Final Group Project

Team 5

2024-04-28

Introduction

Global Super Store is a data set which has around 50000 values. Its a customer centric data set , which has the data of all the orders that have been placed through different vendors and markets , starting from the year 2011 till 2014.

It provides data regarding profit gained over sale of various types of products, and can be used by a company planning to launch a certain type of product in a market.

Link to Project Github - https://github.com/Ayush-Srillex/DataScienceR

Data Columns are self-describing through their name, there data type are as follows:

Table 1: Column Type

Column	Type
Row ID	ID
Order ID	ID
Order Date	Date
Ship Date	Date
Ship Mode	Categorical
Customer ID	ID
Customer Name	Text
Segment	Categorical
City	Categorical
State	Categorical
Country	Categorical
Postal Code	Text
Market	Categorical
Region	Categorical
Product ID	ID
Category	Categorical
Sub-Category	Categorical
Product Name	Text
Sales	Numeric
Quantity	Numeric
Discount	Numeric
Profit	Numeric
Shipping Cost	Numeric
Order Priority	Categorical

Research question

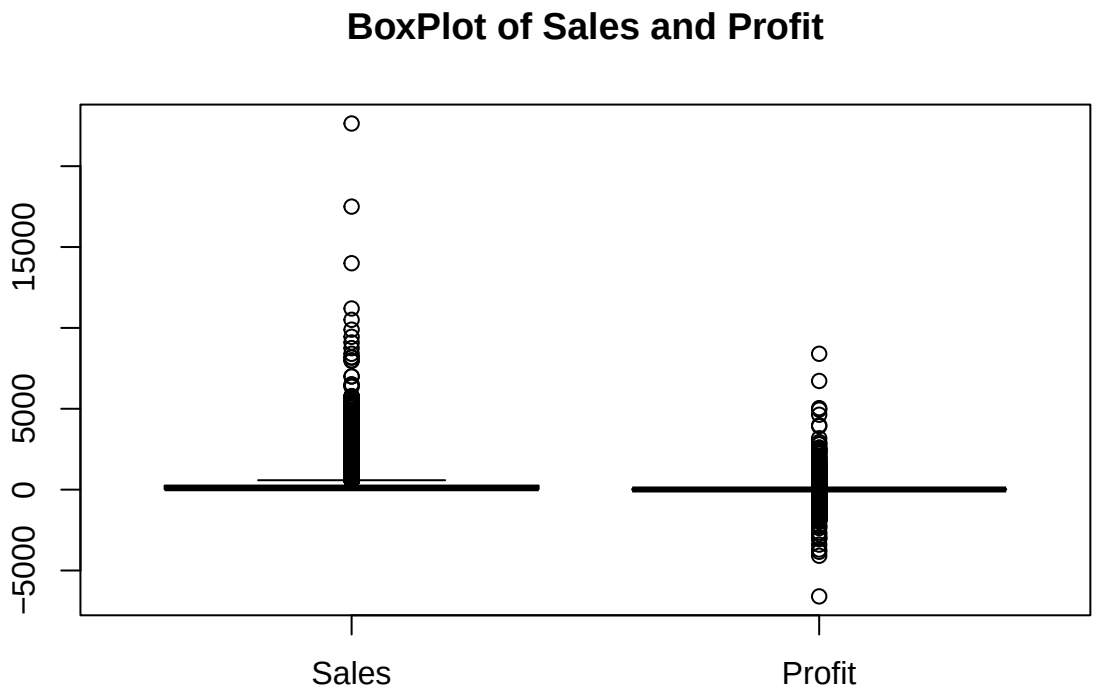
Our research aims to provide a base for companies manufacturing and selling various kinds of products which are within our scope of data. We wish to answer the following common questions any company would have regarding their new product:

1. If company manufactures a certain category of product, which sub category should it target to generate high profits?
2. Which market should it be choosing on priority for product launch?
3. What discount should it offer its consumers moving forward to retain its consumer base and increase sales and hence profits
4. If a company already operates in a certain market and wants to expand itself to different parts of world, how should it move forward? How accurate is the model developed in their own market will be in different market?

We don't aim to provide accurate answer to these questions as these vary with different companies. We only wish to provide a strong base for this analysis through extensive EDA and model development.

#Data Preparation and Cleaning

```
superstore <- read_excel("Global Data Superstore.xls",sheet = "Orders")

#Check Boxplots for Sales and Profit Variables:
boxplot(superstore[c(19,22)],main="BoxPlot of Sales and Profit")
```



```r
#try out different outlier percentage removal
superstore$'Sub-Category' <- as.factor(superstore$'Sub-Category')
superstore$Profit <- as.numeric(superstore$Profit)

# Define the thresholds for outliers
lower_5 <- quantile(superstore$Profit, 0.05)
upper_5 <- quantile(superstore$Profit, 0.95)
lower_10 <- quantile(superstore$Profit, 0.10)
upper_10 <- quantile(superstore$Profit, 0.90)
lower_15 <- quantile(superstore$Profit, 0.15)
upper_15 <- quantile(superstore$Profit, 0.85)

# Filter superstore to remove outliers
superstore_5 <- superstore %>% filter(Profit > lower_5 & Profit < upper_5)
superstore_10 <- superstore %>% filter(Profit > lower_10 & Profit < upper_10)
superstore_15 <- superstore %>% filter(Profit > lower_15 & Profit < upper_15)

# Count how many entries remain in each category
count_5 <- table(superstore_5$'Sub-Category')
count_10 <- table(superstore_10$'Sub-Category')
count_15 <- table(superstore_15$'Sub-Category')
original_count <- table(superstore$'Sub-Category')

comparison<-rbind(original_count,count_5,count_10,count_15)
rownames(comparison)<-c("Original Count","5% Outlier","10% Outlier","15% Outlier")


knitr::kable(
t(comparison),
caption = "Comparison of datapoints removed in Each category with % of outlier removed"
)
```

Table 2: Comparison of datapoints removed in Each category with % of outlier removed

	Original Count	5% Outlier	10% Outlier	15% Outlier
Accessories	3075	2775	2378	1990
Appliances	1755	1319	1057	842
Art	4883	4832	4620	4212
Binders	6152	6027	5792	5394
Bookcases	2411	1693	1149	800
Chairs	3434	2840	2211	1677
Copiers	2223	1497	997	709
Envelopes	2435	2421	2284	2072
Fasteners	2420	2419	2381	2265
Furnishings	3170	3056	2816	2458
Labels	2606	2602	2589	2517
Machines	1486	1094	823	628
Paper	3538	3520	3373	3087
Phones	3357	2708	2127	1665
Storage	5059	4610	3991	3450
Supplies	2425	2377	2216	1994

	Original Count	5% Outlier	10% Outlier	15% Outlier
Tables	861	369	225	139

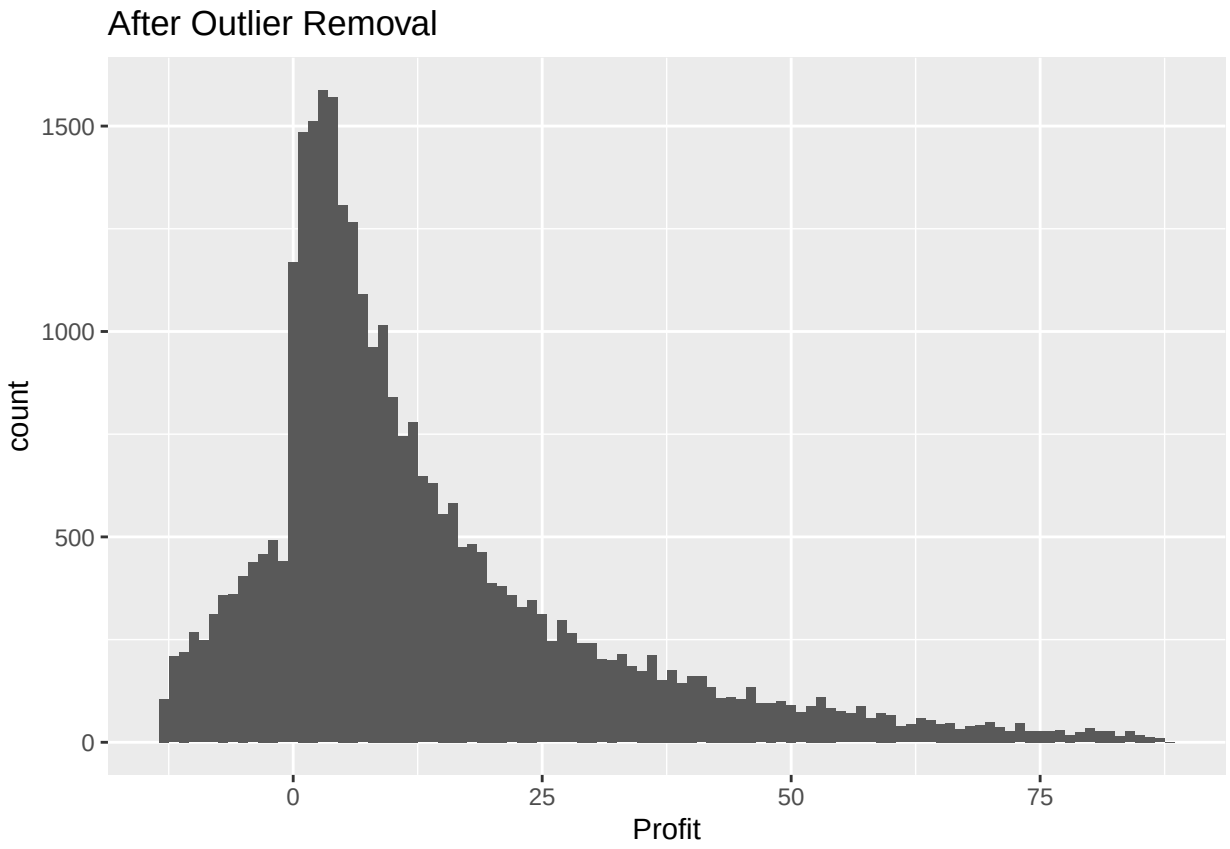
1. Subcategories with the most drop: Bookcases, Chairs, Copiers, and Tables. It indicates that these subcategories have an outstanding or low profit values which was filtered. Bookcases and Tables illustrates huge reductions, and therefore require a deeper analysis on market served.

2. Subcategories with the minimal drop: Art, Envelopes, Labels, Fasteners, and Paper. It suggests that there are fewer extreme profit outliers and that sales in these categories are more regularly within the median profit range. Fasteners and Labels were not affected by the outliers at all. It implies that their profit is stable. Summary:

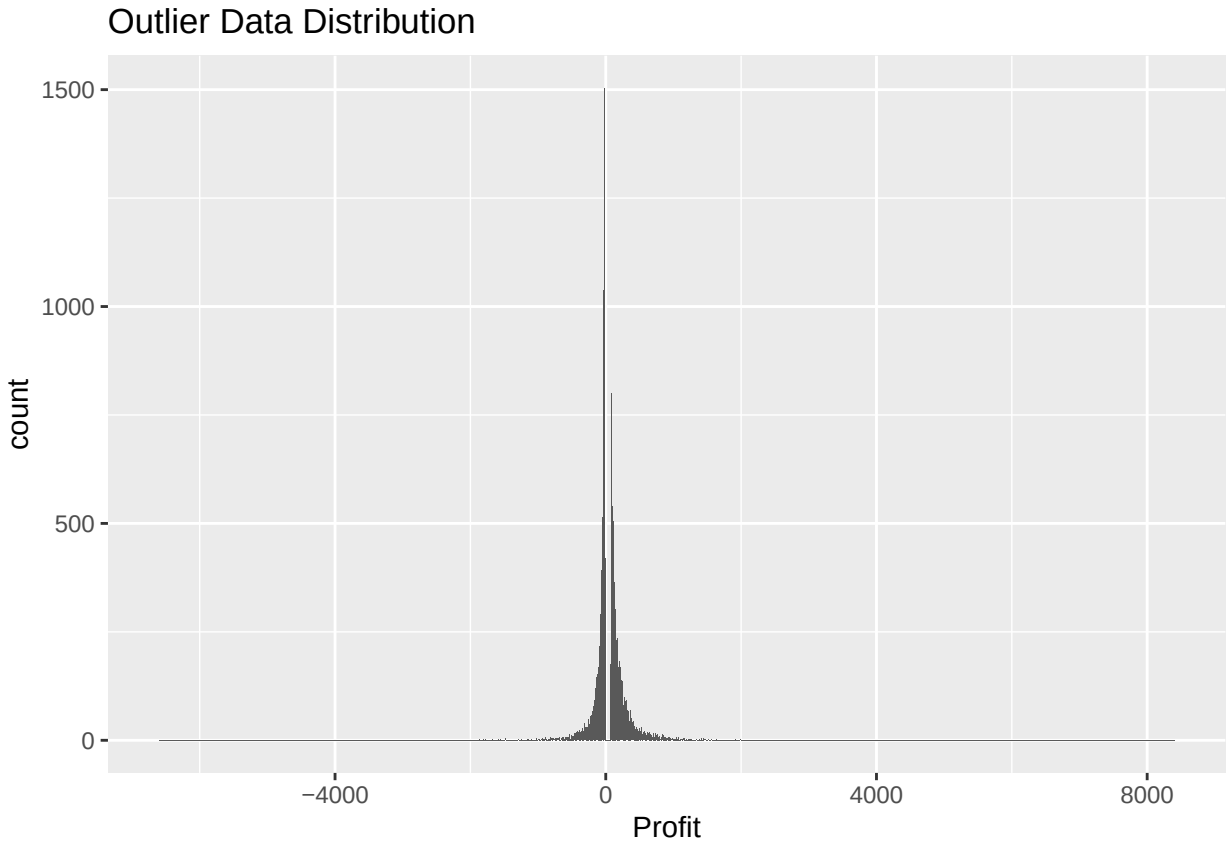
a. Office category shows a stability and could be consistent sources of investment or growth, as well as dependable contributors to total profitability.
b. Review high drop subcategories by other affected factors such as market, discounts, and delivery.

Exploratory Data Analysis (EDA)

```
df<-superstore[ superstore$Profit > quantile(superstore$Profit , 0.15 ) , ]
df<-df[ df$Profit < quantile(df$Profit , 0.85 ) , ]

df<-df[df$Sales<quantile(df$Sales,0.85) ,]

outlierData<-superstore[ superstore$Profit < quantile(superstore$Profit , 0.15 ) | superstore$Profit > q

ggplot(data=df)+
  geom_histogram(mapping = aes(x= Profit),binwidth = 1)+labs(title = "After Outlier Removal")
```



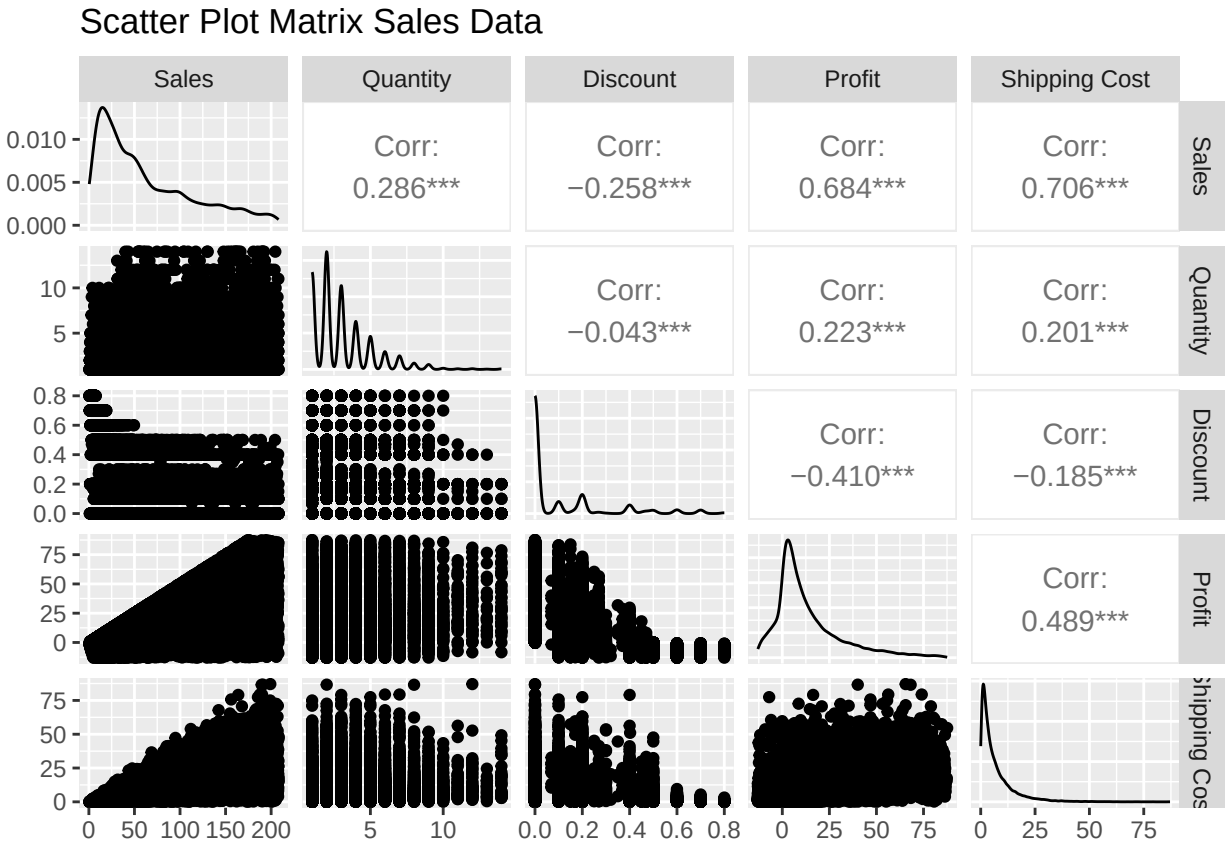

```
ggplot(data=outlierData)+
  geom_histogram(mapping = aes(x= Profit),binwidth = 10)+labs(title = "Outlier Data Distribution")
```



Profit is more evenly distributed. We will not completely neglect the cutoff points, we will deal with separately as to why some products have high profits and high loss.

We now move to visualise all numeric variables in our dataset: Sales, Quantity, Discount, Profit and Shipping Cost. We study the relationship between all the combinations of the variables as well.

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```



For our focus variable Profit, we see that it is positively correlated with Sales, Quantity and Shipping Cost, but negatively correlated with Discount. This makes sense with the common knowledge of commerce and finance.

We move to feature selection

Feature Selection

Before building models, we identify the most relevant features for predicting our target variable.

We start with identifying most important categorical variables. We carry out the similar analysis as before in EDA analysis but with some categorical variables and Profit as primary target variable. We will primarily use linear modeling to identify correlation with Profit. We will carry out test of Profit with Ship Mode, Segment, Market, Country, Sub-Category, Order Priority.

```
library(broom)
models <- list()
cat_variables=c("`Ship Mode`","Segment", "Country", "Market", "`Sub-Category`", "`Order Priority`")
table_list <- list()
i=1
for(c in cat_variables) {
  formula <- as.formula(paste("Profit", "~", paste(c,"-1")))
  reg<-lm(formula,data=df)

  coefs <- tidy(reg)
  a<-coefs[order(coefs$estimate, decreasing = TRUE),]
```

```
  print(knitr::kable(
  head(a),
  caption = c
  ))
  i=i+1
}
```

```
##
##
## Table: ‘Ship Mode‘
##
## |term                       | estimate| std.error| statistic| p.value|
## |:--------------------------|--------:|---------:|---------:|-------:|
## |‘Ship Mode‘First Class     | 13.41009| 0.2594348|  51.68966|       0|
## |‘Ship Mode‘Standard Class  | 13.39333| 0.1283639| 104.33875|       0|
## |‘Ship Mode‘Second Class    | 13.26988| 0.2207273|  60.11890|       0|
## |‘Ship Mode‘Same Day        | 13.09881| 0.4296768|  30.48526|       0|
##
##
## Table: Segment
##
## |term                | estimate| std.error| statistic| p.value|
## |:-------------------|--------:|---------:|---------:|-------:|
## |SegmentHome Office  | 13.40822| 0.2323818|  57.69911|       0|
## |SegmentCorporate    | 13.35007| 0.1811724|  73.68712|       0|
## |SegmentConsumer     | 13.33928| 0.1380052|  96.65776|       0|
##
##
## Table: Country
##
## |term               | estimate| std.error| statistic|   p.value|
## |:------------------|--------:|---------:|---------:|---------:|
## |CountrySwaziland   | 43.38000| 11.711666|  3.703999| 0.0002126|
## |CountryEritrea     | 38.10000| 11.711666|  3.253167| 0.0011425|
## |CountrySlovenia    | 35.37000| 11.711666|  3.020066| 0.0025292|
## |CountryLesotho     | 32.06250|  8.281398|  3.871629| 0.0001083|
## |CountryGuadeloupe  | 29.07500|  8.281398|  3.510880| 0.0004473|
## |CountryParaguay    | 28.61778|  5.520932|  5.183505| 0.0000002|
##
##
## Table: Market
##
## |term          | estimate| std.error| statistic| p.value|
## |:-------------|--------:|---------:|---------:|-------:|
## |MarketEU      | 16.30648| 0.2322729|  70.20400|       0|
## |MarketCanada  | 14.46695| 1.0101575|  14.32148|       0|
## |MarketAPAC    | 13.50755| 0.2280789|  59.22314|       0|
## |MarketAfrica  | 12.76615| 0.3204495|  39.83825|       0|
## |MarketLATAM   | 12.68094| 0.2172714|  58.36452|       0|
## |MarketUS      | 12.42069| 0.2115788|  58.70479|       0|
##
##
## Table: ‘Sub-Category‘
```

```
##
## |term                       | estimate| std.error| statistic| p.value|
## |:--------------------------|--------:|---------:|---------:|-------:|
## |`Sub-Category`Copiers      | 33.42272| 1.1140837|  30.00019|       0|
## |`Sub-Category`Bookcases    | 28.51385| 0.9643008|  29.56945|       0|
## |`Sub-Category`Appliances   | 23.36115| 0.6747579|  34.62153|       0|
## |`Sub-Category`Machines     | 21.66272| 0.9217478|  23.50178|       0|
## |`Sub-Category`Phones       | 21.52848| 0.5369874|  40.09121|       0|
## |`Sub-Category`Chairs       | 21.49838| 0.5216677|  41.21087|       0|
##
##
## Table: `Order Priority`
##
## |term                       | estimate| std.error| statistic| p.value|
## |:--------------------------|--------:|---------:|---------:|-------:|
## |`Order Priority`Critical   | 13.63760| 0.3562714|  38.27869|       0|
## |`Order Priority`Medium     | 13.38136| 0.1310332| 102.12192|       0|
## |`Order Priority`High       | 13.28767| 0.1807418|  73.51740|       0|
## |`Order Priority`Low        | 12.99978| 0.4583847|  28.35998|       0|
```

We see that for categorical variables - Ship Mode, Segment, Order Priority have almost the same estimate for each of its categories. This means the Profit doesn't depend on these categories, each category would give the same Profit. We can run a t test to confirm the hypothesis that these estimates are not statistically significant.

For other categorical variables - Country, Market and Product Sub Category, there is difference in estimate for each category. The estimates are given in decreasing order to identify the categories with most positive impact on Profit. For example:

- In market variable- EU Market comes out to be most profitable one, and EMEA is the least profitable.
- In Country variable, we see that countries like Yemen and UAE as actually giving loss for the products sold there.

Other inferences can be drawn as per reader's wish.

Moving on to identifying most important numeric variables against Profit.

```
cat_variables=c("Sales","Quantity","Discount","Shipping Cost","Profit")
df_train<-df[,cat_variables]
regressor <- lm(Profit ~ . , data = df_train)
relImportance <- calc.relimp(regressor, type = "lmg", rela = TRUE)
sort(relImportance$lmg, decreasing=TRUE)
```

```
##         Sales Shipping Cost      Discount      Quantity
##    0.57238081    0.19717086    0.19408118    0.03636715
```

We see that Sales is the most important feature for Profit.

With the above analysis, we identify the following most important features:

- Categorical Variables: Market, Product Sub-Category
- Numeric variables: Sales, Shipping Cost, Discount

Model Building

We divide the dataset in 70% train test split for model testing:

```
train_indices <- createDataPartition(df$Profit, p = 0.7, list = FALSE)
train_data <- df[train_indices, ]
test_data <- df[-train_indices, ]
```

Our preliminary model would be taking all the important variables: Model 0

```
reg0<-lm(Profit~Market+`Sub-Category`+Sales+`Shipping Cost`+Discount,data=train_data)
summary(reg0)
```

```
##
## Call:
## lm(formula = Profit ~ Market + `Sub-Category` + Sales + `Shipping Cost` +
##     Discount, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.585  -5.046  -0.253   5.240  48.165
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.044956   0.464725   0.097 0.922937
## MarketAPAC                 -1.878317   0.323641  -5.804 6.58e-09 ***
## MarketCanada               -1.945319   0.876113  -2.220 0.026402 *
## MarketEMEA                  0.466241   0.366064   1.274 0.202797
## MarketEU                   -1.484576   0.323353  -4.591 4.43e-06 ***
## MarketLATAM                -0.747139   0.315662  -2.367 0.017947 *
## MarketUS                    2.766359   0.318862   8.676  < 2e-16 ***
## `Sub-Category`Appliances    2.429794   0.666522   3.645 0.000268 ***
## `Sub-Category`Art           1.034160   0.420086   2.462 0.013832 *
## `Sub-Category`Binders       3.396782   0.412175   8.241  < 2e-16 ***
## `Sub-Category`Bookcases    -2.545371   0.900372  -2.827 0.004703 **
## `Sub-Category`Chairs       -2.580623   0.565932  -4.560 5.14e-06 ***
## `Sub-Category`Copiers       1.057421   1.002702   1.055 0.291633
## `Sub-Category`Envelopes     2.844043   0.475631   5.980 2.27e-09 ***
## `Sub-Category`Fasteners     2.264698   0.474557   4.772 1.83e-06 ***
## `Sub-Category`Furnishings   0.937386   0.460997   2.033 0.042025 *
## `Sub-Category`Labels        2.987178   0.468295   6.379 1.82e-10 ***
## `Sub-Category`Machines     -1.572818   0.880116  -1.787 0.073942 .
## `Sub-Category`Paper         4.557266   0.438217  10.400  < 2e-16 ***
## `Sub-Category`Phones       -1.883241   0.576674  -3.266 0.001094 **
## `Sub-Category`Storage      -2.299810   0.442678  -5.195 2.06e-07 ***
## `Sub-Category`Supplies      0.812403   0.485994   1.672 0.094612 .
## `Sub-Category`Tables      -20.079376   2.994035  -6.706 2.04e-11 ***
## Sales                       0.226652   0.002447  92.615  < 2e-16 ***
## `Shipping Cost`             0.035678   0.014089   2.532 0.011336 *
## Discount                  -24.161735   0.474446 -50.926  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11.88 on 22024 degrees of freedom
## Multiple R-squared:  0.5431, Adjusted R-squared:  0.5426
## F-statistic:  1047 on 25 and 22024 DF,  p-value: < 2.2e-16
```

```
pred <- predict(reg0, newdata = test_data)

test0<-test_data
test0$predictions<-pred
```

Analysis of Model 0:

- Residuals: Median of residuals is just -0.242 , very close to 0 (ideal residual) and the 1st, 3rd quantile are also small. But the min and max of residuals varies from -55.103 tpo 50.492, which is a very large range. This means the residuals are distributed in a tapered fashion. This can also be seen from the residual standard error : 11.78. If we go on removing outliers from the dataset, this distribution will tend to be more normal.
- Coefficients: We can interpret each coefficient by comparing with standard 95% p value - 0.05 or by the number of stars in from of the variable. 3 stars indicate high relation of predictor and response variable, while 0 starts indicate very low relation.

Out of the markets, APAC, EU and US are statistically more likely to influence profit than other markets. Similar inference can be made for other variables.

Our next model would be taking the most important categorical and numeric variable. These are - Sub-Category and Sales

Model 1

```
reg1<-lm(Profit~`Sub-Category`+Sales,data=train_data)
summary(reg1)
```

```
##
## Call:
## lm(formula = Profit ~ `Sub-Category` + Sales, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.679  -6.044   0.674   5.907  50.001
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.868892   0.409721  -7.002 2.59e-12 ***
## `Sub-Category`Appliances    2.817897   0.709893   3.969 7.23e-05 ***
## `Sub-Category`Art           0.931712   0.443269   2.102 0.035572 *
## `Sub-Category`Binders       2.171250   0.435699   4.983 6.30e-07 ***
## `Sub-Category`Bookcases    -4.587559   0.957153  -4.793 1.65e-06 ***
## `Sub-Category`Chairs       -4.439466   0.601301  -7.383 1.60e-13 ***
## `Sub-Category`Copiers      -0.352786   1.065792  -0.331 0.740641
## `Sub-Category`Envelopes     2.020259   0.502750   4.018 5.88e-05 ***
## `Sub-Category`Fasteners     1.244347   0.499524   2.491 0.012743 *
## `Sub-Category`Furnishings   0.493894   0.491146   1.006 0.314622
## `Sub-Category`Labels        2.300599   0.492992   4.667 3.08e-06 ***
```

```
## `Sub-Category`Machines     -2.961794   0.935191  -3.167 0.001542 **
## `Sub-Category`Paper         5.251510   0.466213  11.264  < 2e-16 ***
## `Sub-Category`Phones       -2.342592   0.613768  -3.817 0.000136 ***
## `Sub-Category`Storage      -2.983271   0.469247  -6.358 2.09e-10 ***
## `Sub-Category`Supplies      0.149053   0.513788   0.290 0.771738
## `Sub-Category`Tables      -21.445648   3.188943  -6.725 1.80e-11 ***
## Sales                       0.245787   0.001898 129.479  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.67 on 22032 degrees of freedom
## Multiple R-squared:  0.4805, Adjusted R-squared:  0.4801
## F-statistic:  1199 on 17 and 22032 DF,  p-value: < 2.2e-16
```

```
pred <- predict(reg1, newdata = test_data)

test1<-test_data
test1$predictions<-pred
```

Our next model takes Sales, Shipping Cost and Discount only, removing all categorical variables

Model 2

```
reg2<-lm(Profit~Sales+`Shipping Cost`+Discount,data=train_data)
summary(reg2)
```

```
##
## Call:
## lm(formula = Profit ~ Sales + `Shipping Cost` + Discount, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.597  -5.571  -0.141   5.289  47.713
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.795893   0.147191  18.995   <2e-16 ***
## Sales             0.204869   0.002258  90.737   <2e-16 ***
## `Shipping Cost`   0.036389   0.014406   2.526   0.0115 *
## Discount        -24.576259   0.475492 -51.686   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.15 on 22046 degrees of freedom
## Multiple R-squared:  0.5214, Adjusted R-squared:  0.5214
## F-statistic:  8007 on 3 and 22046 DF,  p-value: < 2.2e-16
```

```
pred <- predict(reg2, newdata = test_data)

test2<-test_data
test2$predictions<-pred
```

Our last model takes only Sales and Shipping Cost Variables

Model 3

```
reg3<-lm(Profit~Sales+`Shipping Cost`,data=train_data)
summary(reg3)
```

```
##
## Call:
## lm(formula = Profit ~ Sales + `Shipping Cost`, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.970  -6.180   0.892   5.849  47.755
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.984161   0.135251  -7.277 3.54e-13 ***
## Sales            0.226503   0.002349  96.416  < 2e-16 ***
## `Shipping Cost`  0.036865   0.015253   2.417   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 22047 degrees of freedom
## Multiple R-squared:  0.4634, Adjusted R-squared:  0.4634
## F-statistic:  9521 on 2 and 22047 DF,  p-value: < 2.2e-16
```

```
pred <- predict(reg3, newdata = test_data)

test3<-test_data
test3$predictions<-pred
```

We now move to test our models for their performance

Model Selection

```
model_eval <- data.frame(
  model=character(),
  rms = numeric(),
  r_sq = numeric(),
  adj_r_sq = numeric()
)
#Model 0:
rmse <- sqrt(mean((test0$Profit - test0$predictions)^2))
r_squared <- summary(reg0)$r.squared
adjusted_r_squared <- summary(reg0)$adj.r.squared

model <- data.frame(
  model="Model 0",
  rms = rmse,
  r_sq = r_squared,
  adj_r_sq = adjusted_r_squared
```

```
)

model_eval<-rbind(model_eval,model)

#Model 1:

rmse <- sqrt(mean((test1$Profit - test1$predictions)^2))
r_squared <- summary(reg1)$r.squared
adjusted_r_squared <- summary(reg1)$adj.r.squared

model <- data.frame(
  model="Model 1",
  rms = rmse,
  r_sq = r_squared,
  adj_r_sq = adjusted_r_squared
)

model_eval<-rbind(model_eval,model)

#Model 2:

rmse <- sqrt(mean((test2$Profit - test2$predictions)^2))
r_squared <- summary(reg2)$r.squared
adjusted_r_squared <- summary(reg2)$adj.r.squared

model <- data.frame(
  model="Model 2",
  rms = rmse,
  r_sq = r_squared,
  adj_r_sq = adjusted_r_squared
)

model_eval<-rbind(model_eval,model)

#Model 3

rmse <- sqrt(mean((test3$Profit - test3$predictions)^2))
r_squared <- summary(reg3)$r.squared
adjusted_r_squared <- summary(reg3)$adj.r.squared

model <- data.frame(
  model="Model 3",
  rms = rmse,
  r_sq = r_squared,
  adj_r_sq = adjusted_r_squared
)

model_eval<-rbind(model_eval,model)

knitr::kable(
model_eval,
caption = "Model Evaluation"
)
```

Table 3: Model Evaluation

model	rms	r_sq	adj_r_sq
Model 0	11.77375	0.5431495	0.5426309
Model 1	12.57559	0.4804711	0.4800702
Model 2	12.06626	0.5214259	0.5213608
Model 3	12.81629	0.4634344	0.4633857

We see that our preliminary Model works the best out of the 4 models we tested, mainly because of the R Squared Value for each model. Model 0 gives 55.01% R-Squared value.

- Multiple R squared: We get R squared value as 55.01, which is not very high. This value indicates that 54.89% of variance in our data set can be explained by this model. A value >80% is generally considered to be a good fit. ways to improve this R squared value are:

1. Add more variables in our model which are highly correlated to response variable. Since we have already taken the most correlated variables from our previous steps, this essentially means identifying parameters outside current dataset which might be able to better explain data variation.
2. Removing outliers - Outlier removal is a good way to improve R squared value but this also reduces datapoints which might be important to modeling market variance.

- Adjusted R Squared: Here Adj R squared values are almost equal to R squared and provide no additional information. Adj R squared basically improves itself upon addition of parameter only if the parameter improves the value. # Advanced Model Development

In real life, it is unlikely that a company would have sales and profit data of each market region before actually entering that market. In such cases, company usually have to build a model on regions it currently operates in, and then use that model to predict behaviour of other market. We would simulate one such scenario:

We assume that we only have sales data for EU and US market, and that we're trying to enter APAC market. We'll train our model using best features on EU and US market:

```
df_EUandUS<-subset(df, Market == "US" | Market == "EU")
df_APAC<-subset(df, Market == "APAC")

regNew<-lm(Profit~`Sub-Category`+Sales+`Shipping Cost`+Discount,data=df_EUandUS)
pred <- predict(regNew, newdata = df_APAC)

#test model accuracy
rmse <- sqrt(mean((df_APAC$Profit - pred)^2))
r_squared <- summary(regNew)$r.squared
print(paste("RMSE:", rmse))
```

```
## [1] "RMSE: 14.1962769437432"
```

```
print(paste("R Squared:", r_squared))
```

```
## [1] "R Squared: 0.564457366815323"
```

We see that RMSE for APAC data is 14.19. If we check the test data error when we initally built the model considering all the markets, it was 11.9. This shows that although the performance reduced when we tried to superimpose data from one market to other, it still performs reasonably with R Squared value - 56.44%

Outlier Data Analysis- Extended EDA

We now analyse Outlier data. Further analysis of high drop subcategories:

```
focused_data <- superstore %>%
  filter(`Sub-Category` %in% c("Bookcases", "Chairs", "Copiers", "Tables"))

remove_outliers_5 <- function(df) {
  quantiles <- quantile(df$Profit, probs = c(0.05, 0.95))
  df %>% filter(Profit > quantiles[1] & Profit < quantiles[2])
}

cleaned_data_5 <- focused_data %>%
  group_by(Market, `Sub-Category`) %>%
  do(remove_outliers_5(.))

remove_outliers_10 <- function(df) {
  quantiles <- quantile(df$Profit, probs = c(0.1, 0.9))
  df %>% filter(Profit > quantiles[1] & Profit < quantiles[2])
}

cleaned_data_10 <- focused_data %>%
  group_by(Market, `Sub-Category`) %>%
  do(remove_outliers_10(.))

remove_outliers_15 <- function(df) {
  quantiles <- quantile(df$Profit, probs = c(0.15, 0.85))
  df %>% filter(Profit > quantiles[1] & Profit < quantiles[2])
}

cleaned_data_15 <- focused_data %>%
  group_by(Market, `Sub-Category`) %>%
  do(remove_outliers_15(.))

# Apply the function to each market and sub-category combination


summary_data <- focused_data %>%
  group_by(Market, `Sub-Category`) %>%
  summarise(Count = n(),
            Average_Profit = mean(Profit),
            .groups = 'drop') %>%
  arrange(desc(Average_Profit))

summary_data_5 <- cleaned_data_5 %>%
  group_by(Market, `Sub-Category`) %>%
  summarise(Count = n(),
            Average_Profit = mean(Profit),
            .groups = 'drop') %>%
  arrange(desc(Average_Profit))

summary_data_10 <- cleaned_data_10 %>%
  group_by(Market, `Sub-Category`) %>%
  summarise(Count = n(),
```

```
            Average_Profit = mean(Profit),
            .groups = 'drop') %>%
  arrange(desc(Average_Profit))

summary_data_15 <- cleaned_data_15 %>%
  group_by(Market, `Sub-Category`) %>%
  summarise(Count = n(),
            Average_Profit = mean(Profit),
            .groups = 'drop') %>%
  arrange(desc(Average_Profit))
knitr::kable(
head(summary_data),
caption = "Profit vs Category for complete data"
)
```

Table 4: Profit vs Category for complete data

Market	Sub-Category	Count	Average_Profit
US	Copiers	68	817.9092
Canada	Copiers	13	204.8954
Canada	Tables	2	150.0900
APAC	Copiers	652	124.0093
EU	Copiers	465	120.8595
EU	Bookcases	484	116.5439

```
knitr::kable(
head(summary_data_5),
caption = "Profit vs Category for 5% Outlier Removed Data"
)
```

Table 5: Profit vs Category for 5% Outlier Removed Data

Market	Sub-Category	Count	Average_Profit
US	Copiers	60	521.4651
Africa	Tables	33	142.7157
Canada	Copiers	11	139.5682
EU	Bookcases	434	115.7690
EMEA	Tables	39	114.0805
EU	Copiers	417	112.4960

```
knitr::kable(
head(summary_data_10),
caption = "Profit vs Category for 10% Outlier Removed Data"
)
```

Table 6: Profit vs Category for 10% Outlier Removed Data

Market	Sub-Category	Count	Average_Profit
US	Copiers	51	478.57895
Africa	Tables	29	137.69266
EU	Bookcases	386	112.11706
EU	Copiers	371	104.00662
APAC	Copiers	520	101.40692
EMEA	Tables	35	95.70686

```
knitr::kable(
head(summary_data_15),
caption = "Profit vs Category for 15% Outlier Removed Data"
)
```

Table 7: Profit vs Category for 15% Outlier Removed Data

Market	Sub-Category	Count	Average_Profit
US	Copiers	46	419.77385
Africa	Tables	25	142.94964
EU	Bookcases	338	106.81133
EU	Copiers	325	98.06867
EMEA	Tables	31	95.02142
APAC	Copiers	456	94.96350

Important Findings: 1. Always High Performers: a. US Copiers continually exhibit high average profitability at all outlier removal levels, albeit the average profit declines with increasing outlier removal. This suggests that the US market for Copiers is healthy and may be fueled by profitable high-end sales that continue to do well even after being tamed by eliminating extremes. b. Additionally, Copiers in the EU and APAC frequently rank among the top results, indicating robust demand and profitability in these countries as well.

2. Outlier Removal on Profits:

a. As more outliers are eliminated, Bookcases and Copies in a variety of markets typically see a decline in average profit. This implies that there are a sizable proportion of highly profitable sales in these subcategories that are regarded as anomalies.
b. Tables in Africa and EMEA exhibit intriguing trends, indicating that tables can be lucrative even after a sizable amount of outlier data has been eliminated. This suggests steady demand and perhaps profitable operations.

3. Decrease in Counts:

a. It is assumed that the elimination of outliers will reduce the number of transactions (data points) for each subcategory, however the effect on profit measures varies. Even with fewer transactions, certain categories manage a comparatively high average profit, indicating that the eliminated outliers were, in fact, extreme values that were either too high or too low.

4. Outlier Removal on Market:

a. The original data had 28 different combinations of Market and Subcategories. However, due to the outlier removal, we lost 1 combination, which is Tables in Canada. It has to be mentioned that only two Tables were sold.

Conclusion and suggestion: 1. The regular appearance of Copier in different Market locations suggest this subcategory possibility to be accepted in any market regardless of price, strategies or product itself.

2. Bookcases, Tables and Chairs have to be further analyzed to see other factors affect. Nevertheless, companies with these subcategories served should focus on pricing to be well accepted in any Market. It has to be mentioned that is not a case for Africa market where tables on top 2 by the average profit and Canada with their highest average profit in Chair subcategory.

a. Only 7 Tables out of 37 was sold in Africa with a discount rate of 0.7. Therefore, it is required to use outlier removal of at least 5% to cover some of these purchases.
b. 1284 Bookcases products out of 2411 was sold with a discount across all Markets. The highest discount rate again happened in Africa, and therefore outlier removal is a must.

3. Chairs appear on top 10 list by the average profit only in original data. Outlier removal shifted their appearance in top 10 suggesting that an outstanding or low sales happened. Further investigation into the discount and pricing has to be done.

a. All Chair sales in Canada was without the discount.

Results and Analysis

We completed:

- EDA for our dataset
- Identified key features to model one variable - Profit
- Experimented with model trained on one category data and used it to predict data for other value of the category.

We identified the following most important features:

- Categorical Variables: Market, Product Sub-Category
- Numeric variables: Sales, Shipping Cost, Discount

We achieved best performance or R-Squared = 55.01% with Model 0.

We analysed outlier data as well in depth.

Limitations of Models Developed:

Limited Predictive Power: a. Unexplained Variance: Despite achieving an R-squared value of 55.01%, there's still a significant portion of the variance in profit that the model doesn't account for. This unexplained variance could stem from factors not included in the model or from inherent unpredictability in profit dynamics.

b. Complex Real-world Dynamics: Profitability in a real-world business setting is influenced by multi-faceted interactions among numerous variables, such as market trends, consumer behavior, competitor actions, and macroeconomic conditions. Capturing all these dynamics accurately in a model is challenging, leading to incomplete predictions.

c. Data Limitations: The model's predictive power is constrained by the quality, quantity, and representativeness of the data used for training. If the dataset doesn't encompass the full range of scenarios or lacks granularity, the model may struggle to generalize to new situations.

Model Complexity: a. Overfitting: As models become increasingly complex, they run the risk of overfitting the training data. Overfitting occurs when a model captures noise or random fluctuations in the training data, rather than genuine underlying patterns. This can lead to poor performance when applied to unseen data.

b. Interpretability: More complex models, such as neural networks or ensemble methods, may offer higher predictive accuracy but often sacrifice interpretability. It becomes challenging to understand how the model arrives at its predictions, making it harder to gain insights or trust its outputs.

c. Computational Resources: Complex models typically require more computational resources, both in terms of processing power and memory. This can pose practical challenges, especially for deployment in resource-constrained environments or real-time applications.

d. Training and Maintenance: Complex models may necessitate more extensive training and ongoing maintenance efforts. They often involve fine-tuning numerous hyperparameters and monitoring for performance degradation over time. This increases the complexity of model development and management.