Mega store feature analysis

- **Null Values**

  None

- **Data description**

  1-Row ID:

  count 7995.000000

  mean 5010.902564

  std 2879.692377

Min 2.000000

max 9994.000000

- We drop this column cause of lack of efficiency

[Unique]

2- Order ID

Count 7995

Unique 4445

top CA-2017-100111

Freq 12

3-Order Date

Count 7995

Unique 1208

top 11/10/2016

Freq 3

Extract feature [year-month-day]

Continues data

- dropped after extraction other features from it (duration time)

4-Ship Date

Count 7995

Unique 1305

top 9/26/2017

Freq 29

Extract feature[year-month-day]

- dropped after extraction other features from it(duration time)

5-Ship Mode

Count 7995

Unique 4

Freq 4778

Discrete data [classified into 4 categories]

First Class 1212

Same Day 438

Second Class 1567

Standard Class 4778

- We drop this column cause of lack of efficiency

6-Customer ID

Count 7995

Unique 791

top PP-18955

Freq 31

Dropped because its lack of efficiency

7-Customer Name

Count 7995

Unique 791

Top Paul Prost

Freq 31

- We drop this cause it like customer id

8-Segment

Count 7995

Unique 3

Freq 4162

Discrete data [classified into 4 categories]

Consumer 4162

Corporate 2413

Home Office 1420

- We drop this column cause of lack of efficiency

9-Country

Count 7995

Unique 1

top United States

Feq 7995

One fixed value: United states

- We drop this column because there is One fixed value: United states

10-City

Count 7995

Unique 518

top New York City

Freq 738

Continues data

- We drop this column cause of lack of efficiency

11-State

Count 7995

Unique 49

top California

Freq 1593

Continues data

13- Product ID

Count 7995

Unique 1829

top TEC-AC-10003832

Freq 15

14-CategoryTree

Count 7995

Unique 17

- dropped after extraction other features from it [main category - subcategory]

15-Product Name

Count 7995

Unique 1816

Top Staples

Freq 40

- We drop this column cause of lack of efficiency we use product id instead

16- Sales

count 7995.000000

mean 228.211970

std 570.647158

min

0.444000

max 13999.960000

Continues data

17-Quantity

count 7995.000000

mean 3.764228

Std 2.204703

Min 1.000000

max 14.000000

Continues data

18-Discount

count 7995.000000

mean 0.155885

std

0.205622

Min 0.000000

Max 0.800000

Continues data

19-Region

count 7995

unique 4

top West

freq 2574

Discrete data [classified into 4 categories]

Central 1875

East 2248

South 1298

West 2574

20-[goal] Profit

count 7995.000000

mean 30.130534

std 215.924064

min -3839.990400

max 6719.980800

Discrete VS Continues

Discrete

[Ship Mode- segment – categoryTree - Region]

Continues

[Order Date - Ship Date – City – State - Postal Code – Sales – Quantity – Discount]

Correlation was used to select the highest effective features in the data set and are ['State' ,'MainCategory' ,'Product ID' ,'Region' ,'Sales' ,'Quantity' ,'Discount']

Correlation coefficients for train data:

```
Kendall correlation state: -0.082
Kendall correlation city : 0.058
Kendall correlation MainCategory : 0.144
Kendall correlation SubCategory : 0.014
Kendall correlation Region : 0.108
Kendall correlation Ship Mode : -0.011
Kendall correlation Segment : 0.006
Kendall correlation Order ID : -0.043
Kendall correlation Customer ID : -0.008
Kendall correlation Customer Name : -0.008
Kendall correlation Product ID : 0.131
Kendall correlation Product Name : 0.043
Kendall correlation Postal Code : 0.004
Spearman's correlation coefficient: 0.5175989232786377
Spearman's correlation coefficient: 0.24096240939002042
Spearman's correlation coefficient: -0.5366832764035246
Spearman's correlation coefficient: -0.007153716542985247
```

Correlation coefficients for test data :

```
Kendall correlation state: -0.104
Kendall correlation city : 0.069
Kendall correlation MainCategory : 0.119
Kendall correlation SubCategory : 0.010
Kendall correlation Region : 0.113
Kendall correlation Ship Mode : -0.024
Kendall correlation Segment : 0.053
Kendall correlation Order ID : -0.040
Kendall correlation Customer ID : -0.023
Kendall correlation Customer Name : -0.021
Kendall correlation Product ID : 0.115
Kendall correlation Product Name : 0.043
Kendall correlation Postal Code : -0.012
Spearman's correlation coefficient for sales: 0.5588185835487547
Spearman's correlation coefficient for quantity: 0.23285698014537573
Spearman's correlation coefficient dicount: -0.5314892552627561
Spearman's correlation coefficient time duration: -0.03161575368397155
```

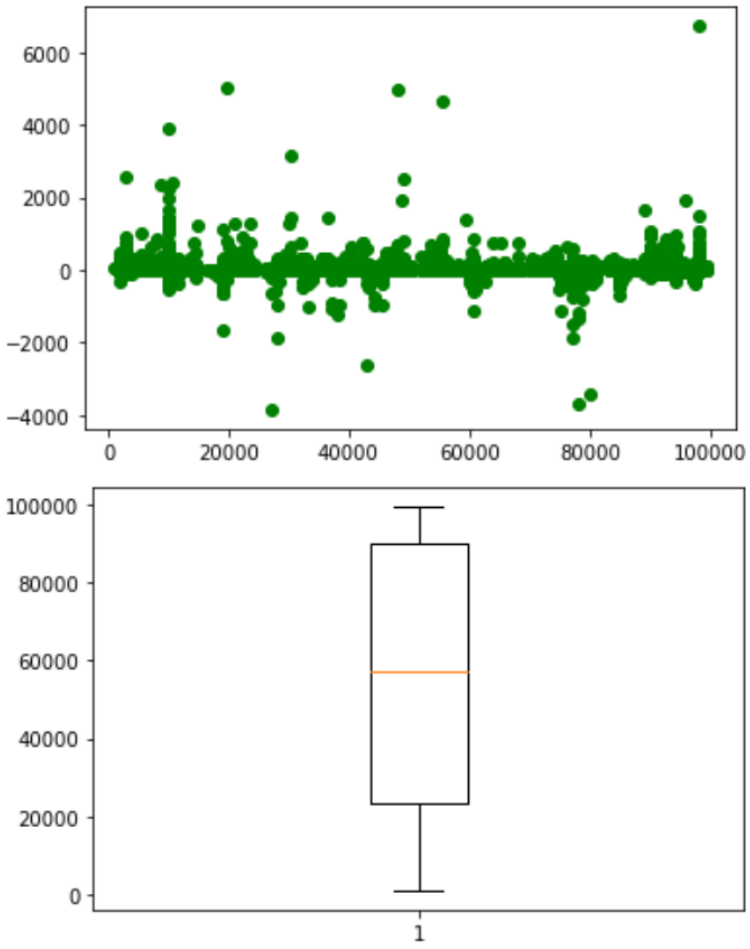
Same effictive features in both.

Scaling:

Max value: 99301.000000 & 13999.999999

Min Value: 0.4 & 0

1-Postal code

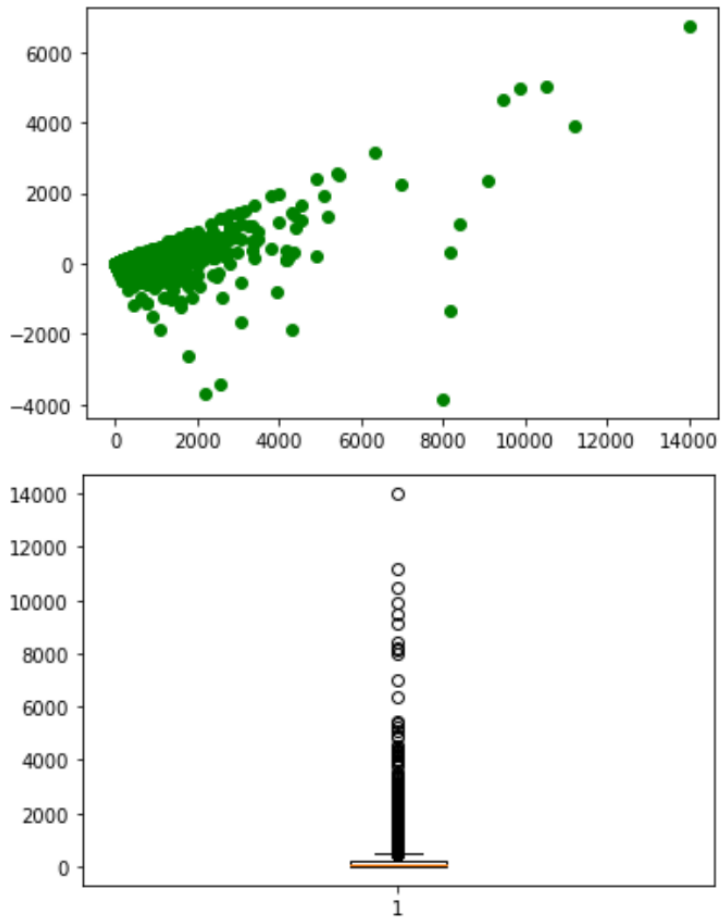


Outliers:

Outliers : mainly the y-axis causes it[Over 2500-Under -2000]

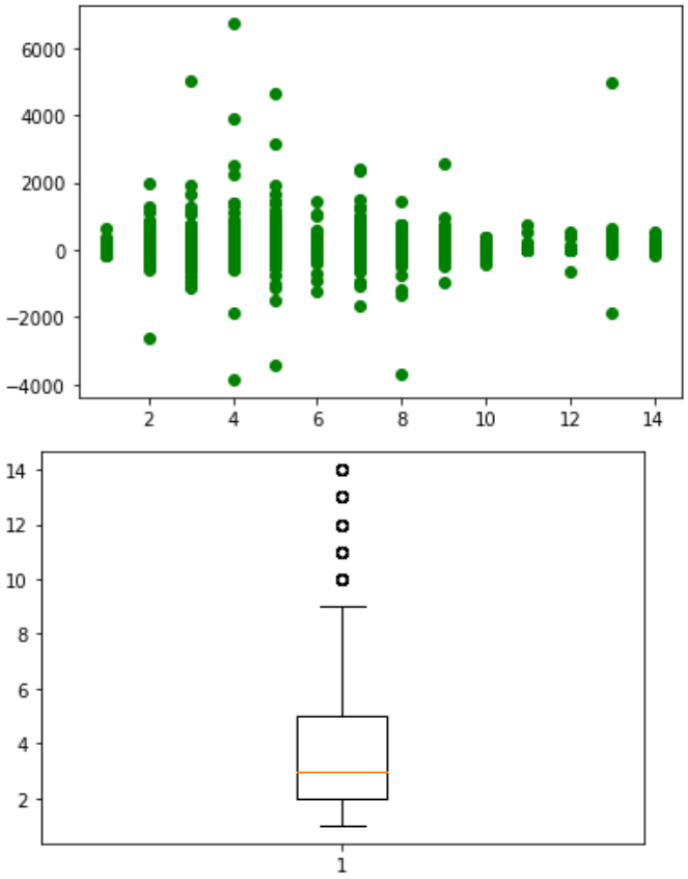
We removed the outliers using IQR and z transform

2- Sales



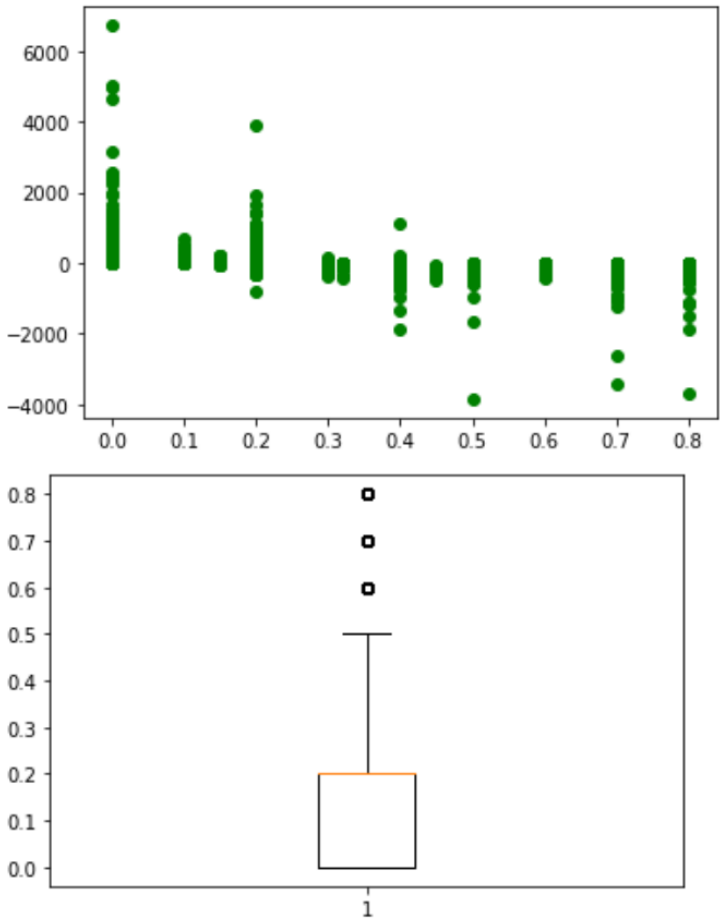
Outlier: mainly the x-axis [After 8000]

3-Quantity:



Outlier: mainly the y-axis [Over 3000 – Under -2000]x-axis [after 10]

4-Discount:



Mainly the y-axis [over 3800] x-axis[after 0.6]

Algorithms we use in the project

1-Linear regression

2-Polynomial regression

3-Lasso

4-Elastic

We splitted the data into training and testing sets before pre-processing :

Train 80%

Test 20%

The train data was then splitted into train and validation sets:

Train 75%

Validation 25%

Accuracy

we removed the scaling due to its lack of efficiency, as we see the accuracy with scaling

For train data

```
linear model cross validation score is 841.9695143448678
Mean Squared Error linear regression : 885.2414543560092
R score: 0.4133780039769157
polynomial model cross validation score is 489.0394135823791
Mean Square Error polynomial regression :  611.9429969874361
R score: 0.5944843967951474
 lasso model cross validation score is 1127.0656259994023
Mean Squared Error lasoo : 1262.1133992663588
R score: 0.1636366803182282
 lasso model cross validation score is 1127.0656259994023
Mean squared error of Elastic: 1419.6804304019086
R score: 0.059221906408425395
```

For test data:

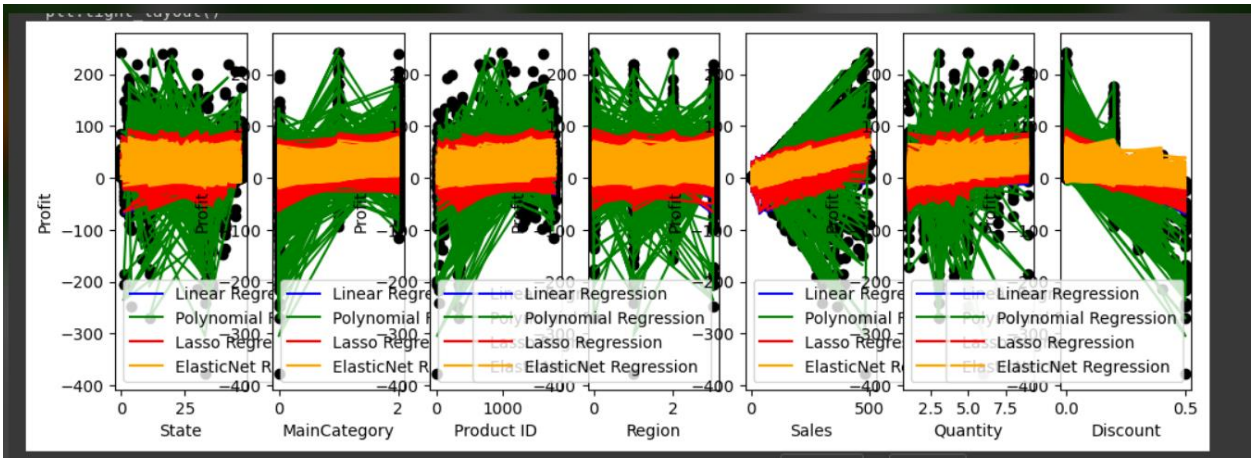
```
Kendall correlation Postal Code : -0.012
Mean Squared Error linear regression : 771.5515589379315
R2 score of linear regression: 0.29747340635606034
Mean Square Error polynomial regression :  2982.235266491376
R2 score of polynomial regression: -1.7154369127281566
Mean Squared Error lasoo : 952.4466125672359
R2 score of lasso regression: 0.13276168442244374
Mean squared error of Elastic: 1059.1713100664679
R2 score of elastic regression: 0.03558484986970967
```

Without scaling:

For test data

```
Mean Squared Error linear regression : 712.1385821226789
R2 score of linear regression: 0.35157114711848103
Mean Square Error polynomial regression :  14329.936592525053
R2 score of polynomial regression: -12.04794400952035
Mean Squared Error lasoo : 759.759934742511
R2 score of lasso regression: 0.3082101218530019
Mean squared error of Elastic: 911.8213588331373
R2 score of elastic regression: 0.1697524996066211
```

Regression plots with the selected features:



Our intuition was that the polynomial regression model would perform better than the linear regression model, since it can capture non-linear relationships between the features and the target variable. We also expected that the Lasso and ElasticNet models would perform better than the linear regression model, since they can help reduce overfitting by adding a penalty term to the loss function. Our results showed that the polynomial regression model had the lowest MSE and highest R-squared score on the test data, followed by the ElasticNet and Lasso models. The linear regression model had the highest MSE and lowest R-squared score. This confirmed our intuition that the polynomial regression model would perform better than the linear regression model, and that the Lasso and ElasticNet models would help reduce overfitting. Overall, this phase of the project was successful in finding a

regression model that fit the MegaStore dataset with minimum error (regression). By comparing different regression techniques, we were able to identify the best model and gain insights into the relationships between the features and the target variable.