IIT School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Group 315: A STATISTICAL ANALYSIS ON GLOBAL SUPERSTORE DATA

First Name	Last Name	Email (hawk.iit.edu)	Student ID
Priyanka	Hosur Mahadevu	pmahadevu@hawk.iit.edu	A20460285
Padmashri	Adgonda Malgonnavar	padgondamalgonnavar@hawk.iit.edu	A20457612

Table of Contents

1. Introduction

The Superstores industry comprises companies that operate by having large size spaces that store and supply large amounts of goods. The superstore industry is comprised of extensive stores that sell a typical product line of grocery items and merchandise products, such as food, pharmaceuticals, cosmetics and personal care items, health products, games and toys, furniture, and appliances. Superstore provides a membership fee for consumers to shop within the store and once a member, the superstore provides consumers a broad array of products for discounted costs.

The superstore industry is part of the retail trade market. Most of the products bought at superstores are used by other wholesalers and smaller retail businesses for their own companies. There is constant competition between the superstores and supercenters with many merchandisers, department stores, wholesalers, and grocery stores. Large superstores and superstore chains are predominant in this market because of their economics of scale in financing, purchasing, and distributing.

To analyze such an industry is of great importance and induced us as it gives insights into the sales and profits of various products. Our analysis is based on global superstore data where the products are ordered between the years 2011-2015. Here, in this superstore data we analyze and discover various aspects that determine the profit of superstore based on some parameters like discount, products, and sales.

2. Data

Our analysis is based on a retail dataset of a global superstore from the year 2011 – 2015 (4 years) and the dataset belongs to the retail domain with **51290 observations**. We are exploring the relationship between sales against different market groups and sales against different product category groups where we are trying to predict the profit (dependent variable) with the help of the information contained in the other variable with a **95% confidence level**. The final model would be useful for the superstore manager to predict the store profit with respect to qualified independent variables such as store market region, category of product, discounts, etc.

To work on this project, we have chosen the dataset from Kaggle and retrieved from the link: https://www.kaggle.com/jr2ngb/superstore-data

Attribute Name	Description	Attribute Data Type
Row ID	Unique ID for each row	Quantitative
Order ID	ID assigned to the Customer's Order	Qualitative
Order Date	Order date of the product	Quantitative
Ship Date	Shipping date of the product	Quantitative
Ship Mode	Mode of shipping (standard, first and second class)	Qualitative
Customer ID	ID assigned to the Customer	Qualitative
Customer Name	Name of a Customer	Qualitative
Segment	Type of business section	Qualitative
City	Location of superstore	Qualitative
State	Location of superstore	Qualitative
Country	Location of superstore	Qualitative
Postal Code	Location postal code	Quantitative
Market	Name of the continent	Qualitative
Region	Geographical business area	Qualitative
Product ID	ID assigned to the Product	Qualitative
Category	Product category name	Qualitative
Sub-Category	Product sub-category name	Qualitative
Product Name	Name of the Product	Qualitative
Sales	Number of sales	Quantitative
Quantity	Number of quantities	Quantitative
Discount	Discount on product	Quantitative
Profit	Profit of a company	Quantitative
Shipping Cost	Shipping cost of a product order	Quantitative
Order Priority	Order priority segments	Qualitative

3. Problems to be Solved

- Do superstore have different Sales with respect to different groups of Market regions.
- Whether the group of product categories in the superstore have the same sales or not.
- Build multiple linear regression models to predict profit of superstore with respect to region-wise, sales-wise, product wise sales etc.

4. Solutions

To address the above problems, we wish to work towards achievement of following solutions:

- **Do superstore have different Sales with respect to different groups of Market regions.**
  - As we observed that the Market has a group within groups, we are implementing ANOVA to solve the problem by building the ANOVA regression model.
  - To check model assumptions, we will perform residual analysis, then perform F-test and compare p-value with significance level to accept or reject the Null Hypothesis.
  - At last, we check the adjusted p value of each slope in the t-test to know the statistically significant difference among all the market group.
  - We consider Market as the categorical variable and Sales as Quantitative variable.

- **Whether the group of product categories in the superstore have the same sales or not.**
  - As we observed that the product category has groups within groups, we are implementing ANOVA to solve the problem by building the ANOVA regression model.
  - We perform F-test and compare the p-value with the significance level to accept or reject the Null Hypothesis.
  - At last, will look at the p-value of each slope in the t-test to know the statistically significant difference among all the category group.
  - We consider Category as the categorical variable and Sales as Quantitative variable.

- **To build multiple regression model for predicting profit.**
  - We consider profit as the dependent variable in building multiple linear regression model.
  - Sales, discount, shipping_cost, ship_mode, segment, region, sub_category, order_priority etc., are considered as independent variable in building multiple linear regression model.
  - As postal code is the only column has missing value and we remove that column as it is not useful for statistical analysis.
  - Also, we remove row_id, order_id, order_date, customer_id, customer_name, product_id, product_name has these columns are not useful for any statistical analysis.
  - According to the size of the dataset, we choose the **hold-out evaluation** method to split the data by considering **70%** of the data as training data and **30%** of the data as testing data.
  - And asses the multicollinearity issue with the VIF method.
  - By using feature selection will build different models by parameter estimates and check the goodness of fit. And perform residual analysis to validate the model is qualified or not and rebuild the model until the model is qualified. Look for influential points if exists remove and rebuild the model.

5. Experiments and Results

5.1. Methods and Process

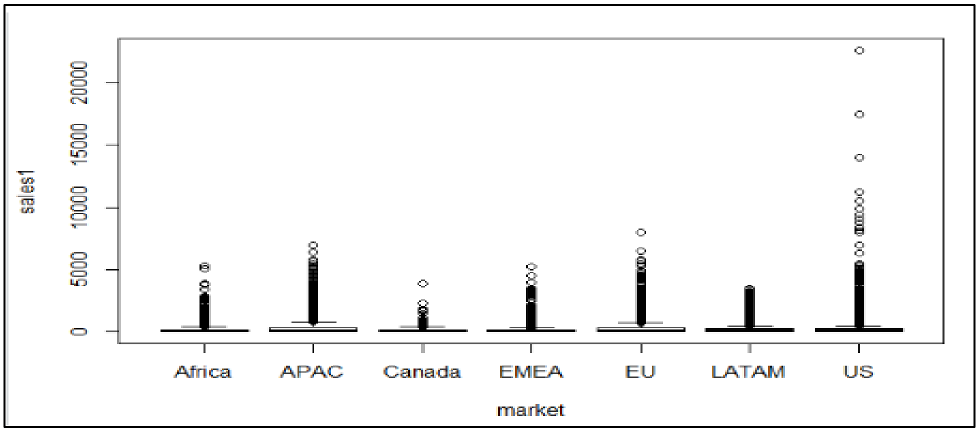
5.1.1 Do superstore have different Sales with respect to different groups of Market regions.

OBJECTIVE: Compare group means among more than two market groups by analyzing the variances to see if they are significantly different.

PRELIMINARY ANALYSIS

A best practice before performing the ANOVA in R is to visualize the data in relation to the research question. The best method to do so is to draw boxplots of the quantitative variable Sales for each market regions.

```
# Creating new anovadf2 dataframe
anovadf2 <- superstore_data[, c("sales", "market")]
head(anovadf2)
sales1 = anovadf2$sales
market = anovadf2$market
# Diffrences among market groups are not visible through side by side box plots
plot(sales1~market)
```



From the above boxplot, sales v/s market we observe that differences among market groups are not visible and are not clear through side by side box plots.

FURTHER ANALYSIS

1. **F-test**

   Null Hypothesis: Average sales in all market groups are the same.

   Alternative Hypothesis: Average sales in all market groups are not same i.e., at least one or two group have different mean.

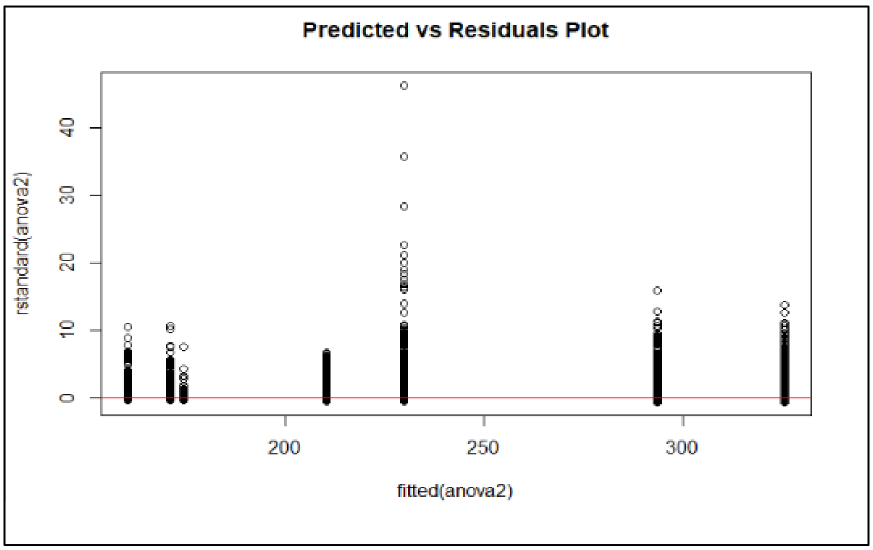
```
> # Build anova model for sales~ market
> anova2=lm(sales1~market)
> summary(anova2)

Call:
lm(formula = sales1 ~ market)

Residuals:
    Min      1Q  Median      3Q     Max 
 -323.0  -205.8  -142.8     8.0 22408.6 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    170.868      7.148  23.905  < 2e-16 ***
marketAPAC     155.049      8.508  18.223  < 2e-16 ***
marketCanada     3.424     25.718   0.133    0.894    
marketEMEA     -10.566      9.884  -1.069    0.285    
marketEU       122.941      8.633  14.241  < 2e-16 ***
marketLATAM     39.410      8.594   4.586 4.54e-06 ***
marketUS        58.990      8.634   6.832 8.44e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 484.1 on 51283 degrees of freedom
Multiple R-squared:  0.01424,	Adjusted R-squared:  0.01413 
F-statistic: 123.5 on 6 and 51283 DF,  p-value: < 2.2e-16
```

2. **Individual parameter test**
   Null Hypothesis: Difference of variables are statistically significant.
   Alternative Hypothesis: Difference of variable are not statistically significant.

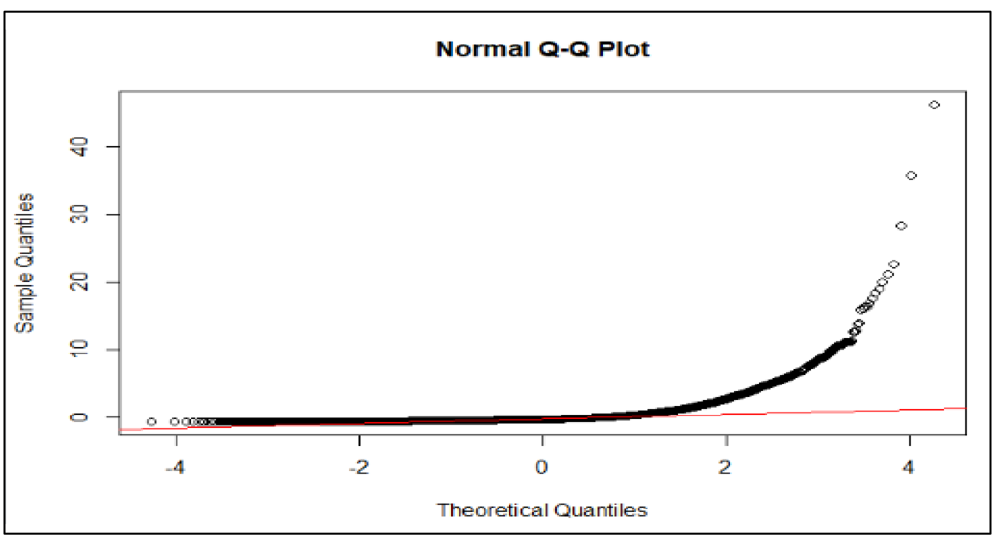
From the above summary of ANOVA model, we observe that p-value (2.2e-16) < α (0.05) but we cannot interpret for F-test and Individual parameter. In order to check the model assumption, we will perform the residual analysis first.

3. **Residual analysis**

- Constance Variance



- Normality Test



INTERPRETATION

- In the plot predicted vs residuals we observe that spread is not constant from the plot.
- From the Q-Q plot, we observe that the points are not around the line and are not normal.

Finally, we can conclude that there are no linearity and normality from the plots. Hence, we perform transformation.

TRANSFORMATION

The replacement of a variable by a function of that variable is called transformation in data analytics. The transformation can be performed by log, square root and inverse.

- Logarithm transformation

```
> # log Transformation on Sales1
> anova21=lm(log(sales1)~market)
> summary(anova21)

Call:
lm(formula = log(sales1) ~ market)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9217 -1.0476 -0.0886  1.0084  5.9177 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.08088    0.02091 195.167   <2e-16 ***
marketAPAC     0.79133    0.02489  31.794   <2e-16 ***
marketCanada   0.17384    0.07523   2.311   0.0209 *  
marketEMEA    -0.03048    0.02891  -1.054   0.2918    
marketEU       0.77551    0.02525  30.708   <2e-16 ***
marketLATAM    0.38781    0.02514  15.426   <2e-16 ***
marketUS       0.02888    0.02526   1.143   0.2529    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.416 on 51283 degrees of freedom
Multiple R-squared:  0.05649,	Adjusted R-squared:  0.05638 
F-statistic: 511.8 on 6 and 51283 DF,  p-value: < 2.2e-16
```

- Square root transformation

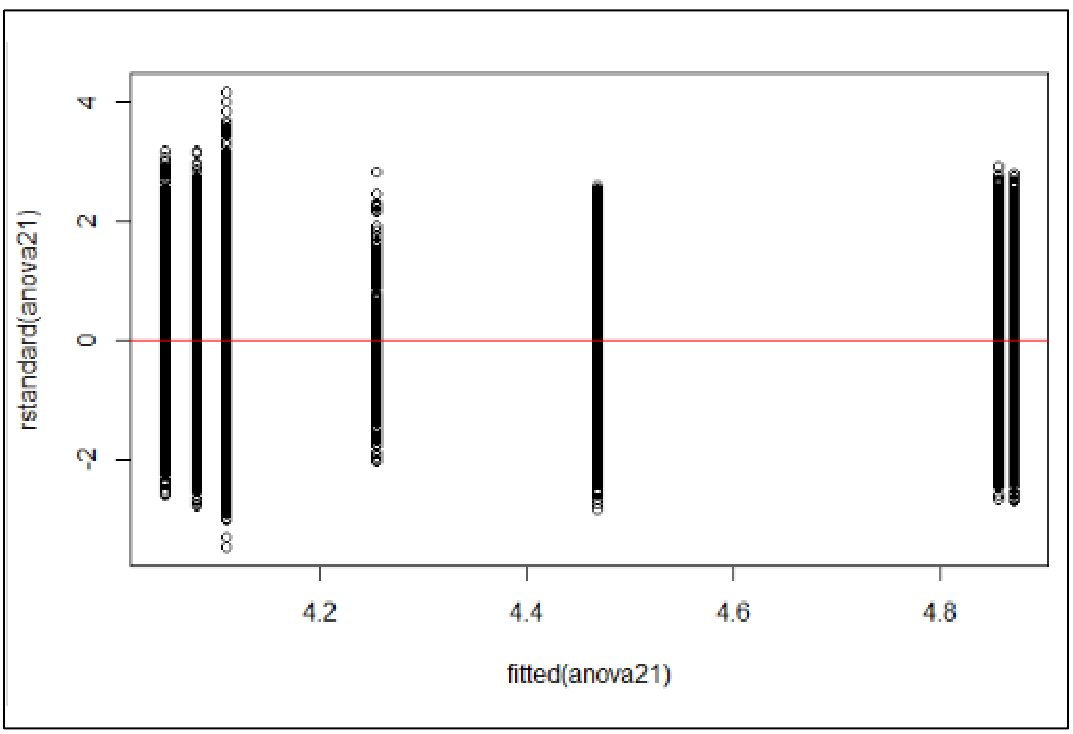
```
> # sqrt Transformation
> anova22=lm(sqrt(sales1)~market)
> summary(anova22)

Call:
lm(formula = sqrt(sales1) ~ market)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.778  -6.488  -3.015   3.481 139.452 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   10.0562     0.1423  70.670  < 2e-16 ***
marketAPAC     4.4186     0.1694  26.087  < 2e-16 ***
marketCanada   0.4811     0.5120   0.940    0.347    
marketEMEA    -0.2382     0.1968  -1.211    0.226    
marketEU       3.9396     0.1719  22.923  < 2e-16 ***
marketLATAM    1.6590     0.1711   9.696  < 2e-16 ***
marketUS       0.9530     0.1719   5.545 2.96e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.637 on 51283 degrees of freedom
Multiple R-squared:  0.03175,	Adjusted R-squared:  0.03164 
F-statistic: 280.3 on 6 and 51283 DF,  p-value: < 2.2e-16
```

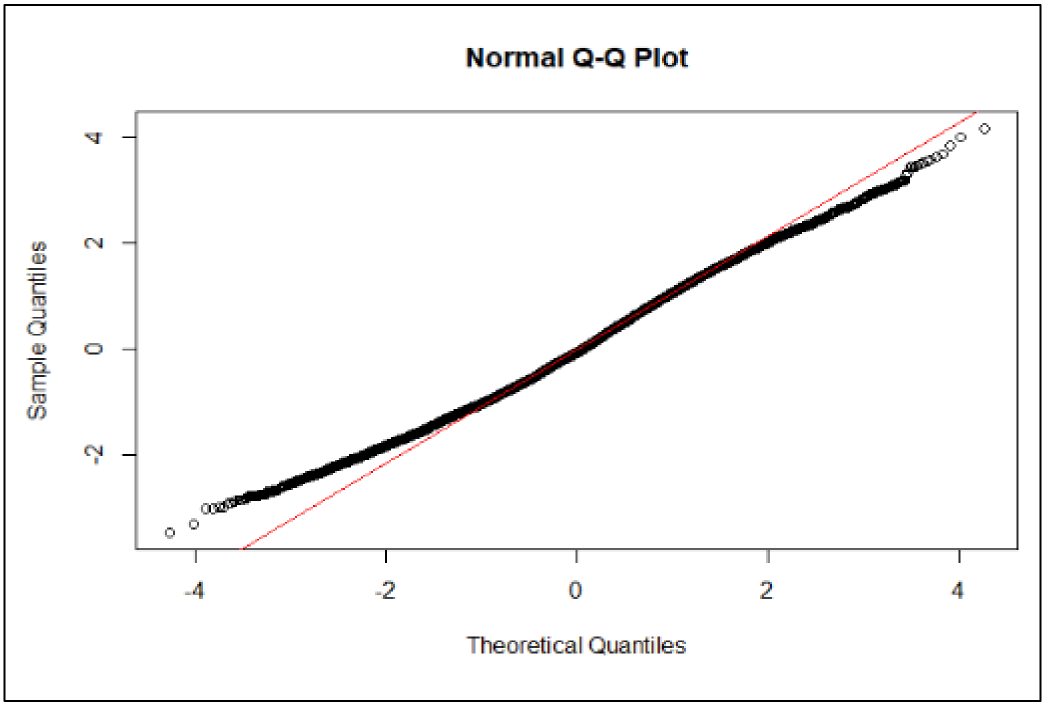
- Inverse transformation

```
> # inverse Transformation
> anova23=lm((1/sales1)~market)
> summary(anova23)

Call:
lm(formula = (1/sales1) ~ market)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.05073 -0.02240 -0.01209  0.00487  2.20148 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.0424067  0.0008241  51.457   <2e-16 ***
marketAPAC   -0.0249902  0.0009810 -25.475   <2e-16 ***
marketCanada -0.0123752  0.0029651  -4.174    3e-05 ***
marketEMEA   -0.0014884  0.0011396  -1.306    0.192    
marketEU     -0.0261913  0.0009953 -26.314   <2e-16 ***
marketLATAM  -0.0167885  0.0009909 -16.943   <2e-16 ***
marketUS      0.0083699  0.0009954   8.408   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05582 on 51283 degrees of freedom
Multiple R-squared:  0.0559,	Adjusted R-squared:  0.05579 
F-statistic: 506.1 on 6 and 51283 DF,  p-value: < 2.2e-16
```

RESIDUAL PLOT FOR ALL THE TRANSFORMATION

Now we perform residual plot for all the above transformation, and we observe for the plot having linearity and normality. The below plots displayed are for log transformation.





INTERPRETATION

- In the plot predicted vs residuals, we observe that variable is scattered around zero line and there is linearity.
- From the Normality distribution Q-Q plot, we observe that the points are distributed around normal line.
- Finally, we can conclude that, there is linearity and normality with log transformation of Sales, and we further use this log transformation for building the ANOVA model.

BUILD ANOVA MODEL

ANOVA model is build using aov() function with log transformation of Sales, which is used to determine if the means of two or more groups are differ significantly from each other. Then perform F-test and compare p-value with significance level to accept or reject the Null Hypothesis.

```
> anovaaa = aov(anova21)
> summary(anovaaa)
               Df Sum Sq Mean Sq F value Pr(>F)    
market          6   6158    1026   511.8 <2e-16 ***
Residuals   51283 102848       2                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

F-TEST INTERPRETATION

- The F-test statistic p-value(2e-16) < α (0.05).
- As p-value < α, we don't have enough evidence to accept Null hypothesis.
- From the F-test result with log transformation of Sales, we can conclude that "With 95% confidence level at least one market group have different average sales".

INDIVIDUAL PARAMETER TEST

To investigate the difference between all market region and to know which ones are different, the type of test that we perform is post-hoc test. However, result of ANOVA do not tell us which market region groups are different from the others. The post-hoc tests mean (in Latin, "after this", so after obtaining statistically significant ANOVA results).

One of the post-hoc test that we are using is **Tukey HSD test**. This test is used to compare all groups to each other. We use **TukeyHSD()** function with log transformation of Sales.

```
> data.test <- TukeyHSD(anovaaa, conf.level=0.95)
> data.test
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = anova21)

$market
                     diff          lwr         upr     p adj
APAC-Africa    0.79133332  0.717950457  0.86471618 0.0000000
Canada-Africa  0.17383585 -0.047972351  0.39564405 0.2386691
EMEA-Africa   -0.03047805 -0.115724783  0.05476868 0.9411534
EU-Africa      0.77550629  0.701049500  0.84996309 0.0000000
LATAM-Africa   0.38781251  0.313690819  0.46193420 0.0000000
US-Africa      0.02887833 -0.045585493  0.10334215 0.9146395
Canada-APAC   -0.61749747 -0.834252834 -0.40074210 0.0000000
EMEA-APAC     -0.82181137 -0.892881869 -0.75074087 0.0000000
EU-APAC       -0.01582703 -0.073514325  0.04186027 0.9841587
LATAM-APAC    -0.40352081 -0.460774940 -0.34626668 0.0000000
US-APAC       -0.76245499 -0.820151361 -0.70475862 0.0000000
EMEA-Canada   -0.20431390 -0.425367856  0.01674005 0.0920491
EU-Canada      0.60167044  0.384549141  0.81879174 0.0000000
LATAM-Canada   0.21397666 -0.003029956  0.43098327 0.0562305
US-Canada     -0.14495752 -0.362081234  0.07216619 0.4349844
EU-EMEA        0.80598434  0.733805499  0.87816319 0.0000000
LATAM-EMEA     0.41829056  0.346457443  0.49012368 0.0000000
US-EMEA        0.05935638 -0.012829716  0.13154247 0.1882601
LATAM-EU      -0.38769378 -0.446318055 -0.32906951 0.0000000
US-EU         -0.74662797 -0.805684216 -0.68757171 0.0000000
US-LATAM      -0.35893418 -0.417567377 -0.30030098 0.0000000
```

T-TEST INTERPRETATION

- From the Tukey's test with log transformation of Sales, we conclude that there is significant difference in all other market group at adjusted p-value < 0.05, except between the groups Canada-Africa, EMEA-Africa, US-Africa, EU-APAC, EMEA-Canada, LATAM-Canada, US-EMEA and US-Canada.
- The APAC Market group has larger sales and Africa has least number of sales.

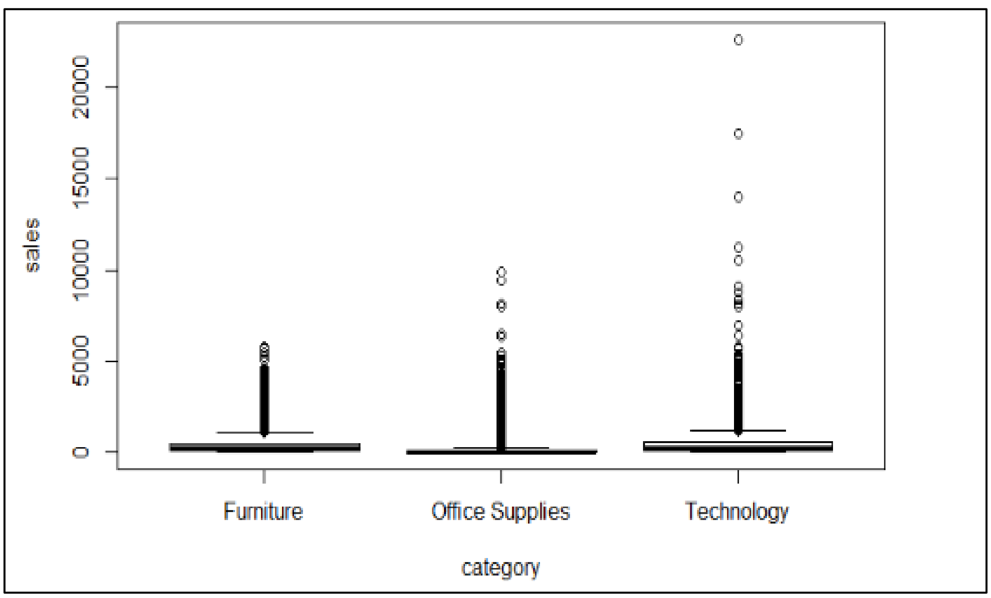
5.1.2 Whether the group of product categories in the superstore have the same sales or not.

OBJECTIVE : Compare group means among more than two category groups by analyzing the variances to see if they are significantly different.

PRELIMINARY ANALYSIS

A best practice before performing the ANOVA in R is to visualize the data in relation to the research question. The best method to do so is to draw boxplots of the quantitative variable Sales for each category.

```
# Creating new anovadf1 dataframe
anovadf1 <- superstore_data[, c("sales", "category")]
# Removing space between category groups
#anovadf1$category = sub(' ', '', anovadf1$category)
head(anovadf1)
anovadf1
sales = anovadf1$sales
category = anovadf1$category
# Diffrences among category groups are not visible through side by side box plots
plot(sales~category)
```



From the above boxplot, sales v/s category, we observe that differences among category groups are not visible and are not clear through side by side box plots.

FURTHER ANALYSIS

1. **F-test**

   Null Hypothesis: The mean of sales in all category is same.

   Alternative Hypothesis: The mean of sales in all category is not same i.e., at least one or two groups have different mean.

```
> anova1=lm(sales~category)
> summary(anova1)

Call:
lm(formula = sales ~ category)

Residuals:
    Min      1Q  Median      3Q     Max 
 -466.9  -116.5   -85.6     2.1 22170.6 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              416.249      4.643  89.651  < 2e-16 ***
categoryOffice Supplies -295.152      5.326 -55.418  < 2e-16 ***
categoryTechnology        51.610      6.523   7.912 2.59e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 461.4 on 51287 degrees of freedom
Multiple R-squared:  0.1044,	Adjusted R-squared:  0.1044 
F-statistic:  2990 on 2 and 51287 DF,  p-value: < 2.2e-16
```

2. **Individual parameter test**

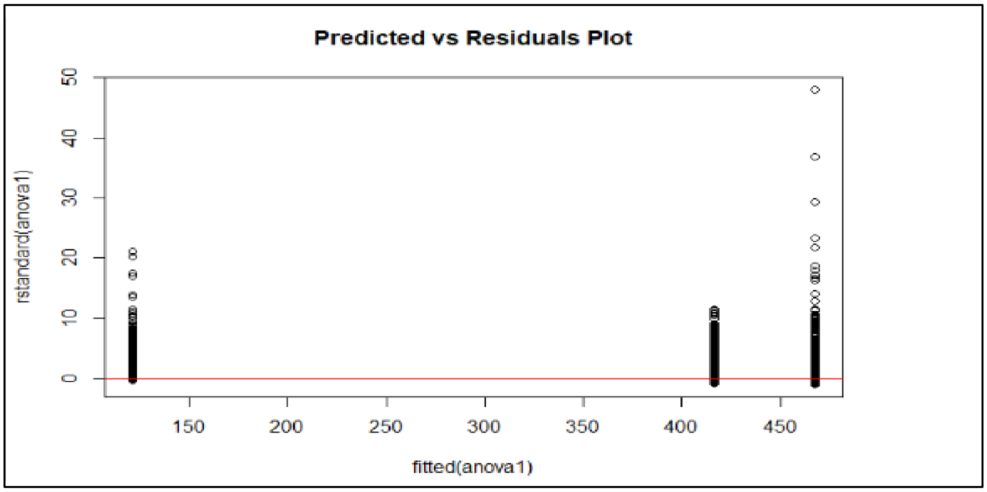
   Null Hypothesis: Difference of variables are statistically significant.

   Alternative Hypothesis: Difference of variable are not statistically significant.

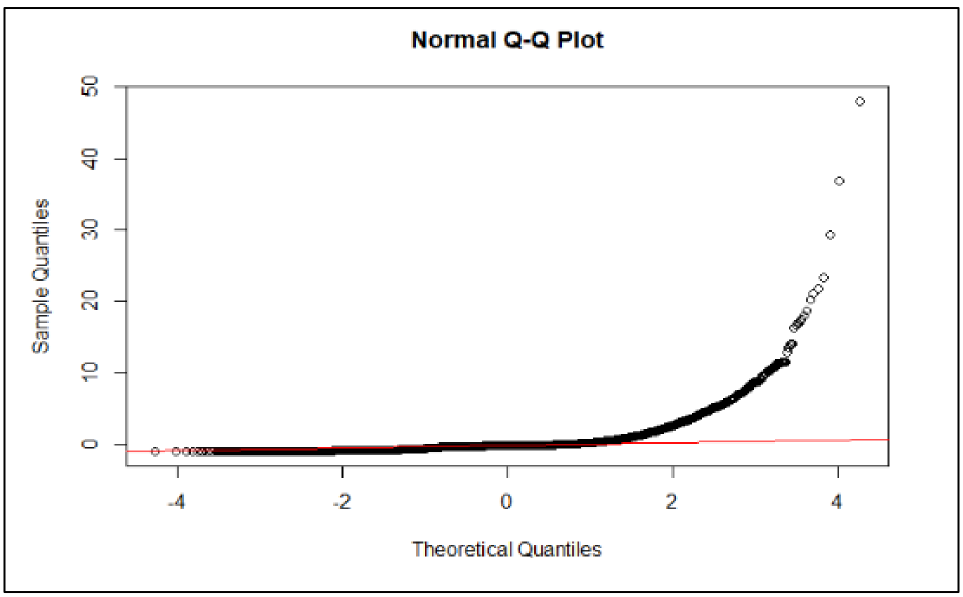
From the above summary of ANOVA model, we observe that p-value (2.2e-16) < α (0.05) but we cannot interpret for F-test and Individual parameter. In order to check the model assumption, we will perform the residual analysis first.

3. **Residual analysis**

- Constance Variance



- Normality Test



INTERPRETATION

- In the plot predicted vs residuals we observe that spread is not constant from the plot.
- From the Q-Q plot, we observe that the points are not around the line and are not normal.

As we observed, there are no linearity and normality from the plots. So, we perform transformation.

TRANSFORMATION

The replacement of a variable by a function of that variable is called transformation in data analytics. The transformation can be performed by log, square root and inverse.

- Logarithm transformation

```
> anova11=lm(log(sales)~category)
> summary(anova11)

Call:
lm(formula = log(sales) ~ category)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5350 -0.8349 -0.0067  0.8075  5.3216 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              5.33217    0.01252  425.81   <2e-16 ***
categoryOffice Supplies -1.45423    0.01436 -101.24   <2e-16 ***
categoryTechnology       0.19276    0.01759   10.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.244 on 51287 degrees of freedom
Multiple R-squared:  0.2714,	Adjusted R-squared:  0.2713 
F-statistic:  9550 on 2 and 51287 DF,  p-value: < 2.2e-16
```

- Square root transformation

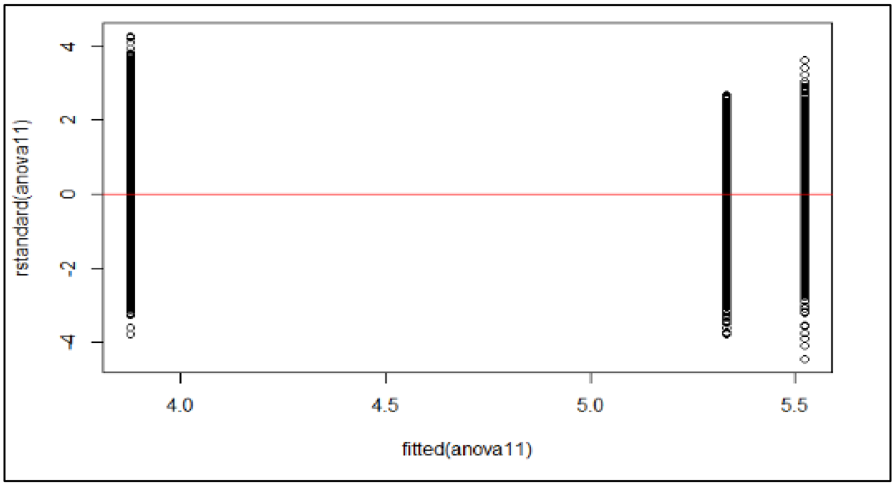
```
> # sqrt Transformation
> anova12=lm(sqrt(sales)~category)
> summary(anova12)

Call:
lm(formula = sqrt(sales) ~ category)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.587  -4.880  -1.935   2.672 131.879 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             17.35538    0.08710  199.25   <2e-16 ***
categoryOffice Supplies -8.73662    0.09992  -87.44   <2e-16 ***
categoryTechnology       1.22699    0.12238   10.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.656 on 51287 degrees of freedom
Multiple R-squared:  0.2188,	Adjusted R-squared:  0.2188 
F-statistic:  7183 on 2 and 51287 DF,  p-value: < 2.2e-16
```

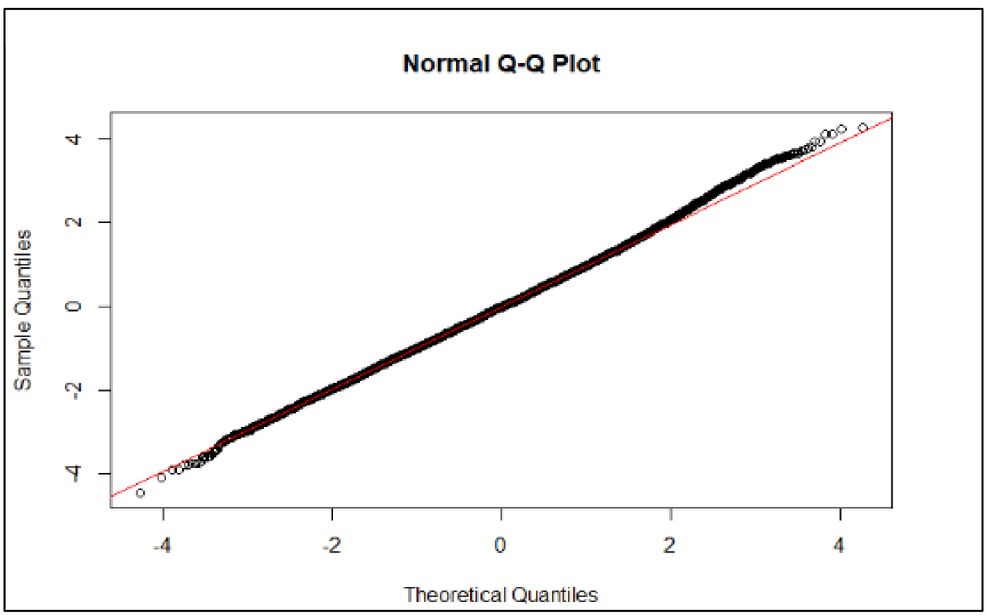
- Inverse transformation

```
> # inverse Transformation
> anova13=lm((1/sales)~category)
> summary(anova13)

Call:
lm(formula = (1/sales) ~ category)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.04272 -0.02592 -0.00751  0.00181  2.20943 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              0.0116080  0.0005548  20.922  < 2e-16 ***
categoryOffice Supplies  0.0312105  0.0006364  49.040  < 2e-16 ***
categoryTechnology      -0.0034163  0.0007795  -4.383 1.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05514 on 51287 degrees of freedom
Multiple R-squared:  0.07861,	Adjusted R-squared:  0.07857 
F-statistic:  2188 on 2 and 51287 DF,  p-value: < 2.2e-16
```

RESIDUAL PLOT FOR ALL THE TRANSFORMATION

Now we perform residual plot for all the above transformation, and we observe for the log transformed plot having linearity and normality.





INTERPRETATION

- In the plot predicted vs residuals, we observe that variable is scattered around zero line and there is linearity.
- From the Normality distribution Q-Q plot, we observe that the points are distributed around normal line.
- Finally, we can conclude that, there is linearity and normality with log transformation of Sales, and we further use this log transformation for building the ANOVA model.

BUILD ANOVA MODEL

- ANOVA model is build using aov() function with log transformation of Sales, which is used to determine if the means of two or more groups are differ significantly from each other. Then perform F-test and compare p-value with significance level to accept or reject the Null Hypothesis.

```
> anovaa = aov(anova11)
> summary(anovaa)
               Df Sum Sq Mean Sq F value Pr(>F)    
category        2  29580   14790    9550 <2e-16 ***
Residuals   51287  79426       2                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

F-TEST INTERPRETATION

- The F-test statistic p-value(2e-16) < α (0.05).
- As p-value < α, we don't have enough evidence to accept Null hypothesis.
- From the F-test result with log transformation, we can conclude that "With 95% confidence level at least one category group have different mean sales".

INDIVIDUAL PARAMETER TEST

To investigate the difference between all category and to know which ones are different, the type of test that we perform is post-hoc test. However, result of ANOVA do not tell us which groups are different from the others. The post-hoc tests mean (in Latin, "after this", so after obtaining statistically significant ANOVA results).

One of the post-hoc test that we are using is **Tukey HSD test**. This test is used to compare all category groups to each other. We use **TukeyHSD()** function with log transformation of Sales.

```
> data.test1 <- TukeyHSD(anovaa, conf.level=0.95)
> data.test1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = anova11)

$category
                                diff       lwr        upr p adj
Office Supplies-Furniture -1.4542272 -1.487893 -1.4205617     0
Technology-Furniture       0.1927634  0.151530  0.2339968     0
Technology-Office Supplies 1.6469906  1.613661  1.6803201     0
```

T-TEST INTERPRETATION

- From the Tukey's test with log transformation of Sales, we conclude that there is significant difference in all category group at adjusted p-value < 0.05.
- The technology category has higher sales and office-supplies category has least sales.

5.1.3 Build Multiple Linear Regression model.

OBJECTIVE: Building multiple linear regression model to predict superstore profit.

PREPROCESSING THE DATA

a) **Check for Missing value**

```
> ###############################################################################
> ### Check for missing Records
> #        Only postal_code has missing records and postal code is not useful for
> #        Statistical Analysis.
> ###############################################################################
> na_count = sapply(superstore_data, function(x) sum(is.na(x)))
> na_count = data.frame(na_count)
> na_count
              na_count
row_id               0
order_id             0
order_date           0
ship_date            0
ship_mode            0
customer_id          0
customer_name        0
segment              0
city                 0
state                0
country              0
postal_code      41296
market               0
region               0
product_id           0
category             0
sub_category         0
product_name         0
sales                0
quantity             0
discount             0
profit               0
shipping_cost        0
order_priority       0
```

In our dataset, we observe that the attribute postal_code as missing value. We remove the column postal code as it is not useful for any statistical analysis.

```
> # Removing these ID columns as they are not useful for statistical Analysis
> superstore_data$row_id <- NULL
> superstore_data$order_id <- NULL
> superstore_data$order_date <- NULL
> superstore_data$customer_id <- NULL
> superstore_data$customer_name <- NULL
> superstore_data$postal_code <- NULL
> superstore_data$product_id <- NULL
> superstore_data$product_name <- NULL
> # Get the number of rows and coulmns of data
> dim(superstore_data)
[1] 51290    16
```

Since, the ID columns are not useful for statistical analysis, we remove them.

b) Check collinearity

To check the collinearity between region and market, we perform Chi-Square test. In order to know if two features are independent or dependent.

Null Hypothesis: Market and Regions are independent
Alternate Hypothesis: Market and regions are dependent/relationship exists.

```
> # Fetching Categorical variables
> categorical=superstore_data %>% select_if(negate(is.numeric))
> # Chi-Square-Test to check the relationship between market and region
> chitest1 = table(categorical$market, categorical$region)
> chisq.test(chitest1)

        Pearson's Chi-squared test

data:  chitest1
X-squared = 239550, df = 72, p-value < 2.2e-16
```

CHI-SQUARE -TEST INTERPRETATION

- From the chi-square-test, we observe that p-value < 0.05. Hence, we reject null hypothesis and accept alternate hypothesis accepting that two features are not independent.
- We ignore the column Market while building the regression model.

c) Create Dummy variable

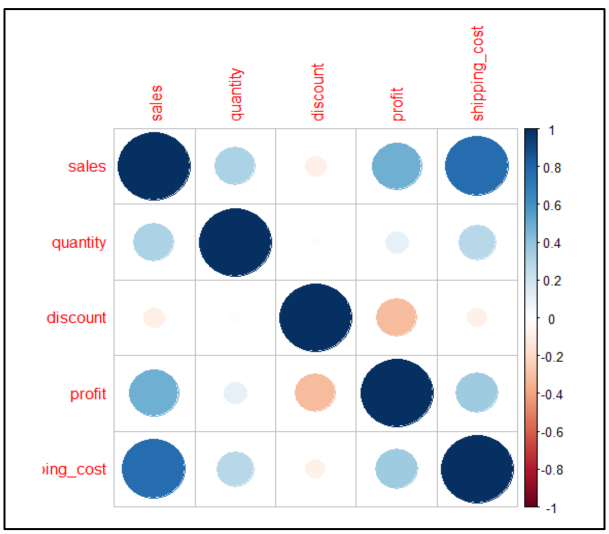
We create N-1 Dummy variables.

```
# Creating n-1 dummy variables
dummydf<- data.frame(sapply(categorical,function(x) data.frame(model.matrix(~x-1,data =categorical))[,-1]))
dim(dummydf)

dummydf = clean_names(dummydf)
names(dummydf)
```

```
> # As p-value is less than 0.05 indicates market and region are not ind
ependent
> # hence ignoring market for further process
> categorical$market <- NULL
> # Creating n-1 dummy variables
> dummydf<- data.frame(sapply(categorical,function(x) data.frame(model.m
atrix(~x-1,data =categorical))[,-1]))
> dim(dummydf)
[1] 51290    36
> dummydf = clean_names(dummydf)
```

d) Check the correlation between variables and transform the variables if necessary.

```
> ########################################################################
> # Correlation
> ########################################################################
> checknumericvar = sapply(superstore_data, is.numeric)
> # Fetching numeric features
> numericvar = superstore_data[checknumericvar]
> # checking Correlation between numeric features
> corr=cor(numericvar)
> #install.packages("corrplot")
> library(corrplot)
> corrplot(corr, method="circle")
> corr
                   sales   quantity    discount     profit shipping_cost
sales          1.00000000  0.3135772 -0.08672187  0.4849181    0.76807284
quantity       0.31357718  1.0000000 -0.01987470  0.1043650    0.27264897
discount      -0.08672187 -0.0198747  1.00000000 -0.3164902   -0.07905555
profit         0.48491811  0.1043650 -0.31649017  1.0000000    0.35444090
shipping_cost  0.76807284  0.2726490 -0.07905555  0.3544409    1.00000000
```



INTERPRETATION

- We have observed that there is a weak correlation. So, we need to do transformations on variables in order to increase the correlation with dependent variable.

TRANSFORMATION

```
> # Applying Transformation to improve weak correlation
> # Transformation Quantity
> quantity1 = log(numericvar$quantity)
> cor(numericvar$profit, quantity1)
[1] 0.1006548
> quantity2 = sqrt(numericvar$quantity)
> cor(numericvar$profit, quantity2)
[1] 0.1043998
> quantity3 = (1/numericvar$quantity)
> cor(numericvar$profit, quantity3)
[1] -0.0862773
> # Transformation Sales
> sales1 = log1p(numericvar$sales)
> cor(numericvar$profit, sales1)
[1] 0.2689784
> sales2 = sqrt(numericvar$sales)
> cor(numericvar$profit, sales2)
[1] 0.3934623
> sales3 = (1/numericvar$sales)
> cor(numericvar$profit, sales3)
[1] -0.08086905
> # Transformation Discount
> discount1 = log1p(numericvar$discount)
> cor(numericvar$profit, discount1)
[1] -0.3149576
> discount2 = sqrt(numericvar$discount)
> cor(numericvar$profit, discount2)
[1] -0.2953832
> discount3 = (1/numericvar$discount)
> cor(numericvar$profit, discount3)
[1] NaN
> # Transformation of shipping cost
> shipping_cost1 = log1p(numericvar$shipping_cost)
> cor(numericvar$profit, shipping_cost1)
[1] 0.257276
> shipping_cost2 = sqrt(numericvar$shipping_cost)
> cor(numericvar$profit, shipping_cost2)
[1] 0.3285132
> shipping_cost3 = (1/numericvar$shipping_cost)
> cor(numericvar$profit, shipping_cost3)
[1] NaN
```

We have applied transformation to improvise the weak correlation. By observing, Correlation of quantity has not been increased.

Hence, we remove quantity from the dataset.

```
> # No improvement in corr of Qunatity hence removing quantity feature
> numericvar$quantity <- NULL
> corr=cor(numericvar)
> final_df=data.frame(numericvar, dummydf)
```

Now, we have completed the preprocessing of Data. The final Dataset is final_df.

DATA SPLIT:

We have two ways of splitting up data 1) Hold-out Evaluation and 2) N-Fold cross validation. We have used Hold out evaluation Techniques to split my data as data size is large. We have used 70% of total rows to train our model and 30% of total rows to test our model.

BUILD THE REGRESSION MODEL

i. **Build full model without transforming any features**

```
> # Building Full Model
> fullmodel = lm(profit~., data = train.data)
> summary(fullmodel) #Adj-R2 = 0.3337

Call:
lm(formula = profit ~ ., data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-5659.2   -27.9    -6.2    36.1  5295.0

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.459e+01  4.955e+00   6.980 3.01e-12 ***
sales                        1.759e-01  2.511e-03  70.074  < 2e-16 ***
discount                    -2.146e+02  3.621e+00 -59.277  < 2e-16 ***
shipping_cost                2.597e-02  2.170e-02   1.197 0.231381
ship_mode_x_same_day        -4.299e+00  3.743e+00  -1.148 0.250809
ship_mode_x_second_class     3.300e-01  2.559e+00   0.129 0.897372
ship_mode_x_standard_class   2.032e-01  2.348e+00   0.087 0.931021
segment_x_corporate         -5.262e-01  1.698e+00  -0.310 0.756577
segment_x_home_office       -8.968e-01  2.002e+00  -0.448 0.654131
region_x_canada             -9.719e+00  8.844e+00  -1.099 0.271789
region_x_caribbean          -4.418e+00  4.744e+00  -0.931 0.351645
region_x_central            -8.665e+00  2.928e+00  -2.959 0.003088 **
region_x_central_asia       -6.407e+00  4.468e+00  -1.434 0.151597
region_x_east                4.271e+00  3.983e+00   1.072 0.283615
region_x_emea                3.440e+00  3.401e+00   1.012 0.311772
region_x_north              -6.897e+00  3.453e+00  -1.998 0.045776 *
region_x_north_asia         -1.061e+00  4.279e+00  -0.248 0.804116
region_x_oceania            -8.305e+00  3.791e+00  -2.190 0.028499 *
region_x_south              -1.241e+01  3.203e+00  -3.875 0.000107 ***
region_x_southeast_asia     -4.578e+00  3.925e+00  -1.166 0.243462
region_x_west               -2.533e+00  3.884e+00  -0.652 0.514273
sub_category_x_appliances   -1.850e+01  5.003e+00  -3.698 0.000218 ***
sub_category_x_art          -1.981e+00  3.843e+00  -0.515 0.606244
sub_category_x_binders       1.070e+01  3.693e+00   2.898 0.003762 **
sub_category_x_bookcases    -3.952e+01  4.585e+00  -8.618  < 2e-16 ***
sub_category_x_chairs       -2.753e+01  4.149e+00  -6.636 3.26e-11 ***
sub_category_x_copiers      -5.074e+00  4.712e+00  -1.077 0.281620
sub_category_x_envelopes     1.656e+00  4.546e+00   0.364 0.715632
sub_category_x_fasteners     2.780e+00  4.554e+00   0.610 0.541592
sub_category_x_furnishings  -4.285e-02  4.210e+00  -0.010 0.991878
sub_category_x_labels        7.615e-01  4.465e+00   0.171 0.864588
sub_category_x_machines     -4.604e+01  5.313e+00  -8.665  < 2e-16 ***
sub_category_x_paper         2.244e+00  4.125e+00   0.544 0.586370
sub_category_x_phones       -1.985e+01  4.169e+00  -4.761 1.93e-06 ***
sub_category_x_storage      -1.503e+01  3.812e+00  -3.942 8.08e-05 ***
sub_category_x_supplies     -6.056e+00  4.588e+00  -1.320 0.186925
sub_category_x_tables       -1.753e+02  6.598e+00 -26.569  < 2e-16 ***
order_priority_x_high       -4.595e+00  3.120e+00  -1.473 0.140771
order_priority_x_low        -2.047e+00  4.741e+00  -0.432 0.665822
order_priority_x_medium     -3.319e+00  3.210e+00  -1.034 0.301171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 139.7 on 35863 degrees of freedom
Multiple R-squared:  0.3345,    Adjusted R-squared:  0.3337
F-statistic: 462.1 on 39 and 35863 DF,  p-value: < 2.2e-16
```

ii. **Check for multicollinearity**

A very simple test to assess multicollinearity in our regression model is by calculating VIF(). The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation.

```
> # Checking for multicollinearity
> vifs=vif(fullmodel)
> vifs
                     sales                  discount             shipping_cost       ship_mode_x_same_day
                  2.688765                  1.090216                  2.634521                  1.287389
  ship_mode_x_second_class ship_mode_x_standard_class       segment_x_corporate      segment_x_home_office
                  1.895156                  2.112500                  1.108582                  1.108380
           region_x_canada         region_x_caribbean           region_x_central      region_x_central_asia
                  1.080995                  1.330678                  2.671942                  1.404337
             region_x_east             region_x_emea             region_x_north        region_x_north_asia
                  1.557080                  1.882662                  1.862237                  1.462461
          region_x_oceania            region_x_south    region_x_southeast_asia             region_x_west
                  1.635891                  2.134926                  1.591354                  1.613918
 sub_category_x_appliances       sub_category_x_art     sub_category_x_binders  sub_category_x_bookcases
                  1.541478                  2.363421                  2.638072                  1.723969
     sub_category_x_chairs     sub_category_x_copiers   sub_category_x_envelopes  sub_category_x_fasteners
                  1.987919                  1.686302                  1.705286                  1.708122
sub_category_x_furnishings     sub_category_x_labels     sub_category_x_machines      sub_category_x_paper
                  1.907513                  1.753884                  1.455823                  1.999829
     sub_category_x_phones     sub_category_x_storage    sub_category_x_supplies     sub_category_x_tables
                  1.977236                  2.370505                  1.682325                  1.301813
```

From the above generated output, we observe that none of the x variable have VIF > 4.

iii. Perform Backward elimination model

The method in backward elimination, we use p-value as metric. We perform backward elimination of non-significant features by p-value > 0.05 and recursively re-build model each time using a custom function.

```
# Function to rebuild model by removing variables having p-value >= 0.05
remove_non_sig_var <- function(model, df)
{
  all_x_variables <- names(model[[1]])[-1]  # names of all X variables
  # Get the summary of variables
  modelsummary <- summary(model)  # fetching summary of model
  pvalues <- modelsummary[[4]][, 4]  # getting all pvalues
  non_sig_x_var <- character()  # init variables that aren't statsitically significant
  non_sig_x_var <- names(which(pvalues >= 0.05)) # fetch records which are having p-value >= 0.05
  non_sig_x_var <- non_sig_x_var[!non_sig_x_var %in% "(Intercept)"]
  # If there are any non-significant variables,
  while(length(non_sig_x_var) > 0){
    all_x_variables <- all_x_variables[!all_x_variables %in% non_sig_x_var[1]]
    regformula <- as.formula(paste("profit ~ ", paste (all_x_variables, collapse=" + "), sep=""))  # new formula
    newmodel <- lm(regformula, data=df)  # re-build model with new formula
    # Get the non-significant vars from the rebuilt model to loop through again.
    newmodelsummary <- summary(newmodel)
    pvalues <- newmodelsummary[[4]][,4]
    non_sig_x_var <- character()
    non_sig_x_var <- names(which(pvalues >= 0.05))
    non_sig_x_var <- non_sig_x_var[!non_sig_x_var %in% "(Intercept)"]
  }
  return(newmodel)
}
```

```
> # Running backward elimination model by P-Value >= 0.05 and rebuild a model
> eliminationmodel = remove_non_sig_var(fullmodel, train.data)
> summary(eliminationmodel) # Adj-R2 = 0.3338

Call:
lm(formula = regformula, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-5691.2   -27.2    -6.6    35.6  5275.0 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               2.616e+01  1.353e+00  19.331  < 2e-16 ***
sales                     1.773e-01  1.643e-03 107.940  < 2e-16 ***
discount                 -2.148e+02  3.531e+00 -60.837  < 2e-16 ***
region_x_central         -4.352e+00  1.900e+00  -2.291  0.02200 *  
region_x_east             8.941e+00  3.276e+00   2.729  0.00636 ** 
region_x_emea             7.515e+00  2.592e+00   2.899  0.00374 ** 
region_x_south           -8.074e+00  2.297e+00  -3.515  0.00044 ***
sub_category_x_appliances -1.729e+01  4.158e+00  -4.159 3.21e-05 ***
sub_category_x_binders    1.129e+01  2.375e+00   4.753 2.01e-06 ***
sub_category_x_bookcases -3.843e+01  3.640e+00 -10.559  < 2e-16 ***
sub_category_x_chairs    -2.674e+01  3.066e+00  -8.724  < 2e-16 ***
sub_category_x_machines  -4.484e+01  4.516e+00  -9.929  < 2e-16 ***
sub_category_x_phones    -1.878e+01  3.103e+00  -6.054 1.43e-09 ***
sub_category_x_storage   -1.416e+01  2.569e+00  -5.513 3.56e-08 ***
sub_category_x_tables    -1.737e+02  5.979e+00 -29.050  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 139.7 on 35888 degrees of freedom
Multiple R-squared:  0.3341,    Adjusted R-squared:  0.3338 
F-statistic:  1286 on 14 and 35888 DF,  p-value: < 2.2e-16
```

Now, to validate the model is qualified or not we perform Goodness of fit.

1) **F-Test**

   NULL Hypothesis: $H_0 = 0$ i.e. No linear relationship. None of the predictors x variables having an association with dependent 'Y' variable.

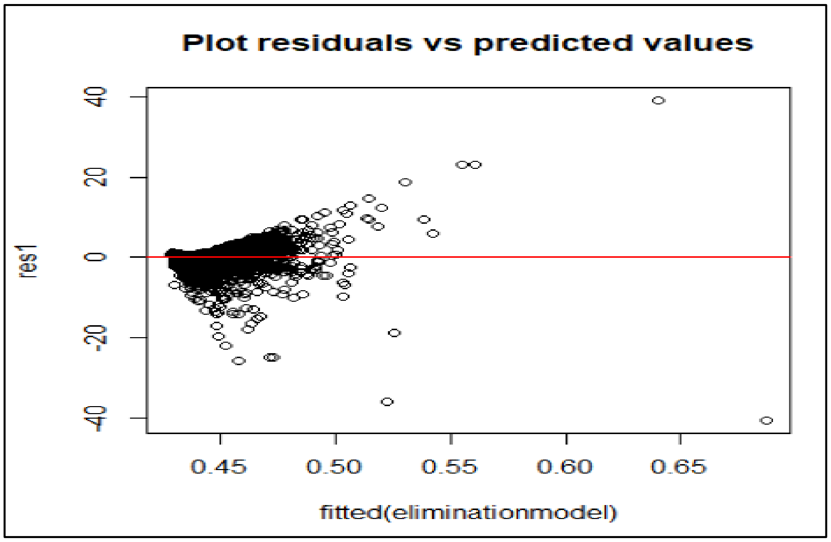
   Alternate Hypothesis: $H_a \neq 0$ i.e. At least one of the predictor variables has a significant linear relationship with dependent variable.

F-TEST INTERPRETATION

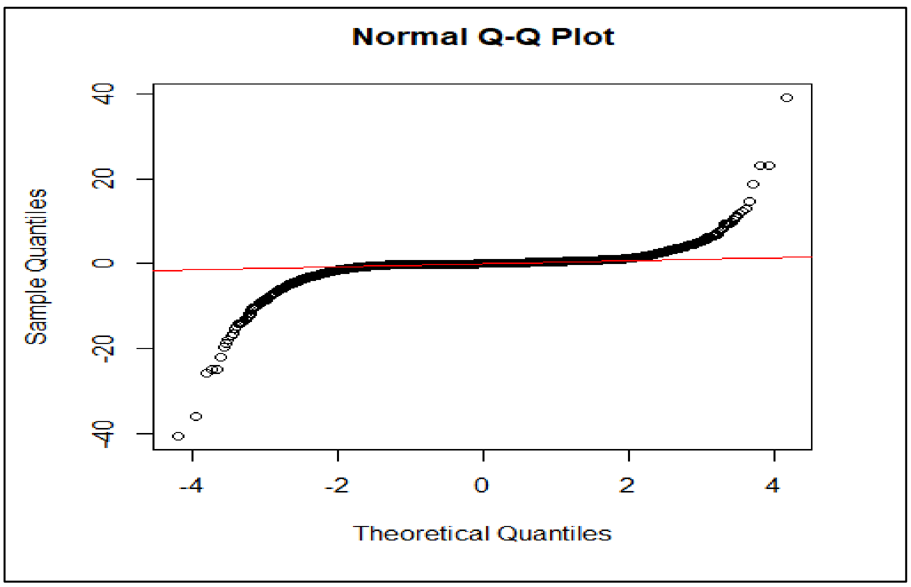
- As the P-value of the F-statistic is < 2.2e-16, which is highly significant. Hence, we reject Null Hypothesis.
- We can see that, at least, one of the predictor variables is significantly related to the outcome variable.

2) **Residual analysis**

- Constance Variance



- Normality Test



INTERPRETATION

- In the plot residuals vs predicted, we observe that spread is not constant from the plot.
- From the Q-Q plot, we observe that the points are not around the line and are not normal.

Finally, we can conclude that there is no linearity and normality from the plots. It requires transformation or other methods to improve the model.

Transformation

The replacement of a variable by a function of that variable is called transformation in data analytics. The transformation can be performed by log, square root and inverse.

- Logarithm transformation on Y variable

```
> logformula1 <- as.formula(paste("log(profit) ~ ", paste (eliminationmodel_var, collapse
=" + "), sep=""))  # new formula
> logeliminationmodel3 <- lm(logformula1, data=train.data)
> summary(logeliminationmodel3)

Call:
lm(formula = logformula1, data = train.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.02902 -0.00455 -0.00132  0.00537  0.42964 

Coefficients:
                           Estimate Std. Error   t value Pr(>|t|)    
(Intercept)              -0.8161555  0.0002093 -3899.760  < 2e-16 ***
sales                     0.5010399  0.0057512    87.119  < 2e-16 ***
discount                 -0.0290954  0.0004642   -62.685  < 2e-16 ***
region_x_central         -0.0007909  0.0002938    -2.692 0.007113 ** 
region_x_east             0.0010203  0.0005067     2.014 0.044043 *  
region_x_emea             0.0008544  0.0004008     2.131 0.033061 *  
region_x_south           -0.0014470  0.0003552    -4.074 4.64e-05 ***
sub_category_x_appliances -0.0016051  0.0006430    -2.496 0.012550 *  
sub_category_x_binders    0.0013902  0.0003673     3.784 0.000154 ***
sub_category_x_bookcases -0.0041554  0.0005629    -7.382 1.59e-13 ***
sub_category_x_chairs    -0.0028711  0.0004741    -6.056 1.41e-09 ***
sub_category_x_machines  -0.0061318  0.0006984    -8.780  < 2e-16 ***
sub_category_x_phones    -0.0017025  0.0004798    -3.548 0.000388 ***
sub_category_x_storage   -0.0017021  0.0003973    -4.284 1.84e-05 ***
sub_category_x_tables    -0.0241630  0.0009246   -26.134  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02161 on 35888 degrees of freedom
Multiple R-squared:  0.2773,	Adjusted R-squared:  0.277 
F-statistic: 983.5 on 14 and 35888 DF,  p-value: < 2.2e-16
```

- Square transformation on Y variable

```
> sqrtformula1 <- as.formula(paste("sqrt(profit) ~ ", paste (eliminationmodel_var, collap
e=" + "), sep=""))  # new formula
> sqrteliminationmodel3 <- lm(sqrtformula1, data=train.data)
> summary(sqrteliminationmodel3)

Call:
lm(formula = sqrtformula1, data = train.data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.292487 -0.001436 -0.000388  0.001781  0.193816 

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)    
(Intercept)               6.648e-01  6.808e-05 9764.896  < 2e-16 ***
sales                     1.832e-01  1.871e-03   97.939  < 2e-16 ***
discount                 -9.409e-03  1.510e-04  -62.315  < 2e-16 ***
region_x_central         -2.419e-04  9.558e-05   -2.531 0.011390 *  
region_x_east             3.910e-04  1.648e-04    2.373 0.017670 *  
region_x_emea             3.272e-04  1.304e-04    2.509 0.012110 *  
region_x_south           -4.418e-04  1.156e-04   -3.823 0.000132 ***
sub_category_x_appliances -6.855e-04  2.092e-04   -3.277 0.001049 ** 
sub_category_x_binders    5.127e-04  1.195e-04    4.290 1.79e-05 ***
sub_category_x_bookcases -1.640e-03  1.831e-04   -8.955  < 2e-16 ***
sub_category_x_chairs    -1.138e-03  1.542e-04   -7.377 1.66e-13 ***
sub_category_x_machines  -2.120e-03  2.272e-04   -9.333  < 2e-16 ***
sub_category_x_phones    -7.423e-04  1.561e-04   -4.756 1.98e-06 ***
sub_category_x_storage   -6.346e-04  1.292e-04   -4.910 9.15e-07 ***
sub_category_x_tables    -8.346e-03  3.008e-04  -27.751  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00703 on 35888 degrees of freedom
Multiple R-squared:  0.3078,	Adjusted R-squared:  0.3076 
F-statistic:  1140 on 14 and 35888 DF,  p-value: < 2.2e-16
```

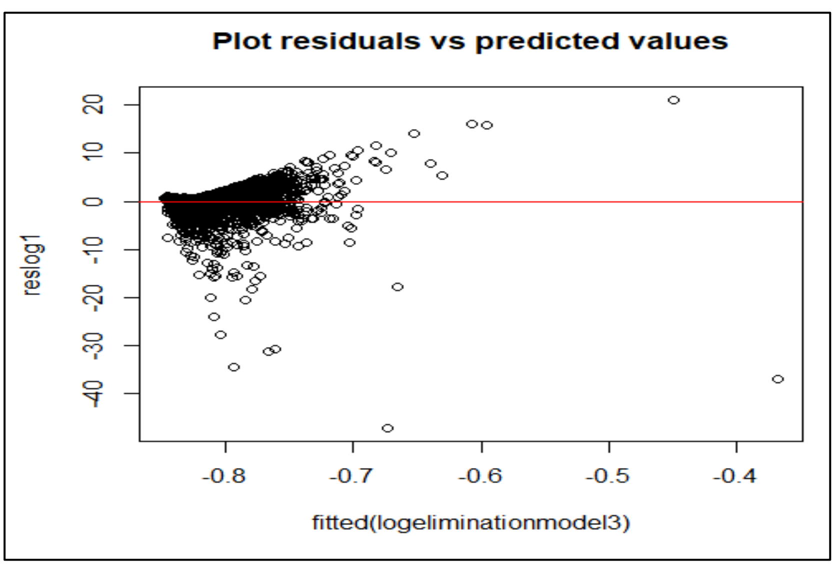
- Inverse transformation on Y variable

```
> invformula1 <- as.formula(paste("(1/profit) ~ ", paste (eliminationmodel
_var, collapse=" + "), sep=""))  # new formula
> inveliminationmodel3 <- lm(invformula1, data=train.data)
> summary(inveliminationmodel3)

Call:
lm(formula = invformula1, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5593 -0.0126  0.0036  0.0114  3.4406 

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                2.2598800  0.0005225 4324.775  < 2e-16 ***
sales                     -0.9176775  0.0143597  -63.906  < 2e-16 ***
discount                   0.0698350  0.0011589   60.260  < 2e-16 ***
region_x_central           0.0020490  0.0007336    2.793 0.005224 ** 
region_x_east             -0.0016728  0.0012651   -1.322 0.186072    
region_x_emea             -0.0014331  0.0010008   -1.432 0.152173    
region_x_south             0.0039051  0.0008869    4.403 1.07e-05 ***
sub_category_x_appliances  0.0018799  0.0016054    1.171 0.241609    
sub_category_x_binders    -0.0024726  0.0009172   -2.696 0.007024 ** 
sub_category_x_bookcases   0.0061312  0.0014055    4.362 1.29e-05 ***
sub_category_x_chairs      0.0041811  0.0011837    3.532 0.000412 ***
sub_category_x_machines    0.0136401  0.0017437    7.822 5.32e-15 ***
sub_category_x_phones      0.0016687  0.0011980    1.393 0.163673    
sub_category_x_storage     0.0029774  0.0009920    3.001 0.002689 ** 
sub_category_x_tables      0.0508118  0.0023085   22.011  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05396 on 35888 degrees of freedom
Multiple R-squared:  0.2071,	Adjusted R-squared:  0.2068 
F-statistic: 669.4 on 14 and 35888 DF,  p-value: < 2.2e-16
```

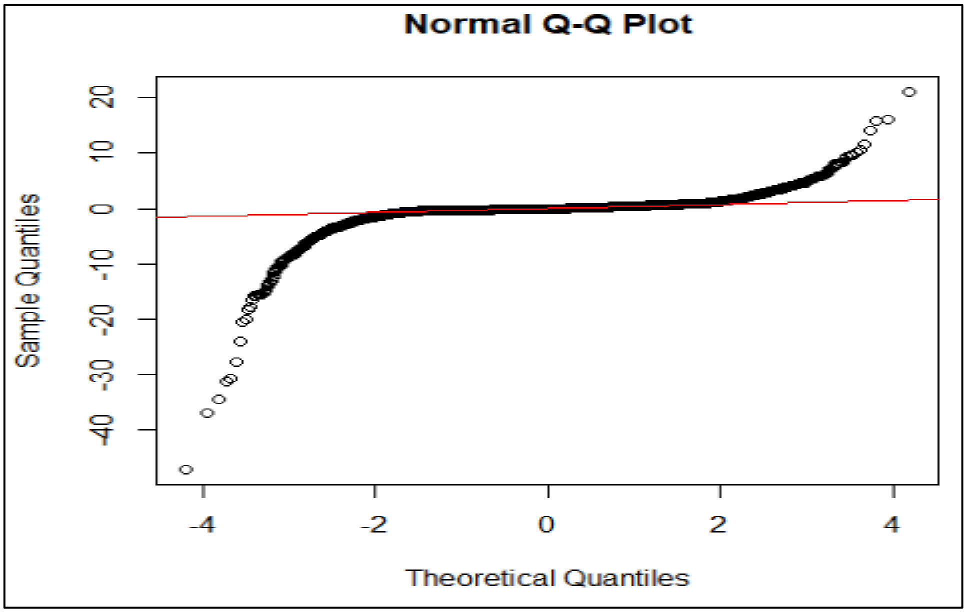
Residual plot for all the transformation

Residual analysis on log transformed model

- Constance variance



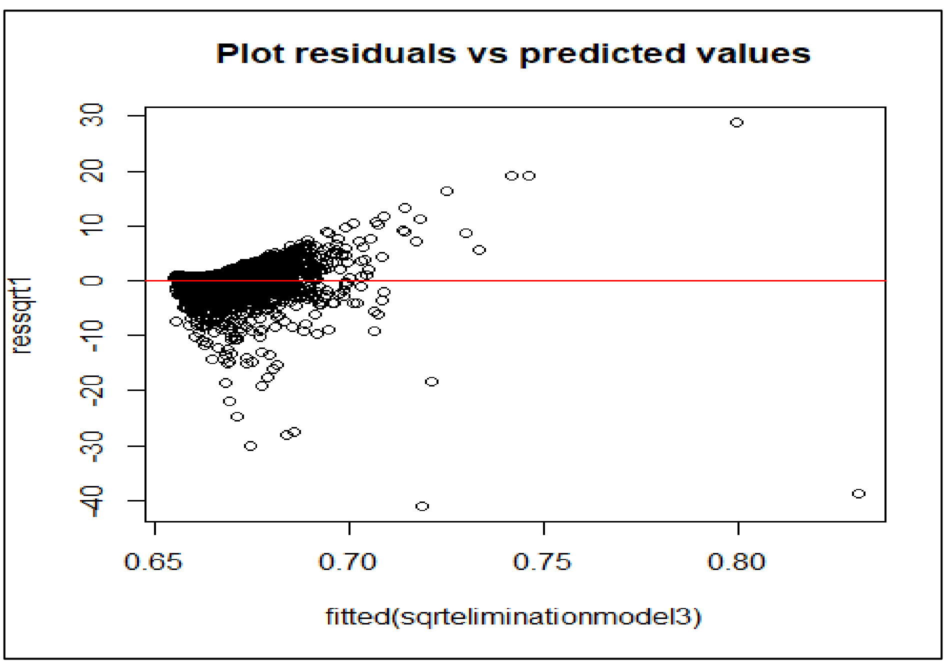
- Normality test



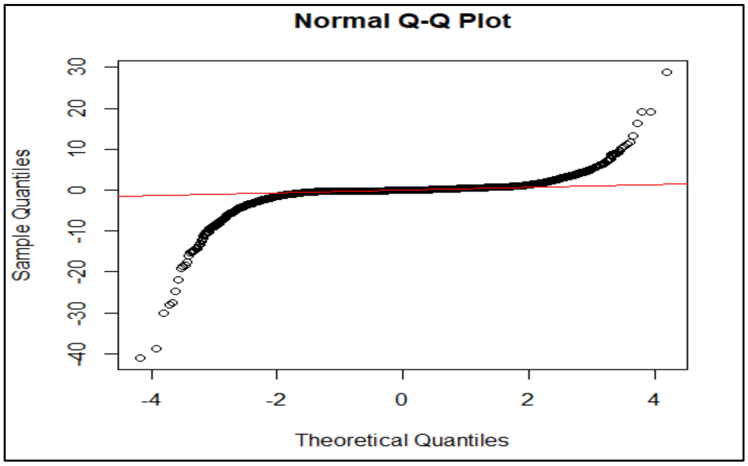
We observe that with log transformed model there are no linearity and normality from the plots.

Residual analysis on sqrt transformed model

- Constance variance



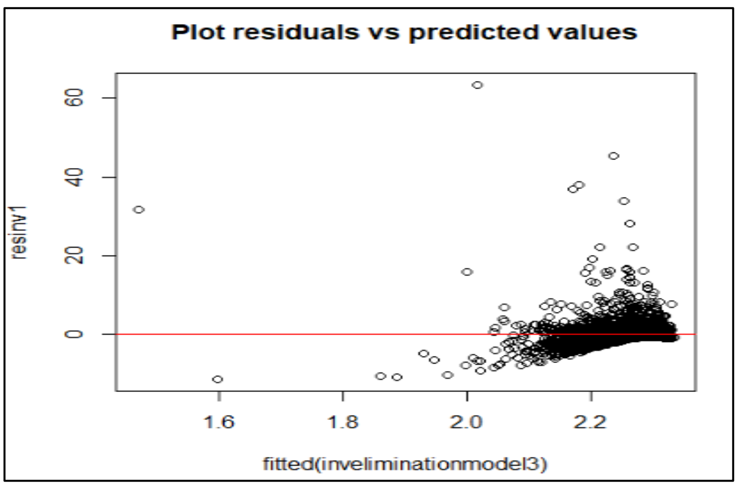
- Normality test



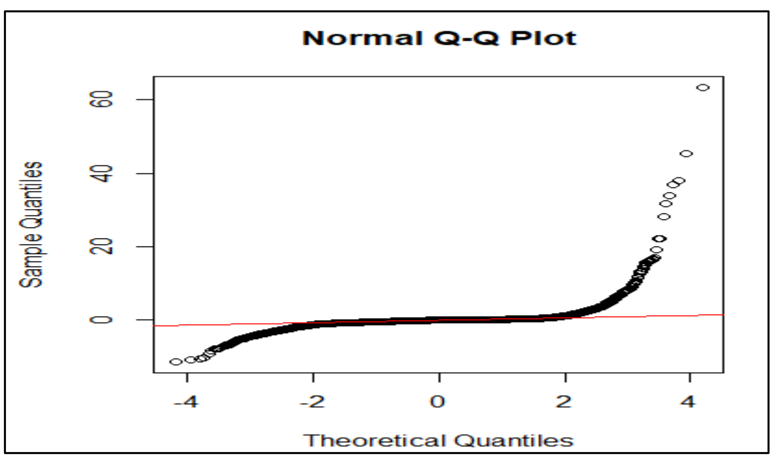
We observe that with sqrt transformed model there are no linearity and normality from the plots.

Residual analysis on Inverse transformed model

- Constance variance



- Normality test

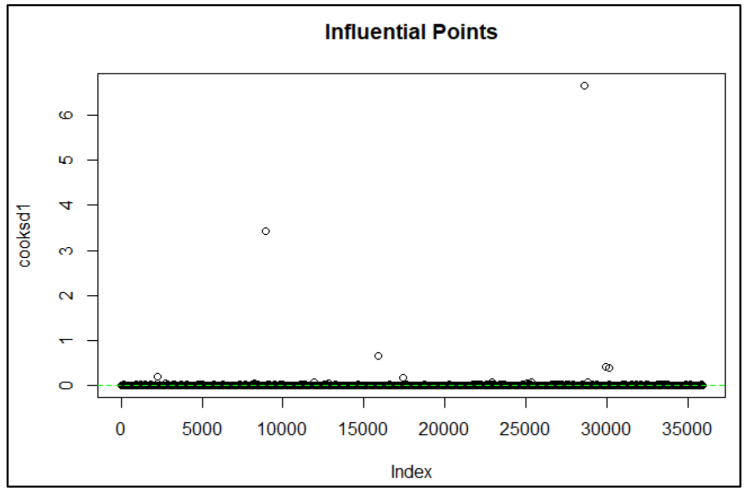


We observe that with inverse transformed model there are no linearity and normality from the plots. Finally, we can conclude that from all the residual plot of transformation on Y has not shown any improvement in linearity and normality.

Check Influential points

We use cooks.distance() to identify the influential points. We check for influential points in elimination model.

```
> # Check for influential points in elimination model
> cooksd1 = cooks.distance(eliminationmodel)
> n = nrow(train.data)
> plot(cooksd1, main="Influential Points")
> abline(h = 4/n, lty=2, col="green")
> influential_points1 = as.numeric(names(cooksd1[cooksd1 > (4/n)]))
> #influential_points1
> newtrain.data12 <- train.data[-influential_points1,]
> nrow(newtrain.data12)
[1] 34394
> # Rebuilding model after removing influential points
> eliminationmodel_var <- names(eliminationmodel[[1]])[-1]
```



In order to improvise regression model, we remove influential points that we found in the output generated.

```
> eliminationmodel3 <- lm(regformula1, data=newtrain.data12)
> summary(eliminationmodel3) # Adj-R2 = 0.557

Call:
lm(formula = regformula1, data = newtrain.data12)

Residuals:
    Min      1Q  Median      3Q     Max 
-437.51  -16.33   -2.85   20.62  361.78 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.288e+01  5.364e-01  24.014  < 2e-16 ***
sales                    1.770e-01  1.087e-03 162.920  < 2e-16 ***
discount                -1.195e+02  1.389e+00 -86.066  < 2e-16 ***
region_x_central        -8.448e-01  7.343e-01  -1.151  0.24991    
region_x_east            6.812e+00  1.272e+00   5.356 8.56e-08 ***
region_x_emea            2.983e+00  9.973e-01   2.991  0.00278 ** 
region_x_south          -1.265e+00  8.915e-01  -1.419  0.15602    
sub_category_x_appliances -1.586e+00  1.694e+00  -0.936  0.34917    
sub_category_x_binders   7.747e+00  9.049e-01   8.561  < 2e-16 ***
sub_category_x_bookcases -2.388e+01  1.484e+00 -16.097  < 2e-16 ***
sub_category_x_chairs   -1.723e+01  1.205e+00 -14.292  < 2e-16 ***
sub_category_x_machines -1.529e+01  1.842e+00  -8.301  < 2e-16 ***
sub_category_x_phones   -1.068e+01  1.224e+00  -8.721  < 2e-16 ***
sub_category_x_storage  -1.089e+01  9.830e-01 -11.079  < 2e-16 ***
sub_category_x_tables   -1.639e+02  3.508e+00 -46.722  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.88 on 34379 degrees of freedom
Multiple R-squared:  0.5572,	Adjusted R-squared:  0.557 
F-statistic:  3090 on 14 and 34379 DF,  p-value: < 2.2e-16
```

Rebuilding the elimination model

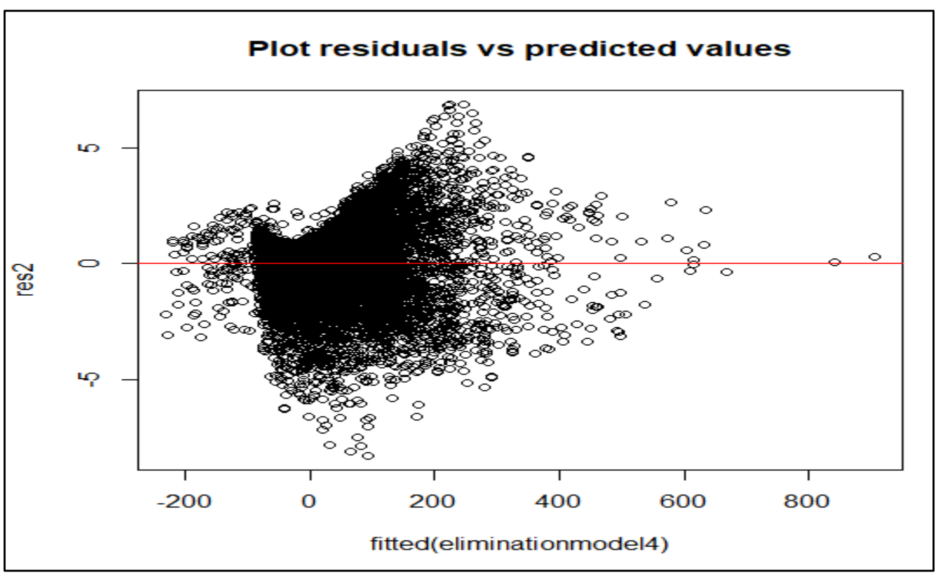
We rebuild the model using custom function, as there are still non-significant ‘x’ variables.

```
> # As there are still non significant variables rebuilding model by removing them
> eliminationmodel4 = remove_non_sig_var(eliminationmodel3, newtrain.data12)
> summary(eliminationmodel4) # Adj-R2 = 0.557

Call:
lm(formula = regformula, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-436.92  -16.28   -2.83   20.59  362.43 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               1.242e+01  4.673e-01  26.571  < 2e-16 ***
sales                     1.769e-01  1.072e-03 164.991  < 2e-16 ***
discount                 -1.196e+02  1.388e+00 -86.171  < 2e-16 ***
region_x_east             7.201e+00  1.245e+00   5.786 7.26e-09 ***
region_x_emea             3.391e+00  9.624e-01   3.523 0.000427 ***
sub_category_x_binders    7.811e+00  9.013e-01   8.666  < 2e-16 ***
sub_category_x_bookcases -2.373e+01  1.476e+00 -16.080  < 2e-16 ***
sub_category_x_chairs    -1.708e+01  1.198e+00 -14.252  < 2e-16 ***
sub_category_x_machines  -1.512e+01  1.837e+00  -8.231  < 2e-16 ***
sub_category_x_phones    -1.056e+01  1.217e+00  -8.678  < 2e-16 ***
sub_category_x_storage   -1.081e+01  9.772e-01 -11.066  < 2e-16 ***
sub_category_x_tables    -1.637e+02  3.503e+00 -46.736  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.88 on 34382 degrees of freedom
Multiple R-squared:  0.5571,	Adjusted R-squared:  0.557 
F-statistic:  3932 on 11 and 34382 DF,  p-value: < 2.2e-16
```

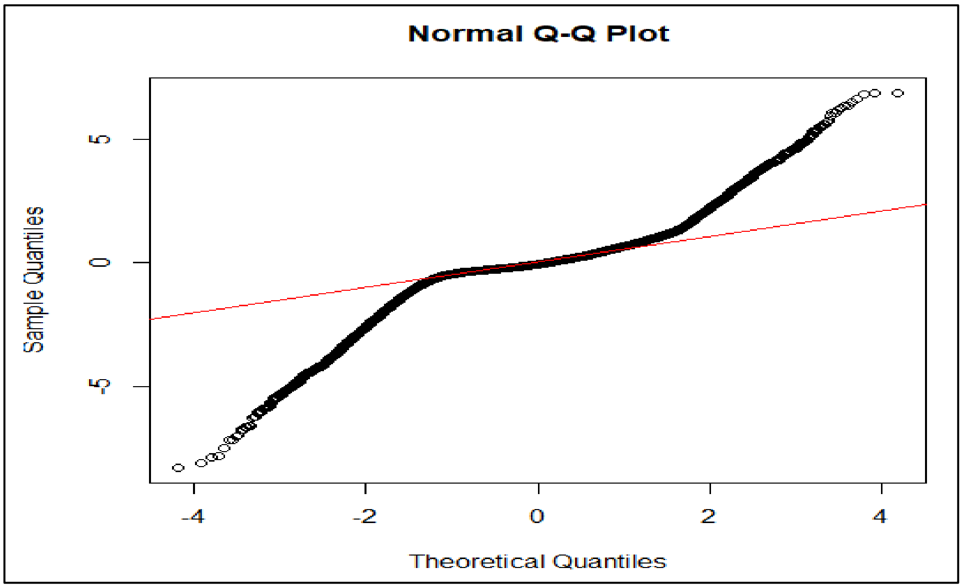
From the output generated we observe that, the rebuilt model shows all x variables are statistically significant to predict profit with p-value < 0.05 and we reject Null hypothesis.

Residual analysis

- Constance variance



- Normality test



INTERPRETATION

- In the plot residuals vs predicted, we observe that the points are scattered.
- From the Q-Q plot, we observe that there is slight normality.
- From Elimination model, with many predictor variables, the adjusted R^2 = 0.557, which means that "55.7% of the variance in the measure of profit can be predicted by significant x variables."

iv. Feature selection

Stepwise selection:

We build linear regression model using Stepwise feature selection with both feature forward and backward model with direction as "both".

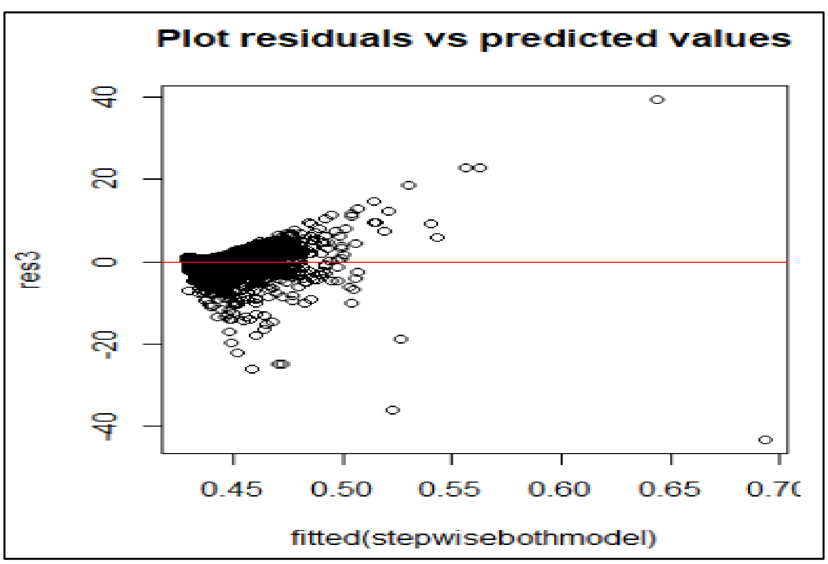
```
> fullmdl <- lm(profit~., data=train1.data)
> stepwisebothmodel = step(fullmdl, direction="both", trace=F)
> summary(stepwisebothmodel) # Adj-R2 0.3339
```

```
lm(formula = profit ~ sales + discount + region_x_central + region_x_east +
    region_x_emea + region_x_north + region_x_oceania + region_x_south +
    sub_category_x_appliances + sub_category_x_binders + sub_category_x_bookcases +
    sub_category_x_chairs + sub_category_x_machines + sub_category_x_phones +
    sub_category_x_storage + sub_category_x_supplies + sub_category_x_tables,
    data = train1.data)

Residuals:
    Min      1Q  Median      3Q     Max
-5692.0   -27.5    -6.2    35.8  5273.8

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.824e+01  1.573e+00  17.959  < 2e-16 ***
sales                     1.774e-01  1.644e-03 107.872  < 2e-16 ***
discount                 -2.149e+02  3.538e+00 -60.750  < 2e-16 ***
region_x_central         -5.899e+00  2.031e+00  -2.905  0.00368 **
region_x_east             7.273e+00  3.355e+00   2.168  0.03016 *
region_x_emea             5.998e+00  2.687e+00   2.232  0.02560 *
region_x_north           -4.151e+00  2.726e+00  -1.523  0.12784
region_x_oceania         -5.373e+00  3.136e+00  -1.714  0.08660 .
region_x_south           -9.598e+00  2.406e+00  -3.989 6.64e-05 ***
sub_category_x_appliances -1.787e+01  4.168e+00  -4.286 1.82e-05 ***
sub_category_x_binders    1.070e+01  2.396e+00   4.466 7.99e-06 ***
sub_category_x_bookcases -3.895e+01  3.653e+00 -10.663  < 2e-16 ***
sub_category_x_chairs    -2.722e+01  3.081e+00  -8.837  < 2e-16 ***
sub_category_x_machines  -4.545e+01  4.526e+00 -10.042  < 2e-16 ***
sub_category_x_phones    -1.938e+01  3.117e+00  -6.217 5.12e-10 ***
sub_category_x_storage   -1.474e+01  2.588e+00  -5.693 1.26e-08 ***
sub_category_x_supplies  -6.115e+00  3.621e+00  -1.689  0.09124 .
sub_category_x_tables    -1.744e+02  5.986e+00 -29.129  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139.7 on 35885 degrees of freedom
Multiple R-squared:  0.3342,    Adjusted R-squared:  0.3339
F-statistic:  1060 on 17 and 35885 DF,  p-value: < 2.2e-16
```

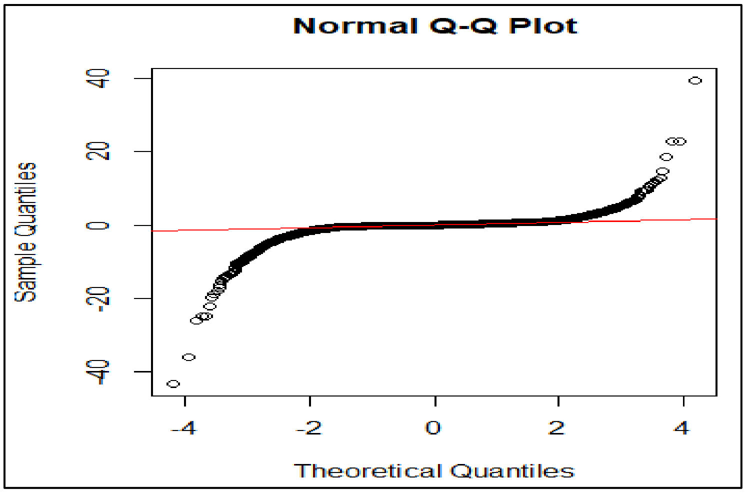
Cross checking for multicollinearity, as full model built by removing collinearity variable there is no issue in feature selection model.

Residual analysis on Stepwise model

- Constance variance



- Normality test



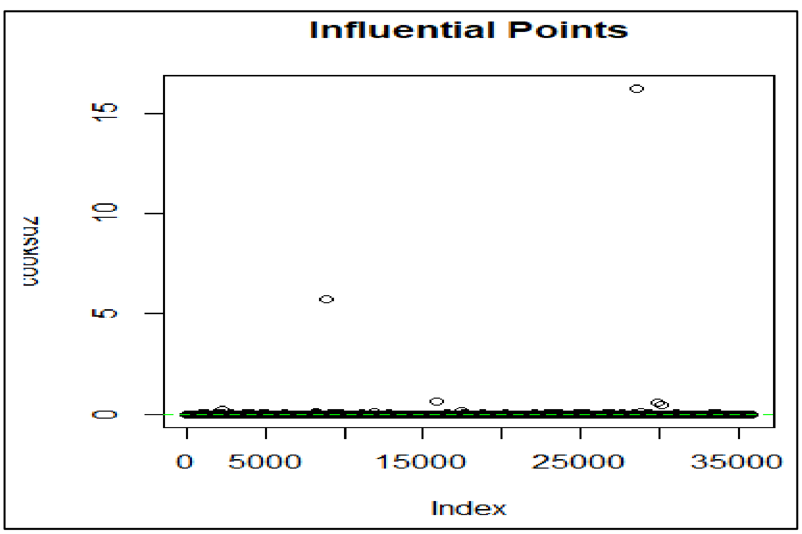
INTERPRETATION

- From the plots we observe that, there is no linearity and normality.

Check Influential points

We use cooks.distance() to identify the influential points. We check for influential points in stepwise model.

```
> # Check for influential points in stepwise both model
> cooksd2 = cooks.distance(stepwisebothmodel)
> n = nrow(train1.data)
> plot(cooksd2, main="Influential Points")
> abline(h = 4/n, lty=2, col="green")
> influential_points2 = as.numeric(names(cooksd2[cooksd2 > (4/n)]))
> newtrain.data2 <- train1.data[-influential_points2,]
> nrow(newtrain.data2)
[1] 34478
```

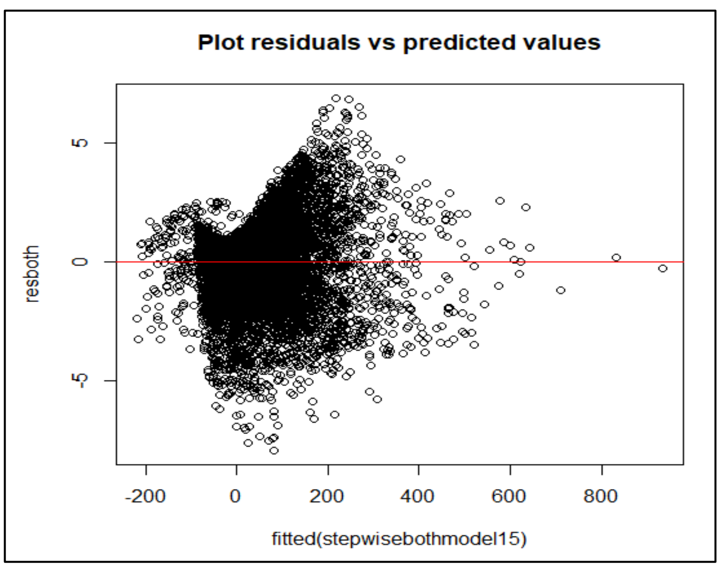


In the Influential points, we could observe that there are potential outliers in plot with cook's distance method and model to be rebuild.

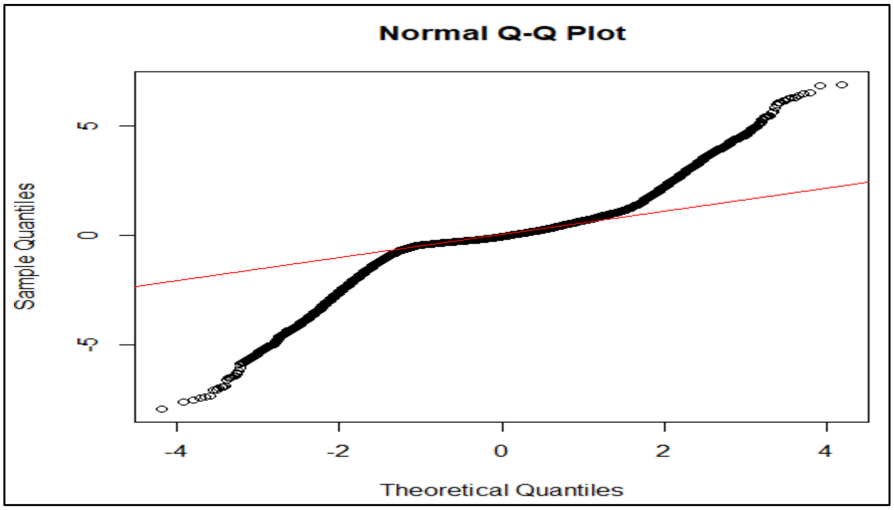
```
> summary(stepwisebothmodel15) # Adj-R2 = 0.5548

Call:
lm(formula = regformula, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-423.16  -17.03   -2.78   21.17  367.91

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.397e+01  6.881e-01  20.303  < 2e-16 ***
sales                     1.745e-01  1.734e-03 100.614  < 2e-16 ***
discount                 -1.198e+02  1.426e+00 -83.999  < 2e-16 ***
shipping_cost             5.129e-02  1.181e-02   4.342 1.42e-05 ***
region_x_central         -3.430e+00  7.776e-01  -4.411 1.03e-05 ***
region_x_east             4.007e+00  1.309e+00   3.060 0.002212 **
region_x_north           -3.956e+00  1.055e+00  -3.749 0.000178 ***
region_x_oceania         -6.487e+00  1.231e+00  -5.272 1.36e-07 ***
region_x_south           -3.579e+00  9.317e-01  -3.841 0.000123 ***
region_x_southeast_asia  -5.991e+00  1.297e+00  -4.618 3.89e-06 ***
sub_category_x_binders    8.958e+00  9.815e-01   9.127  < 2e-16 ***
sub_category_x_bookcases -2.451e+01  1.533e+00 -15.992  < 2e-16 ***
sub_category_x_chairs    -1.679e+01  1.266e+00 -13.262  < 2e-16 ***
sub_category_x_copiers   -9.628e+00  1.583e+00  -6.084 1.19e-09 ***
sub_category_x_envelopes  3.068e+00  1.409e+00   2.178 0.029402 *
sub_category_x_fasteners  3.670e+00  1.413e+00   2.597 0.009402 **
sub_category_x_labels     3.399e+00  1.373e+00   2.476 0.013304 *
sub_category_x_machines  -1.560e+01  1.882e+00  -8.288  < 2e-16 ***
sub_category_x_paper      5.217e+00  1.206e+00   4.328 1.51e-05 ***
sub_category_x_phones    -1.096e+01  1.284e+00  -8.540  < 2e-16 ***
sub_category_x_storage   -9.812e+00  1.052e+00  -9.324  < 2e-16 ***
sub_category_x_tables    -1.542e+02  3.366e+00 -45.814  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.62 on 34456 degrees of freedom
Multiple R-squared:  0.5551,    Adjusted R-squared:  0.5548
F-statistic:  2047 on 21 and 34456 DF,  p-value: < 2.2e-16
```

Residual analysis

- Constance variance



- Normality test



INTERPRETATION

- From the plots, we observe that the points are scattered and there is slight normality.

v. **Finding accuracy**

```
> ####################################################
> # RMSE Calculation on train.data
> ####################################################
> y1 = predict.glm(eliminationmodel4, train.data)
> y2 = predict.glm(stepwisebothmodel15, train.data)
> y = train.data$profit
> rmse_1 = sqrt(((y-y1)%*%(y-y1)) /nrow(train.data))
> rmse_1
         [,1]
[1,] 141.4796
> rmse_2 = sqrt(((y-y2)%*%(y-y2)) /nrow(train.data))
> rmse_2
         [,1]
[1,] 141.4361
```

```
> ##################################################
> # RMSE Calculation on test.data
> ##################################################
> x1 = predict.glm(eliminationmodel4, test.data)
> x2 = predict.glm(stepwisebothmodel15, test.data)
> x = test.data$profit
> rmse_1 = sqrt(((x-x1)%*%(x-x1)) /nrow(test.data))
> rmse_1
        [,1]
[1,] 149.493
> rmse_2 = sqrt(((x-x2)%*%(x-x2)) /nrow(test.data))
> rmse_2
         [,1]
[1,] 149.6765
```

Regression Model	ADJ-R2	ROOT MEAN SQUARE ERROR	
		Train	Test
Elimination full model	0.557	141.4796	149.493
Stepwise Both Model	0.5548	141.4361	149.6765

INTERPRETATION

- The Elimination model, with many predictor variables, the adjusted R2 = 0.557, meaning that "55.7% of the variance in the measure of profit can be predicted by statistically significant x variables.
- The stepwise both model, with many predictor variables, the adjusted R2 = 0.5548, meaning that "55.48% of the variance in the measure of profit can be predicted by statistically significant x variables.
- From RMSE calculation, Model with low RMSE is the best fit model, here elimination model has less RMSE and high R square with test data compared to stepwise model.

So, the best reduced fit model is,

```
Y= profit =1.242e+01+(sales*1.769e-01)+(discount*-
1.196e+02)+(region_x_east*7.201e+00)+(region_x_emea*3.391e+00)+(
sub_category_x_binders*7.811e+00)+(sub_category_x_bookcases*-2.373e+01)+(
sub_category_x_chairs*-1.708e+01)+(sub_category_x_machines*-1.512e+01)+(
sub_category_x_phones*-1.056e+01)+(sub_category_x_storage*-1.081e+01)+(
sub_category_x_tables*-1.637e+02)
```

vi. Perform Regularization

The technique used to reduce the error by fitting a function appropriately on the given training set and to avoid overfitting.

➢ Creating numeric matrix for the training features and a vector of target values.

```
> ##########################################################
> # Regularization
> ##########################################################
> library(caret)
> dummies <- dummyVars(profit~., data = train.data)
> train_dummies = predict(dummies, newdata = train.data)
> test_dummies = predict(dummies, newdata = test.data)
> print(dim(train_dummies)); print(dim(test_dummies))
[1] 35903    36
[1] 15387    36
```

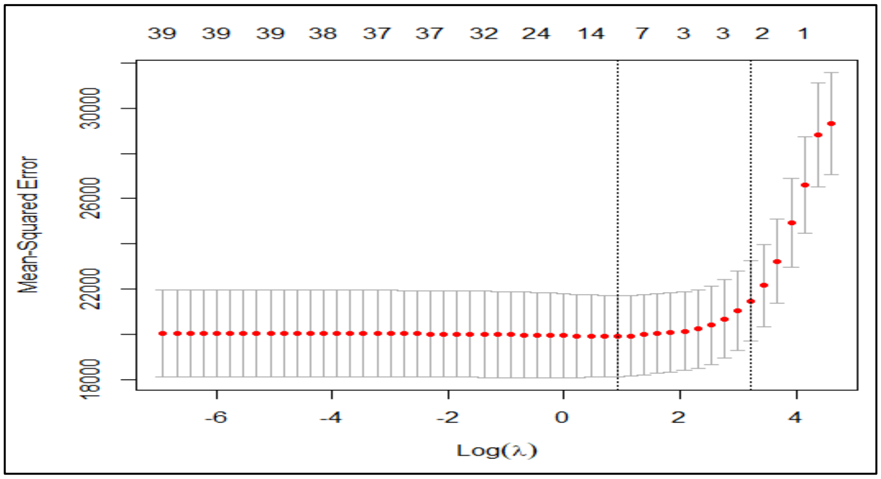
- Create a custom function to compute R-Square from true and predicted values

```
> # Custom function to compute R-square from true and predicted values
> eval_results <- function(true, predicted, df) {
+   SSE <- sum((predicted - true)^2)
+   SST <- sum((true - mean(true))^2)
+   # Calculate R-square value
+   R_square <- 1 - SSE / SST
+   # Calculate RMSE
+   RMSE = sqrt(SSE/nrow(df))
+   # Model performance metrics RMSE and R_square
+   data.frame(
+     RMSE = RMSE,
+     Rsquare = R_square
+   )
+ }
```

Regularization techniques

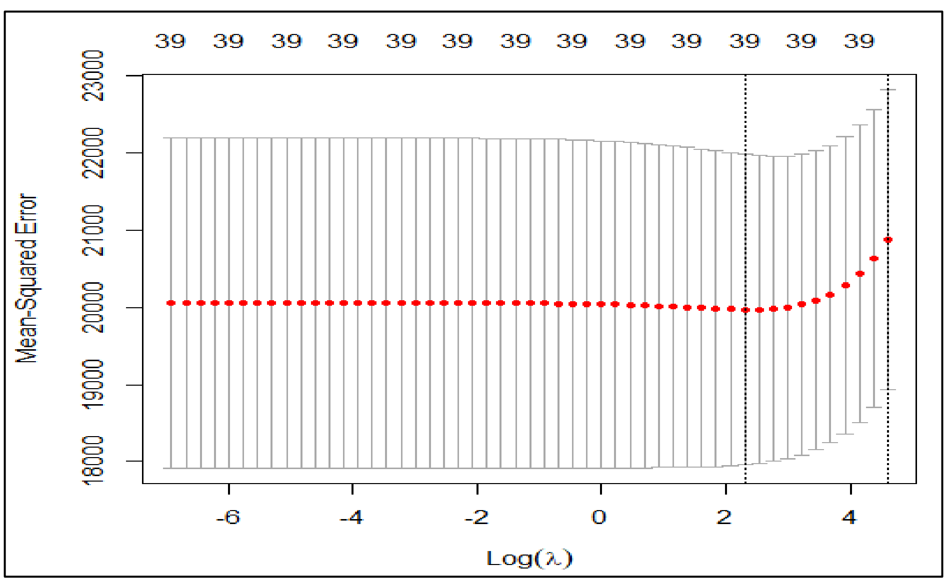
a. Lasso Regression model –
   Lasso is considered as a feature selection process to make use of the most influential features.

```
> ###############################################################################
> # Lasso Regression
> ###############################################################################
> grid <- 10^seq(2, -3, by = -.1)
> #lambdas <-10^seq(10, -2, length = 100)
> # Setting alpha = 1 implies lasso penalty
> lasso_reg <- cv.glmnet(x_train, y_train, alpha=1, lambda=grid, standardize=TRUE, nfolds=10)
> plot(lasso_reg)
> lambda_best <- lasso_reg$lambda.min
> lambda_best
[1] 3.981072
> lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = lambda_best, standardize = TRUE)
> lasso.coef = predict(lasso_model, s=lambda_best, type="coefficients")[1:37, ]
> lasso.coef[lasso.coef !=0]
          (Intercept)                  sales              discount   sub_category_x_binders
           20.4753115              0.1600284          -202.2600545                4.6928964
sub_category_x_bookcases   sub_category_x_chairs  sub_category_x_machines    sub_category_x_tables
           -7.5442475             -1.5223789            -8.8577027             -127.1907043
>
```

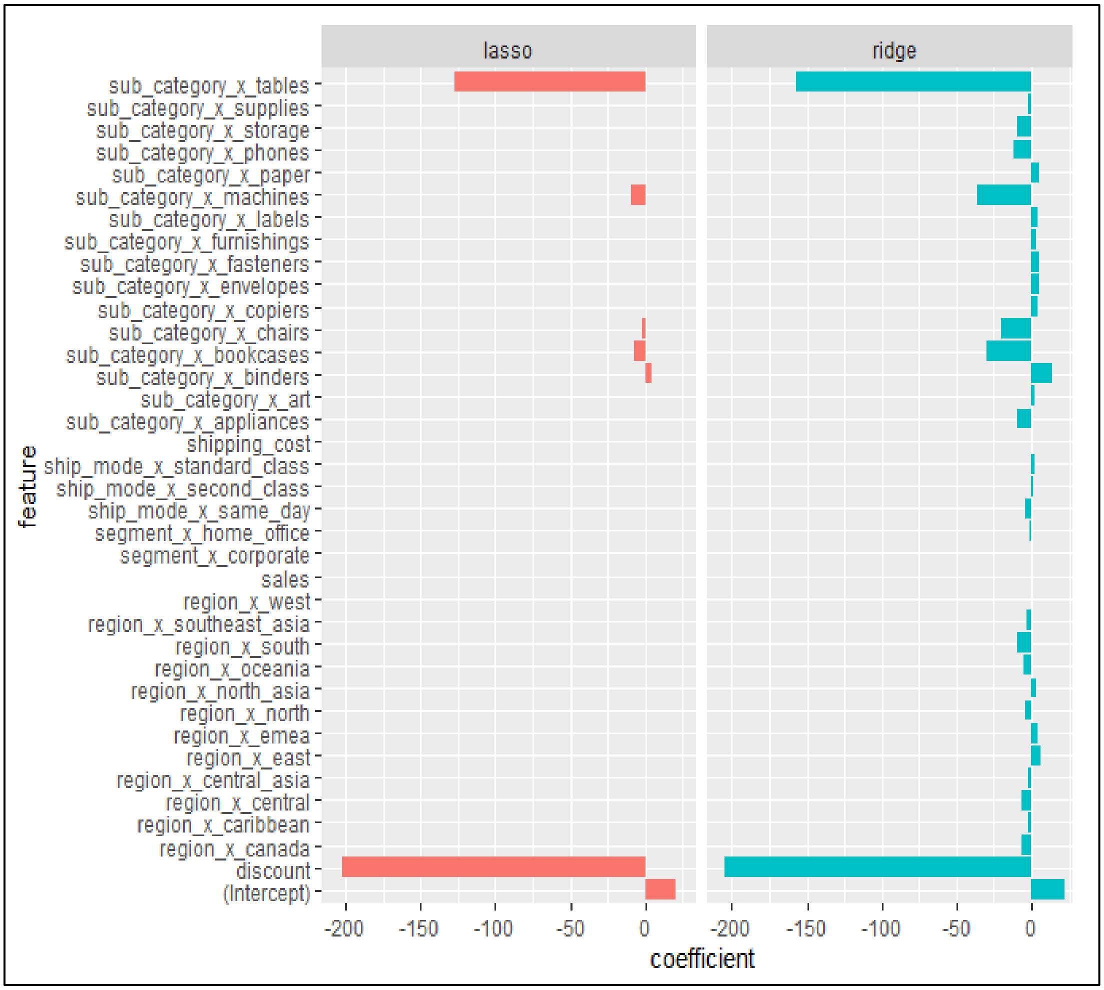


b. Ridge Regression model –
In ridge regression modification is done by adding a penalty parameter that is equivalent to the square of the magnitude of the co-eff.

```
> ###############################################################################
> # Ridge Regression
> ###############################################################################
> grid <- 10^seq(2, -3, by = -.1)
> # The alpha=0 implies Ridge penalty
> ridge_reg = glmnet(x_train, y_train, nlambda = 100, alpha = 0, family = 'gaussian', lambda = grid)
> cv_ridge <- cv.glmnet(x_train, y_train, alpha = 0, lambda = grid)
> plot(cv_ridge)
> optimal_lambda <- cv_ridge$lambda.min
> optimal_lambda
[1] 10
```



We plot for Lasso and Ridge regression model.



From the Lasso and Ridge plot, we could observe that the lasso model shrinkages co-efficients to zero, wherein the ridge model just reduced the co-efficient values but retained all. we observe that coefficients of discount are much higher influence on profit as compared to rest of the coefficient.

Prediction and evaluation from Ridge and lasso regression

The prediction and evaluation using best lambda on train and test data from Ridge and Lasso.

➢ <u>Ridge regression</u>

```
> # Prediction and evaluation on train data
> predictions_train <- predict(ridge_reg, s=optimal_lambda, newx=x_train)
> eval_results(y_train, predictions_train, train.data)
      RMSE  Rsquare
1 139.8677 0.332552
> # Prediction and evaluation on test data
> predictions_test <- predict(ridge_reg, s=optimal_lambda, newx=x_test)
> eval_results(y_test, predictions_test, test.data)
      RMSE   Rsquare
1 149.0366 0.3253518
```

➢ <u>Lasso regression</u>

```
> # Prediction and evaluation on train data
> predictions_train <- predict(lasso_model, s = lambda_best, newx = x_train)
> eval_results(y_train, predictions_train, train.data)
      RMSE   Rsquare
1 140.0986 0.3303457
> #Prediction and evaluation on test data
> predictions_test <- predict(lasso_model, s = lambda_best, newx = x_test)
> eval_results(y_test, predictions_test, test.data)
      RMSE   Rsquare
1 148.5376 0.3298613
```

5.2. Evaluations and Results

- We compare the Ridge and Lasso regression with multiple linear regression model by the RMSE and R-square values.

Regression Model	R-Square		Root Mean Square Error	
	Train	Test	Train	Test
Multiple linear Regression Model	0.557		141.4796	149.493
Lasso Regression Model	0.3303457	0.3298613	140.0986	148.5376
Ridge Regression Model	0.332552	0.3253518	139.8677	149.0366

5.3. Findings

- The Lasso model built by shrinking many features to zero with RMSE as 148.5376 on test data.
- The Ridge model on test data gives RMSE as 149.0366, which is almost less than multiple linear regression model.
- There is no overfitting issue.
- Finally, we can conclude that Lasso model is the best model with less RMSE.

6. Conclusions and Future Work

6.1. Conclusions

- The global superstore has statistically significant difference in sales with respect to different groups of market regions. The APAC Market group has larger sales and Africa has least number of sales.
- The global superstore has statistically significant difference in sales with respect to different category groups. The technology category has higher sales and office-supplies category has least sales.
- The multiple linear regression model is statistically significant in predicting profit of global super store with respect to sales, discount, region, and product subcategory etc.
- The Lasso regression model is the best model with least RMSE out of multiple linear regression model and Ridge model.

6.2. Limitations

The limitations of our project are –

- Our global superstore dataset has data from the year 2011 - 2015.

6.3. Potential Improvements or Future Work

- The global superstore data has insights with respect to city, state, the model can be enhanced by considering these features also to predict profit in more micro level.
- The analysis with respect to months and days would give profit prediction in the season wise.
- In our analysis, we have our data only from 2011 – 2015. Hence, collecting data and analyzing it for more years can be another learning.