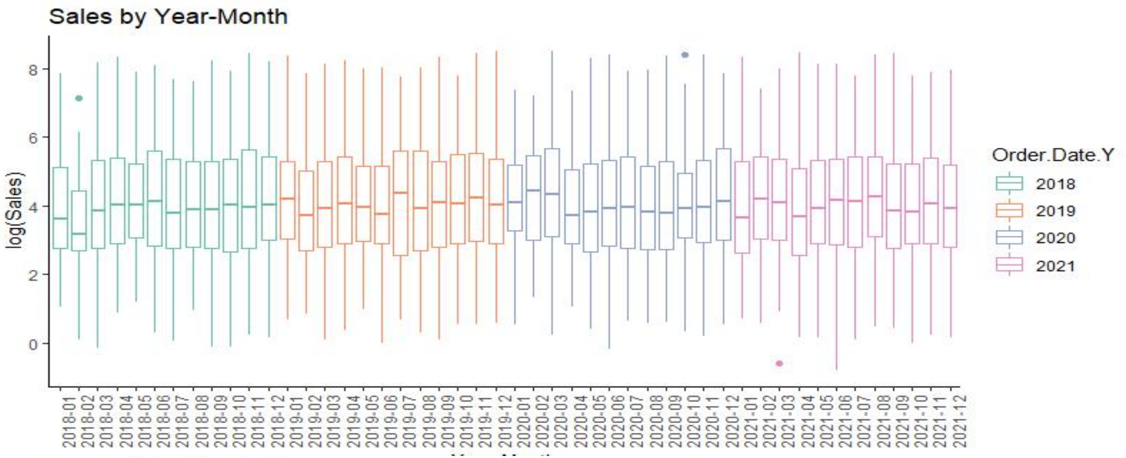
## Total Sales by State

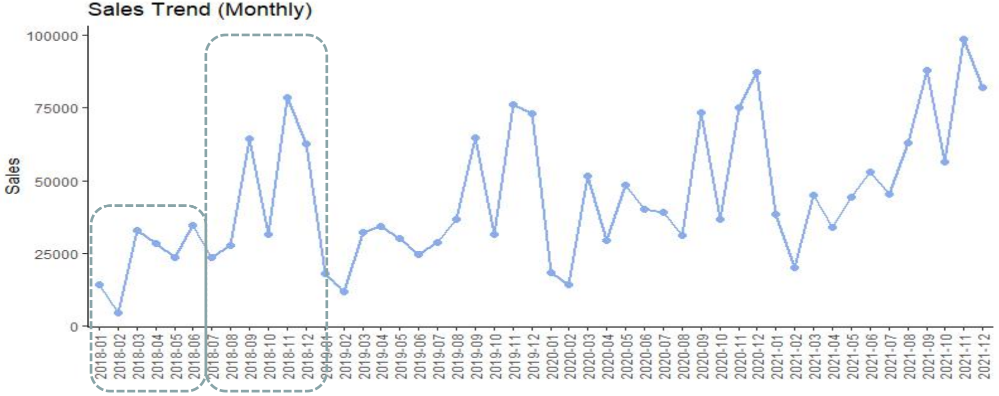



| Top5 States | |
|---|---|
| California | 444416 |
| New York | 287476 |
| Texas | 158325 |
| Washington | 124641 |
| Pennsylvania | 108112 |



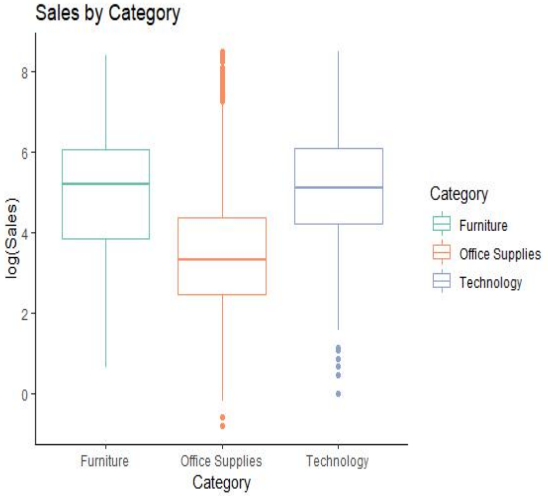
# 04 Experiment - Data Exploration

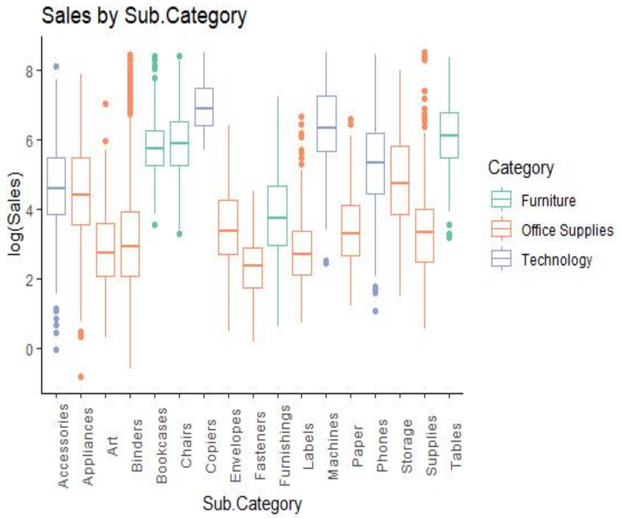


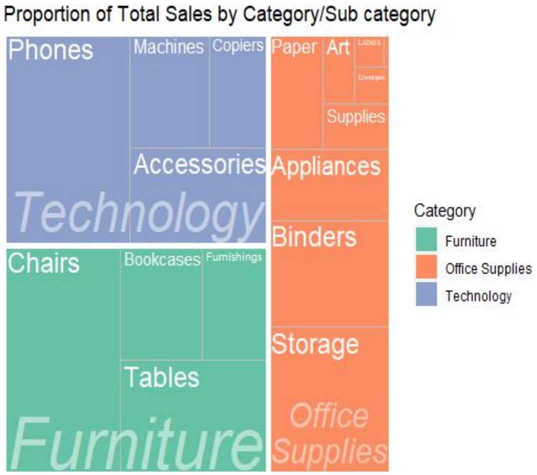




1. Sales shows a similar patten of seasonality from year to year.
2. In the same year, sales shows an evident stronger trend in Sep – Dec, with exception in October.

# 04 Experiment - Data Exploration




Sales value are very different in each sub category which means they are **strongly related**.

# 04 Experiment - Data Explanation

## Correlation Matrix

| | Sales | Quantity | Discount | Profit |
|---|---|---|---|---|
| Sales | 1,0000 | 1 0,2577 | 2 -0,0464 | 3 0,4781 |
| Quantity | 0,2577 | 1,0000 | 0,0079 | 0,0969 |
| Discount | -0,0464 | 0,0079 | 1,0000 | -0,2976 |
| Profit | 0,4781 | 0,0969 | -0,2976 | 1,0000 |


1) There is a positive **linear relationship between Quantity and Sales**, but the linearity is not so strong.
2) The correlation coefficient of Discount vs Sales is negative value close to zero, which indicates there is no linear relationship. Based on the boxplot of discount level and sales, sales distribution are different, there is non-linear relationship possibly.
3) **Profit is positively related to Sales**. We will not consider profit as a predictor because materially speaking profit should depend on Sales

## 04 Experiment - Data Exploration

Next we analyzed the interaction between the related variables

1.1 State vs Sub Category
1.2 State vs Order month
1.3 State vs Quantity
1.4 State vs Discount

2.1 Order Month vs Category/Sub Category
2.2 Order Month vs Quantity
2.2 Order Month vs Discount

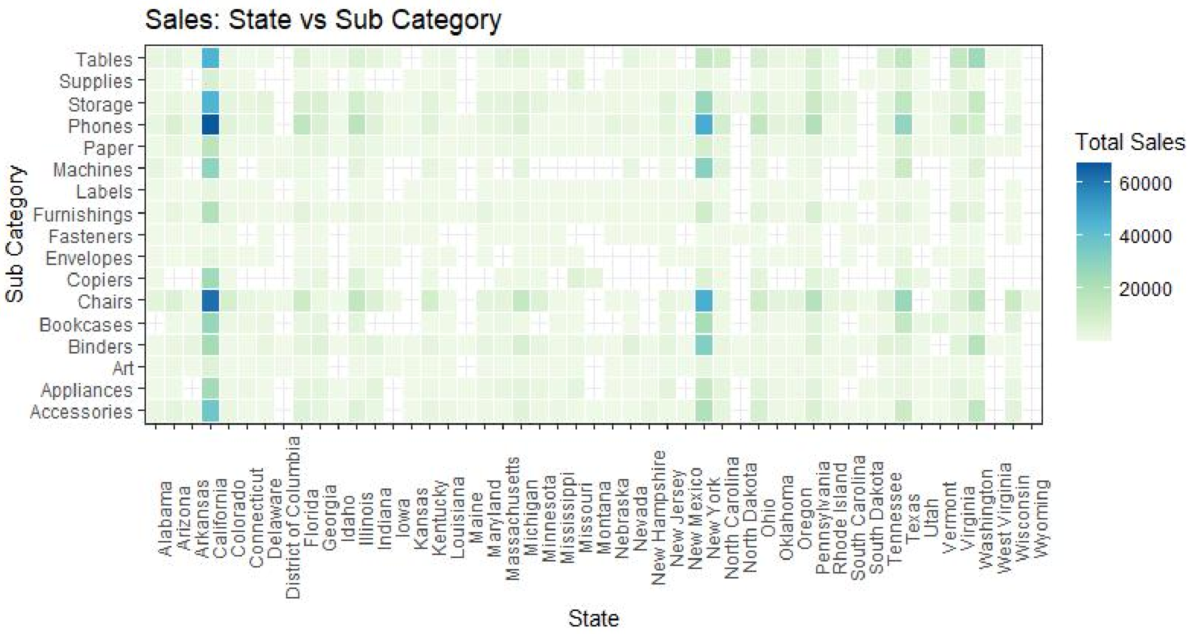
3.1 Sub Category vs Quantity
3.2 Sub Category vs Discount

4. Discount vs Quantity

# 04 Experiment - Data Exploration

## 1.1 State vs Sub Category

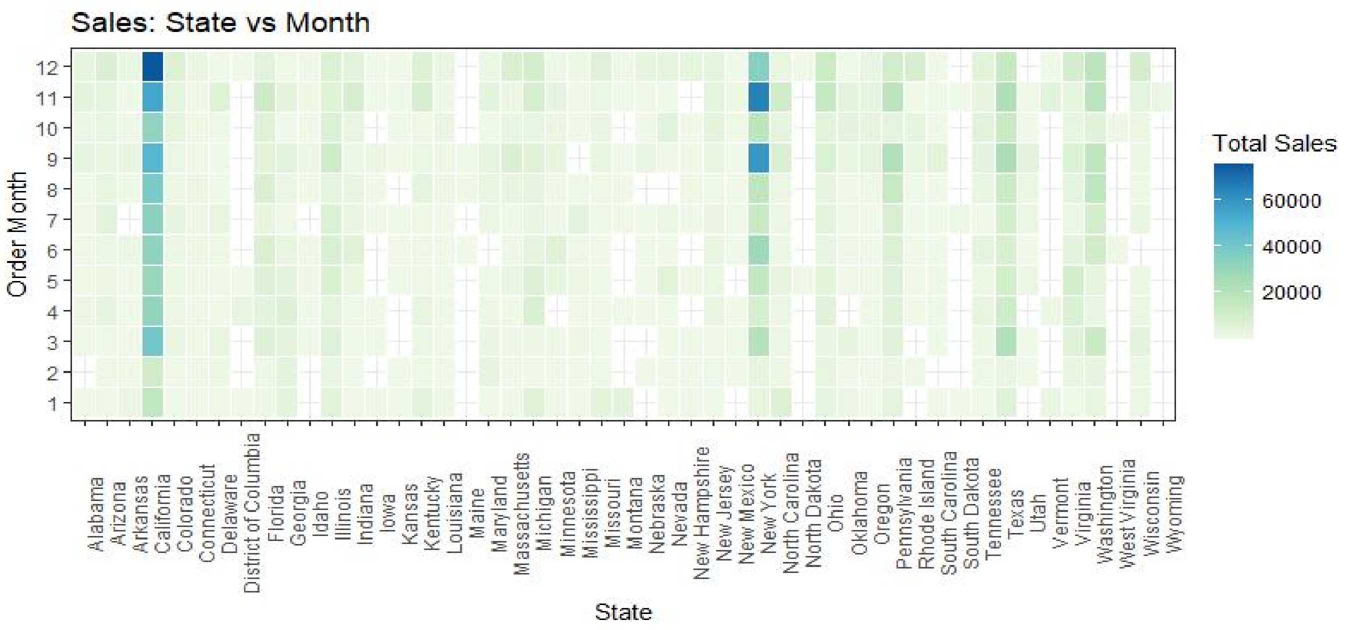
All the States have the similar popular subcategories:
e.g.: phones, chairs



# 04 Experiment - Data Exploration

## 1.2 State vs Order month

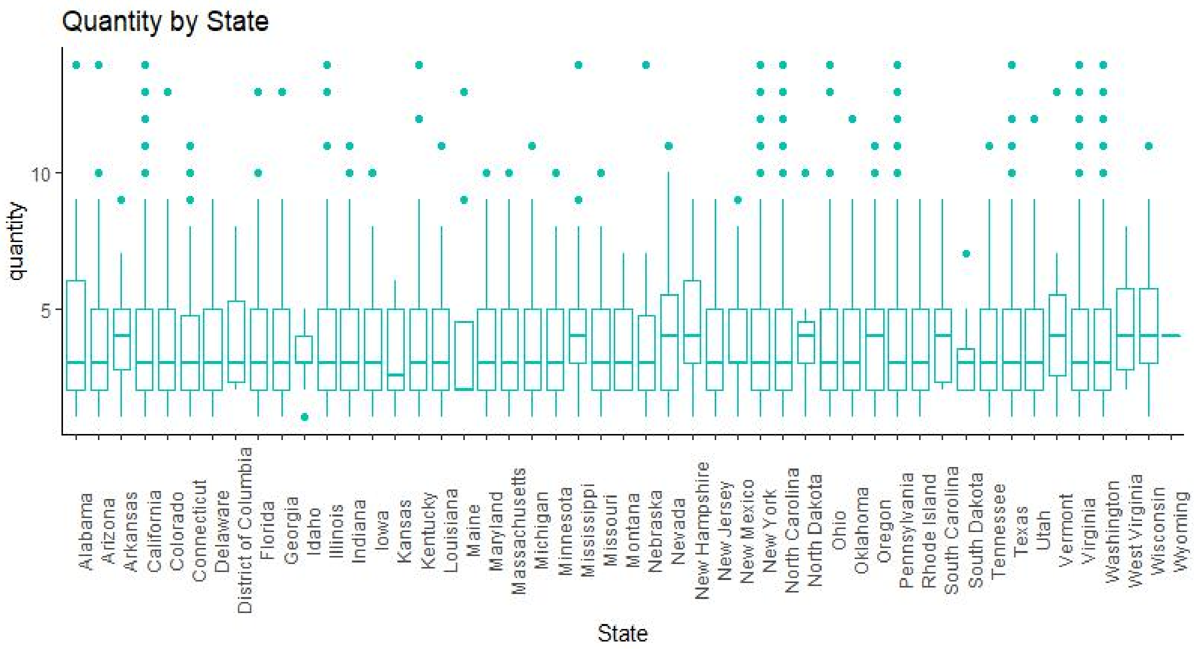
No evident patten in sales with respect to different order month



# 04 Experiment - Data Exploration

## 1.3 State vs Quantity

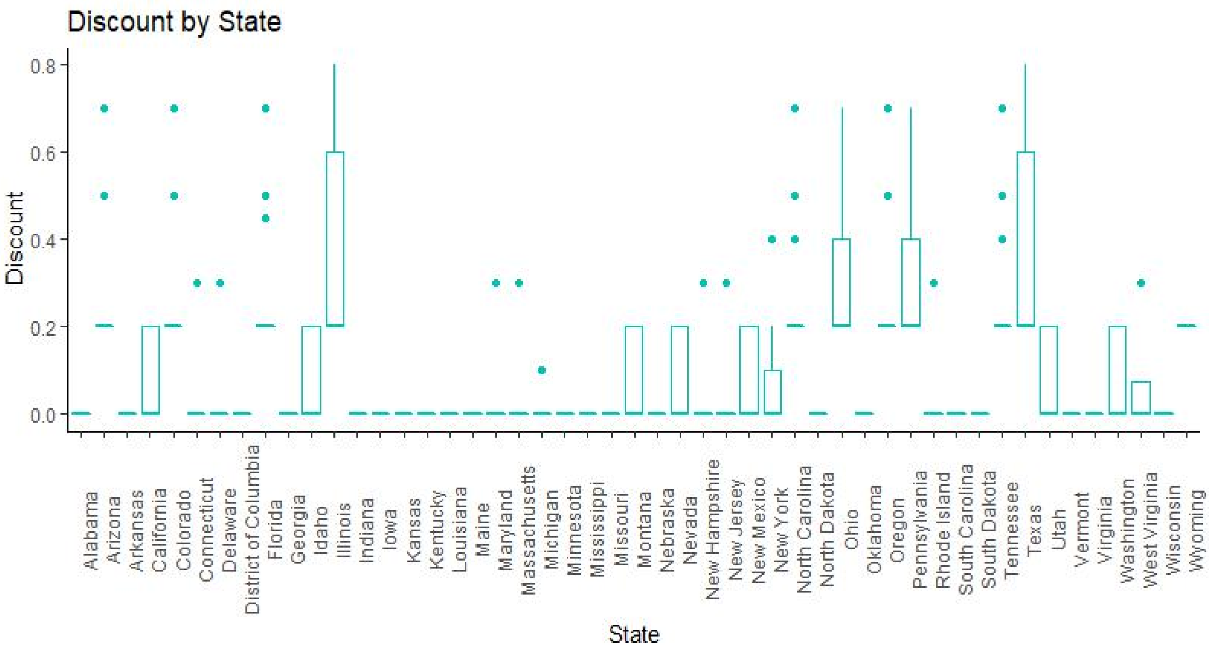
Generally, the quantity does **not** vary much over different States, except few particular cases exist like Idaho, North Dakota, South Dakota



# 04 Experiment - Data Exploration

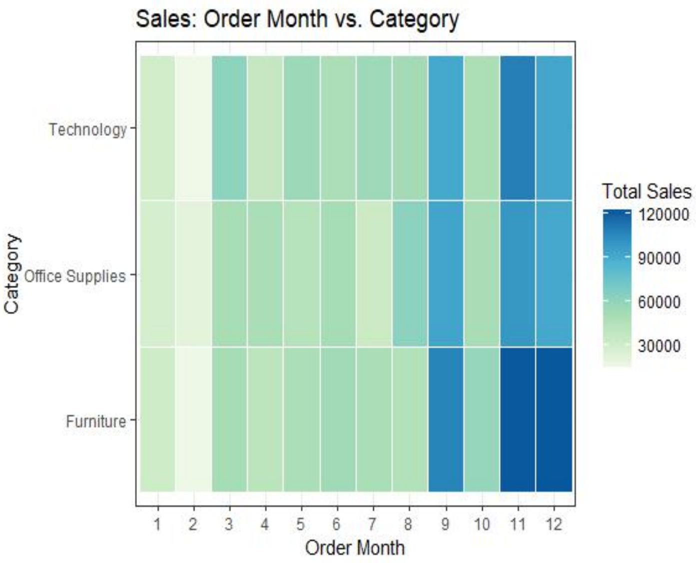
## 1.4 State vs Discount

Discount distribution is **different** in different states.

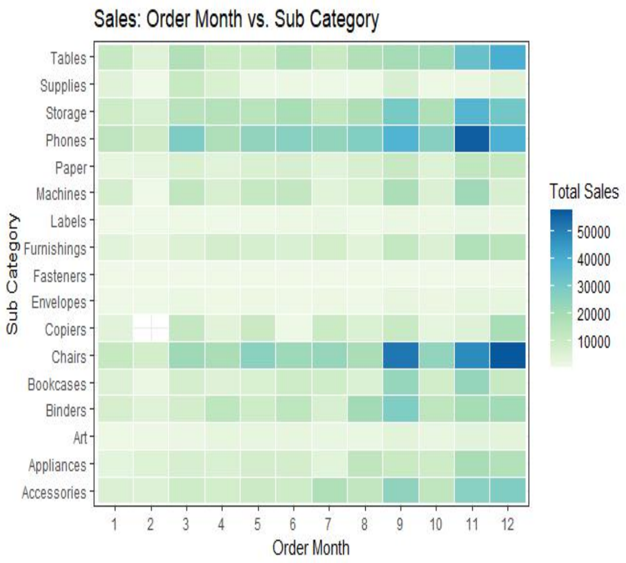


# 04 Experiment - Data Exploration

## 2.1 Order Month vs Category/Sub Category



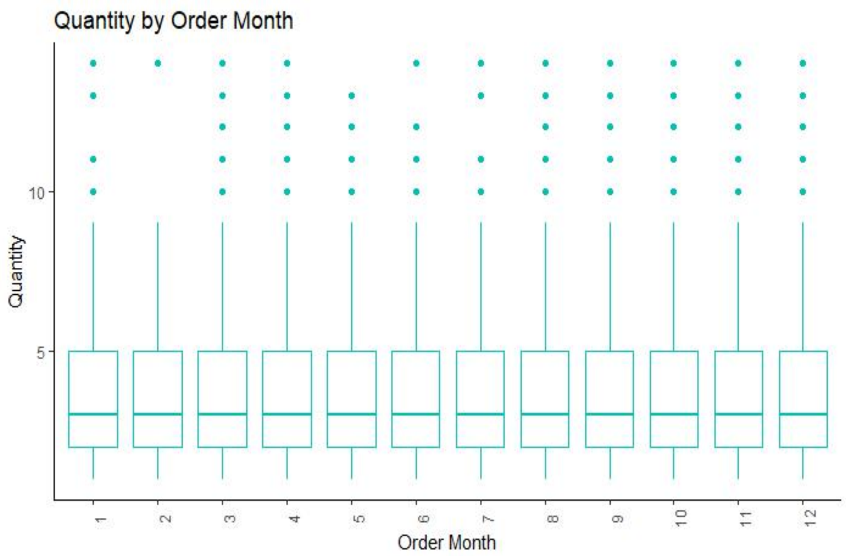

Order month increase in all categories in Sep, Nov, Dec, which is in line with previous analysis on sales




Same pattern is found also in the analysis on subcategory, where stronger in few particular subcategories (e.g.: phones, chairs, etc.) than others

# 04 Experiment - Data Exploration

## 2.2 Order Month vs Quantity

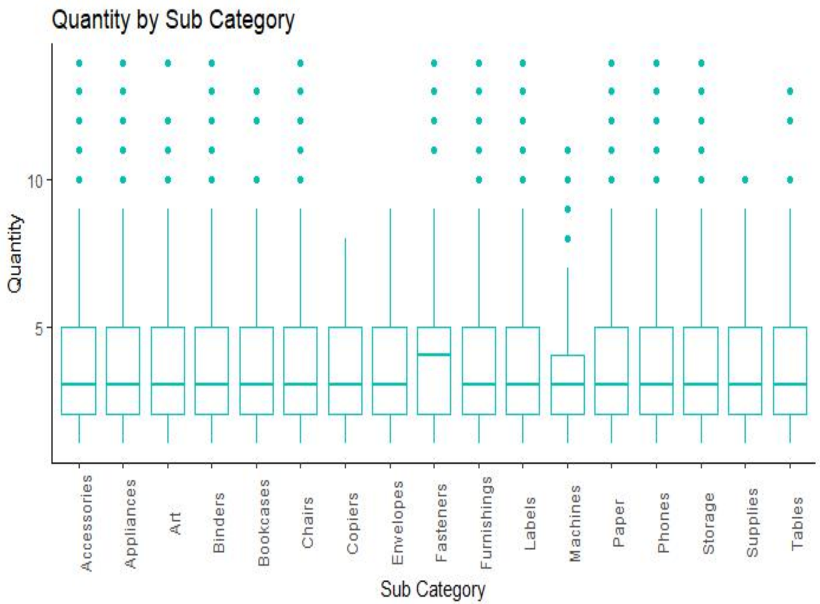



## 2.3 Order Month vs Quantity


Quantity and Discount do **not** depend on Order Month

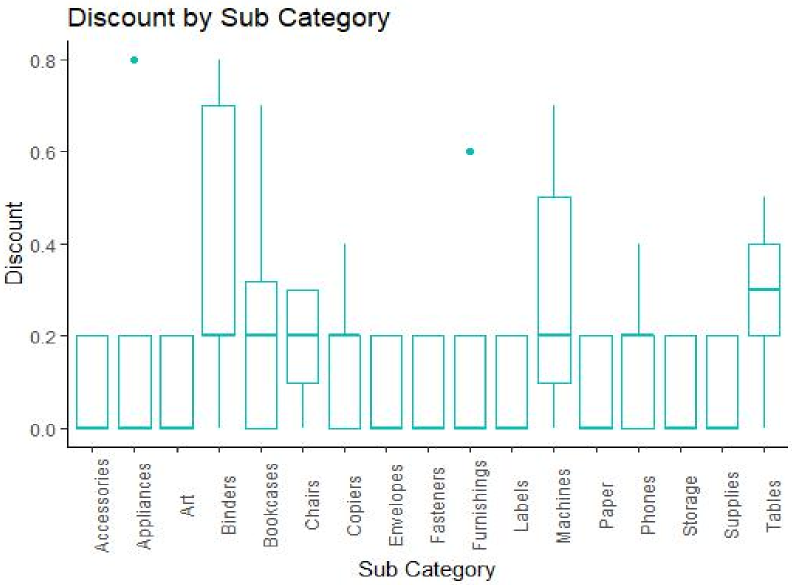
# 04 Experiment - Data Exploration

## 3.1 Sub Category vs Quantity




Discount is almost same distributed.

## 3.2 Sub Category vs Discount




Discount variation for different Sub Category

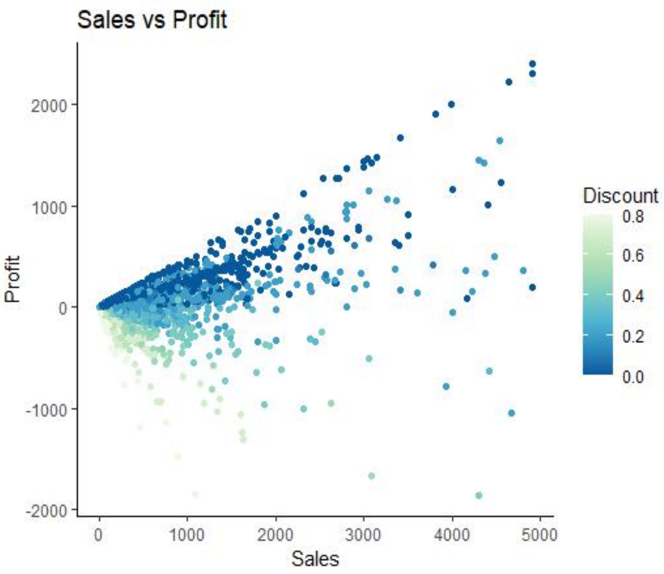
# 04 Experiment - Data Explanation

## 4 Discount vs Quantity


Discount level is **not** creating influence on Quantity.

# 04 Experiment - Data Explanation

Additionally, plotting Sales versus Profit considering also discount level we could see that:




It is clearly showing that:

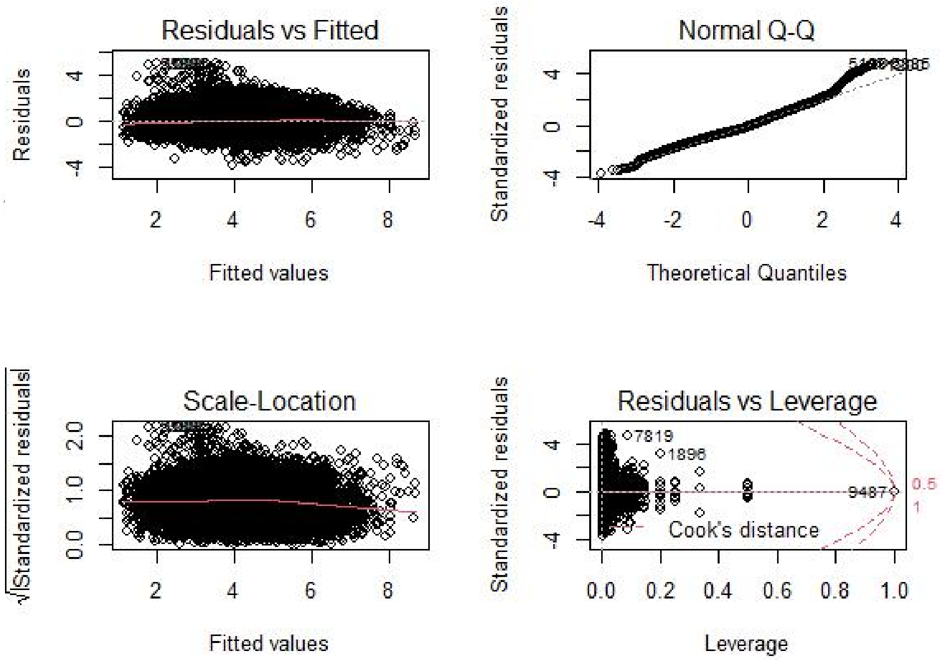
1)no discount: **positive** relationship between sales and profit

2)part of lower discount, median and high discount: **negative** relationship

3)part of lower discount has the **positive** effect on profit is valuable.

## Model 1 - Based on the previous analysis

```
Call:
lm(formula = log(Sales) ~ State + Sub.Category + Order.Date.M
    Discount.Level + Quantity + Discount.Level:Sub.Category +
    Discount.Level:State, data = sub.data)


---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.052 on 9830 degrees of freedom
Multiple R-squared:  0.5887,    Adjusted R-squared:  0.5829
F-statistic:   102 on 138 and 9830 DF,  p-value: < 2.2e-16
```



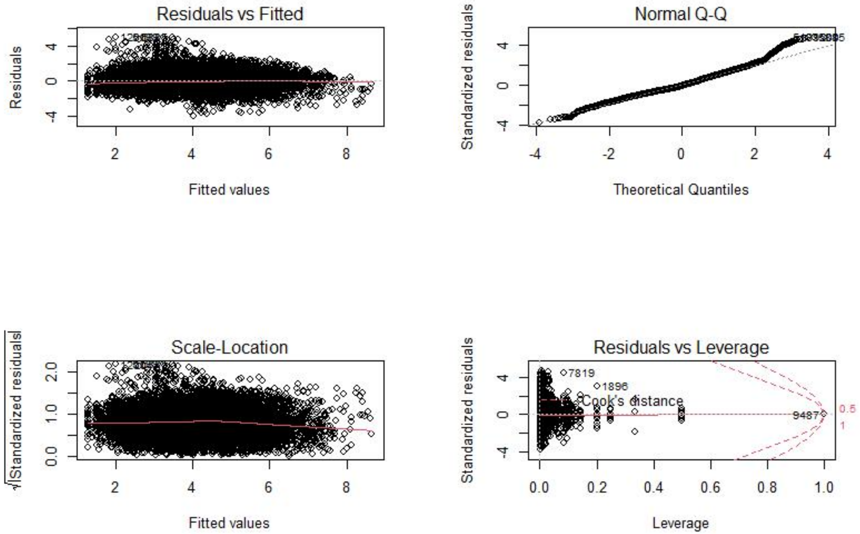
# 04 Experiment - Data Modeling

## Model 2 - Constraint based backward elimination method

```
> mod.R <- update(mod.F, .~.-Order.Date.M)
> anova(mod.R, mod.F)
Analysis of Variance Table

Model 1: log(Sales) ~ State + Sub.Category + Discount.Level + Quantity +
    Sub.Category:Discount.Level + State:Discount.Level
Model 2: log(Sales) ~ State + Sub.Category + Order.Date.M + Discount.Level +
    Quantity + Discount.Level:Sub.Category + Discount.Level:State
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1   9841 10889
2   9830 10876 11    12.429 1.0212  0.424
> mod.R <- update(mod.R, .~.-Discount.Level:Sub.Category)
> anova(mod.R, mod.F)
Analysis of Variance Table

Model 1: log(Sales) ~ State + Sub.Category + Discount.Level + Quantity +
    State:Discount.Level
Model 2: log(Sales) ~ State + Sub.Category + Order.Date.M + Discount.Level +
    Quantity + Discount.Level:Sub.Category + Discount.Level:State
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1   9865 10918
2   9830 10876 35    41.852 1.0807 0.3419
> # step3
> mod.R <- update(mod.R, .~.-Discount.Level:State)
> anova(mod.R, mod.F)
Analysis of Variance Table

Model 1: log(Sales) ~ State + Sub.Category + Discount.Level + Quantity
Model 2: log(Sales) ~ State + Sub.Category + Order.Date.M + Discount.Level +
    Quantity + Discount.Level:Sub.Category + Discount.Level:State
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1   9900 10989
2   9830 10876 70    112.76 1.4559 0.007898 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

lm(log(Sales)~State+Sub.Category+Discount.Level+Quantity+Discount.Level:State,sub.data)

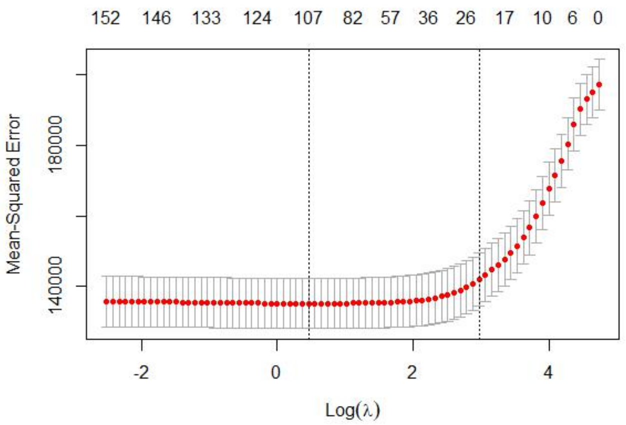
```
Residual standard error: 1.052 on 9865 degrees of freedom
Multiple R-squared:  0.5871,    Adjusted R-squared:  0.5828
F-statistic: 136.2 on 103 and 9865 DF,  p-value: < 2.2e-16
```



# Experiment - Data Modeling

## Model 3 - Applying LASSO on variable selection

```
> library(glmnet)
> # design matrix
> X <- model.matrix(log(Sales)~State+Sub.Category+Order.Date.M+Discount.Level+Quantity
+                   +Discount.Level:Sub.Category+Discount.Level:State, sub.data)
> # remove the first column relative to the intercept
> X <- X[,-1]
> # vector of responses
> y <- sub.data$Sales
> #select 75%*n observation for training set
> set.seed(25)
> train <- sample(1:nrow(X), nrow(X)*0.8)
> test <- (-train)
> y.test <- y[test]
> # apply lasso to the training set without specifying lambda
> lasso.mod <- glmnet(X[train,], y[train], alpha=1)
> plot(lasso.mod, label=TRUE)
> # use 10 folds cross-validation to choose the value of lambda
> cv.out <- cv.glmnet(X[train, ], y[train], alpha = 1, nfold=10)
> plot(cv.out)
> # identify the best lambda value estimated test MSE
> bestlam <- cv.out$lambda.min
> bestlam
[1] 1.58976
> # estimate the test MSE with the best lambda
> lasso.pred <- predict(lasso.mod, s=bestlam, newx=X[test,])
> mean((lasso.pred-y.test)^2)
[1] 104648.3
>
```



## Comparing models

### Model 2 MSE

```
> con.best <- lm(log(Sales)~State+Sub.Category+Discount.Level+Quantity
+                 +Discount.Level:State,sub.data[train,])
> con.pred <- predict(con.best, newdata=sub.data[test,])
> mean((con.pred-y.test)^2)
[1] 182739.5
> summary(con.best)
```

### Model 3 MSE

```
> # estimate the test MSE with the best lambda
> lasso.pred <- predict(lasso.mod, s=bestlam, newx=X[test,])
> mean((lasso.pred-y.test)^2)
[1] 104648.3
```

# 04 Experiment - Data Modeling

## Model 4 - a simpler model

## Model 4 MSE

```
> mod <- lm(log(Sales)~Sub.Category+Discount.Level+Quantity
+               ,sub.data[train,])
> mod.pred <- predict(mod, newdata=sub.data[test,])
> mean((mod.pred-y.test)^2)
[1] 182739.8
> summary(mod)
```

```
Call:
lm(formula = log(Sales) ~ Sub.Category + Discount.Level + Quantity,
    data = sub.data[train, ])

Residuals:
    Min      1Q  Median      3Q     Max
-4.0611 -0.7033 -0.1385  0.6915  5.0953

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.356899   0.072563  32.481  < 2e-16 ***
Sub.CategoryAppliances         0.028710   0.069838   0.411  0.68101
Sub.CategoryArt               -1.782502   0.059708 -29.854  < 2e-16 ***
Sub.CategoryBinders           -0.972564   0.056986 -17.067  < 2e-16 ***
Sub.CategoryBookcases          1.402848   0.090267  15.541  < 2e-16 ***
Sub.CategoryChairs             1.309773   0.063958  20.479  < 2e-16 ***
Sub.CategoryCopiers            2.540948   0.153633  16.539  < 2e-16 ***
Sub.CategoryEnvelopes         -1.053558   0.084504 -12.468  < 2e-16 ***
Sub.CategoryFasteners         -2.407585   0.092362 -26.067  < 2e-16 ***
Sub.CategoryFurnishings       -0.805912   0.057880 -13.924  < 2e-16 ***
Sub.CategoryLabels            -1.815722   0.075429 -24.072  < 2e-16 ***
Sub.CategoryMachines           2.052639   0.125614  16.341  < 2e-16 ***
Sub.CategoryPaper             -1.174840   0.053274 -22.053  < 2e-16 ***
Sub.CategoryPhones             0.729193   0.058758  12.410  < 2e-16 ***
Sub.CategoryStorage            0.189966   0.058935   3.223  0.00127 **
Sub.CategorySupplies          -1.039083   0.095159 -10.919  < 2e-16 ***
Sub.CategoryTables             1.640805   0.082224  19.955  < 2e-16 ***
Discount.LevelLow Discount     1.217632   0.056017  21.737  < 2e-16 ***
Discount.LevelMedian Discount  0.918558   0.082770  11.098  < 2e-16 ***
Discount.LevelNo Discount      1.453338   0.056987  25.503  < 2e-16 ***
Quantity                       0.240842   0.005274  45.662  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.052 on 7954 degrees of freedom
Multiple R-squared:  0.584,     Adjusted R-squared:  0.5829
F-statistic: 558.2 on 20 and 7954 DF,  p-value: < 2.2e-16
```

# 04 Experiment - Data Explanation

**Quantity**

1. Positive relationship
2. one unit quantity increase will lead to 0.24 unit increase in average of sales

**DIscount Level**

1. Positive relationship for all levels of discount
2. The coefficient of different levels means the average difference unit increase of sales compared to high discount level

Sales

**Sub Category**

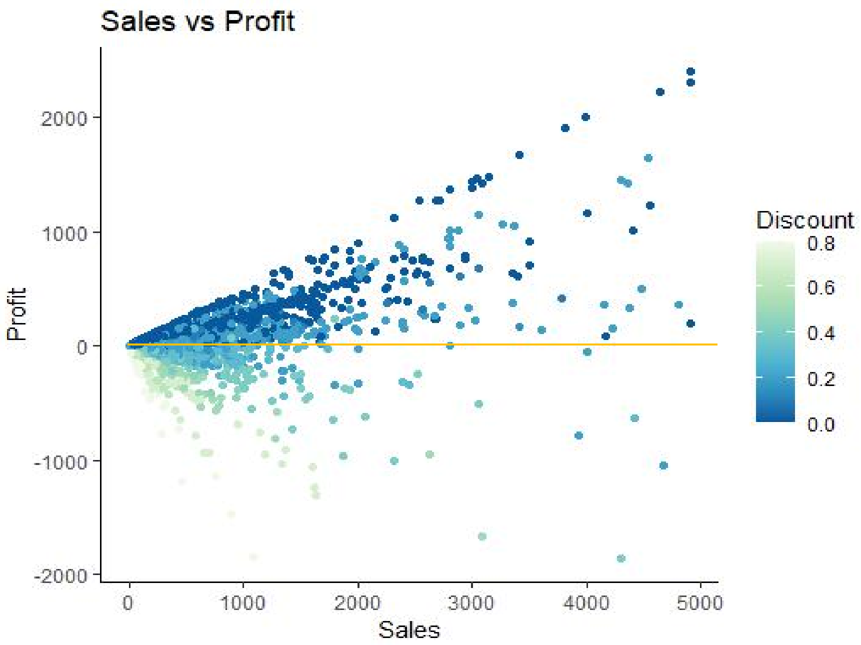
Difference in average of each sub category compared to the basic one - Accessories

# 04 Experiment - Data Explanation

## *Inspiration 01*

- A fit discount level should bring **increase both on sales and profit** contemporarily.

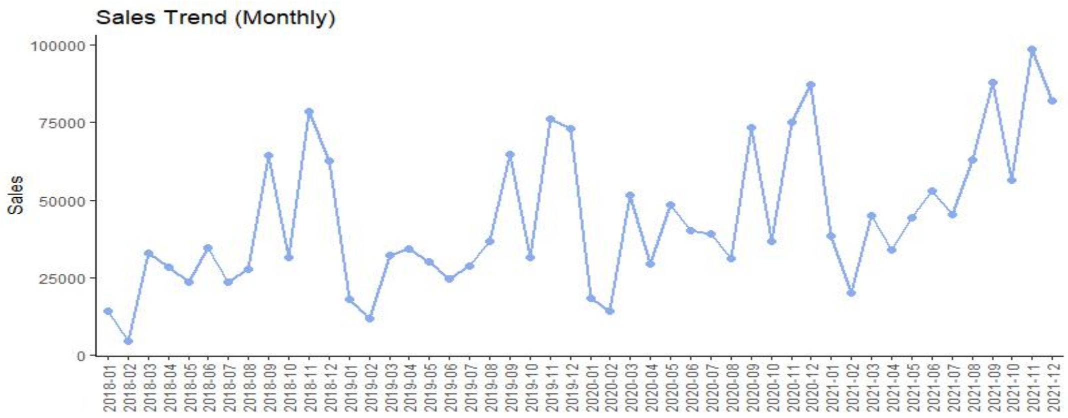
- Our analysis on this dataset shows: **low level discount** (0-0.3) is a range for this super store to guarantee their profitability while boosting the sales.



# Inspiration 02

A **strong seasonality** implies that this superstore may adopt some more flexible strategies to minimize the operational cost in low sales season to maximize their profitability: e.g. reduce stock level in low season, hire temporary employees to cover high season, etc.

Also in low season they could analyze to combine appropriate discount level in low season in order to increase level of sales.



# 04 Experiment - Data Explanation

## *Inspiration 03*

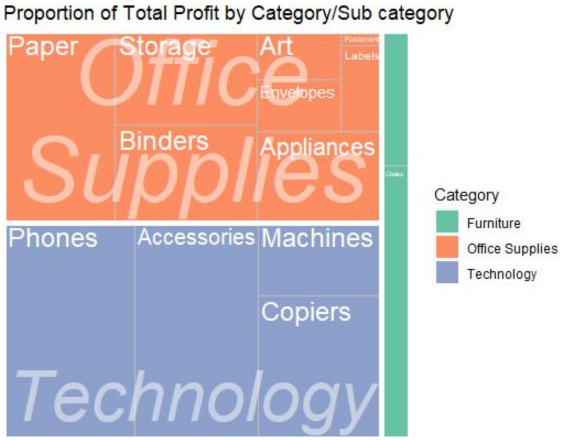
From product category point of view:

There are some products have the equivalent contribution to both sale and profit, for example phones,

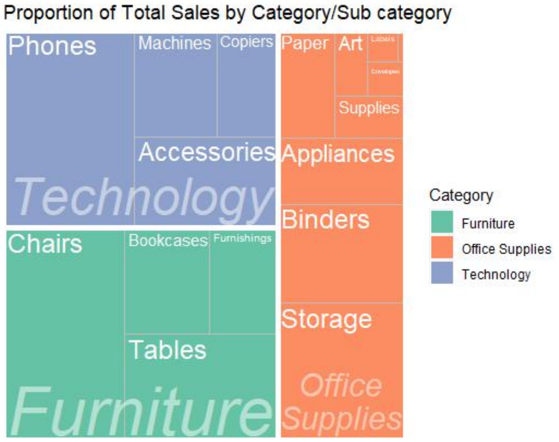
Meanwhile some "best sellers" are not outstanding in profit proportion.

which suggests that this super store could **focus more on products with higher profitability**.

Profit




Sales



# THANKS