TEAM ID : Sc_18

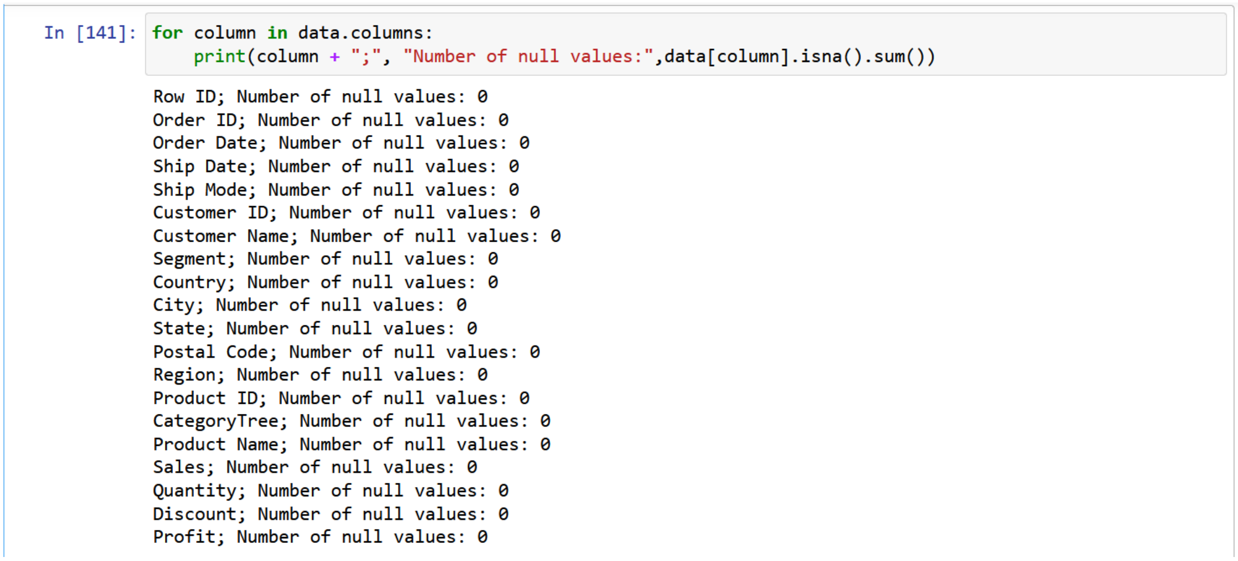
ID	Student Name	Department
20201701051	احمد محمد عبد الحميد محمد	Sc
20201700030	احمد تامر السيد علي	Sc
20201700077	احمد محمد عبد الهادي محمد	Sc
20201700425	طارق محمد سعد السيد	Sc
20201700236	خالد محمود شعبان محمود	Sc
20201700641	مازن حماده بدر محمد	Sc
20201700375	سهيل ابي محمد	Sc

Megastore Profit Regression Report

A- Overview Of the data and Big Picture

The dataset has 7995 rows and 20 columns including the profit.

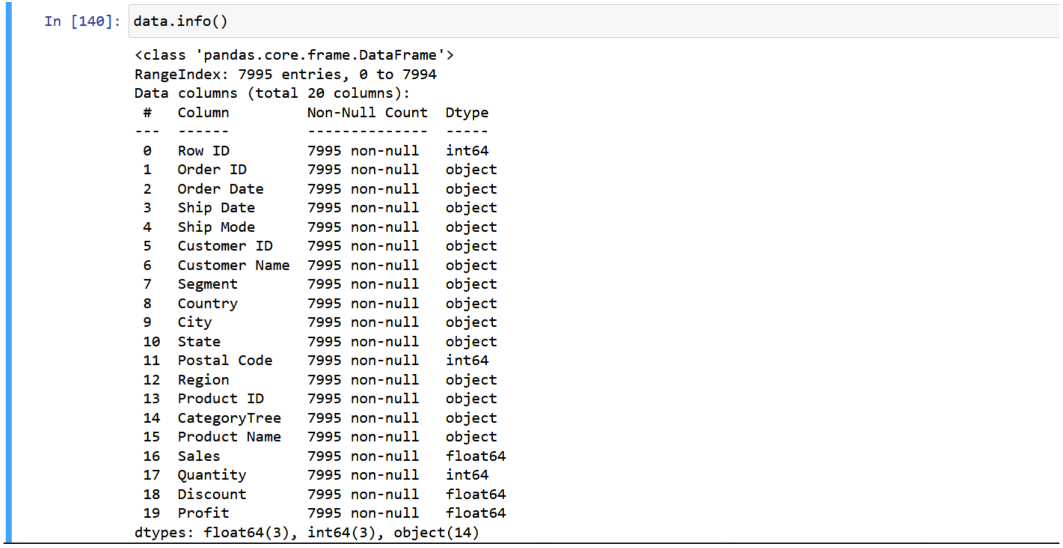
- **The data has no null values.**



```
In [141]: for column in data.columns:
              print(column + ";", "Number of null values:",data[column].isna().sum())
```

```
Row ID; Number of null values: 0
Order ID; Number of null values: 0
Order Date; Number of null values: 0
Ship Date; Number of null values: 0
Ship Mode; Number of null values: 0
Customer ID; Number of null values: 0
Customer Name; Number of null values: 0
Segment; Number of null values: 0
Country; Number of null values: 0
City; Number of null values: 0
State; Number of null values: 0
Postal Code; Number of null values: 0
Region; Number of null values: 0
Product ID; Number of null values: 0
CategoryTree; Number of null values: 0
Product Name; Number of null values: 0
Sales; Number of null values: 0
Quantity; Number of null values: 0
Discount; Number of null values: 0
Profit; Number of null values: 0
```

The datatypes of the columns are (float – int – object)



```
In [140]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7995 entries, 0 to 7994
Data columns (total 20 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Row ID         7995 non-null   int64  
 1   Order ID       7995 non-null   object 
 2   Order Date     7995 non-null   object 
 3   Ship Date      7995 non-null   object 
 4   Ship Mode      7995 non-null   object 
 5   Customer ID    7995 non-null   object 
 6   Customer Name  7995 non-null   object 
 7   Segment        7995 non-null   object 
 8   Country        7995 non-null   object 
 9   City           7995 non-null   object 
 10  State          7995 non-null   object 
 11  Postal Code    7995 non-null   int64  
 12  Region         7995 non-null   object 
 13  Product ID     7995 non-null   object 
 14  CategoryTree   7995 non-null   object 
 15  Product Name   7995 non-null   object 
 16  Sales          7995 non-null   float64
 17  Quantity       7995 non-null   int64  
 18  Discount       7995 non-null   float64
 19  Profit         7995 non-null   float64
dtypes: float64(3), int64(3), object(14)
```

B-Feature Encoding and analyzing the meaning of each feature.

1- Order Date:

It represents the date that the order was checked out.

- Converted into (day – month – year) and (day – month) converted into (month sin – month cos – day sin - day cos) by Cyclical encoding to preserve its periodic nature and be more representable for the model.

2- Ship Date:

It represents the date that the order was shipped.

- Converted into (day – month – year) and (day – month) converted into (month sin – month cos – day sin - day cos) by Cyclical encoding to preserve its periodic nature and be more representable for the model.

Delay days:

An extracted feature from the difference between the ship date and order date.

3- Ship mode:

It represents the type of shipping of the order.

- Encoded according to the type of shipping (Standard - First – Second – corporate) and the encoding according to the frequency of each class by visualization.

4- Order ID:

It represents the ID of each order and it's unique.

- Splitted to country code part, year part and the id part then the country code encoded by ordinal encoding and then the three parts got combined again together.

5- Segment:

It represents the type of the customer and its scale.

- Encoded according to the type of Customer (Home Office - consumer – – same day) and the encoding according to the frequency of each class by visualization.

6- Customer ID:

It represents the id of each Customer and it's unique.

- Splitted into letters part (The first letters of first and second name of the Customer) and id part, Encoded the letters part by ordinal encoding and then combined again.

7- Customer name:

The name of the customer that put the order.

- Encoded by ordinal encoder then dropped because the id represents it very well and id is numerical by its nature, it is more suitable.

8- Country:

Has 1 unique value (United stated) so it is encoded by 0 and then dropped because it's single value feature.

9- City:

It represents the residential city of the customer.

- Encoded by ordinal encoding then dropped because postal code represents it.

10- State:

It represents the residential state of the customer.

- Encoded by ordinal encoding then dropped because postal code represents it.

11- Postal Code:

It represents the postal code of the residential area of the customer.

12- Region:

It represents the residential region of the customer.

- Encoded according to the type of Region (East - West – South – Central) and the encoding according to the frequency of each class by visualization.

13- Product ID:

- Splitted to Main Category code part, Subcategory code part and the id part then the Main Category code and Subcategory code parts encoded by ordinal encoding and then the three parts got combined again together.

14- Category Tree:

It contains a list that has the Main category and Subcategory of each product.

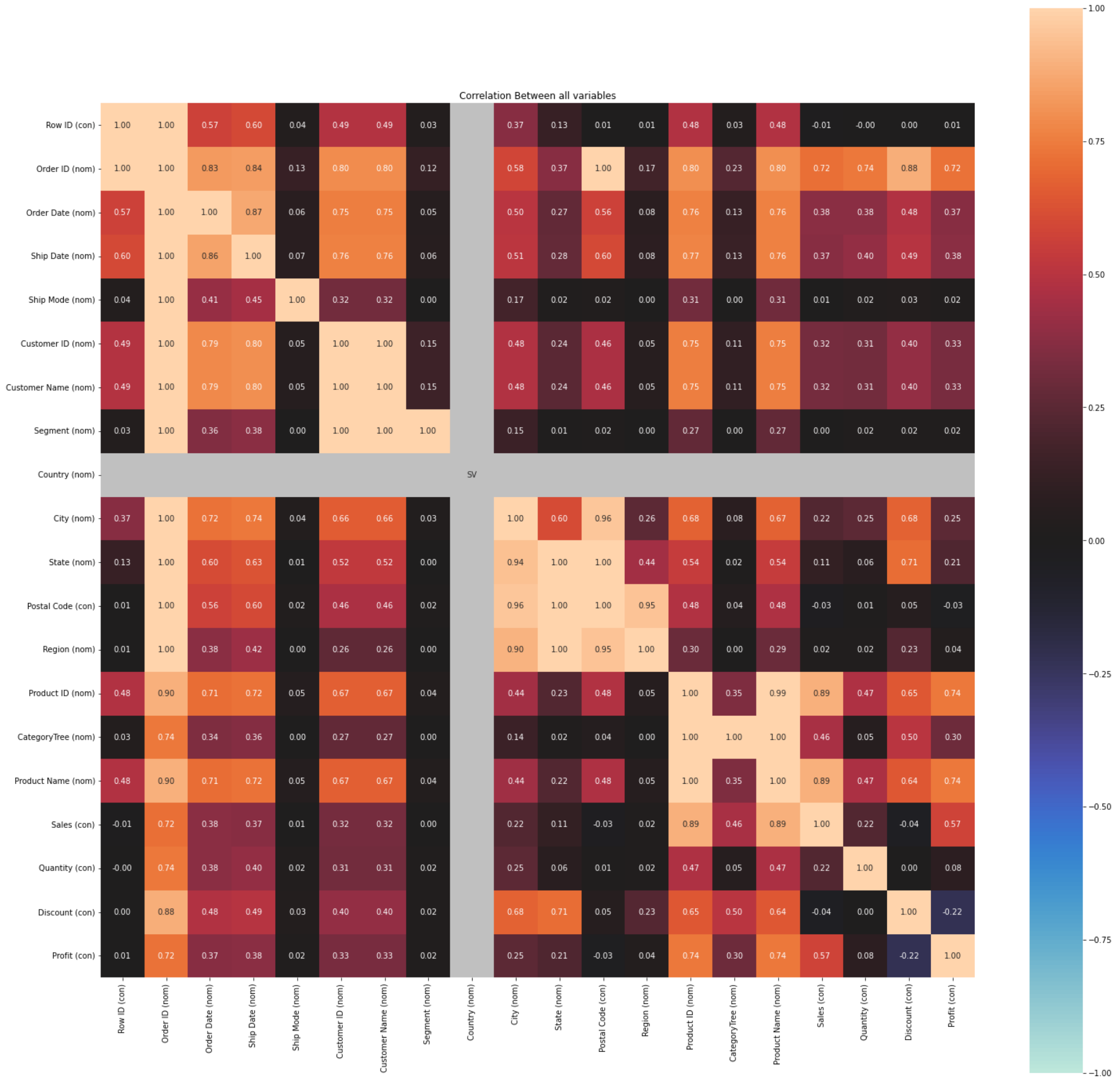
- Extracted the information of the main category and subcategory by Regex code.

15- Product Name:

It represents the name of the product.

- Encoded by ordinal encoding and then dropped because the Product ID represents it very well and the ID is numerical by its nature.

C- Analyzing correlation among the features.



- **It gives us a good representation of the correlation between each two features together to detect which features are high correlated**.

D- Regression Models and feature selection and splitting the data.

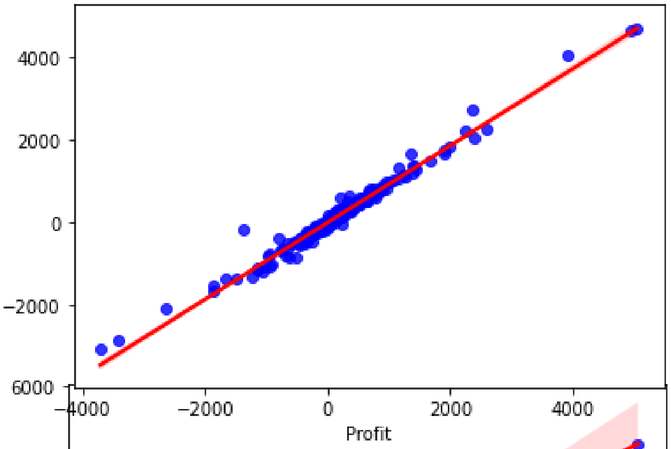
We Splitted the data into 70% training which equals 5596 and 30% test which equal 2399 rows.

1- Random Forrest model

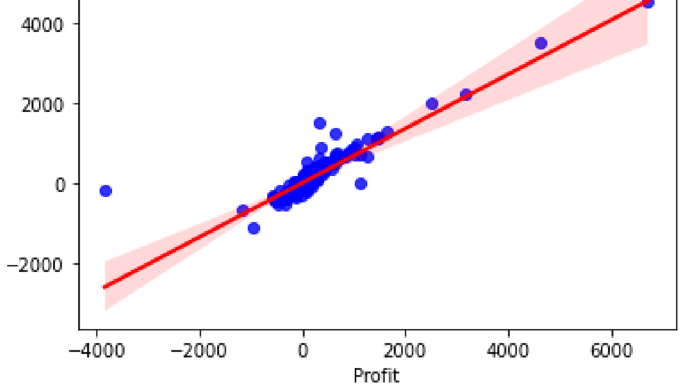
- We used Wrapper Forward selection algorithm to select most suitable features with the Random Forrest model to apply regression on the profit.
- We made the feature selection with cross validation of 5 parts.

- We selected 8 features that give us the best fitting line on the training data ('Sales', 'Discount', 'encoded ship mode', 'labeled Segment', 'Labeled Country', 'product main', 'product sub', 'ProductName')

- The train R2 Score of the Random Forrest model with those features is 97.55%



- The test R2 Score of the Random Forrest model with those features is 79.52%

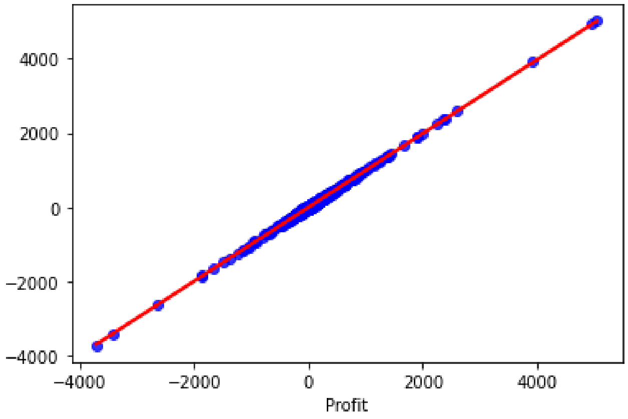


2- XGboost model

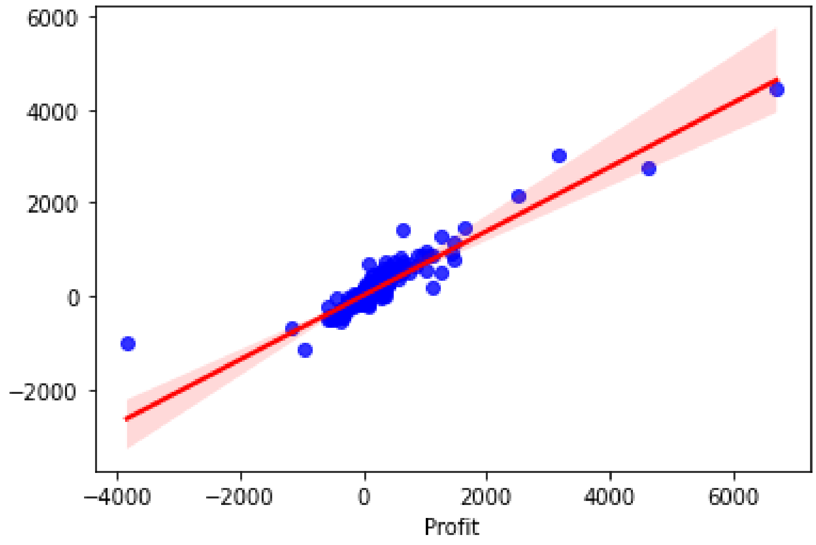
- We used Wrapper Forward selection algorithm to select most suitable features with the XGboost model to apply regression on the profit.
- We made the feature selection with cross validation of 5 parts.

We selected 8 features that give us the best fitting line on the training data ('Sales', 'Discount', 'Ship day_sin', 'Ship day_cos', 'Labeled Country', 'labeled Region', 'productmain', 'productsub')

- The train R2 Score of the XGboost model with those features is 99.96%



- The test R2 Score of the Random Forrest model with those features is 82.36%



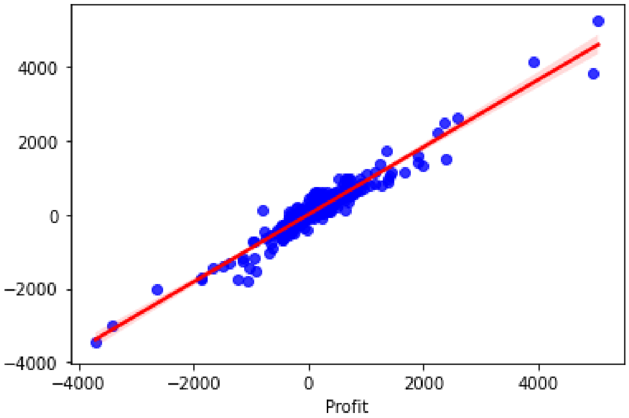
3- Polynomial Regression model

- We used Wrapper Forward selection algorithm to select the most suitable features with the polynomial regression model to apply regression on the profit.
- We made the feature selection with cross validation of 5 parts.

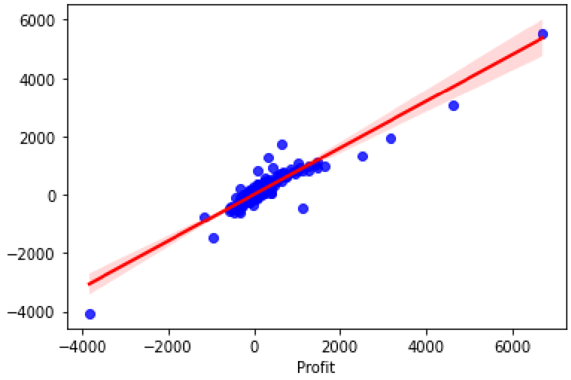
- We selected 10 features that give us the best fitting line on the training data ('Sales productnum', 'Sales Discount^2', 'Sales Ship month productmain', 'Sales Order month Final Product ID', 'Sales Order year encoded ship mode', 'Sales Ship month_cos Labeled Customer names', 'Sales Ship day_cos Order month_sin', 'Sales encoded ship mode productsub', 'Sales labeled Region productmain', 'Sales productsub^2')

- The selected features were transformed into polynomial features with degree = 3

- The train R2 Score of the XGboost model with those features is 91.54%



- The test R2 Score of the Random Forrest model with those features is 86.25%



E- Conclusion on this phase

- **The best regression model on this dataset is the Polynomial regression.**
- **We assumed that the features nature is linear with the profit then the scores of our models proved this assumption.**

	Random Forrest	XGboost	Polynomial Regression
Train R2 Score	**97.55%**	**99.96%**	**91.54%**
Test R2 Score	**79.52%**	**82.36%**	**86.25%**