Global Superstore Data Analysis

Group 315

Priyanka Hosur Mahadevu

Padmashri Adgonda Malgonnavar

Illinois Institute of Technology Chicago, IL, 60616, USA

School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

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 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.
- 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.
- Data Source https://www.kaggle.com/jr2ngb/superstore-data

Research Problems

- 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.

Attributes

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

Proposed Solutions

- Installing Packages and loaded necessary libraries.
- Reading .csv file from google drive.
- Cleaned column names using built in function.

```
> # Loading libraries
> library("googledrive")
> library("janitor")
> library("plyr")
> library("dplyr")
> library(psych)
> library(ggplot2)
> # Assigning google drive file unique id
> id <- "1gA_laz_JwONL8qRt_iCBYloZENA9Y8CZ"
> # Reading the .csv file from google drive link
> superstore_data <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id))
> # Formatting column names by replacing space with underscore
> superstore_data = clean_names(superstore_data)
>
```

Structure of Data

Structure of Data:

```
> str(superstore_data)
'data.frame':   51290 obs. of  24 variables:
 $ row_id         : int  42433 22253 48883 11731 22255 22254 21613 34662
 $ order_id       : Factor w/ 25035 levels "AE-2011-9160",..: 10 10996 9
 10847 539 132 9197 ...
 $ order_date     : Factor w/ 1430 levels "1/1/2011","1/1/2013",..: 1 1
 ...
 $ ship_date      : Factor w/ 1464 levels "1/1/2012","1/1/2013",..: 1271
 1369 1070 1070 1190 1070 ...
 $ ship_mode      : Factor w/ 4 levels "First Class",..: 4 4 3 3 4 4 3 1
 $ customer_id    : Factor w/ 1590 levels "AA-10315","AA-10375",..: 1469
 1208 907 434 1422 ...
 $ customer_name  : Factor w/ 795 levels "Aaron Bergman",..: 752 399 49
10 710 ...
 $ segment        : Factor w/ 3 levels "Consumer","Corporate",..: 1 1 1
 $ city           : Factor w/ 3636 levels "Aachen","Aalen",..: 789 3456
4 2138 1929 3577 ...
 $ state          : Factor w/ 1094 levels "'Ajman","'Amman",..: 256 703
93 583 423 ...
 $ country        : Factor w/ 147 levels "Afghanistan",..: 3 7 57 124 7
 $ postal_code    : int  NA NA NA NA NA NA NA 92691 NA NA ...
 $ market         : Factor w/ 7 levels "Africa","APAC",..: 1 2 4 5 2 2 2
 $ region         : Factor w/ 13 levels "Africa","Canada",..: 1 10 7 8 1
 $ product_id     : Factor w/ 10292 levels "FUR-ADV-10000002",..: 7847 7
80 8777 391 5194 7017 ...
 $ category       : Factor w/ 3 levels "Furniture","Office Supplies",..:
 ...
 $ sub_category   : Factor w/ 17 levels "Accessories",..: 15 16 15 13 10
 $ product_name   : Factor w/ 3788 levels "\"While you Were Out\" Messag
 Page",..: 3415 156 3384 1331 1229 1144 761 3111 1423 3437 ...
 $ sales          : num  408.3 120.4 66.1 44.9 113.7 ...
 $ quantity       : int  2 3 4 3 5 2 2 2 1 3 ...
 $ discount       : num  0 0.1 0 0.5 0.1 0.1 0 0.15 0 0 ...
 $ profit         : num  106.1 36 29.6 -26.1 37.8 ...
 $ shipping_cost  : num  35.46 9.72 8.17 4.82 4.7 ...
 $ order_priority: Factor w/ 4 levels "Critical","High",..: 4 4 2 2 4 4
```

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
```

Deal with Missing Value

- As postal code is the only column has missing value remove the column as postal code is not useful for statistical analysis.

- The ID columns are not useful for statistical analysis hence, removing rest of id columns.

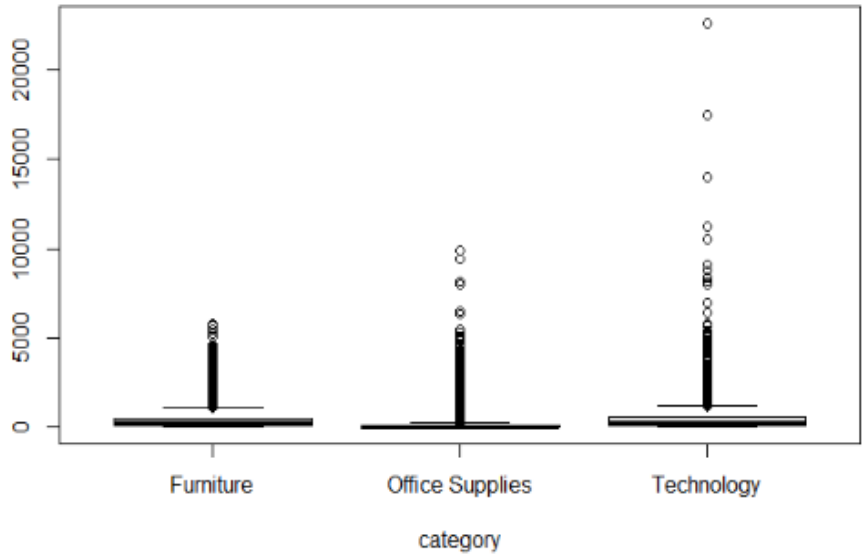
```
> # Removing these ID columns as they are not useful for statistical Analysis
> superstore_data$row_id <- NULL
> superstore_data$order_id <- NULL
> superstore_data$order_date <- NULL
> superstore_data$customer_id <- NULL
> superstore_data$customer_name <- NULL
> superstore_data$postal_code <- NULL
> 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
>
```

Anova Hypothesis 1

Sales vs Category with 95% confidence level

1) F-TEST

- H0: The mean of sales in all category is same ➔ $\mu_A = \mu_B = \mu_C$
- Hα: The mean of sales in all category is not same ➔ At least one or two group have different mean



2) INDIVIDUAL PARAMETER TEST

- H0: Difference of variables are statistically significant.
- Hα: Difference of variables are not statistically significant.

```
> 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
```

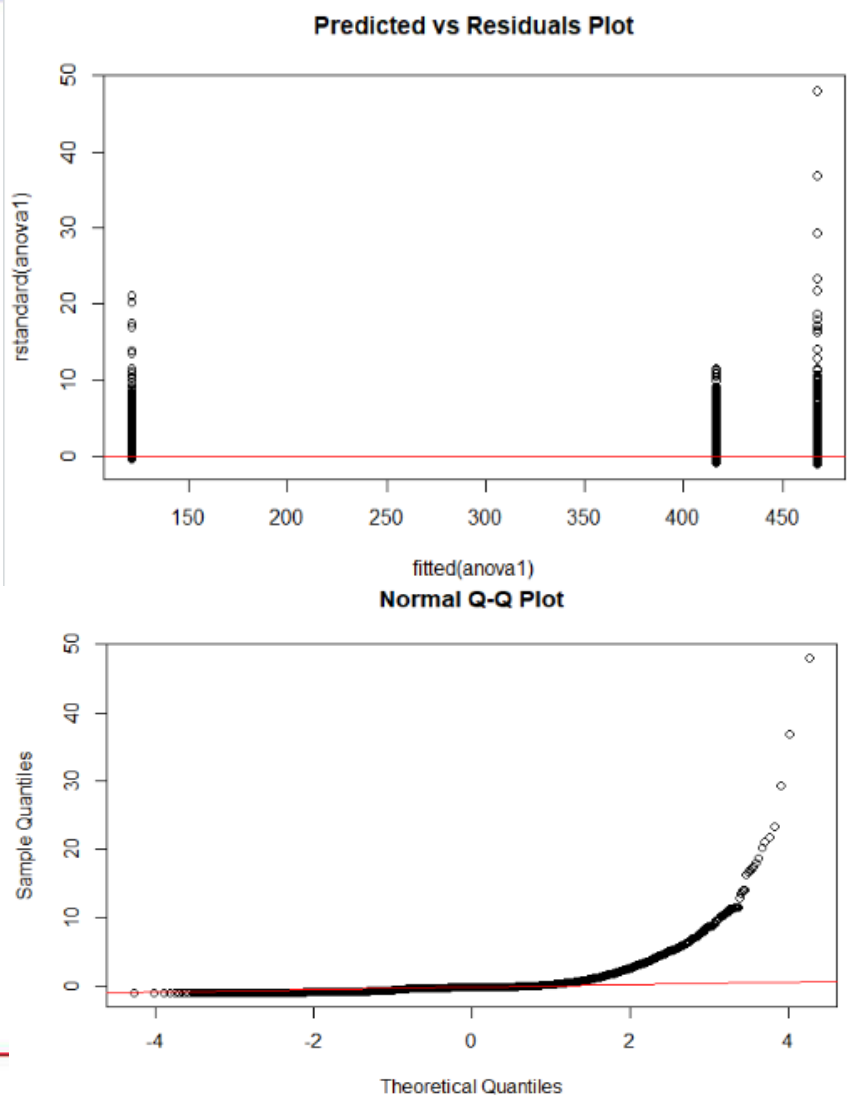
Anova Hypothesis 1

3) RESIDUAL ANALYSIS

- To check constant variance for the residuals by plotting predicted values vs residuals.
- 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.



Anova Hypothesis1

PERFORM TRANSFORMATION

Log 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
```

sqrt 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
```

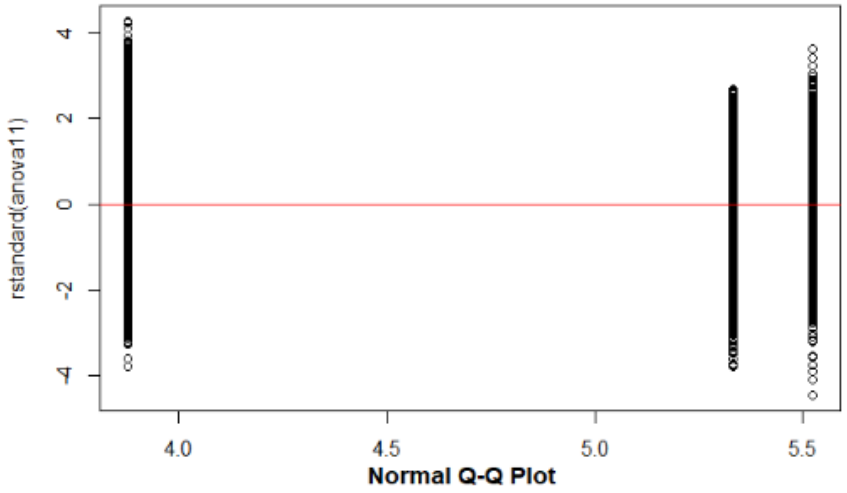
Anova Hypothesis1

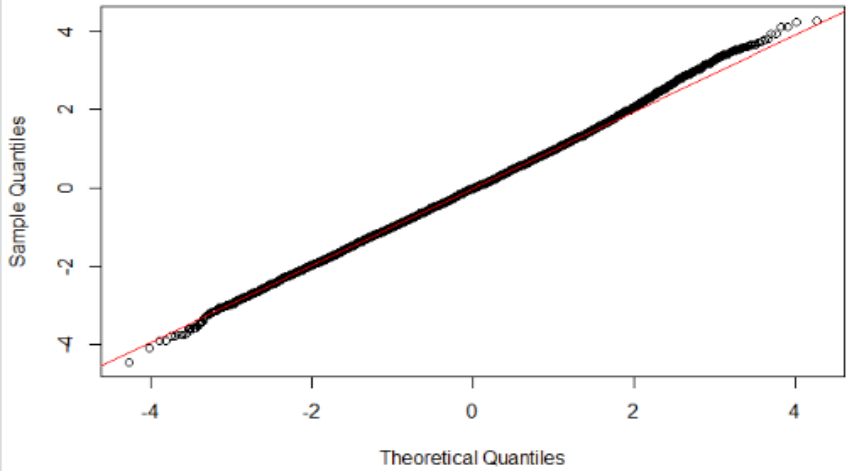
RESIDUAL ANALYSIS

- To check constant variance for the residuals by plotting predicted values vs residuals.
- Normality test

INTERPRETATION

- In the plot predicted vs residuals we observe that variables are scattered around zero line.
- From the Normality distribution Q-Q plot we observe that the points are distributed around normal line.

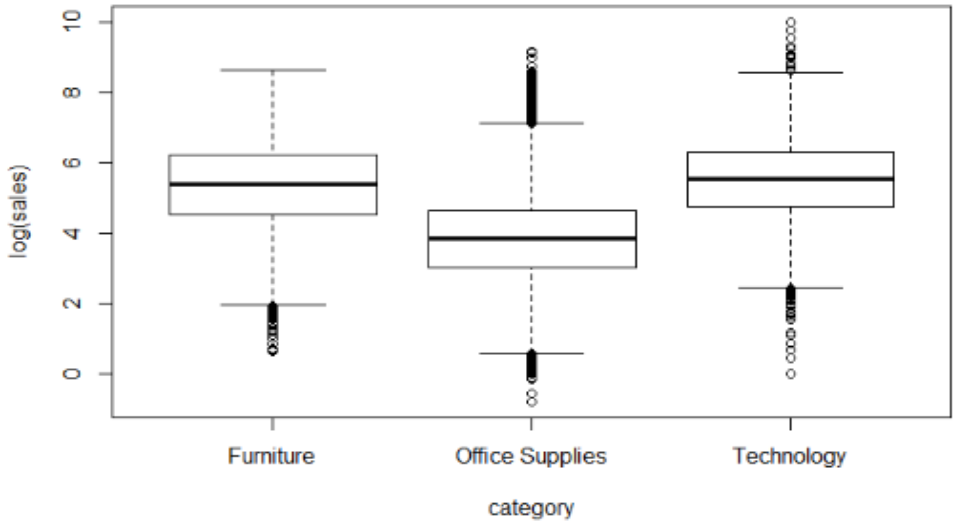




Anova Hypothesis1

ANOVA

- Compare group means among more than two groups by analyzing the variances.



F-TEST INTERPRETATION

- The F-test statistic p-value 2.2e-16(<0.05).
- As P-value < α, we don't have enough evidence to accept NULL hypothesis with 95% confidence level and accept that at least two category groups have different average sales.

```
> 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
>
```

Anova Hypothesis1

INDIVIDUAL PARAMETER TEST

Tukey (T) test

Provide a detail insight between different groups.

```
> 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
```

INTERPRETATION

- From the Tukey's test, we conclude that there is a significant difference in all the category group at adjusted p-value < 0.05.

Anova Hypothesis 2

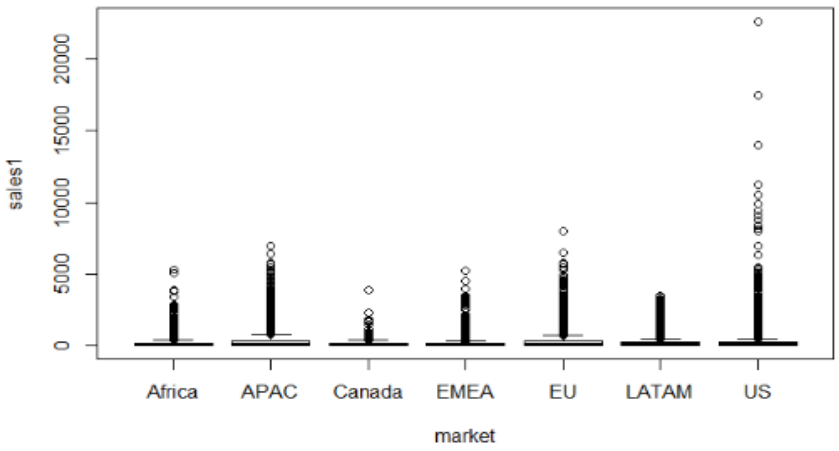
Sales vs Market with 95% confidence level

1) F-TEST

- H0: Average sales in all market groups are the same.
- Hα: : Average sales in all market groups are not the same.

2) INDIVIDUAL PARAMETER TEST

- H0: Difference of variables are statistically significant.
- Hα: Difference of variables are not statistically significant.




```
> # 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
```

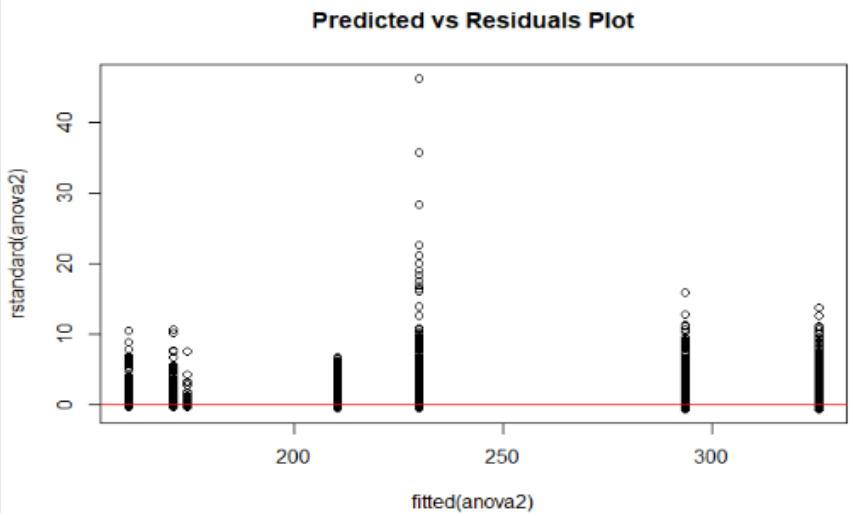
Anova Hypothesis 2

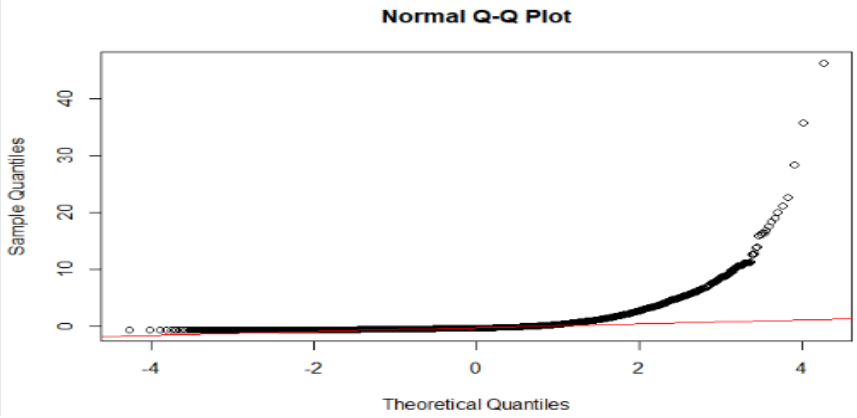
3) RESIDUAL ANALYSIS

- To check constant variance for the residuals by plotting predicted values vs residuals.
- 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.





Anova Hypothesis2

PERFORMING TRANSFORMATION

Log 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
```

sqrt 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
```

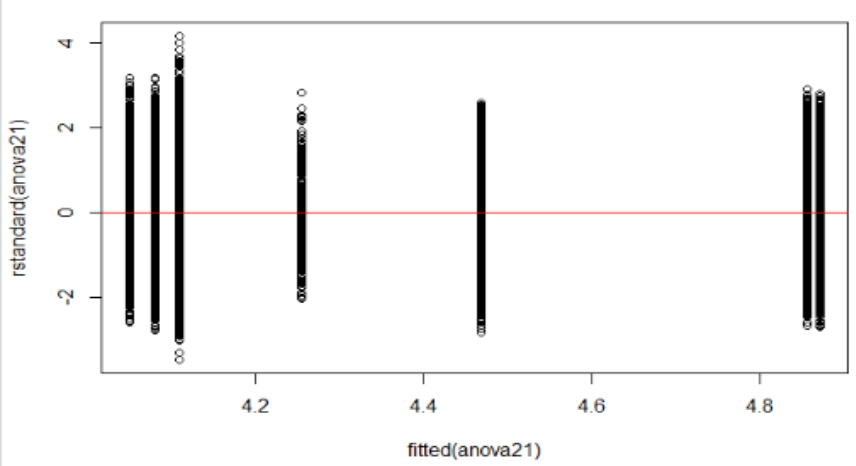
inverse transformation

```
> # 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
```

Anova Hypothesis2

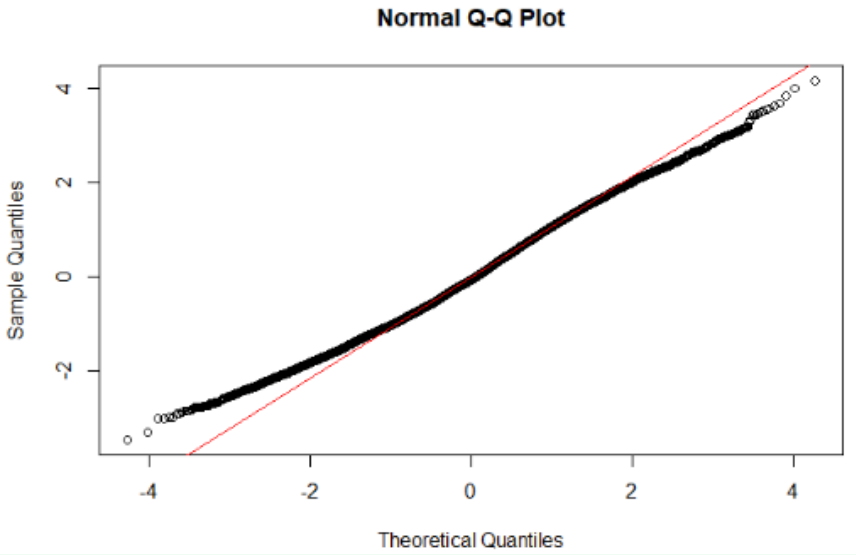
RESIDUAL ANALYSIS

- To check constant variance for the residuals by plotting predicted values vs residuals.
- Normality test



INTERPRETATION

- In the plot predicted vs residuals we observe that variables are scattered around zero line.
- From the Normality distribution Q-Q plot we observe that the points are distributed around normal line.



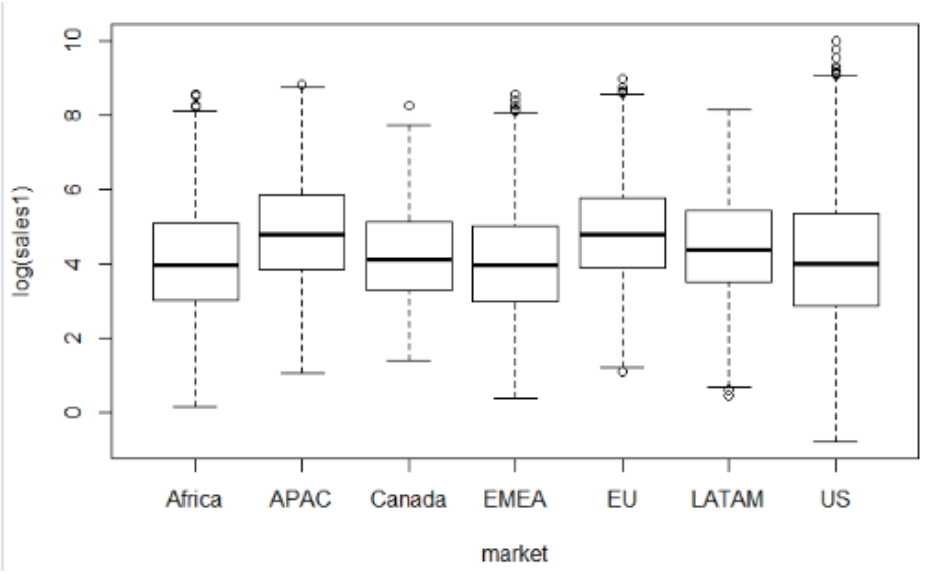
Anova Hypothesis2

ANOVA

- Compare group means among more than two groups by analyzing the variances.

F-TEST INTERPRETATION

- The F-test statistic p-value 2.2e-16(<0.05).
- As P-value < α, we don't have enough evidence to accept NULL hypothesis with 95% confidence level and accept that at least two market groups have different average sales.




```
> 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
>
```

Anova Hypothesis2

INDIVIDUAL PARAMETER TEST

Tukey (T) test

Provide a detail insight between different groups.

INTERPRETATION

- From the Tukey's test, we conclude that there is a 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, US-canada.

```
> 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
```

Data Preprocessing

- **NORMALIZATION:**

```
> ################################################################
> # Normalization
> ################################################################
> checknumericvar = sapply(superstore_data, is.numeric)
> numericvar = superstore_data[checknumericvar]
> head(numericvar)
    sales quantity discount  profit shipping_cost
1 408.300        2      0.0 106.140         35.46
2 120.366        3      0.1  36.036          9.72
3  66.120        4      0.0  29.640          8.17
4  44.865        3      0.5 -26.055          4.82
5 113.670        5      0.1  37.770          4.70
6  55.242        2      0.1  15.342          1.80
> # Min-Max normalization
> normalized_data <- as.data.frame(apply(numericvar, 2,
+                                        FUN=function(x)(x-min(x))/(max(x)-min(x))))
> head(normalized_data)
        sales   quantity  discount    profit shipping_cost
1 0.018016404 0.07692308 0.0000000 0.4470759   0.037983226
2 0.005297368 0.15384615 0.1176471 0.4424023   0.010411646
3 0.002901135 0.23076923 0.0000000 0.4419759   0.008751352
4 0.001962229 0.15384615 0.5882353 0.4382629   0.005162977
5 0.005001582 0.30769231 0.1176471 0.4425179   0.005034438
6 0.002420616 0.07692308 0.1176471 0.4410227   0.001928083
>
```

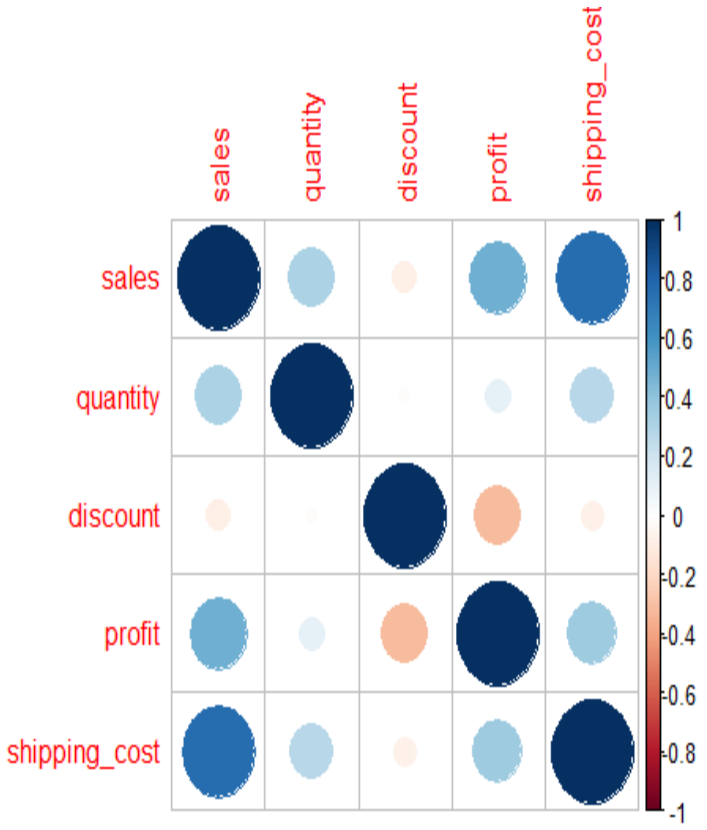
- **N-1 DUMMY VARIABLES:**

```
> ########################################################################
> # Creating N-1 Dummy Variables
> ########################################################################
> # Creating dummy variables
> library(tidyverse)
> # Fetching Categorical Variables
> categorical=superstore_data %>% select_if(negate(is.numeric))
> # Creating n-1 dummy variables
> dummydf<- data.frame(sapply(categorical,function(x) data.frame(model.
matrix(~x-1,data =categorical))[,-1]))
> dim(dummydf)
[1] 51290    44
> # Correcting features names
> dummydf = clean_names(dummydf)
```

Data Preprocessing - Correlation analysis

Correlation Assumption: There is weak correlation exists applied transformation and observed only shipping cost given slight improvement and removed quantity feature.

```
> ######################################################################
> # Correlation
> ######################################################################
> corr=cor(normalized_data)
> 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
>
```



Applied Transformation:

```
> # No improvement in corr of Qunatity hence removing quantity and there is
 slight improvement in shipping cost
> # adding that into data set
> normalized_data$quantity <- NULL
> normalized_data$shipping_cost <- NULL
> normalized_data[,"shipping_cost1"] = shipping_cost1
> corr=cor(normalized_data)
> corr
                  sales    discount     profit shipping_cost1
sales       1.00000000 -0.08672187  0.4849181     0.78158493
discount   -0.08672187  1.00000000 -0.3164902    -0.08474717
profit      0.48491811 -0.31649017  1.0000000     0.35538394
shipping_cost1 0.78158493 -0.08474717 0.3553839   1.00000000
> final_df=data.frame(normalized_data, dummydf)
```

Multiple Linear Regression Model- Split Data

- Splitting Data set into training and testing dataset with 70% and 30% respectively.
- Building regression model with 95% confidence interval.

```
> final_df=data.frame(normalized_data, dummydf)
> #number of total sales data
> dim(final_df)
[1] 51290    48
> #Fetching 70% of total sales row data as train data
> train.data = final_df[1:floor(0.7*nrow(final_df)),]
> nrow(train.data)
[1] 35903
> #Fetching rest 30% of row data as test data
> test.data = final_df[floor(0.7*nrow(final_df)+1):nrow(final_df),]
> nrow(test.data)
[1] 15387
>
```

Full Multiple Regression Model

- Built full model with training data set.
- The model summary has ‘NA’ records which says collinearity.
- Rebuilding model by removing these records.

```
> # Building Full Model
> fullmodel = lm(profit~., data = train.data)
> summary(fullmodel) #Adj-R2 = 0.3344

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.38372 -0.00188 -0.00039  0.00239  0.34857

Coefficients: (6 not defined because of singularities)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.307e-01  4.857e-04 886.825  < 2e-16 ***
sales                        2.713e-01  3.864e-03  70.203  < 2e-16 ***
discount                    -1.230e-02  2.062e-04 -59.672  < 2e-16 ***
shipping_cost1              -1.140e-03  1.654e-03  -0.689  0.49072
ship_mode_x_same_day        -2.968e-04  2.494e-04  -1.190  0.23411
ship_mode_x_second_class    -5.204e-06  1.706e-04  -0.031  0.97566
ship_mode_x_standard_class  -3.092e-05  1.566e-04  -0.197  0.84351
segment_x_corporate         -3.375e-05  1.131e-04  -0.298  0.76546
segment_x_home_office       -5.935e-05  1.334e-04  -0.445  0.65634
market_x_apac               -2.733e-04  2.616e-04  -1.045  0.29612
market_x_canada             -6.772e-04  5.893e-04  -1.149  0.25051
market_x_emea                2.340e-04  2.266e-04   1.032  0.30192
market_x_eu                 -1.453e-03  3.349e-04  -4.337 1.45e-05 ***
market_x_latam              -9.076e-04  3.401e-04  -2.668  0.00762 **
market_x_us                 -1.473e-04  2.588e-04  -0.569  0.56924
region_x_canada                     NA         NA      NA       NA
region_x_caribbean           6.197e-04  4.008e-04   1.546  0.12206
region_x_central             4.587e-04  2.747e-04   1.670  0.09496 .
region_x_central_asia       -1.537e-04  3.216e-04  -0.478  0.63265
region_x_central_asia       -1.537e-04  3.216e-04  -0.478  0.63265
region_x_east                4.547e-04  2.859e-04   1.590  0.11179
region_x_emea                       NA         NA      NA       NA
region_x_north               6.898e-04  3.242e-04   2.127  0.03340 *
region_x_north_asia          2.004e-04  3.104e-04   0.646  0.51854
region_x_oceania            -2.647e-04  2.788e-04  -0.949  0.34243
region_x_south               6.795e-05  2.843e-04   0.239  0.81112
region_x_southeast_asia             NA         NA      NA       NA
region_x_west                       NA         NA      NA       NA
category_x_office_supplies   1.135e-02  4.596e-04  24.698  < 2e-16 ***
category_x_technology        1.168e-02  4.399e-04  26.561  < 2e-16 ***
sub_category_x_appliances   -8.903e-04  3.565e-04  -2.497  0.01253 *
sub_category_x_art           2.884e-04  2.812e-04   1.026  0.30498
sub_category_x_binders       1.075e-03  2.725e-04   3.945 7.99e-05 ***
sub_category_x_bookcases     9.141e-03  4.485e-04  20.381  < 2e-16 ***
sub_category_x_chairs        9.884e-03  4.314e-04  22.913  < 2e-16 ***
sub_category_x_copiers      -2.264e-04  3.147e-04  -0.719  0.47200
sub_category_x_envelopes     4.919e-04  3.232e-04   1.522  0.12796
sub_category_x_fasteners     5.733e-04  3.233e-04   1.773  0.07617 .
sub_category_x_furnishings   1.167e-02  4.409e-04  26.464  < 2e-16 ***
sub_category_x_labels        4.202e-04  3.181e-04   1.321  0.18658
sub_category_x_machines     -2.926e-03  3.549e-04  -8.246  < 2e-16 ***
sub_category_x_paper         4.203e-04  2.999e-04   1.401  0.16114
sub_category_x_phones       -1.298e-03  2.779e-04  -4.669 3.03e-06 ***
sub_category_x_storage      -5.861e-04  2.802e-04  -2.092  0.03647 *
sub_category_x_supplies             NA         NA      NA       NA
sub_category_x_tables               NA         NA      NA       NA
order_priority_x_high       -3.617e-04  2.081e-04  -1.738  0.08224 .
order_priority_x_low        -1.909e-04  3.161e-04  -0.604  0.54578
order_priority_x_medium     -3.126e-04  2.147e-04  -1.456  0.14540
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009312 on 35861 degrees of freedom
Multiple R-squared:  0.3352,    Adjusted R-squared:  0.3344
F-statistic: 440.9 on 41 and 35861 DF,  p-value: < 2.2e-16
```

Full Multiple Regression Model

- Multiple collinearity exists as we could see records with 'NA' value. Removed these records- built model again.

```
> summary(fullmodel1) # Adj-R2 = 0.3344

Call:
lm(formula = profit ~ sales + discount + shipping_cost1 + ship_mode_x_same_day +
    ship_mode_x_second_class + ship_mode_x_standard_class + segment_x_corporate +
    segment_x_home_office + market_x_apac + market_x_canada +
    market_x_emea + market_x_eu + market_x_latam + market_x_us +
    region_x_caribbean + region_x_central + region_x_central_asia +
    region_x_east + region_x_north + region_x_north_asia + region_x_oceania +
    region_x_south + category_x_office_supplies + category_x_technology +
    sub_category_x_appliances + sub_category_x_art + sub_category_x_binders +
    sub_category_x_bookcases + sub_category_x_chairs + sub_category_x_copiers +
    sub_category_x_envelopes + sub_category_x_fasteners + sub_category_x_furnishings +
    sub_category_x_labels + sub_category_x_machines + sub_category_x_paper +
    sub_category_x_phones + sub_category_x_storage + order_priority_x_high +
    order_priority_x_low + order_priority_x_medium, data = train.data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.38372 -0.00188 -0.00039  0.00239  0.34857

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.307e-01  4.857e-04 886.825  < 2e-16 ***
sales                        2.713e-01  3.864e-03  70.203  < 2e-16 ***
discount                    -1.230e-02  2.062e-04 -59.672  < 2e-16 ***
shipping_cost1              -1.140e-03  1.654e-03  -0.689  0.49072
ship_mode_x_same_day        -2.968e-04  2.494e-04  -1.190  0.23411
ship_mode_x_second_class    -5.204e-06  1.706e-04  -0.031  0.97566
ship_mode_x_standard_class  -3.092e-05  1.566e-04  -0.197  0.84351
segment_x_corporate         -3.375e-05  1.131e-04  -0.298  0.76546
segment_x_home_office       -5.935e-05  1.334e-04  -0.445  0.65634
market_x_apac               -2.733e-04  2.616e-04  -1.045  0.29612
market_x_canada             -6.772e-04  5.893e-04  -1.149  0.25051
market_x_emea                2.340e-04  2.266e-04   1.032  0.30192
market_x_eu                 -1.453e-03  3.349e-04  -4.337 1.45e-05 ***
market_x_latam              -9.076e-04  3.401e-04  -2.668  0.00762 **
market_x_us                 -1.473e-04  2.588e-04  -0.569  0.56924
region_x_caribbean           6.197e-04  4.008e-04   1.546  0.12206
region_x_central             4.587e-04  2.747e-04   1.670  0.09496 .
region_x_central_asia       -1.537e-04  3.216e-04  -0.478  0.63265
region_x_east                4.547e-04  2.859e-04   1.590  0.11179
region_x_north               6.898e-04  3.242e-04   2.127  0.03340 *
region_x_north_asia          2.004e-04  3.104e-04   0.646  0.51854
region_x_oceania            -2.647e-04  2.788e-04  -0.949  0.34243
region_x_south               6.795e-05  2.843e-04   0.239  0.81112
category_x_office_supplies   1.135e-02  4.596e-04  24.698  < 2e-16 ***
category_x_technology        1.168e-02  4.399e-04  26.561  < 2e-16 ***
sub_category_x_appliances   -8.903e-04  3.565e-04  -2.497  0.01253 *
sub_category_x_art           2.884e-04  2.812e-04   1.026  0.30498
sub_category_x_binders       1.075e-03  2.725e-04   3.945 7.99e-05 ***
sub_category_x_bookcases     9.141e-03  4.485e-04  20.381  < 2e-16 ***
sub_category_x_chairs        9.884e-03  4.314e-04  22.913  < 2e-16 ***
sub_category_x_copiers      -2.264e-04  3.147e-04  -0.719  0.47200
sub_category_x_envelopes     4.919e-04  3.232e-04   1.522  0.12796
sub_category_x_fasteners     5.733e-04  3.233e-04   1.773  0.07617 .
sub_category_x_furnishings   1.167e-02  4.409e-04  26.464  < 2e-16 ***
sub_category_x_labels        4.202e-04  3.181e-04   1.321  0.18658
sub_category_x_machines     -2.926e-03  3.549e-04  -8.246  < 2e-16 ***
sub_category_x_paper         4.203e-04  2.999e-04   1.401  0.16114
sub_category_x_phones       -1.298e-03  2.779e-04  -4.669 3.03e-06 ***
sub_category_x_storage      -5.861e-04  2.802e-04  -2.092  0.03647 *
order_priority_x_high       -3.617e-04  2.081e-04  -1.738  0.08224 .
order_priority_x_low        -1.909e-04  3.161e-04  -0.604  0.54578
order_priority_x_medium     -3.126e-04  2.147e-04  -1.456  0.14540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009312 on 35861 degrees of freedom
Multiple R-squared:  0.3352,    Adjusted R-squared:  0.3344
F-statistic: 440.9 on 41 and 35861 DF,  p-value: < 2.2e-16
```

Full Multiple Regression Model-Assumptions

- Checking for **variance Inflation factor** to remove multicollinearity as a data post processing.
- As many we could observe many variables has VIF > 4, removing them from highest to lowest and rebuilding model each time.
- Written custom function to rebuild a model each time.

```
> # VIFS
> vifs=vif(fullmodel1)
> vifs
                    sales                  discount             shipping_cost1
                 2.889025                  1.102437                  3.035231
      ship_mode_x_same_day ship_mode_x_second_class ship_mode_x_standard_class
                 1.290062                  1.917332                  2.437302
       segment_x_corporate      segment_x_home_office             market_x_apac
                 1.109756                  1.108736                  4.724344
          market_x_canada             market_x_emea              market_x_eu
                 1.081109                  1.882699                  7.291743
           market_x_latam              market_x_us       region_x_caribbean
                 7.654930                  4.395671                  2.140171
         region_x_central      region_x_central_asia            region_x_east
                 5.297443                  1.639103                  1.807593
           region_x_north        region_x_north_asia         region_x_oceania
                 3.699451                  1.733487                  1.993092
           region_x_south category_x_office_supplies      category_x_technology
                 3.788564                 20.850177                 12.779111
 sub_category_x_appliances       sub_category_x_art     sub_category_x_binders
                 1.763418                  2.850086                  3.236264
  sub_category_x_bookcases      sub_category_x_chairs     sub_category_x_copiers
                 3.715049                  4.841625                  1.694303
 sub_category_x_envelopes   sub_category_x_fasteners sub_category_x_furnishings
                 1.940681                  1.938890                  4.711552
    sub_category_x_labels    sub_category_x_machines       sub_category_x_paper
                 2.004998                  1.462871                  2.381495
    sub_category_x_phones     sub_category_x_storage     order_priority_x_high
                 1.978635                  2.884611                  3.789614
     order_priority_x_low    order_priority_x_medium
                 1.781168                  4.664031
```

```
> # Function to rebuild model by removing variables having VIF>4
> library(car)
> verify_vif <- function(model, df, vifs)
+ {
+   vifcoulmnnames <- names(vifs)
+
+   # Remove predictors with VIF > 4 and re-build model until none of VIFs don't exceed 4
+   while(any(vifs > 4)){
+     maximum_vif <- names(which(vifs == max(vifs)))  # get the var with max vif
+     vifcoulmnnames <- vifcoulmnnames[!(vifcoulmnnames) %in% maximum_vif]  # remove maximum VIF first
+     # Craeting new regression formula by removing max VIF variable
+     formula <- as.formula(paste("profit ~ ", paste (vifcoulmnnames, collapse=" + "), sep=""))
+     # Rebuilding model for each maximum VIF removal
+     rebuiltmodel <- lm(formula, data=df)
+     # Fetching vifs list for each model build
+     vifs <- car::vif(rebuiltmodel)
+   }
+   return(rebuiltmodel)
+ }
>
```

Full Multiple Regression Model

- Calling custom ‘verify_vif’ function remove VIF > 4 records and rebuild model.
- In rebuilt model all features has VIF < 4.

```
> vifs=vif(vifmodel)
> vifs
                     sales                   discount              shipping_cost1
                  2.779681                   1.061307                   2.848204
      ship_mode_x_same_day ship_mode_x_second_class ship_mode_x_standard_class
                  1.287438                   1.899863                   2.203451
       segment_x_corporate      segment_x_home_office             market_x_canada
                  1.109127                   1.108559                   1.042142
             market_x_emea              market_x_eu              market_x_us
                  1.405705                   1.720308                   1.748922
       region_x_caribbean         region_x_central      region_x_central_asia
                  1.152953                   2.042540                   1.192733
            region_x_east            region_x_north       region_x_north_asia
                  1.483168                   1.502132                   1.222728
         region_x_oceania           region_x_south sub_category_x_appliances
                  1.290520                   1.516793                   1.259746
       sub_category_x_art     sub_category_x_binders    sub_category_x_bookcases
                  1.656287                   1.774065                   1.347487
     sub_category_x_chairs     sub_category_x_copiers   sub_category_x_envelopes
                  1.458959                   1.339489                   1.333734
  sub_category_x_fasteners sub_category_x_furnishings      sub_category_x_labels
                  1.337665                   1.426643                   1.359886
   sub_category_x_machines       sub_category_x_paper      sub_category_x_phones
                  1.221618                   1.486494                   1.459282
    sub_category_x_storage      order_priority_x_high       order_priority_x_low
                  1.638520                   1.073561                   1.052052
```

```
> # Verify multicollinearity by checking variance inflation factor and rebuild model if any
> vifmodel = verify_vif(fullmodel1, train.data, vifs)
> summary(vifmodel) # Adj-R2 = 0.3203

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.37388 -0.00193 -0.00031  0.00247  0.35650

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)                4.406e-01  2.327e-04 1893.542  < 2e-16 ***
sales                      2.618e-01  3.830e-03   68.356  < 2e-16 ***
discount                  -1.291e-02  2.045e-04  -63.131  < 2e-16 ***
shipping_cost1            -3.276e-03  1.619e-03   -2.024 0.042986 *
ship_mode_x_same_day      -2.253e-04  2.518e-04   -0.895 0.370881
ship_mode_x_second_class  -3.321e-05  1.716e-04   -0.194 0.846529
ship_mode_x_standard_class -1.437e-04  1.505e-04   -0.955 0.339748
segment_x_corporate       -5.609e-05  1.143e-04   -0.491 0.623578
segment_x_home_office     -4.889e-05  1.348e-04   -0.363 0.716799
market_x_canada           -5.830e-04  5.847e-04   -0.997 0.318736
market_x_emea              4.672e-04  1.979e-04    2.361 0.018253 *
market_x_eu               -6.642e-04  1.644e-04   -4.040 5.35e-05 ***
market_x_us                1.615e-04  1.650e-04    0.979 0.327559
region_x_caribbean        -1.292e-04  2.973e-04   -0.435 0.663722
region_x_central          -2.959e-05  1.724e-04   -0.172 0.863680
region_x_central_asia     -1.900e-04  2.772e-04   -0.685 0.493252
region_x_east              2.609e-04  2.617e-04    0.997 0.318910
region_x_north             2.992e-05  2.088e-04    0.143 0.886063
region_x_north_asia        1.684e-04  2.635e-04    0.639 0.522833
region_x_oceania          -2.300e-04  2.267e-04   -1.014 0.310359
region_x_south            -4.713e-04  1.818e-04   -2.592 0.009539 **
sub_category_x_appliances  7.051e-04  3.045e-04    2.315 0.020602 *
sub_category_x_art         1.527e-03  2.166e-04    7.049 1.83e-12 ***
sub_category_x_binders     2.361e-03  2.039e-04   11.578  < 2e-16 ***
sub_category_x_bookcases  -6.578e-04  2.730e-04   -2.410 0.015955 *
sub_category_x_chairs      4.504e-06  2.393e-04    0.019 0.984983
sub_category_x_copiers     1.671e-03  2.828e-04    5.908 3.49e-09 ***
sub_category_x_envelopes   1.699e-03  2.707e-04    6.277 3.50e-10 ***
sub_category_x_fasteners   1.760e-03  2.713e-04    6.486 8.93e-11 ***
sub_category_x_furnishings 1.607e-03  2.451e-04    6.557 5.58e-11 ***
sub_category_x_labels      1.600e-03  2.648e-04    6.044 1.52e-09 ***
sub_category_x_machines   -1.047e-03  3.277e-04   -3.194 0.001403 **
sub_category_x_paper       1.667e-03  2.394e-04    6.963 3.38e-12 ***
sub_category_x_phones      5.667e-04  2.412e-04    2.350 0.018788 *
sub_category_x_storage     7.556e-04  2.134e-04    3.541 0.000399 ***
order_priority_x_high     -6.240e-05  1.119e-04   -0.557 0.577253
order_priority_x_low       1.253e-04  2.455e-04    0.511 0.609605
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.00941 on 35866 degrees of freedom
Multiple R-squared:  0.321,     Adjusted R-squared:  0.3203
F-statistic:   471 on 36 and 35866 DF,  p-value: < 2.2e-16

>
```

Full Multiple Regression Model - Assumptions

- Backward elimination of non-significant features by p-value > 0.05 and recursively rebuild model each time using custom function.
- **Goodness of Fit: 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.

- **Interpretation:**

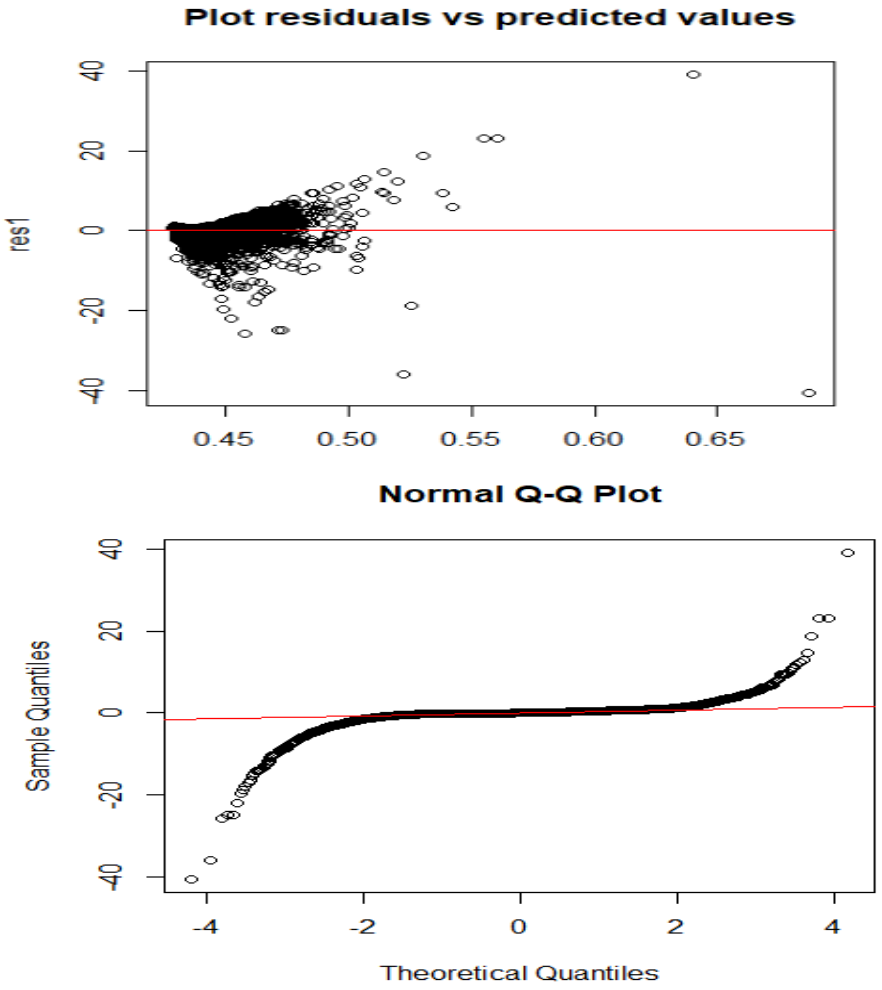
P-value of the F-statistic is < 2.2e-16, which is highly significant means reject Null Hypothesis. This means that, at least, one of the predictor variables is significantly related to the outcome variable.

```
> # 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){
+     print("Insidewhile")
+     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=""))  #
w formula
+     print("before model")
+     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(vifmodel, train.data)
> summary(eliminationmodel) # Adj-R2 = 0.3204

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.36810 -0.00190 -0.00032  0.00248  0.36006

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              0.4404041  0.0001301 3385.390  < 2e-16 ***
sales                    0.2560484  0.0025433  100.674  < 2e-16 ***
discount                -0.0128831  0.0002016  -63.911  < 2e-16 ***
market_x_emea            0.0004440  0.0001718    2.585 0.009738 **
market_x_eu             -0.0007078  0.0001294   -5.468 4.59e-08 ***
region_x_south          -0.0004426  0.0001496   -2.958 0.003099 **
sub_category_x_appliances  0.0007030  0.0002914    2.413 0.015847 *
sub_category_x_art       0.0015572  0.0002000    7.785 7.16e-15 ***
sub_category_x_binders   0.0024065  0.0001859   12.947  < 2e-16 ***
sub_category_x_bookcases -0.0006957  0.0002580   -2.696 0.007011 **
sub_category_x_copiers   0.0016021  0.0002674    5.991 2.11e-09 ***
sub_category_x_envelopes 0.0017032  0.0002575    6.615 3.76e-11 ***
sub_category_x_fasteners 0.0017608  0.0002581    6.823 9.07e-12 ***
sub_category_x_furnishings 0.0016528  0.0002302    7.178 7.19e-13 ***
sub_category_x_labels    0.0016183  0.0002514    6.437 1.24e-10 ***
sub_category_x_machines -0.0010659  0.0003156   -3.378 0.000732 ***
sub_category_x_paper     0.0017379  0.0002235    7.776 7.66e-15 ***
sub_category_x_phones    0.0005754  0.0002248    2.559 0.010487 *
sub_category_x_storage   0.0007699  0.0001962    3.924 8.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009409 on 35884 degrees of freedom
Multiple R-squared:  0.3207,    Adjusted R-squared:  0.3204
F-statistic: 941.2 on 18 and 35884 DF,  p-value: < 2.2e-16
```

Elimination Model

- Performing Residual Analysis: There is no constant variance and normality.
- It requires transformation or other methods to improve model.

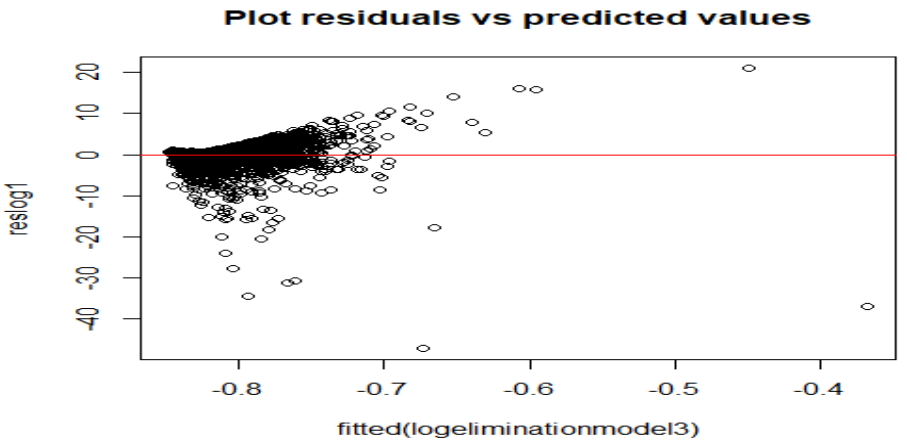


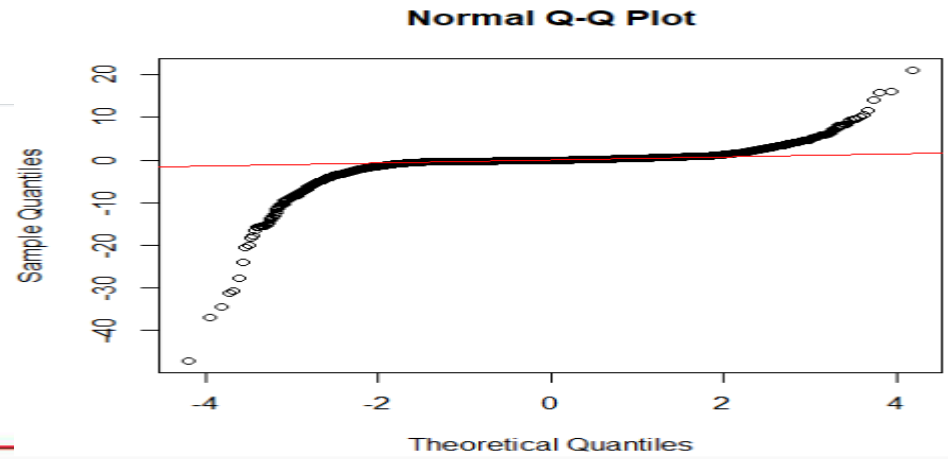
Elimination Model-Transformations

- Log 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
-1.01971 -0.00463 -0.00109  0.00558  0.44863

Coefficients:
                          Estimate Std. Error   t value Pr(>|t|)
(Intercept)             -0.8185643  0.0003012 -2717.487  < 2e-16 ***
sales                    0.4729659  0.0058891    80.313  < 2e-16 ***
discount                -0.0306036  0.0004668   -65.567  < 2e-16 ***
market_x_emea            0.0008013  0.0003977     2.015 0.043955 *
market_x_eu             -0.0014638  0.0002997    -4.884 1.05e-06 ***
region_x_south          -0.0011918  0.0003464    -3.440 0.000582 ***
sub_category_x_appliances  0.0019029  0.0006747    2.820 0.004800 **
sub_category_x_art         0.0025901  0.0004631    5.592 2.26e-08 ***
sub_category_x_binders     0.0043859  0.0004304   10.191  < 2e-16 ***
sub_category_x_bookcases  -0.0006118  0.0005974   -1.024 0.305833
sub_category_x_copiers     0.0043302  0.0006192    6.993 2.74e-12 ***
sub_category_x_envelopes   0.0029180  0.0005962    4.895 9.89e-07 ***
sub_category_x_fasteners   0.0028986  0.0005976    4.851 1.24e-06 ***
sub_category_x_furnishings 0.0030563  0.0005331    5.733 9.97e-09 ***
sub_category_x_labels      0.0025077  0.0005822    4.307 1.66e-05 ***
sub_category_x_machines   -0.0024806  0.0007307   -3.395 0.000687 ***
sub_category_x_paper       0.0029572  0.0005175    5.715 1.11e-08 ***
sub_category_x_phones      0.0017306  0.0005206    3.324 0.000887 ***
sub_category_x_storage     0.0014457  0.0004543    3.182 0.001462 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02179 on 35884 degrees of freedom
Multiple R-squared:  0.2654,    Adjusted R-squared:  0.265
F-statistic: 720.3 on 18 and 35884 DF,  p-value: < 2.2e-16
```

- Performing Residual analysis on log transformed model.



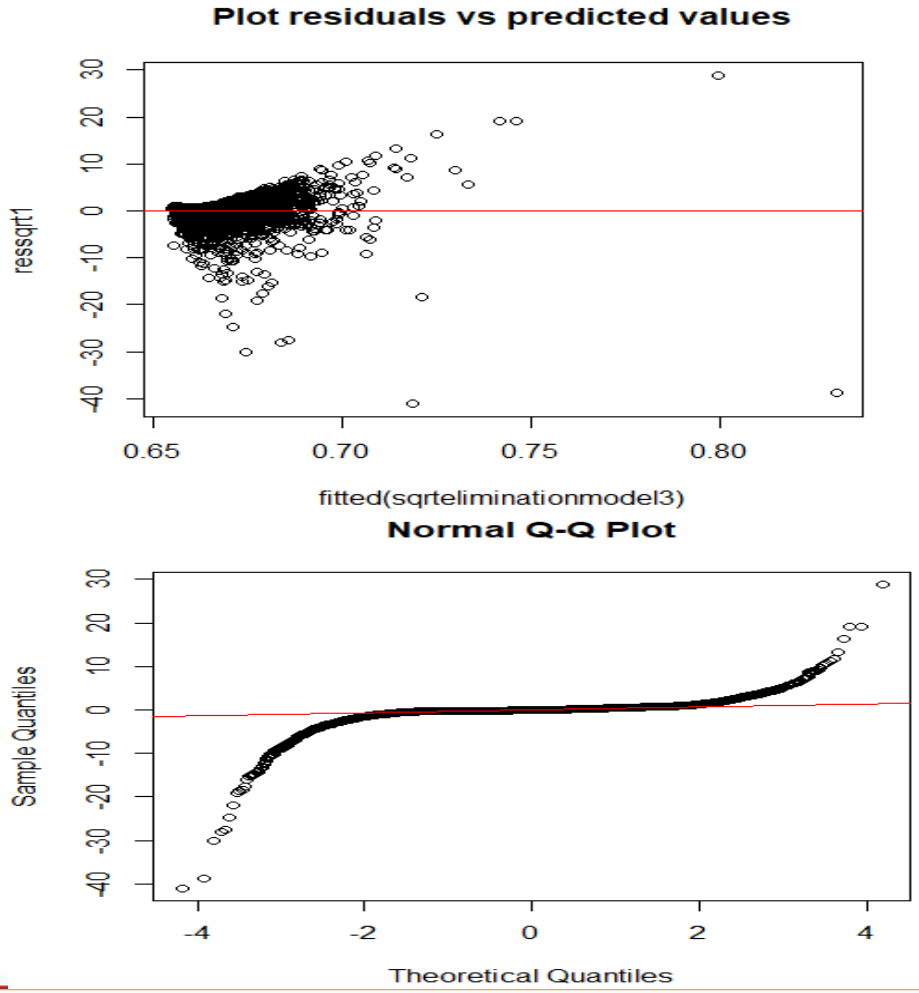


Elimination Model-Transformations

- Sqrt Transformation on Y variable.

```
> #SQRT Transformation on Y variable
> sqrtformula1 <- as.formula(paste("sqrt(profit) ~ ", paste (eliminationmodel_va
r, collapse=" + "), sep=""))  # new formula
> sqrteliminationmodel3 <- lm(sqrtformula1, data=train.data)
> summary(sqrteliminationmodel3)
Call:
lm(formula = sqrtformula1, data = train.data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.289500 -0.001484 -0.000312  0.001850  0.200118

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)                6.639e-01  9.809e-05 6768.441  < 2e-16 ***
sales                      1.742e-01  1.918e-03   90.850  < 2e-16 ***
discount                  -9.925e-03  1.520e-04  -65.301  < 2e-16 ***
market_x_emea              2.978e-04  1.295e-04    2.300 0.021475 *
market_x_eu               -5.073e-04  9.760e-05   -5.198 2.02e-07 ***
region_x_south            -3.631e-04  1.128e-04   -3.218 0.001291 **
sub_category_x_appliances  5.885e-04  2.197e-04    2.678 0.007399 **
sub_category_x_art         1.008e-03  1.508e-04    6.684 2.35e-11 ***
sub_category_x_binders     1.626e-03  1.401e-04   11.599  < 2e-16 ***
sub_category_x_bookcases  -3.557e-04  1.945e-04   -1.828 0.067501 .
sub_category_x_copiers     1.322e-03  2.016e-04    6.556 5.62e-11 ***
sub_category_x_envelopes   1.117e-03  1.941e-04    5.756 8.69e-09 ***
sub_category_x_fasteners   1.135e-03  1.946e-04    5.832 5.54e-09 ***
sub_category_x_furnishings 1.123e-03  1.736e-04    6.468 1.01e-10 ***
sub_category_x_labels      1.016e-03  1.896e-04    5.360 8.37e-08 ***
sub_category_x_machines   -7.972e-04  2.379e-04   -3.351 0.000807 ***
sub_category_x_paper       1.137e-03  1.685e-04    6.748 1.52e-11 ***
sub_category_x_phones      5.087e-04  1.695e-04    3.001 0.002696 **
sub_category_x_storage     5.262e-04  1.479e-04    3.557 0.000376 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007095 on 35884 degrees of freedom
Multiple R-squared:  0.295,     Adjusted R-squared:  0.2947
F-statistic: 834.3 on 18 and 35884 DF,  p-value: < 2.2e-16
```

- Residual Analysis on Sqrt Transformed model.

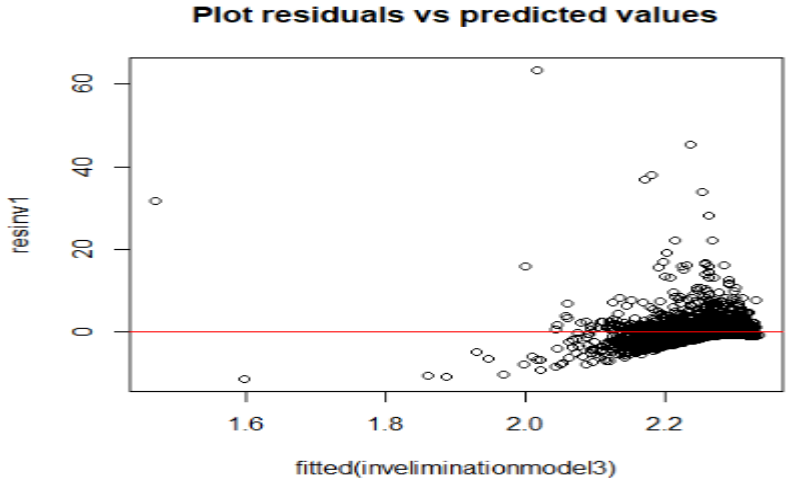


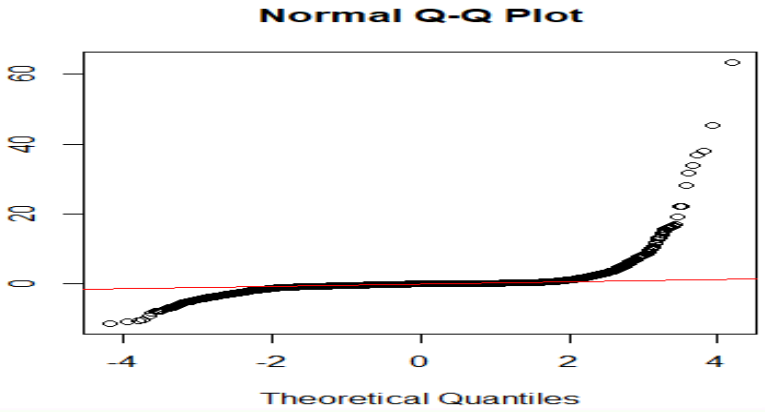
Elimination Model-Transformations

- **Inverse transformation** on Y variable.

```
> # 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
-0.5960 -0.0128  0.0031  0.0112  3.4179

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              2.2640632  0.0007502 3017.759  < 2e-16 ***
sales                   -0.8493552  0.0146678  -57.906  < 2e-16 ***
discount                 0.0730704  0.0011625   62.854  < 2e-16 ***
market_x_emea           -0.0014524  0.0009906   -1.466 0.142625
market_x_eu              0.0030671  0.0007465    4.108 3.99e-05 ***
region_x_south           0.0032529  0.0008629    3.770 0.000164 ***
sub_category_x_appliances -0.0047195  0.0016805   -2.808 0.004981 **
sub_category_x_art       -0.0039817  0.0011535   -3.452 0.000558 *
sub_category_x_binders   -0.0077807  0.0010720   -7.258 4.00e-13 *
sub_category_x_bookcases -0.0005702  0.0014880   -0.383 0.701551
sub_category_x_copiers   -0.0115342  0.0015424   -7.478 7.70e-14 *
sub_category_x_envelopes -0.0047276  0.0014849   -3.184 0.001455 *
sub_category_x_fasteners -0.0043778  0.0014884   -2.941 0.003271 *
sub_category_x_furnishings -0.0055784  0.0013279   -4.201 2.66e-05 *
sub_category_x_labels    -0.0033329  0.0014500   -2.299 0.021539 *
sub_category_x_machines   0.0067359  0.0018199    3.701 0.000215 *
sub_category_x_paper     -0.0047185  0.0012888   -3.661 0.000252 *
sub_category_x_phones    -0.0047336  0.0012966   -3.651 0.000262 *
sub_category_x_storage   -0.0027128  0.0011315   -2.397 0.016514 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05427 on 35884 degrees of freedom
Multiple R-squared:  0.198,     Adjusted R-squared:  0.1976
F-statistic: 492.2 on 18 and 35884 DF,  p-value: < 2.2e-16
```

- **Residual Analysis on inverse Transformed model.**



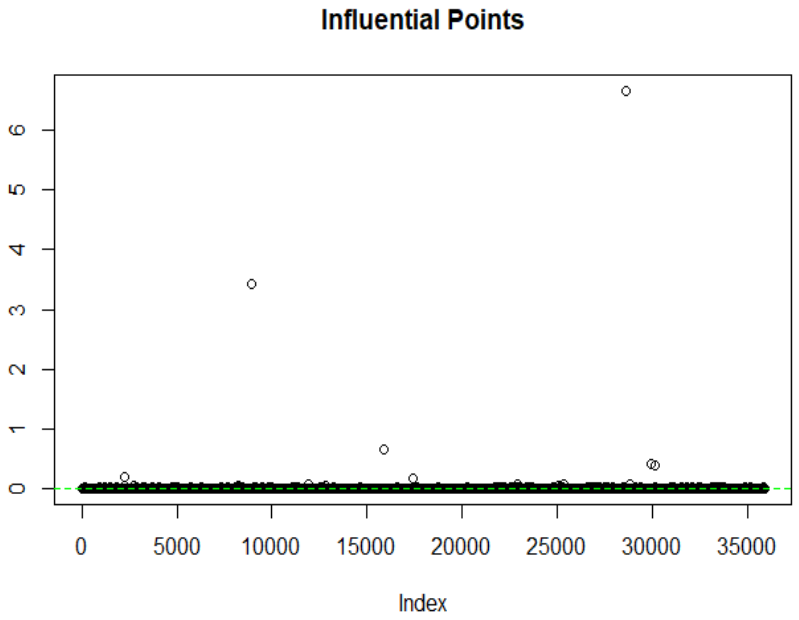


Elimination Model-Influential Points

- Removing influential points in order to improve regression model.

```
> regformula1 <- as.formula(paste("profit ~ ", paste (eliminationmodel_var, col
lapse=" + "), sep=""))  # new formula
> eliminationmodel3 <- lm(regformula1, data=newtrain.data12)
> summary(eliminationmodel3) # Adj-R2 = 0.5201

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

Residuals:
       Min         1Q     Median         3Q        Max
-0.0301702 -0.0011845 -0.0001877  0.0014947  0.0248040

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)                4.403e-01  5.547e-05 7937.933  < 2e-16 ***
sales                      2.491e-01  1.737e-03  143.371  < 2e-16 ***
discount                  -7.435e-03  8.356e-05  -88.974  < 2e-16 ***
market_x_emea              1.941e-04  6.963e-05    2.787  0.00532 **
market_x_eu               -3.604e-04  5.296e-05   -6.804 1.03e-11 ***
region_x_south            -5.330e-05  6.116e-05   -0.872  0.38348
sub_category_x_appliances  8.130e-04  1.239e-04    6.564 5.32e-11 ***
sub_category_x_art         7.739e-04  8.129e-05    9.521  < 2e-16 ***
sub_category_x_binders     1.348e-03  7.615e-05   17.706  < 2e-16 ***
sub_category_x_bookcases  -6.321e-04  1.088e-04   -5.807 6.41e-09 ***
sub_category_x_copiers     5.538e-04  1.141e-04    4.855 1.21e-06 ***
sub_category_x_envelopes   8.849e-04  1.040e-04    8.508  < 2e-16 ***
sub_category_x_fasteners   8.911e-04  1.045e-04    8.528  < 2e-16 ***
sub_category_x_furnishings 7.695e-04  9.298e-05    8.277  < 2e-16 ***
sub_category_x_labels      8.653e-04  1.019e-04    8.488  < 2e-16 ***
sub_category_x_machines    1.705e-05  1.338e-04    0.127  0.89860
sub_category_x_paper       1.047e-03  9.056e-05   11.566  < 2e-16 ***
sub_category_x_phones      2.220e-04  9.278e-05    2.393  0.01671 *
sub_category_x_storage     1.450e-04  7.943e-05    1.825  0.06794 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003766 on 34706 degrees of freedom
Multiple R-squared:  0.5204,    Adjusted R-squared:  0.5201
F-statistic:  2092 on 18 and 34706 DF,  p-value: < 2.2e-16
```

```
> # The transformations on Y reduced R-square value so ignoring transformations
> # 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] 34725
>
```



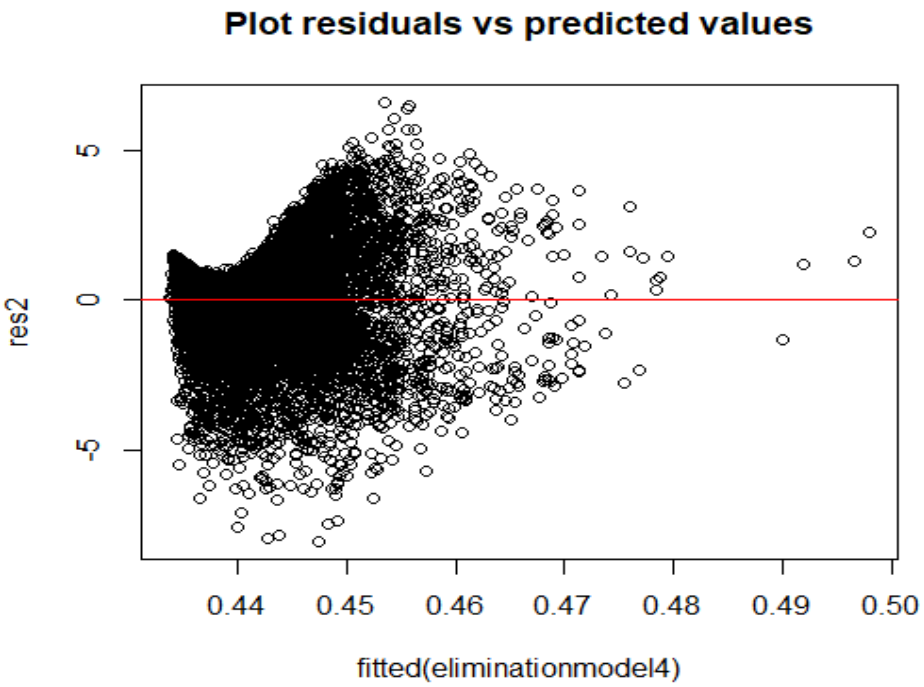
Elimination Model

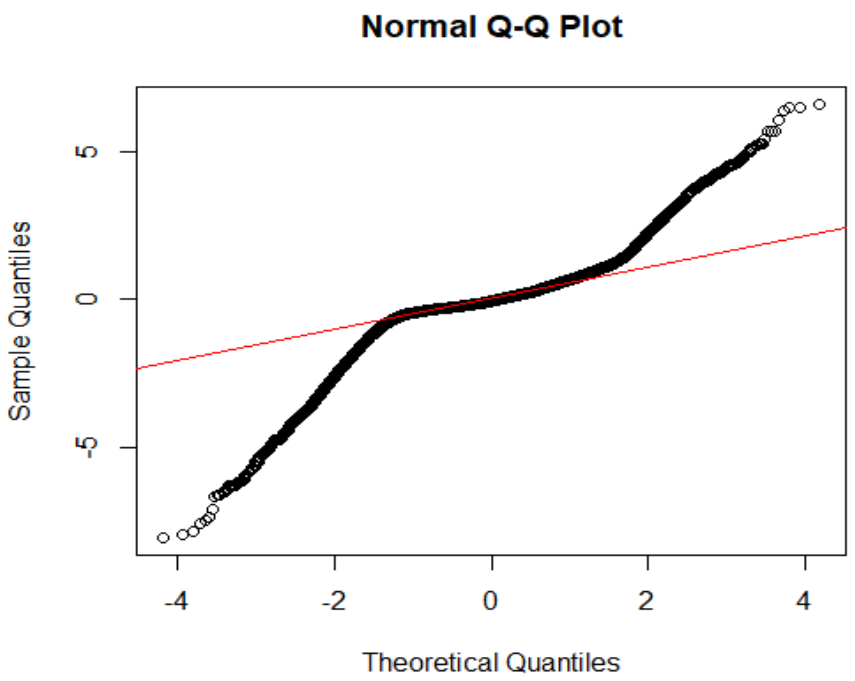
- Rebuilding model as there are still non-significant ‘x’ variables using custom function.
- Rebuilt model shows all x variables are statistically significant to predict profit with p-value less than significance level 0.05 and rejects Null Hypothesis.

```
> eliminationmodel4 = remove_non_sig_var(eliminationmodel3, newtrain.data12)
> summary(eliminationmodel4) # Adj-R2 = 0.5201

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

Residuals:
       Min         1Q     Median         3Q        Max
-0.0302028 -0.0011866 -0.0001851  0.0014924  0.0247725

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              4.404e-01  4.648e-05 9474.889  < 2e-16 ***
sales                    2.488e-01  1.729e-03  143.935  < 2e-16 ***
discount                -7.439e-03  8.352e-05  -89.078  < 2e-16 ***
market_x_emea            2.067e-04  6.901e-05    2.996  0.00274 **
market_x_eu             -3.574e-04  5.249e-05   -6.808 1.00e-11 ***
sub_category_x_appliances  7.662e-04  1.204e-04    6.364 1.99e-10 ***
sub_category_x_art       7.236e-04  7.519e-05    9.624  < 2e-16 ***
sub_category_x_binders   1.299e-03  6.984e-05   18.599  < 2e-16 ***
sub_category_x_bookcases -6.782e-04  1.049e-04   -6.463 1.04e-10 ***
sub_category_x_copiers   5.074e-04  1.103e-04    4.598 4.28e-06 ***
sub_category_x_envelopes 8.346e-04  9.953e-05    8.386  < 2e-16 ***
sub_category_x_fasteners 8.415e-04  1.000e-04    8.414  < 2e-16 ***
sub_category_x_furnishings 7.203e-04  8.810e-05    8.175 3.05e-16 ***
sub_category_x_labels    8.163e-04  9.736e-05    8.385  < 2e-16 ***
sub_category_x_paper     9.974e-04  8.554e-05   11.660  < 2e-16 ***
sub_category_x_phones    1.757e-04  8.811e-05    1.995  0.04610 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003766 on 34709 degrees of freedom
Multiple R-squared:  0.5203,    Adjusted R-squared:  0.5201
F-statistic:  2510 on 15 and 34709 DF,  p-value: < 2.2e-16
```

Elimination Model-Residual Plots

- We could observe points are scattered now and slight normality.
- Here Elimination model , with many predictor variables, the adjusted R2 = 0.5201, meaning that “52.01% of the variance in the measure of profit can be predicted by significant x variables.





Feature Selection Stepwise Both Model

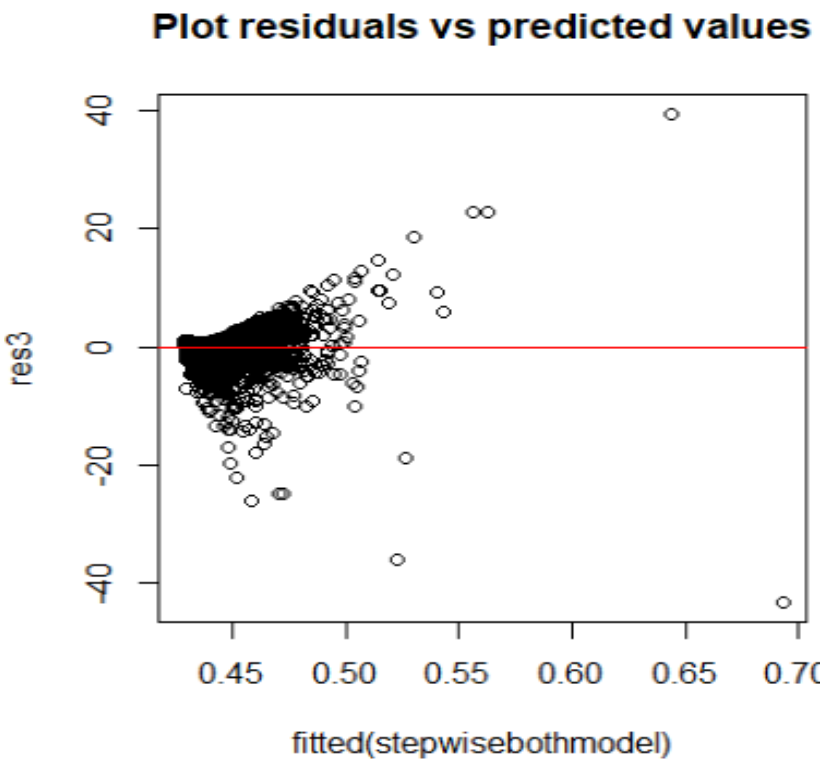
- Building both feature selection forward and backward model with direction = “both.
- Cross checking for multicollinearity, as full model built by removing collinearity variable there is no issue in feature selection model.

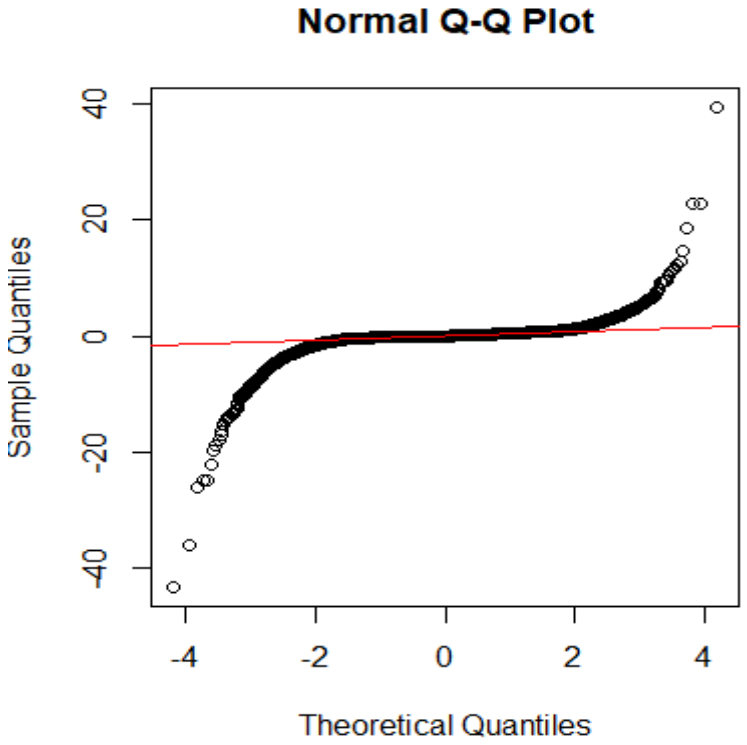
```
> # Checking for multicollinearity one more time
> vifs2=vif(stepwisebothmodel)
> vifs2
                    sales                  discount             shipping_cost1
                 2.650092                  1.033561                  2.620002
            market_x_emea                market_x_eu                market_x_us
                 1.105151                  1.155905                  1.171308
           region_x_south sub_category_x_appliances       sub_category_x_art
                 1.030969                  1.155485                  1.417663
  sub_category_x_binders   sub_category_x_bookcases    sub_category_x_copiers
                 1.486919                  1.207559                  1.207198
sub_category_x_envelopes  sub_category_x_fasteners sub_category_x_furnishings
                 1.211044                  1.216511                  1.266205
   sub_category_x_labels    sub_category_x_machines       sub_category_x_paper
                 1.231130                  1.133368                  1.315354
   sub_category_x_phones     sub_category_x_storage
                 1.271659                  1.386291
>
```

```
> ###########################################################################
> # Feature Selection - Stepwise Both Forward and Backward Model
> ###########################################################################
> # Buidling stepwise both model
> vifmodelvar <- names(vifs)
> # Fetching subset of useful variables from train data set
> train1.data=subset(train.data ,select= vifmodelvar)
> train1.data$profit = train.data$profit
> fullmdl <- lm(profit~., data=train1.data)
> stepwisebothmodel = step(fullmdl, direction="both", trace=F)
> summary(stepwisebothmodel) # Adj-R2 0.3205

Call:
lm(formula = profit ~ sales + discount + shipping_cost1 + market_x_emea +
    market_x_eu + market_x_us + region_x_south + sub_category_x_appliances +
    sub_category_x_art + sub_category_x_binders + sub_category_x_bookcases +
    sub_category_x_copiers + sub_category_x_envelopes + sub_category_x_fasteners +
    sub_category_x_furnishings + sub_category_x_labels + sub_category_x_machines +
    sub_category_x_paper + sub_category_x_phones + sub_category_x_storage,
    data = train1.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37360 -0.00193 -0.00032  0.00246  0.35660

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)               0.4403719  0.0001352 3256.502  < 2e-16 ***
sales                     0.2613931  0.0037391   69.907  < 2e-16 ***
discount                 -0.0129062  0.0002017  -63.975  < 2e-16 ***
region_x_south           -0.0004631  0.0001499   -3.090 0.002003 **
sub_category_x_appliances  0.0006973  0.0002916    2.391 0.016800 *
sub_category_x_art         0.0015266  0.0002004    7.619 2.62e-14 ***
sub_category_x_binders     0.0023573  0.0001867   12.629  < 2e-16 ***
sub_category_x_bookcases  -0.0006564  0.0002584   -2.541 0.011066 *
sub_category_x_copiers     0.0016669  0.0002684    6.210 5.36e-10 ***
sub_category_x_envelopes   0.0017007  0.0002579    6.593 4.36e-11 ***
sub_category_x_fasteners   0.0017579  0.0002587    6.794 1.11e-11 ***
sub_category_x_furnishings 0.0016023  0.0002309    6.939 4.02e-12 ***
sub_category_x_labels      0.0016000  0.0002519    6.352 2.15e-10 ***
sub_category_x_machines   -0.0010469  0.0003156   -3.317 0.000912 ***
sub_category_x_paper       0.0016617  0.0002252    7.378 1.64e-13 ***
sub_category_x_phones      0.0005637  0.0002251    2.504 0.012281 *
sub_category_x_storage     0.0007555  0.0001963    3.850 0.000118 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009409 on 35882 degrees of freedom
Multiple R-squared:  0.3209,    Adjusted R-squared:  0.3205
F-statistic: 847.6 on 20 and 35882 DF,  p-value: < 2.2e-16
```

Feature Selection Stepwise Both Model

- Goodness of Fit: F-Test rejects null hypothesis and accepts that at least one of the x variable useful in predicting profit.
- Residual Analysis: There is no constant variance, linearity and normality.



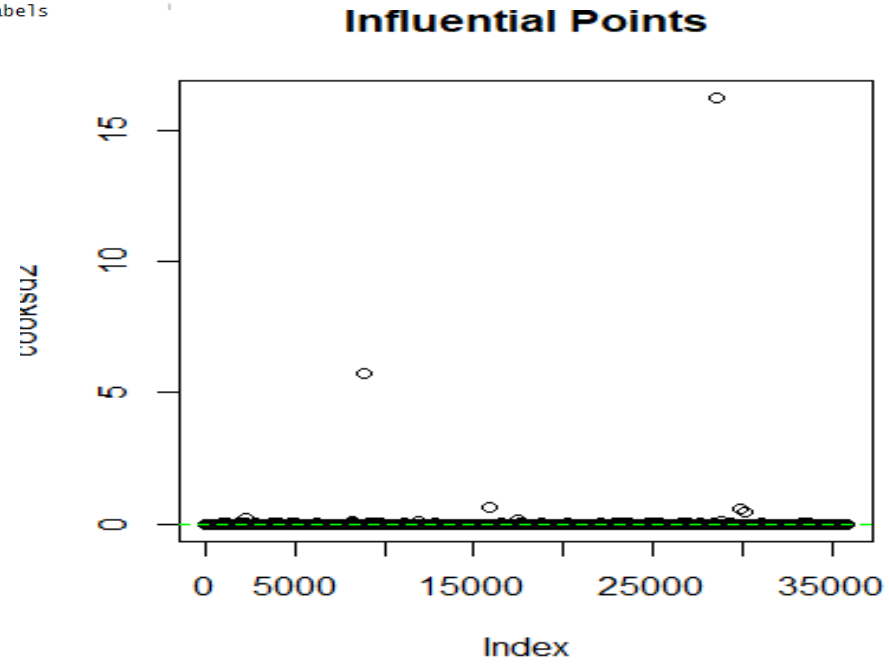


Feature Selection Stepwise Both Model

- Influential Points: Could observe there are potential outliers in plot with cook's distance method.

```
> # Rebuilding final stepwise both model
> fullm12 <- lm(profit~., data=newtrain.data2)
> stepwisebothmodel14 = step(fullm12, direction="both", trace=F)
> summary(stepwisebothmodel14) # Adj-R2 0.5172

Call:
lm(formula = profit ~ sales + discount + shipping_cost1 + market_x_emea +
    market_x_eu + market_x_us + region_x_central + region_x_central_asia +
    region_x_north_asia + region_x_oceania + sub_category_x_appliances +
    sub_category_x_art + sub_category_x_binders + sub_category_x_bookcases +
    sub_category_x_chairs + sub_category_x_copiers + sub_category_x_envelopes +
    sub_category_x_fasteners + sub_category_x_furnishings + sub_category_x_labels
    sub_category_x_paper, data = newtrain.data2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0304544 -0.0011824 -0.0001819  0.0014807  0.0219070

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)               4.404e-01  5.226e-05 8426.345  < 2e-16 ***
sales                     2.410e-01  2.752e-03   87.563  < 2e-16 ***
discount                 -7.390e-03  8.308e-05  -88.961  < 2e-16 ***
shipping_cost1            3.108e-03  9.460e-04    3.285 0.001019 **
market_x_emea             2.732e-04  7.239e-05    3.774 0.000161 ***
market_x_eu              -3.442e-04  6.169e-05   -5.581 2.41e-08 ***
market_x_us               2.012e-04  5.704e-05    3.528 0.000419 ***
region_x_central          1.301e-04  5.511e-05    2.361 0.018219 *
region_x_central_asia     2.449e-04  1.069e-04    2.291 0.021943 *
region_x_north_asia       3.209e-04  1.015e-04    3.162 0.001571 **
region_x_oceania         -2.301e-04  8.538e-05   -2.695 0.007043 **
sub_category_x_appliances 6.658e-04  1.192e-04    5.584 2.37e-08 ***
sub_category_x_art        6.312e-04  7.425e-05    8.500  < 2e-16 ***
sub_category_x_binders    1.186e-03  6.912e-05   17.160  < 2e-16 ***
sub_category_x_bookcases -7.766e-04  1.038e-04   -7.479 7.64e-14 ***
sub_category_x_chairs    -3.983e-04  8.581e-05   -4.642 3.47e-06 ***
sub_category_x_copiers    3.699e-04  1.099e-04    3.366 0.000763 ***
sub_category_x_envelopes  7.533e-04  9.843e-05    7.653 2.01e-14 ***
sub_category_x_fasteners  7.656e-04  9.900e-05    7.733 1.08e-14 ***
sub_category_x_furnishings 6.051e-04 8.715e-05    6.943 3.90e-12 ***
sub_category_x_labels     7.302e-04  9.626e-05    7.586 3.39e-14 ***
sub_category_x_paper      8.680e-04  8.498e-05   10.214  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003715 on 34652 degrees of freedom
Multiple R-squared:  0.5174,    Adjusted R-squared:  0.5172
F-statistic:  1769 on 21 and 34652 DF,  p-value: < 2.2e-16
```

```
> # 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] 34674
```



Feature Selection Stepwise Both Model

- Rebuilt the stepwise both model by removing influential points from data set.

```
> # Rebuilding final stepwise both model
> fullml2 <- lm(profit~., data=newtrain.data2)
> stepwisebothmodel14 = step(fullml2, direction="both", trace=F)
> summary(stepwisebothmodel14) # Adj-R2 0.5172

Call:
lm(formula = profit ~ sales + discount + shipping_cost1 + market_x_emea +
    market_x_eu + market_x_us + region_x_central + region_x_central_asia +
    region_x_north_asia + region_x_oceania + sub_category_x_appliances +
    sub_category_x_art + sub_category_x_binders + sub_category_x_bookcases +
    sub_category_x_chairs + sub_category_x_copiers + sub_category_x_envelopes +
    sub_category_x_fasteners + sub_category_x_furnishings + sub_category_x_labels
    sub_category_x_paper, data = newtrain.data2)

Residuals:
      Min         1Q     Median         3Q        Max
-0.0304544 -0.0011824 -0.0001819  0.0014807  0.0219070

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              4.404e-01  5.226e-05 8426.345  < 2e-16 ***
sales                    2.410e-01  2.752e-03   87.563  < 2e-16 ***
discount                -7.390e-03  8.308e-05  -88.961  < 2e-16 ***
shipping_cost1           3.108e-03  9.460e-04    3.285 0.001019 **
```

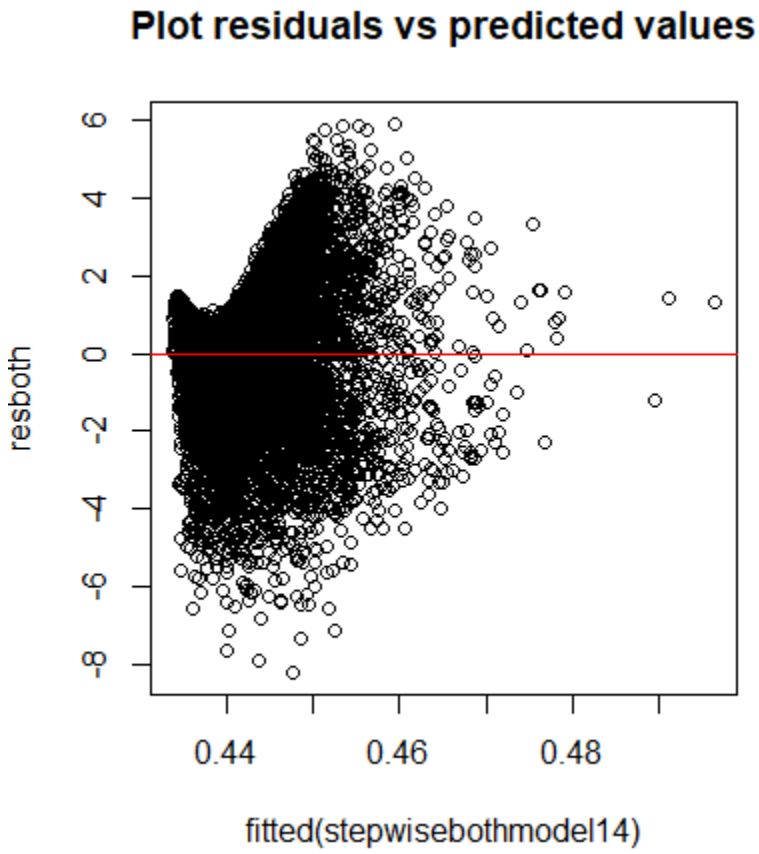
Feature Selection Stepwise Both Model

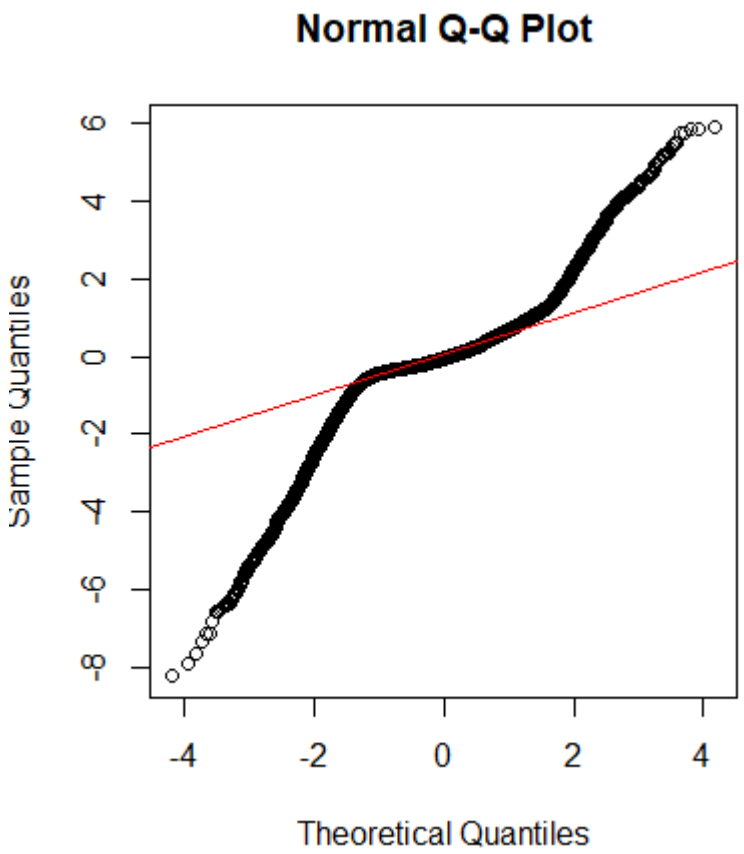
- Rebuilt the stepwise both model contd..

```
market_x_emea                2.732e-04  7.239e-05    3.774 0.000161 ***
market_x_eu                 -3.442e-04  6.169e-05   -5.581 2.41e-08 ***
market_x_us                  2.012e-04  5.704e-05    3.528 0.000419 ***
region_x_central             1.301e-04  5.511e-05    2.361 0.018219 *
region_x_central_asia        2.449e-04  1.069e-04    2.291 0.021943 *
region_x_north_asia          3.209e-04  1.015e-04    3.162 0.001571 **
region_x_oceania            -2.301e-04  8.538e-05   -2.695 0.007043 **
sub_category_x_appliances    6.658e-04  1.192e-04    5.584 2.37e-08 ***
sub_category_x_art           6.312e-04  7.425e-05    8.500  < 2e-16 ***
sub_category_x_binders       1.186e-03  6.912e-05   17.160  < 2e-16 ***
sub_category_x_bookcases    -7.766e-04  1.038e-04   -7.479 7.64e-14 ***
sub_category_x_chairs       -3.983e-04  8.581e-05   -4.642 3.47e-06 ***
sub_category_x_copiers       3.699e-04  1.099e-04    3.366 0.000763 ***
sub_category_x_envelopes     7.533e-04  9.843e-05    7.653 2.01e-14 ***
sub_category_x_fasteners     7.656e-04  9.900e-05    7.733 1.08e-14 ***
sub_category_x_furnishings   6.051e-04  8.715e-05    6.943 3.90e-12 ***
sub_category_x_labels        7.302e-04  9.626e-05    7.586 3.39e-14 ***
sub_category_x_paper         8.680e-04  8.498e-05   10.214  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003715 on 34652 degrees of freedom
Multiple R-squared:  0.5174,    Adjusted R-squared:  0.5172
F-statistic:  1769 on 21 and 34652 DF,  p-value: < 2.2e-16
```

Feature Selection Stepwise Both Model

- Residual Analysis: Points are scattered and there is slight normality.





Evaluation and Accuracy of Models

- RMSE calculation:

```
> ##################################################
> # RMSE Calculation on train.data
> ##################################################
> y1 = predict.glm(eliminationmodel4, train.data)
> y2 = predict.glm(stepwisebothmodel14, train.data)
> y = train.data$profit
> rmse_1 = sqrt(((y-y1)%*%(y-y1)) /nrow(train.data))
> rmse_1
             [,1]
[1,] 0.009526671
> rmse_2 = sqrt(((y-y2)%*%(y-y2)) /nrow(train.data))
> rmse_2
             [,1]
[1,] 0.009529029
```

```
> ##################################################
> # RMSE Calculation on test.data
> ##################################################
> x1 = predict.glm(eliminationmodel4, test.data)
> x2 = predict.glm(stepwisebothmodel14, test.data)
> x = test.data$profit
> rmse_1 = sqrt(((x-x1)%*%(x-x1)) /nrow(test.data))
> rmse_1
             [,1]
[1,] 0.01014852
> rmse_2 = sqrt(((x-x2)%*%(x-x2)) /nrow(test.data))
> rmse_2
             [,1]
[1,] 0.01016472
```

Regression Model	ADJ-R2	ROOT MEAN SQUARE ERROR	
		Train	Test
Elimination Model	0.5201	0.009526671	0.01014852
Stepwise Both Model	0.5172	0.009529029	0.01016472

Evaluation and Accuracy of Models

- The Elimination mode , with many predictor variables, the adjusted R2 = 0.5201, meaning that “52.01% 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 = 0.5172, meaning that “51.72% of the variance in the measure of profit can be predicted by statistically significant x variables.
- **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 best reduced fit model is,

```
y = profit = 4.404e-01 + (2.488e-01*sales) + (-7.439e-03*discount) + (2.067e-04*market_x_emea)
           + (-3.574e-04*sub_category_x_appliances) + (7.236e-04*sub_category_x_art)
           + (1.299e-03*sub_category_x_binders) + (-6.782e-04*sub_category_x_bookcases)
           + (5.074e-04*sub_category_x_copiers) + (8.346e-04*sub_category_x_envelopes)
           + (8.415e-04*sub_category_x_fasteners) + (7.203e-04*sub_category_x_furnishings)
           + (8.163e-04*sub_category_x_labels) + (9.974e-04*sub_category_x_paper)
           + (1.757e-04*sub_category_x_phones)
```

Regularization

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

```
> ##############################################################################
> # Regularization
> ##############################################################################
> 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    47
[1] 15387    47
```

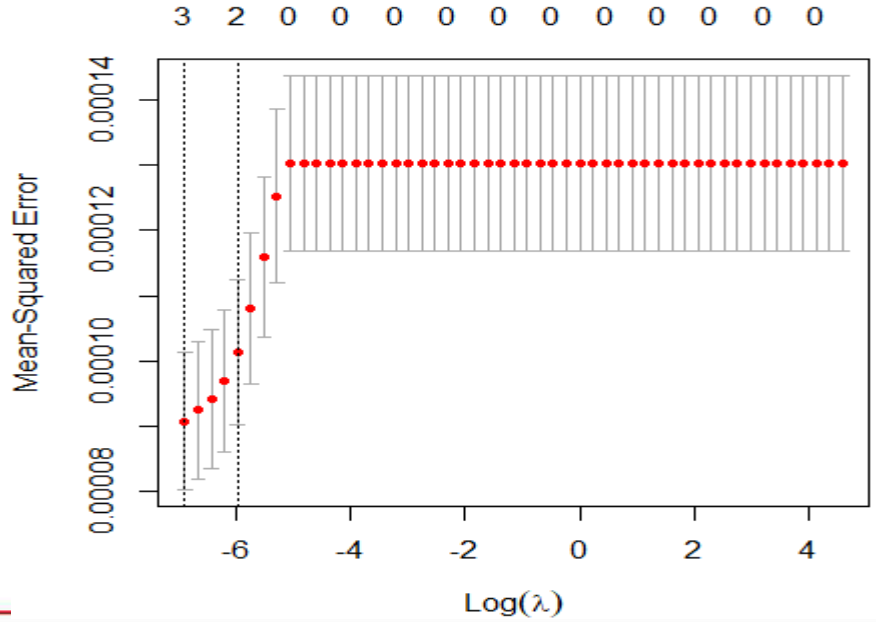
- Created custom function loss function to derive RMSE and R-Square value

```
> # 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
+   )
+ }
```

Lasso Regression Model

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

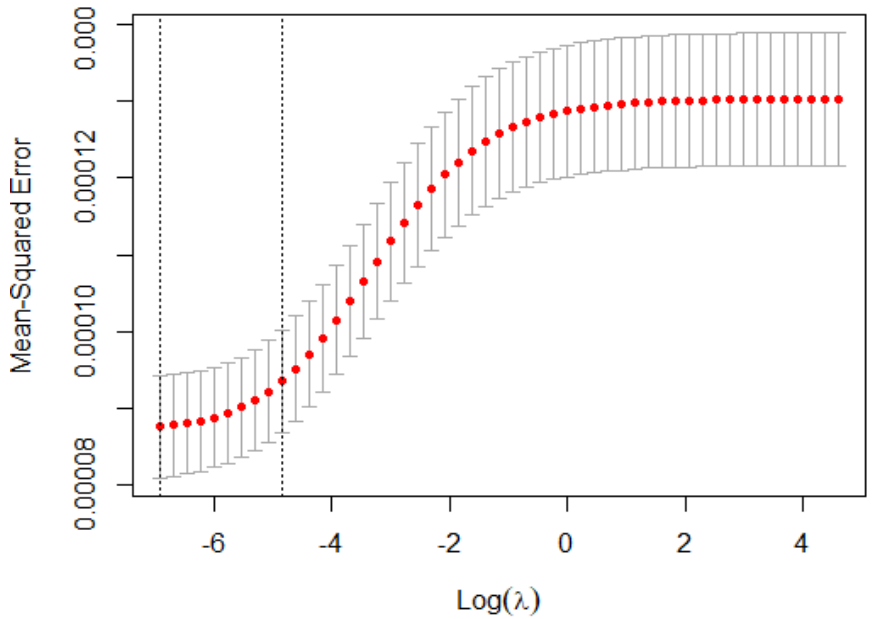
```
> ##############################################################################
> # Lasso Regression
> ##############################################################################
> grid <- 10^seq(2, -3, by = -.1)
> #lambdas <-10^seq(10, -2, length = 100)
> # Setting alpha = 1 implies lasso penalty
> 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] 0.001
> lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = lambda_best, standardize = TRUE)
> losso.coef = predict(lasso_model, s=lambda_best, type="coefficients")[1:48, ]
> losso.coef[losso.coef !=0]
          (Intercept)                 sales              discount sub_category_x_tables
          0.441285749           0.201969056          -0.009103442          -0.002058618
```



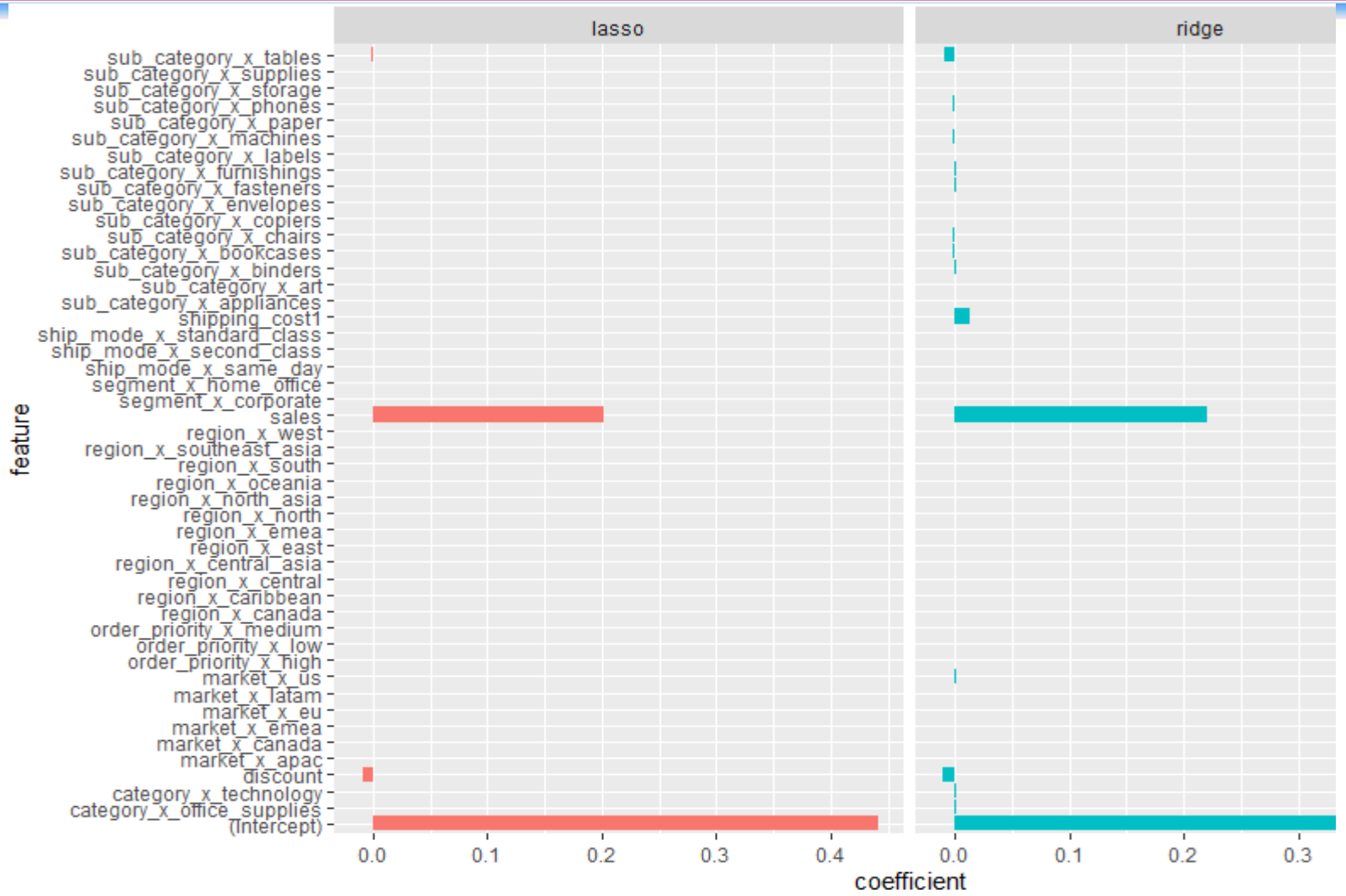
Ridge Regression Model

- 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] 0.001
```



Lasso Ridge Regression Model



Ridge and Lasso Models

- Prediction and evaluation using best lambda on train and test data from Ridge and lasso.

Ridge

```
> # 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 0.009333223 0.3313084
> # 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 0.009954082 0.3228671
```

Lasso

```
> # 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 0.009520469 0.3042082
> #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 0.01016539 0.2938129
```

Experiments Results

- Modeled Ridge and Lasso to compare with multiple linear regression model by comparing RMSE and R-square as there was overfitting chance.
- The Lasso model built by shrinking many features to zero with RMSE of 0.0106 on test data.
- The Ridge model on test data is providing RMSE of 0.009954 which is almost same as multiple linear regression model.
- This shows there is no overfitting issue and ridge model is the best model.

Regression Model	R-Square		Root Mean Square Error	
	Train	Test	Train	Test
Multiple linear Regression Model	0.5201		0.009526671	0.01014852
Lasso Regression Model	0.3042082	0.2938129	0.009520469	0.01016539
Ridge Regression Model	0.3313084	0.3228671	0.009333223	0.009954082

Conclusion & Future Work

- The global superstore have statistically significant difference in sales with respect to different groups of market regions.
- The global superstore have statistically significant difference in sales with respect to different groups of different market groups.
- The multiple linear regression model is statistically significant in predicting profit of global super store with respect to region-wise, sales-wise, product subcategory wise etc.
- The Ridge regression model is the best model with least RMSE out of multiple linear regression model and Lasso model.

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 details in the season wise.