Global Super Store - Project Proposal

2024-03-20

Group - 5

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.

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

Link to Data set - https://data.world/asepetruk/global-superstore

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

Research Question

If the company plans to release a new product in the market to maximize profits:

- Which category of product should it choose?
- Which region should it target the strongest?
- What kind of customer should it look towards?

Once developed, company would also like to know:

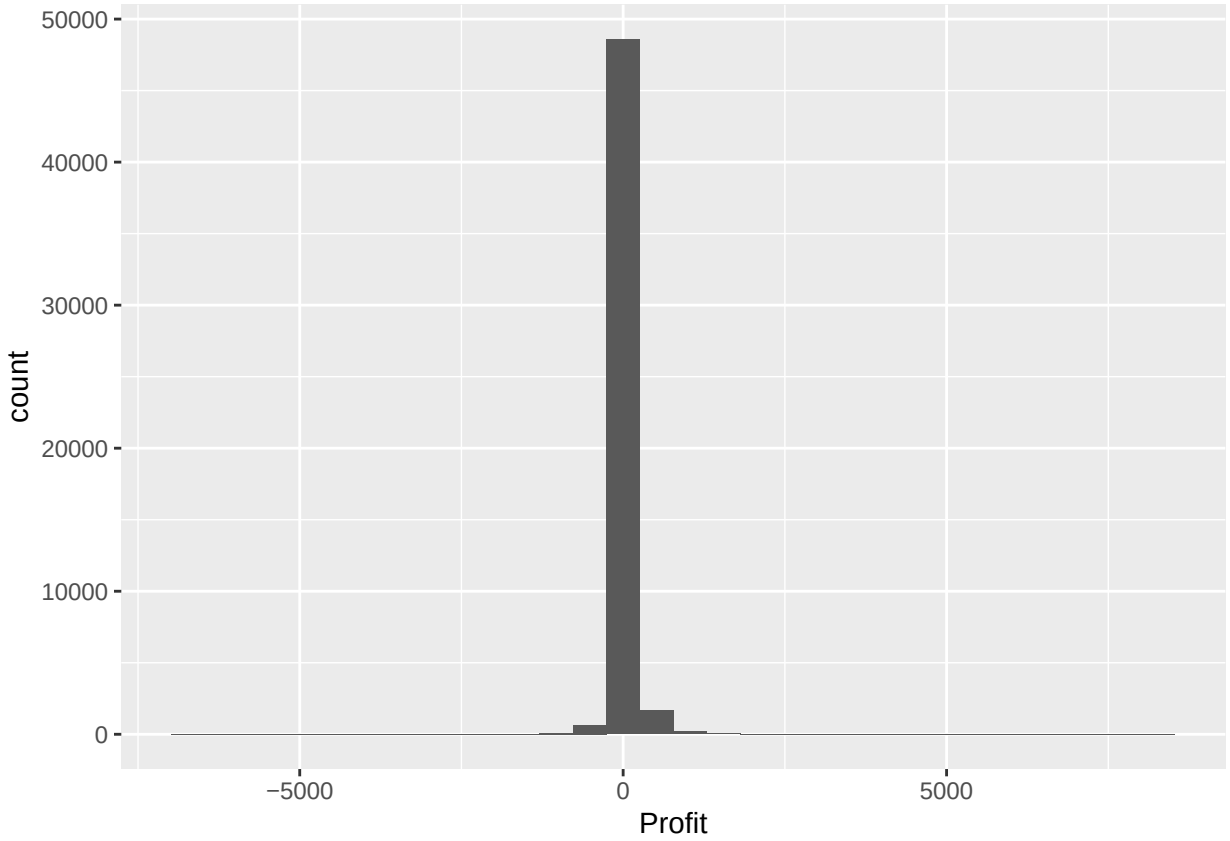
- When would be the best timing to release products to maximize profits?
- What is the optimal quantity to sell in bulk to customers in terms of shipping costs/ profits?

Plots to visualise Data

We can create few plots to better understand data and provide preliminary information

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

ggplot(data=superstore)+
  geom_histogram(mapping = aes(x= Profit))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

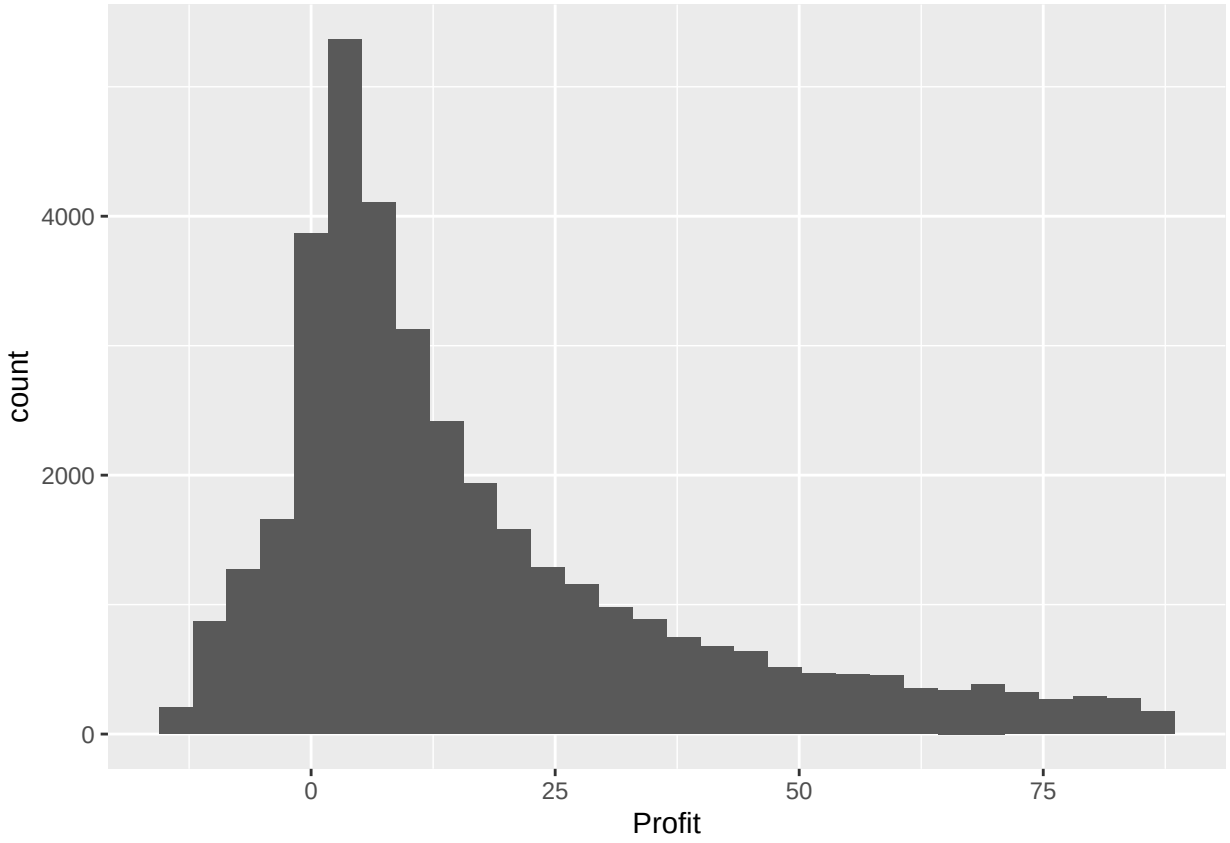


We see that the Profit column of the dataset seems to be concentrated at 0, with some extremely high profit

or extremely high loss points. It seems data has outliers which have to be dealt separatelt. For the purpose of proposal, we will remove the points lower than 15% quartile and above 85% quartile.

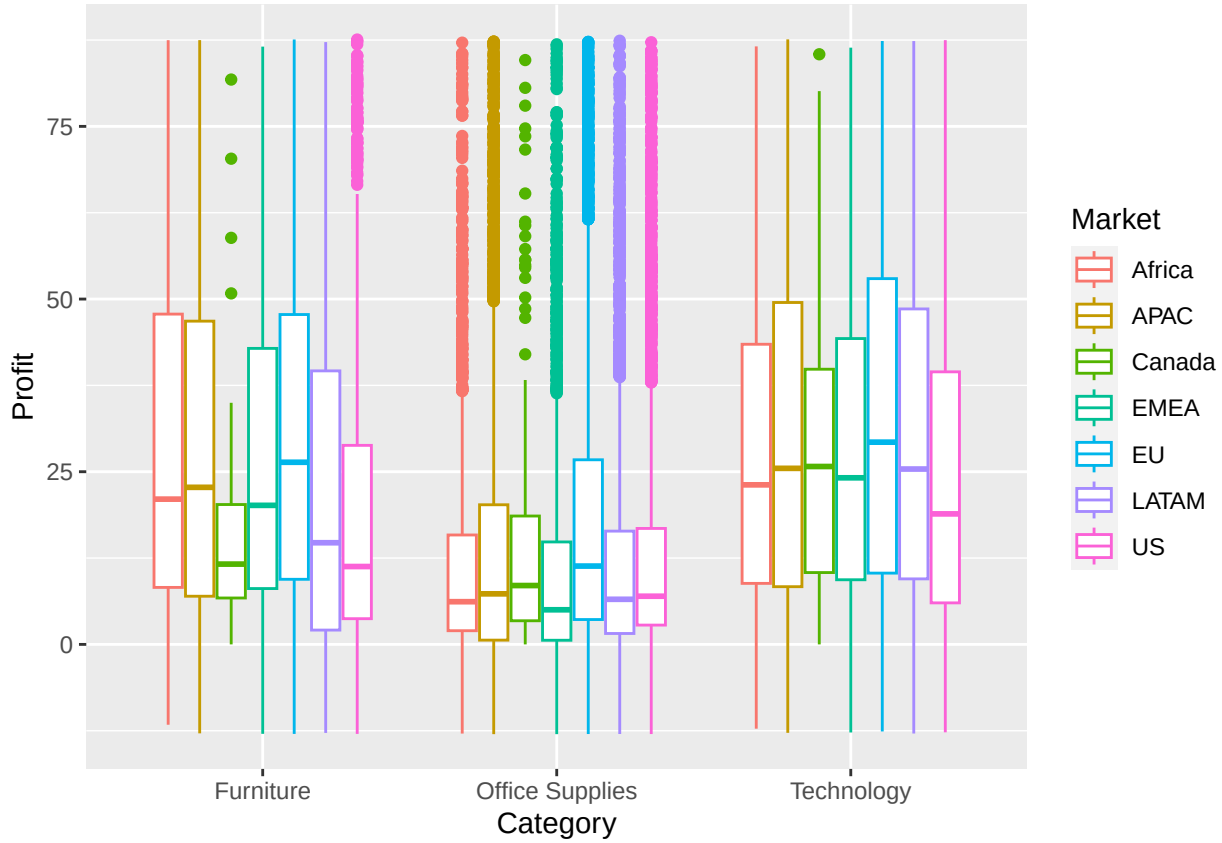
```
df<-superstore[ superstore$Profit > quantile(superstore$Profit , 0.15 ) , ]
df<-df[ df$Profit < quantile(df$Profit , 0.85 ) , ]

ggplot(data=df)+
  geom_histogram(mapping = aes(x= Profit))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Profit is now distributed more evenly distributed. We will not completely neglect these points, we will deal with separately as to why some products have high profits and high loss. Now we can analyse profit with each categorical variable clearly:

```
ggplot(data=df)+
  geom_boxplot(mapping = aes(x= Category,y = Profit,color = Market))
```



We can clearly see dependence of Profit on Category of Product. Further segregating each category into different markets where they were sold, we can see that Furniture has high dependency of profit over market than other category.

These plots give us a starting point in modeling profit given the other parameters. As we dive deep into the variables moving forward, we will find more such relations

We now move to fitting a preliminary linear regression model for Profit

Preliminary Model

```
lm <- lm(Profit ~ Category, data = df)
summary(lm)
```

```
##
## Call:
## lm(formula = Profit ~ Category, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.857 -12.598  -5.972   8.467  74.002
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               24.6777     0.2804   88.02   <2e-16 ***
```

```
## CategoryOffice Supplies -11.2793     0.3078  -36.64   <2e-16 ***
## CategoryTechnology        4.2772     0.3968   10.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.6 on 37053 degrees of freedom
## Multiple R-squared:  0.08307,    Adjusted R-squared:  0.08302
## F-statistic:  1678 on 2 and 37053 DF,  p-value: < 2.2e-16
```

Here Furniture is the reference category. We can see positive estimate of profit for Furniture and Technology but negative estimate for Office Supplies. Maybe selling Office Supplies might n=be non profitable than others. Our further analysis will look to test this hypothesis.

We can fit one more model of Profit with Market:

```
lm2 <- lm(Profit ~ Market, data = df)
summary(lm2)
```

```
##
## Call:
## lm(formula = Profit ~ Market, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.778 -14.318  -7.183   8.839  72.199
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.90451    0.36874  43.132  < 2e-16 ***
## MarketAPAC    2.49658    0.44560   5.603 2.13e-08 ***
## MarketCanada  1.15241    1.22713   0.939   0.3477
## MarketEMEA   -1.22316    0.51655  -2.368   0.0179 *
## MarketEU      4.89872    0.44833  10.927  < 2e-16 ***
## MarketLATAM   0.28247    0.44326   0.637   0.5240
## MarketUS      0.07931    0.44081   0.180   0.8572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.42 on 37049 degrees of freedom
## Multiple R-squared:  0.008697,   Adjusted R-squared:  0.008536
## F-statistic: 54.17 on 6 and 37049 DF,  p-value: < 2.2e-16
```

(Africa is reference category). EU market has a very high estimate of Profit and EMEA has negative estimate. We'll dive further into these categorical variables and look to segment our data even more for in-depth analysis

Future Analysis

As mentioned in the report:

- We will look to analyse Profits in depth with each categorical variable present.
- We will try to make profit into a categorical variable by marking it profitable or loss to simplify further analysis

- We will look to model other numerical variables to identify their relationship with other variables.
- Our final model will look to predict Profit with a combination of best parameters, determined by their correlation with profit and p-value of hypothesis testing.
- We look to build an interactive dashboard for our analysis to facilitate any researcher with their custom analysis.

If we find other datasets from different sources which might help in further studying our underlying dataset, we will include them also.